EVALUACIÓN DEL MODELO

En GPS-Safe evaluamos el modelo para asegurarnos de que, como sistema de alto riesgo, funcione de forma segura, fiable y ética antes de desplegarse en un entorno real. Vamos a definir la evaluación del modelo según los apartados especificados:

a) ¿Cómo se evalúa que el modelo funcione?

La evaluación se basa en revisar cómo calcula el coeficiente de seguridad de cada tramo de la ruta (como se muestra en la explicabilidad de auditoría). Para ello, combinamos pruebas de rendimiento y supervisión ética, en las que nos aseguramos de que el modelo pueda:

1. Integrar correctamente distintos tipos de datos (como criminalidad, reportes de usuario y contexto)
2. Predecir con alta precisión (el objetivo marcado es de un accuracy > 95%)
3. Minimizar errores críticos. Especialmente, los falsos negativos, que podrían señalar como segura una zona que no lo es.
4. Ser explicable, cumpliendo los requisitos de interpretabilidad (XAI) para que sus decisiones puedan entenderse y revisarse.

b) ¿Qué métricas se definen y usan para testearlo?

Se emplean métricas estándar para evaluar la calidad de las predicciones:

- MAE: mide cuánto se equivoca el modelo de media
- MSE: penaliza más los errores grandes
- F1: combina precisión y recall para evaluar el equilibrio del modelo

Pero la métrica más importante en nuestro caso es la tasa de falsos negativos, ya que clasificar como segura una zona peligrosa constituye el fallo más crítico.

c) ¿Qué factores se están midiendo?

Los factores evaluados cubren rendimiento y comportamiento ético:

1. Precisión: que el modelo calcule bien el nivel de riesgo de cada zona
2. Generalización: que funcione bien con nuevos datos y se mantenga estable a largo plazo
3. Ausencia de sesgos (discriminaciones o patrones injustos). Esto será revisado por el Comité Ético.
4. Interpretabilidad (XAI): que sus decisiones puedan entenderse y justificarse y no actúe como una “caja negra”.

d) ¿Hay algún benchmark existente que se pueda usar de referencia?

Sí. La evaluación aquí combina un objetivo interno y dos referencias científicas externas:

Objetivo interno: Como sistema de alto riesgo, GPS-Safe fija un listón propio muy alto:

- Precisión mínima: >95%
- Requisito crítico: cero falsos negativos aceptables

Referencias externas: Usamos dos estudios como guía metodológica:

1. Predicción de accidentes de tráfico (Random Forest + análisis espacial).

   Explica el ≈78% de la varianza (buen rendimiento). Útil como referencia para combinar ML con análisis de red para rutas seguras.

   Referencia: Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks - ScienceDirect

2. Identificación de rutas seguras según crimen (NLP + datos policiales).

   ≈89% de precisión en clasificación de noticias y ≈84% en identificación de ubicación; alta validación de usuarios (8’75/10). Relevante aquí porque muestra cómo mezclar fuentes de datos para estimar el riesgo urbano.

   Referencia: Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths

En conjunto, estos benchmarks demuestran que la metodología de GPS-Safe (IA + datos mixtos + análisis espacial) es viable, aunque nuestro objetivo interno (>95%) es más exigente que los resultados habituales en la literatura. Sin embargo, como es de cara a un Horizonte 2029, nos mostramos optimistas al respecto.

e) Basadas en estas evaluaciones, benchmarks y testing, ¿qué objetivos debe cumplir el modelo para considerarlo adecuado para salir de la fase de testing a un entorno de producción en entorno real? (e.g. llegar a tal umbral “go/no-go” en tal evaluación)

Para que el modelo se considere adecuado para salir de la fase de testing a un entorno de producción en entorno real, debe cumplir cinco condiciones claras:

1. Umbral de precisión (“Go”): Debe superar el 95% de precisión en las pruebas.
2. Seguridad crítica (“No-Go”): No puede generar falsos negativos.
3. Cumplimiento ético y legal (“No-Go”): El modelo no puede ser una caja negra ni mostrar sesgos. Requiere validación del Comité Ético y cumplir los requisitos mínimos de interpretabilidad.
4. Estabilidad operativa (“No-Go”): No debe presentar fallos graves que puedan llevar a recomendar rutas con riesgo alto o conocido.
5. Supervisión continua (requisito permanente): Debe permitir monitorización y pruebas en funcionamiento para asegurar que el nivel de seguridad se mantiene durante su operación real.

En resumen, el modelo no solo debe alcanzar un alto rendimiento, sino demostrar que es seguro, interpretable y estable para poder pasar a producción.

Analogía sugerida por la IA: Pensar en la evaluación del modelo de GPS-Safe es como someter a prueba un nuevo modelo de airbag en un coche. No solo se mide qué tan preciso es el sensor al detectar un choque (la *accuracy* o la capacidad de procesar variables), sino que el umbral de "Go/No-Go" real es la **eliminación total de fallos de tipo falso negativo** (que el airbag no se despliegue cuando *realmente* se necesita). Además, el coche debe pasar una inspección de seguridad humana (Comité Ético) para asegurar que su diseño no perjudica a ningún conductor de manera discriminatoria (evitando sesgos) antes de que se le permita circular en carretera.