Pontificia Universidad JAVERIANA Bogotá

Tópicos Avanzados en Analítica

Proyecto II

Educación **Continua**

Continuas oportunidades para crecer

Descripción General Base de Datos

Base de datos:

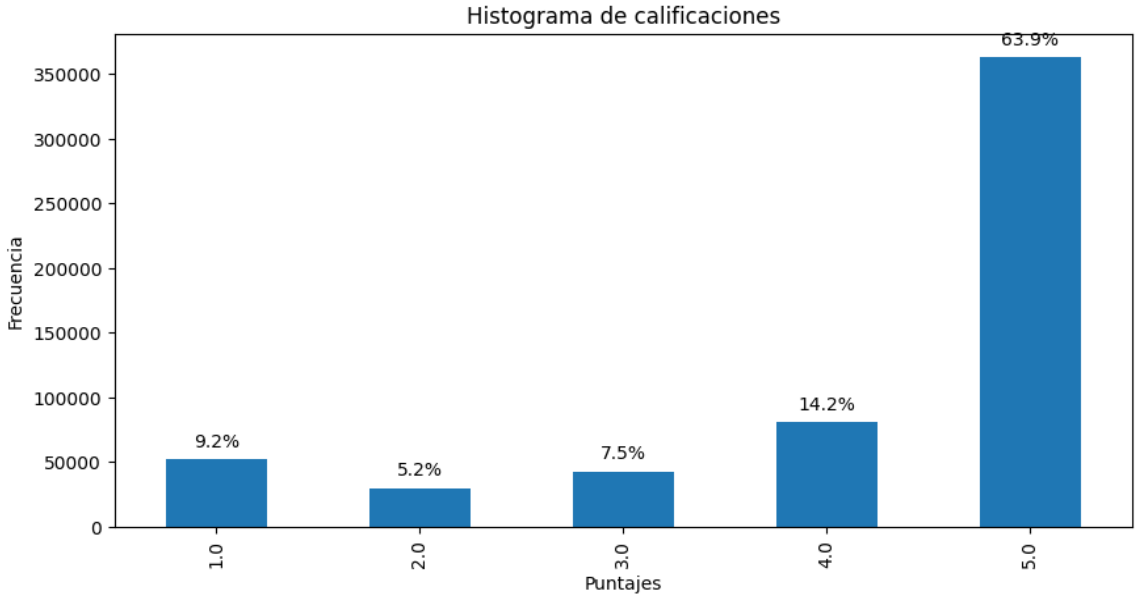
Productos de alimentos gourmet disponibles en Amazon, identificando la percepción de los usuarios frente a estos a través del tiempo y estableciendo las mejores recomendaciones posibles según el contenido de las reseñas y la utilidad de otros consumidores sobre las calificaciones. Para esto se tomará como referencia una base de datos recopilada de Amazon por J. McAuley y J. Leskovec en 2013 y disponible en el siguiente enlace:

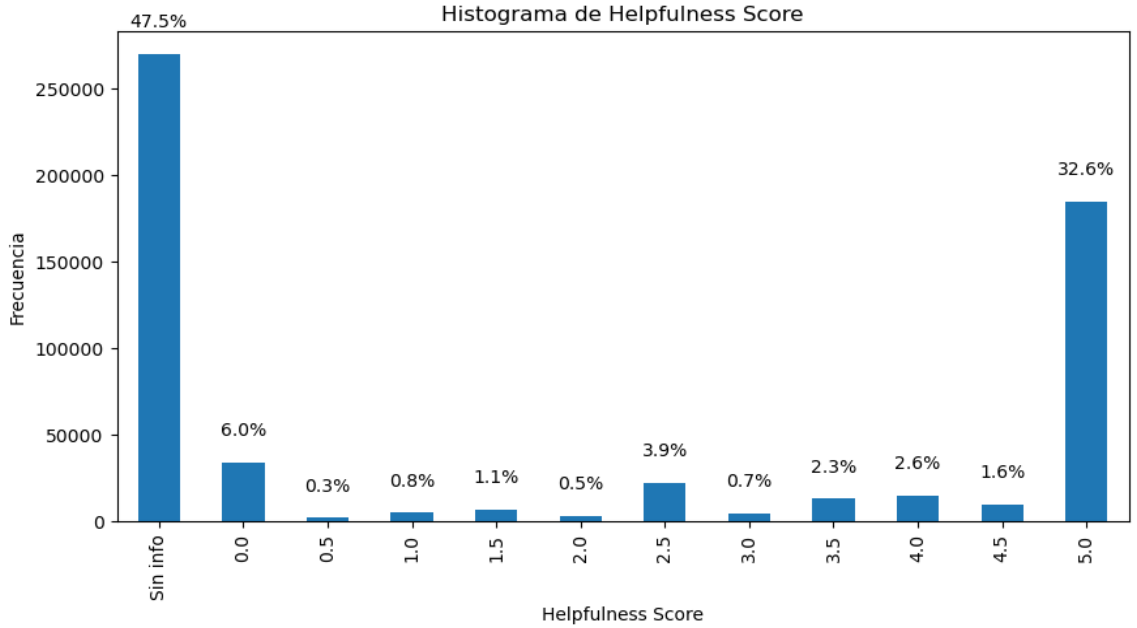
https://snap.stanford.edu/data/web-FineFoods.html

Tabla 1. Diccionario de datos

Variable	Tipo	Rango	Descripción de las variables
ProductID	Texto		Identificador ASIN (particular de Amazon), alfanumérico limitado a 10 caracteres
UserID	Texto		Identificador del usuario
Profilename	Texto		Nombre del perfil del usuario
HelpfulnessNumerator	Numérica	0 – 866	Número de usuarios que encontraron útil la reseña consultada
HelpfulnessDenominator	Numérica	0 – 923	Número total de usuarios que calificaron la reseña
Score	Numérica	1 – 5	Calificación del producto en escala de 1-5
Time	Numérica	*	Serie de 14 años, tiempo en formato UNIX
Summary	Texto		Resumen de la reseña
Text	Texto		Texto completo de la reseña

EDA

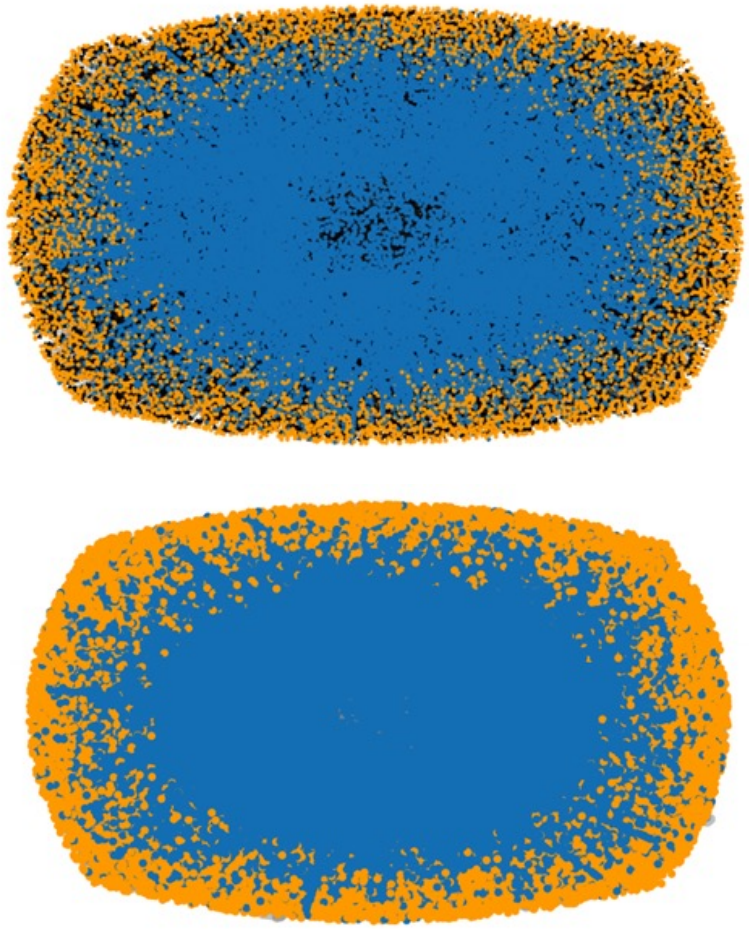




- 568.454 registros, 9 variables
- 74.258 productos
- 256.059 usuarios
- 35% reviews en 2012

Información del grafo

¿ Cuál grafo contiene el peso de los scores?



```
Dataset: Amazon Food Reviews
----------------------------
Numero de grafos: 1
Cantidad de Nodos: 37948
Cantidad de features: 745
Cantidad de clases: N/A

Graph:
------
Se tienen links dirigidos: False
Grafo tiene nodos aislados: False
Grafo tiene self-loops: False
Cantidad de nodos aislados: 0
Cantidad de nodos con grado = 1: 30653
```

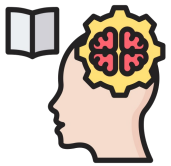
Información del grafo

Análisis de comunidades

Modelo Node2Vec - RecSys

Arquitectura del modelo:

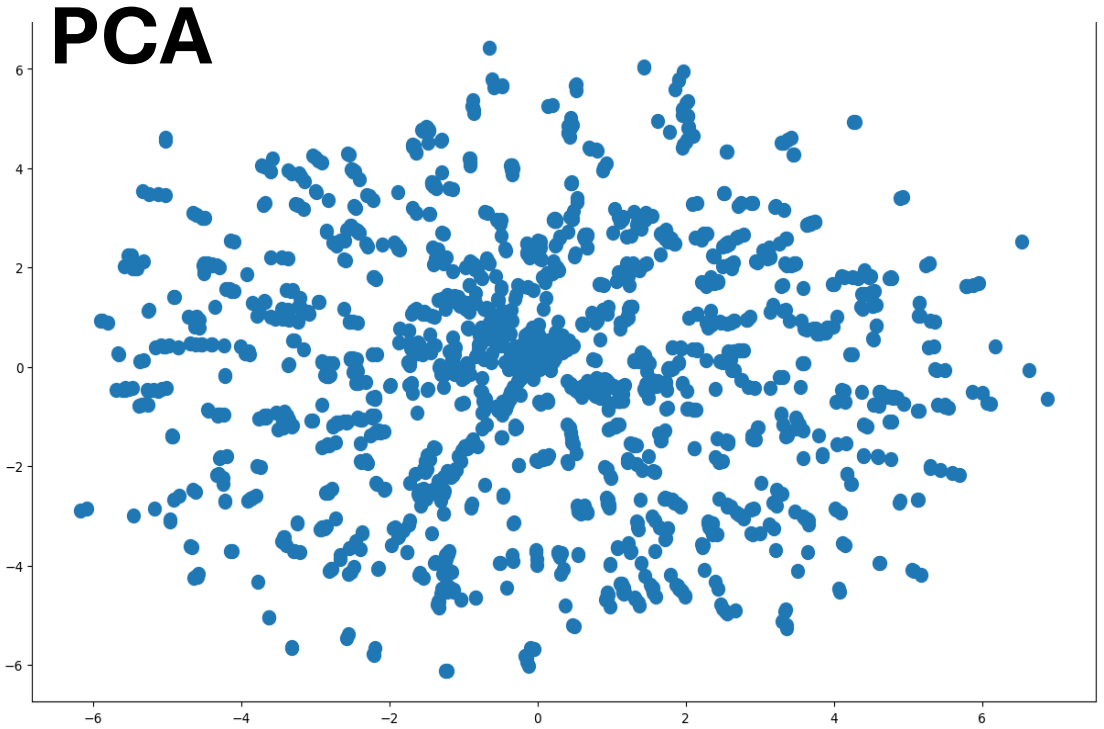
- Filtro por calificación (>= 4) Prob. Compra
- Ocurrencia conjunta (>=5)
- Basado únicamente en productos
- Análisis simplificado



Proceso de Entrenamiento:

- División de Datos 80%-20%
- Secuencias de embeddings
- No cuenta con epochs, caminata aleatoria y longitud de distancia
- Similitud de coseno (pares de productos)
- Generación de datos negativos

	Validación	Prueba
Esc 1 (+)	0,8842	0,8731
Esc 2 (-)	0,8811	0,8696

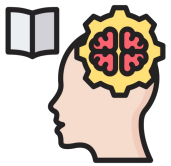


Modelo GCN - Graph Convolutional Network

Arquitectura del modelo:

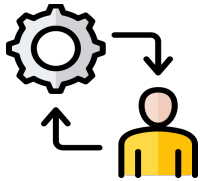
- Convolución con 128 neuronas
- Convolución con 64 neuronas de salida
- Activación ReLU entre las capas

Epoch	Loss	Val	Test
10	0.9946	0.7666	0.7651
20	0.558	0.8009	0.8003
30	0.4007	0.8253	0.8252
...	...	...	...
100	0.3046	0.8397	0.8336



Proceso de Entrenamiento:

- División de datos: 80% - 20%
- Optimización con Adam y tasa de aprendizaje de 0.01.
- Función de pérdida: binary_cross_entropy_with_logits.



Variación hiperparámetros:

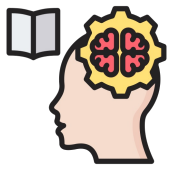
- Activación: ReLu y LeakyReLu
- Wight decay
- Early stoping variando paciencia
- Dropout

Modelo SAGE - Graph Sage

Configuración del modelo:

- Capas de Convolución: 3
- Canales de Entrada: 745
- Canales Ocultos: 128
- Canales de Salida: 64
- Función de Activación: ReLU
- Dropout: 0.5

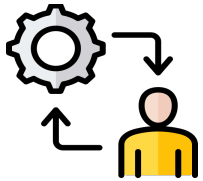
Epoch	Loss	Val	Test
10	0.54	0.48	0.39
20	0.45	0.503	0.5023
30	0.4007	0.53	0.5452
…	…	…	…
100	0.3946	0.59	0.5636



Proceso de Entrenamiento:

- División de datos: 80% - 20%
- Optimización : Adam
- tasa de aprendizaje de 0.005.
- Función de Pérdida: BCEWithLogitsLoss

.



Variación hiperparámetros:

- Se utilizó Optuna para optimizar los hiperparámetros clave.
- Mejores Valores Identificados:
  - Canales Ocultos: 128
  - Canales de Salida: 64
  - Tasa de Aprendizaje: 0.005

Modelo GAT – Graph Attention Network

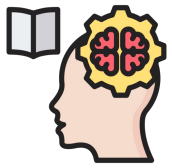
Configuración del modelo:

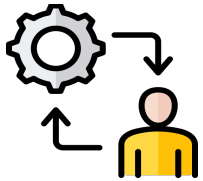
- Capas de Convolución: 2
- Canales de Entrada: 745
- Canales Ocultos: 32
- Dropout: 0.6

Epoch	Loss	Val	Test
10	0.9946	0.7666	0.7651
20	0.558	0.8009	0.8003
30	0.4007	0.8253	0.8252
…	…	…	…
100	0.3046	0.8397	0.8336

Variación hiperparámetros:

Proceso de Entrenamiento:





Evaluación del modelo

Modelo Node2Vec - RecSys

- El modelo realiza un análisis de recomendación mediante la identificación de relaciones entre pares de productos
- Analiza relaciones mediante recorrido desde nodos aleatorios hasta parámetro asignado
- Los resultados del análisis mediante similitud de coseno, incluso al incluir valores negativos muestran buena capacidad de generalización

Modelo GCN - Graph Convolutional Network

- El modelo muestra una notable mejora en la pérdida y precisiones durante las primeras 60 épocas, después de lo cual se estabiliza, lo que sugiere que ha aprendido las características importantes de los datos.
- La cercanía entre las precisiones de validación y prueba indica que el modelo tiene una buena capacidad de generalización.