****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  
Лун Сыхань

Москва — 2024 г.

## Варианты заданий

|  |  |  |
| --- | --- | --- |
| **Номер варианта** | **Номер задачи №1** | **Номер задачи №2** |
| 18 | 18 | 38 |

**Задача №18.**

Для набора данных проведите масштабирование данных для одного (произвольного) числового признака на основе Z-оценки.

**Задача №38.**

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 10 лучших признаков, и метод, основанный на взаимной информации.

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

**Задача №18**

import pandas as pd

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt

data = pd.read\_csv(r'C:\Users\Loong\Desktop\MCU Movies.csv')

numeric\_feature = "Runtime (min)"

scaler = StandardScaler()

data[numeric\_feature + "\_scaled"] = scaler.fit\_transform(data[[numeric\_feature]])

print(data[[numeric\_feature, numeric\_feature + "\_scaled"]].head())

plt.hist(data[numeric\_feature + "\_scaled"], bins=20, color='skyblue', edgecolor='black')

plt.xlabel(numeric\_feature + " (Scaled)")

plt.ylabel('Frequency')

plt.title('Histogram of Scaled ' + numeric\_feature)

plt.show()

图表, 直方图

描述已自动生成

**Задача №38**

import pandas as pd

from sklearn.preprocessing import StandardScaler

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

import matplotlib.pyplot as plt

data = pd.read\_csv(r'C:\Users\Loong\Desktop\MCU Movies.csv')

currency\_columns = ["Production Budget", "Box Office (Local)", "Box Office (International)", "Total Box Office Earnings"]

for col in currency\_columns:

data[col] = data[col].replace('[\$,]', '', regex=True).astype(float)

X = data.drop(["Movie Title", "Release Date (USA)", "Phase", "Genre", "Movie Rating", "Lead Role",

"Production Budget", "Box Office (Local)", "Box Office (International)",

"Total Box Office Earnings", "Rotten Tomatoes Ratings (%)", "IMDb Ratings"], axis=1)

y = data["Total Box Office Earnings"]

selector = SelectKBest(score\_func=mutual\_info\_regression, k=min(10, X.shape[1]))

X\_selected = selector.fit\_transform(X, y)

selected\_features = X.columns[selector.get\_support(indices=True)]

print(selected\_features)

for feature in selected\_features:

plt.hist(data[feature], bins=20, color='skyblue', edgecolor='black')

plt.xlabel(feature)

plt.ylabel('Frequency')

plt.title('Histogram of ' + feature)

plt.show()

图表, 直方图

描述已自动生成

**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