

## Trabajo Fin de Máster

Evaluación comparativa de diferentes métodos para sistemas de recomendación aplicados a la movilidad urbana y su posible inclusión en la cadena de negocio

Comparative evaluation of different methods for recommendation systems applied to urban mobility and their possible inclusion in the business chain

Autor

Diego Sangüesa Pérez

Director

Sergio Ilarri Artigas

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2019



# Agradecimientos

Agradezco a la Universidad de Zaragoza, y más en concreto a la EINA por la oportunidad de llevar a cabo este proyecto. También quiero dar las gracias por supuesto a mi director Sergio Ilarri Artigas, que en todo momento me ha guiado y aconsejado para poder desarrollar el trabajo, y ha estado disponible siempre para que pudiera consultarle mis dudas. Por último, doy las gracias también a mi familia y amigos, que me han apoyado y animado a lo largo de todo el proceso.



# Resumen

El auge de las tecnologías de la comunicación ha hecho que generar y compartir información esté al alcance de todo el mundo, de forma que la cantidad de datos generados está creciendo exponencialmente. Para poder enfrentarse a toda esta información surgen los sistemas de recomendación, que ayudan al usuario a encontrar elementos (productos y servicios, actividades a realizar, etc.) de su interés reduciendo enormemente el coste de su búsqueda, y por lo tanto aportan comodidad y hacen accesible la exploración de datos. Estos sistemas, a partir de información conocida, consiguen predecir los intereses del usuario, para así mostrar los ítems que le resulten relevantes antes que el resto.

En este trabajo se propone utilizar las predicciones obtenidas de este tipo de sistemas para aplicarlas en el ámbito de la movilidad urbana; ya que, si se pueden conocer los intereses de los usuarios de antemano, se podrá prever por qué zonas y cómo se moverán un determinado grupo de turistas por la ciudad, y así se podrán preparar los servicios de movilidad existentes para optimizar su rendimiento.

Para conseguir este objetivo, en primer lugar se han estudiado las bases de funcionamiento de los sistemas de recomendación y se han visto las diferentes metodologías existentes, comprendiendo sus diferencias y las ventajas y desventajas que presentan unas u otras.

Se ha experimentado con algunos de los métodos de recomendación anteriormente nombrados, utilizando el lenguaje de programación R, y operando con un conjunto de datos existente en Internet, para poder conocer de forma práctica cómo se usan los sistemas de recomendación, comprender qué pasos siguen y qué parámetros se utilizan, y analizar los resultados obtenidos comparándolos entre sí.

Además, se han analizado los posibles casos de aplicación de estos sistemas en el contexto de la movilidad urbana, viendo cómo pueden ayudar y qué pueden aportar a este campo, con qué datos pueden trabajar y de dónde se podrían obtener estos, como afectarían a los procesos actuales de toma de decisiones, qué beneficios podrían aportar al mercado y cuáles serían sus costes.



# Índice

<b>1.</b>	<b>Introducción y objetivos</b>	<b>1</b>
1.1	Contexto . . . . .	1
1.2	Motivación . . . . .	1
1.1	Fases del trabajo y estructura del documento . . . . .	4
<b>2.</b>	<b>Sistemas de recomendación</b>	<b>6</b>
2.1	Descripción general . . . . .	6
2.2	Métodos de recomendación . . . . .	7
<b>3.</b>	<b>Comparación experimental de distintos métodos de recomendación</b>	<b>14</b>
3.1	Entorno y conjunto de datos . . . . .	14
3.2	Métodos de recomendación utilizados y tratamiento de datos . . . .	16
3.3	Obtención de predicciones y análisis . . . . .	20
3.4	Influencia del número de valoraciones conocidas . . . . .	23
<b>4.</b>	<b>Aplicación al campo de la movilidad urbana</b>	<b>26</b>
4.1	Casos de aplicación . . . . .	26
4.2	Origen y naturaleza de los datos . . . . .	29
4.3	Influencia en la toma de decisiones y el mercado . . . . .	31
<b>5.</b>	<b>Conclusiones y posibles trabajos futuros</b>	<b>34</b>
5.1	Conclusiones finales . . . . .	34
5.2	Valoración personal . . . . .	35
5.3	Trabajos futuros . . . . .	36
	<b>Bibliografía</b>	<b>37</b>

<b>Lista de figuras</b>	<b>42</b>
<b>Anexos</b>	<b>44</b>
<b>A. Experimentación</b>	<b>46</b>
<b>B. Código utilizado</b>	<b>52</b>
<b>C. Conjuntos de datos</b>	<b>60</b>
<b>D. Métricas de resultados</b>	<b>62</b>





# Capítulo 1

## Introducción y objetivos

En este primer capítulo se va a explicar el contexto que enmarca a este trabajo en la Sección 1.1, los motivos que llevan a realizarlo en la Sección 1.2 y las fases que se han planteado para su estructura en la Sección 1.3.

### 1.1. Contexto

La aparición de nuevas tecnologías es cada vez más rápida, y las aplicaciones de estos avances están muy presentes en nuestra vida cotidiana. Gracias a ellas la transmisión de información es cada vez más fácil y rápida, motivo por el cual la información disponible sobre prácticamente cualquier tema es mucho más abundante hoy en día que hace unos cuantos años. Por este motivo se utilizan los sistemas de recomendación, que ayudan a navegar entre tal cantidad de datos, proporcionando de manera prioritaria los elementos que resultarán más relevantes para el usuario. No obstante, el reto actual consiste en sacar provecho de toda esta cantidad de información, ya que gran parte de la misma no es analizada ni procesada, y se le pueden encontrar muy diversas utilidades.

En este caso, se va a analizar la información con la intención de aplicarla a la movilidad urbana. La cantidad de población que vive en ciudades aumenta constantemente, mientras que las zonas rurales de muchos países desarrollados van quedando cada vez más despobladas. Es por esto que el transporte tanto de mercancías como de personas en la ciudad es un reto cada vez mayor, de manera que hay que aplicar todas las técnicas y tecnologías disponibles e intentar aprovechar toda la información al alcance para optimizarlo, reduciendo así tiempos y costes, y en general, aumentando la calidad de vida de las personas.

### 1.2. Motivación

Uno de los puntos más influyentes en el transporte por ciudad es la planificación urbana, ya que decide cómo van a ser las carreteras, dónde se ubicarán los aparcamientos y cuántos se habilitarán, dónde se van a encontrar las zonas de viviendas y las zonas de interés cultural o de afluencia de público, etc. En general, de este diseño va a depender la fluidez del transporte por toda la ciudad. Este tipo de decisiones son críticas cuando se va a producir algún evento en una ciudad, como podrían ser un gran concierto de un grupo

muy multitudinario, una ceremonia o gala de entrega de premios de algún tipo, la celebración de competiciones deportivas excepcionales o simplemente una festividad en una ciudad con mucho turismo, ya que en estos casos la afluencia de gente por la ciudad se puede multiplicar bastante y además de forma muy repentina, al contrario que el crecimiento natural de una ciudad que sería más paulatino, por lo que hay que planificar y preparar la recepción y la movilidad de todo ese público extra.

Para esta planificación es necesario conocer datos acerca de la cantidad de personas que van a confluír en una misma zona y por dónde se van a mover, para poder prever las partes de la ciudad con más congestión y así poder evitarlo. El problema reside en que en muchas ocasiones estos datos no se pueden conocer de antemano, dificultando enormemente la toma de decisiones. En estos casos una posible opción es el uso de datos históricos de situaciones similares ocurridas en el pasado, buscando eventos que atraigan a la misma cantidad y tipo de personas, y en la misma ciudad o alguna con una estructura parecida; y a partir de estos datos intentar extrapolar cómo se va a comportar la gente en el caso en el que estemos trabajando [31]. Esto puede ser una buena solución en algunas ocasiones, pero en otras es posible que los resultados no se correspondan con el comportamiento posterior de la gente, a pesar de haber encontrado eventos similares en la misma ciudad. Por ejemplo, si ha transcurrido mucho tiempo desde el momento en el que se tomaron los datos, puede que la cultura de la ciudad haya cambiado y las zonas que antes eran modernas y atractivas para los turistas en la actualidad se hayan quedado más desfasadas. Otra opción es que se trate de un evento que atraiga a gente de una determinada edad, pero los gustos y aficiones propios de esta edad hayan cambiado, y las personas que vengan en la actualidad no se interesen por las mismas zonas por las que lo hacían gente con la misma edad, pero hace 20 años. En general el uso de datos históricos puede ser un buen indicador si no se dispone de ningún dato en absoluto, pero no es tan fiable como para tomar decisiones de gran importancia basándose exclusivamente en ello, por lo que es necesario estudiar y comparar en profundidad los casos y aun así nada garantiza que los resultados sean los esperados.

En este trabajo, se propone el uso de otro tipo de información complementaria para estos casos, que si se usa de forma correcta puede ser más fiable, y además lo será aún más en el futuro cercano gracias al crecimiento y mejora de la tecnología. Se trata de las predicciones generadas por los sistemas de recomendación. Esta tecnología es una forma de lidiar con la enorme cantidad de información y de opciones de las que se dispone hoy en día, facilitando a las personas encontrar un producto o servicio que satisfaga sus necesidades de forma mucho más rápida y cómoda. En general, estos sistemas se basan en el filtrado de toda la información disponible para poder mostrar al usuario los datos que más le van a interesar o más relevantes van a ser para él, los cuales calcula a partir de los datos que se dispongan de ese usuario junto con los datos generales de todos los usuarios.

Para este caso relacionado con la optimización de la movilidad urbana, se usarán predicciones generadas por sistemas de recomendación [7] que trabajen con alojamientos turísticos, restaurantes, puntos de interés cultural de la ciudad, zonas de ocio, etc. para así poder conocer cómo se moverán las personas por la ciudad de antemano y poder prepararla en consecuencia. Estos sistemas suelen estar gestionados por empresas privadas, aunque también pueden estarlo por entidades públicas como oficinas de turismo, y están disponibles para que la gente pueda acceder a ellos a través de Internet, normalmente de forma gratuita, para poder encontrar los elementos que desean de forma mucho más fácil y rápida. De este modo, las personas que usan estos servicios encuentran lo que buscan con mayor comodidad y seguridad de que va a cumplir sus expectativas, y las páginas o aplicaciones que lo ofrecen obtienen visitas de forma que pueden obtener beneficio a través de la publicidad, o incluso pueden ofrecer el servicio de contactar o reservar el producto de interés para el usuario, ahorrándole a éste las gestiones y consiguiendo una comisión de parte del producto contratado a cambio de traer clientes que quizá no hubieran llegado de otro modo.

Por todo ello, estos sistemas son cada vez más utilizados por la gente cuando está buscando un determinado servicio o producto. Cuanta más gente los usa de más datos disponen, y por lo tanto mejores pueden ser las recomendaciones, haciéndolos así más fiables. Así, este tipo de sistemas disponen de una gran base de datos con los productos o servicios que se buscan y los usuarios con sus búsquedas y productos contratados, visitados o consumidos anteriormente, y son capaces de generar predicciones para cada usuario acerca de los productos que le interesarán en un futuro. Queda claro pues que esto es de gran interés para los encargados de la planificación urbana que se comentaba anteriormente, por lo que estos podrían ponerse en contacto con las empresas que dispongan de estas bases de datos y predicciones, y adquirir toda esta información que les resulte de interés y utilidad para prever las necesidades de la ciudad de antemano, siempre que los datos compartidos sean anónimos y cumplan con la normativa de protección de datos vigente. Nuevamente, se puede comprobar que se trata de una situación de beneficio mutuo, ya que la empresa o entidad al cargo del sistema de recomendación vende parte de sus datos obteniendo así nuevos ingresos y sin ningún perjuicio, y los trabajadores de la ciudad encargados de gestionar la planificación para estos casos disponen de datos en los que basarse para la toma de decisiones.

Para asegurar la fiabilidad de los datos y evitar que se pueda desconfiar de estos métodos, se va a estudiar el funcionamiento de los sistemas de recomendación, viendo cómo trabajan y tratan los datos y las diferentes opciones disponibles, cada una con sus ventajas y desventajas, y se van a analizar y comparar experimentalmente algunos de los distintos métodos existentes, para comprobar que el error de los datos predichos es aceptable. Tras ello, se podrá garantizar la utilidad de las predicciones obtenidas mediante sistemas de recomendación en el contexto de la planificación urbana y la gestión de ciudades.

Esta experimentación se va a llevar a cabo utilizando el lenguaje de programación R [12], ya que dispone de una librería llamada “recommenderlab” [9] que está totalmente dirigida a los sistemas de recomendación y dispone de diversas funciones y estructuras muy útiles, además de distintos tipos de algoritmos de recomendación y diferentes conjuntos de datos de ejemplo para poder realizar pruebas de forma rápida y sencilla. Además, para hacer pruebas que permitan comparar distintos métodos de recomendación, se ha buscado un conjunto de datos relacionado con el turismo, como se ha comentado anteriormente, y finalmente se ha usado un conjunto de datos procedente de la web “TripAdvisor” [8] que contiene 1890 hoteles, 1202 usuarios y 4606 valoraciones de usuarios.

### 1.3. Fases del trabajo y estructura del documento

De esta forma, el primer paso a llevar a cabo es el estudio de los sistemas de recomendación en general. Se ha de analizar y comprender cómo se tratan los datos y se generan las recomendaciones y la predicción de las valoraciones para cada usuario. Asimismo, también se van a comentar diferentes métodos de recomendación que pueden usar estos sistemas y las diferencias entre ellos, además de algunas ventajas y desventajas. No se trata de realizar un estudio del estado del arte completo de los sistemas de recomendación, ya que el objetivo del trabajo no es el desarrollo de un nuevo tipo de sistema de recomendación ni una mejora de los existentes; simplemente se va a estudiar su funcionamiento y se van a comprender algunos de los métodos más utilizados para validar la aplicación de estos sistemas al campo de la movilidad urbana.

El siguiente paso del trabajo consistirá en la experimentación con algunos de los tipos de sistemas de recomendación comentados anteriormente. Para ello se ha utilizado un conjunto de datos obtenido a través de Internet que contiene las valoraciones de hoteles de un conjunto de usuarios procedentes de la web TripAdvisor [8], ya que son datos relevantes en el contexto de la planificación de la movilidad urbana. Con esta base de datos, se aplicarán distintos sistemas de recomendación y se compararán los resultados obtenidos, para comprobar cuál presenta el menor error en sus predicciones, cómo afectan a éstas los distintos parámetros y las posibles ventajas que pueden presentar unos frente a otros. Para hacer esto, se usará una parte de los datos para entrenar los sistemas de recomendación, a partir de valoraciones conocidas, y luego, se usará la parte restante para hacer las predicciones y compararlas con los datos reales, como se hace en general para evaluar técnicas de aprendizaje automático [6].

Tras esto, se estudiará la posible aplicación de este tipo de datos en el contexto de la movilidad urbana. Se analizará y explorará qué tipo de datos podrían ser útiles y de dónde obtenerlos, en qué casos se podría recurrir a este tipo de datos, y cómo se podrían utilizar los mismos para obtener beneficios en la planificación de la movilidad en ciudades, considerando posibles cadenas de negocio derivadas de ello. Esta será una forma de sacar más provecho de la gran cantidad de información que se genera hoy en día, dándole

utilidad en un campo como es el de la movilidad, consiguiendo así una mejora en ambas áreas de trabajo, ya que las empresas que dispongan de datos podrán venderlos, siempre de acuerdo a la LOPD [20] y acumular más ingresos, y los encargados de la movilidad dispondrán de más información con la que trabajar; generando nuevos puestos de trabajo relacionados con la gestión y el análisis de datos (científico de datos o *data scientist*) [30] en las empresas relacionadas con la movilidad y consiguiendo una mejora para la sociedad.

Por último, se extraerán una serie de conclusiones de los estudios realizados, por una parte en cuanto a la experimentación con los sistemas de recomendación, para comprobar cuáles podrían ser más apropiados para este contexto y cuáles no cumplirían todos los requisitos, y por otra parte en cuanto a la viabilidad y las posibles aplicaciones del uso de estas predicciones generadas por sistemas de recomendación para la toma de decisiones en el ámbito de la movilidad, así como posibilidades de evolución futura de estos temas y posibles trabajos relacionados. También se hará una valoración personal del trabajo, para comprobar cómo ha incidido en mis estudios y mi preparación para el mercado laboral.

De esta forma, en el Capítulo 2 se expondrá el estudio acerca de los sistemas de recomendación, en el Capítulo 3 se explicarán las pruebas realizadas que han dado mejores resultados, cómo se han hecho y las conclusiones a las que se llega. En el Capítulo 4 se explorarán diferentes ideas y opciones para sacar provecho de estas predicciones en el ámbito de la movilidad. En el Capítulo 5 se mostrarán las conclusiones a las que se ha llegado durante la realización del trabajo, así como futuros trabajos posibles y una valoración personal. También se puede consultar el Anexo A, en el que se muestran otras pruebas realizadas pero de menor importancia o con resultados complementarios menos relevantes, así como algunos pormenores del proceso por el que se ha llegado a las pruebas del Capítulo 2. El Anexo B muestra el código utilizado para llevar a cabo estas pruebas. En el Anexo C se comentan los motivos de la elección del conjunto de datos utilizado [8]. En el Anexo D se explican algunas de las métricas utilizadas para analizar los resultados.

# Capítulo 2

## Sistemas de recomendación

En este capítulo se va a hablar sobre los sistemas de recomendación, haciendo primero una descripción general y tratando su utilidad y principios de funcionamiento en la Sección 2.1, y posteriormente se nombrarán y comentarán brevemente distintas tipologías en la Sección 2.2.

### 2.1. Descripción general

En la actualidad se genera mucha más información que hace unos cuantos años, y esto va a más, de forma que cada vez es más difícil gestionar todos esos datos, y gran parte de ellos están sin explotar. A nivel de un usuario individual, la cantidad de información sobre cualquier tema a la que se tiene acceso al entrar a Internet es abrumadora, y no es posible abarcarla toda de forma sencilla, por lo que en muchos casos esta cantidad de datos resulta contraproducente y dificulta acceder a lo que se buscaba en un principio en vez de facilitarlo. Para solucionar esto existen navegadores y buscadores que filtran y ordenan todo este contenido para hacerlo más accesible y permitir encontrar lo que se busca de manera sencilla.

Los sistemas de recomendación [2] surgen con el mismo propósito, ayudar a lidiar con el exceso de información para permitir a los usuarios encontrar lo que puede resultarles relevante de manera más rápida y sencilla. Estos sistemas se suelen encontrar, por ejemplo, en entornos que ofrecen un gran catálogo de productos o servicios, lo que justifica su presencia para poder navegar entre todas estas opciones. Una particularidad que presentan los sistemas de recomendación en general es la necesidad de datos acerca de las valoraciones (en la gran mayoría de casos), no solamente los datos sobre los productos, es decir, conforme se van usando van generando una base de datos que almacena datos de los usuarios y las valoraciones que estos van haciendo de cada producto. Estas valoraciones pueden ser binarias, es decir, si un determinado usuario ha comprado o consumido un determinado producto o no, o si le ha gustado o no; o valoraciones explícitas que se le piden al usuario sobre un producto dentro de una determinada escala, las cuales aportan más riqueza de información al sistema y pueden hacer que los resultados sean más precisos. Ejemplos de estos sistemas de recomendación se pueden encontrar en webs como YouTube [16] o Netflix [17] para los sistemas binarios ya que contabilizan si un usuario ha visto un determinado video o no (aunque en Netflix también se pueden dar valoraciones), o en webs de valoraciones de películas como FilmAffinity [18] o en el propio Google Maps [19] que está disponible en los teléfonos móviles para valoraciones en una escala de 1 a 5. Queda claro pues que los sistemas de recomendación están muy presentes en nuestra vida cotidiana, y van a estarlo cada vez

más. También es importante distinguir entre valoraciones explícitas (*explicit feedback*), que son las que dan los usuarios de forma consciente, y valoraciones implícitas (*implicit feedback*), las cuales son registradas de forma autónoma por el sistema en función de las acciones del usuario.

De esta forma, estos sistemas de recomendación van creando un conjunto de datos en el que aparecen todos los productos o servicios disponibles, así como todos los usuarios que han utilizado alguna vez el sistema y las valoraciones de cada uno de ellos para cada uno de los productos consumidos. A partir de todos estos datos, el sistema es capaz de realizar predicciones acerca de las valoraciones que darán los usuarios a productos que aún no han valorado, y a partir de ellas recomendar productos y ordenarlos de forma que aparezcan antes los que tienen más probabilidades de satisfacer a un determinado usuario o se espera que le vayan a gustar más. La clave reside pues en cómo realiza el sistema de recomendación estas predicciones de las valoraciones antes de que el usuario haya consumido y valorado ese producto realmente, para lo cual existen diversos métodos, cada uno con sus particularidades, y dependiendo de cual use un determinado sistema de recomendación sus prestaciones, necesidades y los resultados proporcionados serán muy distintos.

## 2.2. Distintos métodos de recomendación

A continuación se van a ir nombrando y comentando algunos de los métodos de recomendación más utilizados, así como las ventajas y desventajas que estos presentan.

### **Recomendaciones no personalizadas**

En primer lugar, se encuentran los métodos más sencillos que no requieren datos acerca del usuario al que van a realizar las recomendaciones de productos (o lo que es lo mismo, para el que van a realizar la predicción de sus valoraciones), sino que se basan únicamente en los datos de los que ya disponen acerca de otros usuarios, es decir, no necesitan datos de entrada. Los ejemplos más claros de estos métodos sencillos podrían ser la recomendación aleatoria, la cual recomienda un producto seleccionado de forma aleatoria de entre todo el catálogo disponible y no tiene en cuenta ni el usuario al que la realiza ni las valoraciones previas de los otros usuarios; o la recomendación de los productos más populares, que ordena los productos disponibles en función de las valoraciones almacenadas previamente de todos los demás usuarios, ofreciendo así aquellos que tienen un mayor número de valoraciones y una media de puntuación más alta, de forma que también es un método independiente del usuario objetivo de las recomendaciones. En este método de recomendación por populares, ya existen algunas consideraciones que es necesario hacer, dado que recomendar simplemente el producto con mayor media de puntuación no es adecuado, ya que podría tratarse de uno que dispone de una única valoración pero con la puntuación máxima, lo cual no indica que ese producto guste mucho de forma generalizada sino que ha gustado mucho a un usuario en



concreto. De la misma manera, no es útil recomendar el producto con mayor número de valoraciones, ya que si la puntuación media no es buena indica que es un producto muy consumido pero que no cumple con todas las necesidades que los usuarios esperan satisfacer con él. Así, es necesario llegar a un compromiso entre los productos más consumidos por los usuarios y los productos con mejores puntuaciones, el cual dependerá del caso concreto de uso y del contexto, por lo que habrá que experimentar y ajustar estos parámetros de forma que se consigan los mejores resultados posibles para la situación y el entorno en el que se use. En general, estos métodos de recomendación son los más sencillos de implementar y en cuanto a su funcionamiento, por lo que también suelen ser los que presentan unos peores resultados, sobre todo el método aleatorio, que se suele usar como método base (*baseline*) para comparar con otros y para desarrollar métodos nuevos, ya que si algún método presenta unos resultados peores que el aleatorio no tiene sentido usarlo. No obstante, el método de recomendación por populares no siempre presenta unos resultados tan malos en comparación con otros, y en según qué entornos y contextos puede ser una buena solución y aportar suficiente fiabilidad.

### **Métodos demográficos**

Otro tipo de métodos de recomendación son los sistemas de recomendación demográficos, y es que estos últimos no tienen en cuenta las valoraciones previas del usuario objetivo, sino que utilizan sus datos personales para ubicarlo en un grupo demográfico, y una vez lo han hecho le recomiendan objetos que comparten todas las características en común y que en principio deberían satisfacer las necesidades de ese sector de la población en general. Es decir, no extrae las características de los objetos que ha valorado el usuario, sino que categoriza al usuario en un grupo, y con ello decide qué características deben tener los objetos a recomendar. De esta forma, este tipo de sistemas necesitan partir de una buena base de datos demográficos [28], teniendo bien definidos los distintos grupos a los que puede pertenecer el usuario, para que una vez este proporcione sus datos personales sea ubicado correctamente con la suficiente fiabilidad. Para que los resultados sean satisfactorios, también es de gran importancia seleccionar las características de los objetos a recomendar para cada grupo demográfico, para lo cual no hay que dejarse llevar por los estereotipos y es conveniente basarse en estudios y encuestas. También existe la opción de generar grupos de usuarios (*clusters*) a partir de sus gustos basándose en los datos que ya se conocen, y usar estos *clusters* para ubicar a los nuevos usuarios. Las principales desventajas que pueden presentar estos métodos son dos: por un lado las recomendaciones serán útiles y efectivas para la mayoría de los casos, pero siempre puede haber personas que no se identifiquen en absoluto con los gustos típicos del grupo demográfico al que pertenecen, y para ellos estos sistemas serán completamente inútiles; por el otro lado, para funcionar deben pedir al usuario una serie de datos personales, lo cual hace que parte de los potenciales usuarios eviten usarlos, claro que una vez han accedido al sistema y dado sus datos, si las recomendaciones son correctas, se crea una gran fidelidad por parte de los mismos. Por el contrario, estos sistemas también tienen algunas ventajas, siendo la principal que no requieren información histórica acerca de otras valoraciones, ya que mientras los grupos demográficos sean los adecuados, a los usuarios se les recomendarán objetos de su gusto aunque el sistema lleve funcionando poco tiempo y carezca de mucha información

anterior. Estos métodos no son adecuados para cualquier escenario o contexto, pero por ejemplo son bastante utilizados para sistemas de recomendaciones musicales y para recomendaciones turísticas, que es el caso que nos ocupa, y presentan buenos resultados.

### **Métodos basados en contenido**

Por otro lado, se encuentran los métodos de recomendación basados en el contenido [7], para los cuales se han de conocer valoraciones previas del usuario al que se quiere realizar la recomendación o bien disponer de un perfil de preferencias del usuario. La base de estos métodos consiste en encontrar productos, servicios o actividades, entre todos los disponibles en la base de datos que el usuario no ha valorado, que compartan características con aquellos que el usuario ha valorado positivamente o tengan características similares a aquellas que forman parte de su perfil preferencias, es decir, si a un determinado usuario le han gustado unos productos o servicios concretos, se le recomendarán otros parecidos o que tengan algo en común con aquellos que le habían gustado previamente. Este sistema puede ser apropiado para algunos casos, pero tiene la gran desventaja de que no es capaz de proporcionar recomendaciones de productos diferentes o nuevos que puedan ser de interés para al usuario, sino que recomienda productos similares que ya sabe que le gustarán. Por ejemplo en una web como YouTube puede ser útil este tipo de sistema de recomendación, ya que si un determinado usuario ha visto una gran cantidad de vídeos relacionados con el fútbol, por decir algo, si a ese usuario se le siguen recomendando otros vídeos de fútbol es muy probable que le interese verlos, de forma que pase más tiempo en la plataforma, con lo que YouTube cumple su objetivo (aunque a YouTube también le puede interesar aumentar la variedad de videos que consumen sus usuarios). En cambio, para una web de ventas como Amazon este sistema no es adecuado, ya que si un usuario compra por ejemplo una camiseta, no quiere decir que siempre vaya a comprar camisetas o ropa en general, sino que puede sentir interés por otro tipo de productos no relacionados, como por ejemplo un libro; es decir, que si ha comprado varias camisetas, no por seguir recomendándole camisetas sin parar va a comprar más, de forma que no cumpliría los requisitos que busca esta empresa.

### **Métodos colaborativos**

El siguiente paso son los métodos de recomendación colaborativos [1], los cuales se basan en calcular la similitud entre usuarios o entre productos, y a partir de ahí extrapolar los datos para obtener las predicciones para cada usuario. Por ello, estos métodos sí que requieren datos de entrada, es decir, también necesitan datos acerca del usuario para el que van a realizar las recomendaciones, así como las valoraciones de otros usuarios; por eso se llaman métodos colaborativos. Estos datos pueden ser valoraciones previas dadas por el usuario para otros productos o datos personales introducidos por usuario acerca de sus características o gustos a partir de los cuales poder calcular su similitud con otros. Como se ha adelantado antes, estos métodos colaborativos pueden ser de dos tipos: basado en lo usuarios, o *UBCF User Based Collaborative Filtering* y basados en los objetos, o *IBCF Item Based Collaborative Filtering*.

El *método colaborativo basado en usuarios* [24] actúa de la siguiente manera: para un determinado usuario, analiza las valoraciones previas del mismo, y busca a otros usuarios que hayan dado valoraciones similares a productos similares; así puede calcular los usuarios más parecidos a éste. Una vez se tiene esto, observa los objetos mejor valorados por los usuarios más parecidos al usuario objetivo y que éste no ha valorado, y así construye una lista con los objetos que debe recomendar. Para ello se puede usar la valoración media de entre todos los usuarios similares, o una media ponderada dando más peso a los usuarios más parecidos al usuario objetivo. Este es un buen sistema en cuanto a los resultados que ofrece, ya que las recomendaciones no son siempre de la misma temática como en el caso anterior, sino que están basadas en los gustos de usuarios con valoraciones pasadas similares a las del usuario objetivo, por lo que es de esperar que en el futuro vuelvan a coincidir. No obstante, este método presenta algunas desventajas, como que para cada nuevo usuario es necesario comparar sus valoraciones con todas las de los otros usuarios existentes en la base de datos, por lo que en sistemas muy grandes el tiempo necesario para hacer las recomendaciones puede ser superior al admisible. Otro problema es el derivado de ser un método colaborativo, por lo cual también lo presenta el sistema IBCF comentado a continuación, y es que si se dispone de pocos datos previos, es decir, cuando se comienza a usar el sistema y hay pocos usuarios, va a ser difícil encontrar otros muy similares, de forma que las recomendaciones realizadas tendrán mucha menos precisión. Este problema es conocido como *cold start* o arranque frío [4], y se puede paliar pidiendo a los usuarios unas valoraciones iniciales para así generar datos previos al comenzar el uso de este sistema, o usando un *baseline* como la recomendación de elementos populares mientras no se dispongan de suficientes datos previos.

En cuanto al *método colaborativo basado en ítems*, el IBCF [4], su comportamiento difiere del UBCF aunque el principio en el que se basan es muy parecido. En este caso, no se calcula la similitud entre usuarios, sino que lo que se relacionan entre sí son los ítems, es decir, para todos los casos en los que un determinado ítem tiene una buena valoración se analizan los otros ítems a los que ese usuario ha valorado, y así poniendo en común los datos de muchos usuarios se puede ver la relación entre unos ítems y otros, teniendo una mayor similitud los que tienen puntuaciones más similares de media y más veces aparecen. De esta forma, lo que se crea es una matriz de similitud entre todos los ítems disponibles del catálogo, de forma que cuando se le deben hacer recomendaciones a un nuevo usuario, se observan sus valoraciones previas y se ordenan los ítems restantes en función de las similitudes con los mejor valorados por éste. La diferencia entre este método y los métodos basados en contenido mencionados anteriormente radica en que en el caso anterior las similitudes entre los ítems se calculaban a partir de cualidades y características propias de los objetos, por lo que se acababan recomendando objetos similares, pero en el caso del IBCF la matriz de similitud entre objetos se calcula a partir de las valoraciones que estos reciben de los usuarios (por eso es un método colaborativo) pero los ítems con una alta similitud no tienen por qué compartir características propias. Este método también ofrece unos buenos resultados, aunque suelen ser mejores los del UBCF, y también presenta el problema del *cold start* cuando se dispone de insuficientes datos previos, pero posee una importante ventaja frente al método colaborativo basado en usuarios, y es que en este caso los cálculos para realizar las recomendaciones a un nuevo usuario son mucho más rápidos y conllevan mucho menos tiempo de computación, por

lo que las recomendaciones se hacen más rápido. Lo que lleva bastante tiempo es calcular la matriz de correlación entre todos los objetos, pero una vez se tiene esta matriz se pueden realizar todas las recomendaciones que se desee de forma mucho más rápida, y bastaría con actualizar esta matriz cada cierto tiempo para ir teniendo en cuenta las nuevas valoraciones que se van haciendo y los nuevos ítems que vayan apareciendo.

### **Métodos híbridos**

Cabe comentar que no existe el sistema de recomendación perfecto, presentando todos ellos distintas ventajas y limitaciones, motivo por el cual se utilizan técnicas de hibridación, consistentes en combinar dos o más métodos de recomendación, con el objetivo de mejorar la eficiencia de las recomendaciones y la precisión [7]. Las técnicas de hibridación más conocidas y utilizadas son: la mezcla (*mixed*), que combina los resultados obtenidos de los distintos sistemas de recomendación implicados, el uso de ponderaciones (*weighted*), que es similar al anterior pero utiliza unos pesos para ponderar las recomendaciones en función del sistema del que provengan, las técnicas *switch*, que alternan diferentes métodos de recomendación en función de la situación y el momento en el que se realiza, y las técnicas en cascada (*cascade*), en las que los resultados de un primer sistema son refinados en los siguientes, lo cual aporta más eficiencia computacional que usar todas las técnicas con todos los datos en paralelo. Como se puede ver, cada una de estas técnicas requiere de un ajuste y un balance según el uso para optimizar los resultados, al igual que para los sistemas de recomendación usados individualmente.

### **Métodos contextuales**

Otro tipo de métodos de recomendación que están siendo estudiados y se están desarrollando en la actualidad son los llamados CARS (*Context Aware Recommendation Systems*) o sistemas de recomendación con información del contexto [5]. La idea de estos métodos es usar toda la información disponible para poder realizar las recomendaciones de forma más precisa, por lo que no usan solamente los datos acerca de los usuarios, los productos y las valoraciones pasadas, sino que recopilan toda la información posible acerca de otros aspectos que conforman el contexto en el que se hace esta recomendación. Por ejemplo, para un sistema similar a Netflix datos del contexto que podrían resultar relevantes serían: la hora del día (ya que es probable que un determinado usuario suela ver por ejemplo series de menor duración durante la comida o al mediodía y le resulten más interesantes películas de mayor duración por la noche), la época del año (de forma que las recomendaciones tengan una temática acorde a las distintas estaciones y las celebraciones que van ocurriendo), o incluso si se accede desde un lugar u otro (dado que si un usuario accede desde el ordenador es más posible que vea películas de acuerdo a sus gustos personales, como por ejemplo de temática bélica, o capítulos de las series que suele seguir, pero en cambio si el mismo usuario accede desde la televisión del salón es probable que desee ver algo con la familia o con los amigos, por lo que no le interesen las series que sigue individualmente y en cambio le serán útiles recomendaciones acerca de películas de temática más familiar o más general que las que suele ver por su cuenta).

Otros sistemas que saldrían muy beneficiados con estos métodos de recomendación serían los relacionados con puntos de interés turístico, de forma que en función del motivo del viaje, de qué personas viajan juntas, del medio de transporte con el que se mueven, o del tiempo que se dispone, podría recomendar unos lugares u otros para adaptarse más a los gustos y necesidades de los usuarios. Toda esta información recopilada puede provenir de diferentes lugares; así, en el caso del primer ejemplo para un sistema de contenido de entretenimiento a la carta, sería muy fácil obtener la hora del día, la fecha y el aparato desde el que se accede a través del sistema y comunicándose con otras aplicaciones existentes; en cambio, para el segundo caso de recomendación de puntos de interés turístico, la información sobre el grupo, el tiempo disponible o el medio de transporte no se podría obtener fácilmente, de forma que una solución consistiría en implementar un cuestionario en el propio sistema para que el usuario lo complete con la información pertinente, aunque esto suele percibirse como una incomodidad; otra opción sería extraer toda esta información conectándose a unos sensores. El caso es que cuanto más información esté al alcance del sistema, más ricas y precisas serán las recomendaciones, pero tampoco se debe intentar abarcar más información de la necesaria, ya que puede suponer complicaciones y molestias a cambio de poco o ningún beneficio, de forma que en estos sistemas es importante la experimentación para ver cómo afecta a los resultados cada uno de los parámetros utilizados y saber a cuáles darles más peso y de cuáles se podría prescindir según el caso.

Mención aparte requieren los *mobile CARS* [10], que surgen debido al uso masivo de los teléfonos móviles. La idea de estos sistemas es incrementar el nivel de personalización de las recomendaciones al usar toda la información disponible en el móvil. La meta final sería sustituir el tradicional patrón de uso basado en petición-respuesta para evolucionar hacia unas recomendaciones más proactivas a la vista de toda esta información que recopilan los teléfonos, ya que los llevamos encima en todo momento. Estos sistemas deben identificar el momento más adecuado para generar una recomendación sin una petición explícita del usuario [32], siendo para ello necesario analizar su contexto.

Para concluir, un buen sistema de recomendación es aquel que hace que el usuario no sienta desbordado por la cantidad de información disponible y le ofrezca los resultados más adecuados a su perfil y sus necesidades en ese momento. Como ya se ha dicho antes la cantidad de información al alcance de cualquiera crece de forma exponencial, y las nuevas tecnologías están cada vez más presentes en casi todos los aspectos de la vida cotidiana, motivos por los cuales este tipo de técnicas son de gran importancia y van a desarrollarse cada vez más. A la hora de implementar un sistema de recomendación, es muy importante el ambiente en el que se va a trabajar y tener claros los objetivos que se quieren cumplir, para así poder seleccionar la tipología más adecuada para cada caso concreto (en la Tabla 2.1 se resumen las diferencias entre unos métodos y otros), y además es conveniente llevar a cabo algunas pruebas ya que en gran parte de los casos con un poco de ajuste de parámetros y funcionamiento se optimizan bastante los resultados. Las ventajas de usar sistemas de recomendación son el ahorro de tiempo y esfuerzo para el usuario, el incremento de lo que se conoce como “ventas cruzadas” al recomendar

siempre productos adicionales y el incremento de la fidelidad de los usuarios a un sistema si las recomendaciones son adecuadas, pero también existen algunos inconvenientes derivados de su uso como que pueden generar desconfianza de las recomendaciones entre ciertas personas que no comprendan su funcionamiento, que al introducir usuarios o productos nuevos o si se disponen de muy poco usuarios pueden hacerse recomendaciones poco relevantes, que muchos usuarios no están motivados a introducir sus datos personales y de preferencias por motivos de privacidad y que el mantenimiento de estos sistemas y sus bases de datos es bastante costoso, ya que se requieren equipos muy potentes para que puedan trabajar en tiempo real.

Método	Requisitos	Ventajas	Desventajas
<b>No personalizados</b>	Sin datos usuario objetivo	Muy sencillos	Poco fiables
<b>Basados en contenido</b>	Características de productos	Simple	Limitado por características
<b>Demográficos</b>	Grupos demográficos o <i>clusters</i> de usuarios, info personal del usuario objetivo	No requiere información histórica	No validos para todo el mundo, necesidad de datos demográficos e info personal
<b>UBCF</b>	Valoraciones de otros usuarios y valoraciones conocidas de usuario objetivo	Muy buenos resultados, independiente de los tipos de ítems,	Coste computacional y recomendaciones lentas, <i>cold start</i>
<b>IBCF</b>	Valoraciones de otros usuarios y valoraciones conocidas de usuario objetivo	Buenos resultados, recomendaciones rápidas, independiente de los tipos de ítems	<i>Cold start</i> , coste de creación del modelo
<b>CARS</b>	Información del contexto, histórico de valoraciones	Recomendaciones más precisas y adecuadas según necesidades en ese momento	Dificultad de recopilar información de contexto y aplicarla de forma adecuada

Tabla 2.1: Comparación de requisitos, ventajas y desventajas de distintos sistemas de recomendación

## Capítulo 3

# Comparación experimental de distintos métodos de recomendación

En este Capítulo se va a hablar acerca del conjunto de datos utilizado, algunos de los experimentos más relevantes que se han realizado y cómo se han hecho, y los resultados obtenidos de ellos, así como unas primeras conclusiones extraídas de dichos resultados. Así en la Sección 3.1 se hablará del entorno de trabajo y el conjunto de datos seleccionado. En la Sección 3.2 se enumerarán los métodos de recomendación que se van a utilizar y se hablará del tratamiento del conjunto de datos para poder hacer las recomendaciones. En la Sección 3.3 se explicarán los experimentos hechos y se analizarán los resultados obtenidos. Por último, en la Sección 3.4 se hará otro experimento distinto para ver cómo influye en el resultado el número de valoraciones conocidas por usuario.

### 3.1. Entorno y conjunto de datos

Una vez que se ha comprendido el funcionamiento general de los sistemas de recomendación y se han descrito las distintas tipologías existentes, el siguiente paso es trabajar con algunas técnicas de recomendación para apreciar sus diferencias a nivel práctico, ver cómo afectan los distintos parámetros a los resultados obtenidos, y comparar los distintos resultados para ver en qué casos son más convenientes unos u otros y qué nivel de fiabilidad proporcionan estos sistemas. Para ello, como se comentó en el Capítulo 1, se ha trabajado con el lenguaje de programación R, y se ha utilizado la librería `recommenderlab`, que es una muy buena opción para comenzar a trabajar e introducirse en este campo, ya que posee diversos algoritmos para sistemas de recomendación ya implementados, funciones y estructuras de datos muy útiles para implementar otros, e incluso algunos conjuntos de datos de ejemplo con los que poder comenzar a trabajar y probar las distintas opciones (ver Anexo B para más detalles de las funciones y estructuras utilizadas).

Tras haberse familiarizado con las distintas opciones de la librería `recommenderlab`, el siguiente paso es la búsqueda de un conjunto de datos con el que poder trabajar. Dado que el objetivo es utilizar los datos obtenidos en el contexto de la movilidad urbana, tiene sentido que los objetos a recomendar sean puntos de interés turístico, hoteles, restaurantes o zonas de ocio, dado que serán la mayoría de los casos en los que la información se podrá usar. Lo que interesará, en definitiva, serán todos aquellos datos que ayuden a conocer el movimiento de la gente alrededor de la ciudad, y en especial el de los visitantes o turistas que vengan por un determinado evento, para saber qué zonas se verán más saturadas, qué medios de transporte extra habrá que habilitar, o qué tipo de servicios y dónde serán

necesarios. En este caso, tras explorar algunas de las distintas opciones disponibles para su descarga en Internet (ver Anexo C), se ha decidido que probablemente lo más representativo sea utilizar un conjunto de datos con valoraciones de hoteles, ya que saber dónde estarán ubicados los visitantes será una de las informaciones de mayor interés para la planificación urbana. Por ello, el conjunto de datos seleccionado [8] es uno proveniente de la web TripAdvisor, el cual contiene 4606 valoraciones emitidas por 1202 usuarios acerca de 1890 hoteles o alojamientos.

Al inspeccionar los datos obtenidos, se puede observar que hay más información aparte del usuario, el alojamiento y la valoración (ver Figura 3.1). Lo primero que se ve es que se tratan de valoraciones para distintos hoteles y alojamientos alrededor de todo Estados Unidos, y que dichas valoraciones son hechas también por habitantes de Estados Unidos. Esto no supone un problema, ya que para aplicarlas a la movilidad se extraería la información para los distintos hoteles y alojamientos de una determinada ciudad en concreto, pero los sistemas de recomendación se comportarán de la misma forma en este caso, así que las conclusiones a las que se llegue serán válidas igualmente. Además, este conjunto de datos contiene información sobre el estado en el que habita cada usuario, o por lo menos en el que se encontraba al crear el perfil de usuario, y la zona horaria a la que pertenece dicho estado, al igual que la ciudad en la que se encuentra ubicado cada hotel, el estado el que pertenece y su zona horaria, así como información sobre el tipo de viaje, que en este caso puede ser: familiar, en pareja, de negocios, en solitario o con amigos. Toda esta información podría ser útil para desarrollar sistemas de recomendación demográficos, ya que se podría conocer a qué zonas o estados son más propensos a viajar los turistas según su origen, o incluso para sistemas de recomendación contextuales, que podrían beneficiarse mucho de la información acerca del tipo de viaje.

ID_USER	USER_STATE	USER_TIMEZ	ID_HOTEL	HOTEL_CITY	HOTEL_STAT	HOTEL_TIME	Trip Type	Rating
45	GA	Eastern	105170	Memphis	TN	Central	3	5
45	GA	Eastern	223229	SanAntonio	TX	Central	3	5
45	GA	Eastern	258688	Albuquerque	NM	Mountain	3	5
45	GA	Eastern	98827	ELPaso	TX	Central	3	5
45	GA	Eastern	99518	SanAntonio	TX	Central	3	3
64	TX	Central	224427	Cleveland	OH	Eastern	1	4
64	TX	Central	1751886	Austin	TX	Central	3	4
64	TX	Central	99120	Houston	TX	Central	3	4
100	NY	Eastern	120111	Jacksonville	FL	Eastern	2	5
100	NY	Eastern	91428	Indianapolis	IN	Eastern	2	4
100	NY	Eastern	109399	Dallas	TX	Central	3	3
100	NY	Eastern	98764	Albuquerque	NM	Mountain	4	3
149	OR	Pacific	240681	Phoenix	AZ	Mountain	1	3
149	OR	Pacific	73927	Phoenix	AZ	Mountain	1	3
149	OR	Pacific	100584	Seattle	WA	Pacific	2	1
159	TX	Central	98815	ELPaso	TX	Central	1	3
159	TX	Central	99055	Houston	TX	Central	2	4
159	TX	Central	92773	Louisville	KY	Eastern	3	4
198	VA	Eastern	74590	Phoenix	AZ	Mountain	3	4
198	VA	Eastern	75180	Tucson	AZ	Mountain	3	4

Figura 3.1: Fragmento del conjunto de datos de TripAdvisor



## 3.2. Métodos de recomendación utilizados y tratamiento de los datos

Para obtener las predicciones sobre a qué zonas acudirán los turistas y viajeros, no se van a utilizar sistemas contextuales, dado que para hacer numerosas predicciones a partir de los perfiles de usuario disponibles no se tendrá información del contexto para cada uno; además, hay estudios que señalan la escasez de conjuntos de datos para evaluar sistemas de recomendación dependientes del contexto [11]. En este caso no se va a usar la recomendación basada en contenido, aunque si se dispusiera de datos con las características de los hoteles sería muy buena opción. Tampoco se van a usar las recomendaciones demográficas, ya que para ello es necesario conocer los distintos grupos demográficos en los que dividir a los usuarios, para lo que habría que hacer estudios y pruebas si se quieren buenos resultados; además haría falta disponer de una base de datos con información acerca de los usuarios, lo cual no es fácil de obtener de manera gratuita y a través de Internet debido a las normativas de protección de datos. No obstante, este método también sería válido, y de hecho sería uno de los más adecuados si se conoce el tipo de personas que asistirán al evento para el que se hacen los estudios; ya que si la gran mayoría de asistentes pertenece al mismo grupo demográfico y social, y se disponen de buenos datos acerca de él, se podrían conocer con bastante precisión las zonas con más afluencia de viajeros. De esta forma, los sistemas de recomendación con los que se va a experimentar son, además de los sistemas sencillos que no requieren datos de entrada (sistemas no personalizados), que se usarán como línea base para comparar resultados, los métodos colaborativos. Para aplicar estos métodos no es necesaria información extra aparte de las propias valoraciones, son muy versátiles y se pueden utilizar a gran escala para obtener predicciones generales acerca de grandes grupos de personas.

De esta forma, en los experimentos se van a utilizar 4 métodos de recomendación distintos: recomendación aleatoria, recomendación de populares, UBCF e IBCF. Para todos estos métodos, lo único necesario son datos acerca de un usuario, el objeto que ha puntuado y la valoración en sí; no se requiere de información extra acerca del usuario, ni características del producto ni datos contextuales, por lo que el primer paso tras cargar toda la base de datos completa es crear una tabla que contenga solo estas 3 variables: usuario, producto y valoración. A partir de esta tabla se creará una matriz de valoraciones, en la cual cada fila representa a un usuario y cada columna representa un objeto, de forma que en la intersección de ambas se encuentra la valoración de ese usuario para ese objeto en concreto. Así, la matriz obtenida es una matriz muy dispersa, dado que en la mayoría de las situaciones cada uno de los usuarios ha valorado unos cuantos objetos pero está lejos de haber valorado todos los disponibles, y de la misma forma cada objeto ha sido valorado unas cuantas veces pero no es muy común un objeto valorado por casi todos los usuarios, aunque esto sí que se podría dar por ejemplo en el caso de que fuera un objeto muy popular, como una película o una canción de gran éxito que todo el mundo conoce. En la librería `recommenderlab` existe una estructura llamada *RatingMatrix* que está pensada para crear este tipo de matrices de valoraciones dispersas (ver Figura 3.2), de

forma que los mismos datos, introducidos en una *RatingMatrix* ocupan mucha menos memoria que si se introducen en una matriz o tabla convencional.

```
> getRatingMatrix(RatingMatrixTripAdvisor[1:20,1:30])
20 x 30 sparse Matrix of class "dgCMatrix"
  [[ suppressing 30 column names '72339', '72342', '72359' ... ]]

45 . . . . .
64 . . . . .
100 . . . . .
149 . . . . .
159 . . . . .
198 . . . . .
215 . . . . .
221 . . . . .
247 . . . . .
349 . . . . .
451 . . . . .
730 . . . . .
750 . . . . . 5 . . . . .
753 . . . . . . . . . . 3
763 . . . . .
843 . . . . .
897 . . . . .
958 . . . . .
960 . . . . .
968 . . . . .
```

Figura 3.2: Fragmento de la *RatingMatrix* creada

En este punto surge un problema, y es que a pesar de comprobar que todas las valoraciones en el conjunto de datos se encuentran entre 1 y 5, al crear la *RatingMatrix* se obtienen algunas valoraciones superiores a 5, las cuales no existen realmente y no tienen sentido. Tras investigar los motivos por los que pasaba esto y probar distintas opciones para solucionarlo, se descubrió que estas valoraciones superiores a 5 surgían debido a que había valoraciones repetidas en el conjunto de datos inicial, es decir, un mismo usuario había valorado al mismo objeto en varias ocasiones con distintas puntuaciones. Una vez encontrado el origen del problema, la solución fue sencilla (aunque existen otras soluciones posibles más eficientes): se programó una estructura de bucle que iba leyendo todos los registros del conjunto de datos y las comparaba con los leídos anteriormente, de forma que si una determinada valoración no estaba repetida se introducía a la tabla de valoraciones, pero si se encontraba otra con la misma pareja de usuario-ítem, esa valoración era omitida y se pasaba a la siguiente. Esto se puede interpretar físicamente como que un mismo usuario ha visitado el mismo alojamiento en varias ocasiones y lo ha puntuado cada una de las veces, de manera que la última valoración dada es la que mejor representa la opinión de ese usuario sobre el alojamiento, ya que la misma ha podido variar con el tiempo.

El siguiente paso a seguir es normalizar las valoraciones con el objetivo de evitar los errores debidos al sesgo personal de cada usuario. Esto es necesario ya que algunos usuarios dan valoraciones más altas que otros de forma sistemática, de manera que algunos usuarios darán un 5 a los alojamientos que sean de su agrado y un 3 a los que no lo sean tanto, y en cambio otros darán un 3 cuando estén satisfechos y un 1 en caso de que no, por lo que estas valoraciones son muy distintas para el sistema pero reflejan una misma opinión del usuario. Para evitar los errores derivados de esto la solución es normalizar todas las valoraciones. En este caso lo que se hace es calcular la media de las valoraciones para cada usuario, y lo que se tiene en cuenta es la diferencia entre una determinada valoración y la media de las valoraciones para ese usuario, haciendo que si la valoración coincide con la media de ese usuario se considerará un valor de 0, si está por encima se representará con la diferencia entre esa valoración y la media en positivo, y si está por debajo será la diferencia pero en negativo. Además, esta diferencia se divide entre la desviación típica. Así, se obtiene un histograma de valoraciones o votos bastante más homogéneo, y se ha conseguido evitar, o al menos minimizar, el error debido al sesgo personal (ver Figura 3.3 y Figura 3.4).

Una vez se dispone de las valoraciones normalizadas es el momento de comenzar a utilizar el sistema de recomendación. Para ello, es necesario separar todo el conjunto de datos, de forma que algunos de los datos sean utilizados para entrenar al sistema, y el resto se utilicen para *testear* y verificar los resultados. Esto se puede llevar a cabo de manera sencilla gracias a la función *evaluationScheme* disponible en la librería *recommenderlab* de R, la cual permite precisar qué porcentaje de datos se utilizará para entrenamiento y para evaluación, el número de veces que se llevará a cabo esta

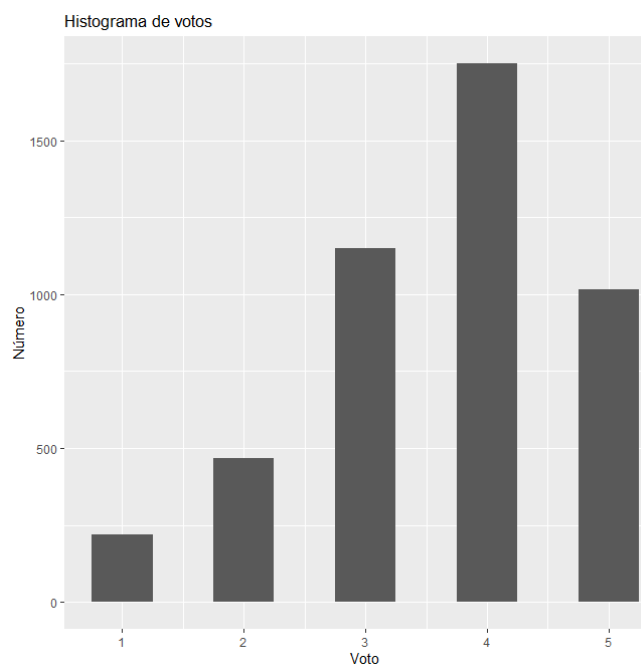


Figura 3.3: Histograma de votos

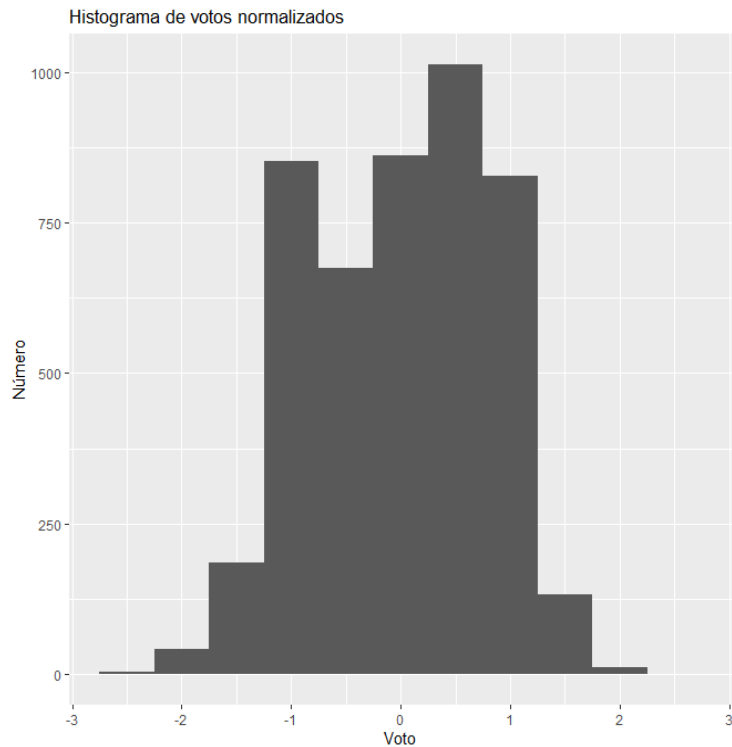


Figura 3.4: Histograma de votos tras la normalización

evaluación, el valor a partir del cual una puntuación se considera buena (por ejemplo, en una escala de 0 a 5, podría asumirse que si la valoración es igual o superior a 3 al usuario le ha gustado ese ítem) y el número de valoraciones que se tomarán como conocidas para cada nuevo usuario del conjunto de test (con el parámetro *given*), dejando el resto de las que se disponga para comparar con las predicciones realizadas para ese usuario. Además, dicha función es capaz de utilizar 3 métodos distintos para realizar esta partición de los datos [9]: el método “split”, que recoge datos de forma aleatoria para los datos de entrenamiento hasta llegar al porcentaje indicado, y deja los restantes para el conjunto de datos de *test*, el método “cross-validation”, que divide los datos en  $k$  partes iguales, siendo  $k$  el número de veces que se va a repetir la evaluación, de forma que en cada evaluación se usan  $k - 1$  partes como datos de entrenamiento, y al final de todas las repeticiones cada una de las partes ha sido usada como datos de *test* una vez; y el método “bootstrap”, que crea el conjunto de datos de entrenamiento cogiendo datos con reemplazamiento hasta llegar al porcentaje indicado, y el resto de los datos no seleccionados son usados para *test*. Tras haber dividido el conjunto de datos seleccionando el método y los parámetros deseados, es el momento de entrenar el sistema de recomendación (ver anexo A).

En *recommenderlab* existe una función llamada *evaluate*, para la cual hay que proporcionar los conjuntos de datos de entrenamiento y test creados con *evaluationScheme*, como se ha comentado anteriormente, el método de recomendación que se desea utilizar (o una lista de varios si se desea compararlos), la normalización de los datos que se precise, y el número de recomendaciones a realizar para cada usuario. Esta función devuelve los resultados tras haber hecho las predicciones para los datos de

*test* con cada uno de los métodos indicados y haberlas comparado con los datos reales. Estos resultados contienen los valores de positivos reales (*true positives* o TP), falsos positivos (FP), negativos reales (TN), falsos negativos (FN), un valor de precisión que se calcula como  $TP/(TP+FP)$ , un valor de *recall* o sensibilidad que se calcula como  $TP/(TP+FN)$ , un ratio de positivos reales (TPR) y un ratio de falsos positivos (FPR) para cada uno de los métodos de recomendación utilizados y para cada caso en función del número de recomendaciones realizado. También existe la opción de hacer los experimentos paso a paso, dado que existe una función para entrenar el sistema de recomendación con el método deseado y a partir de unos datos de entrenamiento, otra para obtener las predicciones utilizando el sistema de recomendación entrenado y los datos que se consideran conocidos para los nuevos usuarios, y por último una para calcular el error de las predicciones comparando éstas con las valoraciones reales de los nuevos usuarios, que se suponían desconocidas. Haciéndolo de esta forma, lo que se obtiene es la media de los errores absolutos (MAE), así como la media de los errores al cuadrado (MSE) y la raíz de la media de los errores al cuadrado (RMSE) (ver Anexo D para más información sobre RMSE), de las predicciones realizadas tras compararlas con los datos reales.

### 3.3. Obtención de predicciones y análisis

Tras haber realizado numerosas pruebas y experimentos se ha comprobado que si los datos de *test* son demasiado pequeños, o si no se repite la evaluación suficientes veces, los resultados obtenidos son muy variables, ya que si se usa el método “*Split*” con unos mismos conjuntos de datos de entrenamiento y test y sin cambiar los parámetros al realizar varias veces las predicciones y evaluarlas los resultados cambian sustancialmente, de modo que no son fiables para llegar a conclusiones. Con el objetivo de conseguir unos resultados consistentes y representativos, se ha utilizado el método de “*cross-validation*” para separar los conjuntos de datos de entrenamiento y test, ya que la validación cruzada de 10 grupos ( $k=10$ ) es el método de evaluación de mayor coste computacional pero también el que mejores resultados aporta, y se va a repetir la evaluación 10 veces. Además, se ha comprobado que el mínimo número de valoraciones por usuario son 3, de forma que se ha ajustado el sistema de recomendación para que en los datos de test se considere que son conocidas 2 valoraciones de cada usuario fijando el parámetro *given* en un valor de 2, haciendo así que en los casos en los que menos valoraciones se hayan dado quede por lo menos 1 de ellas para poder compararla con la predicción y poder obtener el error de la misma. Tras hacer las predicciones con los distintos métodos de recomendación y evaluarlas, se puede obtener una curva ROC y un gráfico que represente precisión y *recall* (ver Figura 3.5 y Figura 3.6). En la curva ROC el eje vertical representa el TPR y el horizontal el FPR, y en ambas gráficas los valores que están junto a las curvas representan el número de ítems recomendados a cada usuario (ver Anexo D para una explicación más detallada de una curva ROC).

Al analizar estas gráficas se observa que, como cabía esperar, al aumentar el número de recomendaciones que se realizan crece el ratio de recomendaciones acertadas, pero

también el de recomendaciones fallidas, aunque mientras el TPR sea significativamente mayor que el FPR se considerará que el método hace recomendaciones válidas en general. En este caso, se comprueba que el mejor método es el IBCF, ya que es el que tiene mayor TPR respecto al FPR, y observando la gráfica parece que cada vez proporcionaría más recomendaciones correctas sin aumentar las incorrectas. De la misma forma, en la otra gráfica se observa que al aumentar la sensibilidad o *recall* disminuirá la precisión, pero igualmente el método que mejores resultados presenta es el IBCF. Cabe comentar que los valores obtenidos de TPR, FPR, precisión y *recall* son bastante pequeños [1], lo cual es debido a que la *RatingMatrix* utilizada es muy dispersa, y si se dispusieran de más datos se obtendrían mejores resultados, como se verá cuando se comparen los resultados en función del número de valoraciones conocidas por usuario. Además, estos resultados dependen del umbral de recomendación escogido para considerar una recomendación adecuada (en este caso una puntuación de 4). También se han realizado las predicciones para cada método paso a paso para poder obtener y comparar el error de las mismas en cada caso con los valores de MAE, MSE y RMSE (ver Figura 3.7).

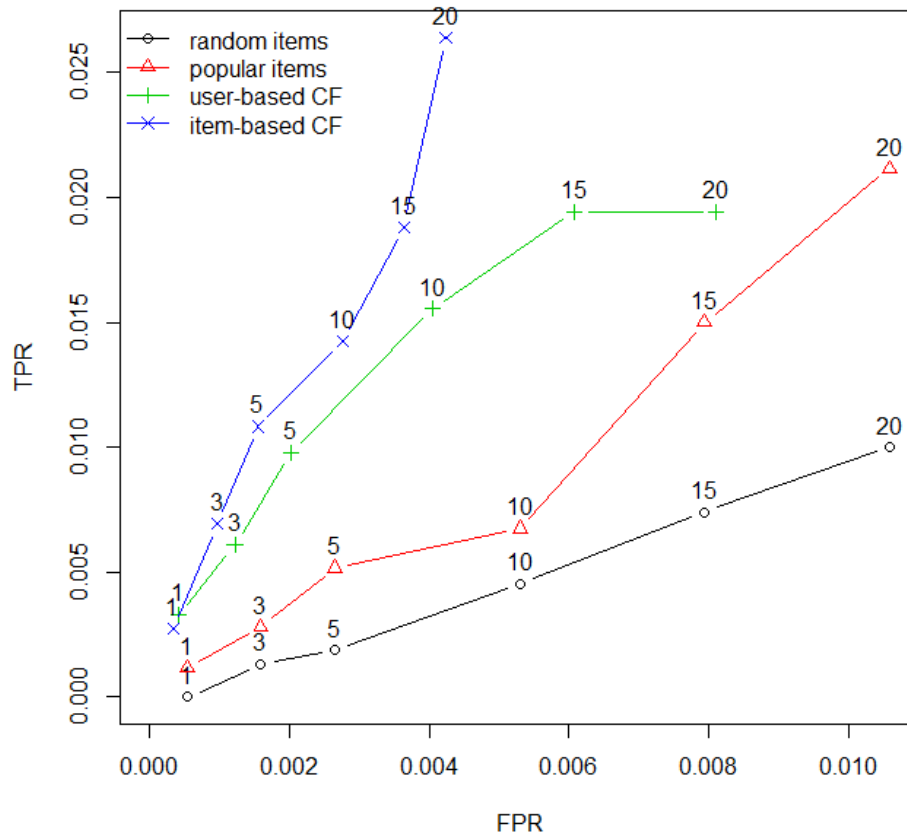


Figura 3.5: Curva ROC

Como se puede ver, los resultados del error en las predicciones siguen la misma tónica que lo que se había apreciado en las gráficas anteriores, y es que el método aleatorio es el que mayor error presenta, seguido por el popular y el UBCF que tienen errores bastante similares, y por último la mejor opción es el IBCF.

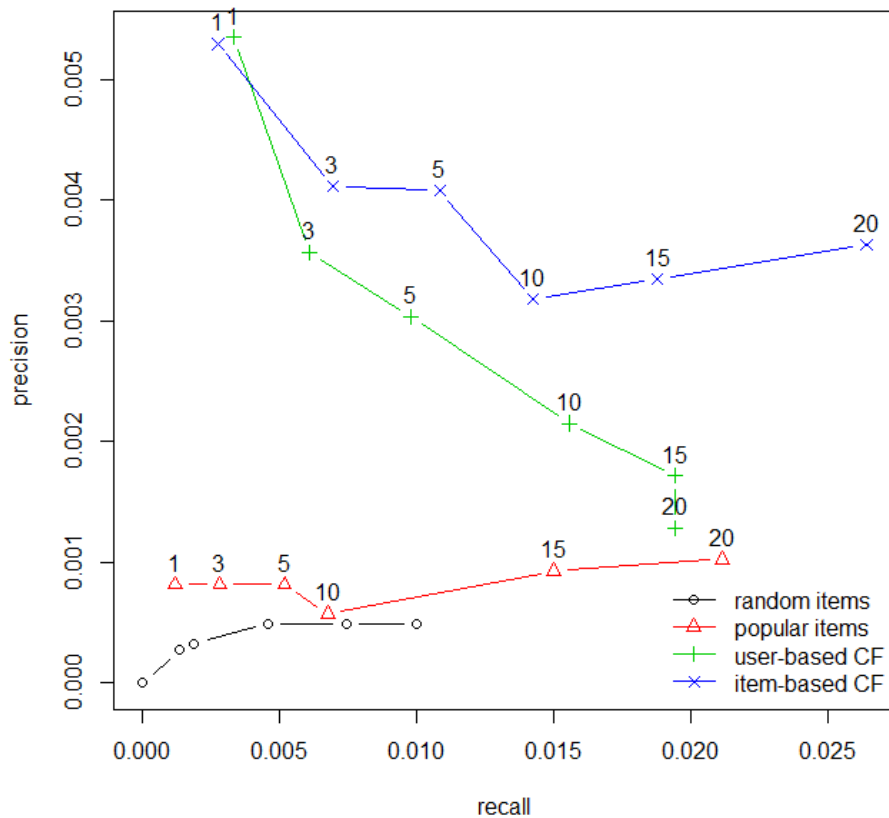


Figura 3.6: Precisión frente a *recall*

	RMSE	MSE	MAE
UBCF	1.1710669	1.371398	0.8986490
IBCF	0.7071068	0.500000	0.6666667
POPULAR	1.2113308	1.467322	0.9337779
RANDOM	1.5311462	2.344409	1.2321279

Figura 3.7: Medidas de error de las predicciones

Esto a priori no encaja con lo que se preveía teóricamente, y es que en principio el método UBCF tenía un mayor coste computacional y era por lo tanto más lento, pero sus resultados deberían ser mejores [1], y en este caso el método colaborativo basado en los objetos, IBCF, presenta unos resultados significativamente mejores; sin embargo como ya se había comentado, dependiendo del caso concreto, del conjunto de datos con el que se trabaje y de los objetivos que se quieran conseguir, será más apropiado un método u otro, y no hay uno que sea estrictamente mejor en cualquier ocasión. Así, muy probablemente este hecho sea debido a que en el conjunto de datos con el que se trabaja los usuarios han dado pocas valoraciones cada uno, de forma que es más complicado calcular la similitud entre ellos, y estos valores sean menos fiables de lo que podrían resultar en otros datos. Y es que si se analizan las valoraciones que ha dado cada usuario (ver Figura 3.8) se aprecia que la gran mayoría solamente tiene 3 alojamientos valorados (más de 700 usuarios de los aproximadamente 1200 totales), y el resto tienen pocos más,

de forma que hay muy pocos usuarios que tengan un alto número de valoraciones. Podría ser interesante, si en un trabajo futuro se utiliza un conjunto de datos con más valoraciones por cada usuario, volver a realizar estas pruebas y comparar los resultados, para ver si el método UBCF presenta resultados mejores.

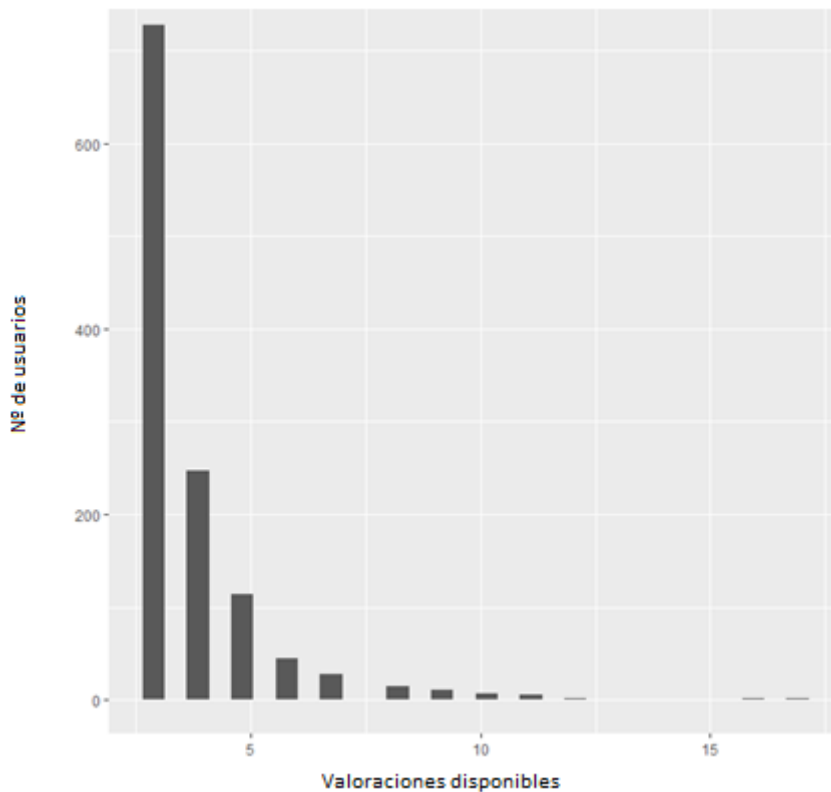


Figura 3.8: Número de usuarios en función de sus valoraciones

### 3.4. Influencia del número de valoraciones conocidas

Ya han sido comparados los resultados obtenidos con distintos métodos de recomendación, pero aun hay más pruebas que pueden aportar información interesante, y quizá la más destacable de ellas sea ver cómo varía el error en las predicciones en función de las valoraciones conocidas que se tengan del usuario objetivo, dado que cuantas más sean con mayor facilidad se podrían encontrar recomendaciones adecuadas ya que se conocerían mejor las preferencias del usuario. Para hacer esto, se han creado los conjuntos de datos de entrenamiento y *test* de una forma similar a la anterior, pero en esta ocasión se ha ajustado el sistema para que considere que en los usuarios del conjunto de *test* se conocen todas sus valoraciones menos dos de ellas, es decir, que se calculará el error de la predicción de 2 valoraciones reales, de forma que el resto se considerarán conocidas para poder realizar las recomendaciones al usuario, si un determinado usuario posee 6 valoraciones, se considerarán 4 conocidas y a partir de ellas se predecirán las otras 2 para compararlas con las valoraciones reales, pero si un usuario posee 3 valoraciones, como es la mayoría de casos, sólo se conocerá 1 de las valoraciones para hacer la predicción y las otras dos se usarán para comparar con los datos reales.



Así pues, tras haber creado los conjuntos de datos de entrenamiento y *test*, se comprueban las valoraciones conocidas y desconocidas para cada usuario del conjunto de *test*, y se seleccionan usuarios concretos con distinto número de valoraciones conocidas. Tras ello, se entrena el sistema de recomendación con los datos de entrenamiento, después se realizan las predicciones para los usuarios concretos seleccionados del conjunto de *test*, y se calcula el error de dichas predicciones en cada caso comparándolo con los valores reales. Para esta ocasión, se han seleccionado 4 usuarios concretos del conjunto de *test*, el primero con solamente 1 valoración conocida, otro con 3 valoraciones conocidas, otro de ellos con 5, y el último con 6; realizando así las predicciones para cada usuario de forma individual y comparando con las 2 valoraciones que se consideran desconocidas de cada uno para conocer el error (ver Figura 3.8, Figura 3.9, Figura 3.10 y Figura 3.11).

```
> rowCounts(getData(scheme,"known")[5,])
2019
  1
> rowCounts(getData(scheme,"unknown")[5,])
2019
  2
> calcPredictionAccuracy(pred,getData(scheme,"unknown")[5,])
      RMSE      MSE      MAE
1.414214 2.000000 1.000000
```

Figura 3.9: Error medio de las predicciones para un usuario con 1 valoración conocida

```
> rowCounts(getData(scheme,"known")[1,])
215
  3
> rowCounts(getData(scheme,"unknown")[1,])
215
  2
> calcPredictionAccuracy(pred,getData(scheme,"unknown")[1,])
      RMSE      MSE      MAE
1.269296 1.611111 1.166667
```

Figura 3.10: Error medio de las predicciones para un usuario con 3 valoraciones conocidas

```
> rowCounts(getData(scheme,"known")[13,])
4784
  5
> rowCounts(getData(scheme,"unknown")[13,])
4784
  2
> calcPredictionAccuracy(pred,getData(scheme,"unknown")[13,])
      RMSE      MSE      MAE
0.6025305 0.3630431 0.6025253
```

Figura 3.11: Error medio de las predicciones para un usuario con 5 valoraciones conocidas

```

> rowCounts(getData(scheme,"known")[40,])
23146
6
> rowCounts(getData(scheme,"unknown")[40,])
23146
2
> calcPredictionAccuracy(pred,getData(scheme,"unknown")[40,])
      RMSE      MSE      MAE
0.16666667 0.02777778 0.16666667

```

Figura 3.12: Error medio de las predicciones para un usuario con 6 valoraciones conocidas

Observando el error de las predicciones para cada usuario concreto, queda muy claro que cuantas más valoraciones se conozcan de un determinado usuario mucho más acertadas serán las recomendaciones para él, que es lo que se esperaba en un principio, pero llama la atención lo mucho que se reduce el error al aumentar sólo un poco el número de valoraciones conocidas, ya que en este caso del usuario con 3 datos conocidos al usuario con 6 el error disminuye en casi un 90%. Por lo tanto, se comprueba que cuanto más se sepa de un usuario objetivo mayor calidad tendrán las recomendaciones para éste, de forma que si se utilizan unas bases de datos ricas y completas los resultados serán sustancialmente mejores que si se usan conjuntos con pocos datos.

Así pues, es importante seleccionar adecuadamente el método de recomendación a utilizar en función de la situación y el caso objetivo, ya que no se puede recurrir siempre al mismo sistema porque en teoría es el más acertado, sino que se debe comprobar cuáles se pueden usar en función de los datos que se tienen y comprobar el que mejor se comporta. También es bastante relevante ajustar los parámetros para obtener los resultados más apropiados según el caso, dado que como se ha visto, aumentar por ejemplo el número de recomendaciones hechas por usuario hace que aumente la sensibilidad y aporta más opciones correctas, pero puede bajar la precisión de las recomendaciones realizadas, por lo que hay que aplicar el sistema de recomendación con cautela. En este aspecto, se podrían llevar a cabo más pruebas para ver cómo afectan otros parámetros como el número de vecinos o el método usado para calcular la similitud (en este caso la similitud de coseno). Por último, ha quedado claro también que los datos utilizados influyen enormemente en la calidad de las predicciones, por lo que se deber ser cuidadoso al adquirir los datos con los que se trabajará, ya que unos datos completos pueden proporcionar unas predicciones de gran fiabilidad, pero unos datos pobres o inadecuados pueden no resultar de gran ayuda.

# Capítulo 4

## Aplicación al campo de la movilidad urbana

En este apartado, se va a explorar y analizar cómo podrían ser aplicadas estas tecnologías y estos conceptos al ámbito de la movilidad, viendo en qué casos pueden ser adecuados y en cuáles no, qué tipo de datos o información previa se debe disponer y de dónde y cómo se podría obtenerla, a qué decisiones pueden afectar estas predicciones, qué impacto tendría esto en el mercado de la información, etc. Así, en la Sección 4.1 se hablará de los posibles casos de aplicación de los sistemas de recomendación a la movilidad. En la Sección 4.2 se comentará la procedencia de los datos utilizados para generar las predicciones. Por último, en la Sección 4.3 se verá cómo utilizar las predicciones obtenidas y los costes y beneficios que podrán suponer.

### 4.1. Casos de aplicación

En primer lugar, hay que dejar claro que por movilidad urbana se entiende el transporte de pasajeros por la ciudad, de manera cualquier aplicación de los sistemas de recomendación que obtenga como resultado un tráfico más fluido se considerará que ayuda a esta movilidad urbana, aunque no sea una aplicación directa. Así, se van a considerar posibles aplicaciones de los sistemas de recomendación en el campo del transporte y el turismo que conduzcan a facilitar el tráfico urbano y en general el movimiento de las personas. No se va a considerar su uso tradicional haciendo recomendaciones a usuarios individuales, aunque esto les ayude a transitar por la ciudad de forma más directa hacia su objetivo. En lugar de eso, se van a explorar casos de uso desde el punto de vista de entidades públicas o empresas que tengan alguna implicación en el transporte, ya sea de forma más directa como podría ser la gestión del transporte público, o de forma más indirecta como el uso de una flota de furgonetas que transitan por la ciudad para repartir productos a los comercios. Así, se va a estudiar cómo podrían utilizarse las predicciones de los sistemas de recomendación en estas entidades o empresas para mejorar sus actividades relacionadas con el transporte, obteniendo por lo tanto una mejora en la movilidad [23].

Como se comentó en el Capítulo 1, el principal caso de aplicación a la movilidad de las predicciones obtenidas de sistemas de recomendación sería la planificación para eventos excepcionales, o cualquier ocasión en la que vaya a darse una afluencia inusitada de turistas o viajeros en una determinada ciudad. El uso de esta técnica para planificar la movilidad urbana de manera general no es necesario. Sería conveniente si existiese una ciudad ya construida pero completamente vacía y se deseara predecir cómo se moverían en un futuro sus habitantes, ya que se podrían encontrar las áreas de más interés comercial o de ocio, así como las áreas en las que la gente es más propensa a alojarse; y se podrían

habilitar los aparcamientos o transportes públicos en función de ello. Sin embargo este no es un caso real, dado que en la práctica en todas las ciudades ya habita gente y tienen sus estructuras de movilidad y sus opciones de transporte público ya definidas; y cuando estas ciudades van creciendo y se van expandiendo, es de forma muy gradual, de manera que no suele ser necesario reestructurar la movilidad, sino que ésta se va adaptando poco a poco y basta con añadir algún transporte público que lleve a las zonas más transitadas. En algún caso de expansión en el que se construya un nuevo barrio residencial periférico sí que sería posible utilizar un sistema de recomendación para ver a qué zonas de la ciudad antigua desearán moverse los habitantes del nuevo barrio, pero en la mayoría de casos no suele ser muy necesario, ya que esto se suele conocer de antemano y no son necesarias predicciones, y además el aumento de habitantes en el nuevo barrio no es repentino, sino que suele ir creciendo de forma gradual y se pueden adaptar las necesidades de movilidad conforme aparezca la demanda [22].

Por todo ello, estos sistemas se podrían usar en algún otro caso si se desea y se estima necesario, no obstante, su principal foco de utilización serán las ocasiones en las que una ciudad reciba, por el motivo que sea, una cantidad elevada de turistas o visitantes en un pequeño lapso de tiempo. Esto podría darse, como ya se ha dicho, debido a algún evento excepcional como un concierto, una competición deportiva o una entrega de premios, pero no tiene por qué tratarse de esos casos, ya que por ejemplo puede ocurrir en determinadas ciudades para un momento o una época del año concreta, y repetirse año tras año. Así, por ejemplo, el uso de sistemas de recomendación sería muy útil para aplicarlo a una ciudad que va a acoger la final del mundial de fútbol, o una exposición internacional, pero también sería adecuado aplicarlo en Madrid cuando llega la Navidad o la Nochevieja, ya que la cantidad de gente en el centro de la ciudad aumenta enormemente, o en alguna ciudad muy turística como París o Nueva York en caso de que se acerque una festividad de carácter nacional y se prevea que va a acudir mucha gente durante los días festivos. A pesar de que sea algo que ocurra de forma periódica, puede ser interesante hacer uso de esta información, ya que la zona con más afluencia de turistas para alojarse, las zonas de ocio a las que suelen acudir, o directamente los gustos de la gente pueden cambiar mucho de un año a otro [28].

Otra ocasión en la que extraer información de los sistemas de recomendación puede ser muy útil de cara a la movilidad, pero está bastante alejada del caso anterior, es para la distribución y logística de un nuevo producto o servicio cuando se lanza al mercado. Toda la información obtenida de estas técnicas se podría utilizar para complementar y contrastar los estudios de mercado; así, si se dispone de la base de datos de una web de venta de objetos, se podría aplicar esto con el nuevo producto para saber a qué usuarios les podría gustar o interesar, y lo mismo si se tratase de un servicio ofrecido en puntos concretos y se dispusiese de una base de datos con valoraciones de este tipo. En función de la información que se conociera de los usuarios, lo cual variará de unos sistemas a otros, se podría conocer el tipo de personas que estarán interesados en adquirir el producto, conociendo su edad, su género o sus aficiones, por ejemplo; o incluso si se dispone de información acerca del domicilio de los usuarios, se podría conocer con

bastante precisión en qué zonas de la ciudad habrá más ventas que en otras o en cuáles los productos serán más atractivos. Como es obvio, esto tiene un enorme interés a la hora de gestionar la logística de distribución del producto, haciendo que éste sea mucho más abundante en las zonas donde ha resultado más popular o donde suelen acudir el tipo de clientes que han resultado ser más propensos, pero no sólo se aplicaría al reparto para comercios, ya que si se tratase de un producto perecedero que se vende in situ o un servicio se podría prever la cantidad del producto que llevar o la cantidad de puestos para recibir dicho servicio que habilitar en una zona, haciendo de su reparto por la ciudad una tarea mucho más rápida y sencilla, y agilizando la movilidad de las mercancías. Incluso si se trata de un servicio relacionado con el transporte de personas por la ciudad, como por ejemplo los patines y bicicletas eléctricas, que tan populares se han vuelto en los últimos tiempos, con sistemas de recomendación se podría conocer en qué zonas es más apropiado colocarlos al inicio del día o dónde se van a acumular al acabar el mismo, facilitando también enormemente el acceso a este medio, y en definitiva haciendo más fluida la movilidad, dado que la logística en estos negocios es un punto clave.

En la Tabla 4.1 se muestra un resumen de los distintos casos de aplicación posibles de los sistemas de recomendación a la movilidad, resaltando qué se necesita en cada uno, qué predicciones se obtienen y cómo afectan éstas a la toma de decisiones.

Caso de aplicación	Datos requeridos	Predicciones obtenidas	Influencia en decisiones
<b>Turistas por festividad</b>	Valoraciones de alojamientos y zonas de ocio	Alojamientos y zonas de ocio de interés general	Gestión de vías de transporte, transporte público y aparcamientos
<b>Turistas por evento excepcional</b>	Valoraciones de alojamientos y zonas de ocio con info de usuarios para conocer perfil concreto	Alojamientos y zonas de ocio de interés para un determinado perfil de usuarios	Gestión de vías de transporte, transporte público y aparcamientos
<b>Distribución nuevo producto</b>	Valoraciones de productos con info sobre los productos y los usuarios que valoran	Productos de interés para distintos grupos de usuarios	Logística y rutas de distribución de productos
<b>Gestión servicio movilidad empresa privada (patines o bicicletas eléctricas)</b>	Valoraciones o afluencia de zonas de ocio y puntos de interés con info de hora del día y localización	Zonas de ocio de interés en función del momento del día o localización del usuario	Puntos de colocación y de recogida, distribución a lo largo del día

Tabla 4.1: Posibles aplicaciones de sistemas de recomendación enfocadas a movilidad

## 4.2. Origen y naturaleza de los datos

Como se ha visto, estos sistemas tienen bastante variedad de aplicaciones posibles en temas relacionados con la movilidad urbana, y llevarlas a la práctica contribuirá a hacer cualquier ciudad más moderna y cómoda, permitiendo que responda y se adapte a las necesidades de sus habitantes, por lo que es un posible camino por el que podría discurrir el diseño y funcionamiento de las ciudades del futuro, pero hay que saber qué es necesario para poder utilizar estas técnicas. Y es que, como se ha concluido tras la experimentación, los conjuntos de datos que se utilicen para poder efectuar las predicciones son de radical importancia para el resultado de las mismas. Según se dispongan de unos datos u otros, se podrá aplicar la técnica deseada para la situación adecuada o no, por lo que será muy importante el intercambio de información a nivel empresarial.

Si se disponen de datos con valoraciones sobre alojamientos, se podrán usar para ocasiones con picos en la cantidad de turistas, pero no es lo mismo el puente en el que va acudir mucha gente, donde se pueden usar los datos de usuarios y valoraciones de forma más general para conocer las tendencias de todos los visitantes, que un concierto de un grupo con público más joven, para lo cual datos de usuarios generales no sería útil, y haría falta tener información sobre los usuarios para poder restringir las recomendaciones al público joven, y así conocer de forma efectiva a qué zonas acudirán los asistentes en realidad. Este mismo ejemplo se puede expandir para cualquier ocasión, haciendo que a veces datos con valoraciones de usuarios y objetos simples sean suficiente, y otras se necesite una base de datos con perfiles de usuario más completos para conseguir información útil, o incluso hagan falta datos con información sobre el contexto para algunos casos [27].

Si la aplicación es la distribución y logística de un producto, lo más conveniente sería una base de datos con las características de los productos disponibles, para poder relacionarlos con el nuevo que va a entrar al mercado (y así poder planificar sus canales y zonas de distribución) usando recomendaciones basadas en contenido, o se podrían utilizar métodos colaborativos si se organizan sesiones de prueba en las que algunos usuarios puedan valorar el producto y otros disponibles, y así a partir de estas valoraciones, se puedan realizar recomendaciones al resto en las que aparezca el nuevo producto gracias a estas valoraciones previas que se tenían de él. Por lo tanto, los datos necesarios pueden ser bastante distintos según el caso, de forma que cuantos más datos estén al alcance y más ricos y completos sean estos datos, en más ocasiones podrán ser útiles, así que es conveniente recopilar todos los datos posibles e intentar mantenerlos actualizados para usarlos cuando sean necesarios.

La siguiente pregunta es inevitable: ¿de dónde provienen todos estos datos? Esta cuestión es de vital importancia, pero existen numerosas opciones para responderla. En el caso de que la entidad que va a usar los datos sea pública, es decir, relacionada con el gobierno

de la propia ciudad, tendrá acceso a gran cantidad de datos que podrían ser de utilidad, como datos históricos de distintos tipos, diversos estudios relacionados por otros entes públicos, datos obtenidos de empresas privadas a través de informes de cuentas que deban presentar por determinadas leyes etc. La cantidad de conjuntos de datos al alcance es inmensa, por lo que será necesaria una búsqueda en profundidad, y además existe la posibilidad de que los datos estén desactualizados o les falte información importante, así que como última opción, siempre se podría recurrir a las fuentes de datos a las que se vería obligada a recurrir una empresa privada.

Y es que, si no se dispone de datos públicos, existen otras opciones, aunque en su mayoría son más caras. Se podría probar a pedir los datos a algún ente público o entes privados pero que tengan sus datos a disposición del público, como podría ser algún tipo de asociación de consumidores o entidad dedicada a la realización de diversas encuestas, pero si no se encuentra esto el siguiente paso sería recolectar los datos uno mismo o comprarlos. Para reunir los datos, lo más común sería contratar a una empresa encargada de hacer un estudio o encuestas a los clientes o la población, aunque si la empresa es muy grande quizá pudiera hacerlos por su cuenta. La última opción, que por otro lado es probable que sea la más utilizada, es la compra de información, de forma que permite adquirir la base de datos deseada a cambio de un desembolso. Cada vez es más común la compra y venta de información, y existen numerosas empresas que se dedican en exclusiva a recopilar información para posteriormente venderla, o incluso se podría adquirir un conjunto de datos de una empresa que utilice sistemas de recomendación en su actividad y por lo tanto disponga de él, como webs de reserva de viajes o de venta de productos por internet [29].

En relación a la compra e intercambio de datos, es conveniente recurrir a la Ley Orgánica de Protección de Datos de Carácter Personal [20] para conocer el marco legal en este ámbito. Esta ley tiene por objeto garantizar y proteger las libertades públicas y los derechos fundamentales de las personas físicas, especialmente el de su honor e intimidad personal y familiar. Así, según el artículo 3 de esta ley: “se entenderá por datos de carácter personal cualquier información concerniente a personas físicas identificadas o identificables”, de forma que mientras la información de los usuarios de un conjunto de datos no contenga nombres, direcciones, números de DNI o algún otro tipo de información que distinga a una persona de manera unívoca, no se verá afectada por esta ley. Esto sucederá en la mayoría de los casos, ya que lo que interesa suelen ser datos como la edad, el sexo, o los gustos y aficiones, los cuales no son suficientes para identificar a personas concretas. Además, en el apartado 2 del artículo 4 dice que “Los datos de carácter personal objeto de tratamiento no podrán usarse para finalidades incompatibles con aquellas para las que los datos hubieran sido recogidos. No se considerará incompatible el tratamiento posterior de éstos con fines históricos, estadísticos o científicos.”, y estos usos para sistemas de recomendación pueden considerarse estadísticos, ya que no interesan datos concretos de un usuario, sino utilizarlos para obtener una visión general de conjunto. De esta forma, la Ley Orgánica de Protección de Datos no supondrá un problema, siempre y cuando no se trabaje con datos sensibles, y lo

mismo pasaría con el Reglamento General de Protección de Datos [33], que es la normativa europea y en el punto 1 del artículo 4 dice “se entenderá por datos personales toda información sobre una persona física identificada o identificable, y se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente”. Además, según la consideración 4 “El tratamiento de datos personales debe estar concebido para servir a la humanidad. El derecho a la protección de los datos personales no es un derecho absoluto sino que debe considerarse en relación con su función en la sociedad y mantener el equilibrio con otros derechos fundamentales, con arreglo al principio de proporcionalidad”, de forma que podrían usarse datos privados de los dispositivos y ordenadores personales para asegurar una movilidad eficiente para la ciudad.

### 4.3. Influencia en la toma de decisiones y el mercado

En general, lo que se desea obtener de los sistemas de recomendación son predicciones acerca de los intereses de las personas, ya sean relacionados con el alojamiento que contratarían, las zonas de ocio que frecuentarían o los productos que estarían dispuestos a consumir. Una vez se disponga de toda esta información, la toma de decisiones relacionadas con la gestión del transporte en cada caso se simplificará en gran medida [34]. Si además se tiene también un estudio de mercado, encuestas realizadas, o datos históricos, toda esta información puede ser útil para comparar y validar los datos. Pero si no se dispone de otra fuente estos datos pueden convertirse en un muy buen punto de partida para comenzar a trabajar y planificar.

En el caso de la gran afluencia de turistas, el primer paso sería conocer en qué zonas se van a alojar. Para ello, los pasos a seguir serían: obtener las predicciones para alojamientos de interés de los usuarios según los datos disponibles, analizar estas predicciones para ver en qué zonas se encuentran estos alojamientos obteniendo así una distribución por la ciudad de los alojamientos de interés para los turistas, y ver en qué zonas se producen una mayor concentración de estos alojamientos de interés.

Si para llegar a estas zonas con puntos de interés hay buenos accesos no habrá problema, pero en caso contrario puede ser interesante proponer una línea de autobús temporal que la conecte con la estación o el aeropuerto, o si no se trata de una zona con muchos aparcamientos disponibles puede ser muy buena opción habilitar un aparcamiento temporal en algún solar o espacio libre cercano. Tras ello, otra posible información sería las zonas de ocio que más caudal de gente tendrán, de forma que habrá que comprobar si las conexiones actuales podrán soportar el exceso de pasajeros, o en caso contrario quizá haya que aumentar la frecuencia de autobuses o metros, o preparar la presencia de otras zonas de ocio creadas para la ocasión y que ofrezcan una alternativa a la gente para que se reparta entre ambas y no haya tanta concentración. Estos serían los ejemplos más claros, aunque podrían encontrarse muchos otros, por lo que es importante estudiar en



profundidad cada caso en cuestión y explorar todos los posibles contratiempos y problemas que puedan surgir para así explotar los datos al máximo y estar preparado para cualquier circunstancia [21].

En el caso de que se utilice en la distribución de mercancía por la ciudad durante la salida de un nuevo producto al mercado, como ya se ha comentado antes, su uso es mucho más intuitivo, influyendo en las rutas de reparto, las cantidades a proveer y la frecuencia de reabastecimiento, pudiendo optimizar así estos procesos mucho más que si no se dispusiera de datos al respecto, y resultando esto en un tráfico más fluido por la ciudad, que es uno de los objetivos a cumplir cuando se trabaja en la movilidad. En general, estas predicciones pueden ser útiles, y serán más o menos precisas en función de cómo y con qué datos se hayan realizado, pero no se deben tomar como hechos seguros, por lo que tampoco hay que despreciar todas las demás fuentes o desechar todas las demás planificaciones para cumplir a rajatabla los que se genere a partir de ellas.

De esta forma, el uso de predicciones realizadas por un sistema de recomendación proporcionará una serie de beneficios en cuanto a la planificación de la movilidad, como pueden ser la validación de los datos históricos o estudios de alojamiento realizados para prever el interés de los turistas, la optimización de las rutas de distribución de mercancías, el ajuste de los transportes públicos para transportar el mayor número de personas posible, etc., contribuyendo con todo ello a hacer el tráfico y el transporte por la ciudad más fluidos. No obstante, la aplicación de los sistemas de recomendación al campo de la movilidad también conllevará unos costes, siendo los principales los costes de gestión y económicos de la adquisición de los datos, los costes derivados del uso del sistema de recomendación (contratación de personal con conocimientos en el campo), y los costes de análisis de las predicciones y comparación con el resto de datos disponibles. Por lo tanto, la aplicación de sistemas de recomendación a la movilidad puede resultar provechosa o no en función de la comparación entre estos costes y beneficios potenciales, de manera que para cada caso concreto en el que se plantee su uso habrá que considerar si es adecuado o no. Por ejemplo, si los datos históricos de alojamientos son muy antiguos o la cantidad de turistas esperada es excepcionalmente grande, puede considerarse muy beneficioso el uso de sistemas de recomendación; por otro lado, si no se pueden obtener los datos necesarios fácilmente o no se dispone de trabajadores con perfiles que incluyan el análisis de datos en plantilla, quizá el coste sea demasiado grande para su uso.

Por último, mencionar que el uso de los sistemas de recomendación en estos contextos supondrá un empujón más para el mercado de la información [27]. Como ya se ha mencionado, el ritmo al que se generan datos e información es cada vez mayor debido a internet y las tecnologías de la comunicación, y gran parte de estos datos no se utilizan para nada. Con iniciativas de este tipo, se fomenta el aprovechamiento de los datos generados, y aunque es cierto que puede ser muy costoso recopilarlos, almacenarlos y mantenerlos actualizados, ofrece nuevas posibilidades de negocio que no se deben desperdiciar, puesto que estos mercados cobrarán más importancia en el futuro. Así, tanto

empresas que utilicen sistemas de recomendación en sus actividades, y que por lo tanto puedan ofrecer sus servicios en nuevos campos de estudio o sacar provecho al compartir unos datos que ya poseen y necesitan; como empresas dedicadas en exclusiva a recopilar y comerciar con información, pueden verse bastante impulsadas. A la vez se mejorará y se renovará el proceso de toma de decisiones en otros campos que a priori no tienen relación, pero que inevitablemente se van a ver afectados por el aumento de información disponible y pueden aprovecharla para mejorar su actividad. Con ello se conseguirán nuevas vías de comercio a nivel general en los mercados, mayor aprovechamiento de los recursos de información disponibles, y una actuación más eficiente, en este caso en el ámbito de la movilidad urbana y las ciudades modernas, pero se puede extrapolar a otros sectores y servicios, aumentando así la calidad de vida y fomentando el progreso social.

# Capítulo 5

## Conclusiones y posibles trabajos futuros

En este Capítulo se van a hacer unas conclusiones globales en la Sección 5.1, así como una valoración del trabajo en la Sección 5.2. En la Sección 5.3 se hablará de futuros trabajos posibles.

### 5.1. Conclusiones finales

A continuación, se van a comentar una serie de conclusiones que se han extraído durante la realización del trabajo. Tras el estudio de los sistemas de recomendación, se ha visto que se trata de un campo enormemente amplio y desarrollado, lo cual tiene sentido considerando la tendencia a aumentar de la información generada a nivel global, una idea que ha sido nombrada varias veces a lo largo de esta memoria ya que es uno de los grandes retos de la actualidad. Así, estos sistemas han tenido una gran importancia y están creciendo y sufriendo grandes cambios en la actualidad, siempre con el objetivo de ayudar a manejar y enfrentarse a toda esta cantidad de información, de manera que se están viendo muy impulsados por los avances en determinados campos de investigación dentro de la minería de datos y el *machine learning*, como la inteligencia artificial, las redes neuronales o el aprendizaje automático. Todo esto está transformando por completo el sector, por lo que se espera que crezca su importancia y desarrollo, y los resultados obtenidos sean cada vez más fáciles y rápidos de obtener, y mucho más precisos y útiles.

Centrándose en los conceptos más generales y en la experimentación realizada, la principal idea que se ha formado tras el trabajo ha sido que los sistemas de recomendación son métodos muy versátiles, que pueden ser útiles en una amplísima variedad de campos y contextos, pero precisamente por ello es muy difícil generalizar en cuanto a su uso. En función del caso o situación que se quiera tratar, será más óptimo el uso de unos métodos u otros, de forma que no se puede asegurar que haya una técnica superior a las demás, cada una encontrará su caso de aplicación, y además normalmente son válidas varias opciones. A lo que conduce todo esto es a que conviene tener conocimiento y una visión general sobre los distintos métodos para poder aplicar unos u otros y encontrar así el óptimo para un determinado objetivo. También es interesante destacar la importancia de la experimentación, ya que los resultados esperados se pueden mejorar cambiando el modo de aplicarlo o algún parámetro utilizado. Tiene igualmente una influencia capital el conjunto de datos que se va a utilizar, puesto que definirá qué información se puede obtener, qué técnica es más apropiado aplicar, y cómo se va a trabajar, así que dirigir esfuerzos a obtener un buen conjunto de datos puede ser la clave para que todo el proyecto triunfe o fracase.

En cuanto a la aplicación en el ámbito de la movilidad, se ha comprobado que las opciones son bastante diversas. Se pueden usar una gran variedad de datos, de diversas procedencias, y encontrar información útil que repercutirá en el transporte de un modo u otro. El uso de predicciones puede ser una muy buena base desde la que apoyarse al comenzar un proyecto, y puede aportar información muy interesante, por lo que se debería tener en cuenta además de otras fuentes de datos más tradicionales. Además, esta iniciativa se encamina hacia un futuro mucho más conectado, en el que los flujos de información se compartan entre unas áreas y otras, propiciando así unos sistemas de transporte más eficientes tanto para personas como para mercancías, y haciéndolos más flexibles para adaptarse a cualquier necesidad.

## 5.2. Valoración personal

A nivel personal, este trabajo me ha resultado muy interesante, porque desconocía por completo el campo de los sistemas de recomendación, y sentía curiosidad por ellos, ya que es algo que casi todos utilizamos de forma bastante asidua en nuestra vida cotidiana. He disfrutado investigando acerca de su funcionamiento y los distintos métodos, y a pesar de no poseer grandes conocimientos informáticos y de programación, creo que he adquirido una base sólida sobre esta área especializada que se enmarca en el ámbito de la minería de datos y el aprendizaje automático. Así, lo primero que tuve que hacer fue leer numerosos artículos, estudios, pruebas y tutoriales relacionados con los sistemas de recomendación para poder comenzar a visualizar el trabajo, y también me surgieron muchos problemas y dudas al utilizar el lenguaje R, de forma que fue necesario leer cursos y tutoriales para poder desenvolverme lo suficiente.

Además, el hecho de aplicarlos a un tema que también me parece de gran interés, como es el transporte y la movilidad, y haber encontrado tantas opciones y posibilidades que podrán ser usadas en un futuro, me ha resultado muy estimulante. Con todo ello, he utilizado conocimientos adquiridos en algunas asignaturas cursadas durante el máster como “Movilidad segura y sostenible” (concepto de movilidad urbana y cómo funciona), “Técnicas avanzadas de producción y logística interna” (distribución y logística de última milla) o “Business intelligence” (gestión y análisis de datos). A pesar de poseer estos conocimientos, también tuve que leer artículos, noticias, blogs y trabajos relacionados con la movilidad para poder aprender cómo aplicar los sistemas de recomendación a este campo.

De esta forma, entre la investigación y el aprendizaje, la realización de pruebas y experimentos, la obtención de conclusiones, y la elaboración y corrección de la memoria, he invertido alrededor de 400 horas de trabajo. Creo que este trabajo me resultará muy útil de cara al futuro, ya que he aprendido acerca de los sistemas de recomendación y la

gestión de información en general, y además haber combinado dos temas tan distintos como éste y la movilidad me ha ayudado a ampliar mi forma de pensar a la hora de afrontar un problema que no sepa resolver.

### 5.3. Trabajos futuros

Futuros trabajos posibles que sigan esta idea pueden ser la realización de unas predicciones completas acerca del alojamiento de turistas en una ciudad y su posterior comparación con datos reales, o el desarrollo de un proyecto de una línea de transporte público o sistema de movilidad privado (como los patines eléctricos) utilizando datos recomendados acerca de las zonas de ocio para la toma de decisiones.

También podría ser de gran interés la validación del uso de predicciones (generadas por sistemas de recomendación) en el campo de la movilidad para un caso concreto, observando lo que ocurre en realidad tras la planificación de la movilidad realizada, y analizando si se habría conseguido un mayor aprovechamiento de los recursos utilizando otro método. Por último, se podría realizar una experimentación más amplia con otros tipos de conjuntos de datos y con varios a la vez para comparar los resultados y ver cómo afectan los datos iniciales y cuáles son más adecuados para cada ocasión.

# Bibliografía

- [1] Jekaterina Novikova, PhD, University of Bath. “Building a Movie Recommendation System”. Junio, 2016. Disponible en Rpubs. <https://rpubs.com/jeknov/movieRec> (Consultado el 13-11-2019)
  
- [2] “Recommender Systems 101 – a step by step practical example in R”. Disponible en el blog Big Data Doctor. Diciembre, 2014. <https://www.r-bloggers.com/recommender-systems-101-a-step-by-step-practical-example-in-r/> (Consultado el 13-11-2019)
  
- [3] Brandon Hoefft, data scientist. “RecommenderLab Tutorial”. Disponible en GitHub. Octubre, 2017. [https://github.com/BrandonHoefft/Recommender-System-R-Tutorial/blob/master/RecommenderLab\\_Tutorial.md](https://github.com/BrandonHoefft/Recommender-System-R-Tutorial/blob/master/RecommenderLab_Tutorial.md) (Consultado el 13-11-2019)
  
- [4] M Hendra Herviawan. “Item-Based Collaborative Filtering Recommendation”. Disponible en Kaggle. 2017. <https://www.kaggle.com/hendraherviawan/itembased-collaborative-filter-recommendation-r> (Consultado el 13-11-2019)
  
- [5] Sergio Ilarri, Raquel Trillo-Lado, Ramón Hermoso, "Datasets for Context-Aware Recommender Systems: Current Context and Possible Directions", First Workshop on Context in Analytics (CiA 2018), in conjunction with the 34th International Conference on Data Engineering (ICDE 2018), Paris (France), IEEE Computer Society, Electronic ISBN 978-1-5386-6306-6, Print on Demand (PoD) ISBN 978-1-5386-6307-3, ISSN 2473-3490, pp. 25-28, April 2018. (DOI: 10.1109/ICDEW.2018.00011) <https://ieeexplore.ieee.org/document/8402014>
  
- [6] María del Carmen Rodríguez-Hernández, Sergio Ilarri, "Pull-Based Recommendations in Mobile Environments", Computer Standards & Interfaces, ISSN 0920-5489, volume 44, pp. 185-204, Elsevier, February 2016. DOI: 10.1016/j.csi.2015.08.002. <http://dx.doi.org/10.1016/j.csi.2015.08.002>
  
- [7] Vargas Pérez, P. y Leiva Olivencia, J. L.: “Prototipo de sistema de recomendación grupal en un destino turístico” Rotur. Revista de Ocio y Turismo, 9: 62-81, (2015), <http://www.rotur.es>, ISSN: 1888-6884.

- [8] Zheng, Y. and Burke, R. and Mobasher, B. "Differential Context Relaxation for Context-aware Travel Recommendation". 13th International Conference on Electronic Commerce and Web Technologies (EC-WEB 2012), pp. 88--99, 2012 [https://github.com/irecsys/CARSKit/tree/master/context-aware\\_data\\_sets](https://github.com/irecsys/CARSKit/tree/master/context-aware_data_sets) (Consultado 17-11-2019)
- [9] Michael Hahsler, Bregt Vereet. "Lab for Developing and Testing Recommender Algorithms". Package 'recommenderlab'. Repositorio CRAN. Agosto, 2019
- [10] María del Carmen Rodríguez-Hernández, Sergio Ilarri, Raquel Trillo-Lado, Ramón Hermoso, "Context-Aware Recommendations Using Mobile P2P", 15th International Conference on Advances in Mobile Computing & Multimedia (MoMM 2017), Salzburg (Austria), ACM Press, ISBN 978-1-4503-5300-7, pp. 82-91, December 2017.
- [11] María del Carmen Rodríguez-Hernández, Sergio Ilarri, Ramón Hermoso, Raquel Trillo-Lado, "DataGenCARS: A Generator of Synthetic Data for the Evaluation of Context-Aware Recommendation Systems", Pervasive and Mobile Computing, ISSN 1574-1192, volume 38, part 2, pp. 516-541, Elsevier, July 2017.
- [12] Estructuras de datos en R. Manual R. <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/ManualR/intro.html> (Consultado el 13-11-2019)
- [13] Cesar Augusto Guzmán Álvarez. "Sistema de recomendación y planificación Turística de la ciudad de Valencia vía web". Tesis. Departamento de Sistemas Informáticos y Computación. Universidad de Valencia. Enero, 2009
- [14] Vicente Coll & Pedro J. Pérez. Curso de introducción a R. Junio, 2017. [https://www.uv.es/pjperez/curso\\_R/index.html](https://www.uv.es/pjperez/curso_R/index.html) (Consultado el 13-11-2019)
- [15] Juan Bosco Mendoza Vega, Universidad Nacional Autónoma de México (UNAM). "R para principiantes". Libro en bookdown.org. 2016. <https://bookdown.org/jboscomendoza/r-principiantes4/operadores-logicos.html> (Consultado el 13-11-2019)

- [16] YouTube. Web para compartir vídeos. [www.youtube.com](http://www.youtube.com) (Consultado el 16-11-2019)
- [17] Netflix. Servicio de contenido multimedia a la carta. [www.Netflix.com](http://www.Netflix.com) (Consultado el 13-11-2019)
- [18] FilmAffinity. Red social de valoraciones para películas y series. <https://filmaffinity.com> (Consultado el 13-11-2019)
- [19] Google Maps. Servicio de localización GPS con información sobre establecimientos. [www.google.com/maps](http://www.google.com/maps) (Consultado el 13-11-2019)
- [20] Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Publicada en BOE núm. 298, de 14/12/1999. Última actualización publicada el 06/12/2018.
- [21] Bernardos, A.B. “Servicios y aplicaciones en movilidad para el sector turístico”. (2007) CITIC, Madrid.
- [22] Castejón, R. & Méndez, E.: “Introducción a la economía para turismo” (2012). Prentice-Hall. España
- [23] Pérez, L.G. (2008), “Modelo de recomendación con falta de información. Aplicaciones al sector turístico”. Tesis doctoral. Universidad de Jaén.
- [24] Schafer, J.; Frankowski, D.; Herlocker, J. & Sen, S.(2007) “Collaborative Filtering Recommender Systems”. In Adaptive Web 2007, LNCS, vol. 4321, Heidelberg, Springer: 291-324
- [25] Wikipedia. [www.wikipedia.org](http://www.wikipedia.org) (Consultado el 15-11-2019)
- [26] Fesenmaier D.R., Ricci F., Schaumlechner E., Wober K., Zanella C. Dietorecs: “Travel advisory for multiple decision styles”. En Information and



Communication Technologies in Tourism 2003: Proceedings of the International Conference in Lausanne (2003). pages 232-241.

- [27] Francisco Femenia-Serra, Sandra Navarro-Ruiz. “Identificación de fuentes de datos para la construcción de un nuevo enfoque de planificación de destinos inteligentes” (2018). Instituto universitario de investigación turística, Universidad de Alicante.
  
- [28] Bigné, J. E., Font, X., y Andreu, L. (2000). “Márketing de destinos turísticos: análisis y estrategias de desarrollo”. Editorial Esic.
  
- [29] Guevara, A. (2008). Sistemas de información de destinos turísticos integrados (SIDTI). En “Investigaciones turísticas. Una perspectiva multidisciplinar: I jornadas de investigación en turismo”.
  
- [30] Ignacio Olmeda Martos, Eugenio Fernández Vicente y M<sup>a</sup> del Mar Miguel Colom. “Sistemas de predicción para la demanda de plazas hoteleras: el proyecto PLAZA”. Estudios Turísticos, n.º 142 (1999), pp. 85-96
  
- [31] Angela de Vitta. “¿Qué es un forecast?”. Artículo en la web Les Hoteliers. <https://www.leshoteliers.com/que-es-un-forecast/> (Consultado el 19-11-2019)
  
- [32] Ramón Hermoso, Sergio Ilarri, Raquel Trillo, María del Carmen Rodríguez-Hernández, "Push-Based Recommendations in Mobile Computing Using a Multi-Layer Contextual Approach", 13th International Conference on Advances in Mobile Computing & Multimedia (MoMM 2015), Brussels (Belgium), ACM Press, ISBN 978-1-4503-3493-8, pp. 149-158, December 2015. DOI: 10.1145/2837126.2837128. <http://dx.doi.org/10.1145/2837126.2837128>
  
- [33] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)

- [34] F. Calabrese, G. Di Lorenzo and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, 2010, pp. 312-317. DOI: 10.1109/ITSC.2010.5625119. URL:<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5625119&isnumber=5624963>
- [35] S. Foell, S. Phithakkitnukoon, G. Kortuem, M. Veloso and C. Bento, "Catch me if you can: Predicting mobility patterns of public transport users," 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, 2014, pp. 1995-2002. DOI: 10.1109/ITSC.2014.6957997 URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6957997&isnumber=6957655>
- [36] Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science*, 45(1), 12–19. [https://doi.org/10.1002/\(sici\)1097-4571\(199401\)45:1<12::aid-asi2>3.0.co;2-1](https://doi.org/10.1002/(sici)1097-4571(199401)45:1<12::aid-asi2>3.0.co;2-1)

# Lista de figuras

3.1	Fragmento del conjunto de datos de TripAdvisor . . . . .	15
3.2	Fragmento de la RatingMatrix creada . . . . .	17
3.3	Histograma de votos . . . . .	18
3.4	Histograma de votos tras la normalización . . . . .	19
3.5	Curva ROC . . . . .	21
3.6	Precisión frente a recall . . . . .	22
3.7	Medidas de error de las predicciones . . . . .	22
3.8	Número de usuarios en función de sus valoraciones . . . . .	23
3.9	Error medio de las predicciones para un usuario con 1 valoración conocida . . . . .	24
3.10	Error medio de las predicciones para un usuario con 3 valoraciones conocidas . . . . .	24
3.11	Error medio de las predicciones para un usuario con 5 valoraciones conocidas . . . . .	24
3.12	Error medio de las predicciones para un usuario con 6 valoraciones conocidas . . . . .	25
A.1	Tabla con usuario, alojamiento y valoración. . . . .	46
A.2	Distribución de las valoraciones de la base de datos . . . . .	47
A.3	Valoraciones disponibles en la RatingMatrix con valores incorrectos. . . . .	47
A.4	Distribución de la media de valoraciones para cada usuario. . . . .	48
A.5	Conjuntos de datos creados con “evaluationScheme” . . . . .	48
A.6	Curva ROC trazada con evaluación inadecuada . . . . .	49
A.7	Ejemplo de resultados obtenidos de la función “evaluate” . . . . .	50
A.8	Valoraciones conocidas de cada usuario en el conjunto de test . . . . .	50



# Anexos



# Anexo A

## Experimentación

Aquí se van a enumerar algunos de los pasos seguidos durante el proceso de realización de los experimentos, que van a mostrar cómo se ha obtenido alguna de la información extraída y cómo se ha trabajado con los datos. A continuación se muestran algunas imágenes y gráficos representativos de todas las pruebas realizadas.

El objetivo de la experimentación es validar el uso de sistemas de recomendación para la obtención de predicciones en cuanto al alojamiento de turistas para su aplicación a la movilidad. En primer lugar se pueden ver algunas de las primeras filas de la tabla creada a partir de la base de datos inicial de TripAdvisor (ver Figura A.1) la cual contenía más información de la necesaria, de forma que esta tabla contiene solamente al usuario, el alojamiento y la valoración. Esta es la estructura a partir de la cual se crea la *RatingMatrix*.

```
> RatingsTripAdvisor[1:15,]
  ID_USER ID_HOTEL Rating
1      45   105170      5
2      45   223229      5
3      45   258688      5
4      45    98827      5
5      45    99518      3
6      64   224427      4
7      64  1751886      4
8      64    99120      4
9     100   120111      5
10     100    91428      4
11     100   109399      3
12     100    98764      3
13     149   240681      3
14     149    73927      3
15     149   100584      1
```

Figura A.1: Tabla con usuario, alojamiento y valoración

A continuación se muestran los recuentos totales de todas las valoraciones de la base de datos (ver Figura A.2). Comentar que aquí todavía se encuentran las valoraciones repetidas que posteriormente son eliminadas, así que el número total de valoraciones no coincidirá con el número de valoraciones de la *RatingMatrix*.

1	2	3	4	5
219	470	1163	1771	1046

Figura A.2: Distribución de las valoraciones de la base de datos

Tras la creación de la *RatingMatrix* de forma directa, sí se observan las valoraciones existentes en ella (ver Figura A.3) se puede apreciar cómo aparecen algunas con valores superiores a 5, de forma que descubrimos que hay algún error al crear esta estructura.

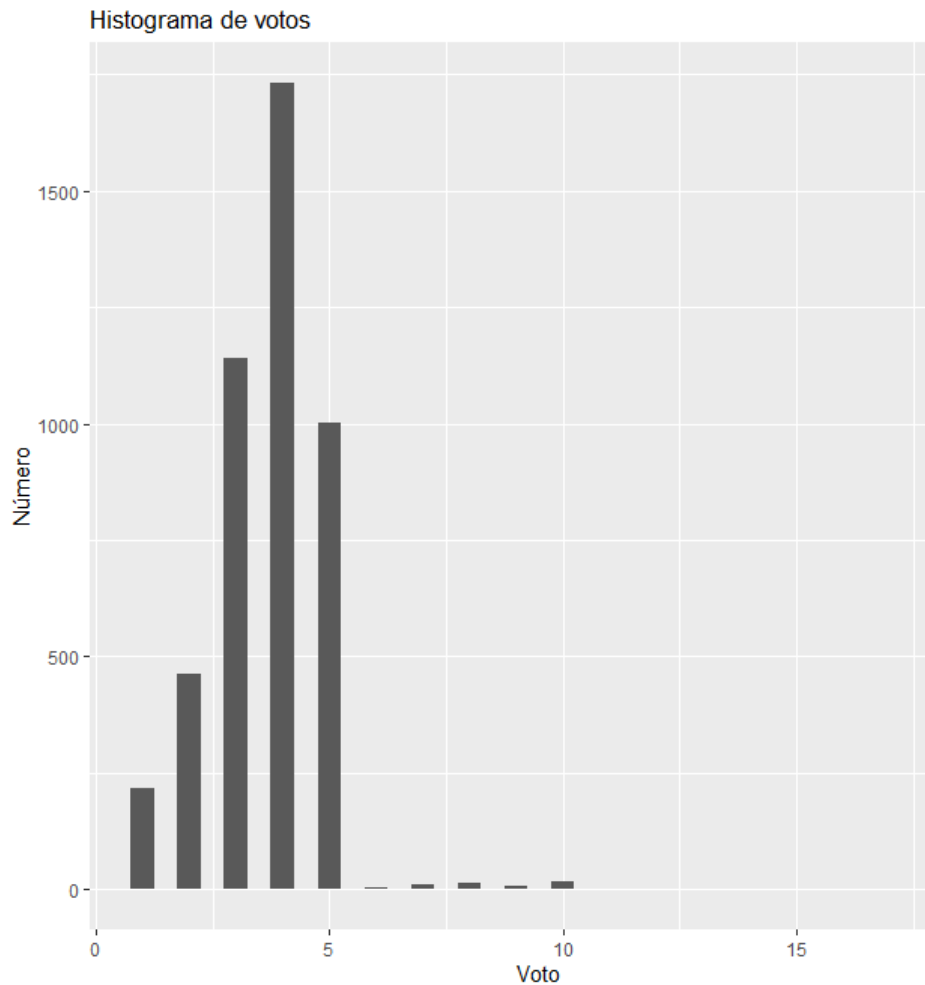


Figura A.3: Valoraciones disponibles en la *RatingMatrix* con valores incorrectos

Tras descubrir el origen del problema y solucionarlo como se ha comentado en la Sección 3.2, se obtiene la *RatingMatrix* con los datos adecuados. De esta forma, si se visualiza en una gráfica las valoraciones de la *RatingMatrix* se obtiene una gráfica con valores de 1 a 5 únicamente (ver Figura 3.2).



Ahora es muy fácil visualizar y trabajar con los datos. Se puede ver por ejemplo las medias de las valoraciones de cada usuario y cómo están distribuidas (ver Figura A.4). Si se analiza esto, se puede llegar a la conclusión de que estas medias de valoraciones dadas son muy dispersas, lo que nos da una pista de la presencia de error de sesgo, motivo por el cual se lleva a cabo la normalización de las valoraciones calculando la media de valoraciones de cada usuario y representando cada valoración como su diferencia respecto a dicha media individual, además dividida por la desviación estándar (z-score).

Tras utilizar la función “evaluationScheme” se puede comprobar cómo se crean distintos conjuntos de datos: uno para entrenamiento, otro con los datos que se consideran conocidos de un nuevo usuario para el *test*, y otro con los datos que se consideran desconocidos de ese mismo usuario a fin de poder comparar las predicciones realizadas con datos reales (ver Figura A.5).

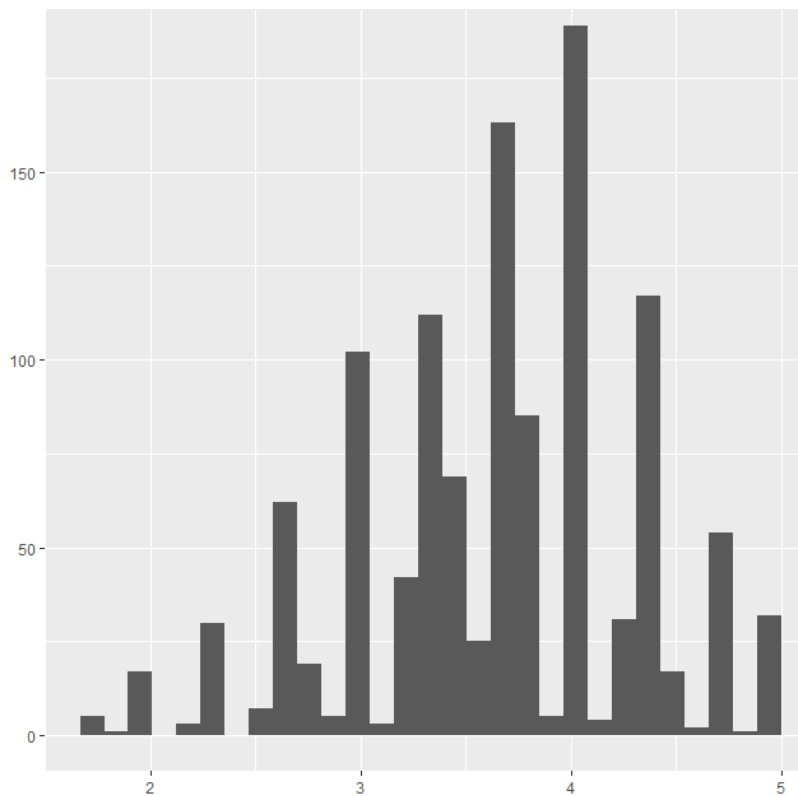


Figura A.4: Distribución de la media de valoraciones para cada usuario

```
> scheme<-evaluationScheme(RatingMatrixTripAdvisor,method="split",train=.9,k=1,given=3,goodRating=4)
> getData(scheme,"train")
1081 x 1890 rating matrix of class 'realRatingMatrix' with 4147 ratings.
> getData(scheme,"known")
121 x 1890 rating matrix of class 'realRatingMatrix' with 363 ratings.
> getData(scheme,"unknown")
121 x 1890 rating matrix of class 'realRatingMatrix' with 96 ratings.
```

Figura A.5: Conjuntos de datos creados con “evaluationScheme”

Si se hacen las predicciones con estos conjuntos de datos, lo que se obtiene no tiene mucho sentido, ya que según esto el método aleatorio no es el que peores recomendaciones hace (ver Figura A.6). Esto es porque se ha utilizado el método “split” que es el más simple, el porcentaje de usuarios de test es muy pequeño, sólo se repite la evaluación una vez, y además se consideran conocidas 3 valoraciones por usuario nuevo, de forma que habrá usuarios en el conjunto de test que no tengan datos que se consideren desconocidos para poder comparar con las predicciones efectuadas.

Si se observa la curva ROC (ver Figura A.6) se ve que el FPR es mayor que el TPR, lo que indica que las predicciones no son buenas, el método de populares mantiene en todo momento su TPR en 0 (lo cual ocurre porque no se disponen de suficientes datos reales para comparar, y los pocos que hay no han coincidido), y el método aleatorio da mejores resultados, lo cual no tiene ningún sentido. Esto varía cada vez que se hace la simulación aunque sea con los mismos parámetros, ya que no es consistente por la variabilidad al dividir el conjunto en datos de entrenamiento y datos de test solo 1 vez. Así pues, estos datos mejorarían si se repitiera la evaluación varias veces, y se hiciera la media de los resultados.

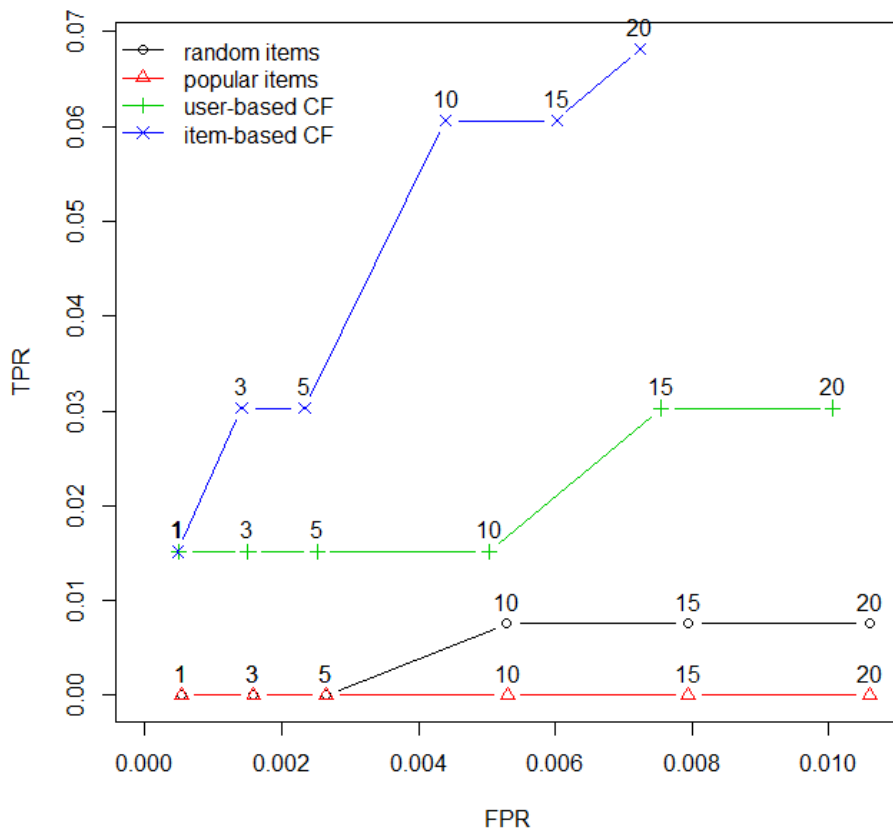


Figura A.6: Curva ROC trazada con evaluación inadecuada

En la Figura A.7 se muestra un ejemplo de los resultados que se obtienen para cada uno de los métodos, donde se puede apreciar que hay valores de positivos reales, falsos positivos, negativos reales y falsos negativos, así como la precisión, el *recall*, el ratio de positivos reales y el ratio de positivos falsos, además de cómo varían al cambiar el número de recomendaciones realizadas.

Por último, se puede ver cómo se comprueban las valoraciones conocidas en cada usuario del conjunto de test (ver Figura A.8) cuando se comparan usuarios por separado para ver cómo afecta al error de predicción el disponer de más o menos información de un usuario, como se ha explicado en la sección 3.4. Los datos se muestran en parejas de 2, donde el superior es el número que representa al usuario, y el inferior el número de valoraciones conocidas.

```
> getConfusionMatrix(results[[4]])
[[1]]
      TP      FP      FN      TN  precision  recall      TPR
1  0.008264463  0.9008264  0.4297521 1885.661 0.009090909 0.01515152 0.01515152
3  0.016528926  2.6694215  0.4214876 1883.893 0.006060606 0.03030303 0.03030303
5  0.016528926  4.4049587  0.4214876 1882.157 0.003636364 0.03030303 0.03030303
10 0.033057851  8.2975207  0.4049587 1878.264 0.003636364 0.06060606 0.06060606
15 0.033057851 11.3636364  0.4049587 1875.198 0.002424242 0.06060606 0.06060606
20 0.041322314 13.6528926  0.3966942 1872.909 0.002352941 0.06818182 0.06818182
      FPR
1  0.0004774901
3  0.0014149539
5  0.0023348919
10 0.0043981597
15 0.0060233871
20 0.0072368290
```

Figura A.7: Ejemplo de resultados obtenidos de la función “evaluate”

```
> rowCounts(getData(scheme,"known"))
159  451  753  2557  2572  4307  4410  5255  5653  5762  5993  6016  6563
  1  1  1  4  1  1  2  1  2  2  1  4  2
6588 6791 7913 8012 8591 8685 9049 10869 12133 12296 12462 12610 13087
  1  2  3  1  2  2  1  1  1  2  4  2  14
13551 14198 14281 15195 15534 15596 15704 16408 16935 17675 18047 19382 20212
  2  1  3  1  1  2  2  1  1  1  1  1  3
21527 21708 22383 24256 24723 26172 26532 26609 29190 29870 30568 30796 31574
  1  1  1  1  2  1  1  1  1  1  1  1  2
33018 33880 33965 34443 34997 35581 35590 35741 36057 36103 36556 37862 38125
  1  1  4  1  2  1  3  1  1  1  4  1  1
38292 38304 39180 39975 41081 41243 41410 41684 41716 41749 41828 42077 42214
  1  1  9  1  1  1  1  1  1  1  2  4  2
42753 42950 44323 45017 45336 45747 47181 47197 47351 47428 48971 49181 49360
  1  1  1  2  2  1  1  1  1  1  4  3  3
49867 50441 51410 51557 52657 53007 53205 53514 53707 53733 55308 55348 56300
  1  1  1  3  3  2  1  1  1  1  7  1  2
56649 58075 58245 58483 58908 60215 61035 61872 62186 62258 62347 62658 63028
  1  1  1  1  1  1  1  2  2  1  5  1  1
63522 64278 64479 65440 65457
  1  3  1  1  1
```

Figura A.8: Valoraciones conocidas de cada usuario en el conjunto de test



# Anexo B

## Código utilizado

```
library(recommenderlab)
```

```
library(ggplot2)
```

```
TripAdvisor <- read.csv(file.choose(),header=TRUE,sep=';')
```

```
RatingsTripAdvisor<-TripAdvisor[,c("ID_USER","ID_HOTEL","Rating")]
```

```
qplot(RatingsTripAdvisor[,3],binwidth=0.5, main="Histograma de  
votos",xlab="Voto",ylab="Número")
```

```
valoraciones<-table(RatingsTripAdvisor[,3])
```

```
valoraciones
```

```
RatingMatrixTripAdvisor<-as(RatingsTripAdvisor,"realRatingMatrix")
```

```
qplot(getRatings(RatingMatrixTripAdvisor), binwidth=0.5, main="Histograma de votos",  
xlab="Voto",ylab="Número")
```

```
##en la matriz hay valoraciones mayores que 5, porque en la base de datos hay datos  
repetidos
```

```
RatingsTripAdvisor[3546,] ##este valor pasa de 5 a 10 al hacer la rating matrix, ya que  
es uno de los que están repetidos
```

```
##para solucionarlo vamos a hacer un bucle que recorra la tabla de valoraciones y las  
vaya pasando a otra tabla siempre que no esté repetida con la anterior (las repetidas  
van seguidas, al estar la tabla ordenada)
```

```
x<-RatingsTripAdvisor[1,]
```

```
i<-2
```

```
while(i<4669){
```

```
if(RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-1,1] |  
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-1,2])
```

```
{x<-rbind(x,RatingsTripAdvisor[i,])
```

```

i<-i+1}

else i<-i+1}

##comprobamos ahora que la ratingmatrix esta bien

RatingMatrixTripAdvisor<-as(x,"realRatingMatrix")

qplot(getRatings(RatingMatrixTripAdvisor), binwidth=0.5, main="Histograma de votos",
xlab="Voto",ylab="Número")

##sigue habiendo algunos valores mal, ya que las valoraciones repetidas no siempre
están juntas como se pensaba al principio; por ejemplo el usuario 43608 tiene una
valoracion repetida y no están juntas

##posible solución->compararlo no solo con el anterior sino con los 5 anteriores. Esto se
podría hacer de forma más generalizada comparando con todos, pero como en este
conjunto de datos las valoraciones de un mismo usuario están todas juntas y la mayoría
de usuarios no han dado más de 5 valoraciones, se ha comprobado que con este
método se eliminan las valoraciones repetidas, ya que la distribución de las valoraciones
de la RatingMatrix es la misma que la de la tabla en la que aparecían todas las
valoraciones. La forma de hacerlo sería con un segundo bucle dentro del primero para
que cada pareja de usuario-item la comparara con todas las ya añadidas, pero esto
tiene un mayor coste computacional, y puesto que en este caso no es necesario, se ha
utilizado este método que es más rápido y para este conjunto de datos es suficiente.

x<-RatingsTripAdvisor[1:5,]

i<-6

while(i<4669){

if((RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-1,1] |
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-
1,2])&(RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-2,1] |
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-
2,2])&(RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-3,1] |
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-
3,2])&(RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-4,1] |
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-
4,2])&(RatingsTripAdvisor[i,1]!=RatingsTripAdvisor[i-5,1] |
RatingsTripAdvisor[i,2]!=RatingsTripAdvisor[i-5,2]))

{x<-rbind(x,RatingsTripAdvisor[i,])

i<-i+1}

```

```

else i<-i+1}

##comprobamos ahora que la ratingmatrix esta bien

RatingMatrixTripAdvisor<-as(x,"realRatingMatrix")

qplot(getRatings(RatingMatrixTripAdvisor), binwidth=0.5, main="Histograma de votos",
xlab="Voto",ylab="Número")

##se observa la distribución de la media de las valoraciones por usuario

qplot(rowMeans(RatingMatrixTripAdvisor))

##se normalizan los datos para eliminar el error derivado de que unos usuarios den
puntuaciones más altas y otros más bajas de forma sistemática

##para ello se hace la media de todas las valoraciones de un usuario y se da la
desviación respecto de esta media, con el método Z-score, además se divide entre la
desviación típica

RMTripAdvisorNormalized<-normalize(RatingMatrixTripAdvisor,method="Z-score")

qplot(getRatings(RMTripAdvisorNormalized), binwidth=0.5, main="Histograma de votos
normalizados", xlab="Voto",ylab="Número")

##se va a comprobar la cantidad de valoraciones dadas por cada usuario

qplot(rowCounts(RatingMatrixTripAdvisor))

##se separan los datos en un conjunto de entrenamiento y otro de evaluación; además
es necesario decir el número de objetos valorados por usuario a partir del cual se harán
las recomendaciones

min(rowCounts(RatingMatrixTripAdvisor)) ## para saber el número de valoraciones
mínimo por usuario

scheme<-
evaluationScheme(RatingMatrixTripAdvisor,method="split",train=.9,k=1,given=3,goodR
ating=4)

getData(scheme,"train")

getData(scheme,"known")

```

##además, el parámetro k representa el número de veces que se repite la evaluación, de forma que cuanto mayor sea más entrenado y más preciso será nuestro recomendador

##se hace una lista con los distintos métodos de recomendación para poder evaluarlos y compararlos a partir de nuestros datos

```
algorithms <- list(
  "random items" = list(name="RANDOM", param=list(normalize = "Z-score")),
  "popular items" = list(name="POPULAR", param=list(normalize = "Z-score")),
  "user-based CF" = list(name="UBCF", param=list(normalize = "Z-score",
  method="Cosine",
  nn=50)),
  "item-based CF" = list(name="IBCF", param=list(normalize = "Z-score"))
)
```

##se aplica cada uno de los métodos al conjunto de datos de evaluación creado anteriormente, variando el número de objetos a recomendar entre 1 y 20

```
results <- evaluate(scheme, algorithms, n=c(1, 3, 5, 10, 15, 20))
```

##se puede observar el tiempo de creación del modelo y el tiempo de predicción para cada método

##a continuación se representa el ratio de positivos reales frente al de falsos positivos para cada método y en función del número de recomendaciones, así como la precisión frente al recall

```
plot(results, annotate=1:4, legend="topleft")
```

```
plot(results, "prec/rec", annotate=2:4)
```

```
getConfusionMatrix(results[[4]])
```

##cuanto más se entrene a los métodos colaborativos, mejores son sus resultados en comparación con el método aleatorio

##si se crean los conjuntos de entrenamiento con "cross-validation" y una k=10 se aprecia bien la diferencia



```
scheme<-evaluationScheme(RatingMatrixTripAdvisor,method="cross-validation",train=.9,k=10,given=3,goodRating=4)
```

```
results <- evaluate(scheme, algorithms, n=c(1, 3, 5, 10, 15, 20))
```

```
plot(results, annotate=1:4, legend="topleft")
```

```
plot(results, "prec/rec", annotate=2:4)
```

##los resultados obtenidos parecen muy variables, puesto que cada vez que se crean los conjuntos de datos de entrenamiento y se evalúan, a pesar de usar los mismos métodos, los resultados obtenidos son bastante distintos

##además, en vez de usar la función "evaluate", es posible ir paso a paso y entrenar un recomendador a partir de los datos de entrenamiento, usarlo para predecir resultados con unos nuevos datos de entrada y calcular el error de esos datos predichos

```
rec_model <- Recommender(getData(scheme, "train"), method = "UBCF")##esto se puede hacer para cualquiera de los métodos de recomendación
```

```
prediccion <- predict(rec_model, getData(scheme, "known"), type="ratings")
```

```
calcPredictionAccuracy(prediccion, getData(scheme, "unknown"))
```

##al crear los conjuntos de entrenamiento y test con el método "split" y usar "calcPredictionAccuracy" con los 4 métodos, los mejores resultados se obtienen con el UBCF, ya que es el que menor error presenta

##para conseguir unos resultados consistentes y representativos, se va a utilizar el método de "cross-validation" con k=10 y un 80% de los datos para entrenamiento y un 20% para test

## además, se le va a decir que para los usuarios de test haga las recomendaciones cogiendo 2 items valorados, ya que hay usuarios que solo han valorado 3 items, así que dejaríamos uno de ellos para comprobar si la predicción ha sido acertada

```
scheme<-evaluationScheme(RatingMatrixTripAdvisor,method="cross-validation",train=.8,k=10,given=2,goodRating=4)
```

```
results <- evaluate(scheme, algorithms, n=c(1, 3, 5, 10, 15, 20))
```

```
plot(results, annotate=1:4, legend="topleft")
```

```
plot(results, "prec/rec", annotate=2:4)
```

##en los gráficos se puede apreciar que los métodos colaborativos ofrecen unos resultados mucho mejores que los métodos simples

##además, en este caso parece que el recomendador IBCF funciona mejor que el basado en usuarios, esto puede ser debido a que en este conjunto de datos cada uno de los usuarios tiene pocas valoraciones, por lo que es difícil calcular la similitud entre usuarios

##a continuación se va a calcular el error medio de las predicciones con cada uno de los métodos

```
rec_model <- Recommender(getData(scheme, "train"), method = "UBCF")
prediccion <- predict(rec_model, getData(scheme, "known"), type="ratings")
error.ubcf<-calcPredictionAccuracy(prediccion, getData(scheme, "unknown"))
```

```
rec_model <- Recommender(getData(scheme, "train"), method = "IBCF")
prediccion <- predict(rec_model, getData(scheme, "known"), type="ratings")
error.ibcf<-calcPredictionAccuracy(prediccion, getData(scheme, "unknown"))
```

```
rec_model <- Recommender(getData(scheme, "train"), method = "POPULAR")
prediccion <- predict(rec_model, getData(scheme, "known"), type="ratings")
error.popular<-calcPredictionAccuracy(prediccion, getData(scheme, "unknown"))
```

```
rec_model <- Recommender(getData(scheme, "train"), method = "RANDOM")
prediccion <- predict(rec_model, getData(scheme, "known"), type="ratings")
error.random<-calcPredictionAccuracy(prediccion, getData(scheme, "unknown"))
```

```
error<-rbind(error.ubcf, error.ibcf, error.popular, error.random)
```

```
rownames(error)<-c("UBCF", "IBCF", "POPULAR", "RANDOM")
```

```
error
```

##el mayor error medio como cabía esperar aparece con el método RANDOM, después el POPULAR, el siguiente el UBFC y el mejor resultado vuelve a ser el del método IBFC

##así pues, para obtener las recomendaciones para un determinado usuario con el método IBCF

```
rec_model <- Recommender(getData(scheme, "train"), method = "IBCF")
```

```
recomendacion <- predict(rec_model, RatingMatrixTripAdvisor[12,], n=10)
```

```
as(recomendacion, "list")
```

##en este caso se han recomendado 10 items para el usuario que ocupa la posición 12 en la RatingMatrix

##a continuación se va a obtener el top 3 de esas recomendaciones

```
recomendacion_top3 <- bestN(recomendacion, n=3)
```

```
as(recomendacion_top3, "list")
```

##también puede ser bastante interesante comparar los resultados obtenidos para determinados usuarios en función del número de valoraciones de los mismos

```
scheme <- evaluationScheme(RatingMatrixTripAdvisor, method="cross-validation", train=.8, k=10, given=-2, goodRating=4)
```

```
rowCounts(getData(scheme, "known"))
```

```
rowCounts(getData(scheme, "unknown"))
```

##ejemplo de cálculo de precisión para método UBCF y un usuario determinado (elegir uno con pocas valoraciones conocidas y otro con muchas valoraciones conocidas)

```
rec_model <- Recommender(getData(scheme, "train"), method = "UBCF")
```

```
pred <- predict(rec_model, getData(scheme, "known")[1,], type="ratings")
```

```
pred
```

```
rowCounts(getData(scheme, "known")[1,])
```

```
rowCounts(getData(scheme, "unknown")[1,])
```

```
calcPredictionAccuracy(pred, getData(scheme, "unknown")[1,])
```



# Anexo C

## Conjuntos de datos

En este anexo se va a comentar de forma rápida algunos de los conjuntos de datos posibles considerados antes de seleccionar el utilizado finalmente.

En primer lugar se consideró utilizar un conjunto de datos de restaurantes (<https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data>) pero posteriormente se decidió que sería más interesante trabajar con alojamientos, ya que es mucho más crucial saber dónde se alojarán los turistas que los restaurantes que pueden ser de su interés, ya que en el caso de que acudan por un evento, es muy probable que no vayan de propio a comer a un restaurante aunque sea el más apropiado para ellos, sino que comerán cerca de donde será el evento o en alguna zona entre su alojamiento y el evento, o incluso en una zona que concentre locales de ocio, pero no se desplazarán a un punto alejado de la ciudad solamente para buscar un restaurante.

El siguiente paso fue consultar conjuntos de datos de hoteles. Se buscaron algunos por España ([https://datos.gob.es/es/catalogo?theme\\_id=turismo](https://datos.gob.es/es/catalogo?theme_id=turismo)), y se vieron conjuntos de datos que enumeraban todos los hoteles existentes en determinados pueblos o municipios (por ejemplo, obtenidos de la web de la comunidad del País Vasco, <https://opendata.euskadi.eus/catalogo/-/alojamientos/>), pero carecían de valoraciones de los clientes, o si las tenían, eran comentarios acerca de las impresiones o la calidad de su estancia, pero no valoraciones numéricas, de forma que no podían usarse directamente para los sistemas de recomendación.

Por último, se recurrió a diversos conjuntos de datos existentes en Github. Estos conjuntos de datos tenían hoteles que no pertenecían a España, pero sí que disponían de valoraciones en una escala de 1 a 5, por lo que eran adecuados, y además disponían de buena documentación y parecía que habían sido usados en numerosas ocasiones, por lo que se pensó que serían fiables. De entre ellos, se seleccionó el de hoteles de TripAdvisor debido a que tenía un tamaño suficiente, pero no era demasiado grande, ya que si se hubiera elegido un conjunto mucho mayor el tiempo de experimentación se habría alargado bastante, aunque también se hubiera dispuesto de una mayor cantidad de datos de entrenamiento y quizá más valoraciones por usuario.

Otro motivo que se tuvo en cuenta durante la decisión fue que este conjunto de datos no tenía muchos datos innecesarios, ya que había otros que también disponían de comentarios o más datos de contexto que no iban a ser utilizados.



# Anexo D

## Métricas de resultados

En este anexo se van a comentar algunas de los métodos utilizados para analizar y comparar los resultados.

**RMSE** – Para medir el error se ha utilizado el valor de RMSE, o *Root Mean Square Error*, que representa la raíz cuadrada del promedio de los errores al cuadrado. Esto es muy utilizado para dar una idea de la diferencia entre unos valores predichos o estimados y unos datos reales. Se puede usar para comparar la predicción de esos valores con diferentes métodos, pero no para comparar predicciones de valores distintos entre sí, ya que el RMSE depende de la escala.

Este método se usa porque el efecto de cada error en el valor de RMSE es proporcional al tamaño del error cuadrado, y por lo tanto, los errores mayores tienen un efecto desproporcionadamente grande, lo que hace esta medida muy sensible a valores atípicos o muy alejados de lo esperado.

**Precisión y recall** – La precisión se calcula como  $TP/(TP+FP)$ , y por lo tanto representa el porcentaje de recomendaciones realmente correctas que hay entre todas las recomendaciones realizadas. El *recall* se calcula como  $TP/(TP+FN)$ , y representa la sensibilidad, es decir, el porcentaje de recomendaciones correctas que se hacen entre todas las recomendaciones correctas posibles que se podrían hacer.

**Curva ROC** – La curva ROC es un acrónimo de *Receiver Operating Characteristic*, o Característica operativa del receptor. Esta curva es una representación gráfica de la sensibilidad frente a la especificidad de un sistema. Esto quiere decir que muestra cómo varían los resultados conforme se varía un parámetro que afecta a lo restrictivo que es el sistema para dar soluciones.

En este caso concreto, lo que se aumenta es el número de recomendaciones realizadas al usuario, de forma que si solamente se hacen unas pocas recomendaciones es más probable que se acierte y las recomendaciones sean realmente de interés para el usuario, y conforme se aumenta el número de recomendaciones realizadas, será más difícil encontrar objetos de su interés. De esta forma hay que llegar a un compromiso entre la precisión, intentando que la mayor parte posible de las recomendaciones hechas sean correctas, y el *recall*,

consiguiendo que se recomienden la mayor cantidad posible de ítems realmente relevantes para el usuario (*trade-off* entre precisión y *recall*) [36].

Así, en este trabajo se representa el ratio de positivos reales y el ratio de positivos falsos, de forma que si al hacer el sistema menos restrictivo (esto es, en vez de hacer solamente la mejor recomendación, hacer las 20 mejores, de manera que se restringen menos los resultados y es posible que aparezcan más resultados incorrectos) no aumenta mucho el ratio de positivos falsos y aumenta más el ratio de positivos reales, quiere decir que este sistema ha hecho bien su trabajo, y la mayor parte de las recomendaciones hechas, aunque sean bastantes, es correcta.