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

Лабораторная работа № 3

по дисциплине «Методы машинного обучения»

Обработка признаков ч.2

ИСПОЛНИТЕЛЬ:

студент ИУ5-25М
Мацнев А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

"__"_____2023 г.

Москва, 2023

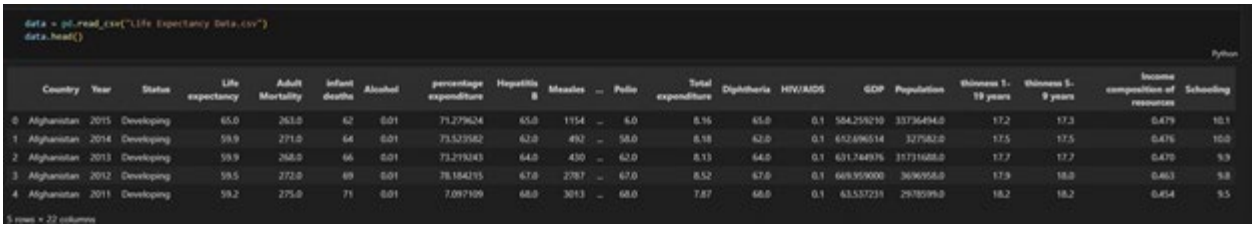
Задание

1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   - масштабирование признаков (не менее чем тремя способами);
   - обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
   - обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
   - отбор признаков:
     - один метод из группы методов фильтрации (filter methods);
     - один метод из группы методов обертывания (wrapper methods);
     - один метод из группы методов вложений (embedded methods).

Выполнение задания

1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.

   Данные – информация об ожидаемой продолжительности жизни:



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

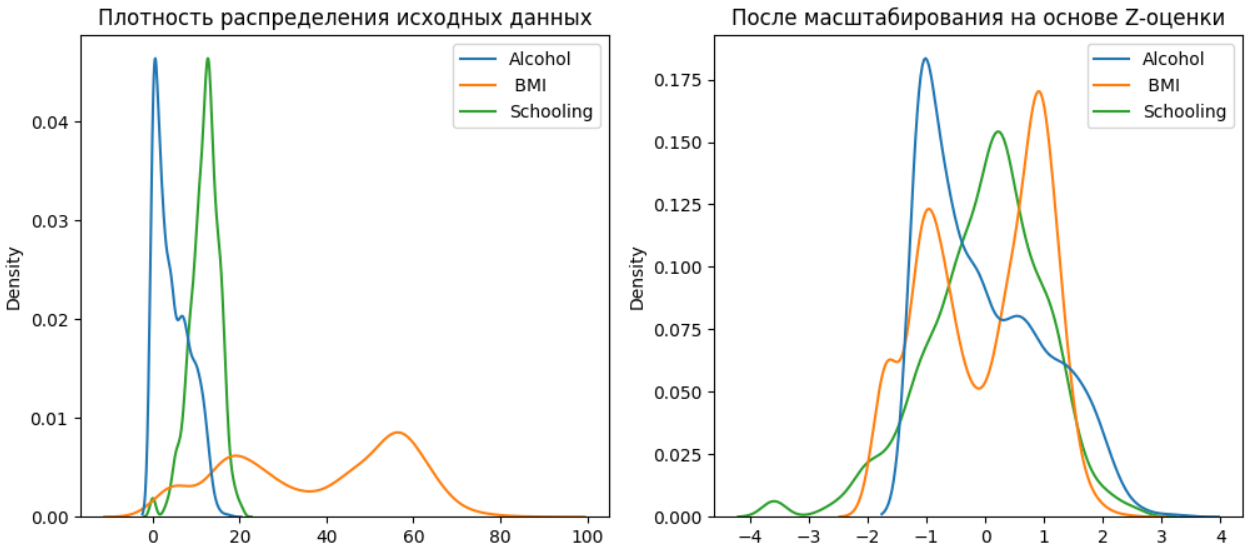
	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	Developing	65.0	263.0	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

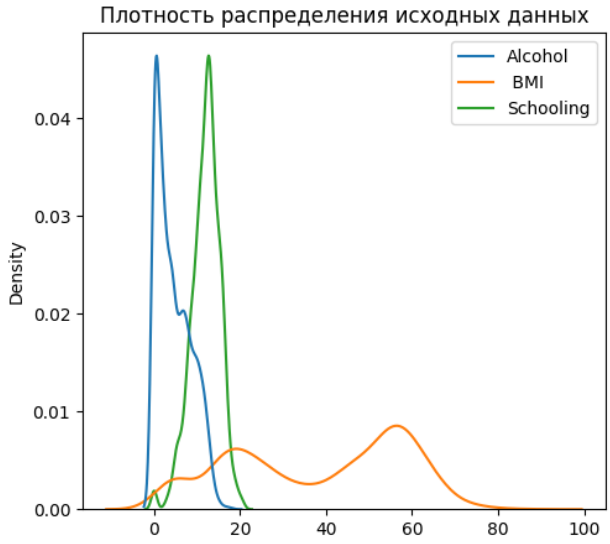
Рис. 1. Набор данных

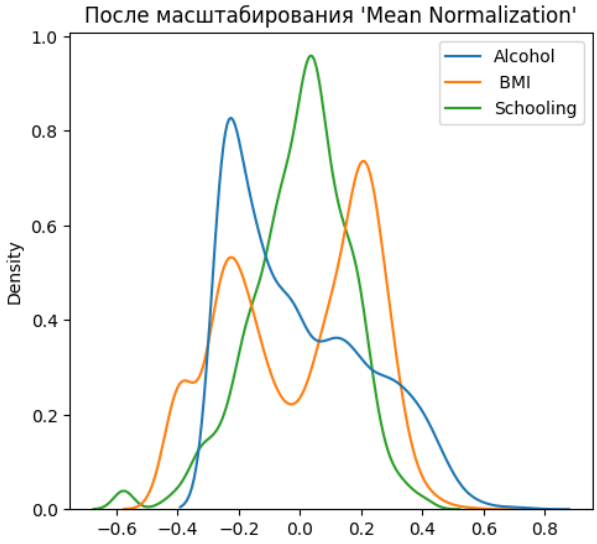
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   - масштабирование признаков (не менее чем тремя способами);

На основе Z-оценки:

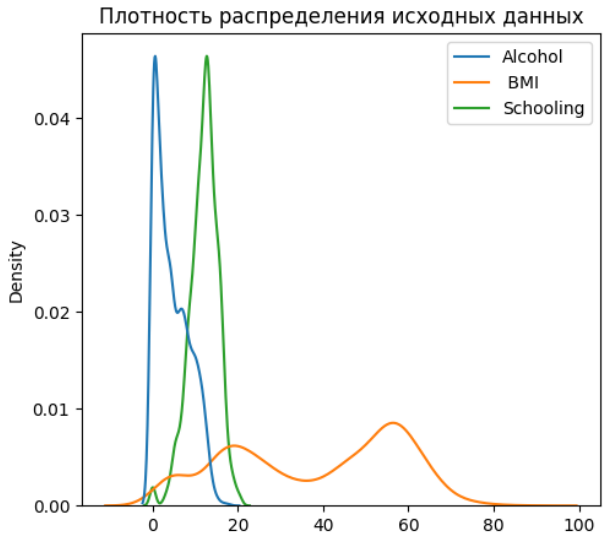


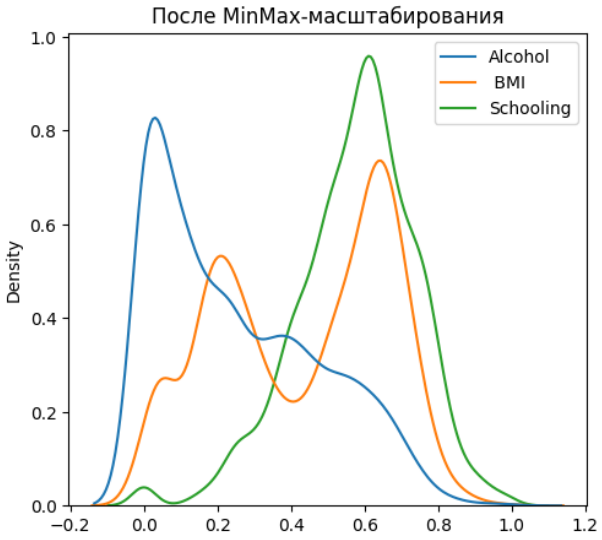
Mean Normalization:





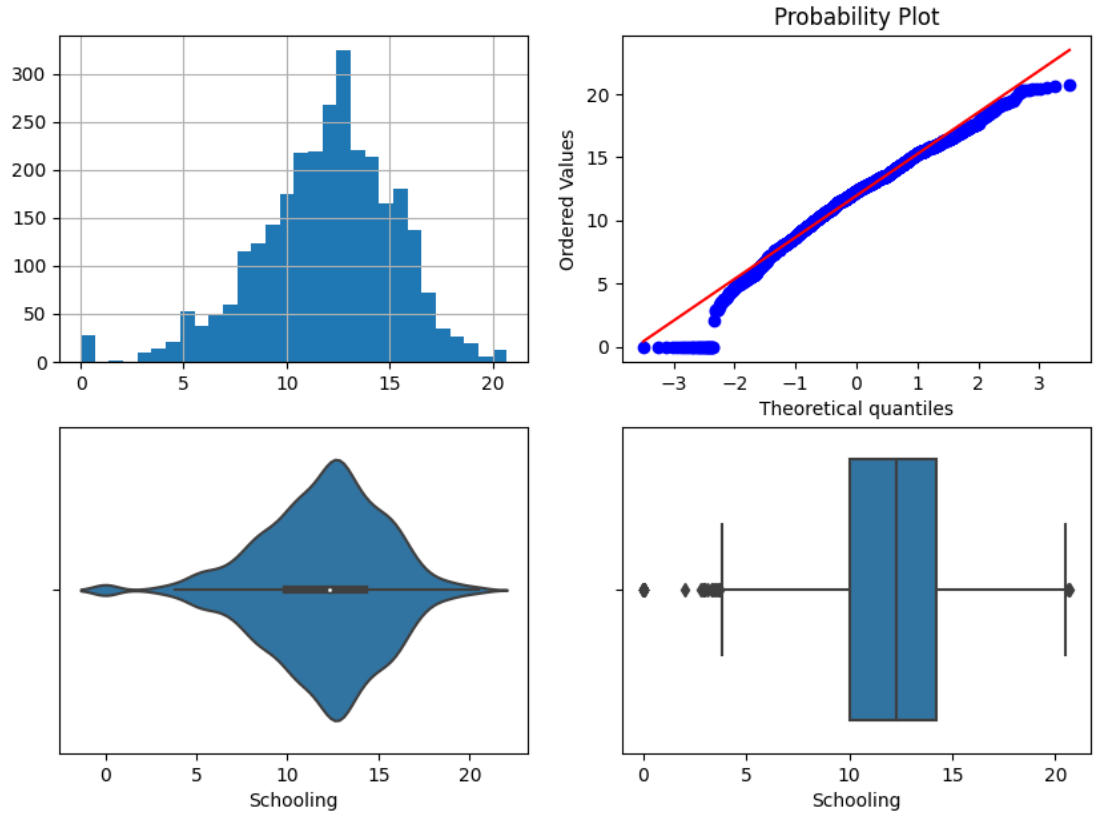
MinMax-масштабирование:



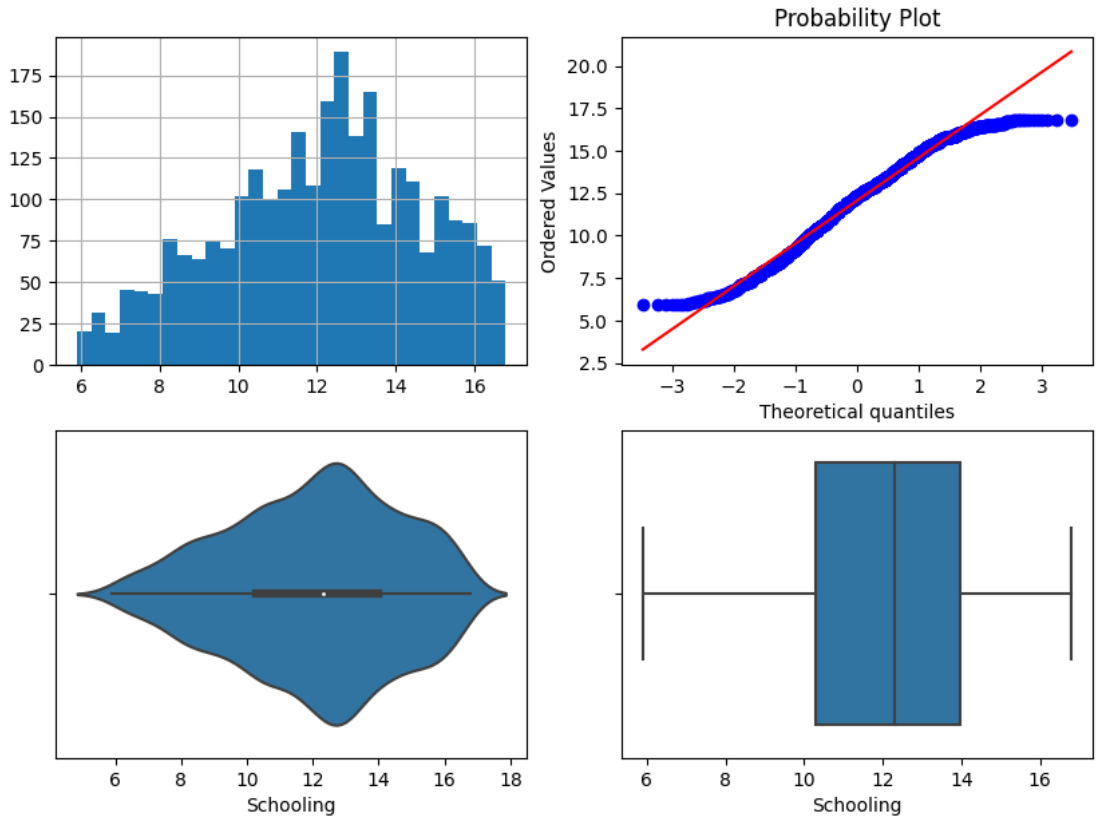


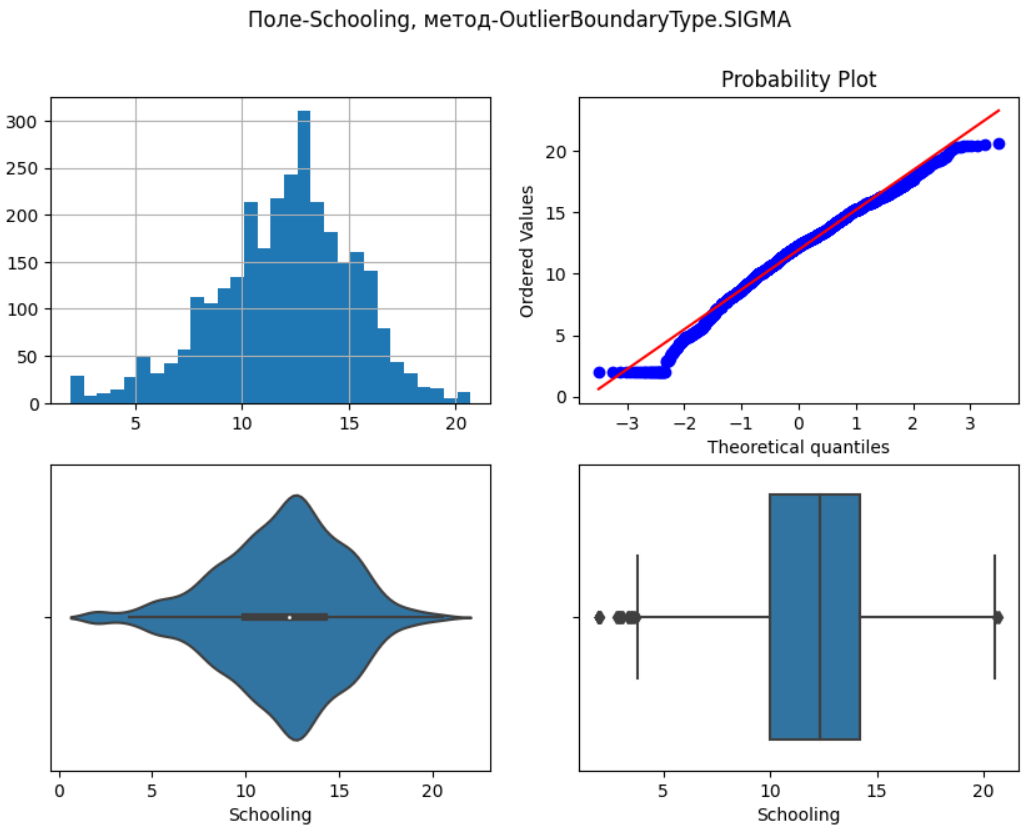
- обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);

Schooling - original



Поле-Schooling, метод-OutlierBoundaryType.QUANTILE, строк-2660





- обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);

```
data_nonstandard
```

	date	home_team	away_team	team	scorer	minute	own_goal	penalty
0	1916-07-02	Chile	Uruguay	Uruguay	José Piendibene	44.0	False	False
1	1916-07-02	Chile	Uruguay	Uruguay	Isabelino Gradín	55.0	False	False
2	1916-07-02	Chile	Uruguay	Uruguay	Isabelino Gradín	70.0	False	False
3	1916-07-02	Chile	Uruguay	Uruguay	José Piendibene	75.0	False	False
4	1916-07-06	Argentina	Chile	Argentina	Alberto Ohaco	2.0	False	False
...	...	...	...	...	...	...	...	...
41003	2022-12-18	Argentina	France	Argentina	Ángel Di María	36.0	False	False
41004	2022-12-18	Argentina	France	France	Kylian Mbappé	80.0	False	True
41005	2022-12-18	Argentina	France	France	Kylian Mbappé	81.0	False	False
41006	2022-12-18	Argentina	France	Argentina	Lionel Messi	109.0	False	False
41007	2022-12-18	Argentina	France	France	Kylian Mbappé	118.0	False	True

41008 rows × 8 columns

```
data_nonstandard['temp_date'] = data_nonstandard.apply(lambda x: pd.to_datetime(x['date'], format='%Y-%m-%d'), axis=1)
data_nonstandard['year'] = data_nonstandard['temp_date'].dt.year
data_nonstandard['month'] = data_nonstandard['temp_date'].dt.month
data_nonstandard['day'] = data_nonstandard['temp_date'].dt.day
data_nonstandard = data_nonstandard.drop('temp_date', axis=1)
data_nonstandard['scorer name'] = data_nonstandard.apply(lambda x: str(x['scorer']).split(' ')[0], axis=1)
data_nonstandard
```

	date	home_team	away_team	team	scorer	minute	own_goal	penalty	year	month	day	scorer name
0	1916-07-02	Chile	Uruguay	Uruguay	José Piendibene	44.0	False	False	1916	7	2	José
1	1916-07-02	Chile	Uruguay	Uruguay	Isabelino Gradín	55.0	False	False	1916	7	2	Isabelino
2	1916-07-02	Chile	Uruguay	Uruguay	Isabelino Gradín	70.0	False	False	1916	7	2	Isabelino
3	1916-07-02	Chile	Uruguay	Uruguay	José Piendibene	75.0	False	False	1916	7	2	José
4	1916-07-06	Argentina	Chile	Argentina	Alberto Ohaco	2.0	False	False	1916	7	6	Alberto
...	...	...	...	...	...	...	...	...	...	...	...	...
41003	2022-12-18	Argentina	France	Argentina	Ángel Di María	36.0	False	False	2022	12	18	Ángel
41004	2022-12-18	Argentina	France	France	Kylian Mbappé	80.0	False	True	2022	12	18	Kylian
41005	2022-12-18	Argentina	France	France	Kylian Mbappé	81.0	False	False	2022	12	18	Kylian
41006	2022-12-18	Argentina	France	Argentina	Lionel Messi	109.0	False	False	2022	12	18	Lionel
41007	2022-12-18	Argentina	France	France	Kylian Mbappé	118.0	False	True	2022	12	18	Kylian

41008 rows × 12 columns

- отбор признаков:
  - один метод из группы методов фильтрации (filter methods);

```
data
```

	BMI	HIV/AIDS	thinness 1-19 years	thinness 5-9 years	Adult Mortality	Alcohol	Diphtheria	GDP	Hepatitis B	Income composition of resources	Life expectancy	Measles	Polio
0	19.1	0.1	17.2	17.3	263.0	0.01	65.0	584.259210	65.0	0.479	65.0	1154.0	6.0
1	18.6	0.1	17.5	17.5	271.0	0.01	62.0	612.696514	62.0	0.476	59.9	492.0	58.0
2	18.1	0.1	17.7	17.7	268.0	0.01	64.0	631.744976	64.0	0.470	59.9	430.0	62.0
3	17.6	0.1	17.9	18.0	272.0	0.01	67.0	669.959000	67.0	0.463	59.5	2787.0	67.0
4	17.2	0.1	18.2	18.2	275.0	0.01	68.0	63.537231	68.0	0.454	59.2	3013.0	68.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	27.1	33.6	9.4	9.4	723.0	4.36	65.0	454.366654	68.0	0.407	44.3	31.0	67.0
2934	26.7	36.7	9.8	9.9	715.0	4.06	68.0	453.351155	7.0	0.418	44.5	998.0	7.0
2935	26.3	39.8	1.2	1.3	73.0	4.43	71.0	57.348340	73.0	0.427	44.8	304.0	73.0
2936	25.9	42.1	1.6	1.7	686.0	1.72	75.0	548.587312	76.0	0.427	45.3	529.0	76.0
2937	25.5	43.5	11.0	11.2	665.0	1.68	78.0	547.358878	79.0	0.434	46.0	1483.0	78.0

2938 rows × 20 columns

```
selected_data1 = arr_to_df(selector1.transform(data))
```

```
selected_data1
```

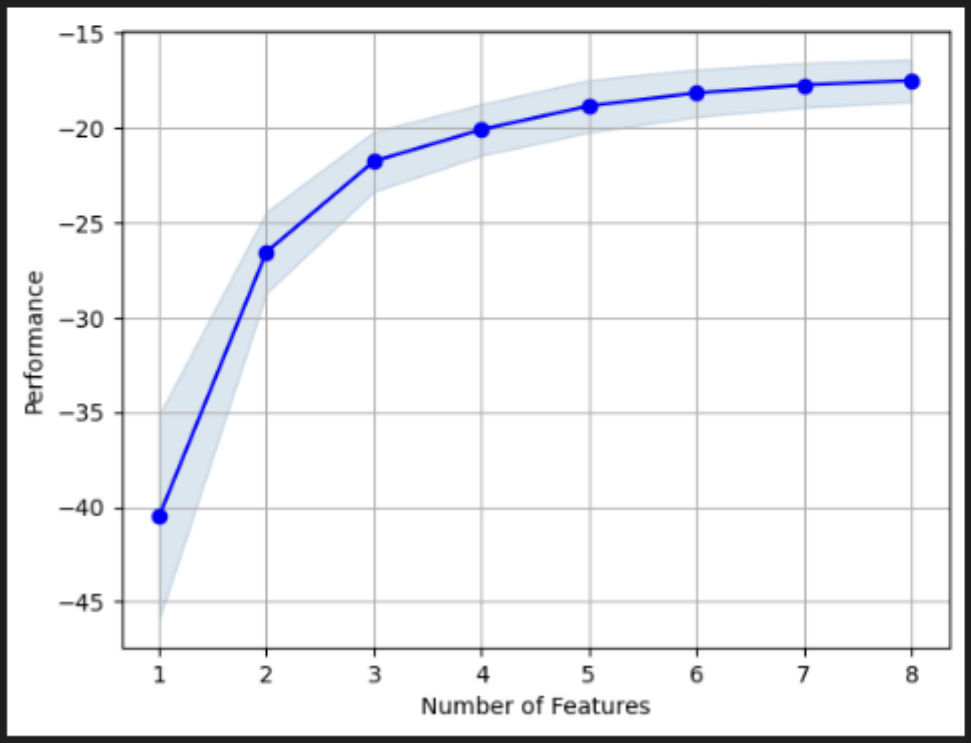
	BMI	HIV/AIDS	thinness 1-19 years	thinness 5-9 years	Adult Mortality	Alcohol	Diphtheria	GDP	Hepatitis B	Income composition of resources	Measles	Polio	Population
0	19.1	0.1	17.2	17.3	263.0	0.01	65.0	584.259210	65.0	65.0	1154.0	6.0	33736494.0
1	18.6	0.1	17.5	17.5	271.0	0.01	62.0	612.696514	62.0	59.9	492.0	58.0	327582.0
2	18.1	0.1	17.7	17.7	268.0	0.01	64.0	631.744976	64.0	59.9	430.0	62.0	31731688.0
3	17.6	0.1	17.9	18.0	272.0	0.01	67.0	669.959000	67.0	59.5	2787.0	67.0	3696958.0
4	17.2	0.1	18.2	18.2	275.0	0.01	68.0	63.537231	68.0	59.2	3013.0	68.0	2978599.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	27.1	33.6	9.4	9.4	723.0	4.36	65.0	454.366654	68.0	44.3	31.0	67.0	12777511.0
2934	26.7	36.7	9.8	9.9	715.0	4.06	68.0	453.351155	7.0	44.5	998.0	7.0	12633897.0
2935	26.3	39.8	1.2	1.3	73.0	4.43	71.0	57.348340	73.0	44.8	304.0	73.0	125525.0
2936	25.9	42.1	1.6	1.7	686.0	1.72	75.0	548.587312	76.0	45.3	529.0	76.0	12366165.0
2937	25.5	43.5	11.0	11.2	665.0	1.68	78.0	547.358878	79.0	46.0	1483.0	78.0	12222251.0

  - один метод из группы методов обертывания (wrapper methods);

```
lr = LinearRegression()
sfs1 = SFS(lr,
           k_features=8,
           forward=True,
           floating=False,
           scoring='neg_mean_squared_error',
           cv=10)
```

```
sfs1=sfs1.fit(X_ALL, Y)
```

```
fig = plot_sfs(sfs1.get_metric_dict(), kind='std_err')
plt.grid()
plt.show()
```



– один метод из группы методов вложений (embedded methods).

```
    linear_reg = Lasso(random_state=1)
    linear_reg.fit(X_ALL, Y)
```

```
c:\Users\Top_p\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\linear_model\_coordinate_descent.py:631: Convergenc
  model = cd_fast.enet_coordinate_descent(
```

```
Lasso
Lasso(random_state=1)
```

```
    list(zip(data.columns, linear_reg.coef_))
```

```
[(' BMI ', 0.05330798705420212),
 (' HIV/AIDS', -0.4354689006583894),
 (' thinness  1-19 years', -0.04243570168805266),
 (' thinness 5-9 years', -0.0),
 ('Adult Mortality', -0.021493113826206327),
 ('Alcohol', 0.06703891483921752),
 ('Diphtheria ', 0.03852312117548346),
 ('GDP', 7.03443220324406e-05),
 ('Hepatitis B', -0.006814763854718093),
 ('Income composition of resources', 0.0),
 ('Life expectancy ', -2.0544927126545947e-05),
 ('Measles ', 0.02861080578908092),
 ('Polio', -1.1824566094900604e-09),
 ('Population', 0.889451030436004),
 ('Schooling', 0.0),
 ('Total expenditure', 0.0),
 ('Year', 0.08774708961805171),
 ('infant deaths', 6.367131577864163e-05),
 ('percentage expenditure', -0.06564397533445436)]
```

```
    select_lr = SelectFromModel(linear_reg)
    select_lr.fit(X_ALL, Y)
    list(zip(data.columns, select_lr.get_support()))
```

```
c:\Users\Top_p\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\linear_model\_coordinate_descent.py:631: Convergenc
  model = cd_fast.enet_coordinate_descent(
```

```
[(' BMI ', True),
 (' HIV/AIDS', True),
 (' thinness  1-19 years', True),
 (' thinness 5-9 years', False),
 ('Adult Mortality', True),
 ('Alcohol', True),
 ('Diphtheria ', True),
 ('GDP', True),
 ('Hepatitis B', True),
 ('Income composition of resources', False),
 ('Life expectancy ', True),
 ('Measles ', True),
 ('Polio', False),
 ('Population', True),
```