Accidentes de tráfico en Madrid 2010-2018

Este proyecto realiza un análisis de los datos de accidentes de Madrid entre los años 2010 y 2018 sobre vehículos, peatones y bicicletas. Los datos han sido tomados del portal de datos abiertos del Ayuntamiento de Madrid:

<https://datos.madrid.es/portal/site/egob/>

En total el conjunto de datos cuenta con alrededor de 253000 registros y 26 variables. Es importante señalar que en estos datos de origen cada REGISTRO se corresponde con información relativa al siniestro POR VÍCTIMA, con lo cual un mismo accidente por ejemplo con 5 víctimas daría lugar a 5 registros, con información común.

Dentro del procesado posterior a la carga de datos, se cambia al formato de un registro por accidente, con los campos relativos al accidente comunes y agregando la información de los campos relativos a víctimas que sean de interés.

Los campos a considerar inicialmente son FECHA, RANGO\_HORARIO (por horas), DIA\_SEMANA, DISTRITO (los 21 de Madrid), LUGAR\_ACCIDENTE, NUM\_CALLE, NUM\_PARTE, Granizo, Hielo, Lluvia, Niebla, Seco, Nieve, supf\_Mojada, supf\_Aceite, supf\_Barro, supf\_Grava\_Suelta, supf\_Hielo, supf\_Seca\_Limpia (todas estas de las condiciones meteorológicas y de suelo booleanas), NUM\_VICTIMAS, TIPO\_ACCIDENTE, Tipo\_Vehiculo, TIPO\_PERSONA, SEXO, LESIVIDAD, Tramo\_Edad.

Por último dada la distribución del TARGET escogido (NUM\_VICTIMAS) donde prevalece de forma predominante los de una sola víctima, se decide discretizar esta variable a partir de esa referencia de una víctima, obteniendo una variable dicotómica de valores LEVE (hasta una víctima) y MEDIA-GRAVE (de 2 víctimas en adelante)

## APRENDIZAJE DESBALANCEADO

Como decíamos en el párrafo anterior, la variable TARGET está claramente desbalanceada (80% LEVE, 20%MEDIA-GRAVE). Para afrontar este problema en este proyecto se ha decidido utilizar distintas técnicas de “resampling”:

* OVERSAMPLING (en concreto SMOTE, Sintetic Minority Oversampling TEcnique, creando nuevas observaciones de la clase minoritaria mediante KNN)
* UNDERSAMPLING (disminuir número de observaciones de la clase mayoritaria mediante submuestreo) y
* COMBINACIÓN secuencial de ambas.

Tomando los resultados con la primera de ellas como referencia.

Se ha probado como modelos Decission Tree, Random Forest y por último Gradient Boost Classifier sobre los datos resultantes de aplicar cada técnica de las anteriormente mencionadas.

## CARPETAS Y ARCHIVOS

* Carpeta “**DATA”**:
* **madrid\_unico.csv**

csv de inicio con los datos descargados de la web, agregados en una única hoja de Excel. Este archivo tiene los datos agregados por víctima.

* **madrid\_accid.csv**

csv resultado de primer tratamiento de los datos sobre los iniciales.

* **madrid\_accidentes\_final\_v1.0.csv**

csv final con el preprocesado y listo para aplicar el premodelado y modelo. Este archivo contiene los datos agregados por accidente.

* **madrid\_accidentes\_modelar\_NVICTIM\_INTERV\_v1\_0.csv**

csv resultado premodelado.

* Carpeta “**NOTEBOOKS”**:
* **madrid\_carga\_archivos\_csv\_1.0.ipynb**

Jupyter notebook cuyo resultado es el csv Madrid\_accid.csv. Primer tratamiento de los datos origen.

* **madrid\_eda.ipynb**

Jupyter notebook utilizado para el estudio de los datos y la posible relación inicial entre ellos.

* **madrid\_uniraccidentes\_v1.0.ipynb**

Jupyter notebook cuya entrada es Madrid\_accid y devuelve el csv listo para el modelado – premodelado y que cargaremos además en la BBDD.

* **premodelado\_NVICTIM\_v1\_0.ipynb**

Jupyter notebook de premodelado

* **modelado\_COMB\_SMOTE\_NEARMISS\_v1\_0.ipynb**
* **modelado\_OVERSAMPL\_SMOTE\_v1\_0.ipynb**
* **modelado\_RESAMPL\_PRESENTACION\_v1\_0.ipynb**
* **modelado\_UNDERSAMPL\_NEARMISS\_v1\_0.ipynb**

Jupyter notebooks del modelado, uno con cada uno aplicando las distintas técnicas de “resampling” mencionadas en este documento sobre los datos del csv “madrid\_accidentes\_modelar\_NVICTIM\_INTERV\_v1\_0.csv”

* Carpeta “**MODELS”**:
* **modelo\_final\_PROYECTO\_joblib.sav**

Archivo con el modelo final con resultado más satisfactorio para los datos que tenemos.

* Archivo “DIAGRAMA\_BLOQUES.docx”

Archivo con el esquema ordenado de ejecución y resultados de los distintos notebooks.