Pseudo-etiquetado en datos tabulares: de los clasificadores clásicos a los modelos fundacionales

Juan J. Rodríguez, José L. Garrido-Labrador, Jesús M. Maudes-Raedo, and César I. García-Osorio

Departamento de Ingeniería Informática. Universidad de Burgos

Resumen Los modelos fundacionales tabulares como *TabPFN* están redefiniendo el aprendizaje sobre datos tabulares. Es razonable preguntarse si las técnicas clásicas de aprendizaje semi-supervisado (SSL) por pseudo-etiquetado (como *self-training* y *tri-training*), concebidas para clasificadores tradicionales, siguen siendo útiles, o incluso se comportan igual, cuando el clasificador base es un modelo fundacional. En datos tabulares, estos métodos de tipo envoltorio (*wrapper*) son atractivos por su carácter *plug-and-play*: pueden combinarse con cualquier clasificador base. Sin embargo, su utilidad real en este dominio dista de ser uniforme. En este trabajo evaluamos empíricamente *self-training* y *tri-training* sobre cuatro clasificadores base de naturaleza muy distinta: *Random Forest*, regresión logística, *gradient boosting* (*LightGBM*) y un modelo fundacional tabular (*TabPFN*), en 31 conjuntos de datos usando solo un 5% de datos etiquetados. Nuestro hallazgo principal es que *el efecto del pseudo-etiquetado depende del clasificador base*. Perjudica de forma significativa a Random Forest, no produce mejora significativa en regresión logística, y beneficia de forma significativa a *LightGBM* y al modelo fundacional *TabPFN*. Un análisis de las pseudo-etiquetas muestra que su calidad no explica el patrón observado, que parece depender más de propiedades del propio clasificador base.

Keywords: Aprendizaje semi-supervisado · Datos tabulares · Pseudo-etiquetado · *Self-training* · Modelos fundacionales · *TabPFN*

1 Introducción

El aprendizaje supervisado requiere conjuntos de datos etiquetados, cuya obtención es costosa en muchos dominios reales. El aprendizaje semi-supervisado (SSL) aborda esta limitación combinando una pequeña cantidad de instancias etiquetadas con un conjunto mayor de instancias no etiquetadas, con el objetivo de mejorar los modelos que se obtendrían usando solo las etiquetadas [3].

Dentro de la taxonomía de van Engelen y Hoos [3], los métodos *wrapper* ocupan un lugar destacado por su sencillez y generalidad: emplean un clasificador base supervisado para etiquetar instancias no etiquetadas (pseudo-etiquetas) y reentrenan incorporando las más fiables. El más conocido es *self-training* [15], donde un único clasificador se entrena inicialmente con los datos etiquetados y va

incorporando de forma iterativa aquellas instancias no etiquetadas para las que produce predicciones con mayor confianza. *Tri-training* [19] extiende la idea con tres modelos que se etiquetan mutuamente. Su principal atractivo es el carácter *plug-and-play*: son agnósticos al clasificador base, por lo que pueden combinarse con cualquier algoritmo supervisado sin modificarlo.

Buena parte de la investigación en SSL se ha centrado en imágenes y texto. En datos tabulares, donde los modelos basados en árboles han dominado tradicionalmente, el panorama es menos claro, y la irrupción reciente de los modelos fundacionales tabulares (en particular *TabPFN*) plantea una pregunta natural: ¿se benefician estos modelos del pseudo-etiquetado clásico igual que los clasificadores tradicionales? Más en general, dado el carácter plug-and-play de estos métodos, cabe preguntarse si su utilidad es uniforme o si depende del clasificador base con el que se combinan.

Este trabajo aborda esa pregunta mediante un estudio empírico comparativo. Las preguntas de investigación que guían el trabajo son:

- **PI1.** ¿Resulta útil el pseudo-etiquetado (*self-training*, *tri-training*) en datos tabulares con pocas etiquetas?
- **PI2.** ¿Depende el efecto del clasificador base empleado?
- **PI3.** ¿Cuándo ayuda y cuándo perjudica el pseudo-etiquetado, y por qué?

Las contribuciones del trabajo son: (i) un estudio comparativo del pseudo-etiquetado *plug-and-play* a través de cuatro clasificadores base en 31 conjuntos tabulares, que muestra que el efecto del pseudo-etiquetado depende del clasificador; (ii) la caracterización de este efecto en relación al número de clases en el conjunto de datos; y (iii) un análisis de la calidad de las pseudo-etiquetas que muestra que esta no explica el patrón observado. Por ejemplo, *Random Forest* genera las pseudo-etiquetas más precisas y aun así es el más perjudicado, lo que parece indicar que el efecto final está muy influenciado por el clasificador base concreto que se está usando. El resto del artículo se organiza como sigue: la Sección 2 revisa el trabajo relacionado; la Sección 3 describe la metodología; la Sección 4 presenta los resultados; la Sección 5 los discute; y la Sección 6 concluye.

2 Trabajo relacionado

2.1 Pseudo-etiquetado y métodos auto-etiquetados

Triguero et al. [15] ofrecen una taxonomía y estudio empírico de los métodos auto-etiquetados, situando a *self-training* y a las variantes de *co-training* en un marco común. Zhou y Li [19] introducen *tri-training*, que evita la necesidad de vistas independientes de los datos entrenando tres clasificadores sobre el mismo conjunto. La idea del pseudo-etiquetado se ha popularizado también en el aprendizaje profundo a partir del trabajo de Lee [11], donde las predicciones de mayor confianza sobre los datos no etiquetados se emplean como etiquetas para el reentrenamiento. Estos métodos son la base de la familia *wrapper* estudiada en este trabajo.

2.2 Panorama del SSL

Para una visión general del campo, van Engelen y Hoos [3] ofrecen la taxonomía de referencia, y Yang et al. [17] revisan los avances más recientes en SSL profundo (regularización por consistencia, métodos generativos y de pseudo-etiquetado). La combinación de métodos de ensemble con SSL ha recibido también atención específica [5].

El SSL aplicado específicamente a datos tabulares ha recibido atención creciente. Yoon et al. [18] proponen VIME (*Value Imputation and Mask Estimation*), un marco de aprendizaje auto y semi-supervisado que induce representaciones útiles a partir de datos no etiquetados mediante la corrupción controlada de las entradas y el entrenamiento del modelo para reconstruir los valores originales e identificar qué posiciones han sido alteradas. Los autores muestran que este enfoque mejora al aprendizaje puramente supervisado cuando se dispone de pocas etiquetas. Ramírez-Sanz et al. [13], por su parte, proponen una versión semi-supervisada de Rotation Forest y la evalúan frente a métodos *wrapper* sobre conjuntos UCI. A diferencia de estos trabajos, que proponen métodos nuevos, el nuestro compara métodos de pseudo-etiquetado *plug-and-play* a través de distintos clasificadores base.

2.3 Modelado de datos tabulares

A pesar del éxito del aprendizaje profundo en visión y lenguaje, su superioridad sobre los métodos basados en árboles no está clara en datos tabulares [14,1]. Estudios comparativos amplios concluyen que los modelos de *gradient boosting* siguen siendo competitivos o superiores en conjuntos de tamaño pequeño y medio [7,12]. En este trabajo incluimos como clasificadores base tanto *Random Forest* como *gradient boosting* (*LightGBM* [10]), junto con un modelo lineal y un modelo fundacional, cubriendo así familias con sesgos inductivos muy distintos.

2.4 Modelos fundacionales tabulares

Aunque los modelos basados en árboles de decisión (en particular el *gradient boosting*) han dominado el aprendizaje sobre datos tabulares durante las últimas dos décadas, recientemente han surgido modelos fundacionales para este tipo de datos. *TabPFN* [9] es un modelo basado en *transformers* que usa aprendizaje en contexto (*In-Context Learning*, *ICL*): en lugar de entrenarse para cada problema, ha sido preentrenado sobre una gran colección de conjuntos sintéticos, de modo que predice condicionando sobre las instancias etiquetadas que recibe como entrada. Resulta especialmente competitivo en conjuntos pequeños, donde la falta de etiquetas puede tener mayor impacto. En este trabajo empleamos *TabPFN-3* [6], la versión más reciente, que maneja de forma nativa los valores ausentes y admite problemas con muchas clases. Un aspecto relevante para la validez del estudio es que *TabPFN-3* se preentrena *exclusivamente* con datos sintéticos, por lo que no puede haber sido expuesto a los conjuntos de datos

utilizados en la comparación experimental de este estudio. Su naturaleza, radicalmente distinta de la de los clasificadores clásicos, motiva preguntarse si su comportamiento frente al pseudo-etiquetado difiere también del de aquellos; esa es una de las cuestiones centrales de este trabajo.

3 Metodología

3.1 Métodos comparados

Comparamos tres configuraciones para cada clasificador base:

- **Supervisado**: el clasificador base entrenado únicamente con las instancias etiquetadas. Sirve de referencia.
- **Self-training**: pseudo-etiquetado iterativo de un único clasificador, incorporando las predicciones cuya confianza supera un umbral (0.75, valor por defecto de la implementación de scikit-learn que empleamos).
- **Tri-training** [19]: tres clasificadores que se etiquetan mutuamente. Empleamos la implementación de `sslearn` [4].

3.2 Clasificadores base

Seleccionamos cuatro clasificadores de naturaleza deliberadamente distinta:

- ***Random Forest* (RF)**: ensemble de *bagging* sobre árboles.
- **Regresión logística (LogReg)**: modelo lineal.
- ***LightGBM* (LGBM)**: *gradient boosting* sobre árboles [10].
- ***TabPFN***: modelo fundacional tabular basado en inferencia en contexto, *ICL* (*TabPFN-3* [6], paquete `tabpfn` 8.0.6).

Cada clasificador recibe el preprocesamiento adecuado a su naturaleza. Random Forest y regresión logística codifican las variables categóricas mediante *one-hot*, lo que evita introducir un orden artificial en variables nominales; la regresión logística además estandariza las numéricas. *LightGBM* y *TabPFN* gestionan las categóricas de forma nativa. Todos los preprocesamientos se ajustan exclusivamente con datos de la partición de entrenamiento (la parte etiquetada en el caso supervisado, y las instancias etiquetadas y pseudo-etiquetadas disponibles en cada reentrenamiento en los métodos SSL).

3.3 Conjuntos de datos

Empleamos 31 conjuntos de datos de OpenML [16], que cubren problemas binarios y multiclase, con variables numéricas, categóricas y mixtas. Las instancias con valores ausentes se eliminan y, a continuación, los conjuntos con más de 3000 instancias se submuestrean a 3000 de forma estratificada para acotar el coste computacional. En `letter` se conservan las 15 primeras clases (orden alfabético) y en `nursery` se descarta la clase residual `recommend` (2 instancias). Se

Tabla 1. Conjuntos de datos empleados (n: instancias tras submuestreo y eliminación de instancias con valores ausentes; d: atributos; K: clases; tipo: num=numérico, cat=categórico, mix=mixto).

Conjunto	n	d	K	tipo	Conjunto	n	d	K	tipo
`adult`	3000	14	2	mix	`mfeat-zernike`	2000	47	10	num
`balance-scale`	625	4	3	num	`mushroom`	3000	22	2	cat
`bank-marketing`	3000	16	2	mix	`nursery`	3000	8	4	cat
`breast-w`	683	9	2	num	`optdigits`	3000	64	10	num
`car`	1728	6	4	cat	`page-blocks`	3000	10	5	num
`credit-approval`	653	15	2	mix	`pendigits`	3000	16	10	num
`credit-g`	1000	20	2	mix	`qsar-biodeg`	1055	41	2	num
`diabetes`	768	8	2	num	`satimage`	3000	36	6	num
`ecoli`	336	7	8	num	`segment`	2310	19	7	num
`heart-statlog`	270	13	2	num	`sonar`	208	60	2	num
`ionosphere`	351	34	2	num	`spambase`	3000	57	2	num
`iris`	150	4	3	num	`tic-tac-toe`	958	9	2	cat
`kr-vs-kp`	3000	36	2	cat	`vehicle`	846	18	4	num
`letter`	3000	16	15	num	`wdbc`	569	30	2	num
`mfeat-factors`	2000	216	10	num	`wine`	178	13	3	num
`mfeat-karhunen`	2000	64	10	num					

adopta esta política por uniformidad: aunque *LightGBM* y *TabPFN-3* manejan los valores ausentes de forma nativa, ni la regresión logística ni las implementaciones de los métodos *wrapper* empleadas los admiten, de modo que eliminarlos da un trato idéntico a los cuatro clasificadores y permite aislar el efecto del pseudoetiquetado de los posibles artefactos introducidos por las técnicas de imputación. Esta eliminación de instancias con valores ausentes solo afecta a cuatro conjuntos (`adult`, `breast-w`, `credit-approval` y `mushroom`), con eliminaciones entre el 2 % y el 31 % de las instancias.

3.4 Protocolo experimental

Para simular el escenario SSL, en cada conjunto se reserva un 20 % como test y, del 80 % restante, solo el 5 % conserva su etiqueta, preservando las proporciones de clase (con un tamaño etiquetado total mínimo de $2K$ instancias, que solo aumenta la fracción etiquetada en `ecoli`). El resto se trata como no etiquetado. Cada experimento se repite con 8 particiones estratificadas distintas (los modelos base usan una semilla fija, de modo que la variabilidad medida es la de la partición de los datos). Como métrica empleamos la *balanced accuracy*, definida como la media de las tasas de acierto por clase (el promedio de los *recalls* de cada clase). Al ponderar todas las clases por igual con independencia de su frecuencia, evita que las clases mayoritarias dominen la métrica en los problemas desequilibrados.

Para cada conjunto de datos se promedian las 8 repeticiones, y todos los contrastes estadísticos se realizan sobre los 31 valores agregados por conjunto (no

Tabla 2. Ganancia media en *balanced accuracy* respecto al supervisado (Δ), por clasificador base y método. Δ se expresa en puntos absolutos de *balanced accuracy* (p.ej.+0.016 significa +1.6 puntos). Valores positivos indican que el pseudo-etiquetado mejora. Entre paréntesis, p del test de Wilcoxon con corrección de Holm. *n.s.*: no significativo.

Clasificador	Δ *self-training*	Δ *tri-training*
RF	-0.048 ($p < 0.001$)	-0.008 (n.s.)
LogReg	-0.013 (n.s.)	-0.003 (n.s.)
LGBM	$+0.012$ ($p = 0.008$)	$+0.007$ ($p = 0.008$)
TabPFN	$+0.016$ ($p = 0.006$)	$+0.008$ ($p = 0.006$)

sobre las repeticiones individuales, para evitar pseudorreplicación). Empleamos el test de los rangos con signo de Wilcoxon (pareado por conjunto) entre cada método SSL y el supervisado; la corrección de Holm se aplica sobre la familia de comparaciones de cada clasificador base (los dos métodos SSL frente al supervisado). Para su comparación usamos el test de Friedman seguido del *post-hoc* de Nemenyi, de nuevo por clasificador base [2].

3.5 Reproducibilidad

Los experimentos se realizan con `lightgbm` 4.6.0, `sslearn` 1.1.0 [4], `tabpfn` 8.0.6 (modelo *TabPFN-3*) y `scikit-learn` 1.6.1; los contrastes estadísticos usan `stats-models` 0.14.6 y `scikit-posthocs` 0.14.0. *Random Forest* y *LightGBM* emplean 100 estimadores, y la regresión logística la regularización por defecto. El *self-training* itera con umbral de confianza 0.75 hasta que no se incorporan nuevas instancias (*tri-training* emplea la configuración por defecto de `sslearn`). Los experimentos se ejecutan en Google Colab; *TabPFN* se ejecuta sobre una GPU NVIDIA T4 (PyTorch 2.11), mientras que el resto de métodos se ejecutan en CPU.

4 Resultados

4.1 Efecto del pseudo-etiquetado según el clasificador base

El resultado central es que el efecto del pseudo-etiquetado depende del clasificador base. La Tabla 2 resume la ganancia media en *balanced accuracy* respecto al supervisado, revelando que el pseudo-etiquetado no tiene un efecto uniforme, sino que interactúa de manera sustancial con el clasificador base. En *Random Forest*, el *self-training* degrada claramente el rendimiento ($p < 0.001$). En regresión logística, las diferencias son pequeñas y no significativas. Por el contrario, *LightGBM* y *TabPFN* se benefician del pseudo-etiquetado, con mejoras significativas y con la mayor ganancia observada en *TabPFN* ($+0.016$, $p = 0.006$). La coincidencia del patrón bajo *self-training* y *tri-training* apunta a una propiedad

Tabla 3. *Balanced accuracy* media (entre los 31 conjuntos) de cada método por clasificador base, y balance de victorias/derrotas del *self-training* frente al supervisado por conjunto (W/L/T).

Clasificador	Supervisado	*Self-training*	*Tri-training*	W/L/T (self)
RF	**0.735**	0.687	0.728	3/28/0
LogReg	**0.746**	0.734	0.744	11/20/0
LGBM	0.593	**0.604**	0.600	15/3/13
TabPFN	0.767	**0.783**	0.775	23/8/0

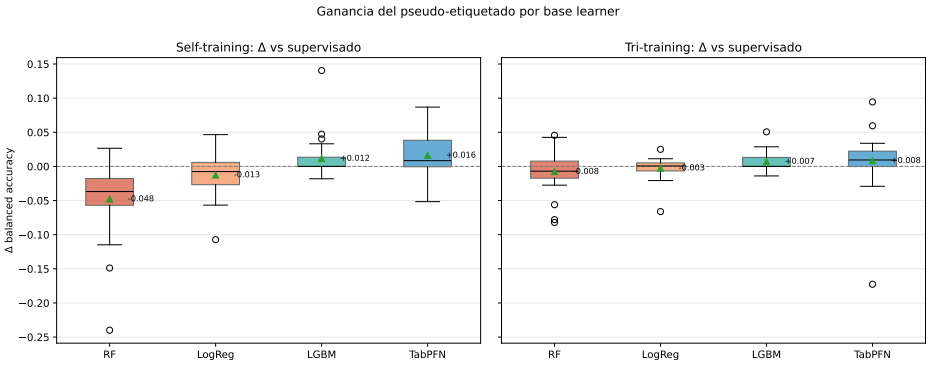


Figura 1. Ganancia Δ en *balanced accuracy* respecto al supervisado, por clasificador base: *self-training* (izquierda) y *tri-training* (derecha).

general del uso de pseudo-etiquetas con cada clasificador base. La Tabla 3 muestra los valores absolutos: *TabPFN* parte ya del mejor rendimiento supervisado (0.767) y alcanza el máximo global con *self-training* (0.783), mientras que el *self-training* degrada a RF de 0.735 a 0.687. El balance de victorias por conjunto refuerza el patrón: el *self-training* supera al supervisado en 23 de 31 conjuntos para *TabPFN*, pero solo en 3 de 31 para RF.

La Figura 1 muestra el efecto del pseudo-etiquetado por clasificador, donde se aprecia el desplazamiento de valores negativos (para el RF) a positivos (para el LGBM y el *TabPFN*). El análisis con el test de Friedman por clasificador, con el *post-hoc* de Nemenyi de la Figura 2, confirma que en RF las diferencias son significativas y el supervisado obtiene el mejor ranking promedio mientras que el *self-training* obtiene el peor; en LGBM y *TabPFN* ese orden se invierte; y en regresión logística el test de Friedman no detecta diferencias significativas ($p = 0.47$).

4.2 Relación con las propiedades del conjunto de datos

Con los resultados obtenidos en nuestros experimentos, la ganancia del *self-training* parece estar relacionada con el número de clases K, y el signo de esa relación coincide con el del efecto que cada clasificador ya exhibe: en *LightGBM*

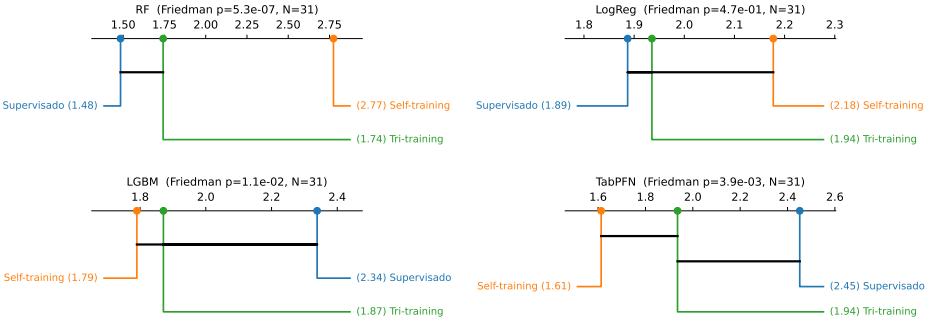


Figura 2. Diagramas de diferencia crítica (*post-hoc* de Nemenyi, $\alpha = 0.05$) de los tres métodos por clasificador base. Las barras horizontales unen métodos sin diferencias significativas; un valor menor indica mejor rango medio.

Tabla 4. Calidad de las pseudo-etiquetas por clasificador base, sobre la primera ronda de pseudo-etiquetado. Macro-promedio sobre los 31 conjuntos y las 8 particiones; los casos sin pseudo-etiquetas aceptadas se excluyen del cálculo de la precisión de las aceptadas.

Clasificador	Precisión aceptadas	Cobertura (% de U)	Confianza media (U)
RF	0.945	46.1 %	0.707
LogReg	0.902	68.2 %	0.813
LGBM	0.883	44.5 %	0.727
TabPFN	0.937	65.8 %	0.816

la ganancia crece con K (Spearman $\rho = +0.43$, $p = 0.015$), mientras que en *Random Forest* el perjuicio se acentúa con K ($\rho = -0.48$, $p = 0.006$). En *TabPFN* la relación es también positiva ($\rho = +0.40$, $p = 0.025$). Es decir, el número de clases se asocia con una amplificación del efecto que cada clasificador presenta de partida: en un régimen de pocas etiquetas, los problemas multiclase disponen de muy pocos ejemplos por clase, de modo que las pseudo-etiquetas tienen mayor peso, para bien o para mal.

4.3 Calidad de las pseudo-etiquetas

Para entender el mecanismo subyacente medimos la fiabilidad de las pseudo-etiquetas que el *self-training* acepta en cada clasificador, comparándolas con las etiquetas reales de las instancias no etiquetadas (conocidas en nuestro protocolo). La Tabla 4 resume, para cada clasificador, tres cantidades sobre la primera ronda de pseudo-etiquetado: la tasa de acierto de las pseudo-etiquetas aceptadas (la fracción de las instancias con confianza superior al umbral de 0.75 que reciben la etiqueta correcta), la cobertura (el porcentaje del conjunto no etiquetado que se acepta), y la confianza media del clasificador (la probabilidad que asigna a la clase más probable, promediada sobre todo el conjunto no etiquetado).

Tabla 5. Error de calibración esperado (ECE) y *Brier score* del clasificador supervisado, por *base learner*. En ambos, un valor menor es mejor; el Brier combina calibración y capacidad discriminativa.

Clasificador	ECE	Brier
RF	0.138	0.295
LogReg	0.083	0.270
LGBM	0.058	0.397
TabPFN	0.075	0.235

La información mostrada en la Tabla 4 es interesante: la calidad de las pseudo-etiquetas no explica el patrón de ganancias. *Random Forest* genera las pseudo-etiquetas *más* precisas (0.945) y es, sin embargo, a quien más perjudica el *self-training*. Más aún, *Random Forest* y *LightGBM* tienen una cobertura casi idéntica (aceptan el 46.1% y el 44.5% del conjunto no etiquetado) con precisiones altas y comparables (0.945 y 0.883), y aun así el *self-training* los afecta de forma opuesta: perjudica al primero y beneficia al segundo. Ni la precisión ni la cobertura de las pseudo-etiquetas, por tanto, separan a los clasificadores que se benefician de los que se perjudican. Esto refuta la hipótesis intuitiva de que mejores pseudo-etiquetas conllevan mayor beneficio.

El efecto parece depender, más que de la calidad de las pseudo-etiquetas generadas, de cómo responde cada clasificador base a la incorporación de datos adicionales con cierto nivel de ruido. Una observación exploratoria en esta dirección es la relación con la calibración, es decir, el grado en que las probabilidades que predice un clasificador se corresponden con las frecuencias reales de acierto [8]. La Tabla 5 reporta el error de calibración esperado (ECE, calculado sobre la clase de máxima probabilidad con 10 intervalos de confianza de anchura uniforme) y, como referencia complementaria, el *Brier score*, que combina calibración y capacidad discriminativa; en ambos, un valor menor es mejor. Los clasificadores mejor calibrados (*LightGBM* y *TabPFN*, con ECE de 0.058 y 0.075) tienden a beneficiarse del pseudo-etiquetado, mientras que el peor calibrado (*Random Forest*, con 0.138) se perjudica; la asociación entre el error de calibración y la ganancia del *self-training* es negativa ($\rho = -0.80$ sobre los cuatro clasificadores). No obstante, con solo cuatro clasificadores esta asociación es meramente indicativa y no permite extraer conclusiones; la formulamos como hipótesis a contrastar con un conjunto mayor de clasificadores base, no como mecanismo establecido.

5 Discusión

Los resultados permiten responder a las tres preguntas de investigación. En cuanto a PI1, el pseudo-etiquetado no es siempre útil en datos tabulares con pocas etiquetas: su utilidad agregada depende por completo del clasificador base. Respecto a PI2, los resultados muestran una dependencia clara del clasificador

base: el pseudo-etiquetado perjudica significativamente a RF ($p < 0.001$), mientras que beneficia significativamente a *TabPFN* ($p < 0.01$). En cuanto a PI3, el análisis de la calidad de las pseudo-etiquetas muestra que esta no explica el patrón: la precisión y la cobertura no separan a los clasificadores que se benefician de los que se perjudican. El efecto parece depender de propiedades del propio clasificador base, entre las que podría estar su calibración.

5.1 El papel de la calibración

Una hipótesis natural es que un clasificador mejor calibrado se beneficie más del pseudo-etiquetado. Nuestros datos son compatibles con esta hipótesis: el error de calibración esperado (ECE) ordena los clasificadores de forma coherente con el efecto observado (*Random Forest*, el peor calibrado y más perjudicado; *LightGBM* y *TabPFN*, mejor calibrados y beneficiados), con una asociación negativa fuerte ($\rho = -0.80$). Sin embargo, esta evidencia se basa en solo cuatro clasificadores (véase además la Sección 5.4 sobre *LightGBM*), por lo que la tratamos como una hipótesis sugerente y no como un mecanismo demostrado. Establecer si la calibración es causal requeriría un estudio con muchos más clasificadores base y, posiblemente, intervenciones (p.ej. recalibrar las probabilidades antes del pseudo-etiquetado y observar si cambia el efecto).

5.2 Casos en los que no ayuda

Incluso para *TabPFN*, el clasificador más beneficiado por el pseudo-etiquetado, el *self-training* perjudica en 8 de los 31 conjuntos (entre ellos `ecoli`, `sonar`, `page-blocks` y `vehicle`), que tienden a ser conjuntos pequeños o con fuerte desbalance. Estos casos muestran que la mejora media debe interpretarse con cautela: el pseudo-etiquetado con un buen clasificador base mejora en promedio, pero no es una garantía en todos los conjuntos.

5.3 Estabilidad

El efecto es robusto globalmente (contrastes de Wilcoxon significativos sobre 31 conjuntos) pero modesto y variable a nivel de conjunto individual. La desviación típica de Δ del *self-training* entre experimentos con un mismo conjunto va de 0.01 (LGBM) a 0.04 (RF) en *balanced accuracy*, comparable en magnitud a las propias ganancias medias. Esto subraya la necesidad de evaluar sobre una batería amplia y de promediar varias repeticiones, en lugar de extraer conclusiones de unos pocos conjuntos.

5.4 Limitaciones

El estudio fija la proporción de etiquetado en el 5 %, un barrido de proporciones permitiría estudiar cuál es el efecto de variar el número de instancias etiquetadas. Otra limitación es que la forma en la que se han tratado los datos ausentes ha

eliminado instancias en cuatro de los conjuntos de datos. Asimismo, en los conjuntos más pequeños el valor por defecto `min_child_samples` = 20 de *LightGBM* impide cualquier partición (basta $|L| < 40$) y su supervisado queda en el nivel de un clasificador constante: ocurre en 13 de los 31 conjuntos (*balanced accuracy* $= 1/K$; de ahí su media anómalamente baja, su ECE bajo y su peor Brier). En 12 de esos 13 el *self-training* tampoco actúa, porque la confianza de un predictor constante (el prior mayoritario) no alcanza el umbral de 0.75; esto explica los empates de la Tabla 3. La mejora de *LightGBM* procede de los 18 conjuntos restantes, donde el modelo es funcional ($\Delta = +0.019$; Wilcoxon $p = 0.006$), no de una recuperación desde el *baseline* degenerado. Finalmente, el análisis de la calidad de las pseudo-etiquetas se ha basado únicamente en la primera ronda de pseudo-etiquetado; el proceso iterativo completo podría mostrar otra tendencia.

6 Conclusiones

Este trabajo ha estudiado empíricamente el efecto del pseudo-etiquetado *plug-and-play* (*self-training*, *tri-training*) en datos tabulares con pocas etiquetas, a través de cuatro clasificadores base de naturaleza distinta. El hallazgo principal es que el efecto del pseudo-etiquetado no es uniforme, sino que depende del clasificador base, con resultados variables que van desde un perjuicio significativo en *Random Forest* hasta un beneficio significativo en *LightGBM* y en el modelo fundacional *TabPFN*. El análisis de la calidad de las pseudo-etiquetas indica que esta no explica el patrón observado: clasificadores con pseudo-etiquetas de precisión y cobertura comparables (como *Random Forest* y *LightGBM*) muestran efectos opuestos, lo que apunta a que el factor determinante reside en propiedades del propio clasificador base. En conjunto, los resultados sugieren que, en datos tabulares con pocas etiquetas, la elección del clasificador base es más determinante que la del método de pseudo-etiquetado.

6.1 Trabajo futuro.

Varias líneas quedan abiertas. En primer lugar, un barrido de la proporción de etiquetado permitiría determinar si la dependencia del efecto respecto al clasificador base se mantiene, se atenúa o se invierte al aumentar la supervisión. En segundo lugar, sería valioso extender el meta-aprendizaje de aplicabilidad del SSL (explorado para ensembles de árboles en otros trabajos [13]) al caso de los modelos fundacionales, estudiando qué propiedades de los conjuntos predicen el beneficio del pseudo-etiquetado. Por último, la incorporación de un segundo modelo fundacional tabular permitiría contrastar si el comportamiento observado en *TabPFN* es general a esta familia de modelos.

Agradecimientos. Este trabajo es parte de los proyectos de I+D+i PID2023-150694OA-I00 y PID2025-171001OB-I00, financiados por MICIU/AEI/10.13039/501100011033 y por "FEDER, UE".

Referencias

1. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Trans Neural Networks Learn Syst **35**(6), 7499–7519 (2022)
2. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
3. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning **109**(2), 373–440 (2020)
4. Garrido-Labrador, J.L., Maudes-Raedo, J.M., Rodríguez, J.J., García-Osorio, C.I.: SSLearn: A semi-supervised learning library for python. SoftwareX **29**, 102024 (2025)
5. Garrido-Labrador, J.L., Serrano-Mamolar, A., Maudes-Raedo, J., Rodríguez, J.J., García-Osorio, C.: Ensemble methods and semi-supervised learning for information fusion: A review and future research directions. Information Fusion **107**, 102310 (2024)
6. Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Manium, M., Hoo, S.B., Bühler, M., Garg, A., et al.: TabPFN-3: Technical report. arXiv preprint arXiv:2605.13986 (2026)
7. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? NeurIPS **35**, 507–520 (2022)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330. PMLR (2017)
9. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. NeurIPS **30** (2017)
11. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. Atlanta (2013)
12. McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., White, C.: When do neural nets outperform boosted trees on tabular data? NeurIPS **36**, 76336–76369 (2023)
13. Ramírez-Sanz, J.M., Martínez-Acha, D., Arnaiz-González, Á., García-Osorio, C., Rodríguez, J.J.: Semi-supervised rotation forest. J Comput Sci **94**, 102777 (2026)
14. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information fusion **81**, 84–90 (2022)
15. Triguero, I., García, S., Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inf Syst **42**(2), 245–284 (2015)
16. Vanschoren, J., Van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: networked science in machine learning. ACM SIGKDD Explorations Newsletter **15**(2), 49–60 (2014)
17. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Trans Knowl Data Eng **35**(9), 8934–8954 (2023)
18. Yoon, J., Zhang, Y., Jordon, J., Van der Schaar, M.: VIME: Extending the success of self-and semi-supervised learning to tabular domain. NeurIPS **33**, 11033–11043 (2020)
19. Zhou, Z.H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng **17**(11), 1529–1541 (2005)