**Аналитический отчет по результатам построения регрессионных моделей с использованием KNIME и Python**

1. Описание набора данных

Выбран датасет с характеристиками красного вина и экспертными оценками качества.

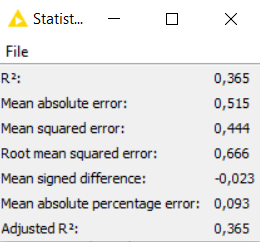
* fixed acidity (фиксированная кислотность) – сумма органических кислот (винная, яблочная и др.), придающих вину свежесть и структуру.
* volatile acidity (летучая кислотность) – концентрация уксусной кислоты. Высокий уровень может придавать вину неприятный уксусный запах.
* citric acid (лимонная кислота) – естественный консервант, который улучшает вкус и свежесть вина.
* residual sugar (остаточный сахар) – количество сахара, оставшегося после ферментации. Влияет на сладость вина.
* chlorides (хлориды) – количество солей, в основном хлорида натрия. Влияет на вкус и восприятие вина.
* free sulfur dioxide (свободный диоксид серы) – SO₂ в свободной форме, предотвращает окисление и рост бактерий.
* total sulfur dioxide (общий диоксид серы) – сумма свободного и связанного SO₂. Высокие значения могут негативно сказаться на аромате вина.
* density (плотность) – масса вина на единицу объема. Связана с содержанием сахара и алкоголя.
* pH (уровень pH) – показатель кислотности вина. Чем ниже pH, тем выше кислотность. Обычно варьируется от 2.9 до 4.0.
* sulphates (сульфаты) – серосодержащие соединения, помогающие предотвратить рост нежелательных бактерий.
* alcohol (алкоголь) – процентное содержание этанола в вине. Влияет на вкус, тело и восприятие вина.
* quality (качество) – целевая переменная, экспертная оценка вина (по шкале от 0 до 10).

Ссылка для скачивания датасета:

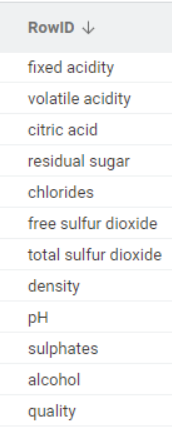
<https://archive.ics.uci.edu/dataset/186/wine+quality>

1. Изображение выглядит как текст, диаграмма, снимок экрана

   Контент, сгенерированный ИИ, может содержать ошибки.Построение регрессии в KNIME

****Изображение выглядит как текст, снимок экрана, График, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки. **Корреляционный анализ** показал, что связь между химическими характеристиками вина и его качеством является слабой. Коэффициенты корреляции между показателем quality и остальными признаками имеют сравнительно низкие значения, находятся в диапазоне от -0.4 до 0.5. Это свидетельствует о том, что отдельные физико-химические свойства вина объясняют лишь небольшую часть вариации в оценках качества.



R^2 = 0.365 – это вполне нормальный результат для исходного датасета, особенно учитывая, что использовался простой алгоритм линейной регрессии (при этом использование, например, Simple Decision Tree дало R^2 равный 0.99)

Итак, 36.5% дисперсии целевой переменной (качества вина) объясняется признаками модели. Остальные 63.6% зависят от факторов, которые модель не учитывает (например, субъективное восприятие дегустаторов, дополнительные характеристики вина). MAE = 0.53. В среднем ошибка предсказания ±0.53 балла. RMSE = 0.677. Ошибки крупнее, чем MAE, что говорит о наличии выбросов.

1. Построение регрессии в Python

Изображение выглядит как текст, снимок экрана, число

Контент, сгенерированный ИИ, может содержать ошибки.

На основе представленной корреляционной матрицы можно отметить наличие мультиколлинеарности между некоторыми факторами. Например, наблюдаются сильные корреляции между fixed acidity и citric acid (0.67), density и fixed acidity (0.67), а также total sulfur dioxide и free sulfur dioxide (0.67). Такие высокие значения корреляции могут указывать на потенциальную проблему мультиколлинеарности, что может привести к нестабильности регрессионной модели и затруднению интерпретации коэффициентов.

Тем не менее, каждый из представленных факторов имеет важную смысловую нагрузку. Кислотность (fixed acidity, citric acid, volatile acidity) влияет на вкус и ощущение вина, при этом разные виды кислот по-разному воспринимаются человеком. Содержание сахара (residual sugar), хлоридов (chlorides) и сернистых соединений (sulfur dioxide) влияет на стабильность и консервацию напитка. Плотность (density) тесно связана с содержанием алкоголя и остаточного сахара, а pH играет ключевую роль в химическом балансе. Сульфаты (sulphates) и алкоголь (alcohol) непосредственно влияют на восприятие качества напитка. Таким образом, даже при наличии корреляций, исключение каких-либо признаков может привести к потере важной информации о характеристиках вина.

Изображение выглядит как текст, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.Модель объясняет 40.3% дисперсии данных, что свидетельствует о среднем уровне качества предсказаний.

**ВЫВОД**

Результаты регрессии, полученные в Python, показывают, что модель объясняет 40,3% дисперсии данных. Среднеквадратичная ошибка (MSE) на тесте составляет 0.390, а на трейне – 0.424, что может свидетельствовать о небольшой недообученности модели, но в целом говорит о ее стабильности.

В KNIME коэффициент детерминации оказался ниже (R² = 0.365), что указывает на худшую способность модели объяснять вариативность данных. Кроме того, MSE составило 0.444, что выше, чем в Python, а это значит, что разброс ошибок больше. Средняя абсолютная ошибка (MAE) и корень из среднеквадратичной ошибки (RMSE) также указывают на более высокий уровень ошибок, чем в Python.

В целом модель, построенная в Python, демонстрирует лучшие результаты, так как имеет более высокое R² и меньшую ошибку предсказаний. Это говорит о большей точности и стабильности модели по сравнению с той, что была обучена в KNIME.