Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Trabajo Fin de Grado

Implementación de técnicas de inferencia causal basadas en redes neuronales para la estimación de efectos de tratamiento

Francisco Javier Gómez Fernández-Getino

24 de junio de 2025

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Trabajo Fin de Grado

Título: Implementación de técnicas de inferencia causal basadas en redes neuronales para la estimación de efectos de tratamientos.

Autor: D. Francisco Javier Gómez Fernández-Getino.

Tutor: Dª. Patricia Alonso de Apellániz.

Ponente: D. Alejandro Almodovar Espeso

Departamento: Señales, sistemas y radiocomunicaciones

Miembros del Tribunal

Presidente: D. . . . . .

Vocal: D. . . . . ..

Secretario: D. . . . . ..

Suplente: D. . . . . ..

Los miembros del tribunal arriba nombrados acuerdan otorgar la calificación de: . . . . . .

Madrid, a de de 20...

Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Trabajo Fin de Grado

Implementación de técnicas de inferencia causal basadas en redes neuronales para la estimación de efectos de tratamiento

Francisco Javier Gómez Fernández-Getino

24 de junio de 2025

Resumen

Este trabajo surge de la necesidad de tomar decisiones informadas en inferencia causal, especialmente en contextos donde los métodos estadísticos no causales son poco prácticos o éticamente inviables. Aunque los ensayos controlados aleatorios (RCT) ofrecen una solución robusta para determinar efectos causales, su implementación se ve limitada por altos costes, largos tiempos de ejecución y consideraciones éticas, lo que dificulta su aplicación en áreas como la salud.

Frente a estas limitaciones, se propone el uso de métodos de análisis causales para estimar efectos de tratamiento a partir de datos observacionales, ofreciendo una alternativa viable en situaciones donde no es posible realizar experimentos aleatorizados. Más concretamente, se diseñan e implementan modelos basados en redes neuronales para estimar los efectos de tratamientos, analizando distintos enfoques: meta-learners como el S-learner y el T-learner, así como métodos de ajuste causal más clásicos que, mediante ajustes probabilísticos, corrigen sesgos en la asignación del tratamiento.

Para validar estos modelos se utilizan conjuntos de datos reconocidos como punto de referencia en inferencia causal. Se ha desarrollado de principio a fin todo el proceso de validación, que incluye la estandarización de variables, la partición estratificada de los datos y el entrenamiento conjunto de submodelos. De este modo, podemos evaluar de forma rigurosa la robustez y generalización de cada técnica.

Finalmente, se compara la precisión y efectividad de cada técnica, identificando sus ventajas y limitaciones. El objetivo principal es verificar que estos métodos de inferencia causal sean fiables y puedan aplicarse con éxito a datos reales, lo que permitirá mejorar la toma de decisiones en entornos complejos.

Summary

This work arises from the need to make informed decisions in causal inference, especially in contexts where non-causal statistical methods are impractical or ethically unfeasible. Although randomized controlled trials (RCTs) offer a robust solution for determining causal effects, their implementation is limited by high costs, long execution times, and ethical considerations, which hinders their application in areas such as health.

Given these limitations, the use of causal analysis methods is proposed to estimate treatment effects from observational data, offering a viable alternative in situations where it is not possible to conduct randomized experiments. More precisely, neural-network-based models are designed and implemented to estimate treatment effects, analyzing different approaches: meta-learners such as the S-learner and the T-learner, as well as more classical causal-adjustment methods that, through probabilistic adjustments, correct biases in treatment assignment.

To validate these models, datasets recognized as benchmarks in causal inference are used.

The entire validation process has been developed end-to-end, including variable standardization, stratified data splitting, and joint training of submodels. In this way, we can rigorously evaluate the robustness and generalization of each technique.

Finally, the accuracy and effectiveness of each technique is compared, identifying their advantages and limitations. The main objective is to verify that these causal inference methods are reliable and can be successfully applied to real-world data, which will allow improving decision-making in complex environments.

Palabras Clave

Inferencia causal; potencial outcomes; meta-learners; Dragonnet; Targeted Regularization; IHDP; estimación de ATE; PEHE; ATT; IPW; métodos doblemente robustos.

Keywords

Causal inference; potential outcomes; meta-learners; Dragonnet; Targeted Regularization; IHDP; ATE estimation; PEHE; ATT; IPW; doubly-robust methods.

Agradecimientos

Quiero expresar, en primer lugar, mi más sincero agradecimiento a Alejandro Almodóvar, por su apoyo constante, dedicación y su excelente dirección a lo largo de este trabajo.

A Patricia Alonso, por su ayuda y sugerencias, que han sido relevantes para el desarrollo y la mejora del trabajo.

A Santiago Zazo, por su valiosa orientación y apoyo, así como por su disposición constante, y a Juan Parras, por su asesoramiento durante mi formación previa, que ha sido de gran utilidad para la realización de este trabajo.

Y, de manera muy especial, a mis padres, Ana Patricia y Luis, y a mi hermano Luis Alberto, por todo su cariño, su plena confianza en mí y su apoyo incondicional.

Índice

Índice de figuras

Índice de tablas

Lista de acrónimos

ATE *Average Treatment Effect* (Efecto promedio de tratamiento)

ATT *Average Treatment effect on the Treated* (Efecto promedio en tratados)

PEHE *Precision in Estimation of Heterogeneous Effect* (Precisión en estimación de efectos individuales)

IPW *Inverse Probability Weighting* (Ponderación por probabilidad inversa)

DR *Doubly-Robust* (Estimador doblemente robusto)

S-learner *Single-learner* (Meta-learner único)

T-learner *Two-learner* (Meta-learner por tratamiento)

T-REG *Targeted Regularization* (Regularización dirigida)

IHDP *Infant Health and Development Program* (Conjunto de datos semi-sintético)

TARNet *Treatment-agnostic Representation Network* (Red de representación agnóstica al tratamiento)

CFRNet *Counterfactual Regression Network* (Red de regresión contrafactual)

SUTVA *Stable Unit Treatment Value Assumption* (Supuesto de tratamiento estable sin interferencia)

1. Introducción y objetivos

1.1. Introducción

La inferencia causal es el proceso para determinar el efecto de una intervención, tratamiento o política sobre un resultado de interés. A diferencia de la mera correlación, que identifica asociaciones estadísticas entre variables, la inferencia causal pretende establecer relaciones de causa-efecto, es decir, cuantificar qué sucedería con el resultado si se cambia deliberadamente el valor de una variable (el "tratamiento") en un sistema real o simulado.

En la práctica, la asignación de tratamientos rara vez se hace de forma aleatoria. Por ejemplo, en medicina, los pacientes más graves pueden recibir un fármaco experimental, y en economía, las políticas sociales tienden a aplicarse en zonas con mayores necesidades. Esta falta de aleatorización introduce *sesgos de selección* y *confusión* (confounding): las covariables que afectan simultáneamente al tratamiento y al resultado distorsionan la estimación del efecto causal. Además, en muchos ámbitos (salud, educación, justicia social) implementar ensayos aleatorizados puede ser éticamente inviable o extremadamente costoso, lo que limita la obtención de datos fiables a través de estudios controlados.

1.2. Objetivos

El propósito de este trabajo es abordar los principales retos que plantea la estimación de efectos causales a partir de datos observacionales, en particular los sesgos de selección y la confusión de ciertas covariables. Para ello, se estudiarán distintas estrategias para mitigar el sesgo de confounding y reducir el impacto de las covariables en la estimación del efecto causal del tratamiento. Las pruebas se llevarán a cabo con datos clínicos de pacientes semi-sintéticos. Estas estrategias incluirán métodos basados en ponderaciones, emparejamientos y ajustes predictivos que permiten equilibrar la distribución de características entre los grupos de tratamiento y control. Se prestará atención a procesos de preprocesado que faciliten la comparación, como la selección de covariables relevantes y técnicas de normalización. El objetivo es construir un conjunto de modelos versátiles que puedan adaptarse a diferentes tipos de datos y niveles de complejidad.

A su vez, cada modelo se evaluará de forma cuantitativa usando ciertas métricas que reflejen la cercanía entre las estimaciones de efecto y la realidad simulada. Además, se analizará la estabilidad de los resultados ante variaciones en la asignación del tratamiento, asegurando que las correcciones propuestas funcionen de manera consistente incluso en escenarios de desequilibrio extremo. De este modo, se obtendrá una visión práctica y comparativa de las fortalezas y limitaciones de cada técnica.

2. Marco teórico

En este capítulo se revisará con detalle los fundamentos de la inferencia causal, basándonos en gran medida en Neal (2020) [1]. Se cubrirá:

- Una breve introducción con la Paradoja de Simpson y la diferencia entre asociación y causalidad.
- El marco de resultados potenciales y el problema fundamental de la inferencia causal.
- Los supuestos de ignorabilidad, intercambiabilidad y estrategias para sortear el problema fundamental.
- Definición de terminología estadística relevante (backdoor adjustment).

2.1. Contexto

La inferencia causal es una rama de la estadística y las ciencias de datos que busca responder preguntas del tipo "¿Qué habría pasado si...?". A diferencia del análisis puramente correlacional, la inferencia causal pretende establecer relaciones de causa y efecto entre variables, permitiendo así evaluar el impacto de diferentes medidas, como intervenciones, políticas públicas o tratamientos médicos entre otros. Esta capacidad es especialmente crucial en contextos y situaciones donde realizar experimentos aleatorizados no es viable o es muy complicado, como en medicina, economía o ciertos campos de las ciencias sociales.

Para comprender el origen de los desafíos y paradojas en la inferencia causal, este primer capítulo comienza con uno de los ejemplos más ilustrativos de los errores que se pueden cometer al no ajustar correctamente por factores de confusión en el establecimiento de relaciones de causa: la Paradoja de Simpson.

2.1.1. Paradoja de Simpson

La Paradoja de Simpson se presenta cuando la tendencia de un efecto en subgrupos difiere de la tendencia global. Consideremos el caso hipotético de un nuevo tratamiento para una nueva enfermedad conocida como COVID-27, y dos posibles tratamientos para reducir su tasa de mortalidad:

Condición	Tratamiento A	Tratamiento B
Leve	15 % (210/1400)	10 % (5/50)
Grave	30 % (30/100)	20 % (100/500)
Total	16 % (240/1500)	19 % (105/550)

Tabla 2.1: Paradoja de Simpson observada en los datos de COVID-27, donde la tendencia conjunta difiere de las tendencias por subgrupo. Lo que se muestra es la tasa de mortalidad de los pacientes en función del tratamiento

Globalmente, el tratamiento A reduce la mortalidad al 16 % frente al 19 % del tratamiento B; sin embargo, al segmentar por condición, B es mejor tanto en pacientes leves (10 % vs. 15 %) como graves (20 % vs. 30 %). Esta contradicción surge porque la condición del paciente confunde la comparación: los pacientes graves, con mayor mortalidad de base, reciben con más frecuencia el tratamiento B. Un análisis causal adecuado que ajusta en este caso por condición revela que B es superior en ambos subgrupos y resuelve la aparente paradoja [1].

2.1.2. Aplicaciones de la inferencia causal

La capacidad de predecir intervenciones es esencial en diferentes campos como:

- **Medicina:** casos donde determinar qué fármaco *causa* mayor supervivencia sin ningún sesgo de selección [2].
- **Políticas públicas:** elegir la política que *provoca* la mayor reducción de emisiones contaminantes o pobreza, por ejemplo. [3]
- **Economía y ciencias sociales:** evaluar el impacto *causal* de la educación o programas sociales en resultados económicos y sociales [4].

Sin un enfoque causal, las decisiones basadas solo en correlaciones pueden resultar ineficaces o contraproducentes, como se ha comprobado con el caso de la Paradoja de Simpson en el apartado anterior [1].

2.1.3. De la asociación a la causalidad

El objetivo del análisis estadístico convencional es estimar parámetros de una distribución a partir de muestras extraídas de ella. Con esos parámetros se infieren asociaciones entre variables, lo que permite calcular probabilidades de eventos pasados o futuros y actualizar dichas probabilidades al incorporar nueva información. El análisis estadístico gestiona adecuadamente estas tareas siempre que las condiciones experimentales permanezcan constantes.

El análisis causal va un paso más allá: pretende inferir probabilidades bajo condiciones que cambian, por ejemplo, al aplicar diferentes tratamientos o intervenciones externas. A diferencia de la pura estadística, que se limita a describir la relación entre variables en el contexto observado, el análisis causal requiere información adicional sobre cómo se comportaría el sistema si alteramos esas condiciones como por ejemplo, al pasar de un escenario observacional a un experimento controlado. Esa información no está contenida

en la distribución de probabilidad en sí misma; hace falta introducir suposiciones causales que especifiquen qué relaciones permanecen invariables ante esos cambios.

Para ilustrar esta idea, consideremos un ejemplo clásico: durante los meses de verano aumenta simultáneamente la venta de helados y el número de ahogamientos en piscinas [5]. Si nos quedamos con la estadística pura, observamos que

$$P\big(\text{ahogamiento} \mid \text{ventas de helados altas}\big) > P\big(\text{ahogamiento} \mid \text{ventas de helados bajas}\big).$$

Sin embargo, sería erróneo concluir que "comprar más helados provoca más ahogamientos". En realidad, existe una variable oculta (la temperatura) que eleva al mismo tiempo el consumo de helados y la afluencia a las piscinas, explicando así la correlación espuria que existía.

Este ejemplo destaca dos puntos fundamentales:

- Una afirmación *asociacional* como

$$P\big(\text{ahogamiento} \mid \text{ventas de helados}\big)$$

  describe simplemente cómo se relacionan los eventos en los datos observados, sin indicar qué sucedería si se cambiara la variable "venta de helados".

- Una afirmación *causal* requiere un formalismo de intervención, por ejemplo

$$P\big(\text{ahogamiento} \mid \text{do}(\text{ventas de helados} = \text{alta})\big),$$

  que representa la probabilidad de ahogamiento en el escenario hipotético donde forzamos un nivel alto de ventas de helados y eliminamos cualquier influencia previa sobre esa variable.

En definitiva, mientras que las herramientas de la estadística tradicional (condicionalización, regresión, etc.) sirven para describir asociaciones, el análisis causal nos permite razonar inequívocamente sobre intervenciones y distinguir "observar X" de "modificar X".

Para poder distinguir dependencia estadística de dependencia estrictamente causal, es necesario introducir nueva notación (por ejemplo, el operador $\text{do}(\cdot)$ [6]). Más adelante presentaremos esta sintaxis de causalidad, y en particular cómo utilizar $P(Y \mid \text{do}(X))$.

2.2. Potential Outcomes

Tras haber destacado la distinción esencial entre asociación y causalidad, y haber visto por qué la mera distribución conjunta de las variables no aporta información sobre el efecto de intervenciones externas [6], se necesita un marco formal que permita definir y razonar acerca de estos efectos causales.

Aunque la probabilidad condicional tradicional no permite expresar intervenciones de forma explícita, existen extensiones formales, como el modelo de resultados potenciales (*potential outcomes*) [7], que permiten construir una teoría causal sobre bases probabilísticas. En esta sección introducimos dicho marco, que aunque se apoya en expectativas y variables aleatorias, permite distinguir claramente entre asociaciones y efectos causales mediante variables contrafactuales.

Antes de entrar en los resultados potenciales, fijemos la notación que utilizaremos a lo largo de este capítulo. Usaremos letras mayúsculas para variables aleatorias y minúsculas para sus realizaciones. En particular:

- T denota la variable aleatoria de tratamiento (usualmente binaria, $T \in \{0, 1\}$).
- Y denota la variable aleatoria de resultado u observación de interés.
- X representa el vector de covariables o atributos de cada unidad.

Aunque en gran parte de nuestro desarrollo T será binario, todos los conceptos se pueden extender sin dificultad a tratamientos con más de dos niveles o incluso continuos.

2.2.1. Resultados potenciales y efecto individual de tratamiento

En el modelo de resultados potenciales (o *potential outcomes*) para cada unidad i de la población se definen dos cantidades hipotéticas [7, 8, 1]:

$Y_i(T = 1)$ = resultado que habríamos observado si la unidad i recibiera el tratamiento,

$Y_i(T = 0)$ = resultado que habríamos observado si la unidad i no recibiera el tratamiento.

De estos se construye el *efecto individual de tratamiento* (ITE):

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

Dado que para cada unidad únicamente registramos uno de los dos resultados (el otro queda como contrafactual), nunca podemos medir directamente τ_i. Sin embargo, esta formulación permite caracterizar con precisión qué entendemos por "efecto causal" y sirve de base para los estimandos promedio.

2.2.2. El problema fundamental de la inferencia causal

El *problema fundamental de la inferencia causal* afirma que es imposible observar simultáneamente ambos potenciales $Y_i(1)$ y $Y_i(0)$ para la misma unidad [8]. Uno de ellos permanece siempre inobservable (contrafactual), convirtiendo el cálculo de τ_i en un problema de datos faltantes.

En las siguientes secciones se presentarán los supuestos y métodos (ignorabilidad, ajuste por covariables, ponderación, etc.) que permiten sortear este obstáculo y estimar los efectos promedio de tratamiento a partir de datos observacionales.

2.3. Sortear el problema fundamental

Se ha visto que nunca es posible observar simultáneamente ambos resultados potenciales para una misma unidad, y por tanto tampoco su efecto individual. No obstante, nuestro interés suele centrarse en estimar efectos promedios, lo cual podemos conseguir si adoptamos ciertas suposiciones y tratamientos de los datos faltantes.

2.3.1. Average Treatment Effect (ATE) e interpretación de datos faltantes

Aunque el *efecto individual* $\tau_i = Y_i(1) - Y_i(0)$ permanezca inobservable, podemos definir el *efecto promedio de tratamiento* (ATE) como

$$\text{ATE} \;=\; \mathbb{E}\big[Y(1) - Y(0)\big] \;=\; \mathbb{E}[Y(1)] \;-\; \mathbb{E}[Y(0)] \tag{2}$$

El problema es que, en los datos reales, cada fila de la tabla ilustrativa 2.2 presenta uno de los dos valores $Y_i(1)$ ó $Y_i(0)$ como "?", de modo que el cálculo directo de la media sobre ambas columnas se topa con datos faltantes.

i	T_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	0	0	?	0
2	1	1	1	?
3	1	1	1	?
4	0	1	?	1
5	0	0	?	0
6	1	0	0	?

Tabla 2.2: Datos de ejemplo utilizados para ilustrar los conceptos de resultados potenciales. Los signos "?" indican los resultados que no son observables para cada individuo.

Un enfoque erróneo sería tomar la diferencia de medias condicionales

$$\mathbb{E}[Y \mid T = 1] \;-\; \mathbb{E}[Y \mid T = 0],$$

porque esa cantidad refleja únicamente asociación y no causalidad, ya que existe un sesgo de confusión (*confounding*) con un flujo de asociación no causal como muestra el camino $T \leftarrow X \rightarrow Y$ en la figura 2.1.

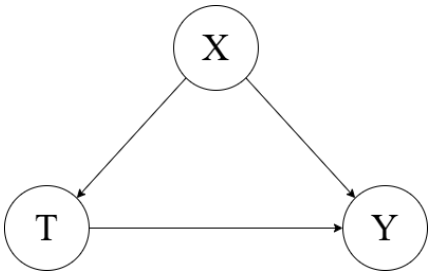

Figura 2.1: Estructura causal donde la variable X actúa como un confusor al influir tanto en el tratamiento T como en el resultado Y.

Solo cuando se satisfacen los supuestos de *ignorabilidad* (que se garantiza únicamente en los RCT) o, más realista, de *ignorabilidad condicional*, esa diferencia asociacional coincide con el ATE. A continuación se verá cómo formalizar y justificar esas condiciones para poder "imputar" los valores faltantes y obtener una estimación válida del ATE.

2.3.2. Ignorabilidad e intercambiabilidad

La única manera de justificar que la sencilla diferencia de medias

$$\mathbb{E}[Y \mid T = 1] \; - \; \mathbb{E}[Y \mid T = 0]$$

sea igual al efecto causal promedio es suponer que, una vez atribuida la asignación del tratamiento, los resultados que no observamos (los contrafactuales) pueden tratarse como si no existiesen sesgos sistemáticos. A esta hipótesis se la denomina *ignorabilidad*. En lenguaje de grafos causales, equivale a eliminar cualquier flecha que conecte covariables X con la asignación T, de modo que la única dependencia relevante sea $T \to Y$.

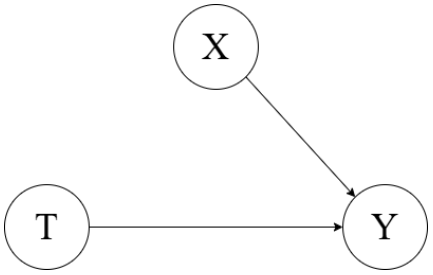

Figura 2.2: Escenario donde la covariable X no afecta al tratamiento T al asumir ignorabilidad.

Formalmente, la ignorabilidad exige que los dos resultados potenciales cumplan lo siguiente:

$$(Y(1),\, Y(0)) \;\perp\!\!\!\perp\; T.$$

Bajo esta condición, se demuestra que

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0].$$

Es decir, el ATE coincide con la diferencia de expectativas observadas en los grupos tratados y de control.

La forma equivalente de esta idea se conoce como *intercambiabilidad de medias*: intercambiar los conjuntos de tratamiento y control no modifica los valores esperados de los resultados potenciales. De manera explícita,

$$\mathbb{E}[Y(1) \mid T = 0] = \mathbb{E}[Y(1) \mid T = 1], \quad \mathbb{E}[Y(0) \mid T = 1] = \mathbb{E}[Y(0) \mid T = 0],$$

lo cual implica de forma inmediata que la diferencia de medias observadas es igual al verdadero ATE.

Decimos que un estimando causal es *identificable* cuando puede expresarse únicamente en términos de la distribución conjunta de las variables observadas (X, T, Y), sin referencia alguna a cantidades contrafactuales no observables.

En la práctica de los estudios observacionales, raramente se cumple la ignorabilidad absoluta, porque casi siempre existen covariables que afectan simultáneamente a T y a Y. Por esta razón, se recurre a la versión más débil, conocida como *ignorabilidad condicional*, que exige independencia entre $\{Y(1), Y(0)\}$ y T una vez que se han fijado ciertos valores de X. Esta última hipótesis será la que se utilizará a lo largo de este trabajo.

2.3.3. Intercambiabilidad condicional y ausencia de confusión

En datos observacionales no cabe esperar que los grupos de tratamiento y control sean directamente comparables en todas las características relevantes. Antes de ajustar, no hay nada que garantice que difieran únicamente en el tratamiento recibido. No obstante, si condicionamos sobre un conjunto adecuado de covariables X, es plausible que dentro de cada nivel de X las unidades sí resulten homogéneas salvo por el tratamiento. A este requisito se le conoce como *intercambiabilidad condicional* o *ausencia de confusión*.

Formalmente se puede expresar de la siguiente manera:

$$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \mid X$$

Bajo este supuesto, cualquier asociación entre T y los resultados potenciales desaparece una vez fijamos X. Gráficamente en las siguientes figuras se observa como la vía de confusión $T \leftarrow X \rightarrow Y$ queda "bloqueada" al condicionar sobre X.

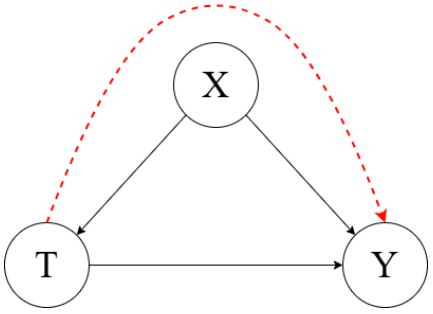

Figura 2.3: Flujo de asociación no causal entre T y Y inducido por la variable confusora X.

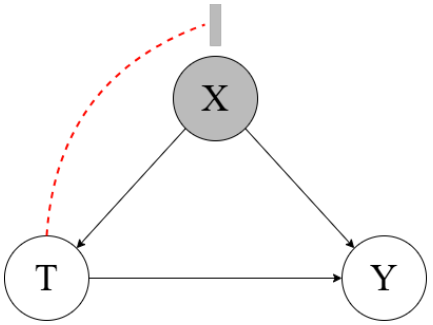

Figura 2.4: Condicionar por la variable X elimina el sesgo por confusión, bloqueando la vías de confusión entre T e Y.

Esta hipótesis permite derivar la fórmula de ajuste, que identifica el efecto promedio de tratamiento a partir de la distribución observacional:

$$\mathbb{E}\big[Y(1) - Y(0)\big] = \mathbb{E}_X\big[\mathbb{E}[Y \mid T = 1, X] \; - \; \mathbb{E}[Y \mid T = 0, X]\big] \tag{3}$$

Con la intercambiabilidad condicional como requisito fundamental, queda por definir bajo qué condiciones adicionales se puede garantizar que todas las combinaciones de tratamiento y covariables tengan representación suficiente en los datos. En la siguiente sección

analizaremos precisamente este aspecto clave, conocido como *positividad* o *overlap*, y su relación con la capacidad de extrapolar de manera fiable los efectos estimados.

2.3.4. Positividad / Solapamiento y Extrapolación

Al incorporar múltiples covariables con el fin de lograr la ausencia de confusión, surge un nuevo requisito: garantizar que cada combinación de tratamiento y características tenga representación en los datos observacionales. Esta asunción se conoce como *positividad* o *overlap*. En el caso de un tratamiento binario, la condición se enuncia de la siguiente manera:

Positividad (o solapamiento / soporte común). Para todo valor de covariables $\mathbf{x}$ que aparezca en la población de estudio (es decir, tal que $P(\mathbf{X} = \mathbf{x}) > 0$), debe cumplirse

$$0 \;<\; P(T = 1 \mid \mathbf{X} = \mathbf{x}) \;<\; 1\,.$$

Esta restricción impide que exista algún subgrupo con covariables $\mathbf{x}$ en el cual todos reciban siempre el mismo tratamiento (o sean todos del grupo de control). Si en alguna región del espacio de covariables $P(T = 1 \mid \mathbf{X} = \mathbf{x}) = 0$ o $P(T = 0 \mid \mathbf{X} = \mathbf{x}) = 0$, la fórmula de ajuste

$$\mathbb{E}\big[Y(1) - Y(0)\big] = \mathbb{E}_{\mathbf{X}}\big[\mathbb{E}[Y \mid T = 1, \mathbf{X}] - \mathbb{E}[Y \mid T = 0, \mathbf{X}]\big]$$

no puede evaluarse, pues en el denominador de alguno de sus términos aparecería un cero.

Desde un punto de vista intuitivo, la positividad equivale a exigir que las distribuciones de covariables en los grupos tratado y de control se solapen completamente. Todo valor de $\mathbf{X}$ observado bajo tratamiento también debe observarse, con probabilidad positiva, bajo control, y viceversa. En ausencia de este solapamiento, cualquier estimador que intente predecir $\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}]$ acabará extrapolando fuera de la región poblada de datos reales, lo que suele inducir sesgos enormes e inestabilidad en la estimación causal.

Por ello, en la práctica existe un compromiso entre la riqueza de covariables que incluimos para controlar el sesgo de confounding y el riesgo de fragmentar excesivamente la muestra hasta violar la positividad (a menudo ligado a la mal llamada "maldición de la dimensionalidad"). En la sección siguiente abordaremos estrategias para paliar este problema cuando las dimensiones de $\mathbf{X}$ son elevadas.

2.3.5. Ausencia de interferencia, consistencia y SUTVA

Además de las condiciones ya expuestas, existen dos supuestos más que se han ido dando por sentado a lo largo de este capítulo. A continuación se formulan de manera explícita:

Ausencia de interferencia. El resultado observado para la unidad i no depende del tratamiento asignado a ninguna otra unidad, sino únicamente de su propio tratamiento. Formalmente, si $\mathbf{t} = (t_1, \ldots, t_n)$ denota el vector de tratamientos de todas las unidades,

$$Y_i(t_1, \ldots, t_{i-1}, t_i, t_{i+1}, \ldots, t_n) \;=\; Y_i(t_i).$$

Esta condición impide que exista influencia cruzada, como por ejemplo, que la felicidad de un individuo dependa de si sus amigos tienen perro.

Consistencia. El valor realmente observado Y_i coincide con el resultado potencial correspondiente al tratamiento efectivamente recibido T_i. Es decir, si para la unidad i se observa $T_i = t$, entonces

$$Y_i = Y_i(t),$$

o de forma equivalente,

$$Y_i = Y_i(T_i).$$

La consistencia garantiza que no existan múltiples versiones no distinguibles del tratamiento dentro de la misma categoría (por ejemplo, "adoptar un perro" podría descomponerse en distintas edades o razas que producirían efectos diferentes).

Estos dos supuestos, junto con la no-confusión condicional y la positividad, conforman la llamada *Stable Unit Treatment Value Assumption* (SUTVA). En la práctica, SUTVA equivale a exigir (1) que no haya interferencia entre unidades, (2) que el tratamiento esté bien definido (no existan versiones ambiguas), y (3) que los resultados potenciales sean deterministas dado el tratamiento.

En las demostraciones anteriores, para reemplazar expresiones con potenciales $Y(1)$ y $Y(0)$ por cantidades que sólo involucren el Y observado, hemos hecho implícito uso de las dos ideas adicionales recién comentadas (Consistencia y ausencia de interferencia).

Gracias a estas dos condiciones se puede sustituir, por ejemplo, $\mathbb{E}[\,Y(1) \mid T = 1\,]$ por $\mathbb{E}[\,Y \mid T = 1\,]$, y lo mismo para el caso $T = 0$. De este modo, junto con la intercambiabilidad condicional y la positividad, ya contamos con todos los supuestos necesarios para llevar la definición causal del ATE a una fórmula que sólo depende de cantidades observables en los datos.

2.3.6. Síntesis de asunciones

A lo largo de esta sección se han presentado los supuestos esenciales que permiten pasar de la definición causal del efecto promedio

$$\text{ATE} = \mathbb{E}\big[Y(1) - Y(0)\big]$$

a una expresión estimable directamente en términos de la distribución observacional. Los cuatro pilares de esta identificación son:

1. *Intercambiabilidad condicional* (unconfoundedness):

   $$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

   que garantiza que, dentro de cada valor de las covariables X, no quedan sesgos de confusión.

2. *Positividad* (overlap o common support): para todo x con $P(X = x) > 0$,

   $$0 < P(T = 1 \mid X = x) < 1,$$

   lo cual asegura que en cada estrato de X haya al menos alguna unidad con $T = 1$ y alguna con $T = 0$.

3. *Ausencia de interferencia*: la respuesta de cada unidad i depende únicamente de su propio tratamiento T_i, no del de las demás.

4. *Consistencia*: el valor observado Y_i coincide con el potencial $Y_i(T_i)$ asociado a su tratamiento recibido.

Bajo estos cuatro supuestos se prueba que

$$\text{ATE} = \mathbb{E}_X\Big[\mathbb{E}[Y \mid T = 1, X] \; - \; \mathbb{E}[Y \mid T = 0, X]\Big] \tag{4}$$

llamada *fórmula de ajuste* o *adjustment formula*. En la práctica, este resultado ofrece la herramienta para estimar efectos causales promedio a partir de datos observacionales, siempre que sea razonable aceptar los supuestos anteriores.

2.4. Modelos causales

Para formalizar el paso de estimandos causales a expresiones basadas solo en la distribución observacional necesitamos un modelo que describa cómo cambian los mecanismos al intervenir. En este capítulo se presentará brevemente el operador `do` y el principio de *backdoor adjustment*, dos conceptos centrales en los modelos causales de Pearl [6].

2.4.1. Operador `do` y distribuciones intervencionales

Mientras que la notación clásica de probabilidad solo permite *condicionar* sobre un evento (por ejemplo, $P(Y \mid T = t)$), el operador `do` expresa *intervenciones* que fuerzan un valor de tratamiento en todo el sistema. Lo escribimos así:

$$P\big(Y = y \mid \text{do}(T = t)\big),$$

y representa la distribución de Y en el escenario hipotético donde todos han recibido $T = t$ de forma forzada. A diferencia de la mera condición $P(Y \mid T = t)$, intervenir con $\text{do}(T = t)$ implica deshacer cualquier influencia previa sobre T y luego observar Y [1].

Nuestro objetivo de identificación consiste en demostrar que esas mismas cantidades (las esperanzas $\mathbb{E}[Y \mid \text{do}(T = t)]$) pueden reescribirse usando solo probabilidades condicionadas y marginales que podemos estimar directamente de los datos observados. Es decir, queremos transformar una expresión que involucra intervenciones en otra que solo requiere contar frecuencias o ajustar modelos sobre (X, T, Y).

2.4.2. Criterio de "backdoor" y fórmula de ajuste

Para saber qué covariables debemos "ajustar" antes de calcular un efecto causal, Pearl introdujo el *criterio de backdoor* [6]. Supongamos que se trabaja con un grafo causal $\mathcal{G}$ cuyos nodos incluyen al tratamiento T, al resultado Y y a un conjunto de covariables Z. Se dice que Z satisface el criterio de backdoor para el efecto de T en Y si se cumplen las dos condiciones siguientes:

1. Ningún $Z \in X$ es descendiente de T.
2. Z bloquea todas las rutas con asociaciones espurias de T a Y, según la regla de *d-separation*.

Más formalmente, en la teoría de grafos causales se dice que Z *d-separa* a T de Y si no existe ninguna *cadena conectiva abierta* entre ellos una vez que se fija (u "observa") el

valor de todas las variables en Z. Bajo esta condición, cualquier asociación espuria que viajase por esas rutas queda completamente anulada.

Cuando Z cumple este criterio, la distribución bajo intervención se puede escribir como:

$$P\big(Y = y \mid \mathrm{do}(T = t)\big) \;=\; \sum_{z} P\big(Y = y \mid T = t,\, Z = z\big)\; P(Z = z). \tag{5}$$

A partir de aquí, la *fórmula de ajuste* para el efecto promedio (ATE) queda:

$$\mathrm{ATE} \;=\; \sum_{z} \Big[\mathbb{E}\big[Y \mid T = 1,\, Z = z\big] - \mathbb{E}\big[Y \mid T = 0,\, Z = z\big]\Big]\, P(Z = z) \tag{6}$$

En palabras sencillas, primero se divide la población según cada valor posible de Z. Dentro de cada estrato estimamos la diferencia de medias entre tratados y controles, y finalmente promediamos esas diferencias según la proporción de cada estrato en la población. Gracias al criterio de backdoor, al calcular primero la diferencia de medias entre tratados y controles dentro de cada estrato de Z, eliminamos cualquier sesgo de confusión. Después, al promediar esas diferencias estratificadas según la proporción de cada estrato en la población, obtenemos el ATE global. Todo el proceso se basa únicamente en probabilidades y medias extraídas de los datos observados. [6].

2.5. Average Treatment Effect on the Treated (ATT)

Hasta ahora nos hemos centrado en el *Average Treatment Effect* (ATE), que busca estimar el efecto causal medio que tendría aplicar el tratamiento a toda la población. Sin embargo, en muchos contextos reales, el interés se centra en evaluar qué impacto ha tenido el tratamiento sobre los individuos que efectivamente recibieron el tratamiento. Este es precisamente el objetivo del *Average Treatment Effect on the Treated* (ATT). Su fórmula general es la siguiente:

$$\mathrm{ATT} \;=\; \mathbb{E}\big[Y(1) - Y(0) \mid T = 1\big] \tag{7}$$

El ATT representa el efecto causal medio del tratamiento en el subgrupo de tratados. Esto se puede interpretar como la diferencia entre el resultado observado en los tratados y el resultado contrafactual que habrían tenido si no hubiesen recibido el tratamiento, en promedio. Aunque no siempre coincide con el ATE, su estimación se vuelve especialmente relevante cuando los tratados no son representativos del conjunto completo de la población. Además, el ATT nos permite enfocar el análisis en la población de tratados, que puede presentar características muy distintas de la población de control y, por tanto, requiere una evaluación específica. Es decir se puede dar el caso de que:

$$P\big(X \mid (T = 0)\big) \;\neq\; P\big(X \mid (T = 1)\big).$$

En este trabajo, considerar tanto el ATE como el ATT permite una evaluación más completa de los modelos, al reflejar distintos objetivos de análisis: uno más general y poblacional, y otro más específico y práctico.

2.6. Resolución de la Paradoja de Simpson

Con todo lo anterior en mente, podemos ahora resolver adecuadamente la Paradoja de Simpson presentada al inicio (Tabla 2.1). El sesgo surge porque al agregar los datos de ambos subgrupos ignoramos la influencia de la condición de los pacientes. Para obtener el verdadero efecto causal:

En la tabla original vemos que, a primera vista, la tasa de mortalidad (Y=1) es mayor para el Tratamiento B cuando agregamos todos los casos:

$$E[Y \mid T = \mathrm{A}] - E[Y \mid T = \mathrm{B}] = 16\,\% - 19\,\% = -3\,\%,$$

lo que incorrectamente sugiere que el Tratamiento B tiene una mayor tasa de mortalidad y es peor tratamiento que el A.

Sin embargo, al estratificar por la condición ($X \in \{\mathrm{Leve},\ \mathrm{Grave}\}$) obtenemos:

$$\begin{aligned} E[Y \mid T = \mathrm{A},\ X = \mathrm{Leve}] - E[Y \mid T = \mathrm{B},\ X = \mathrm{Leve}] &= 0{,}15 - 0{,}10 = 0{,}05, \\ E[Y \mid T = \mathrm{A},\ X = \mathrm{Grave}] - E[Y \mid T = \mathrm{B},\ X = \mathrm{Grave}] &= 0{,}30 - 0{,}20 = 0{,}10. \end{aligned}$$

Para combinar estos efectos necesitamos la distribución de X en la población completa:

$$P(X = \mathrm{Leve}) = \frac{1400 + 50}{1500 + 550} = \frac{1450}{2050} \approx 0{,}7073, \quad P(X = \mathrm{Grave}) \approx 0{,}2927.$$

Aplicando la fórmula de ajuste:

$$\begin{aligned} \mathrm{ATE} &= \sum_{x \in \{\mathrm{Leve},\mathrm{Grave}\}} \big(E[Y \mid T = \mathrm{A}, X = x] - E[Y \mid T = \mathrm{B}, X = x]\big)\, P(X = x) \\ &= 0{,}05 \times 0{,}7073 \ + \ 0{,}10 \times 0{,}2927 \approx 0{,}0354 + 0{,}0293 = 0{,}0647 = 6{,}47\,\%. \end{aligned}$$

El ATE ajustado de 6.47 indica que, una vez controlamos por la condición del caso, el Tratamiento A tiene en promedio una tasa de mortalidad de 6.5 puntos porcentuales adicionales frente al Tratamiento B. De este modo se corrige el sesgo de confusión y se revela que el Tratamiento B es en realidad el más eficaz.

3. Desarrollo y metodología

Una vez sentadas las bases teóricas para la identificación del efecto causal promedio, su fórmula de ajuste y sus supuestos esenciales, se da ahora el salto a la parte práctica: la implementación y comparación de distintos estimadores. En primera instancia nos centraremos en los métodos clásicos, que incluyen los denominados *meta-learners* (S-learner y T-learner) y los estimadores basados en *propensity scores* (IPW y doubly-robust method). Más adelante, y una vez comprendida su mecánica y limitaciones, se presentarán algunas arquitecturas del estado del arte que combinan modelos de resultado condicionado con regularización o redes neuronales especializadas.

3.1. Introducción al aprendizaje automático y reconocimiento de patrones

Antes de describir los métodos clásicos de estimación causal, conviene realizar una breve introducción del aprendizaje automático (machine learning) y su capacidad para reconocer patrones a partir de datos y la motivación que hay para utilizarlo en nuestro caso. A diferencia de los métodos estadísticos tradicionales, que se basan en una fórmula clara, el aprendizaje automático se apoya en algoritmos que extraen directamente la estructura subyacente sin presuponer ninguna forma concreta. Es decir, se aprende de los datos, adaptando la complejidad del modelo a la información disponible.

Esta flexibilidad se traduce en varias ventajas:

1. Los modelos realizan reconocimiento de patrones y son capaces de identificar relaciones no lineales y de alto orden entre covariables y resultados, incluso cuando dichas relaciones son difíciles de observar de antemano.

2. El aprendizaje se centra en extraer todo el conocimiento posible de los datos, maximizando la eficiencia predictiva sobre ejemplos nuevos y empleando validación cruzada o conjuntos de prueba para evitar el sobreajuste.

3. Gracias a su propiedad de aproximación universal de funciones [9] [10], las redes neuronales profundas pueden, en teoría, reproducir con gran precisión cualquier función continua, lo que las convierte en herramientas muy eficaces para modelar efectos de tratamiento complejos.

En el marco de la inferencia causal, las herramientas de aprendizaje automático, y en particular las redes neuronales, han adquirido un rol cada vez más importante como complemento a los métodos tradicionales. Su principal ventaja reside en su capacidad para manejar relaciones complejas entre variables sin necesidad de imponer una forma funcional específica. Esto resulta especialmente útil en contextos donde las relaciones entre

el tratamiento, las covariables y el resultado incluyen interacciones difíciles de modelar con técnicas clásicas.

El **aprendizaje automático** permite además aprovechar mejor conjuntos de datos grandes y de alta dimensión, que suelen ser problemáticos para otros modelos estadísticos convencionales. Al centrarse en la predicción, estos métodos pueden estimar funciones de forma más precisa, lo que a su vez mejora las estimaciones del efecto causal.

Las **redes neuronales**, por su parte, destacan por su capacidad de representar funciones complejas y adaptarse a una gran variedad de patrones en los datos. Aunque no están diseñadas originalmente con fines causales, su uso se ha extendido en inferencia causal precisamente por esta capacidad.

En definitiva, el uso de técnicas de machine learning permite adaptarse mejor a la complejidad de los datos reales, sin dejar de respetar los principios que hacen posible una interpretación causal rigurosa.

3.2. Métodos clásicos

Tras haber introducido el valor que aporta el machine learning para modelar la complejidad de los datos observacionales respetando principios causales, se pasa ahora a la parte práctica. En esta etapa implementaremos y compararemos distintos estimadores. Empezaremos con los métodos clásicos, que incluyen los llamados meta-learners (S-learner y T-learner) [11] y los estimadores basados en *propensity scores* (IPW [12] y método doubly-robust [13]).

3.2.1. Meta-learners

Antes de describir cada uno de estos estimadores, es necesario comentar e indicar las dos familias de estimadores a las que pertenece cada uno de los estimadores que vamos a usar:

- **Conditional Outcome Modeling (COM):** utilizan un único modelo predictivo que aprende la relación conjunta $\hat{f}(x,t) \approx \mathbb{E}[Y \mid X = x, T = t]$. El S-learner es el ejemplo claro de este grupo.
- **Grouped Conditional Outcome Modeling (GCOM):** encajan modelos independientes para cada grupo de tratamiento. El T-learner ajusta dos regresiones separadas $\hat{\mu}_0(x)$ y $\hat{\mu}_1(x)$, de modo que el efecto individual se construye con la diferencia.

Ambas categorías son formas de *meta-learning*, ya que reutilizan esquemas de regresión o clasificación estándar para inferir efectos causales. A continuación profundizaremos en cada uno de estos enfoques [1] [11].

S-learner

El *S-learner* constituye la forma más directa de aplicar un modelo de resultado condicionado (COM) para estimar efectos causales. Su idea clave es tratar la variable de tratamiento T como una característica más del vector de covariables X, ajustando en nuestro caso un único predictor.

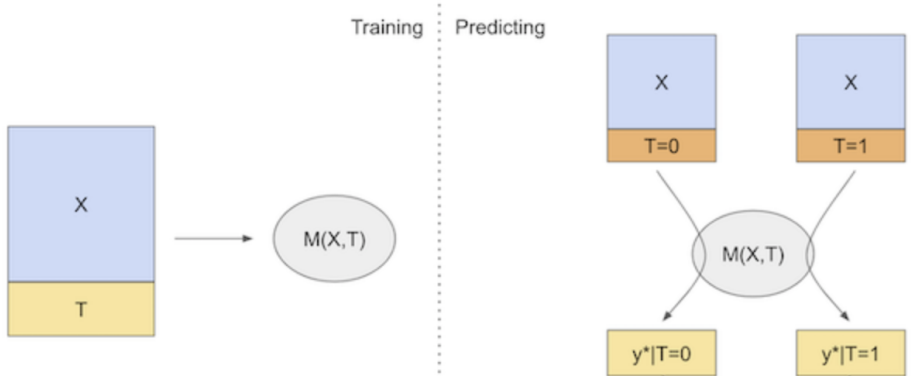

Figura 3.1: Arquitectura del S-learner, donde X y T son las entradas de la red y M es el modelo [13].

Para ello:

$$\hat{\mu}(t, x) \approx \mathbb{E}[Y \mid T = t,\, X = x]$$

mediante una red neuronal de perceptrón multicapa (MLP).

1. Se crea un conjunto de datos $\{(x_i, t_i, y_i)\}_{i=1}^{n}$, donde cada muestra incorpora tanto las covariables x_i como el indicador t_i.

2. Se entrena un modelo $\hat{\mu}$ implementado como un MLP, minimizando el error cuadrático medio en la predicción de Y a partir de (T, X) [1].

3. Una vez obtenido $\hat{\mu}$, el efecto individual estimado para la unidad i se obtiene restando las predicciones:

$$\hat{\tau}_i \;=\; \hat{\mu}(1, x_i) \;-\; \hat{\mu}(0, x_i). \tag{8}$$

Arquitectura empleada El predictor $\hat{\mu}$ se implementa como un perceptrón multicapa con varias capas ocultas, activación ReLU y capas de *dropout*. La red finaliza en una capa lineal que produce la predicción de Y.

Con este procedimiento se obtiene un estimador del ATE o de efectos individuales sin necesidad de separar el ajuste por niveles de tratamiento, aunque su precisión depende de la capacidad del modelo de capturar la contribución de T frente a la de X.

T-learner

En el *T-learner*, perteneciente al enfoque de *Grouped Conditional Outcome Modeling* (GCOM), se opta por entrenar dos modelos separados, uno para el grupo tratado y otro para el de control, en lugar de un único modelo.

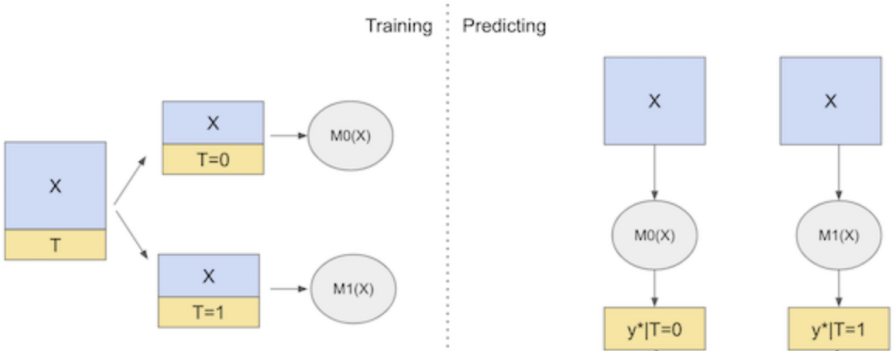

Figura 3.2: Arquitectura del T-learner, donde se la red se divide en dos submodelos $M0$ y $M1$, uno para $T = 0$ y otro para $T = 1$ [13].

Con ello, se evita que el tratamiento pase inadvertido dentro de un vector de características de alta dimensión:

$$\mu_1(x) \approx \mathbb{E}[Y \mid T = 1,\, X = x], \qquad \mu_0(w) \approx \mathbb{E}[Y \mid T = 0,\, X = x].$$

El efecto promedio se estima entonces como

$$\hat{\tau} \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big[\,\hat{\mu}_1(x_i) \;-\; \hat{\mu}_0(x_i)\Big] \tag{9}$$

Para ello:

1. Se separa el conjunto de datos en dos bloques: uno con las unidades tratadas ($T = 1$) y otro con las de control ($T = 0$).
2. En cada bloque se ajusta un MLP independiente, uno para $\hat{\mu}_1$ y otro para $\hat{\mu}_0$ a partir de las covariables X, minimizando el error cuadrático medio.
3. Finalmente, para cada unidad i se calcula la diferencia entre $\hat{\mu}_1(x_i)$ y $\hat{\mu}_0(x_i)$; su promedio sobre todas las unidades da $\hat{\tau}$.

Arquitectura empleada Cada uno de los dos modelos es un perceptrón multicapa con capas ocultas ReLU y *dropout*. Las entradas son únicamente las covariables X, la salida es la predicción de Y. Al disponer de dos redes especializadas, el T-learner evita que la señal de T se diluya en un único predictor y garantiza que la estimación capture por separado la respuesta bajo tratamiento y bajo control.

3.2.2. Propensity Score

Hasta ahora se ha visto que para eliminar el sesgo por confusión, se debe ajustar por todo el conjunto de covariables X que satisface el criterio de backdoor. Sin embargo, cuando X es de gran dimensión puede resultar poco práctico o dar lugar a problemas de solapamiento. Afortunadamente, Rosenbaum y Rubin demostraron que basta con ajustar por el *propensity score*, una única variable escalar que resume toda la información de X relevante para la asignación al tratamiento [14].

Se define

$$e(x) \;=\; P\big(T = 1 \mid X = x\big) \tag{10}$$

la probabilidad de recibir el tratamiento condicionado en $X = x$. El resultado central que se obtiene es que bajo los supuestos de positividad e ignorabilidad condicional respecto a X, se tiene:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X).$$

En otras palabras, si un conjunto de covariables X garantiza la condición de unconfoundedness, entonces condicionar también por el propensity score $e(X)$ mantiene dicha condición. Gráficamente, como se observa en la figura 3.3, esto equivale a que $e(X)$ media completamente el efecto de X en la asignación T, bloqueando los mismos caminos de confusión que bloqueaba X.

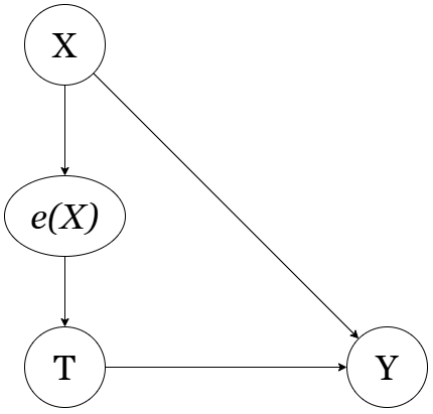

Figura 3.3: Uso del *propensity score* $e(X)$ como variable para bloquear todas las vías de confusión entre tratamiento y resultado.

El principal beneficio del *propensity score* es que condicionar sobre esta única variable escalar equivale a ajustar por todo el conjunto original X, sin necesidad de estratificar en múltiples niveles simultáneos. Esto simplifica enormemente la implementación y mejora el solapamiento efectivo entre los grupos de tratamiento y control, al evitar estratos demasiado pequeños que dificulten el cumplimiento de la positividad.

En la práctica rara vez se conoce $e(X)$ de antemano, así que hay que estimarla por medio de un modelo predictivo ajustándolo a los datos observados (X_i, T_i). El resultado estimado $\hat{e}(X)$ se utiliza luego en los métodos de ponderación (IPW) o en estimadores *doubly robust*, que se verán a continuación.

Inverse Probability Weighting (IPW)

La idea fundamental del *Inverse Probability Weighting* (IPW) es construir una "pseudo-población" en la que la asignación del tratamiento sea independiente de los confusores, de modo que la asociación pase a comportarse como causalidad. En la figura original (Figura 2.1), la variable X confunde la relación entre T y Y porque ambos dependen de X, es decir,

$$P(T \mid X) \neq P(T)\,.$$

Si somos capaces de reponderar cada unidad por el inverso de la probabilidad de recibir el tratamiento que efectivamente obtuvo, anularíamos ese sesgo: en la población reponderada se cumplirá

$$P(T \mid X) = P(T) = \text{constante},$$

y el grafo resultante carecerá de unión $X \to T$ (Figura 2.2).

Precisamente, dichas ponderaciones se obtienen a partir del *propensity score* $e(X) = P(T = 1 \mid X)$. Para cada unidad con $T = 1$ su peso será $1/e(W)$, y para las que tienen $T = 0$ será $1/(1-e(X))$ [15]. Intuitivamente, al multiplicar cada observación por el inverso de $P(T \mid X)$ se "cancela" el mecanismo de asignación $P(T \mid X)$, consiguiendo una muestra sintética donde T y X resultan independientes.

Bajo esta nueva distribución se puede demostrar la siguiente fórmula:

$$\mathbb{E}\big[Y(t)\big] = \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}\, Y}{P(T = t \mid X)}\right],$$

de la cual, para un tratamiento binario, se deduce el estimando del ATE:

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\tfrac{\mathbf{1}\{T=1\}\, Y}{e(X)}\right] \; - \; \mathbb{E}\left[\tfrac{\mathbf{1}\{T=0\}\, Y}{1-e(X)}\right].$$

Para obtener el estimador empírico IPW sustituimos las expectativas por medias muestrales y $e(X)$ por su estimación $\hat{e}(X)$. Así, una forma equivalente del estimador es:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\mathbf{1}\{T_i = 1\}\, Y_i}{\hat{e}(X_i)} \; - \; \frac{\mathbf{1}\{T_i = 0\}\, Y_i}{1 - \hat{e}(X_i)}\Big] \tag{11}$$

Un problema práctico es que si $\hat{e}(X_i)$ está muy cerca de 0 o de 1, los pesos se disparan y aumenta drásticamente la varianza. Por ello, en ocasiones se acotan los valores de $\hat{e}(X)$ que quedan por fuera de $[\varepsilon,\, 1 - \varepsilon]$ para limitar los pesos máximos a $1/\varepsilon$ y mínimos a ε a costa de introducir un pequeño sesgo adicional.

De igual modo, para estimar el efecto promedio condicional

$$\tau(x) \;=\; \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big],$$

basta con aplicar la misma fórmula IPW, pero restringiendo la suma a las unidades que comparten $X_i = x$. No obstante, cuando cada estrato definido por X contiene pocos individuos, los pesos IPW tienden a ser muy dispares y la varianza del estimador aumenta de forma significativa [16].

Arquitectura empleada Para estimar el *propensity score* $\hat{e}(X)$, utilizamos un modelo de regresión logística implementado como una red neuronal sencilla. La arquitectura consta de:

- Una capa lineal que toma como entrada el vector de covariables X y devuelve un escalar.
- Una función sigmoide que transforma la salida lineal en una probabilidad $\hat{e}(X) = P(T = 1 \mid X)$.

Esta elección, que equivale a una regresión logística simple, limita la flexibilidad del modelo para capturar relaciones no lineales en los datos.

Una vez ajustado el modelo, las predicciones $\hat{e}(X_i)$ se incorporan directamente en la fórmula IPW descrita anteriormente, cerrando así el ciclo desde la estimación del *propensity score* hasta el cálculo del estimador de tratamiento.

3.2.3. Métodos doblemente robustos

Los *métodos doblemente robustos* combinan las ideas de modelado del resultado condicionado $\mu(t,x) = \mathbb{E}[Y \mid T = t, X = x]$ y de modelado de la probabilidad de tratamiento $e(x) = P(T = 1 \mid X = x)$. Su principal virtud es que el estimador resultante es consistente para el ATE si el modelo de resultados $\hat{\mu}$ está correctamente estimado *o también* si la estimación del *propensity score* $\hat{e}$ lo está [17][18].

Además, la velocidad de convergencia de este estimador es el producto de las velocidades de convergencia de $\hat{\mu}$ y de $\hat{e}$, lo cual es especialmente útil cuando se emplean modelos de aprendizaje automático flexibles en altas dimensiones, donde cada componente por separado puede converger más despacio que la tasa ideal $n^{-1/2}$ (siendo n el número de muestras) [19][20].

En la práctica, un estimador doblemente robusto típico (también llamado *augmented IPW* o AIPW) combina la parte de ponderación inversa propia del método IPW con un término de corrección basado en la predicción de $\hat{\mu}$. Una formulación frecuente de este método es la siguiente:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbf{1}\{T_i = 1\}Y_i}{\hat{e}(X_i)} - \frac{\mathbf{1}\{T_i = 0\}Y_i}{1-\hat{e}(X_i)} + \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\right\} \tag{12}$$

donde el término final $\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)$ sirve para reducir la varianza y corregir posibles sesgos del componente IPW [17][21].

Sin embargo, su desempeño en la práctica depende de que al menos uno de los dos modelos esté suficientemente bien especificado; de lo contrario, el estimador puede presentar sesgos inesperados o varianzas elevadas [22][23].

Arquitectura empleada Para implementar el estimador *doubly robust* se combinan tres componentes:

- Las predicciones $\hat{\mu}_1(x_i)$ y $\hat{\mu}_0(x_i)$ obtenidas del *T-learner*.
- El vector de *propensity scores* $\hat{e}(x_i)$ estimado por la regresión logística usada para luego calcular el IPW.
- Las observaciones reales T_i y Y_i.

La fórmula A-IPTW para cada paciente i es:

$$\hat{\tau}_i = T_i\,\frac{Y_i - \hat{\mu}_1(x_i)}{\hat{e}(x_i)} - (1 - T_i)\,\frac{Y_i - \hat{\mu}_0(x_i)}{1-\hat{e}(x_i)} + \left[\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right].$$

A continuación promediamos para obtener el ATE:

$$\hat{\tau}_{\text{Doubly}-\text{robust}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i.$$

3.3. Arquitecturas del estado del arte

Una vez familiarizados con los estimadores "clásicos" (meta-learners y métodos basados en *propensity score*), nos centraremos ahora en arquitecturas más modernas. Estos modelos

no se limitan a elegir entre ajustar la respuesta condicionada o el propensity score, sino que los integran simultáneamente y añaden mecanismos de regularización específicos para ganar estabilidad y precisión en problemas de alta dimensión.

Antes de profundizar en Dragonnet y Targeted Regularization, conviene conocer otras propuestas recientes que han influido en esta materia y aportan distintas estrategias de representación y balance.

3.3.1. Otras arquitecturas contemporáneas

A continuación, presentamos de forma concisa algunas de las alternativas más destacadas:

- **TARNet** (Treatment-Agnostic Representation Network) [24]: aprende primero una representación común de las covariables X y luego se divide en dos cabezas (una para el grupo tratado y otra para el de control) evitando que las diferencias de grupo distorsionen la base compartida.
- **CFRNet** (Counterfactual Regression with Integral Probability Metrics) [24]: toma el esquema de TARNet y añade un término que mide la distancia entre las distribuciones latentes de tratados y controles, forzando un mejor equilibrio en el espacio de representación.
- **X-learner** [11]: tras ajustar dos modelos de resultado separados, genera estimaciones contrafactuales cruzadas y las combina según la proporción de cada grupo, mejorando la precisión especialmente cuando los tamaños de muestra están desbalanceados.
- **R-learner** [25]: descompone la tarea en dos etapas (estimación del *propensity score* y regresión de residuos) y formula la inferencia del efecto como un problema de optimización de residuos, permitiendo usar cualquier algoritmo de machine learning en cada fase.

Con esta visión global sobre distintas líneas de trabajo, estamos listos para examinar en detalle dos de las propuestas más influyentes en los últimos años: Dragonnet y Targeted Regularization.

3.3.2. Dragonnet

Dragonnet es una arquitectura de red neuronal diseñada específicamente para la estimación de efectos causales en datos observacionales, propuesta por Shi [26]. Su diseño parte de dos observaciones clave:

(i) **Suficiencia del *propensity score* para el control de confusión.** Rosenbaum y Rubin demostraron que, bajo positividad e ignorabilidad condicional, basta con ajustar por la probabilidad de tratamiento condicionada en las covariables,

$$e(X) = P(T = 1 \mid X),$$

en lugar de por todo el vector X [14]. Este *propensity score* concentra toda la información de X relevante para el sesgo de asignación, de modo que

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X).$$

Dragonnet dedica una de sus tres "cabezas" a estimar $e(X)$ y fuerza la representación interna a retener únicamente aquellas características de X que realmente influyen en la asignación de tratamiento, bloqueando todos los caminos de confusión de manera efectiva.

(ii) **Modelado conjunto de resultado y tratamiento para optimizar la estimación final.** En lugar de aprender primero un *propensity score* y luego, por separado, un modelo de resultados, Dragonnet entrena simultáneamente dos subredes para predecir $\mathbb{E}[Y \mid T = 0, X]$ y $\mathbb{E}[Y \mid T = 1, X]$, además de la subred de propensity score. Al compartir una representación intermedia $\mathbf{Z}(X)$, los gradientes de cada tarea (tratamiento y resultado) interactúan mutuamente. La tarea de tratamiento refuerza en $\mathbf{Z}(X)$ las señales útiles para discriminar quién recibe T, mientras que la tarea de resultado garantiza que esa misma representación capte también cómo cambia Y según T. Este acoplamiento mejora la estabilidad y la precisión final del estimador [26].

Arquitectura A continuación se muestra en esta imagen la arquitectura de Dragonnet, la misma que usa Claudia Shi en su paper *Adapting Neural Networks for the Estimation of Treatment Effects*:

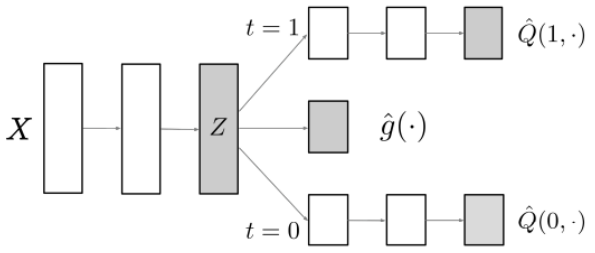

Figura 3.4: Representación esquemática de la red neuronal Dragonnet, que integra modelado de resultado y regularización por *propensity score* [26].

Dragonnet consta de tres "cabezas" que comparten una misma representación intermedia $\mathbf{Z}(X)$:

- **Cabeza de resultados con $T = 0$:** una subred de dos capas ocultas que predice $\hat{Q}(0, x) \approx \mathbb{E}[Y \mid T = 0, X = x]$.
- **Cabeza de resultados con $T = 1$:** idéntica a la anterior, pero entrenada para $\hat{Q}(1, x) \approx \mathbb{E}[Y \mid T = 1, X = x]$.
- **Cabeza de *propensity score*:** un único mapeo lineal seguido de una sigmoide que estima $\hat{g}(x) \approx P(T = 1 \mid X = x)$.

De este modo, la red aprende simultáneamente a representar $\mathbf{Z}(X)$, que son las características relevantes tanto para la asignación al tratamiento como para la predicción del resultado, evitando que la subred de resultados "olvide" la señal del tratamiento ni que la de *propensity score* sufra de alta varianza.

Función de pérdida El entrenamiento se realiza minimizando la suma de dos componentes diferenciables sobre el dataset $\{(x_i, t_i, y_i)\}_{i=1}^n$:

$$\mathcal{L}(\theta, X) = \frac{1}{n}\sum_{i=1}^{n}\Big[\big(\hat{Q}(t_i, x_i; \theta) - y_i\big)^2 \; + \; \alpha\, \text{CrossEntropy}\big(\hat{g}(x_i; \theta),\, t_i\big)\Big] \tag{13}$$

donde $\alpha > 0$ es un hiperparámetro que equilibra la importancia del *propensity score* en la función de pérdida [26], θ son los parámetros de la red y X son las covariables de entrada.

Estimación del efecto causal Una vez ajustados los parámetros θ, *Dragonnet* produce las predicciones $\hat{Q}(1, x)$ y $\hat{Q}(0, x)$. El ATE se calcula mediante

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\Big[\hat{Q}(1, x_i) - \hat{Q}(0, x_i)\Big] \tag{14}$$

3.3.3. Regularización dirigida (Targeted Regularization)

Targeted Regularization es una técnica que modifica la función de pérdida en el entrenamiento de redes neuronales con el fin de asegurar que el estimador de efecto causal cumpla ciertas ecuaciones de ortogonalidad (influencia eficiente) propias de la teoría semiparamétrica. El objetivo es combinar la flexibilidad de los modelos de resultado condicionado con la corrección basada en el *propensity score*, obteniendo así estimadores con buenas propiedades asintóticas (doble robustez y eficiencia) sin sacrificar la estabilidad en muestras finitas [26].

Preliminares de la teoría semiparamétrica Sea $\tau = \mathbb{E}[Y(1) - Y(0)]$ el estimando causal de interés. En el marco no paramétrico, un estimador $\hat{\tau}$ disfruta de propiedades óptimas (convergencia rápida y mínima varianza asintótica) si existe un triplete $(\hat{Q}, \hat{g}, \hat{\tau})$ que satisface la ecuación:

$$0 \;=\; \frac{1}{n}\sum_{i=1}^{n}\varphi\big(y_i,\, t_i,\, x_i;\, \hat{Q},\, \hat{g},\, \hat{\tau}\big) \tag{15}$$

donde

$$\varphi(y, t, x; Q, g, \tau) = \big[Q(1, x) - Q(0, x) - \tau\big] \; + \; \frac{t - g(x)}{g(x)\,(1 - g(x))}\,\big[y - Q(t, x)\big] \tag{16}$$

es la *efficient influence curve* para τ (ver Chernozhukov [27] y Van der Laan & Rose [28]).

Idea principal Supongamos que Q (modelo de resultados) y g (*propensity score*) son aproximados por una red neuronal con parámetros θ, cuyas salidas son

$$Q^{nn}(t, x; \theta) \approx \mathbb{E}[Y \mid T = t, X = x], \quad g^{nn}(x; \theta) \approx P(T = 1 \mid X = x).$$

La regularización dirigida introduce un nuevo parámetro escalar ε y define una versión "perturbada" de Q:

$$\widetilde{Q}(t_i, x_i, \theta, \varepsilon) = Q^{nn}(t_i, x_i;\, \theta) \; + \; \varepsilon\left[\frac{t_i}{g^{nn}(x_i;\, \theta)} \; - \; \frac{1 - t_i}{1 - g^{nn}(x_i;\, \theta)}\right] \tag{17}$$

de modo que el término añadido coincide con la parte de la influence curve que contiene $y - Q$. A continuación, se añade al criterio de entrenamiento habitual un término que penaliza el error cuadrático de esta perturbación:

$$(\hat{\theta}, \hat{\varepsilon}) = \arg\min_{\theta,\varepsilon} \left\{ \underbrace{\mathcal{L}(\theta, X)}_{\substack{\text{pérdida original} \\ \text{(Dragonnet)}}} + \beta \underbrace{\frac{1}{n}\sum_{i=1}^{n} \left[y_i - \widetilde{Q}(t_i, x_i;\, \theta, \varepsilon)\right]^2}_{\substack{\text{término de} \\ \text{regularización dirigida}}} \right\} \tag{18}$$

donde $\beta > 0$ es un hiperparámetro que pondera el peso de la regularización dirigida. Y donde la función de pérdida original de Dragonnet es la siguiente ecuación:

$$\mathcal{L}(\theta, X) = \frac{1}{n}\sum_{i=1}^{n} \Big[\big(\hat{Q}(t_i, x_i; \theta) - y_i\big)^2 \; + \; \alpha \, \text{CrossEntropy}\big(\hat{g}(x_i; \theta),\, t_i\big) \Big],$$

donde recordamos que $\alpha > 0$ es un hiperparámetro que pondera, respectivamente, la predicción del *propensity score*.

Implicación sobre la ecuación de influencia Cuando derivamos la nueva función de pérdida conjunta respecto a ε y fijamos esa derivada a cero en el óptimo $\hat{\varepsilon}$, obtenemos precisamente

$$0 \;=\; \frac{1}{n}\sum_{i=1}^{n} \varphi\big(y_i,\, t_i,\, x_i;\, \widetilde{Q}_{\hat{\theta},\hat{\varepsilon}},\, g_{\hat{\theta}},\, \hat{\tau}^{\text{treg}}\big),$$

donde φ es la *efficient influence curve* definida antes. Esta igualdad significa que el ajuste conjunto de θ y ε hace que el estimador cumpla la condición de ortogonalidad requerida por la teoría semiparamétrica: la corrección introducida con ε anula cualquier sesgo de primer orden, garantizando así la doble robustez y la eficiencia asintótica del estimador [28].

Estimador final del ATE Tras entrenar y obtener $(\hat{\theta}, \hat{\varepsilon})$, definimos el modelo de resultados "perturbado"

$$\hat{Q}^{\text{treg}}(t, x) = \widetilde{Q}\big(t, x;\, \hat{\theta}, \hat{\varepsilon}\big).$$

El ATE se estima entonces de forma natural como la media de las diferencias de predicción:

$$\hat{\tau}^{\text{treg}} = \frac{1}{n}\sum_{i=1}^{n} \Big[\hat{Q}^{\text{treg}}(1,\, x_i) - \hat{Q}^{\text{treg}}(0,\, x_i)\Big] \tag{19}$$

En definitiva, Dragonnet y Targeted Regularization representan dos caminos complementarios para llevar las ideas clásicas de ajuste por *propensity score* y modelado de resultados a entornos de alta dimensión y gran flexibilidad. Dragonnet aprovecha una representación compartida que permite aprender simultáneamente la asignación del tratamiento y el comportamiento de la variable de salida, mientras que Targeted Regularization introduce explícitamente en la función de pérdida los términos necesarios para garantizar las propiedades de doble robustez y eficiencia asintótica.

Los modelos descritos en esta sección, junto con los métodos clásicos vistos anteriormente, serán evaluados empíricamente sobre el conjunto de datos IHDP. De este modo podremos comparar su desempeño bajo condiciones homogéneas y extraer conclusiones sólidas sobre su precisión y robustez.

4. Resultados

Antes de entrar de lleno con los resultados obtenidos, presentamos primero un breve resumen de la base de datos empleada y, a continuación, las métricas utilizadas para comparar los distintos estimadores implementados: S-learner, T-learner, IPW, doubly-robust, Dragonnet y Targeted Regularization.

El código completo de la implementación de los métodos descritos en este trabajo está disponible en el repositorio GitHub: https://github.com/franciscogomez03/tfg-causal-inference.

4.1. Conjunto de datos y métricas

El experimento se realiza sobre el conjunto de datos `IHDP`, en el cual los datos generados son semi-sintéticos. Gracias a este hecho, conocemos además del resultado observado Y_i, también los potenciales contrafactuales $Y_i(0)$ y $Y_i(1)$. Esto nos permite calcular el efecto individual verdadero $\tau_i = Y_i(1) - Y_i(0)$ y el ATE real, que utilizamos como referencia para evaluar la precisión de nuestros estimadores.

Cada una de las cien particiones contiene 747 observaciones independientes, con las siguientes variables:

Columna	Descripción
`T`	Indicador de tratamiento (0 = control, 1 = tratado)
`y_f`	Resultado factual observado Y_i
`y_cf`	Resultado contrafactual $Y_i(1 - T_i)$
`mu0`	Media generadora de $Y_i(0)$ en simulación
`mu1`	Media generadora de $Y_i(1)$ en simulación
`X1`–`X6`	Covariables continuas
`X7`–`X25`	Covariables discretas (binarias)

Tabla 4.1: Descripción de las columnas del dataset IHDP semi-sintético utilizado.

- Un vector de covariables $X_i \in \mathbb{R}^p$.
- Un indicador de tratamiento $T_i \in \{0, 1\}$.
- Una respuesta observada $Y_i \in \mathbb{R}$.
- Además, en estas simulaciones se conoce el efecto individual verdadero $\tau_i = Y_i(1) - Y_i(0)$.

Para evaluar la robustez de los métodos, entrenamos y evaluamos cada estimador en las cien particiones por separado y luego reportamos la media y desviación estándar de las métricas (ATE, PEHE, ATT) sobre estos cien experimentos. La sección del anexo recoge estadísticas descriptivas y el proceso de preprocesado (limpieza, imputación y escalado).

Para más detalles sobre la base de datos, véase el Anexo C.

4.1.1. Métricas de evaluación

Recordamos que para cada estimador obtenemos predicciones de efectos causales que comparamos mediante tres métricas:

- **Error en la estimación del ATE.** Se evalúa el sesgo medio y el error cuadrático medio entre el ATE real y su estimador $\hat{\tau}$. Aquí reportamos la media y desviación estándar del MAE a lo largo de las 100 particiones.
- **Precision in Estimation of Heterogeneous Effect (PEHE).** Esta medida cuantifica la precisión en la estimación de los efectos individuales:

$$\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}_i - \tau_i\right)^2} \tag{20}$$

  Obtenemos su valor promedio y su dispersión sobre las cien bases de datos.
- **Average Treatment effect on the Treated (ATT).** El ATT mide el efecto promedio entre las unidades tratadas:

$$\text{ATT} = \mathbb{E}\big[Y(1) - Y(0) \mid T = 1\big] \;\approx\; \frac{1}{n_1}\sum_{i:T_i=1}\tau_i \tag{21}$$

  Su estimación $\widehat{\text{ATT}}$ se compara al final con el ATT verdadero. Cabe destacar que la fórmula para este estimador varía en función del método ultilizado, por lo que más adelante se tratará con mayor profundidad.

4.1.2. Preprocesado y partición de datos

Antes de ajustar los modelos, se han realizado dos pasos clave de preprocesado:

1. **Ficheros de entrenamiento/validación y test.** Contamos con dos ficheros en formato `.npz` que agrupan los datos de los 100 experimentos:
   - `ihdp_npci_1-100.train.npz`: contiene todas las observaciones destinadas a entrenamiento y validación.
   - `ihdp_npci_1-100.test.npz`: contiene exclusivamente los ejemplos de test.

   Para cada uno de los 100 sub-conjuntos de entrenamiento/validación, separamos un 27 % de las muestras para validación y usamos el 73 % restante como *training*. Esta división se hace de forma aleatoria.
2. **Estandarización de variables.** Aplicamos una transformación *z-score* (media cero y desviación típica uno) únicamente a las seis covariables continuas `X1` a `X6`, con

`StandardScaler` de `scikit-learn`. El resto de columnas con variables discretas se mantienen sin modificar. De este modo, se garantizan que las magnitudes de estas variables no dominen el proceso de entrenamiento y que las redes converjan de forma más estable (véase el script `standardization.py` para más detalles).

En el siguiente apartado se muestra, para cada estimador, sus valores medios y desviaciones estándar de ATE, PEHE y ATT sobre las diez particiones, acompañados de tablas comparativas y gráficos de barras que faciliten la visualización de su rendimiento relativo.

4.2. Comparativa y aprendizaje

A continuación ofrecemos un análisis detallado de los resultados obtenidos con todos los métodos probados (tanto los clásicos como los más avanzados). Para facilitar la comprensión, primero se muestran unas gráficas comparativas y una tabla resumen de errores medios (desviación estándar) de cada modelo:

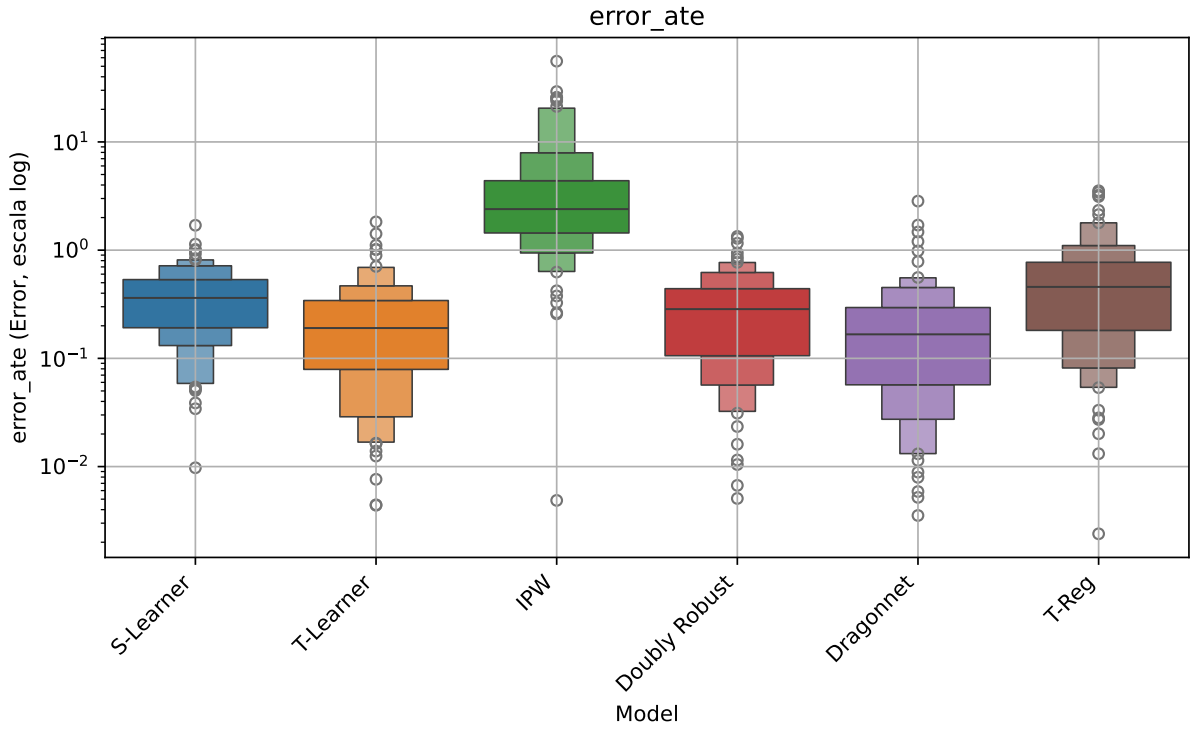

Figura 4.1: Evolución del error medio del estimador ATE en las 100 réplicas sobre IHDP en función del modelo.

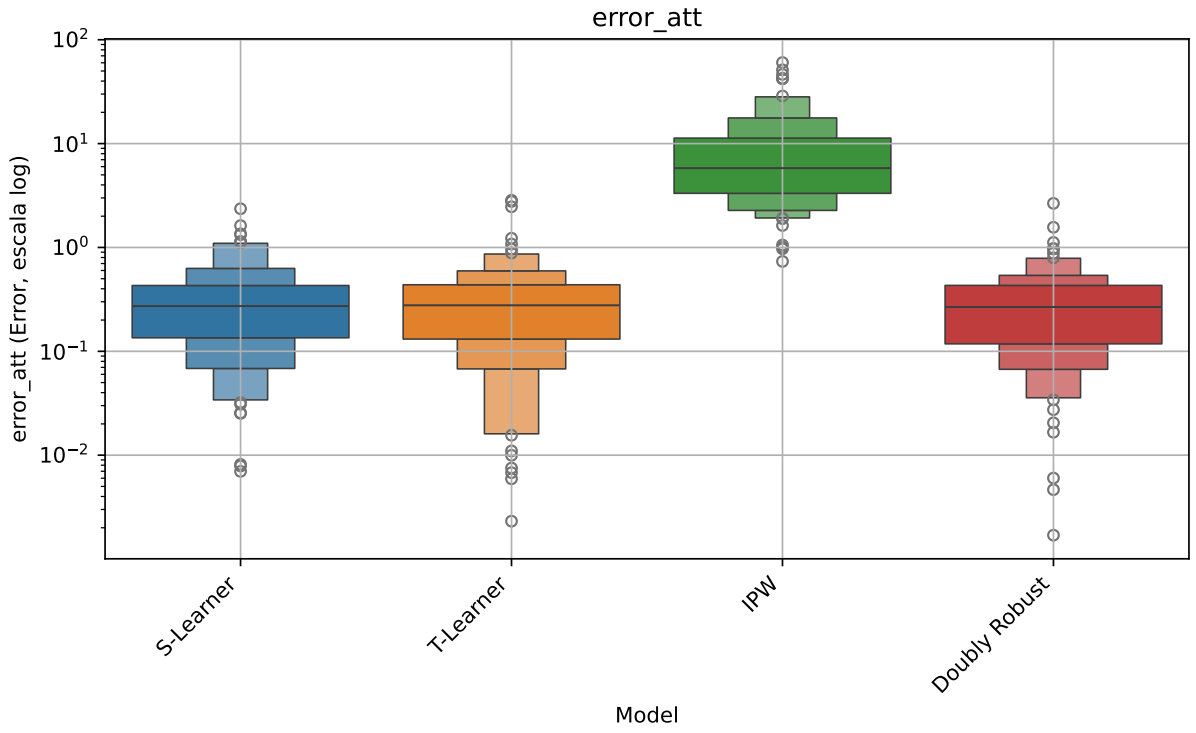

Figura 4.2: Evolución del error medio del estimador ATT en las 100 réplicas sobre IHDP en función del modelo.

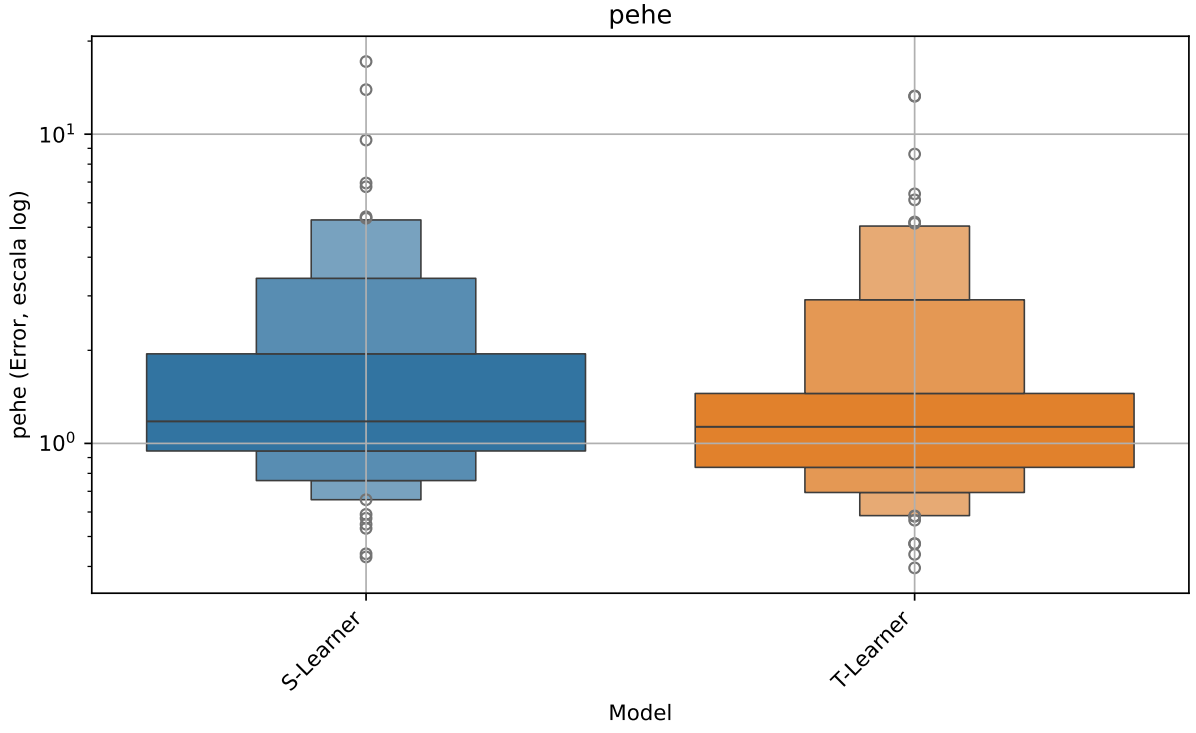

Figura 4.3: Evolución del error medio del PEHE en las 100 réplicas sobre IHDP, en función del modelo.

Modelo	error_ate	error_att	PEHE
S-Learner	0.405 (0.280)	0.370 (0.375)	2.035 (2.483)
T-Learner	0.269 (0.298)	0.380 (0.475)	1.768 (2.145)
IPW	5.009 (7.902)	9.759 (11.285)	–
Doubly Robust	0.338 (0.278)	0.332 (0.353)	–
Dragonnet	0.258 (0.383)	–	–
Targeted regularization	0.646 (0.723)	–	–

Tabla 4.2: Comparativa de rendimiento entre varios métodos para estimación causal. Se reportan los errores medios del ATE, ATT y PEHE, acompañados de su desviación estándar entre paréntesis. Algunos modelos no reportan ciertos estimandos por diseño. Dragonnet fue sometido a un *paired t-test* (α=0.05) frente al resto de modelos, mostrando diferencias estadísticamente significativas exceptuando el T-learner, por lo que se han marcado con un sombreado sus resultados.

4.2.1. Meta-learners

A primera vista, con respecto a los meta-learners, el S-learner y el T-learner constituyen variaciones menores de un mismo esquema: modelar $\mathbb{E}[Y \mid T, X]$. Aun así, en la práctica su comportamiento difiere de manera sustancial.

El **S-learner** entrena un único modelo $\hat{\mu}(t, x)$ que toma como entrada tanto las covariables x como la variable de tratamiento t. Cuando el efecto del tratamiento es pequeño en comparación con la variabilidad de los datos o la complejidad de X, la red suele centrarse más en aprender patrones de X y deja pasar parte de la información de T. En nuestro experimento esto se traduce en un `error_ate` medio de 0.405 y un PEHE de 2.035, cifras que evidencian que, con frecuencia, la red "ignora" la dimensión T y sobreajusta las covariables, perdiendo sensibilidad al efecto del tratamiento.

En cambio, el **T-learner** divide explícitamente la tarea en dos subproblemas: primero ajusta

$$\hat{\mu}_1(x) \approx \mathbb{E}[Y \mid T = 1, X = x],$$

sobre el subconjunto de datos con $T = 1$, y luego

$$\hat{\mu}_0(x) \approx \mathbb{E}[Y \mid T = 0, X = x],$$

sobre el conjunto con $T = 0$. De esta manera, cada red se concentra únicamente en su propio grupo y aprovecha toda su capacidad para capturar la relación entre X e Y sin interferencias. Este aislamiento de tareas reduce drásticamente el `error_ate` a 0.269, prácticamente un 50 % de mejora respecto al S-learner. Además, el T-learner reduce el PEHE a 1.768, mejorando también esta métrica frente al 2.035 obtenido por el S-learner, una vez más debido a su capacidad de mejora con respecto al S-leraner. Dicho de otro modo, la especialización permite extraer de forma más fiable la contribución causal del tratamiento, porque no hay interferencia interna de la estimación de los resultados bajo la otra condición de T.

Sin embargo, esta mejora en la estimación global del ATE no se traslada por completo al cálculo del ATT (*Average Treatment effect on the Treated*). El ATT sólo requiere $\hat{\mu}_1$, es

decir, la rama de la red que modela el grupo tratado. Al haber entrenado $\hat{\mu}_1$ únicamente con las observaciones de $T = 1$, ese modelo dispone de un conjunto de datos más reducido y su varianza puede ser mayor, especialmente, como en nuestro caso, cuando la muestra de tratados no es muy grande. De hecho, se observa un `error_att` de 0.380 para el **T-learner** frente al 0.370 del **S-learner**: aunque el **S-learner** tenía peor `error_ate`, su única red también se entrenaba con todos los datos y, en consecuencia, su subpredicción $\hat{\mu}(1, x)$ puede resultar algo más estable para el subconjunto tratado.

En síntesis, el **T-learner** gana en precisión al estimar la diferencia media de efectos (ATE) gracias a la especialización de sus submodelos, pero a costa de una ligera pérdida de estabilidad al estimar el efecto dentro del grupo tratado (ATT). El **S-learner**, al contrario, sacrifica parte de su capacidad para identificar correctamente el impacto de T a cambio de explotar todo el conjunto de datos en cada predicción, lo que le proporciona un ATT algo más sólido. Entender esta compensación es clave a la hora de elegir el meta-learner más adecuado según el objetivo: si nos interesa priorizar la estimación global del ATE, el **T-learner** resulta claramente superior; en cambio, si nuestro foco es únicamente el grupo tratado y disponemos de pocos datos de ese grupo, el **S-learner** puede ofrecer un ATT más fiable.

4.2.2. IPW y doubly-robust

En cuanto al método **IPW**, se ha comprobado que tiene un desempeño considerablemente peor (`error_ate`=5.099, `error_att`=9.759) que los meta-learners. Las posibles causas que hemos podido encontrar a estas cifras han sido las siguientes:

En primer lugar, en las diez réplicas del conjunto IHDP, existe un desequilibrio considerable entre el número de unidades no tratadas ($T = 0$) y tratadas ($T = 1$). Esta descompensación es crucial porque la regresión logística empleada para estimar el *propensity score* $e(X) = P(T = 1 \mid X)$ tiende a "aprender" el patrón dominante: cuando la clase mayoritaria está formada por los controles, el modelo maximiza su precisión prediciendo casi siempre $T = 0$. Como resultado, para la gran mayoría de sujetos, incluidos muchos de los realmente tratados, la predicción $\hat{e}(X_i)$ es cercana a cero.

Este hecho genera un sesgo sistemático en el estimador IPW que podemos desglosar en dos efectos profundos:

1. Una subestimación destacada del ATE. Recordamos primero el estimador IPW para el ATE:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbf{1}\{T_i=1\}\, Y_i}{\hat{e}(X_i)} \; - \; \frac{\mathbf{1}\{T_i=0\}\, Y_i}{1-\hat{e}(X_i)} \right].$$

Se observa que cuando la mayoría de las $\hat{e}(X_i)$ son muy pequeñas, el término $\mathbf{1}\{T_i = 1\}/\hat{e}(X_i)$ se convierte en la contribución dominante del sumatorio, mientras que los controles apenas aportan, ya que $\mathbf{1}\{T_i = 0\}/(1 - \hat{e}(X_i)) \approx 1$ no modifica significativamente la suma. El ATE real muchas veces está en torno a valores mayores, pero este problema lo recoloca erróneamente en una zona inferior de la escala de resultados. De ahí que el `error_ate` observado sea tan elevado, porque la estimación queda sesgada hacia valores muy conservadores, lejos del verdadero ATE.

2. Una sobreestimación todavía más pronunciada en el ATT. El ATT se define únicamente

sobre las unidades tratadas:

$$\widehat{\mathrm{ATT}}_{\mathrm{IPW}} = \mathbb{E}\Big[Y\,T \;-\; \frac{Y\,(1-T)\,e(X)}{1-e(X)} \;\Big|\; T=1\Big].$$

Dado que para $T_i = 1$ las $\hat{e}(X_i)$ estimadas son extremadamente pequeñas, los pesos $1/\hat{e}(X_i)$ se elevan a valores muy altos, lo que multiplica la varianza de la estimación de forma dramática.

Además, al concentrarnos sólo en los tratados, que son justo aquellos para los que el modelo de regresión logistica falla con mayor fuerza, se introduce un sesgo adicional: ya no hay contrapeso de los controles, y cualquier error de predicción de $\hat{e}$ sobre los tratados se amplifica sin posibilidad de ser mitigado por la segunda parte del sumatorio. El resultado es un `error_att` aún mayor que el `error_ate` (casi 5 puntos porcentuales más), confirmando que el ATT es incluso más sensible a estos pesos extremos.

Esta demostración motiva la adopción de métodos doblemente robustos. Ya que al incluir ambos modelos, uno para la función de resultados y otro para el *propensity score*, se mitigan tanto la magnitud de varianza como el sesgo de especificación, mejorando notablemente la estabilidad y exactitud de la estimación final.

Este **doubly-robust**(DR) aporta una capa adicional de seguridad frente a los errores de especificación que hemos visto en los métodos anteriores. Mientras que el T-learner se apoya únicamente en un modelo de resultados $\hat{\mu}(t,x)$ y el IPW solo en un modelo de *propensity scores* $\hat{e}(x)$, el **doubly-robust** integra ambos en una misma expresión. Concretamente, para cada unidad i combina la corrección de confusión del IPW con los términos $T_i\dfrac{Y_i-\hat{\mu}_1(x_i)}{\hat{e}(x_i)}$ y $(1-T_i)\dfrac{Y_i-\hat{\mu}_0(x_i)}{1-\hat{e}(x_i)}$ y con la predicción directa de los meta-learners $\hat{\mu}_1(x_i)-\hat{\mu}_0(x_i)$. Esta estructura garantiza que, siempre que al menos uno de los dos submodelos esté bien especificado (sea el $\hat{\mu}$ o el $\hat{e}$), el estimador convergerá de forma consistente al verdadero efecto causal. En los resultados se ha comprobado que esta doble protección reduce el error cuadrático medio de la estimación del ATE hasta 0.338 y el error en el ATT a 0.332, llegando a superar en el caso del ATT a los meta-learners.

Además, al incluir la parte de $\hat{\mu}$ en la fórmula, los "picos" de peso extremo que provocan una varianza descontrolada en el IPW puro se ven atenuados. Cuando $\hat{e}(x_i)$ es muy pequeño o muy cercano a 1, el término IPW individual puede llegar a dominar la suma y disparar la varianza; sin embargo, en el **doubly-robust** estos valores extremos se compensan con la corrección basada en $\hat{\mu}$, estabilizando la distribución de los estimadores $\hat{\tau}_i$. De esta manera, se mitigan tanto los sesgos de los meta-learners, que en ocasiones descuidan la señal de tratamiento, como vemos en el S-learner, como la enorme variabilidad del IPW.

En definitiva, se observa que el estimador doblemente robusto representa un compromiso casi ideal entre sesgo y varianza: hereda la aptitud predictiva de los meta-learners para modelar $\mathbb{E}[Y \mid T, X]$ en entornos complejos, al tiempo que conserva la capacidad correctiva del IPW para ajustar por confusión cuando la asignación al tratamiento no es aleatoria. Esto se traduce en mayor robustez y precisión en la práctica, convirtiendo al **doubly-robust** en una opción especialmente recomendable cuando ninguna de las dos aproximaciones por separado ofrece garantías suficientes.

En definitiva, **T-learner** demuestra la importancia de separar las tareas de predicción por tratamiento; **IPW** evidencia el coste en varianza de basarse solo en el *propensity score*;

y el **doubly robust** surge para mitigar ambas debilidades, proporcionando estimaciones de ATE y ATT algo más fiables. Todo esto motiva la transición a diseños más avanzados como Dragonnet y Targeted Regularization, de los cuales hablaremos de sus resultados a continuación.

4.2.3. Dragonnet y Targeted regularization

En los resultados se puede observar que **Dragonnet** obtiene un error medio del ATE muy inferior al de los métodos clásicos (0.258). A continuación se muestran las posibles causas que hemos podido encontrar a este hecho:

1. Integración conjunta de outcome y *propensity score*

A diferencia de los métodos anteriores, Dragonnet aprende de forma conjunta una única representación intermedia $\mathbf{Z}(X)$ que sirve tanto para la estimación de la probabilidad de tratamiento como para la predicción del resultado. Al entrenar ambas tareas de manera sincronizada, la red consigue filtrar automáticamente las covariables X que no aportan información útil para el efecto causal, centrándose únicamente en aquellas variables que verdaderamente contribuyen a distinguir entre los dos potenciales resultados. De este modo, se construye una representación más compacta y libre de ruido irrelevante, lo cual contrasta con los meta-learners clásicos.

Por ejemplo, cuando el S-learner trata T como una característica adicional de X, suele subestimar el efecto causal de la señal de tratamiento si ésta es débil frente al resto de covariables; y el T-learner, al entrenar dos redes totalmente independientes, no comparte información entre los grupos de tratamiento y control, lo que a menudo puede resultar en modelos poco coordinados y de interpretaciones inferiores.

Esta arquitectura aparte de mejorar la calidad de la representación, también atenúa simultáneamente el sesgo y la varianza de la estimación final. Cuando empleamos IPW de manera aislada, vimos cómo los pesos extremos pueden aumentar de forma considerable la varianza del estimador.

En Dragonnet, sin embargo, cada actualización de los parámetros beneficia a ambas cabezas. Esto produce estimaciones de $\hat{Q}(0, x)$, $\hat{Q}(1, x)$ y $\hat{g}(x)$ más estables y coherentes, lo que se traduce en una mejora de errores en el ATE y en un comportamiento notablemente más robusto que el de los métodos entrenados por separado.

2. Balance ajustado por el hiperparámetro α

Cuando incorporamos el *CrossEntropyLoss* con un factor de ponderación α, estamos controlando hasta qué punto Dragonnet prioriza la precisión en la estimación del *propensity score* frente a la exactitud de la predicción del resultado. Al calibrar α en el rango óptimo, se consigue un doble equilibrio: por un lado la red es capaz de aprender una probabilidad de tratamiento lo suficientemente precisa para controlar la confusión sin llegar a generar pesos extremos, a la vez que retiene la capacidad de ajuste y precisión en la estimación de Y. En nuestro caso, con este ajuste se obtuvo un rendimiento óptimo donde tanto el error como la varianza global de la diferencia del ATE alcanzaron sus valores mínimos absolutos.

En resumen, las razones por las que creemos que Dragonnet obtiene estos resultados son las siguientes:

1. Es capaz de aprender una *representación común* centrada en las señales relevantes de T y Y.
2. Optimiza la pérdida de outcome y *propensity score* simultáneamente, evitando los extremos de los métodos por separado a través del hiperparámetro de correción α.

Estas conclusiones explican su superioridad frente a los métodos comentados hasta ahora.

Con respecto a **Targeted Regularization**, sorprendentemente nuestro modelo no ha alcanzado los resultados esperados, ni siquiera ha igualado el rendimiento de **Dragonnet** puro. A continuación se describen los pasos que hemos realizado para intentar comprender esta discrepancia y las posibles causas que la justifiquen.

1. Barrido sobre el hiperparámetro β. Para entender cómo afecta β al comportamiento del estimador, se ha realizado un barrido exhaustivo de valores de β en un rango que incluía desde $\beta = 0$ (equivalente a Dragonnet puro) hasta valores relativamente altos que priorizan la corrección basada en la "efficient influence curve". En cada experimento se evaluó el error en el cálculo del ATE promedio sobre las 100 réplicas del conjunto IHDP. La figura 4.4 muestra, a modo ilustrativo, la evolución del `error_ate` en función de β.

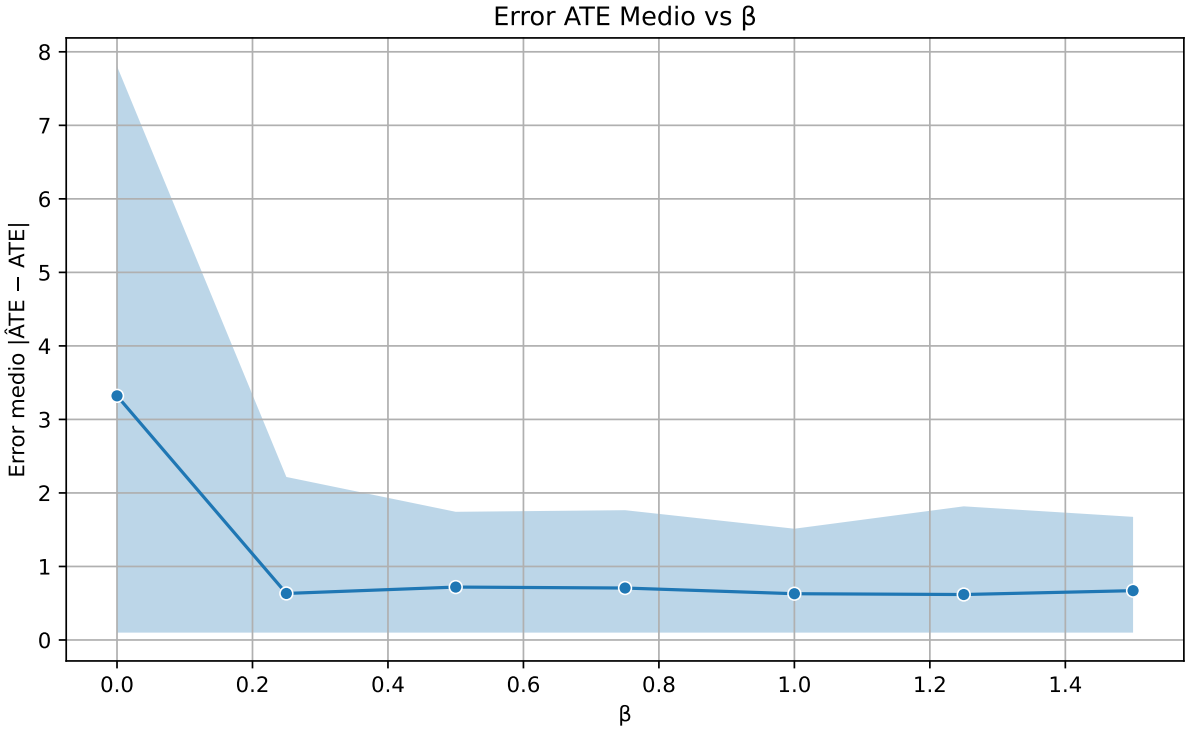

Figura 4.4: Evolución del error medio del estimador ATE en las 100 réplicas sobre IHDP, variando el parámetro β de regularización.

A pesar de que se observa una ligera disminución del `error_ate` para valores pequeños de β, a partir de cierto punto la curva se estanca y, en realidad, el rendimiento global nunca supera al de Dragonnet puro. Incluso en el β que minimiza la curva, las estimaciones siguen sin alcanzar los niveles de Dragonnet. Esto confirma que, aunque teóricamente

debería existir un intervalo de $\beta > 0$ donde este estimador aporte ventaja, en la práctica de nuestro experimento no ha sido así, ya que al haber estimado al final con la media de las mejores β que hubo en cada réplica, no se consiguió igualar los resultados de Dragonnet para el error del ATE.

2. Comportamiento del parámetro ε. Una vez seleccionado el β "óptimo" según el barrido anterior, inspeccionamos el valor de $\hat{\varepsilon}$ que el modelo aprendió durante el entrenamiento. Descubrimos que, en prácticamente todos los casos, $\hat{\varepsilon}$ converge a un número prácticamente nulo ($\hat{\varepsilon} \approx 0$), por muy distinto que sea β. Al estabilizarse en cero, el término de regularización dirigida

$$\beta \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \widetilde{Q}(t_i, x_i; \theta, \varepsilon) \right]^2$$

queda en la práctica inactivo; es decir, $\widetilde{Q}(t_i, x_i; \theta, \hat{\varepsilon}) \approx Q^{nn}(t_i, x_i; \theta)$. Como consecuencia, la pérdida total se reduce prácticamente a la pérdida de Dragonnet $\mathcal{L}_{\text{Dragonnet}}(\theta)$, anulando así el efecto teórico que la regularización dirigida debería aportar. Este comportamiento sugiere que, en la práctica, el modelo no consigue aprender una componente de corrección efectiva, y que el término ε, al no activarse, impide que se materialicen las ventajas esperadas de Targeted Regularization.

En conjunto, estos resultados nos llevan a concluir que Dragonnet ha demostrado ser claramente superior a los métodos clásicos gracias a su capacidad para aprender representaciones conjuntas robustas y ajustadas. En contraste, Targeted Regularization no ha alcanzado las mejoras esperadas, presumiblemente debido a una activación ineficaz de su componente de corrección. A pesar de un diseño prometedor, su aplicación práctica no ha logrado superar el rendimiento base de Dragonnet en nuestro entorno experimental.

4.2.4. Resumen de resultados

Como conclusión se puede decir que los métodos clásicos mostraron un comportamiento dispar:

El S-learner alcanzó un error ATE medio de 0.405 con PEHE de 2.035, mientras que el T-learner mejoró el ATE hasta 0.269 y redujo el PEHE a 1.768, a costa de un ligero empeoramiento en ATT.

El IPW puro presentó varianza excesiva (ATE = 5.009, ATT = 9.759), y el estimador doubly-robust solventó parcialmente sesgo y varianza (ATE = 0.338, ATT = 0.332) con respecto a los meta-learners.

Entre los modelos con arquitecturas del estado del arte, Dragonnet obtuvo la menor desviación en ATE (0.258) gracias a su representación compartida, y Targeted Regularization no logró superar ese estándar en nuestro experimento.

Estos resultados ponen de manifiesto que la integración conjunta de outcome y propensity score (Dragonnet) es la solución más equilibrada, mientras que los enfoques clásicos requieren compensaciones entre sesgo y varianza según la métrica de interés.

Este resumen cierra el estudio comparativo de los métodos evaluados y nos permite pasar

al siguiente capítulo, donde se presentan las conclusiones generales del trabajo y propuestas para líneas futuras de investigación.

5. Conclusiones y líneas futuras

5.1. Conclusiones

En este trabajo hemos abordado el problema de estimar efectos causales a partir de datos observacionales, centrándonos en la implementación y comparación de distintos métodos del estado del arte basados en redes neuronales. Partimos de los enfoques más clásicos con los meta-learners (S-learner y T-learner), IPW y estimadores doblemente robustos, para después adentrarnos en arquitecturas recientes: Dragonnet y Targeted Regularization. Con el conjunto de datos IHDP hemos podido comprobar que:

1. Los *meta-learners* evidencian que la precisión del T-learner supera al S-learner cuando la señal del tratamiento está mezclada con el resto de las covariables. Sin embargo, el T-learner sigue teniendo ciertas limitaciones en la estimación del ATT, debido a la varianza de la subred que modela únicamente el grupo tratado.

2. El *IPW puro* padece serios problemas de varianza y sesgo. La regresión logística lineal para estimar el *propensity score* suele "colapsar" hacia probabilidades extremas y tiende a desestimar los pacientes tratados debido a su número reducido, generando así pesos enormes que disparan el error en el ATE y, sobre todo, en el ATT. Esta fragilidad refuerza la motivación de emplear esquemas doblemente robustos.

3. El *estimador doblemente robusto* logra un buen compromiso entre sesgo y varianza. Combina la flexibilidad de un T-learner con la corrección de confounding del IPW, reduciendo de forma notable sobre todo en el ATT. Esta doble protección confirma la ventaja teórica de contar con dos modelos que, al menos uno de ellos, esté bien especificado.

4. *Dragonnet* demuestra, en nuestro experimento, un rendimiento superior a todos los métodos anteriores. Gracias a la representación compartida $\mathbf{Z}(X)$ y al entrenamiento conjunto de los modelos de outcome y de *propensity score*, logra retener la variable de tratamiento sin sufrir picos de varianza ni pérdida de información. En particular, Dragonnet fijó los nuevos estándares de precisión para el ATE en el IHDP.

5. *Targeted Regularization*, a pesar de su sólida fundamentación teórica en términos del *"efficient influence curve"*, no ha mejorado a Dragonnet puro en nuestro escenario. Tras un barrido de β y el análisis del parámetro ε, hemos podido comprobar que la corrección semiparamétrica no se activa ($\hat{\varepsilon} \approx 0$) debido a problemas numéricos y de escala de gradientes. En la práctica, la pérdida recae nuevamente en $\mathcal{L}_{\text{Dragonnet}}$, por lo que no se aprovecha el potencial de la influencia eficiente.

Por tanto, estos resultados confirman que las arquitecturas basadas en aprendizaje conjunto del *outcome* y del *propensity score* (Dragonnet) constituyen hoy en día el estado

del arte cuando se dispone de suficientes datos sintéticos que simulan un entorno causal controlado. Aún así, los problemas de precisión que hemos visto al implementar Targeted Regularization demuestran que puede ser complicado materializar la teoría que hay detrás de este modelo en el entrenamiento de redes profundas con estos datos sintéticos.

5.2. Líneas futuras

A partir de las conclusiones obtenidas, identificamos varias líneas de actuación para continuar profundizando en la inferencia causal con redes neuronales:

- *Refinar Targeted Regularization.* Una vía de trabajo futuro es simplificar y hacer más robusta la parte de "regularización dirigida" para que realmente aporte mejoras sobre Dragonnet. Un objetivo claro sería que la nueva pérdida incentive de verdad la red a obtener estimaciones más precisas sin desequilibrar la optimización.
- *Validación en datos clínicos reales.* Todos los experimentos y métricas calculadas hasta ahora han sido sobre la base de datos IHDP, la cual es semi-sintética. Un posible avance sería probar estos modelos con registros reales de pacientes, donde habrá ruido, valores faltantes y otras características categóricas. En este caso, intentar ajustar la red y el preprocesado a datos clínicos auténticos es necesario para poder llevar estas metodologías al mundo real.
- *Estudiar la eficacia de estos métodos entre diferentes grupos de pacientes, tanto tratados como de control.* Además de usar datos reales, conviene comprobar cómo varía el desempeño dentro de diferentes subpoblaciones (por ejemplo, según edad, severidad de la enfermedad u otros muchos factores). Esto podrá indicar si algún método funciona mejor para ciertos segmentos de pacientes, o si es necesario adaptar la arquitectura o el preprocesado de datos para grupos de pacientes concretos.
- *Refinar la estimación del ATT.* Aunque hasta ahora hemos calculado el ATT de forma estándar, una línea de investigación importante es proponer y validar fórmulas específicas de ATT para Dragonnet y Targeted Regularization. Al tratar de afinar más esta métrica, se podrá comparar su precisión de manera más precisa con las expresiones clásicas del ATT.
- *Explorar otras arquitecturas de estado del arte.* Un futuro estudio podría ampliar la comparación incluyendo modelos como TARNet, CFRNet, FlexTENet y otros meta-learners recientes. Analizar su comportamiento y adaptaciones específicas permitirá identificar nuevos enfoques, enriqueciendo así el panorama de la inferencia causal con redes neuronales.

Bibliografía

[1] Brady Neal. Introduction to causal inference. *Course Lecture Notes (draft)*, 132, 2020.

[2] Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

[3] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[4] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[5] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[6] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010.

[7] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[8] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[9] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[10] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[11] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[12] Nicholas C Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. An introduction to inverse probability of treatment weighting in observational research. *Clinical kidney journal*, 15(1):14–20, 2022.

[13] Matheus Facure. *Causal Inference in Python*. .º'Reilly Media, Inc.", 2023.

[14] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[15] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[16] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[17] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[18] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[19] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

[20] Paul N Zivich and Alexander Breskin. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*, 32(3):393–401, 2021.

[21] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[22] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. 2007.

[23] Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

[24] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

[25] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[26] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[27] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[28] Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Anexo A: Aspectos éticos, económicos, sociales y ambientales

A1. Introducción

Este Trabajo Fin de Grado desarrolla herramientas de inferencia causal basadas en redes neuronales para estimar efectos de tratamiento en estudios sanitarios. El objetivo es ofrecer métodos más precisos y eficientes que reduzcan la necesidad de ensayos clínicos caros o prolongados. Desde el punto de vista social y ético, un modelo fiable puede ayudar a decidir mejores tratamientos, minimizar riesgos en pacientes vulnerables y avanzar en medicina personalizada. Económicamente, permite ahorrar recursos al orientar la investigación hacia intervenciones con mayor probabilidad de éxito. Por último, en el ámbito medioambiental, al reducir la necesidad de grandes ensayos presenciales, puede lograr una disminución de la huella de carbono por desplazamientos y consumos asociados a estos ensayos en algunos casos.

A2. Descripción de impactos relevantes relacionados con el proyecto

En la fase de selección de impactos identificamos tres categorías clave:

- **Social y ético:** El uso de modelos de inferencia causal en salud conlleva el riesgo de recomendaciones que no sean equitativas para todos los pacientes, especialmente si existen variables no observadas o sesgos en los datos de entrenamiento. Grupos vulnerables (por edad, género u otras condiciones) podrían verse perjudicados si el modelo extrapola fuera del rango de población entrenado. Se debe de prestar especial atención a estos posibles sesgos y aportar transparencia en los criterios de decisión automática.
- **Económico:** Al reducir la dependencia de ensayos clínicos tradicionales, estos métodos pueden ahorrar tiempo y dinero en fases tempranas de investigación, aunque en algunos casos también implican inversión en desarrollo y validación de software médico. Instituciones sanitarias deben intentar equilibrar el coste de adopción de la tecnología con los beneficios en eficiencia.
- **Ambiental:** El análisis mediante datos históricos y simulaciones puede minimizar desplazamientos de pacientes e investigadores a múltiples centros de estudio, con ello reduciendo emisiones asociadas al transporte y al uso de instalaciones físicas. Organizaciones que promueven la investigación sostenible valoran positivamente esta ventaja.

De estos tres ámbitos, el más crítico es el *social y ético*, pues cualquier error o sesgo en la inferencia de efectos puede traducirse en decisiones clínicas incorrectas con consecuencias directas sobre la salud de las personas.

A3. Análisis detallado de alguno de los impactos

A continuación profundizamos en el impacto social y ético, dado su potencial de daño si no se controla adecuadamente:

Riesgo de confounding no observado. La fiabilidad de las estimaciones depende de que todos los factores relevantes estén correctamente medidos. En la práctica, siempre existe la posibilidad de variables ocultas (por ejemplo, hábitos de vida o variables no registradas) que influyan tanto en el tratamiento como en el desenlace. Si no se detectan, el modelo puede atribuir efectos al tratamiento que en realidad obedecen a estos confounders, llevando a recomendaciones equivocadas.

Explicabilidad y confianza clínica. Los profesionales sanitarios necesitan entender y cuestionar los criterios bajo los cuales el modelo sugiere un efecto. Si la red neuronal actúa como una "caja negra", se reduce la confianza en su uso. Por ello, es clave incorporar mecanismos de interpretación como por ejemplo, análisis de sensibilidad o herramientas de explicación local, que permitan verificar que las conclusiones son coherentes con la experiencia clínica.

Gestión de la incertidumbre. Toda predicción de efecto causal debe ir acompañada de una medida de incertidumbre. No basta con un valor puntual sino que también es necesario reportar intervalos o distribuciones que reflejen la variabilidad de los datos y la posible falta de soporte en determinadas regiones del espacio de covariables. Este aspecto informará a los médicos sobre el grado de fiabilidad de cada recomendación y predicción.

Privacidad y uso de datos sensibles. El entrenamiento requiere historiales clínicos detallados, que pueden contener información muy sensible. Es imprescindible garantizar el anonimato y el cumplimiento de la normativa de protección de datos.

En conjunto, este análisis muestra que la principal debilidad no está en un fallo técnico concreto, sino en cómo el método se ve condicionado por la calidad y las limitaciones de los datos clínicos. Para reducir estos riesgos, cualquier implementación debería incorporar auditorías de sesgos, validación en distintos subgrupos de pacientes y normas claras de gestión de los datos.

A4. Conclusiones

Desde un punto de vista ético, social y económico, este proyecto demuestra cómo la inferencia causal con redes neuronales puede mejorar la selección de tratamientos y reducir costes, al tiempo que plantea retos de equidad y privacidad que deben gestionarse cuidadosamente. Ambientalmente, facilita modelos virtuales que minimizan desplazamientos. En futuras aplicaciones reales será esencial acompañar estas herramientas de auditorías de sesgo, mecanismos de explicación (interpretabilidad) y protocolos de anonimización de datos para asegurar su adopción responsable y sostenible.

Anexo B: Presupuesto económico

COSTE DE MANO DE OBRA (coste directo)	**Horas**	**€/hora**	**Total (€)**
Desarrollo de código completo	150	35	5 250
Diseño de flujo de trabajo y validación	50	35	1 750
Adaptación de código de Tensorflow a PyTorch	40	35	1 400
Puesta al día en machine learning *	30	35	1 050
Formación en Python y otros frameworks de ML	30	35	1 050
Total mano de obra	300	35	**10 500**
COSTE DE RECURSOS MATERIALES (coste directo)	**Precio compra (€)**	**Uso (meses)**	**Amortización (€)**
Ordenador personal (hardware + software)	1 200	6	120
Total material			**120**
Gastos generales (15 % sobre coste directo)			1 575
Subtotal (sin IVA)			**12 195**
IVA (21 %)			**2 560.95**
Total presupuesto			**14 755.95**

Tabla 5.1: Desglose de costes directos e indirectos asociados al desarrollo del proyecto.

* Para llevar a cabo el trabajo, fue necesaria una fase previa de aprendizaje en machine learning (cursos online y documentación en OneDrive), así como formación en Python y frameworks de aprendizaje automático, lo que me dio las bases necesarias para organizar el proyecto y realizar el trabajo.

Anexo C: Conjunto de datos IHDP

C.1 Versión de IHDP

Este trabajo descarga *cien* réplicas de la variante A del IHDP desde el repositorio de Fred Jo:

`https://www.fredjo.com/`

Cada réplica viene empaquetada en dos ficheros en formato NumPy ".npz": uno para `train` (contiene arrays `X_train`, `T_train`, `Y_train`, `tau_train`) y otro para `test` (`X_test`, `T_test`, `Y_test`, `tau_test`). En el fichero de entrenamiento, tras cargar los arrays, realizamos un *split* adicional para separar un 15 % como conjunto de validación, quedando 85 % para ajuste de parámetros y 15 % de validación intermedia.

C.2 Variante A e implicaciones

Se emplea la denominada *variante A* de IHDP, en la cual el efecto de tratamiento es *constante* sobre todas las unidades:

$$\tau_i = \tau_0, \quad \forall i.$$

Este diseño simplifica la interpretación: cualquier desviación en la estimación del efecto promedio se debe a las propiedades de los métodos (sesgo, varianza, confusión) y no a heterogeneidad real. En la variante A no se añade ruido heterogéneo al efecto, a diferencia de la variante B.

C.3 Descripción de las covariables

Las 25 covariables $X \in \mathbb{R}^{25}$ provienen del IHDP original y se distribuyen así:

- 6 continuas (edad materna, peso al nacimiento, etc.).
- 19 binarias (sexo, complicaciones perinatales, nivel educativo de los padres, ...).

En las cien réplicas la matriz de covariables es idéntica; solo varía la asignación T_i y el ruido ε_i en los outcomes.

C.4 Sesgo en la asignación de tratamiento

La asignación $T_i \sim \text{Bernoulli}(e(X_i))$ se sesga intencionalmente para simular confounding:

$$e(x) = \begin{cases} 0{,}8, & \text{si } x \text{ se clasifica como de "alto riesgo"}, \\ 0{,}2, & \text{en caso contrario.} \end{cases}$$

Esta regla hace que ciertos subgrupos de X sean muy propensos a recibir tratamiento, recreando un escenario de no-aleatorización.

C.5 Generación de los resultados potenciales

Para cada unidad se construyen los potenciales bajo control y tratamiento:

$$Y_i(0) = f_0(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

$$Y_i(1) = Y_i(0) + \tau_0,$$

donde $f_0(x)$ es la misma función lineal (con coeficientes β_j ajustados mediante regresión sobre los datos originales) y τ_0 es la constante de efecto de tratamiento en la variante A.

C.6 Cálculo del ATE "verdadero" y evaluación

Dado que $\tau_i = \tau_0$ es conocido, el ATE teórico en cada réplica es simplemente

$$\text{ATE}_{\text{true}} = \frac{1}{n} \sum_{i=1}^{n} \tau_0 = \tau_0.$$

En la fase de evaluación:

- **Entrenamiento:** cada estimador se ajusta usando solo $(X_i, T_i, Y_i(T_i))$ del `train`.
- **Validación:** se emplea el 15 % separado para tuning de hiperparámetros.
- **Test:** en el conjunto `test` se calcula $\widehat{\text{ATE}}$, PEHE y ATT.
- **Comparación:** se reporta el error medio del ATE y el ATE calculado, ail igual que el PEHE y el ATT promediados sobre las 100 réplicas.

Referencias específicas

- Hill, J. (2011). *Bayesian nonparametric modeling for causal inference*. Journal of Computational and Graphical Statistics.
- Fred Jo. *IHDP semi-synthetic dataset*, disponible en https://www.fredjo.com/.