PROMiDAT Iberoamericano
Machine Learning con Python II
Métodos NO Supervisados

- Las tareas tienen fecha de entrega una semana después a la clase y deben ser entregadas antes del inicio de la clase siguiente.
- Cada día de atraso en implicará una pérdida de 10 puntos.
- Las tareas son estrictamente de carácter individual, tareas iguales se les asignará cero puntos.
- En nombre del archivo debe tener el siguiente formato: `Tarea1_nombre_apellido.html`. Por ejemplo, si el nombre del estudiante es Luis Pérez: `Tarea1_luis_perez.html`. Para la tarea número 2 sería: `Tarea2_luis_perez.html`, y así sucesivamente.
- Todas las preguntas tienen el mismo valor.
- Esta tarea tiene un valor de un 25 % respecto a la nota total del curso.

Tarea Número 1

1. [25 puntos] En este ejercicio vamos a usar la tabla de datos `SpotifyTop2018_40_V2.csv`, que contiene una lista de 40 de las canciones más reproducidas en Spotify en el año 2018. Los datos incluyen una serie de características importantes del audio de cada canción.

   La tabla contiene 40 filas y 11 columnas, las cuales se explican a continuación.

   - `danceability`: Describe qué tan apta para bailar es la canción
   - `denergy`: Representa una medida de intensidad y actividad.
   - `dloudness`: Sonoridad general de la pista en decibelios.
   - `dspeechiness`: Detecta la presencia de palabras en la canción.
   - `dacousticness`: Indica qué tan acústica es la canción.
   - `dinstrumentalness`: Indica si la canción contiene o no voces.
   - `dliveness`: Detecta la presencia de público en la grabación.
   - `dvalence`: Describe la positividad musical transmitida por la canción.
   - `dtempo`: Es el tempo estimado general de una pista en beats por minuto.
   - `dduration_ms`: Es la duración de la canción en milisegundos.
   - `dtime_signature`: Especifica cuántos beats hay en cada barra o medida.

   Nota: Todas son variables numéricas y no tienen NAs. Realice lo siguiente:

   a) Calcule el resumen numérico, interprete los resultados para dos variables.
   b) Realice el test de normalidad para una variable e interprete el resultado.
   c) Realice un gráfico de dispersión e interprete dos similitudes en el gráfico.
   d) Para dos variables identifique los datos atípicos, si los hay.

e) Calcule la matriz de correlaciones, incluya alguna de las imágenes que ofrece Python e interpréte dos de las correlaciones. Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.

f) Efectúe un ACP y dé una interpretación siguiendo los siguientes pasos:

- En el círculo de correlación determine la correlación entre las variables.
- Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano.
- En el plano de los componentes 1 y 3 interprete las canciones In My Feelings, In My Mind, Havana, Candy Paint y HUMBLE, que son mal representadas en los componentes 1 y 2.

2. [25 puntos] En este ejercicio vamos a usar los datos `TablaAffairs.csv`, los cuales recopilan información sobre infidelidades en parejas casadas, como lo es la edad de la persona, años de casado y el nivel de educación.

   La tabla contiene 601 filas y 9 columnas, las cuales se explican a continuación.

   - `TiempoInfiel`: Medida del tiempo que pasó en la relación fuera del matrimonio.
   - `Genero`: Género del individio
   - `Edad`: Edad del individuo
   - `AnnosCasado`: Años que lleva casado
   - `Hijos`: Indica si hay hijos en el matrimonio o no.
   - `Religioso`: índice del grádo de religiosidad. Entre más alto más religioso.
   - `Educacion`: Índice del grado de educación. Entre más alto, mayor educación.
   - `Ocupacion`: Indica la categoría de ocupación de la persona.
   - `Valoracion`: Valoración que da el individuo al matrimonio.

   Realice lo siguiente:

   a) Calcule el resumen numérico, interprete los resultados para una variable.

   b) Calcule la matriz de correlaciones, incluya alguna de las imágenes que ofrece Python e interpréte dos de las correlaciones. Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.

   c) Usando solo las variables numéricas efectúe un ACP y dé una interpretación siguiendo los siguientes pasos: 1) en el plano principal encuentre 4 clústeres, 2) en el círculo de correlación determine la correlación entre las variables y 3) explique la formación de los clústeres basado en la sobre-posición del círculo y el plano.

   d) Ahora convierta las variables Género e Hijos en Código Disyuntivo Completo y repita el ACP ¿Se gana interpretabilidad al convetir Género e Hijos en Código Disyuntivo Completo?

3. [25 puntos] En este ejercicio vamos a realizar un ACP para la tabla `SAheart.csv` la cual contiene variables numéricas y categóricas mezcladas. La descripción de los datos es la siguiente:

Datos Tomados del libro: The Elements of Statistical Learning Data Mining, Inference, and Prediction de Trevor Hastie, Robert Tibshirani y Jerome Friedman de la Universidad de Stanford. .[Ex]ample: South African Heart Disease: A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of coronary heart disease. Many of the coronary heart disease positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their coronary heart disease event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal. Below is a description of the variables:

- `sbp`: systolic blood pressure (numérica)
- `tobacco`: cumulative tobacco (kg) (numérica)
- `ldl`: low densiity lipoprotein cholesterol (numérica)
- `Adiposity`: adiposity (numérica)
- `famhist`: family history of heart disease (Present, Absent) (categórica)
- `typea`: type-A behavior (numérica)
- `Obesity`: obesity (numérica)
- `alcohol`: current alcohol consumption (numérica)
- `age`: age at onset (numérica)
- `chd`: coronary heart disease" (categórica)

Las dos variables categóricas se explican como sigue: `famhist` significa que hay historia familiar de infarto y que la variable `chd` significa que la persona murió de enfermedad cardíaca coronaria. Realice lo siguiente:

a) Efectúe un ACP usando solo las variables numéricas y dé una interpretación siguiendo los siguientes pasos:
   - En el plano principal encuentre los clústeres.
   - En el círculo de correlación determine la correlación entre las variables,
   - Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano.

b) Efectúe un ACP usando las variables numéricas y las variables categóricas (recuerde recodificar las categóricas usando código disyuntivo completo). Luego dé una interpretación siguiendo los siguientes pasos:
   - En el plano principal encuentre los clústeres.
   - En el círculo de correlación determine la correlación entre las variables,
   - Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano.
   - Explique las diferencias de este ACP respecto al anterior (usando solo las variables numéricas. ¿Cuál le parece más interesante? ¿Por qué?

4. [25 puntos] Programe una clase derivada (que herede) de la clase `class ACP`, que fue presentada durante la lección, que incluya adicionalmente lo siguiente:

a) Que sobrecargue el constructor de la clase `__init__` para seleccionar variables, es decir, que reciba adicionalmente un vector con los números de una columna o nombres de la variables respectivas de manera que el atributo `datos` sea modificado para eliminar esas columnas. Y así todos los cálculos sean realizados eliminando estas columnas.

b) Que sobrecargue los métodos `plot_plano_principal` y `plot_sobreposicion` de manera tal que en estos gráficos se puedan eliminar individuos mal representados, esto basado en en el $cos^2(x)$ mediante el atributo `cos2_ind`. Es decir, que estos métodos reciban un parámetros adicional que el es porcentaje mínimo en el coseno cuadrado aceptable para que los individuos aparezcan en estos gráficos.

c) Verifique la nueva clase programada con los datos del ejercicio 1.

Entregables:

1. Suba en el Aula Virtual en el `Script` generado.

2. Genere desde Jupyter Notebook un documento autoreproducible HTML con la solución de la tarea y súbalo en el Aula Virtual.