Modelo Predictivo para Mejorar La Seguridad Urbana en la Ciudad de Bogotá, D.C

Astrid Carolina Lezama Merizalde

Wilson Vargas Martínez

Universidad Ean

Facultad de Ingeniería

Maestría en Inteligencia de Negocios

Bogotá, Colombia

26/06/2024

Modelo Predictivo para Mejorar La Seguridad Urbana en la Ciudad de Bogotá, D.C

Astrid Carolina Lezama Merizalde

Wilson Vargas Martínez

Trabajo de grado presentado como requisito para optar al título de:

Magister en Inteligencia de negocios

Director (a):

Luis Armando Cobo Campo

Modalidad:

Monografía

Universidad Ean

Facultad de Ingenierías

Maestría en Inteligencia de Negocios

Bogotá, Colombia

15/02/2024

Nota de aceptación:

_____

_____

_____

_____

_____

_____
Firma del jurado

_____
Firma del jurado

_____
Firma del director del trabajo de grado

Ciudad, día/mes/año

“Dedico este trabajo de grado a mi hija, Isabel Milagros Hernández Lezama, que desde el cielo me impulsó a seguir adelante y realizar esta maestría.

A mi esposo, Daniel Fernando Hernández Pinto, y a mi hijo, Joseph Santiago Hernández Lezama, por ser la motivación y el motor de mi vida, inspirándome a ser siempre mejor.

A mis padres, por enseñarme que la disciplina vale más que el talento y que nunca es tarde para alcanzar nuestros sueños.

A mi hermana, Paola Andrea Preciado, por ser mi mejor amiga y aliada en las trasnochadas durante este camino.

A mi abuelita, Beatriz Pardo, por inculcarme el valor de la independencia a través del estudio.”

Carolina Lezama Merizalde

“A mis padres y hermana, cuyo amor incondicional, sacrificio y esfuerzo han sido la base de mi formación y crecimiento. Gracias por enseñarme el valor del trabajo duro y por ser mi ejemplo.

A Juan Pablo Botero y Joan Sebastián Tilagui, quienes más que amigos, son hermanos para mí. Su compañía, apoyo y palabras siempre estuvieron presentes en cada etapa, dándome la fuerza necesaria para seguir adelante.

A Jessica Tatiana Bravo, cuya inspiración llegó sin querer. Gracias por darme ideas y ánimo para encontrar soluciones, sin importar cuántas veces fallara en el intento. Tus palabras me empujaron a persistir y a no rendirme, incluso en los momentos más complicados.

A todos ustedes, les dedico este trabajo con profundo agradecimiento y cariño.”

Wilson Vargas Martínez

Agradecimientos

Queremos expresar nuestro más sincero agradecimiento a todos aquellos que contribuyeron a la realización de esta monografía. Sin su apoyo y orientación, este trabajo no habría sido posible.

En primer lugar, agradecemos profundamente a Luis Armando Cobo, nuestro tutor académico, por su invaluable guía, consejos y compromiso a lo largo de todo el proceso de investigación. Su experiencia y dedicación fueron pilares fundamentales para el desarrollo exitoso de este proyecto. Agradezco a Carolina mi compañera de un viaje en el que pocos creyeron y siempre persistimos, gracias por la confianza.

Extendemos también nuestro agradecimiento a nuestras familias y amigos. En particular, el autor Wilson Vargas Martínez desea expresar su profundo agradecimiento a Juan Pablo Botero Ramírez y Joan Sebastián Tilagui, quienes le brindaron un apoyo constante y lo alentaron en cada paso del camino. Su paciencia y palabras de ánimo fueron esenciales para superar los desafíos enfrentados durante este proceso.

Adicionalmente, Wilson Vargas Martínez quiere agradecer especialmente al Capítulo de Estudiantes de ACOFI encabeza de la ingeniera Luz Marina Patiño por creer en proyectos que parecían imposibles y por inspirarlo con el apoyo necesario para hacerlos realidad. Su confianza lo motiva a transformar lo imposible en posible.

Por su parte, la autora Carolina Lezama extiende un profundo agradecimiento a Edwin Fernando González por su invaluable apoyo en la comprensión y los avances teóricos. Sin su dedicación y guía, este trabajo no habría alcanzado su forma actual. A Wilson, mi compañero, por su comprensión y constante entusiasmo en este gran proyecto. A los profesores de mi carrera y de este posgrado, quienes me enseñaron las bases para llegar tan lejos y buscar la

excelencia. No puedo dejar de agradecer a mi familia, a quienes dedico este trabajo, porque sin su amor y apoyo, no sería la persona que soy hoy.

Resumen

La monografía titulada " Modelo Predictivo para Mejorar La Seguridad Urbana en la Ciudad de Bogotá, D.C" aborda la problemática de la inseguridad en Bogotá mediante el desarrollo de un modelo predictivo basado en técnicas de machine learning e inteligencia artificial. El objetivo principal es diseñar un prototipo que procese datos públicos para identificar y notificar áreas con alta incidencia de criminalidad, reduciendo así la exposición de los habitantes a zonas peligrosas. La metodología empleada incluye la recolección y procesamiento de datos de seguridad pública, la creación de un algoritmo de aprendizaje automático y la evaluación del modelo mediante métricas de desempeño específicas. Los resultados obtenidos demuestran que el modelo desarrollado puede mejorar significativamente la seguridad y la percepción de seguridad en la ciudad, optimizando la asignación de recursos y, en última instancia, mejorando la calidad de vida de los ciudadanos.

Palabras clave: seguridad urbana, modelo predictivo, machine learning, inteligencia artificial, Bogotá.

Abstract

The monograph titled "Predictive Model to Improve Urban Security in the City of Bogotá, D.C." addresses the issue of insecurity in Bogotá through the development of a predictive model based on machine learning and artificial intelligence techniques. The main objective is to design a prototype that processes public data to identify and notify areas with a high incidence of crime, thereby reducing residents' exposure to dangerous zones.The methodology employed includes the collection and processing of public security data, the creation of a machine learning algorithm, and the evaluation of the model using specific performance metrics. The results obtained demonstrate that the developed model can significantly enhance security and the perception of safety in the city, optimizing resource allocation and ultimately improving citizens' quality of life.

Keywords: urban safety, predictive model, machine learning, artificial intelligence, Bogotá.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Introducción

La seguridad es un aspecto importante dentro de un territorio, esta influye en la calidad de vida como uno de los ejes principales en los residentes de la comunidad. La sensación de seguridad que sienten las personas al desplazarse por la ciudad es fundamental.

La creación de modelos predictivos utilizando datos proporcionados por los ciudadanos y autoridades dentro de un territorio, puede ser una herramienta valiosa para mitigar el impacto de la inseguridad, al permitir la optimización de recursos en zonas que requieren de intervención. Este modelo puede ayudar a reducir el miedo y la ansiedad que sienten las personas, mejorando así significativamente su calidad de vida.

De acuerdo con la información revisada, no existen aplicaciones que generen de forma nativa análisis predictivos para identificar áreas de alta criminalidad y ayudar a los ciudadanos a asemejar fácilmente las zonas, que en sus desplazamientos tengan algún tipo de nivel de inseguridad en función de sus necesidades en la realidad de la ciudad.

Sin embargo, existen aplicaciones como Waze la cual es “una aplicación de navegación GPS detallada que proporciona información sobre el tráfico en tiempo real, además de todo tipo de elementos sociales interesantes y de geogaming que añaden diversión a los desplazamientos. Los wazers pueden enviarse información sobre el tráfico, controles policiales, obras, radares y mucho más” (Google,2023), que pueden proporcionar las instrucciones para realizar un viaje sin considerar variables influyentes más allá de la ruta rápida. Estas aplicaciones tienen algunas limitaciones, como advertencias de seguridad inexactas y falta de personalización sobre la ejecución de los trayectos en curso.

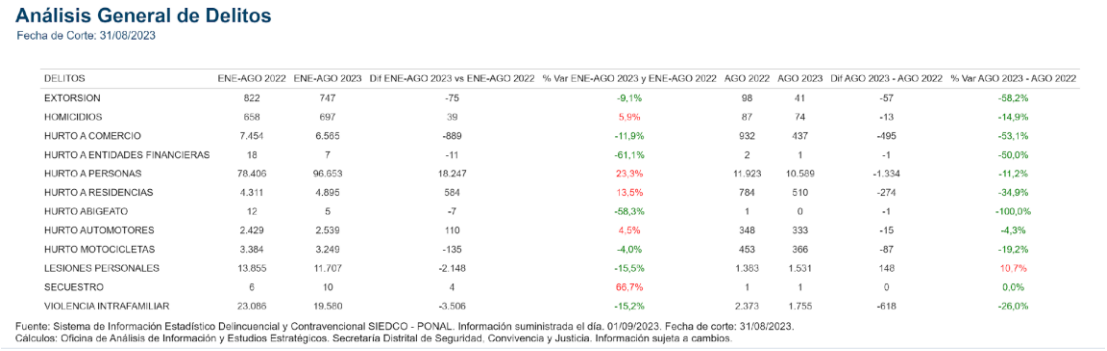
En el contexto de la ciudad de Bogotá, el análisis general de delitos, según la Secretaría Distrital de Seguridad, Convivencia y Justicia (2023), se presenta en la Figura 1 y abarca el período comprendido entre 2022 y agosto de 2023. Este análisis resalta la importancia del diseño del modelo predictivo propuesto, dado que se identifican diversas tendencias delictivas durante el mismo período. En particular, se observa un aumento del 5,9% en el número de asesinatos (39 casos adicionales), mientras que los robos en comercios experimentaron una disminución del 11,9% (-889 casos). A pesar de esta reducción, la Figura 1 también indica una notable disminución en los asesinatos en general. Además, los casos de robo en instituciones financieras se redujeron en un 61,1% (-11 casos). Sin embargo, se reportó un incremento significativo en los hurtos, que aumentaron un 23,3% (18,247 casos más), así como en los secuestros, que crecieron un 66,7% (4 casos adicionales).

Es relevante mencionar la disminución en los índices de daños personales (-15,5%) y de violencia doméstica (-15,2%), tal como se ilustra en la Figura 1. Estos datos evidencian diferencias significativas en la incidencia de distintos tipos de delitos, lo que subraya la necesidad de abordar áreas críticas y de mantener estrategias efectivas de prevención del delito.

La continua fluctuación y el aumento de los delitos analizados en la Figura 1 indican que Bogotá requiere la implementación de medidas efectivas para mejorar la seguridad de sus ciudadanos, con el objetivo de mitigar los impactos negativos en sus residentes y visitantes. En este sentido, una adecuada recolección, análisis y procesamiento de los datos delictivos a través de un modelo predictivo puede contribuir a abordar las deficiencias actuales en la ciudad.

Finalmente, en la Figura 1 se presenta una clasificación cuantitativa de los diferentes delitos correspondientes a los años 2022 y 2023, así como una comparación de los índices delictivos de un año a otro en términos porcentuales.

Figura 1 Análisis de delitos de la secretaría de seguridad de Bogotá 2022-2023



Análisis General de Delitos

Fecha de Corte: 31/08/2023

DELITOS	ENE-AGO 2022	ENE-AGO 2023	Dif ENE-AGO 2023 vs ENE-AGO 2022	% Var ENE-AGO 2023 y ENE-AGO 2022	AGO 2022	AGO 2023	Dif AGO 2023 - AGO 2022	% Var AGO 2023 - AGO 2022
EXTORSION	822	747	-75	-9,1%	98	41	-57	-58,2%
HOMICIDIOS	658	697	39	5,9%	87	74	-13	-14,9%
HURTO A COMERCIO	7.454	6.565	-889	-11,9%	932	437	-495	-53,1%
HURTO A ENTIDADES FINANCIERAS	18	7	-11	-61,1%	2	1	-1	-50,0%
HURTO A PERSONAS	78.406	96.653	18.247	23,3%	11.923	10.589	-1.334	-11,2%
HURTO A RESIDENCIAS	4.311	4.895	584	13,5%	784	510	-274	-34,9%
HURTO ABIGEATO	12	5	-7	-58,3%	1	0	-1	-100,0%
HURTO AUTOMOTORES	2.429	2.539	110	4,5%	348	333	-15	-4,3%
HURTO MOTOCICLETAS	3.384	3.249	-135	-4,0%	453	366	-87	-19,2%
LESIONES PERSONALES	13.855	11.707	-2.148	-15,5%	1.383	1.531	148	10,7%
SECUESTRO	6	10	4	66,7%	1	1	0	0,0%
VIOLENCIA INTRAFAMILIAR	23.086	19.580	-3.506	-15,2%	2.373	1.755	-618	-26,0%

Fuente: Sistema de Información Estadístico Delincuencial y Contravencional SIEDCO - PONAL. Información suministrada el día. 01/09/2023. Fecha de corte: 31/08/2023.
Cálculos: Oficina de Análisis de Información y Estudios Estratégicos. Secretaría Distrital de Seguridad, Convivencia y Justicia. Información sujeta a cambios.

Fuente: *Recuperado de* http://analitica.scj.gov.co/analytics/saw.dll?Portal&PortalPath=/shared/OAIEE/SIEDCO/_portal/An%C3%A1lisis%20de%20datos%20Siedco&NQUser=publico&NQPassword=publico2019

A raíz de lo identificado en las aplicaciones y algunos estudios generados sobre el tema de seguridad para ciudades en el territorio colombiano, surge la pertinencia de plantear y construir un modelo que tenga en cuenta diferentes aristas permitiendo minimizar el impacto de la criminalidad ante la reacción de las autoridades y un sistema de alertas a los ciudadanos en el territorio conforme el desplazamiento de estos en un espacio geográfico mapeado dentro del modelo.

Todo lo anterior lleva a plantear la siguiente pregunta de investigación:

¿Cómo se puede mejorar la seguridad en Bogotá a través de alertas basadas en modelos predictivos que ayuden a evitar áreas de alta criminalidad cuando la gente se desplaza en la cotidianidad?

Objetivos

Objetivo general

Desarrollar un prototipo de un modelo predictivo de alertas para Bogotá, empleando machine learning e inteligencia artificial, que procese datos públicos disponibles para identificar y notificar áreas con alta incidencia de criminalidad, contribuyendo así a reducir la exposición de los habitantes de la ciudad a zonas con altos índices de criminalidad en su cotidianidad.

Objetivos específicos

Diseñar y desarrollar un prototipo de algoritmo de aprendizaje automático capaz de procesar, analizar y predecir zonas seguras e inseguras, considerando la ubicación geoespacial y el contexto histórico local.

Recolectar la información de seguridad de Bogotá disponible y procesarla para poder ser consumida por el prototipo del modelo predictivo desarrollado.

Plantear y evaluar el prototipo modelo predictivo desarrollado, mediante la utilización de métricas de desempeño acordes al tipo de modelo y la comparación con algún sistema existentes, para validar su efectividad y aplicabilidad en el contexto de Bogotá.

Justificación

La inseguridad en varias de las ciudades de Colombia, especialmente en Bogotá, ha sido un tema con un alto impacto en los últimos años, aumentado por factores como lo es la pandemia. Según datos de la Secretaría Distrital de Seguridad (2023), el nivel de inseguridad percibida en Bogotá aumenta considerablemente, mientras que más del 60% de los ciudadanos afirman sentirse inseguros en sus barrios.

Esto supone no solamente un incremento de la incidencia de la criminalidad, sino que las autoridades precisan aplicar y desarrollar estrategias más restrictivas con el fin de solventar de una manera más eficiente este problema. Un sistema mejor desarrollado de seguridad permeado por ML e inteligencia artificial puede tener un buen impacto en el modelo de políticas públicas.

De esta manera, el análisis generaría ensayos precisos a partir de los cuales se generaría el conocimiento de las áreas críticas donde deben patrullar las fuerzas de policía preventiva. Estas, por su parte, ayudarían a la generación instantánea de alerta inmediata y planes de intervenciones con argumentos y hallazgos concretos; en este espacio con poder, los gobiernos locales pueden desarrollar políticas más proactivas y seguras.

Por último, esto contribuiría a la mejora de la seguridad y la reducción de la percepción de inseguridad, reforzando la confianza en las instituciones; en este sentido, debería apoyar la campiña sosa de riqueza.

Marco Teórico

La era de los datos ha transformado los pasos de ciertas áreas en la que están evolucionando, el machine learning e inteligencia artificial, estos son esenciales para el concepto de las ciudades inteligentes. Como conjunto de métodos, el machine learning tiene la capacidad de detectar patrones automáticamente en los datos y, con base en ellos, predecir un resultado futuro o tomar decisiones en estado de incertidumbre. Dichos métodos tienen aplicaciones prácticas en ramas filosóficas como la financiera, donde tratan de predecir conductas y magnificar estrategias. Desde las ciudades inteligentes, el paradigma predictivo resulta ser un reto significativo para la aplicación en la seguridad urbana.

Desde la calidad y disponibilidad de datos fiel hasta el cualitativo sobre sus interrogantes, las limitaciones afectan la precisión predictiva. Además, los sesgos presentes en los datos históricos garantizan un entorno propicio para la persistencia de perentor. Un informe del mismo laboratorio nacional de Argonne (2023) señala que en efecto es fundamental lanzar técnicas de inteligencia artificial transparente (XAI) que permiten interpretar los resultados de modelos en una forma explícita y ética, de forma que se entregue importancia en cuanto al trabajo a autos sustitutivos que son justos y justas. Se puede deducir entonces que, aunque el ML ofrece masa de machine learning, son necesarios estos desafíos de gestión ética de datos y redes de creación para garantizar resultados invariables y responsables.

En el machine learning se pueden diferenciar distintos enfoques, los cuales se pueden agrupar en dos grandes categorías que son el método supervisado y el método no supervisado, en donde el aprendizaje supervisado es uno de los más empleados y su uso puede ser para

sistemas como la clasificación o la predicción. Aunque clásicamente se ha asociado al caso de la regresión lineal, esta metodología tampoco se limita a la predicción de conductas lineales.

Modelos de relaciones no lineales también se pueden ejecutar a través de modelos avanzados como los árboles de decisión, las redes neuronales convolucionales y el modelo recurrente (CRNN), al igual que los modelos de ensamble como son los bosques aleatorios (Random Forest). Un buen ejemplo para ejemplificar este tipo de aprendizaje supervisado es el modelo de clasificación de flores de tipo Iris, tal como lo expone Murphy (2012), en el que se 'learn to distinguish three different kinds of iris flower, called setosa, versicolor and virginica' (p. 7). Este pequeño ejemplo pone de evidencia que los algoritmos supervisados son aplicables para aprender patrones complejos a partir de características específicas, más allá de relaciones lineales simples.

Figura 2 Tipos de flores Iris para modelo de clasificación

Fuente: recuperado de *http://www.statlab.uni-heidelberg.de/data/iris/*

El segundo tipo de aprendizaje es el aprendizaje no supervisado, el cual Murphy (2012) “where we are just given output data, without any inputs” (p. 9). Este tipo de aprendizaje es el más semejante al aprendizaje cognitivo de los animales o humanos, ya que el procedimiento de

aprendizaje no tiene un objetivo explícito; el modelo busca los patrones y estructuras dentro de los datos.

Un caso de ejemplo de aprendizaje no supervisado se puede observar en una tarea de clasificación que consiste en clasificar rostros con o sin gafas como se evidencia en la figura 3, el modelo observa las diferentes características como la forma de los ojos, presencia de complementos, etc. y, aunque no existe un listado con etiquetas previamente definidas por un ser humano, en función de encontrar estructuras, el sistema aplicará algoritmos como el análisis de componentes principales (PCA) o métodos de agrupamiento (clustering), reconociendo posibles patrones comunes dentro de las imágenes y procediendo a clasificar o agrupar las imágenes. Así pues, este modelo poco a poco clasifica la información, mostrando de esta manera la capacidad del aprendizaje no supervisado para poder detectar relaciones ocultas dentro de los datos.

Figura 3 Imágenes de caras aleatorias, b los principales vectores obtenidos.

.

Fuente: Machine Learning Book, Kevin P. Murphy 2012

A partir de los entramientos de datos por medio del machine learning, toma fuerza los modelos de redes neuronales y de ahí, el Deep Learning. Según (Goodfellow, Bengio, & Courville, 2016) “Deep learning known as cybernetics in the 1940s–1960s, deep learning known

as connectionism in the 1980s–1990s, and the current resurgence under the name deep learning beginning in 2006.”(P. 13), de acuerdo con esto, el Deep learning es un entrenamiento que surge entre los años 40 y 60, que se ha ido transformando en lo que hoy conocemos como redes neuronales artificiales, desde la perspectiva de (Goodfellow, Bengio, & Courville, 2016) “is that they are engineered systems inspired by the biological brain (whether the human brain or the brain of another animal)”(P.13) , esto, impulso varias ideas como la ingeniería inversa como una funcionalidad innata del cerebro, el cual su único fin es duplicar la información para aprenderla. Sin embargo, se ha dado una mayor complejidad al terminó, debido a que desde la neurociencia, se considera que el aprendizaje tiene como principio aprender en diferentes niveles de composición, lo cual ha impulsado los frameworks actualmente usados en el machine learning. (Goodfellow, Bengio, & Courville, 2016)

Aun así, el Deep learning se sigue basando en los datos, pero en una alta complejidad, buscando patrones que puedan dar una predicción cercana a lo que haría un cerebro humano o animal (Goodfellow, Bengio, & Courville, 2016), es por eso, que en la búsqueda de los avances tecnológicos, de cómo resolver problemas por medio de los datos o la información, se ha buscado la máxima similitud con la forma de resolución de los humanos. El entendimiento del cerebro humano, no solo en su aprendizaje sino también en su psicología, el actuar, su forma de pensar, las matemáticas y la economía, llevan a la definición de la inteligencia artificial en 4 categorías, como se observa en la figura 4: Piensa Humanamente, que hace referencia a las actividades que surgen de los pensamiento humanos y hacen que estos tomen decisiones para actuar; Actuar humanamente: realizar funciones que requieren de inteligencia y pensamiento humano; Pensar racionalmente: Facultades para razonar y tener percepción de

las cosas ; Actuar racionalmente: Actuar de acuerdo con los pensamientos racionales. (Russell & Norvig, 2010).

Figura 4 Categorías de la definición de inteligencia artificial.

Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Fuente: Recuperado de Artificial Intelligence, Rusell & Norving.

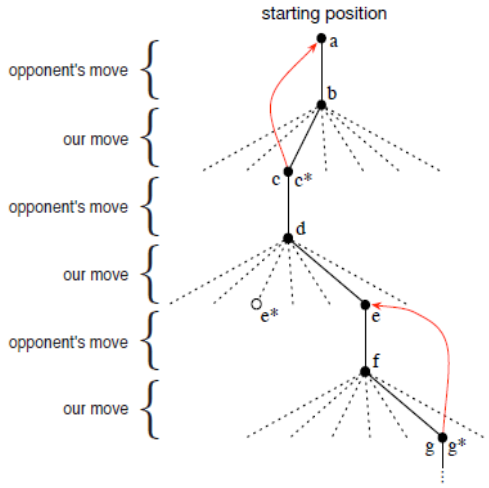
Algunas de las facultades mencionadas por Rusell & Norving (2010) para la categoría de actuar humanamente, estaba el procesamiento de lenguaje natural, representación del aprendizaje, respuesta de preguntas y sacar conclusiones. Para las otras categorías se considera un poco más complejo debido a el entendimiento y la funcionalidad del cuerpo más allá del cerebro.

Con el entendimiento y uso de las diferentes tecnologías, que buscan simular o duplicar el comportamiento humano, se ha empezado a utilizar para optimizar y mejorar diferentes ámbitos de la vida cotidiana, como lo es las finanzas, el aprendizaje, optimización de proceso y hasta la seguridad de las diferentes ciudades.

El Reinforcement Learning, según Sutton y Barto (2018), "is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told

which actions to take, but instead must discover which actions yield the most reward by trying them" (p.1). Este tipo de aprendizaje busca que las acciones realizadas tengan la mayor probabilidad de lograr el objetivo deseado. Un claro ejemplo de esto es el juego del triqui, o Tic-Tac-Toe, como lo llaman Barto y Sutton (2010). En este juego, el objetivo es alinear tres "O" o "X" en una cuadrícula de 3x3. Las acciones tomadas en este juego pueden interferir con las posibilidades de victoria del oponente. Así, el modelo puede tomar decisiones basadas en las acciones del oponente y las diferentes posibilidades de movimiento, tanto para el modelo como para el oponente. Como se muestra en la Figura 5, las posibles acciones se representan con líneas punteadas, mientras que las decisiones tomadas se indican con líneas negras, todo esto en forma de árbol (Sutton y Barto, 2018).

Figura 5 Secuencia de movimientos de juego Triki



Fuente: Recuperado de Reinforcement Learning Sutton & Barto

Con esta probabilidad para ejercer acciones, las diferentes fuerzas públicas podrían realizar un modelo predictivo que permita tomar decisiones a partir de las acciones que realicen en el vandalismo de las ciudades.

Al implementar distintos modelos en las ciudades, incluido el reinforcement learning, ha nacido el termino ciudades inteligentes, la cual (Townsend, 2013) lo define como “places where the information technology is wielded to address problems old and new” (P.2), en el cual se expresa que las muchas tecnologías como las implementadas por IBM, Cisco y muchas otras empresas, sea usadas para los problemas del común de las ciudades. Con el nacimiento de los smartphones y las nuevas tecnologías emergentes, se pueden solucionar muchos problemas como el tráfico, la contaminación, la inseguridad, la ubicación, el turismo, etc. Las ciudades inteligentes, buscan la sinergia de la tecnología con la vida cotidiana, en pro de mejorar la calidad de vida de los ciudadanos.

Como lo menciona Townsend (2013), estas ciudades inteligentes tienen incorporadas en su gran mayoría sensores, dispositivos y sistemas de información que recompilan, la información y la analizan, con el fin de dar información relevante a las autoridades y residentes, que le permitan tomar decisiones. Además, buscan que la naturaleza también se vea beneficiada, por la reducción de consumo de la energía, agua y otros recursos vitales que podrían ser optimizados por medio de las herramientas brindadas por la evolución de la tecnología. (Townsend, 2013).

El machine learning en el uso para mitigar la criminalidad

En la revista Ibérica de Sistemas y Tecnologías de la Información se publicó un artículo en el año 2020 sobre un modelo de machine learning para predecir las tendencias de hurto en Colombia, para ello, los autores hicieron uso de un modelo de máquinas de soporte vectorial,

después de contrastar diferentes opciones de modelos implementados en diferentes ciudades del mundo.

Partiendo de lo anterior, Ordoñez Hugo, Cobos Carlos y Víctor Bucheli (2020) denotan dos grupos de algoritmos que se han aplicado, el primer grupo solo a entender y aprender los patrones de crimen en un lugar tal y como señalan a continuación “Algunos trabajos se centran en encontrar patrones de crimen en ciudades. Para las ciudades de California-Irvine y del estado de Misisipi (McClendon and Meghanathan 2018), datos que están disponibles en https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime y https://www.neighborhoodscout.com. Con base en dichos datos y aplicando técnicas de machine learning se detectan patrones de crimen.

Con los datos obtenidos y las técnicas de regresión lineal, additive regression, decision stump predicen el número de asesinatos, hurtos, entre otros” (P.4), es decir, que el uso del aprendizaje automático se ha vuelto una herramienta para atacar este problema y ser de utilidad para la autoridad, así como ,los investigadores donde tienen cabida algoritmos, tales como, la regresión lineal y la toma de decisión a través de los conjuntos de datos que se recopilan en la ciudades, permitiendo un grado determinado de confiabilidad sobre patrones de criminalidad para formular estrategias de atención y repuesta que permitan mitigar el problema.

Ahora bien, en el segundo grupo se hace uso de algoritmos más robustos y su enfoque se centra en la predicción, para este caso los autores Ordoñez Hugo, Cobos Carlos y Víctor Bucheli (2020) hablan sobre un lugar más específico como Vancouver donde dicen lo siguiente “En Vancouver, donde se busca predecir los delitos a partir de aprendizaje automático. En este trabajo, se recopilaron datos de delitos de Vancouver en los últimos 15 años, los cuales se analizaron mediante modelos K-nearest-neighbour (k-nn) y boosted decision tree, para los

cuales se obtiene una exactitud de la predicción del delito entre 39% y 44% (Kim et al. 2018)" (P.4) de ello se puede inferir, que para el caso de Vancouver y otras ciudades se aplicaron las técnicas como K-nearest-neighbour , Risk Terrain Modeling , Kernel density estimation , redes neuronales y arboles de decisión, sin embargo, los resultados de predicción tuvieron un rango de asertividad significativamente bajo, a pesar del uso de un conjunto de datos de 15 años, aun así, se pueden rescatar los indicadores generados abriendo la posibilidad a usar algoritmos más adecuados o con nuevas variables en su estructura para tener una predicción proactiva en el futuro, con un margen de error más bajo y así, destacar que los modelos siempre tienen márgenes de adecuación y mejora, lo que permite sea puedan refinar en sus futuras versiones.

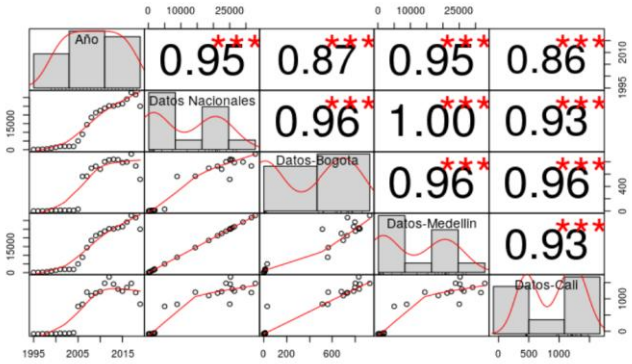
Los autores Ordoñez Hugo, Cobos Carlos y Víctor Bucheli (2020) a través del estudio y análisis de los diferentes casos aplicados de las técnicas de aprendizaje automático generan el siguiente modelo como propuesta "la técnica seleccionada fue las máquinas de soporte vectorial, específicamente las orientadas a problemas de regresión.

El modelo utiliza un dataset tomado del sistema de información de la fiscalía nacional de Colombia, el cual cuenta con un total de 2,662,402 registros de delitos realizados en Colombia desde el año 1960 hasta el 2019." (P.5), para esto, se usa esta técnica como una extensión de un problema de regresión dada la naturaleza de los datos que se recopilan y se dispone de un dataset de una fuente gubernamental que permita aumentar la confiabilidad del modelo que tiene un histórico amplio.

El SVR según los autores Ordoñez Hugo, Cobos Carlos y Víctor Bucheli (2020) "Support Vector Machines (SVM) también se puede utilizar como método de regresión, manteniendo todas las características principales que caracterizan el algoritmo (margen máximo). La regresión basada en soporte vectorial (SVR) utiliza los mismos principios que SVM para la

clasificación, con solo algunas diferencias menores” (P.5), es decir, se habla de SVM por su capacidad de clasificar mediante la regresión en un conjunto de datos entendiendo el margen máximo generando cierta flexibilidad al modelo.

Figura 6 Correlación modelo propuesto



Fuente: Recuperado de *https://www.proquest.com/openview/fb8bfe36673b48be2d035ee8a035c307/1?pq-origsite=gscholar&cbl=1006393*

Frente a la figura 6 y los resultados los autores Ordoñez Hugo, Cobos Carlos y Víctor Bucheli (2020). “Una vez analizadas las relaciones entre las variables, como primera instancia, se evaluó el modelo, con un modelo de regresión lineal estándar (línea negra) frente a un modelo de SVR estándar (línea azul) y el modelo ideal establecido por los datos reales (línea roja). Se puede observar que el modelo predice de manera armónica los valores de la variable dependiente, ya que sus valores guardan mayor cercanía y no están muy dispersos al híper plano trazado para la regresión, esto se debe a que SRV, minimiza la distancia entre cada punto de la variable dependiente al margen de tolerancia épsilon. En este caso, los valores de la variable hechos-año.

En este sentido, el modelo predice, que el punto más alto es en el año 2017 con un valor promedio de 32890 hurtos, pero a medida que aumenta el valor de la variable independiente, los valores de la variable dependiente tienden a bajar, prediciendo así, que para el año 2020 el número de delitos de hurto disminuirá en Colombia estará entre 25027 y 32890" (P.5) Se puede decir que, predice un valor numérico continuo con la finalidad de encontrar un punto que separe los datos de la forma más eficiente posible, reduciendo el margen del error entre los coeficientes de la predicción con los datos que se proporcione al modelo, conforme lo dispuesto por los autores, donde la efectividad armoniza con las variables dependientes, lo que implica que las predicciones están más cerca de la línea de regresión a través del RMSE y la forma en que contrastaron fue con una métrica para medir la dispersión de los residuos en la predicción.

A través del estudio de este modelo, se puede evidenciar que las técnicas de aprendizaje automático, tienen el potencial para desempeñar un papel significativo en el análisis y predicción de tendencias del crimen en diferentes contexto acorde a las variables que se capturen en los conjuntos de datos incluyendo la forma de mitigar los desafíos asociados con la exactitud en la predicción, el aprendizaje automático ha demostrado ser de utilidad en la lectura e interpretación de patrones complejos en datos criminológicos y proporcionar indicadores que podrían ser cruciales para la formulación de estrategias de seguridad pública basadas en datos, permitiendo de esta manera una transición a ciudades inteligentes cada vez más seguras en relación al avance y optimización de los algoritmos que son constantemente evaluados con el mismo propósito mejorar la sociedad.

El machine learning para predecir la tasa de criminalidad en barrios urbanos

En el artículo *Prediction of a crime rate in urban neighborhoods based on machine learning* disponible de manera online desde septiembre del año 2021, He y Zheng, mencionan en el artículo el principal problema y objetivo del modelo predictivo que desarrollan como “only considers crime data and does not consider environmental factors, the importance of considering environmental factors stems from the broken window theory (Wilson and Kelling, 1982) and crime prevention through the environment (Casteel and Peek-Asa, 2000; Cozens et al., 2005). Slightly more advanced techniques, such as multiple regression, do improve predictive power, but still rely only on historical crime data. And the success of these methods relies heavily on the correct selection of indicator variables. And considering additional data sets leads to a significant increase in valid features (Stec and Klabjan, 2018).” (P.2), donde se habla de considerar, no únicamente los datos históricos delictivos de la ciudad, sino también, factores relevantes para el modelo, como el factor ambiental de las diferentes zonas que hacen parte del modelo. Además, también es importante mencionar que las diferentes técnicas utilizadas como las de regresión, se basan únicamente en los datos y no dejan tener en cuenta algunas variables que podrían llegar a ser significativas para el modelo.

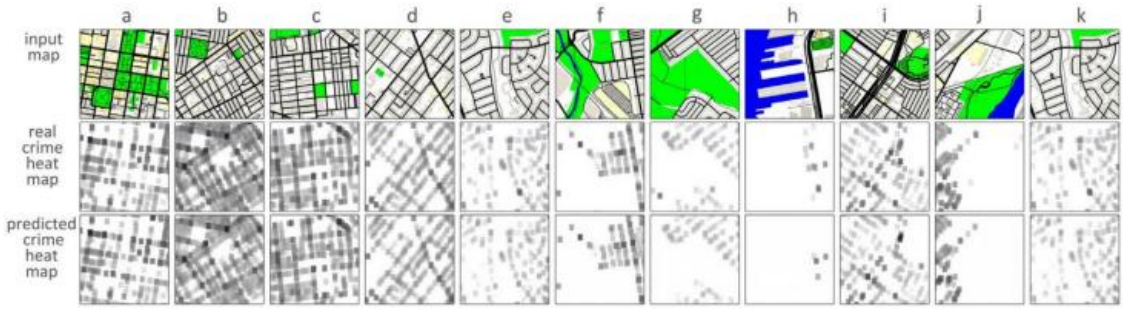
La propuesta de He y Zheng, ronda con respecto a las redes generativas adversarias (GAN), una de las muchas técnicas de modelos de redes neuronales convolucionales, que a diferencia de otra tiene en cuenta la perdida de información, como lo menciona es su artículo He y Zheng “GAN uses both content loss and adversarial loss, it has more unique image simulation and representation capabilities than traditional algorithms (Yangjie et al., 2018). And GAN does not have complex variance lower bounds, which can greatly reduce the difficulty of training and improve training efficiency (Gonog and Zhou, 2019).” (P.2). Y no siendo esto suficiente para la elección de su técnica para modelamiento, también se expone como las GAN son muy

efectivas a la hora de utilizar imágenes de un antes y después de la información relevante para la generación del modelo. El modelo GAN, He y Zheng lo describen como “GAN is a neural network based on inputting graphics and outputting images.” (P.3), la cual en este caso se busca ayudar a los arquitectos a optimizar la planificación y diseño de áreas más seguras.

El objetivo de articulo, es mostrar el desarrollo, prueba y conclusión de un modelo generativo a partir de redes neuronales convolucionales GAN, para predecir puntos críticos de criminalidad en una ciudad, en este caso Filadelfia. Para tras ingresar las entradas del modelo, obtener un mapa de calor que permita ver cuáles son las zonas que se pueden mejorar por medio de la arquitectura y del modelo GAN para generar mayor seguridad a la hora de diseñar las arquitecturas de los diferentes barrios.

Para los datos objetivos, se seleccionó la información proveniente de la ciudad de Filadelfia entre el 2006 y el 2018, debido a su fácil acceso a la información, en dicha información se encuentra información relevante como coordenadas no exactas de sitios donde sucedió el crimen, no se tiene en cuenta los tipo de delitos para el desarrollo del modelo, sim embargo se realiza un procesamiento de datos por medio de scripts de Python , para convertir la información en el mapa de calor de crimen, para buscar las frecuencias de los delitos sobre el mapa. Tras realizar el mapa de los datos, He y Zheng, consideran necesario obtener un mapa real de Filadelfia, que será la entrada de su modelo GAN, de los cuales se obtienen 1500 mini mapas de la ciudad. Tras realizar el entrenamiento, como se puede observar en la figura 7, Se obtiene la información de los diferentes mapas con sus entradas y salidas, según los resultados previstos para los más de 1000 mini mapas, de los que ya se habían mencionado anteriormente.

Figura 7 Resultados previstos para el entrenamiento



Fuente: *https://pdf.sciencedirectassets.com/271095/1-s2.0-S0952197621X00085/1-s2.0-S0952197621003080/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEF4aCXVzLWVhc3QtMSJGMEQCIEt7kFA2f0bnX6NiIrsDabaPWy6EgYbBof8PIB*

Continuando con su investigación He y Zheng, proponen revisar en ciudades de diferentes tamaños, para la cuidad de menor tamaño se eligió Princeton que contiene una distribución compleja de calles, para la ciudad intermedia se eligió Seattle que tiene grandes jardines y edificios y en las grandes ciudades se eligió Nueva York. Y aunque el enfoque en este punto del artículo ya no va dirigido a un modelo exclusivamente de tasas de criminalidad, si se busca que la distribución de las calles y del diseño de la ciudad sea obtenido por medio de los modelos, con el fin de reducir la criminalidad, ya como un área de enfoque hacia la disminución de los crímenes con una redistribución de la ciudad en sus diferentes escalas.

En conclusión, He y Zhen identificaron “Although crime centers are often concentrated on main roads, not all main roads have high crime rates (Beavon et al., 1994). In contrast, in the vicinity of parks and elementary schools, the crime rate is lower due to the presence of neighborhood surveillance (Kim and Shin, 2014). Due to naturally occurring surveillance

mechanisms in these areas, more strangers or passersby on well-traveled, highly visible streets can benefit as a crime prevention strategy (Shu, 2009). Because the extents of the effects of these factors change, the degree of impact on output tends to vary in combination with different urban layouts. This it is not the case that a particular urban layout has a certain effect on crime rates.” (P.11), donde no solo se ve como factor relevante para la tasa de criminalidad de las diferentes ciudades el histórico de delitos, sino también el ambiente, distribución y escala de la ciudad, como un factor relevante a la hora de tener en cuenta un ecosistema global de las diferentes variables que puede llegar a afectar la tasa de criminalidad de una ciudad.

Hipótesis

En respuesta a los modelos que han comenzado a desplegarse en diferentes ciudades del mundo, así como la integración del concepto de ciudades inteligentes, se propone un cambio de enfoque para Bogotá. Se sugiere evolucionar desde un enfoque pasivo, que genera modelos que utilizan solo los datos para determinar patrones en los conjuntos de datos desde una perspectiva meramente descriptiva, hacia un enfoque más activo.

Este enfoque implicaría la implementación de algoritmos más robustos que aprovechen las tecnologías emergentes actuales.

Entre estas tecnologías, se destaca la generación de modelos de redes convolucionales recurrentes (CRNN, por sus siglas en inglés: Convolutional Recurrent Neural Network), que permitirían analizar e integrar datos históricos y en vivo sobre criminalidad, movilidad y percepciones de seguridad. Además, estos modelos podrían generar alertas en un único conjunto de datos, facilitando la identificación temprana de áreas de trabajo en el campo de la seguridad y la atención a puntos geográficos con altas tasas de afectación a la seguridad. A su vez, también permitirían a las autoridades locales generar planes que faciliten la identificación de áreas que requieren más atención y establecer políticas públicas para optimizar los recursos disponibles en seguridad.

Se aspira a que el modelo pueda ser alimentado por un flujo continuo de datos, lo que permitiría establecer que los coeficientes entre las variables tengan valores razonables dentro de las estimaciones. Además, se espera que este conjunto de datos pueda servir como un puente para conectar a las autoridades, comunidades e interesados, y pueda ser regulado para garantizar su confiabilidad. De esta manera, se contribuiría a una mejora gradual de la calidad

de vida de los habitantes del espacio geográfico donde se implemente, ofreciendo una solución más realista que lo observado en propuestas anteriores.

En esas propuestas se expresaba que se requerirían varios cientos de décadas para mitigar el problema. A través de esta nueva propuesta, se busca una reducción proactiva de los incidentes en este eje que afectan a la ciudad, permitiendo que las autoridades puedan tener una respuesta más rápida cuando estos hechos se presenten. Además, se aspira a que el modelo pueda ser optimizado conforme el avance de los algoritmos y la tecnología lo permita.

Variables

Variable: Criminalidad

Definición Conceptual: Este indicador describe el estado, frecuencia y gravedad de los diferentes delitos en la ciudad de Bogotá.

Definición Operacional: Se medirá a través del uso de la información histórica de delitos consolidada en las fuentes abiertas de datos y se calculará un índice ponderado para cada localidad. Además de utilizar la información histórica se sustentará a través de la encuesta de percepción de criminalidad.

Clasificación: Variable cuantitativa.

Variable: Alertas de Seguridad

Definición Conceptual: Disparador de notificaciones basadas en el análisis de datos proporcionado por el modelo para informar a los usuarios sobre áreas de alto riesgo en tiempo real.

Definición Operacional: Se medirá la efectividad de las alertas de seguridad a través de la encuesta, preguntando sobre las herramientas que se utilizan de estas alertas y su percepción sobre el impacto de estas en la seguridad ciudadana.

Clasificación: Variable cualitativa y cuantitativa.

Variable: Mejora en la calidad de vida

Definición Conceptual: Evaluación general del impacto del modelo en la seguridad y percepción de seguridad de los bogotanos.

Definición Operacional: Se realizaría seguimiento a través de encuestas de percepción de seguridad y datos estadísticos sobre la disminución de delitos en las fuentes gubernamentales. Complementada con datos estadísticos sobre la disminución de delitos.

Clasificación: Variable cualitativa y cuantitativa.

Variable: Movilidad

Definición Conceptual: Evaluación de la movilidad enfocada en la seguridad de los ciudadanos de Bogotá, teniendo en cuenta factores como la seguridad zonificada, la criminalidad de la zona y la accesibilidad segura a las diferentes zonas de la ciudad.

Definición Operacional: Se medirá a través de la encuesta que abarca preguntas sobre la percepción de seguridad en diferentes modos de transporte y áreas de la ciudad, además, del análisis de indicadores como el número de incidentes delictivos reportados, la tasa de criminalidad en diferentes áreas de la ciudad y el acceso a servicios de seguridad, como la presencia policial y la iluminación pública.

Clasificación: Variable cualitativa y cuantitativa.

Metodología

Enfoque y alcance de la investigación

Enfoque de la Investigación

El desarrollo de la investigación tendrá un enfoque mixto, haciendo uso integral de elementos cualitativos y cuantitativos de forma que permita estudiar y entender de mejor manera los datos existentes y, al mismo tiempo, integrar las perspectivas cualitativas de investigaciones con un carácter similar que permitan desarrollar un modelo más óptimo. Además del uso de la encuesta de percepción de criminalidad para una comprensión integral del tema.

Diseño de Investigación

La investigación se caracterizará por ser no experimental y transversal. El objetivo es recopilar y analizar datos cuantitativos sobre la seguridad histórica en la ciudad de Bogotá durante un periodo reciente, complementado con datos cualitativos obtenidos de literatura relevante al tema, así como de una encuesta sobre la percepción de seguridad. Esto permitirá establecer correlaciones entre las variables, enriqueciendo así la comprensión del problema y la evaluación del modelo propuesto.

Tipo de Investigación

La investigación será de naturaleza descriptivo, exploratorio y correlacional y se centrará en el análisis de los patrones históricos actuales de seguridad en Bogotá. Se evaluarán las relaciones preliminares entre variables en datos históricos. Además, se diseñará un prototipo

de modelo predictivo básico para identificar posibles correlaciones y predecir tendencias de seguridad permitiendo contrastar la viabilidad de desarrollar un modelo más sólido en el futuro.

Fases de la Investigación

Recolección de Datos Históricos

Recopilación y análisis de fuentes datos públicas de seguridad y literatura disponible de investigaciones de carácter similar.

Objetivo Relacionado

Establecer los métodos, técnicas y algoritmos utilizados en los sistemas de seguridad pública existentes en Bogotá.

Aplicación de Encuesta

Aplicación, recolección y análisis de percepciones actuales de seguridad y calidad de vida en Bogotá.

Objetivo Relacionado

Plantear y evaluar el modelo predictivo desarrollado, mediante la utilización de métricas de desempeño acordes al tipo de modelo y la comparación con algún sistema existente, para validar su efectividad y aplicabilidad en el contexto de Bogotá.

Procesamiento de datos

Procesamiento de datos obtenidos en fuentes de datos disponibles con percepciones actuales proporcionadas por los datos públicos de seguridad de Bogotá para poder ser usados en el modelo.

Objetivo Relacionado

Diseñar y desarrollar un algoritmo de aprendizaje automático capaz de procesar, analizar y predecir zonas seguras e inseguras, considerando la ubicación geoespacial y el contexto local.

Desarrollo del Prototipo del Modelo de Seguridad

Crear un prototipo de modelo predictivo haciendo uso de los datos recopilados y análisis realizados con la mejor tecnología disponible.

Objetivo Relacionado

Diseñar y desarrollar un algoritmo de aprendizaje automático capaz de procesar, analizar y predecir zonas seguras e inseguras, considerando la ubicación geoespacial y el contexto local.

Evaluación y Validación del Modelo Prototipo

Realizar pruebas y mejorar la eficacia del modelo.

Objetivo Relacionado

Plantear y evaluar el modelo predictivo desarrollado, mediante la utilización de métricas de desempeño acordes al tipo de modelo y la comparación con algún sistema existente, para validar su efectividad y aplicabilidad en el contexto de Bogotá.

Correlación de Resultados y Conclusiones

Correlacionar los hallazgos y funcionamiento del prototipo.

Objetivo Relacionado

Plantear y evaluar el modelo predictivo desarrollado, mediante la utilización de métricas de desempeño acordes al tipo de modelo y la comparación con algún sistema existente, para validar su efectividad y aplicabilidad en el contexto de Bogotá.

Población y muestra

La población objetivo estará orientada a residentes actuales de la ciudad de Bogotá mayores de 18 años. Se estimará con base a los datos más recientes del Departamento Administrativo Nacional de Estadística (DANE) alrededor de 8 millones de personas.

La selección de muestra adopta un muestreo no probabilístico, particularmente de la clase por conveniencia, ya que el objetivo esencial de esta investigación está relacionado con la exploración de patrones y percepciones sobre la seguridad en distintos contextos de la ciudad, lo que le confiere pertinencia a una estrategia de este tipo para una investigación exploratoria como la que nos ocupa, en cuya naturaleza se encuentra el interés por identificar tendencias más que por la posibilidad de extrapolar los resultados a la totalidad de la población. El número de entrevistas estimadas es de 112 personas seleccionadas a partir de lugar de residencia y por experiencia directa frente a situaciones de inseguridad.

Si bien el tamaño de la muestra no es relevante con respecto a la población total de Bogotá desde el punto de vista estadístico, el número se justificaría en virtud de su condición para aportar indicadores iniciales sobre patrones de percepción de la seguridad en la ciudad. En cuanto a las unidades de observación, han sido elegidas de acuerdo a su condición de poseer un conocimiento específico o por experiencia relevante sobre la temática de la investigación para favorecer entonces un análisis dirigido y profundo.

El enfoque cualitativo de esta investigación busca identificar dinámicas de la seguridad a partir de una muestra cuya representación desde el punto de vista estadístico no se considera su condición adecuada para el objetivo del estudio. Puesto que se busca explorar y explicar percepciones de seguridad en contextos concretos de la ciudad, la estrategia de muestreo no probabilístico se ajusta a la necesidad de permitir obtener una visión particular y detallada de las experiencias de los participantes. La relación entre el número de la muestra y la estrategia de muestreo no probabilístico se explica en virtud de la identificación de tendencias y patrones en un grupo que tuviera una experiencia relevante con respecto a la problemática de seguridad, lo cual resulta pertinente para este estudio exploratorio.

Instrumentos

La recopilación de datos se realizará a partir de registros oficiales y bases de datos de seguridad pública.

Además, realizara una encuesta sobre la percepción de seguridad de la ciudad. Para mejorar la comprensión de la situación de seguridad en Bogotá en la actualidad, Se buscará obtener una comprensión mejor de lo que la gente espera de un modelo de seguridad e impacto en la ciudad.

Técnicas para el análisis de la información

Ahondando en el detalle del análisis de los datos se investigó y recurrió a la combinación de técnicas descriptivas e inferenciales. Primero, se aplicaron estadísticas descriptivas para comprender la distribución de los incidentes, la variabilidad en las tasas de criminalidad y las

características geoespaciales relacionadas, lo que proporciono un panorama más amplio del comportamiento de los datos.

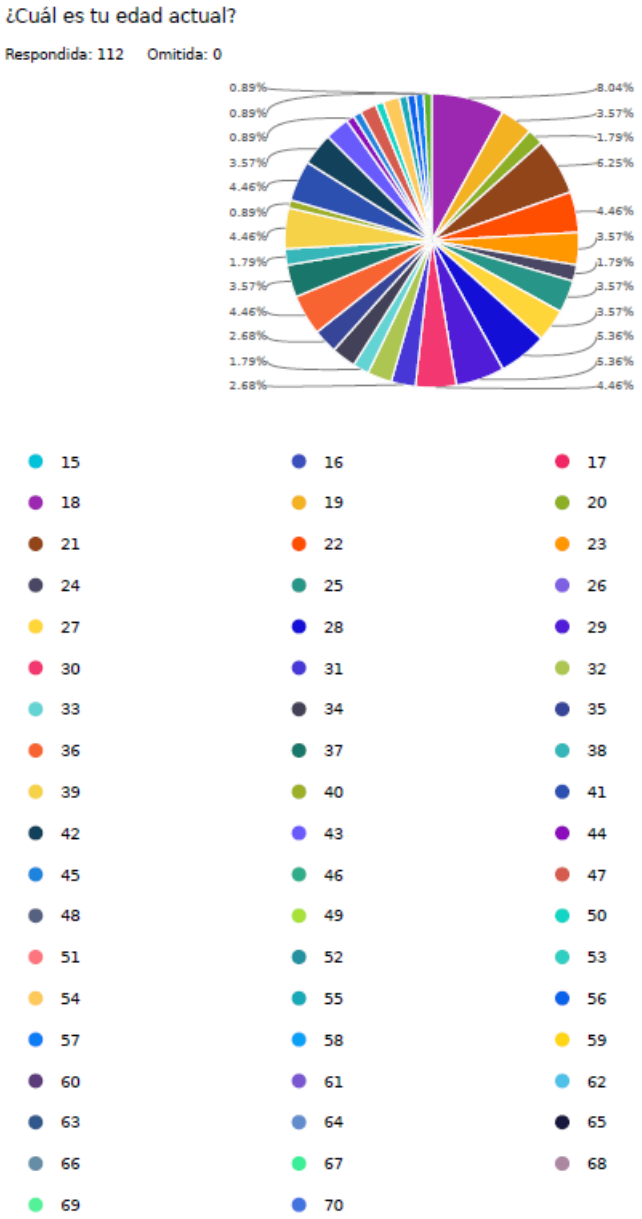
Seguidamente, se realizó la construcción e implementación un modelo de red neuronal recurrente convolucional (CRNN) con el objetivo de predecir los niveles de inseguridad en Bogotá.

El preprocesamiento de los datos incluyó la codificación de variables categóricas, la normalización de variables numéricas y el balanceo de clases para asegurar un entrenamiento más robusto del modelo.

Para la debida y correcta ejecución de lo anteriormente mencionado se recurrió al uso del lenguaje de programación Python y las librerías especializadas tales como pandas, numpy, TensorFlow para el análisis y el modelado, además también se usó Spark para gestionar e integrar grandes volúmenes de datos en formato JSON derivados del datawarehouse construido.

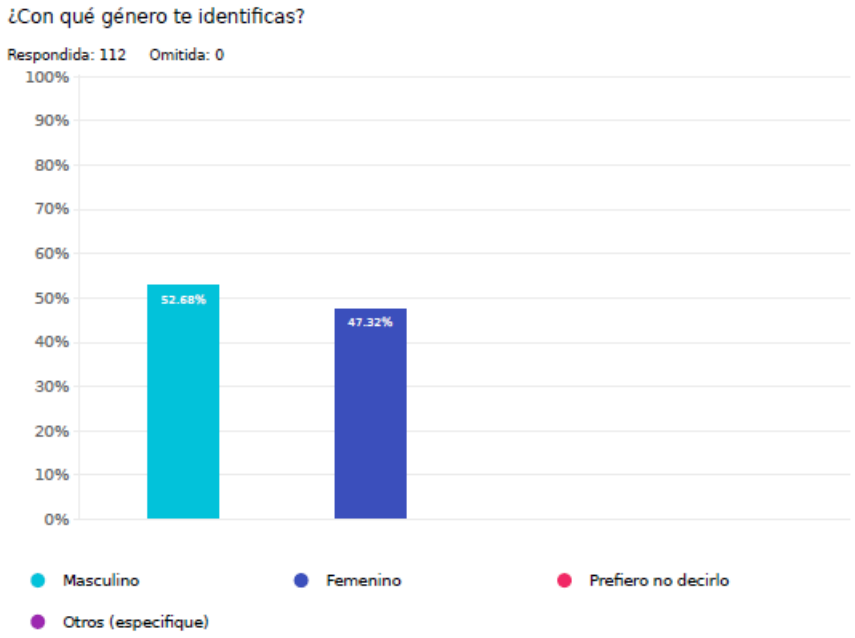
Adicionalmente, se aplicó el instrumento de investigación que se mencionó con anterioridad, la encuesta fue aplicada bajo el título "*Encuesta sobre Percepciones de Seguridad en Bogotá*", que capturó la percepción de 112 personas sobre la seguridad en la ciudad.

Figura 8 Resultados edades encuesta percepción



Fuente: Elaboración Propia

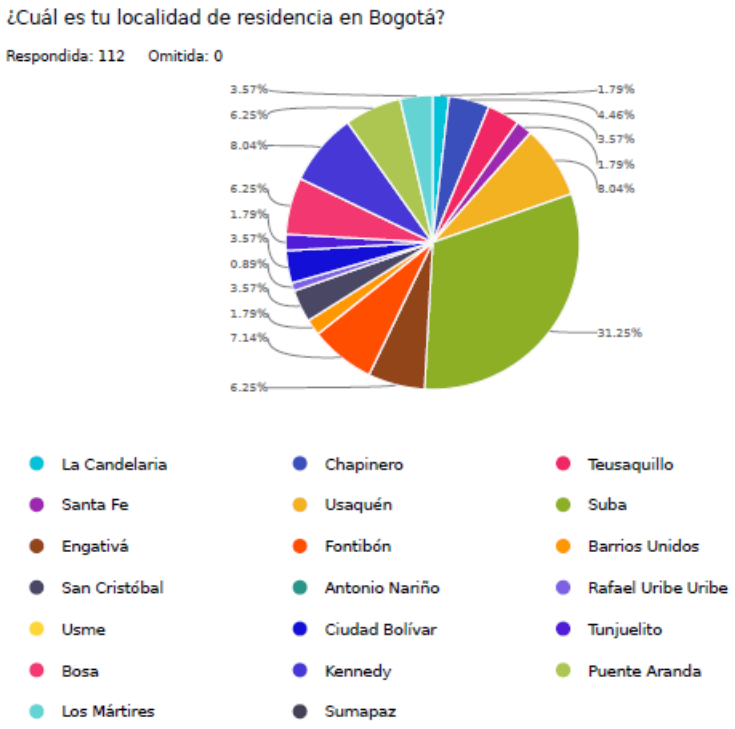
Figura 9. Resultados géneros encuesta percepción



Fuente: Elaboración Propia

En las figuras 8 y Figura 9 relacionadas con preguntas sobre la edad y el género arrojaron una muestra equilibrada, donde se observa un 52.7% de hombres y un 47.3% de mujeres, esto indica que la percepción de inseguridad no se concentra en un grupo específico, sino que es una preocupación común entre los encuestados.

Figura 10. Resultados localidad encuesta percepción



Fuente: *Elaboración Propia*

En la figura 10 sobre el lugar de residencia se exalta que Suba (31.3%) y Kennedy (8.0%) tenían una tasa mayor de representación en las respuestas, lo que indica son localidades que deberían recibir mayor atención a revisar y donde se impulsen esfuerzos por mejorar la seguridad en la ciudad.

Figura 11. Resultados seguridad barrio encuestado- encuesta percepción

Fuente: Elaboración Propia

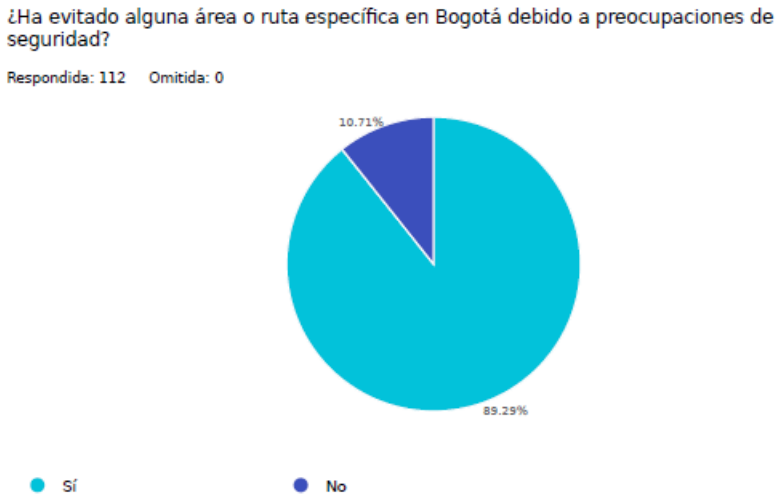
Cuando se pidió a los encuestados calificar la seguridad en sus barrios en la figura 11, se muestran encontrar datos interesantes donde un 47.3% otorgó una calificación de 3 en una escala del 1 al 5, lo que sugiere una percepción de inseguridad moderada en la mayoría de los barrios.

Figura 12. Resultados testigo o víctima de acto delictivo en el último año encuestado-encuesta percepción

Fuente: Elaboración Propia

En contra parte en la figura 12 inmediatamente relacionada con la anterior figura, se evidencia que un 41.9% reportó haber sido testigo o víctima de un delito en el último año, dato que permite inferir acerca de las preocupaciones de los ciudadanos no son solo percepciones infundadas, sino que están sustentadas en experiencias directamente vividas con la criminalidad en la ciudad de Bogotá.

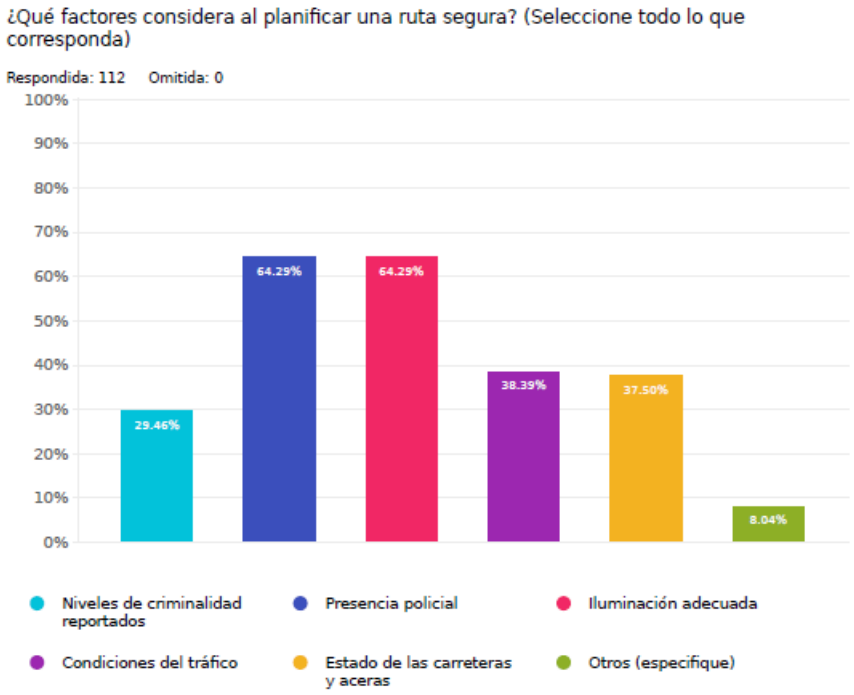
Figura 13. Resultados exclusión de rutas en Bogotá- encuesta percepción



Fuente: Elaboración Propia

La encuesta reveló que el 89.3% de los encuestados había evitado ciertas áreas de la ciudad por razones de seguridad como se observa en la figura 13, que indica acerca del estado actual de percepción de riesgo entre la diferente ciudadanía que tomo esta encuesta, y cómo influye en las decisiones cotidianas de las personas tal como elegir una ruta para llegar un punto a otro en la ciudad.

Figura 14. Resultados factores al planificar una ruta segura - encuesta percepción



Fuente: Elaboración Propia

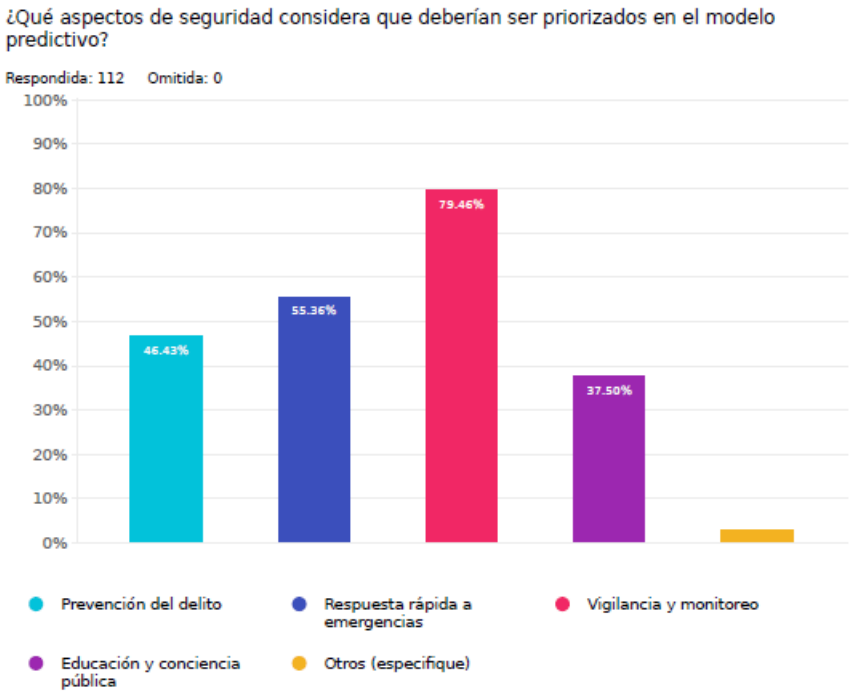
Por otra parte, el 64.3% en esta pregunta relaciono que la presencia policial y la iluminación son factores clave al planificar rutas seguras como se constata en la figura 14, lo que arroja un dato importante que los ciudadanos están considerando dentro de su campo de percepción al desplazarse por la ciudad y como afecta sensación de seguridad dentro de sus criterios.

Figura 15. Resultados modelo predictivo para mejorar la seguridad de la ciudad de Bogotá - encuesta percepción

Fuente: Elaboración Propia

Respecto a sobre si los ciudadanos consideran que el desarrollo de un modelo predictivo podría influir en mejorar la seguridad, el 67.9% de los encuestados consideró que un modelo predictivo podría mejorar la seguridad en Bogotá como se observa en la figura 15.

Figura 16. Resultados aspectos de seguridad que debe priorizar el modelo predictivo para mejorar la seguridad de la ciudad de Bogotá - encuesta percepción



Fuente: Elaboración Propia

Frente a qué aspectos debería considerar el modelo los ciudadanos priorizaron la vigilancia (79.5%) y la respuesta rápida a emergencias (55.4%) como aspectos críticos que el modelo debería abordar de acuerdo con lo consolidado en la figura 16.

Para concluir, se puede denotar que estos resultados refuerzan la idea de desarrollar soluciones con características que consideren un enfoque predictivo cuya meta permita anticipar situaciones de riesgo y optimizar los recursos de seguridad para prevenir el delito en diferentes lugares de la ciudad y permitir reducciones significativas soportadas en la tecnología y los datos. Si desea consultar más información puede revisar **https://survey.zohopublic.com/zs/report/ZhBT4I** con la contraseña EAN2025.

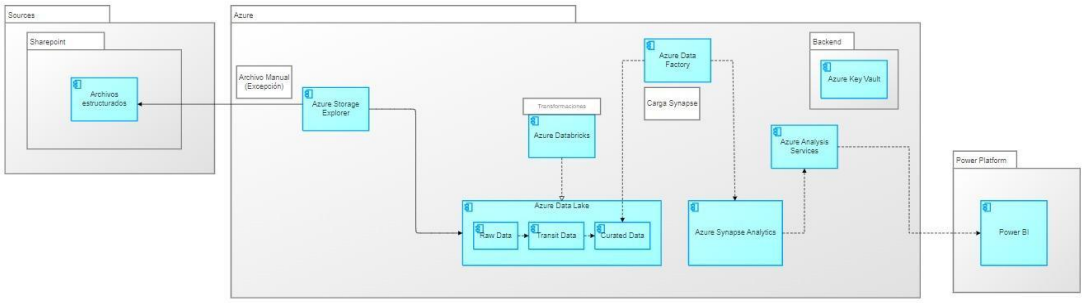
Trabajo de Campo

Para el desarrollo de este proyecto, se llevó a cabo una exhaustiva revisión de la información, enfocada en la recolección de datos de seguridad pública disponibles en Bogotá, principalmente en páginas como la secretaria de seguridad de Bogotá y Datos Abiertos de Bogotá. La recopilación de datos se realizó a partir de fuentes estructuradas disponibles en plataformas del distrito de Bogotá y la Secretaría de Seguridad de Bogotá. Estos datos incluyen reportes de actos delictivos, detalles de los crímenes, ubicaciones geográficas de los eventos y un histórico desde 2020 sobre los incidentes en los diferentes barrios, UPZ y localidades de la ciudad.

Procesamiento de los datos

El procesamiento de los datos recolectados se realizó siguiendo una arquitectura cuidadosamente diseñada, que garantizara la correcta transformación y almacenamiento de la información. La siguiente figura ilustra la arquitectura implementada:

Figura 17. Arquitectura propuesta con lineamientos archimate



Fuente: Elaboración propia

Esta arquitectura se realizó bajo la metodología de arquitectura empresarial TOGAF y con ayuda de la metodología de diagrama Archimate, a continuación, se describirán cada uno de sus componentes:

SharePoint: Fuente principal de donde serán cargados los datos obtenidos de las plataformas oficiales del distrito.

Azure Storage Explorer: Herramienta utilizada para la transferencia de datos desde las fuentes originales hacia el entorno de Azure.

Azure Data Lake: Los datos se almacenaron inicialmente en su forma bruta (raw data) y luego fueron transformados a través de capas intermedias (transit data) hasta alcanzar un estado refinado (curated data).

Azure Databricks: Plataforma utilizada para llevar a cabo las transformaciones necesarias en los datos, aplicando técnicas de limpieza, normalización y enriquecimiento.

Azure Data Factory: Orquestador encargado de gestionar el flujo de datos entre las distintas etapas del procesamiento, asegurando la carga de datos en Azure SQL Server.

Azure SQL Server: Entorno analítico donde los datos refinados fueron cargados para su uso en el modelo.

Azure Analysis Services y Power BI: Herramientas utilizadas para la presentación y visualización de los datos procesados, permitiendo la creación de informes y dashboards interactivos.

A partir de esto se eligieron las fuentes que serían utilizadas para el modelo predictivo y que generaran un mayor valor en las predicciones de acuerdo con esto se generó el siguiente diccionario de los datos elegidos:

DELITO ALTO IMPACTO POR SECTOR CATASTRAL

AIUPZ.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto.

CMIUUPLA: Código de la unidad de planificación.

CMNOMUPLA: Nombre de la unidad de planificación.

Datos Temporales y Conteos:

Desde CMH18CONT hasta CMVI23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMHVAR, CMLPVAR, etc.

Totales: CMHTOTAL, CMLPTOTAL, etc.

Geometría:

SHAPE.AREA: Área de la forma geográfica.

SHAPE.LEN: Longitud de la forma geográfica.

DAILoc.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto.

CMIULOCAL: Código del local.

CMNOMLOCAL: Nombre del local.

Datos Temporales y Conteos:

Desde CMH18CONT hasta CMVI23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMHVAR, CMLPVAR, etc.

Totales: CMHTOTAL, CMLPTOTAL, etc.

Geometría:

SHAPE.AREA: Área de la forma geográfica.

SHAPE.LEN: Longitud de la forma geográfica.

DAISCAT.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto.

CMIUSCAT: Código de la categoría.

CMNOMSCAT: Nombre de la categoría.

Datos Temporales y Conteos:

Desde CMH18CONT hasta CMVI23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMHVAR, CMLPVAR, etc.

Totales: CMHTOTAL, CMLPTOTAL, etc.

Geometría:

SHAPE.AREA: Área de la forma geográfica.

SHAPE.LEN: Longitud de la forma geográfica.

INCIDENTES REPORTADOS POR SECTOR CATASTRAL

IRLoc.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto (int).

CMIULOCAL: Código del local (str).

CMNOMLOCAL: Nombre del local (str).

Datos Temporales y Conteos:

Desde CMR18CONT hasta CMMM23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMRVAR, CMNVAR, CMAOPVAR, CMMMVAR, CMMVAR.

Totales: CMRTOTAL, CMNTOTAL, CMAOPTOTAL, CMMTOTAL, CMMMTOTAL.

Geometría:

SHAPE.AREA: Área de la forma geográfica (float).

SHAPE.LEN: Longitud de la forma geográfica (float).

IRSCAT.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto (int).

CMIUSCAT: Código de categoría (str).

CMNOMSCAT: Nombre de la categoría (str).

Datos Temporales y Conteos:

Desde CMR18CONT hasta CMMM23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMRVAR, CMNVAR, CMAOPVAR, CMMMVAR, CMMVAR.

Totales: CMRTOTAL, CMNTOTAL, CMAOPTOTAL, CMMTOTAL, CMMMTOTAL.

Geometría:

SHAPE.AREA: Área de la forma geográfica (float).

SHAPE.LEN: Longitud de la forma geográfica (float).

IRUPZ.geojson

Identificadores y Ubicación:

OBJECTID: Identificador único del objeto (int).

CMIUUPLA: Código de la unidad de planificación (str).

CMNOMUPLA: Nombre de la unidad de planificación (str).

Datos Temporales y Conteos:

Desde CMR18CONT hasta CMMM23CONT: Conteos o mediciones realizadas entre los años 2018 y 2023.

Variaciones: CMRVAR, CMNVAR, CMAOPVAR, CMMMVAR, CMMVAR.

Totales: CMRTOTAL, CMNTOTAL, CMAOPTOTAL, CMMTOTAL, CMMMTOTAL.

Geometría:

SHAPE.AREA: Área de la forma geográfica (float).

SHAPE.LEN: Longitud de la forma geográfica (float).

DIVIPOLA CENTROS POBLADOS

Código: Representa un código numérico asociado al departamento (int).

Nombre: Nombre del departamento (String).

Código Mun: Representa un segundo código numérico del municipio (int).

Nombre: Nombre del municipio (String).

Código Centro Poblado: Un tercer código numérico con una subdivisión más específica como un distrito o centro poblado (int).

Nombre Centro Poblado: Nombre de la subdivisión, distrito o centro poblado (String).

Tipo: Clasificación del centro poblado, por ejemplo, distrito especial, corregimiento, etc. (String).

Longitud: Longitud geográfica del centro poblado.

Latitud: Latitud geográfica del centro poblado.

HURTO_A_MOTOCICLETAS

ARMAS_MEDIOS: Tipo de arma o medio utilizado en el hurto de motocicletas (object).

DEPARTAMENTO: Nombre del departamento donde ocurrió el incidente (object).

MUNICIPIO: Nombre del municipio donde ocurrió el incidente (object).

FECHA: Fecha en que se reportó el hurto de la motocicleta (datetime64[ns]).

CODIGO_DANE: Código DANE asociado al municipio donde ocurrió el hurto (float64).

CANTIDAD: Número de hurtos reportados en la entrada correspondiente (float64).

HURTO_A_PERSONAS

ARMA MEDIO: Tipo de arma o medio utilizado en el hurto (object).

DEPARTAMENTO: Departamento donde ocurrió el hurto (object).

MUNICIPIO: Municipio donde ocurrió el hurto (object).

FECHA HECHO: Fecha en que se reportó el hurto (datetime64[ns]).

GENERO: Género de la víctima (object).

AGRUPA_EDAD_PERSONA: Grupo de edad de la víctima (object).

CODIGO DANE: Código DANE asociado al municipio donde ocurrió el hurto (float64).

CANTIDAD: Cantidad de hurtos reportados (float64).

HURTO_A_AUTOMOTOR

ARMAS_MEDIOS: Tipo de arma o medio utilizado en el hurto de automotores (object).

DEPARTAMENTO: Departamento donde ocurrió el hurto de automotores (object).

MUNICIPIO: Municipio donde ocurrió el hurto de automotores (object).

FECHA: Fecha en que se reportó el hurto de automotores (datetime64[ns]).

CODIGO_DANE: Código DANE asociado al municipio donde ocurrió el hurto (float64).

CANTIDAD: Cantidad de hurtos reportados en la entrada correspondiente (float64).

Clasificación de Datos

Datos de Incidentes

Descripción: Datos sobre incidentes específicos como hurto a motocicletas, hurto a personas y hurto de automotores.

Ejemplo de Datos: Tipo de arma, fecha del incidente, cantidad de hurtos.

Datos Geográficos

Descripción: Datos sobre ubicaciones geográficas específicas, incluidos nombres, áreas y longitudes de diferentes divisiones.

Ejemplo de Datos: Código de unidad de planificación, nombre de local, área de la forma geográfica.

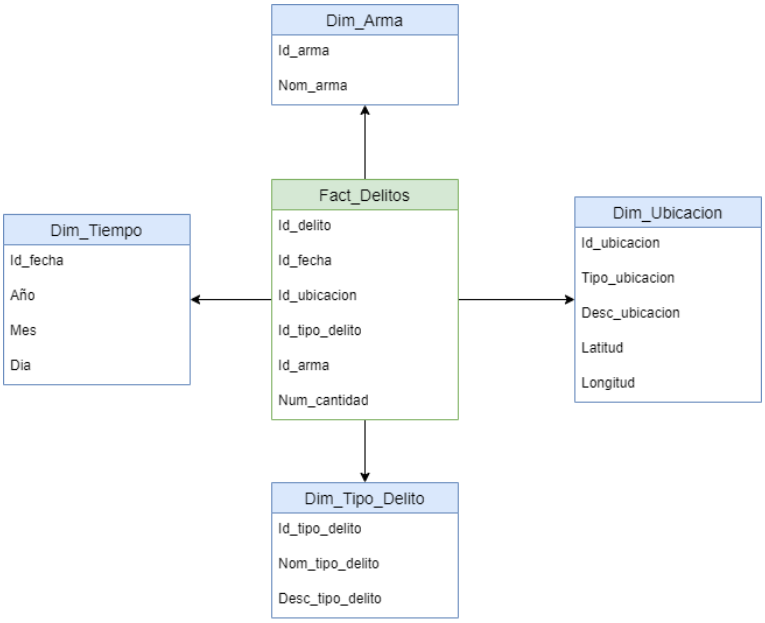
Ya con una identificación de los datos crudos recolectados de las diferentes plataformas y buscando las mejores prácticas, se realizó, de acuerdo con la necesidad del modelo, en primera instancia el modelo conceptual que se utilizaría en la implementación. Esto se hizo con el fin de tener claras las entidades a utilizar y su concepto.

Figura 18. Modelo de datos conceptual

Fuente: Elaboración propia

Ya teniendo claras las diferentes entidades, se identificaron los atributos de acuerdo con el diccionario de datos previamente expuesto, para verificar si teníamos toda la información necesaria y cuáles campos aportarían valor al modelo.

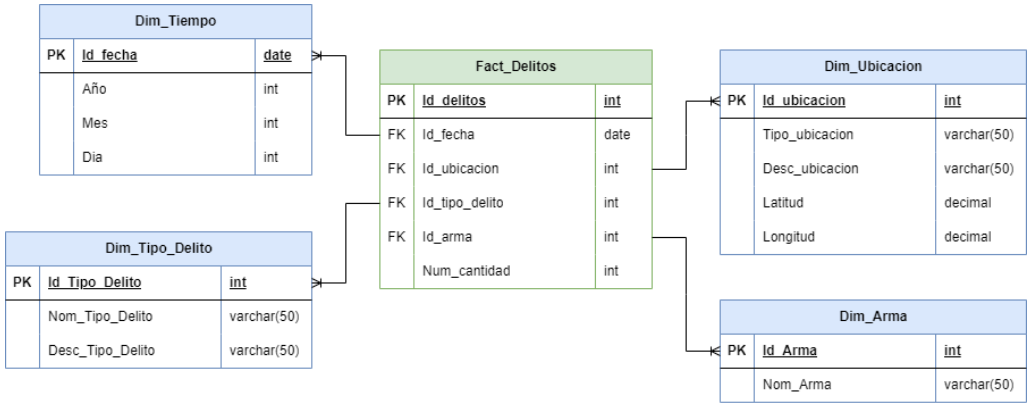
Figura 19. Modelo de datos Lógico



Fuente: Elaboración propia

Para finalizar, ya con los diferentes atributos identificados, se determinaron los tipos de datos y las claves que se utilizarían en la base de datos, dando lugar al modelo físico expuesto a continuación:

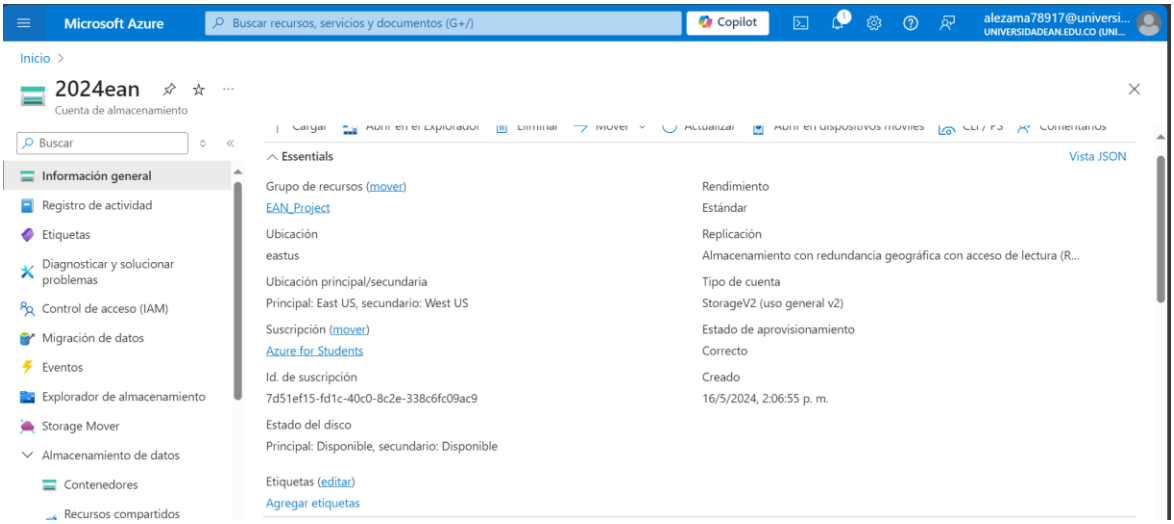
Figura 20. Modelo de datos Físico



Fuente: Elaboración propia

A partir de la definición de la estructura y herramientas expuestas anteriormente, se dio paso al desarrollo y tratamiento de los datos, por medio de las herramientas de Azure Cloud. Lo primero que se realizo fue la puesta a disposición de los recursos necesarios para la implementación en la plataforma, donde se creó la cuenta de almacenamiento 2024ean, de tipo StorageV2 estándar.

Figura 21. Blob Storage en Portal Azure



Fuente*: https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e 338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024 ean/overview*

Aquí, se realizó la creación de cada uno de los contenedores necesarios para el desarrollo, donde se encuentran las siguientes capas:

Capa Flatfile: Cargue manual de archivos, en caso de que no sea posible una integración directa con las fuentes oficiales, o para este caso se encuentra fuera del alcance del proyecto realizar dicha integración bien sea por API o por RPA-

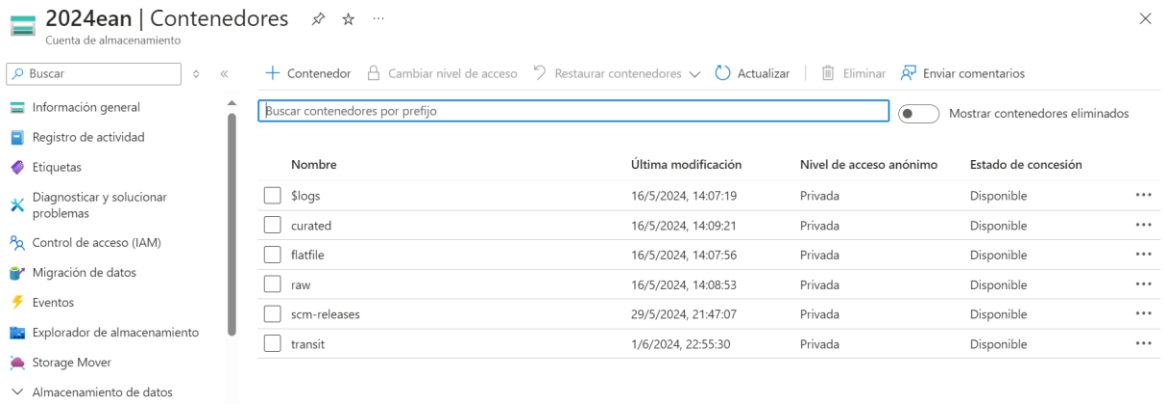
Capa Raw Data: Se realiza el cargue o traspaso de la información en formato csv.

Capa Transit Data: Se busca hacer una limpieza de los datos con las reglas de calidad necesarias y reglas de negocio que sean necesarias aplicar para que la información quede limpia.

Capa Curated Data: En esta capa se estructura la información como se evidencia en el modelo físico anteriormente expuesto.

Base de datos: se realiza el cargue de los datos curados en la base de datos, por esto es importante que los datos curados cuenten con la estructura y limpiezas correctas hasta este punto.

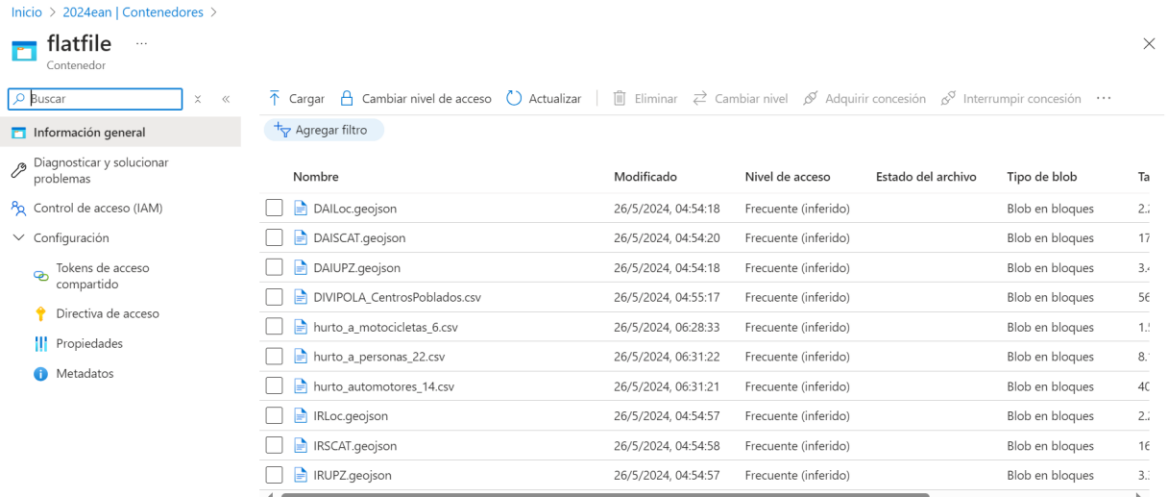
Figura 22. Contenedores Azure



Fuente: *https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024ean/overview*

En la capa flatfile se realizó el cargue de los archivos seleccionados que aportan valor al modelo y son relevantes para un correcto entendimiento del comportamiento actual de los delitos en la ciudad de Bogotá.

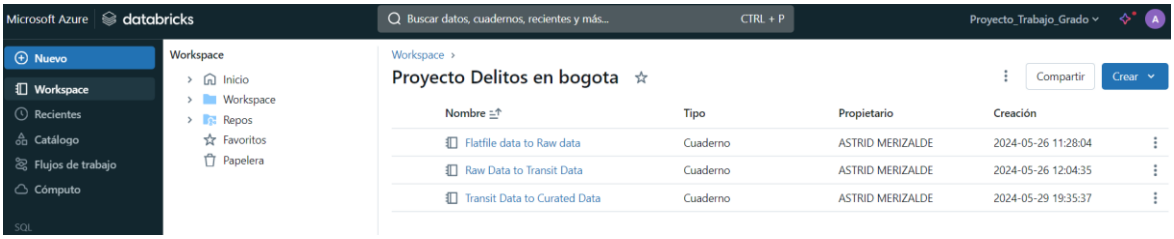
Figura 23. Archivos cargados en el flatfile



Fuente: *https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024ean/containersList*

Siguiendo la arquitectura se realizó el traspaso y transformación entre capas con la herramienta Azure Databricks, donde se realizó un Notebook para cada una de las capas mencionadas anteriormente y siguiendo las buenas prácticas y lineamientos correspondientes, previamente definidos.

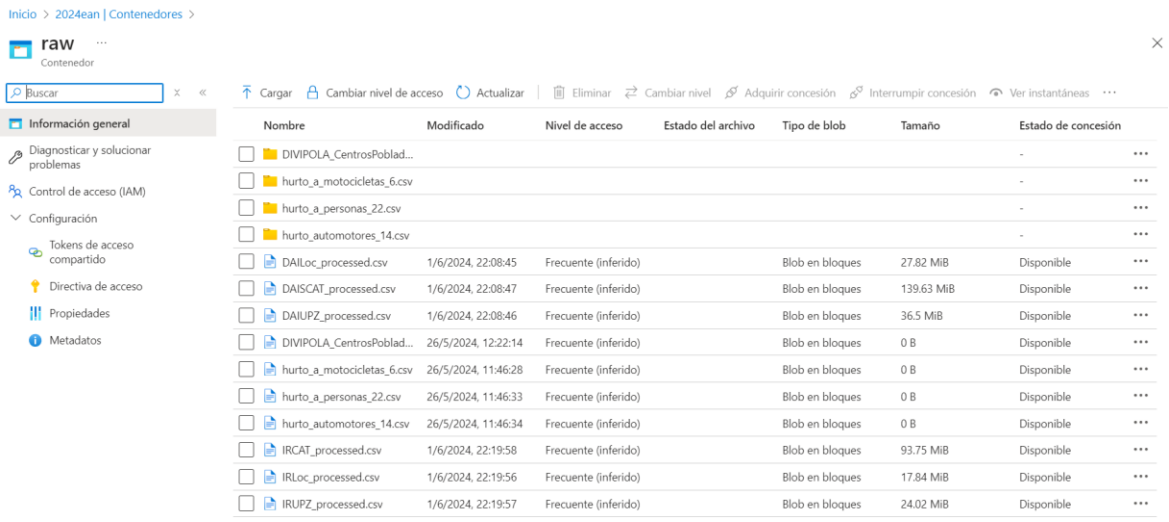
Figura 24. Notebooks creados en Databricks



Fuente: https://adb-83873127444379.19.azuredatabricks.net

A partir de la ejecución de los notebooks se genera los archivos csv es sus capas correspondientes, estos archivos son generados con pyspark y generado en particiones teniendo en cuenta la escalabilidad futura del proyecto, que puede llegar a contener grandes volúmenes de datos.

Figura 25. Archivos de la capa Raw Data

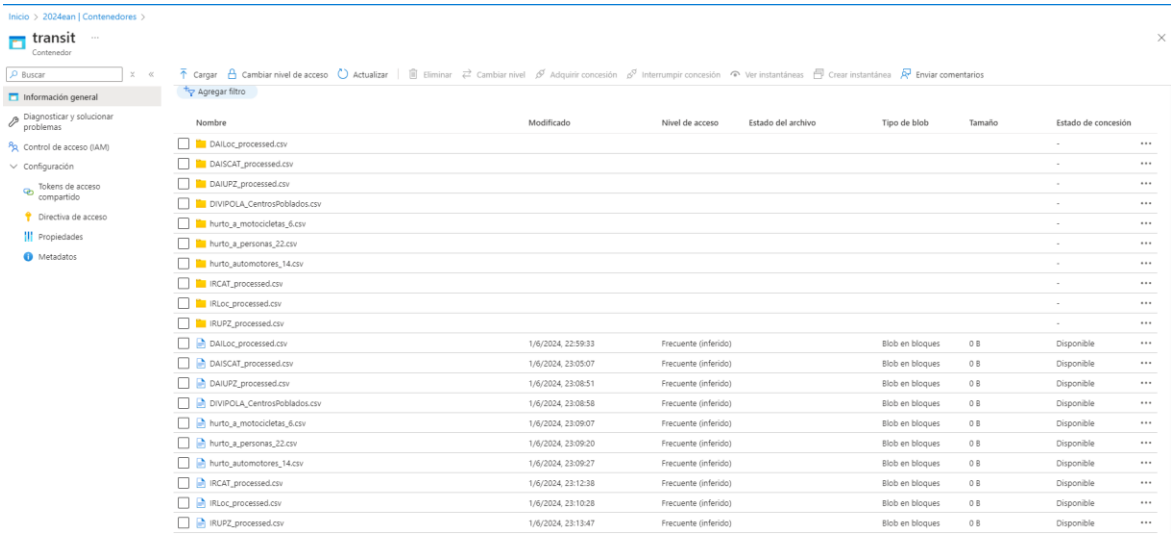


Fuente: https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-

8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024ean/overview

En la Figura 25, se observa que los archivos de la capa raw generado por el script en Databricks "*Flatfile Data to Raw Data*", donde se hace el traspaso de la información y se procedió a que todos los archivos estuvieran en formato CSV.

Figura 26. Archivos de la capa Transit Data

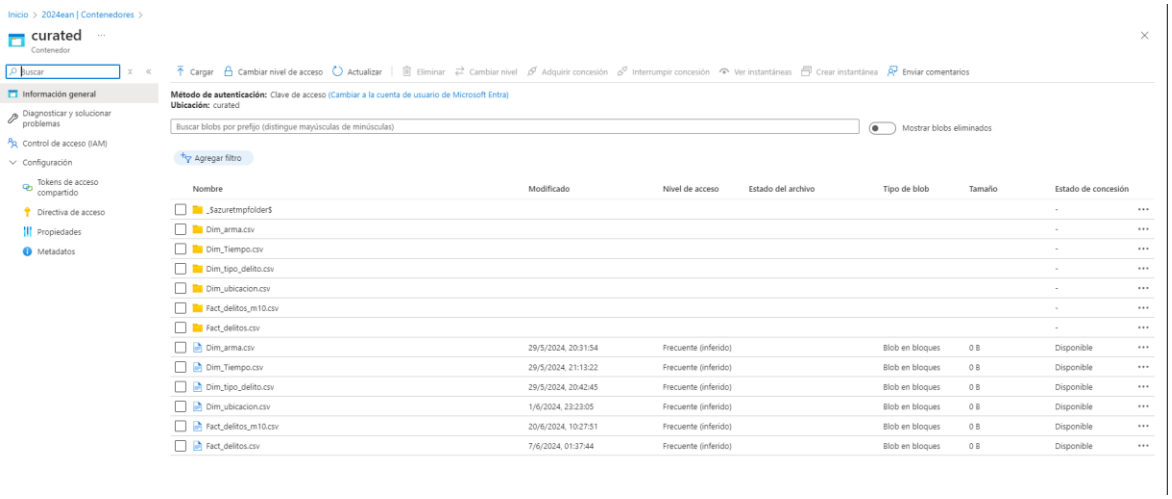


Fuente: *https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024ean/overview*

Después, se tomaron los archivos de la capa Raw, para hacer limpiezas de tildes, caracteres extraños, dar un formato tabular a los archivos y trasformar los datos de acuerdo con las

necesidades del modelo. Los archivos resultantes se guardan en la capa Transita data como se puede observar en la Figura 26.

Figura 27. Archivos de la capa Curated Data

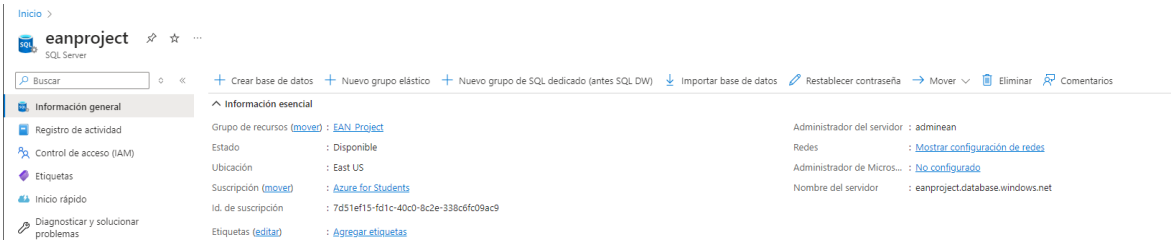


Fuente: https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Storage/storageAccounts/2024ean/overview

En la Figura 27 se observan los archivos de la capa Curated, los cuales fueron generados a partir de los datos de la capa Transit y se realizó la estructuración de estos datos, para que concordara con el modelo de datos en estrella propuesto en la Figura 11.

Ya teniendo los datos totalmente transformados, se creó una base de datos de tipo Azure SQL con un servidor SQL Server como se puede ver en la Figura 19 y Figura 20, debido a los costos se creó con un límite de almacenamiento.

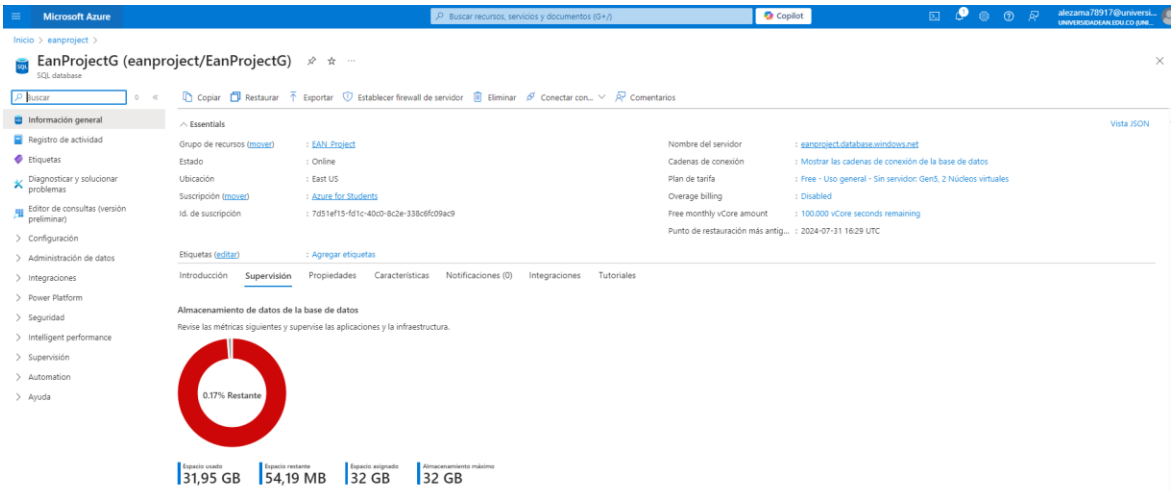
Figura 28. Servidor de la base de datos



Fuente:

https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Sql/servers/eanproject/databases/EanProjectG/overview

Figura 29. Propiedades de la base de datos.

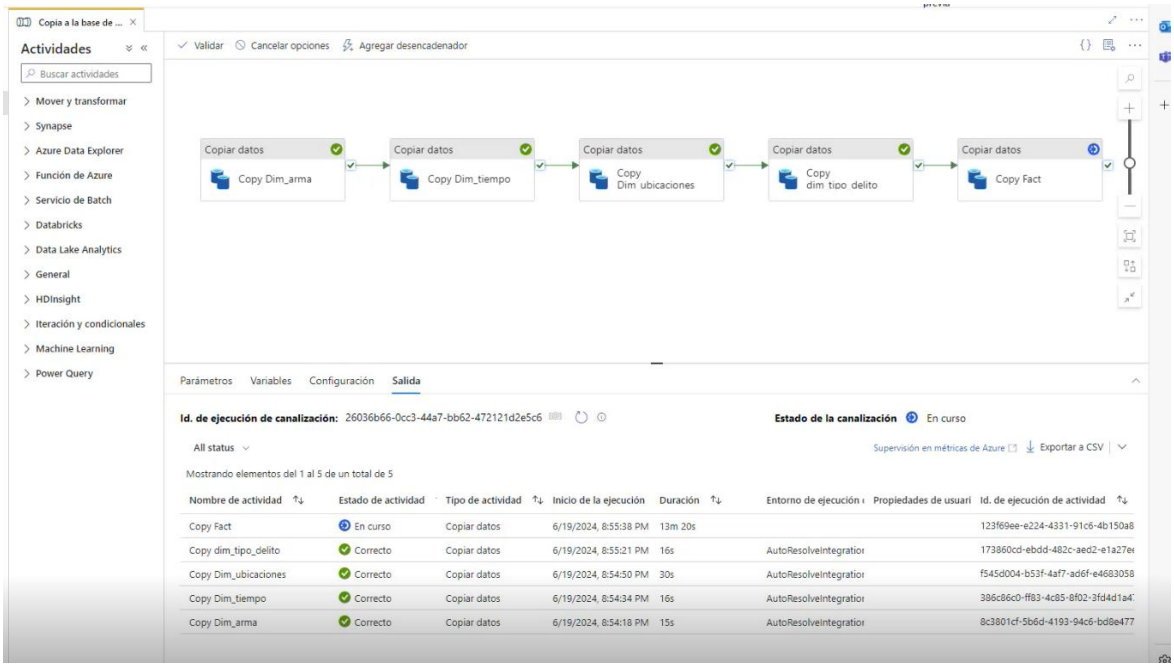


Fuente*:* https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e-

338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.Sql/servers/eanproject/databases/EanProjectG/overview

A partir de la creación de la base de datos, se generaron los DDL de las tablas de dimensión y de hecho de acuerdo con el modelo presentado en la Figura 29, al ser ejecutadas se crearon las tablas en la base de datos. Y posterior mediante un orquestador y una actividad de copia se cargan los datos de la capa curated data a las tablas correspondientes.

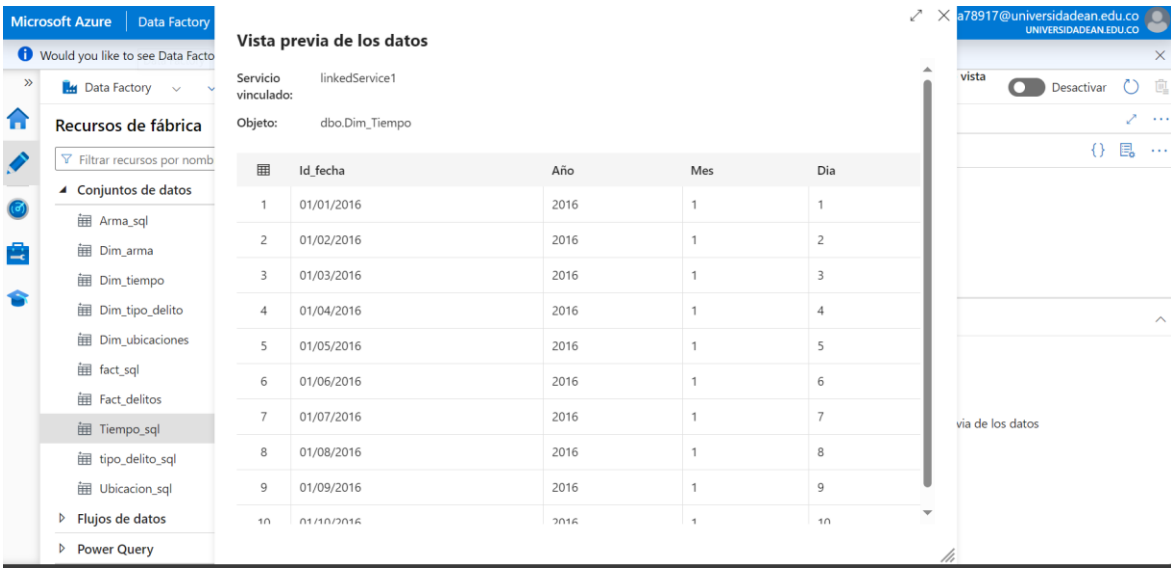
Figura 30. Azure data Factory Orquestador de copia a la Base de datos.



Fuente*:* https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.DataFactory/factories/ProyectoGrado2024Ean/overview

Al realizar la ejecución y verificación de la carga de la información en la base de datos se encuentra que ha sido exitosa como se muestra en la Figura 30.

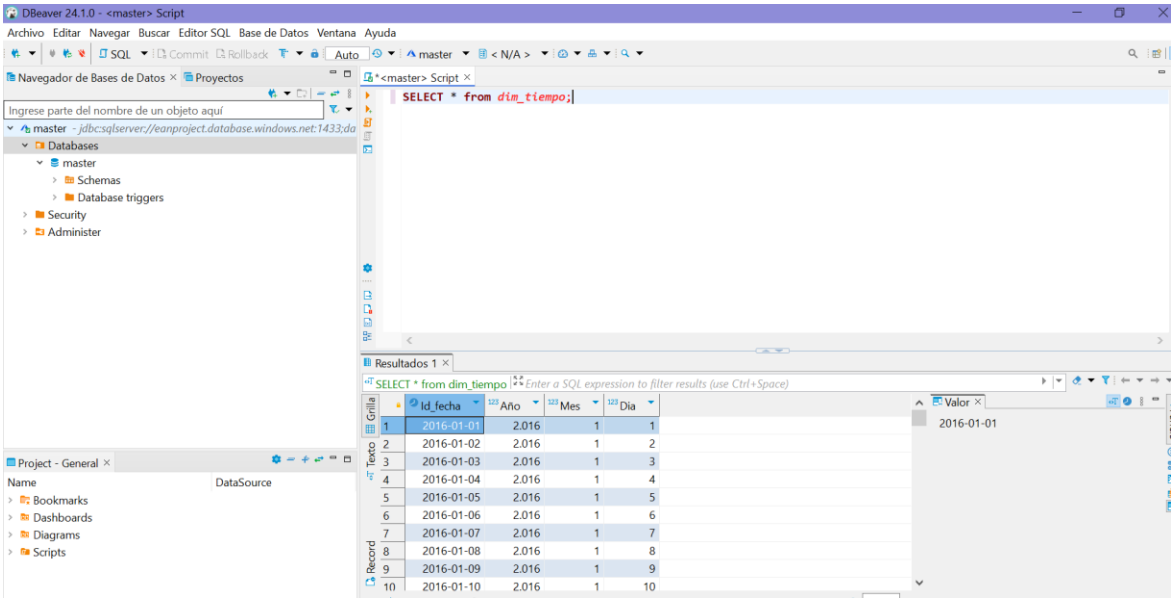
Figura 31. Vista previa de los datos de Dim_Tiempo desde Azure



Fuente: *https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e-38c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.DataFactory/factories/ProyectoGrado2024Ean/overview*

En la Figura 31 se puede ver también directamente ya cargada en la base de datos.

Figura 32. Vista de la Dim_Tiempo en la base de datos



Fuente: https://portal.azure.com/#@universidadean.edu.co/resource/subscriptions/7d51ef15-fd1c-40c0-8c2e338c6fc09ac9/resourceGroups/EAN_Project/providers/Microsoft.DataFactory/factories/ProyectoGrado2024Ean/overview

De acuerdo, con lo anterior se llevó a cabo una técnica de balanceo que consiste en un sobremuestreo con el fin de contrarrestar el desbalanceo que existe entre las frecuencias de las clases de incidentes de seguridad. En este caso, el sobremuestreo consiste en replicar las muestras obtenidas de las clases minoritarias hasta llegar a la misma frecuencia que las de la clase mayoritaria, lo que evita que el modelo esté sesgado hacia la clase mayoritaria. Este proceso se llevó a cabo utilizando la función resample de la librería scikit-learn a partir de las clases de inseguridad, mediante las cuales se replicaron hasta equilibrar su representación en el dataset. Una vez terminado este balanceo, se aplicó la función shuffle para mezclar

aleatoriamente las muestras balanceadas, de forma que el modelo a aprender no sea capaz de aprender patrones muy artificiales que distorsionan la realidad.

Ventajas de efectuar el sobremuestreo en el presente estudio

Reducción del sesgo: Al llevar a cabo la igualación entre las diferentes clases de número de muestras, se favorece la capacidad de detección, la habilidad del modelo para detectar patrones entre las clases de inseguridad (no solo con aquella de mayor tamaño). **El rendimiento en los incidentes de clases de baja frecuencia:** El modelo adquiere mayor precisión clasificadora en aquellas categorías de menos reportes existenciales, lo que constituye una ventaja no despreciable en la detección de las zonas con bajo número de reportes y criminalidad.

Mantenimiento de los patrones de clase original: El sobremuestreo simple mantiene el patrón original en su totalidad, a diferencia de los métodos más sofisticados, dado que permite una comprensión más clara sobre la magnitud de los números de incidentes, lo que se convierte en un argumento adicional a favor de su uso en el análisis de la seguridad urbana.

Comparaciones con el SMOTE

Mientras que el SMOTE (Synthetic Minority Oversampling Technique) es una técnica más avanzada que zarpa hacia la generación de muestras sintéticas para las clases minoritarias. En el presente estudio no se ha decidido emplear el SMOTE por múltiples razones:

Datos numéricos y categóricos combinados

SMOTE es más eficaz con datos numéricos.

En este estudio, la unión de variables categóricas (como tipos de delitos o localizaciones) puede dar lugar a datos irreales si se utiliza SMOTE, empeorando así la precisión del modelo.

Complejidad computacional

SMOTE necesita de más recursos computacionales y tiempo de procesamiento, lo que puede haber sido una limitación en el análisis de grandes cantidades de datos relativos a la seguridad urbana en Bogotá.

El sobremuestreo simple respeta la distribución real de los datos y se presta a un posicionamiento en un contexto práctico.

Impacto del sobremuestreo en la seguridad urbana

Este enfoque garantiza que todas las clases de incidentes estén equilibradas, mejorando la capacidad del sistema en la identificación de patrones de criminalidad en toda la ciudad, incluso en aquellas zonas con menor incidencia reportada. Se contribuye de esta manera a una mayor calidad y equidad de la toma de decisiones por parte de las autoridades locales y se mejora la distribución de los recursos y la confianza de la ciudadanía en los sistemas de inteligencia artificial aplicada a la seguridad.

Resumiendo, el uso del sobremuestreo asegura la validez de los resultados y que resulten representativos, ya que alinea la técnica adoptada con los objetivos del estudio: mejorar la seguridad urbana en Bogotá mediante modelos predicativos precisos y éticos.

El sobremuestreo destaca como la mejor opción para este trabajo frente al resto de opciones como submuestreo o SMOTE, por las ventajas que le proporciona el contexto de los datos analizados. Frente al submuestreo, que realiza una reducción de la cantidad de datos de la clase mayoritaria y puede atentar incluso a la ausencia de información considerada la más importante, el sobremuestreo respeta toda la información que nos proporcionan los datos, permitiendo que el modelo trabaje con la totalidad de los datos. Por su parte, SMOTE (Synthetic Minority Oversampling Technique), aunque es un método aceptado para datos numéricos, se haría complejo en este caso debido a la combinación de datos numéricos y categóricos, trabajando con datos generados en el contexto de esta combinación, probablemente produciendo datos no reales que pueden atentar contra la complejidad a la hora de producir el modelo.

Además, SMOTE es un método computacionalmente más complejo y que consume más tiempo en cuanto a la producción de datos, siendo una limitación en cuanto a la analítica de grandes volúmenes de datos como es el caso dado en este trabajo. El sobremuestreo simple, por el contrario, respeta la estructura de los datos y la distribución obtenida, respetando que las relaciones obtenidas entre variables fueran no solo reales, sino que también fuesen el reflejo directo del lugar que se intenta solo atentar. Estas características dan como resultado que el sobremuestreo sea la opción más directa para este trabajo, puesto que se atiene a los objetivos que se quiere dar, como se daba al sujeto a analizar, no solo mostrar patrones de inseguridad en Bogotá, sino también a su vez intentar replicar el modelo en la práctica.

A fin de comprobar la generalización del modelo, se analizaron diferentes métricas, tales como la precisión y la pérdida, en los conjuntos de validación y de prueba, lo que permitió comprobar que el desempeño del modelo no esté acotado solamente al conjunto de entrenamiento, sino que se mantenga en un nivel acorde a lo lejos de los datos no vistos y así evitar el sobreajuste. Con respecto a las variables más importantes, se podría haber utilizado Grad-CAM para mostrar aquellas características que más impactan en las predicciones en términos de localización o tipo de delito, reforzando así la interpretabilidad del modelo, dado que las CNN procesan de manera implícita las entradas. Finalmente, la más que necesaria relación de las entradas con la salida fue garantizada por el hecho de haber mantenido la estructura original de los datos gracias al sobremuestreo, lo cual garantizaba que variables relativamente obvias, tales como las frecuencias de los incidentes o las categorías de inseguridad, tuvieran una relación con la predicción de los niveles de inseguridad urbana.

Análisis de resultados

Los resultados a analizar son aquellos generados por el prototipo de la red neuronal convolucional recurrente (una CRNN por sus siglas en inglés) modelada y entrenada. Esta red neuronal se diferencia de otros tipos de redes neuronales al contar con capas convolucionales (en este caso, las capas Conv1D) y capas recurrentes (las capas LSTM), las primeras estando encargadas de los datos espaciales y la segunda de los datos temporales.

Dependiendo del resultado, el modelo calificaría las zonas de Bogotá, a través de una función softmax, en cuatro niveles de inseguridad: “Muy seguro”, “seguro”, “inseguro” y “muy inseguro”. Esto permite traducir números y variables a conceptos fáciles de entender,

especialmente en situaciones donde las necesidades de información clara e instrucciones precisas son necesarias, como lo puede ser en el riesgo de ser víctima de un delito.

Los datos por utilizar, primeramente, pasan por un proceso de balanceo. Durante este proceso los datos menos utilizados son mostrados más veces, con el fin de que sean proporcionales (en cantidad) a los más utilizados. Para este caso en particular de implementación, los beneficios de esta técnica sobrepasan la disminución en la rigurosidad de los datos.

La obtención, transformación y cargado de datos puede ser explicado a través del proceso ETL (por sus siglas en inglés: *Extraction*, *Transformation* y *Load*), un paso fundamental para hacer que los datos de los delitos sean aptos para el análisis y el modelo pueda aprovecharlos eficazmente. En este proyecto se utilizó específicamente PySpark para el manejo de los volúmenes de datos con los que se entrenaría la red neuronal necesario en la consolidación del DWH creado previamente con la información histórica.

Los distintos indicadores (como puede ser el arma del delito, el día y la hora, la localización geográfica, el tipo de delito, etc.) y sus respectivos datos, luego de que son extraídos, son guardados paulatinamente en archivo JSON. Estos indicadores y los respectivos datos fueron depositados en un formato específico y eficiente para el consumo de la red neuronal en una tabla a través del uso de SQL. Una vez los datos son transformados (es decir, pasaron de ser información sin un formato a algo que representa la misma data, pero en un formato que la red neuronal puede entender) se cargan en un formato CSV y se puede continuar con el proceso de entrenamiento y evaluación.

Al terminar el proceso de ETL y posteriormente utilizar dichos datos, el modelo alcanzó una precisión del 97%, lo que indica que fue capaz de identificar correctamente los patrones en los

datos de seguridad pública. Las métricas utilizadas para evaluar el desempeño del modelo fueron la precisión, que mide la proporción de predicciones correctas, y la pérdida, que mide el error durante el entrenamiento y la validación. Este proceso de entrenamiento y validación tiende a mejorar con el tiempo, acercándose cada vez más a la realidad que dictan los patrones de los datos proporcionados durante su periodo de entrenamiento.

De acuerdo con lo anterior, es importante plantearse algunas preguntas claves sobre el rendimiento del modelo generalizado y su capacidad para sobreajustar el conjunto de entrenamiento.

Evaluación generalización del modelo

Si bien el modelo alcanza una precisión impresionante del 97% en el conjunto de entrenamiento, es importante saber cómo es la generalización del modelo. Las CNN tienen un aprendizaje aprendible pero complejo que puede dar lugar a sobreajuste (overfitting), es decir, el modelo puede únicamente estar memorizando los datos de entrada de entrenamiento sin aprendizaje de patrones generalizables. Para conocer la respuesta a esta pregunta se han de tener en cuenta algunas métricas (valores) y para ello se propone que las precisiones y las pérdidas se tomen también en cuenta en los conjuntos de validación y test, lo que supondrá un riesgo de la información que trascenderá su poder total discriminatorio con respecto al desempeño del modelo.

Precisión y pérdida en el conjunto de validación

La precisión y la pérdida en el conjunto de validación van a permitir saber si el modelo está comenzando a partir el ajuste (overfit) que hace el modelo de los datos de entrenamiento.

Si la precisión en el conjunto de validación es inferior que en el entrenamiento o si la pérdida en el conjunto de validación es superior, esto sugiere que el modelo no es capaz de generalizar bien en su desempeño.

Precisión y pérdida en el conjunto de prueba

Las métricas en el conjunto de prueba son imprescindibles para una evaluación final. La precisión en el conjunto de test refleja cómo el modelo hace sus predicciones para datos totalmente nuevos y la pérdida en el conjunto refleja cómo es el modelo en general.

Presentar estas métricas contribuirá a abordar si el modelo está mostrando ese buen rendimiento porque realmente es capaz de generalizar o si, por el contrario, se está ajustando a los datos del conjunto de entrenamiento y eso podría comprometer su aplicabilidad y rendimiento en un entorno de trabajo real.

Con el objetivo de evaluar la eficacia del modelo predictivo implementado, se realizó un análisis comparativo donde se aplicaron las métricas más significativas de desempeño para los modelos de machine learning, orientados hacia el mismo problema de la seguridad urbana. En primer lugar, se usaron las métricas de precisión y pérdida de datos tanto en el entrenamiento como en la validación o en las pruebas. De la misma forma, la precisión alcanzada en los sets de validación y pruebas se utilizó para comprobar hasta qué punto se estaba aprendiendo el modelo en base a ejemplos no vistos en los procesos de entrenamiento, de acuerdo con las especificaciones del modelo, evitando el sobreajuste de datos. Junto a esto, la pérdida se

utilizó para evaluar hasta qué punto el modelo estaba ajustando bien la predicción de las salidas a partir de los datos de entrada.

En lo que respecta a la comparación con sistemas existentes, se utilizaron como referencia aplicaciones como Waze, que, a pesar de no estar enfocadas en la seguridad, permiten comprobar la efectividad de su enfoque en lo que respecta a la gestión de rutas, aportando datos sobre tráfico y accidentes. Sin embargo, el sistema implementado fue sustanciado en su capacidad para incorporar la toma de decisión como modo de integrar datos históricos de criminalidad y realizar predicciones sobre zonas de alto riesgo, pasando a ser una herramienta predicativa en lugar de reactiva.

La efectividad del modelo fue objeto de la validación de la capacidad predictiva que este generaba en comparación con la realidad actualizada de los delitos en la ciudad, permitiendo comprobar si las alertas básicas generadas por el modelo coinciden con las zonas de alta incidencia delictiva, tal y como se documentan oficialmente. La comparación fue complementada con pruebas de la sensibilidad y especificidad, para evaluar el balance de falsos positivos y negativos del modelo y así asegurar la validez real del mismo para las situaciones de desplazamiento urbano.

Las variables más importantes del modelo

Si bien las CNN no disponen de "variables" de entrada al uso, es importante proporcionar un análisis de la interpretabilidad del modelo. Para esto se pueden emplear herramientas como Grad-CAM o mapas de activación para visualizar las características más relevantes que el modelo hace usadas para obtener resultados. Estas técnicas permiten identificar patrones en

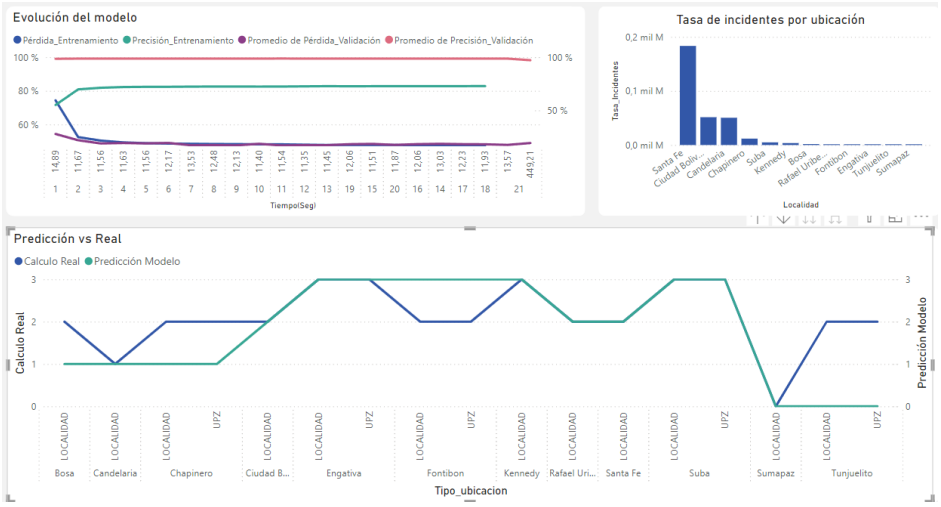
los datos que son importantes para la obtención de predicciones, mejoran la transparencia del modelo y a su vez permiten comprobar si el modelo está aprendiendo lo que se espera de él, es decir, que sea capaz de realizar una clasificación.

Relación entre las entradas y la salida

Las redes neuronales generalmente inferirán de manera implícita las relaciones entre las variables de entrada, pero resulta interesante comprobar que algunas de las características del conjunto de datos tengan relación directa/indirecta con la variable de salida (por ej. el grado de inseguridad). Para esto, la metodología SHAP (SHapley Additive exPlanations) puede proporcionar información sobre hasta qué punto las características inciden en la predicción del modelo. Esto permitiría, en última instancia, asegurar que las decisiones del modelo sean interpretables, evitando que se aprendan correlaciones espurias.

En la Figura 33 se observan los resultados de un entrenamiento previo; Los datos fueron normalizados para reflejar la tasa de incidentes por cada 100.000 habitantes, lo que permitió realizar comparaciones justas entre diferentes zonas de la ciudad. Las zonas con tasas superiores a 200 incidentes por cada 100 mil habitantes fueron clasificadas como “muy inseguras”, mientras que aquellas con tasas bajas se consideran “muy seguras”. Además, se emplearon técnicas de regularización como el “Early Stopping” para evitar que el modelo se acostumbrara demasiado a los datos de entrenamiento y perdiera su capacidad de generalización.

Figura 33. Resultados del entrenamiento.



Fuente: Power BI Dashboard Desempeño. Elaboración Propia

Los resultados gráficos consolidados generados (Figura 34) por el modelo predictivo indican situaciones y realidades interesantes a analizar a la par que se contrasta tanto con el conocimiento general y sociocultural de las zonas de las cuales proviene la data utilizada, al igual que se contrasta con los datos publicados por las distintas autoridades entendiendo que se hace uso de una muestra de la totalidad de la bodega de datos.

Figura 34. Resultados gráficos consolidados

Fuente: Power BI Dashboard Desempeño. Elaboración Propia

Al analizar el panel de información que se genera al volcar todos los datos en el sistema, los análisis que se pueden realizar son varios. Primeramente, es importante entender la evolución por la que ha pasado el modelo, lo que indica la gráfica de líneas de la esquina superior izquierda y la relación que este tiene entre el conjunto de datos de entrenamiento y el conjunto de datos de validación.

Se ve, en orden, una disminución en ambas pérdidas (Tabla 1), tanto en la fase de entrenamiento como en la de validación. Viendo paralelamente un aumento en las tasas de precisión tanto del conjunto de entrenamiento como en el conjunto de precisión. La disminución de las pérdidas indica que la similitud entre la información generada por el sistema inteligente y

la información real aumenta; esto, inevitablemente, genera un aumento en la precisión tal y como está reflejado, puesto que los datos generados son cada vez más parecidos a los que se esperaría de una información obtenida de un entorno real.

Tabla 1 Resultados del entrenamiento (primeras 5 épocas)

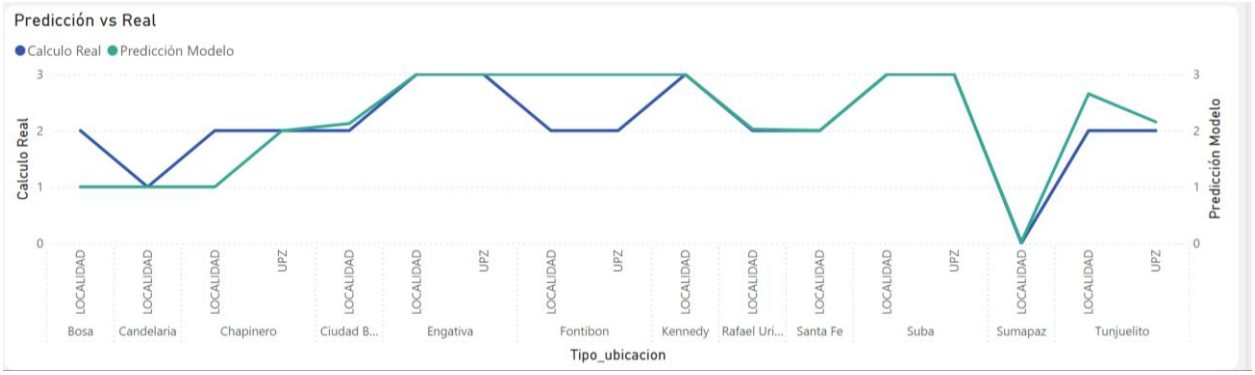
Época	Tiempo de Época (s)	Pérdida de Entrenamiento	Precisión de Entrenamiento	Pérdida de Validación	Precisión de Validación
1	14.38	1.0354	0.5022	0.7977	0.5681
2	11.63	0.9199	0.5398	0.6912	0.7763
3	11.63	0.8589	0.6008	0.6443	0.9689
4	11.64	0.8442	0.6115	0.9497	0.4843
5	11.65	0.8047	0.6534	0.5802	0.9456

Fuente: Elaboración Propia

Estos resultados, además de depender de las funciones intrínsecas del sistema inteligente, se obtienen mediante el uso equilibrado de dos conjuntos de información: el conjunto de entrenamiento y el conjunto de validación. La red neuronal aprende a partir del conjunto de entrenamiento y valida lo aprendido con el conjunto de validación, lo que garantiza que no está simplemente memorizando las respuestas de los datos conocidos, sino identificando los patrones subyacentes a los indicadores utilizados para evaluar los datos proporcionados.

Tomando en cuenta lo anterior, se puede analizar el resto de los resultados sabiendo que provienen de un análisis profundo y soportado por el entrenamiento de un sistema; un proceso que proporciona patrones y no respuestas, tal como lo realizaría un ser humano.

Figura 35. Comparativa sobre los datos reales vs predicción



Fuente: Power BI Dashboard Desempeño. Elaboración propia

En la figura 35 está expuesta una comparativa en la que se enfrentan los datos reales de algunas ubicaciones (azul oscuro) y los datos generados por la red neuronal (turquesa). Es apreciable como en casos puntuales la precisión es del 100%, y como en líneas generales la tendencia de los datos generados concuerda con los datos reales. Dicho contraste es utilizado como representación de la fiabilidad de los datos, demostrando que la red neuronal puede ser utilizada a modo de predicción educada, sirviendo tanto a funcionarios como a público civil como método para entender la criminalidad según la zona que transiten.

Es importante, también, entender que el análisis estadístico/aritmético es posible que genere disparidades que, a nivel porcentual, pueden parecer relativamente altas. Por ejemplo, situaciones en las que una disparidad de 2 a 3 no parece muy alta, pero a nivel aritmético-porcentual es un aumento del 50%. Dicho análisis es contrarrestado por el entendimiento de las

tendencias y los patrones, información que no es fácilmente reflejada con el cómputo y limitaciones que conlleva la realización de este prototipo en un entorno de desarrollo personal.

Por último, otro resultado generado por el sistema son las diversas tasas de incidentes por ubicación, indicando la suma total de todas las tasas de incidentes en cada municipio. Esta información sirve de punto de comparación para, por ejemplo, entender la caída en las predicciones de Sumapaz, donde tanto los datos reales (históricos) como los predichos por el sistema inteligente son los más bajos de todos los presentados.

Adicionalmente, este modelo integra el uso de IA para tomar una muestra de meta data de las predicciones para analizar la confiabilidad del modelo y apreciaciones interesantes que encuentra en el conjunto de datos final devolviendo un pequeño informe del tema.

El análisis de estos resultados genera conclusiones interesantes que serán cubiertas en apartados posterior a profundidad, pero que dejan entrever que el sistema: al contar con información, al indicarse los suficientes indicadores para su entrenamiento y al comparar la data generada con la histórica, cumple con su función y se plantea como una herramienta factible para cumplir con los propósitos planteados en el objetivo general y los diversos objetivos específicos.

Propuesta de solución a la problemática

Al ver los resultados a través del lente de la problemática que se tomó en cuenta a la desarrollar esta red neuronal, se retoma la pregunta, ¿cómo se puede mejorar la seguridad en Bogotá a través de alertas basadas en modelos predictivos que ayuden a evitar áreas de alta criminalidad cuando la gente se desplaza en la cotidiana?

La situación actual en la que se encuentra Bogotá y sus distintos municipios es alarmante, pero el hecho de que existan estos siniestros y que hallan puntos especialmente relevantes implica que hay patrones que se pueden analizar y, por ende, ser aprendidos para su futura predicción educada. En la actualidad el sistema es enteramente reactivo, dependiendo de la víctima y/o testigo el realizar la notificación, en caso de ser posible, a las autoridades que posteriormente realizan los procedimientos necesarios y adecuados.

Esta situación actual genera oportunidades, si se lograse transformar el sistema de reactivo (esperar a que suceda el hecho) a proactivo (evitar que suceda el hecho) el beneficio es múltiple: Las fuerzas de la ley se ven menos saturadas con casos posiblemente evitables, los centros de salud evitan posibles heridos, los ciudadanos pasarían por menos situaciones traumáticas.

El hecho de contar con información que previamente era imposible obtener (patrones y predicciones educadas basadas en información histórica) más allá de recomendaciones y comentarios a pie de calle implica que todos esos beneficios son alcanzables con las herramientas adecuadas disponibles con la tecnología que se tiene hoy día.

El objetivo fundamental, siendo en todo momento la gente, necesita de una información valiosa que, a su vez, puede ser extremadamente escasa y complicada de conseguir a tiempo real. Se plantea entonces una propuesta de solución al problema planteado: la posibilidad de uso de un tipo de información previamente inexistente, una información basada en patrones y predicciones más allá del razonamiento humano promedio.

Habiendo desarrollado la inteligencia artificial, indicados los parámetros, entrenado el modelo, validado y confirmado su adecuada funcionalidad se puede, entonces, proponer la

utilización de un modelo basado en el prototipo presentado como una posible solución a la situación de inseguridad que viven constantemente los bogotanos.

La implementación utilizada para el desarrollo del presente trabajo es escalable y sus parámetros fundamentales pueden verse generosamente beneficiados si se llevase de un entorno donde los recursos son escasos y limitados a donde pueda verdaderamente ser utilizada a su máximo de capacidad.

Si se contase con información privilegiada a tiempo real, con volcados diarios de los eventos sucedidos a lo largo del día el modelo tendería a volverse más preciso, dando con cada iteración mejores indicaciones de alta utilidad para todos los usuarios civiles o de las fuerzas de la ley que encuentren valor en las predicciones.

Se debe construir una infraestructura de captación y gestión de información sólida, que contemple la integración de APIs para la búsqueda de información online, y que permita la obtención de datos en tiempo real desde fuentes de datos como bases de datos gubernamentales, plataformas de monitorización, etc. Se debe utilizar, a su vez, bases de datos distribuidas para almacenar y recuperar grandes volúmenes de información, contar con protocolos de seguridad para la protección de información sensible, implementar un preprocesado automático de la información para asegurar la calidad de esta antes de poder ser utilizada por el modelo, etc. De este modo, se podrá asegurar el funcionamiento ininterrumpido y seguro del sistema en escenarios dinámicos y en escenarios donde haya una gran cantidad de información. Igualmente, en pro de entregar un eficaz acceso a los datos en los escenarios de uso real, estos son los puntos a los que se ha de prestar atención:

Mecanismos de acceso a datos en tiempo real: Asegurar el acceso a las fuentes de datos en tiempo real, sean bases de datos de administraciones públicas o plataformas y servicios de

monitorización, incluso cuando estén bajo alta demanda o puedan caer; utilizando APIs sólidas y protocolos que aseguren el acceso en tiempo real sin comprometer las características del modelo.

Escalabilidad de la infraestructura de datos: Usar tecnologías que permitan escalar la infraestructura de la misma manera como servidores distribuidos, almacenamiento en la nube y carga balanceada para que el modelo pueda manejar altas cargas de datos en condiciones adecuadas sin reducir el rendimiento del sistema.

Manejo de datos en situaciones adversas: Implementar mecanismos de redundancia y backup para asegurarse de que el sistema sigue siendo accesible en situaciones de eventos inesperados o de pérdida del acceso a los datos; de manera que el sistema siga funcionando de manera continua y segura, incluso en condiciones dinámicas.

Por otra parte, los costos de mantenimiento son relativamente reducidos y razonables, especialmente si el funcionamiento de esta se encuentra en la nube y se compara con la posibilidad de evitar siniestros que afectan negativamente la vida de los ciudadanos de la capital. La necesidad de personal calificado se vuelve obligatoria pero las oportunidades de formación actualmente no son escasas, por lo que no se considera una limitante especialmente relevante a tomar en cuenta.

Discusión

El desarrollo del modelo predictivo para la seguridad urbana en Bogotá realizado tuvo varias limitaciones para entender su alcance y aplicabilidad de los resultados que se obtuvieron en el presente trabajo.

Una de las principales limitaciones del actual modelo es la calidad y disponibilidad de los datos utilizados. Los datos se encuentran incompletos o desactualizados, lo que podría actualmente afectar la precisión del modelo. Además, la falta de algunos datos que proporcionen la información contextual de las zonas, como iluminación de las calles y la actividad comercial, esto limita la capacidad del modelo para capturar todos los posibles factores de la criminalidad en Bogotá. Haciendo recolección de nuevos datos enfocados en el entorno e integrándolo al modelo, puede dar un mayor contexto y generar un modelo más asertivo y fiable.

Además, el entendimiento de la criminalidad como un fenómeno complejo, que se ve influenciado por múltiples factores socioeconómicos, culturales y psicológicos; puede ayudar al modelo predictivo actual, el cual está basado en datos históricos y que aún no llega a comprender del todo la criminalidad. La inclusión de factores cualitativos y la colaboración con expertos en criminología y sociología mejoraría la interpretación de los resultados y la formulación de políticas efectivas.

Aunque el modelo actual está enfocado a la ciudad de Bogotá, su aplicabilidad en otras ciudades e incluso en otros países estaría limitado por estructuras urbanas, políticas de seguridad y patrones de la criminalidad o la cultura. Se vería necesario realizar estudios

comparativos en diferentes entornos urbanos para validar la generalización del modelo y adaptarlo a las características específicas de cada región.

Ahora bien, teniendo en cuenta que el uso de algoritmos avanzados y grades volúmenes de datos presenta desafíos técnicos significativos. Es necesario de una infraestructura robusta y una gestión eficiente de los datos son aspectos críticos en la implementación exitosa del modelo. Además, se debe garantizar las seguridad y privacidad de los datos de los usuarios es fundamental para mantener a confianza de los ciudadanos.

Lo anterior nos lleva a que, además, hay algunas implicaciones éticas importantes, dado que, si los datos utilizados tienen sesgos, este podría llevar a la discriminación o clasificación injusta de ciertos grupos o zonas de la población. Lo ideal es aplicar estos modelos con la responsabilidad que este conlleva, asegurando que las decisiones tomadas sean lo más justas y equitativas posibles.

Teniendo en cuenta todos los aspectos mencionados anteriormente, lo ideal sería hacer un encuentro multidisciplinario que combine técnicas de análisis de datos, como actualmente se está realizando, con conocimientos en criminología, sociología y hasta política. Además, el apoyo de las autoridades locales y organizaciones gubernamentales pueden mejorar la recolección de los datos y la implementación de una mejor solución. Además, la continua evaluación y ajuste del modelo basado en los datos nuevos y retroalimentación de los usuarios finales.

Sin embargo, aunque el modelo predictivo muestra un gran potencial, se debe reconocer y abordar los diferentes puntos anteriormente expuestos para garantizar su efectividad y validez. La integración de más datos en tiempo real, la colaboración interdisciplinaria y demás es crucial para el éxito de la implementación de este modelo en la ciudad de Bogotá.

Conclusiones y Trabajo Futuro

Conclusiones

Se recolectaron y procesaron datos de seguridad de múltiples fuentes oficiales y públicas de Bogotá. El procesamiento incluyó técnicas avanzadas de limpieza, normalización y enriquecimiento de datos, permitiendo que la información sea consumida eficientemente por el modelo predictivo. Este proceso garantizó la calidad y la integridad de los datos, lo cual es esencial para la precisión del modelo.

La implementación del modelo predictivo contribuye significativamente a la mejora de la seguridad y la percepción de seguridad en Bogotá. Al reducir la exposición de los habitantes a zonas con altos índices de criminalidad, se espera que la calidad de vida mejore, disminuyendo el miedo y la ansiedad relacionados con la inseguridad.

Durante el procesamiento de datos, se identificó la relevancia de integrar información geoespacial y temporal para mejorar la precisión del modelo predictivo. La utilización de datos detallados sobre la ubicación de incidentes delictivos y su evolución temporal permitió generar mapas de calor y predicciones más exactas. Este enfoque integral es crucial para abordar problemas de seguridad de manera efectiva y adaptativa.

La implementación de una arquitectura robusta utilizando herramientas como Azure Data Lake, Azure Databricks y Azure SQL Server permitió una transformación y almacenamiento eficientes de grandes volúmenes de datos. Esta infraestructura no solo soporta la escala del proyecto actual, sino que también permite la expansión futura, asegurando que el sistema pueda manejar incrementos en la cantidad y variedad de datos sin comprometer el rendimiento.

Dentro del proceso de desarrollo que fue iterativo y sufrió múltiples cambios durante el tiempo en que se trabajó se logró que el modelo fuera capaz de obtener una alta precisión de cerca del 95% a 97% en la identificación de zonas con diferentes niveles de inseguridad.

Lo anterior se da gracias a la forma en que combinaron las capas convolucionales, que analizan la distribución espacial de los incidentes, con las capas LSTM, que se encargan de los patrones temporales. Adicionalmente a la integración de todo el DWH en un único dataset con las características para realizar el análisis predictivo de forma satisfactoria.

Además, se recurrió al uso de técnicas como el Early Stopping cuya decisión fue acertada posteriormente a las pruebas ya que permitió corregir el punto en que el modelo se acostumbraba demasiado a los datos de entrenamiento, lo que luego aseguro que pudiera realizar una mejor generalización a situaciones reales.

Otro punto para destacar fue la aplicación del enfoque de balancear los datos ya que debido a la imposibilidad de usar la totalidad de la muestra fue crucial para poder mejorar la exactitud del modelo durante su ejecución con la muestra proporcionada.

Con el contexto anterior se comprendió que, si las clases menos representadas no tenían suficiente peso en el conjunto de datos, las predicciones podrían estar sesgadas para ello si se incrementaba la frecuencia de las muestras menos comunes, se podría permitir asegurar que el modelo estaría considerando todas las categorías de inseguridad con la misma importancia con propósito de que lo llevara a un mejor desempeño en sus distintas aristas.

Otro punto clave en el funcionamiento del modelo fue la decisión de normalizar los datos por cada 100,000 habitantes debido a la población total lo que contribuyó a que las comparaciones entre las distintas zonas de Bogotá fueran más justas y comprensibles.

Cuando se adopta este enfoque, se puede construir un juicio de una manera clara al interpretar la incidencia delictiva, permitiendo que tanto las autoridades como aquellos interesados del tema pudieran identificar las áreas más y menos seguras con mayor facilidad en este caso apoyado por el dashboard construido que contiene los datos geoespaciales para poder verlo visualmente de las predicciones.

Se debe también exaltar que el proceso de aprendizaje del modelo mostró mejoras constantes, lo que refleja las optimizaciones en las predicciones con cada época. La disminución progresiva en las pérdidas durante el entrenamiento y la validación arrojaron datos interesantes que indicaban que no solo estaba aprendiendo a partir de los datos, sino que estaba afinando sus capacidades para replicar la realidad, de tal manera como se esperaba al plantear los objetivos específicos del proyecto.

La propuesta de esta solución también permite vislumbrar potenciales aplicaciones para continuar su crecimiento, refinamiento con miras a ser aún más precisa si se implementa con datos en tiempo real. Se pudiera incorporar información constante por eventos reportados sobre los incidentes delictivos, el modelo podría proporcionar alertas más útiles tanto para los cuerpos de seguridad como para los ciudadanos.

Por otra parte, se corroboró que el uso de la nube y los costos operativos moderados hacen que esta propuesta sea viable y escalable, siempre y cuando se cuente con personal calificado para realizar el mantenimiento sobre los servicios y la solución.

Para cerrar, se puede decir que a lo largo de esta investigación y desarrollo se pudieron alcanzar todos los objetivos específicos propuestos, permitiendo validar los supuestos que se evaluaron.

Respecto al primer objetivo, que se hablaba sobre el diseño y desarrollo del prototipo del algoritmo de aprendizaje automático, este se cumplió satisfactoriamente al poder desarrollar e implementar un modelo que hizo uso de técnicas avanzadas para procesar tanto la distribución espacial como los patrones temporales de los incidentes delictivos. Facilitando la predicción e identificar las áreas de mayor riesgo con alta precisión teniendo en cuenta la tasa de incidentes conforme a la población de la ciudad de Bogotá.

Frente al segundo objetivo, orientado a la recolección y procesamiento de la información de seguridad de Bogotá, se logró mediante un trabajo de recopilación, análisis y normalización de datos de fuentes oficiales y publicas de los organismos concernientes a estos temas en algunos casos geojson.

Este trabajo investigativo garantizó que el modelo trabajara con información de alta calidad desde su origen y fuera trabajado con la misma responsabilidad en su tratamiento, lo que fue crucial para augurar que los indicadores fueran precisos soportados en la realidad de la ciudad.

Por último, el tercer objetivo, relacionado con la evaluación del prototipo, se alcanzó satisfactoriamente al poder obtener un alto desempeño en la identificación de zonas inseguras, con una precisión de entre 95% y 97% entre las diferentes iteraciones y la muestra con la que se trabajó de la totalidad ya que se tenían cerca de 615963208 registros en la fact.

La comparación con frente a otros sistemas existentes revisado en el marco teórico de este trabajo confirmó la efectividad del modelo, afirmando la capacidad para dar un apoyo en la toma de decisiones en materia de seguridad pública.

Por otra parte, se ejecutó nuestro instrumento de investigación, es decir una encuesta para recopilar percepciones de los ciudadanos sobre la seguridad en Bogotá, lo que brindó una perspectiva complementaria al análisis técnico que ya soportaba el modelo.

Entre las preguntas más relevantes, se incluyó la percepción de seguridad en los barrios, donde el 47.32% calificó su seguridad como "regular" (3 en una escala del 1 al 5), y un 41.96% de los encuestados indicó haber sido testigo o víctima de algún acto delictivo en el último año. Estos resultados reflejan que la percepción de inseguridad está ampliamente extendida y respaldada por experiencias directas de la población.

Los resultados de la encuesta sientan las bases para la importancia de desarrollar un modelo predictivo que permita anticipar áreas de riesgo y mejorar la sensación de seguridad. La herramienta podría enfocarse en identificar zonas que necesitan más atención, aprovechando el uso de datos históricos y características específicas de cada lugar para optimizar la asignación de recursos y las estrategias de prevención.

De esta misma manera, el 67.86% de los participantes consideró que un sistema predictivo de este tipo podría contribuir positivamente a la seguridad en la ciudad, lo que refuerza la aplicabilidad, aceptabilidad e impacto potencial de la implementación de una solución desarrollada con estas características así como la necesidad de investigar y cooperar con estos organismos para investigar sobre la aplicación de las nuevas tecnologías para elevar la calidad de vida en un aspecto tan especial como es seguridad en una ciudad en constante expansión y evolución como lo es la ciudad de Bogotá.

Trabajo Futuro

El desarrollo e investigación del proyecto *Modelo Predictivo para Mejorar La Seguridad Urbana en la Ciudad de Bogotá*, muestra potencial para optimizar la asignación de recursos y mejorar la percepción de seguridad entre la ciudadanía. Sin embargo, se identifican varias

oportunidades para futuras investigaciones y desarrollos que pueden expandir y mejorar los resultados obtenidos.

Para comenzar, es esencial considerar la integración de una gama variada de nuevas fuentes de información. Actualmente, el modelo se basa en información histórica principalmente de seguridad pública. Incorporar datos adicionales como patrones de tráfico, eventos climáticos y datos socioeconómicos puede proporcionar una visión más completa y permitir predicciones más precisas. Algunos factores a considerar, pueden ser la iluminación de las calles, la presencia de CAI de policía, cámaras de seguridad y hasta actividad comercial nocturna que pueden influir significativamente en los índices de criminalidad y se deberían considerar en el modelo para una próxima versión.

También, se podría explorar el uso de otros algoritmos avanzados como redes neuronales profundas, algoritmos de aprendizaje por refuerzo y técnicas de ensemble learning. Estos métodos podrían capturar patrones más complejos y proporcionar predicciones más robustas y precisas. Además, el uso de técnicas como el aprendizaje transferido puede ser investigado para aprovechar modelos pre-entrenados en contextos similares y adaptarlos a las necesidades de Bogotá.

Uno de los puntos más relevantes previstos para el futuro, es la implementación del modelo para su recepción de datos en tiempo real. Esto implicaría no solo predicción de áreas de alta criminalidad, sino también una alta adaptabilidad y aprendizaje de nuevos datos a medida que se reciben. La escalabilidad del modelo validar nuevas y más áreas geográficas e integrarse con aplicativos o sistemas de seguridad en tiempo real podría ayudar a las autoridades a mejorar sus tiempos de respuesta ante los incidentes.

Adicionalmente, se espera que este modelo se pueda aplicar dentro de alguna aplicación de movilidad como Waze o Google maps, para que los ciudadanos puedan recibir alertas o recomendaciones en sus rutas basadas en las predicciones del modelo, buscando que los usuarios sean notificados en tiempo real y sugiera rutas alternativas más seguras. Además, la interacción y reportes continuos de los usuarios, podría también ser parte de los insumos de modelo y así poder refinar más este.

Finalmente, es imperativo realizar estudios a largo plazo para evaluar el impacto real del modelo en la seguridad y calidad de vida de los ciudadanos. Esto tendría en cuenta no solo la reducción de índices de criminalidad, sino también la mejorar de percepción de seguridad y confianza en las autoridades locales.

Aunque el modelo actual tuvo unos excelentes resultados y muy prometedores, este tiene múltiples oportunidades para expandirse y mejorar por medio de la investigación. Integrar más datos, mejorar los algoritmos, obtener datos en tiempo real, implementar en plataformas y analizar en el largo plazo, puede hacer la diferencia en un sistema de seguridad urbana más eficiente y efectivo para la ciudad de Bogotá.

Referencias

AI for science, energy, and security. (2023). *Argonne National Laboratory*. https://www.anl.gov/ai/reference/AI-for-Science-Energy-and-Security-Report-2023

Casteel, C., & Peek-Asa, C. (2000). Effectiveness of crime prevention through environmental design (CPTED) in reducing robberies. *American Journal of Preventive Medicine, 18*(4S), 99-115. https://doi.org/10.1016/S0749-3797(00)00146-X

Cozens, P. M., Saville, G., & Hillier, D. (2005). Crime prevention through environmental design (CPTED): A review and modern bibliography. *Property Management, 23*(5), 328-356. https://doi.org/10.1108/02637470510631483

Gonog, L., & Zhou, Y. (2019). A review: Generative adversarial networks. *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Xi'an, China, 505-510. https://doi.org/10.1109/ICIEA.2019.8833686

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org/

Google. (2023). *Waze*. https://support.google.com/waze/answer/6071177?hl=es&ref_topic=9022747&sjid=14869956915034199756-NA

He, J., & Zheng, H. (2021). Prediction of crime rate in urban neighborhoods based on machine learning. *Chongqing.* https://doi.org/10.1016/j.engappai.2021.104460

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Massachusetts: The MIT Press.

Ng, A. (2018). *Machine learning yearning: Technical strategy for AI engineers in the era of deep learning.* deeplearning.ai. https://github.com/ajaymache/machine-learning-yearning/blob/master/full%20book/machine-learning-yearning.pdf

Ordóñez, H., Cobos, C., & Bucheli, V. (2020). Machine learning model for predicting theft trends in Colombia. *Revista Ibérica de Sistemas e Tecnologias de Informação, 29*(E), 494-506. https://www.researchgate.net/publication/340634796_Machine_learning_model_for_predicting_theft_trends_in_Colombia

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach.* New Jersey: Pearson.

Secretaría de Seguridad de Bogotá. (2023a). *Análisis de datos Siedco*. http://analitica.scj.gov.co/analytics/saw.dll?Portal&PortalPath=/shared/OAIEE/SIEDCO/_portal/An%C3%A1lisis%20de%20datos%20Siedco&NQUser=publico&NQPassword=publico2019

Secretaría de Seguridad de Bogotá. (2023b). *Indicadores de seguridad y convivencia*. https://scj.gov.co/sites/default/files/documentos_oaiee/Reporte_bogota_2023_07.pdf

Secretaría de Seguridad de Bogotá. (2023c). *Tecnología para la seguridad*. https://scj.gov.co/es/noticias/tecnología-cuidamos-bogotá

Stec, A., & Klabjan, D. (2018). Forecasting crime with deep learning. *arXiv preprint arXiv:1806.01486*. https://arxiv.org/abs/1806.01486

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: The MIT Press. https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

Townsend, A. M. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia.* New York, NY: W. W. Norton & Company.

https://books.google.com/books/about/Smart_Cities_Big_Data_Civic_Hackers_and.html?id=PSsGAQAAQBAJ

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows: The police and neighborhood safety. *The Atlantic Monthly, 249*(3), 29-38. https://www.theatlantic.com/past/politics/crime/windows.htm

Anexo. Pipeline de Azure Data Factory

Este anexo contiene el repositorio del pipeline creado en Azure Data Factory. Se realizó un pull directamente a GitHub desde Azure para mantener una copia completa de todo el proceso.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/tree/devdatabricks

Anexo. Script Python "Flatfile data to Raw data"

Este script se utilizó para la transformación de los datos desde la capa flatfile hasta la capa raw, siguiendo la arquitectura propuesta.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Flatfile%20data%20to%20Raw%20data.py

Anexo. Script Python "Raw data to Transit data"

Este script se utilizó para la transformación de los datos desde la capa raw hasta la capa transit, conforme a la arquitectura propuesta.

Enlace:
https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Raw%20Data%20to%20Transit%20Data.py

Anexo. Script Python "Transit data to Curated data"

Este script se utilizó para la transformación de los datos desde la capa transit hasta la capa curated, según la arquitectura propuesta.

Enlace:
https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Transit%20Data%20to%20Curated%20Data.py

Anexo. DDL base de datos Azure Data Base

Este DDL se utilizó para crear las tablas especificadas en los modelos dentro de la base de datos de Azure Database mencionada en el documento.

Enlace:
https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Transit%20Data%20to%20Curated%20Data.py

Anexo. Modelo Físico de datos

Este modelo se utilizó para crear las tablas y estructuras del proyecto, sirviendo como base para su construcción.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Modelo%20de%20datos%20Proyecto-Modelo%20F%C3%ADsico.png

Anexo. Modelo Lógico de datos

Este modelo se utilizó para definir las tablas y estructuras del proyecto, sirviendo como guía para su implementación.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Modelo%20de%20datos%20Proyecto-Modelo%20Logico.png

Anexo. Modelo Conceptual de datos

Este modelo se utilizó para definir el esquema general de los datos y sus relaciones, sirviendo como base para el diseño del proyecto.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Modelo%20de%20datos%20Proyecto-Modelo%20conceptual.png

Anexo. Arquitectura propuesta

Esta arquitectura fue la primera versión planteada, desarrollada bajo la metodología TOGAF y Archimate. Sirvió como guía para la realización de todo el proyecto.

Enlace:

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Arquitectura%20Proyecto%20de%20grado.jpeg

Anexo. Archivos Base para el proyecto

Estos archivos contienen los datos utilizados para crear toda la transformación, base de datos, datalake y modelo predictivo del proyecto.

Enlaces:

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Incidentes%20reportados%20por%20UPZ.zip

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Delito%20de%20alto%20impacto%20por%20sector%20catastral.zip

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Incidentes%20reportados%20por%20UPZ.zip

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Incidentes%20reportados%20por%20sector%20catastral.zip

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/hurto_a_motocicletas_6.xlsx

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/hurto_a_personas_22.xlsx

https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/hurto_automotores_14.xlsx

Anexo. ETL Datalake

Este anexo contiene la carpeta con el código del ETL utilizado para generar el datalake, que proveerá la información necesaria para el modelo predictivo.

Enlace: https://universidadeaneduco-my.sharepoint.com/:f:/g/personal/wvargasm4887_universidadean_edu_co/Ev-YsjFiPuJHikIbv0U0BRkBIY6r5sqrQH74lcktM3W6nA?e=55SWR6

Anexo. Script creación del Modelo Predictivo

Este anexo incluye el script utilizado para la generación y el entrenamiento del modelo predictivo desarrollado para este proyecto.

Enlace: https://universidadeaneduco-my.sharepoint.com/:f:/r/personal/wvargasm4887_universidadean_edu_co/Documents/Repositorio%20-%20Tesis%20-%20Soluci%C3%B3n?csf=1&web=1&e=aAfZvb

Anexo. Encuesta de percepción de seguridad

Este anexo contiene la encuesta realizada para evaluar la percepción de seguridad entre los ciudadanos de Bogotá, utilizada como herramienta para el proyecto.

Enlace: https://github.com/ALEZAMAEAN/ProjectoGradoEan/blob/dev/Encuesta%20sobre%20Percepciones%20de%20Seguridad%20en%20Bogota%CC%81%20(2)%20(1).pdf