|  |  |
| --- | --- |
|  | **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 19***

Студент \_\_ИУ5И-21М\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Энькаэр Уэркэнь\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Ю.Е.Гапанюк\_\_**

(Подпись, дата) (И.О.Фамилия)

2025 г.

**ВВЕДЕНИЕ**

Для студентов групп ИУ5-21М, ИУ5-22М, ИУ5-23М, ИУ5-24М, ИУ5-25М номер варианта = номер в списке группы.

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

Для студентов групп ИУ5-25МВ номер варианта = 20 + номер в списке группы.

Дополнительные требования по группам:

* Для студентов групп ИУ5-21М, ИУ5И-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".
* Для студентов групп ИУ5-22М, ИУ5И-22М - для произвольной колонки данных построить гистограмму.
* Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".
* Для студентов группы ИУ5-24М, ИУ5И-24М - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".
* Для студентов группы ИУ5-25М, ИУ5И-25М, ИУ5-25МВ - для произвольной колонки данных построить парные диаграммы (pairplot).

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 + 4 =19
* Номер задачи №1: 19

Задача №19 - Для набора данных проведите масштабирование данных для одного (произвольного) числового признака с использованием метода "Mean Normalisation".

* Номер задачи №2: 39

Задача №39 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectPercentile для 10% лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

* Для студентов групп ИУ5-21М, ИУ5И-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

**ХОД ВЫПОЛНЕНИЯ РАБОТЫ**

# Часть 1. Текстовое описание набора данных

# Описание атрибутов набора данных forestfires:

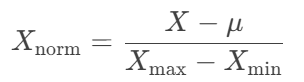
Таблица 1 – описание атрибутов набора данных forestfires

|  |  |
| --- | --- |
| Название атрибута | Описание |
| X | X axis spatial coordinate within the Montesinho park map: 1 to 9 |
| Y | Y axis spatial coordinate within the Montesinho park map: 2 to 9 |
| month | month of the year: "jan" to "dec" |
| day | day of the week: "mon" to "sun" |
| FFMC | FFMC index from the FWI system: 18.7 to 96.20 |
| DMC | DMC index from the FWI system: 1.1 to 291.3 |
| DC | DC index from the FWI system: 7.9 to 860.6 |
| ISI | ISI index from the FWI system: 0.0 to 56.10 |
| temp | temperature in Celsius degrees: 2.2 to 33.30 |
| RH | relative humidity in %: 15.0 to 100 |
| wind | wind speed in km/h: 0.40 to 9.40 |
| rain | outside rain in mm/m2 : 0.0 to 6.4 |
| area | the burned area of the forest (in ha): 0.00 to 1090.84 |

Датасет содержит 517 строк данных, в котором есть 4 дублирующихся значения. После удаления дубликатов осталось 513 строк, которые сохранены в формате CSV.

# Часть 2. Задача №19

Для числового признака **temp** (температура) выполнено масштабирование методом Mean Normalisation. Формула:



где:

X — исходное значение признака,

μ — среднее значение признака,

Xmaxи Xmin — максимальное и минимальное значения признака.

# Код для масштабирования признака temp

import pandas as pd

# Загрузка данных

df = pd.read\_csv(r"C:\Users\86188\Desktop\23forestfires.csv")

# Выбор признака temp

temp = df['temp'].values.reshape(-1, 1)

# Вычисление параметров

mean\_temp = df['temp'].mean()

min\_temp = df['temp'].min()

max\_temp = df['temp'].max()

# Применение Mean Normalisation

df['temp\_mean\_norm'] = (df['temp'] - mean\_temp) / (max\_temp - min\_temp)

# Вывод первых 5 строк

print(df[['temp', 'temp\_mean\_norm']].head())

OUTPUT:

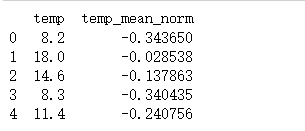


Рисунок 1: Нормализованные результаты

Результат:  
Столбец temp масштабирован, новые значения находятся в диапазоне [-0.5, 0.5].

# 

# Часть 3. Задача №39

Для выбора 10% наиболее значимых признаков использован метод **SelectPercentile** с метрикой **mutual\_info\_regression**, так как целевая переменная area является числовой.

# Код для отбора признаков

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

from sklearn.preprocessing import LabelEncoder

# Преобразование категориальных признаков

le = LabelEncoder()

df['month'] = le.fit\_transform(df['month'])

df['day'] = le.fit\_transform(df['day'])

# Разделение данных на признаки и целевую переменную

X = df.drop('area', axis=1)

y = df['area']

# Отбор признаков

selector = SelectPercentile(mutual\_info\_regression, percentile=10)

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

# Вывод выбранных признаков

selected\_mask = selector.get\_support()

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

print(f'Выбранные признаки: {selected\_features.tolist()}')

OUTPUT:

Выбранные признаки: ['month', 'DC']

# Часть 4. Дополнительные требования

Для визуализации взаимосвязи между температурой (temp) и относительной влажностью (RH) построена диаграмма рассеяния.

# Код для построения графика

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))

plt.scatter(df['temp'], df['RH'], alpha=0.5, c='red')

plt.title('Зависимость влажности от температуры')

plt.xlabel('Температура (°C)')

plt.ylabel('Относительная влажность (%)')

plt.grid(True)

plt.show()

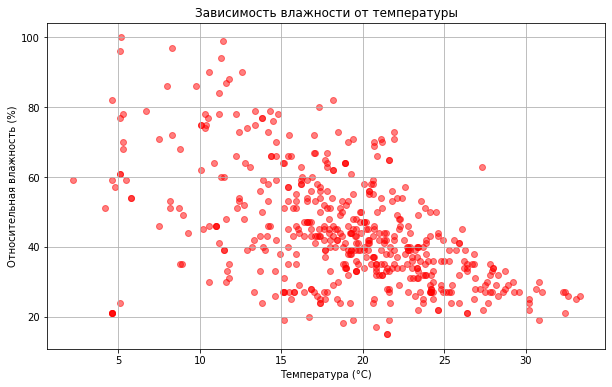


Рисунок 2:Зависимость влажности от температуры

График показывает слабую отрицательную корреляцию: при повышении температуры влажность tends to decrease.

**ЗАКЛЮЧЕНИЕ**

Масштабирование: Признак temp нормализован для улучшения работы алгоритмов машинного обучения.

Отбор признаков: Методом взаимной информации выбраны 2 ключевых признака: month и DC.

Визуализация: Диаграмма рассеяния между температурой и влажностью выявила тенденцию к обратной зависимости.