Proyecto Integrado (EA1). Formulación de una necesidad de ingenieria de datos

JUAN DAVID PEÑALOZA HERNANDEZ

RAFAEL DE JESUS MOLINA HAECKERMANN

Andres Felipe Callejas Jaramillo

Proyecto Integrado V - Linea de Énfasis (andres felipe callejas) - PREICA2502B020074

Institución Universitaria Digital

Facultad de Ingeniería y Ciencias Agropecuarias

Ingeniería de Software y Datos

Semestre X

Índice (Tabla de Contenido)

Lista de Tablas

Lista de Figuras

Resumen

La priorización de recursos para la seguridad vial en Colombia enfrenta el desafío de datos dispersos, lo que dificulta la toma de decisiones informada. Este proyecto de ingeniería de datos aborda esta necesidad mediante la creación de una solución local para el análisis de siniestralidad. El objetivo es diseñar e implementar un pipeline de datos que transforma el dataset público "Sectores Críticos de Siniestralidad Vial" (publicado por el Ministerio de Transporte) desde un formato CSV estático a una base de datos estructurada y relacional en SQLite. La metodología sigue un plan de proyecto definido en un diagrama de Gantt, abarcando la exploración del dataset, el diseño del modelo de datos, la ingesta y la validación. Como resultado principal, se genera una base de datos consultable y un reporte CSV priorizado que identifica los sectores con mayor número de fallecidos. Esta solución técnica proporciona a las secretarías de movilidad una herramienta ágil para fundamentar sus decisiones de intervención, optimizando el gasto en infraestructura y salvando vidas.

Palabras clave: ingeniería de datos, SQLite, seguridad vial, siniestralidad, pipeline, CSV, analítica.

Abstract

Resource prioritization for road safety in Colombia faces the challenge of scattered data, hindering informed decision-making. This data engineering project addresses this need by creating a local solution for road crash analysis. The objective is to design and implement a data pipeline that transforms the public "Critical Road Crash Sectors" dataset (published by the Ministry of Transport) from a static CSV format into a structured, relational SQLite database. The methodology follows a project plan defined in a Gantt chart, covering dataset exploration, data model design, ingestion, and validation. As a primary outcome, a queryable database and a prioritized CSV report are generated, identifying the sectors with the highest number of fatalities. This technical solution provides mobility secretariats with an agile tool to substantiate their intervention decisions, thereby optimizing infrastructure spending and saving lives.

Keywords: data engineering, SQLite, road safety, crash analysis, pipeline, CSV, analytics.

1. Introducción y Objetivos

1.1. Objetivo General

Desarrollar una solución de ingeniería de datos local, utilizando SQLite y Python, que centralice el dataset de "Sectores Críticos de Siniestralidad Vial" para facilitar la consulta, el análisis y la generación de reportes priorizados, apoyando así la toma de decisiones basada en evidencia de las entidades de movilidad.

1.2. Objetivos Específicos

- Formular un caso de uso enfocado en la priorización de intervenciones de seguridad vial, a partir de un dataset público del gobierno colombiano.
- Planificar las fases, tareas y dependencias del proyecto mediante un diagrama de Gantt.
- Diseñar e implementar una base de datos local en SQLite que modele y almacene de forma estructurada los datos de siniestralidad.
- Construir un script (en Python) que automatice la ingesta de los datos desde el archivo CSV original (`SECTORES_CRITICOS_DE_SINIESTRALIDAD_VIAL_20251109.csv`) a la base de datos SQLite.
- Generar un reporte en formato CSV que contenga los 20 sectores viales más críticos, ordenados por el número de fallecidos, como un producto de datos accionable.
- Documentar el proyecto, incluyendo la fuente de los datos, la metodología y los artefactos técnicos (scripts, CSV, BD) en un repositorio de GitHub.

2. Metodología

Para el desarrollo de este proyecto se siguió un enfoque estructurado de ingeniería de datos, dividido en cuatro fases secuenciales que abarcan desde la definición del problema hasta la visualización de resultados.

2.1. Fase 1: Definición y Planificación del Proyecto

En esta etapa inicial se seleccionó el dataset *"Sectores Críticos de Siniestralidad Vial"* y se formuló la necesidad de negocio enfocada en la priorización de recursos de seguridad vial. Se definieron las variables clave (`fallecidos`, `departamento`, `municipio`, `gizscore`) y se estableció el cronograma de trabajo mediante un diagrama de Gantt.

2.2. Fase 2: Diseño y Construcción de la Base de Datos (Ingeniería)

Esta fase concentró el desarrollo técnico del pipeline de datos utilizando **Python (Pandas/NumPy)** y **SQLite**. Se ejecutaron los siguientes procesos de limpieza y enriquecimiento:

1. **Ingesta y Normalización:** Se cargó el dataset original tratando los valores `"<Null>"` como nulos. Se estandarizaron los nombres de las columnas al formato *snake_case* (minúsculas, sin espacios ni tildes, ej. `fecha_accidente`) para garantizar la consistencia técnica.
2. **Limpieza de Datos:** Se implementaron scripts para la detección y eliminación de registros duplicados, asegurando la unicidad de los sectores críticos reportados. 3. Enriquecimiento Temporal (Feature Engineering): Dado que el dataset original carecía de una serie temporal histórica, se generó una variable sintética `fecha` con valores aleatorios entre el 1 de enero de 2022 y el 31 de diciembre de 2024. A partir de esta, se derivaron las nuevas columnas `anio`, `mes` y `dia`, permitiendo transformar un listado estático en un dataset apto para análisis de tendencias. 4. Almacenamiento: El dataset procesado se estructuró y almacenó tanto en una base de datos local SQLite (`proyecto.db`) como en un archivo CSV enriquecido (`dataset_enriquecido.csv`).

2.3. Fase 3: Procesamiento y Generación de Resultados (Análisis)

Se llevó a cabo un Análisis Exploratorio de Datos (EDA) utilizando las librerías **Matplotlib** y **Seaborn** para interpretar el comportamiento de las variables:

- **Análisis Univariado:** Se generó un histograma de la variable `fallecidos`, evidenciando la distribución de la severidad de los accidentes y detectando valores atípicos (sectores con alta mortalidad).
- **Análisis Geográfico:** Mediante gráficos de barras, se identificaron los 10 departamentos con mayor concentración de puntos críticos.
- **Análisis Bivariado:** Se correlacionó la intensidad estadística (`gizscore`) con el número de víctimas mediante diagramas de dispersión.
- **Análisis Temporal:** Utilizando las variables enriquecidas, se graficó la tendencia anual simulada de la siniestralidad para los periodos 2022-2024.

2.4. Fase 4: Documentación y Control de Versiones (Entrega)

Todo el código fuente, los datasets (crudo y enriquecido) y los gráficos generados fueron versionados en un repositorio de **GitHub**. La documentación se consolidó en este informe bajo normas APA y en el archivo `README.md` del repositorio, garantizando la reproducibilidad del ejercicio.

3. Resultados

(Como se solicitó, solo se incluye el título para la Etapa 1)

(Esta sección se desarrollará en la Etapa 3 del proyecto, una vez que la base de datos esté construida y se hayan generado las exportaciones CSV.)