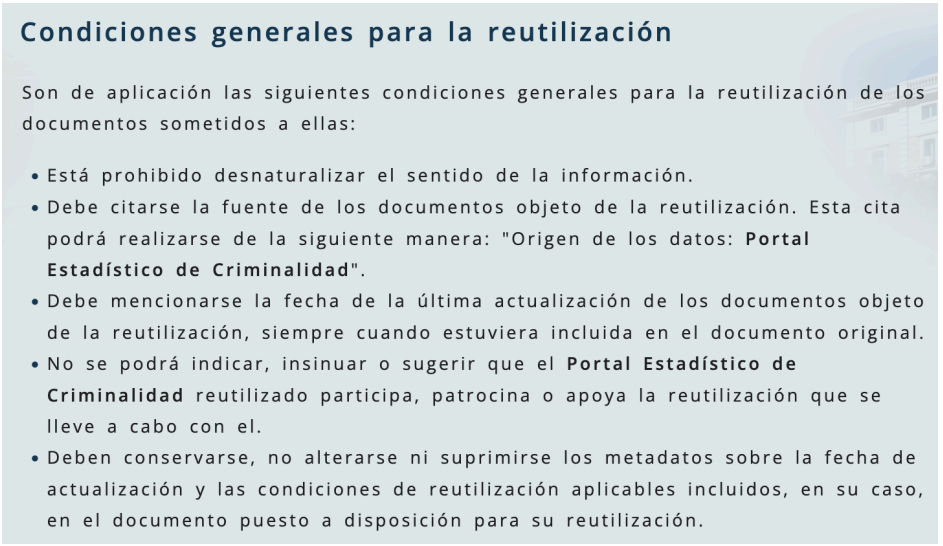
## Entrenamiento del sistema

### 1.1 Datos

El modelo, que es la base fundamental de la aplicación, se usará para el cálculo de rutas seguras en España (nos limitamos a un país por ahora para validar el producto y luego expandirnos internacionalmente). Necesitamos, por tanto, datos de cualquier actividad criminal posible: asesinatos, robos, violaciones sexuales, tráfico de droga… Preferiblemente actualizados cada día y con una localización precisa (al menos el municipio del crimen).

Los datos que usaremos, por lo menos al inicio, serán datos del Ministerio de Interior (no tendríamos datos de reportes todavía), que nos proporcionan información de criminalidad agregada por municipios y trimestralmente (https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/balances).

La licencia de uso/reutilización de dichos datos tiene las siguientes condiciones (sacado de https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/avisoLegal.html):



**Condiciones generales para la reutilización**

Son de aplicación las siguientes condiciones generales para la reutilización de los documentos sometidos a ellas:

- Está prohibido desnaturalizar el sentido de la información.
- Debe citarse la fuente de los documentos objeto de la reutilización. Esta cita podrá realizarse de la siguiente manera: "Origen de los datos: **Portal Estadístico de Criminalidad**".
- Debe mencionarse la fecha de la última actualización de los documentos objeto de la reutilización, siempre cuando estuviera incluida en el documento original.
- No se podrá indicar, insinuar o sugerir que el **Portal Estadístico de Criminalidad** reutilizado participa, patrocina o apoya la reutilización que se lleve a cabo con el.
- Deben conservarse, no alterarse ni suprimirse los metadatos sobre la fecha de actualización y las condiciones de reutilización aplicables incluidos, en su caso, en el documento puesto a disposición para su reutilización.

Para lidiar con el problema de que los datos son publicados con una frecuencia baja (cada 3 meses), usaremos datos de noticias diarias, ya que no estamos utilizando la propiedad intelectual como el texto sino solo extraemos los hechos a partir del texto (https://dudas.derechosdigitales.org/caso/como-pueden-usarse-las-noticias-los-articulos-y-las-fotos-de-prensa/). Para no tener una única fuente de información, nos centraremos en 2 periódicos gratuitos sin muros de pago:

- eldiario.es (https://www.eldiario.es/temas/asesinatos/)
- 20minutos.es (https://www.20minutos.es/tags/temas/asesinatos.html)

A partir de sus avisos legales (20minutos, eldiario), concluimos que hay que ponerse en contacto con dichos periódicos para conseguir la licencia de uso.

En caso de no conseguirlo, podemos usar servicios de agregadores de noticias https://gnews.io/, que proporcionan listado de noticias con encabezados y pequeñas descripciones para uso comercial con una suscripción de pago.

Esta segunda fuente (las noticias), al lanzar la aplicación se podría reemplazar o complementar con los reportes de los usuarios de la app.

**1.2 Evaluación**

- Definir en detalle la **evaluación** del modelo: ¿Cómo se evalúa que el modelo funciona? ¿Qué métricas se definen y usan para testearlo? ¿Qué factores se están midiendo? ¿Hay algún benchmark existente que se pueda usar de referencia? 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)

**1.3 Hardware**

- Estimar el **hardware** necesario para su entrenamiento, si fuera entrenado hoy día, no en 2029. Utilizar varias métricas, incluido potencia, consumo energético, coste económico de alquiler de computación en la nube para los sucesivos entrenamientos. Podéis estimar los FLOPs de entrenamiento -> horas de GPU -> energía (kWh) -> coste en la nube. Tened en cuenta un multiplicador típico en desarrollo (+50-200%) por el coste de iterar mientras hay desarrollo y pruebas. Podéis incluir coste de proveedor (AWS/GoogleCloud/Azure, o coste de e.g. H100) con tarifas actuales. Y en base a todo ello, estimad vuestra huella de carbono (e.g. con alguna herramienta online tipo carbon footprint calculator). Compartid todos vuestros cálculos, con referencias.

Un ejemplo que se puede usar como referencia para ver modelos que hicieron: https://www.sciencedirect.com/science/article/pii/S2590123024009927

Dado que el modelo GPS-Safe combina Random Forest y procesamiento de redes espaciales sobre datos de criminalidad y tráfico, el hardware necesario se puede estimar de la siguiente manera:

## Hardware recomendado

Se recomienda usar al menos una GPU **NVIDIA H100**.
Las H100 son GPUs de última generación, con hasta **989 TFLOPS** en operaciones tensoriales, lo que permite entrenar modelos complejos rápidamente.

## Tiempo estimado

Para entrenar un ciclo completo (50 iteraciones de desarrollo) se estima **10 h de GPU**, considerando un uso intensivo y overhead de operaciones no optimizadas.

Se puede duplicar este tiempo por pruebas, debugging y ajustes durante el desarrollo (+100% margen).

## Coste estimado en AWS

- Instancia con 1 GPU H100 equivalente cuesta ~4,92 USD/h (≈ **4,58 €/h**)
- Por ciclo: 10 h × 4,58 € ≈ **45,8 €**
- Para 50 iteraciones: 45,8 × 50 ≈ **2.290 €**
- Con margen de desarrollo (+100%): hasta **4.580 €**

**Coste eléctrico en España:**

- Potencia GPU: 0,7 kW × 10 h ≈ 7 kWh por ciclo
- Precio electricidad: 0,223 €/kWh → 7 × 0,223 ≈ **1,56 € por ciclo**
- Para 50 iteraciones: 1,56 × 50 ≈ **78 €**
- Con margen: hasta **156 €**

## Huella de carbono

- Energía consumida por ciclo: 7 kWh
- Intensidad de carbono electricidad España: 200 g $CO_2$/kWh
- Emisión: 7 × 200 g ≈ **1,4 kg $CO_2$ por ciclo**
- Para 50 iteraciones: 1,4 × 50 ≈ **70 kg $CO_2$**
- Con margen de desarrollo (+100%): hasta **140 kg $CO_2$**

CIFRAS ANUALES

| Nvidia H100 |
| --- |
| Power Draw:700W |
| Daily Cost:$2.28 |
| Monthly Cost:$68.42 |
| Annual Cost:$832.42 |
| Annual Carbon Cost:2215.18 lbs of $CO_2$ |