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 5

1. [25 puntos] Esta pregunta utiliza los datos sobre la conocida historia y tragedia del Titanic, usando los datos (`titanic.csv`) de los pasajeros se trata de predecir la supervivencia o no de un pasajero.

   La tabla contiene 12 variables y 1309 observaciones, las variables son:

   - `PassegerId`: El código de identificación del pasajero (valor único).
   - `Survived`: Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
   - `Pclass`: En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
   - `Name`: Nombre del pasajero (valor único).
   - `Sex`: Sexo del pasajero.
   - `Age`: Edad del pasajero.
   - `SibSp`: Cantidad de hermanos o cónyuges a bordo del Titanic.
   - `Parch`: Cantidad de padres o hijos a bordo del Titanic.
   - `Ticket`: Número de tiquete (valor único).
   - `Fare`: Tarifa del pasajero.
   - `Cabin`: Número de cabina (valor único).
   - `Embarked`: Puerto donde embarco el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

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

   b) Recodifique las variables cualitativas que están codificadas con números e ignore variables que no se deben usar, es decir, con valor único.

   c) Calcule todas las estadísticas básicas en esta tabla de datos, incluya los análisis vistos en clase para variables numéricas.

d) Realice gráficos de barras que ayuden a determinar la distribución de las variables categóricas.

e) Realice gráficos tipo boxplot para encontrar los datos atípicos en esta tabla de datos.

f) Realice histogramas, gráficos de la función de densidad y test de normalidad para al menos dos variables de esta tabla de datos.

g) Realice al menos dos scatter plot individuales y un gráfico de todas las variables 2 a 2 usando el paquete Seaborn.

h) Calcule y grafique la matriz de correlaciones.

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) Repita el ejercicio 1 usando esta tabla de datos.

3. [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) Efectúe un ACP y dé una interpretación siguiendo los siguientes pasos: 1) En el plano principal encuentre los clústeres o segmentos. 2) En el círculo de correlación determine la correlación entre las variables e interprételas. 3) Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano, interprete (los años 1987 y 1988 no se deben interpretar ya que están muy mal representados en este plano, debido a que los cosenos cuadrados de estos año en las componentes 1 y 2 son muy bajos).

b) En el plano y círculo 1-3 (componentes 1 y 3) interprete los años 1987 y 1988.

4. [25 puntos] Para la tabla del ejercicio 2 realice lo siguiente:

a) Usando solamente las variables numéricas, efectúe un ACP y dé una interpretación siguiendo los siguientes pasos: 1) En el plano principal encuentre los clústeres o segmentos. 2) En el círculo de correlación determine la correlación entre las variables e interprételas. 3) Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano, interprete, utilice para esto 3 clústeres.

b) Usando todas las variables con `famhist` y `chd` recodificadas como códigos disyuntivos completos (variables Dummy), efectúe un ACP y dé una interpretación siguiendo los siguientes pasos: 1) En el plano principal encuentre los clústeres o segmentos. 2) En el círculo de correlación determine la correlación entre las variables e interprételas. 3) Explique la formación de los clústeres basado en la sobre-posición del círculo y el plano, interprete, utilice para esto 3 clústeres.

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

Entregables:

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

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

PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de
Formación en Minería de Datos