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).

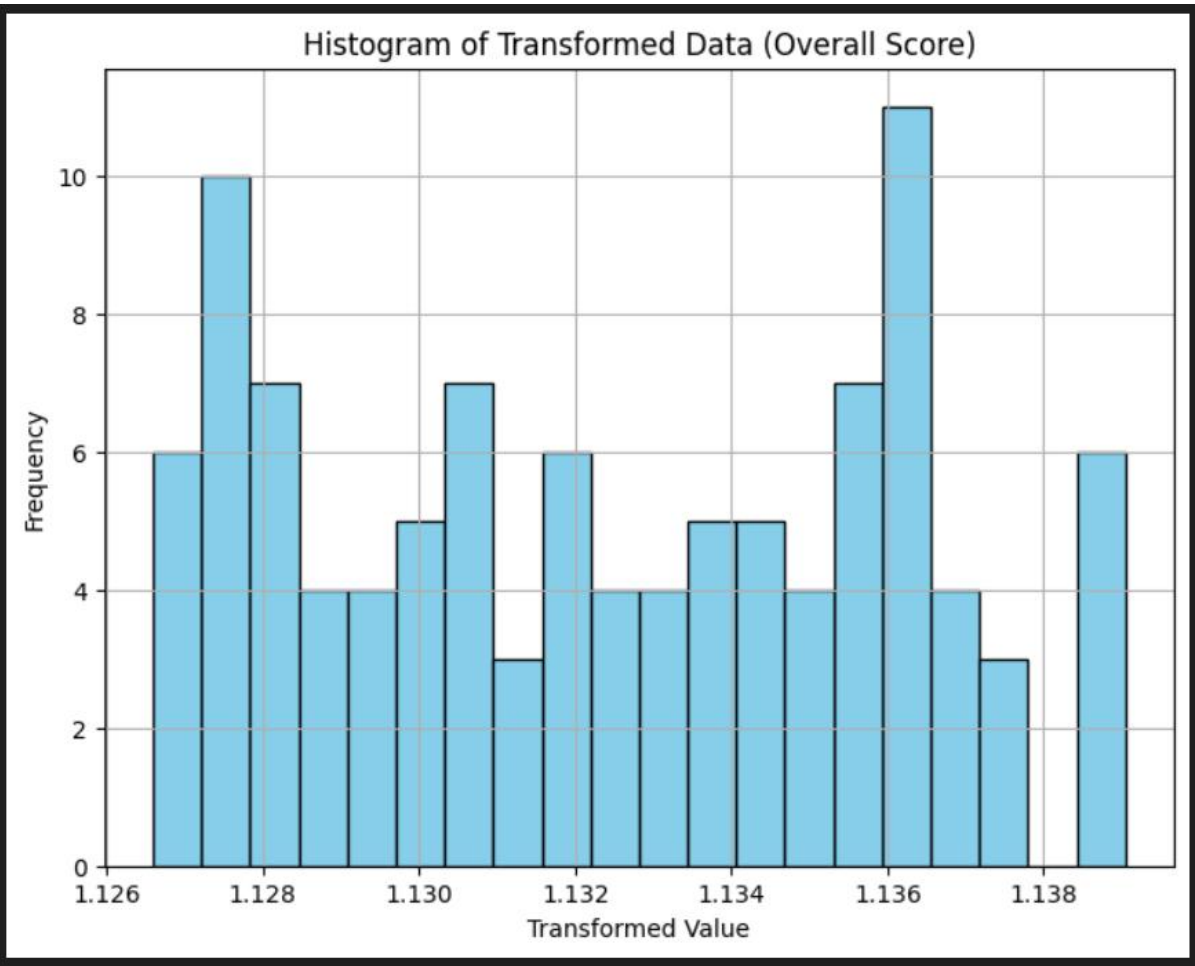
Задача №36.

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

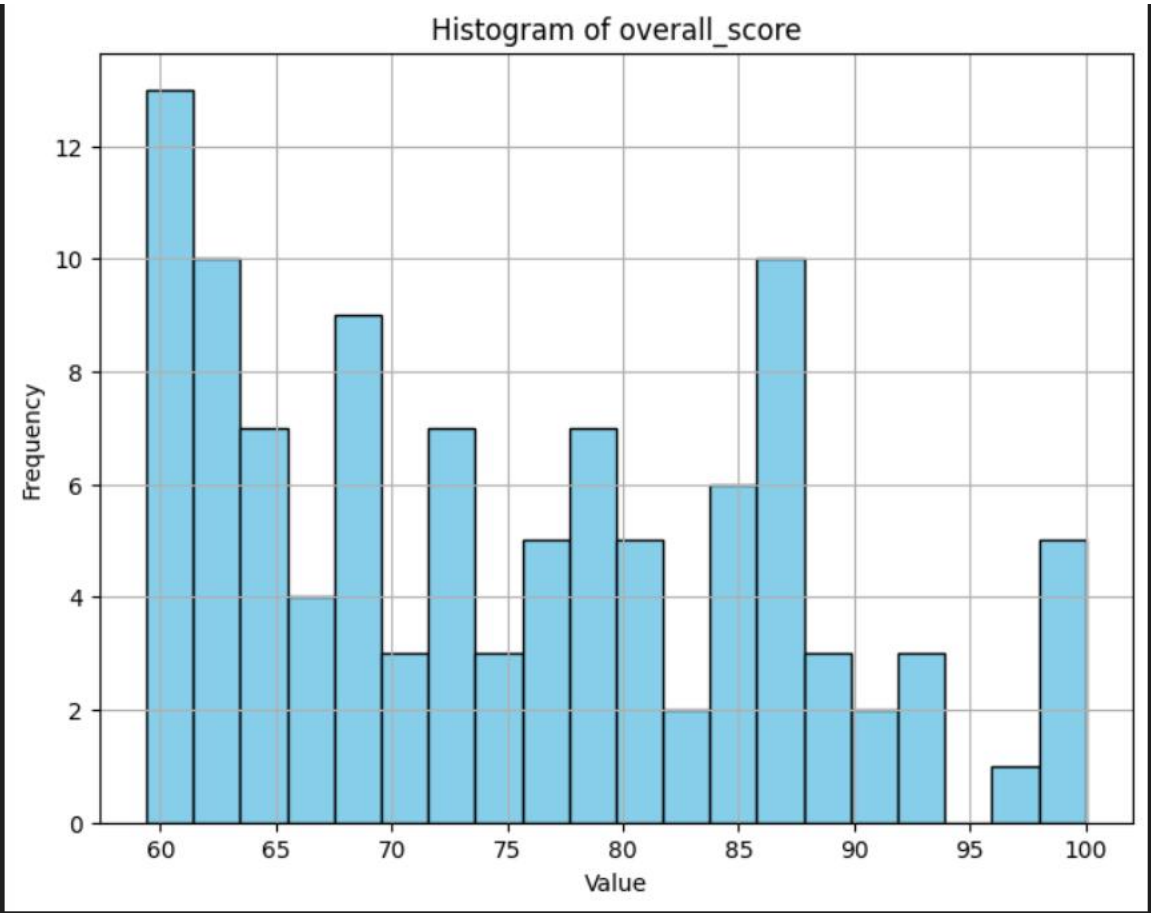
```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
df = pd.read_csv('top 100 world university 2024 new.csv', encoding='ISO-8859-1')
column_name = 'overall_score'
data_to_transform = df[column_name]
transformed_data, lambda_value = stats.boxcox(data_to_transform)
df[column_name] = 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)
```



最优 λ 值: -0.8613006347259206

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)
```




```
selected_features: Index(['faculty_student_ratio', 'academic_reputation', 'employer_reputation',
       'employment_outcomes', 'citations_per_faculty'],
      dtype='object')
```

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