Proyecto Final – Primera Entrega

Yachay Tolosa Bello – 202315750
Kevin Infante Hernández – 201117324
John Vicente Moreno Triviño – 202210162
MINE-4101: Ciencia de Datos Aplicada

1. Introducción

El presente documento tiene como propósito presentar la primera entrega del proyecto final de la asignatura Ciencia de Datos Aplicada, bajo la metodología ASUM-DM.

2. Definición de la problemática y entendimiento del negocio

2.1 Clientes

Este proyecto cuenta con dos clientes, el directo y el final. El cliente directo es GSD+, una empresa consultora en transporte y tecnología aplicada al transporte, que ha realizado múltiples proyectos a nivel internacional en diferentes frentes, tales como la planeación de transporte, el diseño de sistemas de transporte, el diseño de políticas públicas para el transporte, el diseño de sistemas de recaudo y control de flota para transporte público, entre otros.

Actualmente, GSD+ participa en una consultoría ad honorem para la Secretaría de Movilidad de Bogotá (SDM) para desplegar, en conjunto, un ejercicio de analítica de datos enfocado a mitigar la alta siniestralidad en la ciudad. En este orden de ideas, se considera que la SDM es el cliente final para quien este ejercicio será de suma utilidad a la hora de tomar decisiones para implementar acciones enfocadas en mejorar las condiciones de seguridad vial de las personas que se movilizan por Bogotá.

2.2 Problemática

Año a año, en Bogotá se presentan más de 500 muertes en siniestros viales y alrededor de 35.000 heridos. Estas cifras son altamente preocupantes, por no decir que escandalosas, y generan una alerta para que haya acciones inmediatas, y de medio y largo plazo, que busquen mitigar la situación.

Por otra parte, existe una amplia cantidad de datos sobre los siniestros viales en Bogotá, los cuales pueden ser aprovechados para generar políticas públicas estratégicas que permitan disminuir el número de accidentes en la ciudad.

2.3 Estrategia de negocio

La base de la política de Visión Cero de la SDM es que todos los siniestros viales ocurren por causas que son prevenibles y, por lo tanto, conocer de antemano estas causas prevenibles habilitará a los tomadores de decisión para plantear estrategias que permitan mitigar el número de siniestros viales con fallecimientos o heridos.

En ejercicios anteriores, se ha venido trabajando en descubrir algunas de las causas más relevantes de esta problemática. Con base en esto, GSD+ y la SDM han venido pensando en estrategias de política pública que permitan mitigar la alta accidentalidad. Sin embargo, no existe una estimación del impacto que puedan tener estas medidas y de su relación beneficio vs. costo de implementación.

Por lo tanto, la estrategia general alrededor de la cual girará este proyecto es la de tratar de predecir/estimar el impacto en la reducción de accidentes viales con muertos y heridos de diferentes medidas que se tienen contempladas. Con estas estimaciones, y teniendo en cuenta el costo esperado de implementar cada medida, será posible hacer una evaluación de negocio para conocer la conveniencia de cada acción propuesta y priorizar su implementación real.

2.4 Datos clave del sector

En Bogotá se realizan 20 millones de viajes cada día, y en promedio, cada día mueren 1,37 personas en uno de estos viajes, mientras que se tienen 96 heridos. A primera vista, los siniestros viales con fallecimientos o heridos podrían parecer una proporción muy pequeña respecto al total de viajes, sin embargo, de ningún modo es aceptable que exista tal tasa de mortalidad en una ciudad como Bogotá. La visión propuesta es que hayan cero fallecimientos y cero heridos en la ciudad debidos a accidentes evitables.

2.5 Objetivo del proyecto

- Evaluar la conveniencia y priorizar acciones de política pública que permitan disminuir el número de accidentes con fatalidades y/o heridos en Bogotá.

2.6 Métricas de negocio

Las métricas de negocio que se proponen son las siguientes:

- Priorización de al menos 2 acciones de política pública, que según la evaluación hecha, demuestren ser las más convenientes y de mayor impacto.
- Formular un grupo de acciones de política pública que, en su conjunto, permitan disminuir al menos un 10% de los accidentes con fatalidades y/o heridos en Bogotá, en el corto plazo (menos de 12 meses).

3. Ideación

3.1 Usuarios del producto, sus procesos y necesidades

De manera general, el producto tendrá dos tipos de usuarios. Desde GSD+, se tendrá un usuario consultor en transporte, que estará interesado en disponer de una herramienta que le permita estimar el impacto de diversos escenarios hipotéticos de políticas públicas sobre los niveles de siniestralidad vial en la ciudad.

Por otra parte, desde la SDM, existirá un usuario que será funcionario de la Dirección de Inteligencia para la Movilidad y su interés será el disponer de la

información generada por el usuario anterior, para tomar decisiones sobre las acciones estratégicas y tácticas que más le convengan a la ciudad.

3.2 Requerimientos, componentes y mockup

Los requerimientos funcionales de alto nivel del producto se definen a continuación:

- Permitir realizar prediciones/estimaciones basadas en datos, sobre el impacto en los niveles de siniestralidad vial de Bogotá, de medidas tales como:
  - Restricción de la circulación de motocicletas en Bogotá.
  - Imposición de límites de velocidad más estrictos en algunas o todas las zonas y/o avenidas de Bogotá.
  - Instalación de cámaras salvavidas en puntos críticos de la ciudad.
  - Otras medidas podrán ser propuestas durante este ejercicio.

  Nota: Las medidas aquí mencionadas solo son una muestra de las posibles acciones a tomar.
- La predicción/estimación será realizada para periodos anuales. Es decir, se estimará el impacto año a año de los paquetes de medidas planteados, en un horizonte de 5 años. Sin embargo, si por temas de resolución del algoritmo se requiere, las predicciones/estimaciones podrán ser realizadas por día o por mes, para luego agregarse a nivel anual.

Entre tanto, la solución constará de los siguientes componentes:

- Módulo para la ingestión de datos históricos.
- Módulo para la selección de escenarios y medidas de política pública.
- Modelo de Machine Learning para predicción/estimación de accidentes según datos históricos y medidas de política públicas simuladas.
- Aplicación web (se considerará implementar API REST).
- Dashboard web que permita a los usuarios interactuar con la herramienta, siguiendo los lineamientos y buenas prácticas de la analítica visual:

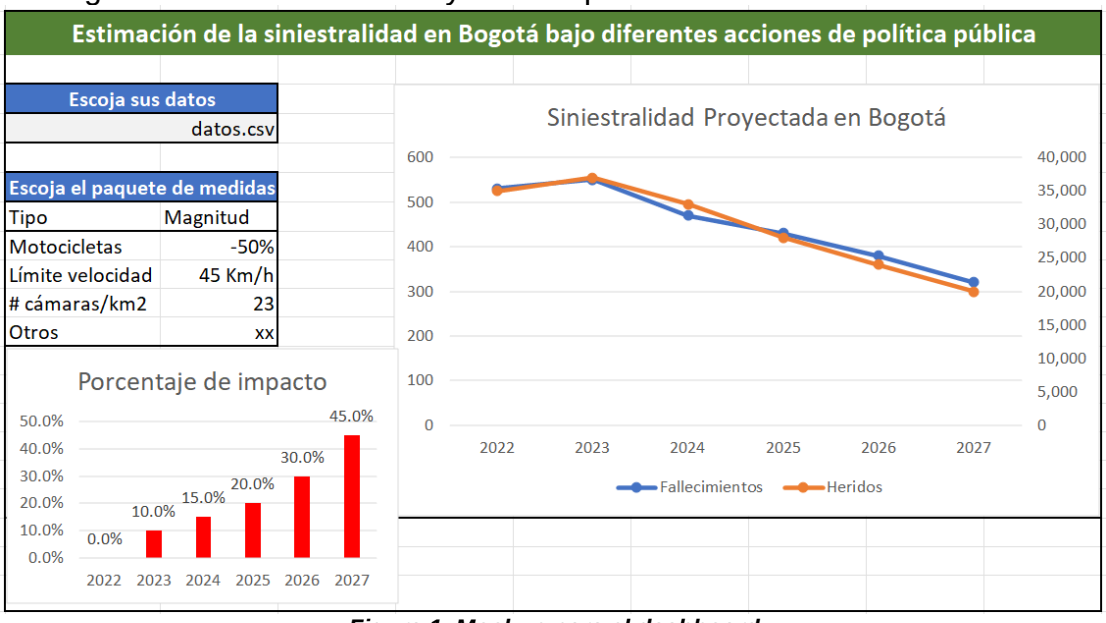

Figura 1. Mockup para el dashboard

4. Responsible

Los datos utilizados son de acceso abierto y ya se encuentran anonimizados. De este modo, se puede afirmar que no existe ninguna restricción para su uso.

Así mismo, en cuanto a las técnicas de inteligencia artificial que se puedan usar tampoco existen restricciones a priori. Sin embargo, hay que mencionar que como consideración ética será necesario que los implementadores de este proyecto siempre estén muy conscientes de que las estimaciones generadas por el modelo se verán reflejadas en vidas humanas salvadas (o no salvadas) en la vida real.

5. Enfoque analítico

5.1 Preguntas de negocio

- ¿Cuál será el impacto (a 12 meses) en la reducción de siniestros viales con fallecimientos y/o heridos de distintas políticas públicas, si estas se implementaran de manera individual?
- Luego de priorizar un paquete de políticas públicas ¿Cuál será su impacto (en 12 meses) en la reducción de accidentes con fallecimientos y/o heridos si las medidas se implementan de manera simultánea?

5.2 Técnicas de modelamiento de datos

A priori, se considera que la técnica de Machine Learning que podría ser usada en este ejercicio es la **Regresión Lineal Múltiple**. La idea es tratar de estimar el número de accidentes viales graves (con muertos y/o heridos) con base en variables tales como los tipos de vehículos involucrados, la hora del día, la época del año, las condiciones de la vía, las condiciones de luz, la presencia de cámaras salvavidas, entre otras.

Como es de notarse, existen variables numéricas y categóricas que deberán procesarse para poder usarse en un modelo de regresión. Así mismo, será necesario probar con diferentes técnicas de **regularización** para el modelo, tales como **Ridge o Lasso**. Dado el caso, también podría ser útil trabajar con un **modelo de regresión de soporte vectorial (SVR)** y enfrentarlo con la regresión lineal múltiple para observar cuál se comporta mejor. Por último, será importante considerar técnicas de balanceo de datos durante la preparación de estos.

5.3 Métricas de calidad para el modelamiento

Para evaluar la calidad del modelo entrenado, la principal métrica utilizada será el **MSE (Error Cuadrático Medio)** por tratarse de una regresión. También será muy importante calcular el **coeficiente de autodeterminación y sesgo** para comprobar que el modelo desarrollado no se encuentre sesgado por los datos de entrada.

6. Recolección de datos

Los datos que se usarán corresponden al historial de los siniestros viales en Bogotá entre 2015 y 2022, con detalles de sus causas y contexto, provisto por la SDM. Este

dataset, junto con otros que podrían ser de utilidad, es de libre acceso en el portal de datos abiertos de la ciudad: https://datosabiertos.bogota.gov.co/dataset/siniestros-viales-consolidados-bogota-d-c

Estos datos se encuentran distribuidos en 7 conjuntos que contienen diversa información de interés, por lo que fue necesario unirlos en un solo dataset que contiene 745.256 registros y 28 columnas. Entre la información más destacada se puede observar: la fecha y hora del siniestro, la modalidad (choque o atropellamiento), la descripción por parte del oficial que atendió el incidente, la causa reconocida, estado de la vía, iluminación, número de carriles y calzadas de la vía, materiales de la vía, tipos de vehículos envueltos en el accidente, gravedad (solo daños, con heridos o fallecidos), entre otros.

En general, es posible decir que los datos presentan una buena consistencia y la información de las columnas que se mantuvieron es totalmente relevante y suficiente para implementar los modelos propuestos.

7. Entendimiento de los datos

El entendimiento de datos se puede encontrar en el notebook anexo a esta entrega.

8. Roles de los implementadores

Se describe brevemente el rol que tendrá cada uno de los implementadores:

- **Yachay Tolosa – Científico de datos:** encargado de experimentar con modelos de ML, colocarlos a prueba y entrenar con los datos de entrada.
- **Kevin Infante – Ingeniero de datos:** encargado de preparar la infraestructura para que los modelos trabajen de manera adecuada. Además, a su cargo también estará la aplicación web.
- **John Moreno – Ingeniero de negocio:** encargado de entender y traducir las necesidades del cliente y su negocio, definir los requerimientos de la herramienta y extraer los insights de valor para los usuarios del producto.

Hay que decir que estos roles son tentativos y no necesariamente mandatorios.

9. Conclusiones, insights y próximos pasos

El entendimiento de datos ha permitido ver que se cuenta con datos de una calidad y consistencia adecuadas para llevar a cabo el proyecto. Sin embargo, aún no se descarta que haya que recurrir a alguno de los datasets complementarios que se encuentran en el repositorio de datos abiertos de la ciudad. Por otra parte, se generan algunos primeros insights sobre las principales causas de los accidentes:

- **Incremento de la proporción de motocicletas en la ciudad:** el número de motocicletas en circulación aumentó 4 veces entre 2007 y 2022.

- **Hora del día:** los accidentes suelen ocurrir más frecuentemente en la madrugada, de “03:00 a 05:00”. También es importante la porción entre las horas pico: “06:00 a 08:00” y “17:00 a 19:00”.
- **Exceso de velocidad:** es común encontrar que una gran parte de los accidentes graves ocurren porque al menos uno de los vehículos involucrados iba a más de 50 km/h.
- **Prácticas inseguras:** adelantamiento indebido, invasión de carril, giros bruscos y saltarse los semáforos en rojo están dentro de las malas prácticas que más generan accidentes.

Finalmente, como próximos pasos se contempla continuar con la metodología ASUM-DM, específicamente con la preparación final de los datos, la construcción y validación de varios modelos de ML, y selección y evaluación del mejor modelo.