ESPECIFICACIÓN V1 | GRUPO PRISION RISK-AI

En esta especificación vamos a aclarar los puntos que se tocaron en el debate los cuales enumeramos a continuación, y además, como se trataron temas legales como la IA Act o el RGPD, pues vamos a incluir la información sobre estos puntos de privacidad en ellos. Luego, profundizaremos en 2 temas más los cuales son el Mapa de datos + minimización y, la retención y borrado. Por último dibujaremos 2 diagramas mockup sobre la GUI y la arquitectura de PrisionRisk.

Código accesible / caja negra

PrisonRisk-AI **no toma decisiones automáticas**, sino que **asiste al personal penitenciario** mediante alertas interpretables.
Cada predicción se acompaña de una **explicación clara de sus causas** (factores de movimiento, tono de voz, densidad de grupo), usando técnicas de **IA explicable**.
Todas las alertas quedan registradas y pueden revisarse, garantizando **derecho a reclamación y trazabilidad**, cumpliendo así con la **supervisión humana y transparencia** exigidas por la **IA Act (2024)**.

Supervisión humana

El sistema **siempre funciona bajo control de un funcionario**.
Las alertas son meramente informativas; **la decisión final es humana**.
Esto cumple con el requisito de **supervisión efectiva** de los sistemas de **alto riesgo** de la **IA Act (Anexo III)**.

Sesgos y auditorías

PrisonRisk-AI evita analizar color de piel, tatuajes o vestimenta.
Las imágenes se transforman en **esqueletos digitales anónimos**, donde solo se interpretan posturas y movimientos.
Los datos de entrenamiento se validarán por **consultorías externas** para garantizar **representatividad y ausencia de sesgos**.
Se aplicarán auditorías periódicas y métricas de equidad (*demographic parity*, *equalized odds*).
Además, el modelo se **adapta a cada centro penitenciario**, aprendiendo sus patrones internos sin asociar “peligrosidad” a personas concretas.

Riesgo de sesgo racial

El sistema **no utiliza variables raciales ni étnicas**, y las características procesadas (movimiento, tono, distancia) son **numéricas y neutras**.
Un **comité ético permanente** y auditorías externas garantizarán la ausencia de discriminación indirecta, cumpliendo el **art. 10 de la IA Act**.

Adaptación / posible engaño del sistema

La IA es **multimodal** (visión, audio y contexto) y se **reentrena periódicamente**.
Incluso si se ocultan gestos, detecta otros patrones como tono de voz o proximidad inusual, reduciendo el riesgo de manipulación.

Relaciones entre presos y “presos peligrosos”

El sistema **evalúa contextos, no personas**.
No etiqueta ni clasifica internos, evitando estigmatización y respetando el **principio de reinserción** (art. 25.2 CE).
Los antecedentes o relaciones previas se usan solo como **contexto secundario** para ajustar sensibilidad, nunca como base de decisión.
El modelo **aprende dinámicas internas de cada prisión**, adaptándose a su entorno sin vulnerar derechos.

Legalidad del uso de IA en prisiones

La **IA Act (2024)** no prohíbe estos sistemas, los considera **de alto riesgo** (Anexo III).
PrisonRisk-AI cumple todos los requisitos:

- **Supervisión humana** (art. 14): los funcionarios aprueban toda acción.
- **Transparencia y explicabilidad** (art. 13): cada alerta se justifica.
- **Control de calidad de datos** (art. 10): auditorías externas y validación ética.
- **Documentación técnica y gestión del riesgo** (arts. 9-11).

El tratamiento de datos se ampara en el **interés público y seguridad pública** (art. 6.1.e y 9.2.g RGPD).

Retención y borrado de los datos

Los datos personales se conservarán durante el tiempo necesario para el fin para el que se recogen (art. 5.1.e RGPD), siendo estos de máximo 6 meses para entrenamiento y de máximo 2 años para validación y auditoría, antes de su eliminación o anonimización.

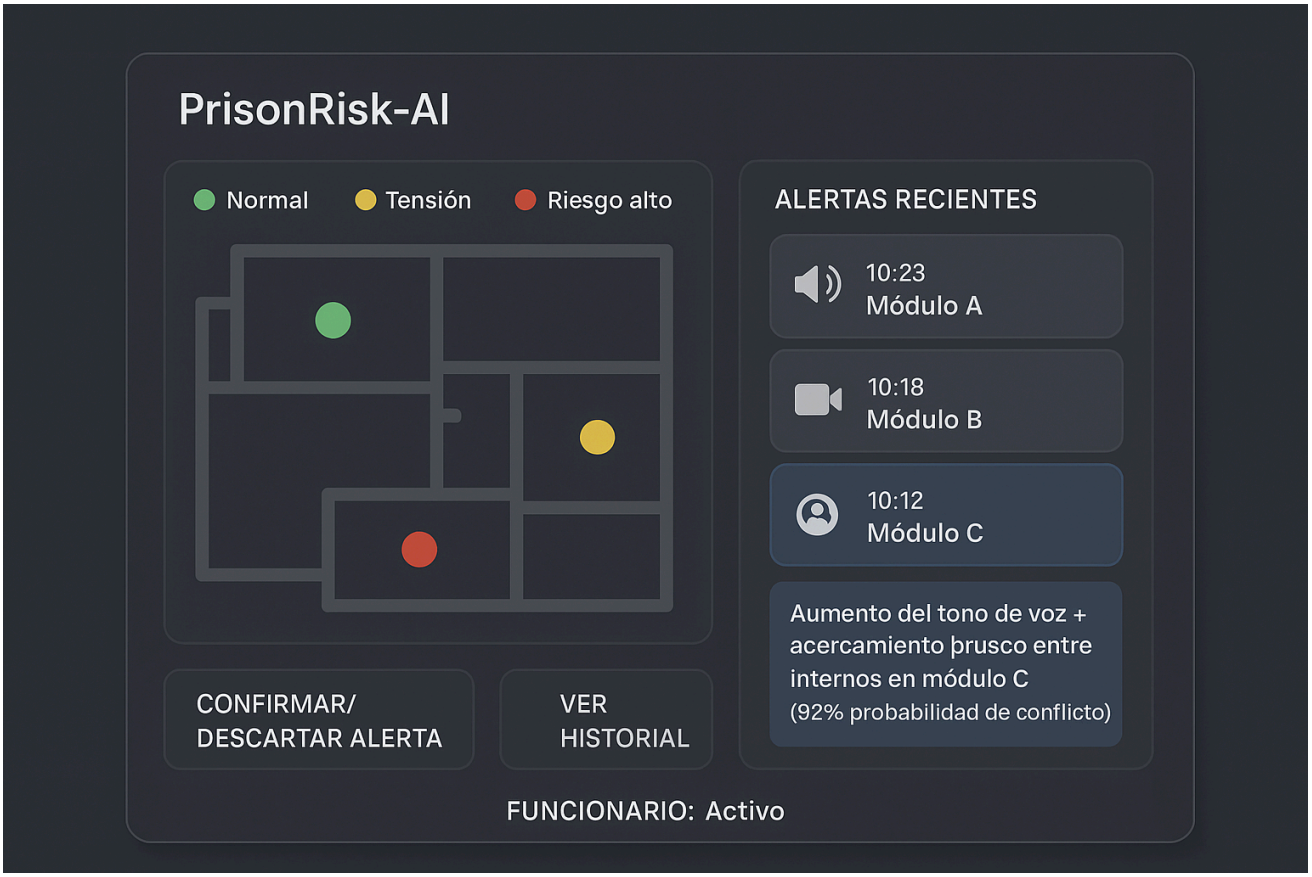
Protección de datos y mapa de minimización

PrisonRisk-AI sigue el principio de minimización del art. 5 RGPD: solo usa datos necesarios para detectar tensiones.

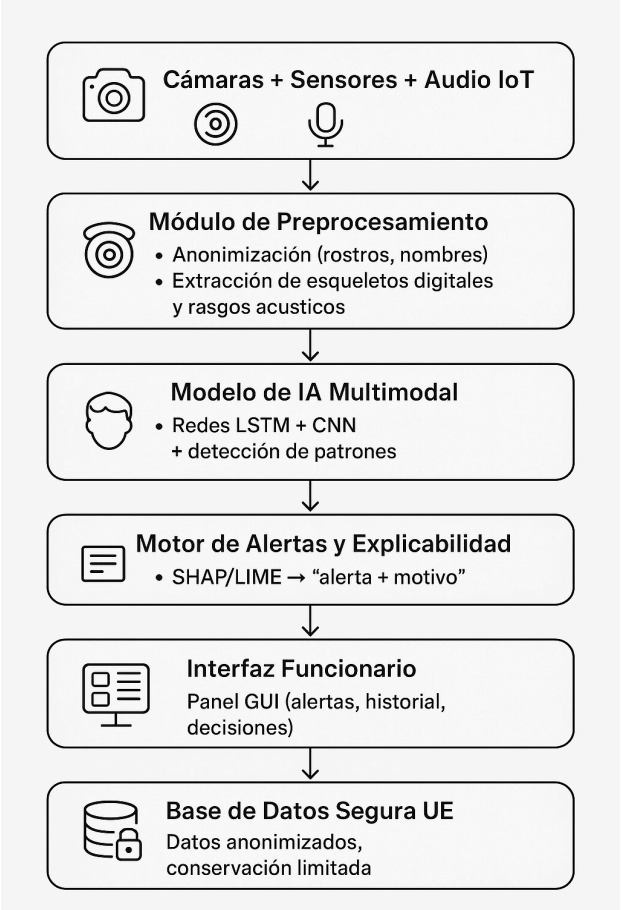
Tipo de dato	**Finalidad**	**Minimización aplicada**
Video vigilancia	Analizar movimientos y distancias	Se extraen esqueletos digitales; sin rostros ni color de piel.
Audio	Detectar picos de ruido o gritos	Solo se procesan rasgos acústicos.
Registros disciplinarios	Entrenamiento y etiquetado	Datos anonimizados sin nombres
Datos ubicación	Analizar proximidades inusuales	Coordenadas relativas sin identificación
Datos psicológicos agregados	Contexto de comportamiento general	Variables codificadas y anónimas

Los datos se alojan **en servidores de la UE**, cifrados, con acceso restringido.
Antes del despliegue se realizará una **Evaluación de Impacto (EIPD)** y supervisión por un **Delegado de Protección de Datos (DPO)**.

Mockup GUI



Mockup Arquitectura



Referencias

RGPD -
https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm

LOPD-RGPD - https://protecciondatos-lopd.com/empresas/cumplimiento-lopd/

IA ACT -
https://maldita.es/malditatecnologia/20240916/ley-inteligencia-artificial-union-europea/

https://artificialintelligenceact.eu/es/ai-act-explorer/

Referencia de los artículos -

https://chatgpt.com/share/68f7ab59-5d14-8000-b170-f54ea94bec97

Información sobre el principio de reinserción -

https://chatgpt.com/share/68f7b0a1-37f8-8000-9179-19111554ef36