


✓ Рубежный контроль №1 по курсу «Методы машинного обучения»

ИУ5-21М Данилин М.С.

✓ Импорт библиотек

```
import numpy as np
import pandas as pd
import seaborn as sns
```

```
# Подгрузим датасет и продемонстрируем его содержимое
data_loaded = pd.read_csv('./cust_beh.csv', sep=",")
data_loaded.head()
```



	Taken_product	Yearly_avg_view_on_travel_page	preferred_device	total_li
0	Yes	307.0	iOS and Android	
1	No	367.0	iOS	
2	Yes	277.0	iOS and Android	
3	No	247.0	iOS	
4	No	202.0	iOS and Android	

```
# Используем только некоторые признаки
cols_filter = ['Taken_product', 'preferred_device', 'Yearly_avg_view_on_travel_p
               'travelling_network_rating']
data = data_loaded[cols_filter]
data.head()
```

	Taken_product	preferred_device	Yearly_avg_view_on_travel_page	Yearly_ε
0	Yes	iOS and Android		307.0
1	No	iOS		367.0
2	Yes	iOS and Android		277.0
3	No	iOS		247.0
4	No	iOS and Android		202.0


✓ Задание 5. Для набора данных проведите кодирование одного (произвольного) категориального признака с использованием метода "one-hot encoding".

Информация из лекции

- One-hot encoding предполагает, что значение категории заменяется на отдельную колонку, которая содержит бинарные значения.
- Преимущества:
 - Простота реализации.
 - Подходит для любых моделей, так как НЕ создает фиктивное отношение порядка между значениями.
- Недостатки:
 - Расширяется признаковое пространство.


✓ Решение

```
pd.get_dummies(data[['preferred_device']]).head()
```



	preferred_device_ANDROID	preferred_device_Android	preferred_device_And:
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False

```
# Добавление отдельной колонки, признака пустых значений
pd.get_dummies(data[['preferred_device']], dummy_na=True).head()
```



	preferred_device_ANDROID	preferred_device_Android	preferred_device_And:
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False

Задание 25. Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе межквартильного размаха.

Информация из лекции

Межквартильный размах IQR (interquartile range, IQR) - это разность третьего квартиля и первого квартиля:

✓ Решение

✓ Обнаруживаем выбросы

```
def remove_outliers_iqr(data, column):  
    Q1 = data[column].quantile(0.25)  
    Q3 = data[column].quantile(0.75)  
    IQR = Q3 - Q1  
    lower_bound = Q1 - 1.5 * IQR  
    upper_bound = Q3 + 1.5 * IQR  
    filtered_data = data[(data[column] >= lower_bound) & (data[column] <= upper_bound)]  
    return filtered_data
```

✓ Удаление выбросов

```
data.shape
```

```
↔ (11756, 5)
```

```
filtered_dataset = remove_outliers_iqr(data, 'Yearly_avg_view_on_travel_page')  
filtered_dataset.shape
```

```
↔ (11165, 5)
```

✓ Диаграмма рассеяния

```
sns.scatterplot(  
    data=data,  
    x="Yearly_avg_comment_on_travel_page",  
    y="Yearly_avg_view_on_travel_page",  
    hue="Taken_product"  
)  
  
[↔] <Axes: xlabel='Yearly_avg_comment_on_travel_page',  
      ylabel='Yearly_avg_view_on_travel_page'>
```

