S05_T01

April 26, 2022

0.0.1 Nivell 1

- Exercici 1

Agafa un conjunt de dades de tema esportiu que t'agradi. Realitza un mostreig de les dades generant una mostra aleatòria simple i una mostra sistemàtica.

[23]:
```
import csv
import numpy as np
import pandas as pd
import sklearn as sk

players = pd.read_csv('C:
↪\\Users\\motxi\\Documents\\Data_Science_IT_Academy\\Sprints\\S05\\players.
↪csv',
                      engine = 'python',
                      header = 0,
                      )
```

[24]:
```
players.head()
```

[24]:
```
   Unnamed: 0          Player  height  weight  \
0           0  Curly Armstrong   180.0    77.0
1           1     Cliff Barker   188.0    83.0
2           2    Leo Barnhorst   193.0    86.0
3           3       Ed Bartels   196.0    88.0
4           4      Ralph Beard   178.0    79.0

                          collage    born   birth_city birth_state
0              Indiana University  1918.0          NaN         NaN
1           University of Kentucky  1921.0     Yorktown     Indiana
2         University of Notre Dame  1924.0          NaN         NaN
3  North Carolina State University  1925.0          NaN         NaN
4           University of Kentucky  1927.0  Hardinsburg    Kentucky
```

[25]:
```
players = players.filter(['Player', 'height' ,'weight' ,'collage'␣
↪,'born','birth_city' ,'birth_state'], axis=1)
```

[26]:
```
players.head()
```

[26]:
```
            Player  height  weight                          collage    born  \
0  Curly Armstrong   180.0    77.0              Indiana University  1918.0
1     Cliff Barker   188.0    83.0           University of Kentucky  1921.0
2    Leo Barnhorst   193.0    86.0         University of Notre Dame  1924.0
3       Ed Bartels   196.0    88.0  North Carolina State University  1925.0
4      Ralph Beard   178.0    79.0           University of Kentucky  1927.0

    birth_city birth_state
0          NaN         NaN
1     Yorktown     Indiana
2          NaN         NaN
3          NaN         NaN
4  Hardinsburg    Kentucky
```

[27]:
```
# Simple Random Sampling

sample_df = players.sample(10)
sample_df
```

[27]:
```
                Player  height  weight                          collage  \
181        Fred Schaus   196.0    92.0           West Virginia University
89          Tony Jaros   190.0    83.0            University of Minnesota
1757     Alex Stivrins   203.0    99.0             University of Colorado
2364   Mark Strickland   206.0    95.0                  Temple University
3360     Lester Hudson   190.0    86.0  University of Tennessee at Martin
902     Charlie Lowery   190.0    83.0          University of Puget Sound
2907   Carmelo Anthony   203.0   108.0                Syracuse University
733           Jim Reid   198.0    95.0     Winston-Salem State University
1055        Don Smith   188.0    86.0            University of Minnesota
3341  Dante Cunningham   203.0   104.0               Villanova University

        born   birth_city birth_state
181   1925.0       Newark        Ohio
89    1920.0  Minneapolis   Minnesota
1757  1962.0      Lincoln    Nebraska
2364  1970.0      Atlanta     Georgia
3360  1984.0      Memphis   Tennessee
902   1949.0          NaN         NaN
2907  1984.0     New York    New York
733   1945.0          NaN         NaN
1055  1920.0    Minnesota         NaN
3341  1987.0      Clinton    Maryland
```

[28]:
```
# Systematic (Lineal) Sampling

def systematic_sampling(df, step):
    indexes = np.arange(0, len(df), step=step)
```

```
    systematic_sample = df.iloc[indexes]
    return systematic_sample



systematic_sampling(players, 500)
```

[28]:

	Player	height	weight	collage	born	\
0	Curly Armstrong	180.0	77.0	Indiana University	1918.0	
500	Maury King	188.0	88.0	University of Kansas	1935.0	
1000	Ben Kelso	190.0	88.0	Central Michigan University	1949.0	
1500	Lewis Lloyd	198.0	92.0	Drake University	1959.0	
2000	Nate Johnston	203.0	95.0	University of Tampa	1966.0	
2500	Malik Rose	201.0	113.0	Drexel University	1974.0	
3000	Dwight Howard	211.0	120.0	NaN	1985.0	
3500	Markieff Morris	188.0	88.0	Northwestern University	1925.0	

	birth_city	birth_state
0	NaN	NaN
500	NaN	NaN
1000	NaN	NaN
1500	Philadelphia	Pennsylvania
2000	Birmingham	Alabama
2500	Philadelphia	Pennsylvania
3000	Atlanta	Georgia
3500	NaN	NaN

0.0.2 - Exercici 2

Continua amb el conjunt de dades de tema esportiu i genera una mostra estratificada i una mostra utilitzant SMOTE (Synthetic Minority Oversampling Technique).

[ ]:
```
# Stratified Sampling

X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y,␣
 ↪test_size=0.1)
```