

**Московский государственный технический
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 18

Выполнил:
Файзуллин К.Х.
группа ИУ5-64Б

Проверил:
Гапанюк Ю.Е.

Дата: 13.06.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

Задание:

Номер варианта: **18**

Номер задачи: **3**

Номер набора данных, указанного в задаче: **2** (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine)

Для студентов группы ИУ5-64Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

Задача №3.

Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

1. Введение

В рамках рубежного контроля была проведена работа с набором данных Wine Dataset. Целью работы являлось масштабирование данных и преобразование категориальных признаков.

2. Описание исходных данных

Набор данных **Wine** из библиотеки scikit-learn содержит результаты химического анализа 178 образцов итальянских вин, относящихся к трем различным классам (**class_0**, **class_1**, **class_2**). Каждый образец характеризуется 13 количественными признаками, включая содержание алкоголя, яблочной кислоты, флавоноидов, цветовой интенсивности и других химических

показателей. Целевая переменная (**target**) представляет категориальный признак с метками классов вин (0, 1, 2), что делает этот набор данных типичным примером задачи многоклассовой классификации. Данные не содержат пропущенных значений, все признаки являются числовыми и имеют разный масштаб, что требует предварительной стандартизации для многих алгоритмов машинного обучения.

3. Ход выполнения:

1) Загрузка и первичный анализ данных

```
[5]: from sklearn.datasets import load_wine
from sklearn.preprocessing import StandardScaler, LabelEncoder, OneHotEncoder
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Загрузка данных
data = load_wine()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = data.target # Добавляем категориальный признак (классы вина)
df.head()
```

```
[5]:
```

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target	
0	14.23	1.71	2.43		15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.0	0
1	13.20	1.78	2.14		11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.0	0
2	13.16	2.36	2.67		18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.0	0
3	14.37	1.95	2.50		16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.0	0
4	13.24	2.59	2.87		21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.0	0

2) Масштабирование признака **alcohol**

```
[6]: scaler = StandardScaler()
df['alcohol_scaled'] = scaler.fit_transform(df[['alcohol']])

# Сравнение исходных и масштабированных значений
df[['alcohol', 'alcohol_scaled']].head()
```

```
[6]:
```

	alcohol	alcohol_scaled
0	14.23	1.518613
1	13.20	0.246290
2	13.16	0.196879
3	14.37	1.691550
4	13.24	0.295700

3) Преобразование категориального признака **target**

```
[8]: onehot_encoder = OneHotEncoder(sparse_output=False)
target_onehot = onehot_encoder.fit_transform(df[['target']])
onehot_df = pd.DataFrame(target_onehot, columns=[f'target_{i}' for i in range(target_onehot.shape[1])])
df = pd.concat([df, onehot_df], axis=1)
df.filter(like='target_').head()
```

```
[8]:
```

	target_label_encoded	target_0	target_1	target_2
0	0	1.0	0.0	0.0
1	0	1.0	0.0	0.0
2	0	1.0	0.0	0.0
3	0	1.0	0.0	0.0
4	0	1.0	0.0	0.0

4) Построение скрипичной диаграммы

```
!:
```

```
alcohol_by_class = [df[df['target'] == i]['alcohol'] for i in sorted(df['target'].unique())]

# Настройка графика
plt.figure(figsize=(10, 6))
plt.violinplot(alcohol_by_class, showmeans=True, showmedians=True)
plt.title('Скрипичная диаграмма: Alcohol по классам вина', fontsize=14)
plt.xlabel('Класс вина', fontsize=12)
plt.ylabel('Alcohol (%)', fontsize=12)
plt.xticks(ticks=[1, 2, 3], labels=['Class 0', 'Class 1', 'Class 2'])
plt.grid(axis='y', linestyle='--', alpha=0.4)
plt.show()
```



4. Используемые методы и причины их выбора

1. Масштабирование данных (StandardScaler)

Метод:

Для масштабирования признака `alcohol` был применен метод `StandardScaler` из библиотеки `scikit-learn`. Этот метод стандартизирует данные, преобразуя их таким образом, чтобы среднее значение стало равным 0, а стандартное отклонение — 1.

Причины выбора:

Нормализация масштаба: Признаки в датасете имеют разный масштаб (например, `alcohol` измеряется в процентах, а `malic_acid` — в других единицах). Масштабирование необходимо для корректной работы алгоритмов, чувствительных к масштабу данных (например, SVM, k-NN, методы кластеризации).

Сохранение интерпретируемости: `StandardScaler` не меняет распределение данных, а только приводит их к единому масштабу, что упрощает интерпретацию результатов.

2. One-Hot Encoding

Метод:

Для устранения недостатков Label Encoding был применен `OneHotEncoder`, который создает бинарные колонки для каждой категории (например, `target_0`, `target_1`, `target_2`).

Причины выбора:

Учет номинальности данных: Классы вина не имеют естественного порядка, и One-Hot Encoding устраняет ложную зависимость между числами.

Совместимость с алгоритмами: Большинство моделей машинного обучения (например, линейная регрессия, нейросети) работают лучше с бинарными признаками.

Недостатки:

Увеличивает размерность данных (проблема для датасетов с множеством категорий).

3. Визуализация (скрипичная диаграмма)

Метод:

Для анализа распределения признака `alcohol` по классам вина была построена скрипичная диаграмма (`violinplot`) с помощью `matplotlib`.

Причины выбора:

Комбинация boxplot и KDE: Диаграмма показывает медиану, межквартильный размах (как boxplot) и плотность распределения (как KDE).

Наглядность: Позволяет сразу оценить различия в распределении алкоголя между классами.

5. Выводы

Выбор методов был обусловлен:

1. Характеристиками данных (разный масштаб признаков, номинальность целевой переменной).
2. Требованиями алгоритмов (необходимость масштабирования для расстояний, чувствительность к порядку категорий).
3. Интерпретируемостью результатов (наглядность скрипичной диаграммы).

Все методы были реализованы средствами scikit-learn и matplotlib, что обеспечило воспроизводимость и минимальные затраты на предобработку.