Introducción a Machine Learning

1. Explicar la diferencia entre un data-hub y un data-lake, y proponer un caso que ejemplifique el uso de cada uno.

2. Un sitio de ventas de autos quiere agregar una nueva funcionalidad de forma que cuando un usuario cargue su auto para la venta, el sitio le recomiende un precio para el mismo. Para ello quiere generar un modelo con XGBoost que al momento de cargar un nuevo aviso prediga el precio de venta.

El set de datos sobre las publicaciones de autos posee estos atributos: fecha, marca, modelo, versión, año, cantidad de puertas, segmento/tamaño, equipamiento, kilometraje, estado, transmisión, color, tipo de combustible, motor, potencia, único dueño, provincia, dueño directo y precio.

Indicar el proceso de feature engineering que realizaría, indicando el detalle de los features que probaría para armar un modelo de predicción mediante XGBoost. Dar un ejemplo de un row del set de datos final.

3. ¿Cuáles de las siguientes técnicas sirve para reducir el overfitting en XGBoost? Justificar.

- Aumentar el número de estimadores;
- aumentar el valor del hiperparámetro gamma;
- reducir el valor del hiperparámetro min-child-weight;
- reducir el learning rate;
- reducir el valor del hiperparámetro max-depth.

4. Dados los siguientes puntos en una dimensión y sus labels para un problema de clasificación binaria:
(RED 0) (RED 0) (GREEN 1) (GREEN 0) (BLUE 0) (GREEN 0) (BLUE 0) (RED 1) (GREEN 1)

aplicar dos algoritmos diferentes de mean encoding indicando cómo quedaría cada registro.

5. Para el siguiente conjunto de registros, aplicar One-hot-encoding para la columna "hijos" y Target-encoding para la columna "¿Trabajo?". Sugerir además una forma de reducir posibles filtraciones de los labels para la columna codificada con Target-encoding.

Reg	Sexo	¿Hijos?	¿Trabajo?	¿Auto?	¿Seguro de vida?
1	M	S	S	S	sí
2	F	N	N	N	no
3	F	N	S	N	no
4	M	N	S	S	no
5	M	S	S	S	si
6	M	S	N	S	sí
7	F	S	S	N	no
8	M	S	N	N	sí
9	F	N	N	N	no
10	F	N	S	S	sí
11	F	S	N	N	no
12	F	S	S	S	sí

Redes neuronales

1. Explicar en qué consiste Transfer Learning.

2. Explicar por qué motivo es conveniente usar RELU como función de activación en lugar de la función sigmoidea para las capas intermedias de una red neuronal.

3. En este ejercicio se plantea la aplicación de Redes Convoluciones para textos. Se tienen oraciones de un texto y se quiere clasificarlas en 10 categorías diferentes. Se quiere usar cada palabra del texto como una unidad indivisible.

a) Explicar cómo se aplicaría una capa de convolución sobre el texto, cuáles serían los hiperparámetros a definir, y cuáles son los parámetros que debe encontrar la red neuronal.
b) Diseñar y explicar una arquitectura para resolver el problema propuesto usando una única capa de convolución, una de pooling, una capa fully-connected y luego la capa final para clasificar. Justificar.

4. Se parte de imágenes RGB de 20x20 píxeles que se quieren clasificar entre 3 clases posibles con una red neuronal convolucional. Se aplica una capa convolucional del tipo "same" con 5 filtros de 3x3, luego se aplica una capa max pooling de 4x4 con stride de 4, a continuación una capa fully-connected y finalmente una capa softmax.

a. Diagramar el modelo de red. ¿Cuántas neuronas tienen las dos últimas capas del modelo?
b. ¿Cuántos parámetros se deben entrenar en total?

5. Se quiere aprender un clasificador tipo Perceptron para 6 puntos en el plano (A, B, C, D, E y F) y sus correspondientes clases (Ca, Cb, Cc, Cd, Ce y Cf) comenzando con el vector de pesos Wo = [1 1]. Se sabe que el algoritmo converge luego de 5 iteraciones:

- en la primera iteración clasifica mal al punto "A";
- en la segunda iteración clasifica mal al punto "B";
- en la tercera iteración clasifica mal al punto "A";
- en la cuarta iteración clasifica mal al punto "F";
- en la quinta iteración todos los puntos quedan bien clasificados.

Expresar el resultado final de W en función de los puntos A, B, C, D, E, F (no olvidar el término independiente).

6. Dados los puntos:
X1 = [-2 2] (Clase +1) X2 = [0 -2] (Clase +1) X3 = [4 4] (Clase -1) X4 = [4 5] (Clase - 1)

a. Entrenar un Perceptron utilizando el vector de pesos inicial Wo = [1 1 1] y learning rate 0,5 hasta que todos los puntos queden correctamente clasificados.
b. Graficar los puntos y la separación de clases encontrada en base al vector de pesos resultante.
c. Indicar si existe alguna mejor separación entre clases y, en caso afirmativo, indicar cómo poder encontrarla.

7. Dados los siguientes puntos en una dimensión (con sus correspondientes clases dadas por el signo):
[1 +1] [5 -1] [0 +1] [6 -1] [4 +1] [7 -1]

se aplica Perceptron para clasificar los puntos a partir del vector inicial de pesos w = [1 1]. Se observa que Perceptron converge luego de 33 pasos, y que en el proceso se clasificó 6 veces mal al primer punto, 13 veces mal al segundo puntos y 14 veces mal al quinto. En base a esto, se pide:

a. indicar el valor final del vector de pesos;
b. graficar el vector final de pesos y los puntos.
c. indicar qué se podría hacer para lograr la convergencia de forma más rápida.

Item adicional: para la propuesta dada en el item c, indicar en cuántos pasos convergería Perceptron y cuál sería el valor final del vector de pesos.

Árboles de decisión

1. Dados los siguientes datos, generar un árbol de decisión utilizando ID3 que permita determinar cuándo una persona es candidata a contratar un seguro de vida.

Reg	Sexo	¿Hijos?	¿Trabajo?	¿Auto?	¿Seguro de vida?
1	M	S	S	S	sí
2	F	N	N	N	no
3	F	N	S	N	no
4	M	N	S	S	no
5	M	S	S	S	si
6	M	S	N	S	sí
7	F	S	S	N	no
8	M	S	N	N	sí
9	F	N	N	N	no
10	F	N	S	S	sí
11	F	S	N	N	no
12	F	S	S	S	sí

2. Se quiere construir un árbol de decisión para clasificar si un conjunto de platos es adecuado para carnívoros (clase C) o veganos (clase V). El algoritmo CART debe analizar si hacer o no un split por la columna "ingrediente 2 en gramos" (columna numérica). Dados los valores de la columna y la clase a predecir, indicar cuál sería el valor de la ganancia de información para dicha columna:
(23 V) (80 C) (85 C) (65 V) (100 V) (11 C)

3. Dada la información sobre los siguientes días:

Día	Clima	Temperatura	Humedad	Viento	¿Jugar al tenis?
D1	Soleado	Caluroso	Alta	Débil	No
D2	Soleado	Caluroso	Alta	Fuerte	No
D3	Nublado	Caluroso	Alta	Débil	Sí
D4	Lluvia	Templado	Alta	Débil	Sí
D5	Lluvia	Frío	Normal	Débil	Sí
D6	Lluvia	Frío	Normal	Fuerte	No
D7	Nublado	Frío	Normal	Fuerte	Sí
D8	Soleado	Templado	Alta	Débil	No
D9	Soleado	Frío	Normal	Débil	Sí
D10	Lluvia	Templado	Normal	Débil	Sí
D11	Soleado	Templado	Normal	Fuerte	Sí
D12	Nublado	Templado	Alta	Fuerte	Sí
D13	Nublado	Caluroso	Normal	Débil	Sí
D14	Lluvia	Templado	Alta	Fuerte	No

utilizar ID3 para determinar si las condiciones serán o no apropiadas para jugar al tenis.

4. Dada la siguiente información sobre préstamos, donde la clase indica si se pagó o no el préstamo, construir un árbol de decisión con ID3.

Trabajo	Estudios	Rango de edad	Pagó
Relación de dependencia	Universitarios	18 a 25	Sí
Monotributo	Universitarios	18 a 25	Sí
Sin trabajo	Universitarios	26 a 35	No
Relación de dependencia	Secundarios	18 a 25	No
Relación de dependencia	Secundarios	36 a 45	Sí
Monotributo	Universitarios	36 a 45	Sí
Sin trabajo	Secundarios	18 a 25	No
Monotributo	Primarios	26 a 35	No

5. Dados los siguientes datos:

Persona	Estudios	¿Trabaja?	Estado civil	¿Préstamo?
1	Universitarios	Sí	Casado	Sí
2	Primarios	Sí	Soltero	No
3	Universitarios	Sí	Casado	Sí
4	Universitarios	No	Casado	No
5	Primarios	Sí	Divorciado	Sí
6	Secundarios	No	Divorciado	No
7	Universitarios	Sí	Divorciado	Sí
8	Universitarios	Sí	Soltero	No
9	Secundarios	No	Divorciado	No
10	Secundarios	Sí	Soltero	Sí
11	Universitarios	Sí	Soltero	No
12	Primarios	Sí	Casado	Sí

a. Armar un árbol de decisión seleccionando en cada split el atributo que mayor ganancia de información da.
b. Explicar cómo hacer para evitar el overfitting y aplicarlo al árbol del punto anterior.

6. Se tienen los siguientes datos acerca de partidos de fútbol americano disputados por un cierto equipo ("¿Rwin?" indica si el rival tiene récord ganador).

Juega de	Mes	¿Rwin?	Resultado
Local	Octubre	Sí	Win
Visitante	Octubre	No	Win
Local	Noviembre	No	Win
Local	Noviembre	Sí	Lose
Visitante	Diciembre	No	Lose
Visitante	Diciembre	Sí	Lose
Local	Diciembre	No	Win

Construir un árbol de decisión para predecir el resultado en función de los otros tres atributos. Sugerir una mejora al modelo construído.

7. Dados los siguientes puntos en una dimensión y sus labels para un problema de regresión (con el formato "[coordenada X label]"):
[1 6] [17 11] [18 14] [26 23] [303 108] [411 109] [511 211]

se quiere usar Gradient Boosting para aproximar los valores de forma tal de minimizar el error cuadrático medio (MSE por sus siglas en inglés).

a. Construir un árbol de decisión de nivel 1 y dibujarlo.
b. Construir un segundo árbol basado en Gradient Boosting para indicar la predicción final que se realizaría para cada punto y el error de entrenamiento en base a estas predicciones. Indicar si se cae en un caso de overfitting o underfitting.

Clustering

1. Demostrar que K-Means converge siempre para la distancia Euclídea. Se recomienda seguir los siguientes pasos.

a. Demostrar que el punto que minimiza la función de costo para un cierto cluster es el promedio de todos los puntos del cluster.
b. Demostrar que en cada paso la función de costo decrece.
c. Terminar la demostración con una aclaración adicional.

2. Un sistema recibe un flujo de datos numéricos que se necesitan asignar a un cluster utilizando K-Means online, con k = 3. Considerando que la inicialización se hace con los primeros 6 datos del stream, y luego a continuación vienen 4 datos más, realizar el seguimiento indicando cómo se comporta el algoritmo para generar los clusters, y cuál es el resultado final luego de procesar los primeros 10 elementos: 2, 6, 7, 18, 1, 9, 0, 3, 4 y 23.

3. Se tiene la siguiente tabla en donde cada fila es un usuario y cada columna es una categoría de películas. la tabla indica con una escala de 0 a 1 si al usuario le gusta o no la categoría.

Usuario	Western	Comedia	Acción	Ciencia Ficción
Alice	0,2	0,6	0,5	0,1
Bob	0,7	0,4	0,7	0,4
Charly	0,1	0,6	0,7	0,9
Diane	0,5	0,1	0,9	0,8
Ethan	0,2	0,4	0,5	0,6

Se quieren encontrar dos clusters. Explicar de qué forma se aplicaría clustering espectral a partir de los datos proporcionados.

4. Se tienen los centroides [0 0] y [100 40]. Dados los siguientes puntos, indicar para cuál de ellos cambiaría el centroide al cual queda asignado según si se usa distancia Manhattan o Euclídea:
[53 15] [51 15] [50 18] [52 13]

5. Dados los siguientes puntos en dos dimensiones:
[5 1] [5 3] [4 4] [9 4] [10 3] [11 6]

aplicar K-Means comenzando con los centroides [8 1] y [7 5].

6. Dados los siguientes puntos en dos dimensiones:
[1 1] [2 1] [4 3] [5 4]

aplicar K-Means tomando como centroides [1 1] y [2 1], realizando un desarrollo descriptivo paso a paso del algoritmo, incluyendo representación gráfica e indicando:

a. posición aproximada final de los centroides al converger;
b. clusters finales obtenidos.

7. Dados los siguientes puntos en 1 dimensión: 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, se quiere aplicar K-Means utilizando la distancia Euclídea para encontrar dos clusters. Para iniciar el algoritmo, se eligen dos puntos al azar de la lista.

a. Indicar cuál de todas es la inicialización que minimiza la función objetivo de K-Means.
b. Para todos los resultados posibles, indicar cuántos puntos cambian de cluster en la segunda iteración de K-Means.

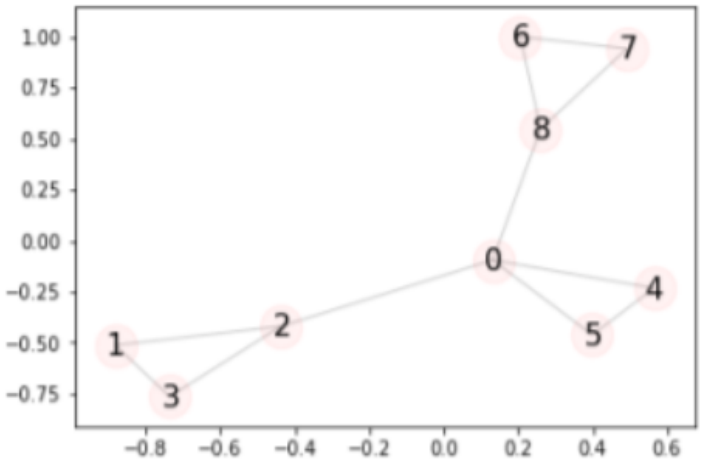
8. Considerar los siguientes puntos en una dimensión y la clase a la que pertenece cada punto (en formato "(número clase)"):
(5 b) (10 b) (2 b) (6 a) (7 b) (8 b) (0 a) (13 b) (16 b) (1 a) (3 a) (17 b)

Se quiere usar KNN para clasificar los puntos con la distancia Manhattan. Con el objetivo de encontrar el valor óptimo de K (cantidad de vecinos), se utiliza Cross-Validation con folds de 4 registros. Considerar los folds en el orden en que fueron presentados los puntos (Ejemplo: primer fold tiene a (5 b) (10 b) (2 b) (6 a)).

Determinar el valor óptimo de K realizando Grid-Search para tres valores posibles: 1, 3 y 5. Usar como métrica la precisión (accuracy).

9. Se desea aplicar clustering espectral sobre el siguiente grafo. Los 4 primeros autovalores de la matriz laplaciana (ordenados de menor a mayor) son: 1 = 0; 2 = 0,27; 3 = 0,55; 4 = 3; y sus autovectores son:
v1 = [-0,33 -0,33 -0,33 -0,33 -0,33 -0,33 -0,33 -0,33 -0,33]
v2 = [0 -0,44 -0,33 -0,44 0 0 0,44 0,44 0,33]
v3 = [0,25 -0,28 -0,12 -0,28 0,55 0,55 -0,28 -0,28 -0,12]
v4 = [-0,07 0,72 -0,07 -0,66 0,15 -0,09 0,03 0,03 -0,07]

Determinar cuántos y cuáles autovectores se deben utilizar para detectar las 3 comunidades del siguiente grafo (justificar).



10. Dados los siguientes puntos:
[1 2] [1 3] [1 5] [2 3] [2 8] [3 3] [4 1] [4 9] [7 8] [9 9] [12 2] [13 4]

agruparlos utilizando Clustering Jerárquico, indicando la cantidad de clusters resultantes, y el criterio utilizado para definir dicho número. Representar el dendrograma mostrando cómo se agrupan los distintos elementos en los clusters obtenidos.

11. En cada uno de los siguientes casos, sugiera qué algoritmo de clustering usaría.

a. 1 000 puntos, 3 clusters de diferentes densidades y formas complejas.
b. 1 340 millones de puntos, 168 clusters de diferentes densidades y forma regular.
c. 20 000 millones de puntos, cantidad indefinida de clusters, densidad variable y formas complejas.

12. A partir de los siguientes puntos en el plano, mostrar el funcionamiento del algoritmo K-Means, con k = 3, hasta que no haya cambios. En cada iteración grafique los puntos, centroides y clusters generados en el plano. Usar los puntos en negrita como centroides iniciales.
A1: [2 10] A2: [2 5] A3: [8 4] **A4: [5 8]** A5: [7 5] A6: [6 4] **A7: [1 2]** A8: [4 9]

13. A partir de los siguientes puntos y 3 centroides, intentar predecir a qué va a converger K-Means.

Streaming

1. "DealExtreme" es una empresa que vende una enorme cantidad de productos importados de bajo costo de China. Se analiza un stream en donde se recibe el ID de cada producto vendido. Cada vez que se venden 100 unidades de un determinado producto, el departamento de marketing publica un pequeño aviso sobre el mismo. Esto permite aumentar aún más las ventas de los productos populares.

A la gente de marketing se le ocurrió también que sería una buena idea publicitar los productos que no han llegado nunca a las 100 ventas. Para ello, crearon una campaña en la cual crean un aviso especial para los productos vendidos los días viernes que nunca tuvieron publicado un aviso.

Diseñar una solución que, a través del procesamiento del stream, permita a la gente de marketing determinar si debe publicar un aviso o no en base al ID del producto.

Observación: notar que no interesa la cantidad total de unidades vendidas para cada producto , sino simplemente la detección de cuándo la cantidad de ventas llega a 100.

2. Dado un stream compuesto por los siguientes números: 3, 1, 4, 1, 5, 9, 2, 6, 5; se quiere aplicar el algoritmo Flajolet-Martin para calcular el momento de orden 0 del stream usando una función de hashing de la familia: $h(x) = ax + b \bmod 32$. El resultado debe tomarse como un número de 5 bits contando la cantidad de ceros a derecha del mismo. No todos los valores de a y b son adecuados, por lo que se debe explicar qué valores son los más adecuados y por qué. A modo de ejemplo, analizar las funciones resultantes de usar: a = 2, b = 1; a = 3, b = 7; y a= 4, b = 0.

3. Se desea utilizar Flajolet-Martin para estimar la cantidad de elementos distintos en un stream. Suponer que de los 10 elementos posibles (números naturales del 1 al 10), solo 4 de ellos aparecen efectivamente en el stream. Para realizar la estimación, se puede hashear cada elemento en binarios de 4 bits con $h(x) = 3x + 7 \bmod 11$. Por ejemplo, para x = :, $h(x) = 31 \bmod 11 = 9$, por lo que en binario: h(x) = 1001. A los efectos de la implementación, contar los ceros del final del código (por ejemplo, para "0100", son 2 ceros). Encontrar todos los subconjuntos de 4 elementos que hacen que la estimación sea exacta.

4. Aplicando Reservoir Sampling con k = 4, se observa el siguiente stream: 5, 6, 7, 8. A continuación, se observa el elemento 3. Justificar cuál es la probabilidad de que 6 continúe en la muestra de Reservoir Sampling.

5. a. Construir un filtro de bloom de 16 bits (m = 16) que contenga los caracteres "C" y "D", para el universo de 5 (n = 5) caracteres: "A", "B", "C", "D" y "E" considerando las siguientes funciones de hashing "h1" y "h2". Tener en cuenta que la probabilidad de falsos positivos para la construcción es de 0,2.

	A	**B**	**C**	**D**	**E**
h1	13	8	4	4	3
h2	15	5	9	15	7

b. Sabiendo que el valor del "m" que se utilizó es óptimo, indicar si la cantidad de funciones de hashing utilizadas es óptima (justificar).

6. Dados los siguientes streams:
E A B D C A E C B D
E A E B C E E D B E

estimar el número sorpresa utilizando AMS con 5 estimadores, y explicar qué conclusiones se pueden obtener de la comparación de los resultados para ambos streams.

7. Se usa el algoritmo Count-Min con 3 filtros de 8 posiciones cada uno. El estado de los filtros está dado por:
[0 2 4 2 0 0 3 5]
[1 0 2 6 1 2 0 4]
[0 0 3 3 3 2 4 1]

Indicar cuáles de las siguientes afirmaciones son verdaderas (justificando adecuadamente).

a. La cantidad total de elementos del stream es 16.
b. No puede tratarse de un stream con 16 elementos diferentes.
c. Puede existir un elemento con frecuencia 5.

8. ¿Cuál es la cantidad máxima que puede estimar el algoritmo Count-Min dados los siguientes filtros? ¿Qué resultados debería dar la función de hashing para lograr dicha estimación?
[2 6 0 0 0 3 0 1]
[0 2 1 1 2 2 4 3]
[3 2 0 0 2 4 3 2]

9. Se usan 3 funciones de hashing 0..4 para el algoritmo Count-Min. Para las siguientes palabras, se da el resultado de las 3 funciones de hashing:
"casa" -> [2 1 2] "canasta" -> [0 3 3] "kilo" -> [4 1 0] "alfa" -> [3 3 0]

si los filtros son los siguientes:
[0 2 3 1 3]
[1 2 2 4 0]
[3 0 1 1 1]

¿Cuál de las cuatro palabras es estimada como la más frecuente?

10. Dado el siguiente stream: 1, 3, 2, 3, 3, 5, 2, 1; indicar una función de hashing de la forma h(x) = ax + b mod 32 tal que el algoritmo de Flajolet-Martin se aproxime de la mejor forma posible al cálculo del momento de orden 0 del stream. Usar 5 bits y tomar los bits a derecha para el algoritmo.

11. Dado el siguiente stream: 1, 3, 4, 3, 4, 2, 2, 4, 2, 2; calcular el momento de orden 2 del stream y luego realizar una estimación usando AMS a partir de los valores del stream 2, 2, 4, 2, 2. Usar el promedio entre estimadores para la estimación final.

Recomendaciones

1. Se tiene la siguiente matriz que relaciona a usuarios (A, B, C y D) y productos comprados (P1, P2, etc.)

	P1	P2	P3	P4	P5	P6	P7	P8
A		1			1	1		
B			1			1	1	1
C	1				1	1		
D		1		1			1	

Usando la semejanza de Jaccard y Collaborative Filtering user-user con k = 2, indicar cuáles serían los 3 productos que le recomendaría al usuario A y en qué orden. En caso de que haya empates, usar Collaborative Filtering item-item para desempatar.

2. Dada la siguiente matriz de utilidad entre películas (de 1 a 6) y usuarios (de 1 a 12).

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

a. Usando Collaborative Filtering item-item calcular la calificación del usuario 2 para la película 1, considerando 2 items similares.
b. Usando Collaborative Filtering en base a desviaciones calcular la calificación del usuario 2 para la película 1, considerando 2 items similares.
c. ¿Qué conclusiones se pueden sacar de la comparación de ambos resultados?

3. Dada la siguiente tabla de calificaciones de libros dada por usuarios:

	B1	B2	B3	B4	B5
Alice	5			4	1
Bobe		3	2	4	4
Charly	1	5	4		3
Diana	5	3	1	2	

Usar Collaborative Filtering user-user tomando los 2 vecinos más cercanos para estimar la calificación de los libros B2 y B3 para Alice.

4. Spotify registra con 1 o 0 si al usuario le ha gustado o no una canción. Se tiene la siguiente matriz en donde se conocen las canciones (C1, C2, etc.) que le gustan a un conjunto de usuarios (A, B, C, D, E). Se pide estimar si al usuario A le van a gustar o no las canciones 5 y 6, usando Collaborative Filtering user-user tomando los 2 vecinos más cercanos.

	C1	C2	C3	C4	C5	C6
A	1	0	0	1		
B	0	1	1	0	1	0
C	1	1	0	0	0	1
D	1	0	1	1	1	0
E	1	0	0	1	0	1

5. Sea la siguiente matriz de utilidad entre usuarios (1, 2, 3, 4, 5) y libros (A, B, C, D, E, F).

	A	B	C	D	E	F
1	3		5		3	
2			4		2	4
3	5		5	5		
4		2				5
5	2	2		2		5

Realizando Collaborative Filtering item-item, determinar qué libro se le debe recomendar en primer lugar al usuario2. Tomar para la recomendación los 2 libros más semejantes.

6. Dada la siguiente matriz de utilidad representando la calificación de usuarios y películas:

	U1	U2	U3	U4
M1	5	4	1	4
M2	4	5	2	4
M3	1	4	4	
M4	2	4	4	1

a. Normalizar la matriz restando a cada columna su promedio. Estimar la calificación faltante usando Collaborative Filtering user-user.
b. ¿Qué modalidad de Collaborative Filtering genera resultados más previsibles o conservadores: user-user o item-item? Justificar.

7. Estimar la calificación de Avatar para el usuario 4 (U4) usando Collaborative Filtering item-item a partir de la siguiente matriz de calificaciones de películas.

	U1	U2	U3	U4
Oblivion	5	2		4
Matrix	5		4	4
Avatar	2	3	4	

8. Se sabe que el promedio global de un conjunto de calificaciones de películas es de 2,28. Sabiendo que el usuario "Ariel" ha realizado las siguientes calificaciones: 1, 4, 2, 1, 1; y que la película "Interstellar" tiene las siguientes calificaciones: 5, 4, 1, 3; ¿cuál sería la estimación de la calificación de "Ariel" para "Interstellar"?

9. Se tienen las calificaciones de Bob para cinco películas: 2, 1, 4, 1, 2; las de Claire para esas mismas películas: 4, 2, 5, 3, 5; y las de "Alice" para las primeras cuatro películas: 3, 1, 5, 2. Usando el coeficiente de correlación de Pearson, indicar cuál es la semejanza de A con los usuarios "Bob" y "Claire", y luego estimar la calificación de Alice para la quinta película.

10. Se conocen las calificaciones que tres usuarios han hecho sobre 5 películas.

	P1	P2	P3	P4	P5
U1	4	2	5	2	2
U2	2	3	4	1	5
U3	5	1	4	2	

Estimar la calificación del usuario 3 para la película 5 usando semejanza user-user y desviaciones sobre el promedio global.

Aplicaciones de grafos

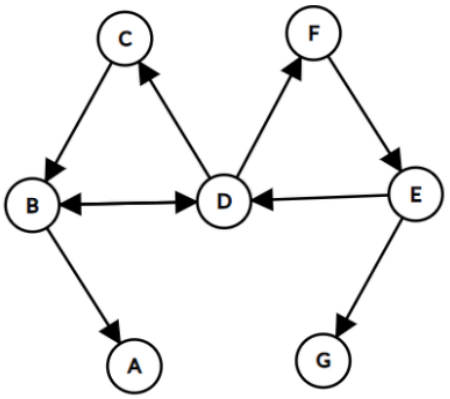
1. Explicar el fenómeno de "mundo pequeño", indicando en qué casos se da y cuáles son las características de un grafo en el que se presenta.

2. Se tiene la siguiente información parcial sobre un grafo dirigido al cual se le ha aplicado PageRank hasta la convergencia:

- el PageRank de A es 0,15;
- A tiene 2 links: uno hacia C y otro hacia B;
- C recibe solo dos links, uno de los cuales viene de B;
- el PageRank de D es 0,1;
- B recibe un solo link;
- D solo recibe un link que viene de C;
- desde C solo sale un link.

¿Cuántos links salen desde B? Justificar detalladamente.

3. Dado el siguiente grafo



a. Hallar la matriz de transición
b. Analizar qué problemas presenta la misma para poder aplicar el modelo de Random Walker, y cómo es que los mismos se detectan.

4. Una empresa de alquiler de monopatines eléctricos tiene 4 estaciones (A, B, C, D) en donde los usuarios pueden alquilar monopatines. En experimentos iniciales, la empresa pudo determinar con qué frecuencia los usuarios van desde una estación a otra. Dicha información se presenta en la siguiente tabla.

	A	B	C	D
A	5	10	2	0
B	8	0	5	6
C	3	3	5	15
D	4	2	23	4

A partir de la información anterior, aplicar PageRank para determinar la cantidad de monopatines que se deben asignar a cada estación al comienzo de cada día, considerando que el total de monopatines de la empresa es 100.

5. Sea el grafo dirigido G cuya lista de aristas es la siguiente: (A B), (B C), (B D), (D E), (E C), (C A), (C D). Se sabe que los valores de PageRank actuales de los nodos son:
A = 0,31 B = 0,30 C = 0,08 D = 0,12 E = 0,19

Calcular el valor de PageRank de cada nodo para la siguiente iteración, usando B = 0,8.

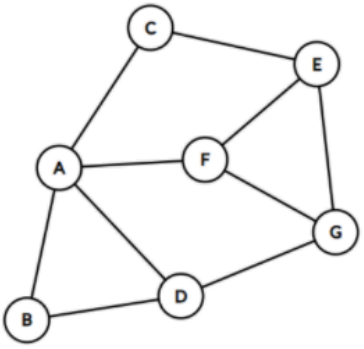
6. a. ¿Es verdad que si al agregar un link en una red social cambia el coeficiente de clustering de 2 nodos, entonces se puede afirmar que dichos nodos tienen al menos 2 amigos en común?
b. Dada una comunidad de N nodos en donde todos son amigos de todos, si una de esas amistades se rompe: ¿cuántos coeficientes de clustering se modifican? ¿Cuál es el nuevo coeficiente de clustering promedio en función de N?

7. Considerar una red social en la cual cada usuario puede tener 0, 1 o 2 amigos, y en donde prácticamente todos los usuarios tienen 2 amigos. Se quiere comparar la estructura de esta red social con respecto a una en donde no existan estas limitaciones. Analizar los siguientes elementos:

- distribución del grado;
- betweenness promedio;
- diámetro
- cantidad de triángulos;
- coeficiente de clustering.

8. Se tiene un grafo dirigido con 5 nodos y las siguientes 7 aristas: (A B), (A D), (D B), (B C), (C B), (C E) , (E A). Calcular cuál sería el vector de teletransportación necesario para que el PageRank de todos los nodos dé igual, tomando beta en 0,5.

9. Dado el siguiente grafo



a. mostrar la distribución de grados;
b. indicar cuál es el diámetro de la red;
c. calcular el coeficiente de clustering promedio;
d. comparar dichos valores con los valores característicos de una red social, analizando semejanzas y diferencias.

10. Dada la siguiente matriz de links M:

	A	B	C	D	E	F
A						
B	1/3		1		1	
C	1/3					
D		½				
E		½		½		
F	1/3			½		

a. dibujar el grafo dirigido asociado a la misma, indicando el peso de las aristas;
b. indicar todos los problemas que hacen que no se pueda aplicar PageRank sin teletransportación sobre el grafo anterior;
c. indicar paso a paso cómo construir una única matriz de PageRank que incluya teletransportación (tomar beta en 0,85).

11. Realizar TopicRank del topic 1 asociado al grafo dado por lo siguiente:

- A pertenece al topic 1 y enlaza con B, C y E;
- B pertenece al topic 1 y enlaza con A y E;
- C pertenece al topic 2 y enlaza con A, C y D;
- D pertenece al topic 1 y enlaza con E;
- E pertenece al topic 2 y no tiene más links.

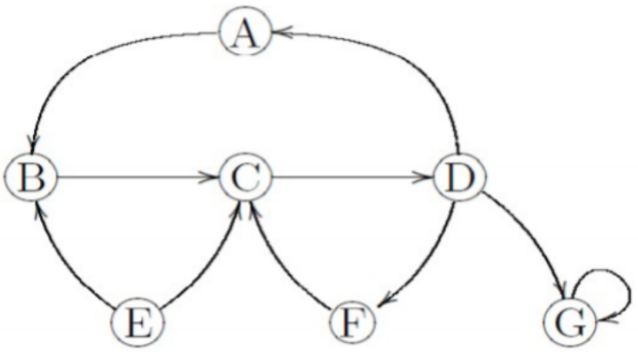
Realizar 3 iteraciones con beta en 0,3 y rankear los resultados. Determinar si puede considerar que los resultados son adecuados y por qué.

12. Sea un grafo dado por las siguientes aristas: (A C), (C D), (D E), (F A), (E A), (A B), (F B). Justificando adecuadamente:

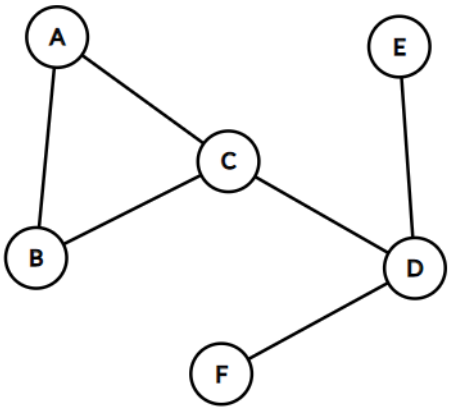
a. plantear el sistema MV inicial y calcular el PR para la primera iteración (con 6 puntos);
b. indicar sin hacer cuentas cuál va a ser el nodo con mayor y menor PR final;
c. finalmente indicar cuál sería la arista a agregar que más posiciones mejoraría al nodo F en su PR final.

13. Se tiene un grafo con m nodos y k aristas. Se quiere conocer si el grafo responde a las características de una red social. Indicar el tipo de análisis a realizar para determinar esto, qué variables analizaría y qué valores debería esperar si el grafo respondiera al comportamiento de una red social.

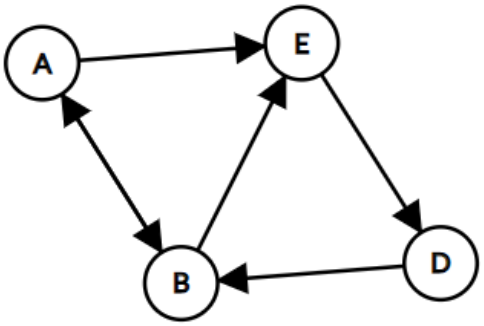
14. Aplicar PageRank para el siguiente grafo hasta la convergencia, indicando el ranking final de cada nodo (tomar beta en 0,8).



15. Para la siguiente grafo, calcular el coeficiente de clustering y el betweenness para el nodo C.



16. Realizar una iteración de PageRank para el siguiente grafo (tomar beta en 0,85).



17. Zolio tiene 4 amigos: Armando, Bárbara, Claudio y Diana. Bárbara y Claudio son amigos, pero los demás no se conocen. ¿Cuál es el coeficiente de clustering de Zolio?

18. Sea el grafo dirigido dado por las siguientes aristas: (B H), (H C), (C H), (C B) y (B C). Indicar el PageRank de B, C y H luego de 3 iteraciones (considerar beta en 0,85).

19. Sean las siguientes relaciones de amistad en una red social no dirigida: (A B), (A C), (B D), (C D), (D E), (D E), (D F), (E C), (E F).

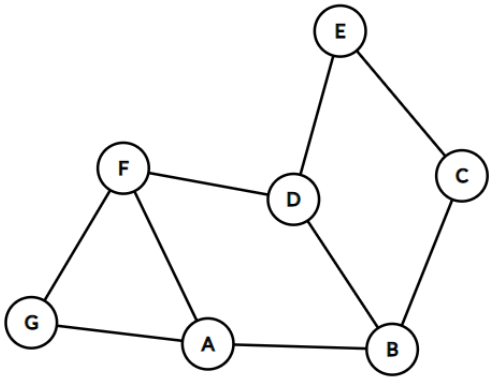
a. Calcular el coeficiente de clustering promedio.
b. Calcular el betweenness de cada nodo.

20. Sea el grafo dirigido dado por las siguientes aristas: (A B), (B D), (D A), (C A), (D E). Calcular el PageRank de cada nodo luego de 3 iteraciones a partir del vector [0,2 0,2 0,2 0,2 0,2] (considerar beta en 0,85).

21. Se usa SimRank para recomendar usuarios a seguir en Twitter. Sin embargo, el algoritmo es demasiado "conservador": los usuarios piensan que las recomendaciones realizadas son obvias, y por lo tanto no permite descubrir usuarios interesantes a seguir. Para mejorar el algoritmo, se plantea la posibilidad de aumentar o reducir el parámetro beta. ¿Cuál de las dos opciones haría que el algoritmo genere recomendaciones un poco más interesantes?

22. Dar un ejemplo de una red en donde el nodo con mayor centralidad por PageRank tenga coeficiente de clustering 0. Analizar si este tipo de red es probable dentro de una red social real.

23. Dado el siguiente grafo:



a. de acuerdo al modelo de Preferential Attachment, ¿cuál o cuáles son las aristas con mayor probabilidad de agregarse a la red?
b. ¿Cuál es el coeficiente de clustering promedio de la red?
c. ¿Cuáles son los nodos con mayor betweenness? Indicarlo sin hacer cuentas.