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

Лабораторная работа №1
По дисциплине: «ОМО»
Тема:” Знакомство с анализом данных: предварительная обработка и визуализация.”

Выполнил:
Студент 3-го курса
Группы АС-66
Осовец А.О.
Проверил:
Крощенко А.А.

Брест 2025

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 7

Выборка Auto MPG. Содержит технические характеристики различных автомобилей и данные о расходе топлива (миль на галлон).
Задачи:
1. Загрузите данные. Обратите внимание, что пропуски в столбце horsepower могут быть обозначены знаком ?.
2. Преобразуйте столбец horsepower в числовой формат и заполните пропуски средним значением.
3. Постройте диаграмму рассеяния, чтобы изучить зависимость расхода топлива (mpg) от веса автомобиля (weight).
4. Преобразуйте категориальный признак origin (страна производства) в числовой.
5. Создайте новый признак age, рассчитав возраст автомобиля относительно года, когда были собраны данные (например, 1983 - model year).
6. Визуализируйте распределение количества цилиндров (cylinders) с помощью столбчатой диаграммы.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# ------------------------------
# 1. Загрузка данных
# ------------------------------
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data'
columns = ['mpg', 'cylinders', 'displacement', 'horsepower', 'weight',
           'acceleration', 'model_year', 'origin', 'car_name']

df = pd.read_csv(url, names=columns, sep=r'\s+', na_values='?')

# ------------------------------
# 2. Обработка пропусков
# ------------------------------
df['horsepower'] = df['horsepower'].astype(float)
df['horsepower'] = df['horsepower'].fillna(df['horsepower'].mean())

# ------------------------------
# 3. Исследовательский анализ
# ------------------------------
print("Типы данных:\n", df.dtypes)
print("\nКоличество пропусков:\n", df.isna().sum())
```

```
print("\nОсновные статистические показатели:\n", df.describe())

# ----------------------------
# 4. Преобразование категориального признака
# ----------------------------
df = pd.get_dummies(df, columns=['origin'], prefix='origin')

# ----------------------------
# 5. Создание нового признака age
# ----------------------------
df['age'] = 1983 - (1900 + df['model_year'])

# ----------------------------
# 6. Визуализация данных
# ----------------------------

# 6.1 Расход топлива vs вес
plt.figure(figsize=(8,6))
plt.scatter(df['weight'], df['mpg'], alpha=0.7)
plt.title('mpg vs weight')
plt.xlabel('Вес автомобиля')
plt.ylabel('Расход топлива (mpg)')
plt.grid(True)
plt.show()

# 6.2 Распределение количества цилиндров
cylinder_counts = df['cylinders'].value_counts().sort_index()
plt.figure(figsize=(8,6))
plt.bar(cylinder_counts.index, cylinder_counts.values, color='skyblue')
plt.title('Распределение цилиндров')
plt.xlabel('Количество цилиндров')
plt.ylabel('Количество автомобилей')
plt.show()

# 6.3 Гистограмма расхода топлива
plt.figure(figsize=(8,6))
plt.hist(df['mpg'], bins=15, color='lightgreen', edgecolor='black')
plt.title('Распределение расхода топлива (mpg)')
plt.xlabel('mpg')
plt.ylabel('Количество автомобилей')
plt.show()

# 6.4 Расход топлива vs horsepower
plt.figure(figsize=(8,6))
plt.scatter(df['horsepower'], df['mpg'], color='orange', alpha=0.7)
plt.title('mpg vs horsepower')
plt.xlabel('Мощность (horsepower)')
plt.ylabel('Расход топлива (mpg)')
plt.grid(True)
plt.show()
```

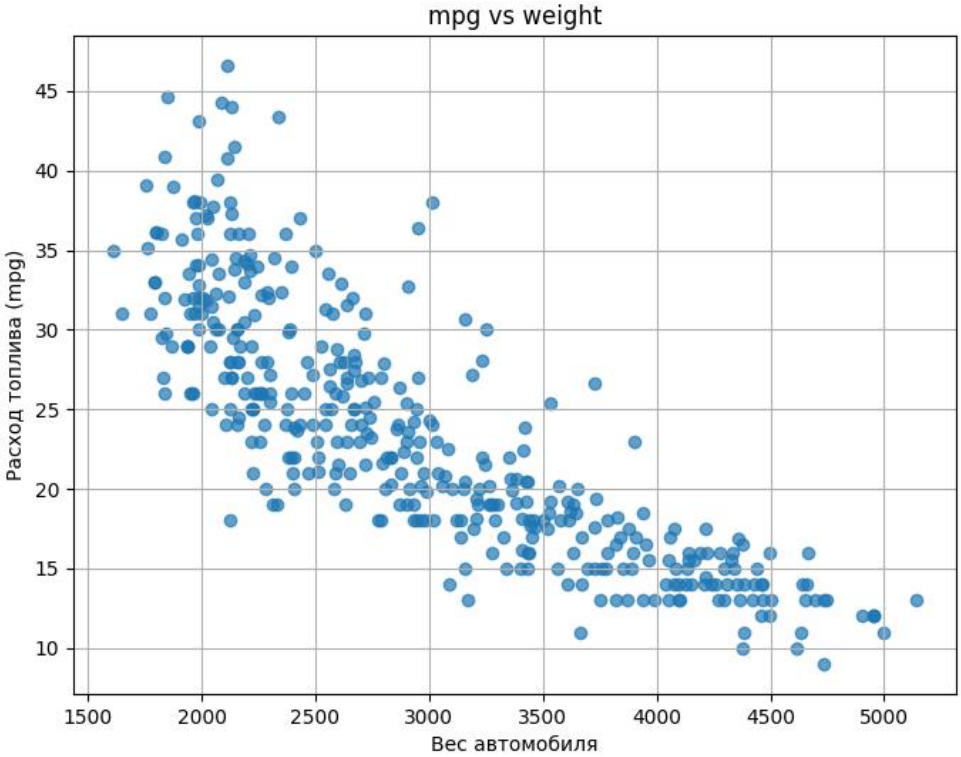
```
# 6.5 Расход топлива vs displacement
plt.figure(figsize=(8,6))
plt.scatter(df['displacement'], df['mpg'], color='red', alpha=0.7)
plt.title('mpg vs displacement')
plt.xlabel('Объём двигателя (displacement)')
plt.ylabel('Расход топлива (mpg)')
plt.grid(True)
plt.show()

# 6.6 Корреляционная матрица
plt.figure(figsize=(10,8))
corr = df[['mpg','cylinders','displacement','horsepower','weight','acceleration','age']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Корреляционная матрица')
plt.show()

# 6.7 Boxplot веса автомобиля по количеству цилиндров
plt.figure(figsize=(8,6))
sns.boxplot(x='cylinders', y='weight', data=df, hue='cylinders', legend=False, palette='Set2')
plt.title('Вес автомобиля по количеству цилиндров')
plt.xlabel('Количество цилиндров')
plt.ylabel('Вес автомобиля')
plt.show()
```

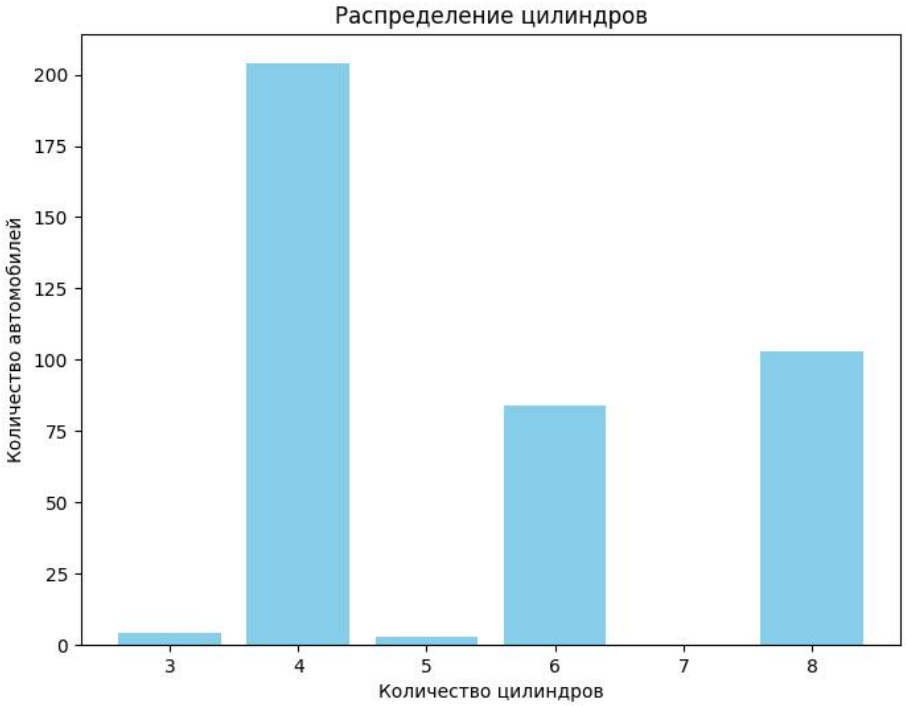
```
Типы данных:
 mpg             float64
cylinders          int64
displacement     float64
horsepower       float64
weight           float64
acceleration     float64
model_year         int64
origin             int64
car_name          object
dtype: object
```

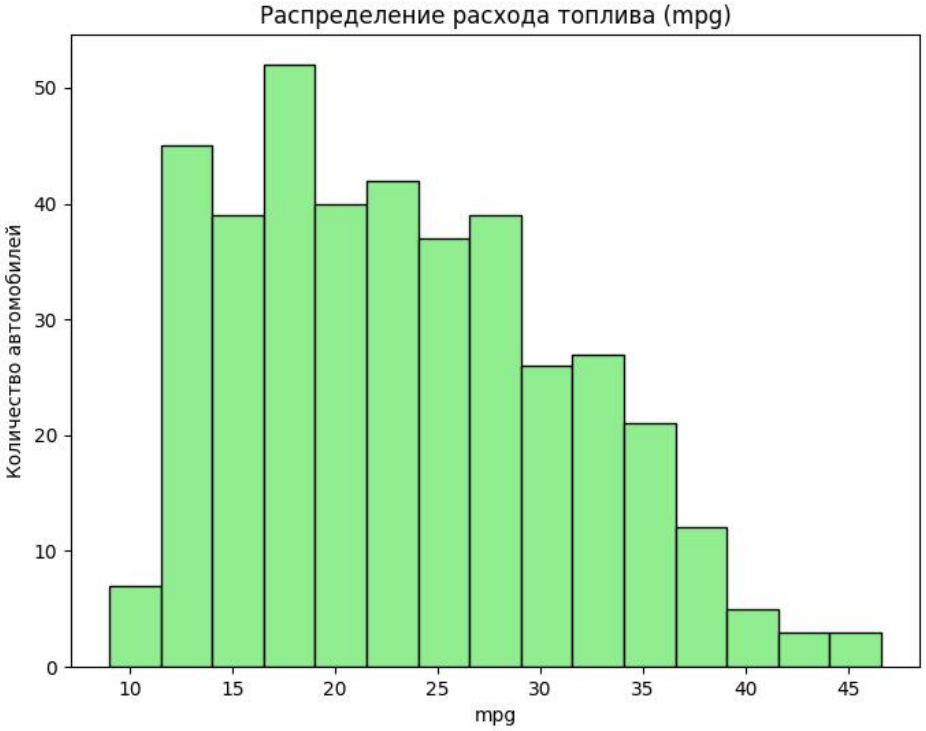
```
Количество пропусков:
 mpg             0
cylinders        0
displacement     0
horsepower       0
weight           0
acceleration     0
model_year       0
origin           0
car_name         0
dtype: int64
```

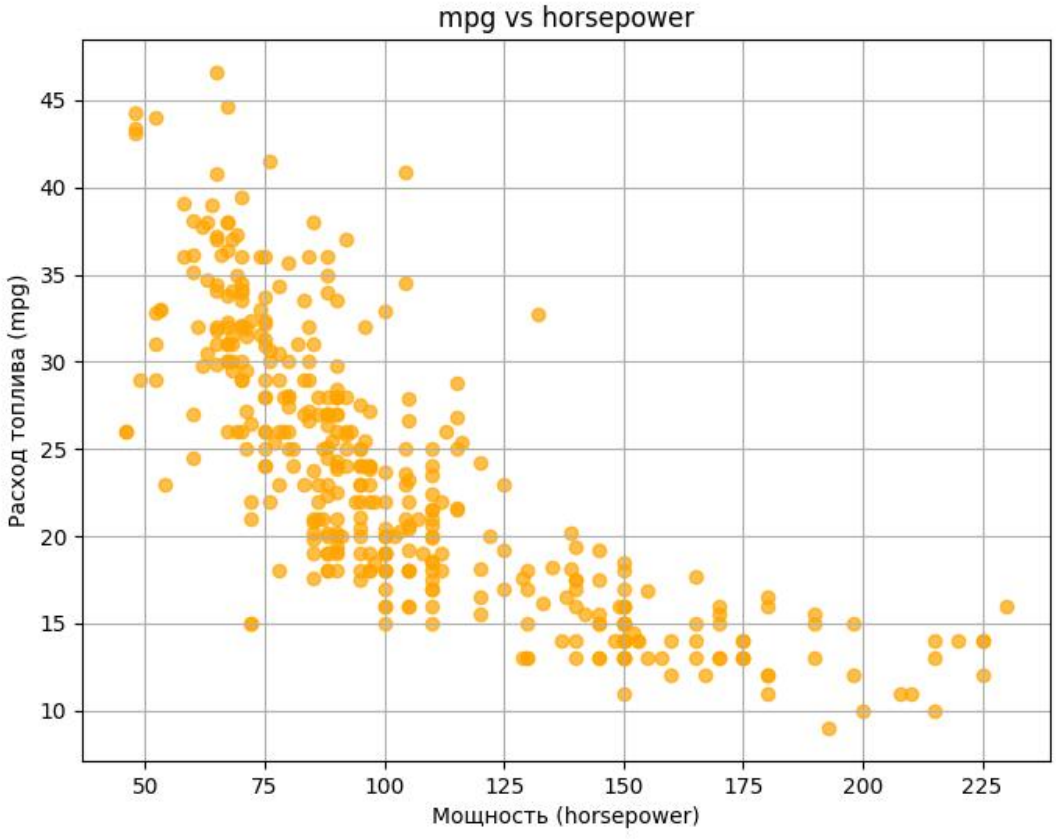
```
Основные статистические показатели:
                mpg   cylinders  ...  model_year      origin
count  398.000000  398.000000  ...  398.000000  398.000000
mean    23.514573    5.454774  ...   76.010050    1.572864
std      7.815984    1.701004  ...    3.697627    0.802055
min      9.000000    3.000000  ...   70.000000    1.000000
25%     17.500000    4.000000  ...   73.000000    1.000000
50%     23.000000    4.000000  ...   76.000000    1.000000
75%     29.000000    8.000000  ...   79.000000    2.000000
max     46.600000    8.000000  ...   82.000000    3.000000
```

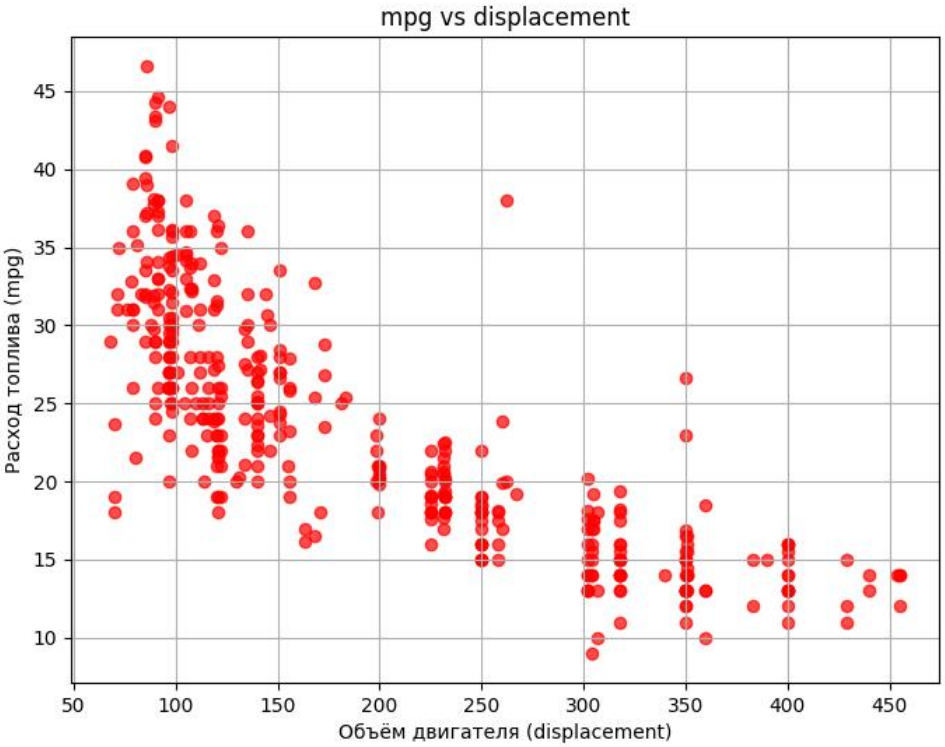
Графики:

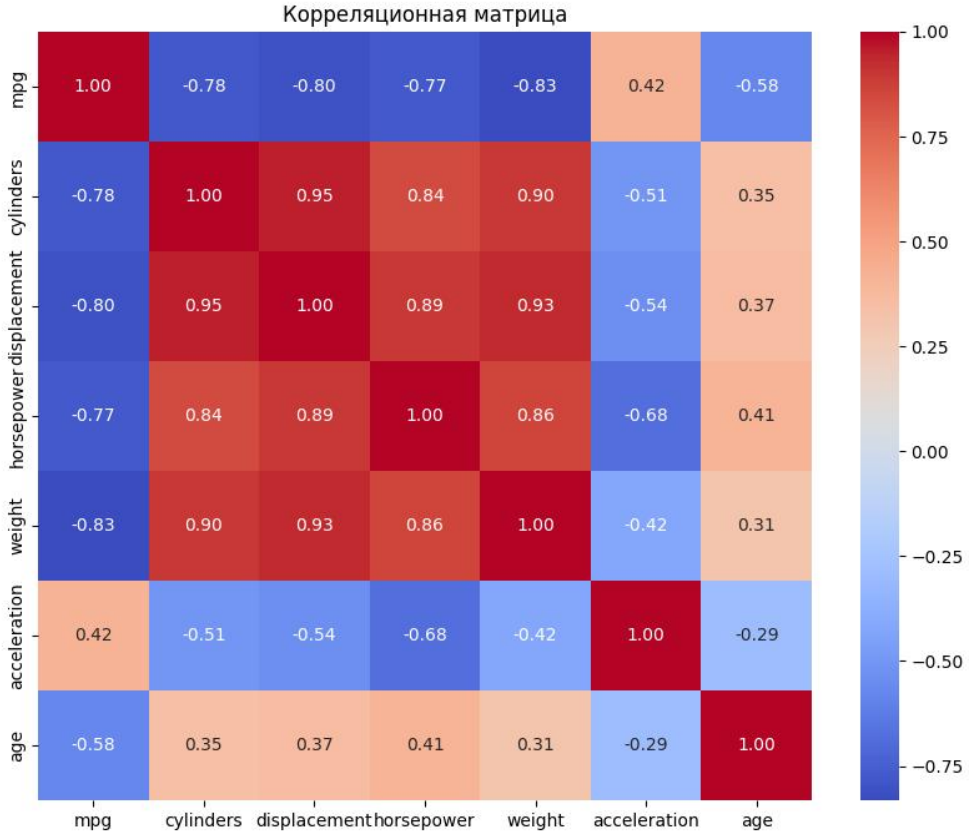

mpg vs weight
Расход топлива (mpg)
Вес автомобиля
45
40
35
30
25
20
15
10
1500
2000
2500
3000
3500
4000
4500
5000

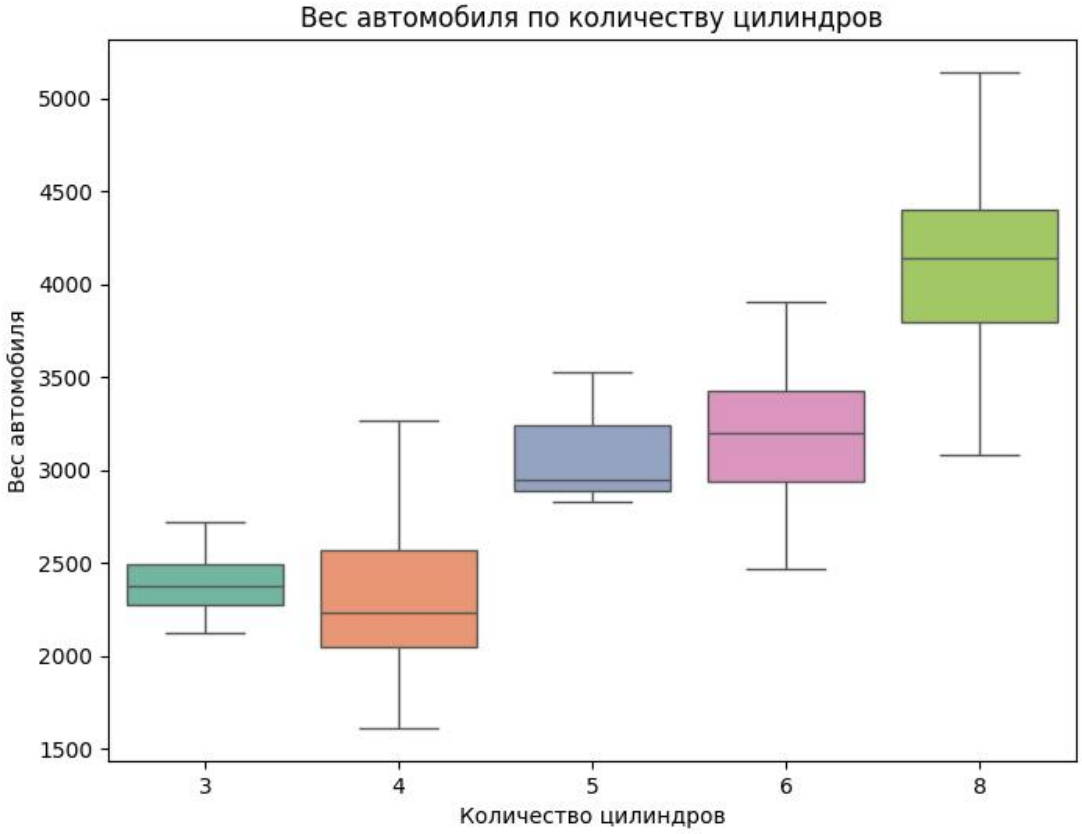

Распределение цилиндров
Количество автомобилей
Количество цилиндров
200
175
150
125
100
75
50
25
0
3
4
5
6
7
8


Распределение расхода топлива (mpg)
Количество автомобилей
mpg
0
10
20
30
40
50
10
15
20
25
30
35
40
45


mpg vs horsepower
Расход топлива (mpg)
Мощность (horsepower)
10
15
20
25
30
35
40
45
50
75
100
125
150
175
200
225


mpg vs displacement
Расход топлива (mpg)
Объём двигателя (displacement)
45
40
35
30
25
20
15
10
50
100
150
200
250
300
350
400
450


Корреляционная матрица
mpg
cylinders
displacement
horsepower
weight
acceleration
age
1.00 -0.78 -0.80 -0.77 -0.83 0.42 -0.58
-0.78 1.00 0.95 0.84 0.90 -0.51 0.35
-0.80 0.95 1.00 0.89 0.93 -0.54 0.37
-0.77 0.84 0.89 1.00 0.86 -0.68 0.41
-0.83 0.90 0.93 0.86 1.00 -0.42 0.31
0.42 -0.51 -0.54 -0.68 -0.42 1.00 -0.29
-0.58 0.35 0.37 0.41 0.31 -0.29 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75



Вывод: в результате выполнения данной лабораторной работы 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.