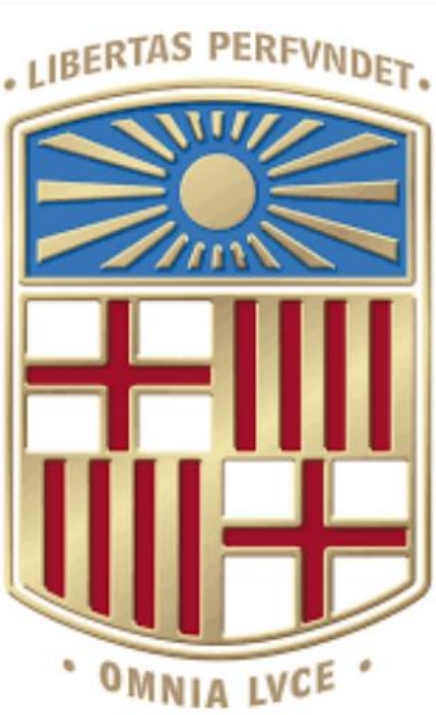


Universitat de Barcelona

Modelamiento de la severidad de los accidentes de tránsito de Maryland

Trabajo final para optar al título de

Máster en Data Science & Big Data

Valerim Daiana Clavijo Amorin

Heiner Romero Leiva

Nicolás Bene Rodríguez

Felipe Orlando Franz Silva

Capítulo 1

Selección del tema y objetivos

En el presente capítulo se presenta la selección de la temática elegida para el trabajo final de máster, así como los objetivos y justificación de la misma. Se presentan también los stakeholders de interés para el tema seleccionado, seguido de los antecedentes investigados sobre el mismo. Por último, se procede a presentar el dataset en cuestión, realizando una breve descripción del mismo.

1.1) Tema del TFM

Analizar datos de accidentes de tránsito geolocalizados, con la finalidad de predecir la severidad de los mismos, es decir si habrá heridos o no. Para ello se usan los datos de accidentes del estado de Maryland, Estados Unidos en el periodo comprendido entre los años 2016 a 2019.

1.2) Objetivos del TFM

1.2.1) Objetivo general

Aplicar modelos de minería de datos en el estado de Maryland en Estados Unidos en el periodo 2016 a 2019 mediante el uso de datos georreferenciados con la finalidad de predecir si, dado que sucede un accidente, si en los mismos habrán o no personas heridas, así como también poder determinar qué variables son las que ejercen mayor influencia en la severidad de estos accidentes.

1.2.2) Objetivos específicos

a. Realizar un análisis exploratorio de datos de los accidentes que han ocurrido en el estado de Maryland.

b. Predecir la severidad de los accidentes de tránsito (existencia o no de heridos) haciendo uso de técnicas avanzadas de minería de datos y estadística espacial.

c. Contrastar los principales factores que más inciden en la severidad de dichos accidentes, tomando como punto de referencia los modelos estadísticos propuestos.

d. Dar a conocer los principales hallazgos que inciden en dichos accidentes para proponer recomendaciones y detectar oportunidades de mejora.

1.3) Justificación

Los accidentes de tránsito son uno de los problemas más grandes que se enfrentan hoy en día a nivel mundial, sobre todo en países en vías de desarrollo. La organización Mundial de la Salud ha denotado este problema y ha indicado que estos accidentes no sólo atentan contra la calidad de la vida de las personas sino también que representan un problema que va más allá de una simple enfermedad o muertes espontáneas, en realidad son muertes causadas (Yassin, 2020).

Con el rápido desarrollo de la urbanización, el número de vehículos utilizados en las ciudades se incrementa velozmente, con lo cual también aumentan los accidentes de tránsito (Fan, Liu, Cai, y Yue, 2019). Los accidentes de tránsito provocan numerosas muertes, heridos, y también pérdidas económicas, por lo que su prevención resulta fundamental para evitar tales incidentes (Santos, Saias, Quaresma, y Nogueira, 2021).

Predecir accidentes de tránsito permite tomar acciones para evitarlos, reducir sus daños en caso que sucedan, dar alertas a los conductores de potenciales peligros, realizar mejoras en las condiciones urbanísticas en ciertos sectores, así como mejorar el sistema de atención de emergencias (Santos, Saias, Quaresma, y Nogueira, 2021).

Las autoridades han sido enfáticas en denotar que es necesario pausar este problema, ya que no solo tiene repercusiones a nivel económico y social, sino también a nivel de desarrollo y calidad de vida. En recientes años y con la ayuda de las matemáticas aplicadas se han realizado diversas aproximaciones a soluciones para pausar este tipo de accidentes, más tarde con la llegada de los algoritmos de aprendizaje automático, se han podido

generar nuevas soluciones. Es por ello que en el presente trabajo se pretenden elaborar modelos que permitan predecir la severidad de los accidentes, a efectos de que se puedan tomar medidas para disminuir su probabilidad de ocurrencia, mitigar su impacto en caso de que ocurran, o reducir su intensidad y frecuencia, y con ello crear una mejora en la calidad de vida de las personas.

1.4) Stakeholders

En el presente proyecto no se trabajará de forma asociada con una empresa privada ni con un organismo público u organización no gubernamental (ONG). Si bien los datos que se van a utilizar son de Maryland, no se pretende trabajar conjuntamente con las autoridades locales de dicho estado, ni con ningún otro tipo de organización pública o privada.

Sin embargo, tal como fue mencionado en la justificación del trabajo, se entiende que un análisis de este tipo puede ser de mucha utilidad para autoridades de gobiernos locales, así como a empresas privadas que se dedican a la tercerización de servicios de la Administración Pública y de los Ministerios de Transporte de diversos países, no solo de Estados Unidos. Cabe destacar que este tipo de análisis no son comunes en América Latina debido a diversos factores que aún enfrentan dichas administraciones estatales como lo son: baja calidad de los datos, ausencia de mecanismos para capturar y resguardar los mismos, desconocimiento de cómo tratarlos y explotarlos, lo cual dificulta en gran medida su obtención. Por lo tanto los resultados que puedan surgir de este trabajo podrían ser de utilidad para diversas autoridades de Latinoamérica para que así puedan comprender la importancia de llevar un registro adecuado y centralizado de datos de accidentes de tránsito y de esta forma generar mecanismos para su captura y obtención. Además, si se logra obtener cuáles son las variables que influyen más en la severidad de dichos accidentes, las autoridades podrían actuar sobre las mismas y ajustar cada una de ellas a su contexto específico.

Con lo mencionado anteriormente, se puede decir que las organizaciones y personas a las que les podría interesar este estudio son:

- Autoridades de gobiernos locales (alcaldías, municipalidades, entre otros)

- Ministerios de Transporte
- Empresas privadas que se dediquen a la tercerización de servicios relacionados al transporte e infraestructura urbana
- Organizaciones público-privadas
- Población en general, tales como: conductores, ciclistas, peatones, etc.
- Compañías de Seguros (para valuar las pólizas de seguros de accidentes de tránsito)
- Académicos que estudian el tráfico y la siniestralidad

1.5) Antecedentes

En la última década la predicción de accidentes de tránsito ha sido activamente estudiada como consecuencia de dos factores: por un lado, ha existido un movimiento activo de solicitud a los gobiernos y administraciones públicas a que publiquen los datos que generan sus actividades (open data), de forma que todo ciudadano interesado pueda acceder a los mismos. Por otra parte, el desarrollo de la Analítica de Big Data y la Ciencia de Datos, ha permitido mejorar el procesamiento de los grandes volúmenes de datos que manejan los organismos públicos, y el desarrollo de modelos de predicción. Es decir, que con la disponibilidad de grandes volúmenes de datos sobre accidentes de tránsito, y con la mejora en las técnicas e infraestructura para tratarlos, ha sido posible que aumente la cantidad y que mejore la precisión de los modelos de predicción de dichos accidentes (Hébert, A., Guédon, T., Glatard, T., y Jaumard, B., 2019).

Por otro lado es imperioso que las administraciones públicas de los países de América Latina avancen hacia un gobierno más abierto en el que sus ciudadanos sean capaces de poder conocer en todo momento todos los esfuerzos que se están realizando por parte de los Estados en la mejora general de su calidad de vida y con ello rendir cuentas. Hoy en día, los datos son el recurso más valioso que existe para lograr dicha consecución e infortunadamente en América Latina no se le ha dado la misma importancia que en otros lugares. Oszlak (2013) defiende que:

> *"Es que poseer (y retener, negar o distorsionar) información equivale a disponer de un inapreciable recurso de poder. (...) la información, es decir, el conocimiento experto o la disponibilidad de datos no conocidos*

> *por otros, que ofrece una capacidad diferencial a quien la posee para decidir o actuar." (*Oszlak, 2013, p. 8-9).

Es así como se puede generar la siguiente premisa: ¿verdaderamente los accidentes de tránsito se pueden frenar si las administraciones públicas capturan dichos datos y los ponen al alcance de la ciudadanía? O, ¿es posible que dichos datos se puedan utilizar para frenar la ola de accidentes que suceden hoy en día? La respuesta es sí, ya que con los últimos avances en ciencia y tecnología y con el estallido que tienen las técnicas de predicción más avanzadas que en la actualidad se conocen como *Machine Learning,* se han dado respuestas a interrogantes que en el pasado no eran tan fáciles de dilucidar o cuya respuesta no era tan rápida de obtener. Ahora bien, Oszlak (2013) menciona lo siguiente*:*

> *"Los sistemas de información suelen ser el talón de Aquiles de la responsabilización. Si no se dispone de los datos necesarios para establecer la distancia entre las metas que deben cumplirse y los efectos conseguidos, resultará imposible que funcione un proceso transparente y objetivo de rendición de cuentas. No podrá saberse qué insumos fueron asignados a qué responsables, cuáles fueron las actividades que se completaron ni, menos todavía, qué efectos se lograron a través de los productos obtenidos. Idealmente, estos sistemas no sólo deberían informar cuál fue el desempeño en el proceso de conversión de insumos en productos (eficiencia), sino también de qué manera se convirtieron los productos en efectos o resultados inmediatos (efectividad), dimensión mucho más difícil de observar frente a la multidimensionalidad de la mayoría de las cuestiones de política pública."* (Oszlak, 2013, p. 27 a través de Norton y Elson, 2002)."

Es por ello que es relevante indicar que los accidentes de tránsito, su ocurrencia y magnitud son un problema serio a nivel de políticas públicas y que los mismos no se deben a cuestiones del azar. En la medida en que los datos de dichos accidentes se puedan tratar y monitorear, es posible generar indicadores alentadores para luchar contra este problema.

En la actualidad existen diversos estudios y artículos académicos que realizan modelos de predicción de accidentes, donde se buscan principalmente dos objetivos:

1. Predecir los lugares más probables donde ocurrirán accidentes (*hotspots*)
2. Determinar los factores que más inciden en la ocurrencia y severidad de esos accidentes

Con respecto al primero de los objetivos mencionados, generalmente la ocurrencia de un accidente es la etiqueta que se usa para entrenar al modelo y con el mismo se intenta predecir dónde es más probable que ocurra uno (Hébert, A., Guédon, T., Glatard, T., y Jaumard, B., 2019). A esos lugares en donde existen ciertos factores que pueden llevar a que ocurra un número significativo de accidentes se les suele llamar *hotspots*, aunque también se los denomina como *blackspots,* puntos de alto riesgo, entre otros (Montella, 2010). Existen varios métodos para detectar estos hotspots, a los que se les denomina *HotSpot IDentification (HSID) methods*, y no necesariamente son algoritmos de Machine Learning. Con estos métodos se busca, entonces, resolver un problema de clasificación binario ya que se quiere clasificar qué puntos o *hotspots* se activan dadas ciertas circunstancias (Santos, Saias, Quaresma, y Nogueira, 2021). La clase "positiva" es por lo tanto la ocurrencia de un accidente en un momento y lugar determinados, mientras que la "negativa" es que no se produzca un accidente en ese momento y lugar (Hébert, A., Guédon, T., Glatard, T., y Jaumard, B., 2019).

Por otro lado, existen autores que mencionan que se pueden utilizar otros modelos como Modelo Autorregresivo Integrado de Media Móvil (ARIMA), ya que los accidentes se pueden tratar como un problema de series temporales y su ocurrencia así como severidad corresponde a ciclos estacionales concretos. (Sangare, M., Gupta, S., Bouzefrane, S., Banerjee S., Mühlethaler, P., 2021). Asimismo, estos autores mencionan que se pueden utilizar otras aproximaciones utilizando modelos gráficos probabilísticos como lo son: redes bayesianas, cadenas de Markov, Markov Random Fields (MRFs) y otros de la familia de métodos no paramétricos como las Redes Neuronales Artificiales y Support Vector Regression.

Sin embargo, existen múltiples razones para la fluctuación en el flujo de tráfico y además los patrones en los datos son multimodales. Esto dificulta el aprendizaje y los enfoques basados en redes para poder modelar mapas complejos, ya que requieren un espacio dimensional alto, por lo que para tener un espacio de alta dimensión es necesario cumplir con el requisito de una gran cantidad de datos anotados. Por lo tanto, en el espacio de alta dimensión, el problema de sobreajuste se vuelve común.

Es por ello que para superar este problema, Sangare, M., Gupta, S., Bouzefrane, S., Banerjee S., y Mühlethaler, P. (2021), recomiendan una estructura no lineal multicapa ya que los enfoques de aprendizaje profundo tienen una gran capacidad para expresar patrones de modelos múltiples en datos utilizando un número reducido de dimensiones. Un ANN (Red Neuronal Artificial) es un tipo de red en aprendizaje automático que ha sido ampliamente utilizado para la predicción de incidentes viales en diferentes entornos (autopista, urbano y carreteras no urbanas, etc.) con el fin de minimizar las lesiones y la pérdida de vidas en las carreteras. En esta misma línea, los autores mencionados recomiendan utilizar métodos para reducción de la dimensionalidad, es por ello que no recomiendan el uso de modelos de redes bayesianas o las Máquinas de Soporte Vectorial (a nivel de regresión), ya que carecen de la capacidad de seleccionar las características más relevantes del conjunto de datos.

Es así como se recomienda el uso de Modelos Mixtos Gaussianos (Gaussian Mixture Model) o Máquinas de Soporte Vectorial pero utilizando el clasificador, ya que fácilmente los puntos se pueden asociar a probabilidades que varían en función de los datos y estos no tienen que venir previamente normalizados o su ocurrencia no debe seguir un patrón normal, sino más bien que estos tratan de encontrar valores atípicos en los datos y de ahí crean sus predicciones. Sin embargo, para Sangare, M., Gupta, S., Bouzefrane, S., Banerjee S., y Mühlethaler, P. (2021), lo ideal es poder hacer combinaciones de diferentes modelos, siendo en este caso una buena aproximación las Máquinas de Soporte Vectorial junto con los Modelos Mixtos Gaussianos.

Cabe destacar que, según diversos autores (Hébert, A., Guédon, T., Glatard, T., y Jaumard, B., 2019; Santos, Saias, Quaresma, y Nogueira, 2021), los algoritmos de Machine Learning que se usan con mayor frecuencia para predecir hotspots son: árboles de decisión, random

forest, XGBoost, K vecinos más cercanos (KNN), regresión logística Bayesiana, naive Bayes, redes Bayesianas, y redes neuronales.

Un aspecto importante de la predicción de hotspots que es mencionado por Hébert, A., Guédon, T., Glatard, T., y Jaumard, B. (2019) es que, al tratarse de un problema de clasificación donde el evento que se pretende detectar es raro, ya que es más frecuente que no haya accidentes de tránsito a que estos ocurran, puede haber problemas de desbalance de datos. Los modelos de machine learning suelen enfocarse en reducir el error general y no en detectar la clase "favorable" que sería el accidente.

En lo que respecta al segundo objetivo, que es el de detectar los factores y las variables que inciden en los accidentes, en la revisión de literatura sobre el tema (Hébert, A., Guédon, T., Glatard, T., y Jaumard, B., 2019; Yassin, S.S., 2020; Santos, Saias, Quaresma, y Nogueira, 2021) se observó que aparecen como significativos generalmente los siguientes:

- flujo del tráfico
- variables temporales que influyen en el tráfico, tales como: hora del día, día de la semana, día del año, horas pico, entre otras.
- temperatura
- visibilidad
- cantidad de accidentes en el mismo lugar en los años anteriores
- orientación de las calles o avenidas
- velocidad y dirección del viento
- experiencia de los conductores
- sexo y edad del conductor
- tipo de vehículo
- año de servicio del vehículo
- estado de la calle

Cabe destacar que algunas de estas variables dependen del contexto de la ciudad a analizar. Este contexto incluye al clima, el estado de la flota de vehículos que circula en la ciudad, la exigencia para obtener la libreta de conducir, el límite de alcoholemia, entre otros. Por ejemplo, la nieve no sería un factor a considerar en una ciudad donde no ocurre dicho

fenómeno climático, así como en el caso de la lluvia como afecte la misma a los accidentes va a depender de los milímetros al año registrados.

En función de los antecedentes reseñados, es que se decidió para el presente trabajo desarrollar modelos que permitan predecir la severidad de los accidentes (si hay o no personas heridas en los mismos), y analizar a partir de algunos de estos cuáles son los factores que más influyen en la aparición de accidentes graves. Para el caso de los factores, es necesario utilizar modelos que no sean de caja negra y sean interpretables, a efectos de que los diferentes stakeholders mencionados puedan tomar decisiones sobre los mismos.

1.6) Dataset

1.6.1) Selección del dataset a utilizar

En un principio, se analizaron varios conjuntos de datos de diferentes regiones geográficas: Montevideo, Antofagasta, Madrid, el Estado de Maryland, entre otras. Para realizar la elección del dataset a utilizar se tuvieron en cuenta los siguientes criterios:

- **Cantidad de observaciones del dataset:** se buscaron bases de datos extensas, que implicarán realizar cruces entre diferentes tablas que llevaran a tener más de dos millones de registros en total entre todas las tablas analizadas.
- **Número de variables**: se examinó que existieran variables relevantes que pudieran ser explicativas del accidente, y que hagan referencia tanto a factores ambientales (como la hora, el clima), como a datos y condiciones de los conductores (edad, alcohol en sangre, por ejemplo), características de los autos involucrados, y señalización de las calles (existencia de semáforos, calidad de las calles), entre otras.
- **Período de tiempo al que hace referencia**: se buscó que hubieran datos de varios años, por diversos motivos: para tener un mayor volumen de datos; para analizar si los hotspots cambian con el tiempo; y porque en la bibliografía revisada se vio que la cantidad de accidentes que ocurrieron en un lugar en años anteriores es un factor que permite predecir si van a ocurrir accidentes nuevamente en dicho lugar.

- **Georeferenciación de los datos**: se analizó que los accidentes estuvieran localizados geográficamente por medio de coordenadas, especialmente que tuvieran la latitud y la longitud, a efectos de facilitar luego la elaboración de mapas.
- **Calidad del diccionario de variables**: se dio importancia a este hecho debido a que resulta trascendental entender qué contiene cada variable y cuáles son sus limitaciones. Asimismo es una herramienta con la cual se obtiene un análisis preliminar de la calidad de las variables.

En función de los criterios seleccionados, se decidió utilizar los datos de los accidentes del estado de Maryland de Estados Unidos. El dataset se obtuvo de la página web de datos abiertos del gobierno de dicho país: https://data.gov/. Más específicamente se filtró para el estado mencionado los datos relativos al tráfico[1].

1.6.2) Características del dataset

El dataset está formado por una tabla principal con todos los accidentes ocurridos entre 2015 y 2021. Para cada uno de esos accidentes hay un número de reporte, y se poseen tres bases de datos que posteriormente se unirán: la base principal que refiere a las circunstancias del accidente; otra base con las características de las personas involucradas; y una tercera base con los vehículos presentes en el accidente. El dataset, y la limpieza que se le realiza al mismo se detallan con mayor profundidad en el próximo capítulo.

1

Ver:https://catalog.data.gov/dataset?tags=traffic&publisher=opendata.maryland.gov&ext_location=&q=traffic&ext_prev_extent=-141.6796875%2C8.754794702435618%2C-59.4140625%2C61.77312286453146&sort=views_recent+desc&ext_bbox=&organization=state-of-maryland&page=1

Capítulo 2

Descripción, limpieza, estructuración y depuración de los datos

En el presente capítulo se describen brevemente las tablas que componen la base de datos seleccionada para el trabajo final del máster, para luego proceder a la unión, limpieza y transformación de la misma, a efectos de tener una base lista para elaborar modelos que permitan cumplir con los objetivos planteados en el apartado anterior.

Durante la limpieza nos daremos cuenta que existen inconsistencias sustanciales, errores de tipografía y valores missing (NaN o NA). Además de una excesiva tipificación de ciertas categorías, en donde se podrían crear nuevas más generales para así facilitar el estudio. Dentro de esta fase, se procede también a realizar una selección justificada de las variables más importantes para los objetivos mencionados.

En el Capítulo 1 se mencionó que el set de datos analizado está compuesto por tres tablas distintas. El objetivo del presente Capítulo 2 es unir las mismas para crear un dataframe que contemple todas las variables, al cual se le realizará una exhaustiva limpieza de datos para aplicarlos de manera óptima a los modelos en los capítulos siguientes.

Para ello, se unen las tres tablas mencionadas, a saber:

1) Maryland Statewide Vehicle Crashes
2) Maryland Statewide Vehicle Crashes Person Details Anonymized
3) Maryland Statewide Vehicle Crashes Vehicle Details

2.1) Base de datos

2.1.1) DF1: Maryland Statewide Vehicle Crashes

Tenemos la base principal llamada *Maryland Statewide Vehicle Crashes*, a la cual denominaremos como DF1. Tiene dimensión 771,145 x 56. Presentamos un resumen de las variables y su descripción en la siguiente tabla.

Tabla 1.1 - Variables y descripción de DF1

Variables	Descripción	ACC TIME	Hora del accidente
YEAR	Año del accidente	LOC CODE	Códigos locales
QUARTER	Cuatrimestre del año	SIGNAL FLAG DESC	Presencia o no de señal de tránsito
LIGHT DESC	Descripción de características de la luz del día	SIGNAL FLAG	Presencia o no de señal de tránsito (misma variable que la anterior)
LIGHT CODE	Código de luz de día	CM ZONE FLAG	Presencia o no de señal de que hay una obra en la calle
COUNTY DESC	Condado	AGENCY CODE	Código de agencia que habilitó el reporte
COUNTY NO	Código del condado	AREA CODE	Código de información del área local
MUNI DESC	Variable vacía, solo tiene valores NA	HARM EVENT DESC1	Descripción de evento de daño 1 (daño a peatón, otro auto, objeto fijo, étc).
MUNI CODE	Código que describe MUNI DESC, pero no está presente el significado del mismo en el diccionario	HARM EVENT CODE1	Código de evento de daño 1 (daño a peatón, otro auto, objeto fijo, étc).
JUNCTION DESC	Tipo de cruce	HARM EVENT DESC2	Descripción de evento de daño 2 (daño a peatón, otro auto, objeto fijo, étc).
JUNCTION CODE	Código de tipo de cruce	HARM EVENT CODE2	Código de evento de daño 2 (daño a peatón, otro auto, objeto fijo, étc).
COLLISION TYPE DESC	Tipo de choque	RTE NO	Número de ruta
COLLISION TYPE CODE	Código de tipo de choque	ROUTE TYPE CODE	Código de tipo de ruta
SURF COND DESC	Condiciones de la superficie de la calle (nieve, húmedo, agua,étc)	RTE SUFFIX	Sufijo de ruta
SURF COND ODE	Código condiciones de la superficie de la calle (nieve, húmedo, agua,étc)	LOG MILE	Milla de la calle
LANE DESC	Tipo de carril	LOGMILE DIR FLAGDESC	Dirección de la calle (este, oeste, sur, norte)
LANE CODE	Código del tipo de carril	LOGMILE DIR FLAG	Código dirección de la calle
RD COND DESC	Condiciones de la calle	MAINROAD NAME	Nombre de la calle principal
RD COND CODE	Código condiciones de la calle	DISTANCE	Distancia de la referencia
RD DIV DESC	División de la calle	FEET MILES FLAG DESC	Descripción de medida de distancia (pies o millas)
RD DIV CODE	Código de división de la calle	FEET MILES FLAG	Código de de medida de distancia
FIX OBJ DESC	Objetos fijos que intervinieron en el accidente (barreras, borde de la calle, cercas, étc).	DISTANCE DIR FLAG	Dirección de la referencia al punto donde ocurrió el choque (este, oeste, norte, sur)
FIX OBJ CODE	Código de objetos fijos	REFERENCE NO	Id del punto de referencia
REPORT NO	Número de reporte del accidente	REFERENCE TYPE CODE	Código del tipo de referencia (no hay diccionario de esta codiguera)
REPORT TYPE	Tipo de accidente: solo daño de propiedad, heridos o muertos.	REFERENCE SUFFIX	Sufijo de la referencia
WEATHER DESC	Descripción del clima	LATITUDE	Latitud del lugar donde ocurrió el accidente
WEATHER CODE	Código de clima	LONGITUDE	Longitud del lugar donde ocurrió el accidente
ACC DATE	Fecha del accidente	LOCATION	Combinancion de latitude y longitude

2.1.2) DF2: Maryland Statewide Vehicle Crashes Person Details Anonymized

Presentamos la segunda base que tiene por nombre *Maryland Statewide Vehicle Crashes Person Details Anonymized*, la cual llamaremos a partir de ahora DF2. Presenta 1,728,652 filas y 48 columnas. Presentamos un resumen de DF2 en la siguiente tabla.

Tabla 1.2 - Variables y descripción de DF2

Variables	Descripción	DRUG TEST DESC	Descripción de aplicación del test de consumo de drogas
SEX DESC	Sexo	DRUG TEST CODE	Código de aplicación del test de consumo de drogas
SEX CODE	Código de sexo	DRUG TESTRESULT DESC	Descripción del resultado del test de consumo de drogas
CONDITION DESC	Condición de la persona (normal, tomó alcohol, presencia de drogas, étc.)	DRUG TESTRESULT CODE	Código del resultado del test de consumo de drogas
CONDITION CODE	Código de condicion de la persona	BAC CODE	Código BAC
INJ SEVER DESC	Tipo de herida de la persona (ninguna, leve, fatal, étc.)	FAULT FLAG DESC	Indicación de si la persona tuvo la culpa o no por el accidente.
INJ SEVER CODE	Código de tipo de herida	FAULT FLAG	Código de si la persona tuvo la culpa o no por el accidente.
REPORT NO	Número de reporte del accidente	EQUIP PROB DESC	Descripción de problemas de equipamiento de seguridad (no tenía abrochado el cinturón, no funcionó el airbag, étc.)
OCC SEAT POS DESC	Descripción de donde estaba sentada la persona	EQUIP PROB CODE	Código de problemas de equipamiento de seguridad
OCC SEAT POS CODE	Código de asiento de la persona	SAF EQUIP DESC	Equipamiento de seguridad usado por la persona (cinturón, silla para niños, étc.)
PED VISIBLE DESC	Visibilidad del peatón	SAF EQUIP CODE	Código de equipamiento de seguridad usado por la persona (cinturón, silla para niños, étc.)
PED VISIBLE CODE	Código de visibilidad del peatón	EJECT DESC	Descripción de expulsión de la persona del auto, si corresponde
PED OBEY DESC	Obediencia del peatón a las señales de tránsito	EJECT CODE	Código de expulsión de la persona del auto
PED OBEY CODE	Código de obediencia del peatón	DATE OF BIRTH	Fecha de nacimiento
PED TYPE DESC	Tipo de peatón	PERSON ID	Código único de identificador de la persona
PED TYPE CODE	Código de tipo de peatón	LICENSE STATE CODE	Código del estado donde obtuvo la licencia
PED LOCATION CODE	Código de ubicación del peatón	CLASS	Clase de la licencia
MOVEMENT DESC	Movimiento de la persona	CDL FLAG DESC	Indicación de si la licencia es o no comercial
MOVEMENT CODE	Código de movimiento de la persona	CDL FLAG	Código de si la licencia es o no comercial
PERSON TYPE DESC	Tipo de persona (peatón, conductor, ocupante)	VEHICLE ID	Id del vehículo al que está asociado la persona
PERSON TYPE	Código de tipo de persona	EMS UNIT LABEL	Etiqueta de la emergencia asociada al reporte
ALCOHOL TEST DESC	Resultado de test de alcoholemia	AIRBAG DEPLOYED	Código de como se activó el airbag
ALCOHOL TEST CODE	Código de resultado de test de alcoholemia	YEAR	Año del accidente
ALCOHOL TESTTYPE DESC	Tipo de test de alcoholemia (orina, aliento, étc.)	Quarter	Cuatrimestre del año
ALCOHOL TESTTYPE CODE	Código de tipo de test de alcoholemia		

Este dataset resulta crucial, ya que contiene la variable dependiente del modelo: *INJ_SEVERE_CODE*. La misma establece el tipo de herida (o no) que la persona sufrió en el accidente.

2.1.3) DF3: Maryland Statewide Vehicle Crashes Vehicle Details

Presentamos la tercera base que tiene por nombre *Maryland Statewide Vehicle Crashes Vehicle Details* (de ahora en más DF3). Presenta 1,438,808 observaciones y 49 variables. Presentamos un resumen de DF3 en la siguiente tabla.

Tabla 1.3 - Variables y descripción de DF3

Variables	Descripción	GOING DIRECTION CODE	Código de dirección de ida del vehículo
HARM EVENT DESC	Descripción de evento de daño(daño a peatón, otro auto, objeto fijo, étc).	BODY TYPE DESC	Tipo de vehículo (vehículo de pasajeros, camión, moto, étc.)
HARM EVENT CODE	Código de evento de daño (daño a peatón, otro auto, objeto fijo, étc).	BODY TYPE CODE	Código de tipo de vehículo
CONTI DIRECTION DESC	Descripción de continuación del movimiento (este, oeste, sur, norte, desconocido)	DRIVERLESS FLAG DESC	Indicación de si es o no un vehículo sin conductor (autónomo o automatizado)
CONTI DIRECTION CODE	Código de continuación del movimiento	DRIVERLESS FLAG	Código de si es o no un vehículo sin conductor (autónomo o automatizado)
DAMAGE DESC	Tipo de daño al vehículo (ninguno, superficial, destruido, étc.)	FIRE FLAG DESC	Indicación de si el auto se prendió o no fuego.
DAMAGE CODE	Código del tipo de daño al vehículo	FIRE FLAG	Código de si el auto se prendió o no fuego.
MOVEMENT DESC	Tipo de movimiento del auto	PARKED FLAG DESC	Indicación de si el auto estaba o no estacionado
MOVEMENT CODE	Código de tipo de movimiento	PARKED FLAG	Código de de si el auto estaba o no estacionado
VIN NO	Número de identificación del vehículo	SPEED LIMIT	Límite de velocidad
REPORT NO	Número de reporte del accidente	HIT AND RUN FLAG DESC	Indicación de si el auto se dio o no a la fuga
CV BODY TYPE DESC	Descripción de tipo de vehículo comercial	HIT AND RUN FLAG	Código de si el auto se dio o no a la fuga
CV BODY TYPE CODE	Código de tipo de vehículo comercial	HAZMAT SPILL FLAG DESC	Indicación de si se derramaron o no sustancias tóxicas en el accidente
VEH YEAR	Año del vehículo	HAZMAT SPILL FLAG	Código de si se derramaron o no sustancias tóxicas en el accidente
VEH MAKE	Marca del vehículo	VEHICLE ID	Código único de identificación del vehículo
VEH MODEL	Modelo del vehículo	TOWED VEHICLE CONFIG DESC	Descripción del remolque del vehículo (en caso de que tenga)
COMMERCIAL FLAG$_DESC$	Indicación de si se trata o no de un vehículo comercial	TOWED VEHICLE CONFIG CODE	Código del remolque del vehículo (en caso de que tenga)
COMMERCIAL FLAG	Código de si se trata o no de un vehículo comercial	AREA DAMAGED CODE IMP1	Código del area dañada del vehículo en primer impacto
HZM NUM	Número identificador de sustancias peligrosas	AREA DAMAGED CODE1	Código de área dañada 1 del vehículo
TOWED AWAY FLAG$_DESC$	Remolcado (sí o no)	AREA DAMAGED CODE2	Código de área dañada 2 del vehículo
TOWED AWAY FLAG	Código de Remolcado (Sí o no)	AREA DAMAGED CODE3	Código de área dañada 3 del vehículo
NUM AXLES	Número de ejes	AREA DAMAGED CODE MAIN DESC	Descripción del área más dañada del vehículo
GVW DESC	Descripción de masa máxima autorizada del vehículo	AREA DAMAGED CODE MAIN	Código de área más dañada del vehículo
GVW CODE	Código de masa máxima autorizada del vehículo	YEAR	Año del accidente
GOING DIRECTION DES	C Dirección de ida del vehículo (este, oeste, sur, norte, desconocido)	Quarter	Cuatrimestre del año

2.2) Unión de bases DF1, DF2 y DF3

Como podemos observar en las tres tablas analizadas, presentan varias columnas con el mismo nombre. Una de esas es *REPORT_NO*, la cual es el número de reporte del accidente,

es decir que es un id que identifica en forma única a cada accidente registrado. Esta variable está presente en todos los datasets a efectos de poder unir todos los datos relativos a un mismo accidente.

Como lo que queremos unir son los datos generales del accidente (DF1), con los datos de las personas y vehículos que intervinieron en el mismo (DF2 y DF3 respectivamente), y como una persona puede haber estado en un solo vehículo en el accidente, es que primero se une DF2 y DF3 por las variables *REPORT_NO* y *VEHICLE_ID* (variable que identifica en forma unívoca a cada vehículo del accidente, y que está presente tanto en DF2 como DF3). De esta unión obtenemos un dataset con los datos de los vehículos de cada accidente y las características de las personas que estaban en los mismos. Esto resulta crucial, ya que las características que tiene el auto pueden influir en el hecho de que la persona haya resultado herida o no.

A este dataset unido, se le aplica posteriormente un merge con el DF1 usando la variable *REPORT_NO,* generando así una nueva gran base de datos llamada *DATA*, con dimensión 1,728,805 x 150. De esta forma obtenemos un dataset con datos específicos del accidente, de los vehículos que intervinieron en el mismo, y de las personas involucradas. Cada observación es una persona que estuvo presente en un determinado accidente, y que pudo haber resultado herida o no en el mismo (señalado en la variable *INJ_SEVERE_CODE).*

2.3) Validación de columnas con registro Null o NaN

Con el objetivo de reducir la cantidad de variables en la presente base de datos, así como para mejorar su calidad, presentamos el conteo de los valores Null o NaN de cada columna, donde tendremos como criterio para eliminar o no considerar las variables que presenten más de un 10% de esos valores en toda la base de datos.

2.3.1) Análisis de nulos

Se eliminan todas aquellas variables que tienen aproximadamente más de 172,000 registros nulos (ya que el dataset tiene una dimensión de 1,728,805). No se van a tomar en cuenta ya que sólo añadirían ruido a los modelos, y si se imputan se crearía un sesgo al redimensionar variables con registros específicos.

Con las restantes columnas se va a proceder a utilizar un método de imputación dirigido en Python. Como la mayoría de las variables son categóricas se decide imputar utilizando la moda.

2.4) Limpieza de variables

Además de las variables que se borraron por tener un número elevado de valores missing o NaN, eliminaremos otras variables porque son descripciones de códigos presentes en otras variables. Como la información ya está contenida en dichos códigos, y además se cuenta con un diccionario de los mismos, no tiene sentido duplicar la información.

Por otra parte, también se eliminan variables duplicadas que se generan con los joins, debido a que los tres datasets que se unieron repiten los mismos datos de algunas variables.

2.5) Transformación de variables categóricas a numéricas

Para transformar las variables categóricas a numéricas hacemos uso del paquete *Label Encoder* de *Scikit-Learn*, aunque para la ejecución de algunos modelos (como el probit y probit espacial) se usa una dummy por cada categoría menos una (la más frecuente). De todas maneras esto será explicado en mayor detalle al analizar dichos modelos.

2.6) Imputación de valores nulos utilizando método de la moda

Al ser variables categóricas hay que imputar usando algún método de imputación que utilice las más frecuentes, para este caso y por la cantidad de valores que se deben imputar vamos a utilizar la moda.

Si se tuviese un cluster dedicado de procesamiento o un servidor se hubiese utilizado el KNNImputer.

2.7) Dataset final

Luego de realizada la limpieza mencionada, queda un dataset ya listo para ser utilizado para modelar. Este dataset contiene 1.054.896 observaciones y 24 variables, las cuales se detallan a continuación:

Tabla 1.4 - Descripción de las variables

Variable	Descripción
LIGHT_CODE	Características de la luz del día
JUNCTION_CODE	Tipo de cruce
COLLISION_TYPE_CODE	Tipo de choque
SURF_COND_CODE	Condiciones de la superficie de la calle (nieve, húmedo, agua,étc)
RD_COND_CODE	Condiciones de la calle
WEATHER_CODE	Condiciones del clima
SIGNAL_FLAG	Presencia o no de señal de tránsito
SEX_CODE	Sexo
EQUIP_PROB_CODE	Problemas de equipamiento de seguridad (no tenía abrochado el cinturón, no funcionó el airbag, étc.)
SAF_EQUIP_CODE	Equipamiento de seguridad usado por la persona (cinturón, silla para niños, étc.)
DAMAGE_CODE	Tipo de daño al vehículo (ninguno, superficial, destruido, étc.)
MOVEMENT_CODE	Tipo de movimiento del auto
COMMERCIAL_FLAG	Indicación de si se trata o no de un vehículo comercial
BODY_TYPE_CODE	Tipo de vehículo (vehículo de pasajeros, camión, moto, étc.)
DRIVERLESS_FLAG	Indicación de si es o no un vehículo sin conductor (autónomo o automatizado)
SPEED_LIMIT	Límite de velocidad
LATITUDE	Latitud del lugar donde ocurrió el accidente
LONGITUDE	Longitud del lugar donde ocurrió el accidente
HOUR	Hora del día del accidente
YEAR	Año del accidente
MONTH	Mes del accidente
DAY	Día del accidente
INJ_SEVER_CODE	Tipo de herida sufrida por la persona

La última variable es *INJ_SEVER_CODE,* que es la variable a predecir. Esta variable venía con 5 códigos:

- 01 - sin heridas
- 02 - herida leve no discapacitante
- 03 - herida potencialmente discapacitante
- 04 - herida discapacitante
- 05 - muerte

Como el objetivo es predecir si hubo heridos o no (incluyendo dentro del concepto de heridos también a los muertos) es que se decide transformar esta variable a binaria, adquiriendo el valor 0 en caso de que la persona no resulte herida (código 01 de la variable anterior), y 1 en caso de que si haya sufrido heridas (códigos 02 a 05). Por lo tanto, nuestra variable "positiva" será el hecho de que la persona resultó herida.

Por otra parte, cabe destacar que se elimina del dataset los años 2020 y 2021 por considerarlos outliers, ya que debido a la pandemia la circulación bajó drásticamente en el 2020, y no se recuperó del todo en el año 2021. También, debido a que son registros más antiguos y para disminuir los tiempos de procesamiento, se eliminan los datos de 2015.

Por último se divide el dataset en train y test, con una proporción de 80% y 20% respectivamente.

En el siguiente capítulo, se realiza el análisis exploratorio de los datos sobre el dataset final definitivo, que será usado para el posterior modelamiento.

Capítulo 3

Análisis exploratorio de los datos

En el presente capítulo se presenta la exploración y análisis del conjunto de datos, aplicado al dataset final, que fue definido luego de realizada la limpieza y depuración del dataset inicial.

3.1) Análisis exploratorio de las variables del dataset

Luego de detallar el listado de variables incluidas en nuestro dataset, como se presentó en el capítulo previo, se procede a presentar un resumen de cada una de las variables:

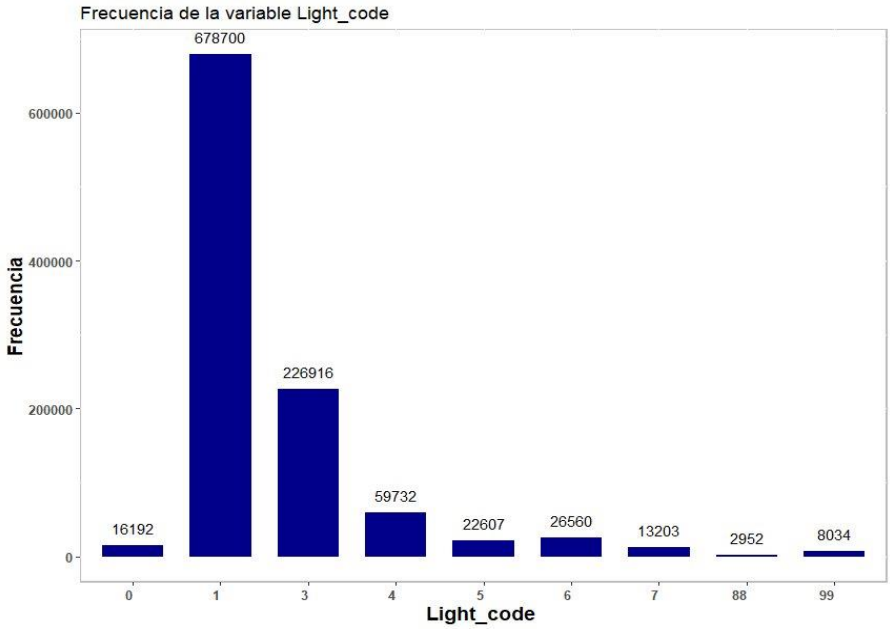
- **Light_code**: Es una variable categórica con 9 categorías, como se indica en la Tabla.

Tabla 1.5 - Frecuencia absoluta y relativa de la variable Light_code

LIGHT_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	16192	0,02
1	Daylight	678700	0,64
3	Dark Lights On	226916	0,22
4	Dark No Lights	59732	0,06
5	Dawn	22607	0,02
6	Dusk	26560	0,03
7	Dark - Unknown Lighting	13203	0,01
88	Other	2952	0,00
99	Unknown	8034	0,01
Total		1054896	1

Se puede observar, al analizar la frecuencia, que se mantiene una gran concentración en la categoría 1, “Daylight”, con 678,700 observaciones, indicando de esta forma que la mayor parte de los accidentes, es decir un 64%, ocurren a la luz del día. Se presenta también el gráfico a continuación para visualizar la frecuencia de la variable.

Gráfico 1.1 - Frecuencia de la variable Light_code



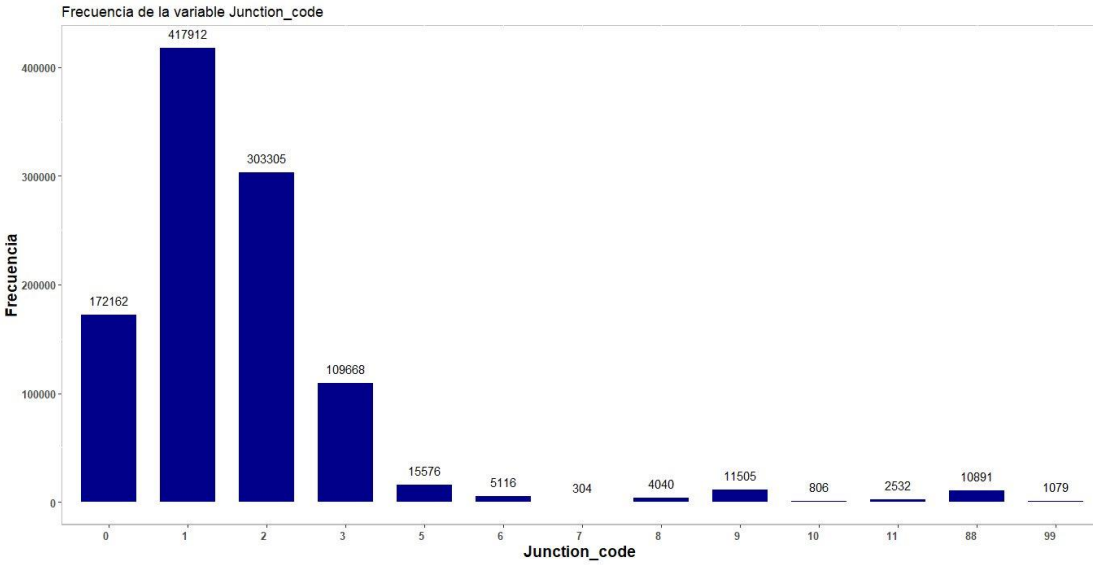
- **Junction_code**: Es una variable categórica con 13 categorías, como se indica en la Tabla.

Tabla 1.6- Frecuencia absoluta y relativa de la variable Junction_code

JUNCTION_CODE	Descripción	Frecuencia	Frecuencia relativa
00	Not Applicable	172162	0,16
01	Non Intersection	417912	0,40
02	Intersection	303305	0,29
03	Intersection Related	109668	0,10
04	Driveway Alley Access Related	15576	0,01
5	Interchange Related	5116	0,00
6	Crossover Related	304	0,00
7	Railway Grade Crossing	4040	0,00
8	Residential Driveway	11505	0,01
9	Commercial Driveway	806	0,00
10	Alley	2532	0,00
88	Other	10891	0,01
99	Unknown	1079	0,00
Total		1054896	1

Se puede ver que se da una gran concentración de observaciones en la categoría ”Non Intersection”, indicando de esta forma que la mayor parte de los accidentes (40%), ocurren en sitios donde no hay intersecciones. Se agrega el gráfico a continuación para visualizarlo mejor.

Gráfico 1.2 - Frecuencia de la variable Junction_code



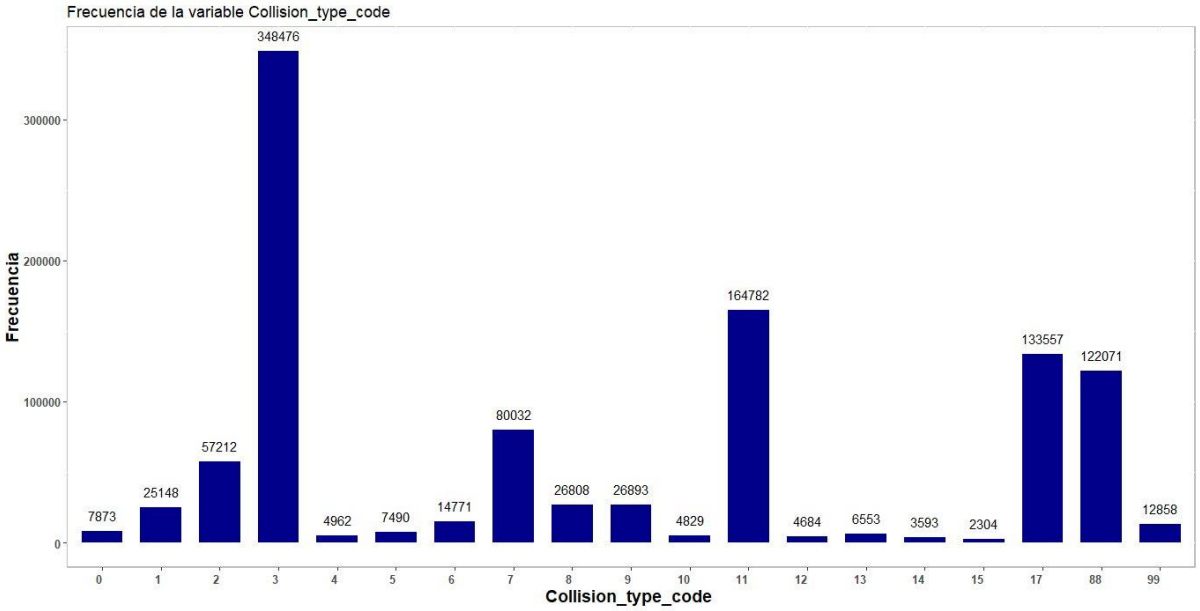
- **Collision_type_code**: Es una variable de origen categórica, con 13 categorías, que indica la posición en que quedó el vehículo luego de ocasionado el accidente. Se presenta tabla de frecuencias.

Tabla 1.7 - Frecuencia absoluta y relativa de la variable Collision_type_code

COLLISION_TYPE_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	7873	0,01
1	Head On	25148	0,02
2	Head On Left Turn	57212	0,05
3	Same Direction Rear End	348476	0,33
4	Same Direction Rear End Right Turn	4962	0,00
5	Same Direction Rear End Left Turn	7490	0,01
6	Opposite Direction Sideswipe	14771	0,01
7	Same Direction Sideswipe	80032	0,08
8	Same Direction Right Turn	26808	0,03
9	Same Direction Left Turn	26893	0,03
10	Same Direction Both Left Turn	4829	0,00
11	Same Movement Angle	164782	0,16
12	Angle Meets Right Turn	4684	0,00
13	Angle Meets Left Turn	6553	0,01
14	Angle Meets Left Turn Head On	3593	0,00
15	Opposite Direction Both Left Turn	2304	0,00
17	Single Vehicle	133557	0,13
88	Other	122071	0,12
99	Unknown	12858	0,01
Total		1054896	1

Se observa que en la mayor parte de los siniestros, en un 33%, el vehículo quedó en posición donde el "Extremo trasero del vehículo terminó ubicado en la misma dirección en que se dirigía". El gráfico a continuación puede mostrar mejor el peso de la categoría 3.

Gráfico 1.3 - Frecuencia de la variable Collision_type_code



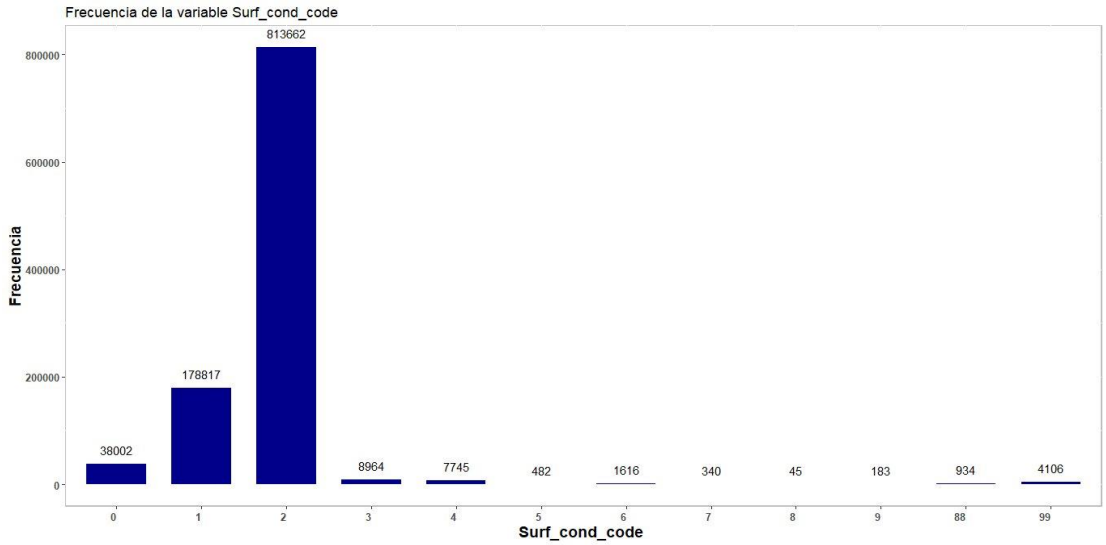
- **Surf_cond_code**: Esta variable es una variable categórica con 12 categorías, como se indica en la Tabla.

Tabla 1.8 - Frecuencia absoluta y relativa de la variable Surf_cond_code

SURF_COND_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	38002	0,04
1	Wet	178817	0,17
2	Dry	813662	0,77
3	Snow	8964	0,01
4	Ice	7745	0,01
5	Mud, Dirt, Gravel	482	0,00
6	Slush	1616	0,00
7	Water (standing/moving)	340	0,00
8	Sand	45	0,00
9	Oil	183	0,00
88	Other	934	0,00
99	Unknown	4106	0,00
Total		1054896	1,00

Se observa que en el caso de esta variable, se mantiene la concentración en la categoría 2, "Seco", con 813,662 observaciones, indicando de esta forma que la gran mayoría de los accidentes, es decir un 77%, ocurrieron con una superficie totalmente seca. Se muestra gráfico para verlo visualmente.

Gráfico 1.4 - Frecuencia de la variable Surf_cond_code



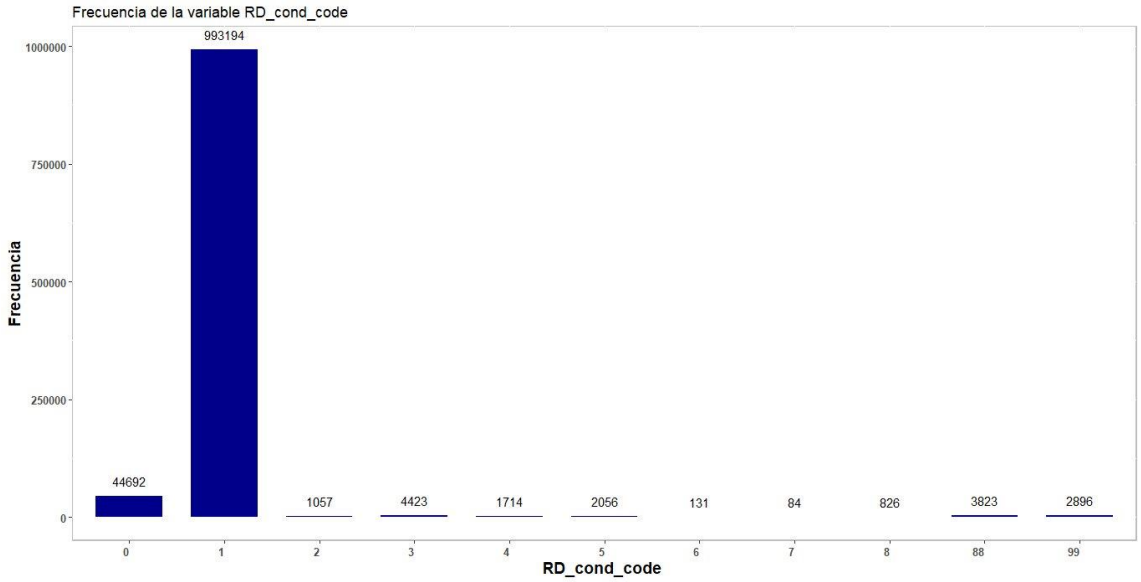
- **RD_cond_code**: Esta variable es una variable de origen categórico con 10 categorías, que indica las condiciones de la calle al momento del accidente.

Tabla 1.9- Frecuencia absoluta y relativa de la variable RD_cond_code

RD_COND_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	44692	0,04
1	No Defects	993194	0,94
2	Shoulder Defect	1057	0,00
3	Holes, Ruts, Etc.	4423	0,00
4	Foreign Material	1714	0,00
5	Loose Surface Material	2056	0,00
6	Obstruction Not Lighted	131	0,00
7	Obstruction Not Signaled	84	0,00
8	View Obstructed	826	0,00
88	Other	3823	0,00
99	Unknown	2896	0,00
Total		1054896	1,00

Se puede ver que se da una gran concentración de observaciones en la categoría ”No Defects”, lo que significa que al momento de ocurrir el accidente de tránsito, la calle donde se dirigía el vehículo, en el 94% de los casos, no presenta ningún defecto. Se agrega el gráfico a continuación para visualizarlo mejor.

Gráfico 1.5 - Frecuencia de la variable RD_cond_code



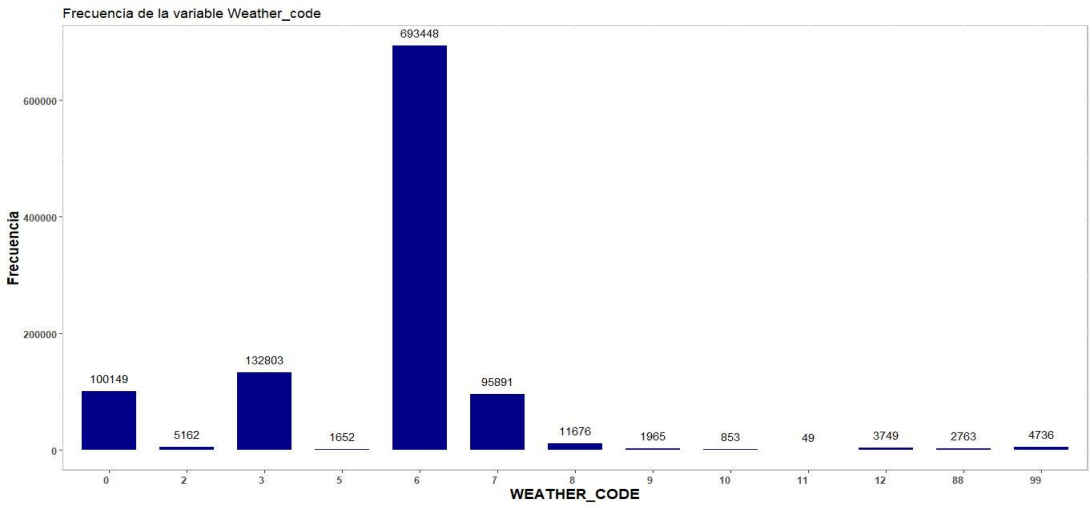
- **Weather_code**: Esta variable indica las condiciones climáticas al momento del accidente. Es una variable que presenta 13 categorías.

Tabla 1.10- Frecuencia absoluta y relativa de la variable Weather_code

WEATHER_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	100149	0,09
2	Foggy	5162	0,00
3	Raining	132803	0,13
5	Severe Winds	1652	0,00
6	Clear	693448	0,66
7	Cloudy	95891	0,09
8	Snow	11676	0,01
9	Sleet	1965	0,00
10	Blowing Snow	853	0,00
11	Blowing Sand, Soil, Dirt	49	0,00
12	Wintry Mix	3749	0,00
88	Other	2763	0,00
99	Unknown	4736	0,00
Total		1054896	1,00

Como se puede observar, en el 66% de los accidentes ocurridos en el periodo analizado, el clima al momento del siniestro se encontraba despejado.

Gráfico 1.6 - Frecuencia de la variable Weather_code



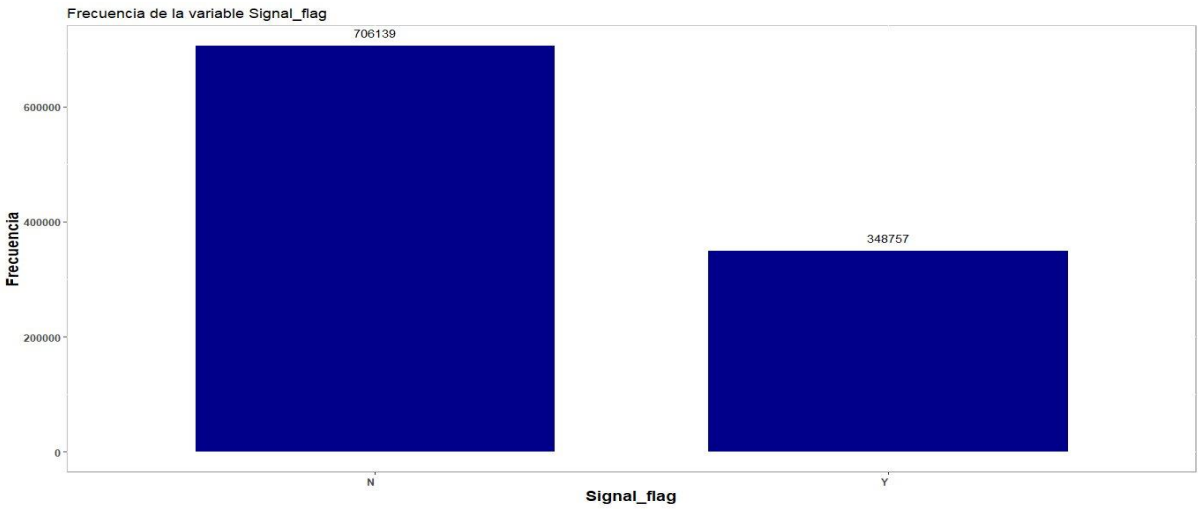
- **Signal_flag**: Esta variable es categórica y presenta dos categorías, “Yes” si había una señal de tránsito en el lugar donde ocurrió el accidente y “No” si no había.

Tabla 1.11- Frecuencia absoluta y relativa de la variable Signal_flag

SIGNAL_FLAG	Descripción	Frecuencia	Frecuencia relativa
N	No	706139	0,67
Y	Yes	348757	0,33
Total		1054896	1,00

Como se desprende de la tabla, en el 67% de los accidentes no existe la presencia de una señal de tránsito en el lugar y momento de ocurrido el mismo.

Gráfico 1.7 - Frecuencia de la variable Signal_flag



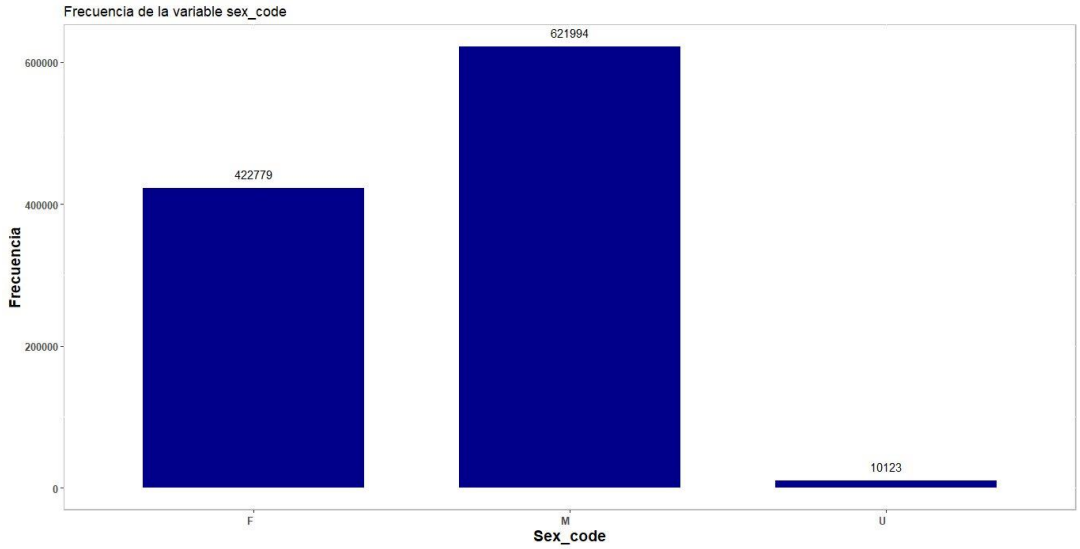
- **Sex_code**: Esta variable hace referencia al sexo de las personas involucradas en el accidente de tránsito. Es una variable también categórica con 3 categorías: F: Femenino, M: Masculino y U: Desconocido.

Tabla 1.12- Frecuencia absoluta y relativa de la variable Sex_code

SEX_CODE	Descripción	Frecuencia	Frecuencia relativa
F	Female	422779	0,40
M	Male	621994	0,59
U	Unknown	10123	0,01
Total		1054896	1,00

La mayoría de las personas involucradas en accidentes son de sexo masculino, representando un 59% del total de observaciones.

Gráfico 1.8 - Frecuencia de la variable Sex_code



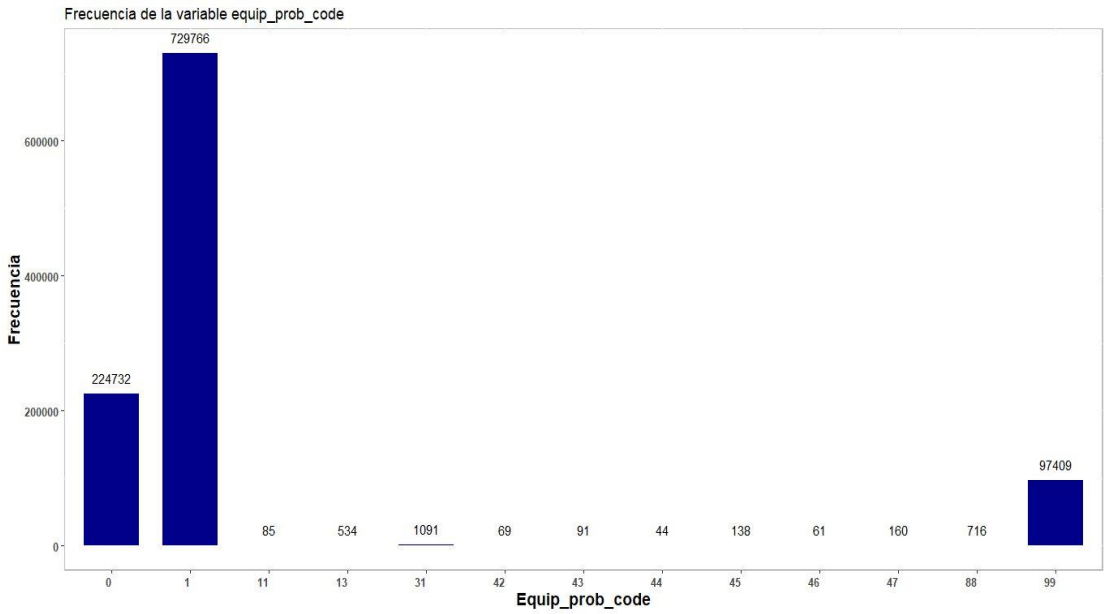
- **Equip_prob_code**: Esta variable hace referencia a los problemas de equipamiento de seguridad que tenía el vehículo al momento del accidente, como pueden ser “no tener abrochado el cinturón de seguridad”, “incorrecto uso del cinturón”, etc. Es una variable de origen categórico, con 13 categorías presentadas a continuación:

Tabla 1.13- Frecuencia absoluta y relativa de la variable Equip_prob_code

EQUIP_PROB_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	224732	0,21
1	No Misuse	729766	0,69
11	Belts/Anchors Broken	85	0,00
13	Belt(s) Misused	534	0,00
31	Air Bag Failed	1091	0,00
42	Facing Wrong Way	69	0,00
43	Not Anchored Right	91	0,00
44	Anchor Not Secure	44	0,00
45	Not Strapped Right	138	0,00
46	Strap/Tether Loose	61	0,00
47	Size/Type Improper	160	0,00
88	Other	716	0,00
99	Unknown	97409	0,09
Total		1054896	1,00

Como se puede ver, en la mayoría de los casos (69%), no hubo un mal uso en el equipamiento de seguridad. Se presenta el gráfico:

Gráfico 1.9 - Frecuencia de la variable equip_prob_code



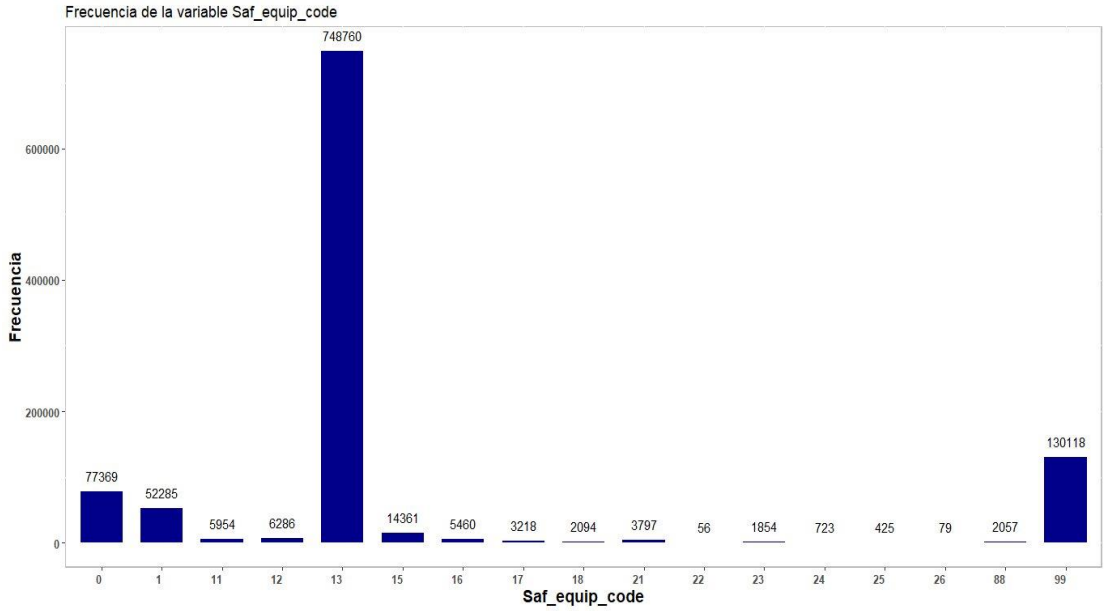
- **Saf_equip_code**: Esta variable hace referencia al equipamiento de seguridad utilizado por el individuo al momento del accidente. Es una variable categórica, que se presenta a continuación sus categorías y tabla de frecuencias:

Tabla 1.14- Frecuencia absoluta y relativa de la variable Saf_equip_code

SAF_EQUIP_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	77369	0,07
1	None	52285	0,05
11	Lap Belt Only	5954	0,01
12	Shoulder Belt Only	6286	0,01
13	Shoulder/Lap Belt(s)	748760	0,71
15	Child Restraint System Forward Facing	14361	0,01
16	Child Restraint System Rear Facing	5460	0,01
17	Booster Seat	3218	0,00
18	Child Restraint Type Unknown	2094	0,00
21	MC/Bike Helmet	3797	0,00
22	MC/Bike Eye Protection Only	56	0,00
23	MC/Bike Helmet and Eye Protection	1854	0,00
24	Protective Pads	723	0,00
25	Reflective Clothing	425	0,00
26	Lighting	79	0,00
88	Other	2057	0,00
99	Unknown	130118	0,12
Total		1054896	1,00

Como se puede observar en los datos, en la mayoría de los casos (71%), la persona involucrada en el accidente presentaba el cinturón de seguridad sujetando tanto el hombro como la cintura. Visualizamos esto en el gráfico:

Gráfico 1.10 - Frecuencia de la variable Saf_equip_code



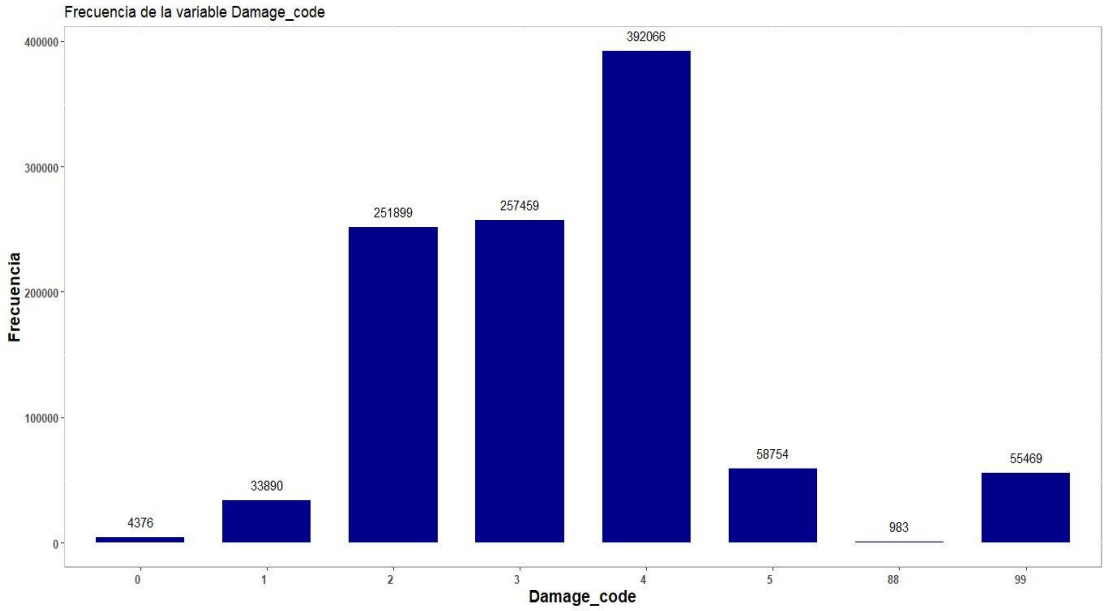
- **Damage_code**: Esta variable hace referencia al tipo de daño que sufre el vehículo, es decir, cómo queda el mismo luego del accidente(“ningún daño”, “daño superficial”, etc), siendo una variable categórica, con 8 categorías que se describen a continuación:

Tabla 1.15- Frecuencia absoluta y relativa de la variable Damage_code

DAMAGE_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	4376	0,00
1	No Damage	33890	0,03
2	Superficial	251899	0,24
3	Functional	257459	0,24
4	Disabling	392066	0,37
5	Destroyed	58754	0,06
88	Other	983	0,00
99	Unknown	55469	0,05
Total		1054896	1,00

En la mayoría de los casos, un 37%, el vehículo quedó inhabilitado luego del accidente. Se presenta el gráfico:

Gráfico 1.11 - Frecuencia de la variable Damage_code



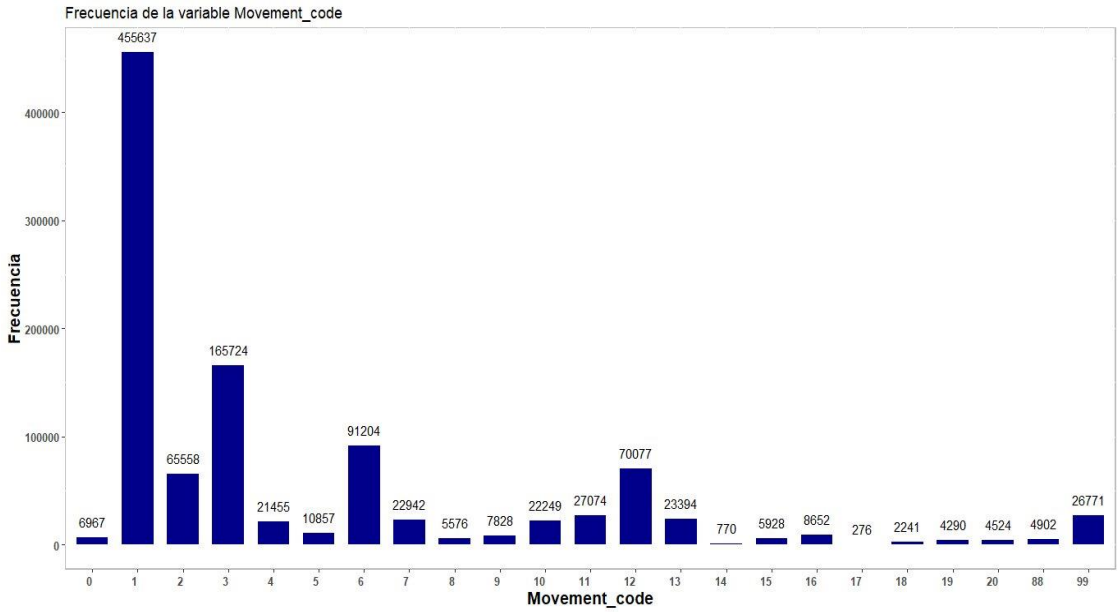
- **Movement_code**: Esta variable hace referencia al movimiento que estaba realizando el vehículo al momento del accidente, como por ejemplo “estacionando”, “doblando en una curva”, etc. Es una variable de origen categórico, con 23 categoría:

Tabla 1.16- Frecuencia absoluta y relativa de la variable Movement_code

MOVEMENT_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	6967	0,01
1	Moving Constant Speed	455637	0,43
2	Accelerating	65558	0,06
3	Slowing or Stopping	165724	0,16
4	Starting From Lane	21455	0,02
5	Starting From Parked	10857	0,01
6	Stopped in Traffic Lane	91204	0,09
7	Changing Lanes	22942	0,02
8	Passing	5576	0,01
9	Parking	7828	0,01
10	Parked	22249	0,02
11	Backing	27074	0,03
12	Making Left Turn	70077	0,07
13	Making Right Turn	23394	0,02
14	Right Turn on Red	770	0,00
15	Making U Turn	5928	0,01
16	Skidding	8652	0,01
17	Driverless Moving Vehicle	276	0,00
18	Leaving Traffic Lane	2241	0,00
19	Entering Traffic Lane	4290	0,00
20	Negotiating a Curve	4524	0,00
88	Other	4902	0,00
99	Unknown	26771	0,03
Total		1054896	1,00

En la mayoría de las observaciones, el vehículo venía moviéndose a una velocidad constante previo a producirse el accidente de tránsito. Esto sucedió en un 43% de los siniestros. Se presenta el gráfico:

Gráfico 1.12 - Frecuencia de la variable Movement_code



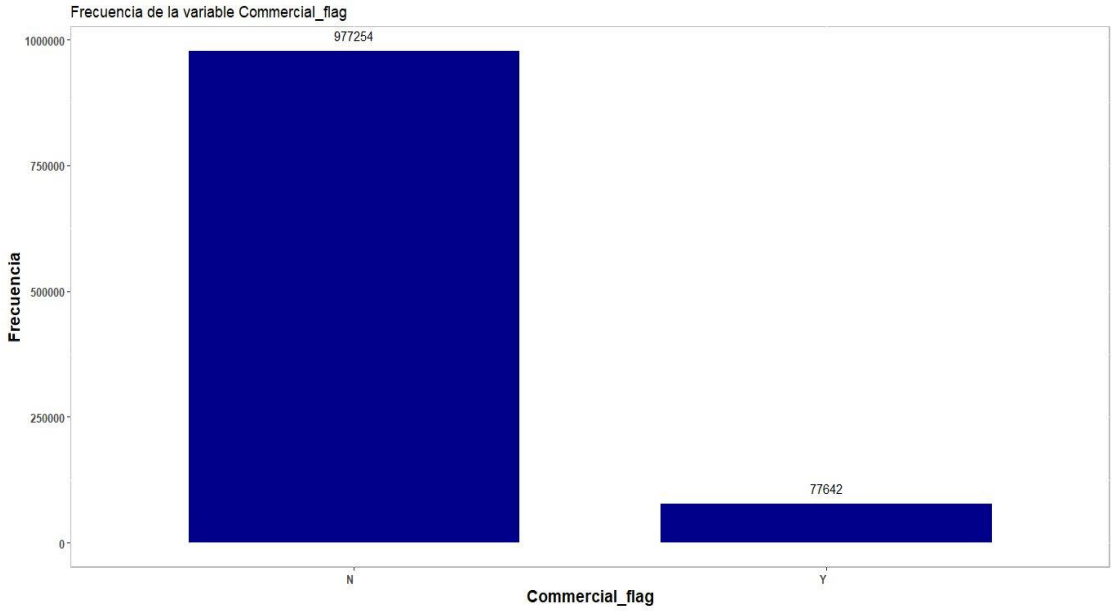
- **Commercial_flag**: Esta variable hace referencia a si el vehículo involucrado se trata de un vehículo comercial, tomando valor de Si o No en caso que no lo sea. Se visualiza la frecuencia:

Tabla 1.17 - Frecuencia absoluta y relativa de la variable Commercial_flag

COMMERCIAL_FLAG	Descripción	Frecuencia	Frecuencia relativa
N	No	977254	0,93
Y	Yes	77642	0,07
Total		1054896	1,00

En el 93% de los casos, los vehículos involucrados en accidentes no eran comerciales.

Gráfico 1.13 - Frecuencia de la variable Commercial_flag



- **Body_type_code**: Esta variable, de origen también categórico, refiere al tipo de vehículo involucrado en el accidente de tránsito. Se presentan las 30 categorías en la siguiente tabla de frecuencias:

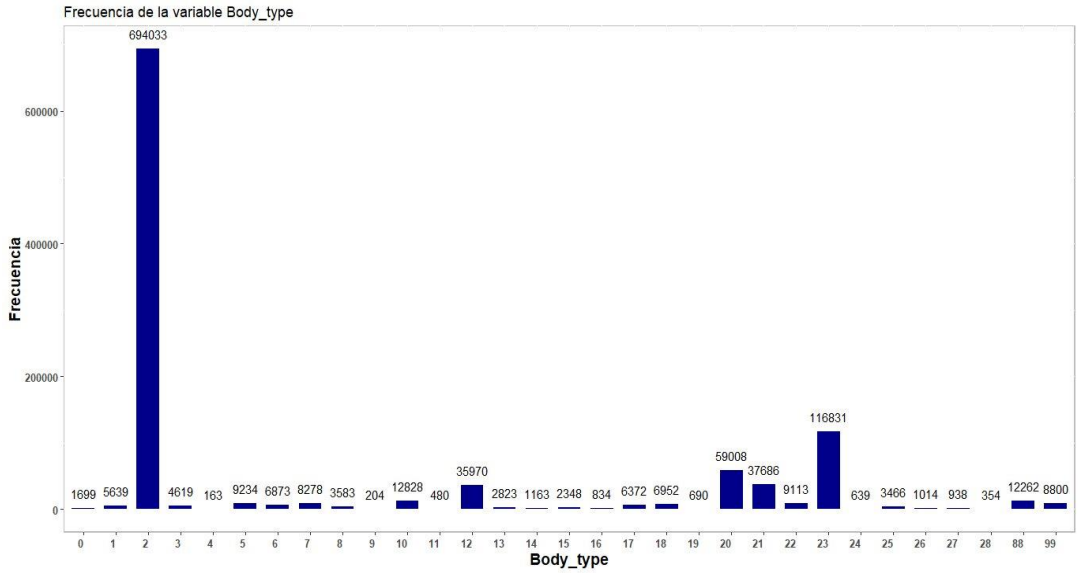
Tabla 1.18- Frecuencia absoluta y relativa de la variable Body_type_code

BODY_TYPE_CODE	Descripción	Frecuencia	Frecuencia relativa
0	Not Applicable	1699	0,00
1	Motorcycle	5639	0,01
2	Passenger Car	694033	0,66
3	Station Wagon	4619	0,00
4	Limousine	163	0,00
5	Cargo Van/Light Truck 2 axles (10,000 lbs (4,536 kg) or less)	9234	0,01
6	Medium/Heavy Truck 2 axles (10,000 lbs (4,536 kg) or less)	6873	0,01
7	Truck Tractor	8278	0,01
8	Recreational Vehicle	3583	0,00
9	Farm Vehicle	204	0,00
10	Transit Bus	12828	0,01
11	Cross Country Bus	480	0,00
12	School Bus	35970	0,03
13	Ambulance/Emergency	2823	0,00
14	Ambulance/Non Emergency	1163	0,00
15	Fire Vehicle/Emergency	2348	0,00
16	Fire Vehicle/Non Emergency	834	0,00
17	Police Vehicle/Emergency	6372	0,01
18	Police Vehicle/Non Emergency	6952	0,01
19	Moped	690	0,00
20	Pickup Truck	59008	0,06
21	Van	37686	0,04
22	Other Light Trucks (10,000 lbs (4,536 kg))	9113	0,01
23	(Sport) Utility Vehicle	116831	0,11
24	Low Speed Vehicle	639	0,00
25	Other Bus	3466	0,00
26	All Terrain Vehicle (ATV)	1014	0,00
27	Snowmobile	938	0,00
88	Other	12616	0,01
99	Unknown	8800	0,01

Total		1054896	1,00

Observando la frecuencia de esta variable, se presenta que en la mayoría de los casos, un 66%, el vehículo involucrado fue un automóvil de pasajeros normal. Visualizando de forma gráfica:

Gráfico 1.14 - Frecuencia de la variable Body_type_code



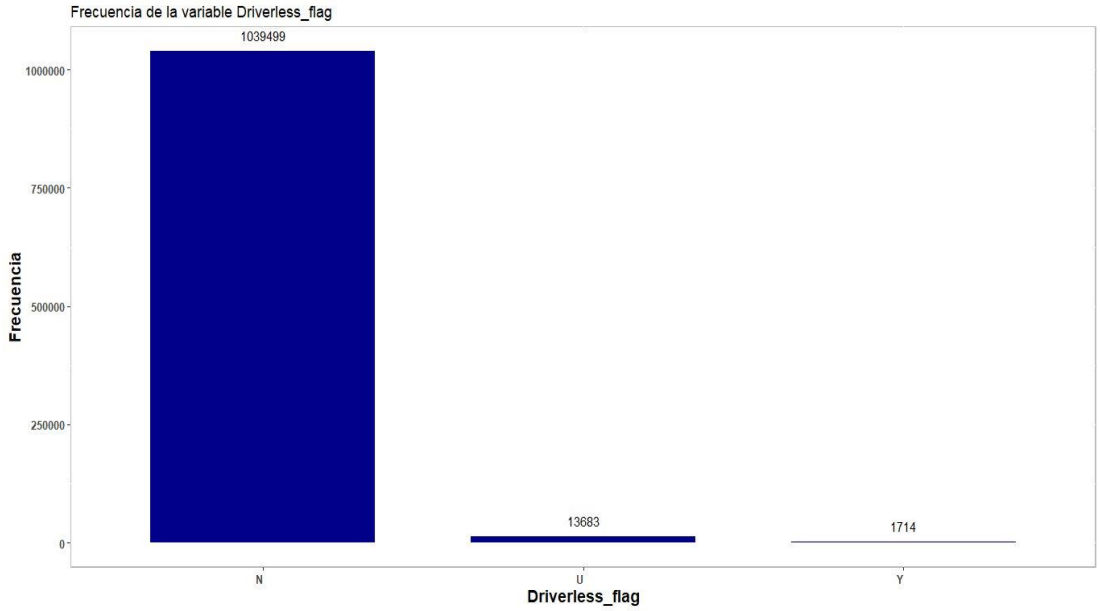
- **Driverless_flag**: Esta variable hace referencia a si el vehículo involucrado, es un vehículo sin conductor, es decir un vehículo autónomo. Es una variable de origen categórico, con 3 categorías siendo las categorías: Si, No o Se desconoce, presentadas a continuación con sus respectivas frecuencias:

Tabla 1.19- Frecuencia absoluta y relativa de la variable Driverless_flag

DRIVERLESS_FLAG	Descripción	Frecuencia	Frecuencia relativa
N	No	1039499	0,99
U	Unknown	13683	0,01
Y	Yes	1714	0,00
Total		1054896	1,00

Como se puede observar, y era de esperarse, en el 99% de los accidentes no se vieron involucrados vehículos autónomos.

Gráfico 1.15 - Frecuencia de la variable Driverless_flag



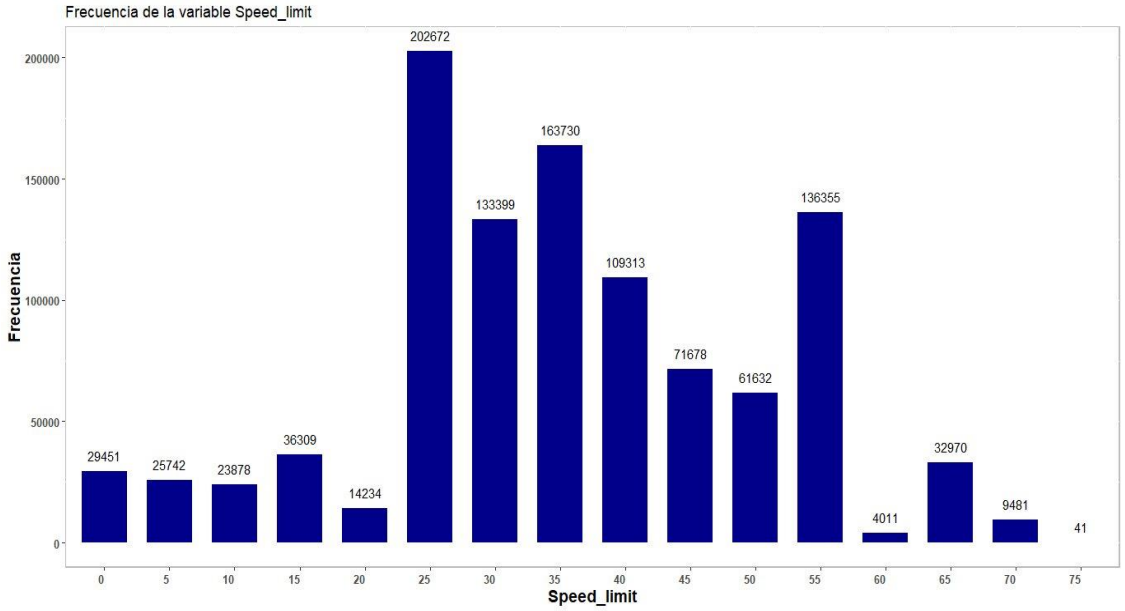
- **Speed_limit**: Esta variable hace referencia al límite de velocidad de la zona por donde circulaba el vehículo, al momento del accidente. Se presenta la tabla de frecuencias de la misma:

Tabla 1.20 - Frecuencia absoluta y relativa de la variable Speed_limit

SPEED_LIMIT	Frecuencia	Frecuencia relativa
0	29451	0,03
5	25742	0,02
10	23878	0,02
15	36309	0,03
20	14234	0,01
25	202672	0,19
30	133399	0,13
35	163730	0,16
40	109313	0,10
45	71678	0,07
50	61632	0,06
55	136355	0,13
60	4011	0,00
65	32970	0,03
70	9481	0,01
75	41	0,00
Total	1054896	1,00

Como se observa de forma clara, la mayor cantidad de accidentes se dan en zonas donde los límites de velocidad se encuentran entre los valores medios, es decir, entre 25 y 55 se dan el 83% de los accidentes (878,779 casos). En zonas con límite de 25 (un 19% de los accidentes) y en segundo lugar en zonas con 35 (un 16% de los accidentes). Luego en tercer lugar vemos que un 13% de los accidentes se da en zonas con un límite más alto, de 55, siguiéndole a este también con un 13% las zonas con límites de velocidad de 30. En los extremos, ya sea zonas con límites de velocidad más bajos, y zonas con límites muy altos, se puede visualizar que se dan la mínima cantidad de accidentes (el 17% restante). Se añade el gráfico para verlo visualmente:

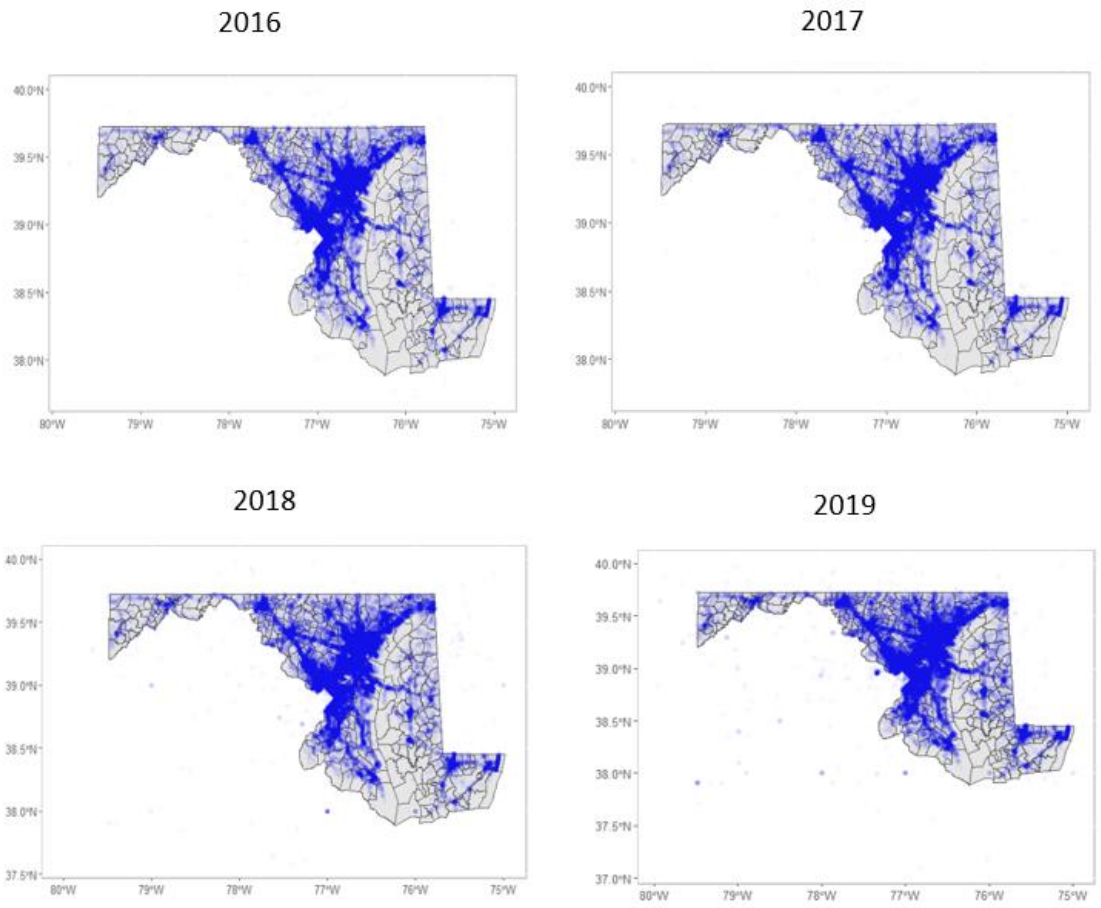
Gráfico 1.16 - Frecuencia de la variable Speed_limit



- **Latitud y Longitud**: Estas variables, refieren exactamente a las coordenadas geográficas que nos permiten ubicar el punto exacto en el mapa en que sucedieron los accidentes de tránsito.

  Como ya fuera mencionado anteriormente, los accidentes de tránsito suelen darse frecuentemente en los mismos puntos (*hotspots*). Por esta razón, se presenta a continuación una secuencia de figuras con el mapa del estado de Maryland, junto con los accidentes ocurridos para cada uno de los años en estudio, detallando las distintas ubicaciones geográficas donde sucedieron los mismos.

Gráfico 1.17 - Distribución geográfica de accidentes en el estado de Maryland 2016-2019



Al comparar los distintos mapas del estado de Maryland a lo largo del tiempo, se puede observar que los accidentes de tránsito parecerían concentrarse en los mismos lugares geográficos, sin notar cambios a lo largo del tiempo en las zonas de ocurrencia. Por lo que la hipótesis de existencia de *hotspots* parece razonable. Por esta razón es que se decide usar un modelo espacial, a efectos de analizar el posible efecto que puede tener la ocurrencia de accidentes en un punto, en la probabilidad de que ocurran otros accidentes en ese mismo lugar. Este modelo se desarrolla en el siguiente capítulo.

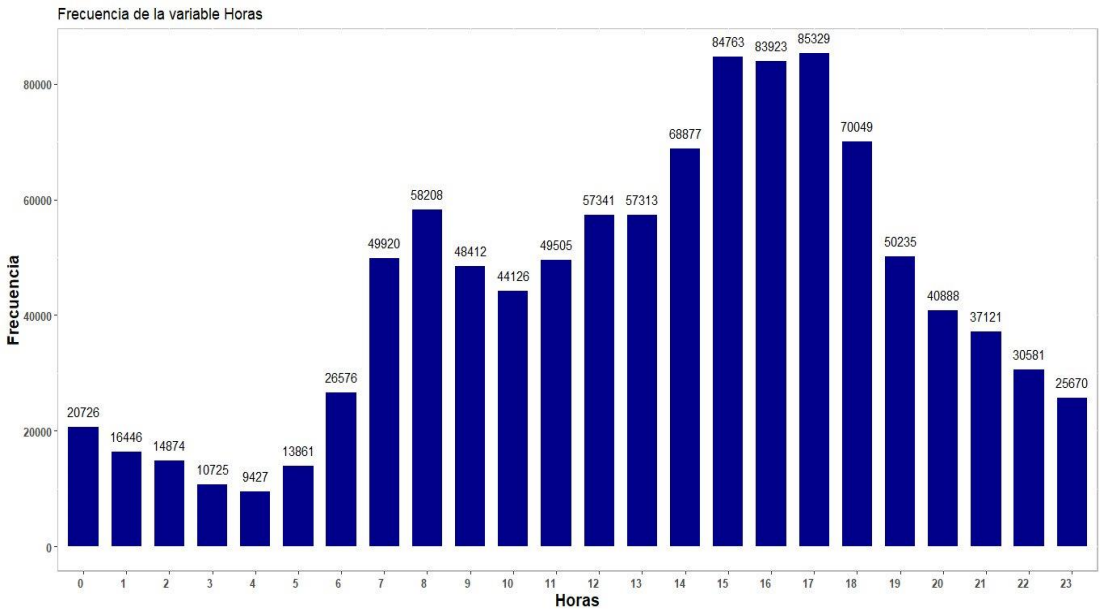
- **Hour**: Luego de realizado un feature engineering para la variable "ACC DATE", se obtuvo la variable "Hour", que detalla la hora del día en que ocurrió el accidente de tránsito. Se divide en 24 horas, y se presenta a continuación su tabla de frecuencias:

Tabla 1.21- Frecuencia absoluta y relativa de la variable Hour

HOUR	Frecuencia	Frecuencia relativa
0	20726	0,02
1	16446	0,02
2	14874	0,01
3	10725	0,01
4	9427	0,01
5	13861	0,01
6	26576	0,03
7	49920	0,05
8	58208	0,06
9	48412	0,05
10	44126	0,04
11	49505	0,05
12	57341	0,05
13	57313	0,05
14	68877	0,07
15	84763	0,08
16	83923	0,08
17	85329	0,08
18	70049	0,07
19	50235	0,05
20	40888	0,04
21	37121	0,04
22	30581	0,03
23	25670	0,02
Total	1054896	1,00

Se observa que la mayor cantidad de accidentes se da en las horas de la tarde, concentrándose la mayor cantidad a las 17 hs, totalizando 85,329 observaciones, seguido de las horas 15, 16 y 18, abarcando en estas 4 horas un 30.72% de los accidentes de tránsito ocurridos en el periodo. Se añade gráfico de la variable:

Gráfico 1.18 - Frecuencia de la variable Hour



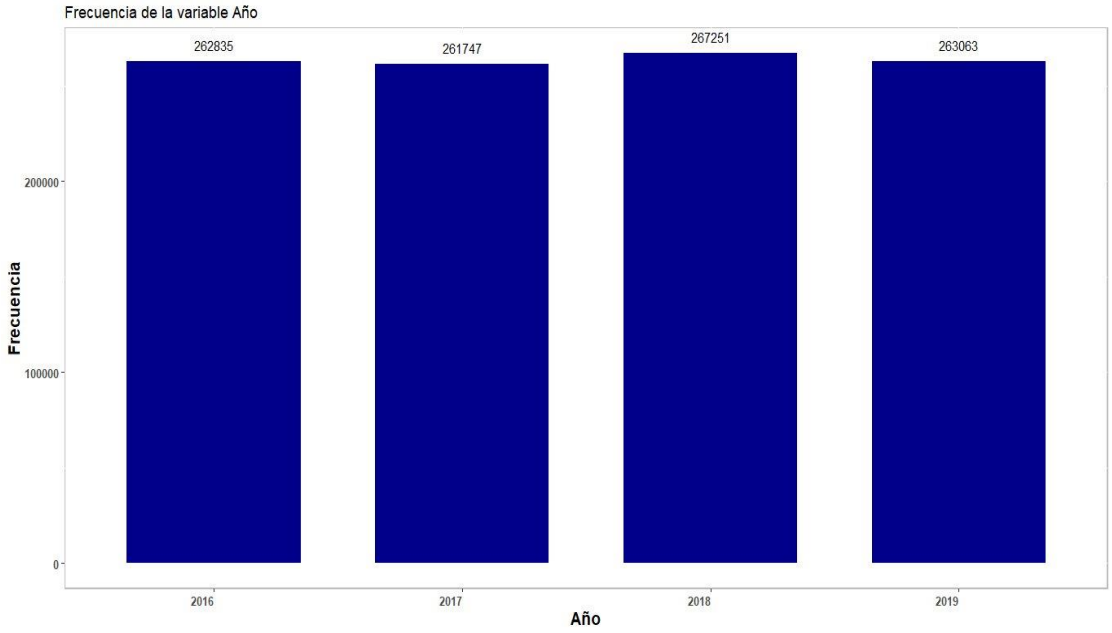
- **Year**: Esta variable indica el año en que sucedió el accidente de tránsito, tomando los valores siguientes:

Tabla 1.22- Frecuencia absoluta y relativa de la variable Year

YEAR	Frecuencia	Frecuencia relativa
2016	262835	0,2492
2017	261747	0,2481
2018	267251	0,2533
2019	263063	0,2494
Total	1054896	1,00

Se observa que la distribución de la variable Year es muy similar entre los distintos años, con un peso de 25% de cada uno de los años analizados, lo que indica que no existen diferencias en la cantidad de accidentes ocurridos en los 4 años.

Gráfico 1.19 - Frecuencia de la variable Year



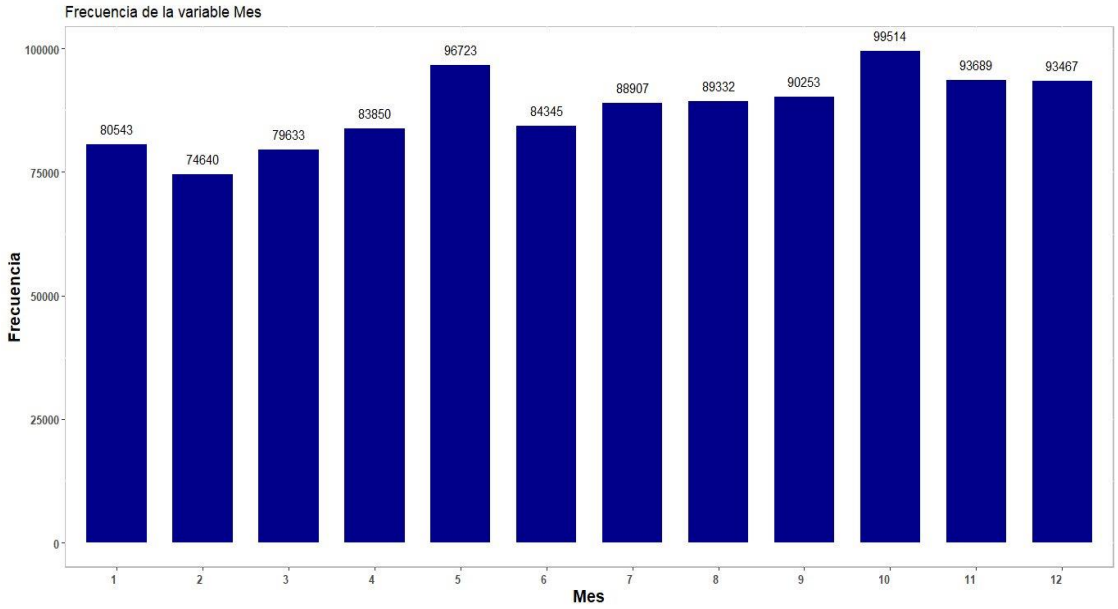
- **Mes**: Luego de realizado un feature engineering para la variable “ACC DATE”, se obtuvo la variable “Mes”, que detalla los distintos meses del año en que ocurrieron los siniestros de tránsito analizados. Se presenta a continuación su tabla de frecuencias:

Tabla 1.23- Frecuencia absoluta y relativa de la variable Mes

MES	Frecuencia	Frecuencia relativa
1	80543	0,076
2	74640	0,071
3	79633	0,075
4	83850	0,079
5	96723	0,092
6	84345	0,080
7	88907	0,084
8	89332	0,085
9	90253	0,086
10	99514	0,094
11	93689	0,089
12	93467	0,089
Total	1054896	1,00

Se muestra el detalle para los 12 meses, y se observa que la mayor cantidad de accidentes se dieron en el mes de Octubre, totalizando 99.514 siniestros, representando un 9,43% de los accidentes de tránsito ocurridos en los distintos meses. Un detalle a destacar es que se observa como en el gráfico podemos ver una tendencia al aumento de la cantidad de accidentes a medida nos movemos sobre el eje de las x de manera creciente con el número de meses (con excepción de Mayo y Octubre que son los meses donde se presentó la mayor cantidad de accidentes). Como se puede observar, los meses donde ocurre la menor cantidad de accidentes, son los meses de Enero, Febrero y Marzo, que coinciden con el invierno en el estado de Maryland.

Gráfico 1.20 - Frecuencia de la variable Mes



- **Día**: Luego de realizado un feature engineering para la variable “ACC DATE”, al igual que las dos variables anteriores, se obtuvo la variable “Día”, que detalla los distintos días del mes en que ocurrieron los accidentes de tránsito analizados. Se presenta a continuación su tabla de frecuencias:

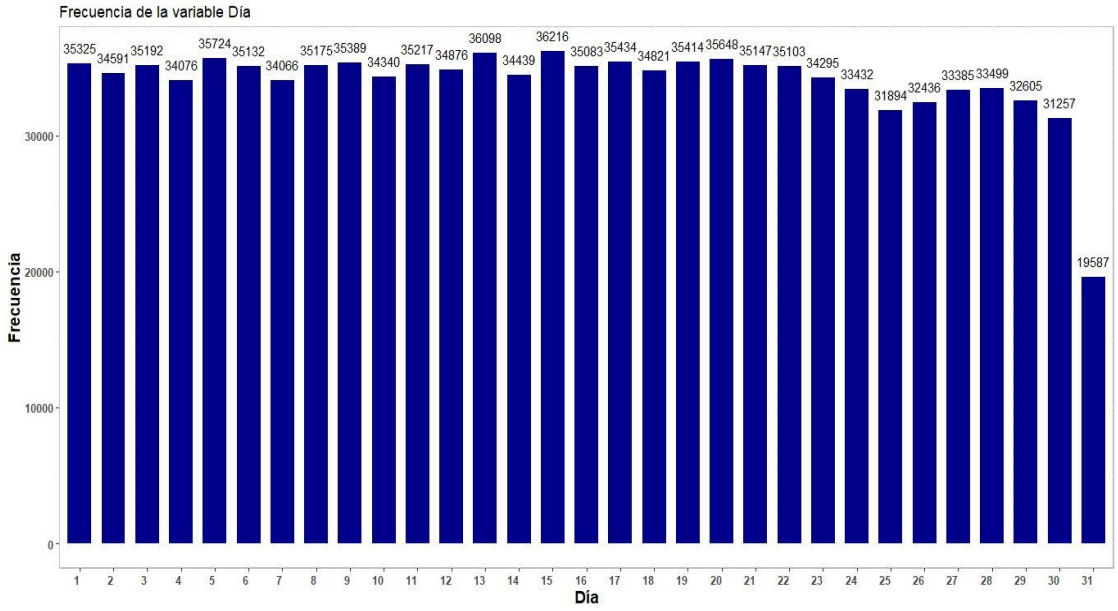
Tabla 1.24 - Frecuencia absoluta y relativa de la variable Día

DÍA	Frecuencia	Frecuencia relativa
1	35325	0,0335
2	34591	0,0328
3	35192	0,0334
4	34076	0,0323
5	35724	0,0339
6	35132	0,0333
7	34066	0,0323
8	35175	0,0333
9	35389	0,0335
10	34340	0,0326
11	35217	0,0334
12	34876	0,0331
13	36098	0,0342
14	34439	0,0326
15	36216	0,0343
16	35083	0,0333
17	35434	0,0336
18	34821	0,0330
19	35414	0,0336
20	35648	0,0338
21	35147	0,0333
22	35103	0,0333
23	34295	0,0325
24	33432	0,0317
25	31894	0,0302
26	32436	0,0307
27	33385	0,0316
28	33499	0,0318
29	32605	0,0309
30	31257	0,0296
31	19587	0,0186

Total	1054896	1,00

No se ven grandes diferencias entre los días en los que suceden los accidentes de tránsito. Con una mínima diferencia se percibe que el día del mes en que se registran más casos, es el día 15, con un total de 36.216 casos, representando un 3,43% del total. Podría detectarse una leve y muy baja tendencia a la disminución de los accidentes en los últimos días del mes. Además cabe aclarar que el día 31 se sitúa tan distante a los demás simplemente por la existencia de los meses con 30 días. Se añade el gráfico para visualizarlo:

Gráfico 1.21 - Frecuencia de la variable Día



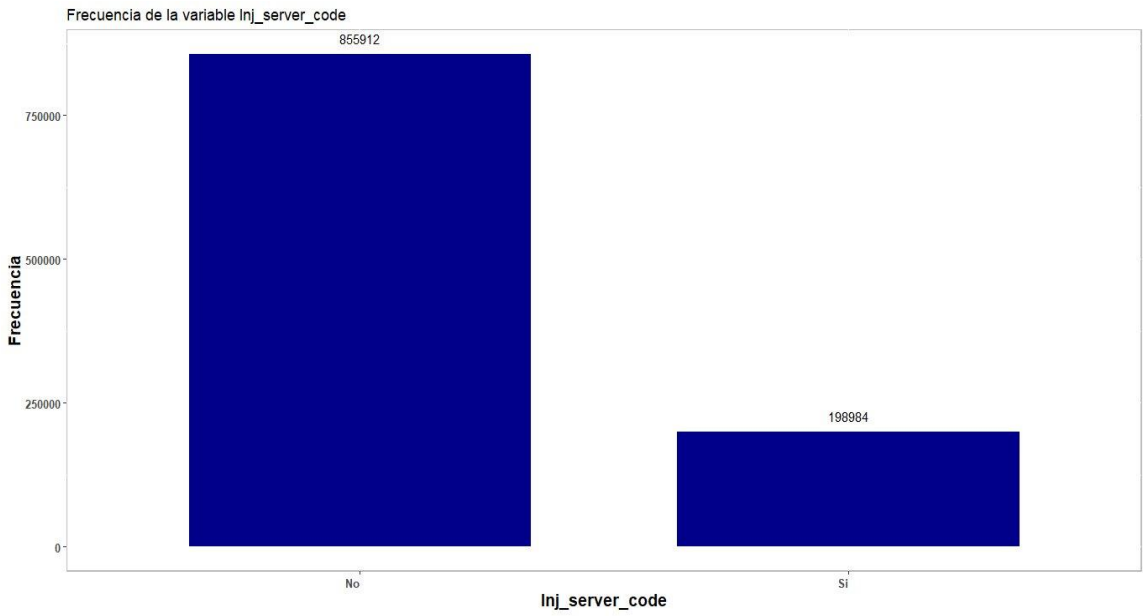
- **INJ_SERVER_CODE**: Como se comentó en el capítulo previo, esta variable sería la variable objetivo, la cual indica si hubo heridos o no dentro de los accidentes ocurridos. Esta variable se transformó a una variable binaria, adquiriendo el valor No en caso de que la persona no resulte herida y Si, en caso de que si haya sufrido heridas. Se presenta la tabla de frecuencias de la variable:

Tabla 1.25 - Frecuencia absoluta y relativa de la variable INJ_SERVER_CODE

INJ_SEVER_CODE	Frecuencia	Frecuencia relativa
No	855912	0,81
Si	198984	0,19
Total	1054896	1,00

Como se visualiza en los datos, en el 81% de los casos, la persona involucrada en el accidente no resultó herida, mientras que en el restante 19% la persona presentó heridas luego de ocurrido el siniestro.

Gráfico 1.22 - Frecuencia de la variable INJ_SERVER_CODE



Capítulo 4

Modelo Probit Espacial

En el presente capítulo se desarrolla un modelo espacial para poder predecir si dado que se dió un accidente, si en el mismo habrán heridos o no. Al ser un problema binario de clasificación, se usó un modelo *probit espacial*, basándonos en la metodología planteada por Wilhelm y de Matos (2013).

En este capítulo explicaremos brevemente en qué consiste este modelo, realizaremos un probit común a efectos de comparar con el probit espacial, y posteriormente ejecutaremos el modelo espacial mencionado.

4.1) Descripción del modelo a usar

Para aplicar el probit se parte del modelo espacial autoregresivo (SAR por sus siglas en inglés), que obedece a la siguiente fórmula:

$$z = \rho W z + X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

Donde:

z = es la variable dependiente latente inobservada

ρ o rho = es el parámetro espacial autorregresivo

W = es una matriz de pesos espaciales de dimensión n x n .

X = es una matriz con las variables regresoras o independientes

β = es un vector de coeficientes asociados a las variables regresoras

ϵ = es el error del modelo

La variable z en este caso es una variable latente que no se observa. La variable observada es la *y*, la cual es binaria e implica la ausencia o presencia de una característica, y que se obtiene de la siguiente forma:

$$y_i = \begin{cases} 1 \text{ si } z_i \geq 0 \\ 0 \text{ si } z_i < 0 \end{cases}$$

En nuestro caso la variable *y* adquirirá el valor 1 si la persona resultó herida en el accidente, y 0 en caso contrario.

Adquiere significancia en este modelo el parámetro espacial ρ (rho). Si existe dependencia espacial adquirirá valores distintos a 0, y entre -1 y 1. En este caso eso implicaría que valores de *y* de una observación estarían influenciados por valores de *y* de observaciones cercanas (Novkaniza, Djuraidah, Fitrianto, Sumertajaya, 2019). Con respecto al tema del presente trabajo final, si el p o rho es significativo al 5% o menos, eso implicaría que la existencia de heridos en un accidente está influenciada por la ocurrencia de accidentes con heridos en esa misma calle o en calles cercanas, produciéndose lo que hemos denominado como *hotspots*.

La distancia entre esas observaciones cercanas se determina por la matriz W, que contiene la información de la relación espacial entre observaciones. Esta matriz se construye o bien calculando la distancia entre las observaciones, o usando los vecinos más cercanos (Novkaniza, Djuraidah, Fitrianto, Sumertajaya, 2019).

Novkaniza, Djuraidah, Fitrianto y Sumertajaya (2019) mencionan que la estructura de la dependencia espacial agrega complejidad en la estimación de los parámetros. Según Wilhelm y de Matos (2013) las técnicas más frecuentemente usadas para estimar los parámetros en un probit espacial han sido la máxima verosimilitud y el método de los momentos generalizado (GMM). No obstante, estos autores se inclinan por una estimación Bayesiana, ya que sostienen que GMM funciona bien solamente con muestras muy grandes.

La idea central de la estimación Bayesiana es muestrear a partir de una distribución posterior de los parámetros del modelo p(z, β, $\rho|y$) en función de los datos y de las distribuciones a priori de p(z), p(β), p(ρ) (Wilhelm y de Matos, 2013). Este muestreo se hace a través de métodos de Monte Carlo con cadenas de Markov y se muestrea a partir de tres densidades condicionales: p(z|β, ρ,*y*); p(β|z, ρ,*y*); y p(ρ|z,β,*y*).

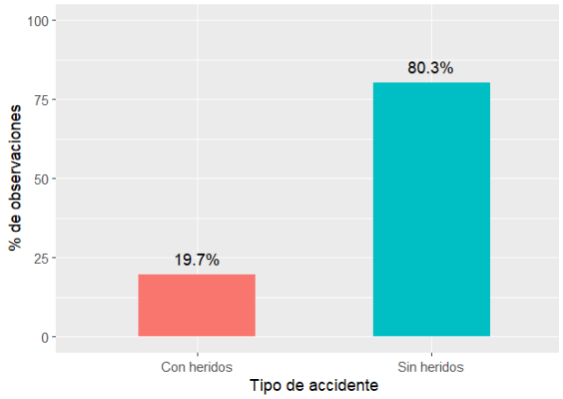
4.2) Estrategia seguida para desarrollar el modelo

Se partió del dataframe ya limpio desde Python según la estrategia mencionada en el capítulo 2, y se realizaron los siguientes pasos:

1) Se cargó en R el dataset de train procesado en Python pero sin el label encoder, creándose una dummy para cada valor de las variables categóricas excepto por el valor más frecuente. Se elimina una de las dummies para cada categoría para que no haya multicolinealidad perfecta con el intercepto.

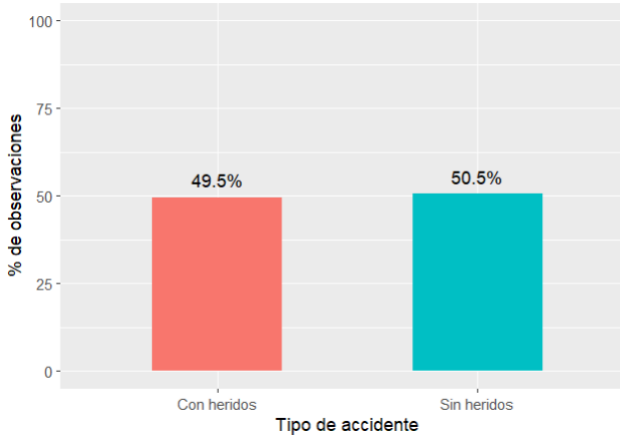
2) Debido al nivel elevado de recursos computacionales que se requieren para correr modelos espaciales en R, se obtuvo una muestra de 10 mil observaciones en train, las cuales presentan la siguiente distribución:

Gráfico 1.23 - Distribución de la muestra de 10 mil en función del tipo de accidente



3) Como las clases definidas no están balanceadas, se aplica el algoritmo SMOTE (Synthetic Minority Over-sampling Technique) con el que se crean observaciones artificiales en la clase minoritaria para con ello tratar de aumentar la predicción que se obtiene al nivel del "si", es decir, de aquellas personas que han sufrido lesiones en los accidentes de tránsito. Aplicando esta técnica se obtiene la siguiente distribución:

Gráfico 1.24 - distribución de la muestra luego de aplicar SMOTE en función del tipo de accidente



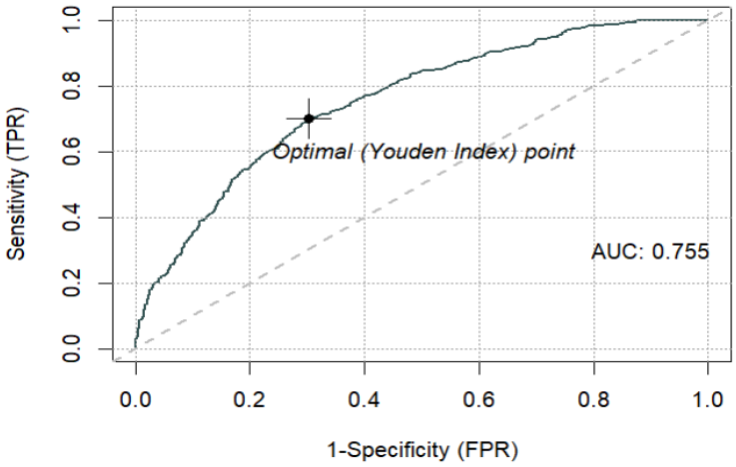
Al haber aplicado SMOTE, la muestra queda más balanceada, prácticamente en 50 y 50%.

4) A efectos de comparar, y ver si hay dependencia espacial, se realiza primero un modelo probit común, usando todas las variables expuestas, salvo la latitud y la longitud. Dado que se creó una dummy por cada categoría menos una de cada variable categórica, se tienen 169 variables para usar como regresoras. Una vez ejecutado este modelo, se removieron todas las variables que no eran significativas con un alfa del 5%, quedando solo 65 variables.

5) El modelo reducido de 65 variables se aplica en testeo para probar su performance. Para ello se carga el dataset de test ya limpio desde Python pero sin el label encoder, se obtienen dummies de forma similar a como se hizo con train, y se usa una muestra. Como se usó una muestra de 10 mil en train, se emplea una muestra de 2,500 en test, para mantener la misma proporción de 80%-20%.

6) Se realiza la predicción en test, y se analiza la Curva ROC con el modelo probit.

Gráfico 1.25 - Curva ROC en test - modelo probit común

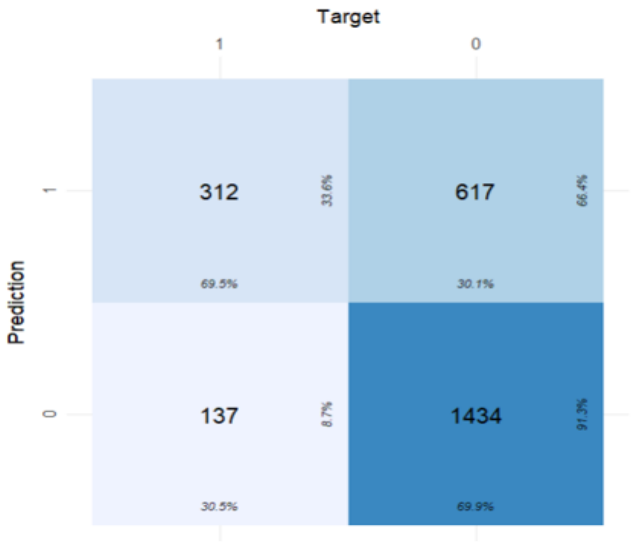


Con este modelo el área bajo la curva ROC en test es de 0.755, por lo que todavía hay espacio para mejorar ya sea mediante el probit espacial o por medio de otros modelos.

Para fijar el *cutoff* o el límite de probabilidad a partir del cual se considera que una observación cae en la clase "positiva" (que haya heridos en el accidente), se utilizó el punto óptimo del índice de Youden expuesto en el gráfico anterior, de forma de tener un equilibrio entre sensibilidad y especificidad. Dicho punto óptimo se calculó con el paquete *cutpointr*, y es de 0.488 aproximadamente. Con este *cutoff* se obtiene las siguientes métricas y matriz de confusión:

Gráfico 1.26 - Matriz de confusión en test - modelo probit

Métrica	Valor
Accuracy	69.8%
Sensibilidad o Recall	69.5%
Especificidad	69.9%
Precision	33.6%



Al haberse utilizado el índice de Youden para definir el límite para definir clases, se obtiene un equilibrio entre especificidad y sensibilidad, siendo estos dos valores, así como el accuracy general del modelo, del entorno del 69%, lo cual no es un valor muy elevado. Además, la precisión es muy baja, siendo del 33.6%, con lo cual hay un número importante de falsos positivos. De todas formas se entiende que si lo que se pretende con el modelo es predecir accidentes más riesgosos con el fin de tomar medidas para evitarlos, es preferible que haya más falsos positivos que falsos negativos. En este caso el peor error es decir que en el accidente no van a haber heridos, cuando sí los hay.

7) Antes de ejecutar el modelo espacial, se realiza un test de Moran para analizar si hay dependencia espacial en el modelo probit realizado, lo cual nos podría estar indicando en que exista una correlación entre accidentes en el mismo lugar o en calles cercanas (*hotspots*). Para ello, usamos una matriz de pesos W usando las 10 observaciones más cercanas.

Ejecutando el test, se obtiene un p-valor muy bajo (de 0.00000000000000022), con lo que se confirma que hay dependencia espacial en los residuos del modelo.

Esto significa que los residuos altos del modelo están cerca geográficamente de los residuos altos, y los bajos se encuentran cerca de otros bajos. Lo que puede estar sucediendo es lo que se ha analizado en la bibliografía citada sobre los *hotspots*, es decir que hay calles, rutas en las que, por ser muy transitadas o estar con desperfectos u otros motivos, ocurren muchos accidentes y sea más probable que hayan heridos, mientras que en otras calles o rutas con menor frecuencia no se observan accidentes de este tipo.

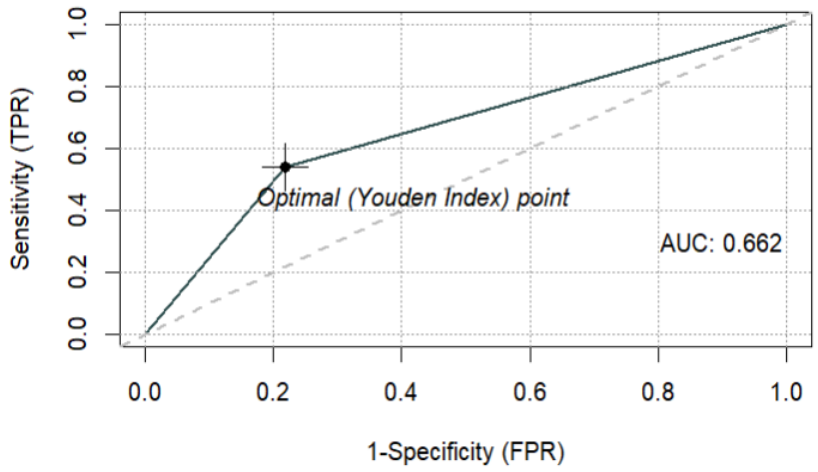
8) Al verificarse la dependencia espacial, una forma de solucionarla es usar un modelo linear generalizado espacial, en este caso un SAR probit. Se ejecuta entonces este modelo descrito anteriormente, usando la misma muestra de train obtenida con SMOTE mencionada en el punto 3, y como variables regresoras (X) las mismas variables significativas usadas para el probit común. La matriz de pesos W es la misma que se usó para el test de Moran.

9) Al ejecutar el modelo, se obtiene un ρ o rho significativo con un alfa del 5% (incluso con 0.01% o menos), y adquiere un valor de 0.5, que está indicando cómo influye sobre la

probabilidad de que haya un accidente con heridos, el hecho de que hayan habido otros accidentes cercanos con heridos.

10) Se realiza la predicción en test por medio de cálculo de matrices y aplicando la fórmula de la variable latente, ya que el paquete *spatialprobit* no tiene función de predict. Para ello primero se calcula la matriz de pesos W en test usando también 10 vecinos más cercanos. Se obtiene la siguiente curva ROC:

Gráfico 1.27 - Curva ROC en test - modelo probit espacial

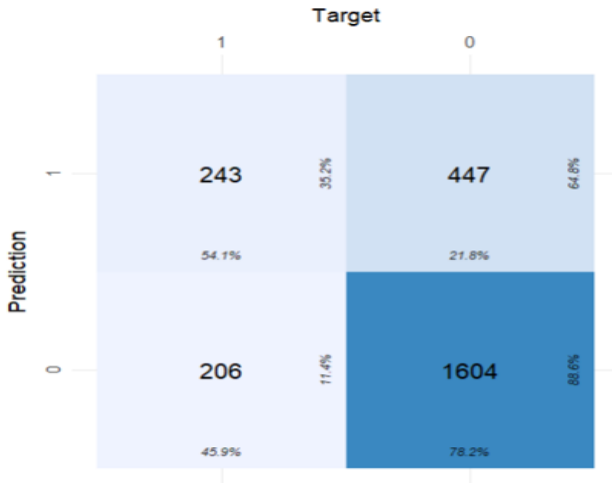


Se observa que con este modelo el área bajo la curva ROC ha disminuido con respecto al modelo probit común, siendo de 0.662, mientras que el probit común era de 0.755.

11) Se analiza para este modelo la matriz de confusión en test.

Gráfico 1.28 - Matriz de confusión en test - modelo probit espacial

Métrica	Valor
Accuracy	73.9%
Sensibilidad o Recall	54.1%
Especificidad	78.2%
Precision	35.2%



Con este modelo se obtiene un accuracy de 73.9%, que es mayor al visto en el probit común, pero la sensibilidad disminuye (de 69.5 a 54.1%) ya que detecta menos accidentes con heridos (los verdaderos positivos). Por otra parte la especificidad aumenta (de 69.9 a 78.2%), debido a que detecta un mayor número de verdaderos negativos, y la precisión mejora levemente.

A pesar de lo mencionado, el modelo no mejora con respecto al probit común debido a que, como fuera mencionado, es más importante que pueda detectar mejor los verdaderos positivos que los verdaderos negativos, es decir importa más la sensibilidad que la especificidad. Debido a este factor, y a que el modelo SAR probit es más complejo y consume más recursos computacionales, entre estos dos modelos expuestos es preferible usar el probit común, que incluso podría ejecutarse con más observaciones ya que no requiere de tantos recursos como el probit espacial.

4.3) Análisis de los factores que inciden en la severidad de los accidentes

Como ya fuera mencionado en el primer capítulo, uno de los objetivos del presente trabajo es analizar cuáles son las variables que más inciden en el hecho de que un accidente tenga o no heridos.

Como en el presente capítulo se desarrolló un modelo probit espacial el análisis de los coeficientes no resulta tan intuitivo, ya que estos no representan el cambio en la variable dependiente ante la variación de una unidad en una variable independiente. Para ello es mejor recurrir a los efectos marginales promedio, es decir el efecto esperado en la probabilidad de que haya heridos en el accidente ante cambios en una de las variables explicativas (X). Cabe destacar que, al ser un modelo espacial el cambio en una variable regresora no solo afectará en forma directa a la variable dependiente para esa observación (efecto directo), sino que lo hará también para el valor de esa variable de otras observaciones, es decir que también hay un efecto indirecto (Wilhelm, y de Matos, 2013). A continuación se presenta exclusivamente los efectos marginales directos promedio de las variables independientes (X):

Tabla 1.26 - Efecto marginal promedio de las variables regresoras (X)

Variables regresoras (X)	Efectos marginal directo promedio
BODY_TYPE_CODE_12	-0,5475
DAMAGE_CODE_99	-0,4971
LIGHT_CODE_88	-0,4110
BODY_TYPE_CODE_15	-0,3351
DAMAGE_CODE_1	-0,2912
BODY_TYPE_CODE_99	-0,2835
DAMAGE_CODE_2	-0,2769
MOVEMENT_CODE_9	-0,2667
DAMAGE_CODE_0	-0,2557
BODY_TYPE_CODE_25.88	-0,2262
BODY_TYPE_CODE_6	-0,2176
RD_COND_CODE_88	-0,2087
MOVEMENT_CODE_10	-0,1921
DAMAGE_CODE_3	-0,1882
DRIVERLESS_FLAG_U	-0,1852
BODY_TYPE_CODE_5	-0,1549
BODY_TYPE_CODE_18	-0,1356
COLLISION_TYPE_CODE_6	-0,1260
MOVEMENT_CODE_16	-0,1185
MOVEMENT_CODE_5	-0,1142
BODY_TYPE_CODE_21	-0,1119
MOVEMENT_CODE_13	-0,1062
COLLISION_TYPE_CODE_7	-0,1034
JUNCTION_CODE_88	-0,1027
BODY_TYPE_CODE_20	-0,1007
MOVEMENT_CODE_11	-0,0943
LIGHT_CODE_4	-0,0930
MOVEMENT_CODE_12	-0,0930
EQUIP_PROB_CODE_99	-0,0908
COLLISION_TYPE_CODE_9	-0,0877
MOVEMENT_CODE_7	-0,0813
MOVEMENT_CODE_4	-0,0670
SURF_COND_CODE_0	-0,0612
JUNCTION_CODE_0	-0,0565
LIGHT_CODE_3	-0,0537
EQUIP_PROB_CODE_0	-0,0506
LIGHT_CODE_6.02	-0,0465
BODY_TYPE_CODE_23.08	-0,0441
MOVEMENT_CODE_2	-0,0264
SPEED_LIMIT	-0,0013
DAY	-0,0012
MONTH	0,0045
SIGNAL_FLAG_Y	0,0313
COLLISION_TYPE_CODE_11	0,0412
SAF_EQUIP_CODE_99	0,0433
COLLISION_TYPE_CODE_1	0,0581
RD_COND_CODE_0	0,0648
COLLISION_TYPE_CODE_2	0,0789
MOVEMENT_CODE_6	0,0842
COLLISION_TYPE_CODE_8	0,0861
COLLISION_TYPE_CODE_5	0,0863
BODY_TYPE_CODE_88	0,0934
SAF_EQUIP_CODE_0	0,1207
SEX_CODE_F	0,1238
BODY_TYPE_CODE_8	0,1358
COLLISION_TYPE_CODE_4	0,1586
SAF_EQUIP_CODE_16.14	0,1705
RD_COND_CODE_2	0,1976
RD_COND_CODE_99	0,2237
DAMAGE_CODE_5	0,2544
BODY_TYPE_CODE_1	0,2808
SAF_EQUIP_CODE_1	0,3794
SAF_EQUIP_CODE_23	0,3969
SAF_EQUIP_CODE_21	0,4301
EQUIP_PROB_CODE_13	0,4858

Se observa que existen algunas variables que disminuyen la probabilidad de que haya heridos en la observación promedio (aquellas con efecto negativo), y otras que tienen un impacto positivo, es decir que incrementan en promedio la probabilidad de que haya heridos.

Al respecto parecería que la probabilidad de que haya heridos disminuye significativamente si la persona iba en determinados tipos de vehículos, como un autobus escolar (BODY_TIPE_CODE_12), o en un camión de bomberos (BODY_TIPE_CODE_15). Puede suceder que estos transportes estén elaborados con medidas extras de seguridad que no tienen los autos comunes. Asimismo, si el vehículo no sufrió daño (DAMAGE_CODE_1) también se reduce la probabilidad de que haya un herido.

Algo importante a resaltar y que fue observado en el análisis exploratorio de datos del capítulo 3, es que muchas observaciones asumen la categoría “Otros” o “Desconocido” para ciertas variables, y eso repercute en el modelo. Por ejemplo, las variables DAMAGE_CODE_99 (daño desconocido), LIGHT_CODE_88 (otros tipos de luz del día), BODY_TYPE_CODE_99 (tipo de vehículo desconocido) tienen todas un efecto marginal promedio negativo y de magnitud considerable, lo cual no aporta mucha información para poder realizar políticas o acciones para disminuir la probabilidad de que haya heridos en accidentes de tránsito.

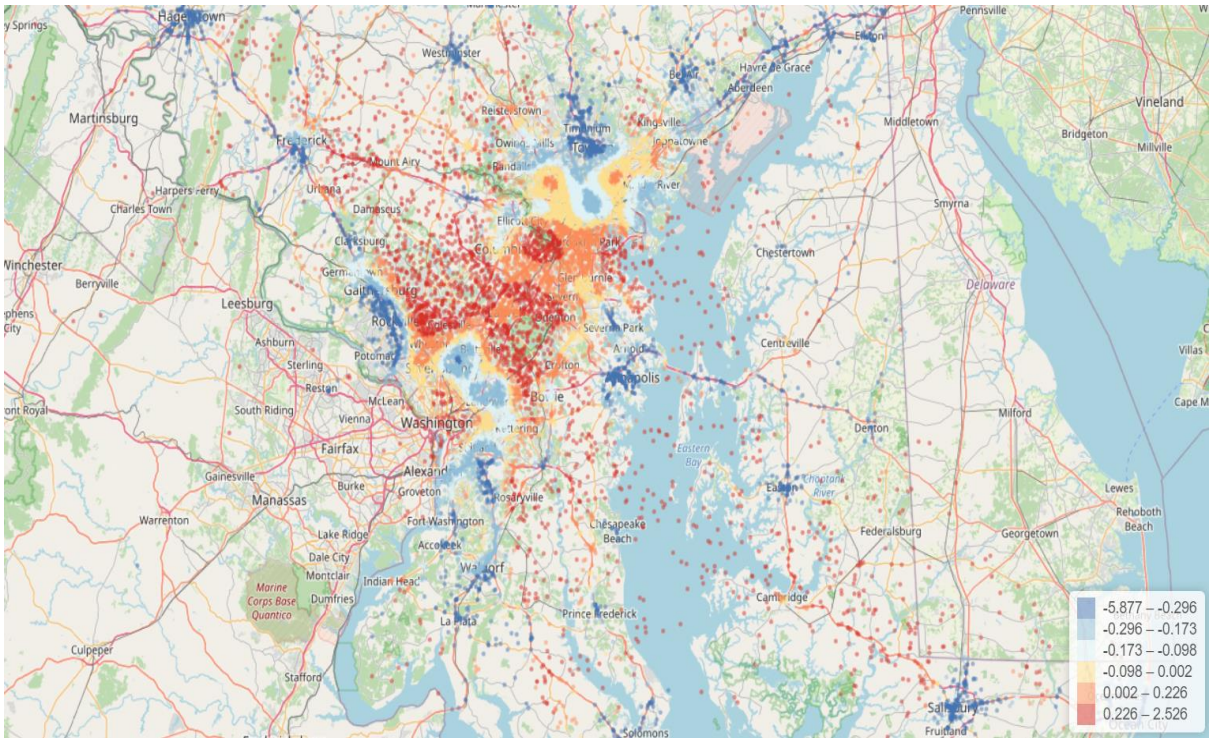
Por otro lado se encuentran variables con un efecto marginal promedio positivo, es decir que tienen un mayor impacto en la probabilidad de que existan heridos en un accidente, aumentando la misma. Un ejemplo de ello es la mala utilización de los cinturones de seguridad (EQUIP_PROB_CODE_13), que incrementaría en promedio la probabilidad de que haya heridos en un accidente. También se deduce que ser motociclista es un factor relevante de riesgo en los accidentes de tránsito incluso usando casco, debido a que las variables de equipamiento de casco, y de casco y protección ocular (SAF_EQUIP_CODE 21 y 23), así como el tipo de vehículo motocicleta (BODY_TYPE_CODE_1) tienen un efecto marginal considerable y de signo positivo en la probabilidad de resultar herido en un accidente. Por último, el no uso de ningún tipo de equipamiento de seguridad (SAF_EQUIP_CODE_1) también implica, en promedio, un incremento significativo en la probabilidad de que haya heridos. Otro aspecto relevante a señalar es que la destrucción completa del auto (DAMAGE_CODE_5) también contribuye en promedio a incrementar la

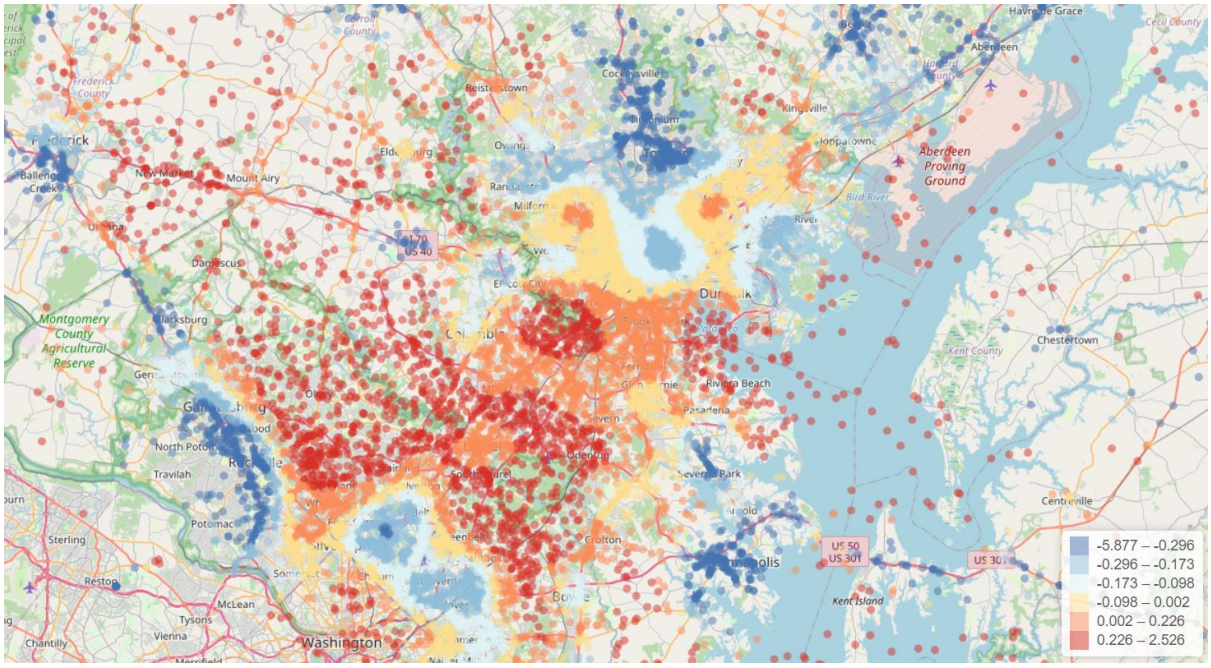
probabilidad de que haya heridos. Las condiciones de la ruta, como por ejemplo la existencia de defectos en el borde de la carretera (RD_COND_CODE_2) es un factor que puede aumentar la probabilidad de que se produzcan heridos.

4.4) Análisis de residuos del modelo probit en el espacio

A efectos de analizar cómo se comportan los residuos en el espacio, se realiza una Regresión Ponderada Geográficamente (GWR por sus siglas en inglés) con los residuos del modelo probit. Se grafican estos residuos en el mapa, obteniéndose los siguientes gráficos.

Gráfico 1.29 - Distribución espacial de residuos del modelo probit





Lo primero que se observa es que hay algunos puntos sobre el agua, lo cual probablemente implique que hay accidentes que tienen mal registradas las coordenadas, lo que significaría un problema en la calidad de los datos.

Por otra parte, se observa que los residuos no se distribuyen equitativamente por todo el Estado de Maryland, existiendo ciertas zonas (como por ejemplo Hagerstown, Westminster, Frederick, Annapolis, entre otras) donde hay mayores residuos (los puntos azules). Esto podría estar indicando la existencia de heterocedasticidad espacial en los residuos. Puede estar sucediendo que en ciertas ciudades estén ocurriendo más accidentes, y que la dependencia espacial de los accidentes no esté sucediendo con la misma intensidad en todas las zonas.

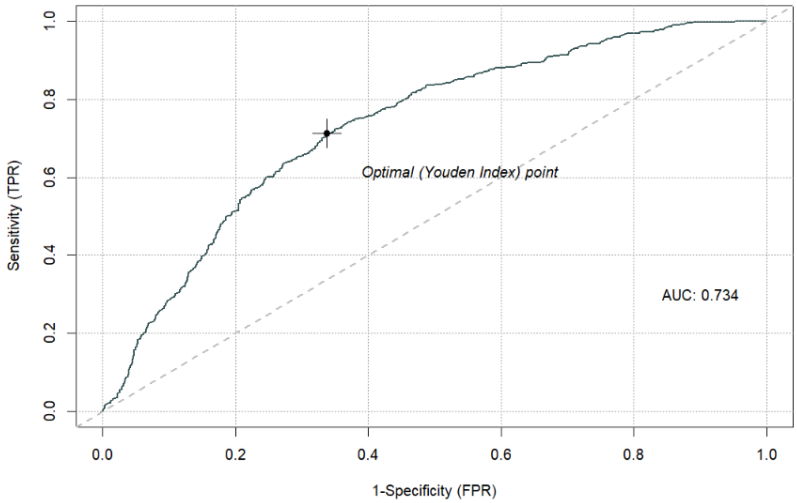
Analizando el dataset de accidentes, se observa que hay una variable que describe el condado donde ocurrió el accidente. Por los motivos mencionados, y para intentar mejorar la predicción del modelo se decide agregar esta variable al modelo probit reducido (de 65 variables) y si es posible al probit espacial. Para ello se utiliza una dummy para cada condado salvo para Baltimore, que es el que tiene más población.

4.5) Modificación del modelo probit inicial

Se ejecuta el modelo probit anteriormente mencionado, agregándole variables dummy por cada uno de los 24 condados de Maryland, excepto por uno. Se encuentra que solo las dummies de 5 condados resultan significativas: Allegany, Dorchester, Frederick, Howard y Queen Anne. A su vez, al haber agregado los condados, dejan de ser significativas 17 variables, las cuales se remueven junto con los condados que también resultan no significativos.

Se ejecuta entonces un probit con las variables resultantes del proceso mencionado (que son 53), y se analiza la performance de este nuevo modelo en test, obteniéndose un AUC menor que el modelo probit anteriormente visto.

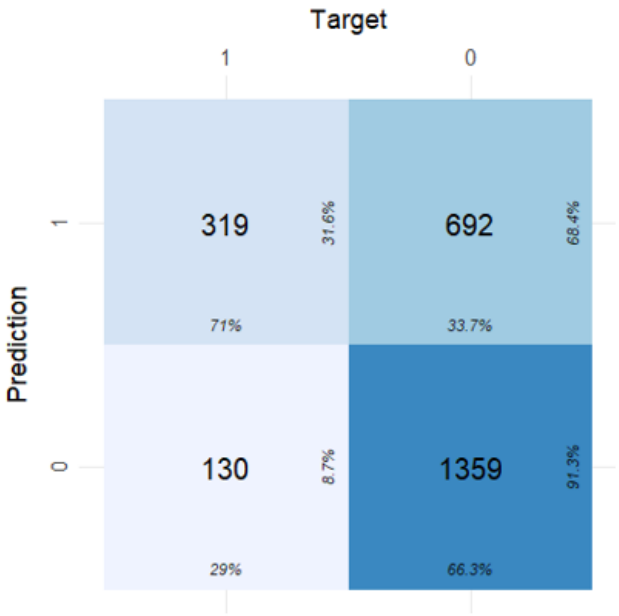
Gráfico 1.30 - Curva ROC en test - modelo probit con variables de condados



Se analiza también la matriz de confusión y las métricas de accuracy, recall, especificidad y precisión para este nuevo modelo.

Gráfico 1.31 - Matriz de confusión en test - modelo probit con variables de condados

Métrica	Valor
Accuracy	67.1%
Sensibilidad o Recall	71.1%
Especificidad	66.2%
Precision	31.6%



Se observa que la accuracy es peor que la de los otros dos modelos vistos en el presente capítulo, pero sin embargo este modelo presenta la mejor sensibilidad, es decir que es el que detecta la mayor proporción de accidentes con heridos. No obstante lo antedicho, este modelo no mejora significativamente los dos anteriores.

Se realiza nuevamente un test de Moran de los residuos de este modelo probit que toma en cuenta los condados, y otra vez el p-valor es muy bajo, con lo cual no se rechaza que pueda haber dependencia espacial. Por lo tanto se realiza un modelo probit espacial incluyendo estas nuevas variables, pero se obtiene un rho (el factor de dependencia espacial) que no resulta significativo al 5% (de hecho el p-valor es de aproximadamente del 25%). Por lo tanto, lo que podría estar sucediendo es que la dependencia espacial está escondiendo una heterocedasticidad espacial. Por este motivo se descarta este modelo.

Por último cabe mencionar que, en virtud de la distribución de residuos observada en el punto 4.4, se decidió realizar una variable dummy que separara en norte y sur del estado, y si bien la misma resultaba significativa, las métricas analizadas (AUC, accuracy, sensibilidad, entre otras) no mejoraban, por lo que también se descartó esa alternativa.

Es por esto que se decidió analizar otros modelos aparte de los vistos en este capítulo, además de que en la literatura revisada se menciona que cierto tipo de modelos de machine

learning pueden predecir adecuadamente los accidentes de tránsito. En el siguiente capítulo se describe el resultado de ejecutar los mismos.

Capítulo 5

Otros Modelos

Para la consecución de los objetivos del presente trabajo, y tras haber presentado el modelo probit espacial, se realizó el análisis y posterior aplicación de distintos modelos de machine learning que en principio se entendían buenos para este tipo de escenarios (debido a la revisión de literatura). A lo largo del desarrollo de los diversos modelos, y a medida que se fueron observando sus resultados, se fueron comprendiendo e intentando mejorar los algoritmos, para de esta forma acercarse cada vez más a una mejor predicción de si los accidentes de tránsito contaban con heridos o no. Más concretamente se trabajó con los modelos de Árboles de decisión, Gradient Boosting, Máquinas de soporte vectorial, y también una red neuronal.

Cabe destacar que para estos modelos se utilizaron todas las observaciones y no una muestra como en el capítulo anterior (aunque sí se mantuvo la separación en train y test). Esto es así debido a que no se requerían tantos recursos computacionales como con el modelo probit espacial.

Por otra parte, se intentó ejecutarlos primero sin balancear las clases (con y sin heridos) y luego, debido a los malos resultados obtenidos, se utilizó la técnica SMOTE de similar forma a como se hizo en el capítulo anterior para poder equilibrar las clases.

A continuación se presenta el detalle de los modelos realizados.

5.1) Árboles de decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico que se utiliza tanto para tareas de clasificación como de regresión.

Los mismos pueden utilizarse de manera individual o en algoritmos ensamblados, del tipo Random Forest o Gradient Boosting (el cual se analiza en el siguiente apartado). Estos

algoritmos combinan los árboles de decisión de diferentes formas para crear otros más potentes y robustos.

Con respecto a los hiperparámetros utilizados, cabe destacar que se decidió modificar con respecto a los que vienen por defecto solamente la profundidad del árbol, debido a que Scikit-Learn por defecto expande todo el árbol. Para evitar overfitting se usa una profundidad máxima (max_depth) de 7.

Con este modelo se observan las siguientes matriz de confusión, reporte de clasificación y curva ROC.

Gráfico 1.32 - Matriz de confusión en test - modelo Árbol de decisión

Métrica	Valor
Accuracy	82,9%
Sensibilidad o Recall	16,3%
Especificidad	98,4%
Precision	69,7%

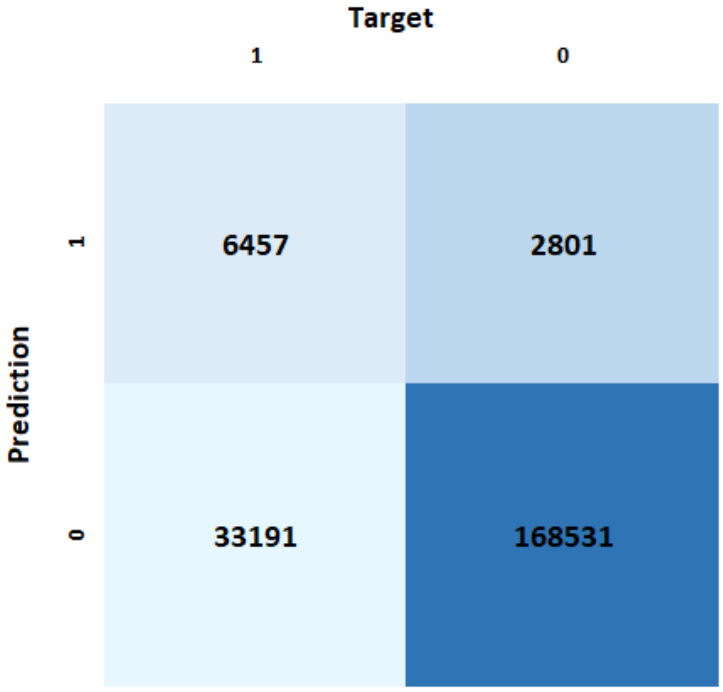
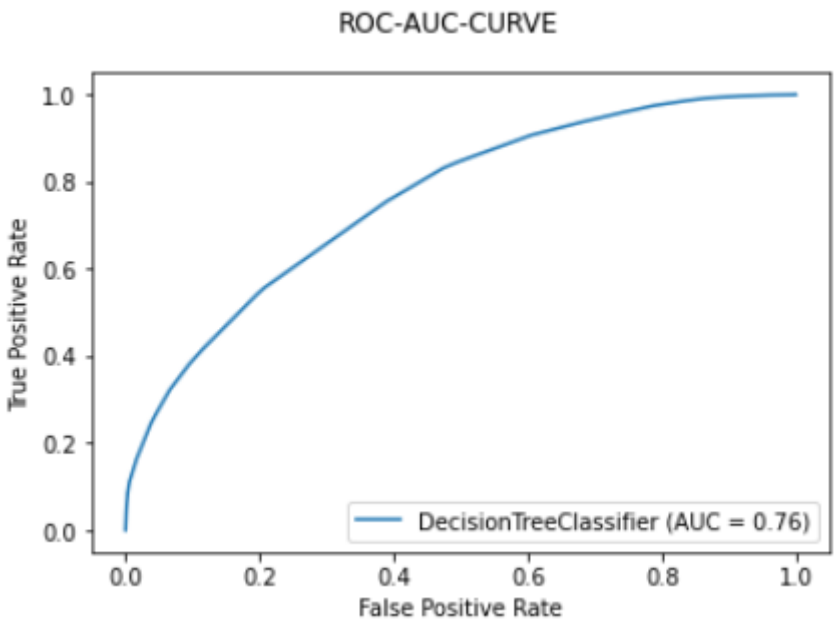


Tabla 1.27 - informe de clasificación - modelo Árbol de decisión

Informe de Clasificación:

	precision	recall	f1-score	support
0.0	0.84	0.98	0.90	171332
1.0	0.70	0.16	0.26	39648
accuracy			0.83	210980
macro avg	0.77	0.57	0.58	210980
weighted avg	0.81	0.83	0.78	210980

Gráfico 1.33 - Curva ROC en test - modelo Árbol de decisión



5.2) Gradient Boosting

Un modelo Gradient Boosting está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de modelizar y mejorar los errores de los árboles anteriores.

De esta forma, asigna ponderaciones a las salidas de los árboles individuales para luego, a las clasificaciones incorrectas del primer árbol de decisión, asignarle una ponderación más alta y una entrada al árbol siguiente. Después de numerosos ciclos, el método boosting combina estas reglas débiles en una única regla de predicción que se vuelve más poderosa.

El único hiperparámetro del modelo que se modificó con respecto a los que viene por defecto fue la profundidad de cada uno de los árboles usados (*max_depth*), que se usó un valor de 5.

Realizando predicciones y una evaluación del dataset con Gradient Boosting se obtuvieron los siguientes resultados para este modelo.

Gráfico 1.34 - Matriz de confusión en test - modelo Gradient Boosting

Métrica	Valor
Accuracy	83,6%
Sensibilidad o Recall	20,5%
Especificidad	98,2%
Precision	71,9%

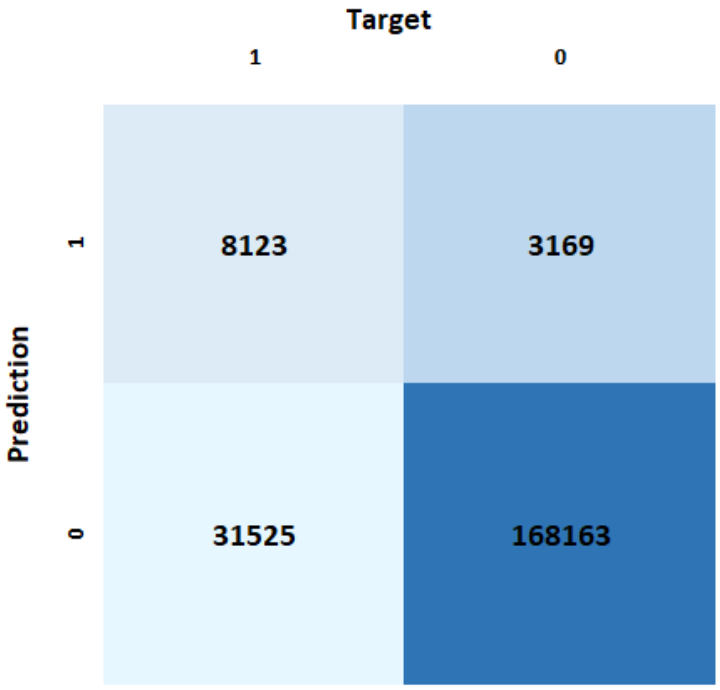
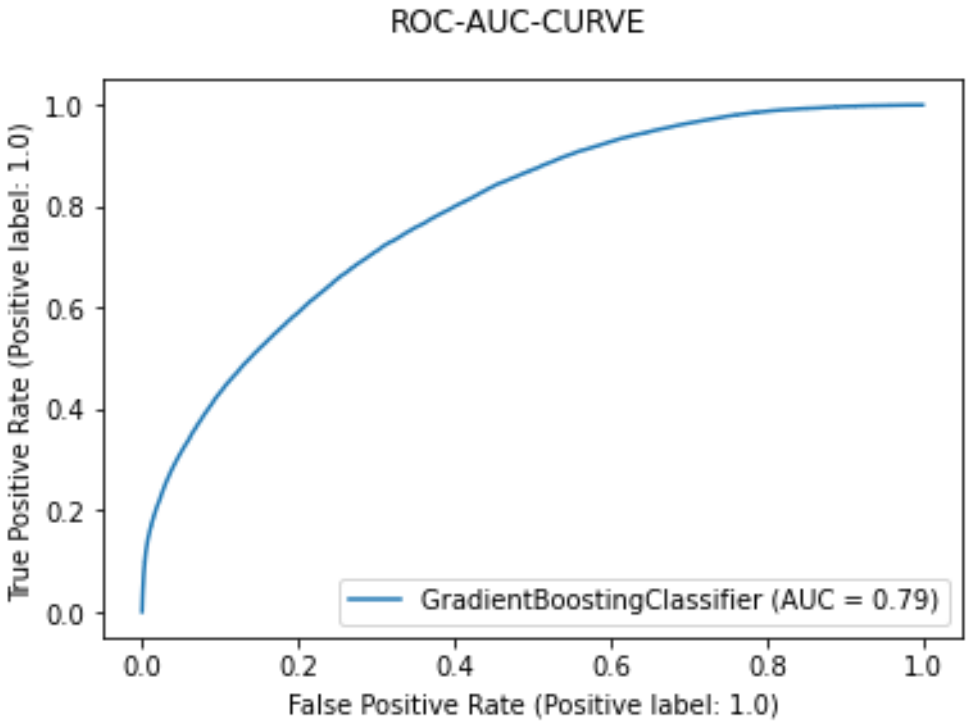


Tabla 1.28 - informe de clasificación - modelo Gradient Boosting

Informe de Clasificación:

	precision	recall	f1-score	support
0.0	0.84	0.98	0.91	171332
1.0	0.72	0.20	0.32	39648
accuracy			0.84	210980
macro avg	0.78	0.59	0.61	210980
weighted avg	0.82	0.84	0.80	210980

Gráfico 1.35 - Curva ROC en test - modelo Gradient Boosting



5.3) Máquinas de soporte vectorial (SVM)

Se procede a evaluar lo mismo que en el caso de los algoritmos anteriores (Árboles de decisión y Gradient Boosting) pero utilizando máquinas de soporte vectorial (SVM) con soporte para clases desbalanceadas para entender cómo se comporta con este algoritmo tan potente.

Las Máquinas de Vectores Soporte constituyen un método basado en aprendizaje para la resolución de problemas de clasificación y regresión. El principal uso de las Máquinas de Soporte Vectorial se da en la clasificación binaria, es decir para separar un set de datos en dos categorías o clases diferentes, como es el caso del problema analizado en el presente trabajo.

Para modelar el caso de estudio se usa un kernel sigmoide, un parámetro de regularización (C) de 1, se limitan las iteraciones en 500 y para balancear las clases se usa el argumento "balanced" para el parámetro class_weight. Con estos hiperparámetros seleccionados obtenemos los siguientes resultados.

Gráfico 1.36 - Matriz de confusión en test - modelo SVM

Métrica	Valor
Accuracy	38,0%
Sensibilidad o Recall	83,8%
Especificidad	27,4%
Precision	21,1%

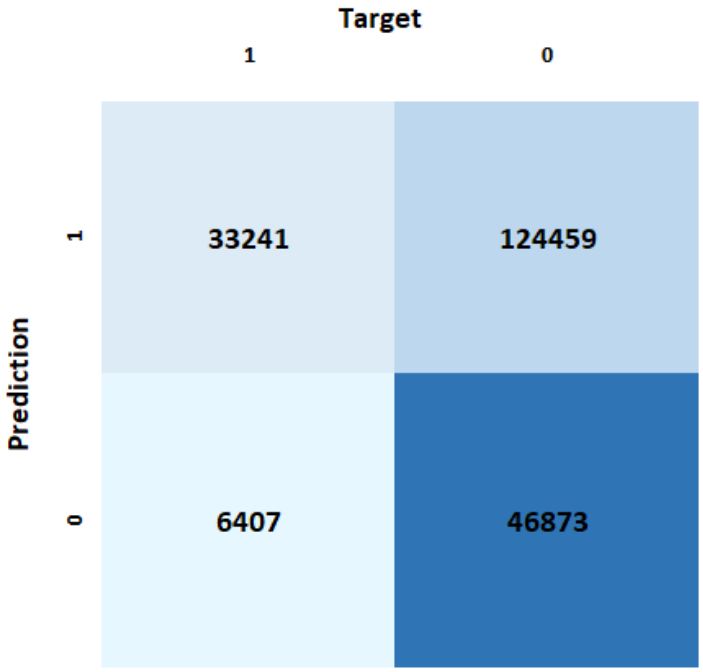
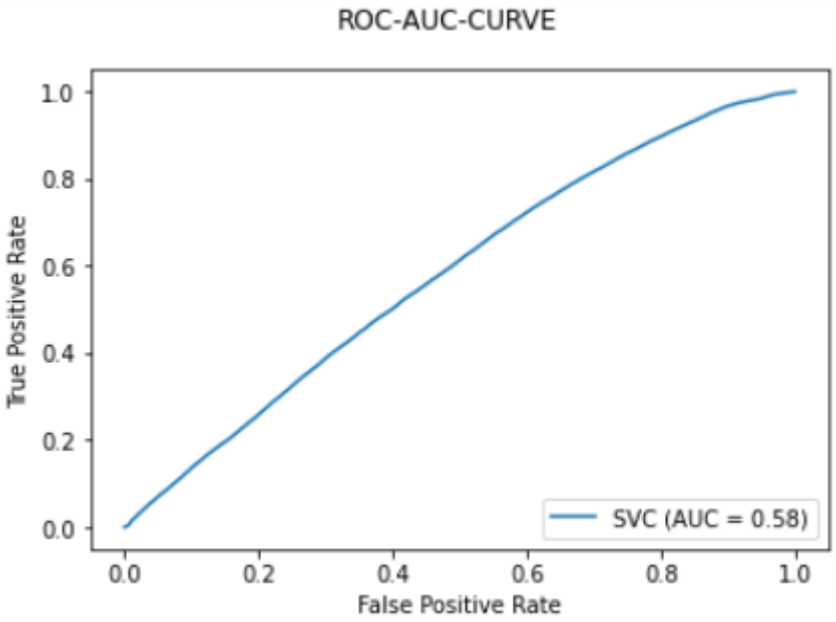


Tabla 1.29 - informe de clasificación - modelo SVM

Informe de Clasificación:

	precision	recall	f1-score	support
0.0	0.88	0.27	0.42	171332
1.0	0.21	0.84	0.34	39648
accuracy			0.38	210980
macro avg	0.55	0.56	0.38	210980
weighted avg	0.75	0.38	0.40	210980

Gráfico 1.37 - Curva ROC en test - modelo SVM



Se puede observar que aunque el modelo no es bueno, con la Máquina de Soporte Vectorial se le trata de dar importancia a la clase minoritaria y ahora el f1 del "No" es 42% y el "Si" es de un 34%. Asimismo la especificidad es de un 27% y el recall de un 84%, dándose la situación opuesta de lo que sucedía en anteriores modelos.

5.4) SMOTE

Considerando que se ha modelado el dataset con diversos algoritmos y técnicas, y no se ha obtenido resultados que garanticen una correcta clasificación y predicción, se indagó y se buscó implementar otras herramientas que ayuden a afrontar el problema del desbalance

de las categorías. Para ello, se procede a utilizar SMOTE, que es el acrónimo de sus siglas en inglés de “Synthetic Minority Over-sampling Technique” con la cual se crean observaciones artificiales en la clase minoritaria para con ello tratar de aumentar la predicción que se obtiene al nivel del “sí”, es decir, de aquellas personas que han sufrido lesiones en los accidentes de tránsito. Esta misma técnica fue utilizada también para aplicar el probit y el probit espacial.

Se tienen los siguientes eventos:

- Antes del oversampling, el conteo de la etiqueta “1” (accidente con heridos): 159.336
- Antes del oversampling, el conteo de la etiqueta “0” (accidente sin heridos): 684.580
- Después del oversampling, el conteo de la etiqueta “1” (accidente con heridos): 684.580
- Después del oversampling, el conteo de la etiqueta “0” (accidente sin heridos): 684.580

Por lo tanto con el SMOTE las clases quedan balanceadas en un 50% y 50%.

5.4.1) Árbol de decisión con SMOTE

Entrenando el modelo de decisión tree classifier con oversampling, se observan algunos cambios en los resultados presentados anteriormente.

Gráfico 1.38 - Matriz de confusión en test - modelo árbol de decisión con SMOTE

Métrica	Valor
Accuracy	82,3%
Sensibilidad o Recall	25,0%
Especificidad	95,6%
Precision	56,6%

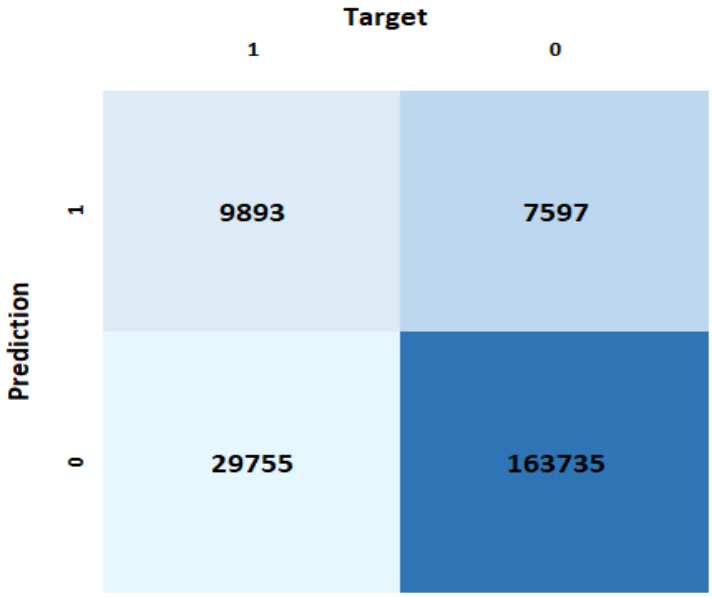


Tabla 1.30 - informe de clasificación - modelo árbol de decisión con SMOTE

	precision	recall	f1-score	support
0.0	0.85	0.96	0.90	171332
1.0	0.57	0.25	0.35	39648
accuracy			0.82	210980
macro avg	0.71	0.60	0.62	210980
weighted avg	0.79	0.82	0.79	210980

5.4.2) Gradient Boosting con SMOTE

Entrenando el nuevo modelo con oversampling con la técnica ya vista de Gradient Boosting, se obtienen las siguientes matriz de confusión, reporte y curva ROC.

Gráfico 1.39 - Matriz de confusión en test - modelo Gradient Boosting con SMOTE

Métrica	Valor
Accuracy	82,8%
Sensibilidad o Recall	29,5%
Especificidad	95,2%
Precision	58,7%

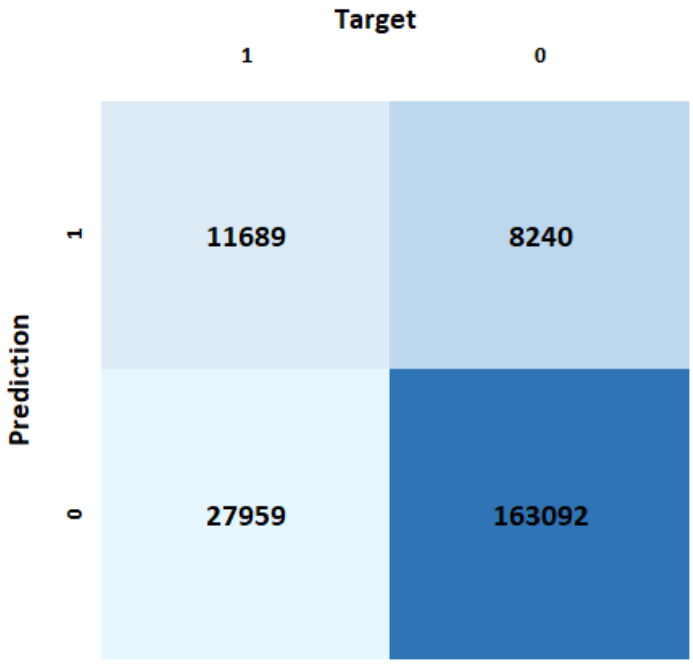


Tabla 1.31 - informe de clasificación - modelo Gradient Boosting con SMOTE

	precision	recall	f1-score	support
0.0	0.85	0.95	0.90	171332
1.0	0.59	0.29	0.39	39648
accuracy			0.83	210980
macro avg	0.72	0.62	0.65	210980
weighted avg	0.80	0.83	0.80	210980

5.4.3) SVM con SMOTE

Entrenando el nuevo modelo con oversampling con la técnica de support vector machines, se observan los siguientes resultados:

Gráfico 1.40 - Matriz de confusión en test - modelo SVM con SMOTE

Métrica	Valor
Accuracy	42,5%
Sensibilidad o Recall	75,0%
Especificidad	35,0%
Precision	21,1%

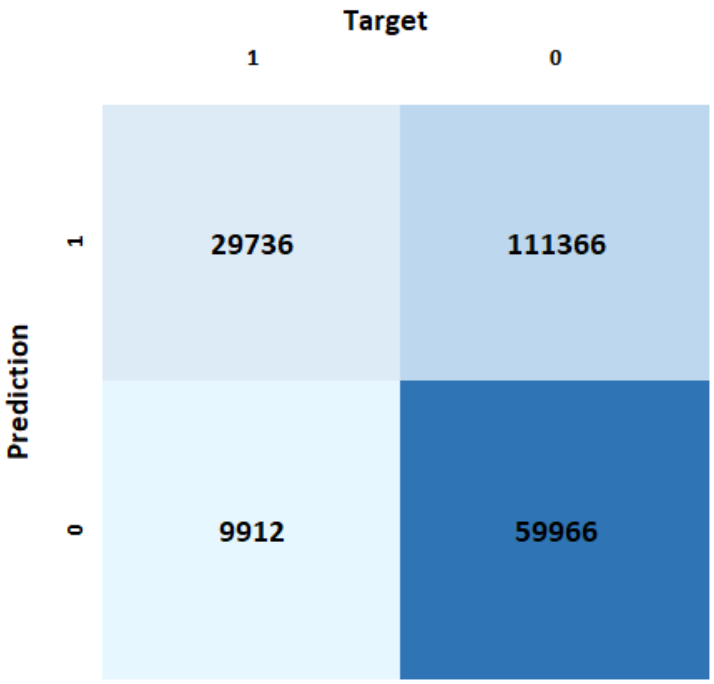


Tabla 1.32 - informe de clasificación - modelo SVM con SMOTE

	precision	recall	f1-score	support
0.0	0.86	0.35	0.50	171332
1.0	0.21	0.75	0.33	39648
accuracy			0.43	210980
macro avg	0.54	0.55	0.42	210980
weighted avg	0.74	0.43	0.47	210980

5.5) Red neuronal

Luego de entrenar los distintos algoritmos presentados anteriormente, y al observar que no se han obtenido mejores resultados, se procede a la creación de una red neuronal para analizar si mejoran las métricas de las predicciones realizadas.

Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: funciona simultaneando un número elevado de unidades de procesamiento, conectadas entre sí, que parecen versiones abstractas de neuronas.

En las redes neuronales, las unidades de procesamiento se organizan en capas: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida, con unidades que representan el campo o los campos de destino. Las unidades se conectan con fuerzas de conexión variables y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente enviándose, finalmente, un resultado desde la capa de salida.

Esta red aprende examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta que se va haciendo cada vez más precisa en la replicación de resultados conocidos, y aplicándose allí, a casos futuros en los que se desconoce el resultado.

Un modelo de red neuronal admite el análisis de regresión, de asociación y de clasificación.

Se diseña una red secuencial de 3 capas usando Keras:

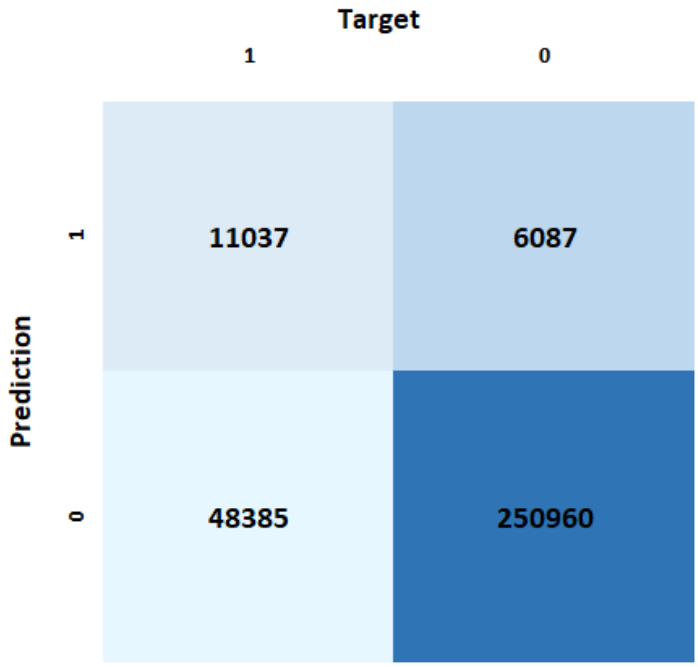
- Una capa lineal de entrada con 10 nodos, donde se aplica una activación de tipo ReLU
- Otra capa lineal con 5 neuronas, donde se aplica una función sigmoide
- Una capa de salida que obtiene la probabilidad de cada clase (con heridos o sin heridos), donde se aplica una activación softmax.

Teniendo en cuenta que la variable de salida es categórica (si se hirió la persona en el accidente o no) la función de pérdida usada es una categorical cross entropy. Como optimizador se usa Adam.

Aplicando esta red se obtiene la siguiente matriz de confusión.

Gráfico 1.41 - Matriz de confusión en test - modelo red neuronal

Métrica	Valor
Accuracy	82,8%
Sensibilidad o Recall	18,6%
Especificidad	97,6%
Precision	64,5%



5.6) Análisis de los modelos

Luego de analizar la totalidad de los modelos desarrollados a lo largo del capítulo se puede comentar que, si bien ninguno de los modelos es mejor que otro en todas las métricas analizadas, podemos tener en cuenta algunos puntos a destacar.

Dentro de los modelos trabajados se puede destacar que el modelo de Árbol de Decisión y el Gradient Boosting común se comportan muy bien en varias de sus métricas respecto a los demás modelos. Son por un lado los que presentan el mejor accuracy de todos (junto con la red neuronal), siendo de un 83% y 84% respectivamente, y adicional a esto presentan el valor más alto del área bajo la curva ROC, siendo esta de un 0,76 y 0,79 respectivamente.

Si bien ambos modelos tienen muy alto el nivel de accuracy (es el porcentaje total de elementos clasificados correctamente) esta medida no es la mejor en muchos casos, como lo es en el presente problema. En este caso se tienen dos clases que están muy desbalanceadas, lo que indica que seguramente se estará prediciendo bien muchos casos de la clase “no heridos”, resultando así en un accuracy alto sin significar que el modelo esté performando mejor. Por esto, la medida de accuracy por sí misma, es buena cuando las clases están casi equilibradas y no sería muy útil en este caso.

El Árbol de Decisión y el Gradient Boosting también tienen muy buena métrica de Precision respecto a los demás modelos, siendo de 70% y 72% respectivamente. Si bien esta métrica es importante, se entiende que no es la más relevante, siendo la métrica clave la sensibilidad o recall, tal como se mencionará más adelante.

Al mencionar la red neuronal, cabe destacar que este modelo presenta una Precision alta, aunque menor que la del Árbol de Decisión y el Gradient Boosting. En el resto de las métricas (salvo accuracy) no se observa ningún valor destacable, ya que en recall y f1-score no presenta valores altos.

Si se hace referencia a una métrica que toma importancia en un problema como el actual, el Recall toma mucho valor, ya que un guarismo muy alto del mismo estará indicando que se predicen muy bien los casos de verdaderos positivos. Es decir que en este caso, es importante tener alto recall o sensibilidad, ya que esta métrica mostrará la capacidad de detectar mayor proporción de accidentes con heridos. Es por esto que un modelo a destacar en este caso es el SVM que presenta un recall de 84%. El problema que tiene este modelo es que presenta una precision del 21%, con lo cual hay muchos falsos positivos. Por otra parte el probit común, desarrollado en el capítulo anterior, si bien tiene un recall menor al SVM (del 70%) presenta una mayor precision (34%) con lo cual tiene menos falsos positivos.

Por otra parte, se debe mencionar que los modelos de Árboles de Decisión y de Gradient Boosting aplicados sobre las clases balanceadas por medio del uso de SMOTE, presentan leves mejoras respecto a las versiones originales en lo que tiene que ver con la sensibilidad, aunque estos no sean cambios significativos.

Capítulo 6

Conclusiones

A lo largo del presente Trabajo Final de Máster se han explorado diversas técnicas modernas de Machine Learning, así como técnicas estadísticas más clásicas para poder responder nuestro planteamiento inicial: predecir la severidad de accidentes de tránsito; es decir, si hubo o no heridos en los mismos. Sin embargo durante el desarrollo del trabajo se han encontrado una serie de puntos importantes para recalcar, ya que utilizamos diferentes tipos de algoritmos, métodos de aumentación y de limpieza y creación de variables artificiales para poder producir un resultado de calidad. No obstante, los resultados no fueron tan alentadores como esperamos y se debe a las siguientes razones:

- Baja calidad en muchas variables que hicieron que las tuviéramos que borrar, y que de haberlas tenido hubieran aportado al modelo en función de la literatura revisada. Esto ahondado por un imbalance de clases.
- Para tratar de paliar el imbalance de clases se usó SMOTE, sin embargo no se logró llegar a resultados óptimos y solamente se logró crear un poco de paridad entre los datos. Esto se debe principalmente a que el método SMOTE es gaussiano y va a intentar crear un oversampling de la clase con menos observaciones y mediante probabilidades crear observaciones artificiales, esto no siempre funciona ya que dicho método siempre intentará crear curvas gaussianas explicativas pero los datos con geolocalización no siempre podrán ser explicativos, de ahí que no se obtuvieron tan buenos resultados.
- Otro problema se encontró al tratar de realizar modelos de estadística clásica ya que se encontró heterocedasticidad espacial. Esto se observó durante el capítulo 4, donde se hizo un test de Moran sobre los residuos del probit, que llevó a no rechazar la dependencia espacial, pero al revisar la distribución espacial de los mismos se vio que no se distribuían equitativamente por el espacio, y además al agregar al probit espacial variables dummies de condados, el factor de dependencia espacial rho dejó de ser

significativo. Por lo tanto la dependencia espacial hallada en una primera instancia podría estar escondiendo una heterocedasticidad espacial.

- Un inconveniente detectado durante el análisis exploratorio, y que tuvo impacto en la elaboración de modelos, era el hecho de que gran parte de las variables contaba con muchas categorías de “Otros” o “No aplica" que hubiesen podido ayudar a una mejor explicación de la severidad de los accidentes. Por lo que, muchos de los datos sólo quedaban en esta categoría y no se podía ahondar en detalles explicativos para el modelo, por ende solo se llegó a resultados no óptimos.

Es importante aclarar que se usaron varios modelos diferentes, y ninguno tuvo buenos resultados para explicar el fenómeno. También resulta interesante comentar que gran parte de la bibliografía consultada usaba árboles de decisión, redes neuronales y máquinas de soporte vectorial, sin embargo no se obtuvieron los mismos resultados. Esto indica que no necesariamente los modelos van a servir en áreas específicas y que más que utilizar un algoritmo u otro, hay que tener mucho cuidado con la calidad de los datos que se están ingestando en el modelo ya que, como se ha dicho, los datos utilizados para este trabajo carecían de variables importantes y había mucha redundancia en algunos de las categorías y en otras más bien eran demasiado generales, por ende el algoritmo no podía encontrar patrones evidentes para crear una buena generalización.

Por otro lado, los algoritmos utilizados son muy buenos prediciendo cuando no va a haber heridos, pero no tan buenos detectando cuando va a haber un accidente con heridos, es decir que fallaban en la sensibilidad o recall, en detectar los verdaderos “positivos”. Si lo que se quiere predecir es siempre los accidentes de tránsito y la ocurrencia de heridos, aquí nuestro algoritmo presenta deficiencias y se necesita más investigación y trabajo con los datos para poder llegar a resultados válidos y más extrapolables en la vida real.

Otro de los objetivos planteados en el presente trabajo fue analizar los principales factores que inciden en la severidad de los accidentes de tránsito. Para ello se utilizaron los modelos que ofrecen más interpretabilidad, es decir el probit y más específicamente el probit espacial. En ese sentido, se observó que ser pasajeros de ciertos vehículos como autobuses escolares o camiones de bomberos puede disminuir la probabilidad de resultar herido en un

accidente, así como también el hecho de que el vehículo no presente ningún tipo de daño. Por otra parte, se confirma que el mal uso de cinturones y otras medidas de seguridad aumenta considerablemente el riesgo de resultar herido en un accidente. También se observó que los motociclistas tienen un riesgo alto de resultar heridos, incluso usando casco.

Entendemos que los resultados mencionados pueden contribuir a tomar políticas públicas para disminuir la siniestralidad vial. En este sentido, parece importante mantener las exigencias de seguridad exigidas a autobuses escolares y camiones de bomberos, así como los requisitos para otorgar licencias especiales a sus conductores, ya que estas medidas parecerían estar funcionando. Asimismo, puede ser necesario realizar más campañas de comunicación en donde se insista en la importancia de los cinturones, cascos y otras medidas de seguridad para evitar lesiones en los accidentes.

A pesar de lo mencionado, el hecho citado de la gran cantidad de observaciones en categorías de otros en numerosas variables, dificultan conocer más factores que podrían ayudar a disminuir la cantidad y el daño de accidentes de tránsito. A esto se le suma el problema citado de haber eliminado por mala calidad en sus datos, a variables que son relevantes según la literatura revisada para poder predecir accidentes. Estos problemas impactan tanto en la precisión de los modelos (ya que faltan variables significativas) como en la interpretabilidad de aquellos que no son de caja negra (las categorías de otros no aportan información que podría utilizarse en políticas públicas, y la falta de variables hace que se pierda información de factores en los que se podría actuar por medio de las mismas).

Por último, se puede mencionar que, dada la distribución espacial de los accidentes vista en el capítulo 3, y la dependencia espacial observada en el capítulo 4, agregado al hecho de que el rho del modelo probit espacial resultó significativo, hay suficientes argumentos como para decir que se podrían estar formando *hotspots* en el estado de Maryland. El problema observado resulta en la especificación de cómo se producen los mismos por medios de modelos. En el presente trabajo se vio que en el SAR probit dio significativo el rho, pero la capacidad para predecir que hay heridos en un accidente era baja, de hecho la sensibilidad o recall del modelo era del 54.1%. Analizando los residuos del probit, se vio que estos no se

comportan igual en todo el territorio, pero el problema tampoco se pudo resolver agregando variables geográficas al modelo (como por ejemplo condados, o división norte-sur). De hecho agregando esas variables el rho dejó de ser significativo, con lo que podría haber heterocedasticidad espacial, tal como fuera mencionado.

Recomendaciones

Como parte del esfuerzo colectivo del grupo para poder encontrar mejores soluciones a futuro para poder replicar un modelo en la región de esta índole y que pueda generar mejores resultados se pueden dar las siguientes recomendaciones:

1. Se puede intentar utilizar un geometric SMOTE (Douzas y Bacao, 2019) para poder explicar de mejor forma el modelo y para garantizar que todas aquellas nuevas observaciones artificiales que se crean correspondan a puntos reales, con ello se puede intentar crear algún tipo de patrón evidente que el algoritmo pueda detectar y utilizar para su generalización.
2. Ser más exhaustivos en la base de datos a utilizar, ya que, no necesariamente una base que tenga datos de accidentes va a funcionar, sino que tiene que cumplir con una serie de criterios como: representatividad, que no haya más de un 5% de los datos en cada categoría nulos o que cuenten con la categoría “otros”, que no hayan variables que sean redundantes y otras que no se entienda bien qué es lo que representan.
3. En el caso del modelo de estadística espacial, se podría seguir realizando un feature engineering, para detectar posibles variables geográficas que puedan estar señalando un comportamiento distinto de la severidad de los accidentes en determinadas zonas, tal como se vio en el mapa de distribución de residuos. Asimismo, quizás se podría explorar el uso de otro tipo de modelos de estadística espacial, y compararlos con el SAR probit.
4. Los modelos de machine learning que se recomiendan para estos casos son: las máquinas de soporte vectorial, árboles de decisión, random forest y redes neuronales, sin embargo también se pueden intentar otros métodos que no nos

digan si va a haber heridos en un accidente o no, sino que más bien nos digan si existen características específicas asociadas a dichos accidentes. De esta forma podríamos convertir dicho problema de clasificación a uno de asociación o clusterización y ahí se puede intentar utilizar por ejemplo un K-means, que es un algoritmo que funciona bien con un gran volumen de datos y que además es muy explicativo para poder tomar conclusiones futuras.

5. Con respecto a los modelos de machine learning usados se podría realizar con mayor profundidad una optimización de hiperparámetros, usando por ejemplo el grid search de Scikit-Learn, a efectos de examinar si es posible mejorar las métricas analizadas.

Bibliografía

Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Information Sciences, 501, 118-135. Disponible en: https://arxiv.org/abs/1709.07377

Fan, Z., Liu, C., Cai, D., & Yue, S. (2019). Research on black spot identification of safety in urban traffic accidents based on machine learning method. Safety science, volumen 118. Disponible en: https://doi.org/10.1016/j.ssci.2019.05.039

Hébert, A., Guédon, T., Glatard, T., & Jaumard, B. (2019). High-resolution road vehicle collision prediction for the city of Montreal. In 2019 IEEE International Conference on Big Data (Big Data). IEEE. Disponible en: https://ieeexplore.ieee.org/document/9006009

Montella, A. (2010). A comparative analysis of hotspot identification methods. Accident Analysis & Prevention, 42(2). Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0001457509002632

Novkaniza, F., Djuraidah, A., Fitrianto, A., & Sumertajaya, I. M. (2019). Simulation study for comparison of spatial autoregressive probit estimation methods. In IOP Conference Series: Earth and Environmental Science (Vol. 299, No. 1, p. 012030). IOP Publishing. Disponible en : https://iopscience.iop.org/article/10.1088/1755-1315/299/1/012030/meta

Oszlak, O. (2013). Gobierno abierto: hacia un nuevo paradigma de gestión pública. Recuperado el, 17. Disponible en: https://www.oas.org/es/sap/dgpe/pub/coleccion5rg.pdf

Sangare, M., Gupta, S., Bouzefrane, S., Banerjee, S., & Muhlethaler, P. (2021). Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning. Expert Systems with Applications, 167, 113855. Disponible en: https://hal.archives-ouvertes.fr/hal-03119076/document

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. Computers volumen 10 N°12. Disponible en: https://doi.org/10.3390/computers10120157

Wilhelm, S., & de Matos, M. G. (2013). Estimating Spatial Probit Models in R. R J., 5(1), 130. Disponible en: https://journal.r-project.org/archive/2013/RJ-2013-013/RJ-2013-013.pdf

Yassin, S.S. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences. Volumen 2 N° 1576. Disponible en: https://doi.org/10.1007/s42452-020-3125-1