Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Predicción de ocurrencia de delitos

Implementación de modelos predictivos en base a los casos registrados en CABA durante el periodo 2017-2019

AUTOR: RAFAEL ALBERTO ZAMBRANO MEDINA

TUTOR: ROBERTO ABALDE

DICIEMBRE 2020

Resumen

La recopilación de datos referentes a la ocurrencia de delitos según una fecha y localización determinada, brinda un amplio abanico de oportunidades para la explotación de información en diversos tipos de organizaciones. El presente trabajo tiene como objetivo implementar modelos predictivos de aprendizaje automático, que permita relacionar una ubicación y un momento determinado con la ocurrencia de delitos, con base en los casos registrados entre los años 2017 y 2019 en la Ciudad Autónoma de Buenos Aires (CABA). El foco principal de la investigación, reside en contextualizar la problemática de la criminalidad dentro de la gestión de datos de las organizaciones.

La estructura del estudio parte de un proceso de recopilación de información, indexando elementos que se consideren importantes en el momento y entorno de la ubicación geográfica del delito. Los factores agregados de "entorno-tiempo", se transforman en posibles variables predictivas, validando previamente la existencia de relaciones entre estos elementos y la incidencia de la criminalidad. Una de las hipótesis principales del estudio, reside en el principio donde el delito al ser efectuado por un ser humano, no sigue un comportamiento aleatorio, todo lo contrario, subyace y responde a patrones sociales que se evidencian en la variación de la incidencia y los tipos de delitos registrados, en función del tiempo, el clima y de las distintas zonas geográficas de la ciudad.

Palabras clave: Ocurrencia de delitos, predicción, modelos predictivos, entorno-tiempo, toma de decisiones.

INDICE

Introducción

Cuando se toman decisiones para la aplicación de políticas públicas de seguridad, o por parte de una empresa cuyo negocio esté afectado por el fenómeno de la inseguridad, surge la necesidad de complementar los análisis descriptivos, el conocimiento del negocio y el juicio experto, con herramientas que brinden una mayor confianza al momento de ejecutar cualquier medida. Dentro de estas herramientas se encuentran los modelos predictivos, estos pueden generar un cambio en el impacto de las estrategias gerenciales, sin embargo, las causas o motivos que influencian las conductas delictivas son multicausales e involucran áreas muy diversas como la sociología, estadística, psicología, entre otras. Por lo tanto, es necesario acotar el tipo de relaciones que se desean analizar y validar que las mismas tengan poder predictivo dentro de los modelos que se utilicen.

La problemática de la ocurrencia de delitos es de índole social y económica, dado lo amplio y transversal que resulta estudiar este tópico, se pretende abordar el estudio elaborando una relación funcional entre la problemática de estudio y la dualidad "entorno-tiempo". Leong & Sung (2015) amplían de manera conceptual las diversas aplicaciones de este término, el entorno hace referencia a los medios físicos más cercanos a la ubicación física de la ocurrencia de delitos, por su parte, el tiempo guarda relación con las condiciones climáticas y el momento temporal, como los días o meses. Con ambos factores se busca verificar la factibilidad de relacionar estos elementos con los hechos delictivos, con la finalidad de predecir su ocurrencia y encontrar posibles variables que tengan mayores contribuciones en las predicciones.

Por tales motivos, se busca implementar modelos de aprendizaje automático que permitan predecir la ocurrencia de delitos en un conjunto de esquinas de CABA, en función de elementos climáticos y del entorno cercano a la ocurrencia del crimen, en base a los delitos registrados durante los años 2017 y 2019. El análisis se reduce al factor tiempo y al entorno físico del delito, permitiendo a través de la localización indexar distintos elementos que puedan ayudar a mejorar la calidad de las predicciones. El resultado pretende mostrar una herramienta que permita complementar la toma de decisiones, permitiendo a su vez construir el siguiente planteamiento, ¿Cómo puede relacionarse el entorno geográfico, el

clima y la ocurrencia delictiva, para brindar herramientas que complementen la ejecución de políticas públicas de seguridad?

Específicamente, se propone analizar la contribución que tienen los elementos más cercanos al entorno geográfico del delito en la ocurrencia del mismo. Además, implementar un flujo de trabajo que permita hacer reproducible el estudio desde el proceso de captura de datos hasta la presentación e interpretación de resultados. Finalmente se plantea, diseñar una metodología que permita consumir de manera sencilla la salida del modelo para la toma de decisiones. Estos son objetivos que se abordarán paulatinamente a medida que se desarrolle el estudio, siendo a su vez condiciones necesarias que permiten generar un proyecto factible y útil para la toma de decisiones.

Como premisa o hipótesis fundamental, se sostiene que mediante la implementación de un modelo predictivo de aprendizaje automático, es factible relacionar el factor "entorno-tiempo" con la ocurrencia de delitos, permitiendo generar predicciones que puedan utilizarse en la aplicación de políticas de seguridad. Adicionalmente se persiguen distintos supuestos, el primero de ellos hace referencia al histórico de delitos ocurridos, afirmando que la adición de los elementos de entorno y climáticos contribuirán positivamente en la generación de mejores predicciones. También se plantean afirmaciones sobre la distribución de la ocurrencia de delitos, sosteniendo que se distribuye geográficamente de manera heterogénea en CABA, además se afirma que los delitos siguen patrones estacionales según los meses del año.

Distintos tópicos dictados durante la especialización serán utilizados para llevar a cabo este estudio, el primer gran desafío es el manejo de grandes volúmenes de datos provenientes de múltiples repositorios, haciendo necesario el uso de distintas técnicas de manipulación y limpieza de datos. El elemento central es la localización del delito, sin embargo, las variables iniciales de entrada están basadas en la geolocalización de las intersecciones de calles de CABA. Las esquinas de la ciudad cumplen la función de permitir la construcción del factor "entorno-tiempo", que busca relacionarse con la incidencia de la criminalidad, mediante la contabilización del número de delitos que ocurren en las cercanías

de las esquinas en distintos espacios de tiempo. Esta investigación se implementará a través del lenguaje estadístico de programación R.

Predecir la ocurrencia delictiva en distintas esquinas de una ciudad es un desafío que no solo involucra aspectos técnicos, también es necesario entender las limitaciones y el alcance del estudio, para así poder contextualizar e interpretar los resultados de manera correcta. Los datos sobre los cuales se basa esta investigación se refieren a los delitos registrados, es decir, en su mayoría hechos delictivos denunciados ante las autoridades, por lo tanto, existe un sesgo importante con respecto a la criminalidad real ocurrida en la ciudad. La aplicación de posibles cambios extraordinarios en las políticas públicas de seguridad durante el periodo de análisis, están fuera del alcance de este estudio, estableciendo como supuesto que dichas políticas se mantienen constantes, o que no tienen impactos significativos en la variación de la ocurrencia del delito.

El cuerpo del presente estudio está constituido por tres apartados principales, cada apartado contiene tres subapartados y estos a su vez presentan tres secciones internas, siendo en definitiva nueve subapartados y veintisiete secciones internas. El primer apartado hace referencia a la gestión de datos en contextos organizacionales, específicamente sobre la problemática organizacional y su vinculación con la ciencia de datos. El segundo apartado se refiere a la descripción metodológica, abordando los tópicos de recopilación, procesamiento y análisis de la información de manera conceptual. EL último apartado hace foco en la etapa de implementación, centrándose en la operacionalización de la metodología planteada. Como sección complementaria se puede consultar el apartado de Anexos, donde se pueden revisar las rutas de los directorios de información utilizados..

1 Gestión de datos en contextos organizacionales

El elemento fundamental de este proyecto reside en la gestión de datos en contextos organizacionales, por lo tanto, esta primera sección permite sentar las bases de la problemática de estudio. Aunque la implementación de modelos predictivos es también un tema central, los algoritmos por sí solos no aportan ningún valor a la organización, es necesario vincularlos con alguna problemática organizacional. En este capítulo se hace una descripción general de la organización involucrada, se analiza el proceso de gestión de datos de esta, finalmente se muestra brevemente la problemática organizacional que se desarrollará, mostrando además algunas de las principales implicaciones.

1.1 Descripción de la organización

La organización vinculada con este estudio corresponde a una institución pública del estado argentino, por lo tanto, el enfoque del proyecto es distinto al que puede plantearse con otra organización. La propuesta para resolver la problemática organizacional no pretende mostrar una solución comercial, se propone la implementación de un estudio que ayude a los tomadores de decisiones, con el propósito de aplicar políticas de seguridad más eficientes basadas en datos. A continuación, se define la organización, se muestran sus objetivos operacionales y se presenta la estructura y organigrama de esta.

1.1.1 Definición de la organización

Debido a que el estudio trata el tópico de la criminalidad en CABA, se propone como institución vinculada o usuario del proyecto a La Policía de la Ciudad Autónoma de Buenos Aires. Esta organización depende jerárquica y funcionalmente del Jefe de Gobierno a través del Ministerio de Justicia y Seguridad, además el Gobierno de la Ciudad Autónoma de Buenos Aires (GCABA), se considera empleador del personal de la institución. Toda la información referente a la organización se extrae de su la página oficial, la misma puede consultarse en la sección de Àpendice.

La institución es la encargada de operacionalizar las políticas públicas de seguridad diseñadas, por lo tanto, se considera como el organismo que está más cercano al tema de la criminalidad. Como se mencionó en la introducción del proyecto, la gran mayoría de las instituciones encargadas del diseño e implementación de políticas de seguridad, sus

decisiones están basadas en datos. Por lo tanto, la construcción de herramientas que ayuden a maximizar la eficiencia de los operativos de seguridad, puede traer beneficios transversales a toda la sociedad, repercutiendo principalmente en la disminución del delito, pero a su vez en el aumento de las actividades comerciales en distintas zonas de la ciudad.

1.1.2 Objetivos

Dentro de los principales objetivos de la institución y que a su vez están más vinculados con este estudio, se puede citar la facilitación de las condiciones que posibiliten el pleno ejercicio de las libertades, derechos y garantías constitucionales. El Mantenimiento del orden y la tranquilidad pública, la protección de la integridad física de las personas, así como sus derechos y bienes. Con respecto a los delitos, promover su investigación, persecución y sanción de sus autores, también el establecimiento de mecanismos de coordinación y colaboración para evitar la comisión de delitos. Finalmente, la promoción de intercambio de información delictiva dentro del marco de la ley, además de dirigir y coordinar los organismos de ejecución de pena a los fines de lograr la reinserción social del condenado.

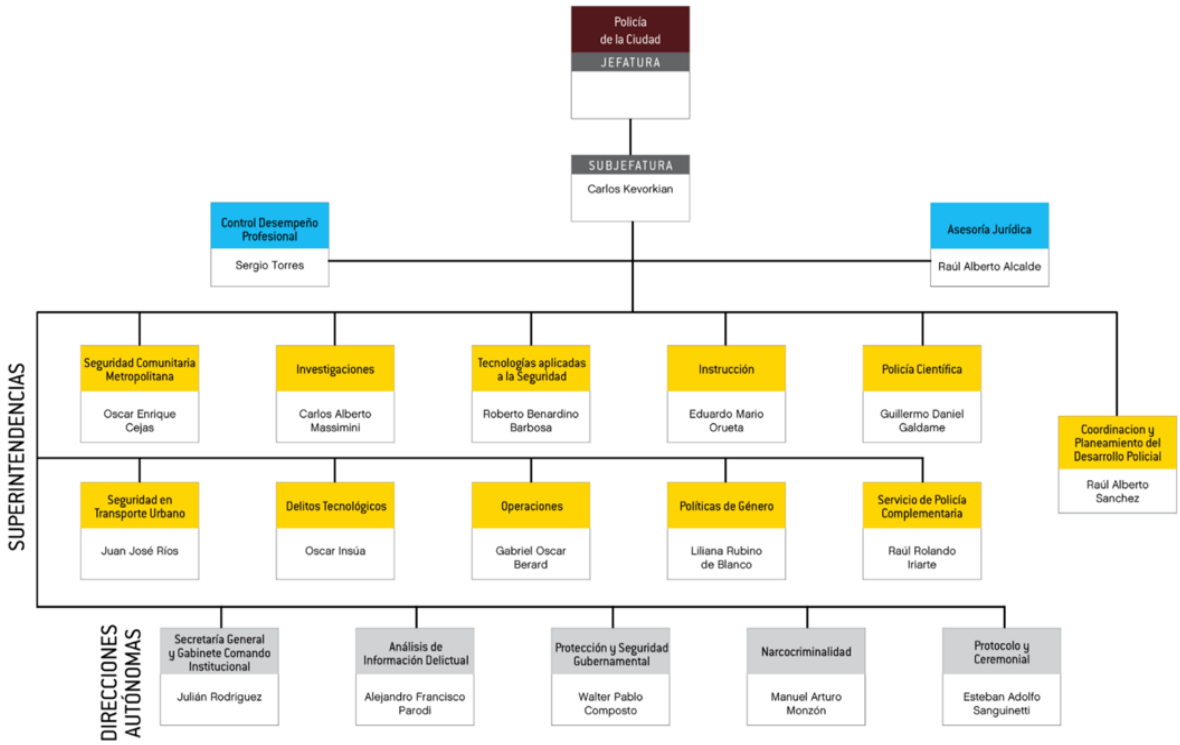
Los principios de una institución también ayudan a dar una visión más amplia acerca del funcionamiento de esta, siendo a su vez un complemento de los objetivos. Los principales principios hacen referencia al control civil sobre la gestión institucional, la cercanía con las comunidades, la transparencia, prevención, eficiencia, eficacia y la confiabilidad de la información estadística. Esto brinda un panorama bastante claro acerca de las prioridades y accionar de la organización, facilitando con esto el diseño de un proyecto que se adapte a sus necesidades.

1.1.3 Estructura

La máxima instancia dentro de la organización corresponde a la jefatura, la cual se encarga de organizar, conducir y controlar los servicios policiales y coordinarlos para el cumplimiento de las funciones. Aunque para efectos de este análisis, solo se hará énfasis sobre esta organización, el funcionamiento de la misma requiere de la interacción con otros organismos públicos. Según extrae desde el portal web de la Policía de CABA, otras de las funciones de las autoridades de la organización están relacionadas con desarrollar y elaborar propuestas de servicios de seguridad y coordinar su accionar con el Ministerio Público Fiscal

y con el Cuerpo de Investigaciones Judiciales. Además de ser uno de los encargados del diseño del Plan General de Seguridad Pública, a continuación se muestra el organigrama de la institución.

Figura 1.*Organigrama de la Policía de CABA*



Fuente: Policía de CABA

Además de la jefatura, la estructura también cuenta con subjefaturas, superintendencias y direcciones autónomas. Analizando el diagrama, las superintendencias de tecnologías aplicadas a la seguridad y operaciones, sumado a la dirección autónoma de análisis de información delictual, se perfilan como los posibles usuarios tomadores de decisiones o unidades más impactadas por el presente estudio. Este aspecto es de suma relevancia en función del alcance futuro que pueda tener el proyecto, en caso de querer contactar a alguna unidad o autoridad que pueda estar interesada en la implementación presentada.

1.2 Gestión de datos por parte de la organización

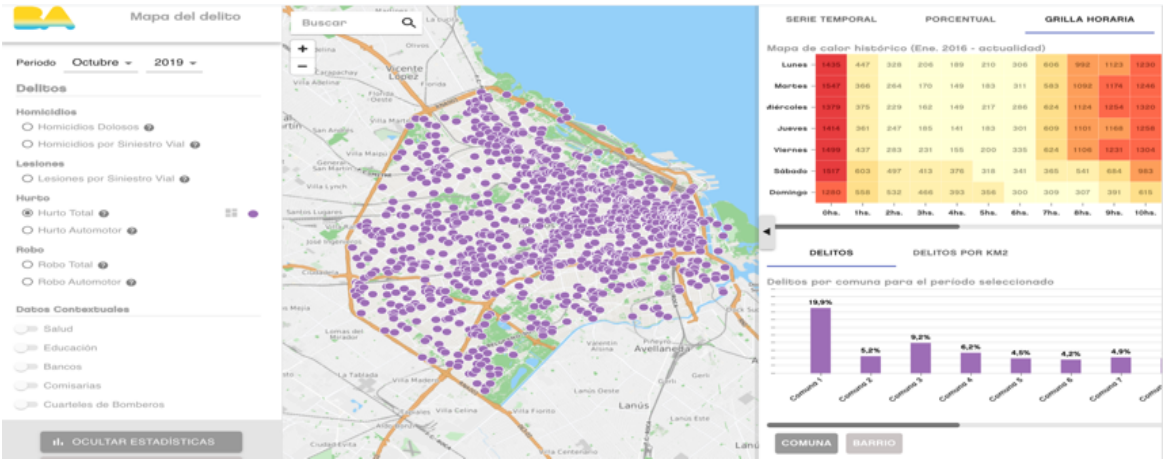
Definida la estructura y los objetivos de la organización, corresponde analizar el funcionamiento de está a través de la gestión de datos. Se evaluarán los desarrollos y soluciones de problemáticas desarrollados por la institución, de esta manera se puede inferir

el grado de utilidad que muestra la institución por el manejo de datos. También se analizarán las fuentes de datos utilizadas y algunos casos de uso de organizaciones similares.

1.2.1 Procesos de tomas de decisiones

Penn (2019) comenta que las organizaciones generalmente no se crean como instituciones basadas en datos, todo lo contrario, es un proceso evolutivo en su cultura y estrategia. Penn (2019) profundiza y plantea, para que una organización pueda tener una cultura basada en datos, se debe transitar por un proceso evolutivo de cinco etapas. La etapa inicial se denomina resistencia al dato, se caracteriza por rechazar la incorporación de la utilización de los datos en base a una serie de argumentos, usualmente basados en miedos a mostrar estrategias desalineadas o fallas graves en el funcionamiento. La etapa final se denomina "Data-driven", en este punto la organización está atravesada por los datos, todas las decisiones se diseñan e implementan en función de los datos. Además, los métodos empleados suelen ser de vanguardia, robustos, basados en rigurosidad una estadística, matemática, computacional y con sentido organizacional. Se describe la etapa inicial y final, para dar un sentido de los puntos extremos del proceso, en la publicación de Penn (2019) se pueden encontrar detalles de las etapas intermedias. Para inferir el grado de utilización de datos actual de la organización, se analizan los desarrollos y soluciones de problemáticas desarrolladas, sirviendo esto como indicadores que permitan ubicar a la institución en algunas de las etapas del proceso.

Figura 2. *Mapa del delito*

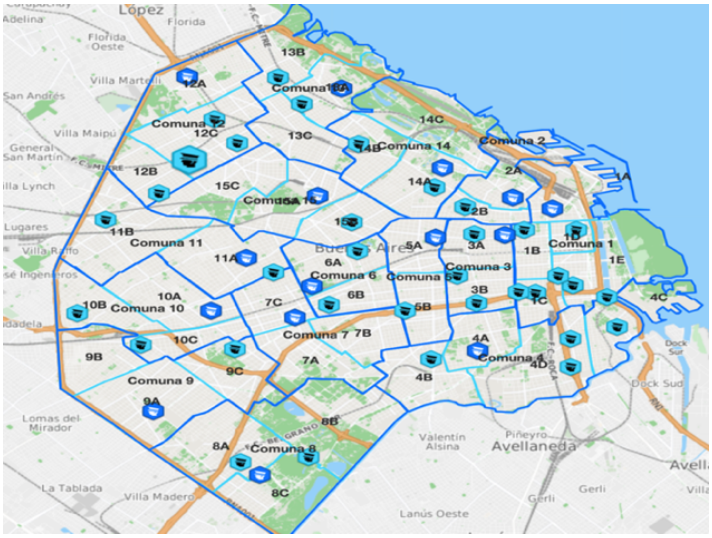


Fuente: Policía de CABA

Según se extrae del portal web de la organización, existe una herramienta analítica ampliamente utilizada y de acceso público denominada “Mapa del delito”. Se define la misma como una herramienta online con estadísticas fiables de criminalidad en la ciudad, que facilita dimensionar el problema de la inseguridad, establecer patrones, realizar diagnósticos certeros y evaluar la efectividad de nuestras respuestas. En la Figura 2 se muestra una imagen de la herramienta mencionada, se aprecia una herramienta con una gran cantidad de información para mostrar, además de múltiples opciones para filtrar y desagregar la ocurrencia de delitos. La información registra los delitos hasta el año 2019, no se han actualizado los delitos del 2020, sin embargo, la herramienta por sí mismas es un indicio de la presencia de una cultura basada en datos. Es importante destacar que este tipo de iniciativas tienen efectos que exceden a la propia institución, especialmente sobre los profesionales que se dedican a analizar y explotar información, haciendo disponible información que permite el desarrollo de análisis como el presente en el actual estudio.

En función a la herramienta mostrada anteriormente, también se encuentra otro desarrollo basado en datos con un impacto directo en las políticas públicas. Un elemento relevante en la estructura organizacional que no fue mencionado en la sección anterior, hace referencia a la distribución geográfica de las comisarías. Este es un aspecto central, ya que son las unidades que operacionalizan y permiten el despliegue activo de políticas de seguridad, en el siguiente mapa se puede apreciar dicha configuración.

Figura 3. *Distribución de comisarías*



Fuente: Policía de CABA

La Figura 3 refleja el mapa de CABA mostrando la división territorial por comunas, los puntos en el mapa de color azul mas intenso representan a las comisarias comunales, los otros puntos señalan a las comisarias vecinales. Actualmente existen 15 comisarías comunales y 28 comisarías vecinales, según se señala en el portal web, esta distribución de comisarías responde a del análisis del Mapa del Delito, del Despliegue Territorial y del Sistema Integral de Videovigilancia. Esto brinda señales acerca del funcionamiento de la organización, permitiendo inferir que toman decisiones basadas en la evidencia de datos, facilitando con esto la posible propuesta técnica que arroje este proyecto. A través de estos desarrollos no es posible identificar la etapa exacta en la cual se encuentra la organización, sin embargo, se puede afirmar que gran parte del proceso de toma de decisiones está basado en datos.

Un aspecto que muchas veces se pasa por alto, está relacionado con la vinculación entre la toma de decisiones y la ciencia de datos, específicamente en la puesta en marcha de los proyectos. Provost & Fawcett (2013), amplía esta idea y plantean la necesidad de ver a la ciencia de datos más allá de los algoritmos e implementaciones, se debe pensar en los principios básicos y conceptos que subyacen a las técnicas. Pero también, en el pensamiento sistémico que fomenta el éxito en la toma de decisiones basada en datos, dicho éxito en los entornos organizacionales, está orientado a diseñar estrategias que muestren cómo los conceptos técnicos se aplican a problemas organizacionales particulares. Este elemento está contenido dentro del espíritu del presente estudio, buscando vincular la ciencia de datos con una institución, a través de la problematización de una necesidad de la organización.

1.2.2 Datos utilizados

Es importante destacar que la naturaleza de la propia organización dificulta encontrar información referente al funcionamiento interno, específicamente en lo referente a las fuentes de información. Recordando que esta institución maneja información sensible, además de cumplir funciones que apoyan operaciones de inteligencias del estado. Dicho esto, para efectos de este proyecto se asumirá como los datos utilizados por la organización, aquellos que son de conocimiento público como se describe en los siguientes párrafos.

Los datos correspondientes a la ocurrencia de delitos surgen de las denuncias que se realizan en las distintas comisarías de la ciudad, los cuales son recopilados y posteriormente procesados. De forma complementaria se utilizan datos relacionados al entorno del delito

como comisarías, centros de salud y educativos, bancos, cuarteles de bomberos, entre otros. Los datos complementarios provienen del repositorio de datos del GCABA, es probable que también se haga uso de repositorios internos que no son de acceso público.

Como se mencionó anteriormente, el funcionamiento de esta organización requiere la interacción con múltiples instancias del estado, por lo tanto, está dada la posibilidad de que exista un flujo de transferencia de información con otras organizaciones. En relación con el presente estudio, se hará uso únicamente de datos que sean de acceso público, siendo la principal fuente de datos el repositorio del GCABA. En consecuencia, el estudio tiene la factibilidad de ser totalmente reproducible, además de poder mejorarse siendo complementario por información interna que maneje la organización.

1.2.3 Casos de uso de organizaciones similares

Se recopilaron una serie de estudios, donde se diseñaron e implementaron herramientas referentes al tema de la criminalidad, con el objetivo de apoyar la toma de decisiones de organizaciones públicas. Sevri, Karacan, & Akcayol (2017) presentaron un estudio realizado en Estados Unidos, donde analizaron elementos geográficos, étnicos y característicos de un conjunto de hechos delictivos históricos, aplicando un algoritmo de reglas de asociación, con el objetivo de crear perfiles que asocian patrones en común que relacionan a las víctimas con los delincuentes y el entorno. Al analizar las reglas de asociación arrojadas por el estudio, se observaron algunos patrones obvios desde el punto de vista psicológico, pero que en este caso se plantean con un soporte matemático, también se descubrieron asociaciones insospechadas entre las víctimas y victimarios según se modifiquen las características étnicas.

Desde una visión distinta, Sommer, Lee, & Abèle (2018) plantearon la posible influencia de las características climáticas en la criminalidad, a través de un modelo de inferencia causal. Los resultados de este estudio realizado en la ciudad de Boston, mostraron menos delitos en los días con temperaturas extremadamente bajas, en comparación con fechas donde las temperaturas eran bajas pero sin ser catalogadas de extremas, también se observó una menor tasa de incidencia de delitos durante los días con lluvia en comparación con días secos. En general sugiere la posible factibilidad de la integración de los pronósticos en los programas de prevención del delito, o en la evaluación del impacto económico,

sociológico o médico del cambio climático. Estos indicios son relevantes al momento de seleccionar variables, con la finalidad de ser indexadas en un modelo predictivo o explicativo.

El entorno es uno de los factores fundamentales a estudiar en esta investigación, Ferreira, João, & Martins (2012), realizaron un estudio donde aplicaron métodos estadísticos-espaciales, basadosen el registro de denuncias policiales de distintos delitos ocurridos en la ciudad de Lisboa. El objetivo consistía en identificar áreas donde los niveles de criminalidad tuvieran una mayor frecuencia, y que a su vez permitieran implementar modelos predictivos más precisos, que apoyaran las tomas de decisiones relacionadas con el despliegue táctico de recursos policiales. La combinación de la ocurrencia delictiva en función de la localización y el tiempo, permite transformar la problemática en un fenómeno dinámico, y a su vez capturar con mayor certeza los patrones del comportamiento humano, que es en esencia el elemento que modela la criminalidad.

Ariel & Partridge (2017) estudiaron la incidencia delictiva en la ciudad de Londres, pero focalizando el análisis en el transporte, específicamente en distintas paradas de autobús cercanas a ubicaciones denominadas “hotspots” o “puntos calientes”. Mediante un ensayo controlado aleatorio, evaluaron la variación de las denuncias de delitos en las zonas de estudio, en función de la presencia policial en distintos intervalos de tiempo. Los resultados del estudio revelaron que la vigilancia de puntos críticos puede resultar contraproducente, si los delincuentes pueden predecir de manera sistemática y precisa el patrón temporal-espacial del patrullaje policial, se produce un aumento de los crímenes registrados en los momentos en los de escasa presencia policial. Este estudio puede resultar de utilidad al momento de analizar los datos, especialmente si se desea indexar variables relacionadas con el transporte público y el patrullaje policial. Hart & Zandbergen (2014), muestran distintas metodologías para estimar los denominados “hotspots” referentes a la temática de la criminalidad, estas estimaciones son los puntos de partida de la mayoría de las investigaciones analizadas.

Duan, Ye, Hu, & Zhu (2017), realizaron también un estudio enfocado en la criminalidad, pero con un objetivo particularmente distinto implementado en la ciudad de Wuhan, aunque utilizaron el factor “entorno-tiempo”, el foco del estudio consiste en realizar predicciones de futuras ubicaciones geográficas de un determinado sospechoso. Los

investigadores utilizaron un modelo de predicción de ubicación llamado CMoB (Crime Multi-order Bayes model) , que de manera resumida, inicia aglomerando a los sospechosos en grupos similares, para luego a través de datos históricos de movilidad, estimar probabilidades de transición geográficas con cadenas de Márkov en conjunto con un enfoque bayesiano. El estudio muestra indicios de aplicaciones de modelos alternativos que se han implementado recientemente, analizando en este caso la futura movilidad geográfica de un posible delincuente. Clark (1965), realizó un estudio de aplicación de cadenas de Márkov en el ámbito del análisis geográfico, si bien no desarrollar el modelo "CmoB", muestra una explicación del modelo más básico de Márkov.

Aunque en una primera instancia se pudiera pensar que, en el pasado reciente, las instituciones encargadas de diseñar e implementar políticas de seguridad no han incorporado la gestión de datos y el aprendizaje automático, las investigaciones presentadas demuestran lo contrario. La temática de la criminalidad, también ha sido una fuente de inspiración para la implementación de los algoritmos más sofisticados de aprendizaje automático. Sin embargo, se deben destacar las iniciativas de políticas de datos abiertos de la mayoría de las principales ciudades del mundo, la información referente a la criminalidad está al alcance de todos. Este importante avance en hacer disponible la información, trae como consecuencia que la creatividad y curiosidad de muchos investigadores y profesionales de la ciencia de datos, sean un canal para tratar la problemática delictiva con técnicas y conocimientos utilizadas en otras áreas.

1.3 Problemática de la organización y la gestión de los datos

Aunque se puede observar que la organización tiene presente una cultura basada en la gestión de datos, es necesario analizar la disponibilidad e implicaciones del uso de los datos. Desde el punto de vista de la institución, pero también desde la perspectiva de un investigador que desee usar las fuentes de datos. En la última parte se hará un breve planteamiento de la problemática a tratar, señalando algunos posibles beneficios que podría traer la propuesta de valor de este proyecto.

1.3.1 Disponibilidad de la información

Siendo esta una organización perteneciente al estado, tiene la obligación de generar estadísticas oficiales, más aún con el tipo de datos que recopila y transmite. Esto trae consigo

una complejidad extra en comparación con las instituciones privadas, debido a los controles y regulaciones más estrictos a los cuales son sometidas estos tipos de instituciones. Aunque este proyecto no está orientado a analizar exhaustivamente la calidad de las estadísticas oficiales generadas, es importante destacar las dificultades existentes, para así comprender las limitaciones y alcance que puede tener un proyecto basado en estos datos.

En este aspecto, Salgado (2016) plantea los principales retos y obstáculos que atraviesan este tipo de instituciones, identificando como el mayor desafío el denominado acceso institucional. Esto hace referencia a la temática del marco legal, involucrando la legislación estadística, protección de datos, temas centrales al manejar datos delictivos, donde se involucran víctimas y victimarios. Salgado (2016) profundiza y comenta que el factor tecnológico supone otro reto de gran envergadura, principalmente por los costos mismos que genera el mantenimiento de las plataformas y sistemas de datos, pero también por las inversiones necesarias para implementar, diseñar e integrar las tecnologías a la estructura compleja de la administración pública. Estos factores son indispensables entendiendo el gran volumen de datos que debe manejar la organización, además de la necesidad de integrar la información con repositorios externos de otras instituciones.

Como se mencionó anteriormente, los datos que maneja la organización son recopilados y almacenados por la misma entidad, haciendo referencia a las denuncias delictivas, por lo tanto, la disponibilidad de la información depende del propio funcionamiento de la institución. La información que luego se publica a toda la ciudadanía en los repositorios pertinentes, provienen de un proceso de transformación de los datos recopilados. La información complementaria a los delitos, provienen de otras instancias del GCABA, esta es una ventaja al momento de integrar y combinar información. Esto se puede corroborar al revisar el repositorio del GCABA, donde se observa una estandarización y normalización general para la gran mayoría de los conjuntos de datos,

Desde el punto de vista del investigador, todos los datos necesarios se encuentran en el repositorio del GCABA, recordando que el mismo es público y de acceso gratuito. Los datos complementarios que hacen referencia al entorno y clima, también están en el mismo repositorio, sin embargo más adelante se comentarán algunas consideraciones a tener en cuenta. Según se extrae de la página oficial de la organización, la información suele actualizarse con cierta periodicidad, por ejemplo, el denominado Mapa del delito, se

actualiza dos veces al año. Esto sugiere que, si algún proyecto se desarrolla desde afuera de la organización, no sería posible tener una actualización en tiempo real, ya que la información puede tardar meses en actualizarse.

1.3.2 Implicaciones con el uso de los datos

La información delictiva es inherentemente sensible y confidencial, por lo tanto, la captura de las denuncias debe sufrir un proceso de anonimización de los datos antes de ser procesada y publicada. Se desconoce la metodología que implementa la institución, sin embargo, se puede corroborar en el registro de delitos del repositorio del GCBA, que los datos ya se encuentran agregados. Es decir, en lugar de mostrar los hechos delictivos puntuales, se agrupan por franjas horarias y ubicaciones geográficas, mostrando únicamente la ubicación, franja horaria y tipo de crimen del delito ocurrido. Münch, Grosselfinger, Krempel, Hebel, & Arens (2019), describen un caso de implementación de métodos anonimización, enfocado en la protección de datos registrados públicamente en Alemania.

Aunque todos los datos se encuentran en el repositorio del GCBA, agregando datos internos que solo tiene acceso la institución, existen algunas implicaciones de formatos y captura de información que se deben tener en cuenta. En primer lugar, no todos los conjuntos de datos tienen estandarizados los nombres y formatos de campos comunes, esto fue detectado en los campos de latitud y longitud. Los conjuntos de datos se encuentran almacenados en diversos formatos, lo cual complejiza el proceso de integración de datos, sin embargo, dichos aspectos se tratarán en detalle en las secciones posteriores.

Un elemento que es necesario destacar, está relacionado con la procedencia de la captura de información, recordando que el delito contabilizado hace referencia a los crímenes denunciados por los canales regulares. Por lo tanto, estos datos capturan una parte importante de los delitos que ocurren, pero aún existe una proporción de crímenes no denunciados que no están siendo estudiados. Esto es de vital importancia al momento de tomar decisiones de seguridad pública, pudiendo existir variaciones en la diferencia entre lo denunciado y lo ocurrido realmente, dichas variaciones pueden estar vinculadas por múltiples factores, como la distribución geográfica y de entorno de los barrios de CABA.

1.3.3 Problemática tratada

Aprovechando la política de datos abiertos del GCABA, a través de los datos registrados de delitos ocurridos durante los años 2017 y 2019, se implementarán modelos de aprendizaje automático para un conjunto de esquinas de CABA. Esto con el objetivo de predecir la ocurrencia de delitos en el mes de diciembre del año 2019, complementando el análisis con elementos climáticos y de entorno. Definida brevemente y a nivel general la propuesta a desarrollar, se deben evaluar la factibilidad del estudio, en función de la posible utilidad que pueda encontrar la organización en los resultados de esta investigación. Si bien es posible que este estudio sea presentado a algunas de las autoridades del GCABA o de la Policía de CABA, siendo estas las que evalúen dicha factibilidad, se propone plantear algunos casos de utilidad de los resultados del estudio.

El caso más directo corresponde a la planificación de operativos, destinando más recursos humanos y técnicos en áreas donde se pronostique una mayor incidencia de delitos, o pudiendo desplegar una menor cantidad de recursos en zonas donde no sea tan prioritario. Produciendo posiblemente una mejora en el proceso de asignación de recursos, basando en datos y pronósticos las decisiones de planificación de la organización. Un beneficio secundario, pero no menos importante, corresponde a la reducción de costos, ya que de ser efectiva la implementación, se podrían evaluar los recursos ociosos e innecesarios que se estaban asignando.

Además del enfoque predictivo, también se pueden evaluar las variables que más contribuyen en la predicción, en función a esto se pueden plantear hipótesis candidatas a ser corroboradas con futuros estudios de causalidad, o implementando pruebas piloto. Entendiendo que el estudio involucra variables de entorno y climáticas, se pueden evaluar cuáles elementos físicos que aspectos climáticos, contribuyen a aumentar o disminuir las predicciones de los modelos. Se aclara que la causalidad está fuera del alcance de este estudio, por esta razón se menciona siempre la contribución de las variables en la predicción, en lugar de mostrar variables como los causales de la ocurrencia de delitos. La organización también podría enriquecer el estudio con datos internos que no son de acceso público, pudiendo generar mejores predicciones, además de poder implementar enfoques distintos o complementarios.

2 Descripción metodológica

La dualidad tiempo-entorno aplicada sobre la ocurrencia de delitos en CABA, constituye la estructura principal de este estudio. Por lo tanto, surge la necesidad de entender la relación entre el entorno geográfico, los días del año y la ocurrencia de delitos a través de diversos modelos predictivos. Esto con la finalidad de brindar herramientas que permitan ejecutar políticas públicas de seguridad, o para tomar decisiones comerciales en distintos rubros. La operacionalización de los objetivos referentes a la problemática organizacional planteada, requiere la definición de una metodología que permita ordenar de forma consistente el flujo de trabajo a implementar. Además, dicho flujo debe tener la propiedad de ser reproducible, desde el proceso de captura de datos hasta la presentación e interpretación de resultados. La metodología está constituida a nivel general por la recopilación de información, el procesamiento de datos y el análisis de información. En esta sección se presenta la estructura metodológica construida, en la sección posterior se muestra la implementación de esta, analizando además los resultados obtenidos.

2.1 Recopilación de información

La primera parte de la estructura metodología está centrada en los datos, siendo los mismos el punto de partida e insumo principal de la problemática abordada. Debido al gran número de conjuntos de datos involucrados, se presentan diversos formatos, estructuras y consideraciones inherentes a cualquier proyecto basado en datos, Taylor (2016) muestra un resumen de estas consideraciones. Estos elementos originan la necesidad de presentar una descripción exhaustiva de los conjuntos de datos involucrados, además de explicar la forma de tratarlos y capturarlos de manera eficiente.

2.1.1 Descripción de las fuentes de datos

El presente estudio no parte de un conjunto de datos suministrado, terminado o parcialmente elaborado, el mismo debe construirse buscando la información en distintos repositorios. Serán utilizados los datos de registros de delitos ocurridos entre diciembre del 2017 y noviembre del 2019, utilizando este periodo como el conjunto de datos de entrenamiento, el mes de diciembre del 2019 será utilizado como el conjunto de datos de

validación. El principal requerimiento que debe cumplir algún conjunto de datos para ser incorporado en el estudio, consiste en contener la geolocalización o fecha de cada observación.

La ocurrencia de delitos se extrae desde el repositorio público del GCABA, el mismo está agrupado dentro de la temática de seguridad. El propietario es el Ministerio de Justicia y Seguridad. Policía de la Ciudad, la fecha de publicación y de relevamiento de los datos fue el 05/03/2020, hasta el momento la última fecha de actualización fue el 29/06/2020. Los recuentos de delitos se encuentran en archivo con formato ".csv", además tienen una estructura tabular que responde de manera directa a datos de series temporales, pero para este estudio se tomará bajo un enfoque multivariado, con el objetivo de hacer un análisis de regresión y clasificación.

El punto de partida no son los delitos ocurridos, en su lugar son seleccionadas la mayoría de las calles y avenidas de CABA, buscando posteriormente sus intersecciones para encontrar las esquinas de la ciudad, esta información se extrae desde el repositorio de OpenStreetMap. Las intersecciones de calles o avenidas, por sí solas no aportan información relevante para cumplir con los objetivos de la investigación. Para transformar estas ubicaciones información, se requiere calcular el número de delitos ocurridos en las cercanías de cada una de las esquinas dentro de un radio de metros determinado.

Además de los delitos, a cada esquina se les indexan variables de entorno, como por ejemplo el número de comisarías, hoteles, estaciones de subte dentro de un radio de metros. Los delitos son indexados por espacios temporales mensuales, es decir, para una esquina determinada se calcula el número de delitos ocurridos en un mes. Para cada una de estas combinaciones "esquina-mes", se agregan variables climáticas como la temperatura, lluvia y viento. Los datos se extraen desde el repositorio público del GCABA, son seleccionados 35 conjuntos de datos para indexar distintas variables relacionadas con el entorno, por lo tanto, es necesario normalizar, estandarizar y depurar los distintos conjuntos de datos. Todas los atributos generados o agregados serán estudiados a través de estadísticas descriptivas que incluyen correlaciones, medidas de tendencia central, dispersión y estadísticos de distribuciones.

2.1.2 Consideraciones técnicas

Existen algunas consideraciones para tener en cuenta, las mismas tienen un impacto significativo en la integración de la información y en la reproducibilidad del estudio. Un aspecto favorable de las fuentes de datos viene dado por el tipo de repositorio, siendo de acceso público y mantenido por un organismo oficial del estado argentino, lo que brinda una mayor confiabilidad con respecto a la veracidad y robustez de la información. El aspecto más importante está relacionado con los distintos formatos, los archivos tienen terminaciones ".csv" y ". shape", siendo archivos planos y objetos geoespaciales respectivamente.

La geolocalización es un elemento fundamental, siendo representado por la longitud y latitud en cada conjunto de datos. Este proceso debe automatizarse, debido a que se debe interactuar con más de treinta conjuntos de datos. Sin embargo, existen problemas de consistencia con el nombre de los campos, nombrando los atributos latitud y longitud de manera distinta, inclusive con formatos distintos en algunas situaciones. La manipulación de los datos geoespaciales requiere un tratamiento distinto, transformando los mismos en una estructura tabular, esto será explicado en la sección de procesamiento de la información. Kuhn & Johnson (2019), exponen el tratamiento común a los valores faltantes y duplicados presentes en los proyectos basados en datos, sin embargo, para este estudio es necesario tener un apoyo de herramientas visuales basadas en mapas a través de la geolocalización. Debido a que un elemento de entorno puede tener legalmente distintas estructuras, pero físicamente la misma ubicación, sin embargo, a fines de este estudio debe contabilizarse como una ubicación única.

2.1.3 Captura de datos

Los datos se capturan estableciendo conexiones con los repositorios, para posteriormente leer la información directamente sin necesidad de descargar individualmente cada uno de los archivos. Los objetos geoespaciales requieren en algunos casos la descarga de la información individual, debido a la estructura interna de los archivos, estando los mismos contenidos en carpetas con niveles múltiples. La información referente al clima presenta una estructura interna distinta al resto, por lo tanto, requiere algunos pasos de preprocesamiento adicionales. Un aspecto central para considerar consiste en la dependencia

de la estructura de los datos con el repositorio del GCABA. Si ocurre algún cambio en las direcciones de los repositorios, provocando por ejemplo, debido a alguna modificación en el diseño de la interfaz web o en el almacenamiento de la información, deben modificarse las rutas de conexión. Además de las direcciones, posibles cambios en los formatos de información también pueden perjudicar la reproducibilidad del estudio, por lo tanto, para ambos escenarios es necesario verificar la existencia de cambios relevantes recientes en los repositorios.

2.2 Procesamiento de la información

El aspecto más desafiante del presente estudio reside en el diseño y planificación de la limpieza, indexación, creación y unificación de variables en un conjunto de datos definitivo, acorde al tipo de problema organizacional seleccionado. Requiriendo la creación de algoritmos y funciones propias, que permitieran automatizar el flujo de trabajo, en algunos casos además eran una condición necesaria para poder continuar con la investigación. En la siguiente sección se expone en detalle la etapa de limpieza y transformación de datos, indexación de variables, y la creación unificación de toda la información en una estructura tabular definitiva.

2.2.1 Limpieza y transformación de datos

La longitud y latitud de los delitos ocurridos, así como de los elementos de entorno no pueden presentar valores vacíos, por lo tanto, serán excluidos aquellos registros que no cumplan con esta condición. Para estos dos atributos, se estandariza el nombre de ambas variables en todos los conjuntos de datos, con el objetivo de facilitar la automatización del flujo de trabajo. Finalmente, los delitos están almacenados en archivos separados de acuerdo con el año, es necesario filtrar y posteriormente unificar las fuentes de datos de acuerdo con el periodo de estudio establecido.

Las calles y avenidas de CABA deben transformarse en datos tabulares para que puedan ser utilizadas en el presente estudio, debido a que estos están almacenados como objetos geoespaciales. El único dato de interés de estos objetos consiste en la intersección de los mismos objetos, es decir, la intersección de las distintas calles y avenidas de la ciudad. Los puntos de intersección resultantes hacen referencias a las esquinas, de estas esquinas se

extrae la longitud y latitud, siendo estos los campos claves para poder incorporar esta información al estudio.

El proceso de transformación de información para llegar al conjunto de datos definitivo, fue inspirado en gran medida sobre el estudio que presentaron Lin, Yen, & Yu (2018). Estos investigadores incorporan el factor temporal a los elementos demográficas y de entorno, pero evaluando los tres factores de manera simultánea para construir un modelo de aprendizaje automático, con el objetivo de capturar desde una perspectiva más completa el entorno criminal. El análisis fue realizado en una ciudad de Taiwán, dividiendo el mapa de la ciudad en cuadrículas de 200 x 200 metros, posteriormente calculaban el número de delitos ocurridos en el polígono. Sin embargo, el presente estudio toma este principio, pero en lugar de cuadrículas se seleccionan esquinas, para luego calcular los delitos que ocurrieron en un radio de metros de cercanía. El enfoque que se utiliza en este proyecto se considera mucho más realista, ya que a priori las esquinas están localizadas en zonas habitadas o mínimamente transitadas, el enfoque de cuadrículas contabiliza polígonos únicamente por estar dentro del espacio geográfico de la ciudad.

2.2.2 Indexación de variables

La agregación de variables debe cumplir una serie de pasos de transformación de datos, partiendo por seleccionar campos claves, sobre los cuales se efectuarán las indexaciones del resto de las variables. El punto de partida será la geolocalización obtenida de las esquinas de la ciudad, sin embargo, primero debe seleccionarse el número de esquinas óptimas a utilizar. Con el objetivo de no contabilizar un delito de manera duplicada en distintas esquinas, se propone seleccionar las esquinas que estén a una distancia de más de 250 metros entre sí, para esto implementarse un algoritmo de optimización.

Definidas las esquinas óptimas, se procede a calcular el número de delitos ocurridos en cada esquina dentro de un radio de 150 metros, realizando el cálculo de manera diferenciada por meses, partiendo de diciembre del año 2017 a diciembre del 2019. Posteriormente para cada mes se agregan las variables climáticas correspondientes, presentando por cada fila una esquina con campos que representan los delitos ocurridos, e indicadores climáticos en cada uno de los meses. Como último paso de depuración, son excluidas aquellas esquinas no presenten delitos en 16 o más de los 24 meses de estudio.

El siguiente grupo de variables a indexar corresponde a los elementos de entorno, para ello se debe calcular el número de observaciones de cada variable de entorno, que están dentro de un radio inferior a 250 metros de cada esquina. Esto requiere la construcción de un algoritmo, el mismo debe automatizar el cálculo de distancias y el conteo de eventos. Siguiendo la metodología planteada hasta este punto, puede generarse un conjunto de datos parcialmente definitivo, faltando únicamente una última transformación para iniciar la etapa de modelado.

2.2.3 Conjunto de datos definitivo

Esta sección de la etapa de transformación de datos es crucial y determinante, con respecto a poder implementar modelos predictivos en la problemática organizacional. El siguiente paso consiste en aplicar una estrategia que en este estudio será denominada como "ventana deslizante", sin embargo, la idea fue tomada del estudio publicado por Lin et al. (2018). La misma consiste en seleccionar el mes intermedio "t" de todo el periodo de estudio. Posteriormente, se calcula el número de delitos ocurridos para cada esquina en el mes anterior, en los últimos 3 meses, 6 meses, 12 meses y el mismo mes del año anterior. La variable objetivo sería el número de delitos ocurridos en el mes de estudio, además este mismo procedimiento es aplicado con las variables climáticas.

El procedimiento anterior es replicado de manera móvil, es decir, rodando la ventana temporal desde "t+1" hasta "n", siendo "n" el último mes de estudio. Estos "n-t" conjuntos de datos se unifican en un conjunto de datos único, representando el mismo el conjunto de entrenamiento, el conjunto número "n" hace referencia a los datos sobre los cuales se hartan las predicciones finales. Para este estudio, este último conjunto de datos está representado por el mes de diciembre del año 2019. Todo este procedimiento requiere ser automatizado a través de la creación de un algoritmo, el mismo es recomendable que esté estructurado como una función, con esto se facilita la legibilidad y reproducibilidad del estudio.

2.3 Análisis de la información

El primer paso de esta etapa se define como el análisis exploratorio de datos, haciendo referencia a la observación de las distribuciones, presencia de valores atípicos y estructura de las variables del conjunto de datos. Posteriormente se realiza el análisis en función de la variable objetivo, identificando patrones o relaciones simples con la ocurrencia de delitos.

A través de esta primera etapa se pueden identificar variables altamente correlacionadas, atributos con baja varianza o posibles creaciones de variables a partir de las ya existentes. Un análisis individual de la variable objetivo también es pertinente, identificando posibles sesgos o transformaciones de esta, pudiendo dar flexibilidad al estudio, en el sentido de permitir abordar la problemática desde un enfoque de regresión y clasificación. En los siguientes apartados se tratará en detalle la estrategia de modelación, definiendo los diferentes enfoques a aplicar, a su vez se hará un breve análisis acerca de los modelos a implementar y las métricas de evaluación seleccionadas.

2.3.1 Estrategia de modelación

Desde el punto de vista de la implementación de modelos predictivos, la problemática organizacional será abordada a través de tres enfoques. Una primera aproximación corresponde al establecimiento de un modelo base, con esto se implementa un modelo poco complejo que establezca el límite inferior de rendimiento. Para esto, se aplicará un modelo de series de tiempo, que considerará únicamente la ocurrencia pasada de delitos en cada esquina, haciendo el pronóstico para el mes de diciembre del año 2019. El objetivo es contrastar la siguiente hipótesis, un modelo de aprendizaje automático que considera el histórico de delitos ocurridos, el clima y el entorno de las esquinas, debería poder superar el rendimiento de un modelo de serie de tiempo que solo considera el delito como variable.

El segundo enfoque, corresponde a la aplicación de modelos de aprendizaje automático con el objetivo de predecir el número de delitos. Es decir, para cada esquina se busca pronosticar el número de delitos que ocurrirán, este enfoque se puede contrastar directamente con el modelo base obtenido. Se requiere previamente analizarla distribución de la variable objetivo, con la finalidad de evaluar la aplicación alguna transformación en la misma, y verificar si dicha transformación mejora el rendimiento del modelo.

El tercer enfoque, corresponde a la aplicación de modelos de aprendizaje automático con el objetivo de predecir el nivel de riesgo de las esquinas. Previamente se debe discretizar la variable objetivo, requiriendo con esto analizar la distribución de los delitos ocurridos. Para el enfoque de clasificación, será implementado un modelo de calcificación de múltiples clases, siendo contrastado con la aplicación de dos modelos de clasificación binarios.

Para los enfoques de aprendizaje automático, se aplicará la misma estrategia de partición de datos. En este estudio no hace falta realizar una partición aleatoria en datos de

entrenamiento y validación, debido a que ambos conjuntos de datos se obtienen a través de las fechas. El balance de las clases de la variable objetivo es un tema central y que debe ser verificado, requiriendo según sea el caso aplicar alguna técnica especial, como el "Under-Over-sampling" detallado en Krawczyk (2016).

2.3.2 Modelos implementados

Para estimar correctamente el error de ajuste en los datos de entrenamiento, fue seleccionada una estrategia de validación cruzada. Kumar (2020) expone que la misma consiste en dividir las observaciones de entrenamiento en "k" conjuntos, de aproximadamente el mismo tamaño, luego el modelo se entrena con todas las particiones menos una partición y se prueba el ajuste con la partición faltante. Este método se repite hasta que todas las particiones hayan sido usadas para evaluar el ajuste del modelo, el promedio de todas las estimaciones tiende a converger con el valor real del error sobre los datos de prueba.

Los modelos seleccionados para este estudio se dividen en dos tipos, modelos de series de tiempo y modelos de aprendizaje automático. Con respecto al primer tipo, será utilizado el modelo "Croton", siendo un modelo especial para regresiones de conteo, que toma en cuenta únicamente el histórico de la ocurrencia de delitos. Para los modelos de aprendizaje automático, se implementaron principalmente dos modelos basados en árboles, siendo el *Extreme Gradient Boosting* (XGB) y el *Light Gradient Boosting Machine* (LightGBM).

Croston

La distribución de ocurrencia de delitos en las esquinas de CABA, no presentan un espacio muestral continuo en la gran mayoría de las series, en su lugar vienen en forma de recuentos. En estas situaciones, resulta conveniente la aplicación de métodos de pronóstico de conteos, en este estudio será utilizado el método de método de "Croston". Hyndman, & Athanasopoulos (2018) afirman:

> Con el método de Croston, construimos dos nuevas series a partir de nuestra serie de tiempo original al señalar qué períodos de tiempo contienen valores cero y qué períodos contienen valores distintos de cero. Sea q_i la i -ésima cantidad distinta de cero, y sea a_i el tiempo entre q_{i-1} y q_i. El método de Croston implica pronósticos de

suavizado exponencial simples separados en las dos nuevas series “a” y “q”. Debido a que el método generalmente se aplica a series de tiempo de demanda de artículos, “q” a menudo se denomina demanda y “a” tiempo entre llegadas (sección 12.2 Time series of counts, párr.4).

Existen métodos más avanzados para estudiar este tipo de fenómenos, pero están fuera del alcance de este proyecto, una introducción acerca del tema puede encontrarse en el artículo publicado por Christou & Fokianos (2015).

***Extreme Gradient Boosting* (XGB)**

Bisong (2019) refiere que es una implementación mejorada del algoritmo de gradiente de árboles reforzados. Este algoritmo se entrena combinando las predicciones de un conjunto de modelos más simples y débiles, pero de forma secuencial en lugar de una aplicación simultánea de todos los árboles. Este modelo será implementado para el enfoque de regresión, contrastando posteriormente con el enfoque de series de tiempo. Chen & Guestrin (2016), presentaron una publicación donde se pueden verificar los detalles técnicos del modelo, además de las diversas implementaciones en distintos lenguajes de programación.

***Light Gradient Boosting Machine* (LightGBM)**

Ruiz (2020), describe este algoritmo como un marco que implementa la técnica de gradient boosting , usando algoritmos basados en árboles con la eficiencia como objetivo principal. A diferencia del XGB, que utiliza algoritmos basados en la clasificación previa, LightGBM usa algoritmos basados en histogramas que agilizan el entrenamiento. Está optimizado para que el árbol crezca en la dirección de los mejores nodos, contribuyendo esto a un mejor manejo de los recursos en memoria. Este algoritmo será implementado únicamente para el caso de clasificación, tanto para la predicción multiclase como en el caso de clasificadores binarios individuales. Ke et al. (2017), presentaron el LightGBM, donde se pueden observar en detalle todos los aspectos técnicos de su funcionamiento.

Los modelos de aprendizaje automático que se seleccionaron tienen un conjunto de parámetros conocidos como hiperparámetros, Obi (2019) presenta una introducción al tema. Malik (2020) expone que la forma más común de encontrar los valores óptimos de los hiperparámetros, es a través de técnicas de búsquedas de cuadrícula. Estos métodos

esencialmente buscan crear un número limitado de combinaciones para los valores de los hiperparámetros, luego para cada combinación el modelo es entrenado en el caso de este estudio con la estrategia de validación cruzada planteada. Finalmente es seleccionada la combinación de hiperparámetros que tenga un mejor rendimiento según la métrica de error seleccionada.

Amat (2020) sostiene que estas técnicas tienen un alto costo computacional, debido a que se deben probar todas las combinaciones suministradas al modelo, además la mayoría no evalúan secuencialmente el error obtenido en cada combinación ajustada. Esto ocasiona que se prueben valores de parámetros innecesarios, para este estudio se aplicará un método de ajuste de hiperparámetros denominado "optimización bayesiana". Según Snoek, Larochelle, & Adams (2012), consiste en construir un modelo probabilístico fijando como función objetivo la métrica de ajuste del modelo, el objetivo es buscar los parámetros óptimos de manera iterativa, seleccionando solo las combinaciones de parámetros que se estima que puedan mejorar la métrica de error.

2.3.3 Métricas de evaluación

Finalizada la etapa de entrenamiento de modelos, aunque también durante el ajuste de hiperparámetros, se requiere definir una serie de métricas que permitan hacer comparaciones entre modelos. Handelman et al. (2019) definen un conjunto de métricas de evaluación, dividiendo las mismas según sea el problema de clasificación o regresión. Entre las métricas de regresión, serán utilizadas el "rmse", "rsq" y "mae", para los casos de calcificación se utilizarán la "auc" y "matriz de confusión". El "rmse" hace referencia al error cuadrático medio, el "mae" al error absoluto medio y "rsq" al "r" cuadrado. Según el tipo de problema y la naturaleza del fenómeno de estudio, puede ser más conveniente el uso de alguna de estas métricas. Hale (2020) describe de forma detallada las definiciones de ambos indicadores, además de la conveniencia de uso de cada uno.

Los modelos de clasificación serán evaluados a través de dos métricas, el AUC y matrices de confusión. Las siglas de AUC significan área bajo la curva "ROC", Lantz (2013) plantea que esta métrica "examina la compensación entre la detección de verdaderos positivos, mientras se evitan los falsos positivos." (p.313). La justificación de la utilización de estas métricas será explicada en la sección de implementación, la segunda métrica que se utilizará para los modelos de clasificación corresponde a la matriz de confusión. Lantz

(2013) afirma “Una matriz de confusión es una tabla que categoriza las predicciones según si coinciden con el valor real de los datos. Una de las dimensiones de la tabla indica las posibles categorías de valores pronosticados, mientras que la otra dimensión indica lo mismo para los valores reales.” (p.298). Esta métrica será utilizada para la clasificación de clases múltiples y binaria, teniendo su aplicación sobre los datos de prueba. Posteriormente serán agregados costes unitarios a los tipos de errores, con la finalidad de diferenciar los tipos de errores según la lógica organización de políticas de seguridad pública.

3 Implementación

Definida la problemática organizacional, identificadas las fuentes de datos y detallado el proceso de limpieza, transformación y modelado de información, se procede a operacionalizar la metodología planteada. En esta sección se narrará de forma cronológica cada una de las etapas de la implementación, combinando la explicación en prosa con código de programación, salidas, gráficos, tablas o diagramas según corresponda. Se parte con el proceso de implementación de modelos, seguida de la presentación y visualización de resultados definitivos, finalmente se plantea una metodología hipotética para llevar a cabo la implementación dentro de la organización.

3.1 Captura, limpieza y transformación de datos

Este subapartado corresponde al más extenso de esta sección, ya que se operacionalización la mayor parte del flujo de trabajo del proyecto. Se parte del proceso de captura y limpieza, detallando todas las consideraciones importantes en el flujo de trabajo. Seguidamente se realiza el análisis exploratorio de datos y la transformación de información, para finalmente definir el conjunto de datos definitivo que será incorporado en los modelos predictivos.

3.1.1 Recopilación de datos

Todos los datos provienen del repositorio de datos del GCBA, por lo tanto, el proceso de captura de la información se ejecuta de forma similar para todos los conjuntos de datos. Todas las muestras de código están escritas en R, las mismas pueden encontrarse en el Apéndice, buscando la referencia del fragmento de código citada en el texto. En el Apéndice,

también se podrá consultar la dirección del repositorio del paquete “raFunctions”, creado para almacenar las funciones creadas, además de la ubicación de la versión extendidad el estudio.

Tabla 1

Formato tabular de los delitos ocurridos

id	fecha	franja_horaria	tipo_delito	subtipo_delito	cantidad_registrada	comuna	barrio	lat	long
374556	2019-01-01	12	Lesiones	Siniestro Vial	1	4	Nueva Pompeya	-34.64839	-58.40475
426152	2019-01-01	6	Robo (con violencia)		1	9	Liniers	-34.64983	-58.51386
371604	2019-01-01	8	Lesiones	Siniestro Vial	1	15	Chacarita	-34.58811	-58.43939
425359	2019-01-01	16	Hurto (sin violencia)	Hurto Automotor	1	10	Floresta	-34.63188	-58.48398
437571	2019-01-01	2	Robo (con violencia)	Robo Automotor	1	4	Parque Patricios	-34.63316	-58.39712
431424	2019-01-01	16	Robo (con violencia)		1	4	Boca	-34.63449	-58.35938

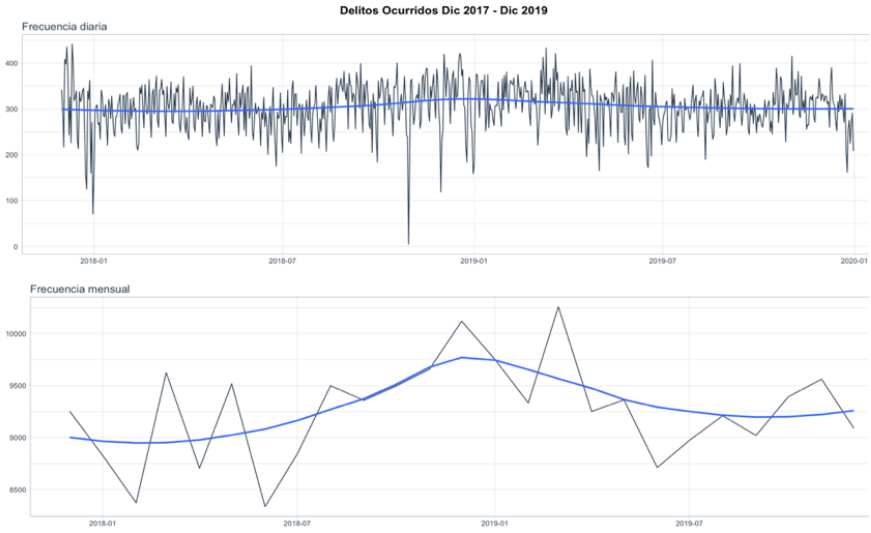
Nota. Fuente: Elaboración propia

En el Fragmento 1, se observa la sintaxis para la descarga de la información , aunque la mayoría de los datos se pueden descargar con la sintaxis presentada, algunos requieren una sintaxis especial, tal es el caso de los objetos geoespaciales, en el Fragmento 2 se muestra la sintaxis para su captura. En la Tabla 1, se presenta una muestra de los datos de los delitos ocurridos, observando los atributos de geolocalización, fecha y tipo de delito. El siguiente paso consiste en estandarizar y depurar los conjuntos de datos, esto con la finalidad de poder automatizar el flujo de trabajo. En el Fragmento 3 se muestra el proceso de limpieza de datos con algunas de las consideraciones más importantes, establecidas en la sección de metodología, por ejemplo, lo referente a formatos, valores faltantes, nombres de columnas y ubicaciones físicas de los elementos. Los datos de delitos ocurridos se unifican y se analizan los tipos y subtipos de delitos, permitiendo analizar el origen de cada denuncia. Para este estudio serán excluidos los crímenes relacionados con siniestros viales y lesiones.

3.1.2 Análisis exploratorio de datos

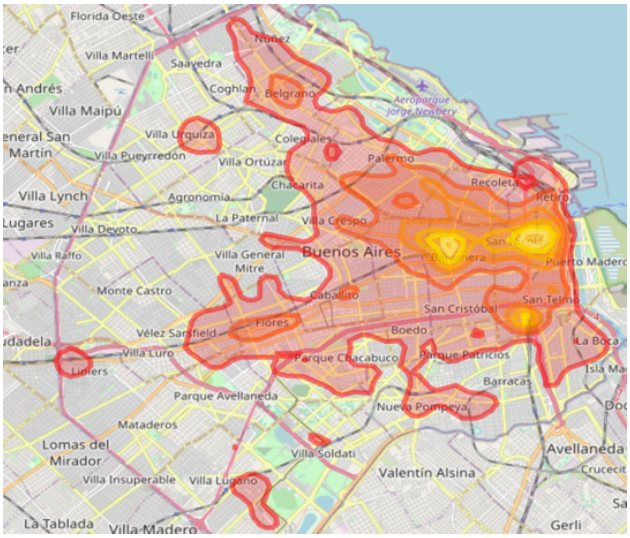
Establecidos los delitos definitivos con los cuales se trabajará, además de agrupar los mismos por franja horaria, se procede a realizar el análisis exploratorio de datos. Se parte por el análisis temporal de los delitos, permitiendo evaluar posibles factores estacionales durante el periodo de estudio. Finalmente se procede a realizar el análisis geográfico, analizando la distribución de los delitos en el mapa de CABA.

Figura 4. *Análisis temporal de los delitos ocurridos*



Fuente: Elaboración propia

Figura 5. *Análisis geográfico de los delitos ocurridos*



Fuente: Elaboración propia / OpenStreetMap

En la Figura 4, se observa un patrón estacional en la ocurrencia de delitos, presentando saltos y caídas en los mismos meses durante los años estudiados. En la Figura 5, se muestra geográficamente una concentración mayor de delitos en la zona este y central de la ciudad, las zonas que no están marcadas en color rojo también registran delitos, pero con una frecuencia bastante baja. Es importante destacar una de las limitaciones planteadas sobre el estudio, las zonas periféricas de la ciudad probablemente registren tasas similares o inferiores a la media de la ciudad, sin embargo, solo se cuenta con los delitos denunciados y registrados, pudiendo ser una aproximación alejada de la realidad para estas zonas. En el

Fragmento 4 se puede observar el proceso de transformación que se aplica sobre las variables climáticas, ya definidas en la sección de metodología.

Figura 6. *Calles y avenidas de CABA*

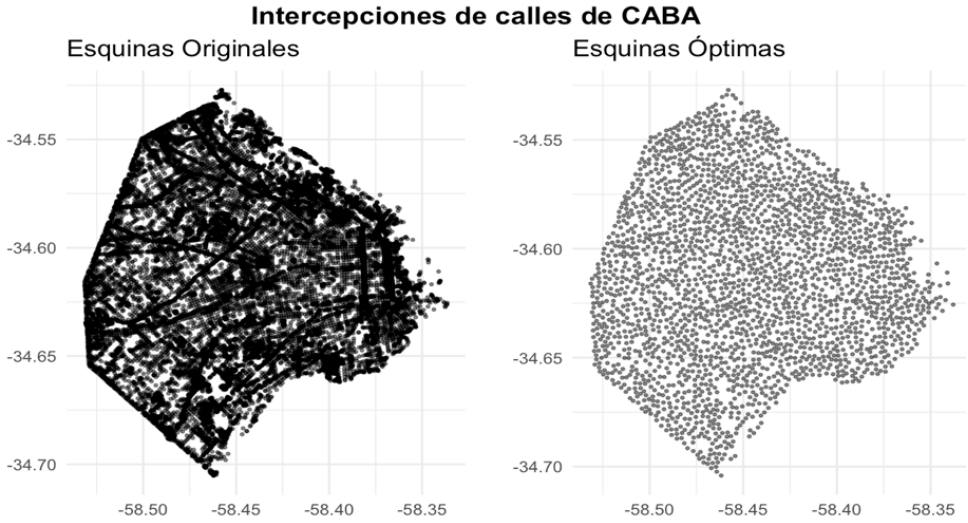
Fuente: Elaboración propia / OpenStreetMap

En la Figura 6 se presentan las calles de CABA extraídas del repositorio de OpenStreetMap, este objeto espacial es junto a los delitos ocurridos el principal insumo del estudio. El objetivo con las calles reside en extraer las intersecciones o esquinas de la ciudad, sin embargo, se aprecia que muchas esquinas están muy próximas entre sí. En este caso, se propone aplicar el algoritmo de optimización definido en la sección de metodología, Rabosky et al. (2016) crearon una función de optimizacion en el lenguaje R, que se adapta al requerimiento del estudio, la misma sera utilizada a traves del paquete "rangeBuilder".

3.1.3 Conjunto de datos definitivo

La justificación para eliminar esquinas que estén muy próximas tiene que ver con el conteo de delitos, la idea es contabilizar los delitos que ocurrieron en las cercanías de las esquinas, durante distintos periodos de tiempo. Por lo tanto, si son seleccionadas esquinas que están muy cercanas, un delito puede ser registrado en más de una esquina. A continuación, en la Figura 7 se muestran las esquinas sobre el mapa de CABA, comparando las intercepciones originales y las resultantes luego de la aplicación del algoritmo de selección de esquinas.

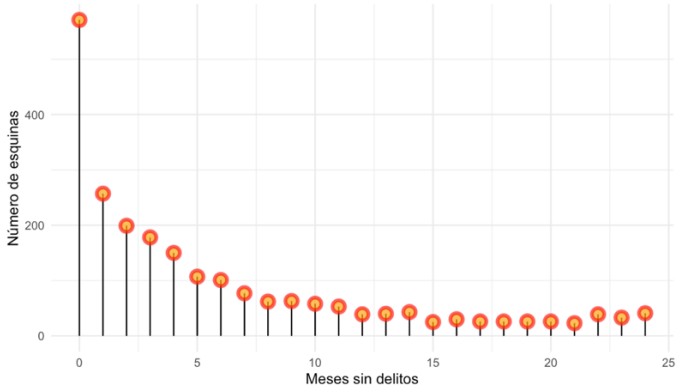
Figura 7. *Comparación de las esquinas de CABA antes y después de la aplicación del algoritmo*



Fuente: Elaboración propia

Originalmente se habían encontrado 22.048 esquinas, luego del proceso de optimización se obtuvieron 2.417 esquinas. El siguiente paso consiste en asignar a cada delito la esquina que esté dentro de una distancia de 150 metros, según el criterio establecido en la sección de metodología. Se estudiaron un total de 231.514 delitos, de los cuales 182.346 tienen asociada una esquina, a través de la fecha, se indexan las cuatro variables climáticas. Para observar este proceso de transformación, se puede revisar el Fragmento 5, creando un conjunto de datos donde cada fila representa una esquina. Las esquinas presentan 24 meses de observación, pero no en todos los meses registran delitos ocurridos, se debe prestar atención aquellas esquinas con ocurrencia de delitos en pocos meses observados.

Figura 8. *Distribución de esquinas con meses sin delitos registrados*



Fuente: Elaboración propia

En la Figura 8 se visualiza la distribución de esquinas con meses sin delitos registrados, las esquinas que registraron delitos en todos los meses totalizan 571. El 88% de las esquinas no mostraron delitos como máximo en 15 de los 24 meses analizados, aquellas esquinas que no presentan delitos en más de 15 meses se excluyen del estudio. Esto debido a que pueden acarrear un desbalance excesivo sobre la variable objetivo, realizando la depuración quedan 2023 esquinas definitivas. Para indexar todas las variables del entorno se implementa el algoritmo de distancias mencionado en la metodología, la aplicación de la misma puede observarse en el Fragmento 6.

La última transformación necesaria sobre los datos consiste en la aplicación de la técnica de "ventana deslizante" explicada en el apartado de metodología. Para esto fue creada una función nombrada como "sliding_window", la misma permite aplicar la metodología al grupo de variables que se requiera. Además, permite parametrizar el inicio del mes de inicio y la extensión de la aplicación de la técnica, en el Fragmento 7 se muestra su implementación. En esencia ya se obtiene el conjunto de datos definitivo para iniciar la etapa de modelado, se cuenta con 69 columnas y 26.299 observaciones. Cada observación hace referencia a una esquina vista desde un mes determinado, las columnas agrupan las variables de entorno, clima y el historial de delitos ocurridos.

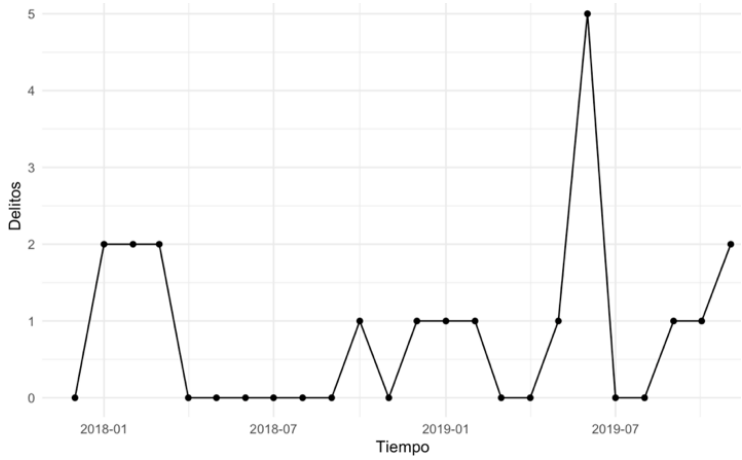
3.2 Modelado

En esta etapa, se implementan los algoritmos de aprendizaje automático, siguiendo los tres enfoques planteados. Se inicia estableciendo un modelo base a través de un modelo de series de tiempo, seguidamente se implementan los enfoques de regresión y clasificación. Finalmente, se realiza la revisión a través de distintas métricas de evaluación y se plantea también la justificación para la elección de estas. A excepción del enfoque de series de tiempo, todos los modelos se implementarán en el marco de aprendizaje automático Tidymodels, perteneciente al lenguaje R, la documentación oficial puede consultarse en Kuhn, & Wickham (2020). Se aplicará una validación cruzada de diez particiones utilizando la técnica "k-Fold-Cross-Validation (CV)", adicionalmente los hiperparámetros de los modelos se ajustarán a través de una optimización bayesiana, con un máximo de 15 iteraciones y una restricción de finalización de 10 iteraciones sin mejora.

Como modelo base para el establecimiento del límite inferior de rendimiento de los modelos, se implementará el modelo de series temporales Croston. Siendo un modelo

sencillo y poco complejo que solo toma en cuenta el conteo de delitos, está diseñado para abordar problemas de predicciones de series intermitentes. En la Figura 9 se ilustra el justificativo para la elección de este modelo.

Figura 9. *Registro histórico de delitos de la esquina "999"*



Fuente: Elaboración propia

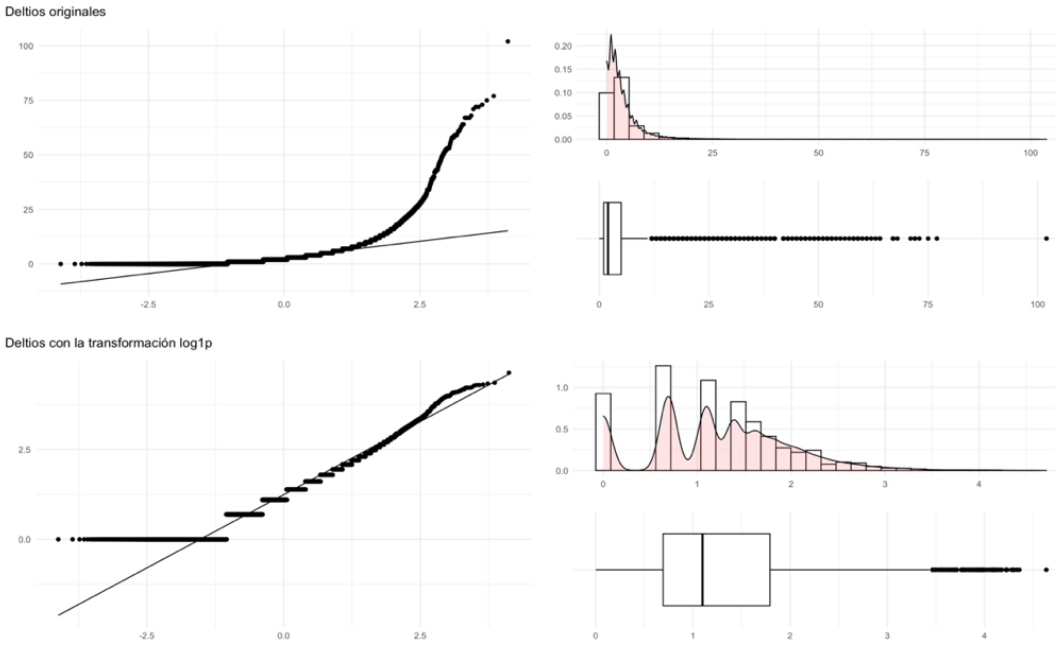
Esta serie de tiempo representa el histórico de delitos ocurridos para una esquina de CABA, se observan muchos ceros presentes en los datos. Este tipo de series no suelen presentar una tendencia definida, además de ser bastante volátiles, lo que dificulta realizar pronósticos con métodos convencionales. Según Hyndman, & Athanasopoulos (2018), aunque el modelo Croston fue diseñado originalmente para el pronóstico de demanda en un período determinado, se usa habitualmente para tratar este tipo de series. Grogan (2020) muestra una implementación de este modelo en el lenguaje R, además de explicaciones con casos de uso. Este modelo se ajusta para todas las esquinas, haciendo la predicción para el mes de validación, el ajuste se realiza con los parámetros por defecto. Finalmente se contrastará con el cálculo de la media de las series, y el modelo de regresión de la siguiente sección, evaluando cual de los modelos tienen un menor error de predicción, en el Fragmento 8 puede observarse la implementación.

3.2.1 Regresión

Los delitos ocurridos representan a la variable objetivo, analizando la distribución de la variable se observa un fuerte sesgo. Boehmke, & Greenwell (2019) exponen esta temática, en este tipo de situaciones suele ser una práctica común aplicar transformaciones sobre la variable, para minimizar la asimetría de la respuesta, con el objetivo de intentar mejorar el

poder predictivo de los modelos. Se propone la aplicación de una transformación logarítmica, con la adición de "1" para evitar el caso de aplicar el logaritmo a cero, A continuación, en la Figura 10 se muestra la variable original y su posterior transformación.

Figura 10. *Distribución de la ocurrencia de delitos y su posterior transformación logarítmica*



Fuente: Elaboración propia

Aunque la transformación minimizó la asimetría de la variable, no se percibe una normalización de esta. Las transformaciones pueden disminuir el sesgo de la variable, pero esto no implica que las predicciones de los modelos tengan menos errores. Esto se evidencia en este estudio, donde el rendimiento de los modelos con la transformación logarítmica no se vio reflejado en ninguna mejoría de las métricas de evaluación. Por lo tanto, no se hará ninguna transformación sobre la variable respuesta, en el Fragmento 9 se presenta el procesamiento de datos implementado en el marco de Tidymodels.

El preprocesamiento se puede crear en un objeto denominado "récipe", el mismo permite generar cualquier tipo de ingeniería de características, entrenarlas y aplicarla sobre cualquier conjunto de datos. A través del récipe no hace falta hacer modificaciones sobre los datos originales. El primer paso consiste en definir la variable a predecir, los predictores y los datos a utilizar, luego a través de un conjunto de *step functions* se aplican las transformaciones requeridas sobre los datos. El "mae" es la métrica que se optimizará en la búsqueda de hiperparámetros, esto debido a que se sabe de la existencia de esquinas que

presentan una alta volatilidad en el registro de delitos. Aunque son pocas las observaciones que presentan esta característica, afectan de manera severa la evaluación global de los modelos, por lo tanto, no interesa penalizar excesivamente los errores en estos datos atípicos. Se seleccionó el modelo XGB, en la tabla de hiperparámetros del Apéndice, se puede consultar el valor y la descripción de los hiperparámetros ajustados.

Tabla 2

Errores de validación cruzada XGB

.metric	.estimator	mean	n	std_err	modelo
mae	standard	1.562	10	0.0118	XGB
rmse	standard	2.330	10	0.0352	XGB
rsq	standard	0.779	10	0.0069	XGB

Nota. Fuente: Elaboración propia

Tabla 3

Errores sobre los datos de validación XGB

.metric	.estimator	.estimate
rmse	standard	2.1774
rsq	standard	0.8026
mae	standard	1.5034

Nota. Fuente: Elaboración propia

En la Tabla 2 se reflejan los errores sobre las particiones de validación cruzada, la Tabla 3 se refiere a los errores sobre el mes de validación. Se observa una mejoría de las tres métricas al comparar ambos errores, los cual es un signo de que el modelo a priori no está sobreajustado. La diferencia existente entre las métricas rmse y mae, se debe principalmente a esquinas que presentaron saltos abruptos en el registro de delitos para el mes de diciembre.

3.2.2 Clasificación

El enfoque de clasificación será abordado y contrastado desde dos perspectivas, contraponiendo las mismas para evaluar cual resulta una mejor aproximación a la problemática organización que se estudia. Se implementará el modelo LightGBM y se utilizará la misma receta de preprocesamiento aplicada en el enfoque de regresión, aunque con una ligera modificación. Se debe agregar un paso extra al final de la receta, esto con el objetivo de discretizar los delitos ocurridos, a continuación, se muestra la discretización para comparar ambos enfoques.

Tabla 4

Criterios de discretización multiclase

Delitos	Frecuencia	Condición
Bajo	9.190	Si los delitos son <= 1 ->> Bajo
Medio	10.511	Si los delitos son <= 4 y >= 2 ->> Medio
Alto	6.598	Si los delitos son >= 5 ->> Alto

Nota. Fuente: Elaboración propia

En la Tabla 4 se agrupan las esquinas según el nivel de riesgo, distribuido en tres niveles, el objetivo reside en implementar modelos que puedan discriminar las tres categorías. El enfoque de clasificación se implementará mediante el LightGBM, Nikhilesh (2019) plantea que es un algoritmo mucho más rápido que el XGB, permitiendo probar una mayor combinación de transformaciones en menos tiempo. Esto abre la posibilidad de plantear una problemática de clasificación multiclase, y posteriormente una clasificación binaria múltiple.

Clasificación multiclase

Una vez discretizada la variable objetivo, se prueba realizar un balanceo de clases a través de un método de sobremuestreo, con la finalidad de mejorar las predicciones. Se implementa el algoritmo de *Synthetic Minority Oversampling Technique* (SMOTE), para mayores detalles consultar a Chawla, Bowyer, Hall, & Kegelmeyer (2002). Sin embargo, las observaciones sintéticas generadas no mejoraron el rendimiento del modelo, por lo tanto, se ajustó el modelo definitivo sin balancear las clases.

La métrica optimizada para la búsqueda de hiperparámetros corresponde a la "roc_auc" en la implementación Hand-Till, debido a que se cuenta con una distribución de clases desequilibrada. Kuhn, & Wickham (2020) afirman que la implementación dentro del marco "Tidymodels", este método no hace suposiciones específicas sobre la distribución de clases o los costos de clasificación erróneas, es decir, es insensible a distribuciones de clases. El objetivo es realizar una evaluación posterior de los modelos a través de matrices de confusión, agregando costos a los errores. Para más detalles, Hand, & Till (2001) exponen la implementación del método.

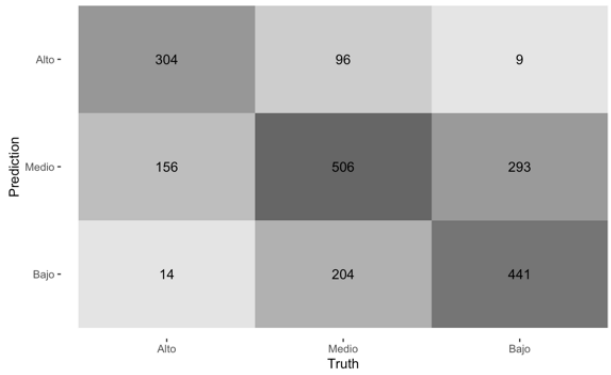
Tabla 5

Rendimiento del modelo de clasificación multiclase

.metric	.estimator	valor	n	std_err	modelo	Error
roc_auc	hand_till	0.8113618	10	0.002411511	LightGBM	validación cruzada
roc_auc	hand_till	0.8147452	NA	NA	LightGBM	Diciembre 2019

Nota. Fuente: Elaboración propia

Figura 11. *Matriz de confusión del modelo multiclase*



Fuente: Elaboración propia

En la Tabla 5, se observa un ligero aumento de la métrica de evaluación entre la validación cruzada y los datos de prueba, significando que el modelo tuvo una buena generalización de los datos. La Figura 11 corresponde a la matriz de confusión, muestra que los errores entre las categorías extremas son bastante bajos, siendo de 14 y 9 errores en ambos pares. En la sección de presentación de resultados, se agregan costos a las matrices resultantes, sin embargo, se percibe que el modelo logra discriminar satisfactoriamente una esquina peligrosa de una de bajo riesgo. Los errores se acumulan principalmente en la categoría intermedia, siendo más difícil discriminar en muchos casos donde el número de delitos está ubicado en los límites de las categorías, aunque se destaca que a pesar de lo mencionado se logró clasificar correctamente al 63% de esta categoría.

Clasificación binaria múltiple

El objetivo de este enfoque es implementar dos clasificadores binarios, para luego combinar las predicciones. Se aplica el mismo proceso de discretización utilizado en el enfoque multiclase, pero generando las categorías “Alto” y “Medio_Bajo” para un primer clasificador, “Bajo” y “Medio_Alto” para un segundo clasificador. Con esto se intenta

capturar el orden jerárquico de las clases (Bajo < Medio < Alto), luego se unifican las predicciones de ambos modelos de forma tabular.

- Modelo A = P(Bajo) ---> True ; P(Medio_Alto) ---> False
- Modelo B = P(Alto) ---> True ; P(Medio_Bajo) ---> False

Tabla 6

Criterios de unificación de predicciones de los modelos binarios

Modelo A	Modelo B	Definitivo
True (Baja)	False (Media_Baja)	Baja
True (Baja)	True (Alta)	No sabe
False (Media_Alta)	True (Alta)	Alta
False (Media_Alta)	False (Media_Baja)	Media

Nota. Fuente: Elaboración propia

Luego de ajustar cada clasificador de manera individual, se combinan las predicciones siguiendo los criterios de la Tabla 6. Para ambos clasificadores se utiliza la misma receta de preprocesamiento aplicado en el modelo multiclase, pero en este caso se utiliza la técnica de SMOTE, igualando la clase minoritaria a igual proporción que la mayoritaria, ya que se está comparando una categoría con dos categorías combinadas, produciendo un desbalance en los datos. La métrica por optimizar en el ajuste de hiperparametros también es la "roc_auc", en este caso no se desea minimizar un tipo de error de clasificación, ya que para los clasificadores individuales es igual de importante clasificar correctamente los dos tipos de esquinas. Además el entrenamiento se hará con clases balanceadas, gracias a la aplicación del método SMOTE.

Tabla 7

Evaluación del clasificador A

modelo	.metric	.estimator	valor	n	std_err	algoritmo	Error
Modelo A	roc_auc	binary	0.8201688	10	0.002398226	lgbm	validación cruzada
Modelo A	roc_auc	binary	0.8128123	NA	NA	lgbm	Diciembre 2019

Nota. Fuente: Elaboración propia

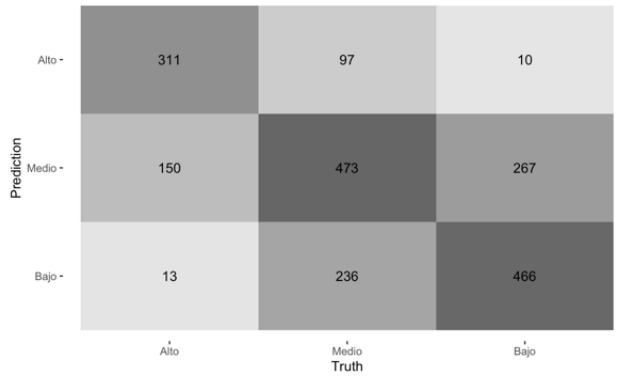
Tabla 8

Evaluación del clasificador B

modelo	.metric	.estimator	valor	n	std_err	algoritmo	Error
Modelo B	roc_auc	binary	0.8981	10	0.00296	lgbm	validación cruzada
Modelo B	roc_auc	binary	0.9108	NA	NA	lgbm	Diciembre 2019

Nota. Fuente: Elaboración propia

Figura 12. *Matriz de confusión de la combinación de los clasificadores binarios*



Fuente: Elaboración propia

El "Modelo A" busca discriminar la categoría "Bajo" del resto, sin embargo, en la Tabla 7 la métrica de error sufre una baja de rendimiento al comparar el error de validación cruzada con los datos de prueba, aunque es leve la diferencia, este es un síntoma de que el modelo no pudo generalizar de la mejor manera. En la Tabla 8, el "Modelo B" muestra el caso opuesto, mejorando su rendimiento, se recuerda que este modelo busca discriminar la categoría "Alto" del resto. Aplicando la metodología planteada en la Tabla 6, se obtiene una combinación de las predicciones de ambos clasificadores, mostrando dicha combinación en la matriz de confusión de la Figura 12. Se observa una mejoría en la predicción de las categorías extremas, la incorporación del orden jerárquico de las categorías a través de dos clasificadores binarios, arrojó resultados satisfactorios, sin embargo, se observó una menor precisión en la categoría media. En el apartado siguiente se comparan ambos enfoques asignando costos a los errores de las matrices.

3.2.3 Presentación de resultados

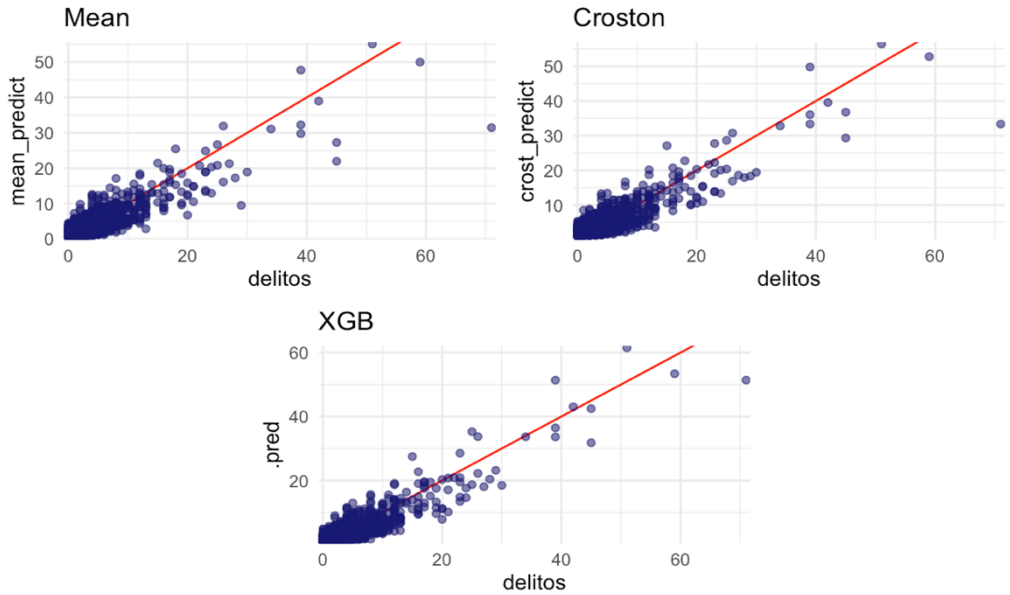
Luego del ajuste de los modelos y cálculo de métricas de rendimiento, se procede a presentar y contrastar las predicciones generadas. Se inicia revisando los resultados de los modelos de clasificación y regresión, mediante visualizaciones y tablas de resumen.

Posteriormente se analiza la contribución global de las variables sobre las predicciones, para finalmente realizar una evaluación de las contribuciones locales seleccionando algunas esquinas.

Regresión

Inicialmente se establecieron dos modelos base para el enfoque de regresión, siendo la media de las series y la aplicación del modelo de series de tiempo Croston. Posteriormente se ajustó un modelo XGB, se comparan las predicciones de estos tres modelos para el mes de prueba. A continuación, se muestran los gráficos comparando los registros observados con los predichos, además de una tabla resumen con las principales métricas.

Figura 13. *Comparación de modelos de regresión*



Fuente: Elaboración propia

Tabla 9

Evaluación del rendimiento de los modelos de regresión

modelo	rmse	mae	rsq
Croston	2.429	1.697	0.7669
MEAN	2.458	1.574	0.7520
XGB	2.177	1.503	0.8027

Nota. Fuente: Elaboración propia

En la Tabla 9 se aprecia que el modelo XGB fue superior en las tres métricas de evaluación, visualmente también se observa un mejor ajuste de los datos en la Figura 13. Existen esquinas que presentaron cambios abruptos en el registro de delitos para el mes de evaluación, afectando el rendimiento de los modelos. Una alternativa para modelar este tipo de problemas es a través de regresiones de conteo, se podrían contrastar estos resultados con este tipo de enfoques, Hilbe (2017) presenta una amplia explicación del tema.

Clasificación

Para los dos enfoques de clasificación utilizados, se mostraron las matrices de confusión para el mes de prueba. Analizando la problemática desde la perspectiva de la organizacional, se puede tolerar de mejor manera un error de predicción de un esquina considerada de riesgo medio y que se clasifique como de bajo o alto riesgo. Sin embargo, una esquina considerada de bajo riesgo, que reciba una clasificación de alto riesgo o viceversa, puede resultar un error sumamente costoso desde el punto de vista logístico, social y económico. Aunque ambos enfoques presentan buenos rendimientos discriminando las categorías extremas, se propone asignar un mayor costo a este tipo de errores a través de una matriz de costos.

Tabla 10

Matriz de Costos

Predict / observado	Alto	Medio	Bajo
Alto	-1	1	10
Medio	1	0	1
Bajo	10	1	-1

Nota. Fuente: Elaboración propia

Figura 14. *Comparación de modelos de clasificación a través de matrices de costos*

Clsificadores binarios

Riesgo	Alto	Medio	Bajo
Alto	-311	97	100
Medio	150	0	267
Bajo	130	236	-466

Costo 203

Clasificador múltiple

Riesgo	Alto	Medio	Bajo
Alto	-304	96	90
Medio	156	0	293
Bajo	140	204	-441

Costo 234

Fuente: Elaboración propia

La Tabla 10 muestra la matriz de costos, asignado los mismos de manera arbitraria, pero a modo de ejemplo puede guiar a la organización para optimizar la problemática. Para los errores extremos fue asignado un costo de 10 unidades, para el resto de los errores el costo fue de 1, los aciertos de los casos extremos generan un beneficio de 1 unidad y la predicción del riesgo medio no presenta costo. En la Figura 14 se observa que los clasificadores binarios presentan un costo más bajo para la organización, a excepción del par "Bajo-Medio", el resto de los errores fueron sustancialmente menores que el clasificador multiclase. En definitiva, el enfoque de clasificadores binario superó al modelo de clasificación multiclase, sin embargo, esto puede variar en función de las prioridades que tenga la organización. Si la predicción correcta del riesgo medio se convierte en una prioridad, asignándole a la misma un valor positivo en la matriz de costos, el modelo multiclase pasaría a ser el de menor costo.

3.3 Metodología de implementación dentro de la organización

La etapa final del estudio está relacionada con implementar todo el desarrollo mostrado hasta ahora, dentro de la organización seleccionada. Primero se explicarán las distintas metodologías para hacer la inferencia de los distintos modelos ajustados, este punto es clave para hacer disponible un modelo en un ambiente productivo. Finalmente se selecciona y desarrolla una metodología para efectuar la implementación dentro de la organización, este aspecto es hipotético, ya que no se cuenta con información certera acerca del funcionamiento de la gestión de datos dentro de la entidad.

3.3.1 Inferencia del modelo

Entrenados y ajustados los distintos modelos, se procede a guardarlos a través de métodos y formatos adecuados. Esto con el objetivo de poder utilizarlos en un futuro de manera independiente al flujo de trabajo utilizado, además de poder guardar distintas versiones de los modelos ajustados. Si bien existen muchos enfoques para guardar modelos, a continuación, se muestran ejemplos de los métodos RDS, binario y la portabilidad entre los lenguajes Python y R.

Método RDS

Este método es único del lenguaje R, es la forma más directa de guardar modelos en R, aunque es ampliamente utilizado para guardar cualquier tipo de objeto. Básicamente comprime automáticamente el objeto y también guardará los metadatos, permitiendo tener un objeto más liviano sin perder información. Una desventaja es que los objetos solo se pueden recuperar en R, a excepción de algunos objetos que pueden cargarse en Python, Grolemund (2014) presenta una introducción al tema. Para los cuatro modelos ajustados, se toma una muestra aleatoria de seis observaciones de los datos originales, eliminando además la variable a delitos, con el objetivo de generar predicciones aisladas del flujo del trabajo. A través de las recetas de transformación, se pueden aplicar modelos con especificaciones distintas como regresión, clasificación multiclase y binaria, sin necesidad de modificar la data original.

```
set.seed(123)

# Datos a predecir

Datos_sample = Data_set_modelos %>% sample_n(6) %>% select(-delitos)

# Regresion

saveRDS(modelo_fit_reg, "modelo_fit_reg.rds"); modelo_fit_reg = readRDS("modelo_fit_reg.rds")

# Multiclass

saveRDS(model_fit_multiclass, "model_fit_multiclass.rds"); model_fit_multiclass = readRDS("model_fit_multiclass.rds")

# Binario A

saveRDS(model_fit_A, "model_fit_A.rds"); model_fit_A = readRDS("model_fit_A.rds")

# Binario B

saveRDS(model_fit_B, "model_fit_B.rds"); model_fit_B = readRDS("model_fit_B.rds")

# Predicciones

bind_cols(regreesion = predict(modelo_xgb_reg, Datos_sample)$.pred,

          multiclase = predict(model_fit_multiclass, Datos_sample)$.pred_class,

          clasificdor_A = predict(model_fit_A, Datos_sample)$.pred_class,
```

regreesion <dbl>	multiclase <fctr>	clasificdor_A <fctr>	clasificdor_B <fctr>
1.7409934	Bajo	Bajo	Media_Bajo
2.3842454	Medio	Medio_Alto	Media_Bajo
2.8863425	Medio	Medio_Alto	Media_Bajo
5.2717018	Alto	Medio_Alto	Alto
0.9696378	Bajo	Bajo	Media_Bajo
4.2494230	Medio	Medio_Alto	Media_Bajo

Los cuatro modelos ajustados se guardan en formato ".rds", posteriormente se pueden cargarse en una nueva sesión independiente. El objeto guardado es un "workflow" que contiene dos elementos, estos se ejecutarán de forma secuencial al momento de predecir

nuevos datos. El primer elemento es el "recipe", este guarda todo el preprocesamiento necesario para que el modelo pueda predecir. El segundo objeto es el modelo ajustado con todas las especificaciones, por lo tanto, los datos a predecir pueden estar en su estado original sin procesamiento, ya que primero se aplicará el preprocesamiento y luego el modelo. En la salida se observa la predicción de los cuatro modelos sobre los datos brutos. Los flujos de trabajos mostrados y la muestra aleatoria de datos, serán reutilizados para explicar los otros métodos.

Método binario

Este enfoque consiste en guardar los modelos como objetos binarios, la mayoría de los modelos tienen un método asociado que permite guardarlos en este formato. El objeto generado tiene la ventaja de ser universal entre las diversas interfaces del modelo, se menciona como desventaja el hecho de tener que descomponer el "workflow" de Tidymodels. A continuación, a modo de ejemplo se muestra la implementación para el modelo de regresión y uno de los modelos de clasificación binario.

```
Modelo = extract_model(modelo_fit_reg)
Receta = extract_recipe(modelo_fit_reg)
xgboost::xgb.save(Modelo, 'modelo_xgb_deploy.model')
regresion_data= bake(Receta, new_data = Datos_sample) %>% as.matrix()
modelo_xgb_deploy = xgboost::xgb.load('modelo_xgb_deploy.model')
predict(modelo_xgb_deploy, regresion_data, reshape = TRUE)
```

```
[1] 1.7409934 2.3842454 2.8863425 5.2717018 0.9696378 4.2494230
```

```
Modelo = extract_model(model_fit_A)
Receta = extract_recipe(model_fit_A)
lightgbm::lgb.save(Modelo , 'modelo_lgb_deploy.model')
multiclass_data = bake(Receta , new_data = Datos_sample_A) %>% select(-delitos) %>% as.matrix()
modelo_lgb_deploy = lightgbm::lgb.load('modelo_lgb_deploy.model')
predict(modelo_lgb_deploy, multiclass_data)
```

```
[1] 0.3894982 0.7603496 0.7807194 0.7306535 0.2176520 0.9414982
```

El primer paso consiste en utilizar las funciones "extract_model" y "extract_recipe", con la finalidad de extraer el modelo ajustado y la receta de preprocesamiento del "workflow". Posteriormente se guarda el modelo como un objeto binario, a través de la

receta aplicamos el preprocesamiento de los datos a predecir con la función "bake". Finalmente se carga el modelo binario y se realiza la predicción sobre los datos previamente preprocesados. Si bien parece poco práctico este método en comparación con el anterior, teniendo que descomponer el "workflow", tiene la ventaja de poder utilizar cualquier modelo fuera del marco de "Tidymodeles", inclusive del propio lenguaje R como se verá en el siguiente método.

Método de portabilidad de modelos R-Python

Este método es una continuación del apartado anterior, pero mostrando la portabilidad entre los lenguajes R y Python. En esta oportunidad se expondrá la aplicación de entrenar el modelo en R y realizar las predicciones en Python, pero el caso inverso también es posible con el mismo enfoque, sin embargo, está fuera del alcance del estudio. A continuación, se muestra la implementación del método, el primer fragmento de código es en R y el siguiente es en Python, se mostrará únicamente la aplicación para el modelo de regresión.

```
write.csv(regresion_data ,'regresion_data.csv', row.names=FALSE)
```

```
!pip install xgboost==1.2.1; import xgboost
import pandas as pd
model = xgboost.Booster()
model.load_model('modelo_xgb_deploy.model')
data = pd.read_csv(regresion_data.csv')
data = xgboost.DMatrix(data.values)
model.predict(data)
```

```
array([1.7409934 , 2.3842454 , 2.8863425 , 5.271702  , 0.96963775,4.249423  ],
        dtype=float32)
```

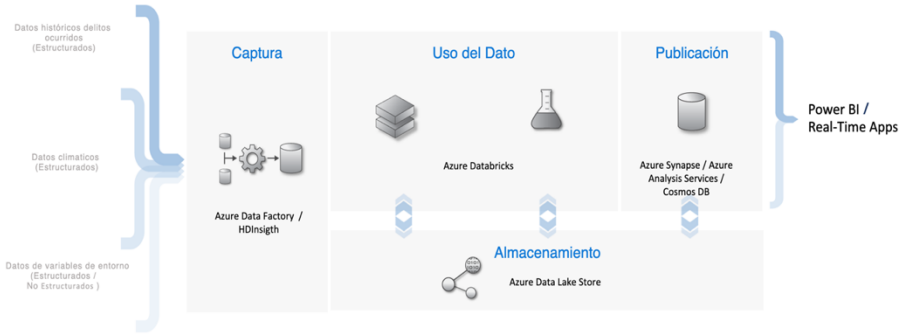
El primer paso consiste en exportar los datos preprocesados a predecir como un archivo ".csv", esto se hace en R. Seguidamente desde Python se instala el paquete del modelo y se carga dentro del mismo, el modelo binario entrenado en R. Finalmente se carga el archivo ".csv" y se transforman los datos en una matriz "sparse", para luego realizar la predicción, se observan los mismos resultados para los tres métodos sobre el modelo de regresión.

3.3.2 Selección de metodología

Este estudio se realiza desde una perspectiva externa a la organización, por lo tanto, es necesario precisar algunas limitaciones con respecto a las propuestas metodológicas. No se cuenta con información referente a los sistemas de gestión de datos específicos que maneja la organización, además los datos que se utilizan para este análisis son capturados desde los repositorios públicos ya mencionados. La organización probablemente tenga acceso a información menos agregada y más específica con respecto a la publicada en los repositorios, además de recibir los datos con una mayor frecuencia a la publicada en los portales oficiales.

En función a esto se propone seleccionar una arquitectura según un sistema de gestión hipotético, aunque no sea el que utiliza la organización, se asume que es una aproximación realista. Por lo tanto, aunque este estudio fue implementado previamente de forma local, se propone una solución en la nube. La arquitectura seleccionada corresponde al servicio de computación en la nube *Microsoft Azure*, a continuación, en la Figura 15 se presenta un esquema de la arquitectura que mejor se adapta a la problemática abordada.

Figura 15. *Arquitectura de datos seleccionada*



Fuente: Elaboración propia

Se asume que la organización tiene acceso al repositorio, los datos en su mayoría estructurados, se asume que están almacenados en una base de datos SQL. Los datos no estructurados hacen referencia a los objetos geoespaciales, se almacenan en la base NoSQL *Apache HBase*, estos tienen una baja frecuencia de actualización, ya que corresponde a delimitaciones de diversas zonas urbanas. Los datos históricos de delitos se capturan mediante *Azure Data Factory,* finalmente todos los datos son almacenados dentro de *Azure Data Lake Store*. A través de *Azure Databricks* se transforman, analizan y combinan todas las fuentes de datos almacenadas según la metodología expuesta.

En este punto se propone realizar una actualización en tiempo real de los delitos ocurridos, con el objetivo de generar estadísticas descriptivas y tableros a través de *Power BI*. Posteriormente se implementan los modelos de aprendizaje automático con una frecuencia mensual, ya que no sería viable generar las predicciones diarias, debido a que se cuentan con muy pocos delitos diarios registrados para las más de 2000 esquinas estudiadas. Estos puntos están cubiertos mediante las herramientas *Azure Synapse* y *Azure Analysis Services*, que permiten guardar los resultados del modelo aplicado y combinarlo con datos guardados en el almacén de datos. La última acción requerida hace referencia a la ejecución de acciones en tiempo real dentro de alguna aplicación, en este caso sería a través de una *Shiny app*. Permitiendo registrar en la aplicación mapas interactivos de CABA, siendo complementarias a las estadísticas descriptivas mostradas en *Power BI*, pero mostrando además las predicciones delictivas generadas mensualmente, esto se puede se implementa a través de la herramienta *Cosmos DB*.

La metodología para la implementación del proyecto se denomina CRISP MD (*Cross Industry Standard Process for Data Mining*), existen varias razones para su elección que serán explicadas a continuación. El primer elemento se refiere a los entregables, siendo principalmente información incrustada en los espacios de decisión, simplemente corresponde a un complemento visual y descriptivo en la toma de decisiones. Aunque parte del entregable necesite generarse en tiempo real y con ello integrarse a alguna pieza funcional de software, no será el eje de este proyecto. El segundo aspecto, citando a Data Science Project Management. (s.f) "El enfoque inicial en la comprensión empresarial es útil para alinear el trabajo técnico con las necesidades comerciales y para evitar que los científicos de datos se sumerjan en un problema sin comprender adecuadamente los objetivos comerciales" (párr.54). Se hace referencia al foco que tiene la metodología en la definición del problema, siendo este aspecto fundamental en este proyecto. Por último, dicha metodología se adapta a equipos de trabajos pequeños, como se menciona en Saltz, Shamshurin & Crowston (2017), este aspecto coincide con las necesidades de este proyecto. En la siguiente sección se describen las distintas etapas de la metodología, además de la explicación de su implementación, en Chapman et al. (2000) puede revisarse la documentación oficial.

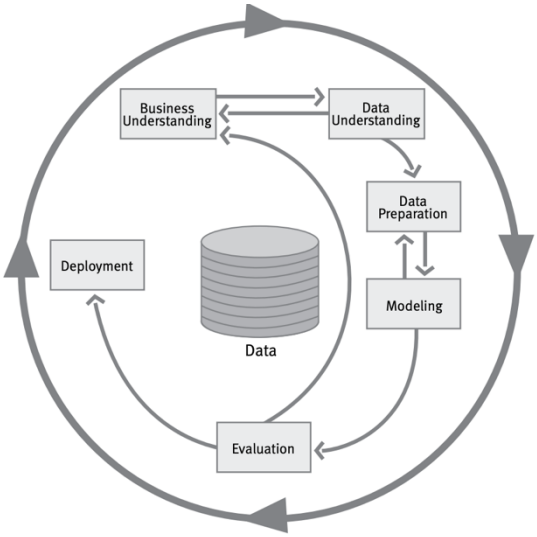
3.3.3 Desarrollo de la metodología

La metodología CRISP-DM está estructurada en seis etapas, siendo las mismas la comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. El flujo de trabajo puede ser iterativo entre alguna de las etapas, aspecto común en los proyectos de ciencia de datos. Antes de adaptar las fases de la metodología a la problemática organizacional, se propone abordar algunas definiciones que dan contexto a la implementación. Se propone un equipo de trabajo compuesto por un arquitecto de datos, un científico de datos, un ingeniero de aprendizaje automático y un analista de datos. El arquitecto tiene el rol de diseñar correctamente la estructura del sistema de gestión de datos, verificando la compatibilidad entre las bases y el sistema de nube seleccionado.

El científico de datos se encarga de implementar el proyecto, sin embargo, a diferencia del presente estudio que fue realizado localmente, dentro de la organización debe implementarse dentro de la arquitectura seleccionada. El analista de datos cumple el rol de prestar soporte al científico de datos, específicamente recopilando hallazgos dentro de la organización, recogiendo las inquietudes e impresiones del usuario, además de encargarse del proceso de visualización de datos comentado anteriormente. El ingeniero de datos cumple la función de hacer funcionar lo desarrollado por el científico y analista de datos en el servicio de nube, recomendando además mejoras a nivel de código y flujos de trabajo que permitan hacer más eficiente la implementación.

Para medir el éxito del proyecto se proponen dos elementos a considerar, en primer lugar la implementación esté en funcionamiento dentro de la organización, atendiendo todas las consideraciones técnica planteadas. Seguidamente, se propone evaluar la efectividad de la propuesta en la solución de la problemática organizacional, fijando indicadores de costos logísticos, monetarios y de efectividad de los operativos policiales efectuados. Esto con el objetivo de aplicar pruebas piloto y comparar si existe una mejoría en los indicadores, comparando políticas organizacionales utilizando o no la herramienta desarrollada. A continuación, en la Figura 16 se muestra el desarrollo de la metodología CRISP-DM dentro de la organización.

Figura 16. Esquema de la metodología CRISP-DM



Fuente: Chapman et al. (2000)

Comprensión del negocio

Para este problema de estudio, se hace referencia a la comprensión del funcionamiento de la organización y su vinculación con la propuesta desarrollada. La principal inquietud surge al medir la ocurrencia de delitos, con el objetivo de generar algún tipo de predicción, ¿Se agrupan por días y se predicen delitos en función de una fecha?, ¿ Se busca generar predicciones en función de la geolocalización de los delitos ocurridos?. Esta fase de diseño es crucial e iterativa, requiere de un desarrollo intensivo del estado del arte y del entendimiento de la organización. No hacer un foco correcto en esta etapa, puede generar una incorrecta definición de la variable a predecir, siendo en este caso los delitos ocurridos en distintas esquinas de CABA, definición que surgió luego de un gran número de iteraciones.

Comprensión de los datos

Esta etapa es iterativa con la anterior, ya que para definir lo que se desea predecir también es necesario conocer las fuentes de datos a las cuales tiene acceso la organización. Se deben estudiar las variables que se incorporarán en el estudio, evaluar si hacen sentido para la organización, si el formato de la información es posible manipularlo e integrarlo. Además de analizar el uso que actualmente les da la organización a estas fuentes de datos, pudiéndose tomar ideas de soluciones que ya están desarrolladas.

Preparación de los datos

Esta fase constituye todo el proceso de integración, limpieza y transformación de datos según lo definido en las etapas anteriores. Existe un primer paso referente a la integración de la información, aprovechando una ventaja debido a que toda la información se encuentra en el mismo repositorio. Posteriormente se realiza el análisis exploratorio de datos, permitiendo efectuar modificaciones en las definiciones del problema, además de generar nuevas ideas que enriquezcan el proyecto. Finalmente se ejecutan todos los algoritmos de optimización y cálculo de distancias, necesarios para generar el conjunto de datos definitivo.

Modelado

Construido el conjunto de datos definitivo para modelar, inicia la etapa de selección de modelos, definición de estrategias de aparición de datos y entrenamiento. El desafío está en escalar en la nube lo desarrollado localmente, pudiendo generar múltiples instancias para evaluar simultáneamente distintos modelos con diversas estrategias de entrenamiento. Es importante primero generar el funcionamiento óptimo del flujo de trabajo, para después sobre el mismo implementar diversos modelos de aprendizaje automático.

Evaluación

Esta etapa está estrechamente relacionada con la fase de modelado, debido a que se deben evaluar los modelos implementados a través de diversas métricas. Por lo tanto, de la aplicación de la evaluación se puede volver a la etapa de preparación de datos e inclusive a la definición del problema. Se definen las métricas de evaluación, también se diseñan pruebas para predecir datos nuevos que pueden surgir en futuras actualizaciones del proyecto, finalmente se presentan los resultados de las evaluaciones a los usuarios.

Despliegue

Si la implementación cumple con todas las etapas de la metodología, incluyendo la aprobación por parte del usuario, se procede al despliegue del proyecto. Esto implica la puesta en producción de los tableros en *Power BI* y la *Shiny app*, para capturar en tiempo real la ocurrencia de los registros registrados. Adicionalmente se despliegan los modelos ajustados con las predicciones para el último periodo de tiempo, pero preparados para generar predicciones al inicio de cada mes para todas las esquinas de la ciudad.

Conclusiones

La aplicación de modelos predictivos en entornos empresariales o en problemáticas sociales, brindan una mejor comprensión del funcionamiento interno de los algoritmos. Las bases conceptuales y los fundamentos teóricos son imprescindibles para responder interrogantes referentes a los distintos modelos, ¿Para qué sirve?, ¿Cómo funciona?, ¿En qué tipo de problemas es más conveniente su implementación?. Preguntas que habitualmente surgen en cualquier proyecto de ciencia de datos, sin embargo, muchos modelos tienen componentes matemáticos y estadísticos que necesitan una visión complementaria desde un enfoque práctico para poder comprenderlos. Este estudio fue una buena oportunidad para implementar distintos algoritmos de aprendizaje automático en una problemática social y económica como la ocurrencia de delitos.

La construcción de una propuesta metodológica como la presentada en este estudio, permiten generar un vínculo entre la ciencia de datos y las problemáticas organizacionales, siendo este uno de los aspectos centrales de la especialización. Aplicar algoritmos sofisticados, utilizar herramientas de vanguardia como los sistemas de computación en la nube, tener un dominio avanzado de algún lenguaje de programación, son elementos necesarios pero no suficientes para generar cambios en las organizaciones. Sin una correcta definición del problema, un mal enfoque de la contextualización del proyecto o una falta de entendimiento de las necesidades de la organización, las habilidades técnicas o herramientas tecnológicas no pueden generar por sí solas propuestas de valor agregado.

A través de los resultados del estudio, se afirma que es factible relacionar el factor "entorno-tiempo" con la ocurrencia de delitos, mediante la implementación de modelos predictivos de aprendizaje automático, permitiendo generar una propuesta de valor para la aplicación de políticas de seguridad. La incorporación de los factores climáticos y de entorno físico, contribuyeron a generar mejores predicciones, siendo evidente al observar el rendimiento inferior de los modelos que únicamente toman en cuenta el historial de delitos ocurridos. Además fue efectivo el planteamiento de enfocar la problemática desde distintas perspectivas, específicamente desde los enfoques de regresión, clasificación multiclase y clasificación binaria múltiple, brindando un abanico más amplio de herramientas para apoyar la toma de decisiones por parte de la organización.

La etapa de planteamiento y enfoque de la problemática a tratar fue el aspecto más desafiante del presente análisis, partiendo desde una perspectiva de predicción del tipo de delito, iterando a través de múltiples propuestas hasta converger en el planteamiento definitivo mostrado en el estudio. Tomar como eje principal las esquinas de CABA para posteriormente vincular los delitos ocurridos en las cercanías, resultó una propuesta creativa y novedosa, no se encontraron antecedentes parecidos en la bibliografía consultada. Lin et al. (2018) presentaron la propuesta más similar, que de hecho como se mencionó en las secciones anteriores, sirvió de inspiración en muchos aspectos para desarrollar esta investigación.

La implementación de un flujo de trabajo que permita hacer reproducible la propuesta de valor, desde el proceso de captura de datos hasta la presentación de los resultados, fue un objetivo que se pudo cumplir de forma satisfactoria. Para esto fueron determinantes algunos elementos, dentro de ellos se puede citar la correcta planificación metodológica desde el inicio del proyecto, así como el seguimiento ordenado del mismo. También la codificación ordenada, la programación funcional utilizada y el complemento de versionado de código, contribuyeron a culminar el proyecto en tiempo y forma. Siendo esta una necesidad debido a la gran cantidad de datos que se tenían que manipular, la creación y aplicación de algoritmos para transformar la información, los diversos enfoques de modelado implementado. Finalmente el marco de aprendizaje automático "Tidymodels", inspirado en el enfoque de *"tidy data"*, expuesto en Wickham (2014), permitió diseñar y aplicar un flujo de trabajo robusto y flexible, aspectos importantes debido a las múltiples iteraciones que sufren este tipo de proyectos.

En función de la propuesta desarrollada, se recomienda a los tomadores de decisiones de la organización, que no vean este tipo de propuestas como una sustitución o solución "mágica" en la toma de decisiones, todo lo contrario, es un complemento para ejecutar mejores políticas de seguridad. Es importante además, que se realice una evaluación referente a los procesos de gestión de datos en la entidad, para así poder determinar las debilidades y prioridades, y que esto a su vez permita que la implementación de modelos predictivos que generen soluciones de impacto dentro de la organización. Finalmente, se espera que independientemente de la adopción de la propuesta dentro de la organización,

este análisis sirva de inspiración para vincular la ciencia de datos, con la solución de problemáticas organizacionales.

A continuación, se presentan algunas propuestas para futuras líneas de investigación relacionadas con el estudio, así como también posibles mejoras o enfoques complementarios. En primer lugar se hace referencia a la implementación del proyecto en instancias con una mayor capacidad de procesamiento, como por ejemplo en algunos de los principales servicios de nube. Esto con la finalidad de poder ajustar los modelos con más alternativas, tanto en la combinación e incorporación de variables, como en métodos más intensivos en el ajuste de hiperparametros. También se propone mejorar el algoritmo que permite calcular el conteo de delitos ocurridos en cada esquina , debido a que la combinación de esquinas y delitos que requiere probar la matriz de distancias suele ser muy extensa , esto trae como consecuencia que este proceso tome muchas horas para ejecutarse.

Además de mejoras relacionadas con el ámbito computacional, también se plantea abordar el problema distintos enfoques. Lin et al. (2018) utilizaron cuadrículas de una determinada dimensión para realizar el conteo de delitos, sin embargo incorporaron en el modelado la ocurrencia de delitos en las cuadrículas vecinas, esta idea puede incorporarse ajustándose sobre las esquinas seleccionadas. Un cambio radical en el estudio pero que puede complementarlo, hace referencia a lo presentado por Sevri et al. (2017), donde enfocan el problema desde una perspectiva descriptiva. Implementando un modelo de reglas de asociación, con esto se podría analizar la relación entre los elementos de entorno, el clima y los tipos de delitos.

Finalmente se pueden incorporar análisis de inferencia causal, con el objetivo de entender las causas de la delictividad en función de los distintos elementos estudiados. También se plantea la aplicación de un enfoque similar al inferencial, pero con el objetivo de evaluar e interpretar los modelos de aprendizaje automático, observando la contribución y relación de las variables dentro de los modelos, al momento de generar predicciones. Los enfoques de *Shapley Values* y *Local Surrogate* (LIME) son los más populares para interpretar los modelos, Molnar (2019) expone ambas metodologías.

Referencias bibliográficas

Amat, J. (2020, abril). Machine Learning con R y tidymodels. *Cienciadedatos.net*. Recuperado de https://www.cienciadedatos.net/documentos/59_machine_learning_con_r_y_tidymodels.html

Ariel, B., Partridge, H. (2017). Predictable Policing: Measuring the Crime Control Benefits of Hotspots Policing at Bus Stops. *J Quant Criminol*, (33), 809–833. doi:10.1007/s10940-016-9312-y

Bisong , E. (2019). Building Machine Learning and Deep Learning Models on Google Cloud Platform. doi:10.1007/978-1-4842-4470-8

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. Step-by-step Data Mining Guide. Recuperado de https://www.the-modeling-agency.com/crispdm.pdf.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16,321-357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. New York, NY, USA: ACM. doi:10.1145/2939672.2939785

Christou, V., & Fokianos, K. (2015). On count time series prediction, *Journal of Statistical Computation and Simulation*, 85(2), 357-373, doi:10.1080/00949655.2013.823612

Clark, W. (1965). Markov Chain Analysis in Geography: An Application to the Movement of Rental Housing Areas. *Annals of the Association of American Geographers*, 55, 351-359. doi:10.1111/j.1467-8306.1965.tb00523.x

Boehmke, B., & Greenwell, B.M. (2019). Hands-On Machine Learning with R. doi:10.1201/9780367816377

Data Science Project Management. (s.f). CRISP-DM. Recuperado de https://www.datascience-pm.com/crisp-dm-2/

Duan, L., Ye, X., Hu, T., & Zhu, X. (2017). Prediction of Suspect Location Based on Spatiotemporal Semantics. *ISPRS International Journal of Geo-Information*, 6(7), 185. doi:10.3390/ijgi6070185

Ferreira, J., João, P., & Martins, J. (2012). GIS for Crime Analysis - Geography for Predictive Models. *The Electronic Journal Information Systems Evaluation*, 15, 36 - 49.

Grogan, M. (7 de septiembre de 2020). Forecasting an Intermittent Time Series. *Medium*. Recuperado de https://medium.com/swlh/forecasting-an-intermittent-time-series-1461de7616fe

Grolemund, G. (2014). *Hands-On Programming with R*. Recuperado de https://rstudio-education.github.io/hopr/index.html

Hale, J. (6 de agosto de 2020). Which Evaluation Metric Should You Use in Machine Learning Regression Problems?. *Towards Data Science*. Recuperado de https://towardsdatascience.com/which-evaluation-metric-should-you-use-in-machine-learning-regression-problems-20cdaef258e

Hand, D. J. & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. doi:10.1023/A:1010920819831

Handelman, G., Kok, H., Chandra, R., Razavi, A., Huang, S., Brooks, M., Lee, M., & Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *American Journal of Roentgenology*, 212(1), 38-43. doi:10.2214/AJR.18.20224

Hart, T., & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting, *Policing: An International Journal*, 37(2), 305-323. doi:10.1108/PIJPSM-04-2013-0039

Hilbe, J. (2017). The statistical analysis of count data / El análisis estadístico de los datos de recuento. *Cultura y Educación*. 29. 1-52. doi:10.1080/11356405.2017.1368162.

Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. Recuperado de https://otexts.com/fpp2/

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 3149–3157. doi:10.5555/3294996.3295074

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Recuperado de http://www.feat.engineering/

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Recuperado de https://www.tidymodels.org

Kumar, S. (13 de septiembre de 2020). Understanding 8 types of Cross-Validation. *Towards Data Science*. Recuperado de https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d

Lantz, B. (2013). *Machine learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Birmingham, UK: Packt Pub.

Leong, K., & Sung, A. (2015). A review of spatio-temporal pattern analysis approaches on crime analysis. *International E-Journal of Criminal Sciences*, 9, 1–33.

Lin, Y.-L., Yen, M.-F., & Yu, L.-C. (2018). Grid-Based Crime Prediction Using Geographical Features. *ISPRS International Journal of Geo-Information*, 7(8), 298. doi:10.3390/ijgi7080298

Malik, F. (18 de febrero de 2020). What is Grid Search?. *Medium*. Recuperado de https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

Molnar, C. (2019*). Interpretable machine learning. A Guide for Making Black Box Models Explainable*. Recuperado de https://christophm.github.io/interpretable-ml-book/

Münch D., Grosselfinger AK., Krempel E., Hebel M., Arens M. (2019). Data Anonymization for Data Protection on Publicly Recorded Data. ICVS 2019. Lecture Notes in Computer Science, Springer, Cham, 11754, 245-258, doi:10.1007/978-3-030-34995-0_23

Nikhilesh, S. (2019, julio). XGBOOST vs LightGBM: Which algorithm wins the race!!!. *Towards Data Science*. Recuperado de https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d

Obi, B. (2019, octubre). Model Parameters and Hyperparameters in Machine Learning — What is the difference?. *Towards Data Science*. Recuperado de https://towardsdatascience.com/model-parameters-and-hyperparameters-in-machine-learning-what-is-the-difference-702d30970f6

Penn, C. (8 de agosto de 2019). The Evolution of the Data-Driven Company. [Publicación de Blog]. Recuperado de https://www.christopherspenn.com/2019/08/the-evolution-of-the-data-driven-company/

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59. doi:10.1089/big.2013.1508

Rabosky, A., Cox, C., Rabosky, D., Title, P., Holmes, I., Feldman, A., & McGuire, J. (2016). Coral snakes predict the evolution of mimicry across New World snakes. *Nature Communications*, *7*, 11484.

Ruiz, L. (29 de agosto de 2020). Introducción a LightGBM [Publicación de Blog]. Recuperado de https://dev.to/ruizleandro/introduccion-a-lightgbm-ij7

Salgado, D. (Julio, 2016). Big Data y la Estadística Oficial: retos. *Índice: revista de estadística y sociedad*, 68, 14-17.

Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the*

Association for Information Science and Technology, 68, 2720-2728. doi:10.1002/asi.23873

Sevri, M., Karacan, H., & Akcayol, A. (2017). Crime Analysis Based on Association Rules Using A priori Algorithm. *International Journal of Information and Electronics Engineering*, 7(3). 99-102. doi: 10.18178/ijiee.2017.7.3.669.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012, (4), 2951-2959.

Sommer, A., Lee, M., & Abèle, M. (2018). Comparing apples to apples: an environmental criminology analysis of the effects of heat and rain on violent crimes in Boston. *Palgrave Commun*, 4(138). doi:10.1057/s41599-018-0188-3.

Taylor, J. (22 de enero de 2016). *14 Key Data Cleansing Pitfalls*. Invensis. Recuperado de https://www.invensis.net/blog/14-key-data-cleansing-pitfalls/

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Anpéndice

Repositorios

Repositorio público del Gobierno de la Ciudad de Buenos Aires (GCABA), https://data.buenosaires.gob.ar/

Repositorio de OpenStreetMap, www.openstreetmap.org

Pagina Oficial de la Policia de CABA, https://policiadelaciudad.gob.ar/

Repositorio de la version extendida del presente estudio, https://rafael-zambrano-blog-ds.netlify.app/posts/2020-12-22-prediccin-de-delitos-en-caba/

Repositorio del paquete creado para almacenar funciones, https://github.com/rafzamb/raFunctions

Fragmentos de código

- Fragmento 1

```
delitos_2019 <- read_csv("http://cdn.buenosaires.gob.ar/datosabiertos/datasets/mapa-del-delito/delitos_2019.csv")

delitos_2018 <- read_csv("http://cdn.buenosaires.gob.ar/datosabiertos/datasets/mapa-del-delito/delitos_2018.csv")

delitos_2017 <- read_csv("http://cdn.buenosaires.gob.ar/datosabiertos/datasets/mapa-del-delito/delitos_2017.csv")
```

- Fragmento 2

```
proyeccion = "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"

# Barrios populares

dir.create(file.path(getwd(), "shape_barrios_populares"))

download.file("https://cdn.buenosaires.gob.ar/datosabiertos/datasets/desarrollo-humano-y-habitat/barrios-populares/barrios-populares-badata.zip", destfile="shape_barrios_populares/barrios-populares-badata.zip")

unzip("shape_barrios_populares/barrios-populares-badata.zip",exdir= "shape_barrios_populares")

shape_barrios_populares = readOGR("./shape_barrios_populares", layer = "barrios_vulnerables")

barrios_populares <- spTransform(shape_barrios_populares, CRS(proyeccion))

leaflet(barrios_populares) %>% addTiles() %>% addPolygons()
```

- Fragmento 3

```
delitos_2019 = delitos_2019 %>% drop_na(long,lat)

universidades = universidades %>% distinct(universida,long,lat)

local_bailables = local_bailables %>% rename(long = longitud,lat = latitud) %>% drop_na(long,lat)

hotel_baja = hotel %>% filter(categoria %in% c("1*","2*","3*"))

hogares_paradores = hogares_paradores %>% rename(long = X, lat = Y) %>% drop_na(long,lat)

hospitales = hospitales %>% mutate(coordenadas = str_trim(gsub("\\(|\\)|[[:alpha:]]","",WKT))) %>%

  select(coordenadas)%>% separate(coordenadas,c("long","lat"), sep = " ") %>% mutate_all(as.double)
```

- Fragmento 4

```
temperatura_mes_histoico = temperatura_mes_histoico %>%  clean_names() %>%
 bind_cols(meses = match(.$mes,meses)) %>% mutate(fecha = paste0(ano,"-",meses)) %>%
  mutate(fecha = as.yearmon(fecha, "%Y-%m")) %>% select(fecha, maxima,minima,media)
```

- Fragmento 5

```
delitos_def_3 = delitos_def_2 %>% pivot_wider(names_from = fecha_split,
             values_from = c(delitos, temperatura, mm_agua, dias_lluvia, veloc_viento),values_fill = 0) %>% clean_names()
```

- Fragmento 6

```
variables_entorno = lapply(datas, F.distancia.punto, metros = 250, data = split_esquina)
```

- Fragmento 7

```
pliegues = 1:13; names(pliegues) = pliegues
variables = c("delitos", "temperatura", "mm_agua", "lluvia", "viento"); names(variables) = variables
Data_set_final_2 = sliding_window(Data_set_final_1 %>% select(-c(long,lat)), 13, pliegues, variables)
```

- Fragmento 8

```
ts_mean=lapply(ts_object ,mean);ts_crost = lapply(ts_object ,function(x) crost(x,type='croston',h=1)$components$c.out[,1])
```

- Fragmento 9

```
Receta_regresion = recipe(formula = delitos ~ ., data =  train) %>%
  step_mutate(bancos_atm = bancos + atm, teatros_cine = teatros + cine) %>%
  step_rm(bancos, atm, teatros, cine, id, pliegue, lat, long) %>%
 step_nzv(all_predictors(),-c(estaciones_ferrocarril,bocas_de_subte,hotel_baja,bancos_atm,teatros_cine))
```

Tabla de Hiperparámetros

Parámetros	XGB	Lgbm	Lgb A	Lgb B	Descripción
mtry	35	20	32	34	número de predictores que se muestrearán aleatoriamente en cada división
trees	1017	906	1071	1461	número de árboles
min_n	6	22	26	8	número mínimo de puntos de datos en un nodo que se requieren para que el nodo se divida más
tree_depth	4	2	10	14	profundidad máxima
learn_rate	0.00288	0.0522	0.00328	0.07	La velocidad a la que el algoritmo de impulso se adapta de una iteración a otra
loss_reduction	14.3	1.13	0.0000056	26.5	La reducción en la función de pérdida requerida para dividir aún más
sample_size	0.604	0.85	0.768	0.897	La cantidad de datos expuestos a la rutina de ajuste