Validar científicamente un modelo de IA para Calidad, Vida Útil y Seguridad Alimentaria (Vol.2)

Validación científica, industrial y jurídica de modelos de IA aplicados a calidad, vida útil y seguridad alimentaria

José Manuel López · Dirección Técnica
INTABIOTECH SL

A todos los que dedican su vida al estudio, el avance, la perfección y la mejora de la vida de los demás, en cualquiera de los campos que estudien o dominen.

El Autor

Cómo validar científicamente un modelo de inteligencia artificial para calidad, vida útil y seguridad alimentaria

Volumen 2

De la precisión aparente a la evidencia de aptitud para el uso: validación externa, errores de clasificación, deriva, incertidumbre y criterios de aceptación (Capítulos 19-32)

José M. López, BSc, MD, PhD, LLD (Cand.)

Manuscrito técnico-científico · Edición editorial · Julio de 2026

Servicio de Publicaciones de
INTABIOTECH

Índice general Vol. 2

19. Errores metodológicos que invalidan resultados aparentemente excelentes

Muchos modelos de inteligencia artificial aplicados a calidad, vida útil o seguridad alimentaria no fracasan porque el algoritmo sea técnicamente deficiente, sino porque el diseño experimental produce una estimación artificialmente optimista de su rendimiento.

El problema es especialmente grave cuando una métrica elevada se interpreta como prueba de aptitud industrial sin analizar cómo se obtuvieron los datos, qué unidades fueron realmente independientes, qué información estuvo disponible durante el entrenamiento y cuántas decisiones metodológicas se adoptaron después de consultar el conjunto de prueba.

Un modelo puede alcanzar:

$$\text{accuracy}=98\%$$

$$\text{AUC}=0{,}97$$

o:

$$R^2=0{,}95$$

y, sin embargo, carecer de capacidad real para generalizar a un lote, campaña, proveedor, planta o formulación nuevos.

Las métricas no son propiedades intrínsecas del algoritmo. Son resultados de una evaluación realizada bajo un diseño concreto. Cuando ese diseño contiene fuga, dependencia, selección retrospectiva o falta de representatividad, el número obtenido deja de medir aquello que se pretende demostrar.

19.1. Fuga de datos entre entrenamiento y prueba

Existe fuga de datos cuando información que no debería estar disponible durante el entrenamiento influye directa o indirectamente en el modelo.

Puede expresarse como:

$$I_{\text{prueba}} \rightarrow \text{desarrollo}$$

cuando la información del conjunto de prueba afecta a:

- selección de variables;
- preprocesamiento;
- ajuste de hiperparámetros;
- elección de arquitectura;
- definición de umbral;
- exclusión de observaciones;
- selección de métricas;
- decisión sobre la versión final.

La fuga puede ser evidente o extremadamente sutil.

Ejemplo directo

Una misma muestra aparece en entrenamiento y prueba.

Ejemplo indirecto

Una muestra genera varios espectros. Algunos se utilizan para entrenar y otros para probar.

Aunque los archivos sean diferentes, comparten:

- composición;
- lote;
- preparación;
- instrumento;
- condiciones;
- ruido específico.

El modelo puede reconocer la identidad de la muestra en lugar de aprender una relación generalizable.

19.2. Fuga por réplicas técnicas

Las réplicas técnicas cuantifican la variabilidad del procedimiento de medida. No representan nuevas unidades biológicas o industriales.

Supóngase que se analizan 100 lotes y se obtienen 20 espectros por lote:

$$100 \times 20 = 2.000 \text{ espectros}$$

Si los espectros se dividen aleatoriamente, el conjunto de entrenamiento y el de prueba contendrán mediciones de los mismos lotes.

El tamaño aparente será:

$$n = 2.000$$

pero el número de unidades independientes continuará siendo aproximadamente:

$$n_{independiente} = 100$$

El algoritmo puede aprovechar rasgos específicos de cada lote y presentar una capacidad predictiva irreal.

La partición correcta debe realizarse antes de generar o distribuir las réplicas:

$$\text{lote} \rightarrow \text{entrenamiento validación prueba}$$

Todas las mediciones del mismo lote deben permanecer en el mismo conjunto.

19.3. Fuga por aumentación de datos

La aumentación genera versiones modificadas de una observación:

- rotación;
- recorte;
- inversión;
- cambio de brillo;
- adición de ruido;
- desplazamiento;
- interpolación.

Es útil para entrenar modelos visuales, pero puede causar fuga si las versiones aumentadas se generan antes de dividir los datos.

Una imagen original podría quedar en entrenamiento y una versión rotada en prueba.

El sistema no está generalizando a una unidad desconocida. Está reconociendo una transformación de una observación conocida.

El procedimiento correcto es:

1. dividir las unidades originales;

2. aplicar aumentación únicamente dentro del conjunto de entrenamiento;

3. mantener la prueba libre de derivados del entrenamiento.

La aumentación tampoco aumenta el número de unidades independientes. Puede mejorar robustez, pero no sustituye la recopilación de nuevos lotes, productos o condiciones.

19.4. Fuga por preprocesamiento global

Numerosos procedimientos aprenden parámetros a partir de los datos:

- normalización;
- estandarización;
- imputación;
- selección de variables;
- reducción dimensional;
- corrección de lotes;
- eliminación de valores atípicos.

Si estos parámetros se calculan utilizando toda la base antes de dividirla, el conjunto de prueba influye en el desarrollo.

Por ejemplo, una estandarización utiliza:

$$z= [x-\mu] / [\sigma]$$

Si (μ) y (σ) se calculan con entrenamiento y prueba conjuntamente, el modelo recibe información sobre la distribución futura.

La secuencia correcta es:

ajustar transformación en entrenamiento

aplicar transformación congelada a validación y prueba

En validación cruzada, el preprocesamiento debe repetirse dentro de cada pliegue, no una sola vez sobre la base completa.

19.5. Fuga por selección de variables

Si se seleccionan variables utilizando su asociación con el resultado en toda la base y después se realiza validación cruzada, la evaluación será optimista.

El proceso correcto debe quedar incluido dentro de cada partición:

1. seleccionar variables en el entrenamiento del pliegue;

2. ajustar el modelo;

3. evaluar en la parte no utilizada.

Esto es particularmente importante en:

- metabolómica;
- espectroscopia;
- datos ómicos;
- sensores de alta dimensión;
- imágenes hiperespectrales.

Cuando existen miles de variables y pocas muestras, algunas correlaciones aparentes aparecerán por azar. Si la selección ve la totalidad de los datos, el modelo explotará esas asociaciones fortuitas.

19.6. Fuga temporal

La fuga temporal aparece cuando se utiliza información futura para predecir una decisión pasada.

Ejemplos:

- emplear resultados analíticos obtenidos después de la liberación;
- incorporar temperatura registrada tras el momento de predicción;
- utilizar clasificación final de calidad;
- introducir reclamaciones posteriores;
- calcular estadísticas con periodos futuros;
- entrenar con campañas posteriores y probar con anteriores.

En una aplicación prospectiva debe cumplirse:

$$t_{entrada} \leq t_{decisión}$$

Ninguna variable disponible después de la decisión puede utilizarse como predictor, salvo que el objetivo real sea retrospectivo.

Las bases históricas suelen contener columnas que resumen el resultado del proceso y que no existirán en tiempo real. Su inclusión puede producir un rendimiento extraordinario y completamente inoperante.

19.7. Fuga a través de identificadores

Los modelos pueden aprender a partir de variables aparentemente inocuas:

- código de producto;
- número de lote;
- fecha;
- identificador de proveedor;
- código de laboratorio;
- nombre de fichero;
- ruta de almacenamiento;
- operador;
- cámara;
- planta.

Estas variables pueden actuar como sustitutos de la etiqueta.

Ejemplo: si los positivos proceden históricamente de una planta específica, el modelo puede aprender el código de planta en lugar de las características del alimento.

Este aprendizaje puede ser útil para describir el pasado, pero será frágil ante cambios de proceso y puede ocultar la verdadera causa.

Deben analizarse variables proxy y realizar validaciones en las que se reserven grupos completos.

19.8. Pseudorreplicación

La pseudorreplicación consiste en tratar observaciones dependientes como si fueran independientes.

Puede aparecer cuando se consideran unidades separadas:

- mediciones del mismo lote;
- unidades del mismo envase;
- imágenes del mismo producto;

• tiempos de la misma curva;

• muestras de una misma producción;

• superficies próximas de una misma zona;

• múltiples microorganismos aislados del mismo evento.

Esto produce:

• intervalos de confianza demasiado estrechos;

• significación exagerada;

• estimaciones optimistas;

• tamaño muestral ficticio.

Si existen (m) observaciones por conglomerado y correlación intraclase (ρ), el efecto de diseño puede aproximarse a:

$$DE=1+(m-1)\rho$$

El tamaño efectivo sería:

$$n\_efectivo \approx [n] / [DE]$$

Cuando la correlación es alta, añadir muchas réplicas del mismo lote aporta menos información que añadir nuevos lotes.

19.9. Validación cruzada aleatoria inadecuada

La validación cruzada convencional divide aleatoriamente las observaciones en varios pliegues.

Puede ser correcta si las observaciones son independientes y proceden de una población homogénea. Sin embargo, en alimentación suele existir estructura:

• lotes;

• plantas;

• campañas;

• líneas;

• proveedores;

• operadores;

• instrumentos;

• formulaciones.

Si estas estructuras se mezclan entre pliegues, la validación evalúa interpolación dentro de contextos conocidos, no generalización a contextos nuevos.

La alternativa es la validación cruzada agrupada:

$$G\_i \cap G\_j = \emptyset$$

para los grupos asignados a pliegues distintos.

Ejemplos:

• GroupKFold por lote;

• leave-one-provider-out;

• leave-one-plant-out;

• leave-one-campaign-out;

• validación temporal progresiva.

La estrategia debe reflejar el uso previsto.

19.10. Validación leave-one-out engañosa

La validación leave-one-out utiliza todas las observaciones salvo una para entrenar y evalúa sobre la restante.

Puede parecer rigurosa porque repite el proceso muchas veces, pero presenta limitaciones:

- conjuntos de entrenamiento casi idénticos;
- alta varianza de la estimación;
- elevado coste;
- fuga si existen réplicas correlacionadas;
- poca representación de cambios de dominio.

Si se deja fuera un espectro mientras otros espectros del mismo lote permanecen en entrenamiento, la independencia es ficticia.

En datos agrupados, debería dejarse fuera el lote, proveedor o centro completo.

19.11. Validación cruzada anidada ausente

Cuando se utilizan los mismos pliegues para seleccionar hiperparámetros y estimar el rendimiento, la métrica será optimista.

La validación cruzada anidada separa dos niveles:

Bucle interno

Selecciona:

- arquitectura;
- hiperparámetros;
- variables;
- umbral.

Bucle externo

Estima el rendimiento del procedimiento completo.

Formalmente:

$$\text{evaluación externa}\ [\ \text{selección interna}\ ]$$

Sin esta separación, el rendimiento refleja parcialmente el ajuste a los pliegues de evaluación.

La validación anidada es especialmente importante cuando se prueban muchas configuraciones.

19.12. Sobreajuste al conjunto de prueba

El conjunto de prueba puede contaminarse sin utilizarlo directamente para entrenar.

Cada vez que el equipo consulta su resultado y modifica el modelo, incorpora información del test.

Supóngase que se prueban 100 configuraciones y se elige la que mejor funciona en el mismo conjunto de prueba. Aunque ninguna haya ajustado sus pesos con esos datos, la configuración final ha sido seleccionada para ese conjunto.

El test se convierte en un conjunto de validación.

La solución es:

- bloquear el test;
- limitar el número de evaluaciones;
- mantener un conjunto final verdaderamente independiente;
- registrar todas las consultas;

- utilizar validación anidada;
- obtener nuevos datos para la confirmación definitiva.

19.13. Multiplicidad de modelos

Cuando se entrenan numerosos modelos, alguno puede obtener un rendimiento elevado por azar.

Si se prueban:

- diez algoritmos;
- veinte combinaciones de variables;
- diez umbrales;
- cinco preprocesamientos;

el espacio de búsqueda contiene:

$$10 \times 20 \times 10 \times 5 = 10.000$$

configuraciones.

Seleccionar la mejor sin una evaluación independiente produce un fuerte sesgo de selección.

Debe documentarse:

- número de modelos probados;
- espacio de hiperparámetros;
- criterio de selección;
- rendimiento de candidatos;
- diferencia entre desarrollo y prueba final.

No debe presentarse la versión seleccionada como si hubiera sido la única hipótesis evaluada.

19.14. Selección retrospectiva del umbral

Un modelo puede generar una probabilidad continua. El rendimiento depende del umbral.

Si se selecciona el umbral que maximiza la sensibilidad y especificidad sobre el conjunto de prueba, las métricas serán optimistas.

El umbral debe ajustarse en:

- entrenamiento;
- conjunto de calibración;
- validación interna.

Después debe congelarse y evaluarse sin cambios en el conjunto externo.

Una práctica incorrecta es comunicar:

«El modelo alcanzó una sensibilidad del 98 % con el umbral óptimo.»

sin indicar que ese umbral fue elegido sobre los mismos casos utilizados para calcular el 98 %.

19.15. Optimización mediante el índice de Youden

El índice de Youden se define como:

$$J = \text{Sensibilidad} + \text{Especificidad} - 1$$

Puede utilizarse para identificar un punto de equilibrio en la curva ROC.

Sin embargo, trata sensibilidad y especificidad de manera simétrica y no considera:

- gravedad diferencial de errores;
- prevalencia;
- capacidad de confirmación;
- volumen de producción;
- costes;
- límites mínimos de seguridad.

En seguridad alimentaria, el umbral que maximiza (J) puede generar una sensibilidad insuficiente.

El criterio correcto podría ser:

$$\text{Especificidad}$$

sujeto a:

$$\text{Sensibilidad} \geq S\_$$

19.16. Selección de métricas después de observar los resultados

Otro sesgo aparece cuando el equipo comunica únicamente las métricas favorables.

Ejemplos:

- presenta accuracy y omite sensibilidad;
- presenta AUC y omite el umbral;
- comunica F1 y no informa falsos negativos;
- muestra (R^2) y no MAE;
- informa el error medio y oculta la cola;
- comunica rendimiento global y omite un producto débil.

La solución es preespecificar:

- métrica principal;
- métricas secundarias;
- subgrupos;
- intervalos;
- análisis de errores.

El informe debe incluir resultados favorables y desfavorables.

19.17. Desequilibrio de clases ignorado

Cuando la clase positiva es rara, la accuracy puede estar dominada por los negativos.

Ejemplo:

$$\text{Prevalencia} = 0{,}5\%$$

Un modelo que clasifica todo como negativo obtiene:

$$\text{Accuracy} = 99{,}5\%$$

pero:

$$\text{Sensibilidad} = 0$$

La evaluación debe incluir:

- matriz de confusión;
- sensibilidad;

- especificidad;
- valores predictivos;
- precision–recall;
- número absoluto de errores.

También debe analizarse la prevalencia del conjunto frente a la real.

19.18. Reequilibrio mal gestionado

Para entrenar con clases raras pueden utilizarse:

- sobremuestreo;
- submuestreo;
- ponderación;
- SMOTE;
- generación sintética.

Estas técnicas deben aplicarse únicamente al entrenamiento.

Si se reequilibra el conjunto de prueba, se altera artificialmente:

- prevalencia;
- VPP;
- VPN;
- carga de alertas;
- utilidad operativa.

La sensibilidad y especificidad pueden estimarse en muestras enriquecidas, pero los valores predictivos deben trasladarse a la prevalencia real.

19.19. Datos sintéticos tratados como evidencia externa

Los datos sintéticos pueden ser útiles para:

- explorar escenarios;
- aumentar entrenamiento;
- probar software;
- estudiar condiciones raras.

No deben considerarse equivalentes a evidencia externa real.

Un generador aprende de los datos disponibles. Puede reproducir sus sesgos, omitir fenómenos desconocidos y crear observaciones excesivamente regulares.

La validación final debe incluir muestras reales e independientes del proceso objetivo.

19.20. Contaminación mediante SMOTE

SMOTE genera observaciones sintéticas interpolando entre positivos.

Si se aplica antes de dividir los datos, observaciones sintéticas relacionadas con una muestra pueden aparecer en entrenamiento y prueba.

La secuencia correcta es:

1. dividir unidades originales;
2. aplicar SMOTE solo al entrenamiento;

3. evaluar sobre datos reales sin sintetizar.

No debe utilizarse SMOTE para aumentar artificialmente el número de positivos del conjunto de validación.

19.21. Exclusión retrospectiva de valores atípicos

Los valores atípicos pueden deberse a:

- error de medida;
- error de registro;
- muestra real extrema;
- nueva condición;
- fallo del modelo.

Eliminar después de observar que empeoran el resultado introduce sesgo.

Los criterios de exclusión deben definirse previamente.

Debe distinguirse:

Dato inválido

Existe evidencia objetiva de error.

Dato válido extremo

Representa una condición real aunque infrecuente.

Dato fuera de dominio

No pertenece al uso autorizado, pero debe evaluarse si el sistema lo detectó correctamente.

Fallo del modelo

No debe eliminarse del análisis.

La exclusión debe documentarse con:

- razón;
- responsable;
- momento;
- impacto;
- análisis con y sin la observación.

19.22. Limpieza de datos basada en la etiqueta

Un procedimiento puede eliminar registros que «no encajan» con la clase esperada.

Ejemplo: excluir una muestra microbiológicamente positiva porque su espectro parece similar a negativos.

Esta decisión puede eliminar precisamente los casos difíciles que el modelo debe detectar.

La limpieza debe basarse en criterios de calidad del dato, no en si el resultado favorece la hipótesis.

19.23. Error del patrón de referencia ignorado

Cuando el modelo discrepa con la etiqueta, suele asumirse que el algoritmo está equivocado.

Pero el patrón de referencia puede fallar por:

- muestreo insuficiente;
- distribución heterogénea;

- límite de detección;
- error de laboratorio;
- panel sensorial inconsistente;
- transcripción;
- contaminación cruzada.

Debe revisarse una muestra de discordancias mediante:

- repetición;
- adjudicación;
- segundo método;
- revisión ciega;
- análisis de trazabilidad.

Sin embargo, no debe reinterpretarse sistemáticamente la referencia para hacer coincidir los resultados con el modelo.

19.24. Referencia incorporada a las entradas

Puede existir circularidad cuando el modelo utiliza como predictor una medida estrechamente relacionada con la definición del resultado.

Ejemplo:

- predecir incumplimiento microbiológico utilizando el propio recuento final;
- predecir vida útil usando una valoración sensorial posterior;
- clasificar calidad con una variable que forma parte de la especificación final.

La pregunta debe ser si el predictor estará disponible en el momento de uso y si aporta realmente capacidad anticipatoria.

19.25. Confusión entre correlación y predicción prospectiva

Un análisis retrospectivo puede encontrar asociaciones fuertes que no se mantienen en el futuro.

Las causas incluyen:

- cambios de proceso;
- variables proxy;
- tendencias temporales;
- intervención;
- causalidad inversa;
- selección de casos.

Por ejemplo, una variable administrativa puede correlacionarse con reclamaciones porque se introduce después de identificar el problema.

La validación prospectiva permite comprobar si la relación funciona en el momento real de decisión.

19.26. Mezcla de periodos anteriores y posteriores a una intervención

Si durante el periodo histórico se modificaron:

- limpieza;
- formulación;
- proveedor;

- método analítico;
- proceso;
- criterio de calidad;

la base puede contener regímenes distintos.

Una partición aleatoria mezcla ambos y facilita que el modelo utilice señales temporales.

Debe documentarse la cronología y realizar análisis por régimen.

Cuando la finalidad es predecir el futuro, el entrenamiento debe representar preferentemente las condiciones actuales o incluir mecanismos explícitos de cambio.

19.27. Validación en una población demasiado similar

Un conjunto denominado «externo» puede seguir siendo casi idéntico al de entrenamiento si procede de:

- la misma planta;
- el mismo periodo;
- los mismos proveedores;
- el mismo equipo;
- el mismo protocolo.

La independencia formal no garantiza la externalidad relevante.

Debe identificarse qué dimensión se pretende generalizar:

- tiempo;
- ubicación;
- equipo;
- matriz;
- población;
- organización.

Una validación externa robusta separa al menos una dimensión material.

19.28. Falta de validación prospectiva

La evaluación retrospectiva utiliza datos ya existentes. Es eficiente, pero puede contener:

- sesgos históricos;
- variables posteriores;
- selección;
- cambios en medición;
- información incompleta.

La validación prospectiva aplica el modelo congelado a casos nuevos antes de conocer el resultado.

Permite comprobar:

- disponibilidad real de entradas;
- tiempos;
- fallos;
- interacción humana;
- deriva;
- prevalencia;

• impacto.

Para aplicaciones críticas, la evidencia retrospectiva debería complementarse con evaluación prospectiva.

19.29. Optimismo por muestras de conveniencia

Las bases de datos suelen formarse con lo que resulta fácil medir.

Pueden sobrerrepresentar:

• lotes disponibles;
• productos estables;
• condiciones de laboratorio;
• muestras claras;
• defectos evidentes;
• equipos bien calibrados.

Y subrepresentar:

• fines de semana;
• turnos nocturnos;
• incidencias;
• productos raros;
• límites de especificación;
• condiciones adversas;
• fallos de sensores.

La muestra debe diseñarse para cubrir la población objetivo, no solo para facilitar el desarrollo.

19.30. Espectro de enfermedad o defecto limitado

Un modelo puede entrenarse con positivos muy evidentes y negativos muy limpios.

En producción encontrará casos fronterizos.

Ejemplo microbiológico:

• positivos de alta concentración;
• negativos completamente libres.

Pero el uso real exige detectar concentraciones próximas al límite.

Ejemplo visual:

• cuerpos extraños grandes;
• producto conforme perfecto.

Pero el reto real son fragmentos pequeños, transparentes o parcialmente ocultos.

La validación debe cubrir la severidad completa, especialmente la proximidad a la frontera de decisión.

19.31. Ausencia de casos difíciles

El conjunto de prueba debería incluir:

• falsos parecidos positivos;
• positivos sutiles;
• condiciones de interferencia;
• productos nuevos;

- entradas de baja calidad;
- combinaciones extremas;
- clases poco frecuentes.

Un test construido únicamente con casos fáciles mide una tarea simplificada, no la aplicación industrial.

19.32. Cambio de definición de la clase

Durante el proyecto puede cambiar el criterio:

- de defecto visible a defecto comercial;
- de presencia microbiológica a incumplimiento al final de vida útil;
- de fecha de duración mínima a caducidad;
- de rechazo sensorial a incumplimiento fisicoquímico.

Si cambia el resultado, el modelo y la validación deben revisarse.

No debe mantenerse la misma etiqueta nominal cuando su significado ha evolucionado.

19.33. Análisis de subgrupos posterior y selectivo

Analizar numerosos subgrupos después de ver los resultados puede generar diferencias fortuitas.

Sin embargo, omitir completamente los subgrupos también puede ocultar fallos.

La solución es:

- preespecificar los subgrupos críticos;
- limitar análisis exploratorios;
- ajustar la interpretación por multiplicidad;
- confirmar hallazgos en nuevos datos.

Los análisis no preespecificados deben presentarse como exploratorios, no como evidencia concluyente.

19.34. Agrupación de subgrupos incompatibles

Puede calcularse una métrica global mezclando:

- productos con mecanismos de deterioro distintos;
- equipos diferentes;
- microorganismos diferentes;
- matrices con propiedades diferentes;
- defectos críticos y menores.

El promedio resultante puede carecer de significado tecnológico.

Antes de agrupar debe justificarse que las poblaciones comparten:

- finalidad;
- relación predictor-resultado;
- patrón de referencia;
- umbral;
- consecuencias del error.

Si no, deben validarse por separado.

19.35. Dominio excesivamente amplio

Una empresa puede intentar aprobar el modelo para «todos los productos refrigerados» basándose en datos de dos o tres referencias.

La amplitud del dominio debe corresponder a la evidencia.

Una validación limitada no debe extrapolarse por similitud comercial.

Productos que parecen similares pueden diferir en:

- pH;
- actividad de agua;
- estructura;
- conservantes;
- microbiota;
- envasado;
- proceso;
- mecanismo de deterioro.

La conclusión correcta puede ser una aprobación restringida.

19.36. Confusión entre interpolación y extrapolación

Un modelo puede funcionar bien dentro del rango observado y fallar fuera de él.

Si fue entrenado con temperaturas entre 2 y 8 °C, predecir a 12 °C es extrapolar.

Los algoritmos no siempre indican que han abandonado el dominio.

Debe documentarse:

$$X\_j^{} \leq X\_j \leq X\_j^{}$$

y analizar también combinaciones multivariantes.

Una predicción fuera del rango no debe considerarse válida aunque el software genere un número.

19.37. Interpretación errónea de un (R^2) elevado

El (R^2) puede ser elevado cuando el rango de la variable es amplio.

Ejemplo:

- vidas útiles entre 5 y 180 días;
- error medio de 10 días.

El modelo puede explicar gran parte de la variación y seguir siendo inaceptable para asignar fechas.

Deben presentarse:

- MAE;
- sesgo;
- percentiles;
- errores extremos;
- sobreestimación;
- límites de acuerdo.

La correlación no demuestra concordancia.

19.38. Error medio próximo a cero

Un sesgo medio nulo puede resultar de la cancelación entre sobreestimaciones e infraestimaciones.

Debe analizarse:

$$E(e|\ e>0)$$

$$E(|e||\ e<0)$$

y la distribución completa.

En vida útil, las sobreestimaciones deben evaluarse separadamente porque su riesgo no queda compensado por predicciones conservadoras en otros lotes.

19.39. Dependencia exclusiva del RMSE

El RMSE penaliza valores extremos, pero no indica la dirección ni la frecuencia del error.

Dos modelos pueden presentar el mismo RMSE:

- uno con errores moderados frecuentes;
- otro con pocos errores catastróficos.

La selección depende de la aplicación.

Debe complementarse con:

- MAE;
- sesgo;
- percentiles;
- máximo;
- tasa de error peligroso;
- análisis por subgrupos.

19.40. Uso indiscriminado del MAPE

El MAPE presenta inestabilidad cuando el valor real se aproxima a cero.

También puede producir interpretaciones problemáticas en:

- concentraciones bajas;
- tiempos cortos;
- escalas logarítmicas;
- resultados censurados.

No debe utilizarse automáticamente porque produzca un porcentaje intuitivo.

La métrica debe corresponder a la escala científica de la variable.

19.41. AUC utilizada como única prueba

El AUC resume la discriminación sobre todos los umbrales, incluidos puntos industrialmente irrelevantes.

Un modelo puede tener AUC elevada y no alcanzar la sensibilidad mínima en la región operativa.

Debe comunicarse:

- sensibilidad al umbral utilizado;
- especificidad;
- valores predictivos;

- curva precision–recall;
- calibración;
- intervalos;
- subgrupos.

En seguridad, interesa el rendimiento donde:

$$\text{Sensibilidad} \geq S\_$$

no el área completa de la curva.

19.42. Interpretación causal de la explicabilidad

Métodos como SHAP identifican cómo contribuyen las variables a la predicción del modelo.

No demuestran:

- causalidad;
- corrección;
- ausencia de sesgo;
- validez científica;
- estabilidad.

Una variable puede ser importante porque actúa como proxy de una planta, campaña o proveedor.

Las explicaciones deben utilizarse para investigar, no como sustituto de la validación externa.

19.43. Confianza interna confundida con probabilidad real

Las redes neuronales y otros modelos pueden emitir valores entre 0 y 1 que parecen probabilidades.

Sin calibración, esos valores son puntuaciones.

Un valor de:

$$0{,}95$$

no significa automáticamente un 95 % de probabilidad real.

Debe evaluarse:

- calibración;
- Brier;
- pendiente;
- intercepto;
- estabilidad por subgrupos.

19.44. Ausencia de intervalos de confianza

Comunicar únicamente estimaciones puntuales impide valorar la precisión.

Debe distinguirse entre:

$$S=98\%$$

obtenida con 50 positivos y con 5.000 positivos.

Los intervalos deben respetar:

- estructura agrupada;
- tamaño;

- prevalencia;
- método de muestreo.

Cuando hay pocos eventos, la incertidumbre puede ser demasiado grande para aprobar el sistema.

19.45. Uso de intervalos incorrectos

La aproximación normal puede producir límites imposibles o poco fiables en proporciones cercanas a cero o uno.

Deben considerarse:

- Wilson;
- Clopper–Pearson;
- bootstrap agrupado;
- modelos jerárquicos.

El método debe documentarse.

19.46. Ignorar la incertidumbre por agrupamiento

Si los errores se concentran por planta o lote, asumir independencia infravalora la incertidumbre.

Debe utilizarse:

- bootstrap por conglomerados;
- efectos aleatorios;
- errores robustos;
- validación por grupos.

Un modelo probado sobre miles de unidades de pocos lotes no posee la misma evidencia que otro probado en numerosos lotes independientes.

19.47. Prueba de significación confundida con utilidad

Una mejora puede ser estadísticamente significativa pero industrialmente trivial.

Con muestras muy grandes, una diferencia mínima puede producir:

$$p<0{,}001$$

La decisión debe considerar:

- magnitud;
- intervalo;
- coste;
- seguridad;
- complejidad;
- capacidad operativa.

La pregunta no es solo si existe diferencia, sino si la diferencia justifica el cambio.

19.48. Ausencia de comparador razonable

Comparar un modelo avanzado con ausencia total de control puede exagerar su beneficio.

Debe compararse con:

- práctica actual;
- método simple;

- regla experta;
- regresión;
- carta de control;
- modelo microbiológico establecido.

Si un método sencillo ofrece rendimiento equivalente, puede ser preferible por:

- interpretabilidad;
- estabilidad;
- coste;
- mantenimiento;
- auditabilidad.

19.49. Comparación injusta entre modelos

Los modelos deben evaluarse con:

- mismos datos;
- mismas particiones;
- misma referencia;
- mismos criterios;
- mismo horizonte;
- misma información disponible.

No es válido comparar un modelo que utiliza variables posteriores con otro limitado a datos disponibles en tiempo real.

19.50. Falta de análisis de errores

Una métrica agregada no explica por qué falla el modelo.

Debe elaborarse una taxonomía:

- producto;
- planta;
- proveedor;
- nivel de concentración;
- severidad;
- calidad de imagen;
- proximidad al umbral;
- condición ambiental;
- dato ausente;
- fuera de dominio.

Los errores pueden revelar que el problema no es aleatorio, sino sistemático.

19.51. Casos discordantes eliminados del informe

Los fallos más importantes deben describirse.

Un informe serio incluye:

- falsos negativos críticos;
- sobreestimaciones extremas;

- fallos fuera de dominio;
- incertidumbre;
- causas;
- restricciones.

Ocultarlos impide una evaluación de riesgo real.

19.52. Validación exclusiva en laboratorio

Las condiciones de laboratorio suelen ser más controladas que la producción:

- iluminación estable;
- temperatura constante;
- muestras homogéneas;
- equipos calibrados;
- operarios expertos;
- ausencia de ruido.

La validación industrial debe comprobar:

- velocidad;
- suciedad;
- vibración;
- cambios de turno;
- retrasos;
- datos faltantes;
- condiciones logísticas;
- interacción humana.

Un modelo válido en laboratorio puede no ser apto en línea.

19.53. Confusión entre prototipo y sistema de producción

El prototipo puede utilizar:

- archivos preparados manualmente;
- variables completas;
- orden correcto;
- intervención experta.

El sistema real depende de:

- sensores;
- integraciones;
- bases;
- interfaces;
- comunicaciones;
- reglas de negocio;
- usuarios.

Debe validarse el sistema completo, no solo el notebook del científico de datos.

19.54. Ignorar la latencia

Una predicción puede ser exacta y llegar demasiado tarde.

Debe medirse:

$$[\ T_{total}$$

$$T_{captura} + T_{procesamiento} + T_{inferencia} + T_{comunicación} + T_{acción}\ ]$$

Si:

$$[\ T_{total}$$

$$T_{ventana\ de\ intervención}\ ]$$

el modelo carece de utilidad para esa decisión.

19.55. No validar el actuador

En sistemas automáticos, el software puede clasificar correctamente y el mecanismo físico fallar.

Debe evaluarse:

- señal;
- sincronización;
- expulsión;
- bloqueo;
- alarma;
- verificación.

El resultado de extremo a extremo es:

$$[\ P(\text{control efectivo})$$

$$P(\text{detección}) \times P(\text{transmisión}) \times P(\text{actuación})\ ]$$

Si cada componente tiene un 99 % de fiabilidad:

$$0{,}99^3 \approx 97{,}0\%$$

La fiabilidad total puede ser inferior a la de cada componente.

19.56. Supervisión humana nominal

Un sistema puede declararse «supervisado» aunque la persona:

- no comprenda la salida;
- no tenga tiempo;
- no pueda anular;
- carezca de información;
- acepte automáticamente.

Debe validarse el comportamiento humano mediante casos reales o simulados.

19.57. Automation bias no evaluado

Cuando los usuarios confían excesivamente en el algoritmo pueden ignorar:

- olor anómalo;
- desviación visible;
- dato imposible;

- historial del lote;
- conocimiento del proceso.

La formación y la interfaz deben fomentar confianza apropiada, no obediencia automática.

19.58. Falta de procedimiento para abstenciones

Un modelo puede identificar un caso como indeterminado, pero si la organización no define qué hacer, la función de seguridad queda anulada.

Debe existir una acción inequívoca:

$$\text{abstención} \Rightarrow \text{retención, revisión o análisis}$$

Nunca debe interpretarse como resultado negativo por defecto.

19.59. Imputación silenciosa

El sistema puede sustituir automáticamente datos ausentes sin informar al usuario.

Esto puede generar una predicción aparentemente completa basada en valores supuestos.

Debe registrarse:

- variable imputada;
- método;
- impacto;
- incertidumbre;
- condición de aceptación.

En aplicaciones críticas, determinados datos ausentes deben impedir la predicción.

19.60. Falta de detección fuera de dominio

Un modelo puede emitir una predicción confiada para un producto nuevo.

Sin detector de dominio o límites, la empresa puede interpretar el número como válido.

Debe comprobarse que:

- rangos inválidos se bloquean;
- combinaciones nuevas se detectan;
- la salida muestra advertencia;
- la decisión automática queda inhibida.

19.61. Aprobación basada en datos del proveedor

Las métricas del proveedor pueden proceder de:

- otra población;
- condiciones controladas;
- prevalencia distinta;
- productos diferentes;
- equipos distintos;
- selección favorable.

La empresa debe realizar una validación local o demostrar transferibilidad.

La certificación del proveedor no sustituye la evidencia en el entorno de uso.

19.62. Opacidad contractual

Cuando la empresa no conoce:

- versión;
- cambios;
- variables;
- límites;
- validación;
- actualizaciones;

no puede gobernar adecuadamente el sistema.

La confidencialidad industrial del proveedor puede proteger detalles, pero no debe impedir que el operador obtenga evidencia suficiente para evaluar el riesgo.

19.63. Cambios no versionados

Modificar:

- umbral;
- cámara;
- preprocesamiento;
- firmware;
- modelo;
- librería;

puede alterar el rendimiento.

Si el cambio no queda registrado, las predicciones posteriores no pueden vincularse a la evidencia de validación.

19.64. Reentrenamiento automático sin control

El aprendizaje continuo puede incorporar:

- etiquetas erróneas;
- datos sesgados;
- efectos de decisiones anteriores;
- manipulación;
- deriva.

Una versión que cambia continuamente no conserva una identidad estable.

Debe existir:

- modelo activo;
- modelo candidato;
- evaluación;
- aprobación;
- despliegue;
- reversión.

19.65. Ausencia de monitorización posterior

El rendimiento inicial puede degradarse.

Sin monitorización, la empresa no sabrá:

- si cambia la prevalencia;
- si aparecen nuevos productos;
- si deriva el sensor;
- si aumentan falsos negativos;
- si se pierde calibración.

La validación inicial sin vigilancia posterior es incompleta.

19.66. Monitorizar solo las entradas

La estabilidad de las variables no garantiza estabilidad de la relación con el resultado.

Puede existir deriva conceptual sin cambios visibles en (P(X)).

Debe verificarse periódicamente:

$$P(Y|X)$$

mediante resultados reales y muestreo independiente.

19.67. Verificar únicamente los positivos

Si solo se confirman las alertas, no se observan falsos negativos.

Esto produce sesgo de verificación.

Debe mantenerse un programa de muestreo de predicciones negativas.

19.68. Ajustar el modelo a sus propias decisiones

Si el modelo decide qué se analiza, las etiquetas futuras dependerán de él.

Este ciclo puede reforzar sus sesgos:

modelo arrow selección arrow datos observados arrow reentrenamiento

Deben preservarse muestras aleatorias o estrategias de exploración independientes.

19.69. Publicación selectiva

Puede existir una tendencia a publicar o comunicar solo los modelos exitosos.

Los intentos fallidos, bases alternativas o métricas desfavorables quedan ocultos.

Esto produce una visión exagerada del estado de la tecnología.

Un informe corporativo debería documentar:

- modelos descartados;
- razones;
- cambios;
- resultados;
- limitaciones;
- análisis no confirmados.

19.70. Comparación con literatura no homogénea

No debe afirmarse que un modelo es superior porque su accuracy supera la publicada en otro estudio si difieren:

- productos;
- clases;
- prevalencia;
- dificultad;
- referencia;
- partición;
- unidad;
- validación externa.

Las métricas solo son comparables bajo diseños suficientemente equivalentes.

19.71. Benchmark demasiado fácil

Los conjuntos académicos pueden contener:

- imágenes limpias;
- clases equilibradas;
- defectos evidentes;
- condiciones uniformes.

El rendimiento en benchmark no demuestra rendimiento industrial.

La validación debe reflejar:

- ruido;
- variabilidad;
- defectos raros;
- cambios;
- condiciones límite;
- costes.

19.72. Sobreinterpretación de resultados exploratorios

Un modelo preliminar puede indicar una dirección prometedora, pero no debe presentarse como herramienta validada.

Deben diferenciarse:

- prueba de concepto;
- validación interna;
- validación externa;
- evaluación prospectiva;
- impacto.

Cada nivel permite afirmaciones distintas.

19.73. Lenguaje excesivamente concluyente

Expresiones como:

- «garantiza»;
- «elimina el riesgo»;
- «predice con certeza»;
- «sustituye al laboratorio»;

• «evita todos los falsos negativos»;

suelen exceder la evidencia.

La redacción debe reflejar:

• población;
• versión;
• condiciones;
• métricas;
• incertidumbre;
• límites.

19.74. Lista de comprobación contra resultados artificiales

Antes de aceptar una métrica elevada debe preguntarse:

1. ¿La unidad de prueba es realmente independiente?
2. ¿Hay réplicas del mismo lote en otros conjuntos?
3. ¿La aumentación se realizó después de la partición?
4. ¿El preprocesamiento se ajustó solo con entrenamiento?
5. ¿La selección de variables estuvo dentro de la validación?
6. ¿Se utilizó información futura?
7. ¿El umbral se fijó antes del test?
8. ¿Cuántas configuraciones se probaron?
9. ¿El test se consultó repetidamente?
10. ¿La prevalencia es realista?
11. ¿Se eliminaron casos después de ver los errores?
12. ¿Se informan intervalos de confianza?
13. ¿Se analizaron subgrupos?
14. ¿Existen suficientes positivos independientes?
15. ¿La referencia es fiable?
16. ¿La validación es temporal o externa?
17. ¿Se probaron casos de frontera?
18. ¿Se validó el sistema completo?
19. ¿Se comparó con la práctica actual?
20. ¿Existe verificación prospectiva?

Una respuesta negativa a varias de estas preguntas reduce sustancialmente la credibilidad del resultado.

19.75. Ejemplo de resultado engañoso en espectroscopia

Supóngase un estudio con:

• 40 lotes;
• 50 espectros por lote;
• 2.000 observaciones;
• partición aleatoria al 80/20.

El modelo alcanza:

$$R^2=0,98$$

y:

$$RMSE=0,4$$

Sin embargo, los espectros de todos los lotes aparecen en entrenamiento y prueba.

Al repetir la evaluación dejando fuera lotes completos, el rendimiento cae a:

$$R^2=0,62$$

y:

$$RMSE=2,8$$

La primera evaluación medía reconocimiento de lotes conocidos. La segunda se aproxima a la capacidad real de predecir lotes nuevos.

19.76. Ejemplo de resultado engañoso en visión artificial

Se dispone de 1.000 imágenes originales, cada una aumentada diez veces.

La base contiene:

$$10.000 \text{ imágenes}$$

Si se divide después de la aumentación, versiones casi idénticas aparecen en entrenamiento y prueba.

El modelo alcanza:

$$accuracy=99\%$$

Cuando se evalúa con fotografías de una campaña posterior, obtiene:

$$accuracy=74\%$$

La diferencia no representa necesariamente deriva posterior. Puede revelar que la validación inicial estaba contaminada.

19.77. Ejemplo de resultado engañoso en vida útil

Un modelo se desarrolla con datos de distintos días de los mismos lotes. Algunos tiempos se utilizan para entrenar y otros para probar.

El modelo reconoce la trayectoria parcial del lote y predice los puntos posteriores con gran precisión.

Pero en el uso real deberá estimar la vida útil de un lote completamente nuevo.

La partición correcta debe reservar curvas completas.

19.78. Ejemplo de resultado engañoso en seguridad microbiológica

Un conjunto contiene:

- 5 % de positivos;
- 95 % de negativos.

El modelo presenta:

$$accuracy=96\%$$

Pero la matriz es:

$$VP=10$$

$$FN=40$$

$$FP=0$$

$$VN=950$$

La sensibilidad es:

$$[10] / [10+40]=20\%$$

El sistema falla en cuatro de cada cinco positivos. La accuracy elevada se debe al predominio de negativos.

19.79. Ejemplo de selección retrospectiva

Se prueban 30 umbrales sobre el conjunto de prueba y se elige aquel que proporciona:

$$Sensibilidad=98\%$$

El resultado se presenta como validación externa.

En realidad, el test se utilizó para optimizar. La sensibilidad debe confirmarse en otro conjunto independiente.

19.80. Ejemplo de exclusión problemática

Un modelo de detección falla en seis muestras positivas. El equipo observa que todas presentan turbidez elevada y las excluye por considerarlas «no representativas».

Si la turbidez puede aparecer en producción, esas muestras son precisamente parte del problema.

La conclusión correcta sería:

El modelo no está validado para productos con turbidez superior al rango especificado.

No:

El modelo alcanzó el 100 % después de eliminar las muestras problemáticas.

19.81. Medidas preventivas

Para evitar estos errores debe implantarse:

- protocolo prospectivo;
- separación por unidad real;
- validación agrupada;
- test bloqueado;
- pipeline reproducible;
- registro de experimentos;
- análisis estadístico independiente;
- informe completo;
- validación prospectiva;
- auditoría de fuga.

19.82. Auditoría de fuga de datos

La auditoría debería revisar:

- identificadores duplicados;
- similitud entre archivos;
- fechas;
- lotes;

- réplicas;
- transformaciones;
- imputación;
- selección de variables;
- historial de acceso al test;
- número de modelos;
- criterios de exclusión;
- correspondencia temporal.

En imágenes puede utilizarse detección de duplicados o similitud perceptual. En espectros, análisis de distancia. En datos tabulares, identificadores y combinaciones únicas.

La auditoría técnica debe complementarse con conocimiento del proceso. Dos registros con identificadores diferentes pueden proceder de la misma unidad física.

19.83. Registro de experimentos

Debe conservarse:

- hipótesis;
- fecha;
- versión de datos;
- variables;
- partición;
- hiperparámetros;
- métricas;
- resultado;
- decisión.

Esto permite conocer cuántos modelos se probaron y evita reconstrucciones selectivas.

19.84. Revisión estadística independiente

Una persona no involucrada en el desarrollo debería revisar:

- unidad;
- partición;
- tamaño;
- intervalos;
- multiplicidad;
- métricas;
- subgrupos;
- exclusiones;
- conclusiones.

La revisión independiente puede detectar supuestos que el equipo de desarrollo considera normales por familiaridad.

19.85. Reproducibilidad

Otro equipo debería poder ejecutar el protocolo y obtener los mismos resultados dentro de una tolerancia definida.

Se requiere conservar:

- datos;
- código;
- entorno;
- semillas;
- pesos;
- instrucciones;
- dependencias;
- versiones.

La imposibilidad de reproducir una métrica debilita la evidencia.

19.86. Validación negativa

La validación no solo debe confirmar dónde funciona el modelo. También debe identificar dónde no funciona.

Debe documentarse:

- productos excluidos;
- rangos no cubiertos;
- condiciones adversas;
- subgrupos débiles;
- errores conocidos;
- incertidumbre.

Esta evidencia negativa es esencial para evitar extensiones indebidas.

19.87. Principio de falsabilidad

Un protocolo científico debe permitir que el modelo fracase.

Si los criterios, exclusiones y análisis pueden adaptarse indefinidamente hasta obtener una conclusión positiva, no existe validación real.

La pregunta metodológica es:

¿Qué resultado conduciría a rechazar el modelo?

Si no existe una respuesta clara, el protocolo está diseñado para confirmar, no para validar.

19.88. Conclusión del bloque

Los resultados artificialmente elevados suelen proceder de cinco fuentes principales:

$$\text{fuga} + \text{dependencia} + \text{selección} + \text{falta de representatividad} + \text{comunicación incompleta}$$

La prevención exige:

- definir la unidad independiente;
- separar lotes, periodos y centros;
- encapsular el preprocesamiento;
- bloquear el conjunto de prueba;
- preespecificar umbrales y métricas;
- informar todos los errores relevantes;
- validar prospectivamente;

- reproducir el análisis;
- someterlo a revisión independiente.

Una cifra elevada no constituye evidencia si el modelo ha visto directa o indirectamente aquello que debía predecir.

La pregunta decisiva no es:

«¿Qué métrica alcanzó el algoritmo?»

Sino:

«¿Qué capacidad de generalización demuestra realmente el diseño mediante el cual se obtuvo esa métrica?»

Cuando esta pregunta no puede responderse con precisión, la aparente excelencia del modelo debe considerarse una hipótesis pendiente de validación, no una prueba de aptitud industrial.

20. Validación externa multicéntrica, temporal y prospectiva: demostrar que el modelo funciona fuera del entorno en el que fue creado

La validación interna responde a una pregunta limitada:

¿Qué rendimiento cabe esperar cuando el modelo se aplica a observaciones suficientemente parecidas a las utilizadas durante su desarrollo?

La validación externa formula una pregunta más exigente:

¿Mantiene el sistema un rendimiento aceptable cuando cambia alguna dimensión relevante del entorno de utilización?

Esta dimensión puede ser:

- temporal;
- geográfica;
- industrial;
- instrumental;
- tecnológica;
- composicional;
- microbiológica;
- organizativa.

La diferencia es esencial. Un modelo puede generalizar correctamente a nuevos registros procedentes de la misma planta, campaña y metodología analítica, pero fracasar cuando se traslada a otra instalación, un proveedor distinto, una formulación modificada o una nueva estación del año.

Por ello, la validación externa no debe definirse únicamente por la ausencia formal de observaciones repetidas. Debe existir una separación material entre el contexto de desarrollo y el de evaluación.

Las directrices TRIPOD+AI, aunque formuladas para modelos predictivos del ámbito sanitario, reflejan un principio metodológico transferible: la evaluación debe describir con precisión el origen, periodo, población, resultados, tratamiento de datos, rendimiento y limitaciones, evitando utilizar la expresión «modelo validado» como una propiedad absoluta e independiente del contexto.

En alimentación, la conclusión correcta no es:

> *«El modelo ha sido validado externamente.»*

Sino:

> *«La versión especificada fue evaluada en plantas, periodos, proveedores, equipos y productos no utilizados durante el desarrollo, obteniendo las prestaciones descritas dentro de esas condiciones.»*

20.1. Externalidad formal y externalidad científica

Un conjunto de prueba puede ser formalmente independiente y, sin embargo, aportar una evidencia externa débil.

Ejemplo

Se reserva aleatoriamente un 20 % de los lotes de una base obtenida en:

- la misma fábrica;
- el mismo trimestre;

- los mismos proveedores;
- el mismo instrumento;
- los mismos operadores;
- la misma formulación.

Las muestras de prueba no se utilizaron para ajustar el algoritmo, pero representan prácticamente el mismo entorno.

El estudio demuestra cierta capacidad de generalización a lotes nuevos dentro de una población muy próxima. No demuestra portabilidad temporal, instrumental o industrial.

La externalidad científica depende de la distancia relevante entre:

$$P_{desarrollo}(X,Y)$$

y:

$$P_{validación}(X,Y)$$

Cuando ambas distribuciones son casi idénticas, la evaluación externa proporciona una prueba limitada. Cuando son excesivamente diferentes, el modelo puede estar siendo evaluado fuera de su finalidad prevista.

La validación debe encontrar una tensión razonable:

$$\text{independencia suficiente} + \text{relevancia para el uso previsto}$$

20.2. Transportabilidad y generalización

La transportabilidad es la capacidad del modelo para conservar su utilidad cuando se traslada a un entorno distinto.

Puede analizarse en varios niveles.

Transportabilidad interna

Aplicación a nuevos lotes de la misma planta.

Transportabilidad temporal

Aplicación a periodos posteriores.

Transportabilidad instrumental

Aplicación a otro equipo, sensor o laboratorio.

Transportabilidad industrial

Aplicación a otra línea o planta.

Transportabilidad composicional

Aplicación a nuevas formulaciones dentro de una familia autorizada.

Transportabilidad geográfica

Aplicación a regiones con distintas materias primas, condiciones ambientales o prácticas productivas.

Transportabilidad organizativa

Aplicación bajo procedimientos, operarios y sistemas informáticos distintos.

No existe una transportabilidad general. Un modelo puede ser temporalmente estable y no ser transferible entre instrumentos, o funcionar en varias plantas y degradarse al cambiar de proveedor.

La validación debe declarar qué dimensión ha sido realmente probada.

20.3. Jerarquía de evidencia externa

Puede establecerse una jerarquía práctica.

Nivel 0. Rendimiento aparente

El modelo se evalúa sobre los mismos datos con los que fue entrenado.

No constituye validación.

Nivel 1. Validación interna aleatoria

Se reservan observaciones de la misma base.

Útil para desarrollo inicial, pero vulnerable a dependencia y similitud excesiva.

Nivel 2. Validación interna agrupada

Se reservan lotes, curvas, proveedores o días completos.

Evalúa generalización a nuevas unidades dentro del mismo entorno.

Nivel 3. Validación temporal

Se utilizan periodos posteriores al desarrollo.

Evalúa estabilidad frente al paso del tiempo.

Nivel 4. Validación en otra línea o equipo

Introduce variabilidad operacional o instrumental.

Nivel 5. Validación externa en otra planta

Evalúa portabilidad industrial.

Nivel 6. Validación multicéntrica

Incluye varias instalaciones independientes.

Permite estimar heterogeneidad entre contextos.

Nivel 7. Evaluación prospectiva en modo sombra

El modelo congelado opera sobre datos reales antes de conocerse el resultado, sin intervenir todavía en la decisión.

Nivel 8. Evaluación prospectiva de impacto

Se estudia si el uso del sistema mejora el resultado industrial o sanitario.

La jerarquía no implica que todo modelo necesite alcanzar el nivel máximo. La exigencia depende de la función, criticidad y amplitud de la declaración comercial o científica.

Un sistema utilizado únicamente en una línea concreta puede validarse localmente. Un proveedor que afirme que su modelo es aplicable a toda una categoría alimentaria debería aportar evidencia multicéntrica, multiproducto e instrumental.

20.4. Validación temporal

La validación temporal reserva datos posteriores a los utilizados para desarrollar el modelo:

$$t_{entrenamiento} < t_{validación}$$

Ejemplo:

- entrenamiento: enero de 2023 a diciembre de 2024;
- ajuste y calibración: enero a junio de 2025;
- prueba temporal: julio de 2025 a junio de 2026.

Este diseño reproduce mejor la utilización real que una división aleatoria, ya que el modelo siempre se empleará para predecir el futuro, no una muestra aleatoria del pasado.

La validación temporal permite detectar cambios asociados a:

- estacionalidad;
- campañas;
- proveedores;
- materias primas;
- equipos;
- prevalencia;
- prácticas de producción;
- microbiota;
- logística.

El Reglamento europeo de IA exige, para los sistemas de alto riesgo incluidos en su ámbito, que los datos de entrenamiento, validación y prueba sean relevantes, suficientemente representativos y adecuados a la finalidad prevista. Este principio refuerza la necesidad de comprobar que los datos utilizados cubren las variaciones temporales previsibles del entorno donde operará el sistema.

20.5. Separación temporal estricta

La división temporal debe impedir cualquier información futura en el desarrollo.

No deberían utilizarse para ajustar el modelo:

- estadísticas calculadas con todo el periodo;
- normalizaciones que incluyan datos futuros;
- categorías identificadas posteriormente;
- imputaciones basadas en años posteriores;
- resultados de campañas futuras;
- parámetros recalibrados con el periodo de prueba.

Debe cumplirse:

$$\_t>t\_0 \text{ desarrollo del modelo}$$

donde (_t>t_0) representa toda información posterior a la fecha de corte.

20.6. Validación temporal móvil

Cuando existe un volumen de datos suficiente puede utilizarse una evaluación progresiva.

Ejemplo:

1. entrenar hasta diciembre;
2. evaluar enero;
3. entrenar o actualizar según protocolo;
4. evaluar febrero;
5. continuar secuencialmente.

Este diseño se denomina a veces *rolling-origin evaluation* o evaluación de origen móvil.

Permite estimar:

- evolución del error;
- degradación gradual;
- respuesta a campañas;

- estabilidad del umbral;
- necesidad de recalibración.

La actualización del modelo no debe mezclarse con la evaluación de la versión original. Deben distinguirse:

$$M_0$$

modelo congelado inicial, y:

$$M_t$$

modelo actualizado en el periodo (t).

20.7. Estacionalidad

Numerosos procesos alimentarios presentan variaciones estacionales.

Ejemplos:

- composición de leche;
- humedad de cereales;
- acidez de frutas;
- carga microbiológica ambiental;
- temperatura logística;
- presión de plagas;
- materias primas agrícolas;
- actividad de agua;
- oxidación.

Un modelo validado durante una única estación puede no haber observado la amplitud real.

La validación temporal debería incluir, cuando sea relevante:

$$\text{invierno,primavera,verano,otoño}$$

o varias campañas completas.

No debe utilizarse el término «validación anual» cuando el periodo cubierto no representa realmente la variabilidad estacional del producto.

20.8. Validación por campañas

En productos agrícolas, pesqueros o de origen animal, la campaña puede constituir una unidad más relevante que el mes.

La separación correcta podría ser:

- entrenamiento: campañas 2022–2024;
- prueba externa: campaña 2025.

Este diseño es más exigente que mezclar aleatoriamente muestras de todas las campañas.

La campaña de validación debe describirse en términos de:

- origen;
- clima;
- variedades;
- composición;
- incidencias;

- condiciones de almacenamiento;
- proveedores;
- prevalencia de defectos.

20.9. Validación geográfica

La validación geográfica analiza la transferencia entre regiones.

Puede resultar necesaria cuando cambian:

- materias primas;
- razas o variedades;
- clima;
- agua;
- prácticas agrícolas;
- logística;
- microbiota;
- proveedores;
- procedimientos.

Una validación geográfica no consiste únicamente en evaluar una fábrica situada en otra localidad si ambas reciben la misma materia prima, utilizan el mismo equipo y comparten todos los procedimientos.

Debe identificarse qué fuentes reales de variabilidad introduce la nueva localización.

20.10. Validación entre plantas

La validación entre plantas es una de las pruebas más relevantes para modelos corporativos.

Supóngase que el modelo se desarrolla en la planta A y se evalúa en B, C y D:

$$M_A \text{ arrow } D_B,D_C,D_D$$

Debe calcularse el rendimiento por planta:

$$M_B, M_C, M_D$$

y no únicamente el rendimiento agregado:

$$M_{total}$$

El promedio puede ocultar una degradación crítica en una instalación.

Ejemplo

Planta	Sensibilidad	Especificidad
B	98 %	91 %
C	97 %	89 %
D	78 %	94 %
Global	93 %	91 %

El valor global podría parecer aceptable, pero la planta D presenta una sensibilidad incompatible con el uso previsto.

20.11. Validación multicéntrica

Un estudio multicéntrico incorpora datos procedentes de varios centros independientes.

Su finalidad no es solo aumentar el tamaño muestral. Permite evaluar:

- variabilidad entre centros;
- estabilidad del rendimiento;
- dependencia de procedimientos locales;
- efectos de equipos;
- diferencias de población;
- necesidad de calibración local.

El diseño debería documentar por centro:

- número de unidades;
- número de positivos;
- productos;
- periodos;
- equipos;
- referencia;
- prevalencia;
- exclusiones;
- rendimiento.

Las recomendaciones metodológicas modernas para modelos predictivos insisten en comunicar el rendimiento y la composición de las poblaciones de evaluación, en lugar de reducir toda la evidencia a una métrica agregada. Aunque TRIPOD+AI se haya elaborado para investigación clínica, su exigencia de transparencia en el origen y evaluación de los datos es plenamente trasladable a la validación alimentaria.

20.12. Centro como efecto aleatorio

Cuando existen varias plantas, el rendimiento puede analizarse mediante modelos jerárquicos.

Para una variable continua:

$$[y_{ij} \alpha + u_j + \beta yij + \varepsilon ij]$$

donde:

- (i) identifica la observación;
- (j), la planta;
- (u_j), el efecto específico del centro.

Esto permite estimar:

- variabilidad entre plantas;
- sesgo medio;
- plantas atípicas;
- incertidumbre de transferencia.

Para resultados binarios pueden utilizarse modelos logísticos mixtos.

La inclusión del centro como efecto aleatorio evita tratar todas las observaciones como independientes e idénticamente distribuidas.

20.13. Heterogeneidad del rendimiento

La heterogeneidad no debe considerarse únicamente ruido estadístico. Puede revelar que la relación aprendida depende del entorno.

Debe analizarse:

$$\mathrm{Var}(M_j)$$

donde (M_j) es la métrica en cada centro.

Puede estudiarse mediante:

- intervalos por centro;
- modelos jerárquicos;
- metaanálisis;
- análisis de interacción;
- gráficos de bosque;
- pruebas de heterogeneidad.

No obstante, una prueba estadística no sustituye la interpretación tecnológica.

Una diferencia entre plantas puede explicarse por:

- formulación;
- proceso térmico;
- instrumento;
- operario;
- prevalencia;
- método de referencia;
- tratamiento de datos.

20.14. Gráfico de bosque

Para cada planta puede representarse:

- estimación;
- intervalo de confianza;
- peso;
- valor combinado.

Ejemplo conceptual:

$$S_j \pm \mathrm{IC}\ 95\ \%$$

La visualización permite identificar:

- centros con baja precisión;
- centros con bajo rendimiento;
- resultados inconsistentes;
- evidencia dominada por una sola planta.

Un promedio ponderado favorable no debe justificar la aprobación de una planta que incumple un criterio crítico.

20.15. Regla de aprobación por centro

La validación multicéntrica puede utilizar dos enfoques.

Aprobación global

Se exige que la estimación combinada cumpla el criterio.

Riesgo: ocultar centros débiles.

Aprobación local

Cada centro debe cumplir el límite.

Riesgo: tamaños reducidos e intervalos amplios.

Una estrategia razonable combina:

$$LCB_{95\%}(M_{global}) \geq M_{}$$

y:

$$M_j \geq M_{local\ mínimo}\ \forall\ j$$

junto con la ausencia de una heterogeneidad tecnológicamente preocupante.

Cuando un centro no dispone de suficientes casos, la conclusión debe ser «evidencia insuficiente», no «centro validado».

20.16. Validación leave-one-site-out

En proyectos con varias plantas puede utilizarse un esquema de exclusión completa por centro:

1. entrenar en todas salvo una;
2. evaluar en la planta excluida;
3. repetir para cada planta.

$$LOSO\text{-}CV$$

Este diseño estima cómo se comportaría el procedimiento en una planta no vista.

Es más exigente que repartir aleatoriamente datos de todas las plantas entre pliegues.

Sin embargo, no sustituye una validación prospectiva final, especialmente si las decisiones de arquitectura se ajustan observando reiteradamente los resultados de todas las plantas.

20.17. Desarrollo multicéntrico frente a validación multicéntrica

No es lo mismo entrenar con datos de varias plantas que validar externamente en varias plantas.

Si A, B y C participan en entrenamiento y prueba, el modelo ha visto características de los tres centros.

La verdadera validación de transferencia requeriría una planta D no utilizada durante el desarrollo:

$$[\ D_{train}\ \{A,B,C\}\ ]$$

$$[\ D_{external}\ \{D\}\ ]$$

Un desarrollo multicéntrico suele mejorar la representatividad, pero no elimina la necesidad de evaluar centros nuevos cuando se pretende una generalización abierta.

20.18. Validación instrumental

Los modelos basados en:

- NIR;
- Raman;

- hiperespectral;
- visión;
- sensores de gases;
- biosensores;
- cromatografía;
- equipos rápidos;

pueden depender intensamente del instrumento.

La validación debe diferenciar:

- mismo instrumento;
- misma marca y modelo;
- unidades distintas del mismo modelo;
- modelos distintos;
- laboratorios distintos.

No debe asumirse equivalencia porque dos equipos compartan denominación comercial.

Deben evaluarse:

- desplazamiento de señal;
- resolución;
- ruido;
- calibración;
- respuesta espectral;
- iluminación;
- software;
- firmware;
- preparación de muestra.

20.19. Transferencia de calibración

En espectroscopia puede requerirse una transferencia entre instrumento maestro y secundario.

La transferencia puede utilizar:

- estandarización directa;
- estandarización por tramos;
- muestras de transferencia;
- corrección de sesgo;
- actualización local.

Pero después de transferir debe validarse el rendimiento en el nuevo instrumento.

No basta con demostrar que los espectros se parecen. Debe comprobarse que las predicciones conservan:

- sesgo;
- precisión;
- sensibilidad;
- estabilidad;
- criterios por producto.

20.20. Validación entre laboratorios

Cuando el patrón de referencia procede de varios laboratorios debe analizarse la comparabilidad.

Diferencias entre laboratorios pueden deberse a:

- método;
- preparación;
- reactivos;
- incubación;
- interpretación;
- calibración;
- límites;
- operadores.

El modelo puede parecer menos transferible cuando parte de la discrepancia procede de la referencia.

Debe incluirse, cuando proceda:

- ensayo interlaboratorio;
- materiales comunes;
- controles;
- estimación de reproducibilidad;
- adjudicación de discordancias.

20.21. Validación entre proveedores

Los proveedores pueden introducir variabilidad en:

- composición;
- granulometría;
- tratamiento;
- humedad;
- microbiota;
- origen;
- contaminantes;
- comportamiento tecnológico.

Para demostrar generalización a proveedores nuevos puede aplicarse:

leave-one-supplier-out

Un modelo entrenado con todos los proveedores conocidos puede funcionar bien dentro de ellos y fallar ante un origen nuevo.

La aprobación debe distinguir entre:

- proveedores incluidos;
- proveedores equivalentes;
- proveedores nuevos sujetos a verificación;
- proveedores excluidos.

20.22. Proveedor como variable predictora

Incluir el identificador de proveedor puede mejorar el rendimiento histórico, pero plantea un riesgo.

El modelo puede aprender:

$$\text{proveedor} \rightarrow \text{resultado}$$

sin utilizar las propiedades reales de la materia prima.

Esta relación puede desaparecer cuando el proveedor mejora, modifica el origen o se incorpora otro.

El identificador puede utilizarse si existe una justificación y una política clara para categorías desconocidas, pero no debe sustituir la caracterización del producto.

20.23. Validación composicional

Un modelo no puede extenderse automáticamente a formulaciones nuevas.

Debe definirse la distancia respecto a la población validada.

Ejemplo:

$$X = (pH, a_w, sal, grasa, conservante)$$

Una nueva formulación puede estar dentro del rango de cada variable y, sin embargo, presentar una combinación nunca observada.

La validación composicional debe analizar:

- rangos;
- interacciones;
- mecanismos;
- ingredientes funcionales;
- estructura;
- proceso;
- flora.

20.24. Familias de producto

Puede agruparse una familia cuando existe evidencia de que comparte:

- mecanismo de deterioro;
- variables relevantes;
- método de referencia;
- proceso;
- dominio;
- consecuencia del error.

No debe agruparse únicamente por una categoría comercial.

Por ejemplo, dos productos denominados «salsas refrigeradas» pueden diferir sustancialmente en:

- emulsión;
- pH;
- tratamiento térmico;
- conservantes;
- envase;
- flora alterante.

20.25. Validación de una nueva referencia

La incorporación de un producto puede gestionarse mediante niveles.

Nivel 1. Dentro del dominio y altamente similar

Puede requerir confirmación limitada.

Nivel 2. Dentro de rangos, pero combinación nueva

Requiere validación adicional.

Nivel 3. Fuera de un rango crítico

No debe utilizarse el modelo sin reentrenamiento o validación completa.

La similitud debe cuantificarse con criterios tecnológicos y estadísticos.

20.26. Validación microbiológica entre matrices

En microbiología predictiva, una relación validada en un alimento no se transfiere automáticamente a otro.

La eficacia de un modelo puede depender de:

- pH;
- actividad de agua;
- estructura;
- grasa;
- fase acuosa;
- microbiota competidora;
- conservantes;
- estado fisiológico del microorganismo;
- envase;
- temperatura.

Codex reconoce el valor de los modelos matemáticos para validar medidas de control, pero señala que deben ser apropiadamente validados para la aplicación alimentaria específica, considerando la incertidumbre y pudiendo requerir ensayos adicionales.

20.27. Matriz, cepa y condiciones

Un modelo microbiológico externo debería probarse con variabilidad suficiente de:

- lotes;
- cepas;
- niveles de inoculación;
- estados fisiológicos;
- condiciones;
- formulaciones.

No es necesario incluir cualquier combinación imaginable, pero sí aquellas que puedan modificar materialmente la conclusión.

La validación debe aclarar si se pretende generalizar a:

- una cepa;
- un cóctel;
- una especie;

• flora natural;
• contaminación real.

20.28. Validación prospectiva

La validación prospectiva aplica el modelo congelado antes de conocer el resultado verdadero.

La secuencia es:

$$X_t \text{ arrow } Yt \text{ arrow registro arrow } Yt+\Delta \text{ arrow comparación}$$

A diferencia de un estudio retrospectivo, demuestra que:

• las entradas están disponibles a tiempo;
• el flujo de datos funciona;
• el modelo puede ejecutarse;
• las predicciones se registran;
• el resultado aparece después;
• no existe información futura.

20.29. Modo sombra

En modo sombra, el modelo opera en producción pero no modifica todavía la decisión.

Permite analizar:

• rendimiento prospectivo;
• latencia;
• disponibilidad;
• diferencias con la práctica;
• errores;
• carga de alertas;
• datos fuera de dominio.

Es especialmente útil antes de automatizar una función de elevada criticidad.

El modo sombra no elimina el riesgo metodológico si los usuarios modifican indirectamente el proceso al conocer las predicciones. Idealmente, las salidas no deberían influir en la práctica durante la fase puramente evaluativa, salvo que exista una alerta cuya omisión sea éticamente o legalmente inadmisible.

20.30. Estudio silencioso ciego

En un diseño más estricto:

• el modelo genera predicciones;
• las predicciones quedan bloqueadas;
• los responsables operativos no las ven;
• se obtiene el resultado real;
• un equipo independiente realiza la comparación.

Este diseño minimiza la retroalimentación.

No siempre será posible en seguridad, ya que una alerta potencialmente grave podría requerir actuación. El protocolo debe definir previamente cómo gestionar esa eventualidad.

20.31. Validación prospectiva con intervención

Después del modo sombra puede evaluarse el sistema como parte del flujo real.

Se comparan:

- procedimiento convencional;
- procedimiento asistido por IA.

El objetivo deja de ser exclusivamente predictivo y pasa a medir impacto.

Resultados posibles

- tiempo hasta detección;
- lotes no conformes;
- desperdicio;
- análisis;
- falsas alarmas;
- incidencias;
- decisiones;
- costes;
- carga humana.

NIST considera la monitorización posterior al despliegue necesaria para comprobar que el sistema opera de forma fiable en escenarios reales, detectar salidas imprevistas y reconocer cambios de distribución o contexto.

20.32. Diseños comparativos

Antes-después

Se compara el periodo previo y posterior.

Limitación: pueden cambiar simultáneamente otras variables.

Líneas paralelas

Una línea utiliza IA y otra mantiene el procedimiento habitual.

Limitación: las líneas pueden no ser equivalentes.

Despliegue escalonado

Las plantas incorporan el sistema en momentos distintos.

Permite utilizar las plantas todavía no desplegadas como comparación temporal.

Aleatorización por lote

Puede ser útil en decisiones de calidad de bajo riesgo.

No resulta aceptable asignar aleatoriamente un control potencialmente inferior cuando existe una implicación sanitaria.

20.33. Validación prospectiva de vida útil

En vida útil, el modelo debe generar la estimación al inicio o en el momento de uso y mantenerse bloqueada hasta observar el evento.

Deben evitarse:

- ajustes posteriores;

- corrección manual no registrada;
- uso de datos futuros;
- cambio de criterio;
- extensión del estudio solo en casos favorables.

Cada lote debe conservar:

- predicción;
- intervalo;
- fecha;
- versión;
- condiciones;
- resultado final.

20.34. Resultados censurados en validación prospectiva

Puede ocurrir que algunos productos no alcancen el final de vida útil durante el seguimiento.

Estos casos no deben asignarse arbitrariamente a la fecha de finalización del estudio.

Debe utilizarse análisis de supervivencia o prolongar el seguimiento según el protocolo.

La censura debe ser independiente, en la medida de lo posible, del resultado.

20.35. Validación prospectiva microbiológica

Cuando el resultado es raro, puede requerirse:

- periodo prolongado;
- varias plantas;
- combinación con estudios enriquecidos;
- challenge tests;
- muestreo dirigido y aleatorio.

La muestra prospectiva real permite estimar utilidad en prevalencia operativa, mientras que los conjuntos enriquecidos permiten estimar sensibilidad con suficientes positivos.

Ambas evidencias son complementarias.

20.36. Tamaño muestral multicéntrico

El tamaño total no basta. Debe existir información suficiente por:

- centro;
- producto;
- clase;
- periodo;
- subgrupo crítico.

Si una planta aporta miles de negativos y otra solo diez observaciones, el resultado global estará dominado por la primera.

La planificación debe considerar:

$$[\, n_{total}\ \Sigma_{j=1}^{J} n_j \,]$$

y no únicamente (n_{total}).

Para sensibilidad, el número relevante es:

$$n_{+,j}$$

positivos reales del centro (j).

20.37. Correlación intracentro

Las unidades producidas en una misma planta pueden compartir factores comunes.

La independencia efectiva se reduce cuando existe correlación intracentro.

El efecto de diseño puede aproximarse como:

$$DE = 1 + (m-1)\rho$$

donde:

- (m) es el tamaño medio por centro;
- (ρ), la correlación intraclase.

La incertidumbre debe estimarse mediante:

- bootstrap por centro;
- modelos jerárquicos;
- errores robustos;
- análisis por conglomerados.

20.38. Número de centros

Muchos datos procedentes de dos plantas no ofrecen la misma evidencia de transportabilidad que datos repartidos entre numerosas plantas.

La generalización industrial depende tanto de:

$$n_{observaciones}$$

como de:

$$J_{centros}$$

Cuando el número de centros es pequeño, la estimación de heterogeneidad será imprecisa.

20.39. Selección de centros

Los centros no deberían elegirse solo por:

- facilidad;
- buena calidad de datos;
- proximidad;
- colaboración;
- rendimiento esperado.

La muestra multicéntrica debería cubrir:

- plantas grandes y pequeñas;
- equipos nuevos y antiguos;
- diferentes productos;
- distintas condiciones;
- diferentes niveles de prevalencia;

• variación organizativa.

Si se seleccionan únicamente instalaciones muy controladas, la generalización será excesivamente optimista.

20.40. Centro de peor caso

Puede incluirse deliberadamente una instalación con condiciones difíciles:

• mayor variabilidad;
• equipo antiguo;
• proveedor complejo;
• clima extremo;
• alta carga;
• productos fronterizos.

El objetivo no es castigar al modelo, sino comprobar sus límites.

Un fallo en el peor caso puede conducir a:

• restricción;
• calibración local;
• mejora de sensores;
• exclusión de esa planta;
• revisión del proceso.

20.41. Calibración global y local

Un modelo puede discriminar adecuadamente en todas las plantas y presentar probabilidades mal calibradas en algunas.

Debe evaluarse:

$$\alpha_j$$

intercepto de calibración por planta, y:

$$\beta_j$$

pendiente por planta.

Puede ser necesaria una recalibración local:

$$\mathbf{[\ logit(p_j)\ \alpha_j + \beta_j logit(p)\ ]}$$

La recalibración no debe utilizarse para ocultar una discriminación deficiente.

20.42. Modelo global o modelos locales

Existen varias estrategias.

Modelo global

Una versión para todas las plantas.

Ventajas:

• mantenimiento simplificado;
• más datos;
• consistencia.

Riesgos:

- promedio que no representa centros específicos;
- pérdida de rendimiento local.

Modelo global con recalibración local

La estructura es común, pero se ajusta la probabilidad o el umbral.

Modelos locales

Cada planta dispone de una versión.

Ventajas:

- adaptación.

Riesgos:

- menos datos;
- múltiples validaciones;
- gobernanza compleja;
- divergencia de versiones.

La elección debe basarse en heterogeneidad, volumen, criticidad y capacidad de mantenimiento.

20.43. Umbral global o local

Un umbral único facilita la gestión, pero puede producir tasas de error diferentes por planta.

Un umbral local puede adaptarse a:

- prevalencia;
- capacidad;
- coste;
- calibración.

Sin embargo, cada umbral constituye una configuración que debe validarse.

No debe modificarse localmente por conveniencia operativa sin aprobación.

20.44. Adaptación de dominio

Las técnicas de adaptación de dominio intentan ajustar el modelo a un nuevo entorno.

Pueden utilizar:

- recalibración;
- reponderación;
- transferencia;
- ajuste fino;
- alineación de distribuciones;
- muestras locales.

Pero la adaptación modifica el sistema.

La secuencia debe ser:

1. detectar la brecha;
2. formular la adaptación;
3. utilizar datos separados;
4. congelar la nueva versión;

5. validar localmente;

6. comparar con el modelo anterior;

7. aprobar.

20.45. Generalización de dominio

La generalización de dominio pretende desarrollar un modelo capaz de funcionar en centros no vistos sin ajuste local.

Para demostrarla debe evaluarse en centros completamente excluidos.

No basta una validación aleatoria dentro de una base multicéntrica.

La afirmación:

«El modelo es independiente de la planta»

requiere evidencia de desempeño consistente en plantas no participantes.

20.46. Externalidad del equipo evaluador

La evaluación puede ser realizada por:

- el equipo desarrollador;
- un equipo interno independiente;
- una planta externa;
- un laboratorio independiente;
- un tercero.

La independencia del evaluador reduce el riesgo de:

- selección favorable;
- exclusiones retrospectivas;
- interpretación benevolente;
- cambios no registrados.

No siempre es necesaria una validación por tercero, pero cuanto mayor sea la criticidad y amplitud comercial, mayor debería ser la independencia.

20.47. Validación del proveedor por el cliente

Un proveedor de IA puede aportar una validación multicéntrica general.

El cliente debe efectuar, como mínimo:

- verificación local;
- análisis de compatibilidad;
- evaluación del dominio;
- prueba de integración;
- confirmación de métricas críticas;
- evaluación de flujo operativo.

La evidencia del proveedor y la del usuario responden a preguntas distintas.

Evidencia del proveedor

¿Puede funcionar el producto en entornos representativos?

Evidencia del usuario

¿Funciona esta configuración en mi planta, con mis matrices, equipos y decisiones?

20.48. Portabilidad de un modelo comercial

Un dossier comercial debería declarar:

- plantas de desarrollo;
- plantas externas;
- productos;
- equipos;
- periodos;
- prevalencia;
- métricas por centro;
- intervalos;
- fallos;
- exclusiones;
- limitaciones;
- actualizaciones.

No debería limitarse a:

«Validado en más de un millón de muestras.»

Un millón de mediciones procedentes de pocos lotes o de una sola instalación puede aportar menos evidencia de portabilidad que una muestra menor distribuida entre centros independientes.

20.49. Análisis de equivalencia

Cuando se traslada un modelo a un nuevo equipo o planta puede plantearse una hipótesis de equivalencia.

Se define un margen:

$$\Delta$$

y se evalúa si la diferencia de rendimiento permanece dentro de:

$$[-\Delta,+\Delta]$$

El margen debe ser tecnológicamente justificado.

No debe elegirse después de observar la diferencia.

20.50. Análisis de no inferioridad

Para una métrica donde valores altos son mejores:

$$[\ M\_nuevo-M\_referencia$$

$$-\Delta\ ]$$

Si el límite inferior del intervalo supera $(-\Delta)$, puede concluirse no inferioridad.

En seguridad, el margen no debe permitir una pérdida relevante de sensibilidad.

La no inferioridad en accuracy global no basta si empeora la tasa de falsos negativos.

20.51. Fallo de transportabilidad

Cuando el modelo funciona en desarrollo y falla externamente deben investigarse cuatro posibilidades.

Cambio de datos

$$P(X)$$

ha variado.

Cambio de prevalencia

$$P(Y)$$

ha variado.

Cambio conceptual

$$P(Y|\ X)$$

ha cambiado.

Cambio de referencia

La medición de (Y) no es equivalente.

La solución depende de la causa.

No debe aplicarse automáticamente un nuevo umbral.

20.52. Investigación de discrepancias entre plantas

Debe revisarse:

- composición;
- proceso;
- equipo;
- operador;
- variables faltantes;
- metodología;
- prevalencia;
- calidad de referencia;
- calibración;
- dominio.

Puede descubrirse que una planta no incumple el modelo, sino que opera bajo un mecanismo diferente.

20.53. Restricción del dominio

Una conclusión negativa en un centro no implica necesariamente descartar el modelo.

Puede aprobarse para:

$$[\ \_\text{aprobado}$$

$$\_\text{total}$$

$$\_\text{no validado}\ ]$$

Debe configurarse el sistema para impedir el uso en el dominio excluido.

No basta con incluir una advertencia en un informe que los usuarios podrían ignorar.

20.54. Reentrenamiento tras validación externa fallida

Utilizar los datos externos para reentrenar transforma esa muestra en parte del desarrollo.

Después será necesario un nuevo conjunto independiente.

La secuencia es:

D_externo 1 arrow reentrenamiento

D_externo 2 arrow confirmación

No debe comunicarse el rendimiento sobre (D_externo 1) después de incorporarlo al modelo como validación final.

20.55. Repetición de evaluaciones externas

Cuando se ajusta repetidamente el modelo a varios centros conocidos, estos dejan de representar entornos verdaderamente nuevos.

Debe conservarse, cuando sea posible:

- un centro final;
- una campaña posterior;
- un periodo prospectivo;
- una muestra de confirmación.

20.56. Validación externa y deriva

La validación externa y la monitorización se relacionan, pero no son equivalentes.

Validación externa

Evalúa un entorno independiente antes o durante la autorización.

Monitorización

Comprueba continuamente el entorno de producción después de la autorización.

NIST sitúa la prueba, evaluación, verificación, validación y monitorización dentro de un ciclo continuo que incluye cambio, incidentes, recuperación y eventual retirada.

Un modelo puede superar una validación externa y degradarse posteriormente.

20.57. Criterios de aceptación multicéntrica

Un protocolo puede exigir:

Rendimiento combinado

$$LCB_{95\%}(S_{global}) \geq S_{}$$

Rendimiento local

$$S_j \geq S_{local\ mínimo}$$

para todas las plantas autorizadas.

Heterogeneidad

Ausencia de diferencias tecnológicamente inaceptables.

Calibración

Probabilidades adecuadas global y localmente.

Robustez

Mantenimiento bajo equipos y condiciones diferentes.

Dominio

Detección de referencias o configuraciones no representadas.

Prospección

Confirmación sobre producción posterior.

20.58. Matriz de validación externa

Dimensión	Desarrollo	Validación externa	Evidencia obtenida
Tiempo	2023–2025	2026	Estabilidad temporal
Plantas	A y B	C y D	Transportabilidad industrial
Proveedores	1–4	5 y 6	Generalización de materia prima
Equipos	Sensor X1	Sensores X2 y X3	Transferencia instrumental
Productos	Referencias 1–5	Referencias 6–7	Generalización composicional
Estaciones	Invierno-primavera	Verano-otoño	Estacionalidad
Laboratorios	Laboratorio interno	Laboratorio externo	Reproducibilidad
Usuarios	Equipo técnico	Operarios habituales	Generalización operacional

La matriz debe señalar también las dimensiones no evaluadas.

20.59. Informe de validación externa

El informe debería incluir:

1. finalidad;
2. versión;
3. diferencia respecto al desarrollo;
4. centros;
5. periodos;
6. productos;
7. equipos;
8. métodos;
9. criterios;
10. tamaño;
11. prevalencia;
12. métricas globales;
13. métricas locales;

14. intervalos;
15. calibración;
16. subgrupos;
17. heterogeneidad;
18. casos discordantes;
19. limitaciones;
20. decisión.

20.60. Declaraciones científicamente válidas

Declaración excesiva

El modelo es universalmente aplicable a productos cárnicos.

Declaración defendible

El modelo mantuvo el rendimiento especificado en cuatro plantas, tres campañas, dos configuraciones instrumentales y las familias de productos descritas. No se ha demostrado su aplicabilidad a productos fermentados, equipos distintos ni formulaciones fuera de los rangos establecidos.

Declaración excesiva

La validación multicéntrica demuestra que no existe deriva.

Declaración defendible

La evaluación multicéntrica demuestra transportabilidad inicial entre los centros estudiados. La estabilidad posterior deberá confirmarse mediante monitorización.

Declaración excesiva

El modelo funciona en cualquier proveedor.

Declaración defendible

El modelo fue probado con seis proveedores. Los proveedores nuevos deberán superar el procedimiento de incorporación y compatibilidad.

20.61. Errores frecuentes en validación externa

- denominar externo a un split aleatorio;
- agregar centros sin informar resultados individuales;
- utilizar los datos externos para recalibrar y evaluar simultáneamente;
- seleccionar centros favorables;
- ocultar subgrupos pequeños;
- no ajustar la incertidumbre por conglomerados;
- evaluar solamente una campaña;
- asumir equivalencia instrumental;
- ignorar cambios del método de referencia;
- afirmar generalización más allá de las dimensiones probadas.

20.62. Estrategia mínima para una empresa alimentaria

Para un modelo de criticidad moderada:

1. partición por lotes;
2. prueba temporal;
3. evaluación en otra línea;
4. modo sombra;
5. monitorización.

Para un modelo de elevada criticidad:

1. desarrollo multicampaña;
2. prueba temporal independiente;
3. validación en otra planta;
4. evaluación multicéntrica;
5. condiciones adversas;
6. prueba prospectiva;
7. verificación independiente;
8. aprobación local;
9. vigilancia continua.

20.63. La externalidad como propiedad de la pregunta

No existe un conjunto universalmente externo.

Un proveedor puede ser externo respecto al tiempo, pero interno respecto a la planta.

Una planta puede ser externa respecto a la organización, pero utilizar el mismo equipo y producto.

La validez externa siempre debe expresarse respecto a una dimensión:

externo respecto a tiempo centro equipo proveedor producto geografía

Esta precisión evita declaraciones ambiguas.

20.64. Conclusión del bloque

La validación externa no consiste en apartar datos al azar. Consiste en someter al modelo a una variabilidad relevante que no haya influido en su desarrollo.

Un modelo científicamente transportable debe demostrar que:

- mantiene el rendimiento en periodos posteriores;
- funciona en unidades industriales independientes;
- no depende de un equipo concreto sin declararlo;
- conserva su comportamiento entre proveedores y campañas;
- mantiene la calibración;
- reconoce condiciones no representadas;
- presenta heterogeneidad controlada;
- puede operar prospectivamente;
- no genera una pérdida de seguridad en el flujo real.

La afirmación:

«El modelo ha sido validado externamente»

solo tiene valor cuando se acompaña de una respuesta completa a cinco preguntas:

1. ¿Externo respecto a qué?
2. ¿En qué población y condiciones?
3. ¿Con qué versión y configuración?
4. ¿Qué rendimiento alcanzó en cada centro y subgrupo?
5. ¿Hasta dónde puede generalizarse legítimamente la conclusión?

La validación externa no convierte un modelo en universal. Delimita, con evidencia, la distancia que puede recorrer fuera del entorno donde fue creado.

21. Validación frente a métodos de referencia, expertos humanos y procedimientos convencionales

Un modelo de inteligencia artificial no puede evaluarse en el vacío. Su rendimiento solo adquiere significado cuando se compara con una referencia capaz de representar, con incertidumbre conocida, el fenómeno que pretende medir o predecir.

En calidad, vida útil y seguridad alimentaria, esa referencia puede adoptar formas muy distintas:

- un método analítico normalizado;
- un método microbiológico de referencia;
- un estudio de desafío;
- un estudio de durabilidad;
- una inspección instrumental;
- un panel sensorial;
- el consenso de especialistas;
- un procedimiento histórico de la empresa;
- una regla tecnológica;
- otro modelo predictivo.

Cada referencia responde a una pregunta diferente. Comparar un modelo con un método de laboratorio permite evaluar concordancia analítica. Compararlo con inspectores humanos permite estimar su reproducibilidad operacional. Compararlo con el procedimiento vigente permite determinar si aporta un beneficio incremental. Ninguna de estas comparaciones, considerada aisladamente, demuestra necesariamente que el sistema sea apto para adoptar una decisión de seguridad.

La validación debe distinguir tres conceptos:

validez frente a la referencia

equivalencia o no inferioridad frente al comparador

utilidad incremental dentro del proceso

Un modelo puede concordar razonablemente con una referencia y no mejorar la práctica actual. También puede superar a los operadores humanos y seguir siendo insuficiente frente a un criterio sanitario exigente.

21.1. Qué significa método de referencia

Un método de referencia es un procedimiento reconocido o previamente validado que se utiliza para establecer el estado de una muestra o para comparar el funcionamiento de un método alternativo.

Codex recomienda dar preferencia a métodos oficiales desarrollados por organizaciones internacionales competentes y cuya fiabilidad haya sido establecida considerando, según proceda, selectividad, exactitud, precisión, sensibilidad, límites de detección y cuantificación, practicabilidad y aplicabilidad.

En microbiología de la cadena alimentaria, la serie ISO 16140 diferencia la validación de métodos alternativos de su posterior verificación en el laboratorio usuario. ISO 16140-2 contempla estudios de comparación del método y estudios interlaboratorio para métodos cualitativos y cuantitativos; ISO 16140-3 regula la verificación local de métodos de referencia y de métodos alternativos ya validados.

Esta distinción resulta directamente trasladable a la IA:

- la validación general demuestra que el método o modelo puede alcanzar determinadas prestaciones;

- la verificación local confirma que esas prestaciones pueden reproducirse en la planta, laboratorio, equipo y matriz concretos del usuario.

Una empresa no debería considerar suficiente que el proveedor haya validado su algoritmo en otras instalaciones. Debe demostrar que la configuración implantada funciona correctamente en su propio contexto.

21.2. El método de referencia no es necesariamente una verdad perfecta

La expresión *ground truth* puede inducir a pensar que la etiqueta utilizada para validar el modelo representa una verdad absoluta.

En realidad, todo patrón de referencia puede contener incertidumbre derivada de:

- muestreo;
- heterogeneidad de la muestra;
- preparación;
- recuperación;
- límite de detección;
- precisión;
- especificidad;
- interpretación;
- variabilidad entre laboratorios;
- variabilidad entre expertos.

Si el método de referencia presenta una repetibilidad o reproducibilidad limitada, no es razonable interpretar toda discordancia como error exclusivo del algoritmo.

Sea:

$$Y^*$$

el estado verdadero, no observable directamente, y:

$$Y_R$$

el resultado del método de referencia.

Puede ocurrir:

$$Y_R \neq Y^*$$

El modelo se evalúa normalmente frente a (Y_R), no frente a (Y^*). Por tanto, el rendimiento observado incorpora simultáneamente:

$$\text{error del modelo} + \text{error de la referencia} + \text{error de muestreo}$$

Esta realidad no invalida la comparación, pero obliga a caracterizar la referencia y a revisar cuidadosamente las discordancias.

21.3. Error de muestreo frente a error analítico

En seguridad microbiológica, el principal problema puede no encontrarse en la técnica analítica, sino en la porción de muestra seleccionada.

Una contaminación puede distribuirse de manera heterogénea. El modelo puede detectar señales asociadas a un problema del lote mientras que la muestra microbiológica concreta resulta negativa, o a la inversa.

Debe distinguirse:

Error analítico

El procedimiento no detecta o cuantifica correctamente el analito presente en la porción analizada.

Error de muestreo

La porción analizada no representa adecuadamente el lote.

La concordancia entre IA y laboratorio estará limitada por ambos componentes.

El protocolo debe documentar:

- unidad de muestreo;
- masa o volumen;
- número de incrementos;
- localización;
- homogeneización;
- conservación;
- tiempo hasta análisis;
- plan estadístico de muestreo.

Un resultado de laboratorio técnicamente perfecto sobre una muestra no representativa no constituye una referencia perfecta para todo el lote.

21.4. Validación analítica y validación predictiva

La validación de un método analítico y la validación de un modelo predictivo comparten principios, pero no son equivalentes.

Un método analítico suele evaluar:

- selectividad;
- exactitud;
- precisión;
- linealidad;
- rango;
- límite de detección;
- límite de cuantificación;
- robustez;
- recuperación.

AOAC estructura sus requisitos de desempeño mediante criterios previamente definidos y contempla atributos como rango, recuperación, precisión, límite de cuantificación y funcionamiento colaborativo o interlaboratorio según el tipo de método.

Un modelo de IA puede requerir, además:

- generalización;
- calibración probabilística;
- comportamiento fuera de dominio;
- deriva;
- desempeño temporal;
- dependencia de prevalencia;
- interacción humana;

• impacto operacional.

Por ello, la validación analítica constituye una fuente metodológica valiosa, pero no cubre por sí sola todos los riesgos de un sistema predictivo.

21.5. Método alternativo frente a método de referencia

Un modelo puede actuar como método alternativo cuando pretende reproducir o aproximar el resultado de una técnica reconocida.

Ejemplos:

• espectroscopia frente a cromatografía;
• visión artificial frente a inspección manual;
• sensor rápido frente a cultivo microbiológico;
• IA sobre imágenes microscópicas frente a recuento convencional;
• modelo quimiométrico frente a análisis composicional.

La pregunta central será:

¿Produce el sistema alternativo resultados suficientemente concordantes con la referencia para la finalidad prevista?

No debe exigirse necesariamente identidad perfecta, pero sí una equivalencia compatible con la decisión.

21.6. Comparación pareada

La comparación más informativa utiliza las mismas muestras para el modelo y para el método de referencia.

Para cada unidad (i):

$$(Yi, YR, i)$$

Este diseño reduce la variabilidad derivada de comparar poblaciones distintas.

En clasificación binaria se construye una tabla pareada:

Referencia	IA positiva	IA negativa
Positiva	(a)	(b)
Negativa	(c)	(d)

Donde:

• (a): concordancia positiva;
• (d): concordancia negativa;
• (b): IA negativa y referencia positiva;
• (c): IA positiva y referencia negativa.

Las celdas discordantes (b) y (c) son esenciales para comparar los métodos.

21.7. Prueba de McNemar

Cuando se comparan resultados binarios pareados puede utilizarse la prueba de McNemar, centrada en las discordancias:

[^2

$$[(b-c)^2] / [b+c]]$$

o una variante corregida o exacta cuando el número de discordancias es reducido.

La prueba analiza si existe asimetría entre:

- casos en los que la IA falla y la referencia no;
- casos en los que la IA identifica un positivo que la referencia clasifica negativamente.

Sin embargo, una ausencia de diferencia estadísticamente significativa no demuestra equivalencia.

Puede deberse a:

- muestra insuficiente;
- pocos positivos;
- baja potencia;
- amplios intervalos.

La equivalencia o no inferioridad requiere un diseño y un margen predefinidos.

21.8. Acuerdo porcentual

El acuerdo global se calcula como:

$$P_o = [a+d] / [a+b+c+d]$$

Es intuitivo, pero puede ser elevado por la prevalencia de una clase.

Si el 99 % de las muestras son negativas, dos métodos que clasifican casi todo como negativo mostrarán un acuerdo elevado aunque su capacidad para detectar positivos sea insuficiente.

Deben calcularse separadamente:

Acuerdo positivo

$$PPA = [a] / [a+b]$$

cuando la referencia define los positivos.

Acuerdo negativo

$$NPA = [d] / [c+d]$$

Estos conceptos se aproximan a sensibilidad y especificidad respecto a la referencia, pero la terminología de acuerdo puede ser más prudente cuando no existe un estándar perfecto.

21.9. Coeficiente kappa

El coeficiente kappa intenta corregir el acuerdo esperado por azar:

$$\mathbf{[P_o-P_e] / [1-P_e]]}$$

donde:

- (P_o) es el acuerdo observado;
- (P_e), el acuerdo esperado según las distribuciones marginales.

Un valor de uno representa acuerdo completo; cero indica un acuerdo similar al esperado por azar.

Sin embargo, kappa presenta limitaciones:

- depende de la prevalencia;
- depende del desequilibrio marginal;
- puede ser bajo pese a un acuerdo porcentual alto;
- no pondera automáticamente la gravedad de los desacuerdos;
- no sustituye sensibilidad y especificidad.

No debe aprobarse un modelo únicamente porque alcance una categoría verbal como «acuerdo sustancial». Los puntos de corte cualitativos de kappa no constituyen criterios universales.

21.10. Kappa ponderado

Cuando las categorías tienen orden puede utilizarse un kappa ponderado.

Ejemplo:

- conforme;
- defecto menor;
- defecto mayor;
- defecto crítico.

Una confusión entre categorías adyacentes no debería penalizarse igual que una confusión entre conforme y defecto crítico.

El kappa ponderado asigna pesos:

$$w_{ij}$$

a cada tipo de desacuerdo.

La elección de los pesos debe reflejar la consecuencia tecnológica y fijarse antes del análisis.

No debe utilizarse una ponderación favorable seleccionada después de observar los resultados.

21.11. Concordancia para variables continuas

Cuando el modelo y el método de referencia producen valores continuos, la correlación no es suficiente.

Deben analizarse:

- sesgo;
- error absoluto;
- pendiente;
- intercepto;
- límites de acuerdo;
- precisión en diferentes concentraciones;
- heterocedasticidad;
- capacidad próxima a los límites de decisión.

Dos métodos pueden presentar:

$$r=0,99$$

y, sin embargo, diferir sistemáticamente.

Por ejemplo:

$$Y=1,2Y+3$$

conservaría una correlación muy elevada, pero produciría un sesgo proporcional y constante.

21.12. Regresión ordinaria y error en ambas variables

La regresión lineal ordinaria presupone habitualmente que la variable independiente se mide sin error relevante.

Cuando tanto el modelo como la referencia contienen error, este supuesto puede ser inadecuado.

Pueden considerarse métodos como:

• regresión de Deming;
• Passing–Bablok;
• modelos de error en variables;
• modelos jerárquicos.

La regresión de Deming incorpora la relación entre las varianzas de error de ambos métodos:

$$[\lambda \sigma^2_{referencia} \sigma^2_{modelo}]$$

La estimación dependerá de conocer o aproximar esa relación.

El objetivo no debe reducirse a obtener una recta con buen ajuste, sino a determinar si el sesgo constante y proporcional son compatibles con la finalidad.

21.13. Análisis de Bland–Altman

Para cada muestra:

$$d_i = Y_i - Y_{R,i}$$

$$m_i = +Y_{R,i2}$$

Se analiza el sesgo medio y los límites de acuerdo:

$$d \pm 1{,}96 s_d$$

El método permite identificar:

• sesgo constante;
• dispersión;
• heterocedasticidad;
• valores extremos;
• errores dependientes de la magnitud.

Los límites estadísticos deben compararse con límites de aceptación tecnológica.

Un intervalo de acuerdo de:

$$-1{,}5 \text{ a } +1{,}7 \text{ UFC/g}$$

puede ser estadísticamente descriptivo, pero inaceptable para una decisión microbiológica.

21.14. Coeficiente de correlación de concordancia

El coeficiente de concordancia de Lin combina correlación y proximidad a la línea de identidad:

$$[\rho_c \rho C_b]$$

donde:

• (ρ) es la correlación de precisión;
• (C_b) mide la desviación respecto a la línea de concordancia.

Puede ser más informativo que la correlación simple, pero sigue siendo una medida resumida.

Debe complementarse con:

• sesgo;
• límites de acuerdo;
• error por rango;
• criterios de decisión.

21.15. Concordancia en torno al límite regulatorio o interno

La concordancia global puede ser menos relevante que el comportamiento próximo al límite.

Sea (L) el criterio de aceptación.

Un error analítico lejos de (L) puede no modificar la decisión:

$$Y\_R\ L$$

o:

$$Y\_R\ L$$

Pero pequeñas discrepancias cerca de:

$$Y\_R \approx L$$

pueden cambiar la clasificación.

La validación debe incluir suficientes muestras dentro de una zona crítica:

$$L-\delta \leq Y\_R \leq L+\delta$$

y evaluar:

- probabilidad de clasificación;
- falsas aceptaciones;
- falsos rechazos;
- incertidumbre;
- repetibilidad.

21.16. Zona gris

Cuando la referencia y el modelo presentan incertidumbre cerca del límite puede definirse una zona gris.

Por ejemplo:

$$Y<L\_1 \Rightarrow \text{conforme}$$

$$L\_1 \leq Y \leq L\_2 \Rightarrow \text{indeterminado}$$

$$Y>L\_2 \Rightarrow \text{no conforme}$$

La amplitud debe derivarse de:

- incertidumbre analítica;
- variabilidad tecnológica;
- consecuencias del error;
- capacidad de confirmación.

La zona gris evita atribuir certeza artificial a resultados fronterizos, pero no debe utilizarse para excluir del análisis los casos más difíciles.

21.17. Probabilidad de detección

Para métodos cualitativos, la probabilidad de detección puede analizarse según el nivel de analito o contaminación:

$$\mathbf{[\ POD(c)\ P(resultado\ positivo|\ c)\ ]}$$

donde (c) es la concentración.

AOAC utiliza marcos de requisitos de desempeño para métodos cualitativos y cuantitativos y contempla la probabilidad de detección como herramienta para evaluar métodos binarios, especialmente en niveles próximos a la capacidad de detección.

En IA, la curva POD puede mostrar:

- baja detección a concentraciones muy reducidas;
- transición alrededor de un nivel;
- saturación en concentraciones altas.

La validación no debería limitarse a positivos muy evidentes.

21.18. Límite de detección del sistema completo

Un modelo de IA puede no tener un límite de detección químico en sentido estricto, pero sí una capacidad funcional mínima.

Ejemplos:

- tamaño mínimo de un cuerpo extraño;
- concentración mínima detectable;
- intensidad mínima de una señal;
- nivel de deterioro necesario;
- proporción mínima de superficie afectada.

Puede definirse:

$$[LOD_{p} c: POD(c) \geq p]$$

Por ejemplo, el nivel al que la probabilidad de detección alcanza el 95 %.

Este límite debe evaluarse con el sistema completo, incluyendo:

- muestreo;
- sensor;
- preprocesamiento;
- modelo;
- umbral;
- actuación.

21.19. Repetibilidad del modelo

La repetibilidad evalúa resultados bajo condiciones esencialmente iguales:

- mismo equipo;
- mismo operador;
- mismo laboratorio;
- corto periodo;
- misma muestra.

Para un sistema determinista, repetir exactamente la misma entrada digital debería producir la misma salida. Pero repetir la medición física puede generar variación.

Debe diferenciarse:

Repetibilidad computacional

$$f(x)=\text{resultado estable}$$

Repetibilidad de medición

$$x_1,x_2,,x_r \text{ arrow } y_1,y_2,,y_r$$

La segunda incluye variabilidad de colocación, muestreo, preparación, iluminación o sensor.

21.20. Reproducibilidad

La reproducibilidad evalúa cambios más amplios:

- operador;
- día;
- equipo;
- laboratorio;
- planta;
- lote de reactivo.

Puede estimarse mediante un diseño factorial o jerárquico:

$$[Y_{ijkl} \mu + O_i + D_j + E_k + L_l + \varepsilon]$$

donde los términos representan fuentes de variación.

La ISO 16140 utiliza estudios interlaboratorio para valorar la reproducibilidad de métodos microbiológicos alternativos, mientras que la verificación local comprueba la correcta implantación del método validado por el usuario.

Un modelo comercial debería demostrar no solo buen funcionamiento en el centro desarrollador, sino reproducibilidad suficiente en los centros donde se pretende utilizar.

21.21. Precisión intermedia

Entre repetibilidad y reproducibilidad completa se encuentra la precisión intermedia.

Evalúa variación dentro de la misma organización al cambiar:

- día;
- operador;
- equipo;
- calibración;
- turno.

En aplicaciones alimentarias puede ser una de las pruebas más relevantes, porque reproduce la operación rutinaria sin exigir todavía un estudio multicéntrico completo.

21.22. Verificación local

La existencia de una validación general no elimina la necesidad de verificación local.

La verificación debe demostrar que el usuario puede reproducir las prestaciones bajo:

- sus matrices;
- su personal;
- su equipo;
- su entorno;

• sus condiciones.

ISO 16140-3 establece precisamente un protocolo para verificar en el laboratorio usuario tanto métodos de referencia como métodos alternativos previamente validados.

Para un modelo de IA, la verificación local debería incluir:

• correcta instalación;
• compatibilidad de sensores;
• integridad de variables;
• reproducción de muestras control;
• rendimiento mínimo;
• gestión de errores;
• correcta aplicación del umbral;
• registros;
• respuesta ante fallo.

21.23. Equivalencia

La equivalencia pretende demostrar que la diferencia entre el modelo y el comparador permanece dentro de un margen aceptable.

Se define un margen:

$$\Delta$$

y una diferencia:

$$D=M\_IA-M\_referencia$$

Para declarar equivalencia, el intervalo de confianza debe quedar dentro de:

$$[-\Delta,+\Delta]$$

La prueba de ausencia de diferencia no demuestra equivalencia.

La equivalencia requiere:

• margen prospectivo;
• tamaño suficiente;
• comparación pareada o adecuadamente controlada;
• análisis por variable crítica.

21.24. No inferioridad

La no inferioridad pretende demostrar que la IA no es peor que el comparador por más de un margen aceptable.

Para una métrica donde valores altos son mejores:

$$D=M\_IA-M\_comparador$$

Se declara no inferioridad cuando:

$$LCB\_1-\alpha(D)>-\Delta$$

El margen (Δ) debe ser tecnológicamente justificable.

No puede elegirse tan amplio que permita una degradación material de seguridad.

21.25. No inferioridad en sensibilidad

Cuando la sensibilidad es crítica:

$$D\_S=S\_IA-S\_referencia$$

El margen debería reflejar el máximo aumento tolerable de falsos negativos.

Por ejemplo, una pérdida absoluta del 2 % puede parecer pequeña, pero sobre 50.000 positivos potenciales representaría:

$$50.000\times0,02=1.000$$

casos adicionales omitidos.

El margen debe evaluarse también en cifras absolutas y consecuencias.

21.26. Superioridad

La superioridad pretende demostrar una mejora:

$$LCB\_1-\alpha(D)>0$$

o superior a una diferencia mínima relevante:

$$LCB\_1-\alpha(D)>\Delta\_relevante$$

Una mejora estadísticamente significativa no implica necesariamente relevancia industrial.

Debe distinguirse:

$$\text{significación estadística}$$

de:

$$\text{beneficio tecnológico}$$

21.27. Comparación con un método imperfecto

Cuando la referencia no es perfecta, la sensibilidad y especificidad observadas del modelo pueden estar sesgadas.

Existen métodos como:

- resolución de discordancias;
- referencia compuesta;
- consenso de métodos;
- modelos de clases latentes;
- adjudicación experta.

Cada uno tiene limitaciones.

Resolución de discordancias

Solo se repiten o analizan con otro método los casos discordantes.

Riesgo: sesgo, porque las concordancias no se verifican con la misma intensidad.

Referencia compuesta

Se combinan varias pruebas.

Riesgo: la definición puede ser compleja y depender de reglas arbitrarias.

Clase latente

Se estima un estado no observado a partir de varios métodos.

Riesgo: depende de supuestos estadísticos difíciles de verificar.

Ningún método elimina completamente la incertidumbre de referencia.

21.28. Revisión de discordancias

Las discordancias deben revisarse de forma ciega cuando sea posible.

El equipo debería desconocer qué resultado produjo la IA para evitar reinterpretaciones favorables.

La revisión puede incluir:

- repetición del método;
- segundo laboratorio;
- segundo método;
- inspección documental;
- revisión de la muestra;
- análisis de trazabilidad.

Debe conservarse el análisis original y presentar:

- resultados antes de adjudicación;
- resultados después;
- reglas utilizadas;
- impacto sobre métricas.

21.29. Sesgo de resolución de discordancias

Si todos los resultados discordantes se resuelven utilizando un método relacionado con la IA, el análisis puede favorecerla.

Ejemplo: un modelo espectroscópico discrepa con un método químico y se adjudica el caso mediante otra técnica espectroscópica.

La adjudicación debería ser suficientemente independiente.

No debe eliminarse una discordancia simplemente porque la explicación del algoritmo parezca plausible.

21.30. Referencia compuesta

Puede definirse positivo cuando se cumple una combinación:

$$Y=1\ (A=1)(B=1)$$

o:

$$Y=1\ (A=1)(B=1)$$

La regla cambia radicalmente el número de positivos.

Una definición mediante “OR” aumenta sensibilidad de la referencia, pero puede reducir especificidad. Una mediante “AND” hace lo contrario.

La regla debe establecerse antes del análisis y justificarse científicamente.

21.31. Comparación con expertos humanos

En visión, sensorial, clasificación de defectos o interpretación de señales, el comparador puede ser una persona o grupo de especialistas.

La afirmación:

«La IA supera al experto»

es incompleta.

Debe especificarse:

- quiénes son los expertos;
- nivel de experiencia;
- formación;
- número;
- condiciones;
- tiempo;
- información disponible;
- patrón de referencia;
- variabilidad individual;
- posibilidad de revisión.

NIST señala que los sistemas que aumentan o sustituyen actividades humanas requieren algún tipo de métrica de referencia humana, aunque reconoce la dificultad de sistematizarla porque humanos y sistemas pueden ejecutar la tarea de manera diferente.

21.32. El experto no constituye automáticamente el estándar de verdad

Una persona puede cometer:

- fatiga;
- error de percepción;
- sesgo;
- inconsistencia;
- variación temporal;
- dependencia del contexto.

El rendimiento humano debe medirse, no asumirse.

Puede ocurrir que:

$$\text{acuerdo entre expertos} < \text{acuerdo IA-experto}$$

En ese caso, la etiqueta humana presenta incertidumbre sustancial.

21.33. Acuerdo intraobservador

El acuerdo intraobservador mide la consistencia de la misma persona cuando evalúa nuevamente las mismas muestras.

Puede analizarse presentando casos repetidos de forma ciega y aleatoria.

Si un experto cambia frecuentemente de decisión, el modelo no puede evaluarse como si la primera etiqueta fuese una verdad exacta.

21.34. Acuerdo interobservador

El acuerdo interobservador mide la consistencia entre expertos.

Puede calcularse mediante:

- acuerdo porcentual;
- kappa;
- kappa multirrater;
- correlación intraclase;
- modelos jerárquicos.

Una baja concordancia puede indicar:

- criterios ambiguos;
- formación insuficiente;
- categoría subjetiva;
- casos fronterizos;
- necesidad de mejorar la referencia.

21.35. Consenso de expertos

El consenso puede establecerse mediante:

- mayoría;
- unanimidad;
- adjudicación por un experto senior;
- discusión conjunta;
- método Delphi;
- panel independiente.

Cada procedimiento introduce sesgos distintos.

El consenso alcanzado después de conocer la salida de la IA no es una referencia independiente.

La adjudicación debe realizarse sin acceso a la predicción del modelo.

21.36. Diseño multilector-multicaso

Cuando varios expertos evalúan múltiples casos puede utilizarse un diseño multilector-multicaso.

Permite separar variabilidad debida a:

- lector;
- caso;
- interacción lector-caso;
- sistema.

Puede compararse:

- experto sin IA;
- experto con IA;
- IA sola;
- consenso.

Este diseño es especialmente útil cuando el objetivo es asistencia, no sustitución.

21.37. IA sola, humano solo y equipo humano-IA

La comparación completa debería incluir:

M_humano

M_IA

M_humano+IA

El equipo combinado no siempre supera a sus componentes.

Puede ocurrir:

- que el humano corrija errores de la IA;
- que la IA reduzca variabilidad humana;
- que el humano acepte errores por automatización;
- que la combinación aumente tiempo sin mejorar rendimiento.

La finalidad debe determinar el comparador relevante.

Si el sistema se utilizará como apoyo, la comparación principal debería ser:

procedimiento asistido frente a procedimiento no asistido

y no únicamente IA frente a experto.

21.38. Rendimiento humano variable

El rendimiento puede variar según:

- experiencia;
- turno;
- fatiga;
- velocidad;
- dificultad;
- prevalencia;
- presión productiva.

Una comparación en laboratorio con expertos descansados puede no representar la práctica industrial.

La evaluación debería reproducir:

- tiempo real;
- volumen;
- entorno;
- interrupciones;
- disponibilidad de información.

21.39. Prevalencia y comportamiento humano

Los expertos pueden modificar su umbral según la frecuencia esperada del defecto.

Cuando el evento es muy raro, pueden disminuir la vigilancia o adoptar un criterio más conservador.

La comparación debe utilizar la misma prevalencia o ajustar su interpretación.

Un estudio artificialmente equilibrado puede mejorar la atención humana y no reproducir el uso real.

21.40. Sesgo de memoria

Si los expertos han visto previamente las muestras, pueden recordarlas.

Debe evitarse:

- utilizar imágenes de formación en evaluación;

- repetir muestras en intervalos cortos;
- mostrar resultados anteriores;
- identificar lotes.

Las presentaciones repetidas deben separarse y aleatorizarse.

21.41. Comparación ciega

El experto no debería conocer:

- resultado de referencia;
- predicción de la IA;
- categoría real;
- procedencia del caso;

salvo que esa información esté disponible en la práctica habitual.

La IA tampoco debe recibir variables que el experto no posee si se pretende una comparación directa de capacidad.

Si dispone de información adicional, debe declararse que se comparan procedimientos, no únicamente agentes.

21.42. Tiempo de decisión

El rendimiento debe acompañarse de:

$$T_{humano}$$

$$T_{IA}$$

$$T_{asistido}$$

Una IA puede ser algo menos exacta pero permitir revisar todos los productos, mientras que una inspección humana solo cubre una muestra.

La utilidad depende de:

- exactitud;
- velocidad;
- cobertura;
- coste;
- consistencia;
- riesgo.

21.43. Cobertura humana e IA

La comparación por unidad puede ocultar diferencias de cobertura.

Ejemplo:

- humano: 99 % de sensibilidad sobre el 5 % de unidades inspeccionadas;
- IA: 95 % sobre el 100 % de unidades.

La capacidad global de detección será distinta.

Puede conceptualizarse:

$$[\ P(\text{detección total})$$

$$P(\text{unidad inspeccionada}) \times P(\text{detección}|\text{inspección})\ ]$$

La evaluación debe considerar el sistema completo.

21.44. Fatiga de alertas

Una IA con sensibilidad alta y especificidad baja puede generar muchas alertas.

El rendimiento del equipo humano-IA puede degradarse cuando los usuarios dejan de revisarlas cuidadosamente.

Debe estudiarse:

- número de alertas;
- tiempo;
- proporción real;
- tasa de revisión;
- errores tras periodos prolongados.

La sensibilidad teórica del modelo no equivale a la sensibilidad efectiva del proceso cuando las alertas no son gestionadas.

21.45. Automation bias

La asistencia puede reducir el rendimiento humano cuando la persona acepta la recomendación incorrecta.

Deben incluirse casos en los que:

- la IA acierta;
- la IA falla;
- el experto acierta;
- ambos fallan;
- existe incertidumbre.

La evaluación debe analizar:

$$P(\text{humano cambia correctamente})$$

y:

$$P(\text{humano cambia incorrectamente})$$

después de recibir la recomendación.

21.46. Explicaciones y rendimiento humano

Las explicaciones pueden mejorar comprensión, pero también generar confianza injustificada.

Debe validarse si una explicación:

- mejora detección;
- reduce tiempo;
- facilita anulación correcta;
- aumenta aceptación ciega;
- confunde al usuario.

Una explicación plausible no garantiza una mejor decisión.

21.47. Comparación con procedimientos convencionales

El comparador más relevante puede ser el proceso completo existente.

Ejemplos:

- plan de muestreo;

- liberación por laboratorio;
- vida útil fija;
- inspección manual;
- carta de control;
- regla de pH y actividad de agua;
- modelo microbiológico convencional.

La comparación debe realizarse sobre las mismas unidades o sobre periodos suficientemente comparables.

21.48. Modelo de IA frente a regla simple

Antes de aprobar una arquitectura compleja debe evaluarse un modelo base:

- regresión lineal;
- regresión logística;
- árbol sencillo;
- regla de límites;
- media histórica;
- peor caso;
- criterio experto.

NIST recomienda considerar métricas frente a referencias humanas u otros benchmarks estándar dentro de la evaluación de riesgo y desempeño.

El modelo complejo debe demostrar un beneficio suficiente para compensar:

- opacidad;
- mantenimiento;
- deriva;
- coste;
- necesidad de infraestructura;
- dificultad de auditoría.

21.49. Benchmark ingenuo

Para regresión puede utilizarse:

$$Ynaive=Ytrain$$

Para series temporales:

$$Yt=Yt-1$$

Para vida útil:

$$T=\text{vida útil fija histórica}$$

El modelo debe superar de manera relevante estas referencias.

Un (R^2) positivo ya implica una mejora respecto a la media en determinadas formulaciones, pero no demuestra utilidad suficiente.

21.50. Comparación con microbiología predictiva convencional

Un modelo de IA puede compararse con:

- Gompertz;

- Baranyi;
- modelos cardinales;
- Ratkowsky;
- modelos secundarios;
- herramientas predictivas establecidas.

La IA debe demostrar qué aporta:

- integración de más variables;
- mejor calibración;
- adaptación a la matriz;
- menor error;
- actualización dinámica;
- mayor robustez.

Si la mejora es marginal, el modelo convencional puede resultar preferible por su interpretabilidad y fundamento mecanístico.

21.51. Comparación con vida útil fija

Una empresa puede asignar una vida útil conservadora única a todos los lotes.

La IA puede ofrecer una estimación dinámica.

La comparación debe evaluar:

Seguridad

- sobreestimaciones;
- incumplimientos;
- cola de error.

Eficiencia

- días recuperados;
- desperdicio evitado;
- flexibilidad logística.

Consistencia

- variabilidad de decisiones;
- estabilidad entre campañas.

El objetivo no debe ser únicamente extender la duración media, sino hacerlo sin aumentar el riesgo.

21.52. Beneficio incremental

El beneficio incremental es la mejora aportada por la IA sobre la información ya disponible.

Puede expresarse como:

[Δ M M_modelo completo M_modelo base]

Pero también debe analizarse si mejora:

- decisiones;
- tiempos;
- costes;

- cobertura;
- seguridad;
- desperdicio.

Un aumento de AUC de 0,90 a 0,92 puede ser estadísticamente real y operativamente irrelevante.

21.53. Reclasificación

Puede analizarse cuántos casos son reclasificados correctamente e incorrectamente al añadir la IA.

Sea:

- (C_+): reclasificación correcta;
- (C_-): reclasificación incorrecta.

El beneficio neto no debe evaluarse únicamente por el número total, sino por la gravedad de los cambios.

Reclasificar correctamente un lote crítico puede tener mayor valor que reclasificar varios defectos menores.

21.54. Curvas de decisión

Las curvas de decisión evalúan el beneficio neto en distintos umbrales de intervención.

Una formulación habitual es:

[NB [VP] / [n] [FP] / [n] [p_t] / [1-p_t]]

donde (p_t) es el umbral de decisión.

El método compara estrategias como:

- intervenir en todos;
- intervenir en ninguno;
- utilizar el modelo.

Sin embargo, la relación implícita entre daños debe ser tecnológicamente defendible.

En seguridad alimentaria, no debe utilizarse para monetizar o compensar de manera simplista consecuencias sanitarias no tolerables.

21.55. Impacto sobre el muestreo

Un modelo puede utilizarse para enriquecer la muestra de análisis.

Debe compararse:

positivos por muestra analizada con IA

frente a:

positivos por muestra analizada sin IA

Pero también:

- positivos omitidos;
- cobertura de zonas no seleccionadas;
- diversidad de muestras;
- sesgo de selección.

Aumentar el rendimiento del muestreo no justifica eliminar por completo la exploración aleatoria.

21.56. Coste por caso verdadero detectado

Puede calcularse:

$$[C_{det} C_{sistema} + C_{confirmaciones} + C_{operación} N_{verdaderos detectados}]$$

Esta medida ayuda a evaluar viabilidad, pero no sustituye los criterios mínimos de seguridad.

Un sistema barato con falsos negativos inaceptables no es una alternativa válida.

21.57. Coste del falso positivo

Puede incluir:

- análisis;
- retención;
- mano de obra;
- retraso;
- pérdida de producto;
- ocupación de almacén;
- logística.

El coste esperado:

$$[C_{FP,total} N_{FP} \times C_{FP}]$$

Debe calcularse en prevalencia real y volumen anual, no únicamente en una muestra equilibrada.

21.58. Coste del falso negativo

Puede incluir:

- reclamación;
- devolución;
- retirada;
- destrucción;
- investigación;
- pérdida de cliente;
- sanción;
- daño sanitario;
- reputación.

No todos los componentes son adecuadamente monetizables.

Por ello, en aplicaciones críticas debe existir una restricción de seguridad previa al análisis económico.

21.59. Comparación con el rendimiento del sistema completo

No es suficiente comparar:

modelo frente a método

Debe compararse:

flujo asistido por IA frente a flujo convencional

El flujo incluye:

- captura;

- análisis;
- comunicación;
- interpretación;
- decisión;
- actuación;
- confirmación.

Una IA rápida puede no mejorar el tiempo total si las alertas esperan horas para revisión.

21.60. Comparación simultánea y secuencial

Comparación simultánea

Ambos procedimientos se aplican a la misma muestra.

Ventaja: comparabilidad directa.

Riesgo: uno puede influir en el otro.

Comparación secuencial

Se utilizan en periodos distintos.

Ventaja: reproduce la práctica.

Riesgo: cambios temporales.

Siempre que sea posible, debe utilizarse un diseño pareado y ciego para rendimiento, complementado con un estudio prospectivo de impacto.

21.61. Orden de ejecución

El orden puede afectar al resultado.

Ejemplo: una inspección destructiva altera la muestra antes de la lectura de IA.

Debe definirse:

- qué método se aplica primero;
- si el orden se aleatoriza;
- si existen muestras replicadas;
- si el procedimiento modifica la propiedad.

21.62. Cegamiento

El analista de referencia no debería conocer el resultado de IA.

El operador de IA no debería conocer la referencia.

El estadístico debería analizar una base bloqueada conforme al protocolo.

El cegamiento reduce el riesgo de:

- reinterpretación;
- repetición selectiva;
- exclusión;
- adjudicación favorable.

21.63. Muestras de control

La comparación debe incluir:

- controles positivos;
- controles negativos;
- materiales de referencia;
- muestras de frontera;
- interferentes;
- niveles altos y bajos;
- matrices representativas.

Los controles no sustituyen las muestras reales, pero permiten verificar estabilidad y detectar fallos.

21.64. Materiales de referencia

Cuando existan materiales certificados pueden utilizarse para evaluar:

- sesgo;
- exactitud;
- trazabilidad metrológica;
- estabilidad.

No siempre estarán disponibles para fenómenos complejos como vida útil, sensorial o contaminación heterogénea.

En esos casos deben utilizarse controles internos bien caracterizados.

21.65. Comparación interlaboratorio

Para un modelo que se comercializará ampliamente puede resultar necesaria una evaluación interlaboratorio o interplanta.

Debe comprobarse que:

- la preparación es reproducible;
- el equipo funciona de manera equivalente;
- las instrucciones son suficientes;
- el rendimiento no depende del desarrollador.

La lógica de los estudios colaborativos utilizados en validación analítica resulta especialmente pertinente para sistemas que pretenden ser transferibles.

21.66. Estudios por lotes de reactivos o consumibles

Cuando la entrada depende de:

- reactivos;
- placas;
- tiras;
- kits;
- medios;
- calibradores;

debe evaluarse la variación entre lotes.

El modelo puede haber aprendido señales específicas de un consumible determinado.

21.67. Cambios del método de referencia

Una actualización del método de referencia puede modificar la etiqueta.

Debe evaluarse:

- equivalencia entre versiones;
- cambio de sensibilidad;
- cambio de límites;
- cambio de interpretación;
- impacto sobre el modelo.

Un modelo validado frente a una versión anterior no queda automáticamente validado frente a la nueva.

21.68. Comparación con métodos reglamentarios

Cuando existe un método especificado legalmente o utilizado para control oficial, el modelo alternativo no debe presentarse como sustituto sin analizar el marco aplicable.

Puede utilizarse para:

- cribado;
- priorización;
- autocontrol;
- alerta;
- apoyo.

La confirmación oficial o contractual puede seguir exigiendo el método reconocido.

Codex mantiene clasificaciones y principios para métodos de referencia y métodos utilizables en control o resolución de controversias, reforzando la necesidad de distinguir entre un método alternativo operativo y el procedimiento que determina oficialmente la conformidad.

21.69. Error total admisible

La aceptación puede basarse en una combinación de sesgo e imprecisión.

Conceptualmente:

$$TE = |\text{sesgo}| + z\sigma$$

donde:

- (TE) es error total;
- (σ), desviación;
- (z), factor relacionado con la cobertura.

El error total admisible debe derivarse del uso.

En torno a un límite crítico, el error tolerable será menor que en una medición exploratoria.

21.70. Incertidumbre combinada

La incertidumbre del resultado integrado puede combinar:

- referencia;
- sensor;
- muestreo;
- modelo;

• proceso.

De forma simplificada, si fueran independientes:

$$[u_c \sqrt{(u_R^2 + u_S^2 + u_M^2 + u_P^2)}]$$

En la práctica pueden existir correlaciones y distribuciones no normales.

El objetivo no es siempre obtener una única cifra, sino evitar atribuir toda la incertidumbre a un solo componente.

21.71. Modelo más preciso que la referencia aparente

En ocasiones la IA puede parecer más reproducible que el método de referencia.

Esto puede ocurrir porque:

• el algoritmo suaviza ruido;

• la referencia tiene alta variabilidad;

• el modelo predice una media latente;

• existe fuga de datos;

• el modelo no responde a variaciones reales.

Una menor variabilidad no demuestra mayor exactitud.

Un sistema que produce siempre el mismo resultado puede ser muy preciso y completamente sesgado.

21.72. Regresión hacia la media

Los modelos pueden reducir valores extremos y producir predicciones próximas a la media.

Esto puede mejorar RMSE, pero degradar la detección de casos críticos.

Debe evaluarse la pendiente:

$$Y_R = \alpha + \beta Y$$

y el rendimiento en extremos.

Un sistema de vida útil que sobreestima duraciones cortas e infraestima las largas puede presentar un error medio moderado y fallar precisamente en los productos de mayor riesgo.

21.73. Discrepancia clínicamente o tecnológicamente relevante

No toda diferencia numérica cambia una decisión.

Debe definirse:

$$\delta_{relevante}$$

como la mínima diferencia tecnológicamente importante.

Ejemplos:

• 0,1 unidades de pH;

• 0,2 log UFC/g;

• un día de vida útil;

• 2 % de humedad;

• 1 mm de defecto.

El valor debe justificarse por:

• especificación;

• incertidumbre;

- proceso;
- consecuencia;
- capacidad de control.

21.74. Diseño de una comparación de no inferioridad

Un protocolo debería especificar:

1. comparador;
2. métrica principal;
3. margen;
4. nivel de confianza;
5. población;
6. tamaño;
7. análisis;
8. tratamiento de indeterminados;
9. subgrupos;
10. conclusión.

Ejemplo:

Se demostrará la no inferioridad del modelo frente a la inspección experta si el límite inferior unilateral del 95 % de la diferencia de sensibilidad es superior a −2 puntos porcentuales, manteniendo una tasa de falsos rechazos inferior al límite operativo.

21.75. Población por protocolo e intención de evaluar

En estudios de no inferioridad, las exclusiones pueden sesgar la conclusión.

Puede realizarse:

Análisis de todas las muestras incluidas

Conserva fallos operativos y datos incompletos según reglas predefinidas.

Análisis por protocolo

Incluye únicamente unidades que cumplen estrictamente el procedimiento.

Ambos pueden ser informativos.

Una conclusión de no inferioridad debería ser robusta a análisis razonables, no depender de excluir fallos incómodos.

21.76. Tratamiento de resultados indeterminados

Los indeterminados pueden:

- excluirse;
- considerarse error;
- resolverse mediante confirmación;
- analizarse por separado.

La decisión debe fijarse antes.

Excluirlos puede favorecer al modelo si aparecen principalmente en casos difíciles.

Debe informarse:

$$[\ R\_indeterminado\_indeterminadosN\ ]$$

21.77. Resultados no evaluables

Debe diferenciarse entre:

- muestra inválida;
- fallo del sensor;
- fallo del software;
- abstención correcta;
- entrada fuera de dominio;
- ausencia de referencia.

Cada categoría tiene una implicación distinta.

Los fallos operativos forman parte del rendimiento real del sistema, aunque no sean errores del clasificador.

21.78. Comparación de disponibilidad

Puede definirse:

$$A= T\_operativo\ T\_total$$

Un método con rendimiento elevado pero baja disponibilidad puede presentar menor utilidad que un método algo menos preciso y constantemente disponible.

Debe considerarse el procedimiento de respaldo.

21.79. Tiempo hasta resultado

La ventaja de la IA suele residir en la rapidez.

Debe medirse:

$$TAT= T\_resultado\text{-}T\_recepción$$

y compararse con:

- laboratorio;
- inspección;
- procedimiento histórico.

La reducción de tiempo debe traducirse en una acción útil.

Un resultado anticipado carece de valor si la empresa no puede intervenir antes.

21.80. Utilidad incremental en seguridad

La IA puede aportar valor aunque no sustituya la referencia si:

- detecta antes;
- aumenta cobertura;
- prioriza;
- reduce tiempo;
- identifica patrones;
- apoya investigación.

La evaluación debería estimar:

$$\Delta T\_detección$$

$$\Delta N\_casos\ identificados$$

$$\Delta C\_control$$

sin asumir que el beneficio autoriza a eliminar la confirmación.

21.81. Utilidad incremental en vida útil

Puede medirse mediante:

- reducción de sobreestimación;
- menor desperdicio;
- días de vida útil recuperados;
- adaptación por lote;
- mejor rotación;
- menos reclamaciones.

Debe comprobarse que el aumento de aprovechamiento no procede de asumir mayor riesgo.

21.82. Utilidad incremental en calidad

Puede incluir:

- cobertura del 100 %;
- menor variación entre inspectores;
- detección temprana;
- clasificación objetiva;
- reducción de reclamaciones;
- menor falso rechazo.

La comparación debe incorporar la actuación física y el resultado final, no solo la puntuación del algoritmo.

21.83. Matriz de comparación

Dimensión	IA	Referencia	Humano	Procedimiento habitual
Sensibilidad	—	—	—	—
Especificidad	—	—	—	—
Sesgo	—	—	—	—
Repetibilidad	—	—	—	—
Reproducibilidad	—	—	—	—
Tiempo	—	—	—	—
Cobertura	—	—	—	—
Indeterminados	—	—	—	—
Coste	—	—	—	—
Trazabilidad	—	—	—	—
Robustez	—	—	—	—
Impacto	—	—	—	—

La matriz obliga a comparar procedimientos completos y evita reducir la decisión a una sola métrica.

21.84. Ejemplo: visión artificial frente a inspectores

Diseño

- unidades independientes;
- múltiples inspectores;
- imágenes o producto real;
- evaluación ciega;
- referencia por panel adjudicador;
- condiciones de línea.

Comparaciones

$$S\_IA$$

$$S\_inspector$$

$$S\_inspector+IA$$

Métricas adicionales

- tiempo;
- fatiga;
- cobertura;
- falsos rechazos;
- consistencia;
- rendimiento por tipo y tamaño de defecto.

Conclusión posible

La IA supera la consistencia media de los inspectores en defectos frecuentes, pero presenta sensibilidad insuficiente para fragmentos transparentes inferiores al tamaño definido. Se aprueba como apoyo y rechazo automático para defectos validados, manteniendo inspección específica para la categoría excluida.

21.85. Ejemplo: modelo NIR frente a método químico

Diseño

- muestras pareadas;
- rango completo;
- varios lotes;
- diferentes instrumentos;
- duplicados;
- referencia caracterizada.

Análisis

- sesgo;
- Deming;
- Bland–Altman;
- repetibilidad;
- precisión intermedia;
- error cerca de especificación.

Conclusión posible

El modelo presenta concordancia suficiente para control rápido de proceso, pero los límites de acuerdo no permiten utilizarlo como único método para resolver lotes situados dentro de la zona de decisión próxima a la especificación. Esos casos requieren confirmación química.

21.86. Ejemplo: IA microbiológica frente a cultivo

Diseño

- positivos a diferentes niveles;
- matrices;
- cultivos y confirmación;
- evaluación de POD;
- resultados indeterminados;
- verificación interlaboratorio.

Conclusión posible

El sistema es apto para cribado, con sensibilidad conforme al criterio y reducción del tiempo de respuesta, pero los positivos y una muestra definida de negativos requieren confirmación por el método autorizado.

21.87. Ejemplo: modelo de vida útil frente a criterio fijo

Diseño

- lotes prospectivos;
- predicción bloqueada;
- fecha fija como comparador;
- resultado real microbiológico y sensorial.

Análisis

- MAE;
- sobreestimación;
- desperdicio;
- días recuperados;
- incumplimientos.

Conclusión posible

El modelo reduce el desperdicio respecto a la fecha fija sin aumentar la tasa de sobreestimación peligrosa dentro del dominio estudiado. Su salida se utiliza aplicando el límite inferior predictivo y el margen de seguridad establecido.

21.88. Criterios de aprobación frente a una referencia

El modelo debería aprobarse únicamente cuando:

1. la referencia está suficientemente caracterizada;
2. las muestras son pareadas y representativas;
3. los límites se fijaron previamente;
4. la concordancia cumple el uso previsto;
5. los casos próximos al límite cumplen;

6. los indeterminados están controlados;
7. la reproducibilidad es suficiente;
8. el beneficio incremental es real;
9. las discordancias graves han sido investigadas;
10. existe verificación local.

21.89. Criterios de aprobación frente a expertos

Debe demostrarse:

- rendimiento humano basal;
- variabilidad entre expertos;
- comparación ciega;
- condiciones realistas;
- beneficio de la asistencia;
- ausencia de degradación por automatización;
- capacidad de anulación;
- formación.

No basta con que la IA alcance la media humana si falla en categorías críticas que los especialistas detectan correctamente.

21.90. Criterios de aprobación frente al procedimiento vigente

Debe probarse:

seguridad no inferior

y, además, al menos una mejora relevante en:

- tiempo;
- cobertura;
- coste;
- consistencia;
- anticipación;
- desperdicio.

La innovación no constituye por sí misma una mejora.

21.91. Declaraciones inadecuadas

Inadecuada

La IA tiene una correlación del 99 % con el laboratorio.

Una correlación elevada no prueba acuerdo.

Inadecuada

No hubo diferencias significativas, por lo que ambos métodos son equivalentes.

La ausencia de significación no demuestra equivalencia.

Inadecuada

El algoritmo supera al ser humano.

No se define qué humanos, tarea, referencia ni condiciones.

Inadecuada

El sistema sustituye al método de referencia.

No se demuestra que el marco normativo o contractual permita esa sustitución.

Inadecuada

Todos los desacuerdos eran errores del laboratorio.

La resolución puede estar sesgada.

21.92. Declaraciones defendibles

La versión evaluada mostró una diferencia de sensibilidad de −0,8 puntos porcentuales respecto al método comparador, con un límite inferior unilateral superior al margen de no inferioridad preespecificado. La conclusión se limita a las matrices, concentraciones y condiciones estudiadas.

El sistema presentó un sesgo medio de 0,12 unidades respecto al método de referencia y límites de acuerdo compatibles con el control de proceso, pero no con la liberación de lotes próximos a la especificación.

La asistencia algorítmica aumentó la sensibilidad de los inspectores y redujo la variabilidad entre operadores, aunque incrementó el tiempo de revisión y produjo dependencia excesiva en los casos en que la recomendación era incorrecta.

21.93. La referencia como parte del sistema de incertidumbre

La validación científicamente rigurosa no debe preguntar únicamente:

¿Cuánto se parece la IA a la referencia?

Debe preguntar también:

- ¿qué mide realmente la referencia?;
- ¿con qué incertidumbre?;
- ¿es representativa la muestra?;
- ¿cómo se resuelven las discordancias?;
- ¿qué diferencias cambian la decisión?;
- ¿qué beneficio aporta la IA?

Esta perspectiva evita dos errores opuestos:

- considerar que la referencia es infalible;
- utilizar sus limitaciones como excusa para aceptar cualquier discordancia del modelo.

21.94. Modelo de informe comparativo

El informe debería contener:

Objetivo

Equivalencia, no inferioridad, superioridad o beneficio incremental.

Comparadores

Descripción completa de:

- IA;

- referencia;
- experto;
- procedimiento.

Población

- matrices;
- lotes;
- niveles;
- plantas;
- periodos.

Diseño

- pareado;
- ciego;
- prospectivo;
- multicéntrico;
- orden de aplicación.

Métricas

- sensibilidad;
- especificidad;
- acuerdo;
- sesgo;
- límites;
- reproducibilidad;
- tiempo;
- cobertura.

Hipótesis

- margen;
- intervalo;
- regla de decisión.

Discordancias

- número;
- tipo;
- revisión;
- adjudicación.

Resultados

- globales;
- subgrupos;
- frontera;
- condiciones adversas.

Impacto

- operativo;
- económico;
- sanitario;
- humano.

Conclusión

- uso autorizado;
- restricciones;
- necesidad de confirmación;
- verificación local.

21.95. Conclusión del bloque

Validar un modelo frente a un método, un experto o un procedimiento convencional no consiste en demostrar que ambos producen números parecidos.

Exige establecer:

- qué representa la referencia;
- cuál es su incertidumbre;
- qué diferencias son tolerables;
- cómo se comporta el modelo en torno a los límites;
- si mantiene la reproducibilidad;
- si es equivalente o no inferior;
- si aporta un beneficio real;
- si las discordancias pueden controlarse.

Un modelo puede superar al promedio humano y seguir siendo inadecuado para una decisión crítica. Puede mostrar elevada correlación con un método químico y presentar un sesgo incompatible con la especificación. Puede reducir el tiempo de análisis y aumentar simultáneamente los falsos negativos.

La conclusión científicamente relevante no es:

> *«La IA funciona igual o mejor que el método tradicional.»*

Sino:

> *«La versión evaluada demuestra, dentro de las matrices, rangos y condiciones especificadas, un nivel de concordancia, reproducibilidad y utilidad incremental compatible con la función autorizada, manteniéndose los métodos confirmatorios y las restricciones definidos para los casos en los que la incertidumbre o las consecuencias del error no permiten su sustitución.»*

La referencia no debe utilizarse como un adorno de validación. Debe constituir el fundamento mediante el cual se demuestra qué puede medir, predecir o decidir legítimamente el sistema y qué debe continuar bajo control humano, analítico o experimental.

22. Interpretación, comunicación y transparencia de los resultados de validación

La validación científica no finaliza cuando se calculan las métricas. Finaliza cuando los resultados se comunican de manera suficientemente completa para que otra persona competente pueda interpretar qué demuestra el estudio, qué no demuestra y bajo qué condiciones puede utilizarse el modelo.

Esta exigencia es especialmente importante en inteligencia artificial porque una misma cifra puede producir impresiones radicalmente diferentes según cómo se presente.

Una afirmación como:

«El modelo alcanza una precisión del 98 %.»

puede inducir al lector a pensar que:

- el sistema acierta el 98 % de todos los casos futuros;
- solo se equivoca en el 2 %;
- la probabilidad de que una alerta sea correcta es del 98 %;
- el rendimiento será idéntico en cualquier producto o planta;
- la precisión se mantendrá indefinidamente;
- el algoritmo puede sustituir al método de referencia.

Ninguna de estas conclusiones se deriva necesariamente de una accuracy del 98 %.

La transparencia exige comunicar, junto con la métrica:

- qué resultado se predijo;
- qué clase se consideró positiva;
- cuál era la prevalencia;
- sobre cuántas unidades independientes se calculó;
- de qué plantas, periodos y productos procedían;
- qué versión del modelo se utilizó;
- cuál fue el umbral;
- qué intervalos de confianza se obtuvieron;
- cuántos falsos positivos y falsos negativos aparecieron;
- qué subgrupos presentaron un rendimiento inferior;
- qué limitaciones permanecen.

La comunicación no es una fase meramente editorial. Forma parte de la gestión del riesgo. Una métrica científicamente correcta puede convertirse en una información peligrosa si se presenta sin contexto o se utiliza para justificar una decisión que el estudio no evaluó.

22.1. Principio de correspondencia entre evidencia y afirmación

La amplitud de una afirmación nunca debe exceder la amplitud de la evidencia.

Puede expresarse conceptualmente como:

$$A_{permitida} \leq E_{demostrada}$$

donde:

- ($A_{permitida}$) es el alcance de la afirmación;
- ($E_{demostrada}$) es el alcance real de la validación.

Si el modelo se evaluó en:

- dos productos;
- una planta;
- un equipo;
- una campaña;
- condiciones controladas;

no debería afirmarse que funciona para:

- todos los productos de la categoría;
- cualquier instalación;
- cualquier sensor;
- cualquier estación;
- cualquier condición logística.

La conclusión debe permanecer dentro del dominio demostrado.

22.2. Diferencia entre informar una métrica e interpretar su significado

La presentación científica debe separar:

Resultado numérico

Sensibilidad=97,4%

Incertidumbre

IC 95 %=94,1-99,0%

Contexto

- 310 casos positivos independientes;
- tres plantas;
- dos campañas;
- umbral congelado;
- productos incluidos en el dominio.

Interpretación

El sistema detectó aproximadamente 97 de cada 100 casos positivos en la población evaluada. El intervalo de confianza indica que el rendimiento real compatible con el estudio podría ser inferior al valor puntual. La conclusión no se extiende a productos o condiciones no incluidos.

La interpretación evita que el lector confunda un estimador muestral con una propiedad exacta e invariable.

22.3. Unidad de comunicación

El informe debe identificar la unidad estadística real.

No es equivalente declarar:

> *«El modelo fue validado con 100.000 imágenes.»*

que:

> *«El modelo fue evaluado sobre 800 unidades de producto independientes, de las que se obtuvieron 100.000 imágenes.»*

La primera formulación puede sugerir un tamaño de evidencia mucho mayor del real.

La comunicación debería indicar simultáneamente:

- número de registros digitales;
- número de muestras;
- número de lotes;
- número de campañas;
- número de plantas;
- número de positivos reales.

En modelos longitudinales también debe informarse:

- número de lotes;
- número de puntos temporales;
- número de curvas completas;
- número de eventos observados;
- número de resultados censurados.

22.4. Presentación obligatoria de la matriz de confusión

En clasificación binaria debería publicarse la matriz completa:

Resultado real	Predicción positiva	Predicción negativa
Positivo	VP	FN
Negativo	FP	VN

Las métricas porcentuales deben acompañarse de números absolutos.

No es equivalente:

$$FNR=1\%$$

obtenido con un falso negativo entre cien positivos, que con cien falsos negativos entre diez mil.

La matriz permite al lector reconstruir:

- sensibilidad;
- especificidad;
- accuracy;
- valores predictivos;
- prevalencia;
- volumen real de errores.

Cuando se omiten las celdas absolutas, resulta difícil valorar la estabilidad estadística y el impacto operacional.

22.5. Declaración inequívoca de la clase positiva

El informe debe indicar qué se consideró positivo.

Ejemplos:

- lote microbiológicamente no conforme;
- cuerpo extraño presente;
- vida útil insuficiente;
- riesgo superior al umbral;

• defecto crítico;

• adulteración detectada.

La etiqueta positiva no debe dejarse implícita.

Si la clase positiva representa el producto conforme, la sensibilidad tendrá un significado distinto del habitual en seguridad.

Una recomendación es utilizar terminología operacional junto con la estadística:

Sensibilidad para detectar lotes no conformes.

Especificidad para reconocer lotes conformes.

Esta redacción reduce la ambigüedad.

22.6. Comunicación de la prevalencia

La prevalencia debe informarse porque condiciona:

• valores predictivos;

• volumen de alertas;

• utilidad operativa;

• calibración.

Debe diferenciarse:

Prevalencia del conjunto de validación

$$[\ \pi_{test}$$

$$[VP+FN] / [N]\ ]$$

Prevalencia esperada en producción

$$\pi_{operativa}$$

Si el conjunto fue enriquecido con positivos, debe declararse expresamente.

Ejemplo:

La muestra de validación contenía un 20 % de positivos para permitir una estimación precisa de la sensibilidad. La prevalencia operativa estimada es del 0,7 %. Los valores predictivos se recalcularon para este escenario.

Ocultar el enriquecimiento puede inducir a interpretar un valor predictivo positivo que no será reproducible en producción.

22.7. Escenarios de prevalencia

Cuando la prevalencia es variable, conviene presentar varios escenarios.

Prevalencia	VPP estimado	VPN estimado	Alertas por 10.000 unidades
0,1 %	—	—	—
0,5 %	—	—	—
1,0 %	—	—	—
5,0 %	—	—	—

Esto permite anticipar cómo cambiará la utilidad del sistema en:

• operación normal;

• campaña de mayor riesgo;

- incidente;
- población seleccionada.

22.8. Intervalos de confianza

Toda métrica principal debería acompañarse de un intervalo.

La presentación recomendada es:

[Sensibilidad

97,4% (IC 95 %:94,1-99,0%)]

y no únicamente:

Sensibilidad: 97,4 %.

En modelos con datos agrupados debe indicarse si el intervalo considera:

- lotes;
- plantas;
- proveedores;
- curvas;
- repeticiones.

Un intervalo calculado suponiendo independencia entre miles de imágenes puede ser artificialmente estrecho si proceden de pocos productos.

22.9. Redondeo y falsa precisión

La presentación de demasiados decimales puede transmitir una exactitud inexistente.

Ejemplo inadecuado:

Sensibilidad=97,4368%

si el estudio contiene pocos positivos.

Una presentación más razonable sería:

97,4%

o incluso:

97%

dependiendo del tamaño y del uso.

La precisión decimal debe ser coherente con:

- incertidumbre;
- método;
- tamaño muestral;
- relevancia industrial.

Mostrar probabilidades como 0,873642 puede inducir a creer que el sistema distingue diferencias que no están científicamente demostradas.

22.10. Intervalo de confianza frente a intervalo de predicción

Debe evitarse confundir ambos conceptos.

Intervalo de confianza

Informa sobre la incertidumbre de un parámetro o promedio.

Intervalo de predicción

Informa sobre la incertidumbre de una observación futura individual.

En vida útil:

Vida útil media predicha: 35 días; IC del promedio: 34–36 días.

no significa que cada lote dure entre 34 y 36 días.

El intervalo individual podría ser:

Intervalo predictivo: 27–42 días.

Para adoptar decisiones sobre lotes individuales, el segundo suele ser más relevante.

22.11. Comunicación de modelos de regresión

Un informe de regresión debería incluir, como mínimo:

- MAE;
- RMSE;
- sesgo medio;
- mediana del error;
- percentiles superiores;
- máximo error relevante;
- (R^2), como medida secundaria;
- análisis de residuos;
- error por rangos;
- intervalos predictivos;
- rendimiento externo.

La afirmación:

> *«El modelo explica el 95 % de la variabilidad.»*

no informa de cuántos días sobreestima la vida útil ni de cuántos lotes clasifica peligrosamente.

La comunicación debe traducir las métricas a unidades industriales:

El error absoluto medio fue de 2,3 días. En el 95 % de los lotes el error absoluto fue inferior a 6,1 días. El modelo sobreestimó la vida útil real en más de dos días en el 4,8 % de los lotes.

22.12. Comunicación asimétrica del error

Cuando una dirección es más peligrosa, debe informarse por separado.

En vida útil:

- sobreestimación;
- infraestimación.

En microbiología:

- infrapredicción del crecimiento;
- sobrepredicción.

En inactivación:

- sobreestimación de la reducción;
- infraestimación.

La media absoluta no refleja esta asimetría.

Ejemplo:

Tipo de error	Frecuencia	Magnitud media	Máximo
Sobreestimación de vida útil	—	—	—
Infraestimación	—	—	—
Sobreestimación >2 días	—	—	—
Sobreestimación >5 días	—	—	—

22.13. Comunicación de calibración

Si la salida se presenta como probabilidad, el informe debe incluir:

- curva de calibración;
- intercepto;
- pendiente;
- puntuación de Brier;
- calibración por subgrupos;
- prevalencia;
- procedimiento de recalibración.

No debería afirmarse:

«Riesgo del 80 %.»

si el modelo no ha demostrado que los casos con esa salida presentan aproximadamente esa frecuencia real.

Cuando la salida no está calibrada debe denominarse:

- puntuación;
- índice;
- score;
- nivel relativo.

No probabilidad.

22.14. Curvas ROC

La curva ROC debe acompañarse de:

- AUC;
- intervalo;
- umbral operativo;
- sensibilidad y especificidad en ese umbral;
- número de casos;
- comparación con el modelo de referencia.

El AUC no debe convertirse en una conclusión autosuficiente.

Una representación útil debe destacar el punto de operación real.

22.15. Curvas precision–recall

En eventos raros conviene informar:

- prevalencia;
- precision;
- recall;
- área bajo la curva;
- punto operativo;
- número de falsas alarmas.

La línea de base depende de la prevalencia.

Una curva obtenida sobre una muestra equilibrada no reproduce el rendimiento de una población donde los positivos representan el 0,1 %.

22.16. Gráficos de calibración

Los gráficos deberían mostrar:

- predicción media;
- frecuencia observada;
- tamaño por intervalo;
- intervalos de incertidumbre;
- diagonal ideal.

Los grupos con pocas observaciones deben identificarse porque su frecuencia observada será inestable.

No debe utilizarse un número excesivo de intervalos para producir una apariencia detallada sin soporte estadístico.

22.17. Gráficos de residuos

En regresión deben incluirse:

- residuo frente a valor predicho;
- residuo frente a valor real;
- residuo frente a tiempo;
- residuo por producto;
- residuo por planta;
- distribución de errores.

Estos gráficos permiten detectar:

- sesgo;
- heterocedasticidad;
- saturación;
- deriva;
- subgrupos.

La comunicación no debería limitarse a una línea de predicción frente a observación con elevada correlación.

22.18. Bland–Altman y límites tecnológicos

El gráfico de Bland–Altman debe acompañarse de:

- sesgo;
- límites de acuerdo;
- intervalo de los límites;
- límites tecnológicos aceptables.

Deben diferenciarse visualmente:

límites estadísticos

y:

límites de aceptación

Un acuerdo estadísticamente descrito puede ser tecnológicamente inaceptable.

22.19. Rendimiento por subgrupos

El informe debe incluir los subgrupos preespecificados.

Subgrupo	N	Positivos	Sensibilidad	Especificidad	IC	Estado
Producto A	—	—	—	—	—	Cumple/no cumple
Producto B	—	—	—	—	—	—
Planta 1	—	—	—	—	—	—
Planta 2	—	—	—	—	—	—
Proveedor X	—	—	—	—	—	—

Debe evitarse informar solo el promedio global.

22.20. Subgrupos con evidencia insuficiente

Cuando un grupo contiene pocos casos debe declararse:

La muestra fue insuficiente para establecer el rendimiento con precisión.

No:

El modelo funciona correctamente porque no se observaron errores.

La ausencia de fallos en cinco casos no demuestra una tasa baja de error.

Puede utilizarse una categoría:

- cumple;
- no cumple;
- evidencia insuficiente;
- fuera del dominio.

22.21. Comunicación de heterogeneidad

En estudios multicéntricos debe informarse:

- rendimiento por centro;
- estimación agregada;
- heterogeneidad;
- posibles causas;
- necesidad de calibración local.

Una gráfica de bosque puede ser más informativa que una única cifra.

La conclusión debe evitar expresiones como:

«Rendimiento consistente.»

si existe una planta con sensibilidad notablemente inferior.

22.22. Comunicación de resultados negativos

Los resultados desfavorables forman parte del valor científico del estudio.

Deben informarse:

- incumplimientos;
- subgrupos débiles;
- fallos de robustez;
- condiciones no evaluables;
- tasas de abstención;
- errores extremos;
- pérdida de calibración.

La transparencia no exige convertir el informe en una enumeración de defectos, pero sí evitar una selección unilateral de resultados positivos.

22.23. Comunicación de casos críticos

Los falsos negativos graves deberían describirse de forma específica.

Para cada caso puede registrarse:

- producto;
- condición;
- nivel real;
- predicción;
- confianza;
- distancia al dominio;
- causa probable;
- barreras;
- consecuencia;
- acción.

No siempre será apropiado publicar datos identificativos, pero el informe interno debe permitir comprender el modo de fallo.

22.24. Taxonomía de errores

Una clasificación útil puede incluir:

- caso próximo al límite;
- entrada degradada;
- condición fuera de dominio;
- error de sensor;
- falta de variable;
- combinación no representada;

- cambio de proceso;
- error de referencia;
- fallo algorítmico;
- error de uso.

La distribución de causas ayuda a decidir si procede:

- ampliar datos;
- mejorar sensor;
- restringir dominio;
- recalibrar;
- modificar procedimiento.

22.25. Afirmaciones sobre ausencia de errores

La expresión:

«No se observaron falsos negativos.»

debe acompañarse del número de positivos y del intervalo.

Ejemplo:

No se observaron falsos negativos entre 120 positivos independientes. La ausencia de errores observados no demuestra una tasa real nula; el límite superior aproximado del 95 % debe interpretarse conforme al tamaño muestral.

Nunca debería traducirse en:

«El modelo elimina los falsos negativos.»

22.26. Afirmaciones sobre el 100 % de sensibilidad

Una sensibilidad observada del 100 % no significa sensibilidad real del 100 %.

Debe informarse:

$$S=100\%$$

junto con:

$$\text{IC } 95\ \%$$

La amplitud dependerá del número de positivos.

En una comunicación comercial puede formularse:

Detectó todos los positivos presentes en la muestra de validación, compuesta por [número] casos independientes.

No:

Detección garantizada.

22.27. Model cards y fichas de modelo

Una ficha de modelo puede resumir:

Identificación

- nombre;
- versión;
- propietario;

- fecha.

Finalidad

- uso autorizado;
- usuario;
- decisión.

Datos

- origen;
- periodo;
- productos;
- centros;
- prevalencia.

Rendimiento

- métricas;
- intervalos;
- umbral;
- subgrupos.

Limitaciones

- productos excluidos;
- rangos;
- condiciones no probadas;
- errores conocidos.

Operación

- abstención;
- revisión;
- respaldo;
- monitorización.

Cambios

- historial;
- revalidación.

La ficha no sustituye el dossier completo, pero facilita una utilización informada.

22.28. Ficha de datos

Una documentación específica de datos debería incluir:

- finalidad de recopilación;
- origen;
- población;
- unidad estadística;
- transformaciones;
- exclusiones;

- calidad;
- distribución;
- representatividad;
- carencias;
- restricciones.

Esto permite evaluar si el modelo se utiliza en poblaciones compatibles.

22.29. Resumen ejecutivo

El resumen ejecutivo debe ser comprensible para dirección, pero no simplificar hasta el punto de ocultar riesgos.

Debería responder:

1. ¿Qué hace?
2. ¿Para qué decisión?
3. ¿En qué productos?
4. ¿Qué rendimiento demostró?
5. ¿Qué error es crítico?
6. ¿Qué limitaciones existen?
7. ¿Qué controles se mantienen?
8. ¿Cuál es la decisión recomendada?

Ejemplo:

El modelo es apto como sistema de cribado para las tres familias de producto evaluadas. Mantiene una sensibilidad externa del 98,1 %, con límite inferior del 95 % superior al criterio preestablecido. No se autoriza su uso para liberar lotes sin confirmación. Los productos con actividad de agua inferior a 0,93 quedan fuera del dominio.

22.30. Informe técnico

El informe técnico debe permitir reproducir la evaluación.

Debe contener:

- protocolo;
- datos;
- partición;
- código o procedimiento;
- preprocesamiento;
- métricas;
- intervalos;
- análisis;
- desviaciones;
- conclusiones.

No debe limitarse a diapositivas o capturas de pantalla.

22.31. Comunicación a operarios

El usuario operacional no necesita conocer toda la arquitectura, pero sí:

- finalidad;

- significado de la salida;
- acción requerida;
- condiciones de abstención;
- límites;
- procedimiento alternativo.

La interfaz y el procedimiento deben utilizar un lenguaje estable.

No deben mezclarse términos como:

- riesgo;
- confianza;
- probabilidad;
- precisión;
- conformidad;

si representan conceptos diferentes.

22.32. Comunicación a auditores

El auditor necesita comprobar:

- trazabilidad;
- autorización;
- validación;
- integración;
- monitorización;
- cambios;
- incidentes.

El paquete debería permitir vincular:

predicción arrow versión arrow evidencia de validación

Una cifra comercial del proveedor no constituye evidencia suficiente.

22.33. Comunicación a clientes

Cuando el modelo se utiliza para respaldar una especificación o una declaración de vida útil, debe evitarse atribuirle una certeza superior a la evidencia.

Una redacción defendible sería:

La estimación se ha obtenido mediante un modelo validado para la categoría y condiciones especificadas y se integra con resultados analíticos y estudios de estabilidad.

No:

La IA garantiza la vida útil.

22.34. Comunicación a autoridades

La organización debería poder explicar:

- qué función cumple el modelo;
- si sustituye o complementa métodos;
- cómo se validó;

- qué controles se mantienen;
- cómo se gestionan errores;
- cómo se conserva trazabilidad.

La utilización de una tecnología compleja no reduce la obligación de ofrecer una explicación verificable.

22.35. Comunicación comercial responsable

El marketing de una herramienta de IA debe mantener una separación clara entre:

- capacidad demostrada;
- beneficio esperado;
- hipótesis;
- aspiración comercial.

Expresiones de riesgo incluyen:

- «100 % fiable»;
- «sin falsos negativos»;
- «garantiza la seguridad»;
- «elimina el laboratorio»;
- «válido para cualquier alimento»;
- «predice exactamente la vida útil»;
- «autónomo y sin supervisión».

Estas afirmaciones rara vez pueden sostenerse científicamente.

22.36. Accuracy como reclamo comercial

Una frase como:

«99 % de accuracy.»

debería considerarse incompleta si no declara:

- prevalencia;
- clases;
- unidad;
- conjunto;
- independencia;
- intervalo;
- sensibilidad;
- falsos negativos.

En problemas desequilibrados, la accuracy puede ser una de las métricas menos informativas.

22.37. Afirmaciones de superioridad

Una afirmación como:

«Superior a los expertos.»

requiere:

- comparador definido;
- muestra;

- cegamiento;
- métrica;
- intervalo;
- condiciones;
- significación y relevancia;
- análisis humano-IA.

No debería basarse en comparar la mejor configuración de IA con el promedio no controlado de operadores.

22.38. Afirmaciones de sustitución

Para afirmar que la IA sustituye un método debe demostrarse:

- equivalencia o no inferioridad;
- marco normativo compatible;
- funcionamiento local;
- reproducibilidad;
- control de casos frontera;
- contingencia;
- trazabilidad.

Una herramienta puede ser muy útil como cribado y no estar autorizada ni científicamente preparada para reemplazar la referencia.

22.39. Afirmaciones de generalización

La afirmación:

«Aplicable a todos los productos cárnicos.»

debe estar respaldada por una evidencia proporcional.

Una formulación correcta podría ser:

Evaluado en las familias A, B y C, dentro de los rangos de formulación y proceso especificados.

La generalización debe formularse por dominio, no por categoría comercial amplia.

22.40. Afirmaciones sobre ahorro

El ahorro debe calcularse considerando:

- licencias;
- sensores;
- integración;
- mantenimiento;
- confirmaciones;
- falsos positivos;
- formación;
- validación;
- vigilancia;
- incidentes.

No debe comunicarse únicamente la reducción teórica de análisis.

Puede expresarse:

$$[\ A_{neto}$$

$$A_{evitado}$$

$$C_{implantación}$$

$$C_{operación}$$

$$C_{errores}\ ]$$

El análisis económico debe realizarse bajo prevalencia real.

22.41. Afirmaciones sobre reducción del desperdicio

Una reducción de desperdicio debe demostrar que no se consiguió aumentando sobreestimaciones peligrosas de vida útil.

Deben presentarse simultáneamente:

- producto recuperado;
- días adicionales;
- errores;
- incumplimientos;
- reclamaciones;
- margen de seguridad.

22.42. Afirmaciones sobre seguridad

La seguridad no puede reducirse a una métrica predictiva.

La afirmación debe considerar:

$$\text{modelo} + \text{barreras} + \text{procedimiento} + \text{verificación}$$

El sistema puede mejorar la seguridad, pero no «garantizarla» de forma aislada.

22.43. Comunicación de limitaciones

Las limitaciones deben ser concretas.

Limitación vaga

El modelo puede no funcionar en todos los casos.

Limitación útil

No se ha validado en productos con pH superior a 6,0, actividad de agua inferior a 0,92, envases de vidrio ni cámaras distintas del modelo especificado.

Las limitaciones deben poder traducirse a controles del sistema.

22.44. Limitación metodológica

Deben informarse aspectos como:

- validación retrospectiva;
- pocos positivos;
- una sola planta;
- ausencia de campaña completa;

- referencia imperfecta;
- subgrupo pequeño;
- falta de impacto prospectivo;
- incertidumbre de calibración.

Estas limitaciones condicionan el nivel de afirmación.

22.45. Limitación operacional

Debe declararse:

- necesidad de conexión;
- tiempo de respuesta;
- datos obligatorios;
- capacidad de confirmación;
- intervención humana;
- tasa de abstención;
- equipo autorizado.

22.46. Limitaciones conocidas frente a limitaciones desconocidas

Las limitaciones conocidas pueden controlarse.

El riesgo más difícil procede de condiciones no anticipadas.

Por ello, además de declarar exclusiones, debe existir:

- detección fuera de dominio;
- vigilancia;
- muestreo independiente;
- gestión de incidentes.

La transparencia documental no sustituye el control técnico.

22.47. Comunicación de cambios de versión

Cuando el modelo cambia, las métricas de la versión anterior no deben aplicarse automáticamente.

Toda comunicación debe vincular el resultado a:

- versión;
- fecha;
- configuración;
- umbral;
- entorno.

Ejemplo:

Los resultados corresponden a la versión 3.2, umbral 0,37, configurada para el sensor X.

No:

Nuestra IA tiene una sensibilidad del 98 %.

22.48. Historial de versiones

Puede utilizarse una tabla:

Versión	Fecha	Cambio	Validación	Estado
1.0	—	Inicial	Interna	Retirada
2.0	—	Nuevos datos	Externa	Sustituida
3.2	—	Recalibración	Local y temporal	Activa

Esto evita confundir resultados históricos con el sistema operativo actual.

22.49. Comunicación de revalidación

La revalidación debe especificar:

- motivo;
- elementos cambiados;
- alcance;
- datos;
- resultados;
- decisión.

Una revalidación parcial no debe presentarse como si hubiese repetido toda la evidencia inicial.

22.50. Paneles de monitorización

El cuadro operativo debe evitar una simplificación excesiva.

Puede incluir:

- versión;
- estado;
- sensibilidad acumulada;
- falsos negativos;
- tasa de alertas;
- calibración;
- abstenciones;
- datos fuera de dominio;
- deriva;
- incidencias.

Debe mostrar:

- valor actual;
- referencia;
- límite;
- tendencia;
- acción.

22.51. Semáforos

Los colores pueden facilitar la gestión, pero deben estar vinculados a reglas.

Verde

Dentro de criterios.

Amarillo

Desviación que requiere investigación.

Naranja

Acción correctiva o restricción.

Rojo

Suspensión.

El color no debe sustituir el valor ni la interpretación.

22.52. Riesgo de indicadores acumulados

Una métrica acumulada desde el despliegue puede ocultar una degradación reciente.

Ejemplo:

- sensibilidad histórica: 98 %;
- sensibilidad últimos 30 días: 89 %.

El cuadro debe mostrar ventanas:

- corta;
- media;
- larga.

22.53. Comunicación de la falta de etiquetas

Cuando el rendimiento no puede calcularse todavía porque los resultados reales llegan con retraso, debe indicarse.

No debería mostrarse una métrica histórica como si fuese actual.

Puede informarse:

Rendimiento confirmado hasta la fecha X. Los resultados posteriores están pendientes de referencia.

22.54. Datos fuera de dominio

El panel debería mostrar:

[R_OOD_fuera de dominioN]

y su distribución por producto, planta y proveedor.

Un aumento puede anticipar una pérdida de validez aunque las métricas etiquetadas todavía no estén disponibles.

22.55. Comunicación de incertidumbre a usuarios

La incertidumbre debe mostrarse de manera comprensible.

Para una predicción continua:

Vida útil estimada: 30 días. Intervalo predictivo: 24–35 días. Fecha operativa autorizada: 22 días tras aplicar el margen de seguridad.

Para clasificación:

Riesgo alto. La observación está dentro del dominio. Requiere retención y confirmación.

Para fuera de dominio:

Predicción no válida. Producto no representado. Aplicar procedimiento alternativo.

22.56. Evitar etiquetas absolutas

En lugar de:

- seguro;
- contaminado;
- garantizado;
- libre de riesgo;

pueden utilizarse:

- cumple criterio del modelo;
- alerta de riesgo;
- requiere confirmación;
- predicción compatible;
- no evaluable.

No obstante, el lenguaje tampoco debe ser tan ambiguo que impida actuar.

22.57. Acción junto a resultado

Toda salida operacional debería incorporar:

$$\text{resultado} + \text{acción}$$

Ejemplo:

Resultado: indeterminado. Acción: retener el lote y repetir la medida.

No basta con mostrar una probabilidad y dejar que cada usuario decida qué significa.

22.58. Transparencia y secreto industrial

La transparencia no exige revelar necesariamente:

- código fuente;
- arquitectura propietaria completa;
- parámetros comerciales sensibles.

Pero sí debe permitir evaluar:

- finalidad;
- datos;
- rendimiento;
- limitaciones;
- versión;
- cambios;
- controles;
- incidentes.

El secreto industrial no debería utilizarse para impedir que el operador conozca si el sistema es adecuado para su uso.

22.59. Transparencia proporcional

El nivel de información puede variar según el destinatario.

Operario

Instrucción clara y límites.

Dirección técnica

Métricas, incertidumbre, riesgos y controles.

Auditor

Evidencia, trazabilidad y procedimientos.

Cliente

Capacidad demostrada y restricciones.

Autoridad

Fundamento, integración y control.

La información puede adaptarse, pero no contradecirse.

22.60. Transparencia interna

La organización debe evitar que únicamente el equipo de datos comprenda el sistema.

Calidad, producción, laboratorio y dirección deben conocer:

- función;
- límites;
- error;
- contingencia;
- responsabilidades.

Un modelo opaco dentro de la propia empresa constituye un riesgo de gobernanza.

22.61. Registro de decisiones humanas

Cuando una persona anula o acepta una recomendación debe conservarse:

- decisión;
- motivo;
- información disponible;
- resultado posterior.

Esto permite estudiar:

- utilidad real;
- automation bias;
- correcciones humanas;
- errores de interfaz;
- necesidad de formación.

22.62. Comunicación de anulaciones

No debe interpretarse toda anulación como error humano.

Puede revelar:

- fallo del modelo;

- dato contextual no capturado;
- condición fuera de dominio;
- conocimiento del proceso.

El análisis debe distinguir anulaciones correctas e incorrectas.

22.63. Informes de incidentes

El informe de incidente debería describir:

- predicción;
- resultado real;
- versión;
- entradas;
- contexto;
- causa;
- impacto;
- lotes afectados;
- contención;
- CAPA;
- revalidación.

La redacción debe evitar atribuir el fallo genéricamente a «la IA» si la causa fue:

- dato;
- sensor;
- usuario;
- integración;
- proceso.

Pero tampoco debe trasladar automáticamente la responsabilidad al usuario cuando el diseño favoreció el error.

22.64. Comunicación de retirada o suspensión

Cuando se suspende una versión debe informarse:

- motivo;
- alcance;
- fecha;
- productos afectados;
- procedimiento alternativo;
- condición de reactivación.

La versión no debe continuar disponible silenciosamente en otros entornos.

22.65. Publicación científica

Un artículo debería describir:

- finalidad;
- datos;
- participantes o unidades;
- referencia;

- partición;
- preprocesamiento;
- métricas;
- intervalos;
- calibración;
- subgrupos;
- resultados completos;
- limitaciones;
- disponibilidad de código o datos cuando proceda.

Debe diferenciar:

- análisis exploratorio;
- validación interna;
- validación externa;
- evaluación prospectiva;
- impacto.

22.66. Comparación con literatura

Las comparaciones deben reconocer diferencias en:

- población;
- prevalencia;
- producto;
- dificultad;
- métrica;
- unidad;
- diseño.

No debería afirmarse:

«Supera el estado del arte.»

basándose únicamente en una accuracy superior obtenida en un conjunto diferente.

22.67. Abstract científico

El resumen debería incluir resultados completos.

Ejemplo:

En validación externa temporal sobre 1.250 lotes independientes, incluidos 180 positivos, el modelo alcanzó una sensibilidad del 96,7 % y una especificidad del 89,5 %. El límite inferior del intervalo de sensibilidad fue inferior al criterio preespecificado en una de las cuatro familias de producto, por lo que la autorización se restringió a las tres restantes.

Este resumen comunica tanto el resultado favorable como la limitación decisiva.

22.68. Evitar lenguaje causal

Un modelo predictivo no demuestra necesariamente causalidad.

No debería afirmarse:

La variable X causa el deterioro porque fue la más importante en SHAP.

La formulación correcta sería:

La variable X contribuyó de forma relevante a las predicciones del modelo en la población estudiada.

La interpretación causal requiere evidencia adicional.

22.69. Separar rendimiento y explicación

Debe informarse por separado:

Rendimiento

Cuánto acierta.

Calibración

Qué significan sus probabilidades.

Explicación

Qué variables utiliza.

Impacto

Qué mejora produce.

Una buena explicación no compensa bajo rendimiento. Una elevada accuracy no demuestra impacto. Una mejora operativa no demuestra causalidad.

22.70. Transparencia sobre selección del modelo

Debe declararse:

- número de algoritmos evaluados;
- criterios de selección;
- ajuste de hiperparámetros;
- conjunto utilizado;
- análisis final independiente.

Esto permite valorar el riesgo de selección optimista.

22.71. Transparencia sobre exclusiones

Toda exclusión debe informarse:

- número;
- causa;
- momento;
- distribución;
- impacto.

Debe diferenciarse:

- dato inválido;
- fuera de dominio;
- fallo técnico;
- resultado indeterminado;

• exclusión retrospectiva.

22.72. Diagrama de flujo de muestras

Puede mostrarse:

N_inicial arrow N_incluido arrow N_evaluable arrow N_analizado

con razones de exclusión.

Esto evita que una métrica se calcule sobre una fracción favorable sin que el lector lo advierta.

22.73. Transparencia sobre datos perdidos

Debe informarse:

• frecuencia;
• variables;
• patrones;
• imputación;
• rendimiento con datos incompletos;
• casos bloqueados.

La imputación silenciosa puede producir una falsa apariencia de completitud.

22.74. Transparencia sobre financiación e intereses

Los informes externos o publicaciones deberían declarar:

• desarrollador;
• financiador;
• propiedad intelectual;
• relación con el proveedor;
• participación en análisis;
• independencia del evaluador.

La existencia de interés comercial no invalida el estudio, pero debe ser visible.

22.75. Separación entre dossier científico y material promocional

El material promocional puede ser más breve, pero sus afirmaciones deben poder rastrearse hasta el informe científico.

Debe existir:

afirmación comercial arrow métrica arrow estudio arrow versión

Si no existe esa trazabilidad, la afirmación no debería utilizarse.

22.76. Control interno de claims

La organización debería revisar afirmaciones mediante:

• calidad;
• regulación;
• ciencia de datos;
• jurídico;

• marketing.

El marketing no debería convertir:

Sensibilidad del 98 % en una validación restringida.

en:

Detecta prácticamente todos los riesgos en cualquier alimento.

22.77. Matriz de afirmaciones permitidas

Evidencia	Afirmación permitida	Afirmación no permitida
Validación interna	Rendimiento interno	Validado industrialmente
Test temporal	Estabilidad temporal estudiada	Sin deriva futura
Una planta	Apto para esa planta	Universal
Cribado	Prioriza análisis	Sustituye laboratorio
Cero FN observados	No hubo FN en la muestra	Cero FN garantizados
MAE bajo	Error medio reducido	Predicción exacta
AUC alta	Buena discriminación	Probabilidad correcta
Recalibración local	Probabilidad ajustada localmente	Modelo universalmente calibrado

22.78. Checklist de comunicación científica

Antes de publicar o aprobar un informe debe comprobarse:

1. ¿Se identifica la versión?
2. ¿Se define la finalidad?
3. ¿Se indica la clase positiva?
4. ¿Se informa la prevalencia?
5. ¿Se muestran números absolutos?
6. ¿Se incluyen intervalos?
7. ¿Se diferencia muestra de registro?
8. ¿Se presentan falsos negativos?
9. ¿Se informa el umbral?
10. ¿Se presenta calibración?
11. ¿Se analizan subgrupos?
12. ¿Se describen exclusiones?
13. ¿Se informan datos faltantes?
14. ¿Se muestran limitaciones?
15. ¿Se evita generalizar?
16. ¿Se distingue IA sola de sistema completo?
17. ¿Se indican métodos de confirmación?
18. ¿Se describen criterios de suspensión?
19. ¿Las afirmaciones comerciales son trazables?
20. ¿Puede un tercero reconstruir la conclusión?

22.79. Modelo de tabla resumida

Elemento	Resultado
Modelo y versión	—
Finalidad	—
Dominio	—
Diseño	—
Unidades independientes	—
Positivos reales	—
Prevalencia	—
Sensibilidad	—
Especificidad	—
Falsos negativos	—
Falsos positivos	—
Calibración	—
Subgrupo más débil	—
Robustez	—
Tasa de abstención	—
Limitaciones	—
Uso aprobado	—
Confirmación requerida	—
Próxima revisión	—

22.80. Modelo de conclusión científica

La versión [X] del modelo fue evaluada mediante [diseño] sobre [número] unidades independientes procedentes de [plantas, periodos y productos]. Alcanzó [métricas e intervalos] en el umbral preespecificado. El rendimiento cumplió los criterios principales en [dominios], pero no se demostró suficientemente en [subgrupos o condiciones]. El sistema se autoriza para [uso] con [confirmación, supervisión y barreras]. Las predicciones fuera del dominio o con incertidumbre elevada deberán considerarse no evaluables. La validez deberá mantenerse mediante el programa de monitorización y revalidación establecido.

22.81. Modelo de claim comercial defendible

Sistema de apoyo validado para priorizar controles en las matrices y condiciones especificadas, con rendimiento documentado mediante validación externa y monitorización posterior.

Este claim evita afirmar:

- sustitución;
- infalibilidad;
- universalidad;
- garantía sanitaria.

22.82. Modelo de advertencia operacional

Esta salida es una predicción del modelo y no sustituye los análisis confirmatorios ni los límites críticos definidos. Utilizar únicamente dentro del dominio autorizado y conforme al procedimiento vigente.

22.83. Modelo de comunicación de fuera de dominio

La observación no pertenece al dominio validado. No debe interpretarse la puntuación como estimación fiable. Aplicar el procedimiento alternativo y registrar el caso para evaluación.

22.84. Modelo de comunicación de incertidumbre elevada

El sistema no puede emitir una predicción suficientemente fiable debido a la incertidumbre de los datos. El lote requiere revisión y confirmación independiente.

22.85. El derecho a una conclusión negativa

La transparencia científica exige aceptar que algunos estudios no permitirán autorizar el uso inicialmente previsto.

Una conclusión negativa puede deberse a:

- sensibilidad insuficiente;
- intervalos amplios;
- mala calibración;
- heterogeneidad;
- sobreestimación peligrosa;
- falta de representatividad;
- robustez deficiente.

La organización no debe reformular retrospectivamente el objetivo para producir una apariencia de éxito.

Puede reconducirse el sistema hacia un uso menos crítico, pero esa nueva finalidad deberá justificarse y documentarse.

22.86. Comunicación de una aprobación restringida

Ejemplo:

El modelo no cumple los criterios para liberación automática. Sí demuestra utilidad como sistema de priorización, siempre que se mantenga el análisis confirmatorio y un muestreo aleatorio independiente de predicciones negativas.

Esta conclusión conserva el valor del proyecto sin atribuirle capacidades no demostradas.

22.87. Comunicación de incertidumbre sin paralizar la decisión

La transparencia no significa que toda incertidumbre impida actuar.

Significa que la incertidumbre debe incorporarse a la decisión mediante:

- margen;
- confirmación;
- abstención;
- restricción;
- vigilancia.

Una organización madura no oculta la incertidumbre ni la convierte automáticamente en inacción. La gestiona.

22.88. La métrica como evidencia contextual

Toda métrica debería interpretarse como:

$$M = M(\text{modelo, datos, umbral, tiempo, centro, referencia})$$

No como una constante propia del algoritmo.

Cuando cualquiera de esos elementos cambia, la métrica histórica puede dejar de ser aplicable.

22.89. Transparencia y confianza

La confianza técnicamente adecuada no se obtiene ocultando limitaciones.

Se obtiene demostrando que:

- los errores se conocen;
- los límites están identificados;
- las afirmaciones están controladas;
- existe respuesta ante fallos;
- el rendimiento se monitoriza;
- la organización puede suspender el sistema.

La transparencia bien gestionada no debilita el producto. Distingue una tecnología científicamente gobernada de una promesa comercial no verificable.

22.90. Conclusión del bloque

La comunicación de una validación debe permitir que el lector comprenda no solo cuánto rindió el modelo, sino también:

- en qué población;
- con qué incertidumbre;
- frente a qué referencia;
- bajo qué umbral;
- con qué errores;
- en qué subgrupos;
- con qué limitaciones;
- para qué uso;
- con qué controles.

Una comunicación científicamente íntegra evita cinco simplificaciones:

$$\text{accuracy} \neq \text{seguridad}$$

$$\text{AUC} \neq \text{calibración}$$

$$\text{correlación} \neq \text{concordancia}$$

$$\text{cero errores observados} \neq \text{cero errores reales}$$

$$\text{validación en un contexto} \neq \text{validez universal}$$

La obligación final no consiste en presentar al modelo de la forma más favorable, sino de la forma más verdadera.

Solo cuando la evidencia, la interpretación y la afirmación mantienen una correspondencia exacta puede considerarse que la validación ha sido comunicada de manera científicamente responsable.

23. Marco regulatorio y jurídico de la validación de modelos de IA en calidad, vida útil y seguridad alimentaria

La validación científica de un modelo de inteligencia artificial no constituye únicamente una buena práctica estadística. Cuando sus resultados influyen sobre la formulación, el control de proceso, la clasificación de producto, la determinación de la vida útil, la liberación de lotes o la gestión de un peligro alimentario, la calidad de esa validación adquiere relevancia jurídica.

Una predicción incorrecta puede contribuir a:

- colocar en el mercado un alimento inseguro;
- mantener un producto más allá de su vida útil justificable;
- aceptar una materia prima fuera de especificación;
- omitir un análisis confirmatorio;
- no identificar una desviación de proceso;
- bloquear injustificadamente producto conforme;
- inducir a error a un cliente o consumidor;
- dificultar una investigación o retirada.

La cuestión jurídica esencial no consiste en determinar si «la IA se equivocó». Consiste en determinar:

1. qué persona u organización decidió utilizarla;
2. para qué finalidad fue implantada;
3. qué evidencia respaldaba esa decisión;
4. qué controles se mantuvieron;
5. qué limitaciones eran conocidas o previsibles;
6. qué registros permiten reconstruir el funcionamiento;
7. qué actuación se adoptó ante la predicción;
8. si el resultado alimentario cumplía la legislación aplicable.

El algoritmo no desplaza las obligaciones del operador alimentario. Añade una nueva fuente de información, decisión y riesgo que debe integrarse en el sistema jurídico y documental de la empresa.

23.1. Superposición de marcos regulatorios

La utilización de IA en alimentación no queda regulada por una única norma. Puede activar simultáneamente distintos cuerpos jurídicos:

Derecho alimentario + Reglamento de IA + responsabilidad por productos + protección de datos + propiedad intelectual + contratación + ciberseguridad

Cada uno responde a una cuestión distinta.

El Derecho alimentario determina si el producto es seguro, quién debe garantizar el cumplimiento, qué controles deben implantarse y qué medidas deben adoptarse ante un riesgo. El Reglamento de IA regula determinadas prácticas, sistemas y operadores de IA en función de su clasificación y finalidad. La normativa de responsabilidad establece quién puede responder por daños. La protección de datos se aplica cuando el sistema procesa información personal. El Derecho contractual distribuye obligaciones entre desarrollador, proveedor, integrador y usuario, aunque esa distribución no pueda perjudicar los derechos de terceros ni eliminar responsabilidades legales indisponibles.

No debe asumirse que el cumplimiento del Reglamento de IA demuestra automáticamente el cumplimiento del Derecho alimentario. Tampoco que un modelo no clasificado como de alto riesgo bajo el Reglamento de IA queda libre de exigencias científicas o jurídicas.

Un sistema puede no ser de alto riesgo en el sentido del Reglamento de IA y, sin embargo, desempeñar una función crítica dentro de un APPCC, contribuir a una decisión defectuosa sobre vida útil o generar una responsabilidad grave para el operador alimentario.

23.2. La responsabilidad primaria del operador alimentario

El Reglamento (CE) n.º 178/2002 establece que los alimentos no deben comercializarse cuando sean inseguros y considera inseguros tanto los alimentos perjudiciales para la salud como los no aptos para el consumo humano. La evaluación de la seguridad debe considerar las condiciones normales de uso, la información proporcionada al consumidor, los posibles efectos inmediatos y a largo plazo, las sensibilidades particulares y el deterioro o contaminación del producto.

El artículo 17 atribuye a los operadores de empresas alimentarias y de piensos la responsabilidad de garantizar, en las actividades bajo su control, que los productos cumplen los requisitos relevantes de la legislación alimentaria y de verificar que dichos requisitos se satisfacen.

Esta obligación no se transfiere al proveedor tecnológico.

La empresa no podría justificar una liberación incorrecta alegando:

«El proveedor garantizaba una precisión del 99 %.»

Tampoco:

«La recomendación fue generada automáticamente.»

Ni:

«El software no mostró ninguna alerta.»

La responsabilidad del operador exige demostrar que la selección, validación, implantación y vigilancia del sistema fueron razonables y proporcionadas a la decisión que se le permitió influir.

Debe distinguirse entre:

responsabilidad técnica del proveedor

y:

responsabilidad alimentaria del operador

El proveedor puede responder por un software defectuoso, por información insuficiente, por incumplimiento contractual o por una actualización que degrade el sistema. Pero el operador continúa siendo responsable de garantizar que el alimento colocado en el mercado cumple los requisitos aplicables a sus actividades.

23.3. El Reglamento europeo de inteligencia artificial

El Reglamento (UE) 2024/1689 establece un marco horizontal basado en riesgo para los sistemas de inteligencia artificial. Entró en vigor el 1 de agosto de 2024 y, conforme a su calendario general, resulta aplicable desde el 2 de agosto de 2026, con excepciones y fechas específicas para determinados capítulos y sistemas. Los capítulos I y II, que incluyen la obligación de alfabetización en IA y las prácticas prohibidas, son aplicables desde el 2 de febrero de 2025; las obligaciones correspondientes a los sistemas del artículo 6.1 vinculados a productos regulados se aplican desde el 2 de agosto de 2027.

A fecha de cierre de este trabajo —25 de julio de 2026— la aplicación general del Reglamento se encuentra, por tanto, inmediatamente próxima. Las empresas no deberían esperar a la fecha formal de aplicación para comenzar a clasificar sus sistemas, identificar sus roles y completar la documentación necesaria.

23.4. Un modelo alimentario no es automáticamente un sistema de alto riesgo

La utilización de un modelo para calidad, vida útil o seguridad alimentaria no determina por sí sola su clasificación como sistema de IA de alto riesgo.

El artículo 6 prevé dos grandes vías de clasificación.

La primera afecta a sistemas utilizados como componente de seguridad de un producto —o que sean ellos mismos un producto— cubierto por la legislación de armonización del anexo I, cuando dicho producto deba someterse a una evaluación de conformidad por tercero. La segunda comprende los casos enumerados en el anexo III.

Las actividades ordinarias de predicción de calidad alimentaria, vida útil o riesgo microbiológico no aparecen, por sí mismas, como una categoría autónoma del anexo III.

En consecuencia, un modelo interno que estima la vida útil de una salsa o prioriza muestras microbiológicas no debe calificarse automáticamente como de alto riesgo solo porque su error pueda tener consecuencias importantes. Su clasificación exige analizar la finalidad prevista y comprobar si encaja en alguno de los supuestos normativos.

Sin embargo, pueden existir escenarios en los que un sistema implantado en una industria alimentaria sí quede dentro del régimen de alto riesgo. Por ejemplo:

- cuando actúe como componente de seguridad de maquinaria incluida en la legislación del anexo I y se cumplan los requisitos del artículo 6.1;
- cuando se utilice para una finalidad distinta incluida en el anexo III, como determinadas decisiones laborales sobre trabajadores;
- cuando una modificación de finalidad transforme un sistema inicialmente no clasificado en otro comprendido en una categoría de alto riesgo;
- cuando futuras modificaciones del anexo III incorporen nuevos supuestos.

Debe evitarse, por tanto, tanto la sobreclasificación como la infraclasificación.

Sobreclasificación

«Todo algoritmo relacionado con la seguridad alimentaria es de alto riesgo bajo el Reglamento de IA.»

Esta afirmación no es correcta.

Infraclasificación

«Como se utiliza dentro de una fábrica de alimentos, nunca puede estar sujeto al régimen de alto riesgo.»

Tampoco es correcta.

La clasificación exige un análisis individualizado y documentado.

23.5. Clasificación jurídica y criticidad alimentaria

La clasificación bajo el Reglamento de IA y la criticidad alimentaria son dos ejes diferentes:

Clasificación Reglamento IA	Criticidad alimentaria	Ejemplo
No alto riesgo	Baja	Clasificación estética orientativa
No alto riesgo	Alta	Predicción interna de vida útil microbiológica

Clasificación Reglamento IA	Criticidad alimentaria	Ejemplo
Alto riesgo	No directamente alimentaria	Selección automatizada de trabajadores
Alto riesgo	Alta	Componente de seguridad de equipo regulado que afecta al proceso

La empresa debería mantener dos análisis separados:

$$C\_AI\ Act$$

clasificación jurídica conforme al Reglamento de IA, y:

$$C\_food\ safety$$

criticidad dentro del sistema alimentario.

Que:

$$C\_AI\ Act \neq alto\ riesgo$$

no significa:

$$C\_food\ safety = baja$$

Este punto es esencial. Un modelo que no quede sujeto a los requisitos de los sistemas de alto riesgo puede, no obstante, requerir una validación externa muy exigente porque influye materialmente en una decisión sanitaria.

23.6. Obligaciones generales de alfabetización en IA

El artículo 4 del Reglamento exige a proveedores y responsables del despliegue adoptar medidas para garantizar, en la mayor medida posible, un nivel suficiente de alfabetización en IA del personal y de las demás personas que operen o utilicen los sistemas por cuenta de la organización. La formación debe considerar conocimientos técnicos, experiencia, educación, contexto de utilización y personas o grupos afectados.

En una empresa alimentaria, esta obligación no debería cubrirse mediante una formación genérica sobre qué es la inteligencia artificial.

La competencia necesaria debe ser funcional y específica:

- significado de la salida;
- diferencia entre puntuación y probabilidad;
- sensibilidad y falsos negativos;
- límites del dominio;
- datos obligatorios;
- abstención;
- gestión de alertas;
- actuación ante discrepancias;
- riesgo de dependencia automática;
- registro de la decisión.

La alfabetización jurídica y técnica debería alcanzar, al menos, a:

- científicos de datos;
- responsables de calidad;
- tecnólogos;
- laboratorio;
- producción;

- mantenimiento;
- tecnologías de la información;
- compras;
- dirección;
- personas autorizadas para anular el sistema.

23.7. Requisitos aplicables cuando el sistema es de alto riesgo

Cuando un sistema quede clasificado como de alto riesgo, deberá cumplir el conjunto de exigencias del Reglamento aplicable a proveedores, responsables del despliegue y otros operadores.

Entre los elementos más relevantes para la validación científica se encuentran:

- gestión de riesgos durante todo el ciclo de vida;
- gobernanza y calidad de datos;
- documentación técnica;
- conservación de registros;
- transparencia e instrucciones de uso;
- supervisión humana;
- exactitud, robustez y ciberseguridad;
- evaluación de conformidad;
- monitorización posterior a la comercialización;
- gestión de incidentes.

El artículo 9 configura la gestión de riesgos como un proceso continuo, iterativo, documentado y mantenido durante todo el ciclo de vida del sistema.

Este enfoque coincide con la tesis científica desarrollada en el presente trabajo:

$$\text{validación inicial} \neq \text{validez permanente}$$

La conformidad exige controlar el sistema después de su implantación, no únicamente demostrar un rendimiento histórico.

23.8. Gobernanza de los datos

Para sistemas de alto riesgo entrenados con datos, el artículo 10 exige prácticas de gobernanza apropiadas a la finalidad prevista. Estas deben abarcar el diseño, origen, recogida, preparación, etiquetado, limpieza, supuestos, disponibilidad, cantidad, idoneidad, posibles sesgos y carencias de los conjuntos de entrenamiento, validación y prueba. Los datos deben ser relevantes, suficientemente representativos y, en la mayor medida posible, completos y libres de errores, considerando el contexto geográfico, funcional y operativo de utilización.

La importancia jurídica de esta disposición es considerable.

Aspectos como:

- representatividad;
- independencia del test;
- calidad de las etiquetas;
- cobertura de subgrupos;
- origen de los datos;
- carencias conocidas;

dejan de ser exclusivamente cuestiones académicas cuando el sistema está sujeto al régimen de alto riesgo. Se convierten en elementos de cumplimiento normativo.

Una base de validación construida con pseudorreplicación, fuga de lotes o condiciones no representativas podría comprometer tanto la credibilidad científica como la conformidad jurídica.

23.9. Exactitud y robustez

El artículo 15 exige que los sistemas de alto riesgo alcancen un nivel apropiado de exactitud, robustez y ciberseguridad y mantengan un funcionamiento consistente durante su ciclo de vida. También exige declarar en las instrucciones las métricas de exactitud relevantes y contempla medidas frente a errores, fallos, inconsistencias, bucles de retroalimentación y ataques dirigidos contra datos, modelos o entradas.

La norma no establece una sensibilidad, especificidad o accuracy universal.

El concepto de «nivel apropiado» debe concretarse según:

- finalidad;
- riesgo;
- población;
- consecuencia del error;
- autonomía;
- posibilidad de supervisión;
- estado de la técnica.

La empresa no puede limitarse a declarar:

> *«El modelo tiene una precisión adecuada.»*

Debe especificar:

- métrica;
- valor;
- intervalo;
- conjunto;
- umbral;
- condiciones;
- limitaciones;
- procedimiento de vigilancia.

23.10. Registros automáticos y trazabilidad

Los sistemas de alto riesgo deben permitir el registro automático de eventos durante su ciclo de vida. Los logs deben facilitar la identificación de situaciones de riesgo, la monitorización posterior y el seguimiento de la operación.

En alimentación, el registro debería permitir reconstruir:

lote arrow datos arrow modelo arrow predicción arrow decisión arrow resultado

Como mínimo, deberían conservarse:

- identificación del lote;
- versión del modelo;
- fecha y hora;
- entradas;

- calidad de los datos;
- condición dentro o fuera del dominio;
- salida;
- umbral;
- incertidumbre;
- usuario;
- anulación;
- acción;
- resultado de referencia posterior.

Para sistemas no clasificados como de alto riesgo, esta trazabilidad puede no venir impuesta por el artículo 12, pero continúa siendo una práctica esencial cuando la predicción influye en una decisión alimentaria.

23.11. Transparencia e instrucciones de uso

Los sistemas de alto riesgo deben suministrarse con información suficientemente clara para que el responsable del despliegue interprete la salida y utilice correctamente el sistema. Las instrucciones deben describir, entre otros elementos, finalidad, capacidades, limitaciones, métricas, circunstancias que puedan afectar al rendimiento, datos de entrada, supervisión, mantenimiento y mecanismos de logs.

En una aplicación alimentaria, las instrucciones no deberían limitarse a un manual informático.

Deberían especificar:

- matrices autorizadas;
- plantas;
- equipos;
- intervalos de pH, actividad de agua o temperatura;
- método de muestreo;
- unidad de entrada;
- definición de positivo;
- umbral;
- casos indeterminados;
- necesidad de confirmación;
- límite temporal de la validación;
- acciones ante deriva.

Una limitación no comunicada adecuadamente puede convertirse en un uso indebido previsible.

23.12. Supervisión humana

El Reglamento exige que la supervisión humana sea efectiva y proporcional al riesgo y al nivel de autonomía. La persona responsable debe poder comprender capacidades y limitaciones, identificar anomalías, interpretar el resultado, evitar la dependencia automática, ignorar o revertir la salida e interrumpir el sistema de forma segura.

La mera presencia formal de un responsable de calidad no garantiza el cumplimiento.

Debe comprobarse que esa persona:

- recibe la predicción a tiempo;
- dispone de información contextual;
- conoce el riesgo de falsos negativos;

• puede ordenar una retención;
• tiene autoridad para anular el sistema;
• no es penalizada por cuestionar la salida;
• registra su decisión;
• mantiene competencia técnica.

La supervisión humana no debe convertirse en una ficción contractual destinada a trasladar toda responsabilidad al usuario.

23.13. Obligaciones del responsable del despliegue

Cuando una empresa utiliza un sistema de alto riesgo como *deployer*, debe adoptar medidas técnicas y organizativas para usarlo conforme a las instrucciones, asignar la supervisión a personas competentes, asegurar la relevancia de los datos bajo su control, monitorizar el funcionamiento, comunicar riesgos o incidentes y suspender el uso cuando existan razones para considerar que presenta un riesgo.

Esto significa que el usuario industrial no puede limitarse a instalar un sistema que ya haya sido evaluado por el proveedor.

Debe efectuar una verificación local que confirme:

• correcta integración;
• compatibilidad de datos;
• adecuación a la población;
• formación;
• monitorización;
• posibilidad de suspensión;
• trazabilidad.

23.14. Cuándo la empresa puede convertirse en proveedor

La empresa alimentaria puede dejar de ser únicamente responsable del despliegue y asumir las obligaciones de proveedor cuando, en relación con un sistema de alto riesgo:

• coloca su nombre o marca sobre el sistema;
• realiza una modificación sustancial;
• modifica la finalidad prevista de un sistema inicialmente no clasificado, de forma que pase a ser de alto riesgo.

El artículo 25 atribuye en esos supuestos las obligaciones del proveedor al distribuidor, importador, responsable del despliegue o tercero que lleve a cabo la actuación.

Esta cuestión es especialmente importante en proyectos internos.

Una empresa puede adquirir una plataforma y posteriormente:

• reentrenarla;
• añadir nuevas variables;
• modificar el umbral;
• cambiar el producto objetivo;
• automatizar la liberación;
• integrarla en una máquina;
• ampliar el dominio.

No toda personalización constituye una modificación sustancial. Pero aquellas modificaciones no previstas en la evaluación original que afecten al cumplimiento o cambien la finalidad pueden exigir una nueva evaluación y alterar el rol jurídico de la empresa.

23.15. Derecho alimentario y APPCC

El Reglamento (CE) n.º 852/2004 exige que los operadores implanten, apliquen y mantengan procedimientos permanentes basados en los principios APPCC.

Estos principios incluyen:

- identificación de peligros;
- determinación de puntos de control crítico;
- límites críticos;
- monitorización;
- acciones correctivas;
- verificación;
- documentación y registros.

Cuando se modifica un producto, proceso o etapa, el operador debe revisar el procedimiento y realizar los cambios necesarios. También debe presentar a la autoridad evidencia del cumplimiento, mantener la documentación actualizada y conservar los registros durante un periodo apropiado.

La introducción de un modelo de IA que influya en cualquiera de esos elementos debe tratarse como una modificación del sistema.

La revisión debería responder:

1. ¿Cambia la identificación de peligros?
2. ¿Modifica un límite o alerta?
3. ¿Sustituye una medición?
4. ¿Modifica la frecuencia de muestreo?
5. ¿Activa una acción correctiva?
6. ¿Excluye casos de análisis?
7. ¿Cambia la liberación del producto?
8. ¿Introduce nuevos modos de fallo?

Si la respuesta es afirmativa, la modificación debe integrarse documentalmente en el APPCC o en el programa de prerrequisitos correspondiente.

23.16. Validación de una medida de control

Codex define la validación como la obtención y evaluación de evidencia científica, técnica y observacional destinada a demostrar que una medida de control puede alcanzar el resultado de seguridad especificado. Distingue la validación inicial de la monitorización y verificación posteriores y recomienda realizarla, siempre que sea posible, antes de la plena implantación y nuevamente cuando los cambios lo justifiquen.

Si el modelo forma parte material de una medida de control, no basta con validar su clasificación estadística.

Debe validarse el efecto completo:

predicción arrow acción arrow control del peligro

Por ejemplo, un modelo puede anticipar correctamente una desviación térmica, pero la medida de control no será eficaz si:

- la alerta llega tarde;
- nadie la recibe;
- no existe acción predefinida;
- la línea no puede detenerse;
- el lote no queda identificado;
- la confirmación no se realiza.

La evidencia debe demostrar que el sistema completo consigue el objetivo de control.

23.17. Utilización como método analítico alternativo

El Reglamento n.º 852/2004 permite, en determinados supuestos en los que no exista un método específicamente establecido, utilizar métodos que ofrezcan resultados equivalentes a los del método de referencia, siempre que estén científicamente validados conforme a reglas o protocolos internacionalmente reconocidos.

Esta disposición no significa que cualquier modelo de IA con buena correlación pueda sustituir a una técnica analítica.

La equivalencia exige demostrar:

- concordancia;
- rango;
- sesgo;
- precisión;
- reproducibilidad;
- comportamiento próximo al límite;
- interferencias;
- aplicabilidad a la matriz;
- incertidumbre.

Cuando exista un método legalmente exigido, contractual u oficial, el modelo puede ser útil para cribado o autocontrol, pero no sustituirlo automáticamente.

La clasificación correcta puede ser:

Uso	Posible función de la IA
Investigación	Exploración
Control de proceso	Estimación rápida
Cribado	Selección de muestras
Autocontrol	Apoyo validado
Liberación	Solo cuando exista fundamento suficiente
Control oficial	Conforme al marco específico aplicable
Resolución de controversias	Método reconocido o aceptado

23.18. Trazabilidad alimentaria y trazabilidad algorítmica

El artículo 18 del Reglamento n.º 178/2002 exige trazabilidad de alimentos, piensos, animales y sustancias a lo largo de las fases de producción, transformación y distribución. El operador debe identificar proveedores y destinatarios y disponer de sistemas que permitan facilitar esa información a la autoridad.

La trazabilidad algorítmica no sustituye la trazabilidad alimentaria.

Ambas deben conectarse:

$$\text{trazabilidad del producto} + \text{trazabilidad de la decisión}$$

Ante un incidente debería ser posible determinar:

- qué lotes fueron evaluados;
- con qué versión;
- qué datos recibieron;
- qué decisión se adoptó;
- qué lotes compartían condiciones equivalentes;
- si otros lotes fueron afectados por el mismo fallo.

Sin esa conexión, una incidencia del modelo puede impedir delimitar correctamente el alcance de una retirada.

23.19. Retirada y comunicación a autoridades

Cuando un operador considere o tenga razones para creer que un alimento no cumple los requisitos de seguridad, debe iniciar su retirada, informar a las autoridades y, cuando sea necesario, comunicar a los consumidores y proceder a la recuperación. También debe informar inmediatamente cuando considere que un alimento comercializado puede ser perjudicial para la salud.

Un fallo de IA puede constituir la información que active ese deber.

Ejemplos:

- se descubre una tasa de falsos negativos superior a la validada;
- se identifica una versión incorrecta;
- el umbral fue modificado;
- el modelo dejó de detectar un subgrupo;
- existió corrupción de datos;
- una actualización alteró la predicción de lotes ya liberados.

La empresa debe evaluar retrospectivamente:

$$\text{periodo afectado} \times \text{versiones} \times \text{productos} \times \text{lotes}$$

y determinar si existe razón suficiente para considerar comprometida la seguridad.

No resulta jurídicamente prudente esperar a que se confirme un daño cuando ya existe evidencia de que una barrera crítica pudo fallar.

23.20. Controles oficiales

El Reglamento (UE) 2017/625 establece el marco de los controles oficiales destinados a verificar el cumplimiento de la legislación de la cadena agroalimentaria. Las autoridades pueden examinar operadores, actividades, procesos, documentación, registros y sistemas relevantes para evaluar la conformidad.

Si una empresa utiliza un modelo para justificar:

- una vida útil;
- una reducción de muestreo;
- una decisión de liberación;
- una medida de control;
- una clasificación de riesgo;

debería estar preparada para presentar una evidencia comprensible y verificable.

El expediente no debería depender de que la autoridad acepte una afirmación genérica como:

«Es una caja negra propietaria.»

El secreto industrial puede proteger detalles técnicos, pero no debería impedir demostrar:

- finalidad;
- validación;
- límites;
- versión;
- resultados;
- control de cambios;
- actuación ante fallos.

23.21. Valor probatorio de los registros

Los logs y expedientes no prueban por sí solos que la predicción fuese correcta. Prueban, cuando son íntegros y fiables, qué información recibió el sistema, qué versión operó y qué salida generó.

Su utilidad probatoria depende de:

- autenticidad;
- integridad;
- marca temporal;
- identificación de la versión;
- conservación;
- control de accesos;
- ausencia de sobrescritura;
- posibilidad de interpretación;
- vinculación con el lote.

Un log que contiene únicamente:

Resultado: 0,74

posee un valor limitado si no puede determinarse:

- qué significaba 0,74;
- cuál era el umbral;
- si era una probabilidad calibrada;
- qué variables se utilizaron;
- qué versión lo produjo;
- si la observación estaba dentro del dominio.

La evidencia debería permitir reproducir o, al menos, reconstruir racionalmente la decisión.

23.22. Expediente probatorio mínimo

Para decisiones críticas deberían conservarse:

Identidad del sistema

- proveedor;
- versión;

- configuración;
- finalidad;
- umbral.

Evidencia científica

- protocolo;
- datos;
- referencia;
- métricas;
- intervalos;
- subgrupos;
- limitaciones.

Evidencia de implantación

- verificación local;
- formación;
- pruebas de integración;
- procedimientos;
- aprobación.

Evidencia operacional

- entradas;
- salida;
- supervisión;
- decisión;
- resultado posterior.

Evidencia de mantenimiento

- monitorización;
- deriva;
- cambios;
- incidentes;
- revalidación.

La ausencia de documentación no demuestra que el modelo fuese incorrecto, pero puede dificultar seriamente acreditar que la empresa actuó con diligencia técnica.

23.23. Protección de datos personales

El Reglamento de IA no sustituye la normativa de protección de datos. Cuando el sistema procese datos personales, continuará siendo aplicable el Reglamento General de Protección de Datos y el resto de legislación pertinente.

En modelos alimentarios puede aparecer tratamiento de datos personales cuando se utilizan:

- imágenes donde aparecen empleados;
- identificadores de operarios;
- productividad individual;
- geolocalización;

- datos de formación;
- patrones de conducta;
- decisiones laborales;
- cámaras en zonas de trabajo.

Debe analizarse:

- base jurídica;
- necesidad;
- proporcionalidad;
- minimización;
- plazos de conservación;
- derechos de los trabajadores;
- información;
- acceso;
- seguridad.

La utilización de la identidad del operario como predictor puede, además, introducir relaciones espurias y convertir un proyecto de calidad en un sistema de evaluación laboral con implicaciones regulatorias diferentes.

23.24. Secreto industrial y transparencia

El Reglamento de IA reconoce la necesidad de proteger propiedad intelectual, información empresarial confidencial y secretos comerciales, pero también impone cooperación y acceso a información suficiente a lo largo de la cadena de valor en los sistemas de alto riesgo.

La solución contractual no consiste en exigir al proveedor la revelación indiscriminada de su código fuente.

Consiste en garantizar acceso a la evidencia necesaria para evaluar:

- datos y alcance;
- métricas;
- limitaciones;
- cambios;
- incidentes;
- comportamiento fuera de dominio;
- interoperabilidad;
- logs;
- continuidad.

Puede utilizarse:

- acceso controlado;
- auditoría por tercero;
- sala segura;
- información agregada;
- depósito de código;
- cláusula de emergencia;
- informes de validación independientes.

El secreto industrial no debe convertirse en opacidad irresponsable.

23.25. Contratación con proveedores de IA

El contrato debería distribuir claramente las obligaciones, pero no limitarse a una licencia de software estándar.

Debe regular, como mínimo:

Finalidad

- uso autorizado;
- usos excluidos;
- matrices;
- equipos;
- decisiones.

Validación

- evidencia facilitada;
- métricas;
- centros;
- versiones;
- límites;
- asistencia a la verificación local.

Cambios

- notificación previa;
- autorización;
- documentación;
- posibilidad de congelar una versión;
- revalidación;
- reversión.

Datos

- titularidad;
- acceso;
- calidad;
- uso para entrenamiento;
- confidencialidad;
- exportación;
- eliminación.

Operación

- disponibilidad;
- latencia;
- soporte;
- continuidad;
- copias;
- recuperación.

Incidentes

- tiempos de comunicación;
- investigación;
- acceso a logs;
- cooperación con autoridades;
- acciones correctivas.

Responsabilidad

- defectos;
- incumplimientos;
- recobro;
- límites contractuales;
- seguros;
- indemnización.

Terminación

- portabilidad;
- conservación de registros;
- entrega de datos;
- funcionamiento alternativo;
- dependencia tecnológica.

La cláusula que permita al proveedor actualizar automáticamente el modelo sin evaluación previa debería considerarse de alto riesgo contractual cuando la versión influye en calidad o seguridad.

23.26. Actualizaciones y modificaciones

Una actualización puede afectar:

- pesos;
- arquitectura;
- variables;
- preprocesamiento;
- calibración;
- umbral;
- interfaz;
- sensor;
- reglas de negocio.

La empresa debe conocer:

$$\Delta\_\text{versión}$$

antes de permitir su entrada en producción.

La pregunta contractual correcta no es:

«¿La actualización mejora el sistema?»

Sino:

«¿Qué cambia, qué evidencia lo respalda y qué parte de la validación anterior continúa siendo aplicable?»

Las actualizaciones deberían pasar por:

1. análisis de impacto;
2. entorno de prueba;
3. comparación;
4. aprobación;
5. despliegue;
6. verificación;
7. posibilidad de reversión.

23.27. Responsabilidad por productos defectuosos

La Directiva (UE) 2024/2853 actualiza el régimen europeo de responsabilidad por productos defectuosos para incluir expresamente el software dentro del concepto de producto. Sus considerandos indican que los sistemas de IA pueden constituir software y que sus proveedores pueden ser tratados como fabricantes a estos efectos. La Directiva deberá transponerse antes del 9 de diciembre de 2026 y se aplicará a productos puestos en el mercado o en servicio después de esa fecha.

Este régimen no elimina la responsabilidad del fabricante del alimento.

Pueden coexistir:

- un alimento defectuoso;
- un software defectuoso;
- una integración incorrecta;
- un uso no conforme;
- una supervisión insuficiente;
- una actualización defectuosa.

La atribución final dependerá de los hechos, la causalidad, la normativa nacional de transposición y las relaciones entre los operadores.

Debe evitarse una interpretación simplista:

«Si la IA falla, responde exclusivamente el proveedor.»

El fabricante alimentario puede continuar respondiendo por el producto que comercializó, sin perjuicio de las acciones contractuales o de repetición que correspondan frente al proveedor tecnológico.

23.28. Software bajo control del fabricante

La nueva Directiva contempla que el software, las actualizaciones y determinados servicios digitales integrados pueden permanecer bajo control del fabricante, y que las modificaciones sustanciales o la falta de actualizaciones necesarias para mantener la seguridad pueden afectar al análisis de defectuosidad y responsabilidad.

Desde una perspectiva preventiva, la empresa debería documentar:

- quién controla las actualizaciones;
- quién decide instalarlas;
- quién monitoriza vulnerabilidades;
- qué ocurre si se pierde conectividad;
- qué versión estaba operativa;
- quién autorizó una modificación.

La ausencia de control de versiones puede convertirse tanto en un fallo técnico como en un problema de imputación jurídica.

23.29. Responsabilidad contractual

Aunque no se produzca un daño personal, un modelo defectuoso puede generar:

- pérdidas de producto;
- incumplimientos de suministro;
- reclamaciones de clientes;
- costes analíticos;
- retrasos;
- reprocesados;
- retirada;
- daño reputacional.

El contrato debe definir:

- niveles de servicio;
- garantías;
- conformidad con especificaciones;
- deber de cooperación;
- límites de responsabilidad;
- exclusiones;
- daños indirectos;
- derecho de recobro;
- seguros.

Una limitación contractual amplia puede dejar al operador alimentario expuesto frente a clientes y consumidores sin una vía efectiva de recuperación frente al proveedor.

23.30. Claims y comunicaciones engañosas

El artículo 16 del Reglamento n.º 178/2002 prohíbe que la presentación, publicidad e información sobre alimentos o piensos induzca a error.

Aunque esta disposición se dirige a la presentación de alimentos y piensos, el principio de veracidad también debe proyectarse sobre las afirmaciones utilizadas para justificar decisiones técnicas y comerciales.

Expresiones como:

- «vida útil garantizada por IA»;
- «cero riesgo microbiológico»;
- «100 % de detección»;
- «sin necesidad de laboratorio»;
- «modelo universal»;

pueden exceder la evidencia y crear una expectativa técnicamente injustificable.

Una comunicación responsable debería vincular toda afirmación a:

$$\text{versión} + \text{uso} + \text{población} + \text{métrica} + \text{limitaciones}$$

23.31. Sanciones del Reglamento de IA

Para los sistemas y operadores comprendidos en su ámbito, el Reglamento de IA establece sanciones administrativas relevantes. El incumplimiento de determinadas obligaciones aplicables a operadores y organismos puede alcanzar, según el supuesto y sin perjuicio de las reglas específicas para empresas, hasta 15 millones de euros o el 3 % del volumen de negocio mundial anual del ejercicio anterior; el suministro de información incorrecta, incompleta o engañosa puede alcanzar hasta 7,5 millones de euros o el 1 %.

Estas cifras no deben presentarse como si cualquier modelo alimentario estuviera automáticamente expuesto a dichas sanciones. Su aplicación depende:

- de que el sistema quede dentro del ámbito correspondiente;
- del rol del operador;
- de la obligación incumplida;
- de las reglas de graduación;
- del marco de ejecución nacional.

Sin embargo, muestran que la calidad de la documentación y la veracidad de la información facilitada a autoridades u organismos de evaluación poseen una relevancia jurídica directa.

23.32. Riesgo de información incompleta o engañosa

Una empresa podría incurrir en una presentación inadecuada si comunica a una autoridad, auditor, cliente u organismo:

«El sistema fue validado sobre 100.000 muestras.»

cuando en realidad se trataba de:

- 100.000 imágenes;
- 500 unidades;
- 20 lotes;
- réplicas técnicas;
- partición contaminada.

La información puede ser literalmente cierta en el número de archivos y, sin embargo, materialmente engañosa respecto a la fuerza de la evidencia.

La documentación debe identificar la unidad independiente y evitar una exageración del tamaño de validación.

23.33. Conservación y disponibilidad de documentación

La legislación alimentaria exige mantener actualizados los documentos APPCC, conservar registros durante un periodo apropiado y facilitar evidencia a la autoridad.

El periodo de conservación de los registros del modelo debería considerar:

- vida útil del producto;
- periodo de distribución;
- posibilidad de reclamaciones;
- obligaciones contractuales;
- investigaciones;
- retiradas;
- plazos legales;

• vida operativa de la versión.

No resulta razonable eliminar logs críticos inmediatamente después de la liberación si el producto continuará en el mercado durante meses.

23.34. Auditoría regulatoria del modelo

Una auditoría jurídica y técnica debería comprobar:

Clasificación

• ¿Es realmente IA bajo el Reglamento?
• ¿Es alto riesgo?
• ¿Qué rol desempeña la empresa?
• ¿Existe análisis documentado?

Derecho alimentario

• ¿Qué decisión afecta?
• ¿Está integrado en APPCC?
• ¿Se mantiene el método obligatorio?
• ¿Existe evidencia científica?

Datos

• ¿Son representativos?
• ¿Existe origen y trazabilidad?
• ¿Se procesan datos personales?
• ¿Hay derechos suficientes de utilización?

Contrato

• ¿Se notifican cambios?
• ¿Existe acceso a logs?
• ¿Se regula el soporte?
• ¿Puede recuperarse la información?

Operación

• ¿Existe supervisión?
• ¿Puede suspenderse?
• ¿Hay respaldo?
• ¿Se gestionan incidentes?

Prueba

• ¿Puede reconstruirse una predicción?
• ¿Se conoce la versión?
• ¿Se conservan entradas y salidas?
• ¿Las limitaciones están documentadas?

23.35. Matriz jurídica de utilización

Función del modelo	Riesgo jurídico principal	Control necesario
Investigación	Sobreinterpretación	Separar exploración de validación
Clasificación de calidad	Incumplimiento contractual	Especificaciones y confirmación
Vida útil	Producto inseguro o deteriorado	Estudios, margen y trazabilidad
Cribado microbiológico	Falso negativo	Muestreo independiente
Liberación de lote	Responsabilidad alimentaria	Validación máxima y barreras
Control de proceso	Fallo de medida	Redundancia y contingencia
Gestión de trabajadores	Clasificación Reglamento IA y RGPD	Evaluación específica
Integración en maquinaria	Seguridad de producto	Análisis del artículo 6.1
Servicio de proveedor	Dependencia contractual	Control de versiones y acceso
Aprendizaje continuo	Modificación no controlada	Modelo candidato y revalidación

23.36. Criterios jurídicos de diligencia

La diligencia empresarial puede evaluarse mediante preguntas como:

1. ¿Se identificó correctamente la finalidad?
2. ¿Se evaluó la legislación aplicable?
3. ¿Se validó con datos independientes?
4. ¿Se conocían las limitaciones?
5. ¿Se mantuvieron controles adicionales?
6. ¿Se monitorizó la deriva?
7. ¿Se formó al personal?
8. ¿Se documentaron los cambios?
9. ¿Se conservaron registros?
10. ¿Se actuó al detectar un riesgo?

Una empresa no demuestra diligencia simplemente porque compró una herramienta reconocida.

La diligencia reside en el proceso mediante el cual decidió utilizarla y en cómo gobernó su comportamiento.

23.37. El expediente como defensa técnica

Un expediente sólido no garantiza la ausencia de responsabilidad. Pero permite demostrar que la empresa:

- analizó el riesgo;
- seleccionó una metodología razonable;
- fijó criterios prospectivos;
- utilizó datos representativos;
- investigó los fallos;
- mantuvo supervisión;
- reaccionó ante la deriva;
- no ocultó las limitaciones.

La inexistencia de documentación puede generar una situación especialmente desfavorable: la organización puede haber actuado correctamente, pero carecer de evidencia para demostrarlo.

23.38. La validación insuficiente como fallo organizativo

Cuando un modelo crítico se implanta sin validación suficiente, el problema no debería atribuirse exclusivamente al departamento de datos.

Puede reflejar fallos de:

- gobierno corporativo;
- calidad;
- compras;
- jurídico;
- dirección;
- producción;
- gestión del cambio.

La aprobación debe ser multidisciplinar porque el riesgo también lo es.

23.39. Requisitos contractuales mínimos de evidencia

Antes de adquirir un modelo destinado a calidad o seguridad, la empresa debería exigir al proveedor:

1. finalidad prevista;
2. definición de las clases o variable;
3. matrices evaluadas;
4. unidad estadística;
5. estrategia de partición;
6. métricas completas;
7. intervalos;
8. falsos positivos y negativos;
9. validación externa;
10. subgrupos;
11. limitaciones;
12. detección fuera de dominio;
13. control de versiones;
14. política de actualización;
15. soporte ante incidentes;
16. acceso a registros;
17. cooperación regulatoria.

La negativa absoluta a facilitar esta información debería tratarse como una señal de riesgo técnico y contractual.

23.40. Flujo jurídico de aprobación

Puede establecerse el siguiente procedimiento:

Paso 1. Clasificación

Determinar:

- finalidad;
- rol;
- nivel de riesgo;

- normativa aplicable.

Paso 2. Validación científica

Comprobar:

- datos;
- referencia;
- diseño;
- métricas;
- criterios.

Paso 3. Integración alimentaria

Revisar:

- APPCC;
- prerrequisitos;
- método;
- acciones;
- trazabilidad.

Paso 4. Contratación

Asegurar:

- acceso;
- versiones;
- cambios;
- responsabilidad;
- continuidad.

Paso 5. Aprobación

Documentar:

- uso autorizado;
- restricciones;
- responsables;
- condiciones.

Paso 6. Vigilancia

Controlar:

- rendimiento;
- deriva;
- incidentes;
- cambios.

Paso 7. Revalidación o retirada

Actuar cuando:

- cambia el producto;
- cambia el modelo;
- cambia la referencia;
- se incumple una métrica;
- aparece un incidente.

23.41. Declaración regulatoria interna

Una autorización interna debería formularse de manera similar a la siguiente:

La versión [X] del sistema [nombre] ha sido evaluada para la finalidad, productos, plantas, equipos y condiciones descritos en el expediente. Su clasificación bajo el Reglamento de IA ha sido analizada y documentada. La utilización no modifica la responsabilidad del operador alimentario ni sustituye los métodos, controles o requisitos legalmente aplicables salvo autorización expresa. El sistema queda sujeto a supervisión, registro, monitorización, gestión de cambios y suspensión conforme a los procedimientos establecidos.

23.42. Causas de rechazo jurídico-técnico

El sistema no debería aprobarse cuando:

- no puede determinarse la finalidad;
- la clasificación jurídica no está analizada;
- el proveedor no identifica la versión;
- no existe validación independiente;
- las métricas son incompletas;
- no se conocen falsos negativos;
- no existe procedimiento de respaldo;
- el usuario no puede anularlo;
- las actualizaciones son opacas;
- no pueden conservarse logs;
- el contrato impide investigar incidentes;
- se pretende sustituir un método obligatorio sin fundamento;
- no puede integrarse en APPCC;
- el sistema utiliza datos personales sin base suficiente;
- la organización no puede suspenderlo.

23.43. Conclusión del bloque

El marco jurídico de la IA alimentaria no debe interpretarse como una sustitución del Derecho alimentario por una nueva regulación tecnológica.

La relación correcta es acumulativa:

$$\text{Reglamento de IA} \cap \text{Derecho alimentario} \cap \text{responsabilidad} \cap \text{contratación}$$

El Reglamento de IA determina obligaciones específicas según la clasificación y el rol del operador. El Derecho alimentario mantiene la responsabilidad primaria de garantizar que el producto sea seguro y conforme. El APPCC exige que las medidas y modificaciones estén justificadas, documentadas y verificadas. La contratación distribuye funciones entre empresas, pero no elimina obligaciones frente a terceros. La trazabilidad probatoria permite reconstruir la decisión y demostrar diligencia.

La afirmación jurídicamente incorrecta sería:

«El modelo cumple el Reglamento de IA; por tanto, la decisión alimentaria es válida.»

La conclusión defendible es:

> *«El sistema ha sido clasificado y gobernado conforme al marco de IA aplicable, científicamente validado para su finalidad, integrado en el sistema alimentario, sujeto a controles independientes y utilizado sin desplazar las responsabilidades legales del operador.»*

En última instancia, el objeto del cumplimiento no es el algoritmo aislado.

Es el conjunto formado por:

datos + modelo + personas + procedimientos + producto

Cuando ese conjunto conduce a una decisión sobre seguridad, calidad o vida útil, la empresa debe estar en condiciones de demostrar no solo que la predicción era técnicamente plausible, sino que confiar en ella resultaba científica, organizativa y jurídicamente razonable.

24. Casos prácticos integrales de validación de modelos de IA en alimentación

La metodología desarrollada en los capítulos anteriores puede parecer abstracta mientras no se aplique a sistemas concretos. En la práctica, la misma sensibilidad, el mismo error medio o una misma AUC pueden conducir a decisiones regulatorias y operativas diferentes según:

- la finalidad del modelo;
- el tipo de error dominante;
- la gravedad de sus consecuencias;
- la prevalencia;
- la existencia de confirmación;
- la capacidad de abstención;
- el dominio validado;
- el comportamiento por subgrupos.

Los tres casos siguientes muestran cómo debe transformarse una colección de métricas en una decisión científicamente defendible.

Se analizarán:

1. un sistema de visión artificial para defectos de calidad;
2. un modelo de predicción de vida útil refrigerada;
3. un modelo de cribado de riesgo microbiológico.

En cada caso se seguirá la misma secuencia:

Finalidad - riesgo - datos - criterios - resultados - errores - decisión

25. Caso práctico I. Visión artificial para la detección de defectos en producto envasado

25.1. Descripción del sistema

Una empresa implanta un sistema de visión artificial en una línea de productos envasados. El sistema captura varias imágenes de cada unidad y clasifica el producto como:

- conforme;
- potencialmente defectuoso;
- no evaluable.

Los defectos de interés son:

1. cierre incompleto;
2. etiqueta mal posicionada;
3. deformación del envase;
4. presencia de residuo visible en la zona de sellado;
5. cuerpo extraño visible.

La salida del modelo está conectada a un expulsor automático.

La finalidad propuesta es:

Detectar y retirar de la línea unidades con defectos visuales previamente definidos, sin sustituir los controles del detector de metales, rayos X, integridad de cierre ni otras medidas de seguridad independientes.

Esta formulación delimita correctamente la función. El modelo no se presenta como una garantía general de ausencia de cuerpos extraños ni como sustituto de cualquier control físico.

25.2. Unidad estadística

Cada unidad de producto genera cuatro imágenes:

- vista superior;
- vista lateral izquierda;
- vista lateral derecha;
- detalle del sellado.

El número de imágenes no constituye el tamaño muestral independiente.

Si se evalúan 12.000 envases:

$$n\_unidades=12.000$$

aunque se hayan procesado:

$$12.000\times4=48.000 \text{ imágenes}$$

Todas las imágenes de un mismo envase deben permanecer en la misma partición.

La unidad de decisión y la unidad estadística principal es:

$$\text{envase individual}$$

25.3. Diseño de los datos

Desarrollo

- 30.000 unidades;

- 150 lotes;
- dos campañas;
- dos turnos;
- una cámara principal;
- tres velocidades de línea.

Validación externa

- 12.000 unidades;
- 60 lotes nuevos;
- campaña posterior;
- operarios diferentes;
- segunda unidad de cámara del mismo modelo;
- condiciones de iluminación industrial habituales.

No se utilizaron imágenes, lotes ni fechas de la validación externa para ajustar:

- arquitectura;
- preprocesamiento;
- aumentación;
- hiperparámetros;
- umbral;
- detector de imágenes no evaluables.

25.4. Patrón de referencia

Cada unidad fue revisada fuera de línea por dos inspectores entrenados.

Cuando existió desacuerdo:

- un tercer inspector realizó adjudicación ciega;
- ninguno conoció la predicción del modelo;
- se conservaron tanto las valoraciones originales como el resultado adjudicado.

La referencia se clasificó en:

- conforme;
- defecto menor;
- defecto mayor;
- defecto crítico.

Para la evaluación primaria, se definió como positivo cualquier defecto mayor o crítico.

Los defectos menores se analizaron separadamente para evitar que una categoría estética de baja gravedad dominara la métrica global.

25.5. Riesgo del sistema

Los errores principales son:

Falso negativo

Una unidad defectuosa continúa en la línea.

Consecuencias posibles:

- reclamación;
- incumplimiento de especificación;
- pérdida de imagen;
- riesgo físico si existe cuerpo extraño visible;
- fallo de integridad si el defecto afecta al cierre.

Falso positivo

Una unidad conforme es expulsada.

Consecuencias:

- desperdicio;
- reprocesado;
- pérdida de capacidad;
- revisión innecesaria.

El coste de ambos errores no es simétrico.

Para defectos de cierre y cuerpos extraños, el falso negativo tiene mayor criticidad. Para etiquetas ligeramente desplazadas, el falso positivo puede tener mayor relevancia económica que sanitaria.

25.6. Criterios preespecificados

Antes de abrir el conjunto externo se establecieron:

Rendimiento global

$$\text{Sensibilidad} \geq 95\%$$

$$\text{Especificidad} \geq 96\%$$

Defectos críticos

$$\text{Sensibilidad crítica} \geq 99\%$$

Incertidumbre

$$\text{LCB\_95\%(sensibilidad global)} \geq 93\%$$

Operación

$$\text{Falsos rechazos} \leq 35 \text{ por } 1.000 \text{ unidades conformes}$$

$$\text{Imágenes no evaluables} \leq 1\%$$

Latencia

$$\text{T\_total} \leq \text{T\_máximo de expulsión}$$

Actuación

$$\text{Eficacia del expulsor} \geq 99{,}9\%$$

Seguridad

- parada o alarma ante pérdida de cámara;
- abstención ante imagen degradada;
- mantenimiento de controles físicos independientes.

25.7. Resultados globales

En la validación externa se identificaron:

• 600 unidades con defecto mayor o crítico;

• 11.400 unidades conformes o con defecto menor.

La matriz de confusión fue:

Referencia	IA positiva	IA negativa	Total
Defecto mayor o crítico	576	24	600
No defectuosa	342	11.058	11.400
Total	918	11.082	12.000

Por tanto:

[Sensibilidad [576] / [576+24] 96{,}0%]

[Especificidad [11.058] / [11.058+342] 97{,}0]

[VPP [576] / [576+342]62{,}7%]

[VPN [11.058] / [11.058+24]99{,}78%]

[Accuracy [576+11.058] / [12.000] 96{,}95%]

[Balanced accuracy 96,0+97,02 96{,}5%]

F1 ≈ 75,9%

La diferencia entre accuracy y F1 demuestra que cada métrica responde a una propiedad distinta. El sistema presenta una elevada exactitud global, pero el valor predictivo positivo queda limitado por la proporción de unidades defectuosas y por el número de falsos rechazos.

25.8. Interpretación operacional

Por cada 1.000 unidades conformes, el sistema rechaza aproximadamente:

$$1.000\times(1-0,97)=30$$

unidades conformes.

Cumple, por tanto, el límite operativo de 35 falsos rechazos por 1.000 unidades conformes.

Sin embargo, entre las 918 unidades expulsadas:

• 576 son verdaderamente defectuosas;

• 342 son conformes.

Esto significa que cerca de un 37 % de los rechazos deberán:

• descartarse;

• revisarse;

• reprocesarse;

• reintroducirse si el procedimiento lo permite.

La empresa debe evaluar si dispone de capacidad para gestionar ese flujo.

Una sensibilidad elevada no basta si el sistema genera una carga de rechazo que termina provocando:

• acumulaciones;

• reintroducciones sin control;

• desactivación del expulsor;

• fatiga del personal.

25.9. Rendimiento por tipo de defecto

Defecto	Casos reales	Detectados	Sensibilidad
Cierre incompleto	120	120	100,0 %
Etiqueta mal posicionada grave	180	174	96,7 %
Deformación	130	126	96,9 %
Residuo en sellado	100	96	96,0 %
Cuerpo extraño visible	70	60	85,7 %

El resultado global del 96 % oculta una deficiencia grave.

La sensibilidad para cuerpos extraños visibles es:

$$[60] / [70]=85,7\%$$

muy inferior al criterio del 99 % establecido para defectos críticos.

La media global no puede compensar este incumplimiento.

25.10. Análisis del fallo

Los diez cuerpos extraños no detectados comparten características:

• material transparente;
• tamaño reducido;
• ubicación próxima a un reflejo;
• contraste insuficiente;
• oclusión parcial por la etiqueta.

El sistema no falla aleatoriamente. Presenta un modo de fallo sistemático.

El análisis de las explicaciones visuales muestra que la red dirige su atención hacia:

• contornos de envase;
• zonas de sellado;
• contraste de etiqueta;

pero no distingue adecuadamente determinados objetos transparentes.

Esta explicación ayuda a investigar el fallo, pero no transforma el resultado en aceptable.

25.11. Pruebas de robustez

Se evaluaron:

• iluminación reducida;
• aumento de velocidad;
• condensación;
• lente parcialmente sucia;
• variación de posición;
• segunda cámara.

Resultados:

Condición	Sensibilidad global	No evaluables
Nominal	96,0 %	0,3 %
Iluminación reducida	94,2 %	0,8 %
Velocidad máxima	95,1 %	0,5 %
Condensación	89,4 %	4,9 %
Lente parcialmente sucia	87,8 %	8,6 %
Segunda cámara	95,5 %	0,4 %

La condición de condensación y la suciedad generan degradación relevante.

Sin embargo, el sistema detecta correctamente gran parte de las imágenes degradadas y se abstiene.

La arquitectura operacional debe establecer:

$$\text{calidad de imagen insuficiente} \Rightarrow \text{alarma y revisión}$$

No:

$$\text{calidad insuficiente} \Rightarrow \text{clasificación conforme}$$

25.12. Validación del actuador

El modelo emitió correctamente 918 señales de rechazo.

El expulsor retiró 914 unidades.

Cuatro unidades señaladas no fueron expulsadas por:

- desincronización;
- variación de separación entre envases;
- retraso neumático.

La eficacia del actuador fue:

$$[914] / [918] = 99{,}56\%$$

Este resultado incumple el criterio de:

$$99{,}9\%$$

Aunque el clasificador cumpliera las métricas globales, el sistema completo no cumpliría el criterio operacional.

La probabilidad de control efectivo depende de:

$$P(\text{detección}) \times P(\text{actuación})$$

Para los defectos globales:

$$0{,}96 \times 0{,}9956 \approx 95{,}58\%$$

La sensibilidad real del proceso extremo a extremo es inferior a la sensibilidad aislada del algoritmo.

25.13. Decisión de validación

No se aprueba para:

- sustitución del detector de cuerpos extraños;
- control exclusivo de objetos transparentes;
- liberación basada en ausencia de alerta;
- operación con condensación no controlada.

Se aprueba condicionalmente para:

- cierre incompleto;
- etiquetado grave;
- deformación;
- residuos visibles de sellado;
- rechazo automático de categorías expresamente validadas.

Condiciones

1. mantener detectores físicos independientes;
2. corregir y revalidar el expulsor;
3. instalar control de calidad de imagen;
4. activar parada ante suciedad o condensación;
5. mantener inspección específica para objetos transparentes;
6. verificar periódicamente unidades no rechazadas;
7. revalidar cualquier cambio de cámara o iluminación.

25.14. Conclusión del caso I

El sistema presenta una buena métrica global, pero no puede aprobarse sin restricciones.

La decisión demuestra el principio de no compensación:

$$\text{buen rendimiento medio} \not\Rightarrow \text{cumplimiento de categorías críticas}$$

También demuestra que la validación debe abarcar:

$$\text{captura} + \text{algoritmo} + \text{señal} + \text{actuador}$$

No únicamente el modelo informático.

26. Caso práctico II. Modelo de predicción de vida útil refrigerada

26.1. Finalidad del sistema

Una empresa fabrica cuatro referencias de producto refrigerado listo para consumo. La vida útil histórica se asigna mediante una fecha fija conservadora.

La empresa desarrolla un modelo que utiliza:

- formulación;
- pH;
- actividad de agua;
- tratamiento térmico;
- concentración de conservantes;
- carga inicial;
- temperatura registrada;
- tipo de envase.

El modelo estima:

[T días hasta alcanzar el criterio de fin de vida útil]

La finalidad propuesta es:

Apoyar la asignación de vida útil por lote dentro de las formulaciones y condiciones validadas, utilizando un límite predictivo conservador y un margen adicional. La predicción no sustituye los estudios microbiológicos, fisicoquímicos y sensoriales requeridos para establecer o modificar la vida útil comercial.

26.2. Definición del evento

El fin de vida útil se define como el primero de los siguientes eventos:

1. incumplimiento microbiológico;
2. rechazo sensorial;
3. pérdida de integridad del envase;
4. superación de una especificación fisicoquímica.

Por tanto:

[T_fin (T_micro, T_sensorial, T_envase, T_fisicoquímico)]

Esta definición compuesta evita que el modelo prediga únicamente estabilidad microbiológica mientras ignore un deterioro sensorial anterior.

26.3. Datos

Desarrollo

- 420 lotes;
- cuatro referencias;
- tres campañas;
- almacenamiento nominal y condiciones dinámicas;
- puntos longitudinales por lote.

Validación externa prospectiva

- 180 lotes independientes;
- campaña posterior;
- distribución entre las cuatro referencias;
- predicción congelada en el día de fabricación;
- seguimiento hasta el evento o fin del estudio.

La partición se realizó por lote completo. Ningún punto temporal del mismo lote se utilizó en conjuntos diferentes.

26.4. Definición del error

Se define:

$$[e_i T_i-T_i]$$

donde:

- (T_i) es la vida útil predicha;
- (T_i) es la observada.

Por tanto:

$$e_i>0$$

representa sobreestimación, potencialmente peligrosa.

Y:

$$e_i<0$$

representa infraestimación, generalmente conservadora pero económicamente desfavorable.

26.5. Criterios preespecificados

Sesgo

$$|ME|\leq 1 \text{ día}$$

Error absoluto

$$MAE\leq 3 \text{ días}$$

Cola del error

$$P_95(|e|)\leq 6 \text{ días}$$

Sobreestimación peligrosa

$$P(e>2 \text{ días})\leq 5\%$$

$$P(e>5 \text{ días})\leq 1{,}5\%$$

Intervalos predictivos

$$\text{Cobertura observada}_95\% \geq 92\%$$

Subgrupos

Cada referencia debe cumplir el límite de sobreestimación peligrosa.

Extrapolación

Predicciones fuera del dominio deberán bloquearse.

26.6. Resultados globales

En los 180 lotes externos se obtuvieron:

$$ME=-0,7 \text{ días}$$

$$MAE=2,4 \text{ días}$$

$$RMSE=3,2 \text{ días}$$

$$P\_95(|e|)=5,8 \text{ días}$$

$$P(e>2)=4,4\%$$

$$P(e>5)=1,1\%$$

$$\text{Cobertura del IP\_95\%}=93,9\%$$

Globalmente, el modelo cumple todos los criterios principales.

El sesgo negativo indica una ligera tendencia conservadora:

$$ME<0$$

Pero esta media no permite concluir que no existan sobreestimaciones.

26.7. Distribución de los errores

Intervalo de error	Lotes	Porcentaje
(e<-5) días	5	2,8 %
(-5≤ e<-2)	35	19,4 %
(-2≤ e≤2)	132	73,3 %
(2<e≤5)	6	3,3 %
(e>5)	2	1,1 %

El modelo ofrece una predicción dentro de ±2 días en aproximadamente tres cuartas partes de los lotes.

Sin embargo, existen dos lotes con sobreestimación superior a cinco días. Estos casos deben analizarse individualmente.

26.8. Rendimiento por referencia

Referencia	Lotes	MAE	ME	(P(e>2))	Cobertura IP
A	50	2,0 días	−0,8	2,0 %	96,0 %
B	45	2,3 días	−1,0	4,4 %	93,3 %
C	45	2,1 días	−0,6	2,2 %	95,6 %
D	40	3,6 días	+0,1	10,0 %	87,5 %

El promedio global oculta un incumplimiento en la referencia D.

Para D:

$$P(e>2)=10\%$$

frente al máximo establecido del 5 %.

Además:

$$\text{Cobertura}=87,5\%$$

inferior al mínimo del 92 %.

26.9. Investigación de la referencia D

La referencia D presenta:

- mayor contenido de grasa;
- emulsión más compleja;
- conservante parcialmente asociado a la fase lipídica;
- flora alterante distinta;
- mayor variabilidad del sellado.

El modelo había aprendido relaciones dominadas por las referencias A, B y C.

Aunque D estaba formalmente dentro de los rangos individuales de pH y actividad de agua, su combinación composicional estaba poco representada.

Esto demuestra que:

$$\text{dentro de mínimos y máximos} \not\Rightarrow \text{dentro del dominio multivariante}$$

26.10. Mecanismo de deterioro

En A, B y C el fin de vida útil se debe principalmente a:

- crecimiento de flora alterante;
- pérdida sensorial gradual.

En D, una fracción relevante de lotes falla antes por:

- inestabilidad de emulsión;
- interacción envase-producto;
- desarrollo localizado de levaduras.

El modelo general no representa adecuadamente este mecanismo.

No basta con recalibrar la probabilidad si la relación entre variables y evento es diferente.

26.11. Condiciones térmicas

Los lotes se evaluaron bajo:

- almacenamiento nominal;
- excursiones breves;
- perfiles dinámicos realistas.

El rendimiento fue:

Condición	MAE	Sobreestimación >2 días
Nominal	2,0 días	3,1 %
Dinámica habitual	2,6 días	4,7 %
Abuso razonablemente adverso	4,1 días	9,5 %

El modelo no cumple bajo abuso térmico adverso.

La empresa no debe interpretar esta limitación como irrelevante si ese perfil puede ocurrir en la cadena logística.

Las opciones son:

- restringir la utilización;
- aplicar un margen mayor;
- incorporar datos;

• desarrollar un modelo específico;
• activar abstención cuando el historial térmico exceda el dominio.

26.12. Intervalos predictivos

El modelo genera:

$$[TL,TU]$$

Un lote presenta:

$$T=35 \text{ días } IP_{95\%}=29\text{-}40 \text{ días}$$

No sería científicamente correcto asignar directamente 35 días como vida útil comercial.

La regla propuesta es:

$$[\ T_{operativa}\ T_{L}\text{-}M\ ]$$

donde:

• (T_L) es el límite inferior;
• (M) es el margen adicional.

Si:

$$M=2 \text{ días}$$

entonces:

$$[\ T_{operativa}\ \mathbf{29\text{-}2} \Rightarrow 27 \text{ días}\ ]$$

La estimación central informa, pero la decisión utiliza una aproximación conservadora.

26.13. Comparación con la vida útil fija

La práctica histórica asignaba:

$$T_{fija}=24 \text{ días}$$

para A, B y C.

El modelo, aplicando límite inferior y margen, permite una media de:

$$T_{operativa\ media}=27 \text{ días}$$

La ganancia media es:

$$27\text{-}24=3 \text{ días}$$

Pero el beneficio solo es admisible si no aumenta la tasa de sobreestimación peligrosa.

En A, B y C:

• se cumplen los criterios;
• se mantiene el margen;
• existe evidencia prospectiva.

En D no puede aplicarse la misma extensión.

26.14. Impacto económico ilustrativo

Supóngase una producción anual de:

$$2.000.000 \text{ unidades}$$

para A, B y C.

Si la ampliación prudente reduce en un 1,5 % el desperdicio por caducidad:

$$2.000.000 \times 0{,}015 = 30.000$$

unidades recuperadas.

Con un coste industrial de:

1,80 € por unidad

el valor bruto potencial sería:

[30.000×1,80 54.000 €]

Esta estimación debe descontar:

- validación;
- análisis;
- integración;
- monitorización;
- incidencias;
- coste de la infraestructura.

El ahorro no puede justificar la extensión si aumenta un riesgo sanitario o de incumplimiento.

26.15. Decisión de validación

Aprobado para:

- referencias A, B y C;
- rangos de formulación documentados;
- perfiles térmicos nominales y dinámicos validados;
- decisión mediante límite inferior y margen;
- utilización conjunta con estudios experimentales.

No aprobado para:

- referencia D;
- abuso térmico fuera del dominio;
- formulaciones nuevas;
- extensión automática de fecha;
- sustitución de estudios de vida útil.

Acciones para D

1. recopilar más lotes;
2. caracterizar el mecanismo específico;
3. incorporar variables de emulsión y envase;
4. desarrollar o ajustar un modelo específico;
5. realizar nueva validación externa.

26.16. Conclusión del caso II

El modelo cumple globalmente, pero no puede aprobarse de forma global.

La decisión correcta es:

aprobación parcial por dominio

y no:

aprobación de toda la familia comercial

El caso demuestra cuatro principios:

1. el error medio no controla la cola peligrosa;
2. la sobreestimación debe analizarse separadamente;
3. la cobertura de intervalos importa para decisiones individuales;
4. una formulación comercialmente similar puede seguir un mecanismo distinto.

27. Caso práctico III. Modelo de cribado microbiológico

27.1. Finalidad

Una empresa desarrolla un modelo que combina:

- proveedor;
- historial microbiológico;
- temperatura;
- pH;
- actividad de agua;
- incidencias de proceso;
- resultados ambientales;
- tiempo de espera;
- variables de limpieza.

El objetivo es estimar el riesgo de que un lote incumpla un criterio microbiológico definido.

La finalidad propuesta es:

Priorizar lotes para análisis microbiológico intensificado y revisión técnica, sin sustituir los análisis obligatorios ni autorizar la liberación automática de predicciones negativas.

La distinción es decisiva. El modelo se concibe como cribado y no como método confirmatorio.

27.2. Riesgo principal

El error crítico es el falso negativo:

$$FN = \text{lote positivo clasificado como bajo riesgo}$$

Un falso positivo provoca:

- análisis;
- retención;
- retraso;
- coste.

Un falso negativo puede provocar:

- liberación de un lote problemático;
- incumplimiento;
- retirada;
- riesgo sanitario.

Los errores son claramente asimétricos.

27.3. Datos

Desarrollo

- 18.000 lotes históricos;
- tres plantas;
- cuatro años;
- datos enriquecidos mediante investigación de incidencias.

Validación externa

- 3.000 lotes independientes;
- periodo posterior;
- dos plantas conocidas y una nueva;
- 150 positivos reales;
- 2.850 negativos.

La prevalencia del conjunto externo es:

$$[\ \pi\_test\ [150]\ /\ [3.000]\ 5\%\ ]$$

La prevalencia operativa habitual estimada es:

$$[\ \pi\_operativa\ 0\{,\}5\%\ ]$$

El conjunto externo fue deliberadamente enriquecido con lotes de alto riesgo para obtener suficientes positivos.

27.4. Criterios preespecificados

Detección

$$\text{Sensibilidad} \geq 98\%$$

Incertidumbre

$$\text{LCB\_95\%(sensibilidad)} \geq 95\%$$

Especificidad operacional

$$\text{Especificidad} \geq 88\%$$

Subgrupos

$$\text{Sensibilidad por planta} \geq 95\%$$

Dominio

- abstención para combinaciones no representadas;
- prohibición de clasificar como negativas entradas incompletas.

Verificación

- confirmación de alertas;
- muestreo aleatorio de predicciones negativas.

27.5. Resultado en el umbral de seguridad

La matriz fue:

Referencia	Riesgo alto	Riesgo bajo	Total
Positivo	147	3	150
Negativo	285	2.565	2.850
Total	432	2.568	3.000

Por tanto:

$$[\ \text{Sensibilidad}\ [147]\ /\ [150]\ 98\{,\}0\%\ ]$$

$$[\ \text{Especificidad}\ [2.565]\ /\ [2.850]\ 90\{,\}0\%\ ]$$

[VPP [147] / [147+285] 34{,}0%]

[VPN [2.565] / [2.565+3] 99{,}88%]

[Accuracy[147+2.565] / [3.000] 90{,}4%]

F1 ≈ 50,5% [Balanced accuracy 94{,}0%]

La accuracy del 90,4 % podría parecer inferior a la del modelo visual, pero el sistema cumple el criterio principal de sensibilidad.

La decisión no debe basarse en comparar cifras entre aplicaciones distintas.

27.6. Interpretación del valor predictivo

En la muestra enriquecida:

$$VPP=34,0\%$$

Esto significa que aproximadamente una de cada tres alertas corresponde a un positivo real.

Sin embargo, en producción la prevalencia es mucho menor.

Aplicando:

$$\mathbf{[\ VPP\ [\ S\pi\ ]\ /\ [\ S\pi+(1-E)(1-\pi)\ ]\ ]}$$

con:

$$S=0,98$$

$$E=0,90$$

$$\pi=0,005$$

se obtiene:

$$VPP \approx 4,69\%$$

En operación habitual, menos de una de cada veinte alertas correspondería a un positivo real.

El valor predictivo negativo sería:

$$VPN \approx 99,989\%$$

Pero este porcentaje no debe interpretarse como autorización automática de liberación, porque el volumen de producción puede convertir una tasa pequeña en un número relevante de falsos negativos.

27.7. Proyección por 100.000 lotes

Con prevalencia del 0,5 %:

$$N\_+=500$$

$$N\_-=99.500$$

Con sensibilidad del 98 %:

$$VP=490$$

$$FN=10$$

Con especificidad del 90 %:

$$VN=89.550$$

$$FP=9.950$$

El modelo generaría:

$$490+9.950=10.440$$

alertas por 100.000 lotes.

De ellas, solo 490 serían positivos reales.

Este volumen podría resultar operacionalmente inviable si cada alerta exige una retención prolongada y un análisis completo.

Al mismo tiempo, existirían diez falsos negativos por cada 100.000 lotes.

La organización debe decidir si:

- la tasa de falsos negativos es aceptable para un sistema de priorización;
- existe suficiente muestreo independiente;
- el volumen de confirmaciones es viable;
- puede añadirse una segunda etapa.

27.8. Umbral alternativo

Se evalúa un umbral más alto que produce:

$$\text{Sensibilidad}=94\%$$

$$\text{Especificidad}=97\%$$

A prevalencia del 0,5 %, por cada 100.000 lotes:

Positivos

$$VP=470$$

$$FN=30$$

Negativos

$$VN=96.515$$

$$FP=2.985$$

Las alertas disminuirían a:

$$470+2.985=3.455$$

Pero los falsos negativos aumentarían de diez a treinta.

La mejora operacional se obtiene a costa de triplicar los positivos omitidos.

El umbral alternativo incumple el criterio de sensibilidad preespecificado.

No debe elegirse únicamente porque reduce el coste.

27.9. Estrategia en dos etapas

Para mejorar la viabilidad se propone:

Etapa 1. Modelo de IA

Identifica aproximadamente el 10,4 % de lotes como alto riesgo.

Etapa 2. Prueba rápida o revisión técnica

Se aplica a las alertas para reducir falsos positivos antes del análisis confirmatorio completo.

El rendimiento combinado dependerá de ambas etapas.

Si la segunda etapa tiene:

$$S_2=98\%$$

la sensibilidad total aproximada será:

$$[S_{total} S_1\times S_2]$$

$$[S_{total} 0{,}98\times0{,}98\ 96\{,\}04\%]$$

La segunda etapa puede reducir falsas alarmas, pero también introducir nuevos falsos negativos.

No debe asumirse que añadir controles siempre aumenta la sensibilidad. Los sistemas secuenciales multiplican las probabilidades de detección cuando ambos deben ser positivos.

Una arquitectura más segura podría utilizar:

- IA para priorización;
- análisis obligatorio según plan base;
- análisis adicional en alto riesgo;

en lugar de utilizar la IA para eliminar controles.

27.10. Rendimiento por planta

Planta	Positivos	Sensibilidad	Especificidad
Planta A	60	100,0 %	91,0 %
Planta B	55	98,2 %	90,5 %
Planta C, nueva	35	94,3 %	87,5 %

La planta C incumple:

$$\text{Sensibilidad mínima local}=95\%$$

y:

$$\text{Especificidad mínima}=88\%$$

La planta nueva presenta:

- proveedor local distinto;
- variable ambiental no disponible;
- mayor frecuencia de datos ausentes;
- distribución de pH desplazada.

El modelo no puede aprobarse para C basándose en el resultado agregado.

27.11. Datos ausentes

En la planta C falta una variable ambiental en el 18 % de los lotes.

El sistema la imputó mediante la mediana histórica.

Esta imputación redujo artificialmente la variabilidad y produjo puntuaciones próximas al riesgo medio.

La validación demuestra que el dato ausente no debe tratarse silenciosamente.

La regla revisada será:

$$\text{variable crítica ausente} \Rightarrow \text{abstención}$$

y no:

$$\text{variable crítica ausente} \Rightarrow \text{imputación automática}$$

27.12. Calibración

Globalmente se obtiene:

$$\alpha=-0,18$$

$$\beta=0,81$$

El intercepto negativo indica cierta sobreestimación global del riesgo.

La pendiente inferior a uno indica probabilidades excesivamente extremas.

Por planta:

Planta	Intercepto	Pendiente
A	−0,05	0,96
B	−0,12	0,88
C	−0,46	0,61

La calibración de C es claramente deficiente.

No debe comunicarse la salida de la planta C como probabilidad real.

27.13. Validación temporal

Durante los primeros tres meses, la sensibilidad se mantiene.

En el cuarto mes aumenta la prevalencia y cambia un proveedor.

Se observa:

- incremento de puntuaciones;
- mayor tasa de alertas;
- pérdida de especificidad;
- aumento de entradas fuera de dominio.

El sistema debe activar una investigación.

No debe modificarse automáticamente el umbral para restaurar el número histórico de alertas, porque el incremento puede reflejar un deterioro real del proceso.

27.14. Muestreo de negativos

Si solo se confirmaran las 432 alertas de la validación, los tres falsos negativos no se detectarían.

El protocolo mantiene:

- confirmación de todos los positivos del modelo;
- muestra aleatoria de predicciones negativas;
- sobremuestreo de negativos próximos al umbral;
- controles por planta y proveedor.

Este diseño permite estimar la tasa real de omisión.

27.15. Decisión de validación

Planta A

Aprobada para priorización, con:

- análisis confirmatorio;

- mantenimiento del plan base;
- muestreo de negativos;
- monitorización mensual.

Planta B

Aprobación condicional:

- misma finalidad;
- revisión de calibración;
- vigilancia intensificada;
- evaluación adicional de subgrupos.

Planta C

No aprobada para operación.

Se requiere:

1. completar variables críticas;
2. recopilar datos locales;
3. recalibrar o reentrenar;
4. congelar nueva versión;
5. realizar validación independiente.

Uso prohibido en todas las plantas

- liberación automática de lotes negativos;
- sustitución del método microbiológico;
- reducción de controles obligatorios sin validación específica;
- utilización de probabilidades no calibradas como riesgo absoluto.

27.16. Conclusión del caso III

El modelo cumple las métricas globales de sensibilidad y especificidad, pero presenta dos problemas:

1. carga de falsas alarmas elevada en prevalencia real;
2. pérdida de transportabilidad en una planta nueva.

La conclusión no debe ser:

El modelo tiene un VPN del 99,99 % y puede liberar producto.

Debe ser:

El modelo aporta valor para priorizar controles en las plantas validadas, pero no sustituye la confirmación ni el plan independiente de muestreo. Su utilización en la planta nueva queda suspendida hasta completar una validación local.

28. Comparación transversal de los tres casos

Dimensión	Visión artificial	Vida útil	Cribado microbiológico
Tipo de salida	Clasificación	Regresión/tiempo	Probabilidad/clasificación
Error crítico	Falso negativo crítico	Sobreestimación	Falso negativo
Métrica primaria	Sensibilidad por defecto	Sobreestimación peligrosa	Sensibilidad
Referencia	Inspección adjudicada	Evento experimental	Método microbiológico
Unidad	Envase	Lote/curva	Lote
Riesgo de prevalencia	Moderado	No principal	Muy elevado
Dominio crítico	Condiciones visuales	Formulación y temperatura	Planta, proveedor y datos
Barrera	Controles físicos	Margen y estudios	Confirmación y muestreo
Fallo oculto	Defecto específico	Referencia D	Planta C
Decisión	Aprobación restringida	Aprobación por producto	Aprobación por planta

La tabla muestra que no existe una métrica universal.

Para visión artificial, la sensibilidad debe desglosarse por defecto.

Para vida útil, el promedio debe complementarse con la dirección y cola del error.

Para cribado microbiológico, la prevalencia transforma radicalmente los valores predictivos y la carga operacional.

29. Lecciones metodológicas derivadas

29.1. La métrica global puede ocultar el fallo decisivo

En el caso visual:

$$S_{global}=96\%$$

pero:

$$S_{cuerpo\ extraño}=85{,}7\%$$

En vida útil:

$$P(e>2)=4{,}4\%$$

globalmente, pero:

$$P_D(e>2)=10\%$$

En microbiología:

$$S_{global}=98\%$$

pero:

$$S_C=94{,}3\%$$

La aprobación debe depender del subgrupo crítico, no únicamente del promedio.

29.2. La prevalencia cambia la utilidad

El modelo microbiológico tiene:

$$VPP=34\%$$

en una muestra con 5 % de positivos, pero aproximadamente:

$$VPP=4{,}69\%$$

con prevalencia del 0,5 %.

El rendimiento discriminativo no ha cambiado. Ha cambiado el significado práctico de la alerta.

29.3. El sistema completo puede rendir peor que el modelo

En visión artificial, el clasificador presenta una sensibilidad del 96 %, pero el expulsor no actúa en todos los casos.

La validación debe abarcar:

$$\text{modelo} + \text{integración} + \text{acción}$$

29.4. La similitud comercial no garantiza equivalencia científica

La referencia D pertenece a la misma familia comercial, pero presenta un mecanismo de deterioro diferente.

Las categorías comerciales no sustituyen la caracterización:

- composicional;
- microbiológica;
- tecnológica.

29.5. La abstención es una función de seguridad

En los tres casos existen condiciones en las que el sistema no debería decidir:

- imagen degradada;
- perfil térmico fuera de dominio;
- variable crítica ausente;
- planta no validada.

El comportamiento seguro es:

$$\text{no sé} \Rightarrow \text{procedimiento alternativo}$$

29.6. Recalibrar no resuelve todos los fallos

La recalibración puede corregir probabilidades sesgadas.

No resuelve:

- un defecto visual que el modelo no detecta;
- un mecanismo de deterioro diferente;
- una planta con datos esenciales ausentes;
- una discriminación insuficiente.

29.7. La aprobación puede ser parcial

Las posibles decisiones son:

- por defecto;
- por referencia;
- por planta;
- por equipo;
- por condición;
- por grado de autonomía.

La aprobación parcial no representa un fracaso. Representa una correspondencia exacta entre evidencia y autorización.

30. Modelo de acta de decisión para los casos prácticos

30.1. Identificación

Sistema:
Versión:
Fecha:
Responsable:
Proveedor:

30.2. Finalidad

Variable o clase:
Decisión influida:
Nivel de autonomía:
Método de confirmación:

30.3. Dominio autorizado

Productos:
Plantas:
Equipos:
Proveedores:
Rangos:
Condiciones:

30.4. Evidencia

Unidades independientes:
Positivos o eventos:
Diseño externo:
Periodo:
Métricas principales:
Intervalos:
Subgrupos:
Robustez:

30.5. Criterios

Criterio	Límite	Resultado	Estado
Métrica crítica 1	—	—	Cumple/no cumple
Métrica crítica 2	—	—	—
Subgrupo	—	—	—
Robustez	—	—	—
Operación	—	—	—

30.6. Limitaciones

30.7. Controles obligatorios

- confirmación;
- muestreo independiente;
- supervisión;
- abstención;
- sistema alternativo;
- monitorización.

30.8. Decisión

- no aprobado;
- investigación;
- modo sombra;
- apoyo;
- uso restringido;
- aprobación condicional;
- aprobación completa.

30.9. Criterios de suspensión

- falso negativo crítico;
- incumplimiento de sensibilidad;
- cambio de versión;
- pérdida de calibración;
- entrada fuera de dominio;
- fallo técnico;
- cambio de proceso.

30.10. Próxima revisión

Fecha o evento activador:

31. Conclusión general de los casos prácticos

Los tres supuestos demuestran que validar un modelo de IA no consiste en calcular una métrica y compararla con un porcentaje genérico.

La decisión científicamente correcta exige identificar:

qué error importa

en qué subgrupo aparece

con qué frecuencia

con qué consecuencia

qué barrera lo controla

El modelo visual no se aprueba para todos los defectos porque falla precisamente en una categoría crítica.

El modelo de vida útil no se extiende a toda la familia porque una formulación sigue un mecanismo de deterioro diferente.

El modelo microbiológico no se utiliza para liberar lotes porque su elevado valor predictivo negativo no elimina los falsos negativos y porque la prevalencia real genera una carga sustancial de falsas alarmas.

En los tres casos, la conclusión adecuada adopta una forma condicionada:

La versión evaluada puede utilizarse únicamente para la finalidad, productos, plantas, equipos y condiciones en los que ha demostrado un rendimiento compatible con los criterios preestablecidos, manteniendo las barreras y verificaciones independientes requeridas.

Esta es la diferencia entre un modelo con buenos resultados y un sistema científicamente validado.

El primero genera predicciones convincentes.

El segundo dispone de evidencia suficiente para saber cuándo esas predicciones pueden utilizarse, cuándo deben confirmarse y cuándo deben rechazarse.

32. Discusión general: de la exactitud algorítmica a la validez científica y operacional

La incorporación de inteligencia artificial a la industria alimentaria está modificando la forma en que se observan, predicen y controlan procesos complejos. Los modelos actuales pueden integrar simultáneamente señales visuales, espectrales, microbiológicas, fisicoquímicas, térmicas, logísticas y productivas que difícilmente podrían ser interpretadas con la misma velocidad mediante procedimientos convencionales.

Sin embargo, esta capacidad computacional no elimina las exigencias del método científico. Las hace más necesarias.

Cuanto mayor es la complejidad del modelo, más difícil resulta determinar:

- qué patrones ha aprendido;
- si esos patrones son generalizables;
- qué condiciones alteran su comportamiento;
- cuándo una salida representa una probabilidad real;
- qué ocurre ante datos no representados;
- cuánto puede degradarse tras el despliegue;
- qué consecuencias produce cada tipo de error.

La cuestión central no es si la inteligencia artificial puede alcanzar métricas elevadas. Numerosos modelos pueden hacerlo dentro de bases históricas cuidadosamente construidas. La cuestión es si esas métricas representan una capacidad estable, reproducible y compatible con una decisión alimentaria concreta.

La validación científica debe impedir que se produzca una sustitución conceptual indebida:

buen ajuste a los datos validez científica

validez científica utilidad industrial

utilidad industrial seguridad jurídica o alimentaria

Cada una de estas afirmaciones requiere evidencia adicional.

32.1. El modelo no es la unidad completa de validación

Uno de los resultados más importantes del análisis desarrollado es que la unidad real de validación no debería ser el algoritmo aislado.

El sistema relevante está constituido por:

$$[\ D+M+H+P+I+A]$$

donde:

- (D) representa los datos;
- (M), el modelo;
- (H), la intervención humana;
- (P), los procedimientos;
- (I), la infraestructura;
- (A), la acción generada.

Un clasificador puede ser estadísticamente correcto y el sistema completo fallar porque:

- el sensor genera una señal degradada;
- la variable de entrada se registra en unidades incorrectas;
- el usuario interpreta mal la salida;

- la alerta llega demasiado tarde;
- el expulsor no actúa;
- el procedimiento no define qué hacer;
- el modelo se utiliza fuera del dominio;
- una actualización modifica silenciosamente el umbral.

Por tanto, la validación debe abarcar la cadena completa:

muestra arrow medición arrow dato arrow predicción arrow interpretación arrow decisión arrow acción arrow resultado

La exactitud del algoritmo solo representa un eslabón.

32.2. La finalidad prevista determina la evidencia exigible

La misma arquitectura puede requerir niveles de validación radicalmente distintos según su uso.

Un modelo puede emplearse para:

- explorar datos;
- generar hipótesis;
- visualizar tendencias;
- priorizar muestras;
- recomendar una revisión;
- retener producto;
- asignar una fecha;
- liberar un lote.

El rendimiento necesario aumenta con:

- la gravedad del error;
- la autonomía;
- la exposición;
- la irreversibilidad;
- la ausencia de barreras.

Puede formularse:

$$E_{requerida} \propto S \times A \times X \times (1-B)$$

donde:

- (S) es la severidad;
- (A), la autonomía;
- (X), la exposición;
- (B), la eficacia de las barreras.

Un modelo con rendimiento moderado puede aportar valor en investigación o priorización. El mismo modelo puede resultar inaceptable para una liberación automática.

No existe, por tanto, una categoría abstracta de «modelo validado». Existe una versión validada para una finalidad, un dominio y un nivel de autonomía determinados.

32.3. La validación no puede reducirse a una métrica

La popularidad de métricas como accuracy, AUC o (R^2) ha favorecido una simplificación excesiva.

Estas medidas pueden ser útiles, pero ninguna responde por sí sola a las preguntas críticas.

Accuracy

Puede quedar dominada por la clase mayoritaria.

AUC

Evalúa discriminación global, pero no el rendimiento en el umbral real.

F1

Combina precision y recall, pero no considera verdaderos negativos ni gravedad diferencial.

(R^2)

Mide variación explicada, pero no concordancia ni magnitud operativa del error.

RMSE

Penaliza errores grandes, pero no identifica su dirección.

MAE

Resume la magnitud media, pero puede ocultar una cola peligrosa.

La validación debe utilizar una arquitectura de métricas:

[discriminación, calibración, error, incertidumbre, robustez, impacto]

y no una cifra única.

32.4. La dirección del error es tan importante como su magnitud

En muchas aplicaciones alimentarias los errores son asimétricos.

Vida útil

T>T

puede ser más peligroso que:

T<T

Crecimiento microbiológico

N<N

puede infravalorar el riesgo.

Inactivación

R> R

puede sobreestimar la eficacia del tratamiento.

Cribado

FN

puede ser más grave que:

FP

La utilización de errores simétricos puede generar una falsa neutralidad.

Un modelo que sobreestima peligrosamente unos lotes e infraestima de forma conservadora otros puede presentar:

$$ME \approx 0$$

porque ambos errores se compensan.

La seguridad no debe evaluarse mediante compensación estadística entre consecuencias incompatibles.

32.5. La representatividad es una condición causal de la validez

El modelo aprende de la distribución observada.

Su capacidad depende de que los datos representen:

- productos;
- plantas;
- proveedores;
- campañas;
- equipos;
- operadores;
- condiciones;
- prevalencias;
- casos frontera;
- mecanismos de deterioro.

Una base puede ser grande y científicamente pobre si contiene:

- muchas réplicas de pocos lotes;
- imágenes similares;
- una sola campaña;
- defectos evidentes;
- una única planta;
- condiciones controladas.

El volumen digital no equivale a diversidad experimental.

Debe distinguirse:

$$n\_registros$$

de:

$$n\_unidades\ independientes$$

y de:

$$n\_contextos\ representados$$

Un millón de espectros de cincuenta lotes no demuestra la misma generalización que varios miles de lotes independientes distribuidos entre plantas, campañas y proveedores.

32.6. La fuga de datos es una amenaza estructural

La fuga no es únicamente un error técnico. Es una amenaza a la validez epistemológica del estudio.

Cuando información del conjunto de prueba influye en:

- selección de variables;
- normalización;
- hiperparámetros;

• umbral;
• arquitectura;
• exclusiones;
• elección final;
la evaluación deja de representar una predicción futura.

El modelo puede parecer excepcional porque ha aprendido características específicas de aquello que debía desconocer.

En alimentación, la fuga aparece con frecuencia por:

• réplicas técnicas;
• puntos temporales del mismo lote;
• imágenes aumentadas;
• espectros de la misma muestra;
• datos históricos posteriores;
• identificadores de planta o proveedor.

La partición debe realizarse conforme a la unidad real de generalización.

32.7. La calibración es imprescindible cuando se habla de riesgo

Un score no calibrado puede ordenar correctamente los casos, pero no representar una probabilidad.

Si una salida de 0,80 se presenta como «80 % de riesgo», debe demostrarse que:

$$P(Y=1|p=0{,}80) \approx 0{,}80$$

La calibración puede variar por:

• planta;
• prevalencia;
• campaña;
• producto;
• periodo;
• subgrupo.

Un modelo globalmente calibrado puede no estarlo localmente.

Cuando la probabilidad influye en:

• priorización;
• retención;
• muestreo;
• asignación de vida útil;
• coste esperado;
la calibración constituye una propiedad esencial.

32.8. Reconocer la ignorancia es una función de seguridad

Un sistema fiable no debería emitir siempre una conclusión.

Debe poder reconocer:

• datos incompletos;
• imagen degradada;

- combinación desconocida;
- producto no validado;
- sensor incompatible;
- incertidumbre elevada;
- observación fuera de distribución.

La abstención transforma:

$$\text{incertidumbre oculta}$$

en:

$$\text{acción controlada}$$

La organización debe definir qué ocurre después:

$$\text{abstención} \Rightarrow \text{retención, revisión o método alternativo}$$

Una abstención sin procedimiento carece de valor operativo.

32.9. La validación externa debe probar una distancia relevante

No basta con denominar externo a un conjunto separado aleatoriamente.

Debe preguntarse:

¿Externo respecto a qué?

La validación puede ser externa respecto a:

- tiempo;
- planta;
- equipo;
- proveedor;
- campaña;
- formulación;
- laboratorio;
- región.

La transportabilidad no es una propiedad indivisible.

Un modelo puede generalizar temporalmente y fallar entre instrumentos. Puede funcionar entre plantas y degradarse con una nueva formulación.

La declaración de validez debe especificar la dimensión evaluada.

32.10. Los subgrupos limitan la autorización

El rendimiento agregado puede ocultar:

- una referencia débil;
- una planta con sensibilidad insuficiente;
- un defecto crítico no detectado;
- una campaña problemática;
- un proveedor no representado.

Cuando un subgrupo crítico incumple, la respuesta científicamente correcta puede ser:

- restringir;

- recalibrar;
- reentrenar;
- ampliar datos;
- rechazar.

No debe permitirse que el volumen de los grupos mayoritarios compense el riesgo de una minoría relevante.

Puede aplicarse:

$$[M\_autorización (M\_1,,M\_K)]$$

cuando la seguridad exige desempeño mínimo en todos los grupos.

32.11. El método de referencia también tiene incertidumbre

La referencia puede fallar por:

- muestreo;
- heterogeneidad;
- límite de detección;
- variación analítica;
- desacuerdo de expertos;
- error de etiquetado.

Por ello, la comparación debe caracterizar tanto:

$$U\_modelo$$

como:

$$U\_referencia$$

Sin embargo, reconocer la imperfección de la referencia no autoriza a reinterpretar todas las discordancias en favor del algoritmo.

Las discrepancias deben investigarse mediante reglas ciegas y preespecificadas.

32.12. La IA debe demostrar beneficio incremental

La complejidad no es un valor en sí misma.

Un modelo debería compararse con:

- regla simple;
- procedimiento actual;
- inspector;
- modelo estadístico;
- microbiología predictiva convencional;
- fecha fija.

La pregunta no es si el modelo es técnicamente sofisticado, sino si aporta:

- mejor detección;
- menor error;
- mayor cobertura;
- menor tiempo;
- menos desperdicio;
- mayor consistencia;

- mejor anticipación.

Si una regla sencilla alcanza prestaciones equivalentes, puede resultar preferible por:

- auditabilidad;
- estabilidad;
- menor mantenimiento;
- menor coste;
- menor riesgo de deriva.

32.13. La validación inicial es temporalmente limitada

El entorno cambia.

Pueden cambiar:

- materias primas;
- proveedores;
- formulación;
- microbiota;
- sensores;
- métodos;
- usuarios;
- prevalencia;
- proceso.

Por ello:

$$P_t(X,Y) \neq P_{t+1}(X,Y)$$

La monitorización debe observar:

- calidad de datos;
- distribución de entradas;
- distribución de salidas;
- calibración;
- métricas;
- falsos negativos;
- abstenciones;
- impacto.

Una validación histórica sin vigilancia posterior constituye una evidencia progresivamente debilitada.

32.14. La actualización no es una operación neutra

Modificar:

- pesos;
- umbral;
- preprocesamiento;
- variables;
- equipo;
- firmware;

• finalidad;

puede invalidar la evidencia anterior.

La organización debe mantener:

$$\text{modelo activo} \neq \text{modelo candidato}$$

El candidato puede aprender y evolucionar. El activo debe permanecer identificado, controlado y validado hasta que una nueva versión sea aprobada.

32.15. La supervisión humana debe ser real

La presencia de una persona no garantiza supervisión.

La persona debe:

• comprender la salida;
• conocer los límites;
• recibir información contextual;
• disponer de tiempo;
• poder anular;
• tener autoridad;
• registrar la decisión.

El sistema humano-IA debe validarse como conjunto.

Puede ocurrir que:

$$M_{humano+IA} < M_{IA}$$

o:

$$M_{humano+IA} < M_{humano}$$

si aparece:

• sesgo de automatización;
• fatiga de alertas;
• interfaz confusa;
• dependencia excesiva.

32.16. La integración en APPCC no puede ser ornamental

Cuando el modelo influye en:

• peligros;
• muestreo;
• límites;
• monitorización;
• acciones;
• liberación;

debe incorporarse al sistema de gestión.

Eso implica:

• análisis de modos de fallo;
• procedimiento;

- registros;
- acciones correctivas;
- verificación;
- revalidación;
- contingencia.

La IA no debe aparecer únicamente en una presentación corporativa mientras permanece fuera del APPCC y del control documental.

32.17. La responsabilidad no se delega al algoritmo

El operador decide:

- adquirir;
- integrar;
- confiar;
- automatizar;
- modificar;
- mantener;
- suspender.

El proveedor puede responder por el producto tecnológico, pero la organización alimentaria conserva sus obligaciones respecto del producto que fabrica o comercializa.

La afirmación:

> *«El algoritmo lo autorizó»*

no constituye una defensa técnica suficiente.

La pregunta será:

¿Era razonable atribuir a esa salida el peso que recibió?

La respuesta dependerá de:

- validación;
- finalidad;
- limitaciones;
- supervisión;
- controles;
- registros.

32.18. La documentación es parte de la validez

Un resultado no documentado difícilmente puede:

- reproducirse;
- auditarse;
- explicarse;
- vincularse a una versión;
- investigarse tras un incidente.

El expediente debería conservar:

$$\text{datos} + \text{protocolo} + \text{modelo} + \text{resultado} + \text{decisión} + \text{seguimiento}$$

La documentación no convierte un sistema deficiente en válido, pero la ausencia de documentación impide demostrar que uno correcto fue utilizado de manera diligente.

33. Recomendaciones para la industria alimentaria

33.1. Comenzar por la decisión, no por el algoritmo

Antes de seleccionar una tecnología debe definirse:

- qué decisión se pretende mejorar;
- qué error resulta crítico;
- qué evidencia existe;
- qué barreras deben mantenerse;
- cuánto riesgo residual es aceptable.

La secuencia correcta es:

$$\text{problema} \Rightarrow \text{finalidad} \Rightarrow \text{riesgo} \Rightarrow \text{datos} \Rightarrow \text{modelo}$$

No:

$$\text{modelo disponible arrow buscar una aplicación}$$

33.2. Definir la finalidad con precisión jurídica y tecnológica

La declaración debería incluir:

- producto;
- planta;
- equipo;
- variable;
- momento;
- usuario;
- acción;
- exclusiones.

Una finalidad ambigua genera una validación ambigua.

33.3. Crear un comité multidisciplinar

El comité debería incluir:

- calidad;
- seguridad alimentaria;
- tecnología;
- microbiología;
- datos;
- estadística;
- producción;
- mantenimiento;
- tecnologías de la información;
- jurídico;
- dirección.

Ninguna disciplina posee por sí sola toda la información necesaria.

33.4. Clasificar la criticidad antes del desarrollo

Debe establecerse si el sistema es:

- exploratorio;
- informativo;
- de cribado;
- decisor de calidad;
- influyente en seguridad.

Esta clasificación determinará:

- tamaño;
- diseño;
- independencia;
- métricas;
- barreras;
- vigilancia.

33.5. Definir la unidad estadística

Debe identificarse si la unidad es:

- lote;
- envase;
- muestra;
- curva;
- planta;
- campaña.

Las réplicas no deben inflar la evidencia.

33.6. Reservar un conjunto verdaderamente independiente

El conjunto final debe permanecer bloqueado y no utilizarse para:

- ajustar;
- seleccionar;
- recalibrar;
- decidir umbrales;
- eliminar casos.

Cuando se consulte repetidamente, deja de ser una prueba (test) final.

33.7. Preespecificar los criterios de aceptación

Los criterios deben fijarse antes de conocer los resultados.

Deben incluir:

- métrica principal;
- límite;
- intervalo;
- subgrupos;

- robustez;
- abstención;
- suspensión.

33.8. Priorizar el error crítico

En seguridad y vida útil debe identificarse la dirección peligrosa.

No debe aceptarse que una métrica promedio compense:

- falsos negativos;
- sobreestimaciones;
- fallos extremos.

33.9. Exigir calibración cuando se comunican probabilidades

Si el sistema no está calibrado, la salida debe denominarse puntuación.

No debe utilizarse como riesgo absoluto.

33.10. Incorporar detección fuera de dominio

Deben existir:

- límites físicos;
- límites multivariantes;
- control de datos faltantes;
- advertencia;
- abstención.

La interfaz debe impedir que una entrada desconocida produzca una decisión automática aparentemente fiable.

33.11. Validar el sistema completo

Debe comprobarse:

- sensor;
- datos;
- integración;
- algoritmo;
- interfaz;
- usuario;
- actuación;
- respaldo.

33.12. Mantener controles independientes

Durante las fases iniciales y en aplicaciones críticas deben mantenerse:

- análisis;
- muestreo;
- inspección;
- límites físicos;

- revisión humana.

La IA puede reforzar el control antes de sustituirlo.

33.13. Desplegar progresivamente

La secuencia recomendada es:

1. investigación;
2. retrospectiva;
3. test externo;
4. modo sombra;
5. apoyo;
6. automatización limitada;
7. autorización completa si procede.

33.14. Implantar monitorización desde el primer día

Debe vigilarse:

- entradas;
- salidas;
- rendimiento;
- deriva;
- abstenciones;
- incidentes;
- anulaciones;
- impacto.

No debe esperarse a que aparezca una reclamación para evaluar la estabilidad.

33.15. Controlar versiones y actualizaciones

Toda versión debe disponer de:

- identificador;
- historial;
- evidencia;
- fecha;
- umbral;
- entorno.

Las actualizaciones automáticas deben estar sujetas a control.

33.16. Preparar un procedimiento de suspensión

Debe definirse qué eventos provocan:

- alerta;
- restricción;
- recalibración;
- revalidación;
- suspensión.

El sistema alternativo debe estar disponible.

33.17. Auditar al proveedor

La compra debería exigir:

- datos de validación;
- unidad estadística;
- falsos negativos;
- limitaciones;
- versiones;
- actualizaciones;
- logs;
- cooperación;
- continuidad.

33.18. Evitar claims absolutos

No deberían utilizarse expresiones como:

- infalible;
- universal;
- garantiza;
- cero riesgo;
- sustituye completamente;
- 100 % fiable.

Las afirmaciones deben vincularse a la evidencia.

33.19. Conservar la trazabilidad de cada decisión crítica

Para cada lote debería poder conocerse:

- versión;
- datos;
- salida;
- incertidumbre;
- usuario;
- acción;
- resultado posterior.

33.20. Revisar periódicamente la finalidad

Un sistema puede comenzar como apoyo y terminar utilizándose de facto como decisor.

La empresa debe comprobar si el uso real coincide con el autorizado.

34. Agenda de investigación pendiente

La validación de IA alimentaria continúa siendo un campo en desarrollo. Existen varias áreas que requieren mayor investigación.

34.1. Estándares sectoriales específicos

La industria necesita protocolos que adapten la validación de IA a:

- microbiología;
- visión;
- espectroscopia;
- vida útil;
- control de proceso;
- sensorial.

Los marcos generales de IA resultan útiles, pero no sustituyen criterios específicos de matrices, peligros y decisiones alimentarias.

34.2. Tamaño muestral para eventos extremadamente raros

Cuando la prevalencia es muy baja, resulta difícil demostrar tasas mínimas de falsos negativos mediante estudios convencionales.

Se requieren enfoques que combinen:

- datos reales;
- enriquecimiento;
- simulación;
- evidencia mecanística;
- vigilancia secuencial;
- métodos bayesianos.

34.3. Incertidumbre en modelos multimodales

Los sistemas que integran imágenes, sensores, texto, historial y laboratorio plantean preguntas sobre:

- propagación de incertidumbre;
- ausencia de modalidades;
- contradicción entre fuentes;
- contribución de cada señal.

34.4. Predicción conformal bajo deriva

Las garantías conformales dependen de condiciones de intercambiabilidad que pueden romperse en producción.

Es necesario desarrollar y validar mecanismos adaptativos que mantengan cobertura sin ocultar cambios conceptuales.

34.5. Evaluación causal

Muchos modelos identifican correlaciones útiles, pero no permiten afirmar qué intervención mejorará el proceso.

La combinación de:

- inferencia causal;
- modelos mecanísticos;
- experimentos;
- IA;

puede mejorar la capacidad de decidir, no únicamente de predecir.

34.6. Validación de modelos fundacionales

La incorporación de modelos generativos y fundacionales a:

- documentación;
- auditoría;
- interpretación de incidentes;
- formulación;
- soporte técnico;

introduce riesgos diferentes de los modelos predictivos cerrados.

Debe evaluarse:

- alucinación;
- variabilidad;
- trazabilidad de fuentes;
- contaminación de contexto;
- seguridad de la información;
- reproducibilidad.

34.7. Interacción humano-IA

Se necesita más evidencia sobre:

- automatización;
- confianza;
- fatiga;
- competencia;
- diseño de interfaz;
- responsabilidad compartida.

34.8. Benchmarks industriales realistas

Los conjuntos públicos suelen simplificar las condiciones.

La industria necesita benchmarks:

- multicéntricos;
- longitudinales;
- con prevalencia real;
- con casos difíciles;
- con cambios de dominio;
- con actuación física.

34.9. Comparación con modelos mecanísticos

Los modelos híbridos pueden combinar:

$$\text{conocimiento científico} + \text{aprendizaje automático}$$

Esto puede mejorar:

- extrapolación;
- coherencia;
- interpretabilidad;
- eficiencia de datos.

Pero requiere protocolos específicos para validar tanto la parte mecanística como la aprendida.

34.10. Evidencia posdespliegue

Debe avanzarse hacia sistemas de vigilancia comparables a los utilizados en otras tecnologías reguladas:

- incidentes;
- alertas;
- tendencia;
- actualización;
- aprendizaje compartido;
- retirada de versiones.

35. Conclusiones

La inteligencia artificial puede aportar un valor extraordinario a la gestión de calidad, vida útil y seguridad alimentaria. Puede anticipar desviaciones, integrar señales complejas, mejorar la cobertura, reducir tiempos y permitir decisiones más ajustadas a la variabilidad real de los productos.

Pero esta capacidad no debe confundirse con fiabilidad demostrada.

La validación científica exige responder, de manera documentada, a las siguientes cuestiones:

1. Qué predice el sistema.
2. Para qué decisión se utilizará.
3. Cuál es la unidad real de análisis.
4. Qué referencia define la verdad operacional.
5. Si los datos representan el dominio.
6. Si la prueba es realmente independiente.
7. Qué errores comete y en qué dirección.
8. Cuál es su incertidumbre.
9. Cómo funciona por subgrupos.
10. Qué ocurre fuera del dominio.
11. Cómo responde ante perturbaciones y fallos.
12. Qué beneficio aporta frente a la práctica existente.
13. Qué barreras permanecen.
14. Cómo se monitorizará la deriva.
15. Qué evidencia permite reconstruir cada decisión.

La validación no debe concluir con una frase genérica como:

«El modelo ha sido validado.»

Debe concluir con una autorización condicionada:

La versión identificada ha demostrado un rendimiento compatible con la finalidad prevista dentro del dominio, las condiciones y el grado de autonomía especificados, con las limitaciones, controles, procedimientos de verificación y criterios de suspensión documentados.

Esta formulación expresa la verdadera naturaleza de la evidencia.

Un modelo nunca se valida en abstracto.

Se valida:

una versión

para:

una finalidad

dentro de:

un dominio

bajo:

unas condiciones

y con:

unas barreras

La principal amenaza no es que la inteligencia artificial cometa errores. Todo sistema los comete.

La amenaza aparece cuando:

- no se conocen;
- no se cuantifican;
- se comunican de forma incompleta;
- se utilizan fuera del dominio;
- no existe respuesta ante ellos;
- se mantienen ocultos por una métrica global favorable.

La validación científica no transforma la predicción en certeza. Transforma una incertidumbre opaca en una incertidumbre medida, limitada y gobernable.

Ese es el umbral que separa la experimentación algorítmica de una aplicación industrial responsable.

35.1. Tesis final

La validez de un modelo de IA alimentaria no reside en su complejidad, en su novedad ni en la magnitud de la base de datos.

Reside en la coherencia entre:

evidencia

afirmación

decisión

riesgo

Cuando estos cuatro elementos permanecen alineados, la IA puede convertirse en una herramienta científica de alto valor.

Cuando se separan, una predicción técnicamente sofisticada puede convertirse en un nuevo punto ciego dentro del sistema de calidad y seguridad alimentaria.

36. Bibliografía

Nota editorial

La bibliografía se divide en cuatro bloques para facilitar su utilización:

1. legislación de la Unión Europea;
2. normas técnicas, directrices y documentos institucionales;
3. metodología estadística, validación y gobernanza de inteligencia artificial;
4. microbiología predictiva, vida útil y tecnologías instrumentales aplicadas a alimentos.

Las referencias científicas se presentan siguiendo los criterios de APA 7. En las disposiciones normativas y normas técnicas se conserva la identificación jurídica o normalizada completa, por ser la forma más precisa de citación.

Antes de la publicación definitiva deberá comprobarse si se ha producido alguna modificación normativa posterior a la fecha de cierre del manuscrito.

36.1. Legislación y documentos jurídicos de la Unión Europea

Comisión Europea. (2005). *Reglamento (CE) n.º 2073/2005 de la Comisión, de 15 de noviembre de 2005, relativo a los criterios microbiológicos aplicables a los productos alimenticios*. Diario Oficial de la Unión Europea, L 338, 1–26. Versión consolidada de 1 de julio de 2026.

Comisión Europea. (2022). *Comunicación de la Comisión sobre la aplicación de sistemas de gestión de la seguridad alimentaria que contemplan buenas prácticas de higiene y procedimientos basados en los principios del APPCC, incluida la facilitación y flexibilidad respecto de su aplicación en determinadas empresas alimentarias (2022/C 355/01)*. Diario Oficial de la Unión Europea, C 355, 1–58.

Comisión Europea. (2024). *Reglamento (UE) 2024/2895 de la Comisión, de 20 de noviembre de 2024, por el que se modifica el Reglamento (CE) n.º 2073/2005 en lo que respecta a Listeria monocytogenes*. Diario Oficial de la Unión Europea, L 2024/2895.

Parlamento Europeo y Consejo de la Unión Europea. (2002). *Reglamento (CE) n.º 178/2002, de 28 de enero de 2002, por el que se establecen los principios y requisitos generales de la legislación alimentaria, se crea la Autoridad Europea de Seguridad Alimentaria y se fijan procedimientos relativos a la seguridad alimentaria*. Diario Oficial de las Comunidades Europeas, L 31, 1–24.

Parlamento Europeo y Consejo de la Unión Europea. (2004). *Reglamento (CE) n.º 852/2004, de 29 de abril de 2004, relativo a la higiene de los productos alimenticios*. Diario Oficial de la Unión Europea, L 139, 1–54.

Parlamento Europeo y Consejo de la Unión Europea. (2011). *Reglamento (UE) n.º 1169/2011, de 25 de octubre de 2011, sobre la información alimentaria facilitada al consumidor*. Diario Oficial de la Unión Europea, L 304, 18–63.

Parlamento Europeo y Consejo de la Unión Europea. (2016a). *Directiva (UE) 2016/943, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados —secretos comerciales— contra su obtención, utilización y revelación ilícitas*. Diario Oficial de la Unión Europea, L 157, 1–18.

Parlamento Europeo y Consejo de la Unión Europea. (2016b). *Reglamento (UE) 2016/679, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos*. Diario Oficial de la Unión Europea, L 119, 1–88.

Parlamento Europeo y Consejo de la Unión Europea. (2017). *Reglamento (UE) 2017/625, de 15 de marzo de 2017, relativo a los controles y otras actividades oficiales realizados para garantizar la aplicación de la legislación sobre alimentos y piensos*. Diario Oficial de la Unión Europea, L 95, 1–142.

Parlamento Europeo y Consejo de la Unión Europea. (2024a). *Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial —Reglamento de Inteligencia Artificial—*. Diario Oficial de la Unión Europea, L 2024/1689.

Parlamento Europeo y Consejo de la Unión Europea. (2024b). *Directiva (UE) 2024/2853, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos y por la que se deroga la Directiva 85/374/CEE del Consejo*. Diario Oficial de la Unión Europea, L 2024/2853.

36.2. Normas técnicas, Codex y documentos institucionales

AOAC International. (2016). *Appendix F: Guidelines for standard method performance requirements*. En *Official methods of analysis of AOAC International* (20.ª ed.). AOAC International.

Codex Alimentarius Commission. (1997). *Principles and guidelines for the establishment and application of microbiological criteria related to foods* (CXG 21-1997; revisadas en 2013). FAO/WHO.

Codex Alimentarius Commission. (1999). *Principles and guidelines for the conduct of microbiological risk assessment* (CXG 30-1999; revisadas en 2021). FAO/WHO.

Codex Alimentarius Commission. (2007). *Principles and guidelines for the conduct of microbiological risk management* (CXG 63-2007). FAO/WHO.

Codex Alimentarius Commission. (2008). *Guidelines for the validation of food safety control measures* (CXG 69-2008; revisión de 2013). FAO/WHO.

Codex Alimentarius Commission. (2022). *General principles of food hygiene* (CXC 1-1969; revisión de 2022). FAO/WHO.

European Food Safety Authority Panel on Biological Hazards. (2020). Guidance on date marking and related food information: Part 1—Date marking. *EFSA Journal, 18*(12), 6306. doi:10.2903/j.efsa.2020.6306

European Food Safety Authority Panel on Biological Hazards. (2021). Guidance on date marking and related food information: Part 2—Food information. *EFSA Journal, 19*(4), 6510. doi:10.2903/j.efsa.2021.6510

European Union Reference Laboratory for *Listeria monocytogenes*. (2021). *Technical guidance document on challenge tests and durability studies for assessing shelf-life of ready-to-eat foods related to Listeria monocytogenes* (Version 4). ANSES. doi:10.5281/zenodo.7760700

Food and Agriculture Organization of the United Nations, & World Health Organization. (2004). *Risk assessment of Listeria monocytogenes in ready-to-eat foods: Technical report*. Microbiological Risk Assessment Series No. 5. FAO/WHO.

International Organization for Standardization. (2016a). *ISO 16140-1:2016. Microbiology of the food chain—Method validation—Part 1: Vocabulary*. ISO.

International Organization for Standardization. (2016b). *ISO 16140-2:2016. Microbiology of the food chain—Method validation—Part 2: Protocol for the validation of alternative proprietary methods against a reference method*. ISO.

International Organization for Standardization. (2017). *ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories*. ISO.

International Organization for Standardization. (2018). *ISO 22000:2018. Food safety management systems—Requirements for any organization in the food chain*. ISO.

International Organization for Standardization. (2019). *ISO 20976-1:2019. Microbiology of the food chain—Requirements and guidelines for conducting challenge tests of food and feed products—Part 1: Challenge tests to study growth potential, lag time and maximum growth rate*. ISO.

International Organization for Standardization. (2020a). *ISO 16140-4:2020. Microbiology of the food chain—Method validation—Part 4: Protocol for method validation in a single laboratory*. ISO.

International Organization for Standardization. (2020b). *ISO/IEC TR 24028:2020. Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence*. ISO.

International Organization for Standardization. (2021a). *ISO 16140-3:2021. Microbiology of the food chain—Method validation—Part 3: Protocol for the verification of reference methods and validated alternative methods in a single laboratory*. ISO.

International Organization for Standardization. (2021b). *ISO/IEC TR 24027:2021. Information technology—Artificial intelligence—Bias in AI systems and AI-aided decision making*. ISO.

International Organization for Standardization. (2021c). *ISO/IEC TR 24029-1:2021. Artificial intelligence—Assessment of the robustness of neural networks—Part 1: Overview*. ISO.

International Organization for Standardization. (2022a). *ISO/IEC 22989:2022. Information technology—Artificial intelligence—Artificial intelligence concepts and terminology*. ISO.

International Organization for Standardization. (2022b). *ISO/IEC 23053:2022. Framework for artificial intelligence systems using machine learning*. ISO.

International Organization for Standardization. (2023a). *ISO/IEC 23894:2023. Information technology—Artificial intelligence—Guidance on risk management*. ISO.

International Organization for Standardization. (2023b). *ISO/IEC 25059:2023. Software engineering—Systems and software quality requirements and evaluation—Quality model for AI systems*. ISO.

International Organization for Standardization. (2023c). *ISO/IEC 42001:2023. Information technology—Artificial intelligence—Management system*. ISO.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework—AI RMF 1.0* (NIST AI 100-1). U.S. Department of Commerce. doi:10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). U.S. Department of Commerce. doi:10.6028/NIST.AI.600-1

36.3. Metodología estadística, validación y gobernanza de modelos de IA

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19*(4), 453–473.

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591. doi:10.1561/2200000101

Archer, L., Snell, K. I. E., Ensor, J., Hudda, M. T., Collins, G. S., & Riley, R. D. (2021). Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine, 40*(1), 133–146. doi:10.1002/sim.8766

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. doi:10.1016/S0140-6736(86)90837-8

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. doi:10.1186/s12864-019-6413-7

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. doi:10.1177/001316446002000104

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. doi:10.1136/bmj-2023-078378

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. En *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). Association for Computing Machinery. doi:10.1145/1143844.1143874

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association, 24*(6), 1052–1061. doi:10.1093/jamia/ocx030

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44. doi:10.1145/2523813

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. En *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning, 45*, 171–186. doi:10.1023/A:1010920819831

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2.ª ed.). Springer.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. En *Proceedings of the International Conference on Learning Representations*.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. doi:10.1016/j.patter.2023.100804

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

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. En *Advances in Neural Information Processing Systems* (Vol. 30).

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar,

K., Riklund, K., Marias, K., ... Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, e081554. doi:10.1136/bmj-2024-081554

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268. doi:10.2307/2532051

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering, 31*(12), 2346–2363. doi:10.1109/TKDE.2018.2876857

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta—Protein Structure, 405*(2), 442–451. doi:10.1016/0005-2795(75)90109-9

Moons, K. G. M., van Smeden, M., Riley, R. D., Collins, G. S., Dhiman, P., Reitsma, J. B., Damen, J. A. A. G., Kaul, T., Hooft, L., Andaur Navarro, C. L., Yang, B., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., ... Wynants, L. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, e082505. doi:10.1136/bmj-2024-082505

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. En *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). Association for Computing Machinery. doi:10.1145/1102351.1102430

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. En *Advances in Neural Information Processing Systems* (Vol. 32).

Riley, R. D., Debray, T. P. A., Collins, G. S., Archer, L., Ensor, J., van Smeden, M., & Snell, K. I. E. (2021). Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine, 40*(19), 4230–4251. doi:10.1002/sim.9025

Riley, R. D., Collins, G. S., Ensor, J., Archer, L., Booth, S., Mozumder, S. I., Rutherford, M. J., Van Calster, B., & Snell, K. I. E. (2022). Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Statistics in Medicine, 41*(7), 1280–1295. doi:10.1002/sim.9275

Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024a). Evaluation of clinical prediction models—Part 1: From development to external validation. *BMJ, 384*, e074819. doi:10.1136/bmj-2023-074819

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2024b). Evaluation of clinical prediction models—Part 2: How to undertake an external validation study. *BMJ, 384*, e074820. doi:10.1136/bmj-2023-074820

Riley, R. D., Snell, K. I. E., Ensor, J., Archer, L., Booth, S., Collins, G. S., & Martin, G. P. (2024c). Evaluation of clinical prediction models—Part 3: Calculating the sample size required for an external validation study. *BMJ, 384*, e074821. doi:10.1136/bmj-2023-074821

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. doi:10.1111/ecog.02881

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. doi:10.1371/journal.pone.0118432

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2.ª ed.). Springer.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138. doi:10.1097/EDE.0b013e3181c30fb2

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology, 74*, 167–176. doi:10.1016/j.jclinepi.2015.12.005

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, 230. doi:10.1186/s12916-019-1466-7

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91. doi:10.1186/1471-2105-7-91

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574. doi:10.1177/0272989X06295361

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery, 30*, 964–994. doi:10.1007/s10618-015-0448-4

36.4. Microbiología predictiva, vida útil y evaluación del riesgo alimentario

Baranyi, J., & Roberts, T. A. (1994). A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology, 23*(3–4), 277–294. doi:10.1016/0168-1605(94)90157-0

Baranyi, J., Roberts, T. A., & McClure, P. (1993). A non-autonomous differential equation to model bacterial growth. *Food Microbiology, 10*(1), 43–59. doi:10.1006/fmic.1993.1005

Buchanan, R. L., Whiting, R. C., & Damert, W. C. (1997). When is simple good enough: A comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves. *Food Microbiology, 14*(4), 313–326. doi:10.1006/fmic.1997.0125

Dalgaard, P., & Jørgensen, L. V. (1998). Predicted and observed growth of *Listeria monocytogenes* in seafood challenge tests and in naturally contaminated cold-smoked salmon. *International Journal of Food Microbiology, 40*(1–2), 105–115. doi:10.1016/S0168-1605(98)00019-1

Koutsoumanis, K. P., Lianou, A., & Gougouli, M. (2016). Latest developments in foodborne pathogens modeling. *Current Opinion in Food Science, 8*, 89–98. doi:10.1016/j.cofs.2016.04.006

McKellar, R. C., & Lu, X. (Eds.). (2004). *Modeling microbial responses in foods*. CRC Press.

McMeekin, T. A., & Ross, T. (2002). Predictive microbiology: Providing a knowledge-based framework for change management. *International Journal of Food Microbiology, 78*(1–2), 133–153. doi:10.1016/S0168-1605(02)00231-3

Membré, J.-M., & Lambert, R. J. W. (2008). Application of predictive modelling techniques in industry: From food design up to risk assessment. *International Journal of Food Microbiology, 128*(1), 10–15. doi:10.1016/j.ijfoodmicro.2008.07.006

Nauta, M. J. (2000). Separation of uncertainty and variability in quantitative microbial risk assessment models. *International Journal of Food Microbiology, 57*(1–2), 9–18. doi:10.1016/S0168-1605(00)00225-7

Pérez-Rodríguez, F., & Valero, A. (2013). *Predictive microbiology in foods*. Springer.

Ratkowsky, D. A., Olley, J., McMeekin, T. A., & Ball, A. (1982). Relationship between temperature and growth rate of bacterial cultures. *Journal of Bacteriology, 149*(1), 1–5. doi:10.1128/jb.149.1.1-5.1982

Ross, T. (1996). Indices for performance evaluation of predictive models in food microbiology. *Journal of Applied Bacteriology, 81*(5), 501–508. doi:10.1111/j.1365-2672.1996.tb03539.x

Ross, T., Dalgaard, P., & Tienungoon, S. (2000). Predictive modelling of the growth and survival of *Listeria* in fishery products. *International Journal of Food Microbiology, 62*(3), 231–245. doi:10.1016/S0168-1605(00)00340-8

Vose, D. (2008). *Risk analysis: A quantitative guide* (3.ª ed.). Wiley.

Zwietering, M. H., Jongenburger, I., Rombouts, F. M., & van 't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied and Environmental Microbiology, 56*(6), 1875–1881. doi:10.1128/AEM.56.6.1875-1881.1990

36.5. Visión artificial, espectroscopia e inteligencia artificial aplicada al control de alimentos

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology, 18*(2), 72–83. doi:10.1016/j.tifs.2006.09.003

Cubero, S., Aleixos, N., Moltó, E., Gómez-Sanchis, J., & Blasco, J. (2011). Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food and Bioprocess Technology, 4*, 487–504. doi:10.1007/s11947-010-0411-8

Feng, Y.-Z., & Sun, D.-W. (2012). Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition, 52*(11), 1039–1058. doi:10.1080/10408398.2011.651542

Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging—An emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology, 18*(12), 590–598. doi:10.1016/j.tifs.2007.06.001

Huang, H., Liu, L., & Ngadi, M. O. (2014). Recent developments in hyperspectral imaging for assessment of food quality and safety. *Sensors, 14*(4), 7248–7276. doi:10.3390/s140407248

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. doi:10.1016/j.compag.2018.02.016

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. doi:10.3390/s18082674

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives—A review. *Analytica Chimica Acta, 1026*, 8–36. doi:10.1016/j.aca.2018.04.004

Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry, 28*(10), 1201–1222. doi:10.1016/j.trac.2009.07.007

Sun, D.-W. (Ed.). (2016). *Computer vision technology for food quality evaluation* (2.ª ed.). Academic Press.

37. Correspondencia recomendada entre capítulos y fuentes

Para facilitar la inserción de citas en el manuscrito, pueden utilizarse las siguientes asociaciones.

Fundamentos de validación y APPCC

- Codex Alimentarius Commission (2008, 2022).
- Comisión Europea (2022).
- ISO 22000:2018.
- Reglamento (CE) n.º 852/2004.

Gestión del riesgo y gobernanza de IA

- Reglamento (UE) 2024/1689.
- NIST (2023).
- ISO/IEC 23894:2023.
- ISO/IEC 42001:2023.
- Lekadir et al. (2025).

Diseño, riesgo de sesgo y comunicación

- Collins et al. (2024).
- Moons et al. (2025).
- Kapoor y Narayanan (2023).
- Cawley y Talbot (2010).
- Varma y Simon (2006).

Validación externa y tamaño muestral

- Altman y Royston (2000).
- Archer et al. (2021).
- Riley et al. (2021, 2022, 2024a, 2024b, 2024c).
- Roberts et al. (2017).

Discriminación y clasificación

- Fawcett (2006).
- Davis y Goadrich (2006).
- Saito y Rehmsmeier (2015).
- Chicco y Jurman (2020).
- Matthews (1975).

Calibración e incertidumbre

- Brier (1950).
- Guo et al. (2017).
- Van Calster et al. (2016, 2019).
- Ovadia et al. (2019).
- Angelopoulos y Bates (2023).

Deriva y monitorización

- Davis et al. (2017).
- Gama et al. (2014).
- Lu et al. (2019).
- Webb et al. (2016).
- NIST (2023).

Concordancia y comparación de métodos

- Bland y Altman (1986).
- Lin (1989).
- ISO 16140-2:2016.
- ISO 16140-3:2021.
- ISO/IEC 17025:2017.

Microbiología predictiva y vida útil

- Baranyi y Roberts (1994).
- McMeekin y Ross (2002).
- Ross (1996).
- Dalgaard y Jørgensen (1998).
- Membré y Lambert (2008).
- Pérez-Rodríguez y Valero (2013).
- EURL Listeria monocytogenes (2021).
- EFSA BIOHAZ Panel (2020, 2021).

Espectroscopia y visión artificial

- Cen y He (2007).
- Cubero et al. (2011).
- Gowen et al. (2007).
- Feng y Sun (2012).
- Huang et al. (2014).
- Pasquini (2018).
- Rinnan et al. (2009).

38. Fórmula editorial recomendada para las citas

Las citas parentéticas deberán utilizarse de la siguiente forma para cualquier clase de búsqueda o verificación:

- un autor: (Ross, 1996);
- dos autores: (Baranyi & Roberts, 1994);
- tres o más autores: (Van Calster et al., 2019);
- varias referencias: (Brier, 1950; Guo et al., 2017; Van Calster et al., 2019);
- autor institucional: (Codex Alimentarius Commission, 2008);
- normativa europea: Reglamento (UE) 2024/1689 o (Reglamento [UE] 2024/1689).

En la primera mención de una entidad puede emplearse:

European Food Safety Authority Panel on Biological Hazards (EFSA BIOHAZ Panel, 2020).

En menciones posteriores:

EFSA BIOHAZ Panel (2020).

39. Nota final de cierre bibliográfico

La bibliografía seleccionada combina:

- fuentes normativas primarias;
- estándares internacionales;
- directrices oficiales;
- literatura fundacional;
- metodología contemporánea de validación de modelos;
- microbiología predictiva;
- evaluación de vida útil;
- visión artificial y espectroscopia alimentaria.

Su función no es únicamente respaldar afirmaciones generales sobre inteligencia artificial. Proporciona el fundamento metodológico para:

- diseñar la validación;
- seleccionar métricas;
- cuantificar incertidumbre;
- evaluar transportabilidad;
- controlar la deriva;
- comparar métodos;
- documentar la decisión;
- integrar el modelo en APPCC;
- justificar jurídicamente su utilización.

Servicio de Publicaciones de INTABIOTECH SL

www.intabiotech.com