

ZUGZWANG

Tu modelo no debería estar obligado a mover

De una trilogía de experimentos a una librería con garantías

Bruno Marentes

10 de julio de 2026

Prefacio

Este es el cuarto tomo de una serie que empezó con un fracaso instructivo, siguió con una promesa estadística y continuó con una herejía constructiva. *MetaClassifier* demostró con matemáticas por qué un combinador de modelos fuertes no puede crear señal. CAUTO cambió la pregunta y certificó estadísticamente cuándo callar. GAMBIT construyó desde cero un modelo donde predecir y abstenerse se aprenden juntos — y su contrato falsable dictaminó, con honestidad mecánica, que la elegancia no pagaba ventaja.

ZUGZWANG es lo que queda cuando se destila todo eso: no un modelo nuevo, sino un *producto*. Una librería de código abierto que envuelve cualquier clasificador existente y le concede el derecho que las reglas del aprendizaje supervisado le niegan: pasar. En ajedrez, [zugzwang](#) es la obligación de mover cuando cualquier movimiento empeora la posición; un clasificador estándar vive en zugzwang permanente. La librería le da la jugada prohibida — la abstención — con garantías estadísticas de riesgo.

Como sus predecesores, el producto se somete a un contrato falsable pre-registrado (B1–B4), y cada veredicto se publica con el mismo detalle, gane o pierda. Este libro cuenta la génesis (capítulo 1), el diseño (capítulo 2), la maquinaria matemática de cada garantía (capítulo 3), los benchmarks con sus veredictos reales (capítulo 4) y la pieza de producción (capítulo 5). Todo corresponde a código real del repositorio, en inglés y público; el libro, como los documentos rectores, está en español.

Los términos técnicos aparecen [coloreados y enlazados](#) en su primera mención y se definen en el glosario final.

Índice general

1. Génesis: lo que la trilogía probó	1
1.1. Tres proyectos, cinco hallazgos	1
1.2. Por qué un producto y no otro experimento	1
2. El diseño del producto	3
2.1. La forma: librería, fases, contrato	3
2.2. La arquitectura: dos capas y una constante	3
2.3. Errores ruidosos y multiclase de nacimiento	3
2.4. El nombre	4
3. Cada garantía por dentro	5
3.1. Tres nociones de riesgo, una promesa	5
3.2. SGR: la cola binomial exacta, dos veces	5
3.3. RCPS: la misma cola, sobre la tasa incondicional	5
3.4. Conformal PAC: inflar el cuantil con la Beta	5
3.5. El margen estructural: la mitad del presupuesto	6
3.6. Las variantes en esperanza, como opt-in	6
4. Los benchmarks y sus veredictos	7
4.1. El contrato corrido	7
4.2. B1: la historia de las enmiendas	7
4.3. B2: nadie alcanza el punto del certificado	7
4.4. B3 y B4: las derrotas, tal cual	7
5. Producción: la compuerta y su vigía	9
5.1. Tres cosas que no deben separarse	9
5.2. El vigía: medir lo medible	9
5.3. Lo que queda fuera, a propósito	9
Glosario	11

Capítulo 1

Génesis: lo que la trilogía probó

1.1. Tres proyectos, cinco hallazgos

La tesis de producto de ZUGZWANG no es una opinión: cada punto es el veredicto de un contrato falsable corrido en uno de los tres proyectos anteriores, sobre los mismos doce datasets congelados (siete binarios, cinco multiclase, seeds 42–46, checksums verificados).

1. **El motor no se compite: se envuelve.** GAMBIT construyó un gradient boosting propio y, tras semanas de optimización, quedó a 0.3 puntos porcentuales del LightGBM de estante. No hay premio en reinventar el motor; el valor está en lo que se pone encima.
2. **La señal de selección es la confianza del propio modelo.** Nada aprendido encima paga: lo dijo H1 de CAUTO (scores de ensamble y desacuerdo no superan a la confianza calibrada), lo repitieron C3, C3-MC y C5 de GAMBIT (la abstención aprendida no compra cobertura certificada), y lo confirmó E-MLP (el resultado es robusto a la familia de modelo).
3. **Para rankear no se paga impuesto de calibración.** La [calibración](#) vuelve legibles las probabilidades, pero no cambia el orden de los ejemplos — y la selección es un problema de orden. C4 de GAMBIT mostró el reembolso: entrenar con todo el presupuesto y certificar sobre confianza cruda supera a entrenar con menos para calibrar en el sobrante.
4. **Las garantías son lo único con respaldo formal.** El umbral elegido “a ojo” viola el riesgo objetivo en más de la mitad de las corridas (CAUTO, tabla de validez); el certificado [SGR](#) lo viola en menos de δ . Esa diferencia — promesa matemática contra costumbre — es el producto entero.
5. **El punto de operación se elige, no emerge.** C2 de GAMBIT: la abstención nativa de un modelo entrenado para abstenerse no encuentra sola su punto de operación útil; la curva [riesgo–cobertura](#) sí lo da, y elegir sobre ella es un acto explícito del dueño del sistema.

Lo que hay que recordar

- Cada punto de la tesis viene de un contrato falsable ya corrido, no de intuición.
- El fracaso diagnosticado (MetaClassifier, C3 de GAMBIT) vale tanto como el éxito: delimita dónde NO buscar.
- ZUGZWANG no inventa ciencia nueva: convierte veredictos en ergonomía.

1.2. Por qué un producto y no otro experimento

Los tres proyectos anteriores terminaron en reportes y libros; el conocimiento quedó demostrado pero no *usable*. La pieza que falta no es otra hipótesis sino empaque: que cualquier

persona con un modelo entrenado pueda escribir tres líneas y obtener un umbral certificado, con los errores ruidosos donde debe haberlos y la teoría citada donde debe citarse. El estado del arte externo (MAPIE, crepes) resuelve conjuntos conformes con APIs maduras; ninguno entrega la tarea completa del usuario final de clasificación selectiva: *riesgo objetivo* $\rightarrow$ *umbral certificado* $\rightarrow$ *predicción o abstención*. Ese hueco es el nicho.

Capítulo 2

El diseño del producto

2.1. La forma: librería, fases, contrato

ZUGZWANG es una librería Python de código abierto (PyPI, licencia MIT, todo lo público en inglés) construida por fases, cada una un release usable: F1 entrega el núcleo certificado y el contrato corrido (v0.1); F2 añade calibración opt-in y reporte HTML (v0.2); F3 cierra con la pieza de producción (v1.0). El contrato falsable B1–B4 quedó congelado antes de escribir la primera línea:

- **B1 — validez** (bloqueante, sin zona gris): los certificados de los tres métodos deben respetar su promesa en test; si se viola sistemáticamente, el producto está roto y no se publica.
- **B2 — piso**: contra el pipeline artesanal (umbral empírico, calibrar-y-confiar). Si el artesanal respeta el riesgo igual de bien, el certificado no aporta y el pitch se re-piensa.
- **B3 — techo**: contra MAPIE (y crepes) en la misma tarea de usuario final, con cobertura certificada, velocidad y ergonomía medidas.
- **B4 — el reembolso**: replicar C4 bajo la implementación nueva; si no replica, el claim se retira del marketing.

2.2. La arquitectura: dos capas y una constante

La capa 1 (`zugzwang.core`) son funciones puras sobre arrays — numpy y scipy, nada más —: scores de confianza, umbrales sin garantía, los tres certificadores, curvas riesgo–cobertura y validación ruidosa de insumos. La capa 2 (`SelectiveClassifier`) es un meta-estimador sklearn que envuelve cualquier cosa con `predict_proba`. La constante es `ABSTAIN = -1`, heredada de GAMBIT.

El detalle de diseño más importante no es visible en el diagrama: `certify()` recibe el split de certificación *fuera* de `fit()`. Eso preserva la firma sklearn estándar (el estimador pasa `check_estimator`), hace natural el caso `prefit=True` (el modelo de producción ya está entrenado y no se re-entrena para ganar abstención) y codifica el protocolo de datos honesto: entrenar con TODO el presupuesto; el único costo real es el split `select`, pequeño y etiquetado, que se usa una sola vez para certificar.

2.3. Errores ruidosos y multiclase de nacimiento

Dos lecciones de la trilogía están cableadas como comportamiento, no como documentación. Primera: nunca umbrales silenciosos — un riesgo incertificable lanza `UncertifiableRiskError` informando el mínimo certificable con esos datos ($1 - \delta^{1/n}$ con cero errores), y un split sospechoso de contaminación (filas de `select` duplicadas del train) lanza excepción antes de emitir un

certificado sin valor. Segunda: el binario es $K=2$ por el mismo código — no hay rama multiclase, y hay tests que verifican la no-cirugía (lección C2).

2.4. El nombre

Stalemate estaba tomado en PyPI; *castling* libre pero enterrado en SEO ajedrecístico. *Zugzwang* es impronunciable y distintivo — como *xgboost*, como *seaborn* — y dice exactamente lo que el producto arregla: la obligación de mover cuando toda jugada empeora la posición.

Lo que hay que recordar

- Cada fase es un release usable; el contrato B1–B4 se congeló antes de la primera corrida.
- `certify()` fuera de `fit()`: firma sklearn intacta, `prefit` natural, protocolo de datos honesto.
- Los errores ruidosos y la K-aridad sin cirugía son comportamiento testado, no promesas de documentación.

Capítulo 3

Cada garantía por dentro

3.1. Tres nociones de riesgo, una promesa

El producto certifica tres cantidades distintas, y la honestidad exige decir cuál es cuál. SGR controla el *riesgo selectivo condicional*: $P(\text{error} \mid \text{aceptado})$. RCPS controla la *tasa incondicional de malas decisiones*: $P(\text{error} \wedge \text{aceptado})$ — “el modelo contestó y se equivocó”, sobre todos los insumos. El conformal selectivo controla la *cobertura de conjuntos*: $P(y \in C(x))$. Lo que las tres comparten en ZUGZWANG es el *sabor* de la promesa: con confianza $\geq 1 - \delta$ sobre el muestreo del split de certificación, la cantidad certificada queda bajo control. Ese sabor común no fue gratis: es la consecuencia del gate B1 (capítulo 4).

3.2. SGR: la cola binomial exacta, dos veces

Ordenados los n puntos del select por confianza descendente, cada prefijo de tamaño m con k errores observados da una cota superior exacta del riesgo condicional verdadero: el límite de Clopper–Pearson unilateral, $\text{Beta}^{-1}(1 - \delta; k + 1, m - k)$. SGR devuelve el prefijo MÁS GRANDE cuya cota queda bajo r^* . Probar muchos prefijos cuesta multiplicidad; el híbrido de CAUTO reparte δ en dos mitades — búsqueda binaria con Bonferroni sobre sus $\log_2 n$ pasos (robusta) y secuencia fija a nivel completo por paso (ajustada) — y se queda con la de mayor cobertura: por union bound, el error familiar total sigue $\leq \delta$.

3.3. RCPS: la misma cola, sobre la tasa incondicional

La pérdida $\ell_i(\theta) = e_i \cdot \mathbf{1}\{s_i \geq \theta\}$ es monótona no creciente en θ : aceptar menos nunca sube la tasa de malas decisiones. El teorema 1 de Bates et al. (2021) dice que, con pérdidas monótonas, basta una cota superior de confianza puntual — Clopper–Pearson otra vez, ahora sobre los n totales — y el barrido del umbral NO paga multiplicidad. El resultado: $P(\text{tasa} \leq r^*) \geq 1 - \delta$.

3.4. Conformal PAC: inflar el cuantil con la Beta

El conformal split clásico toma el cuantil $\lceil (n+1)(1-\alpha) \rceil$ de las no-conformidades $1 - \hat{P}(y|x)$: su garantía es MARGINAL (en promedio sobre los muestreos del select; tu split particular puede sub-cubrir). Para la promesa condicional al entrenamiento se usa que, con scores continuos, la cobertura condicional del k -ésimo estadístico de orden se distribuye $\text{Beta}(k, n+1-k)$: se elige el menor k con $F_{\text{Beta}(k, n+1-k)}(1 - \alpha) \leq \delta$ (Vovk 2012; Park et al. 2020). Si ningún k lo logra, la certificación se niega en voz alta — con pocos datos, prometer es mentir.

3.5. El margen estructural: la mitad del presupuesto

El gate B1 enseñó una lección sutil. Un método válido que gasta TODO su δ deja la cantidad certificada pegada a r^* ; el ruido de muestreo del lado del test empuja lo observado por encima de r^* en una fracción de corridas mayor que δ , aunque el certificado nunca mintió. SGR pasó el gate holgado precisamente porque su híbrido gasta $\delta/2$ por rama. La respuesta de producto: las tres garantías gastan a lo sumo LA MITAD del presupuesto declarado — en SGR por necesidad matemática, en RCPS y conformal PAC por diseño. La confianza anunciada es un piso, nunca una estimación. Ser más conservador que lo anunciado es siempre válido; lo contrario, jamás.

3.6. Las variantes en esperanza, como opt-in

El CRC clásico (Angelopoulos et al. 2022) y el conformal marginal compran más cobertura con una promesa más débil: control EN MEDIA sobre el muestreo, con excedencias puntuales en cerca de la mitad de las corridas — eso significa “en esperanza”, y B1 lo midió: 33 % y 42 % de violaciones puntuales con medias bajo el objetivo. Siguen en el producto vía `delta=None`, con su contrato citado textualmente en el certificado, porque hay usos legítimos (pipelines de gran volumen donde importa el agregado); pero jamás son el default.

Lo que hay que recordar

- Tres cantidades certificadas distintas; UNA promesa: con confianza $\geq 1 - \delta$, control.
- Toda la maquinaria dura es la cola binomial exacta usada tres veces (prefijos condicionales; tasa incondicional; estadísticos de orden).
- Margen estructural $\delta/2$: la confianza declarada es piso.
- En esperanza $\neq$ con alta probabilidad; el certificado lo dice textualmente y B1 midió la diferencia.

Capítulo 4

Los benchmarks y sus veredictos

4.1. El contrato corrido

Quince datasets (los doce congelados de la trilogía más tres externos de OpenML-CC18, elegidos y congelados ANTES de cualquier corrida), cinco seeds, cuatro riesgos objetivo, Wilcoxon pareado con $p < 0,05$ pre-declarado. El test congelado se usó UNA vez. Los cuatro veredictos, con sus derrotas incluidas:

Criterio	Veredicto	Número clave
B1 validez (bloqueante)	PASA	sgr 9/217, crc 15/294, conf. 21/295
B2 piso vs artesanal	GRIS	no-dominado; +13,7 pp vs quien respeta
B3 techo vs MAPIE	REFUTADO	mediana de cobertura favorece a MAPIE
B4 el reembolso	REFUTADO	claim retirado; impuesto ≈ 0

4.2. B1: la historia de las enmiendas

La primera corrida usó las variantes en esperanza para crc y conformal: 99 y 127 violaciones puntuales de 300 — mientras sus PROPIOS contratos se cumplían (medias observadas bajo el objetivo). El criterio congelado es nativo del sabor de alta probabilidad, y la decisión del dueño del proyecto fue endurecer los métodos, no el criterio: RCPS y conformal PAC bajaron las violaciones a 22 y 26 — al filo del permitido (21). El último paso fue el margen estructural $\delta/2$ del capítulo 3: 15 y 21. Los tres métodos pasan y el criterio quedó intacto. Toda la cadena está publicada en el reporte: la honestidad del proceso vale tanto como el resultado.

4.3. B2: nadie alcanza el punto del certificado

El brazo artesanal directo (umbral empírico al riesgo objetivo) iguala la cobertura de ZUGZWANG... violando el riesgo el 38 % de las veces. El brazo calibrar-y-confiar respeta el riesgo... pagando -14 a -16 pp de cobertura. Ninguna receta artesanal es al mismo tiempo segura y cubridora: ZUGZWANG es no-dominado en el plano riesgo-cobertura. El veredicto mecánico quedó GRIS porque las dos condiciones congeladas no se cumplen contra UN MISMO brazo — y porque nuestra regla pre-declarada tenía un error de categoría (comparaba cobertura contra el brazo que hace trampa), que se publica junto con las dos lecturas.

4.4. B3 y B4: las derrotas, tal cual

MAPIE certifica más cobertura en la mediana de pares (aunque pierde en media y en dos datasets con significancia): REFUTADO, y la documentación dice cuándo usar MAPIE — con-

juntos con cobertura marginal — y cuándo ZUGZWANG — el contrato selectivo completo. B4 no replicó el “reembolso” de la trilogía; el diagnóstico exploratorio (mismos splits, mismo motor, solo cambia el score certificado) mostró que el impuesto de calibración sobre cobertura certificada es ≈ 0 y que la variable que importa es el TAMAÑO del split de certificación. El claim se retiró del marketing; la guía práctica quedó en la documentación: no mates tu select por alimentar al motor.

Lo que hay que recordar

- B1 pasó con el criterio congelado intacto; el costo fue endurecer los métodos (RCPS, PAC, margen $\delta/2$).
- B2: el certificado compra un punto riesgo–cobertura que ninguna receta artesanal alcanza.
- Las refutaciones de B3 y B4 se publican con el mismo detalle que los éxitos — y B4 dejó ciencia útil: importa el tamaño del select, no evitar la calibración.

Capítulo 5

Producción: la compuerta y su vigía

5.1. Tres cosas que no deben separarse

Un clasificador certificado son tres piezas que en producción tienden a divorciarse: el modelo (que alguien re-entrena), el umbral (que alguien “ajusta tantito”) y el certificado (que alguien olvida). El `DecisionGate` las empaqueta como UNA unidad serializable con sello de versión de formato: cargar una compuerta es cargar también su promesa, textual, en `gate.certificate.guarantee`. La lección viene de la trilogía entera: el punto de operación se elige y se documenta; una compuerta sin su certificado es un umbral supersticioso.

5.2. El vigía: medir lo medible

El certificado se emitió bajo intercambiabilidad con el split de certificación. Nadie puede prometer detectar el drift distribucional — la extensión E-drift de GAMBIT mostró lo epistémicamente resbaloso que es — pero hay dos cantidades OBSERVABLES cuya desviación es evidencia dura de que la premisa murió: la tasa de aceptación (contra la cobertura certificada, test binomial exacto bilateral) y la tasa de error entre los aceptados ya etiquetados (contra r^* , test unilateral de excedencia). El `GateMonitor` acumula, testea y ALARMA; actuar — re-certificar con select fresco, pausar la compuerta — es del dueño del sistema. Sin claims epistémicos: el monitor no ve el mundo, ve la compuerta.

5.3. Lo que queda fuera, a propósito

v1.0 cierra el alcance declarado en el diseño: sin regresión selectiva, sin scores aprendidos, sin detección de drift, sin entrenar motores. Cada exclusión tiene detrás un veredicto de la trilogía o del contrato B1–B4. El producto es pequeño porque lo que afirma está respaldado; crecer sin contrato sería volver al zugzwang del que veníamos huyendo.

Lo que hay que recordar

- `DecisionGate`: modelo + umbral + certificado viajan juntos o no viajan.
- El monitor vigila cobertura y riesgo observados — lo medible — con tests binomiales exactos; informar no es actuar.
- El anti-alcance es parte del producto: cada “no” tiene un veredicto detrás.

Glosario

Zugzwang En ajedrez, posición donde el jugador está obligado a mover y cualquier movimiento empeora su posición. Metáfora del clasificador estándar: obligado a predecir aun cuando toda predicción es mala.

Calibración Ajuste post-hoc de probabilidades para que “70 %” signifique 70 % de aciertos empíricos. Vuelve legibles las probabilidades; no cambia el ranking (y por eso no mejora la selección).

SGR (Selective Guaranteed Risk) Certificador de Geifman y El-Yaniv (2017): umbral de máxima cobertura cuyo riesgo selectivo, acotado con la cola binomial exacta, queda bajo r^* con confianza $1 - \delta$.

Curva riesgo–cobertura Riesgo selectivo (error entre los aceptados) como función de la cobertura (fracción aceptada), al ordenar por confianza descendente. La herramienta para ELE-GIR el punto de operación.