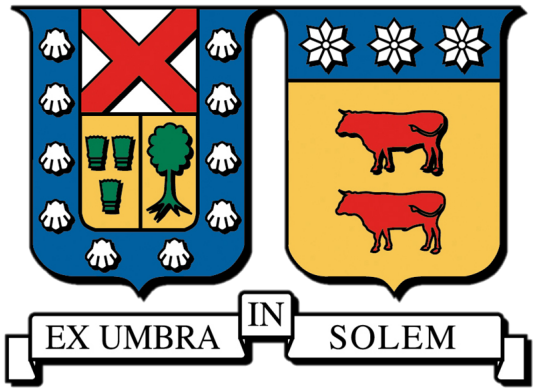
Universidad Técnica Federico Santa María
Departamento de Informática
Valparaíso, Chile



Relevancia de atributos y fronteras epistémicas en reconocimiento de patrones

De la relevancia operacional a la informatividad poblacional

Pablo Neirz

Tesis para optar al grado de

DOCTOR EN INGENIERÍA INFORMÁTICA

Directora de Tesis:
Dra. Carolina Saavedra

Co-Director de Tesis:
Dr. Héctor Allende Olivares

Valparaíso, Chile, 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ■Tesis de Postgrado

Título del trabajo: "Relevancia de atributos y fronteras epistémicas en reconocimiento de patrones:

De la relevancia operacional a la informatividad poblacional"

Nombre del candidato(a): Pablo Antonio Neirz Vergara

Carrera / Grado: Doctorado en Ingeniería Informática

Campus: Casa Central **Departamento:** Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Carolina Saavedra Ruiz, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

■ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 14-08-2026 **Firma:** 

Estudiante o Candidato(a): Pablo Neirz Vergara

Fecha:14-08-2026 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

COMITÉ EVALUADOR

Dra. Carolina Saavedra
Universidad Técnica Federico Santa María, Chile

Directora de Tesis

Dr. Héctor Allende Olivares
Universidad Técnica Federico Santa María, Chile

Co-Director de Tesis

Dr.Enrique Canessa
Universidad de Valparaíso, Chile

Evaluador Externo Nacional

Dr.Julio Sotelo
Universidad Técnica Federico Santa María, Chile

Evaluador Interno

Dr. Alejandro Frery
Victoria University of Wellington, Nueva Zelanda

Evaluador Externo Internacional

Agradecimientos

En primer lugar, a mi Amanda y a nuestra niña Alicia, por su amor incondicional, su paciencia y su apoyo constante a lo largo de este camino. Sin ellas, este logro no habría sido posible. A Ginette Martinez, por su comida, compañia y presencia mientras escribía buena parte de este trabajo. A Leonardo Zambrano, por su confianza y por brindarme sus valiosas recomendaciones. Al profesor Héctor Allende, por su inspiración, sus consejos y su guía intelectual a lo largo de los años. A mi profesora guía Carolina Saavedra, por su paciencia, dedicación y apoyo riguroso durante todo el proceso de tesis. Finalmente, a mis amigos Nataly Pérez y Guilherme Cardim, por ayudarme a mantener el enfoque en momentos más exigentes de este proyecto.

A mi Amanda y a mi hija Alicia, cuyo apoyo y hermosas sonrisas fueron fundamentales para poder completar este trabajo

Resumen

Esta tesis aborda la dificultad de interpretar la relevancia de atributos en reconocimiento de patrones, entendido aquí como el aprendizaje de regularidades que permiten asociar observaciones con clases, decisiones o valores de respuesta. El problema aparece cuando los resultados predictivos se leen, de manera demasiado directa, como evidencia sobre el fenómeno que genera los datos. Para enfrentar esa dificultad, se desarrolla una propuesta metodológica y conceptual que combina el *Attribute Relevance Score* (ARS), orientado a estimar la relevancia de un atributo mediante el contraste *null-swap*, método introducido en este trabajo que estima la relevancia de un atributo mediante la comparación sistemática de versiones permutadas del mismo bajo un mismo protocolo, y un marco de fronteras epistémicas que distingue entre causalidad, informatividad poblacional y predictibilidad operacional.

La tesis sostiene que la relevancia de un atributo puede formularse con mayor precisión si se separan tres niveles de evidencia: por un lado, las relaciones causales del fenómeno, asociadas a intervenciones sobre el proceso generador de datos; por otro, la informatividad poblacional de una variable en la distribución observacional, con especial atención a la información mutua condicional, entendida como la información incremental que una variable aporta sobre la respuesta dado un conjunto de otras variables observadas o covariables; y, finalmente, la predictibilidad operacional que emerge bajo una representación, un modelo, un procedimiento de entrenamiento y un protocolo de evaluación determinados. Desde esta separación, es posible matizar la lectura de medidas de importancia, incluyendo métodos post-hoc, es decir, explicaciones calculadas después del entrenamiento para describir el comportamiento del predictor, como SHAP (*SHapley Additive exPlanations*), cuyas atribuciones pueden ser fieles al modelo entrenado y, a la vez, insuficientes para sostener interpretaciones causales o mecanísticas.

Los resultados sugieren que ARS ofrece una vía operacional estable basada en el contraste

null-swap, especialmente en escenarios con ruido y relaciones no lineales. En experimentos sintéticos y de referencia, la medida tiende a asignar valores cercanos a cero a variables no informativas, mientras conserva sensibilidad ante relaciones lineales y no lineales que aportan predictibilidad operacional bajo el modelo elegido. A su vez, los experimentos sobre dependencia del protocolo muestran que ciertas estrategias de remuestreo pueden inducir efectos persistentes sobre variables nulas; por ello, predictibilidad operacional, informatividad poblacional y causalidad mantienen alcances inferenciales distintos.

En conjunto, la tesis propone comprender la relevancia de atributos como una relación situada que depende del fenómeno, del régimen que genera los datos, ya sea observacional, experimental o intervencional, del protocolo, del tamaño muestral, de la clase de modelos y del alcance de la pregunta científica. En particular, introduce los protocolos Bayes-consistentes, entendidos como procedimientos cuya estimación se orienta hacia el riesgo de Bayes de la cantidad poblacional pertinente, como primera condición para evaluar cuándo una importancia operacional puede leerse como evidencia de informatividad poblacional. Sin embargo, esta condición debe articularse con un contraste *null-swap*, ya que, sin esa comparación con una contraparte no informativa equivalente, una ganancia estable puede reflejar resonancia estocástica, es decir, mejoras predictivas inducidas por ruido, o artefactos específicos del modelo. Así, la formulación propuesta busca ordenar las condiciones de uso del análisis causal y de las medidas clásicas de dependencia, explicitando qué tipo de afirmación queda respaldada por cada resultado.

Palabras clave: relevancia de atributos, selección de variables, explicabilidad, causalidad, predictibilidad, aprendizaje automático

Abstract

This thesis addresses the difficulty of interpreting attribute relevance in pattern recognition, understood here as learning regularities that associate observations with classes, decisions, or response values. The problem arises when predictive results are read too directly as evidence about the phenomenon that generates the data. To address this difficulty, the thesis develops a methodological and conceptual proposal that combines the *Attribute Relevance Score* (ARS), aimed at estimating the relevance of an attribute through the *null-swap* contrast with permuted versions of the attribute under repeated protocols, with a framework of epistemic frontiers that distinguishes between causality, population informativeness, and operational predictability.

The thesis argues that attribute relevance can be formulated more precisely by separating three levels of evidence: causal relations in the phenomenon, associated with interventions on the data-generating process; population informativeness in the observational distribution, with particular attention to conditional mutual information, understood as the incremental information that a variable provides about the response given a set of other observed variables or covariates; and operational predictability emerging under a specific representation, model, training procedure, and evaluation protocol. From this separation, it becomes possible to qualify the interpretation of importance measures, including post-hoc methods, understood here as explanations computed after training to describe the predictor's behavior, such as SHAP (*SHapley Additive exPlanations*), whose attributions may be faithful to the trained model while remaining insufficient to support causal or mechanistic interpretations.

The results suggest that ARS provides a stable operational way to evaluate attributes, especially in settings with noise and nonlinear relationships. In synthetic and benchmark experiments, the measure tends to assign values close to zero to non-informative variables, while preserving sensitivity to linear and nonlinear relationships that provide operational

predictability under the chosen model. In turn, the experiments on protocol dependence show that some resampling strategies may induce persistent effects on null variables; therefore, operational predictability, population informativeness, and causality retain distinct inferential scopes.

Taken together, the thesis proposes understanding attribute relevance as a situated relation that depends on the phenomenon, the data-generating regime, whether observational, experimental, or interventional, the protocol, the sample size, the model class, and the scope of the scientific question. In particular, it introduces Bayes-consistent protocols, understood as procedures whose estimates are oriented toward the Bayes risk of the relevant population quantity, as a first condition for evaluating when operational importance may be read as evidence of population informativeness. This condition, however, must be combined with a *null-swap* contrast, since, without comparison against an equivalent non-informative counterpart, a stable gain may still reflect stochastic resonance, that is, noise-induced predictive improvement, or model-specific artefacts. The proposed formulation therefore seeks to organize the conditions of use of causal analysis and classical measures of dependence, making explicit what kind of claim is supported by each result.

Keywords: attribute relevance, feature selection, explainability, causality, predictability, machine learning

Lista de Abreviaturas

Sigla	**Descripción**
ARS	*Attribute Relevance Score*; medida operacional de relevancia de atributos basada en el contraste *null-swap* aplicado a versiones permutadas del atributo.
DGP	Proceso generador de datos, entendido como el conjunto de mecanismos que produce las observaciones disponibles.
MI	Información mutua, medida de dependencia estadística entre variables.
CMI	Información mutua condicional, información incremental que una variable aporta sobre otra dado un conjunto de covariables.
XAI	Inteligencia artificial explicable, familia de métodos orientados a describir o auditar el comportamiento de modelos predictivos.
HCXAI	Inteligencia artificial explicable centrada en humanos.
SHAP	*SHapley Additive exPlanations*; método post-hoc de atribución de importancia inspirado en valores de Shapley.
LIME	*Local Interpretable Model-agnostic Explanations*; método post-hoc que aproxima localmente un predictor complejo mediante un modelo interpretable.
SCM	*Structural Causal Model*; modelo causal estructural que representa variables mediante ecuaciones y relaciones de intervención.
CV	Validación cruzada, procedimiento de evaluación que rota particiones de entrenamiento y prueba.

Sigla	Descripción
MIC	*Maximal Information Coefficient*; coeficiente máximo de información, usado para detectar dependencias posiblemente no lineales.
MAS	*Maximum Asymmetry Score*; puntaje máximo de asimetría asociado a patrones de dependencia.
MEV	*Maximum Edge Value*; valor máximo de borde, medida asociada a estructuras de dependencia en grillas.
GMIC	*Generalized Mean Information Coefficient*; coeficiente de información media generalizada.
AUC	Área bajo la curva ROC; medida de discriminación para clasificación binaria.
ROC	*Receiver Operating Characteristic*; curva que relaciona sensibilidad y tasa de falsos positivos a distintos umbrales.
RMSE	Raíz del error cuadrático medio, métrica de error para problemas de regresión.
XOR	*Exclusive OR*; patrón lógico donde la señal depende de la combinación de variables.

Notación

Símbolo	Descripción
$\boldsymbol{X} = (X_1, \ldots, X_p)$	Vector aleatorio de covariables observadas.
Y	Variable objetivo, etiqueta o respuesta: discreta en clasificación, continua en regresión. El símbolo denota el mismo objeto en los tres registros epistémicos (causal, informativo, predictivo), cuyo estatuto interpretativo se precisa en la Sección 3.1.
X_j	Variable aleatoria escalar correspondiente a la j-ésima componente de $\boldsymbol{X}$: el atributo o covariable evaluado.
S	Conjunto de atributos considerado en una evaluación, representado por sus posiciones en $\boldsymbol{X}$; formalmente, $S \subseteq \{1, \ldots, p\}$.
$S \setminus \{j\}$	Conjunto que identifica los atributos de S salvo el atributo evaluado X_j.
$S_{\setminus j}$	Abreviatura de $S \setminus \{j\}$ cuando conviene nombrar explícitamente el contexto.
$\boldsymbol{X}_S$	Subvector de $\boldsymbol{X}$ con las componentes indexadas por S; por ejemplo, $\boldsymbol{X}_{\{j\}} = X_j$.
$\boldsymbol{X}_{S\setminus\{j\}}$	Subvector de $\boldsymbol{X}$ formado por los atributos de S salvo X_j; equivalentemente, $\boldsymbol{X}_{S_{\setminus j}}$.
$\boldsymbol{X}_{\setminus j}$	Abreviatura contextual de $\boldsymbol{X}_{S\setminus\{j\}}$ cuando S ya fue declarado; si $S = \{1, \ldots, p\}$, corresponde a todas las covariables salvo X_j.
$\mathcal{X}, \mathcal{Y}$	Espacio de atributos y espacio de respuestas.
$D_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$	Conjunto de entrenamiento con n observaciones.
$\boldsymbol{x}_i$	Vector de atributos observado para la i-ésima unidad.

Símbolo	Descripción
$\boldsymbol{x}_{\cdot j}$	Columna muestral asociada a X_j, es decir, las n realizaciones observadas del atributo j en D_n.
$\mathcal{F}$	Clase de modelos o espacio de hipótesis.
$L(y, \hat{y})$	Función de pérdida entre respuesta real y predicha.
$L_E(M; D_n)$	Pérdida de evaluación del modelo M bajo la reimplementación E del protocolo.
$R(f)$	Riesgo esperado de la función f bajo la distribución poblacional.
$R_{\text{emp}}(f)$	Riesgo empírico de f sobre la muestra.
$R^*(S)$	Riesgo de Bayes asociado al subvector de covariables $\boldsymbol{X}_S$.
$\mathcal{C}_Y$	Espacio causal respecto de Y: variables que pueden modificar la distribución de Y bajo intervención.
$\mathcal{I}$	Espacio de informatividad poblacional: variables que agregan información sobre Y en la distribución observacional.
$\mathcal{P}$	Espacio predictivo operacional: atributos incorporados por un predictor bajo un protocolo de aprendizaje y evaluación.
$I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}})$	Información mutua condicional de X_j respecto de Y dado el subvector de contexto $\boldsymbol{X}_{S\setminus\{j\}}$; con Y discreta, coincide con la brecha de riesgo de Bayes (Proposición 3.2).
$\Delta^*_{S,j}$	Brecha poblacional de riesgo al retirar X_j del conjunto evaluado S.
Δ^*_j	Abreviatura de $\Delta^*_{\{1,\dots,p\},j}$, brecha al retirar X_j del conjunto completo de atributos.
$\widetilde{\Delta}^{\text{ns}}_{n,S,j}$	Contraste empírico *null-swap* que reemplaza X_j por una copia nula dentro del conjunto evaluado S.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto y descripción del problema

El reconocimiento de patrones estudia procedimientos capaces de identificar regularidades en datos y utilizarlas para clasificar, predecir, segmentar, detectar anomalías o apoyar decisiones. En una formulación supervisada, cada observación se describe mediante un conjunto de atributos y una variable de respuesta, de modo que el modelo aprende, a partir de ejemplos disponibles, una relación que luego se evalúa por su capacidad para generalizar a observaciones nuevas. Aunque esta descripción parece estrictamente técnica, contiene una tensión central para la investigación científica: el mismo procedimiento que permite anticipar una respuesta suele usarse también para decir algo sobre el fenómeno que produjo los datos. Esta distinción entre predecir y explicar ha sido discutida con amplitud en estadística y aprendizaje automático. Ejemplo de esto es el trabajo de Breiman (2001b), donde caracterizó dos culturas de modelamiento: una orientada a describir el proceso generador mediante modelos estocásticos y otra centrada en construir algoritmos predictivos de alto desempeño; Shmueli (2010), a su vez, mostró que explicar y predecir responden a objetivos distintos, con criterios de éxito, supuestos y riesgos inferenciales diferentes. Más recientemente, Miller et al. (2021) han propuesto una lectura menos dicotómica, en la que ambas culturas pueden dialogar siempre que se declare con precisión qué tipo de afirmación se está sosteniendo. Esta tesis se ubica precisamente en ese punto de contacto, pues estudia cuándo una ganancia predictiva atribuida a una variable puede leerse como algo más que un resultado operacional

del modelo.

Al reconstruir la lógica de estas fuentes, aparece una historia más exigente que una simple oposición entre modelos interpretables y modelos precisos. Breiman desplaza la atención desde el modelo supuesto hacia la caja negra que conecta atributos y respuesta; Shmueli muestra que el propósito de modelar modifica todas las decisiones del análisis; y la literatura posterior sobre el efecto Rashomon advierte que, aun cuando varios modelos predicen de manera similar, pueden apoyarse en regularidades distintas (D'Amour, 2021; Fisher et al., 2019; Semenova et al., 2022). Para la relevancia de atributos, esta observación es decisiva: una variable puede parecer central en un predictor y secundaria en otro sin que el desempeño permita resolver por sí solo cuál merece mayor alcance inferencial.

En la práctica, la pregunta por la relevancia de atributos aparece cuando se desea saber qué variables contribuyen a resolver una tarea, cuáles podrían descartarse con una pérdida limitada de desempeño y cuáles merecen una interpretación más detallada, si bien una tabla de importancia de variables puede parecer suficiente en un reporte técnico, en investigación científica la palabra "relevancia" exige una lectura más cuidadosa, ya que un atributo puede mejorar una predicción, contener información estadística sobre la respuesta, actuar como proxy de otra variable, estar confundida con el fenómeno de interés (es decir, su efecto se encuentra entremezclado con el de otros factores, impidiendo aislar su contribución real) o participar en un mecanismo causal. Todas estas posibilidades se superponen en los datos observados, aunque remiten a niveles de análisis distintos. Esta ambigüedad se vuelve especialmente delicada cuando los resultados de un modelo se integran en investigación científica, salud, finanzas, física, planificación urbana, seguridad informática o sistemas de apoyo a la decisión. En esos contextos, un ranking de variables rara vez se interpreta como una descripción neutral del predictor. Con frecuencia se lee como evidencia preliminar sobre el fenómeno: factores de riesgo, señales diagnósticas, mecanismos candidatos, sesgos de adquisición o dimensiones relevantes para una política pública.

Por otro lado, la literatura sobre explicabilidad ha desarrollado herramientas valiosas para inspeccionar modelos complejos, como LIME y SHAP (Lundberg and Lee, 2017; Ribeiro et al., 2016); al mismo tiempo, ha señalado que la interpretabilidad reúne objetivos heterogéneos y que una explicación fiel al modelo puede tener un alcance limitado como explicación del fenómeno (Lipton, 2018; Sokol and Flach, 2020). Esta limitación no es solo filosófica: la evidencia empírica muestra que modelos con buen desempeño pueden explotar atajos, co-

rrelaciones espurias o variables de contexto. En imágenes médicas y visión computacional, por ejemplo, se ha observado que algunos modelos aprenden marcas de adquisición, características del hospital, artefactos de escaneo o señales periféricas que mejoran la predicción sin corresponder al mecanismo sustantivo que el investigador esperaba estudiar (Badgeley et al., 2019; DeGrave et al., 2021; Lapuschkin et al., 2019; Roberts et al., 2021). La discusión reciente sobre el efecto Clever Hans en inteligencia artificial (llamado así en honor al caballo que parecía resolver operaciones matemáticas, pero que en realidad detectaba sutiles señales corporales y gestos faciales de sus espectadores) apunta en la misma dirección, al mostrar que un sistema puede producir respuestas correctas por razones poco adecuadas para la interpretación científica o para una transferencia robusta (Pathak et al., 2026); por ello, una medida de importancia debe situarse dentro de un protocolo, una población observacional, una métrica y una pregunta inferencial.

El problema de esta tesis puede formularse entonces como una pregunta sobre el estatuto de la evidencia: cuando un modelo mejora al incorporar una variable, ¿qué se ha aprendido? La respuesta mínima es operacional, pues indica que, bajo cierta muestra, cierto algoritmo, cierta pérdida y cierta forma de evaluación, el modelo logró aprovechar esa variable. Una respuesta más ambiciosa diría que la variable porta información poblacional sobre la respuesta, mientras que una lectura causal sostendría que intervenir sobre ella podría modificar el fenómeno. En este trabajo enfatizamos la necesidad de distinguir con claridad estas tres interpretaciones. Aunque la predicción aporta evidencia valiosa, su alcance y validez dependen de las condiciones en las que se generó y de los supuestos que permiten vincularla con una cantidad poblacional o con un mecanismo causal (Hammerton and Munafò, 2021; Pearl, 2009; Vowels, 2024).

1.2. Motivación y brecha de investigación

La literatura sobre selección de variables, relevancia de atributos y explicabilidad ha producido un conjunto amplio de métodos: los enfoques de filtro evalúan variables mediante medidas estadísticas relativamente independientes del modelo final; los métodos *wrapper* comparan subconjuntos de atributos a través del desempeño de un predictor (Kohavi and John, 1997); los métodos incorporados integran la selección dentro del entrenamiento (Guyon and Elisseeff, 2003); y los métodos post-hoc describen el comportamiento de modelos ya

ajustados (Lundberg and Lee, 2017; Molnar, 2022; Ribeiro et al., 2016). Aunque esta diversidad responde a necesidades legítimas, también genera una dificultad conceptual: se habla de "importancia" o "relevancia" para referirse a cantidades que pueden tener significados distintos.

Una distinción particularmente importante aparece entre el minimal optimal problem (problema de selección mínima) y el problema *all-relevant* (selección de todos los relevantes). El primero busca el subconjunto más pequeño de variables que preserve el desempeño, mientras que el segundo intenta identificar todos los atributos que guardan alguna relación con la respuesta, incluso si esta relación es débil, redundante, o relevante solo en interacción con otros atributos (Kursa and Rudnicki, 2010; Mnich and Rudnicki, 2020; Nilsson et al., 2007). La diferencia importa porque una variable descartable para construir un predictor compacto puede seguir siendo científicamente interesante, y su utilidad se intensifica en dominios de alta dimensionalidad, como datos ómicos, sensores, sistemas financieros o representaciones aprendidas, donde la parsimonia del modelo y el mapeo amplio de señales responden a preguntas distintas (Degenhardt et al., 2019; Passemiers et al., 2024).

Los métodos basados en contraste nulo consisten en comparar el desempeño de las variables observadas contra una o más versiones artificiales sin relación con la variable respuesta (referencias nulas). Ejemplos comunes son la importancia por permutación, las variables aleatorias de referencia y los atributos *shadow* del algoritmo Boruta (Breiman, 2001a; Kursa and Rudnicki, 2010; Stoppiglia et al., 2003). Esta aproximación ofrece una estrategia atractiva para ordenar la discusión sobre la importancia de variables. La intuición subyacente es simple pero potente: si una variable contiene información útil, debería aportar más que una copia nula (Breiman, 2001a; Kursa and Rudnicki, 2010; Stoppiglia et al., 2003). Al introducir un punto de comparación dentro del mismo problema, estos métodos reducen parte de la arbitrariedad inherente a los rankings absolutos. No obstante, una copia nula no transforma automáticamente una diferencia de desempeño en una explicación poblacional o causal. Por ello, la comparación debe interpretarse siempre a la luz del protocolo experimental que la genera, de la variabilidad muestral y de la forma específica en que se construye la copia nula.

La brecha que aborda esta tesis surge en esa zona intermedia, donde conviven una necesidad práctica de medir relevancia de atributos de manera robusta, especialmente ante ruido, redundancia, relaciones no lineales e interacciones, y una necesidad conceptual de precisar qué significa que un atributo "importe". Las herramientas disponibles suelen resolver bien una

parte del problema, pero dejan abierta la transición entre importancia operacional, informatividad poblacional y causalidad: en algunos casos entregan un ranking útil para el modelo; en otros, una medida de dependencia estadística; y, en otros, una explicación local de una predicción. La dificultad aparece cuando esos resultados se presentan como si pertenecieran al mismo registro.

Esta tesis propone abordar la brecha mediante dos movimientos complementarios. El primero es metodológico y consiste en evaluar la contribución de cada atributo a través del contraste *null-swap* propuesto en este trabajo: la variable observada se compara, dentro de un contexto de atributos declarado, con una versión permutada que conserva su forma marginal y rompe su vínculo sistemático con la respuesta. Con ello se busca separar la utilidad atribuible a la asociación con la variable objetivo de aquella que podría surgir por escala, tipo de dato, cardinalidad, dimensionalidad o disponibilidad de un atributo adicional.

El segundo movimiento es conceptual y consiste en distinguir tres espacios de afirmación: causalidad, informatividad poblacional y predictibilidad operacional. Estos dos movimientos no son independientes: el contraste *null-swap* provee la evidencia empírica, mientras que la separación conceptual determina qué puede afirmarse legítimamente a partir de esa evidencia, y bajo qué condiciones una diferencia de desempeño gana o pierde alcance inferencial.

La motivación de fondo es construir un vocabulario que permita reportar relevancia sin sobredimensionar la evidencia. Una variable puede superar a su copia nula en un contexto evaluado y, con ello, mostrar relevancia operacional estable bajo un protocolo explícito; si además el protocolo aproxima una cantidad poblacional pertinente, esa evidencia puede ganar una lectura epistémica, en el sentido de que el resultado sugiere que la variable contiene información sobre la distribución poblacional. La lectura causal, en cambio, requiere supuestos adicionales sobre intervención, identificación o estructura del proceso generador. Esta separación permite conservar la utilidad de las medidas predictivas y, al mismo tiempo, evitar que se transformen por inercia en afirmaciones causales.

1.3. Pregunta de investigación

La pregunta central se concentra en el tipo de evidencia que puede sostener una medida de relevancia cuando se obtiene bajo un protocolo predictivo explícito. En consecuencia, la pregunta que articula esta tesis puede formularse del siguiente modo:

> ¿Bajo qué condiciones metodológicas e inferenciales una diferencia de desempeño atribuida a una variable en un contexto de evaluación, estimada mediante un contraste *null-swap*, puede interpretarse como evidencia de informatividad poblacional, y qué alcance adquiere esa evidencia dentro de la relación entre predicción, información y causalidad?

Formulada de este modo, la pregunta establece una separación analítica entre tres registros. El primero es operacional: qué logra explotar un modelo bajo una configuración concreta. El segundo es poblacional: qué dependencia parece existir en la población que se intenta aproximar. El tercero es causal: qué cambios podrían esperarse bajo intervención. La tesis examina qué puede inferirse desde una diferencia de desempeño, qué permanece en el nivel del predictor y qué requiere supuestos adicionales para sostener afirmaciones sobre información poblacional o sobre mecanismos causales.

La pregunta también delimita el tipo de respuesta buscada: esta tesis no propone una regla universal para transformar importancias predictivas en explicaciones causales, ni una medida capaz de capturar toda forma posible de dependencia. El objetivo es más acotado, pues consiste en formular condiciones bajo las cuales un efecto operacional puede reportarse con un alcance inferencial explícito, especialmente cuando se compara contra una copia nula y se evalúa bajo protocolos diseñados para aproximar cantidades poblacionales pertinentes.

1.4. Hipótesis

La hipótesis de trabajo de esta tesis puede enunciarse del siguiente modo:

> *Una medida de relevancia de atributos construida sobre el contraste* null-swap *(la comparación de cada atributo con una copia permutada que conserva su distribución marginal, evaluada en un contexto declarado, bajo el mismo modelo, la misma métrica y un protocolo repetido con significancia estadística y margen práctico) distinguirá variables informativas de variables nulas con mayor especificidad que las medidas absolutas de dependencia o importancia, asignando relevancia cercana a cero a atributos no informativos incluso en presencia de ruido y relaciones no lineales; y esa relevancia operacional podrá interpretarse como evidencia*

de informatividad poblacional únicamente cuando el protocolo de evaluación sea Bayes-consistente respecto de la cantidad poblacional pertinente.

La hipótesis contiene dos componentes contrastables. El primero es operacional: si el contraste *null-swap* cumple su función, las variables nulas por construcción deberían recibir relevancia cercana a cero y las variables informativas conservar valores positivos, incluso bajo ruido y relaciones no lineales donde las correlaciones clásicas pierden sensibilidad. El segundo es inferencial: si la consistencia del protocolo es efectivamente una condición necesaria para la lectura poblacional, un protocolo que reutiliza una muestra fija debería poder estabilizar contrastes positivos sobre variables nulas conocidas, mientras que un protocolo orientado a nuevas realizaciones del proceso generador debería concentrarlas alrededor de cero. Ambos componentes se evalúan, respectivamente, en los dos bloques experimentales del Capítulo 4.

La hipótesis admite criterios de debilitamiento. La lectura como informatividad poblacional gana respaldo cuando el efecto estimado por el contraste *null-swap* persiste ante reimplementaciones razonables, se mantiene bajo controles de fuga de información, muestra estabilidad frente a variaciones plausibles de modelo o partición y se obtiene bajo un protocolo orientado a una cantidad poblacional interpretable. Cuando esas condiciones se atenúan, el resultado conserva valor descriptivo como efecto operacional del protocolo, aunque su lectura epistémica debe formularse con mayor cautela.

1.5. Alcance y delimitaciones

Este trabajo se ubica en el campo del reconocimiento de patrones y del aprendizaje automático supervisado, con énfasis en la interpretación de relevancia de atributos. Su interés principal es estudiar cómo se construye y se interpreta la evidencia de relevancia cuando los atributos se evalúan mediante desempeño predictivo; por ello, el análisis se organiza alrededor de procesos generadores controlados, datos sintéticos, conjuntos de referencia y escenarios experimentales que permiten aislar ruido, no linealidad, confusión, dependencia del protocolo, redundancia y sinergia entre variables.

Esta delimitación tiene una consecuencia importante: los resultados deben leerse como evidencia conceptual y procedimental, no como validación exhaustiva en todos los dominios posibles. Los escenarios sintéticos permiten saber, con mayor claridad que en datos puramente

observacionales, qué variables son informativas, cuáles son nulas, cuáles actúan por confusión y cuáles adquieren relevancia solo en combinación. Esa claridad facilita examinar los límites de las medidas de importancia, aunque deja abierta la necesidad de estudiar dominios reales con supuestos, costos de medición y estructuras de dependencia propias.

Este trabajo tampoco busca reemplazar los métodos existentes de selección de variables, explicabilidad o inferencia causal. Su propósito es ofrecer una gramática de interpretación que permita ubicarlos con mayor precisión. Una medida de dependencia puede ser adecuada para explorar asociaciones; una medida de importancia operacional puede orientar selección o auditoría de modelos; un método post-hoc puede describir la conducta del predictor; y un análisis causal puede sostener afirmaciones sobre intervención cuando se cumplen supuestos de identificación. El aporte de la tesis consiste en articular esas capas sin reducirlas unas a otras.

1.6. Objetivos

Objetivo general

Desarrollar un marco metodológico y conceptual que permita estimar la relevancia operacional de atributos mediante contrastes *null-swap* y precisar las condiciones bajo las cuales esa relevancia puede interpretarse como evidencia de informatividad poblacional, distinguiéndola de afirmaciones causales.

Objetivos específicos

1. Sistematizar las nociones de relevancia de atributos presentes en la literatura, distinguiendo problema minimal optimal, problema *all-relevant*, dependencia estadística, atribución post-hoc y lectura causal.

2. Diseñar una medida operacional de relevancia basada en el contraste *null-swap* con copias permutadas del atributo, que declare el contexto de evaluación e incorpore evaluación repetida, significancia estadística y un margen mínimo de efecto práctico.

3. Formalizar la distinción entre causalidad, informatividad poblacional y predictibilidad operacional mediante los espacios causal $\mathcal{C}_Y$, informativo $\mathcal{I}$ y predictivo $\mathcal{P}$ definidos

en la Sección 3.1, y definir la relevancia epistémica como condición de transición entre importancia operacional e informatividad poblacional.

4. Caracterizar los protocolos Bayes-consistentes como condición para vincular efectos operacionales con cantidades poblacionales, precisando su articulación con el contraste *null-swap* frente a ruido y artefactos de remuestreo.

5. Construir una estrategia de validación basada en procesos generadores controlados, con variables nulas, informativas, confundidas, causales y sinergéticas conocidas por diseño, y evaluar en ella el comportamiento de la medida propuesta junto a medidas de dependencia, importancias de modelo y atribuciones post-hoc.

6. Proponer criterios de reporte que permitan comunicar la relevancia de atributos con un alcance inferencial explícito, declarando protocolo, contexto, contraste, estabilidad y tipo de afirmación.

1.7. Contribuciones de la tesis

Para responder a la pregunta planteada, la tesis articula contribuciones de tres tipos, cuyo desarrollo detallado se reserva para los capítulos de propuesta y validación. En el plano conceptual, se distinguen tres espacios de afirmación: causalidad, informatividad poblacional y predictibilidad operacional; y se formula la noción de relevancia epistémica como categoría de transición entre el desempeño operacional y la informatividad poblacional, junto con la condición de protocolos orientados al riesgo de Bayes pertinente. En el plano metodológico, se introduce el *Attribute Relevance Score* (ARS), una medida operacional de relevancia basada en comparar cada atributo con una copia nula que conserva su forma marginal y rompe su asociación con la respuesta.

En el plano experimental, se construye una estrategia de validación apoyada en procesos generadores controlados, donde el papel de cada variable se conoce por diseño, que permite examinar cuándo una medida distingue señal de ruido y cuándo una variable predictivamente prominente puede carecer de lectura causal directa. De estas piezas se desprende, finalmente, una gramática de reporte que invita a comunicar la relevancia junto con su protocolo, su

contexto de evaluación, su contraste, su estabilidad y su alcance inferencial. Las conclusiones del Capítulo 5 retoman y precisan estas contribuciones a la luz de la evidencia reunida.

El punto de partida se apoya en distinciones ya conocidas. Predicción y explicación responden a propósitos diferentes; las atribuciones post-hoc describen al modelo; y el efecto Rashomon muestra que varios predictores con desempeño similar pueden usar variables distintas. El aporte de la tesis consiste en reunir estas ideas en tres espacios y vincularlas con ARS mediante el contraste *null-swap*. Así, la contribución se concentra en ordenar los conceptos, precisar las condiciones para interpretar los resultados y proponer una guía de reporte.

1.8. Organización de la tesis

El Capítulo 2 presenta los antecedentes necesarios para situar el problema dentro del aprendizaje supervisado, la selección de variables, las medidas de dependencia e información, la explicabilidad, los datos sintéticos, el ruido, las interacciones, la causalidad, la reproducibilidad y los contrastes nulos. Sobre esa base, el Capítulo 3 desarrolla la propuesta metodológica y conceptual integrada. El Capítulo 4 presenta la validación experimental, con énfasis en diseño, resultados, interpretación y límites. Finalmente, el Capítulo 5 recoge las conclusiones, contribuciones y líneas de trabajo futuro. Los apéndices reúnen materiales de trazabilidad y resultados complementarios.

1.9. Orden de la tesis y de las publicaciones

Los artículos que integran el corpus siguieron un orden cronológico distinto del orden expositivo adoptado en esta tesis. El trabajo sobre ARS fue publicado en 2024 y el marco de fronteras epistémicas en 2026. El manuscrito doctoral presenta primero el marco conceptual y formal, y sitúa después la formulación operacional de ARS y la evidencia experimental. Este orden responde a una decisión pedagógica: delimitar qué significan causalidad, informatividad poblacional y predictibilidad operacional antes de interpretar el alcance de una medida basada en desempeño. Por ello, el orden de los capítulos expresa la dependencia argumental entre conceptos, método y evidencia, mientras que el orden de las publicaciones registra la

trayectoria cronológica de la investigación.

La Tabla 1.1 explicita la procedencia de los desarrollos principales. La correspondencia se presenta por componentes y permite describir los ajustes de notación, idioma y estructura argumental realizados en el manuscrito doctoral.

Tabla 1.1: Procedencia de los desarrollos principales y función que cumplen dentro de la integración doctoral.

Procedencia	**Contenido retomado**	**Tratamiento en la tesis**
Neirz et al. (2024)	Formulación de ARS y experimentos del Bloque I sobre relaciones bivariadas, modelos base y ruido estructurado.	Reexpresión dentro de la notación común del contraste *null-swap* y lectura explícita como evidencia operacional.
Neirz et al. (2026)	Separación de los espacios $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$; resultados formales sobre consistencia e informatividad; experimentos del Bloque II.	Desarrollo en español, orden pedagógico y articulación con ARS, sus supuestos y su alcance.
Integración doctoral	Revisión conjunta del estado del arte, notación transversal, ejercicio Pima, discusión integrada, correspondencia entre objetivos y evidencia, y síntesis de criterios de reporte.	Construcción de un argumento unitario que conecta los dos trabajos, delimita el alcance de sus resultados y hace explícitas sus relaciones metodológicas.

Capítulo 2

Marco teórico y estado del arte

Este capítulo construye, el lenguaje conceptual que este trabajo necesita. Comienza por los fundamentos del aprendizaje supervisado, porque la pregunta por la relevancia de atributos solo adquiere sentido una vez que se ha precisado qué significa aprender una relación entre variables y respuesta, cómo se evalúa esa relación y bajo qué supuestos se interpreta. Sobre esa base se introducen, de lo general a lo específico, la selección de variables, las medidas de dependencia e información, los métodos basados en contraste nulo, los fenómenos de ruido, confusión e interacción, los datos sintéticos, la explicabilidad y, finalmente, la relación entre información, predictibilidad y causalidad. El recorrido busca establecer un lenguaje común para la discusión posterior y, al mismo tiempo, hacer visibles las tensiones metodológicas que la propuesta intentará ordenar.

2.1. Fundamentos del aprendizaje supervisado

Siguiendo la formulación clásica de Mitchell (1997), puede decirse que un programa aprende cuando, a partir de una experiencia E, mejora su desempeño en una clase de tareas T según una medida de desempeño P. Esta forma de definir el aprendizaje permite precisar que los datos, por sí solos, adquieren sentido metodológico al vincularse con una tarea y con un criterio de evaluación. En aplicaciones contemporáneas, esa experiencia suele tomar la forma de conjuntos de datos, pero su interpretación depende del protocolo que conecta observaciones, objetivo predictivo y medida de desempeño (Murphy, 2012). Dentro de este campo, el aprendizaje supervisado se ocupa del caso en que cada ejemplo viene acompañado

de una respuesta conocida, y el objetivo consiste en aprender una función capaz de predecir esa respuesta para observaciones nuevas (Bishop, 2006; Hastie et al., 2009). La situación se formaliza mediante un conjunto de entrenamiento $D_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, donde cada $\boldsymbol{x}_i \in \mathcal{X}$ describe una observación a través de un vector de atributos y cada $y_i \in \mathcal{Y}$ corresponde a la respuesta asociada. Se asume que los pares $(\boldsymbol{x}_i, y_i)$ se obtienen de manera independiente desde una distribución conjunta $P(\boldsymbol{X}, Y)$ desconocida, supuesto que conecta la muestra disponible con la población que se desea describir (Vapnik, 1998).

Cuando la respuesta Y toma valores en un conjunto finito de categorías, el problema se denomina clasificación; cuando Y es una cantidad continua, se habla de regresión (Duda et al., 2001; James et al., 2013). En ambos casos, el aprendizaje requiere comprometerse con una clase de modelos o espacio de hipótesis $\mathcal{F}$, es decir, el conjunto de funciones candidatas entre las que el procedimiento puede elegir. La elección de $\mathcal{F}$ no es neutral: codifica supuestos sobre la forma de la relación entre atributos y respuesta, y determina qué regularidades pueden representarse y cuáles quedan fuera de alcance (Mitchell, 1997).

2.1.1. Pérdida, riesgo y minimización empírica

Para comparar funciones candidatas se introduce una función de pérdida $L(y, \hat{y})$, que cuantifica el costo de producir una salida predictiva $\hat{y}$ cuando la respuesta observada es y. En esta notación, Y denota la variable de respuesta como cantidad aleatoria; y o y_i corresponden a realizaciones observadas de esa variable; y $\hat{y}$ representa la predicción generada por un modelo para una observación dada, de modo que, si la función predictiva es f, entonces $\hat{y} = f(\boldsymbol{x})$. Ejemplos habituales son la pérdida cuadrática en regresión, la pérdida 0–1 en clasificación y la pérdida logarítmica cuando interesa la calidad de las probabilidades estimadas (Hastie et al., 2009). A partir de la pérdida se define el riesgo esperado de una función f, entendido como el costo promedio bajo la distribución poblacional,

$$R(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}\Big[L\Big(Y, f(\boldsymbol{X})\Big)\Big]. \tag{2.1}$$

La función que minimiza este riesgo dentro de todas las funciones posibles recibe el nombre de predictor de Bayes, y su riesgo asociado, el riesgo de Bayes, constituye el límite inferior alcanzable para la tarea y la pérdida consideradas (Duda et al., 2001; Vapnik, 1998). Como la distribución $P(\boldsymbol{X}, Y)$ es desconocida, el riesgo esperado no puede calcularse de manera

directa; en su lugar se estima el riesgo empírico sobre la muestra,

$$R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i, f(\boldsymbol{x}_i)\Big), \tag{2.2}$$

y muchos algoritmos operan, explícita o implícitamente, buscando una función que lo reduzca, principio conocido como minimización del riesgo empírico (Vapnik, 1998). Esta distinción entre riesgo esperado y riesgo empírico es central para la tesis, pues separa lo que se desea estimar de lo que efectivamente se mide en una muestra finita.

2.1.2. Generalización, sobreajuste y el compromiso sesgo–varianza

El propósito del aprendizaje no es reproducir las respuestas del conjunto de entrenamiento, sino generalizar a observaciones no vistas. Un modelo suficientemente flexible puede ajustar casi perfectamente la muestra disponible y, al mismo tiempo, comportarse pobremente frente a datos nuevos, fenómeno conocido como sobreajuste (Bishop, 2006; Domingos, 2012). La diferencia entre el desempeño en entrenamiento y el desempeño fuera de muestra (el desempeño en un conjunto no visto durante el entrenamiento ej: conjunto de testeo) mide, precisamente, la brecha de generalización. El análisis clásico de esta tensión se expresa mediante el compromiso entre sesgo y varianza: modelos demasiado simples incurren en sesgo elevado al no representar la estructura del fenómeno, mientras que modelos demasiado complejos exhiben varianza alta al volverse sensibles a fluctuaciones particulares de la muestra (Geman et al., 1992; Hastie et al., 2009). La capacidad de generalización depende, así, del equilibrio entre la riqueza de la clase de modelos, la cantidad de datos y la intensidad de la señal disponible.

Para contener el sobreajuste se emplean técnicas de regularización, que restringen la complejidad efectiva del modelo mediante penalizaciones sobre sus parámetros, como en la regresión *ridge* o el operador *lasso*, este último capaz de inducir esparsidad y, con ello, una forma de selección de variables integrada al entrenamiento (Bishop, 2006; Tibshirani, 1996). Conviene subrayar que la complejidad relevante no siempre se expresa como una penalización explícita: la arquitectura del algoritmo, el criterio de detención o el procedimiento de optimización pueden imponer restricciones efectivas que actúan como regularización implícita (Goodfellow et al., 2016). Esta observación será importante más adelante, porque una ganancia de desempeño puede provenir tanto de la señal del fenómeno como de las preferencias inductivas del procedimiento.

2.1.3. Estimación del desempeño fuera de muestra

Dado que el riesgo empírico sobre los datos de entrenamiento tiende a subestimar sistemáticamente el riesgo esperado, una evaluación honesta del desempeño de un modelo requiere estimar su comportamiento fuera de muestra. Se entiende por desempeño fuera de muestra la capacidad predictiva del modelo sobre datos nuevos e independientes de aquellos utilizados durante su ajuste (entrenamiento). La estrategia más simple para obtener esta estimación consiste en dividir la muestra en un conjunto de entrenamiento y uno de prueba. Sin embargo, una única partición puede resultar inestable, especialmente cuando los datos son escasos. La validación cruzada mitiga esta inestabilidad al rotar las particiones de entrenamiento y evaluación, promediando el desempeño sobre múltiples divisiones (Hastie et al., 2009; Kohavi, 1995). El *bootstrap*, por su parte, estima la variabilidad de una cantidad remuestreando con reemplazo a partir de la muestra observada (Efron and Tibshirani, 1997). Cada uno de estos procedimientos estima una cantidad ligeramente distinta y captura fuentes de variación diferentes, por lo que la evidencia predictiva siempre queda mediada por el protocolo de evaluación elegido. Esta dependencia, lejos de ser un detalle técnico, reaparecerá como un eje central en el análisis de relevancia de atributos.

2.1.4. Familias de modelos predictivos

La práctica del aprendizaje supervisado dispone de un repertorio amplio de clases de modelos, cada una con supuestos y sesgos inductivos propios. Los modelos lineales y la regresión logística ofrecen interpretabilidad y estabilidad cuando la relación es aproximadamente lineal en los atributos (Hastie et al., 2009). Los vecinos más cercanos predicen a partir de la proximidad en el espacio de atributos, sin asumir una forma funcional explícita (Duda et al., 2001). Los árboles de decisión particionan recursivamente ese espacio y son la base de métodos de ensamble que combinan muchos predictores para reducir varianza o sesgo, como el *bagging* (Breiman, 1996), Random Forest (Bosques Aleatorios) (Breiman, 2001a) y los *Extremely Randomized Trees* (Geurts et al., 2006), o que ajustan modelos de manera secuencial para corregir errores residuales, como el *gradient boosting* y sus implementaciones escalables (Chen and Guestrin, 2016). Las redes neuronales, finalmente, componen transformaciones no lineales capaces de aprender representaciones jerárquicas y han impulsado buena parte de los avances recientes del área (Goodfellow et al., 2016). La disponibilidad de bibliotecas de uso

extendido ha normalizado esta cadena metodológica y ha favorecido prácticas reproducibles y comparables (Pedregosa et al., 2011).

Ningún algoritmo domina a los demás en todos los problemas posibles: los teoremas de tipo *no free lunch* muestran que el desempeño superior de un método sobre otro depende de la correspondencia entre sus supuestos y la estructura del problema (Wolpert, 1996). Esta advertencia es pertinente para la tesis, porque implica que la relevancia atribuida a una variable puede variar con la clase de modelos elegida, y que ninguna medida de importancia puede declararse universalmente correcta con independencia del procedimiento que la produce. Con estos fundamentos establecidos, las secciones siguientes examinan qué significa que un atributo sea relevante y por qué esa noción admite lecturas que conviene distinguir.

2.2. Atributos, evidencia predictiva y aprender sobre el fenómeno

Establecidos los fundamentos del aprendizaje supervisado, conviene detenerse en el estatuto de los atributos y en lo que una predicción permite afirmar. Las componentes del vector $\boldsymbol{x}_i$ reciben distintos nombres según la tradición disciplinar: atributos, variables, características, covariables o predictores; en esta tesis se usa principalmente el término atributo para enfatizar su papel dentro de un procedimiento de modelamiento y evaluación. El reconocimiento de patrones, como campo, estudia cómo identificar regularidades en datos para apoyar tareas de clasificación, regresión, detección, segmentación, agrupamiento o priorización (Bishop, 2006; Duda et al., 2001), y la relevancia de atributos aparece cuando se pregunta cuáles de esas regularidades dependen de qué variables.

Un punto fácil de pasar por alto es que el modelo opera sobre una representación construida del fenómeno, no sobre el fenómeno en sí. Antes de interpretar una variable como relevante conviene precisar qué fue medido, cómo fue codificado, qué regularidades puede explotar el algoritmo y bajo qué protocolo se estimó su desempeño. La distinción entre desempeño en entrenamiento y desempeño fuera de muestra, ya introducida, opera aquí como primer control frente a interpretaciones apresuradas: la evidencia predictiva siempre queda mediada por un protocolo, de modo que una variable que parece relevante bajo una forma de evaluación puede perder ese estatuto bajo otra.

La mediación del protocolo adquiere relevancia cuando el modelo se usa para aprender sobre el fenómeno. La tradición algorítmica destacada por Breiman (2001b) ha mostrado que modelos flexibles pueden capturar regularidades complejas y lograr alto desempeño en problemas difíciles, mientras que la tradición explicativa pregunta por la estructura que genera los datos. La contribución de Shmueli (2010) es decisiva porque no reduce esta diferencia a una preferencia técnica entre modelos simples y modelos complejos: muestra que explicar y predecir ordenan de manera distinta el diseño completo del análisis. En un ejercicio explicativo, la selección de variables, la forma funcional, el control de confusión y la interpretación de parámetros suelen subordinarse a una teoría del fenómeno; en un ejercicio predictivo, las decisiones se evalúan por su capacidad de anticipar observaciones no vistas, por la pérdida fuera de muestra, la calibración y la estabilidad ante nuevos casos.

Esta diferencia altera el estatuto de una variable. Un atributo puede ser legítimo como señal predictiva aunque opere como proxy, huella de medición, variable de contexto o consecuencia de una trayectoria institucional; esa misma utilidad, sin embargo, no basta para presentarlo como parte del mecanismo que genera el fenómeno. El error conceptual no consiste en usar predicción como evidencia, sino en omitir el criterio bajo el cual esa evidencia fue producida. En ese punto, Miller et al. (2021) permiten evitar una lectura dicotómica: ambas orientaciones pueden complementarse si se explicita qué tipo de afirmación queda respaldada. Esta tesis asume esa complementariedad en un sentido preciso: un contraste *null-swap* entrega primero evidencia operacional sobre pérdida predictiva frente a una copia nula; solo bajo condiciones adicionales puede discutirse si esa diferencia aproxima informatividad poblacional o sugiere hipótesis causales.

2.3. Atributos, representación y riesgo

Los atributos funcionan como la interfaz entre el fenómeno observado y el modelo computacional, una idea sencilla solo en apariencia. Un atributo puede ser una medición directa, un índice agregado, una codificación categórica, un resultado de preprocesamiento, una variable derivada, una señal de procedimiento o un proxy de otra variable no observada; y, como el modelo aprende desde esa interfaz, puede incorporar sesgos de adquisición, fugas de información, redundancias, diferencias de escala o trazas del proceso de medición cuando estos elementos ayudan a mejorar el desempeño.

El riesgo esperado de un predictor se define como el costo promedio bajo la distribución poblacional. Como esa distribución suele ser desconocida, se estima mediante muestras finitas y protocolos empíricos. La validación cruzada, el bootstrap y los conjuntos de prueba intentan aproximar el desempeño fuera de muestra, pero lo hacen con supuestos diferentes sobre independencia, estabilidad y variación muestral. En experimentos comparativos de aprendizaje automático, la varianza asociada a semillas, particiones e hiperparámetros puede alterar rankings y conclusiones, lo que sugiere que la evaluación de modelos debe tratar la variabilidad como parte del resultado y no como un ruido administrativo (Bouthillier et al., 2021).

La noción de riesgo también permite distinguir dos preguntas: una operacional, referida a cuánto mejora un modelo, bajo un protocolo dado, cuando dispone de cierto atributo; y otra poblacional, orientada a cuánto debería mejorar el riesgo esperado si la distribución poblacional fuera conocida y el aprendizaje pudiera aproximar el predictor óptimo bajo una pérdida determinada. La distancia entre ambas preguntas es el lugar donde aparecen los problemas principales de esta tesis, pues una mejora empírica puede aproximar una cantidad poblacional, pero también reflejar una fluctuación de muestra, un sesgo de protocolo, una interacción con el algoritmo o un artefacto de representación.

Los experimentos controlados en línea muestran una versión aplicada de este problema: en pruebas A/B, por ejemplo, el diseño experimental intenta estimar efectos bajo asignación controlada y reducir varianza mediante covariables previas o diseños sensibles al contexto (Deng et al., 2013; Kohavi and Longbotham, 2017). Aunque el objetivo de esta tesis es distinto, estos trabajos ilustran una idea útil para la evaluación de relevancia: el protocolo no es un detalle posterior al modelo, sino parte de la evidencia, porque el modo de particionar, repetir, permutar o comparar determina qué afirmación queda razonablemente respaldada.

La representación de variables categóricas ofrece otro ejemplo, ya que algoritmos como CatBoost fueron diseñados para tratar estos atributos reduciendo ciertos sesgos derivados del uso ingenuo de estadísticas de objetivo (Prokhorenkova et al., 2018). La relevancia de una variable puede depender, entonces, tanto de la señal que contiene como del modo en que el algoritmo la codifica y la protege frente a fuga de información; por esa razón, una medida de importancia debería reportarse junto con las decisiones de representación y evaluación que hicieron posible su estimación.

2.4. Las dos culturas, el efecto Rashomon y la autoridad del modelo

Una forma productiva de leer las fuentes centrales de esta tesis consiste en reconstruir la pregunta que cada una deja abierta para la interpretación de atributos. En Breiman (2001b), la pregunta es cómo extraer conocimiento desde datos cuando el mecanismo que conecta $\boldsymbol{X}$ con Y permanece, en gran medida, oculto. Su crítica a la cultura de modelos de datos no se reduce a una preferencia por algoritmos más precisos; apunta a una práctica más profunda, en la que se asume una familia estocástica relativamente simple para el proceso generador y luego se interpretan sus parámetros como si el modelo hubiese capturado la estructura relevante del fenómeno. Frente a ello, la cultura algorítmica trata el mecanismo como una caja negra y juzga el procedimiento por su capacidad predictiva, abriendo un espacio donde árboles, bosques, redes neuronales y ensambles pueden ser científicamente útiles incluso cuando no ofrecen una forma paramétrica fácilmente narrable.

El gesto de Breiman puede entenderse, entonces, como un desplazamiento de autoridad. En la cultura de modelos de datos, la autoridad interpretativa descansa en la forma supuesta del modelo: si la regresión, el modelo lineal generalizado o el modelo de supervivencia se consideran representaciones razonables del fenómeno, sus coeficientes adquieren una lectura sustantiva. En la cultura algorítmica, en cambio, la autoridad se traslada al desempeño fuera de muestra: un modelo merece atención si logra anticipar datos no observados. El problema es que ninguna de esas autoridades basta por sí sola para resolver la pregunta por la relevancia de atributos. Una forma paramétrica puede ser interpretable y estar mal especificada; un predictor flexible puede ser preciso y, aun así, apoyarse en regularidades que no admiten una lectura poblacional o causal directa.

Shmueli (2010) introduce una segunda pieza de la historia al demostrar que explicación y predicción no son etapas intercambiables de un mismo procedimiento, sino propósitos que modifican el diseño completo del análisis. Su argumento es especialmente importante para esta tesis porque desplaza la discusión desde la apariencia del modelo hacia el uso inferencial que se hace de él. Un modelo explicativo busca sostener una historia sobre relaciones entre variables, normalmente apoyada en teoría previa, supuestos sobre confusión, interpretación de parámetros y contrastes de hipótesis. Un modelo predictivo, en cambio, puede prescindir de

una narración mecanística completa si logra anticipar casos nuevos con menor error; por ello puede incorporar transformaciones, regularización, variables auxiliares o proxies que serían problemáticos si se interpretaran directamente como causas o mecanismos.

Esta asimetría permite entender lo que suele ocurrir con la relevancia de atributos. Una variable puede mejorar la predicción porque resume información disponible en el entorno observacional, porque sustituye parcialmente a otra variable no medida o porque captura una práctica de medición que acompaña al desenlace. En un análisis predictivo, ese uso puede ser legítimo y aun así tener una lectura sustantiva limitada. Shmueli ofrece, por tanto, una advertencia metodológica de fondo: la pregunta no es si la predicción sirve o no para aprender, sino bajo qué condiciones una ganancia predictiva puede trasladarse a una afirmación sobre el fenómeno.

La respuesta de Miller et al. (2021) permite evitar una lectura excesivamente polarizada. En vez de elegir entre cultura explicativa y cultura algorítmica, estos autores proponen reconocer un espacio de modelos híbridos, donde conocimiento mecanístico, estructura estadística y flexibilidad computacional pueden combinarse. Esta lectura resulta importante para el estudio de la relevancia, porque sugiere que la pregunta no debería formularse como una competencia entre interpretación y desempeño, sino como una búsqueda de protocolos que indiquen qué tipo de evidencia produce cada método. La relevancia de un atributo puede ser operacionalmente útil, informativa a nivel poblacional o causalmente interpretable, pero esas lecturas deben ganarse mediante el diseño del análisis, no por la mera presencia de una tabla de importancias.

El efecto Rashomon radicaliza esta discusión. Breiman usó la imagen para describir situaciones en que múltiples modelos alcanzan desempeños comparables y, sin embargo, cuentan historias distintas sobre los datos. La literatura posterior formaliza esta intuición mediante el conjunto Rashomon: la colección de modelos casi igualmente precisos para un problema dado (Semenova et al., 2022). Si ese conjunto es amplio, el mejor modelo individual deja de ser una autoridad única; pasa a ser una realización dentro de una familia de explicaciones predictivamente plausibles. Esta observación es decisiva para la interpretación de atributos, porque la importancia de una variable puede variar entre modelos que predicen casi igual. En ese caso, el desempeño no permite adjudicar de manera directa qué variables son estructuralmente relevantes y cuáles son solo útiles para una de las posibles soluciones.

D'Amour (2021) formula esta dificultad en términos particularmente cercanos al proble-

ma de la tesis: modelos con criterios de precisión similares pueden procesar la información de los datos de maneras sustancialmente diferentes, lo que dificulta extraer conclusiones o automatizar decisiones a partir de un único ajuste. Un ejemplo estilizado ayuda a precisar el problema. Supóngase un escenario supervisado en que una respuesta clínica puede predecirse tanto mediante un marcador fisiológico asociado al estado del paciente como mediante una variable de contexto, por ejemplo el hospital, el protocolo de medición o la prioridad de atención, que resume indirectamente diferencias de gravedad ya presentes en los datos. Dos modelos pueden alcanzar errores fuera de muestra muy parecidos: uno puede apoyarse principalmente en el marcador fisiológico, mientras otro puede explotar la variable de contexto porque esta funciona como proxy de múltiples decisiones previas. Desde el punto de vista predictivo, ambos ajustes podrían parecer igualmente aceptables; desde el punto de vista epistémico, sin embargo, conducen a lecturas muy distintas sobre qué atributo es relevante y por qué. La dificultad no reside en que el desempeño agregado puede dejar indeterminado qué regularidad está siendo utilizada.

Fisher et al. (2019) llevan esta idea al estudio de importancia de variables al proponer que, en vez de preguntar cuánto depende un modelo de una variable, puede ser más informativo estudiar cuánto pueden depender de ella todos los modelos bien comportados dentro de una clase. En esa formulación, la relevancia deja de ser una propiedad puntual de un predictor y se convierte en un rango de dependencia dentro de una familia de modelos plausibles.

Los datos sintéticos adquieren aquí un papel central. En datos reales, el efecto Rashomon puede mostrar que varias historias predictivas son compatibles con el mismo desempeño, pero rara vez permite decidir cuál coincide con la estructura generadora. En un proceso sintético, en cambio, se puede diseñar deliberadamente una variable causal con señal observacional débil, una variable confundida con alta importancia predictiva, variables nulas, redundancias e interacciones sinergéticas. De ese modo, es posible observar cómo se separan las tres preguntas: qué explota el modelo, qué informa la población y qué pertenece al mecanismo causal. La historia que va de Breiman a Rashomon, pasando por Shmueli, conduce así a una exigencia metodológica concreta: la relevancia debe evaluarse bajo protocolos capaces de mostrar qué tipo de evidencia está produciendo la predicción.

2.5. Selección de variables y nociones de relevancia

La selección de variables reúne procedimientos orientados a escoger, ponderar o descartar atributos dentro de un problema predictivo o exploratorio (Guyon and Elisseeff, 2003). En aprendizaje automático se utiliza para reducir dimensionalidad, mejorar estabilidad, disminuir costos de medición, acelerar entrenamiento y facilitar la interpretación; en investigación científica, además, puede operar como una primera lectura de qué variables parecen asociarse con el fenómeno observado. Esta segunda función requiere cautela, porque un método de selección puede ser adecuado para construir un predictor compacto y, al mismo tiempo, insuficiente para mapear todos los atributos que contienen información sobre la respuesta.

Kohavi and John (1997) formularon una distinción influyente entre relevancia fuerte y débil en el contexto de selección de subconjuntos. En términos generales, una variable fuertemente relevante no puede eliminarse sin pérdida de información para la tarea, mientras una débilmente relevante puede ser redundante respecto de otras variables, aunque aún contenga señal. Esta idea abrió una dificultad que sigue vigente: si dos variables son copias o proxies cercanos, un predictor puede necesitar solo una, pero una investigación interesada en el fenómeno podría querer conservar ambas como evidencia de estructura.

Así, la selección de atributos puede orientarse tanto a optimizar el desempeño predictivo como a revelar formas de organización latente o evidencia de estructura en los datos. Desde esta perspectiva, es posible distinguir dos objetivos metodológicos: el problema *minimal-optimal* y el problema *all-relevant*. El problema minimal optimal privilegia un subconjunto reducido de atributos capaz de preservar o, eventualmente, mejorar el desempeño del modelo; por ello, se vincula de manera natural con criterios de parsimonia, reducción del costo computacional y eficiencia operacional. El enfoque *all-relevant*, por su parte, desplaza el énfasis hacia la recuperación exhaustiva de atributos informativos, aun cuando su contribución marginal no resulte necesaria para alcanzar el mejor desempeño predictivo posible (Kursa and Rudnicki, 2010; Mnich and Rudnicki, 2020; Nilsson et al., 2007).

La distinción ayuda a entender por qué la selección de variables aparece en dominios tan diversos. En detección de ataques DDoS para redes IoT, la selección dinámica de atributos puede reducir carga computacional y mejorar respuesta operacional (Ullah et al., 2023); en modelos de apoyo a decisiones odontológicas, la selección de características puede facilitar sistemas predictivos clínicamente manejables (Kang et al., 2023); y, en problemas de decisión

multicriterio, planificación, optimización no supervisada o construcción de reglas de despacho, los atributos seleccionados afectan tanto el rendimiento como la interpretabilidad de la solución (AL-Gburi et al., 2024; Durasevic et al., 2024; Kiratsoudis and Tsiantos, 2024). Estos ejemplos sugieren que la relevancia no es una propiedad abstracta aislada, sino que se define en relación con una tarea, un costo y una forma de uso.

Los métodos de selección suelen agruparse en tres familias (Guyon and Elisseeff, 2003; Kohavi and John, 1997). Los métodos de filtro evalúan atributos mediante criterios estadísticos relativamente independientes del modelo final, como correlaciones, pruebas de asociación o medidas basadas en información; un ejemplo influyente combina máxima relevancia respecto de la respuesta y mínima redundancia entre atributos seleccionados (Peng et al., 2005). Los métodos *wrapper* entrenan modelos con distintos subconjuntos y comparan su desempeño, de modo que la relevancia queda ligada a la tarea predictiva y al algoritmo usado (Kohavi and John, 1997). Los métodos incorporados o *embedded* integran la selección dentro del entrenamiento, por ejemplo mediante penalizaciones que inducen esparsidad, como el *lasso*, o mediante la importancia interna de modelos de árbol (Breiman, 2001a; Tibshirani, 1996). Ninguna familia domina a las demás en todos los escenarios, ya que cada una responde a una noción distinta de utilidad, en línea con la ausencia de un método universalmente superior (Wolpert, 1996).

En problemas de alta dimensionalidad esta heterogeneidad se vuelve particularmente visible: en datos ómicos, por ejemplo, el número de variables puede superar ampliamente al número de observaciones y la redundancia biológica puede producir múltiples subconjuntos plausibles (Degenhardt et al., 2019). A ello se suman benchmarks recientes de selección de características en redes neuronales, donde métodos modernos pueden tener dificultades para detectar señales no lineales cuando hay muchas variables distractoras (Passemiers et al., 2024). Estas observaciones refuerzan una idea central para la tesis: la evaluación de relevancia debe considerar la estructura del problema, no solo el ranking final de atributos.

2.6. Medidas de dependencia e información

Las medidas clásicas de dependencia, como la correlación de Pearson, Spearman o Kendall, ofrecen descripciones útiles y de lectura relativamente directa: Pearson resume asociación lineal, mientras Spearman y Kendall describen asociaciones monótonas a partir de rangos.

Su claridad explica su persistencia en análisis exploratorios y reportes científicos; sin embargo, los datos de alta dimensión suelen exhibir no linealidad, asimetrías, redundancias, relaciones por umbrales e interacciones contextuales, de modo que una baja correlación lineal puede coexistir con una dependencia predictivamente relevante.

La teoría de la información amplía esta perspectiva al formalizar la dependencia como reducción de incertidumbre (Cover and Thomas, 2006; Shannon, 1948). Sea Y la variable respuesta y X un atributo predictivo. Si la entropía $H(Y)$ representa la incertidumbre inherente asociada a Y, entonces la información mutua $I(X;Y) = H(Y) - H(Y \mid X)$ cuantifica cuánto se reduce esa incertidumbre al observar X. Esta definición es válida tanto cuando Y es discreta (entropía de Shannon) como cuando es continua (entropía diferencial). A diferencia de la correlación, la información mutua no se limita a relaciones lineales o monótonas, por lo que constituye un marco más general para describir asociaciones estadísticas. Sin embargo, su interpretación requiere cuidado: $I(X;Y)$ es simétrica, refleja únicamente dependencia observacional y, por sí sola, no indica dirección causal ni utilidad predictiva operacional bajo un modelo específico.

La información mutua condicional $I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}})$ introduce un refinamiento especialmente útil para esta tesis, donde S denota el conjunto de atributos considerado en una evaluación, representado por sus posiciones en $\boldsymbol{X}$. La cantidad evalúa si X_j aporta información sobre la respuesta una vez observado el subvector de contexto $\boldsymbol{X}_{S\setminus\{j\}}$, es decir, los atributos de S salvo X_j. En ese sentido, se aproxima a una noción de informatividad poblacional incremental: pregunta si X_j reduce la incertidumbre sobre Y más allá de lo que ya permite conocer el contexto declarado. Cuando S corresponde al conjunto completo de atributos, esta cantidad recupera la formulación $I(Y; X_j \mid \boldsymbol{X}_{\setminus j})$, es decir, la contribución incremental respecto del resto de las covariables. Esta formulación ayuda a distinguir entre dependencia marginal, redundancia y contribución contextual. Una variable puede ser marginalmente informativa y perder aporte condicional si su señal ya está contenida en otro atributo; de modo inverso, puede mostrar poca asociación marginal y volverse informativa al considerar interacciones. Por ello, la información condicional resulta conceptualmente cercana al problema de relevancia, aunque su interpretación depende del conjunto de covariables observado, de la distribución poblacional considerada y de las decisiones de estimación disponibles en la práctica.

Esta familia ha motivado numerosos criterios de selección. Brown et al. (2012) propo-

nen un marco que unifica varios métodos de selección basados en información al equilibrar relevancia respecto de la respuesta y redundancia respecto de variables ya seleccionadas, generalizando criterios previos como el de máxima relevancia y mínima redundancia (Peng et al., 2005). En variables continuas o mixtas, la estimación de información mutua requiere decisiones adicionales, como discretización, estimadores por vecinos cercanos o supuestos de suavidad (Kraskov et al., 2004; Ross, 2014). En grandes conjuntos de datos, el coeficiente de información maximal buscó detectar asociaciones generales y no lineales (Reshef et al., 2011); herramientas posteriores facilitaron su cálculo y aplicación práctica (Albanese et al., 2013, 2018), aunque también surgieron extensiones y debates sobre su sensibilidad y generalización (Luedtke and Tran, 2013).

La estimación empírica de dependencia enfrenta límites importantes. En muestras pequeñas, una medida flexible puede detectar patrones accidentales; en alta dimensión, las comparaciones múltiples pueden inflar hallazgos aparentes; en presencia de confusión, una dependencia observacional puede reflejar causas comunes a ambas variables, fenómeno que la Sección 2.8 desarrolla; y en presencia de interacciones, una variable puede ser marginalmente débil pero condicionalmente informativa.

Un ejemplo ilustrativo aparece en genómica, donde Chen and Zhang (2013) propusieron los *module-guided Random Forests* (mgRF) para integrar marcadores genéticos y expresión génica en el estudio del peso de ratones. En este enfoque, un módulo no designa una unidad biológica asumida de antemano, sino un grupo de variables altamente correlacionadas identificado a partir de una red de correlación entre atributos; en el caso genómico, esos grupos pueden reunir SNPs, es decir, polimorfismos de nucleótido único, en desequilibrio de ligamiento, genes coexpresados o combinaciones de marcadores y expresiones asociadas. El Random Forest se modifica entonces para utilizar esa estructura: en cada división se seleccionan primero módulos candidatos y luego variables representativas dentro de esos módulos, con actualizaciones iterativas de importancia de módulo e importancia de variable. De esta forma, la dependencia entre atributos pasa a organizar la estructura misma del procedimiento de selección y aprendizaje y deja de ser tratada únicamente como una complicación posterior en el ranking de importancia. Otras medidas intentan conservar una lectura cercana a la correlación clásica en contextos categóricos, ordinales o mixtos. El coeficiente ϕ_k propuesto por Baak et al. (2020), por ejemplo, busca mantener propiedades interpretables similares a Pearson para tipos de datos heterogéneos, incluyendo variables categóricas, ordinales e

intervalares.

Una referencia especialmente cercana al problema de esta tesis es el *Predictive Power Score* (PPS), presentado como una alternativa predictiva, direccional y agnóstica al tipo de dato (Wetschoreck, 2020). Su lógica formula la relación entre X e Y de forma univariada: para cada par ordenado de variables, evalúa cuánto mejora la predicción de la variable objetivo Y cuando se dispone únicamente del atributo candidato X como entrada. Debido a su carácter asimétrico, el puntaje de X para predecir Y puede diferir del puntaje de Y para predecir X.

Este enfoque univariado prepara la discusión de la sección siguiente, donde la relevancia se examina como una mejora de desempeño respecto de una referencia explícita: un predictor ingenuo (mediana o moda, dependiendo del problema) en el caso del PPS, y atributos de prueba o copias nulas en los métodos basados en referencias nulas.

2.7. Métodos basados en modelos y contrastes nulos

Las medidas de dependencia descritas en la sección anterior preguntan si dos variables comparten estructura estadística; los métodos basados en modelos desplazan la pregunta hacia el uso predictivo de esa estructura. En este registro, un atributo recibe apoyo como relevante cuando su presencia, ausencia o perturbación modifica el desempeño de un predictor bajo una muestra, una pérdida, una familia de modelos y un protocolo de evaluación. La relevancia deja entonces de presentarse como una propiedad directa del par (X_j, Y) y pasa a formularse como una comparación contrafactual: cuánto cambia la calidad predictiva cuando el atributo observado se sustituye, se elimina o se enfrenta a una referencia que representa ausencia de señal.

La importancia por permutación, asociada a Random Forest de Breiman (2001a), ofrece una formulación influyente de esta intuición. Primero se estima el desempeño del modelo sobre datos fuera de entrenamiento, típicamente observaciones *out-of-bag* en Random Forest o un conjunto de validación. Luego, para cada atributo, se permutan sus valores entre observaciones y se vuelve a evaluar el mismo modelo sobre esa versión perturbada de los datos. Si el desempeño disminuye, los resultados sugieren que el predictor aprovechaba regularidades asociadas a ese atributo; si permanece estable, la variable puede ser redundante, débil bajo la métrica elegida o irrelevante para ese predictor; si mejora, el atributo original pudo haber introducido ruido, sobreajuste o una interacción desfavorable con el procedimiento

de aprendizaje. Esta ambivalencia será retomada en la Sección 2.8, porque muestra que la perturbación no separa por sí sola señal, ruido y protocolo.

La fortaleza de la permutación es también una fuente de cautela. Al permutar X_j, el procedimiento conserva su distribución marginal, pero rompe tanto su correspondencia con Y como su dependencia con las demás covariables. Por ello, el deterioro de desempeño estima una cantidad operacional: cuánto pierde el predictor entrenado cuando se le entrega una copia nula de ese atributo bajo el mismo esquema de evaluación. Esa cantidad puede resultar muy informativa, aunque su interpretación como dependencia poblacional o como efecto causal requiere supuestos adicionales. Además, en Random Forest y métodos afines, ciertas medidas de importancia pueden favorecer variables con más categorías, escalas particulares o estructuras que facilitan divisiones del árbol (Strobl et al., 2007). La pregunta metodológica se vuelve, entonces, más precisa: conviene especificar contra qué referencia se juzga una magnitud de importancia producida por el modelo.

Las referencias nulas cómo contraste aparecen como una respuesta a esa necesidad de comparación explícita. Stoppiglia et al. (2003) propusieron incorporar al conjunto de variables potencialmente relevantes un atributo de prueba (probe) generado aleatoriamente y, por construcción, sin relación sistemática con la variable objetivo. Los atributos observados se evalúan en relación con esa señal de control: un atributo recibe apoyo como relevante si supera al atributo de prueba bajo el criterio usado por el procedimiento. El aporte conceptual es claro, pues la relevancia se interpreta frente a una expectativa nula y no solo frente al ranking interno de las variables observadas. Sin embargo, esa referencia opera como control global; puede orientar la selección, pero no ofrece una contraparte emparejada que conserve las propiedades marginales particulares de cada atributo.

Boruta desarrolla esta intuición en una dirección más cercana a la importancia por permutación. El algoritmo construye atributos *shadow*, es decir, versiones permutadas de los atributos originales que mantienen una estructura marginal comparable y rompen su asociación con la respuesta (Kursa and Rudnicki, 2010). En lugar de preguntar únicamente si un atributo tiene alta importancia dentro del conjunto observado, Boruta evalúa si supera de manera consistente a las mejores referencias *shadow* a través de iteraciones de Random Forest. Esta lógica se ajusta especialmente bien al problema *all-relevant*, porque busca recuperar atributos con evidencia de señal aunque resulten redundantes para un subconjunto predictivo mínimo. Su resultado principal, con todo, es una decisión de aceptación, rechazo

o indeterminación dentro de un procedimiento de selección; el énfasis recae menos en estimar una diferencia de desempeño emparejada, atribuible a cada variable, bajo una métrica reportable y un protocolo reproducible.

El PPS ocupa una posición complementaria porque usa una referencia de desempeño antes que una copia nula del atributo. En su implementación de referencia, ajusta un árbol de decisión con un único atributo candidato como entrada, evalúa el desempeño mediante validación cruzada y normaliza el resultado contra un predictor ingenuo de la variable objetivo. En regresión, esa referencia corresponde a predecir siempre la mediana; en clasificación, al mejor desempeño entre predecir siempre la clase mayoritaria y asignar clases al azar (Wetschoreck, 2020). Esta construcción aporta una lectura direccional y uno-a-uno de utilidad predictiva; su límite para esta tesis es que permanece bivariada, evalúa cada atributo fuera de un contexto multivariado y compara contra un predictor ingenuo antes que contra una copia nula emparejada del propio atributo.

Sobre estos antecedentes persiste, sin embargo, una misma limitación de fondo. La importancia por permutación evalúa una perturbación global del atributo, sin fijar una contraparte nula explícita ni una unidad de evidencia reportable; Stoppiglia y Boruta introducen esa contraparte nula, pero la resuelven como control global o como decisión de aceptación, rechazo o indeterminación, más que como una diferencia de desempeño emparejada bajo un protocolo declarado; y PPS ofrece una comparación direccional atribuible a una variable específica, pero frente a un predictor ingenuo antes que frente a una copia nula del propio atributo. Ninguno de estos antecedentes combina, a la vez, una perturbación bajo el mismo modelo, una contraparte nula emparejada que conserve la forma marginal del atributo y una atribución reportable a una variable concreta bajo un protocolo declarado.

Esta limitación conjunta importa porque, sin una unidad de evidencia que combine esas tres exigencias, un resultado positivo bajo perturbación queda expuesto a lecturas ambiguas: puede reflejar una ventaja genuina del atributo observado, o bien depender de tamaño muestral, partición, repetición, regularización, métrica o familia de modelos, sin que el protocolo permita distinguir entre ambas posibilidades. El Capítulo 3 retoma esta necesidad para formalizar una operación que satisfaga conjuntamente esas tres exigencias.

2.8. Ruido, redundancia, confusión y aprendizaje de atajos

La evaluación de relevancia se vuelve más compleja cuando las variables no actúan de manera aislada, como ocurre habitualmente en datos reales, donde pueden coexistir ruido de medición, atributos redundantes, variables confundidas, fuga de información e interacciones que modifican el valor de una variable según la presencia de otras. Estos fenómenos no solo agregan dificultad empírica, sino que afectan de manera distinta a medidas estadísticas, métodos basados en modelos y explicaciones post-hoc.

El ruido introduce una primera dificultad, pues una variable puede contener una señal débil mezclada con variabilidad irrelevante o ser puramente nula y, aun así, mostrar asociación accidental en una muestra finita. Las medidas sensibles a fluctuaciones muestrales pueden asignar importancia positiva a atributos no informativos, especialmente cuando se evalúan muchos atributos en paralelo; este problema se intensifica en alta dimensión, donde aumenta el número de oportunidades para encontrar patrones aparentes. Además, la literatura sobre entrenamiento con ruido muestra que este no siempre actúa como degradación: puede funcionar como regularizador, modificar el paisaje de optimización o, en casos de resonancia estocástica, elevar el desempeño del sistema al volver más aprovechable una señal débil (Bishop, 1995; Zhai et al., 2023), de modo que una mejora de desempeño admite múltiples lecturas inferenciales: puede reflejar señal del fenómeno, pero también una interacción entre ruido, modelo, métrica y protocolo.

La redundancia plantea un problema distinto: dos atributos pueden portar información semejante sobre la respuesta, de modo que un modelo predictivo necesite solo uno para mantener el desempeño. Desde una perspectiva de selección mínima, descartar uno puede ser razonable; desde una perspectiva *all-relevant*, en cambio, ambos pueden ser relevantes porque describen aspectos asociados al fenómeno. Esta diferencia explica por qué una medida que penaliza redundancia puede ser útil para construir modelos compactos y, al mismo tiempo, menos apropiada para mapear todos los atributos asociados con la respuesta.

La confusión introduce una tercera capa, porque una variable puede estar asociada con la respuesta por compartir con ella una causa común, es decir, por depender ambas de una tercera variable que las origina a la vez, por reflejar un proceso de selección o por

capturar un sesgo de medición. Para un predictor, esa variable puede ser útil; para una interpretación causal directa, en cambio, su uso exige cautela. El caso de neumonía y asma ilustra esta capa con precisión: en modelos de riesgo de neumonía, el asma aparece asociada a un menor riesgo estimado, pese a que clínicamente no constituye un factor protector; la explicación plausible no está en un mecanismo biológico beneficioso, sino en el proceso de atención, ya que los pacientes asmáticos suelen recibir tratamiento más intensivo, monitoreo temprano o decisiones clínicas distintas, y esa trayectoria modifica la asociación observada entre asma y desenlace (Caruana et al., 2015). El ejemplo muestra que una misma variable puede ocupar lugares distintos según el registro inferencial: en el plano operacional, puede mejorar el desempeño; en el plano poblacional, puede informar sobre una distribución ya afectada por prácticas clínicas; en el plano causal, puede carecer del efecto protector que una lectura ingenua sugeriría. De modo similar, modelos de fractura de cadera o COVID-19 han aprendido variables de sistema de salud, hospital o adquisición que aumentan desempeño sin constituir explicación sustantiva del fenómeno biológico (Badgeley et al., 2019; DeGrave et al., 2021; Roberts et al., 2021).

Por otro lado, Lapuschkin et al. (2019) mostraron que las explicaciones pueden revelar estrategias tipo Clever Hans, en las que el modelo acierta por explotar señales periféricas. En este contexto, se entiende por aprendizaje de atajos la tendencia de un predictor a apoyarse en claves fáciles de explotar dentro del conjunto observado, pero frágiles ante cambios de contexto o poco alineadas con la estructura sustantiva que se desea estudiar. Pathak et al. (2026) revisan este fenómeno en términos de correlaciones espurias y robustez, subrayando que la aparente competencia del modelo puede descansar en dependencias débiles desde el punto de vista epistémico. Sin ser el mismo fenómeno, los ejemplos adversariales de Szegedy et al. (2014) agregan otra pieza al argumento, al mostrar que modelos de alto desempeño pueden ser sensibles a perturbaciones que apenas modifican la percepción humana, lo que sugiere que sus fronteras de decisión pueden responder a geometrías difíciles de interpretar sustantivamente.

Estos fenómenos justifican el uso de controles explícitos: si una medida de relevancia asigna alta importancia a una variable confundida, puede estar describiendo fielmente la conducta del predictor, pero el problema aparece cuando esa fidelidad se lee como causalidad o como comprensión del mecanismo. Una tesis sobre relevancia de atributos debe, por tanto, separar la pregunta por utilidad predictiva de la pregunta por el origen de esa utilidad.

2.9. Interacciones, sinergias y estructuras de orden superior

Multiples medidas de relevancia se interpretan de manera univariada, preguntando cuánto aporta una variable por sí sola; sin embargo, algunos patrones dependen de combinaciones, de modo que una variable puede tener poca utilidad marginal y adquirir relevancia cuando se evalúa junto con otra. Los patrones tipo XOR son el ejemplo canónico: cada atributo por separado puede parecer débil, mientras el par contiene una señal clara. Esta estructura desafía tanto filtros marginales como métodos de importancia que no exploran adecuadamente dependencias de orden superior.

La búsqueda de interacciones entre genes ilustra, en genética, la dificultad de detectar efectos que no se reducen a asociaciones marginales (Cordell, 2009); en series financieras y otros dominios con bajo cociente señal-ruido, de modo análogo, las interdependencias entre indicadores pueden emerger solo mediante modelos capaces de explorar combinaciones (Donick and Lera, 2021). Los benchmarks de selección con redes neuronales refuerzan esta preocupación, pues las señales no lineales y escasas pueden quedar ocultas entre variables distractoras, tensionando métodos que se comportan bien en escenarios lineales o de efectos fuertes (Passemiers et al., 2024).

Las interacciones también complican la lectura causal, ya que una variable causal puede mostrar un efecto marginal débil si su impacto depende del valor de otras variables; de modo inverso, una variable no causal puede ser predictivamente fuerte por su posición en una estructura de confusión. La inferencia desde relevancia observada hacia causalidad se vuelve especialmente frágil en estos casos, lo que refuerza la necesidad de distinguir entre una variable que el modelo puede explotar, una variable que aporta información poblacional sobre la respuesta y una variable cuya intervención modifica la respuesta.

Desde el punto de vista metodológico, las interacciones sugieren que una medida univariada debe entenderse como una lectura parcial: puede ser útil para detectar señales individuales, contrastar variables contra referencias nulas o construir un primer mapa de atributos, pero su silencio frente a una variable no debería interpretarse de inmediato como irrelevancia estructural. En diseños sintéticos, esta distinción puede examinarse de manera explícita al construir variables que son nulas marginalmente pero informativas de forma conjunta.

2.10. Datos sintéticos y procesos generadores controlados

El uso de datos sintéticos en esta tesis responde a una necesidad metodológica: evaluar medidas de relevancia requiere escenarios donde la verdad de fondo sea conocida. En datos observacionales reales, rara vez se sabe con certeza qué variables son causalmente activas, cuáles son solo informativas, cuáles funcionan como proxies, cuáles están confundidas y cuáles corresponden a ruido puro. En esta tesis, una variable confundida designa una variable asociada con la respuesta por compartir una causa, un proceso de selección o una trayectoria de medición, sin que esa asociación implique necesariamente un efecto causal directo; una variable nula designa un atributo que, por construcción del proceso generador, carece de relación sistemática con la respuesta. Por eso, aunque un conjunto real puede mostrar que dos métodos discrepan, suele entregar poca evidencia para decidir si la discrepancia se debe a una falla del método, a una regularidad espuria, a una interacción no modelada o a una propiedad genuina del fenómeno.

Los procesos generadores controlados permiten definir una regla explícita para producir atributos y respuesta. En términos de modelos causales estructurales, esa regla puede representar causas directas, causas comunes, mediadores, variables irrelevantes, redundancias e interacciones (Pearl, 2009); desde una perspectiva estadística, permite conocer la distribución objetivo o aproximarla mediante múltiples realizaciones independientes. Así, el experimento deja de depender exclusivamente de una muestra fija y puede evaluar si la medida recupera una relación que persiste cuando el fenómeno se vuelve a generar bajo las mismas condiciones.

Esta capacidad es crucial para estudiar falsos positivos: si se desea saber si una medida asigna importancia a variables nulas, hace falta disponer de variables nulas por construcción; si se quiere examinar confusión, el diseño debe incluir variables asociadas con la respuesta a través de una causa común, pero sin efecto causal directo; y, si se quiere evaluar sinergias, se necesitan variables marginalmente débiles y conjuntamente informativas. Los datos reales pueden contener estas estructuras, pero rara vez las entregan etiquetadas; los datos sintéticos permiten aislarlas, graduarlas y observar la respuesta del protocolo ante cambios de ruido, tamaño muestral, redundancia, no linealidad e interacción.

La objeción más fuerte contra esta estrategia es que un proceso sintético puede parecer

artificial o demasiado diseñado para el argumento que se desea sostener. Esa objeción debe tomarse en serio. La respuesta de esta tesis no consiste en presentar los datos sintéticos como reemplazo de dominios reales, sino como una forma de ingeniería inversa epistemológica: se construyen mecanismos conocidos para observar qué parte de ellos recupera una medida, qué parte queda oculta bajo un protocolo y qué parte se confunde con utilidad predictiva. La artificialidad, en este contexto, no es un defecto accidental, sino una condición de control; su valor depende de que el diseño represente patrones metodológicamente relevantes y de que sus límites se declaren con precisión.

La literatura reciente ofrece buenos argumentos para esta estrategia: Passemiers et al. (2024) muestran que escenarios sintéticos relativamente controlados pueden revelar limitaciones de métodos modernos de selección y saliencia en redes neuronales; Donick and Lera (2021) utilizan relaciones de tipo XOR para mostrar cómo ciertas interdependencias pueden permanecer ocultas en ambientes de alto ruido; y Olivetti et al. (2026) reportan evidencia de que incluso bajo desafíos diseñados para causal discovery, clasificar roles causales desde datos observacionales resulta difícil. A ello se suman ejemplos aplicados que muestran por qué la verdad de fondo rara vez puede suponerse conocida: asociaciones clínicas modificadas por decisiones de tratamiento (Caruana et al., 2015), atajos de adquisición o de institución en imágenes médicas (Badgeley et al., 2019; DeGrave et al., 2021), e indeterminación causal en estudios observacionales donde varias historias mecanísticas siguen siendo compatibles con los mismos datos (Hammerton and Munafò, 2021). En conjunto, estos trabajos sugieren que los datos sintéticos no son una simplificación trivial, sino un laboratorio donde se puede estudiar la frontera entre lo que el algoritmo detecta y lo que el investigador sabe que fue generado.

La sinteticidad también permite examinar protocolos, pues una misma variable nula puede recibir importancia distinta si se evalúa bajo bootstrap, validación cruzada o realizaciones independientes del proceso generador. Esta diferencia no es un accidente menor: indica que el protocolo define la cantidad empírica que se estima. En una muestra fija, el remuestreo puede reutilizar fluctuaciones particulares; en realizaciones independientes, en cambio, la variabilidad del proceso generador queda representada de otra manera. Para distinguir relevancia operacional de informatividad poblacional, esta separación es central.

El ruido merece una atención específica porque, aunque en muchos diseños se trata como residuo que dificulta el aprendizaje, la literatura sugiere que puede actuar como regularizador,

modificar la robustez o incluso mejorar el desempeño mediante mecanismos de resonancia estocástica (Bishop, 1995; Zhai et al., 2023). Un proceso sintético permite convertir esa ambivalencia en un factor experimental: se puede aumentar su intensidad, hacerlo independiente, correlacionarlo con variables informativas, introducirlo como distractor o analizarlo como copia nula. Esto permite estudiar no solo si un método detecta señal, sino también si logra abstenerse frente a variables que no deberían recibir relevancia.

La defensa de los datos sintéticos debe reconocer sus límites, ya que un proceso generador demasiado simple puede favorecer involuntariamente a la medida propuesta, uno demasiado estrecho puede exagerar la generalidad de los resultados y una distribución artificial puede omitir artefactos de medición, sesgos de selección o cambios de dominio presentes en aplicaciones reales. Por ello, los datos sintéticos se entienden aquí como instrumentos de identificación conceptual y validación controlada, cuya fuerza reside en permitir preguntas que los datos reales rara vez responden de forma directa: qué ocurre cuando la verdad de fondo se conoce, cómo cambia la medida ante perturbaciones específicas y qué interpretaciones sobreviven cuando el experimento se repite desde el proceso generador, no solo desde una muestra observada.

Los procesos generadores controlados permiten, además, ordenar la investigación como una secuencia metodológica: definir primero los tipos de relevancia, diseñar luego escenarios donde esos tipos se separan, evaluar las medidas bajo controles explícitos y, finalmente, interpretar los resultados según el alcance de la evidencia. Los datos sintéticos son, en este sentido, el espacio donde se vuelve observable la diferencia entre causas, información y predictibilidad.

2.11. Explicabilidad, atribución y alcance de las explicaciones

La inteligencia artificial explicable (XAI, por *explainable artificial intelligence*) ha producido métodos orientados a describir, auditar o comunicar el comportamiento de modelos complejos, y ha sido objeto de revisiones sistemáticas que ordenan sus enfoques y propósitos (Guidotti et al., 2018; Molnar, 2022). Entre los métodos post-hoc más difundidos cabe mencionar a LIME y SHAP; LIME construye una explicación local alrededor de una observación específica: genera perturbaciones cercanas, consulta el predictor original en ese vecindario y

ajusta un modelo sustituto interpretable, usualmente lineal y esparso, ponderando con mayor fuerza las observaciones más próximas al caso que se desea explicar (Ribeiro et al., 2016). Su resultado ofrece una aproximación local de la variación predictiva alrededor de un punto y sitúa la explicación en el vecindario perturbado que el método construye.

SHAP, por su parte, formula la explicación como una descomposición aditiva de la predicción respecto de un valor de referencia. Inspirado en los valores de Shapley de la teoría de juegos cooperativos, asigna a cada atributo una contribución marginal promedio al considerar distintas coaliciones de variables disponibles (Lundberg and Lee, 2017). Esta formulación entrega una gramática atractiva para repartir crédito predictivo entre atributos, con propiedades formales de consistencia y aditividad bajo ciertos supuestos. Sin embargo, su interpretación depende del valor base, de la distribución de referencia usada para comparar predicciones y del modo en que se tratan las dependencias entre atributos. En consecuencia, LIME y SHAP ofrecen herramientas prácticas para examinar predictores difíciles de inspeccionar directamente, pero sus atribuciones describen ante todo el comportamiento del modelo entrenado, no una medida inmediata de informatividad poblacional ni de causalidad.

La literatura ha señalado, sin embargo, que "explicabilidad" agrupa objetivos distintos: para algunos usuarios, una explicación debe aumentar confianza; para otros, debe permitir auditoría, depuración, acción, rendición de cuentas o generación de hipótesis (Lipton, 2018; Sokol and Flach, 2020). Frente a la proliferación de explicaciones post-hoc, algunos autores defienden que, en decisiones de alto riesgo, conviene preferir modelos intrínsecamente interpretables antes que explicar modelos opacos a posteriori (Rudin, 2019). La perspectiva human-centered XAI enfatiza justamente que las explicaciones deben evaluarse en relación con usuarios, contextos y propósitos, no solo como propiedades algorítmicas (Ehsan and Riedl, 2023; Ridley, 2025). En aplicaciones como física de altas energías, calidad de vida urbana o recursos humanos, SHAP y LIME se han usado para comunicar patrones aprendidos por modelos, aunque su interpretación depende de la representación, el dominio y la pregunta de uso (Arunika et al., 2024; Pezoa et al., 2023; Ríos-Vásquez et al., 2025).

La distinción entre explicar el modelo y explicar el fenómeno es clave, porque una atribución post-hoc puede ser fiel al predictor entrenado y, al mismo tiempo, limitada como explicación del proceso generador. Si el modelo aprende una regularidad espuria, una variable confundida o una señal dependiente del protocolo de adquisición, la explicación puede describir con precisión esa dependencia aprendida; en ese caso, la herramienta cumple una

función valiosa de auditoría, pero su resultado no debe confundirse con evidencia causal.

Esta limitación se observa también en debates recientes sobre la estabilidad de las explicaciones: Hwang et al. (2025) muestran que las explicaciones basadas en SHAP pueden ser sensibles a la representación de las características; Knab et al. (2025) discuten variantes y desafíos de LIME; y Biecek and Samek (2024) proponen orientar la explicabilidad hacia la formulación de preguntas críticas sobre el modelo, más que hacia la justificación retrospectiva de sus decisiones. Estas contribuciones son coherentes con la tesis, pues una explicación útil no es necesariamente una explicación causal, y su valor aumenta cuando ayuda a delimitar qué sabe el modelo, qué usa y qué podría estar explotando indebidamente.

La explicabilidad formal de árboles de decisión muestra otro matiz. Aunque los árboles suelen presentarse como modelos transparentes por su estructura de reglas, la literatura reciente examina explicaciones con requisitos lógicos y probabilísticos más exigentes (Arenas et al., 2024, 2025). Una razón suficiente, por ejemplo, puede entenderse como un subconjunto de condiciones sobre los atributos que basta para garantizar la misma predicción del árbol bajo todas las asignaciones posibles de las variables restantes compatibles con esos valores; su minimalidad exige que cada condición incluida tenga un papel efectivo en esa garantía. Una explicación probabilística, en cambio, relaja esa exigencia universal y evalúa con qué probabilidad una predicción se mantiene cuando las variables restantes varían bajo una distribución de referencia. En ambos casos, explicar exige precisar el objeto formal de la explicación, el criterio de suficiencia, la noción de minimalidad, la distribución considerada y el costo computacional de calcularla. Así, la interpretabilidad aparece como una relación entre modelo, lenguaje explicativo, usuario y propósito.

La discusión sobre representaciones aprendidas en cerebros y máquinas agrega una advertencia conceptual, ya que estudios de alineamiento entre modelos visuales y representaciones cerebrales sugieren que arquitecturas distintas pueden alcanzar desempeños similares o alineamientos próximos bajo ciertas métricas, mientras sus sesgos inductivos y experiencias de entrenamiento difieren (Conwell et al., 2024; Nature Machine Intelligence, 2025). Esta observación prolonga el argumento Rashomon desarrollado en la Sección 2.4: múltiples modelos pueden producir predicciones comparables y apoyarse en representaciones distintas. Por ello, la atribución de importancia a una variable debe entenderse como dependiente del predictor y del protocolo, no como lectura directa y única del fenómeno.

2.12. Información, predictibilidad y causalidad

La información mutua condicional ofrece una referencia formal para hablar de informatividad poblacional. Si se considera una distribución conjunta fija y una pérdida logarítmica, la ganancia poblacional asociada a incluir una variable puede expresarse en términos de la información que esa variable aporta sobre la respuesta dado un contexto de atributos; el caso completo corresponde a condicionar en el resto de las covariables. Esta relación conecta el lenguaje predictivo con la teoría de la información: una variable informativa reduce incertidumbre sobre Y en la distribución poblacional (Brown et al., 2012; Cover and Thomas, 2006; Shannon, 1948).

La causalidad responde a otra pregunta. En la tradición de los modelos causales estructurales, una variable es causalmente relevante respecto de Y si intervenir sobre ella puede modificar la distribución de Y bajo algún contexto (Pearl, 2009; Peters et al., 2017); esta noción depende de supuestos sobre mecanismos, variables no observadas, intervenciones posibles y estructura del sistema, y ha sido formalizada tanto desde el formalismo de grafos e intervenciones (Pearl, 2009; Pearl and Mackenzie, 2018) como desde los algoritmos de descubrimiento de estructura causal a partir de datos (Spirtes et al., 2000). La dependencia estadística puede surgir por causalidad, confusión, selección o medición, mientras que una variable causal puede mostrar una señal observacional débil si su efecto se cancela marginalmente, depende de interacciones o queda mediado por otras variables.

La predictibilidad operacional constituye un tercer registro, referido a lo que un sistema de aprendizaje puede explotar bajo condiciones concretas: una muestra, una representación, un modelo, una pérdida, una métrica y un procedimiento de evaluación. Como este espacio incluye señal genuina, patrones de ruido, confusión, regularización implícita, fuga de información y efectos de protocolo, la predictibilidad operacional es empíricamente importante, pero no agota la interpretación.

Trabajos recientes sobre explicabilidad y causalidad advierten precisamente contra el desplazamiento automático desde importancia de variables hacia inferencia causal (Hammerton and Munafò, 2021; Vowels, 2024). Los desafíos de descubrimiento causal refuerzan el punto: incluso cuando se trabaja con datos diseñados para evaluar causalidad, la identificación de roles causales desde patrones observacionales puede ser limitada (Olivetti et al., 2026). Estos antecedentes permiten formular una tensión que el capítulo siguiente convertirá en

una gramática propia: una variable puede mejorar el desempeño de un predictor por razones de representación, protocolo, confusión o muestra, y una causa puede quedar débilmente visible para un predictor concreto cuando su efecto depende de interacciones, mediaciones, compensaciones marginales o escalas de medición. La separación entre utilidad predictiva y relevancia causal queda así establecida como un problema de alcance inferencial, antes que como una diferencia meramente terminológica.

A partir de esa tensión, esta tesis trata la informatividad poblacional como un registro epistémico autónomo, es decir, como el plano de lectura en el que se pregunta si una variable aporta información sobre la respuesta en la población, qué evidencia sostiene esa lectura y qué alcance inferencial puede atribuírsele. Su interés consiste en identificar dependencias entre atributos y respuesta, con valor propio para la descripción, la comparación de fenómenos y la generación de hipótesis. El aporte del trabajo consiste en articular este registro con otros dos registros, causalidad y predictibilidad operacional, y en proponer que el contraste *null-swap* evalúe cuándo el desempeño asociado al atributo observado se distingue de lo esperable bajo una contraparte nula comparable: una copia que conserva su forma marginal y rompe su vínculo sistemático con la respuesta. En el Capítulo 3, esta distinción se formaliza mediante los espacios $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$, y la relevancia epistémica precisa cuándo esa distinguibilidad operacional puede sostener una lectura sobre $\mathcal{I}$.

2.13. Reproducibilidad y protocolos

La reproducibilidad en aprendizaje automático incluye la posibilidad de volver a ejecutar un procedimiento, regenerar resultados comparables y sostener conclusiones bajo variaciones razonables del experimento. Siguiendo la taxonomía de Goodman et al. (2016), adaptada al contexto computacional por Bouthillier et al. (2019), pueden distinguirse tres niveles. La reproducibilidad de métodos se refiere a la capacidad de regenerar el experimento con el mismo código, los mismos datos y un entorno suficientemente especificado; en este nivel importan semillas, dependencias, versiones de bibliotecas, contenedores y documentación del flujo de ejecución. La reproducibilidad de resultados exige que una reimplementación o reconstrucción admisible del procedimiento produzca valores estadísticamente similares, desplazando la atención desde un número exacto hacia una distribución de métricas. La reproducibilidad inferencial pregunta si la afirmación que se desea defender se mantiene bajo

un diseño experimental razonablemente distinto, por ejemplo otra partición, otra muestra, otra inicialización, otra trayectoria de optimización o una reimplementación del protocolo.

Bouthillier et al. (2019) muestran que estas capas pueden entrar en tensión: fijar una semilla favorece la reproducción exacta de un resultado, pero también puede restringir la conclusión a una realización particular del experimento. En sus comparaciones de modelos, múltiples réplicas por semilla y por conjunto de datos revelan que los rankings pueden variar de manera suficiente como para alterar la conclusión empírica. En la misma línea, Bouthillier et al. (2021) modelan el benchmark como un proceso con fuentes de variación asociadas al muestreo de datos, la inicialización, el aumento de datos y la elección de hiperparámetros, y proponen incorporar explícitamente esa varianza al comparar algoritmos.

Esta literatura es especialmente relevante para la evaluación de atributos mediante contrastes con una copia nula. Cada diferencia entre el atributo observado y su copia nula es una realización de un procedimiento aleatorio que involucra particiones, permutaciones, semillas, hiperparámetros y, en ocasiones, muestras sintéticas independientes. Por ello, las repeticiones cumplen una función inferencial: permiten estimar si la ventaja del atributo observado se distingue de la contraparte nula de manera estable, o si corresponde a una fluctuación inducida por el protocolo. La validación cruzada, el bootstrap y las realizaciones independientes representan formas distintas de promediar variabilidad; al reportarlas, la tesis trata el protocolo como parte de la afirmación de relevancia.

La conexión entre reproducibilidad y datos sintéticos vuelve a ser un punto central ya que, cuando el proceso generador es conocido, se puede repetir el experimento desde la distribución y no solo desde una muestra observada. Esta posibilidad permite separar estabilidad operacional de evidencia epistémica: un método puede ser reproducible en el sentido de producir el mismo sesgo bajo el mismo protocolo, pero la pregunta más exigente es si esa estabilidad se orienta hacia la cantidad que se desea interpretar. El Capítulo 3 retoma esta exigencia al formalizar el contraste *null-swap* y las condiciones bajo las cuales su resultado puede leerse como evidencia poblacional.

Las fuentes revisadas en este capítulo muestran que el problema tiene raíces metodológicas y epistemológicas: las dos culturas explican por qué desempeño e interpretación responden a autoridades distintas; el efecto Rashomon muestra que un único modelo puede ser una base insuficiente para narrar la relevancia; la selección de variables distingue parsimonia de recuperación *all-relevant*; las medidas de información separan dependencia de causalidad;

los métodos de explicación diferencian fidelidad al modelo de comprensión del fenómeno; los datos sintéticos permiten evaluar bajo verdad de fondo conocida; y la reproducibilidad exige reportar la variación inducida por protocolos. En conjunto, estas líneas convergen en una misma exigencia: distinguir, para cada resultado, qué tipo de afirmación autoriza antes de declararlo evidencia de relevancia.

Capítulo 3

Propuesta metodológica y conceptual

3.1. Tres espacios epistémicos

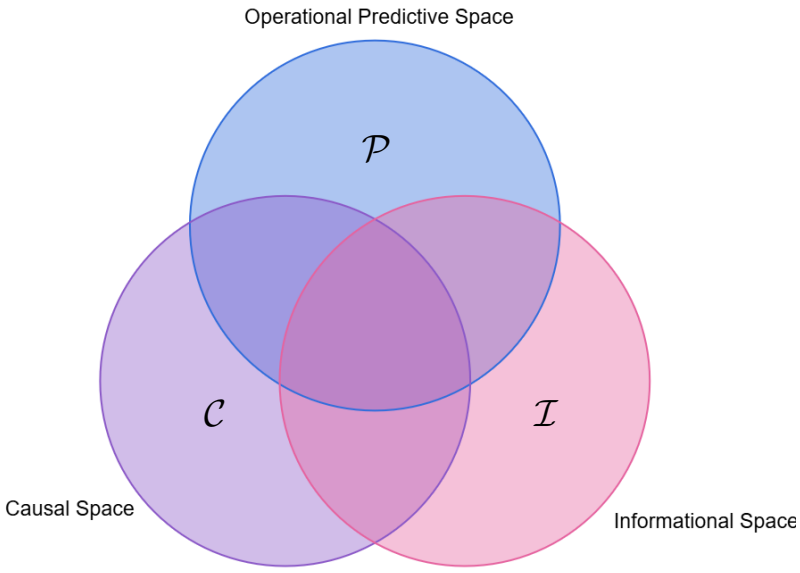


Figura 3.1: Espacios causal, informativo y predictivo. La figura representa de manera esquemática que los espacios causal, informativo y predictivo pueden intersectarse sin volverse equivalentes. Adaptado del marco conceptual publicado en Neirz et al. (2026).

En este trabajo, la noción de registro epistémico designa el tipo de afirmación que se formula sobre una variable y las condiciones de evidencia bajo las cuales esa afirmación resulta defendible. Esta noción permite precisar qué clase de pregunta se plantea cuando una variable es descrita como relevante:

- **Registro causal:** qué variables modificarían la respuesta bajo una intervención sobre ellas.
- **Registro informativo:** qué variables reducen la incertidumbre sobre la respuesta en la distribución poblacional.
- **Registro predictivo:** qué variables mejoran el desempeño de un predictor, o son aprovechadas por este, bajo una muestra, un modelo, una métrica y un protocolo concretos.

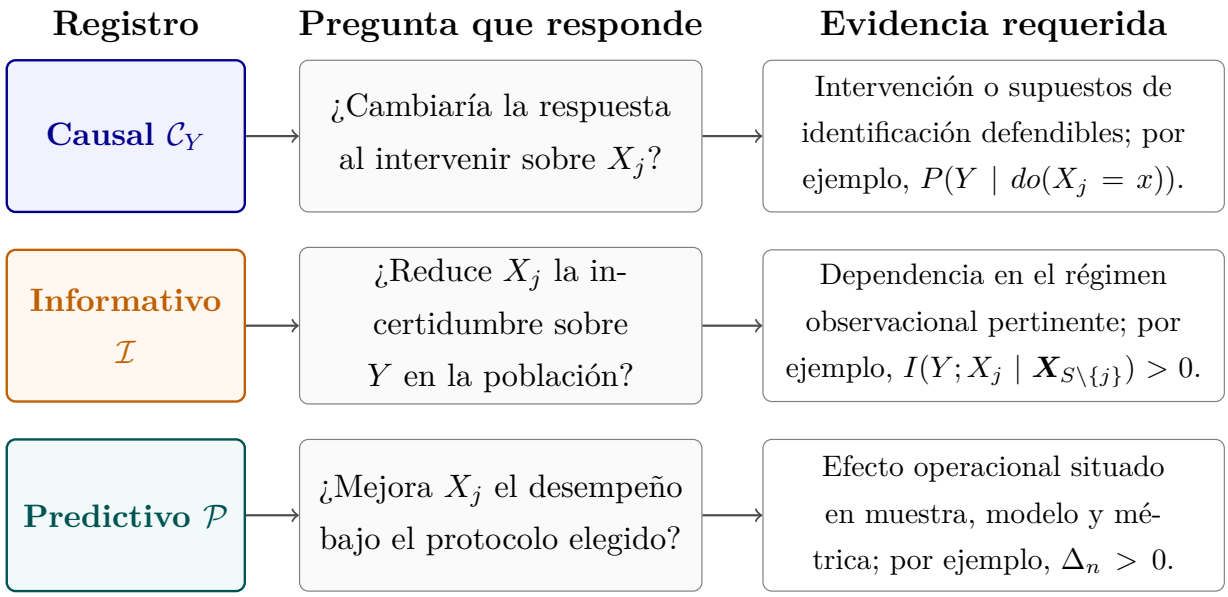


Figura 3.2: Preguntas y evidencias características de los registros causal, informativo y predictivo. Cada fila resume el diseño y el tipo de resultado que sustentan la pertenencia al espacio correspondiente. El esquema presenta sus diferencias; las coincidencias entre registros dependen del proceso generador y de condiciones adicionales.

Cada registro, aplicado a una variable, induce una condición de pertenencia a un espacio epistémico: la variable pertenece al espacio causal, al de informatividad poblacional

o al predictivo operacional cuando la afirmación correspondiente puede sostenerse bajo las condiciones de evidencia propias de ese registro. En aplicaciones reales, una misma variable puede satisfacer más de una de esas condiciones, o satisfacer una sin satisfacer las demás; por ello, reducir estas pertenencias a una única noción de importancia bajo la etiqueta de variable relevante tiende a producir sobreinterpretaciones y malos entendidos. Sobre esta base, proponemos distinguir tres espacios epistémicos: causalidad, informatividad poblacional y predictibilidad operacional. La Figura 3.1 representa de manera esquemática que esos espacios pueden intersectarse y conservar su diferencia conceptual. La Figura 3.2 sintetiza la pregunta y la evidencia propias de cada registro. La Sección 3.2 desarrolla esta separación como una propiedad estructural del problema y presenta sus demostraciones mediante contraejemplos.

Para fijar la convención notacional, en lo que sigue S denota el conjunto de atributos considerado en una evaluación, representado por las posiciones de esos atributos en $\boldsymbol{X}$. Formalmente, $S \subseteq \{1, \ldots, p\}$ y, cuando se evalúa X_j, se asume $j \in S$. La operación $S \setminus \{j\}$ retira del conjunto evaluado la posición correspondiente a X_j; por tanto, $\boldsymbol{X}_{S\setminus\{j\}}$ denota el subvector de $\boldsymbol{X}$ formado por los atributos de S salvo X_j. Cuando S queda fijado por el contexto, escribimos $\boldsymbol{X}_{\setminus j}$ por simplicidad; si $S = \{1, \ldots, p\}$, esta abreviatura corresponde al conjunto completo de covariables excepto X_j.

La notación distingue dos niveles. En el nivel poblacional, X_j designa una variable aleatoria escalar: el atributo j-ésimo del vector aleatorio de atributos $\boldsymbol{X} = (X_1, \ldots, X_p)$. En una muestra $D_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, esa misma variable se observa como una columna $\boldsymbol{x}_{\cdot j} := (x_{1j}, \ldots, x_{nj})^\top$. Por economía de escritura, expresiones como retirar, reemplazar o permutar X_j se refieren al atributo como objeto conceptual; cuando la operación es empírica, actúa sobre la columna muestral correspondiente.

3.1.1. Espacio causal

Para fijar la lectura de la notación causal, el operador $do(\cdot)$ designa una intervención idealizada sobre el sistema, en el sentido del cálculo causal de Pearl: $do(X_j = x)$ representa un régimen en el cual X_j se fija externamente en x y el mecanismo que normalmente genera esa variable es sustituido por dicha fijación. Las demás relaciones estructurales permanecen gobernadas por el modelo causal, salvo aquellas alteradas por la intervención declarada. Esta

notación permite distinguir dos diseños. En un diseño observacional, $P(Y \mid X_j = x)$ representa la distribución condicional de Y entre los casos que presentan $X_j = x$, bajo el régimen de observación que generó los datos. En un diseño experimental o intervencional idealizado, $P(Y \mid do(X_j = x))$ representa la distribución de Y cuando la fijación externa de X_j modifica ese régimen generador. Ambas expresiones describen distribuciones, pero corresponden a preguntas y diseños distintos (Pearl, 2009). En la ecuación siguiente, $do(\boldsymbol{X}_{S\setminus\{j\}} = \boldsymbol{x}_{S\setminus\{j\}})$ cumple el mismo papel para las covariables usadas como contexto.

Con esta convención, definimos el espacio causal $\mathcal{C}_Y$ como aquel que contiene variables cuya intervención puede modificar la distribución de Y bajo algún contexto. Formalmente, $X_j \in \mathcal{C}_Y$ si existen valores $x \neq x'$, un conjunto evaluado $S \subseteq \{1, \ldots, p\}$ con $j \in S$ y una realización $\boldsymbol{x}_{S\setminus\{j\}}$ tales que:

$$P(Y \mid do(X_j = x), do(\boldsymbol{X}_{S\setminus\{j\}} = \boldsymbol{x}_{S\setminus\{j\}})) \neq P(Y \mid do(X_j = x'), do(\boldsymbol{X}_{S\setminus\{j\}} = \boldsymbol{x}_{S\setminus\{j\}})). \quad (3.1)$$

La ecuación expresa una diferencia bajo intervención: se comparan dos regímenes en los que el investigador fija X_j en valores distintos, manteniendo el contexto indicado por $\boldsymbol{X}_{S\setminus\{j\}}$. Si la distribución de Y cambia, la variable tiene relevancia causal en ese contexto. Este espacio no exige que la variable sea fácil de detectar con correlaciones marginales. Una causa puede tener señal observacional débil si su efecto depende de otras variables, si opera solo en ciertos subgrupos o si queda parcialmente compensada por mecanismos del proceso generador.

3.1.2. Espacio de informatividad poblacional

El espacio de informatividad poblacional $\mathcal{I}$ agrupa variables que aportan información sobre Y en la distribución poblacional. Una formulación contextual considera que $X_j \in \mathcal{I}$ si existe algún conjunto evaluado $S \subseteq \{1, \ldots, p\}$ con $j \in S$ tal que:

$$I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) > 0. \quad (3.2)$$

El valor de la ecuación es positivo cuando observar X_j reduce la incertidumbre sobre Y una vez consideradas las covariables de contexto $\boldsymbol{X}_{S\setminus\{j\}}$ (Cover and Thomas, 2006). La condición vale para Y discreta o continua; la Proposición 3.2 añade la restricción a Y discreta únicamente para obtener la identidad exacta con el riesgo de Bayes. Esta información puede

provenir de una causa directa, pero también de una variable confundida, de un proxy, de una práctica de medición o de una consecuencia del fenómeno. Por ejemplo, una variable asociada a un protocolo hospitalario puede ser muy informativa para predecir un desenlace sin representar un mecanismo biológico que deba intervenirse.

3.1.3. Espacio predictivo operacional

El espacio predictivo operacional $\mathcal{P}$ corresponde a aquello que un predictor efectivamente puede explotar bajo un protocolo concreto. En el sentido amplio propuesto por Neirz et al. (2026), $\mathcal{P}$ funciona como un contenedor operacional de lo que una canalización de aprendizaje puede explotar bajo una muestra, una clase de modelos, una pérdida o métrica y un protocolo de evaluación determinados. Si T denota ese conjunto de decisiones de modelamiento y evaluación, y Π una distribución sobre las reimplementaciones admisibles E del protocolo (semillas, particiones de entrenamiento y prueba, inicializaciones), la pertenencia a $\mathcal{P}$ se refiere a la incorporación operacional de una variable por el procedimiento aprendido, no a un contraste específico ni a una certificación poblacional. En esta tesis se interpreta que $X_j \in \mathcal{P}(T, \Pi; D_n)$ cuando, bajo alguna reimplementación admisible y algún conjunto evaluado S que contiene a j, el procedimiento aprendido incorpora X_j como parte de la regla predictiva o de la representación que conduce a la predicción. La idea cubre casos simples: una regresión que conserva un coeficiente asociado a X_j, un árbol que divide nodos según X_j, o un procedimiento de selección que retiene esa variable.

Dado su carácter amplio y laxo, $\mathcal{P}$ recoge una acepción frecuente en la práctica de la ciencia de datos industrial: una variable es "relevante" si el procedimiento logra incorporarla de alguna forma aprovechable bajo el protocolo observado. Esta laxitud es deliberada y cumple una función de punto de partida. Permite registrar qué atributos entran en el funcionamiento del predictor antes de evaluar si esa incorporación es estable, consistente o distinguible de alternativas nulas. En consecuencia, $\mathcal{P}$ incluye tanto señal genuina como regularización implícita, redundancias, fugas de información, sensibilidad algorítmica, efectos de selección y artefactos finitos de muestra. Su valor metodológico radica en organizar esta primera lectura operacional, sobre la cual se pueden luego imponer requisitos evidenciales más exigentes (consistencia del protocolo y distinguibilidad frente al ruido). Esta asimetría respecto de $\mathcal{C}_Y$ y $\mathcal{I}$ responde a su función: $\mathcal{P}$ registra lo que una canalización concreta incorpora. El contraste,

la estabilidad y la consistencia permiten auditar esa incorporación. Cuando el contexto es claro, se escribe simplemente $\mathcal{P}$.

Sin embargo, precisamente por su naturaleza amplia, la mera pertenencia a $\mathcal{P}$ resulta demasiado permisiva para muchas aplicaciones analíticas. Para avanzar hacia mediciones más útiles y discriminativas, utilizaremos la noción de contraste operacional. Este se define como una comparación empírica, realizada dentro de un mismo protocolo T y una misma reimplementación admisible E, entre el desempeño de un modelo entrenado con el atributo observado y el desempeño obtenido tras una modificación controlada de la información disponible. Las modificaciones consideradas en este trabajo consisten en excluir X_j del subvector $\boldsymbol{X}_S$ o en reemplazarlo por una copia nula que preserve sus propiedades marginales. De este modo, los contrastes operacionales operan sobre una pertenencia a $\mathcal{P}$ ya establecida y permiten cuantificar el impacto concreto de modificar controladamente la información predictiva disponible, así como evaluar si un efecto observado resiste una comparación adecuada. En lo que sigue, $L_E(M; D_n)$ denota la pérdida (o métrica) de evaluación del modelo M bajo la reimplementación E. Como primer ejemplo, consideramos el retiro puntual de una variable (*leave-one-variable-out*, denotado loo). Sea $M_E^{(S)}$ el modelo entrenado con las covariables de S bajo E, y $M_E^{(S\setminus\{j\})}$ el modelo entrenado tras excluir X_j. Entonces:

$$\Delta^{\text{loo}}_{n,S,j}(E) = L_E(M_E^{(S\setminus\{j\})}; D_n) - L_E(M_E^{(S)}; D_n), \tag{3.3}$$

Dado que L_E se interpreta aquí como pérdida, un valor positivo indica que retirar la variable aumenta el error y, por tanto, que el modelo entrenado con $\boldsymbol{X}_S$ se beneficia de ella bajo esa reimplementación. Un valor cercano a cero sugiere que el retiro altera poco el desempeño. Un valor negativo indica que el modelo sin la variable obtiene menor pérdida; ese resultado puede aparecer por ruido, redundancia, sobreajuste o interacción con la clase de modelos. La lectura de estos signos queda acotada al protocolo: una mejora puntual registra incidencia operacional, pero su utilidad robusta exige evidencia adicional sobre estabilidad, magnitud y comparación con una copia nula.

Conviene precisar el estatuto de los objetos que intervienen en este contraste. Los modelos $M_E^{(S)}$ y $M_E^{(S\setminus\{j\})}$ son elementos concretos del procedimiento: resultan de entrenar la clase de modelos declarada por T sobre la muestra finita D_n, y la pérdida L_E se evalúa fuera de la muestra de entrenamiento. En general, ambos difieren del predictor de Bayes (el predictor ideal que alcanza el menor riesgo posible bajo la distribución poblacional), y el contraste evita

cualquier supuesto de optimalidad sobre ellos. Por eso, $\Delta^{\text{loo}}_{n,S,j}(E)$ es un hecho empírico acerca de un procedimiento de aprendizaje y evaluación aplicado a una muestra finita. Cuando registra una mejora con probabilidad positiva, documenta una consecuencia de desempeño asociada al atributo, pero su interpretación como señal poblacional o como evidencia causal requiere las condiciones adicionales desarrolladas en las secciones siguientes. Abordaremos en más detalle el predictor de Bayes en la Sección 3.3 como referencia poblacional del marco: allí se pregunta bajo qué condiciones estas cantidades empíricas convergen hacia brechas de riesgo definidas en esos términos.

La amplitud de $\mathcal{P}$ resulta útil porque refleja la práctica real del modelamiento: un predictor puede incorporar una variable por una dependencia poblacional estable, por confusión, por una codificación del proceso de medición o por una regularidad accidental de la muestra. Agrupar esos casos en el espacio predictivo operacional permite reconocer que todos pueden aparecer inicialmente como uso predictivo del atributo, aunque su interpretación posterior requiera criterios distintos.

3.1.4. Contraste null-swap

Una segunda forma de medición, formalizada en este trabajo, es el contraste *null-swap*. Sus antecedentes inmediatos articulan componentes que aquí se integran: la importancia por permutación de Breiman (2001a) estima cuánto se altera el desempeño de un modelo ya ajustado cuando, durante la evaluación, se permutan los valores de un atributo de entrada; la comparación con atributos aleatorios de Stoppiglia et al. (2003) y los atributos *shadow* de Boruta (Kursa and Rudnicki, 2010) introducen copias nulas explícitas; y el *Predictive Power Score* de Wetschoreck (2020) subraya la conveniencia de una comparación direccional, atribuible a una variable específica, frente a una línea de base de desempeño. A partir de esos elementos, el *null-swap* (denotado ns en los superíndices que siguen) evalúa el aporte de un atributo observado frente a una copia nula construida bajo el mismo problema predictivo. En su forma más general, el contraste se define sobre un conjunto evaluado S que contiene a j: el modelo completo usa $\boldsymbol{X}_S$, mientras que el modelo nulo conserva los demás atributos de S y reemplaza X_j por una copia nula $\pi_j(X_j)$:

$$\tilde{\Delta}^{\text{ns}}_{n,S,j}(E) = L_E(M_E^{(S,j\leftarrow\pi_j)}; D_n) - L_E(M_E^{(S)}; D_n). \tag{3.4}$$

La cantidad $\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E)$ es la brecha empírica del contraste *null-swap*: la diferencia de pérdida entre el modelo nulo, entrenado con la copia permutada de X_j, y el modelo completo, entrenado con el atributo observado, bajo la reimplementación E. La virgulilla distingue esta familia de cantidades, asociada al contraste *null-swap*, de $\Delta^{\mathrm{loo}}_{n,S,j}(E)$, que no la lleva; en la Sección 3.5 se reutiliza además para marcar un agregado por mediana sobre reimplementaciones, distinción que se explicita al introducirla. Aquí π_j denota la operación de construcción de la copia nula para el atributo j. En el nivel empírico, suele implementarse como una permutación de la columna observada $\boldsymbol{x}_{\cdot j}$ bajo el protocolo declarado; en el nivel poblacional, $\pi_j(X_j)$ representa la copia nula asociada al atributo. La notación $\pi_j(X_j)$ se reserva para esa copia nula, mientras que Π designa la distribución sobre reimplementaciones admisibles del protocolo. El procedimiento completo, desde la construcción de la copia nula hasta la comparación de los dos modelos entrenados, se ilustra paso a paso en la Figura 3.3.

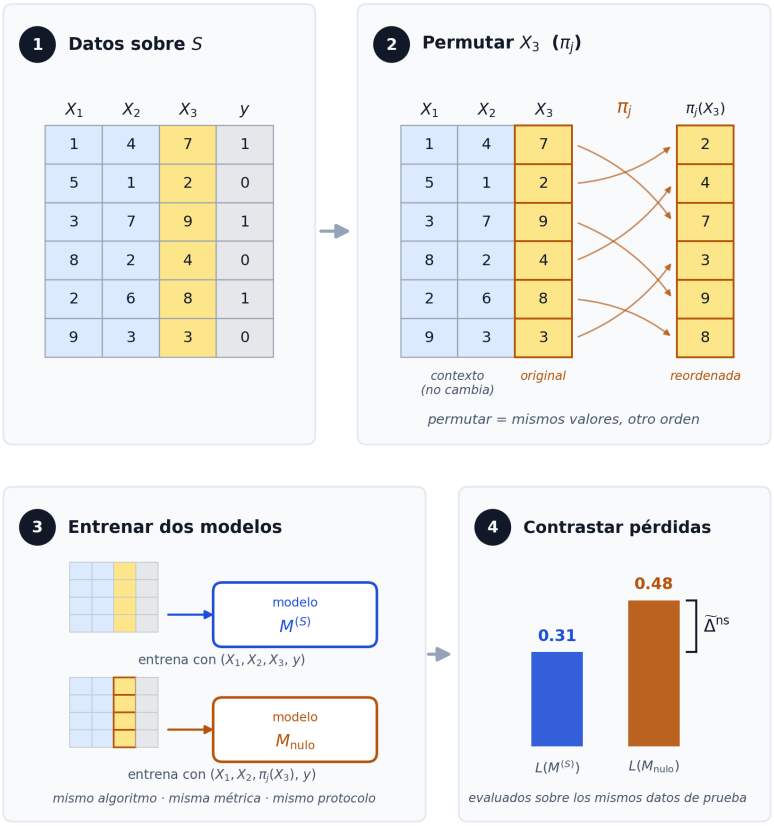


Figura 3.3: El contraste *null-swap* paso a paso. (1) Los datos sobre el conjunto evaluado S contienen las covariables de contexto X_1, X_2 (azul), el atributo evaluado X_3 (amarillo) y la respuesta y (gris). (2) En la muestra finita, permutar la columna observada $\boldsymbol{x}_{\cdot 3}$ consiste en reordenar sus valores entre las filas: la copia nula $\pi_3(\boldsymbol{x}_{\cdot 3})$ conserva exactamente los mismos valores y con ello la distribución marginal empírica, pero en otro orden, lo que rompe su emparejamiento con y; las columnas de X_1 y X_2 se mantienen como contexto. (3) Se entrenan dos modelos con las covariables de S, el mismo algoritmo y el mismo protocolo: $M^{(S)}$ con los datos originales y M_{nulo} con la copia permutada en lugar de la columna observada de X_3. (4) Ambos modelos se evalúan con la misma métrica sobre los mismos datos de prueba; una brecha positiva $\widetilde{\Delta}^{\text{ns}}_{n,S,j}$, definida en la Ecuación (3.4), indica que la asociación observada de X_3 con la respuesta aporta más que una copia nula comparable construida desde el propio problema.

En pérdidas, un valor positivo significa que permutar X_j empeora el desempeño respecto

del atributo real; el resultado sugiere una ventaja operacional de la asociación observada entre X_j e Y. Un valor cercano a cero indica que el atributo real y su copia nula son prácticamente indistinguibles bajo el protocolo. Un valor negativo puede señalar que la copia permutada funcionó mejor en una realización específica, lo que recomienda cautela antes de atribuir señal al atributo original; esta posibilidad es coherente con la discusión sobre ruido y resonancia estocástica presentada en la Sección 2.8. En métricas donde valores mayores son mejores, el signo se invierte.

El orden del contraste corresponde al número de atributos incluidos en el conjunto evaluado S, es decir, a $|S|$. Esta denominación describe el tamaño del contexto de evaluación, no la cantidad de aportes estimados simultáneamente: en cada contraste el aporte se estima para un atributo específico X_j. En primer orden, $S = \{j\}$; se evalúa X_j sin covariables de contexto, comparando su versión observada con la copia nula $\pi_j(X_j)$. En segundo orden, $S = \{i, j\}$; se evalúa el aporte de X_j en presencia de X_i. La dirección importa, porque X_i permanece intacta como contexto y solo X_j se reemplaza por su copia nula. La Figura 3.3 ilustra un contraste de tercer orden: el conjunto evaluado contiene tres atributos, pero el atributo puesto a prueba es X_3, mientras X_1 y X_2 permanecen disponibles como contexto. En órdenes superiores, la misma idea se extiende a conjuntos S más grandes: se permuta el atributo evaluado y las demás variables de S permanecen intactas. En todos los casos, la variable de respuesta Y permanece fija; lo que se permuta es el atributo evaluado.

El contraste por retiro y el *null-swap* no son intercambiables en muestras finitas. Puede ocurrir, por ejemplo, que reemplazar X_j por una copia permutada empeore el modelo, mientras que retirarlo tenga un efecto nulo o incluso mejore el desempeño. La diferencia no constituye una contradicción conceptual: retirar una variable cambia la dimensionalidad del problema, la regularización efectiva, la selección interna de variables y, en algunos algoritmos, la forma misma del ajuste; reemplazarla por una copia nula conserva un atributo comparable, mantiene fijo el resto del procedimiento y pregunta si la asociación observada con la respuesta aporta más que una versión sin vínculo sistemático. Por esa razón, el *null-swap* será el contraste usado más adelante para ARS y para la condición de distinguibilidad del ruido, mientras que LOO queda como una medición auxiliar de consecuencias por exclusión. A nivel poblacional, la Sección 3.3 establece que, bajo una copia nula apropiada y el mismo conjunto evaluado S, ambos contrastes pueden identificarse con la misma brecha de riesgo; sus discrepancias empíricas pertenecen al plano finito-muestral y de diseño.

3.2. Separación formal de los tres espacios

La distinción anterior tendría una utilidad limitada si los tres espacios epistémicos coincidieran siempre salvo en casos excepcionales. El resultado central del marco establece lo contrario: sin supuestos adicionales sobre el proceso generador, la pertenencia a un espacio epistémico no permite inferir la pertenencia a otro.

Proposición 3.1 (separación de los tres espacios). *En ausencia de supuestos adicionales sobre el proceso generador por ejemplo, fidelidad, ausencia de variables latentes o identificabilidad completa, la pertenencia a $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$ no induce relaciones generales de implicación. En particular, ninguna de las seis implicaciones posibles entre los tres espacios se cumple en general.*

Demostración. La estrategia consiste en exhibir contraejemplos mediante modelos causales estructurales. Cada caso muestra que una pertenencia puede sostenerse mientras otra falla; en conjunto, los cuatro casos cubren las seis direcciones posibles.

(i) *Causal sin informatividad ni predictibilidad* ($\mathcal{C}_Y \not\Rightarrow \mathcal{I}$, $\mathcal{C}_Y \not\Rightarrow \mathcal{P}$).

Sea $U \sim \text{Bernoulli}(1/2)$ una variable latente y defínase

$$X_j := U, \qquad Y := X_j(1-U).$$

En la distribución observacional se tiene $Y = U(1-U) = 0$ casi seguramente. Por tanto,

$$I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) = 0$$

para todo $S \subseteq \{1,\dots,p\}$ con $j \in S$, y $X_j \notin \mathcal{I}$. Además, un protocolo orientado a predecir una respuesta constante, con regularización apropiada, puede producir predictores que no incorporan X_j en ninguna reimplementación admisible; bajo ese protocolo, $X_j \notin \mathcal{P}$. Bajo intervención, en cambio, $do(X_j = 1)$ produce $Y = 1-U \sim \text{Bernoulli}(1/2)$, mientras que $do(X_j = 0)$ produce $Y = 0$ de forma determinista. Luego $X_j \in \mathcal{C}_Y$.

(ii) *Informativa y predictiva sin causalidad* ($\mathcal{I} \not\Rightarrow \mathcal{C}_Y$, $\mathcal{P} \not\Rightarrow \mathcal{C}_Y$).

Sea $Z \sim \mathcal{N}(0,1)$ una causa común y sean $e_1, e_2 \sim \mathcal{N}(0,1)$ ruidos independientes. Defínase

$$X_j := Z + e_1, \qquad Y := Z + e_2.$$

La causa común induce dependencia observacional entre X_j e Y, de modo que $I(X_j; Y) > 0$ y $X_j \in \mathcal{I}$. Un predictor puede explotar esa dependencia bajo un protocolo adecuado, por ejemplo mediante un modelo lineal evaluado por pérdida cuadrática, por lo que $X_j \in \mathcal{P}$ en ese protocolo. Sin embargo, intervenir sobre X_j no modifica el mecanismo que genera Y, que depende solo de Z y e_2; así, $do(X_j = x)$ deja invariante la distribución de Y. Luego $X_j \notin \mathcal{C}_Y$.

(iii) *Informativa sin predictibilidad* ($\mathcal{I} \not\Rightarrow \mathcal{P}$).

Sea $X_j \sim \mathcal{N}(0, 1)$ y defínase

$$Y := X_j^2.$$

La dependencia es determinista, de modo que $I(Y; X_j) > 0$ y $X_j \in \mathcal{I}$. Sin embargo,

$$\mathrm{Cov}(X_j, Y) = \mathbb{E}[X_j^3] = 0.$$

Por ello, el mejor predictor lineal de Y bajo pérdida cuadrática es constante y el coeficiente poblacional asociado a X_j es nulo. Bajo un protocolo cuya clase de modelos es lineal, con pérdida cuadrática y selección consistente de coeficientes, por ejemplo, regularización ℓ_1 o un umbral sobre coeficientes estimados, la probabilidad de que el procedimiento retenga un coeficiente para X_j tiende a cero. Para muestras D_n en ese evento, $X_j \notin \mathcal{P}(T, \Pi; D_n)$ en ninguna reimplementación admisible. La variable es informativa en la distribución poblacional, pero su señal cuadrática queda fuera de la representación que el protocolo puede explotar.

(iv) *Predictiva sin informatividad* ($\mathcal{P} \not\Rightarrow \mathcal{I}$).

Sea X_j una variable nula por construcción: un atributo de ruido independiente de Y y de las covariables informativas del proceso generador. Entonces $X_j \notin \mathcal{I}$, pues observarlo no reduce la incertidumbre poblacional sobre la respuesta.

Aun así, bajo un protocolo finito, X_j puede entrar en $\mathcal{P}$: ya sea por una asociación accidental de la muestra que el procedimiento incorpora, ya sea porque actúa como perturbación beneficiosa para una regla de umbral, una clase de modelos o una dinámica de entrenamiento. La Figura 3.4 ilustra esta segunda posibilidad mediante resonancia estocástica: el ruido puede mejorar el desempeño del sistema sin constituir información poblacional sobre Y.

Los cuatro casos muestran que cada espacio puede separarse de los otros bajo condiciones admisibles del problema. Por ello, ninguna de las seis implicaciones se sostiene en general. □

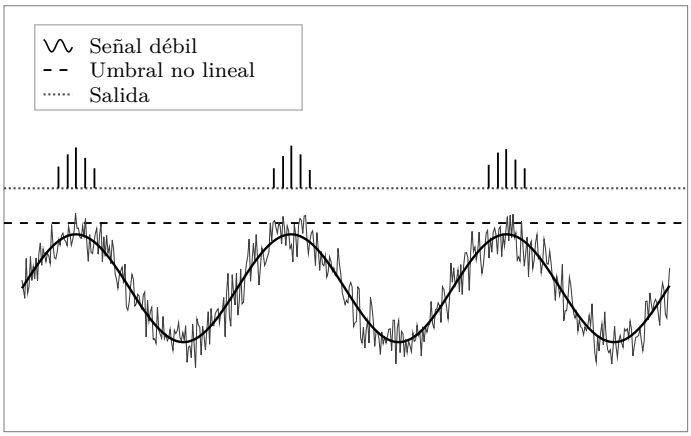


Figura 3.4: Resonancia estocástica y utilidad operacional del ruido. Una señal subumbral mantiene inactivo el detector. Al combinarse con ruido independiente, algunos cruces del umbral pueden alinearse con los máximos de la señal y mejorar el desempeño del sistema. El esquema es una elaboración propia basada en el mecanismo de señal débil, umbral no lineal y fuente de ruido descrito por Gammaitoni et al. (1998). La figura ilustra el caso (iv): el ruido contribuye al protocolo como perturbación operacional y conserva independencia poblacional respecto de la respuesta.

La proposición delimita el contenido lógico del marco: cada espacio responde a una pregunta con condiciones de justificación propias, y las coincidencias entre espacios son contingentes al proceso generador, a la distribución observacional y al protocolo. Establecer una de esas coincidencias exige, en cada caso, evidencia adicional a la pertenencia misma. La dificultad práctica de tender puentes entre los espacios es consistente con la evidencia disponible en desafíos de descubrimiento causal, donde clasificar roles causales desde patrones observacionales resulta difícil incluso bajo condiciones idealizadas (Olivetti et al., 2026). El Capítulo 4 ilustra estas separaciones de manera empírica mediante un modelo causal estructural con verdad de fondo conocida, en el que conviven una variable causal con firma observacional débil, una variable confundida predictivamente fuerte y un par sinergético.

3.3. Consistencia del protocolo y Bayes-consistencia

3.3.1. Consistencia respecto de una cantidad declarada

Para que la transición desde una mejora operacional hacia una lectura poblacional tenga contenido metodológico, conviene distinguir dos niveles de consistencia. El primero es la consistencia respecto de una cantidad de desempeño declarada. Un protocolo puede estimar de manera estable una métrica poblacional, como accuracy, error cuadrático, F_1 o AUC, sin que esa métrica entregue por sí misma una lectura de informatividad en el sentido de $\mathcal{I}$. En tal caso, el resultado sostiene una afirmación de valor predictivo poblacional para la métrica elegida, pero su alcance permanece ligado al objetivo operacional que esa métrica define.

Formalmente, si Q denota una cantidad poblacional de desempeño asociada a la métrica declarada y $\widehat{Q}_{T,n}(S)$ su estimador inducido por el protocolo T, diremos que el protocolo es consistente para Q cuando, para los conjuntos evaluados,

$$\widehat{Q}_{T,n}(S) \xrightarrow[n\to\infty]{p} Q(S), \qquad \widehat{\Delta}^{Q}_{T,n,S,j} \xrightarrow[n\to\infty]{p} \Delta^{Q}_{S,j}, \tag{3.5}$$

donde $\Delta^{Q}_{S,j}$ es la brecha poblacional definida por esa misma cantidad. Esta noción cubre protocolos consistentes para accuracy u otras métricas de tarea: permite hablar de utilidad predictiva poblacional bajo una regla de evaluación, aunque no autoriza todavía una lectura como información mutua condicional.

3.3.2. Bayes-consistencia

El segundo nivel, más fuerte y más específico, es la Bayes-consistencia. Aquí la cantidad objetivo no es cualquier métrica de desempeño, sino el riesgo de Bayes asociado a una pérdida ℓ. Sea entonces ℓ una pérdida estrictamente propia y, para un conjunto evaluado $S \subseteq \{1, \ldots, p\}$, sea

$$R^*(S) := \inf_f \mathbb{E}\Big[\ell(f(\boldsymbol{X}_S), Y)\Big] \tag{3.6}$$

el riesgo de Bayes alcanzable usando las covariables de S. La brecha poblacional asociada a retirar X_j desde un conjunto evaluado S con $j \in S$ es

$$\Delta^*_{S,j} := R^*(S \setminus \{j\}) - R^*(S). \tag{3.7}$$

Cuando $S = \{1, \ldots, p\}$, se escribe Δ_j^* como abreviatura de $\Delta_{\{1,\ldots,p\},j}^*$.

Definición 3.1 (protocolo Bayes-consistente). *Sea $\widehat{R}_{T,n}(S)$ el riesgo empírico inducido por el protocolo T al entrenar sobre D_n con las covariables $\boldsymbol{X}_S$ y evaluar bajo una reimplementación $E \sim \Pi$, y sea $\widehat{\Delta}_{T,n,S,j}$ la brecha empírica correspondiente. El protocolo T es Bayes-consistente para la pérdida ℓ y la distribución $P(\boldsymbol{X}, Y)$ si, para todo $S \subseteq \{1, \ldots, p\}$ y todo $j \in S$,*

$$\widehat{R}_{T,n}(S) \xrightarrow[n\to\infty]{p} R^*(S), \qquad \widehat{\Delta}_{T,n,S,j} \xrightarrow[n\to\infty]{p} \Delta_{S,j}^*, \tag{3.8}$$

donde la convergencia en probabilidad se toma sobre la ley conjunta de (D_n, E).

La Bayes-consistencia distingue protocolos que estabilizan efectos dependientes de una muestra realizada de protocolos cuyos riesgos y brechas empíricas convergen hacia sus contrapartes poblacionales óptimas para la pérdida elegida. En la práctica, cuatro condiciones de diseño aproximan esta propiedad:

- Declarar la cantidad poblacional que se pretende aproximar, por ejemplo una brecha de riesgo al retirar o anular una variable.
- Evaluar modelos dentro de una clase suficientemente expresiva para aproximar el predictor de Bayes asociado a la pérdida elegida, o bien reconocer explícitamente el sesgo inducido por una clase restringida.
- Separar la variación propia del proceso generador de la variación inducida por una muestra fija, de manera que la estimación no estabilice una fluctuación accidental como si fuera propiedad poblacional.
- Aplicar el contraste nulo bajo una regla que conserve las propiedades que se desean controlar, como escala y distribución marginal, sin introducir dependencias artificiales que cambien la pregunta inferencial.

El alcance de la definición es observacional: establece una ruta para que una ganancia operacional aproxime una cantidad poblacional interpretable, y deja las afirmaciones causales a cargo de supuestos de identificación o diseños interventivos. Un bootstrap sobre una muestra fija, por ejemplo, puede ser reproducible y, al mismo tiempo, conservar una fluctuación

accidental de esa muestra; una secuencia de realizaciones independientes del proceso generador aproxima otra fuente de variación y ofrece una lectura poblacional más exigente. En métricas como accuracy, la consistencia del protocolo puede ser suficiente para justificar una afirmación de desempeño poblacional; la lectura como informatividad poblacional requiere, en cambio, las condiciones adicionales que conectan la brecha de riesgo con una cantidad informativa. Bajo pérdida logarítmica y variable objetivo discreta, la brecha de Bayes admite una identidad exacta en términos de información mutua condicional.

Proposición 3.2 (brecha de Bayes e información mutua condicional). *Si ℓ es la pérdida logarítmica, Y es discreta, $S \subseteq \{1, \ldots, p\}$ y $j \in S$, entonces*

$$\Delta^*_{S,j} = R^*(S \setminus \{j\}) - R^*(S) = I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}). \tag{3.9}$$

En particular, $\Delta^*_{S,j} > 0$ si y solo si $I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) > 0$. El caso $S = \{1, \ldots, p\}$ recupera la brecha incremental completa $\Delta^*_j = I(Y; X_j \mid \boldsymbol{X}_{\setminus j})$.

Demostración. Bajo pérdida logarítmica, el riesgo de Bayes coincide con la entropía condicional, $R^*(S) = H(Y \mid \boldsymbol{X}_S)$. La identidad estándar $I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) = H(Y \mid \boldsymbol{X}_{S\setminus\{j\}}) - H(Y \mid \boldsymbol{X}_S)$ (Cover and Thomas, 2006) entrega el resultado. □

3.3.3. Identidad poblacional del contraste null-swap

Bajo estas condiciones, la misma identidad poblacional puede formularse mediante el contraste *null-swap*, lo que ofrece una justificación teórica para usarlo como aproximación empírica alternativa al retiro de variable.

Proposición 3.3 (contraste *null-swap* poblacional). *Sea $\widetilde{X}_j$ una copia nula poblacional de X_j, independiente de Y condicional en $\boldsymbol{X}_{S\setminus\{j\}}$, para algún $S \subseteq \{1, \ldots, p\}$ con $j \in S$. Sea $R^*_{\mathrm{ns}}(S, j)$ el riesgo de Bayes alcanzable usando $(\boldsymbol{X}_{S\setminus\{j\}}, \widetilde{X}_j)$. Entonces*

$$\widetilde{\Delta}^*_{S,j} := R^*_{\mathrm{ns}}(S, j) - R^*(S) = \Delta^*_{S,j} = I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}). \tag{3.10}$$

Demostración. La independencia condicional implica $P(Y \mid \boldsymbol{X}_{S\setminus\{j\}}, \widetilde{X}_j) = P(Y \mid \boldsymbol{X}_{S\setminus\{j\}})$, de modo que la copia nula agrega cero información sobre Y dado el contexto y $R^*_{\mathrm{ns}}(S, j) = R^*(S \setminus \{j\})$. Al sustituir en la definición de $\widetilde{\Delta}^*_{S,j}$ se obtiene la identidad. □

Las dos proposiciones establecen que retiro y *null-swap* identifican la misma cantidad poblacional cuando se refieren al mismo conjunto evaluado S; la preferencia empírica por el segundo se debe a que conserva la dimensionalidad y el resto del procedimiento, y a que en datos reales la partición entre variables informativas y nulas se desconoce de antemano. Con estas piezas disponibles, la sección siguiente formaliza la exigencia empírica sobre el contraste y enuncia el criterio central del marco.

3.4. Relevancia epistémica

La sección anterior estableció cuándo un protocolo estima de manera consistente una cantidad poblacional declarada y, en su forma más fuerte, cuándo sus brechas empíricas convergen hacia brechas de riesgo de Bayes. Sobre esa base, el marco integrado define la relevancia epistémica como una condición más exigente que la predictibilidad operacional inmediata. Una variable ofrece evidencia de informatividad poblacional, en un conjunto evaluado S y respecto del contexto $\boldsymbol{X}_{S\setminus\{j\}}$, cuando presenta predictibilidad operacional bajo ese contexto de evaluación, su efecto se establece mediante un protocolo Bayes-consistente orientado a la cantidad poblacional pertinente y, además, el contraste *null-swap* muestra una ventaja estable frente a una copia nula equivalente. Esta sección formaliza la condición restante, la distinguibilidad frente al ruido, y articula ambas exigencias en el criterio central del marco.

La relevancia epistémica se define aquí como una relación entre variable, distribución, protocolo y pregunta inferencial, antes que como una propiedad intrínseca de la variable aislada. Esta formulación permite conservar el valor práctico de las medidas de importancia sin convertirlas automáticamente en explicaciones del fenómeno. En particular, una variable puede ser operacionalmente relevante para un predictor y, al mismo tiempo, requerir cautela antes de ser interpretada como informativa en sentido poblacional.

El punto puede verse en un ejemplo real. En un estudio de predicción de fractura de cadera a partir de radiografías, el modelo aprovechó variables de proceso hospitalario (como el departamento o el equipo de origen de la imagen) que se asociaban con la probabilidad de fractura sin representar un mecanismo biológico directo (Badgeley et al., 2019). Un predictor puede aprovechar esa clase de regularidad y el contraste *null-swap* puede resultar positivo en el contexto evaluado; sin embargo, la interpretación poblacional debe formularse como dependencia observacional bajo un sistema de medición, no como mecanismo clínico directo.

La relevancia epistémica exige entonces preguntar si el protocolo aproxima una cantidad poblacional defendible y si el efecto supera una copia nula comparable.

La condición sobre el contraste admite una formulación probabilística precisa. Bajo protocolos finitos, una variable puede aparentar aporte por el solo hecho de añadir una dimensión, por regularización implícita, por efectos de selección o por interacciones estocásticas; el criterio exige, por eso, una ventaja de magnitud mínima sostenida en la mayoría de las reimplementaciones admisibles.

Conviene recordar la notación que interviene en el criterio. EN este caso el conjunto evaluado S contiene al atributo X_j mediante su posición j, y $S \setminus \{j\}$ representa los atributos de S salvo X_j; D_n es la muestra disponible; y E denota una reimplementación admisible del protocolo, por ejemplo una partición de entrenamiento y prueba, una semilla o una inicialización, muestreada según la distribución Π. La cantidad $\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E)$ es la brecha empírica del contraste *null-swap* en esa reimplementación: mide la diferencia entre la pérdida del modelo nulo que reemplaza la columna observada $\boldsymbol{x}_{\cdot j}$ por su copia nula $\pi_j(\boldsymbol{x}_{\cdot j})$ dentro del subvector $\boldsymbol{X}_S$ y la pérdida del modelo entrenado con el atributo observado, como se definió en (3.4). En esa brecha, $\pi_j(X_j)$ denota la copia nula conceptual del atributo evaluado, construida empíricamente mediante la operación π_j declarada por el protocolo.

Definición 3.2 (distinguibilidad del ruido). *El atributo X_j es distinguible del ruido en el conjunto evaluado S, con $j \in S$, si existe una copia nula $\pi_j(X_j)$ construida bajo el mismo protocolo, que conserva las propiedades marginales que se desean controlar y rompe la asociación con Y, y si existen $\varepsilon > 0$ y $\tau > 1/2$ tales que*

$$\Pr_{E\sim\Pi}\Big[\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E) > \varepsilon \,\Big|\, D_n\Big] > \tau. \tag{3.11}$$

Los dos parámetros del criterio tienen lecturas complementarias: ε fija un margen práctico bajo el cual una diferencia positiva se considera irrelevante, y τ exige que la ventaja aparezca en la mayoría de las reimplementaciones, lo que descarta efectos confinados a particiones o semillas específicas. La Sección 3.5 presenta el instrumento que implementa este contraste en su versión univariada.

Con esta definición queda formalizada la exigencia propia del contraste: el efecto atribuido al atributo observado debe superar, con estabilidad suficiente, una contraparte no informativa (copia nula) construida bajo las mismas condiciones de evaluación. La transición desde $\mathcal{P}$

hacia $\mathcal{I}$ requiere, además, la condición relativa al protocolo que produce esa brecha: la estimación debe orientarse hacia la cantidad poblacional que se desea interpretar, en el sentido de la Bayes-consistencia de la Definición 3.1. Ambas exigencias se articulan a continuación en un criterio único.

Antes de enunciar el criterio, conviene preservar un estatuto intermedio. En muestras finitas, e incluso bajo protocolos diseñados para converger hacia un riesgo de Bayes, un efecto predictivo reproducible puede tener valor instrumental aunque no sea distinguible del ruido mediante *null-swap*; por ejemplo, cuando una perturbación, una dimensión adicional o una forma de resonancia estocástica mejora el desempeño del sistema sin aportar evidencia clara de informatividad poblacional del atributo observado. Si el objetivo declarado es optimizar un predictor bajo las reglas de un protocolo dado, ese efecto puede ser útil en sus propios términos. Su alcance, sin embargo, permanece en $\mathcal{P}$: describe una ventaja operacional explotable, no una razón suficiente para afirmar pertenencia a $\mathcal{I}$ ni para atribuir un mecanismo causal.

Definición 3.3 (relevancia epistémica). *El atributo X_j es epistémicamente relevante bajo el protocolo T y el conjunto evaluado S, con $j \in S$, si, partiendo de una lectura operacional en $\mathcal{P}$ para ese protocolo y ese conjunto, el atributo es distinguible del ruido en ese conjunto (Definición 3.2) y T es Bayes-consistente para la brecha $\Delta^*_{S,j}$ (Definición 3.1).*

El criterio es deliberadamente unilateral y conviene explicitar su lógica. Superar una copia nula no garantiza que incluir la variable mejore el riesgo respecto de excluirla en el mismo conjunto evaluado: una variable puede ser menos dañina que su copia permutada y aun así empeorar el modelo completo. A la inversa, el incumplimiento del criterio deja abierta la pertenencia a $\mathcal{I}$: una dependencia puede permanecer indetectada por expresividad limitada de la clase de modelos, por tamaño muestral insuficiente o por un protocolo subóptimo. La reproducibilidad a través de reimplementaciones admisibles fortalece la lectura evidencial, aunque la estabilidad por sí sola puede reflejar sesgos compartidos del diseño. El aporte del criterio es, precisamente, fijar las condiciones bajo las cuales un efecto predictivo puede leerse como evidencia y no solo como utilidad acotada al protocolo.

En el régimen asintótico, el criterio admite una caracterización exacta que conecta sus dos condiciones con la cantidad poblacional de la Proposición 3.2.

Proposición 3.4 (caracterización asintótica de la relevancia epistémica). *Bajo pér-*

dida logarítmica, variable objetivo discreta, protocolo Bayes-consistente y copia nula poblacional en el sentido de la Proposición 3.3, fijado un conjunto evaluado S con $j \in S$, $\varepsilon > 0$ y $\tau \in (1/2, 1)$: si $I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) > \varepsilon$, el criterio de la Definición 3.3 se satisface para todo n suficientemente grande; si $I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}) < \varepsilon$, el criterio deja de satisfacerse para todo n suficientemente grande.

Demostración. La prueba sigue tres pasos.

(1) Por la Proposición 3.3, la brecha poblacional del contraste *null-swap* coincide con la información mutua condicional:

$$\widetilde{\Delta}^{*}_{S,j} = I(Y; X_j \mid \boldsymbol{X}_{S\setminus\{j\}}).$$

(2) La Bayes-consistencia del protocolo (Definición 3.1) implica que la brecha empírica converge en probabilidad hacia esa cantidad poblacional, sobre la ley conjunta de (D_n, E):

$$\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E) \xrightarrow{p} \widetilde{\Delta}^{*}_{S,j}.$$

(3) Esta convergencia determina el límite de $\Pr[\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E) > \varepsilon]$ según el signo de $\widetilde{\Delta}^{*}_{S,j} - \varepsilon$:

- Si $\widetilde{\Delta}^{*}_{S,j} > \varepsilon$, entonces $\Pr[\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E) > \varepsilon] \to 1 > \tau$, y el criterio de la Definición 3.3 se satisface para n suficientemente grande.
- Si $\widetilde{\Delta}^{*}_{S,j} < \varepsilon$, entonces $\Pr[\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E) > \varepsilon] \to 0 < \tau$, y el criterio deja de satisfacerse para n suficientemente grande.

□

La caracterización precisa el contenido del criterio: bajo las condiciones declaradas, la relevancia epistémica identifica asintóticamente la informatividad poblacional incremental de X_j en el conjunto evaluado S que supera el margen práctico ε. Cuando un protocolo empírico es Bayes-consistente y el contraste *null-swap* permanece positivo frente a una copia nula equivalente, el efecto observado constituye evidencia de informatividad poblacional incremental; la afirmación permanece en el plano observacional, y una lectura causal requiere supuestos de identificación o intervención.

La misma caracterización aclara por qué el tamaño muestral, por sí solo, no resuelve el problema: una cantidad creciente de datos ayuda cuando el protocolo converge hacia la cantidad poblacional pertinente, mientras que, si el protocolo conserva o amplifica regularidades

de una muestra realizada, el aumento de n puede estabilizar una cantidad equivocada. La convergencia de la Definición 3.1 es una propiedad del diseño, no un beneficio automático de acumular observaciones. Por ello, la tesis trata el diseño del protocolo como parte de la evidencia y no como un detalle técnico secundario; el Capítulo 4 muestra este contraste con protocolos que, ante variables nulas por construcción, se comportan de manera cualitativamente distinta al crecer la muestra.

3.5. Attribute Relevance Score

Enunciado el criterio, esta sección presenta el instrumento que implementa su condición de distinguibilidad del ruido. El *Attribute Relevance Score* (ARS) se ubica dentro de $\mathcal{P}$ como una medida operacional basada en el contraste *null-swap*. Su función es estimar, bajo un modelo y un protocolo declarados, si un atributo observado produce una ventaja de desempeño frente a una copia nula comparable.

Sea $\boldsymbol{X} = (X_1, \ldots, X_p)$ un vector aleatorio de atributos, asociado a una variable objetivo Y. Para evaluar un atributo X_j, ARS compara dos modelos entrenados bajo el mismo protocolo. En coherencia con el contraste *null-swap*, la formulación univariada corresponde al caso $S = \{j\}$; por abreviatura, escribimos $M_{E,j} := M_E^{(\{j\})}$ para el modelo entrenado con el atributo observado y $M_{E,j}^{\pi} := M_E^{(\{j\}, j \leftarrow \pi_j)}$ para el modelo nulo entrenado con una versión permutada del mismo atributo. Dado que la permutación preserva la escala y la distribución marginal del atributo, pero rompe su asociación con la respuesta, la diferencia entre ambos desempeños puede evaluarse sobre múltiples particiones o reimplementaciones como una ventaja operacional frente a una copia nula.

La medida se apoya en una intuición metodológica simple, aunque exigente: para recibir un puntaje positivo en ARS, un atributo debe mostrar una ventaja consistente durante múltiples iteraciones frente a una copia nula evaluada con el mismo modelo base, la misma métrica y el mismo protocolo. Esta condición compara el atributo con una versión que conserva propiedades marginales relevantes y sitúa la operación dentro del contraste *null-swap*, cuyo vocabulario conecta la evaluación metodológica con la pregunta por informatividad poblacional presentada en las secciones anteriores.

Cuando la métrica F se interpreta como "mientras mayor, mejor" (ej Accuracy, Recall, F_1, etc..) y está normalizada en $[0, 1]$, sea $F_E(M; D_n)$ su valor bajo la reimplementación E

y la muestra disponible. Aquí $\mathrm{med}_{E\sim\Pi}(\cdot)$ denota la mediana muestral de esa cantidad sobre el conjunto de reimplementaciones evaluadas, elegida por su robustez frente a repeticiones atípicas. En $\widetilde{F}_j$ y $\widetilde{F}_j^{\pi}$, la virgulilla marca precisamente ese agregado por mediana entre reimplementaciones, un uso distinto del que tiene sobre $\widetilde{\Delta}^{\mathrm{ns}}_{n,S,j}(E)$ en la Sección 3.1.4, donde solo identifica la familia *null-swap* sin implicar agregación. Definimos

$$\widetilde{F}_j = \mathrm{med}_{E\sim\Pi}\, F_E(M_{E,j}; D_n), \qquad \widetilde{F}_j^{\pi} = \mathrm{med}_{E\sim\Pi}\, F_E(M^{\pi}_{E,j}; D_n).$$

Entonces ARS puede escribirse como:

$$\mathrm{ARS}_F(X_j) = \mathrm{máx}\left\{0,\ \frac{\widetilde{F}_j - \left(\widetilde{F}_j^{\pi} + \varepsilon\right)}{1 - \left(\widetilde{F}_j^{\pi} + \varepsilon\right)}\right\}, \tag{3.12}$$

donde ε introduce un margen mínimo de efecto práctico. Cuando la diferencia entre el modelo original y el modelo nulo con versión permutada carece de respaldo estadístico bajo el nivel de significancia elegido, el valor se fija en cero, decisión que evita leer fluctuaciones muestrales como relevancia.

La presencia de ε cumple una función que conviene distinguir de la significancia estadística, porque un contraste puede resultar estadísticamente detectable y, aun así, tener una magnitud práctica reducida. El margen mínimo introduce entonces una exigencia adicional: la diferencia debe ser suficientemente grande respecto de la línea de base nula. De este modo, ARS combina tres criterios: dirección del efecto, respaldo estadístico y tamaño mínimo de la diferencia operacional.

La implementación experimental utilizada para esta tesis evalúa cada atributo mediante repeticiones independientes del mismo protocolo: en cada una se entrena un modelo con el atributo observado y un modelo nulo con una versión permutada del mismo atributo, y ambos desempeños se contrastan mediante una prueba pareada de Wilcoxon unilateral, de modo que el valor final se fija en cero cuando la ventaja del atributo observado frente a su copia nula carece de respaldo estadístico al nivel predefinido. La Tabla 4.2, en el Capítulo 4, detalla la partición, los modelos base, las métricas y el nivel de significancia empleados en esa implementación. En esta implementación, el margen práctico se calcula como

$$\varepsilon = 0.2\,\bar{\Delta}, \tag{3.13}$$

donde $\bar{\Delta}$ es la media de las diferencias pareadas de desempeño entre el atributo observado y su copia nula a través de las repeticiones. El factor 0.2 se inspira en el tamaño de efecto estandarizado de Cohen para diseños pareados, $d_z = \bar{\Delta}/s_\Delta$, donde s_Δ es la desviación estándar de esas mismas diferencias y $d_z = 0.2$ es la referencia convencional para un efecto pequeño. Conviene precisar que la Ecuación (3.13) no impone esa condición estandarizada de manera directa: en vez de exigir $d_z \geq 0.2$ sobre la escala de s_Δ, la implementación aplica el mismo valor numérico como fracción de $\bar{\Delta}$, en la escala original de la métrica. La referencia a $d_z = 0.2$ cumple así una función de calibración, pues toma prestada la convención de efecto pequeño para fijar la magnitud del umbral, pero ε y d_z no son la misma cantidad ni se derivan algebraicamente una de la otra. Se prefiere específicamente el extremo pequeño de la convención de Cohen porque el margen debe ser conservador: alcanza para descartar mejoras negligibles, sin exigir de entrada un efecto grande que dejaría fuera atributos con una ventaja real pero modesta. El margen resultante opera como una exigencia mínima adicional de tamaño de efecto, distinta de la significancia estadística que aporta la prueba de Wilcoxon; junto con la partición 0.6/0.4 y el nivel $\alpha = 0.05$, se trata de una constante operativa fija, cuya sensibilidad sistemática se reconoce como limitación pendiente de caracterizar en la Sección 4.6.

Así construido, ARS implementa la condición de distinguibilidad del ruido (Definición 3.2) en el caso univariado $S = \{j\}$: la mediana sobre repeticiones estima la tendencia central del desempeño, la prueba pareada aporta el respaldo estadístico sobre las reimplementaciones y el margen ε cumple el papel del umbral práctico.

Un ejemplo simple permite fijar la lectura. Si un árbol de clasificación entrenado con X_j obtiene un F_1 mediano de 0.82 y el mismo árbol entrenado con la copia permutada obtiene 0.76, la diferencia sugiere que el atributo observado contiene una asociación aprovechable por ese modelo y ese protocolo. Si, en cambio, ambos valores quedan prácticamente empatados, o la ventaja aparece solo en unas pocas particiones, ARS fija el resultado en cero o lo reduce por debajo del margen práctico. En ambos casos, el resultado ofrece, en primer término, una lectura de utilidad predictiva bajo el protocolo.

Para métricas de pérdida, donde valores menores indican mejor desempeño, el contraste se invierte:

$$\widetilde{L}_j = \mathrm{med}_{E\sim\Pi}\, L_E(M_{E,j}; D_n), \qquad \widetilde{L}_j^{\pi} = \mathrm{med}_{E\sim\Pi}\, L_E(M_{E,j}^{\pi}; D_n).$$

$$\mathrm{ARS}_\ell(X_j) = \max\left\{0,\ \frac{\left(\widetilde{L}_j^{\pi} - \varepsilon\right) - \widetilde{L}_j}{\widetilde{L}_j^{\pi} - \varepsilon}\right\}. \tag{3.14}$$

La medida evalúa atributos de manera univariada en la formulación desarrollada aquí. Esa decisión facilita su interpretación y paralelización, aunque puede subestimar relaciones puramente sinergéticas; por ello, la limitación es metodológicamente relevante: una medida univariada puede ser adecuada para ciertos análisis exploratorios y, a la vez, insuficiente para fenómenos donde la señal emerge solo en combinaciones de variables.

La interpretación de ARS queda acotada por el diseño que lo produce: un valor positivo indica que, bajo un modelo, una métrica, una estrategia de partición y un número de repeticiones, el atributo original produjo un desempeño superior al de su copia nula, con respaldo estadístico y magnitud práctica mínima. Esa afirmación pertenece primero al espacio $\mathcal{P}$ de predictibilidad operacional; su lectura como evidencia de informatividad poblacional requiere las condiciones de la Definición 3.3. En sentido inverso, un valor nulo de ARS no agota las formas posibles de pertenencia a $\mathcal{P}$: una variable puede ser usada por un predictor, aparecer en una atribución post-hoc o intervenir en una regla de selección interna sin superar, bajo este diseño particular, el contraste frente a una copia nula.

El criterio de la Definición 3.3 añade a la distinguibilidad la Bayes-consistencia del diseño (Definición 3.1), una propiedad del protocolo y de la clase de modelos que debe establecerse por separado. La implementación experimental descrita, con modelos base de expresividad acotada y particiones repetidas de una muestra fija, entrega una lectura operacional bajo el diseño declarado; leer sus puntajes como evidencia de informatividad poblacional requiere verificar, además, que el protocolo converja hacia la brecha de riesgo pertinente. El Capítulo 4 examina de manera empírica esa dependencia del protocolo.

3.6. Lectura de métodos post-hoc

Los métodos post-hoc, entre ellos SHAP (*SHapley Additive exPlanations*), se ubican primariamente en $\mathcal{P}$. En esta tesis, el término designa explicaciones calculadas después del entrenamiento para describir cómo el modelo usa las variables bajo una representación dada, una fidelidad al predictor que resulta valiosa para auditoría, depuración, comunicación técnica y comparación entre modelos. El límite aparece cuando una explicación del predictor se

presenta como explicación del fenómeno.

La Proposición 3.1 muestra que ese límite es estructural y no una cuestión de calidad de la herramienta: la fidelidad al predictor es una propiedad en $\mathcal{P}$, y una explicación acotada a $\mathcal{P}$ se convierte en evidencia sobre $\mathcal{I}$ o $\mathcal{C}_Y$ solo mediante las condiciones de las secciones anteriores o mediante supuestos causales explícitos. Si el modelo aprende una representación confundida, una atribución fiel describirá con exactitud esa dependencia confundida; la exactitud descriptiva y el alcance inferencial son propiedades independientes.

Esta lectura conserva el valor de los métodos post-hoc y los sitúa en el nivel que pueden defender con mayor claridad. Una atribución puede ser útil para identificar dependencia de una variable sensible, detectar atajos de medición, comparar modelos o examinar cambios de representación; su alcance cambia cuando se la usa para sostener que una variable es causa del resultado o que describe el mecanismo generador de datos. En ese paso se requieren argumentos que exceden la fidelidad al predictor.

Para reportar estas atribuciones de manera verificable, conviene declarar el protocolo que produjo el modelo y el contraste *null-swap* usado para distinguir señal de estructura espuria, siguiendo el procedimiento que la Sección 3.7 formaliza a continuación. Si el objetivo es predecir, una atribución estable puede ser suficiente; si el objetivo es sostener una afirmación sobre informatividad poblacional, corresponde exigir las condiciones de la Definición 3.3; y, si el objetivo es intervenir, se requieren argumentos causales adicionales.

3.7. Procedimiento integrado de análisis

La tesis propone un procedimiento de lectura en cuatro pasos:

1. Definir el tipo de afirmación: operacional, de informatividad poblacional o causal.

2. Documentar la predictibilidad operacional bajo un protocolo explícito, declarando modelo, métrica, particiones, tamaño muestral y reimplementaciones, evaluando la consistencia respecto de la métrica declarada y, cuando se pretenda una lectura sobre $\mathcal{I}$, la Bayes-consistencia del diseño.

3. Contrastar cada efecto mediante un contraste *null-swap*, declarando el conjunto evaluado S y el contexto $\boldsymbol{X}_{S\setminus\{j\}}$, usando versiones permutadas del atributo, e informar

estabilidad y tamaño de efecto.

4. Reservar las afirmaciones causales para diseños o supuestos que permitan pasar desde dependencia observacional hacia intervención.

El segundo paso fija la pertenencia a $\mathcal{P}$ bajo el protocolo declarado y, cuando corresponde, examina la orientación poblacional del protocolo para la métrica declarada; el tercero agrega la distinguibilidad frente a una copia nula comparable. Cuando la orientación poblacional adopta la forma más fuerte de Bayes-consistencia, ambos pasos corresponden al criterio de relevancia epistémica (Definición 3.3). El cuarto delimita el paso adicional hacia $\mathcal{C}_Y$, que ninguna combinación de los anteriores autoriza por sí sola, como establece la Proposición 3.1. Este procedimiento no busca rigidizar la práctica exploratoria; su propósito es ordenar el lenguaje de reporte para que los resultados predictivos conserven su valor sin asumir, por inercia, un alcance explicativo mayor que el que el protocolo permite defender.

En términos de escritura científica, el procedimiento funciona como una lista de control conceptual: antes de afirmar que una variable es relevante, conviene precisar si se habla de relevancia operacional, informatividad poblacional o causalidad; antes de comparar importancias, conviene declarar el conjunto evaluado S, el contexto $\boldsymbol{X}_{S\setminus\{j\}}$, la copia nula usada, la métrica y las repeticiones; y, antes de extraer una interpretación mecanística, conviene explicitar el diseño causal o las hipótesis de identificación. Esta práctica de reporte es uno de los aportes integradores de la tesis.

Capítulo 4

Validación experimental y discusión integrada

4.1. Criterio de integración experimental

Los experimentos pueden leerse como una secuencia de pruebas sobre una misma dificultad: distinguir cuándo una diferencia observada en el desempeño de un modelo puede tratarse como señal, cuándo debe entenderse como efecto del protocolo y cuándo resulta insuficiente para sostener una lectura causal. La tesis organiza esa secuencia en dos bloques. El primero evalúa ARS como medida operacional de relevancia. El segundo estudia las fronteras entre predictibilidad operacional, informatividad poblacional y causalidad.

Esta organización permite desarrollar una validación progresiva. Los primeros experimentos muestran cómo se comporta una medida basada en contraste *null-swap* frente a relaciones lineales, no lineales, ruido y variables no informativas. Los experimentos posteriores examinan qué ocurre cuando la estimación de esos contrastes depende del protocolo o cuando la importancia del modelo entra en tensión con la estructura causal del proceso generador. La Tabla 4.1 anticipa esa continuidad; la definición operativa de bootstrap, validación cruzada y realizaciones independientes se entrega en la Sección 4.4.2, donde esos protocolos se comparan directamente.

Tabla 4.1: Validación experimental: lectura unificada de los experimentos de validación.

Bloque	**Pregunta experimental**	**Lectura dentro de la tesis**
ARS, relaciones sintéticas	Cómo responde la medida ante relaciones lineales, rotadas, no lineales y estructuras sin señal clara.	Evalúa sensibilidad operacional y comportamiento frente a variables que otras medidas pueden puntuar de manera positiva.
ARS, conjunto de referencia	Cómo cambia la relevancia estimada con modelos distintos, ruido estructurado y variables no informativas.	Muestra que ARS depende del modelo base y que su contraste tiende a ser conservador sobre variables nulas.
Fronteras, protocolo	Cómo se comporta el contraste *null-swap* bajo bootstrap, validación cruzada y realizaciones independientes.	Señala que el protocolo forma parte de la evidencia y que el tamaño muestral no corrige por sí solo una estimación orientada a una cantidad inadecuada.
Fronteras, disociación epistémica	Cómo aparecen causalidad, informatividad poblacional, predictibilidad operacional y atribuciones SHAP bajo un SCM conocido.	Ilustra que una variable puede ser importante para el modelo sin ser causal, y que una causa puede tener baja prominencia observacional.

Los diseños sintéticos permiten observar el comportamiento de las medidas cuando se conoce, por construcción, el papel de cada variable: informativa, nula, confundida, causal o sinergética. El primer grupo de diseños evalúa ARS en relaciones bivariadas lineales, rotadas, no lineales y visualmente engañosas; su papel es calibrar la medida frente a casos simples y frente a relaciones que las correlaciones clásicas suelen subrepresentar. El segundo grupo utiliza una función de referencia con tres variables informativas y tres variables nulas, además de versiones perturbadas por ruido, para estudiar sensibilidad al modelo base y conservadurismo frente a distractores. El tercer diseño examina la dependencia del protocolo mediante una clasificación binaria donde X_1 es informativa y Z_1, Z_2, Z_3 son nulas; allí se

comparan bootstrap, validación cruzada y realizaciones independientes del proceso generador con distintos tamaños muestrales y presupuestos de evaluación. El cuarto diseño construye una disociación epistémica completa: una variable causal con señal observacional débil, una variable confundida sin efecto causal directo, dos variables sinergéticas de tipo XOR y tres variables de ruido.

Cada diseño introduce un tipo de tensión inferencial: sensibilidad a no linealidad, abstención frente a nulos, dependencia del protocolo, confusión, causalidad débil o sinergia. Así, la validación experimental funciona como una serie de pruebas sobre el alcance de la evidencia, además de una comparación de puntajes.

4.1.1. Un contraejemplo real: Pima Indians Diabetes

La Sección 2.10 argumentó que los datos reales rara vez permiten decidir si una diferencia de desempeño refleja señal, redundancia, confusión o un artefacto de protocolo, porque la función generadora y el papel verdadero de cada variable permanecen desconocidos. Antes de presentar los diseños sintéticos que estructuran los Bloques I y II, conviene mostrar esa dificultad en un ejercicio aplicado sobre datos reales, ya con el contraste *null-swap* y los tres espacios epistémicos formalizados en el Capítulo 3.

La elección de Pima Indians Diabetes responde a su utilidad como ejemplo metodológico reproducible. Se trata de un conjunto público, compacto, con respuesta binaria y atributos clínicos heterogéneos. Estas características permiten visualizar contrastes contextuales por pares mediante un procedimiento acotado. Su amplia circulación también facilita reconocer una dificultad frecuente: el conjunto informa sobre las variables registradas, mientras el proceso generador, las relaciones causales y la relevancia verdadera requieren conocimiento adicional. Por ello, el ejercicio cumple una función ilustrativa y metodológica.

El material reproducible trata como faltantes los ceros fuera del rango clínico esperable en glucosa, presión arterial, grosor del pliegue cutáneo, insulina e índice de masa corporal. La imputación usa la mediana estimada dentro de cada partición de entrenamiento, de modo que el preprocesamiento emplea exclusivamente información de entrenamiento. Esta decisión añade una dependencia respecto del protocolo. El tamaño reducido de la muestra, la especificidad de la cohorte y la multiplicidad derivada de examinar pares refuerzan el alcance ilustrativo del ejercicio.

La Figura 4.1 muestra este ejercicio sobre el conjunto Pima Indians Diabetes, disponible en Kaggle (Mathchi, 2020) y asociado históricamente al estudio de Smith et al. (1988). El ejercicio calcula un contraste *null-swap* direccional de segundo orden usando *accuracy*: en cada celda se evalúa la variable de la columna en presencia de la variable de la fila, que permanece intacta como contexto. Solo la variable de la columna se reemplaza por una copia permutada. Un valor positivo indica que, bajo ese contexto, el modelo obtuvo mayor *accuracy* con la variable real que con su versión nula.

El interés de la figura reside precisamente en que produce una lectura plausible y, al mismo tiempo, insuficiente. El mapa sugiere que la glucosa conserva señal predictiva en presencia de distintos atributos acompañantes y que algunos atributos parecen aportar más cuando se observan junto con otros; sin embargo, el conjunto real no permite decidir si esos patrones corresponden a sinergias clínicas, redundancias parciales, proxies, sesgos de medición, prácticas de tratamiento o regularidades propias de la cohorte. El ejemplo funciona, así, como contraejemplo metodológico: los datos reales pueden orientar hipótesis y comunicar patrones operacionales, pero no bastan para validar una medida de relevancia cuando la verdad de fondo permanece desconocida. Esa limitación es precisamente la que motiva los procesos generadores controlados descritos en la Sección 2.10 y los diseños sintéticos de los Bloques I y II que siguen.

4.2. Entorno computacional y reproducibilidad de métodos

Los experimentos se implementaron en el ecosistema científico de Python, apoyándose principalmente en *scikit-learn* para los modelos base, las particiones y las métricas de evaluación (Pedregosa et al., 2011). Esta elección favorece la comparabilidad y la replicación, dado que las implementaciones de referencia son de código abierto y de uso extendido en la comunidad. La aleatoriedad de particiones, permutaciones y remuestreos se controla mediante semillas asociadas a cada iteración, de modo que cada repetición del protocolo queda determinada por su configuración.

La trazabilidad se organiza por bloque. El Bloque I se apoya en el artículo de ARS y en su repositorio público, que contiene la implementación de la medida, los datos sintéticos

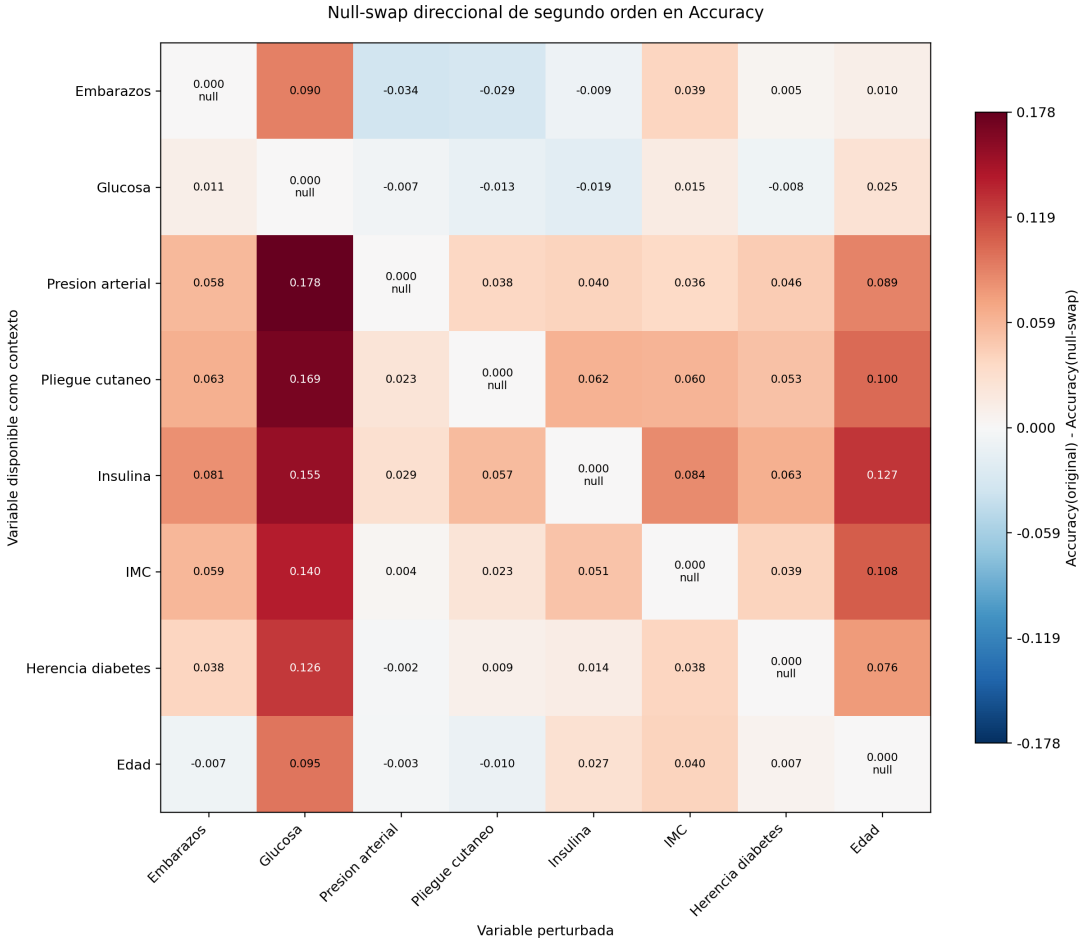


Figura 4.1: Diabetes Pima: contraste *null-swap* direccional de segundo orden medido como diferencia de *accuracy* entre el modelo con la variable real y el modelo con la variable reemplazada por una versión permutada de referencia. Un valor positivo indica que el desempeño con la variable real fue mayor que con la copia nula. Las filas indican la variable disponible como contexto y las columnas la variable perturbada. La diagonal se fija en cero como referencia nula del protocolo.

derivados y los cuadernos de ejemplo; el módulo principal utiliza *scikit-learn*, NumPy, pandas, SciPy, joblib y tqdm, con particiones controladas por semilla. El Bloque II se apoya en el repositorio de experimentos de fronteras epistémicas; según la especificación publicada, esos experimentos fueron ejecutados en Python 3.11 con *scikit-learn* 1.5.2, SHAP 0.46.0 y NumPy 2.1.3. El Apéndice A resume los repositorios, materiales derivados y alcance de cada recurso.

Esta disposición sigue la distinción entre reproducibilidad de métodos, de resultados y de inferencias discutida en la Sección 2.13. Los repositorios permiten inspeccionar y reejecutar procedimientos; la estabilidad de las conclusiones, en cambio, se examina dentro del propio diseño experimental mediante repeticiones, protocolos alternativos y nulos conocidos. Por ello, la reproducibilidad forma parte de la evidencia sobre el alcance de cada afirmación, además de la disponibilidad del código.

4.3. Bloque I: comportamiento operacional de ARS

Los experimentos de este bloque comparten el protocolo descrito en la Sección 3.5; la Tabla 4.2 resume su configuración para facilitar la lectura y la replicación de los resultados. La comparación del Bloque I se orienta a una pregunta acotada: qué comportamientos de ARS aparecen frente a medidas de dependencia que pueden subrepresentar relaciones no lineales o atribuir valores positivos a configuraciones sin señal clara. La comparación con métodos más próximos por familia metodológica, como la importancia por permutación y Boruta, se retoma en la Tabla 4.8, donde esas medidas se examinan bajo un proceso generador conocido y junto al contraste *null-swap* por pares.

Tabla 4.2: Protocolo ARS: configuración del protocolo experimental de ARS.

Elemento	**Configuración**
Partición	Entrenamiento/prueba con proporción 0.6/0.4 y semilla controlada por iteración.
Repeticiones y tamaño muestral	Evaluación repetida del mismo protocolo sobre particiones independientes; en la implementación del repositorio, ARS usa por defecto 1000 iteraciones. Las relaciones bivariadas se generan con $n = 1000$ observaciones por conjunto; el conjunto de referencia con ruido estructurado usa 100 repeticiones de 100 observaciones.
Modelos base (regresión)	Regresión lineal; árbol de decisión con criterio de error absoluto; vecinos cercanos con $k = 5$.
Modelos base (clasificación)	Árbol de clasificación con criterio de entropía; vecinos cercanos.
Métricas	Error absoluto medio (regresión); F_1 (clasificación), entendido como media armónica entre precisión y exhaustividad.
Contraste estadístico	Prueba pareada de Wilcoxon unilateral con $\alpha = 0.05$; sin respaldo estadístico al nivel declarado, ARS se fija en cero.
Margen práctico	Si Δ_r denota la mejora pareada de desempeño en la repetición r, con signo definido según la orientación de la métrica, la implementación del repositorio calcula ε según la Ecuación (3.13), como 0.2 veces la media $\bar{\Delta}$ de esas diferencias. Este margen desplaza el punto de referencia del contraste nulo antes de calcular el puntaje final.

4.3.1. Relaciones sintéticas bivariadas

El primer experimento de ARS utiliza veintiún relaciones bivariadas sintéticas, todas con $n = 1000$ observaciones y escalamiento MinMax al intervalo $[-5, 5]$. La Figura 4.2 funciona como mapa visual del diseño: los paneles A–G cubren distribuciones normales multivariadas con distintos niveles y signos de correlación; los paneles H–N muestran relaciones lineales bajo cambios de orientación; y los paneles O–U reúnen patrones no lineales, geométricamente complejos o visualmente estructurados. Para convertir cada patrón en un problema predictivo, el protocolo trata una coordenada como atributo y la otra como respuesta. ARS evalúa, en ese marco, si el atributo observado aporta más desempeño predictivo que una copia permutada evaluada bajo el mismo modelo y la misma partición.

La selección de estas distribuciones cumple una función diagnóstica. El primer grupo calibra las medidas frente a asociaciones lineales con distinta intensidad y signo; el segundo permite examinar cuánto cambia su lectura ante transformaciones geométricas de una estructura relacionada; y el tercero introduce dependencias curvas, multimodales o cerradas que tensionan los supuestos de linealidad y monotonía. En conjunto, los veintiún escenarios forman un banco de pruebas heredado del estudio de ARS y permiten comparar sensibilidad y abstención. Cada aplicación requerirá una distribución acorde con su dominio.

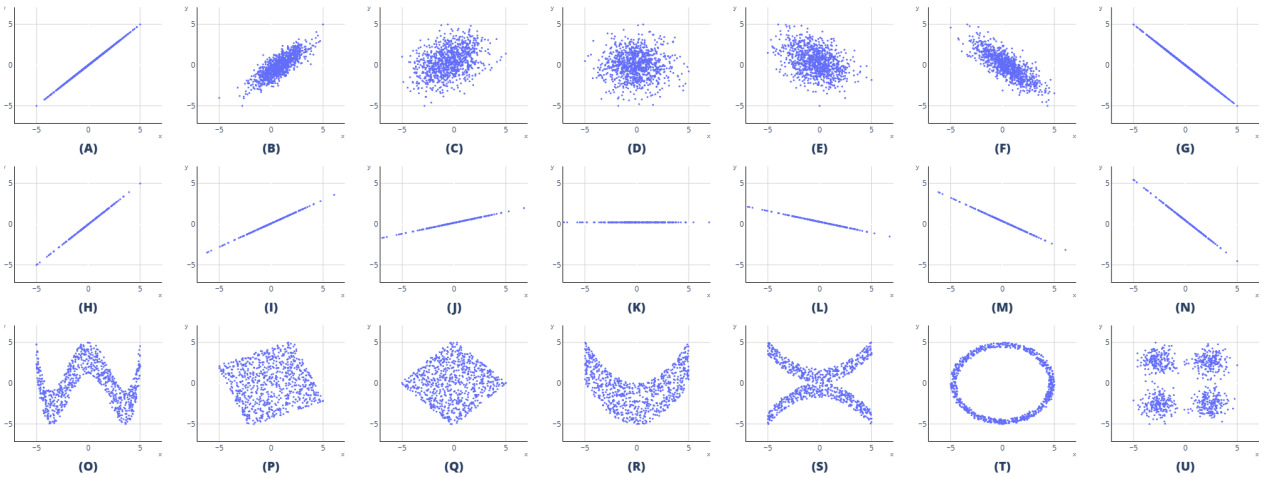


Figura 4.2: Relaciones bivariadas sintéticas usadas para evaluar ARS. Los paneles incluyen distribuciones normales con distintos niveles de correlación, versiones rotadas y patrones no lineales o visualmente engañosos. Figura reproducida de Neirz et al. (2024).

El diseño permite observar tres comportamientos de interés. Primero, calibra la medida en relaciones lineales donde la señal direccional es clara. Segundo, examina si esa lectura se mantiene cuando la misma estructura aparece bajo distintas orientaciones geométricas. Tercero, explora patrones no lineales o visualmente estructurados en los que la predictibilidad operacional puede diferir de la dependencia capturada por otras medidas. La Tabla 4.3 resume las medidas usadas como comparación. Cuatro de ellas, el Coeficiente de Información Maximal (MIC, *Maximal Information Coefficient*), la Puntuación de Asimetría Máxima (MAS, *Maximum Asymmetry Score*), el Valor de Borde Máximo (MEV, *Maximum Edge Value*) y el Coeficiente de Información Media Generalizado (GMIC, *Generalized Mean Information Coefficient*), pertenecen a la familia MINE (*Maximal Information-based Nonparametric Exploration*): exploran particiones de grilla para detectar asociaciones no lineales, pero enfatizan aspectos distintos de esa estructura.

Tabla 4.3: Medidas de dependencia comparadas con ARS en relaciones bivariadas sintéticas.

Medida	**Qué captura**
Pearson	Asociación lineal.
Spearman	Asociación monótona mediante rangos.
Kendall	Asociación monótona mediante concordancia de pares.
Información mutua	Dependencia estadística general, como reducción de incertidumbre.
MIC	Asociación fuerte a través de múltiples resoluciones de grilla.
MAS	Asimetría entre orientaciones de la relación.
MEV	Dependencias localizadas en bordes o regiones específicas de la grilla.
GMIC	Generalización de MIC mediante una media parametrizada sobre la matriz característica, que modula la sensibilidad de la medida en muestras finitas.
ϕ_k	Correlación interpretable para variables categóricas, ordinales, continuas o mixtas, a partir de una tabla de contingencia y una transformación del estadístico χ^2 hacia una escala comparable entre 0 y 1.

Con esa referencia visual, la Tabla 4.4 conserva la misma codificación A–U de la Figura 4.2. La lectura se organiza por bloques: A–G muestran sensibilidad a la fuerza y signo de la correlación; H–N permiten observar estabilidad frente a cambios de orientación; y O–U muestran cómo responde ARS ante formas no lineales, patrones cerrados, cruces, parábolas,

conglomerados y otras geometrías bivariadas.

Tabla 4.4: Comparación de ARS con medidas de dependencia en relaciones bivariadas sintéticas, agrupadas por tipo de relación. Los conjuntos A–G corresponden a normales multivariadas con distintos niveles de correlación; H–N, a normales rotadas; y O–U, a patrones no lineales o visualmente complejos. Valores reproducidos de Neirz et al. (2024).

Tipo	Conjunto	Pearson	Spearman	Kendall	ϕ_k	MI	MIC	MAS	MEV	GMIC	ARS
Normal	A	1.000	1.000	1.000	1.000	5.984	1.000	0.000	1.000	1.000	0.996
Normal	B	0.805	0.783	0.587	0.859	0.575	0.519	0.030	0.519	0.461	0.429
Normal	C	0.378	0.378	0.258	0.373	0.081	0.200	0.017	0.200	0.148	0.049
Normal	D	0.009	0.016	0.010	0.157	0.007	0.126	0.007	0.126	0.054	0.000
Normal	E	0.394	0.383	0.261	0.407	0.054	0.215	0.015	0.215	0.158	0.089
Normal	F	0.809	0.800	0.606	0.817	0.508	0.512	0.015	0.512	0.465	0.394
Normal	G	1.000	1.000	1.000	1.000	5.984	1.000	0.000	1.000	1.000	0.996
Rotada	H	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
Rotada	I	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
Rotada	J	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
Rotada	K	0.606	0.624	0.483	0.784	0.697	0.498	0.038	0.498	0.412	0.000
Rotada	L	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
Rotada	M	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
Rotada	N	1.000	1.000	1.000	1.000	4.373	1.000	0.000	1.000	1.000	0.980
No lineal	O	0.052	0.046	0.033	0.732	0.733	0.773	0.626	0.773	0.670	0.477
No lineal	P	0.014	0.014	0.006	0.525	0.240	0.195	0.026	0.195	0.136	0.046
No lineal	Q	0.036	0.044	0.023	0.526	0.315	0.163	0.025	0.138	0.121	0.024
No lineal	R	0.028	0.028	0.018	0.688	0.534	0.416	0.245	0.416	0.382	0.297
No lineal	S	0.020	0.021	0.018	0.783	0.902	0.450	0.037	0.262	0.208	0.135
No lineal	T	0.025	0.020	0.003	0.836	1.421	0.566	0.079	0.376	0.443	0.215
No lineal	U	0.013	0.025	0.017	0.000	0.012	0.134	0.009	0.134	0.065	0.000

En los paneles A–G, ARS sigue el patrón esperable para relaciones lineales: alcanza valores cercanos a uno en las correlaciones perfectas positiva y negativa (A y G), disminuye en las correlaciones intermedias (B, C, E y F) y se anula en el caso sin relación clara (D). Esta parte del experimento calibra la medida frente a un escenario simple: cuando la estructura bivariada ofrece señal predictiva aprovechable, el atributo observado supera a su copia permutada.

Los paneles H–N agregan una prueba geométrica. En las relaciones lineales reorientadas con señal direccional clara, ARS conserva valores altos, lo que sugiere que el contraste responde a la predictibilidad disponible más que a una orientación visual específica. El panel K ofrece un caso especialmente informativo: varias medidas de dependencia conservan valores

positivos, mientras ARS reporta 0.000. La diferencia precisa la naturaleza operacional del contraste: ARS estima una ganancia de desempeño frente a una copia nula bajo el modelo y el protocolo declarados.

La lectura de los paneles O–T amplía el diagnóstico hacia patrones no lineales o geométricamente complejos. En el caso cuadrático con ruido (O), ARS obtuvo 0.477 mientras Pearson y Spearman se mantuvieron cerca de 0.05; también asignó valores positivos en el patrón parabólico (R) y en el patrón trigonométrico (T), con valores aproximados de 0.297 y 0.215. Estos resultados sugieren que la medida puede captar formas de predictibilidad operacional que las correlaciones lineales o monótonas tienden a subrepresentar.

Los casos D, K y U completan la lectura porque muestran abstención bajo configuraciones distintas. En D, la nube carece de una dirección predictiva clara; en K, la forma rotada produce valores positivos en varias medidas, pero el contraste no respalda una ganancia operacional; y en U, los conglomerados son visualmente nítidos, aunque aportan poca capacidad para predecir una coordenada a partir de la otra bajo el protocolo definido. En conjunto, estos casos muestran el rasgo conservador de ARS: el puntaje se vuelve positivo cuando la variable observada supera a su copia nula en desempeño, y permanece en cero cuando esa ventaja no se estabiliza bajo el criterio estadístico y práctico declarado.

4.3.2. Conjunto de referencia sin ruido

El segundo experimento de ARS trabaja con un conjunto de referencia construido a partir de una función generadora con tres variables informativas y tres variables no informativas. Las variables $X_1, \ldots, X_6$ se generan independientes con distribución uniforme en $[0, 1]$; las tres primeras contribuyen a la respuesta y las restantes actúan como distractores nulos. El proceso generador se define como:

$$Y = 0.25 \exp(4X_1) + \frac{4}{1 + \exp[-20(X_2 - 0.5)]} + 3X_3 + \eta, \qquad \eta \sim \mathcal{N}(0, 0.2^2). \tag{4.1}$$

Así, X_1 introduce una relación exponencial dominante, X_2 una relación logística acotada y X_3 una contribución lineal más moderada, mientras que X_4, X_5 y X_6 no contribuyen directamente a Y.

La versión sin término de ruido permite estudiar el comportamiento de ARS bajo condiciones más controladas, comparando tres modelos base: árbol de decisión, regresión lineal y vecinos cercanos con $k = 5$. Este punto importa porque ARS mide una relación entre atributo, modelo, métrica y protocolo; la comparación entre modelos base vuelve visible esa dependencia.

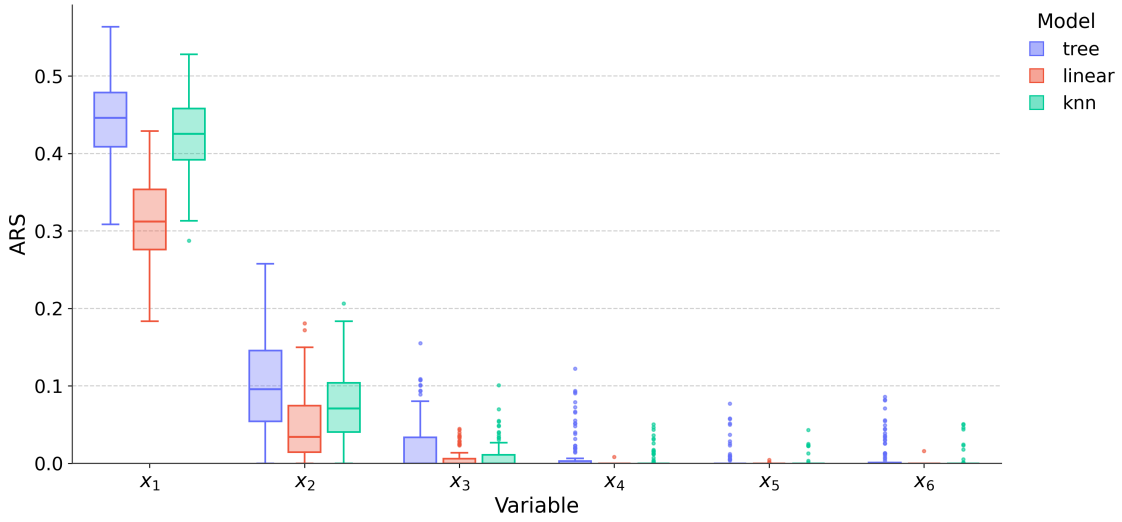


Figura 4.3: ARS bajo distintos modelos base en el conjunto de referencia sin ruido. La figura compara variables informativas y no informativas para árbol de decisión, regresión lineal y vecinos cercanos. Figura reconstruida a partir de los materiales de Neirz et al. (2024).

Los árboles de decisión y los vecinos cercanos capturan mejor las dependencias no lineales asociadas a X_1 y X_2, mientras que la regresión lineal resulta menos sensible a esas estructuras, según se observa en la Figura 4.3. En la salida tabular reproducible asociada a esta figura, el diseño usa seis atributos, con tres informativos y tres nulos, evaluados en 100 repeticiones para cada modelo base. En las 300 evaluaciones correspondientes a atributos nulos, la regresión lineal produjo 295 valores ARS iguales a cero y cinco positivos acotados, con máximo 0.0158; vecinos cercanos produjo 260 ceros y un máximo de 0.0508; y el árbol de decisión produjo 230 ceros y un máximo de 0.1222. La lectura de este resultado es doble: por una parte, ARS refleja la capacidad del modelo base para explotar la relación disponible; por otra, el contraste *null-swap* mantiene una copia nula que ayuda a distinguir señal operacional de estructura irrelevante, aunque la sensibilidad del árbol también deja ver positivos pequeños

sobre distractores en muestras finitas.

4.3.3. Conjunto de referencia con ruido estructurado

La segunda parte del conjunto de referencia introduce ruido estructurado mediante versiones perturbadas de las variables originales. Para cada variable X_i se generan versiones $V_i^{(j)}$ con niveles crecientes de ruido,

$$V_i^{(j)} = X_i + \left(0.01 + \frac{0.5(j-1)}{10-1}\right)\xi_{ij}, \qquad \xi_{ij} \sim \mathcal{N}(0, 0.3^2), \quad j = 1, \ldots, 10. \tag{4.2}$$

De este modo, la correlación entre la variable original y su versión perturbada disminuye progresivamente con j. El diseño considera 100 repeticiones con 100 observaciones, lo que permite comparar medias y desviaciones de distintas medidas de relevancia bajo escenarios de señal más débil.

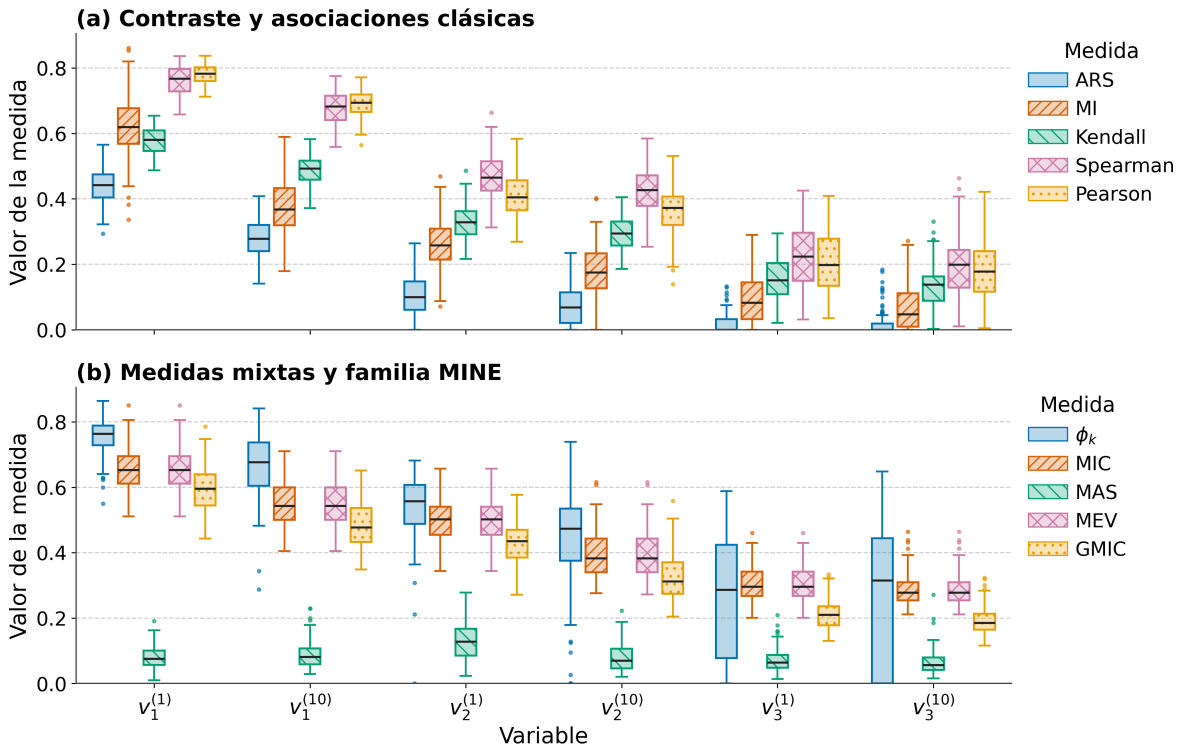


Figura 4.4: Variables informativas perturbadas por ruido: comparación de medidas de dependencia y relevancia bajo niveles crecientes de perturbación. El panel (a) reúne ARS, información mutua y asociaciones clásicas; el panel (b), ϕ_k y las medidas de la familia MINE. Las tramas complementan el color para facilitar la distinción visual. Figura reconstruida a partir de los materiales de Neirz et al. (2024).

En las variables informativas, ARS asignó valores mayores a la variable dominante asociada a la relación exponencial: para $V_1^{(1)}$ y $V_1^{(10)}$, los valores medios reproducidos fueron aproximadamente 0.443 y 0.275, respectivamente; la variable logística $V_2^{(j)}$ mantuvo valores positivos más bajos, en torno a 0.097 y 0.074 en los extremos considerados; y la variable lineal $V_3^{(j)}$ mostró valores menores, cercanos a 0.022 y 0.020, según se aprecia en la Figura 4.4. Esta jerarquía es compatible con la estructura de la función generadora, pues no todas las variables informativas aportan la misma magnitud de predictibilidad operacional bajo el modelo elegido.

En las variables no informativas, el comportamiento fue distinto: ARS se concentró cerca de cero, con valores positivos como casos raros. Las tres variables nulas, $V_4^{(1)}$, $V_5^{(1)}$ y $V_6^{(1)}$,

reproducen el mismo patrón: medias de ARS entre 0.006 y 0.010 con desviaciones estándar entre 0.018 y 0.025, mientras que Spearman se mantiene entre 0.079 y 0.088, MIC entre 0.239 y 0.244 y GMIC entre 0.139 y 0.143. Estos valores permiten formular la comparación de manera más precisa: el contraste *null-swap* produjo una magnitud media sustancialmente menor sobre los distractores y, dado que el cero queda dentro de la variabilidad reportada, esos positivos aislados son compatibles con fluctuaciones alrededor de ausencia de efecto estable. El contraste actúa entonces como filtro conservador frente a asociaciones espurias o fluctuaciones muestrales, tendencia que la Figura 4.5 confirma para el conjunto completo de variables no informativas y medidas comparadas.

Tabla 4.5: Valores trazadores del conjunto de referencia con ruido estructurado. Las celdas reportan media y desviación estándar reproducidas desde los CSV generados a partir de los materiales de Neirz et al. (2024).

Tipo	Variable	ARS	Pearson	Spearman	MIC	GMIC
Informativa	$V_1^{(1)}$	0.443 (0.055)	0.781 (0.028)	0.762 (0.044)	0.652 (0.064)	0.595 (0.065)
Informativa	$V_1^{(10)}$	0.275 (0.058)	0.690 (0.042)	0.679 (0.054)	0.553 (0.070)	0.486 (0.068)
Informativa	$V_2^{(1)}$	0.097 (0.061)	0.412 (0.072)	0.467 (0.073)	0.499 (0.067)	0.430 (0.064)
Informativa	$V_2^{(10)}$	0.074 (0.061)	0.365 (0.076)	0.425 (0.075)	0.397 (0.072)	0.326 (0.070)
Informativa	$V_3^{(1)}$	0.022 (0.033)	0.205 (0.091)	0.222 (0.095)	0.305 (0.053)	0.212 (0.045)
Informativa	$V_3^{(10)}$	0.020 (0.040)	0.178 (0.090)	0.193 (0.094)	0.286 (0.050)	0.195 (0.042)
Nula	$V_4^{(1)}$	0.010 (0.023)	0.092 (0.067)	0.088 (0.068)	0.240 (0.035)	0.139 (0.028)
Nula	$V_5^{(1)}$	0.006 (0.018)	0.085 (0.066)	0.079 (0.059)	0.239 (0.032)	0.140 (0.027)
Nula	$V_6^{(1)}$	0.010 (0.025)	0.079 (0.061)	0.080 (0.064)	0.244 (0.039)	0.143 (0.030)

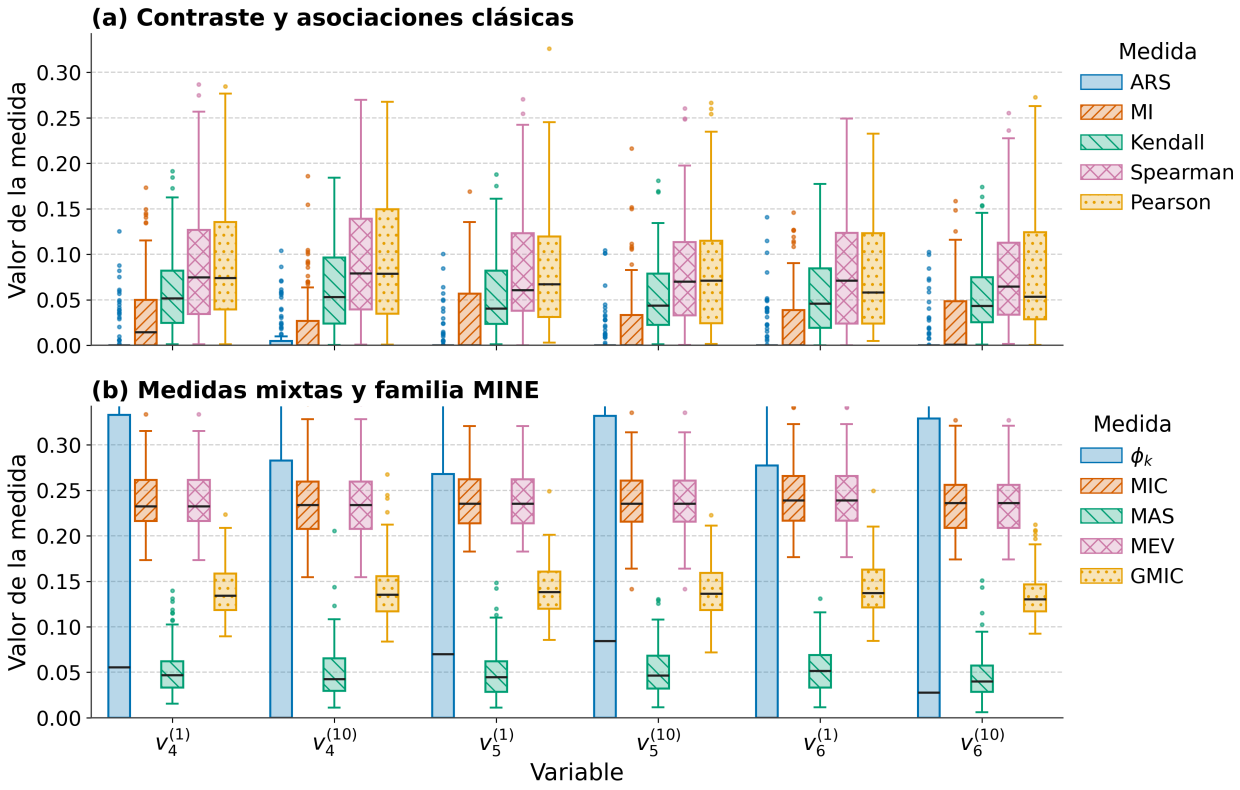


Figura 4.5: Variables no informativas perturbadas por ruido: comparación de medidas de dependencia y relevancia. El panel (a) reúne ARS, información mutua y asociaciones clásicas; el panel (b), ϕ_k y las medidas de la familia MINE. Las tramas complementan el color para facilitar la distinción visual. Algunas medidas asignan valores positivos a variables definidas como nulas en el proceso generador, mientras ARS tiende a concentrarse cerca de cero. Figura reconstruida a partir de los materiales de Neirz et al. (2024).

La interpretación debe evitar dos excesos: por un lado, concluir que ARS reemplaza a todas las medidas de dependencia, cuando las medidas comparadas responden a preguntas distintas y pueden ser útiles para exploración; por otro, tratar el valor de ARS como evidencia automática de informatividad poblacional. La tesis adopta una lectura intermedia, según la cual ARS ofrece evidencia de relevancia operacional bajo un protocolo explícito y su valor aumenta cuando se reporta junto con estabilidad, significancia, tamaño de efecto y una discusión del modelo base.

4.4. Bloque II: fronteras epistémicas

4.4.1. Resultados formales como guía de lectura

La segunda línea de validación parte de una afirmación formal: causalidad, informatividad poblacional y predictibilidad operacional no se implican mutuamente en general. La demostración se construye mediante contraejemplos en modelos causales estructurales y permite ordenar situaciones recurrentes en análisis predictivos, donde puede existir una variable causal con señal observacional débil, una variable informativa por confusión sin efecto causal directo, o una variable con utilidad para un predictor finito sin que esa utilidad represente una propiedad poblacional estable.

El marco también establece un vínculo positivo bajo condiciones específicas: en clasificación con variable objetivo discreta y pérdida logarítmica, la brecha de riesgo de Bayes asociada a retirar una variable desde un contexto de atributos coincide con su información mutua condicional en ese contexto. Esta identidad ofrece una referencia para conectar desempeño poblacional e informatividad poblacional; sin embargo, esa conexión exige protocolos consistentes respecto del riesgo de Bayes y un contraste adecuado frente a una copia nula. La teoría indica, por tanto, qué condiciones harían razonable leer una mejora empírica como información poblacional.

4.4.2. Dependencia del protocolo

El primer experimento de esta línea estudia la dependencia del protocolo al estimar contrastes *null-swap*. Para ello, se generan datos de clasificación binaria con una variable informativa X_1 y tres variables nulas Z_1, Z_2 y Z_3:

$$X_1, Z_1, Z_2, Z_3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad Y \mid X_1 \sim \text{Bernoulli}\{\sigma(\alpha X_1)\}, \tag{4.3}$$

donde $\sigma(t) = (1+\exp(-t))^{-1}$ es la función logística. En la primera parte se fija $\alpha = 2$ y en la segunda se varía $\alpha \in \{0.5, 2, 10\}$ para modificar la intensidad de la señal. Así, X_1 posee informatividad por construcción, mientras las variables Z_j son nulas. En esta sección, una variable nula es un atributo independiente de la respuesta en el proceso generador de datos (DGP, por *data-generating process*); por tanto, cualquier contraste positivo persistente sobre ese grupo debe interpretarse como efecto del protocolo, de la muestra o del estimador.

El diseño compara tres protocolos: el primero usa bootstrap, un procedimiento de remuestreo con reemplazo aplicado sobre una muestra fija; el segundo emplea validación cruzada de 10 particiones, procedimiento que rota la partición usada para evaluación mientras entrena con las restantes hasta completar el número de evaluaciones; y el tercero usa realizaciones independientes del DGP. Una evaluación corresponde, respectivamente, a una remuestra bootstrap, a un fold de validación cruzada o a una muestra fresca del proceso generador. Por ello, cuando $B = 100$, el protocolo de validación cruzada equivale a 10 repeticiones completas de 10 folds, mientras que el protocolo DGP usa 100 conjuntos independientes. El bootstrap sirve como protocolo de referencia: puede conservar regularidades accidentales de una muestra realizada y permite observar qué ocurre cuando la repetición difiere de aproximar el proceso poblacional.

Al variar el tamaño muestral entre 32 y 8192 observaciones y considerar $B \in \{10, 50, 100\}$ evaluaciones por configuración, el contraste estimado compara la información del atributo original con la de su copia nula en el contexto univariado de este diseño y permite examinar si el patrón depende de una elección puntual de presupuesto experimental. En particular, $\widetilde{\Delta}$ se calcula como diferencia entre la información mutua estimada para la columna observada de la variable real y la información mutua estimada para su copia nula, $\widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$; $\widehat{I}$ corresponde al estimador de información mutua para variables mixtas discreto-continuas basado en vecinos cercanos de Ross (2014), con $k = 5$ y reducción a $\min(k, n-1)$ en tamaños muestrales pequeños, según la implementación de *scikit-learn* (Pedregosa et al., 2011).

El resultado central es que el protocolo modifica el significado de la evidencia. Bajo bootstrap, las variables nulas exhiben contrastes positivos persistentes incluso cuando aumenta el tamaño muestral. El patrón sugiere un mecanismo más específico que una fluctuación muestral aislada: al remuestrear con reemplazo, el protocolo introduce filas duplicadas y altera la geometría local que usa el estimador por vecinos cercanos; la copia permutada, en cambio, rompe parte del emparejamiento inducido por esos duplicados. La diferencia resultante puede quedar desplazada hacia valores positivos aun cuando la variable evaluada sea nula por construcción. Bajo realizaciones independientes del proceso generador, en cambio, las variables nulas se concentran alrededor de cero y la variable informativa se separa con mayor claridad a medida que crece n, mientras que la validación cruzada ocupa una posición intermedia: evita el sesgo positivo persistente del bootstrap, pero muestra mayor variabilidad en tamaños pequeños.

La segunda parte del experimento varía la intensidad de la señal mediante el parámetro $\alpha \in \{0.5, 2, 10\}$, lo que permite distinguir dos efectos. Cuando el protocolo es compatible con la cantidad poblacional de interés, el aumento de n reduce la dispersión y ayuda a separar la variable informativa de las nulas; cuando el protocolo conserva una orientación inconsistente, en cambio, el tamaño muestral no corrige el problema de fondo. En el régimen de $n = 8192$, por ejemplo, el contraste de ruido bajo bootstrap permanece alrededor de 0.10, mientras el contraste de la variable informativa aumenta con la fuerza de la señal. La constancia de esa meseta apunta a una interacción sistemática entre remuestreo y estimación, más que a una alineación accidental que debería atenuarse con n. Por eso, todo contraste positivo debe leerse junto con los nulos conocidos.

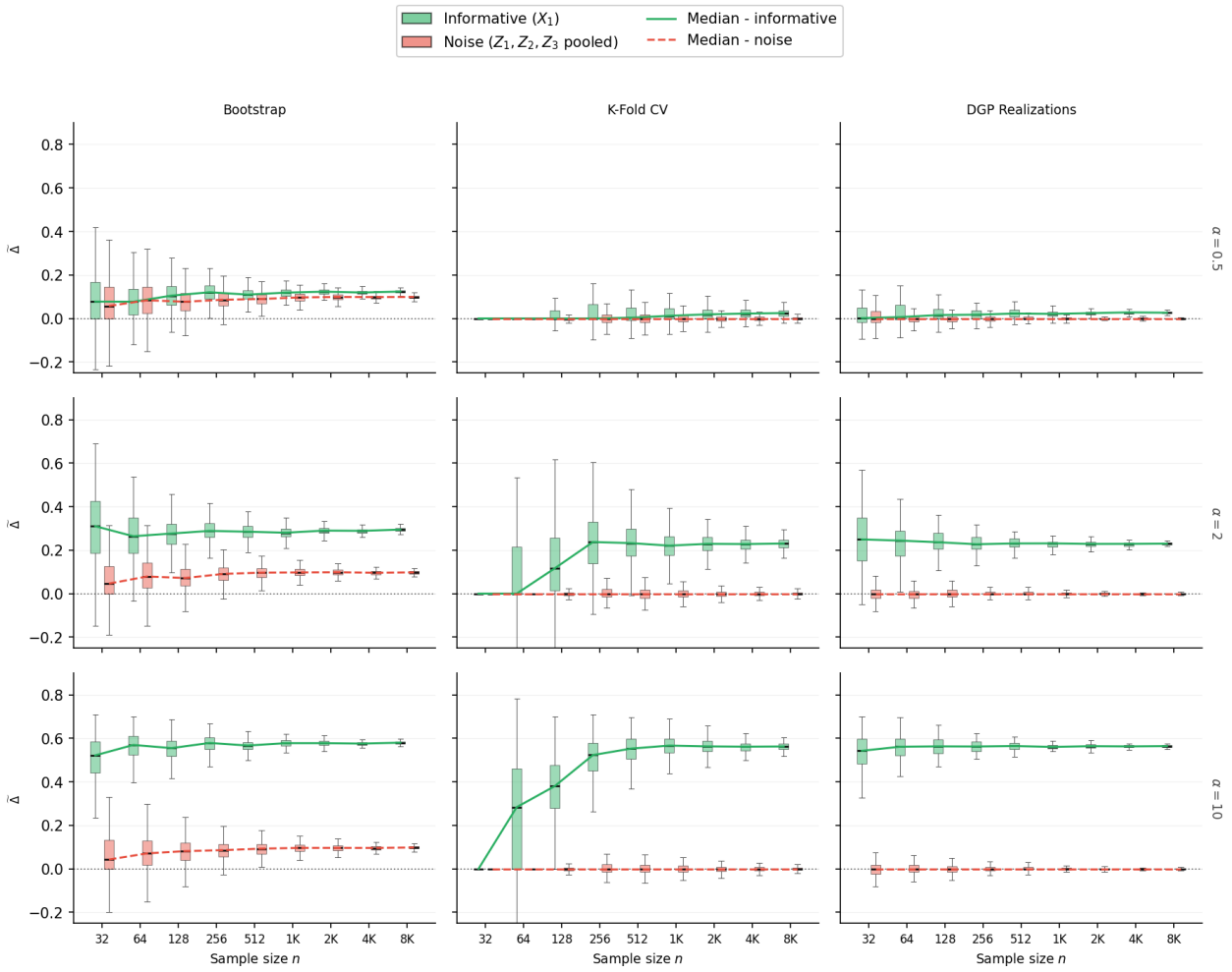


Figura 4.6: Intensidad de señal: contraste *null-swap* bajo variación de $\alpha \in \{0.5, 2, 10\}$ con $B = 100$ evaluaciones. Cada panel compara la variable informativa X_1 con las variables nulas agrupadas, según protocolo (bootstrap, validación cruzada y realizaciones independientes del proceso generador) y tamaño muestral. Las líneas indican las medianas de cada grupo.

La Figura 4.6 resume este análisis. Con señal débil ($\alpha = 0.5$), la variable informativa y las nulas resultan difíciles de separar bajo bootstrap, donde ambas medianas convergen hacia valores positivos semejantes; con señales más fuertes ($\alpha = 2$ y $\alpha = 10$), la variable informativa se distancia en los tres protocolos, pero las variables nulas conservan un contraste positivo persistente solo bajo bootstrap, mientras la validación cruzada y las realizaciones independientes las mantienen alrededor de cero. La figura también muestra que la validación

cruzada requiere tamaños muestrales moderados antes de estabilizar la separación, observación pertinente para regímenes de muestra pequeña. El detalle numérico completo de este experimento, desglosado por protocolo, tamaño muestral e intensidad de señal, se reporta en el Apéndice B.

Esta lectura matiza el resultado formal de separación entre $\mathcal{P}$ e $\mathcal{I}$. La relevancia práctica de la brecha entre predictibilidad operacional e informatividad poblacional depende del régimen. Con señales fuertes, tamaños muestrales grandes y protocolos orientados a nuevas realizaciones del proceso generador, la evidencia operacional puede acercarse a una lectura poblacional. Con señales débiles o protocolos que reiteran una muestra fija, una diferencia positiva puede adquirir estabilidad empírica sin ganar el alcance epistémico que se busca.

Tabla 4.6: Contraste *null-swap* $\widetilde{\Delta} = \widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$ para la variable informativa X_1 y para las variables nulas agrupadas Z_1, Z_2 y Z_3, con $\alpha = 2$ y $B = 100$ evaluaciones. Las celdas reportan media ± desviación estándar, calculadas a partir de los resultados crudos reproducibles del experimento de dependencia del protocolo.

	Bootstrap		Validación cruzada		Realizaciones DGP	
n	X_1	Nulas	X_1	Nulas	X_1	Nulas
32	0.303 ± 0.161	0.068 ± 0.106	0.005 ± 0.083	0.001 ± 0.077	0.253 ± 0.120	0.004 ± 0.072
64	0.267 ± 0.110	0.086 ± 0.080	0.106 ± 0.216	-0.001 ± 0.158	0.234 ± 0.084	-0.003 ± 0.041
128	0.275 ± 0.069	0.076 ± 0.057	0.146 ± 0.162	-0.001 ± 0.080	0.240 ± 0.058	0.001 ± 0.033
256	0.290 ± 0.046	0.091 ± 0.042	0.236 ± 0.135	0.001 ± 0.065	0.227 ± 0.043	0.002 ± 0.019
512	0.285 ± 0.037	0.096 ± 0.030	0.235 ± 0.091	-0.001 ± 0.049	0.233 ± 0.026	0.002 ± 0.015
1024	0.279 ± 0.026	0.098 ± 0.022	0.219 ± 0.065	0.000 ± 0.034	0.227 ± 0.020	0.001 ± 0.011
2048	0.290 ± 0.017	0.098 ± 0.015	0.228 ± 0.045	-0.001 ± 0.024	0.228 ± 0.014	0.000 ± 0.007
4096	0.288 ± 0.011	0.096 ± 0.010	0.227 ± 0.032	0.000 ± 0.017	0.228 ± 0.010	0.000 ± 0.005
8192	0.295 ± 0.009	0.098 ± 0.007	0.230 ± 0.024	0.000 ± 0.012	0.230 ± 0.007	0.000 ± 0.004

La Tabla 4.6 vuelve visible el punto metodológico principal. Un valor positivo de $\widetilde{\Delta}$ indica que la variable real muestra más información mutua estimada con Y que su copia nula; un valor cercano a cero indica que ambas son indistinguibles bajo el protocolo; y un valor negativo refleja variación estimativa o una ventaja accidental de la copia nula. La variable informativa X_1 mantiene contrastes positivos bajo los tres protocolos una vez que el tamaño muestral es suficiente; en tamaños pequeños, la validación cruzada exhibe mayor dispersión porque cada evaluación opera sobre folds reducidos. Las variables nulas, en cambio, separan con claridad los protocolos: las realizaciones independientes del proceso generador

se concentran alrededor de cero y reducen su dispersión al crecer n; la validación cruzada también permanece cerca de cero, aunque con mayor variabilidad en muestras pequeñas; el bootstrap estabiliza un contraste positivo cercano a 0.10. Esta diferencia muestra el límite de la repetición: repetir sobre una muestra fija puede reproducir con precisión una cantidad inducida por el acoplamiento entre remuestreo y estimador. En el Apéndice B se explicita el diagnóstico metodológico que se desprende de este patrón.

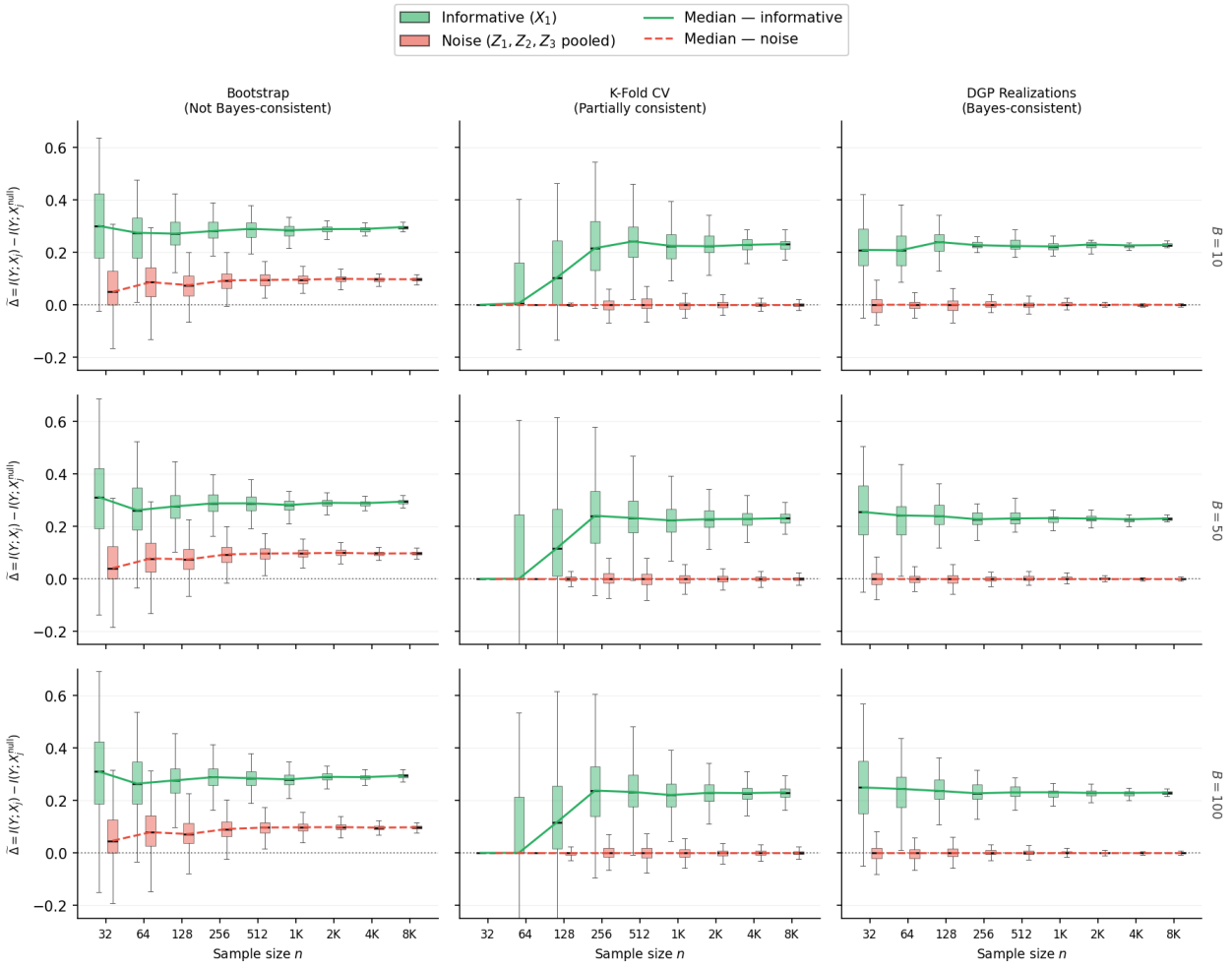


Figura 4.7: Sensibilidad del protocolo: contraste *null-swap* frente a protocolo, tamaño muestral y número de evaluaciones. La figura compara la variable informativa X_1 con las variables nulas agrupadas bajo bootstrap, validación cruzada y realizaciones independientes del proceso generador.

La Figura 4.7 complementa la tabla porque muestra la distribución completa de los contrastes, no solo sus medias. La separación entre X_1 y las variables nulas aparece con mayor claridad cuando el protocolo se aproxima a nuevas realizaciones del proceso generador; bajo bootstrap, en cambio, el grupo de nulos conserva una masa positiva incluso con tamaños muestrales grandes. El patrón cualitativo se conserva al variar el presupuesto de evaluación entre $B = 10$, $B = 50$ y $B = 100$: aumentar B reduce la dispersión visual de las estimaciones, pero mantiene la meseta positiva de las variables nulas bajo bootstrap. Dado que Z_1, Z_2 y Z_3 son nulas por construcción, esa masa positiva no puede atribuirse a una señal omitida del dominio. La lectura más plausible es que el protocolo produce una inflación estable del contraste nulo a través de su interacción con el estimador de información mutua. Por ello, la comparación muestra, además de variabilidad, un sesgo de interpretación: la evidencia operacional puede estabilizarse alrededor de una cantidad que pertenece al protocolo antes que a la distribución poblacional.

4.4.3. Disociación entre causalidad, informatividad y predictibilidad

El segundo experimento construye un modelo causal estructural con cuatro tipos de variables: una variable causal, X_{causal}, con efecto bajo intervención sobre Y pero señal observacional marginal débil; una variable confundida, X_{conf}, asociada a Y a través de una causa común, aunque sin efecto causal directo; dos variables sinergéticas, X_{syn1} y X_{syn2}, que tienen poca informatividad marginal pero aportan información de manera conjunta mediante un patrón tipo XOR (*exclusive OR*), donde la señal aparece en la combinación de variables y no en cada variable por separado; y variables de ruido independientes.

La construcción es estilizada a propósito: aísla tres patrones que aparecen con frecuencia en la práctica científica, a saber, efectos causales cuya firma observacional marginal puede ser débil, variables predictivas que actúan como proxies por confusión, tratamiento o adquisición, e información que se vuelve visible solo en combinaciones de atributos. El ejemplo de neumonía y asma analizado en la Sección 2.8 ilustra el segundo patrón (Caruana et al., 2015); los trabajos sobre atajos en imágenes médicas (Badgeley et al., 2019; DeGrave et al., 2021) y sobre interacción genética (Cordell, 2009) apuntan a dificultades semejantes desde otros dominios. El SCM permite reunir esas tensiones bajo verdad de fondo conocida.

El proceso generador de datos se define a partir de un confusor latente $U \sim \text{Bernoulli}(0.5)$ y una causa común observacional $Z_{\text{common}} \sim \mathcal{N}(0,1)$. La variable causal se genera como $X_{\text{causal}} = U + \varepsilon_c$, con $\varepsilon_c \sim \mathcal{N}(0, 0.1^2)$, pero su contribución a la respuesta se modula por $(1-U)$, de modo que parte de su señal observacional queda cancelada. Esta construcción instancia el contraejemplo (i) de la Proposición 3.1 con cancelación parcial en lugar de total: como $X_{\text{causal}} \approx U$, el componente $U(1-U)$ se anula y solo sobrevive la variación inducida por $\varepsilon_c(1-U)$, mientras que una intervención externa sobre X_{causal} sí modifica el mecanismo de respuesta. La variable confundida se genera como $X_{\text{conf}} = Z_{\text{common}} + \varepsilon_{\text{conf}}$, con $\varepsilon_{\text{conf}} \sim \mathcal{N}(0, 0.3^2)$, mientras que Y recibe un componente $1.5Z_{\text{common}}$. Las variables sinergéticas se generan uniformemente en $[0,1]$ y aportan el término $2\,\text{XOR}(X_{\text{syn1}} > 0.5,\, X_{\text{syn2}} > 0.5)$, donde $\text{XOR}(a,b)$ denota la disyunción exclusiva, igual a 1 si exactamente una de las proposiciones a, b es verdadera y 0 en caso contrario. El diseño incluye además tres variables de ruido independientes de Y. Finalmente, la respuesta binaria se obtiene a partir de una variable latente:

$$Y^* = X_{\text{causal}}(1-U) + 1.5Z_{\text{common}} + 2\,\text{XOR}(X_{\text{syn1}} > 0.5,\, X_{\text{syn2}} > 0.5) + \varepsilon_Y, \tag{4.4}$$

donde $\varepsilon_Y \sim \mathcal{N}(0, 0.5^2)$, $P(Y=1) = \sigma(Y^*)$ y σ denota nuevamente la función logística. Esta construcción separa utilidad predictiva, dependencia observacional, sinergia e intervención causal, de modo que cada medida pueda evaluarse contra un papel conocido del atributo dentro del proceso generador.

El diseño permite observar cómo se separan los espacios $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$ cuando el proceso generador es conocido. En estimaciones marginales de información mutua con $n = 8192$, X_{causal} aparece con un valor cercano a cero, aproximadamente 0.002, pese a su papel causal bajo intervención; en cambio, X_{conf} muestra un valor mayor, aproximadamente 0.095, debido a la confusión. El resultado refleja una señal marginal observacional débil bajo la distribución generada: una medida marginal de informatividad puede subestimar una variable causal cuyo efecto permanece bajo intervención. Las variables sinergéticas permanecen cerca de cero cuando se evalúan marginalmente, aunque su contribución se vuelve visible al considerar el par.

Luego se entrenan cuatro modelos sobre una partición común de entrenamiento y prueba generada con $n = 8192$ observaciones: Random Forest, un ensamble de árboles de decisión

entrenados con aleatorización; Gradient Boosting, un ensamble secuencial que ajusta modelos para corregir errores residuales; regresión logística; y una red neuronal multicapa. Los desempeños predictivos resultan comparables, con *accuracy* entre 0.729 y 0.761, y AUC (área bajo la curva ROC, que resume la capacidad de separar clases a distintos umbrales) entre 0.763 y 0.803. En ese contexto, el análisis SHAP muestra que X_{conf} recibe la mayor importancia en todos los modelos, mientras X_{causal} queda en posiciones bajas. La Figura 4.8 expone esta jerarquía de importancias para los cuatro modelos entrenados, con la variable confundida ocupando siempre el primer lugar y la variable causal relegada por su débil señal observacional marginal. Esta observación delimita el alcance de SHAP, que describe lo que el predictor usa bajo una representación y un protocolo. Si el modelo aprovecha una variable confundida, SHAP puede reportar fielmente esa dependencia.

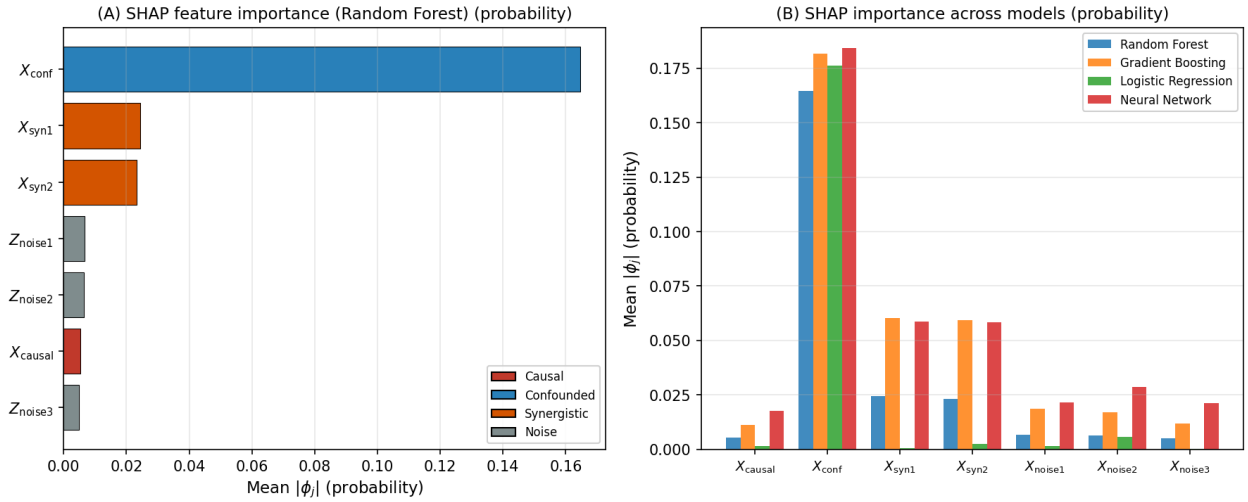


Figura 4.8: Disociación epistémica: importancia SHAP en el diseño de disociación epistémica. La variable confundida aparece como la más prominente para el predictor, mientras la variable causal queda relegada por su señal observacional débil.

La lectura cualitativa de la Figura 4.8 es consistente entre los cuatro modelos entrenados: la magnitud relativa de las barras varía, pero el orden de prominencia se mantiene, con la variable confundida siempre por delante de la causal y de las variables de ruido. SHAP describe la contribución de cada atributo a la predicción del modelo ajustado bajo esa representación y ese protocolo; esta estabilidad cualitativa no debe confundirse con una lectura causal ni con el papel del atributo en el mecanismo generador de la respuesta.

La Tabla 4.7 resume cómo se calcula cada columna de la comparación que sigue. Con esa referencia, la Tabla 4.8 compara medidas de importancia sobre una base común, usando el Random Forest del mismo diseño como modelo de referencia para las importancias globales. Gini, permutación, SHAP y LIME describen prominencia predictiva bajo ese predictor y protocolo; Boruta agrega una lectura de selección *all-relevant* basada también en bosques, y el *null-swap* por pares evalúa una pregunta distinta: si la variable observada supera a su copia nula dentro de un contexto, es decir, con una segunda covariable disponible sin permutar.

Tabla 4.7: Métodos de importancia comparados en el diseño de disociación epistémica: qué reporta cada columna de la Tabla 4.8.

Método	**Cómo se calcula**
Gini	Importancia por impureza del bosque (*mean decrease in impurity*): para cada atributo, suma las reducciones de impureza de Gini producidas por las divisiones que lo usan, ponderadas por la proporción de muestras que alcanza cada nodo y normalizadas a través de los árboles.
Perm. RMSE	`permutation_importance` sobre la partición de prueba, usando como puntuación el negativo del RMSE entre Y y la probabilidad predicha de la clase positiva; valores mayores indican que permutar el atributo aumenta más el error probabilístico.
SHAP	Media de contribuciones absolutas para la clase positiva.
LIME	Promedio de los coeficientes absolutos de sustitutos lineales locales ajustados mediante Ridge sobre perturbaciones, tomando como objetivo la probabilidad estimada $P(Y = 1 \mid x)$.
Boruta	Posiciones devueltas por su procedimiento de selección *all-relevant*, basado también en bosques.
Null-swap por pares	Si la variable observada supera a su copia nula dentro de un contexto, evaluado sobre pares ordenados $S = \{i, j\}$.

Tabla 4.8: Comparación de importancias: medidas predictivas, selección *all-relevant* y contraste *null-swap* por pares en el diseño de disociación epistémica. Gini, permutación, SHAP y LIME se calculan sobre el Random Forest de referencia; Perm. RMSE indica el aumento medio de RMSE al permutar cada atributo, calculado sobre probabilidades predichas. Boruta reporta las posiciones devueltas por su procedimiento de selección; los empates se conservan como tales, de modo que varias variables pueden compartir la misma posición.

Variable	Gini	Perm. RMSE	SHAP	LIME	Boruta (pos.)	*null-swap*
X_{conf}	0.509	0.085	0.177	0.123	1	0.051
X_{syn1}	0.103	0.017	0.023	0.067	1	0.005
X_{syn2}	0.101	0.017	0.022	0.067	1	0.005
Z_{noise1}	0.073	0.000	0.009	0.009	2	0.000
X_{causal}	0.073	-0.000	0.008	0.036	3	0.000
Z_{noise2}	0.072	-0.001	0.010	0.009	3	0.000
Z_{noise3}	0.069	0.000	0.007	0.009	5	-0.000

La Tabla 4.8 permite ver el desacople entre prominencia predictiva y estatuto causal. Casi todas las medidas ubican a X_{conf} en primer lugar; las variables sinergéticas aparecen como relevantes para varios métodos, aunque su señal depende del tratamiento conjunto; y X_{causal} no destaca en la mayoría de los rankings. El comportamiento es consecuencia de la pregunta que estas medidas responden: describen utilidad para el predictor o selección bajo una representación; la intervención causal pertenece a otro registro. La consecuencia metodológica es directa: una tabla de importancia puede ser precisa respecto del modelo y, al mismo tiempo, inducir una interpretación equivocada si se la traslada al mecanismo generador sin el control causal correspondiente.

La verificación mediante intervenciones completa la lectura. Al simular intervenciones, el DGP se regenera fijando externamente la variable intervenida y manteniendo los demás mecanismos estructurales. Bajo ese procedimiento, fijar X_{conf} no modifica la probabilidad media de Y y el cambio reportado es aproximadamente 0.000; en cambio, intervenir sobre X_{causal} produce un cambio de $\Delta P(Y = 1) \approx 0.268$ en la probabilidad media de la respuesta. Así, la variable más prominente para SHAP no es una causa directa, mientras la variable

causal tiene baja prominencia operacional y observacional marginal. El contraste resume el punto de fondo: la prominencia predictiva puede identificar señales útiles para anticipar respuestas, pero su traducción a causalidad requiere un diseño o un supuesto adicional que no está contenido en el ranking.

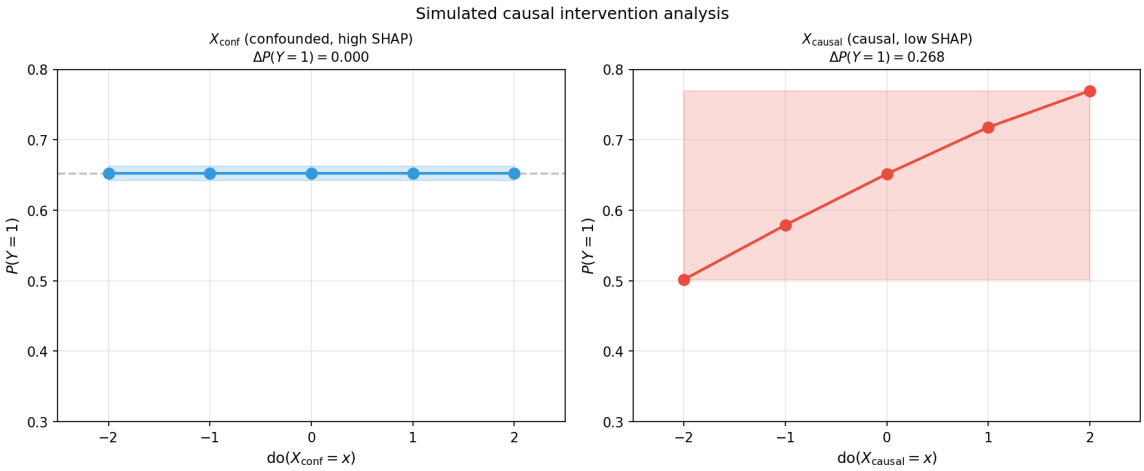


Figura 4.9: Intervenciones simuladas: contraste sobre la variable confundida y la variable causal. El resultado muestra que alta importancia predictiva y efecto causal directo pueden separarse en un proceso generador conocido.

El experimento incorpora además contrastes *null-swap* por pares para evaluar estructuras de orden superior. En cada par ordenado se permuta el atributo evaluado dentro del par, manteniendo intacta la variable de respuesta. La métrica es ΔRMSE calculada sobre probabilidades predichas: un valor positivo indica que la permutación de la variable evaluada aumenta el error respecto de la versión observada. El RMSE evalúa la degradación en el espacio de probabilidades predichas y evita que el contraste dependa de un umbral de clasificación fijo. El efecto de X_{conf} permanece positivo, con una media agregada de ΔRMSE cercana a 0.051 a través de los contextos evaluados, lo que refleja su utilidad operacional bajo el protocolo; el efecto de X_{causal} permanece cerca de cero, con una media agregada cercana a 0.00024, consistente con su señal observacional débil; y el par sinergético muestra contrastes positivos en ambas direcciones, cercanos a 0.0284, lo que indica que su relevancia emerge de manera conjunta. La Tabla 4.9 reproduce una selección de estos contrastes para contextos específicos, por lo que una fila individual puede diferir levemente de la media agregada citada;

la columna $P(\Delta > 0)$ reporta la frecuencia empírica de repeticiones en que el contraste fue positivo para cada fila. Esta parte del experimento muestra por qué una medida univariada como ARS puede requerir extensiones multivariadas cuando la información aparece por interacción.

Tabla 4.9: Contrastes por pares: contrastes *null-swap* seleccionados, estimados como ΔRMSE bajo validación cruzada repetida.

Contexto	Variable anulada	Media	Desv. est.	$P(\Delta > 0)$
X_{causal}	X_{conf}	0.0507	0.0073	1.000
X_{conf}	X_{causal}	-0.0003	0.0025	0.475
X_{syn1}	X_{syn2}	0.0284	0.0055	1.000
X_{syn2}	X_{syn1}	0.0284	0.0053	1.000
X_{causal}	Z_{noise1}	-0.0002	0.0022	0.465

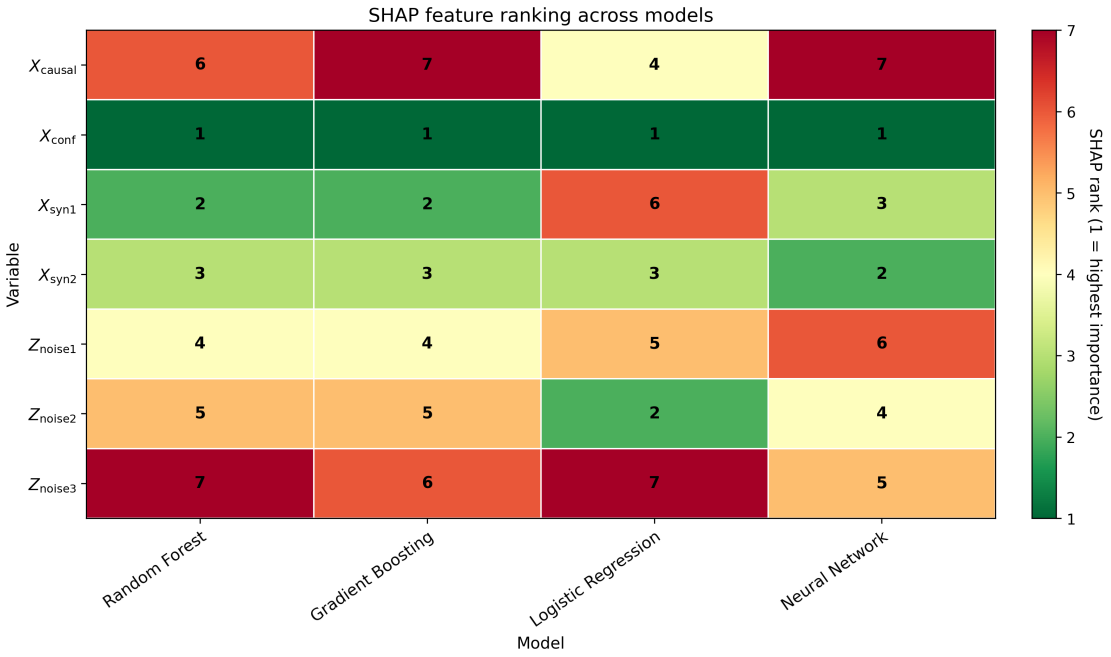


Figura 4.10: Rashomon predictivo: variación de rankings SHAP entre modelos con desempeños comparables. La figura ilustra el componente Rashomon del problema: modelos con resultados predictivos próximos pueden ordenar las variables de manera distinta.

La Tabla 4.9 y la Figura 4.10 cierran el argumento experimental. El *null-swap* por pares detecta la utilidad operacional de X_{conf} y la señal conjunta de las variables sinergéticas, pero mantiene cerca de cero la variable causal cuando se la evalúa bajo una señal observacional débil. A su vez, los rankings SHAP entre modelos hacen visible el componente Rashomon del problema. El rango se obtiene ordenando las variables por su importancia SHAP media absoluta, con rango 1 para la variable más prominente. Todos los modelos asignan rango 1 a X_{conf}, lo que indica estabilidad de la señal confundida; sin embargo, la concordancia entre rankings varía de manera amplia, con Kendall τ entre 0.048 y 0.714. La variable causal permanece en posiciones bajas, pero no de forma idéntica: aparece en rango 4 para la regresión logística y cae hasta rango 7 en Gradient Boosting y en la red neuronal. La consecuencia práctica es directa: una explicación basada en un único predictor puede describir fielmente ese predictor y, aun así, ofrecer una historia de relevancia dependiente de la clase de modelos. Por ello, la gramática de reporte propuesta exige declarar el modelo, el protocolo y el registro epistémico de la importancia antes de trasladar el ranking hacia una afirmación sobre el fenómeno.

4.5. Discusión integrada

Los resultados convergen en la importancia del contraste *null-swap*. En ARS, el contraste permite comparar el atributo original contra una versión permutada bajo el mismo modelo, la misma métrica y un contexto de evaluación declarado; en los experimentos de dependencia del protocolo, la misma idea se usa para evaluar si un efecto operacional se mantiene frente a una copia nula y para diagnosticar cuándo un protocolo produce positivos sobre variables que deberían permanecer en cero. En ambos casos, la copia nula funciona como referencia interna construida desde el mismo problema.

La diferencia principal está en el nivel de lectura: ARS entrega una medida de relevancia operacional, cuyos resultados son valiosos para selección de variables y análisis exploratorio, especialmente cuando se reportan con repeticiones, significancia y tamaño de efecto; el marco epistémico, en cambio, pregunta cuándo esa relevancia operacional puede interpretarse como evidencia de informatividad poblacional. Esa transición exige consistencia del protocolo, estabilidad del contraste y cuidado frente a fuga de información o artefactos de remuestreo.

Los experimentos también muestran que la explicabilidad post-hoc debe ubicarse con

precisión. SHAP puede ser fiel al predictor y, al mismo tiempo, asignar alta importancia a una variable confundida. Esta situación aparece cuando una herramienta de explicación del modelo se usa como herramienta de identificación causal. La tesis propone mantener esa distinción en el reporte: atribución del modelo, informatividad poblacional y causalidad deben declararse como niveles distintos.

Una tercera lectura transversal concierne a las interacciones. ARS, en la formulación presentada, evalúa atributos individualmente, mientras que el experimento con variables tipo XOR muestra que existen situaciones donde la señal emerge solo al considerar combinaciones. Esto sugiere una línea de extensión natural: conservar el principio *null-swap*, pero aplicarlo a subconjuntos o contrastes condicionados; dicha extensión requiere cautela porque el número de combinaciones crece rápidamente y porque la multiplicidad estadística puede inflar hallazgos aparentes.

En términos metodológicos, los resultados invitan a reportar la relevancia como una afirmación con alcance: decir que una variable es relevante bajo ARS significa que produjo un efecto positivo en un contraste *null-swap* bajo una métrica, un modelo, un protocolo y un contexto específicos; decir que una variable es informativa en sentido poblacional requiere condiciones de consistencia para ese contraste; y decir que es causal exige intervención, identificación o supuestos estructurales defendibles. Esta gramática de reporte pide más declaraciones y, a cambio, comunica los hallazgos con mayor precisión.

Un ejemplo de reporte con alcance declarado permite fijar esta distinción. En el diseño de disociación, X_{conf} puede reportarse como una variable de alta prominencia predictiva para el Random Forest de referencia: ocupa el primer lugar en Gini, permutación, SHAP y LIME en la Tabla 4.8, y el contraste *null-swap* por pares conserva un efecto positivo medio cercano a ΔRMSE=0.051 bajo validación cruzada repetida. La lectura poblacional debe formularse como dependencia observacional inducida por una causa común, con alcance diferente al de un mecanismo directo. La lectura causal queda acotada por la intervención simulada: fijar externamente X_{conf} produce un cambio medio aproximado de 0.000 en $P(Y = 1)$, mientras que intervenir sobre X_{causal} produce $\Delta P(Y = 1) \approx 0.268$. Por tanto, el reporte defendible es: "X_{conf} es operacionalmente prominente bajo el predictor, la métrica y el protocolo declarados; su señal es compatible con informatividad observacional por confusión; la evidencia causal directa bajo la intervención simulada se concentra en X_{causal}".

4.6. Limitaciones

La propuesta posee limitaciones que conviene tener presentes al interpretar los resultados. ARS, en su versión presentada, depende del modelo base y evalúa atributos de forma individual, por lo que puede subestimar relaciones que aparecen solo por interacción, redundancia o codificación contextual. Esta dependencia del contexto forma parte del objeto medido: en presencia de variables altamente correlacionadas o sustitutos cercanos, una permutación marginal altera la estructura conjunta y permite observar cuánto depende el desempeño de esa coordinación. Esa lectura es informativa bajo el protocolo declarado; su alcance poblacional exige distinguir entre señal propia del atributo, redundancia con el contexto y efecto de la nulificación elegida. Por ello, un valor bajo de ARS deja abierta la relevancia estructural cuando existen redundancias, sinergias o restricciones de dominio que no han sido modeladas. Además, la caracterización empírica presentada fija constantes operativas de ARS (la partición 0.6/0.4, el nivel unilateral de Wilcoxon $\alpha = 0.05$ y el margen práctico asociado a $d_z = 0.2$), de modo que la sensibilidad sistemática a esas decisiones queda pendiente de caracterización.

Los procesos generadores utilizados se concentran en ruido y relaciones curvas. La robustez ante valores atípicos (*outliers*) y puntos de alto apalancamiento requiere un estudio específico, pues estos valores pueden alterar de manera distinta el ajuste del modelo, la métrica de desempeño y las medidas de dependencia. Este estudio deberá considerar valores atípicos en las covariables, en la respuesta y en configuraciones conjuntas, con varias prevalencias e intensidades.

La estimación de información mutua y de contrastes *null-swap* también depende del tamaño muestral, del estimador y de decisiones de discretización, vecindad, partición, búsqueda de hiperparámetros y calibración probabilística. En muestras pequeñas, la variabilidad del estimador puede dominar la señal; en muestras grandes, un protocolo orientado hacia una cantidad inadecuada puede estabilizar un sesgo con gran precisión. Además, la evaluación repetida de muchos atributos o subconjuntos introduce un problema de multiplicidad que debe tratarse mediante criterios de significancia, estabilidad o control de falsos descubrimientos.

En particular, el patrón positivo observado sobre variables nulas bajo *bootstrap* es compatible con una interacción entre la duplicación de filas y la geometría local del estimador de información mutua basado en vecinos cercanos. Los experimentos actuales documentan el

patrón. La identificación del mecanismo requiere controles que separen el reemplazo, la duplicación y la familia del estimador. Por ello, la atribución conserva un alcance diagnóstico.

Una limitación adicional concierne a la relación entre el puente formal y las métricas empleadas. La identidad de la Ecuación 3.9, que iguala la brecha de riesgo de Bayes con la información mutua condicional en un contexto de atributos, está establecida para pérdida logarítmica y variable objetivo discreta; los experimentos, en cambio, estiman contrastes con F_1, *accuracy*, error absoluto medio o RMSE, métricas para las cuales esa identidad no se sostiene de manera exacta. La conexión entre contraste empírico e informatividad poblacional opera entonces como referencia conceptual y no como equivalencia cuantitativa: un contraste positivo bajo estas métricas sugiere informatividad cuando el protocolo es consistente, pero su magnitud no puede leerse como estimación directa de la información mutua condicional. Extender el puente formal hacia reglas de puntaje propias u otras pérdidas calibradas queda como tarea pendiente del marco propuesto.

Las simulaciones de fronteras epistémicas aíslan estructuras conocidas para clarificar argumentos, de modo que su valor principal es conceptual y metodológico, más que una afirmación directa sobre un dominio aplicado particular. Los escenarios analizados permiten estudiar ruido, confusión, protocolos de remuestreo, efecto Rashomon y sinergia, pero no agotan la diversidad de dominios donde la relevancia de atributos adquiere consecuencias científicas o aplicadas. Falta evaluar con mayor amplitud casos con mediciones sesgadas, covariate shift, variables categóricas de alta cardinalidad, datos faltantes, costos de adquisición y conocimiento experto incorporado al diseño del contraste.

Estas limitaciones delimitan el alcance de la tesis y, al mismo tiempo, orientan su lectura: los resultados apuntan a una forma de inferencia más explícita, donde cada medida se interpreta según el protocolo que la produce y el tipo de pregunta que pretende responder. La contribución consiste en ofrecer condiciones para decir con mayor precisión cuándo una importancia es operacional, cuándo puede ganar una lectura poblacional y cuándo requiere una justificación causal adicional.

Capítulo 5

Conclusiones generales y trabajo futuro

5.1. Conclusiones

Esta tesis desarrolló una propuesta metodológica y conceptual para interpretar la relevancia de atributos en reconocimiento de patrones, cuyo hilo común puede resumirse en una idea: la importancia de una variable adquiere sentido cuando se declara el tipo de evidencia que la sostiene. Así, una mejora operacional de desempeño, una dependencia poblacional y una relación causal pertenecen a registros distintos, aunque puedan relacionarse bajo condiciones específicas.

La tesis mostró cómo una medida operacional, *Attribute Relevance Score* (ARS), puede leerse dentro de un marco más amplio sobre informatividad poblacional, predictibilidad operacional y causalidad. El estudio de ARS sirvió como punto de partida: mostró que el contraste *null-swap* entrega evidencia operacional y que la lectura poblacional requiere una teoría del protocolo que sostiene esa evidencia.

El ARS aporta una estrategia concreta para evaluar atributos mediante el contraste *null-swap* con versiones permutadas del atributo. Los experimentos sintetizados sugieren que esta comparación ayuda a controlar asociaciones espurias y valores inflados en variables no informativas, especialmente en presencia de ruido; al mismo tiempo, la medida conserva una dependencia explícita respecto del modelo, la métrica y el protocolo, por lo que su

interpretación corresponde al registro de la relevancia operacional.

El marco de fronteras epistémicas amplía esta lectura al separar causalidad, informatividad poblacional y predictibilidad operacional. Sus resultados teóricos y simulados muestran que la predictibilidad no garantiza causalidad, que la informatividad poblacional puede surgir por confusión y que una variable causal puede exhibir señal marginal débil bajo observación; además, indican que métodos post-hoc como SHAP pueden describir fielmente la conducta del modelo sin informar, por sí solos, sobre el mecanismo del fenómeno.

En conjunto, los hallazgos apuntan a una noción de relevancia epistémica, según la cual una variable puede leerse como evidencia de informatividad poblacional, en un contexto de evaluación declarado, cuando se cumplen dos condiciones conjuntas: por una parte, el protocolo debe ser Bayes-consistente respecto de la cantidad poblacional pertinente; por otra, el contraste *null-swap* debe mostrar que el efecto operacional supera a una copia nula equivalente. Esta segunda condición es necesaria porque una ganancia estable bajo un protocolo consistente aún puede provenir de fuentes distintas a la informatividad poblacional: resonancia estocástica, ilustrada en la Figura 3.4, artefactos específicos del modelo o interacciones con ruido.

5.2. Respuesta a la pregunta de investigación

La pregunta formulada en la Sección 1.3 planteaba bajo qué condiciones una diferencia de desempeño atribuida a una variable en un contexto de evaluación, estimada mediante un contraste *null-swap*, puede interpretarse como evidencia de informatividad poblacional. La evidencia reunida permite responderla en dos partes. La primera identifica condiciones conjuntas: el protocolo debe orientarse hacia la cantidad poblacional pertinente, en el sentido Bayes-consistente desarrollado en la Sección 3.3; el contraste debe mostrarse positivo frente a una copia nula equivalente, con respaldo estadístico y un margen práctico declarado; y el efecto debe permanecer estable ante reimplementaciones razonables del procedimiento. La segunda parte delimita el alcance de esa lectura: aun cuando las condiciones se cumplen, la afirmación resultante es observacional, de modo que el paso hacia una interpretación causal requiere supuestos de identificación o diseños con intervención que el contraste, por sí solo, no provee.

Los experimentos otorgan contenido empírico a cada condición. La comparación de pro-

tocolos mostró que el bootstrap sobre una muestra fija puede estabilizar contrastes positivos cercanos a 0.10 en variables nulas conocidas, mientras las realizaciones independientes del proceso generador concentran esos mismos contrastes alrededor de cero; el protocolo, por tanto, forma parte de la afirmación, además de su implementación. La variación de la intensidad de señal indicó que la distancia entre predictibilidad operacional e informatividad poblacional depende del régimen: se estrecha con señales fuertes, tamaños muestrales mayores y protocolos orientados a nuevas realizaciones, y se mantiene abierta cuando la señal es débil o la repetición preserva fluctuaciones de una muestra particular. La disociación epistémica, finalmente, mostró que una variable prominente para el predictor puede carecer de efecto bajo intervención, lo que indica que las condiciones observacionales pertenecen a un registro distinto de los requisitos propios de la lectura causal.

Respecto de la hipótesis enunciada en la Sección 1.4, los resultados son compatibles con su formulación en los regímenes explorados: la relevancia operacional estable se sostuvo cuando el contraste *null-swap* se evaluó de forma repetida, con significancia, margen práctico, protocolo y contexto declarados, y su lectura como informatividad poblacional ganó respaldo bajo protocolos orientados al proceso generador. Los criterios de debilitamiento previstos también se observaron: bajo bootstrap, una ganancia estable sobre variables nulas conservó valor descriptivo como efecto del protocolo sin adquirir alcance poblacional. La hipótesis queda respaldada como criterio de trabajo dentro de los escenarios controlados estudiados; su demostración general y su evaluación en dominios aplicados, con métricas y estructuras de dependencia propias, permanecen como tareas abiertas.

5.3. Contribuciones

La tesis realiza seis contribuciones principales, ordenadas según el recorrido conceptual que sostiene la propuesta:

1. Formaliza una distinción entre $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$, útil para precisar afirmaciones sobre causalidad, informatividad poblacional y predictibilidad operacional.

2. Propone la relevancia epistémica como categoría de transición: una importancia operacional puede leerse como evidencia de informatividad poblacional, en un contexto

declarado, cuando combina consistencia Bayesiana del protocolo, estabilidad y distinguibilidad frente a una copia nula equivalente mediante contraste *null-swap*.

3. Formula los protocolos Bayes-consistentes como condición para vincular efectos operacionales con cantidades poblacionales, aclarando que por sí solos pueden identificar utilidad predictiva consistente sin descartar resonancia estocástica o artefactos específicos del modelo.

4. Integra ARS como aproximación operacional a la relevancia de atributos, basada en el contraste *null-swap*, y muestra cómo su estudio conduce a una separación conceptual más amplia entre lo que el predictor usa, lo que la distribución observacional informa y lo que puede sostenerse causalmente.

5. Reinterpreta las atribuciones post-hoc como evidencia situada en el espacio operacional, con valor descriptivo sobre el modelo y límites frente a lecturas mecanísticas.

6. Propone una gramática de reporte para relevancia de atributos: declarar protocolo, contexto de evaluación, contraste *null-swap*, estabilidad, métrica, relación con el riesgo poblacional y alcance inferencial.

Estas contribuciones retoman antecedentes sobre la diferencia entre predicción y explicación, el alcance causal de las atribuciones post-hoc y la multiplicidad asociada al efecto Rashomon. El aporte doctoral los reúne en un marco común, propone condiciones para interpretar el paso desde la relevancia operacional hacia la informatividad poblacional y vincula ese criterio con ARS. La contribución se concentra así en la integración conceptual y metodológica.

5.4. Publicaciones generadas

Los desarrollos metodológicos y conceptuales de esta tesis dieron origen a las siguientes publicaciones en revistas indexadas:

Año	Publicación	Articulación con la tesis
2024	Neirz, P., Allende, H. y Saavedra, C. *Attribute Relevance Score: A Novel Measure for Identifying Attribute Importance. Algorithms*, 17(11), 518. DOI: `https://doi.org/10.3390/a17110518`.	Sostiene la formulación de ARS en la Sección 3.5 y el Bloque I de resultados en la Sección 4.3; la tesis amplía su lectura al situarla dentro del marco epistémico.
2026	Neirz, P., Allende, H. y Saavedra, C. *Epistemic Frontiers and the Distinction between Causality, Information, and Predictability in Pattern Recognition. Scientific Reports.* DOI: `https://doi.org/10.1038/s41598-026-52883-z`.	Sostiene la separación entre $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$ en las Secciones 3.1–3.6, junto con el Bloque II de la Sección 4.4; la tesis integra esa contribución con ARS y con la gramática de reporte.

5.5. Correspondencia entre objetivos y evidencia

La lectura integrada de la tesis permite vincular los objetivos específicos con evidencia desarrollada en el cuerpo del manuscrito. Para mantener el orden argumental de la propuesta sin perder verificabilidad, la Tabla 5.1 conserva el número original de cada objetivo y explicita las secciones, tablas y figuras que lo respaldan.

Tabla 5.1: Correspondencia entre objetivos específicos, evidencia principal y alcance de la afirmación.

Objetivo	**Evidencia principal**	**Alcance**
1	La sistematización conceptual se desarrolla en el marco teórico: problema minimal optimal y problema *all-relevant* en la Sección 2.5, medidas de dependencia en la Sección 2.6, contraste nulo en la Sección 2.7, explicabilidad en la Sección 2.11 y causalidad en la Sección 2.12.	Alcance conceptual. La revisión organiza tradiciones que suelen mezclarse bajo el rótulo general de importancia de variables.
2	El diseño operacional de ARS se presenta en la Sección 3.5, con las formulaciones de las Ecuaciones 3.12 y 3.14. El Bloque I evalúa su comportamiento en relaciones bivariadas, conjunto de referencia sin ruido y ruido estructurado, en particular mediante la Tabla 4.4 y la Tabla 4.5.	Alcance metodológico-operacional. ARS entrega evidencia de ventaja frente a una copia nula bajo modelo, métrica y protocolo declarados; su lectura poblacional requiere las condiciones del marco.
3	La formalización de $\mathcal{C}_Y$, $\mathcal{I}$ y $\mathcal{P}$ se desarrolla en la Sección 3.1; la separación entre espacios se muestra mediante la Proposición 3.1 en la Sección 3.2; la relevancia epistémica se articula en la Definición 3.3 y en la Sección 3.4.	Alcance conceptual-formal. Las afirmaciones resultantes ordenan registros de evidencia y delimitan cuándo una lectura observacional puede aspirar a informatividad poblacional.

Objetivo	Evidencia principal	Alcance
4	La condición de protocolo se desarrolla en la Sección 3.3, con apoyo en la Ecuación 3.9. Empíricamente, la Tabla 4.6 y la Figura 4.7 muestran que bootstrap, validación cruzada y realizaciones independientes producen lecturas distintas ante nulos conocidos.	Alcance metodológico. La evidencia sugiere que el protocolo forma parte de la afirmación, especialmente cuando una cantidad operacional puede estabilizarse lejos de la cantidad poblacional buscada.
5	La estrategia de validación controlada aparece en los DGP de la Sección 4.4: nulos conocidos, variación de señal, confusión, causalidad débil, sinergia, intervención simulada y efecto Rashomon. La evidencia central se concentra en la Tabla 4.8, la Tabla 4.9, la Figura 4.9 y la Figura 4.10.	Alcance experimental controlado. Los resultados permiten observar separaciones que serían difíciles de atribuir en datos reales sin verdad de fondo conocida.
6	La gramática de reporte se formula como procedimiento integrado en la Sección 3.7 y se ejemplifica en la discusión integrada de la Sección 4.5, donde una misma variable se reporta con alcance operacional, observacional y causal explícito.	Alcance comunicacional y metodológico. El aporte consiste en hacer verificable qué se afirma, bajo qué protocolo y con qué límite inferencial.

5.6. Trabajo futuro

Una primera línea de trabajo consiste en explorar extensiones multivariadas de ARS. Los tensores *null-swap* constituyen una posible representación de la relevancia contextual. Su formulación, alcance computacional y protocolo de validación se definirán en trabajos posteriores.

Una segunda línea estudiará los protocolos Bayes-consistentes en escenarios prácticos. Un primer análisis podría emplear una clasificación sintética con información mutua condicional poblacional calculable y comparar el contraste obtenido con pérdida logarítmica, F_1 y *accuracy* al variar el tamaño muestral y la intensidad de señal. Un análisis complementario examinaría, una decisión a la vez, la partición de entrenamiento y prueba, el nivel de significancia y el margen práctico de ARS. Los resultados reportarían estabilidad de selección y falsos positivos sobre variables nulas. Trabajos posteriores también podrán caracterizar familias de protocolos para las cuales $\mathcal{P}$ admita condiciones más precisas.

Una tercera línea estudiará la robustez mediante valores atípicos en covariables y respuesta, con distintas prevalencias e intensidades. También examinará el sesgo observado bajo *bootstrap* mediante comparaciones entre remuestreo con reemplazo, submuestreo de filas, duplicación controlada y estimadores alternativos de información mutua.

Una cuarta línea concierne a aplicaciones en dominios científicos con datos reales. Se considera explorar datos de física de partículas vinculados al CERN o datos genómicos. La selección del conjunto, la pregunta sustantiva y el protocolo se realizará de acuerdo con las condiciones de medición y el conocimiento experto del dominio elegido.

Una quinta línea articula este enfoque con métodos causales, ya que la tesis sugiere que la explicabilidad predictiva y la inferencia causal pueden dialogar productivamente cuando se reconoce su diferencia. Desarrollar protocolos que conecten contrastes *null-swap*, invariancias, conocimiento experto e intervenciones permitiría avanzar hacia interpretaciones más sólidas y conservar la flexibilidad del aprendizaje automático moderno.

Apéndices

Apéndice A

Notas metodológicas complementarias

A.1. Trazabilidad y reproducibilidad

Los experimentos de validación descritos en el Capítulo 4 se apoyan en código y datos derivados de acceso público. La trazabilidad se declara por bloque: el Bloque I remite al repositorio asociado al artículo de ARS, mientras que el Bloque II remite al repositorio de fronteras epistémicas. La Tabla A.1 resume los recursos disponibles para revisión y replicación.

Los repositorios permiten auditar implementaciones, semillas, particiones y datos derivados; las conclusiones de la tesis se sostienen, además, en la comparación entre protocolos, en la presencia de variables nulas por construcción y en la separación explícita entre evidencia operacional, informatividad poblacional y causalidad.

Tabla A.1: Reproducibilidad: recursos de reproducibilidad y materiales auxiliares.

Recurso	**Material disponible**	**Uso dentro de la tesis**
Repositorio ARS `github.com/Pneirz/Attribute-Relevance-Score`	Implementación de ARS, cuadernos de ejemplo y datos sintéticos derivados, incluidos los archivos usados para las relaciones bivariadas y el conjunto de referencia con ruido estructurado.	Fuente reproducible del Bloque I; permite auditar la medida, las particiones por semilla y los datos derivados del artículo Neirz et al. (2024).
Repositorio de fronteras epistémicas `github.com/Pneirz/epistemic_frontiers_experiments`	Código, datos derivados y scripts de regeneración de figuras para los experimentos de dependencia del protocolo y disociación epistémica.	Fuente reproducible del Bloque II; permite reejecutar las simulaciones con DGP conocido, los protocolos de remuestreo y los contrastes *null-swap*.
Repositorio diabetes ARS `github.com/Pneirz/ars_diabetes`	Ejercicio exploratorio sobre Pima Indians Diabetes, con contraste *null-swap* de segundo orden y reportes derivados.	Material auxiliar para ilustrar uso en datos reales; queda fuera de la evidencia principal por DGP desconocido, multiplicidad e interpretación no causal.

Apéndice B

Resultados detallados del experimento de intensidad de señal

Este apéndice reporta, para cada protocolo y tamaño muestral, la media y la desviación estándar del contraste *null-swap* $\widetilde{\Delta}$ obtenidas con $B = 100$ evaluaciones en el experimento de dependencia del protocolo con intensidad de señal variable (Sección 4.4.2). En este experimento, $\widetilde{\Delta}$ corresponde a la diferencia entre la información mutua estimada para la columna observada de la variable real y la información mutua estimada para su copia nula, $\widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$. Por ello, valores positivos indican una ventaja informativa estimada de la variable real frente a su copia nula, valores cercanos a cero sugieren indistinguibilidad práctica bajo el protocolo, y valores negativos reflejan variación estimativa o una ventaja accidental de la copia nula.

Cada tabla corresponde a un valor del parámetro α, que controla la intensidad de la señal de X_1 en el proceso generador. Las columnas distinguen la variable informativa X_1 y el grupo "Nulas", formado por Z_1, Z_2 y Z_3, variables intercambiables por construcción porque no tienen relación sistemática con Y. La notación media $\pm$ desviación estándar resume la distribución de los contrastes estimados a través de las $B = 100$ evaluaciones del protocolo indicado. Los valores provienen de los datos publicados en el repositorio indicado en el Apéndice A.

Nota diagnóstica sobre la meseta bajo bootstrap

El patrón más delicado de estas tablas aparece en las variables nulas bajo bootstrap: el contraste medio permanece cercano a 0.10 al aumentar n, mientras su desviación estándar disminuye. Este patrón parece más estable que una alineación accidental en una muestra pequeña. Una explicación plausible relaciona el remuestreo con reemplazo con el estimador de información mutua utilizado. En este experimento, $\widehat{I}$ se calcula mediante el estimador de Ross para variables mixtas discreto-continuas, basado en vecinos cercanos, con $k = 5$ y reducción a mín$(k, n-1)$ en muestras pequeñas, según la implementación de *scikit-learn* (Pedregosa et al., 2011; Ross, 2014). Las remuestras bootstrap contienen filas duplicadas, que modifican las distancias locales del estimador. La copia nula permutada organiza de otro modo el emparejamiento inducido por la duplicación. Bajo esta lectura, el contraste positivo sobre nulos se interpreta como un posible efecto del acoplamiento entre protocolo y estimador. El mecanismo queda como hipótesis y su evaluación requiere comparar remuestreo con reemplazo, submuestreo de filas, duplicación controlada y estimadores de otras familias.

Tabla B.1: Intensidad $\alpha = 0.5$: contraste *null-swap* $\widetilde{\Delta} = \widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$ (media $\pm$ desviación estándar) con $B = 100$ evaluaciones.

	Bootstrap		Validación cruzada		Realizaciones DGP	
n	X_1	Nulas	X_1	Nulas	X_1	Nulas
32	0.096 ± 0.122	0.079 ± 0.107	0.001 ± 0.062	0.001 ± 0.075	0.022 ± 0.087	0.007 ± 0.068
64	0.081 ± 0.081	0.090 ± 0.085	0.008 ± 0.165	-0.007 ± 0.155	0.021 ± 0.057	-0.001 ± 0.041
128	0.105 ± 0.062	0.078 ± 0.057	0.016 ± 0.098	-0.002 ± 0.080	0.021 ± 0.046	0.000 ± 0.030
256	0.120 ± 0.043	0.087 ± 0.042	0.030 ± 0.089	0.000 ± 0.070	0.020 ± 0.028	-0.001 ± 0.021
512	0.110 ± 0.031	0.090 ± 0.030	0.019 ± 0.060	-0.000 ± 0.049	0.024 ± 0.022	0.001 ± 0.015
1024	0.119 ± 0.022	0.097 ± 0.022	0.019 ± 0.044	-0.000 ± 0.035	0.020 ± 0.016	-0.000 ± 0.010
2048	0.123 ± 0.015	0.098 ± 0.016	0.021 ± 0.033	-0.002 ± 0.024	0.025 ± 0.011	0.000 ± 0.007
4096	0.120 ± 0.011	0.098 ± 0.011	0.023 ± 0.023	-0.000 ± 0.017	0.026 ± 0.008	0.000 ± 0.005
8192	0.124 ± 0.007	0.099 ± 0.008	0.024 ± 0.018	-0.000 ± 0.012	0.027 ± 0.006	-0.000 ± 0.004

Tabla B.2: Intensidad $\alpha = 2$: contraste *null-swap* $\widetilde{\Delta} = \widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$ (media $\pm$ desviación estándar) con $B = 100$ evaluaciones.

	Bootstrap		Validación cruzada		Realizaciones DGP	
n	X_1	Nulas	X_1	Nulas	X_1	Nulas
32	0.303 ± 0.161	0.068 ± 0.106	0.005 ± 0.083	0.001 ± 0.077	0.253 ± 0.120	0.004 ± 0.072
64	0.267 ± 0.110	0.086 ± 0.080	0.106 ± 0.216	-0.001 ± 0.158	0.234 ± 0.084	-0.003 ± 0.041
128	0.275 ± 0.069	0.076 ± 0.057	0.146 ± 0.162	-0.001 ± 0.080	0.240 ± 0.058	0.001 ± 0.033
256	0.290 ± 0.046	0.091 ± 0.042	0.236 ± 0.135	0.001 ± 0.065	0.227 ± 0.043	0.002 ± 0.019
512	0.285 ± 0.037	0.096 ± 0.030	0.235 ± 0.091	-0.001 ± 0.049	0.233 ± 0.026	0.002 ± 0.015
1024	0.279 ± 0.026	0.098 ± 0.022	0.219 ± 0.065	-0.000 ± 0.034	0.227 ± 0.020	0.001 ± 0.011
2048	0.290 ± 0.017	0.098 ± 0.015	0.228 ± 0.045	-0.001 ± 0.024	0.228 ± 0.014	0.000 ± 0.007
4096	0.288 ± 0.011	0.096 ± 0.010	0.227 ± 0.032	-0.000 ± 0.017	0.228 ± 0.010	-0.000 ± 0.005
8192	0.295 ± 0.009	0.098 ± 0.007	0.230 ± 0.024	0.000 ± 0.012	0.230 ± 0.007	-0.000 ± 0.004

Tabla B.3: Intensidad $\alpha = 10$: contraste *null-swap* $\widetilde{\Delta} = \widehat{I}(Y; X_j) - \widehat{I}(Y; X_j^{\text{null}})$ (media $\pm$ desviación estándar) con $B = 100$ evaluaciones.

	Bootstrap		Validación cruzada		Realizaciones DGP	
n	X_1	Nulas	X_1	Nulas	X_1	Nulas
32	0.509 ± 0.104	0.073 ± 0.115	0.016 ± 0.107	0.001 ± 0.075	0.539 ± 0.076	-0.003 ± 0.061
64	0.563 ± 0.066	0.080 ± 0.083	0.263 ± 0.242	-0.002 ± 0.160	0.559 ± 0.063	0.001 ± 0.043
128	0.551 ± 0.051	0.082 ± 0.059	0.371 ± 0.142	-0.002 ± 0.079	0.566 ± 0.045	0.000 ± 0.033
256	0.576 ± 0.039	0.086 ± 0.041	0.510 ± 0.098	0.002 ± 0.065	0.562 ± 0.030	0.000 ± 0.021
512	0.565 ± 0.026	0.092 ± 0.031	0.549 ± 0.071	-0.000 ± 0.049	0.563 ± 0.023	0.001 ± 0.017
1024	0.578 ± 0.017	0.095 ± 0.021	0.563 ± 0.048	0.001 ± 0.034	0.562 ± 0.016	0.000 ± 0.010
2048	0.577 ± 0.015	0.096 ± 0.015	0.562 ± 0.036	-0.001 ± 0.024	0.563 ± 0.012	-0.000 ± 0.007
4096	0.575 ± 0.008	0.096 ± 0.010	0.560 ± 0.024	-0.000 ± 0.017	0.563 ± 0.007	-0.001 ± 0.005
8192	0.580 ± 0.007	0.098 ± 0.008	0.562 ± 0.017	0.000 ± 0.012	0.564 ± 0.006	-0.001 ± 0.004

Bibliografía

AL-Gburi, A. F. J., Nazri, M. Z. A., Yaakub, M. R. B., and Alyasseri, Z. A. A. (2024). Multi-objective unsupervised feature selection and cluster based on symbiotic organism search. *Algorithms*, 17:355.

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., and Furlanello, C. (2013). minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, 29(3):407–408.

Albanese, D., Riccadonna, S., Donati, C., and Franceschi, P. (2018). A practical tool for maximal information coefficient analysis. *GigaScience*, 7:giy032.

Arenas, M., Barceló, P., Bustamante, D., Caraball, J., and Subercaseaux, B. (2024). A uniform language to explain decision trees. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 60–70.

Arenas, M., Barceló, P., Kozachinskiy, A., Romero, M., and Subercaseaux, B. (2025). On computing probabilistic explanations for decision trees. *Journal of Artificial Intelligence Research*.

Arunika, M., Saranya, S., Charulekha, S., Kabilarajan, S., et al. (2024). A survey on explainable AI using machine learning algorithms SHAP and LIME. In *2024 15th International Conference on Computing Communication and Networking Technologies*.

Baak, M., Koopman, R., Snoek, H., and Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2:31.

Biecek, P. and Samek, W. (2024). Position: Explain to question not to justify. In *Proceedings of the 41st International Conference on Machine Learning*, pages 3996–4006.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., et al. (2021). Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769.

Bouthillier, X., Laurent, C., and Vincent, P. (2019). Unreproducible research is reproducible. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734. PMLR.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, Z. and Zhang, W. (2013). Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLOS Computational Biology*, 9(3):e1002956.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15:9383.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10:392–404.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, 2 edition.

D'Amour, A. (2021). Revisiting Rashomon: A comment on "the two cultures". *Observational Studies*, 7(1):59–63.

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2):492–503.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 123–132.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Donick, D. and Lera, S. C. (2021). Uncovering feature interdependencies in high-noise environments with stepwise lookahead decision forests. *Scientific Reports*, 11:9238.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience, 2 edition.

Durasevic, M., Jakobovic, D., Picek, S., and Mariot, L. (2024). Assessing the ability of genetic programming for feature selection in constructing dispatching rules for unrelated machine environments. *Algorithms*, 17:67.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Ehsan, U. and Riedl, M. O. (2023). Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Gammaitoni, L., Hänggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1):223–287.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63:3–42.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hammerton, G. and Munafò, M. R. (2021). Causal inference with observational data: The need for triangulation of evidence. *Psychological Medicine*, 51(4):563–578.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Hwang, H., Bell, A., Fonseca, J., Pliatsika, V., Stoyanovich, J., and Whang, S. E. (2025). SHAP-based explanations are sensitive to feature representation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1588–1601.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Kang, I. A., Njimbouom, S. N., and Kim, J. D. (2023). Optimal feature selection-based dental caries prediction model using machine learning for decision support system. *Bioengineering*, 10:245.

Kiratsoudis, S. and Tsiantos, V. (2024). Enhancing personnel selection through the integration of the entropy synergy analysis of multi-attribute decision making model: A novel approach. *Information*, 15:1.

Knab, P., Marton, S., Schlegel, U., and Bartelt, C. (2025). Which LIME should i trust? concepts, challenges, and solutions. In *Explainable Artificial Intelligence*. Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2:1137–1143.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and A/B testing. In *Encyclopedia of Machine Learning and Data Mining*. Springer.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Luedtke, A. and Tran, L. H. (2013). The generalized mean information coefficient. arXiv preprint arXiv:1308.5712.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Mathchi (2020). Diabetes data set. Kaggle. Pima Indians Diabetes Database; licencia CC0.

Miller, A. C., Foti, N. J., and Fox, E. B. (2021). Breiman's two cultures: You don't have to choose sides. *Observational Studies*, 7:161–169.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mnich, K. and Rudnicki, W. R. (2020). All-relevant feature selection using multidimensional filters with exhaustive search. *Information Sciences*, 524:277–297.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2 edition.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Nature Machine Intelligence (2025). Are neural network representations universal or idiosyncratic? *Nature Machine Intelligence*, 7:1589–1590.

Neirz, P., Allende, H., and Saavedra, C. (2024). Attribute relevance score: A novel measure for identifying attribute importance. *Algorithms*, 17(11):518. DOI: `https://doi.org/10.3390/a17110518`.

Neirz, P., Allende, H., and Saavedra, C. (2026). Epistemic frontiers and the distinction between causality, information, and predictability in pattern recognition. *Scientific Reports*. Published online 13 May 2026. DOI: `https://doi.org/10.1038/s41598-026-52883-z`.

Nilsson, R., Peña, J. M., Björkegren, J., and Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8:589–612.

Olivetti, E., Zoonekynd, V., Yam, P., López de Prado, M., Imbens, G. W., and Hernán, M. A. (2026). Can AI learn causal structure? evidence from ADIA lab's causal discovery challenge. SSRN Electronic Journal.

Passemiers, A., Folco, P., Raimondi, D., Birolo, G., Moreau, Y., and Fariselli, P. (2024). A quantitative benchmark of neural network feature selection methods for detecting nonlinear signals. *Scientific Reports*, 14:31180.

Pathak, A. K., Gupta, M., and Jain, G. (2026). Unmasking the clever hans effect in AI models: Shortcut learning, spurious correlations, and the path toward robust intelligence. *Frontiers in Artificial Intelligence*, 8:1692454.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Pezoa, R., Salinas, L., and Torres, C. (2023). Explainability of high energy physics events classification using SHAP. In *Journal of Physics: Conference Series*, volume 2438, page 012082.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Ridley, M. (2025). Human-centered explainable artificial intelligence: An annual review of information science and technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1):98–120.

Ríos-Vásquez, G., de la Fuente-Mella, H., and Ceroni-Díaz, J. (2025). Group-specific SVM with bilevel programming methods for parameter optimization and explainable AI in urban quality of life prediction. *IEEE Access*, 13:130475–130491.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3:199–217.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):e87357.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Semenova, L., Rudin, C., and Parr, R. (2022). On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858. Association for Computing Machinery.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265. IEEE Computer Society Press.

Sokol, K. and Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2 edition.

Stoppiglia, H., Dreyfus, G., Dubois, R., and Oussar, Y. (2003). Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Ullah, S., Mahmood, Z., Ali, N., Ahmad, T., and Buriro, A. (2023). Machine learning-based dynamic attribute selection technique for DDoS attack classification in IoT networks. *Computers*, 12:115.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Vowels, M. J. (2024). Trying to outrun causality with machine learning: Limitations of model explainability techniques for exploratory research. *Psychological Methods*.

Wetschoreck, F. (2020). RIP correlation: Introducing the predictive power score. Towards Data Science.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Zhai, Z.-M., Kong, L.-W., and Lai, Y.-C. (2023). Emergence of a resonance in machine learning. *Physical Review Research*, 5(3):033127.