**Universidad de Castilla-La Mancha**

Escuela Superior de Informática

# Estudio del factor riesgo-tráfico en rutas Blablacar de la Comunidad de Madrid

*Proyecto de Minería de Datos*

Autores:
Alejandro Ruiz Aranda
Jesús Santiyán Reviriego
Luis Santiyán García

Diciembre 2021

# Índice

# Abstract

En el siguiente proyecto se realizará un estudio de los datos Carpooling de Blablacar, en concreto los de la comunidad de Madrid.
Además, este dataset se complementará con otros datos procedentes las bases de datos públicas del gobierno relacionadas con tráfico y accidentes en esta zona.
El objetivo principal será realizar un análisis exploratorio de las rutas que se dan en la comunidad y realizar categoriaciones en función de factores como el tráfico, riesgo o accidentalidad.

# ENTREGABLE I

## 1. Descripción de los datos

### 1.1. Blablacar Carpooling Dataset

El set de datos base consta de un listado de viajes anonimizados extraidos de la aplicación de rutas BlaBlaCar con información por día y trayecto ofertados de los mismos desde el 01/12/2017 hasta ella 30/11/2019 y localizados tanto en España como en Portugal.[1]
El tamaño del conjunto de datos es de 618MB, se encuentra en formato .txt y está distribuido en 11 en columnas y más de 11 millones de filas. Las columnas representan las variables de estudio y son las siguientes:

| VARIABLE | TIPO | DESCRIPCIÓN |
|---|---|---|
| dia | categórica | fecha dd/mm/aaaa entre el 01/12/2017 y el 30/11/2019. |
| país | categórica | país donde se ha dado de alta la ruta. |
| origen | categórica | ciudad de origen de la ruta. |
| destino | categórica | ciudad de destino de la ruta. |
| imp km | numérica | importe medio por kilómetro/pasajero. |
| asientos ofertados | numérica | total de plazas ofertadas |
| asientos confirmados | numérica | nº de plazas finalmente ocupadas |
| viajes ofertados | variable numérica | nº de viajes ofertados. |
| viajes confirmados | variable numérica | nº de viajes realizados. |
| ofertantes | numérica | nº de conductores distintos que han ofrecido la ruta. |
| ofertantes nuevos | variable numérica | nº de nuevos ofertantes |

[1] Los datos proporcionados pueden haber sido manipulados y multiplicados por un coeficiente, positivo o negativo, durante el proceso de anonimización, pero representan una realidad proporcional de la actividad de BlaBlaCar durante el periodo seleccionado.

En las filas se representan cada registro de viaje de la siguiente forma:

| | | | |
|---|---|---|---|
| Día<br>Imp km<br>Viajes confirmados | País<br>Asientos ofertados<br>Ofertantes | Origen<br>Asientos confirmados<br>Ofertantes nuevos | Destino<br>Viajes ofertados |
| "01/11/2017"<br>NA<br>0 | .es"<br>3<br>1 | .A Cañiza"<br>0<br>0 | .A Lama"<br>1 |

## 1.2. Api Vía Michelín

Se ha hecho uso de esta API rest privada en la versión de prueba de 60 días con el objetivo de obtener datos sobre la distancia, tiempo de conducción y carreteras involucradas en una determinada ruta.
Los datos se obtienen mediante requests con el siguiente formato:

https://secure-apir.viamichelin.com/apir/1/route.json/spa?steps=1:e::;1:e::
multipleIt=truecallback=trueauthkey=RESTGP20211210013403787883414137'
).format(corigen[1], corigen[0], cdestino[1], cdestino[0])

| **PARÁMETRO** | **DESCRIPCIÓN** |
|---|---|
| route.json | formato de la respuesta, en este caso json. |
| spa | idioma de la respuesta, en este caso español |
| step=1 | formato en el que se introducen las rutas, en este caso mediante coordenadas. |
| corigen[1] | latitud de la ciudad origen |
| corigen[0] | longitud de la ciudad origen |
| cdestino[1] | latitud de la ciudad destino |
| cdestino[1] | latitud de la ciudad destino |
| multipleit | permite obtener todas las alternativas de rutas |
| callback=True | permite obtener el fichero respuesta en formato .json |
| authkey | key de identificación para tener acceso a la api. |

Los datos obtenidos tienen el siguiente formato:

- driving dist: 1 h 40 min
- driving distance: 180 km
- roads: [A2, A3, A4]

### 1.3. Tramos de concentración de accidentes DGT

Mediante web Scraping de la web oficial de la DGT[2] se han obtenido datos sobre los puntos kilométricos de concentración de accidentes en la comunidad de Madrid. Los datos tienen el siguiente formato:

| CARRETERA | KM inicio | KM fin |
|---|---|---|
| A2 | 16.9 | 18.2 |

## 2. Trabajos similares

Se tomaron como referencia los proyectos ganadores del reto Cajamar Carpooling de 2020.
[3]

Entre ellos podemos destacar el proyecto del equipo 'Datmen' en el que entre otras cosas se analizaron las probabilidades de que se dieran las rutas, y obtuvieron una solución para cubrir las rutas pocos probables basada en el enlace de otras rutas más probables.

Por otra parte, tambíen destacar el proyecto del equipo 'Datatontos' en el que se realizó un estudio sobre el impacto de los festivales de música a la hora de usar la aplicación de Carpooling de Blablacar. Los cuales cuentan que, inicialmente, hicieron un análisis exploratorio de los datos de Blablacar para determinar que datso les serían útiles y ver en que comunidades había más influencia de tráfico y en que fechas para poder eliminar datos que les fueran innecesarios.

Debido a que estos datos eran insuficientes exploraron otras apis y bases de datos para integrar más datos con los que pudirean trabajar y una vez añadidos, limpiaron y trataron dichos datos para poder sacar los resultados y la información acorde a estos.
Por todo esto, decidimos fijarnos en este proyecto porque su modelo de trabajo va muy acorde al proceso KDD y pensamos que es la mejor forma de trabajar y por tanto usarlo como referencia.

Otro proyecto destacable es el del grupo Data South en el que consiguieron desarrollar Dashboards intuitivos para poder entender mejor los datos de Blablacar y facilitar la toma de decisiones de negocio.

[2] http://mapamovilidad.dgt.es/home.html
[3] https://www.cajamardatalab.com/datathon-cajamar-universityhack-2020/ganadores/

# 3. Planteamiento de la hipótesis y objetivos

Un aspecto de vital importancia en la conducción es la seguridad, y hoy en día gran parte de los accidentes son evitables. Por esta causa, se realizará un estudio exploratorio de los trayectos de Carpooling de Blablacar, y las bases de datos del gobierno de tráfico y accidentes con el se puedan obtener conocimientos para aplicarlos en el sistema de la plataforma y que el usuario tengo una experencia de viaje más segura.

El fin de este proyecto será el de poder estimar en función de algunas variables clave como la fecha, el volumen de tráfico, puntos negros y tramos de riesgo las rutas en las que se están produciendo más accidentes.
Para ello, lo primero sería explorar el dataset dado para encontrar datos puedan resultar útiles.
Lo siguiente, sería investigar otras fuentes de datos como podrían ser las diferentes bases de datos de tráfico que ofrece el gobierno para obtener datos adicionales e integrarlos a nuestra futura tarjeta de datos. Y por último, realizar alguna limpieza y preproceso de datos eliminando aquellos datos que no nos sean útiles y obtener la tarjeta de datos final.

La diferencia frente a un problema estadístico es que con este trabajo se realizarán categorizaciones de las rutas, y también se podrá estimar el factor de riesgo de las mismas.
En un problema estadístico en lugar de categorizaciones tan solo podrían establecer consultas, como por ejemplo de la información de las rutas concurridas o el número de accidentes en una determinada época del año.

# 4. Enrequicimiento de los datos

El set de datos base de Blablacar Carpooling se ha complementado con otras fuentes de datos necesarias para poder conseguir con los objetivos planteados. Entre ellas se encuentran:

- **Api Via Michelín:**[4] api rest a la que dandole una determinada ciudad origen y una ciudad destino ofrece gran cantidad de información relevante sobre el trayecto como: tiempo de conducción, distancia, velocidad media, carreteras, areas de serivico, coste medio del combustible, etc.
- **Tramos de concetración de accidentes DGT:**[5] web oficial de la DGT en la que se obtiene una lista de los puntos negros en carreteras de una determinada provincia o comunidad autónoma.
- **Datos de tráfico Gobierno de España:** [6]base de datos del gobierno de España en las que se pueden obtener datos oficiales de tráfico procedentes de las mediciones de las estaciones estatales permanentes o semipermanentes desde 1963.

[4] https://api.viamichelin.com/
[5] http://mapamovilidad.dgt.es/home.htm
[6] https://datos.gob.es/

# ENTREGABLE II

## 5. Selección de datos, integración de fuentes y selección de infomación base

Inicialmente partíamos del set de datos de Blablacar Carpooling, el cual contenía unos 11 millones de registros y las siguientes variables: dia, país, origen, destino, imp km, asientos ofertados, asientos confirmados, viajes ofertados, viajes confirmados, ofertantes, ofertantes nuevos.
Esto era una cantidad de datos excesiva y manejar ficheros tan pesados ralentizaba su manejo a la hora de procesarlos con otras herramientas como Google Collab. Por ello, se tomo la decisión de reducir el dominio del problema a la comunidad de Madrid.
Antes de tomar esta de decisión hubo que comprobar si realmente había rutas suficientes en Madrid para poder llevar a cabo nuestro proyecto. Esto se conseguió cruzando (operación merge de Pandas) la columna origen y destino con una lista de todos los municipios de Madrid.

Esta lista la obtuvimos de Wikipedia[7] mediante técnicas de web Scraping con la herramienta de python BeatifulSoup. Una vez obtenidas las rutas dentro de la comunidad de Madrid, se observaron los datos, y se descartaron los que no serían relavantes para nuestro proyecto como: ofertantes u ofertantes nuevos, ya que no afectan de forma directa en la accidentalidad o en el tráfico.

En una primera instancia, el objetivo era descartar también asientos y viajes ofertados ya que con los asientos y viajes confirmados sería suficiente, pero se descubrió que en la mayoría de las rutas ofertadas sobre todo en municipios de pocos habitantes dentro Madrid nunca obtenía si quiera una plaza confirmada y tan solo viajaba el conductor. Por lo que finalmente tomamos viajes y asientos ofertados, y descartamos los confirmados.
El siguiente paso fue añadir para cada ruta, el tiempo de conducción, distancia, y carreteras involucradas. Teniendo tiempo y espacio también podríamos calcular la velocidad media.

Para obtener esta información, tan solo había que realizar una request a la api vía Michelin incluyendo los parámetros en cuestión, sin embargo, el parametro ciudad origen y destino había que introducirlos con un formato de coordenadas, es decir, con latitud y longitud origen y latitud y longitud destino.

[7] https://es.wikipedia.org/wiki/Anexo:Municipios$_d e_l a_C omunidad_d e_M adrid$

Esto en un principio se realizó con la api de google geocoding. mediante requests para cada municipio de la lista de municipios de Madrid, pero esta opción no era para nada eficiente y se gastaban muchos recursos realizando tantas requests a la api.
La solución fue volver a utilizar técnicas de web scraping para obtener de la web una lista con las latitutes y longitudes de los municipios de Madrid y añadirlas a nuestra lista.

Tras conseguir las coordenadas, se realizaron las requests a la api y se añadieron las siguientes columnas a la tarjeta de datos: velocidad, tiempo de conducción, distancias, carreteras.

Teniendo las carreteras que involucran una ruta, se decidió realizar un cálculo de cambios de tramo en cada ruta, este dato afecta de forma muy directa a la hora de determinar la seguridad de un trayecto.
Para realizar este calculo bastaría con realizar una cuenta de elementos distintos dentro de cada lista de carreteras.

Finalmente completamos la tarjeta de datos añadiendo una columna con los puntos negros o tramos de concentración de accidentes para cada ruta.

Para conseguirlo, obtuvimos, mediante web Scraping de la web oficial de la DGT, la lista de puntos negros dentro de la comunidad de Madrid. Seguidamente se comprobaron las carreteras que tiene una ruta y si hay algun tramo de concentración de accidentes dentro de estas.

# 6. Preproceso y transformación

Para poder filtrar las rutas de la comunidad de Madrid, se necesitaba una lista con los nombres de todos los municipios. Esta lista se consiguió mediante web scraping de Wikipedia. Sin embargo, algunos nombres de municipios no eran iguales a los nombres que hay en el dataset de Blablcar. Los municipios que contenían un árticulo delante del nombre aparecían en la lista de municipios de forma invertida.
Por ejemplo: La Acebeda aparecía como Acebeda, La. Sin embargo, se observó que esto solo ocurría en 4 o 5 pueblos, por lo que se optó por cambiarlo a mano.

Se eliminaron tambíen los valores nulos del DataSet y en una primera instancia se eliminaron los duplicados. Pero más tarde se descubrió que se estaba perdiendo información, por lo que se optó por agrupar las rutas en función de su origen y su destino y obtener la media de asientos y viajes ofertados de las mismas.

# 7. Características de la tarjeta de datos

| VARIABLE | TIPO | DESCRIPCIÓN |
|---|---|---|
| origen | categórica | indica la ciudad o población desde la que se inicia el viaje |
| destino | categórica | indica la ciudad o población donde finaliza el viaje |
| aofertados | numérica | indica la media de asientos ofertados en una ruta |
| vofertados | numérica | indica los viajes ofertados en una ruta de media |
| distancia | numérica | indica la distancia en km que hay entre origen y destino |
| tiempo | numérica | indica cuanto se tarda en llegar del origen al destino en horas |
| carreteras | categórica | indica cuantas y cuales son las carretras de una determinada ruta |
| ntramos | numérica | indica el número de tramos que tiene una ruta |
| riesgo | categórica | indica si una ruta pasa por un tramo de concentración de accidentes |
| velocidadmedia | numérica | indica la velocidad media de una ruta |

# 8. Lineas de trabajo

A partir de la tarjeta datos ya construida, el siguiente paso será aplicar algoritmos de Clustering para categorizar las rutas en función de variables que tengan relacion directa con la seguridad. Para ello, primero se deberá determinar cuales son estas variables y si son suficentes. De no serlo, una posible solución sería poder añadir más datos sobre accidentes o tráfico a la tarjeta.

Despúes de la clásificación, otra posible linea de trabjo sería establecer un sistema de reglas borroso con el que poder asignar factores de riesgo a las rutas. Esto no deja de ser otro tipo de clasificación, pero en este paso se podría crear algun sistema de puntación en el que se califiquen las rutas en función de su seguridad.

Por último y cómo linea de trabajo final, se podría construir un sistema recomendador con el que poder ofrecer al usuario la ruta blablar más segura o incluso ofrecer nuevas rutas que tomen trayectos alternativos en los que se eviten tramos peligrosos.

# 9. Incidencias

Desde el comienzo del proyecto, el objetivo del mismo ha sido el de categorizar las rutas de Blablacar según su riesgo y poder ofrecer rutas alternativas más seguras. Para ello, se necesitaban datos como la densidad del tráfico de las rutas que se iban a estudiar, el delay porducido por esa congestión, etc.
Con todo esto en mente, se buscaron otras fuentes de datos que sirvieran de ayuda para complementar los datos iniciales. Estudiando posibles apis y bases de datos se tomó la decisión de que los datos a integrar se podían obtener de una api llamada tomtom[8], ya que ofrecía muchos datos que se necesitaban para cumplir el objetivo marcado, pero desistimos de su uso por varios motivos: el primero es que la api era de pago aunque se podían obtener algunas cosas de forma gratuiuta, la segunda era la dificultad de su uso a la hora de obtener los datos y por último, se llego a la conclusión de que esos datos eran en tiempo real y por lo tanto no servían para el uso que queríamos darle.

[8] https://developer.tomtom.com/