# Trabajo Práctico Final: *Dataset Agent*

Palomares Nicolas

Diciembre, 2025

## Definición del Problema y Valor de Negocio

---

El trabajo del científico de datos abarca diferentes áreas: la estadística y la matemática para la limpieza y análisis exploratorio de los datos, la informática y programación para el entendimiento de los algoritmos y la resolución de problemas, y la inteligencia de negocios para el storytelling y comunicación de los resultados.

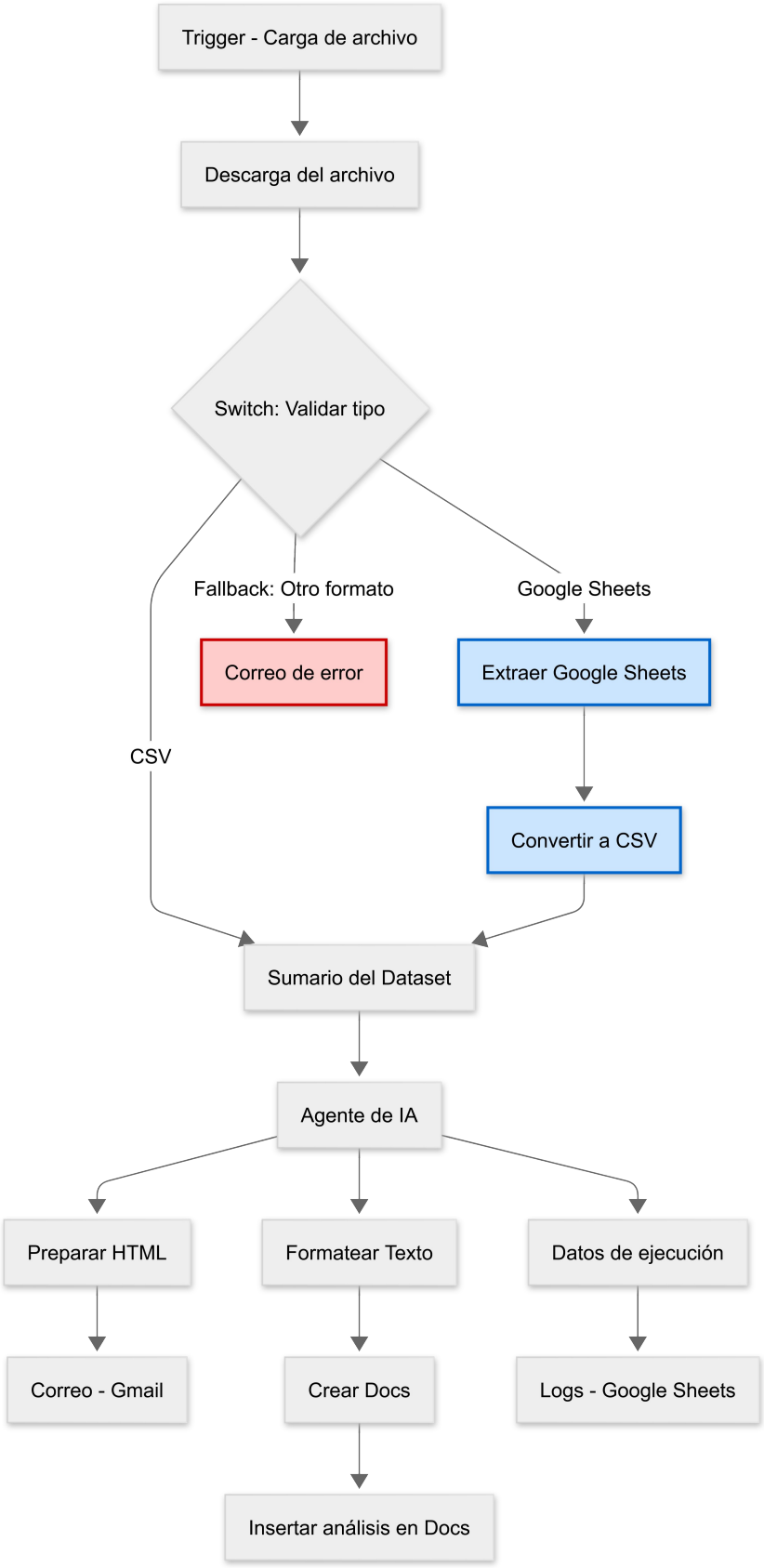
El presente documento explica el armado y desarrollo de un flujo automatizado en la plataforma **N8N** que promete dar soporte al científico de datos, aportando información clave para el análisis de un determinado dataset y ayudando en la toma de decisiones. Principalmente, el flujo se encarga de analizar un dataset mediante un **Agente IA** para luego brindar recomendaciones en cuanto al manejo de los datos (tanto en la limpieza como en el *feature engineering*), y dar apoyo para el modelado de un futuro algoritmo de machine learning o deep learning, especificando posibles modelos a entrenar, métricas a utilizar y posible variable objetivo.

**Valor al negocio:** Este proyecto proporciona soporte técnico integral al equipo de ciencia de datos en las etapas críticas del desarrollo de modelos de machine learning y deep learning. Específicamente, el valor se materializa en:

- **Automatización de procesos de limpieza de datos**: Reducción significativa del tiempo invertido en tareas repetitivas de preprocesamiento, permitiendo al científico de datos enfocarse en aspectos estratégicos del modelado.
- **Ingeniería de variables optimizada**: Generación de recomendaciones basadas en análisis exploratorio automatizado, identificación de correlaciones relevantes y sugerencias de transformaciones de variables que potencien el poder predictivo del modelo.
- **Análisis y diagnóstico inteligente**: Provisión de insights sobre la calidad de los datos, detección de anomalías, evaluación de desbalanceo de clases y análisis de distribuciones, facilitando la toma de decisiones informadas durante el ciclo de modelado.
- **Aceleración del time-to-market**: Al reducir el tiempo dedicado a tareas preparatorias, se acelera el desarrollo e implementación de modelos en producción, generando valor de negocio más rápidamente.

El **impacto esperado** incluye mejoras en la eficiencia operacional del equipo de datos, mayor calidad en los modelos desarrollados y una reducción en el tiempo total del ciclo de desarrollo de proyectos de machine learning.

# Descripción y Funcionamiento del Flujo *Dataset Agent*



El flujo *Dataset Agent* está conformado de la siguiente manera:

**Parte 1: Pipeline de ingesta y procesamiento de datos**

- **Nodo Google Drive Trigger**: Detecta automáticamente archivos nuevos cargados en la carpeta ***Datasets de ejemplo*** del Drive.
- **Nodo Google Drive – Download File**: Descarga el archivo detectado tal como está almacenado en Google Drive.
- **Nodo Switch**: Valida el tipo de archivo mediante MIME type. Si es Google Sheets, lo procesa en una rama específica. Si es CSV nativo, continúa directamente al análisis. Cualquier otro formato activa el fallback que envía un correo de error y finaliza el flujo.
- **Nodo Extraer Google Sheets**: Extrae los datos de la hoja de cálculo de Google Sheets.
- **Nodo Convertir a CSV**: Convierte los datos extraídos de Google Sheets a formato CSV para su procesamiento unificado.
- **Nodo Sumario del Dataset (Python)**: Realiza un análisis estadístico completo del dataset utilizando pandas y NumPy. Genera métricas descriptivas, detecta valores faltantes y outliers, analiza distribuciones categóricas y numéricas, y estructura toda esta información en formato JSON para su posterior evaluación por el agente de IA.

**Parte 2: Agente de IA**

- **AI Agent:** Agente de IA que analiza la información generada en el nodo anterior usando GPT-4.1 Mini. Este nodo implementa un agente inteligente que procesa exclusivamente los resultados agregados y estadísticos generados en la Parte 1 del flujo, no el dataset completo. Esta arquitectura responde a un principio de eficiencia: analizar el dataset completo a través de la API sería inviable debido al alto consumo de tokens y costos asociados. Con esta información condensada, el agente GPT-4.1 Mini genera:
    - Recomendaciones contextualizadas para la limpieza de datos
    - Sugerencias de ingeniería de variables basadas en patrones identificados
    - Insights sobre posibles estrategias de modelado
    - Alertas sobre problemas potenciales en los datos (desbalanceo, multicolinealidad, etc.)
    - Propuestas de transformaciones específicas para optimizar el modelo

**Parte 3: Resultados**

- **Nodo Get Execution:** El nodo extrae los datos de ejecución del flujo.
- **Logs – Google Sheets:** Este nodo registra automáticamente cada ejecución del flujo (utilizando los datos extraídos del nodo anterior) en una hoja de cálculo de Google Sheets destinada al monitoreo y trazabilidad (directorio ***Logs***).
- **Nodo de código – Preparar HTML:** Nodo de transformación que convierte el análisis generado por el AI Agent en un formato HTML estructurado y visualmente atractivo. Aplica estilos CSS para mejorar la legibilidad del correo electrónico, organizando el contenido en secciones claramente diferenciadas con encabezados, listas y tablas formateadas. Esta preparación garantiza que el destinatario reciba un reporte profesional y fácil de interpretar.
- **Nodo GMail – Enviar por correo**: Nodo de Gmail que envía el análisis formateado al correo electrónico de preferencia del usuario.

- **Nodo de código – Formatear texto**: Nodo de procesamiento que adapta el contenido del análisis al formato requerido por Google Docs. Elimina etiquetas HTML, ajusta el formato de texto plano y estructura el contenido para su correcta inserción en el documento.
- **Nodo Google Docs – Creación de documento**: Crea un nuevo documento en Google Docs para el análisis completo, tal cual fue generado por el agente. El documento se almacena en la carpeta ***Docs*** de Google Drive y genera la estructura base del documento con título, fecha y configuración inicial.
- **Nodo Google Docs – Actualizar documento**: Nodo final que inserta el contenido formateado del análisis en el documento de Google Docs creado previamente. Añade todo el texto procesado, mantiene la estructura de secciones y genera un documento completo y profesional que puede ser consultado, editado o compartido posteriormente.

## Referencias

---

### Plataformas y Tecnologías

- **N8N** - Plataforma de automatización de flujos de trabajo de código abierto. https://docs.n8n.io
- **OpenAI GPT-4.1 Mini** - Modelo de lenguaje para análisis y generación de recomendaciones técnicas. https://platform.openai.com/docs
- **Python** - Lenguaje de programación con librerías Pandas y NumPy para procesamiento de datos. https://pandas.pydata.org | https://numpy.org

- **APIs de Google Cloud**
    - **Google Drive API** - Almacenamiento y detección automática de archivos.
    - **Google Sheets API** - Registro de logs y trazabilidad.
    - **Google Docs API** - Generación de documentos con análisis.
    - **Gmail API** - Envío automatizado de reportes. https://developers.google.com