Festividades de Castilla-La Mancha. Estudio de viajes realizados en BlaBlaCar usando técnicas de minería de datos.

Álvaro López, José Carlos Gualo

January 24, 2022

Abstract

El siguiente proyecto tiene como objetivo resolver un problema relacionado con un reto propuesto en la Datathon de Cajamar del año 2019 en el que se pretende hacer un pronóstico sobre en qué fechas cercanas a cualquier festividad en el territorio de Castilla-La Mancha va a haber un mayor aumento de nuevos ofertantes, viajes y asientos confirmados en la aplicación de Bablacar, de manera que si este repentino aumento es demasiado grande, podría acarrear problemas para los usuarios de la aplicación, ya que varios de ellos podrían quedarse sin poder reservar ningún asiento.

Así mismo, en el presente documento, se muestra cómo se han aplicado cada una de las etapas del proceso de KDD, aprendido a lo largo de la asignatura de Minería de Datos, con el fin de alcanzar un pronóstico lo más acertado posible, permitiendo que dicho pronóstico nos sea útil si queremos ayudar a que los usuarios castellano-manchegos puedan saber con certeza que podrán salir de vacaciones en una hora prudente, en la que no habrá problemas de escasez de asientos o de atascos que provoquen numerosas situaciones de estrés entre los pasajeros, al encontrarse todavía lejos de su destino vacacional y entre una gran masa de furibundos conductores que se encuentren en la misma situación.

1 Introducción y descripción de los datos

En este proyecto, nos dispondremos a tratar los datos recogidos a partir de la aplicación BlaBlaCar, la cual consiste en una red social de viajes en coche que permite la compartición de vehículo a aquellas personas que desean desplazarse al mismo lugar. De esta manera, los usuarios que utilizan esta aplicación tendrían diferentes ventajas, como, por ejemplo:

- Redistribución de los gastos. Esta ventaja la tomarían los conductores, ya que, por ejemplo, un viaje que vas a realizar solo te podría salir más barato ahorrándote gastos de combustible si compartes el coche haciendo pagar una tasa a un pasajero.
- Reducción de costes por parte del pasajero. Ya que de esta manera un viaje entre dos puntos saldría más rentable al pasajero que si lo hiciera mediante otro medio de transporte.
- Eficiencia del consumo energético, debido a la compartición de recursos, ya que esta aplicación, indirectamente, permite reducir el consume de gasolina, y por tanto se emitirían menos emisiones de CO2, beneficiando así al medio ambiente.

Para ello, hemos creado un dataset objetivo con los datos recogidos de la aplicación, incluyendo otros datos de fuentes externas, y previo a la selección y ejecución de los algoritmos de minería de datos, se han realizado operaciones de limpieza de datos y preprocesado, al igual que una reducción del número efectivo de variables que nos van a ser útiles de cara a la ejecución de los algoritmos de minería de datos que hemos elegido, para lograr una óptima interpretación de los resultados y poder documentar el conocimiento adquirido.

En cuanto a los datos que se nos ofrecen, podemos observar las siguientes variables:

NOMBRE	**DESCRIPCION**
DÍA	Variable de fecha del trayecto en formato dd/mm/aaaa (entre 01/12/2017 y 30/11/2019)
PAÍS	Variable categórica, que representa el país desde donde se ha dado de alta la ruta. Puede tomar valores ES, (España), o PT, (Portugal).
ORIGEN	Variable categórica, que representa la ciudad de origen del trayecto.
DESTINO	Variable categórica, que representa la ciudad de destino del trayecto.
IMP KM	Variable numérica, que representa el coste medio por kilómetro, que un pasajero paga.
ASIENTOS OFERTADOS	Variable numérica, que representa el número de plazas libres ofertadas a viajeros.
ASIENTOS CONFIRMADOS	Variable numérica, que representa el numero de plazas que han sido ocupadas por viajeros.
VIAJES OFERTADOS	Variable numérica, que representa el número de viajes ofertados.
VIAJES CONFIRMADOS	Variable numérica, que representa el numero de viajes llevados a cabo.
OFERTANTES	Variable numérica, que representa cuantos conductores ofrecen ese trayecto. También incluye a los ofertantes nuevos.
OFERTANTES NUEVOS	Variable numérica, que representa cuantos conductores nuevos ofrecen un servicio.

2 Trabajos anteriores

En cuanto a los previos trabajos que se han realizado sobre este dataset, podemos echar un vistazo tanto al equipo ganador como al equipo subcampeón de la datathon de 2020.

`El equipo ganador` realiza un buen trabajo de visualización además de contrastar los datos proporcionados con el censo de población para obtener la cantidad porcentual de habitantes españoles que usan la aplicación de BlaBlaCar.

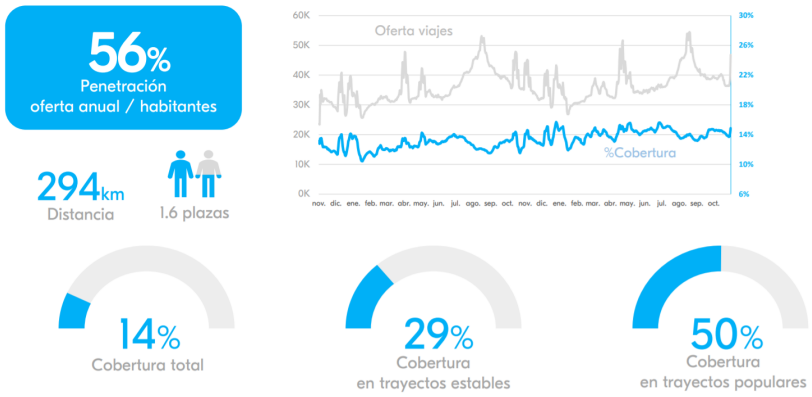


Figure 1: Visualización de los datos del equipo ganador

No obstante, el principal problema que resuelven es el caso de los viajes de un lugar a otro que no se suelen ofertar, obteniendo un conjunto de viajes con un gran porcentaje de oferta a lugares que estén cerca o de camino para poder llegar al destino inicial aunque el usuario tuviera que hacer varias paradas y/o viajes con ofertantes diferentes.

Por otro lado, `El equipo subcampeón` se enfoca en la celebración de diversos festivales musicales en la península y trata de relacionar el uso de la aplicación durante las fechas que se realizan dichos festivales en las zonas próximas para determinar si el número tanto de viajes ofertados como de asientos ocupados aumenta en relación a la media general.

3 Planteamiento de la hipótesis. Objetivos a perseguir

Entre los objetivos que se nos han ocurrido para resolver encontramos las siguientes ideas:

- Pronosticar en qué época del año va a haber un incremento masivo de ofertantes nuevos y de ofertas de viajes para evitar una posible saturación de la red.
- Hacer un análisis sobre aquellas zonas que generan más ingresos con el uso de la aplicación.
- Obtener las ciudades que obtienen un mayor incremento o disminución de población debido a los viajes que se realizan usando BlaBlarCar.

Tras debatir sobre las posibles hipótesis planteadas de las que disponíamos anteriormente, finalmente llegamos a la conclusión de que la que más nos convencía era la siguiente:

Se tratarán los datos que se nos han prestado, así como enriquecer los mismos, como veremos a continuación, para pronosticar los incrementos masivos tanto de ofertantes, como de viajes y viajeros que se pueden llegar a dar en fechas clave del año en la comunidad de Castilla-La Mancha, como son los días festivos del año 2018; con el objetivo de prevenir la congestión y saturación de la red y los servidores de los que hace uso la aplicación.

4 Posibilidades de enriquecimiento de datos

Como hemos mencionado anteriormente, el enriquecimiento de datos nos va a ofrecer la oportunidad de completar la información base de la que partíamos para poder llevar a cabo nuestro objetivo de manera adecuada. Estos datos que van a enriquecer nuestra información base serán los siguientes:

- Fechas de los días festivos en Castilla-La Mancha. Para ello se han obtenido de una página oficial, datos.gob.es tres archivos en formato

Excel con los días festivos de dicha comunidad entre los años 2017 y 2019. Estos datos se han sometido a un procesado para tener un solo archivo con todas las fechas en un formato legible por la máquina.

- Datos del INE de todos los municipios de Castilla-La Mancha. Para ello se ha obtenido un archivo en formato CSV de paginas oficiales del Estado, del cual se obtiene información relativa al nombre del municipio, numero de habitantes, densidad de población, código postal, etcétera.
- Coordenadas de todos los municipios de España y Portugal. Para ello se han obtenido una serie de archivos en formato CSV que contienen amplia cantidad de información sobre cada municipio de la península.

Estos datos serán importantes a la hora del preprocesado de los datos para la obtención del conjunto de datos sobre el cual trabajar, aunque también podrá ser usado en las otras fases restantes.

5 Selección de datos

Una vez que ya sabemos hacia donde queremos orientar el proyecto, y cuáles son los objetivos que perseguimos, vamos a realizar una tarea preliminar al preprocesamiento de los datos. Esta tarea nos va a servir para tener datos aparte de los que se nos ofrece de la aplicación de BlaBlaCar, de tal manera que nos ayuden en la consecución de nuestro objetivo. Estas tareas han consistido principalmente en preguntarnos que datos nos iban a hacer falta en un futuro, de tal modo que, consiguiendo datos sobre ello, sobre los cuales hemos aplicado algoritmos simples, nos permitieran obtener información útil para la hora de realizar el preprocesamiento y transformación de conjunto de datos original de BlaBlaCar.

Todo esto ha sido posible gracias a la integración de distintas fuentes de datos, sobre las que vamos a hablar en el siguiente apartado.

5.1 Integración de varias fuentes

Como hemos dicho, las fuentes de información adicionales que hemos usado han sido necesarias para la realizar correctamente el preprocesamiento de los datos, pero para conocer mejor el por qué de utilizar esas fuentes, vamos a comentar que problema nos ha solucionado el uso de esa fuente.

- Archivos referentes a días no laborables en Castilla-La Mancha. Son tres archivos en formato .xls obtenidos de la página web de la Junta de Comunidades de Castilla-La Mancha en los años 2017, 2018 y 2019, utilizados para saber que días van a ser no laborables en el marco de Castilla-La Mancha.

- Archivos referentes a la localización de municipios de Portugal y España. Estos datos nos han servido para saber a qué zona pertenece cada municipio de estos países.

- Archivo referente a los municipios de Castilla-La Mancha. Utilizado para delimitar aquellos viajes con origen o destino en esta comunidad.

En un apartado posterior explicaremos los procesos que hemos realizado sobre estas fuentes adicionales para obtener la información que hemos mencionado. También hablaremos de como hemos usado esta información sobre el conjunto de datos original a la hora de preprocesar los datos.

Para finalizar este apartado debemos concluir en que la información de base sobre la que se comenzará a trabajar ha sido toda la que hemos mencionado anteriormente, a saber: El conjunto de datos original de BlaBlaCar, los datos sobre las festividades, los archivos sobre la localización de los municipios de España y Portugal, y un archivo con información relativa tan solo sobre los municipios de Castilla-La Mancha.

6 Preprocesado y transformación. Tratamiento de datos.

6.1 Viajes relacionados con Castilla-La Mancha.

Para este paso haremos uso del fichero que mencionamos en el apartado anterior, el cual contiene información referente a todos los municipios de Castilla-La Mancha. Con este archivo, lo que se ha hecho es hacer un listado con el nombre de aquellos municipios pertenecientes a la comunidad, para después comparar con los atributos de‘**ORIGEN’** y **‘DESTINO’** de nuestro conjunto de datos; en el caso de que en alguno de los dos atributos apareciera el nombre de algún municipio de Castilla-La Mancha, ese registro pasaría el filtro, pasando a formar parte de nuestra tarjeta de datos.

6.2 Registros de viajes realizados.

Como hemos dicho antes, encontramos algunos valores nulos en el campo **'IMP KM'**. Estudiando el por qué de esos valores nulos, nos dimos cuenta de que ese atributo aparecía como nulo porque las variables de **'VIAJES CONFIRMADOS'** y **'ASIENTOS CONFIRMADOS'** tenían como valor 0 para esos registros. Por tanto, tomamos como decisión quedarnos solamente con aquellos registros de viajes que se realizaban con seguridad, es decir, aquellos en los que al menos había una confirmación.

6.3 Filtrar fechas relevantes.

6.3.1 Días no laborales y vísperas

En la primera parte de este filtrado, hemos hecho uso de los ficheros referentes a las fechas de los días no laborales en la comunidad de Castilla-La Mancha que mencionamos antes. Mediante su tratamiento se ha obtenido un archivo final que contiene todos los días no laborables de los años 2017, 2018, 2019, y sus vísperas, en un formato legible por la máquina. Una vez obtenido este archivo, nos va a ser útil para hacer el filtrado por estas fechas.

6.3.2 Fines de semana

En la segunda parte del filtrado hemos hecho uso de la librería datetime de Python para filtrar por aquellos días que cumplan la condición de ser viernes, sábado o domingo, ya que consideramos esos días en la misma categoría de días "especiales" como pueden ser los días festivos. Una vez realizado este filtrado, obtendríamos como resultado aquellos viajes que de verdad nos interesan, que son aquellos que tienen que ver con Castilla-La Mancha, y se realizan en días no laborables, sus vísperas, o durante fines de semana.

6.4 Agregación de variables.

6.4.1 SEMANA AÑO

Con el objetivo de tener más información sobre la temporalidad de los viajes, consideramos necesario la obtención de una nueva variable llamada **'SEMANA AÑO'**. Esta variable contiene el valor numérico que representa en que semana del año se realizó ese viaje. Este nuevo valor se ha obtenido haciendo uso del método isocalendar de la librería `datetime`.

6.4.2 DIA SEMANA

Se ha añadido el atributo **'DIA SEMANA'** que representa el día de la semana en el que tiene lugar el viaje, de tal manera que el resultado ya esté discretizado, es decir, el Lunes vendrá representado como 0, el Martes como 1, el Miércoles como 2, etc. Esto se ha hecho con el objetivo de tener más información temporal.

6.4.3 FLUJO

Adicionalmente, para tener constancia del flujo de viajes que tiene lugar entre los territorios que definen el conjunto de datos inicial, se ha decidido añadir una variable que represente si el viaje consiste en **salir, (1), entrar, (2), o mantenerse, (0)**, en territorio castellano manchego. De este modo, podemos detectar fácilmente el entorno en el que se mueven los viajes.

6.5 Discretizar PAIS

Tras entender este atributo, pensamos que era buena idea discretizarlo, es decir, pasarlo a una variable numérica, ya que a la hora de realizar modelos predictivos con la variable categórica original puede dar problemas. Esta discretización se ha realizado con la implementación de scikit-learn de `labelEncoder`.

6.6 Clustering de municipios en zonas

A partir de los datos que tenemos almacenados en los archivos .CSV mencionados anteriormente, nos disponemos a realizar una agrupación de todos los municipios de España y Portugal con el fin de contrastar esta agrupación con nuestro dataset principal y obtener una visión más clara del flujo de viajeros en festividades.
Para ello vamos a ejecutar un algoritmo de clustering llamado DBSCAN con la ayuda de las librerias de `scikit-learn`, unas herramientas open source que facilitan el análisis predictivo de datos.
Para lograr un buen resultado y obtener unas agrupaciones justas y con las que podamos trabajar, es necesario tener en cuenta dos factores:

- **Mínimo de Vecinos.** Este parámetro indica el número mínimo de puntos que deben estar en un área para que su agrupación pueda ser

considerada un cluster

- **Distancia.** Este parámetro marca la distancia máxima que puede haber entre dos puntos para que estos puedan ser agrupados.

En cuanto a la distancia, se ha tenido en cuenta que los datos a agrupar son latitud y longitud, y por ello se ha recurrido al cálculo de la métrica mediante la fórmula del semiverseno.
Para afinar el clustering se han hecho pruebas y comparado resultados variando los dos parámetros antes mencionados hasta llegar a la conclusión de que el clustering más óptimo se obtenía con un mínimo de vecinos de 15 y una distancia de 17Km.

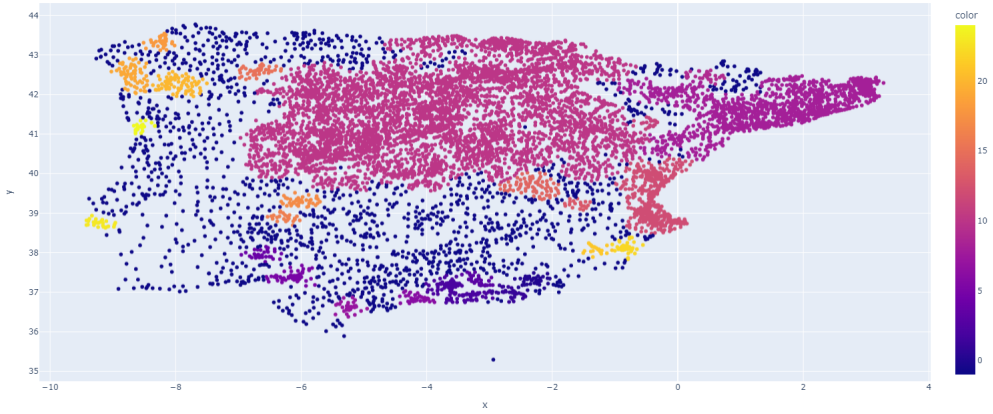


Figure 2: Resultado del clustering por DBSCAN

Las zonas más densas han sido agrupadas, mientras que las zonas menos pobladas se han etiquetado como ruido. Para agrupar todos los municipios considerados ruidos es necesario realizar operaciones adicionales. Para ello, hemos seleccionado un punto fijo tomando como eje horizontal la longitud y como eje vertical la latitud gracias a los datasets anteriormente mencionados. Este punto representa aproximadamente el centro de España. A partir de este punto, se han dividido todos los outliers en 4 grupos por su localización (Noreste, Sureste, Noroeste, Suroeste) De esta forma obtenemos un conjunto de zonas que abarca todos los municipios de la península.

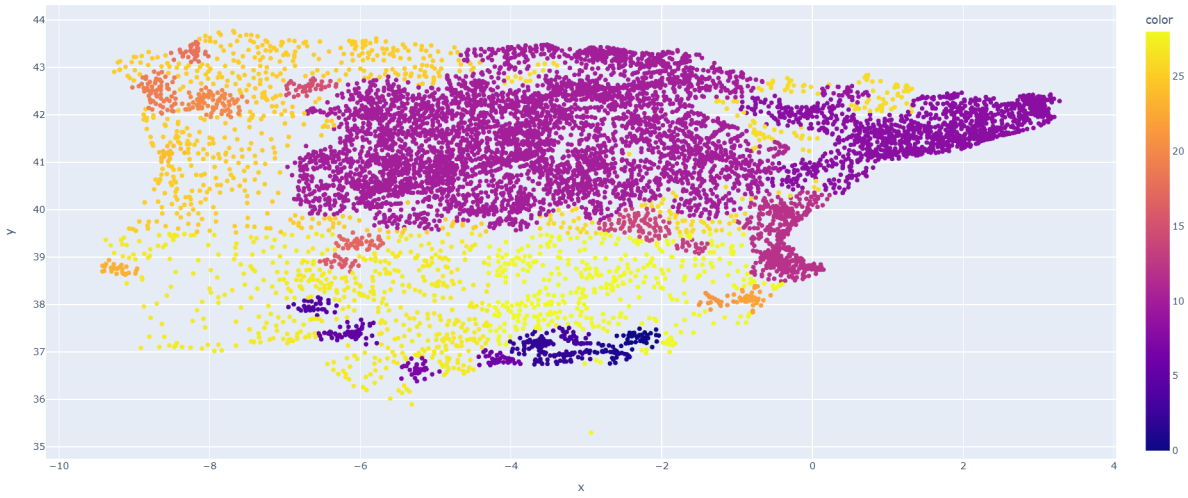


Figure 3: Resultado final de las zonas

7 Tarjeta de Datos. Características.

Como ya hemos visto en el apartado anterior, en nuestra tarjeta de datos mantendríamos todos las columnas que tenía el conjunto de datos original, pero hemos añadido algunos campos que hemos considerado que nos harían falta en un futuro, y estos campos son los que se explican en la tabla siguiente.

Cabe destacar que hemos realizado operaciones con respecto a los datos recogidos en los distintos datasets para obtener una distancia precisa entre el municipio de origen y destino con el fin de ampliar el resultado final del proyecto, sin embargo hemos encontrado limitaciones con respecto a las APIs usadas y hemos decidido no incluir este atributo en la tarjeta de datos.

SEMANA_AÑO	Variable numérica que representa la semana del año en la que se realiza el viaje
DIA_SEMANA	Variable numérica que representa el día de la semana en la que se ha realizado el viaje
FLUJO	Variable numérica, indica si el viaje es interno, saliente si tiene origen en CLM hacia otra comunidad, o entrante si tiene origen en otra comunidad hacia CLM
ZONA_DESTINO	Variable numérica que representa la zona a la que pertenece la ciudad de destino del viaje
ZONA_ORIGEN	Variable numérica que representa la zona a la que pertenece la ciudad de origen del viaje
VOLUMEN	Variable numérica que representa el número total de viajes con misma zona de origen y destino en el mismo día

8 Varias opciones de modelado

Una vez que hemos obtenido nuestra tarjeta de datos, hemos planteado dos opciones para lograr el resultado que deseamos del proyecto. Toma una gran importancia el atributo de Volumen, el cual representará la cantidad de viajes en un día de la semana del año, que se realizan desde una ciudad de una zona de las que hemos hablado anteriormente, a otra ciudad que puede pertenecer o no a la misma zona que la de origen.

8.1 Opción 1. Clasificador Naïve Bayes

Tras esta definición del nuevo atributo, surge la primera idea de utilizar un clasificador Naive Bayes, que haciendo uso de las probabilidades que se calculan a partir del Volumen de los viajes explicados anteriormente. Antes de crear el modelo Naive Bayes, recogimos los datos agrupando los volúmenes de viajes en día concreto, por cada una de las zonas de origen a cada una de las zonas de destino, y obtuvimos el porcentaje de viajes en ese día, desde una zona origen a cada zona de destino, lo cual un nos facilitaría y ayudaría en gran medida le resolución del problema mediante el clasificador Naive Bayes. La tabla resultante sería exactamente igual que la tarjeta de datos anteriormente descrita, con la inclusión de un nuevo campo:

PORCENTAJE	Variable en punto flotante que representa el número de viajes que se realizan desde una zona origen en un día determinado, a una zona destino, sobre el total de viajes que se realizan desde esa zona origen a todas las zonas destino.

8.1.1 Resultados no deseados. Cambio a Random Forest

Una vez que se obtuvo la tabla, se trabajó en la construcción del modelo Naive Bayes para la clasificación de las zonas de destino. Tras la construcción del modelo, pasamos a comprobar la precisión de este, obteniendo resultados bastante negativos, lo cual nos llevó a abortar esta solución y optar por la que finalmente hemos realizado.
Esta nueva solución es bastante parecida a la anterior; consiste básicamente en utilizar un modelo de regresión Random Forest para predecir el volumen de viajes que habrá desde una zona origen, (correspondiente a los municipios de Castilla-La Mancha) a una cualquier zona destino, es decir, teniendo en cuenta únicamente el flujo de salida, en un día determinado. Para esta solución, hemos utilizado una tarjeta de datos que es muy similar a la que acabamos de ver, la cual se explica a continuación.

ZONA_ORIGEN	Variable numérica que equivale a la zona donde comienza un viaje, que corresponde a los clústeres que se habían calculado
ZONA_DESTINO	Variable numérica que equivale a la zona donde termina un viaje, que corresponde a los clústeres que se habían calculado
DIA_SEMANA	Variable numérica que representa el día de la semana en el que se produce un viaje. Los valores se distribuyen de la siguiente manera: 0: Lunes, 1: Martes, 2: Miércoles, 3: Jueves, 4: Viernes, 5: Sábado, 6: Domingo
SEMANA	Variable numérica que representa la semana del año en el que se realiza un viaje. Tiene valores del 1 al 52.
VOLUMEN	Variable numérica que representa el número total de viajes con misma zona de origen y destino en el mismo día

Como vemos, nos desprendemos de la variable 'PORCENTAJE', dejando una tarjeta de datos con **4370 registros** en la que hemos agrupado los viajes entre las distintas zonas en los días determinados, obteniendo la suma de los viajes que se llevan a cabo en la columna del atributo 'Volumen'.
También es importante decir que, a la hora de mirar esta tabla, pueden surgir algunas dudas sobre la correctitud de esta, ya que teniendo en cuenta únicamente el flujo de salida no es posible que se lleve a cabo viajes entre dos mismas zonas. Bien, pues esto no es así del todo, ya que aunque la pueda parecer que no se trata de un flujo de salida de Castilla-La Mancha, si lo es realmente; como pasa con el caso de la zona 28, en la que se llevan a cabo viajes entre dos zonas 28. Esto se debe a que esta zona corresponde con la parte limítrofe del sur de Castilla-La Mancha, con el norte de Andalucía, más concretamente la zona norte de Jaén, por lo que, mirando la tarjeta de datos previa, encontramos por ejemplo como hay viajes que tienen origen en Villamanrique, (Ciudad Real), con destino en Úbeda, (Jaén), de tal manera que se considera como flujo de salida, con lo cual, realmente es correcto.

8.1.2 Construcción del modelo

Como ya hemos comentado, hemos utilizado un modelo de regresión Random Forest para el cálculo del volumen, o, en otras palabras, la cantidad de viajes con flujo de salida que se producen de una zona a cualquier otra, en un mismo día. Para ello, hemos divido la tarjeta de datos anterior en dos partes; una para entrenar al modelo, y otra para probar el funcionamiento y evaluar al mismo. En cuanto a esta evaluación, se ha calculado el error absoluto medio, dando como resultado un valor menor a lo que equivaldría a viaje y medio, lo que es una gran señal, dando a entender que este modelo va a poder sernos de ayuda en cuanto a la predicción.

```
Error Measure  1.3489702517162472
```

Figure 4: Cálculo del error medio absoluto

También se ha dibujado una gráfica, en la cual se muestra las predicciones que ha hecho nuestro modelo en comparación con los datos originales que teníamos en la tarjeta de datos. En esta gráfica, lo que podemos ver, es como el modelo predice bastante bien los viajes que se van a realizar en un día concreto, salvo en algunos casos en los que el numero de viajes en ese día

se disparan. En estos casos, nuestro modelo entrenado, realiza predicciones que se acercan, pero no son tan certeras como las de los movimientos con un volumen bajo. La gráfica de la que se ha hablado se muestra a continuación.

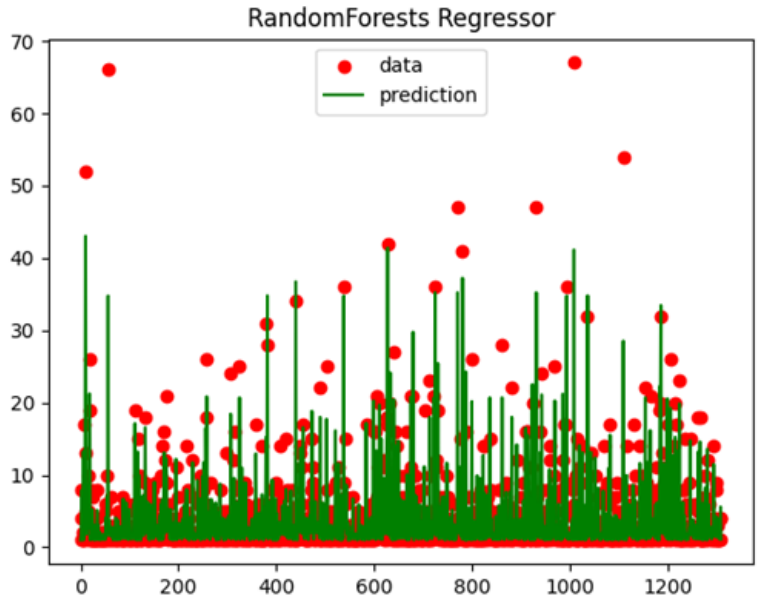


Figure 5: Representación visual de la predicción

Una vez que hemos entrenado el modelo, es hora de predecir el volumen de los viajes para cada una de las zonas. Para llevar a cabo esta tarea, hemos decidido coger un solo punto de origen, y un mismo día, y predecir el volumen de viajes que se llevaran a cabo para cada una de las zonas de destino de las que tenemos registros, de esta manera evitaremos errores en las predicciones. Una vez que hemos obtenido una predicción para cada zona de destino, guardaremos esos datos en un conjunto de datos para ahora trabajar sobre él.
Lo primero que vamos a hacer es sumar todos los valores que se han predicho usando el modelo, obteniendo así el total de viajes que habrá desde una zona en un día concreto; para posteriormente conseguir una nueva columna en la que podamos ver el porcentaje de viajes que corresponden a cada zona de destino partiendo de la zona de origen y día asignados. En el apartado poste-

rior se verán un par de ejemplos para probar el funcionamiento y sacaremos conclusiones sobre los mismos.

8.2 Opción 2. Tabla Pivotada

La otra opción que teníamos en mente era lograr nuestro objetivo principal a partir de una tabla pivotada según el volumen de viajes a cada zona destino distinta. Para esto, fuimos un paso más adelante en el filtrado de datos para reducir el número de instancias con las que trabajar centrándonos solo en el flujo saliente, es decir, todos los viajes con origen en CLM hacia cualquier comunidad autónoma. Además, nos vamos a centrar en los viajes que tienen de origen las capitales de provincia. Esto nos deja una tabla con 52 filas por capital de provincia, una para cada semana del año. El resultado se ve algo así:

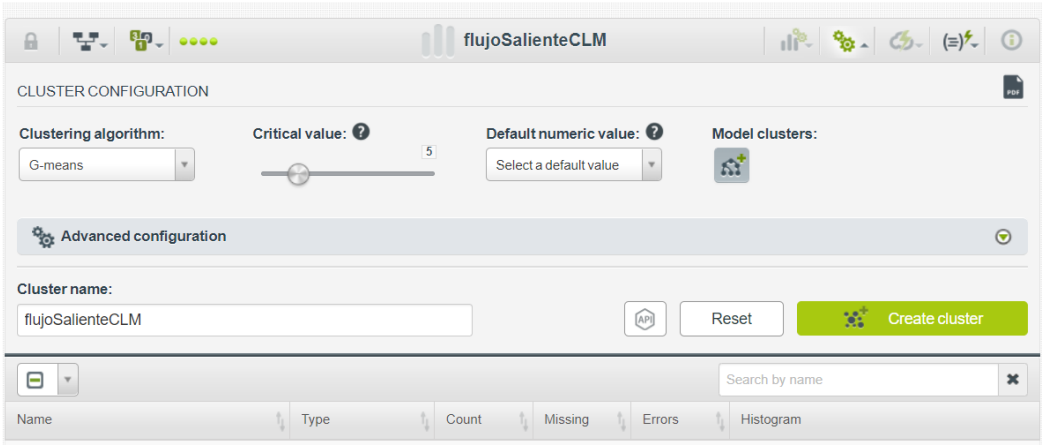
ZONA_DESTINO		0	1	2	3	5	6	7	8	9	10	12	13	14	15	16	17	18	19	20	21	22
ORIGEN	FLUJO	SEMANA																				
Albacete	1	1	0.0	3.0	6.0	0.0	1.0	4.0	0.0	7.0	0.0	40.0	9.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	13.0	5.0
		2	0.0	0.0	5.0	0.0	1.0	0.0	0.0	2.0	0.0	16.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	7.0	6.0
		3	1.0	1.0	3.0	0.0	0.0	2.0	0.0	4.0	0.0	16.0	5.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	7.0	5.0
		4	0.0	3.0	6.0	1.0	3.0	0.0	0.0	3.0	0.0	20.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	4.0
		5	0.0	1.0	5.0	0.0	2.0	0.0	0.0	1.0	0.0	16.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0	2.0

Figure 6: archivo flujoSalienteCLM.csv

Como podemos observar, cada zona destino tiene su propia columna y el indice está compuesto por los campos ORIGEN, FLUJO y SEMANA del archivo *tabla_pivotada.csv* de forma conjunta.

8.2.1 Clustering Jerárquico.

Una vez tenemos en un dataset los datos que necesitamos, realizamos un pequeño clustering jerárquico haciendo uso de las herramientas que nos ofrece BigML para ver el resultado. Subimos el dataset y procedemos a realizar el clustering por volúmenes:



Esto nos proporciona de resultado 3 clusters en los que se quedan agrupadas las capitales Ciudad Real y Toledo por un lado, Cuenca y Guadalajara por otro, y Albacete en el último cluster. En los datos se puede apreciar que Ciudad Real y Toledo tienen más tendencia a las zonas del sur, mientras que el destino de viajes originados en Cuenca y Guadalajara tiende más a ser en el norte. Con albacete se ve una mezcla de ambas. Este dataset se encuentra almacenado en *jerarquico.csv*.

8.2.2 Clasificador, Resultados no deseados

Para este dataset que hemos obtenido también hemos intentado realizar la construcción del modelo de predicción con un clasificador, pero debido a la bajos niveles de confianza que daban los resultados al igual que con la opción anterior, decidimos descartarlo. Por razones de tiempo no hemos podido lograr la construcción de un modelo de predicción para esta opción, pero los datos se encuentran ya transformados y agrupados para fácilmente aplicar las últimas técnicas necesarias para obtener resultados.

9 Resultados e Interpretación.

En este apartado, vamos a mostrar los experimentos realizados con los modelos descritos anteriormente, para ver así el funcionamiento de estos y poder sacar conclusiones sobre ellos. En primer lugar, empezaremos viendo dos ejemplos sobre el algoritmo de regresión que hemos explicado anteriormente,

para ver la diferencia en los volúmenes de viajes de un día normal, y un día relacionado con festividades. Seleccionamos un día normal, por ejemplo, un domingo de marzo. Para ello, el valor del atributo 'DIA_SEMANA' será 6, mientras que el del atributo 'SEMANA' tendrá valor 10, seleccionando como zona origen la zona correspondiente al clúster 28. Con esta configuración, obtenemos como resultado los siguientes datos:

- Un total de 107 viajes en ese día partiendo de esa zona.
- 34,58% son para trasladarse a la zona correspondiente al clúster número 10, (zona del norte de Madrid, correspondiente al Castilla y León, y su continuación hacia el este de la meseta, lo cual tiene sentido, ya que este clúster es más grande que el resto y agrupará más viajes, y su mala conexión de medios de transporte públicos, por lo que la gente busca soluciones alternativas como el uso de BlaBlaCar).
- 16,82% tienen como destino la propia zona 28, para trasladarse entre municipios del sur de Castilla-La Mancha y noreste de Andalucía, (zonas que no destacan por su cobertura de transporte).
- 7,47% tienen como destino la zona 2, (sur de Andalucía, zonas de Almería, Córdoba y Granada), y un 5,60% a la zona 21, correspondiente a la costa de Cartagena, Murcia.
- El resto de los destinos se reparten equitativamente, en torno a 1 y 3 viajes, de manera que ninguno destaca.

Ahora seleccionamos un día relacionado con un festivo, por ejemplo, el día de reyes, (6 de Enero). Para hacer esto, el valor del atributo 'DIA_SEMANA' será 6, mientras que el del atributo 'SEMANA' tendrá valor 1, seleccionando nuevamente como zona origen la zona correspondiente al clúster 28. Con esta nueva configuración, obtenemos como resultado los siguientes datos:

- 127 viajes en ese día partiendo de esa zona. 20 más que con la anterior configuración.
- 25,80% tendrán como destino la zona 10, la cual anteriormente suponía un porcentaje mayor, por lo que podemos observar que ya se están repartiendo los viajes de manera más equitativa.

- 15,32% irán destinados a la zona 28, que al igual la zona anterior, aún habiendo disminuido el porcentaje respecto del total, sigue subiendo el numero de viajes, es decir, el volumen.
- En cuanto al resto de zonas de destino, vemos como sigue aumentando el numero de viajes que reciben con respecto al ejemplo anterior, suponiendo así un incremento global en el número de viajes en esa fecha.

Tras este ejemplo, vemos perfectamente como las predicciones que hace el sistema son distintas para cada uno de los datos de entrada, teniendo como conclusión que, en fechas clave se produce un evidente aumento del numero de viajes de salida, lo que se puede traducir, por ejemplo, en la mejora de campañas de publicidad para determinados destinos.

10 Conclusiones

Gracias al proceso KDD podemos decir a modo de conclusión que, tras aplicar técnicas de preproceso y transformación al conjunto de datos inicial, del cual, en un primer momento parecía que no se podía sacar nada en claro debido a su simplicidad, hemos llegado a producir dos tarjetas de datos que aportan el conocimiento suficiente para poder construir modelos que sirvan para predecir el comportamiento de los movimientos con flujo de salida desde Castilla-La Mancha en fechas de importancia. Esto nos permite también retroceder y volver a realizar cualquier parte del proceso para mejorar los resultados o ampliar conocimientos sin tener que modificar por completo la estructura, lo cual es muy útil. También ha servido para aprender y conocer de primera mano en que consiste el proceso KDD, haciendo que mediante la realización de la práctica se lleguen a conclusiones gracias a los datos que hemos ido transformando, como por ejemplo el hecho de que en fechas cercanas a esas fechas clave de las que hemos ido hablando, haya un aumento en el volumen de los viajes.