

PFG Genérico del Departamento de LSI

Grado en Ingeniería en Tecnologías de la Información

Desarrollo de una herramienta para el análisis de sesgos en datasets

El presente proyecto propone el desarrollo de una herramienta orientada al análisis de datasets, con el objetivo de detectar posibles sesgos presentes en los datos. Este tipo de análisis resulta fundamental en numerosos ámbitos de la informática, donde los datos se utilizan como base para el desarrollo de sistemas de análisis, aprendizaje automático o toma de decisiones.

1. Descripción del sistema

La herramienta permitirá cargar datasets estructurados (por ejemplo, en formato CSV o JSON), realizar un análisis de sus características y ofrecer visualizaciones que faciliten la detección de desbalances o distribuciones no representativas.

El sistema seguirá una arquitectura modular. Deberá incluir un conjunto de componentes encargados del procesamiento de datos, el cálculo de métricas estadísticas y la visualización de resultados mediante gráficos y paneles de control (dashboards).

El objetivo final es proporcionar una herramienta que permita explorar datasets de forma interactiva, facilitando la identificación de problemas relacionados con la calidad y el sesgo de los datos.

2. Generación u obtención de datos

Los datos utilizados en el sistema podrán obtenerse mediante la descarga de datasets procedentes de benchmarks como CLEF (<https://clef2026.clef-initiative.eu/labs/>) o TREC (<https://pages.nist.gov/trec-browser/data/>) , así como de repositorios públicos como <https://www.kaggle.com>.

El conjunto de datos deberá ser lo suficientemente amplio y variado como para permitir realizar análisis representativos de las distribuciones y la posible presencia de sesgos.

Se recomienda trabajar con datasets que incluyan tanto variables categóricas como numéricas.

3. Funcionalidades del sistema

El sistema deberá incluir, al menos, las siguientes funcionalidades:

Gestión de usuarios

El sistema deberá permitir un control de acceso y perfiles, con un rol principal:

- Administrador, este usuario se encarga de gestionar los aspectos relacionados con el análisis de los datos y detección de sesgos.

Carga y preprocesamiento de datos

La aplicación deberá contar con un módulo para la gestión de los datos:

- Importación de datasets en distintos formatos.
- Limpieza y tratamiento básico de datos (valores faltantes, formatos, etc.).

Análisis de datos

La aplicación deberá contar con un módulo para el análisis de los datos:

- Cálculo de estadísticas descriptivas (media, mediana, distribución, etc.).
- Análisis de variables categóricas y numéricas.
- Estudio del balance de clases o categorías.

Detección de sesgos

La aplicación deberá tener en cuenta el trabajo con posibles sesgos:

- Identificación de desbalances en variables relevantes.
- Detección de distribuciones no homogéneas.
- Análisis de representatividad de los datos.

Visualización de datos

La aplicación deberá contar con un módulo para la visualización de los datos:

- Generación de gráficos (barras, histogramas, distribuciones).
- Comparación entre distintas variables del dataset.

- Visualización de resultados mediante dashboards interactivos.

Generación de informes

La aplicación deberá contar con un módulo para la generación de informes:

- Resumen automático del análisis realizado.
- Indicadores básicos de calidad del dataset.
- Exportación de resultados (opcional).

4. Casos de uso

A continuación, se describen algunos casos de uso representativos:

- El usuario carga un dataset en la herramienta y accede al módulo de análisis. A continuación, visualiza la distribución de distintas variables y detecta un desbalance en una de las categorías. Mediante los gráficos proporcionados, identifica posibles problemas en la representatividad del dataset.
- El usuario utiliza la herramienta para analizar un dataset antes de emplearlo en un sistema de análisis o aprendizaje automático. A través del informe generado, detecta sesgos que podrían afectar al comportamiento del sistema y decide aplicar técnicas de preprocesamiento.

5. Plan de trabajo

El estudiante deberá elaborar un plan de trabajo acompañado de un calendario realista, en el que se detallen claramente las tareas a realizar y su planificación temporal, concretando así los aspectos que permanecen indefinidos en este documento. Este anteproyecto será revisado y deberá contar con la aprobación del director.

Posteriormente, deberá llevar a cabo el desarrollo del proyecto siguiendo todas las etapas del ciclo de vida del software, e implementar un prototipo de la solución propuesta. La documentación incluirá el diseño técnico, describiendo la arquitectura general del sistema, su organización en módulos o paquetes, la especificación técnica de cada uno de ellos, así como las interfaces entre módulos y con sistemas externos (si procede). También se deberán detallar las principales estructuras de datos y los algoritmos empleados, ya sean propios o de terceros. Igualmente, se presentará un plan de pruebas que permita evaluar la calidad de la solución. Se valorarán especialmente la escalabilidad, la modularidad y la capacidad de adaptación del desarrollo a problemas similares.

La memoria final deberá recoger el diseño técnico indicado en el plan de trabajo y contar una con estructura (orientativa) como la siguiente:

- **Motivación.** En esta sección se presentará el contexto general del proyecto, explicando las razones que justifican su realización y el problema concreto que se pretende abordar
- **Estado del arte.** Se analizarán las soluciones existentes relacionadas con el ámbito de la herramienta, revisando trabajos previos y destacando sus principales aportaciones, así como sus limitaciones.
- **Objetivos.** Se definirán de manera clara los objetivos que se desean alcanzar con el desarrollo del proyecto, tanto a nivel general como específico.
- **Metodología de desarrollo.** Se describirá el enfoque metodológico adoptado para llevar a cabo el proyecto, justificando su elección en función de las características y necesidades del sistema a desarrollar.
- **Descripción informática.** Esta sección detallará los aspectos técnicos del desarrollo, incluyendo la arquitectura del sistema, la organización en módulos, las tecnologías empleadas y las principales decisiones de diseño.
- **Evaluación.** La evaluación del sistema se realizará considerando el correcto funcionamiento de los módulos desarrollados, su capacidad para identificar desbalances y posibles sesgos en los datos, la adecuación de las métricas estadísticas empleadas, así como la claridad y utilidad de las visualizaciones generadas. Asimismo, se validará la herramienta mediante su aplicación a uno o varios datasets reales.
- **Conclusiones.** En esta sección se resumirán los resultados más relevantes obtenidos durante el desarrollo del proyecto, destacando los objetivos alcanzados, las principales dificultades encontradas y posibles líneas de trabajo futuro.

Se deja en manos del estudiante y de su director o directora la definición del diseño y el desarrollo final del proyecto. Aunque la elección de las tecnologías de implementación corresponde al estudiante, esta deberá estar debidamente justificada en la memoria final. La forma de visualizar e interactuar con los contenidos deberá resultar adecuada y adaptarse al dispositivo desde el que se acceda, en caso de optarse por un diseño responsive.

Finalmente, la evaluación del Proyecto de Fin de Grado tendrá en cuenta tanto el proceso de desarrollo y su correcta documentación, como el producto software obtenido y la calidad de la exposición realizada durante la defensa pública.