Primer examen parcial - INF 354

Jesus Rodolfo Izurieta Veliz

26 de septiembre de 2021

1. Introducción

1.1. Selección del dataset

El dataset seleccionado para la realización práctica de la prueba, muestra datos de mediciones de signos relacionadas a apoplejías o **accidente cerebrovascular**, con la finalidad de poder predecir, según estos indicadores, si una persona es propensa a sufrir este padecimiento.

Según la Organización Mundial de la Salud (OMS), el accidente cerebrovascular es la segunda causa principal de muerte a nivel mundial, responsable de aproximadamente el 11 % del total de muertes. Este conjunto de datos se utiliza para predecir si es probable que un paciente contraiga un accidente cerebrovascular en función de los parámetros de entrada como el sexo, la edad, varias enfermedades y el estado tabáquico. Cada fila de los datos proporciona información relevante sobre el paciente.

El dataset está públicamente disponible en la dirección: `https://www.kaggle.com/fedesoriano/stroke-prediction-dataset`

1.1.1. Información de atributos

1. `id`: identificador único
2. `gender`: «`Male`», «`Female`» u «`Other`»
3. `age`: edad del paciente
4. `hipertenssion`: 0 si el paciente no tiene hipertensión, 1 si el paciente tiene hipertensión
5. `heart_disease`: 0 si el paciente no tiene ninguna enfermedad cardíaca, 1 si el paciente tiene una enfermedad cardíaca

6. `ever_married`: «No» o «Yes»
7. `work_type`: «children», «Govt_jov», «Never_worked», «Private» o «Self-employed»
8. `Residence_type`: «Rural» o «Urban»
9. `avg_grucose_level`: nivel promedio de glucosa en la sangre
10. `bmi`: índice de masa corporal
11. `smoking_status`: «formerly smoked», «never smoked», «smokes» o «unknown»*
12. `stroke`: 1 si el paciente tuvo un accidente cerebrovascular o 0 si no

*Nota: «Unknown» en `smoking_status` significa que la información no está disponible para este paciente.

Figura 1: Sección de cerebro procedente de un difunto que sufrió un ataque cerebrovascular (ACV) a nivel de la arteria cerebral media.

2. Desarrollo

El desarrollo de las preguntas del examen se realizó en anaconda, este documento presenta parte del desarrollo, el código completo se encuentra en los repositorios específicos de cada pregunta.

2.1. Pregunta 1

Dirección del repositorio: `https://github.com/izurietajr/primer-parcial-354-1`

Seleccione un dataset de los propuestos por su persona en una anterior tarea, esta debe ser tabular de al menos 1000 filas y 5 columnas. Realice lo siguiente:

1. La media, moda y la desviación estándar por columna; explique qué significa en cada caso mediante Python sin uso de librerías
2. La media, la moda, la desviación estándar con el uso de numpy y pandas
3. Grafique los datos y explique su comportamiento (PYTHON)

2.1.1. Media, moda y desviación estándar (Python)

Cargado de los datos

Usaremos la función open de python para abrir el archivo del dataset en formato csv y crearemos una estructura orientada a objetos para facilitar el manejo de nuestro dataset.

```
class Dataset:
    def __init__(self, url):
        self.load_csv(url)

    def load_csv(self, url):
        content = []
        with open(url) as file:
            lines = file.readlines()
            for line in lines:
                line = line.replace("\n", "")
                content.append(line.split(","))
        self.headers, self.data = content[0], content[1:]
        return self.headers, self.data

    def get_column(self, column):
        index = self.headers.index(column)
        try:
            return [float(x[index]) for x in self.data]
        except Exception as e:
            print(e)
            return []

dataset = Dataset("stroke-dataset.csv")
```

```
print(dataset.headers)
print(dataset.data[:3])
```

Mostramos los nombres de los campos obtenidos del archivo csv y las primeras tres líneas de datos, la salida de la ejecución muestra:

```
['id', 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status',
'stroke']
[['9046', 'Male', '67', '0', '1', 'Yes', 'Private', 'Urban', '228.69',
'36.6', 'formerly smoked', '1'], ['51676', 'Female', '61', '0', '0', 'Yes',
'Self-employed', 'Rural', '202.21', 'N/A', 'never smoked', '1'], ['31112',
'Male', '80', '0', '1', 'Yes', 'Private', 'Rural', '105.92', '32.5', 'never
smoked', '1']]
```

Los nombres de los campos se nos muestran como una lista de cadenas, mientras que los datos se encuentran en una matriz (lista de listas) donde cada línea corresponde a una lista dentro de la lista `data`.

Cálculo de la media

La media muestral es calculada bajo la siguiente definición, considerando que se tiene una muestra $(X_1, X_2, ..., X_n)$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + ... + X_n}{n}$$

```
def media(lista):
    if lista != []:
        suma = sum(lista)
        return suma/len(lista)
    else:
        return None

age = dataset.get_column("age")
media(age)
```

Como resultado se muestra:

```
43.226614481409015
```

Cálculo de la moda

La moda es el valor más común en cuanto a frecuencia, por lo que para calcularla, obtendremos la frecuencia de cada ocurrencia y seleccionaremos la mayor.

```
from functools import reduce


def moda(lista):
    freq = {}
    for i in lista:
        if i not in freq:
            freq[i] = 1
        else:
            freq[i] = freq[i]+1
    items = [(r, s) for r, s in freq.items()]
    moda = reduce(lambda x, y: x if x[1] > y[1] else y, items)
    return moda[0]


moda(age)
```

Dando como resultado:

```
78.0
```

Cálculo de la desviación estándar

La desviación estándar de una población estadística, conjunto de datos o distribución de probabilidad es la raíz cuadrada de su varianza, que es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

```
from math import sqrt


def sd(lista):
    media_ = media(lista)
    smt = [(i-media_)**2 for i in lista]
    return sqrt(media(smt))


sd(age)
```

Dando como resultado:

```
22.61043402711301
```

2.1.2. Media, moda y desviación estándar (Numpy, Pandas)

Cargado de datos

Usaremos la librería pandas para leer el archivo csv y así poder obtener el dataset.

```
import pandas as pd
import numpy as np
from scipy import stats as st

data = pd.read_csv("stroke-dataset.csv")
print(data)
```

Nos mostrará parte del dataset:

```
          id  gender   age  hypertension  heart_disease ever_married  \
0       9046    Male  67.0             0              1          Yes
1      51676  Female  61.0             0              0          Yes
2      31112    Male  80.0             0              1          Yes
3      60182  Female  49.0             0              0          Yes
4       1665  Female  79.0             1              0          Yes
...      ...     ...   ...           ...            ...          ...
5105   18234  Female  80.0             1              0          Yes
5106   44873  Female  81.0             0              0          Yes
5107   19723  Female  35.0             0              0          Yes
5108   37544    Male  51.0             0              0          Yes
5109   44679  Female  44.0             0              0          Yes

          work_type Residence_type  avg_glucose_level   bmi   smoking_status  \
0           Private          Urban             228.69  36.6  formerly smoked
1     Self-employed          Rural             202.21   NaN     never smoked
2           Private          Rural             105.92  32.5     never smoked
3           Private          Urban             171.23  34.4           smokes
4     Self-employed          Rural             174.12  24.0     never smoked
```

```
...            ...            ...              ...   ...              ...
5105        Private          Urban            83.75   NaN     never smoked
5106  Self-employed          Urban           125.20  40.0     never smoked
5107  Self-employed          Rural            82.99  30.6     never smoked
5108        Private          Rural           166.29  25.6  formerly smoked
5109       Govt_job          Urban            85.28  26.2          Unknown

      stroke
0          1
1          1
2          1
3          1
4          1
...      ...
5105       0
5106       0
5107       0
5108       0
5109       0

[5110 rows x 12 columns]
```

Cálculo de la media, moda y desviación estándar

Para calcular estos estadísticos, usaremos la librería numpy.

```
np.mean(data['age'])
```

salida:

```
43.226614481409015
```

Ya que numpy no provee un método para calcular la moda, usaremos la librería scipy, importaremos el módulo stats de esta, que provee la función mode.

```
st.mode(data['age'])
```

salida:

```
ModeResult(mode=array([78.]), count=array([102]))
```

Desviación estándar:

```
np.std(data['age'])
```

salida:

```
22.61043402711301
```

Como podemos ver, los valores calculados con las librerías de python, coinciden con los valores calculados previamente.

2.1.3. Gráficas e interpretación de datos

Usaremos la librería matplotlib para hacer gráficos con los datos del dataset.

```
import matplotlib.pyplot as plt

strokes = data[data.stroke == 1]
strokes.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

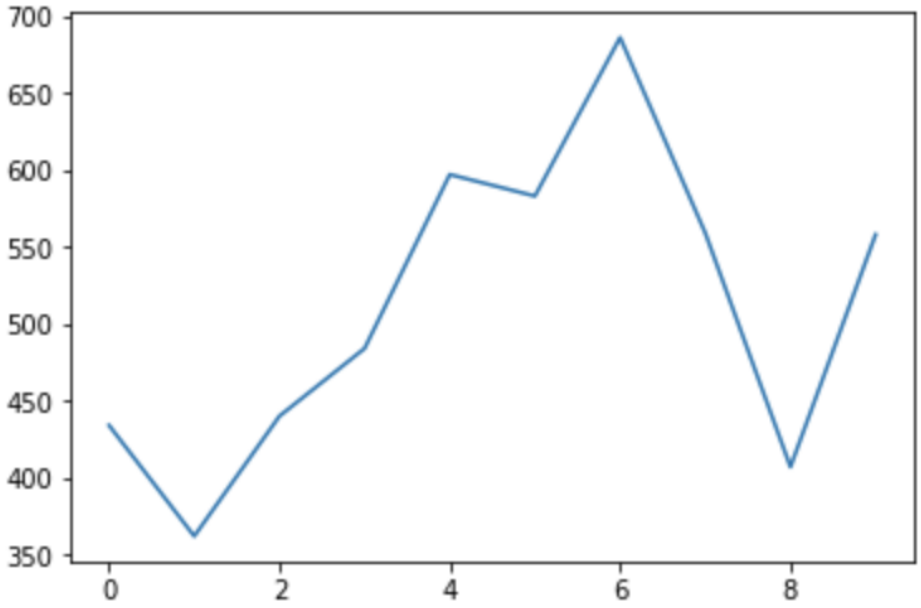
Realizamos el gráfico de un histograma de la siguiente manera:

```
hist, edges = np.histogram(data['age'])

plt.plot(hist)
```

Con esto, obtenemos el gráfico:

```
[<matplotlib.lines.Line2D at 0x7fd3001bf0b8>]
```



2.2. Pregunta 2

Dirección del repositorio: `https://github.com/izurietajr/primer-parcial-354-2`

3. Referencias

`https://www.kaggle.com/fedesoriano/stroke-prediction-dataset`