Proyecto de Integración y Automatización de Datos para IA para el proyecto de "Diseño e implementación de un mecanismo ETL que permita la integración de fuentes públicas del gobierno local para estudios de siniestralidad vial en Bogotá"

Natalia Garzón Arias[1]

Universidad Central
Maestría en Analítica de Datos
Curso de Automatización e Integración de datos
Bogotá, Colombia
ngarzona@ucentral.edu.co

December 1, 2023

Contents

1 Introducción

Para Colombia, la reducción de muertes y lesionados graves derivados de accidentes viales ha sido una prioridad. Muestra de ello es la creación y publicación de la Ley 1702 por la cual se crea la Agencia Nacional de Seguridad Vial (Congreso de Colombia, 2013). Sin embargo, estos esfuerzos del gobierno local y nacional no parecen coincidir con las alarmantes cifras de siniestralidad en las vías.

De acuerdo con estadísticas ofrecidas por el Observatorio Nacional de Seguridad Vial, de la ANSV, en los últimos 5 años las víctimas fatales han aumentado en un 27% a nivel nacional y un 11% para Bogotá. Esto confirma que sigue siendo un reto importante la identificación de zonas con mayor riesgo de accidentalidad y la caracterización de los factores que contribuyen a su ocurrencia dentro de la ciudad (Agencia Nacional de Seguridad Vial [ANSV], s.f.).

Esta preocupación da origen al proyecto de investigación que recibe como título: "Implementation of an intelligent transportation system (ITS) for the prediction of traffic accident risk in Bogotá D.C., Colombia", en adelante nombrado proyecto ITS, con el cual sus investigadores, buscan desarrollar un prototipo de un sistema de transporte inteligente ITS capaz de predecir y mostrar en tiempo real los niveles de riesgo de accidentes de tránsito en la ciudad de Bogotá, Colombia, a través de la ubicación geográfica.

De esta manera, el presente trabajo de grado será enfocado en la participación de dos de las fases mencionadas en el proyecto ITS: (1) la recopilación de datos de las bases de datos de las instituciones del gobierno local y (2) la construcción del dataset a través de la integración y preprocesamiento de los datos recopilados. Con este planteamiento se dará respuesta a la pregunta: ¿En qué condiciones es posible diseñar e implementar un mecanismo ETL que permita la integración de fuentes públicas, heterogéneas y dispuestas por el gobierno local para considerarse como insumo de estudios de siniestralidad vial en Bogotá?

Para abordar la primera fase, se propone integrar datos de acceso público y disponibles a través de diferentes portales web, por mencionar algunos de ellos serían: la Secretaría Distrital de Movilidad (SDM) [**?**], la Infraestructura de Datos Espaciales para el Distrito Capital (IDECA) y la Dirección de Tránsito y Transporte de la Policía Nacional.

Para la segunda fase, se propone desarrollar un mecanismo bajo una política de gobierno de datos con lineamientos claros para la adquisición, tratamiento y almacenamiento de los datos, asegurando su calidad, consistencia y confidencialidad de manera que puedan estar disponibles para estudios posteriores tales como: construcción de modelos predictivos de accidentes de tráfico, identificación de puntos calientes en la ciudad y cualquier tipo de análisis que facilite la toma de decisiones en materia de seguridad vial en Bogotá. Así como la generación

de estrategias que promuevan la reducción de víctimas en las vías derivado de los datos encontrados, pues como lo demuestran las cifras, durante los meses de enero y febrero del año 2023, los accidentes de tráfico en Colombia han resultado en la muerte de 1.317 personas, lo que indica un aumento del 19,1% con respecto al promedio de los últimos cinco años (Nova, D., 2023).

Por tanto, se puede afirmar que con el desarrollo del presente trabajo se aporta al logro de los objetivos del proyecto ITS, al permitir la integración de fuentes de datos relevantes para el estudio de la siniestralidad vial en Bogotá.

En las siguientes páginas el lector encontrará la metodología utilizada para la consecución de los objetivos trazados, así como los resultados y consideraciones de este trabajo, destacando su contribución al avance del conocimiento en el campo de la analítica de datos aplicada a la seguridad vial.

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA

La industria 4.0, donde la analítica sigue jugando un papel fundamental, hoy permea todos los ámbitos, incluidos conceptos sociales y políticos como lo es la salud pública. Pues de acuerdo con la Organización Mundial de la Salud (OMS) los accidentes viales son considerados una de las causas principales de mortalidad en jóvenes. Es entonces como el presente trabajo de grado, propone una iniciativa basada en datos para abordar la problemática de la siniestralidad vial en Bogotá, Colombia y así facilitar la toma de decisiones en esta materia. Pues el propósito se centra en la creación de un mecanismo de Extracción, Transformación y Carga (ETL) para la obtención de información proveniente de diversas fuentes públicas. En la actualidad diferentes organismos gubernamentales han enfocado sus esfuerzos en recolectar, almacenar y poner a disposición de la comunidad información relacionada con la seguridad vial. Ejemplo de ello es el Observatorio Nacional de Seguridad Vial (ONSV), quien se encarga del manejo y la gestión de la información y el conocimiento relacionado con la seguridad vial en el país quien a través de diferentes elementos expone las estadísticas de siniestralidad vial. Por otra parte la Secretaría Distrital de Movilidad, a través de su Sistema Integrado de Información sobre Movilidad Urbana Regional (SIMUR) pone a disposición información asociada al comportamiento de la movilidad en Bogotá. Esto deja en evidencia que no es la falta de información lo que hace que la toma de decisiones no se haga fundamentada en datos, sino que los datos se encuentran aislados dificultando ver las relaciones que unas variables presentan sobre otras. Por esto, establecer un sistema ETL eficiente y adaptable que abarque la recopilación de datos provenientes de fuentes gubernamentales, seguida por la creación de un conjunto de datos coherente mediante procesos de integración y preprocesamiento a través de algoritmos

de limpieza y transformación de datos para asegurar su calidad y consistencia contribuye al análisis de siniestralidad en Bogotá.

2.1 Titulo del proyecto de investigación

Diseño e implementación de un mecanismo ETL que permita la integración de fuentes públicas del gobierno local para estudios de siniestralidad vial en Bogotá.

2.2 Objetivo general

Diseñar e implementar un mecanismo ETL (Extracción, Transformación y Carga) que, apoyado en una política de gobierno de datos, permita la integración de datos de fuentes públicas, heterogéneas y dispuestas por el gobierno local para que puedan ser insumo de estudios de siniestralidad vial en Bogotá.

2.2.1 Objetivos especificos

- Identificar fuentes de datos públicas que puedan contribuir al análisis de siniestralidad en Bogotá, a través de una distinción detallada de su estructura y calidad de la información contenida.
- Desarrollar un mecanismo de extracción de datos que permita obtener la información relevante de las fuentes identificadas y transformarla en un formato adecuado para su integración en el modelo de datos diseñado.
- Definir una política de gobierno de datos que permita establecer los criterios y estándares para la integración de las fuentes de datos heterogéneas en el proyecto, asegurando la consistencia y precisión de la información.
- Implementar las reglas de negocio definidas, garantizando así un proceso de transformación (limpieza, validación, creación, entre otros) de datos que asegure calidad y coherencia de la información integrada.
- Disponer los datos integrados a través de una plataforma que ofrezca información confiable y disponible para ser insumo de estudios de siniestralidad vial en Bogotá.

2.3 Alcance

Al llevar a cabo la propuesta de desarrollo de la ETL, se espera obtener una herramienta que permita integrar y transformar los datos de fuentes heterogéneas y públicas para obtener información útil y valiosa que soporte la toma de decisiones en el ámbito de la movilidad urbana y la seguridad vial. Entre los resultados concretos que se esperan obtener al implementar la ETL se incluyen:

- Aportar al logro de los objetivos del proyecto ITS: al permitir la integración de fuentes de datos relevantes para el estudio de la siniestralidad vial en Bogotá, iniciando con la recopilación de datos de las bases de datos

de las instituciones del gobierno local y concluyendo con la construcción del dataset a través de la integración y preprocesamiento de los datos recopilados en la fase previamente descrita.

- Mejora en la eficiencia y productividad: Al automatizar el proceso de integración de datos, se pueden ahorrar muchas horas de trabajo manual. Esto permite que los analistas de datos y otros profesionales puedan dedicar más tiempo a analizar los datos y tomar decisiones informadas, en lugar de destinar tiempo en tareas repetitivas y tediosas.

- Identificación de patrones y tendencias: al analizar los datos de movilidad urbana y seguridad vial desde diferentes perspectivas, se podrían identificar patrones y tendencias que permitan comprender mejor el comportamiento de los usuarios de la vía, las causas de los accidentes y otros aspectos relevantes para la toma de decisiones.

- Mejora de la toma de decisiones: al contar con información actualizada y detallada sobre el comportamiento de los usuarios del transporte público y la seguridad vial en la ciudad, se podrían tomar decisiones más informadas y eficaces en cuanto a políticas públicas de transporte y seguridad vial.

- Optimización de recursos: con la información proporcionada por la ETL, se podrían identificar áreas de la ciudad que requieren más recursos en términos de transporte público y seguridad vial, lo que permitiría una mejor adquisición de recursos y una mayor eficiencia en la planificación urbana.

- Mejora de la seguridad vial: al analizar los datos de accidentes de tráfico, se podrían identificar las áreas de la ciudad con mayor riesgo.

2.4 Pregunta de investigación

¿En qué condiciones es posible diseñar e implementar un mecanismo ETL que permita la integración de fuentes públicas, heterogéneas y dispuestas por el gobierno local para considerarse como insumo de estudios de siniestralidad vial en Bogotá?

2.5 Hipotesis

A pesar de que el gobierno local recopila, almacena y pone a disposición datos relacionados con accidentes de tráfico y siniestralidad vial a través de diferentes fuentes de acceso público, como el Instituto de Desarrollo Urbano de Bogotá (IDU), la Secretaría Distrital de Movilidad (SDM), la Infraestructura de Datos Espaciales para el Distrito Capital (IDECA) y la Dirección de Tránsito y Transporte de la Policía Nacional, existe un desafío significativo en la integración y consolidación de estos datos dispersos en un formato coherente y accesible. La falta de un mecanismo robusto y eficiente de Extracción, Transformación y Carga (ETL)

dificulta la obtención de información completa y precisa para llevar a cabo estudios detallados sobre la siniestralidad vial en Bogotá.

El proceso de recopilación, limpieza, transformación e integración de datos de diferentes fuentes es esencial para realizar análisis precisos y obtener información valiosa para la toma de decisiones en políticas de seguridad vial. Sin embargo, la ausencia de un mecanismo ETL especializado que permita la integración de las diversas fuentes públicas de datos del gobierno local limita la capacidad de investigadores, planificadores urbanos y responsables de políticas para comprender los patrones de siniestralidad vial, identificar áreas críticas y diseñar estrategias efectivas de prevención.

Por lo tanto, surge la necesidad de abordar este problema mediante la implementación de un mecanismo ETL que facilite la recopilación, transformación e integración de los datos de siniestralidad vial dispersos en diferentes fuentes públicas proporcionadas por el gobierno local de Bogotá. Este mecanismo debe ser capaz de estructurar los datos de manera coherente, eliminar duplicados, tratar inconsistencias y generar una base de datos unificada que permita un análisis más profundo y preciso de la siniestralidad vial en la ciudad.

En este contexto, la presente investigación tiene como objetivo diseñar e implementar un mecanismo ETL (Extracción, Transformación y Carga) que, apoyado en una política de gobierno de datos, aborde la problemática de la integración de datos de fuentes públicas, heterogéneas y dispuestas por el gobierno local para que puedan ser insumo de estudios de siniestralidad vial en Bogotá. El propósito de esta investigación es mejorar la disponibilidad y calidad de los datos utilizados en la toma de decisiones en materia de seguridad vial y planificación urbana.

3 Reflexiones sobre el origen de datos e información

3.1 ¿Cuál es el origen de los datos e información?

Los datos se encuentran distribuidos en diferentes repositorios y publicados a través del portal de datos abiertos de la Secretaría Distrital de Movilidad (SDM).

3.2 ¿Cuáles son las consideraciones legales o éticas del uso de la información?

Los datos abiertos son información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como "todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos". De este modo, la Ley establece la obligatoriedad de las entidades públicas de "divulgar datos abiertos", teniendo en cuenta las excepciones de acceso a la información, asociadas a información clasificada y reservada establecidas en su título tercero, artículos 18 y 19.

3.3 ¿Cuáles son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA?

El poyecto enfrenta los siguientes desafíos:

- **1. Calidad de datos:** Los datos pueden estar incompletos, duplicados o incorrectos. Limpiar y transformar estos datos de manera efectiva es el desafío crítico en este proceso ETL.
- **2. Variedad de datos:** La diversidad de fuentes con recopilación de información similar podría generar redundancia y discrepancia de los datos desde diferentes fuentes
- **3. Gestión de errores:** Los errores pueden ocurrir en cualquier etapa del proceso ETL. Diseñar mecanismos de detección y manejo de errores eficaces es esencial.
- **4. Monitorización y mantenimiento:** La monitorización constante de los flujos de datos y el rendimiento del proceso ETL es crucial para identificar problemas y mantener la integridad de los datos.
- **5. Documentación y trazabilidad:** Mantener un registro detallado de las transformaciones y cambios realizados en los datos es importante para la trazabilidad y la auditoría.

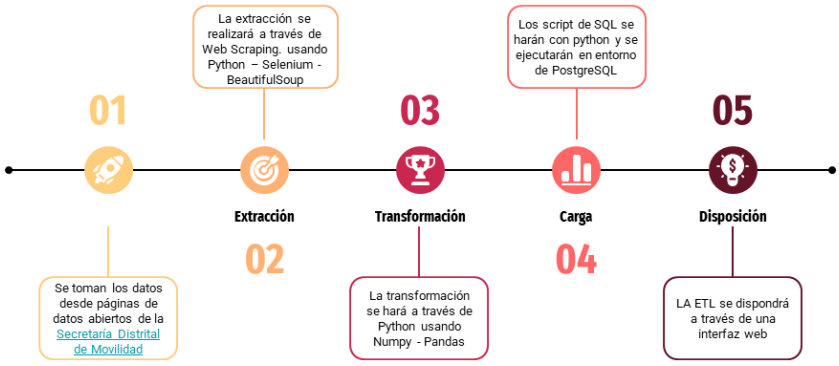
3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto?

Se espera que se libere la carga operativa de descargar, transformar y poner a disposición para su posterior utilización la información desde diferentes platafomas públicas. La obtención de los datos se realizará por medio de un proceso de *web scraping*, a través de Python y haciendo uso de la librería *Selenium*. Esta librería permitirá la automatización web emulando la interacción humana en un navegador. Las transformaciones para convertir, limpiar y enriquecer los datos provenientes de las diversas fuentes extraídas en la fase anterior se realizarán igualmente a través Python y haciendo uso de las librerías *Numpy y Pandas*. Finalmente para la carga se hace la creación de los archivos de sql a través de Python y se llevan a PostgreSQL para ser ejecutados desde Pg Admin o Shell.

4 Diseño de integración y Automatización de Datos para IA (Diagrama)

En la figura 1 se detalla el diagrama que orientará el proyecto en la automatización de extracción, transformación y carga de datos para crear una base de datos de la que se pueda extraer información relevante para analizar la siniestralidad en la ciudad de Bogotá.

Figure 1: Diagrama proyecto ETL



Nota. Conozca los datos abiertos de la Secretaría Distrital de Movilidad en el enlace disponible en el pie de imagen. [a]

[a] `https://datos.movilidadbogota.gov.co/search?groupIds=d3812f8315054cdc84cf744680103713`

5 integración de Datos (*Segunda entrega*)

Vulputate nec hac convallis rutrum eu ante volutpat aliquam ullamcorper pulvinar tristique velit nulla, cubilia felis tempor aptent vitae rhoncus parturient euismod mauris potenti dignissim magna. Nunc nec cum quisque quam tincidunt mauris nascetur conubia placerat fusce consequat eget erat vulputate, est rhoncus etiam dignissim luctus volutpat facilisi molestie torquent at convallis accumsan. Magnis morbi auctor sapien fusce turpis leo ad libero vivamus, sem enim ultrices elementum curae accumsan vel hendrerit. Etiam elementum dui a sodales auctor lacus proin interdum, porttitor netus tortor blandit sociis facilisi ullamcorper, mi aenean euismod diam placerat dignissim class.

6 Automatización de Datos (*Segunda entrega*)

Vulputate nec hac convallis rutrum eu ante volutpat aliquam ullamcorper pulvinar tristique velit nulla, cubilia felis tempor aptent vitae rhoncus parturient euismod mauris potenti dignissim magna. Nunc nec cum quisque quam tincidunt mauris nascetur conubia placerat fusce consequat eget erat vulputate, est rhoncus etiam dignissim luctus volutpat facilisi molestie torquent at convallis accumsan. Magnis morbi auctor sapien fusce turpis leo ad libero vivamus, sem enim ultrices elementum curae accumsan vel hendrerit. Etiam elementum dui a sodales auctor lacus proin interdum, porttitor netus tortor blandit sociis facilisi ullamcorper, mi aenean euismod diam placerat dignissim class.

7 IA (*Segunda entrega*)

Vulputate nec hac convallis rutrum eu ante volutpat aliquam ullamcorper pulvinar tristique velit nulla, cubilia felis tempor aptent vitae rhoncus parturient euismod mauris potenti dignissim magna. Nunc nec cum quisque quam tincidunt mauris nascetur conubia placerat fusce consequat eget erat vulputate, est rhoncus etiam dignissim luctus volutpat facilisi molestie torquent at convallis accumsan. Magnis morbi auctor sapien fusce turpis leo ad libero vivamus, sem enim ultrices elementum curae accumsan vel hendrerit. Etiam elementum dui a sodales auctor lacus proin interdum, porttitor netus tortor blandit sociis facilisi ullamcorper, mi aenean euismod diam placerat dignissim class.

8 Proximos pasos (*Tercera entrega*)

9 Lecciones aprendidas (*Tercera entrega*)

10 Bibliografía