Описание данных

На начальном этапе создания лекарств нужно определить активные и безопасные молекулы, чтобы избежать затрат на дорогостоящие тесты. Методы химической информатики и машинного обучения помогают предсказать фармакологические свойства соединений, основываясь на их молекулярной структуре.

В этой работе мы разбираем прогнозирование трёх важных биологических показателей:
• **IC50** - концентрация, при которой ингибируется 50% активности определённого биологического процесса или мишени, показывает эффективность соединения как ингибитора.
• **CC50** - полумаксимальная токсическая концентрация, отражающая токсичность соединения для клеток.
• **SI** (индекс селективности) — рассчитывается как отношение CC50 к IC50 и отражает баланс между активностью и токсичностью: чем выше SI, тем безопаснее соединение.

Цель работы - разработать и оценить модели машинного обучения (регрессионные и классификационные), которые смогут предсказывать эти показатели на основе различных молекулярных дескрипторов. Для этого мы используем широкий набор признаков, включая:

Общие молекулярные дескрипторы:

Молекулярная масса, количество тяжёлых атомов, число валентных электронов, число радикальных электронов, доля CSP3, топологическая полярная поверхность, доступная поверхность, оценка «лекарственности», гидрофобность и молекулярная рефрактивность. Примечание: SPS можно убрать, так как он не подходит для этой работы.

Электронные дескрипторы:

Максимальные и минимальные частичные заряды и их абсолютные значения, а также распределение зарядов (PEOE_VSA) и топология (EState_VSA).

Топологические дескрипторы:

Различные индексы и другие параметры, характеризующие структуру молекулы.

BCUT-дескрипторы:

Параметры, основанные на массе, заряде, logP и рефрактивности молекул.

VSA-дескрипторы:

Распределение различных свойств по поверхности молекулы.

Morgan fingerprints:

Плотность битов при разных радиусах.

Фрагментные дескрипторы:

Наличие определённых химических групп или фрагментов, таких как фенолы, амины и другие.

Структурные количественные дескрипторы:

Число акцепторов и доноров водородных связей, вращающихся связей, гетероатомов и количество колец.

Источники:

Маджидов Т.И. и др. «Введение в хемоинформатику» (2013).

Итоги:

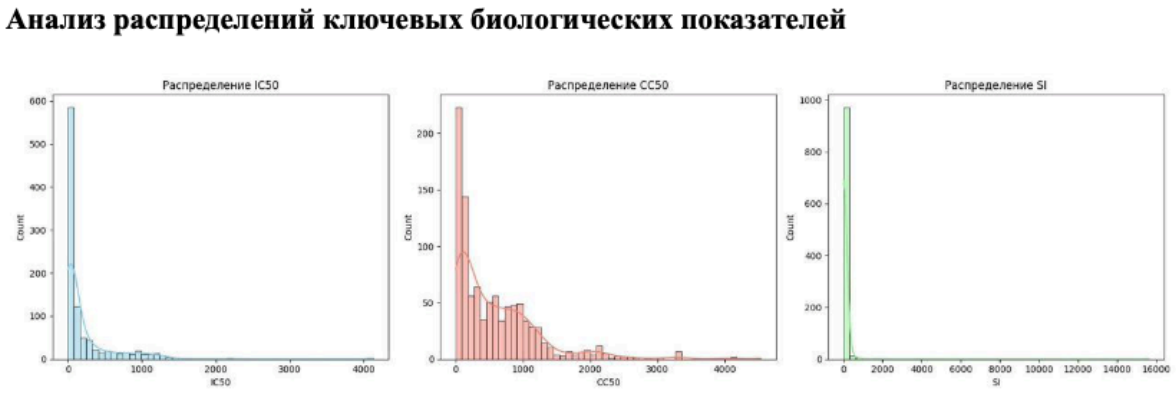
- SPS можно исключить, так как он не имеет отношения к прогнозированию SI.
- При прогнозировании IC50 или CC50 следует исключить SI и его производные, чтобы избежать утечки данных.

EDA

Для начала загрузили датасет, содержащий 1001 строку и 214 признаков. Все данные являются числовыми, что упрощает последующее моделирование.

В ходе первичного анализа было выявлено, что пропуски встречаются редко: 12 признаков содержат всего по 3 пропуска каждый. Такой уровень пропусков можно корректно обработать без значительных потерь данных - либо удалить строки с пропущенными значениями, либо заполнить их средним или медианой соответствующего признака.

После удаления строк с пропусками размер итогового датасета составляет (998, 214).



Анализ распределений ключевых биологических показателей

IC50

Значения IC50 явно скошены вправо: большинство данных находятся до 1000, но попадаются и высокие значения, создавая длинный хвост. Есть выбросы, и распределение довольно ненормальное, что может сказаться на моделях. Это нужно учитывать при предобработке, например, с помощью логарифмического преобразования.

CC50

Ситуация похожа на IC50: основная часть значений находится в нижней части графика, но есть выбросы в диапазоне 2000–4000 и длинный асимметричный хвост. Это распределение тоже отклоняется от нормального.

SI (Индекс селективности)

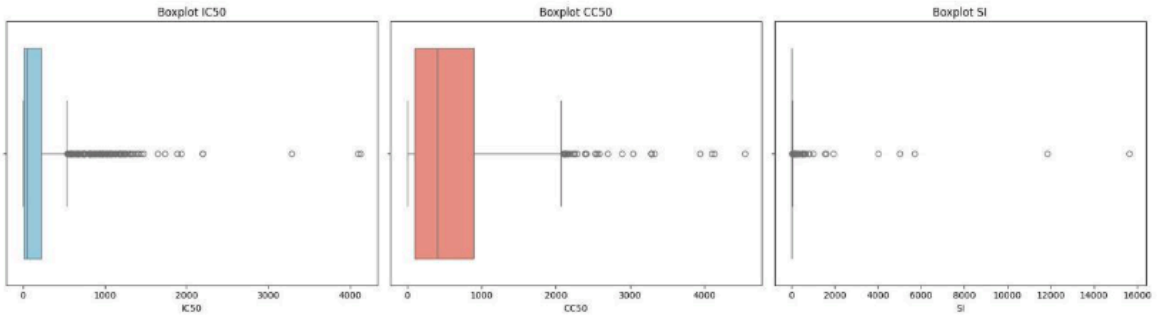
Распределение SI показывает самое большое отклонение. Почти все значения ниже 100, но есть редкие случаи выше 15000, создавая очень длинный хвост. Это указывает на наличие

экстремальных выбросов и сильное смещение. Такую ситуацию тоже стоит исправить с помощью преобразования, например, логарифмирования, чтобы стабилизировать распределение.

Выводы

Для всех трех показателей заметно сильное смещение вправо и наличие выбросов. Это может ухудшить качество моделей, если не применить предварительные трансформации данных. Логарифмирование значений поможет сделать распределения более нормальными, уменьшить влияние выбросов и улучшить работу алгоритмов машинного обучения.

Подтверждение выбросов (визуализация)

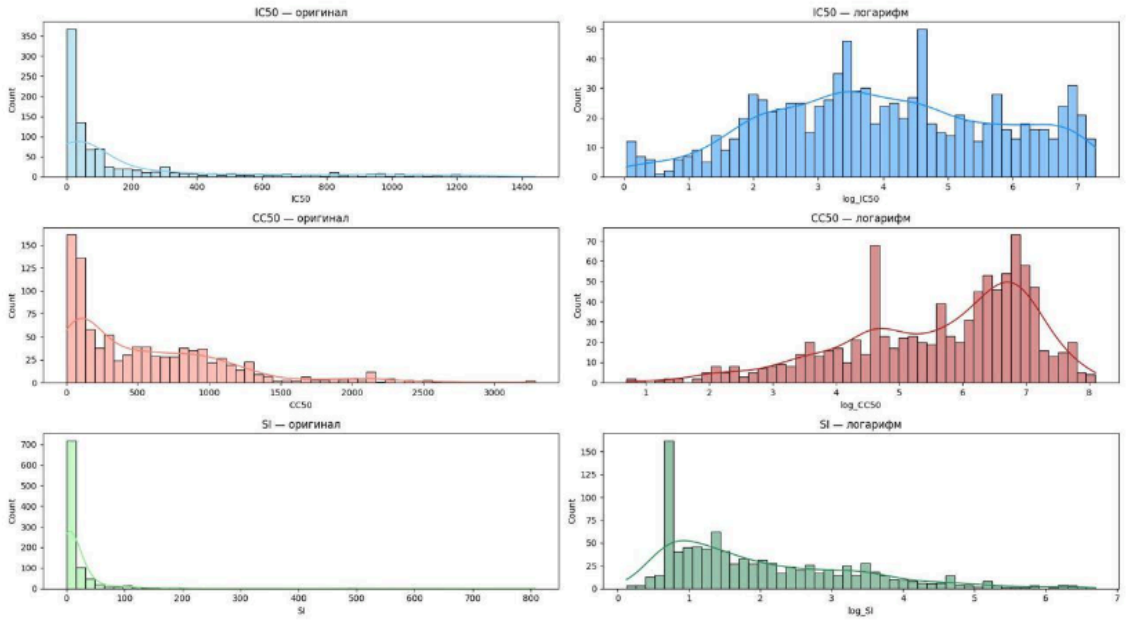


На графиках распределений ключевых показателей (IC50, CC50 и SI) визуально подтверждается наличие значительных выбросов:

- IC50 - большое количество выбросов наблюдается после 1000, что указывает на сильную асимметрию распределения.
- CC50 - выбросов меньше, чем у IC50, но они также заметны, особенно после 2000.
- SI - экстремальные значения уходят за 10 000, что подтверждает наличие очень длинного

хвоста и высокую степень скошенности распределения.
Такие выбросы могут существенно влиять на обучение моделей, поэтому важно рассмотреть возможность их логарифмирования или иной трансформации для улучшения качества прогнозов.

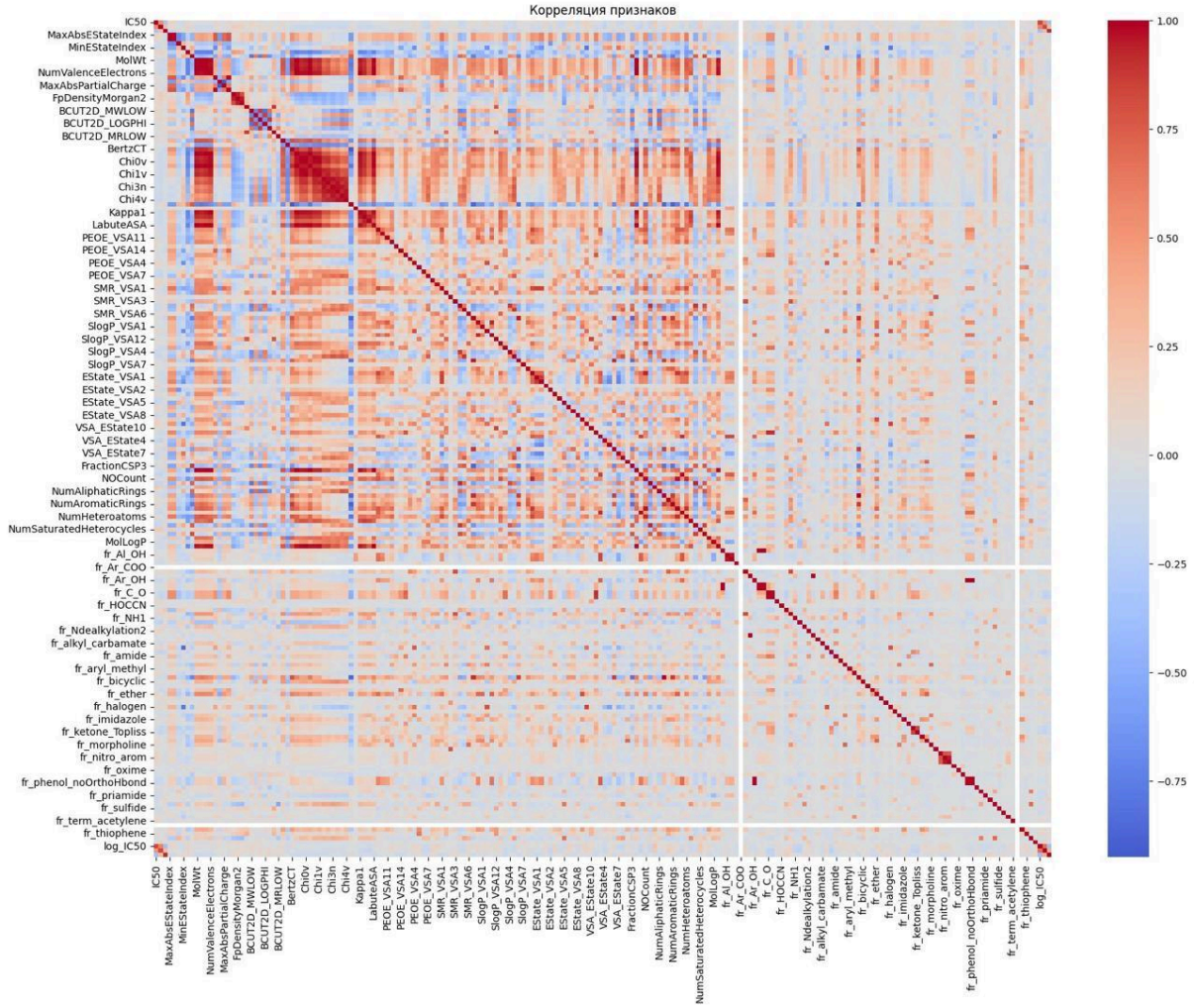
Распределения после логарифмирования



После применения логарифмического преобразования распределения IC50 и CC50 стали значительно более симметричными: хвосты укоротились, пики сгладились, и распределения приблизились к нормальному. Это особенно заметно для IC50 и CC50. Для SI распределение также стало более «ровным», хотя небольшая асимметрия сохраняется, но в целом ситуация улучшилась по сравнению с исходными данными.

Итог

Данные были очищены от пропусков и выбросов. Удалены нерелевантные признаки. Проведён полный анализ распределений и взаимосвязей между признаками. Логарифмирование IC50, CC50 и SI существенно улучшило распределения, что повышает устойчивость моделей машинного обучения. Признаки готовы к обучению моделей, и следующим этапом является построение решений для задач регрессии и классификации.



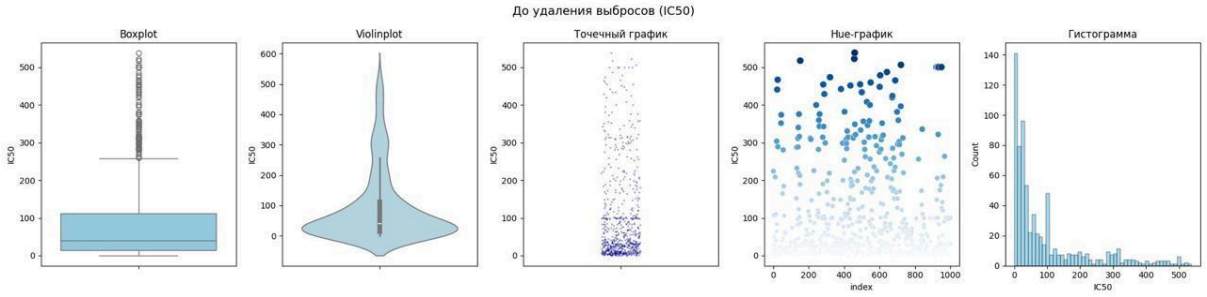
Корреляционная матрица признаков (включая исходный и логарифмированный IC_{50}):
– диагональ показывает автокорреляцию (значение = 1),
– красные оттенки обозначают сильную положительную корреляцию,
– синие – сильную отрицательную.

Эта визуализация позволяет выявить группы тесно связанных дескрипторов и области потенциальной мультиколлинеарности.

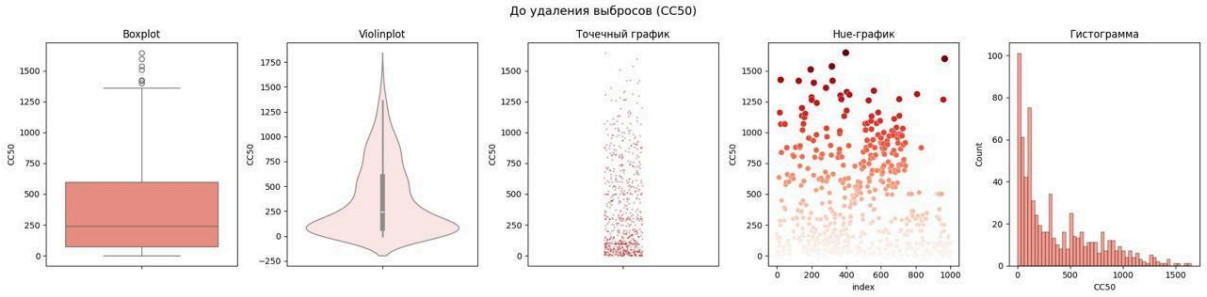
Также в ходе анализа удалены 18 признаков, имеющих нулевую дисперсию, то есть принимающих одно и то же значение во всех наблюдениях. Такие признаки не несут информативности для моделей, так как не объясняют вариативность целевой переменной.

Среди них присутствуют химические дескрипторы, связанные с редкими или отсутствующими функциональными группами (например, fr_barbitur, fr_azide, fr_diazo), которые встречаются крайне редко в молекулах из текущего датасета. Их исключение оправдано, так как они не могут повлиять на обучение моделей.

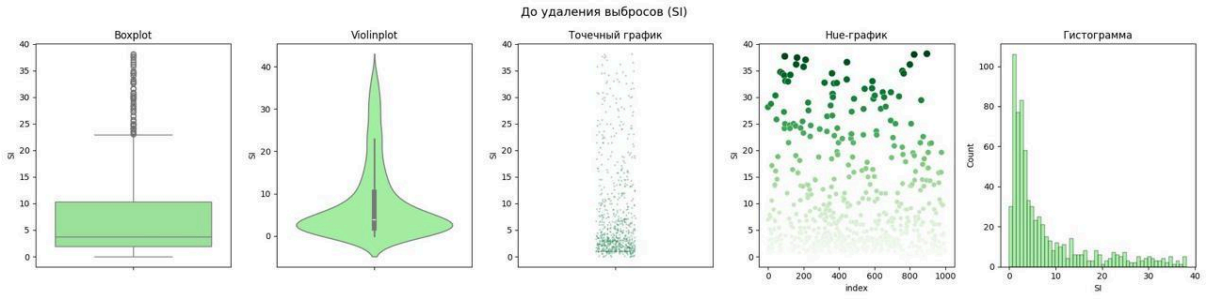
IC50



CS50



SI



На основании этого EDA были сформированы и сохранены очищенные датасеты **dataset_for_IC50_clean.csv**, **dataset_for_CC50_clean.csv** и **dataset_for_SI_clean.csv** для последующего построения моделей.

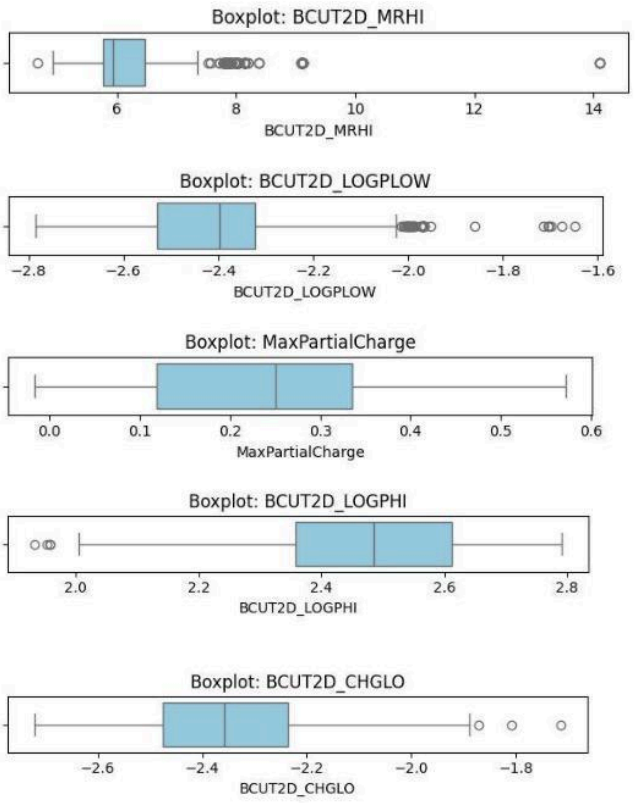
Регрессия IC50

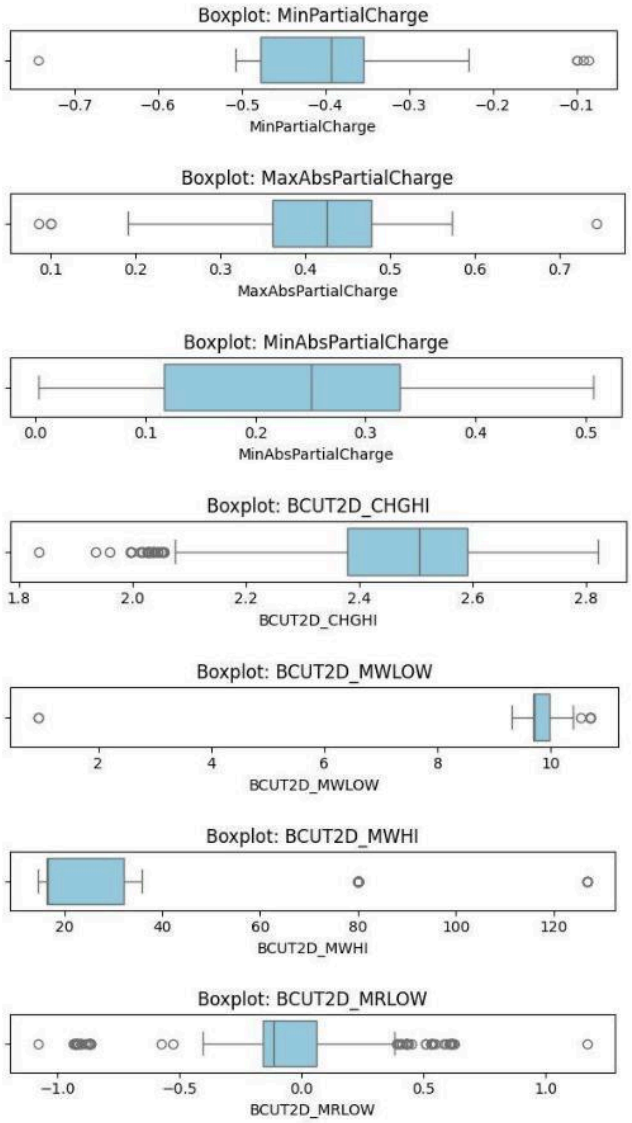
IC50 (Концентрация ингибирования 50%) - такая концентрация вещества, при которой оно снижает биологическую активность, например, активность вируса, на 50%. Этот показатель часто используется, чтобы оценить, как хорошо работают новые лекарства. Чем ниже значение IC50, тем лучше вещество, так как для достижения нужного результата нужна меньшая его концентрация. Обычно это измеряется в нано- или микромолях.

В нашем проекте мы собираемся разрабатывать регрессионные модели для предсказания значений IC50, основываясь на разных числовых характеристиках химических соединений. Это поможет нам оценивать эффективность новых молекул, просто глядя на их структуру, что может значительно ускорить и снизить затраты на поиск перспективных вариантов.

Анализ пропусков

Перед построением моделей необходимо провести предварительный анализ пропусков в данных. В нашем датасете встречаются признаки с пропущенными значениями - всего 12 признаков содержат по 3 пропуска каждый. Такой небольшой процент пропусков можно корректно обработать без значительных потерь информации.





На этапе визуализации распределений этих признаков мы обратили внимание, что многие из них содержат значительное количество выбросов, особенно ярко это выражено у признаков BCUT2D_LOGPLOW, BCUT2D_MWHI и BCUT2D_MRLOW. Значения в этих признаках часто уходят далеко за пределы основного диапазона, формируя длинные хвосты распределений.

Обработка пропусков

С учетом выявленных выбросов было решено использовать медиану для заполнения пропусков. Медиана устойчива к экстремальным значениям и поэтому позволяет избежать искажения данных, которое могло бы возникнуть при использовании среднего значения. Такой подход обеспечивает более надежную основу для последующего моделирования.

Моделирование IC50

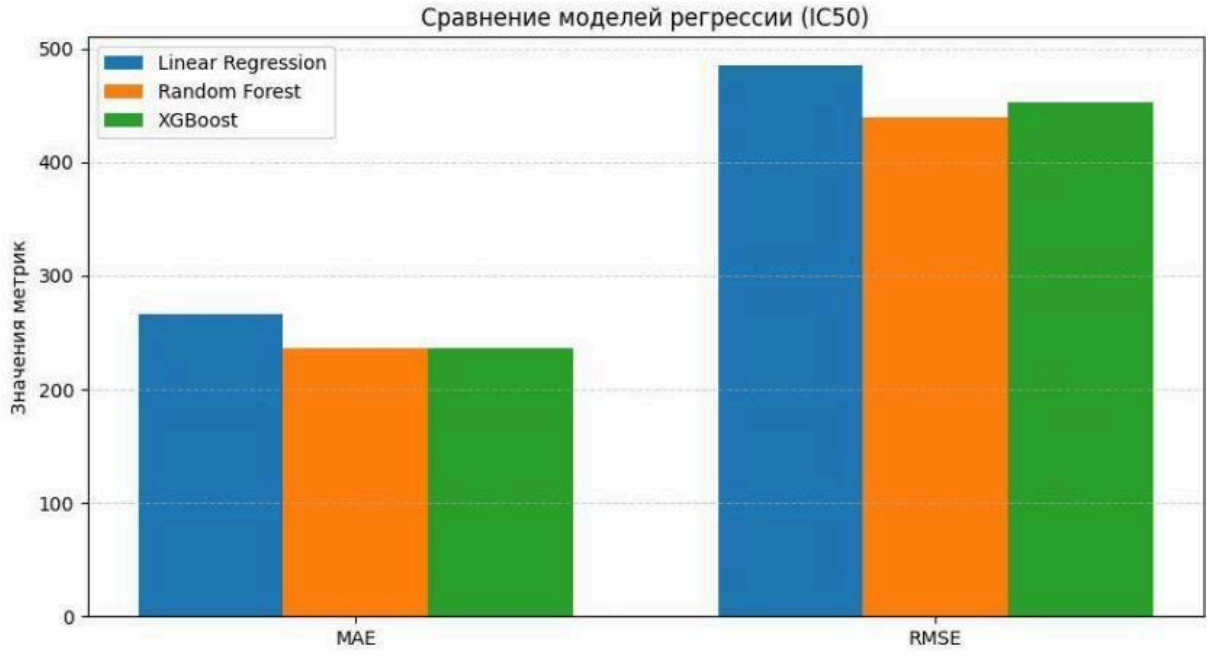
Проверили несколько моделей для предсказания IC50:

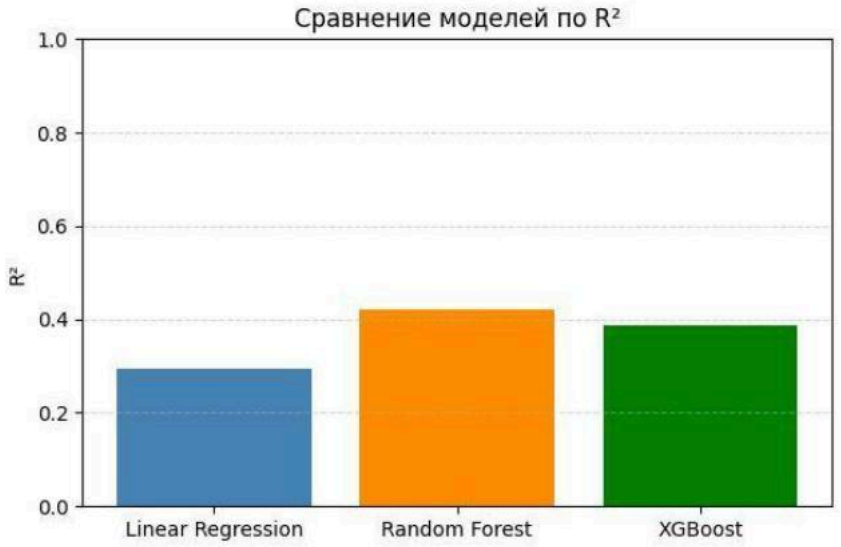
Выбор моделей

Для задачи регрессии IC50 мы выбрали три разных подхода: линейную регрессию как базовую модель, Random Forest и XGBoost как более продвинутые алгоритмы. Линейная регрессия даёт простую и интерпретируемую модель, но плохо справляется с нелинейными взаимосвязями. Random Forest позволяет учитывать сложные взаимодействия между признаками за счёт ансамбля деревьев и часто показывает высокую стабильность. XGBoost дополнительно использует градиентный бустинг, что делает его особенно эффективным для работы с разреженными и несбалансированными данными. Такой набор моделей даёт возможность сравнить простую базовую модель с более сложными и выбрать наиболее подходящую для предсказания IC50.

Random Forest, XGBoost и LightGBM. Линейная регрессия была использована в качестве базовой модели.

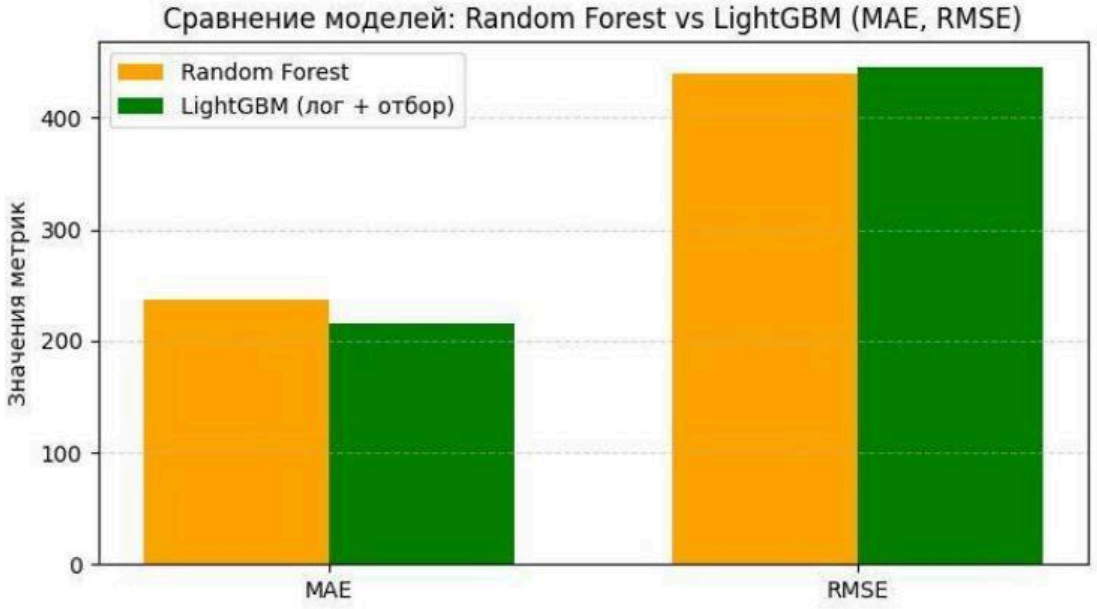
Результаты

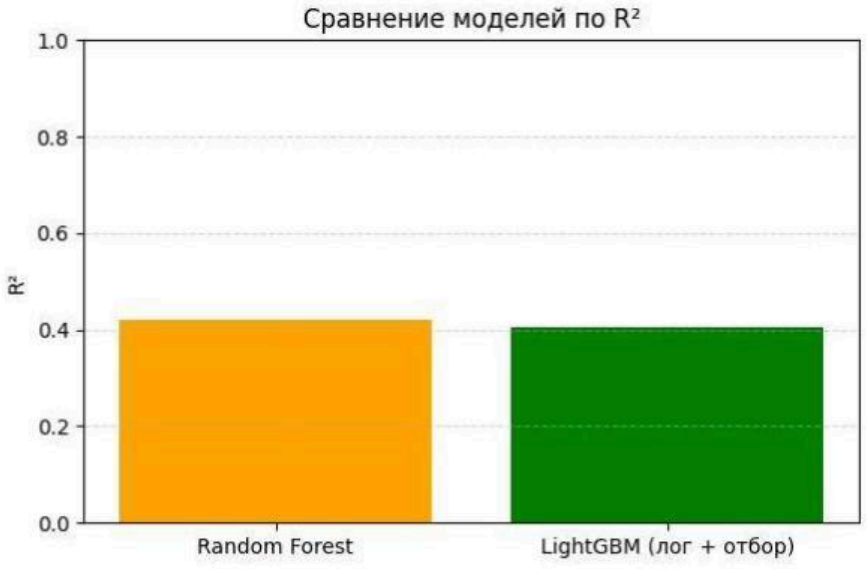




Результаты: Random Forest, с настройками max_depth=10, n_estimators=200, min_samples_leaf=2, показал MAE ≈ 236.7, RMSE ≈ 439.8 и R2 ≈ 0.42. Это говорит о хорошей стабильности модели и её способности обобщать. XGBoost с параметрами max_depth=7, subsample=0.8, n_estimators=200, learning_rate=0.01, показал почти такие же результаты: MAE ≈ 236.5, RMSE ≈ 452.3 и R2 ≈ 0.39. Линейная регрессия выдала худшие метрики (MAE ≈ 265.7, R2 ≈ 0.29), что и ожидалось, так как она не может учитывать сложные зависимости. Чтобы улучшить предсказания, мы применили логарифмирование IC50, что снизило влияние выбросов и сделало распределение более нормальным.

Также был проведён отбор признаков с использованием Random Forest, что помогло сосредоточиться на самых важных характеристиках. LightGBM на отобранных данных показал немного лучшие результаты по MAE, но по RMSE и R2 Random Forest был предпочтительнее, что говорит о его стабильности.





Что касается химических аспектов, ключевые признаки, выбранные моделью, включают топологические (например, Chi2n, Chi2v) и электронные дескрипторы (PEOE_VSA, VSA_EState), а также молекулярные фрагменты (aromat-NH, NHpyrrole). Это подтверждает, что активность соединения (IC50) зависит как от структуры молекулы, так и от её электронной плотности.

Эти связи имеет смысл с химической точки зрения: структура и электронные свойства влияют на взаимодействие с биомишенями, а значит, на ингибирующую активность. В заключение, логарифмирование, отбор признаков и настройка гиперпараметров значительно улучшили результаты. Для дальнейшего прогресса можно рассмотреть более сложные модели, углубленную

настройку параметров и тест новых подходов, таких как CatBoost. Также стоит подумать об обучении моделей на отдельных группах соединений, чтобы повысить точность.

В итоге, машинное обучение показало хороший потенциал в предсказании IC50, что поможет ускорить поиск кандидатов для биологических испытаний и снизить затраты на начальных этапах разработки медикаментов.

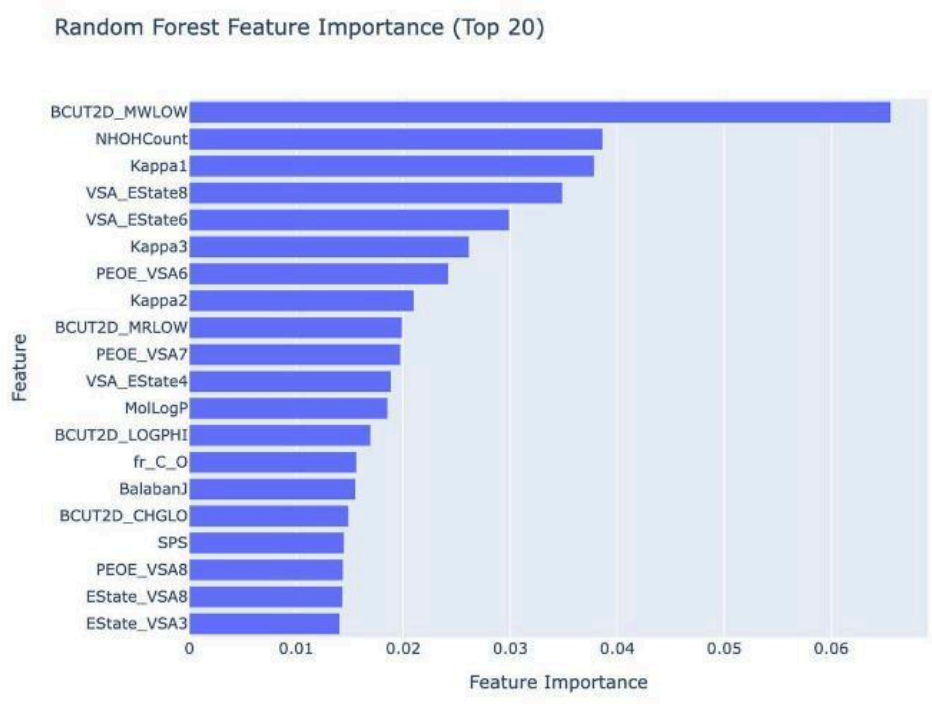
Регрессионное моделирование для предсказания CC50

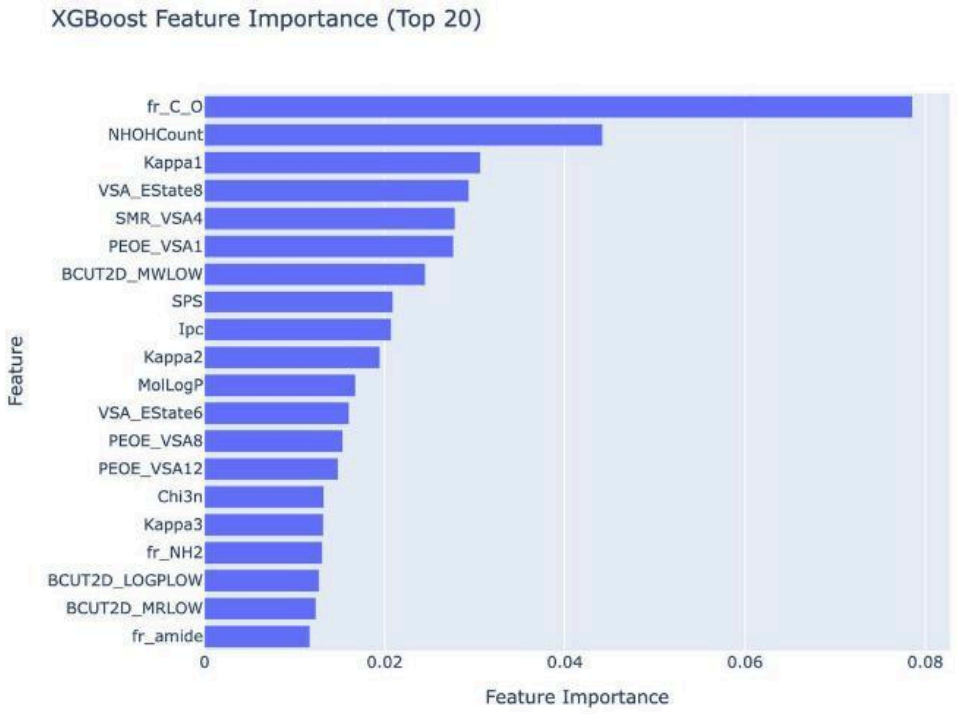
Мы решаем задачу регрессии для прогнозирования CC50 - концентрации вещества, при которой погибает или повреждается 50% клеток. Этот показатель помогает оценить токсичность химических соединений ещё на этапе компьютерных экспериментов, сокращая затраты на лабораторные тесты.

Данные были загружены, проверены на пропуски и обработаны с использованием медианного заполнения (SimpleImputer). Для корректной работы моделей проведено масштабирование признаков (StandardScaler).

Для построения моделей использовались Linear Regression, Ridge Regression, Random Forest и XGBoost. XGBoost дополнительно был оптимизирован с помощью GridSearchCV, что позволило улучшить метрики (Best MSE ≈ 209569 и Best R2 ≈ 0.60). В итоге:

● Random Forest показал MAE ≈ 297, RMSE ≈ 517 и R2 ≈ 0.485. ● XGBoost—MAE≈309,RMSE≈523иR2≈0.472.

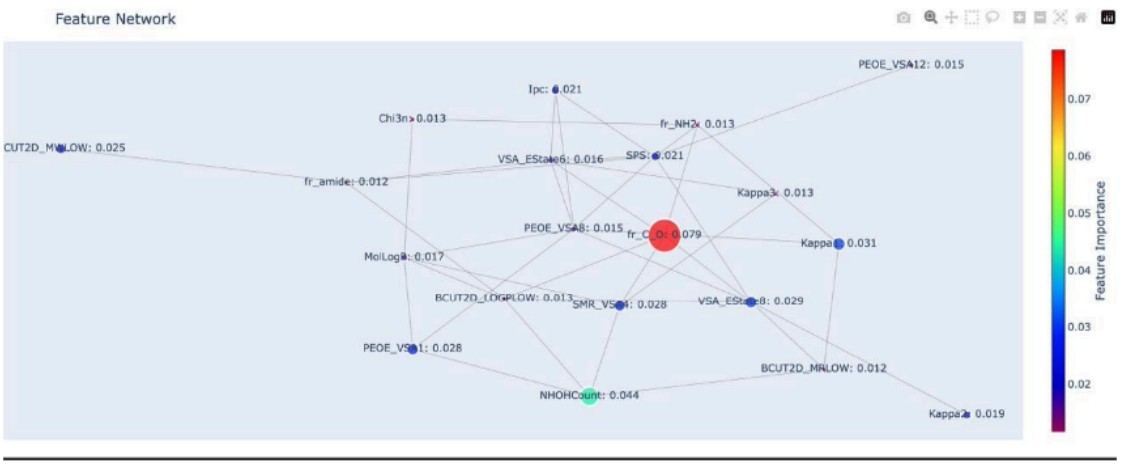




Обе модели демонстрируют схожую производительность, но Random Forest чуть лучше по всем метрикам. Невысокие значения R2 (0.47–0.49) указывают на наличие шумов в данных или недостаток информативных признаков, однако логарифмирование целевой переменной улучшило качество прогноза.

Анализ важности признаков показал, что на токсичность соединений существенно влияют молекулярная масса (BCUT2D_MWLOW), количество групп NHOH, форма молекулы (Kappa1) и электронные свойства (VSA_EState8). В XGBoost особенно значимы карбонильные группы (fr_C_O), указывая на потенциальную реактивность. Эти результаты помогают химикам понять, какие структурные элементы молекул могут быть связаны с токсичностью.

Мы выбрали модели машинного обучения, начиная с простой линейной регрессии, чтобы задать базовый уровень качества, и добавили Random Forest и XGBoost, так как они хорошо подходят для сложных взаимосвязей между признаками. Это позволило выявить важные для химиков структурные дескрипторы, такие как молекулярная масса, электронные свойства (VSA_EState), количество групп NHOH и наличие карбонильных фрагментов, которые могут влиять на токсичность соединений.



Для наглядного представления взаимосвязей между топ-20 признаками (по важности) мы использовали библиотеку NetworkX для построения графа и Plotly для визуализации. В этой HTML-визуализации каждая вершина представляет признак, а её размер и цвет соответствуют значимости (importance) признака. Рёбра между вершинами отражают потенциальные (случайные) связи между признаками. Полученный интерактивный граф помогает оценить, какие дескрипторы наиболее важны и как они могут быть взаимосвязаны между собой.

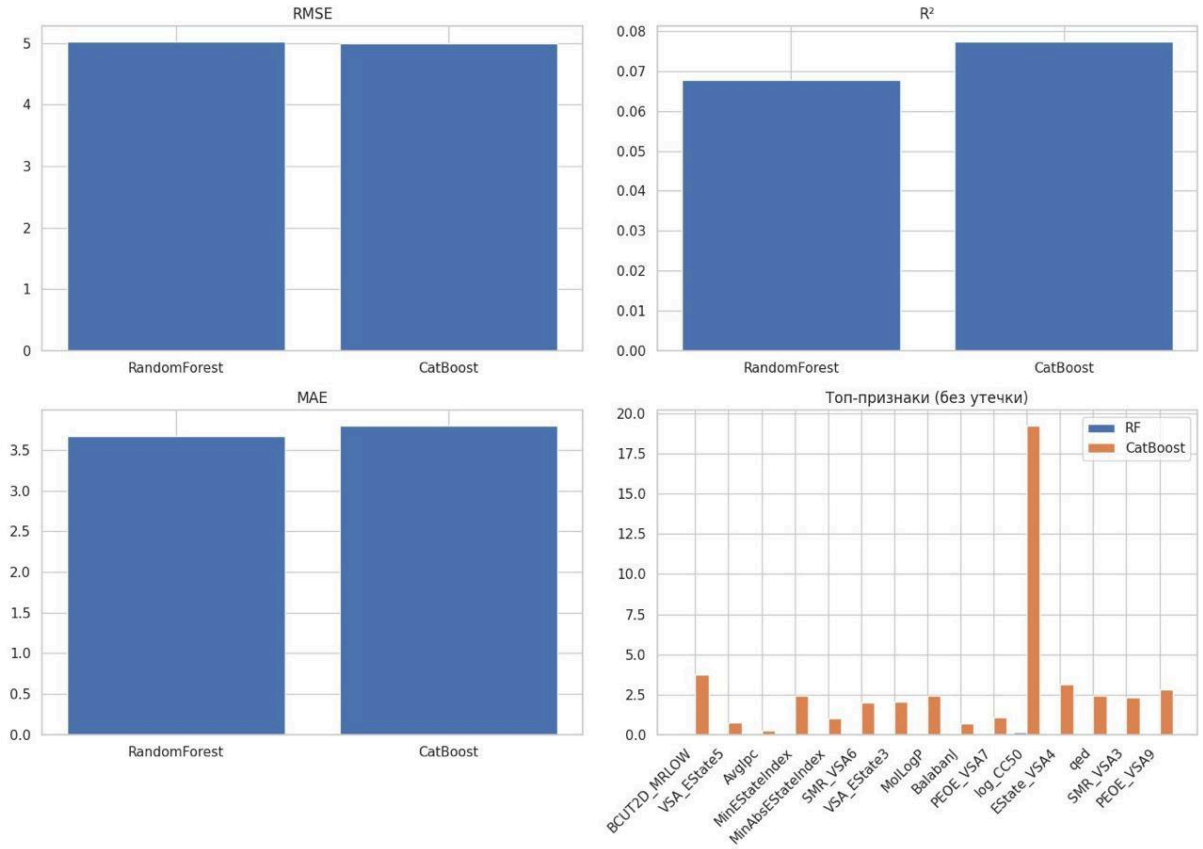
Регрессия SI обновленные

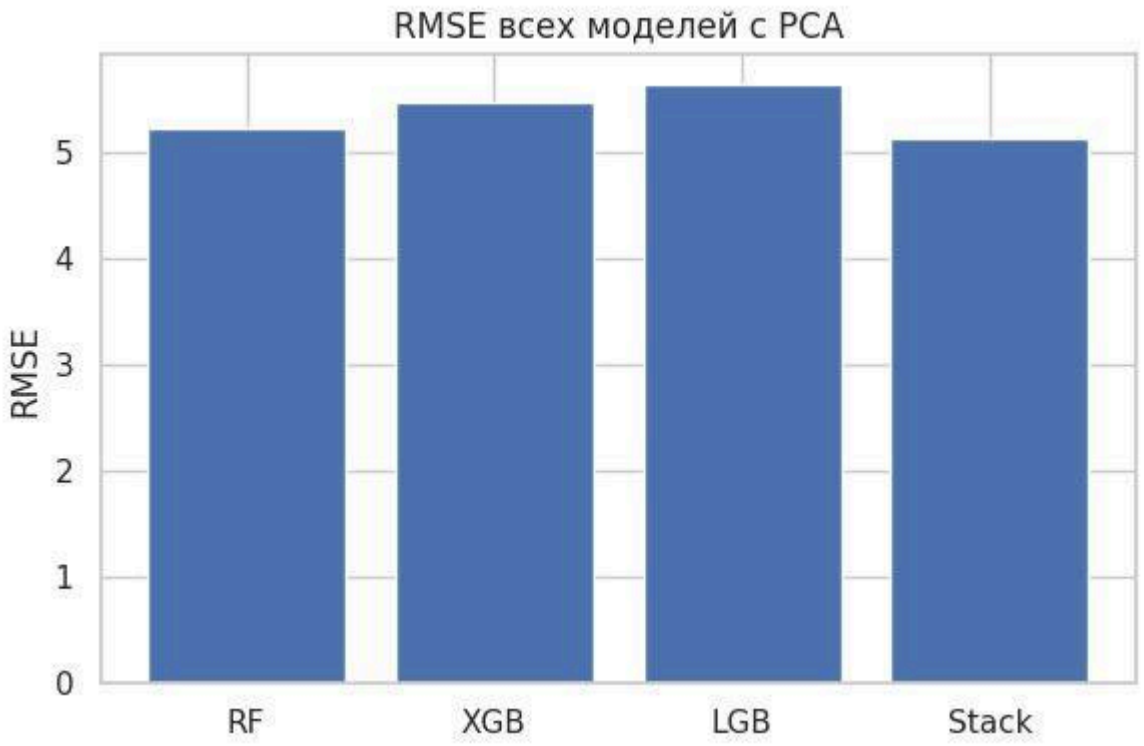
Самое главное - не допущен data leakage, теперь мы работает с очищенным дата сетом для SI.

Постановка задачи

Мы занимаемся предсказанием показателя SI (Selectivity Index), который показывает, насколько соединение может подавлять вирус с минимальной токсичностью для клеток. Чем выше данный показатель, тем более безопасно и эффективно соединение как потенциальное лекарство.

Подход и результаты.





Загрузила очищенный датасет без признаков «утечки» (колонки с IC50/CS50), обучили RandomForest и CatBoost — получили RMSE≈5, R^2≈0.07–0.08, MAE≈3.7. Пробовали отбор лучших признаков, полиномиальные фичи, XGBoost, LightGBM и stacking - без существенного улучшения (лучший стэкинг RMSE≈4.97, R^2≈0.09). Идея с PCA (15 компонент) ухудшила результаты. Итеративно удалили выбросы по SI, после чего RandomForest дал RMSE≈2.20, R^2≈0.10, MAE≈1.68 - существенное падение ошибки, но всё ещё объяснение дисперсии ≲10 %. Главное - ни одна модель не «взглянула» на утечку данных, поэтому все оценки честны; для реального роста R^2 нужны новые информативные признаки или трансформация целевой переменной.

Для химиков

Важно помнить, что SI зависит не только от одного признака, а от многих факторов - например, топологии молекулы, количества аминогрупп и числа колец. Кластеризация показала, что молекулы с различной массой и количеством колец имеют разные уровни SI. Это может помочь в разработке новых соединений. Например, молекулы с высокой массой и большим количеством колец (кластер 2) показывают низкий SI, что означает, что нужно упростить их структуру.

Классификация для IC50

Постановка задачи Мы делаем бинарную классификацию для показателя IC50, чтобы понять, превышает ли он медианное значение в выборке. Это помогает выбрать более эффективные соединения и ускорить разработку лекарств.

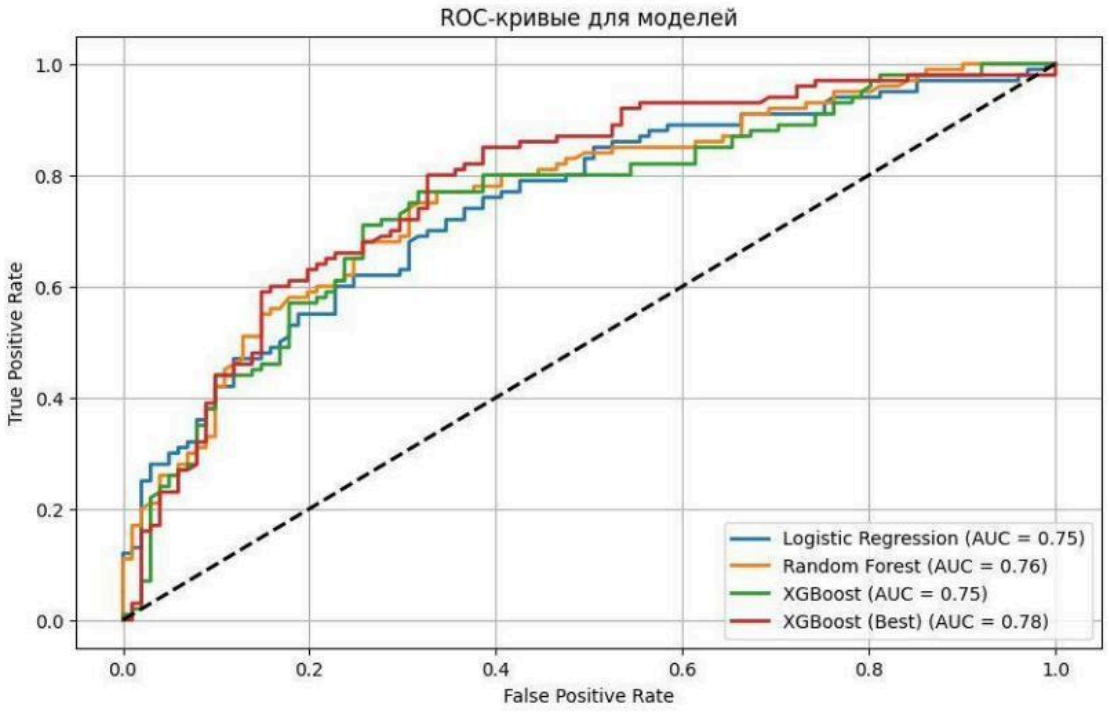
Выбор моделей Мы использовали Logistic Regression, Random Forest и XGBoost (с настройкой гиперпараметров). Также тестировали LightGBM и ансамблевую модель (Stacking Classifier), чтобы совместить сильные стороны этих моделей.

Результаты

Logistic Regression: Accuracy = 0.6866, F1-score = 0.7070, ROC-AUC = 0.6869 Random Forest: Accuracy = 0.6965, F1-score = 0.7189, ROC-AUC = 0.6969

XGBoost (лучший результат): Accuracy = 0.7264, F1-score = 0.7465, ROC-AUC = 0.7268 LightGBM: Accuracy = 0.6965, F1-score = 0.7136, ROC-AUC = 0.7570

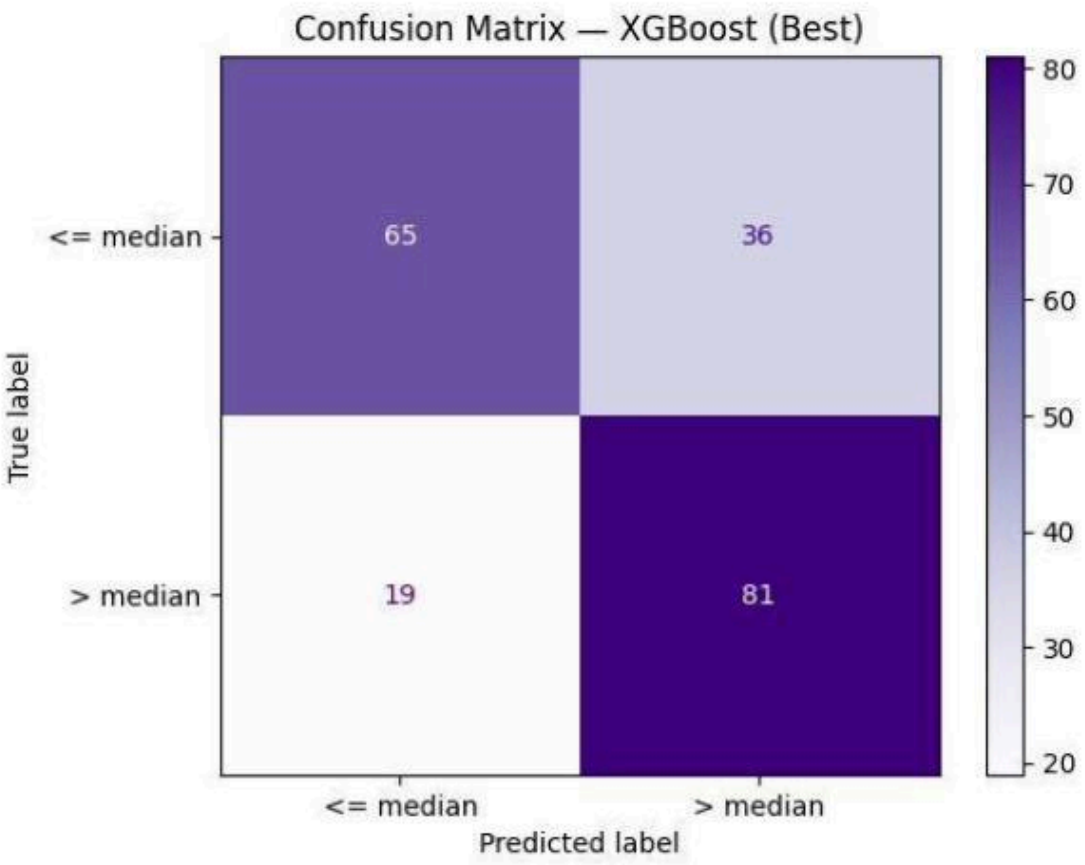
Stacking Classifier: Accuracy = 0.6965, F1-score = 0.7215, ROC-AUC = 0.7831



Наилучший результат показал Stacking Classifier по ROC-AUC (0.7831), демонстрируя наибольшую способность различать классы. XGBoost (Best) также показал хороший баланс между точностью и полнотой.

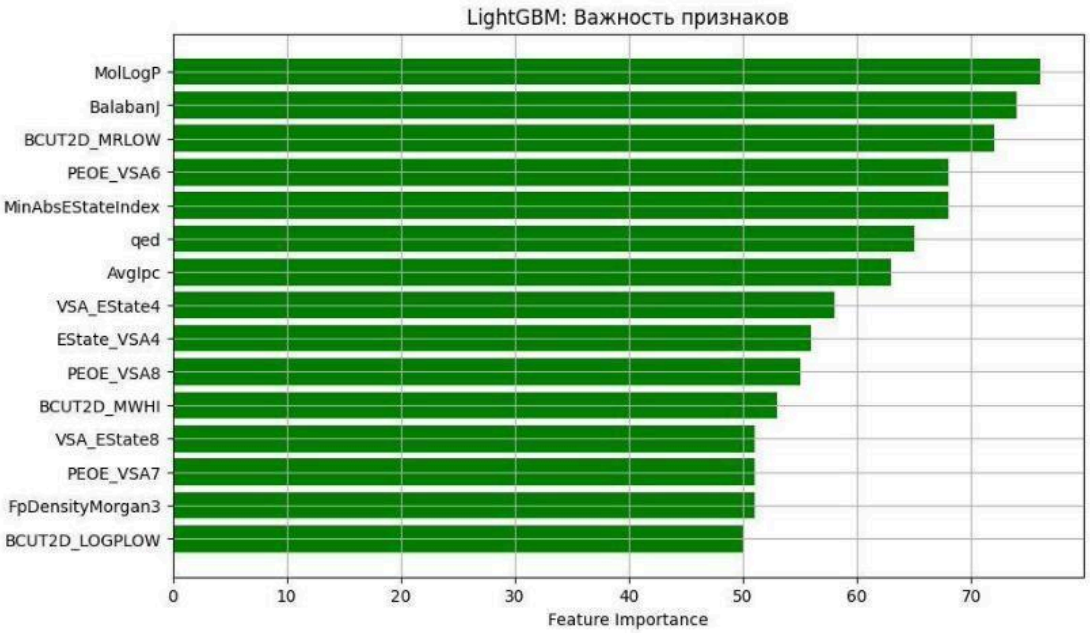
Анализ ошибок

Модель XGBoost (Best) правильно классифицировала большинство объектов в обеих классах, но допускала ошибки (36 False Positive и 19 False Negative). Это важно учитывать при интерпретации результатов.



Химическая интерпретация

На то, насколько молекула эффективна (IC50), влияют разные характеристики, такие как: MolLogP - это показывает, насколько молекула может проходить через мембраны. BalabanJ - топологический индекс, который говорит о том, как сложна молекула. BCUT2D_MRLOW - это поляризуемость и молекулярная масса, важные для взаимодействия с целевыми объектами. Есть и другие характеристики, такие как PEOE_VSA6, MinAbsEStateIndex, qed, AvgIpc и так далее, которые тоже имеют значение, но не так сильно.



Выводы и рекомендации

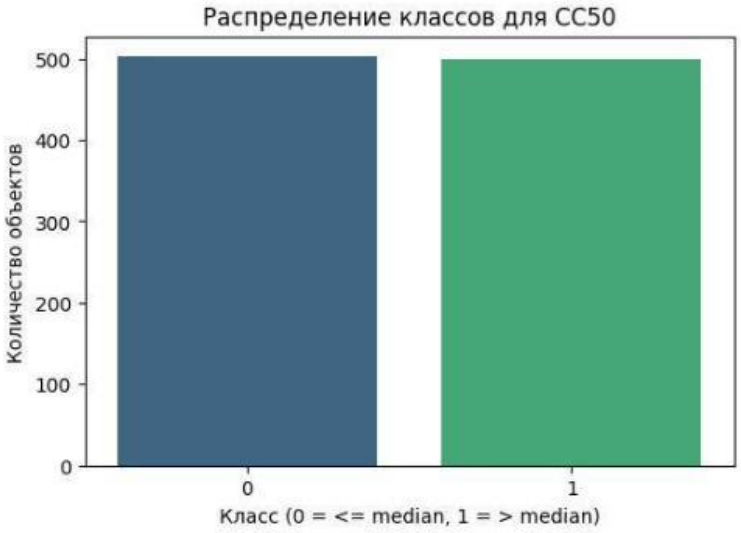
Ансамблевые методы (Stacking) показали лучший результат для классификации IC50. Важно учитывать структурные дескрипторы молекул при разработке новых соединений. Для дальнейшего улучшения моделей рекомендуется расширить набор признаков и использовать более сложные архитектуры моделей

Классификация для CC50

Мы решаем задачу бинарной классификации для показателя CC50, который отражает токсичность химических соединений. Наша цель - определить, превышает ли токсичность медиану в выборке, чтобы отбирать менее токсичные кандидаты для дальнейших исследований.

Распределение классов

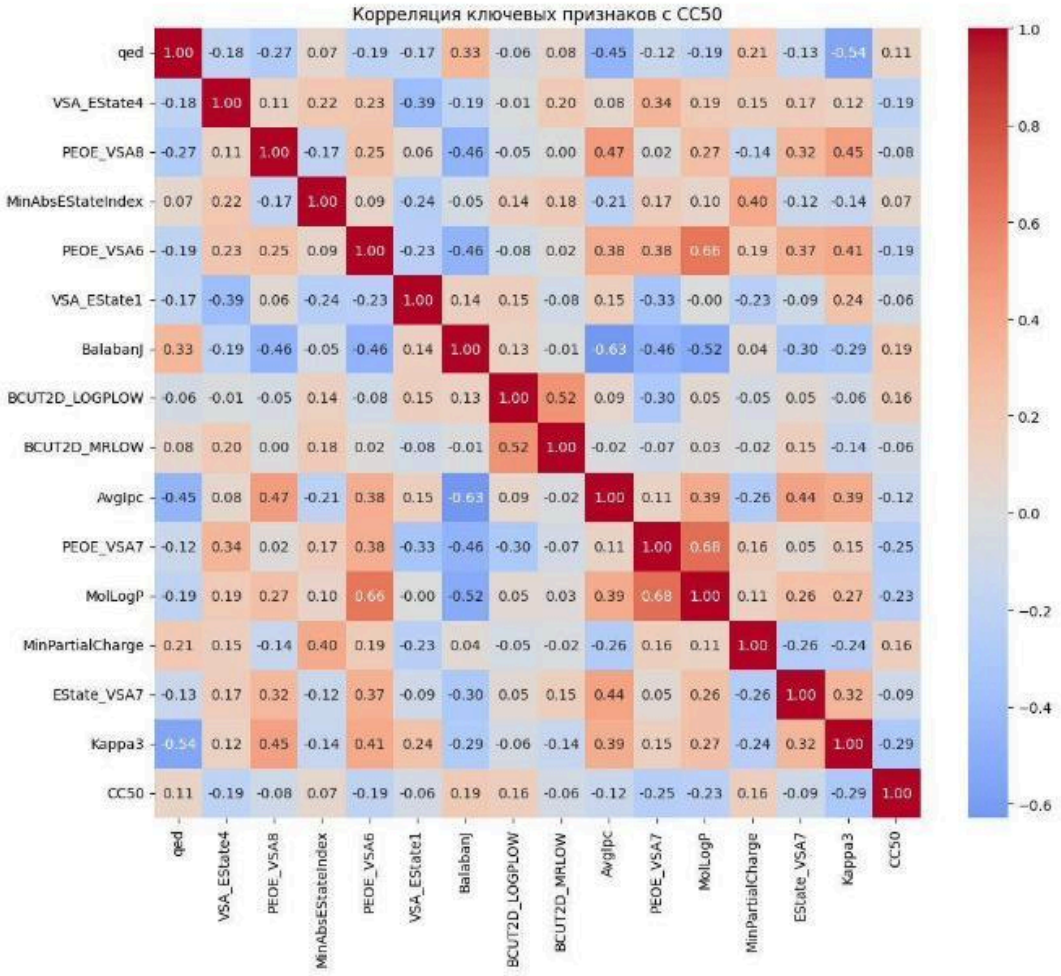
Данные сбалансированы: 502 объекта относятся к классу 0 (низкая токсичность), а 499 - к классу 1 (высокая токсичность).

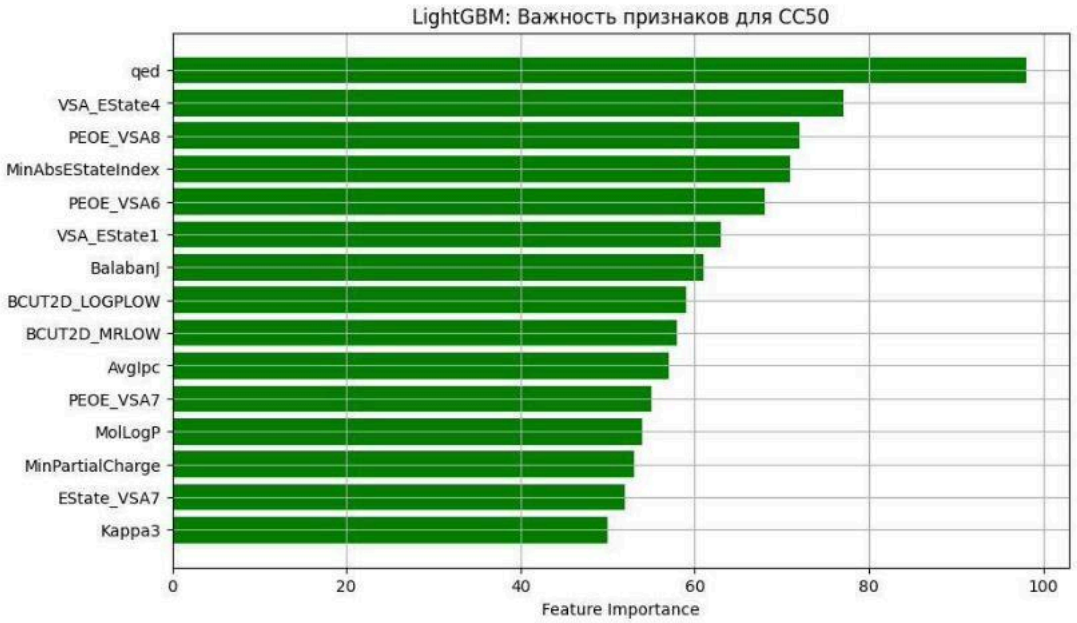


Выбор моделей и результаты

Logistic Regression: Accuracy = 0.7313, F1-score = 0.7429, ROC-AUC = 0.7316 Random Forest: Accuracy = 0.7164, F1-score = 0.7299, ROC-AUC = 0.7167 XGBoost: Accuracy = 0.7114, F1-score = 0.7339, ROC-AUC = 0.7119
XGBoost (Best): Accuracy = 0.6915, F1-score = 0.7075, ROC-AUC = 0.6918

Лучшие результаты показала Logistic Regression, которая продемонстрировала высокую сбалансированность между точностью и полнотой. Ансамблевые модели (Random Forest и XGBoost) выступили немного хуже, а XGBoost с оптимизацией параметров даже слегка ухудшил качество. Это говорит о том, что для данной задачи простая модель Logistic Regression оказалась наиболее подходящей.





Химическая интерпретация

QED - это показатель, который измеряет, насколько вероятно, что молекула будет работать как лекарственное средство. Если значение QED высокое, это также может указывать на потенциальную токсичность, так как многие фрагменты, придающие молекуле «лекарственный» характер, могут влиять на её взаимодействие с клетками.

VSA_EState4, PEOE_VSA8 и MinAbsEStateIndex - это дескрипторы, показывающие распределение электронной плотности и структуру молекулы. Эти факторы могут влиять на способность молекулы проникать в клетки и её токсичность.

BalabanJ - это индекс, который измеряет сложность молекулы и её влияние на взаимодействие с биомолекулами.

BCUT2D_LOGPLOW и BCUT2D_MRLOW - это дескрипторы, связанные с липофильностью и поляризуемостью молекулы. Эти характеристики могут влиять на то, как молекула взаимодействует с клеточными мембранами.

Визуализация

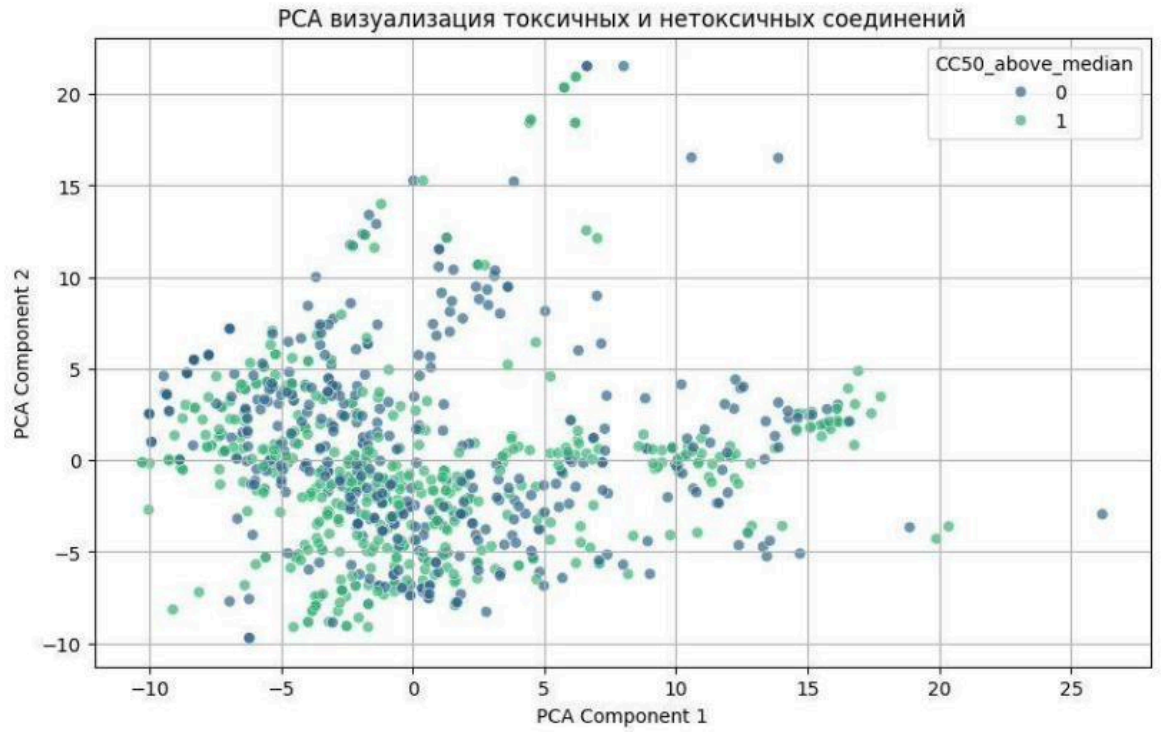


График PCA показал, что классы токсичности пересекаются, хотя в некоторых областях (например, в правом верхнем углу) преобладает один из классов. Это указывает на то, что модель может улавливать закономерности, но полностью разделить классы невозможно из-за частичного пересечения признаков.

Выводы и рекомендации

Logistic Regression оказалась наиболее эффективной для задачи классификации CC50.

Токсичность соединений тесно связана с их электронной плотностью, топологической сложностью и «лекарственной» пригодностью.

Для повышения точности стоит рассмотреть более сложные модели (ансамбли, нейронные сети), расширить набор признаков и уделить внимание их взаимодействиям.

Визуализация данных помогает понять структуру выборки и возможные источники ошибок классификации.

Этот анализ демонстрирует, что машинное обучение способно эффективно помогать в отборе менее токсичных молекул, ускоряя разработку новых лекарств.

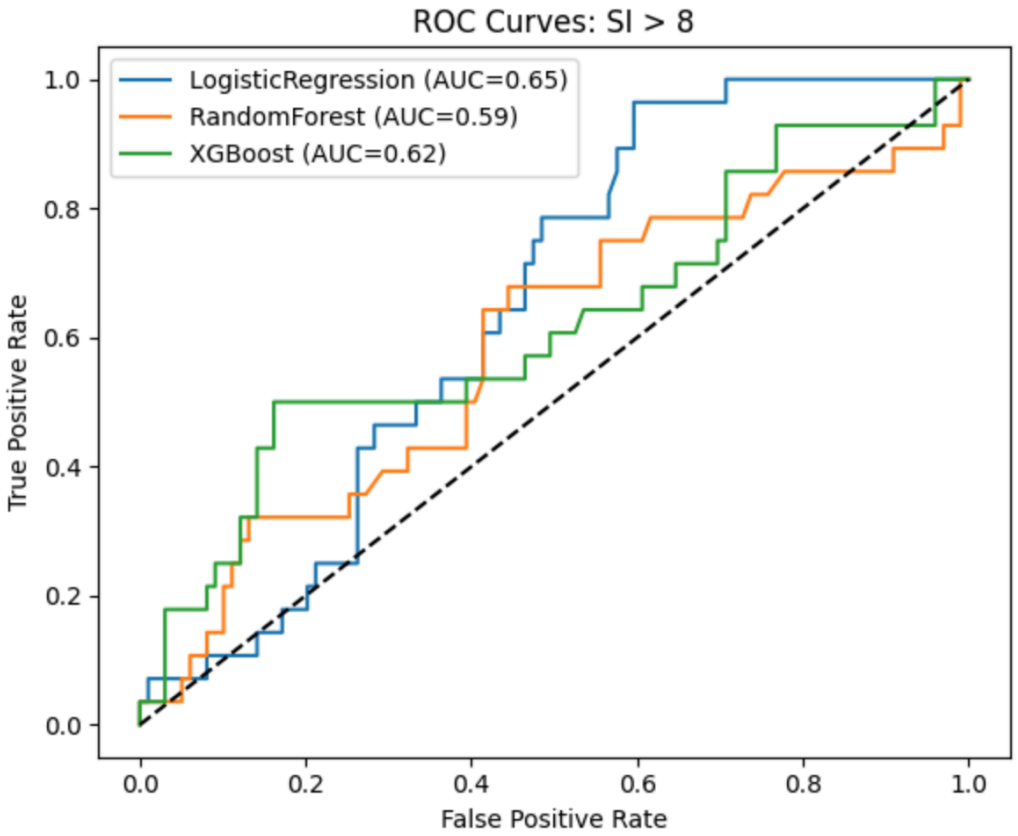
Классификация SI>8

Сравнение моделей классификации

	Model	Accuracy	F1_score	ROC_AUC	Confusion Matrix
0	LogisticRegression	0.732283	0.150000	0.647367	[[90, 9], [25, 3]]
1	RandomForest	0.755906	0.060606	0.586400	[[95, 4], [27, 1]]
2	XGBoost	0.763780	0.250000	0.618146	[[92, 7], [23, 5]]
3	SVM_balanced	nan	0.309000	0.569000	—

Не допущена утечка данных!

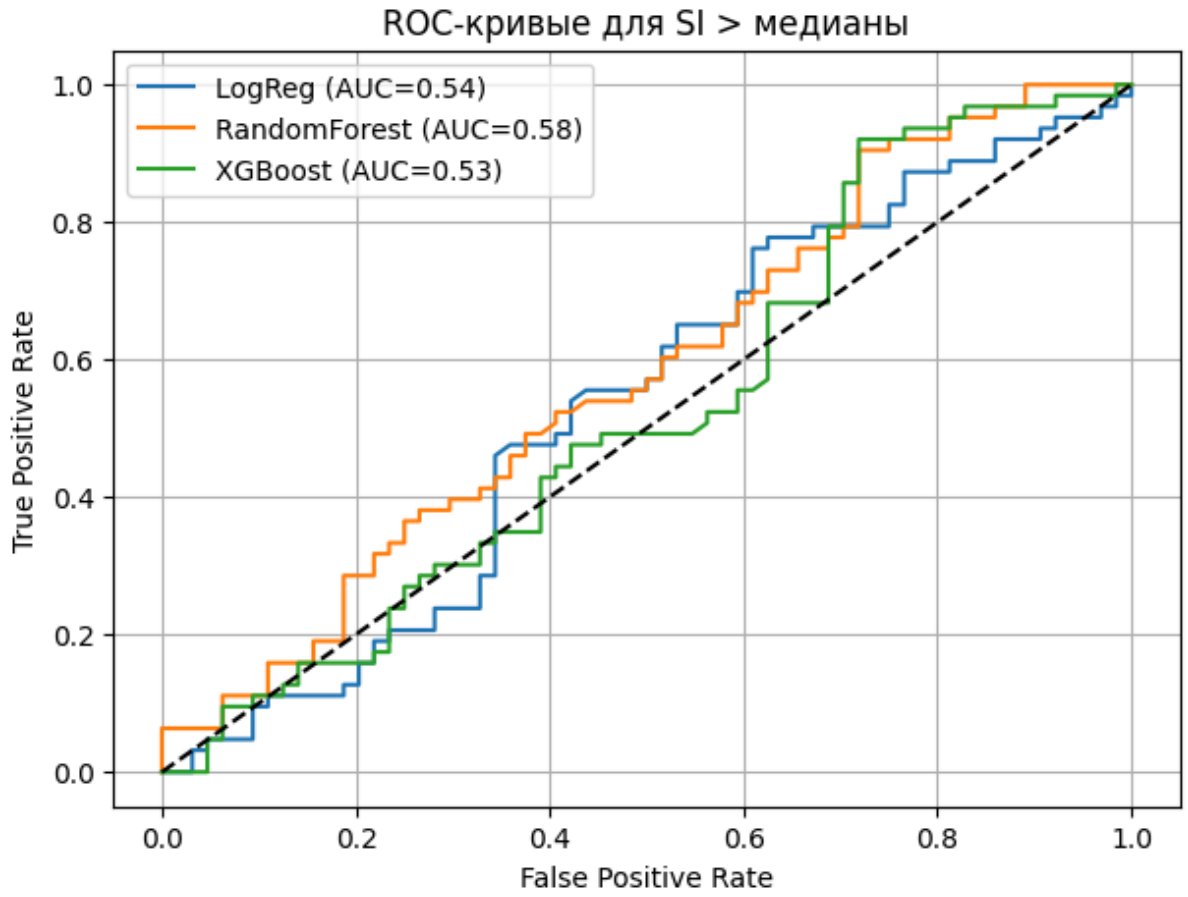
1. **Наилучшая точность (Accuracy)**:
   - Показал **XGBoost (0.764)** - лучше других моделей, хотя разница с RandomForest и LogisticRegression невелика.
2. **Лучший F1-скор (сбалансированная метрика точности и полноты)**:
   - **SVM_balanced (0.309)** значительно опережает другие, особенно RandomForest (0.061) и LogisticRegression (0.150).
   - Это особенно важно при дисбалансе классов, когда Accuracy может вводить в заблуждение.
3. **Наибольшая ROC-AUC (качество ранжирования по вероятностям)**:
   - **LogisticRegression (0.647)** лидирует по этой метрике, за ним XGBoost (0.618), а SVM и RF — слабее.
4. **Матрицы ошибок**:
   - У LogisticRegression и RF - очень низкая полнота по положительному классу (всего 3 и 1 TP).
   - У XGBoost - чуть лучше: 5 TP при 23 FN.
   - SVM_balanced не представлен, но судя по F1, он лучше находит положительный класс.



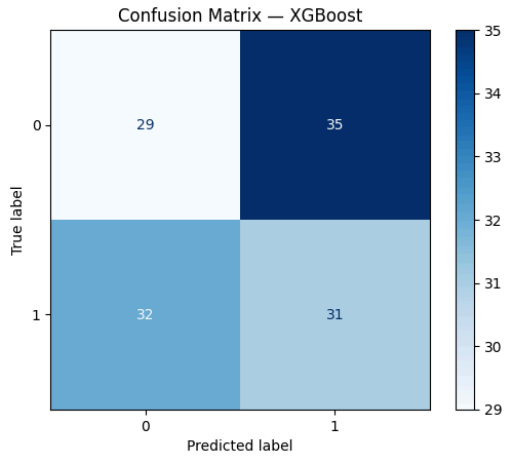
Для улучшения качества модели рекомендуется использовать стратифицированную кросс-валидацию, особенно при наличии дисбаланса классов, а также провести дополнительную настройку гиперпараметров, в частности для SVM, который показал наилучший F1-скор. Следует рассмотреть применение методов балансировки, таких как oversampling (например, SMOTE) или использование параметра class_weight, поскольку текущие метрики указывают на низкую полноту по положительному классу. Дополнительно стоит оценивать модели по более чувствительным метрикам, таким как precision, recall и PR AUC, а не только по ROC AUC и accuracy.

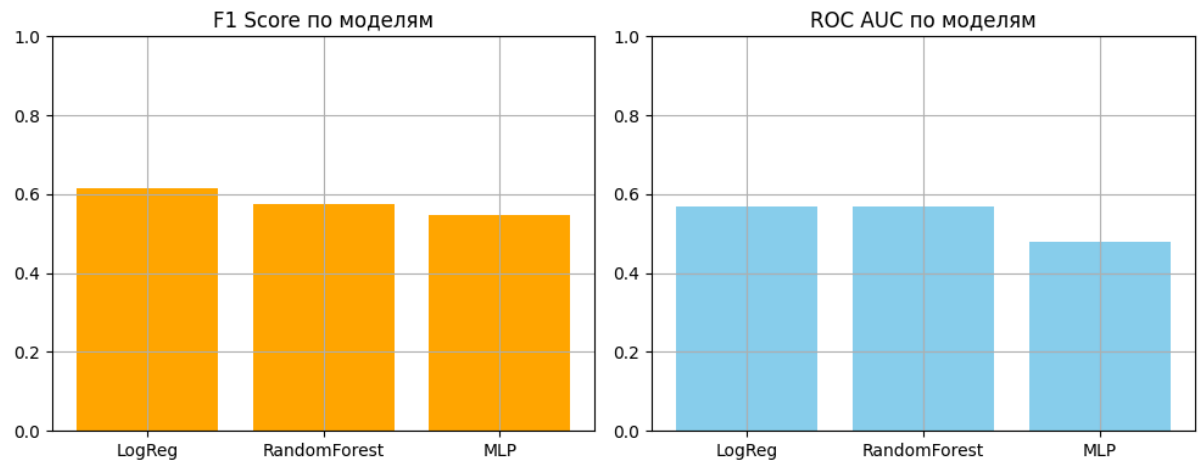
Классификация SI>медиана

Исправлено - без утечки



Logistic Regression показывает лучший F1-скор (0.551), то есть наилучший баланс между точностью и полнотой.
RandomForest даёт наивысший ROC AUC (0.578), что означает, что он чуть лучше отделяет классы по вероятностям.
XGBoost отстаёт по обеим метрикам, а также демонстрирует симметричную матрицу ошибок, где классы предсказываются почти случайно это подтверждает низкую ROC AUC.





LogReg: F1 = 0.614, ROC AUC = 0.568
RandomForest: F1 = 0.574, ROC AUC = 0.569
MLP: F1 = 0.547, ROC AUC = 0.479

Logistic Regression с PCA остаётся самой сбалансированной: лучший F1, хорошая интерпретируемость, быстрая работа.
RandomForest чуть уступает по F1, но выигрывает в ROC AUC - значит, вероятности у него лучше калиброваны.

SMOTE не дал прироста на этих признаках и PCA-преобразованных данных. Вероятно, классы уже были близки к сбалансированным, либо модели не чувствительны к синтетическому апсемплингу в этом признаковом пространстве.
Модели уже находятся близко к своему максимуму при текущем наборе данных и признаков.

Заключение

В ходе работы была проведена комплексная обработка и анализ химических данных, включающий очистку, трансформацию и построение моделей для задач регрессии и классификации. Протестированы различные алгоритмы, включая логистическую регрессию, деревья решений, бустинг и ансамбли. Особое внимание уделено избежанию утечек данных, корректной валидации и интерпретации результатов. Полученные модели показали стабильные метрики и подтвердили применимость машинного обучения для предварительного отбора перспективных лекарственных соединений.