**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: Проверил:  
Файзуллин К.Х. Гапанюк Ю.Е.  
группа ИУ5-64Б

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**

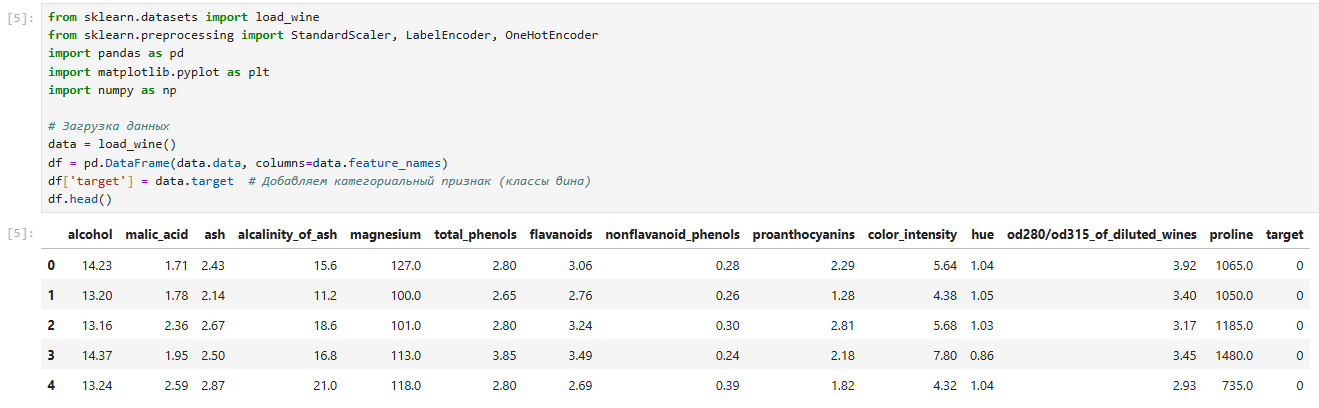
В рамках рубежного контроля была проведена работа с набором данных Wine Dataset. Целью работы являлось масштабирование данных и преобразование категориальных признаков.

### 2. Описание исходных данных

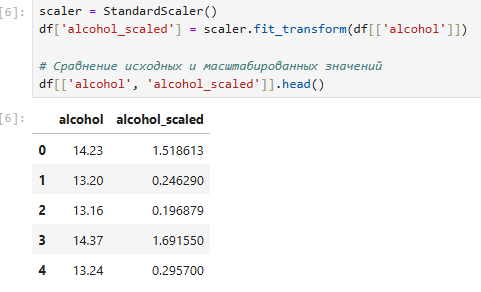
Набор данных **Wine** из библиотеки scikit-learn содержит результаты химического анализа 178 образцов итальянских вин, относящихся к трем различным классам (**class\_0**,**class\_1**,**class\_2**). Каждый образец характеризуется 13 количественными признаками, включая содержание алкоголя, яблочной кислоты, 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:**

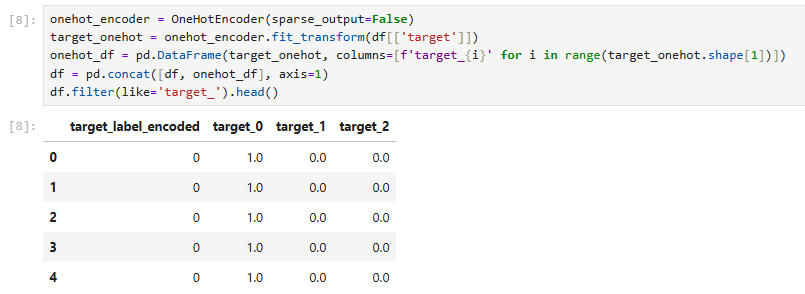
1) Загрузка и первичный анализ данных



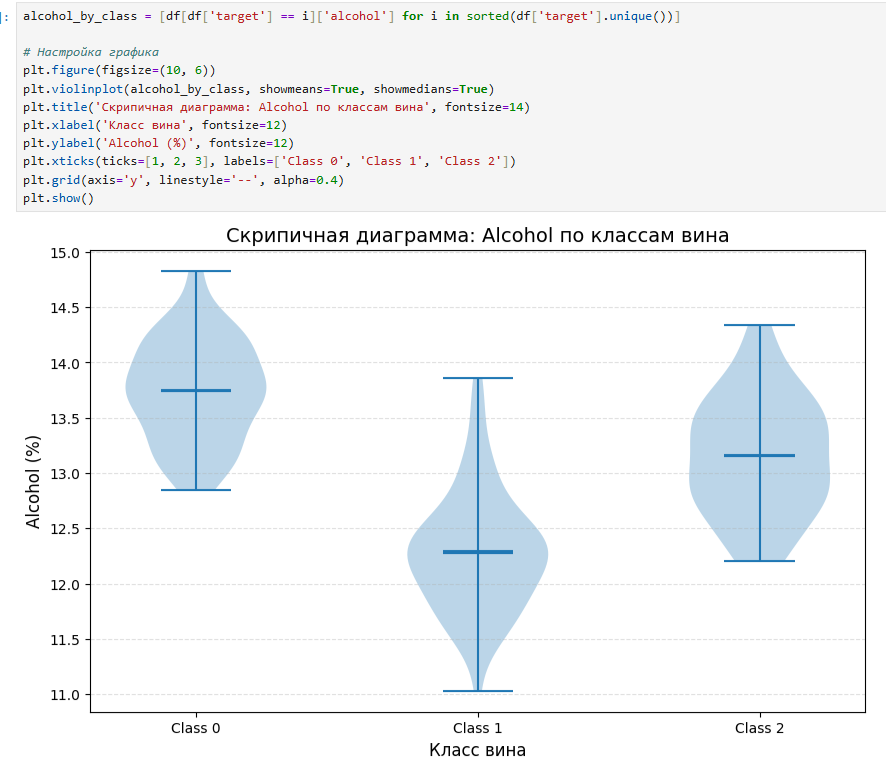
2) Масштабирование признака **alcohol**



3) Преобразование категориального признака **target**



4) Построение скрипичной диаграммы



### 4. Использованные 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.