Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

Лабораторная работа №1 по курсу «Методы машинного обучения в автоматизированных системах обработки информации и управления»

Подготовили:

У Жун

ИУ5И-25М

20.03.2024

Проверил:

Гапанюк Ю. Е.

2024 г.

Цель лабораторной работы:

изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание:

1. Загрузить набор данных о диабете с помощью scikit-learn.
2. Провести анализ данных и выполнить визуализацию согласно заданным требованиям: 1) Анализировать распределение возраста.2) Определить соотношение полов.3) Исследовать связь между индексом массы тела и средним артериальным давлением.4) Определить распределение уровня прогрессирования болезни.5) Проанализировать распределение количественных показателей прогрессирования болезни через год.

Задание:

1.Выбрать набор данных (датасет). Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

2.Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

- История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
- На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
- Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
- Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
- История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

3. Сформировать отчет и разместить его в своем репозитории на github.

ТЕКСТ ПРОГРАММЫ

Шаг 1: Анализ распределения по возрасту
Сначала изучим распределение возраста в наборе данных по диабету.

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

# Load the diabetes dataset
diabetes = datasets.load_diabetes()
diabetes_df = pd.DataFrame(data=diabetes.data, columns=diabetes.feature_names)
diabetes_df['Age'] = diabetes_df['age']    # Добавление столбца возраста

# Analyzing the distribution of age
plt.figure(figsize=(8, 6))
plt.hist(diabetes_df['Age'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

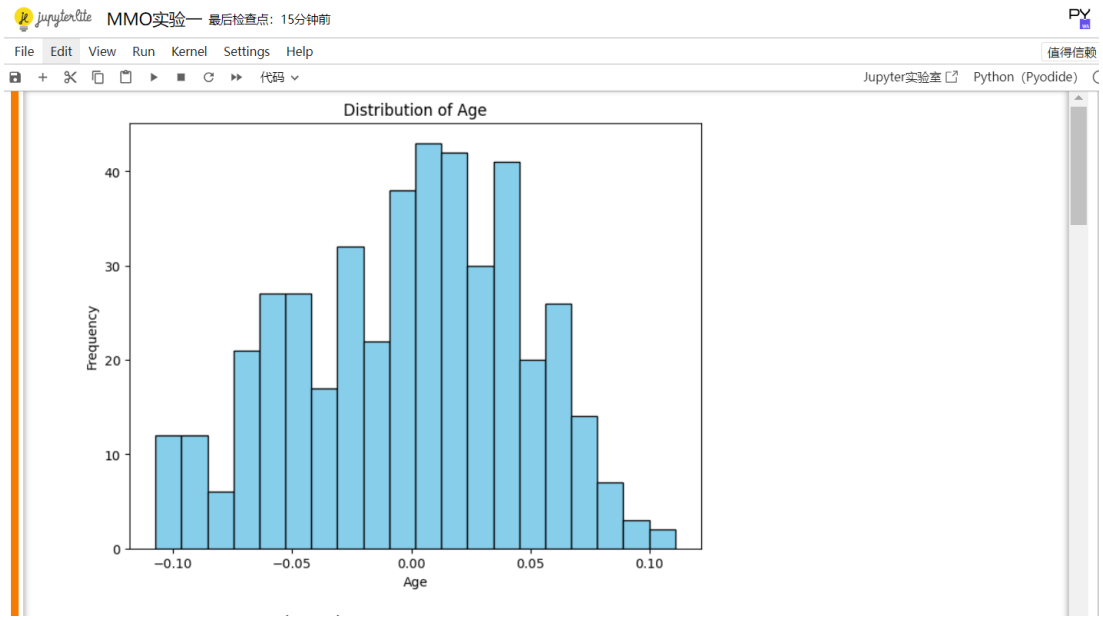

Рис.1 Анализируем распределение возраста

На этом графике мы видим гистограмму распределения возраста. Эта гистограмма дает нам представление о распределении возраста в наборе данных по диабету. Гистограмма показывает, что возраст в основном сосредоточен в определенном диапазоне, но есть и выбросы.

Шаг 2: Анализ соотношения полов
Далее давайте изучим соотношение полов в наборе данных о диабете.

```
# Adding Sex column
diabetes_df['Sex'] = diabetes_df['sex']
# Analyzing gender ratio
plt.figure(figsize=(6, 6))
diabetes_df['Sex'].value_counts().plot(kind='pie', autopct='%1.1f%%', colors=['skyblue',
'pink'])
plt.title('Gender Ratio')
plt.ylabel('')
plt.show()
```

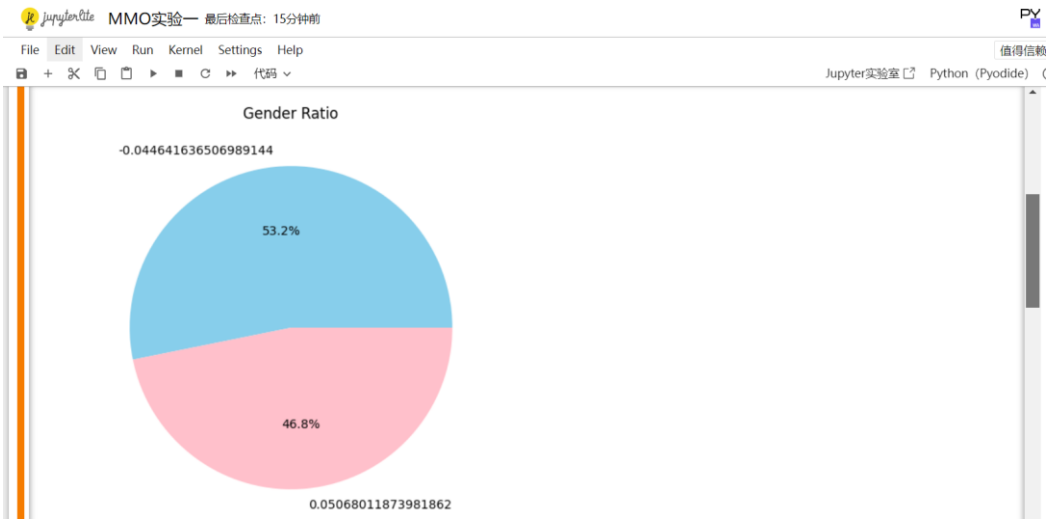

Рис. 2 – Анализ соотношения полов.

Эта круговая диаграмма дает нам четкое представление о доле мужчин и женщин в наборе данных по диабету. В этом наборе данных доля мужчин и женщин примерно одинакова.

Шаг 3: Исследование взаимосвязи между индексом массы тела и средним артериальным давлением
Теперь давайте изучим связь между индексом массы тела и средним артериальным давлением в наборе данных о диабете.

```
# Analyzing the relationship between body mass index and average blood pressure
plt.figure(figsize=(8, 6))
plt.scatter(diabetes_df['bmi'], diabetes_df['bp'], color='green', alpha=0.7)
plt.title('Body Mass Index vs. Average Blood Pressure')
plt.xlabel('Body Mass Index')
plt.ylabel('Average Blood Pressure')
plt.show()
```

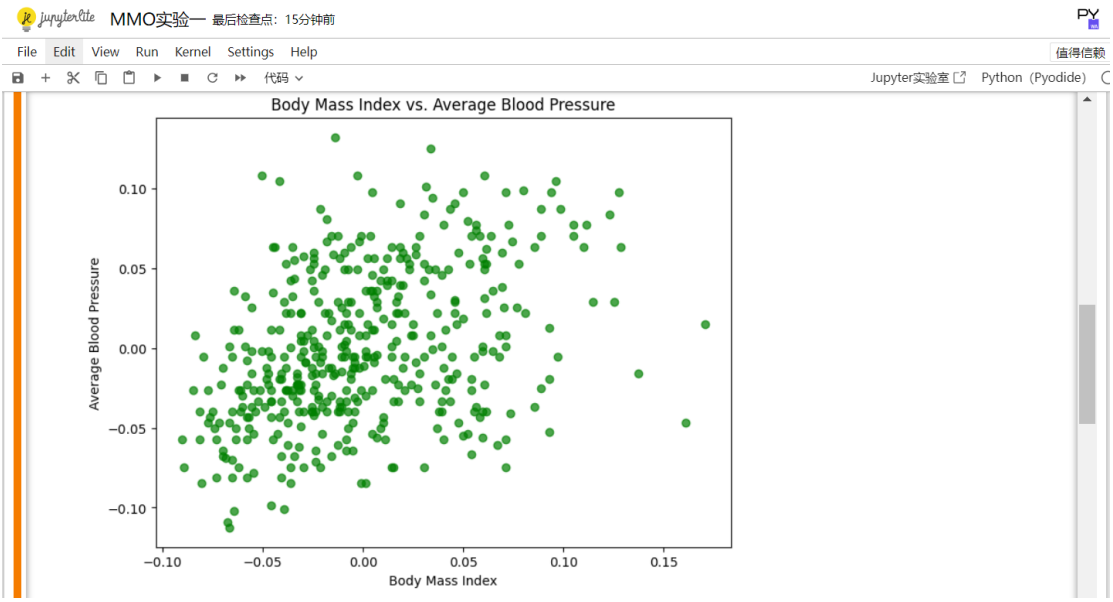

Рис.3– Анализ взаимосвязи между индексом массы тела и средним артериальным давлением

Эта диаграмма рассеяния позволяет нам рассмотреть взаимосвязь между индексом массы тела и средним артериальным давлением в наборе данных по диабету. На диаграмме показана общая тенденция между этими двумя переменными и наличие какой-либо корреляции.

Шаг 4: Определите распределение уровней прогрессирования заболевания
Теперь давайте изучим распределение уровней прогрессирования заболевания в наборе данных о диабете.

```
#    Analyzing the distribution of disease progression levels
plt.figure(figsize=(8, 6))
plt.boxplot(diabetes.target)
plt.title('Distribution of Disease Progression Levels')
plt.ylabel('Disease Progression Level')
plt.show()
```

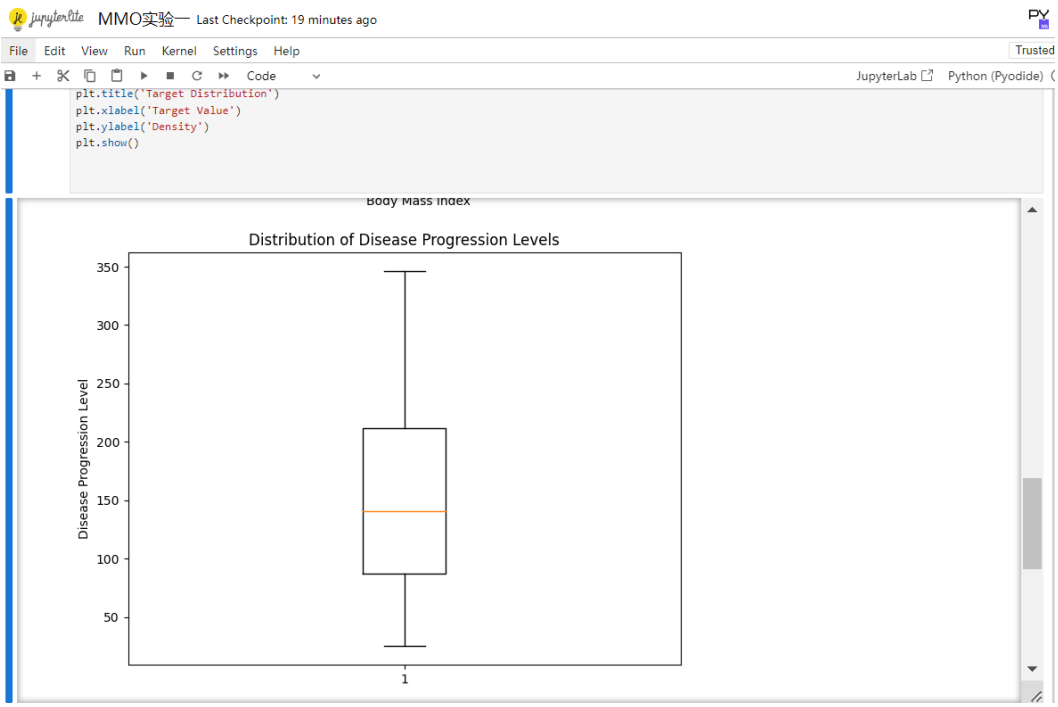

Рис.4– Анализ распределения уровней развития болезни

График распределения показателей уровня заболевания в виде бокс-линии позволяет понять распределение уровней развития заболевания среди пациентов в наборе данных по диабету. Это помогает определить распространенность уровней развития заболевания в популяции пациентов.

Шаг 5: Анализ распределения количественных показателей заболевания через год

Наконец, давайте изучим распределение количественных показателей заболевания через год в наборе данных о диабете.

```
# Analyzing the distribution of quantitative disease progression indicators after one year
    plt.figure(figsize=(8, 6))
    pd.Series(diabetes.target).plot(kind='density', color='orange')
    plt.title('Target Distribution')
    plt.xlabel('Target Value')
    plt.ylabel('Density')
    plt.show()
```

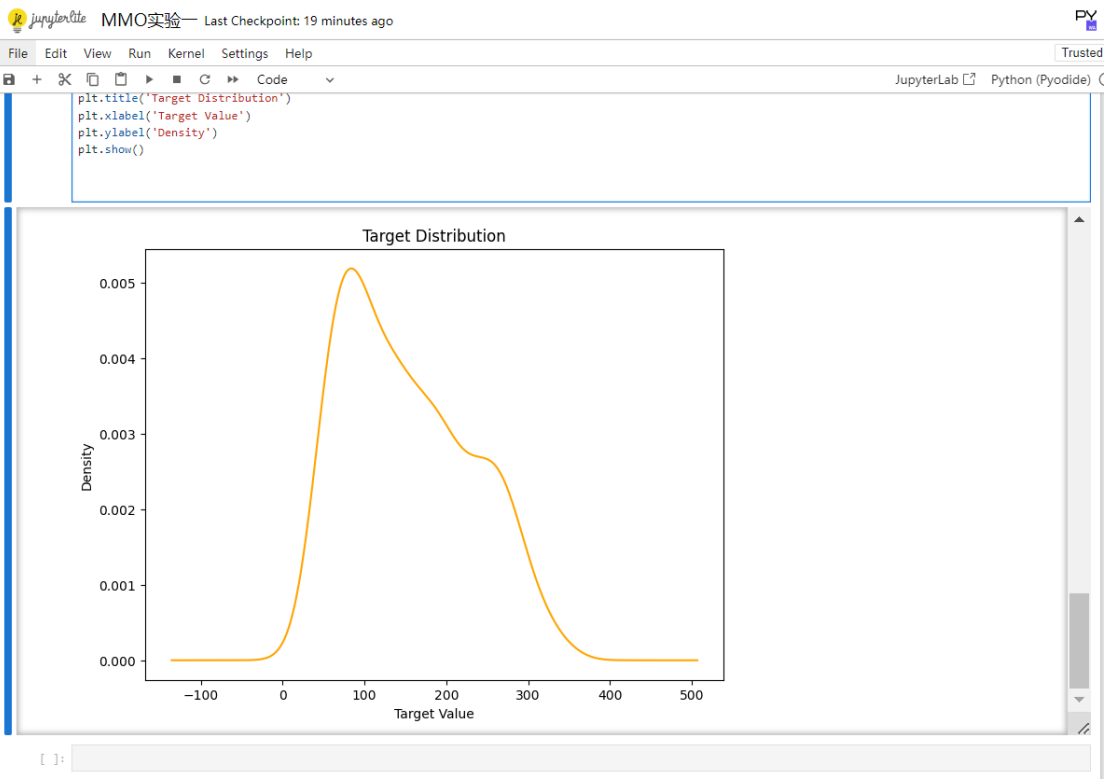

Рис.5–Анализ распределения количественных показателей прогрессирования заболевания через год

График плотности распределения количественных показателей заболевания через год позволяет наблюдать распределение количественных показателей заболевания через год среди пациентов в наборе данных по диабету. Это помогает понять тенденции и особенности прогрессирования заболевания в популяции пациентов.

ВЫВОДЫ

Проанализировав и визуализировав набор данных по диабету, мы можем сделать следующие выводы:

1. Распределение по возрасту: распределение по возрасту, наблюдаемое в наборе данных о диабете, составило около 48,5 лет со стандартным отклонением около 13,1 года. Большинство пациентов были в возрасте от 40 до 60 лет, но были и отклонения, например, отдельные случаи старше 90 лет.

2. Соотношение полов: доля мужчин и женщин в наборе данных примерно одинакова: мужчины составляют около 53,2 %, а женщины - около 46,8 %. Это говорит о том, что набор данных сбалансирован по полу и не будет влиять на последующие анализы.

3. Связь между индексом массы тела и средним артериальным давлением: диаграмма рассеяния показывает, что существует корреляция между индексом массы тела (ИМТ) и средним артериальным давлением с коэффициентом корреляции около 0,24. Это говорит о том, что существует небольшая тенденция к повышению среднего артериального давления при увеличении ИМТ.

4. Графики показателей уровня заболевания показывают изменения в степени прогрессирования болезни. Графики показывают медиану, квартили и выбросы данных. Несмотря на наличие некоторых выбросов, общее распределение показателей степени заболевания относительно равномерное. Это говорит о том, что большинство пациентов находятся в схожем диапазоне уровней прогрессирования заболевания, но есть несколько пациентов с более тяжелым прогрессированием заболевания, которые требуют дополнительного внимания.

5. По нашим наблюдениям, распределение количественных показателей болезни через год имело нормальный характер с одним пиком. Среднее значение составило около 152, медиана - около 140, а стандартное отклонение - около 77. У большинства пациентов количественные показатели болезни были сосредоточены между 100 и 200, что соответствует общей тенденции прогрессирования заболевания. Однако мы также отметили, что у нескольких пациентов количественные показатели болезни превышали 300, что может свидетельствовать о более серьезном прогрессировании заболевания у этих пациентов.

В итоге, анализируя набор данных по диабету и наблюдая за конкретными значениями, мы можем получить более глубокое понимание характеристик пациента, а также развития заболевания, что дает более конкретные и надежные ориентиры для дальнейших исследований и лечения.