Big Data: Diseño y pruebas

1. Introducción

Nuestro proyecto se centra en el análisis y la predicción de accidentes de tráfico en España, utilizando datos abiertos proporcionados por la Dirección General de Tráfico (DGT). El objetivo principal es identificar patrones relevantes que faciliten la toma de decisiones orientadas a mejorar la seguridad vial.

Elegimos este caso de uso por la alta relevancia social y económica, dado que la siniestralidad vial afecta a miles de personas cada año. Además, por la disponibilidad de datasets amplios y detallados accesibles online.

El proyecto integra varias fuentes de datos para conformar un dataset enriquecido con más de 97,000 registros y 76 variables relacionadas con accidentes de tráfico. Para ello, utilizamos Apache Spark para la ingesta, limpieza y transformación de los datos, Python para el análisis estadístico y Power BI para desarrollar un dashboard interactivo. Los objetivos fundamentales son analizar la distribución temporal y espacial de los accidentes, identificar factores de riesgo y predecir el volumen de siniestros por provincia y mes.

2. Fuentes de datos

Usamos un dataset base:

Dataset: Accidentes en España en el Año 2022 (Excel).
Fuente: Datos abiertos de la DGT.
Dimensiones: Sobre 90.000 registros y 73 columnas.
Información columnas:

- Ubicación (provincia, municipio, km, carretera, etc.)
- Instante (año, mes, día de la semana)
- Condiciones del entorno (iluminación, meteorología, visibilidad, aceras, etc.)
- Accidente (tipo de accidente, vehículos implicados, etc.)
- Víctimas y gravedad en 24 horas y en 30 días (muertos, heridos leves, graves)

Y lo enriquecemos con un dataset complementario con los datos de circulación mensuales:

Dataset: Tráfico mensual por estaciones (PDF).
Fuente: Ministerio de Transportes.
Dimensiones: Sobre 900 registros y 18 columnas.
Información columnas:

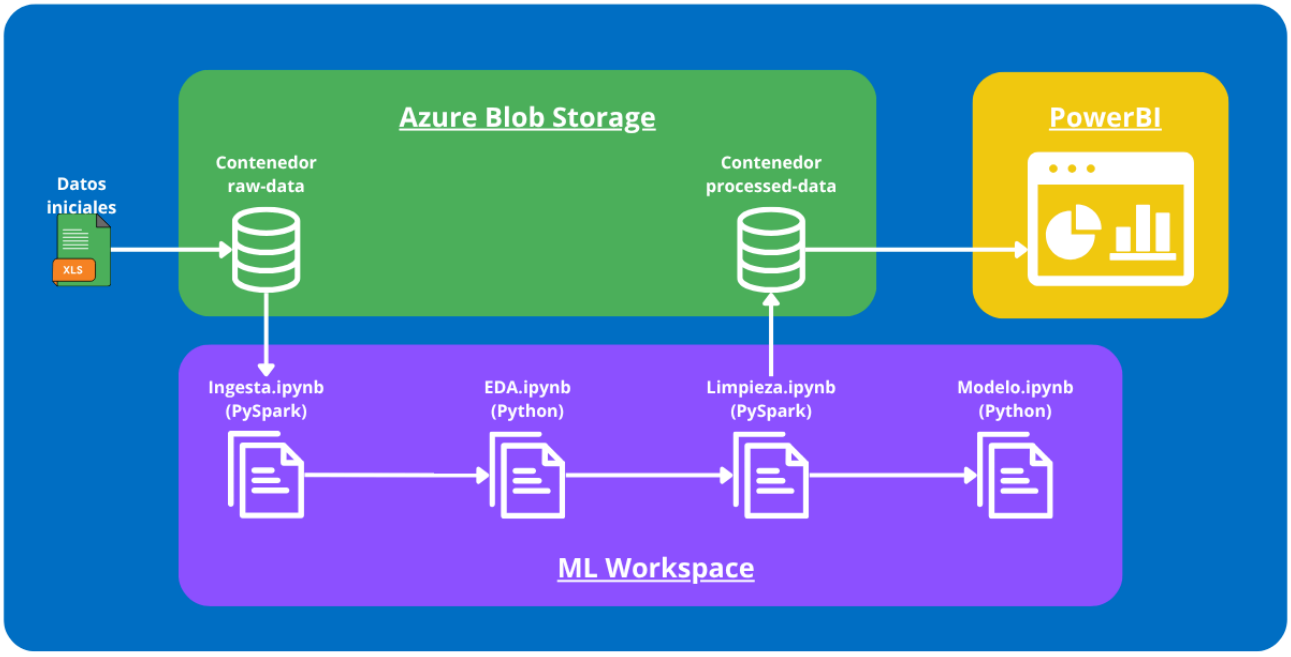
- Ubicación (estación, provincia, vía, punto kilométrico)
- Mes del año con la información del tráfico.
- Tipo de vehículo (ligero o pesado)

Caso de uso: Identificar patrones relevantes que faciliten la toma de decisiones orientadas a mejorar la seguridad vial.
Objetivos del análisis:

- Analizar la distribución temporal y geográfica de los accidentes de tráfico en España.
- Identificar patrones y factores de riesgo asociados a la ocurrencia de accidentes.
- Predecir el volumen de accidentes por provincia y mes para apoyar la planificación y prevención.
- Visualizar insights clave mediante el dashboard interactivo para facilitar la interpretación de los datos.

4. Proceso ETL

Usamos es Stack de Azure para nuestra ETL, nuestro flujo es:



Herramientas utilizadas:

- Apache Spark (PySpark): Procesamiento distribuido de datos, limpieza y transformación.
- Python: EDA y Modelos.
- Azure Blob Storage: Almacenamiento intermedio en la nube.
- Power BI: Visualización interactiva.

5. Plan de pruebas

Tipo de prueba	Descripción	Criterio de éxito
Integridad	Validar que no haya datos nulos en campos clave	Reducción no nulos
Unicidad	Verificar que no haya duplicados	Sin duplicados
Validez	Verificación de formatos válidos	Los campos tienen el formato correcto
Integridad	Campos categóricos están bien referenciados	Coincide con dataset original

6. Revisión de la Calidad del Dato

- Revisión de la calidad del dato:

Prueba	Descripción	Resultado
Unicidad por ID accidente	Comprobar que no existen duplicados	Sin duplicados
Completitud	Deteccion de campos con valores nulos	KM,ISLA,NUDO_INFO, CARRETERA_CRUCE, CONDICION_VIENTO,CONDICION_NIEBLA
Integridad	Son lógicos los valores de COD_MUNICIPIO, de los campos categóricos y temporales	Sin anomalías
Validez	Validar el COD_MUNICIPIO	Sin anomalías

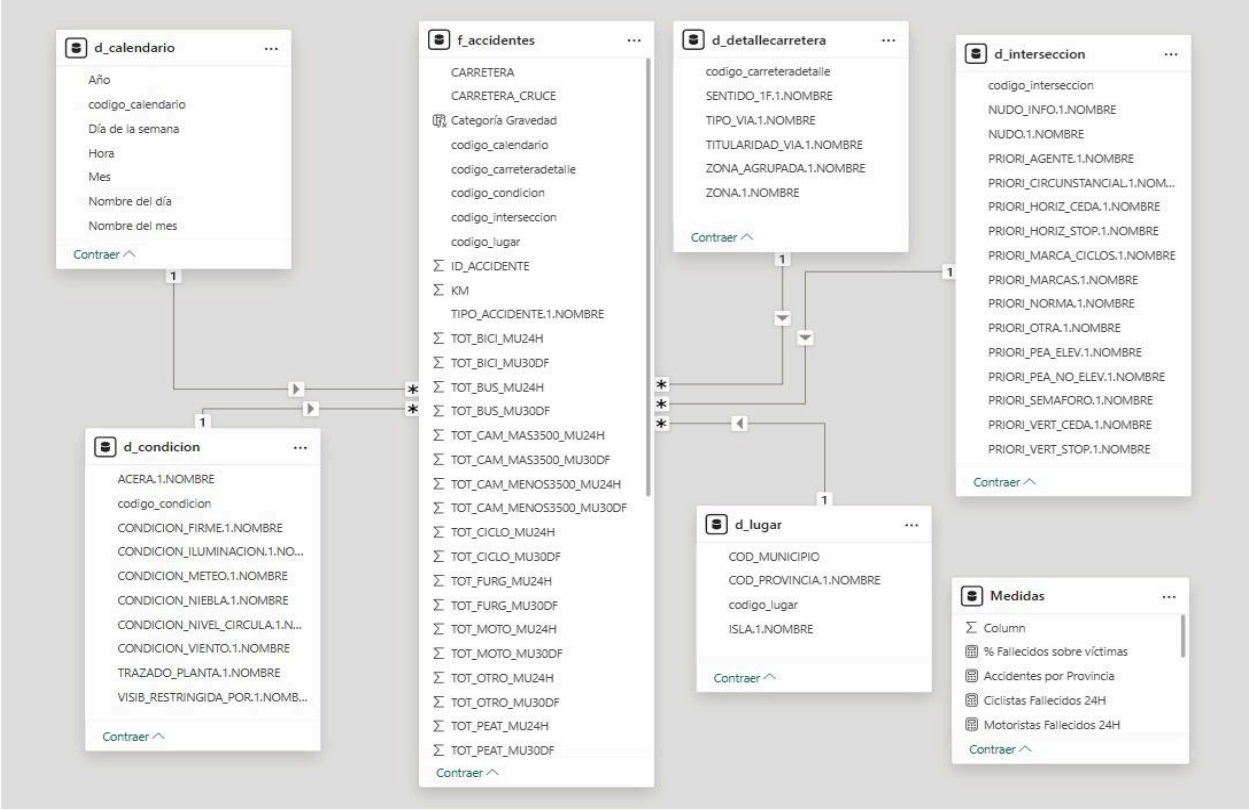
- Herramientas usadas: Pyspark y python.

- Acciones tomadas para limpiar o corregir los datos:
    - Unicidad: Eliminación de duplicados.
    - Completitud: Sustitución de nulos en los campos ISLA, NUDO_INFO, CONDICION_NIEBLA y CONDICION_VIENTO.

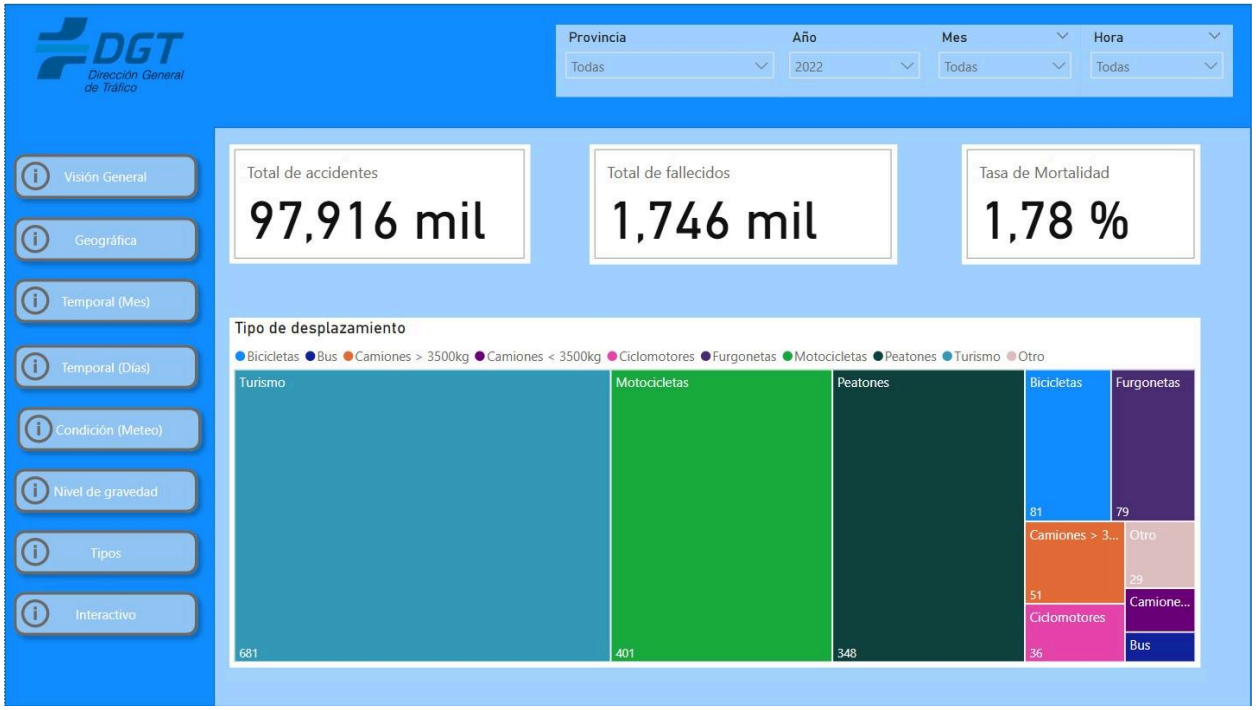
- Integridad: Corrección e imputación de la variable ISLA a partir de código postal y prefijo de carretera para provincias 7, 35 y 38.

7. Dashboard en Power BI

Modelo Estrella



Cuadro de mando principal



8. Conclusiones

El desarrollo del presente proyecto ha permitido construir un pipeline completo de Big Data aplicado al análisis de los accidentes de tráfico en España, desde la ingesta hasta la visualización avanzada con Power BI. A través de este proceso, se ha logrado:

- Estandarizar y enriquecer los datos provenientes de diferentes fuentes (accidentes, tráfico por estaciones, codificación territorial).
- Crear un modelo predictivo capaz de estimar la gravedad de un accidente según múltiples factores, lo que puede ser útil para la toma de decisiones preventivas.
- Diseñar un dashboard de análisis exploratorio y estratégico para diferentes perfiles: ciudadanía, administración pública o cuerpos de seguridad.

Valor de los casos de uso

Los casos de uso planteados son especialmente relevantes por los siguientes motivos:

1. **Impacto social**: Comprender dónde, cómo y por qué ocurren los accidentes más graves ayuda a tomar decisiones con alto valor público (mejoras en infraestructuras, campañas de concienciación...).

2. **Automatización del análisis**: Los dashboards permiten una revisión rápida y visual de patrones mensuales, territoriales y por tipo de vía o vehículo.

3. **Aplicación predictiva**: El modelo de clasificación puede evolucionar hacia sistemas de alerta temprana en función de factores meteorológicos, tráfico o zona.

Posibilidades de mejora

Existen diversas áreas de mejora y expansión:

- **Ampliación histórica del dataset**: El portal de la DGT publica datos desde hace varios años. Incluir históricos permitirá entrenar modelos más robustos y analizar tendencias temporales más amplias.

- **Incorporación de nuevos datasets**: La DGT ofrece información adicional como sanciones, puntos de carnet, atestados, ciclistas, peatones, y datos por tramo de vía. Esto permitiría:

    - Enriquecer las dimensiones del modelo.

    - Explorar nuevas correlaciones (por ejemplo, entre sanciones y siniestralidad).

    - Generar modelos más explicativos y ajustados a la realidad.

- **Integración de datos meteorológicos o eventos**: El cruce con fuentes como AEMET podría abrir líneas de análisis estacionales o de comportamiento bajo condiciones adversas.

Próximos pasos recomendados

- Automatizar el pipeline de ETL para recoger los nuevos datos anuales.

- Versionar los modelos y comparar su rendimiento con nuevos atributos o fuentes externas.

- Explorar la publicación del dashboard como servicio web para permitir acceso público o por perfiles definidos.