МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_1\_\_**

по дисциплине«Методы машинного обучения в автоматизированных системах обработки информации и управления»

Тема: « Истории о данных»

ИСПОЛНИТЕЛЬ: \_\_Се Цзявэнь\_\_\_\_

ФИО

группа ИУ5И-22М\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_15\_"\_\_\_\_\_02\_\_\_\_2024\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Задания

* Выбрать набор данных (датасет).
* 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.

# Программы

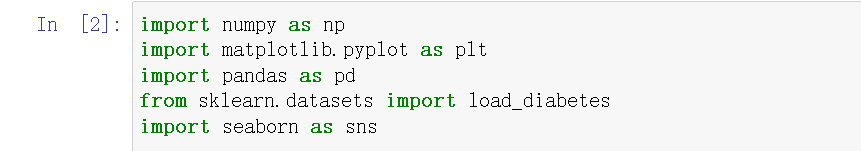
* Текстовое описание набора данных:

Был выбран набор данных о диабете из sklearn. Для каждого из n = 442 пациентов с диабетом были получены десять исходных переменных: возраст, пол, индекс массы тела, среднее артериальное давление и шесть измерений сыворотки крови, а также интересующий ответ - количественный показатель прогрессирования заболевания через год после исходного уровня. Переменные называются «AGE SEX BMI BP S1 S2 S3 S4 S5 S6 Y».

Подробности набора данных следующие:

|  |  |
| --- | --- |
| Количество выборочных данных | 442 |
| Количество функций (размерность данных) | 10 |
| Значение каждой функции | Возраст, пол, ИМТ, среднее артериальное давление, S1, S2, S3, S4, S5, S6. |
| Диапазон значений функции | (-0.2,0.2) |
| Значение тега | На основе количественных измерений через год после прогрессирования заболевания. |
| Диапазон значений тега | [25,346] |

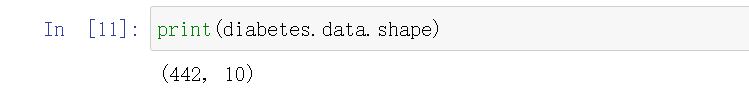
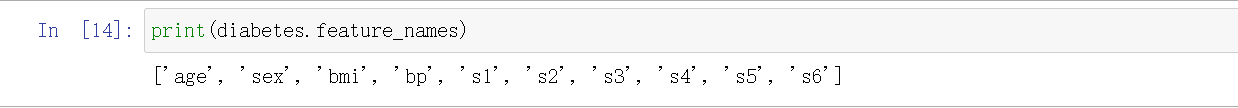
* Импорт библиотек



* Загрузка данных



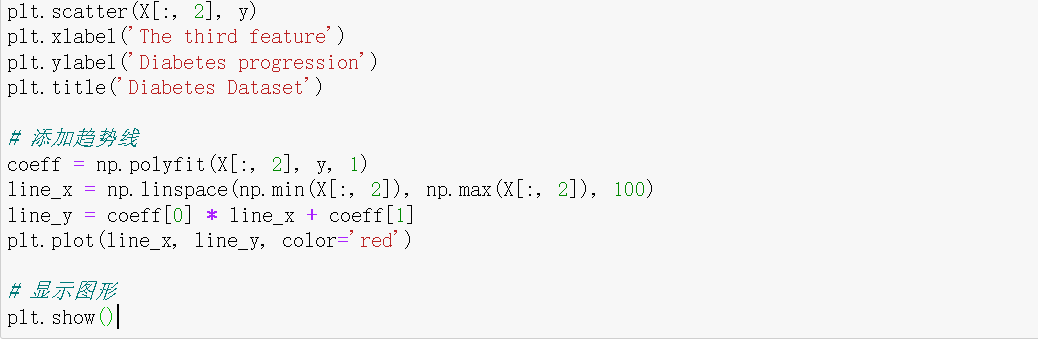
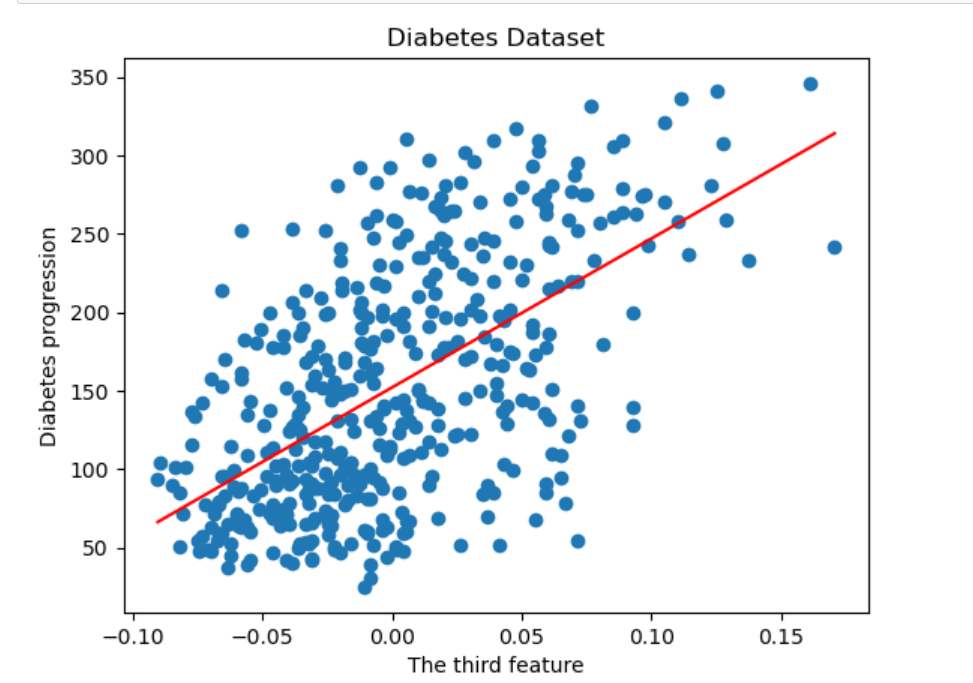
* Основные характеристики датасета

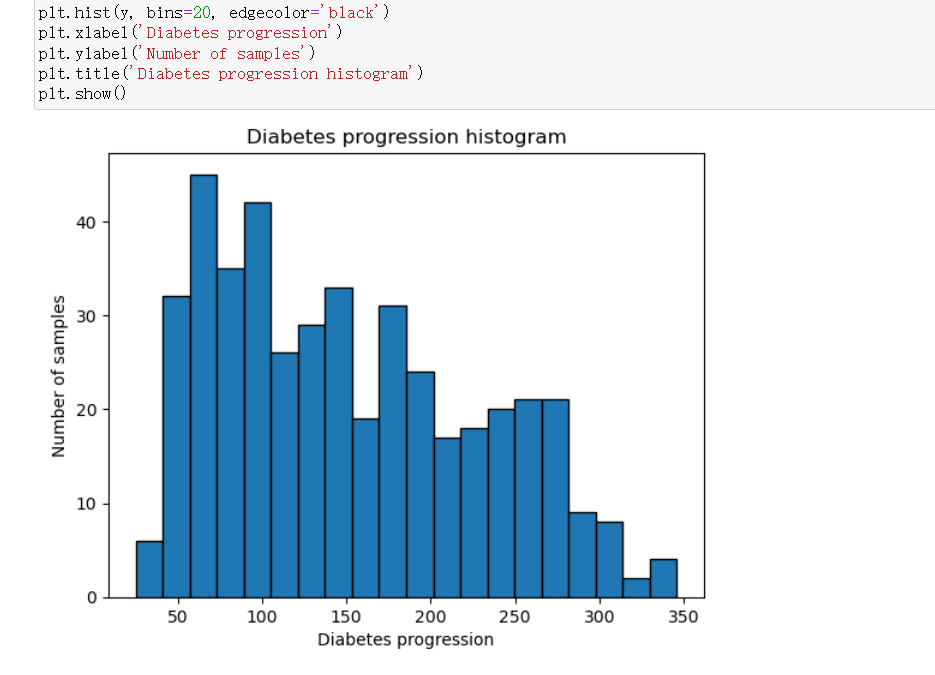
* Визуальное исследование датасета

1. **Диаграмма рассеяния：**

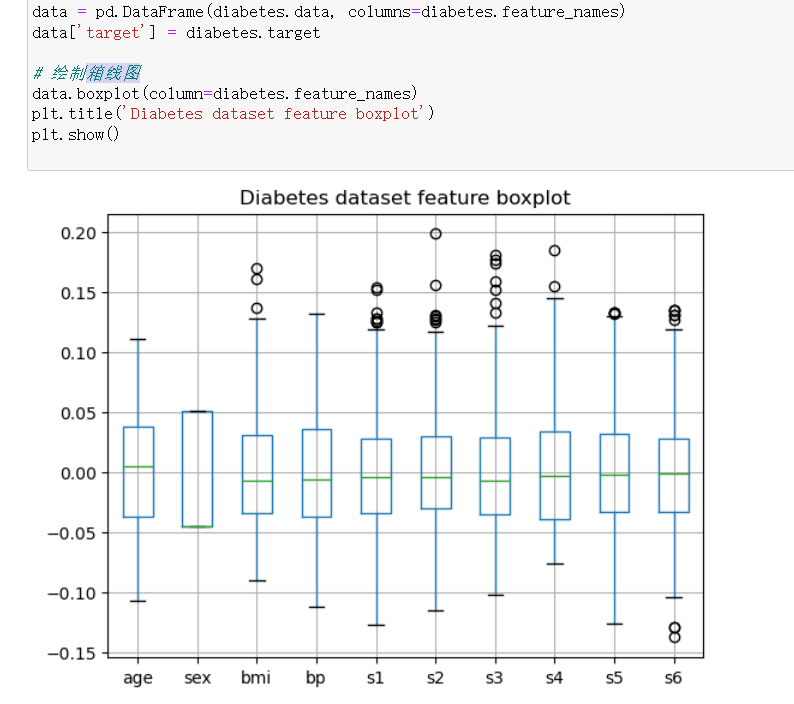
Постройте график распределения третьего признака и прогрессирования диабета в наборе данных и визуально определите наличие зависимостей.

1. **Гистограмма：**

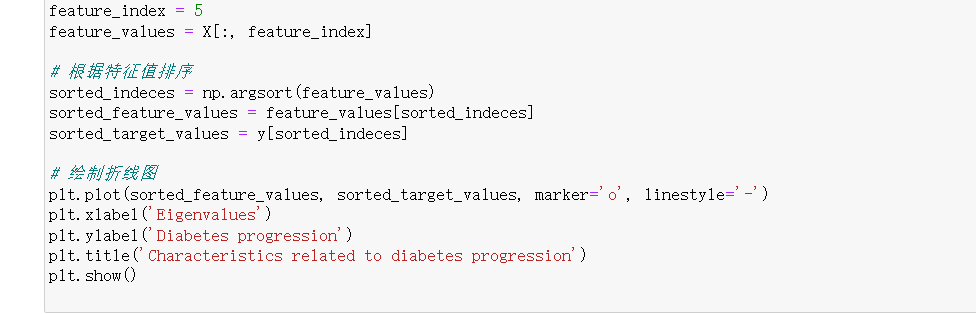
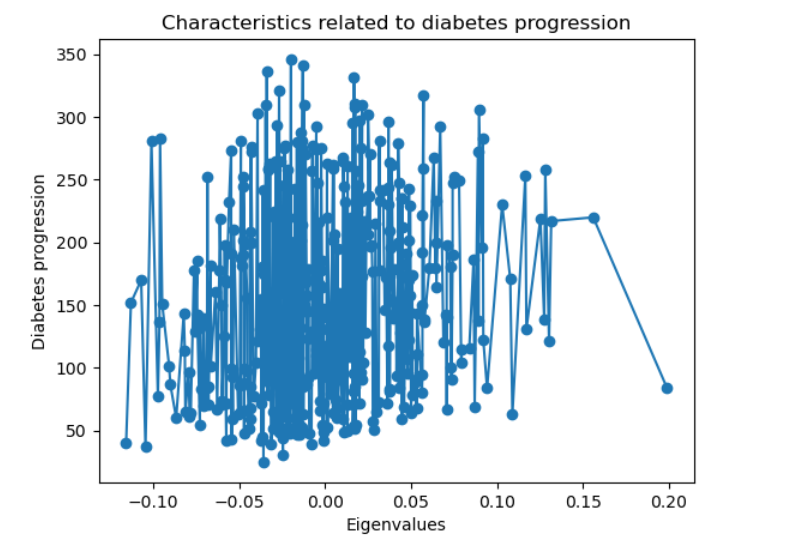
Позволяет оценить плотность вероятности распределения данных. 

1. **Коробчатый сюжет：**

Ящичные диаграммы могут четко отображать распределение данных, выявлять выбросы, сравнивать различные группы данных и отображать статистические характеристики. 

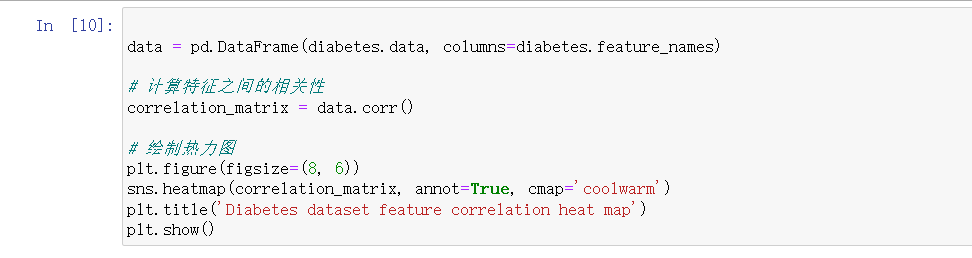
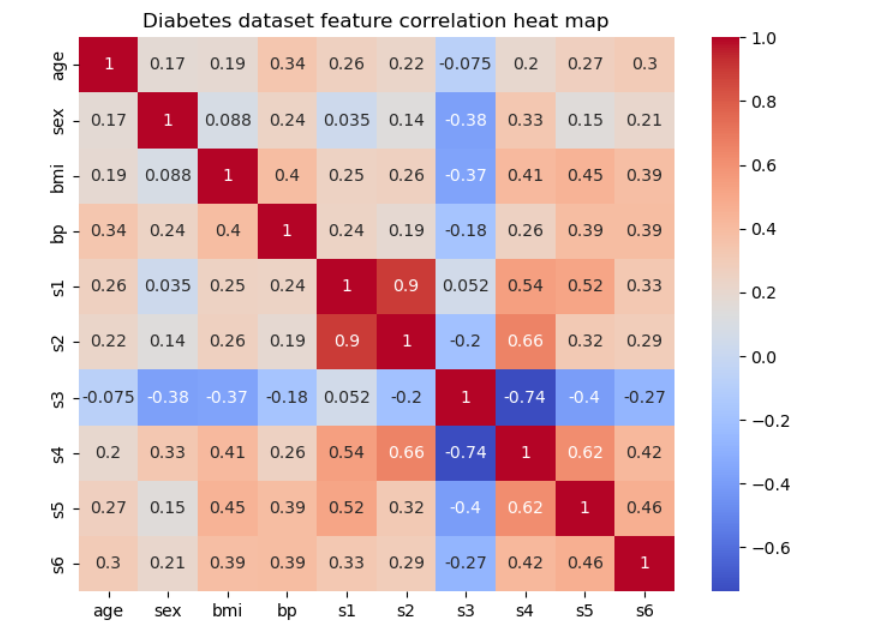
1. **Линейный график:**

Линейный график может четко показать связь тренда цели с другими непрерывными переменными.

1. **Тепловая карта:**

Тепловая карта, визуализирующая корреляционную матрицу.

# Вывод

Используя набор данных о диабете в sklearn для анализа данных, мы можем сделать следующие выводы:

Линейный график показывает тенденцию изменения На основе цифровых измерений через год после прогрессирования заболевания с шестым признаком s2.

Гистограмма показывает распределение чисел на разных уровнях заболевания, при этом наибольшее количество людей имеет уровни заболевания между 70 и 90.

Тепловая карта показывает корреляцию между различными признаками, среди которых положительная корреляция между s1 и s2 достигает 0,9.

На диаграмме показано распределение данных по различным признакам, включая медиану, верхний и нижний квартиль, а также выбросы. Среди них больше всего колеблются данные о поле признака.

Диаграмма рассеяния показывает взаимосвязь между различными характеристиками, показывая корреляцию между ИМТ и распространенностью диабета.