|  |  |
| --- | --- |
|  | **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 ДИСЦ******ИПЛИНЕ***

***«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/data1.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：

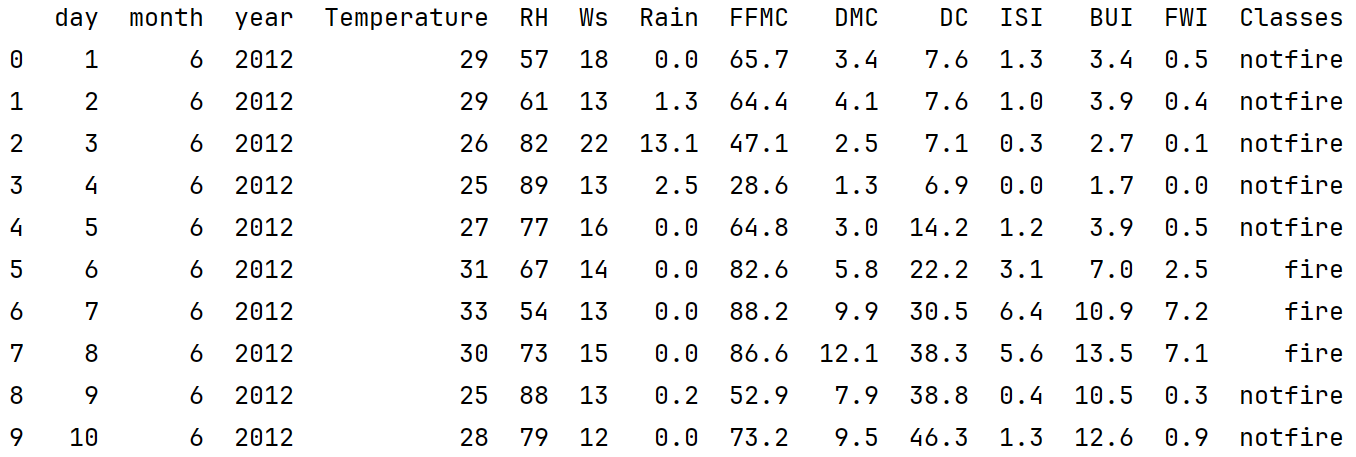


Рис.1 Первые десять строк выборочных данных

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

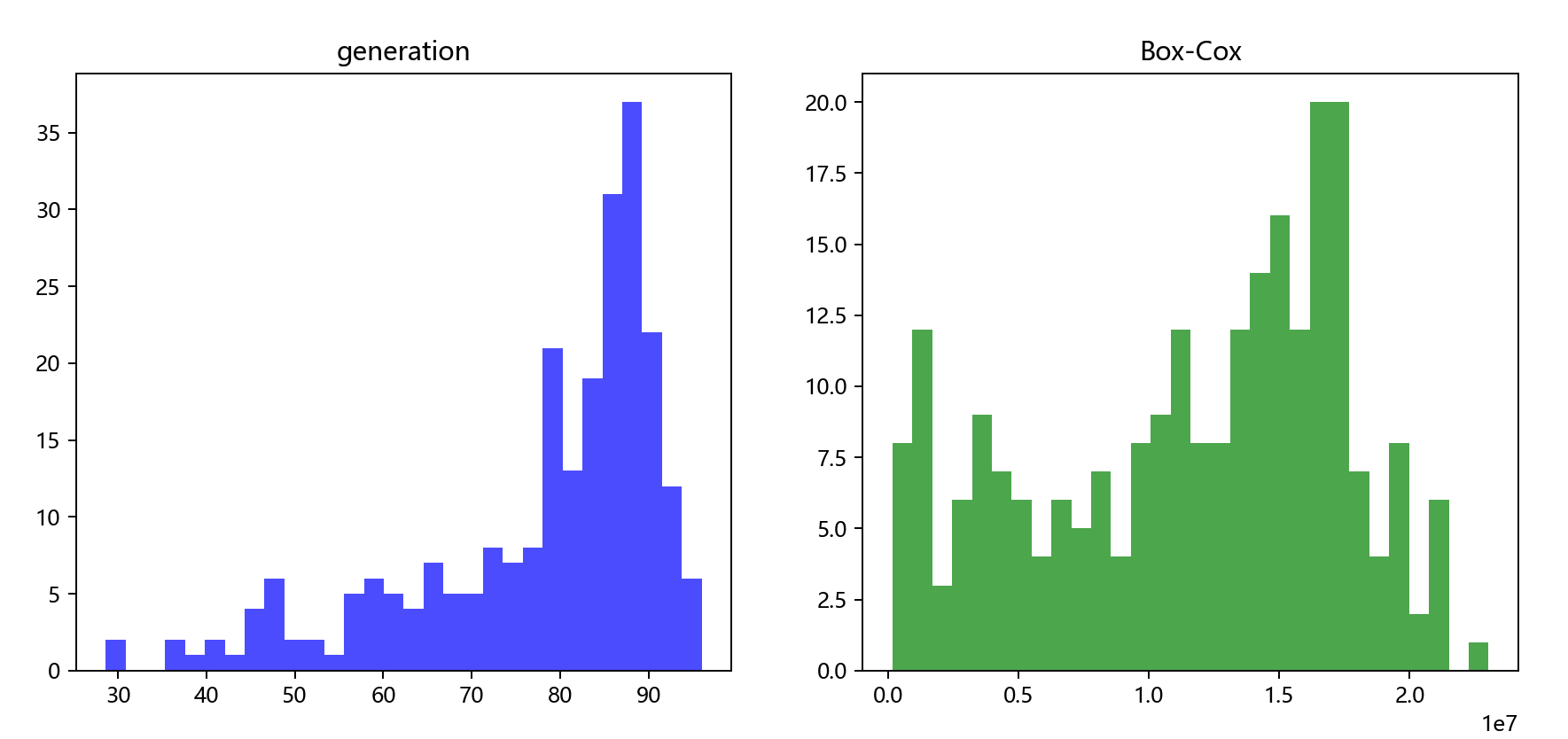
*# 应用Box-Cox变换*  
data['FFMC\_boxcox'], lambda\_param = stats.boxcox(data['FFMC'])  
print(f"best\_λ: {lambda\_param:.3f}")

Выход:

image

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')

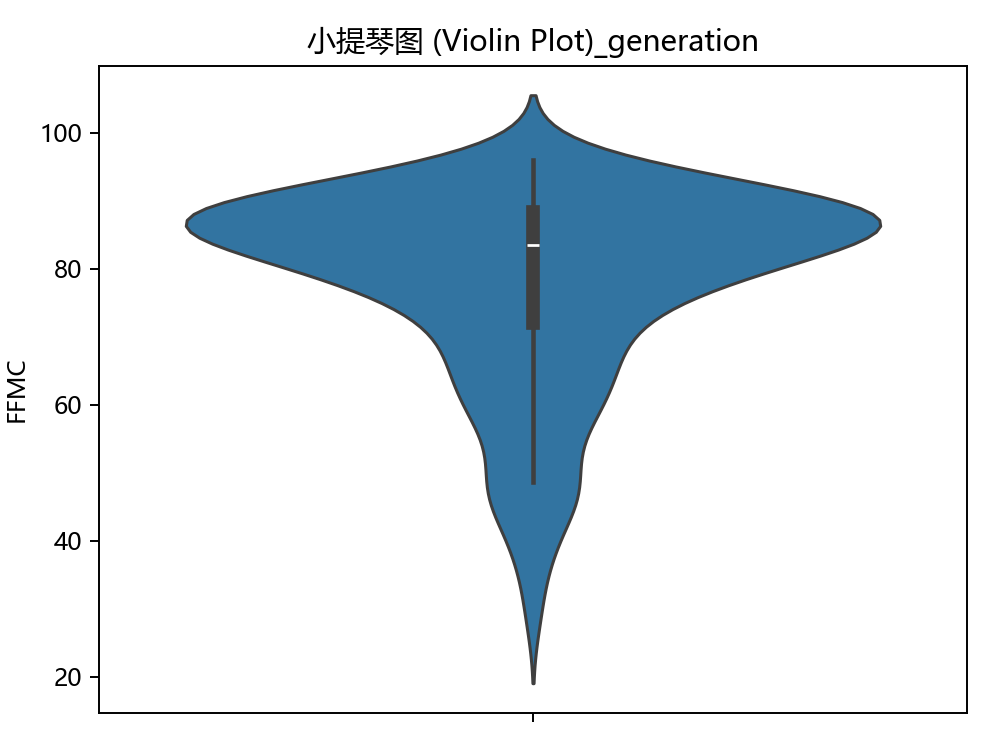


Рис. 3 “FFMC”violin plot

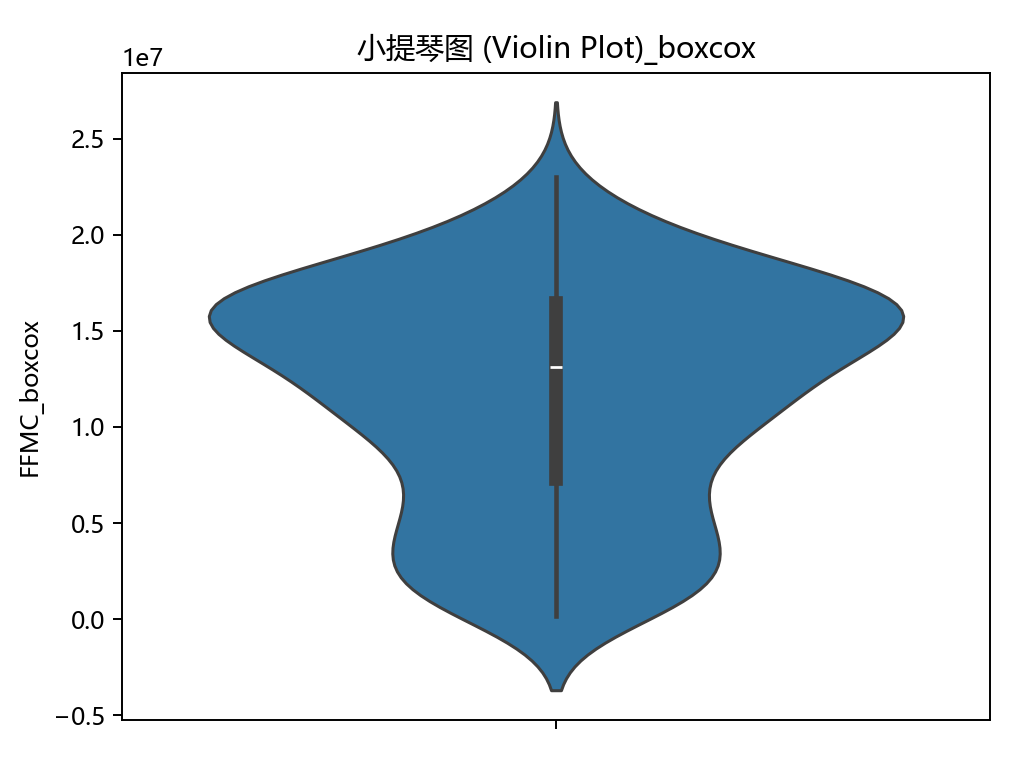


Рис. 4 “FFMC\_boxcox”violin plot

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)

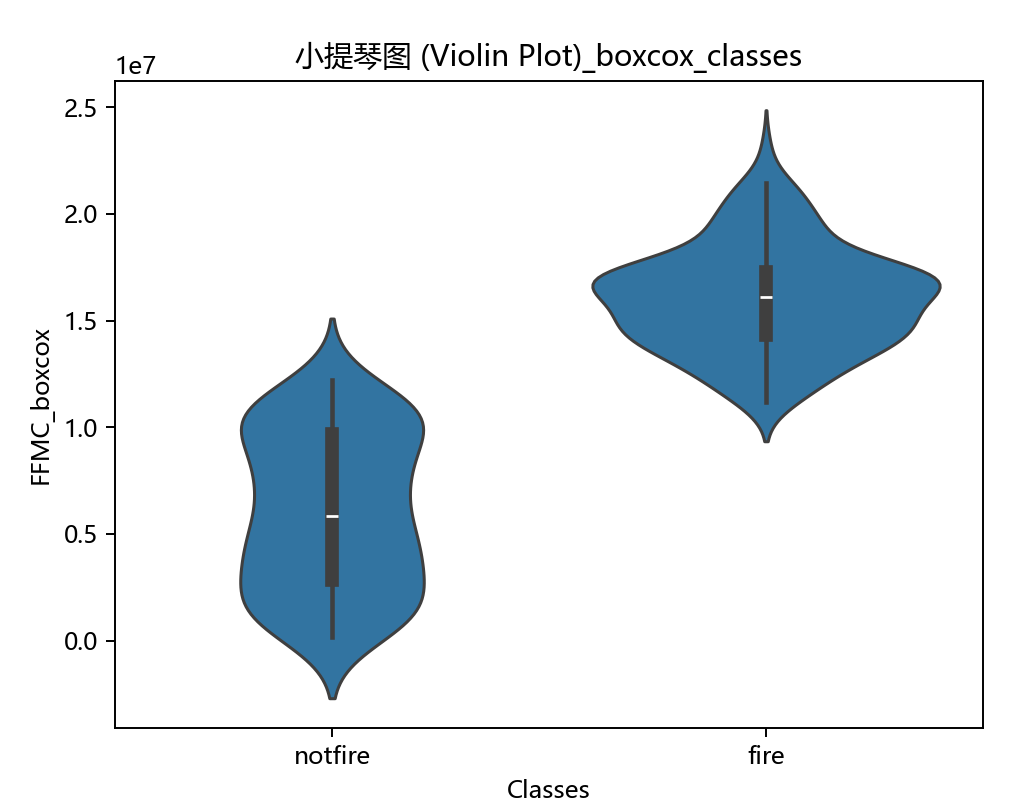


Рис. 6 “FFMC\_boxcox”violin plot(classes)

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：**



(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：

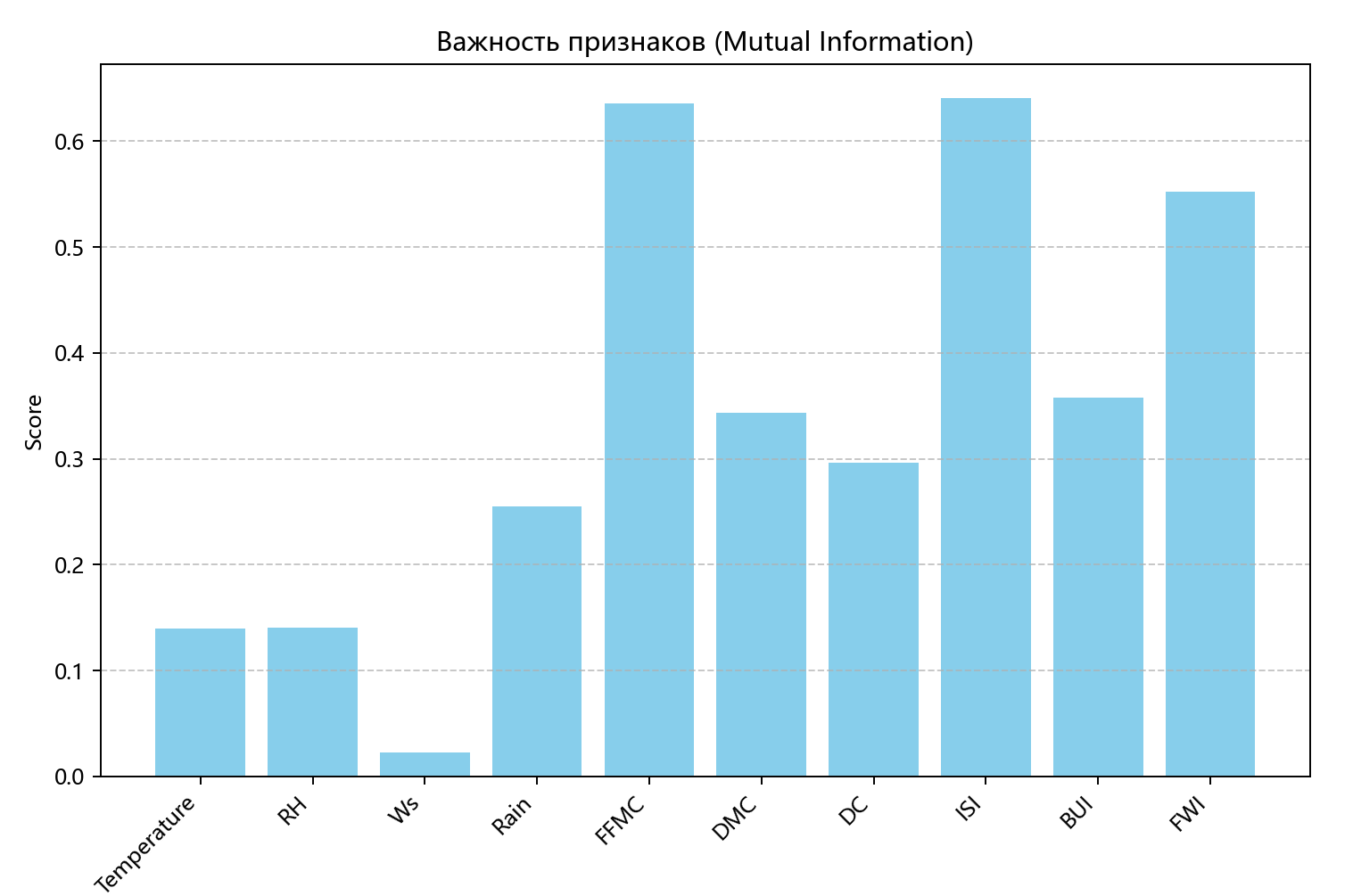


Рис. 7 Оценки важности всех признаков

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：

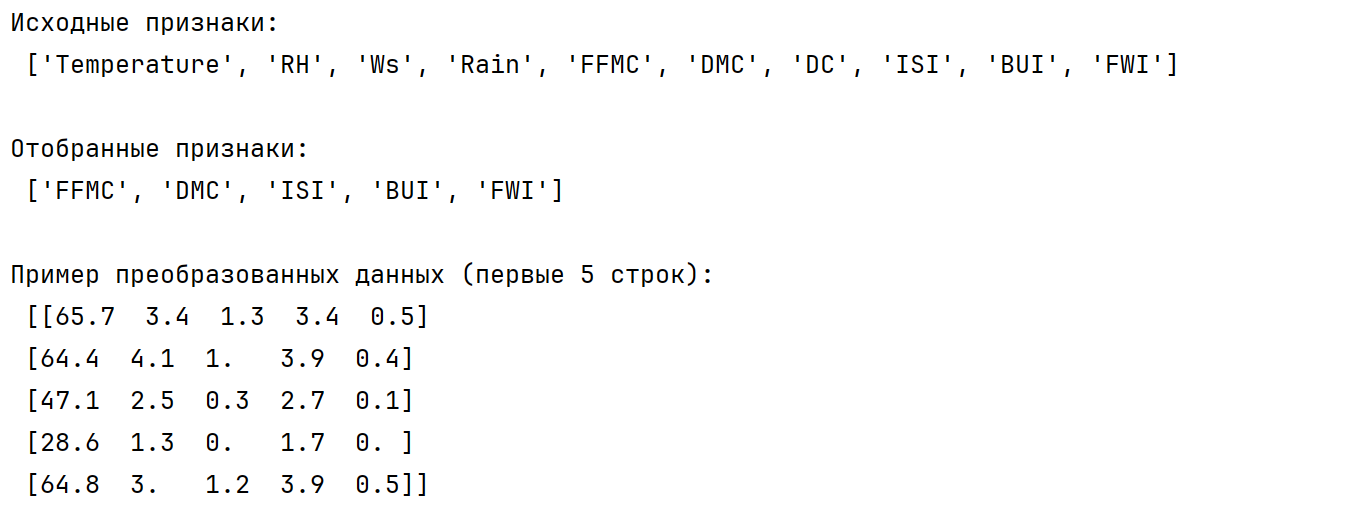


Рис. 7 Проверка отобранных данных

**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.