

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 3 курса
Группы АС-65
Дацкевич М. Г.
Проверил:
Крощенко А. А.

Цель: 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.

Вариант 4

Задачи:

1. Загрузите данные и выведите информацию о типах столбцов.

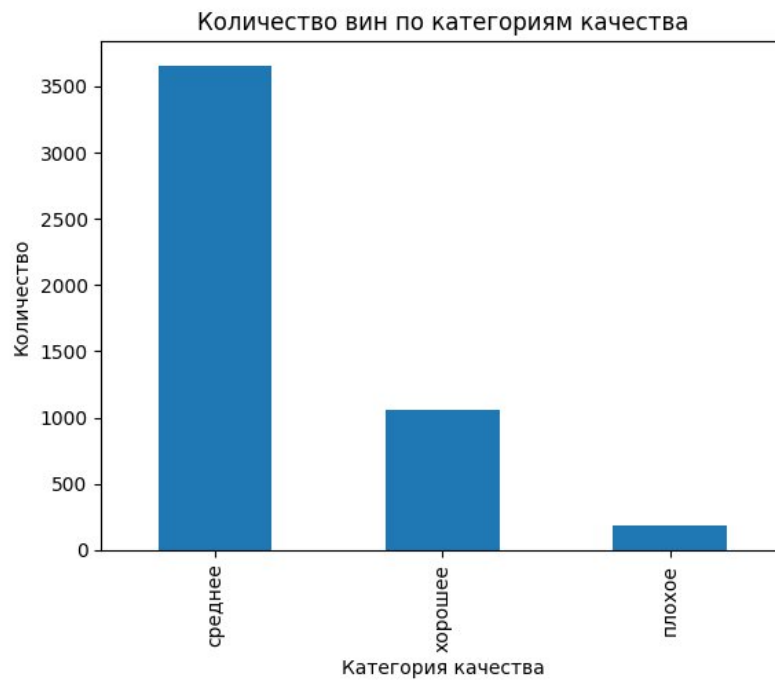
```
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from sklearn.preprocessing import StandardScaler
6.
7. # 1. Загрузка данных и информация о типах столбцов
8. wine_data = pd.read_csv('winequality-white.csv', sep=';')
9. print(wine_data.dtypes)
```

```
fixed acidity      float64
volatile acidity   float64
citric acid        float64
residual sugar     float64
chlorides          float64
free sulfur dioxide float64
total sulfur dioxide float64
density            float64
pH                float64
sulphates          float64
alcohol            float64
quality            int64
dtype: object
```

2. Преобразуйте целевую переменную quality в категориальную: "плохое" (≤ 4), "среднее" (5-6), "хорошее" (≥ 7).

```
# 2. Преобразование целевой переменной quality в категориальную
def categorize_quality(quality):
    if quality <= 4:
        return "плохое"
    elif quality <= 6:
        return "среднее"
    else:
        return "хорошее"

wine_data['quality_category'] = wine_data['quality'].apply(categorize_quality)
```

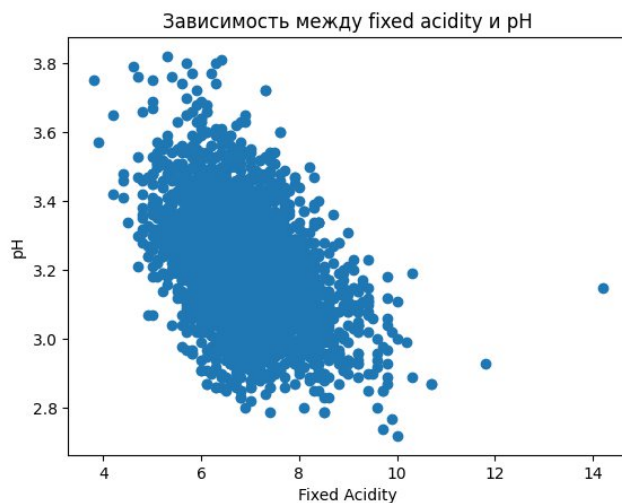


3. Постройте столбчатую диаграмму, показывающую количество вин каждой новой категории качества.

```
# 3. Столбчатая диаграмма количества вин каждой категории качества
wine_data['quality_category'].value_counts().plot(kind='bar')
plt.title('Количество вин по категориям качества')
plt.xlabel('Категория качества')
plt.ylabel('Количество')
plt.show()
```

4. Проверьте корреляцию между fixed acidity и pH. Визуализируйте эту зависимость на диаграмме рассеяния.

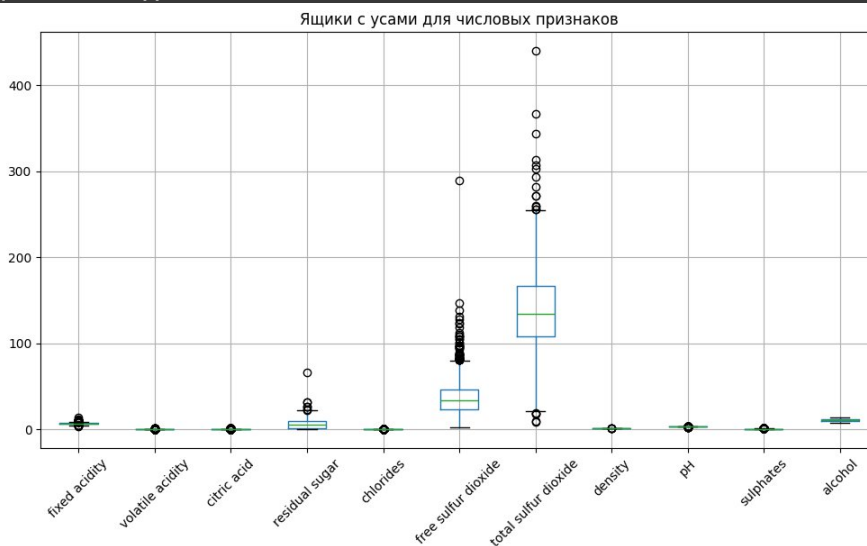
```
# 4. Проверка корреляции между fixed acidity и pH
correlation = wine_data['fixed acidity'].corr(wine_data['pH'])
print(f"Корреляция между fixed acidity и pH: {correlation:.4f}")
plt.scatter(wine_data['fixed acidity'], wine_data['pH'])
plt.title('Зависимость между fixed acidity и pH')
plt.xlabel('Fixed Acidity')
plt.ylabel('pH')
plt.show()
```



5. Найдите признак с наибольшим количеством выбросов, используя "ящик с усами" (box plot).

```
# 5. Поиск признака с наибольшим количеством выбросов
numeric_columns = wine_data.select_dtypes(include=['float64',
'int64']).columns
numeric_columns = numeric_columns.drop('quality')

plt.figure(figsize=(12, 6))
wine_data[numeric_columns].boxplot()
plt.xticks(rotation=45)
plt.title('Ящики с усами для числовых признаков')
plt.show()
```



6. Выполните стандартизацию всех числовых признаков.

```
# 6. Стандартизация всех числовых признаков
scaler = StandardScaler()
wine_data_standardized = wine_data.copy()
wine_data_standardized[numeric_columns] = scaler.fit_transform(
    wine_data_standardized[numeric_columns])
print("Стандартизация выполнена")
```

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