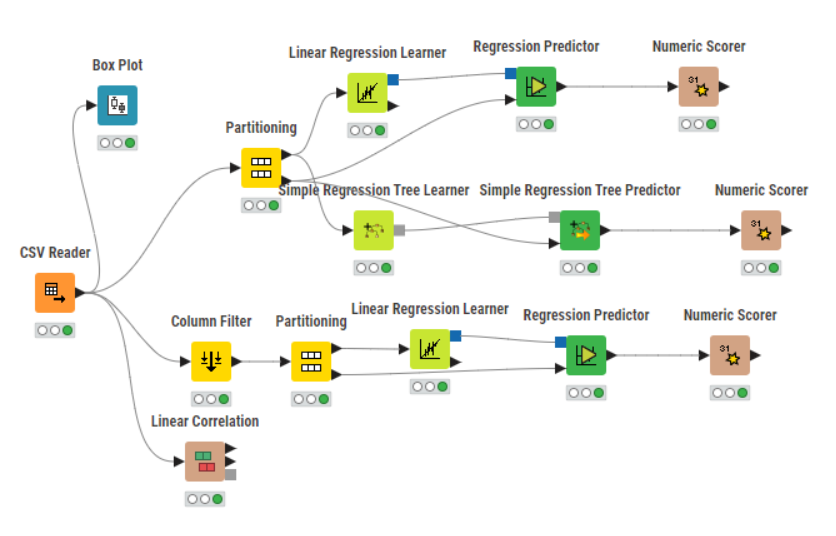
Аналитическая записка по результатам построения регрессионных моделей с использованием KNIME и Python

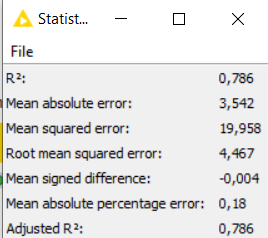


Данные:

* CRIM = уровень преступности
* ZN = разреженность населения (обратно к плотности)
* INDUS = доля предприятий не торгового характера среди всех предприятий квартала
* CHAS = район у реки (1 = да, 0 = нет)
* NOX = загазованность
* RM = среднее число комнат в квартире/доме среди домов этого квартала
* AGE = доля старых домов (построенных до 1940)
* DIS = расстояние до промышленных/деловых районов города (взвешенное)
* RAD = расстояние до радиальных автодорог
* TAX = величина налога на недвижимость в районе в $10,000
* PTRATIO = число учеников не одного учителя (в школах района)
* B = 1000(B - 0.63)^2, где B - доля афроамериканцев
* LSTAT = процент населения с низким соц. статусом
* **MEDV = медиана цен жилых домов в 1000 долларов**

ЦЕЛЕВАЯ ПЕРЕМЕННАЯ – MEDV

**Модель 1. Модель линейной регрессии по всему набору данных**



Коэффициент детерминации 0.786 –модель объясняет 78.6% изменчивости зависимой переменной. Это хороший показатель, но он указывает на то, что часть изменчивости остается неучтенной.

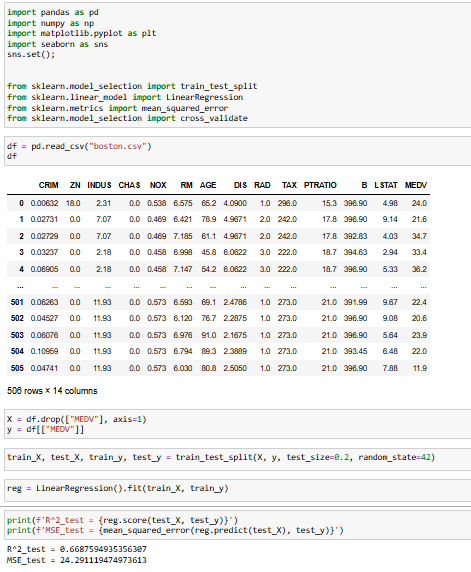
Средняя абсолютная ошибка 3.542 – в среднем, модель ошибается на 3.542 единицы, что говорит о достаточно точном предсказании.

Среднеквадратическая ошибка 19.958 – этот показатель усиленно наказывает большие ошибки, и его значение говорит о наличии некоторых отклонений в предсказаниях.

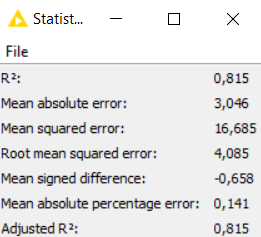
Корень среднеквадратической ошибки 4.467 – в среднем, предсказания модели отклоняются от реальных значений на 4.467 единицы, что указывает на умеренный уровень ошибки.

Средняя абсолютная процентная ошибка 18% – в среднем, модель ошибается на 18% от реального значения.

**Вывод:** Линейная регрессия по всему набору данных показывает хорошее качество предсказаний, но имеет ограниченную способность моделировать сложные нелинейные зависимости. Ошибки достаточно низкие, однако, если в данных присутствуют сложные закономерности, модель может не справляться с их предсказанием.



**Модель 2. Модель регрессии – дерево решений.**



Коэффициент детерминации 0.815 – модель объясняет 81.5% изменчивости зависимой переменной, что лучше, чем у линейной регрессии по всему набору данных.

Средняя абсолютная ошибка 3.046 – уменьшение ошибки по сравнению с первой моделью говорит о лучшем качестве предсказаний.

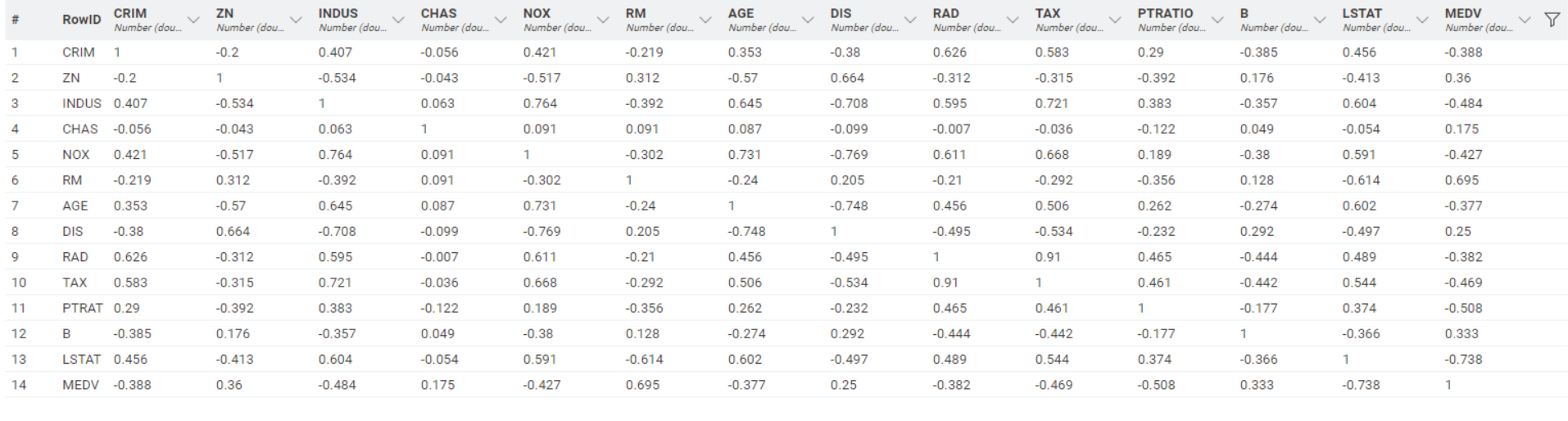
Среднеквадратическая ошибка 16.685 – более низкое значение по сравнению с первой моделью указывает на меньший разброс предсказаний.

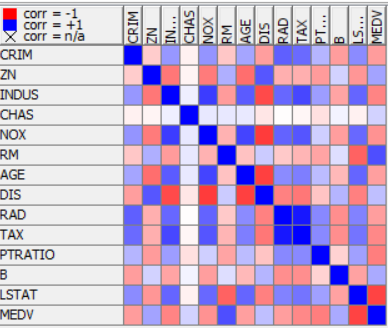
Корень среднеквадратической ошибки 4.085 – меньший показатель по сравнению с предыдущей моделью говорит о повышении точности.

Средняя абсолютная процентная ошибка 14.1% – снижение процентной ошибки указывает на повышение качества предсказаний.

**Вывод:** Линейная регрессия с деревом решений является наиболее точной из всех трех моделей. Она лучше справляется с описанием сложных зависимостей в данных, что приводит к более высоким значениям и меньшим ошибкам.

**3. Модель регресссии с отобранными признаками**



Построим матрицу корреляций признаков и целевой переменной. По результатам корреляционного анализа решено выбрать для построения признаки с корреляцией, превышающей 0.4 по модулю, а именно: 

INDUS(-0.484)

RM(0.695)

TAX(-0.469)

PTRAT(-0.508)

LSTAT(-0.738)

Изображение выглядит как текст, снимок экрана, Шрифт, число

Контент, сгенерированный ИИ, может содержать ошибки.

Коэффициент детерминации 0.655 – модель объясняет 65.5% изменчивости зависимой переменной, что значительно хуже по сравнению с первыми двумя моделями.

Средняя абсолютная ошибка 4.154 – самая высокая ошибка среди всех моделей, что говорит о худшем качестве предсказаний.

Среднеквадратическая ошибка 29.606 – высокий показатель указывает на значительный разброс предсказаний.

Корень среднеквадратической ошибки 5.441 – наибольшая ошибка среди всех моделей.

Средняя абсолютная процентная ошибка 22% – самая высокая процентная ошибка.

**Вывод:** Регрессия по ограниченному количеству признаков показывает худшие результаты. Урезанный набор данных приводит к потере важной информации, что снижает точность предсказаний. Высокие ошибки и низкий коэффициент детерминации подтверждают этот вывод.

**ОБЩИЙ ВЫВОД:**

Лучший результат показала модель линейной регрессии с деревом решений. Она имеет наивысшее значение (0.815) и наименьшие ошибки, что делает ее предпочтительной для предсказаний.

Линейная регрессия по всему набору данных также является достойным вариантом. Она имеет чуть худшие показатели по сравнению с деревом решений, но остается достаточно точной.

Регрессия по ограниченному числу признаков показала наихудший результат. Низкое значение и высокие ошибки говорят о потере значимой информации.

**Рекомендация:** Для получения наилучших результатов следует использовать модель линейной регрессии с деревом решений. Однако при необходимости интерпретируемости можно рассмотреть линейную регрессию по всему набору данных. Модель с ограниченным числом признаков не рекомендуется к использованию, так как она имеет низкую точность.