UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

“DESARROLLO DE UN SISTEMA DE RECOMENDACIÓN PARA UNA PLATAFORMA DE BENEFICIOS.”

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

BENJAMÍN IGNACIO SKOKNIC BINDER

PROFESOR GUÍA:
Pablo Marín Vicuña

MIEMBROS DE LA COMISIÓN:
Alejandra Puente Chandía
Felipe Vildoso Castillo

SANTIAGO DE CHILE
2023

RESUMEN EJECUTIVO

El proyecto se desarrolla en la empresa *SemSo* fundada por consultora estratégica *GeCo*, la cual ha brindado su servicio a más de 115 empresas, dentro de las cuales ha encuestado a más de 60 mil colaboradores (SemSo, 2022), siendo una pequeña empresa de menos de 50 empleados. (Mercantil, 2022)

SemSo brinda el servicio de asesoramiento en la gestión del bienestar para los colaboradores de las empresas asistidas (SemSo, 2022), para lo cual realiza encuestas a los colaboradores de sus clientes, entregando posteriormente recomendaciones de beneficios. Además, genera un reporte de la situación de la empresa con recomendaciones.

Se identificó como problemática, la calidad insatisfactoria del servicio entregado, especialmente con respecto a las recomendaciones realizadas en la plataforma *SemSo*, lo cual causa adicionalmente el incumplimiento de las metas internas de la empresa.

El objetivo del proyecto es, identificar los atributos de los ítems y usuarios con mayor impacto en la conversión en las recomendaciones de la plataforma de *SemSo*, para mejorar la conversión, utilizando modelos estadísticos e incorporando sus resultados en el algoritmo de recomendación.

El modelo SemSo se basa en las preguntas de la *CASEN* respecto de la pobreza multidimensional, incorporando la dimensión económica al modelo de la encuesta y complementando el resto de las dimensiones con preguntas adicionales, con las cuales se calcula el indicador de carencia multidimensional.

Del análisis exploratorio se identificó que el género, el nivel de ingresos, la edad, la carencia y el tiempo de respuesta tienen incidencia en la conversión en la plataforma. También se observó una preferencia por ítems de la dimensión de vivienda y económica, representando alrededor de un 58% de las interacciones. Por otra parte, proporcionalmente se observó un mayor interés por ítems del tipo subsidio, ya que solo un 5% de los ítems corresponden a subsidios, pero un 25% de las interacciones fueron con subsidios.

Se evaluaron sistemas de filtrado colaborativo con factorización matricial con factores implícitos, filtrado basado en contenido y un modelo basado en un clúster. Para este último se utilizó el *K-Means* con 5 clústeres, identificados como: personas carentes, personas con carencia media, personas con carencia baja, personas en camino al bienestar y personas en bienestar.

Se determinó como mejor alternativa utilizar el modelo de clústeres para los usuarios nuevos que no hayan interactuado, ya que tiene obtuvo un *accuracy* de 0,81*,* 20 veces superior a la alternativa evaluada para este segmento. Para usuarios que ya hayan interactuado, pero con menos de 3 ítems diferentes, se decidió hacer una recomendación en función un filtro basado en contenido, el cual usa una asociación por las etiquetas de los temas de las preguntas y la información de las interacciones, con un *acuraccy* de 0,34. Finalmente, para usuarios con más de 3 interacciones de determino utilizar un filtrado colaborativo de factorización matricial no negativo con iniciación basada en una descomposición de valores singulares, con un *accuracy* de 0,815.

Adicionalmente se entrega un plan de acción para implementar y evaluar los resultados del sistema de recomendación propuesto en un entorno de prueba.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1 INTRODUCCIÓN

1.1 MOTIVACIÓN

Este proyecto se realiza a solicitud del director ejecutivo de *SemSo* y toma lugar en el área de operaciones de *SemSo*, en directa interacción con el área de análisis de datos y el área comercial.

Antes de abordar el problema específico identificado dentro de la organización es relevante mencionar el problema general que *SemSo* intenta solucionar, el cual es la pobreza multidimensional en Chile. “En Chile hay aproximadamente un millón y medio de personas consideradas pobres por la carencia de ingresos; pero alrededor de 3,6 millones de personas tienen otro tipo de carencias, siendo pobres desde un punto de vista multidimensional.” (Ministerio de desarrollo Social y Familia, 2018) Estas carencias están divididas por el Ministerio del desarrollo Social en cuatro dimensiones: Educación, Salud, Vivienda, además de Trabajo y Seguridad Social. Estas abarcan desde el nivel de acceso a la educación hasta si cuenta con servicios básicos como agua potable. Para contribuir a una solución SemSo busca ofrecer beneficios personalizados de manera eficiente disminuyendo asimetrías de información en la población.

En este contexto, para ofrecer beneficios personalizados necesita conocer las necesidades e intereses de los colaboradores, para lo cual se cuenta con la información de la encuesta y de los beneficios que visualizan, buscando que la mayor cantidad posible de colaboradores respondan la encuesta y visualicen beneficios.

Adicionalmente, parte de la rentabilidad del negocio está asociada al envío de mensajes personalizados para los usuarios, debido a la existencia un cobro mensual a sus clientes por diligenciar estos. Además, parte de la propuesta de valor de *SemSo* para sus clientes es la capacidad de entregar recomendaciones atingentes para sus colaboradores y la oportunidad de identificar las acciones que tengan mayor impacto para mejorar el bienestar de sus colaboradores. En este contexto, si no se recomienda beneficios atingentes los usuarios no harán clic en estos, no se identificarán correctamente los intereses y los clientes no le verán valor al servicio.

SemSo ha realizado encuestas en terreno para saber la opinión de sus usuarios respecto del servicio entregado, de la cual se pudo identificar que más del 90% confía en la confidencialidad de encuesta SemSo y que más del 84% consideró que la encuesta era fácil de responder. Sin embargo, menos del 55% de los encuestados declaró que contestar la encuesta *SemSo* le trajo alguna utilidad. De aquellos que consideraron que les sirvió la encuesta, la mayoría declaró que era por los beneficios y en segundo lugar debido a que les permitió conocer la situación de su familia, pero menos del 40% de estos quedó satisfecho con los servicios entregados por *SemSo*.

Adicionalmente del grupo de encuestados que terminaron la encuesta *SemSo* menos del 52% revisó sus beneficios y de estos solo el 29,4% declaró que los beneficios entregados les sirvieron. Sin embargo, cerca del 94% estaba interesado en que le enviaran beneficios de acuerdo con sus necesidades y cerca del 85% recomendaría contestar la encuesta a sus compañeros.

Por otra parte, la organización utiliza la tasa de conversión para medir el impacto general de la plataforma, la cual se calcula como el porcentaje de las personas que terminaron la encuesta e hicieron clic sobre al menos un beneficio y cuya meta es de un

60%. Sin embargo, el 2022 la máxima tasa de conversión mensual que se logró fue de 42,4% y la máxima tasa de conversión semanal fue de un 50,5%. Cabe señalar que, la tasa de conversión de septiembre fue solo de un 14,15%, la tasa mínima de conversión semanal fue de un 7,14% y la mediana de la tasa de conversión semanal fue de 33,4%. Esto implica que, algunas semanas menos de 1 de cada 10 personas que completaron la encuesta seleccionaron un beneficio. Esta información puede complementarse con los datos expuesto en el *Anexo B: Gráficos de conversión de usuarios*.

Es preciso señalar que, a partir de noviembre del 2022 se empezó a realizar una encuesta a los usuarios luego de que completaran la encuesta donde indican que tan posible es que recomienden el servicio entre 1 y 10. Con esto se obtiene el puntaje de promoción neto (*NPS*)[1] que se calcula restándole al porcentaje de promotores (aquellos usuarios que evaluaron con una nota superior a 8) el porcentaje de detractores (aquellos usuarios que evaluaron con una nota menor a 7), cuya meta interna es de 50 para los clientes no pagos y 70 para los clientes pagados, pero el año 2022 el *NPS* fue solo de un 37 como puede observarse en el *Anexo C: Gráficos de NPS*.

1.2 SOLUCIÓN PROPUESTA

Dado lo anterior se evidencia interés por conocer su situación y acceder a beneficios por parte de los colaboradores, pero, existen carencias en la calidad del servicio que impiden el cumplimiento de los objetivos organizacionales. Esto puede deberse a que los usuarios no encuentran beneficios que se adecuen a sus necesidades e intereses. Para solucionar esto, existen dos alternativas evidentes, la integración de más beneficios o la mejora de la recomendación de los beneficios.

Para integrar nuevos ítems existen dos alternativas, la integración de beneficios existentes de libre disposición o la creación/adquisición de nuevos beneficios. Por un lado, existe una limitante de beneficios gratuitos existentes, los cuales continuamente se están integrando en la plataforma y aquellos que no son agregados debido a la existencia de beneficios parecidos en la plataforma de beneficios que tienen mejor evaluación. Además, la integración de otro tipo de beneficios conlleva un alto costo y requiere que exista realmente una demanda por él. Esto implica que, para poder crear más beneficios se deba asegurar que es aquel que los usuarios necesitan o que les interesa, siendo fundamental el entendimiento de los intereses y necesidades de los usuarios.

Por otro lado, se tiene que al menos hasta mediados del 2023 la recomendación se realiza en función de un algoritmo que ordena los beneficios por una priorización basada en juicios y corrige situaciones de borde en función de necesidades identificadas en la encuesta.

Para mejorar la calidad de las recomendaciones es necesario tener en cuenta los intereses de los usuarios, por lo que se plantea analizar la información disponible para entender la conversión de los usuarios, utilizando las interacciones de los usuarios para identificar los intereses e incorporar estos resultados en un nuevo sistema de recomendación.

En este ámbito se plantea utilizar un nuevo sistema de recomendación para mejorar la experiencia de los usuarios, donde para los usuarios antiguos se podría aumentar su

[1] ***NPS**: Net Promoter Score*

satisfacción incidiendo en el *NPS* y para los usuarios nuevos se podría mejorar su primera interacción incidiendo en la tasa de conversión.

Usualmente para los usuarios nuevos los sistemas de recomendación utilizan los ítems mejor evaluados y/o con mayor número de visitas para la recomendación. Una vez que ya se tienen las preferencias, evaluaciones u interacciones de los usuarios se utilizan sistemas de recomendación basados en esta información.

Considerando esto, se pretende utilizar técnicas para agrupar a los usuarios e ítems similares de manera de mejorar las recomendaciones realizadas a usuarios nuevos. Para estos se les pretende recomendar los ítems más vistos o mejor evaluados por los usuarios similares a ellos. También se plantea evaluar si algún sistema de recomendación popular básico logra mejores resultados para los usuarios con mayor cantidad de información.

1.3 OBJETIVOS Y ALCANCES

1.3.1 OBJETIVO GENERAL

Identificar los atributos de los ítems y usuarios con mayor impacto en la conversión en las recomendaciones de la plataforma de *SemSo*, para mejorar la conversión, utilizando modelos estadísticos e incorporando sus resultados en el algoritmo de recomendación.

1.3.2 OBJETIVOS ESPECÍFICOS

- Identificar grupos de usuarios con comportamientos y características similares.
- Identificar las variables disponibles que inciden en las interacciones de los usuarios en la plataforma *SemSo*.
- Evaluar distintos sistemas de recomendación conforme a los datos de la plataforma *SemSo* y seleccionar aquellos implementables.
- Desarrollar un algoritmo de recomendación en función de los sistemas de recomendación seleccionados para la plataforma *SemSo*.

1.3.3 ALCANCES

Dado la metodología utilizada para la reconexión de datos no es posible la utilización de todos ellos para la elaboración de los modelos, puesto que solo se tiene la información del tráfico de usuarios por los resultados a partir de junio del 2022, que suman más de 8 mil encuestados a octubre del 2022.

Además, debido a la arquitectura de datos de *SemSo* existen atributos que no se podrán evaluar, ya que la información a la que se accede para la realización de los análisis corresponde a una base de datos que clona y agrupa información de varias tablas, pero que no incluye todas las variables. Asimismo, la arquitectura de la plataforma no se encuentra normalizada, por lo cual pueden ocurrir diferencias que complican la inclusión de algunas variables, siendo necesario dejarlas fuera de los alcances de la memoria.

Adicionalmente, por limitaciones de tiempo no se considera la implementación misma del algoritmo en la plataforma *SemSo*. Sin embargo, se dejará un plan de acción para implementar el algoritmo en un entorno de prueba, realizar una preevaluación del algoritmo y evaluar la conversión real de ambos algoritmos.

1.4 METODOLOGÍA

Los objetivos planteados son abordados con la utilización de herramientas de marketing digital y sistemas de recomendación, lo cual considera tanto la utilización de herramientas de modelamiento para la comprensión y predicción del comportamiento de los usuarios en la plataforma, como la utilización de métodos aplicados de análisis de datos para la validación de hipótesis, la comparación de los modelos y para cuantificar las capacidades predictivas de los modelos.

Adicionalmente la metodología que se utiliza corresponde a *CRISP-DM* (Wirth & Hipp, 2000), la cual abarca entender el negocio asociado a la data, entender los datos, procesar los datos, modelar, evaluar e implementar.

1) **Entender el negocio**; interactuando con el área comercial, se exploran los indicadores de conversión de usuarios, buscando entender las variables que pueden tener influencia sobre el problema y sus correlaciones. Además, se analizan los datos del comportamiento individual de los usuarios para definir las distribuciones de probabilidad que más caractericen el comportamiento individual y de la población.

2) **Entender los datos;** se exploran los datos disponibles de la encuesta para la cuantificación de carencia multidimensional y caracterización de los usuarios y los datos de conversión sobre los ítems u ofertas de beneficios. Para esto se utiliza el lenguaje de programación R, la herramienta R Studio y PowerBI, donde se graficarán los datos, se calcularán estimaciones y métricas, se buscarán anomalías, correlaciones, distribuciones y otros aspectos acordes a un análisis exploratorio de datos. (Exploratory Data Analysis, 2020).

Además, se itera para entender como los datos se relacionan con el negocio y el proceso estudiado.

3) **Procesar los datos**; una vez comprendido los datos se eliminan datos que puedan ensuciar el modelo y se crean muestras de entrenamiento y testeo para los modelos (Data and Sampling Distributions, 2022). Para la eliminación de los datos y creación de muestras se utiliza R.

4) **Modelado**; luego se plantean los diferentes sistemas de recomendación útiles para la resolución del problema, explorando diferentes variables y metodologías, donde se inicia con un sistema de recomendación de referencia, continuando con sistemas de recomendación ampliamente utilizados, hasta la construcción de un sistema de recomendación personalizado (Models for Discrete Choice, 2002) (Limited Dependent Variable and Duration Models, 2002), iterando para entender cuales datos y modelos utilizar. Para esto se usa R y Python.

5) **Evaluación**; a continuación, se evalúan y comparan los algoritmos anteriores aplicados al contexto de la plataforma SemSo para seleccionar los modelos más prácticos y significativos, utilizando diferentes metodologías de acorde a cada modelo (Statistical Experiments and Significance Testing, 2020). Se evalúa como los valores se relacionan con el negocio estudiado y cuáles son las variables más relevantes.

6) **Implementación**; finalmente se utilizarán los aprendizajes anteriores para proponer un sistema de recomendación y un plan de acción.

2 MARCO CONCEPTUAL

2.1 TIPOS DE SISTEMAS DE RECOMENDACIÓN

Los sistemas de recomendación se usan para mostrar ítems a los usuarios en función de preferencias pasadas o las elecciones de otros usuarios, actuando como un filtro de la información relevante para el usuario. (Lu, Wu, Mao, Wnag, & Zhang, 2015) Como se observa en el diagrama de la *Ilustración 1* existen diversos tipos de sistemas de recomendación con variados niveles de complejidad, sin embargo, para una primera iteración solo se explorará la aplicación de sistemas de recomendación básicos.

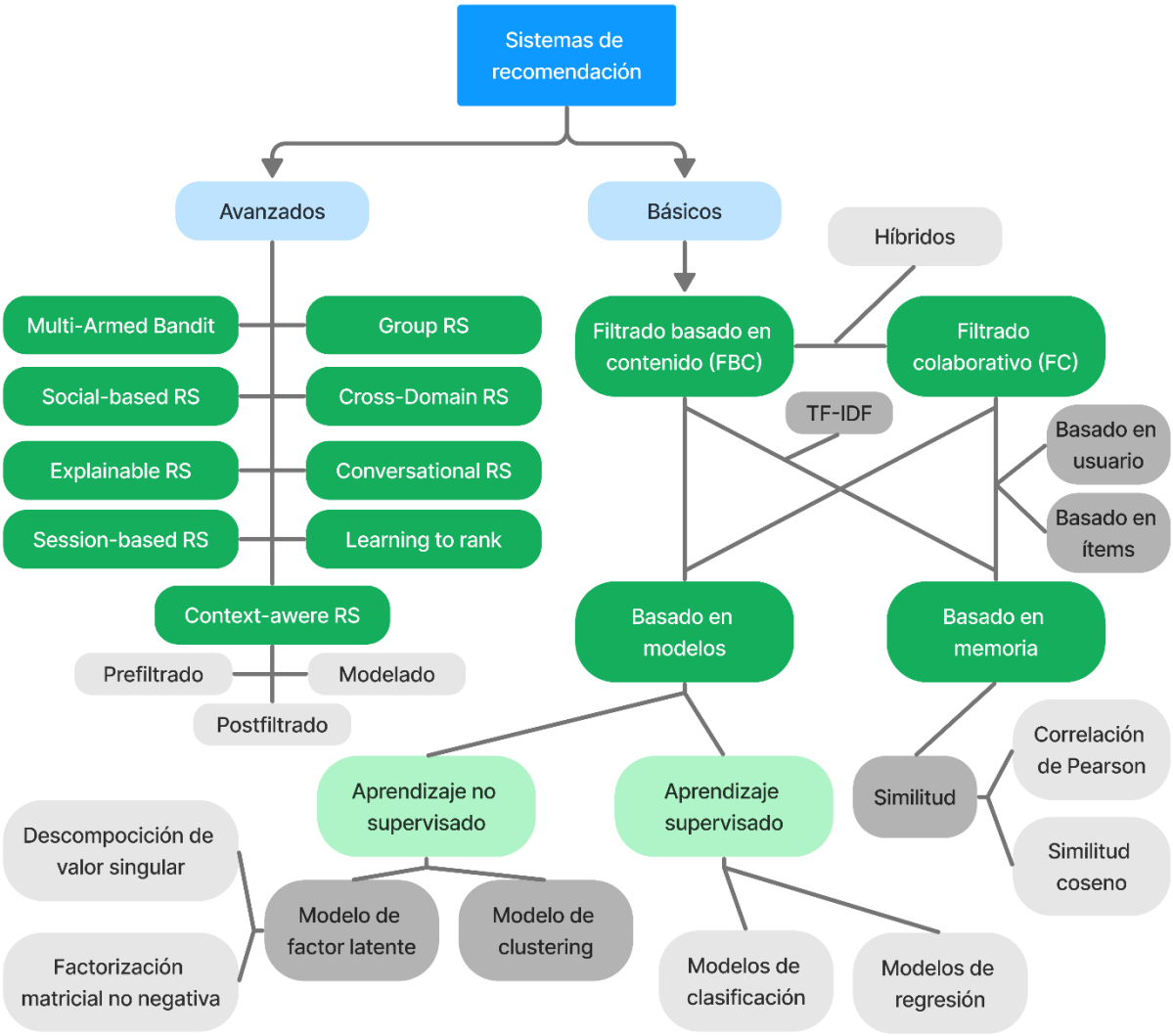


Ilustración 1: Diagrama sistemas de recomendación.

2.1.1 ALGORITMOS DE FILTRADO COLABORATIVO (FC)

Las recomendaciones de algoritmos de filtrado colaborativo se realizan en función de las evaluaciones de usuarios realizadas sobre los diferentes ítems y se pueden dividir en dos categorías principales, aquellas basadas en memoria y aquellas basadas en modelos. A su vez, estos pueden estar centrados en los usuarios o en los ítems. (Bobadilla, Ortega, & Gutiérrez, 2013)

Los filtrados colaborativos basados en memoria con enfoque en los usuarios usan las similitudes entre los comportamientos de evaluación de los usuarios para recomendar

los ítems, usualmente predice la evaluación que tendría un usuario sobre un ítem en función de los usuarios que evaluaron otros ítems de manera similar a este y entrega una recomendación del tipo “usuarios similares a ti le gustaron estos ítems o recomendaciones para ti”.

Los filtrados colaborativos basados en memoria con enfoque en los ítems usan las similitudes entre los ítems para predecir la evaluación de estos, usualmente utiliza las evaluaciones de las interacciones anteriores del usuario para predecir como evaluará los otros ítems y entrega una recomendación del tipo “si te gustó el ítem x puede que te gusten estos ítems o a los usuarios que le gustó el ítem x les gustaron estos ítems”.

Por otro lado, los filtrados colaborativos basados en modelos usan algún modelo de aprendizaje estadístico en función de la matriz de votaciones, siendo algunos ejemplos la factorización matricial, las redes bayesianas y modelos de clustering.

Existen tres aspectos que se deben considerar al usar algoritmos de filtrado colaborativo, listados a continuación. (Aghdam, Analoui, & Kabiri, 2015)

- **Escasez de datos (***Sparsity***):** este problema se da ya que no todos los usuarios evalúan los ítems o no evalúan todos los ítems que visitan, causando falta de información relevante para la recomendación.
- **Escalabilidad:** como considera las votaciones de todos los usuarios e ítems, la cantidad de parámetros crece fuertemente al aumentar la cantidad de ítems y/o usuarios, haciendo los cálculos se vuelven más lentos o requiriendo mejoras de infraestructura.
- **Arranque en frío (***Cold-start***):** el último problema ocurre cuando aparecen nuevos segmentos, ítems o usuarios. En estos casos existen pocas evaluaciones y no se tiene información suficiente para realizar una recomendación o en el caso de los ítems, estos no serán recomendados.

2.1.2 ALGORITMOS DE FILTRADO BASADO EN CONTENIDO (FBC)

Los algoritmos de filtrado basado en contenido se basan en la información de los ítems para hacer una recomendación, buscando que el ítem concuerde con los gustos y preferencias del usuario. (Chang & Hsiao, 2013) Este calcula la similitud entre los ítems por sus características, recomendado ítems similares a los que el usuario vio entregando recomendaciones del tipo “si te gustó el ítem x puede que te gusten estos ítems o si te gusta esta característica puede que te interesen estos ítems”.

Este algoritmo tiene las siguientes limitantes:

- **Análisis de contenido limitado:** este problema surge debido a que las recomendaciones de un ítem están limitadas a las características que se le asociaron, por lo que si se asignan mal las palabras claves o las características el algoritmo lo recomendará erradamente. (Ricci, Rokach, & Shapira, 2015);
- **Sobre especialización:** este problema ocurre cuando los ítems que calzan en un perfil son muy limitados, causando que todas las recomendaciones sean demasiado parecidas entre ellas. (Lops, Gemmis, & Semeraro, 2011)
- **Arranque en frío (***Cold-start***):** el último problema ocurre cuando aparecen nuevos segmentos, ítems o usuarios. En estos casos existen pocas evaluaciones y no se

tiene información suficiente para realizar una recomendación o en el caso de los ítems, estos no serán recomendados. (Adomavicius & Tuzhilin, 2005)

2.1.3 ALGORITMOS DE FILTRADO DEMOGRÁFICO (DM)

En este algoritmo se agrupan a las personas de acuerdo con su información demográfica como edad, género o ubicación. Luego entrega recomendaciones en función de las preferencias de los usuarios pertenecientes al grupo. (Pazzani, 2000)

Las desventajas de este tipo de algoritmo es que se necesita toda la información demográfica del usuario para poder hacerle una recomendación.

2.1.4 ALGORITMOS DE FILTRADO HÍBRIDOS

Los algoritmos de filtrado híbrido integran o combinan dos o más algoritmos de filtrado con el fin de mejorar las recomendaciones y suplir las desventajas que cada filtrado pueda tener.

Sus desventajas dependerán de la combinación de algoritmos que se esté realizando y de la manera en la cual estos sean integrados. (Valdiviezo & Hernando, 2016)

Existen varias estrategias para realizar las combinaciones de múltiples sistemas de recomendación, algunas de las más implementadas para resolver el problema del arranque en frío son: (Hybrid Web Recommender Systems, Robin Burke, 12.2)

- ***Por pesos (****Weighted****)***: Se combinan numéricamente los puntajes obtenidos en los diferentes sistemas de recomendación.
- ***Adaptativa (****Switching****)***: El Sistema selecciona y aplica elementos de los diferentes sistemas de recomendación.
- ***Mixta (****Mixed****)***: Se entregan recomendaciones intercaladas de los diferentes sistemas de recomendación.
- ***Combinación de atributos (****Feature Combination****)***: Características derivadas de diferentes fuentes de información son combinadas y entregadas a un único sistema de recomendación.
- ***Agregación de atributos (****Feature Augmentation****)***: Un método de recomendación es utilizado para calcular atributos, los cuales son utilizados como *input* de otro Sistema de recomendación.
- ***Cascada (****Cascade****)***: Se les atribuyen prioridades a diferentes sistemas de recomendación y se utilizan los sistemas con menor prioridad para desempatar ítems con igual puntaje de sistemas de recomendación con mayor prioridad.
- ***Multiniveles (****Meta-level****)***: Un método de recomendación es aplicado para crear un modelo, el cual es usado como input de otro sistema de recomendación.

2.2 TÉCNICAS UTILIZADAS EN SISTEMAS DE RECOMENDACIÓN

2.2.1 APRENDIZAJE ESTADÍSTICO

El aprendizaje estadístico corresponde a la utilización de diferentes herramientas para entender los datos analizados. Estas herramientas se pueden clasificar en dos categorías: supervisadas y no supervisadas. En la mayoría de los casos el aprendizaje estadístico supervisado requiere la construcción de un modelo estadístico que estime o prediga un resultado en función de parámetros indicados. Por otro lado, el aprendizaje estadístico no supervisado no realiza predicciones ni estimaciones, sino que, en función de los parámetros indicados se buscan relaciones y agrupaciones de datos.

Además, se tiene que las variables utilizadas en los modelos pueden ser de dos tipos, *cuantitativas* o *cualitativas*. Por una parte, las variables cuantitativas son aquellas numéricas que representan alguna característica de los datos. Por otra parte, las variables cualitativas son aquellas que tienen algún valor particular perteneciente a alguna de las posibilidades de su clase.

Generalmente los problemas cuya variable dependiente, aquella predicha o estimada, corresponde a una variable cuantitativa es denominado un *problema de regresión* y aquellos en las cuales esta variable es cualitativa son denominados *problemas de clasificación*. Sin embargo, hay modelos que pueden utilizar ambas categorías de variables e incluso otras que pese a ser de clasificación estiman probabilidades por lo que pueden ser consideradas igualmente de regresión. (What Is Statistical Learning?, 2017)

También se tienen *modelos probabilísticos* y *modelos estructurados*, donde en el primero se modela la toma de decisiones de los agentes como variables aleatorias y en el segundo se asume un comportamiento racional en la toma de decisiones.

Considerando lo anterior se tendrán dos naturalezas de decisión a estudiar, por un lado, *decisiones continuas* como la frecuencia con que se visitan. Por otro lado, se tienen *decisiones discretas* como la selección de un ítem especifico.

Por último, en función de la temporalidad de la decisión se tiene los *modelos estáticos*, donde se asume que la decisión no cambia, ni depende de cuando se hizo, y se tienen los *modelos discretos*, en los cuales las decisiones que se tomaron el pasado afectan a las que se toman en el presente. (Models for Discrete Choice, 2002)

2.2.2 K-MEANS

El modelo de clúster *K-Means* es un modelo simple que particiona los datos en una cantidad especificada de grupos. Este utiliza variables del tipo cuantitativo y asigna las observaciones a los grupos de tal manera que la disimilitud promedio desde la media de cada clúster sea la mínima posible.

Dentro de las ventajas del modelo *K-Mean* se encuentra su simpleza, el hecho de que no es afectado por cómo están ordenado los datos y la incorporación de análisis de varianza.

Sin embargo, dentro de sus desventajas está que: su resultado depende en gran medida de cómo son asignados inicialmente los centros; se pueden tener grandes diferencias entre el óptimo local y el global; su resultado depende del número de grupos que se escoge, cuyo óptimo no siempre es evidente; el modelo es sensible a los *outliers*, pudiendo generar un clúster de *outliers* o afectando a los centros; solo utiliza variables cuantitativas y puede generar clúster desequilibrados. (A Comprehensive Overview of Basic Clustering Algorithms, Glenn Fung, 5.1.5)

2.2.3 ÁRBOLES

Estos se pueden usar como modelos de regresión o como de clasificación. Sin embargo, los árboles de clasificación (*CART*) solo hacen sentido si se tienen múltiples variables de clasificación, estos dividen los atributos en subsecciones y usa reglas de clasificación entre las regiones. Sus particiones son binarias y secuenciales, generando la estructura de un árbol. Suelen ser sensibles a las variaciones pequeñas en los datos,

por lo cual existen variaciones como el *Bootstrap Aggregation* (Bagging), que genera varias muestras y agrega los resultados, y *Random Forests,* que, además de hacer muestras por las observaciones hace muestras por las características.

Además, existen diferentes métodos para clusterizar, los clústeres aglomerativos van juntando o agrupando clústeres recursivamente, donde selecciona los dos grupos con menor disimilitud entre ellos, en cambio los clústeres divisivos comienzan con todos los datos agrupados en un único conjunto, el cual se divide en dos conjuntos y luego progresivamente van dividiendo alguno de los conjuntos, de manera que se produzcan los grupos con la mayor disimilitud posible entre ellos.

Algunas de las ventajas que presentan estos modelos es que entregan un alto grado de flexibilidad en la separación de los datos, son fáciles de manipular y son aplicables a cualquier tipo de atributo.

Sin embargo, estos modelos son subjetivos respecto de donde deben detenerse, no considera la posibilidad de mejorar divisiones que ya fueron realizadas y son de difícil escalabilidad.

2.2.4 TERM-FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF)

Es un método usado para el filtrado basado en contenido utilizado comúnmente cuando se quieren recomendar documentos, pues para utilizar este método primero se extrae todas las palabras de los documentos, suele eliminarse las palabras que son conectores y luego se evalúa la relevancia de las palabras en función de su frecuencia. Luego la matriz *TF-IDF* puede ser utilizada para hacer los cálculos de similitud entre los diferentes documentos.

La matriz *TF-IDF* se construye de una matriz con la frecuencia de apariciones de cada palabra en cada documento, dividiendo la frecuencia de cada palabra por el número de documentos en los que aparece. Esto quedando expresado como:

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D)$$

Donde *tf* representa la frecuencia de un término *t* en un documento *d* e *idf* nos indica qué tan informativo es un término en un documento, siendo expresado como:

$$idf(t, d, D) = log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

Donde el numerador *D* es el espacio de documentos *(d_1, d_2…, d_n)* siendo *n* el número de documentos y el denominador es la cantidad total de documentos en los cuales apareció el término *t*.

Luego para hacer las recomendaciones para el filtrado basado en contenido se ponderan los atributos *TF-IDF* de un ítem por los pesos para cada atributo según las preferencias del usuario. (Jannach, Zanker, Felfernig, & Friedrich, 2011)

2.2.5 SIMILITUD

Para poder hacer las asociaciones en el filtrado basado en contenido es necesario utilizar alguna métrica, entre las más utilizadas esta la *similitud coseno* y la *correlación de Pearson*, las cuales se pueden expresar como:

$$Similitud\ coseno: s(\vec{a}, \vec{b}) = \frac{\Sigma_i q_i b_i}{|\vec{a}||\vec{b}|}$$

$$Correlación\ de\ Pearson: r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

2.2.6 FACTORIZACIÓN MATRICIAL (MF)

La factorización matricial es una técnica utilizada en filtrado colaborativo la cual descompone la matriz de evaluaciones en dos matrices de menor dimensionalidad, de tal manera que la matriz de evaluaciones (*A*) pueda ser recompuesta por la multiplicación de estas dos matrices. Cabe destacar que una de estas representa a los usuarios con factores latentes (*W*) y la otra representa a los ítems con factores latentes (*H*), pudiéndose expresar como:

$$A \approx WH$$

El hecho de que estos factores sean latentes complica la interpretabilidad del modelo, sin embargo, reduce la carga computacional. Adicionalmente, existen variaciones de esta metodología como la factorización matricial con feedback implícito la cual en vez de utilizar la matriz de evaluaciones utiliza una matriz con las interacciones de los usuarios con los ítems en formato binario llamada matriz de preferencias y una matriz de confiabilidad que pondera la cantidad de interacciones por un factor. (Jannach, Zanker, Felfernig, & Friedrich, 2011)

También se tiene la factorización matricial no negativa (*NMF*), la que se caracteriza por tener solo factores positivos, lo que en una predicción de la evaluación tiene sentido pues estas van de 1 a 5.

2.2.6.1 FACTORIZACIÓN MATRICIAL NO NEGATIVA POR ACTUALIZACIÓN DE PESO MULTIPLICATIVO

Es un método iterativo para calcular la *NMF*, para lo cual utiliza como función de costo el cuadrado de la *distancia Euclidiana* entre *A* y *B*, y la función *D(A||B)*, representadas respectivamente como: (Lee & Seung, 1999)

$$\|A - B\|^2 = \sum_{ij}\left(A_{ij} - B_{ij}\right)^2 \qquad D(A||B) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right)$$

Ambas funciones convierten a 0 cuando *A* y *B* son iguales, por lo que se definen dos problemas de optimización:

$$\min\|A - WH\|_F^2 \qquad s.t. W, H > 0 \qquad \min D(A||WH) \qquad s.t. W, H > 0$$

Para lo cual se resuelve:

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^T A)_{a\mu}}{(W^T WH)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia}\frac{(AH^T)_{ia}}{(WHH^T)_{ia}}$$

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}A_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}A_{i\mu}/(WH)_{i\mu}}{\sum_v H_{av}}$$

2.2.6.2 FACTORIZACIÓN MATRICIAL NO NEGATIVA DE DOBLE DESCOMPOSICIÓN DE VALORES SINGULARES (NNDSVD)

La factorización matricial no negativa de doble descomposición de valores singulares es un método que mejora la iniciación de *NMF* basado en la descomposición de valores singulares (*SVD*), debido a la sensibilidad de la iniciación iterativa, en especial cuando hay escasez de datos. (Boutsidis & Gallopoulos, 2007)

Este inicia computando los *k* principales tripletes singulares de *A*, luego forma matrices de rango unitario obtenidos del par de vectores singulares, después extrae su sección positiva y los tripletes singulares respectivos de información y finalmente utiliza estos para iniciar (*W, H*).

2.3 MÉTRICAS DE EVALUACIÓN

Existen varias metodologías para comparar modelos, cuales métricas utilizar dependerá de varios factores como de los modelos que se estén comparando, por ejemplo, hay métricas que para poder ser comparadas requieren que ambos modelos tengan la misma cantidad de variables. En esta situación se desea evaluar la capacidad de los diferentes sistemas de recomendar un ítem que le interese al usuario, para lo cual se puede utilizar la siguiente métrica:

- **Accuracy:** Esta métrica resume el rendimiento de un modelo de clasificación como el número de predicciones correctas divididos por el número total de predicciones. Esta métrica nos permite tener un balance general del rendimiento de la predicción.

Por otra parte, para los modelos de clustering se utilizan otro tipo de métricas, que buscan cuantificar que tan diferentes son los grupos. A su vez estas métricas pueden ser usadas como herramientas que buscan determinar el número óptimo de clúster. Entre ellos se tiene:

- **Método de suma cuadrática dentro de los clústeres (***Within Cluster Sum of Squares, WSS***):** Este busca minimizar la suma total del cuadrado de la distancia entre cada punto y su centroide. Se compara el resultado para diferentes números de clúster y se selecciona aquel donde la suma no aumenta significativamente al aumentar el número de clúster.
- **Método de Silhouette:** Este busca cuantificar la calidad de los clústeres, calificando en una escala de -1 a 1 qué tan bien asignado está cada punto en su clúster en comparación al resto de los clústeres y promediando en busca del número de clúster cuyo coeficiente este más cercano a 1.
- **Método de diferencia estadística (***Gap-Statistic***):** Compara la dispersión de los puntos respecto de un caso aleatorio, donde el óptimo número de clúster es aquel que maximiza la diferencia entre la dispersión observada y la esperada por el caso aleatorio.

3 ENTENDIMIENTO DEL NEGOCIO

3.1 ANTECEDENTES GENERALES

La memoria se desarrolla en el contexto de la empresa *SemSo* la cual comenzó como una iniciativa de la empresa *GeCo*. Si bien la memoria solo se desarrolla en los procesos y servicios de *SemSo*, esta comparte integrantes y recursos con *GeCo*, por lo que se abarcan aspectos generales de *GeCo* y en mayor detalle aspectos de *SemSo*.

3.1.1 CARACTERÍSTICAS DE LA ORGANIZACIÓN

GeCo es una consultora estratégica cuya razón social es "*Gestión de Comunidades LTDA*" y su giro es "*Asesoría Consultoría y Capacitación en Gestión Social, Comunitaria y de Negocios*", siendo su actividad de consultoría de gestión. *GeCo* clasifica como una pequeña empresa de consultoría de gestión. (Mercantil, 2022) Para inicios del 2023 *GeCo* cuenta con 13 colaboradores y brinda el servicio de *SemSo* a nivel nacional, la cual está reconocida como una empresa B certificada y ha recibido reconocimientos como Iniciativas Sustentables 2020 para la reducción de la desigualdad.

GeCo tiene el objetivo de transformar a las empresas en el motor de desarrollo social de Chile, el cual lo declaran como "*Mejores Empresas para una Mejor Sociedad*". Además, resumen sus valores en el acrónimo "*TREPA*" (RAE: "*subir a un lugar alto o poco accesible*"), relacionando la dinámica interna con la escalada de un cerro al indicar "*cómo en cualquier escalada, se parte desde abajo, desde el primer valor Alegría para subir progresivamente a través de Pasión por lo Social, Emprendimiento, Resultados y así llegar a la Transformación*" (GeCo, 2022).

SemSo se rige por los mismos valores, pero con su motivación que "*busca mejorar el bienestar de los trabajadores y de su hogar*" (Geco, 2022). Para esto, busca crear conciencia de la situación de los colaboradores de las empresas, para focalizar las ayudas sociales y mejorar la eficiencia en el uso de los recursos, lo cual se evidencia en declaraciones como "*A través de nuestra plataforma digital conocemos a los colaboradores y entregamos soluciones personalizadas a ellos y su hogar, porque sabemos que mejorar su bienestar no es un tema de recursos, sino de una alocación adecuada de estos*." (Geco, 2022)

3.1.2 Servicios

SemSo brinda el servicio de asesoramiento para la gestión del bienestar de las empresas. Por una parte, realiza mediciones a los colaboradores respecto de 6 dimensiones asociadas a su bienestar y le entrega acceso a una plataforma con beneficios listados de acuerdo a sus necesidades. Por otra parte, le proporciona a la empresa un análisis de su situación, comparándola con la industria y el país, a través de un reporte dinámico acompañado por recomendaciones para desarrollar su estrategia de bienestar.

Adicionalmente se puede personalizar el servicio, ya sea solicitando cambios en el reporte, integrando los beneficios personales de la empresa dentro de las recomendaciones o la posibilidad de comunicar a los colaboradores sobre nuevos beneficios personalizados respecto de sus preferencias. (SemSo, 2022)

De esta manera se tienen dos usuarios, el trabajador social, encargado de recursos humanos o gerente que desea evaluar el estado de los colaboradores asociados para tomar las medidas correspondientes y los colaboradores de los clientes, los cuales

interactúan con la plataforma para evaluar su situación, explorar y acceder a beneficios dispuestos para ellos.

3.1.3 MERCADO

SemSo ofrece un servicio gratuito y otro de pago, habiendo contratado el servicio más de 115 empresas, las cuales suman más de 60.000 colaboradores encuestados. (SemSo, 2022) La encuesta realizada hasta la elaboración de esta memoria inició en julio del 2021, contando para marzo del 2022 con más de 7.000 colaboradores y 45 empresas como clientes.

Existen varios competidores en el mercado enfocados en los diferentes servicios que ofrece *SemSo* entre los cuales se destacan las empresas *Buk*[2] y *IBLe*[3].

Buk es un software de recursos humanos, si bien este permite conocer la situación de los colaboradores su foco está centrado en gestionar las remuneraciones y gestionar a las personas. Al igual que *SemSo* este brinda beneficios a sus usuarios, sin embargo, *SemSo* sobresale al brindar una evaluación más profunda de los colaboradores y por la entrega de recomendaciones acordes a las carencias de la empresa, permitiéndoles identificar qué beneficios les falta integrar.

Por otro lado, *IBLe* al igual que *SemSo* busca cuantificar aspectos del bienestar de los colaboradores, sin embargo, solo se enfocan en el bienestar laboral y deja de lado las otras necesidades que puedan tener los colaboradores. Además, *SemSo* es respaldado por una mayor cantidad de clientes y entrega una versión gratuita del servicio que le permite ser más competitiva al ser de más fácil acceso.

Adicionalmente existe la empresa *Sophia Oxford*[4], la cual ofrece un servicio parecido al de *SemSo*, sin embargo, esta opera en Inglaterra y no tiene cobertura en el territorio nacional.

Con respecto a *GeCo*, pertenece al sector industrial de consultoría de gestión, donde de acuerdo con los datos del *SII* se tenían 30.588 empresas registradas el año 2020, sumando un total de 196.022.182 UF anuales en ventas, con una renta neta informada de 36.242.559 UF y ponderando un total de 67.991 trabajadores mensuales. Cabe destacar que solo hay 11.330 empresas registradas en este rubro que tengan menos de 50 trabajadores dependientes informados, los cuales suman 112.733.008 UF anuales en ventas (SII, 2022). Dentro de los principales actores se destacan *PwC*, *Deloitte*, *EY* y *KPMG* que representaron el 37,4% del mercado en 2018. Siendo sus participaciones de mercado de *Deloitte* 10,9%, *PwC* 10,1%, *EY* 9,3%, *KPMG* 7,1%, *Accenture* 5,5%, *McKinsey and Company* 4,9% y *Boston Consulting Group* 3,4%. (Projectcor, 2022). En cuanto a GeCo, se encuentra en la posición 25753 del ranking de Portalchile, con un puntaje de ventas de 8 de una escala de 1 al 13 donde 13 representa el grupo superior y un capital de +10 de una escala de -10 a +10, donde +10 representa que está en el décimo decil positivo de capital en el balance de las empresas chilenas. (Portalchile, 2022)

[2] https://www.buk.cl/
[3] https://ible.cl/
[4] https://sophiaoxford.org/

3.1.4 DESEMPEÑO ORGANIZACIONAL

A inicios del 2023 *SemSo* se encuentra en una fase de crecimiento, habiendo iniciado el desarrollo de su plataforma a finales del 2019, entrando en operación el actual sistema en julio del 2021. La encuesta ha sido realizada por más de 60 mil colaboradores, donde más de 23 mil han sido en la versión actual de la plataforma y ya se tienen contratos para realizar la encuesta a más de 10 mil nuevos colaboradores.

Dado que la empresa está en una fase de crecimiento, se encuentra activamente buscando e implementando nuevas medidas para potenciar el servicio y poder cumplir sus objetivos. Entre estos destacan objetivos declarados por el director ejecutivo como "lograr a través del BIG Data y la Agregación de Demanda, generar soluciones que permitan el progreso social y la disminución de la desigualdad.".

En este ámbito la organización desea a futuro generar los servicios necesarios para poder ofrecer un sistema de membresía mensual por cada trabajador, donde se pueda identificar la situación de diferentes escenarios, establecer espacios para ofrecer beneficios desarrollados por *SemSo* al poder identificar la demanda y entregar a las personas los beneficios que necesita y le interesan.

3.2 MODELO SEMSO

El modelo *SemSo* parte de la base del modelo de pobreza multidimensional utilizado por la *CASEN*, lo cual permite hacer comparaciones de las muestras. Adicionalmente, incorpora preguntas asociadas a necesidades fisiológicas y de seguridad de las personas, como también preguntas relacionadas con las necesidades de reconocimiento y autorrealización de las personas.

Además, se definen tres niveles de prioridad donde el nivel 1 corresponde a las preguntas de pobreza multidimensional, el nivel 2 a las preguntas de necesidades fisiológicas y de seguridad y el nivel 3 a las preguntas de autorrealización.

Por otra parte, las preguntas están asociadas a una de seis dimensiones, las cuales son las dimensiones: *Emocional, Físico, Educacional, Vivienda, Económico* y *Redes*.

Cabe destacar que, cada pregunta tiene un tema asociado y tres niveles o alternativas, estos niveles son vistos como una analogía de un semáforo representando cada alternativa un color (Rojo, Amarillo o Verde), ya que SemSo es un acrónimo de Semáforo Social, donde rojo representa estar mal y verde estar bien.

Para la dimensión *Emocional* no se tienen preguntas de prioridad uno; de prioridad dos se tiene los temas *Alcohol, Drogas* y *Acceso psicólogo*; de prioridad tres el tema *Actividades de bienestar emocional*.

Para la dimensión *Física, en* prioridad uno tiene los temas *Sistema de salud, Atención médica* y *Nutrición* infantil; en prioridad dos tiene los temas *Salud dental* y *Salud visual*; en prioridad tres se tiene el tema *Actividad física*.

Para la dimensión *Educacional,* en la prioridad uno se tiene los temas *Rezago escolar, Término 4to medio* y *Asistencia escolar*; en la prioridad dos se tiene el tema *Cédula de identidad*; en la prioridad tres se tiene el tema *Capacitaciones*.

Para la dimensión *Vivienda,* en la prioridad uno se tiene los temas *Servicios básicos, Hacinamiento, Estado de la vivienda, Tipo de vivienda, Distancia servicios esenciales,*

Distancia transporte público, Tiempo al trabajo y *Contaminación*; en la prioridad dos se tiene el tema *Situación de la vivienda*; no tiene temas en la prioridad tres.

Para la dimensión *Económico,* en la prioridad uno se tiene los temas *Seguridad social, Jubilación* y *Cesantía*; en la prioridad dos se tiene los temas *Crédito* y *Deuda*; no tiene temas en la prioridad tres.

Para la dimensión *Redes*; en la prioridad uno se tiene los temas *Red de apoyo, Trato* e *Inseguridad del entorno*; en la prioridad dos se tiene el tema *Acceso a internet*; en la prioridad tres tiene el tema *Dispositivos de internet*.

Esta clasificación puede ser complementada con el diagrama expuesto *Anexo D: Modelo SemSo* donde se puede observar gráficamente los niveles de prioridad, las dimensiones y los temas.

Adicionalmente se calcula índices de carencia general de una persona y por dimensiones, los cuales también representan un semáforo, donde rojo es estar en carencia, amarillo estar en camino al bienestar y verde estar en bienestar.

Para calcular el índice de carencia primero se calcula un peso por cada tema y se pondera por cada una de las respuestas obtenidas de manera que cada nivel del semáforo representa 0, 0.5 y 1 respectivamente. El peso de cada tema o pregunta depende de su prioridad, su dimensión y la cantidad de preguntas que hay en esa categoría.

En función de esto se calcula un valor de carencia entre 0 y 1 donde una menor carencia es estar “mejor” y una mayor carencia es tener una mayor pobreza multidimensional. Además, se calculan carencias ponderadas para cada dimensión cuyo valor también es de 0 a 1.

3.3 ALGORITMO DE RECOMENDACIÓN BASADO EN LA ENCUESTA

El algoritmo utilizado en la plataforma consta de un filtro basado en conocimiento, donde los atributos son asignados manualmente a cada ítem. A cada ítem se le atribuye una relación con uno o más temas, los cuales están en su mayoría relacionados con las preguntas de la encuesta, siendo al menos cada pregunta del modelo *SemSo* un tema. Luego se pondera en función las respuestas de los usuarios en la encuesta. Estas respuestas describen su situación y además un panel de expertos le asigno peso entre 0 y 100 a los diferentes temas en función de la relevancia percibida por estos.

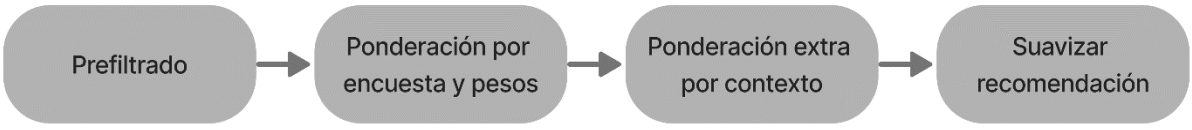


Ilustración 2: Algoritmo de recomendación basado en la encuesta.

Como puede verse en el diagrama de la *Ilustración 2* previo a la ponderación se realiza un prefiltrado de los ítems, excluyendo los ítems específicos de alguna empresa diferente a la del usuario.

Sin embargo, las respuestas de las preguntas son variables categóricas por lo que las respuestas “verde” tienen ponderación 0, las respuestas “amarillo” tienen ponderación 5 y las respuestas “rojo” son ponderadas por 10. Cuando un ítem tiene más de un tema se utiliza el con ponderación mayor.

Por otra parte, se aplica una ponderación de contexto, asociada a las comunas a las que aplican los ítems, esta asociación aumenta un 5% el puntaje obtenido por el beneficio y se hace una priorización de los ítems en función de una evaluación interna de estos de 1 a 5.

Luego para que no agrupe los beneficios similares, se itera sobre todos ellos, ordenados por valor de mayor a menor según el puntaje ponderado usuario ítem, quitando porcentualmente puntuación a medida que se va repitiendo un tema.

Cabe destacar que el algoritmo actual no utiliza la evaluación del ítem ni la popularidad del ítem, los cuales son elementos usualmente utilizados en los sistemas de recomendación y los cuales pueden ser rescatados de las interacciones entre usuarios e ítems.

Finalmente, para identificar los ítems que se muestran en la página de inicio, se seleccionan los 6 beneficios con mayor puntuación. Donde se selecciona el primer beneficio, y luego se excluyen los beneficios que tengan el mismo tema para elegir el siguiente beneficio, repitiendo sucesivamente esta regla para finalmente mostrar aleatoriamente 3 de ellos.

A modo de ejemplo, supongamos que una persona “X” respondió la pregunta asociada al consumo de alcohol “amarillo”, la pregunta asociada a capacitaciones “rojo”, la pregunta de cédula de identidad “rojo” y el resto “verde”. Además, el tema alcohol tiene peso asignado 85, capacitaciones 56 y cedula de identidad 5. Entonces todos los ítems asociados solo a las respuestas “verde” tienen el puntaje 0, luego los ítems asociados solo a cedula de identidad tienen puntaje 50, los asociados a alcohol, pero no a capacitaciones tienen puntaje 425 y los asociados a capacitaciones puntaje 560.

Luego se selecciona aleatoriamente un beneficio de capacitación, si este beneficio no tiene asociado el tema de alcohol se seleccionaría después al azar un ítem de alcohol y por último si el tema de alcohol no tiene asociado el tema de salud mental, se seleccionaría a azar un ítem de salud mental.

4 ENTENDIMIENTO DE LOS DATOS

4.1 INFORMACIÓN DISPONIBLE

A grandes rasgos hay tres fuentes de información con las cuales se trabajará, las que describen la información recolectada de las personas, la información adjudicada a los diferentes ítems y la información del tráfico de los usuarios por la página de beneficios.

La información descriptiva de las personas es recolectada a través de la encuesta para determinar la situación de las personas, adicionalmente a las preguntas explicadas en el modelo *SemSo* se realizan preguntas adicionales y se recopila información de la interacción para identificar a las personas.

Para poder responder la encuesta las personas deben comunicarse a través de *WhatsApp* con un *bot* de *SemSo*, de esta interacción se rescata el número de celular de las personas, además antes de comenzar la encuesta este *bot* les solicita el nombre, su rut, la empresa e instalación en la cual trabaja. Habiendo verificado sus datos el *bot* les manda un enlace para que conteste la encuesta en *typeforme*.

En este el usuario responde las preguntas del modelo *SemSo* y preguntas de caracterización como, su género, su edad, su nacionalidad, el ingreso de su grupo hogar, la cantidad de personas que viven en su grupo hogar, la dirección de su lugar habitacional (asociada una API de Google Maps) y si desean agregar un mail de contacto.

Asociado a estos datos se construyen diversos indicadores, siendo los indicadores más relevantes los de carencia del modelo *SemSo* y el índice de bienestar multidimensional.

La segunda fuente de información relevante corresponde a la de los ítems expuestos en la plataforma *SemSo*. Estos ítems solo pueden ser agregados por personal de la organización y pueden estar sujetos a diversas restricciones. Entre estas restricciones se encuentra la necesidad de pertenecer a una determinada organización, estar en un rango etario determinado, estar suscrito a algún sistema de salud o pertenecer a un grupo vulnerable.

Adicionalmente estos ítems tienen atributos descriptivos como su nombre, una descripción corta, una descripción larga, entidad que la ofrece, atributos de caracterización como tipo, dimensiones, subdimensiones, temas, zona en la que aplica y por último atributos asociados a su funcionamiento como imágenes, links informativos o teléfonos de contacto, entre otros.

También los ítems tienen periodos de duración y su información puede ser modificada, por esto se tienen las fechas de actividad y una fecha de actualización.

Por último, se tiene la información de las interacciones de cada persona con los ítems de la plataforma *SemSo*, esta información tiene asociada un identificador único de la persona, un identificador único del beneficio, el día y la hora en la que se realizó la interacción. Adicionalmente en esta interacción los usuarios tienen la opción de evaluar al ítem con una nota entera entre 1 y 5, donde un 5 es la mejor evaluación. También tiene identificadores del viaje del usuario antes de la interacción, que indican si interactuó con el ítem en la recomendación del inicio, en el buscador o en la recomendación basada en dimensiones.

4.2 ANÁLISIS EXPLORATORIO DE LOS DATOS

En función de los datos disponibles se realizó un análisis exploratorio de los datos para identificar las distribuciones de las diferentes variables e identificar que variables pueden influir en la toma de decisiones de los usuarios en sus interacciones con los ítems expuestos en la plataforma *SemSo*.

Este análisis fue dividido de acuerdo con la información utilizada, sin embargo, antes de realizarlo se eliminaron todos los datos asociados a pruebas. Esto debido a que para probar nuevas funcionalidades o hacer pruebas de concepto se crearon datos ficticios, por lo que no aportan información real de los datos. También se filtraron los ítems de manera de dejar solo la información asociada a su última actualización, evitando que haya atributos repetidos para un beneficio y cualquier información errónea que haya sido corregida.

4.2.1 ENCUESTA

De una exploración inicial de las distribuciones de los datos y sus relaciones se identificó la presencia de outliers como usuarios que reportaron edades superiores a los 120 años, lo cual claramente no es factible en especial al considerar que tendría que estar trabajando en una empresa. También se identificó la existencia de tiempos de respuestas negativos o incluso superiores a un día. Esto implico que se tuviese que limpiar los datos, para lo cual solo se consideraron datos de edades inferiores a 86 años y tiempos de respuestas positivos menores a dos horas.

Se espera que variables como la edad, la nacionalidad y el ingreso puedan incidir en la conversión y para facilitar su análisis se crearon las siguientes variables; una variable banda etaria que agrupa las personas entre 17 y 35 años, las personas entre 36 y 55 años y las personas mayores de 55 años; se creó una variable de agrupación de extranjeros dado que la mayoría de las personas son de nacionalidad chilena y el resto de las nacionalidades tienen representaciones muy bajas; se generalizo los niveles de vulnerabilidad en función del ingreso per cápita en los niveles de pobreza, vulnerabilidad y bienestar; se creó una variable de ingreso per cápita ya que el ingreso estaba expresado en función del hogar.

Posteriormente se analizaron las variables numéricas resumidas en la siguiente tabla.

Variables	Rango	Nulos	Promedio	Desviación estándar	Histograma
Ingresos	[60.000,800.001]	0	273019.08	184951.70	
Edad	[18,81]	0	40.25	12.20	
Personas por hogar	[1,10]	0	3.47	1.68	
Tiempo de respuesta (min)	[1,120]	19	13.23	1294.86	
IPM	[0,100]	0	14.45	13.65	
Carencia	[0,1]	0	0.27	0.14	
Económico	[0,1]	0	0.32	0.23	
Educacional	[0,1]	0	0.22	0.17	
Emocional	[0,1]	0	0.30	0.21	
Físico	[0,1]	0	0.31	0.20	
Redes	[0,1]	0	0.28	0.23	
Vivienda	[0,1]	0	0.20	0.15	

Tabla 1: Variables numéricas datos personas.

Adicionalmente se quiere ver la relación entre las variables, para esto se calcula la correlación entre ellas.

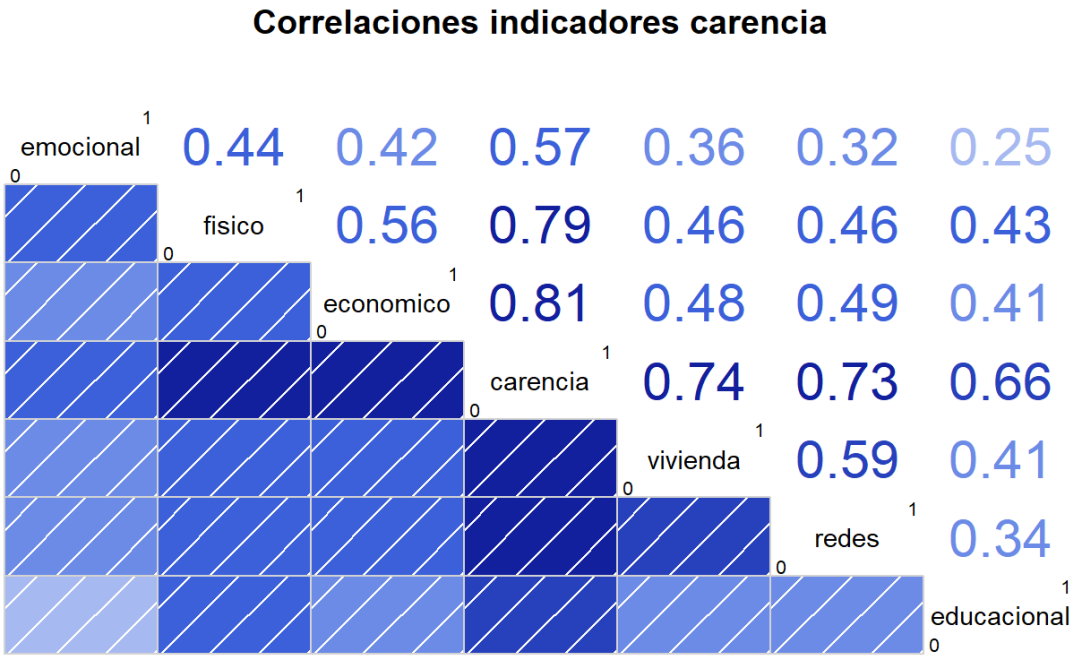


Ilustración 3: Correlaciones indicadores carencia.

Dado que la carencia se calcula en función de las dimensiones es lógico que tengan una correlaciones alta y significativa, sin embargo, esto igual tiene relevancia pues indica que en algunos modelos no se debe incluir todos los indicadores de carencia, ya que las correlaciones afectan el cálculo de los mínimos cuadrados.

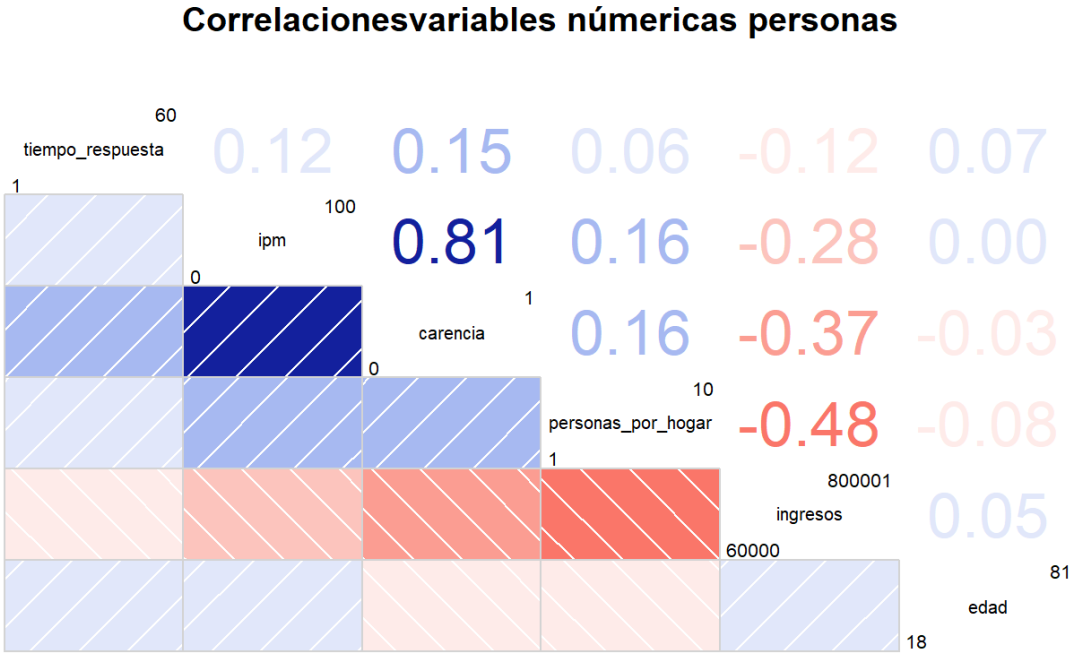


Ilustración 4: Correlaciones variables numéricas personas.

También, se ve que la carencia y el IPM están correlacionados, esto hace sentido, pues, ambos indicadores buscan medir un mismo fenómeno social, sin embargo, con ciertas diferencias como su escala, por lo que no se deben incluir ambas en el modelo. Por otra parte, el ingreso y la cantidad de personas están correlacionados negativamente, esto tiene sentido pues el ingreso per cápita se calcula como el ingreso del hogar dividido por la cantidad de personas del hogar.

Por último, se ve que el resto no están correlacionados, siendo además significativa esta relación, por lo que estas variables numéricamente no debiesen presentar problemas en modelos que supongan que las variables son independientes entre sí.

4.2.2 ÍTEMS

Para poder rescatar más información de los beneficios fue necesario realizar una transformación de sus cualidades de manera que se tenga toda la información de cada beneficio en una fila. Los beneficios se pueden clasificar: en 4 tipos de beneficios, 6 dimensiones, 18 subdimensiones, 37 temas, 9 categorías específicas, 9 públicos, 3 nacionalidades, 6 niveles de vulnerabilidad, si es gratuita, las regiones en las que aplica (16 opciones) y las comunas (346 en total).

Tanto los temas como las subdimensiones y las dimensiones están directamente relacionadas con las preguntas del modelo SemSo. Por otro lado, las regiones y las comunas están asociadas a la cobertura de los ítems, la cual está asociada a zonas del territorio nacional.

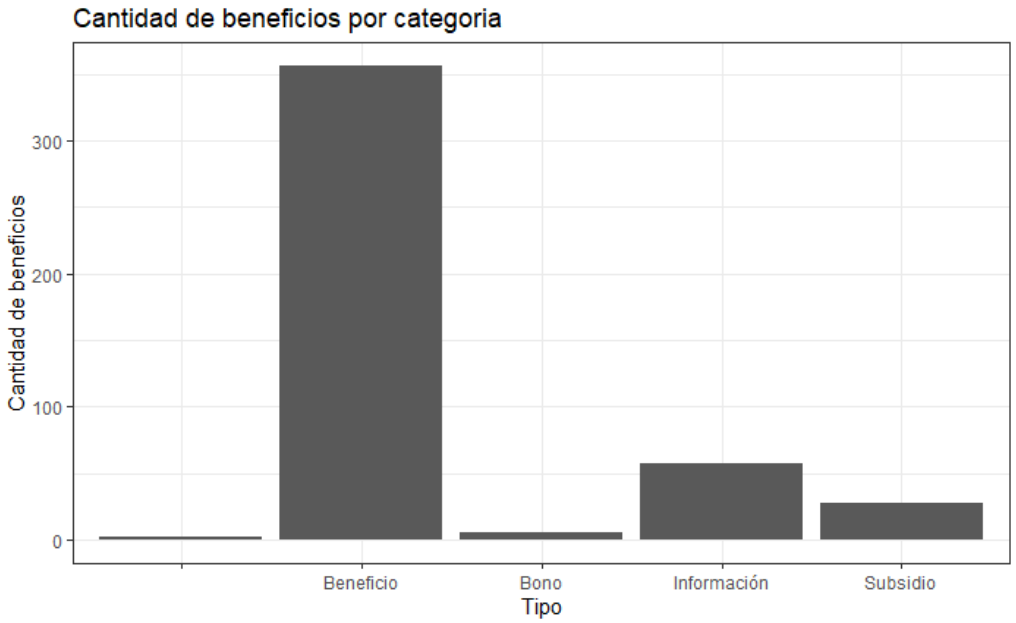


Ilustración 5: Tipos de ítems.

Se percibe en la *Ilustración 5* la mayoría de los ítems corresponden a beneficios, en segundo lugar, se tiene ítems informativos. Sin embargo, se tienen grandes diferencias que puede implicar sesgos en las comparaciones, ya que asumiendo una distribución aleatoria se debiesen ver muchos más ítems de tipo beneficios que otro tipo. A su vez, implica que si la proporción no se cumple en las visualizaciones pueda existir una preferencia por un tipo particular de ítems. Además, existen ítems que no tienen un tipo asociado.

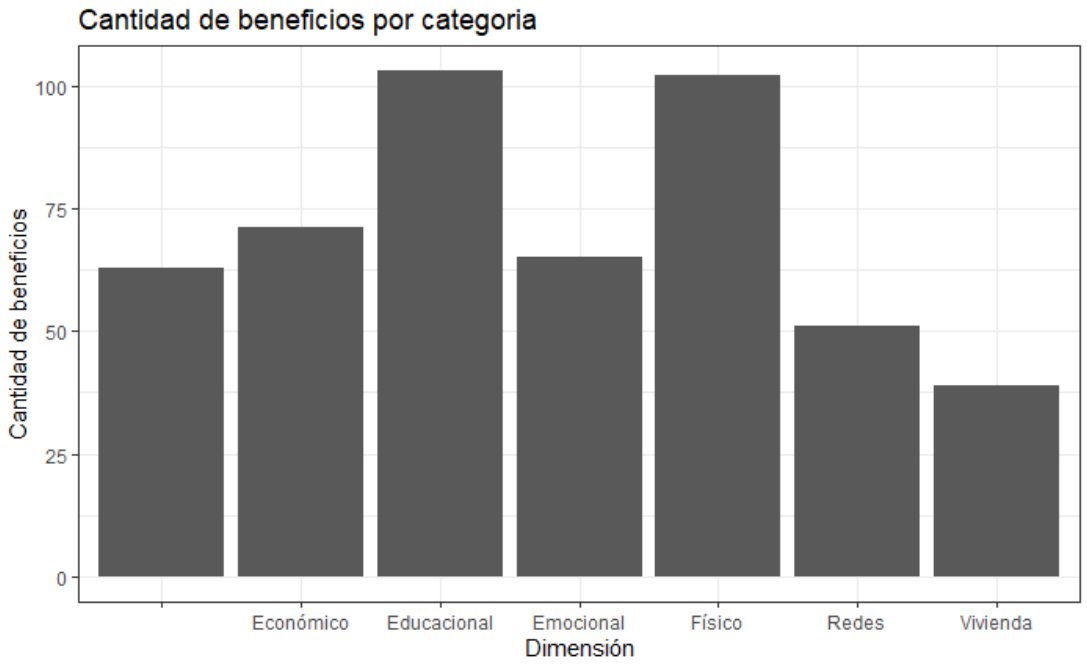


Ilustración 6: Ítems por dimensión.

Como se observa en la *Ilustración 6* se tiene que la mayoría de los ítems están asociados con las dimensiones *Educacional* y *Físico*, siendo la dimensión con menos ítems *Vivienda*. Cabe destacar la existencia de un gran número de beneficios que no tienen ninguna dimensión asociada, esto implica que no tienen un tema del modelo *SemSo* asociado y no serán recomendados en la página inicial de la plataforma *SemSo*.

Alternativamente para asignar o agrupar los ítems se dispone de la descripción y el nombre de estos, sin embargo, al realizar un análisis de frecuencia de las diferentes palabras, los conectores no aportan información de los ítems.

Habiendo eliminado los conectores y palabras que no aporten información se pueden observar palabras que pueden aportar información lógica en función del negocio para los diferentes ítems. Por ejemplo, se observa las palabras vivienda y educación las cuales son asociaciones que también se tiene en las etiquetas por dimensión o las palabras descuento, beneficios e información se pueden relacionar con los tipos de ítems.

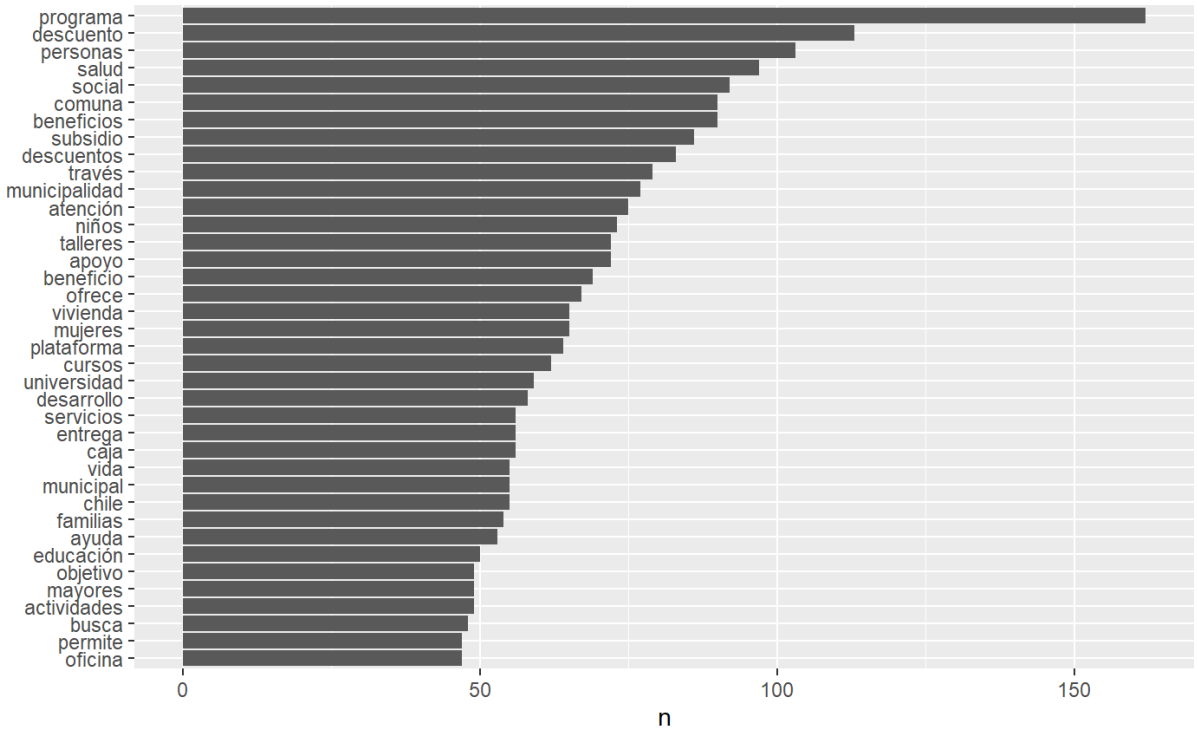


Ilustración 7: Palabras más frecuentes.

4.2.3 INTERACCIONES

La información de la interacción entre personas e ítems es relevante para la creación de indicadores, sin embargo, esta no rescata las cualidades de las personas ni de los ítems. Con esto en consideración igualmente permite calcular la cantidad de ítems que vio cada persona, la cantidad de interacciones que tuvo, cuantas veces se vio cada ítem, la evaluación promedio de cada ítem, la cantidad de evaluaciones de cada ítem y además, permitió identificar “estaciones del día” en que se produjo la interacción (Madrugada]0-6], Mañana]6-12], Tarde]12-18], Noche]18-24]).

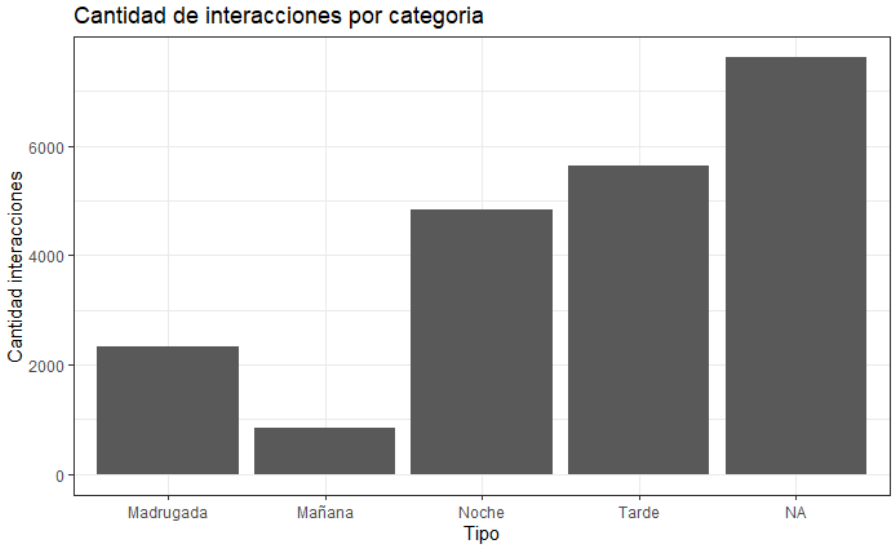


Ilustración 8: Horario de interacción.

En primer lugar, respecto al horario en el que las personas interactúan con los ítems se puede observar una concentración de las interacciones en la tarde y en la noche, esto implica que la mayoría de las interacciones se producen entre las doce del día y las doce de la noche, sin embargo, existe un gran número de interacciones de las cuales no se tiene la hora en que ocurrieron debido a que esta información no fue recolectada en las primeras muestras.

Por otro lado, tiene las diferentes variables cuantitativas que se generaron de las interacciones, las cuales implicaron la agrupación de los datos por los ítems o por los usuarios dependiendo de la variable que se calculara. Los resúmenes de estas variables considerando todas las interacciones del periodo del que se disponen datos pueden ser observados en la *Tabla 2.*

Variables	Nulos	% No Nulo	Promedio	Desviación estándar	Histograma
Interacciones por ítem	0	100%	500.33	442.99	
Visitas por ítem	0	100%	186.80	145.66	
Evaluación ítem	3228	85%	4.43	0.90	
N° Evaluaciones ítem	3228	85%	19.51	21.25	
Interacciones por usuario	0	100%	8.16	6.56	
Ítems vistos por usuario	0	100%	2.95	2.60	
Evaluaciones	20290	3%	4.49	1.15	

Tabla 2: Variables numéricas interacciones.

Se observa que si bien el 85% de los ítems con los cuales hubo interacciones están evaluados, en la mayoría de las interacciones no se tienen evaluaciones. Además, los ítems son en promedio visitados por menos de la mitad de los usuarios que sus interacciones (186.8 visitas y 500.33 interacciones) y que las interacciones tienen una dispersión considerablemente mayor que la cantidad de ítems visualizados (6.56 y 2.6 respectivamente). Sin embargo, los ítems evaluados en promedio son valorados por 19 personas y suelen tener evaluaciones altas (4.49 promedio).

4.3 DEPURACIÓN DE LOS DATOS

A partir de estos datos se crearon dos bases de datos, la primera nace de los ítems e incorpora las características de los estos, las interacciones de los usuarios y los datos de estas.

En esta base se crearon nuevas variables asociadas al cruce de las características del ítem y la persona que lo vio. Entre estas se consideró interesante crear las siguientes variables binarias: si la persona vive en una región/comuna válida del beneficio, si tiene uno o más de los temas del ítem con respuesta roja, si tiene uno o más de los temas del ítem con respuesta amarilla, si tiene la nacionalidad acorde al beneficio, si se está en un nivel de vulnerabilidad dirigido de los beneficios y si se es carente en una de las dimensiones del ítem.

La segunda base parte de la información de las personas, integrando la cantidad de ítems con los que interactuaron. Para la cual se utilizaron los datos a partir del 6 de julio del 2022, pues solo se tiene información de las interacciones a partir de esa fecha y se creó una variable binaria de conversión, la cual indica si la persona vio algún ítem o no.

La primera base tiene la ventaja de incorporar los atributos de los ítems, los cuales pueden tener relevancia a la hora de analizar cuáles son los factores que más influyen

en la elección y tráfico de ítems, sin embargo, no contiene información de las personas que no vieron ítems.

Por otro lado, la segunda base incorpora la cantidad de ítems que vio la persona, pero considerando a las personas inscritas en el periodo que no vieron beneficios. Si bien se pierde la información asociada a los ítems, se obtiene más información de las personas que puede permitir una mejor identificación de perfiles y la búsqueda de los factores que influyen en la conversión de las personas.

También para el análisis por palabras asociado a la descripción y nombre de los ítems solo se consideraron las palabras que aparecieran en más de un ítem, con esto se deja solo las palabras que sirven para realizar relaciones entre los ítems y se disminuye considerablemente la cantidad de variables facilitando los cálculos.

4.4 ANÁLISIS INCIDENCIA VARIABLES EN LA CONVERSIÓN

Para analizar la conversión de los usuarios primero se consideraron los datos de todas las personas, indicando si estas habían visto algún ítem, cuántos y cuántas veces.

Variables	Rango	Nulos	Promedio	Desviación estándar	Histograma
Ítems vistos	[0,501]	0	0.82	1.58	▇▁▁▁
Interacciones	N	0	1.95	6.01	▇▁▁▁

Tabla 3: Variables numéricas de conversión.

Como se observa en la *Tabla 3* se tiene en promedio las personas ven menos de un ítem, sin embargo, el promedio de interacciones es más del doble que el de ítems vistos. De esto se puede deducir que los usuarios ven pocos ítems, pero que suelen interactuar múltiples veces con los ítems que visualizan.

Luego se realizó una exploración gráfica de los datos, para identificar cuales variables y categorías tienen incidencia en la conversión de los usuarios.

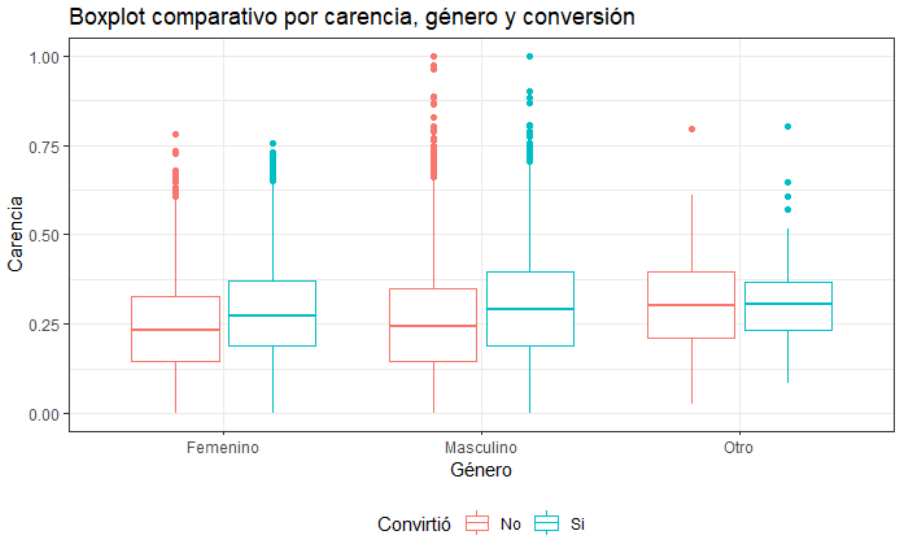


Ilustración 9: Comparación carencia y género.

Como se observa en la *Ilustración 9* existe una relación en la conversión con la carencia, el cual además presenta diferencias entre los géneros, teniendo una diferencia más notoria de carencia en la conversión para el género masculino.

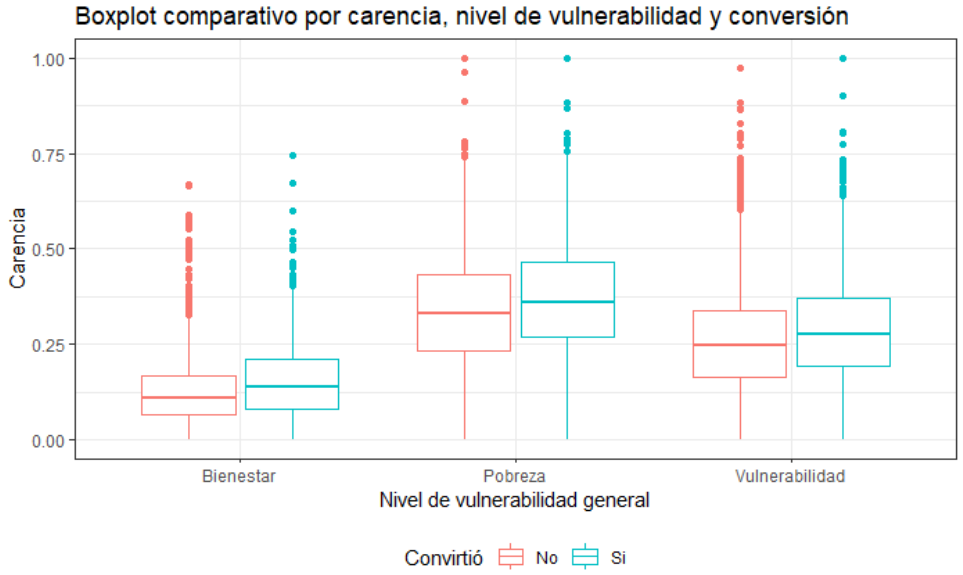


Ilustración 10: Comparación carencia y vulnerabilidad económica.

También, al realizar una comparación de la carencia por los niveles de vulnerabilidad económica se evidencia una clara diferencia entre ellos, teniendo que a mayor nivel de ingresos menor es la carencia. Además, en los niveles más vulnerables se hace más evidente la diferencia de carencia entre los que convierten y los que no.

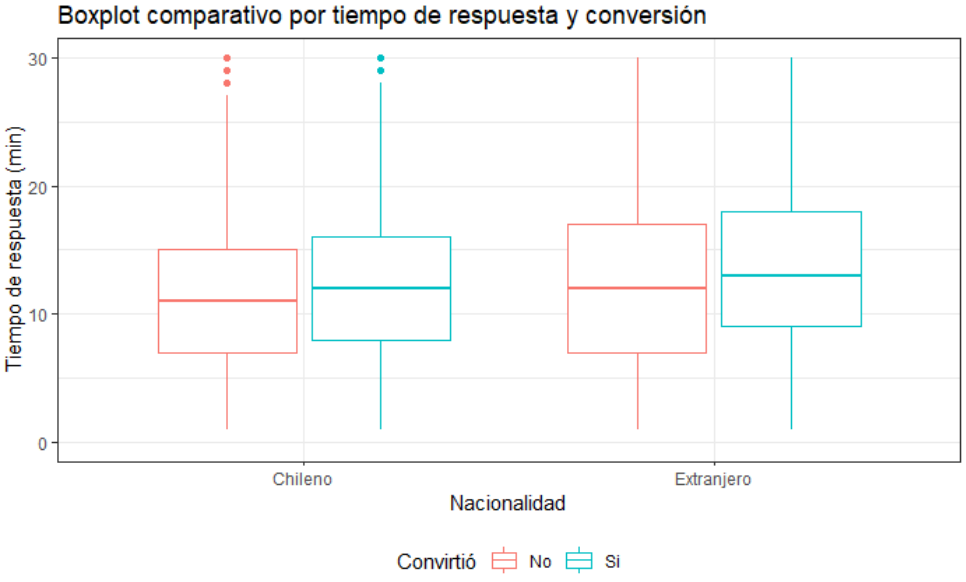


Ilustración 11: Tiempo de respuesta y nacionalidad

Adicionalmente, se observa en la *Ilustración 11* que al hacer una comparación en los tiempos de respuesta entre las nacionalidades buscando ver alguna relación con la conversión se ve una mayor desviación para los extranjeros, siendo el tiempo de los que convierten ligeramente mayor que el de aquellos que no convierten en ningún ítem.

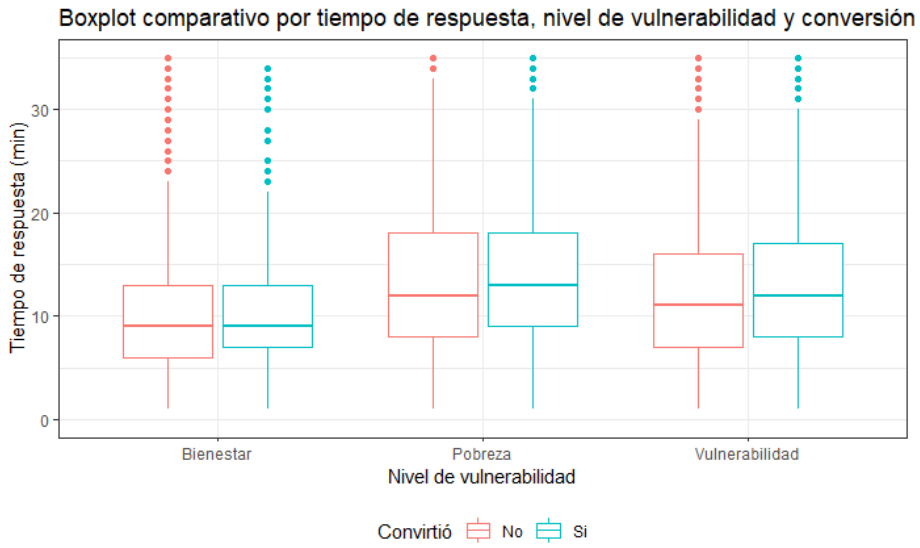


Ilustración 12: Tiempo de respuesta y vulnerabilidad económica.

Como se observa en la *Ilustración 12* , al realizar una comparación del tiempo de respuesta por los niveles de vulnerabilidad económica se evidencia una clara diferencia entre ellos, teniendo que a mayor nivel de ingresos, menor es el tiempo de respuesta. Además, en los niveles más vulnerables se hace más evidente la diferencia del tiempo de respuesta entre los que convierten y los que no.

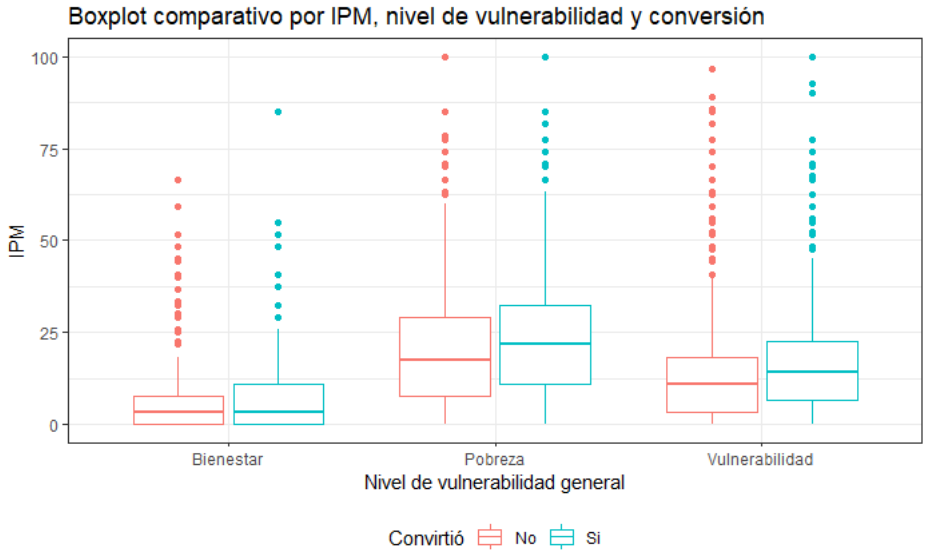


Ilustración 13: IPM y vulnerabilidad económica.

Al realizar la misma comparación por *IPM* como se ve en la *Ilustración 13* igualmente se evidencia una clara diferencia entre los niveles de vulnerabilidad económica, teniendo que a mayor nivel de ingresos menor es el *IPM*. Además, en los niveles más vulnerables se hace más evidente la diferencia del *IPM* entre los que convierten y los que no.

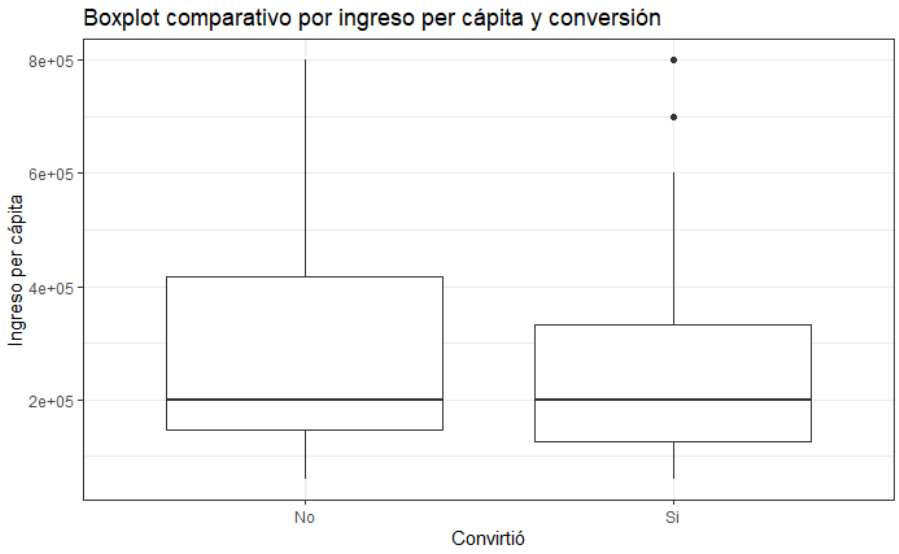


Ilustración 14: Ingreso per cápita y conversión.

Sin embargo, al hacer una comparación directa del ingreso y la conversión, no se hace evidente la relación del ingreso y la conversión.

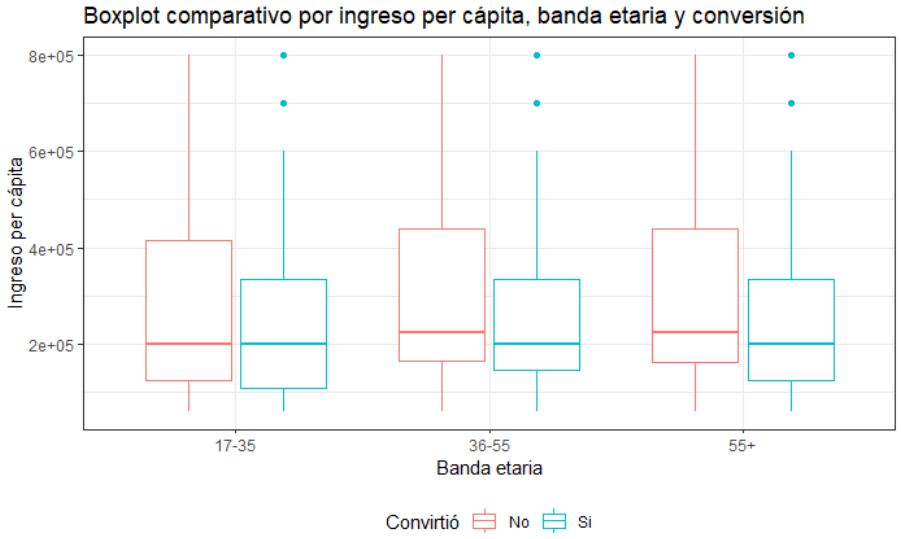


Ilustración 15: Comparación edad e ingreso.

Por otro lado, si se hace una comparación por ingreso, separando por categorías como en la *Ilustración 15* vuelve a observarse una relación entre el ingreso y la conversión, sin embargo, esta diferencia es notoria para los mayores de 36 años, teniendo en las generaciones más jóvenes menos diferencia de ingreso entre los que convierten y los que no.

Al realizar una comparación de las proporciones de cada categoría que convierten se puede observar que atributos categóricos tienen incidencia en la conversión de los usuarios en ítems.

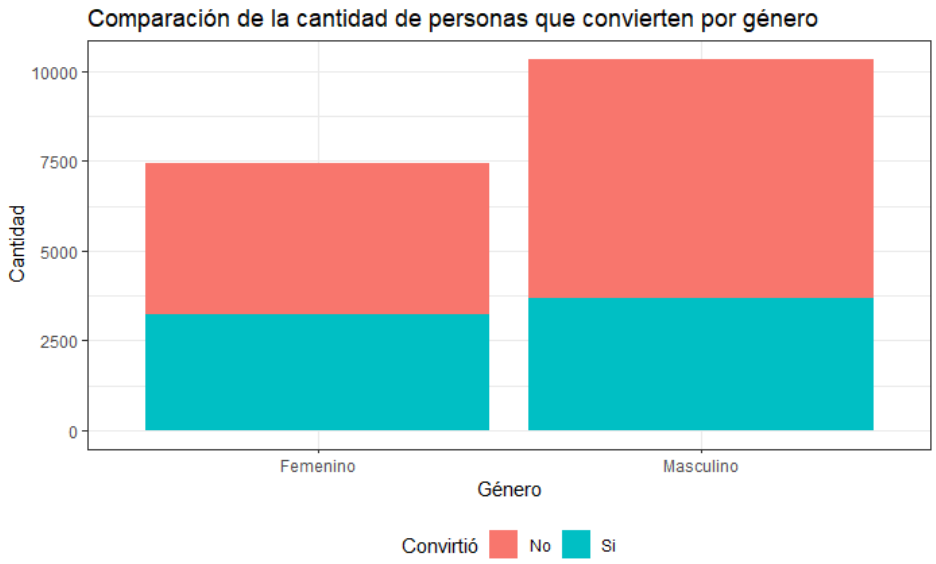


Ilustración 16: Conversión por género.

Al comparar el atributo de género en la *Ilustración 16*, se puede ver una clara diferencia de las proporciones de la muestra que converge de acuerdo con el género. Teniendo que en promedio el género femenino es más propenso a interactuar con algún ítem que el género masculino.

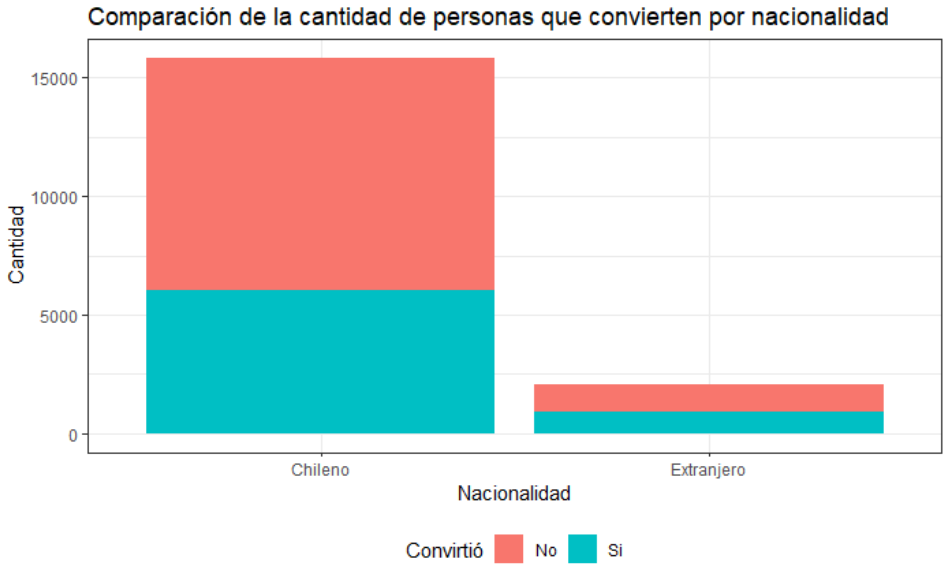


Ilustración 17: Conversión por nacionalidad.

En la *Ilustración 17* se ve que los extranjeros convierten en mayor proporción que los chilenos en ítems, sin embargo, el tamaño de la muestra de extranjeros es considerablemente menor que la muestra de chilenos.

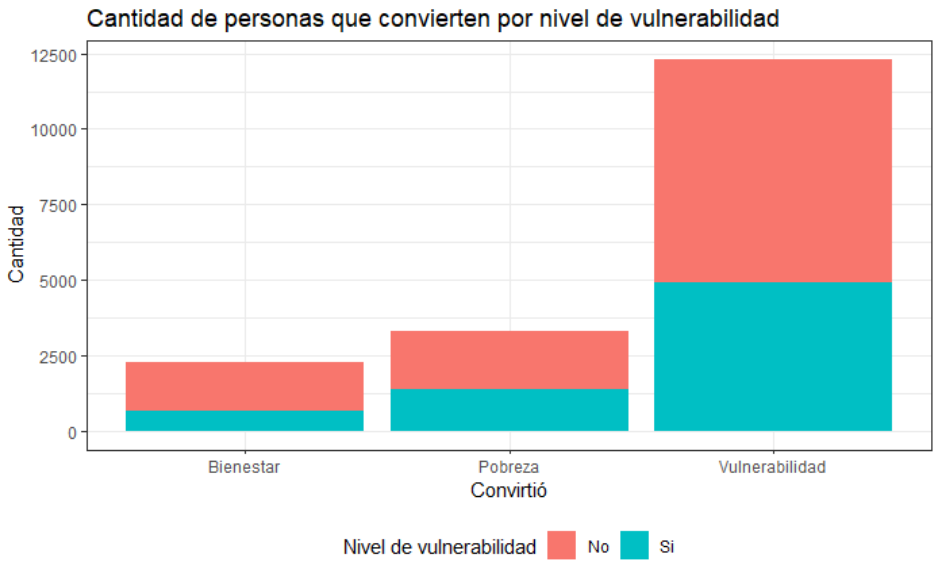


Ilustración 18: Conversión por vulnerabilidad económica.

Al realizar la comparación de las proporciones de acuerdo con la vulnerabilidad económica, como se puede ver en la *Ilustración 18* a mayor vulnerabilidad mayor es la proporción de los usuarios que interactúa con algún ítem.

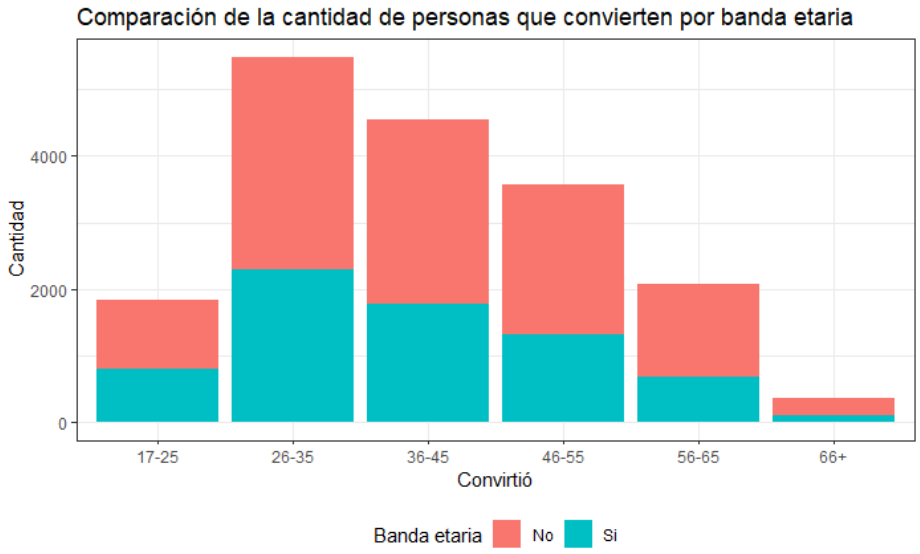


Ilustración 19: Conversión por rango etario.

La última comparación de proporciones se realizó comparando los rangos etarios en la *Ilustración 19* donde se observa que a medida que aumenta la edad de los usuarios, disminuye la proporción de estos que visualiza beneficios. Esto podría deberse a que las generaciones más jóvenes tienen un mayor manejo de la tecnología o también que por ser más jóvenes llevan menos tiempos insertos en el campo laboral y no tienen recursos acumulados.

Habiendo explorado las características de los usuarios que inciden en la conversión es necesario igualmente identificar las características de los ítems que tienen incidencia en la conversión de los usuarios.

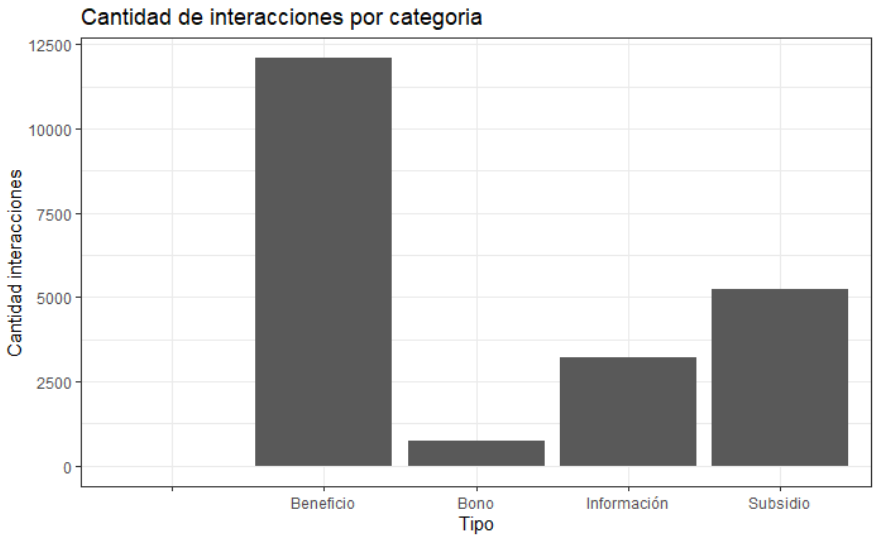


Ilustración 20: Interacciones por tipo de ítem.

En este ámbito, la primera observación que se puede realizar comparando la *Ilustración 5* y la *Ilustración 20* es que los ítems de tipo subsidio tienen la proporción más grande de interacciones en función la cantidad de ítems que clasifica en esa categoría. Esto es un indicio que el tipo de beneficio si puede influenciar la conversión y que no necesariamente por haber más ítems de una categoría, esta tendrá mayor cantidad de interacciones.

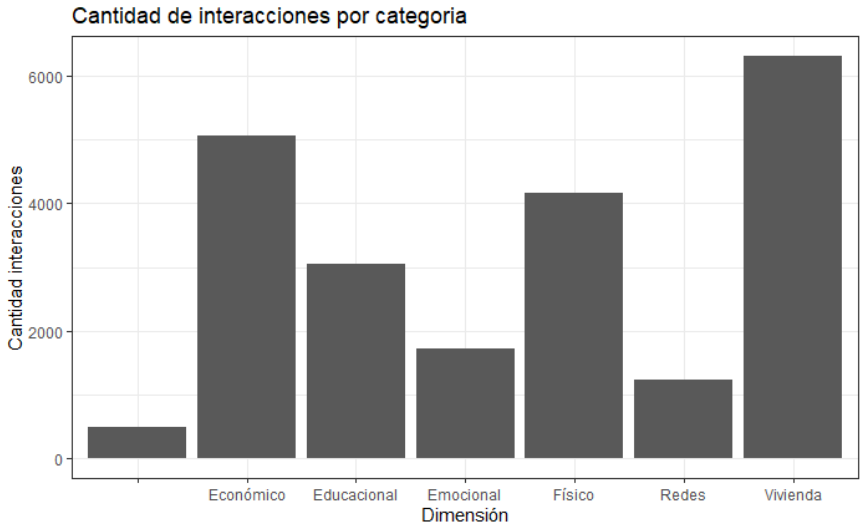


Ilustración 21: Interacciones por dimensión ítem.

Por último, en la *Ilustración 21* se ve que la mayor cantidad de interacciones se tiene para ítems pertenecientes a la dimensión de vivienda, lo cual contrasta con la cantidad de ítems que se observa en la *Ilustración 6*, pues en la dimensión con menor cantidad de ítems. También se tiene que los ítems que no tienen una dimensión asociada tienen menor cantidad de interacciones, lo cual era de esperarse pues estos no son recomendados en la página principal de la plataforma SemSo, sino que solo se pueden acceder por búsqueda. Esto implica que hay usuarios que están buscando ítems con los que desean interactuar pero que no le son recomendados por el sistema de recomendación actual.

5 SISTEMAS DE RECOMENDACIÓN

Para seleccionar y crear el sistema de recomendación para la plataforma *SemSo* lo primero es considerar que estos se basan en diferentes supuestos y requieren de algún tipo de información, esta información varía tanto en su procedencia como en su estructuración. Si bien, ya se explicaron los datos disponibles y se realizó un análisis exploratorio de estos, es relevante mencionar que gran parte de los modelos se basan principalmente en el feedback para entregar las recomendaciones, sin embargo, también utilizan pueden otro tipo de información como puede observarse en el siguiente diagrama.

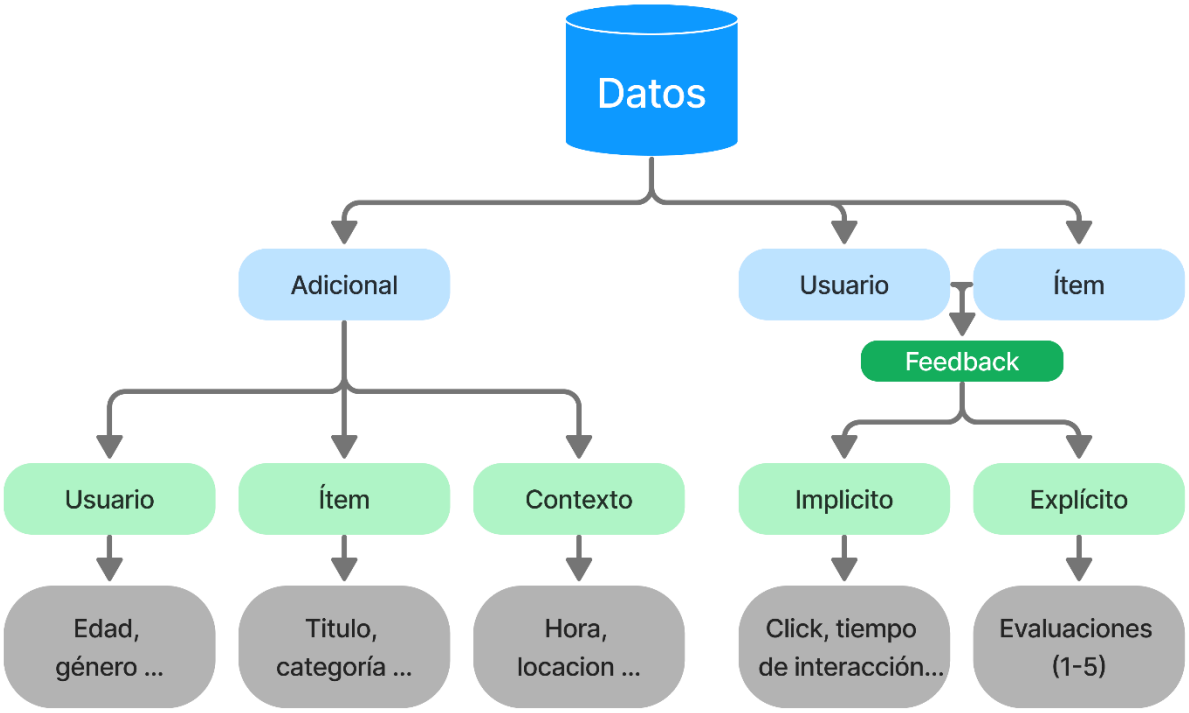


Ilustración 22: Diagrama obtención de datos.

Usualmente tanto los algoritmos de filtrado colaborativo como los algoritmos de filtrado basado en contenido utilizan como base el feedback explícito para realizar las recomendaciones, pero con la incorporación de información asociada a los ítems en el caso del filtrado basado en contenido. El resto de la información suele usarse para sistemas de recomendación avanzados o como información adicional para mejorar los sistemas mencionados.

La data utilizada afecta directamente en la capacidad de recomendación de las diferentes metodologías, usualmente se presenta el problema de *arranque en frío*, este se da por la falta de información asociada a usuarios nuevos, ítems nuevos o por la escasez de datos. Para el *arranque en frío* por usuarios nuevos suele usarse como recomendación los ítems más populares, técnicas de *embedings* o filtrado basado en ítems preguntando por las preferencias. Para el *arranque en frío* por ítems nuevos suele agregarse una selección de ítems nuevos, usar filtrado basado en contenido o incorporar de manera aleatoria el ítem en alguna recomendación. Para el arranque en frío por escasez de datos se usan técnicas avanzadas como la utilización de información contextual, *aprendizaje por transferencia* o *Hashing Trick*.

Para ver cómo enfrentar estos problemas en el contexto de la plataforma lo primero es analizar estos considerando los tipos de usuario que interactúan en esta.

5.1 TIPOS DE USUARIOS

Considerando la información disponible de los usuarios estos pueden ser agrupados en dos categorías, usuarios nuevos y usuarios activos, donde la diferencia de información disponible permite la aplicación de técnicas de recomendación diferentes.

Por un lado, los usuarios nuevos en el contexto de la plataforma de *SemSo* tienen más información disponible que en otras aplicaciones de sistemas de recomendación, ya que para tener acceso a la plataforma se debe completar la encuesta. Esto implica que para los usuarios nuevos se tiene información de caracterización e información de su contexto, pero no se tiene feedback de estos.

Por otro lado, se considera como usuarios activos aquellos que hayan interactuado con ítems de manera tal que se dispone feedback asociados a estos. Este feedback puede ser explícito como puede ser la evaluación de ítems o puede ser implícito considerando las interacciones con diferentes ítems.

Dadas las características descritas tanto el filtrado colaborativo como el filtrado basado en contenido no cuentan con información suficiente para ser implementados para los usuarios nuevos, sin embargo, la presencia de la información contextual permite implementar algún modelo que busque mejorar la recomendación realizada. Para este segmento se explorará realizar clústeres de usuarios y recomendar los ítems más populares dentro de estos, bajo el supuesto de que personas parecidas tendrán gustos parecidos. Además, para tener una referencia de comparación se evaluará un filtrado basado en la encuesta y los pesos utilizados actualmente.

Si bien los usuarios cuentan con feedback esto no implica que lo adecuado sea utilizarlo para realizar recomendaciones ya que se debe considerar la escasez de datos que puede existir tanto por la cantidad de personas con determinado feedback o por no tener suficientes feedback para relacionar a los usuarios.

Al observar la cantidad de datos disponibles de feedback explícito se observa que no se tienen datos suficientes para realizar modelos ya que solo hay 278 usuarios que evaluaron ítems, más aún hay ítems que solo fueron evaluados por un usuario y la mayor parte de los usuarios solo evaluaron un ítem. Sin embargo, al evaluar el uso de feedback implícito donde se observa alrededor de 6.400 usuarios que interactuaron con ítems, representando alrededor de un 35% de los usuarios que ingresaron a la plataforma, siendo una cantidad razonable de datos como para entrenar un modelo y una representación significativa de la muestra de usuarios.

Considerando lo recién mencionado para los usuarios activos se evaluará el uso filtrado basado en contenido y filtrado colaborativo utilizando feedback implícito. Cabe destacar que para hacer la relación entre los ítems para el filtrado basado en contenido se utilizará la información de las etiquetas asociadas a los temas de la encuesta y de los textos descriptivos de estos, y para el filtrado colaborativo se evaluará la utilización de factorización matricial no negativa y la factorización matricial no negativa de doble descomposición de valores singulares, pudiéndose observar los sistemas de recomendación a evaluar en el siguiente diagrama.

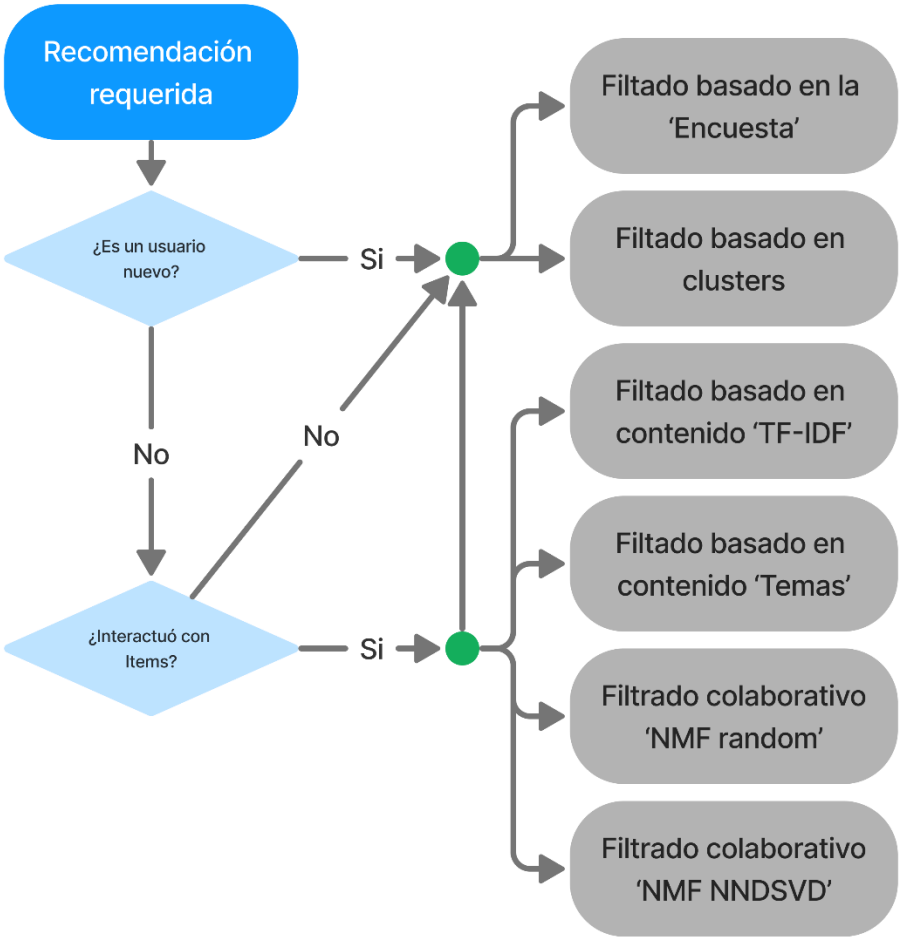


Ilustración 23: Diagrama sistemas de recomendación a evaluar.

5.2 FILTRADO BASADO EN LA ENCUESTA

A modo de referencia se desea tener un punto de comparación del rendimiento del sistema propuesto, para esto se busca una comparación con el sistema utilizado, sin embargo, no se sabe exactamente que ítems fueron recomendados por varios motivos. Primero existe un factor aleatorio en las recomendaciones realizadas, segundo los atributos asociados con los ítems pueden variar con el tiempo, tercero la ponderación de los temas del panel de experto ha variado desde su primera implementación, cuarto los ítems disponibles cambian con el tiempo y quinto existe información utilizada por el modelo actual que no se encuentra disponible en el repositorio utilizado, como exclusividad de ítems y evaluación de interna de los ítems.

Debido a esto se ha planteado un sistema alternativo en función de los pesos de los temas basado en el funcionamiento de un filtrado basado en contenido, el cual en vez de usar las evaluaciones de ítems para calcular ponderaciones de preferencia de las categorías utiliza la encuesta y luego usando la similitud coseno entrega los ítems a recomendación.

Esto implica que, a diferencia del sistema de recomendación utilizado, este no solo considera la ponderación del tema con mayor puntaje de cada intersección ítem-usuario, sino que considera una ponderación por todos los temas asociados a la intersección ítem-usuario.

Para evaluar el rendimiento del sistema se utilizó la muestra de usuarios que interactuaron con algún ítem y se seleccionó aleatoriamente una de las interacciones de cada usuario. Dado que la plataforma recomienda 6 ítems se puede considerar que si el sistema recomienda el ítem seleccionado entre estos, el modelo hizo una predicción correcta. Con esta metodología el filtrado basado en la encuesta logro un *accuracy* promedio de 0,004 con una desviación estándar de 0,00968.

5.3 FILTRADO BASADO EN CLÚSTER

Para la construcción del sistema de recomendación propuesto se parte de los sistemas de recomendación basados en vecindarios (NB), los cuales suponen que usuarios similares prefieren ítems parecidos o que ítems parecidos son preferidos por usuarios similares. (Ricci, Rokach, & Shapira, 2015) Además, el sistema de recomendación busca enfrentar el problema de arranque en frio de los métodos evaluados.

Considerando lo anterior se desea hacer agrupaciones de los usuarios y de los ítems, sin embargo, como se mencionó anteriormente debido a la escasez de datos para los usuarios nuevos no es posible usar el método habitual de asociación por las evaluaciones y ni factores implícitos, lo cual deja la información demográfica para hacer la agrupación de los usuarios y los atributos del contenido para hacer la agrupación de los ítems.

Con la agrupación de los usuarios se busca mejorar la recomendación realizada a los usuarios nuevos, recomendándoles ítems vistos por usuarios similares a ellos. En cambio, con la agrupación de ítems se busca poder tener mayor diversidad en las recomendaciones y poder recomendar ítems nuevos. Al hacer la agrupación de los ítems por una parte se están juntando ítems muy parecidos, por lo que recomendando de agrupaciones diferentes se espera tener mayor diversidad y también, si un ítem deja de estar disponible sigue aportando información de gustos de los usuarios contribuyendo a que se muestren otros ítems parecidos a este por pertenecer a la agrupación.

5.3.1 CLUSTERING USUARIOS

Para realizar el clúster de los usuarios se empleó el método *K-Means*, para el cual se utilizaron solo variables cuantitativas, las cuales fueron la edad del usuario, su ingreso per cápita, su índice de carencia y su tiempo de respuesta.

Estas variables fueron seleccionadas en base al análisis exploratorio, considerando que no están correlacionadas y que con su incorporación se está entregando información respecto de las características demográficas de la persona, características de su situación multidimensional e información respecto a su comportamiento en la plataforma, pero sin incluir variables que puedan estar influenciadas por un *arranque en frío*, pues, se requiere que la persona conteste la encuesta para que pueda acceder a los ítems.

Luego, para decidir cuantos clústeres realizar se utilizaron tres métricas o metodologías, la primera corresponde al método de *WSS*, el segundo al método de *Silhouette* y el ultimo al método de *diferencia estadística*.

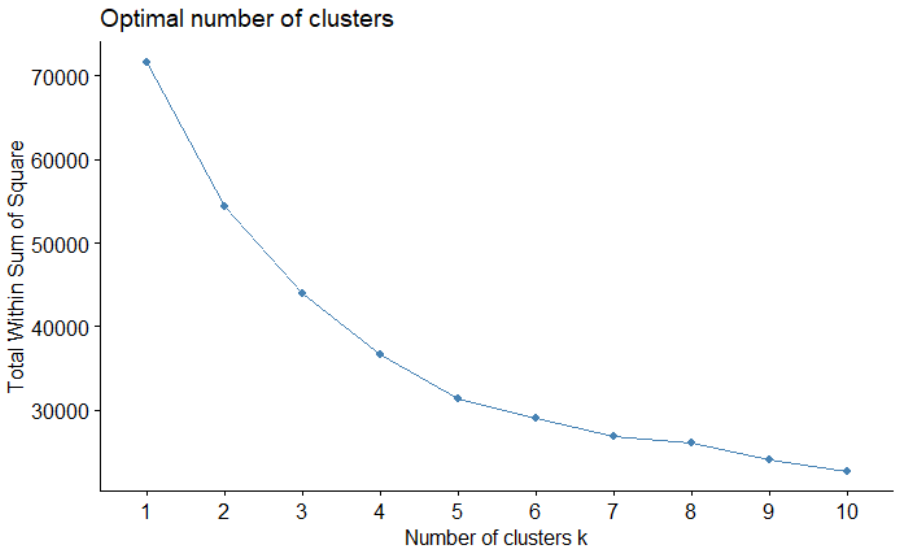


Ilustración 24: WSS para K-Means usuarios.

Como se observa en el gráfico de la *Ilustración 24* correspondiente al *WSS* se tiene el “codo” o punto de infección para 5 clúster, por lo que de acuerdo a esta metodología lo recomendado es realizar 5 clúster.

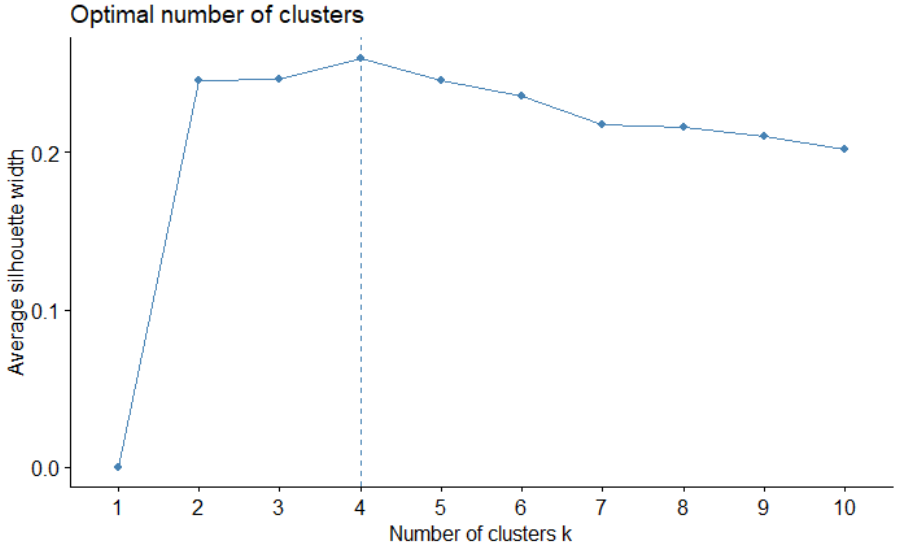


Ilustración 25: Average silhouette width para K-Means usuarios.

Por otra parte, de acuerdo al coeficiente Silhouette graficado en la *Ilustración 25* se tendrá la mayor distancia promedio entre los clústeres si se realizan 4 clúster, por lo que esta metodología recomienda que se realicen 4 clúster.

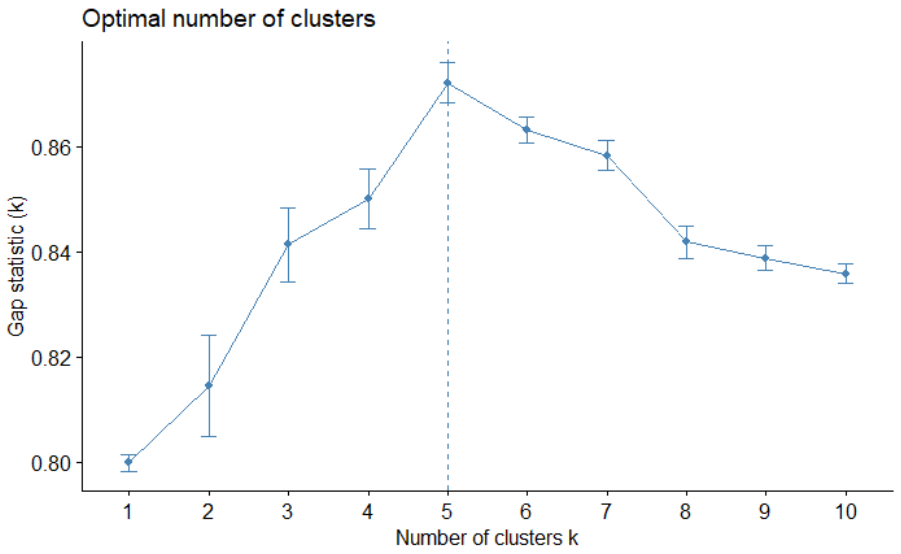


Ilustración 26: Gap statistic para K-Means usuarios.

Por último, de acuerdo a la *Ilustración 26* se tendría la mayor dispersión de los puntos respecto de un caso aleatorio para 5 clúster, pues este maximiza la diferencia entre la dispersión observada y la esperada por el caso aleatorio. Considerando los resultados de las tres metodologías se decidió realizar 5 clúster de los usuarios.

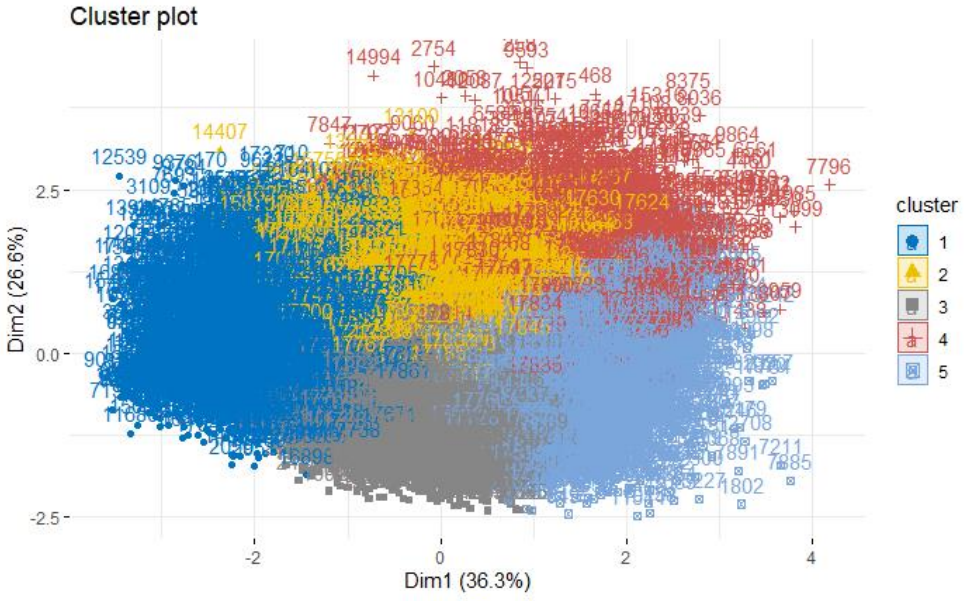


Ilustración 27: Clúster dimensiones K-Means usuarios.

El modelo logró explicar sobre el 80% de la varianza con tres dimensiones, sin embargo, ya con dos dimensiones se es capaz de observar una separación entre los espacios de los diferentes grupos.

Como se puede observar en la *Ilustración 28: Carencia K-Means* usuarios. Existen claras diferencias en los niveles de carencia de los clústeres, pudiendo caracterizarse el primer clúster por un mayor nivel de bienestar y el quinto clúster como el grupo con mayor carencia.

Ilustración 28: Carencia K-Means usuarios.

Adicionalmente de los gráficos presentes en el *Anexo E: Caracterización clúster personas* de la se puede ver que los grupos uno y cinco tienen en promedio menos interacciones, siendo además estos grupos los menos carentes. También se puede observar que los clústeres uno y tres presentan una mayor proporción de hombres que los otros grupos; el clúster uno se caracteriza por tener usuarios de mayor edad; el clúster cuatro se caracteriza por tener mayores tiempos de respuestas; el clúster 2 por tener personas más jóvenes y el clúster tres por tener menor ingreso per cápita.

5.3.2 CLUSTERING ÍTEMS

Para la realización de del clúster de ítem se utilizó la misma metodología utilizada para los clústeres de usuarios, utilizando los temas utilizados en el filtrado basado en la encuesta y usando una matriz *TF-IDF* construida con las palabras procesadas del nombre y descripción de los ítems.

Al evaluar la cantidad de grupos utilizando las metodologías de *WSS*, coeficiente de silueta y diferencia estadística utilizando los temas, cuyos resultados se pueden observar en el *Anexo F: WSS para K-Means por tema*., *Anexo G: Coeficiente Silhouette para K-Means por temas*. y *Anexo H: Diferencia estadística para K-Means por temas*. respectivamente se consideró la utilización de 4, 12, 21, 27, 31 y 34 agrupaciones. Luego, considerando que se busca la mayor cantidad de agrupaciones posible para poder tener más variabilidad en las recomendaciones se determinó utilizar 34 clústeres.

Repitiendo las metodologías anteriores para la matriz *TF-IDF* se obtuvo los resultados presentes en el *Anexo I: WSS para K-Means por TF-IDF*.,

Anexo J: Coeficiente Silhouette para K-Means por TF-IDF. y *Anexo K: Diferencia estadística para K-Means por TF-IDF*., de estos se evalúo utilizar 2, 13 y 18 clústeres, determinándose utilizar 18 clústeres.

5.3.3 RESULTADOS

En función de las agrupaciones anteriores se procedió a calcular para cada clúster de los usuarios la agrupación de ítems más vistas y los ítems más vistos por cada clúster de usuarios, el resultado de las recomendaciones realizadas utilizando la asociación por las etiquetas de temas puede observarse en la *Tabla 4. S*in embargo, el resultado de las recomendaciones utilizando la asociación por la matriz *TF-IDF* no presentó variabilidad entre los grupos por lo que no representaba una mejora su aplicación, una posible explicación a estos resultados es que al ser menor la cantidad de clústeres se estén concentrando los ítems relevantes en los clústeres o que las palabras no permitan determinar las preferencias.

SR Clúster		Clúster				
	Ítems	1	2	3	4	5
Recomendación	1	157	790	790	790	790
	2	57	797	793	794	797
	3	207	5	794	207	5
	4	156	22	207	793	776
	5	776	776	5	5	4
	6	760	55	22	22	55

Tabla 4: Recomendaciones por clúster.

El resultado de utilizar estas recomendaciones aplicando la misma metodología para el cálculo de su efectividad que para el filtrado basado en la encuesta, fue un *accuracy* promedio de 0,081 y una desviación estándar 0,21464, sin embargo, logró un *accuracy* máximo de 0,61 para una muestra aleatoria.

De la *Tabla 4* se puede observar que para el clúster 1 se tienen más recomendaciones únicas que para los otros clústeres, siendo este clúster caracterizado por las personas menos carentes, con mayor ingreso y por tanto por un mayor nivel de bienestar, coincidentemente en el análisis de conversión se había observado que los usuarios en situación de bienestar son los con menores tasas de conversión. Si bien para los otros clústeres no presentan tantas recomendaciones únicas, de igual manera se observan diferencias entre las recomendaciones realizadas a cada uno de ellos.

5.4 FILTRADO BASADO EN CONTENIDO

Como se mencionó anteriormente para el filtrado basado en contenido se evalúa una asociación entre los diferentes ítems utilizando la información de los temas de las preguntas y de sus descripciones. Luego se realiza una recomendación en función de las interacciones pasadas del usuario.

Sin embargo, de las interacciones se puede rescatar más información que solo si una persona accedió a un ítem, ya que un usuario puede interactuar varias veces con un mismo ítem y se supondrá que si un usuario interactúa varias veces con un mismo ítem es porque presenta mayor interés en este. Para reflejar esto en las recomendaciones se construyó una matriz de preferencias la cual hace una ponderación por la cantidad de interacciones, donde las filas son usuarios y las columnas son ítems. Esta toma el valor

cero cuando no ha ocurrido una interacción usuario-ítem y toma el valor uno más *alpha* por el número de interacciones.

A modo de ejemplo si un usuario interactuó cinco veces con un ítem y se usa un *alpha* de 0,2 se tendría el valor 2 para ese usuario-ítem.

También se debe considerar la cantidad de interacciones mínimas con diferentes ítems que un usuario tiene, puesto que a medida que las interacciones aumentan se puede rescatar más información de las preferencias del usuario, pero a medida que se tiene un filtro más alto disminuye la cantidad de usuarios de la muestra. Cabe destacar que se parte de la base de los usuarios que tienen al menos una interacción, sin embargo, para poder calcular las métricas se necesita al menos dos interacciones pues una de estas es eliminada de la muestra para ser usada como comparación de la predicción del modelo.

Adicionalmente el mínimo de interacciones seleccionada impacta directamente en la cantidad de información disponible para entrenar el modelo, lo cual puede evidenciarse en la *Tabla 5*: Cantidad de usuarios por muestra.

	Interacciones mínimas con ítems diferentes						
	0	1	2	3	4	5	6
Muestra	18.228	6.387	3.109	1.546	852	490	312

Tabla 5: Cantidad de usuarios por muestra.

Esto implica que para el análisis de sensibilidad de tienen dos variables que se pueden modificar, el *alpha* de la ponderación de las interacciones y la cantidad mínima de interacciones con diferentes ítems.

5.4.1 ASOCIACIÓN POR TEMAS

Los principales resultados de sensibilidad del filtrado basado en contenido realizando los cálculos de similitud entre los ítems en función de los temas asociados pueden observarse en la *Tabla 6*, teniéndose los mejores resultados para todos los valores de filtrado con *alpha* de 0,4. Adicionalmente se obtuvo el mayor *accuracy* filtrando por que los usuarios hayan interactuado por lómenos con 4 ítems diferentes, pero con el filtrado mínimo para poder realizar la evaluación ya se obtiene mejores resultados que en los modelos anteriores.

SR CBF TEMAS		Interacciones mínimas con ítems diferentes				
	Accuracy	2	3	4	5	6
Alpha	0,2	0,233	0,271	0,12	0,151	0,045
	0,4	0,339	0,186	0,531	0,106	0,054
	0,6	0,231	0,185	0,122	0,04	0,048

Tabla 6: Filtrado basado en contenido por temas.

Por una parte, el *alpha* con mejores resultados puede atribuirse que con un *alpha* más pequeños se le entrega menor valor a al hecho de que un usuario interactúe varias veces con un mismo ítem dejando poca diferencia de ponderación entre los ítems que se interactuó, pero un *alpha* muy grande puede sobre estimar la relevancia que tiene una interacción repetida.

Por otra parte, se espera que con una mayor cantidad de interacciones se tenga más información de las preferencias del usuario, sin embargo, debido a la disminución de la

cantidad de datos de la muestra al filtrar por un mayor número de interacciones mínimas se da que pasado 4 interacciones diferentes disminuye el *accuracy* del modelo.

5.4.2 ASOCIACIÓN POR TF-IDF

Los principales resultados de sensibilidad del filtrado basado en contenido realizando los cálculos de similitud entre los ítems en función de las palabras de su nombre y descripción pueden observarse en la *Tabla 7,* donde se observa los mejores resultados con 2 interacciones y un *alpha* de 0,6.

SR CBF TF-IDF		Interacciones mínimas con ítems diferentes				
	Accuracy	2	3	4	5	6
Alpha	0,2	0,161	0,034	0,055	0,078	0
	0,4	0,151	0,039	0,102	0,012	0,003
	0,6	0,162	0,039	0,045	0	0,006

Tabla 7: Filtrado basado en contenido por TF-IDF.

Para todas las muestras se observó una menor capacidad predictiva que la esperada, especialmente considerando los resultados de la asociación por temas, esto puede deberse a varios factores tales como poca representatividad de las palabras para la asociación de los ítems, poca repetición de palabras entre diferentes ítems o falencias en el filtrado de palabras no relevantes que “ensucian” las relaciones.

5.5 FILTRADO COLABORATIVO

La herramienta más utilizada para el filtrado colaborativo es la factorización matricial explicada en el marco teórico, para su aplicación en el contexto de la plataforma SemSo se utiliza la misma matriz de preferencias utilizada para el filtrado basado en contenido.

Esto deja para el análisis de sensibilidad tres variables a modificar, el *alpha* asociado a las interacciones, el mínimo de interacciones con ítems diferentes de los usuarios y la dimensión de las matrices de factorización. Sin embargo, se decidió estudiar la capacidad predictiva de los modelos variando solo el *alpha* y las interacciones mínimas, dejando la dimensión de las matrices de factorización en 2.

5.5.1 NMF RANDOM

De la factorización matricial no negativa random se obtuvo los resultados de la *Tabla 8* en los cuales se observa el mejor *accuracy* para la muestra filtrando por un mínimo de 4 interacciones y con un *alpha* de 0,4.

SR NMF random		Interacciones mínimas con ítems diferentes				
	Accuracy	2	3	4	5	6
Alpha	0,2	0,145	0,433	0,473	0,596	0,474
	0,4	0,145	0,036	0,815	0,759	0,147
	0,6	0,145	0,037	0,083	0,594	0,387

Tabla 8: Filtrado colaborativo NMF random.

Adicionalmente se observa que para un mínimo de 2 interacciones el *alpha* no implico cambios en la capacidad predictora del modelo y para un mínimo de 3 interacciones se obtuvo el mejor resultado con el menor *alpha* evaluado. Esto puede deberse a que el *alpha* es atribuido a poder diferenciar si el usuario presenta mayor preferencia por alguno de los ítems con los que el usuario interactuó, entonces teniéndose mayor cantidad de

muestras donde no se tiene que discernir entre preferencias un mayor *alpha* puede sobrestimar basado en los usuarios que interactúan mayor cantidad de veces.

Al igual que para el filtrado basado en contenido, pasado 4 interacciones disminuye la capacidad predictiva del modelo, lo cual puede atribuirse a la disminución de la cantidad de información. Además, se tiene que ambos modelos de filtrado basado en contenido lograron mejores resultados para las muestras filtradas por 2 interacciones que las obtenidos por este modelo para estas muestras.

5.5.2 NMF NNDSVD

Al incorporar la utilización de SVD para la iniciación de la factorización matricial no negativa se obtuvo los resultados de la *Tabla 9*, los cuales para un 80% de las muestras fueron mejores que los resultados obtenidos por el método anterior. Sin embargo, una de las diferencias más relevantes fue la obtenida para la muestra de 3 interacciones y un *alpha* de 0,2, pues no solo es mayor que el de la metodología anterior, sino que también entrega mejores resultados que cualquiera de las muestras de los filtrados basado en contenido.

SR NMF nndSVD		Interacciones mínimas con ítems diferentes				
	Accuracy	2	3	4	5	6
Alpha	0,2	0,147	0,702	0,174	0,79	0,34
	0,4	0,151	0,036	0,883	0,692	0,06
	0,6	0,148	0,037	0,009	0,765	0,772

Tabla 9: Filtrado colaborativo NMF nndSVD.

Al igual que el modelo anterior su máximo *accuracy* lo obtuvo con un mínimo de 4 interacciones y con un *alpha* de 0,4, siendo en esta situación cerca de un 8% mejor. Además, tampoco logro mejores resultados que los filtrados basados en contenido para las muestras filtradas por un mínimo de 2 interacciones.

5.6 SISTEMA DE RECOMENDACIÓN PROPUESTO

Ambos sistemas de recomendación evaluados para los usuarios nuevos lograron bajos niveles de predicción, sin embargo, el modelo basado en clúster logró resultados considerablemente mejores que el modelo basado en la encuesta. Considerando esto a la hora de la construcción de un modelo híbrido para realizar las recomendaciones, no hace sentido integrar ambos ya que el modelo basado en la encuesta no estaría agregando valor a la recomendación.

Al evaluar los filtrados basados en contenido se obtuvo para todos los escenarios mejores resultados realizando la asociación por las etiquetas, lo cual descarta el uso del filtrado basado en contenido utilizando la matriz *TF-IDF*.

En cambio, para el filtrado colaborativo no se obtuvo en todas las iteraciones que un modelo fuese mejor que el otro, sin embargo, dado que para el 80% de las muestras y para sus máximos *accuracy* la factorización no negativa con iniciación *SVD* obtuvo un *accuracy* mayor se considera solo utilizar este.

También se obtuvo que para la muestra con al menos dos interacciones y un *alpha* de 0,4 un *accuracy* por filtrado basado en contenido por temas alrededor de un 50% mayor que el obtenido por por el modelo de clúster y más de un 100% mayor que los filtrados colaborativos. Sin embargo, al considerar un mínimo de cuatro interacciones el

filtrado colaborativo logra un *accuracy* cerca de un 50% mayor que el filtrado basado en contenido para esa muestra.

Como se elimina una interacción de la muestra para generar la muestra de testeo y evaluar los modelos, se tiene que al filtrar por un mínimo de dos interacciones se está usando una sola interacción para predecir las recomendaciones de determinados usuarios y equivalentemente para el mínimo de cuatro interacciones utiliza tres interacciones para realizar las recomendaciones.

Considerando lo anterior tiene sentido hacer una separación por la cantidad de interacciones partiendo por el modelo de clúster para los usuarios nuevos, pasando a filtrado basado en contenido una vez que los usuarios interactúen con un ítem. Posteriormente, una vez que un usuario haya interactuado con al menos tres ítems se pasa al modelo de filtrado colaborativo de manera de optimizar el *accuracy* minimizando la cantidad de información a procesar, pues para ambos filtrados se utiliza la matriz de preferencias con un *alpha* de 0,4. Este sistema puede observarse en el siguiente diagrama.

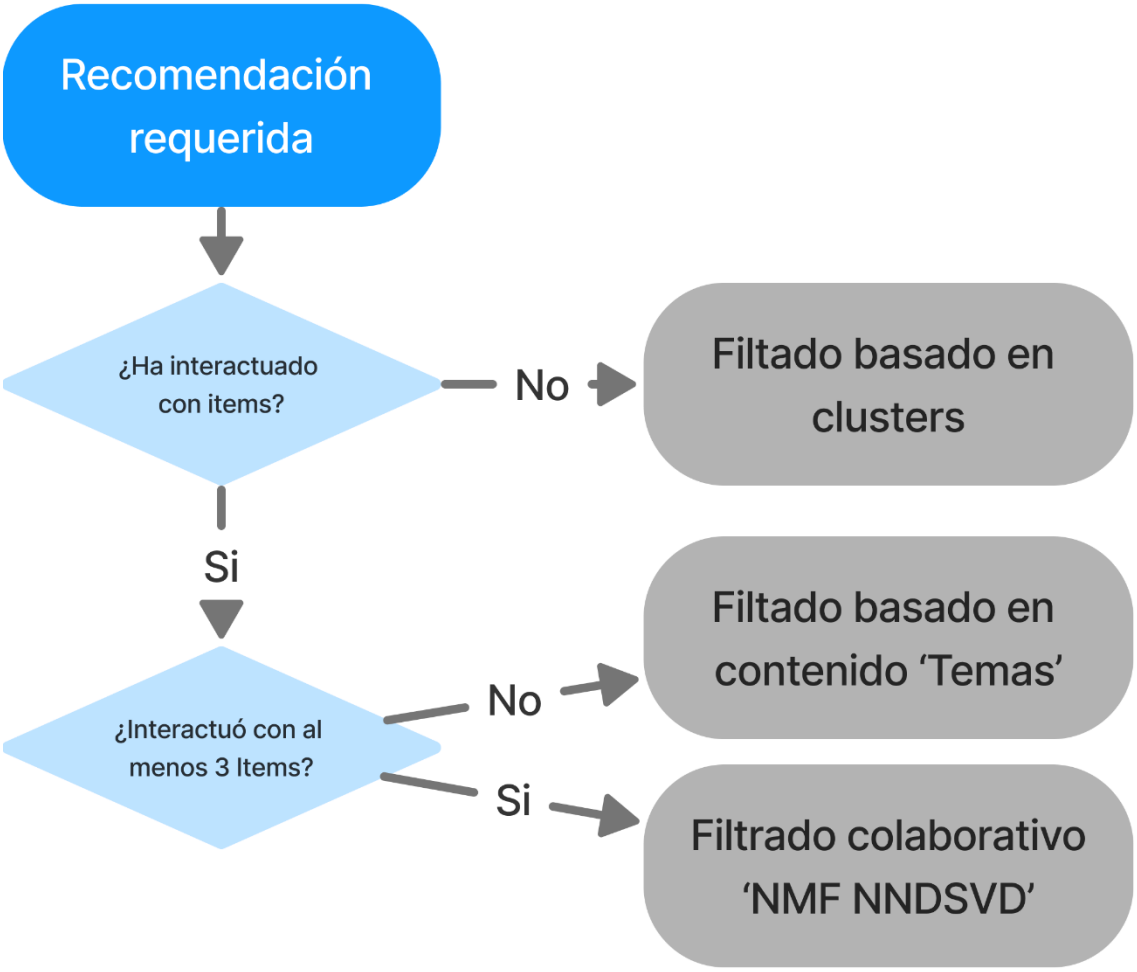


Ilustración 29: Diagrama sistema de recomendación propuesto.

Cabe destacar que la integración del filtrado basado en contenido entre el filtrado basado en clúster y el filtrado colaborativo, permite que el sistema pueda recomendar ítems nuevos para alguna muestra de los usuarios.

6 IMPLEMENTACIÓN

6.1 PLAN DE ACCIÓN

Antes de implementar el sistema de recomendación propuesto es necesario comprobar que este funcione correctamente y para poder validar que tenga un impacto se debe poder comparar sus resultados. Sin embargo, utilizar los datos históricos para comparación puede generar sesgos. Considerando lo anterior se proponen los siguientes pasos a seguir.

1. Crear un entorno de prueba; con esto se busca poder realizar pruebas del sistema de recomendación sin interferir el funcionamiento habitual de la plataforma, para lo cual se puede generar entorno cerrado que simule el funcionamiento de la plataforma.
2. Adaptar el modelo a la base de datos del backend; dado que la base de datos de la cual se alimenta la plataforma es diferente a la utilizada para el análisis es necesario realizar una adaptación del input de la información de manera que se pueda usar la información en tiempo real y no depender de la clonación de la información.
3. Testear el rendimiento del modelo; esto implica desde un análisis exploratorio del sistema probando diferentes situaciones, ajustando ponderaciones especiales que se consideren relevantes para el negocio, evaluando intuitivamente que las recomendaciones sean las esperadas para cada situación y que los tiempos de carga sean adecuados para un correcto funcionamiento de la plataforma.
4. Testear el nuevo modelo en muestras de prueba; antes de implementar en muestras de usuarios, buscar una muestra de personas a quienes se les realice la encuesta, mostrándoles ambas recomendaciones (algoritmo utilizado y algoritmo propuesto) y que se les pregunte cuál de las de las recomendaciones prefiere. Con esto se busca evaluar en un entorno más realista que el algoritmo propuesto no sea peor percibido por los usuarios que el actual.
5. Implementar el nuevo modelo en muestras aleatorias; si el paso anterior es satisfactorio se procede a mostrar de manera aleatoria a los usuarios nuevos alguno de los algoritmos de recomendación, siendo constante en el sistema que se les muestra, con esto se construirán dos muestras aleatorias de usuarios sometidos a los diferentes sistemas.
6. Evaluar los resultados; una vez que se tengan suficientes datos, realizar pruebas estadísticas sobre las muestras utilizando las métricas de rendimiento de manera de determinar si existe un cambio significativo, para luego establecer el sistema con mejor rendimiento como el sistema de recomendación predeterminado.

En función de lo anterior se propone la siguiente *Carta Gantt*:

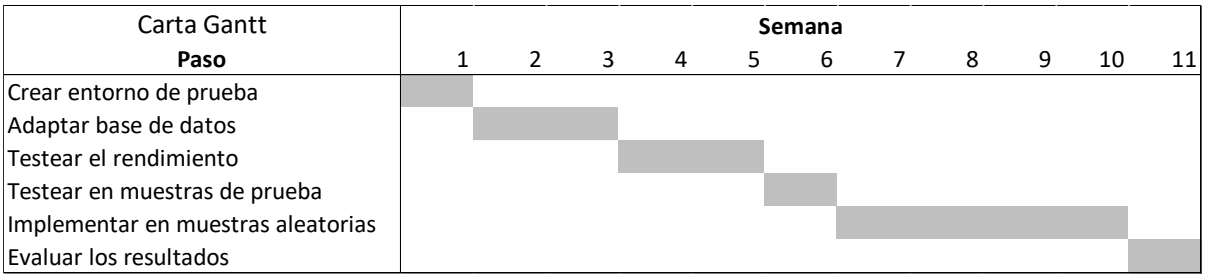


Ilustración 30: Carta Gantt plan de acción.

6.2 TRABAJO FUTURO

Los modelos presentados representan una base para poder seguir mejorando el sistema de recomendación, donde a medida que aumente la cantidad de interacciones se tendrán mejores resultados y también permitirá utilizar otros modelos para continuar mejorando los resultados.

En este ámbito entre las posibles mejoras que se pueden realizar es la integración de más información al sistema de recomendación, en este contexto existe información disponible que no fue utiliza, como la información georreferenciada de la vivienda. Esta podría permitir realizar agrupaciones por manzana, zona o comuna que pueden aportar más información del perfil del usuario.

La utilización de la georreferenciación también permite la integración de datos territoriales externos que están disponible como los datos del *CENSO* u otras encuestas de la *CASEN*, *Paz Ciudadana* u otra institución, estos datos pueden ser usados complementariamente para mejorar la agrupación de los usuarios.

Por otra parte, se dispone información de la empresa en la que trabajan los usuarios, por lo cual si se realiza una caracterización de las empresas se puede rescatar información relevante del contexto en el cual el usuario está inserto, como por ejemplo, se puede rescatar la industria en la que trabaja el usuario o el tamaño de la empresa.

Adicionalmente, una vez que se tengan más periodos de datos se pueden implementar sistemas basados en el contexto que incorporen información del periodo estacionario para mejorar las recomendaciones. Usualmente esta metodología es utilizada por *e-commerce* para mejorar las recomendaciones, priorizando artículos relacionados con “eventos” como si uno busca regalos cerca de San Valentín es razonable pensar que se recomienden artículos como chocolates o flores, en cambio si uno busca regalos cerca de navidad es más probable que se recomienden juguetes o ropa.

En el contexto de la plataforma SemSo igualmente hay estaciones que pueden tener impacto en las preferencias de los usuarios, por ejemplo, es razonable pensar que en marzo se soliciten más ítems relacionados con la educación ya que inicia el pago de aranceles o que en invierno se busquen más ítems relacionados con la salud pues suele aumentar la cantidad de resfríos o beneficios como el subsidio de gas para poder calentar los hogares.

Habiendo ya desarrollado la infraestructura para poder evaluar los sistemas propuestos, a futuro solo debe agregarse un paso de evaluación teóricas de las técnicas futuras propuestas e iterar sobre el plan de acción propuesto.

7 CONCLUSIONES

En función del análisis exploratorio de los datos y de los modelos se identificó que variables como el género, el ingreso y la edad son de las variables con mayor impacto en la conversión de los usuarios, sin embargo, la variable que más incidencia mostró en la conversión fue la carencia. Además, se vio una fuerte preferencia por ítems clasificados dentro de la dimensión de vivienda, en especial los caracterizados como subsidios, habiendo un alto interés por los ítems relacionados con el ámbito económico.

Adicionalmente, se logró identificar satisfactoriamente cinco grupos de usuarios con comportamiento y características similares, clusterizados a partir de la edad de los usuarios, sus ingresos per cápita, índices de carencia y tiempos de respuesta. A excepción de los clústeres tres y cuatro, el resto mostró tener gustos distintos basado en las diferencias de recomendación realizada por el modelo de clúster.

Sin embargo, los clústeres tres y cuatro de igual forma tienen diferencias en sus características donde el primero está compuesto por usuarios más jóvenes y con menores tiempos de respuestas que el segundo. Además, se caracterizó el clúster uno como el compuesto por los usuarios menos carentes, el clúster cinco por los usuarios más carentes y el clúster dos por los usuarios de mayor rango eterio.

Si bien con la *NMF* se logró una capacidad predictiva considerablemente mejor que lo obtenido por el algoritmo actual, esta predicción solo aplica para una proporción menor de los usuarios, cercana a un 8% de la muestra. Además, no tiene un impacto en la tasa de conversión y si bien su aplicación pudiese significar una mejora en la calidad del servicio para estos usuarios, por la proporción de usuarios que representa, su efecto en el *NPS* sería bajo.

Por otra parte, con el *FBC* la capacidad predictora fue menor que con la *NMF*, pero logra mejorar la recomendación para cerca de un 35% de la muestra y pese a que tampoco puede impactar en la conversión, es más probable que la mejora en la calidad de servicio que pueda proporcionar se vea reflejada en el *NPS*.

Pese a que el modelo de clúster no es el con mejor *accuracy*, presenta mejores resultados que los obtenidos por el modelo basado en la encuesta, por lo que teóricamente debería poder mejorar las tasas de conversiones en la plataforma *SemSo*.

Aunque se tiene bastantes muestras de información de usuarios que contestan la encuesta, la cantidad de datos para usuarios que interactúan o evalúan es considerablemente menor y posee una escasez de datos en el sentido de que hay una gran proporción de ítems sin interacciones o con pocas interacciones, causando que los modelos no tengan información suficiente como para recomendar adecuadamente a todos los usuarios.

Finalmente, si bien los resultados del sistema propuesto presentan mejoras teóricas en comparación con el algoritmo utilizado, para poder validar correctamente si tiene un impacto sería necesario hacer pruebas con usuarios reales.

8 BIBLIOGRAFÍA

Adomavicius, G., & Tuzhilin, A. (2005). IEEE Transactions on Knowledge an Data Engineering. En *Toward the next generation of recomender systems* (págs. 734 - 749).

Aghdam, M., Analoui, M., & Kabiri, P. (2015). A Novel Non-Negative Matrix Factorization Method for Recommender Systems. En *Applied Mathematics & Information Sciences* (págs. 2721-2732).

Bobadilla, J., Ortega, F., & Gutiérrez, A. (2013). Knowledge-Based Systems. En *Recommender systems survey* (págs. 46, 109-132).

Boutsidis, C., & Gallopoulos, E. (2007). *SVD based initialization:A head start for nonnegative matrix factorization.*

Casen. (2018). *Situación de pobreza: Síntesis de resultados.*

Chang, T., & Hsiao, W. (2013). En Proceeding of the pacis. En *LDA-based Personalized Document.*

Data and Sampling Distributions. (2022). En P. Bruce, A. Bruce, & P. Gedeck, *Practical Statistics for Data Scientists* (págs. 47-86). Oreilly.

Exploratory Data Analysis. (2020). En P. Bruce, A. Bruce, & P. Gedeck, *Practical Statistics for Data Scientists* (págs. 1-46). Oreilly.

GeCo. (2022). *Cultura GeCo.*

Geco. (3 de Diciembre de 2022). *semso.cl*. Obtenido de https://www.semso.cl/

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender Systems, An Introduction.* Cambridge University Press.

Lee, D., & Seung, H. (1999). Algorithms for Non-negative Matrix. *Advances in Neural Information Processing Systems*.

Limited Dependent Variable and Duration Models. (2002). En W. H. Greene, *Econometric Analysis* (págs. 756-845).

Lops, P., Gemmis, M., & Semeraro, G. (2011). *Recommender Systems Handbook.*

Lu, J., Wu, D., Mao, M., Wnag, W., & Zhang, G. (2015). Decision Support Systems. En *Recomendes System Application Developments: A Survey* (págs. 74, 12-32).

Mercantil. (2022). Obtenido de https://www.mercantil.com/empresa/gestion-de-comunidades-spa/las-condes/300469775/esp/

Ministerio de desarrollo Social y Familia. (2018). *desarrollosocialyfamilia.* Obtenido de www.desarrollosocialyfamilia.gob.cl: https://www.desarrollosocialyfamilia.gob.cl/noticias/magallanes-compromiso-pais-programa-para-combatir-la-pobreza-multidimensional

Models for Discrete Choice. (2002). En W. H. Greene, *Econometric Analysis* (págs. 663-755).

Pazzani, M. (2000). Artificial Intelligence Review. En *Framework for Collaborative, Content-Based and Demographic Filtering* (págs. 393-408).

Portalchile. (2022). Obtenido de https://www.portalchile.org/empresa/gestion-de-comunidades-limitada-76328282

Projectcor. (2022). Obtenido de https://projectcor.com/es/blog/cuarto-trimestre-estadisticas-de-consultoria-de-gestion/

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook.*

SemSo. (2022). Manual de implementación.

SemSo. (2022). *semso.cl*. Obtenido de https://www.semso.cl/

SII. (2022). Obtenido de https://www.sii.cl/sobre_el_sii/estadisticas_de_empresas.html

Sistema B. (s.f.). *Sistemab*. Recuperado el 5 de Octubre de 2022, de https://www.sistemab.org/

Statistical Experiments and Significance Testing. (2020). En P. Bruce, A. Bruce, & P. Gedeck, *Practical Statistics for Data Scientists* (págs. 87-141). Oreilly.

Unsupervised Learning. (2017). En J. Gareth, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)* (págs. 373-418). Springer.

Valdiviezo, P., & Hernando, A. (2016). Iberian Conference on Information Systems and Technologies. En *A Comprhensive View of Recommendation Methods base on Probabilistic Techniques* (págs. 604-609).

What Is Statistical Learning? (2017). En J. Gareth, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)* (págs. 15-29). Springer.

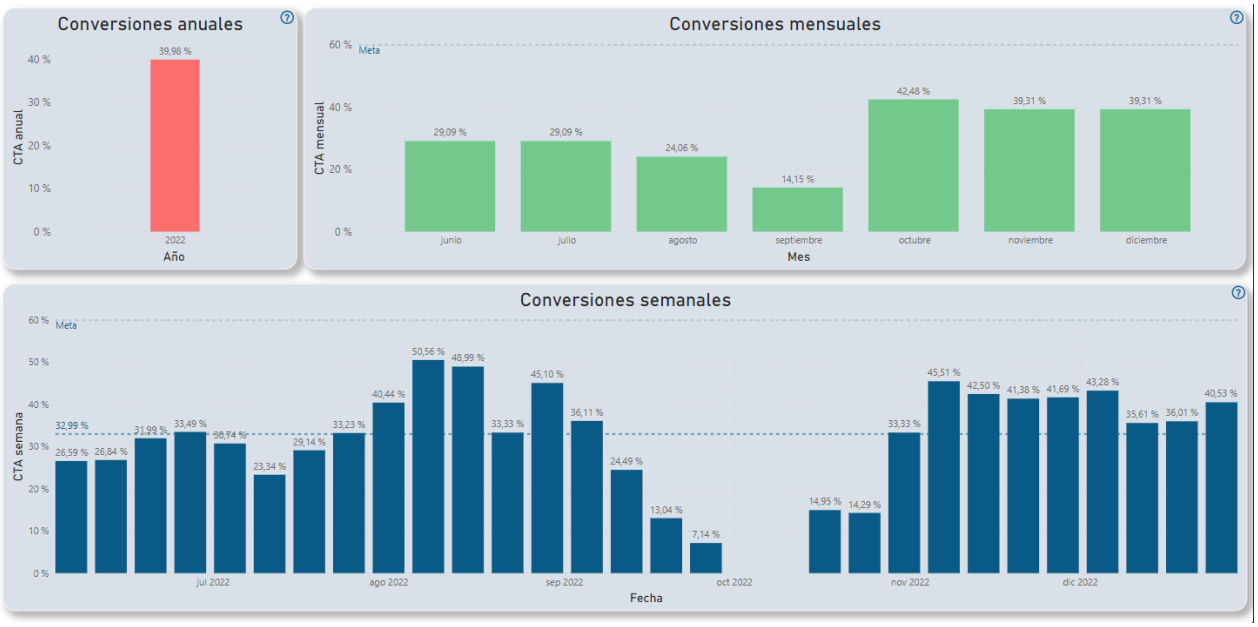
Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data.*

9 ANEXOS

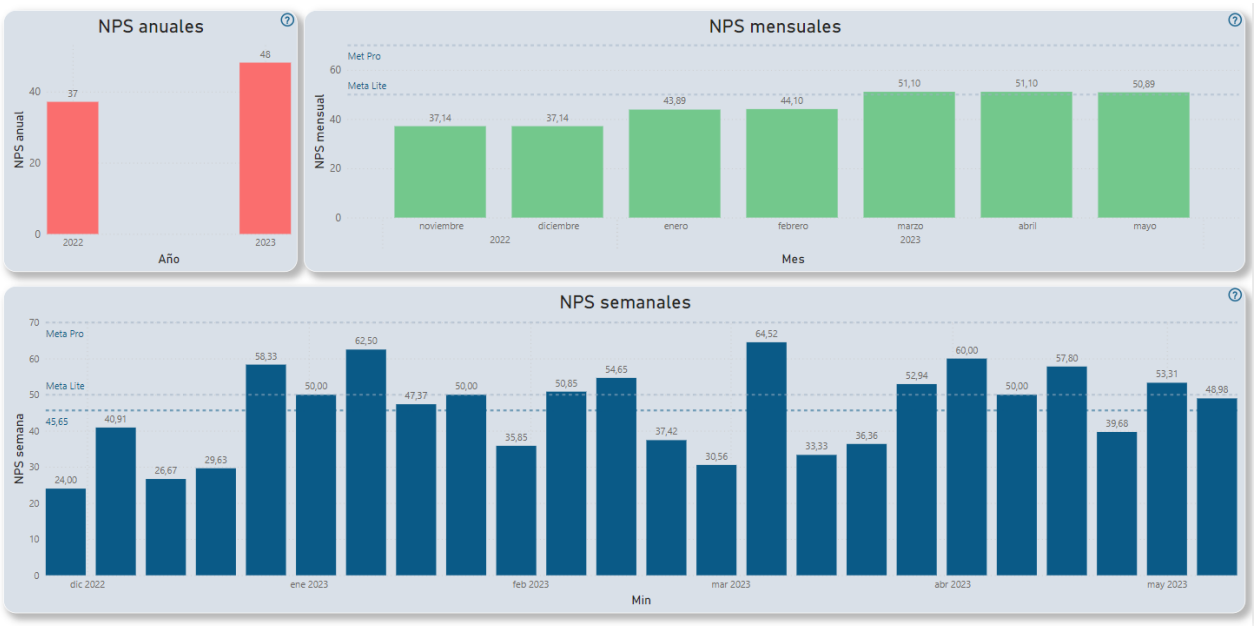
Anexo A: Encuesta de satisfacción en terreno (SemSo)

- Al responder la encuesta ¡Confiaste en que la encuesta era confidencial? 90,9% Si; 9,1% No.
- ¿Qué te pareció la duración de la encuesta? 51,5% Larga; 36%4% Normal; 12,1% Corta.
- ¿Cómo fue responder la encuesta en el celular? 84,8% Fácil; 6,1% Difícil, pero lo logré solo(a); 6,1% Tuve que pedir ayuda; 3% Indiferente.
- ¿Encuentras que te sirvió contestar SemSo? 54,5% Si; 45,5% No.
- ¿Para qué te sirvió contestarla? 44,4% Por los beneficios; 38,9% Para conocer la situación de mi familia; 16,7% Para cumplir con la obligación; 11,1% Otra.
- En conclusión ¿Te ayudó SemSo a lo que esperabas? 39,4% Si; 60,6% No.
- Luego de terminar la encuesta ¿Revisaste tus resultados? 63,6% Si; 36,4% No.
- Luego de terminar la encuesta ¿Revisaste tus beneficios? 51,5% Si; 48,5% No.
- ¿Te sirvieron los beneficios entregados? 29,4% Si; 70,6% No.
- ¿Sabes cómo volver a consultar tus beneficios? 21,2% Si; 78,8% No.
- ¿Te gustaría que SemSo te enviara beneficios adicionales de acuerdo a tus necesidades? 93,9% Si; 6,1% No.
- ¿Recomendarías SemSo a tus compañeros? 84,8% Si; 15,2% No.
- ¿Por qué? Resumen: Para saber cómo esta uno y su familia y para conocer beneficios que pueden servir y que no sabes de su existencia.

Anexo B: Gráficos de conversión de usuarios

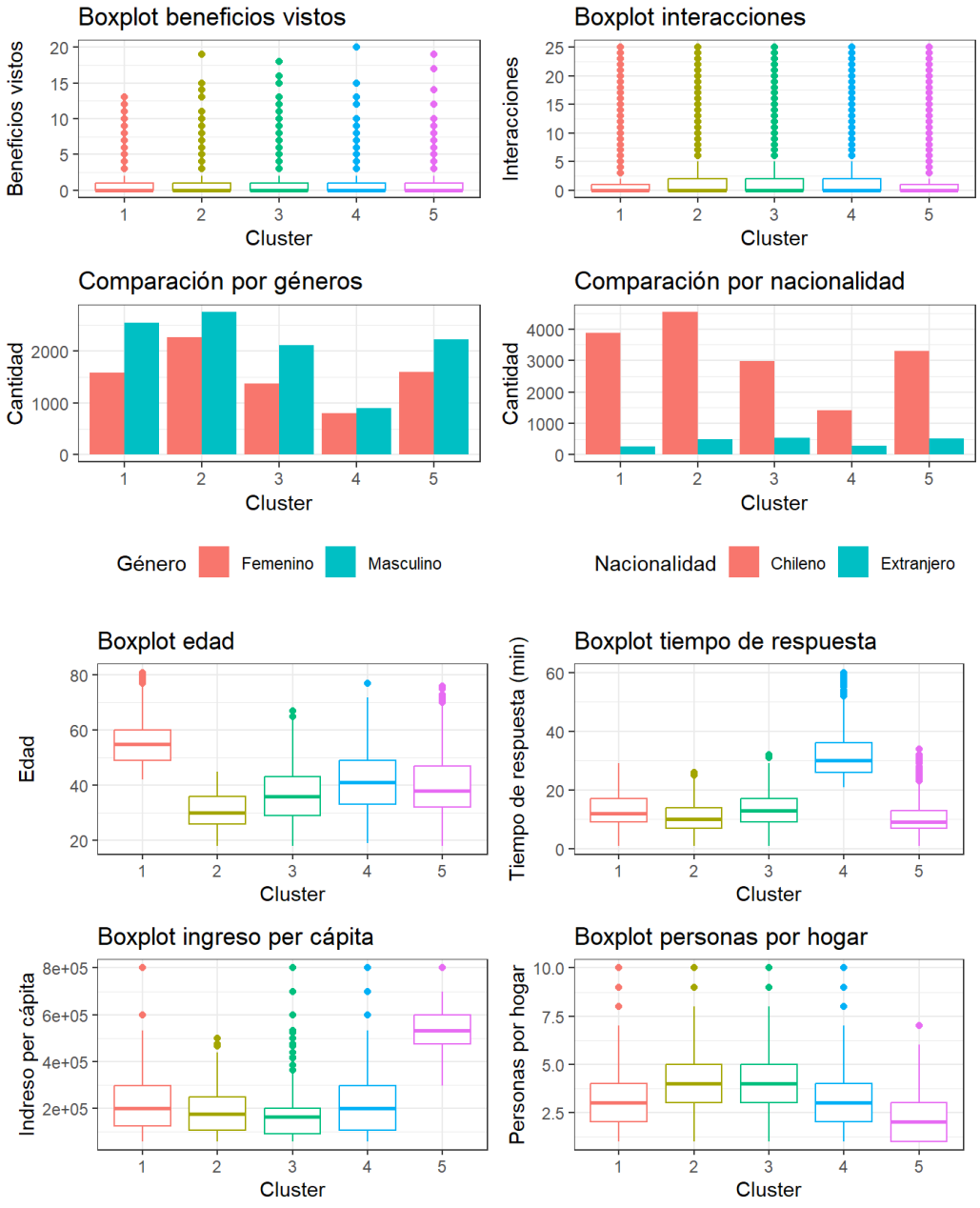


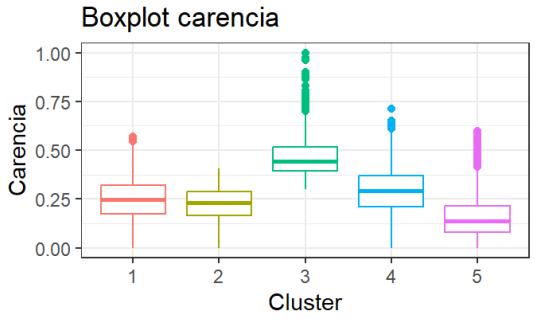
Anexo C: Gráficos de NPS

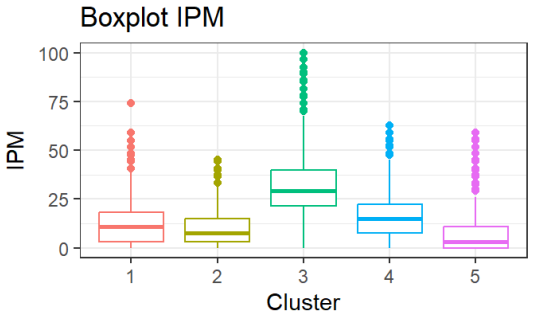


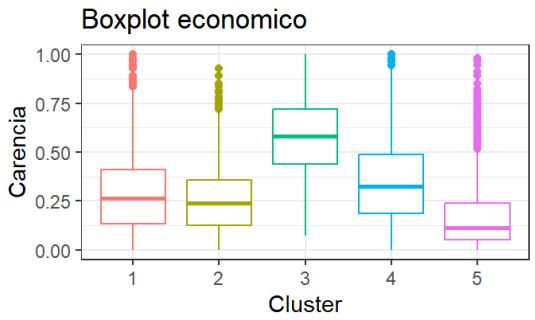
Anexo D: Modelo SemSo

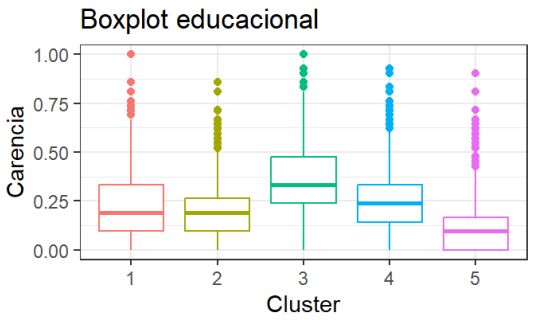
Anexo E: Caracterización clúster personas

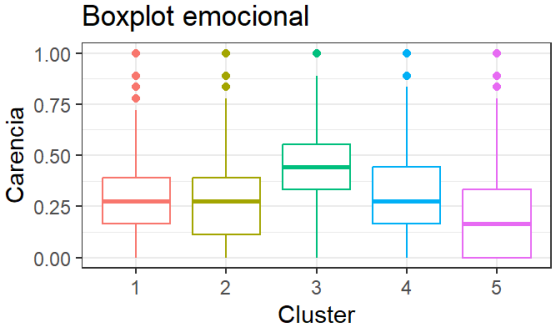


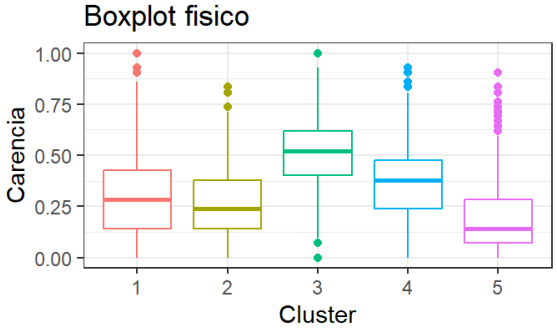

Boxplot carencia
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster

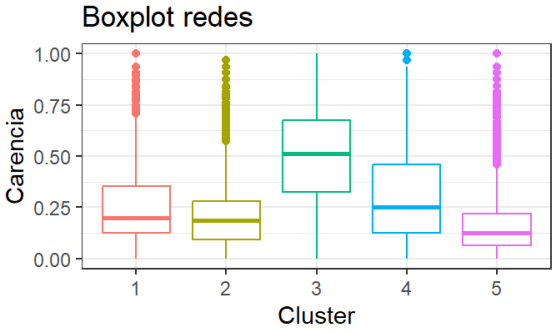

Boxplot IPM
IPM
100
75
50
25
0
1
2
3
4
5
Cluster

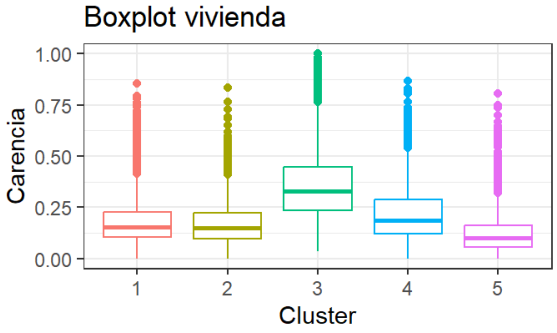

Boxplot economico
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster


Boxplot educacional
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster

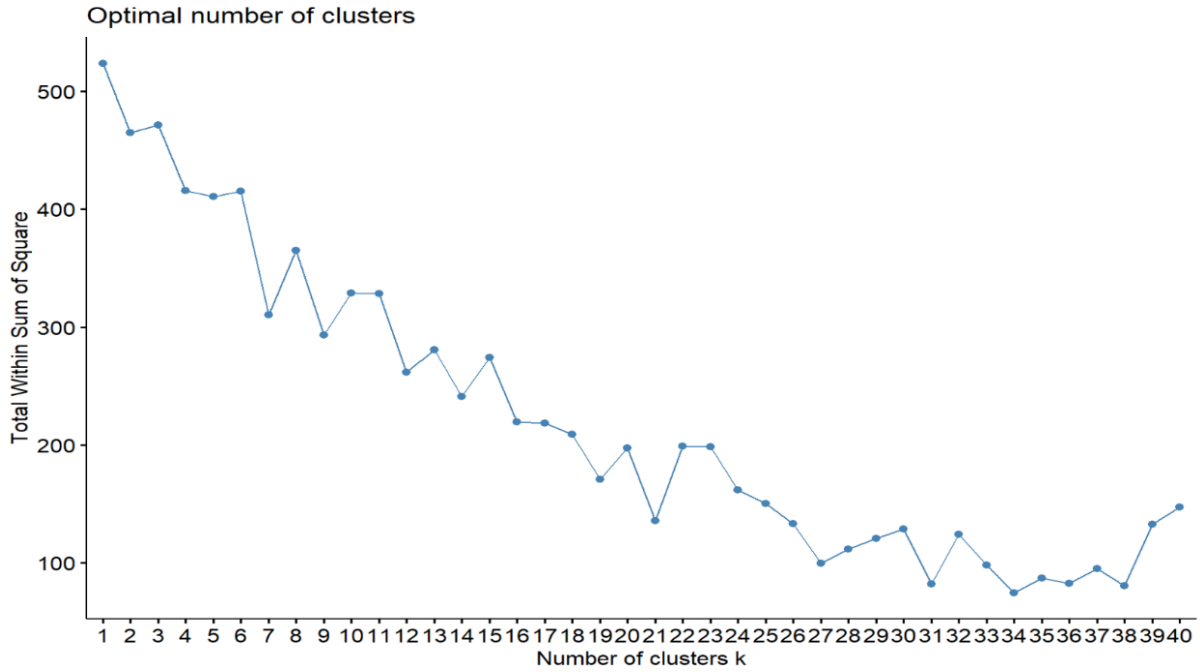

Boxplot emocional
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster


Boxplot fisico
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster

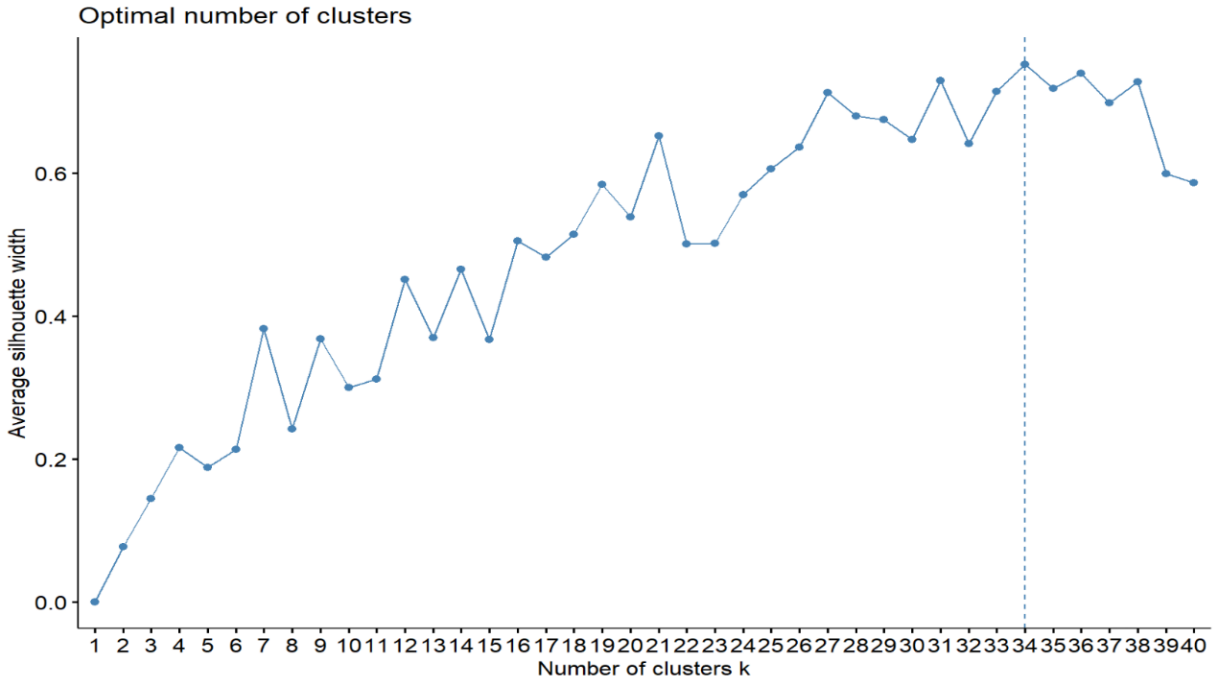

Boxplot redes
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster


Boxplot vivienda
Carencia
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Cluster

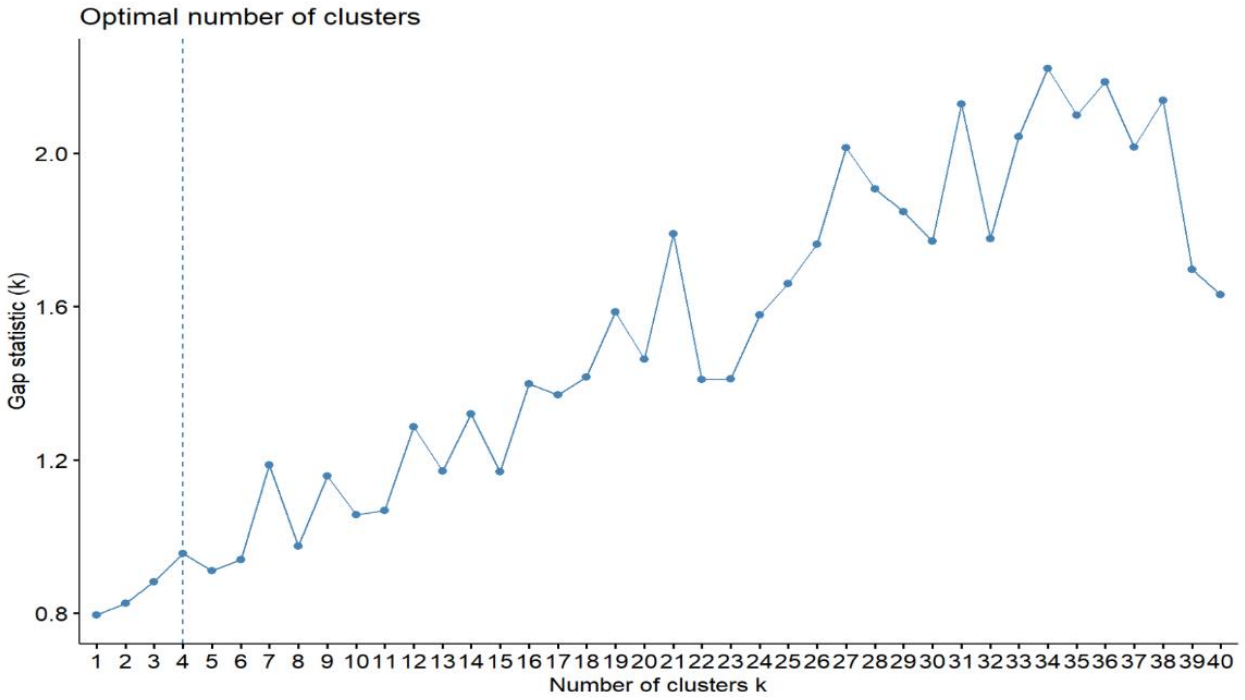
Anexo F: WSS para K-Means por tema.



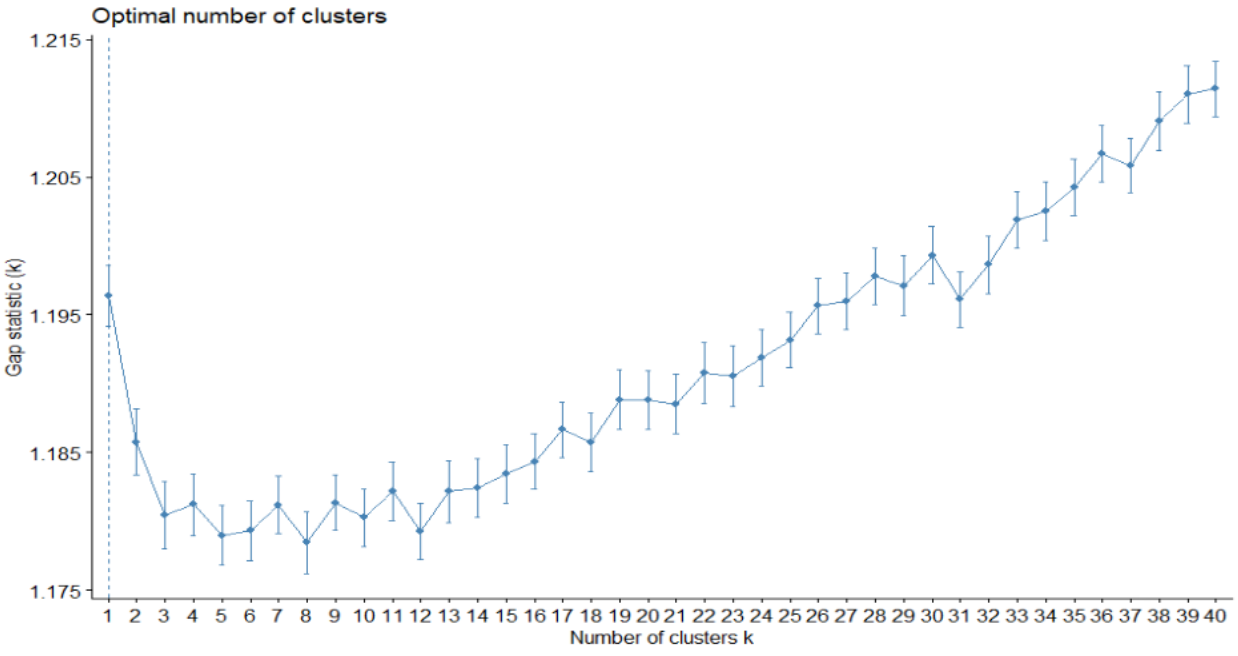
Anexo G: Coeficiente Silhouette para K-Means por temas.



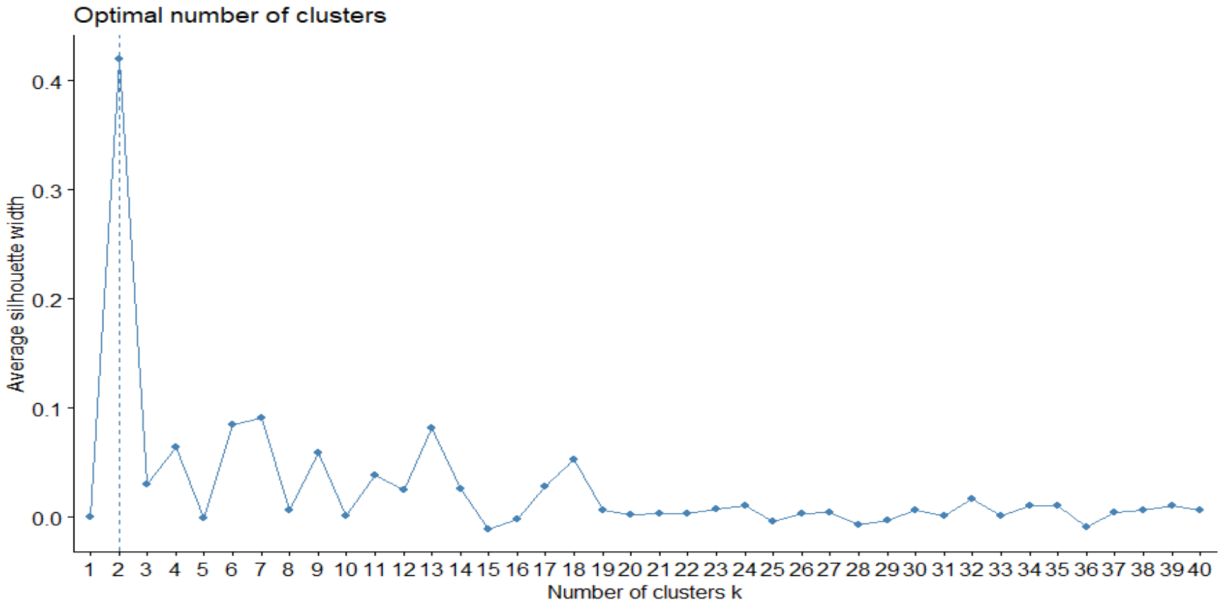
Anexo H: Diferencia estadística para K-Means por temas.



Anexo I: WSS para K-Means por TF-IDF.



Anexo J: Coeficiente Silhouette para K-Means por TF-IDF.



Anexo K: Diferencia estadística para K-Means por TF-IDF.