Python for Data Analysis Project

Eya El Hachemi

Yanyu Chen

Plan

1- Description of the dataset

2- Understanding the problem

3- Feature Encoding

4- Feature Engineering

Description of the dataset

Drug consumption (quantified) Data Set

	ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Alcohol	Amphet	Amyl	Benzos
0	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	CL5	CL2	CL2	CL0
1	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	CL6	CL0	CL0	CL0
2	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	CL4	CL0	CL0	CL3
3	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	CL4	CL1	CL1	CL0
4	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858	CL2	CL0	CL0	CL0

- 1884 Rows and 32 columns
- All input attributes are originally categorical and are quantified
- 0 missing values

Information about the columns

To summarize, it contains:

- an ID column
- 5 demographic columns (features)
- 7 personality traits (features)
- 18 drugs with their usage frequency (target)
- a fake drug called Semeron to verify reliability of answers

Each drug variable can take 6 different values:

- CL0 Never Used
- CL1 Used over a Decade
- CL2 Used in the Last Decade
- CL3 Used in the Last Year
- CL4 Used in the Last Month
- CL5 Used in the Last Week
- CL6 Used in the Last Day

Understanding the problem

- Classification Problem
- Problem will be transformed to binary classification by union of part of classes into one new class. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".
-

Feature encoding

Creating a drug encoder to convert nominal drug values (CL0, CL1, CL2 …) into ordered numerical data

```
Entrée [432]: data.head()
```

Out[432]:

	Age	Gender	Education	Country	Ethnicity	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LS
0	25-34	Male	Doctorate degree	UK	Other	6	2	2	0	7	4	7	3	0	4	0	2	0	
1	35-44	Male	Professional certificate/ diploma	UK	Other	7	0	0	0	7	3	4	0	0	0	0	0	0	
2	18-24	Female	Masters degree	UK	Other	4	0	0	3	6	2	4	2	0	0	0	2	0	
3	35-44	Female	Doctorate degree	UK	Other	4	1	1	0	7	3	7	0	0	1	0	0	1	
4	65+	Female	Left school at 18 years	Canada	Other	2	0	0	0	7	0	4	0	0	0	0	0	0	

Feature Engineering

We transformed the quantified categorical data back to a clearer, nominal, form (Columns : Age, Gender, Education, Country, Ethnicity) in order to explore our data and gain more information using encoders from the data description

	Age	Gender	Education	Country	Ethnicity	Ns
0	25-34	Male	Doctorate degree	UK	Other	-0.6
1	35-44	Male	Professional certificate/ diploma	UK	Other	-0.4
2	18-24	Female	Masters degree	UK	Other	-0.1
3	35-44	Female	Doctorate degree	UK	Other	0.7
4	65+	Female	Left school at 18 years	Canada	Other	-0.6

2. Age (Real) is age of participant and has one of the values:
Value Meaning Cases Fraction
-0.95197 18-24 643 34.11%
-0.07854 25-34 481 25.52%
0.49788 35-44 356 18.89%
1.09449 45-54 294 15.60%
1.82213 55-64 93 4.93%
2.59171 65+ 18 0.95%
Descriptive statistics
Min Max Mean Std.dev.
-0.95197 2.59171 0.03461 0.87813

3. Gender (Real) is gender of participant:
Value Meaning Cases Fraction
0.48246 Female 942 49.97%
-0.48246 Male 943 50.03%
Descriptive statistics
Min Max Mean Std.dev.
-0.48246 0.48246 -0.00026 0.48246

Feature encoding

We will create separate datasets to asses predict whether an individual uses cocaine, methamphetamines, or heroin…

Entrée [114]:
```
#feature engeering
#meth
meth_data = data.copy()
meth_data['Meth_User'] = meth_data['Meth'].apply(lambda x: 1 if x not in [0,1] else 0)
meth_data = meth_data.drop(['Meth'], axis=1)
```

Entrée [115]:
```
#cocaine
coc_data = data3.copy()
coc_data['Cocaine_User'] = coc_data['Coke'].apply(lambda x: 1 if x not in [0,1] else 0)
coc_data = coc_data.drop(['Coke'], axis=1)
```

Entrée [116]:
```
#crack
crack_data = data3.copy()
crack_data['Crack_User'] = crack_data['Crack'].apply(lambda x: 1 if x not in [0,1] else 0)
crack_data = crack_data.drop(['Crack'], axis=1)
```

Entrée [117]:
```
#nicotine
nico_data = data3.copy()
nico_data['Nico_User'] = nico_data['Nicotine'].apply(lambda x: 1 if x not in [0,1] else 0)
nico_data = nico_data.drop(['Nicotine'], axis=1)
```

Entrée [118]:
```
#heroin
hero_data = data3.copy()
hero_data['Hero_User'] = nico_data['Heroin'].apply(lambda x: 1 if x not in [0,1] else 0)
hero_data = nico_data.drop(['Heroin'], axis=1)
```

Thank you for your attention