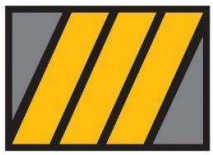


UNIVERSIDAD
TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICCIÓN DE SINIESTROS VIALES EN C.A.B.A.

TESIS

Jennifer Mercy Novello

Mayo 2022

Tutor: Cecilia Ruz

Resumen

La seguridad vial constituye un ámbito de relevante importancia a nivel internacional, y sobre el cual se trabaja de manera continua con el objetivo de disminuir el número de siniestros viales.

En este sentido, el presente trabajo se plantea como objetivo general el uso de técnicas analíticas para explotar masivamente la información y extraer conocimiento sobre los siniestros viales que sucedieron en la Ciudad Autónoma de Buenos Aires entre 2015 y 2018, para así analizar acciones concretas para prevenir siniestros viales.

Para la consecución del objetivo propuesto de esta tesis, en primer lugar, se busca encontrar características y factores (climáticos, demográficos, estacionales) que pueden ser determinantes para que ocurra un siniestro vial.

En la fase de modelado, se busca analizar y predecir mediante el empleo de un modelo Random Forest (Bosque Aleatorio), una de las técnicas supervisadas de aprendizaje automático, la cantidad de siniestros por día, hora, mes. Obteniendo resultados que son de utilidad para la generación de políticas y planes de acción para la prevención de siniestros viales.

Abstract

Road safety is an area of relevant importance at the international level, and on which work is being carried out continuously with the aim of reducing the number of traffic accidents.

In this way, the present work considers as a general objective the use of analytical techniques to explore the information and extract knowledge about the road accidents that happen in the City of Buenos Aires between 2015 and 2018, in order to analyze concrete actions to prevent traffic accidents

To achieve the proposed objective of this thesis, first of all, we seek to find characteristics and factors (climatic, demographic, seasonal) that can be decisive for the occurrence of a traffic accident.

In the modeling phase, the aim is to analyze and predict, through the use of a Random Forest model, one of the supervised machine learning techniques, the number of accidents per day, hour, month. Obtaining results that are useful for the generation of policies and action plans for the prevention of road accidents.

Índice

Índice de Figuras

Índice de tablas

1. Introducción

1.1. Contexto

La seguridad vial sigue siendo un importante problema a nivel mundial, generando un gran impacto social. Según la Organización Mundial de la Salud (OMS) se estima que cerca de 1,3 millones de personas mueren anualmente como consecuencia de los siniestros viales, y entre 20 y 50 millones padecen traumatismos no mortales.[1] A partir de las estadísticas provistas por el Observatorio vial de la Agencia Nacional de Seguridad Vial del Ministerio de Transporte, en Argentina de 2015 a 2019 se registró un promedio de 5.306 muertes[2] por siniestros viales al año, lo que equivale a una tasa de mortalidad[3] de 12,1 cada 100.000 habitantes. Mientras que en la Ciudad Autónoma de Buenos Aires según el Ministerio de Justicia y Seguridad (MJYS) [4] registró un promedio de 149 víctimas fatales anuales, lo que representa una tasa de mortalidad de 4,8 cada 100.000 habitantes. En conjunto, para el periodo 2015-2019 la Ciudad registró un total de 744 víctimas fatales por siniestros viales.

Distintos organismos hacen un constante esfuerzo en materia de seguridad vial para disminuir éstos índices. Al respecto, se han creado numerosas campañas y planes de acción, con los que se ha logrado una reducción de la siniestralidad y mortalidad. Con este objetivo, diversos organismos internacionales toman medidas y presentan planes de acción en el tiempo. La Organización de las Naciones Unidas (ONU) publicó en 2011 el Decenio de Acción de la Seguridad Vial 2011-2020, con la finalidad de disminuir los fallecidos en siniestros viales a partir de estrategias y programas de seguridad vial sostenibles. En agosto 2020, la Asamblea General de la ONU adoptó la resolución A/RES/74/299 proclamando la Década de Acción para la Seguridad Vial 2021-2030, la cual pretende la disminución a la mitad de las muertes y lesiones por causadas por el tránsito para 2030. Esta nueva Década de Acción brinda la oportunidad de aprovechar los éxitos y las experiencias de años anteriores para salvar más vidas. La OMS y las comisiones

[1] https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries

[2] Aquella persona que fallece de inmediato o dentro de los 30 días siguientes como consecuencia de un traumatismo causado por el siniestro vial (se exceptúan los suicidios).

[3] Expresa la relación entre el número de víctimas fatales que ocurren en una unidad geográfica considerada cada cien mil habitantes de la misma unidad geográfica, para un período de tiempo determinado.

[4] El MJyS incluye casos de víctimas fatales que ocurren en el lugar del hecho y hasta 7 días posteriores al siniestro. En este sentido, para alcanzar con la definición propuesta de víctimas a 30 días se emplea un factor de ajuste normalizado, recomendado por la Conferencia Europea de Ministros de Transportes (CEMT), este consiste en multiplicar por 1,08 las víctimas a 7 días.

regionales de la ONU, han desarrollado un Plan Global para esta Década de Acción 2021-2030.

El Plan Global describe las medidas prácticas y efectivas que todos los países y comunidades pueden implementar para salvar vidas y convoca a los gobiernos y asociados para aplicar un enfoque de sistemas integrado garantizando:

- uso seguro de vías de tránsito
- infraestructura vial, vehículos y comportamientos seguros
- respuesta de emergencia oportuna y eficaz

Figura 1: Plan Mundial para reducir por lo menos en un 50% las muertes y traumatismo para 2030

Fuente: OMS

Para aplicar estas medidas, explica la estrategia, que se requieren límites estrictos de velocidad y tecnología para monitorearla, lo que implica encuadrar el Plan en un marco legal y otorgarle financiamiento suficiente. Con respecto a quiénes deben encargarse de la implementación y monitoreo del Plan, la agencia menciona no sólo a los responsables políticos de alto nivel, sino también a otras partes que pueden influir en la seguridad vial, como la sociedad, el sector privado, las instituciones financieras y donantes, y los líderes

comunitarios y juveniles. La OMS ratificó que las medidas del Plan Global se basan en estrategias comprobadas y efectivas para prevenir los siniestros viales.

Bajo este contexto, surge el proyecto en cuestión, que busca aprovechar la información pública para poder analizar y predecir la probabilidad de siniestros viales en la Ciudad Autónoma de Buenos Aires teniendo en cuenta ciertos factores, como el clima, la ubicación del hecho, el horario del suceso, entre otras, de manera que sea posible conocer en qué aspectos o acciones de prevención es más útil trabajar con el objetivo de reducir la mortalidad y lesiones de tráfico en el ámbito de la Ciudad Autónoma de Buenos Aires.

1.2. Problema

La Organización Mundial de la Salud (OMS) en el Plan Global destaca que la mayoría de las colisiones son predecibles y prevenibles. En la actualidad los siniestros viales se encuentran dentro de las 10 principales causas de muerte en el mundo, por lo que la OMS reconoce la importancia del problema y la necesidad de actuar con el objetivo explícito de reducir el número de muertes y lesiones por esta causa.[5]

Actualmente, se registran numerosos datos y se genera una enorme cantidad de información acerca de los siniestros viales que se producen. Por ello, se llevan a cabo muchas investigaciones y estudios que analizan las colisiones causadas por el tránsito para hallar los factores que pueden haber influido en los mismos, como así también los puntos clave o ciertas variables que condicionan estos sucesos. Muchos estudios implican técnicas analíticas que generan información de gran valor en línea con la reducción de la siniestralidad vial y su mortalidad, permitiendo identificar el comportamiento de distintas variables y su influencia en los siniestros viales.

Además, estas técnicas permiten estimar la influencia o importancia de cada uno de los factores de los siniestros viales, de manera que sea posible conocer en qué aspectos es más importante trabajar con el objetivo de reducir los mismos.

1.2.1. Antecedentes

Se relevó que actualmente existe un amplio abanico de estudios realizados para los siniestros viales de distintas ciudades fuera de Argentina, donde por un lado se trabaja en predecir el nivel de siniestrabilidad o severidad (heridas leves o graves, o muertos) y por otro lado se han desarrollado modelos para predecir la cantidad de siniestros viales.

[5] https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

Para predicciones del grado de lesividad de siniestros viales se destaca la utilización de modelos de ensamble como random forest para datos de Madrid (Juan Herrera Briones, 2021) en el cual se concluye que variables de factores atmosféricos tienen menor importancia frente a datos como las infracciones cometidas por el conductor. También se trabajó con modelos de regresión logística para estimar la gravedad de siniestros viales en Bogotá, Colombia (Monroy Varela y Hernando Díaz, 2018), donde se calculan los chances (odds) de que un siniestro sea mortal comparado con que el siniestro no lo sea y se obtuvieron resultados comparativos en las variables día de la semana, localidad y forma de desplazamiento (peatón, ciclista, etc.). Por ejemplo, una colisión causada por el tránsito tiene más probabilidades de tener víctimas mortales en fin de semana, que sí ocurre en día de semana. Otro trabajo utiliza un modelo de árbol de decisión como técnica para el análisis de accidentalidad aplicados a siniestros viales ocurridos en la localidad de Kennedy (ciudad de Bogotá, Colombia) donde tuvo mejores resultados de precisión y rendimiento frente a otras técnicas (como vecinos cercanos o regresión logística) y las variables ambientales tuvieron poca influencia (Diaz y Vargas, 2019).

Por otro lado, hay literatura que muestra cómo se han desarrollado modelos para predecir la cantidad de siniestros viales, por ejemplo, para la ciudad de Valparaíso (Chile) utilizando redes neuronales artificiales, obteniendo buenos resultados (Rojas Godoy, 2015; Acta Polytechnica Hungarica, 2014). También se destaca un sistema de predicción con utilización de regresión logística donde se concluye que los siniestros viales se ven afectados por factores como los tipos de vehículos, la edad del conductor, la antigüedad del vehículo, las condiciones climáticas, la estructura de la carretera, entre otras (IRJET, 2019).

En síntesis, regresión logística, redes neuronales, árboles de decisión, random forest, son técnicas predominantes dentro de los estudios realizados referentes a los siniestros viales.

Asimismo, se destacan técnicas de backcasting[6] en la elaboración de estrategias de seguridad vial o como input necesario para la aplicación de otras metodologías analíticas. El enfoque empírico de Bayes para la estimación de la seguridad vial casi se ha convertido en el estándar para los estudios observacionales de antes y después de las medidas de seguridad vial en los últimos 10 años. El estudio presentado por Rune Elvik muestra que el enfoque empírico de bayes casi siempre proporciona mejores

[6] Un método de planificación que comienza con la definición de un futuro deseable y luego funciona hacia atrás para identificar políticas y programas que conectarán ese futuro específico con el presente

predicciones del número de siniestros viales que el enfoque tradicional de suponer que el número registrado de siniestros es un estimador imparcial del número esperado de siniestros. Sin embargo, las estimaciones no siempre son precisas, ya que, si las diferencias entre las estimaciones y el número real de siniestros son pequeñas y aleatorias, tales imprecisiones son totalmente aceptables, más bien inevitables, dada la aleatoriedad inherente a los recuentos de siniestros. (Rune Elvik - Issue 6, 2008)

También se analizaron pronósticos de víctimas realizados que debían tener en cuenta las predicciones sobre cómo podría cambiar el volumen de viaje por carretera de los distintos modos de transporte y el tipo de nuevas medidas de seguridad vial que podría introducir el Gobierno. Por ejemplo, se destaca lo trabajado para Gran Bretaña, donde se propone identificar nuevas medidas que podrían incluirse en una futura estrategia de seguridad vial y evaluar su probable eficacia. La metodología no puede tener en cuenta desarrollos futuros imprevistos, por lo que las previsiones no han resultado ser correctas en todos los aspectos. Sin embargo, proporciona un marco para examinar las incertidumbres inherentes a los desarrollos futuros y para hacer pleno uso de los conocimientos disponibles, por lo que ha desempeñado un papel valioso en el proceso de formulación de una nueva estrategia de seguridad vial y seguimiento del desarrollo de la seguridad vial. (Broughton, Knowles; 2010)

Como en muchos otros países, Holanda formuló objetivos cuantificados de seguridad vial. Cuando se establecieron estos objetivos, se decidió evaluarlos cada 4 años. Como parte de la primera evaluación, se hicieron previsiones sobre el número de muertes y heridos graves 2020. A partir de estas previsiones se concluyó que el objetivo de heridos graves en 2020 probablemente no se cumpliría sin medidas políticas adicionales. Especialmente los ciclistas y las personas mayores necesitan atención adicional, ya que el número de heridos graves para estos grupos está aumentando. El cumplimiento del objetivo de número de víctimas mortales depende de las inversiones en medidas infraestructurales de seguridad vial y del desarrollo de la movilidad. (Weijermars, Wesemann; 2013)

La literatura reporta numerosos estudios que han demostrado la eficacia de medidas específicas de seguridad vial, pero todas ellas se han basado en datos que fueron recolectados especialmente o estaban disponibles en ciertas áreas.

En línea con lo expuesto anteriormente, se investigaron los proyectos de distintos organismos públicos argentinos que buscan disminuir la accidentabilidad y mejorar la calidad de vida de las personas. A continuación, se explican algunos de estos proyectos:

El Ministerio de Transporte Nacional, posee un Plan Federal de Seguridad Vial

denominado Movilidad Segura, que tiene como objetivo principal disminuir la cantidad de siniestros en las rutas y ciudades del país. En este plan se hace especial hincapié en obtener más y mejor información sobre cómo y dónde se producen las colisiones para poder orientar las políticas de prevención, de controles y de obras (Ministerio de transporte, 2020).

La Secretaría de Transporte y Obras Públicas del Gobierno de la Ciudad de Buenos Aires trabaja en el Plan de Seguridad Vial 2020-2023 de la Ciudad Autónoma de Buenos Aires, el cual tiene como objetivo una reducción del 50% en las víctimas fatales para el año 2030. Éste propone la implementación de acciones eficaces bajo el principio de la Visión Cero, por el cual ninguna muerte ni herido grave en el tránsito es aceptable (Secretaría de Transporte y Obras Públicas GCBA, 2020-2023).

1.3. Objetivo

Los siniestros viales y sus consecuencias pueden ser evitables si se interviene sobre los factores que incrementan la probabilidad de ocurrencia, ya sean estos propios de la conducta humana, condiciones del entorno o propios del funcionamiento de los vehículos, entre otros. Por lo que identificar los factores que los condicionan y predecir la cantidad que ocurrirá en una ventana de tiempo, resulta ser una herramienta de gran ayuda al momento de implementar planes de seguridad vial y/o campañas para la prevención y así evitar que el número de este tipo siniestros siga aumentando.

A partir de información generada por las diferentes entidades públicas y privadas, los avances del análisis de datos y la minería de datos han posibilitado la explotación de ésta para la toma de decisiones a niveles administrativos y gerenciales.

El propósito de esta tesis es el análisis de siniestros viales ocurridos en la Ciudad Autónoma de Buenos Aires durante los períodos 2015 y mayo 2018. Investigar y describir los sucesos con el objetivo de analizar los factores y variables que influyen en los mismos mediante la implementación de algoritmos de machine learning para detectar probabilidad de futuros eventos.

Con la información que analizaremos, la idea es implementar un algoritmo de regresión que permita predecir la cantidad de siniestros viales en determinada zona (barrio/comuna), horario, condición climática y demás variables a analizar.

Este documento se encuentra estructurado de la siguiente manera: en la sección 2 y 3 se presentarán los datos y métodos respectivamente; en la sección 4 se hará un análisis

descriptivo a nivel general de los siniestros viales ocurridos en la Ciudad Autónoma de Buenos Aires, en la sección 5 se mostrarán los resultados de los modelos con sus performance e interpretación, y, por último, en la sección 6, se presentarán las conclusiones del trabajo y recomendaciones.

2. Datos

2.1 Estructura de datos originales

El Observatorio de Movilidad y Seguridad Vial (OMSV) de la Ciudad de Buenos Aires toma como su principal fuente de información datos policiales, tal y como recomiendan los estándares internacionales. Las estadísticas elaboradas se realizan en base a los sumarios que instruye la Policía de la Ciudad con dos clases de delitos que involucran la seguridad vial: lesiones culposas y homicidios culposos.

Hasta 2016 el OMSV se basó en la información provista por la Dirección General de Comisarías (DGC) de la Policía Federal Argentina (PFA). En el año 2017 comenzó a operar la Policía de la Ciudad, dependiente del Ministerio de Justicia y Seguridad del GCBA (MJyS); fuerza que reunió a las dependencias de la PFA (entre ellas la DGC) junto con las de la Policía Metropolitana (Autopistas, Comunas, etc.). La Policía de la Ciudad cubrió así la totalidad del territorio e integró la información que hasta entonces recopilaban las fuerzas de manera independiente.

Gracias a esto, a partir del segundo cuatrimestre de 2018, la principal fuente de información sobre siniestros viales con la que trabaja el Observatorio es la del MJyS; fuente que permite conocer una mayor cantidad de casos por cubrir la totalidad del territorio de la ciudad[7].

La cifra de víctimas fatales informada por el MJyS es cotejada y validada por el Observatorio, que también se nutre de la información proporcionada por múltiples organismos (SAME, AUSA, AUSOL, Hospitales de Agudos de la Ciudad de Buenos Aires, Medios de Comunicación, Cámara Nacional de Apelaciones, Ministerio Público Fiscal de la Ciudad de Buenos Aires), lo que permite que el registro sea lo más completo posible.

Para nuestro trabajo, utilizaremos la base de datos de víctimas lesionadas en siniestros viales ocurridos en la Ciudad Autónoma de Buenos informadas por la Dirección General

[7] https://www.buenosaires.gob.ar/movilidad/plan-de-seguridad-vial/informesestadisticosymapas

de Comisarías (DGC) de la Ex PFA que se obtuvo del sitio web BA DATA[8], los datos son de acceso libre y público. La mencionada base contiene los datos de sucesos ocurridos entre enero del 2015 y mayo del 2018 con 29 columnas y un total de 33.234 registros.

Asimismo, en el sitio web BA DATA se encuentran los datasets con los delitos por siniestro vial informados por el Ministerio de Justicia y Seguridad de 2016 a 2021. Cabe destacar que se analizaron las mismas, y por tratarse de fuentes de datos distintas los registros difieren entre ambas bases para los mismos períodos. Dado que en principio se relevó la información de 2015 a 2018 provista por la DGC (Ex PFA), se trabajaron estos datos, y luego se analizó la posibilidad de utilizar los datos posteriores para validación de nuestro modelo. Ya que, si bien son fuentes distintas, las diferencias no son significativas y se trata de una misma unidad de análisis, que son los siniestros viales, y hoy en día esa información la trabaja el MJyS.

La base de datos a utilizar contiene información referente a cada siniestro vial que comprende fecha y hora exacta del mismo, ubicación del hecho, tipo de transporte involucrado (moto, bicicleta, automóvil, etc.), además si la causa se trató de homicidio o lesiones, también contiene el género, edad y rol de la víctima (peatón, conductor, pasajero). En cada hecho puede haber una o más víctimas involucradas y uno o más vehículos involucrados.

Es útil exponer algunas especificaciones conceptuales necesarias para el entendimiento de las variables disponibles en nuestros datos. Las mismas fueron obtenidas del Glosario del Observatorio de Seguridad Vial de la Ciudad de Buenos Aires. En el Apéndice A se describen en profundidad.

- **Siniestro vial:** cualquier hecho de tránsito con implicación de al menos un vehículo en movimiento, que tenga lugar en una vía pública o en una vía privada a la que la población tenga derecho de acceso, y que tenga como consecuencia al menos una persona herida o muerta.
- **Víctima**: cualquier persona muerta o herida como consecuencia de un siniestro vial. En este trabajo consideramos víctima a la persona registrada como lesionada o muerta en los registros y no al dato de cantidad de víctimas expuesto en nuestro set, que indica la cantidad de víctimas del siniestro vial.

[8] https://data.buenosaires.gob.ar/dataset/victimas-siniestros-viales

- **Homicidio:** se considera la definición avalada por organismos internacionales de víctima fatal del siniestro de tránsito como aquella persona que fallece de inmediato o dentro de los 30 días siguientes como consecuencia de un traumatismo causado por el siniestro (se exceptúan los suicidios).[9]
- **Lesionados:** víctimas que, como consecuencia de un siniestro vial con víctimas, no resulte muerta en el acto o dentro de los 30 días siguientes, pero sufra lesiones. Normalmente, estas lesiones requieren tratamiento médico.
- **Rol:** corresponde a la forma de desplazamiento de la víctima ya sea ciclista, conductor, pasajero o peatón, o una combinación de estas.
- **Tipo (corresponde a tipo de vehículo):** caracteriza a las víctimas fatales y lesionadas según el medio en el que se transportan –tipo de vehículo o forma de desplazamiento– en el momento del siniestro. En el caso de las víctimas de rodados motorizados, se diferencia a las mismas según el tipo de vehículo.

A continuación, se puede ver una muestra de los datos obtenidos:

Tabla 1:Muestra de víctimas de siniestros viales de la base de datos BA Data

Primeras 13 columnas

causa	rol	tipo	sexo	edad	mes	periodo	fecha	hora	lugar_hecho	direccion_normalizada	tipo_calle	direccion_normalizada_arcgis
homicidio	conductor	moto			2	2015	2/14/2015	19:00:00	cafayate y severo garcia gra	cafayate y garcia grande de zeq	calle	cafayate & garcia grande de zeq
homicidio					2	2015	2/25/2015	3:00:00	lugones, leopoldo av. y udaor	lugones, leopoldo av. y udaondo,	avenida	lugones, leopoldo av. & udaondo,
homicidio	peaton	peaton	femenino		2	2015	2/27/2015	8:00:00	avda jujuy y avda independer	jujuy av. e independencia av.	avenida	jujuy av. & independencia av.
homicidio					3	2015	3/2/2015	18:30:00	lavalle 1730	lavalle 1730	calle	1730 lavalle
homicidio	pasajero	camion	masculino		4	2015	4/9/2015	1:20:00	ave salvador m del carril 243	carril, salvador maria del av. 243	avenida	2434 carril, salvador maria del av
homicidio	conductor	moto			4	2015	4/30/2015	23:30:00	lima y carlos calvo	lima y calvo, carlos	calle	lima & calvo, carlos
homicidio	conductor	moto	masculino	18	1	2015	1/1/2015	2:20:00	piedras y av independencia	piedras e independencia av.	avenida	piedras & independencia av.
lesiones	pasajero	automovil	masculino	26	1	2015	1/1/2015	7:10:00	2850 yerbal	yerbal 2850	calle	2850 yerbal
homicidio	conductor	moto	masculino	24	3	2015	3/14/2015	4:30:00	av gral paz colectora y jorge	paz, gral. av. y chavez, jorge	avenida	paz, gral. av. & chavez, jorge
homicidio		moto	masculino		9	2015	9/4/2015	21:14:00	avda del campo y avda punta	del campo av. y punta arenas	avenida	del campo av. & punta arenas
lesiones	conductor	moto	masculino	37	4	2015	4/1/2015	15:25:00	av. coronel diaz 1520	diaz, cnel. av. 1520	avenida	1520 diaz, cnel. av.
homicidio	pasajero	automovil	masculino	19	8	2015	8/14/2015	2:30:00	av pte figueroa alcorta y av d	figueroa alcorta, pres. av. y de la	avenida	figueroa alcorta, pres. av. & de lo
homicidio	pasajero	automovil	masculino	23	8	2015	8/14/2015	2:30:00	av pte figueroa alcorta y av d	figueroa alcorta, pres. av. y de la	avenida	figueroa alcorta, pres. av. & de lo
homicidio	pasajero	transporte pu	masculino	20	10	2016	10/31/2016	2:58:00	av juramento y vuelta de oblig	juramento av. y vuelta de obligad	avenida	juramento av. & vuelta de obligad
homicidio	conductor	moto	masculino	23	1	2017	1/16/2017	16:30:00	av. directorio y riglos	directorio av. y riglos	avenida	directorio av. & riglos
homicidio	pasajero	automovil	femenino	23	2	2017	2/26/2017	5:15:00	av. perito moreno y fournier	moreno, perito av. y fournier	avenida	moreno, perito av. & fournier

Siguientes 16 columnas

calle1	altura	calle2	codigo_calle	codigo_cruce	geocodificacion	semestre	x	y	geom	cantidad_victimas	comuna	geom_3857	tipo_colision1	participantes_victimas	participantes_acusados
cafayate		garcia gra	3015	27019	point(95841.9588	1	-58,5	-35	0101000020E61	1	9	0101000020	11 motovehiculo - v	moto	automovil
lugones, leopoldo a		udaondo, g	12152	22003	point(101433.767	1	-58,4	-35	0101000020E61	1	13	0101000020	110F0000BFBC45	NULL	
jujuy av.		independe	10013	9010	point(105482.771	1	-58,4	-35	0101000020E61	1	3	0101000020	11 peaton - vehicul	peaton	"transporte publico"
lavalle	1730		12089		point(106601.685	1	-58,4	-35	0101000020E61	1	1	0101000020	110F00005621683	NULL	
carril, salva	2434		4039		point(97284.7646	1	-58,5	-35	0101000020E61	1	15	0101000020	11 vehiculo - vehici	camion	camion
lima		calvo, carl	12112	3028	point(107506.679	1	-58,4	-35	0101000020E61	1	1	0101000020	11 motovehiculo - v	moto	automovil
piedras		independe	17078	9010	point(107881.450	1	-58,4	-35	0101000020E61	1	1	0101000020	11 motovehiculo - v	moto	automovil
yerbal	2850		26003	0	POINT(99434.113	1	-58,5	-35	0101000020E61	1	7	0101000020	11 vehiculo - vehici	automovil	automovil
paz, gral. av.		chavez, jor	17041	3224	point(94030.4783	1	-58,5	-35	0101000020E61	1	9	0101000020	11 motovehiculo - r	moto	moto
del campo av.		punta aren	4037	17138	point(99486.6219	2	-58,5	-35	0101000020E61	1	15	0101000020	11 motovehiculo - v	moto	automovil
diaz, cnel. a	1520		4073	0	POINT(104578.72	1	-58,4	-35	0101000020E61	1	14	0101000020	11 vehiculo - motov	moto	automovil
figueroa alcorta, pr		de los omb	6025	4515	point(103286.903	2	-58,4	-35	0101000020E61	3	14	0101000020	11 vehiculo - vehici	automovil	automovil
figueroa alcorta, pr		de los omb	6025	4515	point(103286.903	2	-58,4	-35	0101000020E61	3	14	0101000020	11 vehiculo - vehici	automovil	automovil
juramento av.		vuelta de o	10017	16002	point(100700.716	2	-58,5	-35	0101000020E61	1	13	0101000020	11 vehiculo - vehici	"transporte publico"	"transporte publico"

En el Apéndice B se describe en profundidad el perfilado de cada variable (quality report) donde se expone, de qué tipo son las variables (data type), así como su cantidad de valores

[9] El MJyS incluye casos de víctimas fatales que ocurren en el lugar del hecho y hasta 7 días posteriores al siniestro, y como lesionados, víctimas que no resultan muertas en el acto o dentro los 6 días siguientes.

faltantes y únicos, dado que nuestras variables no implican valores numéricos no analizamos sus mínimos, máximos, promedios y demás medidas estadísticas aplicables sólo a este tipo de datos.

2.2 Transformación de datos y agregado de variables

Para el armado de nuestro modelo nos pareció interesante crear algunas variables que caracterizan al siniestro vial y pueden ser condicionantes del suceso del hecho. Es por eso que a continuación se detallan algunas transformaciones de los datos originales y las variables que crearemos para un mejor análisis. Todas estas variables junto con los datos originales serán analizadas en la sección siguiente para determinar cuáles serán las variables a incorporar en nuestro modelo.

Barrio

Si bien en nuestros datos originales contamos con el dato de las direcciones donde ocurrieron los sucesos, éstas son muy granulares, y el dato de las comunas correspondientes es muy general, por lo que nos pareció más preciso contar con el dato del barrio. Con el dataset publicado en BA DATA de barrios[10], obtuvimos los límites y ubicación geográfica de los barrios de la Ciudad Autónoma de Buenos Aires. A partir de nuestros datos geoespaciales obtuvimos los datos geométricos de geometrías WKT[11] para poder unirlo con los datos de barrios y así obtener el barrio del siniestro vial.

Fecha

Observando el dataset, notamos que la variable fecha no se encontraba estandarizada, dado que algunas fechas estaban consignadas con separación por “-” y otras por “/”, y a su vez tenían distintos formatos por lo que estandarizamos el dato en formato fecha separando año, mes y día con una "/" para su posterior manipulación en operaciones aritméticas de fechas.

Hora

[10] https://data.buenosaires.gob.ar/dataset/barrios

[11] El formato WKT o Well Known Text es un formato de codificación específicamente diseñado para la caracterización y almacenamiento de objetos geométricos espaciales en formato vectorial. Se trata de un formato muy expandido dentro del mundo de las geotecnologías ya que se trata de un estándar definido por el OGC y por ello se ha adoptado por una gran cantidad de Sistemas de Información Geográfica y es aceptado por otras herramientas y librerías geoespaciales.

Dado que los datos de la hora del siniestro vial se encontraban especificados con hora y minutos, optamos por estandarizar la hora creando una variable que contenga solo el dato de la hora sin minutos.

Para un mayor análisis agregamos variables que pueden influir en la ocurrencia de un siniestro vial, las mismas se detallan a continuación:

Días

Día: a partir del dato de la fecha estandarizada generamos una columna solo con el día del suceso.

Día de la semana: se incorporó a la base de datos el día de la semana (lunes, martes, miércoles, jueves, viernes, sábado o domingo)

Otro dato que consideramos que podría aportar en nuestro análisis y modelo, es si el día del suceso correspondía a un día hábil o no.

Fin de semana: para esto utilizamos el dato generado "día de la semana" y generamos la variable dummy fin de semana, considerando sábado y domingo como valor = 1, y lunes, martes, miércoles, jueves y viernes con valor = 0.

Feriados: en línea con lo expuesto anteriormente agregamos en nuestro dataset otra variable dummy correspondiente a sí la fecha del siniestro vial corresponde a un feriado nacional argentino o no. Para esto generamos un archivo con las fechas de los días feriados para determinar si el día del suceso corresponde o no a un día feriado.

Hábil: a partir de las variables fin de semana y feriados, integramos en una única variable los días que son hábiles, es decir laborales, de los que no, como los días de fin de semana (sábado y domingo) y festivos/feriados.

Clima

Se obtuvo del Centro de Información del Servicio Meteorológico Nacional Argentino las precipitaciones diarias y por hora de las precipitaciones en mm correspondientes a las 6 horas anteriores, desde 2015 en adelante, provenientes de la estación meteorológica "Buenos Aires Observatorio" que se encuentra en las inmediaciones de la Facultad de Agronomía.

Lluvia: se agregó como dato una variable dummy donde toma valor 1 si la precipitación en mm en las 6 horas anteriores al suceso es mayor a cero, y valor 0 de lo contrario.

Intensidad lluvia: para discretizar los mm de precipitación se agregó como dato una variable categórica donde se expresa la intensidad de la lluvia según los siguientes valores de mm. Los mismos fueron definidos en función del Manual de uso de términos meteorológicos del AEMET.[12]

Tabla 2:Intensidad de lluvia según acumulación en mm

Intensidad	Acumulación
Sin lluvia	entre 0 a 0,09 mm
Débil	entre 0,1 a 2 mm
Moderado	entre 2,1 a 15 mm
Fuerte	entre 15,1 a 30 mm
Muy fuerte	entre 30,1 a 60 mm
Torrencial	más de 60 mm

Fuente: AEMET

Estación del año

Por otro lado, nos pareció útil a partir de la fecha del suceso agregar como dato la estación del año en la que ocurrió el siniestro vial, esto para analizar si hay alguna época del año con mayores colisiones causadas por el tránsito.

- Verano (21 de diciembre a 20 de marzo)
- Otoño (21 de marzo a 20 de junio)
- Invierno (21 de junio a 20 de septiembre)
- Primavera (21 de septiembre a 20 de diciembre).

Diurno/nocturno

Asimismo, se generó la variable "luz" para evaluar la visibilidad en función de si era de día o de noche a la hora del suceso, considerando la noche como un entorno más peligroso por varias razones: es más difícil ver en la oscuridad; uno puede cegarse temporalmente por el resplandor de las luces de otros vehículos; y es probable que haya más conductores que estén cansados o bajo la influencia de sustancias.

[12] La Agencia Estatal de Meteorología de España es una agencia estatal, cuyo objetivo básico es la prestación de servicios meteorológicos, que sean competencia del Estado.

Si bien según lo investigado el Observatorio de Seguridad Vial de la Ciudad de Buenos Aires define una clasificación diurno/nocturno de los siniestros viales según el momento del día en que ocurren con horarios fijos[13], consideramos que sería más enriquecedor analizar los sucesos en función de la salida y puesta del sol para definir con mayor precisión si en la hora del suceso había luz natural o no. Considerando como "luz = 1" si la hora del acontecimiento se encuentra dentro del rango de la salida y puesta de sol, y de "luz = 0" si se encuentra por fuera de este rango. Los datos de salida y puesta del sol fueron obtenidos de manera manual del sitio web sunrise-and-sunset.com[14].

Todas estas estas variables nos servirán luego para hacer un análisis descriptivo de los datos y para su posible incorporación en el modelo.

3. Metodología

Existen diversas técnicas para descubrir patrones en los datos, en este caso utilizaremos técnicas de aprendizaje automático las cuales se centran en una categoría de algoritmos que aprenden, o mejoran el rendimiento, en función de los datos que consumen.

Existen tres grandes familias de algoritmos de aprendizaje automático. El aprendizaje supervisado en el que el proceso de aprendizaje está dirigido por una variable objetivo, a diferencia del aprendizaje no supervisado, en el que los datos no tienen etiquetas y, por lo tanto, los algoritmos sólo buscan cualquier tipo de patrón. Por último, el aprendizaje por refuerzos se trata de algoritmos que se enfocan en aprender mediante la interacción con un ambiente, por lo que el sistema aprende a base de ensayo-error basado en obtención de recompensas (Sanjiv Ranjan Das, 2017).

En este proyecto abordaremos técnicas de aprendizaje supervisado que se caracterizan por predecir una variable, donde denominamos a los problemas con una respuesta cuantitativa como problemas de regresión, mientras que aquellos que involucran una respuesta cualitativa problemas de clasificación (James, Witten, Hastie, Tibshirani, 2017).

Al momento de modelar se toman muchas decisiones, por ejemplo:

- ✓ Qué datos tomar como inputs para entrenar el modelo;
- ✓ Qué procesamiento de datos hacer;

[13] Diurnos: siniestros ocurridos entre las 07:00 hs y las 18:59 hs y nocturnos: entre las 19:01 hs y las 06:59 hs.

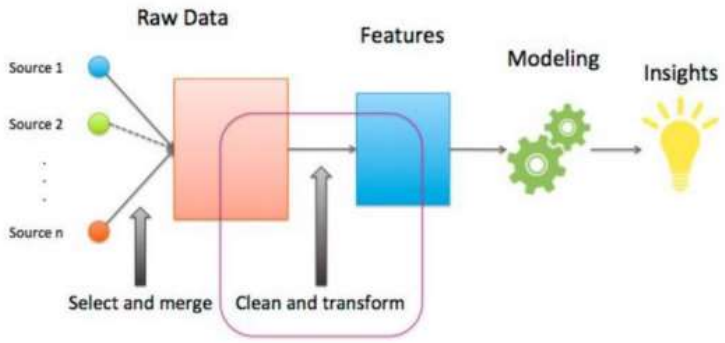
[14] https://www.sunrise-and-sunset.com/es/sun/argentina/buenos-aires

- ✓ Qué modelo usar para predecir y;
- ✓ Con qué valores de hiperparámetros entrenar al modelo

La eficacia del algoritmo de aprendizaje dependerá en buena medida de los atributos de los que disponga para aprender, por lo que generar atributos para entrenar el modelo de aprendizaje es una etapa importante. Las buenas funciones no solamente deben representar aspectos destacados de los datos, sino también ajustarse a las suposiciones del modelo (Zheng A. y Casari A).

Las técnicas tradicionales de ingeniería de atributos varían en función de si se trata de atributos numéricos o categóricos. Otro ejemplo es, a partir de datos como fechas y hora considerar año, mes, día, hora, día de la semana, semana del año, etc.

Figura 2: Etapas del modelado



Fuente: Zheng A. y Casari A. Feature Engineering for Machine Learning

Una vez realizadas las transformaciones necesarias para nuestro modelado, se debe determinar qué atributos se considerarán. Una estrategia utilizada para la selección de atributos se denomina filtering, que consiste en eliminar atributos antes de entrenar por considerar que tienen poca probabilidad de tener poder predictivo, ya sea por exploración manual, o por algún criterio estadístico.

Por lo que antes de entrenar un modelo predictivo, es muy importante realizar una exploración descriptiva de los atributos. Este proceso permite entender mejor qué información contiene cada variable, así como detectar posibles errores (Kelleher, Mac Namee y D'Arcy, 2015). En esta etapa, se comprenden características tales como los tipos y rangos de valores de una variable, y se pueden detectar problemas de calidad que podrían afectar negativamente a los modelos.

Junto con el estudio de variables, es importante analizar los datos faltantes. Los conjuntos de datos pueden tener valores perdidos, y esto puede causar problemas para muchos algoritmos de aprendizaje. Asimismo, se debe considerar si es necesario realizar transformaciones sobre los datos con el objetivo de que puedan ser interpretados por el algoritmo lo más eficientemente posible. Por ejemplo, la transformación de variables categóricas a número. Como así también identificar e imputar los valores faltantes para cada columna en sus datos de entrada antes de modelar o bien eliminarlos.

Binarización de variables cualitativas (Encoding Categorical Variables)

Un tipo de preprocesado que se suele hacer en los datos es la binarización (one-hot encoding), que consiste en crear nuevas variables dummy con cada uno de los niveles de las variables cualitativas.

El siguiente paso tras definir y transformar los datos para nuestro modelo, es seleccionar el algoritmo que se va a utilizar.

3.1 Modelos

Los métodos basados en árboles se han convertido en uno de los referentes dentro del ámbito predictivo debido a los buenos resultados que generan en problemas muy diversos.

3.1.1 Classification and Regression Trees (CART)

Los Classification and Regression Trees (CART) conocidos como los árboles de decisión de clasificación o regresión fueron propuestos por Breiman en 1984. Se trata de modelos predictivos que consisten principalmente en la realización de particiones binarias con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta (dependiente) del modelo.

Son árboles de regresión el subtipo de árboles de predicción que se aplica cuando la variable respuesta es continua y árboles de clasificación cuando la variable dependiente es cualitativa. En el entrenamiento de un árbol, las observaciones se van distribuyendo por los nodos generando la estructura del árbol hasta alcanzar un nodo terminal.

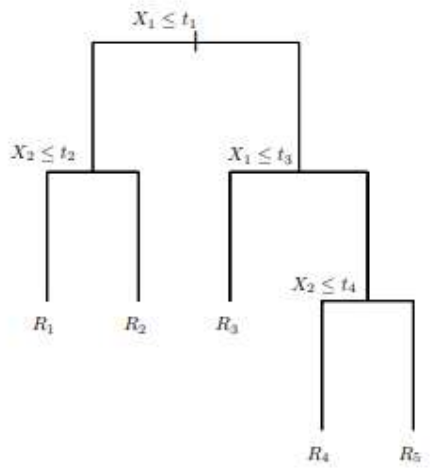
El proceso de entrenamiento de un árbol de regresión se divide en dos etapas (James, Witten, Hastie, Tibshirani, 2017):

- División del espacio predictor, es decir el conjunto de valores posibles para X_1, X_2, …, X_p, en J regiones distintas y no superpuestas (nodos terminales) R_1, R_2, R_3, ..., R_j.

- Predicción de la variable respuesta en cada región. Para cada observación que cae en la región R_j, hacemos la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones de entrenamiento en R_j.

Asimismo, es necesario establecer una metodología que permita crear las regiones R_1, R_2, R_3, ..., R_j o lo que es equivalente, decidir donde se introducen las divisiones: en que predictores y en qué valores de estos.

Figura 3: Ejemplo de un árbol con dos variables independientes (X1; X2) con cinco particiones



Fuente: James, Witten, Hastie, Tibshirani, 2017

En los árboles de regresión, el criterio empleado con más frecuencia para identificar las divisiones es el Residual Sum of Squares (RSS). El objetivo es encontrar las J regiones (R_1, ..., R_j) que minimizan el RSS total:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

donde $\hat{y}_{R_j}$es la media de la variable respuesta en la región R_j. Es decir, que se busca una distribución de regiones tal que, la sumatoria de las desviaciones al cuadrado entre las observaciones y la media de la región a la que pertenecen sea lo menor posible.

Desafortunadamente, es computacionalmente inviable considerar todas las particiones posibles del espacio de características en cajas *J*. Por esta razón, se recurre a lo que se conoce como recursive binary splitting (división binaria recursiva) donde el objetivo es

encontrar en cada iteración, el predictor X_j y el umbral s tal que, si se distribuyen las observaciones en las regiones $\{X_j < s\}$ y $\{X_j \geq s\}$ se consigue la mayor reducción del RSS.

Tras la creación de un árbol, las observaciones de entrenamiento quedan agrupadas en los nodos terminales. Para predecir una nueva observación, se recorre el árbol en función de los valores que tienen sus predictores hasta llegar a uno de los nodos terminales. En el caso de regresión, el valor predicho suele ser la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo. Si bien la media es el valor más empleado, se puede utilizar cualquier otro (mediana, cuantil). (James, Witten, Hastie, Tibshirani, 2017).

El proceso de construcción de árboles se ajusta muy bien a las observaciones empleadas como entrenamiento, por lo que se puede generar overfitting, es decir, el modelo se ajusta tanto a los datos de entrenamiento que es incapaz de predecir correctamente nuevas observaciones; lo que reduce su capacidad predictiva al aplicarlo a nuevos datos. La razón de este comportamiento radica en la facilidad con la que los árboles se ramifican adquiriendo estructuras complejas. Existen dos estrategias para prevenir el problema de overfitting de los árboles: limitar el tamaño del árbol y el proceso de podado.

El tamaño final que adquiere un árbol puede controlarse mediante reglas de "early stopping" que detengan la división de los nodos dependiendo de si se cumplen o no determinadas condiciones. Como, por ejemplo, observaciones mínimas para división, profundidad máxima del árbol, número máximo de nodos terminales, entre otras. Al evaluar las divisiones sin tener en cuenta las que vendrán después, puede no elegirse la opción que resulta en el mejor árbol final. A este tipo de estrategias se les conoce como greedy. Una alternativa no greedy que consigue evitar el overfitting consiste en generar árboles grandes, sin condiciones de parada más allá de las necesarias por las limitaciones computacionales, y después podarlos (tree pruning), manteniendo únicamente la estructura robusta que consigue un error bajo (Hastie, Tibshirani, Friedman, 2008).

3.1.2 Random Forest

Un modelo Random Forest está formado por un conjunto (ensamble) de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping. En cada árbol individual, las observaciones se van distribuyendo por los nodos generando la estructura del árbol hasta alcanzar un nodo terminal por lo que la predicción de una nueva observación se obtiene

agregando las predicciones de todos los árboles individuales que forman el modelo. Para entender cómo funcionan los modelos Random Forest es necesario conocer primero los conceptos de ensamble y bagging.

Todos los modelos de aprendizaje estadístico y machine learning sufren el problema de equilibrio entre sesgo y varianza. El término sesgo hace referencia a cuánto se alejan en promedio las predicciones de un modelo respecto a los valores reales y el término varianza hace referencia a cuánto cambia el modelo dependiendo de los datos utilizados en su entrenamiento.

Los métodos de ensamble combinan múltiples modelos en uno nuevo con el objetivo de lograr un equilibrio entre sesgo y varianza, consiguiendo así mejores predicciones que cualquiera de los modelos individuales originales. La clave para que consigan mejores resultados es que los modelos que los forman sean lo más diversos posibles (Hastie, Tibshirani, Friedman, 2008).

A continuación, se exponen dos de las estrategias de ensamble más utilizadas en las que cada una reduce una parte del error total:

- Boosting: los modelos simples son utilizados secuencialmente, de forma que cada modelo aprende de los errores del anterior. Se emplean modelos con muy poca varianza, pero mucho sesgo, ajustando secuencialmente los modelos se reduce el sesgo. Para predecir, todos los modelos que forman el ensamble participan aportando su predicción y como valor final para variables continuas, se toma la media de todas las predicciones, y para clasificación, la clase más frecuente. Tres de los métodos de boosting más empleados son AdaBoost, Gradient Boosting y Stochastic Gradient Boosting.
- Bagging: proviene del concepto de "agregación de Bootstrap", y hace referencia al empleo del muestreo repetido con reposición bootstrapping. Se emplean modelos con muy poco sesgo pero mucha varianza, agregándolos se consigue reducir la varianza sin incrementar el sesgo. Al igual que boosting se toma la media para modelos de regresión y la clase más frecuente para modelos de clasificación.

El algoritmo de Random Forest es una modificación del proceso de bagging que consigue mejorar los resultados gracias a que decorrelaciona aún más los árboles generados en el proceso (James, Witten, Hastie, Tibshirani, 2017). Random forest realiza una selección aleatoria de m predictores antes de evaluar cada división, este número de predictores será siempre menor que el número total de variables de entrada del modelo y va a ser constante en todo el proceso de formación del árbol.

Por lo que los dos parámetros más importantes de un modelo Random Forest son los siguientes:

- Ntree: número total de árboles de decisión que componen el conjunto.
- Mtry: número de m variables que se tienen en cuenta en cada nodo para su partición.

El resultado de predicción de cada modelo y su tasa de error dependen en gran medida de los valores escogidos para estos parámetros. A mayor número de árboles que conformen el modelo, mayor será su precisión, ya que cuenta con mayor cantidad de datos, pero en cierta cantidad de árboles el error se estabiliza y ya no disminuye más. Una reducción del valor m de variables tenidas en cuenta para la partición de cada nodo provoca que disminuya la correlación existente entre los distintos árboles que forman el modelo, pero esto va acompañado de una disminución de la precisión de los árboles generados, al contar con menos posibilidades de bifurcación de cada nodo.

Por lo que se recomienda utilizar, siendo n el número total de variables de entrada (independientes) del modelo (Hastie, Tibshirani, Friedman):

- Para problemas de clasificación: $Mtry = \sqrt{n}$
- Para problemas de regresión: $Mtry = \frac{n}{3}$

Encoding Categorical Variables: los modelos basados en árboles de decisión, entre ellos Random Forest, son capaces de utilizar predictores categóricos en su forma natural sin necesidad de convertirlos en variables dummy mediante one-hot-encoding. Sin embargo, en la práctica, depende de la implementación que tenga la librería o software utilizado. Por lo que binarizar tiene impacto directo en la estructura de los árboles generados y, en consecuencia, en los resultados predictivos del modelo y en la importancia calculada para los predictores.

3.2 Conjuntos de entrenamiento y test

Para estimar cómo impactan las decisiones de modelado en la performance de nuestro modelo, se requiere comprobar cuán próximas son las predicciones a los verdaderos valores de la variable respuesta.

3.2.1. Validation/Holdout Set

La manera más simple de simular el comportamiento sobre datos desconocidos es separar una submuestra al azar de observaciones del training set como conjunto de validación.

A continuación, se expone una visualización esquemática del enfoque del conjunto de validación. Un conjunto de *n* observaciones se divide aleatoriamente en un conjunto de entrenamiento (se muestra en azul, que contiene las observaciones 7, 22 y 13, entre otras) y un conjunto de validación (que se muestra en naranja y que contiene la observación 91, entre otras). El método de aprendizaje estadístico se ajusta al conjunto de entrenamiento y su rendimiento se evalúa en el conjunto de validación.

Figura 4:Conjunto de entrenamiento y validación (Validation/Holdout Set)

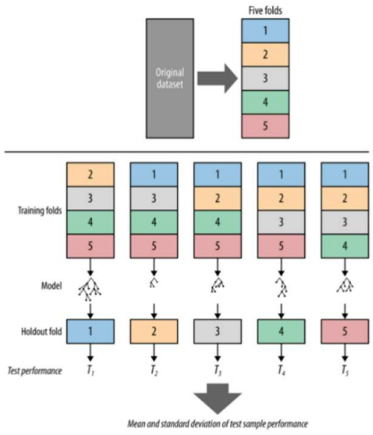
Fuente: James, Witten, Hastie, Tibshirani, 2017

El tamaño adecuado de las particiones depende en gran medida de la cantidad de datos disponibles y la seguridad que se necesite en la estimación del error, 80%-20% suele dar buenos resultados.

3.2.2 K-fold cross-validation

Este enfoque implica la división de validación cruzada (cv) en grupos de muestras aleatorias, llamados pliegues de igual tamaño (si es posible). La función de predicción se aprende usando (*k-1)* pliegues, y el pliegue que queda fuera se usa para la validación. Típicamente los valores de *k* que se usan son 3, 5 ó 10. Como el resultado también depende de particiones al azar, para mejorar la estimación de la performance en testeo, a veces se repite el proceso varias veces.

Figura 5:Imagen ilustrativa del proceso de validación K-fold cross-validation



Fuente: Machine Learning con python (sitio web)[15]

3.3 Métrica de evaluación de modelos

Con un modelo de regresión, predecimos o estimamos el valor numérico de una cantidad desconocida, de acuerdo con unas características dadas. La diferencia entre la predicción y el valor real es el error.

Para estimar el rendimiento y evaluar el ajuste del modelo se considerará el RMSE (Root Mean Squared Error) que es la raíz cuadrada del error cuadrático medio, es decir, la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Se toma la raíz cuadrada para que el error tenga las mismas unidades de lo que estamos midiendo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Indica el ajuste absoluto del modelo a los datos, cuán cerca están los puntos de datos observados de los valores predichos por el modelo. Los valores más bajos de RMSE indican un mejor ajuste. RMSE es una buena medida de la precisión con que el modelo predice la respuesta, y es el criterio más importante para ajustar si el propósito principal del modelo es la predicción.

3.4 Optimización de hiperparámetros

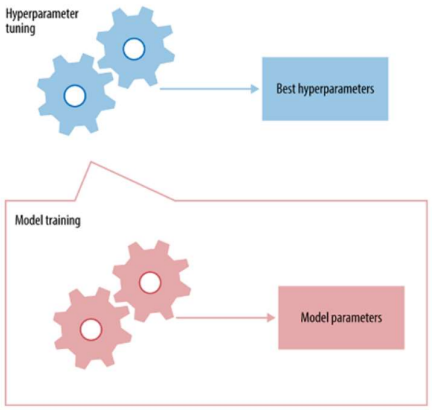
Un hiperparámetro es una característica que es externa al modelo y cuyo valor no se puede estimar a partir de los datos. En Random Forest se destacan aquellos que detienen el crecimiento de los árboles, los que controlan el número de árboles y predictores incluidos, y los que gestionan la paralización.

- **Número de árboles:** en Random Forest, el número de árboles no es un hiperparámetro crítico, dado que no se produce overfitting por exceso de árboles, mejora el resultado, pero añadir árboles una vez que la mejora se estabiliza implica un costo computacional sin sentido.
- **Max features:** el número de predictores considerados a en cada división, si es uno de los hiperparámetros más importantes de Random Forest, dado que es el que permite controlar cuánto decorrelacionan los árboles entre sí.

[15] https://relopezbriega.github.io/blog/2016/05/29/machine-learning-con-python-sobreajuste/

En Apéndice C se exponen los hiperparámetros a optimizar en Random Forest. El análisis individual de los hiperparámetros es útil para entender el impacto en el modelo e identificar el rango de interés, pero es preferible recurrir a búsquedas aleatorias, que son técnicas que se utilizan para encontrar el óptimo de hiperpárametros de un modelo analizando varias combinaciones de hiperpárametros que da como resultado las predicciones más precisas, de esta forma, se consigue explorar el espacio de búsqueda de una forma más distribuida.

Figura 6:Imagen ilustrativa del proceso de ajuste de hiperparámetros



Fuente: sitio web[16]

Para ello existen dos formas:

- Grid search: se hace una búsqueda exhaustiva sobre un conjunto de valores previamente definidos por el usuario.
- Random search: se evalúan valores aleatorios dentro de unos límites definidos por el usuario.

Para cada combinación de valores, se entrena el modelo y se estima su error mediante un método de validación.

Estas estrategias mencionadas generan buenos resultados, pero no tienen en cuenta los resultados obtenidos hasta el momento, lo que impide focalizar en búsqueda de regiones de mayor interés y evitar regiones innecesarias. Para ellos existe la alternativa de búsqueda de hiperparámetros mediante optimización bayesiana. Esta última consiste en

[16] https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

crear un modelo probabilístico en el que la función objetivo es la métrica de validación del modelo (rmse). Con esta estrategia, se consigue que la búsqueda se vaya redirigiendo en cada iteración hacia las regiones de mayor interés. El objetivo final es reducir el número de combinaciones de hiperparámetros con las que se evalúa el modelo, eligiendo únicamente los mejores candidatos.

4. Análisis Exploratorio

En función de las transformaciones realizadas a las variables disponibles, y los atributos creados mencionados en la sección 2, tenemos un dataset de 33.234 registros y 43 columnas.

En esta sección realizaremos un análisis descriptivo y exploratorio de las variables para, por un lado, obtener las primeras características relevantes de los siniestros viales en la Ciudad Autónoma de Buenos Aires y sus víctimas; y, por otro lado, obtener información de los datos para futuras recomendaciones para la toma de decisiones y generación de planes de acción para la prevención de siniestros viales.

Parte de los resultados de este primer análisis nos llevarán a realizar transformaciones en las variables para luego utilizarlas en la etapa del modelo.

Datos faltantes:

Uno de los primeros pasos al momento de analizar datos, es hacer un conteo de los llamados valores faltantes o missing values ya que podrían afectar negativamente a los modelos. En función de nuestro quality report, a continuación, puede observarse las cuatro primeras variables con mayor cantidad de missing values que posee el dataset.

Figura 7:Primeras cuatro variables con mayor porcentaje de valores nulos del dataset

altura	78.658702
codigo_cruce	29.883926
calle2	24.898181
codigo_calle	18.918002

En el Apéndice D se expone un gráfico con la distribución de nulos por variables.

Considerando la cantidad de nulos de las variables geográficas y características de la ubicación del hecho, no las incluiremos en el posterior armado del modelo. Dado que todas estas variables que exponen ubicaciones específicas del suceso, como calle y

numeración, toman demasiados valores únicos es muy poco probable que se detecte un patrón interesante. No obstante, si consideraremos y analizaremos la comuna y/o barrio de los sucesos.

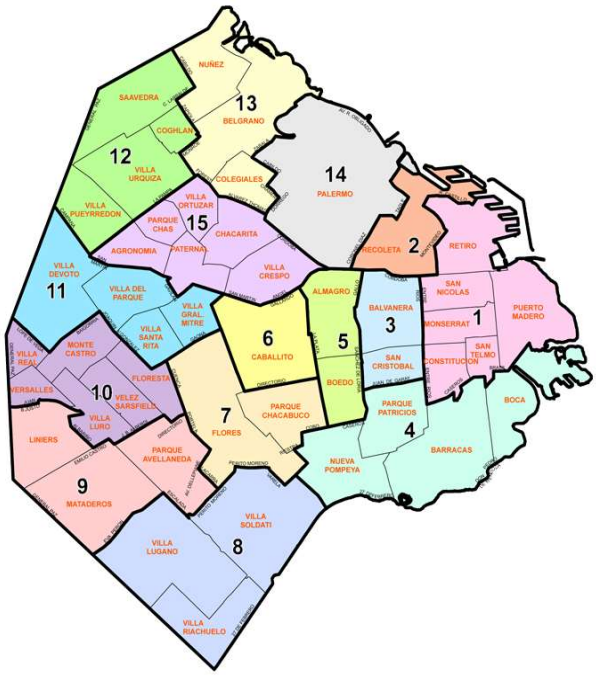
A partir de este primer filtro de variables, comenzamos a analizar las variables que sí nos parecieron interesantes para poder entender las características de los siniestros viales y sus víctimas. Para el siguiente análisis consideramos los siniestros viales de 2015 a 2017 dado que 2018 no se encuentra completo y eso altera algunas estadísticas generales.

4.1 Análisis de los siniestros viales

Comuna y barrio

La Ciudad de Buenos Aires se conforma de 48 barrios que se encuentra organizados en 15 comunas[17] que se corresponden con uno o más barrios de la ciudad.

Figura 8:Mapa de los barrios y comunas de la Ciudad Autónoma de Buenos Aires



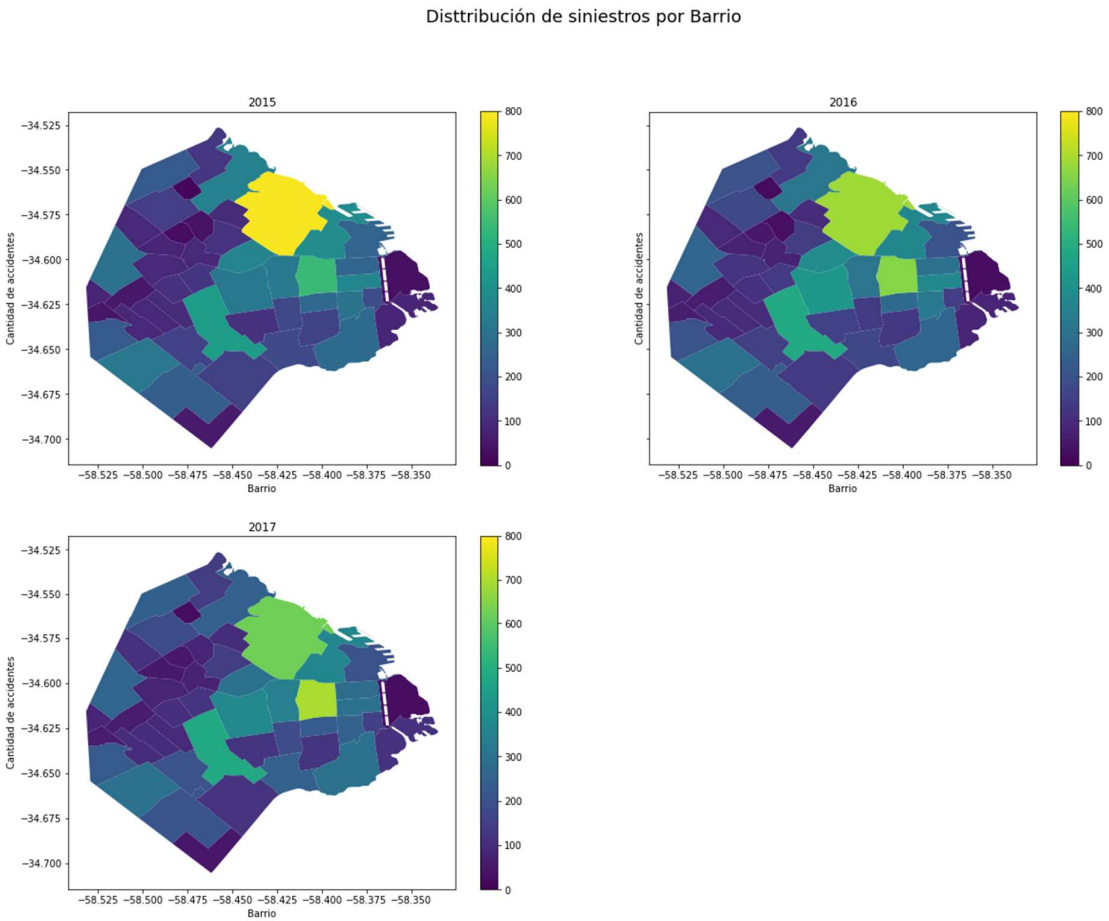
Fuente: buenosaires.gob.ar

[17]Es una subdivisión administrativa que se rige bajo la Ley 1.777 sancionada en 2005. Se trata de unidades descentralizadas de gestión política y administrativa.

A partir del dato de la comuna donde ocurrió el siniestro vial observamos que la comuna 1, la cual se compone de los barrios de Retiro, San Nicolás, Puerto Madero, San Telmo, Montserrat y Constitución, registra la mayor cantidad de siniestros viales. Siendo la comuna 3 (Balvanera y San Cristóbal) y la 4 (La Boca, Barracas, Parque Patricios y Nueva Pompeya) las que siguen en orden descendente. En el Apéndice E se exponen gráficos con distribución de siniestros viales por comuna. Debido a que algunas comunas agrupan más de un barrio y esto puede sesgar el análisis en función de que comuna presenta mayores siniestros viales, es que analizamos los mismos por barrio.

Utilizando como variable el barrio donde ocurrió el siniestro vial podemos destacar que los barrios de Palermo, Flores y Balvanera son donde ocurren a lo largo de los años analizados la mayor cantidad de siniestros viales. En función de lo relevado consideramos útil incorporar en nuestro armado del modelo el dato del barrio donde ocurrió el siniestro vial.

Figura 9:Mapa de calor con distribución de siniestros por BARRIO de C.A.B.A. por año



Mes

Con el dato del mes se puede observar que enero, febrero, julio y diciembre son meses con menor cantidad de siniestros viales. Es posible que exista una relación entre los meses y las vacaciones dado que estos son meses típicos donde muchas personas salen de la ciudad y por lo tanto hay menor circulación de vehículos.

Figura 10:Distribución de siniestros viales totales por mes 2015-2017

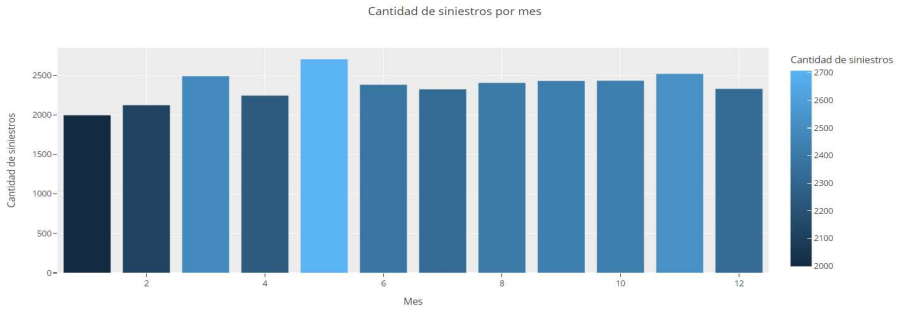
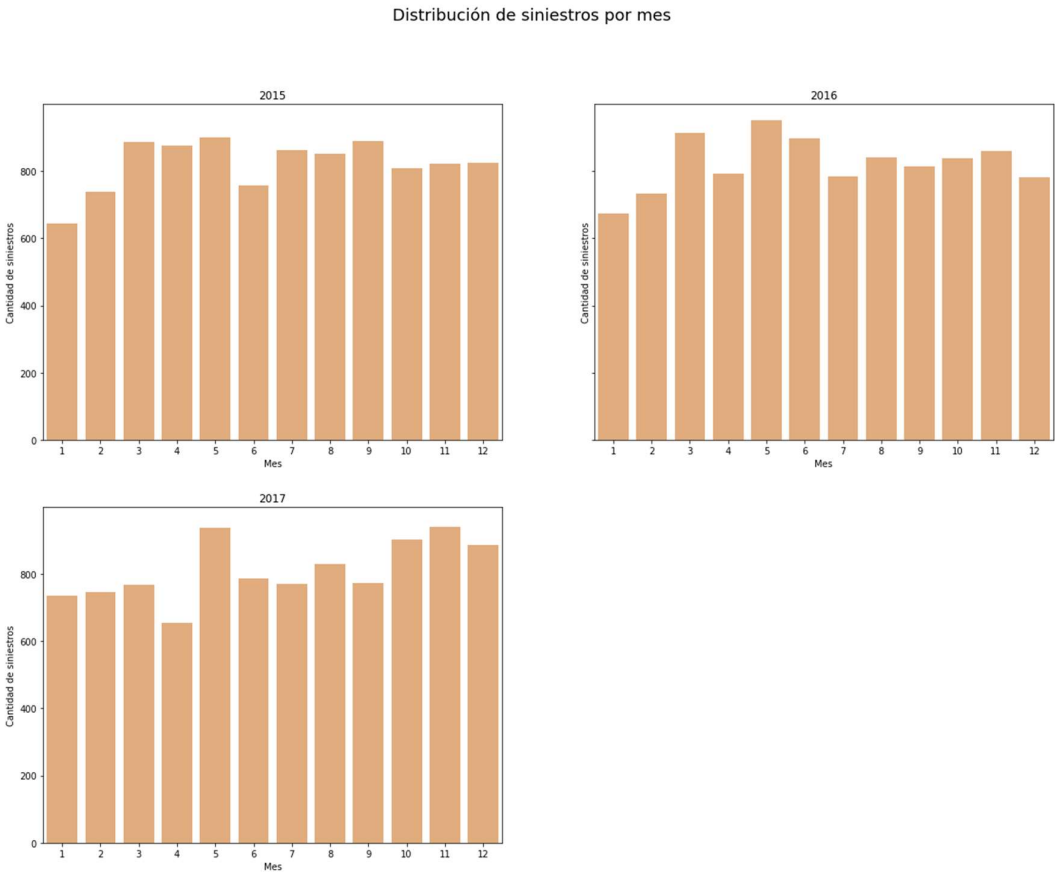


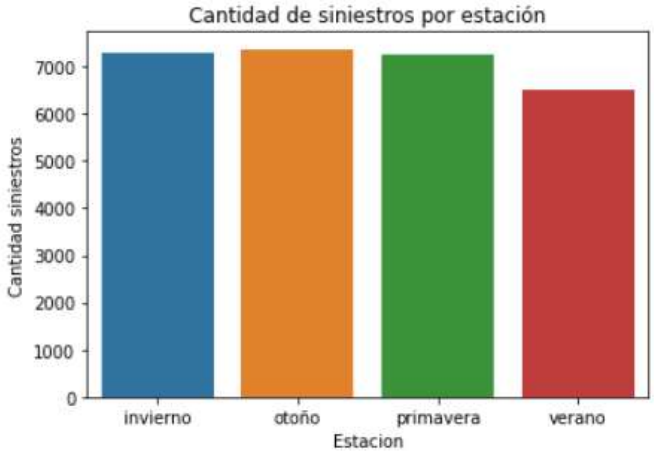
Figura 11:Distribución de siniestros por MES y por año



Estación del año

En línea con la cantidad de siniestros viales por mes y teniendo en cuenta la variable que hemos agregado de la estación del año en que ocurren, analizamos los mismos según la época del año en la que ocurren, donde podemos observar que existe concordancia con lo analizado por mes. La cantidad de siniestros viales se mantiene constante en la época de otoño, primavera e invierno, siendo el verano (diciembre, enero, febrero, marzo) la estación del año con menor cantidad de , colisiones causadas por el tránsito. Dado que esta variable tiene relación directa con el dato del mes, deberíamos considerar utilizar una de estas dos.

Figura 12:Distribución de siniestros viales totales por estación del año 2015-2017

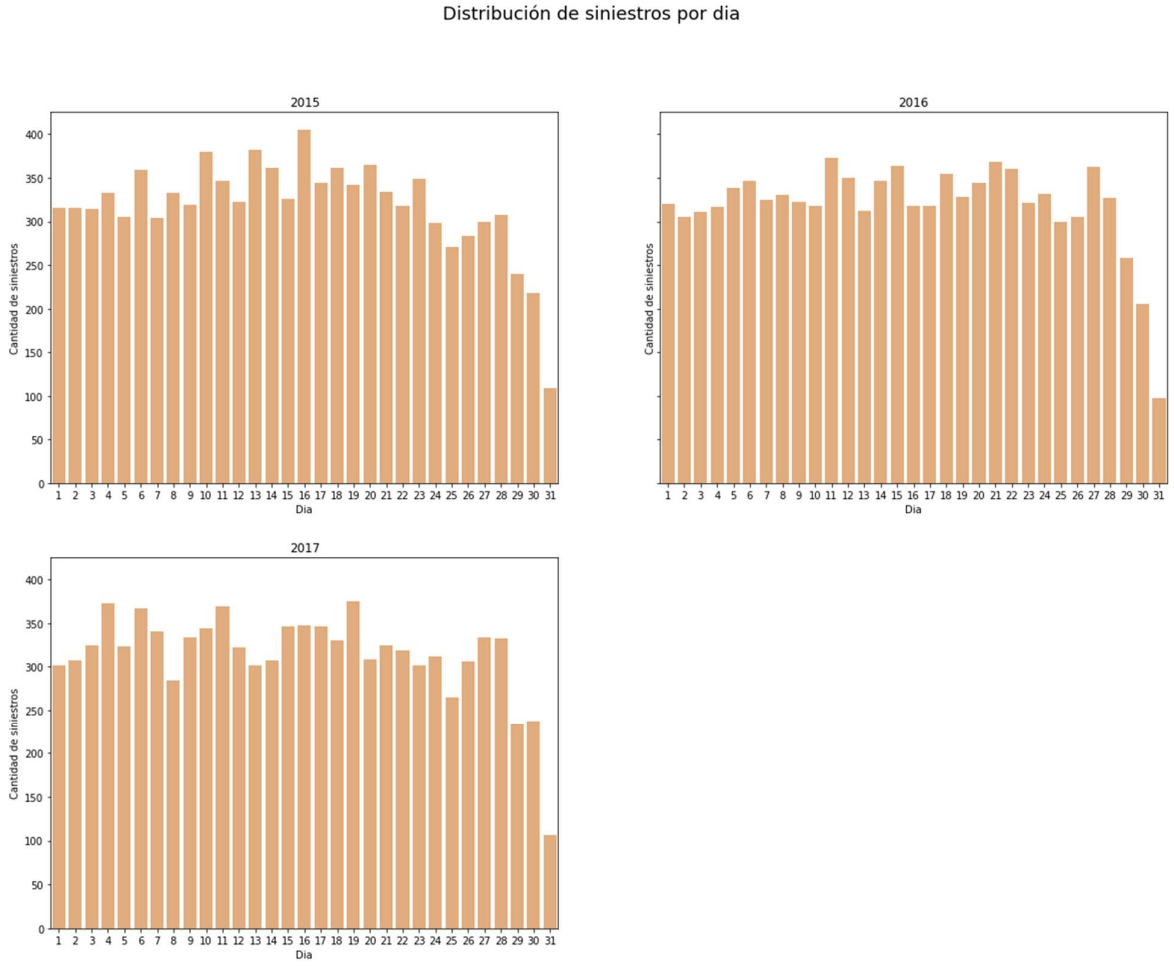


Día

Esta variable nos indica el número del día en el que ocurrió el siniestro vial. Se observa que para los últimos días del mes la frecuencia de siniestros baja, esto se da porque febrero cuenta con 28 días, salvo año bisiesto[18] como sucede en 2016 y porque solo existen siete meses con 31 días, por lo que la cantidad de siniestros para los días 29,30 y 31 disminuye por la inexistencia de estos días en el total de datos analizados. Si bien se observa cierta tendencia constante de los siniestros viales por día a lo largo de los años analizados, cabe destacar que existe cierta fluctuación diaria de los mismos.

[18] Los llamados años bisiestos son aquellos que suceden cada cuatro años y se caracterizan por tener un día más: el 29 de febrero.

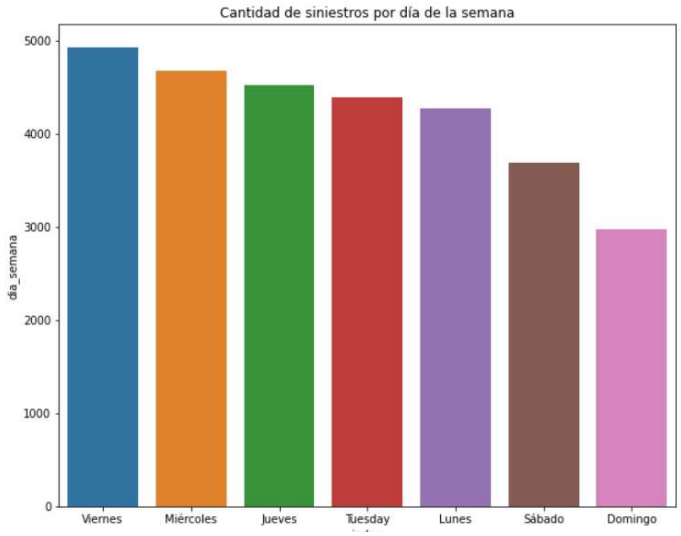
Figura 13:Distribución de siniestros por DÍA y por año



Día de la semana

Con la variable agregada “día semana” la cual indica en qué día de la semana ocurrió el siniestro vial, se puede observar que los días con menor cantidad de siniestros viales son sábado y domingo. Entendemos que existe una relación entre los días de la semana y los días no hábiles como son sábado y domingo, donde el tráfico en la ciudad es menor y por lo tanto hay menor circulación de vehículos. Es por esto que consideramos que la variable día de la semana es una variable que influye en la cantidad de siniestros viales para incorporar en el armado de nuestro modelo.

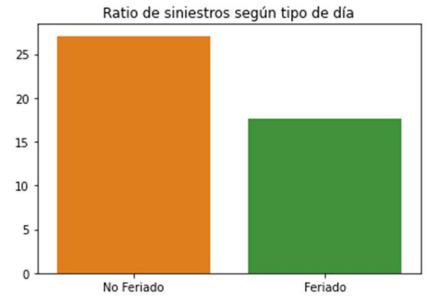
Figura 14:Distribución de siniestros por DÍA DE LA SEMANA 2015-2017



Feriado

Considerando la variable "feriado" agregada en nuestros datos, la cual indica, si el siniestro vial ocurrió en un día feriado o no, analizamos los ratios de siniestros viales en día no feriado y día feriado. Es decir, la cantidad de siniestros ocurridos en función de la cantidad de días feriados o no. Considerando el periodo analizado 2015-2017, por día ocurrieron en promedio 27 siniestros en días no feriados, mientras que 19 en día feriado.

Figura 15:Ratio de siniestros viales según tipo de día feriado o no 2015-2017

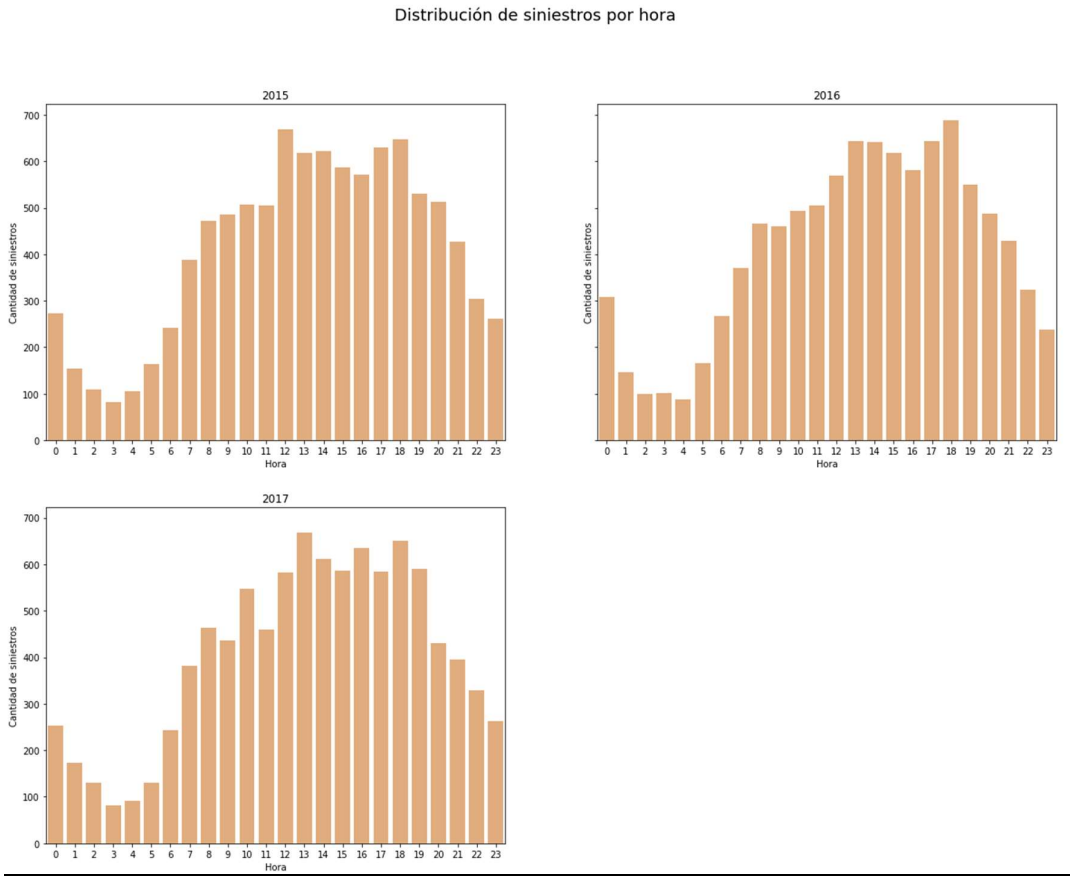


A partir de lo analizado con la variable día de semana y feriado, consideramos útil incorporar a nuestro modelo una variable que unifique los días hábiles con los que no lo son.

Hora

A partir de la variable generada que contiene sólo la hora, sin minutos y segundos del momento en que ocurrió el siniestro vial, generamos el siguiente gráfico para observar la distribución de siniestros por hora. Se puede observar que hay una clara tendencia a lo largo de los años: la cantidad de siniestros es menor durante la madrugada, teniendo picos por las tardes, de 16:00 hs a 20:00 hs, por lo que consideramos relevante dicha variable para incorporar en el modelo.

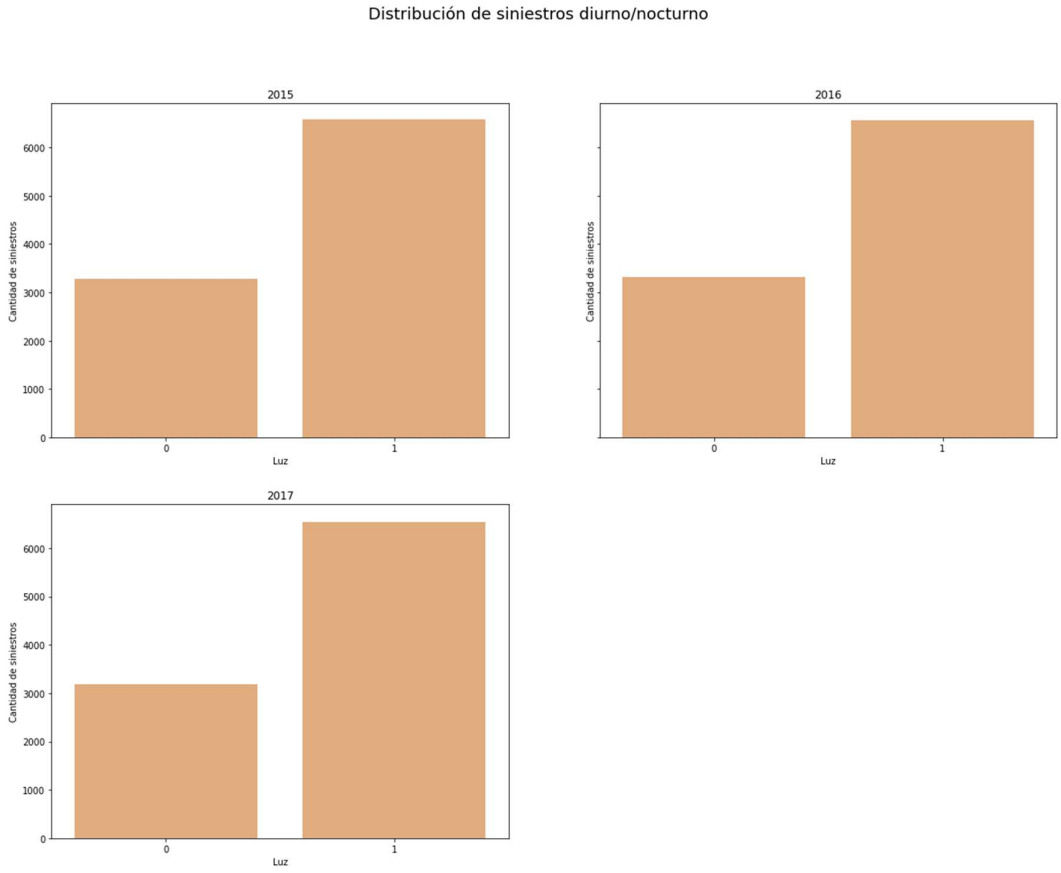
Figura 16:Distribución de siniestros por HORA y por año



Diurno/Nocturno

A partir de nuestra variable “luz”, generada en función del horario de salida y puesta del sol, analizamos los siniestros viales según si se dan en momentos de luz o no. En línea con lo analizado en el punto anterior, donde en horas de la tarde se dan la mayor cantidad de siniestros viales, se puede observar que la mayoría de ellos (67%) ocurren durante el día con luz (= 1), mientras que el resto durante el horario nocturno donde hay menor visibilidad (luz = 0).

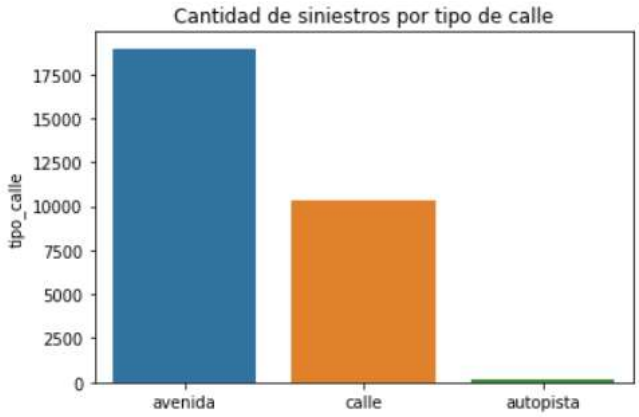
Figura 17:Distribución de siniestros viales en diurno/nocturno según luz



Tipo de calle

Considerando el dato de si el siniestro vial ocurrió en una avenida, calle o autopista, identificamos que las avenidas contemplan el principal “tipo de calle” donde suceden las colisiones causadas por el tránsito. Esto podría relacionarse con el alto tránsito que involucran las mismas.

Figura 18:Distribución de siniestros viales por tipo de calle 2015-2017



Esquina

Con los datos de la localización del hecho generamos una variable categórica correspondiente a si la dirección normalizada se trata de una dirección con número y calle, o una esquina como intersección de dos calles separadas con un "&". A partir de esto, para el periodo analizado 2015-2017 en la Ciudad autónoma de Buenos Aires el 77% de los siniestros viales ocurrieron en una esquina, y para que eso suceda suponemos que alguno de los involucrados no respetó los semáforos o la prioridad de paso, independientemente de su forma de desplazamiento o rol en la vía (ciclista, motociclista, peatón o automovilista).

Tabla 3:Cantidad de siniestros viales según lugar del hecho (dirección/esquina)

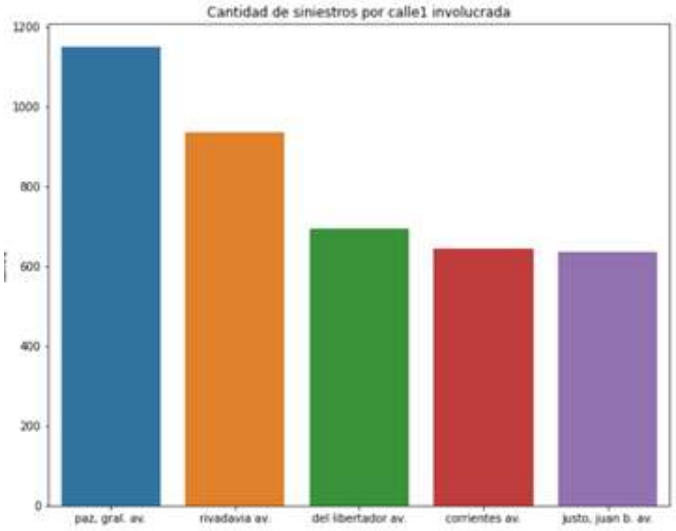
Cantidad de siniestros viales ocurridos en una dirección específica o esquina
En porcentaje

	Año		
	2015	**2016**	**2017**
Dirección	24%	23%	23%
Esquina	76%	77%	77%
Total	**100%**	**100%**	**100%**

Si bien el lugar específico del hecho no será una variable a incorporar en nuestro modelo por la granularidad y especificidad de ésta, consideramos que será útil analizar ciertas características para luego abordar conclusiones pertinentes a este dato.

A partir del dato disponible “calle1” que identifica la primera calle del lugar del hecho, podemos identificar el ranking de las avenidas más conflictivas, siendo Av. Gral Paz, Av. Rivadavia, y Av. Libertador las que lideran el ranking de siniestros viales, como así también de cantidad de víctimas fallecidas. Esto guardaría relación con que son avenidas que tienen alto tránsito y puede estar también vinculado con la velocidad permitida en las mismas, que es mayor que en calles.

Figura 19:Ranking 5 principales calles1 de siniestros viales según lugar del hecho



Considerando que las esquinas conflictivas serían las intersecciones que no sólo tienen alto tránsito, sino también una mayor acumulación de siniestros viales, analizamos, por ejemplo, que a lo largo de los tres años la intersección de “paz, gral. av. & balbin, ricardo, dr. av.” lidera el ranking de siniestros viales, y considerando también “balbin, ricardo, dr. av. & paz, gral. av.”, ocurrieron más de 30 siniestros viales por año a razón de un siniestro cada 12 días.

Sigue en el ranking, la esquina de “paseo colon av. & belgrano av.”/“belgrano av. & paseo colon av.” , dónde ocurrieron en promedio 17 siniestros viales por año, a razón de más de un siniestro por mes.

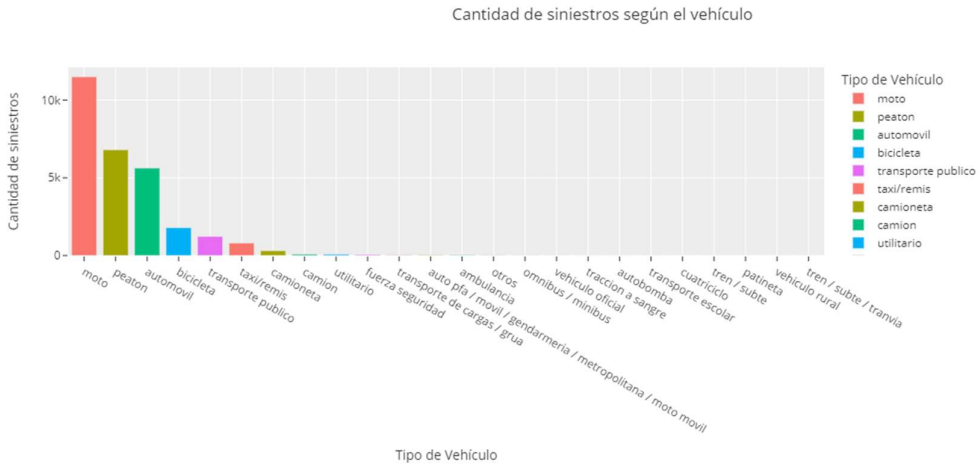
Asimismo, en la intersección, “paz, gral. av. & de los corrales av.”/“de los corrales av. & paz, gral. av.” se registraron en promedio 9 siniestros por año, dejando al menos una muerte anual, lo cual indica que esta intersección sería un punto a considerar para futuras campañas o planes de acción.

En función de lo expuesto anteriormente y en concordancia con el ranking de avenidas con mayor cantidad de siniestros viales, la “av. general paz” con sus distintas intersecciones es predominante de siniestros viales, y esto puede darse por embotellamientos, maniobras o velocidad que se pueden dar en estos tipos de lugares.

Tipo de vehículo

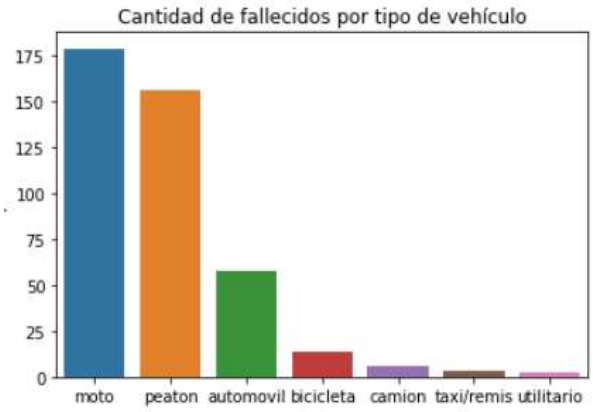
A lo largo de los 3 períodos analizados, la moto, el peatón y el automóvil representan el 80% de los tipos de vehículos involucrados en los siniestros viales. La moto representa casi el 40% de los siniestros viales en los 3 años, mientras que peatón y automóvil casi el 20% respectivamente. Si bien la granularidad de esta variable podría disminuirse en cantidad de opciones unificando ciertos tipos de vehículos o formas de desplazamiento, es una variable que en principio no consideraremos en nuestro armado del modelo.

Figura 20:Distribución del tipo de vehículo o forma de desplazamiento de las víctimas de siniestros viales 2015-2017



Asimismo, la tendencia del tipo de vehículo se mantiene en las víctimas fallecidas a causa de siniestros viales. Moto, peatón y automóvil lideran el ranking de los principales tipos de vehículos o formas de desplazamiento de los fallecidos.

Figura 21:Distribución del tipo de vehículo o forma de desplazamiento de los fallecidos por siniestros viales 2015-2017 (cantidad de fallecidos >1)

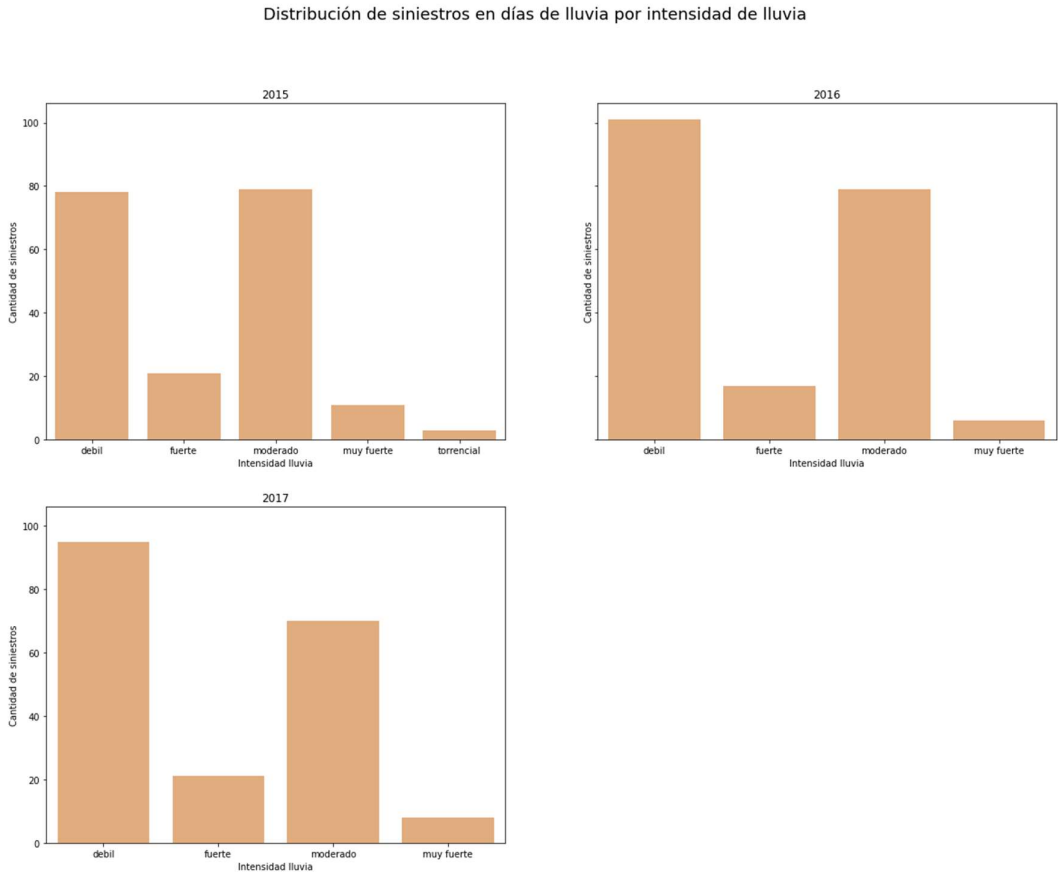


Lluvia

A partir de nuestra variable de si el suceso ocurrió en un día de lluvia o no, analizamos los ratios de cantidad de siniestros en días de lluvia y días de no lluvia. Donde obtuvimos que ocurrieron 2 siniestros por día con lluvia (cantidad de siniestros viales en días de lluvia sobre cantidad de días de lluvia), mientras que para los días sin lluvia sucedieron 35 siniestros viales. Por lo que la gran cantidad de sucesos ocurren en días sin lluvia.

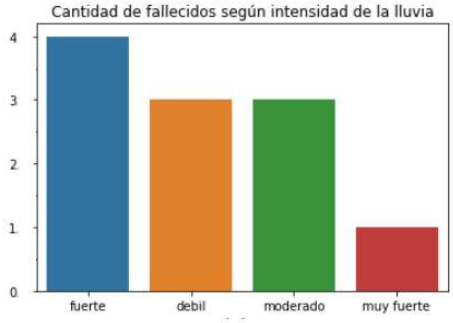
Analizando los siniestros viales que se dan en días de lluvia se observa que la gran mayoría de ellos se dan en días de lluvias débiles o moderadas, y que para 2016 y 2017 no se registraron siniestros viales en días de lluvias torrenciales.

Figura 22:Distribución de siniestros en día de lluvia según intensidad de la lluvia



En función de los ratios mencionados anteriormente, donde ocurren 2 siniestros por día de lluvia, se destaca que las víctimas fatales que resultaron de los siniestros viales en días de lluvia, el 36% ocurrió en días de lluvias fuertes. Por lo que las lluvias fuertes podrían estar asociadas a un factor condicionante de siniestros viales que resultan en víctimas fatales.

Figura 23:Distribución de fallecidos en día de lluvia según intensidad de la lluvia (2015-2017)



4.2 Análisis de las víctimas de siniestros viales

Las víctimas de los siniestros viales analizados se diferencian en homicidios y lesionados. Se observa una tendencia en la cantidad de homicidios y lesiones a lo largo de los años analizados, en promedio el 99% de los sucesos corresponden a lesionados. Aunque la variación entre períodos no es muy amplia, en las siguientes gráficas se puede observar que la tendencia empieza a decrecer levemente para el año 2017. Esta variable no la incluiremos en nuestro modelo dado que la mayoría de los siniestros ocurridos en este periodo tuvieron lesionados, por lo que al tener este desbalanceo de clases puede influir en los resultados de los modelos.

Figura 24:Distribución de causa de siniestros viales (homicidio/lesionados) 2015-2017

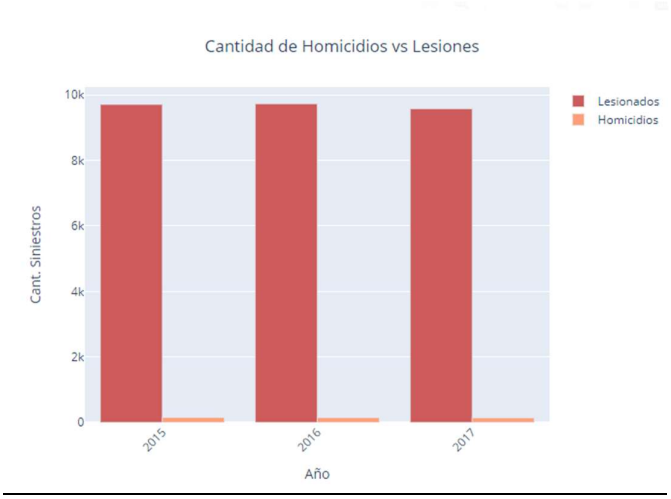
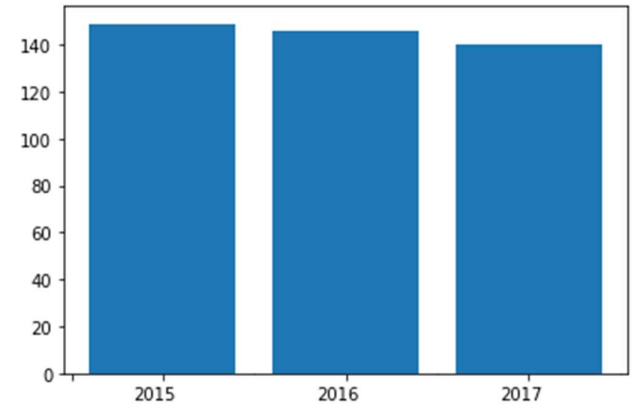


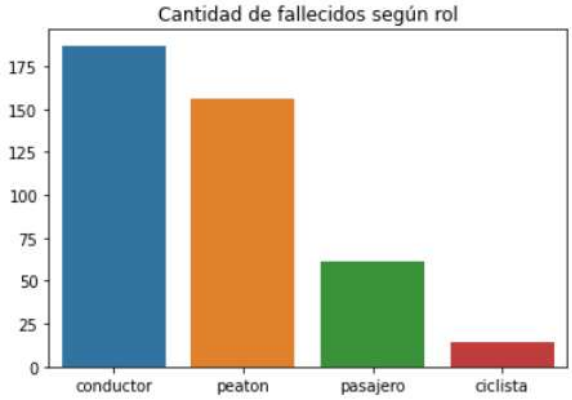
Figura 25:Distribución de víctimas fatales por año



Rol

Otra variable que analizamos fue el rol de las víctimas fallecidas en el siniestro vial, donde la gran mayoría corresponde a conductor, peatón, pasajero y ciclista en concordancia con la mayoría de los tipos de vehículos involucrados, moto, peatón, automóvil y bicicleta. Los conductores, peatones y pasajeros representan casi el 90% del total de los fallecimientos. Al igual que tipo de vehículo esta no es una variable que incorporaremos en nuestro modelo.

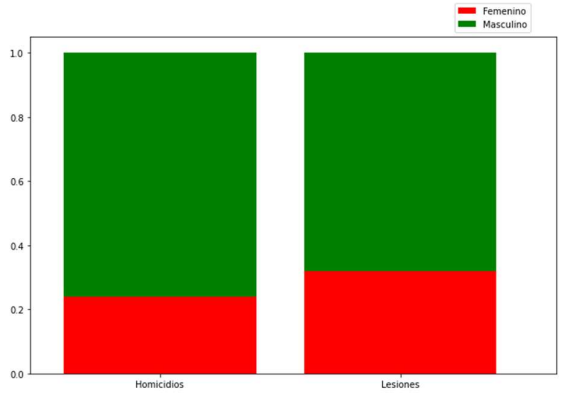
Figura 26:Distribución de fallecidos según rol en la vía pública



Género

Si bien el género no es una variable que consideraremos para nuestro modelo, se puede apreciar que los varones son el grupo más numeroso de víctimas de siniestros viales. Tres cuartas partes de víctimas fatales fueron de sexo masculino (74%).

Figura 27:Distribución de víctimas por género 2015-2017



Esto se observa en contraposición a la relación entre ambos sexos en la Ciudad de Buenos Aires (Estadística CABA, 2017), la cual es medida a través del índice de masculinidad que estima que hay 88 varones cada 100 mujeres y da cuenta del predominio femenino en el total de población. Siendo el índice de masculinidad para los homicidios de 324 varones fallecidos cada 100 mujeres fallecidas; y para los lesionados, cada 200 varones lesionados hay 100 mujeres lesionadas.

La mayor proporción masculina de víctimas se replica en igual proporción para todos los tipos de vehículos, salvo el "transporte público" y "peatón" donde el sexo femenino representa mayor proporción de víctimas.

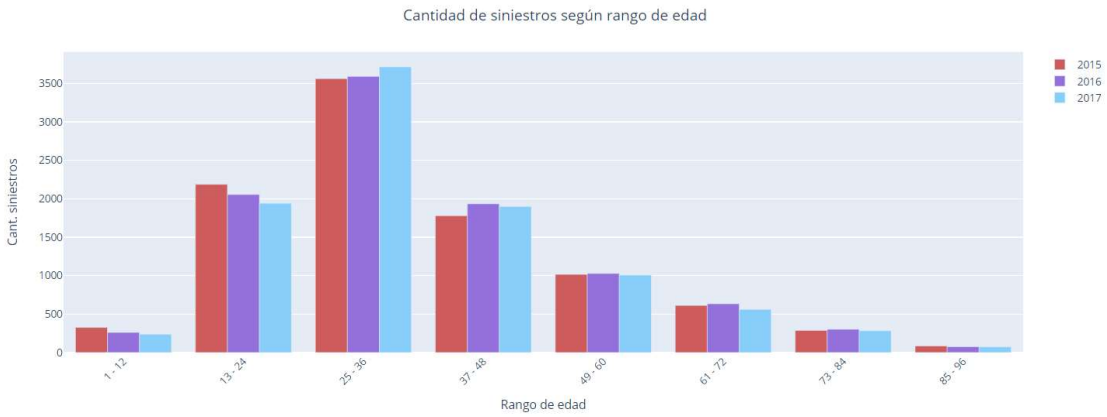
Figura 28:Distribución de género de las víctimas de siniestros viales por tipo de vehículo 2015-2017 (víctimas >50)

Rango etario

Nos pareció relevante analizar la edad de las víctimas de siniestros viales, para lo que generamos un gráfico por rango etario, donde se visualiza que la edad de la mayoría de las víctimas se encuentra de entre 25 y 36 años, como así también predominan víctimas de edad entre 13 -24, y 37-48 años. El promedio de edad de lesionados es de 36 años,

mientras que para fallecidos es de 40 años. Entendiéndose a partir de este análisis, se debe hacer hincapié en personas de este rango de edad (13 a 48 años).

Figura 29:Distribución de siniestros viales según rango de edad



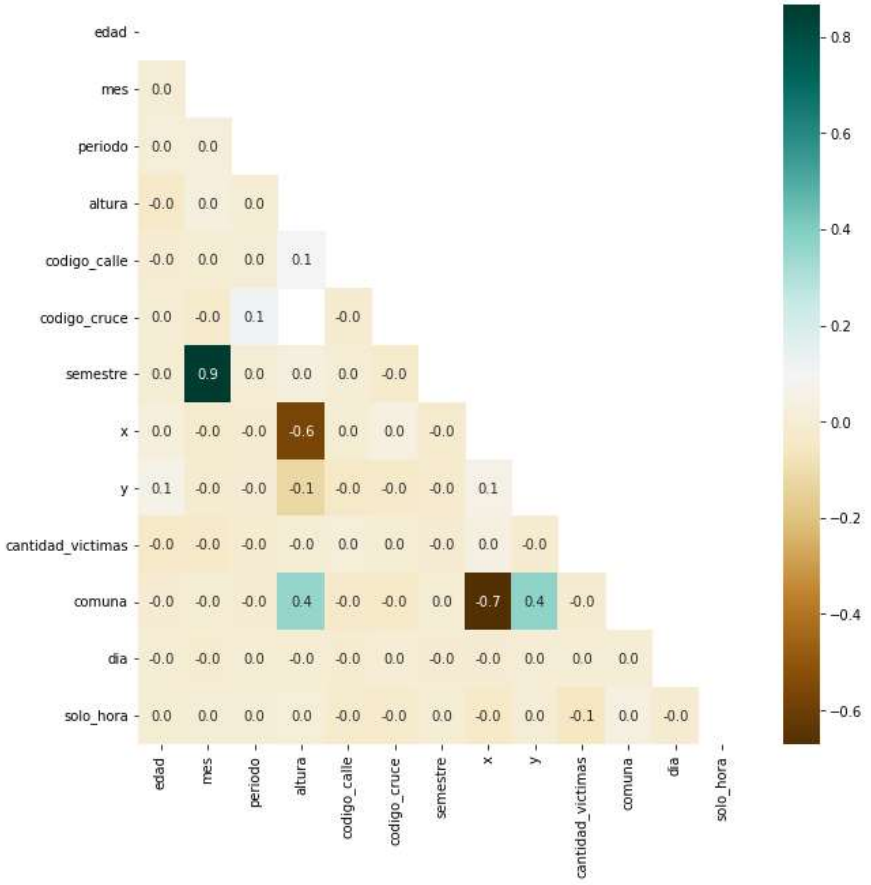
Análisis de correlación entre variables

Algunos modelos se ven perjudicados si incorporan predictores altamente correlacionados. Por esta razón, es conveniente estudiar el grado de correlación entre las variables disponibles. Para ello analizamos la matriz de correlación, la cual mide el grado de relación entre cada par de variables y evalúa patrones en los datos.

La matriz de correlación, basada en la correlación de Pearson, está especialmente diseñada para las mediciones continuas para captar las correlaciones entre variables de diferentes tipos de datos, por lo que analizamos las variables numéricas, sin considerar las binarias que contienen valor 1 ó 0.

Con nuestras variables numéricas, obtuvimos la siguiente matriz, la cual denota una correlación en variables como semestre y mes; latitud y longitud con altura; y está última con comuna también lo cual suena lógico por tratarse de variables que están relacionadas en cuestión de período de tiempo y ubicación.

Figura 30:Matriz de correlación de variables numéricas (incluidas binarias)



Atributos y armado de dataset

A partir de nuestro análisis descriptivo decidimos, por un lado, abordar conclusiones y recomendaciones de planes de acción con variables de los siniestros viales como: el tipo de vehículo involucrado, la forma de desplazamiento, el tipo de calle donde ocurrieron, esquinas con mayor frecuencia de siniestros viales; el rol, la edad y género de las víctimas, ya que a partir de estos se puede determinar sobre qué estrategias y grupos de personas se debe hacer hincapié para la prevención de siniestros viales.

Por otro lado, a partir de lo que hemos ido investigando y analizando de los datos disponibles, consideramos que sería útil generar modelos que permitan descubrir, comprender y analizar la frecuencia de los siniestros en el tiempo, para así poder determinar momentos y/o lugares de mayor probabilidad de siniestros.

En función de las variables que caracterizan al momento del suceso elegimos como atributos para nuestro modelo, variables como día, mes, hora (sin minutos) y barrio

del suceso; si se corresponde con un día hábil o no (unificando como no hábil los días de fin de semanas y feriados); y la intensidad de la lluvia en función de la precipitación registrada las últimas 6 horas antes del hecho.

Antes de comenzar con nuestro modelado tuvimos que armar nuestro dataset. Dado que queríamos trabajar a partir del dato de la fecha y hora del suceso del siniestro vial, y existían un 2% de los registros con estos datos incompletos, decidimos eliminar esos registros dado que los datos faltantes podrían alterar nuestros modelos.

Partiendo de nuestro dataset con datos de 2015 - mayo 2018 (33.234 registros) eliminamos las observaciones donde tuvieran nulo en al menos una de las variables "fecha" u "hora", ya que para estos registros también quedaron sin dato los atributos creados como, intensidad de la lluvia, si se trataba de un día hábil o no, entre otras, ya que ante la falta del dato de la fecha o la hora no pudimos asignarles valores a esas variables creadas. Es así como nos quedamos con un total de 33.162 registros para avanzar con nuestro armado de los modelos.

5. Resultados

En esta sección se expondrán los resultados de los modelos aplicados. Se darán detalles acerca de los hiperparámetros elegidos y su performance.

Decidimos configurar un primer modelo donde se agrupe la cantidad de siniestros viales (registros) por día, mes, hora, barrio, intensidad lluvia, día hábil; siendo el número de siniestros nuestra variable a predecir.

Modelo 1:

Agrupando por día, mes, solo hora, barrio, intensidad de la lluvia, día hábil, y cantidad de siniestros obtuvimos la siguiente estructura de datos:

Tabla 4:Data frame de modelo agrupado por día, mes y hora

	dia	mes	solo_hora	Barrio	intensidad_lluvia	habil	NoAcc
0	1	1	0	BOCA	sin lluvia	1	1
1	1	1	0	VILLA LUGANO	sin lluvia	1	1
2	1	1	1	MATADEROS	sin lluvia	1	2
3	1	1	1	PARQUE PATRICIOS	sin lluvia	1	1
4	1	1	1	RECOLETA	sin lluvia	1	1

Modelo 1.1: Variables categóricas binarizadas:

Determinamos las variables categóricas y numéricas para binarizar solo las variables categóricas y así poder tener nuestro dataset para entrenar el modelo. En una primera iteración, entrenamos nuestro modelo dejando como validación un 20% de nuestros datos de forma aleatoria y con los hiperparámetros predefinidos obtuvimos un error de test (RMSE) de **0,38.**

Hiperparametros definidos

- n_estimators = 100
- criterion = mse
- max_depth = 20
- max_features = auto (utiliza todos los predictores)

Modelo 1.2: Variables categóricas en número:

A su vez, probamos un modelo trabajando no sobre variables categóricas binarizadas, sino sobre variables categóricas transformadas en número. Con un split de 80%-20%, y los hiperparametros definidos anteriormente, obtuvimos un RMSE de **0,37.**

Si bien definimos valores por defecto para sus hiperparámetros, no se puede saber de antemano si estos son los más adecuados, la forma más común de identificar los valores óptimos es probando diferentes posibilidades. Por lo que para afinar hiperparámetros utilizamos *RandomizedSearchCV*[19], una búsqueda aleatoria en hiperparámetros mediante validación cruzada sobre la configuración de los parámetros.

Tabla 5:Hiperparámetros a evaluar

Hiperparámetro	Valor
n_estimators	[100, 200, 300, 400, 500, 600]
max_depth	[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60]
max_features	auto, sqrt
min_samples_split	2,5,10
min_sample_leaf	1,2,4
bootstrap(*)	(True, False)

(*)Método de selección de muestras para entrenar cada árbol.

Definiendo una búsqueda aleatoria de parámetros, utilizando validación cruzada de 3 pliegues con búsqueda en 10 combinaciones diferentes, se obtienen como resultado las siguientes cinco mejores combinaciones de hiperparámetros.

Tabla 6:Resultados de las cinco mejores combinaciones modelo 1 con variables categóricas binarizadas (1.1)

param_n_estimators	param_min_samples_split	param_min_samples_leaf	param_max_features	param_max_depth	param_bootstrap	mean_test_score	std_test_score
100	5	2	sqrt	10	False	0.003049	0.002566
300	10	4	sqrt	40	True	-0.000313	0.003242
500	10	4	sqrt	60	True	-0.000576	0.004037
300	10	4	auto	55	True	-0.035416	0.007686
300	10	4	auto	60	True	-0.035916	0.007912

Tabla 7:Resultados de las cinco mejores combinaciones modelo 1 con variables categóricas transformadas en números (1.2)

[19] Scikit-Learn (Python) ofrece esta funcionalidad en la clase sklearn.model_selection.RandomizedSearchCV que pasa aleatoriamente el conjunto de hiperparámetros y calcula la puntuación y proporciona el mejor conjunto de hiperparámetros que da la mejor puntuación como resultado.
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

param_n_estimators	param_min_samples_split	param_min_samples_leaf	param_max_features	param_max_depth	param_bootstrap	mean_test_score	std_test_score
100	5	2	sqrt	10	False	-0.007483	0.004147
500	10	4	sqrt	60	True	-0.010599	0.005462
300	10	4	sqrt	40	True	-0.012565	0.005460
300	10	4	auto	60	True	-0.027260	0.008963
300	10	4	auto	55	True	-0.027495	0.008808

Para ambas opciones modeladas obtuvimos los mismos mejores hiperparámetros y con estos un mejor error de test de **0,33** para ambos, ya que cómo definimos en nuestro apartado "**Métrica de evaluación de modelos"**, cuánto más pequeño es un valor RMSE, más cercanos son los valores predichos y observados.

A continuación, se expone un resumen de los modelos e hiperparámetros evaluados hasta el momento, donde el algoritmo Random Forest entrenado con la mejor combinación de hiperparámetros encontrados durante el proceso de ajuste informa un mejor rendimiento (RMSE).

Tabla 8:Resumen de modelo 1

Modelo 1	Variables	Train/Test Split	n_estimators	max_depth	max_features	min_samples_leaf	min_samples_split	RMSE
1.1	Categóricas binarizadas	80% - 20%	100	20	auto	1 (default)	2 (default)	**0,38**
1.2	Categóricas en número							**0,37**
1.1.1	Categóricas binarizadas		100	10	sqrt	2	5	**0,33**
1.2.1	Categóricas en número							**0,33**

Asimismo, realizando un ciclo (loop) para ajustar el modelo con cada combinación de hiperparámetros empleando k-cross-validation y teniendo control sobre los resultados secuenciales de las iteraciones; identificamos que la mejor combinación de hiperparámetros continuaba siendo la obtenida con RandomSearchCV.

Figura 31:Gráfico con la evolución de los errores con las distintas combinaciones de hiperparámetros

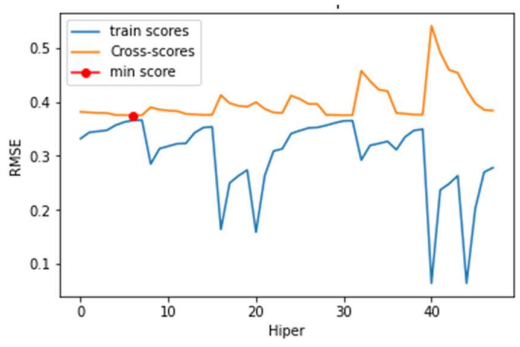


Tabla 9:Resultados de las mejores hiperparámetros

	train_scores	cv_scores	bootstrap	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators
7	0.366200	0.375123	True	10.0	sqrt	2	5	100
6	0.365631	0.375074	True	10.0	sqrt	2	2	100
31	0.365133	0.375284	False	10.0	sqrt	2	5	100

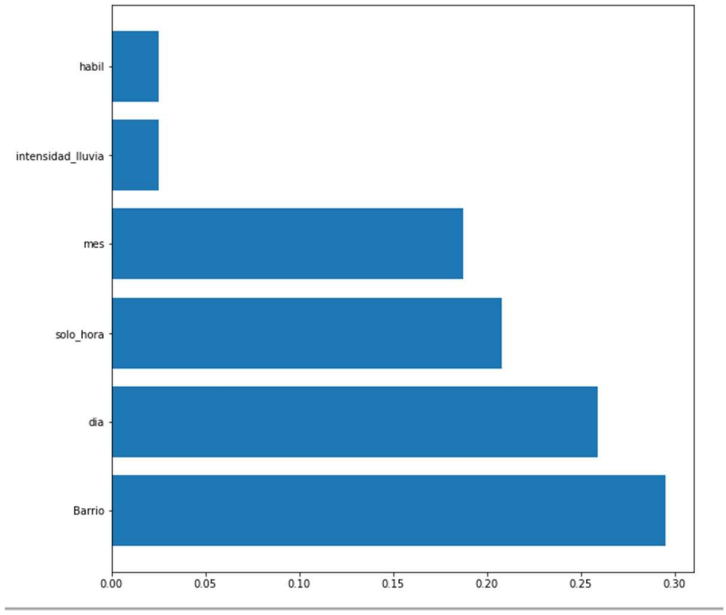
En cuanto a las características más importantes de los modelos evaluados, para un modelo con variables categóricas binarizadas obtuvimos como atributos más importantes el día, la hora y el mes. En Apéndice F se expone un gráfico con el feature importance de todas las variables donde también se destacan como características categóricas el barrio Nueva Pompeya, intensidad lluvia moderado, barrio Puerto Madero, barrio San Telmo, entre otras.

Figura 32:Feature importance modelo 1 variables categóricas binarizadas

predictor	importancia
dia	0.310792
solo_hora	0.285693
mes	0.233137

En el modelo con las variables categóricas transformadas en número, se observa como características más importantes el barrio, día y la hora.

Figura 33:Gráfico feature importance modelo 1 con variables categóricas transformadas en número



Modelo 2

En una segunda iteración, decidimos, entrenar nuestro primer modelo planteado, pero ahora considerando que nuestro modelo aprenda de datos pasados y sus predicciones se apliquen a datos futuros. Por lo tanto, nuestros criterios de división de entrenamiento y validación deben ser temporales en lugar de aleatorios.

De esa forma, nos aseguramos de entrenar el modelo con datos anteriores a los que validamos el modelo. Al tener disponible tres años completos y parte del 2018, optamos por utilizar 2015 y 2016 para entrenamiento, mientras que 2017 hasta mayo 2018 para validación.

Con esta división de entrenamiento-validación temporal, utilizando las variables categóricas transformadas en números, y configurando como hiperparámetros aquellos con los que obtuvimos mejores resultados en las iteraciones anteriores (ver Tabla 8 punto 1.1.1 y 1.1.2) obtuvimos un error de test (RMSE) de: **0,36.**

Con los resultados obtenidos en esta etapa, decidimos recurrir a datos más actuales para validar nuestro modelo. Como expusimos en la sección 2, los datos de siniestros viales posteriores a mayo 2018 se encuentran disponibles públicamente pero son registrados por otro organismo y con otra estructura distinta a la que analizamos de los períodos 2015 a mayo 2018. Esta estructura cuenta con el dato de la fecha y hora, variables que sirven para nuestro modelo. Es por eso que realizamos una validación utilizando datos de 2018 a 2019 (ver en Apéndice G un descriptivo y armado del dataset), el período 2020 no lo incluímos dado que fue un periodo atípico en términos de movilidad debido a las restricciones para circular en el marco de la pandemia de coronavirus[20].

Utilizando estos datos para validación con nuestro modelo planteado y considerando los hiperparámetros que mejor se ajustaron al modelo 1, obtuvimos un error de test de **0,29.**

Tabla 10:Hiperparámetros y RMSE validando con datos 2018-2019

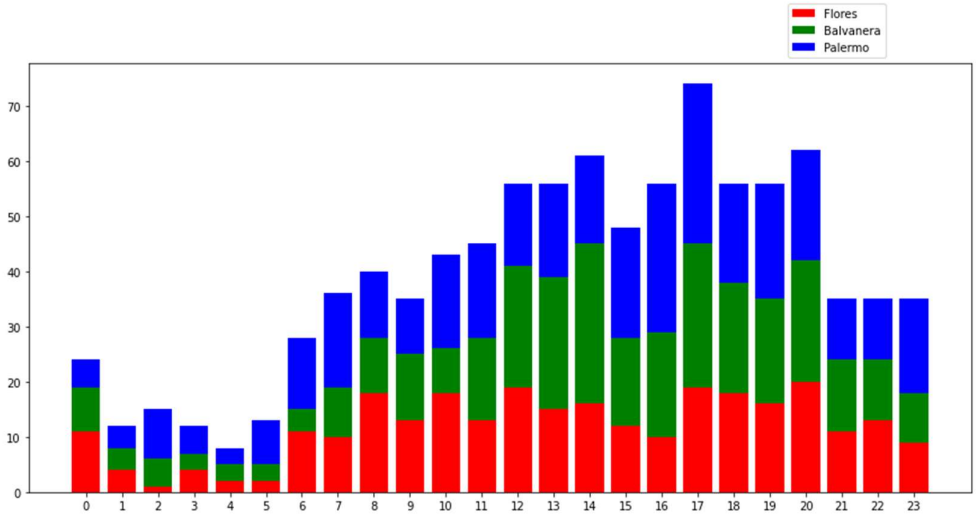
[20] Con fecha 11 de marzo de 2020 la OMS, declaró el brote del virus SARS-CoV-2 como una pandemia lo que requirió un Aislamiento Social, Preventivo y Obligatorio (ASPO).

Hiperparámetro	Valor	RMSE
n_estimators	100	**0,29**
max_depth	10	
max_features	sqrt	
min_samples_split	5	
min_sample_leaf	2	

Con este modelo generamos una salida con las predicciones, que consideramos que podría ser de utilidad para, por ejemplo, generar una interfaz de usuario donde se puede observar por día, mes, hora, barrio y según el estado del tiempo (sin lluvia, o con probabilidad de lluvias débiles o fuertes), la cantidad de siniestros probables. Por ejemplo, si consideramos el día 11 de mayo en el horario de las 20hs, sin lluvia, en el barrio de Mataderos predice 1 siniestro.

Con los resultados obtenidos para el período junio 2018 - diciembre 2019 podemos observar que se predice una mayor cantidad de siniestros viales en barrios como Palermo, Balvanera, Flores, Caballito y Viila Crespo. Y si analizamos cuáles son las horas con mayor cantidad de siniestros predichos se encuentra que la franja horaria de 12 a 20hs lidera el ranking.

Figura 34:Gráfico de distribución de cantidad de siniestros predichos para junio 2018 - diciembre 2019 por hora y barrio



6. Conclusiones y recomendaciones

6.1 Aplicaciones

El presente trabajo se ha centrado en lograr un análisis de los siniestros viales que suceden en la Ciudad Autónoma de Buenos Aires a partir de utilizar una serie de técnicas analíticas que pueden ser de utilidad al momento de generar políticas y planes de acción para la prevención de siniestros viales.

Como consecuencia del análisis descriptivo concluimos:

- ➔ Las esquinas porteñas se constituyen como los puntos de mayor siniestralidad de la ciudad, más aún si una de esas intersecciones contempla una avenida de alto tránsito como ser Av. General Paz, Av. Rivadavia, Av. del Libertador, entre otras. Por lo que estas deberían ser zonas de control constantes, se podría considerar instalar cámaras de fiscalización en las intersecciones de las avenidas/calles más recurrentes de siniestros viales mencionadas en el análisis descriptivo. Existen evidencias[21] que indican reducciones en incidentes viales en puntos con cámaras, ya que esto corrige y mejora las conductas de quienes transitan la vía pública, por lo que sumar herramientas de fiscalización donde se detecta mayor concentración de siniestros viales reduciría la probabilidad de ocurrencia de siniestros viales.
- ➔ Dado que se supone que los siniestros viales se dan en mayor proporción en las esquinas porque alguno de los involucrados no respetó los semáforos o la prioridad de paso; y del análisis descriptivo se destaca que el rol de conductor y peatón representan las mayores víctimas fallecidas; en este punto se recomienda hacer foco en medidas preventivas. Los peatones tienen prioridad al cruzar, por las esquinas o sendas peatonales, con el semáforo a su favor donde lo hay; lo cual se debe respetar por parte de los conductores, pero también concientizar a los peatones que no deben cruzar cuando el semáforo no está a su favor. Se deberían generar campañas de concientización o carteles que comuniquen e informen sobre la importancia de este punto.

[21] http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-35842016000200005 https://www.cochrane.org/es/CD004607/INJ_las-camaras-de-control-de-velocidad-reducen-los-accidentes-de-trafico-las-lesiones-y-las-muertes

- ➔ Se debe hacer hincapié en medidas de concientización o controles a motociclistas, dado que representan el 40% de los tipos de vehículos involucrados en los siniestros viales.
- ➔ Se evidencia también que un alto porcentaje de siniestros viales, tanto en muertos como en lesionados, involucran personas de entre 25 y 36 años en su mayoría de género masculino, lo que sugiere que, a pesar de que la población en C.A.B.A. tiene una mayor cantidad de mujeres, son los hombres los que corren más peligro de verse involucrados en este tipo de siniestros. Las colisiones causadas por el tránsito siguen siendo la causa principal de muerte en jóvenes.[22] Por lo que sería útil a la hora de generar líneas de acción, tener en cuenta como público objetivo personas de estas características.

El modelado de cantidad de siniestros viales busca encontrar cuáles son los barrios y horas donde se debe hacer foco de control por probabilidad de que se produzca un siniestro. Asimismo, un modelo calibrado correctamente, puede llegar a predecir con un buen nivel de precisión la cantidad de siniestros que se producirán en un barrio determinado, en períodos posteriores.

La comparación de los patrones analizados en el análisis descriptivo ayuda a la validación del modelo, dado que fundamenta los siguientes puntos:

- ➔ Hay barrios como Palermo, Flores y Balvanera, donde se dan y predicen mayor cantidad de siniestros, por lo que en estos barrios se debe tener más control. Esto seguramente esté relacionado a que son barrios concurridos que constituyen zonas de alto tránsito.
- ➔ En línea con los barrios de mayor siniestralidad vial se observa que en horas pico de la tarde se presentan mayor cantidad de siniestros. Por lo que se debería hacer foco en determinados barrios y horas por la concurrencia en esos lugares y franjas horarias de mayor circulación.

La capacidad predictiva y los resultados del modelo pueden no ser transferibles a otros conjuntos de datos porque los resultados son observación especifica. En su mayor parte, se ha realizado un esfuerzo extraordinario en el desarrollo de modelos con un ajuste estadístico superior y/o capacidades predictivas. Pero es importante tener en cuenta que este trabajo se ha visto inherentemente limitado por los datos disponibles, que han sido

[22] https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries

demasiado restrictivos en cuanto a los conocimientos que se pueden obtener y las metodologías estadísticas que se pueden desarrollar para generar estos conocimientos. La disponibilidad anticipada de nuevos datos, como, por ejemplo, datos de conducción detallados y datos de accidentes es una promesa considerable para la mejora futura.

La metodología no puede tener en cuenta desarrollos futuros imprevistos, por lo que las previsiones no resultan ser correctas en todos los aspectos. Sin embargo, proporciona un marco para examinar las incertidumbres inherentes a los desarrollos futuros y para hacer pleno uso de los conocimientos disponibles, por lo que desempeña un papel valioso en el proceso de formulación nuevas estrategias de seguridad vial y seguimiento del desarrollo de la seguridad vial. Las previsiones de seguridad vial son necesarias para establecer objetivos ambiciosos y alcanzables y monitorear el progreso hacia estos objetivos.

Aprender sobre causa y efecto a partir de datos casuales es difícil, quizás imposible. Para la seguridad vial es importante intentarlo porque, en muchos casos, no se dispone de otros enfoques de investigación. Una oportunidad para el progreso ha sido creada por un poderoso conjunto de factores, incluyendo el anuncio de una Segunda Década de Acción para la Seguridad Vial por la Asamblea General de las Naciones Unidas. En este contexto la seguridad vial no debe abordarse como un tema aislado, sino como un componente integrado de muchas agendas políticas diferentes. La necesidad de movilidad en sí misma, sin duda evolucionará en la próxima década y eso inevitablemente impulsará cambios en los sistemas de transporte de formas tanto esperadas como inesperadas. Garantizar que estos cambios no provoquen la muerte o lesiones requerirá una revisión constante.

La movilidad es parte integral de casi todos los aspectos de nuestra vida diaria. Salimos de nuestros hogares a un sistema de carreteras que nos lleva al trabajo, a la escuela, a conseguir nuestros alimentos y a muchas de nuestras necesidades familiares y sociales diarias. La influencia del sistema de transporte en las carreteras es tan generalizada que su seguridad, o la falta de ella, afecta a una amplia gama de funciones básicas. Como tal, garantizar la seguridad de las carreteras y permitir la movilidad sostenible juega un papel importante en la reducción de siniestros viales. De hecho, la eficiencia, accesibilidad y seguridad de los sistemas de transporte directamente e indirectamente contribuyen a un enfoque de un sistema seguro.

En línea con lo planteado en el Plan Global propuesto por la OMS, se entiende que se debe seguir trabajando y esforzándose en alcanzar los objetivos a partir de las pautas o estrategias definidas en este trabajo con el fin de disminuir los siniestros viales. Trabajar

en seguridad vial como prioridad porque ninguna víctima de un siniestro vial debería ser aceptable.

6.2 Posibles futuras mejoras

Pronosticar siniestros viales en el futuro es inevitablemente un proceso impreciso, porque el resultado dependerá del comportamiento futuro de los viajeros y el crecimiento de los viajes por carretera como así también de los medios utilizados durante ese periodo.

Obtener una mejor comprensión de los factores que afectan la probabilidad de un accidente automovilístico es un área de investigación constante. Desafortunadamente, los datos detallados de conducción (aceleración, frenado e información de dirección, respuesta del conductor a estímulos, etc.), datos de accidentes (por ejemplo, lo que podría estar disponible en las cajas negras de los vehículos), volumen de tránsito y distintas características de la infraestructura que permitirían una mejor identificación de las relaciones de causa y efecto con las probabilidades de choques generalmente no están disponibles. Sin embargo, hay muchas variables que se podrían investigar e incluir en la recopilación de datos de siniestros viales que ayudarían a mejorar la identificación de las relaciones de causa y efecto con los choques de vehículos individuales. Por ejemplo, si el suceso se corresponde con exceso de velocidad, con cruzar en semáforo en rojo o con alcohol al volante. Otros aspectos para considerar pueden ser las velocidades máximas permitidas en los lugares donde ocurrieron los siniestros viales, si había semáforo o no, si el cruce estaba señalizado o no. Sería interesante analizar para este tipo de variables y otras su relación con la posible ocurrencia o no de un siniestro vial.

Asimismo, podría trabajarse en predecir el número de siniestros en algún espacio geográfico durante un tiempo, para así también analizar si las víctimas fatales son producto del suceso en sí o puede estar relacionado con la demora de la atención médica en llegar.

El calendario necesario para mejorar la seguridad vial significa que es importante preparar planes a largo plazo que se basen sólidamente en los conocimientos actuales y que tengan en cuenta la probable evolución futura. Para ello se requiere de un enfoque mejorado para predecir el futuro, predicciones que tengan en cuenta los efectos de la política de seguridad vial en las tendencias de siniestros en la medida que pudieran establecerse de forma fiable los datos. Promover objetivos basados en previsiones sistemáticas de siniestros que tengan en cuenta, en la medida de lo posible, los efectos de las medidas de seguridad vial.

Una línea de trabajo interesante sería identificar nuevas medidas que podrían incluirse en una futura estrategia de seguridad vial y evaluar su probable eficacia. Esto podría darse aplicando políticas de prevención en algunas zonas y no en otras similares, de tal forma de tener grupos testigos que puedan servir para medir la efectividad de las políticas implementadas.

Existen numerosos desafíos metodológicos del análisis estadístico de los datos de siniestralidad vial, de los cuales el presente estudio no se encuentra exento y por lo tanto es necesario considerarlos a los efectos de reconocer las limitaciones y aspectos de futura mejora.

7. Referencias

- Agencia Estatal de Meteorología de España (pág. 21) http://www.aemet.es/documentos/es/eltiempo/prediccion/comun/Manual_de_uso_de_terminos_met_2015.pdf
- Ali S. Al-Ghamdi (2002) "Using logistic regression to estimate the influence of accident factors on accident severity". King Saud University, Saudi Arabia. Accident Analysis and Prevention, 34. pp. 729-741.
- Calderón Diaz D. H., Sora Vargas D. F. (2009). Universidad Distrital Francisco José de Caldas Faculta de Ingeniería. *Análisis de accidentalidad vehicular usando técnicas de minería de datos.*
- Dirección General de Estadística y Censos del Ministerio de Economía y Finanzas. Informe de resultados. *Característica de la población y sus hogares. Ciudad de Buenos Aires. Año 2017.* (Marzo 2019). https://www.estadisticaciudad.gob.ar/eyc/wp-content/uploads/2019/03/ir_2019_1350.pdf
- Dirección Nacional de Política Criminal en materia de Justicia y Legislación Pena. Subsecretaría de Justicia y Política Criminal. Secretaría de Justicia. Ministerio de Justicia y Derechos Humanos. *Programa de Estudios sobre Siniestros Viales.* (Julio 2018) https://www.argentina.gob.ar/sites/default/files/programa_de_estudios_sobre_siniestros_viales.pdf
- Dominique Lord, Fred Mannering, *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives*, Transportation Research Part A: Policy and Practice, Volume 44, Issue 5, 2010, Pages 291-305, ISSN 0965-8564
- Estadísticas del Observatorio vial de la Agencia Nacional de la Seguridad Vial del Ministerio de Transporte de Argentina. https://www.argentina.gob.ar/seguridadvial/observatoriovialnacional/estadisticas-observatorio
- Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (Second Edition*). Springer
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017*). An introduction to statistical learning (8 edition).* Springer.

- Jeremy Broughton, Jackie Knowles, Providing the numerical context for British casualty reduction targets, Safety Science, Volume 48, Issue 9, 2010, Pages 1134-1141, ISSN 0925-7535

- Joaquín A. Rodrigo (Octubre 2020). *Random Forest con Python* https://www.cienciadedatos.net/documentos/py08_random_forest_python.html
- John D. Kelleher, Brian M. Namee y Aoife D'Arcy (2015). F*undamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.*
- Juan Herrera Briones (Junio 2021). Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Industriales. *Análisis y predicción de la lesividad en accidentes de tráfico mediante la aplicación de Random Forest.*
- Ministerio de transporte, 2020. Presidencia de la Nación. *Plan federal de seguridad vial. Movilidad Segura.* https://movilidadsegura.transporte.gob.ar/media/PlanSeguridadVial.pdf
- Monroy Varela S. E., Díaz H. (2018). *Modelo de predicción de gravedad de accidentes de tránsito: un análisis de los siniestros en Bogotá, Colombia.*
- Rune Elvik, The predictive validity of empirical Bayes estimates of road safety, Accident Analysis & Prevention, Volume 40, Issue 6, 2008, Pages 1964-1969, ISSN 0001-4575
- Sanjiv Ranjan Das (2017). *Data Science: Theories, Models, Algorithms, and Analytics.*
- Sanjiv Ranjan Das (2017). *Data Science: Theories, Models, Algorithms, and Analytics.*
- Secretaría de Transporte, Subsecretaría de Movilidad Sustentable, Dirección General del Cuerpo de Agentes de Tránsito y Seguridad Vial, Observatorio de Seguridad Vial de la Ciudad de Buenos Aires. *Glosario*. https://www.buenosaires.gob.ar/sites/gcaba/files/glosario_2019_1_0.pdf
- Secretaría de Transporte y Obras Públicas GCBA, 2020-2023. *Plan de Seguridad Vial Ciudad Autónoma de Buenos Aires 2020-2023* - Secretaría de Transporte y Obras Públicas Gobierno de la Ciudad de Buenos Aires. https://www.buenosaires.gob.ar/sites/gcaba/files/plan_seg-vial_2020-2023_1_0.pdf

- Ogwueleka Nonyelum F., Department of Computer Science, Federal University – Wukari, Taraba State, Nigeria. *An artificial Neural Network Model for Road Accident Prediction: A case Study of a Developing Country.* Vol. 11, No. 5, 2014
- United Nations General Assembly (2020) *Resolution adopted by the General Assembly on 31 August 2020. Seventy-fourth session, Agenda item 12, Improving global road safety.* https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/226/30/PDF/N2022630.pdf?OpenElement
- Vipul R., Hemant J., Deepark P., Pradnya J., Monika K.. Compute Engineering Department, Shah and Anchor Kutchhi Engineering College, Mumbai. *Road Accident Prediction using Machine Learning Algorithm*. Volume: 06 Issue: 03 | Mar 2019.
- Wendy Weijermars, Paul Wesemann, Road safety forecasting and ex-ante evaluation of policy in the Netherlands, Transportation Research Part A: Policy and Practice, Volume 52, 2013, Pages 64-72, ISSN 0965-8564
- Zheng A. y Casari A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists 1st Edition.*

Apéndice A. Especificaciones conceptuales

Siniestro vial: cualquier hecho de tránsito con implicación de al menos un vehículo en movimiento, que tenga lugar en una vía pública o en una vía privada a la que la población tenga derecho de acceso, y que tenga como consecuencia al menos una persona herida o muerta. Un suicidio o intento de suicidio no se considera un accidente, sino un incidente causado por un acto deliberado de infligirse lesiones mortales. Sin embargo, si un suicidio o un intento de suicidio causan heridas a otro usuario, entonces el incidente debe ser considerado un accidente con víctimas. Se incluyen: las colisiones entre vehículos; entre vehículos y peatones; entre vehículos y animales u obstáculos fijos; los siniestros viales con la intervención de sólo un vehículo; y las colisiones entre vehículos y trenes. Las colisiones múltiples se contabilizan como un solo hecho de tránsito si las colisiones se suceden en un periodo de tiempo muy corto. Se excluyen los hechos de tránsito con sólo daños materiales. Se excluyen los actos terroristas.

Víctima: cualquier persona muerta o herida como consecuencia de un siniestro vial.

Homicidio: se considera la definición avalada por organismos internacionales de víctima fatal de siniestro de tránsito como aquella persona que fallece de inmediato o dentro de los 30 días siguientes como consecuencia de un traumatismo causado por el siniestro (se exceptúan los suicidios). Hoy en día, la fuente oficial de información sobre víctimas fatales en siniestros viales que ocurren en la Ciudad de Buenos Aires es el Ministerio de Justicia y Seguridad (MJYS) del Gobierno de la Ciudad Autónoma de Buenos Aires. El MJyS incluye casos de víctimas fatales que ocurren en el lugar del hecho y hasta 7 días posteriores al siniestro. En este sentido, para alcanzar con la definición propuesta de víctimas a 30 días se suele emplear un factor de ajuste normalizado, recomendado por la Conferencia Europea de Ministros de Transportes (CEMT), este consiste en multiplicar por 1,08 las víctimas a 7 días.

Lesionados: cualquier persona que, como consecuencia de un siniestro vial con víctimas, no resulte muerta en el acto o dentro de los 30 días siguientes (7 días siguientes para MJyS), pero sufra lesiones. Normalmente, estas lesiones requieren tratamiento médico. Se excluyen los intentos de suicidio. Las personas con lesiones muy leves, como pequeños cortes o magulladuras, no suelen ser registradas como heridas. Se excluyen los casos en los que la autoridad competente declara que la causa de la herida ha sido un intento de suicidio.

Rol: corresponde a la forma de desplazamiento de la víctima ya sea ciclista, conductor, pasajero o peatón, o una combinación de estas.

- **Conductor**: cualquier persona implicada en un siniestro vial con víctimas, que estuviera conduciendo un vehículo en el momento del hecho.
- **Pasajero**: Toda persona que, sin ser conductor, se encuentra dentro o sobre un vehículo en el momento del siniestro vial, o es arrollada mientras está subiendo o bajando del vehículo.
- **Peatón**: cualquier persona implicada en un hecho de tránsito con víctimas, distinta de un conductor o un pasajero. Se incluyen los ocupantes o personas que empujan o arrastran un coche de bebé o de una silla de ruedas o cualquier otro vehículo sin motor de pequeñas dimensiones. Se incluye también las personas que caminan, empujan una bicicleta, un ciclomotor, o se desplazan sobre patines, skates y otros artefactos parecidos.

Tipo (corresponde a tipo de vehículo): caracteriza a las víctimas fatales y lesionadas según el medio en el que se transportan –tipo de vehículo o forma de desplazamiento– en el momento del siniestro. En el caso de las víctimas de rodados motorizados, se diferencia a las mismas según el tipo de vehículo.

Se consideran como tipos de vehículo:

1) Automóvil
2) Camión
3) Camioneta
4) Utilitarios
5) Taxi y remis
6) Transporte público,
7) Moto,
8) Bicicleta,
9) Peatón,
10) Transporte público
11) Otros (ambulancia, autobomba, cuatriciclo, fuerza de seguridad, cuatriciclo, etc)

Apéndice B. Variables del datasets original

Quality report

index	Data Type	Missing Values	Unique Values
causa	object	-	2
rol	object	335	9
tipo	object	1.100	24
sexo	object	197	2
edad	float64	559	104
mes	float64	37	12
periodo	int64	-	4
fecha	object	61	1.247
hora	object	17	1.304
lugar_hecho	object	3.716	20.110
direccion_normalizada	object	765	15.521
tipo_calle	object	-	3
direccion_normalizada_arcgis	object	772	15.519
calle1	object	765	1.197
altura	float64	26.123	3.521
calle2	object	8.247	1.391
codigo_calle	float64	9.290	1.033
codigo_cruce	float64	12.525	1.215
geocodificacion	object	1.095	17.959
semestre	int64	-	2
x	float64	1.095	15.600
y	float64	1.095	15.584
geom	object	1.095	16.914
cantidad_victimas	int64	-	12
comuna	float64	1.098	15
geom_3857	object	1.095	16.550
tipo_colision1	object	1.343	18
participantes_victimas	object	1.131	140
participantes_acusados	object	677	183

causa: variable del tipo string que expone si el siniestro vial se trata de un homicidio o lesión.

rol: variable del tipo string que representa el rol de la víctima en el siniestro vial, ya sea ciclista, conductor, pasajero o peatón.

tipo: variable del tipo string que caracteriza a la víctima según el tipo de vehículo o forma de desplazamiento en el momento del siniestro.

sexo: variable de tipo factor que representa el género de la víctima, posee 2 niveles, masculino y femenino.

edad: variable de tipo numérico que representa la edad de la víctima

mes: variable de tipo numérico que representa el número del mes correspondiente al siniestro vial, siendo enero = 1 y diciembre = 12.

periodo: variable de tipo numérico que representa el número del año correspondiente al siniestro vial.

fecha: variable del tipo string que expone la fecha (día, mes y año) en que ocurrió el siniestro vial.

hora: variable del tipo string que expone la hora exacta en que ocurrió el siniestro vial.

lugar_hecho: variable del tipo string que expone el lugar del hecho, con la dirección donde ocurrió el suceso.

direccion_normalizada: variable del tipo string que corresponde al lugar_hecho normalizado.

tipo_calle: variable del tipo string que corresponde al tipo de calle donde ocurrió el hecho, avenida, calle, autopista.

direccion_normalizada_arcgis: variable del tipo string que corresponde al lugar_hecho y direccion_normalizada pero normalizada con la herramienta de geoprocesamiento de ArcGIS que estandariza la información sobre direcciones en una tabla o clase de entidad.

calle1: variable del tipo string que corresponde a la primera calle involucrada del lugar del hecho (por ejemplo, para la siguiente direccion_normalizada_arcgis “cafayate & garcia grande de zequeira, severo”; la calle1 es “cafayate”).

altura: variable del tipo numérico que corresponde a la altura correspondiente a la primera calle involucrada del lugar del hecho (por ejemplo, para la siguiente direccion_normaliza_arcgis: “1730 lavalle”; la calle1 es “lavalle” y la altura: “1730”)

calle2: variable del tipo string que corresponde a la segunda calle involucrada del lugar del hecho, generalmente para los hechos identificados como intersección de calles (por ejemplo para la siguiente direccion_normalizada_arcgis “cafayate & garcia grande de zequeira, severo”; la calle2 es “garcia grande de zequeira, severo”).

codigo_calle: variable de tipo numérico que corresponde al código de la calle del suceso (lugar_hecho).

codigo_cruce: variable de tipo numérico que corresponde al código del cruce donde ocurrió el siniestro vial (lugar_hecho)

geocodificacion: variable del tipo string que corresponde a la geocodificación de las direcciones normalizadas obteniendo las coordenadas en el sistema de referencias utilizados por la CABA[23] del siniestro vial.

semestre: variable del tipo numérico que corresponde al semestre correspondiente al siniestro vial, siendo el primer semestre (enero, febrero, marzo, abril, mayo y junio) = 1 y segundo semestre (julio, agosto, septiembre, octubre, noviembre y diciembre) = 2.

[23] https://usig.buenosaires.gob.ar/manual_normalizador_de_direcciones_USIG.pdf

x: variable del tipo numérico que corresponde a la longitud de las coordenadas geográficas de la ubicación del siniestro vial.

y: variable del tipo numérico que corresponde a la latitud de las coordenadas geográficas de la ubicación del siniestro vial.

geom: variable del tipo string que corresponde a una geo codificación con números y letras de la ubicación del siniestro vial.

cantidad_victimas: variable del tipo numérico que corresponde a la cantidad de victimas involucradas en el siniestro vial.

comuna: variable del tipo numérico que corresponde a la comuna donde ocurrió el siniestro vial.

geom_3857: variable del tipo string que corresponde a una geo codificación con números y letras de la ubicación del siniestro vial.

tipo_colision1: variable del tipo string que expone el tipo de colisión según los participantes involucrados y su forma/vehículo de desplazamiento. (ej. Bicicleta-Vehículo, Vehículo - Motovehículo), también puede identificarse como múltiple, como el siniestro en el que intervienen más de dos vehículos

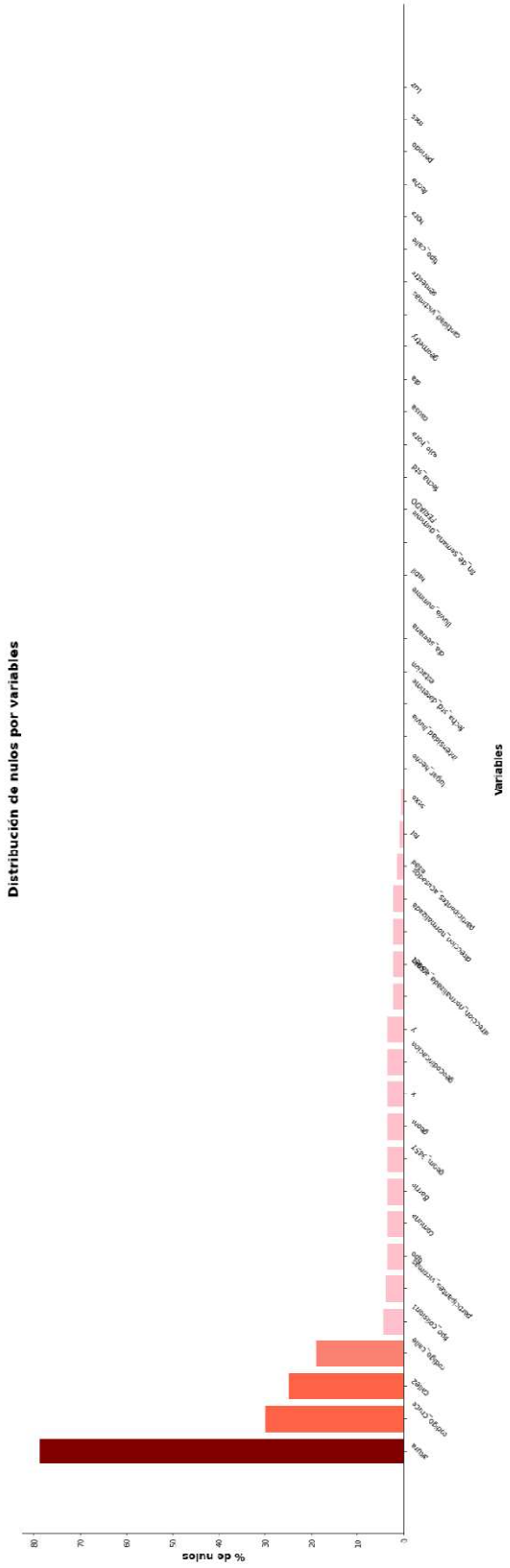
participantes_victimas: variable del tipo string que expone el tipo de vehículo o forma de desplazamiento de la víctima.

participantes_acusados: variable del tipo string que expone el tipo de vehículo o forma de desplazamiento del acusado.

Apéndice C. Tabla Hiperparámetros de Random Forest

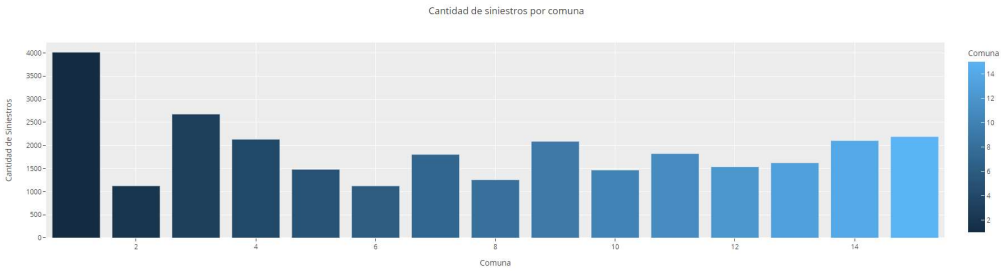
Hiperparámetro	Descripción
n_estimators	número de árboles incluidos en el modelo
max_depth	profundidad máxima que pueden alcanzar los árboles
min_samples_splt	número mínimo de observaciones que debe de tener un nodo para que pueda dividirse. Si es un valor decimal se interpreta como fracción del total de observaciones de entrenamiento.
min_samples_leaf	número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división. Si es un valor decimal se interpreta como fracción del total de observaciones de entrenamiento.
max_leaf_nodes	número máximo de nodos terminales que pueden tener los árboles.
max_features	número de predictores considerados a en cada división: - Un valor entero - Una fracción del total de predictores - "auto", utiliza todos los predictores - "sqrt", raiz cuadrada del número total de predictores - "log2", log2 del número total de predictores - None, utiliza todos los predictores
oob_score	Si se calcula o no el out-of-bag R^2. Por defecto es False ya que aumenta el tiempo de entrenamiento.
n_jobs	número de cores empleados para el entrenamiento. En random forest los árboles se ajustan de forma independiente, por lo la paralelización reduce notablemente el tiempo de entrenamiento. Con -1 se utilizan todos los cores disponibles.
random_state	semilla para que los resultados sean reproducibles. Tiene que ser un valor entero.

Apéndice D. Gráfico de distribución de nulos por variables

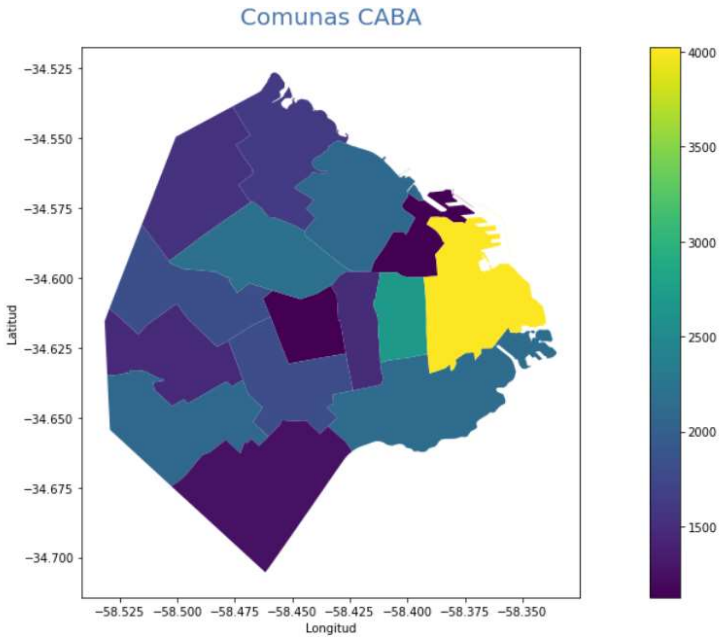


Apéndice E. Distribución de siniestros por comuna

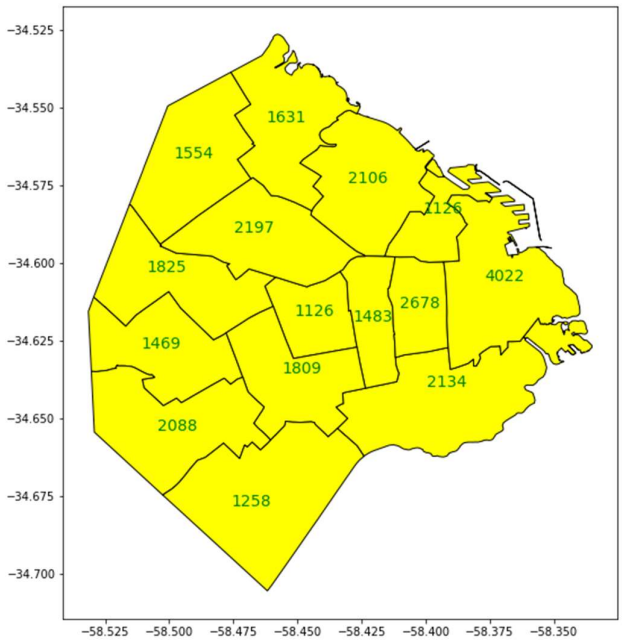
Cantidad de siniestros por COMUNA de C.A.B.A. 2015-2017



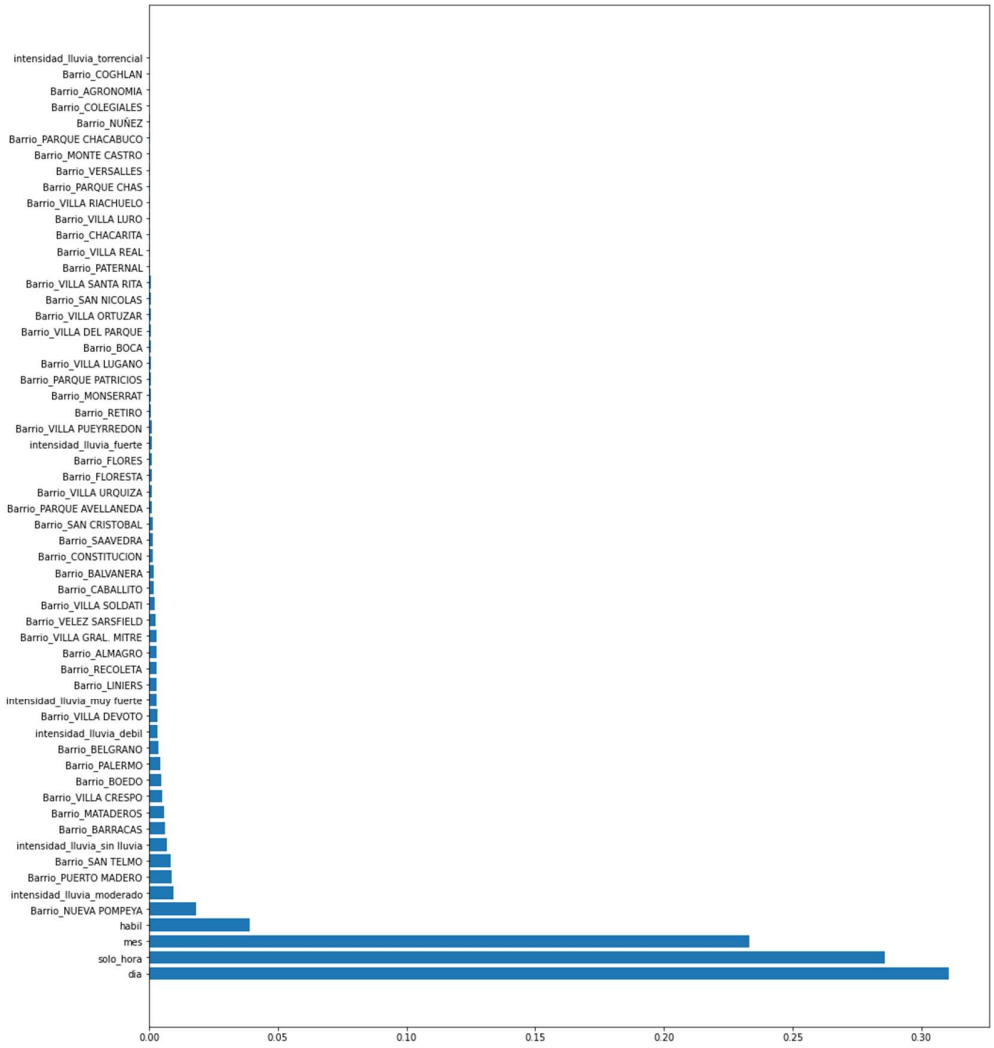
Mapa de calor de siniestros viales por comuna de C.A.B.A 2015-2017



Mapa con la cantidad de siniestros viales por comuna de C.A.B.A 2015-2017



Apéndice F. Gráfico Feature Importante con variables categóricas binarizadas



Apéndice G. Dataset siniestros viales 2018-2019

Los datasets con los delitos por siniestro vial informados por el Ministerio de Justicia y Seguridad se obtuvieron de los datasets de delitos por "subtipo_delito = Siniestro Vial" disponibles en el sitio web BA DATA[24].

Estos tienen la siguiente estructura original:

Dataset original de siniestros viales con datos de junio 2018 a diciembre 2019 (MJyS)

id	fecha	franja_horaria	tipo_delito	subtipo_deli	cantidad_reg	comuna	barrio	lat	long
299386	8/1/2018	15	Lesiones	Siniestro Vial	1.0	1.0	ConstituciÃ³r	-34.619.755	-58.381.045
293134	8/1/2018	13	Lesiones	Siniestro Vial	1.0	4.0	Barracas	-346.407	-58.386.729
298851	8/1/2018	15	Lesiones	Siniestro Vial	1.0				
296099	8/1/2018	10	Lesiones	Siniestro Vial	1.0	14.0	Palermo	-34.596.537	-58.425.979
294894	8/1/2018	7	Lesiones	Siniestro Vial	1.0				
293871	8/1/2018	14	Lesiones	Siniestro Vial	1.0	3.0	Balvanera	-3.461.538	-58.397.817

Se describe el perfilado de cada variable (quality report) del dataset que contiene 6.139 registros y 13 columnas.

Perfilado de cada variable (quality report) del dataset junio 2018 - diciembre 2019.

	Data Type	Missing Values	Unique Values
id	int64	0	6139
fecha	datetime64[ns]	0	303
dia	int64	0	31
mes	int64	0	12
año	int64	0	2
franja_horaria	object	0	25
tipo_delito	object	0	2
subtipo_delito	object	0	1
cantidad_registrada	float64	0	2
comuna	float64	1173	15
barrio	object	1173	48
lat	float64	1173	3734
long	float64	1173	3731

A partir de este análisis, dado que el dato del barrio tenía datos faltantes (19%), decidimos eliminar dichos registros y continuar con la incorporación de nuestras variables definidas en el apartado **"Transformación de datos y agregado de variables"**.

Luego de esto obtuvimos nuestro dataset a utilizar para validación con datos actuales, el cual contenía 4.965 registros.

[24] https://data.buenosaires.gob.ar/dataset/delitos