ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

DESARROLLO DE UN MODELO DE RECOMENDACIÓN GRUPAL QUE CONSIDERE LA PRIVACIDAD DE LOS USUARIOS USANDO DATOS SINTÉTICOS Y AGREGACIÓN DE PREFERENCIAS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN SISTEMAS DE INFORMACIÓN MENCIÓN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

YÉPEZ CASTILLO CAROLINA ELIZABETH
carolina.yepez@epn.edu.ec

Directora: RECALDE CERDA LORENA KATHERINE PhD.
lorena.recalde@epn.edu.ec

Codirector: LOZA AGUIRRE EDISON FERNANDO PhD.
edison.loza@epn.edu.ec

QUITO, junio 2022

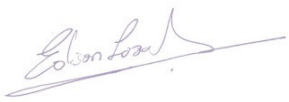
CERTIFICACIÓN DE LA DIRECTORA

Certifico que el presente trabajo fue desarrollado por Carolina Elizabeth Yépez Castillo, bajo mi supervisión.

Lorena Katherine Recalde Cerda, PhD.
DIRECTORA DE PROYECTO

CERTIFICACIÓN DEL CODIRECTOR

Certifico que el presente trabajo fue desarrollado por Carolina Elizabeth Yépez Castillo, bajo mi supervisión.



Edison Fernando Loza Aguirre, PhD
CODIRECTOR DE PROYECTO

DECLARACIÓN

Yo, Carolina Elizabeth Yépez Castillo, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Carolina Elizabeth Yépez Castillo

DEDICATORIA

Este trabajo se lo dedico a mis padres quienes han sido el pilar de mi vida y siempre han estado conmigo en cada etapa apoyándome, dándome ánimos y siendo mi guía.

También se lo dedico a mi esposo por su paciencia y motivación para terminar este trabajo.

Finalmente, se lo dedico a mis hijos Zoe y Snoopy, quienes estuvieron conmigo todas las noches de desvelo hasta culminar el proyecto.

AGRADECIMIENTOS

Doy un profundo agradecimiento a mi Directora de tesis, la Dra. Lorena Recalde por su guía, apoyo y enseñanza en todo este proceso. Le agradezco por darme el tema de tesis y por responder cada una de mis dudas en el camino.

Agradezco también al Dr. Edison Loza, el codirector de mi tesis, quien también me dio la confianza y la guía para poder culminar mi proyecto. Le agradezco por estar al pendiente y por motivarme en la última etapa.

Agradezco al Dr. Ludovico Boratto quien nos ha guiado con sus conocimientos para el desarrollo y finalización del proyecto y la entrega del paper.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

RESUMEN

En este trabajo se propone realizar un modelo de recomendación grupal y se demuestra que usando datos sintéticos se mantiene la protección de los datos en sistemas de recomendación para grupos. Para ello se usaron 2 enfoques para la generación de los datos sintéticos, y así poder ocultar la información privada de los usuarios del grupo. El primer enfoque que fue usado es el de «privacidad diferencial» el cual usa un modelo de red bayesiana que combina distribuciones de baja dimensión para aproximarse a la distribución de dimensión completa de un conjunto de datos. El segundo enfoque utilizado fue el método de «CART» el cual usa modelos secuenciales y aplica una transformación a los datos originales de tal forma que se cambian solo algunos valores. Los resultados con los datos sintéticos de ambos enfoques son bastante prometedores y se comprueba que manipulan información menos específica acerca de las preferencias de los usuarios en el grupo. Nuestro trabajo también muestra que existe viabilidad para poder garantizar privacidad de los datos sin una pérdida significativa de precisión en la recomendación grupal.

Keywords: Sistemas de recomendación grupal, datos sintéticos, privacidad diferencial, CART.

ABSTRACT

In this work, we propose a group recommendation model is proposed it is shown that using synthetic data, data protection is maintained. For this, two approaches were used to generate synthetic data and thus be able to hide the private information of the group's users. The first one is «differential privacy» which uses a Bayesian network model that combines low-dimensional distributions to approximate the full-dimensional distribution of a data set. The second approach used was the «CART» method which uses sequential models and applies a transformation to the original data in such a way that only some values are changed. The results with the synthetic data of both approaches are quite promising and it is verified that they manipulate less specific information about the preferences of the users in the group. Our work also shows that it is possible to guarantee data privacy without a significant loss of accuracy in group recommendation.

Keywords: Group recommender systems, synthetic data, differential privacy, CART.

1 INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

Con la popularidad y uso creciente del contenido digital por parte de los usuarios, el uso de sistemas de recomendación (SR) se ha convertido en una pieza clave para aquellas empresas que deseen explotar los datos relacionados con el comportamiento de sus consumidores [1]. Por ejemplo, algunos SR guían para determinar qué películas son las favoritas de un usuario, qué productos éste está comprando, o cuáles son sus intereses a partir de búsquedas recientes. Con todos estos datos y modelos de preferencias, se pueden generar recomendaciones personalizadas para cada uno de los usuarios o para grupos de ellos. Para generar las recomendaciones, un SR analiza y procesa:

- información histórica de los usuarios (compras previas, calificaciones, elementos cliqueados o visitados, etc.),
- sus datos demográficos, y/o
- características de los productos o de los contenidos (marcas, modelos, precios, contenidos similares).

Luego, transforma estos datos en conocimiento accionable, es decir, predice qué producto puede ser interesante para el usuario y se lo presenta como una recomendación personalizada. Para ello, el SR debe tener acceso a una amplia información de los usuarios, por ejemplo: edad, estado marital, información financiera, o cualquier otro tipo de información personal. Debido a esta necesidad de contar con información privada de los usuarios, se corre el riesgo de que esta información sea fácilmente descifrada. Quebrantar la privacidad de un usuario implicaría su identificación precisa lo cual engendra un problema de privacidad.

Debido a la posible divulgación de esta información, las empresas dudan en proporcionar estos datos a consultoras externas [2]. Como una medida de mitigación a este problema, se pueden generar datos artificiales, o sintéticos, para un control de divulgación estadística, y la creación de conjuntos de datos privados [3], cuyo propósito es mantener la privacidad.

Con la generación de estos datos sintéticos se puede reducir el riesgo de divulgación de información sobre los usuarios [4].

Uno de los tantos escenarios de aplicación de los SR corresponde a los sistemas de recomendación grupales (SRG), en los cuales se usan las preferencias individuales de los miembros del grupo y mediante funciones de agregación se obtienen las recomendaciones grupales. En ocasiones se requiere cierto grado de privacidad entre los miembros de un mismo grupo, de tal forma que no se muestren los intereses o preferenciales individuales de cada usuario.

En [1], los autores sostienen que cuando se presentan recomendaciones a un grupo de usuarios, se puede revelar información privada que de otro modo se desearía ocultar; y muestran que, para cualquier SRG, las representaciones del usuario siempre tendrán que revelar la información suficiente para generar unas buenas predicciones.

Algunos autores han realizado experimentos con varios algoritmos de SR que mezclan datos originales con conjuntos de datos parcialmente sintéticos [5], y han logrado comprobar que el rendimiento relativo de los algoritmos de SR con los datos parcialmente sintéticos refleja el rendimiento de los datos originales. Sin embargo, la aplicación de datos sintéticos y de análisis del rendimiento de los algoritmos en SR no se lo ha aplicado a grupos, es por esa razón que en este proyecto se desea cuantificar en qué grado se afecta el rendimiento de las recomendaciones de un SRG cuando se introducen datos parcialmente sintéticos.

En este proyecto se propone desarrollar un modelo de recomendación grupal que se base en la aplicación de métodos para la generación de ratings sintéticos, de tal forma que se conserve la privacidad de los miembros del grupo. Para la evaluación de este modelo se considerará su comportamiento al ser implementado usando varias estrategias de agregación de preferencias y sobre dos conjuntos de datos distintos.

1.2 OBJETIVOS

1.2.1 Objetivo General

Desarrollar un sistema de recomendación grupal con datos sintéticos para evitar la fuga de información acerca de las preferencias de los usuarios del grupo.

1.2.2 Objetivos Específicos

- ❒ Realizar una revisión de literatura sobre trabajos enfocados en privacidad en SR y SRG.
- ❒ Recolectar dos datasets que contengan información sobre usuarios, ratings e ítems, para la detección de grupos con el fin de implementar un SRG.
- ❒ Verificar el comportamiento del SRG con el uso de datos originales versus el comportamiento del SRG con el uso de datos sintéticos.
- ❒ Crear el modelo de recomendación grupal combinando el mejor método de conservación de privacidad con varias estrategias de agregación de preferencias.
- ❒ Evaluar el rendimiento relativo del SRG considerando métricas de error (RMSE).

1.3 ALCANCE

En este proyecto se desea superar el problema de divulgación de los datos de los usuarios al implementarse un SRG. En la actualidad han sido evaluadas e implementadas ciertas técnicas en SR (para individuos) más no para SRG (para grupos), por tal razón, en este trabajo se propone una solución que abarca desde la elección del conjunto de datos, su preparación y transformación, creación de grupos y finalmente el análisis de los resultados obtenidos en datos originales y en datos sintéticos.

Los conjuntos de datos sobre los cuales se va a trabajar en este proyecto son: Movielens y Goodbooks. Estos conjuntos están a nivel de usuario, es decir que en cada registro se presenta una recomendación (rating) para un ítem (película o libro) dado por un usuario específico.

Posterior a la elección de los datos, se procederá a realizar la predicción de ratings faltantes, es decir, se usará algoritmos de recomendación para generar la predicción de los ratings a un ítem (película o libro) que no ha sido visto por un usuario dado. Las técnicas a ser usadas son: KNN, MeansKNN, SlopeOne y CoClustering.

Luego se procederá a hacer la detección de grupos, usando como mecanismo la similaridad entre las recomendaciones de los usuarios con el coeficiente de Pearson.

Una vez ya obtenidos los grupos, y conociendo qué usuarios pertenecen a cada grupo, se aplicarán funciones de agregación para obtener ratings grupales. Las estrategias de agregación a ser usadas son: Additive Utilitarian (ADD), Least Misery (LMS), Most Pleasure

(MPS) y Multiplicative (MULT). Estas cuatro funciones de agregación son las más utilizadas y conocidas dentro del ámbito de los SRG [6].

Para realizar la sintetización de datos y poder tener menos divulgación de información de los usuarios y de sus grupos, se usaron dos técnicas distintas para evaluarlas, implementarlas y tomar una mejor decisión. Las técnicas a ser usadas para sintetizar los datos son: PrivBayes y CART.

Finalmente, se procederá a hacer un análisis del rendimiento absoluto y relativo de las recomendaciones grupales, poniéndole un mayor enfoque en el rendimiento relativo para dar conclusiones. En este paso se analiza y evalúa qué algoritmo de recomendación presenta mejor rendimiento, qué estrategia de agregación modela de mejor manera los ratings grupales y qué técnica de sintetización se adapta más al conjunto de datos original. Con todo este análisis se propondrá un modelo que se compone de todos estos pasos con el mejor rendimiento relativo para cada conjunto de datos.

No se plantea la creación de un modelo en términos de analítica predictiva, sino más bien ese modelo es aquel que abarca una serie de pasos que mejor se ajustan para SRG en particular. Tampoco se contempla la implementación de una interfaz con el usuario para la interacción con el modelo.

1.4 MARCO TEÓRICO

1.4.1 Sistema de recomendación

Un SR es un sistema que genera sugerencias de artículos (ítems) que pueden ser de interés para un determinado usuario [7], por ejemplo: un libro para leer, una canción para escuchar, un destino turístico para ir, una película para ver, etc. Los SR proporcionan predicciones de ítems que el usuario puede encontrar interesantes y que aún no los ha consumido [8], por tal razón estos SR han demostrado ser útiles para las empresas que busquen impulsar ventas; así como para los usuarios, ya que, facilitan su proceso de toma de decisiones [9].

Uno de los principales elementos de un SR, es la cantidad de información de los usuarios que requieren dado que mientras más interacciones previas entre usuarios e ítems se consideran, la calidad de las recomendaciones mejora [10] [11].

1.4.2 Conceptos y notación

En esta sección se tratarán de expresar los conceptos básicos y las notaciones usadas en algoritmos de recomendación.

En los SR un *usuario* representa una lista de personas que pertenecen a un conjunto de datos, se lo puede representar como $U = u_1, u_2, ..., u_n$.

Un *ítem* representa un elemento que el usuario puede recomendar, por ejemplo, un usuario puede calificar a un conjunto de películas. Los ítems se los puede representar como $I = i_1, i_2, ..., i_m$.

Las *calificaciones,* o también conocidas como *ratings*, son aquellas puntuaciones que el usuario u_i puede dar al item i_j. Los ratings se los puede representar como r_{ui}.

1.4.3 Tipos de sistema de recomendación

1.4.3.1 Sistemas de recomendación basados en filtrado colaborativo

Este tipo de SR utiliza los datos de ciertos usuarios para determinar las preferencias de otro individuo. El principal objetivo de este método de filtrado colaborativo es hacer uso de la matriz User-Item-Rating[1] para calcular los ratings de ítems de los cuales los usuarios no han dado una valoración, es decir, cuyos valores de ratings no están disponibles en la matriz.

Esta matriz se obtiene directamente del conjunto de datos, en donde aparecen los usuarios y los ítems que han sido calificados por cada uno de ellos [13]. A continuación, en la Tabla 1.1 se muestra un ejemplo de la matriz User-Item-Rating:

En la Tabla 1.1 se puede observar que hay valores vacíos, los cuales son aquellos ítems que los usuarios no han calificado, por lo tanto, con el método de filtrado colaborativo se trata de predecir aquellos ratings faltantes basándonos en la similitud de preferencias entre usuarios o ítems [14].

Filtrado colaborativo basado en usuarios

Este tipo de algoritmos predice los ratings de acuerdo con la calificación de los usuarios

[1] User-Item-Rating: En esta matriz cada fila representa a un usuario, cada columna representa un elemento (ítem) y cada celda representa la calificación (rating) otorgada por un usuario a un elemento [12].

Tabla 1.1: Matriz User-Item-Rating

Usuarios	Items		
	i_1	i_2	i_3
U_1	2	1	4
U_2	4		5
U_3	3	5	1
U_4		1	4
U_5			5
U_6	1	3	5

y su similaridad, es decir, se identifica a usuarios que compartan intereses con el usuario actual y se usa sus calificaciones para generar las predicciones [13].

La idea del filtrado colaborativo basado en usuarios es la siguiente: para un usuario u y un ítem i que no tiene calificación, se debe generar un conjunto de usuarios que sean similares a u y que previamente ya hayan calificado al ítem i. Una vez que ya se tiene el conjunto de vecinos más cercanos a u, se puede predecir la calificación r_{ui} tomando un promedio ponderado de los ratings de usuarios similares [15].

Filtrado colaborativo basado en ítems

Este algoritmo está estrechamente relacionado con el filtrado colaborativo basado en usuarios [16], debido a que en este caso se observan ítems similares para el usuario u dado.

Para aplicar el filtrado colaborativo basado en ítems, en lugar de analizar la similaridad entre usuarios, se analiza la similitud entre ítems. Es decir, para un usuario u y un ítem i que no tiene un rating, se genera un conjunto de ítems que son similares a i y que previamente ya hayan sido calificados por el usuario u. Con el conjunto de vecinos más cercanos a i, se procede a predecir la calificación r_{ui} tomando un promedio ponderado de las calificaciones del usuario u en los ítems similares a i [15].

1.4.3.2 Sistemas de recomendación basados en contenido

Este tipo de SR no usan solamente las calificaciones que el usuario ha dado para ciertos ítems, sino que también usan los parámetros de información del ítem (producto) o del perfil del usuario [17].

Estos métodos son bastante usados para generar una recomendación de ítems nuevos [13], ya que, utilizan los atributos descriptivos de los ítems y así generan recomendaciones basadas en la similitud entre ítems.

1.4.3.3 Sistemas de recomendación híbridos

Diversos SR también usan enfoques híbridos de tal forma que combinan métodos de filtrado colaborativo y métodos basados en contenido. Este tipo de SR son usados para contrarrestar las limitaciones de los sistemas antes mencionados.

1.4.4 Sistema de recomendación grupal

En la actualidad, los SR que se usan en varios dominios solo se centran en la interacción de un solo usuario. Sin embargo, existen diversas aplicaciones en las que un usuario interactúa socialmente con otros usuarios [18], por ejemplo: eventos con compañeros del trabajo, cenas con familiares o amigos, o cuando varios usuarios usan el mismo transporte público. Estas situaciones obligan a los usuarios a realizar una participación en grupos [18].

Por tal razón es importante analizar cómo funcionan las recomendaciones en el contexto del grupo. Un SRG usa las preferencias de un usuario y también las preferencias de los demás usuarios del grupo para generar una recomendación a nivel grupal [19].

Para generar una recomendación grupal hay que basarse en 2 pasos [19]:

1. Usar un algoritmo de filtrado colaborativo, para poder tener predicciones de ratings de usuarios individuales para los ítems de cada conjunto de datos.
2. Usar un método de agregación de preferencias, el cual integre las predicciones de los ratings de los usuarios calculadas en el paso anterior, y devuelva un valor de recomendación para cada grupo.

1.4.5 Predicción de Ratings

Se pueden usar varios métodos de filtrado colaborativo para generar la predicción de calificaciones que faltan en la matriz User-Item-Rating. Entre ellos, los que principalmente se usan son los siguientes:

- ❒ KNN básico (KNN): es un método de clasificación no paramétrico, el cual se basa en encontrar los k vecinos más cercanos [20]. Una vez que se tiene la lista de los k vecinos más cercanos, se procede a combinar con los ratings para generar aquellos ítems más relevantes para el usuario actual [21].

- ❒ KNN con medias (MeansKNN): también es un método de clasificación no paramétrico bastante parecido al KNN. Sin embargo, este método, luego de generar los k vecinos más cercanos considera las calificaciones promedio de cada usuario (o ítem).

- ❒ Slope One (SlopeOne): es un método que usa un principio intuitivo de «diferencial de popularidad» [22], entre los ítems para los usuarios. En otras palabras, este método determina cuanto más le gusta un ítem a un usuario que a otro, en forma de pares. Al calcular esta diferencia se puede generar la predicción del rating de un ítem para un usuario u_1 viendo a otro usuario u_2 que ha calificado previamente uno de esos ítems.

- ❒ Co-Clustering (CoClustering): Esta técnica se basa en obtener simultáneamente los k vecinos más cercanos de usuarios y los l vecinos más cercanos de ítems [23] usando co-clustering. Con estas agrupaciones, se generan predicciones de los ratings basados en las calificaciones promedio de los co-clusters.

1.4.6 Modelado para recomendación grupal

Las recomendaciones grupales en SRG se las genera mediante un proceso de tres pasos [19]:

1. Se ejecuta un algoritmo de filtrado colaborativo, para obtener las predicciones de los ratings de usuarios individuales.
2. Se crean grupos de distintos tamaños para poder hacer las evaluaciones pertinentes.
3. Se usan métodos de agregación (funciones de agregación), para poder adaptar al grupo las preferencias individuales.
4. En torno al último paso, existen varias estrategias para agregar las preferencias individuales de los usuarios a una calificación grupal, entre ellas están [24]:

   Estrategias basadas en la mayoría: Son aquellos métodos de agregación que utilizan los ítems más populares. Por ejemplo:

   - ❒ La estrategia de *Plurality Voting*, en la cual el ganador es el ítem que tiene mayor número de votos.
   - ❒ La estrategia de *Borda Count*, en la que el ganador es el ítem que tiene la mejor calificación derivada de la clasificación de ítems.

Estrategias basadas en el consenso: Son aquellos métodos de agregación que consideran las preferencias de todos los miembros del grupo [25]. Por ejemplo:

- En la estrategia *Additive Utilitarian* el ganador es el ítem con la suma máxima de todas las calificaciones individuales de los miembros del grupo.
- En la estrategia *Multiplicative* el ganador es el ítem con el producto máximo de las calificaciones individuales de los miembros del grupo.

Estrategias límite: Son mecanismos de agregación que consideran solamente a un subconjunto de las preferencias de los miembros del grupo [26]. Por ejemplo [25]:

- En la estrategia *Least Misery* el ganador es el ítem con la calificación más alta de todas las calificaciones más bajas que se han dado al ítem.
- En la estrategia *Most Pleasure* el ganador es el ítem con la calificación más alta de todas las calificaciones individuales de los miembros del grupo.

1.4.7 Sintetización de datos

En [1], los autores sostienen que cuando se presentan recomendaciones a los miembros de un grupo, se puede revelar información privada que de otro modo se desearía ocultar. En este trabajo, para poder sintetizar los ratings grupales se usó las técnicas de CART y PrivBayes que ya han sido previamente técnicas usadas para sintetizar ratings individuales.

1.4.7.1 CART

Este algoritmo aplica una técnica de Machine learning (CART) para crear un conjunto de datos parcialmente sintetizados [27], que ocurre en 3 pasos [5]:

1. Se designa ratings en el conjunto de datos, los cuales serán retenidos en los datos sintéticos y se usan como datos de entrenamiento.
2. Se entrena un árbol de decisión, separando los ratings de los usuarios del grupo, optimizando el índice de Gini para poder minimizar la heterogeneidad de los valores dentro de los grupos.
3. El árbol clasifica a cada par de usuario-ítem o grupo-ítem en una hoja, de tal forma que el valor generado se extrae de los valores de esa hoja mediante el uso de bootstrap bayesiano [28].

1.4.7.2 Privacidad Diferencial - PrivBayes

Para obtener los datos sintéticos mediante el método de PrivBayes se siguen los siguientes pasos [29]:

1. Aprendizaje en red: Se calcula una red bayesiana diferencialmente privada tal que se aproxima a la distribución de dimensión completa a través del mecanismo exponencial (EM).
2. Aprendizaje de distribución: Se calculan las distribuciones diferencialmente privadas de los datos en los subespacios de la red bayesiana, usando la técnica de Laplace.
3. Datos Sintéticos: Se generan datos sintéticos a partir de la red bayesiana que está diferencialmente privada, sin considerar de forma explícita a la distribución de dimensión completa.

1.4.8 Background y trabajos relacionados

En esta sección revisamos los antecedentes necesarios para poder entender sobre los algoritmos de recomendación grupal, y de paso revisar trabajos relacionados que han usado la generación datos sintéticos con el fin de proteger la información confidencial en la matriz de User-Item-Rating.

Hay que considerar que el enfoque de este proyecto es generar datos sintéticos en las recomendaciones de un SRG, sin embargo, esto no implica la protección a nivel general del SR, sino que más bien implica la protección de los datos en el caso de que la información de las preferencias de los usuarios se haga pública o tal vez sea utilizada por la competencia (adversarios) con otros fines. La idea principal es observar si los datos sintetizados siguen siendo útiles para poder hacer comparaciones sobre el rendimiento de los algoritmos de SRG o si se presenta algún grado de afectación en el rendimiento de las recomendaciones.

1.4.8.1 Sistemas de Recomendación Grupal

Hay diversos escenarios en donde se usan los SRG para generar recomendaciones de ítems a un grupo de usuarios [30]. Las aplicaciones en donde se ha hecho uso de los SRG son las siguientes: para turismo [31] o vacaciones [32], para restaurantes [33], para propósitos de entretenimiento [34], para ver películas [35], para escuchar canciones [34], entre otros.

Las recomendaciones grupales son creadas a partir de las preferencias individuales de los

miembros del grupo, por esa razón en la literatura se menciona que recomendar a un grupo de usuarios es mucho más complicado que recomendar a usuarios individuales [36]. Sin embargo, hay muchas situaciones en las cuales se deben tomar decisiones en grupo en lugar de forma individual. Por ejemplo, en [37], los autores mencionaron que, para dar un mejor diagnóstico médico, este proceso debería hacerse en grupos.

En la literatura se mencionan varios métodos para combinar las preferencias de los miembros del grupo. En [3], los autores mencionan que las estrategias más comunes para SRG son Additive Utilitarian, Most Pleasure, y Least Misery. Sin embargo, otros autores [38] han reportado que la agregación de las preferencias individuales de los usuarios del grupo no generan recomendaciones eficientes. Otros autores como [39] han aplicado otras estrategias de agrupación, en donde priorizan a los miembros del grupo, otorgándoles pesos a cada uno de ellos de acuerdo a su contribución. En [40] los autores proponen un modelo de espacio de factor latente para hacer la agrupación.

Nuestro enfoque es poder generar recomendaciones grupales que satisfagan las necesidades de los miembros del grupo, por eso nos centramos en aplicar las estrategias más comunes de SRG (Additive Utilitarian, Most Pleasure, y Least Misery), y luego hacer la sintetización de datos para hacer los análisis respectivos.

1.4.8.2 Privacidad en Sistemas de Recomendación

Como ya se mencionó previamente la información de los usuarios de un SR puede ser descifrada por atacantes o adversarios. Por ese motivo los investigadores de la comunidad de SR han realizado varios estudios en el ámbito de la privacidad.

Una forma para mitigar este riesgo es usar técnicas de control de divulgación estadística [5], en las cuales se usan ciertos métodos para modificar o reducir los datos que son públicos. Estos métodos de protección de datos se clasifican en las siguientes categorías [41]:

1. **Métodos de enmascaramiento:** Generan una versión modificada de los datos originales usando una función de perturbación [42]. Este método fue usado por primera vez en [43], combinando la función de perturbación con la aleatorización en el filtrado colaborativo. Los autores demostraron que con esta técnica, pese al usar los datos perturbados, se podían generar recomendaciones aceptables para los usuarios.
2. **Métodos de privacidad diferencial:** Usan el mecanismo de la transformada de Laplace agregando ruido a la matriz de covarianza, con el objetivo de proteger a los vecinos más cercanos de los datos originales [3]. En [44], los autores introdujeron

ruido en los cálculos de las recomendaciones, en donde intercambiaron precisión por privacidad, y descubrieron que sí se puede proporcionar garantías formales de privacidad usando esta técnica.

3. **Métodos de generación de datos sintéticos:** Consiste en construir un modelo con los datos originales y después generan valores artificiales para este modelo [3]. Los métodos de generación de datos sintéticos se dividen en [27] [45]:

   - ❒ Métodos parcialmente sintéticos
   - ❒ Métodos totalmente sintéticos
   - ❒ Métodos híbridos

En [5], los autores usaron el método de CART como técnica de generación de datos parcialmente sintéticos, y lo aplicaron a dos conjuntos de datos Movielens 100K y GoodBooks-10K. En concreto, ellos mostraron que el rendimiento relativo de un conjunto de algoritmos de SR en los datos sintéticos se refleja en el rendimiento relativo de los datos originales. Con esto descubrieron que es posible ocultar información de los usuarios acerca de sus preferencias, la cual podría ser accesible desde la matriz user-item. A diferencia de nuestro trabajo, ellos no generaron predicciones de ratings para ítems que no han sido calificados en el conjunto de datos original. Además, nosotros hicimos estas mismas pruebas, pero en SRG y también, consideramos un segundo algoritmo para la generación de datos sintéticos que fue PrivBayes. En cuanto a los conjuntos de datos, usamos los mismos datasets (Movielens 100K y GoodBooks-10K) y luego hicimos la creación de grupos para realizar los análisis a nivel grupal.

En [46], se construyó un modelo de recomendación con protección de atributos, el cual da recomendaciones de ítems relevantes y evita los ataques de inferencia de atributos privados. Esta investigación muestra que el modelo propuesto preserva el rendimiento de la recomendación y protege a los usuarios contra posibles ataques de privacidad. En cambio, en nuestro estudio, usando datos sintéticos damos una protección a las preferencias de los usuarios de un grupo y no a sus atributos. Además, nosotros realizamos la evaluación del SR a nivel de grupos, y no a nivel de usuario. En la investigación, ellos usaron el conjunto de datos de Movielens, mientras que nosotros usamos dos datasets de diferente contexto.

El uso de datos sintéticos ha estado presente también en investigaciones sobre la recomendación consciente del contexto. Por ejemplo en [47], los autores crearon un generador de conjuntos de datos sintéticos Java, con el fin de evaluar SR conscientes del contexto.

Esto quiere decir que lo usaron para completar la cantidad de información de contexto (que se caracteriza por las preferencias originales del usuario) o para hacer un recálculo de los ratings de acuerdo con los perfiles de otros usuarios.

1.4.8.3 Privacidad en Sistemas de Recomendación Grupal

Las investigaciones sobre el uso de datos sintéticos en SR con fines de privacidad han sido limitadas [3], y hasta donde nosotros conocemos nunca se lo había explorado en SRG.

En [48], los autores presentan un framework de recomendación en donde preservan la privacidad basada en grupos. Proponen un algoritmo de intercambio de preferencias distribuidas con el fin de garantizar el anonimato de los datos, usando un método de agregación de preferencias llamado Kemeny Ranking para agregar las preferencias de los miembros del grupo, y para generar las recomendaciones personalizadas usan un algoritmo de filtrado colaborativo híbrido basado en random walk. Ellos mostraron que usando este método propuesto obtienen un buen rendimiento de la recomendación y preservan la privacidad. En contraste en nuestro trabajo, generamos datos sintéticos con el método de CART y PrivBayes. Además, usamos 4 funciones de agregación distintas para evaluar la que mejor rendimiento presenta, y finalmente usamos 4 algoritmos de filtrado colaborativo para obtener las predicciones de las recomendaciones faltantes.

Cabe recalcar que en [5] sí hacen una comparación de los SR usando datos sintéticos. Es por eso que basados en esta investigación surgió nuestra idea de ver si esto es aplicable para SRG o no.

1.4.8.4 Generación de Datos Sintéticos

En este proyecto nuestro enfoque es aplicar el método parcialmente sintético de CART y usar el método de privacidad diferencial PrivBayes. Usamos los dos métodos con el fin de hacer la prueba de concepto con dos técnicas de control de divulgación estadística distinta.

El método de síntesis de datos CART ha sido usado previamente en varias aplicaciones de SR, como por ejemplo en [3], [4] y[5].

Así mismo en varias aplicaciones de SR colaborativa se ha usado la privacidad diferencial, como por ejemplo en: [29] y en [44]. Estos autores han demostrado que hay viabilidad para garantizar que al usar esta técnica no hay una pérdida significativa en la precisión de la recomendación.

2 METODOLOGÍA

Para el desarrollo de este proyecto se siguió una aproximación basada en el método de Ciencia del Diseño (Design Science Research, DSR). DSR es una de las estrategias de investigación más utilizadas en la disciplina de los IS (sistemas de información) y de la informática [49]. DSR se basa en el desarrollo de nuevos productos de TI (tecnología de la información), que se los conoce como artefactos [49], y que pueden ser de diferentes tipos [50]:

- ❒ Constructos: Hace referencia a las nociones de entidades, objetos o flujos de datos.
- ❒ Modelos: Son combinaciones de constructos los cuales representan situaciones y se los usa para la comprensión y resolución de problemas.
- ❒ Métodos: Son las metodologías para seguir, sobre los modelos que se van a producir y sobre las etapas del proceso para la resolución de problemas utilizando TI.
- ❒ Instancias: Son consideradas un sistema de trabajo en el cual se demuestra que los constructos, modelos, métodos o ideas, pueden ser implementados en un sistema informático.

El entregable de este proyecto será un modelo y a continuación se detalla paso a paso la aplicación de DSR para su generación.

2.1 DESIGN SCIENCE RESEARCH

Mediante la estrategia de investigación DSR nos vamos a centrar en el desarrollo de un modelo de SRG, en el cual se mantenga la privacidad de la información en la recomendación para los usuarios de grupos.
DSR incluye seis pasos, los cuales son detallados a continuación.

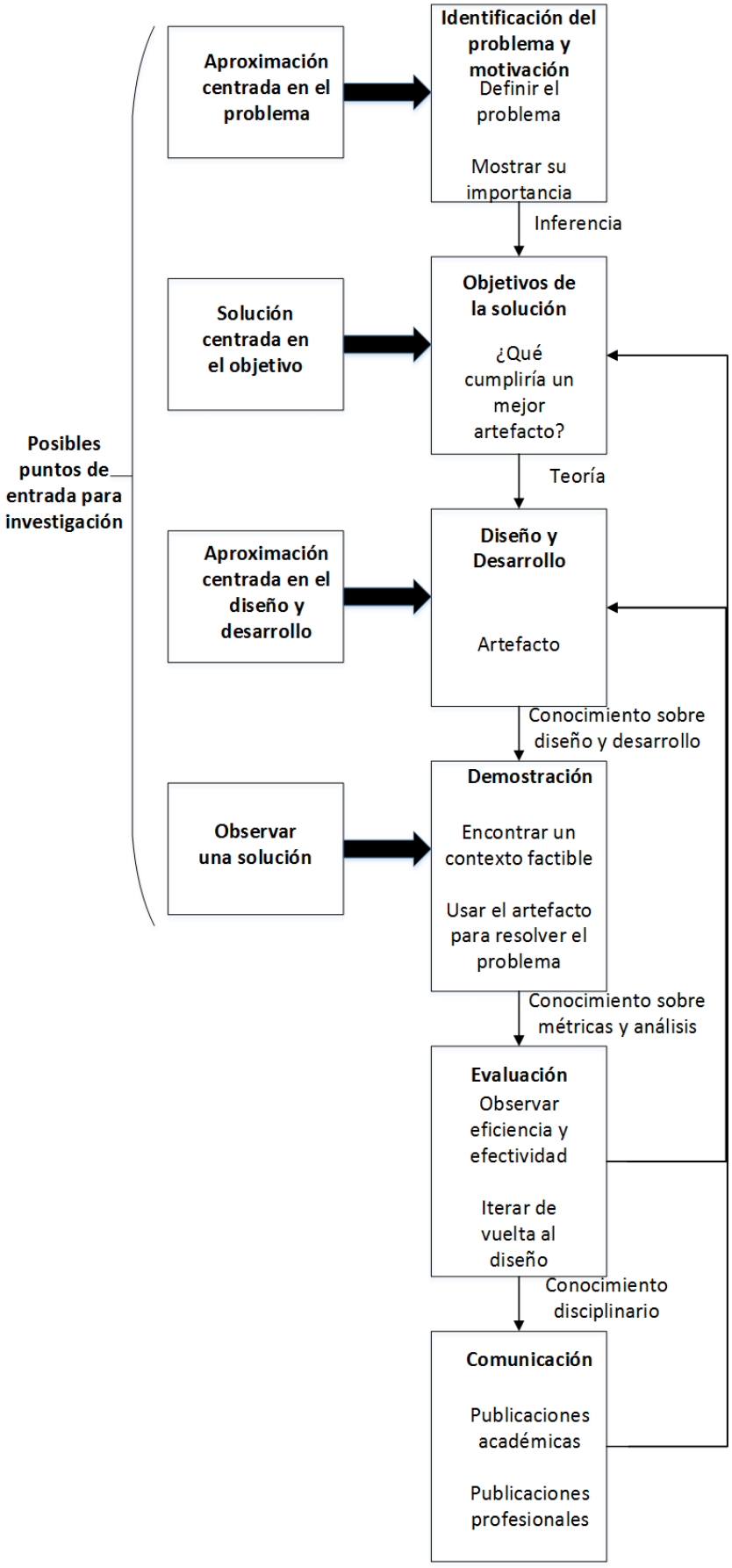


Figura 2.1: Modelo de proceso Design Science Research DSR [51]

1. Identificación y motivación del problema: permite definir el problema y justificar el valor de una solución.
2. Definición de los objetivos de la solución.
3. Diseño y desarrollo: en esta fase se implementa el artefacto en sí, en este caso el modelo de recomendación. Para el desarrollo del modelo se seguirá el proceso establecido por CRISP-DM [52].
4. Demostración: Demostrar el uso del artefacto para resolver el problema. Esto podría implicar su uso en experimentación, simulación, estudio de casos, prueba u otra actividad apropiada. Los recursos necesarios para la demostración incluyen el conocimiento efectivo de cómo utilizar el artefacto para resolver el problema.
5. Evaluación: Observar y medir lo bien que el artefacto soporta una solución al problema. En esta fase, los investigadores pueden decidir si iterar de nuevo al paso 3 para mejorar la eficacia del artefacto o continuar con la comunicación y dejar las mejoras ulteriores para proyectos subsiguientes.
6. Comunicación: Comunicar la solución al problema, su importancia, utilidad, novedad, rigor de diseño y su eficacia para los investigadores y otros públicos pertinentes.

2.2 CRISP-DM

Por otro lado, como se mencionó en la sección precedente, en este proyecto se aplicará el método de CRISP-DM para el desarrollo del modelo de recomendación grupal. CRISP-DM es un método general de minería de datos aplicable en contextos de analítica de datos en general [52] y además permite llevar a cabo las actividades necesarias para proceso de minería [53]. Su proceso se define por las siguientes fases [54]:

1. Comprensión del negocio: en esta fase se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio.
2. Comprensión de los datos: fase que consiste en la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca de la información que podría estar oculta.
3. Preparación de los datos: comprende aquellas actividades de tratamiento de los datos para construir la vista minable o conjunto de datos final sobre el cual se aplicarán las técnicas de minería.

4. Modelado: en esta etapa se aplican los diversos algoritmos de agregación de preferencias para el SRG, generación de datos sintéticos para conservación de la privacidad y modelos de recomendación equitativa (validación y prueba).
5. Evaluación: fase en la que se analizan los resultados obtenidos en función de los objetivos del proyecto. En esta etapa se debería determinar si se ha omitido algún objeto importante del negocio y si el nuevo conocimiento será combinado e implementado.
6. Implementación: realizar el despliegue del modelo obtenido para su uso en SRG.

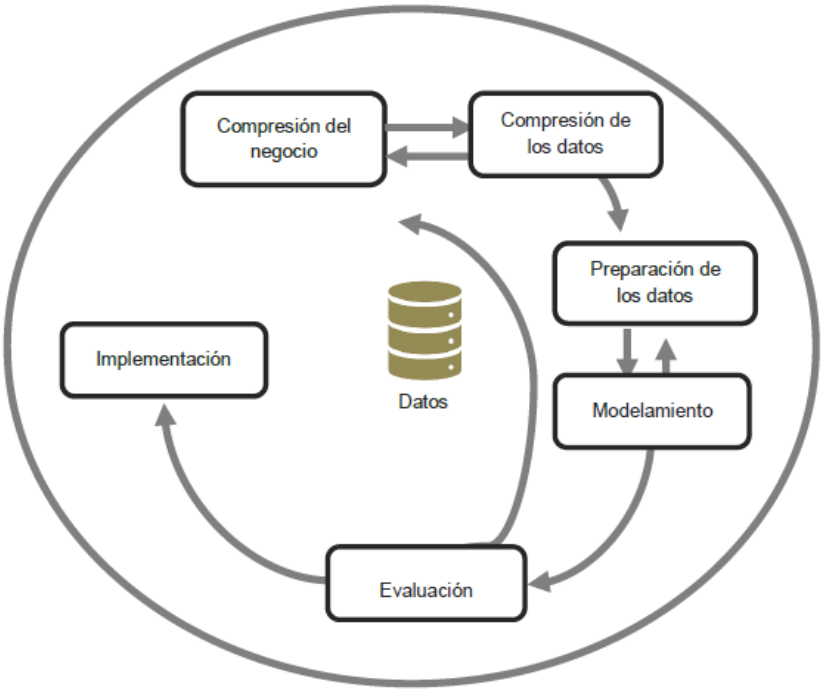


Figura 2.2: Ciclo del modelo CRISP-DM [55]

2.3 COMBINACIÓN DE DSR Y CRISP-DM

En este proyecto se hizo una fusión de DSR con CRISP-DM, ya que ambas tienen aspectos en común, y además porque, por un lado, el entregable del proyecto es un modelo (CRISP-DM) y, por otro lado, se deben responder a los objetivos del proyecto de investigación en sí. Al hacer esta combinación de metodologías se observó que hay fases que son parecidas en las dos, de tal forma que, al combinarlas, los procesos se realizarán una sola vez. A continuación, se presenta la relación existente entre las fases de DSR y CRISP-DM [56].

2.3.1 Fase 1

En esta fase se entiende las necesidades del negocio, se realiza la identificación del problema, motivación del proyecto, y se establecen objetivos.

<table>
<tr><th>Metodología</th><th colspan="7">DSR</th></tr>
<tr><td rowspan="7">CRISP-DM</td><td>Procesos</td><td>Identificar el problema y la motivación</td><td>Definir objetivos para la solución</td><td>Diseño y desarrollo</td><td>Evaluación</td><td>Demostración</td><td>Comunicación</td></tr>
<tr><td>Comprensión del negocio</td><td colspan="2">Fase 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Comprensión de los datos</td><td></td><td></td><td>Fase 2</td><td></td><td></td><td></td></tr>
<tr><td>Preparación de los datos</td><td></td><td></td><td>Fase 3</td><td></td><td></td><td></td></tr>
<tr><td>Modelado</td><td></td><td></td><td>Fase 4</td><td></td><td></td><td></td></tr>
<tr><td>Evaluación</td><td></td><td></td><td></td><td>Fase 5</td><td></td><td></td></tr>
<tr><td>Despliegue</td><td></td><td></td><td></td><td></td><td colspan="2">Fase 6</td></tr>
</table>

Figura 2.3: Relación entre las fases de las metodologías DSR y CRISP-DM

Para identificar el problema se investigó sobre aquellos algoritmos que se usan con mayor frecuencia en el contenido digital. Como se discutió en el Capítulo 1, debido a la popularidad del contenido digitalizado y personalizado se están usando con mayor demanda los algoritmos de SR y SRG. Los algoritmos de SR y SRG usan la información de los usuarios y de sus grupos con el fin de darles recomendaciones precisas, ya sea en el ámbito de películas, de alimentos, de libros, de publicidad de mascotas, etc.

En particular, los SRG usan las preferencias individuales de los usuarios para luego mediante estrategias de agregación obtener una recomendación a nivel de grupos. Estas funciones de agregaciones de preferencias suelen aplicar ciertas fórmulas matemáticas para obtener el valor de la recomendación grupal, por ejemplo, la suma de las preferencias individuales, el valor máximo de las preferencias de los miembros de un grupo, el valor mínimo de las preferencias de los miembros de un grupo, etc.

Sin embargo, a pesar de que esta recomendación esté a nivel de grupos, puede ser fácilmente descifrada, afectando incluso la privacidad de cada uno de los miembros del grupo. Por ello se propone una medida de mitigación para este problema, mediante el uso de datos sintéticos generados a partir de los datos originales bajo el reto de no sacrificar del todo el rendimiento de las predicciones. Para esto se emplearán dos métodos de la literatura, usados para la sintetización parcial de datos.

Además, se analizaron qué datasets están disponibles para el análisis de sistemas de recomendación. En esta fase no se encontraron datasets de grupos, por lo cual se decidió trabajar con datasets que tengan disponibles usuarios, ítems y ratings *individuales* para luego hacer la creación de grupos basado en trabajos previos que usan la similaridad entre usuarios [19] [57], es decir los usuarios que tienen preferencias o gustos parecidos, son

detectados como miembros del mismo grupo.

Con la introducción vista en el Capítulo 1 y con la identificación del problema ya se logró definir cuáles son los objetivos del presente proyecto que se refieren al desarrollo de un SRG con datos sintéticos para evitar la divulgación de información acerca de las preferencias de los usuarios del grupo.
Los objetivos específicos se refieren a buscar al menos dos datasets que contengan los datos de usuarios, ítems y ratings individuales, posterior a ello implementar un SRG con datos originales y con datos sintéticos en donde se observará el comportamiento y el rendimiento relativo de ellos, luego se evaluará el SRG considerando la métrica de error RMSE y finalmente se propondrá un modelo de recomendación grupal que consta de un conjunto de pasos que combinados generaron el menor error posible.

2.3.2 Fase 2

En esta fase se hace una comprensión de los datos, tratando de familiarizarnos con ellos, se realiza una exploración, descripción y gráfica de los datos disponibles.

Se decidió trabajar con dos datasets. El primero de ellos, el de Movielens 100k[1] contiene interacciones de usuarios con películas. La interacción se basa en ratings que los usuarios asignan a las películas una vez que son «consumidas» y que va de una puntuación de 1 (como menor calificación) a 5 (siendo la mayor calificación). El dataset de Movielens 100k está compuesto de 943 usuarios, 1682 películas, y tiene un tamaño de $100,000$ registros o ratings. Se consideró el dataset Movielens 100K por ser bien conocido por la comunidad científica, y además por sus usos previos en SR.

En términos de comparación y evaluación del comportamiento de los métodos usados, también se decidió usar el dataset de Goodbooks[2] el cual tiene $53,000$ usuarios, $10,000$ libros, y tiene un tamaño de $6'000,000$ de registros o ratings, con puntuaciones que van de 1 (como menor calificación) a 5 (siendo la mayor calificación). Se decidió trabajar con el dataset de Goodbooks ya que es un conjunto de datos mucho más grande y además difiere del dominio del otro dataset.

[1] https://grouplens.org/datasets/movielens/

[2] https://www.kaggle.com/philippsp/book-recommender-collaborative-filtering- shiny/data

2.3.3 Fase 3

Esta sección es uno de los aspectos más importantes dentro de la minería de datos [54], ya que en ella se realiza el análisis de los datos, limpieza, agregación, eliminación y también selección de las variables relevantes que van a ser usadas en el modelado.
En este proyecto se realizó lo siguiente:

- ❒ Dataset Movielens 100k:
  1. Se consideró a los usuarios que hayan calificado a más de dos películas.
  2. Se descartó del análisis a la variable $Timestamp$ (fecha y hora de cuándo se dió el rating), ya que no tenía relevancia para nuestro requerimiento.
  3. Se consideró a las variables

     $UserID$: tiene valores que oscilan entre 1 y 943.

     $MovieID$: tiene valores que oscilan entre 1 y 1682.

     $Rating$: tiene una escala de 1 a 5.
- ❒ Dataset Goodbooks
  1. Se consideró a los usuarios que hayan calificado a más de dos libros.
  2. Se consideró a las variables

     $UserID$: tiene valores que oscilan entre 1 y $53,000$.

     $BookID$: tiene valores que oscilan entre 1 y $10,000$.

     $Rating$: tiene una escala que va del 1 a 5.
  3. Se hizo una limpieza de datos, ya que había variables con NA o valores nulos. Por ejemplo, en la variable $Rating$, muchos registros estaban vacíos, por ende, lo que se hizo fue eliminar del análisis a esos registros. También se identificó en la variable de $UserID$ y $BookID$ valores de -1, los cuales también fueron descartados del análisis.

2.3.4 Fase 4

En esta sección se procede a realizar el modelado de los datos de acuerdo con la siguiente estructura.

Predicción de Ratings

Como primer paso se realizó la predicción de los ratings que faltaban, es decir, se buscó completar en la matriz User-Item-Ratings los ratings de los usuarios que no habían calificado a alguna película (Movielens) o libro (Goodbooks). Para ello se procedió a hacer uso de

los algoritmos de recomendación mencionados en la Sección 1.4.5, tanto para el dataset de Movielens como para el dataset de Goodbooks. Estos 4 algoritmos fueron los siguientes:

1. KNN
2. MeansKNN
3. SlopeOne
4. CoClustering

Una vez que se obtuvo la matriz completa con los ratings obtenidos de las predicciones y los ratings dados por los usuarios, se procedió a realizar la creación de grupos.

Detección de grupos

Dado que no se encontraron datasets que tengan ratings grupales, se decidió como segundo paso crear grupos en los datasets encontrados para ejecutar nuestros experimentos. Además, como queremos probar si es más difícil mantener la privacidad de los usuarios para grupos según el tamaño de los mismos, decidimos crear grupos de tamaños desde $n = 2$ hasta $n = 8$, para los dos conjuntos de datos mencionados anteriormente. Creamos tipos de grupos que se basan en la *similaridad*, siguiendo el trabajo de los autores en [58]. Bajo esta premisa, en la similaridad se eligen los miembros del grupo basados en aquellos que tienen preferencias similares, usando el cálculo de similitud entre pares de usuarios en base al *Coeficiente de Correlación de Pearson* [57].

Estrategias de Agregación

Después de haber realizado la detección de grupos, se procedió a hacer el cálculo de la recomendación grupal agregando las preferencias individuales por usuario [59]. En los SRG, el valor de la recomendación va a depender del tipo de estrategia de agregación que se use [60].

En este proyecto se usaron las siguientes cuatro funciones de agregación para SRG:

1. Additive Utilitarian (ADD)
2. Least Misery (LMS)
3. Most Pleasure (MPS)
4. Multiplicative (MULT)

Luego de que se obtuvieron las recomendaciones grupales con cada estrategia de agregación para los dos conjuntos de datos, se procedió a realizar el último paso que es la generación de datos sintéticos.

Generación de datos sintéticos

Se realizó la síntesis de las recomendaciones grupales por dos métodos distintos, el primero es el de *CART* [5], el cual se basa en entrenar un árbol de decisión, separando los ratings de los usuarios del grupo, optimizando el índice de Gini. El segundo método es el de *differential privacy-PrivBayes* que usa redes bayesianas para generar los datos sintéticos basándonos en [29].

2.3.5 Fase 5

En esta sección es donde se evalúan los distintos modelos que se han aplicado, se observan las métricas de error como es el RMSE.

Se compararon los resultados obtenidos con los distintos tamaños de grupos, los distintos algoritmos de recomendación, las distintas funciones de agregación y con las dos formas de sintetización de datos.

Con esta evaluación de resultados se tomó una decisión sobre el modelo final seleccionado para cada conjunto de datos. Esta decisión se la tomó considerando el rendimiento relativo de los SRG y la relevancia de los resultados obtenidos.

2.3.6 Fase 6

En esta última fase del proceso de CRISP-DM, se realizó el despliegue y el informe final de los resultados obtenidos. En esta sección es importante mencionar qué modelo funciona mejor con cada dataset, con cuál se presenta menor error y difundir esta información en el proyecto.

Se ha demostrado que, siguiendo una serie de pasos, el rendimiento relativo de los SRG es el mismo tanto en los datos originales como en los datos sintéticos.

Sin embargo, existe una variación al aplicar distintos algoritmos de predicción para las recomendaciones individuales, por ejemplo, el algoritmo KNN es el que menor RMSE presenta tanto en los datos originales como en los datos sintéticos.

También existe una variación en el rendimiento con los distintos tamaños de grupos, ya que se observó que a mayor tamaño del grupo, el RMSE se incrementa tanto en los datos originales como en los datos sintéticos.

De igual manera, en las estrategias de agregación, hay una variación en el rendimiento con cada una de las funciones aplicadas, por ejemplo, la estrategia Multiplicative (MULT) es la que tiene mayor RMSE comparadas con las demás.

Finalmente, en este proyecto se analizó la capacidad que poseen los datos sintéticos para ocultar las preferencias de los usuarios individuales y grupales, de tal forma que con esta sintetización de datos se ha logrado conservar todos los aspectos de preferencias de usuarios grupales y se ha probado que son útiles estos datos en algoritmos de recomendación. Además, el aporte de nuestro trabajo es que esto ha sido probado en algoritmos de recomendación grupal.

Se trabajó con grupos de distinto tamaño $n = 2, .., 8$, también con 4 métodos de recomendación, 4 métodos de agregación de preferencias y 2 métodos de sintetización de datos.

3 RESULTADOS

Todos los experimentos fueron probados en dos conjuntos de datos: Movielens y Goodbooks.

Como primer paso se hizo una selección de algoritmos de recomendación y se usaron: KNN, MeansKNN, SlopeOne y CoClustering. Estos algoritmos fueron implementados en *PYTHON* usando la librería *SURPRISE*[1].

Como siguiente paso se procedió a realizar la creación de grupos de tamaño $n = 2, .., 8$. Para realizar este paso usamos la implementación de [58], pero únicamente consideramos la similaridad de los usuarios para la creación de grupos[2]. Hay que tomar en cuenta que, un usuario no puede aparecer más de una vez en un grupo determinado, pero este usuario puede ser miembro de varios grupos.

Luego de que se obtuvieron los grupos, se procedió a hacer la selección de las estrategias de agregación. Para esta implementación se usó las siguientes funciones de agregación: Additive Utilitarian (ADD), Least Misery (LMS), Most Pleasure (MPS) y Multiplicative (MULT).

Finalmente, se procedió a generar los datos sintéticos de los ratings grupales. Esto se lo hizo usando dos métodos distintos: CART[3], para el cual usamos la implementación de [5]; y el de PrivBayes[4], para el cual usamos la implementación de [29].
Al tener 4 distintos algoritmos de recomendación, 7 tamaños de grupos, 4 funciones de agregación y 2 métodos para la sintetización de datos, se obtuvo 224 escenarios posibles en cada conjunto de datos.

Para cada escenario se obtuvo un valor de RMSE con el fin de evaluar si cuando introducimos datos sintéticos en el SRG se ve perjudicado el rendimiento de las predicciones de

[1] http://surpriselib.com/

[2] Este trabajo está disponible al público en: https://github.com/mesutkaya/recsys2020

[3] El código está disponible en: https://github.com/SlokomManel/SynRec

[4] El código está disponible en: https://github.com/SAP/data-synthesis-for-machine-learning

ratings o rendimiento del sistema. La idea principal es poder cuantificar el grado en que se ve afectado el rendimiento relativo de las predicciones en un SRG con la métrica de error RMSE. Este análisis es de gran importancia en el proyecto debido a que con esto se puede llegar a dar conclusiones sobre el uso de los datos originales vs. el uso de los datos sintéticos en SRG.

3.1 VALIDEZ ANALÍTICA

Nuestro interés es poder medir la calidad de los datos sintéticos con respecto a los datos originales, y esto se hizo a través de la validez analítica, es decir, observando las propiedades estadísticas de los datos originales vs. los sintéticos, y viendo el grado de correspondencia entre las propiedades estadísticas globales de cada conjunto de datos.

A continuación, las Figuras 3.1 y 3.2 presentan la distribución de los ratings tanto en los datos originales como en los datos sintéticos usando el método CART:

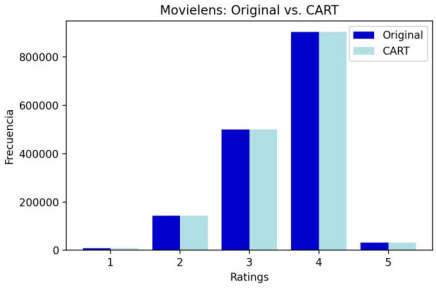


Figura 3.1: Distribución de ratings de usuarios para el conjunto de datos de MovieLens.

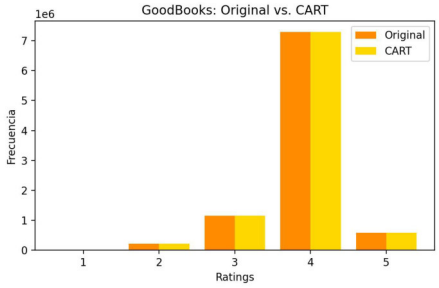


Figura 3.2: Distribución de ratings de usuarios para el conjunto de datos de GoodBooks.

Las Figuras 3.3 y 3.4 detallan la distribución de los ratings tanto en los datos originales como en los datos sintéticos usando el método PrivBayes.

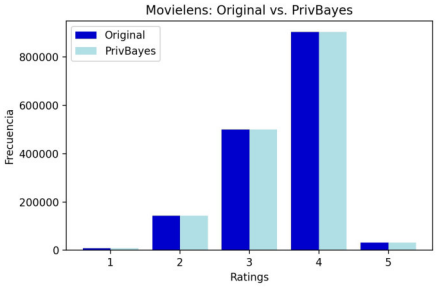


Figura 3.3: Distribución de ratings de usuarios para el conjunto de datos de MovieLens.

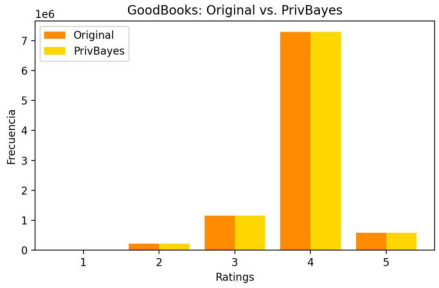


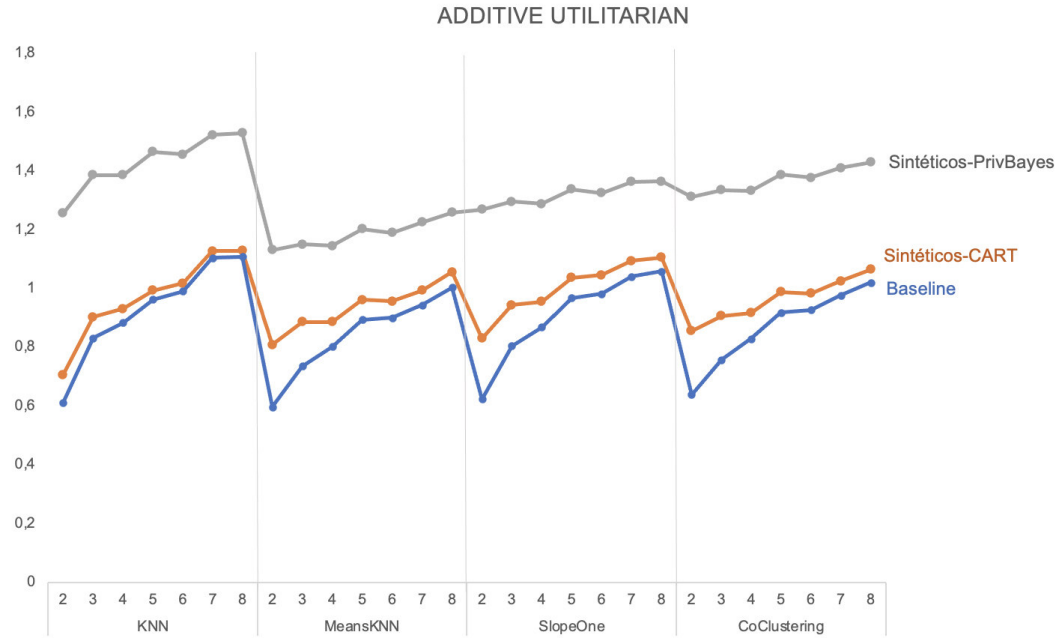
Figura 3.4: Distribución de ratings de usuarios para el conjunto de datos de GoodBooks.

3.2 RENDIMIENTO DE LA RECOMENDACIÓN GRUPAL

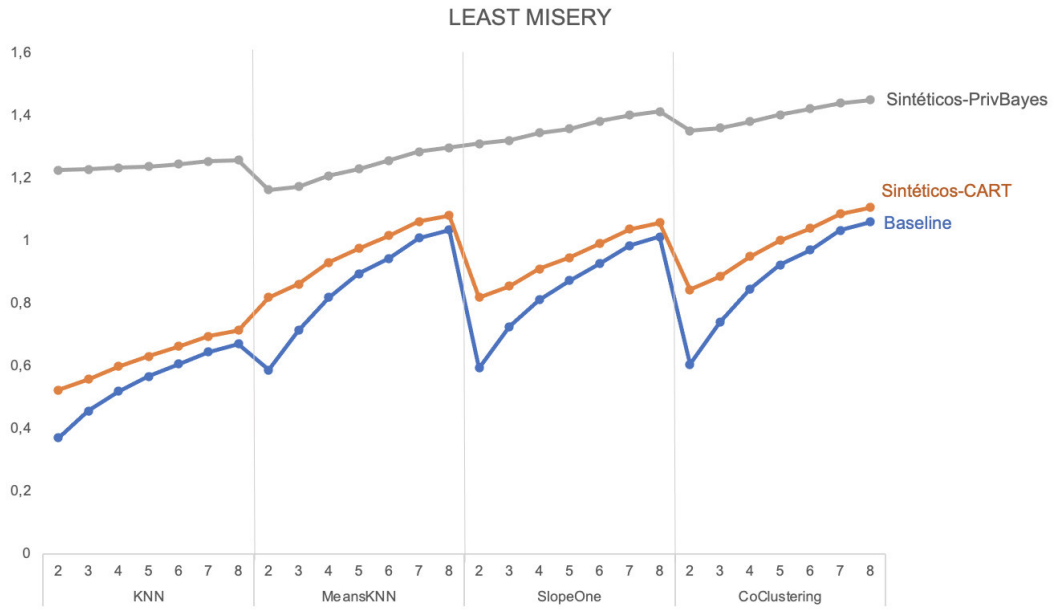
Se realizaron experimentos con distintos algoritmos de recomendación, donde se entrenaron y probaron los conjuntos de datos originales para poder ser comparados con los conjuntos de datos sintéticos generados por PrivBayes y CART respectivamente.

Dentro de estos experimentos se usaron los algoritmos de filtrado colaborativo, que fueron mencionados en la Sección 1.4.5, para obtener las predicciones de las recomendaciones individuales que se encontraban faltantes en los datasets. Posterior a ello, se aplicó cada una de las estrategias de agregación (Sección 1.4.6) que se usan para generar la recomendación grupal.

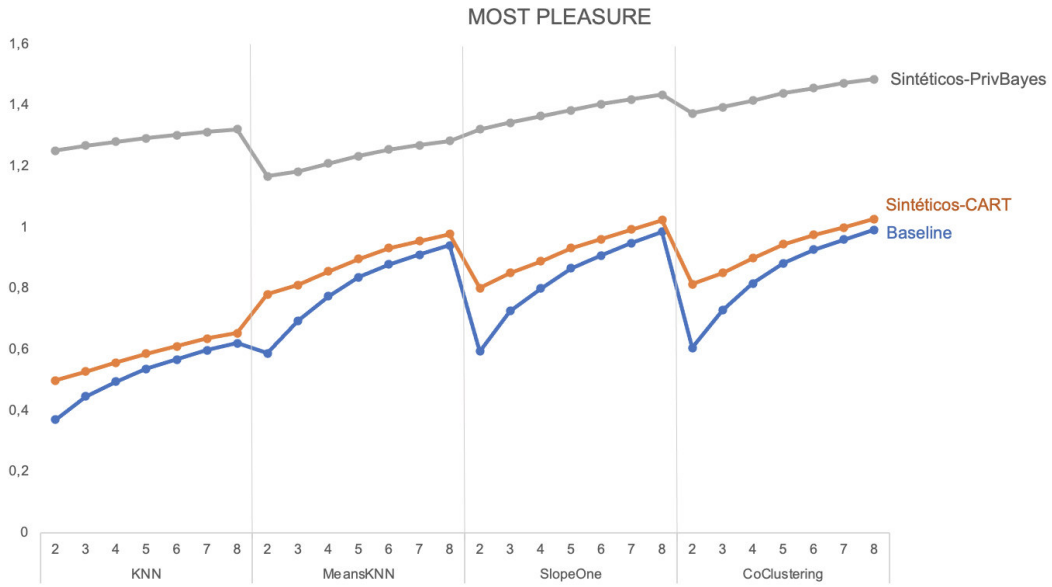
Los resultados de la aplicación de los algoritmos de filtrado colaborativo para el conjunto de datos de Movielens en distintos tamaños de grupo se muestran en la Figura 3.5, separadas por función de agregación.



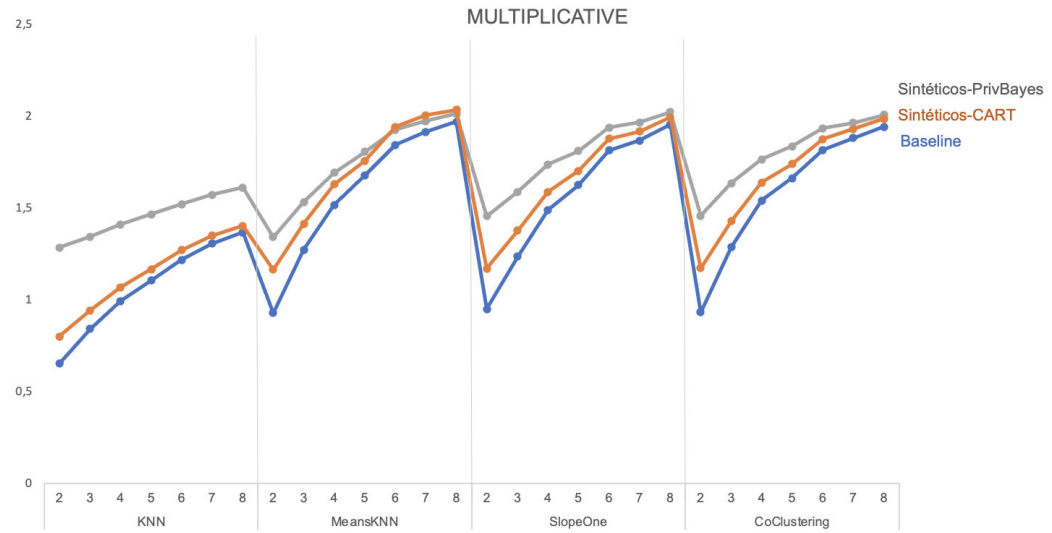
a) Función de agregación Additive Utilitarian



b) Función de agregación Least Misery



c) Función de agregación Most Pleasure

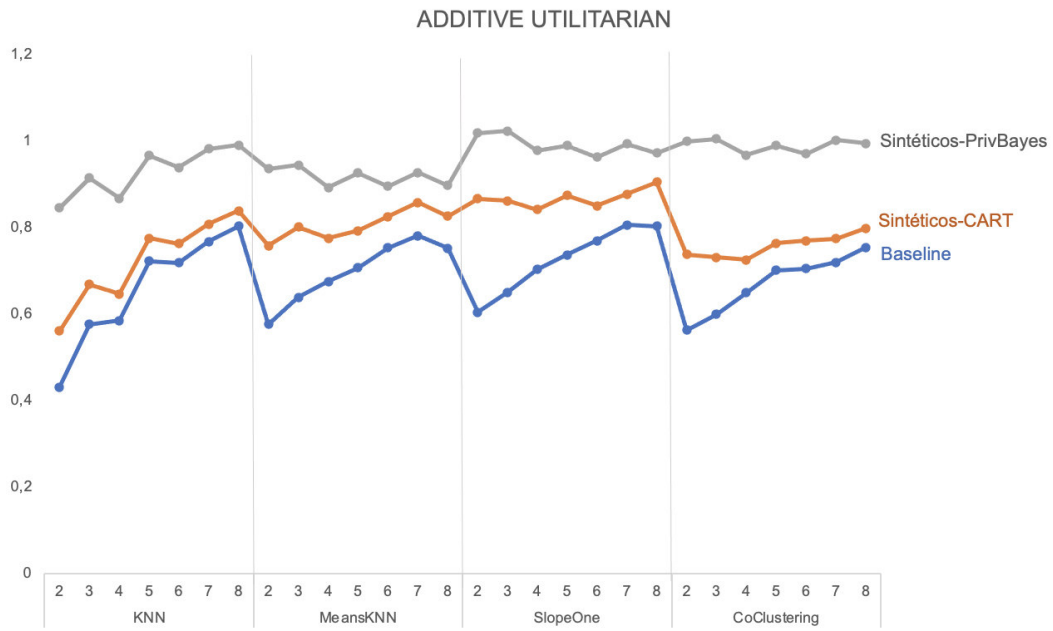


d) Función de agregación Multiplicative

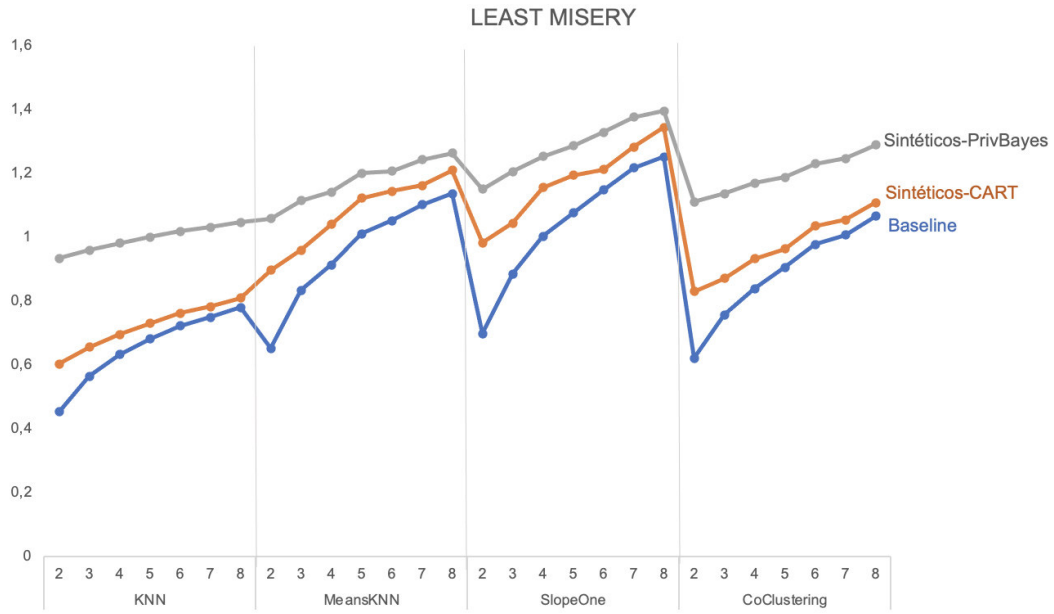
Figura 3.5: RMSE para el conjunto de datos de Movielens

Ahora, para el conjunto de datos de Goodbooks, también se usaron los algoritmos de filtrado colaborativo, para así obtener las predicciones de los ratings individuales faltantes en el dataset. Luego, se aplicó cada una de las estrategias de agregación, para obtener la recomendación grupal.

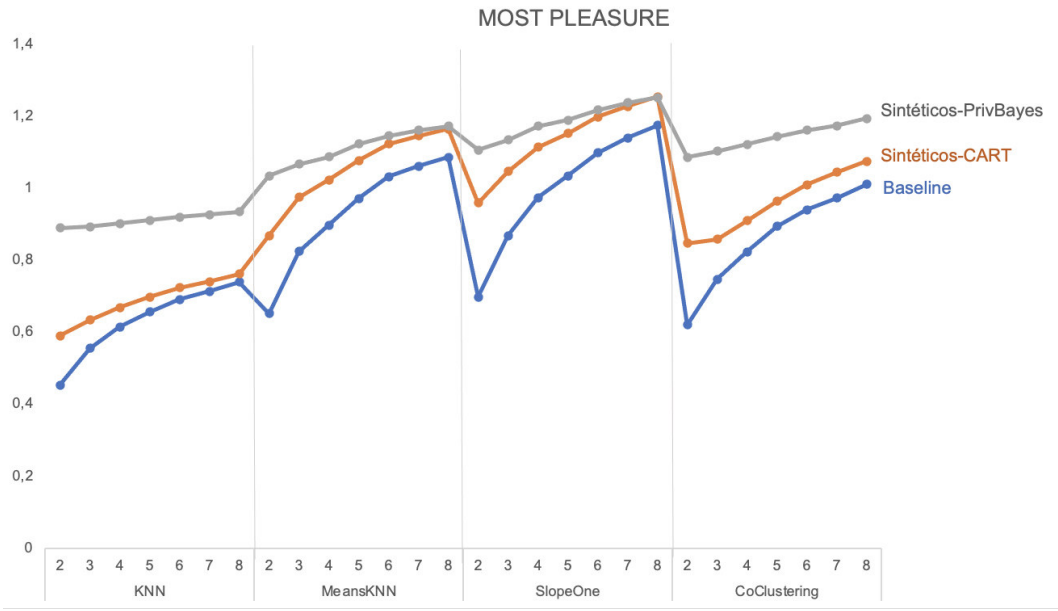
Los resultados de haber aplicado los algoritmos de filtrado colaborativo en los distintos tamaños de grupos se muestran en la Figura 3.6, separados por función de agregación.



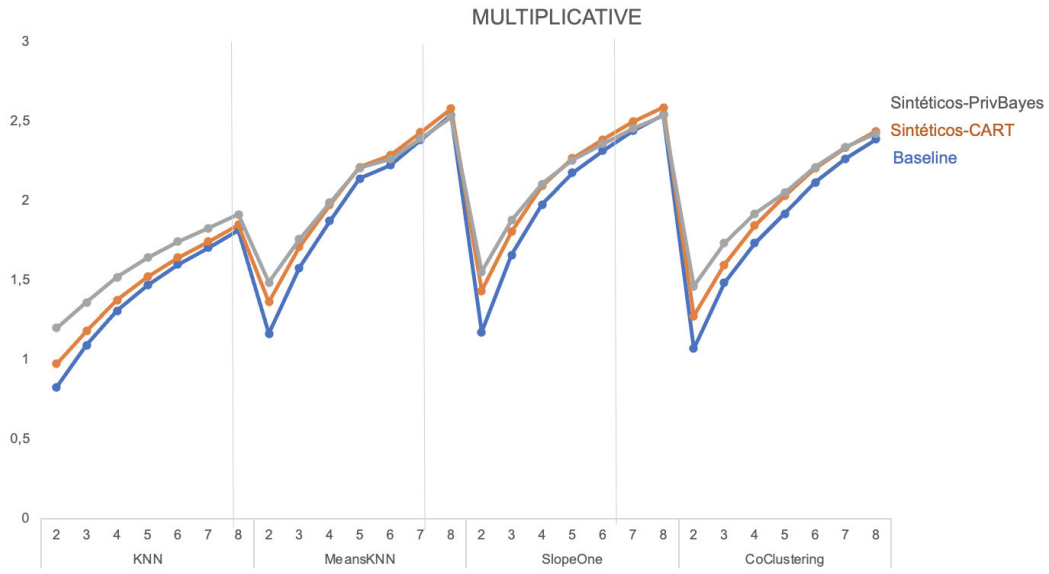
a) Función de agregación Additive Utilitarian



b) Función de agregación Least Misery



c) Función de agregación Most Pleasure



d) Función de agregación Multiplicative

Figura 3.6: RMSE para el conjunto de datos de Goodbooks

Como era de esperarse, el rendimiento de la predicción absoluta, medida en términos de error cuadrático medio (RMSE), es menor en los datos sintéticos (con cualquiera de los dos métodos PrivBayes y CART) que con los datos originales (baseline). Esto es evidente, ya que al trabajar con los datos sintéticos se está ocultando información de preferencias de los usuarios. Por ende, al ocultar y sintetizar los ratings se tendrá un menor rendimiento en los datos sintéticos que en los datos originales que no tienen ninguna sintetización de datos aplicada.

Además, se puede observar que la sintetización de CART presenta un mejor rendimiento absoluto en relación con la sintetización de PrivBayes. Esto se cumple en los dos conjuntos de datos, para los distintos tamaños de grupos, distintas funciones de agregación y para la variedad de algoritmos de recomendación con los cuales se experimentó.

Sin embargo, en este proyecto nos interesa enfocarnos en el rendimiento relativo de los datos originales vs. los sintéticos. Esto debido a que lo que indica el rendimiento relativo es que tan bien se siga ese patrón de error del baseline por parte del método sintético. Así, se debería seguir el mismo patrón de error para poder concluir que los datos sintéticos modelan el error que siguen los datos originales.

Tomando en cuenta lo antes mencionado, podemos decir que el método de sintetización de CART es el que mejor modela el patrón de error de los datos originales. La curva de error se aproxima mucho entre el método de CART y el baseline, más que con el método de PrivBa-

yes. Con eso se puede concluir que usar los datos sintéticos de CART es la mejor opción, porque será la combinación que genere menor error que el resto de las posibilidades.

También se puede evidenciar en las Figuras 3.5 y 3.6 que hay una variación del rendimiento relativo en los distintos tamaños de grupo. Es decir, a mayor tamaño de grupo, mayor RMSE se presenta. En otras palabras, cuando el tamaño de grupo se incrementa, el rendimiento relativo de las predicciones es menor tanto en los datos originales como en los datos sintéticos.

Otro resultado que se pudo determinar es que el algoritmo de KNN es el que presenta mejor rendimiento tanto en los datos originales (baseline) como en los datos sintéticos (PrivBayes y CART), independientemente del tamaño de grupo y de las funciones de agregación aplicadas. Esto se lo puede observar en la Figura 3.7 y en la Figura 3.8, las cuales muestran mapas de calor con el rendimiento promedio de los distintos tamaños de grupos, por cada estrategia de agregación y por algoritmo elegido (KNN, MeansKNN, SlopeOne, CoClustering).

Para Movielens:

Algoritmo		Baseline	Sintéticos-CART	Sintéticos-PrivBayes
KNN	ADD	0.926	0.971	1.427
	LMS	0.548	0.626	1.240
	MPS	0.519	0.581	1.289
	MULT	1.067	1.140	1.456
MeansKNN	ADD	0.838	0.934	1.185
	LMS	0.857	0.963	1.230
	MPS	0.803	0.887	1.228
	MULT	1.587	1.705	1.753
SlopeOne	ADD	0.905	1.000	1.319
	LMS	0.847	0.945	1.361
	MPS	0.832	0.921	1.381
	MULT	1.560	1.658	1.786
CoClustering	ADD	0.866	0.962	1.368
	LMS	0.883	0.987	1.400
	MPS	0.844	0.930	1.433
	MULT	1.578	1.679	1.797

Figura 3.7: Mapa de Calor - Rendimiento promedio de los distintos tamaños de grupo para el conjunto de datos de Movielens

Para Goodbooks:

Algoritmo		Baseline	Sintéticos-CART	Sintéticos-PrivBayes
KNN	ADD	0.657	0.723	0.929
	LMS	0.654	0.719	0.995
	MPS	0.632	0.688	0.911
	MULT	1.401	1.469	1.600
MeansKNN	ADD	0.697	0.805	0.917
	LMS	0.956	1.076	1.175
	MPS	0.932	1.054	1.113
	MULT	1.985	2.079	2.088
SlopeOne	ADD	0.724	0.868	0.991
	LMS	1.039	1.173	1.285
	MPS	0.998	1.136	1.187
	MULT	2.040	2.152	2.162
CoClustering	ADD	0.670	0.757	0.989
	LMS	0.881	0.970	1.195
	MPS	0.858	0.958	1.140
	MULT	1.852	1.960	2.019

Figura 3.8: Mapa de Calor - Rendimiento promedio de los distintos tamaños de grupo para el conjunto de datos de Goodbooks

En cuanto a las estrategias de agregación, se pudo determinar que:

- Para el conjunto de datos de Movielens la estrategia de Most Pleasure (MPS) es la que mejor rendimiento promedio presenta, independientemente del algoritmo de recomendación aplicado.
- Para el conjunto de datos de Goodbooks la estrategia de Additive Utilitarian (ADD) es la que mejor rendimiento promedio presenta en tres de los cuatro algoritmos de recomendación aplicados.
- Para los dos conjuntos de datos, la estrategia Multiplicative (MULT) es la que peor rendimiento promedio reporta. Esto se cumple para los 4 algoritmos de recomendación aplicados.

Luego de haber realizado todo el análisis, se ha obtenido que el modelo de recomendación grupal con mejor rendimiento relativo es el que se compone de:

KNN, MPS, CART: Movielens

KNN, ADD, CART: Goodbooks

Sobre la base de estos resultados, se puede decir que el objetivo del proyecto se ha probado con éxito, ya que, los datos sintéticos tienen potencial uso en algoritmos de SRG y en futuras investigaciones dado que ha sido demostrada su factibilidad.

4 CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

- ❒ En el presente trabajo se desarrolló un SRG usando datos sintéticos de CART y de PrivBayes, con el fin de mantener la privacidad de las preferencias de los usuarios. Se usaron dos enfoques distintos para medir el error relativo entre ellos y los datos originales para así poder concluir sobre cuál método ofrece mejores resultados.

- ❒ Se hizo una revisión de la literatura sobre trabajos relacionados de SR y SRG, sus aplicaciones en la vida real. También se investigó sobre cómo mantener la privacidad de la información en SR y SRG.

- ❒ Se trabajó con dos conjuntos de datos distintos, los cuales tienen distinto contexto. El dataset de Movielens contiene información de usuarios, películas y ratings; mientras que el dataset de GoodBooks contiene información de usuarios, libros y ratings. Posterior a ello, y dado que los datasets escogidos no incluían información a nivel grupal, se crearon grupos de tamaño variado ($n = 2, 3, 4, 5, 6, 7, 8$).

- ❒ Se aplicaron cuatro estrategias de agregación (ADD, LMS, MPS, MULT) para obtener las recomendaciones grupales. Luego, se hizo la generación de datos sintéticos con el método de CART y con el método de PrivBayes. A continuación, se verificó el comportamiento del SRG con los datos originales y con ambos métodos de sintetización de datos. Se obtuvo que el método de CART es el que mejor modela el error que siguen los datos originales, esto debido a que la curva de error se aproxima mucho más entre CART y los datos reales, que lo que sucede con el método de PrivBayes.

En los análisis experimentales se analizaron 224 escenarios por cada dataset. En Movielens se determinó que el modelo que mejor se ajusta para un SRG es el que abarca todos estos pasos: *1) KNN, 2) MPS, 3) CART*. En Goodbooks, en cambio, el modelo que mejor se ajusta para un SRG es el que está compuesto por: *1) KNN, 2) ADD, 3) CART*.

Siendo que el paso 1) corresponde al algoritmo de filtrado colaborativo que menor error

reportó, el paso 2) se refiere a la estrategia de agregación que menor error dio, y el paso 3) abarca al método de sintetización de datos que mejor se ajustó a la curva de error de los datos originales.

En cada dataset se hizo un análisis y evaluación de la métrica de error RMSE para poder determinar el error relativo de las predicciones del SRG. El punto de interés del proyecto no fue enfocarnos en que el error absoluto sea el menor al momento de generar predicciones para el SRG (como es el caso de evaluación general para un sistema de recomendación); sino medir el error relativo, de manera que se observó el grado de error del modelo basado en la privacidad, y se determinó que el error relativo con los dos métodos de sintetización de datos siguen un mismo patrón que el modelo de recomendación grupal con los datos originales.

Por último, se observó que a mayor tamaño de grupo, menor rendimiento relativo se presenta en la recomendación grupal, esto es evidente, ya que al tener mayor número de personas en el grupo, más difícil será encontrar un consenso entre ellos. Se sugiere que cuando se realicen recomendaciones grupales el tamaño del grupo no sea muy elevado.

4.2 RECOMENDACIONES

Nuestros resultados sugieren que la investigación futura y los científicos de datos pueden probar, desarrollar e implementar algoritmos sobre datos sintéticos con nuestro enfoque, así como en datos sin procesar, ya que ambos seguirán un comportamiento similar. Además, cualquier mejora realizada sobre los algoritmos que utilizan los datos originales se reflejará y transferirá a los datos sintéticos, ya que el comportamiento y los patrones de los datos originales se transferirán al generar los datos sintéticos.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] Y. Resheff, Y. Elazar, M. Shahar y O. Shalom, «Privacy and Fairness in Recommender Systems via Adversarial Training of User Representations,» *arXib*, 2018.

[2] A. Narayanan y V. Shmatikov, «Robust de-anonymization of large sparse datasets,» *IEEE Symposium on Security and Privacy*, 2008.

[3] M. Slokom, «Comparing recommender systems using synthetic data,» en *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018, págs. 548-552.

[4] M. Templ, *Statistical disclosure control for microdata*. Springer, 2017, ISBN: 978-3-319-50272-4.

[5] M. Slokom, M. Larson y A. Hanjalic, «Partially Synthetic Data for Recommender Systems: Prediction Performance and Preference Hiding,» *USB/INTRANET proceedings*, 2020.

[6] A. Pujahari y D. S. Sisodia, «Aggregation of preference relations to enhance the ranking quality of collaborative filtering based group recommender system,» *Expert Systems with Applications*, vol. 156, pág. 113 476, 2020.

[7] L. Recalde, «A social framework for set recommendation in group recommender systems,» en *European Conference on Information Retrieval*, Springer, 2017, págs. 735-743.

[8] G. Adomavicius y A. Tuzhilin, «Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,» *IEEE transactions on knowledge and data engineering*, vol. 17, n.º 6, págs. 734-749, 2005.

[9] T. Silveira, M. Zhang, X. Lin, Y. Liu y S. Ma, «How good your recommender system is? A survey on evaluations in recommendation,» *International Journal of Machine Learning and Cybernetics*, vol. 10, n.º 5, págs. 813-831, 2019.

[10] M. Caro-Martínez, «Sistemas de Recomendación basados en técnicas de predicción de enlaces para jueces en línea,» 2017.

[11] C. C. Aggarwal y col., *Recommender systems*. Springer, 2016, vol. 1.

[12] G. Ricci, M. de Gemmis y G. Semeraro, «Matrix and tensor factorization techniques applied to recommender systems: a survey,» *Matrix*, vol. 1, n.º 01, 2012.

[13] F. A. Coronel Flores, «Desarrollo de un sistema de recomendación para identificar información relevante en vigilancia estratégica,» B.S. thesis, Quito, 2021, 2021.

[14] K. Falk, *Practical recommender systems*. Simon y Schuster, 2019.

[15] D. Kluver, M. D. Ekstrand y J. A. Konstan, «Rating-based collaborative filtering: algorithms and evaluation,» *Social Information Access*, págs. 344-390, 2018.

[16] B. Sarwar, G. Karypis, J. Konstan y J. Riedl, «Item-based collaborative filtering recommendation algorithms,» en *Proceedings of the 10th international conference on World Wide Web*, 2001, págs. 285-295.

[17] C. A. Ayala Tipán y K. O. Jiménez Saraguro, «Sistema de recomendación para Ciberseguridad basado en un enfoque de ratings,» B.S. thesis, Quito, 2020., 2020.

[18] M. Kompan y M. Bielikova, «Group recommendations: Survey and perspectives,» *Computing and Informatics*, vol. 33, n.º 2, págs. 446-476, 2014.

[19] L. Baltrunas, T. Makcinskas y F. Ricci, «Group recommendations with rank aggregation and collaborative filtering,» en *Proceedings of the fourth ACM conference on Recommender systems*, 2010, págs. 119-126.

[20] G. Guo, H. Wang, D. Bell, Y. Bi y K. Greer, «KNN model-based approach in classification,» en *OTM Confederated International Conferences."On the Move to Meaningful Internet Systems"*, Springer, 2003, págs. 986-996.

[21] S. Galán, «Filtrado colaborativo y sistemas de recomendación,» *Madrid: Universidad Carlos III de Madrid*, 2007.

[22] D. Lemire y A. Maclachlan, «Slope one predictors for online rating-based collaborative filtering,» en *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, 2005, págs. 471-475.

[23] T. George y S. Merugu, «A scalable collaborative filtering framework based on co-clustering,» en *Fifth IEEE International Conference on Data Mining (ICDM'05)*, IEEE, 2005, 4-pp.

[24] J. Masthoff, «Group recommender systems: aggregation, satisfaction and group attributes,» en *recommender systems handbook*, Springer, 2015, págs. 743-776.

[25] A. Felfernig, L. Boratto, M. Stettinger y M. Tkalčič, *Group recommender systems: An introduction*. Springer, 2018.

[26] C. Senot, D. Kostadinov, M. Bouzid, J. Picault y A. Aghasaryan, «Evaluation of group profiling strategies,» en *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[27] J. Drechsler y J. P. Reiter, «An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets,» *Computational Statistics & Data Analysis*, vol. 55, n.º 12, págs. 3232-3243, 2011.

[28] D. B. Rubin, «The bayesian bootstrap,» *The annals of statistics*, págs. 130-134, 1981.

[29] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava y X. Xiao, «Privbayes: Private data release via bayesian networks,» *ACM Transactions on Database Systems (TODS)*, vol. 42, n.º 4, págs. 1-41, 2017.

[30] J. Masthoff, «Group recommender systems: Combining individual models,» en *Recommender systems handbook*, Springer, 2011, págs. 677-702.

[31] I. Garcia, L. Sebastia y E. Onaindia, «On the design of individual and group recommender systems for tourism,» *Expert systems with applications*, vol. 38, n.º 6, págs. 7683-7692, 2011.

[32] K. McCarthy, M. Salamó, L. Coyle, L. McGinty, B. Smyth y P. Nixon, «Group recommender systems: a critiquing based approach,» en *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, págs. 267-269.

[33] M.-H. Park, H.-S. Park y S.-B. Cho, «Restaurant recommendation for group of people in mobile environments using probabilistic multi-criteria decision making,» en *Asia-pacific conference on computer human interaction*, Springer, 2008, págs. 114-122.

[34] I. A. Christensen y S. Schiaffino, «Entertainment recommender systems for group of users,» *Expert systems with applications*, vol. 38, n.º 11, págs. 14 127-14 135, 2011.

[35] S. Basu Roy, S. Amer-Yahia, A. Chawla, G. Das y C. Yu, «Space efficiency in group recommendation,» *The VLDB Journal*, vol. 19, n.º 6, págs. 877-900, 2010.

[36] M. Sharma, F. M. Harper y G. Karypis, «Learning from sets of items in recommender systems,» *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9, n.º 4, págs. 1-26, 2019.

[37] J. C. Glick y K. Staley, «Inflicted traumatic brain injury: advances in evaluation and collaborative diagnosis,» *Pediatric neurosurgery*, vol. 43, n.º 5, págs. 436-441, 2007.

[38] A. Jameson, «More than the sum of its members,» en *Proceedings of the working conference on Advanced visual interfaces-AVI*, vol. 4, 2004.

[39] W. Wang, G. Zhang y J. Lu, «Member contribution-based group recommender system,» *Decision Support Systems*, vol. 87, págs. 80-93, 2016.

[40] T. V. Vo y H. Soh, «Generation meets recommendation: proposing novel items for groups of users,» en *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018, págs. 145-153.

[41] V. Torra, «Data privacy: foundations, new developments and the big data challenge,» 2017.

[42] N. Shlomo y T. De Waal, «Protection of micro-data subject to edit constraints against statistical disclosure,» 2006.

[43] H. Polat y W. Du, «Privacy-preserving collaborative filtering using randomized perturbation techniques,» en *Third IEEE International Conference on Data Mining*, IEEE, 2003, págs. 625-628.

[44] F. McSherry e I. Mironov, «Differentially private recommender systems: Building privacy into the netflix prize contenders,» en *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, págs. 627-636.

[45] N. Patki, R. Wedge y K. Veeramachaneni, «The synthetic data vault,» en *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2016, págs. 399-410.

[46] G. Beigi, A. Mosallanezhad, R. Guo, H. Alvari, A. Nou y H. Liu, «Privacy-aware recommendation with private-attribute protection using adversarial learning,» en *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, págs. 34-42.

[47] M. del Carmen Rodrıguez-Hernández, S. Ilarri, R. Hermoso y R. Trillo-Lado, «DataGenCARS: A generator of synthetic data for the evaluation of context-aware recommendation systems,» *Pervasive and Mobile Computing*, vol. 38, págs. 516-541, 2017.

[48] S. Shang, Y. Hui, P. Hui, P. Cuff y S. Kulkarni, «Privacy preserving recommendation system based on groups,» *arXiv preprint arXiv:1305.0540*, 2013.

[49] B. J. Oates, «Researching Information Systems and Computing,» *SAGE Publications Ltd.*, 2006.

[50] S. T. March y G. F. Smith, «Design and natural science research on information technology,» *Decision support systems*, vol. 15, n.º 4, págs. 251-266, 1995.

[51] K. Peffers, T. Tuunanen y M. A. Rothenberger, «Methodology for Information Systems Research,» 2008.

[52] D. B. Fernández y S. L. Mora, «Uso de la metodologıa CRISP-DM para guiar el proceso de minerıa de datos en LMS,» en *Tecnologıa, innovación e investigación en los procesos de enseñanza-aprendizaje*, Octaedro, 2016, págs. 2385-2393.

[53] P. Chapman, J. Clinton, R. Kerber y col., «CRISP-DM 1.0,» *CRISP-DM Consortium*, vol. 76, n.º 3, 2000.

[54] ——, «The CRISP-DM user guide,» en *4th CRISP-DM SIG Workshop in Brussels in March*, sn, vol. 1999, 1999.

[55] *IBM Conceptos básicos de ayuda de CRISP-DM*, https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview, Accessed: 2022-05-01.

[56] A. T. Ferreira, C. Fernandes, J. Vieira y F. Portela, «Pervasive intelligent models to predict the outcome of COVID-19 patients,» *Future Internet*, vol. 13, n.º 4, pág. 102, 2021.

[57] D. Herzog y W. Wörndl, «A user study on groups interacting with tourist trip recommender systems in public spaces,» en *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 2019, págs. 130-138.

[58] M. Kaya, D. Bridge y N. Tintarev, «Ensuring fairness in group recommendations by rank-sensitive balancing of relevance,» en *Fourteenth ACM Conference on Recommender Systems*, 2020, págs. 101-110.

[59] A. Jameson y B. Smyth, «Recommendation to groups,» en *The adaptive web*, Springer, 2007, págs. 596-627.

[60] L. Boratto, S. Carta y G. Fenu, «Investigating the role of the rating prediction task in granularity-based group recommender systems and big data scenarios,» *Information Sciences*, vol. 378, págs. 424-443, 2017.