**Tarea TIA-04**

**Fecha entrega: 06-10-2023**

**Peso: 20%**

**Tarea. Reducción de dimensiones utilizando las técnicas de Componentes**

**Jonathan Álvarez Arismendi**

**Héctor Enrique Diaz García**

**Diego Alejandro Rios Vasquez**

**Instrucciones**

* Descargue todos los archivos al PC local o trabaje en Colab en la Nube según su preferencia.
* Renombre el formato Word cambiando la “X” por la letra del equipo de los documentos
* Además del documento Word, debe elaborar 3 Cuadernos Jupyter. Uno para cada ejercicio: 1, 2 y 3.
* Suba los documentos a una carpeta de Google Drive y comparta el link en el Classroom de la asignatura.
* Leer atentamente el enunciado y cada una de las preguntas.

**Conceptos teóricos:**

PCA permite encontrar un número de factores subyacentes (z<p) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de esta z nueva variable recibe el nombre de componente principal.

El Análisis de Componentes Principales (PCA, en su versión inglesa) es un método estadístico cuya utilidad radica en la reducción de la dimensionalidad de la base de datos (BDD) con la que estamos trabajando.

El PCA es un método de análisis multivariante que transforma un conjunto de datos con múltiples variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Esta técnica permite transformar y reducir la dimensionalidad de los datos mientras conserva la mayor cantidad posible de información relevante.

**En cada ejercicio, realice las siguientes actividades:**

1.- Crear un Jupyter Notebook que funcione en Google Colab con el nombre “***cd-2024-1-et0203-tia4-equipo\_x-ejercicio\_n***”. Cambie la letra X por el equipo. Cambie la letra “n” por el número de ejercicio. Al interno, coloquen los nombres de los integrantes.

2.- El código tiene que estar organizado (separado) en secciones diferentes. Tal como se describen a continuación:

1. Enunciado del problema
2. Carga de las librerías necesarias
3. Cargar los datos.
4. Mostrar:
   1. Total, de filas y columnas
   2. Los primeros 20 registros.
   3. Tipos de datos
   4. Operación de borrar columnas (si así lo requiere explicando el porqué)
   5. Categorizar columnas si es necesario (transformar datos nominales en números). Nota: Esto se debe hacer antes de normalizar los datos. Por ejemplo: “One-Hot Encoding”
5. Normalizar/Estandarizar los datos.
6. Obtener los autovectores y autovalores.
7. Obtener las siguientes matrices (Explicar la información contenida brevemente en cada una):
   1. Matriz de covarianza.
   2. Matriz de correlación
   3. Matriz de autovectores
   4. Matriz de autovalores
8. Seleccionar los autovectores correspondientes a las componentes principales
9. Proyectar el dataset original sobre el nuevo espacio de dimensión < 4
10. Elaborar los siguientes gráficos:
    1. Gráfico de Codo
    2. Gráfico de varianza explicada
    3. Gráficos de Barras (variables)
    4. Gráfico de Caja y Bigotes

3.- Describir en detalle cada una de las secciones anteriores. Describa con sus propias palabras de qué se trata y cuál es el propósito de cada una de estas secciones.

4.- De los resultados obtenidos en los diferentes gráficos y las matrices ¿Cuáles son sus conclusiones al respecto?

**Ejercicio 1: Profesor (*cd-2024-1-tia4-equipo\_X-ejercicio\_1*)**

Analizar el problema del profesor través de la técnica de Análisis de Componentes Principales (PCA). El Científico de Datos es libre de utilizar todas las técnicas conocidas para realizar un buen análisis de los datos. Nota: las columnas del dataset se explican por sí solas.

Enunciado del problema: El sistema escolar de una gran ciudad quería determinar las características de un gran profesor, por lo que pidió a 120 estudiantes que califican la importancia de cada uno de los 9 criterios siguientes utilizando una escala Likert de 1 a 10, en la que 10 representa que una característica concreta es extremadamente importante y 1 que la característica no es importante.

Data: ***data-profesor.xlsx***

**Ejercicio 2: Estudiantes (*cd-2024-1-tia4-equipo\_X-ejercicio\_2*)**

Analizar el problema de los estudiantes en relación al rendimiento en el curso de matemáticas a través de la técnica de Análisis de Componentes Principales (PCA). El Científico de Datos es libre de utilizar todas las técnicas conocidas para realizar un buen análisis de los datos. **Nota**: Utilice el archivo adicional para la descripción de los datos (columnas): ***data-estudiantes-info.txt***

Data: ***data-estudiantes.csv***

**Ejercicio 3: Películas (*cd-2024-1-tia4-equipo\_X-ejercicio\_3*)**

Analizar el problema de las películas (gustos, actores, rating) a través de la técnica de Análisis de Componentes Principales (PCA). El Científico de Datos es libre de utilizar todas las técnicas conocidas para realizar un buen análisis de los datos. **Nota**: Utilice el archivo adicional para la descripción de los datos (columnas): ***data-peliculas-info.pdf***

Data: ***data-peliculas.csv***

**RUBRICA**

El trabajo se hará obligatoriamente en grupo; por lo tanto, si el estudiante no ha conformado un grupo, debe hacerlo antes de comenzar esta actividad. La dinámica de la TIA es la siguiente:

1.- Descargar este documento. Utilizar este documento para responder a las preguntas y devolver al profesor para su evaluación.

2.- Conformar un grupo de trabajo y colocar los integrantes del mismo en este documento.

3.- Antes de responder cada pregunta, revise el material propuesto en la clase y la plataforma.

4.- Al terminar la tarea, solamente se entregará un (1) trabajo por Grupo. Es decir, no hagan entregas individuales.

**Rúbrica: Criterios de Evaluación de la Tarea**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Criterio** | | | | | **Peso** | **Calificación** |
| **1** | * **Elabora un documento de entrega en el formato y presentación solicitados (bien organizado y presentable). Cada ejercicio por separado.** * **Presenta descripción de la tarea de manera clara y con una buena redacción.** * **Presenta un buen análisis de resultados y conclusiones coherentes y bien redactadas sobre el trabajo realizado** | | | | | **15** |  |
| **2** | **Presenta un video de todas las actividades realizadas. El vídeo debe tener una duración mínima de 5 minutos y máxima de 8 minutos. Se demuestra el trabajo colaborativo. (Si no aparece en el video, no tiene calificación). ATENCIÓN: Buena calidad y buen sonido.** | | | | | **15** |  |
| **3** | **Elabora el Ejercicio #1 en cada una de sus partes. El ejercicio presenta todos los elementos solicitados, bien comentados y un buen análisis del problema** | | | | | **30** |  |
| **4** | **Elabora el Ejercicio #2 en cada una de sus partes. El ejercicio presenta todos los elementos solicitados, bien comentados y un buen análisis del problema** | | | | | **30** |  |
| **5** | **Elabora el Ejercicio #3 en cada una de sus partes. El ejercicio presenta todos los elementos solicitados, bien comentados y un buen análisis del problema** | | | | | **30** |  |
| **6** | **Los gráficos son de buena calidad, bien diseñados, con todas las etiquetas necesarias y la información correcta** | | | | | **10** |  |
| **7** | **Los matrices son correctas y se explica su contenido de manera correcta** | | | | | **10** |  |
|  | **NOTA** |  | **PUNTOS** |  | **TOTAL** | **150** |  |

**Nota**: El máximo de puntuación del trabajo es 150 puntos. Se evaluará sobre 140. Se requieren un mínimo de 84 puntos para aprobar. **ATENCIÓN**: El no participar ni colaborar con el equipo al que pertenece tiene una penalización del 40% de la nota.