Minería de datos: Carpooling Blablacar

Alejandro Ruiz Aranda, Luis Santiyán García y Jesús Santiyán Reviriego

Escuela Superior de Informática, UCLM, Paseo de la Universidad 4, Ciudad Real, 13071, España.

Abstract

En el siguiente proyecto se realizará un estudio exploratorio de los datos de Carpooling de Blablacar junto con las bases de datos públicas del gobierno relacionadas con tráfico y accidentes de tráfico, con el fín de obtener rutas alternativas más seguras y categorizarlas en función de un factor riesgo-tráfico.

Keywords: Carpooling, Tráfico, Rutas

1 Descripción de los datos

El set de datos a estudiar 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 el 30/11/2019 y localizados tanto en España como en Portugal.[1]

El tamaño del conjunto de datos es de 681MB, se encuentra en formato .txt y está distribuido 11 columnas y más de 11300000 filas. Las columnas representan las variables y son las siguientes:

- DÍA: variable fecha dd/mm/aaaa entre el 01/12/2017 y el 30/11/2019.
- PAÍS: variable categórica, país donde se ha dado de alta la ruta. (ES, PT).
- ORIGEN: variable categórica, ciudad de origen de la ruta.
- DESTINO: variable categórica, ciudad de destino de la ruta.

[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.

- IMP KM: variable numérica, importe medio por kilómetro y pasajero de los viajes realizados.
- ASIENTOS OFERTADOS: variable numérica, número total de plazas ofertadas (sin incluir al conductor).
- ASIENTOS CONFIRMADOS: variable numérica, número total de plazas finalmente ocupadas (sin incluir al conductor, solo plazas ofertadas).
- VIAJES OFERTADOS: variable numérica, número de viajes ofertados.
- VIAJES CONFIRMADOS: variable numérica, número de viajes realizados.
- OFERTANTES: variable numérica, número de conductores distintos que han ofrecido la ruta. Esta variable incluye a los ofertantes nuevos.
- OFERTANTES NUEVOS: variable numérica, número de nuevos ofertantes (primera vez que ofrecen un servicio).

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

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

2 Trabajos similares

Se tomará como referencia los proyectos ganadores del reto Cajamar Carpooling de 2020. [2]. 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. Y por 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. Hemos analizado con detalle estos trabajos y los tendremos en cuenta a la hora de realizar nuestro proyecto aunque tambíen consultaremos otras fuentes para tener aún más referencias.

3 Planteamiento de la hipótesis y objetivos a perseguir

Un aspecto de vital importancia en la conducción es la seguridad, y hoy en día gran parte de los accidentes son evitables, por ello el objetivo de este proyecto será realizar un estudio exploratorio de los trayectos de Carpooling de Blablacar, y las bases de datos del gobierno de tráfico y accidentes. El fin

[2] https://www.cajamardatalab.com/datathon-cajamar-universityhack-2020/ganadores/

de este estudio será el de poder estimar en función de algunas variables clave como la fecha, el volumen de tráfico y puntos negros y tramos de riesgo dónde se esten produciendo más accidentes. De esta forma, se podrán buscar rutas alternativas más seguras con tráfico más fluido o categorizarlas según el riesgo de accidente. La diferencia de esto a un problema estadístico es que con este trabajo vamos a predecir el tráfico que puede haber en las diferentes zonas de la peninsula para a través de estos datos proporcinar a los usuarios ruta seguras o alternativas. En un problema estadístico solo se nos daría la información de las rutas concurridas o el número de accidentes en una determinada época del año.

4 Enriquecimiento de los datos

Este set de datos puede complementarse con bases de datos del gobierno sobre tráfico y accidentes/accidentes de tráfico con el objetivo de tener más información del tráfico en la red carreteras y los puntos clave en los que se producen más accidentes.[3]

[3] https://datos.gob.es/