Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент Лю Цзычжан

группы ИУ5И-23М

Москва — 2025 г.

1. Цель лабораторной работы

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

2. Задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

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

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. 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.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

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

3. Ход выполнения работы

3.1. Текстовое описание набора данных

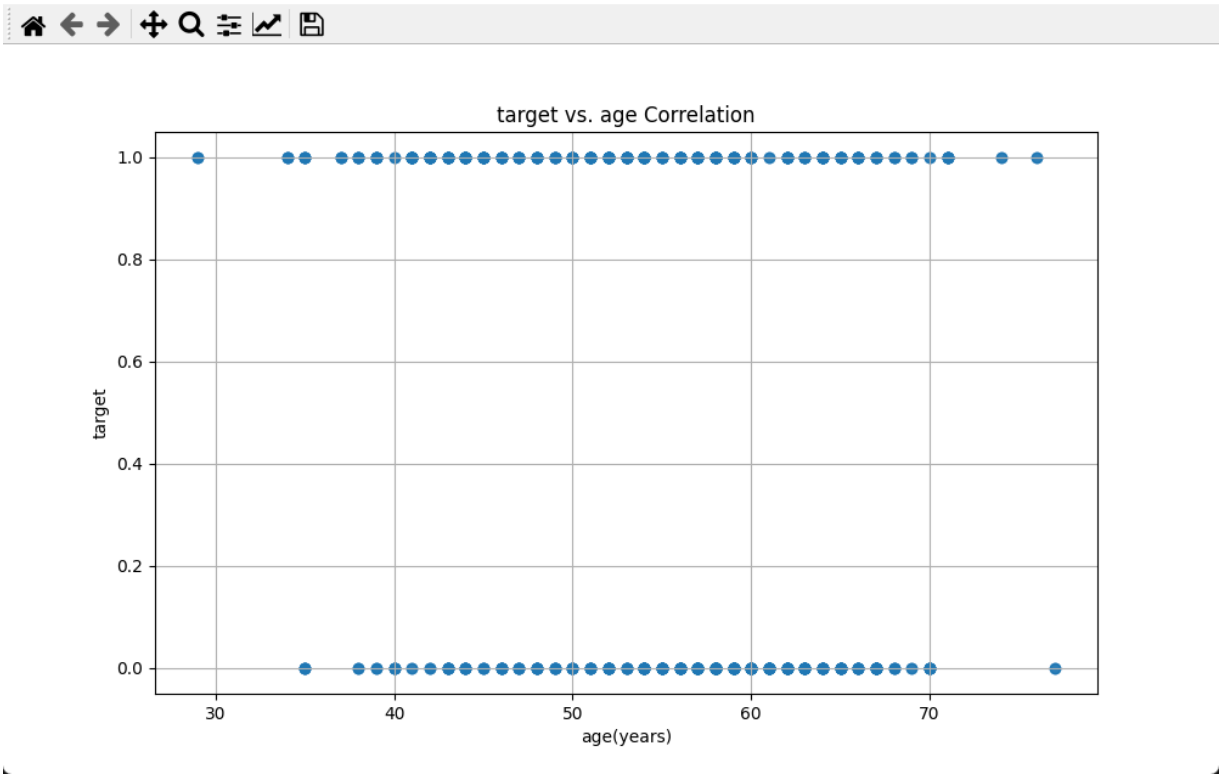
Этот набор данных датируется 1988 годом и состоит из четырех баз данных: Кливленд, Венгрия, Швейцария и Лонг-Бич V. Он содержит 76 атрибутов, включая прогнозируемый атрибут, но все опубликованные эксперименты ссылаются на использование подмножества из 14 из них. Поле «target» относится к наличию у пациента заболевания сердца. Оно имеет целочисленное значение 0 = нет заболевания и 1 = заболевание.

- **The dataset includes the following columns:**
- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestoral in mg/dl

- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

3.2. Основные характеристики набора данных

```
# target和age的相关性散点图
plt.figure(figsize=(10, 6))
plt.scatter(df['age'], df['target'], alpha=0.6)
plt.title('target vs. age Correlation')
plt.xlabel('age(years)')
plt.ylabel('target')
plt.grid(True)
plt.show()
```



```python
# trestbps和age的相关性散点图
plt.figure(figsize=(10, 6))
plt.scatter(df['age'], df['trestbps'], alpha=0.6)
plt.title('trestbps vs. age Correlation')
plt.xlabel('age(years)')
plt.ylabel('trestbps')
plt.grid(True)
plt.show()
```

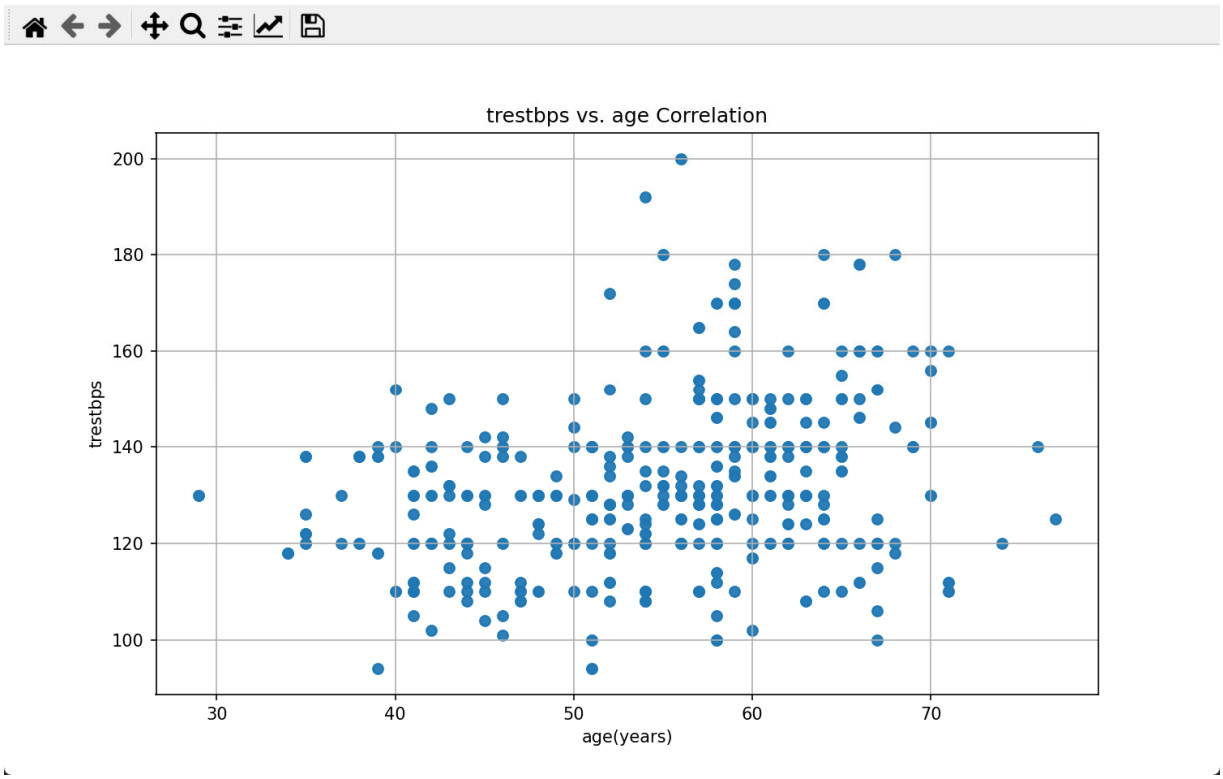


Рисунок 1:Точечная диаграмма корреляции между артериальным давлением и BIM

```python
# target占比饼状图
outcome_counts = df['target'].value_counts()
plt.figure(figsize=(7, 7))
plt.pie(outcome_counts, labels=['target (0)', 'target (1)'], autopct='%1.1f%%', startangle=140, colors=['skyblue', 'salmon'])
plt.title('target')
plt.show()
```

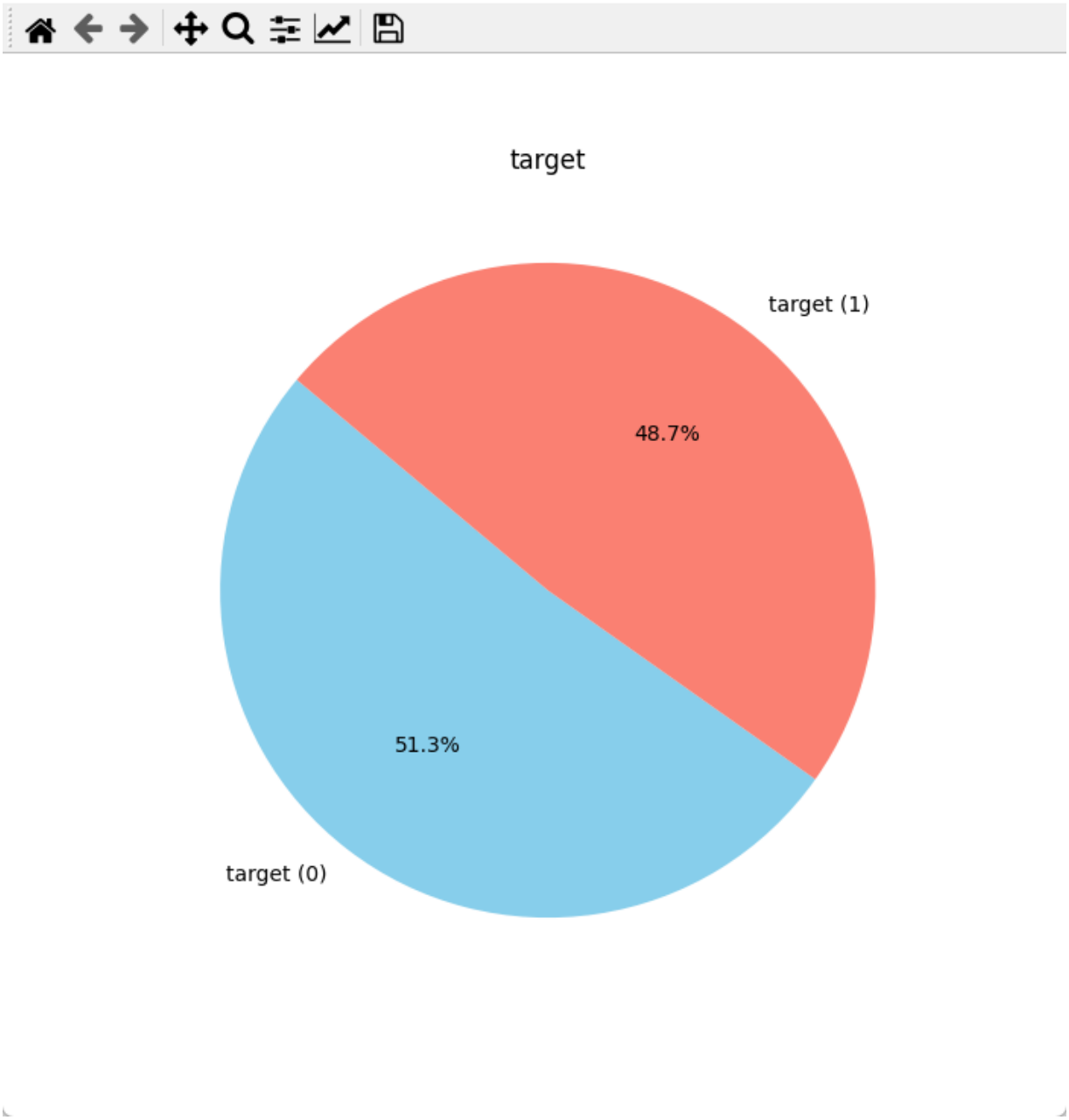


Рисунок 2:Круговая диаграмма результатов лечения сахарным диабетом

```python
# resting blood pressure 的分布箱线图
plt.figure(figsize=(10, 6))
plt.boxplot(df['trestbps'].dropna())
plt.title('resting blood pressure  Distribution')
plt.xticks([1], ['trestbps'])
plt.ylabel('target')
plt.grid(True)
plt.show()
```

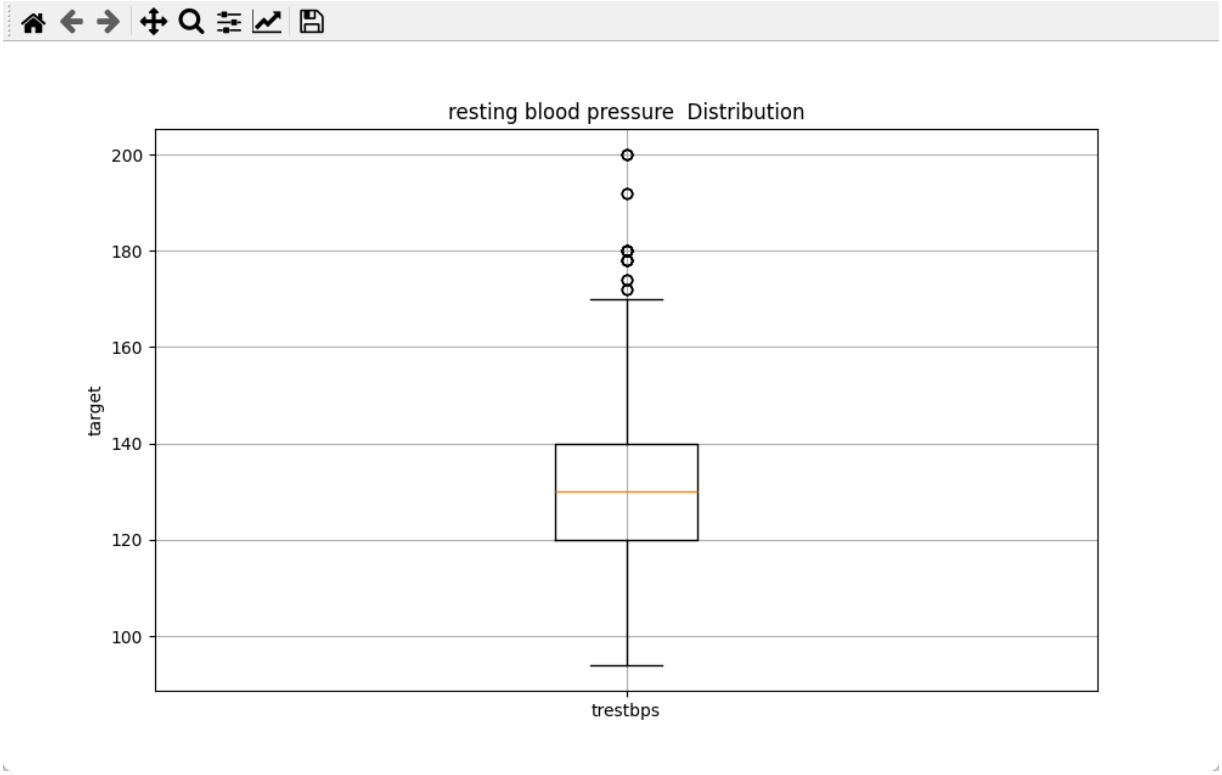


Рисунок 3:Прямоугольный график распределения артериального давления

```
# serum cholestoral in mg/dl和target的直方图
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.hist(df['chol'], bins=30, color='skyblue', edgecolor='black')
plt.title('Skin Thickness Distribution')
plt.xlabel('chol')
plt.ylabel('chol')
plt.subplot(1, 2, 2)
plt.hist(df['target'], bins=30, color='salmon', edgecolor='black')
plt.title('target')
plt.xlabel('target')
plt.ylabel('chol')
plt.tight_layout()
plt.show()
```

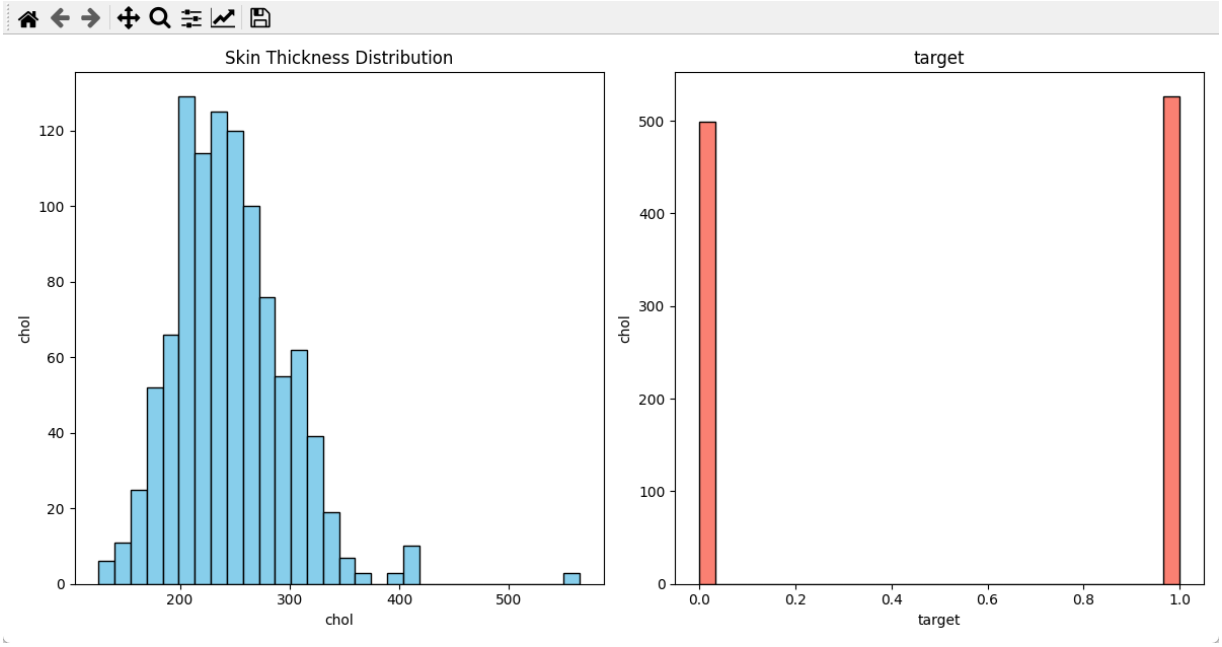


Рисунок 4:Гистограмма толщины кожи и ИМТ

```
columns_of_interest = ['trestbps', 'age', 'chol']
corr = df[columns_of_interest].corr().values
plt.figure(figsize=(8, 6))
plt.imshow(corr, cmap='coolwarm', interpolation='nearest')
plt.colorbar()
plt.xticks(range(len(columns_of_interest)), columns_of_interest, rotation=45)
plt.yticks(range(len(columns_of_interest)), columns_of_interest)
plt.title('Correlation Heatmap')
for i in range(len(corr)):
    for j in range(len(corr)):
        text = plt.text(j, i, f"{corr[i, j]:.2f}", ha="center", va="center", color="w")
plt.show()
```

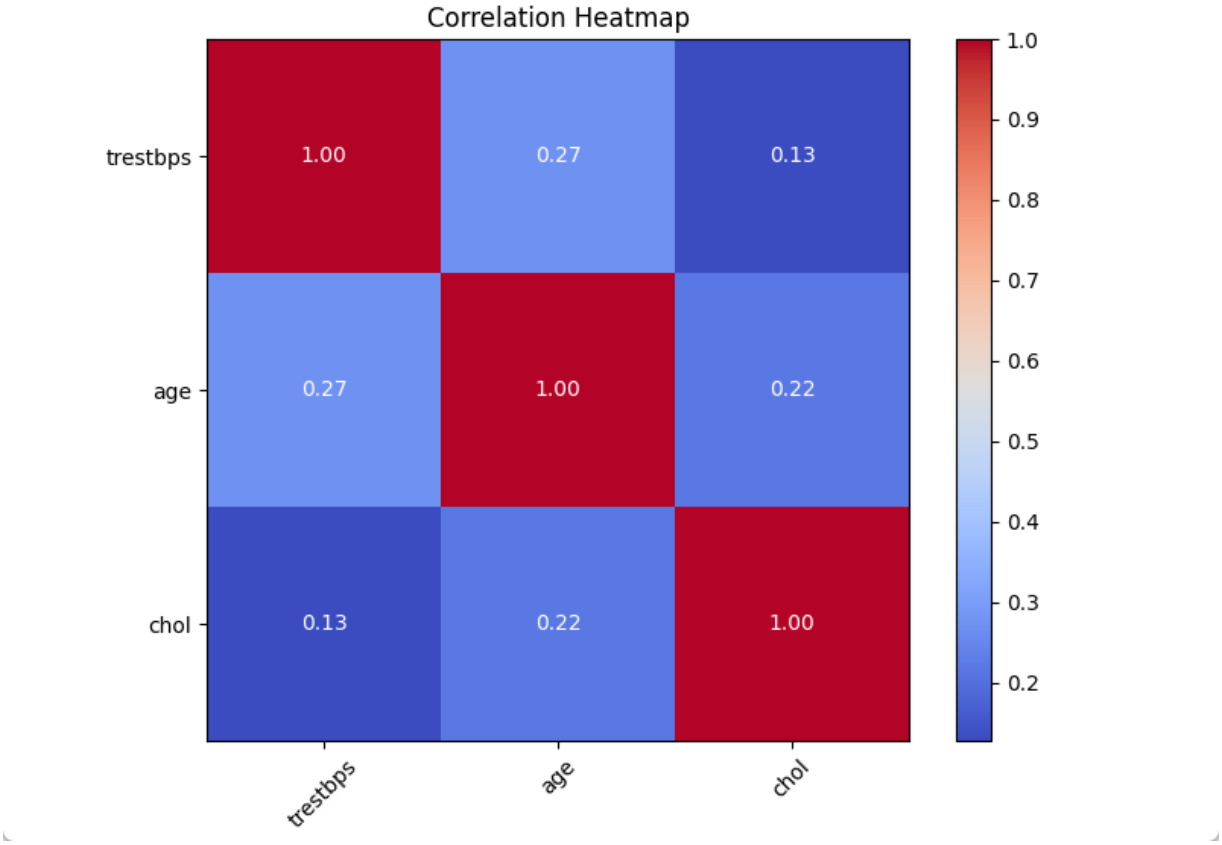


Рисунок 5:Корреляция между функцией спектра диабета, возрастом и тепловой картой артериального давления