Proyecto Final – Tercera Entrega – Ciencia de Datos Aplicada

Yachay Tolosa Bello: 202315750
Kevin Infante Hernández: 201117324
John Vicente Moreno Triviño – 202210162

Entrega 1:

1. Introducción

El presente documento tiene como propósito presentar el 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, uno directo y uno final. El cliente directo es GSD+, una empresa consultora en transporte y tecnología aplicada al transporte, que ha realizado múltiples proyectos a nivel internacional. 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 implementar acciones enfocadas en mejorar las condiciones de seguridad vial en 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, para mitigar la situación.

Por otra parte, a pesar de los esfuerzos realizados por las autoridades, las cifras de accidentes graves vienen creciendo año a año, y se espera que en 2023 este problema crezca aún más. Por lo tanto, es necesario que las políticas relativas a la seguridad vial deriven en acciones focalizadas que permitan ver resultados.

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, entender 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. Esto ha permitido plantear tres estrategias de política pública de alto nivel:

- Controlar más rigorosamente la circulación de motocicletas, estableciendo controles en campo destinados a minimizar las malas prácticas de motociclistas o imponiendo zonas y horarios de restricción.
- Llevar a cero el número de personas que deciden conducir bajo los efectos del alcohol, mediante el establecimiento de controles en lugares y momentos apropiados y el fortalecimiento de las leyes aplicables.
- Minimizar los vehículos que circulan a alta velocidad, en momentos y lugares concretos, mediante la optimización de la ubicación de las Cámaras Salvavidas con las que cuenta la ciudad.

Sin embargo, no existe aún una metodología que permita priorizar la aplicación de una u otra estrategia en momentos y lugares determinados para así maximizar su impacto. La idea de fondo es que tener tal metodología habilitaría a las autoridades para desplegar acciones concretas, en lugares y momentos específicos en los que es probable que se repitan los accidentes.

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 haya cero fallecimientos y cero heridos en la ciudad debido a accidentes evitables.

2.5 Objetivo del proyecto

- Determinar los lugares y momentos prioritarios para llevar a cabo intervenciones enfocadas en reducir la siniestralidad y orientar el tipo de acciones a desplegar allí.

2.6 Métricas de negocio

Para medir el éxito, a nivel de negocio, de las estrategias de política pública que se prioricen como resultado de este ejercicio, se propone el siguiente KPI. Para una política “i”, el indicador de reducción de siniestralidad se definirá como:

$$Reducc\_Siniestros(i)\ \% = [Siniestros(i, 2023\text{-}I) - Siniestros(i, 2024\text{-}I)] / Siniestros(i, 2023\text{-}I)$$

Donde:

- Siniestros(i, año-I) representa el número de accidentes con muertos o heridos asociados a la causa que busca mitigarse con la política “i”, durante el primer trimestre del año en cuestión.

Se considerará que el resultado de este ejercicio es exitoso si Reducc_Siniestros(i) > 20%, para al menos dos de las políticas recomendadas. Se ha escogido el primer trimestre de 2023 y el primero de 2024 para la definición del KPI, puesto que se propondrá implementar un primer piloto de uso del producto de datos al comenzar el siguiente año.

3. Ideación

3.1. Usuarios, sus procesos y necesidades

Los principales usuarios del producto serán funcionarios de la SDM encargados de priorizar intervenciones de diferente índole en la ciudad (controles de alcohol, de velocidad, campañas pedagógicas, señalización, entre otras). No obstante, estas personas se encuentran con el problema del gran volumen de información existente sobre siniestros viales, lo que les dificulta decidir dónde, cómo y cuándo intervenir.

De esta manera, se ha identificado que estos funcionarios requieren de una herramienta que les permita identificar las zonas con mayor probabilidad de accidentes viales graves en diferentes franjas horarias de cada día del año, para así poder desplegar acciones concretas enfocadas en mitigar una o varias causas relevantes.

3.2 Requerimientos, componentes y mockup

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

- Estimar la posibilidad de ocurrencia de un accidente grave (con muertos o heridos) en cuadrantes de tamaño variable, según la hora y el día del año en Bogotá.
- Permitir una visualización basada en una interfaz cartográfica que permita a los usuarios un entendimiento intuitivo de las zonas con mayor riesgo, según hora y día.

Entre tanto, la solución constará de los siguientes componentes:

- Modelo de Machine Learning para la estimación de la probabilidad de ocurrencia de siniestros viales graves en cuadrantes de tamaño variable, según la hora y el día del año.
- Aplicación web (se considerará implementar una API REST).
- Dashboard cartográfico que permita identificar intuitivamente las zonas con mayor probabilidad de accidentes, dados una hora y un día del año. La probabilidad de accidente en una zona será representada por la intensidad del color asignado a un cuadrante.

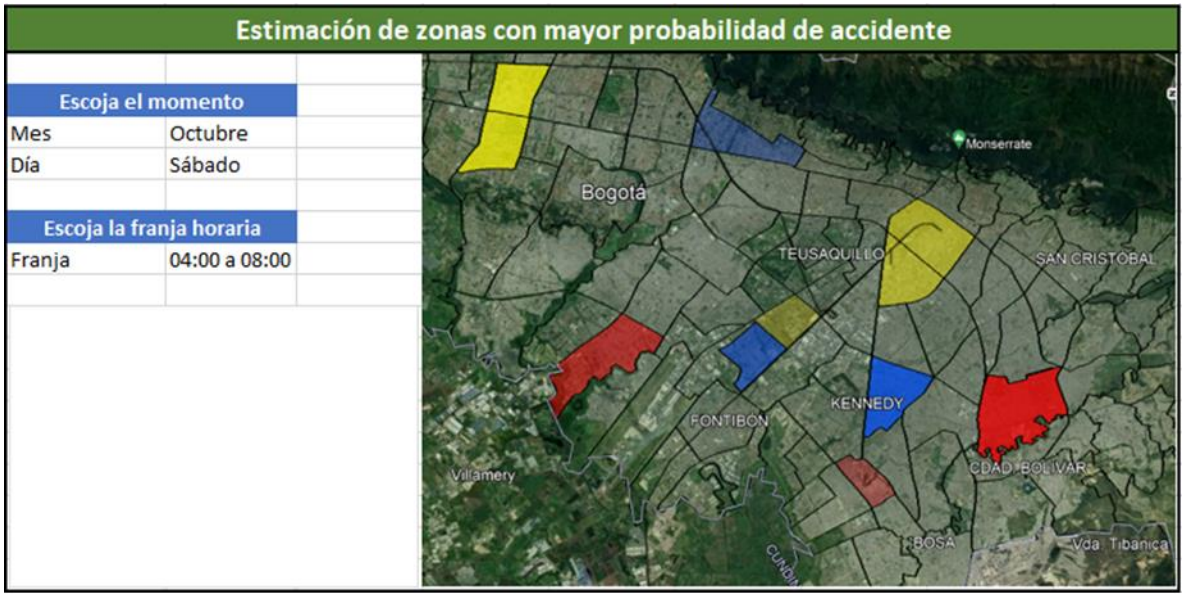

Figura 1. Mockup para el dashboard

4. Declaración de responsabilidad

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 podrán ver reflejadas en vidas humanas salvadas (o no salvadas) en la vida real.

5. Enfoque analítico

5.1 Pregunta de negocio

- ¿Dado un momento del año (mes, día y hora) cuáles son las zonas en las que más probablemente se presenten accidentes y en las que se deba concentrar la ejecución de acciones concretas?

5.2 Técnicas de modelamiento de datos

Se propone realizar un modelo de clasificación que permita estimar la probabilidad de ocurrencia de accidentes en diferentes zonas. Para ello, la ciudad se dividirá en cuadrantes de tamaño variable y se buscará que el modelo determine si un cuadrante es propenso a la ocurrencia de siniestros viales, dados una hora y día específico. La idea es que, con base al histórico de accidentes, y según día del año y hora, el modelo de clasificación pueda decidir si en un cuadrante se deben esperar accidentes o no en un momento concreto.

Se considera que algunos modelos como Regresión Logística, Random Forest y XGBoost pueden ser adecuados para este problema, en tanto: 1) permiten estimar

la probabilidad y 2) escalan relativamente bien al volumen de datos que se maneja en el proyecto.

5.3 Métricas de calidad para el modelamiento

La métrica elegida es el **F1-score.** Por un lado, es importante clasificar con precisión aquellas zonas y momentos (día del año y hora) en los que es más probable que suceda un accidente. Así mismo, para el negocio es importante que aquellas zonas clasificadas como de alto riesgo, en realidad lo sean para no desperdiciar recursos en cuadrantes en los que la probabilidad de accidentes realmente era baja.

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 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

7. Entendimiento de los datos

El entendimiento de datos se encuentra en el notebook de la primera entrega.

8. Roles de los implementadores

Se describe el rol tentativo 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 – Inteligencia 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.

9. Conclusiones e insights del entendimiento de datos

El entendimiento de datos ha permitido ver que se cuenta con datos de una calidad y consistencia adecuadas para llevar a cabo el proyecto. 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.

Entrega 2:

10. Preparación de datos

La primera fase de preparación de los datos ha implicado la eliminación de las columnas con gran cantidad de datos nulos (datos geométricos de la vía, los carriles y la calzada) que no eran relevantes para el negocio. Adicionalmente, se ha eliminado el 6.5% de los registros en los que faltaba la coordenada “Y”, la fecha y la causa del accidente. Estas tres variables resultan cruciales para el modelo, por lo que fue necesario eliminar estos registros, al no existir forma de imputarlos coherentemente ni de complementar los datos con información externa. También se ha eliminado el 21.4% de datos duplicados, lo cual se puede hacer sin perder información relevante para el negocio.

En la segunda fase, se ha procedido a dividir la ciudad de Bogotá en cuadrantes de tamaño variable, de modo que todos los accidentes registrados en el dataset puedan ser agrupados en alguno de estos cuadrantes. Como aclaración, se ha tenido en cuenta solo el área urbana de Bogotá sin Sumapaz ni Soacha. Por otra parte, es en este punto donde ocurre una de las transformaciones más importantes. El dataset, que venía siendo una lista de accidentes, es transformado en una lista donde cada registro representa un cuadrante, en una franja horaria y fecha determinadas, indicando el número de accidentes que hubo. Un ejemplo típico de uno de estos nuevos registros sería el que indica cuántos accidentes se dieron en el cuadrante #15, el día 31 de octubre de 2021, entre las 08:00 y 12:00 horas.

Hasta aquí, se tiene un dataset que muestra cuantos accidentes hubo en un cuadrante, dados una fecha y franja horaria (clase positiva). Sin embargo, si no hubo accidentes, no existen registros de esto. Por lo tanto, en un tercer nivel del proceso se complementa la lista con todos aquellos cuadrantes de la ciudad, dada una fecha y franja horaria, en los que no se registraron accidentes, siendo esta la clase negativa para el algoritmo de clasificación.

Adicionalmente en un cuarto nivel, y para enriquecer los datos, a cada registro se le ha agregado información sobre si era día festivo y el acumulado de accidentes en los últimos 15 días previos en la misma franja horaria. También se aplicó una transformación sinusoidal a los datos periódicos (hora, día y semana), de modo que el modelo pueda tener una noción del comportamiento cíclico de estas variables. Además, se aplicó una estandarización z-score a los datos para evitar sesgos. El siguiente diagrama de bloques resume el proceso de preparación de datos:

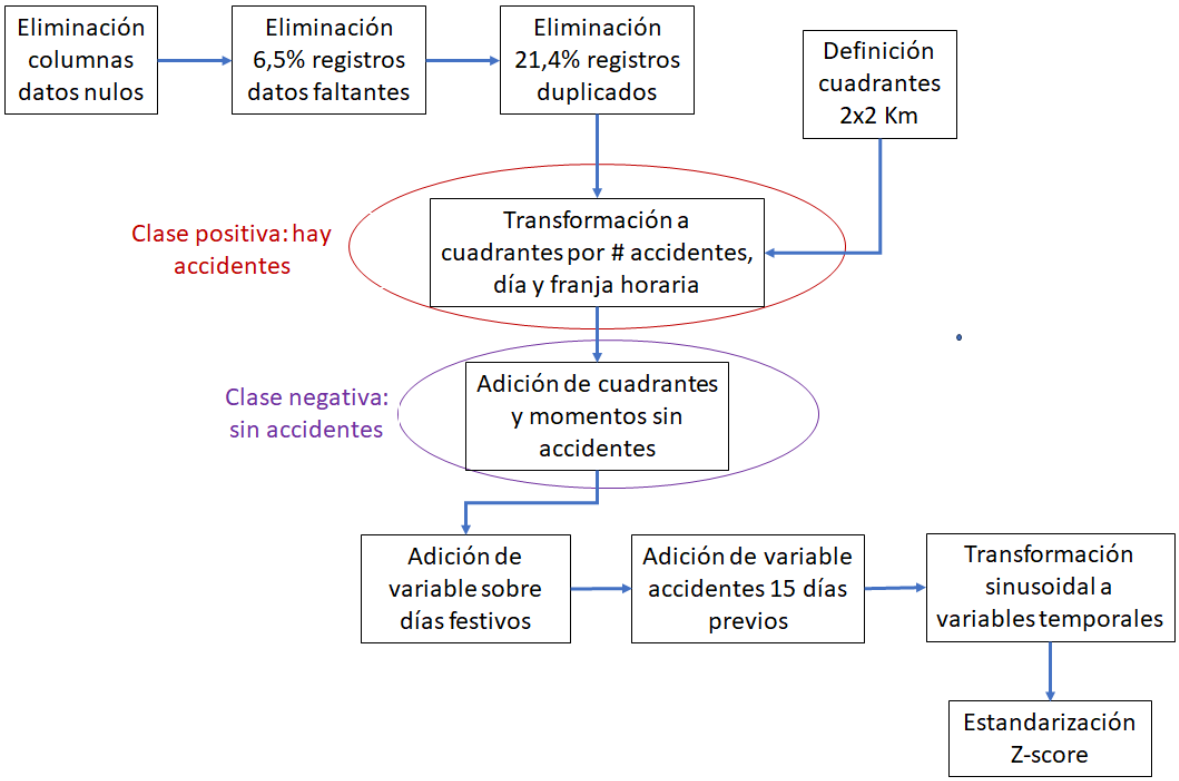

Figura 2. Proceso de preparación de datos

11. Estrategia de prueba y selección del modelo

Se han separado los datos de entrenamiento y prueba en una proporción 70-30. Teniendo en cuenta que se dispone de datos sobre los accidentes entre 2015 y 2022, la idea es entrenar el modelo con el 70% de los primeros datos (ordenados por fecha), para testearlo con el 30% de los datos más recientes, correspondientes estos últimos a 2022.

Aquí es necesario recordar que el objetivo es construir un modelo que permita estimar la probabilidad de ocurrencia de accidentes en un cuadrante de la ciudad, dados un día y una hora determinados. Para esto, se probarán 3 modelos de clasificación (cuya salida pueda leerse en términos de la probabilidad de que haya accidente o no) y la métrica para su selección será el F1-Score.

12. Construcción del modelo

Se han implementado tres modelos de clasificación para determinar la probabilidad de que ocurra un accidente grave en un cuadrante de la ciudad en un momento determinado. En primer lugar, se inicia con un modelo de regresión logística como modelo base. Desafortunadamente, con este modelo se obtuvo un F1-score promedio ponderado de 0.42, por lo cual se optó por incrementar el tamaño de los cuadrantes de 2x2 Km a 4x4 Km, lo que, junto con la optimización de hiperparámetros, ha permitido aumentar esta métrica a 0.68.

Posteriormente, se ha probado con los modelos XGBoost y Random Forest (ver notebook adjunto en esta entrega). Para facilitar la comparación entre modelos, se ha usado la herramienta “Classification experiment” (de la librería “Pycaret”).

Entre tanto, es importante notar que se utilizó un pipeline con todos los pasos de procesamiento discutidos previamente. Adicionalmente, se han generado instancias sintéticas de la clase minoritaria (es decir, la clase positiva cuando y donde sí hay accidentes) usando la técnica SMOTE (Synthetic Minority Oversampling Technique). Además, se ha optimizado la búsqueda de hiperparámetros para cada modelo usando validación cruzada con “Grid Search”, tal como se observa en el notebook.

13. Evaluación del modelo

A continuación, se muestra la matriz de confusión del modelo base de regresión logística, obtenida con los datos de test.

	Precisión	Cobertura	F1-score
No accidente	0.56	0.98	0.72
Accidente	0.61	0.03	0.06
Accuracy			0.56
Promedio	0.59	0.51	0.39
Promedio ponderado	0.58	0.56	0.42

Tabla 1. Matriz de confusión del modelo base (regresión logística)

Dado que el F1-score, que es la principal métrica en este caso, no superó ni siquiera la probabilidad de un clasificador aleatorio, se tomó la decisión de variar los tamaños de cuadrícula y franja de tiempo para mejorar el balance de clases y, así mismo, mejorar el F1-score. Adicionalmente, se realizó un proceso de optimización de hiperparámetros con “Grid search cross validation”, lo cual permitió mejorar el promedio ponderado del F1-score desde 0.42 a 0.68, una mejora del 50% aproximadamente.

Partiendo de estas mejoras, se procedió a construir los modelos correspondientes a los otros dos algoritmos con “Classification experiment”: XGBoost y Random Forest. El algoritmo que demostró mejor desempeño fue XGBoost, para el cual se muestra su matriz de confusión obtenida con los datos de test:

	Precisión	Cobertura	F1-score
No accidente	0.86	0.88	0.87
Accidente	0.66	0.60	0.62
Accuracy			0.81
Promedio	0.76	0.74	0.75
Promedio ponderado	0.80	0.81	0.80

Tabla 2. Matriz de confusión del mejor modelo XGBoost

Puede observarse que mediante la búsqueda de hiperparámetros para el modelo XGBoost se logró mejorar el promedio ponderado del F1-score de 0.42 (modelo base) a 0.80 con XGBoost. Además, el F1-score de la clase positiva (cuadrantes con accidentes) ha mejorado desde 0.06 a un aceptable 0.62.

14. Conclusiones sobre la construcción y evaluación del modelo

Sobre la suficiencia del mejor modelo obtenido:
El modelo desarrollado hasta este momento ya es útil para la predicción de sí habrá accidente o no en un determinado cuadrante, en un día y hora determinados. Sin embargo, durante el tercer sprint se trabajará en mejorar las métricas obtenidas, para obtener una herramienta con mayor precisión y exhaustividad, adecuada para los objetivos de negocio.

Dificultades y estrategias de mitigación:
El desbalanceo de clases es la mayor dificultad que se ha tenido que abordar durante el desarrollo de este proyecto. Este desbalanceo se da porque en los diferentes cuadrantes, en cada intervalo de tiempo de interés, lo más probable es que no haya accidentes, por lo que la clase positiva (accidentes en un cuadrante durante una franja de tiempo) es mucho más escasa dentro de la muestra.

Hasta el momento, ampliar las franjas de tiempo y el tamaño de los cuadrantes y realizar técnicas de sobremuestreo ha permitido obtener una mejora significativa en los resultados. Sin embargo, tener cuadrantes demasiado grandes (4x4 Km) dificulta un poco el cumplimiento de los objetivos de negocio, en el sentido de que la SDM necesita ser capaz de desplegar acciones concretas en áreas territoriales homogéneas y relativamente pequeñas. Por lo anterior, es necesario trabajar en mejorar las métricas del modelo de clasificación, incluso para cuadrantes de menor tamaño.

Condiciones y mejoras en los datos:
Finalmente, se considera que hasta el momento los datos con los que se cuenta son suficientes para obtener resultados aceptables de cara a la identificación de los cuadrantes de mayor riesgo de accidentalidad, por lo cual las estrategias de mejora aquí estarán orientadas principalmente en el preprocesamiento de estos datos y en la búsqueda de un modelo que permita obtener mejores métricas de desempeño en cuadrantes de 2x2 Km.

Entrega 3:

15. Resultados de la tercera iteración de modelación

En esta tercera entrega el esfuerzo se ha enfocado en mejorar las métricas del modelo para cuadrantes de diferente tamaño. Para esto, se ha realizado un trabajo importante de ingeniería de características en el cual se ha buscado incluir diferentes resúmenes de la información histórica de los accidentes tales como: la media móvil de accidentes en la misma hora y lugar, durante todo el día para múltiples ventanas de tiempo. Para seleccionar la mejor combinación se tomaron las características que tienen mayor importancia en promedio en un modelo Random Forest.

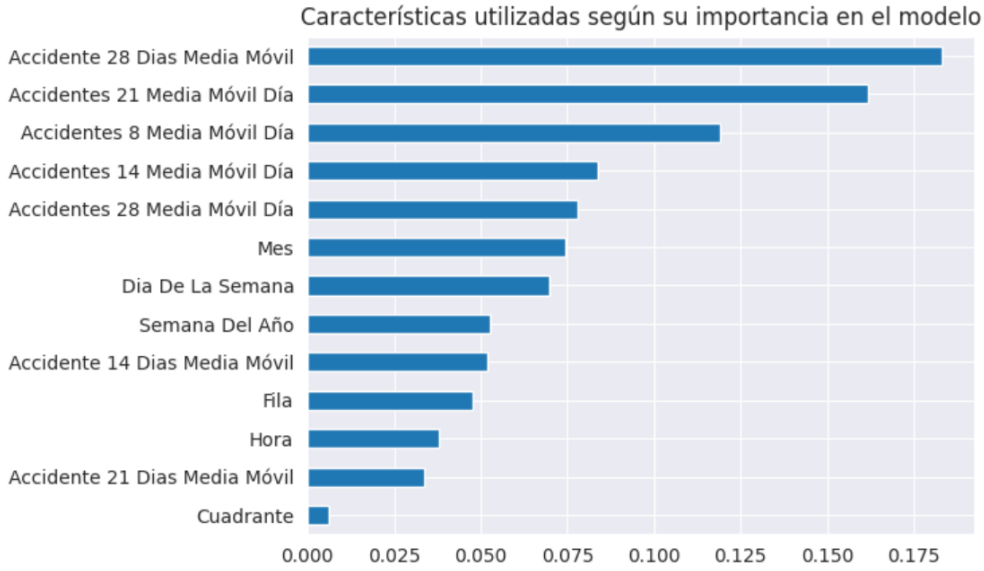

Figura 3. Características por relevancia según el modelo Random Forest

Igualmente, se optó por hacer una búsqueda más amplia de hiperparámetros. Adicionalmente, se han vuelto a entrenar instancias del modelo XGBoost (que nuevamente genero mejores resultados) utilizando 15 tamaños diferentes de cuadrantes. A modo de referencia, el tamaño de los cuadrantes para diferentes grillas se ilustra a continuación (ver también imagen en la siguiente página):

- 2 x 4: equivale a cuadrantes de 10 Km/lado.
- 5 x 10: cuadrantes de 4 Km/lado.
- 8 x 16: cuadrantes de 2.5 Km/lado.
- 15 x 30: cuadrantes de 1.33 Km/lado.

Al presentar esta idea de experimentación al cliente, éste solicitó una herramienta con la cual se pudiera escoger entre distintos tamaños de grilla. Es decir, al ver la imagen de la siguiente página, se dieron cuenta de que sería de gran interés para el negocio poder realizar predicciones en zonas de distintos tamaños: para algunas de sus actividades estratégicas es más útil tener un alcance geográfico amplio por

cuadrante, mientras que en otros casos (tareas tácticas) es mejor tener zonas más pequeñas.

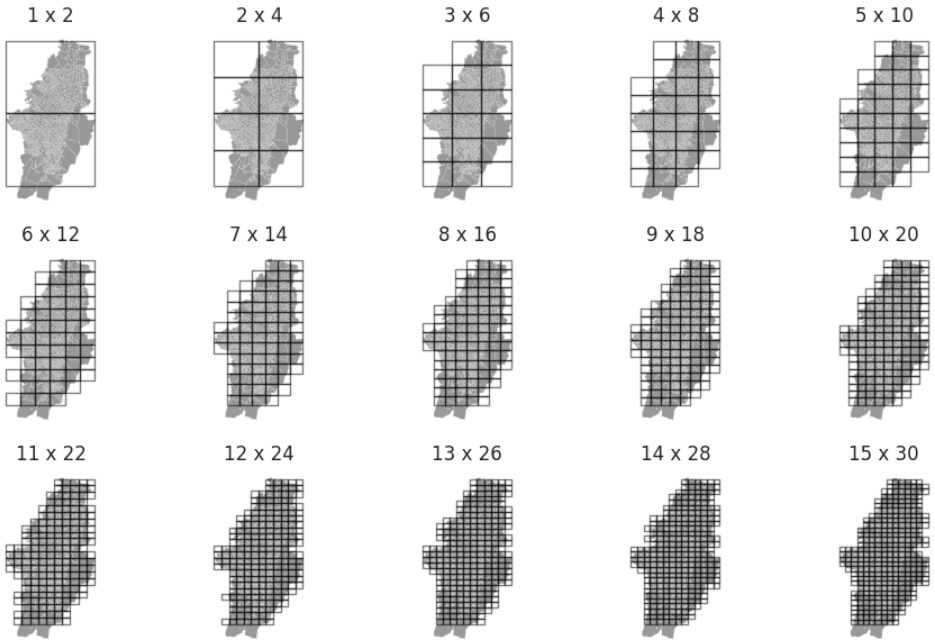

Figura 4. Tamaños de cuadrante para el entrenamiento de 15 modelos

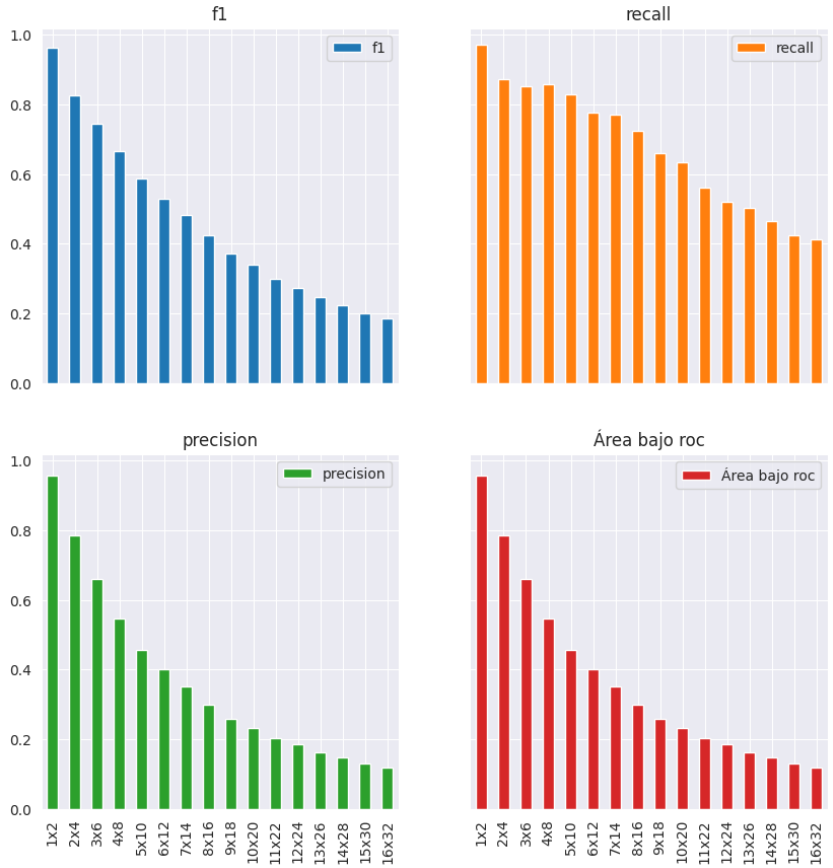

Figura 5. Resultados de las métricas de evaluación para diversos tamaños de cuadrante

Teniendo en cuenta esto, y tal como ya se mencionó, se entrenaron 15 instancias del modelo para cada uno de los tamaños de cuadrante propuestos. En general, y tal como es de esperarse, entre menor es el tamaño del cuadrante, menor es el desempeño del modelo medido a través de las diferentes métricas.

En la gráfica anterior, se puede observar la pérdida de rendimiento que implica aumentar el detalle de la grilla, medido a través de las diferentes métricas, en particular el F1-Score. Este hecho pone de manifiesto una disyuntiva entre el tamaño del cuadrante y la calidad de las métricas. Esto significa que se debe hacer un trade-off entre una pérdida de rendimiento del modelo y el beneficio de una mayor precisión geográfica en la predicción. A modo de ejemplo, para la grilla de 8x16 se predijeron accidentes en el 74% de los cuadrantes donde efectivamente ocurrieron, mientras que este porcentaje se reduce a 44% con la grilla de 16x32.

Finalmente, la siguiente imagen permite ver un compilatorio de ejemplos de resultados obtenidos para una hora y un día concretos, con diferentes tamaños de cuadrante:

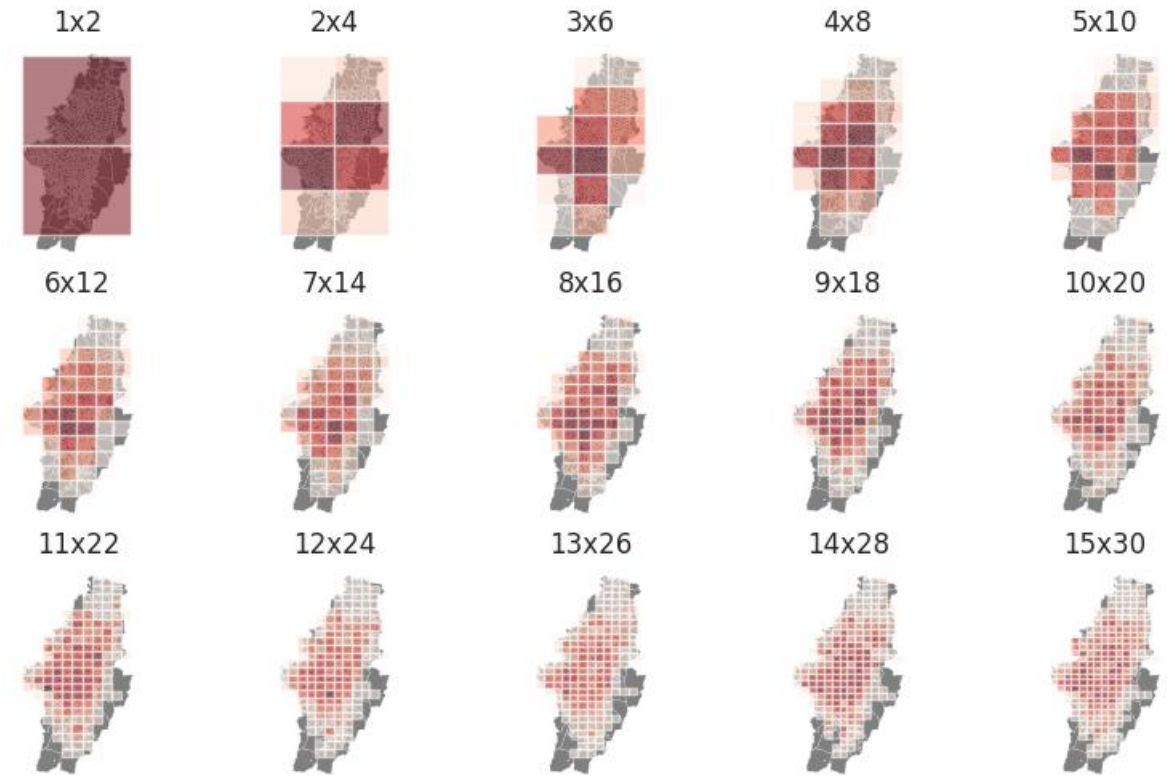

Figura 6. Ejemplos de las predicciones de los 15 modelos para 15 tamaños de cuadrantes

Esta imagen permite dar un primer vistazo a la nueva funcionalidad de poder variar el tamaño de los cuadrantes. Además, entre más oscuro sea el tono de un cuadrante, significa que es más probable que se presente un accidente allí. Como datos relevantes para el negocio a nivel estratégico, es posible ver que, en general,

los cuadrantes en los que más se predicen accidentes futuros están concentrados en las localidades de Bosa, Kennedy, Puente Aranda y Barrios Unidos. Adicionalmente, en las grillas con mayor resolución geográfica es posible identificar que las zonas más rojas coinciden con intersecciones entre avenidas importantes, intersecciones en las cuales se predice que ocurrirán más accidentes.

16. Especificación y construcción del producto de datos

El producto de datos consiste en un modelo basado en Machine Learning, el cual, a través de una interfaz web alimenta un dashboard, que a su vez le permite al usuario de la SDM priorizar la ejecución de actividades estratégicas para la reducción de la accidentalidad en Bogotá, en aquellas zonas geográficas (cuadrantes) con una mayor probabilidad de accidentes. El tamaño de la zona geográfica es parametrizable, dado que dependiendo de la actividad a ejecutar se requiere, o bien de un mayor alcance geográfico (Planeación estratégica), o de una mayor precisión a nivel territorial (Acciones tácticas).

En el dashboard el usuario debe seleccionar la fecha y hora para las cuales desea ver la predicción de accidentalidad por cuadrante, además de elegir un tamaño de los cuadrantes mismos (i.e. el número de divisiones de la grilla). El prototipo construido se encuentra desplegado en la nube, y se puede consultar en el siguiente enlace: https://dashboardmovilidad.streamlit.app/.

Cabe aclarar que para la construcción del prototipo se cuentan con datos de accidentalidad hasta el 2 de febrero de 2023, los cuales fueron proporcionados por la SDM. Además, dado que el prototipo permite la predicción de accidentes de hasta una semana en el futuro, las fechas habilitadas para la consulta de predicciones son del 3 al 9 de febrero de 2023. Una vez se despliegue el prototipo en la infraestructura de la SDM y se tenga conexión con la base de datos de accidentes actualizada, se podrán predecir los siniestros futuros con una anticipación de hasta una semana.

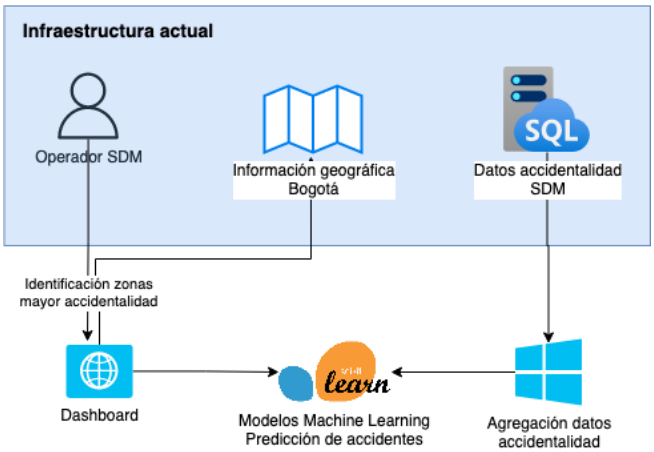

Figura 7. Diagrama arquitectura de solución

Para efectos de la primera demostración, el prototipo se desplegó en Streamlit, un proveedor de nube público. Sin embargo, el despliegue final se deberá realizar en

la nube de la SDM. En la figura anterior, se muestra el diagrama que muestra la arquitectura de la solución y su integración con la infraestructura y recursos de la entidad.
Por una parte, la SDM dispone de una base de datos con la información de accidentalidad en la ciudad, información geográfica de la ciudad, y personal encargado de la toma de decisiones con respecto a las estrategias a implementar para la reducción de accidentalidad (operador). La solución se integra con la arquitectura existente al agregar los datos de accidentalidad en una instancia Azure, calcular las predicciones de accidentalidad por cuadrante, día y hora usando los modelos construidos para este propósito, y mostrar el producto de datos en una instancia web de Azure, que le permite al operador potenciar su toma de decisiones con respecto a las políticas a implementar para la reducción de accidentes por cuadrante. Se especifican instancias Azure en el diagrama dado que es el proveedor de nube que usa la SDM.

17. Retroalimentación por parte de la organización

A continuación, se muestra una tabla con la bitácora de retroalimentación:

Bitácora Retroalimentación SDM			
Fase y Fecha	**Representantes SDM**	**Actividades**	**Retroalimentación y acuerdos**
Fase 1 - 08 sept 2023	Asesores de la Dirección de Inteligencia para la Movilidad	Kick off, conceptualización y definición de alcances	-Conceptualización de la idea de negocio. -Definición de requerimientos de la SDM. -Entrega de datos bases por parte de la SDM. -Acuerdos sobre el alcance de la oportunidad analítica.
Fase 2 - 22 sept 2023	Asesores de la Dirección de Inteligencia para la Movilidad	Presentación de la propuesta concreta relativa a la oportunidad análitica y aprobación de esta propuesta por parte de la SDM	-Ideación y acuerdo sobre el producto de datos a entregar. -Acuerdo sobre el enfoque analítico y los posibles modelos de Machine Learning para utilizar.
Fase 3 - 20 oct 2023	Asesores de la Dirección de Inteligencia para la Movilidad	Reunión de seguimiento mensual	-Reunión de seguimiento para evaluar el avance de la oportunidad analítica y obtener retroalimentación desde el negocio, que permitiese alinear el proceso de modelación con los requerimientos de la SDM.
Fase 4 - 10 nov 2023	Asesores de la Dirección de Inteligencia para la Movilidad	Presentación de la primera versión del modelo a la SDM, generación de insights para el negocio y aportes por parte de la entidad	-Con la primera versión del modelo, la SDM ha solicitado que el producto de datos tenga la posibilidad de parametrizar el tamaño de los cuadrantes. El interés de la entidad es tener tanto grillas grandes como pequeñas. -La SDM considera relevante que se puedan mejorar las métricas de predicción para cuadrantes de 2x2 Km o menores.
Fase 5 - 28 nov 2023	Asesores de la Dirección de Inteligencia para la Movilidad	Pressentación de la versión final del producto de datos 2023 ante la SDM. Aportes de la entidad y retroalimentación.	-Presentación final del producto de datos, desplegado en Streamlit. -La SDM manifiesta que le gustaría tener unos lineamientos o mapa de ruta para el despliegue del producto en su nube Azure. -La SDM también indica que le gustaría potenciar la utilidad del producto de datos añadiendo la predicción de las causas de los accidentes por cuadrante. -La SDM manifiesta que valora de gran manera el resultado obtenido y que desea hacer uso efectivo de la herramienta. -El consultor propone la ejecución de un piloto usando el producto de datos para el primer trimestre de 2024. La SDM evaluará esta posibilidad.

Tabla 3. Bitácora de interacciones y retroalimentación con la SDM

18. Conclusiones Generales

- En su estado actual, el producto de datos permite predecir la probabilidad de accidentes futuros por cuadrante, con una semana de anticipación. Además, permite escoger entre diversos tamaños de estos cuadrantes, según la necesidad del negocio. Para actividades de corte estratégico (dígase, a nivel de localidad), será muy útil escoger cuadrantes de amplio tamaño. Mientras que, para acciones tácticas (UPZ o ZAT) será mejor tomar cuadrantes más pequeños.
- El usuario del producto de datos deberá tener muy claro que entre más pequeños los cuadrantes, habrá una mayor degradación de las métricas del modelo. Dicho en lenguaje de negocio, entre mayor resolución geográfica se requiera, el modelo será menos confiable. Sin embargo, es posible decir que el F1-Score obtenido permite ver que se ofrecen predicciones útiles para el negocio con cuadrantes de 2,8 Km/lado en adelante, estando esta utilidad para el negocio definida como la generación de predicciones creíbles (probabilidad mayor al 50%) sobre las zonas en las que va a haber accidentes, de modo que se amerite la toma de acción por parte de la SDM.
- El producto ha sido muy bien recibido por la SDM, considerándolo como un avance importante en los ejercicios de analítica de datos que se vienen desarrollando y como una herramienta cuyo uso valdrá la pena potenciar.

19.1 Próximos pasos y retos

- Para la SDM existen tres temas de relevancia que podrían potenciar el impacto del producto: mejorar el rendimiento del modelo predictor de accidentes para cuadrantes relativamente pequeños, incluir en la predicción el detalle sobre las causas que ocasionarán los accidentes, y automatizar la actualización automática del dashboard, una vez este se despliegue en la nube de la entidad.
- La propuesta que se ha establecido ante la SDM es ejecutar un piloto durante el primer trimestre de 2024, para evaluar el impacto real del producto de datos. La idea sería estimar la disminución de accidentes graves en tres localidades seleccionadas (Kennedy, Puente Aranda y Barrios Unidos) entre el primer trimestre de 2024 y el primero de 2023. Esto en línea con el KPI de negocio, definido en la sección 2.6 de este documento.