**Московский Государственный Технический Университет имениН.Э.Баумана**

**Факультет Информатика и системы управления**

**Кафедра ИУ-5**

**«Системы обработки информации и управления»**



**Рубежный контроль по дисциплине**

**«Методы машинного обучения»**

**Тема: Методы обработки данных**

Студент: Лу Жуньда

Группа:ИУ5И-22М

**Москва 2024г**

**Варианты заданий**

номер варианта = 2+15=17

Номер задачи №1: 17

Номер задачи №2: 37

**Задача №17.**

Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Йео-Джонсона (Yeo-Johnson transformation).

**Задача №37.**

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectPercentile для 5% лучших признаков, и метод, основанный на взаимной информации.

**Дополнительные требования по группам:**

Для произвольной колонки данных построить гистограмму.

import pandas as pd

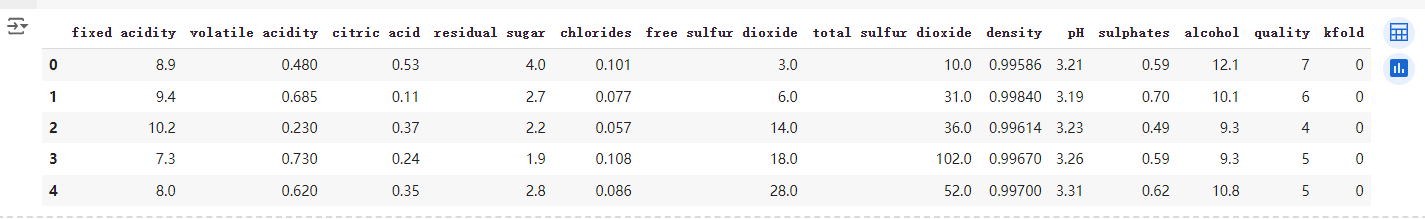
# Load the dataset

file\_path = '/content/winequality-red-folds.csv'

data = pd.read\_csv(file\_path)

# Display the first few rows of the dataset to understand its structure

data.head()



# Задача №17: Нормализация с использованием преобразования Йео-Джонсона

Для этого выберем произвольный числовой признак. Пусть это будет колонка fixed acidity

from sklearn.preprocessing import PowerTransformer

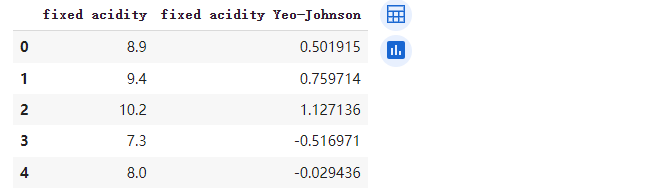
# Преобразование Йео-Джонсона для колонки 'fixed acidity'

yeo\_transformer = PowerTransformer(method='yeo-johnson')

data['fixed acidity Yeo-Johnson'] = yeo\_transformer.fit\_transform(data[['fixed acidity']])

# Сравним исходное и преобразованное значение

data[['fixed acidity', 'fixed acidity Yeo-Johnson']].head()



# Задача №37: Отбор признаков с использованием класса SelectPercentile и метода на основе взаимной информации

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

# Отбор признаков

X = data.drop(columns=['quality', 'kfold'])

y = data['quality']

selector = SelectPercentile(mutual\_info\_regression, percentile=5)

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

# Выбранные признаки

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

# Вывод выбранных признаков

print(selected\_features)



Используя класс SelectPercentile и метод на основе взаимной информации, был выбран признак alcohol как один из лучших 5% признаков.

**Дополнительное требование: Построение гистограммы**

Построим гистограмму для произвольной колонки данных. Пусть это будет колонка alcohol.

import matplotlib.pyplot as plt

# Построение гистограммы

plt.hist(data['alcohol'], bins=10, edgecolor='black')

plt.title('Histogram of Alcohol Content')

plt.xlabel('Alcohol')

plt.ylabel('Frequency')

plt.show()

