Informe de Evaluación Diagnóstica Multimodelo (v4)

Ciencia de Datos — Fundación 29 de Febrero

Julio 2025

Abstract

Este informe detalla la evaluación de cuatro modelos de diagnóstico diferencial mediante el pipeline v4. Se analiza el rendimiento sobre un corpus de 450 casos clínicos, caracterizado por una codificación heterogénea y una brecha semántica significativa. La metodología se diseñó para maximizar el uso de ontologías (SNOMED CT, ICD-10) y recurrir a la verificación semántica (BERT, LLM) como respaldo. Los resultados confirman el liderazgo del modelo **o3**, que demuestra mayor precisión y eficiencia al capitalizar la estructura de códigos existente, minimizando la dependencia de los costosos juicios de LLM.

Contents

1 Análisis del Dataset de Evaluación

El corpus de evaluación consta de **450 casos clínicos**, cada uno con un diagnóstico de referencia (`GDX`) y cinco diagnósticos diferenciales propuestos (`DDX`). El análisis de su estructura de codificación revela dos hechos fundamentales que condicionan el diseño de la evaluación:

1. **SNOMED CT es la columna vertebral semántica.** Aunque la codificación es multi-ontológica, SNOMED CT es el sistema más prevalente y robusto, cubriendo el 76% de los `GDX` y el 87% de los `DDX` codificados. Su rol es central para establecer coincidencias directas.
2. **Existe una brecha de cobertura del 43.3%.** En 195 de los 450 casos, no existe ningún código compartido (SNOMED, ICD-10 u OMIM) entre el diagnóstico de referencia y las cinco propuestas del modelo. Esta "brecha semántica" hace indispensable contar con métodos de validación que no dependan de ontologías, justificando el uso de modelos como BERT y LLMs para rescatar aciertos semánticamente cercanos pero no codificados.

2 Metodología

El pipeline v4 procesa los 450 casos contrastando los 5 `DDX` de cada modelo con el `GDX` de referencia.

2.1 Matriz de Transiciones & Diseño del *Scoring*

Dada la granularidad variable en la codificación ICD-10 (a menudo a nivel de bloque), se construyó una **matriz de transiciones GDX → DDX** (Fig. 1) para cuantificar la proximidad de los aciertos en la jerarquía. El análisis reveló que más del 92% de las coincidencias útiles se ubicaban a *una* arista de distancia (hijo ↔ padre inmediato). En consecuencia, el pipeline v4 considera como acierto válido **cualquier hijo o padre inmediato** en la jerarquía ICD-10.

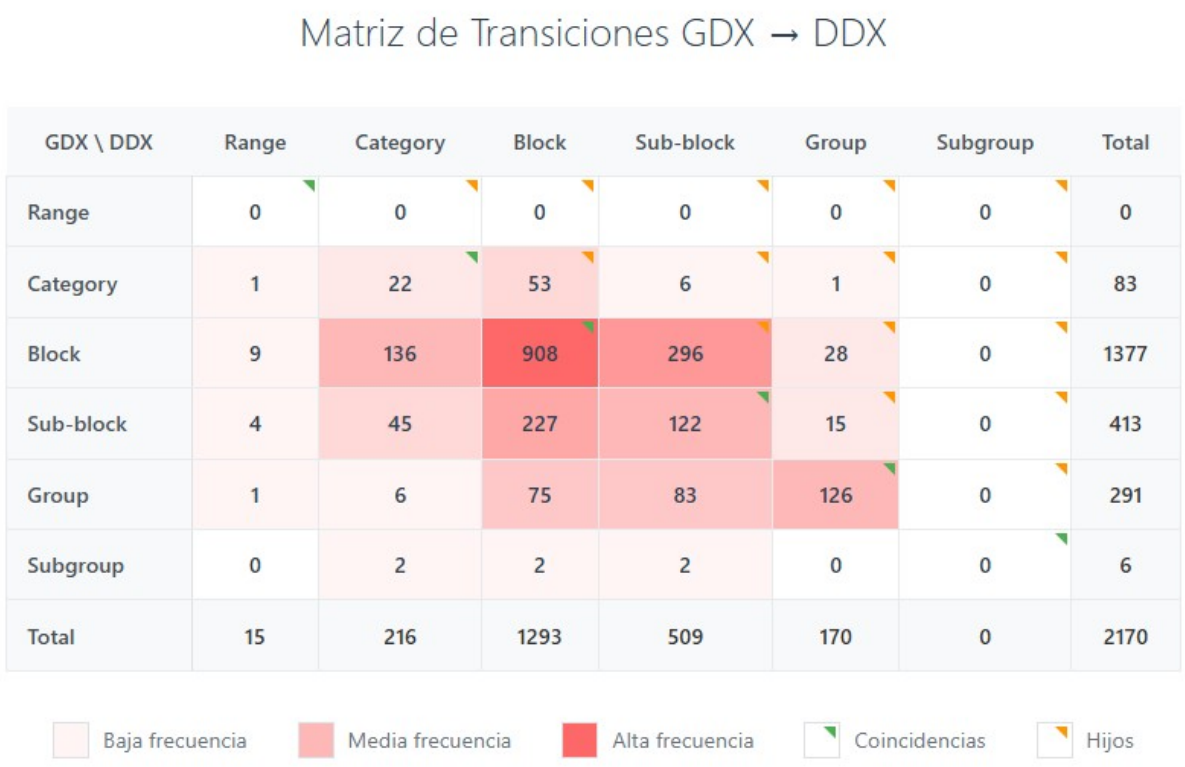

Figure 1: Estudio jerárquico ICD-10 usado para la matriz de transiciones.

Nota: en el paso `ICD10_` los cuatro modos (`EXACT`, `CHILD`, `PARENT`, `SIBLING`) se marcan simplemente como `FIND = true`; no se aplican estratos ni puntuaciones decimales adicionales.

2.2 Jerarquía de Coincidencias

La evaluación es determinista y se detiene en el primer *match* encontrado, siguiendo este orden:

1. **SNOMED_MATCH** — Coincidencia exacta de códigos SNOMED CT.
2. **ICD10_** — Coincidencia jerárquica en ICD-10 (EXACT → CHILD → PARENT → SIBLING).
3. **Verificación Semántica** (solo para casos sin coincidencia de código):
   - BERT: $\cos\theta > 0.925 \Rightarrow$ `BERT_AUTOCONFIRM`.
   - $0.89 < \cos\theta < 0.925 \Rightarrow$ Desempate BERT vs LLM → `BERT_MATCH` / `LLM_JUDGMENT`.

2.3 Flujo de Decisión

3 Resultados

Table 1: Rendimiento comparado (v4).

Modelo	Score %	Pos. Prom.	TOP 1	P5 Fails	LLM Uso
o3	**89.98**	**1.501**	**311**	**9**	**123**
o1	88.19	1.590	305	17	132
o3pro	87.73	1.613	299	24	134
4o	87.70	1.615	299	20	131

4 Conclusiones

La evaluación v4, fundamentada en las características del dataset, permite extraer las siguientes conclusiones:

- **El modelo ‘o3‘ muestra un rendimiento superior y más eficiente.** Lidera en todas las métricas clave (Tabla 1) porque capitaliza mejor la información estructurada disponible. Como se observa en la Fig. 4, la suma de sus aciertos por `SNOMED_MATCH` e `ICD10_*` supera el 80% del total, lo que se traduce en una menor dependencia del LLM (solo 123 usos) y, por tanto, en un menor coste computacional y mayor confianza en los resultados basados en ontologías.
- **La estrategia jerárquica del pipeline es clave para el éxito.** La decisión de aceptar padres/hijos inmediatos en ICD-10 (Sec. 2), justificada por el análisis de transiciones (Fig. 1), junto con la priorización de SNOMED CT, permite resolver la mayoría de los casos. Los métodos semánticos (BERT/LLM) actúan como una red de seguridad indispensable para cubrir la "brecha semántica" del 43.3% del dataset donde no hay códigos coincidentes.
- **Los modelos ‘o3pro‘ y ‘4o‘ ofrecen cobertura a costa de la precisión.** Aunque ‘o3pro‘ tiene un buen desempeño general, su perfil en Fig. 3 muestra una cola más larga de aciertos en posiciones tardías (P4-P5) y un uso más intensivo del LLM. Esto sugiere que, si bien es capaz de encontrar respuestas en casos difíciles, lo hace con menor certeza que ‘o3‘. El modelo ‘4o‘ se comporta como una línea base competente pero sin destacar.

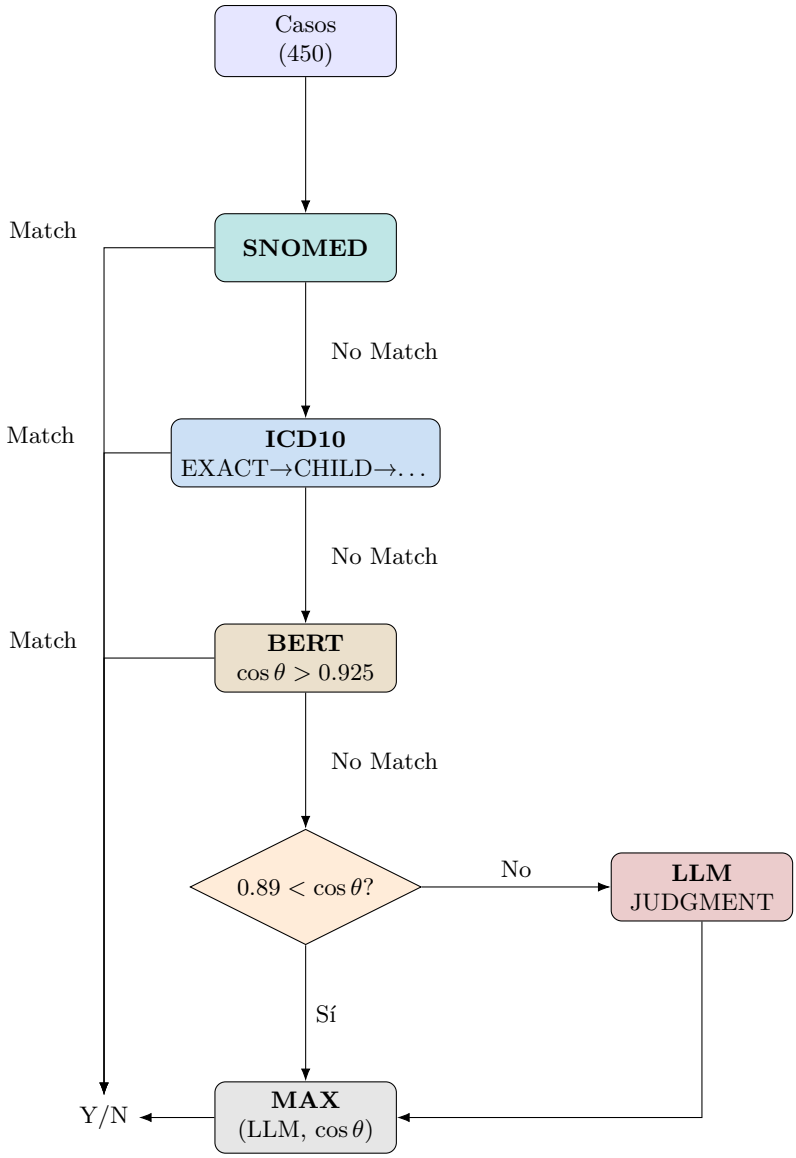

Figure 2: Flujo de decisión del pipeline v4 (con salidas explícitas).

- **El modelo ‘o1‘ presenta un perfil equilibrado.** Se posiciona como una alternativa sólida, con el segundo mejor registro de aciertos en TOP 1 y una distribución homogénea de métodos de acierto, demostrando ser un competidor robusto aunque no supere la eficiencia de ‘o3‘.

5 Anexos Gráficos

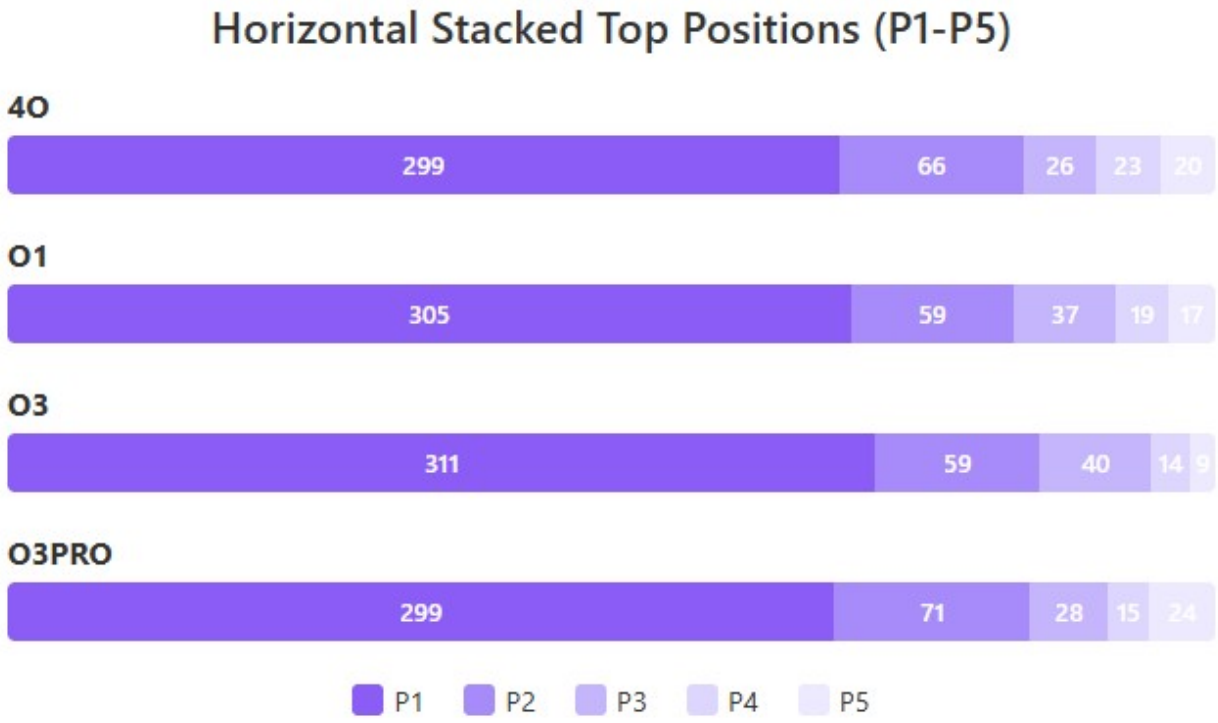

Figure 3: Distribución horizontal apilada de aciertos por posición (TOP 1–5).

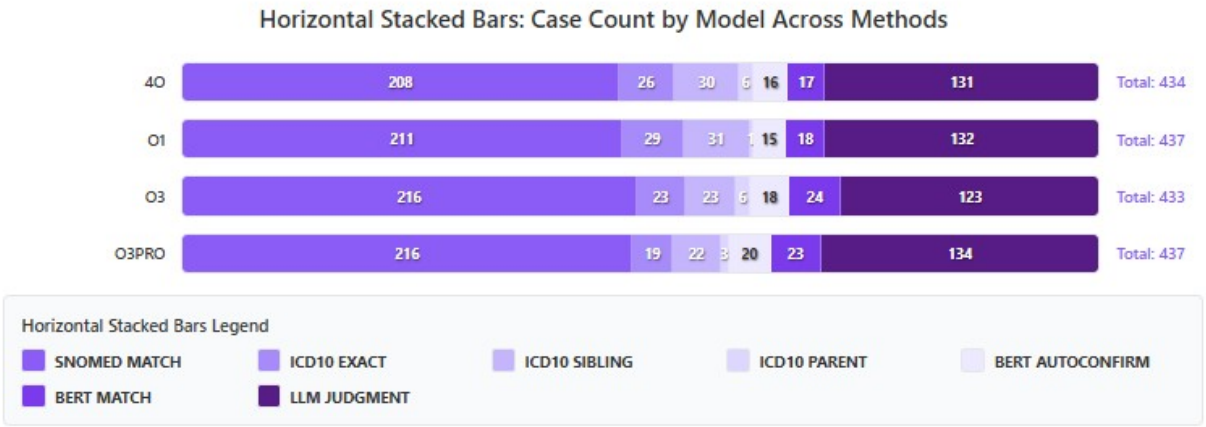

Figure 4: Método de coincidencia dominante por modelo.

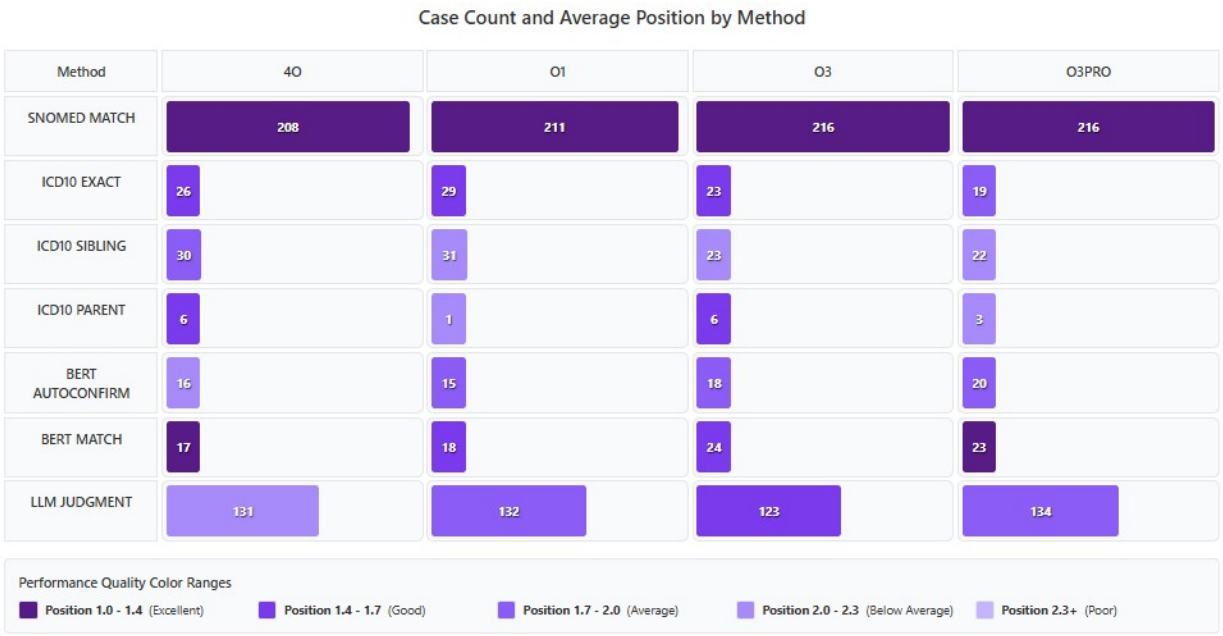

Figure 5: Casos resueltos y posición media por método de coincidencia.