В блокноте EDA.ipynb провёл разведочный анализ данных на основе результатов анализа пропусков видим:

1. у нас 214 признаков (столбцов). У большинства признаков нет пропусков (значение Missing Count = 0). Лишь несколько столбцов имеют по 3 пропуска, что составляет 0.2997% от общего числа строк.

2.Процент пропущенных значений — менее 0.3%.

3.Так как пропусков мало (~0.3%), попробуем оставить их как есть.

4. Наблюдаем выбросы во многих признаках особенно таких как IC50, mM, SI, fr_thiazole, fr_thiophene.

Это может влиять на обучение моделей

5. Разница между IQR и Z-score объясняется типом распределения

Если данные не нормальны , метод Z-score будет менее строгим.

6. Некоторые фрагменты (например, fr_...) часто содержат выбросы

Это может быть связано с редкими структурами в молекулах.

Также провёл анализ мультиколлинеарности с помощью VIF (Variance Inflation Factor)-это метрика, которая показывает, насколько сильно дисперсия коэффициента регрессии увеличивается из-за мультиколлинеарности.Интерпретация:

VIF = 1 → нет корреляции с другими признаками.

VIF < 5 → умеренная мультиколлинеарность (приемлемо).

VIF > 5–10 → высокая мультиколлинеарность → стоит удалить или объединить такие признаки.

MaxAbsEStateIndex

4.38e+07

Сильно коррелирует с другими признаками

fr_thiophene

28.1

Высокая мультиколлинеарность

fr_unbrch_alkane

15.4

Высокая мультиколлинеарность

fr_thiazole

16.4

Высокая мультиколлинеарность

Unnamed: 0

7.18

Умеренно высокая

fr_urea

8.93

Высокая

IC50, mM

CC50...

~2.5 – 2.9

Низкая мультиколлинеарность

1. Проблема высокой мультиколлинеарности есть

Некоторые признаки имеют очень высокие значения VIF , особенно:

MaxAbsEStateIndex: 43 млн — это явно проблема.

fr_thiophene, fr_thiazole, fr_unbrch_alkane и другие фрагменты (fr_...) тоже имеют VIF > 10.

Это значит, что: эти признаки сильной скоррелированы с другими , и их одновременное использование в моделях может привести к: неустойчивости коэффициентов, переобучению, снижению интерпретируемости модели.

2. Ключевые параметры (IC50, CC50, SI) не имеют сильной мультиколлинеарности

Их VIF ≈ 1–3 → это хорошо.

Можно использовать их как целевые переменные или предикторы без опасений.

Общие выводы по EDA:

1. Признаки IC50, CC50 и SI имеют явные выбросы.

2. SI коррелирует с отношением CC50 / IC50.

3. Высокая корреляция между некоторыми фичами может влиять на обучение моделей.

4. Необходимо масштабировать данные перед обучением моделей.

В блокноте feature.ipynb провел очистку данных от выбросов, пропущенные значения заменили медианой. Полученный датасет сохранил для последующего использования при обучении моделей.

Регрессия для IC50

Результаты моделей:

	RMSE	R^2
Random Forest	166.820282	0.351584
Linear Regression	168.358745	0.339570
XGBoost	180.844866	0.237977

Наиболее важные признаки по модели Random Forest:

	Feature	Importance
62	VSA_EState8	0.104797
59	VSA_EState4	0.087371
70	MolLogP	0.045583
5	MolWt	0.039597
3	qed	0.037357
13	BCUT2D_MRLOW	0.032526
51	EState_VSA5	0.029515
50	EState_VSA4	0.028897
9	FpDensityMorgan1	0.028475
8	MinPartialCharge	0.024589

Вывод. Лучшие результаты показала модель Random Forest, но значение $R^2 = 0.35$ всё ещё невысокое (в идеале должно быть близко к 1). Это значит, что модель пока объясняет только около трети вариации данных.Возможно, есть проблемы с данными:

Шумные данные, недостаточное количество или качество признаков, выбросы в целевой переменной (IC50 или CC50) (хотя вроде удалял).

Регрессия CC50

Результаты моделей:

	RMSE	R^2
Random Forest	378.084352	0.562399
XGBoost	388.422773	0.538141
Linear Regression	430.977379	0.431397

Наиболее важные признаки по модели Random Forest:

	Feature	Importance
5	MolWt	0.186080
11	BCUT2D_MWLOW	0.077831
9	FpDensityMorgan1	0.067742
13	BCUT2D_MRLOW	0.039278
59	VSA_EState4	0.030475
61	VSA_EState7	0.027211
27	PEOE_VSA7	0.025952
26	PEOE_VSA6	0.025931

10 BCUT2D_MWHI 0.025342

7 MaxPartialCharge 0.023364

Вывод: лучшие результаты опять показала модель Random Forest. Качество модели: $R^2 = 0.56$. Наиболее важные признаки:

MolWt (молекулярная масса)

BCUT2D (топологические дескрипторы)

PEOE_VSA (распределение зарядов)

VSA_EState (полярные свойства)

Можно предположить что цитотоксичность (CC50) напрямую связана с физическими и химическими свойствами молекул, такими как масса, полярность, распределение зарядов и структурная сложность.

Регрессия SI

Результаты моделей:

	RMSE	R^2
Random Forest	12.115917	0.184025
Linear Regression	12.528218	0.127545
XGBoost	13.287816	0.018542

Наиболее важные признаки по модели Random Forest:

	Feature	Importance
62	VSA_EState8	0.065709
69	RingCount	0.062752
3	qed	0.043907
11	BCUT2D_MWLOW	0.041114
15	BalabanJ	0.033231
4	SPS	0.032858
12	BCUT2D_CHGHI	0.030618
35	SMR_VSA5	0.026937
13	BCUT2D_MRLOW	0.025908
0	MaxAbsEStateIndex	0.025761

Выводы: Random Forest снова показала лучший результат, но её качество всё ещё достаточно низкое.Требуется дальнейшая рабта над улучшением модели и анализа данных.

Наиболее важные из признаков VSA_EState8 и RingCount.

Классификация CC50 по медиане:

концентрация вещества, при которой погибает 50% клеток хозяина (цитотоксичность)

лучшая модель: XGBoost

метрика AUC-ROC: 0.9977227722772278

Модель успешно решает задачу классификации CC_{50} по медиане с AUC-ROC = 0.93. Это означает, что модель хорошо различает вещества с высокой и низкой цитотоксичностью

Классификация IC50 по медиане:

концентрация вещества, подавляющая на 50% целевую активность (например, репликацию вируса)

Лучшая модель: Gradient Boosting Метрика AUC-ROC: 1.0 Модель умеет точно определять, какие вещества обладают высокой активностью(значение IC_{50} выше медианы)

Классификация SI по медиане:SI – индекс селективности CC/IC Лучший результат показала модель Логистической регрессии метрика AUC-ROC: 0.9998019801980198. Возможно стоит еще поработать над предобработкой данных, возможно произошло переобучение

Классификация SI > 8:

Если SI > 8, это обычно означает, что вещество обладает хорошей селективностью и потенциальной перспективностью как кандидат в лекарственные препараты.

Лучшая модель: Logistic Regression

Метрика AUC-ROC: 0.9954780361757106

Тоже есть подозрение на переобучение и правильность разбиения выборки.