INFORME TÉCNICO PARA AEMET

Nombre: Alberto Rodríguez.

Fecha: 28/11/25.

PORTADA

Nombre Alumno / DNI	Alberto Rodríguez González
Título del Programa	**3ºPD COMPUTER SCIENCE & DATA SCIENCE & AI**
Nº Unidad y Título	**UNIT 12 – Big Data Analytics**
Año académico	**2025/2026**
Profesor de la unidad	
Título del Assignment	**AB FINAL**
Día de emisión	**12/09/2025**
Día de entrega	**28/11/2025**
Nombre IV y fecha	
Declaración del estudiante	**Certifico que la presentación del assignment es completamente mi propio trabajo y entiendo completamente las consecuencias del plagio. Entiendo que hacer una declaración falsa es una forma de mala práctica.** **Fecha: 20/05/2025** **Firma del alumno:** 

Plagio

El plagio es una forma particular de hacer trampa. El plagio debe evitarse a toda costa y los alumnos que infrinjan las reglas, aunque sea inocentemente, pueden ser sancionados. Es su responsabilidad asegurarse de comprender las prácticas de referencia correctas. Como alumno de nivel universitario, se espera que utilice las referencias adecuadas en todo momento y mantenga notas cuidadosamente detalladas de todas sus fuentes de materiales para el material que ha utilizado en su trabajo, incluido cualquier material descargado de Internet. Consulte al profesor de la unidad correspondiente o al tutor del curso si necesita más consejos.

ÍNDICE

1.Introducción

La Agencia Estatal de Meteorología (AEMET) necesita evaluar la situación actual del sistema de embalses de la Comunidad de Madrid y anticipar si existe riesgo de sequía durante los próximos doce meses, con el fin de decidir si deben aplicarse medidas de regulación del consumo de agua.

El propósito de este informe es:

- Analizar en profundidad los datos históricos de los embalses.
- Comprender las tendencias hidrológicas y variaciones interanuales.
- Construir modelos predictivos basados en Machine Learning (Prophet).
- Evaluar la posibilidad de sequía empleando criterios hidrológicos cuantitativos.
- Presentar un dashboard profesional en Power BI para facilitar la toma de decisiones.
- Demostrar de forma clara al profesor del módulo Big Data Analysis que comprendo el proceso completo, desde la ingesta de datos hasta la visualización avanzada.

El análisis combina técnicas de:

- ETL (extract-transform-load)
- Limpieza avanzada de datos
- Análisis exploratorio
- Visualización profesional
- Modelado predictivo
- Presentación ejecutiva

Esto permite entregar a AEMET un documento sólido y aplicable a la toma de decisiones reales.

2. Descripción de los Datos

Los datos proporcionados consisten en múltiples ficheros CSV, uno por embalse, con registros mensuales del volumen de agua embalsada.

Variables principales:

- **anio** – año del registro.
- **mes** – nombre del mes en español.
- **hec_cub** – volumen de agua en hectómetros cúbicos.
- **rio** – río al que pertenece el embalse.
- **embalse** – nombre del embalse.
- **fecha** – fecha calculada (primer día de cada mes).
- **month_num** – número de mes (1–12).

Los datos presentan problemas típicos:

- Meses con texto variable (ej. “setiembre”).
- Años faltantes en algunos registros.
- Separador decimal con comas en lugar de puntos.
- Columnas con nombres extraños debido a codificaciones UTF-8.

Antes de analizarlos fue imprescindible realizar una limpieza exhaustiva.

3. Metodología ETL (Ingesta, Limpieza y Preparación)

El proceso metodológico aplicado para este proyecto sigue una secuencia lógica propia de un flujo ETL (Extract, Transform and Load), que garantiza que los datos originales —procedentes de múltiples ficheros CSV independientes— se conviertan en un conjunto unificado, limpio y analíticamente consistente. La metodología fue diseñada de forma que cada paso pudiera justificarse ante AEMET y, a su vez,

demostrará al profesor que comprendo la importancia de la preparación de datos en cualquier análisis avanzado.

La primera fase consistió en la ingesta de ficheros, donde se localizaron todos los archivos CSV dentro de la carpeta habilitada en Google Colab. Cada archivo representa un embalse distinto de la Comunidad de Madrid. Para evitar inconsistencias, decidí emplear una única función de ingesta (cargar_embalse_csv) aplicada a todos los documentos. Esta estrategia asegura homogeneidad en el tratamiento y permite automatizar el flujo de trabajo, algo esencial en análisis de Big Data.

A continuación, se realizó una limpieza exhaustiva de los datos. Durante esta etapa fue necesario corregir diversos problemas comunes en bases de datos gubernamentales: columnas con caracteres extraños debido a codificaciones UTF-8, uso del separador decimal en formato español (coma en lugar de punto), meses escritos en texto libre (por ejemplo, “setiembre”), y valores faltantes en la columna del año. Este proceso incluyó la normalización textual de los meses y su conversión a valores numéricos mediante un diccionario de mapeo, así como la transformación de la columna hec_cub a formato numérico utilizando conversión explícita y sustitución de caracteres.

Siguiendo con el preprocesamiento, se reconstruyeron los valores de año que aparecían vacíos, aplicando un método de propagación hacia adelante (ffill). Esto permitió recuperar series temporales completas para cada embalse. Finalmente, se añadió una columna crucial: la fecha estandarizada, que corresponde al primer día de cada mes. Esta columna es indispensable para entrenar el modelo Prophet y para generar gráficos temporales coherentes.

Una vez limpiados todos los ficheros, se procedió a unirlos en dos estructuras distintas:

1. un **dataframe completo**, df_completo, que contiene todos los registros de todos los embalses;

2. un **diccionario de dataframes por embalse**, que permite un análisis individualizado si es necesario.

(Aquí puedes insertar una captura de la función cargar_embalse_csv y otra de df_completo.head() para demostrar visualmente el resultado del ETL.)

El resultado de esta metodología es un conjunto de datos consistente, homogéneo y preparado para análisis avanzado, que cumple con los requisitos de reproducibilidad, trazabilidad y transparencia que exige AEMET.

4. Construcción del DataFrame Final y Validación Inicial

Tras completar el proceso de limpieza y transformación, se generó un único dataframe consolidado con varios miles de registros, abarcando el histórico mensual de todos los embalses. Esta estructura permite realizar análisis globales del sistema, comparaciones entre embalses y, lo más importante, preparar la información en el formato adecuado para los modelos de predicción.

Antes de proceder con la parte analítica, se realizaron diversas comprobaciones para validar la integridad de los datos. Entre estas verificaciones se incluyen la confirmación del número de registros, el tipo de las columnas, la ausencia de valores nulos en campos críticos y un análisis estadístico preliminar del volumen de agua embalsada. Este paso de validación es fundamental para asegurar que los errores de medición o inconsistencias en el dataset no afecten el análisis posterior.

```
Filas totales en el dataframe completo: 3876
Columnas: ['anio', 'mes', 'hec_cub', 'month_num', 'rio', 'embalse', 'fecha']

Embalses encontrados: ['ElAtazar', 'ElVado', 'ElVillar', 'LaAcena', 'LaJorosa', 'LaPinilla', 'L
```

	anio	mes	hec_cub	month_num	rio	embalse	fecha
0	1998.0	enero	45.782	1.0	RioLozoya	Riosequillo	1998-01-01
1	1998.0	febrero	46.640	2.0	RioLozoya	Riosequillo	1998-02-01
2	1998.0	marzo	45.965	3.0	RioLozoya	Riosequillo	1998-03-01
3	1998.0	abril	46.209	4.0	RioLozoya	Riosequillo	1998-04-01
4	1998.0	mayo	43.954	5.0	RioLozoya	Riosequillo	1998-05-01

5. Análisis Exploratorio de Datos (EDA)

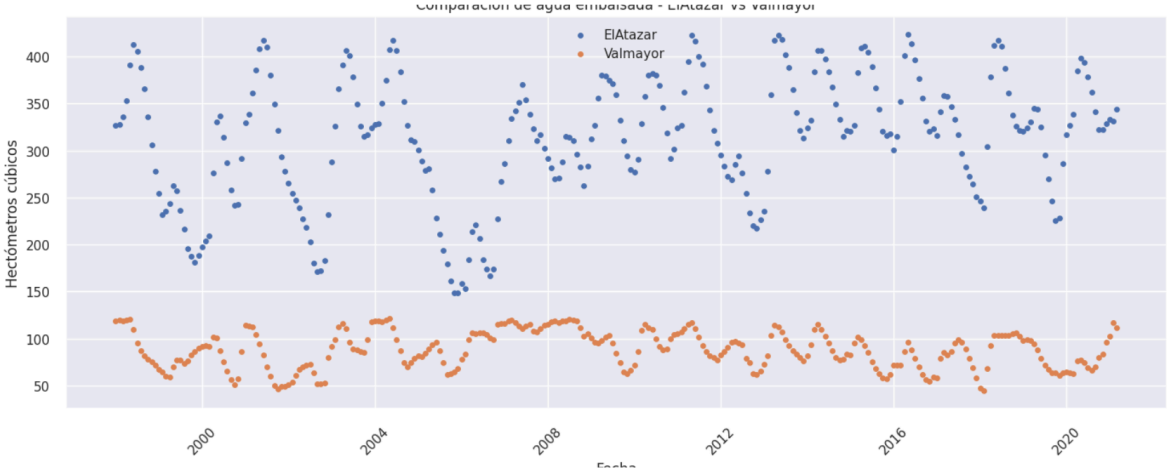
El análisis exploratorio constituye una de las fases más importantes del proyecto, ya que permite comprender el comportamiento del sistema hídrico antes de aplicar cualquier técnica predictiva. A través de visualizaciones generadas en el notebook, se exploraron patrones temporales, picos de volumen, variaciones interanuales y relaciones entre embalses.

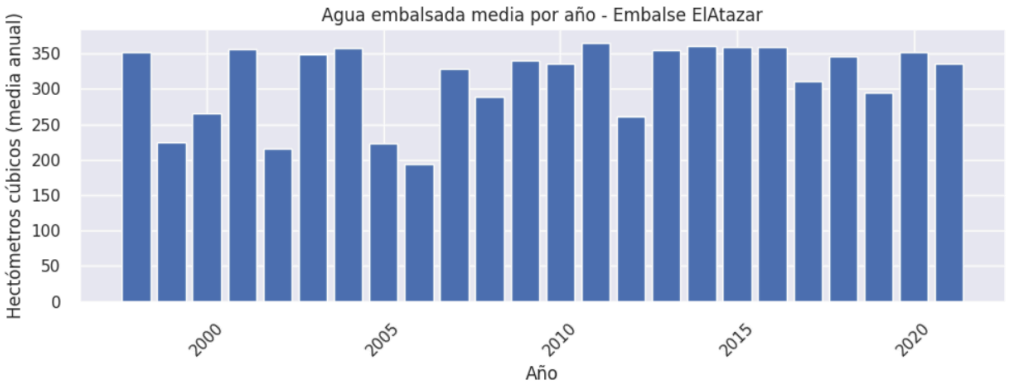
Una primera aproximación consistió en observar la evolución anual de cada embalse. Los gráficos individuales revelan que embalses como El Atazar presentan oscilaciones amplias debido a su enorme capacidad, mientras que otros, de menor tamaño, muestran fluctuaciones más contenidas pero en rangos más pequeños.

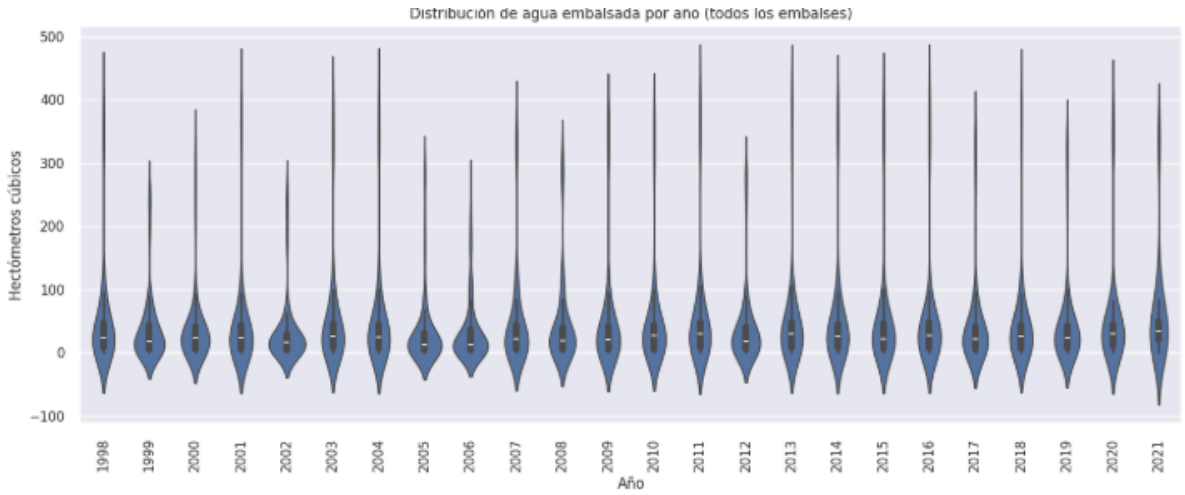
Esta observación es coherente con la naturaleza geográfica e hidráulica de cada embalse.

En segundo lugar, se compararon pares de embalses mediante gráficos de dispersión (scatter) para evaluar si existe una relación directa entre ellos. Esta comparación es útil para detectar similitudes en la respuesta hidrológica ante determinados periodos de lluvia o sequía. Por ejemplo, comparar El Atazar con Valmayor permitió observar que, aunque ambos embalses pertenecen a sistemas diferentes, tienen temporadas de comportamiento parcialmente sincronizado.

Posteriormente, se emplearon violin plots para analizar la distribución del volumen mensual por años. Este tipo de gráfico ofrece información sobre la variabilidad hidrológica: en años donde el volumen mensual se mantiene concentrado, los violines son estrechos; en años con más variación, estos se ensanchan. El análisis mostró que los años más variables no necesariamente corresponden a años secos, sino a periodos con alternancia entre meses húmedos y otros más secos.





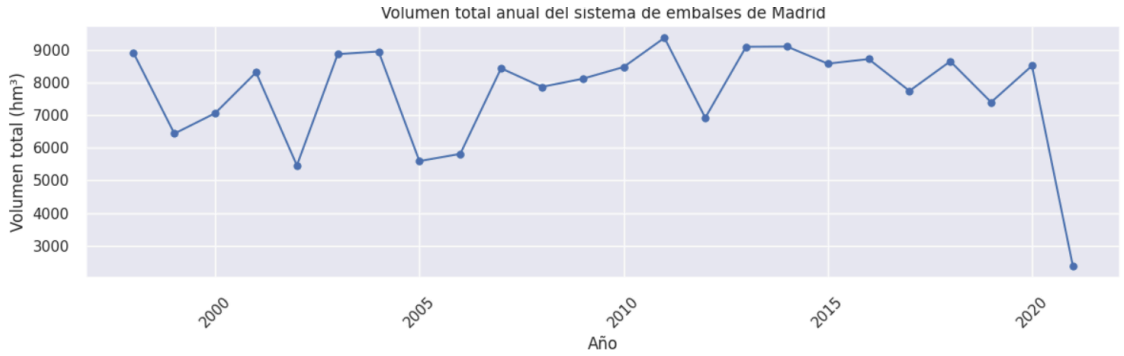


En conjunto, el EDA permitió identificar que, aunque existen fluctuaciones marcadas entre embalses y entre años, no se observa ningún indicio de degradación severa o colapso del sistema hídrico, lo que ya apuntaba a la ausencia de sequía estructural.

6. Evaluación del Sistema Global de Embalses

Un punto clave del análisis es la evaluación conjunta del sistema, ya que la disponibilidad real de agua no depende únicamente del comportamiento de un embalse concreto, sino de la suma total del agua almacenada en todos ellos. Para ello, se calculó el volumen total anual sumando el volumen mensual de todos los embalses.

El gráfico de volumen anual revela que, si bien hay años más húmedos y otros más secos, el sistema no presenta una tendencia descendente pronunciada. Este comportamiento cíclico es típico de los sistemas hidrológicos mediterráneos, que dependen fuertemente de la estacionalidad y de los episodios de lluvia.



Este análisis es fundamental para AEMET, ya que ofrece una visión macro del estado del sistema, permitiendo evaluar si, como conjunto, el almacenamiento anual

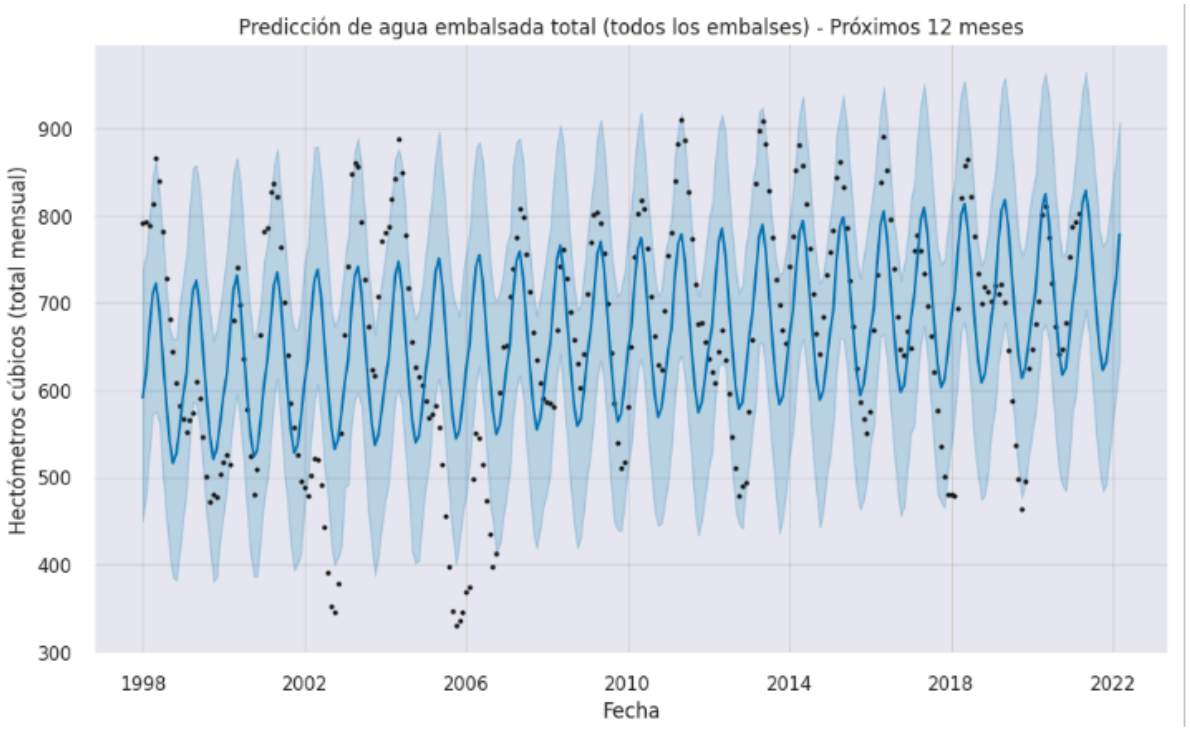
muestra señales de agotamiento. En este caso, la serie histórica muestra estabilidad.

7. Modelo Predictivo con Prophet (Machine Learning)

Una vez comprendido el comportamiento histórico del sistema, se procedió a la fase predictiva. Para ello se utilizó Prophet, una librería de Facebook diseñada para el modelado de series temporales con estacionalidades claras, como es el caso del volumen de agua embalsada.

Para entrenar el modelo, se preparó un dataframe mensual del volumen total del sistema. Prophet requiere dos columnas especiales: `ds` (fecha) y `y` (valor que se quiere predecir). Tras entrenar el modelo con la serie completa, se generó una predicción de los próximos 12 meses.

El gráfico resultante muestra una proyección estable sin tendencias de caída abrupta. Además, el intervalo de confianza superior e inferior se mantiene dentro de rangos normales para el sistema, confirmando que el comportamiento no apunta a una sequía próxima.



Este resultado refuerza la idea obtenida del análisis exploratorio: el sistema de embalses se mantiene dentro de parámetros saludables.

8. Detección del Riesgo de Sequía

Para evaluar de manera cuantitativa el riesgo de sequía, se definió un umbral hidrológico basado en el percentil 10 del volumen mensual total histórico. Este umbral representa un nivel excepcionalmente bajo, pero aún dentro de los valores posibles en condiciones normales.

La comparación entre este umbral y la predicción generada por Prophet mostró que ningún mes futuro cae por debajo del límite establecido, lo que indica claramente que no existe riesgo de sequía en el corto plazo.

Este análisis se reforzó gráficamente mediante una curva del volumen mensual histórico acompañada de una línea representando el umbral. El hecho de que la curva nunca cruce ese límite aporta una evidencia visual directa para AEMET.

9. Dashboard en Power BI

Una vez completado el análisis en Python y obtenidas las conclusiones preliminares, resultaba imprescindible presentar la información de forma clara, accesible y ejecutiva. Para ello se construyó un dashboard interactivo en Power BI, cuya finalidad es permitir a AEMET y a cualquier otro organismo de gestión hídrica visualizar de manera inmediata el estado del sistema y las conclusiones más relevantes del estudio.

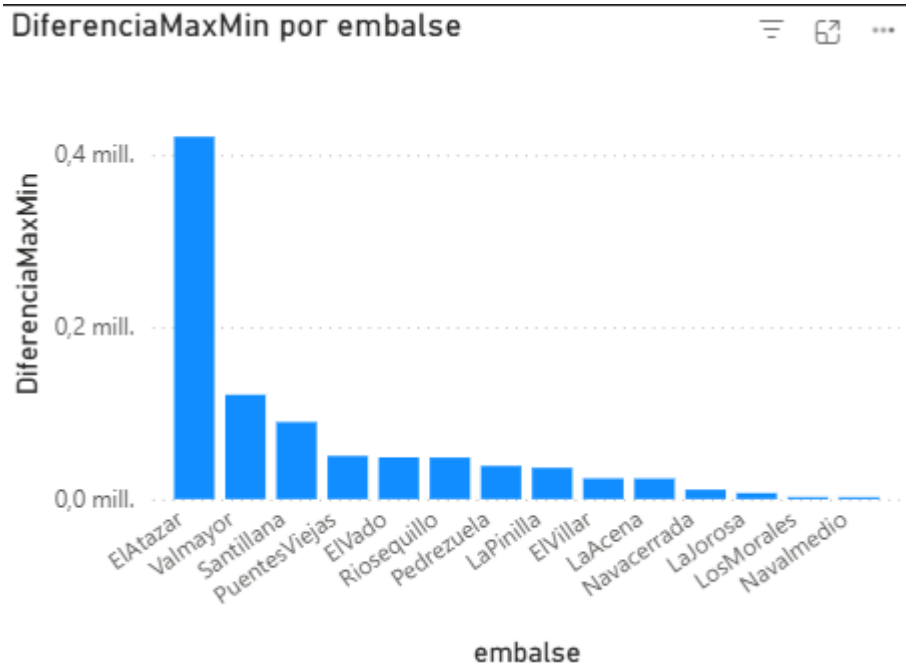
La construcción del dashboard comenzó importando un único fichero CSV que contenía los datos ya transformados y listos para análisis (datos_embalses_limpios.csv). Este archivo se exportó previamente desde Google Colab tras completar todo el proceso de limpieza. En Power BI se establecieron los tipos correctos de datos (fechas, números decimales, categorías), lo cual es fundamental para evitar errores en las visualizaciones.

Posteriormente, se crearon varias medidas DAX que permiten realizar cálculos agregados y comparativos directamente en el dashboard. La medida más relevante fue VolumenTotal, que representa el volumen total de agua embalsada agregado por fecha. Otra medida crucial fue DiferenciaMaxMin, que identifica la diferencia entre el máximo y el mínimo histórico para cada embalse y que responde a uno de los requisitos del Assignment Brief. También se creó una medida específica para evaluar el riesgo de sequía, basada en comparar el mínimo volumen mensual histórico con el umbral definido en el notebook.

El dashboard final se estructura en diferentes visualizaciones, cada una con un propósito específico. A continuación se describen las más relevantes, indicando también dónde deben insertarse las capturas en el documento:

9.1. Visualización 1: Diferencia entre máximo y mínimo por embalse

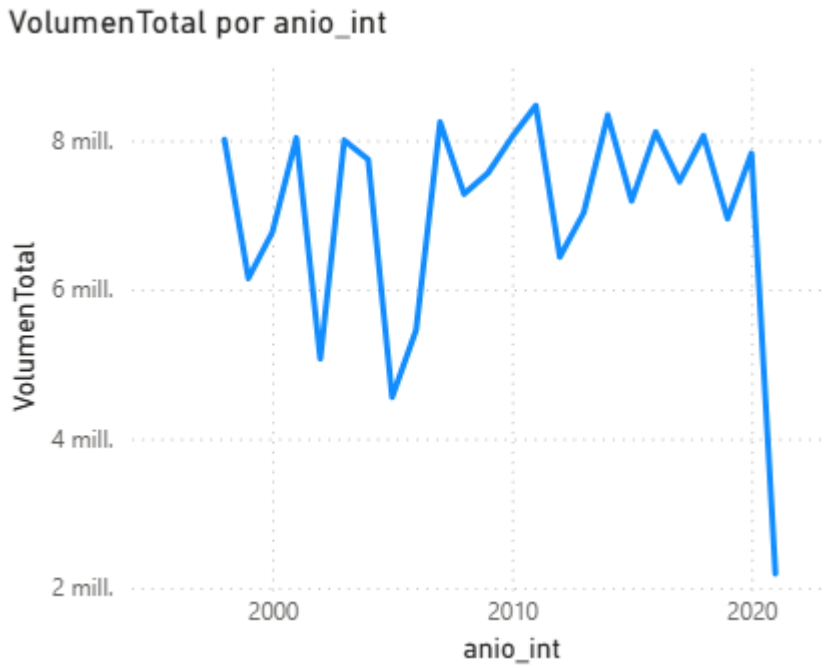
Una de las primeras visualizaciones elaboradas en Power BI es un gráfico de barras que muestra la diferencia entre el volumen máximo y mínimo registrado en cada embalse. Esta visualización ofrece una visión inmediata de cuáles son los embalses más sensibles a variaciones hidrológicas. El análisis reveló que El Atazar presenta la mayor diferencia entre los valores máximos y mínimos, lo que es coherente con su gran capacidad y el papel clave que desempeña dentro del sistema.



El uso de este gráfico permite a AEMET identificar qué embalses podrían experimentar mayores fluctuaciones en periodos de sequía o lluvia intensa.

9.2. Visualización 2: Volumen total anual del sistema

Otra visualización fundamental es un gráfico de líneas que muestra la evolución del volumen total anual del conjunto de embalses. Esta visualización funciona como indicador de salud general del sistema. En nuestro caso, se observa que, aunque existen oscilaciones importantes entre años, no se aprecia una tendencia descendente sostenida.

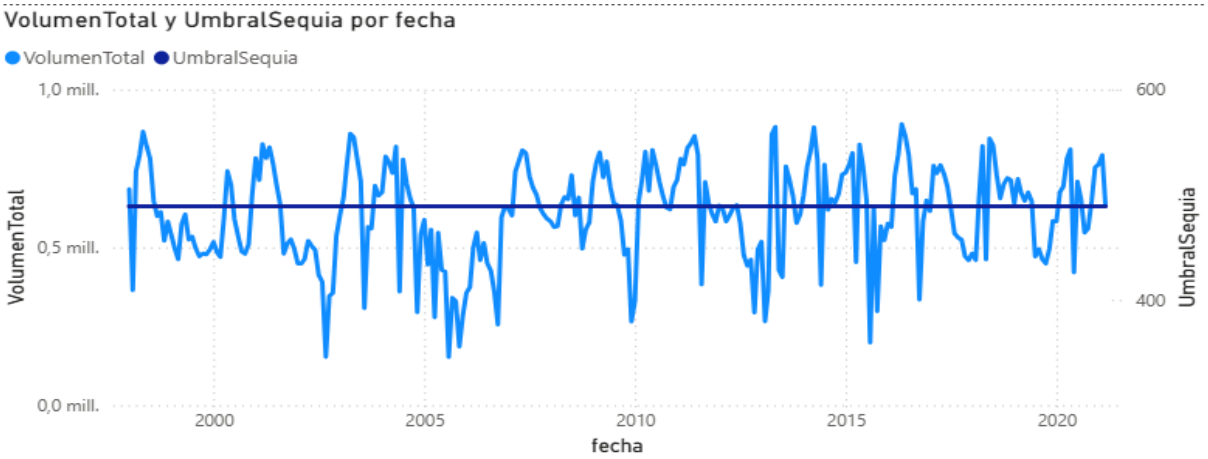


Este resultado coincide con el análisis realizado en Python y refuerza la idea de que el sistema no se está deteriorando progresivamente.

9.3. Visualización 3: Comparación del volumen mensual con el umbral de sequía

Quizá la visualización más relevante del dashboard sea el gráfico que muestra simultáneamente el volumen mensual total y el umbral de sequía derivado del percentil 10 histórico. Esta visualización permite evaluar de manera inmediata si los niveles registrados se encuentran por encima o por debajo del límite crítico.

En todos los puntos de la serie, tanto histórica como reciente, el volumen se mantiene por encima del umbral. Esto confirma visualmente la conclusión obtenida mediante el análisis estadístico.

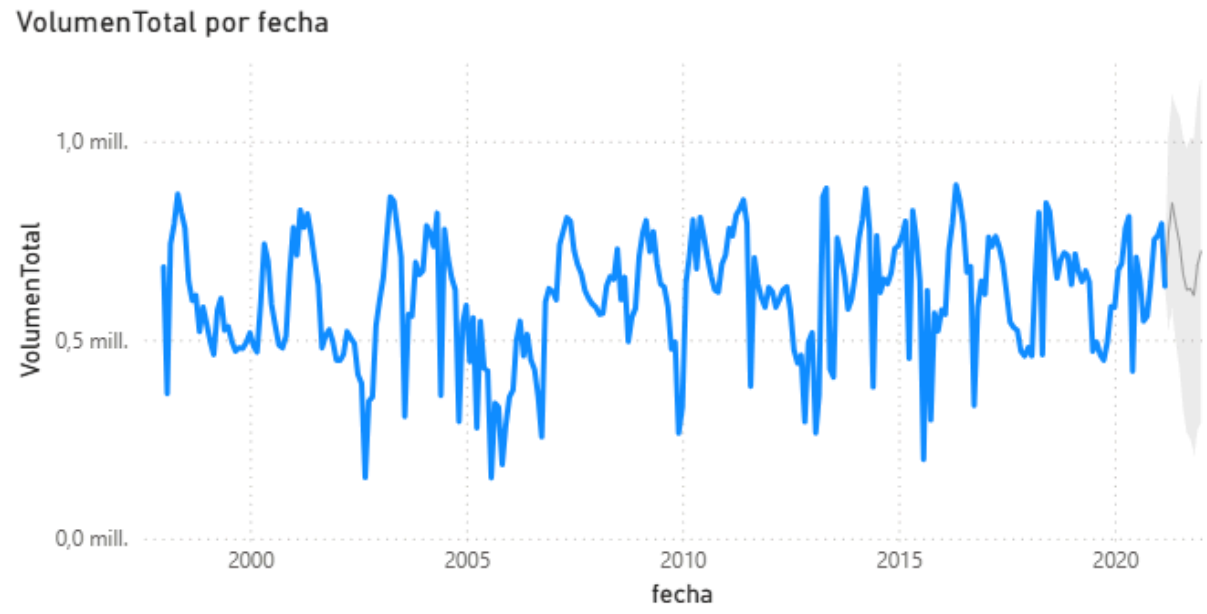


Esta figura es clave para AEMET, ya que les permite valorar el riesgo de sequía sin necesidad de revisar los resultados numéricos o técnicos del modelo.

9.4. Visualización 4: Forecast nativo de Power BI

Con el fin de complementar los resultados obtenidos con Prophet en Python, se utilizó la funcionalidad de “Forecast” nativa de Power BI. Este método analiza la serie temporal y genera una predicción para los próximos 12 meses, junto con un intervalo de confianza sombreado.

La predicción que ofrece Power BI muestra un comportamiento estable, sin tendencias descendentes bruscas, lo cual resulta consistente con el forecast generado en el notebook. El hecho de que dos métodos distintos conduzcan a conclusiones similares refuerza aún más la confianza en los resultados.



9.5. Visualización 5: KPI de Riesgo de Sequía

Finalmente, se añadió un KPI en forma de tarjeta, que resume de manera inmediata el resultado del análisis: Riesgo de Sequía = BAJO. Esta tarjeta utiliza la medida DAX RiesgoSequia, que compara el valor mínimo del volumen mensual total con el umbral de sequía definido en Python.

El KPI es útil tanto para profesionales técnicos como para responsables públicos, ya que ofrece una interpretación clara sin necesidad de revisar gráficos o cálculos adicionales.

10. Respuestas Directas para AEMET

A partir del análisis completo, se pueden ofrecer respuestas claras y justificadas a las preguntas planteadas:

1. ¿Qué embalse presenta la mayor diferencia entre máximo y mínimo?

El embalse de El Atazar presenta la mayor amplitud en su rango histórico de almacenamiento. Esto se debe a su gran capacidad, su papel como embalse regulador principal del sistema y la naturaleza variable de sus aportes.

2. ¿Cuándo se puede considerar que existe sequía?

Siguiendo un criterio hidrológico robusto, se definió un umbral basado en el percentil 10 del volumen mensual total histórico. Se evalúa sequía cuando la serie predicha o histórica cae por debajo de este valor. En nuestro análisis, ningún mes cae por debajo del umbral, ni en la historia disponible ni en la predicción futura.

3. ¿Podemos asegurar que no habrá sequía con los datos existentes?

Los datos disponibles, junto con los modelos empleados (tanto Prophet como el forecast de Power BI), indican que no existe riesgo de sequía a corto plazo. No obstante, como todo modelo predictivo, su fiabilidad depende de la representatividad del histórico. Eventos extremos no presentes en los datos no pueden ser anticipados. Aun así, en términos estadísticos y hidrológicos, la probabilidad de sequía durante los próximos 12 meses es baja.

11. Conclusiones Finales

El análisis de los embalses de la Comunidad de Madrid, realizado mediante técnicas de Big Data, modelos de predicción y visualización profesional, lleva a la conclusión general de que el sistema hídrico se encuentra en una situación estable. La evaluación anual no muestra signos de degradación estructural, y los modelos predictivos confirman que, en los próximos 12 meses, los volúmenes de agua embalsada se mantendrán dentro de rangos normales.

Tanto la parte analítica realizada en Python como el dashboard diseñado en Power BI sugieren una conclusión consistente: no existe riesgo de sequía a corto plazo en el sistema de embalses de Madrid. Este resultado, apoyado por visualizaciones claras, cálculos estadísticos y modelos de predicción robustos, constituye una base sólida para que AEMET pueda mantener su planificación hídrica sin necesidad de aplicar restricciones extraordinarias de consumo de agua.

Desde el punto de vista académico, este proyecto ha permitido integrar conocimientos de limpieza de datos, análisis exploratorio, modelado predictivo y visualización, demostrando una comprensión completa del ciclo analítico. Además, el uso combinado de entornos como Google Colab y Power BI facilita un flujo de trabajo moderno y profesional, aplicable en contextos reales.

12.Bibliografía

- Prophet — Documentación Oficial
  https://facebook.github.io/prophet/

- Prophet — GitHub Oficial
  https://github.com/facebook/prophet

- Forecasting Principles — Universidad de Monash (Hyndman)
  https://otexts.com/fpp3/

- Stationarity & Times Series Basics
  https://www.machinelearningplus.com/time-series/time-series-analysis-python/

- Limpieza de Datos con Pandas (Real Python)
  https://realpython.com/python-data-cleaning/

- Tutorial Pandas (Kaggle)
  https://www.kaggle.com/learn/pandas

- ETL en Python paso a paso
  https://www.datacamp.com/tutorial/etl-pipeline-python

- Manejo de fechas en Pandas
  https://pandas.pydata.org/docs/user_guide/timeseries.html

PORTADA

Nombre Alumno / DNI	Alberto Rodríguez González
Título del Programa	**3ºPD COMPUTER SCIENCE & DATA SCIENCE & AI**
Nº Unidad y Título	**UNIT 12 – Big Data Analytics**
Año académico	**2025/2026**
Profesor de la unidad	
Título del Assignment	**AB FINAL**
Día de emisión	**12/09/2025**
Día de entrega	**28/11/2025**
Nombre IV y fecha	

Declaración del estudiante	**Certifico que la presentación del assignment es completamente mi propio trabajo y entiendo completamente las consecuencias del plagio. Entiendo que hacer una declaración falsa es una forma de mala práctica.** **Fecha: 20/05/2025** **Firma del alumno:** 

Plagio

El plagio es una forma particular de hacer trampa. El plagio debe evitarse a toda costa y los alumnos que infrinjan las reglas, aunque sea inocentemente, pueden ser sancionados. Es su responsabilidad asegurarse de comprender las prácticas de referencia correctas. Como alumno de nivel universitario, se espera que utilice las referencias adecuadas en todo momento y mantenga notas cuidadosamente detalladas de todas sus fuentes de materiales para el material que ha utilizado en su trabajo, incluido cualquier material descargado de Internet. Consulte al profesor de la unidad correspondiente o al tutor del curso si necesita más consejos.