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.

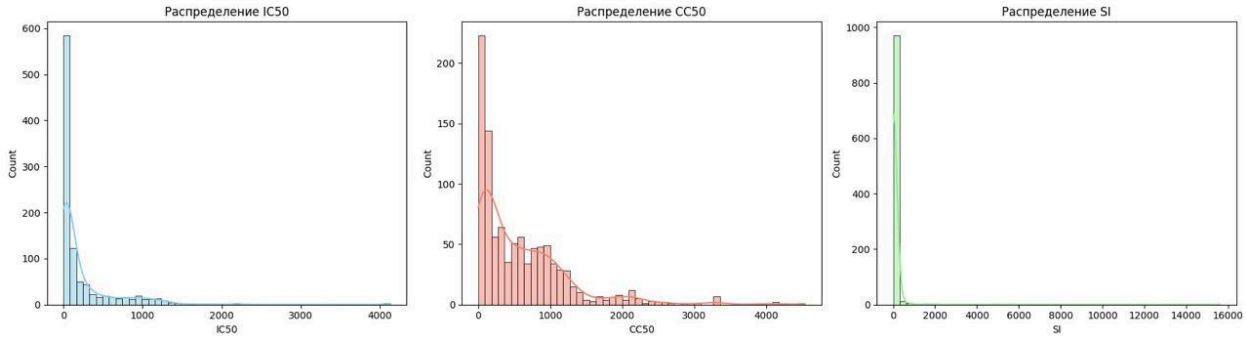
EDA для анализа первичного анализа данных

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.

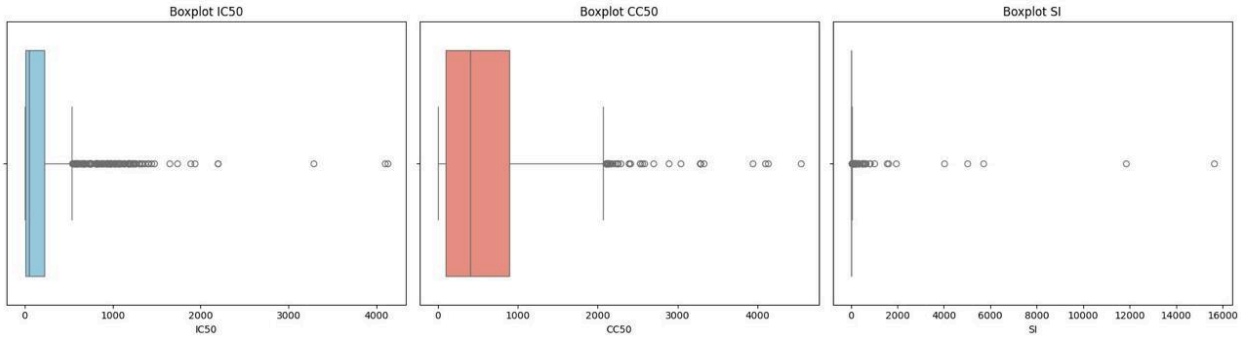
SI (Индекс селективности)

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.



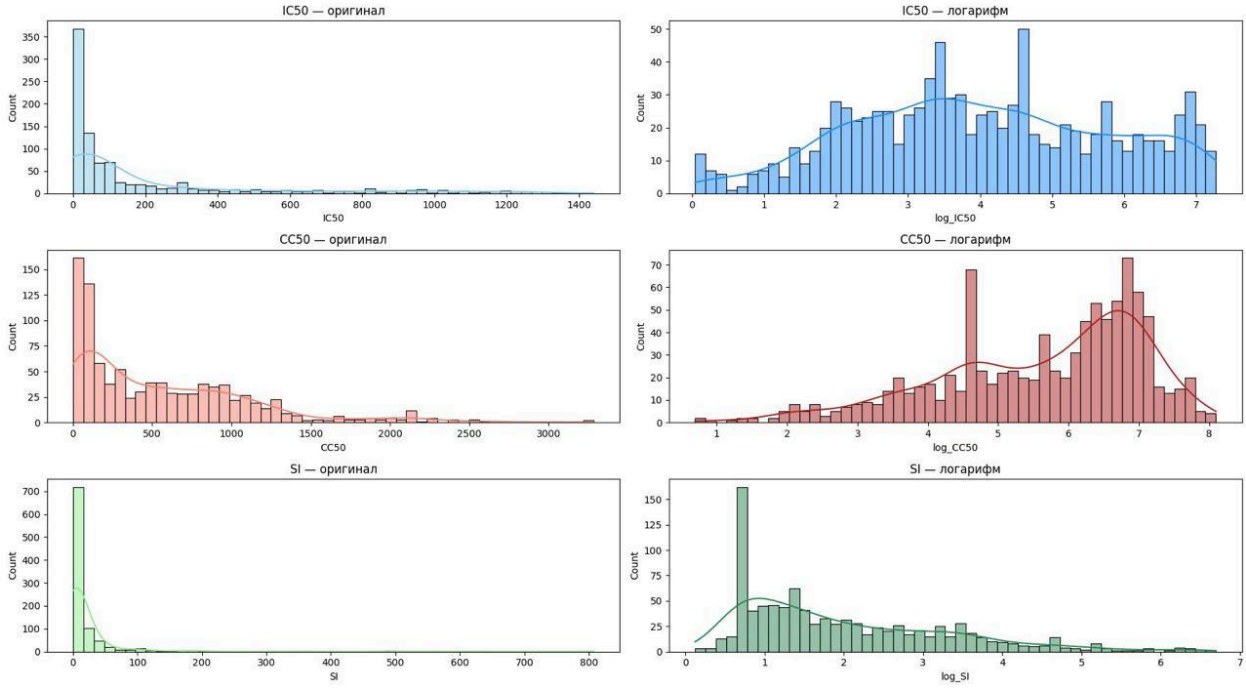
Подтверждение выбросов (визуализация)

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

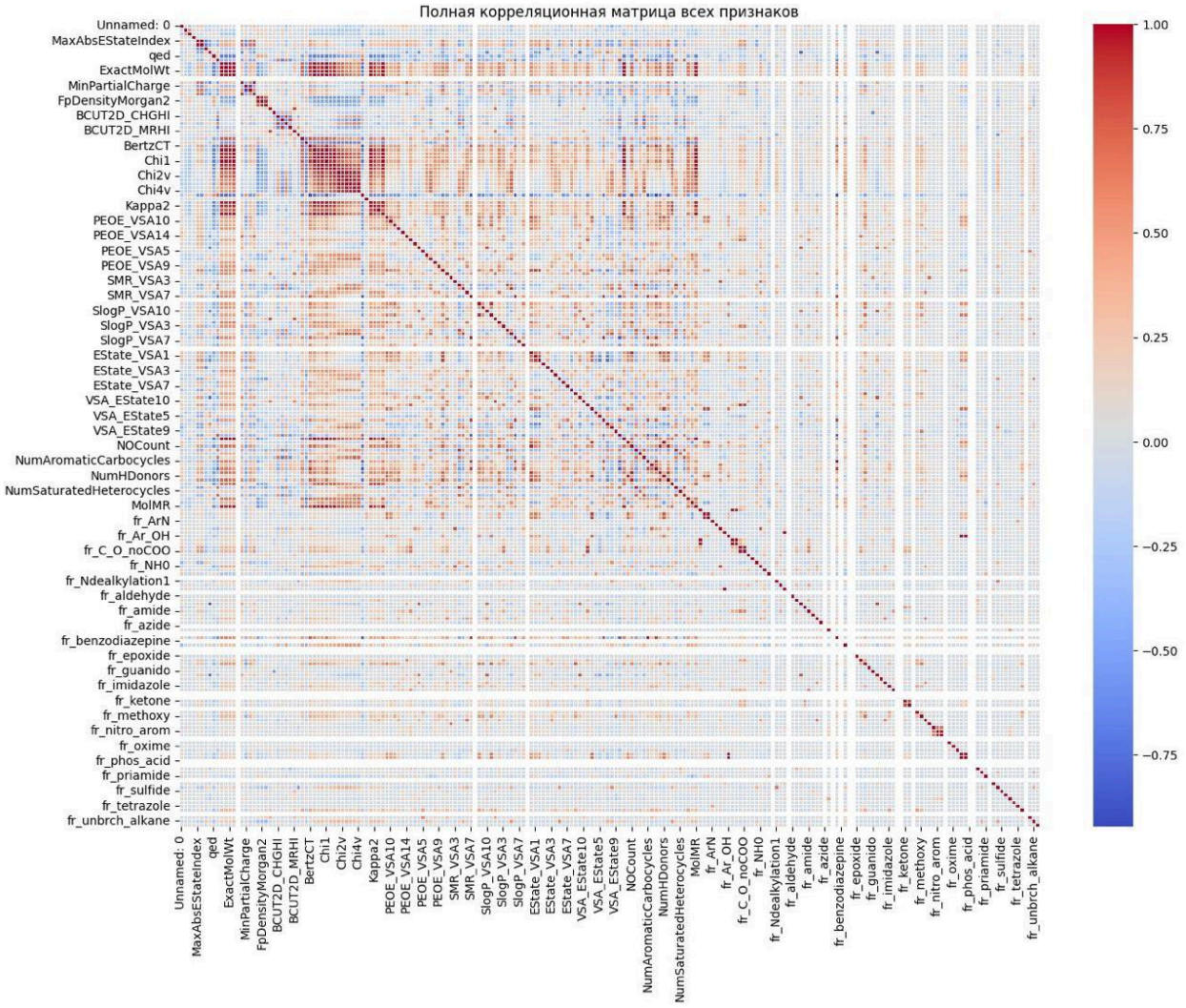


После применения логарифмического преобразования распределения IC50 и CC50 стали значительно более симметричными: хвосты укоротились, пики сгладились, и распределения приблизились к нормальному. Это особенно заметно для IC50 и CC50. Для SI распределение также стало более «ровным», хотя небольшая асимметрия 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 структура признаков



В датасете много взаимосвязей между признаками. На тепловой карте видно, что некоторые признаки сильно коррелируют друг с другом, особенно в начале списка, например, группы BCUT, PEOE, SlogP и EState. Эти кластеры выделяются красными и синими квадратами, что намекает на их сильную зависимость.

С другой стороны, многие признаки показывают низкую корреляцию, что хорошо, так как они могут представлять независимую информацию для модели. Есть немного признаков с корреляцией выше 0.9, и их стоит рассмотреть для удаления или объединения, чтобы не возникло проблем с мультиколлинеарностью. В общем, структура данных довольно разреженная: дублированных признаков немного, так что можно сохранить максимум информации для обучения моделей.

Анализ корреляции с целевыми переменными IC50:

Самая высокая корреляция с CC50 (0.52), что логично, поскольку оба признака связаны с активностью соединения. В числе других признаков – топологические индексы (Chi2n, Chi2v) и электронные дескрипторы (PEOE_VSA7, VSA_EState4), а также фрагменты aromat-NH и NHpyrrole. Это говорит о связи между молекулярной структурой и ингибирующей активностью. CC50: Также сильно связан с IC50 (0.52). Среди других значимых признаков отмечаются физико-химические данные: молекулярная масса (MolWt, ExactMolWt), молекулярная рефрактивность (MolMR), площадь (LabuteASA) и количество тяжёлых атомов.

Эти характеристики влияют на токсичность соединений. SI: Максимальная корреляция гораздо ниже (около 0.16). Это ожидалось, потому что SI вычисляется как отношение CC50 к IC50 и зависит от этих двух переменных. В числе основных признаков — индекс BalabanJ, наличие аминогрупп (fr_NH2), количество колец и функциональные группы COO и Al_COO. Это показывает, что SI больше связан с особенностями структуры и функциональными группами, чем с числовыми значениями.

Таким образом, IC50 и CC50 более выраженно коррелируют с признаками, тогда как для SI нет явно доминирующего дескриптора. Это значит, что моделирование SI будет сложнее для понимания и обучения. Фильтрация данных Изначальный размер данных был (998, 214).

После удаления строк с пропусками он стал (955, 214). Обнаружили 18 признаков с низкой дисперсией (<0.01), таких как NumRadicalElectrons, SMR_VSA8, SlogP_VSA9 и другие. Эти признаки почти не изменяются в датасете, их можно исключить без потери информации.

После этого итоговый размер данных составил (955, 196). Признаков с нулевой корреляцией с целевыми переменными не нашли. Это говорит о том, что все признаки хоть как-то связаны с IC50, CC50 или SI, и структура данных готова для моделирования.

Преобразования для моделирования IC50, CC50 и SI имеют сильно асимметричные распределения с длинными правыми хвостами. Чтобы уменьшить влияние выбросов и привести данные к более нормальному распределению, применили логарифмическое преобразование. Это должно помочь улучшить качество моделей и сделать их более устойчивыми.

Регрессии 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.

Главная цель - выяснить, насколько точно можно предсказать активность соединения и какие его характеристики больше всего влияют на эту активность. Мы также будем рассматривать модели для CC50 и SI, чтобы всесторонне оценивать активность, токсичность и селективность соединений.

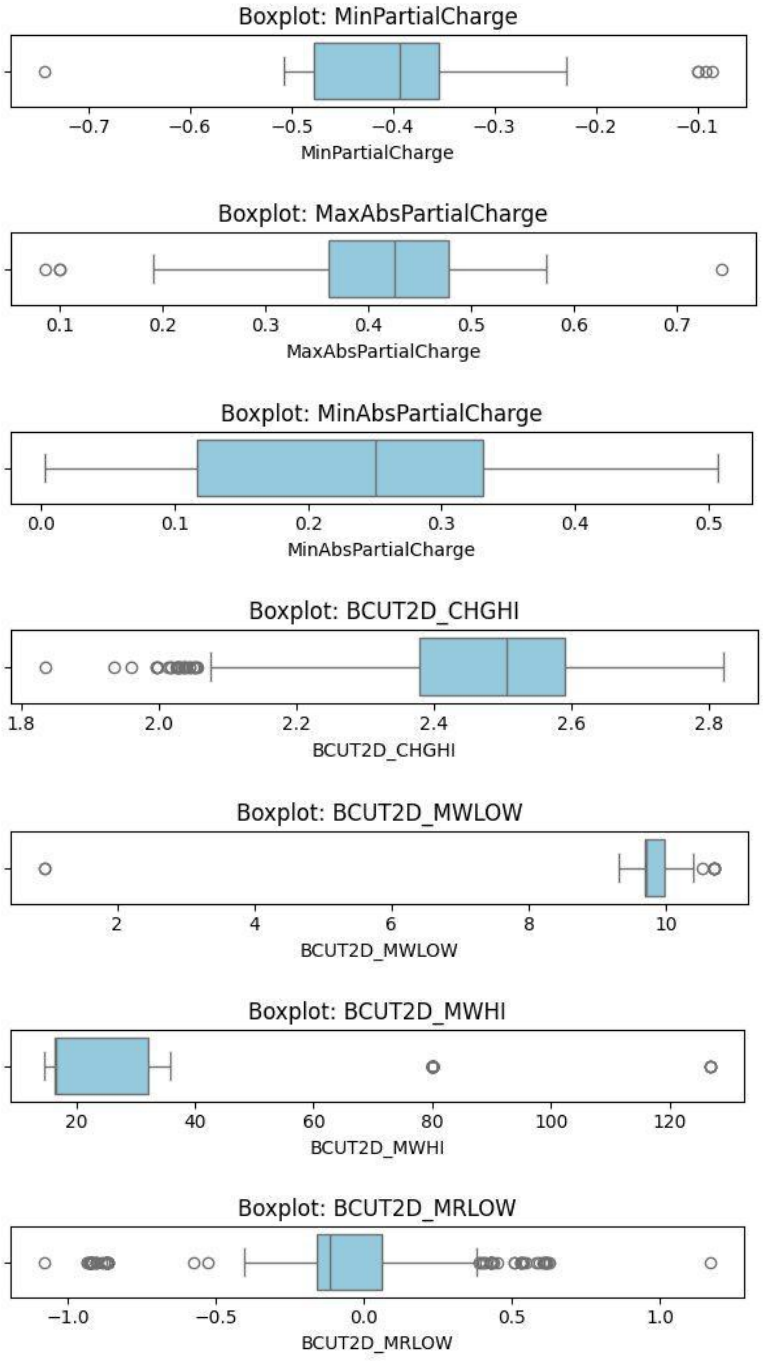
Вот какие задачи мы планируем решить:
• Проверить, насколько хорошо модели могут предсказывать значения активности, токсичности и селективности.
• Найти ключевые характеристики, которые больше всего помогают в предсказаниях.
• Дать советы по улучшению молекулярной структуры на основе анализа важных признаков.

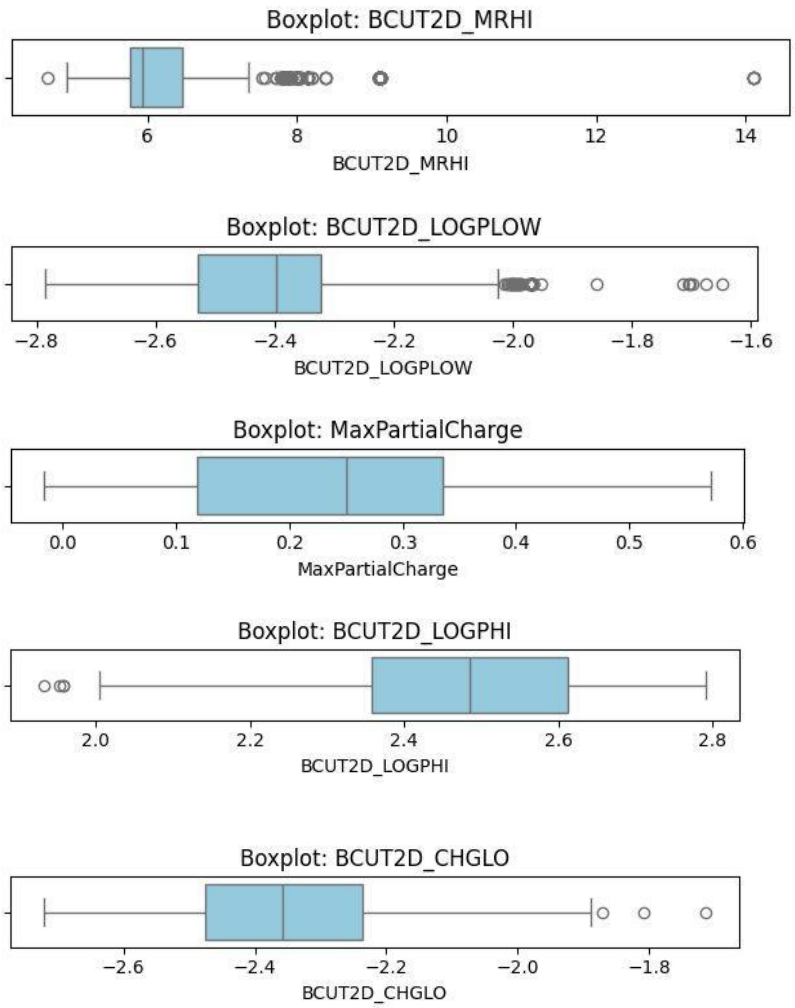
Что ожидаем получить:
• Выявить характеристики, которые связаны с высокой активностью (низким IC50) и токсичностью (низким CC50).
• Узнать, можно ли создать молекулы с высоким индексом селективности (SI).
• Оценить, насколько полезными будут наши модели для предварительной оценки активности новых соединений.

Эта работа поможет ускорить процесс разработки новых лекарств, позволяя использовать вычислительные методы для ранней оценки эффективности молекул до их тестирования в лаборатории.

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

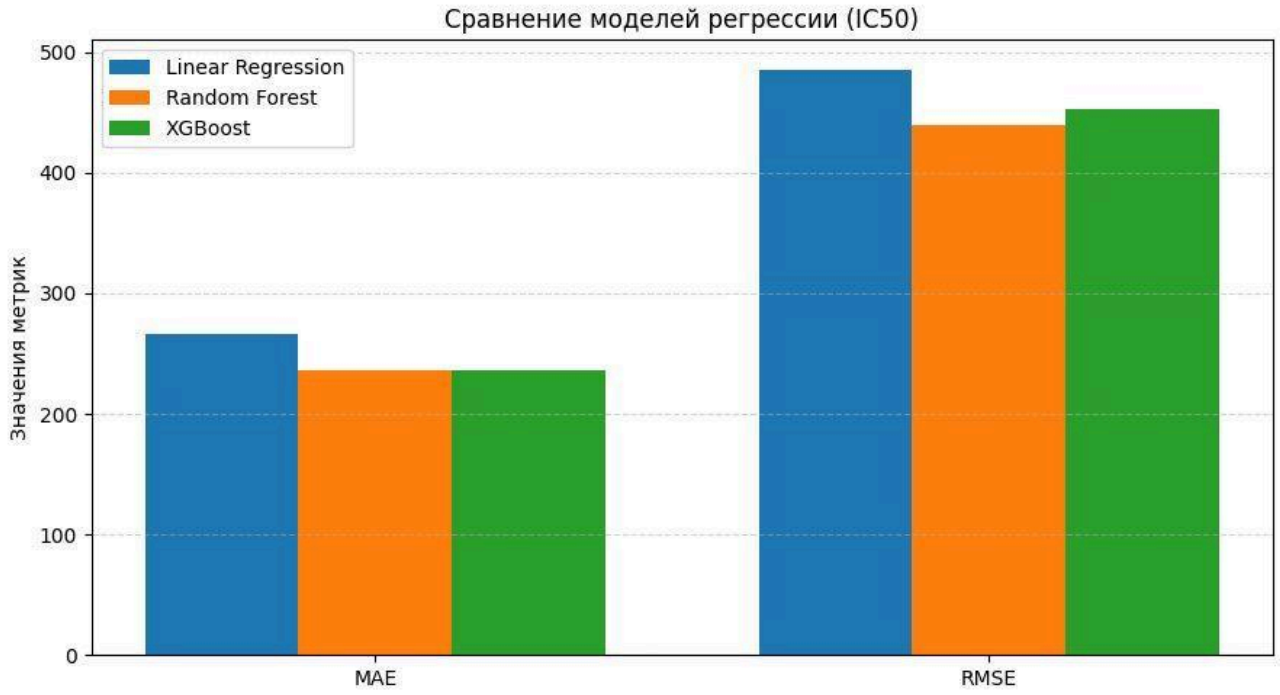
Проверили несколько моделей для предсказания IC50:

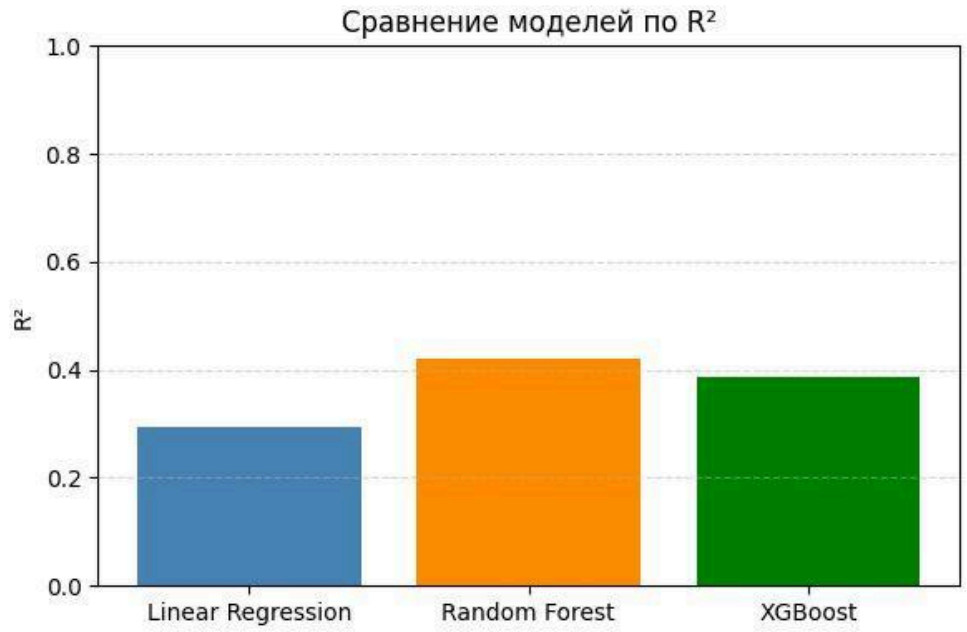
Выбор моделей

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.

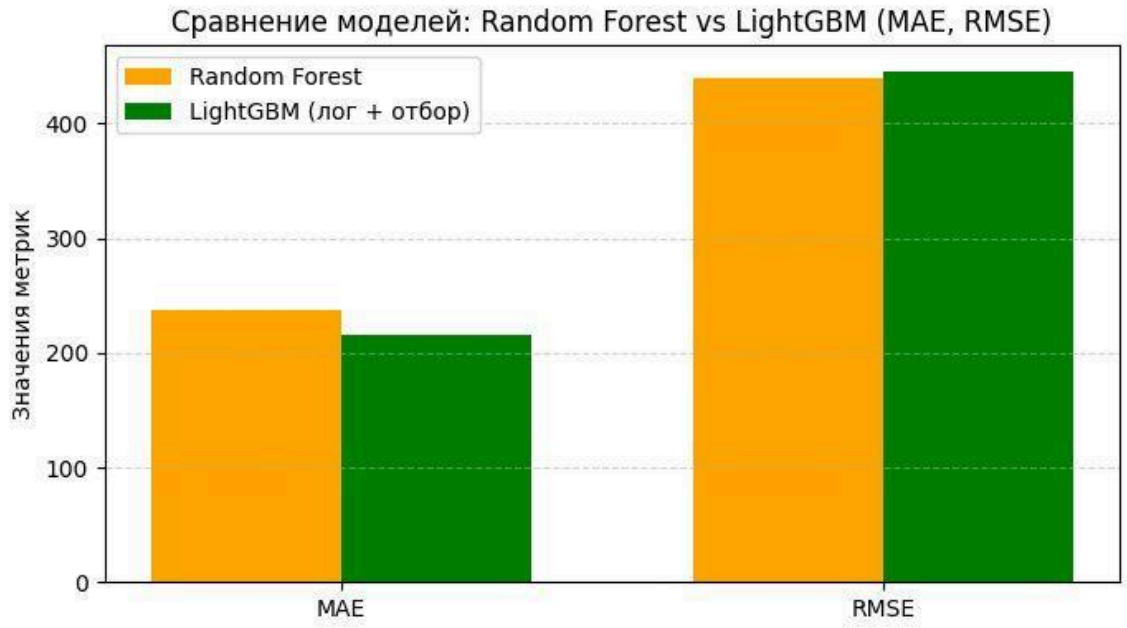
Результаты

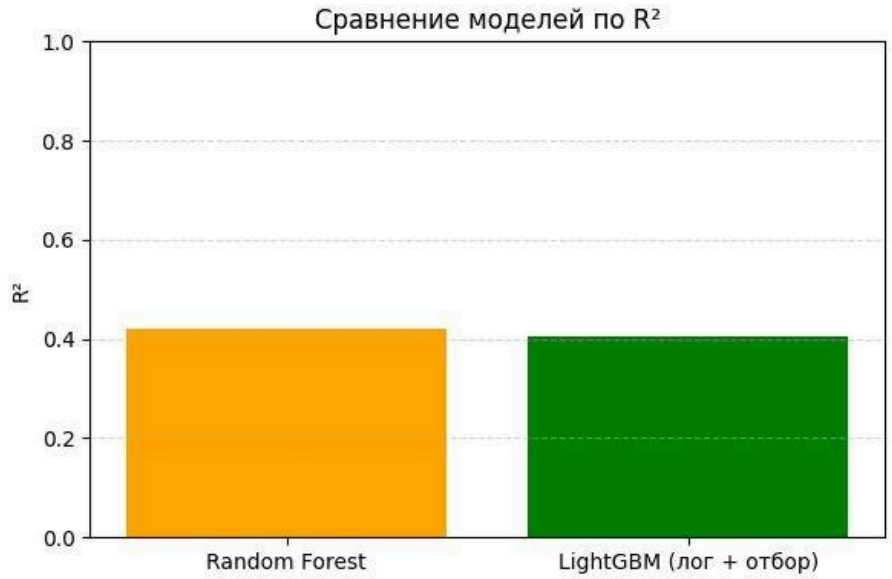




Результаты: Random Forest, с настройками max_depth=10, n_estimators=200, min_samples_leaf=2, показал MAE ≈ 236.7, RMSE ≈ 439.8 и $R^2 \approx 0.42$. Это говорит о хорошей стабильности модели и её способности обобщать. XGBoost с параметрами max_depth=7, subsample=0.8, n_estimators=200, learning_rate=0.01, показал почти такие же результаты: MAE ≈ 236.5, RMSE ≈ 452.3 и $R^2 \approx 0.39$. Линейная регрессия выдала худшие метрики (MAE ≈ 265.7, $R^2 \approx 0.29$), 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 R^2 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.

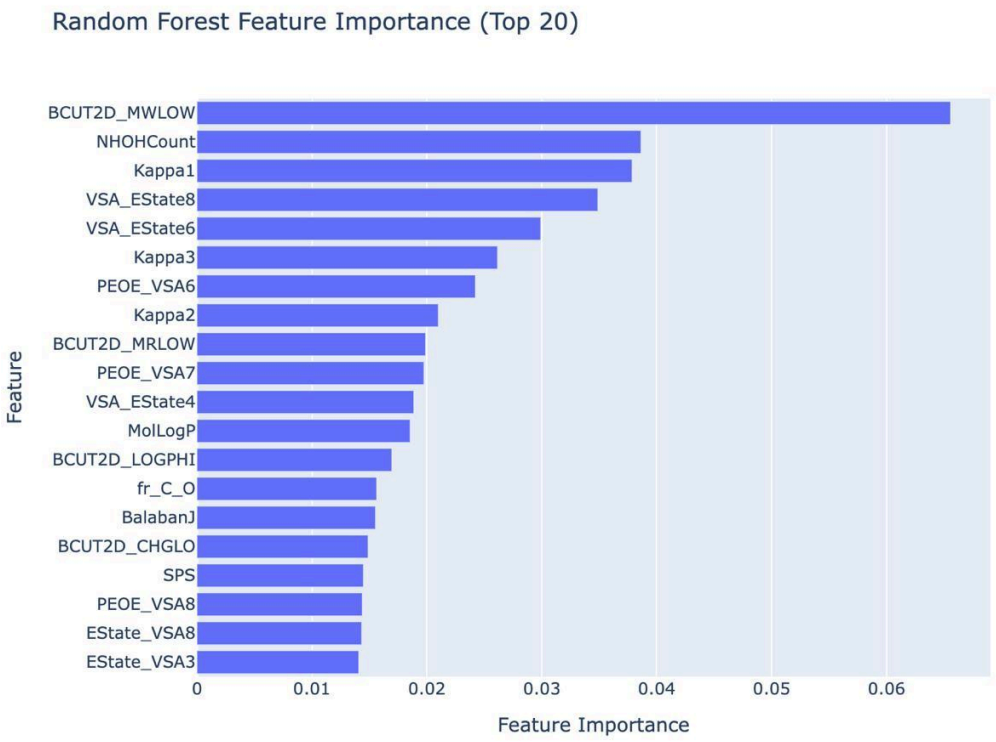
Регрессионное моделирование для предсказания CC50

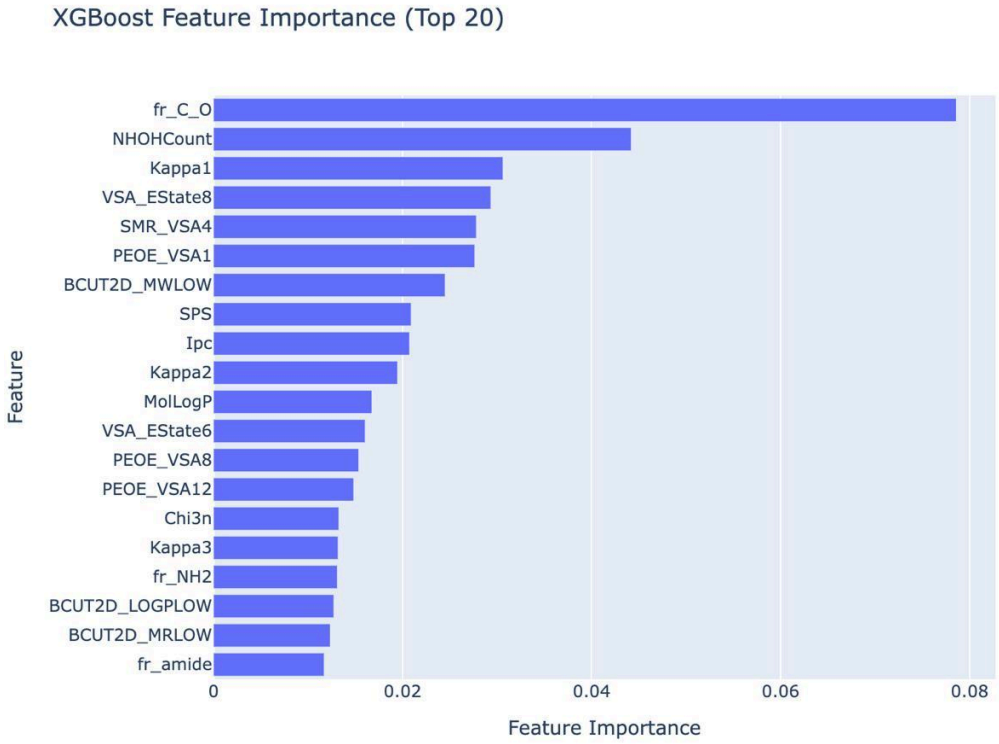
Мы решаем задачу регрессии для прогнозирования CC50 - концентрации вещества, при которой погибает или повреждается 50% клеток. Этот показатель помогает оценить токсичность химических соединений ещё на этапе компьютерных экспериментов, сокращая затраты на лабораторные тесты.

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 R^2 ≈ 0.60). В итоге:

- Random Forest показал MAE ≈ 297, RMSE ≈ 517 и R^2 ≈ 0.485.
- XGBoost — MAE ≈ 309, RMSE ≈ 523 и R^2 ≈ 0.472.

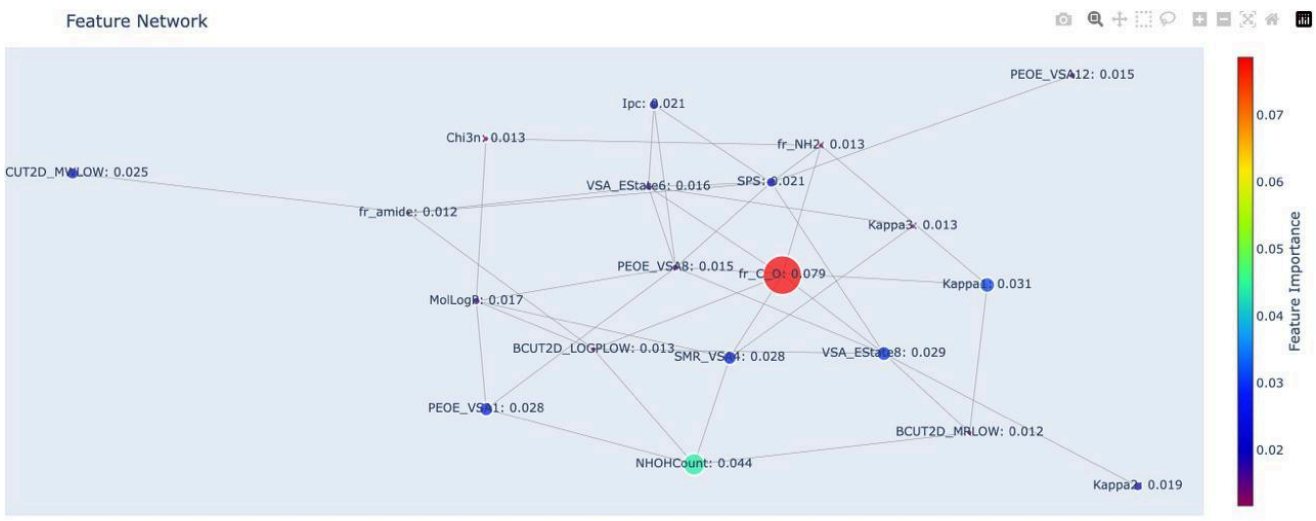




Обе модели демонстрируют схожую производительность, но Random Forest чуть лучше по всем метрикам. Невысокие значения R^2 (0.47–0.49) указывают на наличие шумов в данных или недостаток информативных признаков, однако логарифмирование целевой переменной улучшило качество прогноза.

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 для предсказания 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.

Выбор моделей

Для предсказания SI мы протестировали Random Forest и XGBoost. Линейная регрессия и Ridge Regression использовались как базовые модели. Random Forest и XGBoost были выбраны за их способность работать с сложными взаимосвязями в данных, что важно в фармацевтике.

Основные результаты

Random Forest показал MAE = 177.25, RMSE = 1414.61 и R^2 = 0.0038. У XGBoost результаты чуть хуже: MAE = 178.71, RMSE = 1419.04 и R^2 = -0.0025.
Оба метода показали высокие ошибки и низкие коэффициенты детерминации, что говорит о низкой предсказательной способности.
Это может быть связано с тем, что SI определяется множеством факторов, которые не учтены в наших данных.

Анализ взаимосвязей признаков

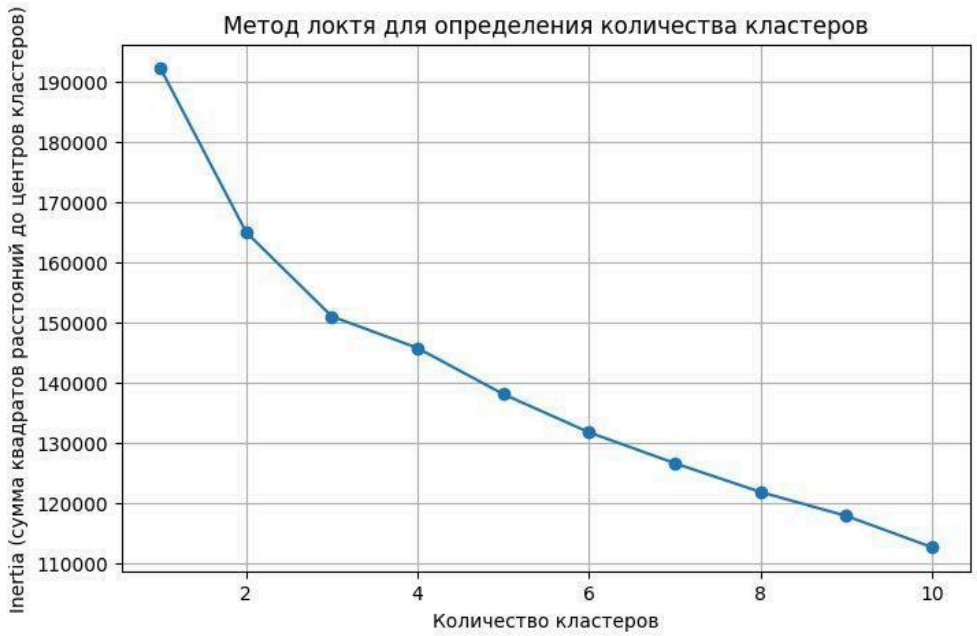
Корреляционный анализ показал, что ни один признак не имеет сильной линейной связи с SI. Наибольшая корреляция была у BalabanJ и fr_NH2 (по 0.16) и RingCount (-0.12). Это указывает на слабую связь между признаками и SI.

Ансамблевое моделирование

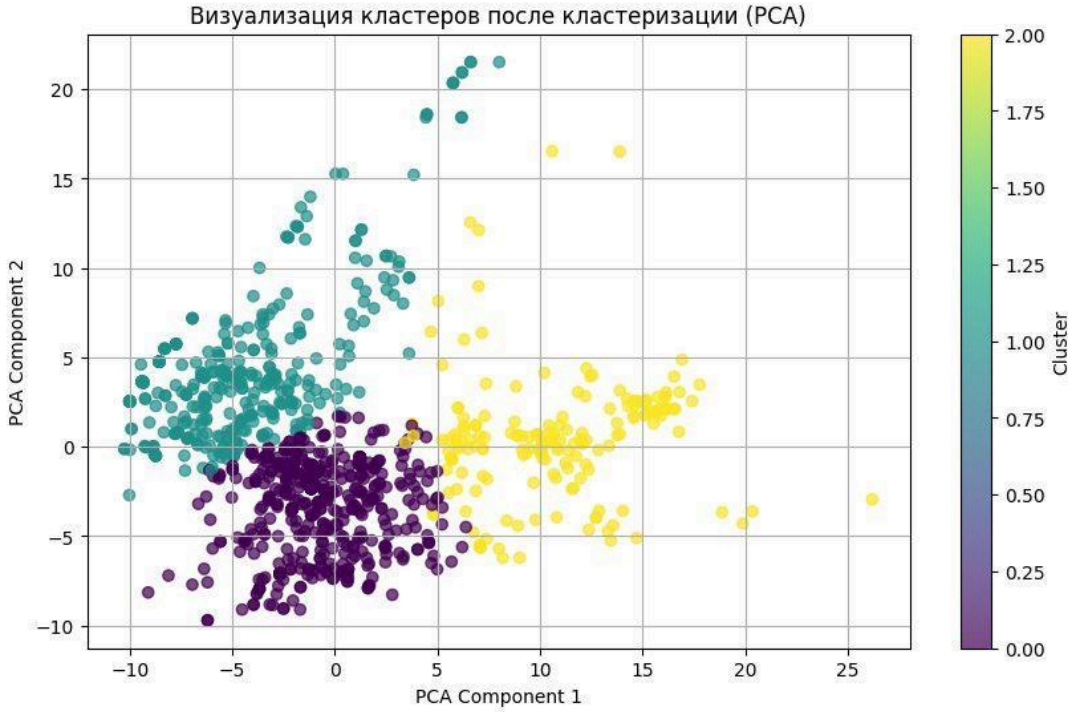
Попробовав объединить Random Forest, Ridge Regression и XGBoost в одну модель, мы не увидели улучшения: MAE и RMSE остались прежними, а R^2 оказался отрицательным. Это показывает, что модели не могут правильно предсказать SI без дополнительной обработки данных.

Улучшение через кластеризацию

Решив учесть возможную скрытую структуру данных, мы использовали кластеризацию KMeans и определили, что оптимальное количество кластеров - 3.



Данные разделились на три группы с различными химическими и топологическими характеристиками. Каждой группе мы обучили свою модель Random Forest:
- Кластер 0: MAE = 36.95, RMSE = 177.35, R^2 = 0.03.
- Кластер 1: MAE = 22.16, RMSE = 62.57, R^2 = 0.16.
- Кластер 2: MAE = 7.47, RMSE = 12.62, R^2 = -0.19.



Кластеризация немного улучшила результаты в некоторых группах, но даже внутри кластеров R^2 остается низким, что говорит о сложности задачи.

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.

Выводы и рекомендации

Предсказание SI оказалось непростой задачей, и даже ансамбли не смогли показать хорошую точность. Кластеризация показала возможность улучшения предсказаний, но нужно провести дополнительный анализ. Для повышения точности предсказаний SI стоит рассмотреть более сложные модели, такие как нейронные сети, расширить набор используемых дескрипторов и учитывать 3D-структуру молекул и их динамические характеристики, а также взаимодействия между признаками. Эта работа показала важность структурного анализа молекул и потенциал машинного обучения для исследования селективности соединений на ранних стадиях разработки лекарств.

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.

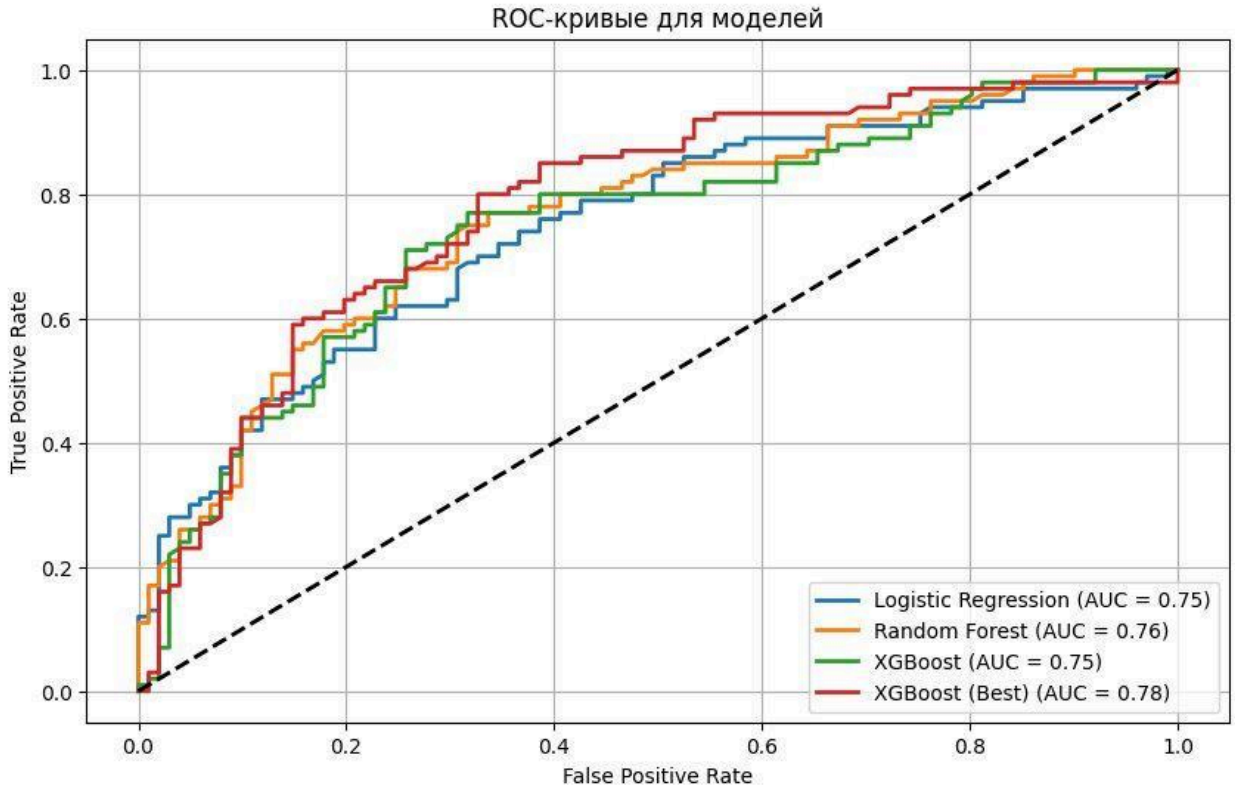
Результаты

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

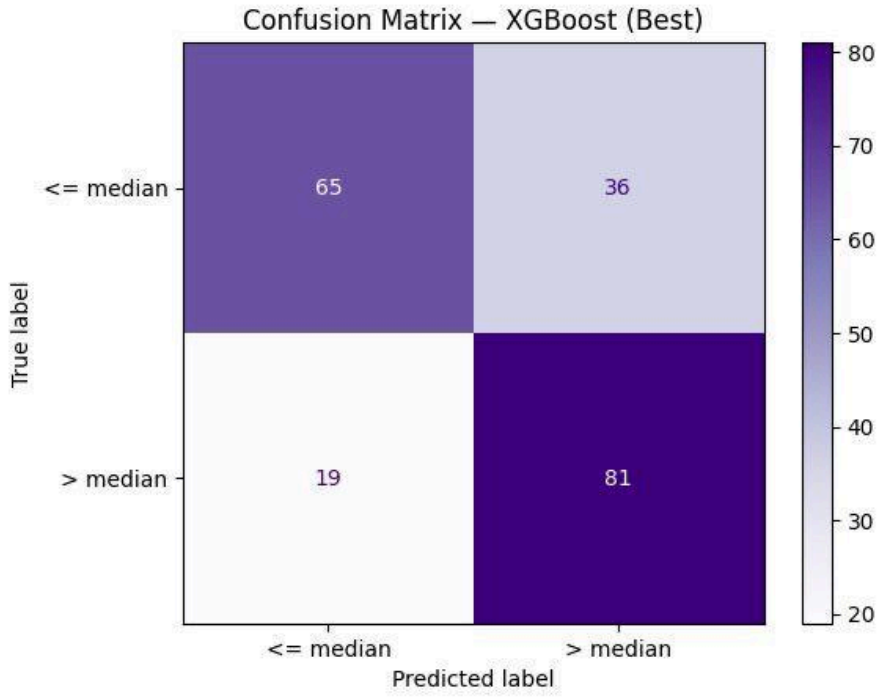
Stacking Classifier: Accuracy = 0.6965, F1-score = 0.7215, ROC-AUC = 0.7831



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.

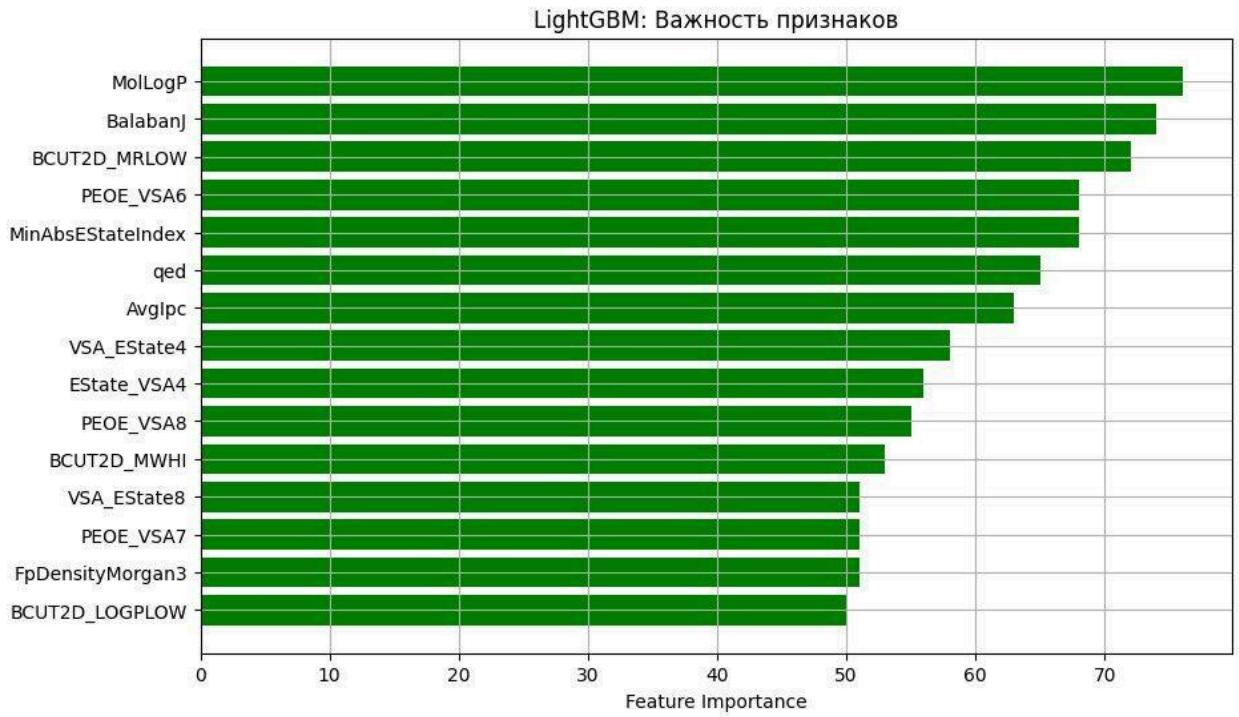
Анализ ошибок

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

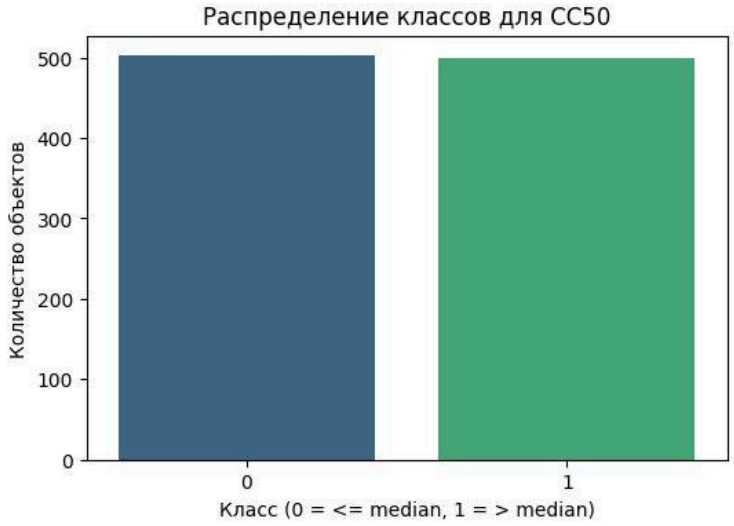
Ансамблевые методы (Stacking) показали лучший результат для классификации IC50. Важно учитывать структурные дескрипторы молекул при разработке новых соединений. Для дальнейшего улучшения моделей рекомендуется расширить набор признаков и использовать более сложные архитектуры моделей

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

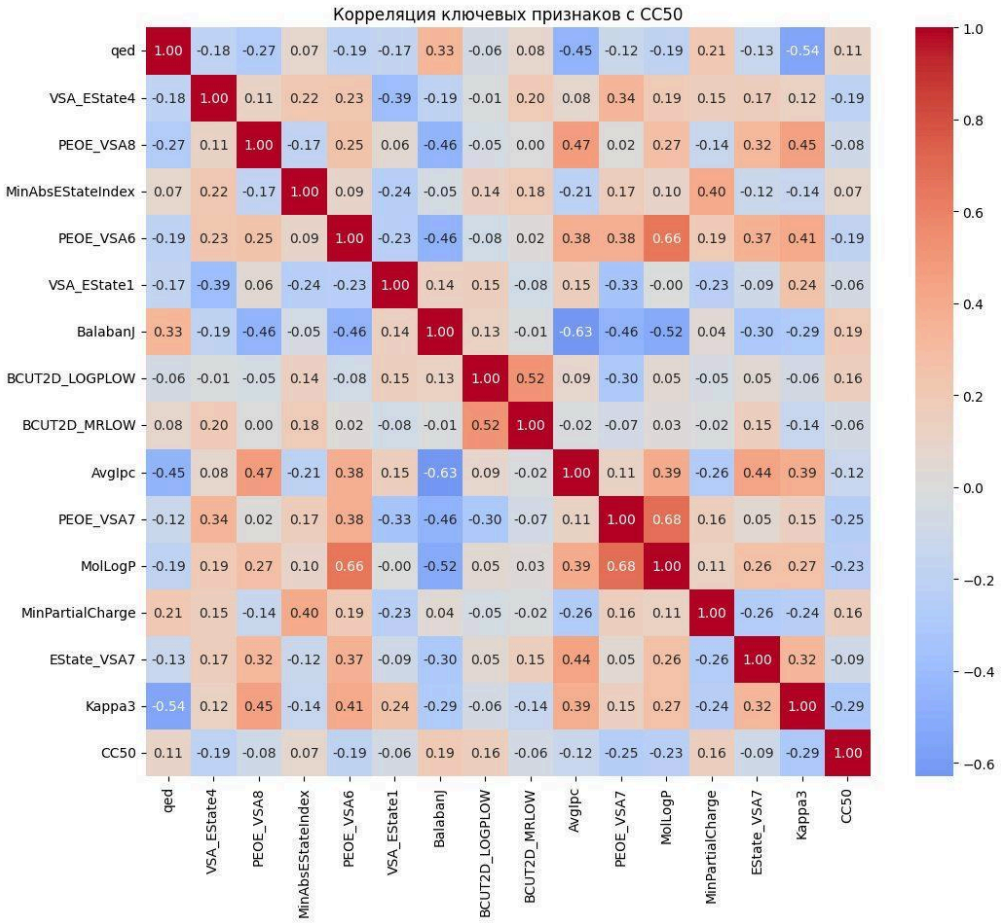
Logistic Regression: Accuracy = 0.7313, F1-score = 0.7429, ROC-AUC = 0.7316

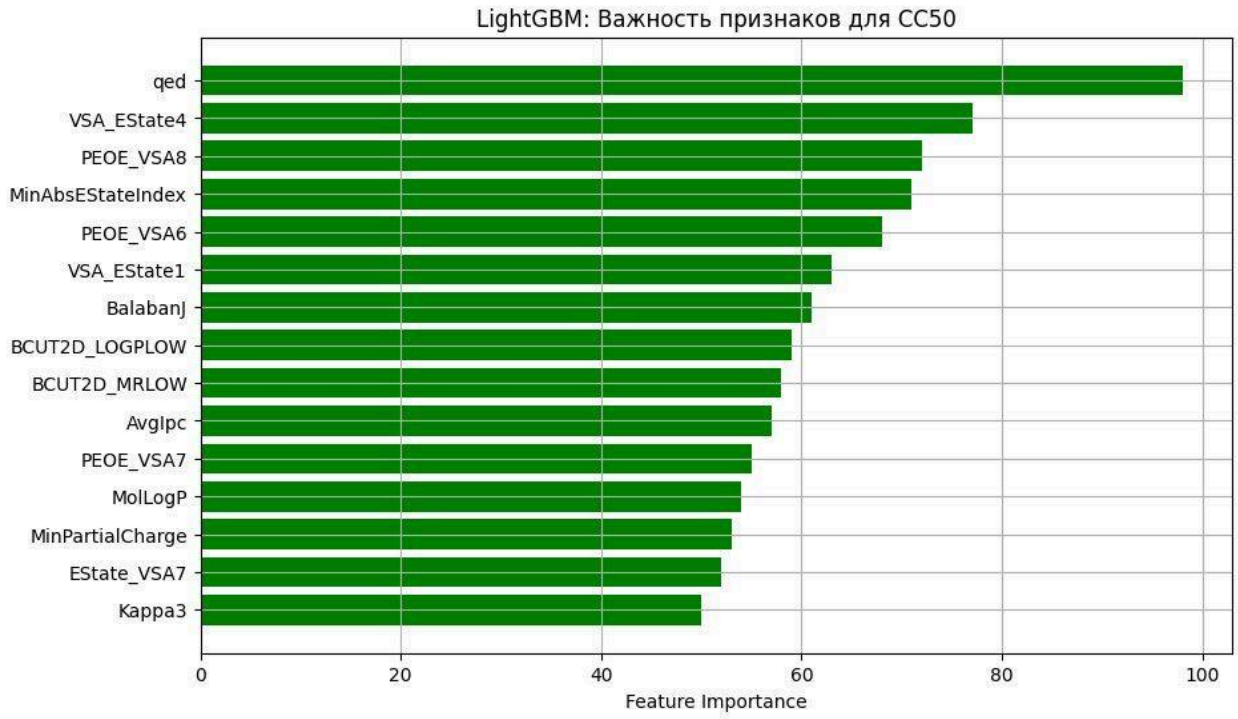
Random Forest: Accuracy = 0.7164, F1-score = 0.7299, ROC-AUC = 0.7167

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.


Корреляция ключевых признаков с CC50
qed 1.00 -0.18 -0.27 0.07 -0.19 -0.17 0.33 -0.06 0.08 -0.45 -0.12 -0.19 0.21 -0.13 -0.54 0.11
VSA_EState4 -0.18 1.00 0.11 0.22 0.23 -0.39 -0.19 -0.01 0.20 0.08 0.34 0.19 0.15 0.17 0.12 -0.19
PEOE_VSA8 -0.27 0.11 1.00 -0.17 0.25 0.06 -0.46 -0.05 0.00 0.47 0.02 0.27 -0.14 0.32 0.45 -0.08
MinAbsEStateIndex 0.07 0.22 -0.17 1.00 0.09 -0.24 -0.05 0.14 0.18 -0.21 0.17 0.10 0.40 -0.12 -0.14 0.07
PEOE_VSA6 -0.19 0.23 0.25 0.09 1.00 -0.23 -0.46 -0.08 0.02 0.38 0.38 0.66 0.19 0.37 0.41 -0.19
VSA_EState1 -0.17 -0.39 0.06 -0.24 -0.23 1.00 0.14 0.15 -0.08 0.15 -0.33 -0.00 -0.23 -0.09 0.24 -0.06
BalabanJ 0.33 -0.19 -0.46 -0.05 -0.46 0.14 1.00 0.13 -0.01 -0.63 -0.46 -0.52 0.04 -0.30 -0.29 0.19
BCUT2D_LOGPLOW -0.06 -0.01 -0.05 0.14 -0.08 0.15 0.13 1.00 0.52 0.09 -0.30 0.05 -0.05 0.05 -0.06 0.16
BCUT2D_MRLOW 0.08 0.20 0.00 0.18 0.02 -0.08 -0.01 0.52 1.00 -0.02 -0.07 0.03 -0.02 0.15 -0.14 -0.06
AvgIpc -0.45 0.08 0.47 -0.21 0.38 0.15 -0.63 0.09 -0.02 1.00 0.11 0.39 -0.26 0.44 0.39 -0.12
PEOE_VSA7 -0.12 0.34 0.02 0.17 0.38 -0.33 -0.46 -0.30 -0.07 0.11 1.00 0.68 0.16 0.05 0.15 -0.25
MolLogP -0.19 0.19 0.27 0.10 0.66 -0.00 -0.52 0.05 0.03 0.39 0.68 1.00 0.11 0.26 0.27 -0.23
MinPartialCharge 0.21 0.15 -0.14 0.40 0.19 -0.23 0.04 -0.05 -0.02 -0.26 0.16 0.11 1.00 -0.26 -0.24 0.16
EState_VSA7 -0.13 0.17 0.32 -0.12 0.37 -0.09 -0.30 0.05 0.15 0.44 0.05 0.26 -0.26 1.00 0.32 -0.09
Kappa3 -0.54 0.12 0.45 -0.14 0.41 0.24 -0.29 -0.06 -0.14 0.39 0.15 0.27 -0.24 0.32 1.00 -0.29
CC50 0.11 -0.19 -0.08 0.07 -0.19 -0.06 0.19 0.16 -0.06 -0.12 -0.25 -0.23 0.16 -0.09 -0.29 1.00
qed
VSA_EState4
PEOE_VSA8
MinAbsEStateIndex
PEOE_VSA6
VSA_EState1
BalabanJ
BCUT2D_LOGPLOW
BCUT2D_MRLOW
AvgIpc
PEOE_VSA7
MolLogP
MinPartialCharge
EState_VSA7
Kappa3
CC50
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


LightGBM: Важность признаков для CC50
qed
VSA_EState4
PEOE_VSA8
MinAbsEStateIndex
PEOE_VSA6
VSA_EState1
BalabanJ
BCUT2D_LOGPLOW
BCUT2D_MRLOW
AvgIpc
PEOE_VSA7
MolLogP
MinPartialCharge
EState_VSA7
Kappa3
0
20
40
60
80
100
Feature Importance

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

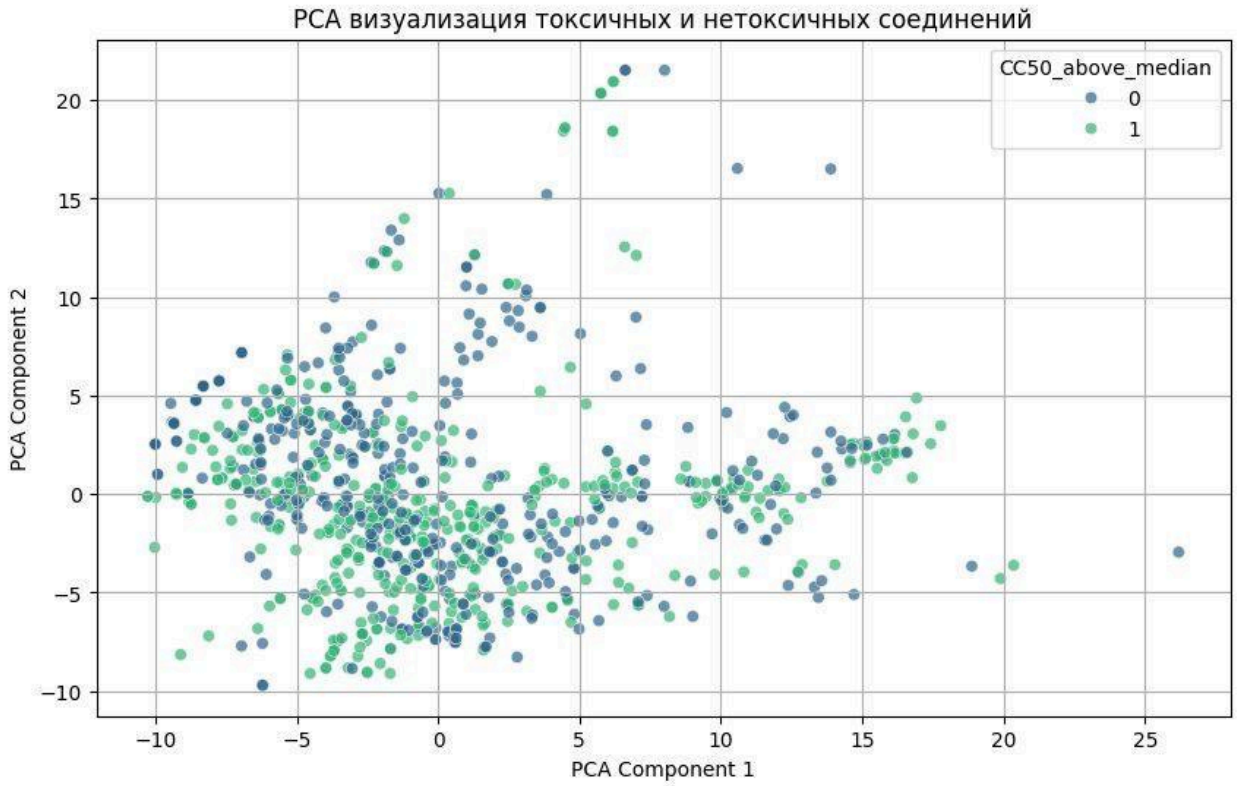


График PCA показал, что классы токсичности пересекаются, хотя в некоторых областях (например, в правом верхнем углу) преобладает один из классов. Это указывает на то, что модель может улавливать закономерности, но полностью разделить классы невозможно из-за частичного пересечения признаков.

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.

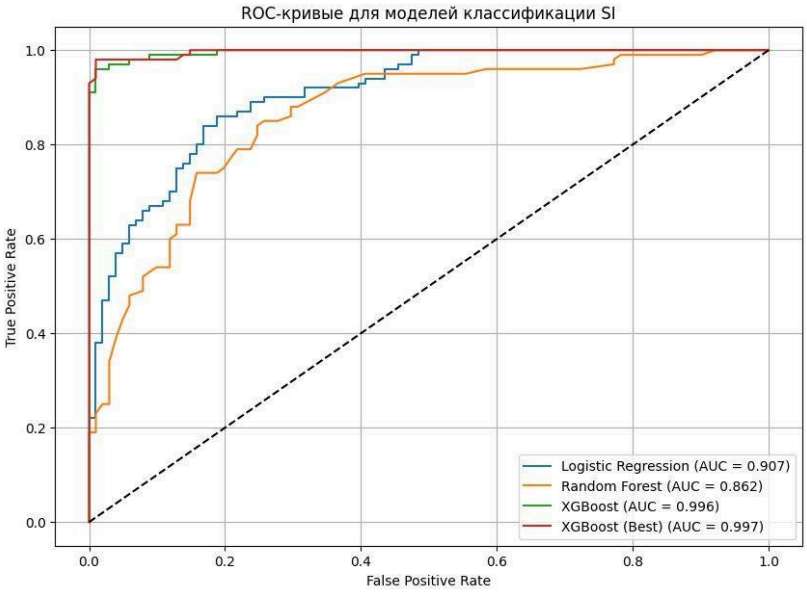
Этот анализ демонстрирует, что машинное обучение способно эффективно помогать в отборе менее токсичных молекул, ускоряя разработку новых лекарств.

Классификация SI > медианы (Selectivity Index)

В этой части проекта мы решаем задачу классификации для показателя SI (Selectivity Index) - важного фармакологического параметра, который показывает, насколько соединение эффективно подавляет вирус, не повреждая здоровые клетки. Наша цель — определить, превышает ли SI медианное значение, чтобы отобрать лучшие кандидаты для дальнейших исследований.

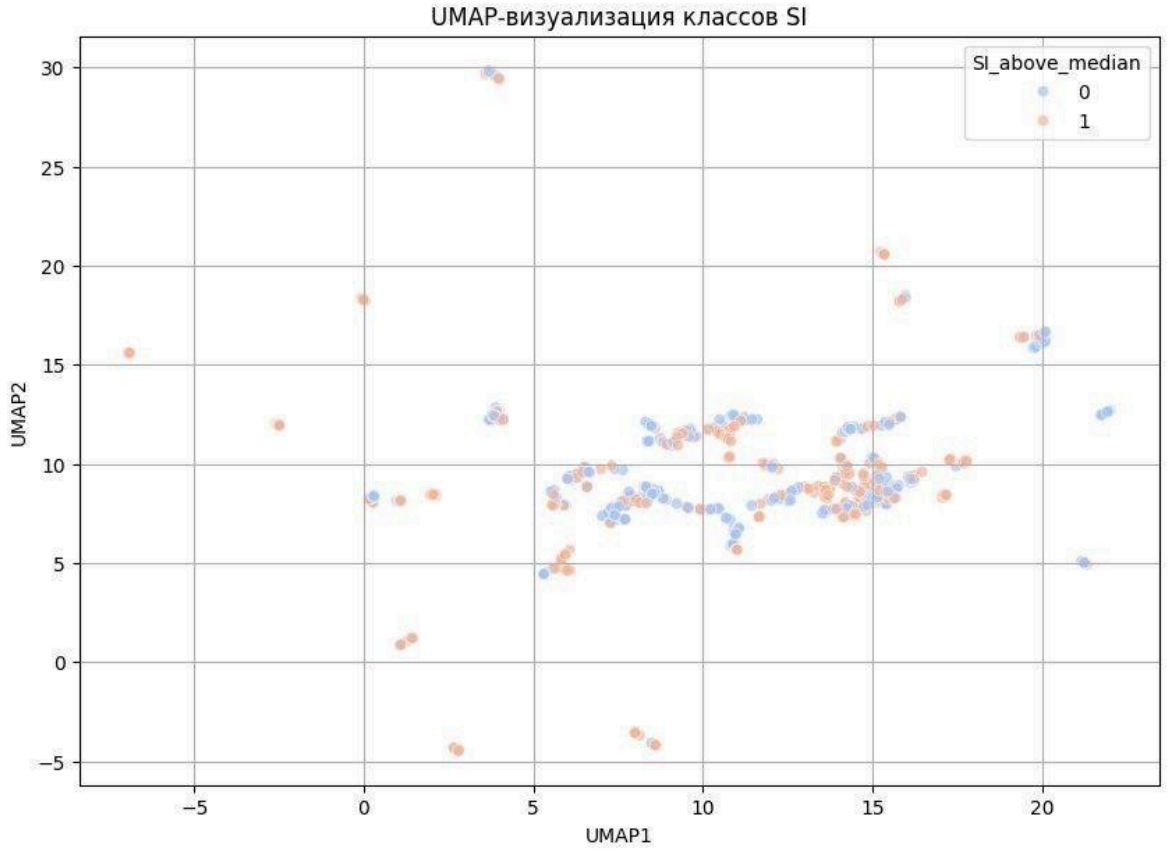
Начали с проверки распределения классов, которое оказалось сбалансированным: 501 молекула в классе SI ≤ медиана и 500 молекул - SI > медиана. После импутации пропущенных значений медианой и масштабирования данных перешли к обучению моделей: Logistic Regression, Random Forest, XGBoost. Для XGBoost дополнительно провели подбор гиперпараметров с помощью GridSearchCV.

Результаты показали, что Logistic Regression продемонстрировала хорошие показатели (Accuracy = 0.8159, F1-score = 0.8083, ROC-AUC = 0.9072), подтверждая свою надежность и интерпретируемость. Random Forest показал немного ниже результаты (Accuracy = 0.7662, F1-score = 0.7432, ROC-AUC = 0.8616). XGBoost продемонстрировал выдающиеся результаты (Accuracy = 0.9652, F1-score = 0.9641, ROC-AUC = 0.9958), а после настройки гипер параметров показатели стали ещё лучше (ROC-AUC = 0.9967), что говорит о высокой способности модели различать классы.



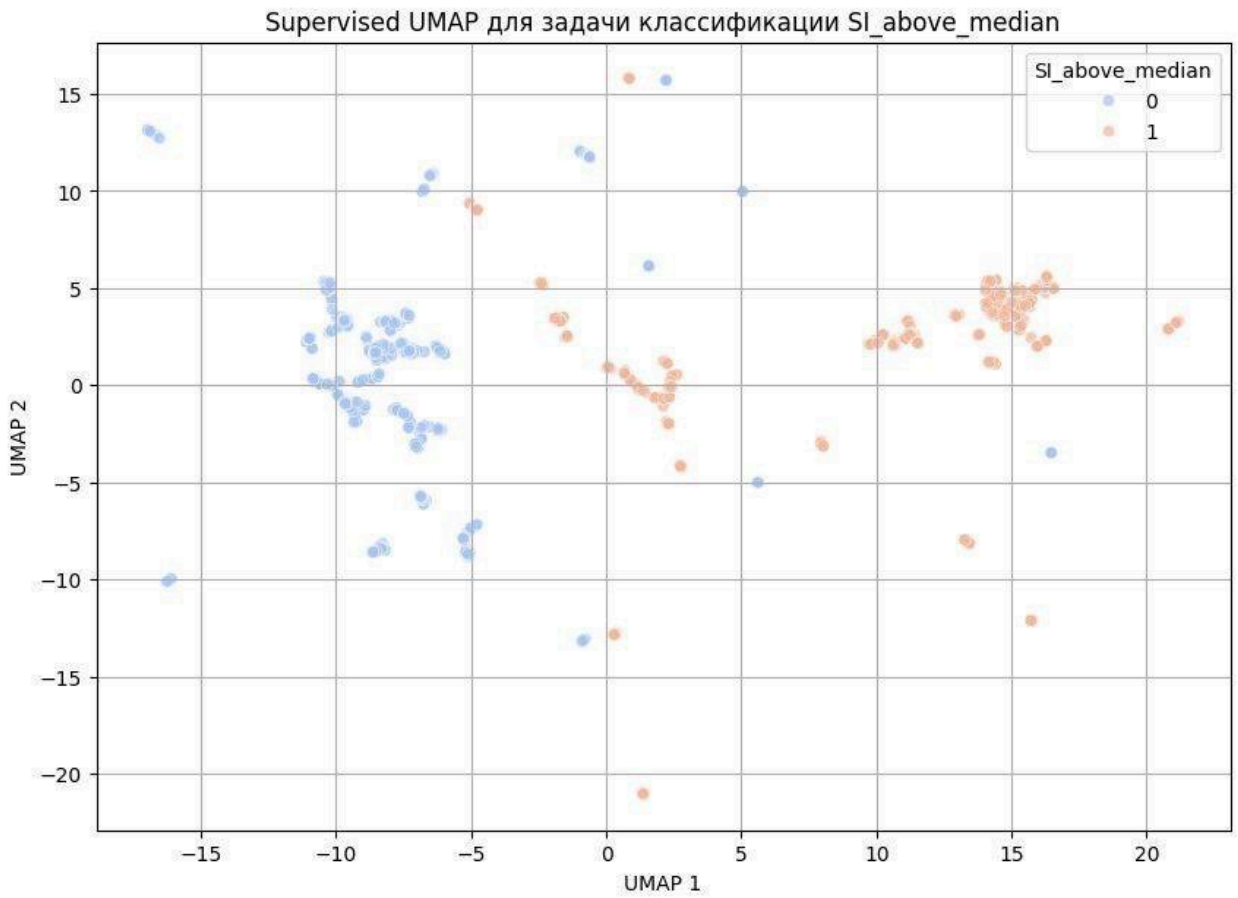
ROC-кривые для всех моделей показали, что XGBoost и его оптимизированная версия превосходят остальные модели, обеспечивая лучшую способность различать классы. Logistic Regression также показал хорошие результаты, а Random Forest уступает по всем метрикам.

Визуализация данных через PCA и UMAP подтвердила, что классы частично пересекаются в пространстве признаков, хотя локальные кластеры всё же видны. Supervised UMAP показал наиболее отчетливое разделение классов, демонстрируя, что модель действительно научилась выделять молекулы с высокой селективностью.



SHAP-анализ выявил ключевые признаки для предсказания SI: логарифмированные значения IC50 и CC50, а также MinEStateIndex, Kappa2, Chi3n, NumHAcceptors, BCUT2D_MRHI, EState_VSA8. Эти дескрипторы связаны с электронной плотностью, топологической структурой и физико-химическими характеристиками молекулы. Это важно для химиков, так как помогает понять, на какие свойства стоит обращать внимание при разработке новых соединений.

Stacking Classifier, объединяющий Logistic Regression, LightGBM и XGBoost, продемонстрировал отличные результаты, почти без ошибок определяя класс SI_above_median. Это показывает, что ансамблевые методы позволяют объединить сильные стороны базовых моделей и добиться высокой точности.

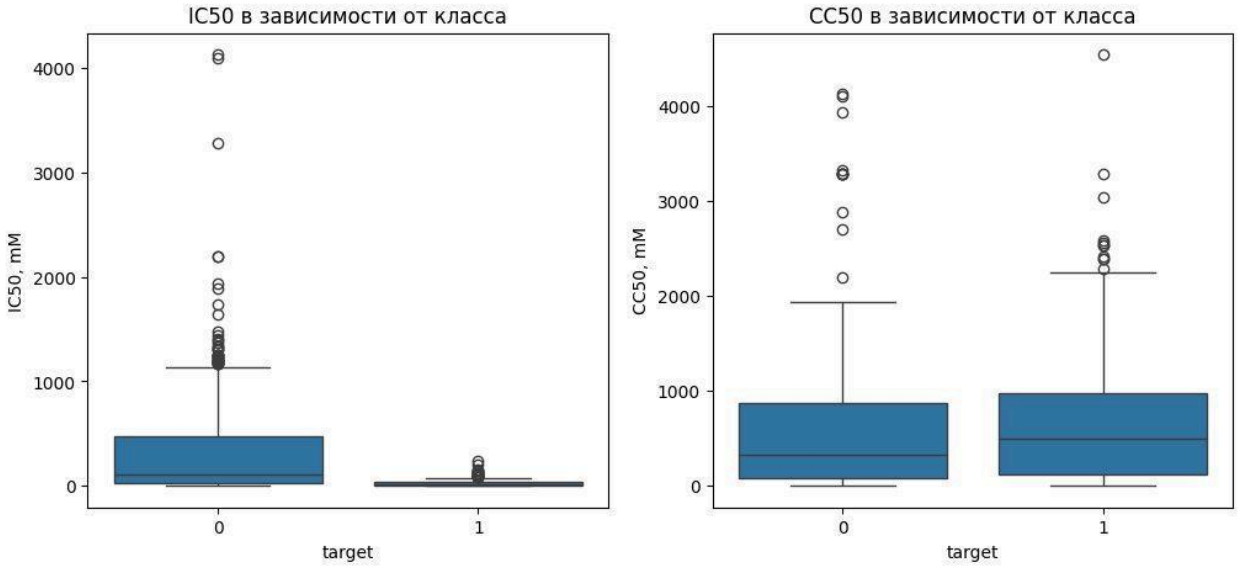


Таким образом, задача классификации SI_above_median решается успешно, а добавление полиномиальных и взаимодействующих признаков позволяет моделям уловить более сложные закономерности. Такой анализ помогает химикам ещё на этапе компьютерного моделирования отбирать соединения с высокой селективностью и снижать ресурсы на лабораторные исследования.

Классификация SI > 8 (Selectivity Index)

Мы занимаемся задачей классификации для показателя SI (Selectivity Index), который показывает, насколько соединение влияет на вирус, не нанося вреда клеткам. Чем выше значение SI, тем безопаснее такое соединение.

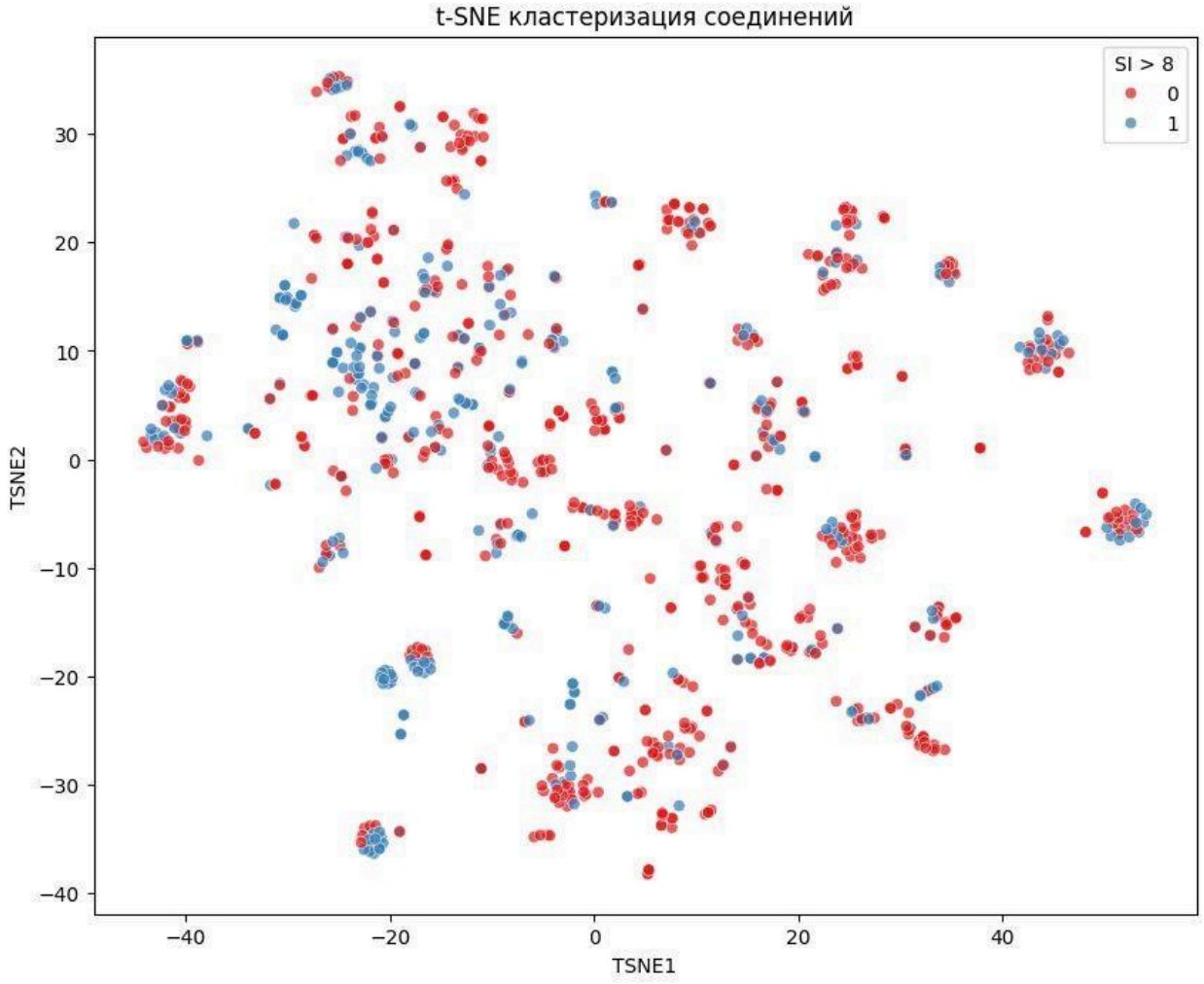
Для отборов мы используем порог SI > 8, чтобы найти перспективные соединения для дальнейших исследований. В процессе анализа мы загрузили данные, рассчитали SI, создали бинарную метку и проверили пропуски, распределения и соотношение классов, а также построили основные визуализации. Это помогло лучше понять данные и подготовить их для обучения моделей. При разборе IC50 оказалось, что значения IC50 для класса target = 1 (SI > 8) в среднем значительно ниже, что ожидаемо: высокая эффективность ингибирования и низкая токсичность дают высокий SI. Распределение CC50 между классами оказалось похожим, но вместе с IC50 оно помогает предсказать SI.



На тепловой карте корреляций мы заметили слабые связи между признаками. Признаки PEOE_VSA и Estate_VSA образуют кластеры с сильной корреляцией, показывая схожесть в распределении зарядов и свойствах молекулы. Слабая корреляция между признаками и целевой меткой указывает на сложность задачи классификации SI > 8.

Мы протестировали несколько моделей:
Logistic Regression, Random Forest и XGBoost. Logistic Regression показала хорошие результаты (точность 94%, ROC-AUC примерно 0.99), что говорит о её способности различать классы, даже будучи простой моделью. Random Forest работала хуже (точность 78%, ROC-AUC около 0.87), вероятно, из-за дисбаланса классов и необходимости корректировки гиперпараметров. XGBoost превзошел всех (точность 96%, ROC-AUC около 0.99), доказав свою устойчивость к сложным данным. Для анализа важности признаков мы использовали Random Forest и XGBoost.

В обеих моделях наибольшее значение имели log_IC50 и log_CC50, что подтверждает правильность выбора задачи - ведь SI зависит от этих значений. Среди других значимых признаков были дескрипторы, связанные с электронной плотностью (PEOE_VSA, Estate_VSA), топологией (Chi, Kappa) и энергетическими характеристиками (MaxPartialCharge). Это показывает, что молекулярная структура и распределение зарядов играют важную роль в предсказании SI.



Чтобы улучшить модель, мы добавили полиномиальные и взаимодействующие признаки, расширив данные до (1001, 22578). Эти признаки помогают учесть нелинейные зависимости и взаимодействия, что важно для фармакологических данных. На графиках UMAP видно, что классы начали образовывать локальные группы, хотя всё ещё перекрываются - это подтверждает сложность задачи и необходимость использования таких моделей, как XGBoost. После добавления взаимодействий и масштабирования, модель XGBoost показала отличные результаты: точность 95%, высокие precision и recall для обоих классов и ROC-AUC около 0.991.

Это говорит о том, что современные подходы машинного обучения и отбор признаков значительно улучшают способность модели различать соединения с высокой и низкой селективностью.

Для химиков это значит, что при разработке новых молекул стоит учитывать не только IC50 и CC50, но и характеристики, такие как электронная плотность, топология и взаимодействия признаков. Это поможет создать более селективные и безопасные соединения для дальнейших исследований.

Заключение

В этой курсовой работе мы провели полный анализ данных и создали модели машинного обучения, чтобы предсказать три важных фармакологических показателя: IC50, CC50 и SI. На этапе предварительного анализа данных мы обнаружили выбросы и использовали логарифмирование, что помогло улучшить распределение и повысить качество моделей. Построили регрессионные модели для IC50 и CC50, и они показали неплохие результаты. А вот задача с SI оказалась сложнее - даже ансамблевые модели не смогли достичь высоких значений R^2, что говорит о том, что нужно еще поработать в этом направлении. В задаче классификации показателей SI, IC50 и CC50 модели XGBoost и Stacking показали отличные результаты, что подтвердило их пользу в биоинформатике. Анализ важности признаков показал, что молекулярные дескрипторы, связанные с электронной плотностью, топологией и структурой, играют ключевую роль в предсказании активности и токсичности соединений. Эти выводы подчеркивают, как машинное обучение может быть полезным на ранних этапах разработки лекарств, что помогает экономить ресурсы и ускорить поиск подходящих кандидатов.