Ciencia de Datos II

Guía de estudio

110 preguntas y respuestas desarrolladas

Material elaborado por el profesor

Sergio Gevatschnaider

Segundo cuatrimestre de 2026

Preguntas 1 a 100 · Temas del parcial

Preguntas 101 a 110 · Complemento de autoencoders

Presentación y forma de estudio

Esta guía desarrolla los temas indicados para el parcial mediante 100 preguntas y respuestas, y añade un complemento de diez preguntas sobre autoencoders. Incluye fundamentos, fórmulas explicadas, interpretación, ejercicios numéricos resueltos, ejemplos de Python y criterios para justificar decisiones.

Las respuestas amplían y reorganizan los contenidos y cuestionarios de los módulos del repositorio de la materia. Los ejemplos numéricos se elaboraron para esta guía y representan situaciones didácticas, no resultados empíricos de empresas. La selección cubre el programa proporcionado; las preguntas no se presentan como una reproducción de un examen específico.

Para estudiar, intentá responder primero sin mirar y después reconstruí el razonamiento. En un ejercicio, escribí datos, fórmula, cálculo, unidades e interpretación. En una pregunta conceptual, explicá mecanismo, condiciones de uso y límites. Cuando compares modelos, indicá cómo separarías la selección de la evaluación y qué costo o decisión representa la métrica.

El índice permite acceder a cada bloque y pregunta. Las figuras acompañan conceptos donde la representación facilita comprender el mecanismo. Los fragmentos de Python se explican en sus respuestas y requieren las variables indicadas; no se presentan como programas completos de un proyecto. Las fuentes y la correspondencia con los módulos se encuentran al final.

Cobertura del cuestionario

Bloque	Preguntas	Alcance
Análisis Exploratorio de Datos	1 a 10	Estructura, distribuciones, visualización, segmentos y ejercicios
Datos faltantes	11 a 20	Mecanismos de ausencia, imputación, calidad y ejercicios
Outliers	21 a 30	Detección, robustez, influencia, métodos y tratamiento
Transformación y regresión lineal	31 a 40	Escala, forma, codificación, correlación, OLS y métricas
Validación cruzada	41 a 50	Regresión, regularización, pipelines y generalización
Series de tiempo	51 a 60	Componentes, modelos, backtesting y pronósticos
Regresión logística	61 a 70	Probabilidades, interpretación, métricas, costos y decisión
Árboles	71 a 80	CART, impureza, regresión, complejidad y poda
Random forest	81 a 90	Bagging, bootstrap, OOB, diversidad e interpretación
Clustering	91 a 100	K-Means, jerárquico, PCA, validación y segmentación
Autoencoders	101 a 110	Complemento de arquitectura, aprendizaje y aplicaciones

Las preguntas 1 a 100 constituyen el cuestionario de los temas del parcial. Las preguntas 101 a 110 constituyen el complemento solicitado sobre autoencoders. Los ejercicios prácticos están integrados dentro de cada bloque, en lugar de separarse de las explicaciones que permiten resolverlos.

La notación usa x para predictores, y para respuesta observada y ŷ para predicción. El símbolo Σ indica suma; ln es logaritmo natural; n representa cantidad de observaciones. En las respuestas se define la notación específica adicional. Los decimales de la explicación se expresan con coma, mientras el código Python utiliza punto decimal.

Índice de bloques y preguntas

Los títulos y números de página del índice permiten navegar dentro del PDF. Los marcadores incluyen los once bloques y sus 110 preguntas.

Presentación y forma de estudio 2

Cobertura del cuestionario 3

1 Análisis Exploratorio de Datos 9

1. ¿Qué es el EDA y cómo se relaciona con la formulación del problema? 9
2. ¿Cómo se audita la estructura de una base antes de calcular estadísticas? 9
3. ¿Por qué el tipo de variable determina las medidas y los gráficos adecuados? 10
4. ¿Cómo se interpretan media, mediana, dispersión y cuantiles? 11
5. ¿Qué diferencias existen entre exploración univariada, bivariada y multivariada? 12
6. ¿Qué muestran un histograma, una densidad y un boxplot, y qué pueden ocultar? 12
7. ¿Cómo se reconocen asimetría, colas pesadas y mezclas de poblaciones? 13
8. ¿Qué es la paradoja de Simpson y cómo se demuestra con un ejemplo? 13
9. Ejercicio resuelto: ¿cómo se analiza una caída de ventas con ticket creciente? 14
10. ¿Cómo se redacta un informe de EDA útil para estudiar y decidir? 14

2 Datos faltantes y calidad de datos 15

11. ¿Qué es un dato faltante y cómo se diferencia de un cero o una ausencia estructural? 15
12. ¿Qué significan MCAR, MAR y MNAR y por qué importan? 15
13. ¿Cómo se diagnostican los patrones de ausencia en una base? 16
14. ¿Cuándo conviene eliminar filas o columnas y qué riesgos tiene esa decisión? 16
15. ¿Cómo funcionan la imputación por media, mediana y moda, y cuáles son sus límites? 17
16. ¿Qué aportan la imputación por grupo, KNN y modelos iterativos? 17
17. ¿Qué es la imputación múltiple y cómo incorpora incertidumbre? 18
18. ¿Qué utilidad tiene un indicador de ausencia y cuándo puede causar problemas? 18
19. Ejercicio resuelto: ¿cómo se imputa sin filtrar información del test? 19

20. ¿Cómo se evalúa una estrategia de tratamiento de faltantes? 19

3 Outliers robustez e influencia 20

21. ¿Qué es un outlier y por qué no debe eliminarse automáticamente? 20

22. ¿Cómo se distinguen anomalías univariadas, multivariadas, locales y contextuales? 20

23. Ejercicio resuelto: ¿cómo se detectan valores mediante la regla IQR? 21

24. ¿Cómo se comparan z-score y MAD para detectar extremos? 21

25. ¿Qué diferencias existen entre residuo, leverage e influencia en regresión? 22

26. ¿Cómo funcionan Mahalanobis y los métodos multivariados de referencia? 22

27. ¿Cómo detectan anomalías Isolation Forest, LOF y One-Class SVM? 23

28. ¿Qué opciones existen para tratar outliers y cómo se elige entre ellas? 23

29. Ejercicio resuelto: ¿cómo cambian los resúmenes y el costo con un extremo? 24

30. ¿Cómo se evalúa un detector y se convierte su score en una decisión? 24

4 Transformación de variables correlación y regresión lineal 25

31. ¿Qué diferencia existe entre escalar, transformar la forma y normalizar filas? 25

32. ¿Cuándo conviene usar logaritmos, Box-Cox o Yeo-Johnson y qué ocurre al transformar el objetivo? . 26

33. ¿Cómo se codifican categorías, interacciones y variables temporales sin introducir relaciones falsas? 26

34. ¿Qué miden covarianza, Pearson y Spearman y cómo se calculan? 27

35. ¿Por qué correlación no implica causalidad y qué es la multicolinealidad? 27

36. ¿Qué es la regresión lineal y qué optimiza mínimos cuadrados? 28

37. Ejercicio resuelto: ¿cómo se estima e interpreta una regresión simple? 28

38. ¿Qué supuestos y diagnósticos deben revisarse en regresión lineal? 29

39. Ejercicio resuelto: ¿cómo se interpretan MAE, MSE, RMSE y R^2? 29

40. ¿Cómo se interpretan modelos con logaritmos, polinomios e interacciones? 30

5 Regresión lineal validación cruzada y generalización 31

41. ¿Para qué se separan entrenamiento, validación y test? 31

42. ¿Cómo funciona K-Fold y qué se aprende de sus resultados? 32

43. ¿Cómo se elige K y qué diferencias hay con holdout, LOOCV y repeticiones? 33

44. ¿Qué son subajuste, sobreajuste y el compromiso entre sesgo y varianza? . . . 33
45. ¿Cómo se usan Ridge, Lasso y Elastic Net dentro de la validación? . . . 34
46. ¿Cómo evita un pipeline el leakage durante la validación cruzada? . . . 35
47. ¿Cuándo se necesitan particiones por grupos o por tiempo? . . . 36
48. ¿Por qué la búsqueda de hiperparámetros puede ser optimista y qué resuelve la validación anidada? 36
49. Ejercicio resuelto: ¿cómo se comparan dos regresiones con scores de CV? . . . 37
50. ¿Cómo se diseña y comunica un experimento completo de regresión? . . . 37

6 Series de tiempo y ejercicios de pronóstico . . . 38

51. ¿Qué distingue una serie de tiempo de una muestra transversal? . . . 38
52. ¿Qué son tendencia, estacionalidad, ciclo y residuo? . . . 38
53. ¿Qué es estacionariedad y qué aportan diferencias y transformaciones? . . . 39
54. ¿Qué representan ACF, PACF y los componentes de ARIMA y SARIMA? . . . 39
55. ¿Qué baselines temporales deben compararse antes de adoptar un modelo complejo? . . . 40
56. ¿Cómo se construyen rezagos y medias móviles sin leakage temporal? . . . 40
57. ¿Cómo funcionan el backtesting y las ventanas expansivas o móviles? . . . 41
58. ¿Qué métricas son apropiadas para pronósticos y por qué MAPE puede fallar? . . . 41
59. Ejercicio resuelto: ¿cómo se evalúa un pronóstico y se elige un cuantil de inventario? . . . 42
60. ¿Cómo se diseña un proyecto completo de series de tiempo y se diagnostican fallas? . . . 42

7 Regresión logística y decisiones de clasificación . . . 43

61. ¿Qué estima la regresión logística y por qué utiliza una sigmoide? . . . 44
62. ¿Qué son odds, logit y odds ratio y cómo se relacionan con probabilidad? . . . 45
63. ¿Cómo se interpretan coeficientes, efectos marginales e interacciones? . . . 45
64. ¿Cómo se estima la regresión logística mediante verosimilitud y log-loss? . . . 46
65. ¿Qué papel tienen regularización, escalado y desbalance de clases? . . . 46
66. Ejercicio resuelto: ¿cómo se calculan las métricas de una matriz de confusión? . . . 47
67. ¿Qué diferencias existen entre ROC-AUC y precision-recall? . . . 47
68. ¿Qué diferencia hay entre discriminación y calibración? . . . 48
69. ¿Cómo se elige un umbral según costos y capacidad? . . . 48

70. ¿Cómo se extiende la logística a varias clases y qué se monitorea en producción? 49

8 Árboles de decisión y ejercicios prácticos 50

71. ¿Cómo representa un árbol de decisión un problema de clasificación o regresión? 50

72. ¿Qué miden Gini y entropía y cómo elige CART una división? 50

73. Ejercicio resuelto: ¿cómo se calcula la ganancia de una división con Gini? 51

74. ¿Cómo construye predicciones y divisiones un árbol de regresión? 51

75. ¿Cómo afectan escala, categorías e interacciones al comportamiento de un árbol? 52

76. ¿Cómo se controla el sobreajuste con profundidad y tamaño de nodos? 52

77. ¿Qué es la poda por costo-complejidad y cómo se selecciona su intensidad? 53

78. ¿Por qué un árbol puede ser inestable y qué consecuencias tiene? 53

79. ¿Cómo se evalúan e interpretan hojas, probabilidades e importancia de variables? 54

80. Ejercicio resuelto: ¿cómo se decide entre un árbol pequeño y uno profundo? 54

9 Random forest y ejercicios prácticos 55

81. ¿Qué es bagging y por qué puede mejorar a un árbol individual? 55

82. ¿Qué añade random forest al bagging de árboles? 55

83. Ejercicio resuelto: ¿qué proporción de casos queda fuera de una muestra bootstrap y qué es OOB? . 56

84. ¿Cómo explican la varianza y la correlación la ventaja de combinar árboles? 57

85. ¿Qué controlan los principales hiperparámetros de random forest? 58

86. ¿Cómo se comparan importancia por impureza e importancia por permutación? 58

87. ¿Qué muestran PDP e ICE y qué límites tienen con predictores correlacionados? 59

88. Ejercicio resuelto: ¿por qué promediar probabilidades puede diferir del voto mayoritario? 59

89. ¿Qué diferencia hay entre bagging, random forest y boosting? 60

90. ¿Cómo se desarrolla una práctica completa de random forest sin perder interpretabilidad? 60

10 Clustering K-Means y jerárquico 61

91. ¿Qué es clustering y cómo se diferencia de clasificación? 61

92. ¿Cómo funciona K-Means y qué función objetivo minimiza? 61

93. Ejercicio resuelto: ¿cómo se realiza una iteración de K-Means? 62

94. ¿Por qué escala, redundancia y outliers afectan una segmentación? 62

95. ¿Cómo se elige K con inercia, silhouette y estabilidad? 63

96. ¿Cómo funciona el clustering jerárquico y cómo se interpreta un dendrograma? 64

97. ¿Cómo se comparan single, complete, average y Ward linkage? 65

98. ¿Qué aporta PCA antes de clustering y qué límites tiene? 65

99. ¿Cómo se valida y describe una segmentación sin inventar perfiles? 66

100. Ejercicio integrador: ¿cómo se comparan K-Means y jerárquico para segmentar clientes? 66

11 Autoencoders complemento de diez preguntas 67

101. ¿Qué es un autoencoder y por qué reconstruir la entrada puede producir aprendizaje útil? 67

102. ¿Qué funciones cumplen el encoder, el espacio latente, el bottleneck y el decoder? 68

103. Ejercicio resuelto: ¿cómo se calcula el error de reconstrucción y se elige una pérdida? 69

104. ¿Cómo se entrena un autoencoder y cómo se evita sobreajuste? 69

105. ¿Cómo se relaciona un autoencoder lineal con PCA? 70

106. ¿Qué diferencias existen entre autoencoders sparse, denoising y contractive? 70

107. ¿Cómo se usa un autoencoder para detectar anomalías y qué supuestos necesita? 71

108. ¿Cómo se preparan datos numéricos, categóricos y faltantes para un autoencoder? 71

109. ¿Qué distingue a un autoencoder variacional de uno convencional? 72

110. Ejercicio integrador: ¿cómo se diseña y evalúa un autoencoder de anomalías de principio a fin? . . 72

Fuentes y correspondencia con el repositorio 73

1 Análisis Exploratorio de Datos

1. ¿Qué es el EDA y cómo se relaciona con la formulación del problema?

El Análisis Exploratorio de Datos, o EDA, es un proceso de inspección, resumen y visualización que permite comprender qué representan los datos, cómo fueron producidos y qué conclusiones podrían sostener. Incluye dimensiones y tipos de variables, distribuciones, relaciones, calidad, segmentos y evolución temporal. Su propósito no es acumular gráficos, sino convertir un problema de decisión en preguntas que la evidencia pueda responder.

Antes de explorar hay que definir la unidad de análisis, la población de interés, el período y la decisión. Una fila puede representar un cliente, una compra o un cliente por mes; esa elección modifica el significado de promedios, frecuencias y relaciones. También debe distinguirse entre describir, explicar, predecir y evaluar una intervención. Predecir quién comprará no demuestra que una campaña le hará comprar.

Por ejemplo, ante una caída de ventas conviene preguntar si disminuyeron los clientes activos, la frecuencia o el ticket, y en qué canales ocurrió. Luego se examinan esas variables y sus componentes. El cierre del EDA debe expresar hallazgos con magnitudes, hipótesis pendientes y límites de cobertura. Si la base solo registra ventas realizadas, no permite medir directamente la demanda que se perdió por falta de stock.

El EDA orienta el modelado, pero no sustituye una evaluación independiente. Cuando una exploración determina transformaciones o selección de variables para un modelo, debe apoyarse en entrenamiento y reservar evidencia de evaluación que no haya guiado esas decisiones.

2. ¿Cómo se audita la estructura de una base antes de calcular estadísticas?

La auditoría comienza con la semántica: identificar qué representa cada fila, qué mide cada columna, su unidad, el momento de captura y su procedencia. Después se verifican cantidad de filas y columnas, tipos, claves, cardinalidades, rangos y reglas entre variables. Un identificador numérico no se convierte por eso en una cantidad sobre la cual tenga sentido calcular una media.

La clave permite determinar si dos registros representan el mismo evento. Si la unidad es una línea de factura, una factura puede repetirse legítimamente porque contiene varios productos. Eliminar esas repeticiones como duplicados destruiría información. En cambio, dos filas con el mismo identificador de línea y los mismos atributos pueden indicar una doble carga. La regla debe surgir del modelo de datos, no de la apariencia de la tabla.

Supongamos que se integran ventas con clientes mediante una unión. Si clientes contiene dos filas por identificador, la unión puede multiplicar las ventas y aumentar artificialmente la facturación. Se comprueba la cardinalidad esperada y se contrastan conteos y totales antes y después. También se revisan monedas, fechas, separadores decimales y códigos como 999 que podrían representar faltantes.

El resultado debe ser un diccionario de datos y un registro de incidencias con regla, cantidad afectada y tratamiento. La trazabilidad permite explicar por qué cambió la base. Conservar una versión original evita que una corrección irreversible impida reconstruir el análisis o descubrir que una supuesta anomalía era válida.

3. ¿Por qué el tipo de variable determina las medidas y los gráficos adecuados?

Una variable nominal distingue categorías sin orden, como canal de venta. Una ordinal incorpora orden, como satisfacción baja, media y alta, aunque no garantiza distancias iguales. Una cuantitativa discreta cuenta unidades, mientras que una continua mide magnitudes. Las fechas y horas tienen una estructura temporal que debe conservarse al construir variables.

Para categorías se utilizan frecuencias absolutas y relativas, moda y gráficos de barras. En una ordinal conviene respetar el orden y puede ser útil presentar porcentajes acumulados. Para cantidades se emplean medidas de centro, dispersión y cuantiles, junto con histogramas o diagramas de caja. Tratar códigos de categorías nominales como números genera distancias y promedios arbitrarios.

Por ejemplo, si se codifica efectivo = 1, tarjeta = 2 y transferencia = 3, una media de 2,4 carece de interpretación económica. Además, un modelo basado en distancias asumiría que efectivo y transferencia están más separados que efectivo y tarjeta, aunque esa relación fue inventada por la codificación. Una representación one-hot evita introducir ese orden.

El tipo almacenado y el tipo conceptual pueden diferir: un ingreso puede estar guardado como texto por incluir un símbolo monetario, y un código postal como entero. Por eso la conversión debe acompañarse de validaciones. Identificar correctamente el tipo permite elegir herramientas y, sobre todo, evitar que operaciones matemáticamente posibles respondan una pregunta sin significado.

4. ¿Cómo se interpretan media, mediana, dispersión y cuantiles?

La media es la suma de valores dividida por el número de observaciones. Resume el centro de gravedad y resulta útil para calcular totales esperados. La mediana divide los datos ordenados en dos mitades y cambia poco ante extremos. La elección depende del objetivo: para describir el cliente típico puede convenir la mediana; para presupuestar el ingreso total puede ser necesaria la media.

La dispersión muestra cuánto difieren las observaciones. La varianza promedia desviaciones cuadráticas respecto de la media, y el desvío estándar devuelve la unidad original. En una muestra se suele usar el divisor n - 1 para estimar la varianza poblacional. El rango intercuartílico, Q3 - Q1, describe el 50 % central y es más resistente a valores extremos. Los percentiles permiten estudiar también colas relevantes.

Con tickets de 10, 10, 10, 10 y 100, la media es 28 y la mediana 10. Ninguna está equivocada: describen aspectos distintos. Informar solo 28 ocultaría que cuatro de cinco operaciones fueron de 10. Para estudiar riesgo operativo conviene agregar percentiles altos y participación del mayor ticket en la facturación.

Dos grupos con igual media pueden tener dispersiones muy diferentes y requerir decisiones distintas. El coeficiente de variación, desvío dividido por media, compara variabilidad relativa cuando existe un cero significativo y la media es positiva y alejada de cero. No debe aplicarse automáticamente a temperaturas en Celsius o variables con medias próximas a cero.

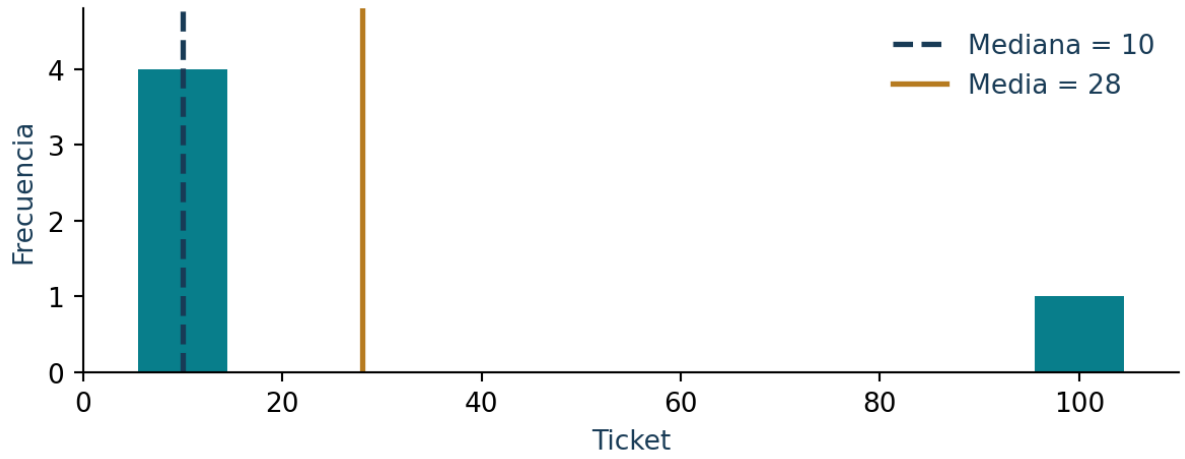


Figura 1. El extremo desplaza la media; la mediana conserva el centro de los tickets habituales. Datos de la pregunta 4.

5. ¿Qué diferencias existen entre exploración univariada, bivariada y multivariada?

La exploración univariada examina una variable por vez: categorías, concentración, asimetría, colas y valores faltantes. Es el punto de partida para entender escalas y detectar problemas básicos. La bivariada analiza relaciones entre dos variables, mientras que la multivariada estudia la estructura conjunta y el papel de varias características simultáneamente.

La herramienta bivariada depende de los tipos involucrados. Dos variables numéricas pueden compararse con un diagrama de dispersión; una numérica y una categórica, con distribuciones por grupo; dos categóricas, mediante una tabla de contingencia con porcentajes condicionados. El denominador importa: porcentaje de abandono dentro de cada canal y porcentaje de cada canal entre quienes abandonan contestan preguntas diferentes.

Un análisis multivariado puede mostrar que la relación entre gasto y satisfacción cambia según antigüedad y canal. También puede detectar variables redundantes, interacciones y observaciones con combinaciones inusuales. Sin embargo, incorporar muchas variables no elimina automáticamente confusión ni sesgos de selección. Aumenta la necesidad de justificar la especificación y controlar la interpretación.

Por ejemplo, una asociación global entre edad y gasto puede reflejar que las personas mayores se concentran en un canal de mayor ticket. Al separar canales, la relación podría debilitarse. El recorrido de lo simple a lo conjunto permite reconocer ese mecanismo. La conclusión debe indicar qué relaciones son marginales, cuáles son condicionadas y qué información faltaría para atribuir un efecto causal.

6. ¿Qué muestran un histograma, una densidad y un boxplot, y qué pueden ocultar?

El histograma agrupa una variable numérica en intervalos y representa frecuencias o densidades. Permite detectar asimetría, colas, huecos y posibles modos. Su forma depende del ancho y origen de los intervalos: pocos pueden esconder estructura y demasiados producir ruido. Si los intervalos tienen anchos diferentes, debe usarse una altura que haga que el área represente frecuencia.

Una densidad suavizada ofrece una representación continua cuya área total es uno. Facilita comparar formas, pero el ancho de banda modifica el resultado: un suavizado excesivo elimina modos y uno muy pequeño crea oscilaciones artificiales. Además, puede asignar densidad a valores imposibles cerca de un límite, como tiempos negativos, si no se ajusta el método.

El boxplot resume mediana, cuartiles y valores fuera de las cercas definidas por una regla. Los bigotes no son siempre mínimo y máximo: en la convención de Tukey llegan a las observaciones más extremas dentro de las cercas. Un punto señalado indica rareza bajo esa regla, no error comprobado. El boxplot tampoco revela bien una distribución con varios modos.

En tiempos de entrega, un histograma puede mostrar dos grupos que un boxplot resume en una sola caja. Conviene contrastar ambos y separar por tipo de servicio. Para comparar segmentos deben mantenerse unidades y escalas, informar tamaños muestrales y distinguir conteos de porcentajes. Una visualización útil permite reconstruir la pregunta y el denominador sin adivinarlos.

7. ¿Cómo se reconocen asimetría, colas pesadas y mezclas de poblaciones?

Una distribución es asimétrica cuando sus lados no tienen una forma equilibrada alrededor del centro. La asimetría positiva suele mostrar una cola extendida hacia valores altos y puede desplazar la media por encima de la mediana, aunque esa relación no es una regla universal. Las colas pesadas implican una mayor presencia de extremos respecto de una referencia, como la distribución normal.

Una mezcla de poblaciones combina grupos con procesos diferentes. Dos modos en un histograma pueden corresponder a clientes minoristas y corporativos; también pueden surgir de intervalos mal elegidos o de una muestra pequeña. Por eso se contrastan gráficos, cuantiles y variables de segmentación antes de concluir que hay grupos reales.

En ingresos, una cola derecha puede ser parte natural del fenómeno y no una falla de medición. Un logaritmo puede facilitar la representación y expresar diferencias multiplicativas, pero no garantiza normalidad ni vuelve erróneos los valores altos. Si la variable admite ceros, debe explicarse el tratamiento; sumar una constante cambia la interpretación y no es una decisión neutra.

El diagnóstico debe comparar centro, dispersión y extremos por segmento. Aplicar una regla simétrica a una distribución muy asimétrica puede producir falsas alertas. Asimismo, la normalidad de cada predictor no es un requisito general de la regresión lineal. Lo relevante es la relación modelada, el comportamiento de los errores y los supuestos de la tarea, no transformar todas las columnas hasta que parezcan gaussianas.

8. ¿Qué es la paradoja de Simpson y cómo se demuestra con un ejemplo?

La paradoja de Simpson aparece cuando una asociación agregada difiere, e incluso se invierte, respecto de las asociaciones dentro de grupos. El mecanismo es una composición desigual: las alternativas comparadas participan en distinta proporción en segmentos con dificultades o probabilidades de base diferentes.

Supongamos dos campañas. En clientes fáciles, A obtiene 81 respuestas de 90 contactos, es decir, 90 %; B obtiene 19 de 20, o 95 %. En clientes difíciles, A obtiene 2 de 10, o 20 %; B obtiene 24 de 80, o 30 %. B tiene mejor tasa en ambos segmentos. Sin embargo, A suma 83 respuestas de 100 contactos y B suma 43 de 100: al agregar, A parece superior.

La inversión surge porque A contactó principalmente clientes fáciles, mientras que B se concentró en difíciles. El promedio global incorpora ese reparto y no compara poblaciones equivalentes. Se puede contrastar usando pesos comunes. Si ambos segmentos pesan 50 %, A alcanza 55 % y B 62,5 %, preservando la ventaja de B en la comparación estandarizada.

La respuesta analítica no consiste en controlar indiscriminadamente cualquier variable. Debe identificarse su papel en el proceso: un confusor previo a la intervención, una consecuencia de ella o una variable de selección no se tratan de igual modo. El ejemplo muestra por qué segmentar y explicitar denominadores es indispensable, pero la atribución causal requiere además un diseño y supuestos defendibles.

9. Ejercicio resuelto: ¿cómo se analiza una caída de ventas con ticket creciente?

Una empresa pasa de 1.000 operaciones con ticket medio de 100 a 800 operaciones con ticket medio de 110. La facturación inicial es 100.000 y la final 88.000, por lo que cae 12 %. Las operaciones disminuyen 20 % y el ticket aumenta 10 %. Sus variaciones no se suman directamente: 0,80 × 1,10 = 0,88 explica el resultado conjunto.

Una descomposición secuencial separa mecanismos. Manteniendo el ticket inicial, la pérdida de 200 operaciones reduce ventas en 20.000. Aplicando luego el aumento de 10 a las 800 operaciones finales, se recuperan 8.000. El cambio neto es -12.000. El orden de la descomposición determina a qué componente se asigna el término de interacción; por eso debe declararse.

El EDA debería examinar clientes activos, frecuencia, unidades, precio y mezcla de productos por canal y período. Un ticket mayor puede indicar suba de precios, compras más grandes o pérdida de clientes de bajo gasto. Si hubo inflación, también interesa comparar valores a precios constantes mediante un índice adecuado. La mejora nominal del ticket no garantiza crecimiento real.

La conclusión defendible es que el ticket no compensó la reducción de operaciones. Atribuir la caída a una campaña exigiría comparar poblaciones y controlar otros cambios. Si faltan registros de un canal durante varios días, primero debe corregirse la cobertura. Una recomendación comercial basada en una base incompleta podría responder a una falla de captura y no a un deterioro del negocio.

10. ¿Cómo se redacta un informe de EDA útil para estudiar y decidir?

Un informe de EDA debe iniciar con la pregunta, la unidad de análisis, la población y el período. Luego presenta la calidad y cobertura de la base, las distribuciones relevantes y las relaciones que responden al problema. Cada gráfico debe incorporar un título informativo, unidades, denominadores y una interpretación que explique la magnitud observada.

Conviene distinguir tres niveles. El hallazgo describe lo observado: la tasa de cancelación fue 18 % en un canal y 11 % en otro. La hipótesis propone un mecanismo: diferencias de anticipación de reserva podrían explicar parte de la brecha. La recomendación plantea el siguiente paso: contrastar por anticipación, cohorte y tipo de cliente antes de modificar condiciones. Mezclar estos niveles convierte una descripción en una afirmación causal injustificada.

La evidencia debe incluir tamaños muestrales y límites. Una diferencia de tasas en grupos de diez casos es menos estable que en grupos de miles, y ambas pueden sufrir sesgos de selección. Es preferible presentar también dispersión o incertidumbre pertinente y aclarar qué casos quedaron fuera del análisis.

Para una entrega práctica, deben conservarse fuente, fecha de extracción, código y reglas de limpieza. El informe final relaciona cada decisión con evidencia, reconoce explicaciones alternativas y establece qué dato o prueba resolvería la incertidumbre. Estudiar EDA exige aprender ese razonamiento: calcular una media o construir un gráfico es una herramienta; justificar qué significa y qué permite hacer es la competencia central.

2 Datos faltantes y calidad de datos

11. ¿Qué es un dato faltante y cómo se diferencia de un cero o una ausencia estructural?

Un dato faltante es un valor no disponible para una variable que se esperaba observar. No equivale a cero: cero unidades vendidas es una medición válida, mientras que una venta sin cantidad registrada es desconocida. Tampoco equivale necesariamente a ausencia estructural, que ocurre cuando un atributo no aplica a la entidad considerada.

Por ejemplo, en viviendas, la superficie del garaje puede ser cero porque no existe garaje, desconocida porque no se midió o no aplicable porque el campo refiere a una característica de un garaje existente. Esas situaciones requieren un diccionario y reglas coherentes. Imputar la superficie media a una vivienda sin garaje crea un recurso inexistente y altera tanto la predicción como la interpretación.

Los faltantes pueden estar representados como NaN, cadenas vacías, espacios, códigos especiales o valores fuera de rango. Antes de cuantificarlos debe homogeneizarse su representación sin borrar diferencias semánticas. También interesa si faltan por columna, por fila, por grupo o en bloques temporales, porque cada patrón sugiere problemas distintos.

Una estrategia defendible conserva el dato original y añade, cuando corresponde, una variable que indique la razón de ausencia. El objetivo es reconstruir la información disponible y sus límites. Un tratamiento estadístico no reemplaza investigar por qué faltó el valor ni autoriza a presentar una estimación como si hubiera sido una observación real.

12. ¿Qué significan MCAR, MAR y MNAR y por qué importan?

MCAR significa ausencia completamente al azar: la probabilidad de faltar no depende de valores observados ni no observados. Una pérdida aleatoria de formularios por un fallo independiente de sus contenidos puede aproximarse a ese mecanismo. Bajo MCAR, analizar casos completos puede conservar representatividad, aunque reduce tamaño y precisión.

MAR significa ausencia al azar condicionada a la información observada. Si el ingreso falta más entre independientes y la ocupación está registrada, la probabilidad de ausencia puede depender de la ocupación, pero no del ingreso no observado una vez condicionada esa información. No significa que la ausencia sea totalmente aleatoria: exige modelar o considerar las variables observadas que explican el patrón.

MNAR significa que la ausencia sigue dependiendo del valor desconocido después de considerar lo observado. Si quienes tienen ingresos excepcionalmente altos omiten declararlos por esa razón, un modelo basado solo en el resto de las variables puede subestimar el extremo superior. Se necesitan hipótesis adicionales, recolección complementaria o análisis de sensibilidad.

Estos mecanismos son propiedades del proceso generador, no etiquetas que un gráfico pueda demostrar de forma definitiva. La asociación de ausencia con una variable observada hace poco plausible MCAR, pero no prueba MAR. En general, MAR y MNAR no pueden distinguirse solo con los datos observados sin supuestos adicionales. El diagnóstico debe formular una hipótesis y evaluar cuánto depende de ella la conclusión.

13. ¿Cómo se diagnostican los patrones de ausencia en una base?

Se calcula primero la proporción de faltantes por variable y por fila, y se identifica su concentración. Después se comparan tasas por segmento, fecha, canal, origen y versión del formulario. Un mapa de ausencia ayuda a reconocer columnas que faltan juntas, pero la interpretación exige revisar el proceso de captura.

Supongamos que faltan ingresos en 30 de 100 independientes y en 10 de 200 asalariados. Las tasas son 30 % y 5 %, aunque los conteos brutos no muestran por sí solos la intensidad del problema. El patrón sugiere una relación con ocupación. También conviene comparar otras características observadas de quienes declaran y no declaran, y estudiar si la brecha cambia por fecha o canal.

Si varias columnas comienzan a faltar después de una actualización del sistema, puede haber un problema de integración. Si faltan solo atributos de productos que no tienen esa característica, podría tratarse de ausencia estructural. Una entrevista con responsables de captura puede aportar evidencia que no aparece en la tabla.

Una prueba estadística puede detectar incompatibilidad con una hipótesis concreta, pero no certificar el mecanismo correcto. Además, con muestras grandes pueden aparecer diferencias pequeñas de escasa relevancia. El diagnóstico debe combinar magnitud, consecuencias y conocimiento del dominio. El producto final es una descripción del patrón, una hipótesis sobre su origen y un plan de tratamiento y sensibilidad.

14. ¿Cuándo conviene eliminar filas o columnas y qué riesgos tiene esa decisión?

Eliminar filas incompletas simplifica el análisis, pero puede reducir precisión y cambiar la población representada. Bajo MCAR, los casos completos constituyen una submuestra aleatoria del conjunto, lo que permite estimaciones descriptivas representativas en condiciones adecuadas. Bajo otros mecanismos, la validez depende del estimando y del modelo; no debe suponerse automáticamente.

Un porcentaje pequeño por columna puede implicar una pérdida grande al exigir completitud conjunta. Si diez variables faltan independientemente en 5 % de los casos cada una, la probabilidad de que una fila esté completa es 0,95 elevado a 10, aproximadamente 59,9 %. Se perdería alrededor de 40,1 % de las filas, mucho más que el 5 % inicialmente observado por variable.

Eliminar una columna depende de su relevancia, recuperabilidad, mecanismo de ausencia y disponibilidad futura. No existe un porcentaje universal que obligue a descartarla. Una variable con 80 % de ausencia podría ser muy informativa en un segmento específico; otra casi completa puede estar contaminada por información posterior a la decisión y ser inválida como predictor.

Antes de eliminar deben compararse perfiles y resultados con alternativas razonables. También conviene explicar qué población queda representada después de la exclusión. En un modelo de crédito, eliminar a quienes no declaran ingresos puede excluir precisamente a un grupo con necesidades o riesgo diferentes y producir una evaluación que no representa el uso real.

15. ¿Cómo funcionan la imputación por media, mediana y moda, y cuáles son sus límites?

La imputación simple reemplaza valores faltantes por una estadística estimada con datos observados. La media minimiza la suma de errores cuadrados respecto de una constante; la mediana minimiza la suma de errores absolutos; la moda selecciona la categoría más frecuente. Son estrategias económicas y útiles como referencia, pero no reconstruyen la variabilidad perdida.

La media suele funcionar como punto de partida en variables aproximadamente simétricas. La mediana resiste mejor extremos y asimetría. En categorías, la moda puede reforzar artificialmente la categoría dominante. Una categoría explícita de ausencia puede preservar información sobre el proceso, siempre que no se confunda con una modalidad sustantiva.

Por ejemplo, edades observadas de 20, 22 y 80 tienen media 40,67 y mediana 22. Imputar 40,67 o 22 a dos casos produce bases distintas y modifica relaciones con otras variables. Repetir una constante genera una concentración artificial y normalmente reduce variación cuando la ausencia es comparable al resto de los casos. No debe utilizarse el desvío de la base imputada como si todos los valores fueran observaciones independientes completas.

En predicción se evalúa la estrategia fuera de muestra, dentro de un pipeline. En inferencia también debe atenderse la incertidumbre por los valores desconocidos. La elección no se justifica solo porque desaparezcan los NaN: debe considerarse distribución, relaciones, desempeño y consecuencias sobre los grupos afectados.

16. ¿Qué aportan la imputación por grupo, KNN y modelos iterativos?

La imputación por grupo usa estadísticas condicionadas a categorías observadas, como mediana de ingresos por ocupación. Puede preservar heterogeneidad que una constante global borra. Requiere grupos con suficiente información y una regla de respaldo para categorías nuevas o sin observaciones. Los grupos deben construirse con variables disponibles al momento de decidir.

KNN imputa a partir de observaciones próximas según variables disponibles. Puede capturar similitudes, pero depende de escala, elección de vecinos y definición de distancia. En muchas dimensiones, la proximidad puede perder significado; además, variables ausentes o ruidosas dificultan encontrar vecinos adecuados. Escalar con entrenamiento y revisar casos donantes son parte del procedimiento.

La imputación iterativa modela una variable incompleta utilizando otras y repite el proceso para diferentes columnas. Permite aprovechar relaciones multivariadas, pero su resultado depende del modelo, restricciones y supuestos. Una regresión sin límites puede imputar edades negativas, por lo que hay que verificar validez y coherencia entre atributos.

Un método más complejo no es automáticamente superior. Conviene comparar con una estrategia simple utilizando el mismo protocolo de evaluación y estudiar errores por segmento. Si se ocultan artificialmente valores conocidos para medir reconstrucción, ese patrón de ocultamiento puede diferir del mecanismo real de ausencia. El ejercicio ayuda a comparar técnicas, pero no demuestra que se recuperarán correctamente los verdaderos valores desconocidos.

17. ¿Qué es la imputación múltiple y cómo incorpora incertidumbre?

La imputación múltiple genera varios conjuntos completos plausibles en lugar de reemplazar cada faltante una única vez. En cada conjunto se realiza el análisis y luego se combinan estimaciones e incertidumbres. Así se representa que el valor desconocido podría tomar varias realizaciones compatibles con información y supuestos.

Si se generan m bases, cada una produce una estimación Q y una varianza U. El promedio de las estimaciones es $\bar{Q}$. La varianza dentro de imputaciones, W, es el promedio de U; la varianza entre imputaciones, B, mide la dispersión de las estimaciones Q. Bajo el marco habitual de Rubin, la varianza total es $T = W + (1 + 1/m)B$. El segundo término incorpora incertidumbre por ausencia.

Por ejemplo, si $W = 4$, $B = 1$ y $m = 5$, $T = 5{,}2$ y el error estándar es aproximadamente 2,28. Usar solo W daría 2 y subestimaría incertidumbre. No basta con repetir una imputación determinista idéntica: la variación debe surgir de un procedimiento probabilístico apropiado.

La validez depende de que el modelo de imputación, el análisis y las hipótesis de ausencia sean compatibles. No resuelve MNAR automáticamente. En predicción su implementación debe respetar la separación de entrenamiento y evaluación; en inferencia es especialmente útil cuando interesa estimar parámetros y sus intervalos. La complejidad exige justificar qué incertidumbre se incorporó y cuál permanece fuera del modelo.

18. ¿Qué utilidad tiene un indicador de ausencia y cuándo puede causar problemas?

Un indicador de ausencia toma valor uno cuando el dato original falta y cero cuando está observado. Se crea antes de imputar para conservar una información que el reemplazo eliminaría. El modelo puede utilizar tanto el valor imputado como el hecho de que el campo no fue completado.

En una solicitud de crédito, no declarar ingreso puede relacionarse con tipo de empleo, dificultades de documentación o cambios del formulario. El indicador permite capturar esa asociación predictiva, pero no explica su causa. Tampoco convierte una imputación incorrecta en una estimación correcta del ingreso desconocido.

El beneficio puede desaparecer si el proceso cambia. Si una aplicación nueva obliga a completar el campo, un modelo que dependía fuertemente de la ausencia recibirá una distribución distinta. Por eso deben monitorearse las tasas de faltantes y la estabilidad del desempeño. También hay que asegurar que el indicador pueda calcularse de la misma manera en producción.

Usarlo puede reproducir diferencias de acceso al proceso de captura. Conviene evaluar errores por grupos y verificar si representa una característica relevante o una deficiencia corregible del sistema. Además, si el campo solo se completa después de conocer el resultado, su ausencia o presencia podría revelar el objetivo y generar leakage. El indicador necesita el mismo examen de disponibilidad temporal que cualquier predictor.

19. Ejercicio resuelto: ¿cómo se imputa sin filtrar información del test?

Los ingresos de entrenamiento son 20, 30, 40 y un faltante. La mediana observada es 30, por lo que se imputa ese valor. En test aparecen 100 y un faltante. El valor desconocido de test también se reemplaza por 30: la regla fue aprendida exclusivamente con entrenamiento y se aplica sin recalcularla.

Si se usaran todos los valores observados, 20, 30, 40 y 100, la mediana sería 35. Ese cambio muestra que test influyó en la representación utilizada para entrenar. Aunque no se haya consultado su variable objetivo, existe filtración porque se incorporó información de la distribución reservada para evaluar.

En validación cruzada, el imputador debe ajustarse nuevamente dentro de cada fold de entrenamiento. No alcanza con imputar una vez antes de dividir en folds. Un pipeline coloca imputación, escalado y modelo dentro del procedimiento que recibe cada partición; la transformación de validación usa las estadísticas aprendidas en su entrenamiento correspondiente.

El mismo principio se aplica a imputación por grupo, selección de variables y transformaciones ajustadas. Las reglas fijas de dominio, como reconocer un código oficialmente documentado de ausencia, son distintas de estadísticas aprendidas. El objetivo es reproducir qué habría sabido el sistema ante un caso nuevo. Una evaluación contaminada puede parecer mejor o simplemente cambiar de forma imprevisible; en ambos casos pierde independencia.

20. ¿Cómo se evalúa una estrategia de tratamiento de faltantes?

La evaluación debe considerar validez de los valores, distribución, relaciones, desempeño y consecuencias operativas. Primero se verifica que las imputaciones respeten unidades y restricciones: superficies no negativas, categorías admitidas y coherencia entre atributos. Después se comparan histogramas, cuantiles y relaciones relevantes antes y después, sin esperar que la distribución imputada replique mecánicamente la observada.

En predicción se comparan eliminación, imputación simple y alternativas más complejas con las mismas particiones. El imputador se ajusta dentro de entrenamiento y se estudian métricas globales y por patrón de ausencia. Un aumento promedio de desempeño puede ocultar errores mayores en el grupo que más faltantes tiene.

También debe hacerse sensibilidad. Si los ingresos desconocidos pudieran ser 20 % mayores que lo imputado, se recalculan resultados bajo ese escenario y se observa si cambia la decisión. Esto no identifica el mecanismo verdadero, pero muestra la dependencia respecto de supuestos. En una estimación inferencial conviene incorporar además la incertidumbre asociada a los faltantes.

La estrategia elegida debe incluir una respuesta al origen del problema. Si la ausencia proviene de un fallo de captura, imputar es una solución provisional; corregir el proceso puede aportar más valor. Se documentan regla, variables utilizadas, parámetros, versión, proporción afectada y límites. El criterio final es si la decisión sigue siendo defendible con información incompleta y supuestos explícitos.

3 Outliers robustez e influencia

21. ¿Qué es un outlier y por qué no debe eliminarse automáticamente?

Un outlier es una observación poco compatible con un patrón de referencia. Ese patrón puede ser una distribución, una relación entre variables, una densidad local o un comportamiento temporal. La rareza estadística no demuestra invalidez: puede señalar un error, un evento legítimo, otro segmento o un cambio de régimen.

Una compra corporativa veinte veces mayor que el ticket habitual puede ser válida y relevante para el negocio. Una edad de 250 años, en cambio, contradice una restricción del dominio. Ambos casos podrían recibir una alerta numérica, pero las acciones difieren. Primero se revisan unidades, origen, reglas, temporalidad y posibilidad de duplicación; después se evalúa impacto analítico.

Eliminar valores altos porque empeoran una métrica puede producir un modelo que nunca aprendió los casos costosos. Si son errores comprobados, se corrigen o excluyen con una regla documentada. Si son válidos, pueden conservarse, transformarse, segmentarse o tratarse con métodos robustos. La respuesta debe depender de la tarea y la población que se pretende representar.

La decisión se contrasta con un análisis de sensibilidad: cuánto cambian parámetros, métricas o recomendaciones con tratamientos alternativos. Conservar el registro original y la justificación permite auditar el proceso. Detectar es identificar una señal; tratar requiere explicar qué significa y qué costo tendría actuar equivocadamente sobre ella.

22. ¿Cómo se distinguen anomalías univariadas, multivariadas, locales y contextuales?

Una anomalía univariada es extrema en una columna. Una multivariada aparece en la combinación de atributos, aunque cada valor aislado sea común. Por ejemplo, un gasto de 5.000 y una capacidad de pago de 500 pueden encontrarse separadamente en la base, pero su combinación podría resultar poco compatible con el patrón habitual.

Una anomalía global se compara con toda la población. Una local se compara con vecinos o con una región de densidad semejante. Si hay un grupo compacto y otro disperso, un punto aparentemente cercano al conjunto puede estar aislado respecto de su entorno inmediato. Los detectores globales y locales pueden dar respuestas distintas sin que uno esté necesariamente mal.

La anomalía contextual depende de condiciones como fecha, estación, ubicación o segmento. Una demanda de 500 unidades puede ser normal en una promoción y extraordinaria durante una semana habitual. Ignorar el contexto conduce a alertar sobre cambios esperables o a pasar por alto eventos extraños dentro de un grupo específico.

Antes de elegir el método hay que definir respecto de qué referencia se juzga rareza. Aplicar una cerca IQR a cada columna no reemplaza evaluar combinaciones. Tampoco basta con escalar variables si se mezclan procesos distintos. El diagnóstico puede requerir segmentación, variables temporales o modelos de relaciones, además de una medida de distancia.

23. Ejercicio resuelto: ¿cómo se detectan valores mediante la regla IQR?

El rango intercuartílico es IQR = Q3 - Q1. La regla usual de Tukey establece una cerca inferior Q1 - 1,5 × IQR y una superior Q3 + 1,5 × IQR. Las observaciones fuera de ellas se marcan como potenciales outliers. Los cuartiles dependen de una convención de cálculo, por lo que en ejercicios y software conviene declararla.

Si Q1 = 20 y Q3 = 40, el IQR es 20. La cerca inferior es -10 y la superior 70. Un valor de 75 queda señalado; uno de 65 no. En una variable necesariamente no negativa, la cerca inferior negativa simplemente puede no detectar ningún caso bajo: no obliga a aceptar valores negativos que contradigan el dominio.

La robustez proviene de utilizar posiciones de la distribución en lugar de magnitudes extremas. Sin embargo, la regla no modela asimetría ni relaciones entre variables. En una distribución de ingresos con cola derecha larga puede marcar observaciones válidas; en una mezcla de segmentos puede confundir heterogeneidad con anomalía.

Se informa el número y proporción señalados, se revisan los registros y se compara por segmentos. Si las cercas se usan como parte de un modelo predictivo, deben aprenderse con entrenamiento y luego aplicarse a evaluación. La cerca es un umbral exploratorio, no una prueba de error ni una autorización para borrar los puntos.

24. ¿Cómo se comparan z-score y MAD para detectar extremos?

El z-score expresa cuántas unidades de desvío separan un valor del centro: z = (x - μ)/σ. En una normal de referencia, un valor absoluto mayor que tres es poco frecuente. Si μ y σ se estiman de una muestra contaminada por extremos, ambos pueden desplazarse y ocultar justamente los casos que se quiere detectar, fenómeno de enmascaramiento.

MAD es la mediana de las desviaciones absolutas respecto de la mediana: MAD = mediana(|x - mediana(x)|). Una puntuación modificada habitual es z* = 0,67449 × (x - mediana(x))/MAD. La constante aproxima una escala comparable al desvío estándar bajo normalidad. Un corte como 3,5 es una convención orientativa, no universal.

Con mediana 10 y MAD 2, un valor 22 tiene z* aproximadamente 4,05 y quedaría señalado bajo ese corte. La cuenta es 0,67449 × 12/2. Si MAD es cero, la fórmula no puede aplicarse directamente: ocurre, por ejemplo, cuando muchos valores son idénticos. Debe revisarse la variable y elegir otra escala o una regla de dominio.

MAD tolera mejor contaminación, pero tampoco convierte una distribución asimétrica en normal ni resuelve anomalías conjuntas. Conviene contrastar gráficos, segmentos y tasas de alerta. La pregunta central es qué patrón define normalidad, no qué umbral produce el mayor número de observaciones eliminadas.

25. ¿Qué diferencias existen entre residuo, leverage e influencia en regresión?

El residuo es la diferencia entre respuesta observada y predicha. Un residuo grande señala que el modelo explica mal una observación. Leverage mide cuán extrema es su posición en el espacio de predictores respecto de los demás datos. En mínimos cuadrados, sus valores son los elementos diagonales de $H = X(X^TX)^{-1}X^T$, cuando la matriz tiene rango completo.

Una observación influyente modifica de manera importante el ajuste al retirarla. La influencia suele combinar leverage y residuo, aunque depende del modelo y de la cantidad de casos semejantes. Un punto lejano en X pero cercano a la tendencia puede tener leverage alto y residuo pequeño. Otro cerca del centro de X puede tener residuo grande sin cambiar mucho la pendiente.

La distancia de Cook resume el efecto sobre los valores ajustados. En OLS, una expresión es $D_i = [e_i^2/(p \times MSE)] \times [h_{ii}/(1 - h_{ii})^2]$, donde p incluye el intercepto y MSE estima la varianza residual. Umbrales como 4/n son señales de revisión y no decisiones automáticas.

El análisis útil compara coeficientes, predicciones y conclusiones con y sin el caso, y verifica su validez. Si una venta excepcional legítima cambia mucho la pendiente, puede estar revelando falta de soporte en ese rango o una relación distinta por segmento. Eliminarla sin explicación ocultaría una fragilidad del modelo.

26. ¿Cómo funcionan Mahalanobis y los métodos multivariados de referencia?

La distancia de Mahalanobis incorpora escalas y correlaciones: $d^2 = (x - \mu)^T\Sigma^{-1}(x - \mu)$, donde μ es el vector de medias y Σ la matriz de covarianzas. A diferencia de la distancia euclídea sin preparación, reconoce que algunas combinaciones son habituales porque las variables se mueven conjuntamente.

Si ingreso y gasto suelen aumentar juntos, una persona con ambos altos puede estar cerca del patrón; otra con ingreso bajo y gasto muy alto puede estar lejos, aunque ninguno sea extremo por separado. La geometría de referencia es elipsoidal. Bajo una normal multivariada con parámetros conocidos, la distancia cuadrática sigue una distribución chi-cuadrado con dimensión p; con parámetros estimados, la calibración exacta requiere cuidado.

La media y covarianza clásicas pueden contaminarse por los mismos outliers. Estimadores robustos, como los basados en subconjuntos de covarianza mínima, intentan limitar ese efecto. También hay dificultades cuando existen pocas observaciones, demasiadas variables o colinealidad que vuelve singular o inestable la matriz.

Mahalanobis funciona mejor cuando una referencia aproximadamente elíptica representa el proceso. No es una solución universal para grupos con densidades diferentes o geometrías complejas. Se comparan diagnósticos, se reduce redundancia cuando corresponda y se validan alertas. Una distancia grande indica incompatibilidad con esa referencia concreta, no evidencia suficiente de fraude o error.

27. ¿Cómo detectan anomalías Isolation Forest, LOF y One-Class SVM?

Isolation Forest construye particiones aleatorias mediante variables y umbrales. Las observaciones que se aíslan en pocas divisiones suelen recibir mayor anomalía. Su intuición es que casos raros requieren menos cortes para separarse. Produce un score cuya orientación y umbral deben verificarse en la implementación: un valor mayor no significa siempre más anomalía.

LOF compara la densidad local de un punto con la de sus vecinos. Es útil cuando existen regiones con densidades distintas, pero depende del número de vecinos, la distancia y el escalado. Puede señalar un punto aislado dentro de una región sin considerarlo extremo global. Para evaluar nuevos casos se necesita una configuración apropiada de detección de novedades, diferente del uso transductivo sobre la muestra original.

One-Class SVM aprende una frontera de referencia, frecuentemente no lineal mediante kernels. Requiere atención a escala y parámetros. El parámetro nu controla restricciones relacionadas con errores de entrenamiento y vectores de soporte; no debe interpretarse como conocimiento cierto de la proporción real de anomalías.

Los tres representan normalidad de forma distinta, por lo que sus alertas no son intercambiables. Se comparan a una tasa de alerta o capacidad de revisión coherente, con etiquetas si existen y con examen de casos. Un parámetro de contaminación fija o ayuda a fijar un corte; no demuestra que esa fracción de los datos sea inválida.

28. ¿Qué opciones existen para tratar outliers y cómo se elige entre ellas?

La corrección procede cuando hay evidencia del error y puede recuperarse el valor verdadero. La exclusión se justifica cuando el registro no pertenece a la población de interés o es inválido e irrecuperable. Conservarlo es apropiado si representa un evento legítimo que el análisis debe cubrir. La segmentación permite modelar procesos distintos en lugar de forzar una sola distribución.

Una transformación logarítmica comprime diferencias altas y puede mejorar una relación multiplicativa, pero modifica la escala y la función de error. La winsorización reemplaza extremos por umbrales, como percentiles. Reduce su influencia, aunque también oculta magnitudes verdaderas y puede sesgar totales o riesgos de cola. Debe explicitarse qué se pierde.

Los métodos robustos modifican el ajuste sin necesariamente borrar datos. Una pérdida Huber trata cuadráticamente residuos pequeños y aproximadamente de forma lineal los grandes, reduciendo su peso relativo. Aun así, robustez respecto de residuos no resuelve todos los problemas de leverage, sesgo de selección o mezcla de segmentos.

Se comparan tratamientos con el mismo protocolo y se revisan errores en las colas. Si la tarea consiste en detectar fraude, los outliers pueden ser el objetivo y no un obstáculo. Los umbrales aprendidos se estiman con entrenamiento. Una decisión transparente registra qué casos se afectaron, por qué, cuánto cambiaron resultados y qué limitaciones conserva la solución elegida.

29. Ejercicio resuelto: ¿cómo cambian los resúmenes y el costo con un extremo?

Considérese una serie de tiempos de entrega de 2, 3, 3, 4 y 28 días. La media es 40/5 = 8 días y la mediana 3. Si se excluye el 28, la media baja a 3. La diferencia es importante, pero no demuestra que deba eliminarse: ese caso podría representar una falla real que interesa reducir.

Si el informe comunicara solo la mediana, describiría bien el servicio típico, pero ocultaría el costo del retraso. Si cada día posterior al quinto genera una penalización de 100, el caso de 28 produce 2.300 de costo y los otros ninguno. Para la decisión deben presentarse centro y cola, no elegir la medida que haga parecer mejor el servicio.

La investigación debe revisar si 28 es un error de fecha, un pedido internacional o un evento excepcional dentro del servicio habitual. En el primer caso se corrige con evidencia; en el segundo puede segmentarse; en el tercero se conserva y se analiza el mecanismo de demora. Si muchos casos similares aparecen en semanas recientes, puede existir un cambio del proceso.

El ejemplo muestra que robustez y riesgo responden preguntas diferentes. Una medida robusta evita que el resumen típico quede dominado por un caso; una medida de costo debe reflejar las consecuencias reales de ese caso. El análisis completo puede necesitar ambas, junto con la frecuencia de incumplimiento y la magnitud de las penalizaciones.

30. ¿Cómo se evalúa un detector y se convierte su score en una decisión?

Un score de anomalía ordena casos según una referencia aprendida. Convertirlo en alerta requiere un umbral y una política de actuación. Si hay etiquetas fiables, se estudian precision, recall y la curva precision-recall; si no las hay, se revisan muestras de alertas y no alertas, estabilidad y coherencia con el dominio. La ausencia de etiquetas limita lo que puede afirmarse sobre exactitud.

La capacidad operativa importa. Si pueden revisarse veinte casos diarios, puede elegirse un corte que priorice ese volumen, pero no corresponde presentarlo como una tasa de fraude estimada. También deben distinguirse costos de omitir una anomalía y de investigar un caso normal. El umbral se selecciona con validación y después se evalúa en evidencia reservada.

La detección de outliers sobre una muestra potencialmente contaminada y la detección de novedades respecto de una referencia mayormente normal son problemas diferentes. El conjunto de entrenamiento y el uso del detector deben ser coherentes con esa elección. Un modelo puede fallar si aprendió anomalías frecuentes como comportamiento habitual.

En producción se monitorean tasa de alertas, distribución de scores, rendimiento de revisiones y composición por segmento. Un aumento sostenido puede señalar un fallo de captura, nuevas anomalías o drift. La respuesta debe investigar cuál ocurrió antes de recalibrar indiscriminadamente. El detector aporta evidencia para una decisión; no reemplaza la verificación del caso.

4 Transformación de variables correlación y regresión lineal

31. ¿Qué diferencia existe entre escalar, transformar la forma y normalizar filas?

Escalar una columna modifica su centro o amplitud para expresar magnitudes de forma comparable. StandardScaler utiliza z = (x - μ)/σ; MinMax utiliza (x - mínimo)/(máximo - mínimo); RobustScaler suele restar la mediana y dividir por el IQR. Cada método aprende estadísticas de entrenamiento. La estandarización no garantiza normalidad, y los valores nuevos pueden quedar fuera de [0, 1] después de MinMax.

Una transformación de forma, como el logaritmo, es no lineal y modifica distancias relativas. Puede comprimir una cola o simplificar una relación multiplicativa. Escalar y transformar responden problemas distintos y pueden combinarse. Una variable asimétrica conserva su asimetría bajo un escalado lineal positivo.

Normalizar filas divide cada observación por una norma, como $\sqrt{(x1^2 + x2^2)}$. Convierte una fila en un perfil de magnitud controlada y puede ser útil para comparar composición o dirección. La fila [3, 4] se convierte en [0,6; 0,8] con norma L2 igual a 5. Esa operación pierde información sobre magnitud total: [30, 40] produce el mismo perfil. No equivale a estandarizar columnas.

La elección depende del algoritmo y el significado. K-Means, KNN y penalizaciones sobre coeficientes son sensibles a escala; los árboles basados en umbrales son generalmente invariantes a transformaciones estrictamente monótonas de cada predictor, salvo efectos numéricos o particularidades de implementación. Si media 50 y desvío 10, x = 70 produce z = 2: indica posición relativa, no probabilidad de que el dato sea correcto.

32. ¿Cuándo conviene usar logaritmos, Box-Cox o Yeo-Johnson y qué ocurre al transformar el objetivo?

El logaritmo comprime diferencias multiplicativas y puede estabilizar variabilidad o simplificar relaciones. Está definido para valores positivos. log1p(x) calcula log(1 + x), útil cuando x es no negativo, pero su interpretación difiere del logaritmo puro. Sumar una constante arbitraria para acomodar ceros o negativos modifica el modelo y necesita justificación.

Box-Cox aplica una familia de transformaciones a valores estrictamente positivos: para λ distinto de cero, (x^λ - 1)/λ; para λ = 0, log(x). Yeo-Johnson admite también cero y negativos mediante una definición por tramos. El parámetro se estima con entrenamiento y no obliga a que el resultado sea normal. El criterio de utilidad debe relacionarse con la tarea, no con producir histogramas visualmente simétricos.

Transformar y cambia qué errores se penalizan. Minimizar errores cuadrados en log(y) da más importancia a diferencias relativas que hacerlo en unidades de y. Si ln(y) = α + βx, un aumento unitario en x multiplica la predicción retransformada por exp(β). El cambio porcentual exacto es 100 × [exp(β) - 1], aproximadamente 100β cuando β es pequeño.

La retransformación a la escala original requiere atención: exp(E[ln Y|X]) no es generalmente E[Y|X]. Con errores lognormales de varianza constante σ^2, la media condicional incluye el factor exp(σ^2/2); bajo otros supuestos puede considerarse un ajuste de smearing. La predicción sin ajuste puede representar una mediana bajo condiciones apropiadas. Debe evaluarse también en unidades originales si allí se toma la decisión.

33. ¿Cómo se codifican categorías, interacciones y variables temporales sin introducir relaciones falsas?

Las categorías nominales suelen representarse con indicadores one-hot. En una regresión con intercepto se omite habitualmente una categoría de referencia para evitar dependencia lineal exacta. El coeficiente de cada indicador compara su categoría con esa referencia, manteniendo constantes los demás predictores. Un número asignado arbitrariamente a una categoría no representa una distancia sustantiva.

La codificación ordinal es adecuada cuando existe un orden real, pero asignar 1, 2 y 3 también puede imponer incrementos iguales si el modelo trata esa variable linealmente. Si esa igualdad no es razonable, pueden usarse indicadores u otra representación. La codificación por objetivo requiere procedimientos fuera de fold para evitar que la respuesta de un caso se filtre a su propia representación.

Una interacción x1 × x2 permite que la pendiente de x1 dependa de x2. En y = β0 + β1x1 + β2x2 + β3x1x2, el cambio marginal respecto de x1 es β1 + β3x2. Los términos polinómicos representan curvatura manteniendo linealidad en los parámetros, aunque aumentan complejidad y pueden provocar colinealidad.

En variables cíclicas, seno y coseno evitan que 23 horas y 0 horas queden artificialmente muy separadas: se usan sin(2πh/24) y cos(2πh/24). Las variables de recencia se calculan con información disponible al origen. El pipeline reúne imputación, codificación, escalado y modelo; en cada partición aprende solo con entrenamiento y debe prever categorías nuevas y reglas reproducibles.

34. ¿Qué miden covarianza, Pearson y Spearman y cómo se calculan?

La covarianza indica si dos variables se desvían de sus medias en el mismo sentido. Su magnitud depende de las unidades. Pearson divide la covarianza por los desvíos y mide asociación lineal en una escala de -1 a 1. Está indefinido si alguna variable es constante y puede cambiar mucho con observaciones extremas.

Para X = [1, 2, 3] e Y = [2, 4, 5], las medias son 2 y 11/3. La suma de productos de desviaciones es 3; la suma de cuadrados de X es 2 y la de Y es 14/3. Por tanto, $r = 3/\sqrt{[2 \times (14/3)]} \approx 0{,}982$. El resultado indica asociación lineal fuerte en esos tres datos; la muestra pequeña no permite generalizar sin más.

Spearman calcula Pearson sobre rangos y mide asociación monótona. En ese ejemplo ambos órdenes coinciden, por lo que Spearman vale 1 aunque la relación no sea una línea exacta. Sin empates puede calcularse como $1 - 6\Sigma d_i^2/[n(n^2 - 1)]$; con empates conviene utilizar rangos promedio y correlacionarlos directamente.

Pearson cercano a cero no demuestra independencia: si X es simétrico alrededor de cero y $Y = X^2$, puede haber una relación determinista con correlación lineal nula. Spearman tampoco identifica todas las dependencias no monótonas. Por eso se acompaña el coeficiente con un scatterplot, contexto, tamaño y revisión de segmentos. Ninguno prueba causalidad, y ambos pueden distorsionarse por selección, faltantes o mezclas de poblaciones.

35. ¿Por qué correlación no implica causalidad y qué es la multicolinealidad?

Una asociación puede surgir porque X causa Y, porque Y afecta X, porque ambas dependen de otra variable o por selección y tendencias compartidas. Encontrar que descuento y ventas crecen juntos no identifica el efecto del descuento: podrían haberse aplicado promociones precisamente a productos con demanda alta o en fechas especiales.

La multicolinealidad aparece cuando predictores contienen información lineal redundante. Si es exacta, no pueden identificarse coeficientes únicos bajo la parametrización completa. Si es fuerte pero imperfecta, puede aumentar la incertidumbre y volver inestables los coeficientes. Las predicciones dentro del rango observado pueden seguir siendo razonables, por lo que interpretación y predicción no enfrentan el mismo problema.

El factor de inflación de varianza de Xj es $VIFj = 1/(1 - Rj^2)$, donde Rj^2 surge al regresarla sobre los otros predictores. Si $Rj^2 = 0{,}90$, VIF = 10. Los cortes usados como alerta son convenciones; no reemplazan evaluar el objetivo y la estructura de datos. Las correlaciones por pares pueden no detectar redundancias que involucran varias variables.

Se puede revisar definición de atributos, combinar medidas redundantes, aumentar información o usar regularización para estabilizar predicción. Eliminar automáticamente todo predictor correlacionado también puede descartar información útil o cambiar la interpretación del modelo. Para causalidad se necesita un diseño apropiado y explicitar supuestos; incluir muchas variables no garantiza remover confusión y puede introducir sesgo si se condiciona sobre variables inadecuadas.

36. ¿Qué es la regresión lineal y qué optimiza mínimos cuadrados?

La regresión lineal representa la media condicional, o una aproximación a ella, como una combinación de predictores: $y_i = \beta 0 + \beta 1 x_i 1 + \ldots + \beta p x_i p + \varepsilon_i$. Es lineal en los parámetros, aunque los predictores pueden incluir logaritmos, polinomios e interacciones. No exige que cada variable tenga distribución normal.

Mínimos cuadrados ordinarios elige coeficientes que minimizan $SSE = \Sigma(y_i - \hat{y}_i)^2$. Elevar errores al cuadrado evita cancelaciones y da más peso a desviaciones grandes. Si X tiene rango completo, la solución matricial se expresa como $\beta = (X^T X)^{-1} X^T y$; en cálculo numérico suelen preferirse métodos más estables que formar explícitamente la inversa.

En regresión simple con intercepto, $\beta 1 = \Sigma(x_i - \bar{x})(y_i - \bar{y})/\Sigma(x_i - \bar{x})^2$ y $\beta 0 = \bar{y} - \beta 1 \bar{x}$. La pendiente tiene unidades de y por unidad de x. En regresión múltiple, cada coeficiente describe una comparación condicionada a los demás términos del modelo, no necesariamente un efecto causal.

El intercepto representa la predicción cuando todos los predictores valen cero, lo cual puede quedar fuera del soporte real. Centrar variables puede facilitar su interpretación sin cambiar las predicciones de un modelo equivalente. El ajuste debe contrastarse con residuos y desempeño independiente. Una línea que explica bien entrenamiento puede extrapolar mal o reflejar una relación confusa que no se mantiene en la población de uso.

37. Ejercicio resuelto: ¿cómo se estima e interpreta una regresión simple?

Supongamos gasto publicitario $X = [1, 2, 3]$ y ventas $Y = [2, 4, 5]$, expresados en miles de unidades monetarias. Se obtiene $\bar{x} = 2$, $\bar{y} = 11/3$, $\Sigma(x - \bar{x})(y - \bar{y}) = 3$ y $\Sigma(x - \bar{x})^2 = 2$. La pendiente estimada es 1,5 y el intercepto 2/3. El modelo es $\hat{y} = 0{,}6667 + 1{,}5x$.

Las predicciones son 2,1667; 3,6667 y 5,1667. Los residuos, aproximadamente -0,1667; 0,3333 y -0,1667, suman cero porque el ajuste incluye intercepto. SSE es 1/6 y SST, la suma de cuadrados respecto de la media de Y, es 14/3. El R^2 de entrenamiento vale $1 - (1/6)/(14/3) = 27/28 \approx 0{,}9643$.

La pendiente indica que, dentro de esta relación ajustada, mil unidades monetarias adicionales de publicidad se asocian con mil quinientas de ventas. No significa que el incremento sea causal ni que pueda extrapolarse a cualquier presupuesto. Tres puntos ofrecen muy poco soporte para evaluar estabilidad y supuestos.

Para X = 4, la predicción sería 6,6667, pero es una extrapolación fuera del rango 1 a 3. Una respuesta completa informa el cálculo, las unidades, el carácter asociativo y la necesidad de validar con otras observaciones. Si ventas y publicidad crecieron por estacionalidad compartida, el coeficiente podría describir esa tendencia y no la eficacia de invertir más.

38. ¿Qué supuestos y diagnósticos deben revisarse en regresión lineal?

Hay que distinguir condiciones para calcular un ajuste, para interpretar parámetros y para realizar inferencia clásica. La ausencia de colinealidad exacta permite identificar coeficientes. Una especificación apropiada y $E[\varepsilon|X] = 0$ son centrales para interpretar el modelo y, bajo condiciones, obtener estimaciones insesgadas. La homocedasticidad y ausencia de correlación de errores sustentan la eficiencia y las fórmulas clásicas de varianza.

La normalidad de errores facilita inferencia exacta en muestras pequeñas, pero no es necesaria para calcular OLS ni para toda forma de predicción. Tampoco se exige normalidad de X. Con heterocedasticidad pueden utilizarse errores estándar robustos para inferencia; eso no corrige una forma funcional equivocada ni convierte el coeficiente en causal.

El gráfico de residuos frente a predicción permite observar curvatura, dispersión creciente y grupos. Un patrón curvo sugiere falta de términos relevantes o una relación mal representada; un embudo sugiere variación no constante. En datos temporales se revisan residuos según tiempo y autocorrelación. También se examinan leverage, influencia y soporte de los predictores.

Un intervalo de confianza para la media condicional y un intervalo de predicción para una nueva observación responden preguntas distintas; el segundo incorpora también variabilidad individual y suele ser más ancho. Los diagnósticos orientan revisión, pero ningún gráfico prueba que todos los supuestos se cumplan. Deben relacionarse con el proceso de muestreo, la finalidad y la evaluación fuera de muestra.

39. Ejercicio resuelto: ¿cómo se interpretan MAE, MSE, RMSE y R^2?

Para valores reales [10, 20, 30] y predicciones [12, 18, 33], los errores reales menos predichos son [-2, 2, -3]. MAE = (2 + 2 + 3)/3 = 2,333; MSE = (4 + 4 + 9)/3 = 5,667 y RMSE = $\sqrt{5,667} \approx 2,380$. MAE y RMSE tienen unidades de y, mientras MSE tiene unidades al cuadrado.

MAE promedia magnitud absoluta y no aumenta cuadráticamente el peso de errores grandes. RMSE destaca más los errores de cola. No corresponde compararlos entre bases con escalas o poblaciones diferentes sin contexto. Una diferencia pequeña puede ser relevante si afecta decisiones costosas, y una mejora promedio puede ocultar deterioro en un segmento.

Para R^2 se usa 1 - SSE/SST. Aquí la media real es 20, SST = 200 y SSE = 17, por lo que $R^2 = 0,915$. En test, R^2 puede ser negativo si las predicciones son peores, en suma de cuadrados, que usar la media del conjunto evaluado. No mide porcentaje de aciertos ni demuestra causalidad. Si y es constante, la definición convencional requiere tratamiento especial.

En OLS con intercepto, el R^2 de entrenamiento no disminuye al añadir predictores. R^2 ajustado = $1 - (1 - R^2)(n - 1)/(n - p - 1)$, donde p excluye intercepto, penaliza esa incorporación bajo la formulación habitual, pero no sustituye validación. Para decidir se contrasta con un baseline y con errores por rango, período y segmento, además de la función de costo relevante.

40. ¿Cómo se interpretan modelos con logaritmos, polinomios e interacciones?

Transformar predictores cambia la relación representada y la interpretación. En $y = \alpha + \beta \ln(x)$, un aumento de 1 % en x produce aproximadamente un cambio de $0{,}01\beta$ unidades de y; exactamente, el cambio al multiplicar x por 1,01 es $\beta \ln(1{,}01)$. En $\ln(y) = \alpha + \beta \ln(x)$, β es la elasticidad de la relación retransformada bajo su interpretación apropiada.

Un modelo cuadrático $y = \beta_0 + \beta_1 x + \beta_2 x^2$ sigue siendo lineal en parámetros. Su pendiente depende de x: $\beta_1 + 2\beta_2 x$. Si $\beta_1 = 4$ y $\beta_2 = -0{,}5$, la pendiente es 4 - x: el efecto marginal ajustado es positivo bajo 4, cero en 4 y negativo por encima. Esa curva no autoriza extrapolar lejos del rango observado ni demuestra un máximo causal.

En una interacción entre precio y canal, el término adicional permite pendientes distintas. Si $y = \beta_0 + \beta_1 \text{precio} + \beta_2 \text{online} + \beta_3 \text{precio} \times \text{online}$, el canal presencial tiene pendiente β_1 y el online $\beta_1 + \beta_3$. β_2 compara canales cuando precio vale cero; centrar precio puede volver esa comparación más interpretable.

La elección de términos se valida y debe guardar coherencia con el dominio. Aumentar grado polinómico puede reducir error de entrenamiento y producir oscilaciones fuera de muestra. Se revisan residuos, estabilidad y penalización cuando corresponde. La calidad del modelo depende de su capacidad de representar una relación útil y generalizable, no de cuántas transformaciones contiene.

5 Regresión lineal validación cruzada y generalización

41. ¿Para qué se separan entrenamiento, validación y test?

Entrenamiento permite estimar parámetros, como coeficientes de una regresión. Validación orienta decisiones de diseño: variables, transformaciones, regularización y otros hiperparámetros. Test estima el desempeño del procedimiento elegido sobre evidencia que no participó en esas decisiones. La distinción es funcional y no depende de que existan exactamente tres archivos.

En validación cruzada, los folds dentro de entrenamiento pueden cumplir la función de validación. Luego se reajusta el procedimiento seleccionado usando todos los datos permitidos de desarrollo y se evalúa una vez en test. Si se vuelve a modificar el modelo después de mirar ese resultado, test pasa a influir en la selección y deja de tener su independencia original.

Un baseline permite evaluar si la complejidad aporta valor. Para regresión bajo MSE puede predecirse la media de entrenamiento; bajo MAE, la mediana. En test se aplica esa constante aprendida, no una media recalculada para mejorar la referencia. El baseline debe representar una alternativa realista y respetar la misma disponibilidad de información.

La separación imita el uso futuro. Si se predicen clientes nuevos, no deben compartirse clientes entre conjuntos; si se pronostica el futuro, debe respetarse el orden temporal. Un porcentaje habitual de división no es una garantía de validez. Importan independencia, tamaño de evaluación, representatividad y correspondencia entre el protocolo y la decisión que se pretende tomar.

42. ¿Cómo funciona K-Fold y qué se aprende de sus resultados?

K-Fold divide los datos en K particiones. En cada iteración se entrena con K - 1 y se evalúa con la restante. Cada caso se evalúa una vez en la versión estándar, y el procedimiento se ajusta K veces. Se obtienen métricas por fold y una síntesis que estima el desempeño bajo ese esquema de muestreo.

Con 100 observaciones y K = 5, cada iteración usa aproximadamente 80 para entrenamiento y 20 para evaluación. El modelo de cada fold es diferente; la validación cruzada no consiste en evaluar cinco veces un único ajuste. Las transformaciones aprendidas también deben reajustarse en cada entrenamiento.

Si las muestras son independientes y el orden carece de significado, puede usarse una mezcla reproducible antes de dividir. Si existe dependencia temporal o por entidad, el K-Fold aleatorio puede dar una estimación engañosa. En clasificación, la estratificación ayuda a mantener representación de clases, pero no corrige por sí sola dependencia o sesgos de selección.

La media resume el rendimiento y la dispersión muestra sensibilidad a las particiones. Los folds comparten datos de entrenamiento, de modo que sus scores no son observaciones estadísticamente independientes. Su desvío no debe transformarse automáticamente en un intervalo de confianza usando la fórmula de una muestra independiente. Se informa el protocolo, tamaños y variación, y se interpretan las diferencias con prudencia.

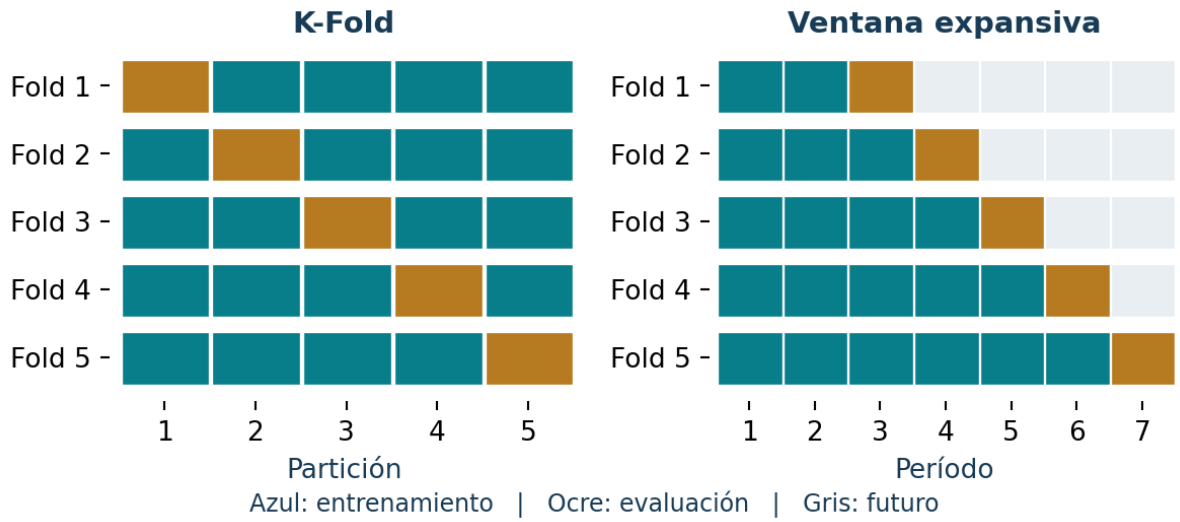


Figura 2. K-Fold rota la evaluación. La ventana expansiva conserva el orden temporal y utiliza solo el pasado de cada origen.

43. ¿Cómo se elige K y qué diferencias hay con holdout, LOOCV y repeticiones?

Holdout usa una única partición y tiene bajo costo, pero el resultado puede depender mucho del reparto. K-Fold utiliza varias y permite observar esa variación. Valores como cinco o diez son puntos de partida comunes, no reglas óptimas universales. La elección debe considerar tamaño, dependencia, costo y estabilidad de la evaluación.

Aumentar K deja más observaciones en cada entrenamiento y menos en cada validación. Eso puede acercar el ajuste al entrenamiento final, pero los scores de folds pequeños pueden ser inestables y el costo aumenta. LOOCV es el extremo K = n: cada caso se evalúa con un ajuste entrenado en los n - 1 restantes. Su estimación puede tener alta variabilidad según el problema, pese a usar casi todos los casos para cada ajuste.

Repeated K-Fold repite particiones con repartos distintos y ayuda a medir sensibilidad al muestreo. No crea nuevos datos independientes ni elimina el optimismo de seleccionar la mejor configuración sobre esas mismas repeticiones. También requiere contabilizar el costo total de ajustes.

Si hay 20 configuraciones, cinco folds y tres repeticiones, se realizan 300 ajustes, además del final. Un incremento del presupuesto debe aportar información útil. Conviene elegir un protocolo razonable, mantenerlo común entre modelos y ampliar la evaluación cuando una diferencia pequeña o una inestabilidad relevante justifique el esfuerzo.

44. ¿Qué son subajuste, sobreajuste y el compromiso entre sesgo y varianza?

El subajuste ocurre cuando el modelo no representa patrones importantes y tiene errores elevados incluso en entrenamiento. El sobreajuste aparece cuando captura peculiaridades o ruido de entrenamiento que no se mantienen en datos nuevos. Un error de entrenamiento muy bajo no garantiza utilidad predictiva; interesa la capacidad de generalizar.

El sesgo describe errores sistemáticos asociados a restricciones del procedimiento, y la varianza refleja cuánto cambia el ajuste al cambiar la muestra de entrenamiento. Bajo una formulación habitual del error cuadrático esperado, se distingue sesgo al cuadrado, varianza e incertidumbre irreducible. Esa descomposición es conceptual y depende de la distribución y el procedimiento considerado.

Un modelo lineal puede subajustar una relación curva. Un polinomio de grado alto puede aproximar entrenamiento casi exactamente y oscilar en test. La regularización, una menor complejidad o más información útil pueden reducir la segunda dificultad, mientras que mejores variables o una forma funcional más adecuada pueden resolver la primera.

Las curvas de aprendizaje comparan error de entrenamiento y validación según tamaño de muestra. Si ambos siguen altos y próximos, conviene revisar representación y capacidad; si entrenamiento es bajo y validación mucho mayor, puede existir alta varianza. Esos patrones son diagnósticos, no pruebas definitivas. Fuga de información, cambios de población y datos ruidosos también afectan curvas y deben descartarse antes de atribuir todo a complejidad.

45. ¿Cómo se usan Ridge, Lasso y Elastic Net dentro de la validación?

Ridge añade una penalización cuadrática sobre coeficientes: se minimiza SSE + $\lambda\Sigma\beta j^2$ bajo una convención de escala. Lasso usa $\lambda\Sigma|\beta j|$ y puede llevar coeficientes exactamente a cero. Elastic Net combina ambas penalizaciones. El intercepto suele quedar sin penalizar. Las convenciones numéricas de λ cambian entre implementaciones, por lo que sus valores no son directamente comparables sin revisar la función objetivo.

La penalización introduce sesgo para estabilizar el ajuste y reducir varianza. Resulta especialmente útil con predictores redundantes o muchos términos. Ridge distribuye el peso entre variables correlacionadas; Lasso puede seleccionar una y descartar otra, de modo que la selección puede ser inestable. Un coeficiente cero no demuestra que la variable carezca de relación causal o sustantiva.

El escalado importa porque penalizar el mismo coeficiente numérico no equivale a penalizar la misma contribución si las unidades difieren. Se aprende imputación y escalado dentro de cada fold y se compara una grilla de penalizaciones con la métrica elegida. Una penalización excesiva puede subajustar; una casi nula puede mantener la inestabilidad de OLS.

El mejor λ se selecciona en validación, luego se reajusta el pipeline con los datos de desarrollo y se evalúa en test. Si dos configuraciones rinden de forma similar, puede preferirse la más estable o simple. La justificación debe incluir desempeño independiente e interpretación del costo de complejidad, no solo el mínimo de una curva ruidosa.

46. ¿Cómo evita un pipeline el leakage durante la validación cruzada?

Un pipeline encadena transformaciones y modelo como un único procedimiento. Al recibir un fold de entrenamiento, ajusta allí imputación, escalado y estimador. Al evaluar el fold reservado, aplica las transformaciones ya aprendidas y predice. Esa organización evita que estadísticas de validación influyan en el ajuste cuando se utiliza correctamente.

El error frecuente es calcular medias de imputación, parámetros de escalado o selección de variables con toda la base antes de ejecutar K-Fold. Aunque después se dividan las filas, ya se compartió información. PCA y selección de predictores también deben ubicarse dentro del proceso. Un pipeline no resuelve una variable temporal calculada con información futura ni una división por grupos equivocada: esos problemas requieren diseño previo.

Una estructura mínima para predictores numéricos es:

```
from sklearn.pipeline import make_pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge
from sklearn.model_selection import KFold, cross_val_score

modelo = make_pipeline(SimpleImputer(strategy="median"),
                       StandardScaler(), Ridge(alpha=1.0))
cv = KFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(modelo, X, y, cv=cv,
                         scoring="neg_mean_absolute_error")
mae_por_fold = -scores
```

X e y representan únicamente datos de desarrollo. El signo negativo se usa porque las herramientas de selección maximizan scores: negar una pérdida convierte menor error en mayor score. Para presentar MAE se cambia nuevamente el signo. En datos mixtos se utiliza un ColumnTransformer para aplicar tratamientos adecuados a columnas numéricas y categóricas.

47. ¿Cuándo se necesitan particiones por grupos o por tiempo?

La partición debe reflejar qué significa un caso nuevo. Si varias filas pertenecen al mismo cliente, distribuirlas entre entrenamiento y validación puede permitir que el modelo reconozca características de ese cliente. Para evaluar generalización a clientes no vistos se usa una separación que mantenga cada grupo completo en un único lado.

GroupKFold es una opción para entidades repetidas. Si interesa predecir nuevas visitas de clientes conocidos, puede ser válido un diseño diferente, pero debe conservar el orden de información y definir claramente el alcance. Generalizar a entidades nuevas y a nuevas observaciones de entidades conocidas son problemas distintos; sus métricas no son intercambiables.

En series, entrenar con observaciones posteriores a las evaluadas contradice el uso de un pronóstico. Se emplean cortes temporales con entrenamiento anterior y validación posterior, y pueden introducirse brechas para evitar superposición de ventanas o etiquetas. Si el objetivo usa un período futuro de treinta días, casos cercanos a la frontera pueden compartir información del resultado.

La estratificación mantiene proporciones de clases, pero no evita compartir personas ni mirar el futuro. Cuando existen simultáneamente grupos y tiempo, debe construirse un esquema que respete ambas restricciones relevantes. La calidad de una métrica depende de esa correspondencia. Un score obtenido con una división fácil puede medir reconocimiento de casos relacionados y no la capacidad que necesita la decisión real.

48. ¿Por qué la búsqueda de hiperparámetros puede ser optimista y qué resuelve la validación anidada?

Cada score de validación contiene variación de muestreo. Al probar muchas configuraciones y elegir la mejor, se selecciona tanto calidad real como fluctuación favorable. Informar ese máximo como estimación independiente del rendimiento puede ser optimista. El problema aumenta si también se cambian variables, métricas y reglas después de mirar resultados.

La validación anidada separa selección y evaluación mediante dos capas. En cada fold externo se reserva una parte que no participa en la búsqueda. Sobre el entrenamiento externo, los folds internos eligen configuración. Luego se reajusta la ganadora con todo ese entrenamiento externo y se evalúa en el fold externo reservado.

Con cinco folds externos, cuatro internos y diez configuraciones, la búsqueda requiere aproximadamente 200 ajustes internos, además de reajustes externos y final. Los scores externos evalúan el procedimiento de selección, no una única configuración fija elegida de antemano. Después puede ejecutarse la búsqueda sobre todos los datos de desarrollo para construir el modelo final.

Un test verdaderamente reservado también puede ofrecer evaluación final independiente después de seleccionar con CV. No siempre se necesitan ambos mecanismos. La elección depende de tamaño, presupuesto y finalidad. Lo esencial es no usar la misma evidencia para escoger reiteradamente y para sostener que el resultado es una medida imparcial del desempeño futuro.

49. Ejercicio resuelto: ¿cómo se comparan dos regresiones con scores de CV?

Dos modelos obtienen MAE en cinco folds. A presenta [8, 9, 10, 9, 9] y B [7, 7, 7, 7, 17]. Ambos tienen media 9. Con desvío muestral, A tiene aproximadamente 0,707 y B 4,472. La igualdad de medias oculta que B falla mucho en una partición, mientras A resulta más estable bajo ese reparto.

Las diferencias A - B por fold son [1, 2, 3, 2, -8]. Conviene estudiar qué caracteriza al quinto fold: un segmento distinto, un período especial, pocos casos extremos o un error de preparación. Comparar con particiones comunes permite un contraste más informativo que comparar medias obtenidas de repartos distintos.

No se debe concluir automáticamente que A es superior en toda población ni aplicar una prueba de independencia ingenua a cinco scores correlacionados. La diferencia relevante depende también del costo. Si el quinto fold incluye clientes con pérdidas grandes por errores de predicción, la inestabilidad de B puede tener una consecuencia operativa importante.

La decisión puede favorecer A por robustez o justificar investigar B y repetir bajo un esquema apropiado. Se informa media, dispersión, métricas por segmentos y protocolo. Después de elegir se reajusta con datos de desarrollo y se evalúa en test reservado. El análisis completo responde tanto cuánto error se espera como dónde y por qué podría concentrarse.

50. ¿Cómo se diseña y comunica un experimento completo de regresión?

Se define primero el objetivo, la unidad, el horizonte y la información disponible al predecir. Luego se fija un protocolo de partición compatible con entidades y tiempo, una métrica principal y un baseline. También se establecen criterios complementarios, como errores altos, segmentos, estabilidad y costo operativo.

El EDA y la preparación orientados a selección se realizan con datos de desarrollo. Se construye un pipeline y se comparan un conjunto acotado de especificaciones: OLS, transformaciones justificadas y regularización, por ejemplo. Todas utilizan las mismas particiones y disponibilidad de variables. El resultado no debe depender de escoger una métrica diferente para cada competidor.

La búsqueda se realiza en validación; la evaluación final se mantiene independiente. Se documentan semillas, versión de datos, parámetros, transformaciones y código. Repetir el proceso con una semilla fija facilita reproducibilidad, pero no elimina incertidumbre: cuando interesa estabilidad puede evaluarse sensibilidad a otras particiones permitidas.

El informe final presenta desempeño frente a referencia, interpretación de errores, supuestos y límites. Un intervalo de predicción o un análisis de costo puede ser más útil que un R^2 alto aislado. También indica cómo se monitorearán cambios y cuándo se revisará el modelo. Un experimento bien diseñado produce una comparación defendible; una secuencia de ajustes sobre el mismo test produce una cifra cada vez más adaptada a ese test.

6 Series de tiempo y ejercicios de pronóstico

51. ¿Qué distingue una serie de tiempo de una muestra transversal?

Una serie de tiempo registra valores ordenados por fecha y frecuencia. El orden contiene información: ventas de hoy pueden depender de ayer, de la estación y de eventos anteriores. Una muestra transversal describe distintas entidades en un período; una serie estudia evolución. Un panel combina entidades y tiempo, por lo que puede contener dependencias de ambos tipos.

Antes de modelar se revisan frecuencia, calendario, duplicados, intervalos irregulares y cambios de medición. Un día sin registro no equivale a ventas cero: podría ser cierre del local o pérdida de datos. Agregar por semana requiere definir qué días se incluyen y si la variable se suma, promedia o toma al cierre. Facturación es un flujo; stock al final del día es una magnitud de estado y no se interpreta igual al sumarlo.

El origen del pronóstico es el instante en que se dispone de información. El horizonte indica cuántos períodos adelante se predice. Si se pronostican ventas a siete días, toda variable utilizada debe ser calculable en el origen, incluyendo cuándo se publicaron datos externos y no solo a qué fecha corresponden.

La dependencia temporal impide usar sin análisis una división aleatoria convencional. También exige distinguir predicción con variables futuras conocidas, como calendario, de variables que deberán pronosticarse, como precios de un competidor. El diseño debe reproducir la información y la frecuencia reales de la decisión.

52. ¿Qué son tendencia, estacionalidad, ciclo y residuo?

La tendencia es una evolución persistente del nivel, como crecimiento de demanda. La estacionalidad es un patrón recurrente asociado a un calendario o período conocido: días de la semana o meses del año. Un ciclo presenta fluctuaciones de duración no necesariamente fija, como fases económicas. El residuo es la parte que no explica la descomposición elegida; no debe asumirse automáticamente que sea ruido independiente.

En una representación aditiva, $y(t) = T(t) + S(t) + R(t)$. Es apropiada cuando la amplitud estacional es relativamente constante en unidades de y. En una multiplicativa, $y(t) = T(t) \times S(t) \times R(t)$, la amplitud cambia aproximadamente en proporción al nivel. Para valores positivos, un logaritmo transforma productos en sumas y puede facilitar el modelado.

Si una tienda pasa de 100 a 200 unidades medias y su incremento navideño pasa de 30 a 60, la estacionalidad parece proporcional. Si sigue agregando unas 30, una representación aditiva puede ser más adecuada. Se necesitan suficientes ciclos y revisión de cambios, porque patrones visuales con pocos períodos pueden resultar inestables.

Una descomposición retrospectiva puede usar información posterior a cada punto, por ejemplo mediante suavizado centrado. Eso es válido para describir toda la historia, pero sus componentes no pueden incorporarse sin adaptación a un pronóstico que pretende haberse realizado en el pasado. Deben diferenciarse explicación histórica y transformación disponible en tiempo real.

53. ¿Qué es estacionariedad y qué aportan diferencias y transformaciones?

La estacionariedad débil supone una media constante, varianza finita constante y covarianza que depende del rezago, no de la fecha específica. Muchos modelos temporales se apoyan en una estructura aproximadamente estable. Tendencias, cambios de régimen o variación de amplitud pueden violar esa condición.

La primera diferencia es $\Delta y(t) = y(t) - y(t - 1)$. Puede remover una tendencia estocástica o estabilizar el nivel. La diferencia estacional es $y(t) - y(t - s)$, donde s es la longitud del ciclo. Un logaritmo puede ayudar si la variabilidad crece con el nivel. Cada transformación responde una dificultad distinta y no debe aplicarse por rutina.

Por ejemplo, para [100, 110, 121], las diferencias son [10, 11]. Las diferencias logarítmicas son $\ln(110/100)$ y $\ln(121/110)$, ambas aproximadamente 0,0953. El crecimiento porcentual exacto es 10 %, mientras la diferencia logarítmica multiplicada por cien aproxima esa tasa. La elección depende de si interesan cambios absolutos o relativos.

Pruebas como ADF y KPSS usan hipótesis distintas y dependen de especificación y tamaño. ADF suele contrastar una raíz unitaria; KPSS parte de una hipótesis de estacionariedad de nivel o tendencia. No reemplazan gráficos y dominio. Sobrediferenciar puede introducir dependencia innecesaria y perder estructura. Después de transformar hay que prever cómo reconstruir pronósticos e incertidumbre en la escala original.

54. ¿Qué representan ACF, PACF y los componentes de ARIMA y SARIMA?

La ACF mide correlación entre una serie y sus valores rezagados. La PACF mide asociación con un rezago una vez controlados los intermedios bajo la proyección correspondiente. Ayudan a reconocer persistencia y patrones estacionales, pero pueden ser engañosas si una tendencia no tratada domina las correlaciones.

Un AR(p) utiliza p rezagos de la variable, mientras un MA(q) utiliza q rezagos de innovaciones. En este contexto MA significa media móvil de errores, no el promedio móvil descriptivo de observaciones. ARIMA(p, d, q) combina ambos sobre una serie diferenciada d veces. SARIMA añade componentes estacionales con órdenes P, D y Q y período s.

Un AR(1), por ejemplo, puede expresarse como $y(t) = c + \varphi y(t - 1) + \varepsilon(t)$. Si $|\varphi| < 1$ y se cumplen condiciones apropiadas sobre innovaciones, existe una estructura estacionaria con media $c/(1 - \varphi)$. Esa condición no debe extrapolarse sin más a cualquier modelo o serie real.

Los patrones clásicos de corte o decaimiento de ACF y PACF orientan candidatos ideales, pero en muestras reales pueden no ser nítidos. Se combinan con validación temporal, parsimonia y diagnóstico de residuos. Un criterio de información compara ajuste y complejidad bajo sus supuestos; no garantiza el menor error en el horizonte de negocio. El modelo debe competir además con referencias simples.

55. ¿Qué baselines temporales deben compararse antes de adoptar un modelo complejo?

El pronóstico ingenuo usa el último valor observado: $\hat{y}(T + h) = y(T)$. El ingenuo estacional utiliza el valor más reciente del mismo punto del ciclo, por ejemplo el lunes anterior o el mismo mes del año previo. Una media histórica o móvil puede servir como referencia cuando el nivel es relativamente estable.

El modelo con drift prolonga una pendiente promedio: $\hat{y}(T + h) = y(T) + h[y(T) - y(1)]/(T - 1)$. Es una referencia de tendencia, no una garantía de crecimiento constante. Las medias móviles deben calcularse con datos anteriores al origen y elegir su ventana en validación.

Supongamos ventas de lunes de 100, 105 y 110, con última venta diaria del domingo igual a 60. Para el lunes siguiente, el naive diario predice 60 y el naive semanal 110. Si existe un patrón fuerte por día, el segundo puede ser más realista. Un modelo complejo debe superar esa referencia en los mismos orígenes y horizontes.

Los baselines muestran cuánto valor aporta aprender estructura adicional. Una mejora frente a una referencia débil puede ser irrelevante frente a la política operativa vigente. Se evalúan por horizonte y segmento, con métricas y costos coherentes. La comparación exige que ninguno utilice información futura que otro no tenga y que todos reciban el mismo calendario y tratamiento de faltantes.

56. ¿Cómo se construyen rezagos y medias móviles sin leakage temporal?

Una feature temporal es válida si puede calcularse con información disponible en el origen. Para predecir y(t), un rezago y(t - 1) y una media de los siete valores anteriores pueden ser válidos; una media que incluye y(t) o valores posteriores no lo es. El rezago debe interpretarse según la frecuencia y los huecos de calendario.

En una tabla diaria ordenada, una construcción habitual es:

```
datos = datos.sort_values("fecha")
datos["lag1"] = datos["ventas"].shift(1)
datos["media7"] = datos["ventas"].shift(1).rolling(7).mean()
```

El desplazamiento previo evita incluir el valor objetivo del día. Si hay varias tiendas, las operaciones se realizan dentro de cada tienda y respetando su calendario; desplazar toda la tabla sin grupos podría tomar la venta de otra entidad.

Para horizontes mayores, las features deben corresponder al origen. Si se pronostican siete días de una vez, no se pueden usar como rezagos los valores reales de los días intermedios, todavía desconocidos. Pueden usarse predicciones recursivas o modelos directos por horizonte.

También hay leakage por medias centradas, imputación con observaciones posteriores, escaladores ajustados a toda la historia o variables externas publicadas con retraso. El calendario futuro suele conocerse, pero la demanda futura no. Un backtest serio reconstruye el conocimiento histórico y la disponibilidad real de cada variable.

57. ¿Cómo funcionan el backtesting y las ventanas expansivas o móviles?

El backtesting simula pronósticos emitidos en varios orígenes históricos. En cada uno se entrena con información previa y se evalúan períodos posteriores. Permite medir desempeño bajo condiciones parecidas a las del uso real y observar cambios entre períodos. No es una partición aleatoria de fechas.

Una ventana expansiva incorpora cada vez más historia: se entrena hasta diciembre, luego hasta enero y así sucesivamente. Una móvil conserva una longitud fija y descarta datos antiguos. La primera aprovecha más información; la segunda puede adaptarse a cambios, pero usa menos observaciones. La longitud se selecciona sin consultar el período final reservado.

Si se pronostican tres meses, cada origen genera tres errores, uno por horizonte. Se informa rendimiento a un mes, dos y tres, porque las tareas difieren. Si las ventanas de evaluación se superponen, algunos resultados reales aparecen en varios pronósticos y los errores quedan relacionados; no deben tratarse como observaciones totalmente independientes.

El backtest debe reproducir frecuencia de reentrenamiento, publicación de datos y política de variables. Puede requerir un gap cuando etiquetas o features abarcan períodos cercanos a la frontera. Si se seleccionan parámetros sobre muchos backtests y luego se informa la mejor cifra allí, reaparece el sesgo de selección. Conviene separar desarrollo temporal y una evaluación final posterior.

58. ¿Qué métricas son apropiadas para pronósticos y por qué MAPE puede fallar?

MAE y RMSE mantienen unidades de la serie y permiten interpretar errores operativos. RMSE da más peso a desviaciones grandes. MAPE promedia errores absolutos relativos al valor real, pero no está definido cuando este es cero y puede dispararse con valores pequeños. Además, puede dar comparaciones poco equilibradas entre series con niveles muy distintos.

MASE escala el MAE respecto de un error de referencia calculado en entrenamiento. Para período estacional s, el denominador usual es el promedio de $|y(t) - y(t - s)|$ en la historia disponible. Un MASE menor que uno indica un error medio inferior a esa escala de referencia, pero no equivale exactamente a superar un pronóstico ingenuo específico en cada test. Si el denominador es cero, la métrica presenta un problema.

Si interesa una predicción de cuantil q, la pérdida pinball es q × error cuando $y - \hat{y}$ es no negativo, y (q - 1) × error cuando es negativo. Penaliza de forma asimétrica subestimación y sobreestimación. Para intervalos se revisan cobertura y amplitud: un intervalo muy ancho puede cubrir mucho y ser poco útil.

Las métricas se calculan por horizonte, temporada y segmento, y se relacionan con costos. Elegir un modelo solo por error promedio puede ocultar quiebres de stock en picos de demanda. Una evaluación completa explica qué tipo de error castiga y por qué eso representa la decisión prevista.

59. Ejercicio resuelto: ¿cómo se evalúa un pronóstico y se elige un cuantil de inventario?

La demanda real de cuatro semanas es [100, 120, 80, 100] y el pronóstico [90, 110, 100, 100]. Los errores son [10, 10, -20, 0]. MAE = 40/4 = 10 y RMSE = √(600/4) ≈ 12,247. El error medio firmado es cero, pero eso no significa que el pronóstico sea perfecto: las desviaciones positivas y negativas se cancelan.

Si cada unidad faltante cuesta 4 y cada unidad excedente cuesta 1, el costo es 4 × (10 + 10) + 1 × 20 = 100. La métrica simétrica no expresa esa asimetría. En un problema básico de inventario con demanda incierta, costos lineales y sin otras restricciones, el cuantil óptimo es q = Cu/(Cu + Co) = 4/5 = 0,80, donde Cu es costo de faltante y Co de excedente.

Esto sugiere usar el percentil 80 de la distribución predictiva, no agregar arbitrariamente 20 % a una media. La distribución debe estimarse y validarse; una política de inventario puede incorporar además plazos, stock disponible, lotes y vencimientos que modifican el problema básico.

El ejercicio distingue calidad puntual y utilidad de decisión. También muestra que ausencia de sesgo medio no garantiza buen desempeño. La evaluación final compara costo y servicio en backtests con los mismos horizontes y restricciones que la operación real.

60. ¿Cómo se diseña un proyecto completo de series de tiempo y se diagnostican fallas?

Se define frecuencia, origen, horizonte, variable objetivo y política de actualización. Se audita calendario, calidad y publicación de información. Luego se exploran tendencia, estacionalidad, cambios y posibles variables externas. Las transformaciones se eligen con datos históricos de desarrollo y se comparan baselines y modelos bajo el mismo backtest.

Para varios pasos, una estrategia recursiva usa predicciones intermedias como entradas posteriores y puede acumular error. Una directa ajusta modelos para horizontes específicos; una estrategia multisalida estima varios pasos conjuntamente. La elección debe evaluarse, porque menor error a un paso no garantiza mejor pronóstico a doce.

Los residuos se revisan por tiempo, nivel y horizonte. Autocorrelación persistente puede indicar estructura no capturada; errores concentrados en promociones pueden señalar variables omitidas; un deterioro reciente puede responder a drift. Residuos poco autocorrelacionados son deseables en muchos modelos, pero no prueban que las predicciones estén calibradas o que la especificación sea correcta.

El informe presenta comparación con referencias, incertidumbre, errores de cola y consecuencias operativas. En producción se monitorean cambios de distribución, cobertura de intervalos y costo. El procedimiento debe poder reconstruirse en cada origen. Si el desempeño mejora solo cuando se calculan variables con toda la serie, la prioridad es corregir la evaluación, porque esa ventaja no estaría disponible al tomar la decisión futura.

7 Regresión logística y decisiones de clasificación

61. ¿Qué estima la regresión logística y por qué utiliza una sigmoide?

La regresión logística binaria estima P(Y = 1|X). Define un score z = β0 + Σβjxj y lo transforma mediante p = 1/(1 + exp(-z)). La sigmoide convierte cualquier score real en una probabilidad entre cero y uno. Una regresión lineal directa sobre etiquetas 0 y 1 podría producir valores fuera de ese rango y una estructura de errores poco adecuada.

Si z = 0, p = 0,5; si z = 2, p ≈ 0,881; si z = -2, p ≈ 0,119. La relación entre z y probabilidad no es lineal: cambios iguales del score tienen consecuencias diferentes cerca de 0,5 y cerca de los extremos. El modelo es lineal en el logit, no en la probabilidad.

La probabilidad y la decisión son objetos diferentes. Para clasificar se aplica un umbral: positivo si p es al menos t. Con t = 0,5, la frontera equivale a z = 0; con otro corte, z = ln[t/(1 - t)]. Cambiar t modifica decisiones y métricas, pero no las probabilidades estimadas por el ajuste fijo.

El modelo puede incorporar transformaciones e interacciones para representar relaciones más complejas. Su estimación no garantiza causalidad ni calibración fuera de muestra. Hay que considerar población, definición de la clase positiva y disponibilidad de variables. Predecir riesgo alto no demuestra que una intervención particular reducirá ese riesgo.

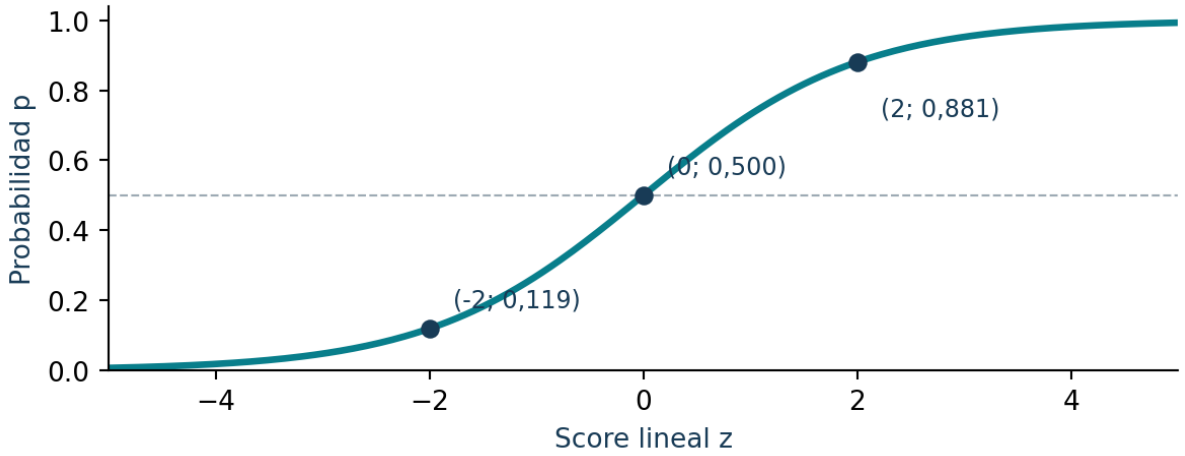


Figura 3. La sigmoide transforma el score en probabilidad. Un mismo cambio de score no produce un cambio constante de probabilidad.

62. ¿Qué son odds, logit y odds ratio y cómo se relacionan con probabilidad?

Los odds comparan probabilidad de ocurrencia con la de no ocurrencia: odds = p/(1 - p). El logit es su logaritmo: ln[p/(1 - p)] = z. La probabilidad se recupera como odds/(1 + odds). Aunque relacionados, probabilidad y odds no son equivalentes y sus cambios no deben intercambiarse al comunicar resultados.

Si p = 0,20, los odds son 0,20/0,80 = 0,25: una ocurrencia por cada cuatro no ocurrencias en la interpretación de proporciones. Si un predictor aumenta una unidad y su coeficiente es ln(2), los odds se multiplican por dos manteniendo constantes los demás términos, siempre que no haya interacciones que alteren esa pendiente.

Partiendo de 0,20, odds 0,25 pasan a 0,50 y la nueva probabilidad es 0,50/1,50 = 0,3333. Duplicar odds no duplicó la probabilidad. Partiendo de 0,80, odds 4 pasan a 8 y la probabilidad llega a 8/9 ≈ 0,8889. El mismo odds ratio tiene cambios absolutos diferentes según el punto inicial.

Un odds ratio de 2 tampoco demuestra que un riesgo causal se haya duplicado. Describe una relación condicionada dentro del modelo. Para una interpretación útil se informa la unidad del predictor, la referencia y, cuando sea pertinente, probabilidades calculadas en perfiles representativos en lugar de trasladar mecánicamente el odds ratio a un porcentaje de aumento de probabilidad.

63. ¿Cómo se interpretan coeficientes, efectos marginales e interacciones?

Un coeficiente βj representa el cambio del logit ante una unidad de xj, manteniendo los otros términos constantes. Sin interacción, exp(βj) es el odds ratio correspondiente. La unidad importa: un coeficiente por peso y otro por mil pesos son numéricamente diferentes aunque representen la misma relación.

Para un predictor continuo que entra linealmente, la derivada de p respecto de xj es βj × p(1 - p). Por eso el efecto marginal depende del perfil. Si βj = 0,8, en p = 0,5 la derivada vale 0,20; en p = 0,1 vale 0,072. Son cambios locales por unidad, no incrementos exactos para grandes variaciones de xj.

Para una variable binaria se utiliza una diferencia discreta: probabilidad con indicador uno menos probabilidad con indicador cero, manteniendo el resto del perfil. Si existen interacciones, el coeficiente aislado no representa un cambio uniforme. En z = β0 + β1x + β2g + β3xg, el cambio del score respecto de x es β1 + β3g, y su efecto sobre p incorpora además p(1 - p).

El intercepto describe un perfil con predictores cero y categorías de referencia, que podría ser poco realista. Centrar variables facilita interpretaciones. Los efectos condicionados del modelo no deben presentarse como causales sin diseño apropiado. También conviene contrastar estabilidad de coeficientes y comunicar efectos sobre probabilidades con incertidumbre cuando la finalidad sea inferencial.

64. ¿Cómo se estima la regresión logística mediante verosimilitud y log-loss?

Para una observación binaria, la probabilidad asignada al resultado real es p si y = 1 y 1 - p si y = 0. Bajo el modelo de observaciones independientes, la verosimilitud conjunta es el producto de esas cantidades. Maximizar su logaritmo equivale a minimizar la entropía cruzada binaria o log-loss promedio.

La pérdida es -[y ln(p) + (1 - y)ln(1 - p)]. Si y = 1 y p = 0,90, vale aproximadamente 0,105; si se predice p = 0,10, vale 2,303. La pérdida penaliza con fuerza errores muy confiados y evalúa probabilidades, no solo etiquetas. Dos modelos pueden tener igual accuracy y distinta log-loss.

El ajuste se realiza mediante optimización numérica. Una separación perfecta de clases puede hacer que el máximo no exista en coeficientes finitos bajo el modelo no penalizado. La regularización limita magnitudes y estabiliza el procedimiento, aunque cambia la estimación y debe seleccionarse mediante validación.

Para evaluar se utiliza información reservada y se mantiene la definición de clase positiva. Ponderar clases modifica la función de entrenamiento y puede ser útil, pero las probabilidades resultantes necesitan evaluación de calibración respecto de la población de interés. Un score probabilístico aparentemente bueno en una muestra balanceada artificialmente no asegura que las probabilidades correspondan a la frecuencia real del evento en producción.

65. ¿Qué papel tienen regularización, escalado y desbalance de clases?

Las penalizaciones L1 y L2 restringen coeficientes y ayudan a controlar complejidad. L1 puede inducir ceros; L2 reduce magnitudes sin producir generalmente selección exacta. En algunas implementaciones, C representa el inverso de la intensidad de regularización: menor C implica mayor penalización. Debe revisarse la convención antes de interpretar una grilla.

El escalado hace comparables las penalizaciones y facilita optimización cuando predictores tienen unidades muy distintas. Se aprende dentro del pipeline y de cada fold. La representación de categorías e interacciones también se valida. Un mejor ajuste en entrenamiento no justifica agregar términos si no mejoran generalización o interpretación pertinente.

Con eventos raros, accuracy puede ser engañosa. Si 1 % de los casos es positivo, predecir siempre negativo obtiene 99 % de accuracy y recall positivo cero. Se consideran ponderación de clases, muestreo y cambio de umbral, pero son intervenciones diferentes. Muestrear antes de dividir puede duplicar o relacionar casos entre entrenamiento y evaluación y contaminar resultados.

Las técnicas de balanceo se aplican únicamente al entrenamiento correspondiente. La evaluación mantiene una prevalencia coherente con el uso real, y se revisan precision, recall, calibración y costos. Cambiar pesos puede mejorar cobertura, pero no elimina la necesidad de elegir una política de decisión. El objetivo no es producir igual cantidad de positivos y negativos, sino actuar bien sobre la población prevista.

66. Ejercicio resuelto: ¿cómo se calculan las métricas de una matriz de confusión?

Definamos TP = positivos reales detectados, FN = positivos omitidos, FP = negativos marcados como positivos y TN = negativos correctamente descartados. Supongamos TP = 30, FN = 10, FP = 20 y TN = 140, sobre 200 casos. La prevalencia real es 40/200 = 20 %.

Accuracy = (TP + TN)/n = 170/200 = 0,85. Precision = TP/(TP + FP) = 30/50 = 0,60: seis de diez alertas son positivas reales. Recall o sensibilidad = TP/(TP + FN) = 30/40 = 0,75: se detectaron tres de cada cuatro positivos. Especificidad = TN/(TN + FP) = 140/160 = 0,875.

F1 = 2 × precision × recall/(precision + recall) ≈ 0,6667. La tasa de falsos positivos es FP/(FP + TN) = 0,125. No debe confundirse con 1 - precision = 0,40, que usa como denominador las alertas. Balanced accuracy promedia sensibilidad y especificidad y vale 0,8125.

Ninguna métrica por sí sola selecciona una decisión. Si omitir un positivo cuesta mucho, recall adquiere mayor importancia; si revisar alertas es caro, precision y volumen son relevantes. F1 no incorpora TN ni costos específicos. Al cambiar el umbral cambia la matriz, por lo que se elige el corte en validación. También se revisa qué sucede por segmento y si los casos etiquetados representan toda la población o solo los que ya fueron investigados.

67. ¿Qué diferencias existen entre ROC-AUC y precision-recall?

La curva ROC representa sensibilidad frente a tasa de falsos positivos para distintos umbrales. Su AUC resume la capacidad de ordenar positivos por encima de negativos y puede interpretarse como una probabilidad de ordenamiento, con manejo de empates. No evalúa calibración ni determina directamente el umbral operativo.

La curva precision-recall relaciona pureza de alertas y cobertura de positivos. Es especialmente informativa cuando la clase positiva es escasa. La precision depende de la prevalencia: con sensibilidad y especificidad fijas, una menor frecuencia del evento puede aumentar mucho la proporción de falsas alertas. La referencia de precision de un ranking aleatorio se relaciona con esa prevalencia.

Por ejemplo, con 1.000 negativos, una tasa de falsos positivos de 5 % genera 50 alertas falsas. Si solo existen 10 positivos y se detectan ocho, recall es 80 % y precision 8/58 ≈ 13,8 %. Un rendimiento ROC aparentemente razonable puede implicar una carga de revisión elevada.

Average precision y el área trapezoidal bajo una curva PR no son idénticas en general, por lo que conviene informar la definición utilizada. Las comparaciones se realizan en la misma población y protocolo. Para decidir interesa además el tramo compatible con capacidad y costo; una AUC global mayor puede no mejorar la precision entre los veinte casos que realmente pueden revisarse.

68. ¿Qué diferencia hay entre discriminación y calibración?

Discriminación es la capacidad de ordenar casos por riesgo. Calibración es la correspondencia entre probabilidades predichas y frecuencias observadas. Entre muchos casos con p cercana a 0,70, una calibración adecuada implica una proporción positiva cercana a 70 %. No significa que cada caso individual sea 70 % verdadero ni que siete de cualquier grupo de diez deban ser positivos.

Un modelo puede ordenar correctamente y exagerar probabilidades. Una transformación estrictamente creciente de scores preserva el ranking y, en condiciones sin cambios de empates, ROC-AUC, pero puede modificar calibración. Por eso un buen AUC no basta para decisiones basadas en valor esperado.

Se usan curvas de confiabilidad y medidas como Brier score, promedio de $(p - y)^2$, y log-loss. Brier refleja más que calibración aislada, incluyendo resolución, por lo que debe interpretarse junto con gráficos y prevalencia. Los grupos de probabilidad con pocos casos producen estimaciones ruidosas; conviene informar soporte.

La recalibración puede aprender una función sigmoidal o una relación isotónica sobre evidencia de calibración independiente del ajuste base, o mediante un esquema de CV apropiado. Ajustarla con test y después reportar ese mismo test como independiente contamina la evaluación. Cambios de prevalencia y de relación entre predictores y evento pueden degradar calibración; el monitoreo debe verificarla en el entorno donde las probabilidades se usan para decidir.

69. ¿Cómo se elige un umbral según costos y capacidad?

El umbral óptimo depende de las consecuencias de cada acción. En un problema binario simple, con costo C_FP por actuar sobre un negativo, C_FN por omitir un positivo y costos de decisiones correctas iguales a cero, actuar tiene costo esperado C_FP(1 - p), mientras no actuar cuesta C_FNp. Se elige positivo cuando $p \geq C_FP/(C_FP + C_FN)$.

Si C_FP = 10 y C_FN = 90, el corte es 0,10. Esa fórmula requiere probabilidades apropiadas para la población y costos comparables. Si hay beneficios adicionales, costos individuales o efectos de la intervención, debe construirse la función de valor correspondiente, no aplicar la fórmula sin adaptar supuestos.

En una campaña, contactar cuesta 4 y cada respuesta genera un ingreso de 30, suponiendo ese ingreso bruto por éxito y sin otros efectos. El valor esperado es 30p - 4, positivo cuando $p > 4/30 \approx 0{,}1333$. Si solo se puede contactar al 10 %, se priorizan casos por valor esperado dentro de la capacidad; el corte efectivo puede superar ese mínimo.

La probabilidad de compra no equivale al efecto incremental de contactar. Si algunos comprarían igualmente, para optimizar una campaña causal se necesita estimar uplift o utilizar un diseño experimental. El umbral se selecciona con validación, se evalúa con test y se revisa por estabilidad y restricciones. Elegir 0,5 por defecto ignora estos mecanismos.

70. ¿Cómo se extiende la logística a varias clases y qué se monitorea en producción?

La formulación multinomial utiliza scores z_k y softmax: p_k = exp(z_k)/Σexp(z_j). Las probabilidades son positivas para scores finitos y suman uno. Para scores [2, 1, 0], se obtienen aproximadamente [0,6652; 0,2447; 0,0900]. Sumar la misma constante a todos no cambia el resultado porque el factor común se cancela; restar el máximo mejora estabilidad numérica.

La decisión puede elegir la clase más probable o minimizar una matriz de costos. En una política one-vs-rest se ajustan clasificadores binarios separados; su interpretación y normalización difieren de una optimización multinomial conjunta. Hay que especificar qué enfoque se utiliza y evaluar errores entre clases concretas, además de promedios macro o ponderados.

En producción se observa distribución de predictores, faltantes, probabilidades, decisiones, prevalencia y rendimiento cuando llegan etiquetas. Data drift es un cambio en P(X); concept drift es un cambio en P(Y|X). Un descenso de desempeño no demuestra por sí solo cuál ocurrió. También puede haber un cambio de captura o de definición del objetivo.

El proceso necesita una referencia, ventanas y acciones ante deterioro. Se distingue recalibrar probabilidades, cambiar umbral y reentrenar el modelo, porque resuelven problemas diferentes. Si las etiquetas llegan con demora o solo para casos revisados, la evaluación puede estar sesgada. La política final se documenta y se monitorea con la misma claridad que la estimación.

8 Árboles de decisión y ejercicios prácticos

71. ¿Cómo representa un árbol de decisión un problema de clasificación o regresión?

Un árbol divide el espacio de predictores mediante reglas sucesivas, como ingreso ≤ 50 o antigüedad > 12. Cada nodo interno plantea una condición, las ramas representan resultados y las hojas reúnen observaciones que cumplen una secuencia de reglas. CART construye habitualmente divisiones binarias y regiones alineadas con ejes de los predictores.

En clasificación, una hoja puede estimar probabilidades mediante proporciones de clases, eventualmente con pesos, y tomar una decisión según la regla elegida. En regresión con pérdida cuadrática, predice la media de los valores de entrenamiento de la hoja. Un mismo árbol produce así una función por regiones, capaz de representar umbrales e interacciones.

Por ejemplo, ingreso puede importar solo entre clientes con deuda alta. El árbol puede dividir primero por deuda y después por ingreso dentro de una rama. Esa capacidad contrasta con un modelo aditivo que mantiene una pendiente uniforme sin términos adicionales. Sin embargo, representar una interacción no demuestra su causalidad.

Un árbol pequeño facilita explicar una ruta de decisión. Uno grande puede resultar difícil de leer y memorizar grupos mínimos. La interpretación debe indicar soporte y variabilidad de las hojas. Una regla obtenida en diez casos no tiene la misma estabilidad que otra respaldada por miles, aunque ambas sean fáciles de expresar en una frase.

72. ¿Qué miden Gini y entropía y cómo elige CART una división?

Gini mide impureza como $1 - \Sigma p_k^2$. Vale cero cuando todos los casos pertenecen a una sola clase. En clasificación binaria alcanza 0,5 cuando ambas proporciones son iguales. La entropía es $-\Sigma p_k \log_2(p_k)$, usando cero para el término de una clase ausente por continuidad; en dos clases equilibradas vale un bit.

CART compara variables y umbrales candidatos y elige una reducción de impureza. Se calcula la impureza del padre menos la media ponderada de los hijos: $\Delta I = I_{padre} - (n_{izq}/n) I_{izq} - (n_{der}/n) I_{der}$. Con pesos de casos, se utilizan los pesos correspondientes. Importa la calidad conjunta de ambos hijos, no solo la pureza de uno.

Gini cero en una hoja significa pureza de las etiquetas de entrenamiento, no error nulo en casos futuros. Una hoja con un único caso es necesariamente pura, pero puede tener muy poca capacidad de generalización. Además, Gini y entropía no siempre eligen exactamente la misma división, aunque frecuentemente produzcan resultados parecidos.

La búsqueda es greedy: toma una buena decisión local y continúa recursivamente, sin garantizar el árbol globalmente óptimo para cualquier objetivo. Las restricciones y la poda ayudan a controlar complejidad. La medida de impureza organiza el entrenamiento, mientras la selección final debe guiarse por validación y por la métrica de la tarea.

73. Ejercicio resuelto: ¿cómo se calcula la ganancia de una división con Gini?

Un nodo tiene 100 observaciones, 50 positivas y 50 negativas. Su Gini es $1 - 0{,}5^2 - 0{,}5^2 = 0{,}5$. Una división genera un hijo de 40 casos, con 30 positivos y 10 negativos, y otro de 60, con 20 positivos y 40 negativos.

El hijo izquierdo tiene proporciones 0,75 y 0,25 y Gini 0,375. El derecho tiene 1/3 y 2/3 y Gini $4/9 \approx 0{,}4444$. La impureza ponderada es $0{,}4 \times 0{,}375 + 0{,}6 \times 0{,}4444 \approx 0{,}4167$. La ganancia es $0{,}5 - 0{,}4167 \approx 0{,}0833$.

Otra división separa dos positivos puros y deja 98 casos con 48 positivos y 50 negativos. Aunque el primer hijo tiene Gini cero, la impureza ponderada es aproximadamente 0,4898 y la ganancia 0,0102. La primera división es mejor según el criterio porque mejora más la mezcla total, no porque produzca la hoja más pura.

La cuenta exige verificar que los hijos conserven todos los casos y conteos del padre. También hay que respetar requisitos como tamaño mínimo de hoja: la segunda división podría estar prohibida por crear solo dos casos. La ganancia de entrenamiento no garantiza mejora en test. La elección de profundidad y restricciones se valida con evidencia separada.

74. ¿Cómo construye predicciones y divisiones un árbol de regresión?

Un árbol de regresión busca regiones cuyos valores objetivo sean homogéneos. Con criterio cuadrático, cada hoja predice la media y la división reduce la suma de errores cuadrados dentro de sus hijos. La media es la constante que minimiza esa suma. Otros criterios pueden usar una pérdida diferente y cambiar la predicción óptima, como la mediana bajo error absoluto.

Para valores $y = [2, 4, 10, 12]$, la media general es 7 y $SSE = 25 + 9 + 9 + 25 = 68$. Si una regla de X separa [2, 4] de [10, 12], las medias son 3 y 11. Cada grupo suma 2 de SSE, por lo que la suma posterior es 4 y la reducción 64.

La predicción es constante dentro de cada región. A diferencia de una regresión lineal, un árbol convencional no extrapola una pendiente: un nuevo valor de X por encima de los observados sigue alguna rama y recibe una predicción basada en una hoja existente. Esto puede ser útil para relaciones de umbral y una limitación para tendencias sostenidas.

Un árbol muy profundo puede crear hojas con un caso y error de entrenamiento cero. Eso no asegura buen pronóstico. Se revisan errores fuera de muestra, tamaños de hoja y soporte del rango. Si el objetivo es demanda creciente, conviene contrastar con modelos que representen tendencia y no asumir que la flexibilidad de particiones permite extrapolarla.

75. ¿Cómo afectan escala, categorías e interacciones al comportamiento de un árbol?

Los árboles basados en umbrales usan el orden de los valores. Una transformación estrictamente monótona preserva los grupos que pueden separarse por un corte, por lo que normalmente no necesitan estandarización por razones de distancia. Puede haber diferencias por precisión, histogramas o detalles de implementación. Cambiar la unidad de una variable no equivale a cambiar su importancia como ocurre en una distancia sin escalar.

Las categorías necesitan una representación compatible con el algoritmo. Codificar nominales como 1, 2 y 3 impone un orden en los cortes; algunas particiones posibles serán más fáciles que otras por una asignación arbitraria. One-hot evita ese orden, aunque puede aumentar dimensiones. Algunas implementaciones admiten categorías nativas; no debe suponerse soporte sin verificarlo.

Las interacciones aparecen al dividir una variable dentro de una rama definida por otra. Por ejemplo, un nivel de deuda puede ser riesgoso solo entre clientes de baja antigüedad. El árbol puede representar esa regla sin crear explícitamente un producto de variables, aunque quizá requiera suficiente profundidad y observaciones.

La forma rectangular de las regiones también impone límites. Una frontera diagonal puede requerir muchos escalones, y una alta dimensión dificulta obtener hojas con suficiente soporte. La elección entre árbol y modelo lineal depende de estructura, estabilidad e interpretación. La capacidad de capturar no linealidad es una ventaja potencial que debe confirmarse mediante validación.

76. ¿Cómo se controla el sobreajuste con profundidad y tamaño de nodos?

La profundidad limita cuántas decisiones sucesivas puede contener una ruta. Un máximo pequeño restringe la función y puede subajustar; uno grande permite reglas específicas y aumenta el riesgo de memorizar ruido. El tamaño mínimo de hoja evita predicciones basadas en muy pocos casos. El tamaño mínimo para dividir un nodo regula cuándo puede proponerse una nueva partición.

Estas restricciones no son equivalentes. Exigir suficientes casos para dividir no garantiza por sí solo hijos grandes, mientras que un mínimo por hoja limita directamente el soporte de las regiones terminales. También puede imponerse una mejora mínima de impureza o un número máximo de hojas. Su efecto depende de distribución y pesos, no solo de la cantidad total de filas.

Si un árbol de profundidad tres obtiene accuracy de entrenamiento 85 % y validación 82 %, mientras uno de profundidad veinte obtiene 100 % y 73 %, la segunda configuración muestra una brecha compatible con sobreajuste. No basta con la brecha: se revisan partición, tamaño y posibles cambios de población.

Los parámetros se eligen mediante CV apropiada o validación temporal, con el mismo protocolo para competidores. Se puede preferir una configuración simple si su rendimiento es similar y sus reglas más estables. Una restricción no garantiza generalización, pero limita la capacidad de crear hojas que describen accidentes de la muestra en lugar de patrones reproducibles.

77. ¿Qué es la poda por costo-complejidad y cómo se selecciona su intensidad?

La poda elimina ramas de un árbol previamente desarrollado para equilibrar ajuste y complejidad. En el esquema de costo-complejidad se considera R(T) + α|hojas(T)|, donde R representa el error o impureza agregada según la formulación y α penaliza cantidad de hojas. A mayor α, se favorecen árboles más pequeños.

La idea es que una rama debe justificar su complejidad mediante suficiente mejora. Una rama que reduce muy poco el error y añade muchas hojas puede desaparecer antes que otra con mayor aporte. El algoritmo construye una secuencia de subárboles candidatos; la intensidad final se elige con validación, no observando qué árbol obtiene menor error sobre el mismo test.

Por ejemplo, si dos candidatos tienen errores de referencia 20 y 18, con cuatro y ocho hojas, para α = 1 sus objetivos son 24 y 26. El árbol de cuatro hojas resulta preferible bajo esa función, aunque ajuste menos. Las cifras ilustran el equilibrio; en software debe revisarse la escala concreta de la impureza y del parámetro.

La poda y las restricciones previas pueden combinarse. No restauran información perdida por una división equivocada ni aseguran causalidad de las reglas. El resultado se evalúa por desempeño, estabilidad y facilidad de explicación. También se revisa si las hojas siguen representando segmentos suficientemente grandes y si las rutas cambian mucho al modificar la muestra.

78. ¿Por qué un árbol puede ser inestable y qué consecuencias tiene?

La inestabilidad significa que pequeños cambios de datos generan estructuras distintas. El entrenamiento greedy puede elegir entre divisiones de calidad muy parecida. Una modificación mínima cambia la primera decisión y, al hacerlo, altera qué casos llegan a cada nodo posterior. El efecto se propaga por el árbol.

Dos árboles pueden tener desempeño parecido y reglas diferentes. Esto importa si se pretende interpretar cada división como una regularidad firme del negocio. Una regla que desaparece al retirar pocos casos no debe presentarse como ley. Se contrastan árboles en remuestreos o folds, predicciones y variables utilizadas, además de la métrica promedio.

Por ejemplo, ingreso y patrimonio pueden aportar información similar. Un árbol elige ingreso en la raíz y otro patrimonio. La diferencia no demuestra que uno descubrió la causa verdadera ni que el otro sea incorrecto; puede revelar redundancia y selección local sensible a la muestra.

Bagging y random forest promedian modelos variados para reducir parte de esa varianza, pero sacrifican la representación mediante una única ruta sencilla. Un árbol pequeño también puede mejorar estabilidad al evitar ramas muy específicas. La respuesta depende de si se necesita una regla operativa compacta, máximo rendimiento o ambas cosas. La validación de estabilidad debe acompañar cualquier interpretación fuerte de una estructura aprendida.

79. ¿Cómo se evalúan e interpretan hojas, probabilidades e importancia de variables?

Una hoja de clasificación estima proporciones según los casos que contiene. Si reúne ocho positivos y dos negativos, la proporción es 0,8, pero con diez casos hay incertidumbre considerable y posibles sesgos. Esa cifra no garantiza una probabilidad calibrada en población nueva. Se revisa tamaño, pureza y estabilidad, y se evalúa calibración fuera de muestra cuando la probabilidad participa en decisiones.

La importancia basada en reducción de impureza suma mejoras asociadas a cada predictor. Describe uso dentro del ajuste, no causalidad. Puede favorecer variables con muchos puntos posibles de corte y distribuir relevancia de forma difícil de interpretar entre predictores correlacionados. Por eso no debe utilizarse como prueba de efecto ni como sustituto de evaluación.

La importancia por permutación altera una variable en datos de evaluación y mide pérdida de score del modelo. Evalúa dependencia predictiva bajo esa perturbación, pero también tiene límites por correlación y combinaciones fuera de la distribución. Se repite para medir variación y se relaciona con calidad del modelo base.

Para explicar una predicción se muestra la ruta, las condiciones y el soporte de la hoja. Para explicar el comportamiento global se examinan reglas y desempeño por regiones. Un umbral aprendido no necesariamente identifica una discontinuidad real del proceso. La interpretación completa conserva la distinción entre mecanismo del algoritmo y mecanismo causal del fenómeno.

80. Ejercicio resuelto: ¿cómo se decide entre un árbol pequeño y uno profundo?

Una empresa compara dos árboles para revisar operaciones riesgosas. En 1.000 casos de test, el pequeño obtiene TP = 60, FN = 40, FP = 40 y TN = 860. El profundo obtiene TP = 80, FN = 20, FP = 120 y TN = 780. Ambos enfrentan 100 positivos y 900 negativos.

El pequeño tiene accuracy 92 %, recall 60 % y precision 60 %. El profundo tiene accuracy 86 %, recall 80 % y precision 40 %. Si omitir un positivo cuesta 50 y revisar un negativo cuesta 5, el costo del pequeño es 40 × 50 + 40 × 5 = 2.200. El del profundo es 20 × 50 + 120 × 5 = 1.600. Bajo esos costos, el profundo es preferible pese a su menor accuracy.

Sin embargo, genera 200 alertas frente a 100 del pequeño. Si la capacidad máxima es 150, esa política no es factible tal como está. Debe elegirse otro umbral, limitar la revisión por valor esperado o usar un modelo alternativo, con esas decisiones determinadas previamente en validación.

El ejercicio ilustra la comparación en un conjunto final: no autoriza cambiar la configuración tras mirar test y seguir reportándolo como independiente. La selección previa debe incluir profundidad, calibración si corresponde, umbral, costo y capacidad. Interpretabilidad y rendimiento tienen valor, pero su peso debe explicarse según la operación concreta.

9 Random forest y ejercicios prácticos

81. ¿Qué es bagging y por qué puede mejorar a un árbol individual?

Bagging, o bootstrap aggregating, entrena varios modelos sobre muestras obtenidas con reemplazo y combina sus predicciones. Cada muestra tiene habitualmente el mismo tamaño que la base original, pero incluye repeticiones y omite algunas observaciones. La variación entre muestras produce modelos distintos, especialmente si el estimador base es inestable, como un árbol profundo.

En regresión se promedian predicciones. En clasificación se puede votar o promediar probabilidades según la implementación. La combinación reduce parte de la varianza cuando los errores no son perfectamente correlacionados. No elimina automáticamente sesgos compartidos por todos los modelos ni corrige información inválida.

Por ejemplo, tres árboles predicen una demanda de 90, 110 y 100. El promedio es 100. Si sus desviaciones tienen componentes distintos, la combinación puede ser más estable que cada predicción. Si todos usan una variable filtrada desde el futuro, promediarlos conserva el problema y podría reforzar una apariencia engañosa de buen desempeño.

El procedimiento se evalúa con la misma disponibilidad y particiones que el árbol individual. Más modelos suelen estabilizar el promedio hasta rendimientos decrecientes, pero aumentan tiempo y memoria. El objetivo es combinar estimadores suficientemente informativos y diversos. Entrenar cien copias idénticas sobre la misma muestra, sin ninguna fuente de variación, no obtiene la ventaja característica del bagging.

82. ¿Qué añade random forest al bagging de árboles?

Random forest incorpora selección aleatoria de un subconjunto de variables candidatas en cada división, además del remuestreo utilizado en su formulación habitual. En bagging de árboles con todas las variables disponibles, un predictor dominante puede aparecer repetidamente en las primeras divisiones y generar modelos muy correlacionados. Restringir candidatos permite explorar otros patrones.

La aleatoriedad busca reducir correlación entre árboles manteniendo capacidad predictiva. No se elige necesariamente un único subconjunto fijo de variables para todo el árbol: en la formulación usual se vuelve a seleccionar candidatos por nodo. La estrategia concreta depende de implementación y parámetros.

Si ingreso domina casi todas las divisiones, permitir que algunos nodos consideren solo antigüedad y deuda genera rutas alternativas. Eso puede hacer la combinación más estable, aunque cada árbol individual sea algo menos fuerte. Si se restringen demasiado las variables, algunos nodos reciben candidatos poco informativos y el sesgo puede aumentar.

Random forest modela relaciones no lineales e interacciones y suele requerir poca ingeniería de escala para los cortes. Aun así, necesita calidad, una representación adecuada de categorías y validación. No es inmune a extrapolación limitada, cambios de distribución ni costos de errores. Su ventaja potencial frente a un árbol se demuestra con desempeño y estabilidad fuera de muestra, no con la cantidad de árboles por sí sola.

83. Ejercicio resuelto: ¿qué proporción de casos queda fuera de una muestra bootstrap y qué es OOB?

En una muestra bootstrap de n extracciones con reemplazo sobre n casos, la probabilidad de no seleccionar una observación en una extracción es 1 - 1/n. La probabilidad de no elegirla en ninguna es (1 - 1/n)^n, que se aproxima a exp(-1) ≈ 0,3679 cuando n crece.

Por tanto, cada árbol deja fuera aproximadamente 36,8 % de casos originales y utiliza alrededor de 63,2 % de casos únicos, aunque su muestra contenga n filas por las repeticiones. Con n = 100, la probabilidad exacta de exclusión es 0,99^100 ≈ 0,3660. Los casos omitidos para ese árbol son out-of-bag, u OOB.

Una predicción OOB de un caso se construye solo con árboles que no lo utilizaron al entrenar. Esto ofrece una evaluación interna útil en condiciones apropiadas y permite seguir cómo cambia error al aumentar árboles. Con pocos árboles, algunos casos reciben muy pocas predicciones OOB o ninguna, lo que limita estabilidad.

OOB no sustituye automáticamente una evaluación temporal o por grupos. Si distintas filas del mismo cliente entran en árboles, un caso OOB puede seguir relacionado con otros usados. Tampoco corrige un preprocesamiento global que incorporó su información ni un hiperparámetro elegido tras muchas consultas al score OOB. Debe interpretarse según independencia, preparación y finalidad del experimento.

84. ¿Cómo explican la varianza y la correlación la ventaja de combinar árboles?

Si T predictores tienen varianza σ^2 y correlación común ρ, la varianza de su promedio es $\sigma^2[\rho + (1 - \rho)/T]$. La expresión se obtiene al sumar varianzas y covarianzas del promedio. Es una aproximación conceptual con supuestos homogéneos, no una fórmula exacta para todo bosque real.

Si $\sigma^2 = 100$, $\rho = 0{,}20$ y $T = 10$, la varianza es $100 \times (0{,}20 + 0{,}80/10) = 28$. Con $T = 100$, baja a 20,8. Al crecer T, el límite es $\rho\sigma^2 = 20$. Aumentar árboles no elimina el componente correlacionado.

Esto explica por qué la diversidad importa. Reducir correlación puede ser más valioso que agregar muchos modelos casi iguales. La selección aleatoria de candidatos y el bootstrap aportan diversidad, pero pueden reducir la fuerza individual si se usan de forma extrema. La calidad final equilibra ambos aspectos.

La reducción de varianza no garantiza eliminar sesgo ni superar toda alternativa. Si cada árbol representa mal una tendencia extrapolada, el promedio también puede hacerlo. En práctica se observa convergencia de scores y estabilidad de predicciones con una validación apropiada. La fórmula ayuda a interpretar rendimientos decrecientes y por qué un bosque más grande puede consumir más recursos sin resolver el principal límite del problema.

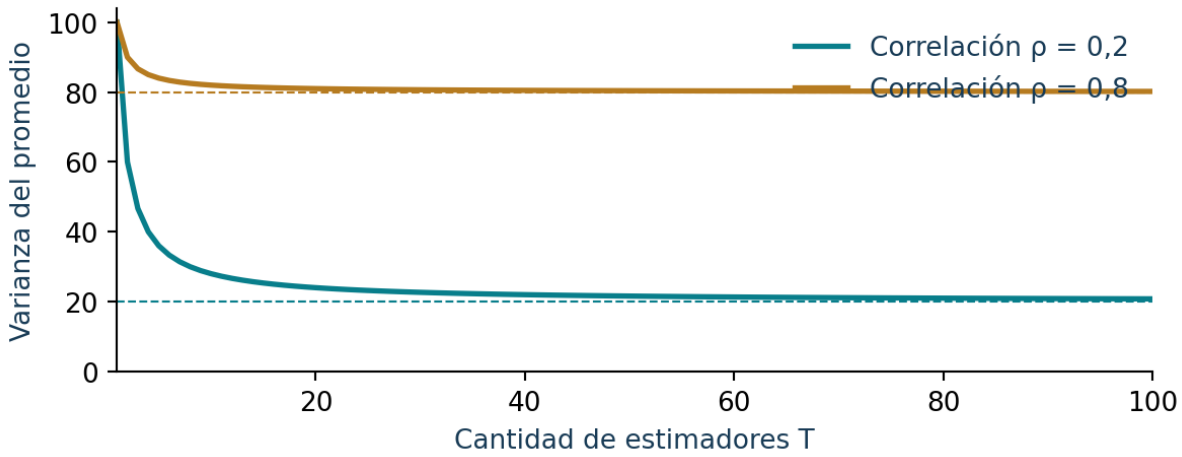


Figura 4. Aumentar estimadores reduce el componente independiente de varianza. La correlación fija un límite en este ejemplo idealizado.

85. ¿Qué controlan los principales hiperparámetros de random forest?

El número de árboles determina cuántos estimadores se combinan. Incrementarlo suele estabilizar el promedio y el score OOB hasta rendimientos decrecientes, con mayor costo. No cumple la misma función que profundidad: más árboles no equivale a hacer cada árbol más complejo y no suele provocar el mismo sobreajuste que profundizar una estructura individual.

Profundidad y tamaño mínimo de hoja regulan la flexibilidad de cada árbol. Hojas mayores pueden suavizar predicciones y mejorar estabilidad, pero omitir segmentos pequeños relevantes. max_features controla cuántos predictores pueden competir por nodo; valores menores aportan diversidad y pueden debilitar divisiones. max_samples, cuando se utiliza bootstrap, modifica el tamaño de cada muestra.

Otros parámetros afectan criterio, pesos de clases o casos y aleatoriedad. Las semillas facilitan reproducibilidad, pero no sustituyen verificar estabilidad. El soporte de faltantes y categorías puede variar por versión e implementación, por lo que debe definirse un tratamiento consistente en el pipeline.

La búsqueda se guía por una métrica ligada al uso y por particiones comunes. No conviene explorar combinaciones sin límite y reportar el mejor score como independiente. Primero puede elegirse una cantidad de árboles suficiente para estabilidad y después revisar complejidad y diversidad. Además de score, se mide memoria, latencia y comportamiento por segmento cuando esas restricciones afectan la operación.

86. ¿Cómo se comparan importancia por impureza e importancia por permutación?

La importancia por disminución de impureza acumula mejoras de divisiones y las combina entre árboles. Es rápida y está ligada al entrenamiento, pero puede favorecer variables con muchos valores o cortes posibles. También describe el uso que hizo el bosque de los predictores, no su efecto causal ni su valor en otros modelos.

La importancia por permutación mide cuánto empeora un score cuando se desordena una variable manteniendo el modelo fijo. Si el MAE base es 10 y pasa a 14 al permutar ingreso, el incremento es 4 bajo esa evaluación. Se repite la permutación para describir variabilidad y se prefiere evidencia independiente del ajuste para examinar generalización.

Variables correlacionadas complican la interpretación. Una puede compensar la perturbación de otra y ambas parecer poco importantes de forma individual. Además, permutar rompe relaciones conjuntas y puede crear combinaciones irreales. Permutar un grupo de variables o usar métodos condicionados puede aportar otra perspectiva, pero responde una pregunta diferente.

Antes de interpretar importancia se confirma que el modelo tenga rendimiento útil. Si predice mal, que una variable sea la más importante dentro de él dice poco sobre el fenómeno. Cuando la importancia participa en selección, debe evaluarse dentro del proceso de desarrollo sin utilizar test para decidir y después seguir llamándolo evaluación independiente.

87. ¿Qué muestran PDP e ICE y qué límites tienen con predictores correlacionados?

Un gráfico de dependencia parcial, PDP, varía un predictor y promedia predicciones sobre los demás atributos de una muestra. Describe el comportamiento del modelo bajo esas sustituciones. ICE realiza el recorrido para cada observación y permite ver heterogeneidad que el promedio oculta. Una curva global plana puede resultar de efectos individuales opuestos.

Si se estudia precio, un PDP podría mostrar que la demanda predicha disminuye al aumentarlo. Eso no demuestra un efecto causal del precio: el modelo aprendió asociaciones y el procedimiento modifica entradas sin simular necesariamente una intervención válida. Variables omitidas y selección comercial pueden explicar la relación.

Cuando precio y calidad están muy correlacionados, fijar precio alto en productos de baja calidad puede crear combinaciones casi ausentes del entrenamiento. El modelo entonces se evalúa en regiones sin soporte suficiente. El promedio puede parecer claro aunque sus predicciones se basen en extrapolaciones de la representación conjunta.

Se revisan rangos observados, distribuciones, ICE y restricciones del dominio. Métodos como efectos locales acumulados pueden reducir ciertas dificultades de correlación, pero tampoco convierten el análisis en causalidad. Una explicación responsable aclara qué variable se perturbó, sobre qué población se promedió y dónde hay evidencia. El gráfico permite inspeccionar una función aprendida; no identifica por sí solo cómo cambiará el mundo al modificar una variable.

88. Ejercicio resuelto: ¿por qué promediar probabilidades puede diferir del voto mayoritario?

Tres árboles asignan probabilidades positivas de 0,45; 0,45 y 0,90 a un caso. Si cada uno clasifica con corte 0,5, las etiquetas son negativo, negativo y positivo. El voto mayoritario produce negativo. Si se promedian probabilidades, la media es 1,80/3 = 0,60 y aplicar 0,5 produce positivo.

Las reglas difieren porque el voto descarta la intensidad de probabilidades dentro de cada árbol, mientras el promedio la conserva. En implementaciones habituales de random forest de clasificación, la predicción combina probabilidades de árboles y elige la clase correspondiente, por lo que no debe explicarse siempre como una votación de etiquetas duras.

La probabilidad promedio no es automáticamente calibrada. Si las hojas son pequeñas o el entorno difiere, el promedio puede sobreestimar o subestimar riesgos. Debe evaluarse con curvas de confiabilidad y pérdidas probabilísticas si se usará para calcular valor esperado. Posteriormente se elige un umbral según costo y capacidad.

En regresión, si árboles predicen 80, 100 y 120, la predicción media es 100. Su dispersión interna no constituye por sí sola un intervalo de predicción calibrado: los árboles están relacionados y no representan directamente todas las fuentes de incertidumbre. El ejercicio obliga a separar combinación, estimación probabilística y decisión, tres etapas que suelen confundirse al interpretar ensembles.

89. ¿Qué diferencia hay entre bagging, random forest y boosting?

Bagging ajusta modelos sobre remuestreos y combina sus predicciones para reducir varianza. Random forest aplica esa lógica a árboles y agrega selección aleatoria de variables candidatas por división para reducir correlación. Sus árboles pueden entrenarse en gran parte de forma independiente.

Boosting construye una secuencia en la que cada etapa atiende aspectos del error del modelo acumulado. En gradient boosting se ajustan aproximaciones al gradiente negativo de una pérdida. Profundidad, tasa de aprendizaje y cantidad de etapas controlan capacidad; no es correcto describirlo simplemente como otro bootstrap de árboles independientes.

Un bosque suele utilizar árboles más flexibles y compensar varianza mediante promedio. Un boosting puede combinar árboles pequeños para representar una función progresivamente, pero esas elecciones son configurables. La categoría del algoritmo no garantiza un patrón universal de sesgo, rendimiento o costo.

Para comparar se mantienen datos, features, particiones y métricas. Si un boosting mejora AUC apenas 0,004 pero multiplica por ocho latencia, debe examinarse valor operativo, incertidumbre y capacidad. Un incremento estadístico pequeño no justifica automáticamente adoptar más complejidad. Este contraste es un complemento al estudio de random forest: ayuda a reconocer qué mecanismo introduce diversidad y cómo se controla generalización, sin confundir dos procedimientos que combinan árboles de maneras diferentes.

90. ¿Cómo se desarrolla una práctica completa de random forest sin perder interpretabilidad?

La práctica comienza con una referencia simple, como un árbol podado o una regresión apropiada. Se define una división compatible con entidad y tiempo, una métrica principal y costos. Después se construye la preparación requerida y se ajusta el bosque con una búsqueda acotada de complejidad y diversidad dentro de validación.

Se compara rendimiento, dispersión entre folds, errores por segmentos y recursos. Si se utiliza OOB, se explican sus condiciones y se mantiene evaluación externa cuando el problema lo requiere. Una diferencia pequeña debe interpretarse junto con variación y costo. Más árboles pueden estabilizar, pero no compensan leakage o una base no representativa.

La interpretación incluye importancia por permutación y examen de relaciones del modelo con soporte observado. Para una decisión individual pueden analizarse rutas de árboles o explicaciones específicas, aclarando que un bosque no equivale a una única regla y que una atribución predictiva no demuestra causalidad.

El informe final presenta tanto ventajas como límites: extrapolación, calibración, dependencia de variables y cambios de distribución. Se documentan parámetros, versión, semillas y política de decisión. En producción se monitorean scores, entradas y resultados, y se define cuándo recalibrar o reentrenar. Una práctica completa conecta el mecanismo de combinación con una mejora verificable de la decisión, en lugar de concluir únicamente que el bosque obtuvo la mayor cifra de una tabla.

10 Clustering K-Means y jerárquico

91. ¿Qué es clustering y cómo se diferencia de clasificación?

Clustering busca agrupar observaciones según una noción de similitud sin utilizar etiquetas objetivo para definir la partición. Clasificación aprende a predecir clases conocidas a partir de ejemplos etiquetados. Un cluster no equivale automáticamente a una categoría real ni a una clase que deba recibir una acción determinada.

La similitud depende de variables, escalas y distancia. Segmentar clientes por frecuencia y gasto responde una pregunta distinta de agruparlos por composición de productos. Incluir identificadores numéricos o características no pertinentes puede producir grupos artificiales aunque el algoritmo converja perfectamente.

Por ejemplo, marketing podría buscar clientes frecuentes de bajo ticket y clientes esporádicos de alto gasto. Una segmentación es útil si sus perfiles son estables, pueden identificarse y permiten acciones diferenciadas. El algoritmo no demuestra que cada grupo responderá a una campaña específica; esa respuesta debe evaluarse, idealmente mediante un diseño de intervención.

La validación combina estructura interna, estabilidad y utilidad. Si hay etiquetas externas pertinentes, pueden compararse como evidencia adicional, pero no deben utilizarse para afirmar que cualquier desacuerdo invalida la segmentación: podrían representar otro concepto. El análisis debe declarar qué se buscó agrupar y por qué. Forzar un número de clusters siempre produce una partición, incluso cuando no existen grupos nítidos en el fenómeno.

92. ¿Cómo funciona K-Means y qué función objetivo minimiza?

K-Means busca K centroides y una asignación que minimicen la suma de distancias euclídeas cuadradas dentro de grupos: $J = \Sigma\Sigma||x_i - \mu k||^2$. Cada centroide es la media de su cluster, que minimiza la suma cuadrática para una asignación fija. El algoritmo alterna asignar cada punto al centro más cercano y recalcular centros.

Las etapas no aumentan la función objetivo en el esquema estándar y el procedimiento llega a una solución local o estable bajo criterios de detención. No garantiza el mínimo global. La inicialización importa, por lo que se utilizan métodos como K-Means++ y varios reinicios. K-Means++ favorece centros inicialmente separados mediante probabilidades relacionadas con distancias cuadradas, pero tampoco garantiza la mejor partición final.

La geometría favorece grupos compactos según la distancia elegida, con fronteras de Voronoi. Puede tener dificultades con formas curvas, densidades muy distintas o outliers. Tampoco es naturalmente adecuado para mezclar variables nominales con numéricas sin una representación y distancia justificadas.

K debe definirse antes del ajuste y seleccionarse mediante diagnóstico, estabilidad y finalidad. Los números de etiquetas son arbitrarios: cluster 0 no es inferior a cluster 1, y el mismo grupo puede recibir otro número en una ejecución distinta. Interpretar exige examinar perfiles y correspondencia entre soluciones, no comparar números de etiquetas literalmente.

93. Ejercicio resuelto: ¿cómo se realiza una iteración de K-Means?

Considérese una variable con valores [1, 2, 8, 9], K = 2 y centros iniciales 1 y 9. En la etapa de asignación, 1 y 2 están más cerca de 1; 8 y 9, de 9. Se forman grupos {1, 2} y {8, 9}.

Al actualizar, sus medias son 1,5 y 8,5. Las distancias cuadradas dentro del primer grupo son 0,25 y 0,25; en el segundo, también 0,25 y 0,25. La inercia total queda en 1. Una nueva asignación conserva los grupos, por lo que el procedimiento está estable para esa inicialización.

Si K = 1, el centro sería 5 y J = 16 + 9 + 9 + 16 = 50. Si K = 4, cada punto podría tener su centro y J = 0. El descenso de inercia al aumentar K es esperado y no demuestra por sí solo que cuatro grupos sean una mejor explicación o una segmentación útil.

Los centroides son medias, no necesariamente observaciones existentes. Un valor extremo puede desplazarlos y alterar asignaciones. En varios predictores se realiza el mismo razonamiento mediante vectores y distancias, pero el escalado determina qué dimensiones dominan. El ejemplo permite reconstruir el algoritmo y distinguir convergencia matemática de validez del número de grupos elegido.

94. ¿Por qué escala, redundancia y outliers afectan una segmentación?

En distancia euclídea, cada diferencia contribuye al cuadrado. Si ingreso varía en cientos de miles y frecuencia en unidades, ingreso puede dominar aunque ambas dimensiones interesen. Estandarizar hace comparables sus dispersiones, pero también define un criterio: una desviación estándar en cada predictor recibe un peso semejante. Esa decisión debe corresponder al problema.

La redundancia puede sobreponderar una dimensión. Si se incluyen gasto anual y gasto mensual casi equivalentes, el comportamiento monetario participa dos veces en la distancia. Revisar correlaciones, combinar variables o usar una representación reducida evita que la lista de columnas determine accidentalmente la ponderación.

Los outliers afectan centros de K-Means porque la función cuadrática penaliza mucho distancias grandes. Pueden atraer un centro o distorsionar perfiles. Se revisan validez y segmentos, y se comparan escalado robusto o métodos alternativos. Eliminar casos raros sin análisis puede borrar clientes importantes que merecen un grupo propio.

No se deben estandarizar indiscriminadamente todas las codificaciones. Una categoría nominal convertida a entero sigue teniendo una distancia arbitraria después del escalado. Tampoco debe incluirse información posterior si la segmentación se usará para asignar acciones futuras. La preparación define la geometría que el algoritmo aprende; por eso su justificación es parte central de la interpretación y de la reproducibilidad.

95. ¿Cómo se elige K con inercia, silhouette y estabilidad?

La inercia disminuye al aumentar K porque hay más centros para aproximar datos. El método del codo busca un punto de rendimientos decrecientes, pero puede no existir un codo claro. No es una prueba estadística de que haya exactamente K grupos verdaderos.

Silhouette compara cohesión y separación. Para un punto, a es la distancia media a su propio grupo y b la menor distancia media a otro grupo. Se define s = (b - a)/max(a, b), entre -1 y 1. Si a = 2 y b = 6, s = 4/6 ≈ 0,667. Valores negativos sugieren que el punto está, en promedio, más cerca de otro grupo. Para grupos de un único punto suelen usarse convenciones específicas, como s = 0.

El promedio resume, pero puede ocultar un cluster problemático. Se observan tamaños y distribución por grupo. Un score alto favorece cierta geometría y distancia, y no garantiza accionabilidad ni superioridad para todos los usos. Tampoco permite comparar sin cautela representaciones muy distintas.

La estabilidad se evalúa con reinicios y remuestreos, alineando etiquetas o usando medidas invariantes como adjusted Rand index cuando corresponde. Si quitar pocos casos cambia todo, la segmentación es frágil. La selección final considera también tamaños mínimos, perfiles comprensibles y acciones. Una partición matemática convincente puede ser inútil si no permite distinguir ofertas o identificar clientes nuevos de manera consistente.

96. ¿Cómo funciona el clustering jerárquico y cómo se interpreta un dendrograma?

El clustering jerárquico aglomerativo comienza con cada observación como un grupo y fusiona sucesivamente los más próximos según un criterio. Produce una estructura anidada. El divisivo recorre la lógica inversa, partiendo de un grupo y separándolo. La versión aglomerativa es frecuente y necesita especificar distancia y linkage.

El dendrograma representa fusiones. Las hojas corresponden a observaciones o grupos comprimidos y la altura de una unión expresa la disimilitud o criterio utilizado por la implementación. El orden horizontal no debe interpretarse como una distancia numérica general: distintas ordenaciones pueden representar la misma jerarquía.

Un corte horizontal define una partición según el nivel de fusión permitido. Cortar alto produce menos grupos y cortar bajo más. La elección se relaciona con saltos de altura, estabilidad y finalidad. El número de grupos no queda validado simplemente por dibujar un dendrograma atractivo.

Las fusiones tempranas suelen ser irreversibles en el procedimiento básico, por lo que errores o outliers pueden afectar la estructura. El costo de memoria y tiempo puede ser considerable en bases grandes. Frente a K-Means, la jerarquía permite inspeccionar varios niveles sin fijar inicialmente K, pero no elimina decisiones: distancia, preparación y linkage determinan qué estructura se construye y cómo debe interpretarse.

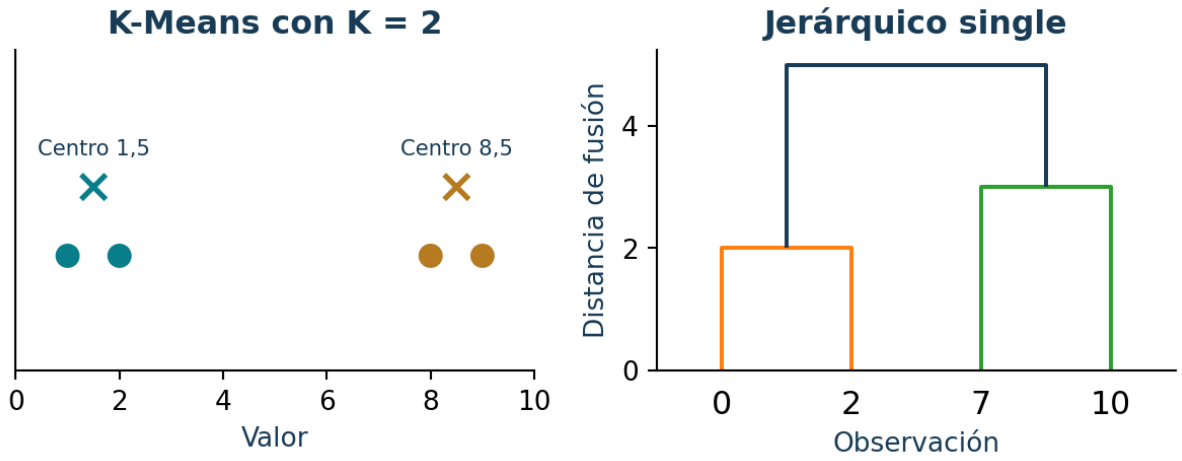


Figura 5. A la izquierda, los centros del ejercicio de K-Means. A la derecha, las fusiones single para 0, 2, 7 y 10; la altura expresa distancia.

97. ¿Cómo se comparan single, complete, average y Ward linkage?

Single linkage usa la menor distancia entre puntos de dos grupos. Puede capturar formas alargadas, pero sufre encadenamiento: puntos intermedios conectan grupos que visualmente parecen distintos. Complete usa la mayor distancia y favorece grupos compactos, aunque puede ser sensible a extremos. Average utiliza un promedio de distancias entre pares y ofrece otra noción de proximidad.

Ward elige fusiones con menor incremento de suma cuadrática dentro de grupos. Para grupos A y B, el incremento es $[nA\ nB/(nA + nB)] \times ||\mu A - \mu B||^2$ en la formulación euclídea habitual. Debe usarse una geometría compatible: aplicar Ward con una distancia arbitraria puede carecer de esa interpretación.

Con grupos A = {0, 2} y B = {7, 10}, single vale 5, complete 10 y average (7 + 10 + 5 + 8)/4 = 7,5. Sus medias son 1 y 8,5, así que el incremento Ward es $1 \times 7{,}5^2 = 56{,}25$. Las alturas reportadas por una biblioteca pueden transformar o escalar ese incremento; no se deben comparar numéricamente entre linkages sin verificar su definición.

No existe un linkage superior para toda estructura. Se examinan sensibilidad, tamaños y perfiles bajo una preparación común. La elección describe qué considera cercano el procedimiento y puede cambiar radicalmente la segmentación. Interpretar grupos sin informar esa decisión oculta una parte esencial del análisis.

98. ¿Qué aporta PCA antes de clustering y qué límites tiene?

PCA construye combinaciones lineales ortogonales que capturan sucesivamente la mayor varianza de los datos centrados. Reducir dimensiones puede eliminar redundancia y facilitar visualización o cálculo. La representación cambia si se usa covarianza sobre unidades originales o correlación mediante variables estandarizadas.

La varianza explicada de un componente es su autovalor dividido por la suma total. Si dos componentes explican 80 %, se conserva esa proporción de variación bajo la preparación elegida; no significa que preserven 80 % de la información útil para distinguir clientes ni que haya grupos reales.

Un patrón discriminante puede estar en una dirección de poca varianza. Reducirla podría borrar un segmento relevante. Asimismo, clusters que parecen separados en dos componentes pueden solaparse en otras dimensiones, y una figura sin separación visual no demuestra ausencia de estructura. Los centroides en componentes se interpretan mediante cargas o reconstrucción y comparación de atributos originales.

Si PCA forma parte de un sistema para asignar casos nuevos, sus parámetros se aprenden en la base de referencia y se aplican sin reestimarlos con cada lote. Se evalúan soluciones con y sin reducción y su estabilidad. PCA organiza variación lineal; no es un algoritmo que certifique número, validez ni utilidad de clusters.

99. ¿Cómo se valida y describe una segmentación sin inventar perfiles?

La descripción utiliza estadísticas de variables originales por cluster: tamaño, medias o medianas, dispersión, categorías y evolución. Un nombre debe surgir de diferencias observadas y no de estereotipos. Si un grupo tiene alto gasto, eso no prueba alta rentabilidad; se necesita información sobre margen y costo de servicio.

La validación interna examina cohesión y separación según la geometría usada. La estabilidad revisa si perfiles y asignaciones se mantienen en reinicios, remuestreos y períodos. La validación externa puede comparar indicadores pertinentes no utilizados en el ajuste, pero su elección debe evitar circularidad: definir grupos por gasto y después afirmar que son buenos porque difieren en gasto no aporta evidencia independiente.

La utilidad operativa requiere que los grupos puedan identificarse, sean suficientemente grandes y permitan acciones diferenciadas. Si un segmento representa 0,1 % de clientes, puede ser valioso o impracticable según el negocio. También debe establecerse cómo se asignan nuevos casos y qué hacer con observaciones fuera de referencia.

Finalmente, una campaña por segmentos necesita evaluar respuesta y valor incremental. El clustering describe similitud, no el beneficio causal de intervenir. Un informe completo muestra perfiles, estabilidad, límites y una hipótesis de acción que pueda probarse. No concluye que la segmentación genera valor solo porque tiene un silhouette alto o nombres atractivos.

100. Ejercicio integrador: ¿cómo se comparan K-Means y jerárquico para segmentar clientes?

Supongamos una base con recencia, frecuencia y gasto. Primero se define una fecha de corte común: recencia usa días desde la última compra hasta ese corte, y frecuencia y gasto abarcan una ventana anterior. Se auditan clientes duplicados, faltantes, monedas y extremos válidos. Un logaritmo sobre gasto puede ser razonable si interesa reducir dominio de la cola, y el escalado define pesos comparables.

Se ajusta K-Means para varios K con reinicios y se construye una jerarquía con distancia y linkage declarados. Se comparan tamaños, silhouette, perfiles y estabilidad. Si K-Means genera un grupo casi exclusivo de una compra extrema, se investiga si es un segmento útil o una distorsión de la función cuadrática. No se elimina el cliente automáticamente para mejorar el gráfico.

Para casos nuevos, K-Means permite aplicar las transformaciones de referencia y asignar al centro más cercano. Una jerarquía aglomerativa ordinaria no proporciona siempre una regla nativa de predict; se necesita definir una extensión, como asignación a perfiles, y evaluar sus consecuencias. Reajustar toda la jerarquía con nuevos casos puede cambiar grupos previos.

El resultado elige una solución defendible por estabilidad y uso, y documenta una alternativa. Si hay formas no compactas o ruido, un método basado en densidad como DBSCAN puede aportar contraste, aunque tiene sus propios parámetros y límites. Con esta pregunta termina el cuestionario de 100 preguntas sobre los temas del parcial.

11 Autoencoders complemento de diez preguntas

101. ¿Qué es un autoencoder y por qué reconstruir la entrada puede producir aprendizaje útil?

Un autoencoder es una red que transforma una entrada x en una representación z mediante un encoder y reconstruye una salida $\hat{x}$ mediante un decoder. Se escribe $z = f\theta(x)$ y $\hat{x} = g\varphi(z)$. El entrenamiento busca que la reconstrucción se aproxime a la entrada bajo una pérdida, con restricciones que favorezcan aprender estructura y no simplemente copiar cada dato.

Se considera aprendizaje no supervisado respecto de etiquetas externas, o auto-supervisado porque la propia entrada proporciona el objetivo. No requiere una clase de fraude o una etiqueta de cliente para entrenar reconstrucción. Eso no significa que pueda aprender sin supuestos: arquitectura, pérdida y regularización definen qué información resulta útil conservar.

Por ejemplo, ocho medidas de consumo pueden depender de pocos patrones latentes, como intensidad y horario. Un código de menor dimensión obliga a resumir parte de esa estructura. Si las medidas contienen correlaciones estables, la reconstrucción puede ser buena sin almacenar las ocho cantidades de manera independiente.

Una red con suficiente capacidad y sin restricciones puede aproximar la identidad y producir una representación poco útil. Bajo error bajo no debe concluirse automáticamente que existe un buen embedding o un buen detector de anomalías. La utilidad se evalúa según una tarea posterior, reconstrucción independiente, estabilidad y sensibilidad. Aprender a reproducir la muestra no equivale a comprender causas ni garantiza generalización.

102. ¿Qué funciones cumplen el encoder, el espacio latente, el bottleneck y el decoder?

El encoder transforma variables observadas en características latentes. El espacio latente contiene esas representaciones y su dimensión controla parte de la capacidad del sistema. El bottleneck es una restricción, frecuentemente dimensional, que limita la información transmitida. El decoder aprende a reconstruir entradas desde ese código.

Una arquitectura 8 → 4 → 2 → 4 → 8 recibe ocho atributos, genera dos coordenadas latentes y vuelve a producir ocho. En la versión undercomplete, la dimensión del código es menor que la entrada. En una overcomplete puede ser igual o mayor, pero se necesitan otras restricciones para evitar una copia trivial, como sparsity, ruido o penalización de sensibilidad.

La dimensión no es el único control de capacidad. Capas profundas y no lineales pueden memorizar estructuras complejas incluso con un código pequeño en una muestra finita. La calidad depende de tamaño de datos, regularización, entrenamiento y evaluación. Un bottleneck reduce libertad, pero no garantiza una representación semántica determinada.

Las coordenadas latentes no son automáticamente conceptos interpretables. La primera no tiene por qué representar gasto y la segunda riesgo. Rotaciones o reparametrizaciones pueden conservar reconstrucción. Para interpretarlas se relacionan con atributos y tareas, revisando estabilidad. También se decide si la aplicación necesita reconstruir, comparar embeddings, reducir dimensión o detectar anomalías, porque esos objetivos pueden requerir arquitecturas y pérdidas diferentes.

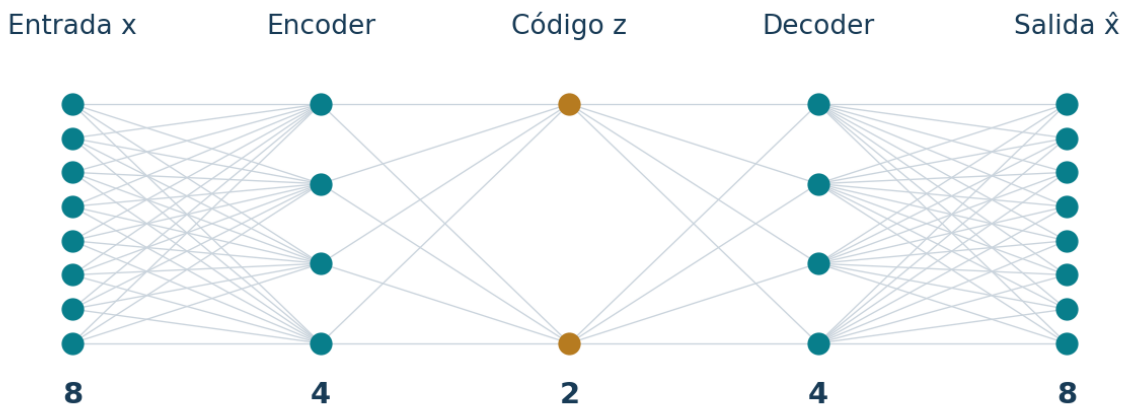


Figura 6. Arquitectura del ejemplo 8 → 4 → 2 → 4 → 8. La restricción central fuerza una representación de dimensión menor que la entrada.

103. Ejercicio resuelto: ¿cómo se calcula el error de reconstrucción y se elige una pérdida?

Para una entrada estandarizada x = [1, 0, -1] y una reconstrucción $\hat{x}$ = [0,8; 0,1; -0,7], las diferencias son [0,2; -0,1; -0,3]. La suma cuadrática es 0,04 + 0,01 + 0,09 = 0,14 y el MSE por atributo es 0,14/3 ≈ 0,0467. El MAE es (0,2 + 0,1 + 0,3)/3 = 0,20.

MSE penaliza más errores grandes y suele ser apropiado para cantidades continuas bajo una interpretación de ruido gaussiano. MAE corresponde a otra función de error y puede ser menos sensible a desviaciones extremas. Para variables binarias puede utilizarse entropía cruzada con probabilidades; para categorías, salidas y pérdidas específicas que respeten sus modalidades.

Si atributos están en escalas distintas, el error puede quedar dominado por uno. Estandarizar o ponderar pérdidas permite controlar contribuciones, pero define un criterio de relevancia. Una diferencia de un desvío estándar en cada variable no tiene necesariamente igual costo económico. En datos mixtos pueden combinarse pérdidas por tipo de atributo.

La activación de salida debe ser coherente con el dominio. Una salida lineal permite números negativos y es razonable para variables estandarizadas. Una sigmoide restringe a [0, 1], pero no debe imponerse si la representación contiene valores fuera de ese rango. Un error bajo se interpreta bajo la pérdida y escala elegidas, no como un porcentaje universal de información preservada.

104. ¿Cómo se entrena un autoencoder y cómo se evita sobreajuste?

El entrenamiento procesa lotes, calcula reconstrucciones y deriva la pérdida respecto de parámetros del encoder y decoder mediante backpropagation. Un optimizador actualiza pesos para reducirla. Los parámetros son pesos aprendidos; los hiperparámetros incluyen arquitectura, dimensión latente, tasa de aprendizaje, tamaño de lote, regularización y número de épocas.

En una regresión cuadrática elemental, la derivada de un error $(x - \hat{x})^2$ respecto de $\hat{x}$ es $2(\hat{x} - x)$. En la red, la regla de la cadena propaga ese gradiente a las capas anteriores. Las no linealidades permiten relaciones complejas, pero también incrementan capacidad y dificultad de optimización.

Se separan entrenamiento y validación de reconstrucción con un esquema compatible con grupos y tiempo. Early stopping detiene o selecciona una época según validación y conserva pesos adecuados. Si la pérdida de entrenamiento disminuye mientras la de validación empeora, puede haber sobreajuste. Regularización, arquitectura más pequeña o más datos representativos pueden ayudar.

La normalización se aprende con entrenamiento. Consultar test para elegir épocas o dimensiones elimina su independencia. También conviene repetir con semillas cuando la estabilidad sea relevante y evaluar la tarea final: una red que reconstruye mejor puede producir embeddings menos útiles para segmentación. El entrenamiento optimiza una pérdida concreta; la validación debe comprobar si ese objetivo genera la utilidad pretendida.

105. ¿Cómo se relaciona un autoencoder lineal con PCA?

En datos centrados, un autoencoder lineal con código de dimensión k y pérdida cuadrática puede aprender el mismo subespacio principal que PCA bajo condiciones adecuadas y un óptimo global. Ambas técnicas buscan una reconstrucción de rango reducido que minimice error cuadrático. Esa equivalencia se refiere al subespacio reconstruido, no a cualquier configuración de red.

Las coordenadas del encoder no tienen por qué coincidir con los componentes de PCA: pueden representar una base transformada del mismo subespacio. PCA además ordena componentes por varianza y utiliza direcciones ortogonales. Un autoencoder lineal sin restricciones adicionales no impone necesariamente esa parametrización en sus pesos.

Con activaciones no lineales, un autoencoder puede representar una variedad curva que una proyección lineal no reconstruye bien. Eso amplía capacidad, pero demanda más datos, entrenamiento y validación. No garantiza superar PCA. Si los datos tienen estructura aproximadamente lineal, PCA puede ser más estable, simple y fácil de interpretar.

Se compara reconstrucción fuera de muestra, tamaño de representación y rendimiento en la tarea posterior. También se verifica que ambos reciban la misma preparación: aplicar PCA a variables originales y el autoencoder a variables estandarizadas confunde algoritmo con escala. La relación ayuda a comprender autoencoders como reducción y reconstrucción aprendidas, y evita presentar toda representación neuronal como un descubrimiento necesariamente más profundo.

106. ¿Qué diferencias existen entre autoencoders sparse, denoising y contractive?

Un sparse autoencoder penaliza o restringe la activación de unidades para que pocas participen simultáneamente. Puede utilizar penalizaciones sobre activaciones o un objetivo de frecuencia de activación. La sparsity no equivale a que el código tenga pocas dimensiones: una representación grande puede activar solo una parte por observación.

Un denoising autoencoder recibe una entrada corrompida $\tilde{x}$, pero se entrena para reconstruir la entrada limpia x. El ruido puede ser gaussiano, enmascaramiento u otra perturbación coherente con los datos. La tarea obliga a utilizar relaciones entre atributos para recuperar información, en lugar de copiar exactamente la entrada recibida.

Un contractive autoencoder penaliza sensibilidad del encoder a pequeñas variaciones de entrada, por ejemplo mediante la norma del Jacobiano $\partial f/\partial x$. Busca representaciones localmente estables. La resistencia a ciertas perturbaciones no garantiza robustez a cualquier cambio de población o anomalía.

Si ocho sensores miden un mismo proceso, enmascarar uno y reconstruirlo puede aprender relaciones redundantes. Pero una corrupción irreal, como alterar categorías sin respetar restricciones, enseña una tarea poco relacionada con el uso. La regularización se selecciona con validación y se compara con referencias. Las variantes introducen propiedades deseadas en el código; no son intercambiables ni deben elegirse solo porque tengan una arquitectura más compleja.

107. ¿Cómo se usa un autoencoder para detectar anomalías y qué supuestos necesita?

Se entrena una reconstrucción sobre una referencia representativa del comportamiento normal y se calcula un score, como MSE por caso. Si una observación no pertenece a la estructura aprendida, puede reconstruirse peor y recibir una alerta. La estrategia depende de que el modelo aprenda normalidad y de que las anomalías relevantes generen un error distinguible.

Ese supuesto puede fallar. Una red muy flexible puede reconstruir también anomalías; una referencia contaminada puede aprenderlas; un caso normal de un segmento poco representado puede tener error alto. También hay anomalías que ocurren dentro de la variedad aprendida y no generan gran error. Por eso reconstrucción alta no equivale a fraude y reconstrucción baja no prueba normalidad.

El umbral se calibra con validación normal y, si existen, anomalías etiquetadas. Usar el percentil 99 del error normal fija aproximadamente una tasa nominal de alertas en esa referencia, no una precision de 99 % ni una garantía de tasa futura. Se evalúan precision, recall, capacidad y costo cuando hay etiquetas adecuadas.

Los errores por atributo ayudan a revisar el caso, aunque no son automáticamente causas del evento. En producción se monitorean distribución de scores y drift. El autoencoder debe competir con métodos más simples, como una referencia estadística o Isolation Forest. La elección se justifica por desempeño independiente y utilidad de revisión, no por el carácter neuronal del algoritmo.

108. ¿Cómo se preparan datos numéricos, categóricos y faltantes para un autoencoder?

La preparación debe respetar tipos y disponibilidad. Los atributos numéricos pueden estandarizarse o transformarse según distribución, con parámetros de entrenamiento. Si se usa salida lineal sobre variables estandarizadas, la red puede reconstruir cualquier valor real. Una salida acotada requiere una escala compatible y revisar qué ocurre con casos futuros fuera del rango histórico.

Categorías pueden representarse mediante indicadores o embeddings y salidas específicas. Si hay varias categorías mutuamente excluyentes, una salida softmax y una pérdida correspondiente preservan esa estructura mejor que tratar cada código como una cantidad continua. La decisión depende de arquitectura y objetivo.

Los faltantes no deben introducirse como NaN en una pérdida ordinaria. Puede imputarse con reglas aprendidas y añadir indicadores, o utilizar una pérdida enmascarada que no penalice atributos no observados. Si se usa el error imputado para detectar anomalías, debe explicarse que parte de la entrada fue estimada y no medida.

Para comparar scores entre casos con distinto número de atributos observados, una pérdida enmascarada puede normalizar por cantidad disponible. Aun así, poca información limita la confiabilidad del score. Nunca se debe confundir una entrada realmente faltante con una corrupción artificial de un denoising autoencoder. El pipeline completo se conserva para producción y se revisa si patrones de ausencia o categorías nuevas alteran la representación.

109. ¿Qué distingue a un autoencoder variacional de uno convencional?

Un autoencoder convencional suele producir un código determinista. Un VAE modela una distribución latente aproximada q(z|x), frecuentemente parametrizada por media y varianza, y un decoder probabilístico p(x|z). Su entrenamiento combina reconstrucción esperada con un término de divergencia respecto de una distribución previa sobre z.

En la formulación básica se minimiza el negativo de una cota variacional: error de reconstrucción probabilístico más KL[q(z|x) || p(z)]. La reparametrización $z = \mu(x) + \sigma(x) \odot \varepsilon$, con ε normal estándar, permite propagar gradientes. El término KL regulariza la distribución latente y hace posible muestrear códigos del prior para generar datos.

Para un VAE gaussiano diagonal con prior normal estándar, la KL incluye $0{,}5\Sigma[\mu j^2 + \sigma j^2 - \ln(\sigma j^2) - 1]$. Reconstrucción y regularización deben equilibrarse; una representación puede colapsar o reconstruir pobremente según el problema.

No toda red que comprime datos es generativa de la misma manera, y un VAE no es automáticamente un mejor detector de anomalías. Se evalúan calidad y utilidad de muestras o scores de acuerdo con la finalidad. Esta pregunta introduce el mecanismo probabilístico; su implementación exige especificar distribuciones y pérdidas compatibles con los atributos.

110. Ejercicio integrador: ¿cómo se diseña y evalúa un autoencoder de anomalías de principio a fin?

Supongamos ocho medidas operativas y una referencia histórica verificada. Se reserva entrenamiento normal, validación normal para seleccionar capacidad y umbral, y un test independiente con casos normales y anómalos etiquetados. Si hay repetición por máquina o dependencia temporal, la división se ajusta para reproducir el uso previsto. La estandarización se aprende únicamente con entrenamiento.

Se compara PCA de dos componentes, una arquitectura 8 → 4 → 2 → 4 → 8 y un detector alternativo. Se entrena el autoencoder con MSE, regularización razonable y early stopping usando validación. Los hiperparámetros no se eligen mirando test. Después se fija, por ejemplo, un umbral según el percentil 99 del score normal de validación o mediante una función de costo si hay anomalías disponibles para selección.

En 200 casos de test, hay 20 anomalías. Si se detectan 16 y se alertan 9 normales, TP = 16, FN = 4, FP = 9 y TN = 171. Precision = 16/25 = 0,64; recall = 16/20 = 0,80 y accuracy = 187/200 = 0,935. El 1 % nominal elegido en validación no garantizó esa misma tasa en test: hay que investigar variación, distribución y soporte.

Se informa costo, volumen de revisión, resultados por tipo de anomalía y atributos con mayor reconstrucción. Un caso señalado requiere verificación. Finalmente se conservan pipeline, pesos, umbral y referencia, y se establece monitoreo de drift. Así termina la guía: el objetivo es poder explicar el mecanismo, resolver el cálculo y defender la decisión y sus límites en cada una de las 110 preguntas.

Fuentes y correspondencia con el repositorio

La base de revisión es el repositorio data-science-business-decisions, en su estado consultado el 11 de octubre de 2026. Se revisaron las guías de módulos y los cuestionarios básicos, junto con los materiales avanzados disponibles de EDA, outliers, transformación, series de tiempo, regresión logística y árboles y ensembles. La referencia de revisión del repositorio es el commit d26008cff43adf8cae6b8aaf065b17bdd38e02ed.

Las respuestas de esta guía son desarrollos propios para el estudio: amplían mecanismos, ejemplos, cálculos y límites de interpretación. Las referencias técnicas complementarias se utilizaron para contrastar detalles de implementación y conceptos específicos. No se reproducen artículos ni capítulos completos de las fuentes externas.

Preguntas	Material del repositorio revisado
1 a 10	Módulo 01 EDA y cuestionario avanzado de Paradigma EDA
11 a 20	Módulo 02 Calidad de datos y valores faltantes
21 a 30	Módulo 03 Outliers, robustez e influencia y cuestionario avanzado
31 a 40	Módulos 04 Transformación y pipelines y 05 Regresión lineal
41 a 50	Módulos 05 Regresión lineal y 06 Validación de modelos
51 a 60	Módulo 07 Series de tiempo y sitio avanzado
61 a 70	Módulo 08 Regresión logística y cuestionario avanzado
71 a 90	Módulo 09 Árboles y ensembles y cuestionario razonado
91 a 100	Módulo 10 PCA, K-Means y clustering jerárquico
101 a 110	Complemento solicitado sobre autoencoders y referencias técnicas

Referencias técnicas complementarias

Scikit-learn. Documentación oficial sobre modelos lineales. Contraste de formulaciones y regularización. Consultar documentación.

Scikit-learn. Documentación oficial sobre validación cruzada. Contraste de los esquemas de partición y su correspondencia con grupos y tiempo. Consultar documentación.

Scikit-learn. Common pitfalls and recommended practices. Contraste del tratamiento de preprocesamiento y leakage. Consultar documentación.

Scikit-learn. Metrics and scoring. Contraste de definiciones e interpretación de métricas. Consultar documentación.

Scikit-learn. Imputation of missing values. Contraste de familias de imputadores y preparación reproducible. Consultar documentación.

Scikit-learn. Ensembles. Contraste de bootstrap, random forest y combinación de predicciones. Consultar documentación.

Scikit-learn. Clustering. Contraste de K-Means, jerárquico y criterios de evaluación. Consultar documentación.

Goodfellow, I., Bengio, Y. y Courville, A. (2016). Deep Learning, capítulo 14, Autoencoders. Referencia conceptual para el complemento. Consultar capítulo.

Kingma, D. P. y Welling, M. (2014). Auto-Encoding Variational Bayes. Introducción al mecanismo probabilístico de los VAE de la pregunta 109. Consultar trabajo.

Keras. Convolutional autoencoder for image denoising. Ejemplo oficial de entrenamiento con entrada corrompida y objetivo limpio. Consultar ejemplo.

Consulta de referencias en línea: 11 de octubre de 2026. Las explicaciones matemáticas se expresan de manera independiente de una versión específica de software; los parámetros y capacidades concretas deben contrastarse con la documentación del entorno utilizado.