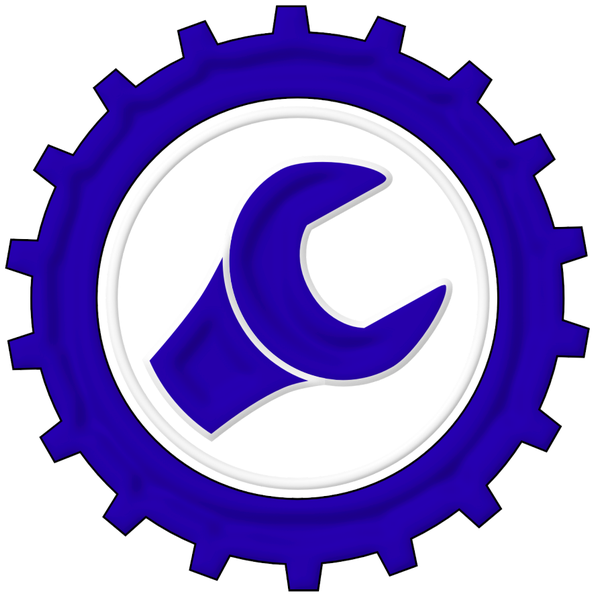
amazon

Sistema de clasificación

John Sebastian Martinez
Carlos Felipe Mora
Andres Parra
Ian Nicolas Rincon
Rosemary Rios Pulido

Descripción General



Amazon desea mejorar su sistema de clasificación de 10 categorias de grupos con el fin de incrementar las ventas y mejorar la experiencia del usuario.

Estado actual:

El sistema de clasificación se basa únicamente en productos comprados juntos.

¿Qué desean?

Implementar un sistema más avanzado que también tenga en cuenta las características de los productos y las relaciones entre ellos.

Objetivos

General:
Implementar un sistema de clasificación basado en grafos utilizando el dataset Amazon de PyTorch Geometric.

Específicos:

- Mejorar la precisión
- Incrementar las ventas
- Mejorar la experiencia del usuario

Relevancia

- Competitividad en el mercado
- Volumen y diversidad de productos
- Comportamiento del Usuario

Utilidad Modelo de Grafos

- Permiten modelar las relaciones complejas entre los productos
- Cada producto puede ser representado como un nodo en el grafo con atributos
- Son escalables y adaptables
- se puede aprovechar el poder del aprendizaje profundo para extraer patrones y representaciones latentes de alta calidad

Conjunto de Datos a Emplear

- Productos como Nodos: Cada producto en Amazon es representado como un nodo en el grafo.
- Relaciones de Compra Conjunta como Bordes: Los bordes entre nodos indican que dos productos se compran frecuentemente juntos.

Primera parte descriptiva

Se presentan 491.722 posibles combinaciones de productos relacionados a los computadores

13.752 nodos (productos de computadores) en el grafo

Análisis Descriptivo

Las etiquetas de los nodos están en el rango de 0 a 9.

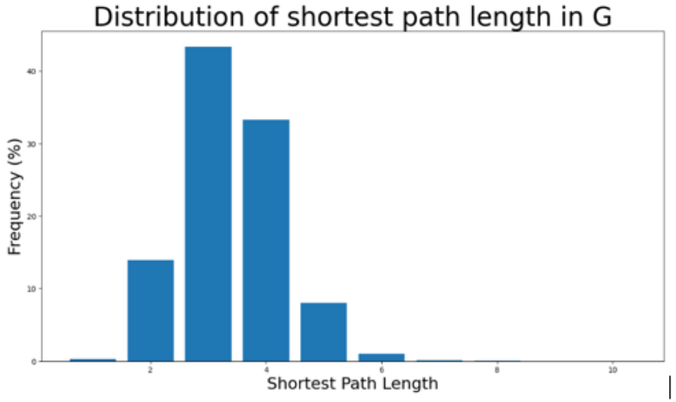
Cada nodo tiene 767 características.

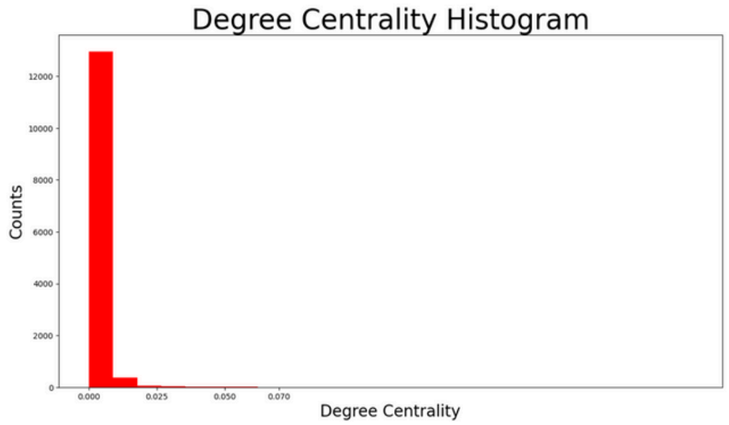
ANÁLISIS DESCRIPTIVO

El camino más corto del producto cero al producto 13.471 se necesita en promedio 3 relaciones

El grafo se divide en 33 grupos de productos interconectados, pero que no tienen conexiones directas entre ellos.

Los productos en Amazon forman diferentes grupos o comunidades que están conectados internamente pero no tienen conexiones directas con otros grupos.





Se necesita en promedio 3 relaciones del producto cero con otros productos para que se elija al mismo tiempo el producto intermedio (6736)

las máximas relaciones que se pueden dar entre productos es de 10 ocasiones

Grado de densidad del 0,27% indicando que los productos no están altamente interconectados en el grafo

Top 5 de productos Amazon

Top de productos	Grado de centralidad	Frecuencia de compra con otros productos
12.888	22,2%	2.992
8.210	18,6%	2.508
8.140	18,5%	2.495
4.528	13,7%	1.850
1.524	11,6%	1.565

Segunda parte descriptiva

Se toma una muestra aleatoria de 1000 nodos

Centralidad intermedia	Centralidad de cercanía	Centralidad del vector propio
8.140	8.140	8.140
8.872	7.939	4.135
904	11.262	13.160
7.797	4.135	3.067
11.262	3.697	5.715

Se identificaron 136 comunidades en el grafo

ANÁLISIS DESCRIPTIVO

En promedio Un 35,1% de los productos están relativamente bien conectados con sus vecinos.

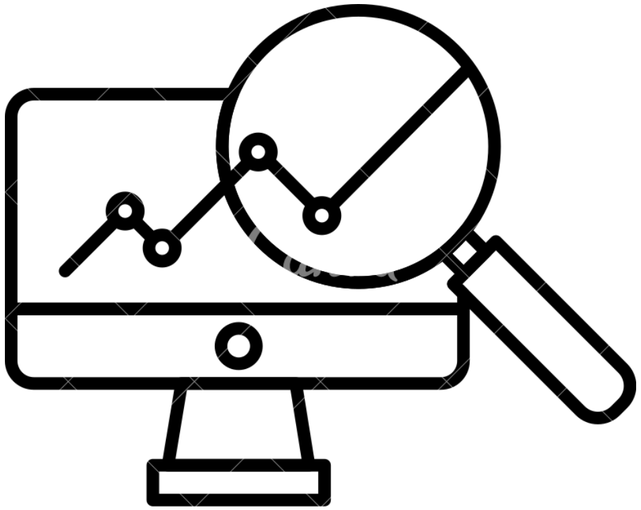
Se presenta 4.582.407 combinaciones de 3 productos que se compran al mismo tiempo.

Existen 342 puentes en el grafo de compra de Amazon, lo que quiere decir que hay muchas conexiones críticas en el grafo de productos de Amazon.

Mas de 700 casos tienen una probabilidad de agruparse cerca de un 38%.

En promedio hay 340 canastas de compras donde se compran los mismos 3 productos a la vez.

Se presenta una tendencia de disasortativity, lo que significa que los productos populares (con alto grado) tienden a estar conectados con productos menos populares (con bajo grado)



Preprocesamiento de los datos

1 CREACIÓN DEL DATAFRAME DE ARISTAS

Construir un DataFrame de pandas con la información de las aristas del grafo.

2 DEFINICIÓN DE MÁSCARAS DE DATOS

Crear máscaras para dividir los nodos en:

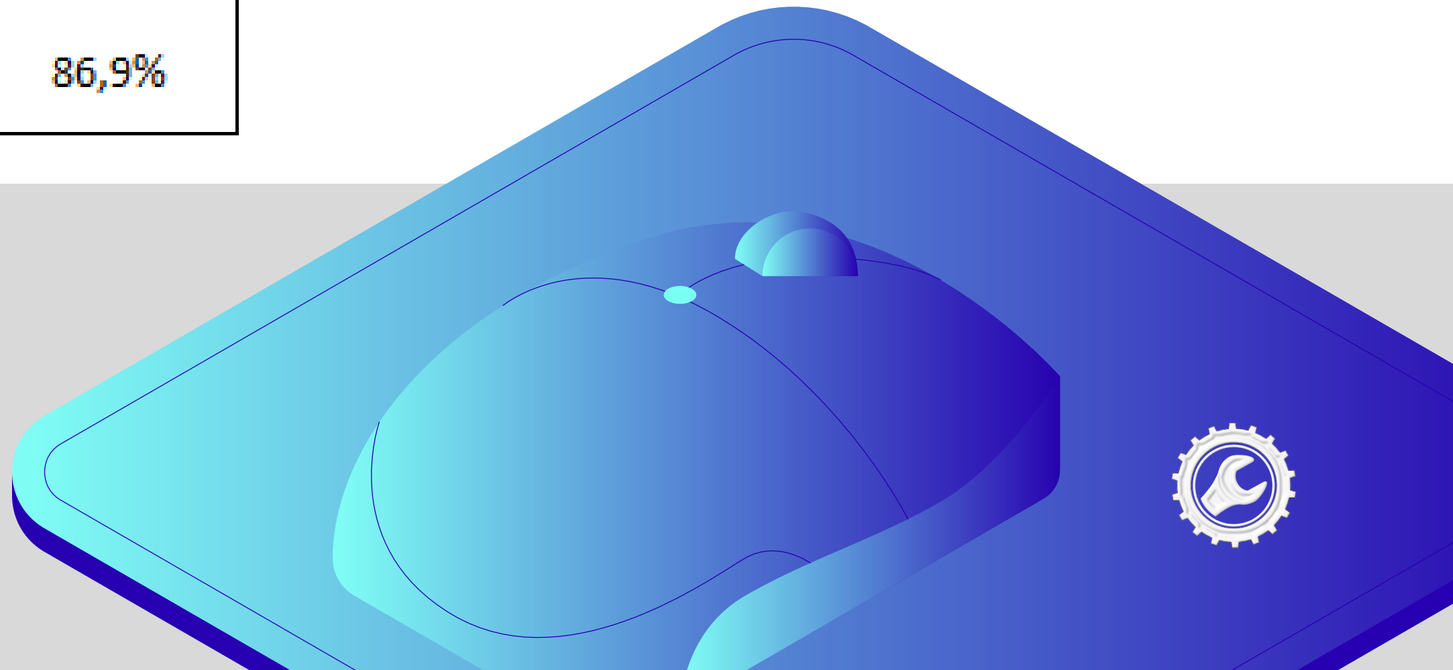
- La máscara de entrenamiento abarcará el 60% de los nodos.
- La máscara de validación cubrirá el 30% de los nodos.
- La máscara de prueba incluirá el 10% de los nodos.

3 DIVISIÓN DEL CONJUNTO DE DATOS

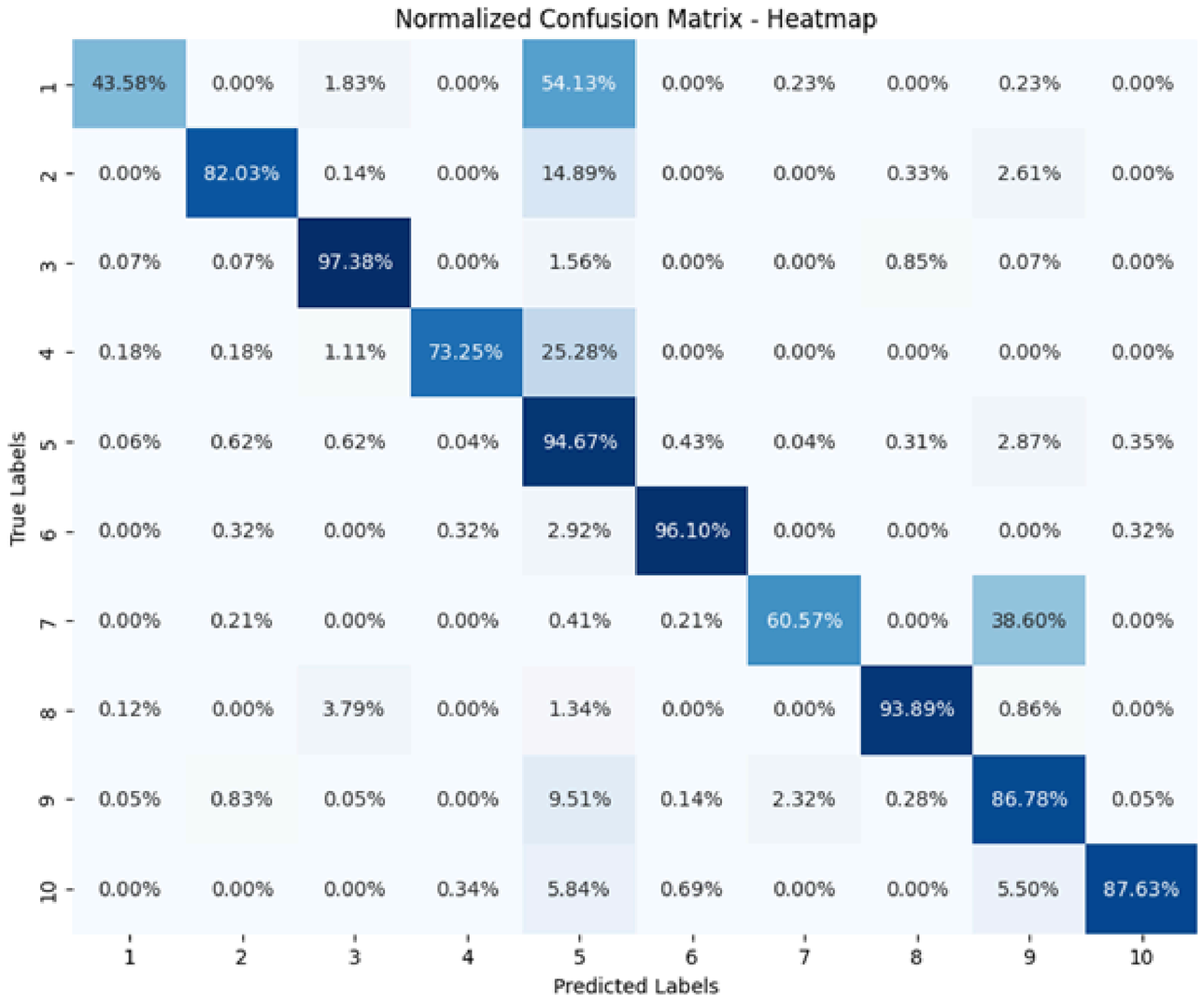
Aplicando cada una de las máscaras al conjunto de datos, se seleccionan los nodos correspondientes para cada conjunto (entrenamiento, validación y prueba)

Modelo de Grafos

Modelo	Acurracy
Red neuronal (GNN sin early stop)	3,4%
Red neuronal 2 (GNN con early stop)	45,4%
Red neuronal convolucional 1	72,0%
Red neuronal convolucional 2	83,3%
Red convolucional 5 capas con regularización	63,0%
Modelo de atención con múltiples	81,6%
Modelo de atención con dos capas de convolución (GAT 2 capas)	86,9%



Matriz de Confusión Modelo Ganador



¡Gracias!