Examen Primer Interciclo Estadística para CC (Parte teórica)

Nombre:

1. Si (1.5Pts):

	X	**Y**
u1	5	1
u2	5	2
u3	1	5

Calcular las correlaciones: corr(X,Y), corr(u1,u2), corr(u1,u3).

2. Si: (2Pts)

$$\mathbf{P_{X,Y}(x,y)=}\begin{cases}\frac{1}{2} & x=3,\ y=4\\ \frac{1}{3} & x=3,\ y=6\\ \frac{1}{6} & x=5,\ y=6\\ 0 & otro\ caso\end{cases}$$

Calcule la cov(X,Y) y la correlación corr(X,Y)

3. Menciones dos objetivos del análisis multivariante (1Pt)
   a. _____
   b. _____
4. Indique tres técnicas de análisis multivariante (1.5Pts)
   a. _____
   b. _____
   c. _____

Proyecto: Sistema de Recomendación de libros (Examen Parte Práctica)

En grupos de 4 personas. Enviar al correo: rhurtadoo@ups.edu.ec los integrantes del grupo, hasta mañana jueves 23 de mayo.

Dataset:

BookCrossing Dataset: 1,149,780 integer ratings (from 0-10) of 271,379 books from 278,858 users.

El conjunto de datos esta dividido en tres partes:

- Información de los usuarios
- Información de los libros
- Información de los votos que los usuarios han dado a los libros (primera columna corresponde al código del usuario. La segunda columna corresponde al código del libro. La tercera columna corresponde al voto del usuario sobre un libro)

Enlace del dataset: http://www2.informatik.uni-freiburg.de/~cziegler/BX/

El proyecto consta de cuatro fases:

Para el jueves 30 de mayo se revisarán las dos primeras fases. Ante cualquier inquietud escribirme al correo para poder orientarles o coordinamos una hora para que se acerquen a mi oficina en el GIIATA.

Primera fase: Análisis de datos (Carga, Transformación y Análisis)

1. Cargar el dataset (Ojo: hay que crear un método para poder cargar el dataset a la estructura tradicional)
   Ejemplo:
   Archivo de votos

1	1	5
1	3	4
2	2	2
2	3	3

   Matriz de Votos (Estructura tradicional): se llena con cero la ausencia de voto.

	Libro 1	Libro 2	Libro 3
Usuario 1	5	0	4
Usuario 2	0	2	3

2. Realizar un histograma de votos. Ejemplo: cantidad de votos con 1, cantidad de votos con 2, etc.
3. Obtener las medidas descriptivas de cada variable (libro) y obtener el libro mejor puntuado por todos los usuarios (es decir, el libro con la media más alta).
4. Transformar los datos
5. Nuevamente obtener las medidas descriptivas de cada variable (libro) y obtener el libro mejor puntuado por todos los usuarios (es decir, la media más alta).

6. Aplicar PCA y determinar la cantidad óptima de componentes principales. Para ello tomar la cantidad de componentes principales en que la sumatoria de las varianzas de los componentes superen el 90%. Mostrar la gráfica de varianzas frente al número de componentes principales (Eje X identificador del componente principal y eje Y la varianza).
7. Desde la matriz de componentes principales, obtener la correlación entre los usuarios, de tal manera que se obtenga la tabla de correlaciones.

Segunda fase: Determinar los k vecinos de un usuario (Desde PHP a Python)

8. Crear una interfaz en PHP que permita que un usuario pueda solicitar los K usuarios con características similares, para ello enviará su código de usuario y el parámetro K al módulo de Python.
9. El módulo de Python determinará los k vecinos del usuario (es decir, los k usuarios con más alta correlación con respecto al usuario conectado), para ello utilizará la tabla de correlaciones de la primera fase. La salida será un arreglo con los códigos de los usuarios vecinos.
10. En la interfaz de PHP se presentará la lista de los usuarios vecinos. En la lista se podrá seleccionar un usuario y ver su información personal.
11. Elaborar un informe con las gráficas y resultados solicitados.

Tercera fase: Generar recomendaciones

12. Aplicar una función para obtener las predicciones y posteriormente las recomendaciones

Cuarta fase: Evaluar el método

13. Medir el error de las predicciones
14. Intentar mejorar el error del método anterior aplicando alguna otra técnica multivariante (LDA, MF, Clustering, etc) o alguna medida de similaridad.
15. Elaborar un informe con las gráficas y resultados solicitados.

Rúbrica para el Examen

ID Actividad	Puntaje
1	1
2	0.5
3	0.5
4	0.5
5	0.5
6	5
7	1
8	1
9	1
10	2
11	1
Total:	14Pts