

Proyecto de Integración y Automatización de Datos para el seguimiento y monitoreo de sinistros viales en la ciudad de Bogotá D.C

Javier Antonio Gonzalez¹,

Universidad Central

Maestría en Analítica de Datos

Curso de Automatización e Integración de Datos para Inteligencia Artificial

Bogotá, Colombia

jgonzalezg27@ucentral.edu.co

October 7, 2023

Contents

1	Introducción (Max 250 Palabras) - (<i>Primera entrega</i>)	3
2	Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA (Max 500 Palabras) - (<i>Primera entrega</i>)	3
2.1	Titulo del proyecto de investigación	3
2.2	Objetivo general (Max 100 Palabras) - (<i>Primera entrega</i>)	3
2.2.1	Objetivos especificos (Max 100 Palabras) - (<i>Primera entrega</i>)	3
2.3	Alcance (Max 200 Palabras) - (<i>Primera entrega</i>)	4
2.4	Pregunta de investigación (Max 100 Palabras) - (<i>Primera entrega</i>) .	4
2.5	Hipotesis (Max 100 Palabras) - (<i>Primera entrega</i>)	5
3	Reflexiones sobre el origen de datos e información (Max 400 Palabras) - (<i>Primera entrega</i>)	6
3.0.1	Datos geográficos	6
3.1	¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (<i>Primera entrega</i>)	7
3.2	¿Cuales son las consideraciones legales o éticas del uso de la información? (Max 100 Palabras) - (<i>Primera entrega</i>)	7
3.3	¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA? (Max 100 Palabras) - (<i>Primera entrega</i>)	7

3.4	¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto? (Max 100 Palabras) - (<i>Primera entrega</i>)	7
4	Diseño de integración y Automatización de Datos para IA (Diagrama) (<i>Primera entrega</i>)	8
5	Proximos pasos (<i>Tercera entrega</i>)	10
6	Lecciones aprendidas (<i>Tercera entrega</i>)	11
7	Bibliografía	12

1 Introducción (Max 250 Palabras) - (Primera entrega)

Uno de los principales efectos de la accidentalidad vial es su alto costo económico. Según datos de la Federación de Aseguradores Colombianos en 2016 (Federación de Aseguradores Colombianos [FASECOLDA], 2018), esto ascendió a 2,3 billones de pesos, equivalente al 0,3% del Producto Interno Bruto (PIB) de Colombia. La mayoría de este costo se destinó a la asistencia médica y pensiones para las víctimas.

Este problema se considera un asunto de salud pública, y las Naciones Unidas han instado a los gobiernos de todo el mundo a establecer metas para reducir los siniestros viales. Esto se reflejó en la Década de Acción para la Seguridad Vial 2021-2030, que busca reducir las muertes y lesiones por tráfico en un 50

Colombia también se ha comprometido con estas metas a través del Plan Nacional de Seguridad Vial 2022-2031. En Bogotá, se busca reducir las víctimas fatales y heridos graves en un 35% para 2026.

Para monitorear el progreso, Bogotá cuenta con el Observatorio de Movilidad, que proporciona datos actualizados sobre siniestros viales. Además, organizaciones no gubernamentales como Bloomberg Philanthropies trabajan en conjunto con las autoridades para mejorar la seguridad vial.

La academia también desempeña un papel importante al contribuir con investigaciones y soluciones. Este estudio se enfoca en abordar la problemática de la accidentalidad vial en Bogotá y construir una herramienta para realizar el monitoreo, identificando los factores que inciden en la ocurrencia de los siniestros viales.

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA (Max 500 Palabras) - (Primera entrega)

2.1 Título del proyecto de investigación

Proyecto de Integración y Automatización de Datos para IA para el seguimiento y monitoreo de siniestros viales en la ciudad de Bogotá D.C.

2.2 Objetivo general (Max 100 Palabras) - (Primera entrega)

Construir un tablero de control para monitorear las características de siniestros viales, que permita recomendar acciones para su seguimiento y prevención.

2.2.1 Objetivos específicos (Max 100 Palabras) - (Primera entrega)

- Realizar una integración de los datos recopilados por la Secretaría Distrital de movilidad que permita realizar análisis descriptivo.
- Determinar las características de los siniestros viales reportados por la Secretaría de Movilidad de la ciudad de Bogotá D.C. en el periodo 2018 a 2021.

- Proponer un modelo de machine learning, que sirva para determinar las causas principales asociadas al grado de severidad de los siniestros viales en Bogotá D.C.

2.3 Alcance (Max 200 Palabras) - (*Primera entrega*)

En las últimas décadas, la industria tecnológica ha experimentado un rápido avance en el desarrollo de capacidades computacionales para el procesamiento y almacenamiento de grandes volúmenes de datos. Este progreso tecnológico ha allanado el camino para la aplicación de técnicas estadísticas y de modelado de datos que son computacionalmente exigentes.

La revisión bibliográfica revela que investigadores han abordado la problemática de la accidentalidad vial desde diversos enfoques y han hecho recomendaciones significativas.

Tianyu et al. (2022) aplicaron modelos de series de tiempo y redes neuronales recurrentes para predecir la cantidad de lesionados en siniestros viales en el Noreste de China. Recomendaron el uso de redes neuronales recurrentes Tianyu et al. (2022).

Islam et al. (2022) utilizaron modelos Random Forest y Regresión Logística para predecir la severidad de siniestros viales en Al-Ahsa, Arabia Saudita, y también identificaron clusters espaciales de alta accidentalidad. Recomendaron la inclusión de variables adicionales en futuros estudios Islam et al. (2022).

Kumar y Melempat Kalapurayil (2022) exploraron modelos de clasificación multietiqueta para abordar la severidad y el tipo de colisión en siniestros viales en Texas, EE. UU. Concluyeron que estos modelos superan a los enfoques tradicionales.

Lia et al. (2022) adoptaron un enfoque innovador al combinar el análisis de redes sociales con modelos de ML para identificar y clasificar incidentes y siniestros viales en Panamá. El modelo Random Forest resultó ser el más efectivo Lia et al. (2022).

Cantillo et al. (2016) investigaron tramos críticos de accidentalidad en Cartagena, Colombia, utilizando técnicas como el Análisis de Componentes Principales y la Regresión Poisson. Identificaron relaciones empíricas entre la frecuencia de accidentes y variables como la densidad del tráfico y la geometría de las vías Cantillo et al. (2016).

Estos estudios resaltan la importancia de utilizar técnicas avanzadas de procesamiento de datos y la necesidad de considerar diversas variables para comprender y predecir la accidentalidad vial de manera más efectiva. Además, enfatizan la importancia de la integridad de los datos y la necesidad de explorar enfoques innovadores para abordar este desafío.

2.4 Pregunta de investigación (Max 100 Palabras) - (*Primera entrega*)

¿Cómo podemos detectar patrones que me permitan comprender y predecir la ocurrencia de siniestros viales y los factores que influyen en esta?

2.5 Hipotesis (Max 100 Palabras) - (*Primera entrega*)

El presente estudio tiene como objetivo abordar la problemática de la siniestralidad vial en Bogotá, analizando las características subyacentes a la gravedad de los accidentes. También busca desarrollar una herramienta actualizada y acorde a las dinámicas recientes de los siniestros viales, que contribuya a mitigar sus consecuencias, beneficiando a los diversos actores viales en la ciudad.

Se pretende responder a la pregunta de cuáles son los factores principales que influyeron en la gravedad de los siniestros viales en Bogotá, utilizando los eventos documentados en el período de 2018 a 2022.

Es importante resaltar que con los resultados del presente estudio se identificarán los factores influyentes en la siniestralidad vial en Bogotá, lo cual permitirá formular, desde la academia, los cambios en el instrumento de seguimiento a los siniestros viales con que actualmente cuenta las autoridades distritales. Lo cual busca redundar en la reducción de los siniestros viales y su gravedad.

3 Reflexiones sobre el origen de datos e información

(Max 400 Palabras) - (Primera entrega)

Con la información capturada en el sitio de ocurrencia del siniestro, la autoridad de tránsito debe reportar la información al Registro Nacional de Accidentes de Tránsito (RNAT). Para el caso de la ciudad de Bogotá, la Secretaría de Movilidad de Distrital ha puesto a disposición de los ciudadanos estos datos desde el año 2015 a través del portal de datos abierto SIMUR.

Los datos publicados realizan una síntesis de las características inherentes a la ocurrencia del siniestro vial. Dentro de estos datos se encuentran: características demográficas de los actores viales involucrados, hipótesis o causas del siniestro vial, características de los vehículos vinculados y características de las vías donde ocurrió el siniestro.

Los datos obtenidos corresponden al corte del 3 de julio de 2023 y se dividen en 5 grupos. En la Tabla 1 se muestra el detalle.

Table 1: Datos de siniestralidad publicados por Secretaría Distrital de Movilidad - SIMUR -

Tabla	Columnas	Registros
Accidente	19	464699
Actor Vial	13	1657681
Causa	10	672815
Vehículo	9	1427204
Vía	19	483952

3.0.1 Datos geográficos

La información recopilada en el formulario IPAT y puesta a disposición por la SDM, presenta una falencia en datos del entorno próximo al sitio de ocurrencia del siniestro. Por lo tanto, a partir de la información disponible en la Infraestructura de Datos Espaciales de Bogotá D.C. y su procesamiento por medio de Sistemas de Información Geográfica (SIG), se incluirán a la base de datos variables exógenas de los siniestros viales, con lo cual se espera a mejorar los resultados del presente estudio.

Las características a incluir son las siguientes:

- Cantidad de Comparendos impuestos entre 2015 - 2021 en la zona del siniestro.
- Si se cuenta con red de cicloinfraestructura a menos de 200 mts del siniestro.
- El estrato predominante en el sector del siniestro.
- Tipo de categoría de la vía donde ocurre el siniestro.
- El uso del suelo predominante en el sector donde ocurre el siniestro.

3.1 ¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (Primera entrega)

El Ministerio de transporte de Colombia por medio de la Resolución No. 11268 de 2012 adoptó el Informe Policial de Accidentes de Tránsito (IPAT), el cual es de obligatorio cumplimiento para la autoridad de tránsito.

El formato adoptado, asigna un número único para cada accidente de tránsito, recolecta información como ubicación, características del lugar, características de las vías y la identificación de los vehículos y sus conductores involucrados.

Esta información es recopilada y puesta a disposición por la Secretaría Distrital de Movilidad a través del portal de datos abierto SIMUR.

3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información? (Max 100 Palabras) - (Primera entrega)

”La Resolución No. 11268 de 2012 estipula que *‘Las autoridades de tránsito velarán por la seguridad de las personas y las cosas en la vía pública y privadas abiertas al público. Sus funciones serán de carácter regulatorio y sancionatorio y sus acciones deben ser orientadas a la prevención y la asistencia técnica y humana a los usuarios de las vías’* (Resolución No. 11268 de 2012).

Esto da lugar a la creación del Registro Único Nacional de Tránsito (RUNT) y el Registro Nacional de Accidentes de Tránsito (RNAT).

El marco técnico de dicha resolución dictamina *‘Que con el objetivo que el RNAT sea la herramienta que permita identificar claramente las hipótesis de las causas de accidentalidad, es necesario adoptar un nuevo Informe Policial de Accidentes de Tránsito, incluyendo datos que permitan un mejor análisis para la elaboración de diagnósticos y la planeación en materia de seguridad vial’* (Resolución No. 11268 de 2012).

Así pues, al ser información que permite prevenir la frecuencia en siniestros viales, y generar hipótesis, su uso es público, y por tanto su análisis no constituye una razón para inducir a un uso irregular, que pudiera violentar la integridad de cualquier entidad pública o privada” (Resolución No. 11268 de 2012).

3.3 ¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA? (Max 100 Palabras) - (Primera entrega)

Es necesario señalar que durante la ejecución del proyecto se deben sortear tanto las dificultades que trae consigo la temática, como el reto que impone el tratamiento de los datos de diversas fuentes. De igual manera, se espera encontrar dificultades en la conciliación de las cifras entre los siniestros viales documentados y las cifras oficiales que se encuentran en los observatorios, anuarios y cifras del sector asegurador (Varela, 2018).

3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto? (Max 100 Palabras) - (Primera entrega)

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

Figure 1: Procesamiento datos siniestros viales



Proceso de Automatización de Datos con Python, Oracle Cloud y Power BI

- Origen: El proceso comienza con la extracción de datos desde cinco tablas diferentes que actúan como origen de datos (Accidente, Actor Vial, Causa, Vehículo, Vía). Estas tablas pueden estar en una variedad de formatos, como bases de datos, hojas de cálculo o archivos CSV.
- Limpieza de Datos con Python: Los datos extraídos pasan por una fase de limpieza y preprocesamiento utilizando scripts en Python. En esta etapa, se pueden realizar tareas como la eliminación de datos duplicados, la imputación de valores faltantes, la estandarización de formatos y la eliminación de valores atípicos utilizando bibliotecas como Pandas y NumPy.
- Creación de Nuevos Datos con Python: Después de la limpieza de datos, se pueden generar nuevos datos o características (features) utilizando técnicas de análisis y modelado disponibles en Python. Esto podría incluir la creación de variables derivadas, la segmentación de datos o la generación de características adicionales para el análisis.
- Aplicación de un modelo de Machine Learning para variables categóricas modelado de la gravedad de siniestros viales. Entre estos modelos se destacan Random Forest, SVM y Árboles de Decisión, estas técnicas serán las utilizadas para lograr el cumplimiento de los objetivos del presente estudio.

- ETL con Oracle Cloud: Los datos preparados y enriquecidos se transfieren a Oracle Cloud para el proceso de Extracción, Transformación y Carga (ETL). En esta etapa, los datos pueden ser almacenados en una base de datos en la nube, transformados según sea necesario y preparados para su análisis posterior.
- Visualización con la herramienta de visualización Power BI: Una vez que los datos se han cargado en Oracle Cloud y están listos para su análisis, se conectan a Power BI, una herramienta de visualización que permite crear paneles interactivos, informes y gráficos que permiten a los usuarios explorar y comprender los datos de manera efectiva.

Este proceso de automatización permite tomar datos crudos de múltiples fuentes, limpiarlos, enriquecerlos, almacenarlos en una plataforma segura y presentarlos de manera visualmente atractiva para tomar decisiones informadas. La automatización de este flujo de trabajo puede aumentar la eficiencia y la calidad del análisis de datos.

5 Proximos pasos (*Tercera entrega*)

6 Lecciones aprendidas (*Tercera entrega*)

7 Bibliografía