INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO
NOMBRE: Cerda García Gustavo
Materia: Inteligencia Artificial

Ejercicio de Laboratorio 7. Apertura de DataSets

Ejercicios

- Elabora un programa, en el lenguaje que quieras, que abra y cargue los contenidos del archivo adjunto.
- De cada columna numérica calcula el promedio, la varianza y la desviación estándar.
- Separa los datos en diferentes matrices, de acuerdo con la categoría de los datos y repite los cálculos del paso anterior.

Código

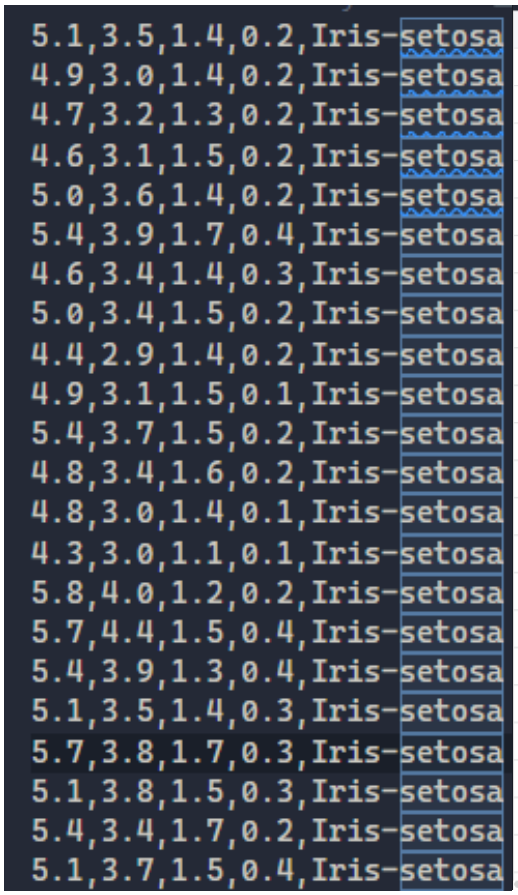
Función `convertir_a_csv (archivo)`

```python
#Nombre: convertir_a_csv
#Desc: Convierte un archivo a CSV
#Entrada:
# - archivo: Nombre del archivo a convertir
def convertir_a_csv(archivo):
    try:
        archivo_salida = archivo.split('.')[0] + '.csv' # Nombre del archivo de salida

        # Leer el archivo
        df = pd.read_csv(archivo, header=None)

        # Convertir a CSV
        df.to_csv(archivo_salida, index=False)

        # Borrar la primera fila, ya que son los encabezados
        df = pd.read_csv(archivo_salida)
        df.to_csv(archivo_salida, index=False, header=False)

        print('Archivo convertido a CSV, con el nombre de {}\n'.format(archivo_salida))
    except FileNotFoundError:
        print('El archivo no existe\n')
    except Exception as e:
        print(f'Error: {e}\n')
```

Esta función toma un archivo de entrada y lo convierte al formato `CSV` utilizando la biblioteca `Pandas.` Primero lee el archivo con '`pd.read_csv`', luego lo guarda como un archivo CSV con '`to_csv`'. Después elimina la fila de los encabezados. Si existe algún error durante el proceso, se maneja adecuadamente.

Salida

Archivo dataset.data | Archivo dataset.csv



```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
```

	A	B	C	D	E
1	**5.1**	**3.5**	**1.4**	**0.2**	**Iris-setosa**
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa
22	5.1	3.7	1.5	0.4	Iris-setosa

Función `main ()`

Parte 1:

```python
# Nombre del archivo
archivo = 'dataset.data'

# Convertir a CSV
convertir_a_csv(archivo)

# ! De cada columna calcular el promedio, la varianza y la desviación estándar
# Leer el archivo
df = pd.read_csv('dataset.csv', header=None)

# print(df)

# Ignoramos la primera fila, ya que son los encabezados, empezamos en la columna 0
for i in range(0, df.shape[1] - 1):
    # Si no es numerico, no se puede calcular (es por las categorias)
    if not df[i].dtype == 'float64':
        continue
    print(f'Columna {i}')
    print(f'Promedio: {df[i].mean()}')
    print(f'Varianza: {df[i].var()}')
    print(f'Desviación estándar: {df[i].std()}')
    print()
```

Ya con el archivo convertido a `CSV`, el código lee el archivo nuevamente con `Pandas` y realiza los cálculos estadísticos sobre cada columna numérica utilizando los métodos '`mean ()`', '`var ()`' y '`std ()`'. Esta función ignora las columnas que no son numéricas (por la columna de las categorías).

Salida

```
Archivo convertido a CSV, con el nombre de dataset.csv

Columna 0
Promedio: 5.843333333333334
Varianza: 0.6856935123042507
Desviación estándar: 0.828066127977863

Columna 1
Promedio: 3.0573333333333337
Varianza: 0.189979418344519
Desviación estándar: 0.4358662849366982

Columna 2
Promedio: 3.7580000000000005
Varianza: 3.116277852348993
Desviación estándar: 1.7652982332594662

Columna 3
Promedio: 1.1993333333333336
Varianza: 0.5810062639821029
Desviación estándar: 0.7622376689603465
```

Parte 2:

```python
# ! Separar los datos en diferentes matrices, de acuerdo a la categoria de la última columna y
repetir el cálculo de promedio, varianza y desviación estándar
# Obtener las categorias, de la ultima columna, menos la primera fila que son los encabezados
categorias = df[df.shape[1] - 1][1:].unique()

# Iterar sobre las categorias
for categoria in categorias:
    print(f'Categoría {categoria}')
    # Obtener las filas que coincidan con la categoria
    df_categoria = df[df[df.shape[1] - 1] == categoria]
    # Ignoramos la primera fila, ya que son los encabezados
    for i in range(0, df_categoria.shape[1] - 1):
        # Si no es numerico, no se puede calcular (es por las categorias)
        if not df_categoria[i].dtype == 'float64':
            continue
        print(f'Columna {i}')
        print(f'Promedio: {df_categoria[i].mean()}')
        print(f'Varianza: {df_categoria[i].var()}')
        print(f'Desviación estándar: {df_categoria[i].std()}')
        print()
    print()
```

El código identifica las categorías en la última columna del dataframe y separa los datos en diferentes dataframes según la categoría. Luego, repite los cálculos estadísticos para cada categoría.

Salida

```
Categoría Iris-setosa
Columna 0
Promedio: 5.006
Varianza: 0.12424897959183677
Desviación estándar: 0.35248968721345136

Columna 1
Promedio: 3.428
Varianza: 0.1436897959183674
Desviación estándar: 0.3790643690962887

Columna 2
Promedio: 1.4620000000000002
Varianza: 0.030159183673469384
Desviación estándar: 0.17366399648018407

Columna 3
Promedio: 0.24599999999999997
Varianza: 0.01110612244897959
Desviación estándar: 0.10538558938004565
```

```
Categoría Iris-versicolor
Columna 0
Promedio: 5.936
Varianza: 0.2664326530612245
Desviación estándar: 0.5161711470638634

Columna 1
Promedio: 2.7700000000000005
Varianza: 0.09846938775510206
Desviación estándar: 0.3137983233784114

Columna 2
Promedio: 4.26
Varianza: 0.22081632653061228
Desviación estándar: 0.46991097723995795

Columna 3
Promedio: 1.3259999999999998
Varianza: 0.03910612244897959
Desviación estándar: 0.19775268000454405
```

```
Categoría Iris-virginica
Columna 0
Promedio: 6.587999999999998
Varianza: 0.4043428571428573
Desviación estándar: 0.6358795932744322

Columna 1
Promedio: 2.974
Varianza: 0.10400408163265305
Desviación estándar: 0.32249663817263746

Columna 2
Promedio: 5.552000000000005
Varianza: 0.30458775510204084
Desviación estándar: 0.5518946956639834

Columna 3
Promedio: 2.0260000000000002
Varianza: 0.07543265306122449
Desviación estándar: 0.2746500556366674
```

Pruebas

Hicimos los cálculos en una hoja de Excel para corroborar que los cálculos son correctos.

	Promedio	Varianza	Desv. Estándar
Col 0	5.843333333	0.685694	0.828066128
Col 1	3.057333333	0.189979	0.435866285
Col 2	3.758	3.116278	1.765298233
Col 3	1.199333333	0.581006	0.762237669

	Iris-Setosa		
	Promedio	**Varianza**	**Desv. Estándar**
Col 0	5.006	0.124249	0.352489687
Col 1	3.428	0.14369	0.379064369
Col 2	1.462	0.030159	0.173663996
Col 3	0.246	0.011106	0.105385589

	Iris-versicolor		
	Promedio	**Varianza**	**Desv. Estándar**
Col 0	5.936	0.266433	0.516171147
Col 1	2.77	0.098469	0.313798323
Col 2	4.26	0.220816	0.469910977
Col 3	1.326	0.039106	0.19775268

	Iris-virginica		
	Promedio	**Varianza**	**Desv. Estándar**
Col 0	6.588	0.404343	0.635879593
Col 1	2.974	0.104004	0.322496638
Col 2	5.552	0.304588	0.551894696
Col 3	2.026	0.075433	0.274650056