Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC3633 — Sistemas Recomendadores — 2024-02

Informe intermedio

Nombre:	Fernanda	Mathias	Felipe	Beatriz
Apellidos:	Pérez Hargreaves	Madsen Sánchez	Olivares Labarca	Errázuriz Camus
RUT:	20.428.286-2	20.415.980-7	20.075.251-1	20.165.959-0
Número de alumno:	19640315	19623925	19642644	19638906

Solución Propuesta

Descripción del Problema y Objetivos del Proyecto

Los sistemas de recomendación actuales suelen priorizar el comportamiento pasado de los usuarios, generando recomendaciones personalizadas pero con poca diversidad, lo que limita el descubrimiento de nuevos contenidos y puede resultar en una experiencia repetitiva. Este proyecto desarrolla la métrica **User Diversity**, que evalúa la diversidad en las recomendaciones en función de los intereses específicos del usuario. A diferencia de las métricas tradicionales, **User Diversity** busca promover la exploración de nuevos géneros y categorías, mejorando la experiencia del usuario al equilibrar personalización y diversidad.

El objetivo principal es desarrollar y evaluar esta métrica, formalizándola en términos de variedad de géneros y categorías, e implementarla en datasets como Last.fm y MovieLens. Se comparará su desempeño con otras métricas, evaluando su impacto en la satisfacción y el descubrimiento de contenido, con el propósito de ofrecer una herramienta que optimice la experiencia de recomendación y fomente la exploración de nuevos contenidos.

Solución propuesta Actual

Hasta el momento, el proyecto ha avanzado en el diseño e implementación de la métrica de diversidad propuesta, denominada **User Diversity**. Esta métrica se basa en el concepto de Diversity Coverage, adaptándolo para capturar con mayor precisión la dispersión de las recomendaciones en las categorías más relevantes para el usuario.

Para implementar esta métrica, se definió un parámetro k que limita el número de categorías a considerar en el cálculo de diversidad. Se estableció un valor de $k = 5$, ya que, con valores superiores, las categorías de recomendación no variaban significativamente, mientras que, con valores menores, el número de categorías consideradas resultaba insuficiente para una evaluación completa. Este top-5 representa las categorías de mayor preferencia del usuario, capturando los géneros de música que consumen con mayor frecuencia y que mejor reflejan sus intereses.

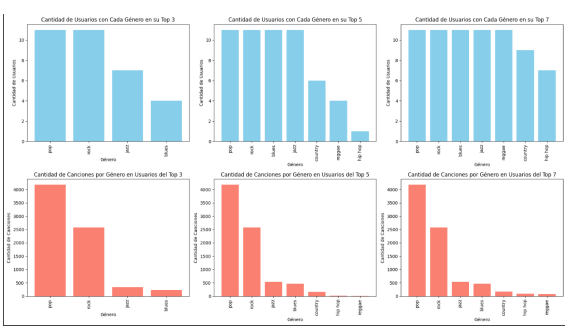

Figura 1: Cantidad de usuarios y canciones por género con distintos K

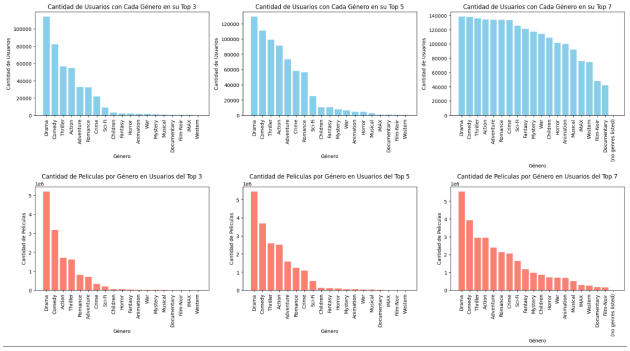

Figura 2: Cantidad de usuarios y películas por género con distintos K

Incrementar el valor de k permite incluir géneros con características auditivas diversas en las recomendaciones. Esto es esencial para una métrica de diversidad, ya que no solo se mide la frecuencia de los géneros recomendados, sino también la variabilidad en sus características de audio, lo cual enriquece la experiencia del usuario.

Una métrica de diversidad puede, por ejemplo, analizar cómo valores como *danceability*, *acousticness* o *instrumentalness* varían entre los géneros recomendados, incentivando la exploración de perfiles auditivos únicos y personalizados para cada usuario.

Observamos en la Figura 3 (incluida en anexo 1) que las estructuras de preferencias de la mayoría de los usuarios tienden a concentrarse en los géneros principales, como *pop*, *rock* y *blues*. Sin embargo, también existen usuarios que valoran propiedades auditivas específicas, como la acústica, que solo se encuentran en géneros recomendados con K igual a 5 o 7, lo que permite un enfoque más flexible y personalizado en la recomendación.

Definimos la formula de User Diversity.

$$UD = 1 - \frac{\left|\sum_{j=1}^{k} \left(\frac{R_j}{R}\right) \cdot \log\left(\frac{R_j}{R}\right)\right|}{\log(k)}$$

Donde:

- k: Número total de categorías para un usuario.
- R: Número total de recomendaciones realizadas.
- R_j: Cantidad de recomendaciones en la categoría j.
- $\frac{R_j}{R}$: Proporción de recomendaciones en la categoría j.
- $\log\left(\frac{R_j}{R}\right)$: Penalización para concentraciones altas.
- $\log(k)$: Normalización para garantizar que $UD \in [0, 1]$.

Se ocupo el valor absoluto en la formula, debido a que $\left(\frac{R_j}{R}\right) \cdot \log\left(\frac{R_j}{R}\right)$ puede generar valores negativos, ya que $\log(x) < 0$ cuando $0 < x < 1$. Esto podría llevar a una suma negativa, afectando la normalización y dando como resultado valores fuera del rango esperado para la métrica. El uso del valor absoluto $|\cdot|$ asegura que el numerador sea no negativo, permitiendo que el cálculo de la métrica esté correctamente escalado entre 0 y 1.

La métrica de User Diversity mide la diversidad de recomendaciones considerando cómo se distribuyen en las principales categorías de interés del usuario. A diferencia de métricas generales, esta se centra en evaluar si las recomendaciones abarcan las categorías más relevantes para el usuario, evitando sesgos hacia una sola categoría o un conjunto limitado de preferencias.

Al enfocarse en los géneros más representativos para cada usuario, User Diversity permite una visión más precisa de la diversidad desde su perspectiva, promoviendo un balance entre personalización y exploración. Esto ayuda a los desarrolladores a mejorar la satisfacción y el descubrimiento de contenido en plataformas de recomendación.

Experimentación realizada y evaluación intermedia

Se llevó a cabo una experimentación utilizando los datasets de Last.fm y MovieLens, que contienen información sobre las preferencias de usuarios en música y películas. La métrica **User Diversity** fue aplicada en cuatro tipos de sistemas de recomendación: **Most Popular** (basado en ítems populares sin personalización), **Random** (selección aleatoria de ítems), **Método colaborativo** (basado en similitudes de usuarios o ítems) y **Método híbrido** (combinación de popularidad y similitud de géneros ajustable mediante el parámetro alpha).

Para comparar el rendimiento de estos métodos, se utilizaron diversas métricas. La **MAP (Mean Average Precision)** evalúa la precisión en función de las preferencias previas del usuario, mientras que **NDCG@N** mide la calidad en el top-N ponderando las posiciones de recomendaciones relevantes. **Precision@N** y **Recall@N** miden la relevancia y cobertura de los ítems recomendados, respectivamente.

Además, se emplearon métricas de diversidad como **Long Tail** (inclusión de ítems menos populares), **Shannon Entropy** (dispersión en diferentes categorías), **Intra List Diversity** (variedad dentro de la lista de recomendaciones), **Diversity Coverage** (proporción de categorías cubiertas) y **Inverse Propensity Score** (favorece recomendaciones de bajo perfil). Estas métricas permitieron una evaluación integral del desempeño de cada método en cuanto a precisión, personalización y diversidad.

Análisis de datos

En este análisis, el gráfico de radar (figura 4 encontrada en anexo 2) se utiliza para comparar el rendimiento de los diferentes métodos de recomendación mencionados a través de un conjunto de métricas clave. Este gráfico de radar permite observar visualmente cómo cada método se desempeña en términos de cada métrica, facilitando una comparación rápida de sus fortalezas y debilidades.

El código de análisis calcula el promedio y la desviación estándar de estas métricas para cada método, permitiendo cuantificar las diferencias en el rendimiento. Además, se utiliza un análisis de varianza (ANOVA) para determinar la significancia estadística de estas diferencias, comparando los resultados obtenidos para cada métrica entre los diferentes métodos de recomendación. El ANOVA calcula un valor F y un valor p para cada métrica, lo cual ayuda a identificar si las diferencias observadas son significativas o si podrían deberse al azar.

En términos de los resultados obtenidos, se observa que el método híbrido destaca en la métrica de User Diversity, lo que implica que es más eficaz para proporcionar recomendaciones que abarquen géneros de interés específico del usuario en comparación con los otros métodos. Esto sugiere que el enfoque híbrido logra un mejor balance entre precisión y diversidad en los gustos del usuario, personalizando la experiencia de recomendación de manera más efectiva.

El figura 5 (encontrada en anexo 3) muestra la variación de diferentes métricas de recomendación para cuatro métodos distintos (most popular, collaborative, hybrid, random) en función del tamaño de las recomendaciones (Top N). La métrica de User Diversity, que captura la diversidad de los ítems recomendados a los usuarios, es de particular interés en el contexto de nuestro análisis.

La métrica User Diversity revela cómo los distintos algoritmos de recomendación logran diversificar las recomendaciones según el valor de Top N. En el caso del método collaborative, observamos que muestra una tendencia a mantener una diversidad de usuario relativamente alta en comparación con los otros métodos, especialmente cuando el número de recomendaciones es pequeño (por ejemplo, Top 5 y Top 10). Esto indica que el enfoque colaborativo tiende a sugerir una variedad de ítems más amplia, probablemente debido a que este tipo de método explora más los intereses de cada usuario en particular.

Por otro lado, el método most popular mantiene una User Diversity baja en los valores de Top N más altos. Este resultado era esperado, ya que este enfoque se basa en recomendar los ítems más populares, lo cual reduce la diversidad en las recomendaciones.

Finalmente, el método hybrid presenta un comportamiento intermedio en términos de diversidad de usuario. Es el que alcanza los valores más altos, lo cual sugiere que la combinación de los enfoques collaborative y most popular logra una diversidad adecuada sin caer en recomendaciones demasiado genéricas ni excesivamente diversificadas.

La métrica *User Diversity* es esencial para evaluar la capacidad de los sistemas de recomendación de ofrecer opciones variadas y personalizadas. Una alta *User Diversity* mejora la experiencia del usuario al proporcionarle contenidos diversos, lo que contribuye a su satisfacción y fomenta la lealtad, evitando la monotonía y el "efecto de burbuja". En contraste, un sistema centrado solo en ítems populares, con baja diversidad, puede resultar repetitivo y poco personalizado, afectando el interés de los usuarios, especialmente aquellos con intereses menos comunes. En conclusión, *User Diversity* es crucial para lograr un balance entre personalización y novedad, promoviendo una experiencia rica y satisfactoria, especialmente en aplicaciones enfocadas en el descubrimiento.

Problemas Identificados Durante el Proceso

Durante el desarrollo del proyecto, se presentaron diversos desafíos técnicos que impactaron directamente la implementación de los modelos y el análisis de los datos.
En primer lugar, se planificó el uso de la librería PyRecLab para la implementación de los modelos propuestos. Sin embargo, los datos de música proporcionados por Last.fm no contenían información sobre rankings, lo cual dificultó el uso directo de dicha librería. Como consecuencia, fue necesario desarrollar funciones personalizadas en un entorno *Jupyter Notebook* para implementar tanto los modelos como las métricas de evaluación. Este proceso implicó la creación de métodos desde cero para las siguientes estrategias de recomendación: Recomendaciones aleatorias (*Random Recommendations*), Recomendaciones basadas en popularidad (*Most Popular*), Filtrado colaborativo mediante factorización matricial (*Collaborative Filtering - Matrix Factorization*) y Sistemas híbridos con algoritmos avanzados (*Hybrid Systems*).

En segundo lugar, al trabajar con el conjunto de datos de películas, surgieron limitaciones en la capacidad de memoria RAM disponible en Google Colab, lo que impidió la ejecución de los métodos previamente definidos para los datos de música. Esta restricción obligó a buscar alternativas y a simplificar algunos de los procedimientos y tener que ajustar el plan propuesto por el tiempo de busquedad de otra alternativa.

Revisión del Plan Propuesto y Justificación de Ajustes

Durante la ejecución del proyecto, se realizaron varios ajustes al plan inicialmente propuesto, tanto en las tareas como en los enfoques metodológicos. Estos cambios fueron necesarios para abordar limitaciones prácticas y refinar el análisis en función de los objetivos del proyecto.

En la etapa inicial, en lugar de realizar un *survey* general sobre métricas de diversidad, se llevó a cabo una investigación exhaustiva de estudios previos donde se desarrollaron métricas de diversidad. A raíz del trabajo de Kunaver y Požrl (2017), se decidió no realizar un *survey* completo, sino explorar métricas específicas que se han creado en investigaciones previas para la medición de diversidad. Este enfoque permitió identificar las características clave que buscábamos en una métrica de diversidad y establecer cómo nuestra métrica, *User Diversity*, podría diferenciarse de las ya existentes. Esta modificación fue clave para garantizar que la métrica propuesta abordara las limitaciones específicas de las métricas actuales, como la falta de personalización y el enfoque limitado en la exploración de contenido.

Sin embargo, algunos desafíos afectaron la implementación según el cronograma previsto. Aunque para los datos de música se lograron implementar los modelos base (*Random Recommendations*, *Most Popular*, *Collaborative Filtering*) y la métrica de diversidad, así como un modelo avanzado híbrido, estos avances estuvieron restringidos exclusivamente a los datos de música. Para los datos de películas, no fue posible extraer la información requerida ni realizar las pruebas necesarias debido a limitaciones técnicas, como la memoria RAM limitada de Google Colab, que en varias ocasiones no permitió ejecutar los análisis necesarios.

Bibliografía

Kunaver, M., & Požrl, T. (2017). *Diversity in recommender systems – A survey*. Knowledge-Based Systems, 123, 154-162. `https://doi.org/10.1016/j.knosys.2017.02.009`

Pérez Hargreaves, F., Madsen Sánchez, M., Olivares Labarca, F., & Errázuriz Camus, B. (2024). *Evaluación de métrica de diversidad en música*. Google Colab. Recuperado de `https://colab.research.google.com/github/ferperezh/IIC3633_ProyectoMetricaDiversidad/blob/main/musica/musica_metrica_diversidad.ipynb`

Adomavicius, G., & Kwon, Y. (2012). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*.

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and diversity in recommender systems. *Recommender Systems Handbook*.

Kaminskas, M., & Bridge, D. (2017). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.

Ziegler, C. N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. *Proceedings of the 14th International Conference on World Wide Web*.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. *IEEE International Conference on Data Mining*.

Johnson, J., & Ranganathan, K. (2019). DeepFM: A factorization-machine based neural network for CTR prediction. *Proceedings of the 26th International Conference on Neural Information Processing.*

Anexo

Anexo 1:

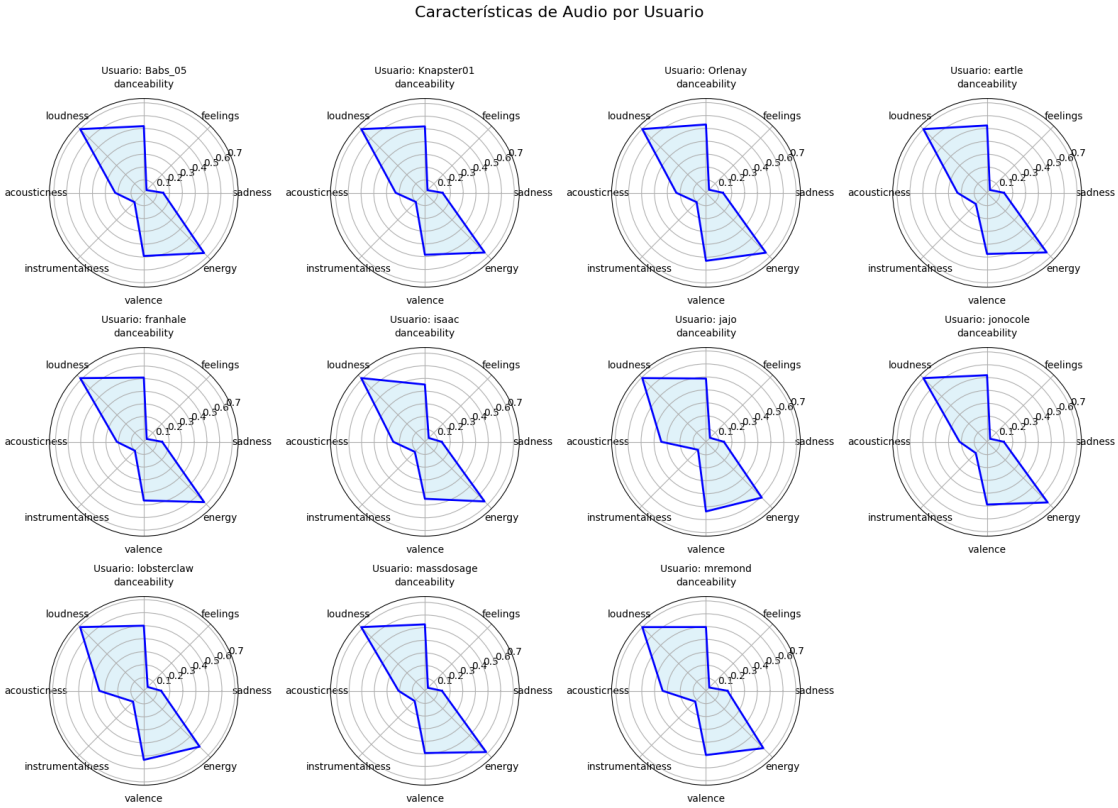

Figura 3: Variación en las características de Audio por Usuario

Anexo 2:

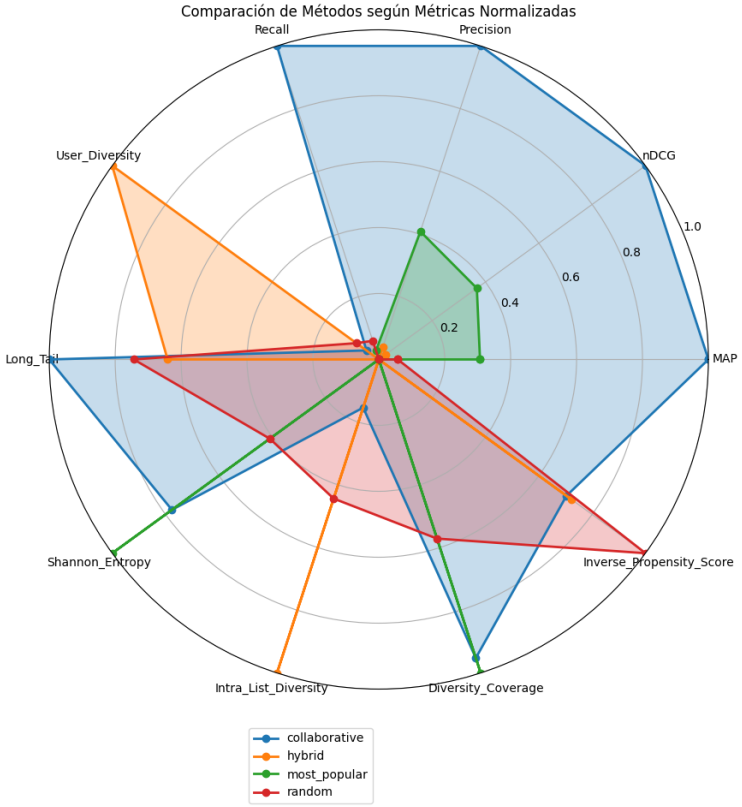

Figura 4: Comparación de métodos de recomendación según métricas normalizadas.

Anexo 3:

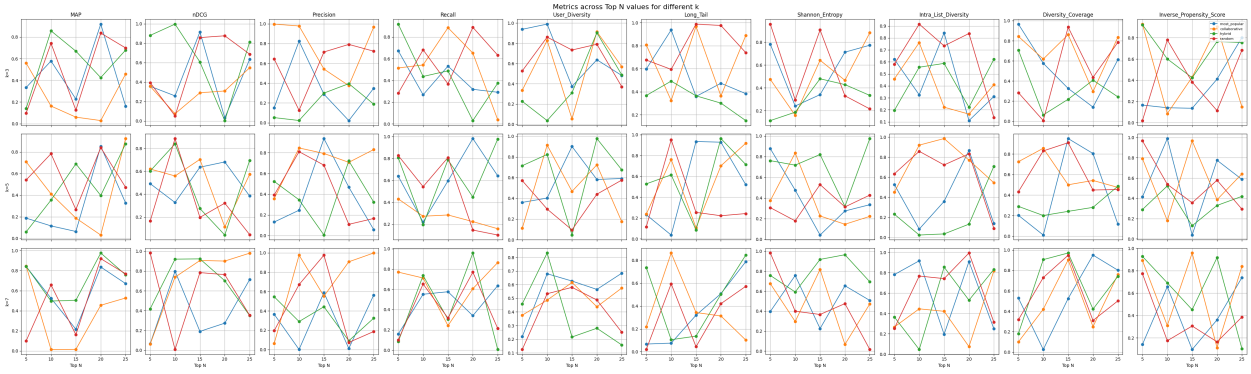

Figura 5: Comparación de Métricas de Rendimiento de Modelos de Recomendación para Diferentes Valores de Top N y distintos K.

Anexo 4:

Métrica	Most Popular	Random	Collaborative	Método Híbrido
MAP	0.3103	0.0387	0.7007	0.1374
nDCG	0.2345	0.0215	0.5413	0.0686
Precision	0.2364	0.0273	0.5000	0.0727
Recall	0.0101	0.0021	0.2476	0.0033
User Diversity	0.4263	0.4825	0.4517	0.9717
Long Tail	0.0000	0.1455	0.3545	0.2273
Shannon Entropy	0.6576	0.2810	0.5618	0.2295
Intra-List Diversity	0.4028	1.0121	0.5225	1.1818
Diversity Coverage	0.4286	0.2078	0.4156	0.1558
Inverse Propensity Score	361.0272	850.9579	793.0523	804.3782

Cuadro 1: Comparación de métricas de rendimiento entre métodos de recomendación

Anexo 5:

Métrica	Características Principales	Comparación con Üser Diversity"
User Diversity (nuestra)	Diversidad entre categorías relevantes para el usuario, promueve el equilibrio entre categorías más populares y menos populares para una experiencia personalizada.	Se enfoca en las categorías de interés del usuario individual, permitiendo una diversidad más personalizada.
Entropía de Shannon	Busca maximizar la diversidad en todos los ítems o categorías, asume una distribución uniforme.	A diferencia de User Diversity, no se centra en las categorías individuales del usuario sino en una diversidad general.
Long Tail	Fomenta la recomendación de ítems de baja popularidad para aumentar la exposición de nichos.	Similar en el enfoque de promover diversidad, pero User Diversity se enfoca en un equilibrio más ajustado a los intereses del usuario.
Intra-List Diversity (ILD)	Mide la variedad dentro de una lista de recomendaciones, penalizando similitudes entre ítems.	Puede complementarse con User Diversity, ya que ambas promueven variedad, aunque ILD no necesariamente toma en cuenta las categorías preferidas del usuario.
Diversity Coverage	Evalúa la diversidad entre categorías recomendadas sin considerar similitudes específicas de ítems.	Es más amplio y no considera tanto las preferencias del usuario individual como lo hace User Diversity.
Inverse Propensity Score (IPS)	Similar a Long Tail, asigna más peso a ítems menos populares, balanceando popularidad.	Compatible con User Diversity, pero IPS no evalúa el equilibrio entre categorías específicas del usuario.

Cuadro 2: Comparación de la métrica User Diversity con otras métricas de diversidad

Anexo 6:

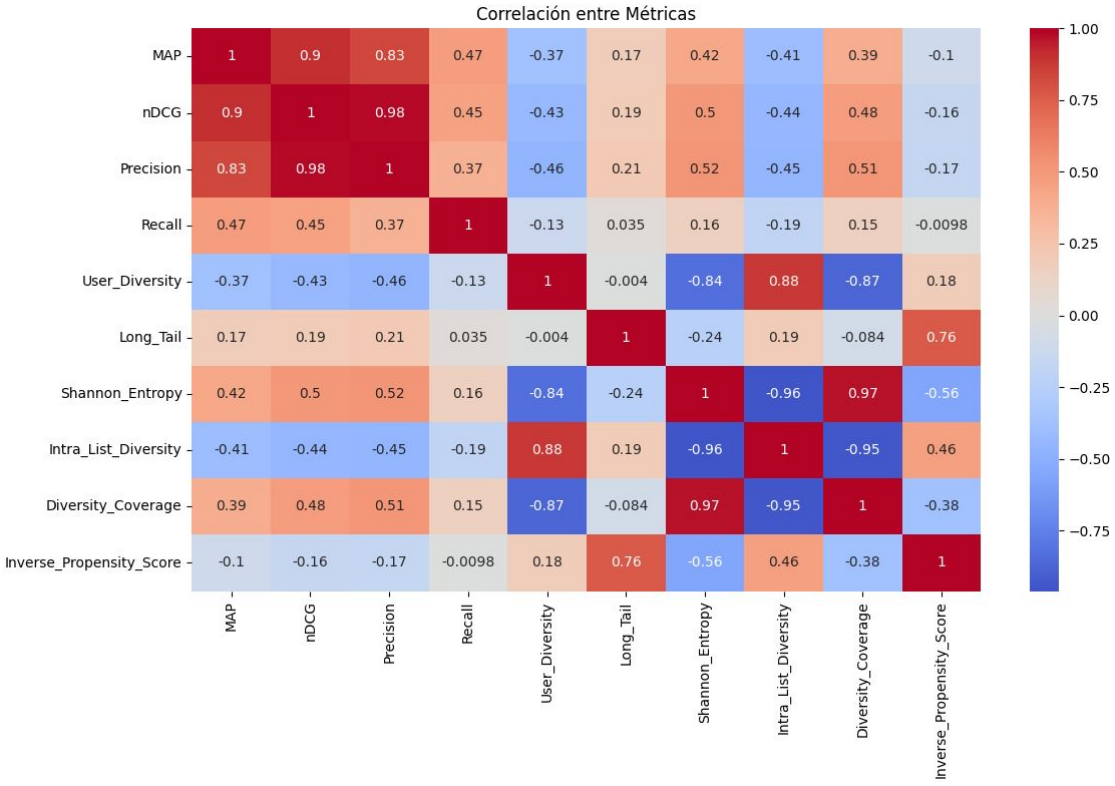

Figura 6: Correlación entre métricas para topn $= 10$ y $K = 5$

Anexo 7:

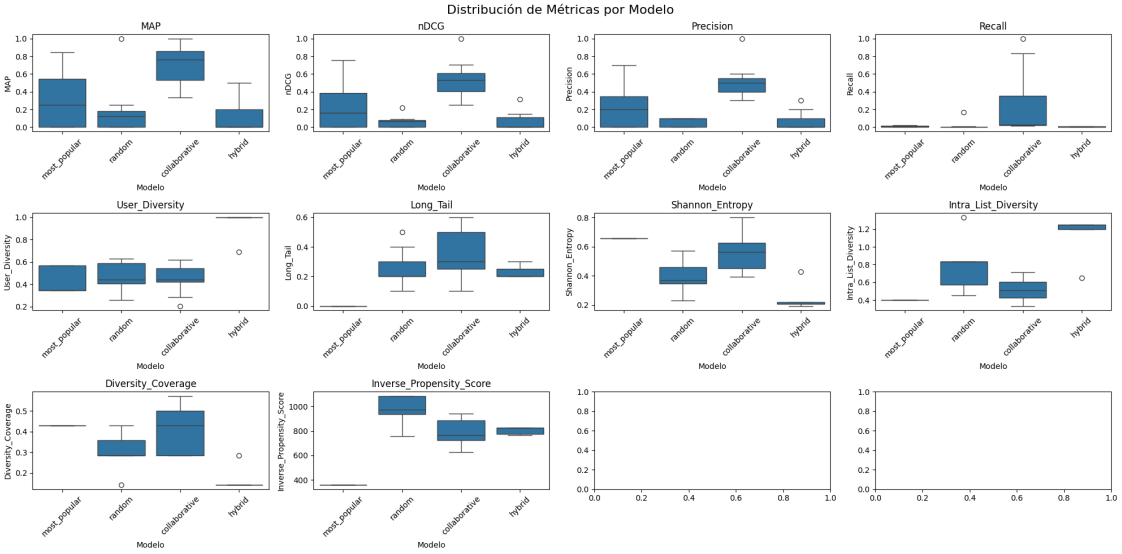

Figura 7: Distribución de métricas de rendimiento para cada modelo de recomendación: Most Popular, Random, Collaborative, y Hybrid para topn $= 10$ y $K = 5$