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 18

Студент ИУ5И-23М _____ **Чжан Мяо**
(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 + 3 =18
- Номер задачи №1: 18

  Задача №18 - Для набора данных проведите масштабирование данных для одного (произвольного) числового признака на основе Z-оценки.
- Номер задачи №2: 38

  Задача №38 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 10 лучших признаков, и метод, основанный на взаимной информации.

Дополнительные требования по группам:

- Для студентов групп ИУ5-23М, ИУ5И-23М - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

ХОД ВЫПОЛНЕНИЯ РАБОТЫ

Часть 1. Задача №18

Задача №18 - Для набора данных проведите масштабирование данных для одного (произвольного) числового признака на основе Z-оценки.

```python
# ------------------- 1. 基于 Z 分数的数据标准化 -------------------
# 选择一个任意的数值特征（例如第一个列）
feature_to_scale = data.columns[0]

# 使用 Z 分数进行标准化
scaler = StandardScaler()
scaled_feature = scaler.fit_transform(data[[feature_to_scale]])

# 可视化原始特征和标准化后的特征分布
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.hist(data[feature_to_scale], bins=20, color='blue', alpha=0.7)
plt.title(f'Histogram of raw features:\n{feature_to_scale}')
plt.xlabel(feature_to_scale)
plt.ylabel('frequency')

plt.subplot(1, 2, 2)
plt.hist(scaled_feature, bins=20, color='green', alpha=0.7)
plt.title(f'Histogram of standardised (Z-score) features:\n{feature_to_scale}')
plt.xlabel(f'{feature_to_scale} (standardisation)')
plt.ylabel('frequency')
plt.tight_layout()
plt.show()
```

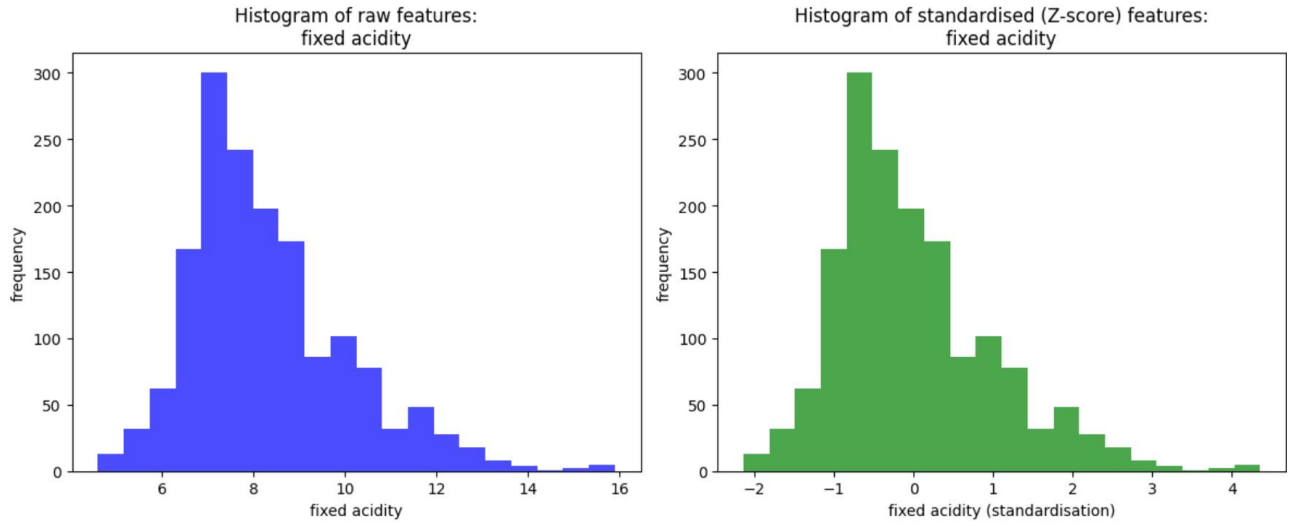


Рисунок 1

Часть 2. Задача №38

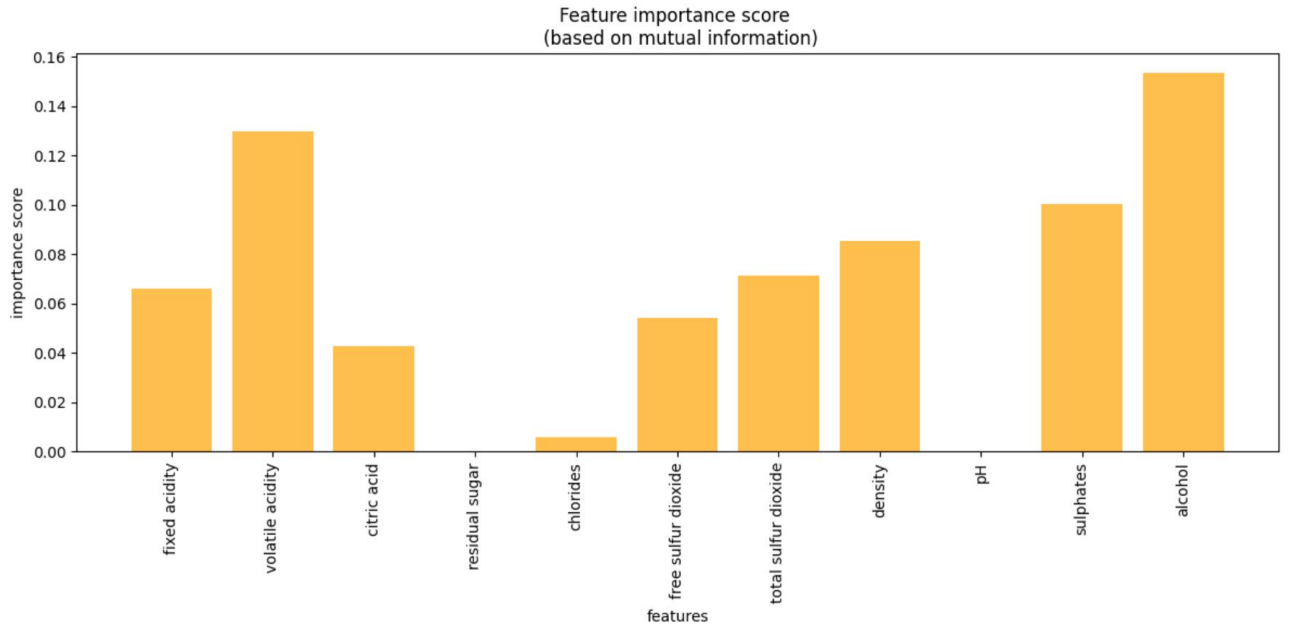
Задача №38 - Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectKBest для 10 лучших признаков, и метод, основанный на взаимной информации.

```python
# -------------------2. 使用 SelectKBest 进行特征选择-------------------
# 使用互信息法选择最好的 10 个特征
selector = SelectKBest(score_func=mutual_info_regression, k=10)
X_new = selector.fit_transform(data.drop(columns=['quality']), data['quality'])

# 获取被选中的特征名称
selected_features = data.drop(columns=['quality']).columns[selector.get_support()]

# 可视化特征重要性得分
scores = selector.scores_
plt.figure(figsize=(12, 6))
plt.bar(range(len(scores)), scores, color='orange', alpha=0.7)
plt.xticks(range(len(scores)), data.drop(columns=['quality']).columns, rotation=90)
plt.title('Feature importance score \n (based on mutual information)')
plt.xlabel('features')
plt.ylabel('importance score')
plt.tight_layout()
plt.show()

print(f"Best features selected: {list(selected_features)}")
```



OUTPUT:

Best features selected: ['fixed acidity', 'volatile acidity', 'citric acid', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol']

Часть 3. Дополнительные требования

для произвольной колонки данных построить график "Ящик с усами (boxplot)".

```
# resting blood pressure 的分布箱线图
plt.figure(figsize=(10, 6))
plt.boxplot(df['trestbps'].dropna())
plt.title('resting blood pressure  Distribution')
plt.xticks([1], ['trestbps'])
plt.ylabel('target')
plt.grid(True)
plt.show()
```