Міністерство освіти і науки України
Національний університет «Львівська політехніка»
Інститут комп’ютерних наук та інформаційних технологій
Кафедра «Системи штучного інтелекту»

Лабораторна Робота №1

З предмету: «Видобування великих даних»

Виконав
студент групи КН-311
Ткачук Орест
Прийняла :
Вовк О. Б.

Львів-2021

Лабораторна робота №1

Завдання: зібрати дані з датасетів. Використовувати як мінімум 2 різних датасети. Попередньо проаналізувати та очистити дані

Хід роботи

Завантажуємо 2 датасети. Перший містить дані по кількості суїцидів на 100000 людей для кожної країни, другий – дані по кількості соц. працівників на 100000 людей для кожної країни.

```
suicide = pd.read_csv("C:\\Users\\Orest\\Desktop\\lpnu\\VVD\\Crude suicide rates.csv")
```

	Country	Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19
0	Afghanistan	Both sexes	42.0	11.0	5.5	5.6	6.6	9.2	10.2	3.1
1	Afghanistan	Male	70.4	20.9	9.8	9.3	10.5	15.1	16.3	4.8
2	Afghanistan	Female	20.1	2.3	1.4	1.6	2.3	2.7	3.5	1.2
3	Albania	Both sexes	16.3	8.3	6.0	7.8	9.1	6.1	6.5	5.0
4	Albania	Male	23.2	11.9	8.1	11.4	13.5	8.8	6.3	3.1
5	Albania	Female	10.9	4.9	3.9	4.4	5.0	3.4	6.6	7.0
6	Algeria	Both sexes	9.4	5.6	4.2	4.1	4.7	5.3	4.2	1.3
7	Algeria	Male	12.7	8.4	6.2	6.2	7.0	7.9	6.2	1.6
8	Algeria	Female	6.4	3.0	2.2	2.0	2.4	2.6	2.1	1.0
9	Angola	Both sexes	63.5	42.1	23.8	14.8	7.0	5.4	6.6	2.6
10	Angola	Male	123.2	68.2	34.7	21.8	10.7	8.4	10.4	3.8

Рис. 1 Датасет з інформацією про суїциди

```
hr = pd.read_csv("C:\\Users\\Orest\\Desktop\\lpnu\\VVD\\Human Resources.csv")
```

Оскільки дані лише за 2016 рік, можна викинути колонку ‘Year’.

```
hr = hr.drop('Year', axis = 1)
```

	Country	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	0.231	0.098	NaN	0.296
1	Albania	1.471	6.876	1.060	1.231
2	Angola	0.057	0.660	0.022	0.179
3	Antigua and Barbuda	1.001	7.005	4.003	NaN
4	Argentina	21.705	NaN	NaN	222.572
5	Armenia	3.840	11.245	0.274	0.788
6	Azerbaijan	3.452	6.717	0.114	1.165
7	Bangladesh	0.130	0.873	NaN	0.124
8	Belarus	13.504	NaN	NaN	5.514
9	Belize	1.392	3.340	NaN	1.113
10	Bhutan	0.508	0.127	NaN	NaN

Рис. 2 Датасет з інформацією про соц. працівників

Знаходимо кореляцію між датасетами.

```
corr = pd.concat([suicide, hr], axis=1, keys=['suicide', 'hr']).corr().loc['suicide', 'hr']
corr
```

	Psychiatrists	Nurses	Social_workers	Psychologists
80_above	-0.103822	0.040800	-0.053800	-0.077333
70to79	-0.065401	0.103956	-0.038626	-0.105002
60to69	-0.048811	0.112087	0.040885	-0.077691
50to59	-0.046169	0.056048	0.125882	0.031454
40to49	-0.024925	0.004177	0.246964	0.110156
30to39	-0.025376	0.028686	0.312060	0.071681
20to29	-0.022002	0.048341	0.399680	0.061256
10to19	-0.066034	0.029492	0.333376	0.041000

Рис. 3 Кореляція між колонками двох датасетів

Перевіряємо датасети на NaN значення

```
suicide.isna().sum().sum()
0

hr.isna().sum().sum()
81
```

Перед об'єднанням необхідно заповнити пропущені дані. Використовуємо алгоритм MICE щоб заповнити пропущені дані в другому датасеті

```
imr = IterativeImputer(min_value = 0)
imr = imr.fit(hr.drop('Country', axis = 1).values)
imputed_data = imr.transform(hr.drop('Country', axis = 1).values)


prof = ['Psychologists', 'Social_workers', 'Nurses', 'Psychiatrists']
for i in range(4):
    hr[prof[i]] = [ele[-i-1] for ele in imputed_data]
hr
```

	Country	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	0.231000	0.098000	0.000000	0.296000
1	Albania	1.471000	6.876000	1.060000	1.231000
2	Angola	0.057000	0.660000	0.022000	0.179000
3	Antigua and Barbuda	1.001000	7.005000	4.003000	8.291635
4	Argentina	21.705000	41.758903	148.557053	222.572000
5	Armenia	3.840000	11.245000	0.274000	0.788000
6	Azerbaijan	3.452000	6.717000	0.114000	1.165000
7	Bangladesh	0.130000	0.873000	0.000000	0.124000
8	Belarus	13.504000	33.400737	19.123188	5.514000
9	Belize	1.392000	3.340000	0.000000	1.113000
10	Bhutan	0.508000	0.127000	0.000000	3.435868

Рис. 4 Результат заповнення пустих значень за допомогою алгоритму MICE

Знову знаходимо кореляцію між датасетами після заповнення даних

	Psychiatrists	Nurses	Social_workers	Psychologists
80_above	-0.095497	-0.000239	-0.084142	-0.072269
70to79	-0.058518	0.056634	-0.064468	-0.086655
60to69	-0.039836	0.060323	-0.028101	-0.062469
50to59	-0.037339	0.020751	0.037183	0.030610
40to49	-0.016243	-0.011810	0.107562	0.102263
30to39	-0.015792	-0.003780	0.095927	0.068014
20to29	-0.011496	0.014577	0.111493	0.052762
10to19	-0.054660	-0.018680	0.069961	0.030237

Рис. 5 Кореляція між колонками двох датасетів після заповнення пропусків

Об'єднуємо датасети

```
merge = pd.merge(suicide, hr, on="Country")
merge
```

	Country	Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	Both sexes	42.0	11.0	5.5	5.6	6.6	9.2	10.2	3.1	0.231000	0.098000	0.000000	0.296000
1	Afghanistan	Male	70.4	20.9	9.8	9.3	10.5	15.1	16.3	4.8	0.231000	0.098000	0.000000	0.296000
2	Afghanistan	Female	20.1	2.3	1.4	1.6	2.3	2.7	3.5	1.2	0.231000	0.098000	0.000000	0.296000
3	Albania	Both sexes	16.3	8.3	6.0	7.8	9.1	6.1	6.5	5.0	1.471000	6.876000	1.060000	1.231000
4	Albania	Male	23.2	11.9	8.1	11.4	13.5	8.8	6.3	3.1	1.471000	6.876000	1.060000	1.231000
5	Albania	Female	10.9	4.9	3.9	4.4	5.0	3.4	6.6	7.0	1.471000	6.876000	1.060000	1.231000
6	Angola	Both sexes	63.5	42.1	23.8	14.8	7.0	5.4	6.6	2.6	0.057000	0.660000	0.022000	0.179000
7	Angola	Male	123.2	68.2	34.7	21.8	10.7	8.4	10.4	3.8	0.057000	0.660000	0.022000	0.179000
8	Angola	Female	25.3	21.7	14.5	8.6	3.5	2.5	2.9	1.4	0.057000	0.660000	0.022000	0.179000

Рис. 6 Об'єднання двох датасетів

Проводимо feature_selection, з умовою щоб варіативність в кожній колонці була більша 80%.

```
country = merge['Country']
sex = merge['Sex']
merge = merge.drop(['Country','Sex'],axis=1)

sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
sel.fit_transform(merge)
feature_sel_merge = merge[merge.columns[sel.get_support(indices=True)]]

feature_sel_merge.insert(loc=0, column='Sex', value = sex)
feature_sel_merge.insert(loc=0, column='Country', value = country)

feature_sel_merge
```

	Country	Sex	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists
0	Afghanistan	Both sexes	42.0	11.0	5.5	5.6	6.6	9.2	10.2	3.1	0.231000	0.098000	0.000000	0.296000
1	Afghanistan	Male	70.4	20.9	9.8	9.3	10.5	15.1	16.3	4.8	0.231000	0.098000	0.000000	0.296000
2	Afghanistan	Female	20.1	2.3	1.4	1.6	2.3	2.7	3.5	1.2	0.231000	0.098000	0.000000	0.296000
3	Albania	Both sexes	16.3	8.3	6.0	7.8	9.1	6.1	6.5	5.0	1.471000	6.876000	1.060000	1.231000
4	Albania	Male	23.2	11.9	8.1	11.4	13.5	8.8	6.3	3.1	1.471000	6.876000	1.060000	1.231000
5	Albania	Female	10.9	4.9	3.9	4.4	5.0	3.4	6.6	7.0	1.471000	6.876000	1.060000	1.231000
6	Angola	Both sexes	63.5	42.1	23.8	14.8	7.0	5.4	6.6	2.6	0.057000	0.660000	0.022000	0.179000
7	Angola	Male	123.2	68.2	34.7	21.8	10.7	8.4	10.4	3.8	0.057000	0.660000	0.022000	0.179000
8	Angola	Female	25.3	21.7	14.5	8.6	3.5	2.5	2.9	1.4	0.057000	0.660000	0.022000	0.179000

Рис. 7 Об'єднання двох датасетів після feature_selection

Усі дані мають достатню варіативність.

Instance_selection проводить зменшення датасету зменшуючи кількість даних, що повторюються. Оскільки цей датасет зразу погрупований по країнах, то цей крок можна опустити.

Загальна кореляція в об'єданому датасеті

	80_above	70to79	60to69	50to59	40to49	30to39	20to29	10to19	Psychiatrists	Nurses	Social_workers	Psychologists
80_above	1.000000	0.941117	0.827987	0.612695	0.401143	0.327255	0.241552	0.115678	-0.102774	-0.123983	-0.112995	-0.132048
70to79	0.941117	1.000000	0.925891	0.723530	0.504677	0.422736	0.321733	0.187010	-0.091199	-0.113153	-0.124040	-0.144504
60to69	0.827987	0.925891	1.000000	0.888028	0.720295	0.635479	0.519286	0.335042	0.048289	-0.018556	-0.044387	-0.100152
50to59	0.612695	0.723530	0.888028	1.000000	0.935646	0.859798	0.724948	0.527571	0.224902	0.123504	0.082497	-0.020774
40to49	0.401143	0.504677	0.720295	0.935646	1.000000	0.956327	0.836143	0.643873	0.255411	0.184627	0.123588	0.017032
30to39	0.327255	0.422736	0.635479	0.859798	0.956327	1.000000	0.929058	0.728000	0.218927	0.194338	0.123025	0.025538
20to29	0.241552	0.321733	0.519286	0.724948	0.836143	0.929058	1.000000	0.841213	0.182531	0.184268	0.151110	0.073516
10to19	0.115678	0.187010	0.335042	0.527571	0.643873	0.728000	0.841213	1.000000	0.144333	0.127861	0.157486	0.115752
Psychiatrists	-0.102774	-0.091199	0.048289	0.224902	0.255411	0.218927	0.182531	0.144333	1.000000	0.659817	0.754050	0.508858
Nurses	-0.123983	-0.113153	-0.018556	0.123504	0.184627	0.194338	0.184268	0.127861	0.659817	1.000000	0.448319	0.283516
Social_workers	-0.112995	-0.124040	-0.044387	0.082497	0.123588	0.123025	0.151110	0.157486	0.754050	0.448319	1.000000	0.900775
Psychologists	-0.132048	-0.144504	-0.100152	-0.020774	0.017032	0.025538	0.073516	0.115752	0.508858	0.283516	0.900775	1.000000

Рис. 8 Загальна кореляція в об'єднаному датасеті

Висновок: у даній лабораторній роботі мені потрібно було зробити об'єднання двох датасетів. Основною проблемою був вибір зв'язаних між собою датасетів, які б можна було з'єднати по відповідній колонці. Також при виконанні лабораторної виникли проблеми з заповненням пропущених даних. Заповнив я їх за допомогою вбудованого в модуль skipy алгоритму MICE, а також спробував провести feature_selection, щоб вибрати лише основні дані. В лабораторній роботі №2 я планую знайти залежність між соц. працівниками та кількістю самогубств в країнах, а також знайти оптимальну кількість працівників для мінімізації рівня самогубств.