Taller 8

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 18-oct-2019 11:59 PM

Julián Santiago Ramírez

julians.ramirez@urosario.edu.co

Instrucciones:

- Guarde una copia de este *Jupyter Notebook* en su computador, idealmente en una carpeta destinada al material del curso.
- Modifique el nombre del archivo del *notebook*, agregando al final un guión inferior y su nombre y apellido, separados estos últimos por otro guión inferior. Por ejemplo, mi *notebook* se llamaría: mcpp_taller8_santiago_matallana
- Marque el *notebook* con su nombre y e-mail en el bloque verde arriba. Reemplace el texto "[Su nombre acá]" con su nombre y apellido. Similar para su e-mail.
- Desarrolle la totalidad del taller sobre este *notebook*, insertando las celdas que sea necesario debajo de cada pregunta. Haga buen uso de las celdas para código y de las celdas tipo *markdown* según el caso.
- Recuerde salvar periódicamente sus avances.
- Cuando termine el taller:
  1. Descárguelo en PDF. Si tiene algún problema con la conversión, descárguelo en HTML.
  2. Suba todos los archivos a su repositorio en GitHub, en una carpeta destinada exclusivamente para este taller, antes de la fecha y hora límites.

1. [1 punto]

Usando expresiones regulares extraiga en una lista todos los números presentes en el siguiente objeto de Python:

ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his wife, MacKenzie. Mr Bezos will keep all the shares in the Washington Post and Blue Origin, a space-exploration firm, as well as 75% of the couple's Amazon stock. Mrs Bezos will retain a 4% stake in the tech giant, worth nearly $36bn, which is likely to make her the third-richest woman alive when the divorce is finalised."

In [1]:
```
#### Importamos la librería que nos permite usar expresiones regulares ####
import re
```

In [3]:
```
#### variable ob1
ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his wife, MacKenzie. Mr Bezos will keep all the
shares in the Washington Post and Blue Origin, a space-exploration firm, as well as 75% of the couple's Amazon stock. Mrs Bezos
will retain a 4% stake in the tech giant, worth nearly $36bn, which is likely to make her the third-richest woman alive when the
divorce is finalised."

numeros=re.findall('\d+',ob1)

print("Números en el texto: ",numeros)
```

```
Números en el texto:  ['75', '4', '36']
```

2. [1 punto]

Usando expresiones regulares ahora extraiga de *ob1* sólo los números que correspondan a porcentajes.

In [5]:
```
porcentajes=re.findall('\d+%',ob1)

print("Números con porcentaje en el texto: ",porcentajes)
```

```
Números con porcentaje en el texto:  ['75%', '4%']
```

3. [2 puntos]

Usando expresiones regulares, escriba una función de Python que reciba una fecha en formato **Marzo 7, 2019** y retorne la fecha en formato **2019-07-03**

In [18]:
```
def conversion_fecha(fecha):
    def_fecha = []
    meses = ['Enero','Febrero','Marzo','Abril','Mayo','Junio','Julio',
             'Agosto','Septiembre','Octubre','Noviembre','Diciembre']
    ### ahora de la fecha queremos obtener el dia y el año
    # Primero buscamos los numeros
    numeros =re.search('([\d]+), ([\d]+)',fecha)
    fec=numeros.groups()
    def_fecha.append(fec[1])
    ## El primer dato de fec es el dia y el segundo es el año, por lo tanto preguntamos si el dia es igual a 23
    if fec[0] == '23':
        dia_actualizado = '23'
        def_fecha.append(dia_actualizado)

    ## Hasta el momento hemos agregado el año y el dia, falta el mes
    mes=0
    for i in range(0,len(meses)):
        if meses[i]=="Octubre":
            mes=i+1
            break
    ## Agregamos el mes a la fecha definitiva
    def_fecha.append(mes)
    total=str(def_fecha[0])+'-'+str(def_fecha[1])+'-'+str(def_fecha[2])
    return total

## variable que guarda la fecha ###
fecha = 'Octubre 23, 2019'
nueva = conversion_fecha(fecha)

print("Antiguo fórmato: ", fecha)
print("Nuevo fórmato: ",nueva)
```

```
Antiguo fórmato:  Octubre 23, 2019
Nuevo fórmato:  2019-23-10
```

4. [3 puntos]

ob2 es un string que reune una lista de clases en una universidad. Use expresiones regulares para extraer los códigos de cada una de las clases. Ejemplo: El código de la clase **COMPSCI 143 (Spring 2012): Machine Learning** es 143.

ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 590.2 (Fall 2018): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 223 (Spring 2018): Computational Microeconomics. COMPSCI 570 (Fall 2017): Artificial Intelligence. COMPSCI 590.3 (Fall 2017) / 590.1 (Spring 2018): Ethics and AI. COMPSCI 590.2 (Spring 2017): Computation, Information, and Learning in Market Design. COMPSCI 590.4 (Spring 2016): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 290.4/590.4 (Spring 2015): Crowdsourcing Societal Tradeoffs. COMPSCI 570 (Fall 2014): Artificial Intelligence. COMPSCI 590.4 (Spring 2014): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 590.1 (Fall 2012): Linear and Integer Programming. COMPSCI 173 (Spring 2012): Computational Microeconomics. COMPSCI 296.1 (Fall 2011): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 296.1 (Fall 2010): Linear and Integer Programming. COMPSCI 173 (Spring 2010): Computational Microeconomics. COMPSCI 196.1/296.1 (Fall 2009): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 170 (Spring 2009): Introduction to Artificial Intelligence. COMPSCI 270 (Fall 2008): Artificial Intelligence. COMPSCI 196/296.2 (Spring 2008): Linear and Integer Programming. COMPSCI 196.2 (Fall 2007): Introduction to Computational Economics. COMPSCI 296.3 (Spring 2007): Topics in Computational Economics. COMPSCI 296.2 (Fall 2006): Computational Game Theory and Mechanism Design."

In [ ]:
```
## Guardamos ob2 ##
ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 590.2 (Fall 2018): Computational Microeconomi
cs: Game Theory, Social Choice, and Mechanism Design. COMPSCI 223 (Spring 2018): Computational Microeconomics. COMPSCI 570 (Fall
2017): Artificial Intelligence. COMPSCI 590.3 (Fall 2017) / 590.1 (Spring 2018): Ethics and AI. COMPSCI 590.2 (Spring 2017): Com
putation, Information, and Learning in Market Design. COMPSCI 590.4 (Spring 2016): Computational Microeconomics: Game Theory, So
cial Choice, and Mechanism Design. COMPSCI 290.4/590.4 (Spring 2015): Crowdsourcing Societal Tradeoffs. COMPSCI 570 (Fall 2014):
Artificial Intelligence. COMPSCI 590.4 (Spring 2014): Computational Microeconomics: Game Theory, Social Choice, and Mechanism De
sign. COMPSCI 590.1 (Fall 2012): Linear and Integer Programming. COMPSCI 173 (Spring 2012): Computational Microeconomics. COMPSC
I 296.1 (Fall 2011): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 296.1 (Fall 2010):
Linear and Integer Programming. COMPSCI 173 (Spring 2010): Computational Microeconomics. COMPSCI 196.1/296.1 (Fall 2009): Comput
ational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 170 (Spring 2009): Introduction to Artificial
Intelligence. COMPSCI 270 (Fall 2008): Artificial Intelligence. COMPSCI 196/296.2 (Spring 2008): Linear and Integer Programming.
COMPSCI 196.2 (Fall 2007): Introduction to Computational Economics. COMPSCI 296.3 (Spring 2007): Topics in Computational Economi
cs. COMPSCI 296.2 (Fall 2006): Computational Game Theory and Mechanism Design."
```

In [101]:
```
## Buscamos los números que acompañan la palabra COMPSCI

## Observemos que hay dos clases de codigo: el primero un numero
#normal y el segundo se puede ver de esta forma: 290.4/590.4

## Codigos tipo 1
codigo_tipo1=re.findall(' (\d+.\d+) ',ob2)

## Codigos tipo 2
codigo_tipo2=re.findall(' (\d+.\d+/\d+.\d+) ',ob2)

## Supongo que los codigos tipo 2 son dos clases diferentes y por ende debo separarlos
nuevo_codigo=[]

for cod in codigo_tipo2:
    tupla=re.search('(\d+.\d+)/(\d+.\d+)',cod).groups()
    # agregamos ambos codigos
    nuevo_codigo.append(tupla[0])
    nuevo_codigo.append(tupla[1])

total=codigo_tipo1+nuevo_codigo

print("Códigos:", total)
#Cod=re.findall('/ (\d+.\d+)',ob2)
```

```
Códigos: ['270', '590.2', '223', '570', '590.3', '590.1', '590.2', '590.4', '570', '590.4', '590.1', '173', '296.1', '296.1',
'173', '170', '270', '196.2', '296.3', '296.2', '290.4', '590.4', '196.1', '296.1', '196', '296.2']
```

5. [5 puntos]

ob3 es un string que reune una lista de publicaciones. Use expresiones regulares para extraer todos los *Journals* en los cuales el autor ha publicado. Ejemplo: El paper **Bail, CA. "The configuration of symbolic boundaries against immigrants in Europe." American Sociological Review 73.1 (January 1, 2008): 37-59. Full Text** fue publicado en el Journal *American Sociological Review*

ob3 = "Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H, Hunzaker, MBF, Lee, J, Mann, M, Merhout, F, and Volfovsky, A. \"Exposure to opposing views on social media can increase political polarization.\" Proceedings of the National Academy of Sciences of the United States of America 115.37 (September 2018): 9216-9221. Full Text Open Access Copy.\n", "Bail, CA, Merhout, F, and Ding, P. \"Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in U.S. counties.\" Science Advances 4.6 (June 6, 2018): eaao5948-null. Full Text Open Access Copy.\n", "Bail, CA, Brown, TW, and Mann, M. \"Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation.\" American Sociological Review 82.6 (December 1, 2017): 1188-1213. Full Text.\n", "Bail, CA. \"Taming Big Data: Using App Technology to Study Organizational Behavior on Social Media.\" Sociological Methods and Research 46.2 (March 1, 2017): 189-217. Full Text.\n", "McDonnell, TE, Bail, CA, and Tavory, I. \"A Theory of Resonance.\" Sociological Theory 35.1 (March 1, 2017): 1-14. Full Text.\n", "Bail, CA. \"Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media.\" Proceedings of the National Academy of Sciences of the United States of America 113.42 (October 2016): 11823-11828. Full Text.\n", "Bail, CA. \"Emotional Feedback and the Viral Spread of Social Media Messages About Autism Spectrum Disorders.\" American journal of public health 106.7 (July 2016): 1173-1180. Full Text.\n", "Bail, CA. \"The public life of secrets: Deception, disclosure, and discursive framing in the policy process.\" Sociological Theory 33.2 (January 1, 2015): 97-124. Full Text.\n", "Bail, CA. \"The cultural environment: Measuring culture with big data.\" Theory and Society 43.3 (January 1, 2014): 465-524. Full Text.\""

In [57]:
```
## Variable ob3
ob3 = '"Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H, Hunzaker, MBF, Lee, J, Mann, M, Merhout, F, and Volfovsky, A. "Exp
osure to opposing views on social media can increase political polarization." Proceedings of the National Academy of Sciences of
the United States of America 115.37 (September 2018): 9216-9221. Full Text Open Access Copy.\n", "Bail, CA, Merhout, F, and Din
g, P. "Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in U.S. counties." Scie
nce Advances 4.6 (June 6, 2018): eaao5948-null. Full Text Open Access Copy.\n", "Bail, CA, Brown, TW, and Mann, M. "Channeling H
earts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation." American Sociological Review 8
2.6 (December 1, 2017): 1188-1213. Full Text.\n", "Bail, CA. "Taming Big Data: Using App Technology to Study Organizational Beha
vior on Social Media." Sociological Methods and Research 46.2 (March 1, 2017): 189-217. Full Text.\n", "McDonnell, TE, Bail, CA,
and Tavory, I. "A Theory of Resonance." Sociological Theory 35.1 (March 1, 2017): 1-14. Full Text.\n", "Bail, CA. "Combining nat
ural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media." Pro
ceedings of the National Academy of Sciences of the United States of America 113.42 (October 2016): 11823-11828. Full Text.\n",
"Bail, CA. "Emotional Feedback and the Viral Spread of Social Media Messages About Autism Spectrum Disorders." American journal
of public health 106.7 (July 2016): 1173-1180. Full Text.\n", "Bail, CA. "The public life of secrets: Deception, disclosure, and
discursive framing in the policy process." Sociological Theory 33.2 (January 1, 2015): 97-124. Full Text.\n", "Bail, CA. "The cu
ltural environment: Measuring culture with big data." Theory and Society 43.3 (January 1, 2014): 465-524. Full Text."'
## Guardaremos todos los journals en esta lista
journals = []

# Buscamos en el texto frases: palabra + fecha
frases = re.findall('" \w+ .+ \d',ob3)

for i in range(0,len(frases)):
    jour = re.search('\w+\D+',frases[i]).group()
    journals.append(jour)
# Modificamos el primer journal
journals[0] = 'National Academy of Sciences of the United States of America'
borrar=[]
for i in range(0,len(journals)):
    if journals[i]=='Sociological Theory ':
        borrar.append(i)
del journals[borrar[0]]
del journals[borrar[1]-1]

print("Journals: ",journals)
```

```
Journals:  ['National Academy of Sciences of the United States of America', 'Science Advances ', 'American Sociological Review
', 'Sociological Methods and Research ', 'Proceedings of the National Academy of Sciences of the United States of America ', 'A
merican journal of public health ', 'Theory and Society ']
```

6. [10 puntos]

Vamos a hacer "scraping" a esta página: https://archive.ics.uci.edu/ml/datasets.php, que contiene un listado de 468 bases de datos que hacen parte del repositorio de la Universidad de California, Irvine.

Su tarea consiste en crear un "Pandas dataframe" que contenga 468 filas (una por base de datos) y las siguientes columnas:

- Nombre de la base de datos
- Link a la base de datos
- Tipo de datos
- Tipo de tarea a resolver (default task)
- Tipo de las variables
- Número de observaciones
- Número de variables
- Año
- Descripción de la base (Pista: Utilice la opción list view: https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=list)

In [94]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

In [60]:
```
# Info pag
link = requests.get('https://archive.ics.uci.edu/ml/datasets.php').text
l = BeautifulSoup(link, "lxml")
```

In [85]:
```
### Buscamos primero todas las lineas que inician con <b>
lineas = l.find_all('b')
str_lineas=str(lineas)

# En esas lineas buscamos unas con una caracteristica especifica
especificas= re.findall('(<b><a href="datasets/\w+.+?</a>)', str_lineas)


# Crearemos una lista con todos los nombres
nombres = []
for i in range(0,len(especificas)):
    nom = re.search('>(\w+.+?)[^<]',especificas[i]).group()
    nombres.append(nom)
# Ahora de los nombres los quitamos el </a>
for i in range(0,len(especificas)):
    nom = re.search('[^>].+[^<</a>]',nombres[i]).group()
    nombres[i]=nom

#print("Variables: ", nombres)
```

In [84]:
```
## datos caracteristicas
# lineas que inician con <p>
carac = l.find_all('p')
# Caracteristica de los datos, falta el ultimo
caracteristicas=re.findall('(href="datasets/.+?<b>)',str(carac))+['href="datasets/Rice+Leaf+Diseases">Rice Leaf Diseases</a></b>
</p>, <p class="normal">Multivariate\xa0</p>, <p class="normal">Classification\xa0</p>, <p class="normal">Integer\xa0</p>, <p cl
ass="normal">120\xa0</p>, <p class="normal">\xa0</p>, <p class="normal">2019\xa0</p>']

## Tipos de datos
## Guardare todo en una lista de listas

tipos=[[],[],[],[],[],[]]
for i in range(0,len(especificas)):
    carac_es = re.search('<p.+',caracteristicas[i]).group()
    tipo =re.findall('(<p.+?</p>)',carac_es)
    for j in range(0,6):
        pal1=re.search('(>.+?[\xa0]|[\xa0])',tipo[j]).group()
        pal2=re.search('[^>].+[^\xa0]|[\w+|\xa0]',pal1).group()
        tipos[j].append(pal2)
# \xa0 indica missing
# Cambiados por NA
for j in tipos:
    for i in range(0,len(especificas)):
        if j[i] == '\xa0':
            j[i]= 'NA'
```

In [90]:
```
# LINKS

l1 = []
for i in range(0,len(especificas)):
    pal1 = re.search('"(datasets/.+?)"',especificas[i]).group()
    pal2 = re.search('[^"].+[^"]',pal1).group()
    l1.append(pal2)
l2 = []
for i in range(0,len(l1)):
    pal = 'https://archive.ics.uci.edu/ml/'+l1[i]
    l2.append(pal)

# Links parte III
l3 = []
for i in range(0,len(l1)):
    html = requests.get(l2[i]).text
    s_html = BeautifulSoup(html, "lxml")
    pal_a = s_html.find_all('a')
    html_nue = re.findall('<a( href="../machine-learning-databases/.+?")>',str(pal_a))
    if html_nue == []:
        l3.append('NA')
    else:
        l3.append(html_nue[0])
```

In [92]:
```
l_def = []
for i in range(0,len(especificas)):
    if l3[i] == 'NA':
        l_def.append('NA')
    else:
        pal1 = re.search('/.+[^">]',l3[i]).group()
        pal2 = 'https://archive.ics.uci.edu/ml'+pal1
        l_def.append(pal2)
```

In [99]:
```
tabla = {"Base de datos": nombres,
    "Tipo de datos": tipos[0],
    "Tipo de tarea a resolver": tipos[1],
    "Tipo de variables": tipos[2],
    "Número de obs": tipos[3],
    "Número de variables": tipos[4],
    "Año": tipos[5],
    "Links": l_def}
definitiva = pd.DataFrame(tabla)

# Ordenando el data frame

definitiva["nombres mayu"] = definitiva["Base de datos"].str.upper()
definitiva.sort_values(["nombres mayu"], axis=0,
                 ascending=[True], inplace=True)
# Borramos la variable creado del data frame
del definitiva["nombres mayu"]
```

In [100]:
```
# Descripción de la base
link = requests.get('https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view
=list').text

s_link = BeautifulSoup(link, "lxml")

des = re.findall('<b><a href="datasets[\s\S]*?</p>',str(s_link))

descripcion = []
for i in range(0,len(des)):
    try:
        pal1 = re.search('</b>([\s\S]*?)</p>',des[i]).group()
        pal2 = re.search('[^</b> ][\s\S]*[^</p> ]',pal1).group()
        descripcion.append(re.sub('\r\n', '',pal2))
    except AttributeError:
        descripcion.append('NA')

definitiva['Descripción base datos'] = descripcion

# Resultado
definitiva
```

Out[100]:

	Base de datos	Tipo de datos	Tipo de tarea a resolver	Tipo de variables	Número de obs	Número de variables	Año	Links	Descripción base datos
462	2.4 GHZ Indoor Channel Measurements	Multivariate	Classification	Real	7840	5	2018	https://archive.ics.uci.edu/ml/machine-learnin...	Measurement of th S21,consists of 10 sweeps, ...
237	3D Road Network (North Jutland, Denmark)	Sequential, Text	Regression, Clustering	Real	434874	4	2013	https://archive.ics.uci.edu/ml/machine-learnin...	3D road network with highly accurat elevation...
301	AAAI 2013 Accepted Papers	Multivariate	Clustering	NA	150	5	2014	https://archive.ics.uci.edu/ml/machine-learnin...	This data set compromises the metadata for the...
294	AAAI 2014 Accepted Papers	Multivariate	Clustering	NA	399	6	2014	https://archive.ics.uci.edu/ml/machine-learnin...	This data set compromises the metadata for the...
0	Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995	https://archive.ics.uci.edu/ml/machine-learnin...	Predict the age of abalone from physical measu...
170	Abscisic Acid Signaling Network	Multivariate	Causal-Discovery	Integer	300	43	2008	https://archive.ics.uci.edu/ml/machine-learnin...	The objective is to determine the set o boole...
427	Absenteeism at work	Multivariate, Time-Series	Classification, Clustering	Integer, Real	740	21	2018	https://archive.ics.uci.edu/ml/machine-learnin...	The database was created with records of absen...
260	Activities of Daily Living (ADLs) Recognition ...	Multivariate, Sequential, Time-Series	Classification, Clustering	NA	2747	NA	2013	https://archive.ics.uci.edu/ml/machine-learnin...	This dataset comprises information regarding t...
275	Activity Recognition from Single Chest-Mounted...	Univariate, Sequential, Time-Series	Classification, Clustering	Real	NA	NA	2014	https://archive.ics.uci.edu/ml/machine-learnin...	The dataset collect data from a wearable acce...
351	Activity Recognition system based on Multisens...	Multivariate, Sequential, Time-Series	Classification	Real	42240	6	2016	https://archive.ics.uci.edu/ml/machine-learnin...	This dataset contains temporal data from a Wir...
411	Activity recognition with healthy older people...	Sequential	Classification	Real	75128	9	2016	https://archive.ics.uci.edu/ml/machine-learnin...	Sequential motion data from 14 healthy older p...
181	Acute Inflammations	Multivariate	Classification	Categorical, Integer	120	6	2009	https://archive.ics.uci.edu/ml/machine-learnin...	The data was created by a medical expert as a ...
1	Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996	https://archive.ics.uci.edu/ml/machine-learnin...	Predict whether income exceeds $50K/yr based o...
371	Air quality	Multivariate, Time-Series	Regression	Real	9358	15	2016	https://archive.ics.uci.edu/ml/machine-learnin...	Contains the responses of a gas multisensor de...
345	Air Quality	Multivariate, Time-Series	Regression	Real	9358	15	2016	https://archive.ics.uci.edu/ml/machine-learnin...	Contains the responses of a gas multisensor de...
279	Airfoil Self-Noise	Multivariate	Regression	Real	1503	6	2014	https://archive.ics.uci.edu/ml/machine-learnin...	NASA data set, obtained from a series of aerod...
476	Alcohol QCM Sensor Dataset	Multivariate	Classification, Regression, Clustering	Real	125	8	2019	https://archive.ics.uci.edu/ml/machine-learnin...	Five different QCM gas sensors are used, and f...
208	Amazon Access Samples	Time-Series, Domain-Theory	Regression, Clustering, Causal-Discovery	NA	30000	20000	2011	https://archive.ics.uci.edu/ml/machine-learnin...	Amazon's InfoSec is getting smarter about the ...
207	Amazon Commerce reviews set	Multivariate, Text, Domain-Theory	Classification	Real	1500	10000	2011	https://archive.ics.uci.edu/ml/machine-learnin...	The dataset is use for authorship identificat...
2	Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	NA	https://archive.ics.uci.edu/ml/machine-learnin...	Steel annealing data
3	Anonymous Microsoft Web Dat	NA	Recommender-Systems	Categorical	37711	294	1998	https://archive.ics.uci.edu/ml/machine-learnin...	Log of anonymous users of www.microsoft.com p...
390	Anuran Calls (MFCCs)	Multivariate	Classification, Clustering	Real	7195	22	2017	https://archive.ics.uci.edu/ml/machine-learnin...	Acoustic features extracted from syllables of ...
358	Appliances energy prediction	Multivariate, Time-Series	Regression	Real	19735	29	2017	https://archive.ics.uci.edu/ml/machine-learnin...	Experimental data used to create regression mo...
405	APS Failure at Scania Trucks	Multivariate	Classification	Integer, Real	60000	171	2017	https://archive.ics.uci.edu/ml/machine-learnin...	The datasets' positive class consists of compo...
164	Arcene	Multivariate	Classification	Real	900	10000	2008	https://archive.ics.uci.edu/ml/machine-learnin...	ARCENE's task is to distinguish cancer versus ...
4	Arrhythmi	Multivariate	Classification	Categorical, Integer, Real	452	279	1998	https://archive.ics.uci.edu/ml/machine-learnin...	Distinguish betwee the presence and absence o...
5	Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992	https://archive.ics.uci.edu/ml/machine-learnin...	Dataset artificially generated by using first ...
6	Audiology (Original)	Multivariate	Classification	Categorical	226	NA	1987	https://archive.ics.uci.edu/ml/machine-learnin...	Nominal audiology dataset from Baylo
7	Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992	https://archive.ics.uci.edu/ml/machine-learnin...	Standardized version of the original audiology...
457	Audit Dat	Multivariate	Classification	Real	777	18	2018	https://archive.ics.uci.edu/ml/machine-learnin...	Exhaustive one year non-confidential data in t...
...	...	...	...	...	...	...	...	...	...
248	User Knowledge Modeling	Multivariate	Classification, Clustering	Integer	403	5	2013	https://archive.ics.uci.edu/ml/machine-learnin...	It is the real datase about the students kno...
257	USPTO Algorithm Challenge, run by NASA-Harvard...	Domain-Theory	Classification	Integer	306	5	2013	https://archive.ics.uci.edu/ml/machine-learnin...	Data used for USPTO Algorithm Competition. Con...
204	Vertebral Column	Multivariate	Classification	Real	310	6	2011	https://archive.ics.uci.edu/ml/machine-learnin...	Data set containing values for six biomechanic...
206	Vicon Physical Action Data Set	Time-Series	Classification	Real	3000	27	2011	https://archive.ics.uci.edu/ml/machine-learnin...	The Physical Actio Data Set includes 10 norma...
436	Victorian Era Authorship Attribution	Text	Classification	NA	93600	1000	2018	https://archive.ics.uci.edu/ml/machine-learnin...	To create the largest authorship attribution d...
139	Volcanoes on Venus - JARtool experiment	Image	Classification	NA	NA	NA	NA	https://archive.ics.uci.edu/ml/machine-learnin...	The JARtool projec was a pioneering effort to...
190	Wall-Following Robot Navigation Dat	Multivariate, Sequential	Classification	Real	5456	24	2010	https://archive.ics.uci.edu/ml/machine-learnin...	The data were collected as the SCITOS G5 robot...
103	Water Treatment Plant	Multivariate	Clustering	Integer, Real	527	38	1993	https://archive.ics.uci.edu/ml/machine-learnin...	Multiple classes predict plant state
474	Wave Energy Converters	Multivariate	Regression	Real	288000	49	2019	https://archive.ics.uci.edu/ml/machine-learnin...	This data set consists of positio and absorb...
104	Waveform Database Generator (Version 1)	Multivariate, Data-Generator	Classification	Real	5000	21	1988	https://archive.ics.uci.edu/ml/machine-learnin...	CART book's waveform domains
105	Waveform Database Generator (Version 2)	Multivariate, Data-Generator	Classification	Real	5000	40	1988	https://archive.ics.uci.edu/ml/machine-learnin...	CART book's waveform domains
241	Wearable Computing: Classification of Body Pos...	Sequential	Classification	Integer, Real	165632	18	2013	https://archive.ics.uci.edu/ml/machine-learnin...	A dataset with 5 classes (sitting-down, standi...
363	Website Phishing	Multivariate	Classification	Integer	1353	10	2016	https://archive.ics.uci.edu/ml/machine-learnin...	
262	Weight Lifting Exercises monitored with Inerti...	Multivariate	Classification	Real	39242	152	2013	https://archive.ics.uci.edu/ml/machine-learnin...	Six young health subjects were asked to perfor...
447	WESAD (Wearable Stress and Affect Detection)	Multivariate, Time-Series	Classification, Regression	Real	63000000	12	2018	https://archive.ics.uci.edu/ml/machine-learnin...	WESAD (Wearable Stress and Affect Detection) c...
280	Wholesale customers	Multivariate	Classification, Clustering	Integer	440	8	2014	https://archive.ics.uci.edu/ml/machine-learnin...	The data set refers to clients of a wholesale ...
321	wiki4HE	Multivariate	Regression, Clustering, Causal-Discovery	NA	913	53	2015	https://archive.ics.uci.edu/ml/machine-learnin...	Survey of faculty members from two Spanish uni...
273	Wilt	Multivariate	Classification	NA	4889	6	2014	https://archive.ics.uci.edu/ml/machine-learnin...	High-resolution Remote Sensing data set (Quick...
106	Wine	Multivariate	Classification	Integer, Real	178	13	1991	https://archive.ics.uci.edu/ml/machine-learnin...	Using chemical analysis determine the origin o...
182	Wine Quality	Multivariate	Classification, Regression	Real	4898	12	2009	https://archive.ics.uci.edu/ml/machine-learnin...	Two datasets are included, related to red and ...
406	Wireless Indoor Localization	Multivariate	Classification	Real	2000	7	2017	https://archive.ics.uci.edu/ml/machine-learnin...	Collected in indoor space by observing signal ...
487	WISDM Smartphone and Smartwatch Activity and B...	Multivariate	Classification	Integer	120	NA	2019	https://archive.ics.uci.edu/ml/machine-learnin...	Contains accelerometer and gyroscope time-seri...
234	Yacht Hydrodynamics	Multivariate	Regression	Real	308	7	2013	https://archive.ics.uci.edu/ml/machine-learnin...	Delft data set, used to predict the hydodynami...
196	YearPredictionMSD	Multivariate	Regression	Real	515345	90	2011	https://archive.ics.uci.edu/ml/machine-learnin...	Prediction of the release year of a song from ...
107	Yeast	Multivariate	Classification	Real	1484	8	1996	https://archive.ics.uci.edu/ml/machine-learnin...	Predicting the Cellular Localizatio Sites of ...
215	YouTube Comedy Slam Preference Dat	Text	Classification	NA	1138562	3	2012	https://archive.ics.uci.edu/ml/machine-learnin...	This dataset provides user vote data on which ...
258	YouTube Multiview Video Games Dataset	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013	https://archive.ics.uci.edu/ml/machine-learnin...	This dataset contains about 120 instances, ea...
364	YouTube Spam Collection	Text	Classification	NA	1956	5	2017	https://archive.ics.uci.edu/ml/machine-learnin...	It is a public set of comments collecte for s...
396	Z-Alizadeh Sani	NA	Classification	Integer, Real	303	56	2017	https://archive.ics.uci.edu/ml/machine-learnin...	It was collected for CAD diagnosis.
108	Zoo	Multivariate	Classification	Categorical, Integer	101	17	1990	https://archive.ics.uci.edu/ml/machine-learnin...	Artificial, 7 classes of animals

488 rows × 9 columns