DIAPOSITIVA 6

Pasamos ahora a la **Evaluación de la Reparación**. Aquí respondemos a la pregunta clave: ¿Son fiables los datos que hemos reconstruido?

En la gráfica pueden ver un ejemplo real del sensor C309 midiendo temperatura. La línea azul muestra el DATO INICIAL DONDE SE HA RECORTADO un tramo plano o perdido, y los puntos rojos son nuestra reconstrucción. Como ven, el modelo ha sido capaz de recuperar perfectamente la **tendencia diurna y nocturna**, basándose en la correlación con sensores vecinos.

Nuestras métricas validan este éxito:

1. Tenemos un **Error Absoluto Medio (MAE) de solo 0.25 grados**, lo que significa que la desviación entre nuestro cálculo y la realidad es mínima.
2. El **R^2 Score medio es de 0.96** , confirmando una altísima precisión gracias a la multicolinealidad entre variables cíclicas.
3. Nuestra PRECISIÓN ES PRÁCTICAMENTE LA RESOLUCIÓN DEL SENSOR

Un punto importante es la **Tasa de Recuperación del 7.40%**. Puede parecer baja, pero es una decisión de diseño: nuestro sistema es conservador. Preferimos descartar un hueco irrecuperable antes que inventar un dato falso. Solo exportamos registros cuando la **continuidad física** y estadística está garantizada

DIAPOSITIVA 7

Avanzamos hacia nuestra **Gobernanza de Datos**. En este proyecto, hemos establecido una 'Política Integral de Calidad' muy estricta: no tratamos todos los fallos igual. Para lograr la máxima fiabilidad, hemos diseñado un **Pipeline de Reconstrucción Diferenciada que actúa en tres capas en cascada**.

Esto significa que, ante un dato perdido, el sistema intenta recuperarlo pasando por tres motores de inteligencia, del más específico al más general:

1. La Primera Capa es el Motor de Imputación Unidimensional. Este es nuestro especialista. Lo usamos exclusivamente para **variables cíclicas** que sabemos que están muy correlacionadas entre sí, como la temperatura.

- *Se recuperan los sensores vecinos desde el grafo*
- *¿Cómo funciona?* Aprende la relación directa entre la variable que falta y la misma variable en los sensores vecinos.
- Si el sensor de al lado sube de temperatura, el sistema sabe cuánto debe subir este. Es muy preciso para fenómenos físicos compartidos de ahí el R^2 de 0.99

2. La Segunda Capa es el Motor de Imputación Multidimensional. Si el análisis anterior no es aplicable, activamos este motor más complejo. Aquí ya no miramos una sola variable, sino **todas a la vez**.
+1

- El sistema aprende relaciones cruzadas: por ejemplo, cómo el pH y la conductividad de los vecinos pueden explicar el oxígeno disuelto que nos falta.
- Para garantizar la seguridad, este motor tiene un 'candado': solo actúa si valida que la capacidad de predicción (el R^2 Score) es superior a **0.70**. Si no hay certeza matemática, no imputa.

3. La Tercera Capa es la Imputación por Interpolación. Esta es nuestra red de seguridad final. Se utiliza solo para **ventanas de tiempo muy pequeñas** que han quedado aisladas o 'huérfanas'.

- Aquí aplicamos matemáticas lineales o cúbicas para unir los puntos extremos de un hueco breve. Es una reparación quirúrgica para mantener la continuidad de la serie sin inventar tendencias complejas.

El Resultado: Gracias a aplicar esta cascada lógica y no un simple 'relleno automático', hemos conseguido recuperar más de **7,800 puntos de datos críticos** con una calidad excepcional: mantenemos un **R^2 Score global superior a 0.98**.
+1

Esto es lo que llamamos nuestro **Estándar de Integridad Crítica**: datos recuperados en los que podemos confiar ciegamente para la toma de decisiones.

El resto de datos no son recuperables pero no es todo culpa nuestra, la administración ha dejado muchos sensores sin mantenimiento y existen ventanas de apagado de incluso meses

Autorizados a realizar vertidos

Dejamos atrás la limpieza técnica de las series temporales y pasamos a entender el contexto físico y administrativo: **¿Quién está autorizado a verter en nuestros ríos y dónde lo hace?**

En esta diapositiva 8, lo que ven en pantalla no es una simple lista, sino un **Grafo de Relaciones**. Hemos modelado la base de datos de autorizaciones como una red compleja para entender las conexiones entre titulares, ubicaciones y tipos de vertido.

Quiero destacar tres hallazgos clave que definen la complejidad del problema al que se enfrenta nuestra IA:

1. Heterogeneidad de las Fuentes Aunque solemos pensar solo en depuradoras, nuestro análisis revela un ecosistema muy diverso. Tenemos desde grandes Estaciones de Depuración de Aguas Residuales (EDAR) municipales e industriales, hasta complejos agroindustriales y operaciones minero-energéticas.

- Sin embargo, el dato es contundente: la inmensa mayoría, **1,711 de los 1,835 casos analizados**, son de naturaleza **'Urbano o Asimilable'**. Esto nos indica que el patrón principal que debe aprender el modelo es el ciclo de vida de las aguas residuales urbanas.

2. Concentración Geográfica y el 'Punto Ciego' del Terreno Geográficamente, **Madrid** es el foco principal, acumulando la mayor frecuencia de registros (585).

- Pero aquí surge un desafío técnico interesante: el medio receptor más común no es siempre el 'Río', sino el **'Terreno'**, con 912 casos. Esto complica la trazabilidad directa, ya que muchos vertidos se filtran o no van a un cauce principal inmediato, lo que justifica por qué usamos grafos para trazar relaciones indirectas mediante una proyección de la posición del autorizado hacia el río y recorriendo aguas abajo los datos.

3. La Gran Dispersión de Volúmenes (El problema de la Escala) Finalmente, quiero que se fijen en la estadística de volumen. Tenemos una media de casi un millón de metros cúbicos (914,000 m^3), pero una **desviación estándar gigantesca**.

- Esto significa que conviven pequeños ayuntamientos con gigantes como el **Canal de Isabel II o la cervecera Mahou**
- Para nuestro sistema de inferencia, esto es crítico: no podemos tratar igual a un pequeño vertido rural que a una gran infraestructura metropolitana. El modelo necesita entender esta **desigualdad de escala** para no generar falsos positivos en zonas de bajo caudal.

En resumen, este mapa nos permite saber **dónde mirar y qué esperar** antes de que el sensor detecte la primera anomalía

Autorizados a realizar vertidos

Si en la diapositiva anterior vimos *quiénes* son, en esta analizamos *cuánto* vierten y *dónde* ocurren las anomalías. Los datos nos revelan una realidad extremadamente asimétrica que condiciona toda nuestra estrategia de vigilancia.

Quiero destacar tres conclusiones fundamentales que hemos extraído del análisis estadístico y de explicabilidad (xAI):

1. La Ley del 1% (Concentración Extrema) Fíjense en la gráfica de la izquierda, la **Curva de Lorenz**. Nos muestra un desequilibrio radical: **el 1% de los titulares autorizados es responsable del 80% del volumen total vertido** ,.
+1

- Esto es un hallazgo operativo crítico. Significa que no necesitamos vigilar con la misma intensidad los 1,800 puntos. Si controlamos estrictamente a ese **1% de grandes actores**, tenemos controlado el 80% del caudal de agua residual del sistema. Es un problema de 'pocos actores con mucho impacto'.

2. El Punto Crítico del Sistema Al bajar al detalle territorial, los datos señalan un claro 'punto caliente': el tributario con mayor presión por contaminación es el **río Jarama, específicamente en la zona de Rivas**.

- Esto valida lo que comentábamos antes: aunque el perfil típico es un vertido mediano en Madrid, la acumulación en ciertos afluentes crea zonas de sacrificio que requieren sensores dedicados.

3. Lecciones de la IA: El Falso Positivo Geográfico Por último, observen el **Mapa de Anomalías** a la derecha. Los puntos rojos indican qué vertidos el modelo consideró 'anómalos'.

- Aquí nos llevamos una sorpresa. Usando técnicas de **Explicabilidad (SHAP)**, descubrimos que el modelo estaba marcando como anomalías vertidos en **Badajoz** simplemente por estar muy lejos del núcleo de Madrid ,.
  +1
- Como la gran mayoría de datos son de Madrid, la IA aprendió erróneamente que 'lo normal es estar en Madrid' y que 'estar lejos es sospechoso'.
- Esto no es un error de los datos, sino un **sesgo de contexto**. Gracias a este análisis, hemos podido recalibrar el sistema para que entienda que un vertido lejano no es necesariamente ilegal, evitando alertas falsas en zonas rurales periféricas