PREDICCIÓN DE CRIMINALIDAD URBANA CON AZURE Y OPEN DATA

AUTOR: ISMAEL ARDOY GARCÍA

Asignatura: Análisis de información para Big Dara

Índice

1. Informe de Resultados

1.1. Definición del Problema de Negocio

La gestión policial tradicional es reactiva. Este proyecto valida la **Teoría de "Ventanas Rotas"** (especialmente Grafitis) como precursor del crimen grave, con el **objetivo** de predecir riesgos futuros cruzando quejas 311 y datos delictivos. Esto aporta **valor** al facilitar el paso de un modelo de respuesta a uno de **prevención temprana**.

1.2. Enfoque, Fuentes y Consideraciones de Negocio

Integramos **OpenData (311 y NYPD)** bajo una premisa de eficiencia operativa. Considerando que los recursos son finitos, filtramos el "ruido" para centrarnos exclusivamente en Grafitis vs. Alto Riesgo, aplicando un enfoque **Lag T+1** para demostrar que la inversión en limpieza mejora la eficiencia en la asignación de recursos de prevención criminal.

1.3. Resultados Obtenidos (KPIs y Visualización)

El análisis ha permitido aislar un patrón predictivo claro al filtrar por tipologías específicas. La teoría se valida con mayor robustez al correlacionar **Grafitis** con de **Alto Riesgo.**

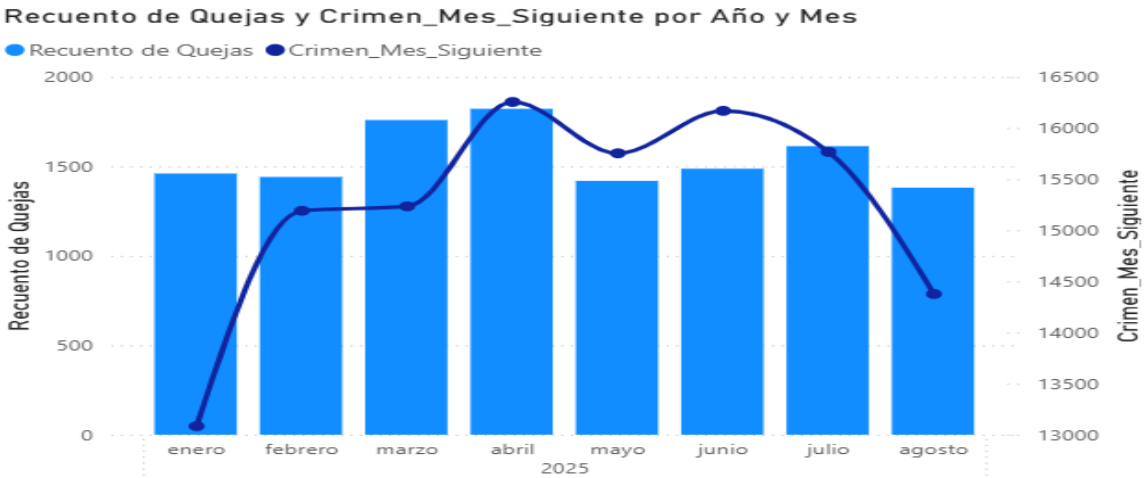

Figura 1.a *Validación de la Hipótesis. Las barras azules (Quejas por Grafitis) anticipan la tendencia de la línea (Crímenes de Alto Riesgo en T+1).*

Evidencia: La fuerte sincronía visual valida la hipótesis: el deterioro del entorno fluctúa en sintonía con el riesgo, actuando como señal de alerta temprana.

1.4. Impacto Operativo

El modelo facilita el paso de la reacción a la **prevención**, permitiendo una asignación dinámica de patrullas en focos de vandalismo. Esto genera un **doble impacto**:

- **Operativo:** Reduce la criminalidad atacando su causa raíz (el entorno) antes de que escale.
- **Económico:** Permite ahorrar dinero público, ya que es mucho más barato limpiar una pared a tiempo que tener que enviar patrullas de policía urgentes cuando el problema ya es grave.

2. Documentación Técnica

2.1. Arquitectura y Herramientas

Se ha implementado una arquitectura ELT (Extract, Load, Transform) nativa en la nube utilizando **Microsoft Azure**, lo que garantiza la escalabilidad y automatización del procesamiento.

Componentes del Stack Tecnológico:

1. **Origen de Datos:** Archivos planos (.csv) de *NYC OpenData*, ingesta manual a un contenedor de almacenamiento (Blob Storage) que actúa como *Data Lake*.
2. **Orquestación y Limpieza: Azure Data Factory (ADF)**. Se utiliza el motor de **Mapping Data Flows** para realizar las transformaciones complejas sin necesidad de código (No-Code/Low-Code).
3. **Almacenamiento Estructurado: Azure SQL Database** como repositorio final para el consumo analítico.
4. **Visualización:** Power BI conectado a la base de datos SQL.

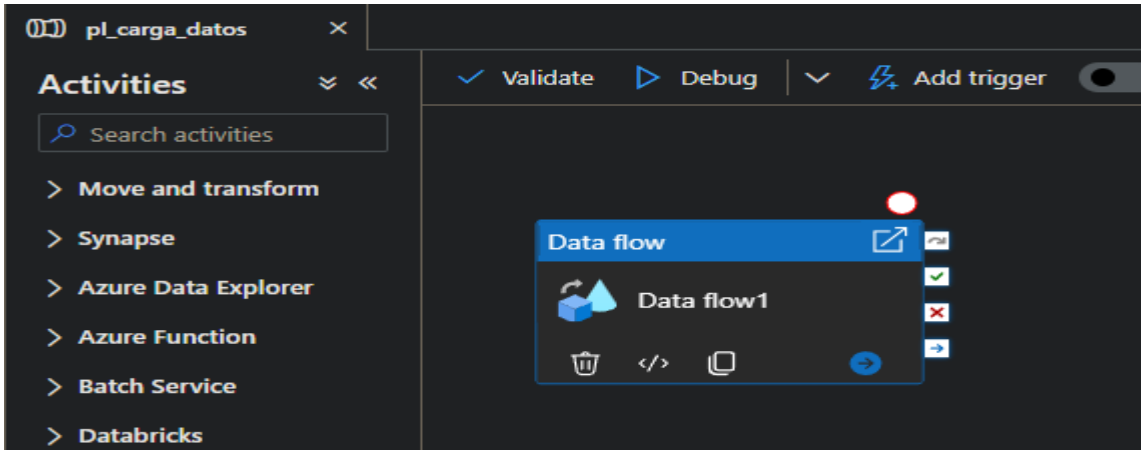

Figura 2. Pipeline principal en Azure Data Factory encargado de la ejecución del proceso.

2.2. Modelo de Datos y Relaciones (Star Schema)

Para poder analizar conjuntamente los datos de criminalidad y las quejas ciudadanas, he diseñado en Power BI un modelo relacional tipo **Estrella (Star Schema)**. Este diseño facilita que los filtros funcionen de forma fluida y rápida.

Estructura del Modelo (ver Figura 3): El modelo conecta dos tablas de hechos mediante dimensiones compartidas ("conformed dimensions"), lo que permite cruzar la información:

1. **Tablas de Hechos (Facts):** Son las tablas transaccionales cargadas desde SQL.
    - TablaCrimenes: Contiene cada delito registrado con su ubicación y fecha.
    - TablaQuejas: Contiene los reportes del 311.

- *Nota:* Ambas tablas almacenan las métricas clave como el Indice_Criminalidad y el Indice_Deterioro.

2. **Dimensiones Compartidas (Dims):**
    - Calendario: Es la tabla maestra de fechas. Se relaciona con ambas hechos (relación one-to-many) para poder ver la evolución temporal y aplicar el desfase (Lag).
    - Dim_Barrio y Maestra_Zonas: Permiten filtrar geográficamente. Al seleccionar un barrio (ej: Bronx), el filtro viaja hacia abajo y recorta tanto la tabla de crímenes como la de quejas simultáneamente.

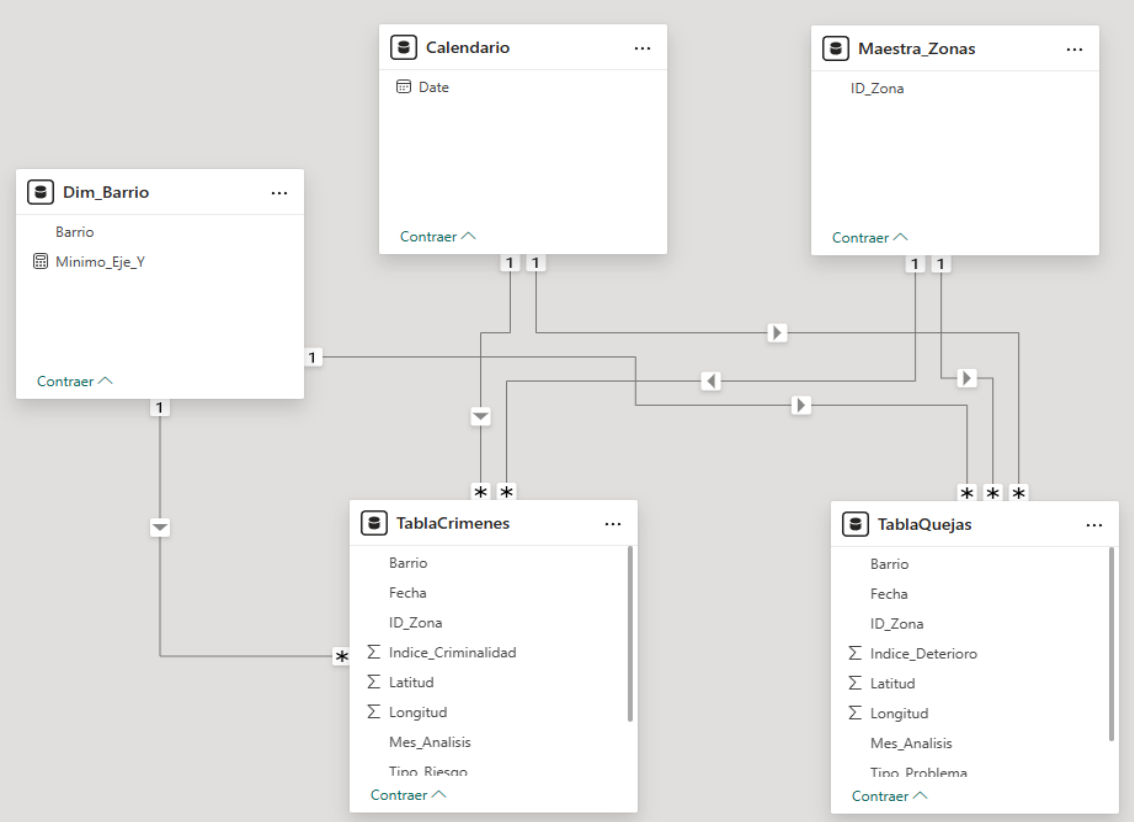

Figura 3. Vista del modelo de datos en Power BI. Las dimensiones superiores filtran a las dos tablas de hechos inferiores.

2.3. Procesos ETL y Limpieza con Data Flows

Para la fase de limpieza y transformación, he utilizado los **Mapping Data Flows** de Azure Data Factory. Esta herramienta me ha permitido diseñar visualmente el flujo de datos sin necesidad de escribir código complejo, procesando las dos tablas (Quejas y Crímenes) de forma paralela.

Lógica del Flujo de Datos (ver Fig. 4): Tal y como se observa en el diagrama diseñado, el proceso sigue 5 pasos secuenciales para ambas ramas:

1. **Fuentes (Source):** Conexión y lectura de los archivos CSV originales (FuenteCrimenes y FuenteQuejas) desde el almacenamiento.

2. **Limpieza y Transformación (Derived Column):** En el paso *LimpiarDatos*, he creado columnas nuevas y corregido formatos (por ejemplo, convertir textos a fechas válidas) para homogeneizar los datos.
3. **Filtrado de Calidad (Filter):** En el paso *FiltrarNulos*, he aplicado una regla para eliminar las filas que no tenían coordenadas geográficas (Lat/Lon), ya que no sirven para el mapa.
4. **Selección de Columnas (Select):** En el paso *EliminarColumnas*, he descartado todas las variables que no aportaban valor al análisis, quedándome solo con lo esencial para aligerar el modelo.
5. **Carga (Sink):** Finalmente, los datos limpios se insertan en las tablas de destino de Azure SQL (DestinoSQLCrimenes y DestinoSQLQuejas).

Figura 4. Diagrama del Data Flow en Azure Data Factory. Se muestra la limpieza paralela de ambas fuentes de datos.

2.4. Lógica del Análisis: Implementación de Desfase Temporal

Para validar la hipótesis predictiva, se ha aplicado una técnica de **correlación con desfase temporal (Time-Lag Analysis)**.

Esta técnica consiste en realizar un desplazamiento temporal de los datos de criminalidad. En lugar de cruzar las fuentes por su fecha natural, el modelo asocia los volúmenes de quejas del periodo actual (T) con los índices delictivos del mes siguiente (T+1).

Esta metodología permite visualizar en el mismo eje temporal la relación entre **causa** (deterioro urbano) y **efecto** (incidencia delictiva futura), facilitando la identificación de patrones preventivos.

2.5. Problemas de Calidad Detectados y Soluciones

Durante la exploración inicial de los archivos (Profiling), se detectaron bloqueos técnicos que impedían la carga directa. Se resolvieron íntegramente dentro del Data Flow:

1. **Exceso de Dimensionalidad (48 columnas):** Los archivos originales contenían un exceso de columnas irrelevantes (48 campos) que ralentizaban el proceso.

- *Solución:* Se aplicó una transformación Select para podar el dataset y mantener estrictamente los campos necesarios.

2. **Inconsistencia de Tipos (Casting):** Las coordenadas (Latitude/Longitude) venían formateadas como texto (*String*) y las fechas presentaban formatos mixtos imposible de ordenar.
    - *Solución:* Se realizó una conversión de tipo a Decimal y una estandarización de fechas antes de la carga.

3. **Filtrado Cruzado (Cross-Filtering): I**nicialmente, al intentar visualizar Quejas y Crímenes en el mismo gráfico, los filtros de fecha y zona no funcionaban correctamente entre ambas tablas.
    - *Solución:* Se implementó un **Modelo en Estrella (Star Schema)** con dimensiones compartidas. Esto permitió que un solo filtro (ej: "Bronx") propagara la selección a ambas tablas de hechos simultáneamente y de forma fluida.

3. Conclusiones y Líneas Futuras

3.1. Conclusiones del Proyecto

El análisis de los datos históricos de Nueva York ha permitido validar satisfactoriamente la hipótesis planteada. Tras procesar y correlacionar las bases de datos de servicios urbanos y criminalidad, se extraen las siguientes conclusiones:

1. **Validación de la Teoría:** Se confirma que el vandalismo (especialmente grafitis) actúa como un indicador adelantado del delito. Los picos de quejas en una zona suelen preceder a un aumento de la criminalidad grave en el mes siguiente (T+1).
2. **Valor del Dato Abierto:** Los portales *OpenData* han demostrado ser una fuente fiable y rica para la inteligencia empresarial, permitiendo generar valor real sin coste de adquisición de datos.
3. **Nuevo Enfoque de Gestión:** La herramienta demuestra la viabilidad técnica de transicionar de un modelo de seguridad reactivo (esperar a la llamada del 911) a uno preventivo (actuar sobre el entorno urbano).

3.2. Próximos Pasos

Para escalar esta prueba de concepto (PoC) a un entorno de producción real, se proponen las siguientes evoluciones:

- **Automatización Total (API):** Sustituir la descarga manual de archivos CSV por una conexión automatizada con la API de *NYC OpenData* mediante **Azure Logic Apps**. Esto permitiría una actualización diaria de los datos sin intervención humana.
- **Alertas en Tiempo Real:** Configurar el servicio de **Power BI Service** para enviar alertas automáticas al correo de los responsables de zona cuando el volumen de grafitis supere un umbral crítico de riesgo.