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

Рубежный контроль №1  
по дисциплине  
«Методы машинного обучения»  
на тему

# **«Методы обработки данных.»**

Выполнил:  
студент группы ИУ5и-22М  
Джин Шуо

Москва — 2024 г.

## Варианты заданий

|  |  |  |
| --- | --- | --- |
| **Номер варианта** | **Номер задачи №1** | **Номер задачи №2** |
| 16 | 16 | 36 |

**Задача №16.**

Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

**Задача №36.**

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Для студентов групп ИУ5-22М, ИУ5И-22М - для произвольной колонки данных построить гистограмму.

**Задача №1**

import pandas as pd

from scipy import stats

import matplotlib.pyplot as plt

df = pd.read\_csv('top 100 world university 2024 new.csv', encoding='ISO-8859-1')

column\_name = 'overall\_score'

data\_to\_transform = df[column\_name]

transformed\_data, lambda\_value = stats.boxcox(data\_to\_transform)

df[column\_name] = transformed\_data

plt.figure(figsize=(8, 6))

plt.hist(transformed\_data, bins=20, color='skyblue', edgecolor='black')

plt.title('Histogram of Transformed Data (Overall Score)')

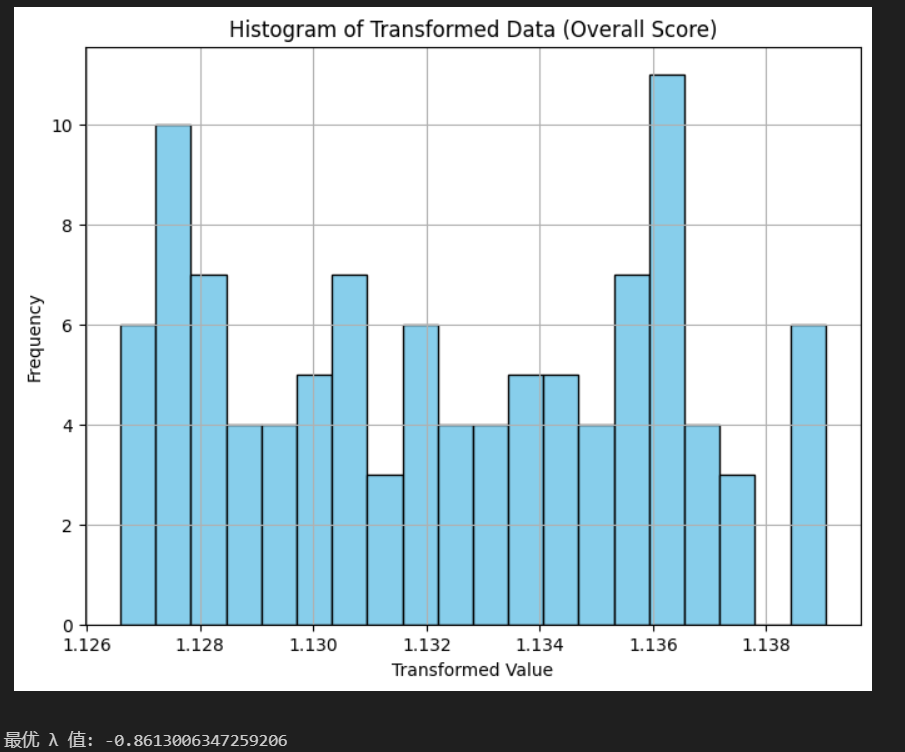
plt.xlabel('Transformed Value')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()

print("最优 λ 值:", lambda\_value)



**Задача №2**

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.feature\_selection import SelectKBest

from sklearn.feature\_selection import mutual\_info\_regression

df = pd.read\_csv('top 100 world university 2024 new.csv', encoding='ISO-8859-1')

numeric\_columns = df.select\_dtypes(include=['int64', 'float64']).columns

df\_numeric = df[numeric\_columns]

X = df[['faculty\_student\_ratio', 'academic\_reputation', 'employer\_reputation', 'employment\_outcomes', 'citations\_per\_faculty']] # 排除标识列

y = df\_numeric['overall\_score']

k\_best\_features = SelectKBest(score\_func=mutual\_info\_regression, k=5)

k\_best\_features.fit(X, y)

selected\_features\_indices = k\_best\_features.get\_support(indices=True)

selected\_features = X.columns[selected\_features\_indices]

selected\_column = 'overall\_score'

plt.figure(figsize=(8, 6))

plt.hist(df[selected\_column], bins=20, color='skyblue', edgecolor='black')

plt.title(f'Histogram of {selected\_column}')

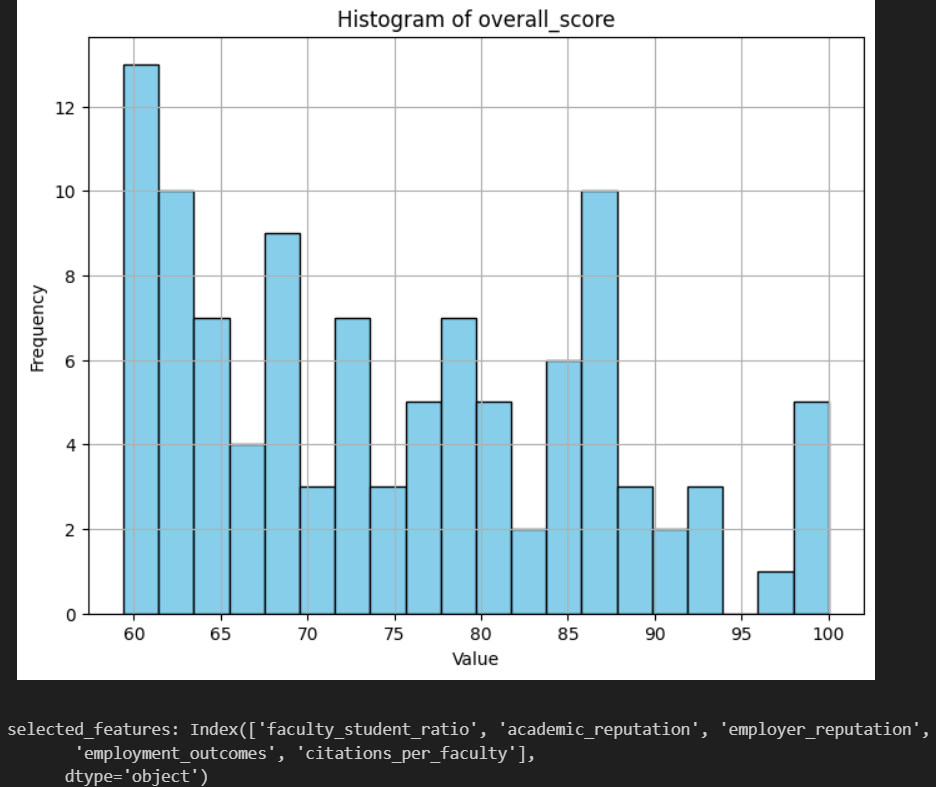
plt.xlabel('Value')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()

print("selected\_features:", selected\_features)



**Список литературы**

[1] Гапанюк Ю. Е. LAB\_MMO\_\_DATA\_STORYЛабораторная работа №1Создание "истории о данных" (Data Storytelling)// GitHub. –– 2024. –– Режим доступа:https://github.com/ugapanyuk/courses\_current/wiki/MMO\_RK\_1