

# Sistema de Recomendaciones

## Proyecto Final

Alexis Rodrigo Arce Delgadillo  
Ingeniería en Informática  
Facultad Politécnica / Universidad Nacional de  
Asunción  
San Lorenzo, Paraguay  
alexisroarce2000@fpuna.edu.py

Bethania Ángel Chávez García  
Ingeniería en Informática  
Facultad Politécnica / Universidad Nacional de  
Asunción  
San Lorenzo, Paraguay  
chavezbethania@fpuna.edu.py

### ABSTRACT

Este proyecto se centra en la creación de un sistema de recomendación aplicable a varios conjuntos de datos, utilizando un enfoque de lógica difusa. El sistema utiliza un conjunto de datos que contiene reseñas de productos en formato Gzip, emplea información de críticas y calificaciones de ítems, integrando módulos para calcular puntajes de calificación, puntajes de similitud e implementar un sistema de inferencia difusa mediante un conjunto de reglas de difusas y un defusificador para representar la recomendación final del producto. El módulo lector de conjuntos de datos es el encargado de gestionar los conjuntos de datos introducidos por el usuario en el formato mencionado anteriormente. El módulo rating score promedia las calificaciones de cada ítem, mientras que el módulo de similarity score calcula la similaridad de contenidos entre reseñas y resúmenes obtenidos del conjunto de datos utilizando el método de similitud del coseno mediante el empleo de librerías externas. El módulo sistema de inferencia difusa incorpora reglas basadas en las similarity score y rating score, categorizando los productos en niveles de recomendación como “Highly Recommended”, “Likely to recommend”, “Recommended” y “Not recommend”. El módulo defusificador, a través del método del centroide ponderado, combina los similarity score y rating score para proporcionar una recomendación concreta. Los módulos del proyecto, en conjunto, tienen el objetivo de mejorar las recomendaciones personalizadas en diversos conjuntos de datos, considerando tanto las preferencias del usuario como las similitudes entre productos. Este enfoque busca optimizar la utilidad y precisión de recomendaciones en diferentes contextos de conjuntos de datos.

### 1. INTRODUCCIÓN

La investigación de este proyecto pone énfasis en el desarrollo de un sistema de recomendación empleable a varios conjuntos de datos, utilizando un enfoque en lógica difusa. Este enfoque tiene como objetivo abordar los desafíos asociados con la mejora de las recomendaciones personalizadas en entornos con gran cantidad y diversidad de información.

La creciente disponibilidad de conjuntos de datos que contienen reseñas de productos en formato Gz ha generado la necesidad de desarrollar sistemas de recomendación más avanzados y adaptables. Este proyecto surge de la necesidad de perfeccionar la utilidad y precisión de las recomendaciones, considerando tanto las preferencias del usuario como las similitudes entre productos.

La motivación detrás de este proyecto radica en la creciente importancia de proporcionar experiencias de usuario altamente personalizadas y pertinentes. Los sistemas de recomendación desempeñan un papel fundamental al facilitar la toma de decisiones de los usuarios y mejorar su satisfacción. La aplicación de técnicas de lógica difusa aborda la naturaleza subjetiva y ambigua de las preferencias del usuario, mejorando la capacidad del sistema para comprender y adaptarse a sus necesidades.

Los resultados obtenidos en el proyecto, abordaron la precisión de las recomendaciones como la capacidad del sistema para adaptarse a los diversos conjuntos de datos que fueron cargados, así también mediante el uso de los módulos se obtuvo un conjunto de recomendaciones para diversos ítems, facilitando la toma de decisiones por medio de las reglas difusas.

### 2. PRELIMINARES

En el núcleo de nuestro sistema de recomendación reside un componente fundamental: la lógica difusa. Este

concepto es esencial para capturar la complejidad de las preferencias del usuario y la similitud entre elementos de manera más realista. En lugar de simplemente clasificar las recomendaciones como verdaderas o falsas, la lógica difusa nos permite trabajar en un espacio continuo de veracidad. La lógica difusa nos brinda la capacidad de modelar la ambigüedad inherente a las preferencias del usuario. Nuestro sistema no se limita a recomendaciones rígidas; en cambio, se ajusta y adapta a la diversidad de preferencias, ofreciendo sugerencias que se alinean con la complejidad del comportamiento humano.

En secciones posteriores, explicaremos cómo estos principios de lógica difusa influyen directamente en las decisiones dentro de nuestro sistema de recomendación, desde la evaluación de similitudes hasta el proceso de inferencia y defusificación.

El sistema de recomendación propuesto utiliza una métrica llamada rating score que en esencia captura la percepción colectiva de los usuarios hacia un producto específico y calcula tomando la media de todas las calificaciones proporcionadas por la comunidad de usuarios. Expresada por la siguiente fórmula:

$$RS_{promedio} = \frac{1}{n} \sum_{i=1}^n R_i \quad (1)$$

donde  $n$  representa el número total de calificaciones para el producto y  $R_i$  denota la calificación individual dada por el usuario  $i$ . Otro parámetro utilizado en el sistema es el Similarity Score [1] (Puntaje de Similitud) determinando la relación entre diferentes elementos del conjunto de datos. Este puntaje, calculado mediante similitud del coseno, proporciona información sobre la proximidad conceptual entre elementos, como reseñas de productos y resúmenes de revisión de usuarios. Se expresa con la ecuación:

$$CosenoSimilitud(P_i, P_j) = \frac{P_i \cdot P_j}{||P_i|| \times ||P_j||} \quad (2)$$

Donde  $P_i$  y  $P_j$  son vectores de TF-IDF (Frecuencia de Término-Frecuencia Inversa de Documento) que representan las características de los elementos  $P_i$  y  $P_j$ , y  $||P_i||$  y  $||P_j||$  son las normas euclidianas de los vectores. En resumen, esta fórmula mide la similitud entre dos vectores, normalizando el producto punto por las normas euclidianas de los vectores involucrados. Cuanto más cercano sea el resultado a 1, mayor será la similitud entre los vectores. Ecuación de la norma de un vector:

$$||P|| = \sqrt{p_1^2 + p_2^2 + \dots + p_N^2} \quad (3)$$

esto normaliza el producto punto  $||P_i|| \times ||P_j||$

dividiéndolo por la magnitud de ambos vectores. Esto garantiza que la similitud no esté dominada por la longitud de los vectores y se centre en la dirección (o ángulo) entre ellos.

La función de sistema de inferencia difusa son reglas difusas que podrían reflejar relaciones entre el "Rating Score promedio" y el "Similarity Score". Tabla 1.

El defusificador es un componente clave en nuestro sistema de recomendación basado en lógica difusa. Su función principal es convertir los resultados difusos ponderados, generados por el sistema de inferencia, en decisiones concretas y acciones específicas. En nuestro enfoque, utilizamos el método del centroide ponderado como métrica de decisión.[2] Este método calcula un punto central, o centroide, de los resultados ponderados, teniendo en cuenta tanto la similitud como el rating score de los productos. Este centroide representa la recomendación final. Utiliza la siguiente fórmula:

$$CentroidePonderado_i = \frac{\sum_{i=1}^n SimilarityScore_i \cdot GradoDePertenencia_i}{\sum_{i=1}^n GradoDePertenencia_i} \quad (4)$$

$$CentroidePonderado_j = \frac{\sum_{j=1}^n RatingScore_j \cdot GradoDePertenencia_j}{\sum_{j=1}^n GradoDePertenencia_j} \quad (5)$$

donde  $n$  representa la cantidad total tanto del Similarity Score como del Rating Score el grado de pertenencia.

#### Pasos del algoritmo elaborado:

1. Leer el dataset: obtener los datos del dataset descargados
2. Rating Score: se obtiene calculando el promedio de las calificaciones de los usuarios para cada producto.
3. Similarity Score: se calcula mediante el uso de técnicas de procesamiento de texto, donde se combina el texto de las reseñas y el resumen de los usuarios. La matriz TF-IDF (Frecuencia de Términos - Frecuencia Inversa de Documentos) se utiliza para representar las reseñas como

vectores, y la similitud del coseno se emplea para medir la similitud entre los elementos.

4. Sistema de inferencia difusa: se utiliza para clasificar las recomendaciones en diferentes niveles.
5. Defusificador: se realiza mediante un método de centroide ponderado, que combina de manera ponderada la similitud y el rating score para tomar una decisión final sobre la recomendación.

### 3. CONTENIDOS

Esta sección describe los datasets que fueron seleccionados para llevar a cabo las pruebas del funcionamiento del algoritmo de recomendaciones que fue desarrollado, así como también se presenta en detalle todos los módulos que fueron implementados en el sistema.

#### 3.1 EXPERIMENTOS.

De los distintos datasets presentados en “Amazon Review Data” [3], se escogieron tres para realizar las distintas pruebas de funcionamiento del sistema. Dichos conjuntos de datos se describen a continuación:

**3.1.1 Amazon Fashion:** Conjunto de datos que contiene información de productos de moda y que cuenta con 3,176 reviews.

**3.1.2 Gifts Cards:** Este conjunto de datos contiene información de tarjetas de regalo y cuenta con 2,972 reviews.

**3.1.3 Magazine Subscriptions:** Conjunto de datos que contiene información sobre suscripciones a revistas y cuenta con 2,375 reviews.

Estos conjuntos de datos incluyen información como:

- a. Identificadores del producto(ASIN).
- b. Categorías de los productos.
- c. Reseñas de los usuarios.
- d. Calificaciones de los productos.

#### 3.2 Detalles Técnicos.

El sistema de recomendación propuesto cuenta con cinco módulos, a saber, un lector de datasets, rating score, similarity score, un sistema de inferencia y un defusificador.

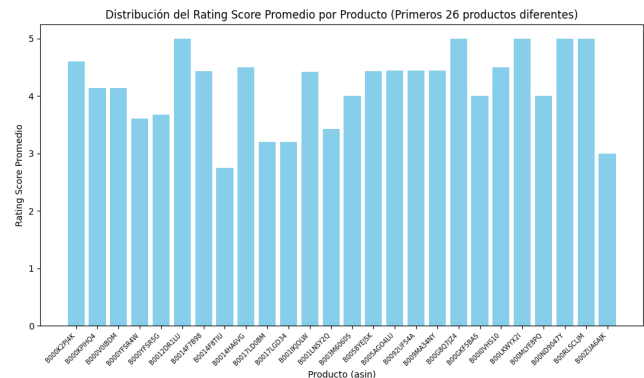
##### 3.2.1 Módulo 1. Lector de Datasets.

Este módulo se encarga de realizar la lectura de cualquier dataset proporcionado para luego crear una estructura de datos adecuada para la representación del conjunto de datos.

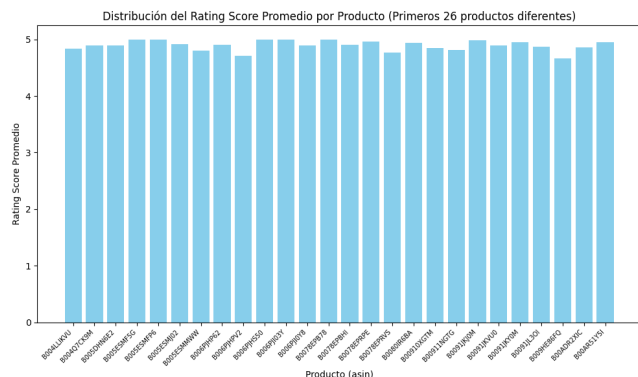
Recibe como parámetro una ruta de archivo y devuelve un dataframe de Pandas.

##### 3.2.2 Módulo 2. Rating Score.

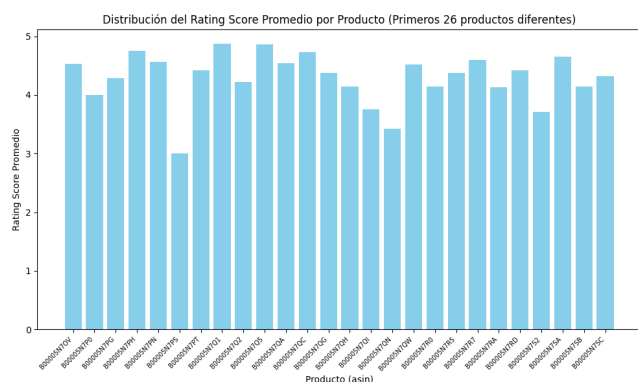
El objetivo de este módulo es encontrar el promedio del rating de cada producto dentro de un conjunto de datos. La función recibe como parámetro el dataframe que fue creado con el módulo anterior y nos devuelve otro dataframe, el cual contiene el rating score de cada producto. Las figuras 1, 2 y 3 representan ejemplos limitados de la salida del rating score.



**Figura 1.** Gráfico del rating score promedio por producto de 26 elementos utilizando el dataset de Amazon Fashion.



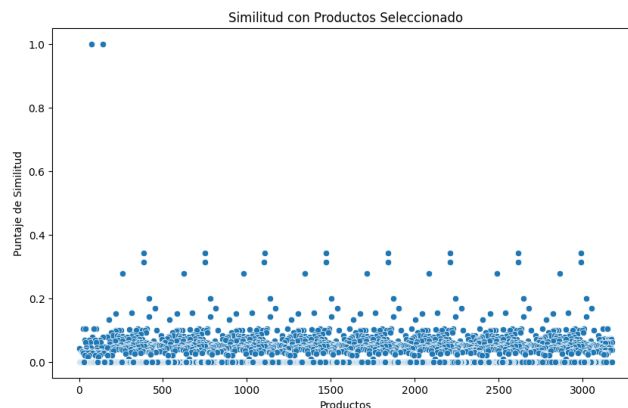
**Figura 2.** Gráfico del rating score promedio por producto de 26 elementos utilizando el dataset de Gifts Cards.



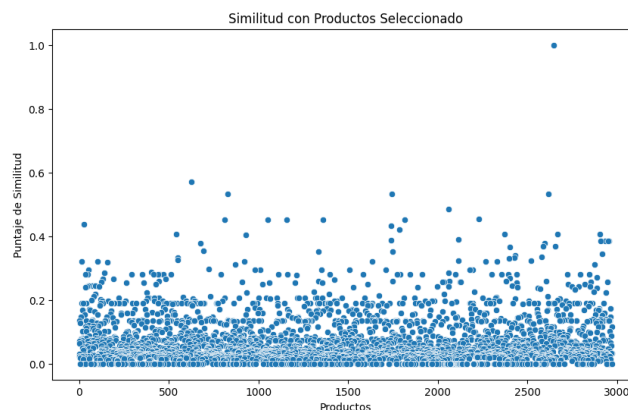
**Figura 3.** Gráfico del rating score promedio por producto de 26 elementos utilizando el dataset de Magazine Subscriptions.

### 3.2.3 Módulo 3. Similarity Score.

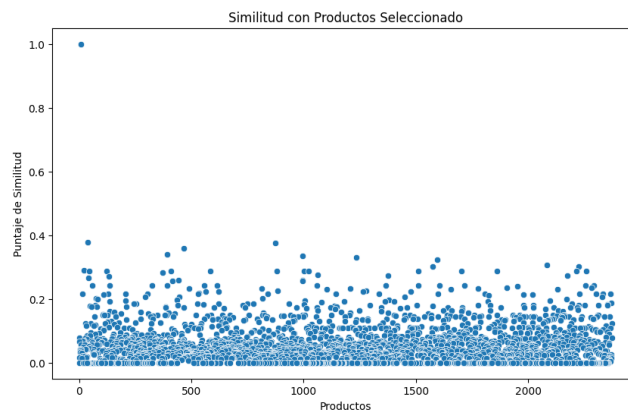
El propósito de este módulo es calcular la similitud entre productos teniendo en cuenta las reseñas de los usuarios utilizando el método del coseno y la representación TF-IDF. Recibe un dataframe como parámetro y devuelve una lista que contiene la similitud entre productos. La similitud se calcula por medio del método del coseno que consiste en retornar valores entre 1 y -1, en los que los valores próximos a 1 indican una similitud perfecta, próximos a 0 representan que no existe similitud y -1 indican similitud negativa. Las figuras 4, 5 y 6 representan las similitudes de manera a visualizar como considerando un producto este realiza las similitudes con el resto de productos pertenecientes al dataset.



**Figura 4.** Gráfico del Similarity Score de un producto con todos los demás productos usando el dataset Amazon Fashion.



**Figura 5.** Gráfico del Similarity Score de un producto con todos los demás productos usando el dataset Gift Cards.



**Figura 6.** Gráfico del Similarity Score de un producto con todos los demás productos usando el dataset Magazine Subscriptions.

**Pasos para implementar el módulo:**

1. Concatenar texto de reseñas y resumen de revisión.
2. Preprocesamiento de datos.
3. Calcular similitud del coseno.
4. Crear una matriz de similitud.
5. Filtrar y ordenar similitudes.
6. Almacenar los resultados.
7. Mostrar resultados.
8. Devolver resultados.

### 3.2.4 Módulo 4. Sistema de Inferencia Difusa.

Este módulo implementa un sistema de inferencia difusa para obtener el nivel de recomendación para productos basados en la similitud del coseno y el rating de productos. Por ende, recibimos como parámetros el similarity score y el rating score, para luego devolver su nivel de recomendación. El sistema está basado en un conjunto de reglas difusas, las cuales se observan en la Tabla 1.

REGLAS DIFUSAS	
REGLA 1	<i>If rating_score_product = excellent and similarity_score=excellent then Recommendation_Level = Muy recomendable</i>
REGLA 2	<i>If rating_score_product = excellent and similarity_score=good then Recommendation_Level = Muy recomendable</i>
REGLA 3	<i>If rating_score_product = good and similarity_score=excellent then Recommendation_Level = Muy recomendable</i>
REGLA 4	<i>If rating_score_product = average and similarity_score=excellent then Recommendation_Level = Muy recomendable</i>
REGLA 5	<i>If rating_score_product = average and similarity_score=good then Recommendation_Level = Recomendable</i>
REGLA 6	<i>If rating_score_product = average and similarity_score=average then Recommendation_Level = Recomendable</i>
REGLA 7	<i>If rating_score_product = excellent and similarity_score=good then Recommendation_Level = Recomendable</i>
REGLA 8	<i>If rating_score_product = avergae and similarity_score=average then Recommendation_Level = Es probable que recomiende</i>

REGLA 9	<i>If rating_score_product = average and similarity_score=average then Recommendation_Level = Recomendable</i>
REGLA 10	<i>If rating_score_product = average and similarity_score=average then Recommendation_Level = No recomendable</i>
REGLA 11	<i>If rating_score_product = poor and similarity_score=average then Recommendation_Level = No recomendable</i>
REGLA 12	<i>If rating_score_product = excellent and similarity_score= good then Recommendation_Level = Recomendable</i>

Tabla 1. Reglas difusas

### Tablas de Resultados obtenidos aplicando las reglas:

Asin	Recommen dation Level	Similarity score	Rating score
B00BXLTUBA	Highly Recommen ded	0.559	4.8518518 51851852
B006PJHP62	Highly Recommen ded	1.000	4.9047619 04761905
B01E4QPDV6	Highly Recommen ded	0.649	4.8717948 71794871
B00CXZPG00	Highly Recommen ded	0.565	5.0

Tabla 2. Algunas salidas utilizando el dataset Gift Card.

Asin	Recomm endation Level	Similarity score	Rating score
B000ILVRO4	Recomm ended	0.627	2.571428571 4285716
B00005N7Q1	Highly	0.757	4.875

	Recomm ended		
B00005N7QW	Highly Recomm ended	0.679	4.517241379 310345
B00007G2X5	Recomm ended	0.405	2.8

**Tabla 3.** Algunas salidas utilizando el dataset Magazine Subscriptions

Asin	Recom mendat ion Level	Similarity score	Rating score
B00GKF5BAS	Highly Recom mended	1.000	4.0
B010RRWKT4	Highly Recom mended	1.000	4.44148936 1702128
B009MA34NY	Highly Recom mended	1.000	4.44297082 22811675
B014IBJKNO	Highly Recom mended	0.692	4.44266666 6666667

**Tabla 4.** Algunas salidas utilizando el dataset Amazon Fashion..

### 3.4.5 Módulo 5. Defusificador.

Este módulo proporciona una forma de defusificación de los resultados obtenidos en el sistema de inferencia proporcionado en el módulo 4. En términos más simples, el módulo recibe como parámetro un conjunto difuso de los niveles de recomendación de los productos y nos devuelve un valor concreto en relación a ese conjunto difuso.

#### Pasos para desarrollar el defusificador:

1. Definir universos y las funciones de membresía.
2. Calcular los grados de pertenencia.

3. Calcular manualmente el centroide ponderado.
4. Imprimir el resultado.

### Resultados con los datasets del experimento.

Rating Score	4.7847
Similarity Score	0.5055

**Tabla 5.** Método del centroide ponderado salida con el dataset Gift Card.

Rating Score	4.1493
Similarity Score	0.3970

**Tabla 6.** Método del centroide ponderado salida con el dataset Magazine Subscriptions..

Rating Score	4.3408
Similarity Score	0.3823

**Tabla 7.** Método del centroide ponderado salida con el dataset Amazon Fashion.

## 4. CONCLUSIÓN

A través de la investigación, nuestro proyecto ha explorado con éxito la implementación de un sistema de recomendación difusa, destacando la eficacia de la lógica difusa en la gestión de expresiones inciertas para personalizar recomendaciones de productos. Al considerar tanto la similitud de contenido como las calificaciones de los usuarios, nuestro enfoque ha demostrado ser flexible y adaptativo, ofreciendo recomendaciones más precisas y relevantes.

En cuanto a los resultados, la aplicación del método del centroide ponderado en el proceso de defusificación ha permitido una toma de decisiones final que refleja de manera equitativa la importancia de los scores de similitud y las calificaciones de los usuarios. La ponderación cuidadosa de estos elementos ha mejorado significativamente la comprensión de las preferencias individuales, proporcionando recomendaciones más alineadas con las expectativas del usuario.

Para futuras mejoras, se sugiere una atención continua a la calidad y diversidad de los datos de entrada, así como la exploración de nuevas técnicas para la expansión y refinamiento de la base de datos. Además, la aplicación de

este sistema en contextos específicos y la adaptación a diferentes sectores podrían ser áreas de enfoque para proyectos futuros.

En última instancia, nuestro sistema de recomendación difusa ha sentado las bases para un enfoque más preciso y holístico en la entrega de recomendaciones personalizadas. Este proyecto destaca el potencial de la lógica difusa en la mejora de la experiencia del usuario en entornos donde la comprensión profunda de las preferencias individuales es esencial.

## REFERENCES

- [1] R.V. Karthik and S. Ganapathy. 2021. Applied Soft Computing Journal ELSEVIER. .DOI: [Applied Soft Computing | Journal | ScienceDirect.com by Elsevier](https://doi.org/10.1016/j.asoc.2021.106788)
- [2] Timothy J. Ross. Fuzzy Logic with Engineering Applications. McGraw-Hill. 2010.
- [3] Jianmo Ni. 2018. Amazon Review Data. Amazon Product Review Data.DOI: [Datos de revisión de Amazon \(nijianmo.github.io\)](https://doi.org/10.26434/chemrxiv-2018-08-00000)