TFM Master Data Science III - Kschool

Análisis de accidentes DGT 2008-2015

Rafael Martínez de Castilla Diez

I. Introducción

Se parte de la información proporcionada por la DGT en su portal estadístico, sobre accidentes de trafico, personas y vehículos relacionados, ocurridos desde los años 2008 a 2015, ambos inclusive. El objetivo es realizar un estudio de estos tres tipos de data set, aplicando los conocimientos de, tanto de técnicas de tratamiento de datos (Python y R), como de técnicas de machine learning, adquiridas durante el master.

El resultado del estudio se encuentra replicado en github:

En el repositorio se encuentra la siguiente relación de ficheros:

- Carpeta DATOS:

  - CLEANED_MICRODATOS→Datos limpiados con proceso TFM_DGT.ipynb
  - VISTAS_08_15→Vistas generadas con la acumulación de datos limpios por cada tipo de dataset.
  - Raw.zip→ Datos en bruto descargados de la DGT
  - Diccionario_accidentes*.xls→Diccionarios del dataset de accidentes, desde 2008 a 2010 y desde 2011 a 2015.
  - Diccionario_personas*.xls→Diccionarios del dataset de accidentes, desde 2008 a 2010 y desde 2011 a 2015.
  - Diccionario_vehiculoss*.xls→Diccionarios del dataset de accidentes, desde 2008 a 2010 y desde 2011 a 2015.

- RAIZ:

    - TFM_DGT.ipynb: Programa para refinamiento de los datos de los dataset. Una vez terminado el tratamiento, genera una vista por cada tipo de fichero, con todos los años acumulados.
    - Analisis_accidentes.Rmd→R markdown con el análisis y código para el dataset de accidentes.
    - Analisis_personas.Rmd→R markdown con el análisis y código para el dataset de personas.
    - Analisis_vehiculos.Rmd→R markdown con el análisis y código para el dataset de vehiculos.

A. Origen de datos

La información se puede descargar desde la siguiente url:

https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces

- Las tablas de microdatos de accidentes con víctimas (TABLA_ACCVICT_XXXX), Vehículos (TABLA_VEHIC_XXXX) y Personas (TABLA_PERS_XXXX) se relacionan entre sí de la siguiente manera:

    - Accidentes con Víctimas y Vehículos: relación 1:N, Primary_key ID_ ACCIDENTE.
    - Vehículos y Personas: relación 1:N, Primary_key ID_ACCIDENTE e ID_VEHICULO.
    - Accidentes con Víctimas y Personas: relación 1:N, Primary_key ID_ACCIDENTE

Esas relaciones se fundamentan en la estructura de la base de datos de accidentes con víctimas en la que para cada registro de la tabla Accidente puede haber 1 o más vehículos implicados. Además para cada vehículo implicado puede haber 1 o más ocupantes del vehículo.

Los datos están divididos en dos grupos, anteriores y posteriores a 2010:

La estructura de los ficheros y el contenido de los datos difiere de un grupo al otros (por ejemplo, diferente codificación para un mismo campo, incluido en los dos grupos).

Data dictionary de datos anteriores a 2010:

Data dictionary desde 2011 a 2015:

Metodología

1. PREPARACION DE LOS DATOS

Una vez descargados o descomprimidos los datos del fichero Raw.zip, se ejecuta el notebook de Python TFM_DGT.ipynb en python 3.

Este proceso limpia cada uno de los dataset y genera las vistas con los años 2008 a 2015 acumulados. Solo con las columnas comunes en todos los años, a fin de tener el mayor rango de tiempo posible para el estudio.

2. ANALISIS

A. ACCIDENTES:

Fichero RMARKDOWN **Analisis_accidentes.Rmd.** En él, se analiza el dataset de accidentes, cada registro de este dataset es un accidente con sus características. El análisis de compone de:

1. Evoluciones por tipología de victimas (totales, muertes, heridos graves y leves)
2. Estudio de correlaciones entre tipología de victimas en los distinto años y regresión lineal.
3. Evolución del numero de accidentes y victimas por año, meses y dia de la semana, a fin de determinar que meses son mas propensos a la siniestralidad.
4. Numero de accidentes y victimas por comunidad autónoma.
5. Arboles de clasificación, que toman como input, en primer lugar dos características principales los accidentes (tipo de calzada y factores atmosféricos). Y un segundo árbol que toma como input todas das las features de cada accidente.

B. PERSONAS:

Fichero RMARKDOWN **Analisis_personas.Rmd.** En él, se analiza el dataset de personas, cada registro de este dataset es un accidente con sus características.

1. Determinar mediante cluster que distintos perfiles de conductor y causas están asociadas con los accidentes. Se han utilizado kmeans, kmedoids y hclust(jerarquico) y método del codo para cada tipo de cluster, para determinar el numero optimo de clusters.

C. VEHICULOS:

Fichero RMARKDOWN **Analisis_vehiculos.Rmd.** Muestra la relación entre el numero de accidentes y la antigüedad de vehículo, en cada año.

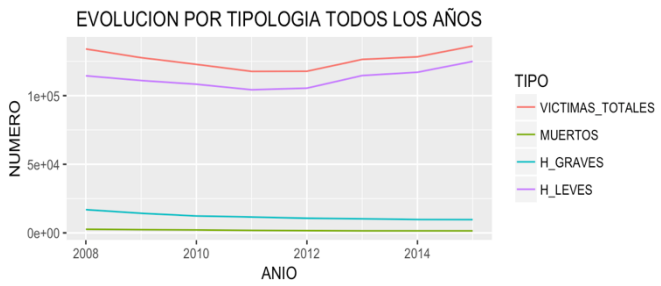
Resumen del resultado

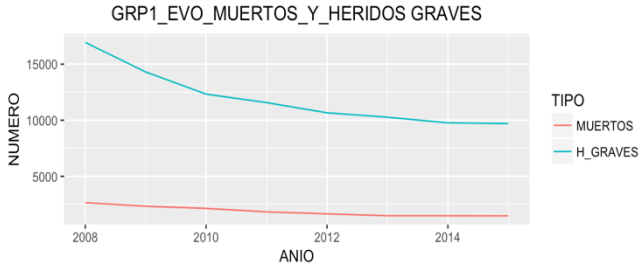
A. ACCIDENTES (mas detalles y gráficos en rmd):

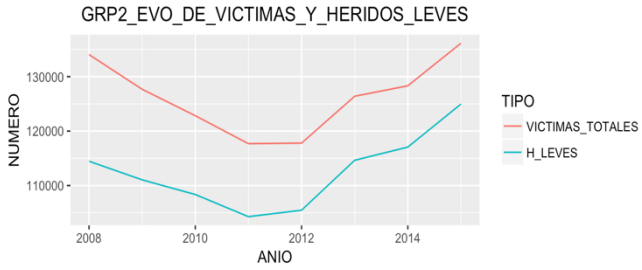
Para el periodo del estudio 2008-2015. Hay un total de 1.010.892 victimas en 8 años, de las cuales:

- MUERTES son 15.092, que representan un 9,45% sobre el total.
- HERIDOS_GRAVES son 95.535, que representan un 1,49% sobre el total.
- HERIDOS_GRAVES son 900.265, que representan un 89,06% sobre el total.

El numero de accidentes decrece desde 2008 a 2012, año en el que empieza a aumentar. En el mismo periodo se aprecia un incremento del numero total de victimas, en cambio en numero de muertes y herido leves, disminuyes. Este aumento de victimas, esta directamente relacionado con el aumento de heridos leves. Podríamos decir que hay mas accidentes, con mas victimas, pero con menos muertos.







Repartiendo el numero de accidentes (gráficos en detalle en fichero .rmd):

-MESES: Vemos de nuevo que los años con mayor siniestralidad son 2015,2014 y 2008. El mas con mayor siniestralidad y victimas es Julio. En 2015 se aprecia un notable incremento en los meses de Mayo a Julio. 2011 y 2012 son los años con numero de accidentes y victimas.

-DIAS: El día con mayor siniestralidad y mayor numero de victimas, es el Viernes, y con menor, el Domingo. 2015 es el año que tiene el top de ambos indicadores 5 de 7 días.

-CCAA: Cataluña es la CCAA con mas accidentes, seguidas de Madrid y Andalucía. En cuanto a menos siniestralidad se encuentran Navarra, Ceuta y Melilla Y la Rioja.

-TIPOLOGIA: En contra de lo que parece lógico, el mayor numero de accidentes, se han producido de día, con la calzada seca y limpia, y con buen tiempo. Posiblemente la velocidad media en estas condiciones sea mayor.

En los arboles de clasificación

1. Teniendo el cuenta solo las variables:

ZONA_AGRUPADA

SUPERFICIE_CALZADA

FACTORES_ATMOSFERICOS

Se obtiene como resultado que la la mayor probabilidad de muerte se da en:

VIAS INTERURBANAS→NIEBLA_INTENSA→ASFALTO CON GRAVILLA SUELTA→ CARRETERA SECA_Y_LIMPIA O UMBRIA

2. Teniendo en cuenta todas las características del dataset de accidentes, la mayor probabilidad de muerte, con un 42% se da en este tipo de circunstancias:

FACTORES_ATMOSFERICOS:

LLOVIZNANDO,
LLUVIA_FUERTE,
NIEBLA_INTENSA,
NIEBLA_LIGERA,
VIENTO FUERTE

RED_CARRETERA:

OTRAS_TITULARIDADES,
TITULARIDAD_PROVINCIAL

VISIBILIDAD_RESTRINGIDA:

5:'DESLUMBRAMIENTO',
7:'OTRA_CAUSA'

LUMINOSIDAD :

CREPUSCULO,
NOCHE_ILUMINACION_INSUFICIENTE,
NOCHE: SIN ILUMINACION

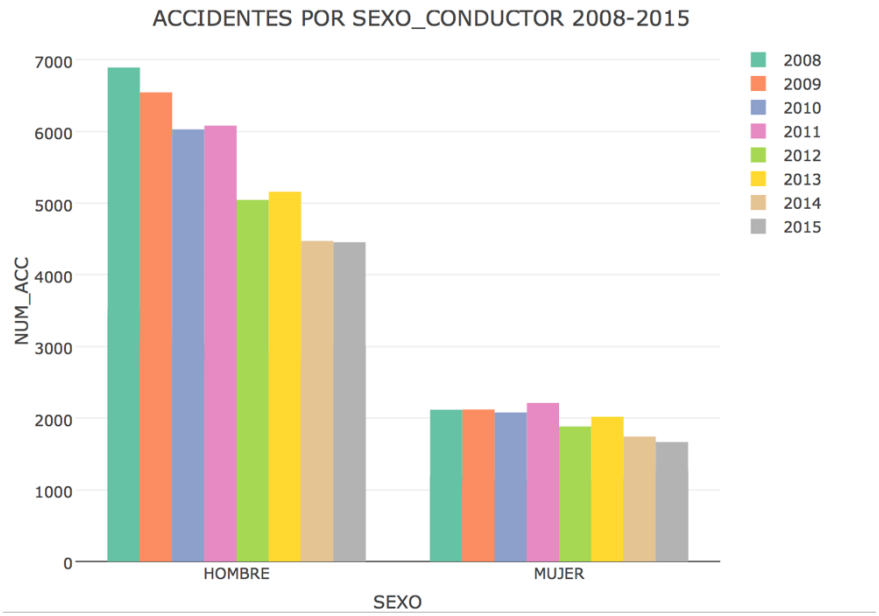
SUPERFICIE_CALZADA:

SECA_Y_LIMPIA, UMBRIA

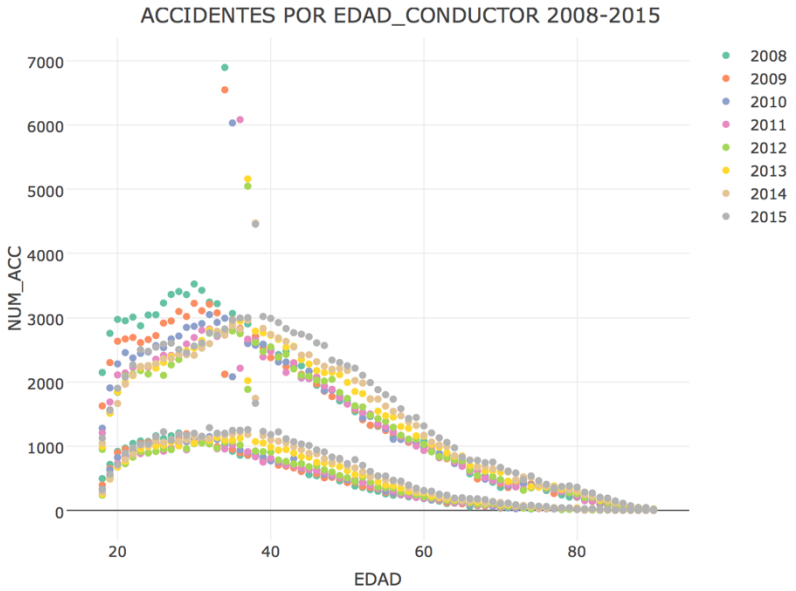
B. PERSONAS (mas detalles y gráficos en rmd):

En todos los años

Por sexo, son los hombre los que tienen mayor siniestralidad.
Por edad, la siniestralidad se concentra en el rango de 30 a 40 años. Se aprecia que la siniestralidad de los hombres va mejorado, mientras la de las mujeres es estable.

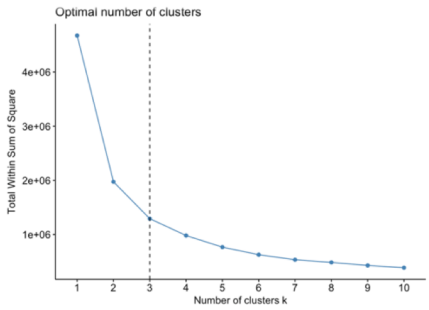


Ha un efecto de disminución de la siniestralidad relacionada con la edad del conductor



CLUSTERS DE PERSONAS

Para los 3 métodos de clustering, kmeans, medoids y hcclust, el método del codo, indica un optimo de 3 clusters, y el año 2008 como ejemplo.



K MEANS:

Cluster 1: Agrupa a hombres de 42 de media de edad y media del carne de 1996, cuyos accidentes tenian mas relacion con el exceso de velocidad o una velocidad inadecuada para el tipo de via.

Cluster 2: Agrupa a hombres 66 años de media de edad y media del carne de 1977, en su mayoría no cometieron infracción de velocidad (NINGUNA), este grupo de hombre tiene la media de EXCESO, sensiblemente menor que el grupo 1.

Cluster 3: Agrupa a mujeres de 46 años de media y media del carne de 1992, no se aprecia un nivel muy alto en la columnas de infracción, si en cambio bajo en exceso de velocidad.

KMEDOIDS

Cluster 1: Agrupa a mujeres de 46 de media de edad y media del carne de 1992 (grupo 3 en KMEANS), se incluyen dentro del grupo de no infracción de velocidad.

Cluster 2: Agrupa a hombres y mujeres en relación 80%-20% respectivamente, de 45 años de media de edad y media del carne de 1993, se desconoce si esos accidentes conllevaron infracción de velocidad.

Cluster 3: Agrupa a hombres de 59 años de media y media del carne de 1983, también con ningún tipo de infracción en cuanto a la velocidad, con el grupo 1.

HCLUST

No distingue tan claramente entre hombres y mujeres como los modelos anteriores, pero tambien diferencia dos grupos con predominancia de hombres (cluster 1 y 2). Y el 3 casi el 50%.

Cluster1: Se lleva todo el peso en cuanto a infracciones por exceso, velocidad adecuada, o ninguna infracción.

Cluster2: Acumula al 80% de hombres y 20% de mujeres con infracciones por defecto de velocidad

C. VEHICULOS

En este caso, debido al poco contenido del dataset, simplemente he creado grupos por año de matriculación del vehiculo. Podemos ver que los vehículos matriculados desde 2000 a 2010, son los que acumulan mayor siniestralidad