Profesor: Dr. Oldemar Rodríguez Rojas
PROMiDAT Iberoamericano
Ciencia de Datos con Python

- 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.pdf`. Por ejemplo, si el nombre del estudiante es Luis Pérez: `Tarea1_luis_perez.pdf`. Para la tarea número 2 sería: `Tarea2_luis_perez.pdf`, y así sucesivamente.
- Todas las preguntas tienen el mismo valor.
- Esta tarea tiene un valor de un 12.5 % respecto a la nota total del curso.

Tarea Número 6

1. [25 puntos] Considérese la siguiente tabla datos, la cual contiene las importaciones hechas por los países centroamericanos, provenientes de México, entre 1979 y 1988 (Está en el aula virtual con el nombre `ImportacionesMexico.csv`):

	Costa Rica	El Salvador	Guatemala	Honduras	Nicaragua	Panama
1979	44,4	27,2	45,6	20	6	14,1
1980	75,5	11,8	58,9	22,6	17,8	14,4
1981	110,7	50,6	128,3	17,2	119,4	118,5
1982	80,3	70,6	102,2	15,2	154,9	146,1
1983	81,6	82,3	89	35,1	169,4	127,1
1984	76,4	97,4	185	51	75,5	129
1985	32	89,5	195,3	31,1	33,4	110,2
1986	55,5	63,1	66,3	24,4	9,7	66,7
1987	74,3	72,6	76,3	28,1	11,2	110,7
1988	84,5	76,2	80,1	29,5	11,8	110,2

   Para esta tabla realice lo siguiente:

   a) Cargue la tabla de datos `ImportacionesMexico.csv`.

   b) Ejecute un Clustering Jerárquico con la agregación del Salto Máximo, Salto Mínimo, Promedio y Ward. Grafique el dendograma con cortes para dos y tres clústeres.

   c) Usando tres clústeres interprete los resultados del ejercicio anterior para el caso de agregación de Ward mediante gráficos de barras y gráficos tipo Radar.

   d) Ejecute el método $k-$medias para $k = 3$.

   e) Interprete los resultados del ejercicio anterior usando gráficos de barras y gráficos tipo Radar.

f) Agregue en el archivo de datos el cluster en el quedó cada individuo y suba el archivo resultante en el aula virtual como parte de la solución de la tarea.

2. [25 puntos] 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. Example: 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`: (numérica)
   - `famhist`: family history of heart disease (Present, Absent) (categórica)
   - `typea`: type-A behavior (numérica)
   - `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.

   a) Cargue la tabla de datos `SAheart.csv`.

   b) Usando solamente las variables numéricas, ejecute un Clustering Jerárquico con la agregación del Salto Máximo, Salto Mínimo, Promedio y Ward. Grafique el dendograma con cortes para dos y tres clústeres.

   c) Usando 3 clústeres interprete los resultados del ejercicio anterior para el caso de agregación de Ward mediante gráficos de barras y gráficos tipo Radar.

   d) Usando solamente las variables numéricas, ejecute el método $k-$medias para $k = 3$.

   e) Interprete los resultados del ejercicio anterior usando gráficos de barras y gráficos tipo Radar.

   f) Usando todas las variables con `famhist` y `chd` recodificadas como códigos disyuntivos completos (variables Dummy), repita los ejercicios anteriores, igual con 3 clústeres en ambos métodos. Para esto en Clustering Jerárquico utilice una distancia especial para datos binarios:

```
grupos = fcluster(linkage(pdist(datos), method = 'ward', metric='binary'),
         3, criterion = 'maxclust')
```

g) ¿Cuál de los análisis anteriores le parece más interesante? ¿Porqué?

3. [25 puntos] El conjunto de datos `DatosBeijing.csv` contiene datos por hora de la concentración de la partícula PM2.5 en la ciudad de Beijing, también incluye datos meteorológicos del Aeropuerto Internacional de Beijing. Contiene un `Id` y 12 variables que se explican seguidamente:

   - `Id`: Número de fila.
   - `Anno`: Año de datos en esta fila.
   - `Mes`: Mes de datos en esta fila.
   - `Dia`: Día de datos en esta fila.
   - `Hora`: Hora de datos en esta fila
   - `ConcetracionParticula_pm2.5`: Concentración de PM2.5.
   - `PuntoRocio`: Punto de rocío.
   - `Temperatura`: Temperatura.
   - `Presion`: Presión (hPa).
   - `DireccionViento`: Dirección del viento combinado.
   - `VelocidadViento`: Velocidad del viento acumulada.
   - `HorasNieve`: Horas acumuladas de nieve.
   - `HorasLluvia`: Horas acumuladas de lluvia.

   Efectúe un análisis de k-medias siguiendo los siguientes pasos:

   a) Cargue la tabla de datos y ejecute un `dropna().describe()`, y encuentre la dimensión de la tabla de datos (número de filas y columnas) `datos.shape`, con esto verifique la correcta lectura de los datos.

   b) Elimine las filas con `NA`. ¿Cuántas filas se eliminaron?

   c) Elimine de la tabla de datos la variable `DireccionViento`. ¿Por qué se debe eliminar? ¿Qué otra alternativa se tiene en lugar de eliminarla?

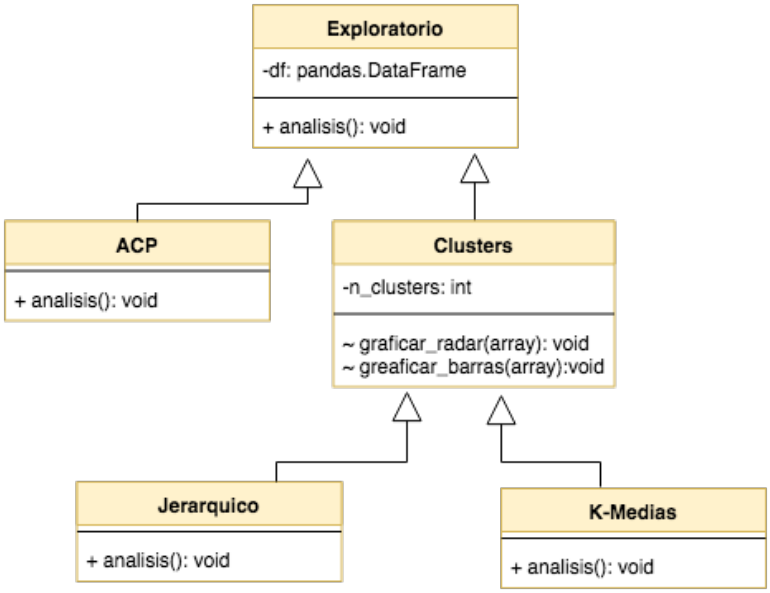
   d) ¿Qué pasa si ejecutamos un clustering jerárquico para esta tabla de datos. ¿Por qué sucede esto?

   e) Investigue y explique para que sirven los atributos `max_iter` (similar a `iter.max` en **R**) y `n_init` (similar a `nstart` en **R**), ambos de la clase `sklearn.cluster.KMeans`. Luego ejecute un k-medias con $k = 3$ usando `max_iter=5000` y `n_init=10`.

   f) Dé una interpretación de los resultados usando un gráfico de barras.

   g) Construya el Codo de Jambu usando `max_iter=1000` y `n_init=1`, ¿cuántos conglomerados (clústeres) sugiere el codo?

4. [25 puntos] Programe la jerarquía de clases que se muestra en el siguiente gráfico, la cual fue diseñada especialmente para facilitar los análisis exploratorios de datos vistos hasta ahora en el curso:



La idea es que a través de una instancia de alguna de las clases `Exploratorio`, `ACP`, `Jerárquico` o k-`medias` con solamente ejecutar el método `análisis` automáticamente se despliegan todos los análisis correspondientes a cada caso vistos en el curso (todos hacen por defecto el análisis exploratorio básico). Para esto la clase Base `Exploratorio` tiene un atributo que es una `Data Frame` de `Pandas` y un método `análisis` que realiza al menos los siguiente: Despliega un encabezado de los datos(head), dimensión de la tabla, estadísticas básicas, los percentiles, valores atípicos, boxplot, distribucion de densidad, histogramas y Tests de normalidad. La clase `ACP` agrega al método `analisis` gráficos para el Plano principal, el Círculo de Correlaciones y la inercia acumulada. La clase `Clústeres` agrega un atributo para la cantidad de clústeres y métodos para los gráficos de Radar y de Barras que son usados en Clasificación Jerárquica y k-medias. La clase `Jerárquico` agrega en el método `análisis` los gráficos y análisis vistos en clase. La clase `Jerárquico` agrega en el método k-`medias` agrega los gráficos y análisis vistos en clase para este método.

Entregables:

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

2. Suba al Aula Virtual el archivo generado en el ejercicio 1f).

PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de
Formación en Minería de Datos