MEDIDAS PRÁCTICAS PARA LA REDUCCIÓN DE ACCIDENTES DE TRÁNSITO

DIPLOMADO EN MINERÍA DE DATOS

Brenda Jiménez González

Miguel Ángel Martínez Potenciano

Mayo 2018

Tabla de Contenido

1. Introducción

Los accidentes de tránsito cobran la vida de más de 1:2 millones de personas en todo el mundo cada año y reducen más años de vida productivos que cualquier otra enfermedad, incluyendo el cáncer y enfermedades cardíacas combinadas.

En Australia, ha habido más de 189 000 muertes en las carreteras desde 1925, cuando comenzó el registro de accidentes. Sin embargo, en las últimas cuatro décadas y media, no solo han disminuido las tasas de mortalidad per cápita y por exposición, sino que el número absoluto de muertes ha disminuido en alrededor de dos tercios [1].

Victoria es el estado geográficamente más pequeño de Australia, con casi 240; 000 km^2, pero el más densamente poblado con 5.9 millones de personas. En 1979, las muertes en las carreteras alcanzaron un máximo histórico, cuando el número absoluto superó las 1; 000 por primera vez. Lo que resultó en una demanda pública de cambio, y el primer resultado fue una legislación que hacía obligatorio el cinturón de seguridad, la primera en el mundo. Para 1972, esta ley se había extendido a los seis estados australianos. Médicos y otros defensores de la seguridad vial continuaron presionando para una mayor legislación de seguridad vial en los años ochenta y noventa, y se introdujeron otras intervenciones, incluidas las pruebas de aliento alcohólico; una concentración de alcohol en la sangre cero para conductores principiantes, vehículos pesados y transporte público; uso obligatorio de cascos de seguridad para ciclistas; la aplicación automatizada (cámara) del exceso de velocidad, principalmente.

Así, Victoria fue líder en la regulación y aplicación del comportamiento de los usuarios en la carretera, apoyando este enfoque con la educación. A pesar de los avances en la seguridad del tráfico durante este tiempo, el foco principal de la política pública se mantuvo en el cambio de comportamiento y la búsqueda del comportamiento “perfecto" de los usuarios de la carretera. Si bien muchas de las intervenciones adoptadas demostraron impacto, los investigadores comenzaron a identificaron problemas de diseño y problemas operacionales en el sistema que incluyeron en el comportamiento del usuario en la carretera; en desacuerdo con la posición de política pública que culpaba únicamente a los usuarios.

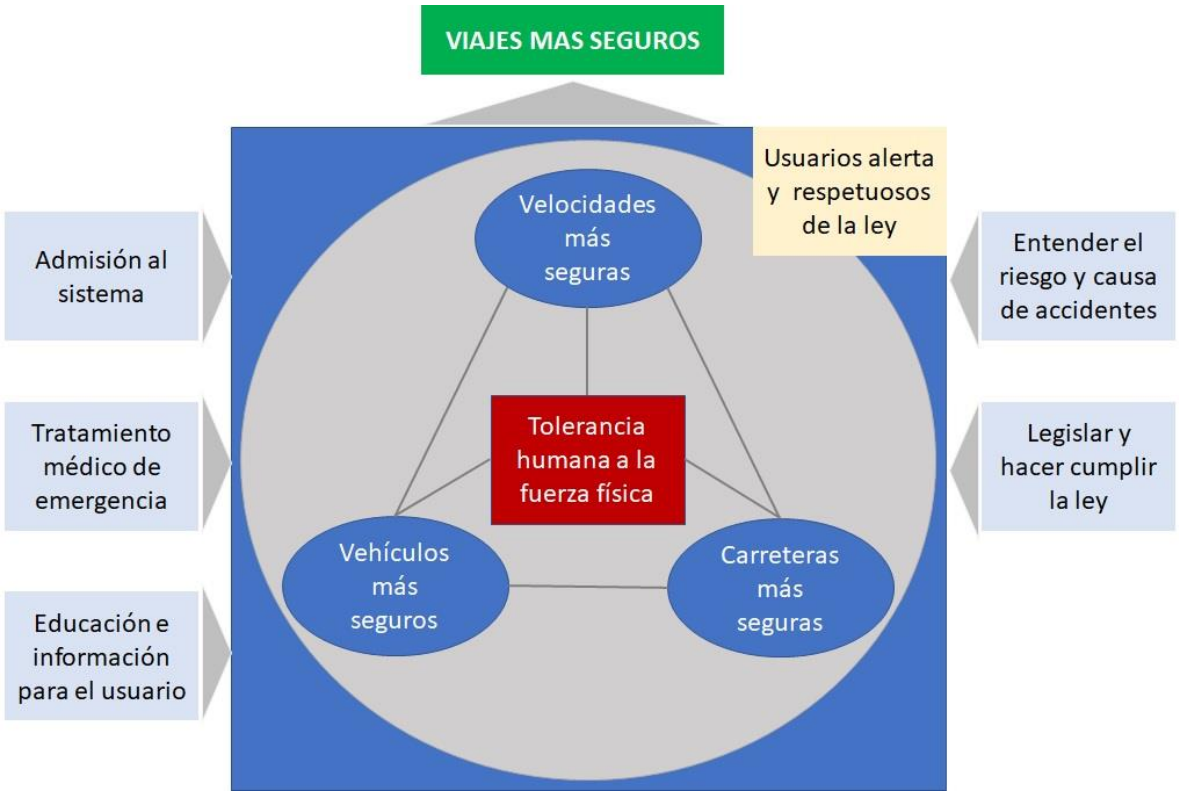
Entre los profesionales de la seguridad del tráfico, el usuario individual de la carretera pasó a ser considerado cada vez menos como la causa predominante y más como el eslabón débil del sistema.

Entonces, la atención se centró en la consideración de cómo el diseño del sistema vial y de tráfico podrían aminorar errores humanos comunes.

La filosofía de *Vision Zero* tenía como objetivo lograr un sistema de transporte por carretera sin víctimas mortales o lesiones graves. A finales de 1990, hubo un fuerte apoyo para *Vision* Zero entre los investigadores australianos. A través del tiempo, quedó claro que los gobiernos australianos apoyarían completamente la Vision Zero en la forma en que se adoptó en Suecia.

Para 2004, se propuso un enfoque de cambio menos "radical" para Victoria, que surgió de una combinación de la filosofía *Vision Zero* y el modelo de Seguridad Sostenible desarrollado por los holandeses. El enfoque de *Safe System* replantea la forma en que se ve y se gestiona la seguridad vial (figura 1). El punto central de este enfoque es la proposición de que el traumatismo vial no se puede intercambiar para mejorar la movilidad. Su objetivo es abordar todos los elementos del sistema de transporte carretero en una manera integrada de garantizar que los niveles de energía cinética en los choques permanezcan por debajo de lo que el cuerpo humano puede tolerar. Mientras que los gobiernos parecen operar principalmente en el nivel reactivo, las instituciones líderes en Victoria ahora planifican y gestionan a nivel proactivo. La adopción de los principios de Sistema Seguro y el desarrollo de estrategias científicas de seguridad vial demuestran esto.

Figura 1. Enfoque del sistema seguro Safe System



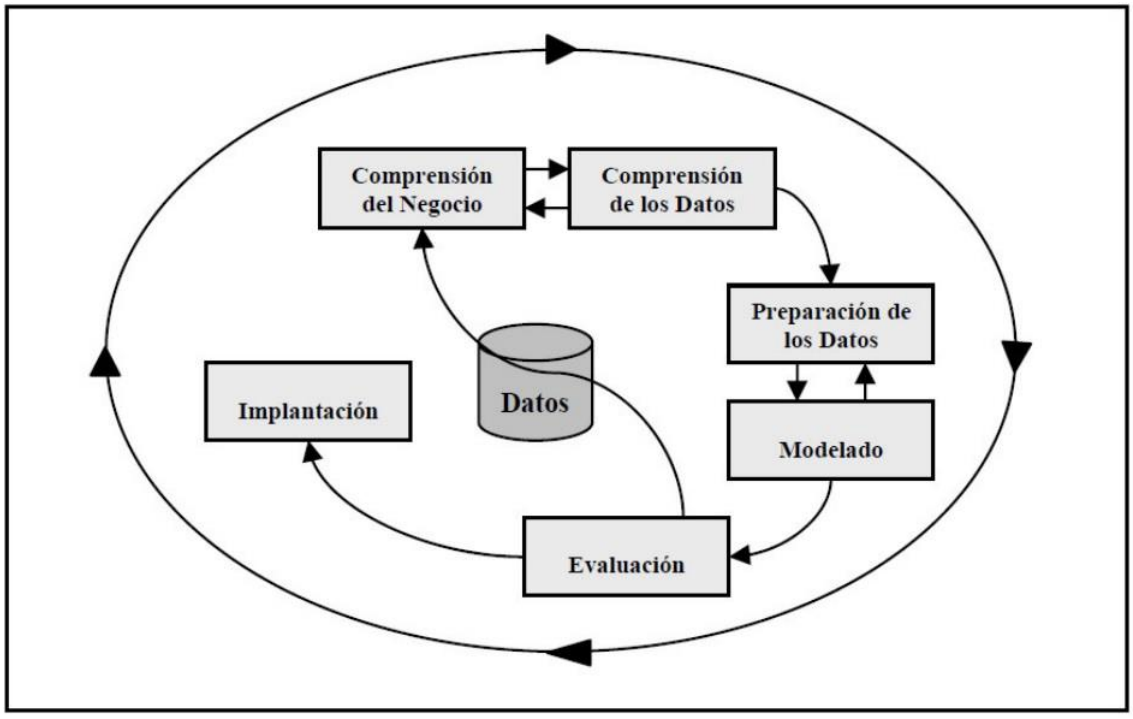
2. Objetivo

El objetivo de este documento es desarrollar un sistema de minería de datos, utilizando la metodología CRISP, para proponer medidas prácticas que disminuyan la cantidad de accidentes de tránsito.

3. Metodología CRISP 4

CRISP-DM (*Cross Industry Standard Process for Data Mining*), es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining. La metodología CRISP-DM está estructurada en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada. En la figura 2 se puede observar las fases en las que se divide y las posibles secuencias a seguir entre ellas.

Figura 2: Metodología CRISP-DM



En las siguientes secciones se explicará la aplicación de cada una de las fases de la metodología en el contexto de nuestra aplicación, pero a modo de resumen podemos decir que los objetivos en cada una de las fases son los siguientes.

1. Comprensión del negocio
   - Entendimiento de los objetivos y requerimientos del proyecto
   - Definición del problema de Minería de Datos
2. Comprensión de los datos
   - Obtención conjunto inicial de datos
   - Exploración del conjunto de datos
   - Identificar las características de calidad de los datos
   - Identificar los resultados iniciales obvios
3. Preparación de Datos
   - Selección de datos
   - Limpieza de datos
4. Modelado
   - Implementación de herramientas de Minería de Datos
5. Evaluación
   - Determinar si los resultados coinciden con los objetivos del negocio
   - Identificar los temas de negocio que deberían haberse abordado
6. Despliegue
   - Instalar los modelos resultantes en la práctica
   - Configuración para minería de datos de forma repetida o continua

3.1. Comprensión del negocio

Deseamos proponer medidas prácticas para disminuir la cantidad de accidentes de tránsito ocurridos. Analizaremos la base de datos para encontrar asociaciones entre las variables, para determinar si la reducción de dimensiones es posible. Una vez seleccionadas las variables de mayor incidencia en accidentes de tránsito, construiremos dos modelos de minería de datos:

- Árbol de decisión
- Red neuronal

Evaluaremos estos dos modelos para determinar el que mejor clasifique los accidentes.

La variable objetivo que deseamos clasificar es el **riesgo de colisión**. Es decir, queremos clasificar los vehículos, y por consiguiente, a los individuos cuyo riesgo de colisión sea más alto, comparado con los demás.

3.2. Comprensión de los datos

La base de datos utilizada, *CrashStats*, es proporcionada de modo abierto por *VicRoads*, en su *website* oficial.

La base consta de 9 tablas o archivos, listados en el Cuadro 1. Primero se realizó el análisis de la base de datos para obtener el diagrama Entidad-Relación. Después se efectuó un análisis exploratorio de datos las tres principales tablas: Accidentes, Personas y Vehículos.

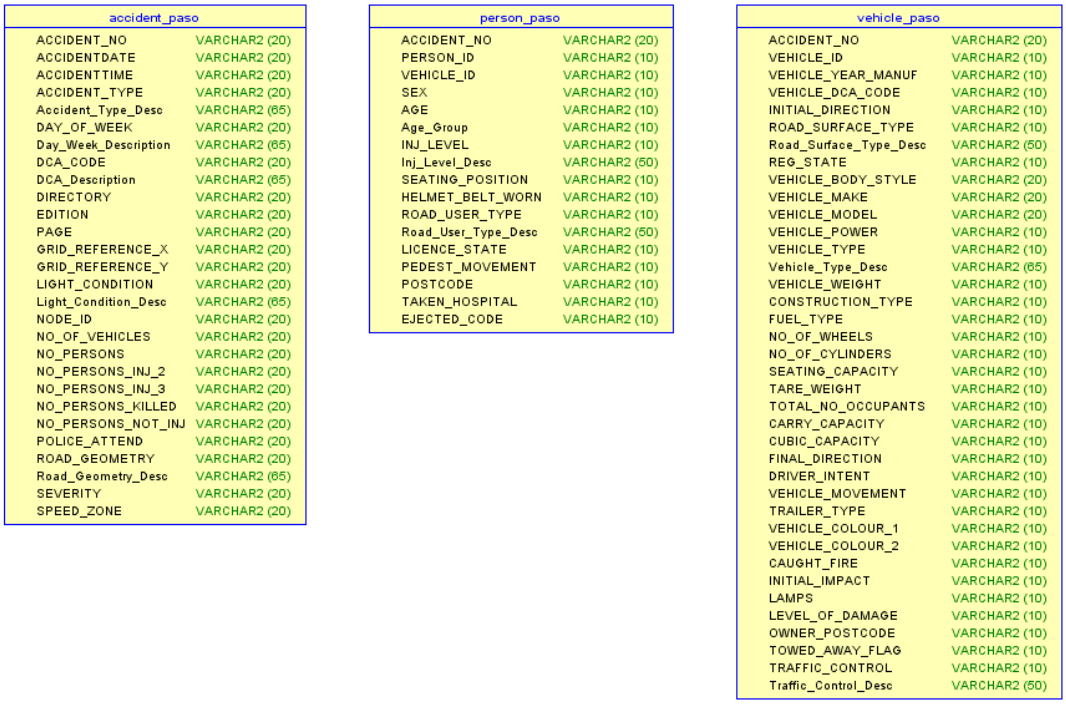
3.2.1. Limpieza de datos y Modelo Entidad-Relación

Debido a la cantidad de información en los archivos que se van a analizar, se tomó la decisión de hacer un proceso de migración de base de datos siguiendo lo siguiente

1. Se crearon tablas de paso, donde almacenamos los datos como vienen en los archivos, sin sufrir ningún tipo de modificación, para evitar tener contratiempos relacionados con:
   - Tipos de datos no coincidentes (fecha, número): en este caso habría que trabajar por mantener la integridad de la información
   - Inconsistencia en de la información: identificadores distintos en cada columna para un mismo catálogo
2. Posteriormente se realizó el modelo entidad relación llegando a la tercera forma normal, se identificaron las tablas consideradas como catálogos (que sufren poca transaccionalidad de la información)
3. Continuamos con la aplicación de los **procesos ETL** (extracción, transformación y carga):
   - Creando procedimientos para verificar que los datos coincidieran con su respectivo catálogo, y colocando un identificador para aquellos registros que no lo fueran para evitar la pérdida de datos
   - Identificación de cada tipo de dato que le corresponde a cada elemento de las tablas
   - Por integridad referencial se eliminaron
     - 3 registros de la tabla Accidentes
     - 87 registros de la tabla Personas

En la figura 3 se muestran el modelo de las tablas de paso mencionado anteriormente

Figura 3: Modelo de tablas de paso



En estas tablas se cargan los datos completos y sin tener ningún tipo de modificación en la información. Teniendo la comprensión de los datos es necesario construir el modelo Entidad Relación cubriendo la tercera forma normal.

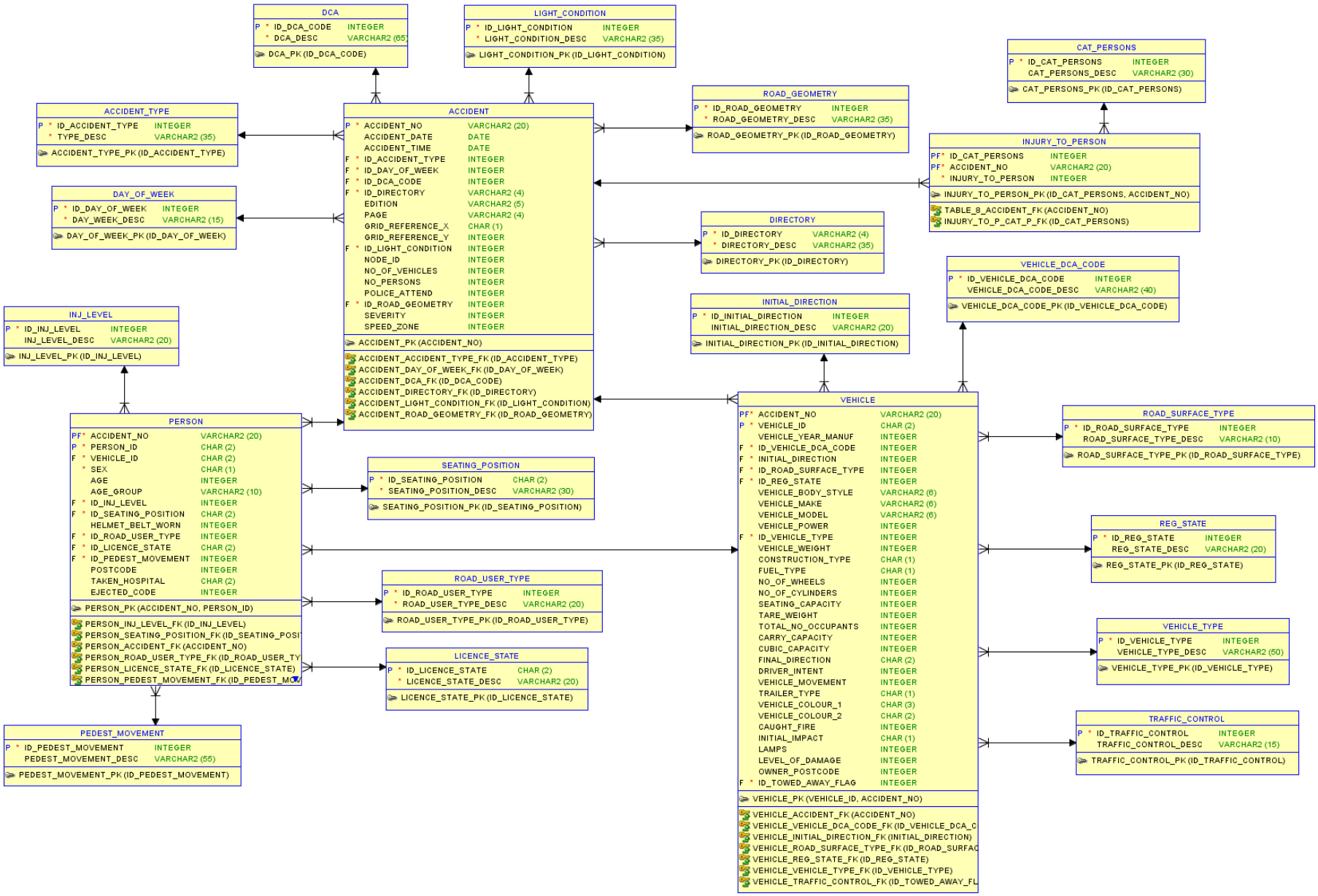
En la figura 4 se muestra el modelo Entidad Relación.

Teniendo la base de datos normalizada, se realiza la carga de información inicial, poblando el modelo de datos que se construyó. Para lo cual se llevaron a cabo una serie de tareas básicas, tales como la limpieza de datos, calidad de datos, procesos ETL, etc.

Primero se cargaron los datos en las tablas identificadas como catálogos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento de las tablas relacionadas para mantener la integridad referencial.

Posteriormente se realizó la carga de la información en las tablas de trabajo, sin perder atención en el tipo de dato que le correspondía a cada elemento.

Figura 4: Modelo entidad relación



3.2.2. Análisis exploratorio de datos

Para el análisis exploratorio se usaron las tablas de Accidente, Persona y Vehículos, cuya descripción puede consultarse en el Cuadro 1.

Cuadro 1: Archivos de la base CrashStats

Archivo	Descripción
Accidente	Detalles básicos del accidente
Persona	Detalles de las personas
Vehículos	Vehículos Detalle de los vehículos
Evento_Accidente	Evento_Accidente Secuencia de eventos
Condiciones_Superficie	Condiciones de la carretera
Condiciones Atmosféricas	Condiciones atmosféricas
DCA_código	Código detallado describiendo el accidente
Nodo_Accidente	Localización
Ruta_Accidente	Ruta_Accidente Ruta detallada y PK

Los atributos de la Tabla Accidentes se muestran en el Cuadro 2. De las 28 variables en la tabla, 6 son variables numéricas y 22 son variables cualitativas.

Cuadro 2: Atributos de la tabla Accidentes

No.	Nombre de la variable	Tipo de variables		N	Valores perdidos
1	**ACCIDENT_NO**	Cualitativa	Nominal	164,789	0
2	**ACCIDENTDATE**	Cualitativa	Ordinal	164,789	0
3	**ACCIDENTTIME**	Cualitativa	Ordinal	164,789	0
4	**ACCIDENT_TYPE**	Cualitativa	Nominal	164,789	0
5	**Accident Type Desc**	Cualitativa	Nominal	164,789	0
6	**DAY WEEK**	Cualitativa	Ordinal	164,789	0
7	**DAY WEEK Description**	Cualitativa	Ordinal	164,789	0
8	**DCA_CODE**	Cualitativa	Nominal	164,789	0
9	**DCA Description**	Cualitativa	Nominal	164,789	0
10	**DIRECTORY**	Cualitativa	Nominal	163,889	900
11	**EDITION**	Cualitativa	Nominal	163,889	900
12	**PAGE**	Cualitativa	Nominal	163,889	900
13	**GRID_REFERENCE_X**	Cualitativa	Nominal	163,889	900
14	**GRID_REFERENCE_Y**	Cualitativa	Nominal	159,611	5,178
15	**LIGHT_CONDITION**	Cualitativa	Nominal	164,789	0
16	**Light Condition Desc**	Cualitativa	Nominal	164,789	0

17	**NODE_ID**	Cualitativa	Nominal	164,789	0
18	**NO_OF_VEHICLES**	Cuantitativa	Razon	164,789	0
19	**NO_PERSONS**	Cuantitativa	Razon	164,789	0
20	**NO_PERSONS_INJ_2**	Cuantitativa	Razon	164,789	0
21	**NO_PERSONS_INJ_3**	Cuantitativa	Razon	164,789	0
22	**NO_PERSONS_KILLED**	Cuantitativa	Razon	164,789	0
23	**NO_PERSONS_NOT_INJ**	Cuantitativa	Razon	164,789	0
24	**POLICE_ATTEND**	Cualitativa	Nominal	164,789	0
25	**ROAD_GEOMETRY**	Cualitativa	Nominal	164,789	0
26	**Road Geometry Desc**	Cualitativa	Nominal	164,789	0
27	**SEVERITY**	Cualitativa	Ordinal	164,789	0
28	**SPEED_ZONE**	Cualitativa	Ordinal	164,789	0

En la Tabla Persona se tienen 19 atributos, de los cuales solo 1 es se refiere a una variable numérica. Se muestran en el Cuadro 3.

Cuadro 3. Atributos de la tabla Personas

No.	Nombre_Variable	Tipo variable	Escala Medición	% Valores nulos
1	**ACCIDENT_NO**	Cualitativa	Nominal	0.00%
2	**PERSON_ID**	Cualitativa	Nominal	0.00%
3	**VEHICLE_ID**	Cualitativa	Nominal	0.00%
4	**AGE**	Cuantitativa	Razón	0.00%
5	**SEATING_POSITION**	Cualitativa	Nominal	0.00%
6	**LICENCE_STATE**	Cualitativa	Nominal	0.00%
7	**EJECTED_CODE**	Cualitativa	Nominal	0.00%
8	**AGE.GROUP**	Cualitativa	Ordinal	0.00%
9	**HELMET_BELT_WORN**	Cualitativa	Nominal	0.00%
10	**PEDEST_MOVEMENT**	Cualitativa	Nominal	0.00%
11	**VEHICLE_ID**	Cualitativa	Nominal	0.00%
12	**INJ_LEVEL**	Cualitativa	Ordinal	0.00%
13	**ROAD_USER_TYPE**	Cualitativa	Nominal	0.00%
14	**POSTCODE**	Cualitativa	Nominal	17.40%
15	**SEX**	Cualitativa	Nominal	0.00%
16	**INJ_LEVEL_DESC**	Cualitativa	Ordinal	0.00%
17	**ROAD_USER_TYPE_DESC**	Cualitativa	Nominal	0.00%
19	**TAKEN_HOSPITAL**	Cualitativa	Nominal	0.00%

La tabla Vehículos cuenta con 38 atributos, de las cuales 9 son variables numéricas y 29 son variables cualitativas.

Cuadro 4. Atributos de la tabla Vehicle

No.	Nombre de la variable	Tipo de variables		N	% Valores nulos
1	**Accident_No**	Cualitativa	Nominal	293,036	0.00%
2	**Vehicle_Id**	Cualitativa	Nominal	293,036	0.00%
3	**Vehicle_Year_Manuf**	Cuantitativa	Intervalo	293,036	7.80%
4	**Vehicle_Dca_Code**	Cualitativa	Nominal	293,036	0.04%
5	**Initial_Direction**	Cualitativa	Nominal	293,036	0.00%
6	**Road_Surface_Type**	Cualitativa	Nominal	293,036	0.00%
7	**Road Surface Type Desc**	Cualitativa	Nominal	293,036	0.00%
8	**Reg_State**	Cualitativa	Nominal	293,036	0.00%
9	**Vehicle_Body_Style**	Cualitativa	Nominal	293,036	0.00%
10	**Vehicle_Make**	Cualitativa	Nominal	293,036	0.00%
11	**Vehicle_Model**	Cualitativa	Nominal	293,036	0.00%
12	**Vehicle_Power**	Cualitativa	Ordinal	293,036	100.00%
13	**Vehicle_Type**	Cualitativa	Nominal	293,036	0.00%
14	**Vehicle Type Desc**	Cualitativa	Nominal	293,036	0.00%
15	**Vehicle_Weight**	Cuantitativa	Intervalo	293,036	87.36%
16	**Construction_Type**	Cualitativa	Nominal	293,036	0.00%
17	**Fuel_Type**	Cualitativa	Nominal	293,036	0.00%
18	**No_Of_Wheels**	Cuantitativa	Intervalo	293,036	21.31%
19	**No_Of_Cylinders**	Cuantitativa	Intervalo	293,036	19.19%
20	**Seating_Capacity**	Cuantitativa	Intervalo	293,036	25.19%
21	**Tare_Weight**	Cuantitativa	Intervalo	293,036	13.92%
22	**Total_No_Occupants**	Cuantitativa	Intervalo	293,036	0.00%
23	**Carry_Capacity**	Cuantitativa	Intervalo	293,036	79.31%
24	**Cubic_Capacity**	Cuantitativa	Intervalo	293,036	88.49%
25	**Final_Direction**	Cualitativa	Nominal	293,036	0.00%
26	**Driver_Intent**	Cualitativa	Nominal	293,036	0.00%
27	**Vehicle_Movement**	Cualitativa	Nominal	293,036	0.00%
28	**Trailer_Type**	Cualitativa	Nominal	293,036	0.00%
29	**Vehicle_Colour_1**	Cualitativa	Nominal	293,036	0.00%
30	**Vehicle_Colour_2**	Cualitativa	Nominal	293,036	0.00%
31	**Caught_Fire**	Cualitativa	Nominal	293,036	0.00%
32	**Initial_Impact**	Cualitativa	Nominal	293,036	0.00%
33	**Lamps**	Cualitativa	Intervalo	293,036	0.00%
34	**Level_Of_Damage**	Cualitativa	Ordinal	293,036	0.00%
35	**Owner_Postcode**	Cualitativa	Nominal	293,036	10.22%
36	**Towed_Away_Flag**	Cualitativa	Nominal	293,036	0.00%

37	**Traffic_Control**	Cualitativa	Nominal	293,036	0.00%
38	**Traffic Control Desc**	Cualitativa	Nominal	293,036	0.00%

Las estadísticas descriptivas principales de las variables numéricas se presentan en el Cuadro 5.

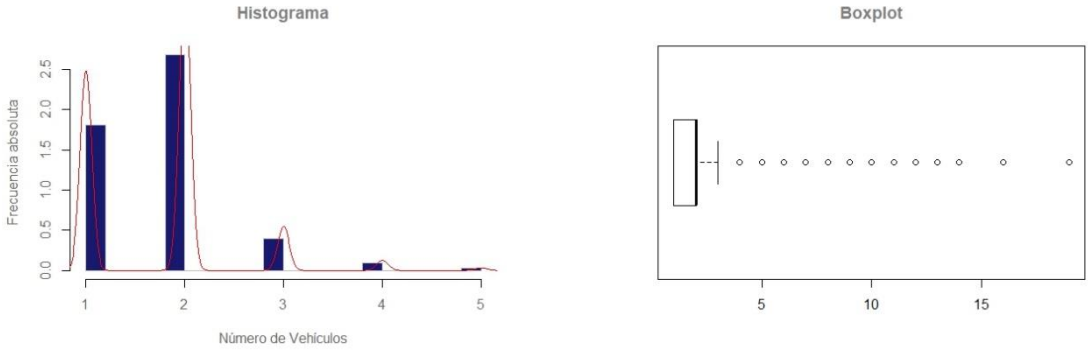
Cuadro 5: Estadísticas descriptivas de las variables numéricas

Tabla	Variable	N	Media	Desv. Stand.	Mín	Máx
Accidentes	**NO_OF_VEHICLES**	164,789	1.8	0.7	1	19
Accidentes	**NO_PERSONS**	164,789	2.4	1.5	1	97
Accidentes	**NO_PERSONS_INJ_2**	164,789	0.4	0.6	0	16
Accidentes	**NO_PERSONS_INJ_3**	164,789	0.9	0.8	0	26
Accidentes	**NO_PERSONS_KILLED**	164,789	0.02	0.2	0	11
Accidentes	**NO_PERSONS_NOT_INJ**	164,789	1.1	1.3	0	87
Personas	**EDAD**	164,789	37.05	19	0	109
Vehículos	**Vehicle_Year_Manuf**	164,789	1,930	372.42964	0	3001
Vehículos	**Vehicle_Weight**	164,789	5,064	6473.5416	0	90000
Vehículos	**No_Of_Wheels**	164,789	4	0.6750115	0	61
Vehículos	**No_Of_Cylinders**	164,789	5	1.5009065	0	93
Vehículos	**Seating_Capacity**	164,789	5	3.1026813	0	70
Vehículos	**Tare_Weight**	164,789	1,609	1680.8287	0	96000
Vehículos	**Total_No_Occupants**	164,789	1	0.9127757	0	96
Vehículos	**Carry_Capacity**	164,789	1,945	3066.2067	0	88200
Vehículos	**Cubic_Capacity**	164,789	162	257	0	999

A continuación, se hace un análisis de las distribuciones que siguen las principales variables.

El Número de Vehículos es una variable numérica discreta que evidentemente no sigue ninguna distribución continua, aunque podemos proponer una distribución Poisson.

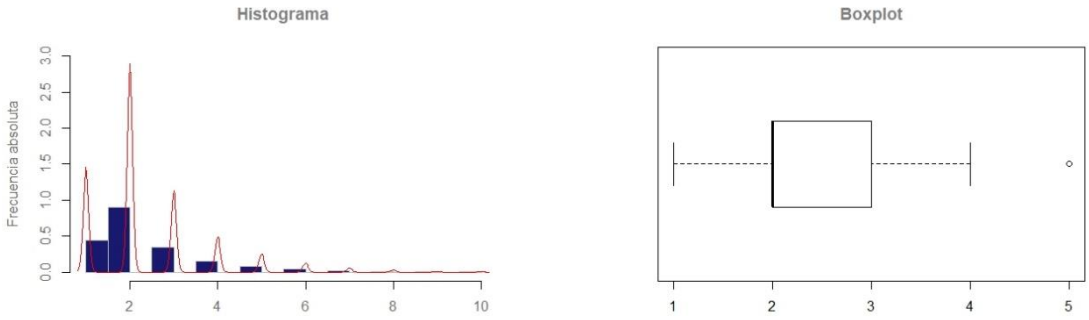
Figura 5: Número de Vehículos



Con todos los datos de la variable

El Número de personas involucradas en el accidente es también una variable discreta y como se observa en la gráfica de Box-Plot, la distribución acumula su mayor probabilidad en los valores 2 y 3.

Figura 6: Número de Personas involucradas en el accidente

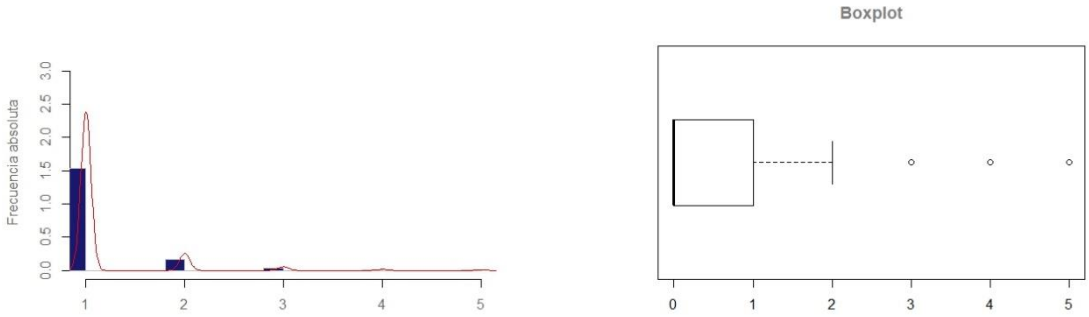


Dejando fuera algunos *outliers*

En cuanto al Número de Personas lesionadas, podemos observar que la gran mayoría de los accidentes sólo tiene una persona lesionada. Notemos también la presencia de varios *outliers* en las gráficas BoxPlot.

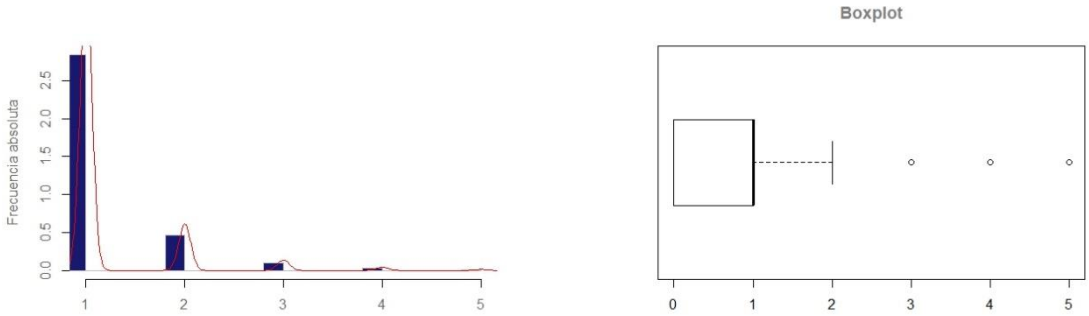
Figura 7: Número de Personas Lesionadas

(a) En el vehículo principal



Dejando fuera algunos *outliers*

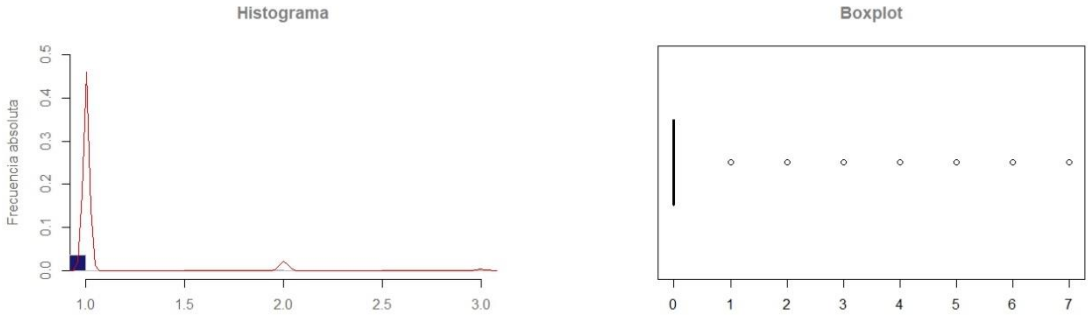
(b) En el vehículo secundario



Dejando fuera algunos *outliers*

Similarmente, para el Número de Personas Fallecidas, se tiene que es, en promedio 0, aunque se observan muchos outliers.

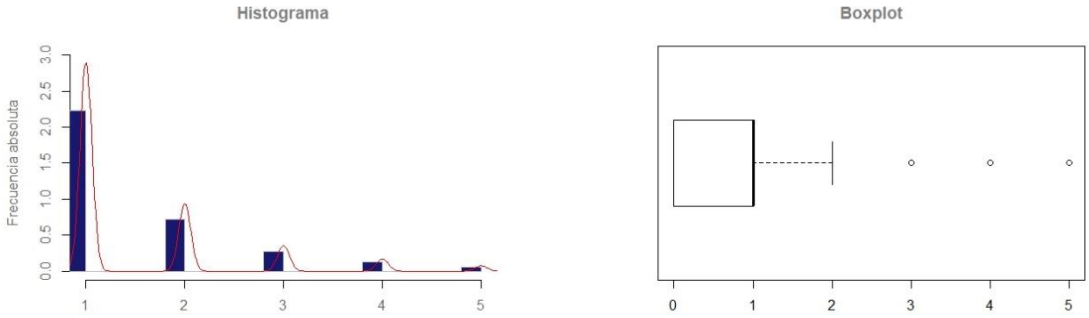
Figura 8: Número de Personas Fallecidas en el accidente



Dejando fuera algunos *outliers*

Para el número de Personas No Lesionadas es una, en promedio. Aunque observando el histograma podemos notar que también hay varios casos donde el número se eleva a 2 y 3.

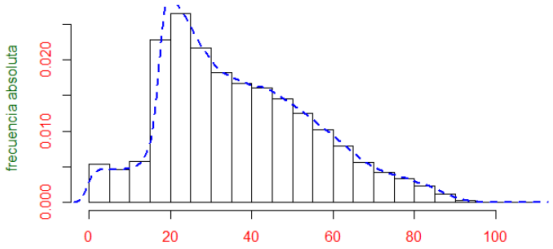
Figura 9: Número de Personas NO lesionadas



Dejando fuera algunos *outliers*

La edad promedio de las personas involucradas en los accidentes es de 37 años. Sin embargo, podemos ver en el histograma que la distribución de las edades no se parece a una normal, pues tiene *colas* más pesadas hacia la derecha y la media no está centrada.

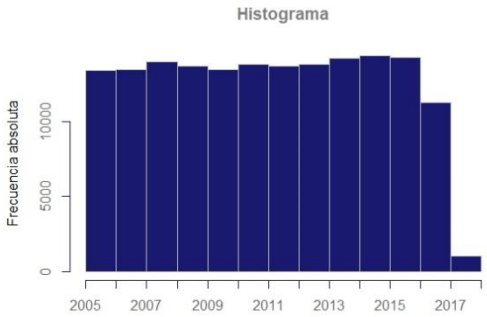
Figura 10. Edad de las Personas involucradas en los accidentes



En cuanto a la temporalidad de los accidentes, podemos observar en el cuadro siguiente que durante el periodo 2016-2017, el año de 2017 fue el año con menor número de accidentes, registrándose solo 6.9%. Por otro lado, 2015 fue el año con el mayor porcentaje de accidente en el periodo. En cuanto al mes, también podemos notar que el mayor porcentaje de accidentes sucedieron en el mes de Marzo, mientras que Junio y Septiembre son los meses de menor intensidad.

Cuadro 6: Años de los accidentes

Año	Porcentaje
2006	8.20%
2007	8.20%
2008	8.50%
2009	8.30%
2010	8.20%
2011	8.40%
2012	8.30%
2013	8.40%
2014	8.70%
2015	8.80%
2016	8.70%
2017	6.90%
2018	0.60%



Cuadro 7: Mes de los accidentes

Mes	Porcentaje
Enero	8.20%
Febrero	8.60%
Marzo	9.20%
Abril	8.40%
Mayo	8.70%
Junio	7.90%
Julio	8.00%
Agosto	8.00%
Septiembre	7.60%
Octubre	8.70%
Noviembre	8.50%
Diciembre	8.30%

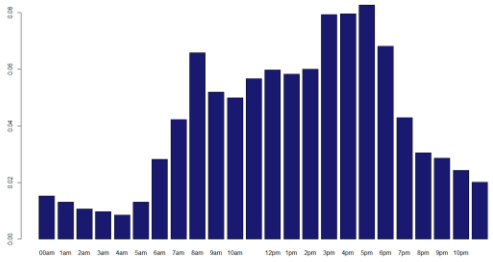
En cuanto al Día de la semana, podemos notar que los días viernes ocurrieron el mayor porcentaje de accidentes, con 15.8% del total. Mientras que los Domingos fueron los días con menor número de accidentes, reportándose solo el 12.6%. También podemos observar que las horas más conflictivas son alrededor de las 8 por la mañana y entre las 4 y las 5 por la tarde.

Cuadro 8: (a)Día de la semana

Día de la semana	Porcentaje
Lunes	13.40%
Martes	14.20%
Miércoles	14.60%
Jueves	15.10%
Viernes	15.80%
Sabado	14.30%
Domingo	12.60%

Cuadro 9: (b) Hora del día

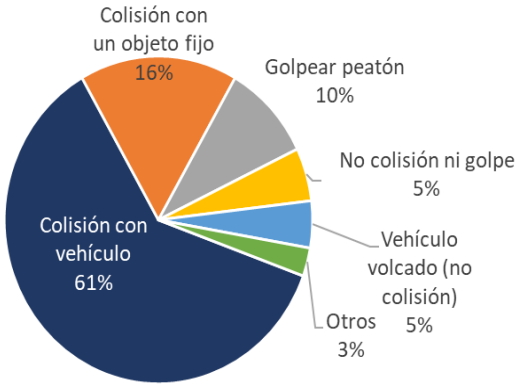
Hora	Porcentaje	Hora	Porcentaje
12:00 AM	1.50%	**12:00 PM**	6.00%
01:00 AM	1.30%	**01:00 PM**	5.80%
02:00 AM	1.10%	**02:00 PM**	6.00%
03:00 AM	1.00%	**03:00 PM**	7.90%
04:00 AM	0.90%	**04:00 PM**	8.00%
05:00 AM	1.30%	**05:00 PM**	8.30%
06:00 AM	2.80%	**06:00 PM**	6.80%
07:00 AM	4.20%	**07:00 PM**	4.30%
08:00 AM	6.60%	**08:00 PM**	3.10%
09:00 AM	5.20%	**09:00 PM**	2.90%
10:00 AM	5.00%	**10:00 PM**	2.40%
11:00 AM	5.70%	**11:00 PM**	2.00%



El 61% de los accidentes, se refieren a Colisiones con auto, aunque la base también incluye colisiones con otros objetos, vehículos volcados y otro tipo de golpes. Información detallada se observa en la siguiente tabla.

Cuadro 10: Tipo de Accidente (TYPE_DESC)

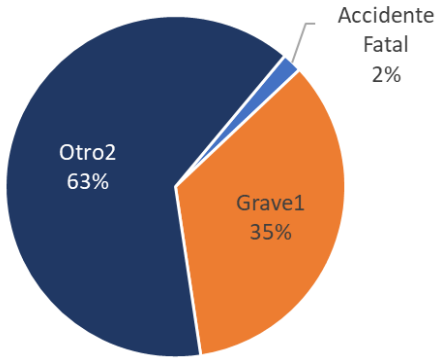
Tipo de Accidente	Porcentaje
Colisión con un objeto fijo	16.60%
Colisión con otro objeto	1.00%
Colisión con vehículo	61.20%
Caer desde o en un vehículo en movimiento	0.90%
Golpear animal	0.90%
Golpear peatón	9.60%
No colisión ni golpe	5.30%
Vehículo volcado (no colisión)	4.60%
Otros accidentes	0.10%



De la siguiente gráfica podemos ver que, aunque se producen muchos accidentes, solo el 36.5% tienen fatalidad grave o que haya requerido hospitalización.

Cuadro 11: Severidad del Accidente

Severidad	Porcentaje
Accidente Fatal	1.90%
Grave[1]	34.60%
Otro[2]	63.50%
Sin lesionados	0.00%



De la tabla de personas, podemos observar que en un 55.8% de los accidentes, están involucradas personas del sexo masculino, mientras que en un 39.9% se trata de personas del sexo femenino.

Cuadro 12: Sexo

Sexo	Porcentaje
Femenino	39.90%
Masculino	55.80%
Desconocido	4.30%

También observamos que en el 70% de los casos, las licencias de conducir de los involucrados fueron expedidas localmente (en Victoria).

Cuadro 13: Licencia

Licencia_edo	Porcentaje
Act	0.06%
Commonwealth	0.00%
Northern Territory	0.03%
New Soth Wales	0.64%
Overseas	1.01%
Queensland	0.27%
South Australia	0.26%
Tasmania	0.05%
Victoria	70.04%
West Australia	0.12%
Desconocido	22.17%
No disponible	5.36%

Aunque la base tenía un campo de descripción acerca de lo que estaba haciendo el peatón antes del accidente, esta información solo aplica para menos del 5% de los casos. Por lo tanto, se consideró que esta variable no era relevante para el modelo en general.

Cuadro 14: Acción del peatón antes del accidente

Pedest_Mov	Porcentaje
Not applicable	95.70%
Crossin carriageway	2.90%
Working, playing..	0.30%
Walking on carriageway with traffic	0.20%
Walking on carriageway against traffic	0.10%
Pushing or working on vehicle	0.10%
Pushing or working on vehicle	0.10%
Walking to from or boarding tram	0.10%
Walking to, from or boarding other vehicle	0.40%
Not on carriageway	0.10%
Not kwown	0.00%

Se contaba con la variable *Seating Position*, que indica la posición de los individuos al momento del accidente. En el 74% de los casos se trataba de los conductores.

También se indica el tipo de seguridad de los individuos en los diferentes vehículos. Cabe destacar que el 57.4% de los casos, las personas sí llevaban el cinturón de seguridad.

Cuadro 15: Posición de la persona dentro del vehículo

Seating Pos	Porcentaje
Cen_Front	0.10%
Cen_Rear	1.20%
Driver	74.30%
Lef_Front	13.70%
Lef_Rear	4.10%
NotApplicable	
Not known	1.80%
Other_rear	1.40%
Pillion_Pass	0.30%
PS	0.00%
RR	3.10%

Cuadro 16: La persona llevaba cinturón de seguridad o casco

SeatBealt	Porcentaje
Seatbelt worn	57.40%
Seatbelt no worn	1.90%
ChildRestreain worn	2.10%
ChildRestreaintNotWorn	0.00%
Seatbelt/not fitted	0.40%
CrashHelmetWorn	8.10%
CrashHelmetNotWorn	0.50%
NotAppropiate	5.20%
NotKnown	24.20%

La variable tipo de usuario de la carretera, indica que el 60% fueron conductores y un 23% pasajeros. Solo el 36% de los registros tenía información acerca de la variable Llevados al hospital, por lo que se consideró no relevante para el modelo.

Cuadro 17: Tipo de usuario de la carretera

Road_user_type	Porcentaje
Pedestrians	4.50%
Drivers	59.70%
Passengers	23.30%
Motorcyclists	6.20%
Pillion Passengers	0.30%
Bicyclists	4.40%
Drivers	0.20%
Passengers	0.20%
Unknown	1.30%

Cuadro 18: Llevados al hospital

Taken_Hosp	Porcentaje
Yes	17.70%
No	18.70%
Unknown	63.60%

Para los códigos DCA del vehículo la frecuencia mayor fue de 56.19% y la frecuencia menor fue de 0.04% como se muestra en el siguiente cuadro

Cuadro 19: Código DCA del vehículo

VEHICLE_DCA_CODE	VEHICLE_DCA_CODE_DESC	FRECUENCIA
1	**Vehicle 1**	56.19%
2	**Vehicle 2**	34.10%
3	**Not known which vehicle was number 1**	0.03%
8	**Not involved in initial event**	9.63%
vacíos		0.04%

Para la dirección inicial la frecuencia mayor es de 17.59% y la frecuencia menor es de 1.93% como se muestra en el siguiente cuadro

Cuadro 20: Dirección inicial

INITIAL_DIRECTION	DIRECTION	INITIAL_DIRECTION_DESC	FINAL_DIRECTION
SW	**South-west**	6.63%	6.78%
NW	**North-west**	7.89%	7.97%
S	**South**	17.59%	16.96%
N	**North**	17.23%	16.43%
E	**East**	17.05%	16.57%
W	**West**	16.91%	16.42%
SE	**South-east**	8.13%	8.31%
NE	**North-east**	6.64%	6.77%
NK	**Unknow**	1.93%	3.79%

En el caso del tipo de superficie de la carretera la frecuencia mayor fue del 94.93% y la frecuencia menor fue de 0.64

Cuadro 21: Tipo de superficie de la carretera

ROAD_SURFACE_TYPE	ROAD_SURFACE_TYPE_DESC	FRECUENCIA
1	**Paved**	94.93%
2	**Unpaved**	0.64%
3	**Gravel**	3.78%
9	**Unknown**	0.65%

La frecuencia del estado de registró, el estado con una frecuencia mayor es Victoria, y el estado de registro con menor frecuencia es Commonwealth, como se muestra en el siguiente cuadro

Cuadro 22: Estado reg

REG_STATE	REG_STATE_DESC	FRECUENCIA
A	**ACT**	0.08%
B	**Commonwealth**	0.00%
D	**Northern Territory**	0.02%
N	**New South Wales**	1.03%
O	**Overseas**	0.00%
Q	**Queensland**	0.50%
S	**South Australia**	0.49%
T	**Tasmania**	0.10%
V	**Victoria**	88.80%
W	**West Australia**	0.17%
Z	**Not known**	1.00%
	not available	7.82%

Las frecuencias para el tipo de vehículo con mayor frecuencia fue el Car con un 53.55% y con el menor porcentaje de frecuencia fue Parked trailers con un 0.02%

Cuadro 23: Tipo de vehículo

VEHICLE_TYPE	VEHICLE_TYPE_DESC	FRECUENCIA
1	**Car**	53.55%
2	**Station Wagon**	14.87%
3	**Taxi**	1.25%
4	**Utility**	7.41%

5	**Panel Van**	2.22%
6	**Prime Mover (No of Trailers Unknown)**	0.22%
7	**Rigid Truck(Weight Unknown)**	0.06%
8	**Bus/Coach**	0.52%
9	**Mini Bus(9-13 seats)**	0.08%
10	**Motor Cycle**	7.90%
11	**Moped**	0.07%
12	**Motor Scooter**	0.47%
13	**Bicycle**	5.93%
14	**Horse (ridden or drawn)**	0.01%
15	**Tram**	0.25%
16	**Train**	0.03%
17	**Other Vehicle**	0.25%
18	**Not Applicable**	0.01%
19	**Parked trailers**	0.02%
20	**Quad Bike**	0.02%
27	**Plant machinery and Agricultural equipment**	0.03%
60	**Prime Mover Only**	0.15%
61	**Prime Mover - Single Trailer**	0.72%
62	**Prime Mover B-Double**	0.29%
63	**Prime Mover B-Triple**	0.01%
71	**Light Commercial Vehicle (Rigid) <= 4.5 Tonnes GVM**	0.94%
72	**Heavy Vehicle (Rigid) > 4.5 Tonnes**	1.29%
99	**Unknown**	1.46%

Las frecuencias para el control de tráfico mayor fueron del 63.69% y con menor es de 0%

Cuadro 24: Control de trafico

TRAFFIC_CONTROL	TRAFFIC_CONTROL_DESC	FRECUENCIA
0	**No control**	63.69%
1	**Stop-go lights**	18.67%
2	**Flashing lights**	0.12%
3	**Out of order**	0.12%
4	**Ped. lights**	0.30%
5	**Ped. crossing**	0.62%
6	**RX Gates/Booms**	0.13%
7	**RX Bells/Lights**	0.02%
8	**RX No control**	0.02%

9	**Roundabout**	3.40%
10	**Stop sign**	2.26%
11	**Giveway sign**	6.23%
12	**School Flags**	0.05%
13	**School No flags**	0.03%
14	**Police**	0.10%
15	**Other**	1.05%
99	**Unknown**	3.19%
	Unknown	0.00%

La intención del conductor con mayor frecuencia fue del 60.22% y con menor es de 0%, como se muestra en el siguiente cuadro

Cuadro 25: Intención del conductor

DRIVER_INTENT	FRECUENCIA
1	60.22%
2	13.13%
3	4.43%
4	1.37%
5	1.28%
6	1.63%
7	0.69%
8	0.47%
9	0.53%
10	0.68%
11	3.19%
12	0.12%
13	0.13%
14	0.13%
15	2.90%
16	0.37%
17	4.92%
18	0.32%
19	0.05%
99	3.43%
vacíos	0.00%

El movimiento del vehículo con mayor frecuencia es de 48.41% y con menor frecuencia es de 0%, se muestra en el siguiente cuadro

Cuadro 26: Movimiento del vehículo

VEHICLE_MOVEMENT	FRECUENCIA
1	48.41%
2	11.04%
3	3.33%
4	1.02%
5	1.18%
6	2.14%
7	0.74%
8	0.55%
9	0.73%
10	0.48%
11	3.07%
12	0.13%
13	0.77%
14	0.14%
15	5.49%
16	0.58%
17	7.15%
18	10.71%
19	0.18%
99	2.16%
vacíos	0.00%

La frecuencia mayor del tipo de tráiler es de 98.043% y con menor es de 0.001%, como se puede ver en el siguiente cuadro

Cuadro 27: Tipo de tráiler

TRAILER_TYPE	FRECUENCIA
	0.001%
A	0.101%
B	1.090%
C	0.058%
D	0.037%
E	0.059%
F	0.052%

G	0.121%
H	98.043%
I	0.024%
J	0.337%
K	0.013%
L	0.063%

Para el color del vehículo con mayor frecuencia es de 12.724%, se muestra en el siguiente cuadro

Cuadro 28: Color del vehículo

COLOR	VEHICLE_COLOUR_1	VEHICLE_COLOUR_2
BLK	10.495%	0.536%
BLU	12.740%	0.263%
BRN	0.679%	0.023%
CRM	0.225%	0.016%
FWN	0.654%	0.019%
GLD	2.357%	0.026%
GRN	5.625%	0.215%
GRY	6.194%	0.123%
MRN	1.446%	0.020%
MVE	0.016%	0.000%
OGE	0.863%	0.058%
PNK	0.075%	0.004%
PUR	0.463%	0.026%
RED	10.029%	0.244%
SIL	14.971%	0.399%
WHI	22.437%	0.565%
YLW	2.338%	0.109%
ZZ	8.392%	97.353%
vacíos	0.000%	0.001%

Las frecuencias para los vehículos que se llegan a incendiar es de 92.628%

Cuadro 29: Se incendió

CAUGHT_FIRE	FRECUENCIA
0	6.094%
1	0.335%
2	92.628%
9	0.943%

En el impacto inicial con mayor frecuencia fue de 37.366%, y con menor frecuencia fue de 0.012%

Cuadro 30: Impacto inicial

INITIAL_IMPACT	FRECUENCIA
0	0.383%
1	11.549%
2	4.293%
3	1.311%
4	3.235%
5	11.305%
6	3.070%
7	1.326%
8	2.828%
9	5.456%
F	37.366%
N	2.608%
R	14.440%
S	0.027%
T	0.555%
U	0.236%
vacíos	0.012%

La frecuencia mayor del nivel de daño del vehículo es de 21.885%, como se muestra en el siguiente cuadro

Cuadro: Nivel de daño

LEVEL_OF_DAMAGE	FRECUENCIA
1	21.885%
2	13.873%
3	18.132%
4	18.106%
5	13.157%
6	8.559%
9	6.288%

3.2.3 Conclusión

De las 85 variables disponibles en la base, y después del análisis exploratorio, hicimos una tabla consolidada con la información de las tres tablas (Accidentes, Vehículos y Personas) con los atributos que consideramos más asociados al riesgo de colisión. Esta tabla contenía información para 38 atributos y con aproximadamente 390 mil registros.

Sin embargo, realizamos nuevamente un análisis de las variables y logramos reducir a las que se muestran en el Cuadro 19, pues encontramos que varias variables se referían a información relacionada.

Cuadro 31: Atributos más ligados al riesgo de colisión

No.	Tabla original	Atributo
1	Accidentes	**ACCIDENT_DATE**
2	Accidentes	**DAY_WEEK_DESC**
3	Accidentes	**TYPE_DESC**
4	Accidentes	**DCA_DESC**
5	Accidentes	**LIGHT_CONDITION_DESC**
6	Accidentes	**ROAD_GEOMETRY_DESC**
7	Accidentes	**SPEED_ZONE**
8	Vehiculos	**ROAD_SURFACE_TYPE_DESC**
9	Personas	**AGE**
10	Personas	**LICENCE_STATE_DESC**
11	Personas	**HELMET_BELT_WORN**
12	Personas	**SEX**
13	Vehículos	**VEHICLE_YEAR_MANUF**
14	Vehículos	**VEHICLE_MAKE**
15	Vehículos	**VEHICLE_TYPE_DESC**
16	Vehículos	**NO_OF_CYLINDERS**
17	Vehículos	**TOTAL_NO_OCCUPANTS**
18	Vehículos	**COLOR**
19	Vehículos	**TRAFFIC_CONTROL_DESC**

En el Cuadro 20 se muestran las variables que parecen estar menos ligados al riesgo de colisión.

Por ejemplo, aunque la hora del accidente (ACCIDENT_TIME) está ligada al riesgo de colisión, nosotros consideramos que no era por la hora en sí, sino más bien por la presencia o ausencia de luz. Así que, en lugar de considerar esa variable para el modelo, consideramos si había luz en el lugar (LIGHT_CONDITION_DESC).

Las variables que se refieren a información de personas (variables 4-7,12, 15 y16 del Cuadro 20) no parecen estar directamente ligadas al riesgo de colisión, pues se trata de las

personas una vez sucedido el accidente. Similarmente, con las variables en los renglones 7 y 8 del Cuadro 20.

Aunque el modelo del vehículo (10 en el Cuadro 20) puede ser un factor asociado al riesgo de colisiones, esta variable está muy relacionada con la marca del vehículo. Preferimos quedarnos con la marca y no con e modelo para reducir el número de categorías con las que trabaja el modelo.

En lo que se refiere a la variable peso del vehículo (TARE_WEIGHT), está relacionada con el tipo de vehículo.

Similarmente, nosotros consideramos la variable Age en vez de la Age_group.

Cuadro 32: Atributos menos ligados al riesgo de colisión

No.	Tabla Original	Atributo
1	Accidentes	ACCIDENT_TIME
2	Accidentes	NO_OF_VEHICLES
3	Accidentes	NO_PERSONS
4	Accidentes	NO_PERSONS_INJ_2
5	Accidentes	NO_PERSONS_KILLED
6	Accidentes	NO_PERSONS_INJ_3
7	Accidentes	NO_PERSONS_NOT_INJ
8	Accidentes	POLICE_ATTEND
9	Accidentes	SEVERITY
10	Vehículos	VEHICLE_DCA_CODE_DESC
11	Vehículos	VEHICLE_MODEL
12	Vehículos	VEHICLE_COLOUR_2
13	Vehículos	LEVEL_OF_DAMAGE
14	Vehículos	TARE_WEIGHT
15	Personas	SEATING_POSITION_DESC
16	Personas	AGE_GROUP
17	Personas	INJ_LEVEL_DESC
18	Personas	TAKEN_HOSPITAL

También usamos Weka para saber si nuestra selección de variables estaba bien caracterizada.

Primero usamos el evaluador ClassifierSubsetEval con el método BestFit y GreedyStepwise. Las variables seleccionadas fueron 4 (HORA_ACCIDENTE, DCA_DESC, ROAD_SURFACE Y TIPO_VEHICULO) y las mismas para ambos métodos. Se muestran en el siguiente cuadro.

Cuadro 33. Salidas de Weka para la Selección de variables

(a) Método BestFit

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 104
        Merit of best subset found:    0.276

Attribute Subset Evaluator (supervised, Class (nominal): 1 ï..CHOQUE):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,4,9,11 : 4
                     HORA_ACCIDENTE
                     DCA_DESC
                     ROAD_SURFACE_TYPE_DESC
                     TIPO_VEHICULO
```

(b) Método Greedy Stepwise

```
=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.276

Attribute Subset Evaluator (supervised, Class (nominal): 1 ï..CHOQUE):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,4,9,11 : 4
                     HORA_ACCIDENTE
                     DCA_DESC
                     ROAD_SURFACE_TYPE_DESC
                     TIPO_VEHICULO
```

También usamos la selección de atributos ***CorrelationAttributeEval*** con el método ***Ranker***. Según este método las variables de peso mayor a 0.1 son: 3 de las variables seleccionadas por los métodos anteriores (*DCA_DESC, ROAD_SURFACE* Y TIPO_VEHICULO) y 4 variables adicionales (ROAD_GEOMETRY, SEÑAL_TRAFICO, CILINDROS, SEGURIDAD y LIGHT_CONDITION).

Los resultados de WEKA se muestran en el Cuadro 34.

Cuadro 34. Salida de Weka para la selección de atributos con el método Ranker

```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     Choques2_VF
Instances:    350650
Attributes:   20
              i..CHOQUE
              HORA_ACCIDENTE
              DAY_WEEK_DESC
              DCA_DESC
              LIGHT_CONDITION_DESC
              ROAD_GEOMETRY_DESC
              SPEED_ZONE
              VEHICLE_YEAR_MANUF
              ROAD_SURFACE_TYPE_DESC
              MARCA
              TIPO_VEHICULO
              Cilindros
              INDICADOR_OCUPANTES
              COLOR
              GRUPO_EDAD
              SEÃ.AL_TRAFICO
              LICENCE_STATE_DESC
              SEGURIDAD
              SEX
              GRUPO_EDAD2
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 1 i..CHOQUE):
        Correlation Ranking Filter
Ranked attributes:
 0.3069    4 DCA_DESC
 0.2747    6 ROAD_GEOMETRY_DESC
 0.1899   16 SEÃ.AL_TRAFICO
 0.1634    9 ROAD_SURFACE_TYPE_DESC
 0.1459   12 Cilindros
 0.1278   18 SEGURIDAD
 0.1026    5 LIGHT_CONDITION_DESC
 0.1011   11 TIPO_VEHICULO
 0.0793    2 HORA_ACCIDENTE
 0.0712   17 LICENCE_STATE_DESC
 0.0685   19 SEX
 0.0629    7 SPEED_ZONE
 0.0489   13 INDICADOR_OCUPANTES
 0.0356   15 GRUPO_EDAD
 0.0353   10 MARCA
 0.0332   20 GRUPO_EDAD2
 0.0289    3 DAY_WEEK_DESC
 0.0225    8 VEHICLE_YEAR_MANUF
 0.01     14 COLOR

Selected attributes: 4,6,16,9,12,18,5,11,2,17,19,7,13,15,10,20,3,8,14 : 19
```

Las variables que selecciona este método son las 19 que estamos contemplando, por lo que nos quedamos con esta base.

3.3. Preparación de Datos

Del proceso anterior nos quedamos con una tabla consolidada de 19 atributos y 390 mil registros. Nuestra variable por predecir es el riesgo de colisiones.

Cuadro 35: Justificación de la relación entre los atributos y el riesgo de colisión

No.	Atributo	Justificación
1	**ACCIDENT_DATE**	Hay meses que muestran mayor número de incidencias que otros
2	**DAY_WEEK_DESC**	El fin de semana es el día que mostro históricamente más accidentes
3	**TYPE_DESC**	Lo usamos para definir nuestra variable Riesgo de Colisión
4	**DCA_DESC**	El tipo de intersección en el qué sucedió el accidente puede incidir en los accidentes
5	**LIGHT_CONDITION_DESC**	Las condiciones de luz pueden elevar el riesgo de los accidentes
6	**ROAD_GEOMETRY_DESC**	La forma de la carretera (si es curva, recta, etc) puede elevar el riesgo de colisión
7	**SPEED_ZONE**	La velocidad permitida puede influir en el número de accidentes
8	**ROAD_SURFACE_TYPE_DESC**	El tipo de pavimento puede influir en el riesgo
9	**AGE**	La edad, si los conductores jóvenes son más propensos a asumir más riesgos
10	**LICENCE_STATE_DESC**	Si los conductores foráneos causan más accidentes al no conocer las reglas locales
11	**HELMET_BELT_WORN**	Si no usas algún tipo de seguridad puedes ser más propenso a asumir riesgos al conducir
12	**SEX**	Posiblemente los hombres son más propensos a sufrir accidentes
13	**VEHICLE_YEAR_MANUF**	Entre más antiguo el auto puede ser más propenso a tener accidentes
14	**VEHICLE_MAKE**	Algunos modelos/marcas de autos se han identificado con mayor número de accidentes
15	**VEHICLE_TYPE_DESC**	Los autos tienen mayor número de accidentes en comparación con las bicicletas o los autos
16	**NO_OF_CYLINDERS**	Está relacionada con la velocidad que pueden alcanzar los autos
17	**TOTAL_NO_OCCUPANTS**	Entre más pasajeros dentro del vehículo, puede haber mayor riesgo de accidentes
18	**COLOR**	Posiblemente el color del automóvil aumente el riesgo de colisión
19	**TRAFFIC_CONTROL_DESC**	La señalización de tránsito, o falta de, en el lugar puede influir en la incidencia de los accidentes

Una vez con esta selección de atributos, procedimos a corregir la base por valores perdidos y por valores atípicos.

3.3.1. Manejo de valores perdidos

Se identificaron 7 variables con diferentes proporciones de valores perdidos. Aunque solo algunas de ellas se consideraron para el modelo, se modificaron los valores perdidos para todas las variables.

Los valores perdidos se reemplazaron utilizando la moda en todos los casos.

Cuadro 36. Valores perdidos

Nombre de la Variable	No. de valores perdidos	% del total de la base	Valor por el que se sustituyo	En modelo[1]
AGE	17,786	4%	Moda	Sí
NO_OF_CYLINDERS	71,647	18%	Moda	Sí
TARE_WEIGHT	47,190	12%	Moda	Sí
VEHICLE_COLOUR_2	2	0%	Moda	No
NO_OF_VEHICLES	17,608	4%	Moda	No
VEHICLE_MODEL	40,243	10%	Moda	No
TAKEN_HOSPITAL	252,954	64%	Moda	No

1. Si la variable fue usada en el modelo
2. Tamaño original de la base consolidada: 397,884 registros

3.3.2. Manejo de valores atípicos

De las 19 variables de nuestra base final se identificaron 9 con valores atípicos muy grandes y que podían sesgar la información. Limpiamos la base eliminando los registros que contenían los valores atípicos de estas 9 variables, cuidando que los datos eliminados de cada variable no pasaran del 5%.

Así, los registros eliminados de todas las variables representan el 12% del total de registros en la base.

Cuadro 37. Valores atípicos eliminados

Nombre de la Variable	Condición de atípico	No. de registros eliminados	% del total de la base	Tamaño de base resultante	En modelo[1]
NO_OF_VEHICLES	> 3	21,732	5.50%	376,152	No
NO_PERSONS	>= 10	3,685	0.90%	372,467	No
VEHICLE_YEAR_MANUF	> 2018 y < 1980	13,354	3.40%	359,113	Sí
TOTAL_NO_OCCUPANTS	> 8	674	0.20%	358,439	Sí
AGE	> 90	520	0.10%	357,919	Sí
NO_OF_CYLINDERS	> 8	88	0.00%	357,831	Sí

NO_PERSONS_INJ_3	> 2	6,981	1.80%	350,850	No
NO_PERSONS_NOT_INJ	> 8	159	0.00%	350,691	No
NO_PERSONS_KILLED	> 3	51	0.00%	350,640	No

1. Si la variable fue usada en el modelo
2. Tamaño original de la base consolidada: 397,884 registros

Después de la eliminación de valores atípicos, la base resultante quedo con 350, 640 registros y 19 variables.

3.3.3. Reducción de Datos

Algunas de las variables originales contaban con muchas categorías. Así que normalizamos las variables para reducir el número de categorías en cada variable.

A continuación, mostramos las variables normalizadas y cuántas categorías quedaron después de la normalización. El detalle de cuales categorías quedaron se puede ver en el Anexo 1.

Cuadro 38. Variables normalizadas

No.	Nombre de la variable		No. Categorías	
	Antes	Después	Antes	Después
1	ACCIDENT_TIME	HORA_ACCIDENTE	+12	5
2	DCA_DESC	DCA_DESC	81	4
3	LIGHT_CONDITION_DESC	LIGHT_CONDITION_DESC	7	3
4	ROAD_GEOMETRY_DESC	ROAD_GEOMETRY_DESC	9	3
5	VEHICLE_YEAR_MANUF	VEHICLE_YEAR_MANUF	+10	3
6	VEHICLE_MAKE	MARCA	+20	8 (Top 8)
7	VEHICLE_TYPE_DESC	TIPO_VEHICULO	28	6
8	TOTAL_NO_PERSONS	INDICADOR_OCUPANTES	5	3
9	VEHICLE_COLOUR_1	COLOR	18	3
10	AGE	GRUPO_EDAD	100	5
11	TRAFFIC_CONTROL_DESC	SEÑAL_TRAFICO	17	3
12	LICENCE_STATE_DESC	LICENCE_STATE_DESC	12	3
13	HELMET_BELT_WORN	SEGURIDAD	6	4
14	NO_CYLINDERS	Cilindros	5	4
15	TYPE_DESC	Choque	9	2

3.3.4. Normalización de variables

Para nuestra base no aplicamos ningún tipo de normalización, pues nos quedamos con solo una variable numérica (**SPEED_ZONE**).

3.4. Modelado

Para nuestros modelos, se consideró el **riesgo de colisión** como variable de respuesta. El riesgo de colisión se definió mediante la variable tipo de accidente (TYPE_DESC). Se considero como ***Colisión*** todas las colisiones con vehículos, un objeto fijo u otros objetos y el golpear a peatones o animales.

Cuadro 39. Definición de la variable de respuesta

TYPE_DESC	Colisión
Collision with vehicle	SI
Fall from or in moving vehicle	NO
Collision with a fixed object	SI
Struck Pedestrian	SI
Collision with some other object	SI
Vehicle overturned (no collision)	NO
No collision and no object struck	NO
Struck animal	SI
Other accident	NO

Entonces, la distribución de la variable de respuesta **Colisión** en la base de datos resulto de la siguiente manera:

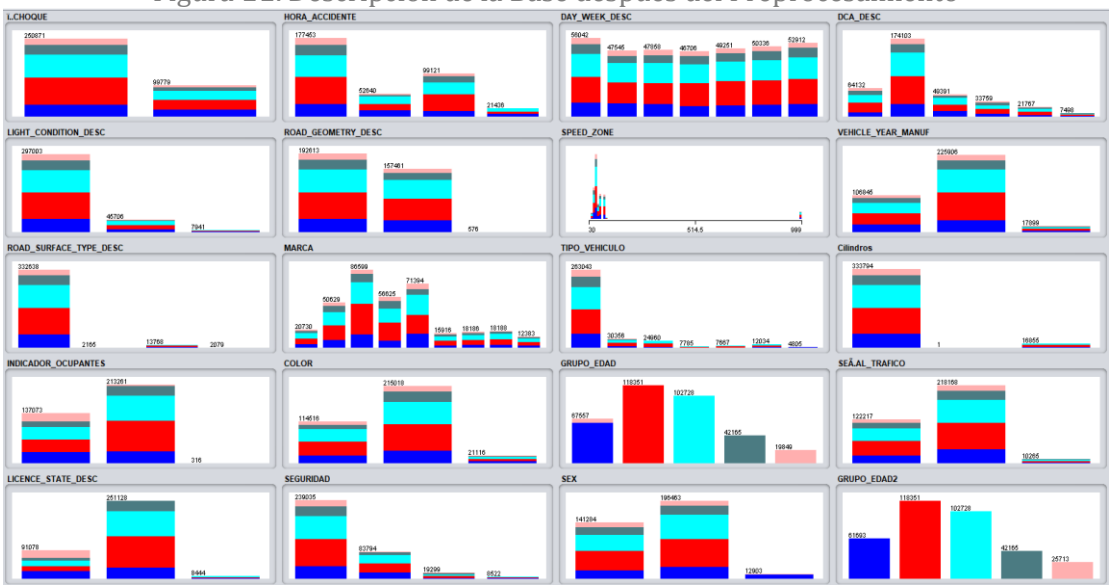
Cuadro 40. Distribución de la variable de respuesta

Colisión	No. de registros	Porcentaje
SI	250,871	71.5%
NO	99,779	28.5%

Como podemos observar, los registros están en casi en proporción 3 a 1.

La siguiente figura muestra los histogramas de los 19 atributos considerados más la variable de respuesta.

Figura 11. Descripción de la Base después del Preprocesamiento



Utilizamos Weka para correr los diferentes modelos

3.4.1 OneR

El clasificador OneR se corrió con el total de registros de la base (350,650) y los 20 atributos. El primer modelo, haciendo el Split de la base (2/3-1/3), es relativamente bueno pues logro clasificar el 88% de las instancias correctamente.

Cuadro 41. Salida de Weka para el modelo OneR Split – Modelo 1

```
Time taken to build model: 0.66 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.23 seconds

=== Summary ===

Correctly Classified Instances      104800               87.904  %
Incorrectly Classified Instances     14421               12.096  %
Kappa statistic                          0.6868
Mean absolute error                      0.121
Root mean squared error                  0.3478
Relative absolute error                 29.7035 %
Root relative squared error             77.0589 %
Total Number of Instances           119221

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,950    0,299    0,889      0,950    0,918      0,692    0,826     0,880     SI
               0,701    0,050    0,848      0,701    0,768      0,692    0,826     0,680     NO
Weighted Avg.  0,879    0,228    0,877      0,879    0,875      0,692    0,826     0,823

=== Confusion Matrix ===

     a     b   <-- classified as
 80985  4278 |     a = SI
 10143 23815 |     b = NO
```

El segundo modelo (Modelo 2) se corrió con 10 Folds y en comparación con el Modelo 1 no hubo cambios significativos. El modelo 2 también clasifico el 88% de los datos correctamente.

Cuadro 422. Salida de Weka para el modelo OneR 10 Folds – Modelo 2

```
=== Classifier model (full training set) ===

DCA_DESC:
        INTERSECCION    -> SI
        OTROS   -> SI
        ESTACIONADO     -> NO
        PEATON  -> NO
        NO REBASAR      -> SI
        REBASAR -> SI
(308529/350650 instances correct)


Time taken to build model: 0.38 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      308529               87.9877 %
Incorrectly Classified Instances     42121               12.0123 %
Kappa statistic                          0.6894
Mean absolute error                      0.1201
Root mean squared error                  0.3466
Relative absolute error                 29.5021 %
Root relative squared error             76.8142 %
Total Number of Instances           350650

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,949    0,294    0,890      0,949    0,919      0,695    0,827     0,881     SI
                0,706    0,051    0,847      0,706    0,770      0,695    0,827     0,681     NO
Weighted Avg.   0,880    0,225    0,878      0,880    0,876      0,695    0,827     0,824


=== Confusion Matrix ===

      a      b   <-- classified as
 238125  12746 |      a = SI
  29375  70404 |      b = NO
```

3.4.2 RIPPER (podado y sin podar)

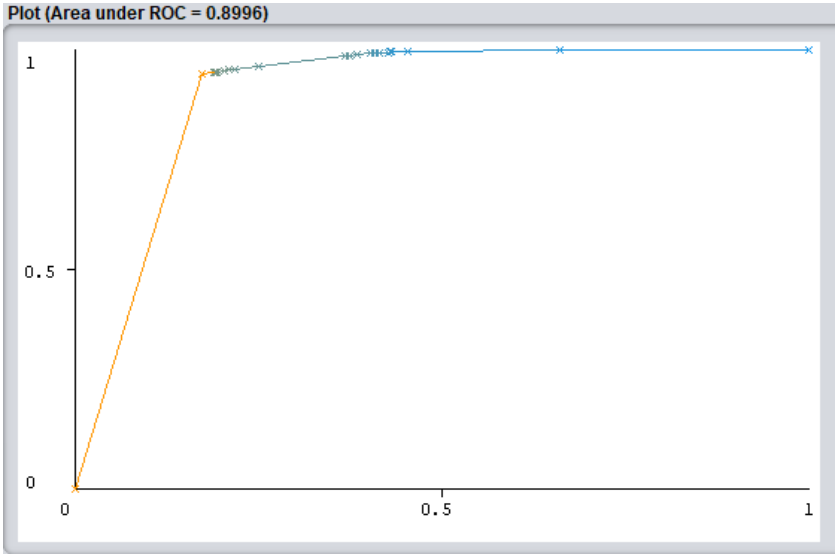
Para el modelo Ripper, se utilizó una muestra del 30% de la base pues no fue posible correr el 100% de la base en Weka para este modelo. El modelo *podado* identifico el 91% de los registros correctamente.

Cuadro 43. Salida de Weka para el modelo RIPER podado

```
=== Run information ===

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:     Choques2_Sample30%
Instances:    100000
Attributes:   20

Number of Rules : 44


Time taken to build model: 293.97 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.39 seconds

=== Summary ===

Correctly Classified Instances       30932               90.9765 %
Incorrectly Classified Instances      3068                9.0235 %
Kappa statistic                          0.7751
Mean absolute error                      0.1484
Root mean squared error                  0.2723
Relative absolute error                 36.5119 %
Root relative squared error             60.537  %
Total Number of Instances            34000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,826    0,057    0,850      0,826    0,838      0,775    0,900     0,842     NO
                 0,943    0,174    0,932      0,943    0,938      0,775    0,900     0,929     SI
Weighted Avg.    0,910    0,141    0,909      0,910    0,909      0,775    0,900     0,905

=== Confusion Matrix ===

     a     b   <-- classified as
  7907  1670 |     a = NO
  1398 23025 |     b = SI
```

El estimador Kappa es de 0.7751 que se puede interpretar como un modelo Sustancial, aunque el ROC es de 0.8996, lo que lo ubica como un modelo regular.

Figura 12. ROC para el modelo Ripper Podado



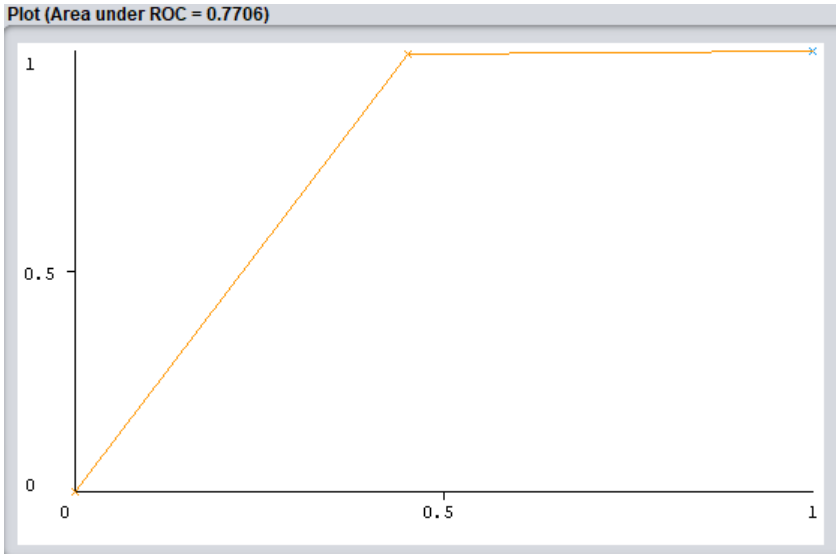
También corrimos el Ripper sin podar. En este caso, el número de instancias correctamente identificadas bajo a 87%.

Cuadro 44. Salida de Weka para el modelo RIPER sin podar

```
=== Run information ===

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1 -P
Relation:     Choques2_Sample30%
Instances:    100000
Attributes:   20
Time taken to build model: 85.78 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.23 seconds

=== Summary ===

Correctly Classified Instances       29494               86.7471 %
Incorrectly Classified Instances      4506               13.2529 %
Kappa statistic                          0.623
Mean absolute error                      0.2177
Root mean squared error                  0.3366
Relative absolute error                 53.5392 %
Root relative squared error             74.8349 %
Total Number of Instances            34000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,549    0,008    0,966      0,549    0,700      0,664    0,771     0,657     NO
                 0,992    0,451    0,849      0,992    0,915      0,664    0,771     0,848     SI
Weighted Avg.    0,867    0,326    0,882      0,867    0,854      0,664    0,771     0,794

=== Confusion Matrix ===

     a     b   <-- classified as
  5255  4322 |     a = NO
   184 24239 |     b = SI
```

El estimador Kappa es de 0.6230 que se puede interpretar como un modelo Sustancial, aunque el ROC, de 0.7706, lo ubica también como un modelo regular.

Figura 13. ROC para el modelo Ripper sin Podar



3.4.3 Árbol de decisión C4.5

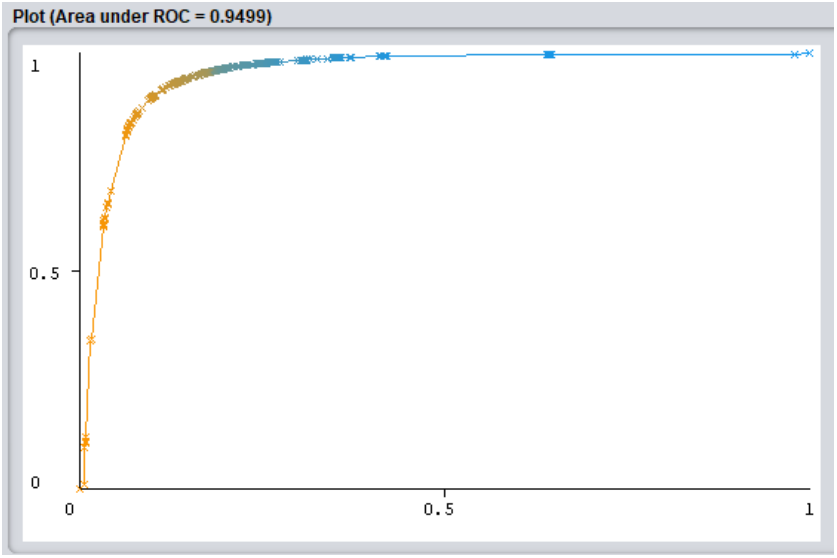
Para el árbol de decisión, nuestro primer intento fue correr el modelo con los valores por default. El modelo resultante clasifico el 92% de las instancias correctamente.

Cuadro 45. Salida de Weka para el Árbol C4.5 – Modelo 1

```
Number of Leaves  :     5036

Size of the tree :      6695


Time taken to build model: 25.82 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.41 seconds

=== Summary ===

Correctly Classified Instances      109353               91.7229 %
Incorrectly Classified Instances      9868                8.2771 %
Kappa statistic                          0.7924
Mean absolute error                      0.1186
Root mean squared error                  0.2541
Relative absolute error                 29.1282 %
Root relative squared error             56.2999 %
Total Number of Instances           119221

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,956    0,181    0,930      0,956    0,943      0,793    0,950     0,967     SI
               0,819    0,044    0,882      0,819    0,849      0,793    0,950     0,902     NO
Weighted Avg.  0,917    0,142    0,916      0,917    0,916      0,793    0,950     0,949

=== Confusion Matrix ===

     a     b   <-- classified as
 81540  3723 |     a = SI
  6145 27813 |     b = NO
```

El estimador Kappa (0.7924) y el ROC (0.9499) indican que es un modelo sustancial y excelente. Sin embargo, dado que el número mínimo de instancias en las hojas es de 0, el árbol resultante tiene 10 niveles y no es fácil de interpretar.

Figura 14. ROC para el Árbol C4.5 – Modelo 1



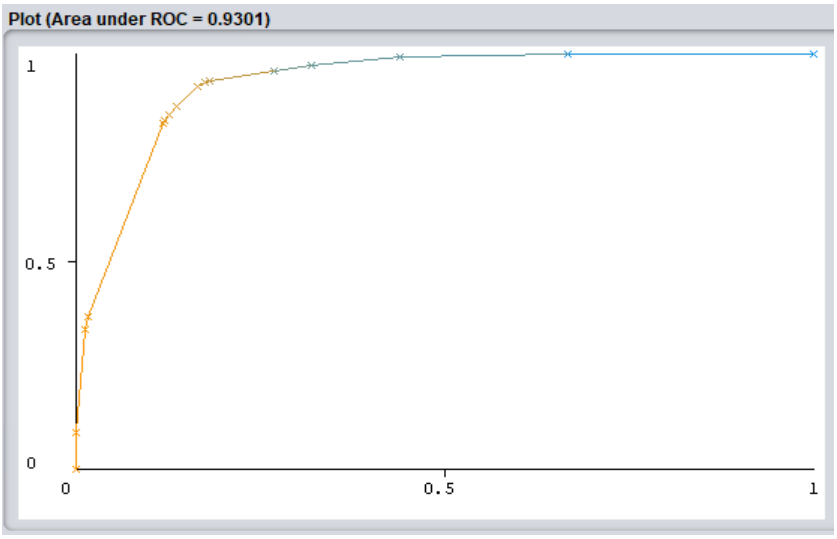
Probamos entonces, un segundo modelo de árbol C4.5 donde el número mínimo de instancias en las hojas era de 5,000. La precisión fue del 90% de las instancias correctamente clasificadas.

Cuadro 46. Salida de Weka para el Árbol C4.5 – Modelo 2
(100% Base, 5 000 instancias en hoja, 3 folds)

```
Time taken to build model: 5.06 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.12 seconds

=== Summary ===

Correctly Classified Instances      107530               90.1938 %
Incorrectly Classified Instances     11691                9.8062 %
Kappa statistic                          0.7579
Mean absolute error                      0.1532
Root mean squared error                  0.2767
Relative absolute error                 37.6242 %
Root relative squared error             61.3152 %
Total Number of Instances           119221

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,935    0,182    0,928      0,935    0,932      0,758    0,930     0,955     SI
               0,818    0,065    0,834      0,818    0,826      0,758    0,930     0,869     NO
Weighted Avg.  0,902    0,148    0,901      0,902    0,902      0,758    0,930     0,931

=== Confusion Matrix ===

     a     b   <-- classified as
 79745  5518 |     a = SI
  6173 27785 |     b = NO
```

El estimador Kappa (0.7579) dice que es un modelo sustancial, mientras que el estimador ROC (0.9301) dice que es un modelo excelente. Además, este modelo es fácil de interpretar y tiene la misma potencia que el árbol cuyos nodos hojas tienen 0 instancias.

Figura 15. ROC para el Árbol C4.5 - Modelo 2
(100% Base, 5 000 instancias en hoja, 3 folds)



También se probó el árbol C4.5 con un número mínimo de hojas de 1000. Para este modelo, el número de instancias bien clasificadas fue de 91%.

Cuadro 47. Salida de Weka para el Árbol C4.5 – Modelo 3
(100% Base, 1 000 instancias en hoja, 3 folds)

```
                    Number of Leaves  :     64    Size of the tree :      89

Time taken to build model: 6.24 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.19 seconds

=== Summary ===

Correctly Classified Instances      108251               90.7986 %
Incorrectly Classified Instances     10970                9.2014 %
Kappa statistic                          0.7679
Mean absolute error                      0.1358
Root mean squared error                  0.2609
Relative absolute error                 33.3536 %
Root relative squared error             57.7991 %
Total Number of Instances           119221

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,954    0,206    0,921      0,954    0,937      0,769    0,948     0,969     SI
                 0,794    0,046    0,872      0,794    0,831      0,769    0,948     0,906     NO
Weighted Avg.    0,908    0,161    0,907      0,908    0,907      0,769    0,948     0,951

=== Confusion Matrix ===

     a     b   <-- classified as
 81302  3961 |     a = SI
  7009 26949 |     b = NO
```

Con un estimador kappa (0.7679) que sugiere un modelo sustancial y un ROC de 0.9483 que indica que es excelente.

Figura 167. ROC para el Árbol C4.5 - Modelo 3
(100% Base, 1 000 instancias en hoja, 3 folds)

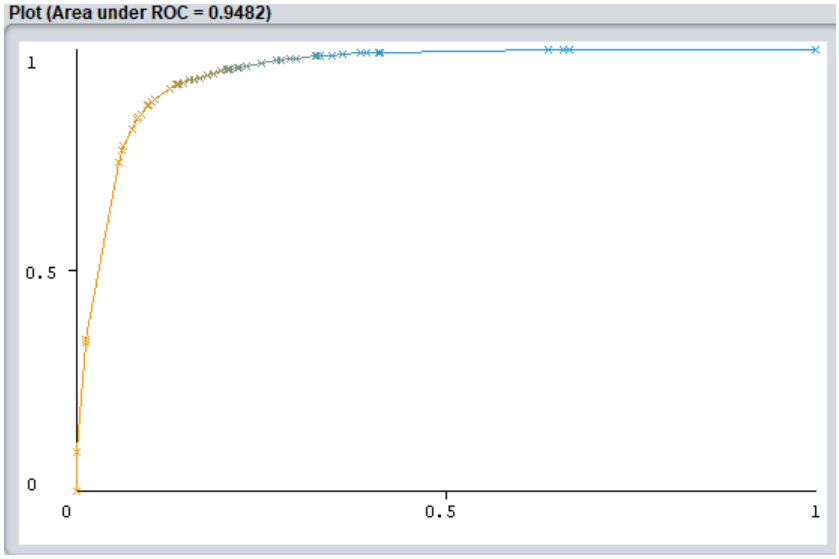


Figura 17 Árbol C4.5 – Modelo 2 (mín no. Instancias=5,000)

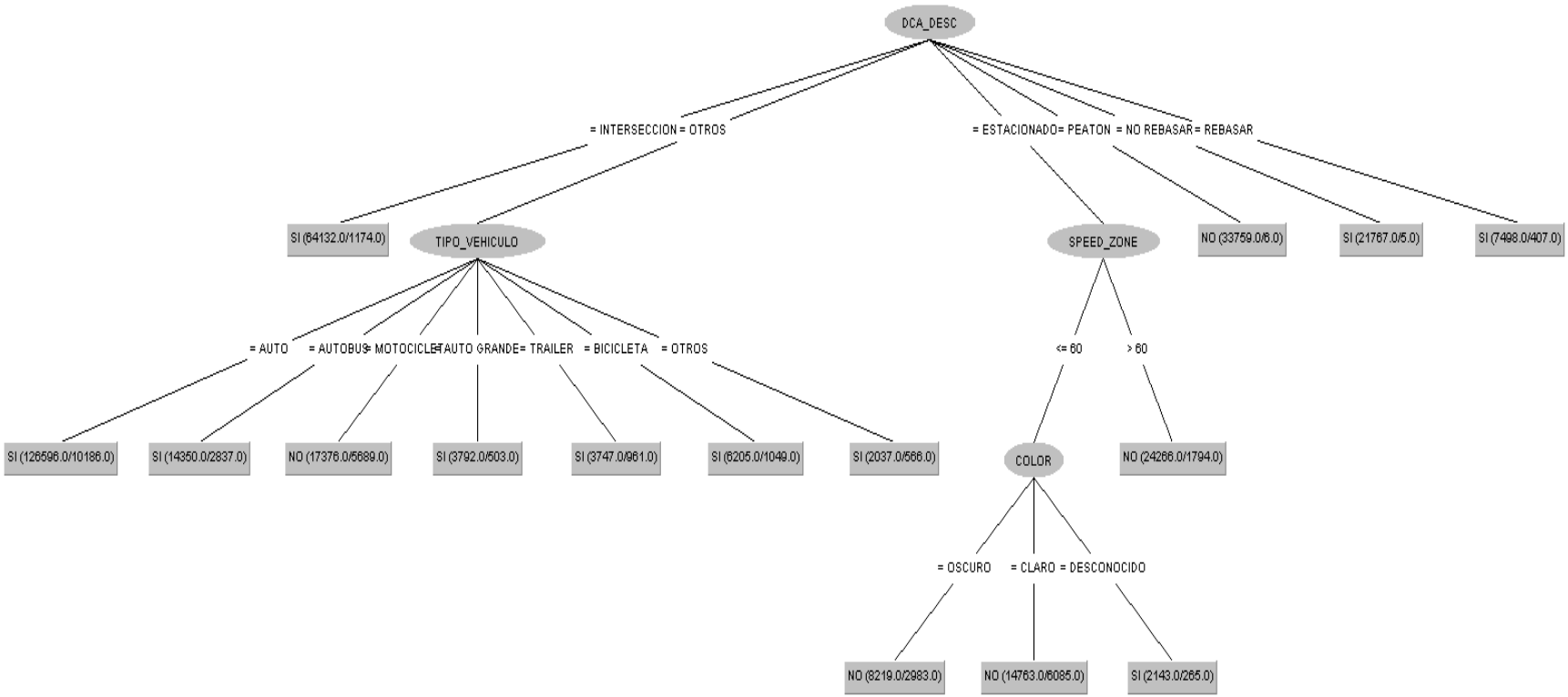
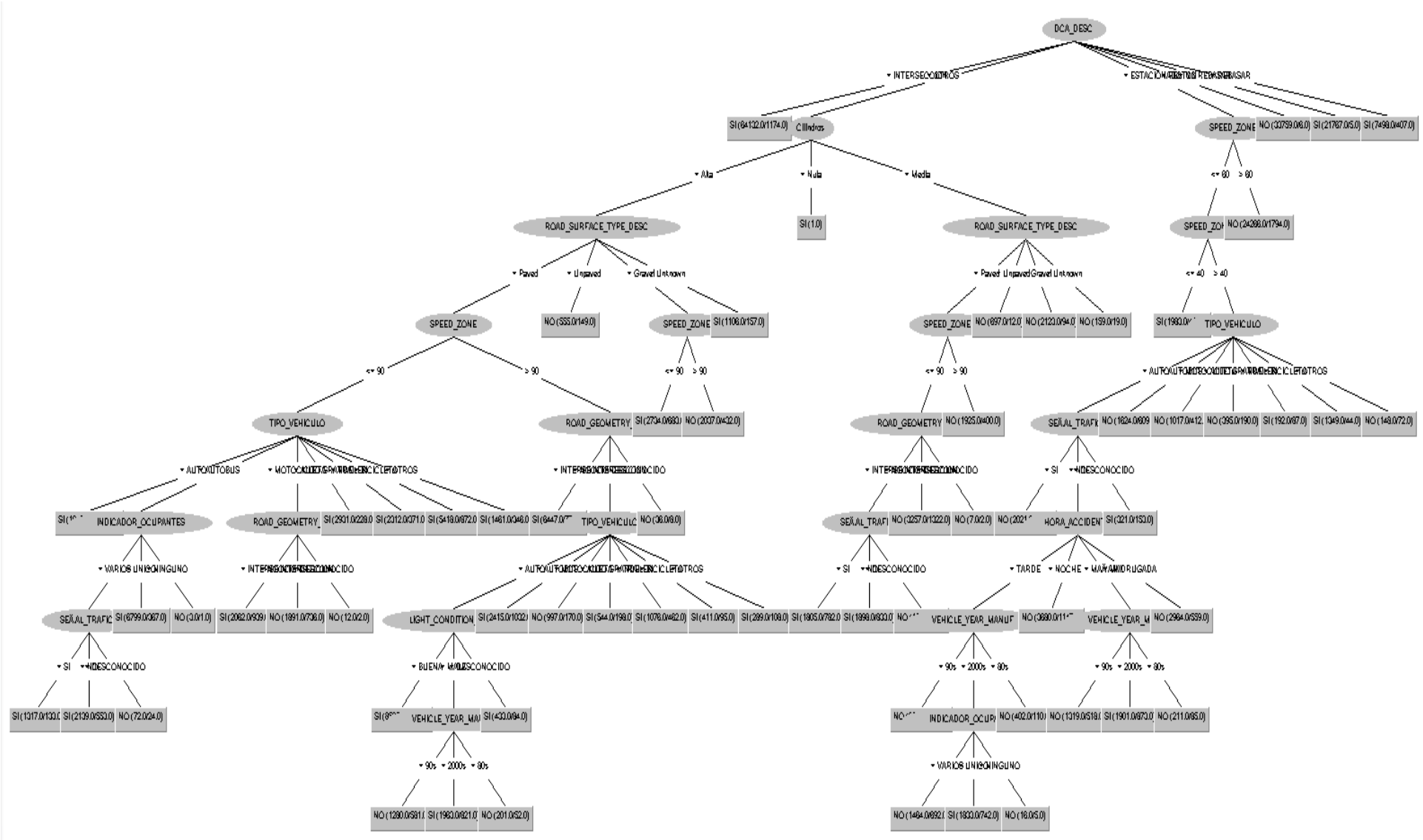


Figura 18. Árbol C4.5 - Modelo 3 (mín no. Instancias=1,000)



Además, este modelo es fácil de interpretar y tiene la misma potencia que el árbol cuyos nodos hojas tienen 0 instancias.

En conclusión, de los modelos de árbol C4.5, el preferido por el nivel de sus estimadores y fácil interpretación es el modelo 3.

3.4.4 Red Neuronal

Para la red neuronal, corrimos dos modelos en Weka. El primer modelo incluyo el 100% de los registros de la base y los parámetros por default.

Para este modelo, el número de instancias correctamente clasificadas es de 91%, el estimador Kappa es de 0.7767, lo que indica que el modelo es sustancial. El estimador ROC es de 0.9552, lo que dice que el modelo es excelente.

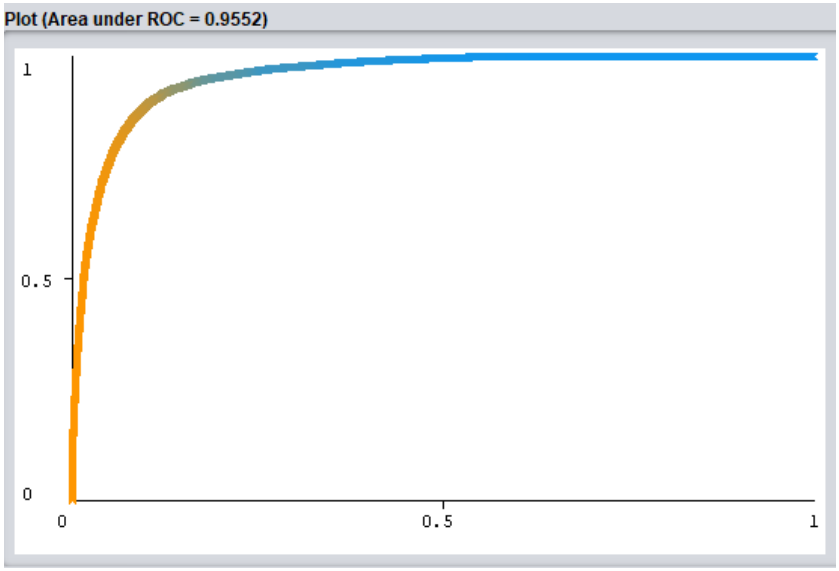
Cabe mencionar que esta primera red neuronal, con aproximadamente 350 mil registros, tardo 6 horas en correr.

Para reducir el tiempo de ejecución, decidimos correr el segundo modelo de red neuronal con una muestra del 30% de la base y 2 *hidden layers*.

Cuadro 48. Salida de Weka para la Red Neuronal – Modelo 1
(100% de la base, parámetros por Default)

```
=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     Choques2_VF
Instances:    350650
Attributes:   20

Time taken to build model: 12900.99 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 4.25 seconds

=== Summary ===

Correctly Classified Instances      108338               90.8716 %
Incorrectly Classified Instances     10883                9.1284 %
Kappa statistic                          0.7767
Mean absolute error                      0.1212
Root mean squared error                  0.2646
Relative absolute error                 29.7642 %
Root relative squared error             58.6281 %
Total Number of Instances           119221

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,934    0,154    0,938      0,934    0,936      0,777    0,955     0,978     SI
                 0,846    0,066    0,836      0,846    0,841      0,777    0,955     0,918     NO
Weighted Avg.    0,909    0,129    0,909      0,909    0,909      0,777    0,955     0,961

=== Confusion Matrix ===

     a     b   <-- classified as
 79618  5645 |     a = SI
  5238 28720 |     b = NO
```

Figura 19. ROC para la Red Neuronal – Modelo 1
(100% de la base, parámetros por Default)

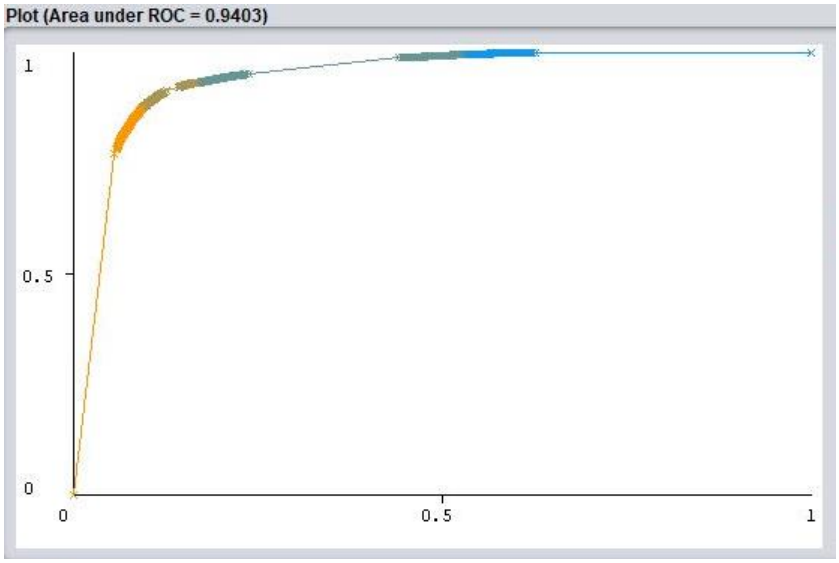


Para el segundo modelo de red neuronal se clasificaron correctamente el 90% de los registros. El estimadora Kappa (0.7604) indica que es un modelo sustancial mientras que el ROC (0.9403) dice que es modelo Excelente.

Cuadro 49. Salida de Weka para la Red Neuronal
(30% de la base, 2 Hidden Layers)

```
=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.5 -N 500 -V 0 -S 0 -E 20 -H 2
Relation:     Choques2_Sample30%
Instances:    100000
Attributes:   20

Time taken to build model: 129.85 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.23 seconds

=== Summary ===

Correctly Classified Instances       30697               90.2853 %
Incorrectly Classified Instances      3303                9.7147 %
Kappa statistic                          0.7604
Mean absolute error                      0.1453
Root mean squared error                  0.2746
Relative absolute error                 35.7384 %
Root relative squared error             61.0444 %
Total Number of Instances            34000

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.831    0.069    0.825      0.831    0.828      0.760    0.940     0.882     NO
               0.931    0.169    0.934      0.931    0.932      0.760    0.940     0.963     SI
Weighted Avg.  0.903    0.141    0.903      0.903    0.903      0.760    0.940     0.940

=== Confusion Matrix ===

     a     b   <-- classified as
  7958  1619 |     a = NO
  1684 22739 |     b = SI
```

Figura 20. ROC para la Red Neuronal
(30% de la base, 2 Hidden Layers)



3.4.5 Reglas de asociación

Con el método a priori construimos reglas de asociación mostradas en el siguiente cuadro. Observemos que la confianza de cada regla es muy buena, mayor a 0.9, lo que nos indica de dichas reglas son muy probables. Además, el *Lift* es mayor que 1 para 9 de las 10 reglas presentadas, lo que indica que estas variables aparecen juntas con más frecuencia de lo indica el azar.

Cuadro 50. Reglas de Asociación. Método A-priori

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.7 (245455 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

 1. ROAD_SURFACE_TYPE_DESC=Paved TIPO_VEHICULO=AUTO 252713 ==> Cilindros=Alta 251847    <conf:(1)> lift:(1.05) lev:(0.03) [11282] conv:(14.01)
 2. TIPO_VEHICULO=AUTO 263043 ==> Cilindros=Alta 262128    <conf:(1)> lift:(1.05) lev:(0.03) [11729] conv:(13.8)
 3. LIGHT_CONDITION_DESC=BUENA Cilindros=Alta 282361 ==> ROAD_SURFACE_TYPE_DESC=Paved 271710    <conf:(0.96)> lift:(1.01) lev:(0.01) [3853] conv:(1.36)
 4. TIPO_VEHICULO=AUTO Cilindros=Alta 262128 ==> ROAD_SURFACE_TYPE_DESC=Paved 251847    <conf:(0.96)> lift:(1.01) lev:(0.01) [3183] conv:(1.31)
 5. TIPO_VEHICULO=AUTO 263043 ==> ROAD_SURFACE_TYPE_DESC=Paved 252713    <conf:(0.96)> lift:(1.01) lev:(0.01) [3181] conv:(1.31)
 6. ROAD_SURFACE_TYPE_DESC=Paved 332638 ==> Cilindros=Alta 319314    <conf:(0.96)> lift:(1.01) lev:(0.01) [2666] conv:(1.2)
 7. LIGHT_CONDITION_DESC=BUENA ROAD_SURFACE_TYPE_DESC=Paved 283113 ==> Cilindros=Alta 271710    <conf:(0.96)> lift:(1.01) lev:(0.01) [2206] conv:(1.19)
 8. TIPO_VEHICULO=AUTO 263043 ==> ROAD_SURFACE_TYPE_DESC=Paved Cilindros=Alta 251847    <conf:(0.96)> lift:(1.05) lev:(0.04) [12310] conv:(2.1)
 9. Cilindros=Alta 333794 ==> ROAD_SURFACE_TYPE_DESC=Paved 319314    <conf:(0.96)> lift:(1.01) lev:(0.01) [2666] conv:(1.18)
10. LIGHT_CONDITION_DESC=BUENA 297003 ==> ROAD_SURFACE_TYPE_DESC=Paved 283113    <conf:(0.95)> lift:(1) lev:(0) [1366] conv:(1.1)
```

3.5. Evaluación

De los modelos ejecutados los mejores, en términos de los estimadores, son las redes neuronales y el árbol con 0 y 5,000 instancias en las hojas (Modelos 1 y 3 respectivamente).

Sin embargo, por su simplicidad, escogeríamos el árbol de 5,000 instancias en las hojas (Modelo 2).

La comparación de los modelos puede verse en el Cuadro 52.

3.5.1 Tareas de agrupamiento (clustering)

Ejecutamos el algoritmo k-medias en Weka para averiguar si había una tendencia de agrupamiento en la base. Comenzamos con 2 clústeres (k=2) y fuimos aumentando el número de clústeres hasta 7. Sin embargo, a medida que aumentamos el número de clústeres, el cuadrado del error también aumentaba, así como el número de instancias mal clasificadas.

Cuadro 51. Salida de Weka para el Cluster (k=2)

```
                    Number of iterations: 5
                    Within cluster sum of squared errors: 2468238.327862641

                Missing values globally replaced with mean/mode

                Final cluster centroids:
                                                        Cluster#
                Attribute                 Full Data           0           1
                                         (350650.0)   (139554.0)  (211096.0)
                ============================================================
                ï..CHOQUE                        SI          SI          SI
                HORA_ACCIDENTE                TARDE       TARDE       TARDE
                DAY_WEEK_DESC                Friday    Thursday      Friday
                DCA_DESC                      OTROS       OTROS       OTROS
                LIGHT_CONDITION_DESC          BUENA       BUENA       BUENA
                ROAD_GEOMETRY_DESC       INTERSECCION INTERSECCION INTERSECCION
                SPEED_ZONE                108.9379    115.1364    104.8401
                VEHICLE_YEAR_MANUF           2000s       2000s       2000s
                ROAD_SURFACE_TYPE_DESC       Paved       Paved       Paved
                MARCA                        OTROS      HOLDEN       OTROS
                TIPO_VEHICULO                 AUTO        AUTO        AUTO
                Cilindros                     Alta        Alta        Alta
                INDICADOR_OCUPANTES          UNICO      VARIOS       UNICO
                COLOR                        CLARO       CLARO       CLARO
                GRUPO_EDAD                ADULTO ADULTO JOVEN      ADULTO
                SEÃ.AL_TRAFICO               NO          NO          NO
                LICENCE_STATE_DESC        Victoria    Victoria    Victoria
                SEGURIDAD                 ADECUADA    ADECUADA    ADECUADA
                SEX                              M           M           M
                GRUPO_EDAD2               ADULTO ADULTO JOVEN      ADULTO

                                          Class attribute: CHOQUE
                                          Classes to Clusters:

Time taken to build model (full training data) : 1.28 seconds
                                                0      1  <-- assigned to cluster
=== Model and evaluation on training set ===  101303 149568 | SI
                                               38251  61528 | NO

Clustered Instances                       Cluster 0 <-- NO
                                          Cluster 1 <-- SI
0      139554 ( 40%)
1      211096 ( 60%)                      Incorrectly clustered instances :       162831.0        46.4369 %
```

Cuadro 52. Comparación de modelos

Modelo	% de la muestra usada	Características	Instancias correctamente clasificadas	Kappa	Interpretación	ROC	Interpretación
OneR – Modelo 1	100%	66% entrenamiento 34% test	87.90%	-	-	-	-
OneR – Modelo 2		10 Folds	87.99%	-	-	-	-
Ripper sin podar	30%	66% entrenamiento 34% test	86.75%	0.6230	Sustancial	0.7706	Regular
Ripper podado	30%	66% entrenamiento 34% test	90.98%	0.7751	Sustancial	0.8996	Regular
Árbol C4.5 – Modelo 1	**100%**	**66% entrenamiento 34% test Instancias por hoja=0 Folds=3**	**91.72%**	**0.7924**	**Sustancial**	**0.9499**	**Excelente**
Árbol C4.6 - Modelo 2	**100%**	**66% entrenamiento 34% test Instancias por hoja=5,000 Folds=3**	**90.19%**	**0.7579**	**Sustancial**	**0.9031**	**Excelente**
Árbol C4.6 – Modelo 3	**100%**	**66% entrenamiento 34% test Instancias por hoja=1,000 Folds=3**	**90.80%**	**0.7679**	**Sustancial**	**0.9482**	**Excelente**
Red neuronal - Modelo 1	**100%**	**66% entrenamiento 34% test Momentum=0.2 Hidden Layers=1**	**90.87%**	**0.7767**	**Sustancial**	**0.9552**	**Excelente**
Red neuronal - Modelo 2	**30%**	**66% entrenamiento 34% test Momentum=0.5 Hidden Layers=2**	**90.29%**	**0.7604**	**Sustancial**	**0.9403**	**Excelente**

3.6. Conclusiones

Analizando la base de datos original, se obtuvieron 5 modelos plausibles para clasificar el riesgo de colisión. Dos de ellos son con redes neuronales y los 3 restantes con árboles de decisión.

En general, los arboles de decisión son preferidos a las redes neuronales por su fácil interpretación y visualización. Es por esto que decidimos elegir el árbol de decisión con un número mínimo de instancias de 1,000 para cada hoja (Modelo 3).

En este modelo el nodo raíz es la variable DCA_DESC, que se refiere a qué estaba haciendo el vehículo justo antes del accidente. En cierta medida, la aparición de esta variable es intuitiva, pues hay maniobras de mayor riesgo -como las intersecciones viales o rebasar por la derecha – que otras.

Otras variables que resultaron significativas para el modelo son: la velocidad cuando el DCA_DESC involucra interacciones con el peatón y el número de cilindros, cuando el DCA_DESC es una intersección. En un tercer nivel del árbol están las variables que se refieren al límite de velocidad en la zona del accidente, al tipo de intersección del lugar y si hay señalización vial. En un cuarto nivel del árbol, las variables son las condiciones de luz, el tipo de vehículo, el año del vehículo, y el indicador de cuántas personas iban en el vehículo al momento del accidente.

Dada la importancia de estas variables, podemos ver que los accidentes vehiculares no solo involucran al conductor, sino también a factores externos como son las condiciones de luz, la señalización vial y el tipo de vialidad en la que se presenta el accidente.

Podemos intuir que, sería deseable, ubicar físicamente los lugares en los que se producen mayor número de accidentes para realizar acciones de mejora como señalización vial o mayor espacio para incorporación vehicular en las intersecciones.

4. Referencias

1. Muir, C., Johnston, I. R. & Howard, E. Evolution of a holistic systems approach to planning and managing road safety: the Victorian case study, 1970-2015. *Inj. Prev.* (2018). doi:10.1136/injuryprev-2017-042358

5. Anexo: Normalización de variables

A continuación, se muestra el detalle de las categorías antes y después de la normalización.

TYPE_DESC	Choque
Collision with vehicle	SI
Fall from or in moving vehicle	NO
Collision with a fixed object	NO
Struck Pedestrian	NO
collision with some other object	NO
Vehicle overturned (no collision)	NO
No collision and no object struck	NO
Struck animal	NO
Other accident	NO

INDICADOR_OCUPANTES	
VARIOS	2+
UNICO	1
NINGUNO	0

AGE	GRUPO_EDAD
0-12	NIÑO
13-21	ADOLESCENTE
22-35	ADULTO JOVEN
36-60	ADULTO
61+	ADULTO MAYOR

ACCIDENT_TIME	HORA_ACCIDENTE
12pm-07pm	TARDE
06am-12pm	MAÑANA
12am-06am	MADRUGADA
07pm-12am	NOCHE

LIGHT_CONDITION_DESC	
Day	BUENA
Dark Street lights on	BUENA
Unknown	DESCONOCIDO
Dark Street lights unknown	DESCONOCIDO
Dusk/Dawn	MALA
Dark No street lights	MALA
Dark Street lights off	MALA

ROAD_GEOMETRY_DESC	
Unknown	DESCONOCIDO
Cross intersection	INTERSECCION
T intersection	INTERSECCION
Multiple intersection	INTERSECCION
Y intersection	INTERSECCION
Not at intersection	NO INTERSECCION
Dead end	NO INTERSECCION
Private property	NO INTERSECCION
Road closure	NO INTERSECCION

VEHICLE_YEAR_MANUF	
[0,1989]	80´s
[1990, 1999]	90´s
[2000, 2018]	2000´s

VEHICLE_MAKE	MARCA
Para la marca se dejaron las 8 marcas más frecuentes y las otras se clasificaron en otros	FORD
	HOLDEN
	HONDA
	HYNDAI
	MAZDA
	MITSUB
	NISSAN
	OTROS
	TOYOTA

VEHICLE_COLOUR_1	COLOR
YLW	CLARO
GRY	CLARO
RED	CLARO
WHI	CLARO
SIL	CLARO
GLD	CLARO
OGE	CLARO
PNK	CLARO
ZZ	DESCONOCIDO
FWN	DESCONOCIDO
MVE	DESCONOCIDO
MRN	OSCURO
BLU	OSCURO
BLK	OSCURO
GRN	OSCURO
BRN	OSCURO
PUR	OSCURO
CRM	OSCURO

HELMET_BELT_WORN	SEGURIDAD
1	ADECUADA
6	ADECUADA
3	ADECUADA
9	DESCONOCIDO
7	NO
2	NO
4	NO
	NO
8	ADECUADA
	NO
5	ADECUADA

LICENCE_STATE_DESC	LICENCE_STATE_DESC
not available	Desconocido
Not known	Desconocido
ACT	Foraneo
Overseas	Foraneo
South Australia	Foraneo
Queensland	Foraneo
New South Wales	Foraneo
West Australia	Foraneo
Northern Territory	Foraneo
Tasmania	Foraneo
Commonwealth	Foraneo
Victoria	Victoria

DCA_DESC	DCA_DESC
LEFT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	ESTACIONADO
RIGHT OFF CARRIAGEWAY INTO OBJECT/PARKED VEHICLE	ESTACIONADO
VEHICLE COLLIDES WITH VEHICLE PARKED ON LEFT OF ROAD	ESTACIONADO
LEAVING PARKING	ESTACIONADO
OFF RIGHT BEND INTO OBJECT/PARKED VEHICLE	ESTACIONADO
REVERSING INTO FIXED OBJECT/PARKED VEHICLE	ESTACIONADO
OFF LEFT BEND INTO OBJECT/PARKED VEHICLE	ESTACIONADO
ENTERING PARKING	ESTACIONADO
VEHICLE STRIKES DOOR OF PARKED/STATIONARY VEHICLE	ESTACIONADO
U TURN INTO FIXED OBJECT/PARKED VEHICLE	ESTACIONADO
PARKED VEHICLES ONLY	ESTACIONADO
PARKED CAR RUN AWAY	ESTACIONADO
DOUBLE PARKED	ESTACIONADO
RIGHT NEAR (INTERSECTIONS ONLY)	INTERSECCION
LEFT NEAR (INTERSECTIONS ONLY)	INTERSECCION
CROSS TRAFFIC(INTERSECTIONS ONLY)	INTERSECCION
RIGHT FAR (INTERSECTIONS ONLY)	INTERSECCION
TWO RIGHT TURNING (INTERSECTIONS ONLY)	INTERSECCION
OFF END OF ROAD/T-INTERSECTION.	INTERSECCION
LEFT FAR (INTERSECTIONS ONLY)	INTERSECCION
OTHER ADJACENT (INTERSECTIONS ONLY)	INTERSECCION
RIGHT/LEFT FAR (INTERSECTIONS ONLY)	INTERSECCION
LEFT/RIGHT FAR (INTERSECTIONS ONLY)	INTERSECCION
TWO LEFT TURNING (INTERSECTIONS ONLY)	INTERSECCION
LANE CHANGE LEFT (NOT OVERTAKING)	NO REBASAR
HEAD ON (NOT OVERTAKING)	NO REBASAR
LANE CHANGE RIGHT (NOT OVERTAKING)	NO REBASAR
FELL IN/FROM VEHICLE	OTROS
REAR END(VEHICLES IN SAME LANE)	OTROS
RIGHT THROUGH	OTROS
U TURN	OTROS
LEFT REAR	OTROS
RIGHT TURN SIDESWIPE	OTROS
TEMPORARY ROADWORKS	OTROS
OFF CARRIAGEWAY TO LEFT	OTROS
OFF CARRIAGEWAY TO RIGHT	OTROS
VEHICLE OFF FOOTPATH STRIKES VEH ON CARRIAGEWAY	OTROS
OFF CARRIAGEWAY ON RIGHT BEND	OTROS
OUT OF CONTROL ON CARRIAGEWAY (ON STRAIGHT)	OTROS
STRUCK OBJECT ON CARRIAGEWAY	OTROS
OTHER ACCIDENTS-OFF STRAIGHT NOT INCLUDED IN DCAs 170-175	OTROS
OUT OF CONTROL ON CARRIAGEWAY (ON BEND)	OTROS
VEHICLE STRIKES ANOTHER VEH WHILE EMERGING FROM DRIVEWAY	OTROS

OFF CARRIAGEWAY ON LEFT BEND	OTROS
LEFT TURN SIDESWIPE	OTROS
RIGHT REAR.	OTROS
OTHER ACCIDENTS ON CURVE NOT INCLUDED IN DCAs 180-184	OTROS
STRUCK ANIMAL	OTROS
OTHER OPPOSING MANOEUVRES NOT INCLUDED IN DCAs 120-125.	OTROS
OTHER SAME DIRECTION-MANOUEVRES NOT INCLUDED IN DCAs 130-137	OTROS
PERMANENT OBSTRUCTION ON CARRIAGEWAY	OTROS
RIGHT/LEFT. ONE VEH TURNING RIGHT THE OTHER LEFT.	OTROS
REVERSING IN STREAM OF TRAFFIC	OTROS
LOAD OR MISSILE STRUCK VEHICLE	OTROS
OTHER ACCIDENTS NOT CLASSIFIABLE ELSEWHERE	OTROS
ACCIDENT OR BROKEN DOWN	OTROS
OTHER ON PATH	OTROS
OTHER MANOEUVRING NOT INCLUDED IN DCAs 140-148	OTROS
UNKNOWN-NO DETAILS ON MANOEUVRES OF ROAD-USERS IN ACCIDENT	OTROS
PULLING OUT -REAR END	OTROS
STRUCK TRAIN	OTROS
LEFT THROUGH	OTROS
STRUCK RAILWAY CROSSING FURNITURE	OTROS
RIGHT/RIGHT BOTH VEHs FROM OPPOSITE DIRECTIONS TURNING RIGHT	OTROS
LEFT/LEFT. BOTH VEHs FROM OPPOSITE DIRECTIONS TURNING LEFT.	OTROS
ANY MANOEUVRE INVOLVING PED NOT INCLUDED IN DCAs 100-108.	PEATON
PED WALKING WITH TRAFFIC	PEATON
PED NEAR SIDE. PED HIT BY VEHICLE FROM THE RIGHT.	PEATON
FAR SIDE. PED HIT BY VEHICLE FROM THE LEFT	PEATON
PED EMERGES FROM IN FRONT OF PARKED OR STATIONARY VEHICLE	PEATON
VEH STRIKES PED ON FOOTPATH/MEDIAN/TRAFFIC ISLAND.	PEATON
PED PLAYING/LYING/WORKING/STANDING ON CARRIAGEWAY.	PEATON
PED ON FOOTHPATH STRUCK BY VEHENTERING/LEAVING DRIVEWAY.	PEATON
PED STRUCK WALKING TO/FROM OR BOARDING/ALIGHTING VEHICLE.	PEATON
PED WALKING AGAINST TRAFFIC.	PEATON
LANE SIDE SWIPE (VEHICLES IN PARALLEL LANES)	REBASAR
PULLING OUT (OVERTAKING)	REBASAR
OUT OF CONTROL (OVERTAKING)	REBASAR
HEAD ON(OVERTAKING)	REBASAR
CUTTING IN (OVERTAKING)	REBASAR
OTHER OVERTAKING MANOEUVRES NOT INCLUDED IN DCAs 150-154	REBASAR