**Taller Final - Big Data con Apache Airflow**

**1. Nombre del Proyecto**: Flujo de datos del IDEAM - Precipitación mensual 2020-2025. Departamento del Meta, estación SENA (Villavicencio)

**2. Descripción general del flujo**: este proyecto implementa un pipeline de Big Data orquestado con Apache Airflow, que automatiza la captura, transformación, visualización y análisis de datos de precipitación. Se simula una fuente tipo Kafka, se transforman los datos con Dask, se almacenan en PostgreSQL, se generan gráficos y se entrena un modelo de regresión lineal.

También, se ha incorporado un módulo de predicción para simular los valores futuros de precipitación en los próximos meses, extendiendo el modelo a una visión prospectiva automatizada.

**Fases del flujo:**

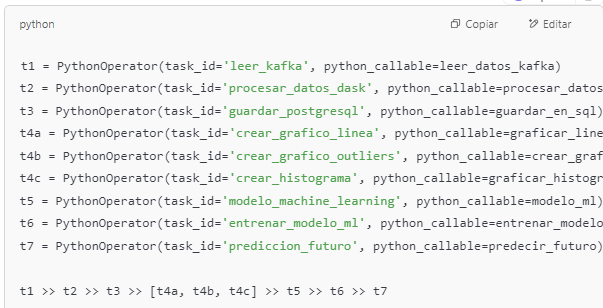
1. **Ingesta de datos**: Simulación de Kafka para cargar un archivo CSV con datos originales.
2. **Procesamiento**: Uso de Dask para transformar los datos (duplicación de valores para simular pérdidas o errores).
3. **Persistencia**: Almacenamiento de los datos procesados en una base de datos PostgreSQL.
4. **Visualización**: Generación de gráficos (línea e histograma) usando Matplotlib.
5. **Modelo de Machine Learning**: Entrenamiento de una regresión lineal con scikit-learn para predecir la precipitación.
6. **Exportación**: Guardado del modelo en formato .joblib y de las predicciones en un archivo .csv.

**2. Diagrama del flujo en Airflow**



**3. Fragmento destacado del DAG**



****

**4. Justificación técnica y logros**

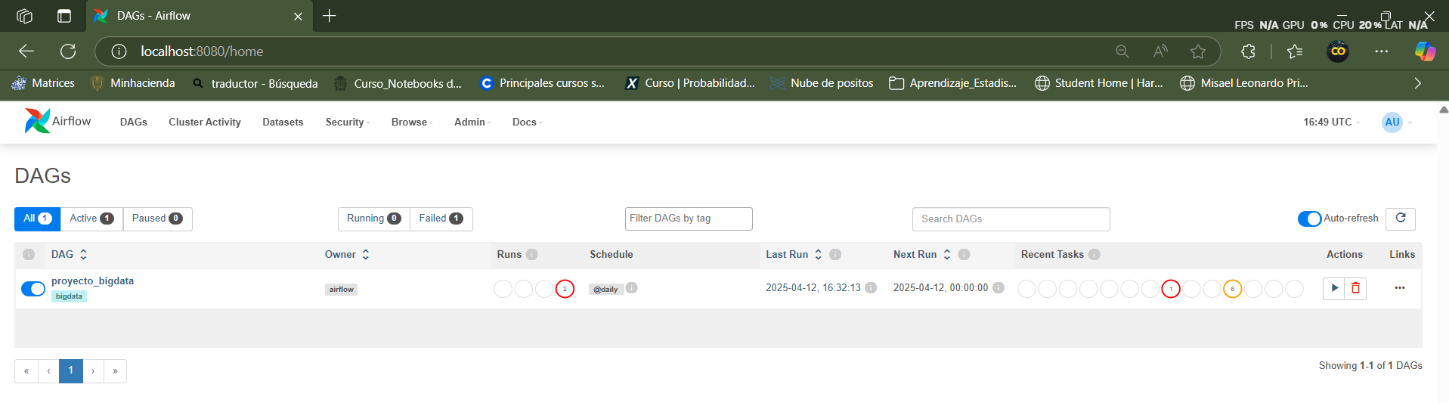
* Se integraron herramientas reales del ecosistema Big Data: Kafka (simulado), Dask, PostgreSQL y Airflow.
* Se generaron visualizaciones automáticas dentro del flujo.
* Se incluyó un modelo de Machine Learning simple (regresión lineal), entrenado automáticamente.
* Extensión del flujo para predecir valores futuros a partir del modelo entrenado.
* Todos los productos (CSV, PNG, modelo .joblib) se almacenan en /opt/airflow/salidas, permitiendo auditar el flujo.
* Airflow permite observar ejecuciones, tiempos, y logs por cada tarea.

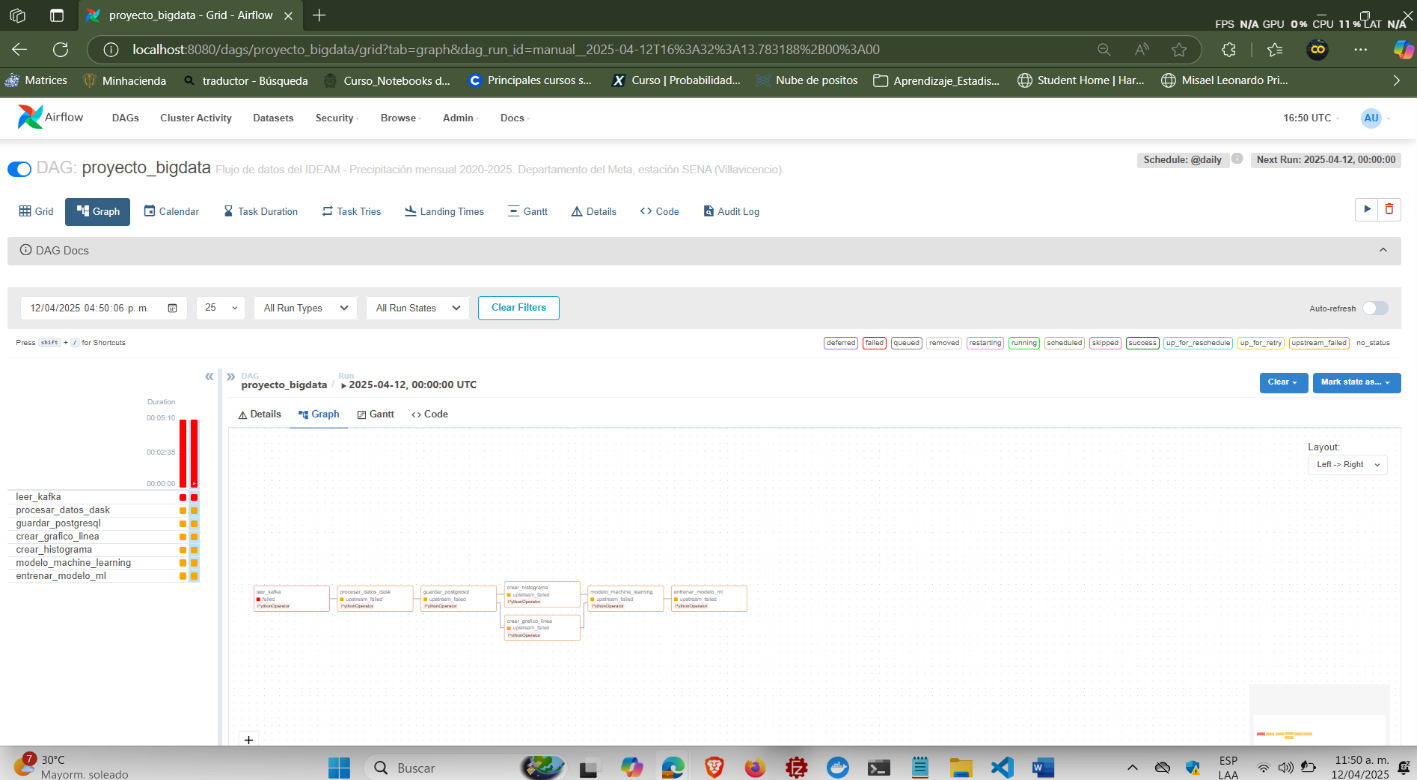
**Anexos:**

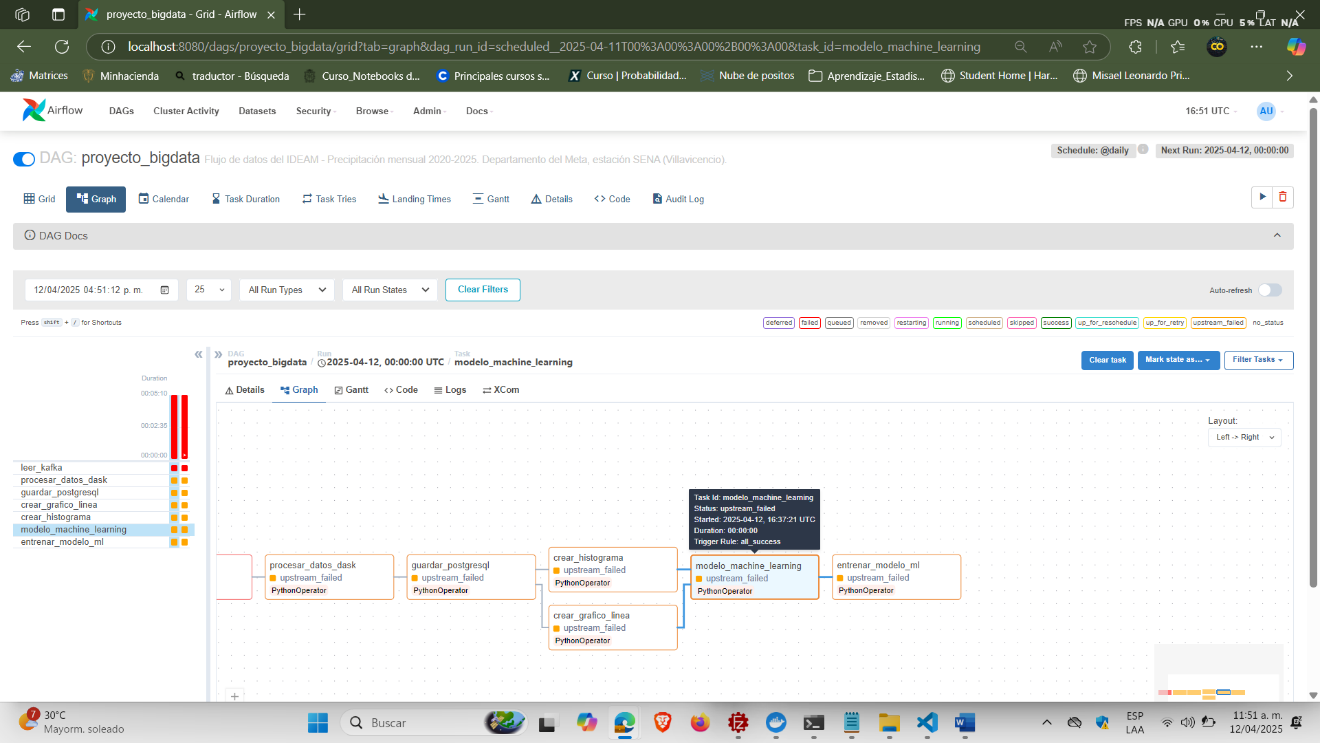
* Capturas de la interfaz de Airflow funcionando
* Archivos: proyecto\_bigdata\_dag.py, modelo\_ml.py, datos, gráficos y predicciones.

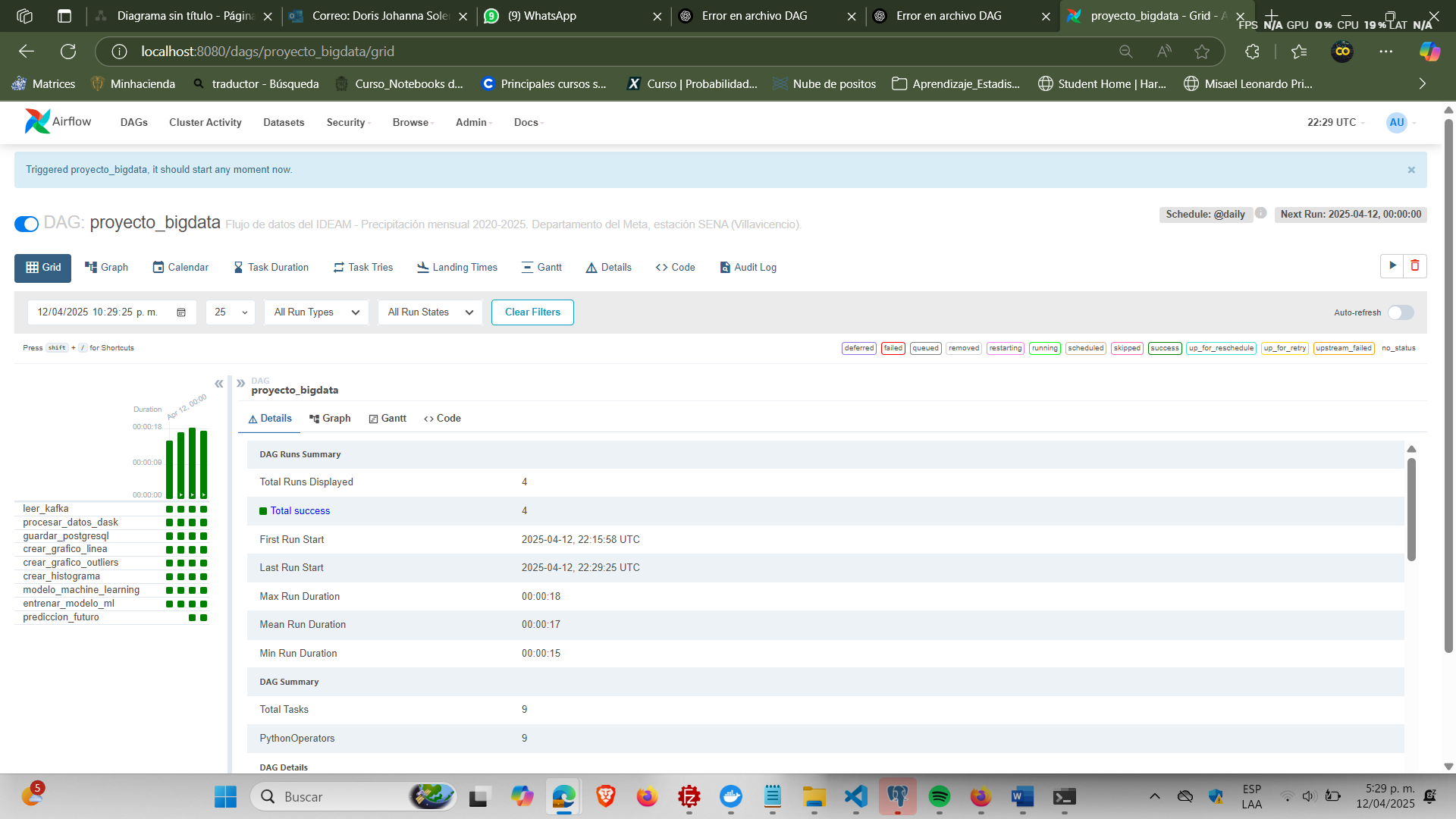
Estudiante: Doris Johanna Soler Castro CC 40330258 MINE III Semestre

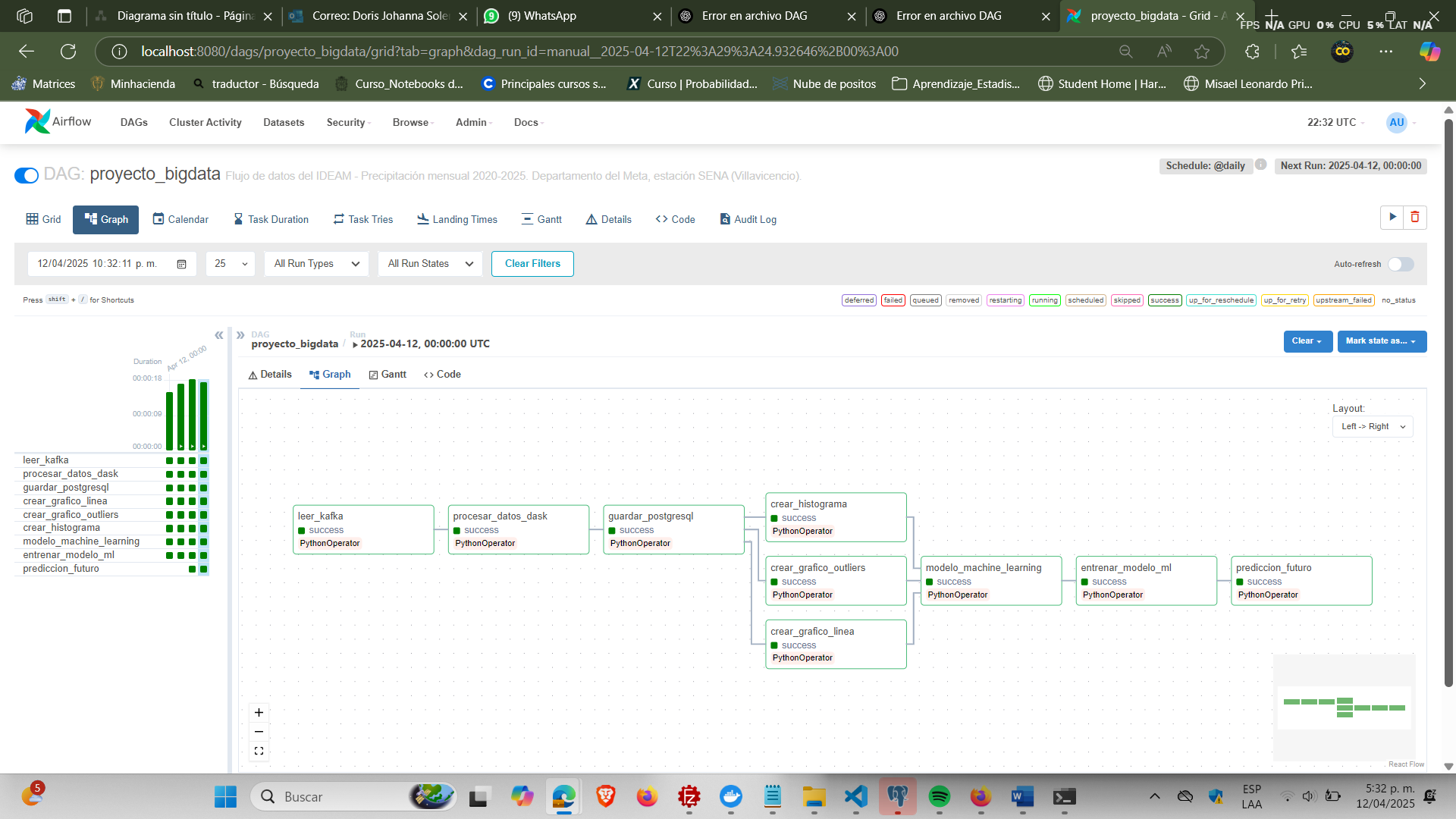
**CAPTURAS DE LA INTERFAZ DE AIRFLOW FUNCIONANDO**

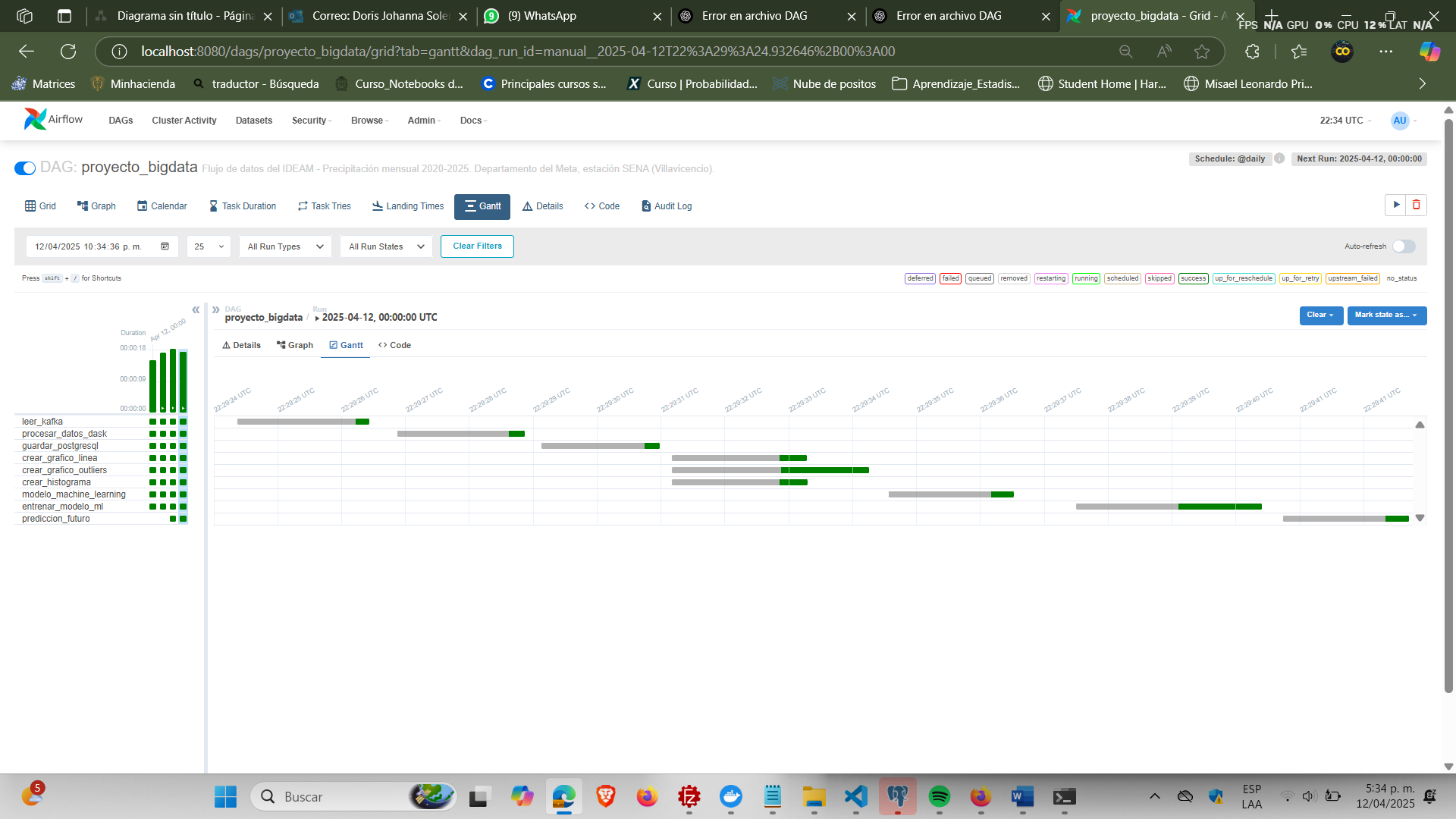


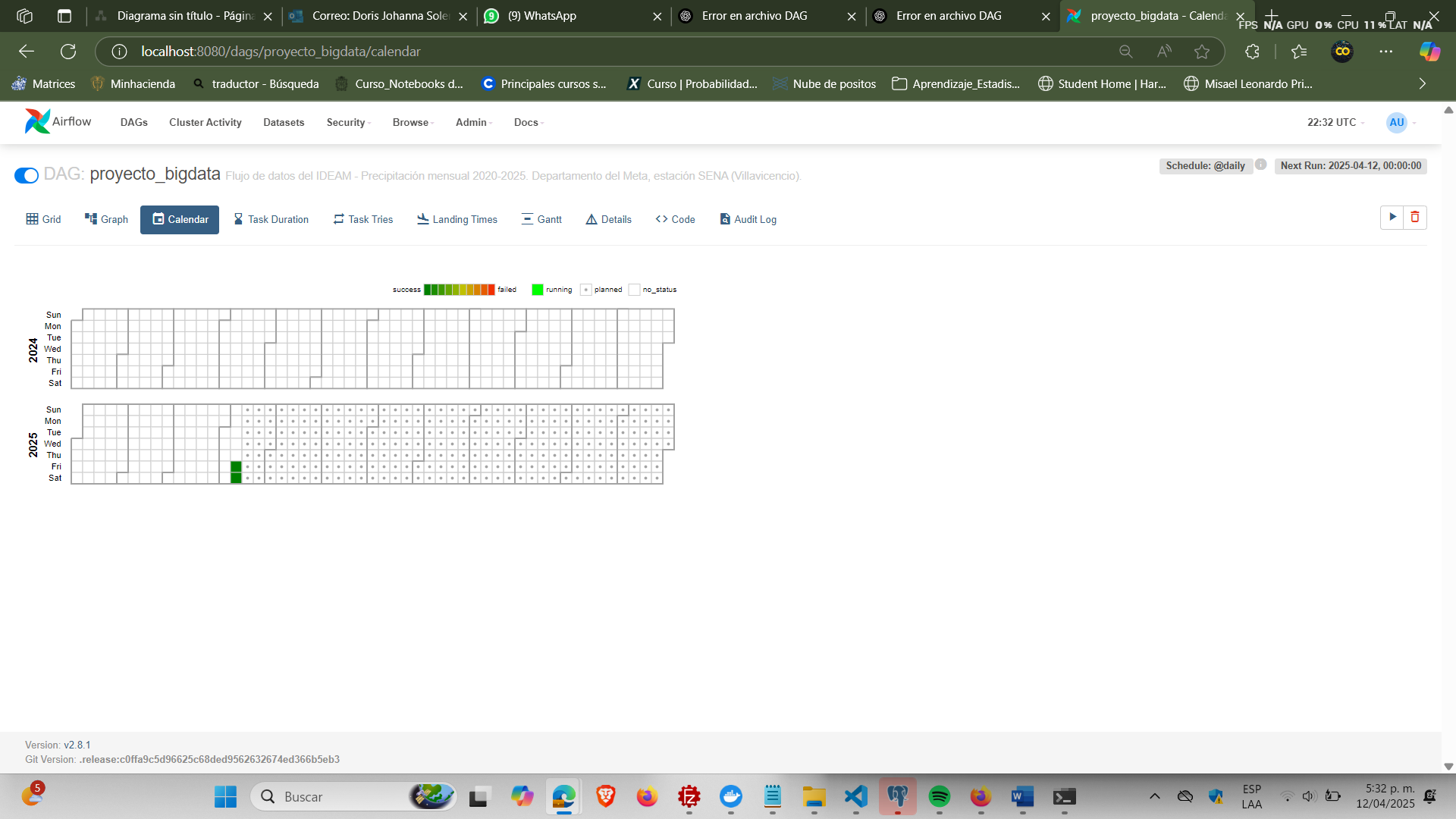


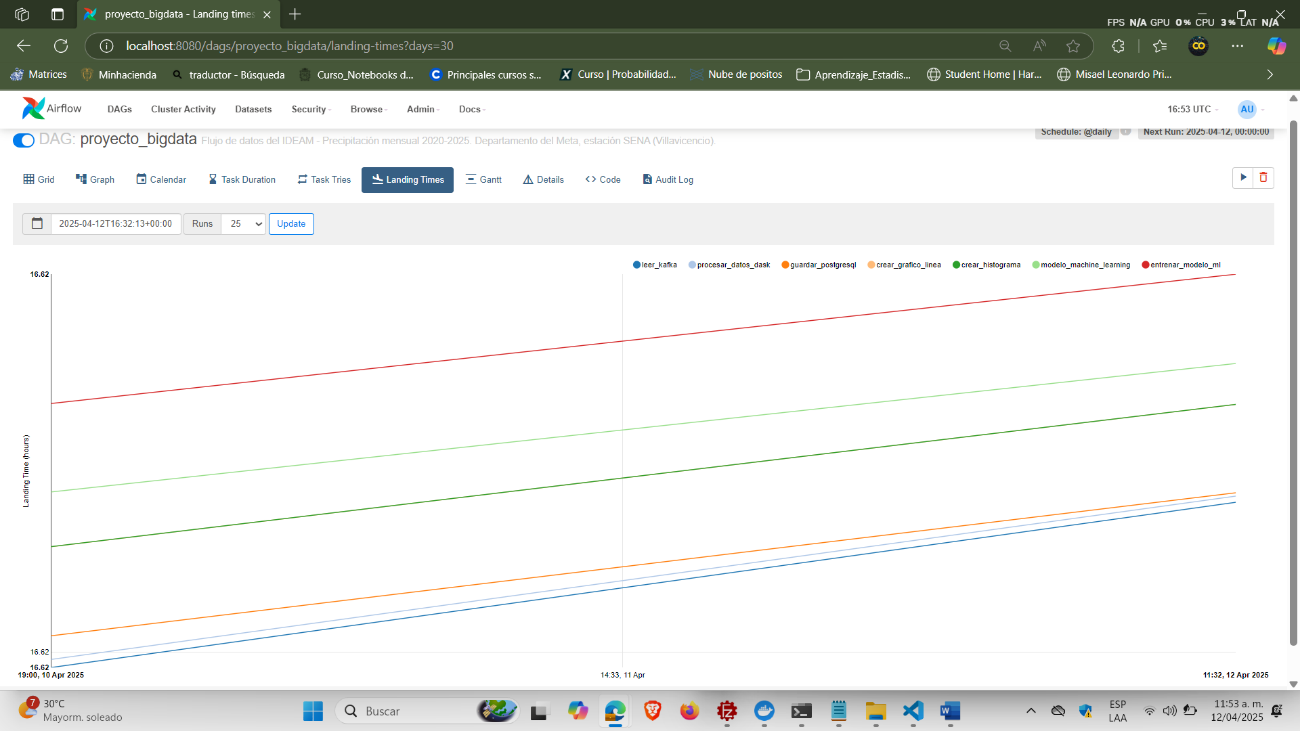


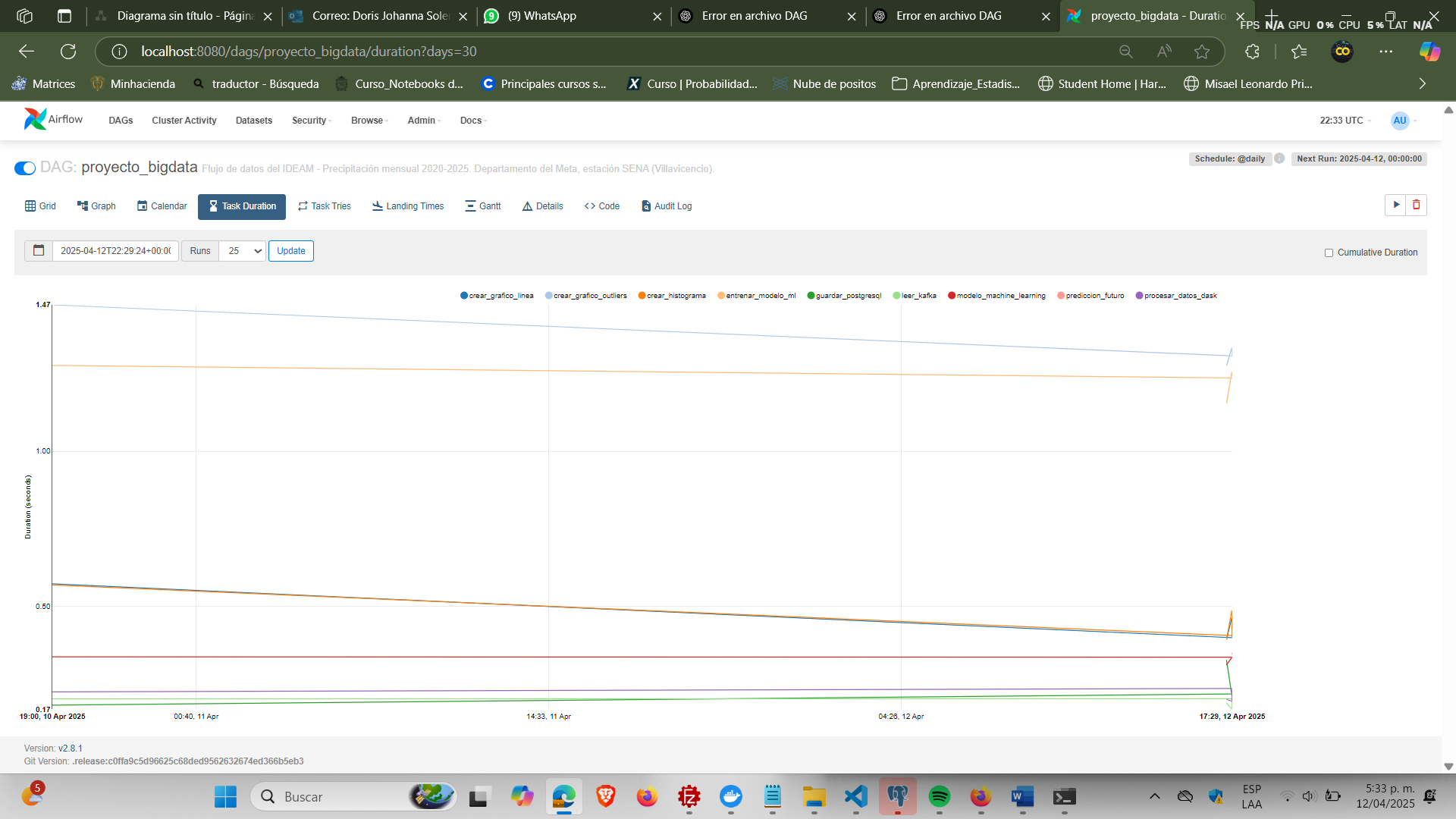


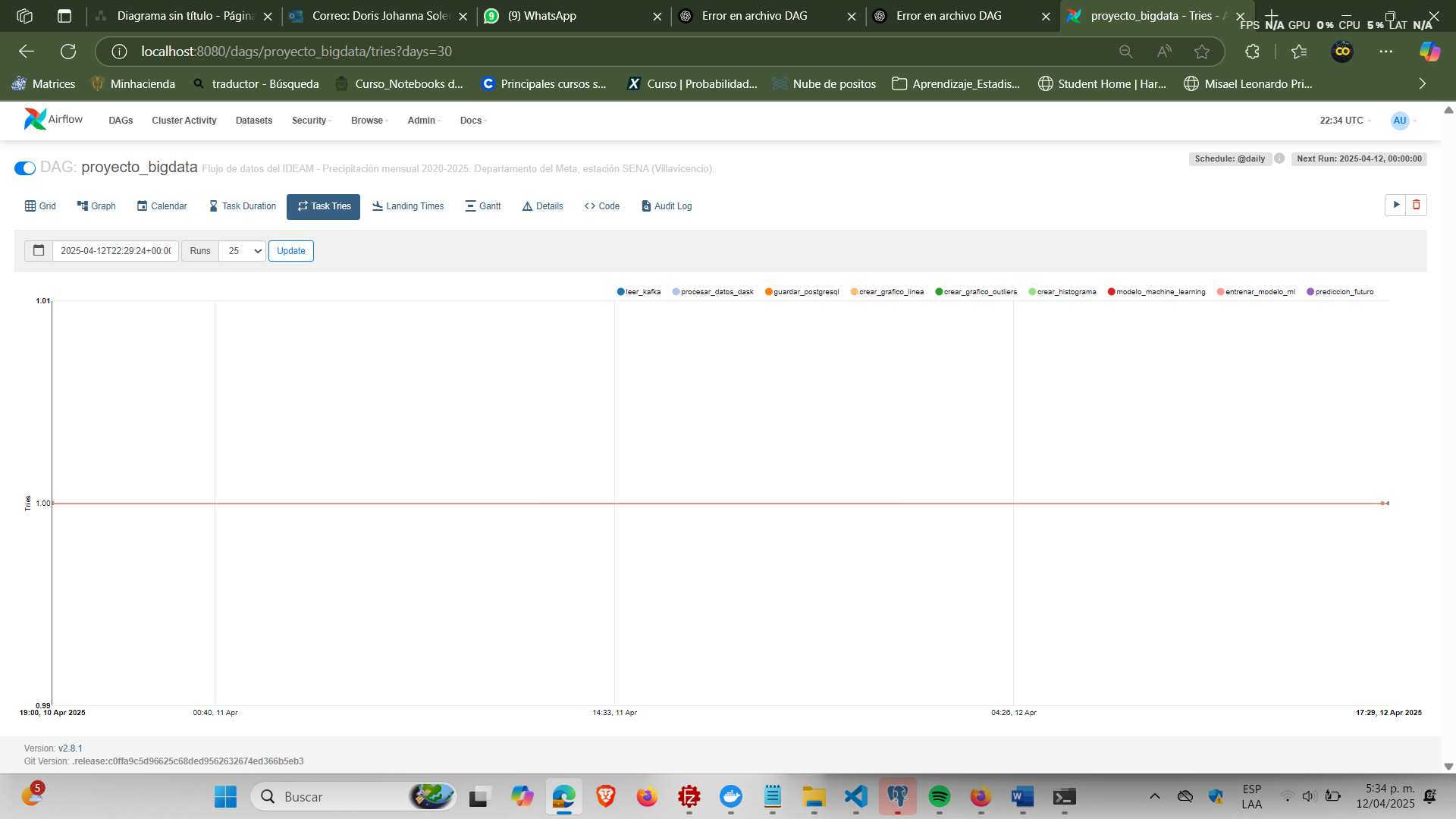


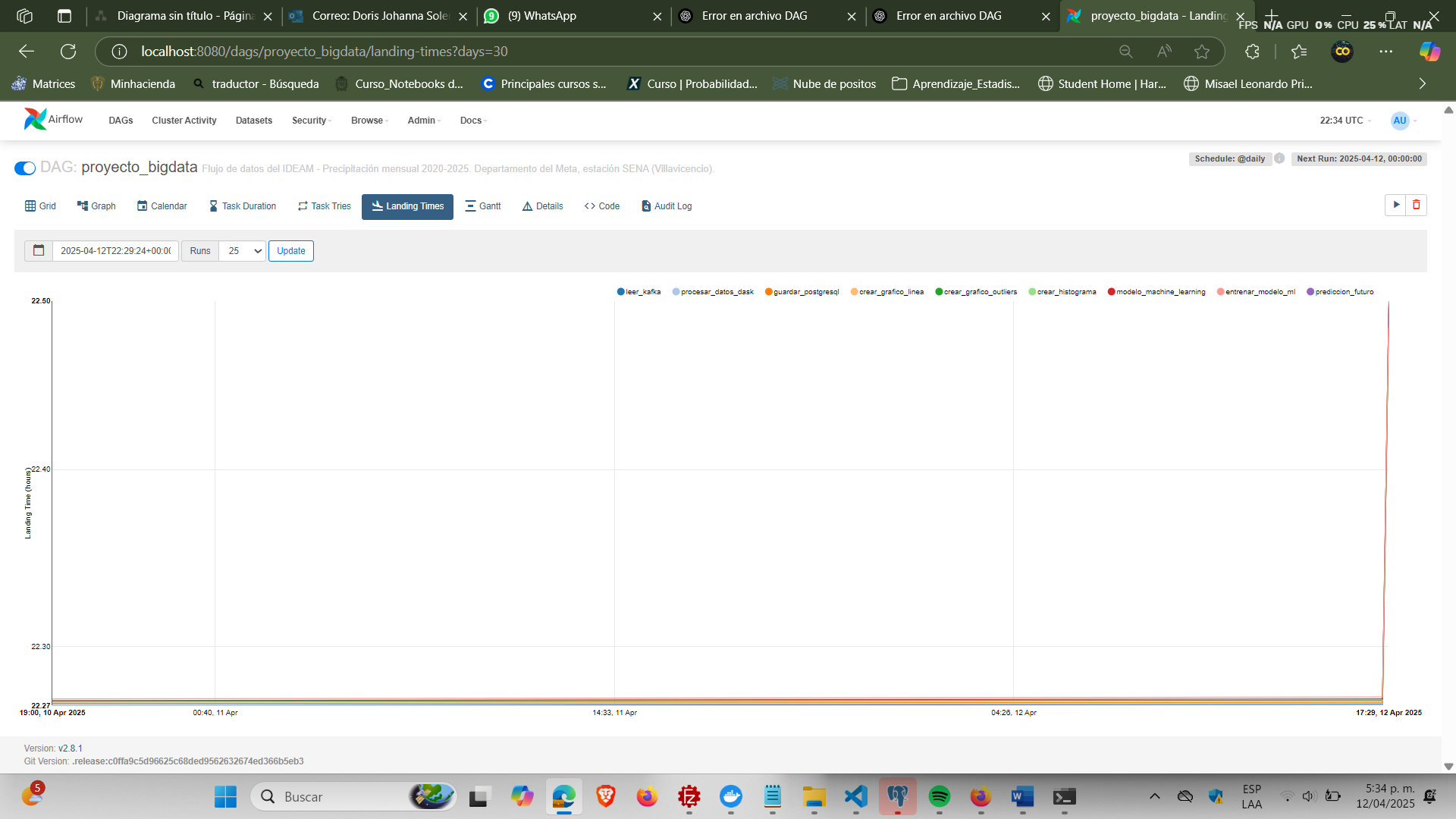


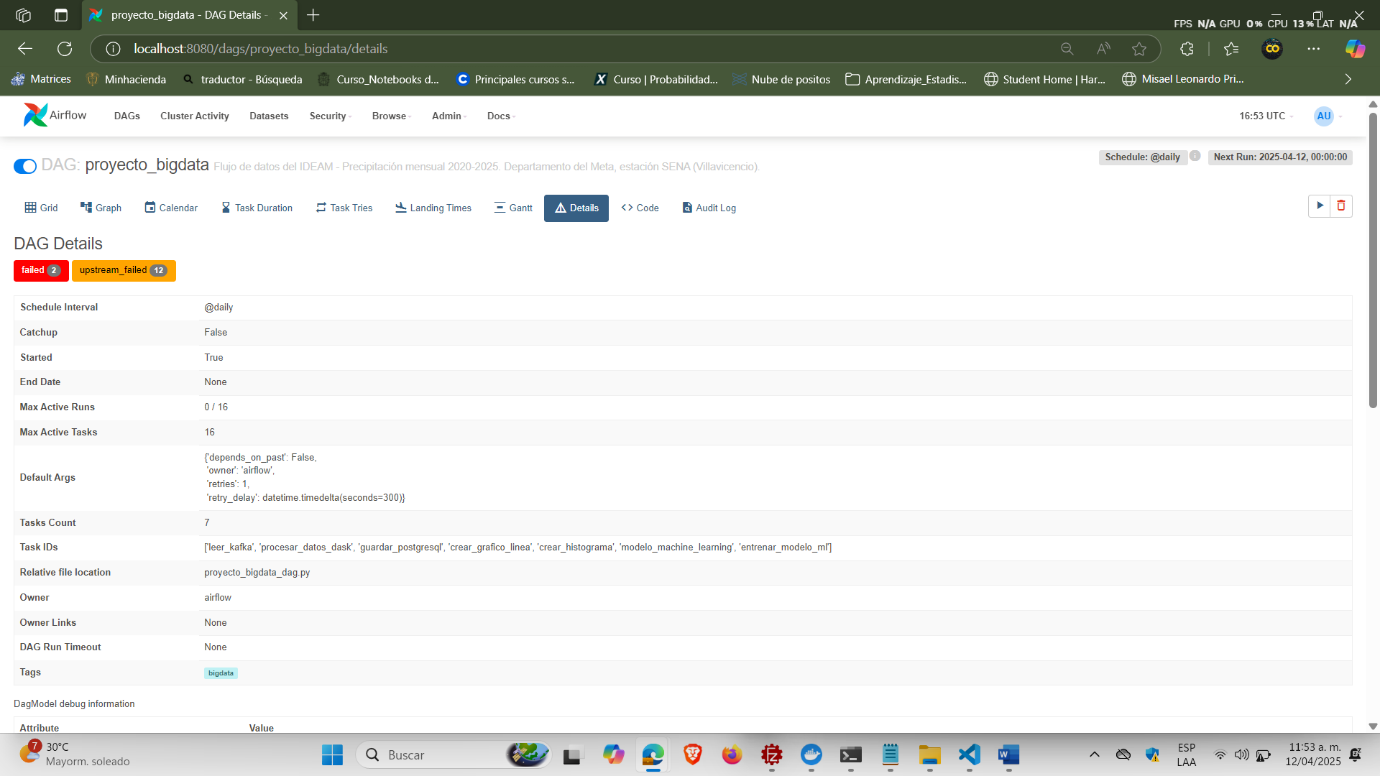


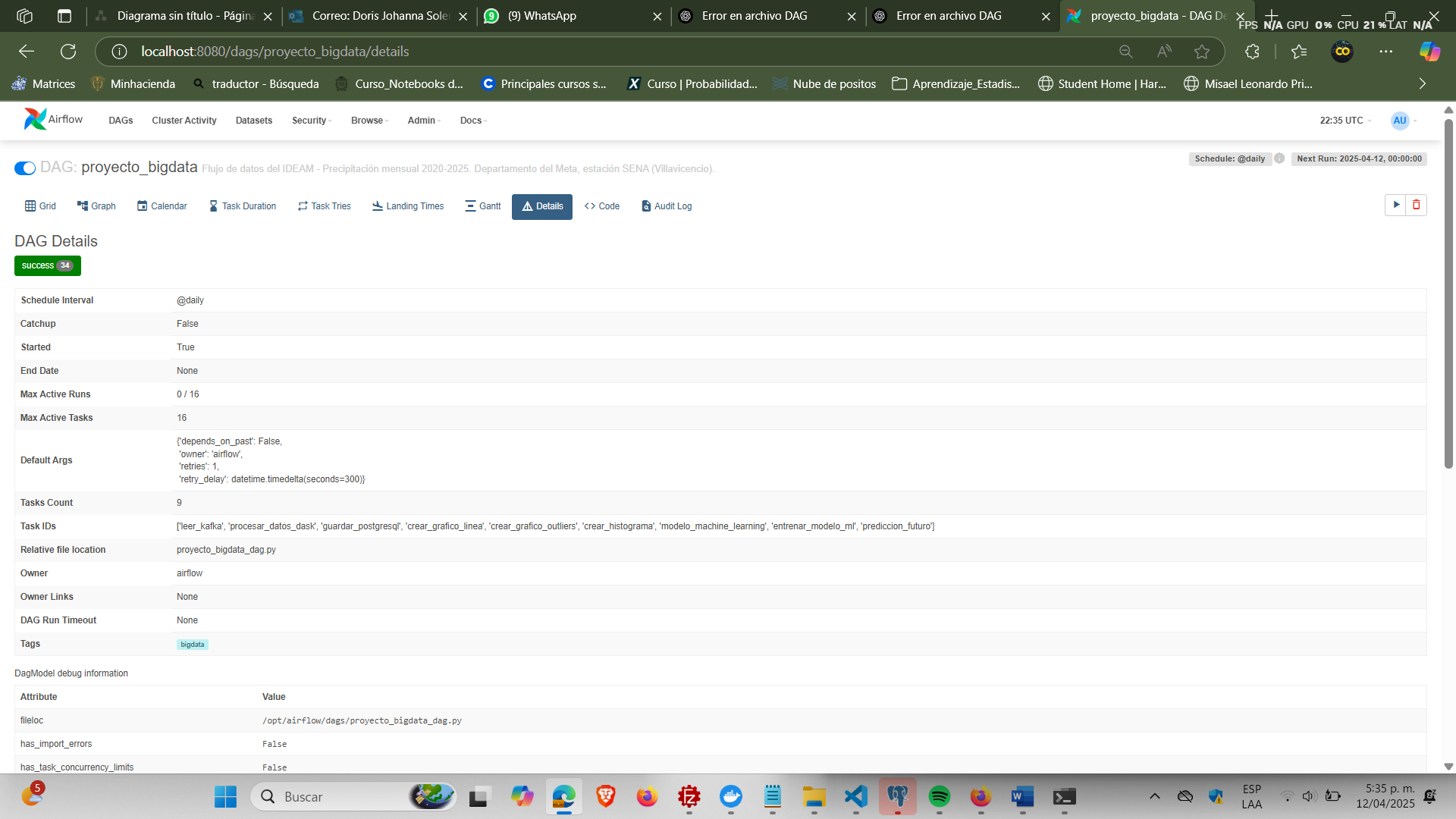


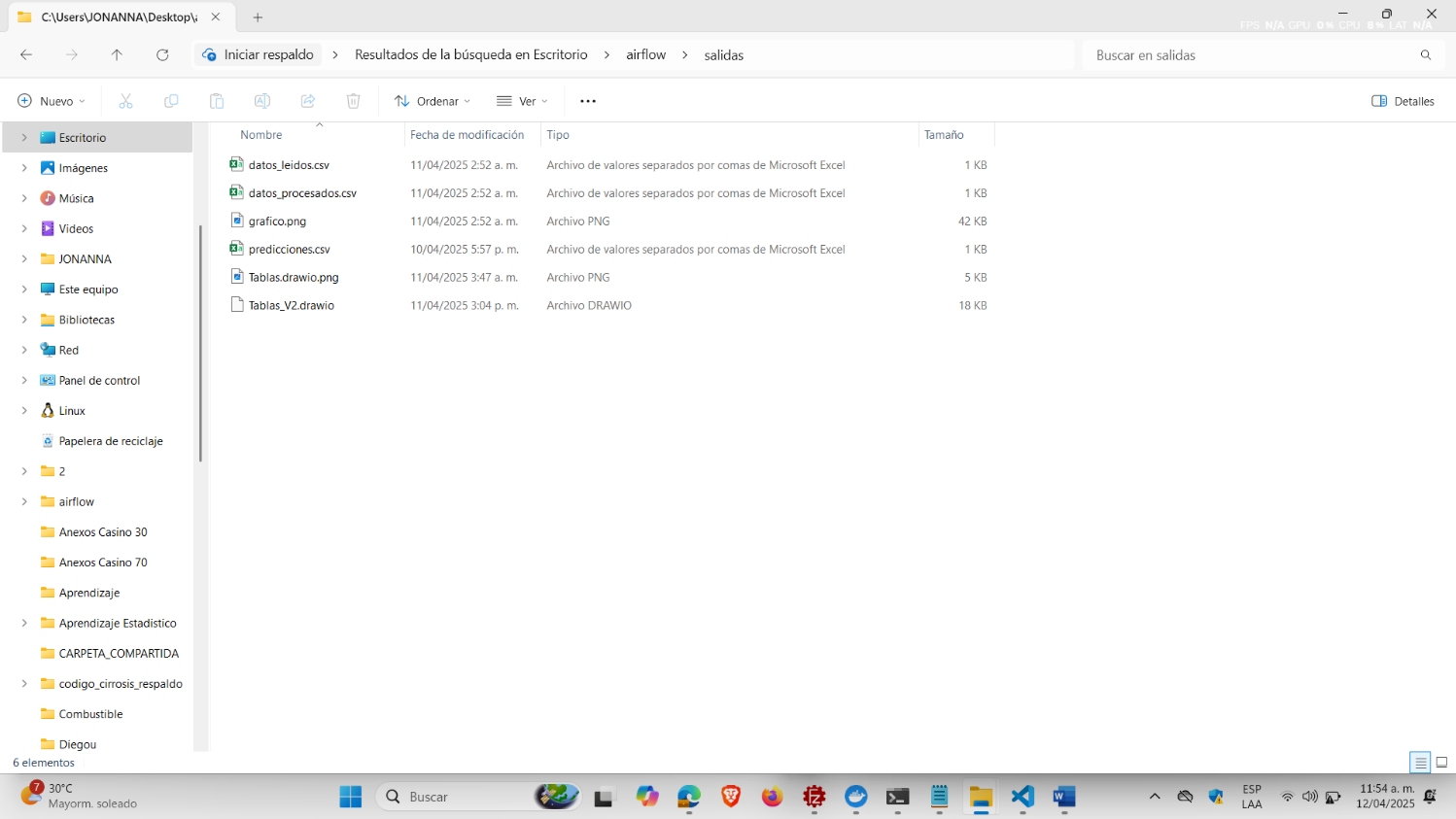


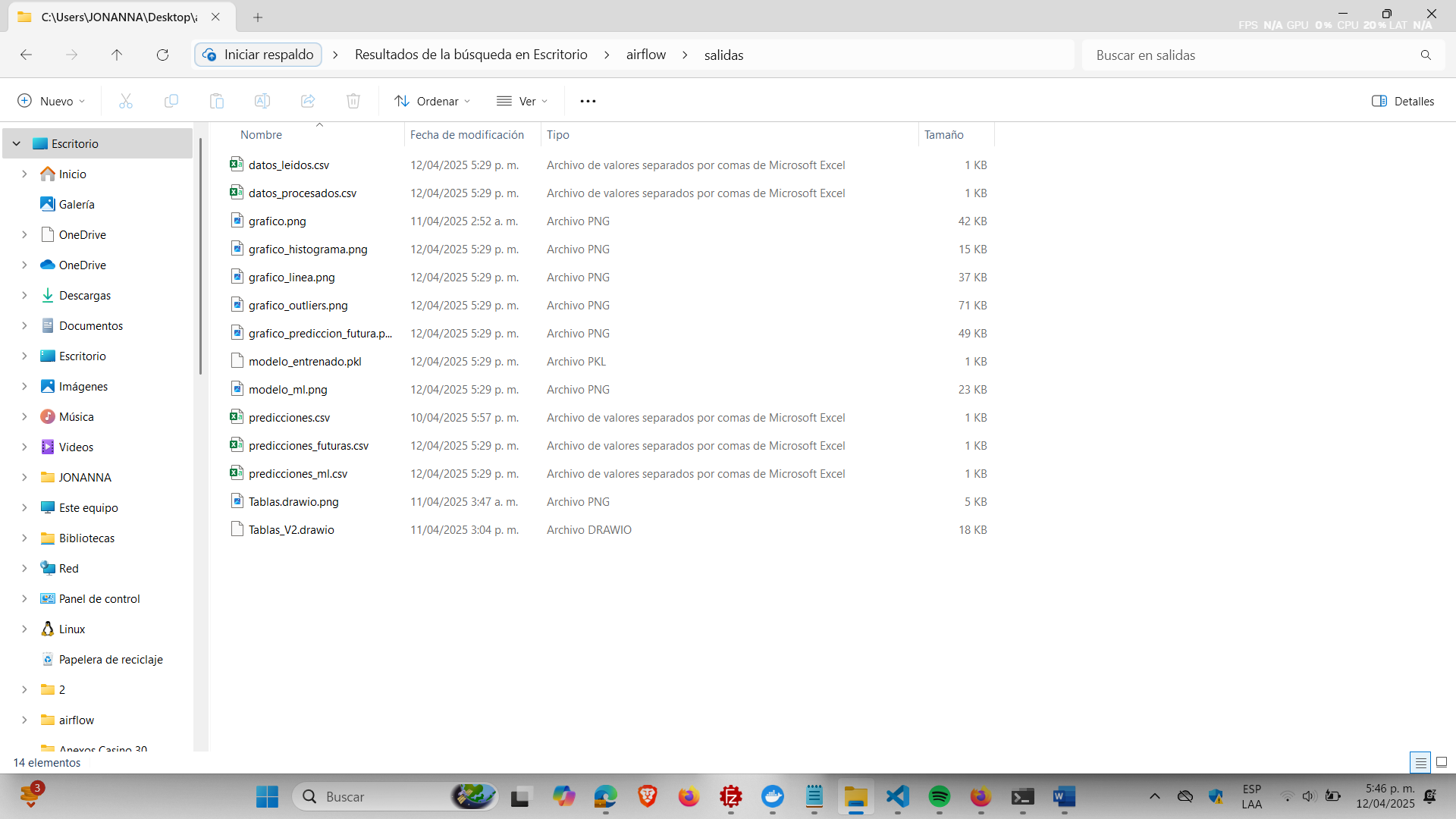


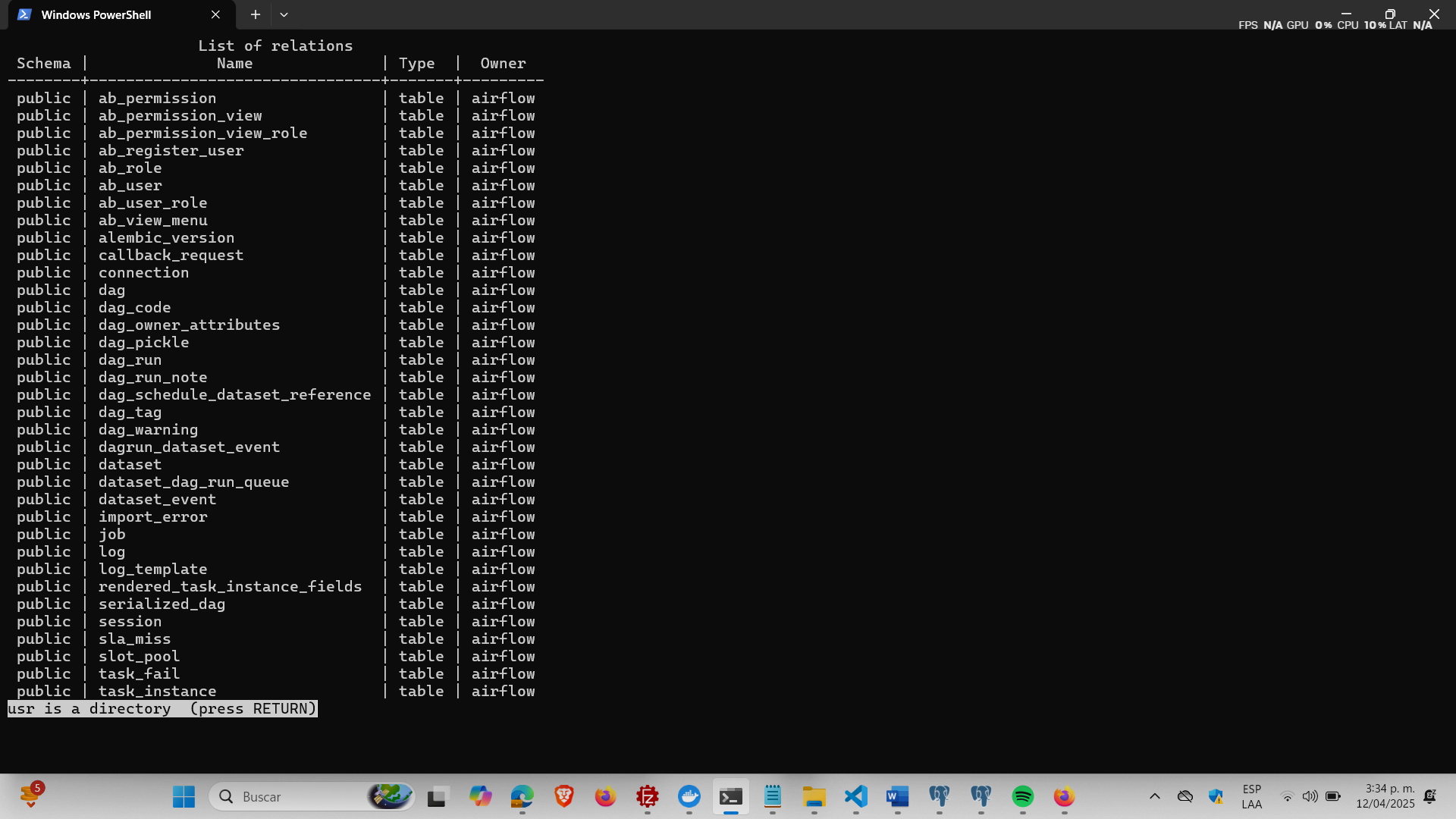












**ARCHIVOS: PROYECTO\_BIGDATA\_DAG.PY, MODELO\_ML.PY, DATOS, GRÁFICOS Y PREDICCIONES**