Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

ОТЧЕТ

ПО РУБЕЖНЫЙ КОНТРОЛЬ №1

ПО ДИСЦИПЛИНЕ «МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»

ВАРИАНТ 16

Студент ИУ5И-24М _____ **Сюэ Цзинъюй**
(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель _____ **Ю.Е.Гапанюк**
(Подпись, дата) (И.О.Фамилия)

2025 г.

ВВЕДЕНИЕ

Для студентов групп ИУ5И-21М, ИУ5И-22М, ИУ5И-23М, ИУ5И-24М, ИУ5И-25М номер варианта = 15 + номер в списке группы.

Каждая задача предполагает использование набора данных. Набор данных выбирается Вами произвольно с учетом следующих условий:

- Вы можете использовать один набор данных для решения всех задач, или решать каждую задачу на своем наборе данных.
- Набор данных должен отличаться от набора данных, который использовался в лекции для решения рассматриваемой задачи.
- Вы можете выбрать произвольный набор данных (например тот, который Вы использовали в лабораторных работах) или создать собственный набор данных (что актуально для некоторых задач, например, для задач удаления псевдоконстантных или повторяющихся признаков).
- Выбранный или созданный Вами набор данных должен удовлетворять условиям поставленной задачи. Например, если решается задача устранения пропусков, то набор данных должен содержать пропуски.

Полученные варианты:

- Номер варианта = 15 + 1 =16
- Номер задачи №1: 16

  Задача №16 - Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).
- Номер задачи №2: 36

  Задача №36 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 5 лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

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

ВЫХОД РАБОТЫ

1. Преобразование Бокса-Кокса

```
    # -*- coding: utf-8 -*-
import numpy as np
import pandas as pd
import os
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_selection import SelectKBest, mutual_info_classif
plt.rcParams['font.sans-serif'] = ['Microsoft YaHei']

# 生成示例数据
file_path = os.path.abspath('C:/Users/xue_j/Desktop/2024-2025-2/MMO/PK1/data
1.csv')  # Windows
data = pd.read_csv(file_path)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
print(data.head(10))
```

Вывести первые десять строк данных:

	day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
0	1	6	2012	29	57	18	0.0	65.7	3.4	7.6	1.3	3.4	0.5	notfire
1	2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1.0	3.9	0.4	notfire
2	3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	notfire
3	4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0.0	1.7	0.0	notfire
4	5	6	2012	27	77	16	0.0	64.8	3.0	14.2	1.2	3.9	0.5	notfire
5	6	6	2012	31	67	14	0.0	82.6	5.8	22.2	3.1	7.0	2.5	fire
6	7	6	2012	33	54	13	0.0	88.2	9.9	30.5	6.4	10.9	7.2	fire
7	8	6	2012	30	73	15	0.0	86.6	12.1	38.3	5.6	13.5	7.1	fire
8	9	6	2012	25	88	13	0.2	52.9	7.9	38.8	0.4	10.5	0.3	notfire
9	10	6	2012	28	79	12	0.0	73.2	9.5	46.3	1.3	12.6	0.9	notfire

Рис.1 Первые десять строк выборочных данных

Выберите столбец «FFMC» для преобразования boxcox и выведите оптимальные параметры:

```
    # 应用 Box-Cox 变换
data['FFMC_boxcox'], lambda_param = stats.boxcox(data['FFMC'])
print(f"best_λ: {lambda_param:.3f}")
```

Выход:

```
best_λ: 4.019
```

Сравните изменения данных «FFMC» до и после нормализации (гистограмма):

```
# 变换前的分布检查
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.hist(data['FFMC'], bins=30, color='blue', alpha=0.7)
plt.title('generation')

# 变换后的分布检查
plt.figure(1)
plt.subplot(1, 2, 2)
plt.hist(data['FFMC_boxcox'], bins=30, color='green', alpha=0.7)
plt.title('Box-Cox')
```

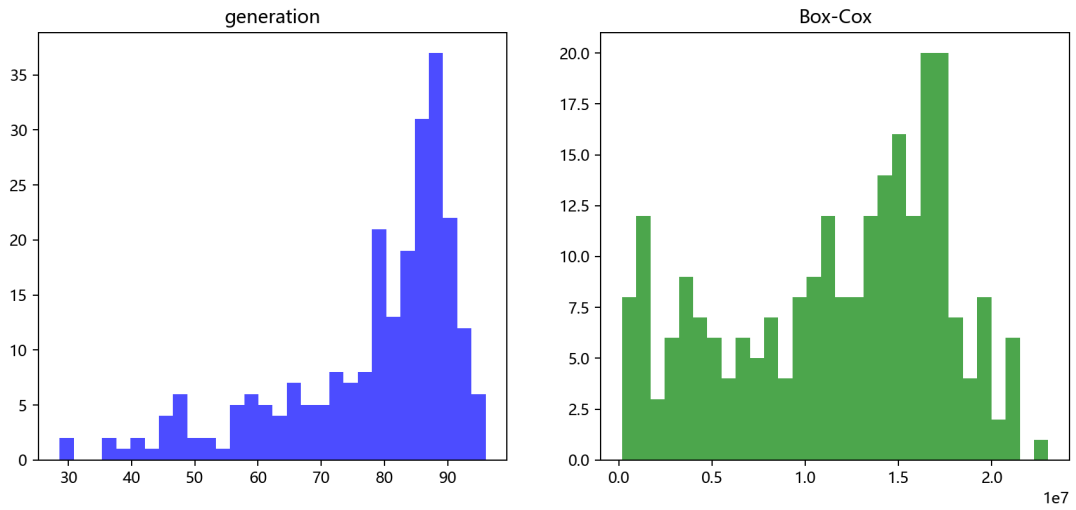

Рис. 2 Сравнение гистограмм

Сравните изменения данных «FFMC» до и после нормализации (violin plot график):

```
    # 垂直小提琴图（针对'数值列'）
plt.figure(2)
sns.violinplot(data=data, y='FFMC')
plt.title('小提琴图 (Violin Plot)_generation')
plt.ylabel('FFMC')

plt.figure(3)
sns.violinplot(data=data, y='FFMC_boxcox')
plt.title('小提琴图 (Violin Plot)_boxcox')
plt.ylabel('FFMC_boxcox')
```

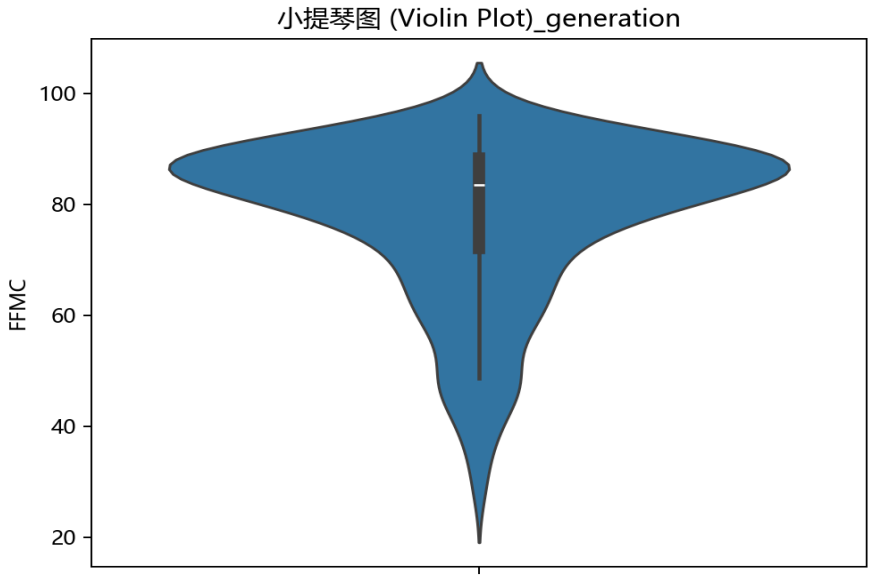

Рис. 3 “FFMC”violin plot

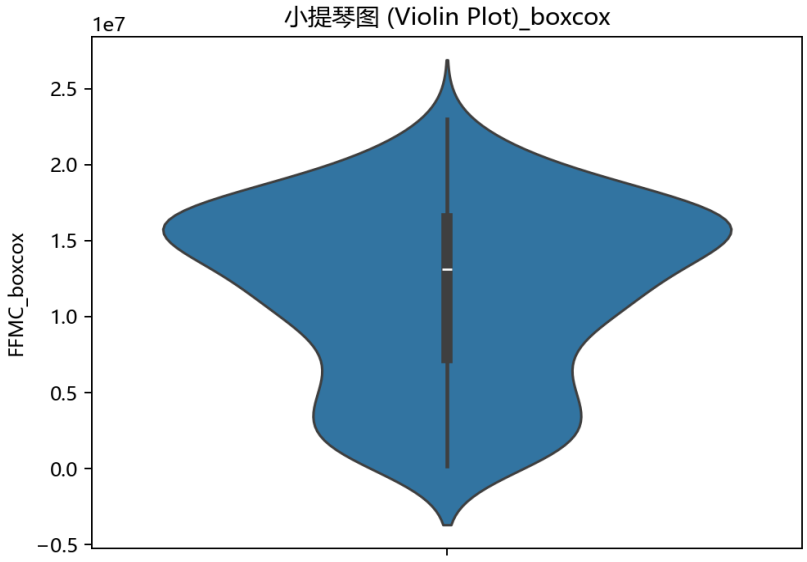

Рис. 4 “FFMC_boxcox”violin plot

Начертите график скрипки по столбцу классификации «классы»:

```
    #按类划分
plt.figure(4)
sns.violinplot(data=data, y='FFMC',x='Classes')
plt.title('小提琴图 (Violin Plot)_classes')
plt.ylabel('FFMC')
plt.xlabel('Classes')
plt.figure(5)
sns.violinplot(data=data, y='FFMC_boxcox',x='Classes')
plt.title('小提琴图 (Violin Plot)_boxcox_classes')
plt.xlabel('Classes')
plt.ylabel('FFMC_boxcox')
plt.show()
```

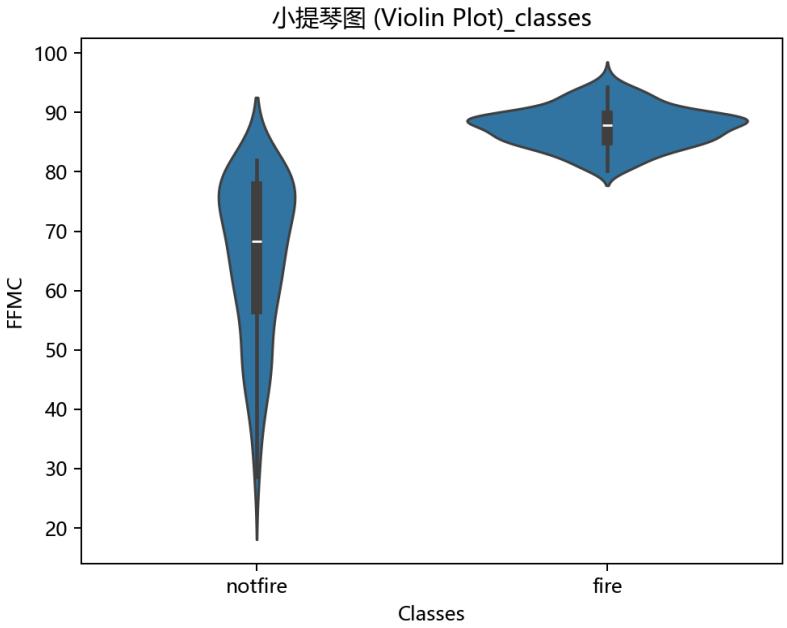

Рис. 5 “FFMC”violin plot(classes)

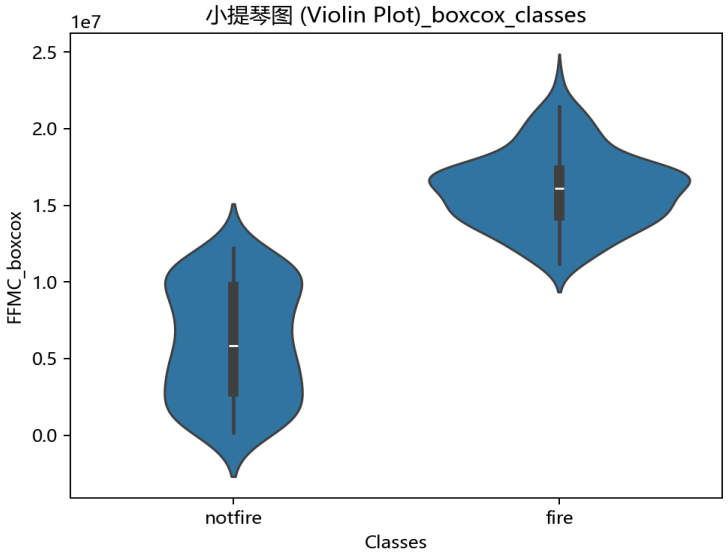

Рис. 6 “FFMC_boxcox”violin plot(classes)

3.2. Процедура отбора признаков

Используем метод SelectKBest с mutual_info_classif, чтобы выбрать 5 наиболее информативных признаков для предсказания класса (Classes — fine/notfine). Данные содержат метеорологические показатели и индексы пожароопасности.

(1) Подготовка данных:

```
# Преобразуем целевой признак в числовой формат (0 для 'notfine', 1 для 'fine
')
data['Classes'] = data['Classes'].map({'notfire': 0, 'fire': 1})

# Разделяем данные на признаки (X) и целевую переменную (y)
X = data.drop(['day', 'month', 'year', 'Classes'], axis=1)  # Исключаем даты
и целевой признак
y = data['Classes']
```

(2) Отбор 5 лучших признаков:

```
# Инициализация SelectKBest с mutual_info_classif
selector = SelectKBest(score_func=mutual_info_classif, k=5)
X_selected = selector.fit_transform(X, y)

# Получение имен выбранных признаков
selected_features = X.columns[selector.get_support()]
print("Лучшие 5 признаков:\n", selected_features.tolist())
```

Выход:

```
Лучшие 5 признаков:
 ['FFMC', 'DMC', 'ISI', 'BUI', 'FWI']
```

(3) Визуализация значимости признаков:

```
    # Оценки важности всех признаков
scores = selector.scores_

plt.figure(figsize=(10, 5))
plt.bar(X.columns, scores, color='skyblue')
plt.xticks(rotation=45, ha='right')
plt.title("Важность признаков (Mutual Information)")
plt.ylabel("Score")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

Выход:

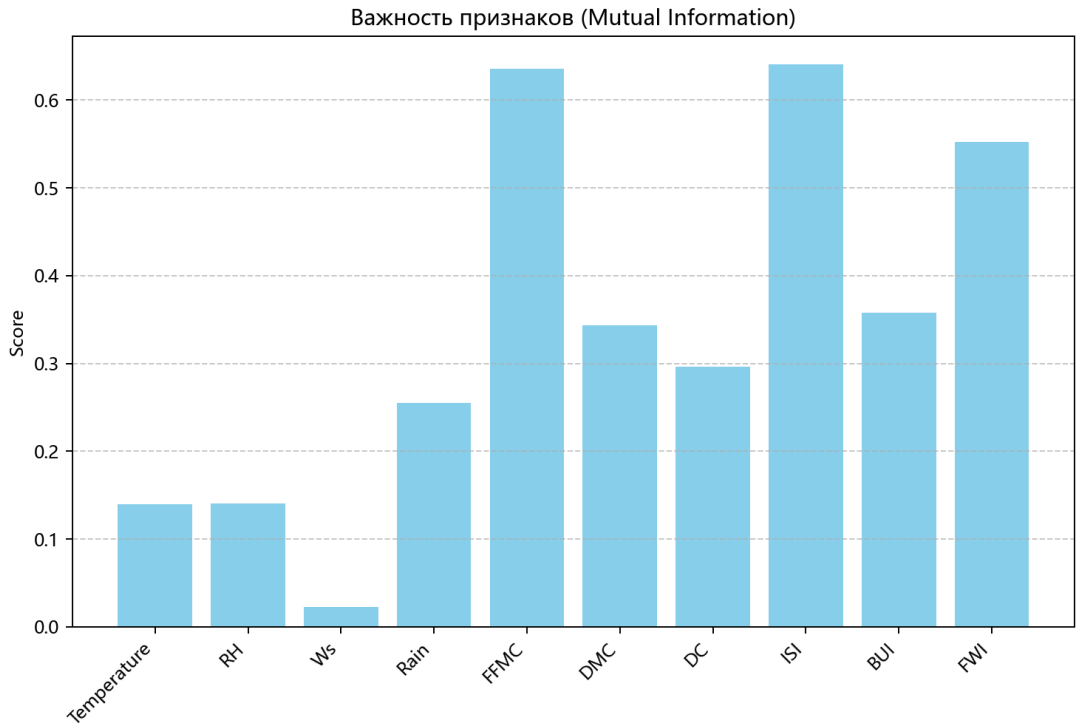

Рис. 7 Оценки важности всех признаков

Интерпретация:

Признаки FFMC, DMC, DC, ISI, FWI (индексы пожароопасности) имеют наибольшую взаимную информацию с целевой переменной Classes.

(4) Проверка отобранных данных:

```
print("Исходные признаки:\n", X.columns.tolist())
print("\nОтобранные признаки:\n", selected_features.tolist())
print("\nПример преобразованных данных (первые 5 строк):\n", X_selected[:5])
```

Выход:

```
Исходные признаки:
 ['Temperature', 'RH', 'Ws', 'Rain', 'FFMC', 'DMC', 'DC', 'ISI', 'BUI', 'FWI']

Отобранные признаки:
 ['FFMC', 'DMC', 'ISI', 'BUI', 'FWI']

Пример преобразованных данных (первые 5 строк):
 [[65.7  3.4  1.3  3.4  0.5]
 [64.4  4.1  1.   3.9  0.4]
 [47.1  2.5  0.3  2.7  0.1]
 [28.6  1.3  0.   1.7  0. ]
 [64.8  3.   1.2  3.9  0.5]]
```

Рис. 7 Проверка отобранных данных

ЗАКЛЮЧЕНИЕ

В ходе выполнения итогового контроля №1 по дисциплине «Методы машинного обучения» была выполнена сложная работа по обработке и анализу данных двух различных массивов, связанных с лесными пожарами в Алжире.

В задании № 16 нормализация данных с использованием преобразования Бокса-Кокса была успешно применена к числовому признаку «FFMC» (код влажности мелкодисперсного топлива) в наборе данных data1.csv. Уменьшить асимметрию распределения данных и приблизить данные к нормальному распределению (что соответствует предположениям многих статистических методов и моделей машинного обучения). Гистограмма и графики скрипки подтверждают, что распределение преобразованных данных значительно улучшено, а параметр λ определяется путем автоматической оптимизации.

Для задачи № 36 была выполнена процедура выбора признаков на наборе данных с использованием класса SelectKBest и метода, основанного на взаимной информации. В результате были выявлены пять наиболее важных признаков для прогнозирования лесных пожаров, которые продемонстрировали эффективность данных методов в задаче прогнозирования и наглядно продемонстрировали оценки важности признаков. Упрощение модели путем исключения менее важных характеристик может улучшить ее производительность и интерпретируемость. Этот метод позволяет фиксировать взаимодействие между признаками и является более эффективным с вычислительной точки зрения, чем методы фильтрации.

Для признака FFMC был построен график скрипки, объединяющий диаграмму ящиков (показывающую медиану и межквартильный размах); график оценки плотности ядра (показывающий схему распределения данных). Анализ показал, что различные категории (например, пожар/не пожар) демонстрируют

бимодальное распределение; выбросы и различия в плотности данных между группами были четко идентифицированы.