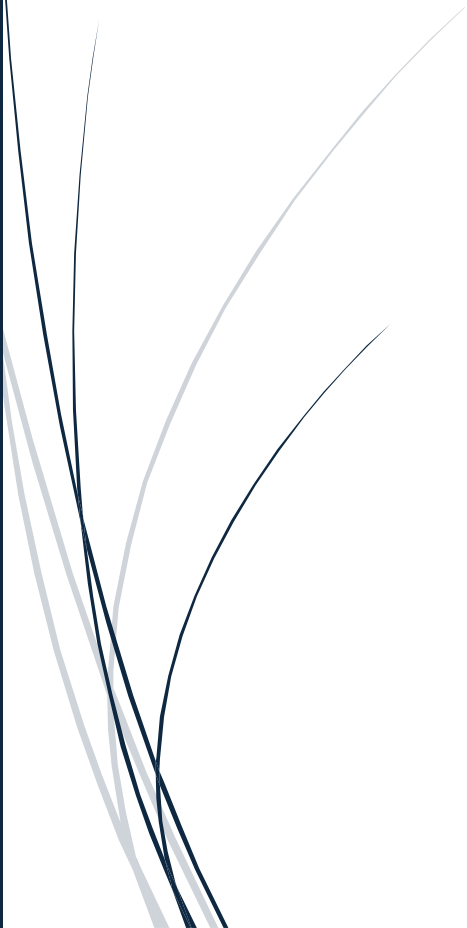


Julio de 2026

Validar científicamente un modelo de IA para Calidad, Vida Útil y Seguridad Alimentaria (Vol. 1)

*Validación científica, externa y operacional
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 1

*De la precisión aparente a la evidencia de aptitud para el uso:
validación externa, errores de clasificación, deriva, incertidumbre y
criterios de aceptación (Capítulos 1-18)*

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

Manuscrito técnico-científico · Edición editorial · Julio de 2026

Servicio de Publicaciones de
INTABIOTECH

Resumen

La incorporación de modelos de inteligencia artificial a la industria alimentaria ha avanzado más rápidamente que las metodologías destinadas a demostrar que sus predicciones son científicamente fiables, industrialmente reproducibles y suficientemente seguras para sustentar decisiones sobre calidad, liberación de lotes, vida útil o control de peligros. En numerosos proyectos, la denominada validación se limita a comunicar una accuracy elevada, una curva ROC obtenida mediante validación cruzada o un error medio calculado sobre datos procedentes del mismo entorno experimental utilizado para desarrollar el algoritmo. Tales resultados pueden acreditar capacidad de ajuste, pero no demuestran necesariamente generalización, robustez frente a variaciones de proceso ni aptitud para intervenir en decisiones con consecuencias sanitarias o económicas.

La validación científica de un modelo alimentario debe comenzar por una definición inequívoca de su finalidad prevista, la población de productos y procesos sobre la que podrá utilizarse, el resultado que pretende predecir y las consecuencias de cada modalidad de error. A partir de esa definición, debe evaluarse la representatividad de los datos, la independencia real entre entrenamiento y ensayo, el rendimiento en conjuntos externos, la calibración de las probabilidades, la sensibilidad y especificidad, el balance entre falsos positivos y falsos negativos, la incertidumbre de las predicciones, la estabilidad frente a cambios de matriz, campaña, proveedor, instalación o equipo analítico y la posibilidad de deriva después del despliegue.

El presente trabajo propone un marco de validación basado en riesgo para modelos aplicados a calidad, vida útil y seguridad alimentaria. Integra principios procedentes de la validación de medidas de control alimentario, la microbiología predictiva, la metrología, la evaluación de modelos estadísticos y el gobierno de sistemas de inteligencia artificial. Se defiende que los criterios de aceptación no deben establecerse mediante umbrales universales de precisión, sino a partir del peligro controlado, la prevalencia del evento, el coste y gravedad de cada error, la capacidad de intervención humana, la incertidumbre residual y la función concreta que el modelo desempeña dentro del sistema APPCC o de gestión de calidad.

Palabras clave: *inteligencia artificial; seguridad alimentaria; vida útil; validación externa; sensibilidad; especificidad; falsos negativos; deriva del modelo; microbiología predictiva; incertidumbre; representatividad de datos; APPCC.*

Índice general Vol. 1

CÓMO VALIDAR CIENTÍFICAMENTE UN MODELO DE INTELIGENCIA ARTIFICIAL PARA CALIDAD, VIDA ÚTIL Y SEGURIDAD ALIMENTARIA	2
VOLUMEN 1.....	2
<i>De la precisión aparente a la evidencia de aptitud para el uso: validación externa, errores de clasificación, deriva, incertidumbre y criterios de aceptación (Capítulos 1-18).....</i>	<i>2</i>
RESUMEN	3
1. INTRODUCCIÓN: EL PROBLEMA DE CONFUNDIR RENDIMIENTO ALGORÍTMICO CON EVIDENCIA CIENTÍFICA.....	15
1.1. UN MODELO NO ESTÁ VALIDADO EN ABSTRACTO	16
1.2. LA FINALIDAD PREVISTA DETERMINA LA VALIDACIÓN	16
1.3. LOS TRES DOMINIOS REQUIEREN METODOLOGÍAS DIFERENTES	17
1.4. HIPÓTESIS CENTRAL DE LA PUBLICACIÓN	17
2. ARQUITECTURA CIENTÍFICA DE LA VALIDACIÓN: QUÉ DEBE DEMOSTRARSE ANTES DE EVALUAR LAS MÉTRICAS.....	19
2.1. LA FINALIDAD PREVISTA COMO NÚCLEO DE LA VALIDACIÓN	19
2.2. EL OBJETO DE VALIDACIÓN NO ES ÚNICAMENTE EL ALGORITMO	20
2.3. DEFINICIÓN DEL RESULTADO VERDADERO: EL PATRÓN DE REFERENCIA.....	21
2.3.1. Validez analítica.....	21
2.3.2. Validez tecnológica.....	22
2.3.3. Reproducibilidad	22
2.3.4. Ausencia de contaminación de la etiqueta	22
2.4. LA UNIDAD ESTADÍSTICA: MUESTRA, LOTE, PRODUCTO O PROCESO	22
2.5. DOMINIO DE APLICACIÓN Y POBLACIÓN OBJETIVO	23
2.6. REPRESENTATIVIDAD NO SIGNIFICA MERA PROPORCIONALIDAD	24
2.7. SEPARACIÓN ENTRE ENTRENAMIENTO, VALIDACIÓN Y PRUEBA.....	25
Conjunto de entrenamiento	25
Conjunto de validación	25
Conjunto de prueba.....	26
2.8. VALIDACIÓN INTERNA Y VALIDACIÓN EXTERNA	26
Nivel 1. Separación aleatoria dentro de la misma base.....	27
Nivel 2. Separación agrupada	27
Nivel 3. Validación temporal	27
Nivel 4. Validación en otra línea o planta	27
Nivel 5. Validación independiente multicéntrica	27
Nivel 6. Evaluación prospectiva de impacto.....	27
2.9. VALIDACIÓN DEL MODELO Y VALIDACIÓN DEL IMPACTO.....	27
2.10. INDEPENDENCIA DE LA VALIDACIÓN Y CONTROL DE CONFLICTOS	28
3. EL PROTOCOLO MAESTRO DE VALIDACIÓN	30
3.1. IDENTIFICACIÓN DEL SISTEMA	30
3.2. PREGUNTA DE VALIDACIÓN.....	30
3.3. DISEÑO DEL ESTUDIO	31
3.4. CRITERIOS DE ACEPTACIÓN	31
3.5. RESULTADO DE LA VALIDACIÓN	32

4. SENSIBILIDAD, ESPECIFICIDAD Y ERRORES DE CLASIFICACIÓN: LA ANATOMÍA REAL DEL RENDIMIENTO.....	33
4.1. LA MATRIZ DE CONFUSIÓN.....	33
• Verdadero positivo	33
• Verdadero negativo	33
• Falso positivo.....	34
• Falso negativo.....	34
4.2. SENSIBILIDAD	34
TASA DE FALSOS NEGATIVOS 1-SENSIBILIDAD [FN] / [VP+FN]	35
4.3. ESPECIFICIDAD.....	35
ESPECIFICIDAD [VN] / [VN+FP].....	35
TASA DE FALSOS POSITIVOS 1-ESPECIFICIDAD [FP] / [VN+FP]	35
4.4. SENSIBILIDAD Y ESPECIFICIDAD NO SON PROBABILIDADES POSTERIORES	36
4.5. VALOR PREDICTIVO POSITIVO	36
4.6. VALOR PREDICTIVO NEGATIVO	36
4.7. EL EFECTO DE LA PREVALENCIA	37
4.8. ACCURACY O EXACTITUD GLOBAL.....	38
ACCURACY [VP+VN] / [VP+VN+FP+FN]	38
4.9. EXACTITUD BALANCEADA	38
BALANCED ACCURACY [SENSIBILIDAD+ESPECIFICIDAD] / [2]	38
4.10. PRECISIÓN Y RECALL	39
PRECISION [VP] / [VP+FP]	39
RECALL [VP] / [VP+FN]	39
4.11. F1-SCORE	39
4.12. F-BETA	40
4.13. COEFICIENTE DE CORRELACIÓN DE MATTHEWS	40
4.14. CURVA ROC	40
4.15. CURVA PRECISION-RECALL	41
4.16. SELECCIÓN DEL UMBRAL DE DECISIÓN	42
4.17. INTERVALOS DE CONFIANZA.....	43
4.18. EL PROBLEMA DE OBSERVAR CERO FALSOS NEGATIVOS	44
4.19. RENDIMIENTO GLOBAL Y RENDIMIENTO POR SUBGRUPOS	44
4.20. FALSOS NEGATIVOS PRÓXIMOS Y LEJANOS AL LÍMITE.....	45
4.21. SISTEMAS CON OPCIÓN DE ABSTENCIÓN	46
TASA DE ABSTENCIÓN [N_INDETERMINADOS] / [N_TOTAL].....	46
4.22. MATRICES MULTICLASE	46
• Macro-promedio	47
M_MACRO [1] / [K] $\sum_{K=1}^K M_K$.....	47
Micro-promedio	47
M_MICRO [$\sum_K VP_K$] / [$\sum_K VP_K + \sum_K FN_K$]	47
Promedio ponderado	47
4.23. MODELOS JERÁRQUICOS DE DECISIÓN	47
4.24. SELECCIÓN DE MÉTRICAS SEGÚN LA FINALIDAD	48
Detección de peligro o lote inseguro	48

• Sistema de cribado	48
• Control de falsas alarmas.....	48
• Clasificación comercial.....	48
• Detección de defectos raros	49
5. VALIDACIÓN DE MODELOS DE REGRESIÓN Y PREDICCIÓN DE VIDA ÚTIL	50
5.1. DEFINICIÓN DEL RESULTADO EN VIDA ÚTIL	50
5.2. ERROR DE PREDICCIÓN	51
5.3. ERROR MEDIO	51
5.4. ERROR ABSOLUTO MEDIO	52
5.5. ERROR CUADRÁTICO MEDIO Y RAÍZ DEL ERROR CUADRÁTICO MEDIO	52
5.6. ERROR MEDIANO Y PERCENTILES	53
5.7. ERROR PORCENTUAL	53
5.8. COEFICIENTE DE DETERMINACIÓN	54
5.9. CORRELACIÓN NO EQUIVALE A CONCORDANCIA	54
5.10. GRÁFICOS DE RESIDUOS	55
5.11. HETEROCEDASTICIDAD.....	55
5.12. LÍMITES DE ACUERDO.....	56
5.13. SOBREESTIMACIÓN E INFRAESTIMACIÓN	56
5.14. ERROR PELIGROSO	57
5.15. MÁRGENES DE SEGURIDAD	57
5.16. INTERVALOS DE PREDICCIÓN	58
5.17. CALIBRACIÓN DE PREDICCIONES CONTINUAS	58
5.18. PREDICCIÓN MICROBIOLÓGICA.....	59
5.19. FACTOR DE SESGO	59
5.20. FACTOR DE EXACTITUD.....	60
5.21. DATOS CENSURADOS	60
5.22. ANÁLISIS DE SUPERVIVENCIA	61
5.23. ÍNDICE DE CONCORDANCIA	61
5.24. PUNTUACIÓN DE BRIER	61
5.25. MODELOS DE HORIZONTE FIJO	62
5.26. VALIDACIÓN TEMPORAL	62
5.27. VALIDACIÓN POR LOTES Y CURVAS LONGITUDINALES	63
5.28. EXTRAPOLACIÓN	63
5.29. VALIDACIÓN EN CONDICIONES RAZONABLEMENTE ADVERSAS	63
5.30. COMPARACIÓN CON MODELOS DE REFERENCIA.....	64
5.31. CRITERIOS DE ACEPTACIÓN PARA VIDA ÚTIL.....	64
5.32. VIDA ÚTIL DINÁMICA	65
5.33. LA PREDICCIÓN NO SUSTITUYE LA EVIDENCIA EXPERIMENTAL	65
6. CALIBRACIÓN, INCERTIDUMBRE Y ROBUSTEZ: VALIDAR QUE EL MODELO CONOCE LOS LÍMITES DE SU PROPIA PREDICCIÓN.....	67
6.1. DISCRIMINACIÓN Y CALIBRACIÓN SON PROPIEDADES DIFERENTES.....	67
6.2. CURVA DE CALIBRACIÓN	68
6.3. INTERCEPTO Y PENDIENTE DE CALIBRACIÓN	68
6.4. PUNTUACIÓN DE <i>BRIER</i>	69
6.5. ERROR ESPERADO DE CALIBRACIÓN	70
6.6. RECALIBRACIÓN	70
6.7. CAMBIO DE PREVALENCIA Y PROBABILIDAD POSTERIOR	70
6.8. INCERTIDUMBRE ALEATORIA Y EPISTÉMICA	71
<i>Incetidumbre aleatoria</i>	71
<i>Incetidumbre epistémica</i>	71
6.9. INCERTIDUMBRE DE MEDIDA.....	72

6.10. ROBUSTEZ FRENTE A PERTURBACIONES	72
6.11. PRUEBAS DE ESTRÉS	73
6.12. COHERENCIA TECNOLÓGICA	74
6.13. PRUEBAS CONTRAFCTUALES.....	74
6.14. DATOS FUERA DE DISTRIBUCIÓN	75
6.15. DISTANCIA AL DOMINIO DE ENTRENAMIENTO	75
6.16. LÍMITES DUROS Y LÍMITES BLANDOS.....	76
<i>Límites duros</i>	76
<i>Límites blandos</i>	76
6.17. PREDICCIÓN SELECTIVA Y COBERTURA.....	76
6.18. UMBRALES DE CONFIANZA.....	77
6.19. CONJUNTOS DE MODELOS	77
6.20. MÉTODOS BAYESIANOS Y APROXIMACIONES	77
6.21. PREDICCIÓN <i>CONFORMAL</i>	78
6.22. INCERTIDUMBRE POR DATOS FALTANTES	78
6.23. ROBUSTEZ FRENTE A ERRORES DE UNIDADES Y CODIFICACIÓN	79
6.24. ROBUSTEZ FRENTE A ATAQUES O MANIPULACIONES	79
6.25. ROBUSTEZ OPERATIVA	80
6.26. SUPERVISIÓN HUMANA Y CONFIANZA APROPIADA	80
6.27. EXPLICABILIDAD E INCERTIDUMBRE.....	81
6.28. CRITERIOS DE ACEPTACIÓN PARA CALIBRACIÓN Y ROBUSTEZ	81
<i>Calibración</i>	81
<i>Puntuación probabilística</i>	82
<i>Cobertura de intervalos</i>	82
<i>Robustez</i>	82
<i>Abstención</i>	82
<i>Coherencia</i>	82
6.29. LA OBLIGACIÓN DE RECONOCER LA IGNORANCIA.....	82
6.30. CONCLUSIÓN DEL BLOQUE	83
7. DERIVA DEL MODELO: POR QUÉ UNA VALIDACIÓN CORRECTA PUEDE DEJAR DE SER VÁLIDA ...	84
7.1. LA DERIVA COMO PROBLEMA ALIMENTARIO, NO EXCLUSIVAMENTE INFORMÁTICO	84
7.2. DERIVA DE COVARIABLES	85
7.3. DERIVA DEL RESULTADO O DE PREVALENCIA	85
7.4. DERIVA CONCEPTUAL.....	86
7.5. DERIVA DE ETIQUETA.....	87
7.6. DERIVA DEL SENSOR O DEL SISTEMA DE MEDIDA	87
7.7. DERIVA TEMPORAL GRADUAL, ABRUPTA Y RECURRENTE.....	88
<i>Deriva gradual</i>	88
<i>Deriva abrupta</i>	88
<i>Deriva recurrente</i>	88
7.8. DERIVA REAL Y DERIVA APARENTE	89
7.9. QUÉ DEBE MONITORIZARSE	89
<i>Calidad de datos</i>	89
<i>Variables de entrada</i>	89
<i>Salidas del modelo</i>	90
<i>Rendimiento</i>	90
<i>Impacto</i>	90
7.10. VENTANAS DE MONITORIZACIÓN	91
<i>Ventana fija</i>	91
<i>Ventana móvil</i>	91
<i>Ventana acumulativa</i>	91
<i>Ventanas múltiples</i>	91

7.11. INDICADORES UNIVARIANTES	91
7.12. <i>POPULATION STABILITY INDEX</i>	92
7.13. DISTANCIAS ENTRE DISTRIBUCIONES	92
7.14. DERIVA MULTIVARIANTE	92
7.15. MONITORIZACIÓN DE PROBABILIDADES Y ALERTAS	93
7.16. MONITORIZACIÓN DEL RENDIMIENTO CON ETIQUETAS RETRASADAS	93
<i>Monitorización sin resultado verdadero</i>	94
<i>Monitorización con resultado verdadero</i>	94
7.17. MUESTREO DE VERIFICACIÓN	94
7.18. SESGO DE RETROALIMENTACIÓN	94
7.19. CONTROL ESTADÍSTICO DEL RENDIMIENTO	95
7.20. INDICADORES ADELANTADOS Y RETRASADOS	95
7.21. UMBRALES DE ALERTA, ACCIÓN Y SUSPENSIÓN	96
<i>Nivel verde: funcionamiento normal</i>	96
<i>Nivel amarillo: alerta</i>	96
<i>Nivel naranja: acción correctiva</i>	96
<i>Nivel rojo: suspensión</i>	97
7.22. MODIFICACIÓN DEL UMBRAL DURANTE LA OPERACIÓN	97
7.23. RECALIBRACIÓN, ACTUALIZACIÓN Y REENTRENAMIENTO	97
<i>Recalibración</i>	97
<i>Actualización parcial</i>	98
<i>Reentrenamiento</i>	98
7.24. APRENDIZAJE CONTINUO	98
7.25. COMPARACIÓN ENTRE VERSIÓN ACTIVA Y CANDIDATA	99
7.26. DESPLIEGUE PARALELO Y MODO SOMBRA	99
7.27. ENSAYOS A/B Y DESPLIEGUE GRADUAL	100
7.28. REVALIDACIÓN PERIÓDICA	100
7.29. REVALIDACIÓN ACTIVADA POR CAMBIOS	100
7.30. REVALIDACIÓN PARCIAL Y COMPLETA	101
<i>Revalidación parcial</i>	101
<i>Revalidación completa</i>	101
7.31. GESTIÓN DE INCIDENTES	101
7.32. CONSERVACIÓN DE REGISTROS	102
7.33. CUADRO MÍNIMO DE VIGILANCIA	103
7.34. RESPONSABILIDADES	103
<i>Propietario del modelo</i>	103
<i>Responsable del proceso</i>	104
<i>Calidad y seguridad alimentaria</i>	104
<i>Tecnologías de la información</i>	104
<i>Laboratorio</i>	104
<i>Comité de aprobación</i>	104
7.35. RELACIÓN CON APPCC Y SISTEMAS DE GESTIÓN	104
7.36. CRITERIOS CIENTÍFICOS PARA MANTENER LA VALIDEZ	105
7.37. LA DERIVA COMO SEÑAL CIENTÍFICA	105
7.38. CONCLUSIÓN DEL BLOQUE	106
8. CRITERIOS DE ACEPTACIÓN BASADOS EN RIESGO: CUÁNDO PUEDE APROBARSE UN MODELO DE IA ALIMENTARIA	107
8.1. ACEPTAR NO SIGNIFICA DEMOSTRAR QUE EL MODELO ES PERFECTO	107
8.2. LOS CRITERIOS DEBEN FIJARSE ANTES DE ABRIR EL CONJUNTO DE PRUEBA	108
8.3. CRITERIOS APROPIADOS, NO NECESARIAMENTE IDÉNTICOS	108
8.4. VARIABLES QUE DETERMINAN EL NIVEL DE EXIGENCIA	109
<i>Severidad</i>	109

<i>Exposición</i>	109
<i>Detectabilidad</i>	109
<i>Reversibilidad</i>	110
<i>Autonomía</i>	110
<i>Barreras independientes</i>	110
8.5. JERARQUÍA FUNCIONAL DE MODELOS ALIMENTARIOS	110
<i>Nivel I. Exploración</i>	110
<i>Nivel II. Apoyo informativo</i>	111
<i>Nivel III. Cribado</i>	111
<i>Nivel IV. Decisión automatizada de calidad</i>	111
<i>Nivel V. Seguridad alimentaria</i>	111
8.6. CRITERIOS PARA MODELOS DE CONTROL DE CALIDAD	112
<i>Falsa aceptación</i>	112
<i>Falso rechazo</i>	112
8.7. CRITERIOS PARA MODELOS DE CRIBADO	113
8.8. CRITERIOS PARA PREDICCIÓN DE VIDA ÚTIL	113
8.9. CRITERIOS PARA MODELOS MICROBIOLÓGICOS	114
<i>Predicción de crecimiento</i>	114
<i>Predicción de tiempo hasta alcanzar un límite</i>	114
<i>Predicción de inactivación</i>	114
8.10. CRITERIOS PARA DECISIONES VINCULADAS A SEGURIDAD ALIMENTARIA	115
<i>Primer nivel: seguridad mínima no compensable</i>	115
<i>Segundo nivel: incertidumbre</i>	115
<i>Tercer nivel: subgrupos</i>	115
<i>Cuarto nivel: robustez</i>	115
<i>Quinto nivel: operación segura</i>	115
8.11. EL PRINCIPIO DE NO COMPENSACIÓN	116
8.12. ERROR MEDIO FRENTE A ERROR CATASTRÓFICO	116
8.13. INTERVALOS DE CONFIANZA Y REGLA DE APROBACIÓN	117
8.14. TAMAÑO MUESTRAL ORIENTADO AL RIESGO	117
8.15. PREVALENCIA Y ESCENARIOS DE UTILIZACIÓN	118
8.16. CRITERIOS POR SUBGRUPOS Y REGLA DEL PEOR CASO	118
8.17. ABSTENCIÓN COMO REQUISITO DE ACEPTACIÓN	118
8.18. FUNCIONAMIENTO SEGURO ANTE FALLOS	119
8.19. COMPARACIÓN CON EL ESTÁNDAR DE PRÁCTICA	119
8.20. BENEFICIO NETO Y UTILIDAD OPERATIVA	120
8.21. APROBACIÓN ESCALONADA	120
<i>No aprobado</i>	120
<i>Aprobado para investigación</i>	120
<i>Aprobado en modo sombra</i>	120
<i>Aprobado como apoyo</i>	120
<i>Aprobado con confirmación obligatoria</i>	120
<i>Aprobado para dominio restringido</i>	120
<i>Aprobado para automatización limitada</i>	121
<i>Aprobado para uso previsto completo</i>	121
8.22. APROBACIÓN CONDICIONAL	121
8.23. CRITERIOS DE SUSPENSIÓN	121
8.24. MATRIZ MAESTRA DE ACEPTACIÓN	122
8.25. EJEMPLO DE ACEPTACIÓN: DETECTOR VISUAL DE CUERPOS EXTRAÑOS	122
8.26. EJEMPLO DE ACEPTACIÓN: VIDA ÚTIL REFRIGERADA	123
8.27. EJEMPLO DE ACEPTACIÓN: CRIBADO MICROBIOLÓGICO	123
8.28. EJEMPLO DE ACEPTACIÓN: CLASIFICACIÓN SENSORIAL	123

8.29. CRITERIOS INVÁLIDOS FRECUENTES	124
8.30. INFORME DE DECISIÓN	124
<i>Sistema</i>	124
<i>Uso autorizado</i>	125
<i>Evidencia</i>	125
<i>Limitaciones</i>	125
<i>Controles obligatorios</i>	125
<i>Decisión</i>	125
<i>Responsables</i>	126
8.31. CONCLUSIÓN DEL BLOQUE	126
9. INTEGRACIÓN DEL MODELO VALIDADO EN APPCC, PRERREQUISITOS Y SISTEMA DE GESTIÓN DE LA SEGURIDAD ALIMENTARIA	127
9.1. LA IA NO SUSTITUYE EL SISTEMA DE GESTIÓN	127
9.2. DETERMINACIÓN DE LA FUNCIÓN DENTRO DEL FSMS	128
9.3. INTEGRACIÓN EN PROGRAMAS DE PRERREQUISITOS	128
<i>Ejemplo: predicción de riesgo de contaminación ambiental</i>	129
9.4. INTEGRACIÓN EN PRERREQUISITOS OPERATIVOS	129
9.5. INTEGRACIÓN EN PUNTOS DE CONTROL CRÍTICO	130
<i>Ejemplo</i>	130
9.6. ¿PUEDE UNA SALIDA ALGORÍTMICA SER UN LÍMITE CRÍTICO?	131
9.7. MODIFICACIÓN DEL ANÁLISIS DE PELIGROS	131
<i>Peligros o desviaciones que el modelo pretende controlar</i>	131
<i>Fallos propios del sistema de IA</i>	131
9.8. ANÁLISIS DE MODOS Y EFECTOS DE FALLO	132
9.9. VALIDACIÓN, MONITORIZACIÓN Y VERIFICACIÓN	132
<i>Validación</i>	132
<i>Monitorización</i>	133
<i>Verificación</i>	133
9.10. PLAN DE MONITORIZACIÓN DEL MODELO	133
9.11. VERIFICACIÓN INDEPENDIENTE	133
9.12. PROCEDIMIENTO DE TRABAJO NORMALIZADO	134
<i>Objeto</i>	134
<i>Alcance</i>	134
<i>Definiciones</i>	134
<i>Responsabilidades</i>	134
<i>Procedimiento</i>	135
<i>Fallos</i>	135
<i>Registros</i>	135
<i>Cambios</i>	135
9.13. DOSSIER MAESTRO DEL MODELO	135
9.14. FICHA DE IDENTIFICACIÓN	136
9.15. DESCRIPCIÓN DE DATOS	136
9.16. TRAZABILIDAD DE LA PREDICCIÓN	137
9.17. INTEGRIDAD DEL REGISTRO	137
9.18. CONTROL DE VERSIONES	138
9.19. GESTIÓN DE CAMBIOS	138
<i>Clasificación posible</i>	139
9.20. MODIFICACIONES NO AUTORIZADAS	139
9.21. GESTIÓN DE PROVEEDORES	140
9.22. CLÁUSULAS CONTRACTUALES MÍNIMAS	140
9.23. PROPIEDAD DE LOS DATOS	141
9.24. SEGURIDAD Y CIBERSEGURIDAD	141

9.25. PLAN DE CONTINUIDAD	142
9.26. SUPERVISIÓN HUMANA EFECTIVA	142
9.27. DISEÑO DE LA INTERFAZ	143
9.28. FORMACIÓN	143
9.29. SESGO DE AUTOMATIZACIÓN	144
9.30. CAPA ASOCIADA AL MODELO	144
<i>Corrección inmediata</i>	144
<i>Acción correctiva</i>	144
<i>Acción preventiva</i>	145
9.31. INVESTIGACIÓN DEL FALSO NEGATIVO	145
9.32. AUDITORÍA INTERNA	145
<i>Preguntas de auditoría</i>	145
9.33. AUDITORÍA DEL PROVEEDOR	146
9.34. REVISIÓN POR LA DIRECCIÓN.....	146
9.35. PAQUETE MÍNIMO DE EVIDENCIA PARA AUDITORÍA	147
9.36. INTEGRACIÓN CON CERTIFICACIONES DE SEGURIDAD ALIMENTARIA	147
9.37. MODELO DE RESPONSABILIDAD RACI	148
9.38. MODELO DE FICHA DE CONTROL OPERATIVO	148
9.39. MODELO DE REGISTRO DE PREDICCIÓN	149
9.40. MODELO DE GESTIÓN DE CAMBIO	149
9.41. MODELO DE EVALUACIÓN DE INCIDENTES.....	149
9.42. INDICADORES PARA REVISIÓN PERIÓDICA	150
9.43. INTEGRACIÓN CON LA CULTURA DE SEGURIDAD ALIMENTARIA	150
9.44. PRINCIPIO DE RESPONSABILIDAD NO DELEGABLE.....	151
9.45. EL MODELO COMO EVIDENCIA Y NO COMO ORÁCULO	151
9.46. CONCLUSIÓN DEL BLOQUE	151
10. PROTOCOLO PRÁCTICO PARA LA VALIDACIÓN CIENTÍFICA DE MODELOS DE IA EN CALIDAD, VIDA ÚTIL Y SEGURIDAD ALIMENTARIA.....	153
10.1. FASE 0. CONSTITUCIÓN DEL EQUIPO Y GOBIERNO DEL PROYECTO.....	153
<i>Composición mínima</i>	153
<i>Preguntas iniciales</i>	154
<i>Entregables</i>	154
<i>Criterio de avance</i>	154
10.2. FASE 1. DEFINICIÓN DEL USO PREVISTO	154
<i>Plantilla de finalidad prevista</i>	155
<i>Ejemplo de calidad</i>	155
<i>Ejemplo de vida útil</i>	155
<i>Ejemplo de seguridad</i>	155
<i>Entregables</i>	155
<i>Criterio de avance</i>	155
10.3. FASE 2. ANÁLISIS DE RIESGO Y CRITICIDAD	155
<i>Identificación de errores</i>	155
<i>Evaluación del riesgo</i>	156
<i>Clasificación de criticidad</i>	156
<i>Consecuencias metodológicas</i>	156
<i>Entregables</i>	156
<i>Criterio de avance</i>	157
10.4. FASE 3. DEFINICIÓN DEL PATRÓN DE REFERENCIA	157
<i>Para calidad</i>	157
<i>Para vida útil</i>	157
<i>Para seguridad microbiológica</i>	157
<i>Validación de la referencia</i>	157

<i>Desacuerdos entre expertos</i>	158
<i>Entregables</i>	158
<i>Criterio de avance</i>	158
10.5. FASE 4. PLAN DE DATOS Y REPRESENTATIVIDAD	158
<i>Dimensiones mínimas</i>	158
<i>Matriz de cobertura</i>	159
<i>Unidad independiente</i>	159
<i>Datos positivos raros</i>	159
<i>Entregables</i>	160
<i>Criterio de avance</i>	160
10.6. FASE 5. DISEÑO ESTADÍSTICO Y PARTICIÓN	160
<i>Principios</i>	160
<i>Diseños posibles</i>	160
<i>Diseño recomendado por criticidad</i>	161
<i>Entregables</i>	161
<i>Criterio de avance</i>	161
10.7. FASE 6. PRE-ESPECIFICACIÓN DE MÉTRICAS Y CRITERIOS DE ACEPTACIÓN	161
<i>Clasificación</i>	161
<i>Regresión y vida útil</i>	161
<i>Seguridad microbiológica</i>	162
<i>Reglas jerárquicas</i>	162
<i>Entregables</i>	162
<i>Criterio de avance</i>	162
10.8. FASE 7. CONGELACIÓN DEL SISTEMA	163
<i>Elementos que deben bloquearse</i>	163
<i>Identificación de versión</i>	163
<i>Reproducibilidad</i>	163
<i>Entregables</i>	163
<i>Criterio de avance</i>	163
10.9. FASE 8. EJECUCIÓN CIEGA DE LA VALIDACIÓN EXTERNA	164
<i>Procedimiento</i>	164
<i>Gestión de exclusiones</i>	164
<i>Análisis primario</i>	164
<i>Análisis secundarios</i>	164
<i>Entregables</i>	165
<i>Criterio de avance</i>	165
10.10. FASE 9. PRUEBAS DE ROBUSTEZ Y ESTRÉS	165
<i>Robustez de entradas</i>	165
<i>Robustez instrumental</i>	165
<i>Robustez alimentaria</i>	165
<i>Robustez operacional</i>	166
<i>Robustez frente a datos fuera de dominio</i>	166
<i>Criterio de degradación</i>	166
<i>Entregables</i>	166
<i>Criterio de avance</i>	166
10.11. FASE 10. EVALUACIÓN HUMANA Y OPERACIONAL	166
<i>Preguntas</i>	167
<i>Diseño</i>	167
<i>Métricas operacionales</i>	167
<i>Supervisión humana</i>	167
<i>Entregables</i>	168
<i>Criterio de avance</i>	168
10.12. FASE 11. EVALUACIÓN COMPARATIVA Y DE IMPACTO	168

<i>Comparadores</i>	168
<i>Hipótesis posibles</i>	168
<i>Impacto en calidad</i>	168
<i>Impacto en vida útil</i>	168
<i>Impacto en seguridad</i>	169
<i>Impactos adversos</i>	169
<i>Entregables</i>	169
<i>Criterio de avance</i>	169
10.13. FASE 12. DECISIÓN DEL COMITÉ DE VALIDACIÓN	169
<i>Opciones</i>	169
<i>Regla de decisión</i>	170
<i>Acta de aprobación</i>	170
<i>Criterio de avance</i>	170
11. CRONOGRAMA TÉCNICO ORIENTATIVO	171
12. PAQUETE DOCUMENTAL MÍNIMO	172
13. MATRIZ DE ACEPTACIÓN PRÁCTICA	173
14. PROTOCOLO ESPECÍFICO PARA UN MODELO DE CALIDAD VISUAL	174
14.1. FINALIDAD	174
14.2. UNIDAD INDEPENDIENTE	174
14.3. CONJUNTO EXTERNO	174
14.4. MÉTRICAS.....	174
14.5. PRUEBAS DE ESTRÉS	174
14.6. CONDICIÓN DE APROBACIÓN	174
15. PROTOCOLO ESPECÍFICO PARA VIDA ÚTIL.....	175
15.1. FINALIDAD	175
15.2. DISEÑO	175
15.3. MÉTRICAS PRIORITARIAS	175
15.4. CRITERIO ESENCIAL	175
15.5. INTEGRACIÓN	175
15.6. EVIDENCIA COMPLEMENTARIA	175
16. PROTOCOLO ESPECÍFICO PARA SEGURIDAD MICROBIOLÓGICA.....	177
16.1. FINALIDAD	177
16.2. PRIORIDAD.....	177
16.3. DATOS	177
16.4. REGLA DE ACEPTACIÓN.....	177
16.5. BARRERAS	177
16.6. PROHIBICIÓN METODOLÓGICA	177
17. PRINCIPALES PUNTOS DE DECISIÓN	178
PUERTA 1. VIABILIDAD CIENTÍFICA	178
PUERTA 2. REPRESENTATIVIDAD	178
PUERTA 3. INDEPENDENCIA	178
PUERTA 4. RENDIMIENTO	178
PUERTA 5. ROBUSTEZ	178
PUERTA 6. IMPACTO.....	178
PUERTA 7. GOBERNANZA.....	178
18. CONCLUSIÓN DEL PROTOCOLO.....	179

1. Introducción: el problema de confundir rendimiento algorítmico con evidencia científica

La industria alimentaria está entrando en una etapa en la que los modelos de inteligencia artificial ya no se limitan a clasificar imágenes en proyectos experimentales. Se utilizan, o se pretende utilizarlos, para anticipar el deterioro microbiológico, calcular la vida útil remanente, detectar contaminantes o defectos, interpretar señales espectroscópicas, identificar desviaciones de proceso, predecir la aparición de *Listeria monocytogenes*, estimar parámetros sensoriales y decidir si un producto debe ser liberado, retenido, reprocesado o destruido.

En estos contextos, el error del modelo deja de ser una cuestión meramente informática. Se convierte en una desviación de calidad, una pérdida económica, una decisión incorrecta sobre la fecha de duración mínima o de caducidad y, en determinadas circunstancias, un fallo potencial del sistema de seguridad alimentaria.

El problema no reside solamente en que algunos modelos presenten un rendimiento insuficiente. El problema más grave es que pueden aparentar un rendimiento excelente durante el desarrollo y fracasar cuando se enfrentan a nuevas campañas, materias primas, proveedores, condiciones ambientales, equipos, fábricas o perfiles microbiológicos. La literatura reciente sobre aprendizaje automático aplicado al análisis de alimentos ha confirmado que determinados diseños de validación cruzada y el tratamiento incorrecto de réplicas analíticas pueden sobreestimar de forma sustancial la capacidad predictiva real. Cuando mediciones repetidas de una misma muestra aparecen simultáneamente en entrenamiento y ensayo, el modelo no demuestra que reconoce una categoría alimentaria desconocida; demuestra, en buena medida, que reconoce características de una muestra que ya había visto.

Esta distinción es decisiva. Un modelo puede memorizar particularidades de un lote, una cámara, un espectrómetro, un laboratorio o una línea de fabricación y, sin embargo, mostrar una aparente accuracy del 95 %, 98 % o incluso superior. El porcentaje, considerado aisladamente, no permite saber si el algoritmo ha aprendido una relación causal o tecnológicamente estable, si se está apoyando en variables espurias o si mantendrá su comportamiento cuando cambie el contexto.

La validación científica no consiste, por tanto, en encontrar la métrica más favorable para el modelo. Consiste en intentar demostrar, mediante un protocolo previamente definido, que el sistema conserva un rendimiento compatible con su finalidad cuando se somete a datos independientes, variabilidad realista y condiciones adversas previsibles.

Esta concepción coincide con la distinción clásica de Codex Alimentarius entre validación, monitorización y verificación. La validación obtiene evidencia de que una medida o combinación de medidas es capaz de controlar un peligro hasta un resultado especificado; la monitorización comprueba durante la operación que el proceso permanece dentro de los límites establecidos, y la verificación determina si el sistema está siendo implementado y funciona conforme a lo previsto. Aplicado a la inteligencia artificial, esto significa que:

- la validación inicial demuestra que el modelo puede cumplir el uso previsto;
- la monitorización vigila sus entradas, salidas y rendimiento durante la operación;
- la verificación periódica confirma que sigue utilizándose correctamente;
- la revalidación se activa cuando aparecen cambios relevantes o señales de degradación.

La evaluación de un modelo no termina, por tanto, en el momento de su despliegue. El NIST sitúa expresamente las actividades de prueba, evaluación, verificación y validación a lo largo de todo el ciclo de vida de la inteligencia artificial: desde la validación de los supuestos y datos durante el diseño, hasta la integración en producción, la recalibración, la vigilancia continuada, el seguimiento de incidentes y la detección de propiedades o impactos emergentes.

1.1. Un modelo no está validado en abstracto

La expresión «modelo validado» suele utilizarse de manera excesivamente categórica. Ningún algoritmo queda validado para cualquier producto, instalación, población o condición futura. Lo que puede demostrarse es que una versión determinada del modelo presenta unas prestaciones concretas, dentro de un dominio de aplicación definido y bajo unas condiciones de utilización especificadas.

Esta cautela terminológica ha sido subrayada por las guías contemporáneas de evaluación de modelos predictivos: no existe un modelo validado de forma absoluta, sino evaluaciones de su rendimiento en conjuntos, lugares, periodos y contextos determinados.

Por ello, una declaración científicamente defendible no debería formularse así:

«El modelo está validado con una precisión del 96 %.»

Debería formularse, por ejemplo, de este modo:

«La versión 2.3 del modelo alcanzó, en tres plantas no utilizadas durante el entrenamiento y durante dos campañas productivas independientes, una sensibilidad del 97,2 % para detectar lotes microbiológicamente no conformes, con una especificidad del 91,4 %, una tasa de falsos negativos del 2,8 % y los intervalos de confianza previamente establecidos. Su uso se limita a las matrices, rangos de pH, actividad de agua, temperaturas, equipos y condiciones de muestreo incluidos en el protocolo de validación.»

La diferencia no es retórica. La segunda formulación identifica el modelo, el contexto, las prestaciones, el evento crítico y los límites de aplicabilidad. Permite auditar la evidencia y evita extender ilegítimamente los resultados a condiciones no estudiadas.

1.2. La finalidad prevista determina la validación

Antes de seleccionar algoritmos o métricas debe definirse la afirmación que se pretende sostener. No es lo mismo afirmar que un modelo:

- ayuda a priorizar qué muestras serán analizadas;
- emite una alerta temprana que debe revisar un técnico;
- clasifica automáticamente productos por categoría comercial;
- estima la vida útil remanente;
- recomienda una fecha de caducidad;
- decide la liberación de un lote;
- actúa como barrera o medida de control dentro de un sistema APPCC.

A medida que aumenta la influencia del algoritmo sobre la seguridad o la conformidad del producto, aumenta también la exigencia de evidencia, independencia de la validación, control de incertidumbre, supervisión humana, trazabilidad y capacidad de intervención.

Un sistema utilizado exclusivamente para ordenar muestras según su probabilidad de no conformidad puede tolerar cierto número de falsos positivos, siempre que ninguna muestra crítica deje de analizarse. Un sistema encargado de descartar automáticamente la presencia de un patógeno no puede evaluarse con la misma lógica, porque un falso negativo puede implicar la liberación de un producto inseguro.

La gravedad del error, y no la sofisticación del algoritmo, debe determinar el rigor de la validación.

1.3. Los tres dominios requieren metodologías diferentes

Aunque calidad, vida útil y seguridad alimentaria están relacionadas, no constituyen el mismo problema estadístico ni tecnológico.

En **control de calidad**, el modelo puede clasificar defectos, estimar composición, detectar adulteraciones o asignar categorías comerciales. El rendimiento debe compararse con un método de referencia, considerar la variabilidad entre operarios o laboratorios y cuantificar los falsos rechazos y falsas aceptaciones.

En **predicción de vida útil**, el resultado suele ser continuo o temporal. No basta con conocer el error medio: es necesario determinar si existe sesgo sistemático, si el modelo sobreestima peligrosamente la duración, cómo se comporta en los extremos y qué intervalo de incertidumbre acompaña a cada predicción.

En **seguridad alimentaria**, el objetivo puede ser anticipar crecimiento, supervivencia o inactivación microbiana, detectar un peligro o identificar una condición de proceso incompatible con el control establecido. En estos casos debe diferenciarse entre errores *fail-safe*, que conducen a una decisión conservadora, y errores *fail-dangerous*, que subestiman el riesgo.

La microbiología predictiva dispone desde hace años de métricas específicas como el factor de sesgo, el factor de exactitud y las zonas de predicción aceptable. No obstante, incluso un modelo matemático bien ajustado debe demostrar su aplicabilidad a la matriz alimentaria concreta y, cuando proceda, complementarse con ensayos de desafío, estudios de durabilidad o datos industriales. Codex advierte expresamente que el uso efectivo de modelos matemáticos requiere una validación apropiada para la aplicación alimentaria específica y puede exigir pruebas adicionales.

En el ámbito de *Listeria monocytogenes*, la versión modificada en febrero de 2026 de la guía técnica del Laboratorio de Referencia de la Unión Europea mantiene precisamente esta lógica de integración entre modelización, ensayos de desafío y estudios de durabilidad, en coordinación con EN ISO 20976-1 y con los criterios microbiológicos europeos.

1.4. Hipótesis central de la publicación

La hipótesis que vertebra este trabajo es la siguiente:

La validez de un modelo de inteligencia artificial aplicado a alimentos no depende de una métrica aislada, sino de la concordancia entre su finalidad prevista, la calidad y representatividad de los datos, el diseño de

evaluación, el coste de los errores, la incertidumbre, la validación externa y la vigilancia durante todo su ciclo de vida.

De esta hipótesis se derivan varias consecuencias.

Una *accuracy* (precisión) elevada puede ser irrelevante cuando las clases están desequilibradas. Una sensibilidad excelente puede coexistir con un volumen inasumible de falsos positivos. Una curva ROC favorable no demuestra que las probabilidades estén calibradas. Una validación cruzada no sustituye a una validación externa. Un error medio reducido puede ocultar sobreestimaciones peligrosas de vida útil. Una explicación SHAP plausible no prueba que el modelo sea correcto. Y un modelo que funcionó durante su implantación puede dejar de hacerlo cuando cambian la materia prima, el proceso o el sistema de medida.

La validación científica debe convertir todas estas cuestiones en un protocolo prospectivo, documentado y auditable. Solo entonces la inteligencia artificial puede pasar de ser una herramienta exploratoria a una herramienta que pueda constituir evidencia técnicamente defendible para la toma de decisiones alimentarias.

2. Arquitectura científica de la validación: qué debe demostrarse antes de evaluar las métricas

La validación científica de un modelo de inteligencia artificial no comienza calculando sensibilidad, especificidad o error medio. Comienza definiendo con precisión **qué afirmación pretende demostrar la empresa**, en qué condiciones será válida y qué consecuencias tendrá una predicción incorrecta.

Este principio, aparentemente elemental, se incumple con frecuencia. Numerosos proyectos alimentarios comienzan acumulando datos y entrenando algoritmos sin haber establecido previamente el uso previsto, la población de aplicación, el método de referencia, la unidad estadística, el horizonte temporal, los límites de funcionamiento ni los criterios de aceptación. El resultado puede ser un modelo matemáticamente sofisticado, pero científicamente indeterminado: predice algo, aunque no está claro qué representa exactamente su salida, contra qué verdad debe compararse ni qué nivel de error resulta admisible.

Por ello, la validación debe estructurarse como un proceso de demostración prospectiva. Antes de acceder al conjunto de prueba definitivo deben quedar documentados:

1. • la finalidad prevista del sistema;
2. • la variable o evento que se pretende predecir;
3. • la población de productos y procesos a la que se aplicará;
4. • el patrón o método de referencia;
5. • la unidad de análisis;
6. • la estrategia de muestreo;
7. • las métricas principales y secundarias;
8. • los subgrupos críticos;
9. • los criterios de aceptación;
10. • las condiciones que obligarán a rechazar, recalibrar o revalidar el modelo.

La documentación previa no constituye una formalidad administrativa. Evita que, una vez conocidos los resultados, se seleccionen retrospectivamente las métricas, subgrupos o umbrales que presentan al modelo de la manera más favorable. En términos metodológicos, reduce el riesgo de sobreajuste no solo del algoritmo, sino también del propio procedimiento de evaluación.

2.1. La finalidad prevista como núcleo de la validación

Todo protocolo debe comenzar con una declaración inequívoca del *intended purpose* o finalidad prevista. Esta declaración debe describir qué hace el modelo, para quién, sobre qué productos, utilizando qué datos y con qué efecto dentro del proceso de decisión.

Una finalidad correctamente formulada podría ser la siguiente:

El sistema estima la probabilidad de que un lote de producto cárnico cocido exceda el límite microbiológico interno establecido para *Listeria monocytogenes* al final de una vida útil de 35 días a 4 °C, utilizando datos de formulación, pH,

actividad de agua, tratamiento térmico, temperatura de almacenamiento y resultados ambientales. La predicción se utiliza para priorizar lotes para investigación y ensayo microbiológico adicional, pero no sustituye el análisis legalmente exigible ni autoriza por sí sola la liberación del producto.

Esta descripción permite identificar inmediatamente varios elementos esenciales:

- • el producto;
- • el microorganismo;
- • el horizonte temporal;
- • las condiciones de conservación;
- • las variables de entrada;
- • la función del algoritmo;
- • su relación con el método microbiológico;
- • el nivel de autonomía decisoria.

Una formulación insuficiente sería afirmar simplemente que el modelo «predice el riesgo de *Listeria*». No precisa qué evento define el riesgo, en qué momento, en qué matriz, en qué condiciones de temperatura ni para qué decisión se utilizará.

La finalidad determina el diseño completo de la validación. Un modelo de priorización puede evaluarse atendiendo a su capacidad para concentrar casos problemáticos en un subconjunto manejable. Un modelo que interviene en la liberación de producto debe demostrar que la probabilidad de aceptar un lote no conforme permanece por debajo de un límite extremadamente exigente. Un modelo de vida útil debe evaluarse frente al tiempo real hasta el deterioro o incumplimiento, no solo frente a valores microbiológicos intermedios.

El Reglamento de Inteligencia Artificial de la Unión Europea utiliza también la finalidad prevista como eje para determinar la adecuación de los datos, el rendimiento esperado, las limitaciones y las condiciones razonablemente previsibles de utilización. Sus exigencias específicas sobre gobernanza de datos, exactitud y robustez se aplican jurídicamente a los sistemas incluidos en su ámbito y, particularmente, a los de alto riesgo. No todo modelo empleado en una industria alimentaria tendrá automáticamente esa clasificación. Sin embargo, los principios del Reglamento ofrecen un referente metodológico sólido: los conjuntos de entrenamiento, validación y prueba deben ser relevantes, suficientemente representativos y adecuados al entorno geográfico, contextual y funcional donde se utilizará el sistema.

2.2. El objeto de validación no es únicamente el algoritmo

En una aplicación industrial, el objeto que debe validarse no es un archivo matemático aislado, sino el **sistema sociotécnico completo** que genera y utiliza la predicción.

Ese sistema comprende, como mínimo:

- • sensores, cámaras, espectrómetros o equipos de laboratorio;
- • procedimientos de muestreo;
- • pretratamiento y transformación de datos;
- • variables introducidas manualmente;
- • arquitectura y parámetros del modelo;
- • umbral de decisión;

- interfaz de usuario;
- reglas de interpretación;
- intervención de operarios;
- integración con sistemas MES, ERP, LIMS o SCADA;
- acciones desencadenadas por la salida;
- procedimientos de revisión y anulación.

Un modelo espectroscópico puede haber sido correctamente entrenado, pero resultar inválido en producción si el instrumento no está adecuadamente calibrado, si cambia la geometría de lectura, si las muestras se presentan a una temperatura diferente o si el preprocesamiento aplicado en línea no coincide con el utilizado durante el desarrollo.

Del mismo modo, un clasificador visual puede presentar una sensibilidad elevada en imágenes controladas y fracasar cuando aparecen condensación, suciedad en la lente, variaciones de iluminación, cambios de velocidad de la cinta, nuevos formatos de envase o productos parcialmente ocultos.

La validación debe abarcar la cadena completa:

muestra física → captura de datos → preprocesamiento → modelo → salida → interpretación humana → decisión

Un fallo en cualquiera de estas etapas puede invalidar el resultado, aunque el algoritmo conserve intactas sus propiedades matemáticas.

El NIST recomienda documentar los conjuntos de prueba, las métricas, los instrumentos, los procesos y los materiales utilizados durante las actividades de prueba, evaluación, verificación y validación. También propone identificar las limitaciones, las condiciones de uso, los resultados de validación, los planes de despliegue, la monitorización y la gestión de cambios.

2.3. Definición del resultado verdadero: el patrón de referencia

Todo modelo supervisado aprende y se evalúa frente a una variable considerada verdadera. Sin embargo, en alimentación esa «verdad» puede contener error.

La calidad del patrón de referencia establece un límite superior a la calidad que puede atribuirse al algoritmo. Un sistema no puede validarse rigurosamente frente a etiquetas ambiguas, métodos analíticos inadecuados, paneles sensoriales no entrenados o resultados microbiológicos obtenidos con un muestreo insuficiente.

El patrón de referencia debe analizarse en cuatro dimensiones.

2.3.1. Validez analítica

Debe demostrarse que el método mide realmente el analito, microorganismo, atributo o condición de interés con precisión, exactitud y límites de detección adecuados.

Cuando el modelo predice composición, fraude o contaminación química, la referencia debería proceder de un método normalizado, validado o adecuadamente caracterizado. Cuando predice estado microbiológico, deben

especificarse el método de cultivo o detección, la porción analítica, el plan de muestreo, el límite de cuantificación y el tratamiento de resultados censurados.

2.3.2. Validez tecnológica

El resultado medido debe representar el fenómeno industrial que se pretende controlar.

Por ejemplo, un modelo de oxidación lipídica no queda necesariamente validado porque reproduzca el índice de peróxidos si el deterioro sensorial del producto depende también de aldehídos secundarios, interacción con aromas, tipo de envase o percepción del consumidor.

De igual modo, predecir recuentos de aerobios mesófilos no equivale automáticamente a predecir vida útil comercial. El fin de vida útil puede estar determinado por levaduras, mohos, bacterias ácido-lácticas, pérdida de textura, separación de fases, rancidez, decoloración o rechazo sensorial.

2.3.3. Reproducibilidad

Debe conocerse la variabilidad del patrón de referencia entre réplicas, días, operarios, laboratorios o lotes de reactivos. Cuando la incertidumbre del método de referencia es elevada, parte de la aparente discrepancia del modelo no procede del algoritmo, sino del propio sistema de medida.

Por ello, resulta incorrecto exigir al modelo una precisión aparente superior a la reproducibilidad de la técnica utilizada como referencia.

2.3.4. Ausencia de contaminación de la etiqueta

La etiqueta no debe incorporar directa o indirectamente información que solo estará disponible después de la predicción.

Si un modelo pretende anticipar el deterioro en el día 0, sus variables de entrada no pueden contener datos obtenidos en el día 20. Tampoco debe incluir variables administrativas que reflejen decisiones posteriores, como la clasificación final asignada por el departamento de calidad, si esa clasificación está causalmente vinculada al resultado que se pretende predecir.

Este fenómeno, denominado *target leakage* o fuga de información hacia la variable objetivo, puede producir modelos con resultados extraordinarios durante la evaluación y completamente inútiles en condiciones prospectivas.

2.4. La unidad estadística: muestra, lote, producto o proceso

Uno de los errores más graves en la validación de IA alimentaria consiste en confundir el número de observaciones digitales con el número de unidades independientes.

Una misma muestra puede generar:

- decenas de imágenes;
- centenas de espectros;

- • múltiples píxeles hiperespectrales;
 - • mediciones repetidas;
 - • varios tiempos de lectura;
 - • diferentes recortes o aumentos artificiales;
 - • réplicas técnicas y analíticas.
- Estas observaciones no equivalen a muestras biológicas o industriales independientes.

Si 500 espectros proceden de 20 lotes, el tamaño efectivo del estudio se aproxima más a 20 unidades industriales que a 500 observaciones completamente independientes. La correlación entre mediciones de una misma muestra o lote reduce la información estadística adicional aportada por cada repetición.

La unidad de partición debe coincidir con la unidad sobre la que se pretende generalizar. Si el modelo se utilizará en lotes nuevos, los datos de un mismo lote no pueden distribuirse entre entrenamiento y prueba. Si se pretende generalizar a proveedores desconocidos, la validación debe reservar proveedores completos. Si la predicción se aplicará a futuras campañas, debe realizarse una separación temporal real.

La partición aleatoria por imagen, espectro o lectura suele generar fuga de identidad. El modelo encuentra en el conjunto de prueba señales procedentes de las mismas muestras que utilizó para aprender y obtiene una estimación excesivamente optimista.

La estrategia correcta puede requerir:

- • separación por lote;
 - • separación por día de producción;
 - • separación por proveedor;
 - • separación por campaña;
 - • separación por planta;
 - • separación por instrumento;
 - • separación por formulación;
 - • separación por periodo temporal.
- La elección depende de la afirmación de generalización que se quiera demostrar.

2.5. Dominio de aplicación y población objetivo

La población objetivo no puede definirse de forma genérica como «productos alimentarios» o «productos cárnicos». Debe delimitarse mediante las variables que condicionan el comportamiento del fenómeno y del sistema de medida.

En un modelo de vida útil microbiológica, el dominio podría quedar definido por:

- • familia de producto;
- • formulación;
- • intervalo de pH;
- • intervalo de actividad de agua;
- • contenido de sal y conservantes;

- • tratamiento térmico;
- • sistema de envasado;
- • concentración de oxígeno residual;
- • carga microbiológica inicial;
- • condiciones de almacenamiento;
- • variabilidad estacional;
- • flora alterante predominante.

• **En un modelo espectroscópico deben añadirse:**

- • modelo y configuración del instrumento;
- • región espectral;
- • resolución;
- • preparación de la muestra;
- • temperatura;
- • humedad;
- • granulometría;
- • espesor;
- • procedimiento de calibración;
- • algoritmo de preprocesamiento.

• **En un sistema visual serán críticas variables como:**

- • iluminación;
- • distancia y ángulo;
- • velocidad de línea;
- • resolución;
- • color del fondo;
- • formato del envase;
- orientación;
- oclusiones;
- reflejos;
- condensación;
- suciedad.

El dominio de aplicación debe formularse de modo que permita responder a una pregunta operativa: **¿esta nueva observación pertenece a la población en la que el modelo demostró su rendimiento?**

Si la respuesta no puede determinarse, la empresa carece de un mecanismo fiable para identificar usos fuera de validación.

2.6. Representatividad no significa mera proporcionalidad

Un conjunto puede reproducir las proporciones habituales de producción y, sin embargo, no ser adecuado para validar un sistema de seguridad.

Supóngase una planta en la que los lotes no conformes representan el 0,2 %. Una muestra puramente proporcional de 10.000 lotes contendría aproximadamente 20 positivos. Ese número puede ser insuficiente para estimar con precisión la sensibilidad, especialmente si se pretende demostrar una tasa de falsos negativos muy baja.

La validación debe combinar dos objetivos diferentes:

1. • reproducir el contexto real de utilización;
2. • reunir suficientes casos críticos para estimar el rendimiento con incertidumbre aceptable.

• **Por ello, puede ser necesario enriquecer deliberadamente la muestra de validación con:**

- • contaminaciones poco frecuentes;
- • desviaciones de temperatura;
- • formulaciones extremas;
- • materias primas atípicas;
- • defectos raros;
- • concentraciones próximas al límite analítico;
- • condiciones de frontera;
- • nuevos proveedores;
- • fallos de sensor;
- • imágenes de baja calidad.

Este enriquecimiento no invalida el estudio, siempre que se documente y que las métricas dependientes de la prevalencia —como los valores predictivos— se recalculen para escenarios operativos realistas.

La representatividad debe evaluarse en varias dimensiones:

$R = f(\text{producto, proceso, tiempo, planta, proveedor, equipo, ambiente, evento crítico})$

No existe una única medida universal de representatividad. Debe demostrarse mediante una matriz de cobertura que compare la población prevista con la composición de los conjuntos de entrenamiento, validación interna y prueba externa.

2.7. Separación entre entrenamiento, validación y prueba

Los términos se utilizan con frecuencia de manera imprecisa, pero representan funciones distintas.

Conjunto de entrenamiento

Se utiliza para ajustar los parámetros internos del modelo: coeficientes, pesos, árboles, reglas de partición o representaciones latentes.

Conjunto de validación

Se emplea durante el desarrollo para seleccionar hiperparámetros, comparar arquitecturas, elegir variables, ajustar preprocesamientos y establecer provisionalmente umbrales.

Conjunto de prueba

Se reserva para obtener una evaluación final del sistema ya congelado. No debe participar directa ni indirectamente en la selección del modelo.

Cuando el equipo consulta reiteradamente el rendimiento del conjunto de prueba y modifica después el algoritmo, ese conjunto deja de ser verdaderamente independiente. Aunque no se utilice formalmente para entrenar los pesos, ha influido en las decisiones de desarrollo.

El conjunto de prueba debe permanecer bloqueado hasta que se hayan fijado:

- • la versión del código;
- • los pesos del modelo;
- • el preprocesamiento;
- • las variables;
- • las reglas de exclusión;
- • el umbral de decisión;
- • las métricas;
- • los criterios de aceptación.

Una vez abierto, no debe reutilizarse indefinidamente para las siguientes versiones. La repetición convierte el test en un instrumento de optimización y favorece un nuevo sobreajuste.

2.8. Validación interna y validación externa

La validación interna estima cuánto puede generalizar el modelo dentro de la población de desarrollo. Incluye técnicas como:

- • partición entrenamiento-prueba;
 - • validación cruzada;
 - • *bootstrap*;
 - • validación cruzada agrupada;
 - • validación temporal interna.
- Estas técnicas son necesarias, pero no suficientes cuando el modelo se utilizará fuera del contexto exacto en el que se desarrolló.

La **validación externa** evalúa el modelo en datos que presentan una separación relevante respecto a la población de desarrollo. Esta separación puede ser:

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

No todo conjunto denominado externo proporciona el mismo nivel de evidencia. Existe una jerarquía práctica.

Nivel 1. Separación aleatoria dentro de la misma base

Los datos proceden de las mismas campañas, instalaciones y procedimientos. Es el nivel más débil.

Nivel 2. Separación agrupada

Se reservan lotes, días, productos o proveedores completos. Reduce la fuga entre observaciones correlacionadas.

Nivel 3. Validación temporal

El modelo se desarrolla con datos históricos y se evalúa prospectivamente sobre producción posterior.

Nivel 4. Validación en otra línea o planta

Permite comprobar la portabilidad frente a diferencias de proceso, ambiente, equipamiento y organización.

Nivel 5. Validación independiente multicéntrica

La evaluación se realiza en varias instalaciones no participantes en el desarrollo, preferiblemente por un equipo que no haya intervenido en la construcción del algoritmo.

Nivel 6. Evaluación prospectiva de impacto

Además del rendimiento predictivo, se analiza si el uso del sistema mejora realmente los resultados industriales o sanitarios sin generar nuevos daños.

La validación externa no consiste simplemente en que los datos «no hayan sido vistos» por el modelo. Deben representar una fuente de variabilidad relevante para el uso futuro.

2.9. Validación del modelo y validación del impacto

Un modelo puede predecir correctamente y, sin embargo, no mejorar el sistema de calidad o seguridad.

Puede ocurrir que:

- • las alertas lleguen demasiado tarde;
 - • los responsables no comprendan la salida;
 - • el volumen de falsos positivos sea inasumible;
 - • los operadores ignoren las recomendaciones;
 - • la intervención activada no sea eficaz;
 - • el sistema desplace recursos desde controles más útiles;
 - • se genere dependencia excesiva del algoritmo;
 - • se relajen medidas preventivas ya consolidadas.
- Por ello deben distinguirse dos preguntas.

- **Validación del rendimiento predictivo:**

- ¿El modelo estima correctamente el resultado?

- **Validación del impacto operativo:**

- ¿Su utilización mejora la decisión o el control respecto al procedimiento existente?

La segunda pregunta exige comparar estrategias completas. Puede requerir un diseño prospectivo entre líneas, plantas o periodos con y sin asistencia algorítmica, evaluando:

- • reducción de producto no conforme;
- • disminución de tiempo de respuesta;
- • reducción de desperdicio;
- • número de alertas investigadas;
- • falsos rechazos;
- • reclamaciones;
- • incidencias microbiológicas;
- • coste total del control;
- • carga de trabajo;
- • comportamiento de los usuarios.

El algoritmo no debe compararse únicamente con la ausencia de modelo. Debe compararse con la práctica vigente y, cuando exista, con métodos estadísticos o reglas de decisión más sencillas. Una red neuronal compleja carece de justificación si una regresión, una carta de control o una regla basada en límites críticos ofrece resultados equivalentes con mayor transparencia y menor coste.

2.10. Independencia de la validación y control de conflictos

La validación dirigida exclusivamente por el equipo que desarrolló el modelo presenta un riesgo evidente de sesgo de confirmación. Los desarrolladores conocen las fortalezas, las debilidades y las decisiones adoptadas durante el ajuste; pueden seleccionar inadvertidamente escenarios favorables o interpretar de forma benevolente resultados ambiguos.

La independencia no exige necesariamente contratar siempre a una entidad externa, pero sí separar funciones y responsabilidades. Una estructura razonable incluiría:

- • equipo de desarrollo;
- • responsables del proceso alimentario;
- • calidad y seguridad alimentaria;
- • estadística o metrología;
- • tecnologías de la información;
- • cumplimiento normativo;
- • validadores independientes;
- • responsable final de aprobación.

El NIST recomienda separar, cuando sea posible, la gestión del desarrollo de las funciones de prueba y evaluación, de manera que estas últimas puedan cuestionar el sistema y promover correcciones independientes. También aconseja establecer responsabilidades claras sobre quién mantendrá, verificará, monitorizará y actualizará el modelo después de su despliegue.

La persona o comité que aprueba el modelo debe disponer de autoridad real para concluir:

- • aprobado;
 - • aprobado con restricciones;
 - • aprobado para uso experimental;
 - • sometido a recalibración;
 - • no aprobado.
-
- La validación pierde su función si el rechazo del modelo no constituye una posibilidad organizativa real.

3. El protocolo maestro de validación

El protocolo maestro es el documento que transforma la evaluación en una investigación reproducible. Debe elaborarse antes de ejecutar el ensayo final e identificar inequívocamente el sistema evaluado.

Como mínimo, deberá contener los siguientes elementos.

3.1. Identificación del sistema

- nombre del modelo;
- versión;
- fecha de congelación;
- código fuente o repositorio;
- pesos y parámetros;
- librerías y dependencias;
- hardware;
- configuración del entorno;
- preprocesamiento;
- variables de entrada;
- formato de salida;
- umbral de decisión.

Una modificación aparentemente menor puede alterar el rendimiento. Cambiar una librería, la resolución de una imagen, una técnica de normalización o un umbral genera, en términos de validación, una versión potencialmente distinta.

3.2. Pregunta de validación

Debe expresarse como una hipótesis verificable.

Ejemplo:

La versión 3.1 del modelo identifica lotes que superarán el criterio microbiológico interno al final de la vida útil con una sensibilidad no inferior al 97 %, manteniendo una especificidad no inferior al 85 %, en productos pertenecientes al dominio de aplicación definido.

Cuando se trate de vida útil:

La estimación del tiempo hasta el incumplimiento presentará un error absoluto medio no superior a tres días y no sobreestimaré la vida útil real en más de dos días en más del 5 % de los lotes evaluados.

La segunda condición es esencial. Un error medio aceptable puede ocultar una cola de sobreestimaciones peligrosas.

3.3. Diseño del estudio

El protocolo debe establecer:

- • naturaleza retrospectiva o prospectiva;
- • centros o plantas participantes;
- • periodo de obtención;
- • criterios de inclusión y exclusión;
- • unidad estadística;
- • tamaño muestral;
- • método de referencia;
- • plan de muestreo;
- • gestión de datos perdidos;
- • tratamiento de resultados indeterminados;
- • análisis por subgrupos;
- • intervalos de confianza;
- • análisis de sensibilidad;
- • reglas de interrupción.

3.4. Criterios de aceptación

Los criterios deben quedar fijados antes de conocer los resultados finales y estar relacionados con el riesgo.

No basta con exigir que la sensibilidad «sea alta». Debe especificarse:

$$\text{Sensibilidad estimada} \geq 0,97$$

y, cuando el riesgo lo justifique:

$$\text{IC 95 \%}, \text{ límite inferior(sensibilidad)} \geq 0,93$$

La segunda condición evita aprobar un modelo cuyo valor puntual parece elevado, pero cuya incertidumbre permite un rendimiento real inaceptable.

Los criterios deben incluir no solo métricas globales, sino también:

- • rendimiento mínimo por producto;
- • rendimiento por planta;
- • comportamiento en condiciones extremas;
- • tasa máxima de falsos negativos;
- • tasa máxima de predicciones no concluyentes;
- • calibración;
- • estabilidad temporal;
- • capacidad de detección de datos fuera de dominio;
- • tiempo de respuesta;
- • disponibilidad;
- • posibilidad de intervención humana.

3.5. Resultado de la validación

La conclusión debe ser explícita y limitarse a la evidencia obtenida.

Un informe válido no debería concluir únicamente que «los resultados son satisfactorios». Debería declarar:

El modelo cumple los criterios principales de sensibilidad y especificidad en la población global, pero no alcanza la sensibilidad mínima preespecificada en productos con actividad de agua inferior a 0,94. Se aprueba exclusivamente para los productos y condiciones incluidos en el dominio restringido descrito. La utilización en matrices de menor actividad de agua queda prohibida hasta completar una validación adicional.

Esta forma de conclusión convierte la validación en un instrumento de control real. Permite aprobar el modelo allí donde existe evidencia y restringirlo donde esa evidencia es insuficiente.

La validación científica no consiste en decidir si el modelo es «bueno» o «malo». Consiste en determinar con precisión **para qué es apto, con qué incertidumbre, en qué condiciones y hasta qué nivel de riesgo puede permitirse que influya en una decisión alimentaria.**

4. Sensibilidad, especificidad y errores de clasificación: la anatomía real del rendimiento

La evaluación de un modelo clasificador comienza comparando sus predicciones con el patrón de referencia. Sin embargo, la utilidad de esa comparación depende de cómo se definan la clase positiva, la clase negativa y el evento que se desea evitar.

En seguridad alimentaria, la clase positiva debería corresponder, por regla general, al evento adverso que el sistema debe detectar: presencia de un patógeno, lote no conforme, contaminación, defecto crítico, vida útil insuficiente o condición de proceso insegura.

Esta convención no es meramente terminológica. Determina la interpretación de todas las métricas. Si se invierten las clases, una sensibilidad aparentemente elevada podría estar midiendo la capacidad de reconocer lotes seguros, no la de detectar los peligrosos.

4.1. La matriz de confusión

Para un problema binario, las predicciones se distribuyen en cuatro categorías:

Resultado real	Predicción positiva	Predicción negativa
Evento presente	Verdadero positivo, VP	Falso negativo, FN
Evento ausente	Falso positivo, FP	Verdadero negativo, VN

- Cada celda representa una consecuencia distinta.
- **Verdadero positivo**
 - El modelo identifica correctamente un producto, lote o condición problemática.
- **Ejemplos:**
 - detecta un cuerpo extraño realmente presente;
 - clasifica como no conforme un lote que excede el criterio microbiológico;
 - predice una vida útil insuficiente cuando efectivamente lo es;
 - alerta sobre una desviación de temperatura que compromete la seguridad.
- El verdadero positivo representa la detección eficaz del evento.
- **Verdadero negativo**
 - El modelo clasifica correctamente como conforme o segura una observación que realmente lo es.

- Esta capacidad resulta esencial para evitar retenciones, análisis adicionales, reprocesados o destrucciones innecesarias.
- **Falso positivo**
 - El modelo emite una alerta cuando el evento adverso no está presente.
- **En términos industriales, puede ocasionar:**
 - • bloqueo innecesario de lotes;
 - • aumento de análisis confirmatorios;
 - • reprocesados evitables;
 - • pérdida de materia prima;
 - • retrasos logísticos;
 - • fatiga de alertas;
 - • pérdida de confianza en el sistema.
- Un exceso de falsos positivos puede hacer inviable un modelo aunque su sensibilidad sea muy elevada.
- **Falso negativo**
 - El modelo clasifica como conforme o segura una observación que realmente presenta el evento adverso.

En calidad, puede suponer la aceptación de un producto defectuoso. En vida útil, puede traducirse en una sobreestimación del periodo comercial. En seguridad alimentaria, puede permitir la liberación de un lote peligroso.

El falso negativo es, habitualmente, el error de mayor gravedad.

No obstante, tampoco debe afirmarse de forma absoluta que todo falso negativo tiene siempre mayor coste que cualquier falso positivo. La relación depende del proceso. Un falso positivo que provoque la destrucción de un lote de gran valor, la interrupción de una línea crítica o la retirada preventiva de producto puede generar consecuencias muy relevantes. La validación debe cuantificar ambos tipos de error en el contexto concreto.

4.2. Sensibilidad

La sensibilidad mide la proporción de eventos positivos reales que el modelo identifica correctamente:

$$\text{Sensibilidad} = \text{[VP]} / \text{[VP+FN]}$$

También se denomina:

- • tasa de verdaderos positivos;
- • *recall* de la clase positiva;
- • *true positive rate*;
- • capacidad de detección.

Una sensibilidad del 98 % significa que el modelo detecta 98 de cada 100 casos positivos presentes en la muestra evaluada. No significa que una predicción positiva tenga un 98 % de probabilidad de ser correcta, ni que el sistema acierte el 98 % de las veces en producción.

En seguridad alimentaria, la sensibilidad suele constituir una métrica prioritaria porque está directamente relacionada con la tasa de falsos negativos:

$$\text{Tasa de falsos negativos} = 1 - \text{Sensibilidad} = \text{FN} / [\text{VP} + \text{FN}]$$

Si la sensibilidad es del 98 %, la tasa de falsos negativos es del 2 %.

Este porcentaje debe interpretarse en conjunto con la frecuencia y gravedad del peligro. Una tasa del 2 % puede parecer baja, pero resultar inaceptable cuando el modelo participa en miles de decisiones y el error puede liberar un producto inseguro.

Supóngase que una planta utiliza el sistema sobre 500.000 unidades anuales y que 1.000 contienen realmente el defecto o peligro objetivo. Con una sensibilidad del 98 %, se esperarían aproximadamente:

$$1.000 \times 0,02 = 20$$

falsos negativos anuales, siempre que el rendimiento observado se mantuviese estable en producción.

La cifra absoluta revela mejor el riesgo que el porcentaje aislado.

4.3. Especificidad

La especificidad mide la proporción de casos negativos reales correctamente identificados:

$$\text{Especificidad} = \text{VN} / [\text{VN} + \text{FP}]$$

También se denomina tasa de verdaderos negativos.

Una especificidad del 90 % indica que el modelo clasifica correctamente como negativos 90 de cada 100 casos en los que el evento no está presente. El 10 % restante genera falsos positivos.

La tasa de falsos positivos se expresa como:

$$\text{Tasa de falsos positivos} = 1 - \text{Especificidad} = \text{FP} / [\text{VN} + \text{FP}]$$

La especificidad tiene una importancia operacional enorme cuando la mayoría de la producción es conforme. Incluso una tasa de falsos positivos aparentemente reducida puede producir un volumen muy elevado de alertas.

Si el 99,5 % de 100.000 lotes o unidades son negativos y la especificidad del modelo es del 95 %, el número esperado de falsos positivos sería:

$$99.500 \times 0,05 = 4.975$$

Un sistema de estas características podría ser industrialmente inviable, aunque detectase prácticamente todos los positivos.

4.4. Sensibilidad y especificidad no son probabilidades posteriores

Sensibilidad y especificidad se calculan condicionando sobre el estado real:

$$P(\text{predicción positiva} | \text{evento presente})$$

y

$$P(\text{predicción negativa} | \text{evento ausente})$$

Sin embargo, en la práctica el usuario suele formular una pregunta diferente:

- Cuando el modelo emite una alerta, ¿qué probabilidad existe de que el peligro esté realmente presente?
- Esa pregunta no se responde con la sensibilidad, sino con el valor predictivo positivo.

De manera análoga, cuando el sistema declara que un lote es conforme, interesa conocer la probabilidad de que realmente lo sea. Esa información corresponde al valor predictivo negativo.

4.5. Valor predictivo positivo

El valor predictivo positivo expresa la proporción de predicciones positivas que son verdaderamente positivas:

$$\text{VPP} = [\text{VP}] / [\text{VP} + \text{FP}]$$

También recibe el nombre de precisión positiva o *precision*.

Un VPP del 40 % indica que, de cada 100 alertas emitidas, aproximadamente 40 corresponden a eventos reales y 60 son falsas alarmas.

El VPP depende de:

- la sensibilidad;
 - la especificidad;
 - la prevalencia del evento.
- **Puede expresarse mediante el teorema de Bayes:**

$$\text{VPP} = [\text{Sensibilidad} \times \text{Prevalencia}] / [\text{Sensibilidad} \times \text{Prevalencia} + (1 - \text{Especificidad}) \times (1 - \text{Prevalencia})]$$

Esta dependencia explica por qué un modelo con excelentes prestaciones puede producir numerosas falsas alarmas cuando el evento es muy infrecuente.

4.6. Valor predictivo negativo

El valor predictivo negativo representa la proporción de predicciones negativas que son realmente negativas:

$$VPN = [VN] / [VN+FN]$$

Su formulación bayesiana es:

$$VPN = [Especificidad \times (1 - Prevalencia)] / [(1 - Sensibilidad) \times Prevalencia + Especificidad \times (1 - Prevalencia)]$$

Un VPN muy elevado puede resultar engañoso cuando la prevalencia del evento es extremadamente baja. Si casi todos los lotes son conformes, incluso un sistema poco informativo puede alcanzar un valor predictivo negativo aparentemente excelente.

Por ello, el VPN no debe utilizarse aisladamente para justificar la seguridad del sistema.

4.7. El efecto de la prevalencia

Considérese un modelo con:

$$\text{Sensibilidad} = 99\%$$

$$\text{Especificidad} = 95\%$$

A primera vista, su rendimiento parece sobresaliente.

Supóngase ahora que el evento adverso aparece en uno de cada mil casos:

$$\text{Prevalencia} = 0,1\%$$

En una población de 100.000 observaciones habría aproximadamente:

- 100 positivos reales;
- 99.900 negativos reales.

• El modelo produciría:

$$VP = 100 \times 0,99 = 99$$

$$FN = 100 \times 0,01 = 1$$

$$VN = 99.900 \times 0,95 = 94.905$$

$$FP = 99.900 \times 0,05 = 4.995$$

El valor predictivo positivo sería:

$$VPP = [99] / [99 + 4.995] \approx 1,94\%$$

Aunque la sensibilidad sea del 99 % y la especificidad del 95 %, solo alrededor de dos de cada cien alertas corresponderían a eventos reales.

Este resultado no significa que el modelo carezca necesariamente de utilidad. Podría seguir siendo valioso para seleccionar muestras, siempre que la empresa disponga de capacidad suficiente para confirmar miles de alertas y que el

coste de no detectar el evento sea muy alto. Pero demuestra que una métrica aparentemente excelente no garantiza viabilidad operativa.

4.8. Accuracy o exactitud global

La *accuracy* mide la proporción total de predicciones correctas:

$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN}$$

Es una medida intuitiva, pero frecuentemente inadecuada cuando las clases están desequilibradas.

Si solo el 0,2 % de los lotes son no conformes, un sistema trivial que clasifique todos los lotes como conformes alcanzaría:

$$Accuracy = 99,8\%$$

Sin embargo:

$$Sensibilidad = 0\%$$

El modelo no detectaría ningún lote problemático.

Por ello, una elevada *accuracy* no debe utilizarse como evidencia principal en problemas de seguridad alimentaria, fraude, contaminación o defectos infrecuentes.

La exactitud global puede resultar útil cuando:

- las clases están razonablemente equilibradas;
- los costes de los errores son similares;
- se presenta junto con métricas desagregadas;
- no oculta subgrupos con bajo rendimiento.

4.9. Exactitud balanceada

La *balanced accuracy* o exactitud balanceada calcula el promedio entre sensibilidad y especificidad:

$$Balanced Accuracy = \frac{Sensibilidad + Especificidad}{2}$$

Esta métrica concede el mismo peso a las dos clases, con independencia de su frecuencia.

Resulta más informativa que la *accuracy* convencional en conjuntos desequilibrados, pero tampoco incorpora directamente la gravedad diferencial de falsos positivos y falsos negativos.

Dos modelos pueden presentar la misma exactitud balanceada y, sin embargo, distribuir sus errores de forma muy diferente. Cuando el falso negativo es crítico, debe priorizarse la sensibilidad mínima, no solo el promedio.

4.10. Precisión y *recall*

En aprendizaje automático se utilizan frecuentemente las métricas *precision* y *recall*:

$$\textbf{Precision} [VP] / [VP+FP]$$

$$\textbf{Recall} [VP] / [VP+FN]$$

La *precision* coincide con el valor predictivo positivo y el *recall* con la sensibilidad.

Existe normalmente una tensión entre ambas. Reducir el umbral de decisión suele aumentar el *recall* porque el sistema emite más alertas, pero también incrementa los falsos positivos y puede disminuir la *precision*.

Aumentar el umbral produce el efecto contrario: menos alertas y mayor proporción de positivos reales entre ellas, pero mayor riesgo de omitir eventos.

No existe un punto óptimo universal. El umbral debe elegirse en función de:

- • la gravedad del peligro;
- • la prevalencia;
- • la capacidad de confirmación;
- • el coste de retención;
- • el coste del falso negativo;
- • la existencia de barreras adicionales;
- • el papel del modelo dentro del sistema.

4.11. F1-score

La puntuación F1 es la media armónica entre *precision* y *recall*:

$$F1 = 2 [Precision \times Recall] / [Precision + Recall]$$

La media armónica penaliza los casos en los que una de las dos métricas es muy baja.

El F1-score puede resultar útil para comparar modelos cuando existe desequilibrio de clases y se desea equilibrar la capacidad de detección con la fiabilidad de las alertas. Sin embargo, presenta varias limitaciones:

- • ignora los verdaderos negativos;
 - • asigna implícitamente importancia similar a *precision* y *recall*;
 - • depende de la prevalencia;
 - • no informa sobre calibración;
 - • no refleja necesariamente el coste industrial de los errores.
- En seguridad alimentaria, un modelo con mayor F1 no debe preferirse automáticamente si presenta una sensibilidad inferior al mínimo requerido.

4.12. F-beta

Cuando se desea dar mayor peso a la sensibilidad o a la *precision* puede utilizarse la familia de métricas (F_{β}):

$$F_{\beta} = (1 + \beta^2) [\text{Precision} \times \text{Recall}] / [\beta^2 \text{Precision} + \text{Recall}]$$

Cuando:

$$\beta > 1$$

se concede mayor importancia al *recall* o sensibilidad.

Cuando:

$$\beta < 1$$

se prioriza la *precision*.

En un sistema destinado a detectar lotes potencialmente peligrosos podría justificarse un (F_2), que otorga más peso a la sensibilidad. No obstante, la elección de (β) debe justificarse a partir del riesgo, no seleccionarse después de observar qué valor favorece más al modelo.

4.13. Coeficiente de correlación de Matthews

El coeficiente de correlación de Matthews integra las cuatro celdas de la matriz de confusión:

$$MCC = [VP \times VN - FP \times FN] / [\sqrt{((VP + FP)(VP + FN)(VN + FP)(VN + FN))}]$$

Su valor se sitúa entre:

$$-1 \leq MCC \leq 1$$

donde:

- • (1) indica clasificación perfecta;
- • (0) indica ausencia de asociación mejor que el azar;
- • (-1) indica clasificación completamente inversa.
- El MCC es particularmente útil en conjuntos desequilibrados porque incorpora simultáneamente verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Sin embargo, su interpretación operacional es menos intuitiva que la de sensibilidad o especificidad. Debe utilizarse como métrica complementaria, no como sustituto de las tasas de error relevantes para seguridad.

4.14. Curva ROC

La curva ROC representa la sensibilidad frente a la tasa de falsos positivos para todos los posibles umbrales de decisión.

En el eje vertical se sitúa:

$$\text{TPR} = \text{Sensibilidad}$$

En el eje horizontal:

$$\text{FPR} = 1 - \text{Especificidad}$$

El área bajo la curva ROC, AUC-ROC, expresa la capacidad de discriminación del modelo. Puede interpretarse como la probabilidad de que una observación positiva elegida al azar reciba una puntuación mayor que una observación negativa elegida al azar.

Un valor de:

$$\text{AUC} = 0,5$$

indica una capacidad similar al azar.

Un valor próximo a:

$$\text{AUC} = 1$$

indica discriminación excelente.

Sin embargo, el AUC-ROC presenta limitaciones relevantes.

Primero, resume el rendimiento sobre todos los umbrales, incluidos muchos que nunca se utilizarán industrialmente.

Segundo, no informa sobre la calibración de las probabilidades.

Tercero, puede parecer favorable en problemas con clases extremadamente desequilibradas.

Cuarto, dos modelos con AUC similar pueden presentar rendimientos muy distintos en la zona crítica de alta sensibilidad.

En una aplicación de seguridad, interesa especialmente el comportamiento del modelo en el intervalo operativo. Por ejemplo:

$$\text{Sensibilidad} \geq 0,98$$

La comparación debería centrarse en la especificidad alcanzada cuando se exige esa sensibilidad, no únicamente en el área global.

4.15. Curva *precision-recall*

La curva *precision-recall* representa la relación entre:

$$\text{Precision}$$

y

Recall

para diferentes umbrales.

Suele ser más informativa que la ROC cuando el evento positivo es raro, porque muestra directamente cuánto disminuye la fiabilidad de las alertas al intentar aumentar la sensibilidad.

La línea de referencia no es constante en 0,5, sino que depende de la prevalencia. Un clasificador sin capacidad predictiva presentará una *precision* aproximadamente igual a la proporción de positivos.

Por ello, el área bajo la curva *precision-recall* no debe compararse sin considerar la prevalencia de la población evaluada.

4.16. Selección del umbral de decisión

Muchos modelos generan una puntuación continua o una probabilidad:

$$0 \leq \hat{p} \leq 1$$

La clasificación final depende de un umbral (t):

$$\hat{y} = 1, \hat{p} \geq t \quad 0, \hat{p} < t$$

El uso de 0,5 como umbral automático carece, en muchos casos, de fundamento científico. Ese valor puede derivarse del procedimiento de entrenamiento, pero no necesariamente del riesgo industrial.

El umbral debe establecerse antes de la validación final mediante un análisis que considere:

- • sensibilidad mínima requerida;
 - • especificidad mínima viable;
 - • coste de confirmación;
 - • consecuencias del falso negativo;
 - • prevalencia esperada;
 - • disponibilidad de otras barreras;
 - • capacidad de intervención;
 - • incertidumbre de la predicción.
- **Una función de coste simplificada puede expresarse como:**

$$C(t) = C_{FN} \times FN(t) + C_{FP} \times FP(t)$$

donde:

- • (C_{FN}) es el coste atribuido a un falso negativo;
- • (C_{FP}) es el coste atribuido a un falso positivo.

Sin embargo, en seguridad alimentaria no todos los costes pueden monetizarse adecuadamente. El daño sanitario, la retirada de producto, la exposición legal y la pérdida de confianza no deben reducirse de manera simplista a un coste económico esperado.

Por ello, suele ser más prudente aplicar una restricción jerárquica:

1. • establecer una sensibilidad mínima no negociable;
2. • seleccionar, entre los umbrales que la cumplen, aquel que proporcione la mejor especificidad operativamente viable;
3. • verificar el rendimiento en el conjunto externo;
4. • comprobar que el intervalo de confianza cumple el criterio.

4.17. Intervalos de confianza

Toda métrica estimada sobre una muestra está sometida a incertidumbre.

No es suficiente informar:

Sensibilidad=98%

Debe acompañarse de un intervalo de confianza:

Sensibilidad=98% IC_{95%}: 92,9-99,8%

El intervalo indica el rango de valores compatibles con los datos y el método de estimación.

Una sensibilidad del 98 % calculada sobre 50 positivos representa:

49 \ VP

y

1 \ FN

El mismo porcentaje obtenido sobre 5.000 positivos tendría una precisión estadística mucho mayor.

Los intervalos deben calcularse mediante métodos adecuados para proporciones binomiales, como:

- • intervalo de Wilson;
- • método exacto de *Clopper-Pearson*;
- • *bootstrap*, cuando la estructura de los datos lo requiera.
- La aproximación normal puede ser inadecuada cuando el tamaño muestral es reducido o la proporción está próxima a cero o uno.

Además, cuando existen observaciones agrupadas por lote, planta o proveedor, los intervalos convencionales pueden infravalorar la incertidumbre. En esos casos deben utilizarse métodos que respeten la estructura de dependencia, como *bootstrap* por conglomerados o modelos jerárquicos.

4.18. El problema de observar cero falsos negativos

Supóngase que durante la validación externa se evalúan 50 casos positivos y el modelo identifica todos:

$$VP=50$$

$$FN=0$$

La sensibilidad observada es del 100 %.

No obstante, no puede concluirse que la tasa real de falsos negativos sea cero. La ausencia de errores puede deberse al tamaño limitado de la muestra.

La denominada regla de tres ofrece una aproximación útil: cuando no se observa ningún evento en (n) ensayos, el límite superior aproximado del intervalo de confianza del 95 % para la tasa real es:

$$[3] / [n]$$

Con 50 positivos:

$$[3] / [50]=6\%$$

Aunque no se haya observado ningún falso negativo, los datos todavía serían compatibles con una tasa real de hasta aproximadamente el 6 %.

Con 300 positivos sin falsos negativos:

$$[3] / [300]=1\%$$

Para demostrar que una tasa de error crítica es muy baja se requieren tamaños muestrales considerablemente mayores de lo que suele suponerse.

4.19. Rendimiento global y rendimiento por subgrupos

Un valor global aceptable puede ocultar fallos graves en productos minoritarios.

Un modelo puede alcanzar una sensibilidad total del 97 %, pero presentar:

- • 99 % en producto A;
- • 98 % en producto B;
- • 76 % en producto C.
- Si el producto C representa una proporción pequeña del conjunto, su mal comportamiento apenas afectará a la métrica global.
- **La validación debe analizar, siempre que sea relevante:**
 - • familia de producto;
 - • formulación;

- • planta;
- • línea;
- • proveedor;
- • campaña;
- • turno;
- • equipo;
- • intervalo de concentración;
- • temperatura;
- • pH;
- • actividad de agua;
- • tipo de envase;
- • nivel de contaminación;
- • operador;
- • condición ambiental.

Los análisis por subgrupos deben preespecificarse. La exploración retrospectiva de decenas de categorías aumenta la probabilidad de resultados espurios y puede conducir a interpretaciones oportunistas.

Cuando un subgrupo contiene pocos casos, debe reconocerse que la evidencia es insuficiente. No observar fallos en una muestra pequeña no demuestra equivalencia.

4.20. Falsos negativos próximos y lejanos al límite

No todos los errores tienen la misma magnitud.

En un problema microbiológico cuantitativo, un falso negativo puede corresponder a:

- • una muestra apenas por encima del criterio;
- • una muestra que supera ampliamente el límite;
- • una muestra con una concentración potencialmente peligrosa.
- En visión artificial, un defecto crítico muy pequeño y parcialmente oculto no es equivalente a un cuerpo extraño grande y claramente visible.

La validación debe analizar la distancia respecto al límite o la severidad del evento. Un modelo que falla principalmente en casos fronterizos presenta un perfil distinto de otro que omite positivos inequívocos.

Puede definirse una zona de incertidumbre alrededor del límite:

$$L - \delta < x < L + \delta$$

donde:

- • (L) es el límite de decisión;
- • (δ) representa la incertidumbre analítica o tecnológica.

- **Los casos situados dentro de esa zona pueden requerir una categoría adicional:**

- • positivo;
- • negativo;
- • indeterminado.

- Forzar una clasificación binaria cuando la información no permite distinguir con seguridad puede producir decisiones excesivamente confiadas.

4.21. Sistemas con opción de abstención

Un modelo industrial no debería estar obligado a emitir siempre una respuesta categórica. Puede incorporar una regla de rechazo o abstención cuando:

- • la probabilidad está próxima al umbral;
- • la incertidumbre es elevada;
- • faltan variables;
- • la calidad de la imagen es insuficiente;
- • la observación está fuera del dominio;
- • existe desacuerdo entre modelos;
- • el sensor presenta anomalías.

- **Formalmente, podría establecerse:**

$$\hat{p} < t_1 \Rightarrow \text{negativo}$$

$$t_1 \leq \hat{p} \leq t_2 \Rightarrow \text{indeterminado}$$

$$\hat{p} > t_2 \Rightarrow \text{positivo}$$

La categoría indeterminada debe activar revisión humana, repetición de la medida o análisis confirmatorio.

Esta estrategia puede reducir errores peligrosos, pero introduce una nueva métrica: la tasa de abstención.

$$\text{Tasa de abstención} = [N_{\text{indeterminados}}] / [N_{\text{total}}]$$

Un sistema que evita errores declarando indeterminada una proporción excesiva de casos puede resultar inútil. Por ello deben validarse conjuntamente:

- • sensibilidad;
- • especificidad;
- • tasa de abstención;
- • distribución de errores entre los casos no abstencionistas;
- • capacidad operativa para resolver los casos dudosos.

4.22. Matrices multiclase

Algunos modelos no clasifican simplemente conforme/no conforme, sino varias categorías:

- defecto microbiológico;
- oxidación;
- daño físico;
- error de etiquetado;
- materia extraña;
- producto correcto.

• En estos casos debe construirse una matriz de confusión multiclase y calcular las métricas para cada categoría.

• **Pueden utilizarse promedios:**

• **Macro-promedio**

• **Calcula la métrica para cada clase y obtiene la media sin ponderar:**

$$M_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K M_k$$

Concede el mismo peso a clases frecuentes y raras.

Micro-promedio

Agrega todos los verdaderos y falsos resultados antes de calcular la métrica:

$$M_{\text{micro}} = \frac{\sum_k VP_k}{\sum_k VP_k + \sum_k FN_k}$$

Está más influido por las clases frecuentes.

Promedio ponderado

Pondera cada clase por su frecuencia.

Ninguno debe sustituir el análisis individual de las clases críticas. Una categoría de peligro grave no puede quedar oculta dentro de un promedio favorable.

4.23. Modelos jerárquicos de decisión

En determinados procesos, el modelo puede operar en varias etapas:

1. • distinguir producto conforme de potencialmente no conforme;
2. • clasificar el tipo de defecto;
3. • estimar la gravedad;
4. • recomendar una acción.

• Cada etapa debe validarse separadamente y en conjunto.

Una sensibilidad del 98 % en la primera etapa y del 90 % en la segunda no genera necesariamente una sensibilidad final del 90 %. Si los errores son secuenciales, la probabilidad de clasificación correcta completa puede aproximarse a:

$$0,98 \times 0,90 = 0,882$$

es decir, un 88,2 %.

Los errores se propagan a través del sistema. La validación no debe limitarse a cada componente aislado, sino evaluar el rendimiento de extremo a extremo.

4.24. Selección de métricas según la finalidad

No existe una métrica universalmente superior. La elección debe responder al uso previsto.

Detección de peligro o lote inseguro

Métricas prioritarias:

- • sensibilidad;
- • tasa de falsos negativos;
- • límite inferior del intervalo de confianza;
- • sensibilidad por subgrupos críticos;
- • capacidad de detección próxima al límite.

• Sistema de cribado

• Métricas prioritarias:

- • sensibilidad;
- • valor predictivo negativo;
- • tasa de reducción de muestras;
- • volumen de casos derivados a confirmación;
- • tasa de abstención.

• Control de falsas alarmas

• Métricas prioritarias:

- • especificidad;
- • valor predictivo positivo;
- • número de alertas por unidad de producción;
- • coste de investigación por positivo real.

• Clasificación comercial

• Métricas prioritarias:

- • matriz de confusión;
- • precisión por categoría;
- • acuerdo ponderado;
- • coste de sobreclasificación e infra-clasificación.

- **Detección de defectos raros**

- **Métricas prioritarias:**

- • *precision-recall*;
- • sensibilidad;
- • VPP en prevalencia real;
- • falsos positivos por mil unidades;
- rendimiento por severidad.

La métrica debe expresar la decisión industrial y su riesgo. Un modelo no es aceptable porque presente un número estadístico elevado, sino porque mantiene los errores relevantes dentro de límites previamente definidos.

5. Validación de modelos de regresión y predicción de vida útil

La validación de un modelo que estima una variable continua exige una lógica diferente de la utilizada en clasificación. En estos casos, el sistema no emite únicamente una categoría —conforme o no conforme—, sino una cantidad: días de vida útil, concentración microbiana, contenido de humedad, índice de oxidación, puntuación sensorial, probabilidad de deterioro o tiempo restante hasta alcanzar un límite crítico.

La dificultad fundamental reside en que dos predicciones pueden presentar errores numéricamente similares y, sin embargo, tener consecuencias tecnológicas muy distintas.

Predecir una vida útil de 27 días cuando la duración real es de 30 días constituye un error conservador de tres días. Predecir 33 días frente a una duración real de 30 días produce el mismo error absoluto, pero puede permitir que el producto permanezca tres días más en el mercado después de haber superado el límite microbiológico, fisicoquímico o sensorial aceptable.

Por tanto, la validación no debe limitarse a cuantificar cuánto se equivoca el modelo. Debe determinar también:

- • en qué dirección se equivoca;
- • en qué condiciones se concentran los errores;
- • qué ocurre en los extremos de la distribución;
- • si la incertidumbre aumenta cerca del final de vida útil;
- • con qué frecuencia el sistema sobreestima de forma peligrosa;
- • si el rendimiento se mantiene en campañas, lotes y plantas posteriores.
- **La pregunta científica correcta no es simplemente:**
 - ¿Cuál es el error medio del modelo?
- **La pregunta relevante es:**
 - ¿Qué magnitud, dirección y distribución presentan los errores, y son compatibles con la decisión industrial para la que se utilizará el modelo?

5.1. Definición del resultado en vida útil

Antes de seleccionar métricas, debe definirse qué representa exactamente la vida útil.

La vida útil no constituye una propiedad única y universal del alimento. Es el periodo durante el cual el producto, bajo unas condiciones definidas, permanece dentro de los criterios de seguridad, calidad y aceptabilidad establecidos.

El final de vida útil puede estar determinado por:

- • superación de un criterio microbiológico;
- • crecimiento de un microorganismo alterante;
- • producción de metabolitos;
- • rancidez oxidativa;
- • pérdida de textura;
- • separación de fases;
- • cambios de color;
- • pérdida de aroma;
- • rechazo sensorial;
- • disminución de la funcionalidad;
- • incumplimiento de una especificación legal o comercial.

En muchos productos coexisten varios mecanismos de finalización. La vida útil efectiva será la correspondiente al primer criterio que se incumpla:

T_vida útil min (T_seguridad, T_microbiología, T_sensorial, T_fisicoquímica, T_funcionalidad)

Un modelo entrenado únicamente para predecir un recuento microbiológico no puede presentarse automáticamente como modelo general de vida útil si la principal causa de rechazo comercial es sensorial o fisicoquímica.

Debe definirse además si el objetivo es predecir:

- • vida útil total desde fabricación;
- • vida útil remanente desde el momento de la medición;
- • tiempo hasta un límite microbiológico;
- • fecha probable de rechazo sensorial;
- • probabilidad de incumplimiento antes de una fecha determinada;
- • intervalo temporal dentro del cual se producirá el deterioro.
- Cada formulación exige un enfoque estadístico distinto.

5.2. Error de predicción

Sea:

$$y_i$$

el valor real observado para la muestra (i), y:

$$\hat{y}_i$$

la predicción generada por el modelo.

El error individual se define como:

$$e_i = \hat{y}_i - y_i$$

Esta convención permite interpretar la dirección:

- • ($e_i > 0$): el modelo sobreestima el valor real;
- • ($e_i < 0$): el modelo lo infraestima;
- • ($e_i = 0$): predicción exacta.
- En vida útil, un error positivo suele ser más peligroso porque atribuye al producto una duración mayor de la realmente demostrada.

Sin embargo, cuando la variable predicha es una concentración microbiana, la interpretación puede invertirse. Si:

$$\hat{y}_i < y_i$$

el modelo infravalora la contaminación y puede generar una falsa impresión de seguridad.

Por ello, la dirección peligrosa del error debe definirse expresamente para cada variable.

5.3. Error medio

El error medio o *mean error* se expresa como:

$$ME = [1] / [n] \sum_{i=1}^n (\hat{y}_i - y_i)$$

Esta métrica cuantifica el sesgo promedio.

Si:

$$ME > 0$$

el modelo tiende a sobreestimar.

Si:

$$ME < 0$$

tiende a infraestimar.

Un valor próximo a cero no implica necesariamente buen rendimiento. Los errores positivos y negativos pueden cancelarse entre sí.

Por ejemplo, si un modelo presenta los siguientes errores en días:

$$+8, +7, -7, -8$$

el error medio es:

$$ME = 0$$

pero todas las predicciones se desvían de manera considerable.

El error medio debe interpretarse junto con medidas de magnitud y dispersión.

5.4. Error absoluto medio

El error absoluto medio, MAE, se calcula como:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

Expresa la desviación absoluta promedio en las mismas unidades que la variable.

Un MAE de 2,5 días significa que, en promedio, las predicciones se alejan 2,5 días del valor real, con independencia de la dirección.

Entre sus ventajas destacan:

- interpretación directa;
 - menor sensibilidad a valores extremos que el RMSE;
 - utilidad para comparar el error con la tolerancia industrial;
 - mantenimiento de las unidades originales.
- **No obstante, el MAE presenta limitaciones importantes:**
 - no distingue sobreestimación de infraestimación;
 - no revela la distribución de los errores;
 - puede ocultar una minoría de fallos extremos;
 - no informa sobre la calibración temporal;
 - asigna el mismo peso a todos los errores absolutos.
 - En vida útil, no debería aceptarse un modelo únicamente porque su MAE global sea reducido.

5.5. Error cuadrático medio y raíz del error cuadrático medio

El error cuadrático medio se define como:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Su raíz cuadrada es el RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

El RMSE mantiene las unidades originales y penaliza más intensamente los errores grandes.

Esta propiedad resulta útil cuando los fallos extremos son especialmente perjudiciales. Unas pocas sobreestimaciones graves aumentarán sustancialmente el RMSE.

Sin embargo, el RMSE también puede estar dominado por valores atípicos, errores del patrón de referencia o casos excepcionalmente difíciles. Por ello, debe analizarse junto con el MAE, los percentiles del error y la inspección de los casos extremos.

La comparación entre MAE y RMSE aporta información sobre la distribución del error. Cuando ambos valores son similares, los errores tienden a ser relativamente homogéneos. Cuando el RMSE es mucho mayor que el MAE, existen probablemente desviaciones extremas.

5.6. Error mediano y percentiles

La mediana del error absoluto es:

$$\text{MedAE} = \text{mediana} (|\hat{y}_i - y_i|)$$

Esta métrica es robusta frente a valores extremos y describe el error típico de una observación central.

No obstante, una validación de seguridad no puede concentrarse exclusivamente en el comportamiento mediano. Deben presentarse también percentiles superiores:

$$P_{90}(|e|)$$

$$P_{95}(|e|)$$

$$P_{99}(|e|)$$

El percentil 95 del error absoluto indica que el 95 % de las predicciones presenta un error igual o inferior a ese valor.

Por ejemplo:

$$\text{MAE} = 2,1 \text{ días}$$

$$P_{95} = 7,8 \text{ días}$$

Esta combinación revela que el comportamiento promedio puede ser aceptable, pero existe una cola de errores considerable.

En decisiones críticas, el criterio de aceptación puede formularse sobre la cola de la distribución:

- Al menos el 95 % de las predicciones deberá presentar un error absoluto inferior a cuatro días.
- **Y, adicionalmente:**
 - Ninguna sobreestimación podrá superar siete días sin activar una revisión específica.

5.7. Error porcentual

El error porcentual absoluto se expresa como:

$$\text{APE}_i = |[\hat{y}_i - y_i] / [y_i]| \times 100$$

El promedio de estos errores genera el MAPE:

$$\text{MAPE} = [100] / [n] \sum_{i=1}^n |[\hat{y}_i - y_i] / [y_i]|$$

El MAPE facilita la comparación entre productos con escalas distintas. Un error de tres días no tiene la misma importancia relativa en una vida útil de diez días que en otra de doce meses.

Sin embargo, presenta graves limitaciones cuando los valores reales son próximos a cero. También penaliza de forma asimétrica determinadas desviaciones y puede resultar difícil de interpretar en variables logarítmicas o censuradas.

En microbiología, donde los recuentos pueden expresarse como log UFC/g, no siempre es adecuado calcular porcentajes sobre la escala logarítmica. Debe indicarse con claridad si la evaluación se realiza:

- en escala original;
- en escala logarítmica;

- sobre incrementos respecto al valor inicial;
- sobre tiempo hasta alcanzar un límite.

5.8. Coeficiente de determinación

El coeficiente (R^2) se utiliza con frecuencia para resumir el ajuste:

$$R^2 = 1 - [\sum_{i=1}^n (y_i - \hat{y}_i)^2] / [\sum_{i=1}^n (y_i - \bar{y})^2]$$

Indica qué proporción de la variabilidad observada explica el modelo respecto a una predicción basada en la media.

Un (R^2) elevado no implica necesariamente que las predicciones sean suficientemente exactas.

Un modelo puede presentar:

$$R^2 = 0,95$$

y, aun así, cometer errores de varios días que resulten inaceptables. Esto puede ocurrir cuando el rango total de la variable es muy amplio.

Además, el (R^2):

- no informa sobre el sesgo;
 - no mide concordancia;
 - no revela errores extremos;
 - depende del rango de valores;
 - puede ser negativo en validación externa;
 - no indica si la pendiente y el intercepto son adecuados.
- Dos métodos pueden estar fuertemente correlacionados y, sin embargo, discrepar sistemáticamente.
- **Por ejemplo:**

$$\hat{y} = y + 5$$

produciría una correlación perfecta, aunque todas las predicciones sobreestimasen la vida útil en cinco días.

El (R^2) debe considerarse una medida complementaria, nunca el criterio principal de aceptación.

5.9. Correlación no equivale a concordancia

La correlación mide si dos variables cambian de forma relacionada. La concordancia evalúa si sus valores son suficientemente próximos.

Un modelo puede ordenar correctamente los productos de menor a mayor vida útil, pero sobreestimar todos los valores. En ese caso, la correlación será elevada, aunque la utilidad para asignar fechas sea deficiente.

La validación debe analizar:

- intercepto de la relación predicción-observación;
 - pendiente;
 - sesgo medio;
 - dispersión de los residuos;
 - concordancia individual.
- **Idealmente, la relación debería aproximarse a:**

$$\hat{y} = y$$

con:

$$\text{intercepto} = 0$$

y:

pendiente=1

Una pendiente inferior a uno puede indicar regresión excesiva hacia la media: el modelo sobreestima los valores bajos e infraestima los altos.

5.10. Gráficos de residuos

Los residuos deben representarse frente a:

- • valor real;
- • valor predicho;
- • tiempo;
- • producto;
- • planta;
- • temperatura;
- • pH;
- • actividad de agua;
- • lote;
- • equipo.
- Un modelo adecuadamente especificado debería mostrar residuos distribuidos sin patrones evidentes alrededor de cero.
- **La aparición de estructuras puede revelar:**
 - • no linealidad;
 - • heterocedasticidad;
 - • variables omitidas;
 - • deriva temporal;
 - • diferencias entre productos;
 - • saturación del sensor;
 - • errores dependientes de la concentración;
 - • funcionamiento deficiente en extremos.

Si la dispersión aumenta con la vida útil, el modelo no presenta una incertidumbre constante. En tal caso, un único MAE global resulta insuficiente.

5.11. Heterocedasticidad

Existe heterocedasticidad cuando la variabilidad del error cambia a lo largo del rango de predicción.

Por ejemplo, el modelo puede estimar con precisión vidas útiles de 10 a 20 días, pero presentar errores crecientes por encima de 45 días.

Esta situación puede deberse a:

- • menor número de observaciones en los extremos;
- • aumento de la variabilidad biológica;
- • mecanismos de deterioro diferentes;
- • acumulación de incertidumbre temporal;
- • cambios de régimen microbiológico;
- • extrapolación.

- La validación debe cuantificar el error por intervalos:

Vida útil real	MAE	RMSE	Sesgo	Sobreestimaciones críticas
0–15 días	—	—	—	—
16–30 días	—	—	—	—
31–45 días	—	—	—	—
>45 días	—	—	—	—

- No debe asumirse que el rendimiento global es aplicable uniformemente a todo el dominio.

5.12. Límites de acuerdo

El análisis de *Bland–Altman* permite evaluar la concordancia entre predicción y referencia.

Para cada observación se calcula:

$$d_i = \hat{y}_i - y_i$$

y:

$$m_i = [\hat{y}_i + y_i] / [2]$$

El sesgo medio es:

$$\bar{d}$$

Los límites de acuerdo aproximados del 95 % son:

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

donde (s_d) es la desviación estándar de las diferencias.

Estos límites indican el intervalo en el que se espera que se sitúe la mayoría de las discrepancias individuales.

La aceptación no depende de que los límites sean estrechos en términos abstractos, sino de que permanezcan dentro de una tolerancia tecnológicamente justificada.

Si una diferencia de ± 8 días es incompatible con la gestión comercial o sanitaria, el modelo no será apto aunque el gráfico muestre un patrón estadísticamente estable.

El análisis debe comprobar además si:

- el sesgo cambia con la magnitud;
- existen valores extremos;
- la variabilidad aumenta con la media;
- la distribución de diferencias es aproximadamente simétrica;
- los datos están agrupados por lote o producto.

5.13. Sobreestimación e infraestimación

En vida útil, conviene separar dos componentes:

$$MAE_{sobre} [1] / [n_{sobre}] \sum_{e_i > 0} e_i$$

$$MAE_{infra} [1] / [n_{infra}] \sum_{e_i < 0} |e_i|$$

También deben calcularse las tasas:

$$P(e_i > 0)$$

y:

$$P(e_i < 0)$$

Una distribución simétrica puede no ser deseable. Desde el punto de vista de seguridad, puede preferirse un modelo moderadamente conservador, con cierta tendencia a infraestimar la vida útil.

Sin embargo, un conservadurismo excesivo también puede resultar perjudicial:

- • incrementa desperdicio alimentario;
- • reduce competitividad;
- • acorta innecesariamente la distribución;
- • distorsiona inventarios;
- • genera decisiones inconsistentes;
- • puede ocultar un modelo mal calibrado.
- La finalidad no es obtener el modelo más pesimista, sino uno suficientemente exacto y con un margen de seguridad explícito.

5.14. Error peligroso

Puede definirse un error peligroso de vida útil como:

$$e_i > \delta_{\text{crítico}}$$

donde ($\delta_{\text{crítico}}$) representa la máxima sobreestimación tolerable.

Por ejemplo:

$$\delta_{\text{crítico}} = 2 \text{ días}$$

La tasa de sobreestimación peligrosa sería:

$$R_{\text{peligroso}} = [\sum I(e_i > \delta_{\text{crítico}})] / [n]$$

donde (I) es una función indicadora.

Esta métrica puede ser más importante que el MAE global.

Un criterio de aceptación razonable podría establecer:

$$R_{\text{peligroso}} < 5\%$$

y:

$$IC_{95\%, \text{superior}}(R_{\text{peligroso}}) < 8\%$$

En aplicaciones de seguridad, los límites deberían ser más exigentes y basarse en la severidad del peligro y en las medidas de control adicionales.

5.15. Márgenes de seguridad

La predicción del modelo no tiene por qué coincidir directamente con la fecha asignada al producto.

Puede aplicarse un margen de seguridad:

$$T_{\text{declarado}} \hat{T}_{\text{vida útil}} - M$$

donde:

- ($\hat{T}_{\text{vida útil}}$) es la estimación;
- (M) es el margen de seguridad.

Este margen debe derivarse de la incertidumbre y del riesgo, no seleccionarse arbitrariamente.

Puede depender de:

- • límite inferior del intervalo predictivo;

- • variabilidad entre lotes;
 - • severidad del peligro;
 - • desviaciones previsibles de temperatura;
 - • incertidumbre del método de referencia;
 - • condiciones de distribución;
 - • población vulnerable;
 - • capacidad de crecimiento microbiano.
- **Una estrategia conservadora es utilizar un percentil inferior:**

$$T_{\text{declarado } Q_{\alpha}}(T|X)$$

donde (Q_{α}) representa un cuantil bajo de la distribución predictiva, como el percentil 5.

En lugar de utilizar la estimación media, se selecciona una duración que el modelo considera superada únicamente con una probabilidad limitada.

5.16. Intervalos de predicción

Una estimación puntual de 30 días no expresa la incertidumbre asociada.

Un modelo debería poder proporcionar:

$$\hat{T}=30 \text{ días}$$

junto con un intervalo de predicción:

$$IP_{95\%}=24-36 \text{ días}$$

El intervalo de predicción estima el rango probable para una observación individual futura. Es más amplio que un intervalo de confianza sobre la media.

La amplitud del intervalo puede variar entre muestras. Un producto situado en una región muy representada del espacio de entrenamiento puede generar un intervalo estrecho. Una formulación atípica o próxima al límite del dominio debería producir mayor incertidumbre.

La validación de intervalos requiere evaluar su cobertura:

$$\text{Cobertura observada } [\sum I(y_i \in [L_i, U_i])] / [n]$$

Si el modelo declara intervalos del 95 %, aproximadamente el 95 % de los valores reales debería quedar dentro de ellos.

No basta con que la cobertura sea correcta. Los intervalos no deben ser tan amplios que resulten inútiles. Debe analizarse simultáneamente:

- • cobertura;
- • amplitud media;
- • amplitud por subgrupos;
- • cobertura en extremos;
- • cobertura temporal.

5.17. Calibración de predicciones continuas

La calibración evalúa si las predicciones corresponden sistemáticamente a los valores observados.

Puede analizarse mediante una regresión:

$$y_i = \alpha + \beta \hat{y}_i + \varepsilon_i$$

Una calibración ideal presenta:

$$\alpha=0$$

$$\beta=1$$

Si:

$$\alpha \neq 0$$

existe un desplazamiento sistemático.

Si:

$$\beta < 1$$

las predicciones pueden ser demasiado extremas o presentar una relación deformada, dependiendo de la especificación.

La calibración debe evaluarse en datos externos. Recalibrar el modelo sobre el conjunto de prueba y comunicar después el rendimiento en ese mismo conjunto invalida la independencia de la evaluación.

5.18. Predicción microbiológica

Los modelos microbiológicos pueden estimar:

- • tasa máxima de crecimiento;
- • fase de latencia;
- • tiempo hasta alcanzar un nivel;
- • concentración en un momento determinado;
- • supervivencia;
- • inactivación;
- • probabilidad de crecimiento;
- • incremento logarítmico durante la vida útil.
- Las métricas deben adecuarse a la variable.
- **Si se predice un recuento:**

$$\log_{10} \text{UFC/g}$$

un error de:

$$1 \setminus \log_{10}$$

equivale a una diferencia de un factor diez en la escala original.

Por ello, un MAE de 0,5 log UFC/g no debe interpretarse como un error del 0,5 %, sino como una discrepancia multiplicativa relevante.

En microbiología predictiva se utilizan también el factor de sesgo y el factor de exactitud.

5.19. Factor de sesgo

El factor de sesgo puede expresarse como:

$$B_f = 10^{[1] / [n] \sum \log_{10} ([\text{predicho}] / [\text{observado}])}$$

Su interpretación depende de la variable modelizada.

Un valor:

$$B_f = 1$$

indica ausencia de sesgo multiplicativo.

Valores superiores o inferiores a uno reflejan sobre-predicción o infra-predicción.

Cuando se predice tiempo de crecimiento o vida útil, la dirección conservadora debe interpretarse cuidadosamente. Sobrepredecir el tiempo necesario para alcanzar un nivel peligroso puede ser no conservador, mientras que sobrepredecir el crecimiento microbiano puede ser conservador.

No basta, por tanto, con afirmar que el factor de sesgo es mayor o menor que uno. Debe explicarse qué significa esa dirección para la seguridad.

5.20. Factor de exactitud

El factor de exactitud se define de forma similar, utilizando el valor absoluto de los logaritmos:

$$A_f = 10^{\frac{1}{n} \sum |\log_{10} ([predicho] / [observado])|}$$

Siempre cumple:

$$A_f \geq 1$$

Cuanto más próximo a uno, mayor concordancia.

Un factor de exactitud de 1,20 indica, de forma aproximada, una desviación multiplicativa media del 20 %, aunque su interpretación exacta depende de la formulación.

Estos factores permiten resumir sesgo y dispersión, pero no sustituyen el análisis de residuos, los intervalos ni la evaluación por condiciones ambientales y matrices.

5.21. Datos censurados

En estudios de vida útil es frecuente que no se observe el momento exacto de deterioro.

Si el estudio termina en el día 60 y el producto sigue siendo aceptable, solo puede afirmarse:

$$T > 60$$

El resultado está censurado a la derecha.

Si el producto ya está deteriorado en la primera evaluación, puede existir censura a la izquierda:

$$T < 5$$

Cuando se conoce que el evento ocurrió entre dos visitas, existe censura por intervalo:

$$20 < T \leq 25$$

Asignar arbitrariamente el último día del estudio como vida útil real introduce sesgo. Los modelos deben utilizar métodos que traten correctamente la censura, como:

- • análisis de supervivencia;
- • modelos de riesgos proporcionales;
- • modelos paramétricos de tiempo hasta evento;
- • bosques de supervivencia;
- • redes neuronales de supervivencia;
- • estimación por intervalos.

La métrica también debe adaptarse. El MAE convencional no puede calcularse correctamente si el tiempo real no se conoce.

5.22. Análisis de supervivencia

El análisis de supervivencia modeliza el tiempo hasta un evento.

La función de supervivencia es:

$$S(t)=P(T>t)$$

y representa la probabilidad de que el producto permanezca dentro del criterio más allá del tiempo (t).

Un modelo puede estimar, por ejemplo:

$$S(30)=0,92$$

lo que indicaría una probabilidad estimada del 92 % de seguir siendo aceptable después de 30 días, bajo las condiciones especificadas.

La validación puede incluir:

- • índice de concordancia;
- • puntuación de Brier dependiente del tiempo;
- • calibración temporal;
- • curvas de supervivencia predichas y observadas;
- • error integrado de predicción;
- • rendimiento en horizontes específicos.
- La utilidad práctica puede ser mayor que la de una fecha puntual, ya que permite decidir qué nivel de riesgo resulta aceptable.

5.23. Índice de concordancia

El índice de concordancia evalúa si el modelo ordena correctamente los tiempos de evento.

De forma simplificada, mide la probabilidad de que, entre dos productos comparables, aquel al que el modelo asigna mayor vida útil sea realmente el que tarda más en alcanzar el evento.

Un valor de:

$$0,5$$

equivale aproximadamente al azar.

Un valor de:

$$1$$

indica ordenación perfecta.

No obstante, el índice de concordancia mide discriminación temporal, no calibración. Un modelo puede ordenar correctamente los productos y asignar fechas sistemáticamente equivocadas.

5.24. Puntuación de Brier

Para un horizonte temporal (t), la puntuación de Brier evalúa la diferencia cuadrática entre la probabilidad predicha y el resultado observado:

$$BS(t)= [1] / [n] \sum_{i=1}^n (I(T_i>t)-\hat{S}_i(t))^2$$

Valores inferiores indican mejor rendimiento.

La puntuación integra discriminación y calibración, pero debe ajustarse adecuadamente en presencia de censura.

Puede calcularse para diferentes horizontes:

- • 15 días;
- • 30 días;
- • 45 días;
- • fecha de caducidad propuesta.
- Esto permite identificar si el modelo funciona bien a corto plazo pero pierde precisión cuando aumenta el horizonte.

5.25. Modelos de horizonte fijo

En ocasiones no interesa conocer el día exacto de deterioro, sino responder a una pregunta concreta:

- ¿Cumplirá el producto el criterio al día 30?

- **El problema puede reformularse como clasificación binaria:**

$$Y = 1, T \geq 30 \setminus 0, T < 30$$

En ese caso deberán evaluarse:

- • sensibilidad para detectar vidas útiles insuficientes;
- • especificidad;
- • valores predictivos;
- • calibración de la probabilidad;
- • errores próximos al horizonte.

Esta estrategia simplifica la decisión, pero pierde información temporal. Un producto que falla en el día 29 y otro que falla en el día 5 quedarían en la misma categoría.

5.26. Validación temporal

La predicción de vida útil está especialmente expuesta a cambios temporales:

- • estaciones;
- • temperatura ambiental;
- • campañas agrícolas;
- • variación de materias primas;
- • cambios de proveedor;
- • envejecimiento de equipos;
- • modificaciones de formulación;
- • flora residente;
- • prácticas de limpieza;
- • condiciones logísticas.
- Una partición aleatoria puede mezclar observaciones de todos los periodos y producir una estimación optimista.

- **La evaluación temporal debe respetar el orden real:**

$$\text{entrenamiento} < \text{validación} < \text{prueba}$$

Por ejemplo:

- • entrenamiento: campañas 2023–2024;
- • ajuste: primer semestre de 2025;
- • validación externa: segundo semestre de 2025;
- • evaluación prospectiva: producción de 2026.

- Nunca debe entrenarse con datos futuros para predecir el pasado cuando la aplicación real será prospectiva.

5.27. Validación por lotes y curvas longitudinales

Cuando se realizan mediciones repetidas del mismo lote a lo largo del tiempo, las observaciones están correlacionadas.

No debe dividirse una curva de vida útil de forma que algunos días entrenen el modelo y otros días del mismo lote se utilicen como prueba. El modelo podría reconocer el perfil particular del lote y sobreestimar su capacidad para generalizar.

La unidad de partición debe ser la curva completa o el lote completo.

Si existen diez mediciones de cien lotes, el tamaño independiente se aproxima a cien lotes, no a mil puntos temporales.

Los intervalos de confianza y los modelos deben considerar la estructura longitudinal mediante:

- efectos aleatorios;
- modelos mixtos;
- *bootstrap* por lote;
- validación cruzada agrupada;
- modelos de series temporales jerárquicas.

5.28. Extrapolación

Un modelo entrenado en temperaturas entre 2 y 8 °C no puede asumirse válido a 12 °C.

Del mismo modo, un modelo desarrollado con:

$$4,5 \leq \text{pH} \leq 5,5$$

no debe aplicarse sin evaluación a un producto con:

$$\text{pH}=6,2$$

Los algoritmos pueden generar predicciones numéricas fuera de su dominio sin advertir que están extrapolando.

La validación debe establecer límites explícitos para cada variable crítica y desarrollar mecanismos de detección de condiciones fuera de rango.

Cuando una observación queda fuera del dominio, el sistema debe:

- abstenerse;
- emitir una advertencia;
- exigir revisión;
- aplicar un modelo alternativo validado;
- impedir la decisión automática.

5.29. Validación en condiciones razonablemente adversas

La evaluación no debe limitarse a las condiciones ideales de almacenamiento.

Debe incluir escenarios razonablemente previsibles, como:

- fluctuaciones térmicas;
- retrasos logísticos;
- apertura parcial del envase;
- oxígeno residual elevado;

- desviación del pH;
- carga inicial alta;
- variación del conservante;
- abuso térmico moderado;
- heterogeneidad entre unidades.
- Un modelo que únicamente funciona a temperatura constante de laboratorio puede no ser adecuado para predecir la vida útil en la cadena de suministro real.

No obstante, la validación tampoco debe extenderse a escenarios físicamente irrelevantes. Las condiciones adversas deben derivarse de datos logísticos, históricos de proceso y análisis de riesgo.

5.30. Comparación con modelos de referencia

La IA debe compararse con alternativas razonables:

- regla fija de caducidad;
- regresión tradicional;
- modelo microbiológico establecido;
- ecuaciones secundarias de temperatura;
- criterio de experto;
- modelo basado en peor caso;
- procedimiento histórico de la empresa.
- La comparación debe realizarse sobre el mismo conjunto externo.

Un modelo de inteligencia artificial no queda justificado porque sea complejo o novedoso. Debe demostrar una mejora relevante en:

- exactitud;
- seguridad;
- capacidad de adaptación;
- anticipación;
- reducción de desperdicio;
- estabilidad;
- utilidad operativa.
- Si un modelo más sencillo ofrece prestaciones equivalentes y mayor interpretabilidad, puede ser la opción científicamente preferible.

5.31. Criterios de aceptación para vida útil

Los criterios deben combinar medidas promedio, sesgo, errores extremos y seguridad.

Un protocolo podría establecer:

$$MAE \leq 3 \text{ días}$$

$$|ME| \leq 1 \text{ día}$$

$$P_{95}(|e|) \leq 6 \text{ días}$$

$$P(e > 2 \text{ días}) \leq 5\%$$

$$\text{Cobertura del IP}_{95\%} \geq 92\%$$

Rendimiento mínimo por familia de producto

La aprobación debería exigir que se cumplan simultáneamente los criterios prioritarios. No resulta correcto compensar una tasa elevada de sobreestimaciones peligrosas con un buen (R^2).

En un contexto de seguridad microbiológica, los límites deben ser más restrictivos y complementarse con *challenge tests*, estudios de durabilidad, validación del proceso y revisión experta.

5.32. Vida útil dinámica

Los sistemas más avanzados pueden actualizar la vida útil remanente a medida que reciben nueva información:

$$\hat{T}_{\text{remanente}}, t f(X_0, X_1, \dots, X_t)$$

Este enfoque permite incorporar:

- • historial de temperatura;
- • cambios de humedad;
- • lecturas de sensores;
- • evolución de gases;
- • resultados analíticos;
- • señales de deterioro.

La validación debe evaluar cada momento de actualización. No basta con demostrar buen rendimiento en el día de fabricación.

Debe analizarse:

- • error al inicio;
- • error a mitad de vida útil;
- • error próximo al final;
- • estabilidad de las revisiones;
- • frecuencia de cambios bruscos;
- • capacidad para responder a abusos térmicos;
- • consistencia entre actualizaciones.

Una predicción dinámica no debería aumentar inexplicablemente la vida útil remanente después de una desviación que objetivamente empeora el producto, salvo que exista una justificación física demostrable.

5.33. La predicción no sustituye la evidencia experimental

Un modelo puede reducir ensayos, priorizar muestras, integrar información y mejorar la toma de decisiones. Pero no debe utilizarse para eludir la validación experimental exigible.

En particular, la predicción de vida útil no sustituye automáticamente:

- • estudios de durabilidad;
- • ensayos de desafío microbiológico;
- • validación de tratamientos;
- • controles de estabilidad;
- • evaluación sensorial;
- • cumplimiento de criterios microbiológicos;
- • verificación de condiciones de distribución.
- La IA puede formar parte de la evidencia, pero su peso debe corresponder a la calidad de la validación y al riesgo de la decisión.

- **La afirmación científicamente correcta no es:**

«La IA ha determinado que el producto dura 60 días.»

Sino:

«El modelo, validado para esta categoría de producto y estas condiciones de proceso, estima una vida útil de 60 días con un intervalo predictivo determinado. Esta estimación se integra con los resultados microbiológicos, fisicoquímicos, sensoriales y de validación experimental establecidos en el protocolo.»

La diferencia separa una inferencia algorítmica de una conclusión tecnológicamente defendible.

6. Calibración, incertidumbre y robustez: validar que el modelo conoce los límites de su propia predicción

La capacidad de discriminación no es suficiente para justificar el uso industrial de un modelo de inteligencia artificial. Un sistema puede ordenar correctamente los lotes desde los menos hasta los más problemáticos y, al mismo tiempo, asignar probabilidades profundamente equivocadas.

Supóngase que un modelo identifica correctamente que el lote A presenta mayor riesgo que el lote B. Esta capacidad discriminativa puede resultar útil. Sin embargo, si asigna al lote A una probabilidad de contaminación del 90 % cuando la frecuencia real en casos equivalentes es del 25 %, sus resultados no están calibrados. Del mismo modo, si declara un riesgo del 5 % para un conjunto de lotes en el que el evento ocurre realmente en el 20 %, el sistema produce una falsa sensación de seguridad.

La calibración responde a una pregunta distinta de la sensibilidad, la especificidad o el área bajo la curva ROC:

- Entre las observaciones a las que el modelo asigna una determinada probabilidad, ¿con qué frecuencia ocurre realmente el evento?
- **Un modelo bien calibrado debería cumplir, de forma aproximada, que:**

$$P(Y=1|\hat{p}=p)=p$$

Esto significa que, entre los casos a los que se asigna una probabilidad del 20 %, el evento debería observarse aproximadamente en el 20 %; entre aquellos con una probabilidad del 80 %, debería producirse alrededor del 80 %.

La calibración es esencial cuando la salida del modelo se utiliza para:

- • priorizar controles;
- • asignar niveles de riesgo;
- • decidir la intensidad del muestreo;
- • estimar la probabilidad de incumplimiento;
- • comparar escenarios;
- • calcular costes esperados;
- • activar medidas de retención o confirmación;
- • establecer una vida útil dinámica.
- Una probabilidad no calibrada puede seguir siendo útil como puntuación relativa, pero no debe interpretarse literalmente como riesgo.

6.1. Discriminación y calibración son propiedades diferentes

La discriminación mide la capacidad del modelo para separar observaciones con y sin evento.

La calibración mide la correspondencia entre las probabilidades predichas y las frecuencias observadas.

Un modelo puede presentar:

- • discriminación elevada y calibración deficiente;
- • discriminación moderada y calibración razonable;
- • buena calibración global y fallos importantes en subgrupos;
- • buen rendimiento inicial y pérdida de calibración con el tiempo.

La curva ROC y su área evalúan discriminación. No demuestran calibración.

El problema puede ilustrarse mediante una transformación extrema. Si un modelo asigna correctamente puntuaciones de mayor a menor riesgo, pero eleva todas sus probabilidades al cuadrado, puede conservar prácticamente

el mismo orden y, por tanto, la misma AUC. Sin embargo, las probabilidades resultantes serán sistemáticamente inferiores y dejarán de representar la frecuencia real del evento.

Por ello, un informe que comunica únicamente:

$$AUC=0,94$$

no permite concluir que un lote con una salida de 0,80 posea un 80 % de probabilidad real de no conformidad.

6.2. Curva de calibración

La curva de calibración compara las probabilidades predichas con las frecuencias observadas.

Un procedimiento habitual consiste en agrupar las observaciones por intervalos de probabilidad:

- • 0–10 %;
- • 10–20 %;
- • 20–30 %;
- • etc.

- **Para cada grupo se calcula:**

$$\bar{p}_k = [1] / [n_k] \sum_{i \in k} \hat{p}_i$$

y:

$$\bar{y}_k = [1] / [n_k] \sum_{i \in k} y_i$$

La curva representa:

$$\bar{p}_k \text{ frente a } \bar{y}_k$$

La calibración ideal sigue la diagonal:

$$\bar{p}_k = \bar{y}_k$$

Si la frecuencia observada supera sistemáticamente la probabilidad predicha, el modelo infravalora el riesgo.

Si la probabilidad predicha es superior a la frecuencia observada, el sistema sobreestima el riesgo.

Sin embargo, el agrupamiento en intervalos introduce decisiones metodológicas. El resultado puede variar dependiendo de:

- • número de grupos;
- • tamaño de cada grupo;
- • distribución de las probabilidades;
- • escasez de casos positivos;
- • presencia de subgrupos heterogéneos.
- Por ello, la curva debe complementarse con medidas continuas y con intervalos de confianza.

6.3. Intercepto y pendiente de calibración

La calibración puede evaluarse mediante una regresión entre el resultado observado y el logit de la probabilidad predicha:

$$\text{logit}(P(Y=1)) \alpha + \beta \text{logit}(\hat{p})$$

donde:

$$\text{logit}(p) \ln ([p] / [1-p])$$

En un modelo ideal:

$$\alpha=0$$

y:

$$\beta=1$$

El intercepto de calibración indica si el modelo sobreestima o infravalora globalmente el riesgo.

Cuando:

$$\alpha>0$$

la frecuencia observada tiende a ser mayor que la predicha, por lo que el modelo infravalora el riesgo.

Cuando:

$$\alpha<0$$

el modelo tiende a sobreestimarlos.

La pendiente informa sobre la extremidad de las predicciones.

Una pendiente:

$$\beta<1$$

suele indicar que las probabilidades son excesivamente extremas: las bajas son demasiado bajas y las altas demasiado altas. Este comportamiento es frecuente en modelos sobreajustados.

Una pendiente:

$$\beta>1$$

puede indicar que las predicciones varían menos de lo necesario y se concentran excesivamente alrededor de la media.

Estos parámetros deben estimarse en datos externos. Evaluarlos en el conjunto utilizado para ajustar el modelo ofrece una imagen artificialmente favorable.

6.4. Puntuación de *Brier*

La puntuación de *Brier* evalúa el error cuadrático de las probabilidades:

$$BS = [1] / [n] \sum_{i=1}^n (\hat{p}_i - y_i)^2$$

donde:

- $(\hat{p}_i)$ es la probabilidad predicha;
- (y_i) toma el valor 0 o 1.

El resultado se sitúa normalmente entre 0 y 1. Un valor inferior indica mejor rendimiento.

Una predicción perfecta presenta:

$$BS=0$$

La puntuación penaliza tanto probabilidades mal calibradas como discriminación deficiente. Sin embargo, depende de la prevalencia del evento. No debe compararse directamente entre poblaciones con prevalencias muy diferentes sin utilizar referencias o descomposiciones adecuadas.

Puede calcularse una puntuación de habilidad:

$$BSS = 1 - [BS_{\text{modelo}}] / [BS_{\text{referencia}}]$$

donde el modelo de referencia podría asignar a todos los casos la prevalencia observada.

Un valor positivo indica mejora respecto a la referencia; un valor negativo significa que el modelo probabilístico funciona peor que una predicción constante basada en la frecuencia media.

6.5. Error esperado de calibración

El *expected calibration error*, ECE, resume la discrepancia entre probabilidad predicha y frecuencia observada:

$$ECE = \sum_{k=1}^K [n_k / n] | \hat{p}_k - \bar{y}_k |$$

Cuanto menor sea el valor, mejor será la calibración aparente.

No obstante, el ECE depende intensamente del número y diseño de los intervalos. Dos modelos pueden intercambiar su orden relativo si se modifica el agrupamiento.

No debe utilizarse como única evidencia de calibración. Resulta preferible acompañarlo de:

- • curva de calibración;
- • intercepto;
- • pendiente;
- • puntuación de *Brier*;
- • análisis por subgrupos;
- • intervalos de confianza.

6.6. Recalibración

Un modelo con buena discriminación pero probabilidades sesgadas puede recalibrarse.

Entre los métodos habituales se encuentran:

- • corrección del intercepto;
- • recalibración logística;
- • *Platt scaling*;
- • regresión isotónica;
- • *beta calibration*;
- • métodos no paramétricos.

La recalibración debe realizarse con datos independientes del conjunto de prueba final. Si se utiliza el mismo conjunto para recalibrar y evaluar, el rendimiento resultante será optimista.

Una estructura correcta sería:

1. • desarrollar el modelo;
2. • calibrarlo en un conjunto específico;
3. • congelar la versión;
4. • evaluarlo en una muestra externa independiente.

La recalibración local puede ser necesaria cuando cambia la prevalencia o la población, pero no corrige automáticamente todos los problemas. Si las relaciones entre variables y resultado han cambiado, puede ser necesario reentrenar el modelo.

6.7. Cambio de prevalencia y probabilidad posterior

La prevalencia de un evento influye de manera decisiva en las probabilidades posteriores y en los valores predictivos.

Un modelo desarrollado en una base enriquecida con un 30 % de lotes no conformes no puede desplegarse directamente en una planta donde la prevalencia real sea del 0,5 % sin analizar la calibración.

El enriquecimiento es útil para entrenar y evaluar la sensibilidad, pero altera la distribución de clases.

Si el algoritmo interpreta la frecuencia artificial de la muestra como frecuencia real, sus probabilidades estarán infladas. Puede seguir ordenando correctamente los casos, pero no reflejar el riesgo operativo.

La corrección requiere:

- • conocer la prevalencia real;
- • documentar el método de muestreo;
- • ajustar probabilidades cuando proceda;
- • validar nuevamente en datos representativos;
- • monitorizar cambios posteriores.

La prevalencia tampoco debe tratarse como una constante inmutable. Puede variar por:

- • campaña;
- • estación;
- • proveedor;
- • línea;
- • producto;
- • cambios de limpieza;
- • intervención correctiva;
- • brote o incidente;
- • modificación de criterios de muestreo.
- Un sistema calibrado para una situación histórica puede perder validez después de una mejora significativa del proceso o de la aparición de un nuevo foco de contaminación.

6.8. Incertidumbre aleatoria y epistémica

No toda incertidumbre tiene el mismo origen.

Incertidumbre aleatoria

La incertidumbre aleatoria, o aleatoria irreducible, procede de la variabilidad inherente al fenómeno.

En alimentación puede derivarse de:

- • heterogeneidad microbiológica;
- • variación entre unidades;
- • distribución irregular de contaminantes;
- • diferencias biológicas;
- • fluctuaciones sensoriales;
- • variabilidad de proceso no controlable;
- • error inherente del método analítico.
- Aunque se recopilen más datos, parte de esta incertidumbre puede permanecer.

Incertidumbre epistémica

La incertidumbre epistémica procede de un conocimiento insuficiente del sistema.

Puede deberse a:

- • escasez de datos;
- • falta de representatividad;
- • ausencia de determinados productos;
- • regiones no cubiertas del espacio de variables;
- • arquitectura inadecuada;
- • parámetros mal estimados;

- cambio de dominio;
- variables relevantes no observadas.
- Esta incertidumbre puede reducirse mediante nuevos datos, mejor diseño experimental, modificación del modelo o restricción del dominio.

La distinción es importante. Un modelo no debería presentar la misma confianza en una formulación bien representada que en otra situada en una región apenas observada.

6.9. Incertidumbre de medida

Las variables de entrada y el patrón de referencia también contienen incertidumbre.

Por ejemplo:

$$\text{pH}=5,20\pm0,05$$

$$a_w=0,945\pm0,003$$

$$N=2,0\pm0,4\ \log_{10}\text{UFC/g}$$

Si pequeñas variaciones dentro de esos intervalos modifican sustancialmente la predicción, el sistema es inestable frente a la incertidumbre analítica ordinaria.

La validación debe propagar la incertidumbre de las entradas. Puede realizarse mediante:

- simulación de Monte Carlo;
 - análisis de sensibilidad;
 - perturbaciones controladas;
 - modelos probabilísticos;
 - repetición de medidas;
 - intervalos sobre las entradas.
- Si:

$$X_{jsimmathcal{D}_j}$$

representa la distribución de una variable medida, puede generarse un conjunto de valores plausibles y obtener la distribución de predicciones:

$$\hat{Y}^*(1), \hat{Y}^*(2), \ldots, \hat{Y}^*(m)$$

La dispersión resultante informa de cuánto depende la salida del error analítico.

6.10. Robustez frente a perturbaciones

La robustez expresa la capacidad del modelo para conservar un rendimiento suficiente cuando las condiciones varían dentro de márgenes previsibles.

No significa que el sistema deba funcionar bajo cualquier modificación. Significa que pequeñas variaciones no deberían provocar cambios desproporcionados o incoherentes.

En una aplicación visual, deben evaluarse alteraciones como:

- luminosidad;
- contraste;
- enfoque;
- rotación;
- distancia;
- velocidad;

- • condensación;
- • suciedad;
- • reflejos;
- • oclusión parcial.
- **En espectroscopia:**
 - • deriva del instrumento;
 - • temperatura;
 - • humedad;
 - • granulometría;
 - • espesor de muestra;
 - • posición;
 - • ruido;
 - • calibración;
 - • cambio de accesorio.
- **En modelos tabulares:**
 - • redondeo;
 - • datos faltantes;
 - • valores próximos al límite;
 - • errores de introducción;
 - • cambios de unidades;
 - • pequeñas desviaciones de pH o actividad de agua;
 - • sustitución de un ingrediente por otro equivalente.
- **La prueba de robustez puede expresar la variación de la salida ante una perturbación (Δx):**

$$\Delta \hat{y} f(x+\Delta x)-f(x)$$

El cambio debe ser coherente con el conocimiento tecnológico.

Si una variación de 0,01 unidades de pH, dentro del error normal del método, provoca que el modelo pase de una probabilidad de riesgo del 5 % al 90 %, existe un problema de inestabilidad o de proximidad a una frontera de decisión que debe gestionarse.

6.11. Pruebas de estrés

Las pruebas de estrés someten el sistema a condiciones difíciles pero plausibles.

Su objetivo no es obtener una métrica promedio, sino identificar modos de fallo.

Deben incluir escenarios como:

- • combinación de variables extremas;
- • pérdida parcial de sensores;
- • retraso en la recepción de datos;
- • imágenes degradadas;
- • concentraciones próximas al límite de detección;
- • formulaciones infrecuentes;
- • productos con alta variabilidad natural;
- • condiciones límite de temperatura;
- • incremento repentino de prevalencia;

- cambio de proveedor;
- mezcla de datos procedentes de distintos equipos;
- entradas contradictorias.

Las pruebas deben diseñarse con participación de tecnólogos, microbiólogos, responsables de calidad, operarios y especialistas del proceso. El equipo informático por sí solo puede no identificar los fallos alimentarios más relevantes.

6.12. Coherencia tecnológica

Un modelo puede presentar buen rendimiento estadístico y, sin embargo, producir respuestas incompatibles con el conocimiento científico.

En determinados dominios deberían cumplirse relaciones razonables, como:

- un aumento de la temperatura dentro del rango de crecimiento no debería reducir sistemáticamente el riesgo microbiano;
- una reducción efectiva de pH no debería incrementar sin explicación la velocidad de crecimiento de un microorganismo ácido-lábil;
- una mayor exposición al oxígeno no debería prolongar la estabilidad oxidativa de una matriz sensible;
- una interrupción de la cadena de frío no debería aumentar la vida útil remanente;
- una concentración superior de un conservante eficaz no debería empeorar sistemáticamente la predicción de estabilidad.

No todas las relaciones son estrictamente monótonas. Pueden existir interacciones, rangos óptimos y fenómenos no lineales. Pero las salidas contraintuitivas deben investigarse.

La validación debe incorporar pruebas de coherencia basadas en hipótesis tecnológicas.

Por ejemplo:

$$T_2 > T_1$$

y condiciones iguales deberían implicar, dentro de un rango microbiológicamente relevante:

$$\hat{\mu}(T_2) \geq \hat{\mu}(T_1)$$

donde ($\hat{\mu}$) representa la tasa de crecimiento predicha.

Si el modelo viola sistemáticamente este comportamiento, puede estar utilizando correlaciones espurias.

6.13. Pruebas contrafactuales

Las pruebas contrafactuales analizan cómo cambia la predicción cuando se modifica una variable manteniendo las demás constantes.

Ejemplo:

- predicción inicial: vida útil de 32 días;
- se incrementa la temperatura de almacenamiento;
- nueva predicción: 38 días.

Este resultado exigiría una explicación. Puede indicar:

- interacción real no evidente;
- problema de codificación;
- variable correlacionada;
- extrapolación;
- falta de cobertura;

- comportamiento no físico.
- Los contrafactuales son especialmente útiles para detectar dependencias indeseadas.

No obstante, algunos escenarios contrafactuales pueden ser imposibles. No debe modificarse una variable aislada hacia combinaciones que nunca podrían existir tecnológicamente.

Por ejemplo, alterar el tratamiento térmico sin modificar ninguna variable asociada puede generar un producto virtual incompatible con la realidad. Las pruebas deben respetar las relaciones causales y de formulación.

6.14. Datos fuera de distribución

Un modelo debería reconocer, en la medida de lo posible, cuándo una observación difiere de los datos con los que fue desarrollado.

El problema se denomina detección *out-of-distribution* u OOD.

Una muestra puede quedar fuera de distribución por:

- nueva materia prima;
- nuevo producto;
- proveedor desconocido;
- equipo distinto;
- cambio de formulación;
- rango de concentración no observado;
- condiciones ambientales extremas;
- imagen anómala;
- fallo de sensor;
- nueva flora microbiana;
- modificación del proceso.

El algoritmo puede seguir emitiendo una predicción numérica, incluso con aparente confianza. Esta seguridad artificial constituye uno de los riesgos más importantes del despliegue.

6.15. Distancia al dominio de entrenamiento

En modelos tabulares puede calcularse la distancia entre una observación nueva y la población de entrenamiento.

Métodos posibles incluyen:

- distancia de *Mahalanobis*;
 - vecinos más próximos;
 - densidad local;
 - modelos de una clase;
 - *isolation forest*;
 - *auto-encoders*;
 - análisis de componentes principales;
 - profundidad estadística.
- La distancia de *Mahalanobis* considera la correlación entre variables:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

donde:

- (μ) es el vector medio;

- Σ es la matriz de covarianza.
- Una distancia elevada sugiere que la observación se encuentra lejos de la población de referencia.

Sin embargo, ningún método garantiza por sí solo la detección de todos los casos fuera de dominio. Deben combinarse controles estadísticos con reglas tecnológicas explícitas.

6.16. Límites duros y límites blandos

El dominio puede controlarse mediante dos tipos de límites.

Límites duros

Son condiciones que invalidan automáticamente la predicción:

$$\text{pH} < 4,2$$

o:

$$\text{pH} > 6,0$$

si esos valores quedan fuera de la validación.

Otros ejemplos:

- modelo de equipo no autorizado;
- producto no registrado;
- datos esenciales ausentes;
- temperatura fuera del intervalo;
- imagen con resolución insuficiente;
- versión de sensor incompatible.

Límites blandos

Identifican observaciones poco representadas, aunque todavía incluidas formalmente en el rango.

Por ejemplo, un producto puede situarse dentro de los límites individuales de todas las variables, pero presentar una combinación nunca observada.

Los límites blandos pueden activar:

- advertencia;
- reducción de confianza;
- revisión humana;
- análisis adicional;
- abstención.

Ambos tipos de control son necesarios. Los límites mínimos y máximos por variable no detectan interacciones atípicas.

6.17. Predicción selectiva y cobertura

Un sistema con capacidad de abstención puede decidir únicamente en los casos para los que dispone de confianza suficiente.

La cobertura se define como:

$$\text{Cobertura} = \frac{[\text{N_predicciones aceptadas}]}{[\text{N_total}]}$$

El riesgo selectivo mide el error entre las predicciones aceptadas.

Al aumentar el nivel de confianza exigido, normalmente:

- disminuye la cobertura;
- mejora el rendimiento en los casos aceptados;
- aumenta el número de casos remitidos a revisión.
- Debe evaluarse la curva riesgo-cobertura.

Un sistema podría alcanzar una tasa de error muy baja decidiendo solo sobre el 20 % de los lotes. Aunque científicamente correcto, podría carecer de utilidad industrial.

Los criterios de aceptación deben establecer conjuntamente:

$$\text{Error selectivo} \leq E_{\text{max}}$$

y:

$$\text{Cobertura} \geq C_{\text{min}}$$

6.18. Umbrales de confianza

La confianza puede utilizarse para definir tres zonas:

$$\hat{p} < t_{\text{bajo}} \Rightarrow \text{riesgo bajo}$$

$$t_{\text{bajo}} \leq \hat{p} \leq t_{\text{alto}} \Rightarrow \text{indeterminado}$$

$$\hat{p} > t_{\text{alto}} \Rightarrow \text{riesgo alto}$$

Pero los umbrales solo tienen sentido cuando las probabilidades están calibradas. Una puntuación interna de 0,90 no representa necesariamente una confianza del 90 %.

La empresa debe distinguir entre:

- probabilidad calibrada;
- puntuación de decisión;
- confianza del clasificador;
- incertidumbre estimada;
- distancia al dominio.
- Mezclar estos conceptos conduce a interpretaciones erróneas.

6.19. Conjuntos de modelos

Los métodos de ensemble combinan múltiples modelos.

La variabilidad entre sus predicciones puede utilizarse como aproximación de incertidumbre:

$$\hat{y} [1] / [M] \sum_{m=1}^M \hat{y}^{(m)}$$

y:

$$s^2 = [1] / [M-1] \sum_{m=1}^M (\hat{y}^{(m)} - \hat{y})^2$$

Una dispersión elevada puede indicar que la observación es difícil o está poco representada.

Sin embargo, el desacuerdo entre modelos no equivale automáticamente a incertidumbre correctamente calibrada. Si todos comparten los mismos datos sesgados o la misma arquitectura, pueden equivocarse de forma coincidente y con aparente seguridad.

6.20. Métodos bayesianos y aproximaciones

Los enfoques bayesianos representan distribuciones sobre parámetros o predicciones.

En lugar de generar un único valor, pueden estimar:

$$p(y|x,D)$$

donde (D) representa los datos observados.

En redes neuronales profundas se utilizan aproximaciones como:

- • *Monte Carlo dropout*;
- • *variational inference*;
- • *deep ensembles*;
- • *Laplace approximation*.
- Estas técnicas pueden mejorar la estimación de incertidumbre, pero no eliminan la necesidad de validación externa.
- Una incertidumbre matemáticamente sofisticada puede seguir estando mal calibrada si los datos no son representativos.

6.21. Predicción conformal

La predicción *conformal* permite construir conjuntos o intervalos predictivos con garantías de cobertura bajo determinados supuestos de intercambiabilidad.

En regresión, puede proporcionar un intervalo:

$$C(x)=[L(x),U(x)]$$

tal que, aproximadamente:

$$P(Y \in C(X)) \geq 1-\alpha$$

En clasificación, puede generar un conjunto de clases plausibles:

$$C(x) \subseteq \{0,1,\dots,K\}$$

Una observación puede recibir:

- • una única clase;
- • varias clases posibles;
- • un conjunto vacío, dependiendo del método.
- Su ventaja principal es que transforma la incertidumbre en un resultado explícito.

Sin embargo, las garantías dependen de que los datos de calibración y los futuros sean suficientemente intercambiables. Si existe deriva, la cobertura puede degradarse.

La predicción *conformal* no debe presentarse como una garantía universal frente a cambios de proceso o de población.

6.22. Incertidumbre por datos faltantes

Los datos incompletos son inevitables en producción.

El modelo puede recibir:

- • una lectura de temperatura ausente;
- • un valor de actividad de agua no disponible;
- • una imagen corrupta;
- • un resultado analítico retrasado;
- • un proveedor sin codificación;
- • un dato manual erróneo.
- La validación debe definir cómo se gestionan.

- **Las opciones incluyen:**

- • impedir la predicción;
- • imputar el dato;
- • utilizar un modelo alternativo;
- • asignar una categoría desconocida;
- • remitir a revisión.

La imputación no es neutra. Sustituir un dato ausente por la media puede reducir artificialmente el riesgo y generar una predicción excesivamente segura.

Debe evaluarse el rendimiento según patrones realistas de ausencia, no únicamente eliminando registros incompletos.

6.23. Robustez frente a errores de unidades y codificación

Una parte sustancial de los fallos no procede del algoritmo, sino de la interfaz de datos.

Ejemplos:

- • temperatura en Fahrenheit interpretada como Celsius;
- • concentración expresada en porcentaje frente a g/kg;
- • actividad de agua codificada como 95 en lugar de 0,95;
- • fechas intercambiadas;
- • coma decimal interpretada como separador de miles;
- • categorías de proveedor reasignadas;
- • cambios de nomenclatura;
- • valores negativos imposibles.

- **La validación debe probar controles de plausibilidad:**

$$0 \leq a_w \leq 1$$

$$pH \in [0, 14]$$

y límites más restrictivos según el producto.

También deben comprobarse unidades, escalas, codificación y consistencia entre sistemas.

La robustez informática forma parte de la robustez científica porque un modelo solo puede ser tan fiable como los datos que recibe.

6.24. Robustez frente a ataques o manipulaciones

Cuando un sistema influye en liberación, clasificación o valoración económica, puede existir incentivo para manipular sus entradas.

Entre los escenarios relevantes se incluyen:

- • alteración deliberada de imágenes;
- • modificación de variables;
- • selección de la zona de muestra más favorable;
- • repetición hasta obtener un resultado aceptable;
- • suplantación de sensor;
- • adulteración de metadatos;
- • uso de una versión no autorizada.

- **La validación y el diseño deben incluir:**

- • autenticación;
- • integridad de datos;
- • registros de auditoría;
- • control de versiones;
- • detección de anomalías;
- • separación de funciones;
- • bloqueo de entradas imposibles;
- • conservación de evidencia.
- No toda aplicación alimentaria estará expuesta al mismo riesgo, pero el análisis no debe descartarse por considerar que el modelo es exclusivamente técnico.

6.25. Robustez operativa

El sistema debe mantener su función bajo condiciones ordinarias de producción:

- • carga de trabajo elevada;
- • latencia;
- • desconexión temporal;
- • fallo de red;
- • actualización de software;
- • indisponibilidad de un servicio;
- • reinicio;
- • cambio de turno;
- • uso simultáneo;
- • acumulación de datos.
- **Deben validarse:**
 - • tiempo máximo de respuesta;
 - • disponibilidad;
 - • recuperación;
 - • almacenamiento de resultados;
 - • ausencia de duplicados;
 - • gestión de errores;
 - • comportamiento en modo degradado.
- Una predicción que llega después de que el lote haya sido expedido carece de utilidad, aunque sea estadísticamente exacta.

6.26. Supervisión humana y confianza apropiada

La supervisión humana no consiste simplemente en mostrar el resultado a un operario.

Para que sea efectiva, la persona debe:

- • comprender la finalidad del modelo;
- • conocer sus límites;
- • disponer de información suficiente;
- • poder cuestionar la salida;

- tener autoridad para anularla;
- registrar el motivo;
- recibir formación;
- no estar sometida a una carga de alertas imposible.
- Existe el riesgo de *automation bias*, por el cual el usuario acepta la recomendación algorítmica incluso cuando contradice señales relevantes.
- También existe el efecto contrario: el rechazo sistemático del sistema cuando genera demasiadas falsas alarmas.
- **La validación de uso debe analizar:**
 - tasa de aceptación de recomendaciones;
 - anulaciones correctas e incorrectas;
 - tiempo de revisión;
 - comprensión de la interfaz;
 - dependencia excesiva;
 - fatiga de alertas;
 - diferencias entre usuarios.

6.27. Explicabilidad e incertidumbre

Una explicación no sustituye la validación.

Métodos como:

- importancia de variables;
- SHAP;
- LIME;
- mapas de activación;
- atención;
- reglas locales;
- pueden ayudar a investigar el comportamiento, pero no demuestran que la predicción sea correcta.

Una explicación plausible puede acompañar a un resultado erróneo. Del mismo modo, una variable importante en el modelo no tiene por qué ser causa del fenómeno.

Las explicaciones deben utilizarse para:

- detectar dependencias espurias;
- comprobar coherencia;
- apoyar la revisión humana;
- identificar cambios;
- documentar decisiones.
- No deben utilizarse como prueba aislada de fiabilidad científica.

6.28. Criterios de aceptación para calibración y robustez

Los criterios deben adaptarse al uso previsto, pero pueden incluir:

Calibración

$$|\alpha| \leq 0,1$$

$$0,9 \leq \beta \leq 1,1$$

junto con intervalos de confianza adecuados.

Puntuación probabilística

$$BSS > 0$$

respecto a un modelo de referencia previamente definido.

Cobertura de intervalos

$$\text{Cobertura observada del 95 \%} \geq 92\%$$

sin amplitud operativamente inaceptable.

Robustez

- ausencia de degradación superior a un límite predefinido ante perturbaciones ordinarias;
- mantenimiento de sensibilidad mínima en condiciones adversas;
- detección de entradas fuera de dominio;
- tasa máxima de fallos técnicos;
- estabilidad de las salidas ante incertidumbre analítica.

Abstención

$$\text{Cobertura} \geq C_{\min}$$

y:

$$\text{Tasa de error en casos aceptados} \leq E_{\max}$$

Coherencia

- ausencia de violaciones tecnológicas graves;
- investigación documentada de comportamientos contrafactuales anómalos;
- revisión de casos extremos.

Los límites numéricos no deben copiarse mecánicamente. Deben derivarse de la gravedad del resultado, la incertidumbre tolerable y la función del modelo.

6.29. La obligación de reconocer la ignorancia

Uno de los atributos más valiosos de un sistema fiable es su capacidad para no responder cuando la evidencia es insuficiente.

Un modelo que siempre emite una decisión aparenta mayor utilidad, pero puede ocultar su incertidumbre.

En aplicaciones críticas, debería poder declarar:

- • datos insuficientes;
- • observación fuera del dominio;
- • resultado indeterminado;
- • baja confianza;
- • necesidad de análisis confirmatorio;
- • revisión humana obligatoria.

Esta abstención no constituye un fallo. Es una función de seguridad.

El verdadero fallo aparece cuando el sistema produce una conclusión categórica en una situación para la que no fue validado.

6.30. Conclusión del bloque

La calibración, la incertidumbre y la robustez convierten una puntuación algorítmica en una herramienta científicamente interpretable.

Un modelo no puede considerarse apto para decisiones alimentarias porque discrimine bien en una base histórica. Debe demostrar que:

- sus probabilidades representan frecuencias reales;
- su incertidumbre está cuantificada;
- identifica condiciones no representadas;
- tolera perturbaciones razonables;
- evita respuestas excesivamente confiadas;
- mantiene coherencia tecnológica;
- permite la abstención;
- conserva su rendimiento en el sistema completo.

La validación científicamente madura no pregunta únicamente si el modelo acierta. Pregunta también si sabe cuándo puede equivocarse, cuánto puede equivocarse y qué debe ocurrir cuando no existe evidencia suficiente para confiar en su salida.

7. Deriva del modelo: por qué una validación correcta puede dejar de ser válida

La validación inicial demuestra el rendimiento de una versión concreta del modelo sobre una población y unas condiciones determinadas. No demuestra que ese rendimiento vaya a mantenerse indefinidamente.

Una vez desplegado, el sistema entra en contacto con una realidad industrial dinámica. Cambian las materias primas, las campañas agrícolas, los proveedores, las formulaciones, los equipos, las condiciones ambientales, la microbiota residente, los procedimientos de limpieza, los operarios, los métodos analíticos y la propia prevalencia de los defectos o peligros.

Cuando estos cambios modifican la distribución de los datos o la relación entre las variables y el resultado, aparece la **deriva del modelo** o *model drift*.

La deriva puede definirse como la pérdida progresiva o abrupta de validez predictiva provocada por la diferencia entre las condiciones actuales de utilización y aquellas bajo las cuales el modelo fue desarrollado y validado.

Formalmente, durante el desarrollo se asume una distribución:

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

mientras que, después del despliegue, puede aparecer:

$$P_{\text{producción}}(X,Y) \neq P_{\text{desarrollo}}(X,Y)$$

donde:

- (X) representa las variables de entrada;
- (Y) representa el resultado real.

La deriva no implica necesariamente que el algoritmo haya cambiado. Puede permanecer exactamente igual mientras el mundo al que se aplica deja de ser el mismo.

Este hecho tiene una consecuencia fundamental:

La validación no es un certificado permanente, sino una afirmación condicionada que debe mantenerse mediante vigilancia, verificación y revalidación.

7.1. La deriva como problema alimentario, no exclusivamente informático

En entornos alimentarios, la deriva puede producirse por causas profundamente ligadas a la biología, la química y la tecnología de proceso.

Entre otras:

- variabilidad estacional de materias primas;
- cambio de origen geográfico;
- nuevas cosechas;
- modificaciones en la alimentación animal;
- variaciones en grasa, proteína, humedad o fibra;
- cambios de flora microbiana;
- adaptación de microorganismos al entorno;
- aparición de nuevos *biofilms*;
- modificación de conservantes;
- sustitución de ingredientes;
- cambios de envase;

- • modificaciones de atmósfera protectora;
 - • nuevas temperaturas logísticas;
 - • envejecimiento de equipos;
 - • sustitución de reactivos;
 - • recalibración de instrumentos;
 - • actualización de software;
 - • cambio de protocolos de muestreo.
- Un sistema de visión entrenado con tomates de una determinada campaña puede degradarse al cambiar tamaño, pigmentación, brillo superficial o frecuencia de defectos.

Un modelo NIR desarrollado con proteína vegetal de un proveedor puede perder exactitud al incorporar materias primas de otro origen con distinta granulometría, humedad, estructura o tratamiento térmico.

Un modelo de vida útil entrenado antes de una modificación de formulación puede dejar de representar el nuevo comportamiento microbiológico, aunque la modificación parezca menor desde el punto de vista comercial.

La vigilancia de la deriva debe diseñarse, por tanto, conjuntamente por científicos de datos, tecnólogos alimentarios, microbiólogos, responsables de calidad, mantenimiento, producción y laboratorio.

7.2. Deriva de covariables

La deriva de covariables o *covariate shift* se produce cuando cambia la distribución de las variables de entrada:

$$P_{\text{producción}}(X) \neq P_{\text{desarrollo}}(X)$$

aunque la relación entre las entradas y el resultado pueda mantenerse aproximadamente estable:

$$P(Y|X) \approx \text{constante}$$

Ejemplos:

- • aumento de la humedad media de una materia prima;
- • cambio en la distribución de pH;
- • incorporación de lotes con mayor contenido de grasa;
- • utilización de un nuevo formato de envase;
- • incremento de la temperatura media de distribución;
- • cambio de frecuencia de determinados defectos visuales;
- • uso de un nuevo instrumento analítico.

La deriva de covariables no implica automáticamente pérdida de rendimiento. Si el modelo sigue siendo válido en la nueva región, podría conservar sus prestaciones.

Sin embargo, constituye una señal de alerta porque el sistema está recibiendo observaciones distintas de las utilizadas para validarlo.

Debe investigarse especialmente cuando las nuevas observaciones se acercan a los límites del dominio o se concentran en regiones poco representadas.

7.3. Deriva del resultado o de prevalencia

La deriva del resultado se produce cuando cambia la frecuencia del evento:

$$P_{\text{producción}}(Y) \neq P_{\text{desarrollo}}(Y)$$

Por ejemplo:

- • disminuye la tasa de contaminación después de mejorar la higiene;

- • aumenta la incidencia de defectos por un problema de proveedor;
- • cambia la proporción de productos rechazados;
- • aparece un incremento estacional de deterioro;
- • se endurece el criterio interno de conformidad;
- • se modifica la definición de positivo.

• **Un cambio de prevalencia afecta especialmente a:**

- • valor predictivo positivo;
- • valor predictivo negativo;
- • calibración;
- • volumen de alertas;
- • utilidad del umbral.
- Un modelo puede conservar sensibilidad y especificidad y, sin embargo, presentar valores predictivos muy diferentes.
- Si el evento se vuelve más raro, disminuirá normalmente el valor predictivo positivo y aumentará la proporción de falsas alarmas entre las alertas.
- Si el evento se vuelve más frecuente, puede disminuir el valor predictivo negativo y aumentar el riesgo de que una predicción negativa sea incorrecta.

La vigilancia no debe limitarse, por tanto, a las métricas teóricamente independientes de la prevalencia. Debe analizar cómo cambia la utilidad real del sistema.

7.4. Deriva conceptual

La deriva conceptual o *concept drift* se produce cuando cambia la relación entre las variables y el resultado:

$$P_{\text{producción}}(Y|X) \neq P_{\text{desarrollo}}(Y|X)$$

Es una de las formas más graves porque la lógica aprendida por el modelo deja de representar el proceso.

Puede aparecer cuando:

- • se modifica una formulación;
- • cambia la flora microbiana predominante;
- • se introduce un nuevo tratamiento;
- • un conservante pierde eficacia por interacción con otros ingredientes;
- • cambia el mecanismo principal de deterioro;
- • aparece un nuevo defecto;
- • se modifica la respuesta del proceso a la temperatura;
- • una variable deja de ser representativa;
- • se introduce una medida correctora que altera el riesgo.

Ejemplo: un modelo de vida útil puede haber aprendido que la reducción del pH está asociada a mayor estabilidad. Si se incorpora un nuevo producto en el que el principal deterioro procede de levaduras acidotolerantes, la relación puede cambiar.

La deriva conceptual es difícil de detectar sin disponer de resultados verdaderos posteriores. Las entradas pueden parecer normales mientras la asociación con el resultado se ha modificado.

Por ello, la monitorización de datos de entrada no sustituye la verificación periódica contra el patrón de referencia.

7.5. Deriva de etiqueta

La deriva de etiqueta aparece cuando cambia la definición, medición o asignación del resultado verdadero.

Puede producirse por:

- • nuevo método microbiológico;
 - • cambio de límite de detección;
 - • modificación del criterio de aceptación;
 - • sustitución del panel sensorial;
 - • cambio de escala;
 - • nueva clasificación comercial;
 - • modificación de la fecha de lectura;
 - • cambio de laboratorio;
 - • nueva interpretación de resultados indeterminados.
- **Supóngase que un modelo fue entrenado para detectar lotes con:**

$$N > 100 \backslash \text{UFC/g}$$

y posteriormente la organización redefine el evento como:

$$N > 10 \backslash \text{UFC/g}$$

El modelo ya no está resolviendo el problema actual, aunque continúe reproduciendo correctamente la definición histórica.

Cualquier cambio en el patrón de referencia o en el criterio de decisión debe someterse a control de cambios y evaluación de impacto.

7.6. Deriva del sensor o del sistema de medida

Los datos pueden cambiar porque cambia el instrumento que los produce.

Entre las causas:

- • envejecimiento;
 - • suciedad;
 - • desgaste;
 - • recalibración;
 - • sustitución;
 - • cambio de software;
 - • modificación de resolución;
 - • cambio de iluminación;
 - • nueva configuración;
 - • deriva de longitud de onda;
 - • variación de sensibilidad;
 - • actualización del firmware.
- En espectroscopia, pequeñas diferencias instrumentales pueden desplazar la distribución de señales.
 - En visión artificial, una cámara de sustitución puede presentar diferente balance de blancos, rango dinámico o respuesta de color.
 - En sensores de gases, la sensibilidad puede degradarse progresivamente.

- **La validación debe definir:**

- • instrumentos autorizados;
- • procedimientos de calibración;
- • tolerancias;
- • pruebas de equivalencia;
- • frecuencia de verificación;
- • necesidad de transferencia de calibración;
- • condiciones para revalidar.
- No debe asumirse que dos equipos del mismo modelo producen datos perfectamente intercambiables.

7.7. Deriva temporal gradual, abrupta y recurrente

La deriva puede clasificarse según su dinámica.

Deriva gradual

El cambio se acumula lentamente.

Ejemplos:

- • envejecimiento de un sensor;
- • modificación progresiva de proveedor;
- • evolución de la flora ambiental;
- • desgaste de una línea;
- • cambio sostenido de condiciones logísticas.
- Es difícil de detectar visualmente porque cada variación individual parece pequeña.

Deriva abrupta

Se produce de forma repentina.

Ejemplos:

- • cambio de formulación;
- • sustitución de equipo;
- • entrada de un nuevo proveedor;
- • modificación del método analítico;
- • avería;
- • cambio de criterio de calidad.
- Puede generar una caída inmediata del rendimiento.

Deriva recurrente

Aparece periódicamente.

Ejemplos:

- • variación estacional;
- • campañas de cosecha;
- • producción de referencias especiales;
- • alternancia entre proveedores;
- • cambios de temperatura verano-invierno;
- • turnos con diferentes prácticas.

- La deriva recurrente puede confundirse con ruido si no se analiza el contexto temporal.

7.8. Deriva real y deriva aparente

No todo cambio estadístico implica que el modelo esté degradándose.

Puede observarse una deriva aparente por:

- • modificación del plan de muestreo;
- • mayor intensidad de controles;
- • selección de casos más difíciles;
- • incorporación de nuevas líneas;
- • cambios administrativos;
- • corrección de errores históricos;
- • aumento de la vigilancia tras un incidente.
- Por ejemplo, si la empresa comienza a analizar preferentemente lotes de alto riesgo, aumentará la prevalencia observada sin que necesariamente haya empeorado el proceso.
- La interpretación debe considerar cómo se generan los datos.

Detectar una diferencia estadística es solo el primer paso. Debe investigarse su origen y relevancia tecnológica.

7.9. Qué debe monitorizarse

La vigilancia debe abarcar al menos cinco capas:

1. • calidad de los datos;
2. • distribución de entradas;
3. • distribución de salidas;
4. • rendimiento contra resultados reales;
5. • impacto operativo.

Calidad de datos

- valores ausentes;
- unidades;
- rangos imposibles;
- registros duplicados;
- errores de codificación;
- fallos de sensor;
- retrasos;
- integridad de archivos;
- calidad de imagen;
- coherencia temporal.

Variables de entrada

- medias;
- medianas;
- desviaciones;

- percentiles;
- categorías;
- correlaciones;
- combinaciones;
- distancia al dominio;
- frecuencia de valores extremos.

Salidas del modelo

- distribución de probabilidades;
- frecuencia de alertas;
- tasa de positivos;
- abstenciones;
- confianza;
- incertidumbre;
- decisiones por categoría.

Rendimiento

- sensibilidad;
- especificidad;
- valores predictivos;
- MAE;
- sesgo;
- calibración;
- errores críticos;
- rendimiento por subgrupos.

Impacto

- lotes retenidos;
- análisis confirmatorios;
- desperdicio;
- tiempo de respuesta;
- incidencias;
- reclamaciones;
- anulaciones humanas;
- carga de trabajo;
- costes.

Una deriva de salida sin deriva visible de entradas puede indicar cambios en el sistema, en la prevalencia o en la relación entre variables.

7.10. Ventanas de monitorización

Las métricas deben calcularse sobre ventanas temporales definidas.

Ventana fija

Por ejemplo:

- • últimos 30 días;
- • último trimestre;
- • últimos 1.000 lotes.
- Es fácil de interpretar, pero puede responder lentamente a cambios abruptos si la ventana es grande.

Ventana móvil

Se actualiza continuamente:

$$W_t = t-n+1, \dots, t$$

Permite observar tendencias recientes.

Ventana acumulativa

Incluye todos los datos desde el despliegue.

Es estable, pero puede ocultar degradaciones recientes porque el volumen histórico domina el cálculo.

Ventanas múltiples

Una estrategia robusta combina:

- • ventana corta para alertas tempranas;
- • ventana media para confirmación;
- • ventana larga para tendencia.
- **Por ejemplo:**
 - • 7 días;
 - • 30 días;
 - • 12 meses.

La longitud debe adaptarse al volumen y frecuencia de los eventos. En situaciones con positivos raros, una ventana demasiado corta no permitirá estimar la sensibilidad con precisión.

7.11. Indicadores univariantes

Para una variable continua puede compararse la distribución actual con la de referencia mediante:

- • diferencia de medias;
- • diferencia de medianas;
- • cambio de desviación;
- • percentiles;
- • prueba de *Kolmogorov–Smirnov*;
- • distancia de *Wasserstein*;
- • índice de estabilidad poblacional;
- • divergencia de Jensen–Shannon.
- **Para variables categóricas:**
 - • cambio de frecuencias;

- prueba chi-cuadrado;
- divergencia entre distribuciones;
- aparición de categorías desconocidas.
- Estas técnicas permiten detectar cambios, pero no deben interpretarse de forma automática.

Con muestras grandes, diferencias irrelevantes pueden resultar estadísticamente significativas. Con muestras pequeñas, cambios importantes pueden no alcanzar significación.

La evaluación debe combinar magnitud estadística y relevancia tecnológica.

7.12. Population Stability Index

El PSI se utiliza para comparar la distribución de una variable entre una población de referencia y otra actual.

De forma simplificada:

$$PSI = \sum_{k=1}^K (A_k - E_k) \ln ([A_k] / [E_k])$$

donde:

- (A_k) es la proporción actual en el intervalo (k) ;
- (E_k) es la proporción esperada o de referencia.
- Valores mayores indican mayor cambio.

Sin embargo, no deben adoptarse umbrales universales sin validación. El PSI depende de:

- número de intervalos;
- tamaño muestral;
- distribución;
- tratamiento de valores nulos;
- elección de la referencia.

Una variación pequeña en pH puede ser microbiológicamente importante, aunque el PSI sea bajo. Una variación amplia en una variable irrelevante puede generar un PSI alto sin afectar el rendimiento.

7.13. Distancias entre distribuciones

La distancia de *Wasserstein* cuantifica el desplazamiento necesario para transformar una distribución en otra.

Para distribuciones unidimensionales:

$$W(P,Q) \in \int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)| du$$

Mantiene relación con la escala de la variable, por lo que puede ser interpretable si se normaliza adecuadamente.

La divergencia de Jensen–Shannon compara distribuciones de probabilidad y es simétrica:

$$JS(P,Q) = \frac{1}{2} KL(P|M) + \frac{1}{2} KL(Q|M)$$

donde:

$$M = [P+Q] / 2$$

Estas medidas son útiles para vigilancia automática, pero no indican directamente si el cambio deteriorará el modelo.

7.14. Deriva multivariante

Las variables pueden conservar individualmente distribuciones similares mientras cambian sus relaciones.

Por ejemplo, la distribución de pH y temperatura puede permanecer estable por separado, pero aparecer una combinación nueva de pH elevado y temperatura alta.

La monitorización multivariante puede utilizar:

- • distancia de *Mahalanobis*;
- • análisis de componentes principales;
- • control estadístico multivariante;
- • clasificadores de dominio;
- • *maximum mean discrepancy*;
- • modelos de densidad;
- • métodos de detección de anomalías.
- Un clasificador de dominio intenta distinguir si una observación procede del conjunto histórico o de la población actual.
- Si puede separarlas con elevada capacidad, existe evidencia de cambio en la distribución conjunta.

Sin embargo, el clasificador puede detectar diferencias irrelevantes. La investigación debe identificar qué variables y combinaciones explican el cambio.

7.15. Monitorización de probabilidades y alertas

La distribución de salidas puede proporcionar señales tempranas.

Deben vigilarse:

- • probabilidad media;
- • percentiles;
- • concentración cerca del umbral;
- • porcentaje de positivos;
- • frecuencia de valores extremos;
- • tasa de abstención;
- • incertidumbre media;
- • número de entradas fuera de dominio.
- Un aumento de observaciones cercanas al umbral puede incrementar la inestabilidad de las decisiones.
- **Un incremento repentino de alertas puede significar:**
 - • deterioro real del proceso;
 - • deriva del sensor;
 - • cambio de prevalencia;
 - • error de datos;
 - • fallo de calibración;
 - • nueva materia prima;
 - • problema del modelo.

No debe corregirse automáticamente ajustando el umbral. Primero debe investigarse la causa.

7.16. Monitorización del rendimiento con etiquetas retrasadas

En muchas aplicaciones, el resultado real no se conoce inmediatamente.

La contaminación puede confirmarse días después. La vida útil real solo se observa al finalizar el estudio. Las reclamaciones pueden aparecer semanas o meses más tarde.

Debe distinguirse entre:

Monitorización sin resultado verdadero

Utiliza:

- • entradas;
- • salidas;
- • incertidumbre;
- • distancia al dominio;
- • calidad de datos;
- • indicadores de proceso.
- Permite detectar señales indirectas, pero no demuestra pérdida de rendimiento.

Monitorización con resultado verdadero

Compara las predicciones con:

- • análisis confirmatorios;
- • resultados microbiológicos;
- • inspecciones;
- • deterioro real;
- • datos sensoriales;
- • reclamaciones;
- • decisiones finales.
- Es la única que permite verificar directamente sensibilidad, especificidad, sesgo o calibración.
- La empresa debe diseñar un flujo para incorporar los resultados reales al sistema de vigilancia.

7.17. Muestreo de verificación

Aunque el modelo clasifique un lote como seguro, debe mantenerse un programa de muestreo independiente.

Si solo se analizan los casos que el modelo marca como positivos, no podrán detectarse falsos negativos.

Este sesgo de verificación produce una evaluación circular: el sistema parece correcto porque solo se comprueban los casos que él mismo selecciona.

Debe analizarse una muestra de predicciones negativas mediante:

- • muestreo aleatorio;
- • muestreo estratificado;
- • sobremuestreo de baja confianza;
- • sobremuestreo de condiciones extremas;
- • muestreo por producto y planta;
- • controles ciegos.
- La proporción debe ser suficiente para detectar degradaciones relevantes.

7.18. Sesgo de retroalimentación

El despliegue del modelo puede modificar el proceso que pretende predecir.

Por ejemplo:

1. • el modelo identifica lotes de alto riesgo;
2. • la empresa aplica medidas correctoras;
3. • disminuye la tasa de no conformidad en esos lotes;

4. • los datos posteriores muestran menos eventos;
5. • el modelo parece sobreestimar el riesgo.

Este fenómeno no significa necesariamente que estuviera equivocado. Su intervención ha cambiado el resultado.

También puede ocurrir que se dejen de analizar lotes clasificados como seguros, reduciendo la disponibilidad de negativos y falsos negativos verificados.

La evaluación debe distinguir entre:

- • rendimiento predictivo;
- • efecto de la intervención;
- • cambios en el proceso de recogida de etiquetas.

7.19. Control estadístico del rendimiento

Las métricas pueden vigilarse mediante gráficos de control.

Por ejemplo, la tasa de falsos negativos observada en cada periodo:

$$\text{FNR}_t = [\text{FN}_t] / [\text{VP}_t + \text{FN}_t]$$

puede compararse con un límite de alerta.

Sin embargo, cuando los positivos son escasos, la proporción será inestable. Deben utilizarse:

- • intervalos binomiales;
- • acumulación secuencial;
- • métodos bayesianos;
- • CUSUM;
- • EWMA;
- • modelos beta-binomiales.

• El CUSUM acumula pequeñas desviaciones:

$$S_t = \max [0, S_{t-1} + (x_t - \mu_0 - k)]$$

donde:

- • (μ_0) es el rendimiento esperado;
- • (k) es una tolerancia.
- Puede detectar degradaciones persistentes antes que un análisis periódico convencional.

7.20. Indicadores adelantados y retrasados

Los indicadores adelantados aparecen antes de que se confirme una caída del rendimiento:

- • aumento de datos fuera de dominio;
- • cambio de proveedor;
- • desplazamiento de variables;
- • aumento de incertidumbre;
- • cambios de equipo;
- • modificación de formulación;
- • deterioro de calidad de imagen;
- • nuevas categorías.

• Los indicadores retrasados confirman el efecto:

- • aumento de falsos negativos;

- • disminución de sensibilidad;
- • incremento de MAE;
- • pérdida de calibración;
- • más reclamaciones;
- • mayor frecuencia de producto no conforme.

Un sistema robusto utiliza ambos. Esperar exclusivamente a los indicadores retrasados puede significar detectar el problema después de haber afectado a lotes comerciales.

7.21. Umbrales de alerta, acción y suspensión

No todas las señales deben provocar la misma respuesta.

Puede establecerse una estructura escalonada.

Nivel verde: funcionamiento normal

- métricas dentro de límites;
- entradas representativas;
- ausencia de señales relevantes;
- operación ordinaria.

Nivel amarillo: alerta

- desplazamiento moderado;
- aumento de incertidumbre;
- nuevo proveedor;
- incremento de abstenciones;
- desviación de una métrica secundaria.

Acciones:

- • investigación;
- • aumento de muestreo;
- • revisión técnica;
- • seguimiento intensivo.

Nivel naranja: acción correctiva

- incumplimiento de una métrica;
- degradación persistente;
- pérdida de calibración;
- nuevo subgrupo con rendimiento insuficiente.

Acciones:

- • restringir el dominio;
- • recalibrar;
- • modificar umbral;
- • exigir confirmación;
- • iniciar revalidación.

Nivel rojo: suspensión

- sensibilidad inferior al mínimo;
- falso negativo grave;
- fallo de integridad;
- entradas no controladas;
- deriva conceptual confirmada;
- cambio no autorizado;
- imposibilidad de verificar el rendimiento.

Acciones:

- • suspender decisiones automáticas;
- • volver al procedimiento convencional;
- • investigar;
- • corregir;
- • revalidar antes de reactivar.
- Los umbrales y acciones deben quedar documentados antes de que aparezca el incidente.

7.22. Modificación del umbral durante la operación

Cambiar el umbral modifica la relación entre sensibilidad y especificidad.

Por ello, no debe considerarse un simple ajuste operativo.

Un cambio de:

$$t=0,50$$

a:

$$t=0,35$$

puede aumentar las alertas, reducir falsos negativos y multiplicar falsos positivos.

Antes de aplicar el nuevo umbral deben evaluarse:

- • rendimiento histórico;
- • rendimiento externo;
- • impacto en capacidad;
- • calibración;
- • volumen de confirmaciones;
- • criterios de aceptación;
- • documentación y aprobación.
- Un umbral modificado genera una nueva configuración del sistema y puede requerir validación específica.

7.23. Recalibración, actualización y reentrenamiento

Estas actuaciones no son equivalentes.

Recalibración

Ajusta la correspondencia entre puntuaciones y probabilidades sin modificar sustancialmente la capacidad discriminativa.

Es adecuada cuando:

- • cambia la prevalencia;
- • existe desplazamiento sistemático;
- • se conserva la ordenación de riesgo.

Actualización parcial

Modifica algunos parámetros o incorpora nuevas observaciones.

Puede utilizarse cuando:

- • el proceso ha cambiado moderadamente;
- • se amplía la población;
- • aparecen nuevos productos;
- • se dispone de datos recientes.

Reentrenamiento

Reconstruye el modelo utilizando un conjunto actualizado.

Puede ser necesario cuando:

- • existe deriva conceptual;
- • aparecen nuevas variables;
- • cambia el método de referencia;
- • se modifica la formulación;
- • la degradación es sustancial.

• En todos los casos debe mantenerse:

- • control de versiones;
- • trazabilidad de datos;
- • separación de conjuntos;
- • evaluación comparativa;
- • posibilidad de reversión;
- • aprobación documentada.

7.24. Aprendizaje continuo

Algunos sistemas pueden actualizarse automáticamente con nuevos datos.

Esta capacidad puede resultar atractiva, pero plantea riesgos significativos.

Un modelo que aprende continuamente puede cambiar sin que exista una versión estable claramente validada.

Puede incorporar:

- • etiquetas erróneas;
- • datos manipulados;
- • sesgos de selección;
- • anomalías;
- • efectos temporales;
- • resultados influidos por sus propias decisiones.
- En aplicaciones alimentarias de riesgo, el aprendizaje continuo no debería significar modificación libre e inmediata en producción.

- **Una arquitectura prudente separa:**

1. • modelo activo validado;
 2. • modelo candidato actualizado;
 3. • entorno de evaluación;
 4. • aprobación;
 5. • despliegue controlado.
- El modelo candidato se entrena con nuevos datos, pero no sustituye al activo hasta completar la evaluación correspondiente.

7.25. Comparación entre versión activa y candidata

La nueva versión debe compararse con la anterior sobre:

- • conjunto externo común;
- • datos recientes;
- • subgrupos críticos;
- • condiciones adversas;
- • métricas de seguridad;
- • calibración;
- • robustez;
- • impacto operativo.
- No basta con demostrar que mejora la métrica media.
- **Debe comprobarse que no introduce regresiones:**
 - • menor sensibilidad en un producto;
 - • mayor tasa de falsos negativos;
 - • peor rendimiento en otra planta;
 - • pérdida de robustez;
 - • aumento de abstenciones;
 - • mayor dependencia de una variable inestable.
- Puede establecerse un criterio de no inferioridad para métricas críticas y superioridad para el objetivo principal.

7.26. Despliegue paralelo y modo sombra

Antes de sustituir una versión puede utilizarse un despliegue en modo sombra.

El modelo candidato recibe los mismos datos que el sistema activo, pero sus predicciones no influyen todavía en la decisión.

Esto permite comparar:

- • resultados;
- • discrepancias;
- • estabilidad;
- • tiempos de respuesta;
- • comportamiento en producción;
- • casos en los que uno detecta y el otro no.
- El modo sombra no sustituye una validación formal, pero aporta evidencia prospectiva y reduce riesgo.

7.27. Ensayos A/B y despliegue gradual

Cuando el riesgo lo permita, puede realizarse un despliegue gradual:

- • una línea;
- • una planta;
- • una familia de producto;
- • un porcentaje limitado de casos;
- • decisiones asistidas, no automáticas.
- Debe existir un procedimiento de reversión inmediata.

Los ensayos A/B requieren precaución cuando una de las estrategias podría ofrecer menor seguridad. No sería ético ni tecnológicamente aceptable exponer deliberadamente producto a un control inferior.

La comparación puede centrarse en variables operativas cuando ambas versiones ya cumplen los mínimos de seguridad.

7.28. Revalidación periódica

Aunque no se detecten señales evidentes, debe establecerse una revisión periódica.

La frecuencia puede depender de:

- • criticidad;
- • volumen;
- • velocidad de cambio;
- • estacionalidad;
- • frecuencia de eventos;
- • complejidad;
- • historial de incidencias;
- • dependencia del modelo.
- **La revisión puede ser:**
 - • mensual para indicadores operativos;
 - • trimestral para rendimiento;
 - • anual para revalidación integral;
 - • por campaña;
 - • tras cambios relevantes.

No existe una periodicidad universal. Debe justificarse mediante análisis de riesgo.

7.29. Revalidación activada por cambios

Además de la revisión periódica, determinados eventos deben activar una evaluación extraordinaria.

Entre ellos:

- • nueva formulación;
- • cambio de proveedor crítico;
- • nueva planta;
- • cambio de equipo;
- • modificación del método analítico;
- • actualización del software;

- • cambio de preprocesamiento;
- • modificación del umbral;
- • nueva población;
- • incidente de seguridad;
- • deterioro de métricas;
- • cambio regulatorio;
- • aparición de datos fuera de dominio;
- • cambio de definición del resultado.
- **Debe existir una matriz de control de cambios que determine:**
 - • impacto esperado;
 - • responsable;
 - • pruebas requeridas;
 - • necesidad de recalibración;
 - • revalidación parcial o total;
 - • autorización de puesta en servicio.

7.30. Revalidación parcial y completa

No todo cambio exige repetir íntegramente el estudio.

Revalidación parcial

Puede ser adecuada cuando:

- • se sustituye un equipo por otro demostrado equivalente;
- • cambia una interfaz sin modificar el cálculo;
- • se recalibra la probabilidad;
- • se añade un producto muy próximo al dominio existente;
- • se modifica un componente no crítico.
- Debe justificarse por qué las pruebas seleccionadas cubren el riesgo.

Revalidación completa

Resulta necesaria cuando:

- • cambia el objetivo;
- • se modifica la arquitectura;
- • se reentrena con nuevos datos;
- • cambia el patrón de referencia;
- • se amplía sustancialmente el dominio;
- • aparece deriva conceptual;
- • se modifica la decisión automatizada;
- • existe un incidente grave.
- La extensión de la revalidación debe ser proporcional al cambio, pero nunca inferior a la necesaria para demostrar que se mantienen los criterios críticos.

7.31. Gestión de incidentes

Debe existir un procedimiento específico para incidentes asociados al modelo.

Un incidente puede incluir:

- • falso negativo con impacto;
- • liberación incorrecta;
- • bloqueo masivo injustificado;
- • fallo técnico;
- • datos corruptos;
- • uso fuera del dominio;
- • versión incorrecta;
- • manipulación;
- • pérdida de registros;
- • error de interfaz.

• **El procedimiento debería contemplar:**

1. • identificación;
2. • contención;
3. • evaluación del producto afectado;
4. • trazabilidad de decisiones;
5. • análisis de causa raíz;
6. • corrección;
7. • verificación;
8. • revalidación;
9. • comunicación interna;
10. • evaluación de obligaciones externas.

La respuesta no debe limitarse a corregir el código. Puede ser necesario revisar:

- • lotes;
- • productos en mercado;
- • resultados analíticos;
- • decisiones humanas;
- • configuración;
- • formación;
- • procedimientos.

7.32. Conservación de registros

La trazabilidad debe permitir reconstruir cada predicción relevante.

Deberían conservarse:

- • fecha y hora;
- • versión del modelo;
- • variables de entrada;
- • origen de datos;
- • preprocesamiento;
- • salida;
- • probabilidad;
- • incertidumbre;

- • umbral;
 - • resultado final;
 - • usuario;
 - • decisión humana;
 - • anulaciones;
 - • resultado de referencia posterior.
- Sin estos registros no puede evaluarse adecuadamente un incidente ni verificar la deriva.
- **Debe garantizarse además que el registro sea:**
- • íntegro;
 - • accesible;
 - • protegido;
 - • versionado;
 - • interpretable;
 - • conservado durante un periodo coherente con la vida comercial y las obligaciones aplicables.

7.33. Cuadro mínimo de vigilancia

Un cuadro de mando de monitorización podría incluir:

Dimensión	Indicador	Frecuencia	Límite	Acción
Datos	Registros incompletos	Diaria	Máximo predefinido	Bloqueo o revisión
Dominio	Entradas fuera de rango	Continua	Cero en decisión automática	Abstención
Distribución	Deriva de variables críticas	Semanal	Umbral validado	Investigación
Salida	Tasa de alertas	Diaria/semanal	Banda histórica	Revisar proceso
Rendimiento	Sensibilidad	Mensual/trimestral	Mínimo aprobado	Suspensión
Rendimiento	Falsos negativos críticos	Inmediata	Tolerancia definida	Incidente
Calibración	Pendiente/intercepto	Trimestral	Rango aprobado	Recalibrar
Vida útil	Sobreestimación peligrosa	Por estudio	Máximo aprobado	Restringir
Operación	Tasa de abstención	Mensual	Máximo viable	Investigar
Humano	Anulaciones	Mensual	Desviación relevante	Formación/revisión

Los valores deben adaptarse a la finalidad y no copiarse de forma genérica.

7.34. Responsabilidades

La vigilancia requiere asignar responsabilidades concretas.

Propietario del modelo

Responsable de:

- • versión;
- • rendimiento;
- • documentación;
- • cambios;
- • incidencias.

Responsable del proceso

Evalúa:

- • cambios tecnológicos;
- • materias primas;
- • formulación;
- • condiciones de producción.

Calidad y seguridad alimentaria

Determina:

- • criterios;
- • acciones;
- • impacto sobre APPCC;
- • necesidad de retención;
- • verificación.

Tecnologías de la información

Garantiza:

- • disponibilidad;
- • integridad;
- • seguridad;
- • registros;
- • configuración.

Laboratorio

Asegura:

- • método de referencia;
- • calidad de etiquetas;
- • resultados confirmatorios;
- • incertidumbre analítica.

Comité de aprobación

Decide:

- • continuidad;
- • restricción;
- • recalibración;
- • revalidación;
- • suspensión.
- La responsabilidad no puede diluirse entre departamentos.

7.35. Relación con APPCC y sistemas de gestión

La monitorización del modelo debe integrarse con el sistema de seguridad alimentaria.

Cuando el algoritmo:

- • controla una variable relacionada con un peligro;

- • genera una alerta;
- • prioriza un análisis;
- • recomienda una acción;
- • interviene en liberación;
- sus fallos deben incorporarse al análisis de peligros.
- **Deben evaluarse peligros derivados del propio sistema:**
 - • no detección;
 - • falsa confianza;
 - • exceso de alertas;
 - • datos incorrectos;
 - • uso fuera de dominio;
 - • caída del sistema;
 - • actualización no validada;
 - • omisión humana.
- El plan debe definir medidas preventivas, monitorización, acciones correctivas, verificación y registros.

La IA no elimina la lógica APPCC. Introduce un componente adicional que también puede fallar y debe ser controlado.

7.36. Criterios científicos para mantener la validez

La validez de un modelo en producción puede considerarse mantenida cuando:

- • las entradas permanecen dentro del dominio;
- • la calidad de datos cumple los límites;
- • no existe deriva relevante no explicada;
- • las métricas continúan dentro de criterios;
- • la calibración se conserva;
- • los subgrupos mantienen rendimiento;
- • los falsos negativos permanecen bajo control;
- • la tasa de abstención es viable;
- • los cambios están documentados;
- • las verificaciones independientes son satisfactorias.
- La pérdida de cualquiera de estas condiciones no implica automáticamente que el modelo deba destruirse, pero sí que debe abrirse una evaluación formal.

7.37. La deriva como señal científica

La deriva no debe entenderse únicamente como un defecto del modelo.

Puede revelar cambios importantes en el propio sistema alimentario:

- • deterioro de materias primas;
- • cambio de microbiota;
- • problema de proceso;
- • desviación logística;
- • nueva fuente de contaminación;
- • modificación sensorial;

- fallo de limpieza;
- alteración de proveedores.
- Un sistema bien monitorizado puede actuar como detector temprano del cambio industrial.

La clave reside en no asumir que toda señal debe resolverse modificando el algoritmo. En ocasiones, el modelo está mostrando correctamente que el proceso ha cambiado.

7.38. Conclusión del bloque

El modelo desplegado debe considerarse un componente vivo dentro de un sistema productivo cambiante.

Su validez no se mantiene por inercia. Debe sostenerse mediante:

- vigilancia continua;
- verificación independiente;
- muestreo de predicciones negativas;
- seguimiento de entradas y resultados;
- control de versiones;
- gestión de cambios;
- revalidación;
- capacidad de suspensión.

• La afirmación:

«El modelo fue validado en 2025»

no demuestra que siga siendo válido en 2026 después de cambiar materias primas, proveedores, formulación, equipos o prevalencia.

La afirmación científicamente defendible es:

«El modelo fue validado para un dominio definido y permanece bajo un programa documentado de monitorización, verificación y revalidación que demuestra que sus prestaciones continúan dentro de los criterios de aceptación.»

En inteligencia artificial alimentaria, la deriva no es una eventualidad excepcional. Es una consecuencia previsible del paso del tiempo. Ignorarla convierte una validación inicial correcta en una seguridad progresivamente ficticia.

8. Criterios de aceptación basados en riesgo: cuándo puede aprobarse un modelo de IA alimentaria

La validación científica culmina en una decisión: determinar si el modelo puede utilizarse, para qué finalidad, bajo qué restricciones y con qué mecanismos de control.

Esta fase no puede reducirse a comprobar si una métrica supera un porcentaje convencional. No existe un valor universal de *accuracy*, sensibilidad, especificidad, MAE o AUC que convierta automáticamente un modelo en aceptable. El nivel admisible depende del contexto, de la gravedad de las consecuencias, del número de decisiones afectadas, de la posibilidad de detectar el error antes de que produzca daño y de la existencia de barreras independientes.

El NIST subraya que la tolerancia al riesgo es específica de cada organización, aplicación y contexto, y que puede estar condicionada por requisitos jurídicos y sectoriales. Su marco no prescribe un umbral general, sino que exige que la organización defina y documente qué riesgo está dispuesta a soportar.

En el ámbito alimentario, esta determinación debe estar subordinada a una regla principal:

- La empresa no puede aceptar como “riesgo algorítmico” aquello que el Derecho alimentario, el sistema APPCC o la protección de la salud no permiten aceptar como riesgo del producto.

- **Un criterio estadístico interno no puede rebajar:**

- un límite legal;
 - un criterio microbiológico;
 - una especificación de seguridad;
 - una medida de control validada;
 - una obligación de retirada;
 - una exigencia de trazabilidad;
 - una condición de uso seguro.
- Por tanto, la aceptación del modelo debe construirse desde la finalidad y el riesgo hacia las métricas, y no desde las métricas disponibles hacia una conclusión favorable.

8.1. Aceptar no significa demostrar que el modelo es perfecto

Ningún modelo predictivo funciona sin error. Incluso los métodos analíticos de referencia, los paneles sensoriales y las decisiones humanas presentan incertidumbre.

La aceptación no exige demostrar que el sistema nunca fallará. Exige demostrar que:

1. los errores previsibles han sido identificados;
2. su frecuencia está dentro de límites previamente justificados;
3. sus consecuencias están controladas;
4. existe una respuesta segura cuando el modelo no puede decidir;
5. el beneficio supera al riesgo residual;
6. el rendimiento continuará siendo vigilado.

- **Una afirmación como:**

«El modelo debe tener una precisión del 100 %»

no constituye normalmente un criterio científico. Puede expresar una aspiración, pero no un protocolo verificable con muestras finitas.

Si se observan cero errores en una validación reducida, no se demuestra que la tasa real sea cero. Lo que se demuestra es que no se detectaron errores dentro del tamaño muestral y de las condiciones evaluadas.

La decisión debe basarse en:

- • estimación puntual;
- • intervalo de confianza;
- • tamaño de muestra;
- • severidad de los errores;
- • rendimiento por subgrupos;
- • incertidumbre residual;
- • condiciones de utilización.

8.2. Los criterios deben fijarse antes de abrir el conjunto de prueba

Los límites de aceptación deben establecerse antes de conocer el resultado final.

No resulta metodológicamente válido:

1. • evaluar el modelo;
 2. • observar una sensibilidad del 93,4 %;
 3. • definir después como aceptable una sensibilidad mínima del 93 %.
- Este procedimiento adapta el criterio al resultado y elimina su función discriminatoria.

• El protocolo debería declarar previamente:

Sensibilidad mínima= S_{min}

Especificidad mínima= E_{min}

Tasa máxima de FN= FNR_{max}

MAE máximo= MAE_{max}

Sobreestimación peligrosa máxima= R_{max}

y las condiciones adicionales relativas a:

- • intervalos de confianza;
- • subgrupos;
- • calibración;
- • robustez;
- • abstención;
- • datos fuera de dominio;
- • rendimiento operativo.
- El resultado final debe compararse con esos límites sin modificarlos retrospectivamente.

8.3. Criterios apropiados, no necesariamente idénticos

El Reglamento europeo de inteligencia artificial exige, para los sistemas de alto riesgo comprendidos en su ámbito, niveles apropiados de exactitud, robustez y ciberseguridad, consistencia durante el ciclo de vida y declaración de las métricas relevantes en las instrucciones de uso. No establece un porcentaje universal de rendimiento; vincula la adecuación a la finalidad prevista y a los riesgos.

Esta lógica resulta útil como principio técnico incluso para modelos alimentarios que no sean jurídicamente clasificados como sistemas de alto riesgo.

Dos sistemas que utilizan el mismo algoritmo pueden necesitar criterios diferentes:

- • un modelo que ordena fotografías para investigación;
- • un modelo que prioriza muestras;
- • un modelo que bloquea automáticamente lotes;
- • un modelo que descarta la presencia de peligro;
- • un modelo que asigna una fecha de caducidad.
- El error estadístico puede ser idéntico, pero su significado cambia según la función.

8.4. Variables que determinan el nivel de exigencia

El nivel de evidencia requerido puede expresarse conceptualmente como:

$$E_{\text{requerida}} f(S, P, X, D, R, A, B)$$

donde:

- • (S) representa la severidad de la consecuencia;
- • (P), la probabilidad o frecuencia del error;
- • (X), la exposición o número de decisiones;
- • (D), la detectabilidad antes del daño;
- • (R), la reversibilidad;
- • (A), el grado de autonomía;
- • (B), la existencia de barreras independientes.

Severidad

No es equivalente:

- • clasificar erróneamente un defecto estético;
- • asignar una categoría comercial incorrecta;
- • subestimar una desviación nutricional;
- • prolongar indebidamente una vida útil;
- • omitir un peligro microbiológico.
- A mayor severidad, menor tolerancia al falso negativo y mayor exigencia sobre la evidencia.

Exposición

Una tasa de error del 0,1 % puede representar:

- • un error anual en una producción reducida;
- • miles de errores en una línea de gran volumen.
- **Debe calcularse el número esperado:**

$$N_{\text{error}} = N_{\text{decisiones}} \times P(\text{error})$$

El porcentaje aislado puede ocultar la dimensión real.

Detectabilidad

Un error es menos peligroso si existe una comprobación independiente antes de la liberación.

Un modelo de cribado cuyos negativos siguen sometiendo a muestreo de verificación presenta un perfil distinto de otro que sustituye por completo el análisis.

Reversibilidad

Retener preventivamente un lote puede revertirse después de un resultado confirmatorio. Liberar un producto inseguro puede no ser reversible una vez distribuido o consumido.

Autonomía

La exigencia aumenta cuando el modelo:

- informa;
- recomienda;
- prioriza;
- decide;
- ejecuta una acción;
- sustituye una barrera.

Barreras independientes

El riesgo residual será menor si la decisión está protegida por:

- análisis confirmatorio;
- revisión humana;
- límite crítico independiente;
- sistema redundante;
- bloqueo físico;
- muestreo de verificación;
- regla conservadora.

Las barreras no deben compartir el mismo modo de fallo. Dos modelos entrenados con la misma base y los mismos sensores no constituyen necesariamente una redundancia independiente.

8.5. Jerarquía funcional de modelos alimentarios

Puede establecerse una clasificación práctica en cinco niveles.

Nivel	Función	Ejemplo	Consecuencia del error	Exigencia
I	Exploración	Investigación de patrones	Sin decisión directa	Moderada
II	Apoyo informativo	Panel de tendencias	Revisión humana completa	Moderada-alta
III	Cribado o priorización	Selección de muestras	Riesgo de omitir casos	Alta
IV	Decisión automatizada de calidad	Clasificación o bloqueo	Impacto económico y comercial	Muy alta
V	Influencia directa sobre seguridad	Liberación, vida útil o control de peligro	Posible daño sanitario	Máxima

Nivel I. Exploración

El modelo se utiliza para:

- investigación;
- generación de hipótesis;
- segmentación;
- descubrimiento de variables;
- análisis retrospectivo.
- No adopta decisiones de producto.

- Puede admitirse mayor incertidumbre, siempre que no se presenten sus resultados como evidencia confirmatoria.

Nivel II. Apoyo informativo

El modelo muestra:

- • tendencias;
- • alertas orientativas;
- • puntuaciones;
- • comparaciones;
- • posibles causas.
- La decisión continúa siendo humana y se apoya en información adicional.
- Debe validarse que el sistema no induce errores sistemáticos ni genera confianza inapropiada.

Nivel III. Cribado

El modelo decide qué casos requieren análisis, revisión o intervención.

La principal amenaza es que los falsos negativos queden excluidos del control.

Debe priorizarse:

- • sensibilidad;
- • límite inferior de confianza;
- • muestreo de negativos;
- • abstención;
- • capacidad de detección fuera de dominio.

Nivel IV. Decisión automatizada de calidad

El sistema clasifica, rechaza, retiene o valora producto sin revisión ordinaria de cada caso.

Además del rendimiento, deben validarse:

- • consecuencias económicas;
- • consistencia;
- • robustez;
- • posibilidad de anulación;
- • registro;
- • recuperación ante fallos.

Nivel V. Seguridad alimentaria

El modelo influye directamente en:

- • liberación de lote;
- • asignación de caducidad;
- • control de un peligro;
- • interpretación de un límite;
- • omisión de un análisis;
- • decisión de retirada.
- La aceptación exige la evidencia más fuerte y un diseño conservador.

Si el algoritmo se integra materialmente en una medida de control, debe demostrarse que el sistema completo es capaz de alcanzar el resultado de control especificado. Esta conclusión es coherente con Codex, que define la validación

como la obtención de evidencia de que una medida o combinación de medidas controla el peligro hasta el resultado previsto.

8.6. Criterios para modelos de control de calidad

En calidad comercial o tecnológica, los errores pueden traducirse en:

- • aceptación de producto fuera de especificación;
- • rechazo de producto conforme;
- • clasificación incorrecta;
- • desviación de composición;
- • fraude no detectado;
- • conflicto con clientes.

- **Los criterios deberían distinguir entre:**

Falsa aceptación

Producto no conforme clasificado como conforme:

$$FA = [FN] / [VP + FN]$$

Puede ocasionar reclamaciones, devoluciones o incumplimientos contractuales.

Falso rechazo

Producto conforme clasificado como no conforme:

$$FR = [FP] / [VN + FP]$$

Puede provocar desperdicio, reprocesado o pérdida económica.

La aceptación debe responder a la criticidad de la especificación.

No es equivalente una desviación:

- • estética;
- • sensorial menor;
- • composicional;
- • declarativa;
- • alergénica;
- • legal.

Para defectos no críticos puede aceptarse un equilibrio económico entre falsa aceptación y falso rechazo. Para desviaciones con implicaciones legales o de alérgenos, la lógica debe aproximarse a la de seguridad.

Un protocolo podría exigir:

$$\text{Sensibilidad para defectos críticos} \geq 0,98$$

$$\text{Especificidad} \geq 0,90$$

$$IC_{95\%, \text{inferior}}(\text{sensibilidad}) \geq 0,95$$

junto con:

- • rendimiento mínimo por defecto;
- • ausencia de fallos graves repetidos;
- • estabilidad entre líneas;
- • tasa máxima de abstención;

- • revisión humana de casos fronterizos.

Estos números son ilustrativos. Deben justificarse para cada producto y decisión.

8.7. Criterios para modelos de cribado

Un modelo de cribado no pretende sustituir el método confirmatorio. Su función es reducir el universo de casos que requieren análisis o intervención.

La utilidad puede expresarse mediante la tasa de reducción:

$$RR = 1 - [N_{\text{derivados a confirmación}}] / [N_{\text{total}}]$$

Pero una elevada reducción carece de valor si se consigue omitiendo casos positivos.

El orden de aceptación debería ser:

1. • cumplir una sensibilidad mínima;
2. • demostrar un límite inferior suficientemente elevado;
3. • comprobar los negativos mediante muestreo independiente;
4. • verificar que el volumen derivado es manejable;
5. • evaluar el valor predictivo en prevalencia real;
6. • demostrar beneficio frente al proceso convencional.

• Ejemplo:

Sensibilidad $\geq 99\%$

IC_{95%}, inferior $\geq 97\%$

RR $\geq 40\%$

Abstención $\leq 10\%$

La reducción de análisis debe considerarse un criterio secundario. La detección del evento es prioritaria.

8.8. Criterios para predicción de vida útil

La vida útil requiere criterios asimétricos.

No basta con:

$$MAE \leq 3 \text{ días}$$

Debe controlarse específicamente la sobreestimación.

Una matriz de aceptación puede incluir:

$$|ME| \leq 1 \text{ día}$$

$$MAE \leq 3 \text{ días}$$

$$P_{95}(|e|) \leq 6 \text{ días}$$

$$P(e > 2 \text{ días}) \leq 5\%$$

$$\max(e) \leq E_{\text{máximo crítico}}$$

Además:

- • cobertura adecuada de intervalos predictivos;
- • validación por producto;
- • validación bajo condiciones logísticas;
- • ausencia de sesgo en extremos;
- • margen de seguridad;

- • integración con estudios experimentales.
- El modelo no debería utilizar directamente su estimación central como fecha comercial cuando existe incertidumbre relevante.
- **Puede establecerse:**

$$T_{\text{asignada}} L_{\alpha}(\hat{T}) M_{\text{adicional}}$$

donde:

- $(L_{\{\alpha\}})$ es un límite inferior predictivo;
- $(M_{\text{adicional}})$ es un margen derivado del riesgo residual.

La aceptación debe demostrar que este procedimiento produce una fecha conservadora sin generar una reducción injustificada.

8.9. Criterios para modelos microbiológicos

Un modelo microbiológico puede predecir:

- • crecimiento;
- • supervivencia;
- • inactivación;
- • probabilidad de presencia;
- • tiempo hasta alcanzar un límite.
- El criterio debe definirse según la dirección peligrosa.

Predicción de crecimiento

La infrapredicción puede ser peligrosa:

$$\hat{N}(t) < N(t)$$

Predicción de tiempo hasta alcanzar un límite

La sobrepredicción puede ser peligrosa:

$$\hat{T}_L > T_L$$

Predicción de inactivación

Sobreestimar la reducción puede ser peligroso:

$$\widehat{\log R} > \log R_{\text{real}}$$

La validación debe evaluar específicamente estos errores, no limitarse a una media simétrica.

Puede definirse una zona aceptable de predicción:

$$L_{\text{inferior}}(x) \leq e(x) \leq L_{\text{superior}}(x)$$

con un límite más estricto en la dirección no conservadora.

La FAO señala que un plan APPCC debe validarse antes de su implantación para demostrar el control consistente de los peligros en las condiciones productivas específicas, y que los modelos matemáticos pueden formar parte de la evidencia, pero deben ser aplicables a la operación y producto concretos.

Por tanto, un modelo microbiológico general no puede aprobarse para una formulación específica sin comprobar:

- • matriz;
- • pH;
- • actividad de agua;

- conservantes;
- flora;
- envase;
- temperatura;
- variabilidad del proceso.

8.10. Criterios para decisiones vinculadas a seguridad alimentaria

Cuando un modelo influye en una decisión sanitaria, los criterios deben ser jerárquicos.

Primer nivel: seguridad mínima no compensable

Determinadas condiciones no pueden compensarse con otras métricas.

Por ejemplo:

$$\text{Sensibilidad} < S_{\min} \Rightarrow \text{no aprobado}$$

aunque el modelo presente:

- AUC elevada;
- especificidad excelente;
- menor coste;
- mayor velocidad;
- mejor F1 global.

Segundo nivel: incertidumbre

Debe cumplirse:

$$IC_{95\%, \text{inferior}}(\text{sensibilidad}) \geq S_{\text{límite}}$$

No basta el valor puntual.

Tercer nivel: subgrupos

Todos los subgrupos críticos deben cumplir:

$$S_k \geq S_{k, \min}$$

No debe permitirse que el promedio oculte una familia de productos insegura.

Cuarto nivel: robustez

El rendimiento debe mantenerse frente a:

- variabilidad prevista;
- perturbaciones;
- condiciones límite;
- fallos parciales;
- cambios razonables.

Quinto nivel: operación segura

Debe existir:

- abstención;
- revisión;
- sistema alternativo;

- registro;
 - parada segura;
 - monitorización;
 - revalidación.
- La exactitud estadística no basta si el sistema carece de una respuesta segura ante fallo.

8.11. El principio de no compensación

En muchos procesos de selección de modelos se construye una puntuación total:

$$\text{Score} = w_1M_1 + w_2M_2 + \dots + w_nM_n$$

Este método puede ser útil para comparar alternativas, pero resulta peligroso si permite que una métrica favorable compense un incumplimiento crítico.

Ejemplo:

Métrica	Modelo A	Modelo B
Sensibilidad	92 %	98 %
Especificidad	99 %	90 %
AUC	0,98	0,96
Coste	Bajo	Medio

- Una puntuación ponderada podría seleccionar el modelo A, pese a que omite más casos positivos.
- **En seguridad debe aplicarse una estructura de restricciones:**

$$\text{Elegible} \Leftrightarrow S \geq S_{\min} \wedge \text{FNR} \leq \text{FNR}_{\max} \wedge R_{\text{crítico}} \leq R_{\max} \wedge C_{\text{robustez}} = \text{cumple} \wedge C_{\text{dominio}} = \text{cumple}$$

Solo entre modelos elegibles procede optimizar especificidad, coste o velocidad.

8.12. Error medio frente a error catastrófico

Las métricas promedio pueden ocultar fallos poco frecuentes pero graves.

Debe distinguirse entre:

E_{promedio}

y:

$E_{\text{catastrófico}}$

Un modelo puede presentar:

$$\text{MAE} = 1,5 \text{ días}$$

pero sobreestimar un lote en 18 días.

En una aplicación crítica, la aceptación debe limitar:

- frecuencia de errores extremos;
- magnitud máxima;
- gravedad;
- concentración en subgrupos;
- circunstancias causales.

- **Puede establecerse:**

$$P(|e| > E_{\text{grave}}) \leq \alpha$$

y:

$$P(e > E_{\text{peligroso}}) \leq \beta$$

Los casos extremos deben revisarse individualmente, incluso cuando el modelo cumpla las métricas globales.

8.13. Intervalos de confianza y regla de aprobación

Una regla científicamente sólida debería utilizar el límite desfavorable del intervalo.

Para sensibilidad:

$$LCB_{95\%}(S) \geq S_{\text{min}}$$

Para tasa de error:

$$UCB_{95\%}(E) \leq E_{\text{max}}$$

donde:

- • (LCB) es el límite inferior;
 - • (UCB), el límite superior.
 - Esto introduce una protección frente a la incertidumbre muestral.
- **Si:**

$$\hat{S} = 98\%$$

pero:

$$IC_{95\%} = 89\% - 100\%$$

el estudio no demuestra con precisión suficiente que la sensibilidad real supere el 95 %.

La solución no consiste en ignorar el intervalo, sino en:

- • aumentar el tamaño muestral;
- • acumular más positivos;
- • reducir el dominio;
- • aceptar un uso menos crítico;
- • mantener confirmación independiente.

8.14. Tamaño muestral orientado al riesgo

El tamaño debe determinarse a partir de la precisión requerida.

Si se desea estimar una sensibilidad esperada (p) con un margen aproximado (d), puede utilizarse:

$$n \approx [z_{1-\alpha/2}^2 p(1-p)] / [d^2]$$

Este cálculo debe ajustarse cuando:

- • el evento es raro;
- • existen conglomerados;
- • se esperan pérdidas;
- • hay múltiples subgrupos;
- • se exige demostrar no inferioridad;
- • se desea limitar una tasa de error muy baja.
- El número relevante para sensibilidad es el de positivos reales, no el tamaño total.

- Un estudio con 100.000 negativos y 20 positivos puede estimar muy bien la especificidad, pero no la sensibilidad.

8.15. Prevalencia y escenarios de utilización

Los valores predictivos deben calcularse para la prevalencia real o para varios escenarios plausibles.

Puede utilizarse:

$$VPP(\pi) [S\pi] / [S\pi + (1-E)(1-\pi)]$$

donde (π) es la prevalencia.

La aceptación debería analizar:

- • prevalencia histórica;
- • escenario bajo;
- • escenario habitual;
- • escenario de incidente;
- • diferencias por producto o planta.
- Un modelo puede ser viable en una población de alto riesgo y generar demasiadas falsas alarmas en una población de bajo riesgo.

8.16. Criterios por subgrupos y regla del peor caso

El rendimiento global no debe ser el único criterio.

Para cada subgrupo crítico (k):

$$M_k \geq M_{k,min}$$

Cuando los tamaños son insuficientes, la conclusión correcta es:

- evidencia insuficiente para aprobar este subgrupo.
- **No debe sustituirse por:**
 - no se observaron diferencias.
- **En determinados contextos puede aplicarse una regla del peor caso:**

$$M_{aprobación} \min(M_1, M_2, \dots, M_K)$$

Esta regla evita que la población mayoritaria diluya fallos en referencias minoritarias.

8.17. Abstención como requisito de aceptación

El sistema debe poder rechazar una predicción cuando:

- • la entrada está fuera de dominio;
- • la calidad es insuficiente;
- • faltan variables;
- • la incertidumbre es elevada;
- • la salida está próxima al umbral;
- • existe inconsistencia.
- **Los criterios deben definir:**

$$A \leq A_{max}$$

donde (A) es la tasa de abstención.

Y simultáneamente:

$$E_{\text{aceptadas}} \leq E_{\text{max}}$$

No debe premiarse a un modelo por reducir errores absteniéndose en casi todos los casos.

También debe comprobarse que las abstenciones se concentran en situaciones realmente difíciles, no de forma desigual o arbitraria.

8.18. Funcionamiento seguro ante fallos

La aceptación debe incluir requisitos no estadísticos.

Ante un fallo, el sistema debería:

- • no liberar automáticamente;
- • conservar la última situación segura;
- • transferir la decisión a una persona;
- • activar el procedimiento convencional;
- • registrar el incidente;
- • impedir datos incompletos;
- • evitar decisiones silenciosas.

El Reglamento de IA contempla, para los sistemas de alto riesgo, resiliencia frente a errores e inconsistencias y la utilización de medidas técnicas y organizativas, incluidas soluciones redundantes o planes de respaldo y funcionamiento seguro.

En alimentación, esto puede traducirse en:

Fallo del modelo $\Rightarrow$ retención o procedimiento alternativo

y nunca en:

Fallo del modelo $\Rightarrow$ liberación por defecto

cuando existe una implicación sanitaria.

8.19. Comparación con el estándar de práctica

La aceptación debe demostrar que el modelo añade valor frente al procedimiento existente.

La comparación puede ser:

$$\Delta M_{\text{M_IA}} M_{\text{referencia}}$$

La referencia puede consistir en:

- • regla fija;
- • modelo estadístico tradicional;
- • experto;
- • inspección humana;
- • procedimiento de muestreo;
- • modelo microbiológico establecido.

• Debe comprobarse:

- • superioridad;
- • no inferioridad;
- • reducción de tiempo;
- • reducción de coste;
- • mejora de anticipación;

- • disminución de desperdicio;
- • aumento de cobertura.
- Un modelo más complejo no debe aprobarse si ofrece una mejora estadísticamente detectable pero industrialmente irrelevante.

8.20. Beneficio neto y utilidad operativa

La utilidad puede conceptualizarse como:

$$U = B_{VP} - C_{FP} - C_{FN} - C_{operación} - C_{riesgo\ residual}$$

No todos estos términos son monetizables.

El análisis debe considerar:

- • producto salvado;
- • análisis evitados;
- • tiempo ganado;
- • lotes retenidos;
- • coste de confirmación;
- • desperdicio;
- • impacto sanitario;
- • retirada;
- • reputación;
- • exposición jurídica.
- El beneficio económico nunca debe justificar una tasa de error incompatible con seguridad o legalidad.

8.21. Aprobación escalonada

La aceptación no tiene que ser binaria.

Puede establecerse:

No aprobado

No cumple criterios críticos.

Aprobado para investigación

Puede utilizarse sin influir en producto comercial.

Aprobado en modo sombra

Opera prospectivamente, pero no adopta decisiones.

Aprobado como apoyo

La salida debe ser revisada por personal competente.

Aprobado con confirmación obligatoria

Toda decisión requiere método independiente.

Aprobado para dominio restringido

Solo ciertas plantas, productos, equipos o rangos.

Aprobado para automatización limitada

Solo decisiones de bajo riesgo o alta confianza.

Aprobado para uso previsto completo

Cumple todos los criterios y controles.

Esta estructura permite utilizar la evidencia disponible sin extenderla más allá de lo demostrado.

8.22. Aprobación condicional

Puede concederse una aprobación limitada cuando:

- la métrica principal cumple;
- una métrica secundaria queda próxima al límite;
- existe una barrera adicional;
- se requiere ampliar muestra;
- el dominio puede restringirse.

• Ejemplo:

Aprobado durante seis meses exclusivamente como sistema de cribado, con confirmación microbiológica de todos los positivos y muestreo aleatorio del 10 % de los negativos. La continuidad queda condicionada a mantener sensibilidad superior al límite y completar 300 casos positivos verificados.

La aprobación condicional debe especificar:

- duración;
- dominio;
- barreras;
- datos a recoger;
- responsable;
- condición de retirada;
- revisión final.

8.23. Criterios de suspensión

La autorización debe definir de antemano cuándo dejará de utilizarse el sistema.

Entre los criterios:

- falso negativo grave;
- sensibilidad bajo mínimo;
- pérdida de calibración;
- aumento de datos fuera de dominio;
- cambio no validado;
- fallo de integridad;
- imposibilidad de obtener etiquetas;
- deriva conceptual;
- tasa de abstención excesiva;
- caída del sistema de respaldo.

• Puede establecerse:

$M_t < M_{\text{suspensión}} \Rightarrow \text{suspender}$

La suspensión no debe depender de una decisión improvisada cuando ya existe un incidente.

8.24. Matriz maestra de aceptación

Una matriz completa puede estructurarse así:

Dimensión	Métrica o prueba	Criterio	Prioridad	Resultado
Finalidad	Coincidencia con uso previsto	Completa	Crítica	—
Datos	Representatividad	Cumple matriz de cobertura	Crítica	—
Independencia	Test externo	Sin fuga	Crítica	—
Detección	Sensibilidad	$\geq$ límite	Crítica	—
Seguridad	Falsos negativos	$\leq$ límite	Crítica	—
Operación	Especificidad	$\geq$ límite viable	Alta	—
Incertidumbre	IC	Límite desfavorable cumple	Crítica	—
Calibración	Intercepto y pendiente	Dentro de rango	Alta	—
Robustez	Pruebas de estrés	Sin degradación crítica	Crítica	—
Dominio	OOD	Detectado correctamente	Crítica	—
Abstención	Cobertura	$\geq$ mínimo	Alta	—
Subgrupos	Rendimiento mínimo	Todos cumplen	Crítica	—
Impacto	Mejora frente a referencia	Relevante	Media/alta	—
Operación	Latencia y disponibilidad	Cumple	Alta	—
Humano	Supervisión	Efectiva	Alta	—
Deriva	Plan de monitorización	Aprobado	Crítica	—
Respaldo	Modo seguro	Verificado	Crítica	—

Las dimensiones críticas deben cumplir todas. Las no críticas pueden admitir acciones correctivas, siempre que no alteren la seguridad.

8.25. Ejemplo de aceptación: detector visual de cuerpos extraños

Finalidad:

- Detectar cuerpos extraños visibles en producto envasado y retirar automáticamente la unidad.

• Criterios posibles:

- sensibilidad global mínima;
- sensibilidad específica por tipo de cuerpo extraño;
- detección por tamaño;
- desempeño con reflejos y condensación;
- falsos rechazos por mil unidades;
- latencia compatible con la línea;
- verificación del expulsor;
- parada ante fallo de cámara;
- registro de imágenes;

- • muestreo de unidades no rechazadas.
- No bastaría una accuracy elevada si el sistema falla con fragmentos transparentes o pequeños.

8.26. Ejemplo de aceptación: vida útil refrigerada

Finalidad:

- Estimar vida útil remanente a partir de formulación, proceso e historial térmico.

• Criterios posibles:

$$MAE \leq 3 \text{ días}$$

$$P(e > 2) \leq 5\%$$

$$P_{95}(|e|) \leq 6 \text{ días}$$

Además:

- • validación en varias campañas;
- • abuso térmico razonable;
- • cobertura de intervalos;
- • margen conservador;
- • prohibición de extrapolación;
- • estudio experimental complementario;
- • revisión ante cambio de formulación.

8.27. Ejemplo de aceptación: cribado microbiológico

Finalidad:

- Priorizar lotes para análisis de confirmación.

• Criterios:

- • sensibilidad muy elevada;
- • límite inferior predefinido;
- • tasa máxima de negativos omitidos;
- • VPP en prevalencia real;
- • volumen de confirmaciones viable;
- • muestreo aleatorio de negativos;
- • prohibición de liberación basada exclusivamente en la predicción;
- • revisión de deriva mensual.

8.28. Ejemplo de aceptación: clasificación sensorial

Finalidad:

- Predecir rechazo sensorial de un panel entrenado.

• Criterios:

- • concordancia con panel;
- • reproducibilidad;
- • error por atributo;
- • estabilidad entre sesiones;
- • análisis de desacuerdo;

- ausencia de sesgo hacia productos mayoritarios;
- rendimiento frente a variabilidad del panel.

La aceptación debe reconocer que el patrón sensorial contiene incertidumbre y que el modelo no puede demostrar una precisión superior a una referencia inestable sin un análisis específico.

8.29. Criterios inválidos frecuentes

No deberían aceptarse criterios como:

«Accuracy superior al 90 %.»

Sin definir:

- clases;
- prevalencia;
- coste de errores;
- subgrupos;
- intervalo de confianza.

• Tampoco:

«AUC superior a 0,80.»

El AUC no establece el rendimiento en el umbral operativo.

Ni:

«Error inferior al publicado por otros autores.»

Las poblaciones y condiciones pueden no ser comparables.

Ni:

«El modelo funciona mejor que un humano.»

Debe definirse qué humanos, con qué experiencia, en qué muestra y frente a qué referencia.

Ni:

«No se observaron falsos negativos.»

Sin tamaño muestral e intervalo.

Ni:

«Las variables importantes son científicamente razonables.»

La plausibilidad no prueba exactitud.

Ni:

«El proveedor certifica una precisión del 99 %.»

Debe verificarse en el entorno específico de la empresa.

8.30. Informe de decisión

El informe final debería declarar:

Sistema

- nombre;

- versión;
- configuración;
- fecha.

Uso autorizado

- productos;
- plantas;
- equipos;
- condiciones;
- usuarios;
- grado de autonomía.

Evidencia

- muestras;
- periodo;
- centros;
- referencia;
- métricas;
- intervalos;
- subgrupos;
- robustez.

Limitaciones

- exclusiones;
- rangos;
- condiciones no evaluadas;
- incertidumbre;
- fallos conocidos.

Controles obligatorios

- revisión;
- confirmación;
- abstención;
- respaldo;
- monitorización;
- muestreo.

Decisión

- aprobado;
- restringido;
- condicional;

- no aprobado.

Responsables

- validación;
- aprobación;
- operación;
- vigilancia;
- revalidación.

El NIST plantea la gestión del riesgo como una actividad continua durante todo el ciclo de vida y recomienda integrar perspectivas multidisciplinarias en las funciones de gobierno, mapeo, medición y gestión. Esta aproximación resulta especialmente necesaria en alimentos, donde una decisión algorítmica puede involucrar simultáneamente estadística, microbiología, tecnología, metrología, producción, legislación y comportamiento humano.

8.31. Conclusión del bloque

Un modelo no debe aprobarse porque sus métricas resulten impresionantes, sino porque se ha demostrado que sus errores residuales son compatibles con la finalidad prevista y permanecen controlados por el sistema completo.

Los criterios científicamente sólidos deben ser:

- • prospectivos;
- • específicos;
- • cuantificables;
- • basados en riesgo;
- • sensibles a la dirección del error;
- • aplicables por subgrupos;
- • acompañados de incertidumbre;
- • vinculados a acciones;
- • sometidos a vigilancia.

• La decisión de aceptación debe responder a cinco preguntas:

1. • ¿Qué ocurre cuando el modelo acierta?
 2. • ¿Qué ocurre cuando se equivoca?
 3. • ¿Quién o qué detecta el error?
 4. • ¿Qué barrera impide que el error se convierta en daño?
 5. • ¿Qué evidencia demuestra que esa protección funcionará de forma continuada?
- Si estas preguntas no pueden responderse, el sistema todavía no está validado para la decisión pretendida, aunque presente una métrica elevada.

La validación científica no convierte la incertidumbre en certeza. La identifica, la mide, la limita y establece hasta dónde puede permitirse que el algoritmo intervenga.

9. Integración del modelo validado en APPCC, prerequisites y sistema de gestión de la seguridad alimentaria

La validación estadística no autoriza por sí sola el uso industrial de un modelo de inteligencia artificial. Después de demostrar su rendimiento, el sistema debe integrarse en la estructura real mediante la cual la empresa controla la calidad, la legalidad y la seguridad de sus productos.

Esta integración exige responder a una cuestión fundamental:

- ¿Qué función desempeña exactamente la salida del modelo dentro del sistema de gestión alimentaria y qué ocurre cuando esa salida es incorrecta, incompleta o no está disponible?

- **La IA puede utilizarse como:**

- • instrumento de observación;
- • sistema de alerta;
- • herramienta de priorización;
- • apoyo a una decisión;
- • mecanismo de monitorización;
- • componente de una medida de control;
- • sistema de clasificación o bloqueo automático.

Estas funciones no son equivalentes. Cada una modifica de forma distinta el análisis de peligros, la documentación, las responsabilidades, los registros y las actuaciones necesarias ante un fallo.

El Reglamento (CE) n.º 853/2004 exige que los operadores alimentarios establezcan, apliquen y mantengan procedimientos permanentes basados en los principios APPCC. La Comunicación de la Comisión de 2022 desarrolla la relación entre buenas prácticas de higiene, programas de prerequisites, prerequisites operativos, procedimientos APPCC y otros elementos del sistema de gestión de la seguridad alimentaria. La incorporación de IA debe realizarse dentro de esa arquitectura y no como una capa tecnológica separada de ella.

9.1. La IA no sustituye el sistema de gestión

Un error habitual consiste en presentar el modelo como sustituto del APPCC, del laboratorio, de la inspección o de la experiencia del tecnólogo.

Esta concepción es incorrecta. El modelo constituye un componente adicional del sistema. Su introducción puede:

- • mejorar la capacidad de detección;
- • anticipar desviaciones;
- • integrar variables complejas;
- • reducir tiempos;
- • priorizar recursos;
- • aumentar la consistencia.

- **Pero también puede introducir nuevos modos de fallo:**

- • datos incorrectos;
- • pérdida de representatividad;
- • falsa confianza;
- • deriva;
- • error de umbral;
- • dependencia tecnológica;

- • indisponibilidad;
- • automatización inadecuada;
- • interpretación errónea.

• **La organización debe analizar simultáneamente:**

riesgo alimentario controlado por la IA

y:

riesgo introducido por la propia IA

El modelo no debe quedar fuera del análisis de peligros por considerarse una herramienta informática. Si su salida influye en una medida preventiva, una retención, una inspección o una liberación, su funcionamiento puede afectar materialmente a la seguridad del alimento.

9.2. Determinación de la función dentro del FSMS

Antes del despliegue debe clasificarse la función del modelo dentro del sistema de gestión de la seguridad alimentaria o FSMS.

Una matriz inicial puede diferenciar las siguientes posiciones.

Función del modelo	Ejemplo	Consecuencia del fallo	Integración requerida
Analítica exploratoria	Identificación de patrones históricos	Limitada, sin decisión directa	Gobierno de datos y revisión técnica
Alerta temprana	Tendencia de temperatura o higiene	Retraso en investigación	Procedimiento de respuesta
Priorización	Selección de muestras	Omisión de casos relevantes	Muestreo independiente de negativos
Monitorización	Vigilancia de variable de proceso	Pérdida de desviación	Respaldo y límites de alarma
Clasificación de calidad	Rechazo automático	Falso rechazo o aceptación	Verificación de expulsión y auditoría
Estimación de vida útil	Asignación de días	Sobreestimación del periodo seguro	Evidencia experimental y margen
Control de peligro	Influencia sobre PCC o medida de control	Liberación de producto inseguro	Máximo nivel de validación y redundancia

La clasificación debe basarse en la función efectiva, no en la denominación comercial del software.

Un proveedor puede describir su solución como «asistente», pero si la empresa acepta sistemáticamente su salida sin una revisión real, el modelo funciona de facto como un decisor.

9.3. Integración en programas de prerrequisitos

Muchos modelos de IA se incorporarán inicialmente a programas de prerrequisitos, por ejemplo:

- • limpieza y desinfección;
- • control de plagas;
- • mantenimiento;
- • calibración;
- • control de proveedores;
- • higiene del personal;
- • control ambiental;
- • almacenamiento;

- trazabilidad;
- gestión de alérgenos.

Ejemplo: predicción de riesgo de contaminación ambiental

Un algoritmo puede combinar:

- resultados de superficies;
 - historial de positivos;
 - temperatura;
 - humedad;
 - zonas higiénicas;
 - rutas de personal;
 - frecuencia de limpieza;
 - incidencias de mantenimiento.
- La salida puede utilizarse para intensificar el muestreo en determinadas zonas.

En este caso, el modelo no sustituye el programa ambiental. Lo modifica de manera dinámica.

La documentación deberá especificar:

- qué datos utiliza;
 - cómo se priorizan las zonas;
 - qué porcentaje permanece bajo muestreo aleatorio;
 - qué ocurre si el modelo no está disponible;
 - qué resultados activan investigación;
 - cómo se comprueba que no omite ubicaciones relevantes.
- Un modelo de priorización que elimina por completo el muestreo aleatorio corre el riesgo de dejar de observar las áreas que clasifica erróneamente como de bajo riesgo.

9.4. Integración en prerrequisitos operativos

Cuando el modelo controla o anticipa una condición directamente vinculada a un peligro significativo, puede actuar como parte de un prerrequisito operativo.

Ejemplos:

- predicción de riesgo de crecimiento durante almacenamiento;
 - detección de desvíos en concentración de conservante;
 - alerta por combinaciones inadecuadas de pH y actividad de agua;
 - identificación de patrones de limpieza insuficiente;
 - vigilancia de integridad del cierre;
 - predicción de contaminación cruzada.
- **En esta situación deben definirse:**
 - criterio operativo;
 - límite de alerta;
 - acción;
 - responsable;
 - frecuencia;
 - método de verificación;

- registro;
- respuesta ante indisponibilidad.
- **La salida del modelo no debe quedar expresada mediante instrucciones vagas como:**

«Riesgo elevado: revisar.»

Debe traducirse en una acción inequívoca:

«Cuando la probabilidad calibrada supere 0,35 o el sistema indique condición fuera de dominio, retener el lote, verificar pH y actividad de agua mediante los métodos autorizados y remitir la decisión al responsable de calidad.»

9.5. Integración en puntos de control crítico

La incorporación de IA a un PCC requiere una prudencia considerablemente mayor.

Un PCC está asociado a un peligro significativo cuyo control resulta esencial. El límite crítico separa la aceptabilidad de la inaceptabilidad y debe ser científicamente válido.

Un modelo puede intervenir en un PCC de varias maneras:

- vigilar una variable medida;
- detectar patrones de desviación;
- anticipar la pérdida de control;
- comprobar la respuesta del proceso;
- ayudar a interpretar múltiples señales.
- Sin embargo, no debe asumirse automáticamente que una probabilidad algorítmica puede sustituir a un límite crítico físico o microbiológico.

Ejemplo

Un tratamiento térmico puede tener como límite crítico una combinación validada de tiempo y temperatura.

Un modelo puede predecir la probabilidad de que el tratamiento resulte insuficiente considerando:

- temperatura;
- tiempo;
- caudal;
- carga;
- viscosidad;
- presión;
- distribución térmica.
- Pero si el modelo deja de funcionar, el proceso debe conservar un mecanismo de control basado en parámetros validados.

- **La arquitectura segura sería:**

límite crítico validado + monitorización instrumental + IA predictiva complementaria

y no:

IA única barrera

salvo que exista evidencia extraordinariamente sólida de equivalencia, control de fallos, redundancia, trazabilidad y cumplimiento regulatorio.

9.6. ¿Puede una salida algorítmica ser un límite crítico?

En términos conceptuales, podría imaginarse un límite como:

$$P(\text{fallo}) < 0,01$$

Pero convertirlo en límite crítico plantea problemas:

- • dependencia de la calibración;
- • variación con la prevalencia;
- • complejidad de interpretación;
- • incertidumbre;
- • deriva;
- • dificultad de verificación directa;
- • dependencia de múltiples entradas;
- • posibilidad de cambio de versión.

Un límite crítico debe ser claro, aplicable y susceptible de monitorización efectiva. Codex diferencia la validación de las medidas de control, la monitorización y la verificación del sistema, y exige que la organización disponga de evidencia de que las medidas alcanzan el resultado de control previsto.

En la mayoría de las aplicaciones actuales, resulta más defendible utilizar la IA como:

- • alarma predictiva previa al límite;
- • sistema complementario;
- • verificación reforzada;
- • herramienta de investigación.

9.7. Modificación del análisis de peligros

La implantación del modelo debe provocar una revisión formal del análisis de peligros.

El análisis debe incluir dos categorías.

Peligros o desviaciones que el modelo pretende controlar

Por ejemplo:

- • crecimiento microbiano;
- • contaminación;
- • cuerpos extraños;
- • deterioro;
- • incumplimiento composicional;
- • fraude;
- • fallo de proceso.

Fallos propios del sistema de IA

Entre ellos:

- • falso negativo;
- • falso positivo;
- • dato ausente;
- • dato erróneo;
- • uso fuera de dominio;

- modelo desactualizado;
 - deriva;
 - fallo de comunicación;
 - fallo de sensor;
 - umbral incorrecto;
 - versión no autorizada;
 - interpretación humana errónea;
 - manipulación.
- **Puede aplicarse una matriz de riesgo específica:**

$$R_{IA} = S \times P \times D$$

donde:

- (S) es la severidad;
- (P), la probabilidad;
- (D), la dificultad de detección.

La valoración no debe limitarse al error histórico. Debe considerar modos de fallo razonablemente previsibles.

9.8. Análisis de modos y efectos de fallo

Una metodología tipo FMEA puede estructurar el análisis.

Modo de fallo	Causa	Efecto	Detección	Control
Falso negativo	Umbral inadecuado	Lote inseguro no retenido	Muestreo de negativos	Sensibilidad mínima y respaldo
Falso positivo	Deriva del sensor	Retención innecesaria	Confirmación	Recalibración
Datos fuera de dominio	Nueva formulación	Predicción no válida	Detector OOD	Abstención
Pérdida de conexión	Fallo de red	Sin predicción	Alarma técnica	Procedimiento manual
Versión incorrecta	Despliegue no controlado	Rendimiento desconocido	Verificación de versión	Control de cambios
Entrada manual errónea	Unidad incorrecta	Resultado incoherente	Reglas de plausibilidad	Validación de datos
Deriva conceptual	Cambio de proceso	Degradación silenciosa	Verificación contra referencia	Revalidación
Automation bias	Confianza excesiva	Aceptación de salida incorrecta	Auditoría de decisiones	Formación y supervisión

El análisis debe revisarse cuando aparezcan incidentes, cambios o nuevos datos.

9.9. Validación, monitorización y verificación

Estos conceptos deben permanecer claramente diferenciados.

Validación

Demuestra antes del uso que el sistema puede alcanzar el rendimiento y el resultado de control especificados.

Pregunta:

¿Puede funcionar adecuadamente?

Monitorización

Observa durante la operación si las variables y salidas se mantienen dentro de los límites establecidos.

Pregunta:

¿Está funcionando ahora dentro de las condiciones autorizadas?

Verificación

Comprueba mediante métodos adicionales que el sistema se aplica correctamente y sigue siendo eficaz.

Pregunta:

- ¿Existe evidencia independiente de que continúa funcionando como estaba previsto?
- Esta distinción coincide con la estructura de Codex y con la lógica de los sistemas APPCC.

9.10. Plan de monitorización del modelo

El plan debe especificar:

- variable monitorizada;
- fuente;
- frecuencia;
- responsable;
- límite;
- acción;
- registro;
- escalado.

• Ejemplo:

Elemento	Frecuencia	Límite	Acción
Datos ausentes	Continua	>1 %	Bloquear predicción
Entradas fuera de dominio	Continua	Cero en decisión automática	Abstención
Tasa de alertas	Semanal	Fuera de banda histórica	Investigar
Falsos negativos	Según confirmación	Por debajo del límite aprobado	Suspender
Sensibilidad acumulada	Trimestral	≥ criterio	Revalidar si incumple
Calibración	Trimestral/semestral	Dentro de rango	Recalibrar
Deriva de entradas	Semanal/mensual	Umbral validado	Evaluación técnica
Tasa de abstención	Mensual	≤ máximo	Revisar dominio
Disponibilidad	Mensual	≥ objetivo	Activar plan técnico

- La frecuencia debe ser coherente con la velocidad a la que puede aparecer el riesgo.

9.11. Verificación independiente

La verificación no debe depender exclusivamente de las señales del propio algoritmo.

Debe incluir:

- análisis confirmatorios;
- muestreo aleatorio de negativos;
- revisión de casos discordantes;

- ensayos ciegos;
- controles positivos y negativos;
- comparación con método de referencia;
- auditoría de configuración;
- revisión de registros;
- verificación del dispositivo físico asociado.

En un sistema visual de rechazo, no basta con confirmar que el software identificó el defecto. Debe comprobarse que:

1. • la cámara lo detectó;
 2. • la señal llegó al controlador;
 3. • el expulsor actuó;
 4. • la unidad correcta fue retirada;
 5. • el rechazo quedó registrado.
- La validación del algoritmo y la verificación del mecanismo físico son elementos distintos.

9.12. Procedimiento de trabajo normalizado

El uso del modelo debe estar regulado por un procedimiento aprobado.

Un PNT o SOP debería incluir:

Objeto

Describir la finalidad del sistema.

Alcance

Definir:

- productos;
- plantas;
- líneas;
- equipos;
- usuarios;
- decisiones autorizadas.

Definiciones

Aclarar:

- predicción;
- probabilidad;
- alerta;
- abstención;
- fuera de dominio;
- falso negativo;
- deriva.

Responsabilidades

Asignar:

- operación;

- • revisión;
- • mantenimiento;
- • aprobación;
- • monitorización;
- • gestión de incidentes.

Procedimiento

Detallar:

1. • adquisición de datos;
2. • controles de calidad;
3. • ejecución;
4. • interpretación;
5. • acciones;
6. • confirmación;
7. • registro;
8. • escalado.

Fallos

Definir el comportamiento ante:

- • indisponibilidad;
- • datos incompletos;
- • resultado indeterminado;
- • alerta crítica;
- • discrepancia;
- • versión incorrecta.

Registros

Especificar qué debe conservarse.

Cambios

Establecer el procedimiento de modificación y revalidación.

9.13. Dossier maestro del modelo

Cada sistema debería disponer de un expediente técnico único y controlado.

El dossier maestro debe contener, al menos:

1. • ficha de identificación;
2. • finalidad prevista;
3. • propietario y responsables;
4. • descripción de arquitectura;
5. • variables de entrada;
6. • fuentes de datos;
7. • preprocesamiento;
8. • patrón de referencia;
9. • dominio de aplicación;

10. • protocolo de validación;
11. • informe de resultados;
12. • criterios de aceptación;
13. • limitaciones;
14. • análisis de riesgos;
15. • procedimiento de uso;
16. • plan de monitorización;
17. • plan de verificación;
18. • plan de revalidación;
19. • gestión de cambios;
20. • gestión de incidentes;
21. • requisitos de seguridad;
22. • formación;
23. • historial de versiones.

El NIST AI RMF vincula la gestión fiable de IA con una documentación sistemática, responsabilidades claras, pruebas, vigilancia continua y decisiones formales de autorización o no autorización.

9.14. Ficha de identificación

La ficha debe permitir identificar inequívocamente el sistema:

- • nombre;
 - • código interno;
 - • versión;
 - • fecha de aprobación;
 - • propietario;
 - • proveedor;
 - • ubicación;
 - • software;
 - • hardware;
 - • dependencias;
 - • datos utilizados;
 - • umbral;
 - • finalidad;
 - • nivel de autonomía;
 - • estado de validación.
- No debe utilizarse una denominación genérica como «modelo de vida útil» cuando existen varias configuraciones.

9.15. Descripción de datos

La documentación de datos debe incluir:

- • origen;
- • periodo;
- • plantas;
- • productos;

- • lotes;
 - • proveedores;
 - • criterios de inclusión;
 - • exclusiones;
 - • frecuencia de positivos;
 - • datos faltantes;
 - • transformaciones;
 - • calidad;
 - • limitaciones;
 - • derechos de utilización.
- Debe identificarse la unidad independiente y evitar que el número de imágenes, espectros o lecturas se presente engañosamente como número de muestras.

9.16. Trazabilidad de la predicción

Para cada decisión relevante debe poder reconstruirse:

entrada → procesamiento → modelo → salida → decisión → resultado real

El registro debería incluir:

- • identificación del lote;
- • fecha y hora;
- • datos de entrada;
- • equipo;
- • versión;
- • puntuación;
- • probabilidad;
- • incertidumbre;
- • umbral;
- • condición de dominio;
- • decisión automática;
- • revisión humana;
- • acción;
- • resultado confirmatorio.

La trazabilidad general de alimentos y piensos sigue siendo responsabilidad del operador conforme al Reglamento (CE) n.º 178/2002; el uso de IA no sustituye la identificación de proveedores, destinatarios y movimientos del producto.

9.17. Integridad del registro

Un registro solo tiene valor probatorio si es:

- • completo;
- • legible;
- • atribuible;
- • contemporáneo;
- • original o fiel;

- • exacto;
- • protegido contra alteración;
- • recuperable.
- **Debe controlarse:**
 - • quién accede;
 - • quién modifica;
 - • qué cambios se realizan;
 - • cuándo;
 - • por qué;
 - • bajo qué autorización.
- Cuando el sistema influye en una decisión crítica, no debería permitirse sobrescribir silenciosamente una predicción anterior.

9.18. Control de versiones

Cada modificación debe generar una identificación de versión.

Pueden cambiar:

- • código;
- • pesos;
- • variables;
- • transformación;
- • librería;
- • sensor;
- • resolución;
- • umbral;
- • interfaz;
- • reglas de negocio;
- • método de calibración.
- La versión utilizada en una predicción debe quedar registrada.
- **Un esquema podría ser:**

$V_{\text{modelo}} = (\text{datos}, \text{código}, \text{pesos}, \text{umbral}, \text{entorno})$

No basta con conservar únicamente el nombre comercial del sistema.

9.19. Gestión de cambios

Todo cambio debe evaluarse antes del despliegue.

El procedimiento debe contener:

1. • solicitud;
2. • descripción;
3. • justificación;
4. • análisis de impacto;
5. • pruebas;
6. • clasificación;
7. • revalidación;

8. • aprobación;
9. • despliegue;
10. • verificación posterior;
11. • plan de reversión.

Clasificación posible

- menor;
- moderado;
- mayor;
- crítico.

Cambio menor

No afecta a la predicción ni al uso previsto.

Ejemplo:

- corrección de texto en interfaz.

Cambio moderado

Puede afectar a utilización o presentación.

Ejemplo:

- modificación del formato de entrada.

Cambio mayor

Afecta al rendimiento o dominio.

Ejemplo:

- • nuevo producto;
- • nueva cámara;
- • recalibración;
- • nuevo umbral.

Cambio crítico

Afecta a finalidad, arquitectura o seguridad.

Ejemplo:

- • reentrenamiento;
- • nuevo método de referencia;
- • automatización de una decisión antes humana.

9.20. Modificaciones no autorizadas

Deben prevenirse:

- • actualización automática del proveedor;
- • cambio de umbral por un usuario;
- • uso de datos no aprobados;
- • sustitución de sensores;
- • modificación de unidades;
- • incorporación de productos no validados;

- • aprendizaje automático no controlado.
- Una actualización técnicamente beneficiosa puede invalidar la evidencia anterior.

9.21. Gestión de proveedores

Cuando el modelo procede de un tercero, la empresa alimentaria no puede delegar por completo su responsabilidad.

Debe obtener información suficiente sobre:

- • finalidad prevista;
 - • arquitectura;
 - • entrenamiento;
 - • validación;
 - • limitaciones;
 - • actualización;
 - • seguridad;
 - • registros;
 - • soporte;
 - • incidentes;
 - • continuidad.
- **El proveedor debería facilitar:**
 - • métricas completas;
 - • composición de los conjuntos;
 - • prevalencia;
 - • subgrupos;
 - • intervalos de confianza;
 - • condiciones de prueba;
 - • versiones;
 - • historial de cambios;
 - • instrucciones;
 - • restricciones.

- **La afirmación comercial:**

«Precisión del 99 %»

carece de valor suficiente sin conocer:

- • qué se midió;
- • sobre qué población;
- • con qué prevalencia;
- • qué error se produjo;
- • si existía independencia;
- • qué intervalo acompaña al resultado.

9.22. Cláusulas contractuales mínimas

El contrato con el proveedor debería regular:

- • acceso a documentación;
 - • notificación previa de cambios;
 - • conservación de versiones;
 - • disponibilidad de registros;
 - • cooperación en incidentes;
 - • tiempos de respuesta;
 - • ciberseguridad;
 - • propiedad y uso de datos;
 - • reversibilidad;
 - • continuidad del servicio;
 - • derecho de auditoría;
 - • responsabilidad;
 - • subcontratistas.
- Debe evitarse que la empresa dependa de una caja negra sin capacidad de obtener evidencia suficiente para justificar su utilización.

9.23. Propiedad de los datos

Debe definirse:

- • quién es titular;
 - • quién puede utilizarlos;
 - • para qué;
 - • durante cuánto tiempo;
 - • si se incorporan a entrenamiento;
 - • si se comparten;
 - • cómo se eliminan;
 - • cómo se exportan al finalizar el contrato.
- Un proveedor no debería utilizar automáticamente datos de producción sensibles para entrenar modelos generales si no existe una base contractual clara.

9.24. Seguridad y ciberseguridad

La integridad de la predicción depende de la seguridad del sistema.

Deben analizarse:

- • acceso no autorizado;
- • alteración de datos;
- • suplantación de sensores;
- • malware;
- • pérdida de servicio;
- • manipulación de modelos;
- • extracción de información;
- • actualización comprometida;
- • dependencia de servicios externos.

- **Los controles pueden incluir:**

- • autenticación;
- • autorización por roles;
- • cifrado;
- • firma;
- • segmentación de red;
- • copias;
- • registro;
- • monitorización;
- • gestión de vulnerabilidades;
- • pruebas de recuperación.

El Reglamento de IA exige, para los sistemas de alto riesgo incluidos en su ámbito, niveles apropiados de exactitud, robustez y ciberseguridad durante el ciclo de vida. Aunque no todo modelo alimentario será jurídicamente de alto riesgo, esta exigencia refleja una buena práctica aplicable a sistemas que influyen materialmente en calidad o seguridad.

9.25. Plan de continuidad

La empresa debe responder:

- ¿Qué ocurre si el modelo deja de estar disponible durante cuatro horas?

- **Las opciones seguras incluyen:**

- • procedimiento manual;
- • retención temporal;
- • regla conservadora;
- • método alternativo;
- • reducción de producción;
- • parada.
- No debe improvisarse tras el fallo.

- **El plan debe establecer:**

- • tiempo máximo de recuperación;
- • responsable;
- • respaldo;
- • datos que deben recuperarse;
- • validación posterior;
- • gestión de decisiones pendientes.

9.26. Supervisión humana efectiva

La existencia formal de un operario no garantiza supervisión real.

La persona debe:

- • conocer la finalidad;
- • entender la salida;
- • identificar resultados anómalos;
- • acceder a datos relevantes;
- • poder detener el proceso;

- • poder anular la recomendación;
- • documentar la decisión.

El Reglamento de IA contempla mecanismos de supervisión humana en los sistemas de alto riesgo para que una persona pueda comprender las capacidades y limitaciones, vigilar anomalías e intervenir o detener el sistema cuando sea necesario.

9.27. Diseño de la interfaz

La interfaz debe evitar:

- • precisión aparente excesiva;
- • colores engañosos;
- • categorías ambiguas;
- • ausencia de incertidumbre;
- • mensajes técnicos incomprensibles;
- • recomendación sin contexto.

• No debería mostrar únicamente:

«Seguro: 94,7 %.»

Debería presentar:

- • tipo de salida;
- • umbral;
- • incertidumbre;
- • condición de dominio;
- • variables faltantes;
- • acción requerida;
- • limitación.

• Ejemplo:

Riesgo bajo según el modelo, con datos dentro del dominio validado. Resultado no equivalente a análisis microbiológico. Mantener el plan de muestreo aplicable.

9.28. Formación

La formación debe cubrir:

- • qué predice;
 - • qué no predice;
 - • errores posibles;
 - • falsos positivos;
 - • falsos negativos;
 - • datos fuera de dominio;
 - • abstención;
 - • respuesta ante alerta;
 - • anulación;
 - • registro;
 - • incidentes.
- Debe verificarse la competencia, no solo la asistencia.

- **Puede evaluarse mediante casos:**

- • resultado contradictorio con el proceso;
- • dato imposible;
- • baja confianza;
- • versión incorrecta;
- • alerta crítica;
- • fallo de conexión.

9.29. Sesgo de automatización

Los usuarios pueden aceptar una salida porque procede de un sistema sofisticado.

Este sesgo aumenta cuando:

- • el algoritmo suele acertar;
- • la interfaz transmite certeza;
- • la persona tiene poco tiempo;
- • el criterio técnico se ha debilitado;
- • la anulación requiere justificación compleja.
- La empresa debe evitar que la supervisión se convierta en una ratificación automática.

9.30. CAPA asociada al modelo

Cuando aparece una desviación debe iniciarse una investigación de causa raíz.

Las causas pueden estar en:

- • datos;
- • sensor;
- • proceso;
- • modelo;
- • umbral;
- • interfaz;
- • usuario;
- • documentación;
- • proveedor;
- • deriva.

- **La CAPA debería diferenciar:**

Corrección inmediata

- retener producto;
- repetir análisis;
- suspender sistema;
- usar respaldo.

Acción correctiva

- corregir causa;
- recalibrar;

- modificar procedimiento;
- formar;
- revalidar.

Acción preventiva

- ampliar monitorización;
- mejorar detección de dominio;
- añadir controles;
- revisar contratos;
- aumentar muestreo.

9.31. Investigación del falso negativo

Un falso negativo crítico debe reconstruirse por completo.

Preguntas:

1. • ¿La entrada era correcta?
 2. • ¿Pertenece al dominio?
 3. • ¿La versión era la autorizada?
 4. • ¿El umbral era correcto?
 5. • ¿La probabilidad estaba calibrada?
 6. • ¿Existían señales de deriva?
 7. • ¿El patrón de referencia es fiable?
 8. • ¿El usuario actuó correctamente?
 9. • ¿Falló una barrera adicional?
 10. • ¿Existen otros lotes afectados?
- No debe clasificarse automáticamente como «caso aislado».

9.32. Auditoría interna

La auditoría debería revisar:

- • coincidencia entre uso real y autorizado;
- • versiones;
- • registros;
- • cambios;
- • formación;
- • monitorización;
- • acciones;
- • verificaciones;
- • proveedores;
- • incidencias;
- • revalidación.

Preguntas de auditoría

- ¿Puede identificarse la versión utilizada en un lote?

- ¿Se conservan entradas y salida?
- ¿Hay productos fuera del dominio?
- ¿Se verifican los negativos?
- ¿El umbral fue aprobado?
- ¿Se investigan abstenciones?
- ¿Existen actualizaciones automáticas?
- ¿Puede detenerse el sistema?
- ¿Se ha probado el respaldo?
- ¿Los usuarios comprenden las limitaciones?

9.33. Auditoría del proveedor

La auditoría puede abarcar:

- • gobierno de desarrollo;
- • calidad de datos;
- • control de código;
- • validación;
- • seguridad;
- • cambios;
- • incidentes;
- • continuidad;
- • subcontratación.
- En sistemas críticos, puede ser necesario revisar evidencia primaria, no únicamente certificados comerciales.

9.34. Revisión por la dirección

La revisión debería incluir:

- • rendimiento;
- • deriva;
- • incidentes;
- • CAPA;
- • cambios;
- • nuevos productos;
- • proveedores;
- • recursos;
- • formación;
- • beneficio;
- • riesgo residual.
- **La decisión puede ser:**
 - • mantener;
 - • restringir;
 - • mejorar;
 - • recalibrar;

- • revalidar;
- • retirar.
- La continuidad no debe darse por supuesta porque el sistema ya esté implantado.

9.35. Paquete mínimo de evidencia para auditoría

La empresa debería poder presentar:

1. • finalidad y alcance;
 2. • análisis de riesgo;
 3. • dominio;
 4. • protocolo;
 5. • informe de validación;
 6. • criterios;
 7. • aprobación;
 8. • procedimiento;
 9. • formación;
 10. • registros;
 11. • monitorización;
 12. • verificación;
 13. • cambios;
 14. • incidentes;
 15. • revalidaciones;
 16. • contratos relevantes.
- **La documentación debe permitir que un tercero competente comprenda:**
 - • qué hace el modelo;
 - • por qué se confía en él;
 - • dónde puede utilizarse;
 - • cómo se controla;
 - • qué sucede si falla.

9.36. Integración con certificaciones de seguridad alimentaria

Los esquemas de certificación pueden exigir evidencias relacionadas con:

- • control documental;
- • validación;
- • verificación;
- • calibración;
- • proveedores;
- • cambios;
- • trazabilidad;
- • acciones correctivas;
- • competencia.
- Aunque no contengan todavía capítulos extensos específicamente dedicados a IA, un modelo que influye en estos procesos debe quedar integrado en ellos.

- No debería presentarse al auditor como un proyecto del departamento digital separado del sistema de calidad.

9.37. Modelo de responsabilidad RACI

Una asignación posible sería:

Actividad	Desarrollo	Calidad	Producción	TI	Laboratorio	Dirección
Definir finalidad	C	R	C	C	C	A
Validar datos	R	C	C	C	R	I
Fijar criterios	C	R	C	I	C	A
Aprobar modelo	I	R	C	C	C	A
Operar	I	C	R	C	I	I
Monitorizar	R	R	C	R	C	I
Gestionar incidente	C	R	R	R	C	A
Autorizar cambio	C	R	C	C	C	A

• **Donde:**

- R: responsable de ejecución;
- A: responsable final;
- C: consultado;
- I: informado.
- La matriz debe adaptarse a la organización, pero siempre debe existir un responsable final.

9.38. Modelo de ficha de control operativo

Modelo:

Versión:

Finalidad:

Producto/línea:

Variables obligatorias:

Rango autorizado:

Umbral:

Resultado:

Incertidumbre:

Estado del dominio:

Acción automática:

Confirmación requerida:

Responsable:

Procedimiento alternativo:

Registro asociado:

Criterio de suspensión:

Esta ficha convierte la validación en instrucciones operativas.

9.39. Modelo de registro de predicción

Campo	Contenido
Lote	Identificación
Fecha/hora	Marca temporal
Producto	Referencia
Planta/línea	Ubicación
Modelo	Nombre y versión
Entradas	Valores
Calidad de datos	Conforme/no conforme
Dominio	Dentro/fuera
Salida	Clase o estimación
Probabilidad	Valor
Incertidumbre	Intervalo o nivel
Umbral	Valor aplicado
Decisión	Acción
Usuario	Identificación
Anulación	Sí/no y motivo
Referencia posterior	Resultado
Cierre	Responsable y fecha

9.40. Modelo de gestión de cambio

Cambio propuesto:

Motivo:

Elementos afectados:

Impacto sobre datos:

Impacto sobre rendimiento:

Impacto sobre seguridad:

Pruebas requeridas:

Revalidación: parcial/completa

Criterios:

Aprobaciones:

Fecha de despliegue:

Plan de reversión:

Verificación posterior:

9.41. Modelo de evaluación de incidentes

Descripción:

Fecha:

Lotes afectados:

Consecuencia:

Versión:

Entradas:

Resultado del modelo:

Resultado real:

Tipo de error:

Causa raíz:

Producto en mercado:

Contención:

CAPA:

Necesidad de revalidación:

Cierre:

9.42. Indicadores para revisión periódica

Un cuadro ejecutivo debería incluir:

- • número de predicciones;
- • distribución por producto;
- • sensibilidad;
- • especificidad;
- • falsos negativos;
- • falsos positivos;
- • MAE;
- • sobreestimaciones peligrosas;
- • calibración;
- • abstenciones;
- • datos fuera de dominio;
- • incidencias;
- • anulaciones;
- • disponibilidad;
- • acciones pendientes.
- No debe limitarse a una única cifra de accuracy.

9.43. Integración con la cultura de seguridad alimentaria

La implantación de IA puede fortalecer o debilitar la cultura de seguridad.

La fortalece cuando:

- • mejora la información;
- • permite aprender;
- • hace visibles las desviaciones;
- • facilita decisiones;
- • conserva la responsabilidad.
- **La debilita cuando:**
 - • desplaza el juicio crítico;

- • oculta la incertidumbre;
- • centraliza conocimiento en un proveedor;
- • reduce la investigación;
- • permite culpar al algoritmo;
- • convierte la tecnología en autoridad incuestionable.
- La responsabilidad sigue siendo humana y organizativa.

9.44. Principio de responsabilidad no delegable

La empresa alimentaria no puede defender una decisión incorrecta afirmando simplemente:

«El algoritmo lo indicó.»

Debe poder demostrar:

- • por qué utilizó el sistema;
- • cómo lo validó;
- • cómo lo vigiló;
- • qué limitaciones conocía;
- • qué respuesta estableció;
- • por qué la decisión era razonable.

La IA produce una salida. La organización decide la función jurídica, técnica y operativa que atribuye a esa salida.

9.45. El modelo como evidencia y no como oráculo

Un resultado algorítmico debe integrarse con:

- • datos analíticos;
- • conocimiento del proceso;
- • información histórica;
- • condiciones reales;
- • incertidumbre;
- • requisitos legales;
- • juicio profesional.

- **El peso atribuido al modelo debe ser proporcional a:**

peso de la predicción $\propto$ calidad de la validación $\times$ representatividad $\times$ control del riesgo

No debe basarse en la novedad de la tecnología ni en la complejidad matemática.

9.46. Conclusión del bloque

La validación científica termina únicamente cuando el modelo se ha convertido en un componente controlado del sistema alimentario.

Eso exige:

- • definir su función;
- • integrarlo en APPCC o prerrequisitos;
- • revisar el análisis de peligros;
- • documentar los modos de fallo;
- • establecer monitorización y verificación;

- • conservar registros;
- • controlar versiones;
- • gestionar cambios;
- • asegurar supervisión humana;
- • preparar contingencias;
- • auditar su uso;
- • revalidarlo cuando proceda.
- Un modelo estadísticamente excelente pero desconectado del sistema documental, de las responsabilidades y de las medidas de respaldo no está adecuadamente implantado.
- **Por el contrario, un sistema científicamente maduro permite responder en cualquier momento:**
 1. • Qué versión está operando.
 2. • Para qué fue aprobada.
 3. • Qué datos recibió.
 4. • Qué predijo.
 5. • Con qué incertidumbre.
 6. • Qué decisión produjo.
 7. • Quién la supervisó.
 8. • Cómo se verificó.
 9. • Qué ocurrió realmente.
 10. • Qué se hará si el rendimiento cambia.
- Esta trazabilidad transforma la predicción en evidencia auditable y evita que la IA se convierta en una zona opaca dentro del sistema de seguridad alimentaria.

10. Protocolo práctico para la validación científica de modelos de IA en calidad, vida útil y seguridad alimentaria

La validación de un modelo de inteligencia artificial debe gestionarse como un proyecto científico, tecnológico y documental independiente del desarrollo informático. Su objetivo no es confirmar que el algoritmo «funciona», sino reunir evidencia suficiente para decidir si una versión determinada del sistema puede utilizarse en una finalidad alimentaria concreta, dentro de un dominio definido y bajo unas condiciones de control previamente establecidas.

Las Directrices del Codex sobre validación de medidas de control distinguen la demostración inicial de capacidad, la monitorización durante el funcionamiento y la verificación posterior. Por su parte, el NIST recomienda que los conjuntos de prueba, las métricas, los instrumentos, los procedimientos de evaluación y los incidentes queden documentados dentro de procesos de prueba, evaluación, verificación y validación desarrollados durante todo el ciclo de vida.

El Reglamento (UE) 2024/1689 establece, para los sistemas de alto riesgo incluidos en su ámbito, que los conjuntos de entrenamiento, validación y prueba deben someterse a prácticas de gobernanza adecuadas a la finalidad prevista, y exige niveles apropiados de exactitud, robustez y ciberseguridad mantenidos durante el ciclo de vida. Aunque un modelo alimentario concreto no tenga necesariamente esa clasificación jurídica, estos requisitos constituyen una referencia metodológica útil para estructurar su validación.

El protocolo que se propone a continuación se articula en trece fases:

Definición → Riesgo → Datos → Diseño → Congelación → Validación → Robustez → Impacto → Decisión → Despliegue → Vigilancia → Revalidación → Retirada

Cada fase debe generar entregables, responsables y criterios de avance. El proyecto no debe pasar a la fase siguiente cuando permanezcan incumplimientos críticos sin resolver.

10.1. Fase 0. Constitución del equipo y gobierno del proyecto

Antes de analizar datos debe constituirse un equipo multidisciplinar con autoridad suficiente para aceptar, restringir o rechazar el sistema.

Composición mínima

- responsable del proceso alimentario;
- responsable de calidad y seguridad alimentaria;
- tecnólogo de alimentos;
- microbiólogo o especialista analítico, cuando proceda;
- científico de datos;
- estadístico o especialista en validación;
- responsable de tecnologías de la información;
- mantenimiento o metrología;
- asesoría regulatoria y jurídica;
- representante de los usuarios;
- responsable final de aprobación.

La validación no debería quedar bajo control exclusivo del equipo que desarrolló el modelo. El NIST recomienda disponer de procedimientos objetivos, repetibles y documentados de prueba y evaluación, así como de responsabilidades diferenciadas durante el ciclo de vida.

Preguntas iniciales

1. ¿Quién es propietario del modelo?
2. ¿Quién es propietario del riesgo?
3. ¿Quién define la verdad de referencia?
4. ¿Quién puede modificar el umbral?
5. ¿Quién autoriza una nueva versión?
6. ¿Quién vigila el rendimiento?
7. ¿Quién puede suspender el sistema?
8. ¿Quién responde ante un incidente?

Entregables

- acta de constitución;
- matriz RACI;
- identificación del propietario;
- composición del comité de validación;
- política de conflictos de interés;
- procedimiento de aprobación.

Criterio de avance

No debe iniciarse la validación formal sin una asignación inequívoca de responsabilidades.

10.2. Fase 1. Definición del uso previsto

La primera obligación científica consiste en formular con exactitud la finalidad del sistema.

La declaración debe responder:

- • qué predice;
- • para qué producto;
- • en qué proceso;
- • en qué momento;
- • con qué variables;
- • para qué usuario;
- • sobre qué resultado real;
- • con qué grado de autonomía;
- • qué decisión puede provocar;
- • qué usos quedan prohibidos.

Plantilla de finalidad prevista

La versión [X] del sistema [nombre] utiliza [variables] para estimar [resultado] en [productos, líneas o plantas], dentro de [rangos y condiciones]. La salida se utilizará para [decisión]. No sustituirá [análisis, control o decisión excluidos]. Los resultados fuera del dominio, incompletos o de incertidumbre elevada se remitirán a [procedimiento alternativo].

Ejemplo de calidad

El modelo clasifica unidades envasadas como conformes o potencialmente defectuosas a partir de imágenes captadas en línea. Activa el rechazo preventivo, pero no modifica las especificaciones comerciales ni la definición de defecto.

Ejemplo de vida útil

El modelo estima la vida útil remanente de una salsa refrigerada utilizando formulación, pH, actividad de agua, tratamiento térmico e historial de temperatura. Su salida se integra con estudios microbiológicos, fisicoquímicos y sensoriales y no constituye, por sí sola, una autorización para ampliar la fecha de caducidad.

Ejemplo de seguridad

El modelo estima la probabilidad de que un lote supere un criterio microbiológico al final de su vida útil. Se utiliza para intensificar el muestreo y la investigación; no sustituye los análisis reglamentarios ni autoriza automáticamente la liberación.

Entregables

- declaración de finalidad;
- diagrama de decisiones;
- lista de usos autorizados;
- lista de usos prohibidos;
- identificación de usuarios;
- definición del nivel de autonomía.

Criterio de avance

La finalidad debe ser suficientemente concreta para diseñar un ensayo capaz de demostrarla o refutarla.

10.3. Fase 2. Análisis de riesgo y criticidad

Debe analizarse tanto el riesgo alimentario que el modelo pretende controlar como los riesgos que introduce su utilización.

Identificación de errores

Para cada salida deben evaluarse:

- • falso positivo;
- • falso negativo;
- • sobreestimación;
- • infraestimación;
- • resultado indeterminado;
- • dato fuera de dominio;

- fallo de disponibilidad;
- error de interfaz;
- uso incorrecto;
- deriva;
- manipulación.

Evaluación del riesgo

Puede utilizarse una matriz:

$$R=S \times P \times D$$

donde:

- (S) es la severidad;
- (P), la probabilidad;
- (D), la dificultad de detección antes del daño.

No obstante, la puntuación numérica debe acompañarse de una justificación técnica. Una cifra agregada no debe ocultar un modo de fallo sanitario crítico.

Clasificación de criticidad

Clase A. Exploratoria

No afecta directamente a producto ni proceso.

Clase B. Apoyo informativo

La decisión permanece bajo revisión humana completa.

Clase C. Cribado o priorización

Puede excluir casos de controles posteriores.

Clase D. Decisión automática de calidad

Retiene, rechaza, clasifica o reprocesa producto.

Clase E. Influencia sobre seguridad

Interviene en liberación, vida útil, control de peligros o medidas APPCC.

Consecuencias metodológicas

A mayor criticidad deberán aumentar:

- independencia de la validación;
- tamaño muestral;
- exigencia del límite inferior de confianza;
- validación externa;
- pruebas de robustez;
- vigilancia;
- redundancia;
- control de cambios;
- capacidad de abstención.

Entregables

- análisis de riesgos;

- FMEA del sistema;
- clasificación de criticidad;
- identificación de barreras;
- definición del riesgo residual;
- criterios preliminares de suspensión.

Criterio de avance

No debe aprobarse un diseño que permita que un fallo crítico provoque directamente una liberación insegura sin una barrera independiente.

10.4. Fase 3. Definición del patrón de referencia

El patrón de referencia debe representar de forma válida el fenómeno que se pretende predecir.

Para calidad

- método analítico validado;
- especificación contractual;
- inspección normalizada;
- panel sensorial entrenado;
- clasificación por expertos con reglas documentadas.

Para vida útil

- día de incumplimiento microbiológico;
- momento de rechazo sensorial;
- pérdida de una propiedad fisicoquímica;
- tiempo hasta exceder una especificación;
- evento compuesto previamente definido.

Para seguridad microbiológica

- método de referencia reconocido;
- ensayo de desafío;
- estudio de durabilidad;
- resultado microbiológico confirmado;
- criterio reglamentario o interno validado.

La microbiología predictiva y los modelos de IA pueden contribuir a la evaluación de vida útil, pero la conclusión debe integrarse con evidencia experimental y con las características específicas de la matriz, el proceso y las condiciones de conservación. La información actual de EFSA sobre *Listeria monocytogenes* sigue vinculando la evaluación del riesgo con los criterios microbiológicos, la temperatura, la vida útil y las técnicas de microbiología predictiva.

Validación de la referencia

Debe documentarse:

- exactitud;

- • precisión;
- • reproducibilidad;
- • límite de detección;
- • incertidumbre;
- • frecuencia de resultados indeterminados;
- • variación entre operadores;
- • variación entre laboratorios;
- • reglas de adjudicación.

Desacuerdos entre expertos

Cuando no existe una medida objetiva única, puede utilizarse:

- • evaluación por varios expertos;
 - • consenso;
 - • adjudicación ciega;
 - • coeficiente de acuerdo;
 - • definición de categoría indeterminada.
- El modelo no debe evaluarse como si la etiqueta fuese infalible cuando el acuerdo entre expertos es reducido.

Entregables

- procedimiento de referencia;
- cualificación del método;
- reglas de etiquetado;
- análisis de incertidumbre;
- gestión de resultados indeterminados.

Criterio de avance

La variable objetivo debe estar definida antes de recopilar el conjunto de prueba final.

10.5. Fase 4. Plan de datos y representatividad

Debe elaborarse una matriz de cobertura que compare la población de utilización con los datos disponibles.

Dimensiones mínimas

- familias de producto;
- formulaciones;
- plantas;
- líneas;
- equipos;
- proveedores;
- campañas;
- estaciones;
- turnos;

- operadores;
- intervalos de pH;
- actividad de agua;
- humedad;
- temperatura;
- concentración;
- envase;
- prevalencia del evento;
- severidad del defecto.

Matriz de cobertura

Dimensión	Población prevista	Entrenamiento	Ajuste	Prueba externa	Brecha
Productos	—	—	—	—	—
Plantas	—	—	—	—	—
Campañas	—	—	—	—	—
Proveedores	—	—	—	—	—
Rangos críticos	—	—	—	—	—
Positivos	—	—	—	—	—
Condiciones extremas	—	—	—	—	—

El Reglamento de IA exige, para los sistemas de alto riesgo a los que resulte aplicable, que los datos sean pertinentes, suficientemente representativos y adecuados a la finalidad prevista, y que se identifiquen las carencias que puedan impedir el cumplimiento.

Unidad independiente

Debe definirse si la unidad es:

- • unidad de producto;
- • muestra;
- • lote;
- • día de fabricación;
- • curva de vida útil;
- • proveedor;
- • planta;
- • campaña.
- Las imágenes, espectros, píxeles o mediciones repetidas no deben contabilizarse como unidades biológicas independientes cuando proceden de la misma muestra.

Datos positivos raros

Cuando el evento sea infrecuente puede utilizarse un conjunto enriquecido, pero deben diferenciarse:

- • muestra analítica enriquecida;
- • prevalencia operativa real.
- Las métricas dependientes de la prevalencia deberán recalcularse para el escenario real.

Entregables

- inventario de datos;
- matriz de cobertura;
- análisis de brechas;
- diccionario de variables;
- evaluación de calidad;
- definición de unidad estadística;
- justificación del tamaño muestral.

Criterio de avance

Las carencias relevantes deben resolverse, limitar el dominio o quedar reconocidas como restricciones explícitas.

10.6. Fase 5. Diseño estadístico y partición

La partición debe reproducir la forma en la que el modelo encontrará datos futuros.

Principios

- ningún lote debe aparecer simultáneamente en entrenamiento y prueba;
- ninguna réplica de la misma muestra debe repartirse entre conjuntos;
- las transformaciones deben ajustarse únicamente con entrenamiento;
- el conjunto de prueba no debe utilizarse para seleccionar variables;
- los umbrales deben fijarse antes de abrir el test final.

Diseños posibles

Separación por lotes

Adecuada cuando el modelo se aplicará a lotes nuevos.

Separación temporal

Adecuada para utilización prospectiva:

$$T_{\text{entrenamiento}} < T_{\text{validación}} < T_{\text{prueba}}$$

Separación por planta

Permite evaluar portabilidad.

Separación por proveedor

Evalúa generalización a nuevas materias primas.

Separación por instrumento

Necesaria cuando existen varios sensores o laboratorios.

Validación multicéntrica

Preferible para usos críticos y transferibles.

Diseño recomendado por criticidad

Criticidad	Diseño mínimo
Exploratoria	Validación interna agrupada
Apoyo	Test temporal independiente
Cribado	Test externo y prospectivo
Calidad automática	Multilínea o multiplanta
Seguridad	Externa, prospectiva, por subgrupos y condiciones adversas

Entregables

- plan estadístico;
- algoritmo de partición;
- lista de unidades por conjunto;
- comprobación de ausencia de fuga;
- procedimiento de bloqueo del test.

Criterio de avance

El evaluador debe poder demostrar que ninguna información de las unidades de prueba influyó en el desarrollo.

10.7. Fase 6. Pre-especificación de métricas y criterios de aceptación

Las métricas deben corresponder a la decisión.

Clasificación

- sensibilidad;
- especificidad;
- falsos positivos;
- falsos negativos;
- valores predictivos;
- F1 o F_1 o (F_{β}) ;
- MCC;
- AUC-ROC;
- área *precision-recall*;
- calibración;
- tasa de abstención.

Regresión y vida útil

- sesgo;
- MAE;
- RMSE;

- percentiles del error;
- sobreestimación peligrosa;
- límites de acuerdo;
- cobertura de intervalos;
- desempeño por horizonte.

Seguridad microbiológica

- error en la dirección no conservadora;
- sensibilidad cerca del límite;
- factor de sesgo;
- factor de exactitud;
- tiempo hasta límite;
- tasa de clasificación insegura.

Reglas jerárquicas

Los criterios críticos no deben compensarse con resultados secundarios.

Ejemplo:

$$\text{LCB}_{95\%}(\text{sensibilidad}) \geq 0,95$$

$$\text{Tasa de FN} \leq 0,03$$

$$\text{Especificidad} \geq 0,85$$

$$\text{Abstención} \leq 0,10$$

Todos los subgrupos críticos cumplen

Para vida útil:

$$\text{MAE} \leq 3 \text{ días}$$

$$|\text{ME}| \leq 1 \text{ día}$$

$$P(e > 2 \text{ días}) \leq 5\%$$

$$\text{Cobertura del IP}_{95\%} \geq 92\%$$

Estos valores son ejemplos metodológicos, no umbrales universales.

Entregables

- tabla de métricas;
- criterios prioritarios;
- criterios secundarios;
- intervalos requeridos;
- análisis por subgrupos;
- criterios de suspensión.

Criterio de avance

Todos los criterios deben quedar firmados antes de acceder al resultado final.

10.8. Fase 7. Congelación del sistema

Antes de la evaluación final debe congelarse la versión.

Elementos que deben bloquearse

- código;
- pesos;
- arquitectura;
- variables;
- preprocesamiento;
- imputación;
- umbral;
- reglas de negocio;
- versión de librerías;
- hardware;
- firmware;
- interfaz;
- detector de dominio;
- reglas de abstención.

Identificación de versión

Puede utilizarse:

$$ID = H(\text{código, pesos, configuración, umbral, entorno})$$

donde (H) representa una huella digital o identificador de integridad.

Reproducibilidad

Debe comprobarse que la misma entrada genera el mismo resultado bajo las condiciones especificadas, salvo que el sistema sea deliberadamente estocástico. En ese caso, debe definirse la variabilidad permitida.

Entregables

- acta de congelación;
- identificador de versión;
- imagen del entorno;
- repositorio controlado;
- copia de seguridad;
- procedimiento de reproducción.

Criterio de avance

Cualquier modificación posterior obligará a documentar una nueva versión y a evaluar su impacto.

10.9. Fase 8. Ejecución ciega de la validación externa

El conjunto de prueba debe procesarse sin modificar el modelo y, preferiblemente, sin que los operadores conozcan el resultado de referencia.

Procedimiento

1. verificar integridad del conjunto;
2. comprobar criterios de inclusión;
3. registrar exclusiones;
4. ejecutar el sistema congelado;
5. conservar salidas brutas;
6. aplicar el umbral predefinido;
7. vincular predicción y referencia;
8. calcular las métricas;
9. estimar intervalos;
10. analizar subgrupos;
11. revisar casos extremos;
12. documentar desviaciones.

Gestión de exclusiones

No deben eliminarse retrospectivamente los casos difíciles salvo que incumplan criterios preespecificados.

Deben diferenciarse:

- • exclusión válida por fallo de referencia;
- • entrada fuera de dominio correctamente detectada;
- • fallo técnico;
- • predicción indeterminada;
- • error del modelo.
- Excluir sistemáticamente los fallos produce una evaluación artificial.

Análisis primario

Debe responder a la hipótesis principal.

Análisis secundarios

Pueden incluir:

- • producto;
- • planta;
- • campaña;
- • proveedor;
- • concentración;
- • severidad;
- • condiciones ambientales.

Entregables

- base de resultados bloqueada;
- matriz de confusión;
- métricas;
- intervalos;
- gráficos;
- análisis de subgrupos;
- registro de exclusiones;
- listado de errores críticos.

Criterio de avance

El modelo debe cumplir todos los criterios críticos. El incumplimiento no puede corregirse modificando retrospectivamente el protocolo.

10.10. Fase 9. Pruebas de robustez y estrés

La validación externa ordinaria debe complementarse con pruebas dirigidas a los límites.

El Reglamento de IA exige, para los sistemas de alto riesgo afectados, un nivel de robustez apropiado y funcionamiento consistente durante su ciclo de vida; el NIST recomienda explorar limitaciones, riesgos emergentes y modos de fallo mediante TEVV documentado.

Robustez de entradas

- ruido;
- redondeo;
- datos ausentes;
- variación de unidades;
- pequeñas desviaciones;
- errores razonables de medida.

Robustez instrumental

- diferentes cámaras;
- iluminación;
- suciedad;
- calibración;
- deriva;
- resolución;
- velocidad.

Robustez alimentaria

- campañas;
- proveedores;

- formulaciones extremas;
- temperaturas adversas;
- pH límite;
- actividad de agua límite;
- concentraciones próximas al criterio.

Robustez operacional

- caída de red;
- latencia;
- reinicio;
- pérdida parcial de datos;
- carga elevada;
- modo degradado.

Robustez frente a datos fuera de dominio

Debe comprobarse que el sistema:

- • identifica valores fuera de rango;
- • reconoce combinaciones atípicas;
- • se abstiene;
- • no emite una decisión automática segura por defecto.

Criterio de degradación

Puede definirse:

$$\Delta M = M_{\text{condición adversa}} - M_{\text{condición nominal}}$$

y exigir:

$$\Delta M \geq -\delta$$

donde (δ) representa la degradación máxima permitida.

Entregables

- plan de estrés;
- resultados;
- degradación por escenario;
- modos de fallo;
- acciones;
- restricciones adicionales.

Criterio de avance

Ningún escenario razonablemente previsible debe provocar un fallo peligroso sin detección o respuesta segura.

10.11. Fase 10. Evaluación humana y operacional

El rendimiento matemático debe evaluarse dentro del flujo de trabajo real.

Preguntas

- ¿El usuario comprende la salida?
- ¿Interpreta correctamente la probabilidad?
- ¿Distingue alerta, predicción e incertidumbre?
- ¿Sabe cuándo anular?
- ¿Puede detener el sistema?
- ¿Las alertas llegan a tiempo?
- ¿Existe fatiga?
- ¿La carga de falsos positivos es asumible?
- ¿El sistema altera comportamientos de manera indeseada?

Diseño

Puede utilizarse un estudio en modo sombra:

- el modelo predice;
- la práctica habitual continúa;
- se comparan decisiones;
- no se permite todavía impacto automático.
- **Posteriormente puede realizarse un despliegue limitado:**
 - una línea;
 - un turno;
 - una planta;
 - una familia de producto;
 - revisión humana obligatoria.

Métricas operacionales

- tiempo hasta decisión;
- alertas por turno;
- porcentaje investigado;
- tasa de anulaciones;
- anulaciones correctas;
- errores humanos;
- latencia;
- disponibilidad;
- volumen de confirmaciones;
- desperdicio;
- producto recuperado.

Supervisión humana

En los sistemas a los que resulta aplicable, el Reglamento de IA exige que la supervisión permita comprender las capacidades y limitaciones, vigilar anomalías, intervenir y detener el sistema cuando no actúe como está previsto. Este principio es igualmente pertinente en sistemas alimentarios de elevada criticidad.

Entregables

- evaluación de usabilidad;
- resultados en modo sombra;
- análisis de automatización;
- carga operativa;
- plan de formación;
- instrucciones definitivas.

Criterio de avance

El sistema no debe aprobarse si la interfaz o el flujo de trabajo inducen errores aunque el modelo matemático cumpla las métricas.

10.12. Fase 11. Evaluación comparativa y de impacto

El modelo debe compararse con la práctica existente y con alternativas razonables.

Comparadores

- inspección humana;
- regla de decisión;
- modelo estadístico;
- microbiología predictiva convencional;
- muestreo actual;
- procedimiento histórico;
- modelo anterior.

Hipótesis posibles

Superioridad

$$M_{IA} > M_{referencia} + \Delta$$

No inferioridad

$$M_{IA} - M_{referencia} > -\Delta$$

donde (Δ) representa la diferencia máxima aceptable.

Impacto en calidad

- reducción de falsas aceptaciones;
- reducción de falsas rechazos;
- uniformidad;
- tiempo de inspección;
- reclamaciones.

Impacto en vida útil

- menor sobreestimación;

- reducción de desperdicio;
- mejor adaptación logística;
- menor variabilidad entre lotes;
- anticipación de deterioro.

Impacto en seguridad

- detección más temprana;
- mayor cobertura;
- menor tasa de falsos negativos;
- reducción del tiempo hasta intervención;
- mejora del muestreo.

Impactos adversos

- exceso de alertas;
- aumento de desperdicio;
- pérdida de juicio experto;
- reducción de controles independientes;
- dependencia de proveedor;
- nuevos fallos técnicos.

Entregables

- estudio comparativo;
- análisis de impacto;
- beneficio neto;
- riesgos residuales;
- recomendación de uso.

Criterio de avance

El sistema debe aportar una mejora relevante o una ventaja operacional justificable sin degradar las garantías críticas.

10.13. Fase 12. Decisión del comité de validación

El comité debe revisar la evidencia completa, no solo un resumen ejecutivo.

Opciones

No aprobado

Existe incumplimiento crítico o evidencia insuficiente.

Aprobado para investigación

No puede influir en producto comercial.

Aprobado en modo sombra

Puede operar sin ejecutar decisiones.

Aprobado como apoyo

Requiere revisión humana completa.

Aprobado con confirmación

Toda salida relevante debe confirmarse independientemente.

Aprobado para dominio restringido

Solo productos, plantas, equipos y rangos expresamente autorizados.

Aprobado condicionalmente

Requiere completar datos, mantener barreras y superar una revisión posterior.

Aprobado para uso completo

Cumple todos los criterios y dispone de controles durante el ciclo de vida.

Regla de decisión

$$\text{Aprobación} \Leftrightarrow \text{cap}_j = 1^m \text{ C}_j = \text{cumple}$$

para todos los criterios críticos (C_j).

Acta de aprobación

Debe contener:

- • versión;
- • finalidad;
- • dominio;
- • métricas;
- • limitaciones;
- • condiciones;
- • controles;
- • responsables;
- • fecha;
- • revisión prevista;
- • criterios de suspensión.

Criterio de avance

No debe desplegarse el sistema sin autorización formal, trazable y limitada a la evidencia obtenida.

11. Cronograma técnico orientativo

La duración real dependerá de la disponibilidad de datos, del evento estudiado y del tiempo necesario para obtener resultados de referencia.

Fase	Actividad principal	Entregable
0	Gobierno	Equipo y RACI
1	Uso previsto	Declaración de finalidad
2	Riesgo	FMEA y criticidad
3	Referencia	Método y etiquetado
4	Datos	Matriz de cobertura
5	Diseño	Plan estadístico
6	Criterios	Protocolo aprobado
7	Congelación	Versión bloqueada
8	Validación externa	Informe estadístico
9	Robustez	Informe de estrés
10	Operación	Estudio en sombra
11	Impacto	Comparación
12	Decisión	Acta de aprobación
Posterior	Vigilancia	Cuadro de mando
Por cambio	Revalidación	Nueva autorización

No debe confundirse el tiempo de desarrollo informático con el tiempo de validación científica. Un modelo puede entrenarse rápidamente y requerir meses para obtener suficientes positivos, completar estudios de vida útil o reunir campañas independientes.

12. Paquete documental mínimo

El expediente final debería contener:

1. • carta de proyecto;
2. • finalidad prevista;
3. • análisis de riesgos;
4. • descripción del sistema;
5. • inventario de variables;
6. • documentación de datos;
7. • método de referencia;
8. • matriz de representatividad;
9. • plan estadístico;
10. • protocolo firmado;
11. • acta de congelación;
12. • resultados externos;
13. • análisis de subgrupos;
14. • calibración;
15. • robustez;
16. • prueba de dominio;
17. • evaluación operacional;
18. • comparación con referencia;
19. • informe de limitaciones;
20. • acta de aprobación;
21. • procedimiento de uso;
22. • plan de monitorización;
23. • procedimiento de incidentes;
24. • control de cambios;
25. • plan de revalidación;
26. • plan de retirada.

El NIST considera la documentación de conjuntos de prueba, métricas, herramientas, procedimientos, resultados, incidentes y acciones de respuesta parte esencial de un sistema de TEVV y gestión continuada.

13. Matriz de aceptación práctica

Dimensión	Pregunta	Evidencia	Criterio
Finalidad	¿Está claramente definida?	Declaración aprobada	Sin ambigüedad
Referencia	¿La verdad es fiable?	Validación analítica	Cumple
Datos	¿Representan el uso?	Matriz de cobertura	Sin brechas críticas
Independencia	¿Existe fuga?	Auditoría de partición	Cero fugas
Detección	¿Cumple sensibilidad?	Test externo	Límite inferior aceptable
Falsos negativos	¿Son tolerables?	Matriz y análisis	Bajo máximo
Especificidad	¿Es operable?	Test y volumen	Sobre mínimo
Calibración	¿La probabilidad es interpretable?	Curva, intercepto, pendiente	Dentro del rango
Regresión	¿El error es aceptable?	MAE, RMSE, sesgo	Cumple
Vida útil	¿Controla sobreestimación?	Cola del error	Bajo máximo
Subgrupos	¿Existe grupo débil?	Análisis estratificado	Todos cumplen
Dominio	¿Reconoce lo desconocido?	Pruebas OOD	Abstención eficaz
Robustez	¿Tolera variaciones?	Estrés	Sin fallo crítico
Operación	¿Funciona en producción?	Modo sombra	Cumple
Humano	¿Se supervisa realmente?	Estudio de usuarios	Cumple
Seguridad	¿Existe modo seguro?	Prueba de contingencia	Verificado
Impacto	¿Mejora la práctica?	Comparación	Beneficio relevante
Deriva	¿Puede vigilarse?	Plan y cuadro de mando	Aprobado
Cambios	¿Existe control?	Procedimiento	Aprobado
Incidentes	¿Puede reconstruirse?	Trazabilidad	Completa

14. Protocolo específico para un modelo de calidad visual

14.1. Finalidad

Detectar y retirar unidades con defectos visibles.

14.2. Unidad independiente

Unidad física de producto, no número de imágenes o recortes.

14.3. Conjunto externo

Debe incluir:

- turnos;
- iluminaciones;
- velocidades;
- formatos;
- niveles de condensación;
- defectos raros;
- tamaños próximos al límite;
- productos conformes visualmente difíciles.

14.4. Métricas

- sensibilidad por defecto;
- falsos negativos por millón de unidades;
- falsos rechazos por mil;
- sensibilidad por tamaño;
- latencia;
- eficacia del expulsor;
- tasa de imágenes no evaluables.

14.5. Pruebas de estrés

- lente parcialmente sucia;
- cambio de iluminación;
- vibración;
- velocidad máxima;
- oclusión;
- reflejos;
- sustitución de cámara.

14.6. Condición de aprobación

No basta con que la red neuronal detecte el defecto. Debe validarse el sistema completo:

detección → señal → expulsión → confirmación

15. Protocolo específico para vida útil

15.1. Finalidad

Estimar el tiempo hasta alcanzar un criterio microbiológico, sensorial o fisicoquímico.

15.2. Diseño

- lotes completos;
- varias campañas;
- condiciones nominales;
- desviaciones térmicas;
- productos de frontera;
- datos longitudinales;
- correcta gestión de censura.

15.3. Métricas prioritarias

MAE
ME
 $P_{95}(|e|)$
 $P(e > \delta_{\text{peligroso}})$
Cobertura del intervalo

15.4. Criterio esencial

La sobreestimación debe controlarse separadamente:

$$P(\hat{T} - T > \delta) \leq R_{\text{max}}$$

15.5. Integración

La fecha asignada no debe depender únicamente de la estimación central:

$$T_{\text{asignada}} = L_{\alpha}(\hat{T}) - M$$

donde:

- (L_{α}) es un límite predictivo conservador;
- (M) , un margen adicional justificado.

15.6. Evidencia complementaria

- microbiología;
- challenge test cuando proceda;
- durabilidad;
- sensorial;
- fisicoquímica;
- condiciones logísticas;

- verificación industrial.

16. Protocolo específico para seguridad microbiológica

16.1. Finalidad

Detectar o anticipar un lote que incumplirá un criterio de seguridad.

16.2. Prioridad

La sensibilidad y la tasa de falsos negativos prevalecen sobre la accuracy global.

16.3. Datos

Deben incluir:

- • positivos suficientes;
- • niveles próximos al límite;
- • matrices diferentes;
- • variabilidad de pH y actividad de agua;
- • distintos lotes;
- • condiciones adversas;
- • plantas o periodos externos.

16.4. Regla de aceptación

$$LCB_{95\%}(\text{sensibilidad}) \geq S_{\min}$$

$$UCB_{95\%}(FNR) \leq FNR_{\max}$$

$$S_k \geq S_{k,\min} \forall k$$

16.5. Barreras

- confirmación microbiológica;
- muestreo de negativos;
- abstención;
- bloqueo fuera de dominio;
- procedimiento manual;
- revisión humana;
- suspensión ante fallo.

16.6. Prohibición metodológica

No debe verificarse únicamente lo que el modelo señala como positivo. Debe conservarse un muestreo independiente de predicciones negativas para estimar los falsos negativos.

17. Principales puntos de decisión

Puerta 1. Viabilidad científica

¿Existe un resultado de referencia fiable?

Si no existe, no debe desarrollarse una afirmación predictiva fuerte.

Puerta 2. Representatividad

¿Los datos cubren el dominio?

Si no, debe ampliarse la muestra o restringirse la finalidad.

Puerta 3. Independencia

¿El test es realmente externo?

Si no, solo puede afirmarse rendimiento interno.

Puerta 4. Rendimiento

¿Se cumplen todos los criterios críticos?

Si no, el modelo no es elegible para el uso previsto.

Puerta 5. Robustez

¿Los fallos se detectan y conducen a un estado seguro?

Si no, no debe automatizarse.

Puerta 6. Impacto

¿Mejora la práctica sin introducir riesgos desproporcionados?

Si no, la complejidad no está justificada.

Puerta 7. Gobernanza

¿Puede controlarse, auditarse, suspenderse y revalidarse?

Si no, el sistema no está preparado para producción.

18. Conclusión del protocolo

La validación de inteligencia artificial en la industria alimentaria debe ejecutarse como un proceso de demostración y no como una recopilación retrospectiva de métricas favorables.

Un protocolo completo debe demostrar que:

- la finalidad está delimitada;
- la referencia es válida;
- los datos representan el uso real;
- el conjunto externo es independiente;
- las métricas corresponden al riesgo;
- los criterios se fijaron previamente;
- la versión fue congelada;
- la incertidumbre está cuantificada;
- los subgrupos críticos cumplen;
- las condiciones adversas no provocan fallos silenciosos;
- el sistema reconoce cuándo no puede decidir;
- el flujo humano es seguro;
- el beneficio frente a la práctica actual es real;
- la organización puede monitorizar, suspender y revalidar.

- **El resultado de la validación no debería expresarse como una certificación absoluta:**

«Modelo validado.»

Debería expresarse como una autorización condicionada:

«La versión identificada del modelo ha demostrado, con la incertidumbre y limitaciones documentadas, un rendimiento compatible con la finalidad prevista dentro del dominio especificado. Su utilización queda condicionada al mantenimiento de las barreras, verificaciones, procedimientos y criterios de suspensión establecidos.»

Esta formulación representa con mayor fidelidad la naturaleza de la evidencia científica. La validación no demuestra que el algoritmo vaya a acertar siempre. Demuestra que, dentro de condiciones determinadas, sus errores son conocidos, cuantificados, suficientemente controlados y compatibles con la decisión que la organización pretende confiarle.

