Московский Государственный Технический Университет имени Н.Э. Баумана

Факультет Информатика и системы управления

**Кафедра ИУ-5 «Системы обработки информации и управления»**

# Отчёт по рубежному контролю № 1

**По дисциплине**

«Методы Машинного Обучения»

Выполнила студентка Хэ Синьчэнь
Группа ИУ5И-24М

**Москва 2024г**

# Номер варианта: 21

# Номер задачи №1:5

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

## Задача 5

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

```
[26] #Загрузка и предобработка данных
     import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
     from sklearn.datasets import load_iris
```

```
[27] # 加载 titanic 数据集
     titanic = sns.load_dataset('titanic')
```

```
[28] # Несколько первых строк данных
     print("Несколько первых строк данных:")
     display(titanic.head())
```

Несколько первых строк данных:

| | survived | pclass | sex | age | sibsp | parch | fare | embarked | class | who | adult_male | deck | embark_town | alive | alone |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 3 | male | 22.0 | 1 | 0 | 7.2500 | S | Third | man | True | NaN | Southampton | no | False |
| **1** | 1 | 1 | female | 38.0 | 1 | 0 | 71.2833 | C | First | woman | False | C | Cherbourg | yes | False |
| **2** | 1 | 3 | female | 26.0 | 0 | 0 | 7.9250 | S | Third | woman | False | NaN | Southampton | yes | True |
| **3** | 1 | 1 | female | 35.0 | 1 | 0 | 53.1000 | S | First | woman | False | C | Southampton | yes | False |
| **4** | 0 | 3 | male | 35.0 | 0 | 0 | 8.0500 | S | Third | man | True | NaN | Southampton | no | True |

```
# Например sex
encoded_feature = pd.get_dummies(titanic['sex'], prefix='sex')
```

```
[30] print("\nКодированные функции:")
     display(encoded_feature.head())
```

Кодированные функции:

| | sex_female | sex_male |
|---|---|---|
| **0** | False | True |
| **1** | True | False |
| **2** | True | False |
| **3** | True | False |
| **4** | False | True |

# Номер задачи №2: 23

Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе правила трех сигм.

## Задача 23

Для набора данных для одного (произвольного) числового признака проведите обнаружение и удаление выбросов на основе правила трех сигм.

```
[33] # 任务二：对于单个数字特征的数据集，根据三西格玛规则进行离群点检测和移除
     # Например age
     feature_to_check = 'age'
     data_to_check = titanic[feature_to_check]
```

```
[34] # Рассчитать среднее значение и стандартное отклонение
     mean_value = np.mean(data_to_check)
     std_deviation = np.std(data_to_check)
```

```
[35] # Определение сферы применения Три сигмы
     lower_bound = mean_value - 3 * std_deviation
     upper_bound = mean_value + 3 * std_deviation
```

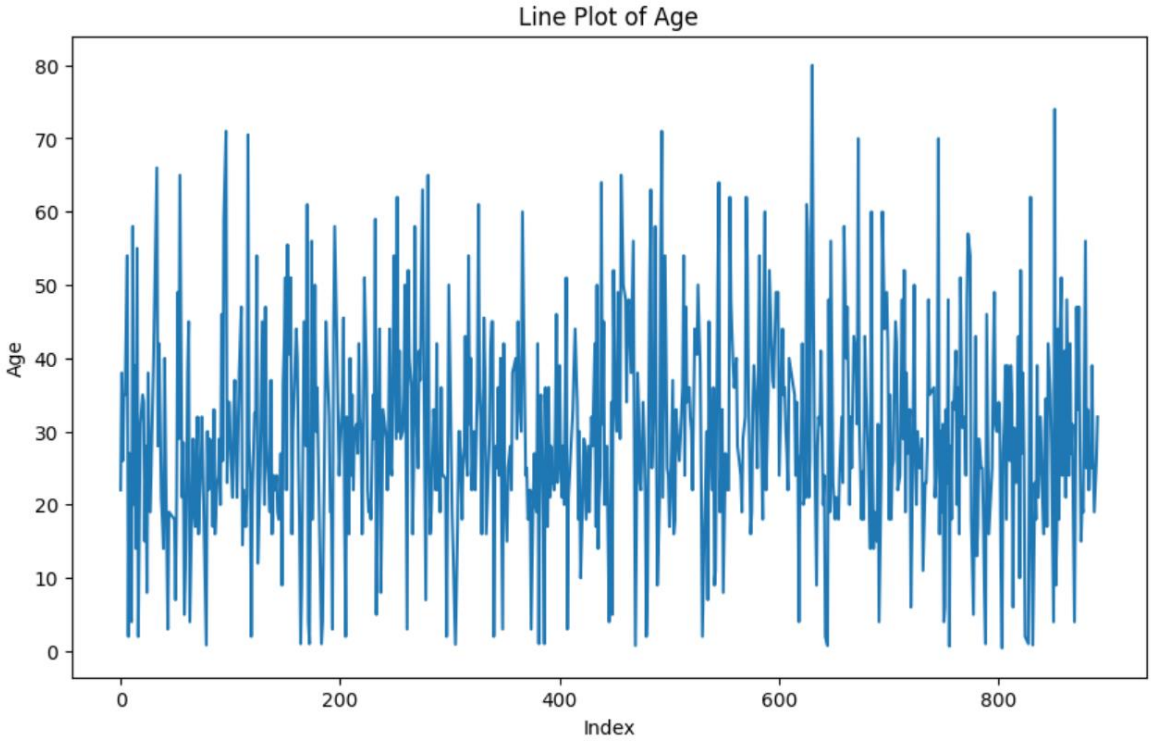
```
[36] # Индекс отмеченных выбросов
     outliers_index = (data_to_check < lower_bound) | (data_to_check > upper_bound)

     # Удаление выбросов
     cleaned_data = titanic[~outliers_index]
```

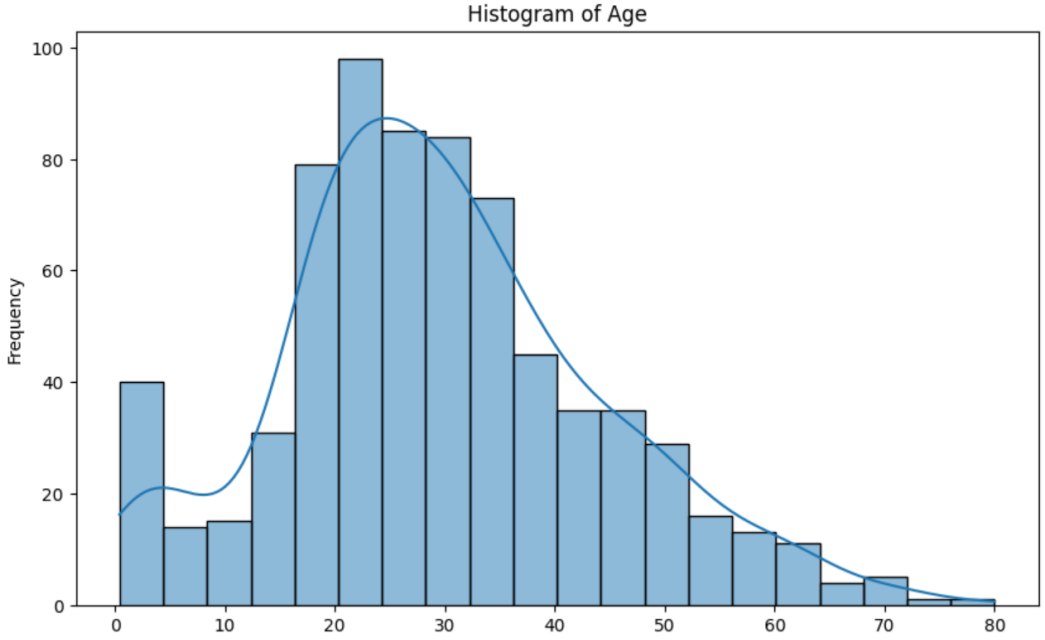
```
# Показывает форму данных до и после удаления выбросов.
print("\nФорма данных до удаления выбросов:", titanic.shape)
print("Форма данных после удаления выбросов:", cleaned_data.shape)
```

```
Форма данных до удаления выбросов: (891, 15)
Форма данных после удаления выбросов: (889, 15)
```

```
[38] # 绘制年龄的线型图
     plt.figure(figsize=(10, 6))
     sns.lineplot(data=titanic, x=titanic.index, y='age')
     plt.title("Line Plot of Age")
     plt.xlabel("Index")
     plt.ylabel("Age")
     plt.show()
```




```
# Построение гистограмм возраста
plt.figure(figsize=(10, 6))
sns.histplot(data=titanic, x='age', bins=20, kde=True)
plt.title("Histogram of Age")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```



## Дополнительные требования по группам:

- Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Дополнительное

```
[40] # В качестве примера здесь выбран тариф (fare).
     plt.figure(figsize=(10, 6))
     sns.violinplot(data=titanic['fare'])
     plt.title("Violin Plot of Fare")
     plt.xlabel("Values")
     plt.ylabel("Frequency")
     plt.show()
```