Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №5

«Линейные модели, SVM и деревья решений»

Студент:

Гаранин А.В.

Группа ИУ5Ц-81Б

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение линейных моделей, SVM и деревьев решений.

**Задание:**

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите следующие модели:
   * одну из линейных моделей;
   * SVM;
   * дерево решений.
5. Оцените качество моделей с помощью двух подходящих для задачи метрик. Сравните качество полученных моделей.

Дополнительные задания:

1. Проведите эксперименты с важностью признаков в дереве решений.
2. Визуализируйте дерево решений.

**Выполнение работы:**

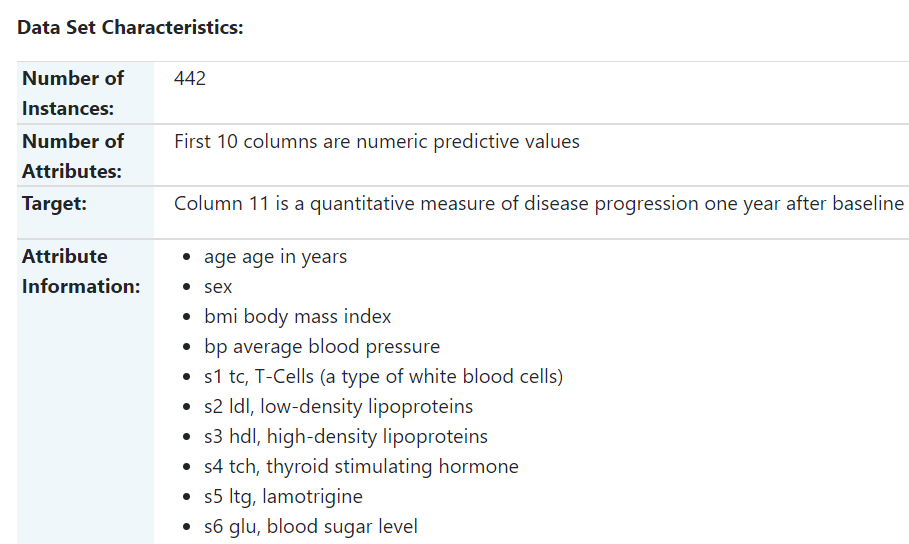
Выберем следующий датасет для решения задач регрессии:

*sklearn.datasets.load\_diabetes*

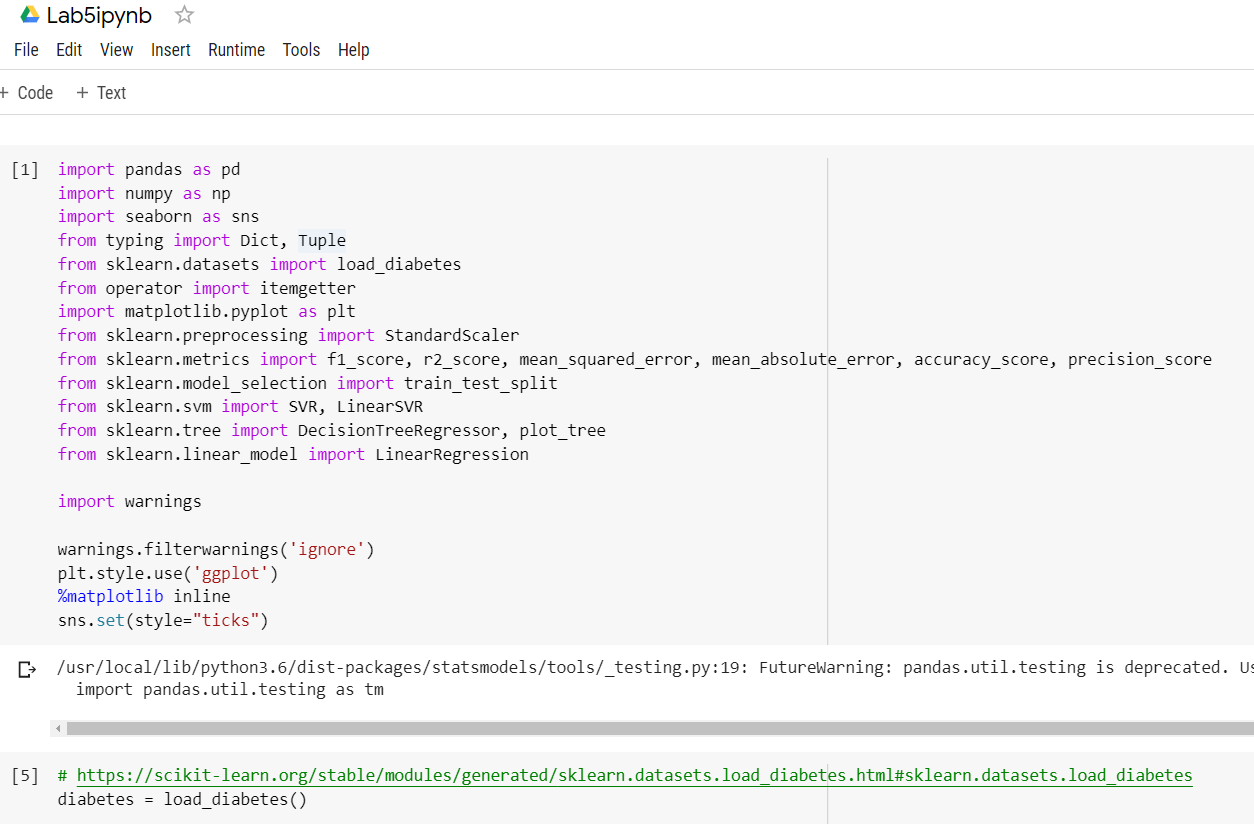
<https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html>

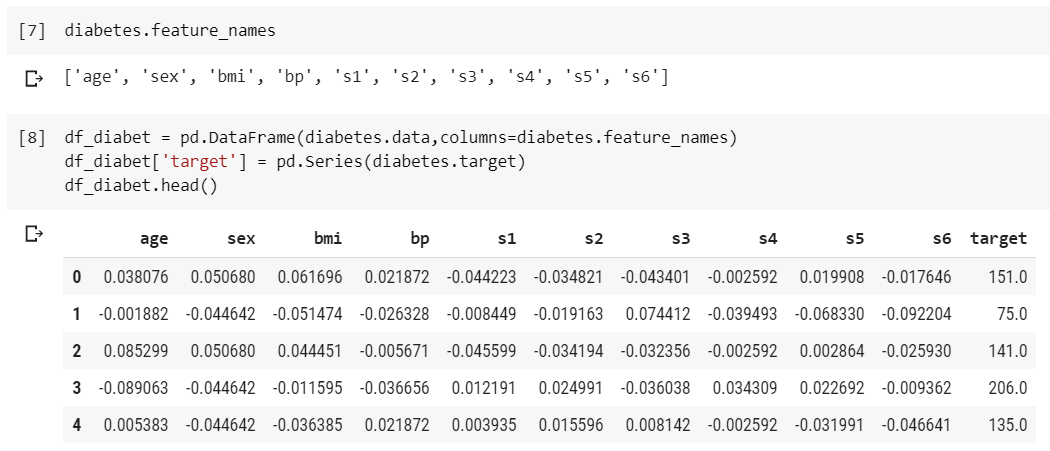
Описание датасета:

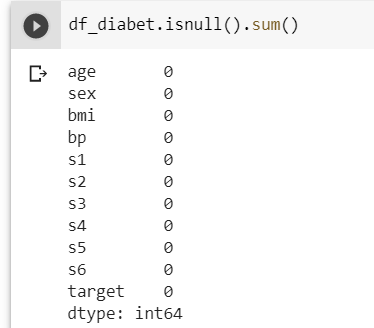
Для каждого из n = 442 больных сахарным диабетом были получены десять исходных переменных, возраст, пол, индекс массы тела, среднее артериальное давление и шесть измерений сыворотки крови, а также количественная мера прогрессирования заболевания через год после фиксации исходного уровня (целевой признак).



Подготовка данных

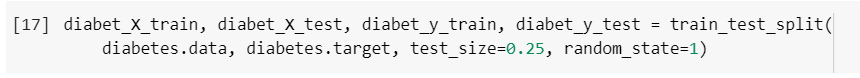




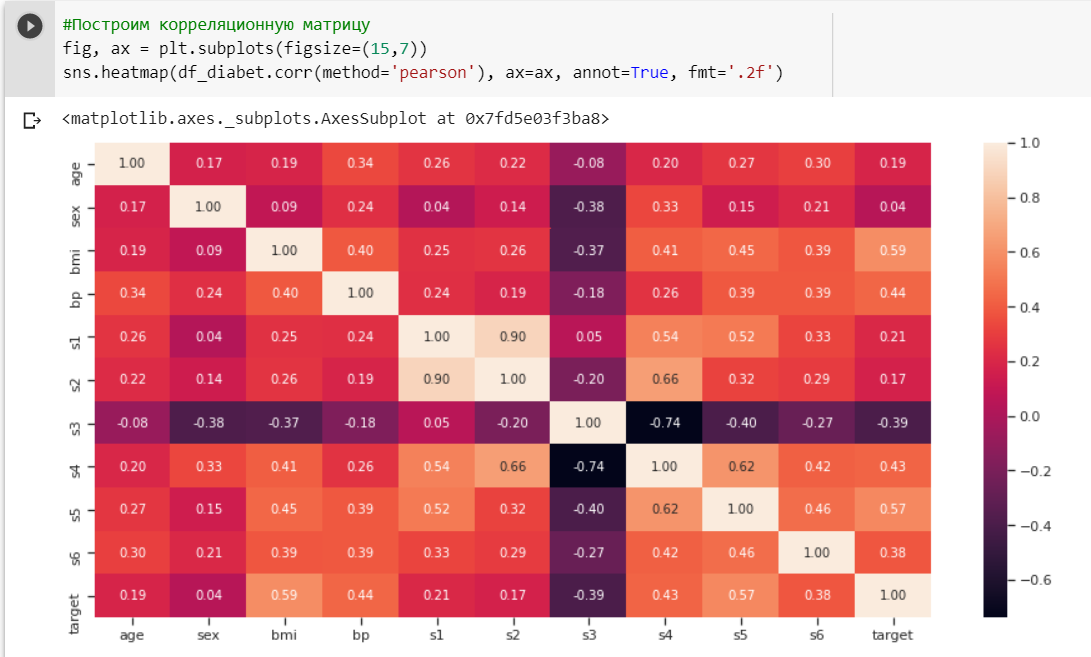


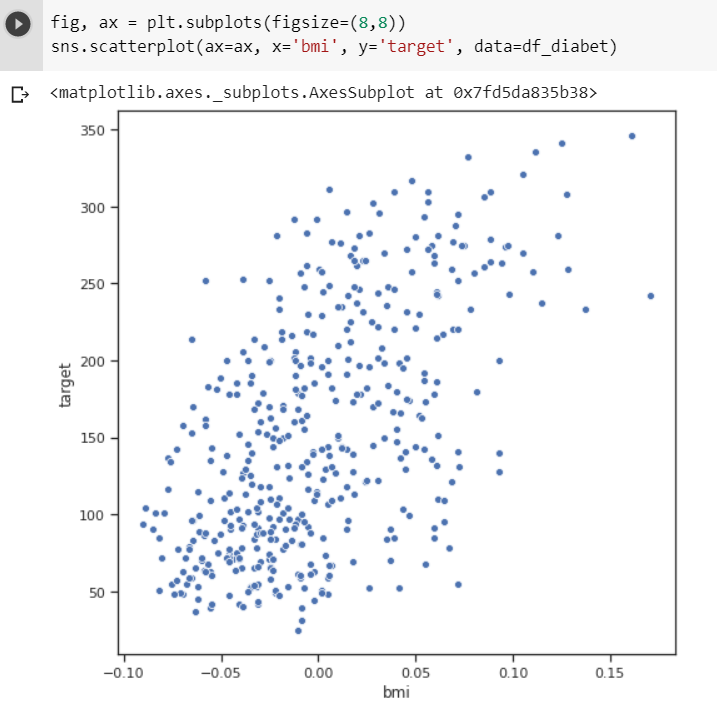
Видим, что пропусков данных нет

Разделим выборку на обучающую и тестовую:



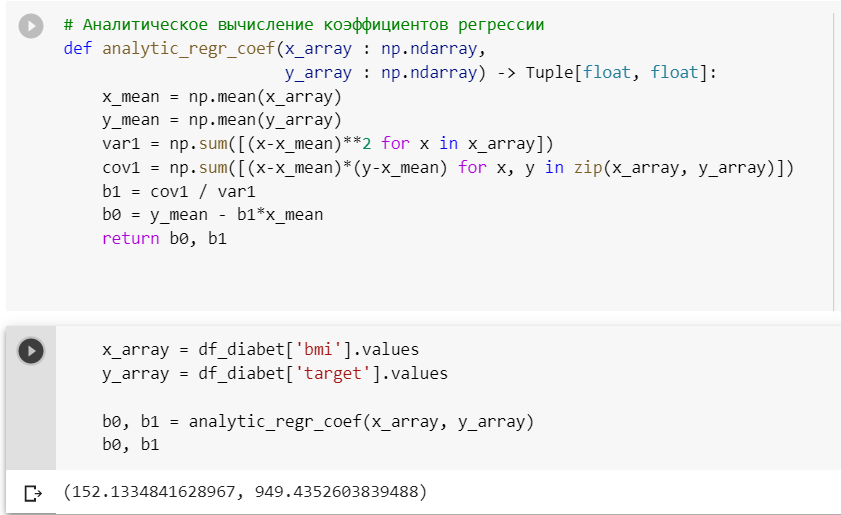
**Линейная модель**

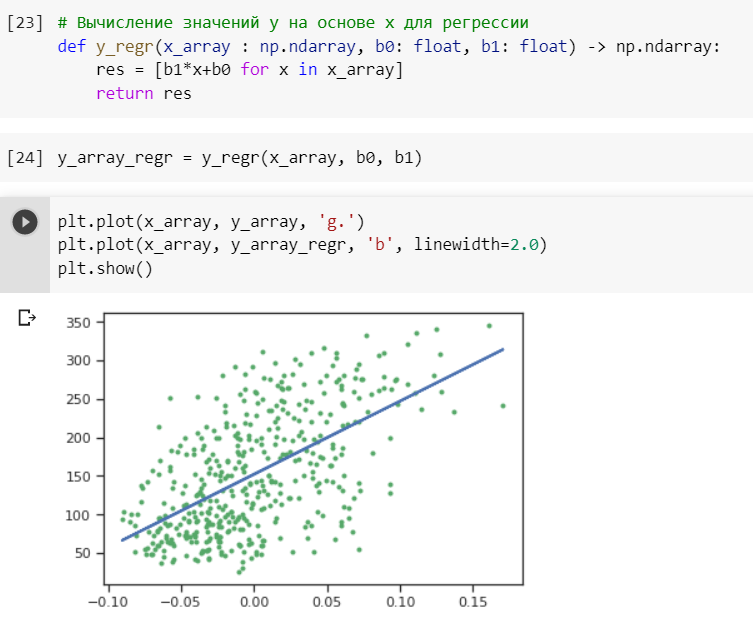




Между признаком "mbmi" и целевым признаком "target" существует зависимость, наиболее близкая к линейной, коэффициент корреляции = 0,59

Попробуем восстановить данную линейную зависимость.





Градиентный спуск

### 

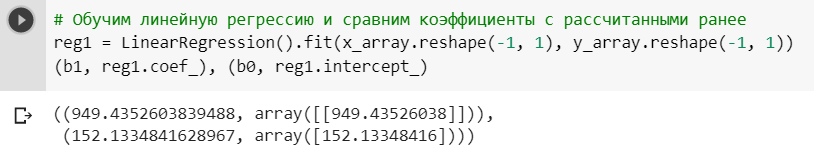
### 

### 

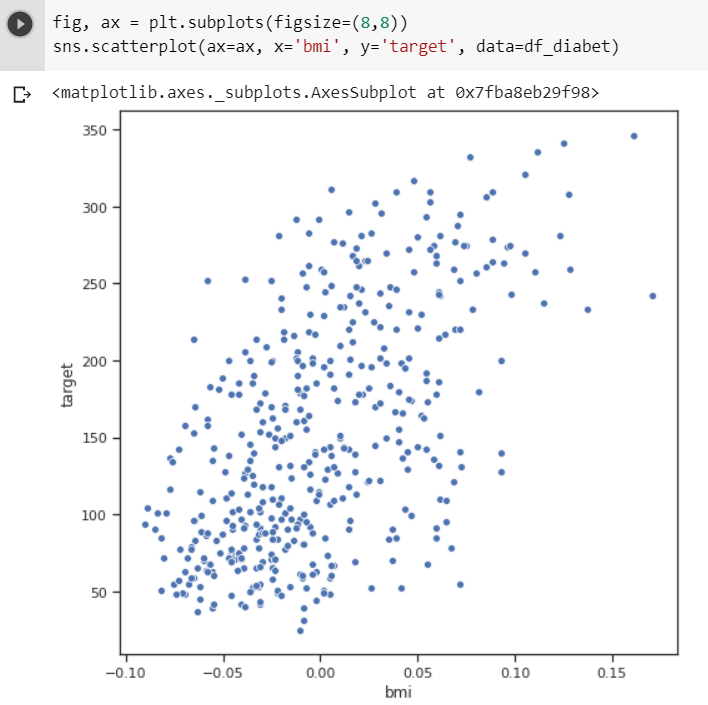
### 

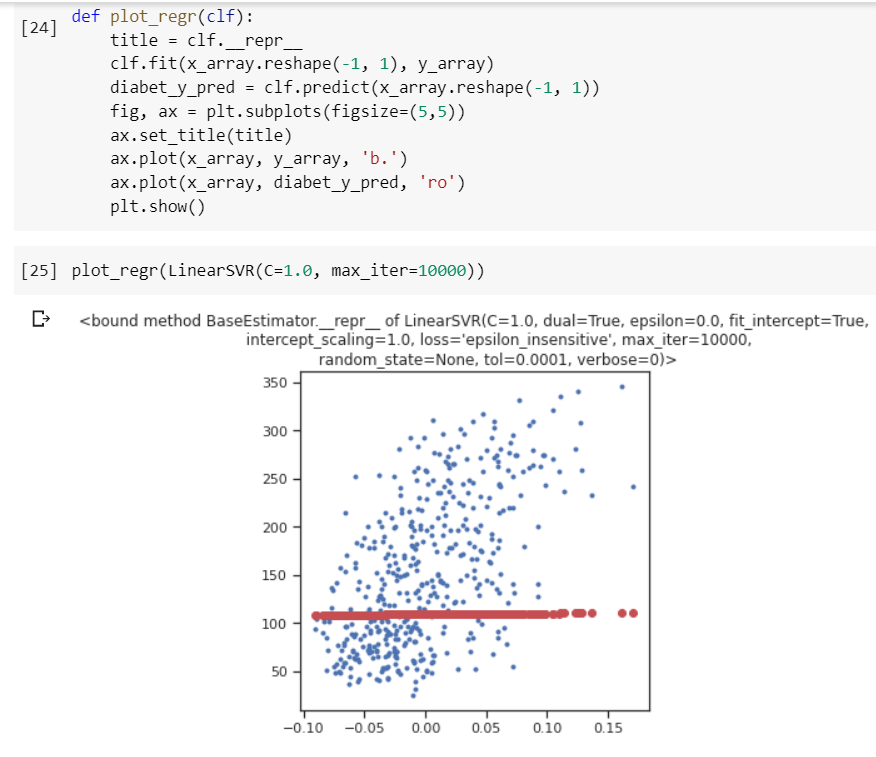
### 

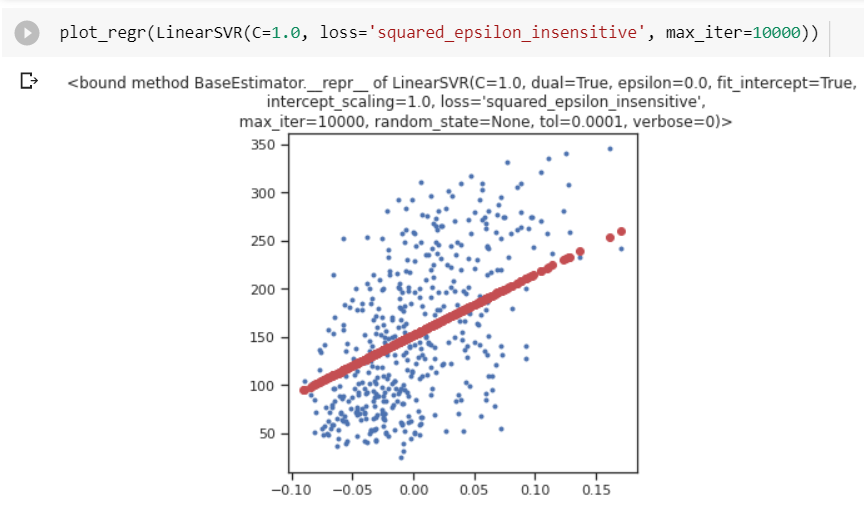
### Использование класса [LinearRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) библиотеки scikit-learn.

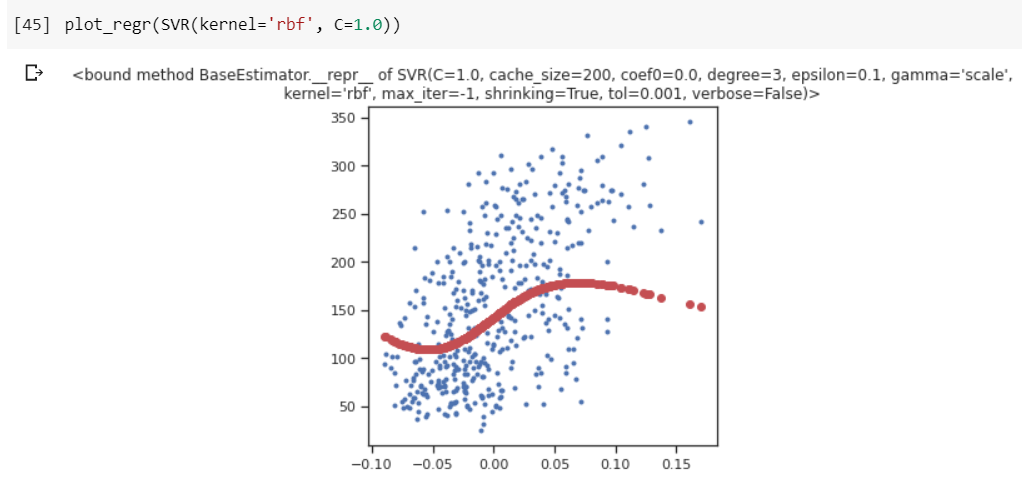


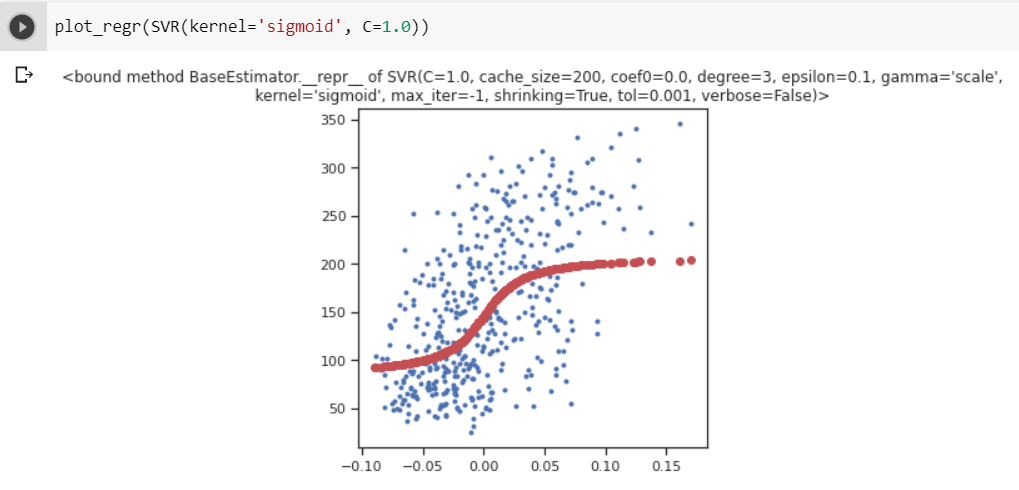
**Машина (метод) опорных векторов - SVM**

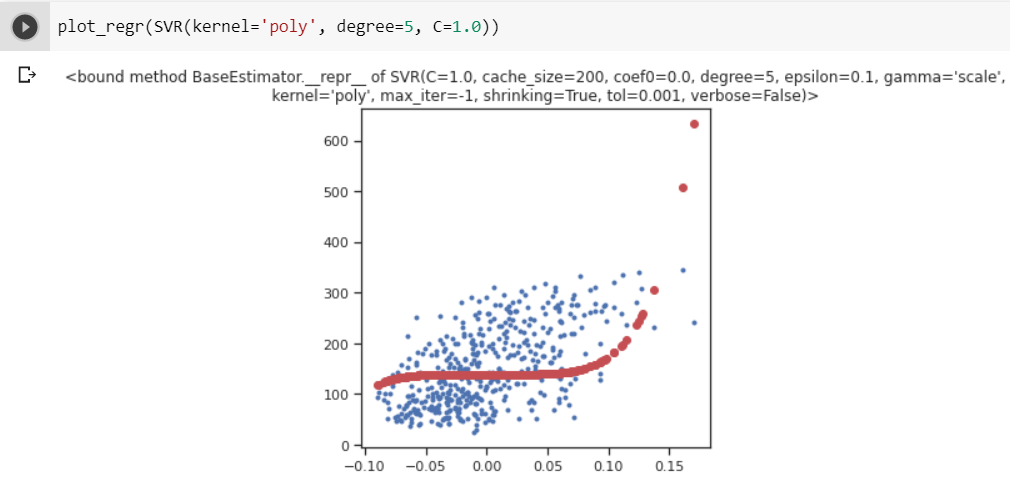


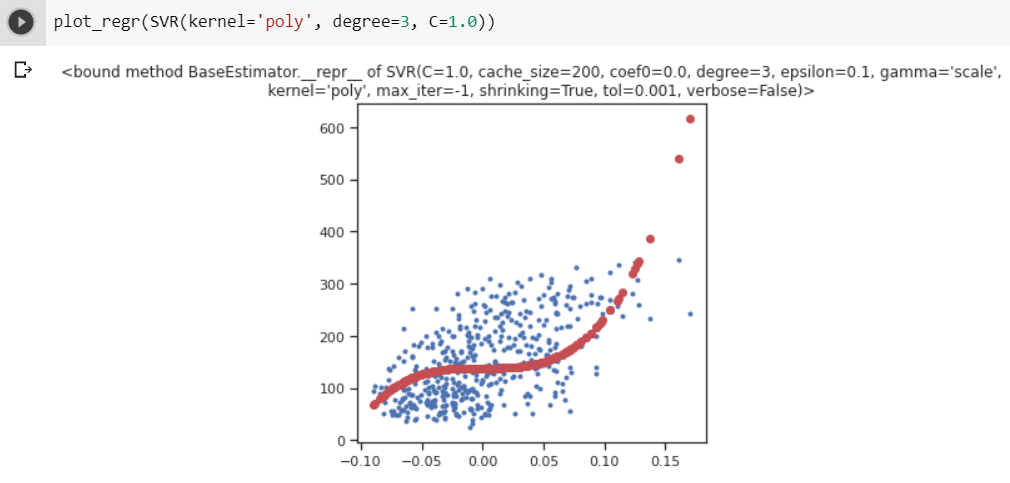




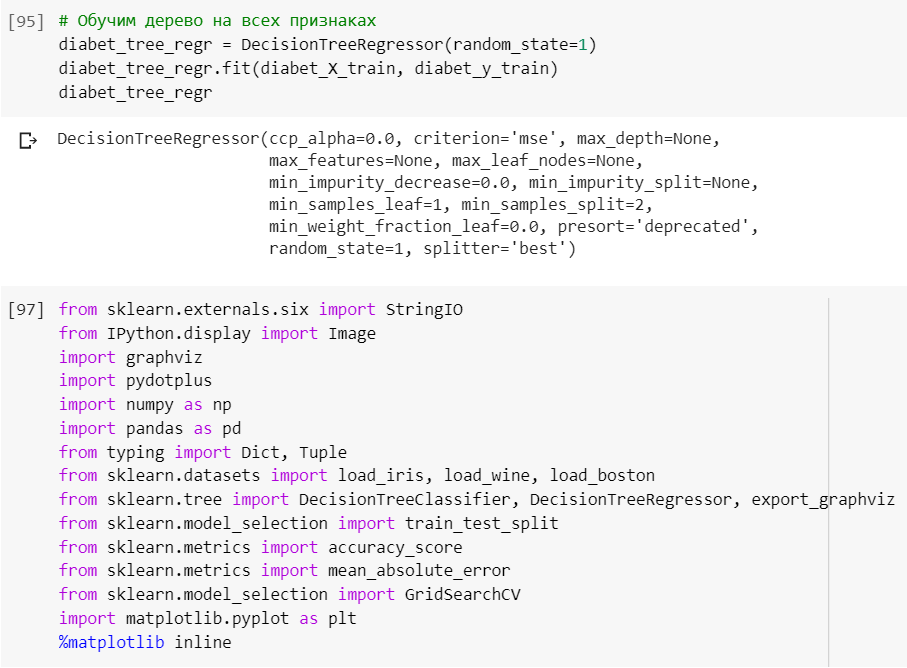


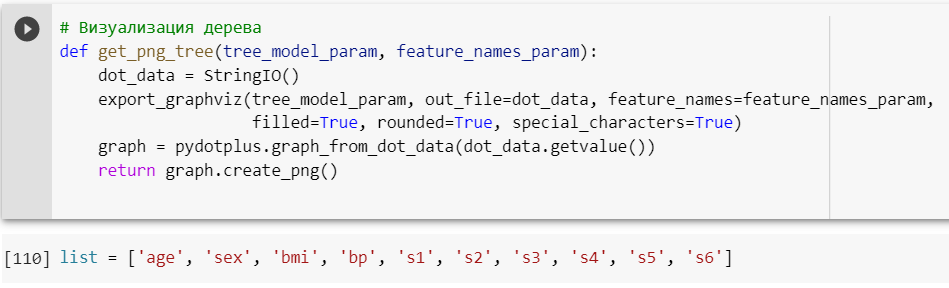


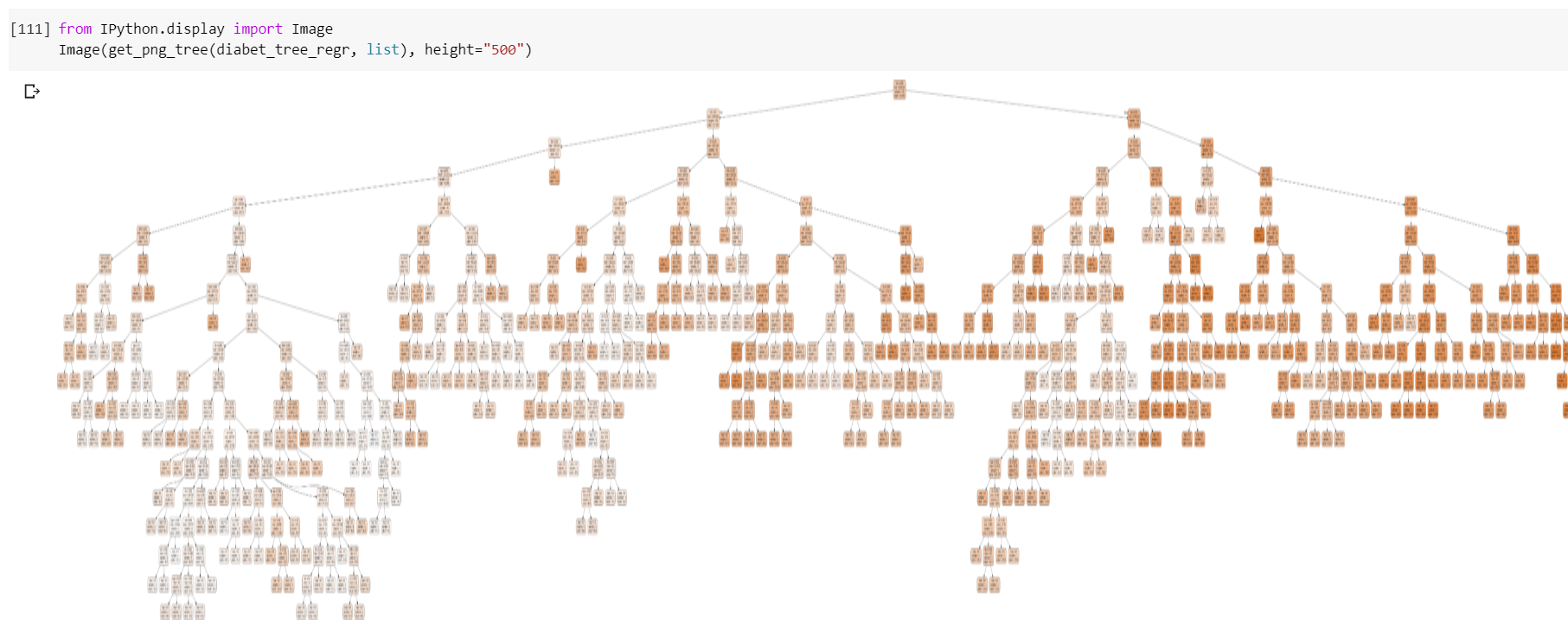


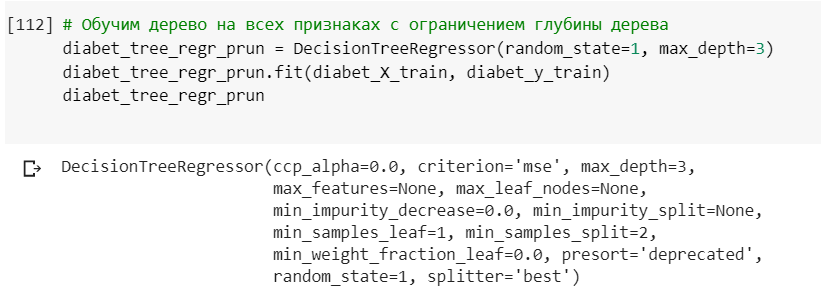


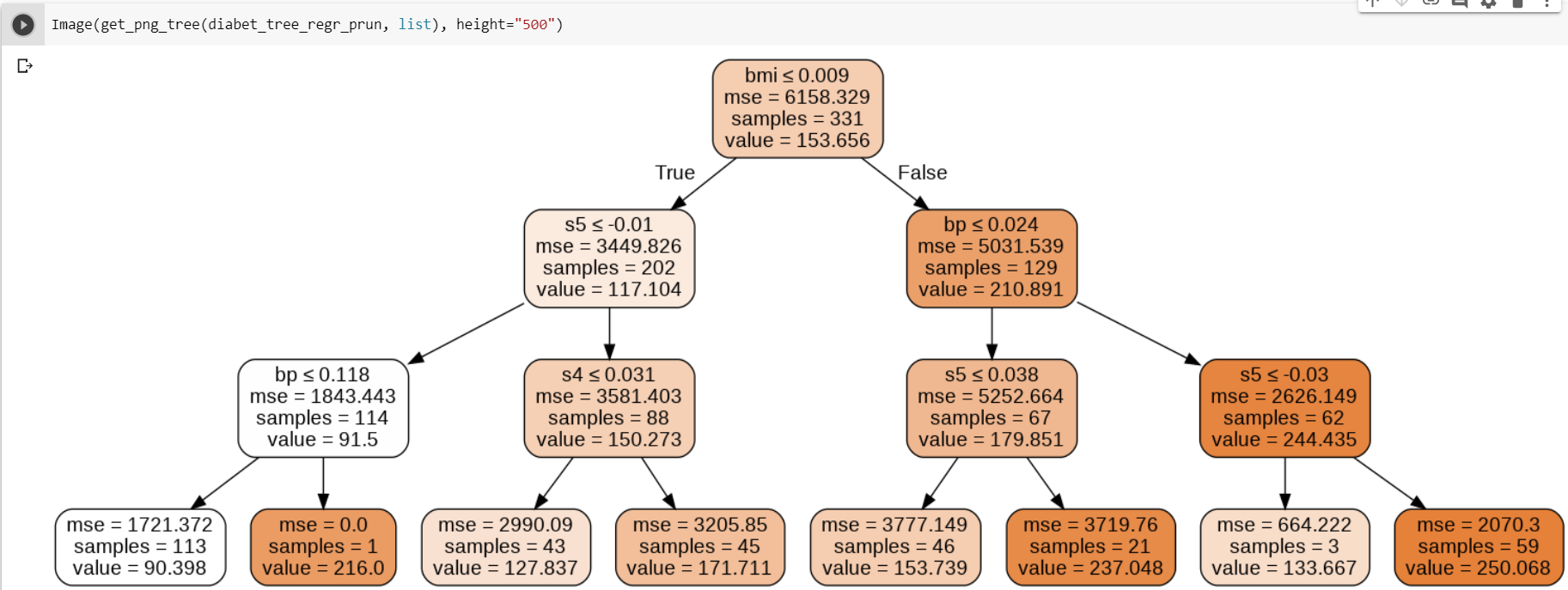
**Дерево решений**

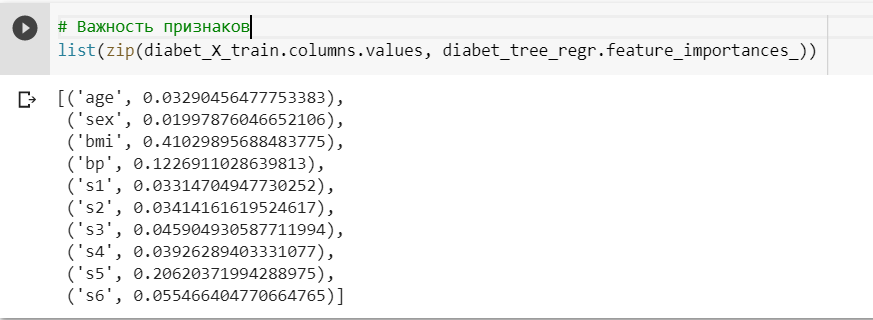


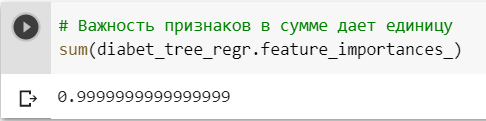




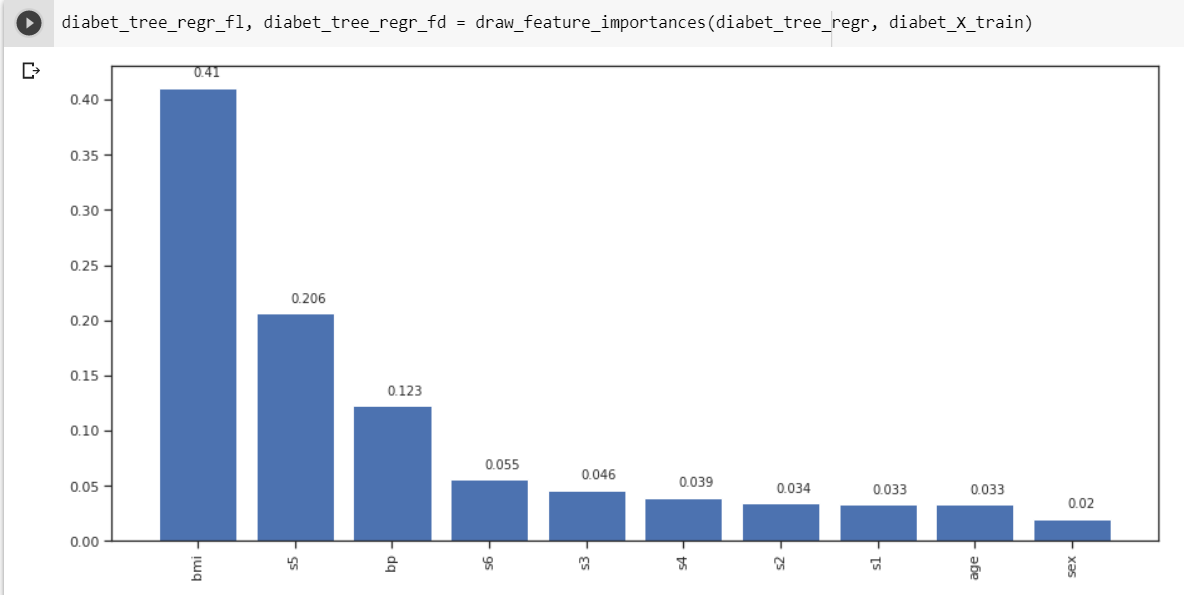


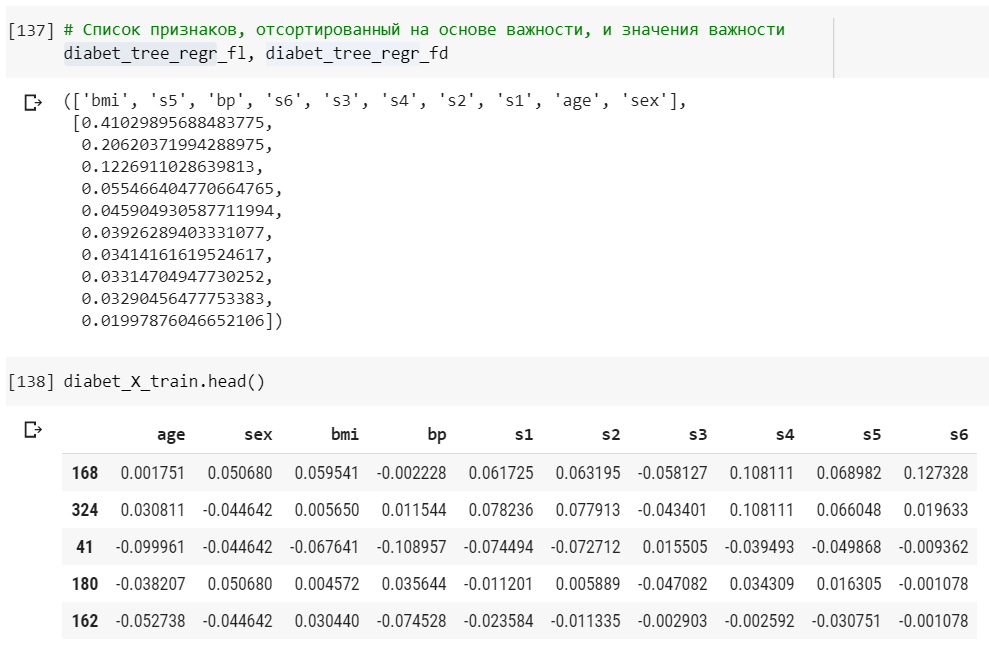


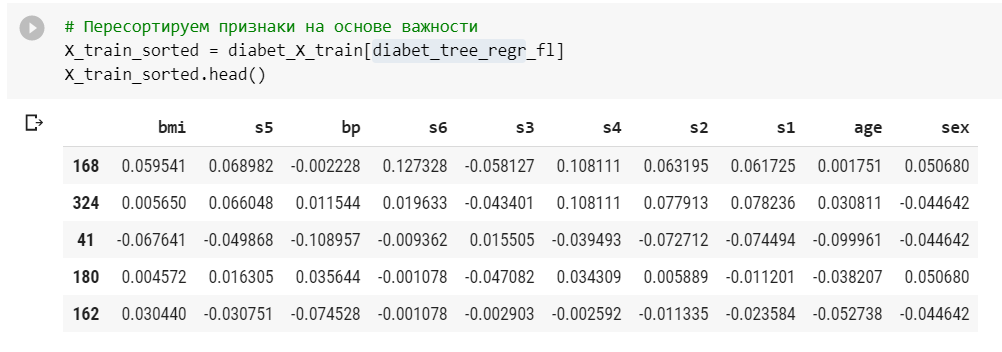


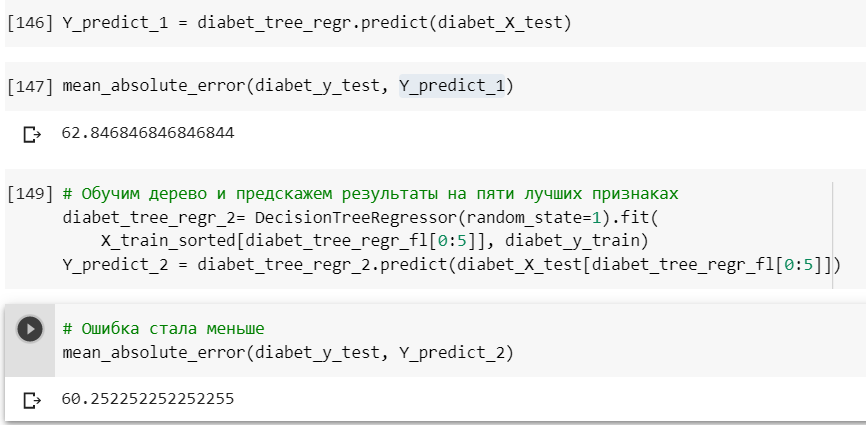


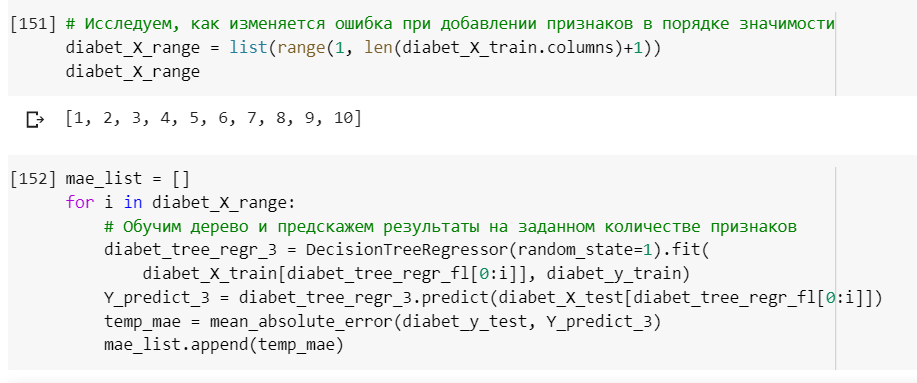


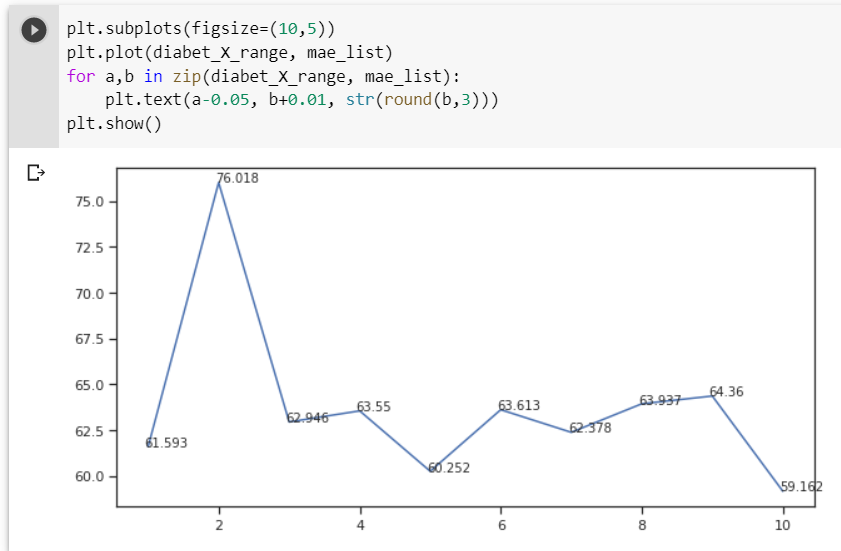






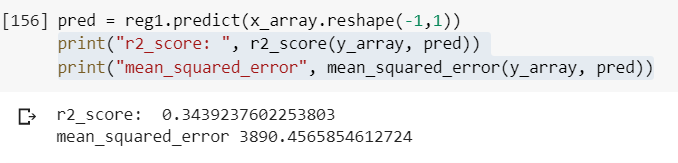




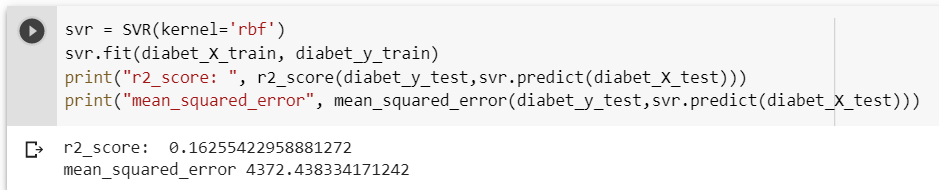


**Оценка качества моделей**

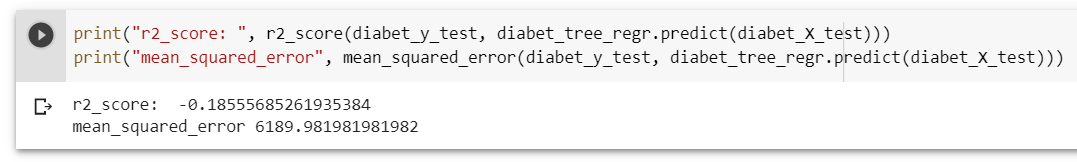
Линейная регрессия



Машина (метод) опорных векторов



Дерево решений



При использовании каждого из 3 методов мы получили значительную ошибку, но по метрике ошибки меньшее значение было получено при использовании метода линейной регрессии.