Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа № 2

по дисциплине «Методы машинного обучения»

Обработка признаков ч.1

ИСПОЛНИТЕЛЬ:

студент ИУ5-25М
Мацнев А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

"__"_____2023 г.

Москва, 2023

Задание

1. Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
    - устранение пропусков в данных;
    - кодирование категориальных признаков;
    - нормализация числовых признаков.

Выполнение задания

1. Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)

   Данные – информация об ожидаемой продолжительности жизни:

```
data = pd.read_csv("Life Expectancy Data.csv")
data.head()
```

Python

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259210	33736494.0	17.2	17.3	0.479	10.1
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696514	327582.0	17.5	17.5	0.476	10.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744976	31731688.0	17.7	17.7	0.470	9.9
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959000	3696958.0	17.9	18.0	0.463	9.8
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537231	2978599.0	18.2	18.2	0.454	9.5

5 rows × 22 columns

Рис. 1. Набор данных

2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   - устранение пропусков в данных;

```
columns_with_md = [c for c in data.columns if data[c].isnull().sum() > 0]
print("Missing data columns count: ",columns_with_md.__len__())
[(column_name, data[column_name].isnull().sum(), data[column_name].isnull().mean()*100) for column_name in
```

Python

```
Missing data columns count:  14

[('Life expectancy ', 10, 0.3403675970047651),
 ('Adult Mortality', 10, 0.3403675970047651),
 ('Alcohol', 194, 6.603131381892443),
 ('Hepatitis B', 553, 18.82232811436351),
 (' BMI ', 34, 1.1572498298162015),
 ('Polio', 19, 0.6466984343090538),
 ('Total expenditure', 226, 7.6923076923076925),
 ('Diphtheria ', 19, 0.6466984343090538),
 ('GDP', 448, 15.248468345813478),
 ('Population', 652, 22.19196732471069),
 (' thinness  1-19 years', 34, 1.1572498298162015),
 (' thinness 5-9 years', 34, 1.1572498298162015),
 ('Income composition of resources', 167, 5.684138869979578),
 ('Schooling', 163, 5.547991831177672)]
```

Рис. 2. Пропуски в данных

Заполнение пропусков в данных:

```python
knnimputer = KNNImputer(
    n_neighbors=7,
    weights='distance',
    metric='nan_euclidean',
    add_indicator=False
    )
imputed_data_temp = knnimputer.fit_transform(missing_data)
imputed_data = pd.DataFrame(imputed_data_temp, columns=missing_data.columns)
imputed_data.head()
```

Python

	BMI	HIV/AIDS	thinness 1-19 years	thinness 5-9 years	Adult Mortality	Alcohol	Diphtheria	GDP	Hepatitis B	Income composition of resources	L expectan
0	19.1	0.1	17.2	17.3	263.0	0.01	65.0	584.259210	65.0	0.479	6
1	18.6	0.1	17.5	17.5	271.0	0.01	62.0	612.696514	62.0	0.476	5
2	18.1	0.1	17.7	17.7	268.0	0.01	64.0	631.744976	64.0	0.470	5
3	17.6	0.1	17.9	18.0	272.0	0.01	67.0	669.959000	67.0	0.463	5
4	17.2	0.1	18.2	18.2	275.0	0.01	68.0	63.537231	68.0	0.454	5

```python
columns_with_md = [c for c in imputed_data.columns if imputed_data[c].isnull().sum() > 0]
print("Missing data columns count: ",columns_with_md.__len__())
[(column_name, imputed_data[column_name].isnull().sum(), imputed_data[column_name].isnull().mean()*100) fo
```

Python

```
Missing data columns count:  0

[]
```

- кодирование категориальных признаков;

```python
Target_ENC = TargetEncoder()
target_features = ['Life expectancy ']
not_target_features = data.columns.difference(target_features)
not_target_features
encoded_data = Target_ENC.fit_transform(data[not_target_features], data[target_features])
encoded_data
```

Python

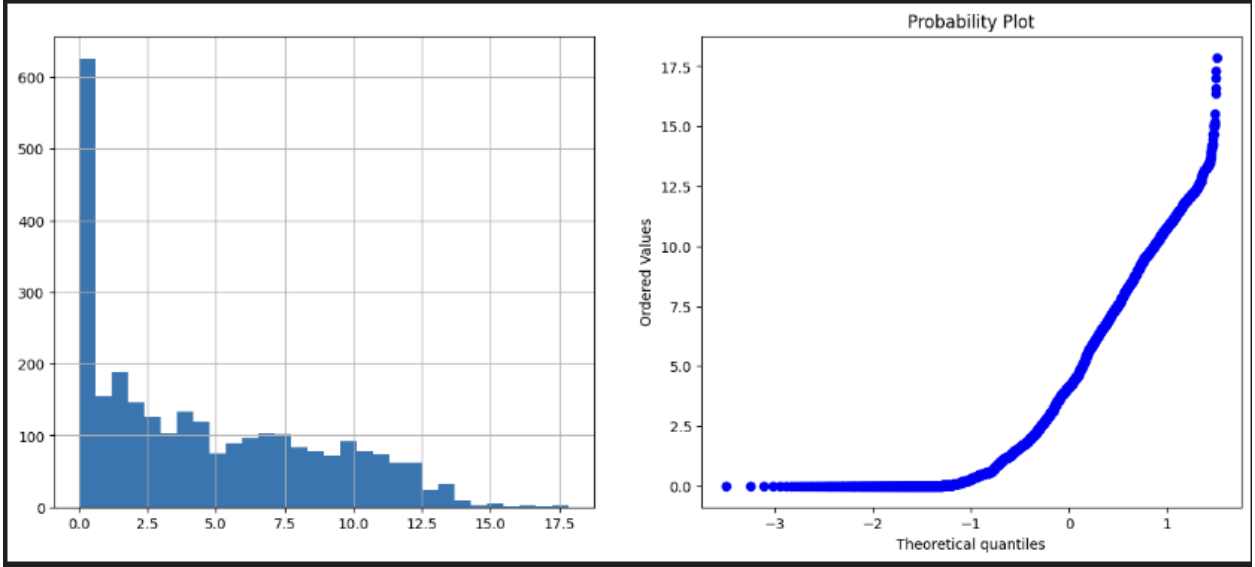
	BMI	HIV/AIDS	thinness 1-19 years	thinness 5-9 years	Adult Mortality	Alcohol	Country	Diphtheria	GDP	Hepatitis B	...	Me
0	19.1	0.1	17.2	17.3	263.0	0.01	64.797982	65.0	584.259210	65.0	...	
1	18.6	0.1	17.5	17.5	271.0	0.01	64.797982	62.0	612.696514	62.0	...	
2	18.1	0.1	17.7	17.7	268.0	0.01	64.797982	64.0	631.744976	64.0	...	
3	17.6	0.1	17.9	18.0	272.0	0.01	64.797982	67.0	669.959000	67.0	...	
4	17.2	0.1	18.2	18.2	275.0	0.01	64.797982	68.0	63.537231	68.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
2933	27.1	33.6	9.4	9.4	723.0	4.36	61.705369	65.0	454.366654	68.0	...	
2934	26.7	36.7	9.8	9.9	715.0	4.06	61.705369	68.0	453.351155	7.0	...	
2935	26.3	39.8	1.2	1.3	73.0	4.43	61.705369	71.0	57.348340	73.0	...	
2936	25.9	42.1	1.6	1.7	686.0	1.72	61.705369	75.0	548.587312	76.0	...	
2937	25.5	43.5	11.0	11.2	665.0	1.68	61.705369	78.0	547.358878	79.0	...	

2938 rows × 21 columns

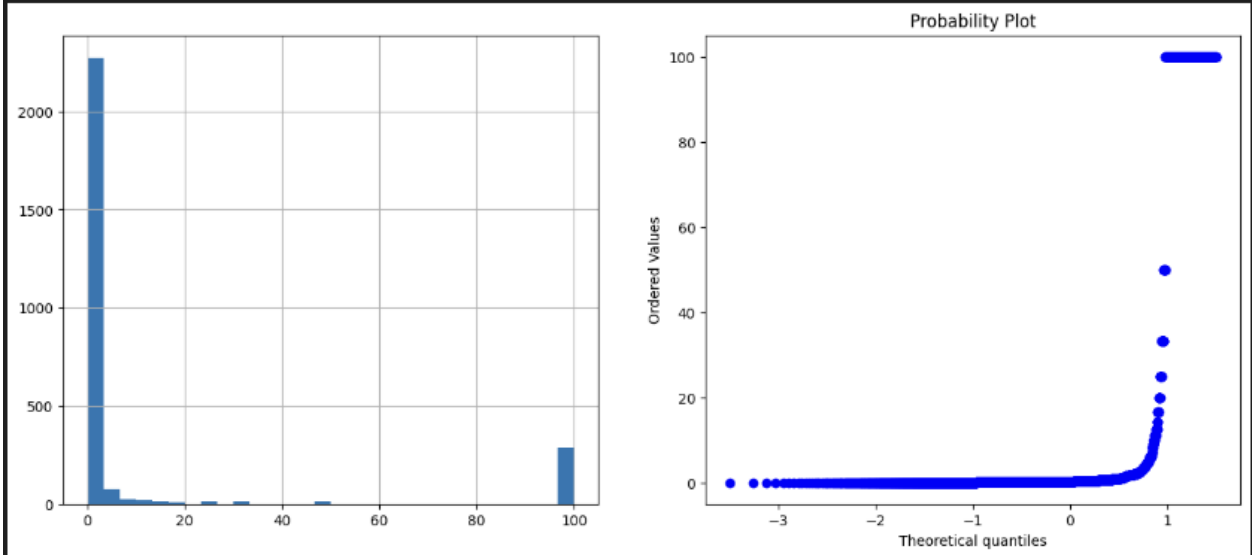
- нормализация числовых признаков.

```
plt.figure(figsize=(15,6))
diagnostic_plots(data, 'Alcohol')
```

```
<Figure size 1500x600 with 0 Axes>
```

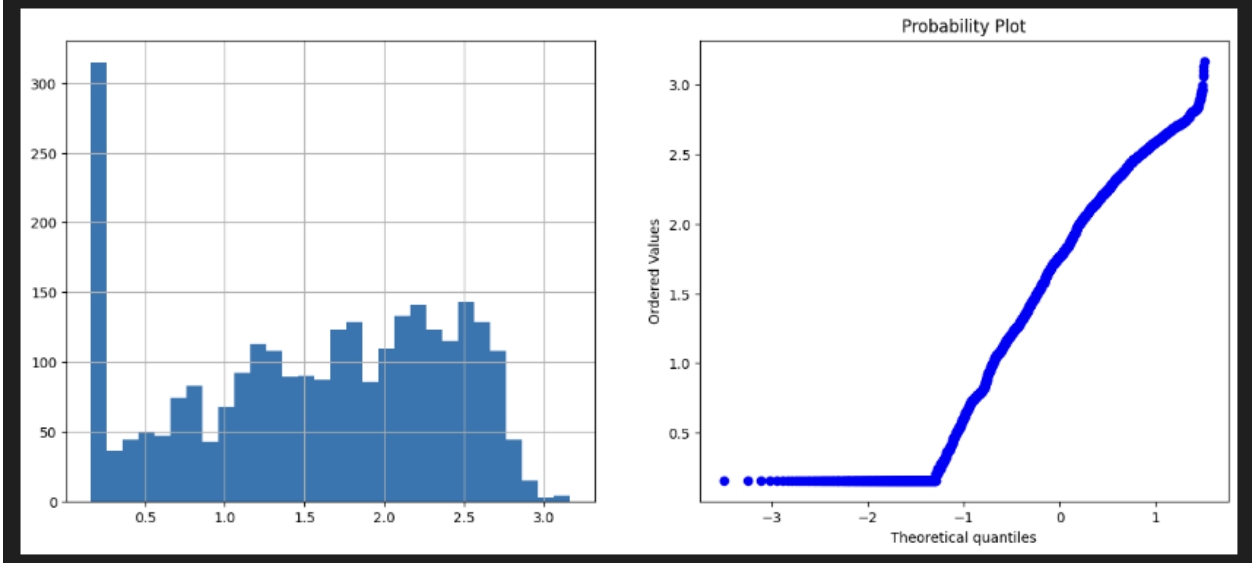



```
data['Alcohol_reciprocal'] = 1 / (data['Alcohol'])
diagnostic_plots(data, 'Alcohol_reciprocal')
```



```
data['Alcohol_exp0p4'] = data['Alcohol']**(0.4)
diagnostic_plots(data, 'Alcohol_exp0p4')
```

Pytho




```
data['Alcohol_log'] = np.log(data['Alcohol'])
diagnostic_plots(data, 'Alcohol_log')
```

Pytho