Tecnicatura en Ciencia de Datos e Inteligencia Artificial
Segunda Entrega de Práctica Profesionalizante I

Solución de Análisis Predictivo para el Plan Nacional de Planificación Familiar

DOCENTE
Charletti, Carlos Ignacio

GRUPO DATAMINDS
1. Ayán, María Trinidad
2. Giordano, Ariel Eduardo
3. Herrera, Edgar Fabián
4. Quiroga, Fernanda

EMPRESA
DataVista Analytics

COLAB
https://colab.research.google.com/drive/1WSoorvK0bv9fxRAyAMfYOAG_0mfoKSob?usp=sharing

ÍNDICE

Segunda Entrega - Exploración y Análisis de Datos (EDA)

1. Objetivo

Teniendo en cuenta que los datasets que presentan mayor continuidad en el tiempo son los vinculados a salud reproductiva y planificación familiar, nuestra propuesta es centrarse en esta subárea del tema general salud. La finalidad del proyecto es evaluar si la cantidad de métodos anticonceptivos efectivamente distribuidos por el sistema de atención público en el Gran Córdoba son suficientes para la cantidad de personas sin acceso a la cobertura médica privada. Para ello realizaremos un análisis utilizando técnicas de EDA para identificar patrones. Nos valdremos de las bibliotecas Pandas, Matplotlib, y Seaborn.

2. Desarrollo

2.1 Exploración Inicial de Datos

El dataset fue cargado y revisado para asegurarse de que los datos se encuentran en el formato adecuado.

```
import pandas as pd

# Cargar el archivo .xlsx

dataset = pd.read_excel('dataset practica prof.xlsx')

print(dataset.info())  # Información general sobre el dataset

print(dataset.head())  # Ver las primeras filas
```

El conjunto de datos tiene 120 filas y 7 columnas. Las columnas proporcionan información acerca de las jurisdicciones, la cantidad de unidades distribuidas de dos tipos de métodos anticonceptivos, el trimestre, el año, y el tipo de dato. Cada fila representa una observación de la cantidad de métodos anticonceptivos distribuidos por trimestre y jurisdicción durante el año 2021.Las columnas principales son:

Columnas del dataset

- jurisdicion_id: ID de la jurisdicción
- jurisdicion_nombre: Nombre de la jurisdicción
- cantidad_unidades_DIUT_cobre_380mm: Cantidad de unidades del DIU de cobre de 380mm

- cantidad_unidades_etonogestrel_68mg: Cantidad de unidades del implante de etonogestrel de 68mg
- trimestre: Trimestre en que se reportaron los datos.
- anio: Año de reporte (en este caso, todos los datos son del año 2021)
- tipo_dato: Tipo de dato

2.2 Análisis Univariado

Se realizó un análisis de las variables numéricas para entender su distribución, utilizando elementos de la estadística descriptiva, como las medidas de tendencia central, las medidas de dispersión, histogramas y boxplots.

Variables Numéricas

1. Cantidad de unidades DIU de cobre 380 mm:

- Media: 89.58
- Desviación estándar: 105.54
- Mínimo: 0
- Máximo: 450

2. Cantidad de unidades de etonogestrel 68mg:

- Media: 1070.83
- Desviación estándar: 1032.19
- Mínimo: 200
- Máximo: 6150

Variables Categóricas

- Jurisdicción: 12 jurisdicciones distintas
- Trimestre: 4 trimestres diferentes
- Año: Solo datos de un año, el 2021
- Tipo de dato: Solo datos consolidados

```
#Estadísticas descriptivas para las variables numéricas:

 print(dataset.describe())

 #Visualización de distribuciones con histogramas y boxplots:

 import matplotlib.pyplot as plt

 dataset['nombre_columna'].hist()

 plt.show()
```

```
  dataset.boxplot(column='nombre_columna')

  plt.show()

  #Para variables categóricas:

  print(dataset['nombre_columna_categorica'].value_counts())
```

Identificación de Outliers

Boxplots

En los gráficos de boxplots, se observan algunos valores atípicos en ambas variables numéricas:

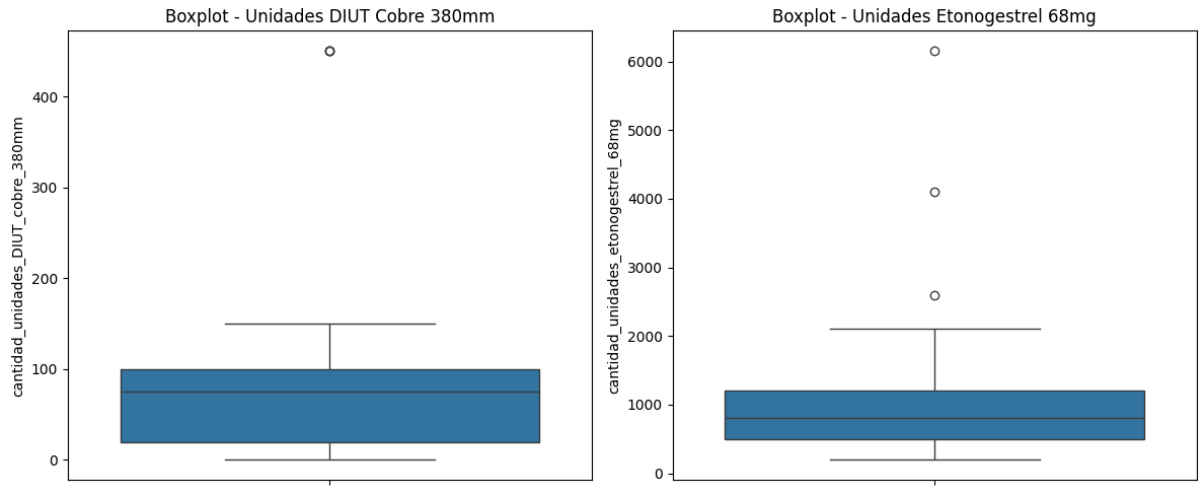
- Unidades de DIU Cobre 380 mm: El valor 450 es un posible outlier.
- Unidades de Etonogestrel 68 mg: Se identifican posibles outliers, particularmente los valores más altos (4100 y 6150).

Cálculo del Z-score

- Los valores de 450 unidades de DIUT Cobre 380 mm aparecen como outliers (z-score > 3).
- Para unidades de etonogestrel 68 mg, el valor 6150 es un outlier significativo con un z-score de 4.97.

```
import matplotlib.pyplot as plt #generar los gráficos.

import seaborn as sns

# Configuración del tamaño de la figura

plt.figure(figsize=(12, 5))

# Boxplot para 'cantidad_unidades_DIUT_cobre_380mm'

plt.subplot(1, 2, 1)

sns.boxplot(y=dataset['cantidad_unidades_DIUT_cobre_380mm'])

plt.title('Boxplot - Unidades DIUT Cobre 380mm')

# Boxplot para 'cantidad_unidades_etonogestrel_68mg'

plt.subplot(1, 2, 2)

sns.boxplot(y=dataset['cantidad_unidades_etonogestrel_68mg'])
```

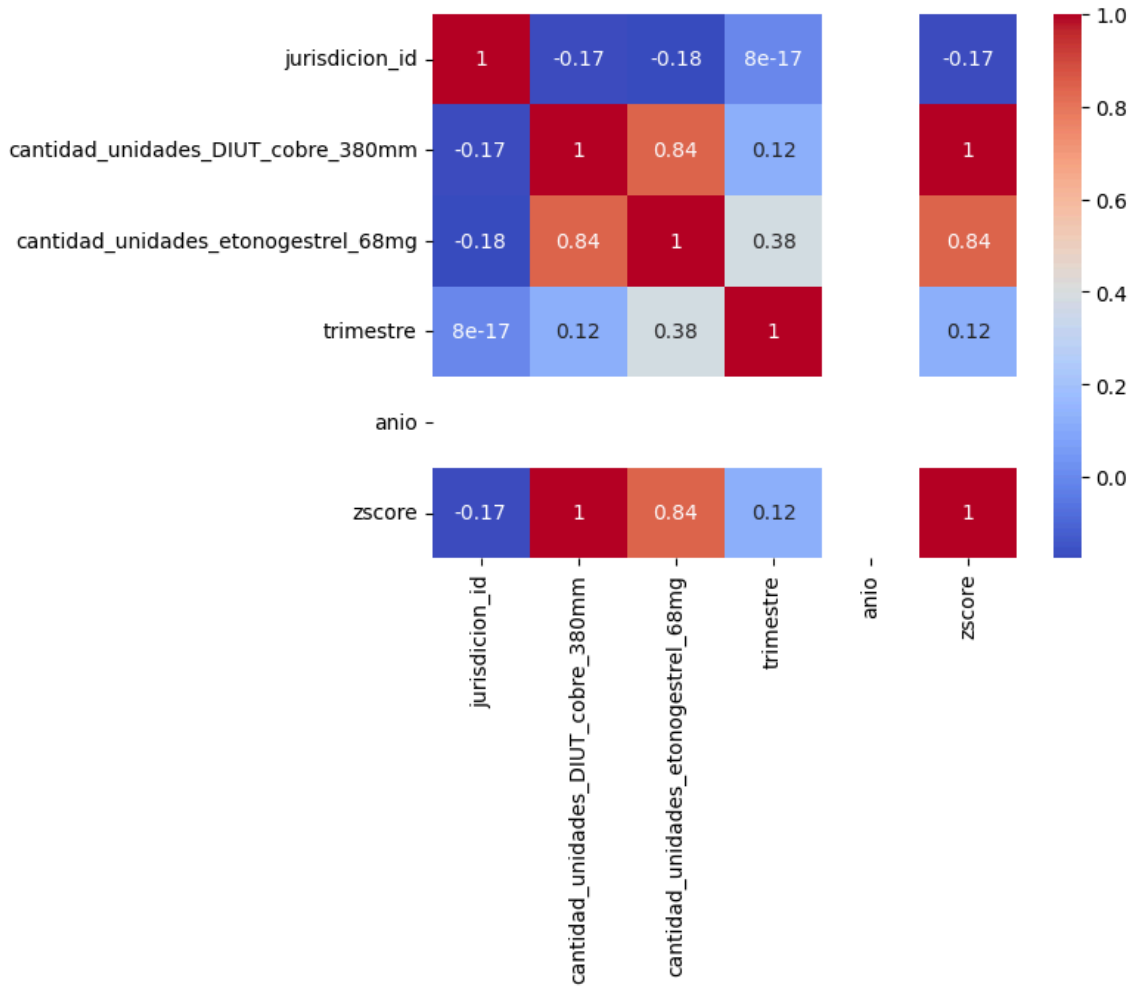
```
plt.title('Boxplot - Unidades Etonogestrel 68mg')

# Ajustar el diseño para que no se solapen los gráficos

plt.tight_layout()

plt.show()
```



2.3 Análisis Bivariado

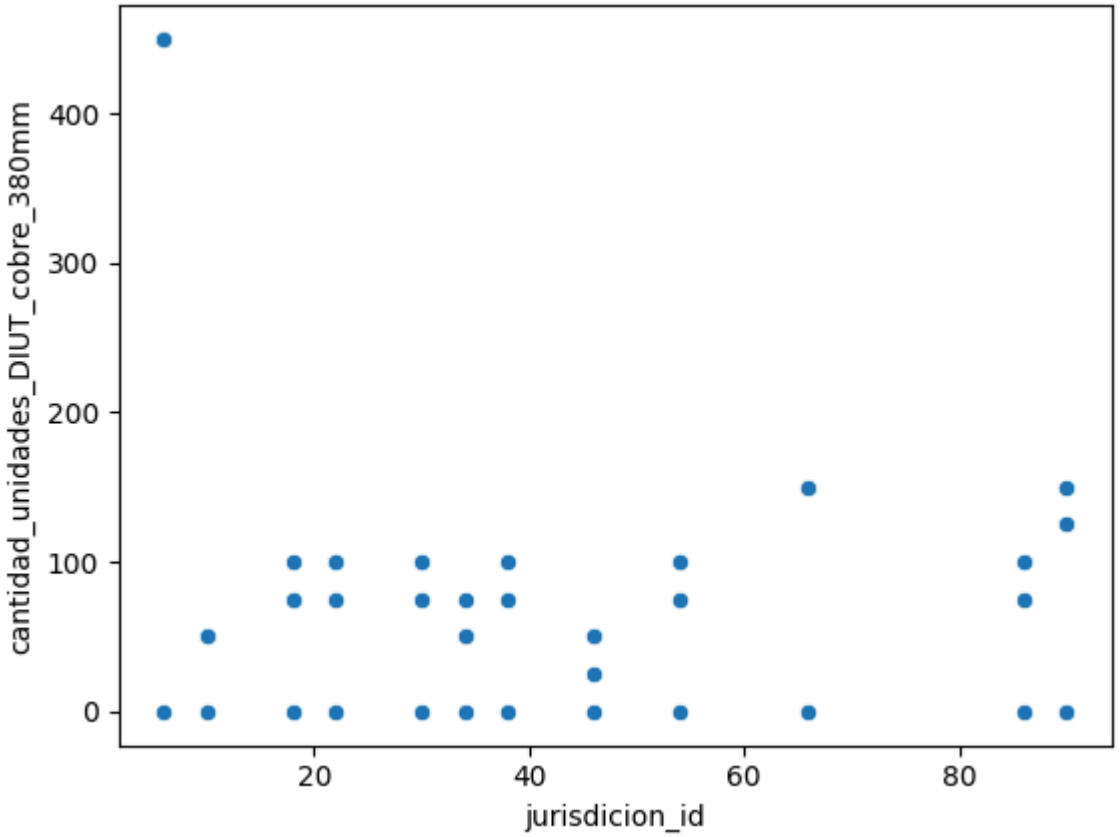
Se realizaron análisis bivariados para identificar posibles relaciones entre las variables numéricas y categóricas. En el presente análisis, no se encontraron correlaciones significativas entre las dos variables numéricas.

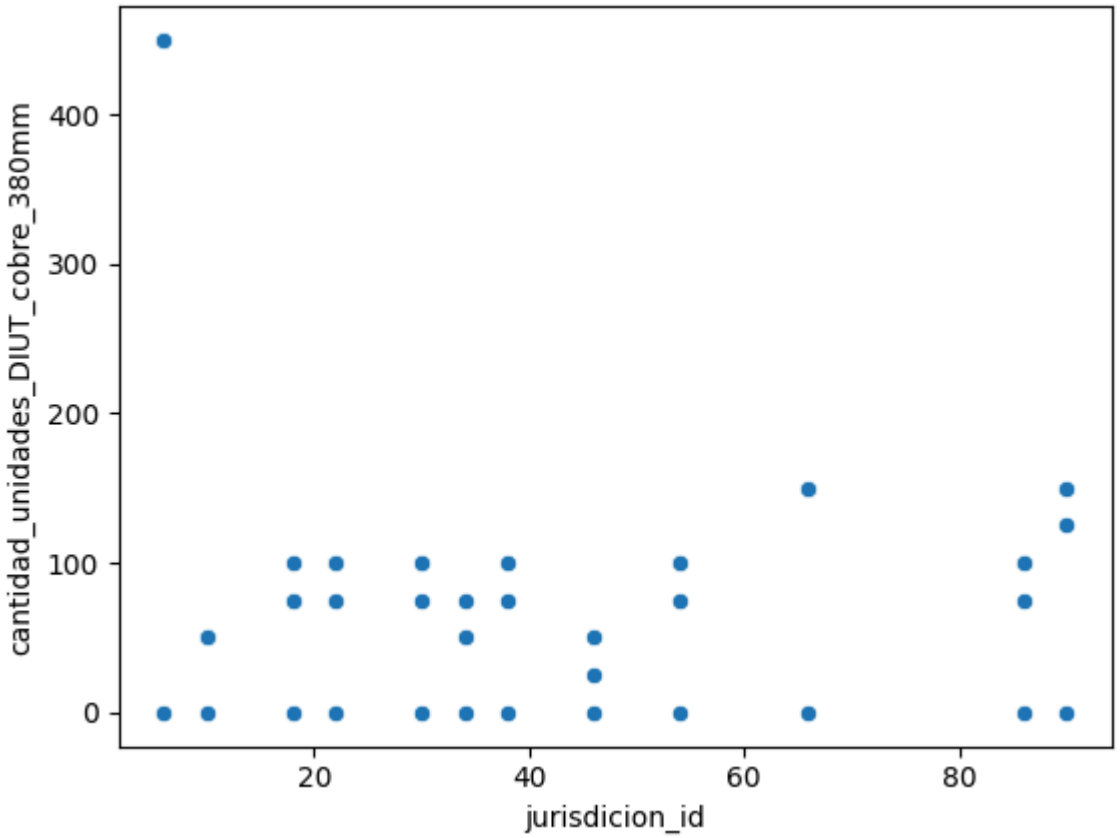
```
#Análisis Bivariado:Evalúa las relaciones entre las variables:

#Mapa de calor de correlación para variables numéricas:

import seaborn as sns

# Seleciona solo columnas mumericas

numeric_dataset = dataset.select_dtypes(include=['number'])

correlation_matrix = numeric_dataset.corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')

plt.show()
```



Comparaciones Categóricas

```
#Gráfico de dispersión para dos variables numéricas:

sns.scatterplot(x='jurisdicion_id',
y='cantidad_unidades_DIUT_cobre_380mm', data=dataset)

plt.show()

#Para comparar variables categóricas con numéricas:

sns.boxplot(x='trimestre', y='cantidad_unidades_DIUT_cobre_380mm',
data=dataset)

plt.show()
```


cantidad_unidades_DIUT_cobre_380mm
400
300
200
100
0
20
40
60
80
jurisdicion_id


cantidad_unidades_DIUT_cobre_380mm
400
300
200
100
0
20
40
60
80
jurisdicion_id

2.4 Limpieza de Datos

Tratamiento de Valores Faltantes

Con la finalidad de identificar valores faltantes, se verificó la presencia de los valores ausente y se utilizó imputación con la mediana de los valores numéricos.

```
#Limpieza de Datos:

#Valores faltantes:

print(dataset.isnull().sum())

#Manejo de valores faltantes:

dataset.fillna(dataset.median(), inplace=True)
```

Normalización y Transformación

Para las variables numéricas, se aplicó una estandarización para aplicar en las futuras etapas del análisis.

```
dataset_scaled =
scaler.fit_transform(dataset[['cantidad_unidades_DIUT_cobre_380mm',
'cantidad_unidades_etonogestrel_68mg']])
```

3. Conclusiones

La Exploración y el Análisis de los Datos (EDA) permitió una visión inicial de la distribución de métodos anticonceptivos en las distintas jurisdicciones. Los métodos de DIU de Cobre 380 mm y los implantes de Etonogestrel mostraron una amplia variabilidad en la cantidad de unidades distribuidas, con algunos valores atípicos. Este análisis sienta las bases para una evaluación más profunda sobre la relación entre la cantidad de métodos distribuidos y la población sin cobertura médica privada.