ISPC
INSTITUTO SUPERIOR POLITÉCNICO CÓRDOBA

TECNICATURA EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL - COHORTE 2020

PRÁCTICA PROFECIONALIZANTE

TRABAJO FINAL

Análisis de accidentalidad en Medellín, Colombia

Profesor: Ing. Narciso Pérez

Grupo formado por:

- Árevalo, Iván;
- Giovine, Carina;
- Gómez, Octavio;
- Herrero Rivero, Eduardo.

ÍNDICE GENERAL

Contenido

INTRODUCCIÓN

Descripción del Problema

Los accidentes de tránsito se han catalogado como un problema social por el daño que produce en las personas, las familias y la comunidad; razón por la cual los investigadores, instituciones académicas y estatales anudan fuerzas en los estudios relacionados con la predicción de accidentes de tránsito y así estabilizar y reducir las cifras de accidentalidad.

Los avances que ha tenido el análisis de datos y los sistemas de minería de datos han permitido la explotación de la información generada por las diferentes entidades públicas y privadas, para permitir la toma de decisiones a niveles administrativos y gerenciales.

Dentro de las técnicas de minería de datos para el análisis de información relacionada a temas de accidentes vehiculares, se encuentran los *árboles de decisión*, *random forest*, *perceptron* entre otros. Esta propuesta de investigación pretende estudiar el comportamiento de los siniestros viales en función de las variables externas presentes en los mismos, el estudio se realizará sobre la Alcaldía de Medellín.

A medida que nuestros sistemas de información evolucionan, se ha facilitado la recolección, administración y almacenamiento de información acerca de estos eventos. Por su parte, la Secretaría de Movilidad de la Alcaldía de Medellín ha dado un paso adelante al liberar estos datos para su análisis, los mismos se pueden consultar en el siguiente enlace http://medata.gov.co/dataset/incidentes-viales.

DESCRIPCÓN DE LA INVESTIGACIÓN

En esta práctica se realiza el análisis del conjunto de datos de accidentes de la Alcaldía de Medellín registrados desde el enero de 2014 a septiembre de 2021. Dicho análisis nos permitirá analizar las variables que intervienen en dichos incidentes, para así entender su comportamiento, lo que podrá servir de apoyo para la toma de decisiones relacionadas con la Seguridad vial, ayudando a disminuir dichos sucesos como lo pueden ser a través de planes de acción, puntos de atención o mejoramiento de la misma infraestructura vial sobre los puntos más críticos.

Los algoritmos y técnicas de machine learning ayudan a generar conocimiento valioso para la toma de decisiones.

2.1 Objetivo general

El objetivo de este estudio es predecir la gravedad de los accidentes viales de acuerdo a las variables registradas.

Nuestra Variable objetivo es 'Gravedad de la víctima', que es variable categórica y tiene como categorías la etiqueta de cada uno de los grupos: 'Herido' – 'Muerto'.

Luego nuestras variables predictoras se mencionarán a medida que avancemos con el estudio de nuestro set de datos.

La elección de algoritmos de clasificación apropiado para una problemática concreta requiere práctica, tal como se menciona en el libro "Python Machine Learning" de la editorial Macombo; siempre se recomienda comparar el rendimiento de un puñado de algoritmos de aprendizaje distintos para seleccionar el mejor modelo para un problema concreto.

Dentro de los algoritmos que se utilizaran mencionamos:

- *Arboles de decisión*

Se pueden definir como un mapa de posibles resultados de una serie de decisiones relacionadas, que comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados posibles crea nodos adicionales que se ramifican en otras posibilidades.

Existen 4 tipos de nodos en un árbol de decisión.

1. Nodo raíz: inicio del árbol de decisión, se plantea la decisión entre alternativas con resultados desconocidos.

2. Nodos de decisión: caminos de acción que pueden ser elegidos por el tomador de decisión después de haber analizado los resultados de decisiones previas.

3. Nodos de eventos: representan los posibles resultados en una decisión. Es necesario determinar los posibles resultados y la probabilidad de ocurrencia de cada uno basados en la información disponible al momento de plantear el árbol de decisión. Puede ocurrir que después de nodos de eventos continúe el proceso de selección entre alternativas probables.

4. Nodos finales: resultados finales generados por la serie de decisiones y resultados previos.

Para determinar qué tan correcta es la predicción o clasificación realizada por modelos como los árboles de decisión, se emplean métricas para determinar la eficacia y si es correcto el resultado obtenido, como la *exactitud* y *precisión*, en donde la primera representa cuantos casos sobre el total han sido clasificados correctamente en la clase a la que pertenecen, y la segunda mide cuantos casos que han sido clasificados en una clase pertenecen realmente a esa clase.

- *Random Forest*

Consiste en producir múltiples árboles que luego se combinan para producir una sola predicción de consenso. La idea de estos métodos de consenso es tomar m muestras aleatorias con reemplazo (boostraps) de los datos originales y luego aplicar en cada una de ellas un método predictivo para luego con algún criterio, establecer un consenso de todos los resultados, que generalmente es un promedio ponderado. Random Forest es un método de consenso, pero con algunas diferencias, primero que todos los modelos predictivos que se aplican a cada muestra es un árbol de decisión y que, a diferencia de los métodos de consenso, solo algunas subcolumnas son seleccionadas. De esta forma, se obtiene un amplio conjunto de clasificadores, cada uno de ellos con distintas calidades y en algunos casos con diferente asignación de la clase. Para la asignación de la clase se establece un sistema de voto mayoritario (James, Witten, Hastie , & Tibshirani, 2009).

2.2 Marco Referencial

Para alcanzar nuestro objetivo vamos a trabajar con los datos en las siguientes etapas:

- *Fase I, Entendimiento del negocio*: Se concentra en la comprensión de los objetivos del proyecto. Esto con el objetivo de desarrollar un plan de trabajo, y enfocarse en el problema que se desea abordar.

- *Fase II, Entendimiento de los datos*: Estudio y comprensión de los datos, calidad, conocimiento preliminar de los mismos. En esta fase se describen los datos con los que se cuenta, lo cual implica establecer el volumen, significado de los campos, y la descripción formal de los formatos.

- *Fase III, Preparación de los datos*: Pasos necesarios para crear el conjunto de datos finales que se utilizarán en la herramienta de modelado. En esta fase se procede a la preparación de los datos para adaptar a las técnicas de aprendizaje de máquina que se usarán posteriormente, esto incluye selección de datos, limpieza, curación, cambios de formato entre otros.

- *Fase IV, Modelación*: La selección de las técnicas se debe hacer acorde a nuestros objetivos. En esta fase se separan los datos en conjunto de entrenamiento y conjunto de prueba para medir la exactitud del modelo generado con el conjunto de prueba. Se genera uno o más modelos y se seleccionan los mejores parámetros que determinen las características del modelo.

- *Fase V, Evaluación*: Se tienen los modelos y se evalúan para ver si corresponde a los objetivos del proyecto. Teniendo en cuenta los criterios de éxito previamente establecidos se evalúa el modelo.

- *Fase VI. Despliegue (puesta en producción)*: Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

RECOPILACIÓN Y PREPROCESAMIENTO DE DATOS

3.1. Fase I. Entendimiento del negocio

Esta investigación se centra en el análisis del comportamiento de la accidentalidad vial y la incidencia que tienen las variables externas en la ocurrencia de estos sucesos, es por ello que decidimos realizar nuestra práctica sobre datos reales.

Debido al motivo anteriormente expuesto, se recurren a los datos oficiales provistos por la Secretaría de Movilidad de la Alcaldía de Medellín, Colombia.

Nuestro estudio se basa en un set de datos que contiene información de incidentes viales ocurridos desde enero de 2014 a septiembre de 2021. Nuestro csv posee un formato de 235843 filas por 19 columnas.

Las variables de estudio son: 'Gravedad_victima', 'Fecha_incidente', 'Hora_incidente', 'Clase_incidente', 'Direccion_incidente',' Sexo', 'Edad', 'Condicion', 'Mes', 'Dia', 'Num_dia', 'Hora', 'Grupo_edad', 'Año', 'Radicado', 'Latitud', 'Longitud', 'Comuna', 'Barrio'.

La cantidad de registros que posee cada una de nuestras variables se detallan a continuación:

Gravedad_victima	235843
Fecha_incidente	235843
Hora_incidente	235843
Clase_incidente	235843
Direccion_incidente	235831
Sexo	235843
Edad	235335
Condicion	235843
Mes	235843
Dia	235843
Num_dia	235843
Hora	235843
Grupo_edad	235843
Año	235843
Radicado	235838
Latitud	235843
Longitud	235843
Comuna	235843
Barrio	235225

3.2 Fase II, Entendimiento de los datos

En esta etapa nos proponemos conocer más a nuestras variables y sus valores.

Descripción de nuestras variables:

- **Gravedad_victima**: Toma valores: *'Heridos' o 'Muertos'.* Sin datos faltantes o nulos. Tipo object.
- **Fecha_incidente:** fecha en la que ocurrió el accidente. Sin datos nulos. Tipo object.
- **Hora_incidente:** Hora del incidente. Sin datos nulos. Tipo object.
- **Clase_incidente:** Tipo de incidente, toma los valores: '*Otro*', '*Atropello*', '*Choque*', '*Caida Ocupante*', '*Volcamiento*', '*Incendio*'. Sin datos nulos. Tipo obejct.
- **Direccion_incidente**: lugar del accidente, CR(carretera) y CL(calle). 12 registros nulos. Tipo object.
- **Sexo**: sexo de la persona involucrada en el accidente. Toma valores: 'M' , 'F', 'Sin Inf', 'Sin inf'. Si n datos nulos. Tipo object.
- **Edad**: Edad de la persona involucrada en el accidente. Variable numérica, posee datos con valor 0 y otros registros con rangos de edades. Tiene 508 valores nulos. Tipo object.
- **Condicion**: toma los siguientes valores: '*Motociclista*', '*Peatón*', '*Acompañante de Motocicleta*', '*Conductor*', '*Ciclista*', '*Pasajero*', '*Acompañante de motocicleta*'. Sin datos nulos. Tipo object.
- **Mes**: Mes del accidente. Toma los valores: '*Ene*', '*Feb*', '*Mar*', '*Abr*¨', '*May*', '*Jun*', '*Jul*', '*Ago*', '*Sept*', '*Oct*', '*Nov*', '*Dic*', '*Sep*'. Sin datos nulos. Tipo object.
- **Dia**: Día de la semana en que se produce el accidente. Toma los siguientes valores: '*Mié*', '*Jue*', '*Vie*', '*Sáb*', '*Dom*', '*Lun*', '*Mar*'. Sin datos nulos. Tipo object.

- **Num_dia**: Número del día del mes que se produce el accidente. La variable toma valores enteros del *1* al *31*. Sin datos nulos. Tipo int.

- **Hora**: Hora del día en la que ocurrió el accidente. Toma los valores entre *0* a *23* y 'Sin Inf'. Sin datos nulos. Tipo object

- **Grupo_edad**: rango etario de la persona que participa en el accidente. Toma los valores: '*oct-19*', '*20 - 29*', '*30 - 39*', '*40 - 49*', '*0 - 9*', '*50 - 59*', '*Sin Inf*', '*60 - 69*', '*70 - 79*', '*80 o más*'. Sin datos nulos. Tipo object.

- **Año**: Año del incidente. La variable toma los siguientes valores: *2014*, *2015*, *2016*, *2017*, *2018*, *2019*, *2020*, *2021*. Sin datos nulos. Tipo int

- **Radicado**: Representa el N° de Expediente que se genera por cada accidente. Hay 5 valores nulos. Tipo object

- **Latitud**: Identifica una latitud. Alguno de valores '6,26691466' '6,289353458' '6,234327372' ... '-75,57582422' '-75,53631071' '-75,54867484' .Sin datos nulos. Tipo object.
- **Longitud**: Identifica longitud. Alguno de sus valores: '-75,5590994' '-75,55329197' '-75,60761079' ... '6,2178952' '6,23426695' '6,272697' .Sin datos nulos. Tipo object.
- **Comuna**: Comuna donde ocurrió el accidente. Toma los siguientes valores: '*04 - Aranjuez*', '*01 - Popular*', '*16 - Belén*', '*10 - La Candelaria*', '*03 - Manrique*', '*07 - Robledo*', '*11 - Laureles Estadio*', '*Sin Inf*', '*14 - El Poblado*', '*15 - Guayabal*', '*09 - Buenos*

Aires', '*06 - Doce de Octubre*', '*05 - Castilla*', '*12 - La América*', '*08 - Villa Hermosa*', '*13 - San Javier*', '*60 - Corregimiento de San Cristóbal*', '*02 - Santa Cruz*', '*90 - Corregimiento de Santa Elena*', '*70 - Corregimiento de Altavista*', '*80 - Corregimiento de San Antonio de Prado*', '*50 - Corregimiento de San Sebastián de P.*' Tipo object.

➔ ***Barrio***: Barrio donde ocurre el accidente. Toma los siguientes valores: 'Manrique Central No. 1' 'Moscú No. 2' 'Las Mercedes' 'Jesús Nazareno' 'Manrique Oriental' 'Villa Flora' 'U.D. Atanasio Girardot' 'Sin Inf' 'Villa Carlota' 'Loma de los Bernal. 618 valores nulos. Tipo object.

Resumen en tabla de los valores nulos del set de datos:

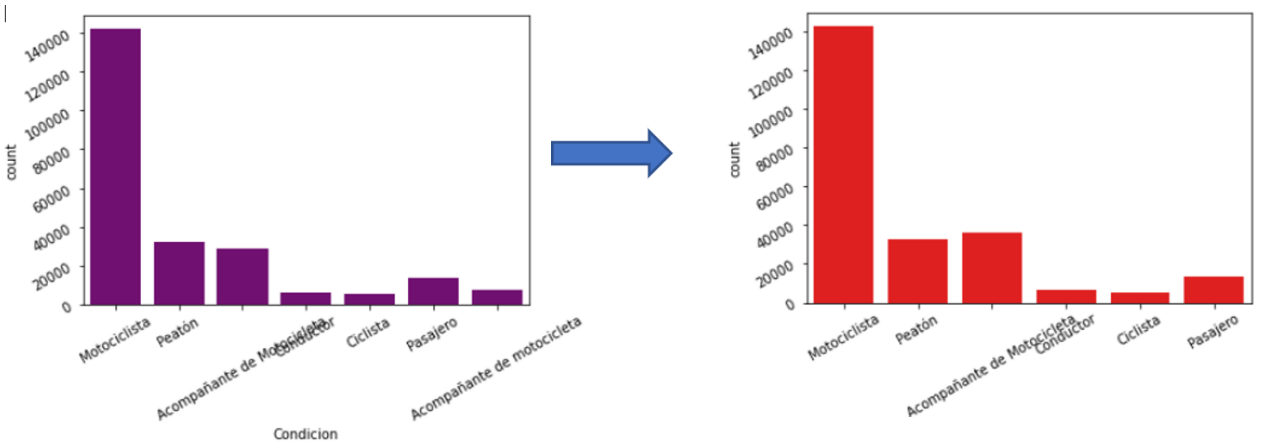
Gravedad_victima	0
Fecha_incidente	0
Hora_incidente	0
Clase_incidente	0
Direccion_incidente	12
Sexo	0
Edad	508
Condicion	0
Mes	0
Dia	0
Num_dia	0
Hora	0
Grupo_edad	0
AÃ±o	0
Radicado	5
Latitud	0
Longitud	0
Comuna	0
Barrio	618

Resumen en tabla del formato de nuestros datos:

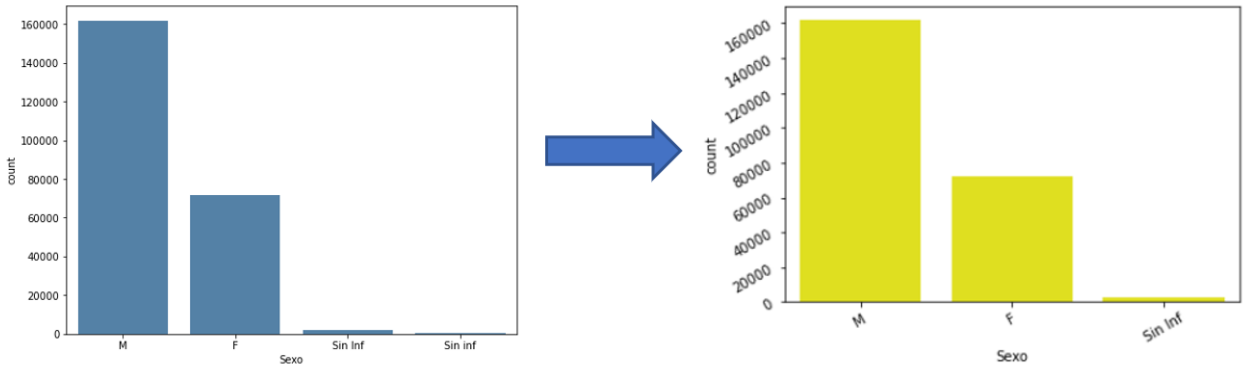
Gravedad_victima	object
Fecha_incidente	object
Hora_incidente	object
Clase_incidente	object
Direccion_incidente	object
Sexo	object
Edad	object
Condicion	object
Mes	object
Dia	object
Num_dia	int64
Hora	object
Grupo_edad	object
AÃ±o	int64
Radicado	object
Latitud	object
Longitud	object
Comuna	object
Barrio	object

La mayoría de nuestras variables son object por lo cual vamos a transformar algunas de ellas a un tipo más conveniente para trabajarlas. Además, se tiene que corregir algunos valores que toman las variables como 'Sexo', 'Condición', 'Mes', 'Grupo_edad', 'Latitud' y 'Longitud'.

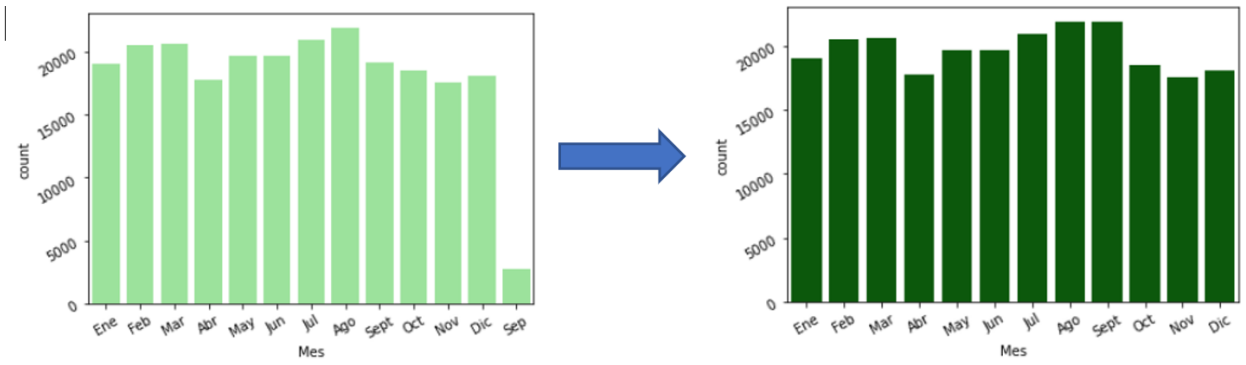
En los gráficos siguientes se mostrarán como quedan algunas de estas transformaciones:



En el caso de 'Condición' se unifican los valores 'Acompañante de Motocicleta' y 'Acompañante de motocicleta'.



En la variable 'Sexo' se unificaron los campos 'Sin Inf' y 'Sin inf'.

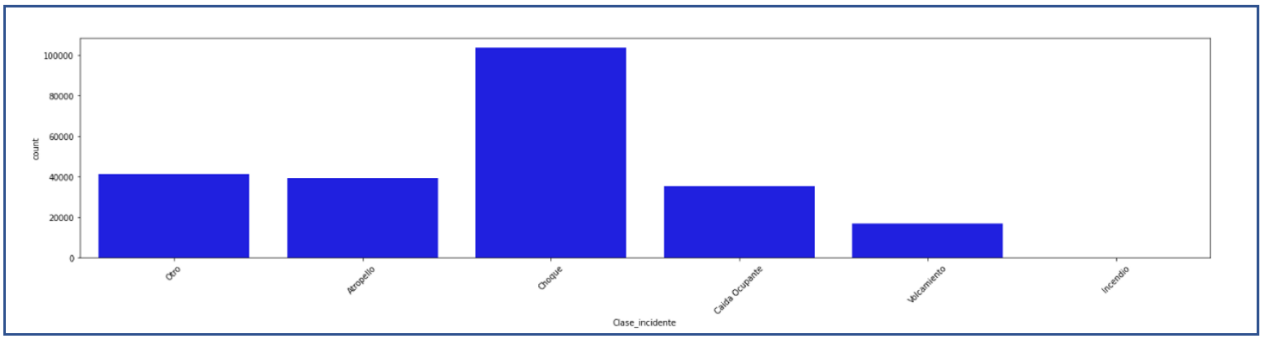


En cuanto a los meses se unificó también el valor 'Sep' con 'Sept'.

Se corrigen los valores invertidos de Latitud y Longitud, creemos que por error de tipeo en la carga de los datos se produjo este error entonces se procede a corregirlos en el set de datos. También se completan los datos faltantes de 'Edad' con la media.

En cuanto a la variable 'Grupo_edad' se realizan transformaciones para corregir los valores cargados incorrectos.

Se realizan gráficas y tablas para observar la ocurrencia de algunas variables, tal es el caso de la variable 'Clase_incidente' donde se observa que la más frecuente es choque.



A continuación, se consulta por la cantidad total de incidentes de acuerdo a los días de la semana y notamos que no hay gran variación de la ocurrencia de accidentes en los distintos días de la semana, salvo una leve baja los días domingos.

	Dia	Cantidad_Incidentes
0	Dom	28570
1	Jue	34568
2	Lun	34078
3	Mar	34750
4	Mié	34914
5	Sáb	34107
6	Vie	34856

En cuanto a la cantidad de incidentes que se registran en los meses y años que estamos analizando, observamos la siguiente distribución:

	Mes	Cantidad_Incidentes
0	Abr	17713
1	Ago	21893
2	Dic	18115
3	Ene	18981
4	Feb	20513
5	Jul	20951
6	Jun	19612
7	Mar	20560
8	May	19649
9	Nov	17519
10	Oct	18515
11	Sept	21822

	Año	Cantidad_Incidentes
0	2014	31411
1	2015	32622
2	2016	33791
3	2017	31658
4	2018	29082
5	2019	31876
6	2020	23676
7	2021	21727

Por último, se hace una consulta en dónde se muestra el total por Clase de incidente y por Comuna.

Clase_incidente Comuna	Atropello	Caida Ocupante	Choque	Incendio	Otro	Volcamiento
01 - Popular	1520	747	1045	1	770	365
02 - Santa Cruz	1191	544	1157	0	585	273
03 - Manrique	2390	1419	3342	1	1545	724
04 - Aranjuez	2852	2052	6398	2	2326	904
05 - Castilla	2850	3551	10761	0	4173	1548
06 - Doce de Octubre	1921	1787	2504	0	1499	478
07 - Robledo	2158	3291	6696	0	3445	1103
08 - Villa Hermosa	1647	1248	2892	1	1295	630
09 - Buenos Aires	1514	1351	3921	0	1837	809
10 - La Candelaria	7744	4434	17472	3	5176	2156
11 - Laureles Estadio	2588	2540	10186	2	3398	1208
12 - La América	974	946	3433	1	1081	365
13 - San Javier	1131	872	1741	0	840	430
14 - El Poblado	1152	1381	6803	1	2017	891
15 - Guayabal	1782	1715	7480	8	2347	1059
16 - Belén	1848	1817	6692	2	2292	992
50 - Corregimiento de San Sebastián de Palmitas	4	0	15	0	2	6
60 - Corregimiento de San Cristóbal	530	613	1048	0	628	275
70 - Corregimiento de Altavista	149	119	237	0	147	65
80 - Corregimiento de San Antonio de Prado	833	442	1568	0	535	224
90 - Corregimiento de Santa Elena	90	116	184	0	125	99
Sin Inf	2587	4113	7808	2	5099	2117

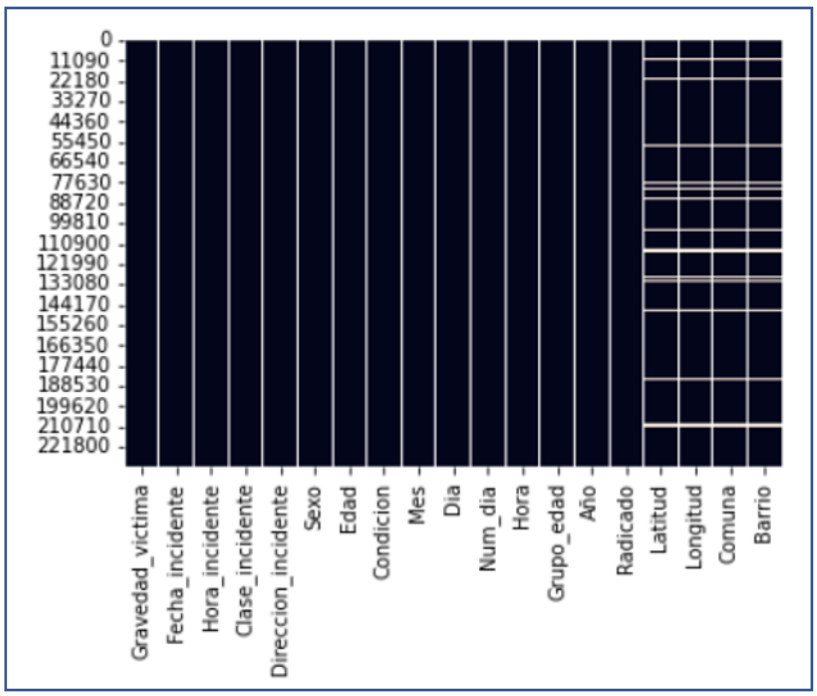
3.3 Fase III, Preparación de los datos

Luego de la limpieza y curación de los datos, nos ocupamos de preparar los mismos para entrenar los modelos de aprendizaje supervisado que me permitan clasificar la gravedad de las víctimas.

Se cambia el formato de algunas variables como se muestra a continuación:

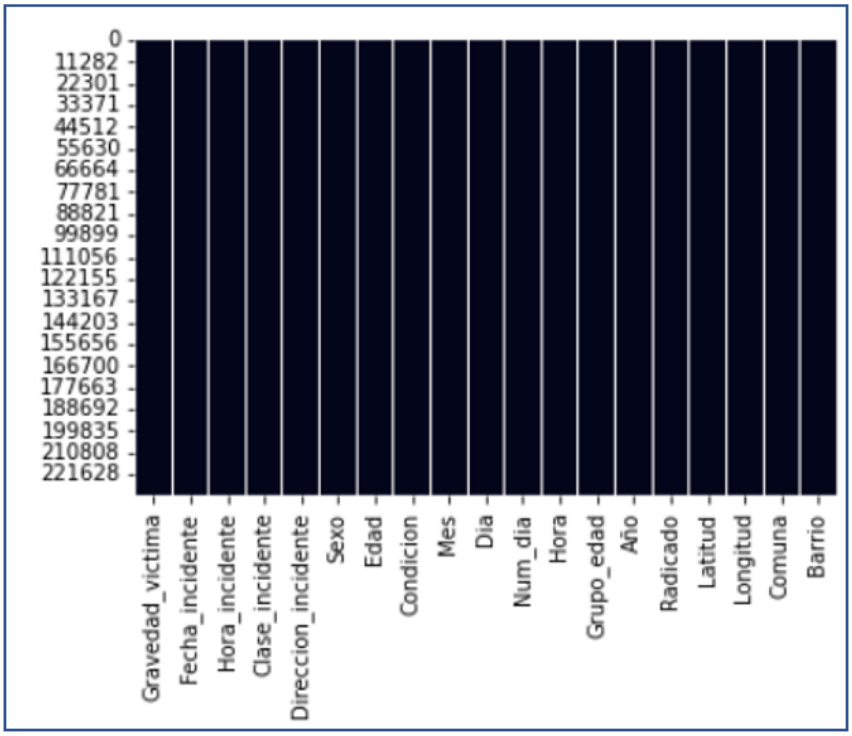
```
Gravedad_victima              object
Fecha_incidente       datetime64[ns]
Hora_incidente                object
Clase_incidente               object
Direccion_incidente           object
Sexo                          object
Edad                           int64
Condicion                     object
Mes                           object
Dia                           object
Num_dia                      float64
Hora                         float64
Grupo_edad                    object
Año                            int64
Radicado                     float64
Latitud                      float64
Longitud                     float64
Comuna                        object
Barrio                        object
```

A pesar de la curación de los datos, aún nos queda tratar con algunos datos faltantes como muestra el gráfico a continuación:



Aquí se nos presenta un nuevo desafío con los datos faltantes, ya que notamos la dificultad de imputar dichos valores ya sea de Latitud y Longitud en base a la Comuna, por ejemplo, ya que como se observa los datos faltantes corresponden a las columnas que nos ayudarían a imputar cada uno de estos datos, por lo tanto, se procede a eliminar las filas que poseen faltantes.

Quedando ahora sí el set de datos listos para convertir a numéricas las variables que usaré en mis modelos de clasificación.



Como indicamos al principio nuestra variable objetivo es 'Gravedad_victima' que hasta este momento es una variable categórica, entonces a través del método de pandas, pandas.get_dummies() que se utiliza para la manipulación de datos; convierte datos categóricos en variables ficticias o indicadoras.

Básicamente lo que hará la función es crear una columna por cada valor diferente de cada celda, separándolos por el caracter que nosotros especifiquemos, y rellenar dicha columna con ceros y unos. 0 si no está el valor presente en una celda y 1 si está presente.

De esta manera, decidimos representar a 'Heridos'=1 y 'Muertos'=0, por lo tanto nos quedamos con la columna que represente los datos de esta manera y la otra columna la eliminamos.

En esta etapa ya estamos con condiciones de seleccionar a nuestras variables predictoras para trabajar con ellas y convertirlas a numéricas.

Las columnas que se consideran como predictoras son:

*Clase_incidente;

*Sexo;

*Edad;

*Condicion;

*Año;

*Comuna.

A dichas variables las procesamos a través del método LabelEncoder() de la librería sklearn, que codifica etiquetas de una característica categórica en valores numéricos entre 0 y el número de clases menos 1.

Una vez realizado este tratamiento a las variables, nos quedaría un nuevo data set con las siguientes características:

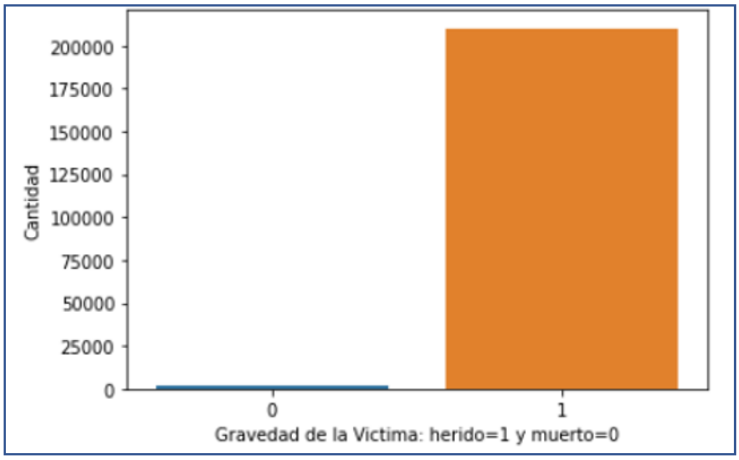
	Clase_incidente	Sexo	Edad	Condicion	Año	Comuna	Gravedad_victima
0	4	1	17	3	2014	3	1
1	0	1	20	3	2014	0	1
2	0	0	18	5	2014	0	1
3	0	1	19	3	2014	15	1
4	0	1	39	5	2014	15	1
...	...	...	...	...	...	...	...
232865	4	0	44	3	2021	5	1
232866	2	0	38	3	2021	13	1
232867	4	1	32	3	2021	7	1
232868	4	0	29	0	2021	7	1
232869	1	1	41	0	2021	2	1

211855 rows × 7 columns

3.4 Fase IV, Modelación

Estamos interesados en predecir la gravedad del incidente. Donde para nuestro data set, herido=1 y muerto=0.

En esta etapa del proyecto es importante aclarar que nuestra variable objetivo tiene un gran desbalance de los datos, como se muestra a continuación:



En donde contabilizamos Heridos (1) por 210131 registros y Muertos (0) por 1724 registros. Lo que nos propusimos en esta etapa tan importante de nuestro proyecto era correr los algoritmos supervisados con nuestro data set desbalanceado y luego, generar data para balancear los datos y volver a correr los algoritmos y comparar los resultados.

Lo primero que debemos hacer es separar nuestros datos en un 80% para train y un 20% para test.

En cuanto a la elección de modelos seleccionamos los siguientes:

Para el conjunto de datos sin balancear: Árboles de Desición (DecisionTreeClassifier) y Random Forest (RandomForestClassifier).

Luego equilibramos nuestros datos de nuestra variable objetivo con el método SMOTE(). El método SMOTE se basa en el sobremuestreo generando datos sintéticos, a partir de características comunes entre las muestras de la clase minoritaria.

Posterior a balancear nuestros datos corremos los algoritmos de Árboles de Desición (DecisionTreeClassifier), Random Forest (RandomForestClassifier), Vecinos Cercanos (KNeighborsClassifier) y Perceptron (Perceptron).

3.4 Fase V, Evaluación

En cuanto a las métricas de nuestros modelos, nos basamos en lo que nos arroja la precisión (accuracy) de los modelos ejecutados.

- La precisión resultante obtenida de ejecutar 'DecisionTreeClassifier ' en los datos originales sin balancear es la siguiente:

```
Árbol de decisión
Accuracy train Árbol de decisión: 99.41%
Accuracy test Árbol de decisión: 98.51%
```

- La precisión resultante obtenida de ejecutar 'RandomForestClassifier' en los datos originales sin balancear en el conjunto de test es del 99%.

Luego de balancear nuestros datos y correr los algoritmos las métricas que obtuvimos son las siguientes:

- DecisionTreeClassifier: train: 95.27% / test: 89.83%
- RandomForestClassifier: test: 0.90%
- KNN Classifier: train: 79.29% / test: 95.86%
- Perceptron: train: 65.44% / test : 94.40%

3.4 Fase VI, Despliegue (puesta en producción)

Al tratarse de un caso de estudio en este caso sólo nos valemos de la práctica que nos generó estar trabajando con datos reales y poder cumplir con todas las etapas del Análisis de Datos.

CONCLUSIONES

Se concluye nuestro trabajo de estudio y práctica profesionalizante de la Tecnicatura de Ciencia de Datos e Inteligencia Artificial con la satisfacción de hacer estado trabajando con datos reales y haber cumplido las etapas necesarias para el 'Análisis de datos'

Hemos analizado, curado, explorado, entendido y hemos podido correr algoritmos de aprendizaje supervisado en dónde hemos obtenidas métricas muy buenas, en la etapa de los datos sin curar creemos que el algoritmo arrojó buenas métricas porque al tener una clase notablemente mayoritaria podía identificar con facilidad estos datos, descartando los datos minoritarios.

Luego de correr algoritmos con datos balanceados el algoritmo también arrojó buenas métricas, cabe aclarar que al generar data sintética los datos se duplican entonces esto no genera un desafío para el entrenamiento y posterior test del algoritmo.

Tomamos como desafío lo propuesto por nuestro profesor de tomar una muestra equivalente a la clase minoritaria y ejecutar los modelos.

Agradecemos el acompañamiento de nuestro profesor, Ing. Narciso Pérez, en la orientación constante y apoyo para lograr nuestros objetivos de aprendizaje.

Carina, Eduardo, Iván,Octavio.