Facultad de Ingeniería

Tópicos Avanzados de Analítica

CamilaAndrea Arias Vargas
Felipe Clavijo Acosta
Joel Alfredo Márquez Álvarez
Juan Camilo Ramírez Restrepo
Juan Pablo Cuellar Solano

Contenido

PROYECTO FINAL - GRAPH MACHINE LEARNING: Amazon by PyTorch Geometric Datasets

1. Entendimiento del Negocio

Amazon fue fundada en 1994 como una librería en línea, ofreciendo una selección extensa de libros que se organizaban en categorías y subcategorías para facilitar la búsqueda y la navegación. Rápidamente se fue expandiendo a una línea más amplia de productos incrementando su catálogo para incluir diferentes tipos de productos que abarcan electrónica, moda, hogar, alimentos, entre otros, su estrategia de organización y clasificación de productos se volvió fundamental para soportar el crecimiento. Esta expansión hizo que su sistema inicial de categorías evolucionara hacia una estructura más compleja que hacía indispensable una correcta indexación y documentación, ya que sería indispensable manejar el crecimiento de su inventario y asegurar una experiencia de usuario fluida. La estructura inicial de categorías sirvió como una forma de navegación intuitiva, permitiendo a los usuarios acceder a productos mediante filtros y subcategorías dentro su plataforma de comercio electrónico.

El crecimiento de Amazon no solo se quedó en venta de catálogos de productos, sino que se amplió a una gama de servicios adicionales como lo es la transformación tecnológica, llegando a servicios en la nube, dispositivos inteligentes, servicios de streaming y más.

Actualmente, utiliza una combinación de enfoques tradicionales de clasificación y algoritmos avanzados de aprendizaje automático para categorizar productos, para lo cual toma en cuenta información de gran relevancia como datos de los vendedores, búsqueda por palabras clave y principalmente las interacciones de sus clientes en sus plataformas, recopilando estas relaciones de manera efectiva.

A pesar de los avances, el sistema de clasificación actual enfrenta desafíos con la escala y la heterogeneidad de los productos. La incorporación de millones de productos nuevos, la variedad de etiquetas aportadas por los vendedores y los cambios constantes en las tendencias de compra de los clientes generan problemas de consistencia y precisión. Además, a medida que Amazon sigue diversificándose, se requiere mayor desarrollo y desempeño de tecnologías que a s vez puedan ser empleadas para realizar clasificaciones cada vez más eficientes y escalables para organizar este inventario masivo.

Para mantener su liderazgo y eficiencia, Amazon se enfrenta a la necesidad de desarrollar un sistema de clasificación de productos más avanzado y autónomo que pueda adaptarse en tiempo real, escalar sin problemas en la infraestructura y optimizar costos dado el tamaño de su inventario, teniendo en cuenta la relación y conexión que se presenta entre los diferentes productos.

JUSTIFICACION: USO DE GRAFOS.

Utilizar un modelo de grafos para la clasificación de productos en Amazon es relevante y bastante interesante, ya que este enfoque aborda de manera efectiva las complejidades y el tamaño del inventario, por lo cual la empresa podría beneficiarse en varios puntos clave:

- Por la complejidad en sus interacciones un modelo de grafos permite capturar las relaciones entre productos de forma natural y detallada, algo que los métodos tradicionales encuentran difícil. Por ejemplo, Amazon puede representar en el grafo cómo los productos se relacionan no solo por las compras conjuntas de usuarios, sino también por sus características compartidas, como marcas, categorías, o similitud en reseñas y valoraciones. Esto ayuda a generar recomendaciones de productos que reflejan mejor las relaciones reales entre ellos y optimiza el descubrimiento de productos para los clientes.

- En los grafos cada producto se convierte en un nodo con atributos específicos y las conexiones representan su relación con otros productos. Esta estructura permite que se considere simultáneamente las características individuales de cada producto y sus relaciones, proporcionando una perspectiva integral que mejora la precisión en la clasificación y la relevancia de las recomendaciones, lo cual es clave para mantener la lealtad de los clientes en una plataforma con tanta diversidad de productos.
- Los grafos ofrecen la ventaja de una alta capacidad de ser escalables, lo cual es fundamental dada la constante incorporación de nuevos productos y las variaciones en las relaciones entre ellos, teniendo en cuanta el crecimiento potencial que una plataforma como esta tiene, permitiendo que nuevos productos y clientes, con productos nuevos se puedan integrar a las conexiones del sistema sin necesidad de reconstruir el modelo nuevamente, manteniendo una clasificación y recomendaciones eficientes y actualizadas en tiempo real.
- La mejor representación de las conexiones en los productos, además de generar una mejor clasificación y potenciar unas mejores recomendaciones de producto, puede influenciar positivamente en los clientes, mejorando la relación con los mismos, al dar una mayor experiencia de usuario, generando mayor satisfacción en el uso de la plataforma, reflejando una mayor lealtad.

2. Objetivos de Negocio

- Mejorar la experiencia del usuario a través de recomendaciones personalizadas que anticipen y se ajusten a sus preferencias, aumentando la satisfacción y fidelidad hacia la plataforma, e impactando positivamente en las ventas y en la relación cliente-compañía. Es por esto que se espera que las recomendaciones generadas sean capaces de cautivar tanto a potenciales clientes como a nuevos retailers, fomentando su compromiso a largo plazo con la plataforma. Para esto se tendrá en cuentan tres indicadores importantes (De mayor prioridad a menor) en relación con cliente y la satisfacción de este, como es capacidad de retención de este.

  Indicador 1: **Customer Lifetime Value.**

  Descripción: Es un estimador del valor total que un cliente generará para la empresa durante toda su relación.

  Fórmula:

$$CLV = \frac{Ingreso\ promedio\ por\ cliente}{Tasa\ de\ desercion} * Duracion\ media\ de\ la\ relacion$$

Meta: Incrementar del 15-20% en el CLV en un período de 1 año.

Indicador 2: **Recommendation Diversity**

Descripción: Mide la variedad de productos recomendados a los usuarios, evaluando si las recomendaciones son variadas o si el sistema tiende a recomendar productos similares repetidamente.

Fórmula:

$$RD = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} similitud\ (P_i, P_j)$$

donde:
- P_i, P_j son los productos recomendados.
- N es el número total de productos recomendados.
- $similitud\ (P_i, P_j)$ es la similitud entre los productos P_i, P_j, medida usando técnicas como la similitud de coseno.

Meta: Mantener la diversidad en un rango alto, es decir, por encima del 0,6.

Indicador 3: **Tiempo de Duración en la APP**

Descripción: Mide la duración en minutos y segundos de las personas dentro de la APP

Meta: Aumentar un 10% el tiempo en uso de la APP, superando la barrera de los 4.35 minutos en promedio de uso.

3. Objetivo de minería

- Desarrollar un sistema de clasificación de productos (computadores) basado en grafos que permita categorizarlos correctamente (10 productos). Este sistema se diseñará para ser escalable y adaptable para su uso en tiempo real, contribuyendo a mejorar la experiencia de usuario mediante categorías precisas y relevantes, al facilitar la navegación y exploración de productos en la plataforma.

Indicador: **Accuracy**

Descripción: Mide la exactitud del sistema de clasificación en la categorización de los productos, expresado como el porcentaje de productos correctamente clasificados sobre el total de productos clasificados.

Meta: Mantener la métrica en un rango alto, es decir, por encima del 0,85.

4. Entendimiento de los datos

Los datos utilizados provienen del dataset Amazon, proporcionado por PyTorch Geometric. Este dataset contiene información detallada sobre productos, así como reseñas de usuarios, las cuales se emplean como características de los nodos en el modelo. El dataset está disponible en el siguiente enlace: **Amazon Dataset Documentation - PyTorch Geometric.**

PyTorch Geometric es una extensión de PyTorch optimizada para el aprendizaje profundo en datos con estructura de grafo. Esta librería ofrece una serie de herramientas y modelos predefinidos que simplifican la implementación y experimentación con técnicas avanzadas de análisis de grafos, como GNNs, GCNs y GATs.

Se utilizó la base de datos Amazon Computers, que contiene 13,752 nodos que representan productos relacionados con computadores y 491,722 aristas que representan las conexiones o relaciones entre ellos, como productos que se compran frecuentemente juntos. Cada nodo cuenta con 767 características que corresponden a las reseñas de productos transformadas en una representación de bolsa de palabras (bag-of-words), y el conjunto de datos incluye 10 categorías de productos diferentes.

La imagen 1 muestra la representación gráfica del conjunto de datos. En el grafo, cada punto azul representa un producto y las líneas que conectan los puntos indican relaciones entre productos que suelen comprarse juntos. Este grafo refleja la alta densidad de conexiones entre los productos y evidencia cómo ciertos artículos están más relacionados que otros, lo cual es útil para identificar patrones de compra y asociaciones fuertes entre productos.

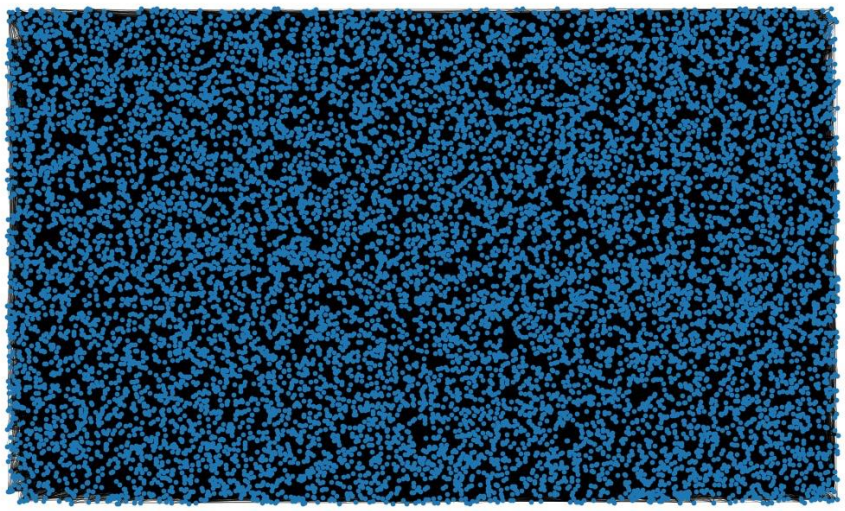

Imagen 1. Grafo productos de computación Amazon

El análisis del grafo usando el spring layout, muestra las conexiones de un nodo específico con otros productos, es más efectivo que el análisis del grafo completo ya que permite un enfoque más detallado en las relaciones directas de un producto particular. Muestra cómo algunos productos tienen una gran cantidad de relaciones directas, mientras que otros están más aislados. Este tipo de

análisis es particularmente útil para identificar productos con muchas conexiones, lo que sugiere una alta popularidad o relevancia con relación a los productos de computación. Además, permite identificar productos menos conectados donde se podrían aplicar estrategias de marketing específicas para incrementar su visibilidad y ventas.

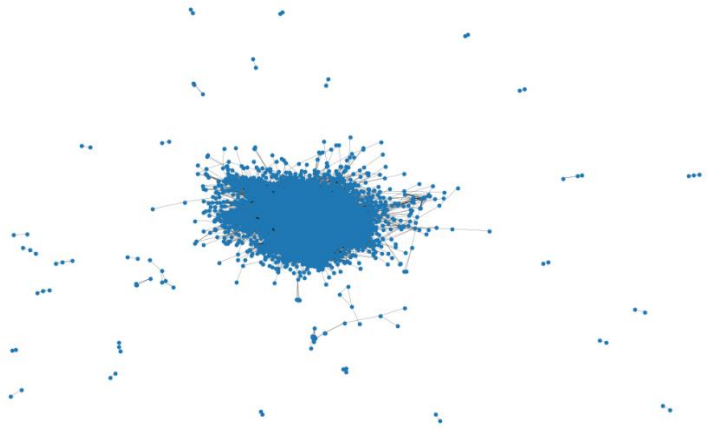

Imagen 2. Grafo de ego con Spring Layout

En promedio, cada nodo está conectado a 57.69 otros nodos, lo cual indica que los productos tienden a tener multiples conexiones entre sí. Esta métrica permite evaluar el grado de correlación entre productos y entender mejor las posibles asociaciones.

La longitud promedio de las rutas más cortas entre todos los pares de nodos fue de aproximadamente 1.98, lo cual significa que, en promedio, se necesitan menos de aristas para conectar cualquier par de productos, lo que refuerza la idea de que los productos está altamente conectados entre sí. El diámetro del grafo, que representa la mayor distancia posible entre dos nodos, fue de 2. Esto indica que, para conectar cualquier par de nodos, se necesitan a lo sumo dos aristas, subrayando la alta conectividad del grafo. La distribución de las longitudes de los caminos más cortos muestra que la mayoría de los productos están conectados en una o dos aristas Imagen 3.

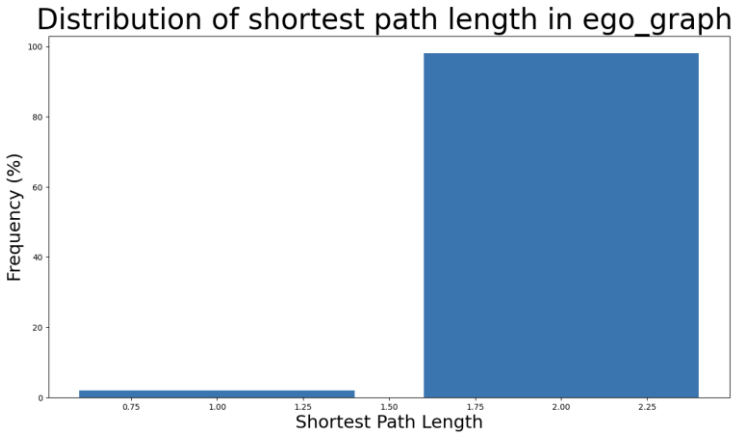

Imagen 3 Longitud de ruta más corta

La densidad del grafo se calculó en 0.019, lo cual indica que solo el 1.9% de todas las conexiones posibles entre los nodos están presentes. Aunque el grafo no está completamente conectado, la densidad es suficiente para identificar patrones significativos y relaciones entre productos. La centralidad de grado mide la cantidad de conexiones de cada nodo. En este caso, se identificaron los productos con mayor cantidad de enlaces dentro del grafo, es decir, aquellos productos con mayor conexión con otros, que se podrían considerar como los más populares o con mayor influencia en el contexto del conjunto de datos. El histograma de la centralidad de grado muestra que la gran mayoría de los nodos tiene un grado de conexión bajo, mientras que solo unos pocos nodos tienen un alto grado de conexión Imagen 4 y 5.

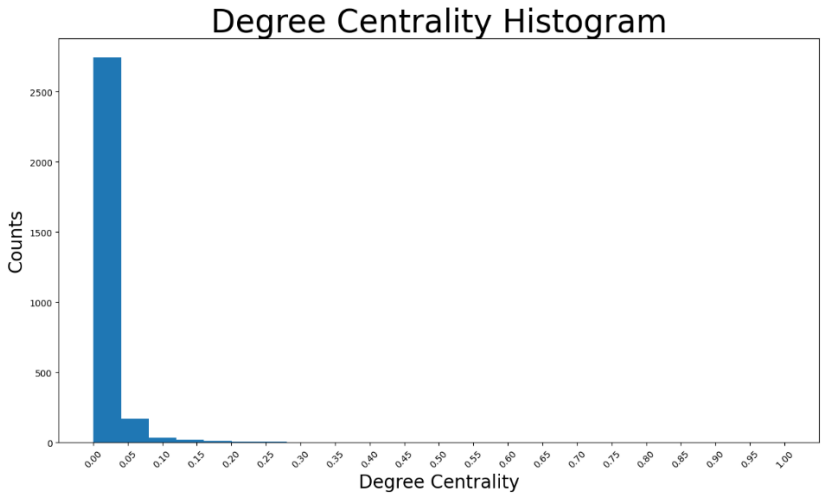

Imagen 4 Grado de centralidad

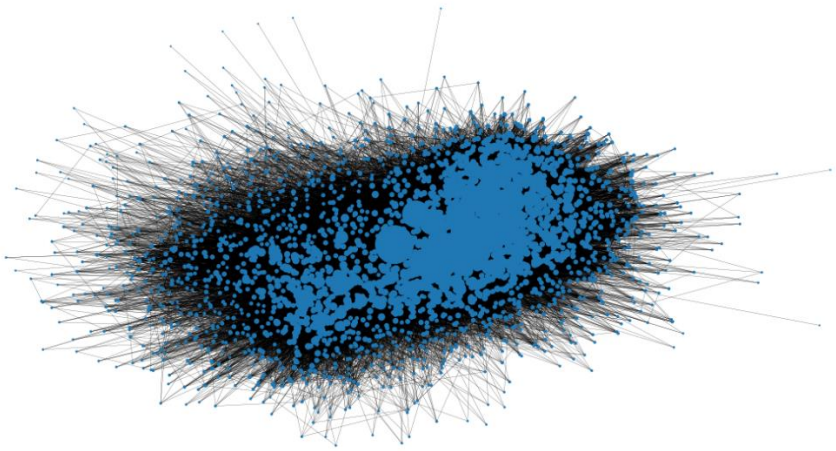

Imagen 5. Grafo de ego

La centralidad de intermediación se refiere a la capacidad de un nodo de actuar como puente entre otros nodos, es decir, medir cuántas veces un nodo aparece en los caminos más cortos entre otros nodos. En este grafo, se identificaron aquellos productos que tienen un alto grado de intermediación, sugiriendo que estos

productos actúan como puntos clave para conectar diferentes partes de la red. La Imagen 6 muestra la distribución de la centralidad de intermediación.

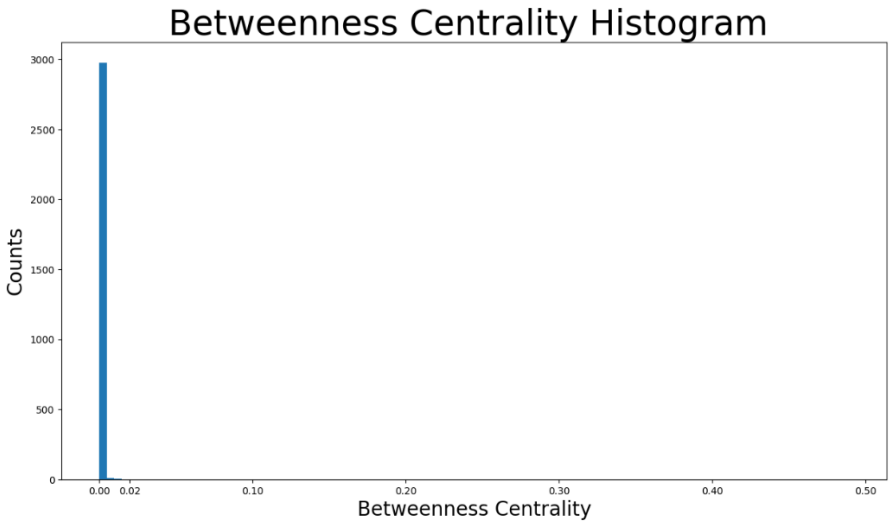

Imagen 6 Histograma de centralidad intermedia

Finalmente, se muestra una representación visual del grafo con los nodos dimensionados según su centralidad de grado y centralidad de intermediación. La visualización permite observar cómo se distribuyen las conexiones en la red, destacando los nodos más importantes que actúan como centros de conexión dentro del grafo. Imagen 7.

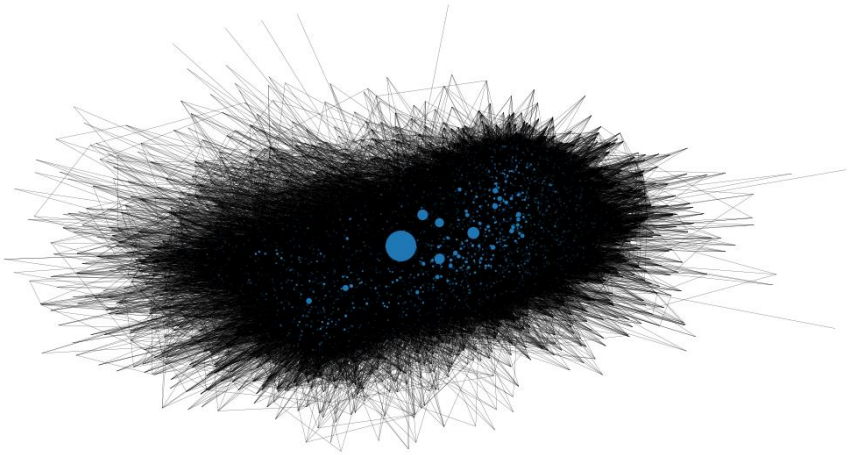

Imagen 7. Grafo de centralidad

El coeficiente de agrupación promedio del grafo es de 0.4257, lo cual indica una tendencia moderada de los nodos a formar clústeres o comunidades locales. Además, la transitividad de la red, con un valor de 0.1136, sugiere que la estructura general del grafo es menos cohesiva, con comunidades más fragmentadas.

Podemos observar que el grafo tiene una estructura densa en el centro, lo cual indica que existen muchos productos estrechamente relacionados, formando grupos claramente definidas. Estas son los grupos de productos que los usuarios suelen comprar juntos, lo cual es particularmente útil para hacer recomendaciones. Los colores diferentes ayudan a visualizar cómo los productos se agrupan en función de sus relaciones con otros.

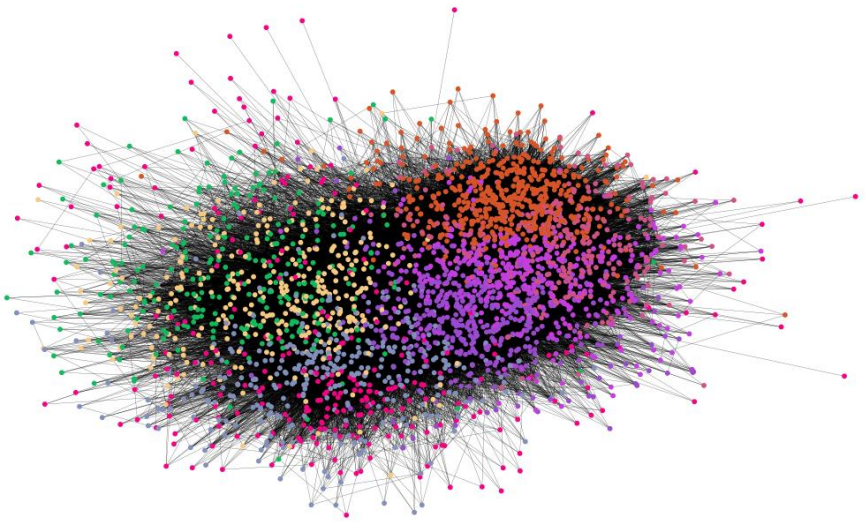

Imagen 8 Grafo de comunidades

Los nodos periféricos, que están conectados a la estructura principal por pocas aristas, indican productos que tienen menos relaciones en comparación con otros. Estos productos podrían beneficiarse de estrategias específicas de marketing para incrementar su visibilidad, por ejemplo, promociones que los asocien a productos más populares.

5. Preparación de los datos

En la fase de preparación de los datos, se realizó la generación de la matriz de adyacencia y la obtención de embeddings de nodos mediante Node2Vec.

5.1 Matriz de Adyacencia

Se generó la matriz de adyacencia del grafo, la cual representa las conexiones entre los nodos en forma de matriz binaria. Para ello, se utilizó el índice de aristas del conjunto de datos, convirtiéndolo en un DataFrame con las columnas 'start_node' y 'end_node'. Luego, se creó una representación densa de la matriz de adyacencia. Adicionalmente, se añadió una matriz identidad para incluir los lazos propios de los nodos, asegurando que cada nodo se conecte consigo mismo.

5.2 Node2Vec

Para la generación de embeddings de los nodos, se utilizó el método Node2Vec. Se definieron los parámetros necesarios para ejecutar las caminatas aleatorias sobre el grafo, incluyendo el número de caminatas (num_walks = 10), la longitud de las caminatas (walk_length = 80), y los parámetros de exploración y retorno (p = 1.0 y q = 0.5). se definió la función next_node para calcular la probabilidad de transición a los nodos vecinos. Posteriormente, la función random_walk permitió realizar caminatas aleatorias a lo largo del grafo, generando secuencias de nodos que se utilizaron como base para entrenar el modelo de Node2Vec.

6. Modelos

Para el desarrollo del sistema de clasificación de productos basados en gráficos, se implementarán y evaluarán distintos modelos de aprendizaje automático y de aprendizaje profundo, tanto tradicionales como basados en gráficos. El objetivo es seleccionar el modelo que mejor se ajuste a la tarea de categorización de productos (computadores) y optimizar la experiencia del usuario. Todos los modelos implementados incluyen un proceso de búsqueda de hiperparámetros para asegurar un rendimiento óptimo. Los modelos para ejecutar son:

- **Regresión Logística:** Un modelo lineal simple pero efectivo para clasificación binaria y multiclase, que servirá como base de comparación para evaluar el rendimiento de otros enfoques.
- **Random Forest:** Un método de conjunto basado en múltiples árboles de decisión, conocido por su capacidad de manejar datos no lineales y ofrecer mayor precisión que los modelos individuales.
- **GCN (Graph Convolutional Network)**: Un tipo de GNN que aplica operaciones de convolución sobre los gráficos para extraer características enriquecidas de los nodos, útil para clasificar productos basados en sus conexiones y atributos.
- **GAT (Graph Attention Network):** Un modelo que mejora el GCN al incorporar mecanismos de atención, asignando diferentes pesos a las conexiones de los nodos para enfocarse en las relaciones más importantes al clasificar los productos.
- **GATv2Conv (Graph Attention Network v2 Convolution)**: una versión mejorada de GAT que introduce un mecanismo de atención más flexible y efectivo. GATv2Cov permite que los pesos de atención se aprendan de manera más eficiente, capturando mejor las relaciones complejas entre los nodos mediante la matriz de adyacencia y las conexiones relevantes en el grafo.

MODELO	ACCURACY
Regresión Logística	0.8612
Regresión Logística con Matriz de Adyacencia	0.9019
Random Forest	0.8169
Random Forest con Matriz de Adyacencia	0.8292
GCN (Graph Convolutional Network)	0.8844
GAT (Graph Attention Network)	0.9106
GATv2Cov (Graph Attention Network v2 Convolution)	0.9193

7. Evaluación y selección

El modelo Graph Attention Network v2 Convolution (GAT) ha sido seleccionado como el mejor modelo por su rendimiento superior, logrando una precisión de 0,91, el más alto entre los modelos evaluados. Esta elección se debe a su mecanismo de atención avanzada, que asigna diferentes pesos a las conexiones en la matriz de adyacencia, permitiendo al modelo enfocarse en las relaciones más relevantes entre los nodos. Esta capacidad mejora significativamente la precisión en la clasificación, superando modelos como la Regresión Logística y Random Forest. Además, GAT es adaptable y escalable, lo que lo hace ideal para aplicaciones en tiempo real y mejora la experiencia de usuario mediante una categorización más precisa y relevante de los productos en la plataforma.

8. Conclusiones

El desarrollo de modelos basados en grafos es crucial para Amazon, ya que permite una gestión más eficaz y detallada de su vasto inventario de productos al capturar las complejas interacciones entre ellos. Estos modelos, como las Graph Convolutional Networks (GCNs) y las Graph Attention Networks (GATs), mejoran la precisión de la clasificación y las recomendaciones al considerar tanto las características individuales de los productos como sus relaciones en el grafo. Esto contribuye a una experiencia de usuario mejorada, con recomendaciones personalizadas que refuerzan la fidelidad del cliente y optimizan el potencial de ventas.

Sin embargo, implementar estos modelos implica un costo computacional significativo. La ejecución y el mantenimiento de grafos densamente conectados, como los que maneja Amazon, requieren recursos computacionales intensivos y optimización continua. Por ello, es esencial balancear los beneficios en precisión y escalabilidad con la eficiencia computacional, empleando técnicas de optimización y selección de modelos para minimizar costos sin comprometer el rendimiento, así como el correcto procesamiento y análisis de los datos para darle solución a las necesidades del negocio.

9. Reporte

TAREA DESARROLLADA / INTEGRANTE	ENTENDIMIENTO DE NEGOCIO	OBJETIVOS DE NEGOCIO	ENTENDIMIENTO DE LOS DATOS	PREPARACIÓN DE LOS DATOS	MODELOS					EVALUACIÓN Y SELECCIÓN	CONCLUSIONES
					Regresión Logística	Random Forest	GNN	GCN	GAT		
Camila Andrea Arias Vargas				X			X	X	X	X	
Felipe Clavijo Acosta			X		X					X	
Joel Alfredo Márquez Álvarez	X	X				X				X	X
Juan Camilo Ramírez Restrepo				X			X	X	X	X	
Juan Pablo Cuellar Solano	X	X				X				X	X

I. Bibliografía

PyTorch Geometric. (s.f.). Amazon Dataset Documentation. PyTorch Geometric Documentation. Recuperado de https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Amazon.html#torch_geometric.datasets.Amazon

Shchur, O., Mumme, M., Bojchevski, A., & Günnemann, S. (2019). Pitfalls of Graph Neural Network Evaluation. NeurIPS 2018 Workshop on Relational Representation Learning (R2L). Recuperado de https://arxiv.org/abs/1811.05868