1- Conjunto/s de datos utilizado durante el desarrollo

ROSAL.IA usa bases de datos mixtas: se generan integrando APIs con información de especies (IEPNB) y literatura científica (Crossreference y Semantic Scholar) beneficiando de las ventajas de ambas informacion robusta y flexibilidad y adaptación a los requerimientos del usuario Listas catalogadas de especie con normativa IEPNB. Fruto de ello surgen dos perspectivas: una más orientada a generar un corpus de conocimiento general y otra de creación adaptativa (a través de peticiones) de bases de datos de informes científicos. En el Github podemos observar una de estas bases de datos generadas después de estas peticiones a modo de ejemplo. por otro lado, tenemos también alojada en Github la base de datos general creada gracias a la implementación en azure. Ambos tipos de bases de datos tienen una gran capacidad de actualización por un lado, debido a que en ambos casos la selección de especies se hace sobre el dataset alojado en la página web y no sobre el excel descargado (estático), si el excel se actualiza, el script igualmente se actualiza. Por otro lado, la búsqueda de la literatura científica se ordena por artículos más recientes en vez de los más relevantes, por lo que un uso programado de estos scripts pueden añadir y actualizar de forma sistemática el corpus de conocimiento.

IEPNB_API = "https://iepnb.gob.es/api/catalogo/v_listapatronespecie_normas"

CROSSREF_API = " https://api.crossref.org/work"

Semantic scholar = " https://api.semanticscholar.org/graph/v1/paper/search?query="

AURA utiliza un conjunto de datos mixto, construido a partir de la integración de listados oficiales de núcleos urbanos, datos de conectividad y rutas de transporte, así como coeficientes institucionales para el cálculo de emisiones de gases de efecto invernadero (GEI). Se combinan, por orden de relevancia:

- Government conversion factors for company reporting of greenhouse gas emissions, publicados por el Department for Energy Security and Net Zero y el Department for Business, Energy & Industrial Strategy del Reino Unido. Este conjunto proporciona factores de conversión oficiales para estimar emisiones a partir de distancias recorridas por diversos medios de transporte, y constituye la base para la modelización de la huella de carbono.

- Bases de datos de rutas de transporte reales, incluyendo:

    - Rutas aéreas europeas de giocomai, utilizadas para limitar la simulación a trayectos plausibles.

    - Horarios y trayectos ferroviarios de Renfe, que permiten modelar la conectividad interurbana dentro de España.

- Listados de ciudades por población y coordenadas geográficas, empleados para construir escenarios realistas de movilidad:

- España: Padrón municipal del INE.
- Europa: "Some 212 million people live in Europe's 500 largest cities" de CityMajors.
- Mundo: Urban Agglomerations estimadas por Naciones Unidas.
- Coordenadas: countries-states-cities-database de Aakash315.

- Airline Passenger Satisfaction de TJ Klein (Kaggle), que permite inferir la distribución de pasajeros por clase (económica, business, primera) y ajustar los cálculos de emisiones en función del tipo de asiento.

2- Justifica la selección del conjunto de datos utilizado, describiendo la metodología empleada para su construcción, así como el análisis realizado sobre su calidad y representatividad respecto al problema abordado.

ROSAL.IA

Selección del conjunto de datos

El objetivo de ROSAL.IA es precisamente el enriquecimiento de la base de datos ya existente del IEPNB, así como su aprovechamiento a nivel de usuario. Por ello, el empleo de la base de datos del IEPNB es indispensable y cumple una función vertebradora sobre la que se ramifican los distintos objetivos de ROSAL.IA. Por otro lado, la asistencia de las APIs de búsqueda de información científica (Crossreference y en menor medida Semantic Scholar) permiten no sólo el enriquecimiento de la base de datos original sino, además, otorgan el potencial para una actualización sistemática y programada del corpus de conocimiento.

Metodología de construcción

Con esta intención inicial, la metodología de ROSAL.IA se planteó para generar una base de datos dinámica a la que poder realizar consultas de forma flexible y usar para la elaboración de informes científicos. Para mantener los estándares y ser fieles a la transparencia y compromiso con el rigor científico se optimizó la búsqueda y limpieza de datos de forma exhaustiva, permitiéndonos extraer una gran cantidad de información de calidad y tener un gran control y seguimiento sobre la misma. La recuperación de datos se realizo mediante recolección automatizada a través de *web scraping* con la librería BeautifulSoup, y mediante el uso de *fetchers* específicos para consultas a APIs como:

- **IEPNB**: para la obtención estructurada de especies protegidas con normativa aplicable.
- **Crossref**: para literatura científica de acceso abierto, seleccionada por su eficiencia y estandarización.
- **Semantic Scholar**: como fuente adicional, empleada ocasionalmente por su capacidad de recuperación basada en modelos de lenguaje,

Fruto de nuestro compromiso con la calidad del dato tambien integramos criterios de evaluación de la calidad de los datos. A continuación se citan las métricas empleadas:

- Recencia: Año promedio de publicación de los artículos.

- Cantidad: Número total de artículos recuperados.

- Precisión: Proporción de artículos 'Exacto' entre todos los artículos con abstract.

- Cobertura: Proporción de artículos que contienen abstract.

- Diversidad temporal: Dispersión en años de publicación (desviación estándar).

Los datos obtenidos también permitienla elaboración de informes mediante el empleo de NLP, adaptados a las necesidades del usuario final a través de una interfaz construida en Streamlit, lo que permite personalizar la visualización y el tipo de información presentada. esta información también se puede considerar una generación de conocimiento y de nuevas bases de datos.

AURA
La construcción del conjunto de datos de AURA se basó en fuentes accesibles y representativas del sector de eventos, incorporando registros de desplazamientos reales, perfiles de asistentes, y especificaciones técnicas de equipos multimedia. La metodología empleada combinó recopilación manual, datos abiertos (por ejemplo, bases de transporte aéreo y ferroviario) y simulaciones controladas para representar distintos tipos de eventos. Se aplicaron técnicas de limpieza, normalización y validación cruzada para garantizar la calidad de los datos. Su representatividad se evalúa en función de la cobertura geográfica, la diversidad de perfiles y la variabilidad de escenarios, asegurando una base robusta para entrenar modelos predictivos con aplicación real.

3- **Indica el enlace a la carpeta de Github donde se ha incluido el conjunto de datos utilizado durante el reto 2**

   https://github.com/AEDI-IA/Ai.dea/tree/main/ROSAL.IA_project

4- **Enlace al repositorio de Github, donde se ha incluido el código desarrollado durante todo el reto 2**

   https://github.com/AEDI-IA/Ai.dea/tree/main/ROSAL.IA_project

   https://github.com/AEDI-IA/Ai.dea/tree/main/Aura_Project

5- **Incluye alguna justificación que consideres adicional**

6- **Describe los modelos de IA empleados, explicando su idoneidad respecto al problema planteado, la arquitectura utilizada y los criterios seguidos para su selección e implementación.**

ROSAL.IA (Repository Of Scientific Articles on Listed species): Inicialmente, se exploró el uso de modelos de la familia RAG (Retrieval-Augmented Generation), incluyendo pruebas con modelos Salamandra 2B, tanto instruidos como no instruidos, a través de la librería Hugging Face Transformers. Sin embargo, en fases posteriores del desarrollo se optó por abandonar

esta estrategia en favor de un enfoque basado en técnicas de NLP clásicas. La decisión se basó en una evaluación de eficiencia: los modelos RAG requerían un consumo computacional considerablemente mayor sin ofrecer mejoras proporcionales en la calidad de los resultados. Por el contrario, el uso de NLP proporcionó resultados más estables, controlables y alineados con los objetivos del proyecto, especialmente en cuanto a rigor científico, velocidad de respuesta y adaptabilidad a distintas necesidades del usuario. La relación entre coste computacional y valor añadido fue determinante para esta elección.

En el marco del desarrollo se exploraron diversas estrategias basadas en IA (fundamentalmente RAG) y DL (principalmente NLP) con el fin de conseguir informes recopilatorios de literatura científica. Las ventajas que presentaba el RAG eran una mayor fluidez y conexión entre distintas piezas de información científica (abstracts), pero era una opción menos robusta a pesar de la implementación de protocolos para controlar las alucinaciones y el acceso regulado a la base de datos a procesar. Por otro lado, el uso de NLP para la síntesis de piezas de resúmenes científicos aportaba un mayor control de los outputs y velocidad de procesamiento, sacrificando parte de la fluidez en favor de un mayor rigor científico, razones por la cuales acabó siendo empleado.

En este sentido, se incorporó el uso de la librería spaCy —con el modelo en_core_web_lg—, que ofrecía un rendimiento notablemente superior para las tareas específicas del proyecto, especialmente en lo relativo a la detección de entidades (nombres científicos y comunes de especies), limpieza de textos, deduplicación de abstracts y extracción de palabras clave. Esta decisión permitió un procesamiento más ligero, rápido y controlado, sin renunciar a la precisión en la interpretación del contenido.

Gracias a continua revisión y planteamiento del proyecto decidimos dividirlo en dos enfoques: uno dirigido hacia la experiencia del usuario y otro cuyo objetivo es la generación de un corpus de conocimiento actualizable para su incorporación en la infraestructura del IEPNB y aprovechamiento por parte del sector público.

AURA nace como una solución práctica para facilitar la toma de decisiones estratégicas por parte de los organizadores de eventos, con el objetivo de reducir y compensar su impacto ambiental. Desde sus primeras fases, el proyecto adoptó un enfoque centrado en el usuario, priorizando el desarrollo de herramientas escalables, accesibles y alineadas con las necesidades reales del sector.

La herramienta integra modelos de aprendizaje automático orientados a estimar la huella de carbono generada tanto por los asistentes como por el transporte de equipos multimedia. Utiliza redes neuronales artificiales (*MLPRegressor*) para predecir las emisiones individuales en función de la procedencia, tipo de vuelo, distancia recorrida y clase de viaje. Además, aplica bosques aleatorios (*RandomForestRegressor*) para calcular las emisiones derivadas del uso de equipos multimedia, considerando variables como el número de pantallas, altavoces, consumo energético y duración. Ambos modelos se integran en pipelines de procesamiento que combinan datos numéricos y categóricos, seleccionados por su capacidad para adaptarse a relaciones no lineales, manejar datos heterogéneos y ofrecer un buen equilibrio entre precisión, interpretabilidad y eficiencia computacional.

AURA ha sido diseñada bajo principios de modularidad y escalabilidad, con una arquitectura basada en módulos independientes que permiten incorporar nuevas fuentes de datos —como

catálogos urbanos, conjuntos externos o parámetros personalizados— de forma ágil. Esta flexibilidad garantiza su adaptabilidad a distintos tipos de eventos y facilita su evolución.

En esta fase, el desarrollo se ha centrado en el análisis cuantitativo de la movilidad de los asistentes, el transporte de equipos multimedia y los mecanismos de compensación mediante reforestación. En etapas futuras, se prevé ampliar el análisis con nuevas dimensiones como los hábitos de transporte, el tipo de alimentación ofrecida y los productos promocionales utilizados, enriqueciendo así la evaluación cualitativa del impacto ambiental.

La selección de modelos y tecnologías se basó en criterios de equilibrio entre precisión, eficiencia y usabilidad. Debía ser operativa en contextos reales, incluso con recursos limitados, y ofrecer resultados comprensibles para perfiles técnicos y no especializados. No se limita a cuantificar el impacto, sino que impulsa un cambio de mentalidad, proporcionando métricas accionables que promueven prácticas más responsables.

En síntesis, AURA representa una convergencia entre innovación tecnológica y compromiso ambiental. Más allá de una plataforma de cálculo, actúa como un aliado estratégico para fomentar una cultura de sostenibilidad en la organización de eventos, promoviendo decisiones informadas y responsables.

7- ¿Se utilizan modelos de la familia ALIA?

No

8- En caso de no utilizar modelos de la familia ALIA, argumenta por qué.

Inicialmente, se exploró el uso de modelos de la familia RAG (Retrieval-Augmented Generation), incluyendo pruebas con modelos Salamandra 2B, tanto instruidos como no instruidos, a través de la librería Hugging Face Transformers. Sin embargo, en fases posteriores del desarrollo se optó por abandonar esta estrategia en favor de un enfoque basado en técnicas de NLP clásicas. La decisión se basó en una evaluación de eficiencia: los modelos RAG requerían un consumo computacional considerablemente mayor sin ofrecer mejoras proporcionales en la calidad de los resultados. Por el contrario, el uso de NLP proporcionó resultados más estables, controlables y alineados con los objetivos del proyecto, especialmente en cuanto a rigor científico, velocidad de respuesta y adaptabilidad a distintas necesidades del usuario. La relación entre coste computacional y valor añadido fue determinante para esta elección.

9- Incluye en la siguiente casilla el enlace a la carpeta de GitHub, donde se han incluido las mediciones del consumo energético

https://github.com/AEDI-IA/Ai.dea/blob/main/logger_carbon.py

10- ¿En qué fases del desarrollo se ha realizado la medición energética?

Durante todo el desarrollo, el proceso se implementa mediante un script Python independiente, basado en la librería CodeCarbon, que extrae de los logs la cantidad de CO_2 registrada y calcula la suma acumulada de las emisiones, analizando más de 250 logs. En total, se registró una emisión acumulada de 0.412776389512317 kg CO_2eq.

11- Escribe el valor numérico del coste en €uros que se han consumido en vuestra cuenta de Microsoft Founders (entre 0 ~ 5000).

El valor numérico 141.12$ se encuentra el desplegable en el directorio Github.

12- ¿Hay alguna otra información adicional que te gustaría comentarnos (Opcional)?

Sugerimos una formación grupal o píldora formativa sobre el uso de la plataforma Azure para mejorar su aprovechamiento en proyectos como este.