Universidad de Playa Ancha

CIF 8458 Ing. Conocimiento
Pauta Prueba Integral I
Lunes 13 de Noviembre de 2023

1 Preguntas

1. (0,5 pts) ¿A qué se refiere el proceso KDD?

   R: Se refiere al proceso de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles. Consiste en integrar varias tecnologías para la gestión de los datos.

2. (0,5 pts) ¿Qué tecnologías se integran en el proceso KDD?

   R: Manejo de Bases de Datos, data ware-housing, máquinas de aprendizaje, visualización, computación paralela, etc.

3. (0,5 pts) ¿Qué significa que un resultado sea válido?

   R: Los patrones deben seguir siendo precisos para datos nuevos, y no sólo para aquellos que han sido usados en su obtención.

4. (0,5 pts) ¿Por qué es importante tener una evaluación subjetiva por parte del usuario?

   R: La retroalimentación con el usuario final es muy importante para el éxito del proyecto.

	Country	Co2-Emissions	Fertility Rate	Infant mortality	Life expectancy	Physicians per thousand	Urban_population	continent
0	Afghanistan	8,672	4.47	47.9	64.5	0.28	9,797,273	Asia
1	Albania	4,536	1.62	7.8	78.5	1.20	1,747,593	Europe
2	Algeria	150,006	3.02	20.1	76.7	1.72	31,510,100	Africa
3	Andorra	469	1.27	2.7	NaN	3.33	67,873	Europe
4	Angola	34,693	5.52	51.6	60.8	0.21	21,061,025	Africa
...	...	...	...	...	...	...	...	...
190	Venezuela	164,175	2.27	21.4	72.1	1.92	25,162,368	Americas
191	Vietnam	192,668	2.05	16.5	75.3	0.82	35,332,140	Asia
192	Yemen	10,609	3.79	42.9	66.1	0.31	10,869,523	Asia
193	Zambia	5,141	4.63	40.4	63.5	1.19	7,871,713	Africa
194	Zimbabwe	10,983	3.62	33.9	61.2	0.21	4,717,305	Africa

Figura 1: Dataframe con información sobre países

5. Se tiene un dataframe llamado *df* (que se muestra parcialmente en la Figura 1) que contiene información de las expectativas de vida de la gente en distintos países del mundo, a lo largo del tiempo.

   (a) (0,5 pts) ¿Con qué línea de comando obtengo cuantos países distintos tiene las emisiones de CO_2 por sobre el promedio?

   ```
   len(df3[df3['Co2-Emissions'] > df3['Co2-Emissions'].mean()]['Co2-Emissions'])
   ```

   (b) (0,5 pts) ¿Con qué línea de comando obtengo el continente con la más alta expectativa de vida en promedio?

   ```
   df3['Life expectancy'].groupby(df3['continent']).mean().idxmax()
   ```

   (c) (0,5 pts) Complete el código para generar el violinplot de la izquierda en la Figura 2.

   ```
   import seaborn as sns
   import matplotlib.pyplot as plt
   sns.violinplot(data=df ,x='Co2-Emissions', y='continent', hue='continent')
   ```

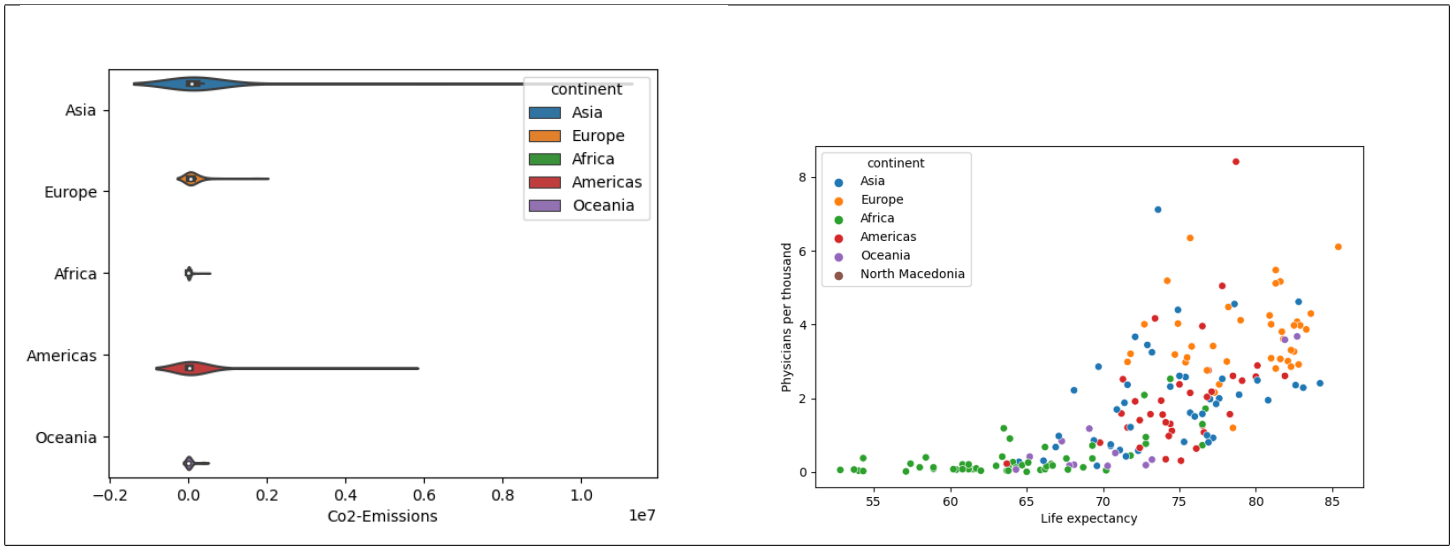

Figura 2: Izquierda: Violinplot de las emisiones de CO_2 cada continente. Derecha: Scatter plot de expectativas de vida versus el número de médicos cada mil habitantes.

(d) (0,5 pts) Complete el código para generar el scatter plot de la derecha en la Figura 1.

```
fig = plt.figure(figsize =(8 ,5))
sns.scatterplot(x='Life expectancy', y='Physicians per thousand',data=df ,hue='continent')
```

(e) (1,0 pto) De los gráficos de la Figura 2, ¿Qué puede concluir?

Respecto de las emisiones de CO_2, los países de las américas y Asia son los que generan más emisiones, teniendo ambos continentes al menos un país que tiene una emisión notablemente mayor.
Oceanía y África tienen las menores emisiones. Es notable notar que apesar de su urbanización, Europa tiene niveles de emisiones más cercanos a estos últimos que a América y Asía, por lo que se asume que hay más preocupación departe de este continente en bajar las emisiones de CO_2.

Respecto del diagrama de dispersión, en áfrica, la relación es practicamente nula, ya que la proporción de médicos se mantiene baja, y al parecer no incide mayormente en las expectativas de vida de los países.

En las Américas y Europa se aprecia una tendencia lineal, es decir, mientras más médicos hay por habitante, más es la expectativa de vida. Es notable mencionar que en Asia se logran expectativas de vida similares a las de Europa con menos médicos por población, esto se puede deber a una cultura de alimentación y ejercicios que ayuda a algunos países asiáticos a tener más longevidad.

(f) (0,5 pts) Respecto de los boxplot de la Figura 3, ¿Qué puede concluir?

En general todos los continentes presentan una mediana muy cercana al Q1 (asimetría hacia la izquierda). Es decir la mitad de los países con menor población urbana tiene menor dispersión que el 50% de los países con mayor población. Asia es la que presenta la mayor población urbana, teniendo 2 outliers más destacados entre los 4 que presenta. Europa es el continente con menor dispersión, sin embargo, muestra 6 outliers que presentan mayor población urbana.

(a) (0,5 pts) Complete el código para generar el gráfico de la Figura 4.

```
corr = df[['Fertility Rate','Infant mortality','Life expectancy','Physicians per thousand']].c
cmap = sns.diverging_palette (220 , 10, as_cmap = True)
sns_plot = sns.heatmap (corr , cmap = cmap , center =0, square = True)
```

(b) (0,5 pts) Comente el gráfico de la Figura 4.

A mayor tasa de fertilidad hay una tendencia a una mayor mortalidad infantil, esto podría producirse al haber más niños y niñas, hay más riesgo de que estos fallezcan. También se observa la relación entre la expectativa de vida y la cantidad de médicos cada mil habiltantes, esto corrobora la relación lineal que se veía en el scatter plot de la pregunta anterior. Otra relación a destacar es quela mortalidad infantil está inversamente relacionada con la expectativa de vida, lo que parece lógico, ya que al haber mortalidad infantil, el promedio de la expectativa de vida tiende a bajar.

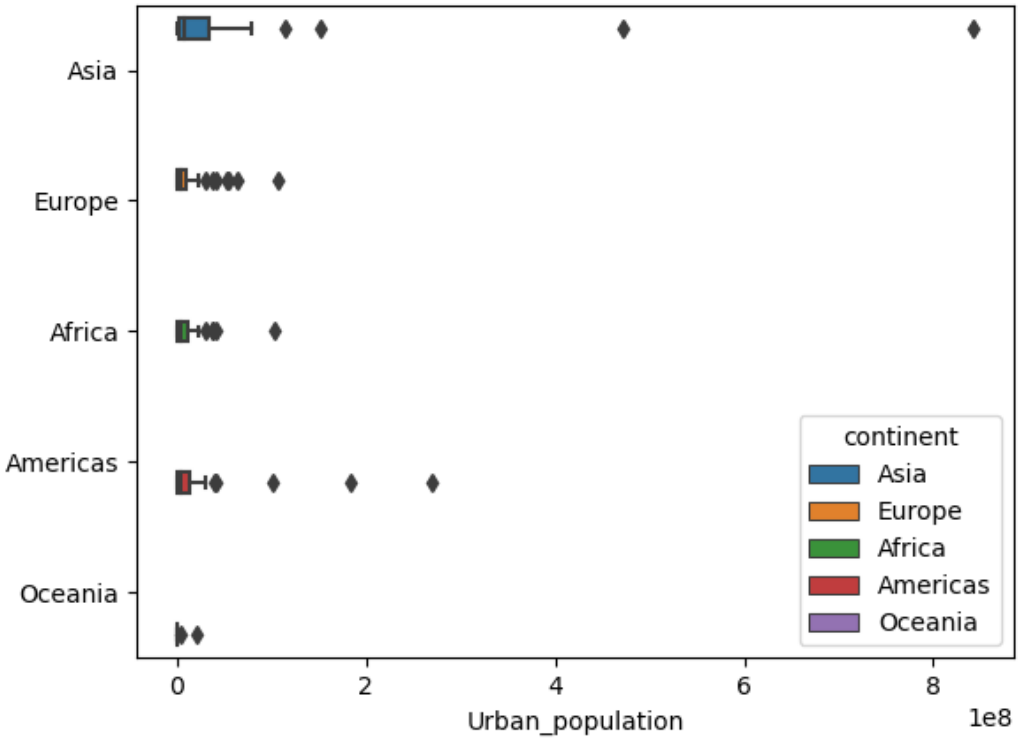

Figura 3: Boxplots de la cantidad de población urbana por país

Figura 4: Matrices de correlación.