Tool: Perfil de datos

El perfil de datos es un análisis exploratorio que brinda información para evaluar la calidad, integridad, temporalidad, consistencia y posibles sesgos de un conjunto de datos que se uti lizará para entrenar un modelo de aprendizaje automático (Gebru et al., 2018).

Fuente de recolección y origen de los datos	
Nombre del conjunto de datos utilizado.	
¿Qué institución creó el conjunto de datos?	
¿Con qué propósito creó la institución el conjunto de datos empleado?	
¿Qué mecanismos o procedimientos se usaron para recoger los datos (por ejemplo, encuesta de hogares, sensor, software, API)? ¿Cumplen la normativa vigente en materia de protección de datos?	
¿Cuál es la escala del conjunto de datos?	
Obtener documentación para cada variable del conjunto de datos. Proporcionar una breve descripción, incluyendo su nombre y tipo, lo que representa, cómo se mide, etc.	

Gobernanza de los datos	
¿Cuál es el dominio de los datos (por ejemplo, propietario, público, personal)?	
Si son personales, ¿los datos están identificados, tienen seudónimos, cuentan con seudónimos no vinculados, son anónimos o agregados?	
Si son privados, ¿se tiene en cuenta los derechos de propiedad intelectual y la protección de datos personales?	

Estructura de los datos	
¿Los datos son estáticos o dinámicos? Si son dinámicos, ¿con qué frecuencia se actualizarán?	
Captar la frecuencia (diaria, semanal, mensual) o el número medio de observaciones por individuo. ¿Qué versión del conjunto de datos se está utilizando?	
¿Es el conjunto de datos más adecuado disponible, teniendo en cuenta el problema en cuestión?	

Calidad de los datos	
¿Cómo se han obtenido los datos (observados, derivados, sintéticos o proporcionados por personas u organizaciones)?	
¿Son los datos representativos de la población de interés?	
Describir el tipo de muestreo utilizado para obtener los datos.	
Analizar la cobertura espacial y temporal de los datos.	
Analizar la cobertura de los grupos protegidos (sexo, raza, edad, etc.).	

Describir las dimensiones importantes en las que la muestra de datos puede diferir de la población, en particular los sesgos de selección no medidos. Utilizar la literatura relacionada con el tema y la información de los expertos.	
Identificar los posibles "estados indeseables" en los datos, que podrían dar lugar a sesgos e inequidades perjudiciales para un determinado subgrupo, o cualquier otro patrón que se considere subóptimo o indeseable desde el punto de vista de la política social.	
¿Hay valores faltantes? Si es así, explicar las razones por las que no se dispone de esa información (esto incluye la información eliminada intencionalmente). Identificar las razones de los datos que faltan y pensar si los datos faltantes se asocian a la variable que va a predecirse. Documentar cualquier proceso de imputación utilizado para sustituir los datos que faltan.	