# La siniestralidad en los accidentes de tráfico de Barcelona

## Tabla de contenidos




Proyecto de Machine Learning
Francisco Fletcher Sanfeliu

# Introducción

El ayuntamiento de Barcelona, en su proyecto de digitalización, ha lanzado a concurso varias licitaciones para proyectos tecnológicos que aporten soluciones a problemas ya existentes o ayuden a mejorar la gestión de los mismos.

En nuestro caso, vamos a utilizar modelos de aprendizaje automático para tratar de optimizar el tiempo de respuesta y los recursos empleados ante un accidente de tráfico en la ciudad a partir de parámetros que sean fácilmente extraibles para cualquier usuario que de aviso al servicio de emergencias.

De esta manera, se pretende clasificar el accidente en diferentes categorías según el grado de siniestralidad para así poder establecer un plan de actuación según el escenario que se de.

# Fuente de datos

Los datos han sido extraídos del banco de datos público del ayuntamiento de Barcelona (https://opendata-ajuntament.barcelona.cat/es/).

Se han utilizado los registros de accidentes por parte de la Guàrdia Urbana desde el año 2010 hasta el 2024 en el que se recoge información como:

- Número de muertos, heridos graves y leves.
- Número de personas y vehículos implicados.
- Descripción del tipo de accidente y de la causa del peatón (si es el caso).
- Otros datos como fecha y hora, turno (mañana, tarde y noche), Distrito, Barrio, Calle y Coordenadas, entre otras.

Tras la homogeneización de variables y conjuntos de datos, la limpieza y el procesado de estos, ha quedado un conjunto de datos con 135000 entradas y 50 columnas numéricas aproximadamente. Dentro del proceso se han descartado algunas variables categóricas con alta variabilidad como por ejemplo el nombre del barrio y la calle donde ocurrió el accidente.

# Objetivos

Una vez establecido el conjunto de datos sobre el que se va a trabajar el proyecto consta de dos objetivos principales y estrechamente relacionados.

* Clasificar los accidentes según su grado de emergencia.

* Predecir el nivel de siniestralidad.

El primer objetivo requiere de un criterio lo más objetivo posible que establezca una jerarquía en el impacto que tiene un accidente en las personas afectadas

Respecto al segundo objetivo, una vez clasificado este grado de emergencia, se pretende acotar más la clasificación para poder anticiparse a las consecuencias que acarrea un cierto accidente.

# Procesamiento de datos y modelos

En esta sección se explican los aspectos más destacables del proceso de construcción y entrenamiento de los modelos de predicción.

## Métrica

El punto de partida del desarrollo del proyecto es la métrica que determina un grado de emergencia para así poder dar respaldo a una futura clasificación.

Para escoger esta métrica se han tenido en cuenta varios factores.

1. La función matemática que la determina tiene que ser creciente para cada variable y no lineal.

2. El crecimiento de la función tiene que estabilizarse para reducir la variabilidad en casos excepcionales con muchos afectados.

De esta manera, la función que determina la métrica a utilizar es la siguiente:

$$f(F, G, L, V) = ln\left(10 \cdot F + 2 \cdot G + \frac{1}{2} \cdot L + \frac{1}{2} \cdot V + 1\right)$$

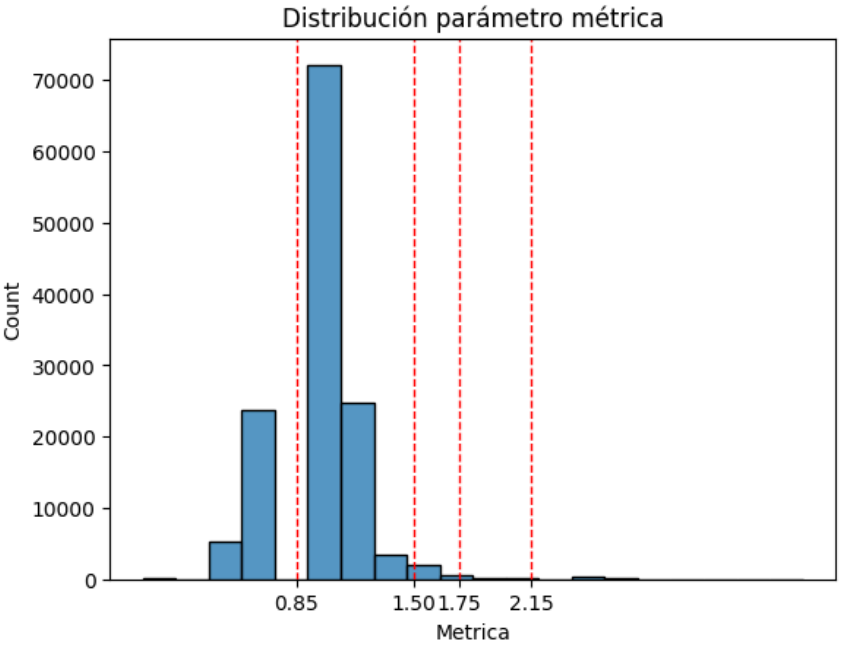
donde $F$ es el número de fallecidos, $G$ es el número de heridos graves, $L$ es en número de heridos leves y $V$ es el número de vehículos implicados.

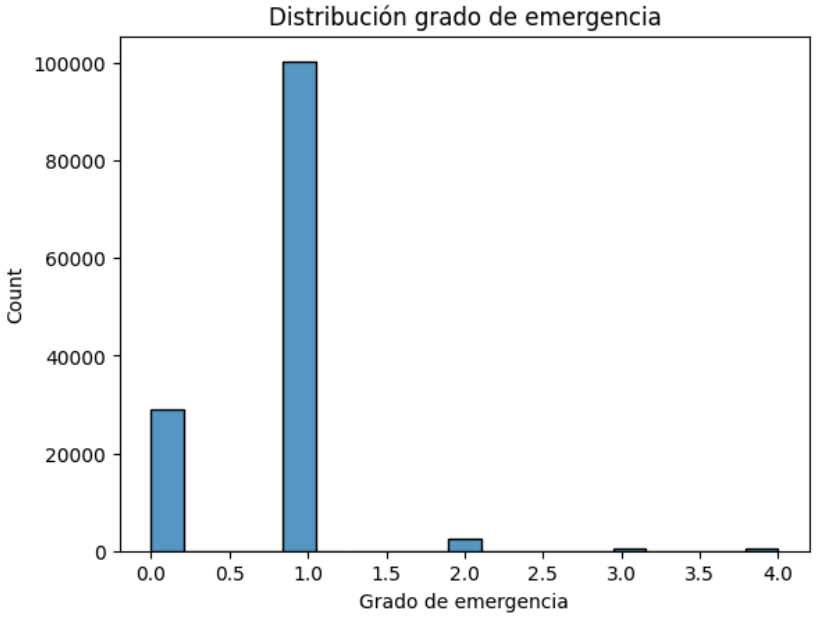
A partir de esta métrica se realiza la siguiente clasificación:

- Clase 0: Todos los accidentes que tengan un valor $m \in [0, 0.85)$
- Clase 1: Los accidentes que tengan un valor $m \in [0.85, 1.5)$
- Clase 2: Los accidentes que tengan un valor $m \in [1.5, 1.75)$
- Clase 3: Los accidentes que tengan un valor $m \in [1.75, 2.15)$
- Clase 4: Los accidentes que tengan un valor $m \in [2.15, +\infty)$

Siendo $m$ el resultado de la función.

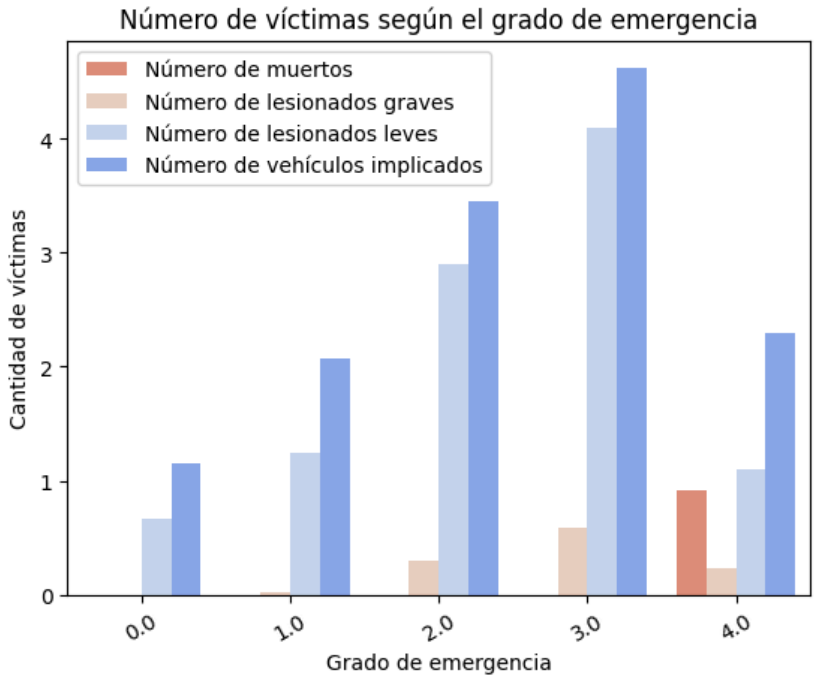
Así pues, la distribución de la métrica y su organización en categorías resulta de la siguiente manera:



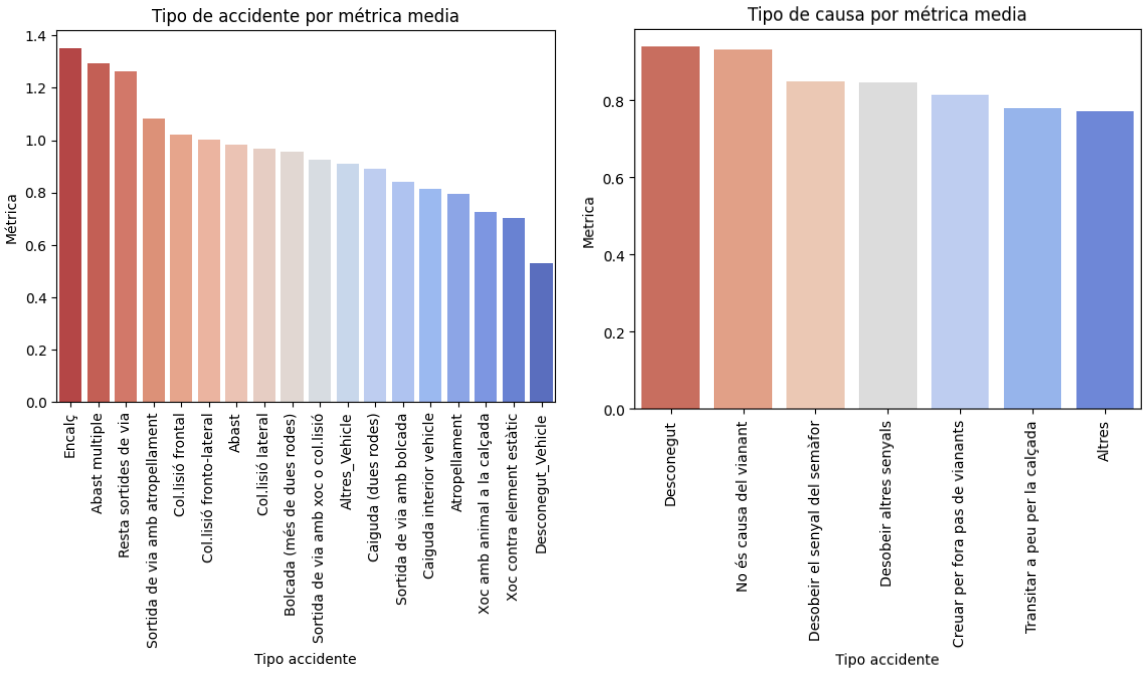

***Nota:*** Si bien tanto la función como la clasificación han sido creadas de manera artificial, sirven como punto de partida para unificar criterios y a partir de los resultados obtenidos. Los parámetros de la función y los puntos de separación pueden ser modificados para adaptarlos a los protocolos de actuación teniendo en cuenta los recursos personales y económicos empleados ante cada situación.

De esta manera, cada clase recoge una distribución media de afectados de cada tipo, como se aprecia en la siguiente gráfica:




# Feature engineering

Una vez establecido el criterio de agrupación y tras realizar unas pruebas preliminares de los modelos predictivos, se ha realizado una transformación del one-hot-encoding de los tipos de accidente y la causa del peatón para asignarles la métrica media asociada a cada descripción, obteniendo así los siguientes resultados:



# Modelos

Para entrenar los modelos se ha separado el problema en dos fases.

Debido al gran desbalance entre clases, se realiza una primera clasificación que determina si un accidente se agrupa entre las dos primeras categorías o las tres con mayor impacto. A continuación, las que son de mayor impacto se asignan entre las tres clases existentes.

Dado que la variable target tiene un carácter muy sensible al tratarse de vidas humanas, la métrica que se ha tenido en cuenta en el momento de evaluar los modelos ha sido el recall para tratar de minimizar el máximo posible las consecuencias del accidente.

$$recall = \frac{Positivos\ verdaderos}{Total\ positivos}$$

## Fase 1: Primer filtro

En la primera fase se clasificará el accidente en función de si pertenece a las categorías 0 y 1 (aquellas con menor impacto), o si por el contrario requiere de un grado de atención mayor (grupos 2, 3 y 4).

Para ello se ha entrenado una regresión logística como punto de referencia y varios RandomForest con pequeñas variaciones para determinar cuál de ellos ofrece un mejor rendimiento.

Finalmente, el modelo que aporta un mejor resultado es un RandomForest con los siguientes parámetros y métricas:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.85 | 0.92 | 103460 |
| 1 | 0.13 | 0.90 | 0.22 | 2564 |
| accuracy | | | 0.85 | 106024 |
| macro avg | 0.56 | 0.87 | 0.57 | 106024 |
| weighted avg | 0.98 | 0.85 | 0.90 | 106024 |

Partiendo de un resultado aparentemente sólido, avanzamos con el siguiente punto del proyecto.

## Fase 2: Clasificación de los accidentes de mayor gravedad

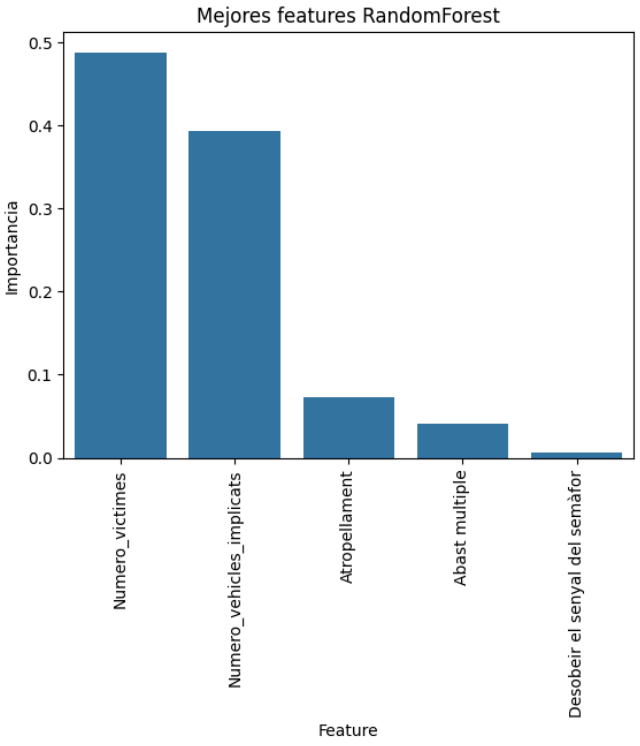
En esta segunda fase se pretende buscar un modelo de clasificación que nos permita diferenciar los 3 tipos de accidente de mayor gravedad para intentar estimar la cantidad de recursos que destinar para hacer frente a la situación.

De esta manera, se han implementado varios Pipelines con una selección de las k-mejores variables y GridSearch adaptados a diferentes modelos para poder comparar el rendimiento de los mismos.

Estos han sido Regresión Logística, Random Forest, Gradient Boosting y XGBoost. Obteniendo nuevamente un mejor desempeño con el Random Forest, en esta ocasión con las siguientes métricas:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.92 | 0.91 | 470 |
| 1 | 0.66 | 0.70 | 0.68 | 96 |
| 2 | 0.76 | 0.67 | 0.71 | 75 |
| accuracy | | | 0.85 | 641 |
| macro avg | 0.78 | 0.76 | 0.77 | 641 |
| weighted avg | 0.85 | 0.85 | 0.85 | 641 |

También se ha extraído la importancia de las variables, obteniendo en todos los modelos que las mismas 5 variables entraban en las más significativas, aunque con algunas variaciones en el orden, pero manteniéndose el Número de víctimas y el número de vehículos implicados como las más destacadas tal y como se observa en la siguiente imagen:




Por este motivo, se han creado nuevas variables para intentar dotar de variabilidad a los modelos. Estas son:

- Víctimas x vehículo
- LogVíctimas
- Víctimas/vehículo
- LogVíctimas/Vehículo

Aunque al entrenar los modelos con estas nuevas variables no se ha obtenido una mejora en el rendimiento, sí es cierto que la variable Víctimas x vehículo y LogVíctimas ha pasado a ser algunas de las de mayor significación en estos, por lo que se tendrán en consideración para los nuevos algoritmos a entrenar.

## Fase 3: Modelo de regresión y red neuronal

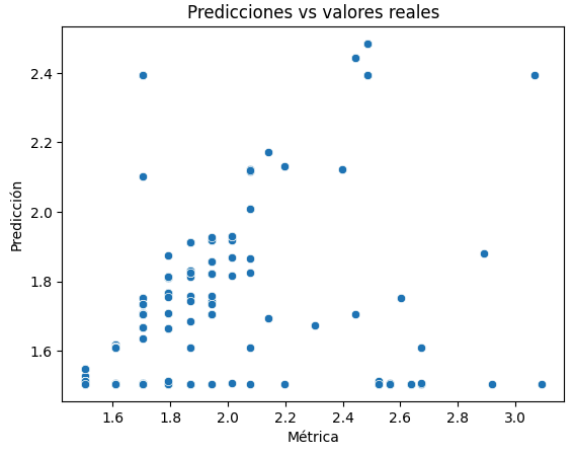
Por último, se plantean dos nuevos modelos para disponer de otro enfoque del problema.

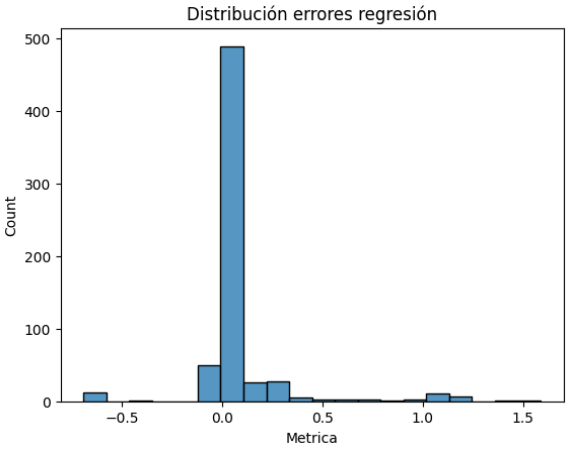
Por un lado, se entrena un modelo de regresión para predecir el valor numérico de la métrica y, a partir del mismo, clasificar el accidente siguiendo el mismo criterio que en la asignación original.

Para entrenar el modelo se han mantenido las nuevas variables construidas para los modelos anteriores y se han obtenido las siguientes métricas:

- MAE: 0.092
- MSE: 0.063
- $R^2$ : 0.390

Analizando la nube de puntos de valores reales vs predicciones y los residuos parece que estos se agrupen en torno al 0 tal y como se aprecia en las siguientes gráficas:




Sin embargo, al clasificar los resultados según la métrica, el rendimiento no alcanza a ser tan bueno como los anteriores algoritmos de clasificación.

Por otro lado, la red neuronal de clasificación tiene un muy mal rendimiento, clasificando casi todas las entradas como la categoría más baja y no asignando ningún evento a la categoría mayor.

# Futuras implementaciones

- Explorar nuevas combinaciones de parámetros e hiperparámetros que mejoren el rendimiento de los algoritmos.
- Dotar a las cámaras de tráfico de la tecnología necesaria para obtener los parámetros predictivos necesarios a través de reconocimiento de imagen.
- Entrenar una red neuronal para predecir la distribución de afectados.

# Conclusiones

1. El primer método de clasificación es un buen punto de partida para un primer triaje de la emergencia.
2. La contaminación entre clases de mayor gravedad dificulta la diferenciacion de estas por los algoritmos de aprendizaje automático.
3. El conjunto de datos necesita otras variables predictivas, por ejemplo:
    a. Clima
    b. Estado del tráfico
    c. Recursos empleados en la atención de los accidentes (ambulancias, bomberos, etc.)