Министерство образования Республики Беларусь
Учреждение образования
«Брестский Государственный технический университет»
Кафедра ИИТ

Лабораторная работа №1
По дисциплине: «Основы машинного обучения»
Тема: «Знакомство с анализом данных: предварительная обработка и визуализация»

Выполнил:
Студент 2 курса
Группы АС-66
Лысюк Р. А.
Проверил:
Крощенко А. А.

Брест 2025

Цель работы: получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Ход работы

Общее задание:

1. Загрузить предложенный набор данных (по вариантам) в DataFrame библиотеки Pandas.
2. Провести исследовательский анализ: изучить типы данных, количество пропусков, основные статистические показатели (среднее, медиана, стандартное отклонение).
3. Обработать пропущенные значения (например, заполнить средним значением или удалить строки/столбцы).
4. Преобразовать категориальные признаки в числовые с помощью метода One-Hot Encoding.
5. Выполнить нормализацию или стандартизацию числовых признаков.
6. Построить несколько графиков для визуализации данных (гистограммы, диаграммы рассеяния) и сделать выводы о зависимостях между признаками.
7. **Написать отчет, создать пул-реквест в репозиторий с кодом решения и отчетом в формате pdf.**

Используемые инструменты: Python, Pandas, Matplotlib, NumPy, Jupyter Notebook / Google Colab / PyCharm

Вариант 4

Выборка Wine Quality. Содержит физико-химические показатели красного и белого вина и оценку его качества по шкале.

Задачи:

1. Загрузите данные и выведите информацию о типах столбцов.
2. Преобразуйте целевую переменную quality в категориальную: "плохое" (<=4), "среднее" (5-6), "хорошее" (>=7).
3. Постройте столбчатую диаграмму, показывающую количество вин каждой новой категории качества.
4. Проверьте корреляцию между fixed acidity и pH. Визуализируйте эту зависимость на диаграмме рассеяния.
5. Найдите признак с наибольшим количеством выбросов, используя "ящик с усами" (box plot).
6. Выполните стандартизацию всех числовых признаков.

Код программы:

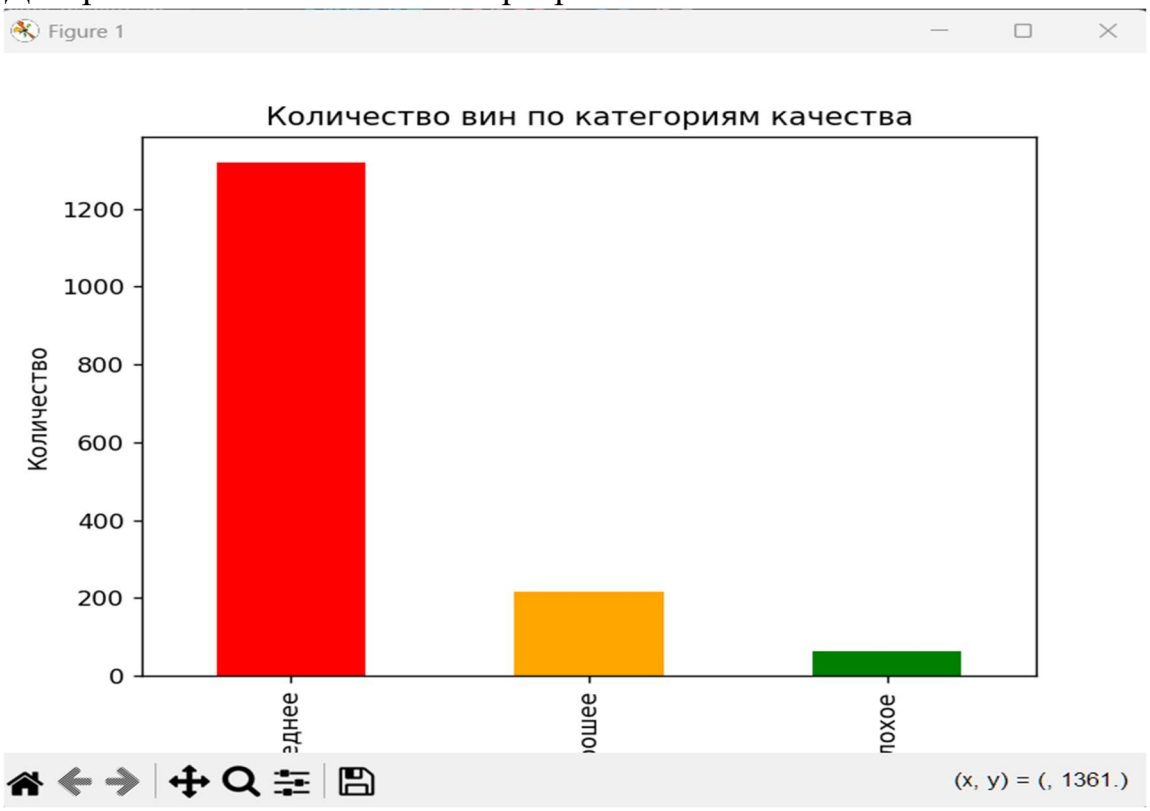
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import StandardScaler

file_path = 'winequality-red.csv'
df = pd.read_csv(file_path, sep=';')
print('Информация о данных:')
print(df.info())

def quality_to_category(q):
    if q <= 4:
        return 'плохое'
    elif 5 <= q <= 6:
        return 'среднее'
    else:
        return 'хорошее'

df['quality_cat'] = df['quality'].apply(quality_to_category)
print('\nРаспределение по категориям качества:')
print(df['quality_cat'].value_counts())

df['quality_cat'].value_counts().plot(
    kind='bar', color=['red', 'orange', 'green'])
plt.title('Количество вин по категориям качества')
plt.xlabel('Категория качества')
plt.ylabel('Количество')
plt.show()

corr = df['fixed acidity'].corr(df['pH'])
print(f'\nКорреляция fixed acidity и pH: {corr:.3f}')

plt.scatter(df['fixed acidity'], df['pH'], alpha=0.5)
```

```
plt.title('Диаграмма рассеяния fixed acidity vs pH')
plt.xlabel('fixed acidity')
plt.ylabel('pH')
plt.show()

def count_outliers(series):
    Q1 = series.quantile(0.25)
    Q3 = series.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = series[(series < lower_bound) | (series > upper_bound)]
    return len(outliers)

numeric_cols = df.select_dtypes(include=np.number).columns.drop('quality')
outliers_counts = {col: count_outliers(df[col]) for col in numeric_cols}
max_outliers_feature = max(outliers_counts, key=outliers_counts.get)

print(
        f'\nПризнак с наибольшим количеством выбросов: {max_outliers_feature} ({outliers_counts[max_outliers_feature]} выбросов)')

plt.boxplot(df[max_outliers_feature])
plt.title(f'Box plot для {max_outliers_feature}')
plt.show()

scaler = StandardScaler()
features_to_scale = numeric_cols
df_scaled = df.copy()
df_scaled[features_to_scale] = scaler.fit_transform(df[features_to_scale])

print('\nПример стандартизированных данных:')
print(df_scaled.head())
```

Часть выборки для примера:

```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5
7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5
7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7
7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5
5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5
7.8;0.61;0.29;1.6;0.114;9;29;0.9974;3.26;1.56;9.1;5
8.9;0.62;0.18;3.8;0.176;52;145;0.9986;3.16;0.88;9.2;5
8.9;0.62;0.19;3.9;0.17;51;148;0.9986;3.17;0.93;9.2;5
8.5;0.28;0.56;1.8;0.092;35;103;0.9969;3.3;0.75;10.5;7
8.1;0.56;0.28;1.7;0.368;16;56;0.9968;3.11;1.28;9.3;5
7.4;0.59;0.08;4.4;0.086;6;29;0.9974;3.38;0.5;9;4
7.9;0.32;0.51;1.8;0.341;17;56;0.9969;3.04;1.08;9.2;6
8.9;0.22;0.48;1.8;0.077;29;60;0.9968;3.39;0.53;9.4;6
7.6;0.39;0.31;2.3;0.082;23;71;0.9982;3.52;0.65;9.7;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
8.5;0.49;0.11;2.3;0.084;9;67;0.9968;3.17;0.53;9.4;5
6.9;0.4;0.14;2.4;0.085;21;40;0.9968;3.43;0.63;9.7;6
6.3;0.39;0.16;1.4;0.08;11;23;0.9955;3.34;0.56;9.3;5
7.6;0.41;0.24;1.8;0.08;4;11;0.9962;3.28;0.59;9.5;5
7.9;0.43;0.21;1.6;0.106;10;37;0.9966;3.17;0.91;9.5;5
7.1;0.71;0;1.9;0.08;14;35;0.9972;3.47;0.55;9.4;5
7.8;0.645;0;2;0.082;8;16;0.9964;3.38;0.59;9.8;6
6.7;0.675;0.07;2.4;0.089;17;82;0.9958;3.35;0.54;10.1;5
6.9;0.685;0;2.5;0.105;22;37;0.9966;3.46;0.57;10.6;6
8.3;0.655;0.12;2.3;0.083;15;113;0.9966;3.17;0.66;9.8;5
6.9;0.605;0.12;10.7;0.073;40;83;0.9993;3.45;0.52;9.4;6
5.2;0.32;0.25;1.8;0.103;13;50;0.9957;3.38;0.55;9.2;5
7.8;0.645;0;5.5;0.086;5;18;0.9986;3.4;0.55;9.6;6
7.8;0.6;0.14;2.4;0.086;3;15;0.9975;3.42;0.6;10.8;6
8.1;0.38;0.28;2.1;0.066;13;30;0.9968;3.23;0.73;9.7;7
5.7;1.13;0.09;1.5;0.172;7;19;0.994;3.5;0.48;9.8;4
```

Диаграммы после выполения программы:



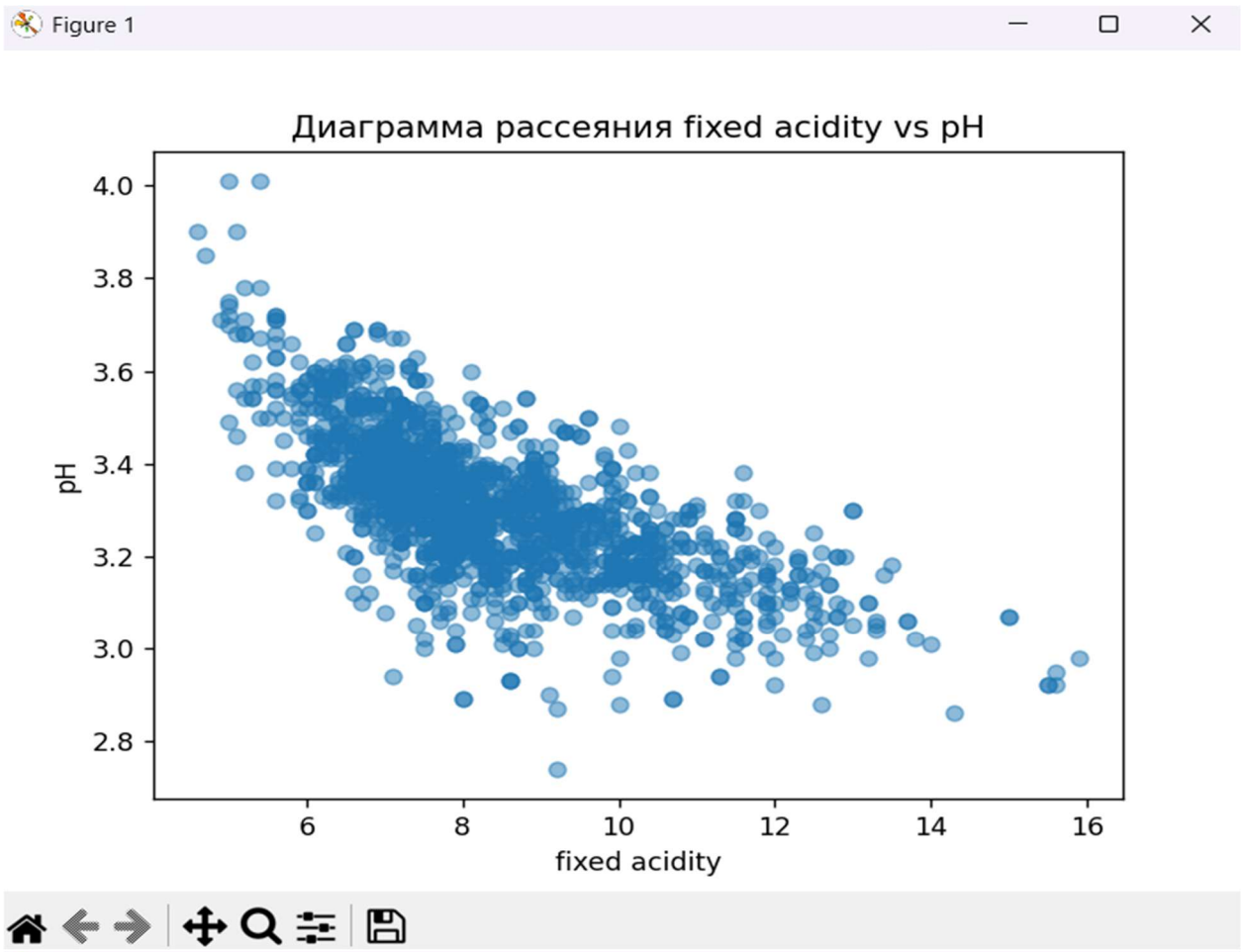


Figure 1
Диаграмма рассеяния fixed acidity vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
6
8
10
12
14
16
fixed acidity

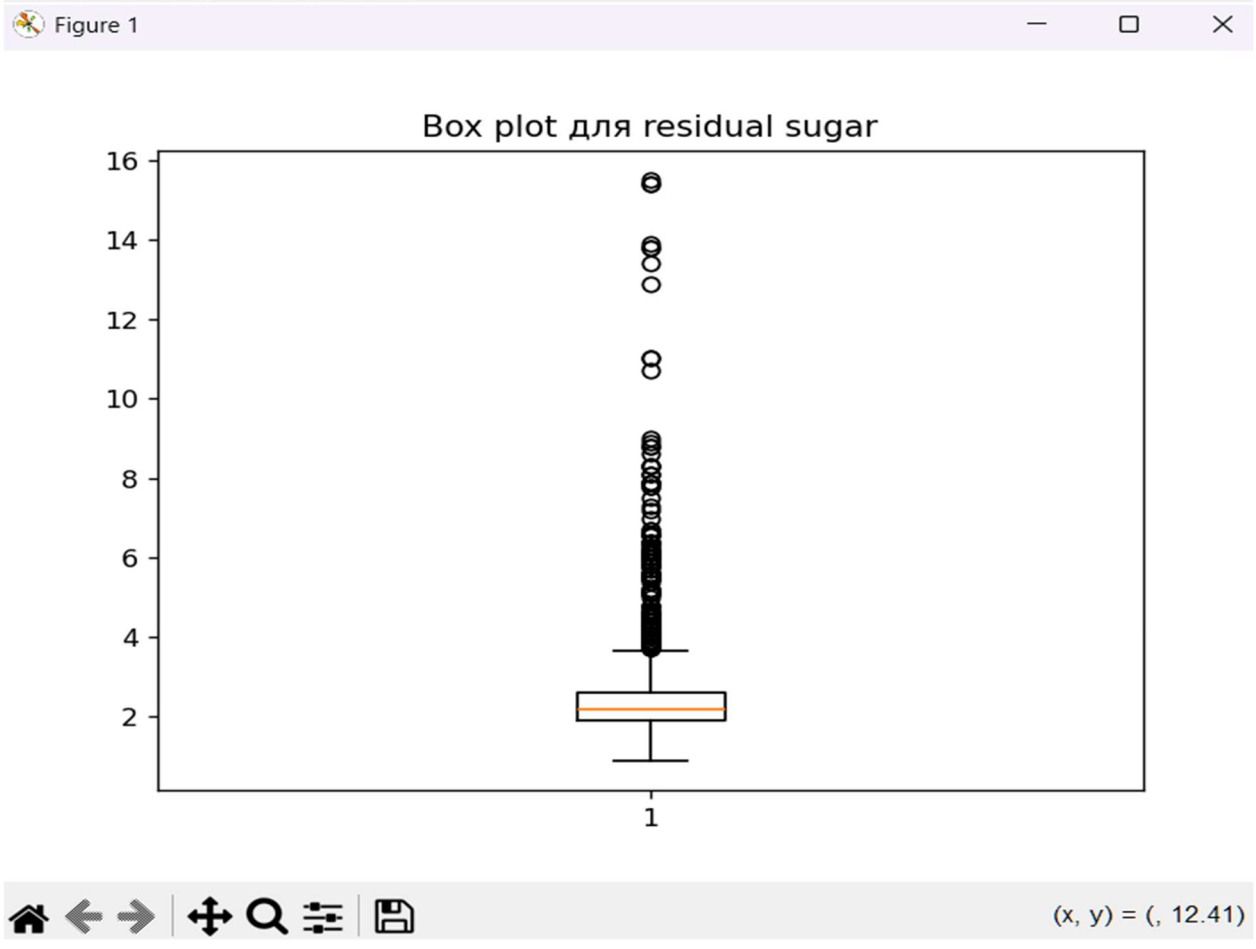


Figure 1
Box plot для residual sugar
16
14
12
10
8
6
4
2
1
(x, y) = (, 12.41)

Вывод в консоли:

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
None

Распределение по категориям качества:
quality_cat
среднее    1319
хорошее     217
плохое       63
Name: count, dtype: int64
```

```
Корреляция fixed acidity и pH: -0.683

Признак с наибольшим количеством выбросов: residual sugar (155 выбросов)

Пример стандартизированных данных:
   fixed acidity  volatile acidity  citric acid  residual sugar  ...  sulphates   alcohol  quality  quality_cat
0      -0.528360          0.961877    -1.391472       -0.453218  ...  -0.579207 -0.960246        5      среднее
1      -0.298547          1.967442    -1.391472        0.043416  ...   0.128950 -0.584777        5      среднее
2      -0.298547          1.297065    -1.186070       -0.169427  ...  -0.048089 -0.584777        5      среднее
3       1.654856         -1.384443     1.484154       -0.453218  ...  -0.461180 -0.584777        6      среднее
4      -0.528360          0.961877    -1.391472       -0.453218  ...  -0.579207 -0.960246        5      среднее

[5 rows x 13 columns]
```

Вывод: получил практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научился выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.