Курсовая работа по дисциплине “Классическое машинное обучение”

Разработал: Барышев В.В.

НИЯУ МИФИ 2025

Структура работы

1. Актуальность темы;
2. Объект и предмет исследования;
3. Литературный обзор;
4. Выводы по результатам литературного обзора;
5. Исследовательский анализ данных
6. Формулировака гипотезы;
7. Цели и задачи диссертационного исследования;
8. Методы и средства;
9. Проверка гипотезы;
10. Практическое значение;
11. Библиограафия

Актуальность

Процесс создания нового лекарственного препарата является сложным и многоэтапным. Необходимо определить его химическую формулу, синтезировать соединение, провести первичные биологические испытания и организовать тестирование. Все эти этапы требуют значительного времени, однако современные методы машинного обучения способны существенно ускорить данный процесс.

Так, например, с помощью различных моделей можно спрогнозировать эффективность соединений и подобрать наиболее подходящие сочетания параметров для разработки лекарственных средств.

Тем не менее, для достижения качественного результата в подобного рода задачах важно наладить эффективное взаимодействие между химиками и специалистами по машинному обучению, что зачастую оказывается непростой задачей.

Представим следующую ситуацию: химиками были предоставлены конфиденциальные данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность, обозначаются как IC50, CC50 и SI.
Данные доступны .

Обратите внимание, что значение SI рассчитывается на основе параметров IC50 и CC50. Подробную информацию об этих показателях можно найти в открытых источниках для более глубокого понимания контекста задачи.
Все остальные представленные признаки являются числовыми характеристиками химических соединений.

Объект и предмет

Предмет исследования – молекулярные соединения, обладающие биологической активностью (например, ингибиторы ферментов, лекарственные вещества).
Границы исследования – количественные взаимосвязи между структурными, электронными и физико-химическими характеристиками молекул (дескрипторами) и их биологической активностью (IC_{50}, CC_{50}, SI) в рамках методов QSAR (Quantitative Structure-Activity Relationship).

Описание дескрипторов (из курса хемоинформатики и QSAR)

– Дескрипторы – числовые или категориальные параметры, описывающие молекулярные свойства.
В QSAR они используются для построения моделей, предсказывающих активность соединений. Рисунок 1.

1. Целевые и контрольные переменны

IC50 (мМ) – концентрация ингибитора, подавляющая 50% активности фермента (мера эффективности);

CC50 (мМ) – концентрация, вызывающая 50% токсичности (мера цитотоксичности);

SI (Selectivity Index) = CC50 / IC50 – чем выше, тем лучше (высокая активность + низкая токсичность);

2. Дескрипторы в QSAR позволяют численно описать молекулярные свойства и связать их с биологической активностью. Выбор дескрипторов зависит от задачи:

Топологические – для анализа структуры;

Электронные – для реакционной способности;

Фрагментные – для выявления ключевых функциональных групп;

BCUT и VSA – для скрининга и ADME-прогнозирования;

Оптимальный набор дескрипторов повышает точность QSAR-моделей

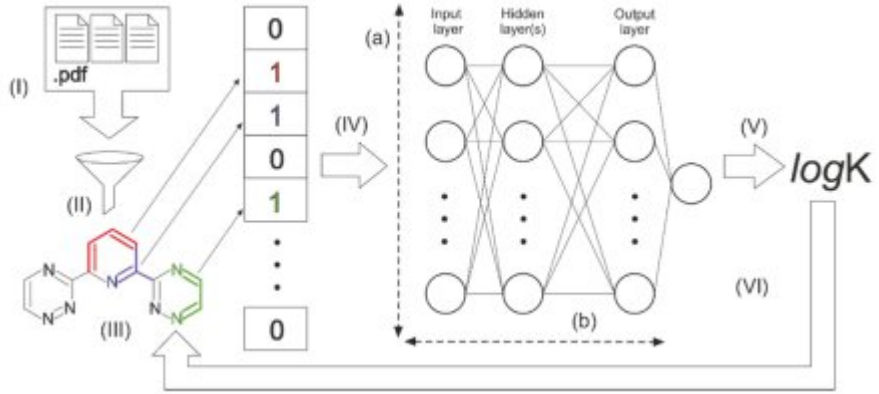


Рисунок 1. Общая схема. Источник

Литературный обзор

выводы из литературного обзора

Рассмотрены работы:

1. Курс лекци: “Искусственный интеллект в химии и материаловедении” https://teach-in.ru/course/ai-in-chemistry-and-materials-science/material

Работы исследовательской группы, в которую входят А. А. Митрофанов, Е. В. Матазова, Б. В. Егорова и соавт., охватывают два на первый взгляд разнесённых направления: (i) разработку макроциклических хелаторов для терапевтических радиоизотопов висмута-212/213 и (ii) применение методов машинного обучения (ML) для описания и прогнозирования свойств гибридных кристаллических материалов. Несмотря на тематические различия, оба направления решают общую задачу ускоренного дизайна функциональных соединений, опираясь на рациональный синтез, вычислительную химию и экспериментальную валидацию.

2. E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin, A. B. Tarasov
“Database of 2D Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps and Atomic Partial Charges Predicted by Machine Learning.” Chemistry of Materials, 2020.

Краткое описание
Цель работы — создать открытую, легко-используемую базу данных двухмерных органо-неорганических перовскитов (Ruddlesden–Popper и Dion–Jacobson типов), в которой каждая структура снабжена тремя ключевыми характеристиками:
1) оптимизированная кристаллическая геометрия,
2) ширина запрещённой зоны (band gap),
3) набор атомных частичных зарядов.

Выводы

из литературного обзора

Из обзора статей можно сделать следующие выводы:

1. Созданы две перспективные архитектуры макроциклических лигандов (benzoaza- и пиридин-азакрауны), обеспечивающие быстрый, высокоэффективный и биосовместимый перенос терапевтических изотопов 212/213Bi. Они превосходят классические DOTA-производные по скорости метки и устойчивости в сыворотке, что критично для клинической ТАТ.

2. Разработаны и открыто опубликованы ML-инструменты (база 2D-перовскитов; переносимые частичные заряды), позволяющие на порядки ускорить первичный скрининг и моделирование гибридных материалов. Это облегчает поиск соединений с заданными оптоэлектронными и сорбционными характеристиками.

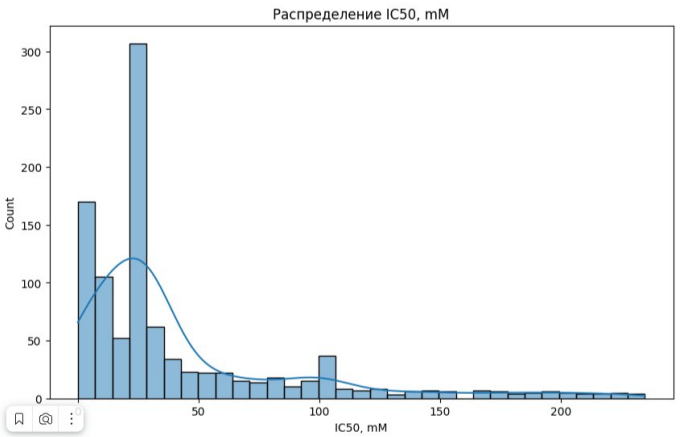
3. Работы демонстрируют сильный методический «транслейт» — от синтетической химии и радиохимии до материаловедения и искусственного интеллекта. Общая стратегия: сочетание целевого синтеза, экспериментальной валидации и машинного обучения.

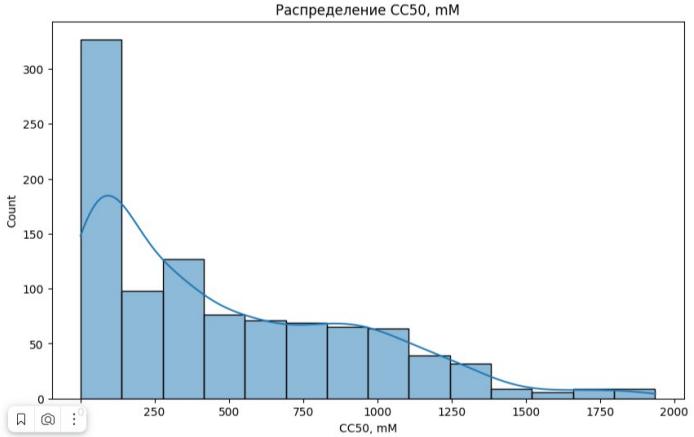
4. Перспективы дальнейших исследований:
   • интеграция ML-дизайна в разработку новых хелаторов (предсказание констант комплексообразования, ин vitro стабильности);
   • расширение ML-зарядов на биомолекулярные системы, что упростит гибридное моделирование «хелатор-радиоизотоп-белок»;
   • создание единой платформы данных, объединяющей радиометрические, структурные и квантово-химические сведения для мультизадачной оптимизации лекарственных и функциональных материалов.

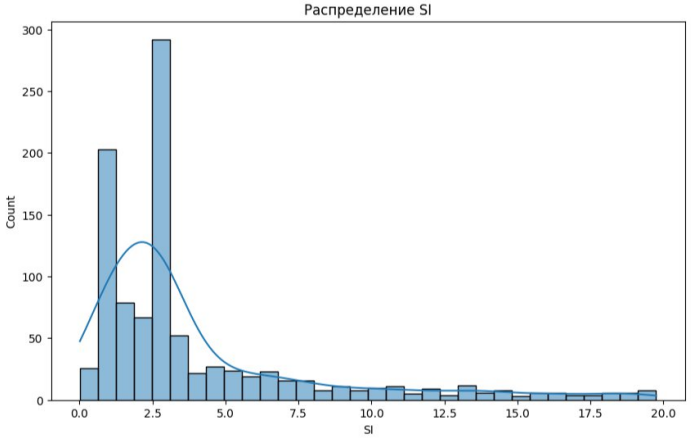
Исследовательский анализ

целевые признаки

Из обзора статей можно сделать следующие выводы:
Частота распределения





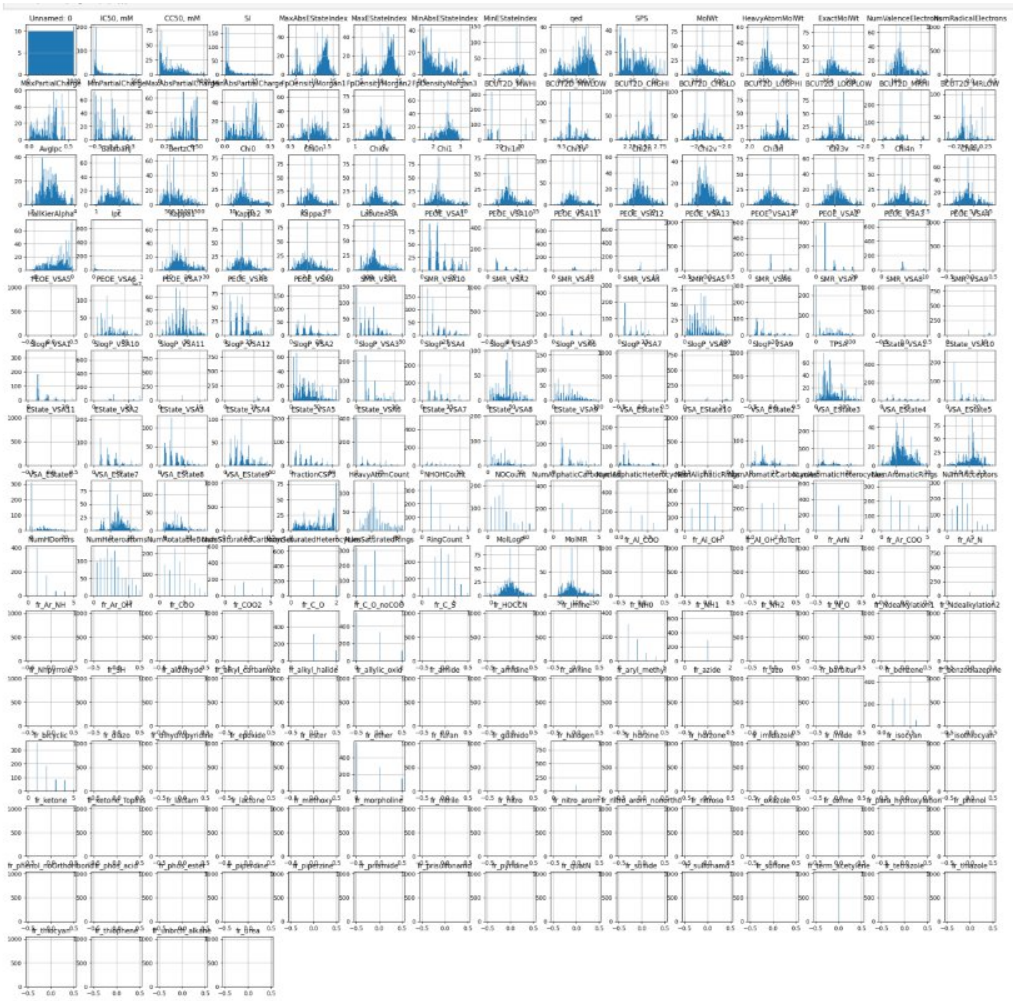


Выводы: визуальновидно видно, что распределения не имеют нормального распределения

Исследовательский анализ

из литературного обзора

Из обзора статей можно сделать следующие выводы:
Частота распределения



Выводы: визуальновидно видно

Исследовательский анализ

целевые признаки

ГИПОТЕЗА

исследования

Основная гипотеза:

С помощью методов QSAR-моделирования и современных алгоритмов машинного обучения (ML) можно построить статистически значимые модели, которые на основе молекулярных дескрипторов предсказывают противогриппозную активность соединений (IC50), цитотоксичность (CC50) и селективность (SI) с точностью, не уступающей экспериментальной погрешности in-vitro тестов

H0

Соотношение SI = CC50/IC50 не поддаётся адекватному прогнозу из дескрипторов лучше, чем случайная модель

H1

ML-модели способны воспроизводить SI с прескажущей способностью, превышающей случайный уровень (например, $R^2 > 0.5$ или MAE ниже выбранного порога)

Цель и задачи

НИЯУ МИФИ

Цель исследования

Разработка и валидация QSAR-моделей на основе методов машинного обучения (ML), способных предсказывать противогриппозную активность (IC_{50}), цитотоксичность (CC_{50}) и селективность (SI) химических соединений с точностью, сопоставимой с экспериментальными данными.

Для достижения этой цели необходимо определить приоритеты развития рекреационных зон и выделить наиболее перспективные территории для формирования новой рекреационной инфраструктуры. Также важно разработать и апробировать цифровые системы поддержки принятия решений.

I

Анализ и предобработка данных

- Проверка данных на полноту, аномалии и корреляции между признаками.
- Нормализация и стандартизация числовых дескрипторов.
- Исследование мультиколлинеарности и отбор наиболее значимых признаков.

II

Построение и обучение ML-моделей

- Тестирование различных алгоритмов (регрессия, ансамбли, нейросети) для предсказания IC_{50}, CC_{50} и SI.
- Оптимизация гиперпараметров с использованием кросс-валидации.
- Оценка важности молекулярных дескрипторов для интерпретируемости моделей.

III

Сравнение с нулевой гипотезой (H_0)

- Проверка, превосходят ли ML-модели случайное предсказание (например, по R^2, MAE, RMSE).
- Определение, можно ли предсказать $SI = CC_{50} / IC_{50}$ лучше, чем случайная модель.

IV

4. Интерпретация и валидация результатов

- Сравнение с литературными данными и экспериментальными погрешностями.
- Анализ химической значимости выявленных закономерностей.
- Проверка устойчивости моделей на новых данных (если доступны).

V

5. Практические рекомендации

- Разработка рекомендаций для химиков по оптимизации структуры соединений.
- Предложение методов интеграции ML в процесс разработки лекарств.

Ожидаемые результаты:

- Подтверждение или опровержение гипотезы H_1 (ML-модели превосходят случайное предсказание).
- Выявление ключевых молекулярных дескрипторов, влияющих на активность соединений.
- Практические рекомендации для ускорения скрининга новых противовирусных препаратов.

Научная новизна

ожидаемые результаты исследования

НИЯУ МИФИ

Научная новизна

1. Разработка гибридного подхода к прогнозированию биологической активности соединений

2. Верификация гипотезы о предсказуемости SI

3. Ожидаемые результаты

Если H1 верна:

• ML-модель сможет предсказывать SI с $R^2 > 0.5$.

• Удастся выявить ключевые дескрипторы, влияющие на активность.

• Сократится время скрининга новых соединений.

Если H0 не отвергается:

• Нужен пересмотр дескрипторов или сбор дополнительных данных.

МЕТОДЫ И СРЕДСТВА

Для достижения поставленных целей исследования будут использованы следующие методы и средства:

1. Модели регресии

2. Модели классификации

Регрессия IC50

методы и средства

Результаты регрессии:

Linear Regression:
Средняя MSE: 10922.6587, стандартное отклонение: 1832.4109
Средняя MAE: 71.4531, стандартное отклонение: 8.7184
Средний R^2: 0.0424

Random Forest:
Средняя MSE: 4337.8913, стандартное отклонение: 934.7052
Средняя MAE: 29.7003, стандартное отклонение: 2.2212
Средний R^2: 0.6094

SVR:
Средняя MSE: 13734.1913, стандартное отклонение: 3978.6531
Средняя MAE: 60.5300, стандартное отклонение: 12.4955
Средний R^2: -0.1728

KNN:
Средняя MSE: 11776.7719, стандартное отклонение: 1787.5656
Средняя MAE: 70.9781, стандартное отклонение: 6.1093
Средний R^2: -0.0418

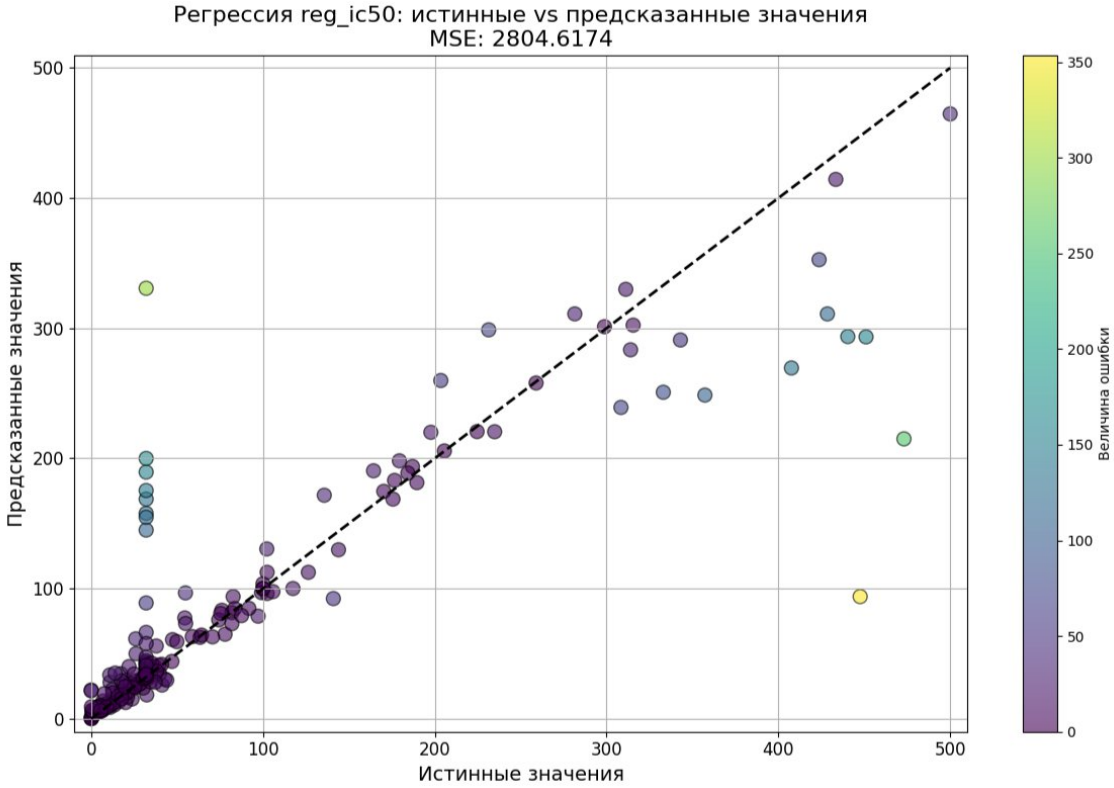
XGBoost:
Средняя MSE: 4344.4585, стандартное отклонение: 1425.9756
Средняя MAE: 31.6786, стандартное отклонение: 5.7581
Средний R^2: 0.6085

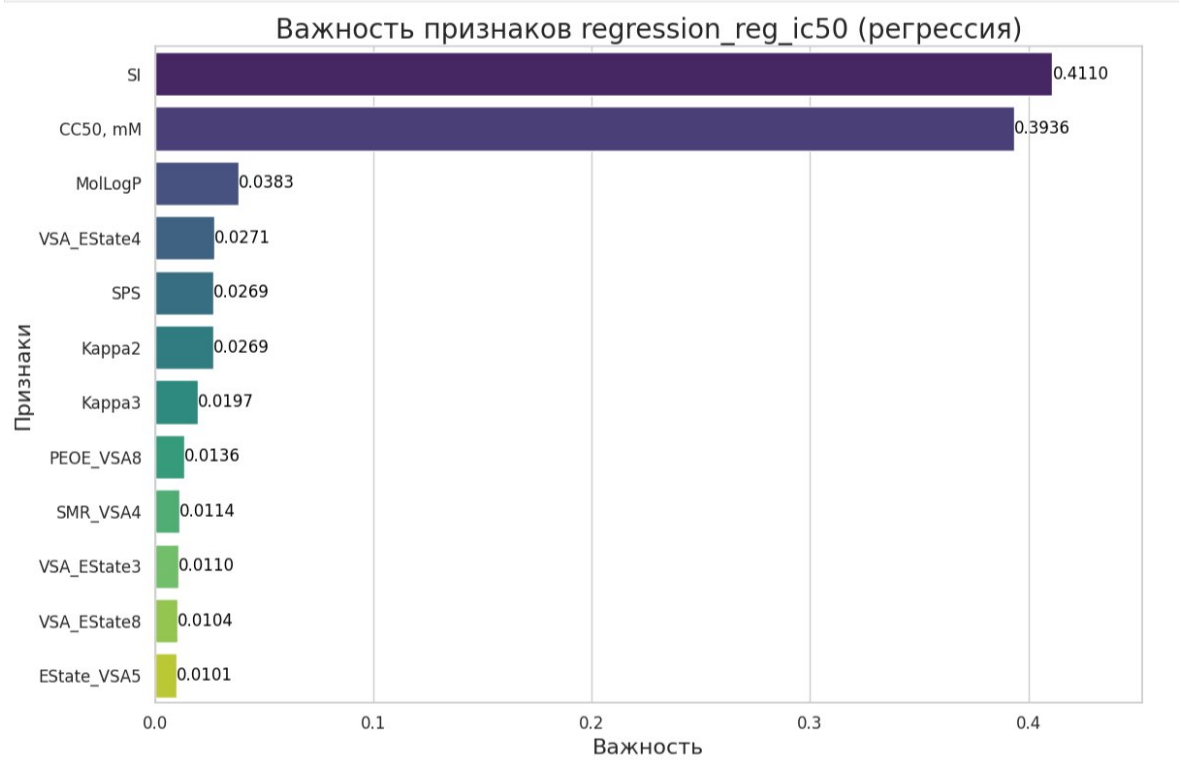
Вывод

На основании проведенного анализа можно сделать вывод, что модели Random Forest и XGBoost являются наиболее эффективными для данной задачи, обеспечивая низкие значения ошибок и высокие коэффициенты детерминации. В то время как Linear Regression, SVR и KNN показывают значительно худшие результаты.

Регрессия IC50

методы и средства





Регрессия IC50

методы и средства

Выводы о важности признаков

SI: Важность = 0.4110
CC50;
mM: Важность = 0.3936;
MolLogP: Важность = 0.0383;
VSA_EState4: Важность = 0.0271;
SPS: Важность = 0.0269;
Kappa2: Важность = 0.0269;
Kappa3: Важность = 0.0197;
PEOE_VSA8: Важность = 0.0136;
SMR_VSA4: Важность = 0.0114;
VSA_EState3: Важность = 0.0110;
VSA_EState8: Важность = 0.0104;
EState_VSA5: Важность = 0.0101;

Наиболее важный признак: SI с важностью 0.4110;
Наименее важный признак: EState_VSA5 с важностью 0.0101

Регрессия CC50

методы и средства

Результаты регрессии:

Linear Regression:
Средняя MSE: 179455.7250, стандартное отклонение: 27148.0167
Средняя MAE: 332.8122, стандартное отклонение: 13.5813
Средний R^2: 0.1742

Random Forest:
Средняя MSE: 66981.8069, стандартное отклонение: 25026.1832
Средняя MAE: 142.9900,
стандартное отклонение: 25.5236
Средний R^2: 0.6982

SVR:
Средняя MSE: 233817.6213, стандартное отклонение: 29583.0243
Средняя MAE: 364.1830, стандартное отклонение: 14.8665
Средний R^2: -0.0782

KNN:
Средняя MSE: 164587.9535, стандартное отклонение: 23806.3988
Средняя MAE: 287.9583, стандартное отклонение: 19.0182
Средний R^2: 0.2332

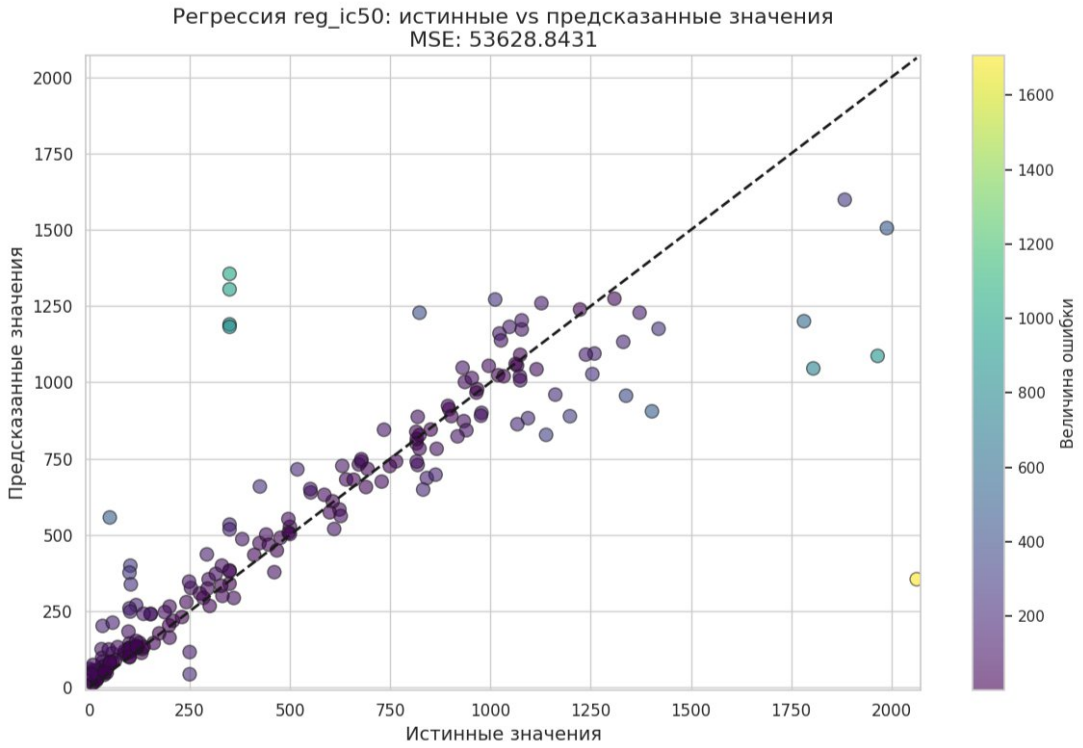
XGBoost:
Средняя MSE: 66556.3811, стандартное отклонение: 27009.1213
Средняя MAE: 133.3069, стандартное отклонение: 23.3576
Средний R^2: 0.7005

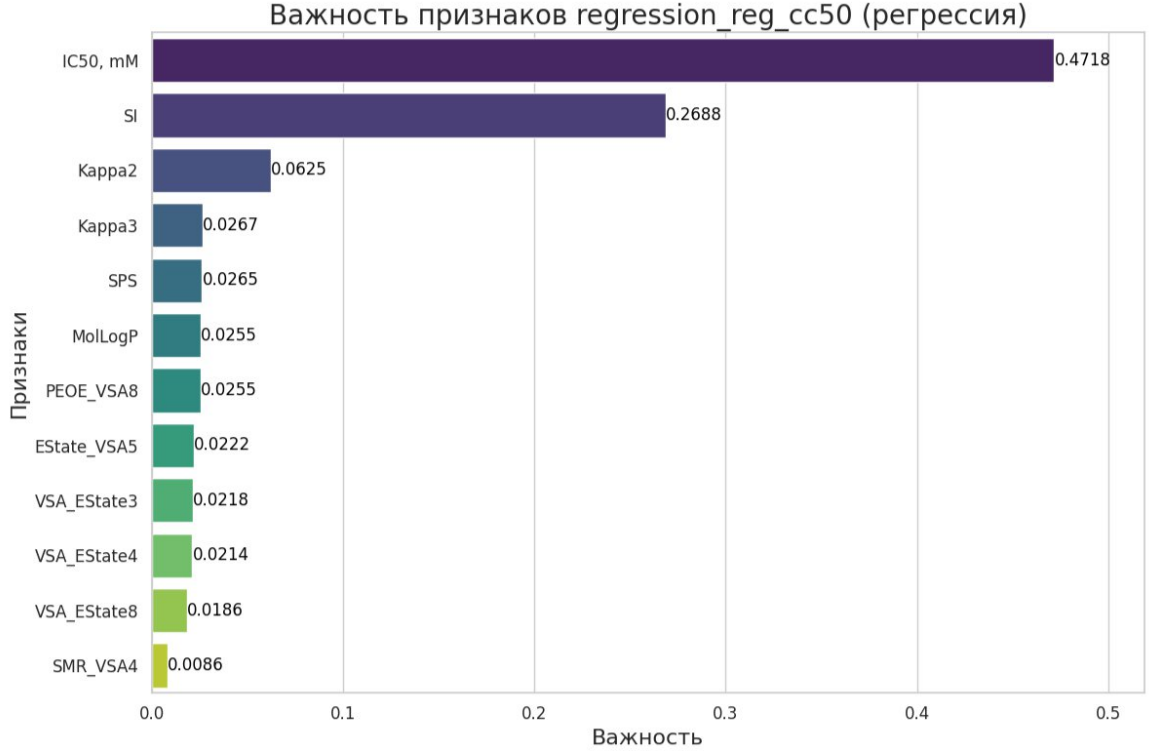
Вывод

На основании проведенного анализа можно сделать вывод, что модели Random Forest и XGBoost являются наиболее эффективными для данной задачи, обеспечивая низкие значения ошибок и высокие коэффициенты детерминации. В то время как Linear Regression, SVR и KNN показывают значительно худшие результаты.

Регрессия CC50

методы и средства





Регрессия IC50

методы и средства

Выводы о важности признаков

IC50, mM: Важность = 0.4718
SI: Важность = 0.2688
Kappa2: Важность = 0.0625
Kappa3: Важность = 0.0267
SPS: Важность = 0.0265
MolLogP: Важность = 0.0255
PEOE_VSA8: Важность = 0.0255
EState_VSA5: Важность = 0.0222
VSA_EState3: Важность = 0.0218
VSA_EState4: Важность = 0.0214
VSA_EState8: Важность = 0.0186
SMR_VSA4: Важность = 0.0086

Наиболее важный признак: IC50, mM с важностью 0.4718
Наименее важный признак: SMR_VSA4 с важностью 0.0086

Регрессия SI50

методы и средства

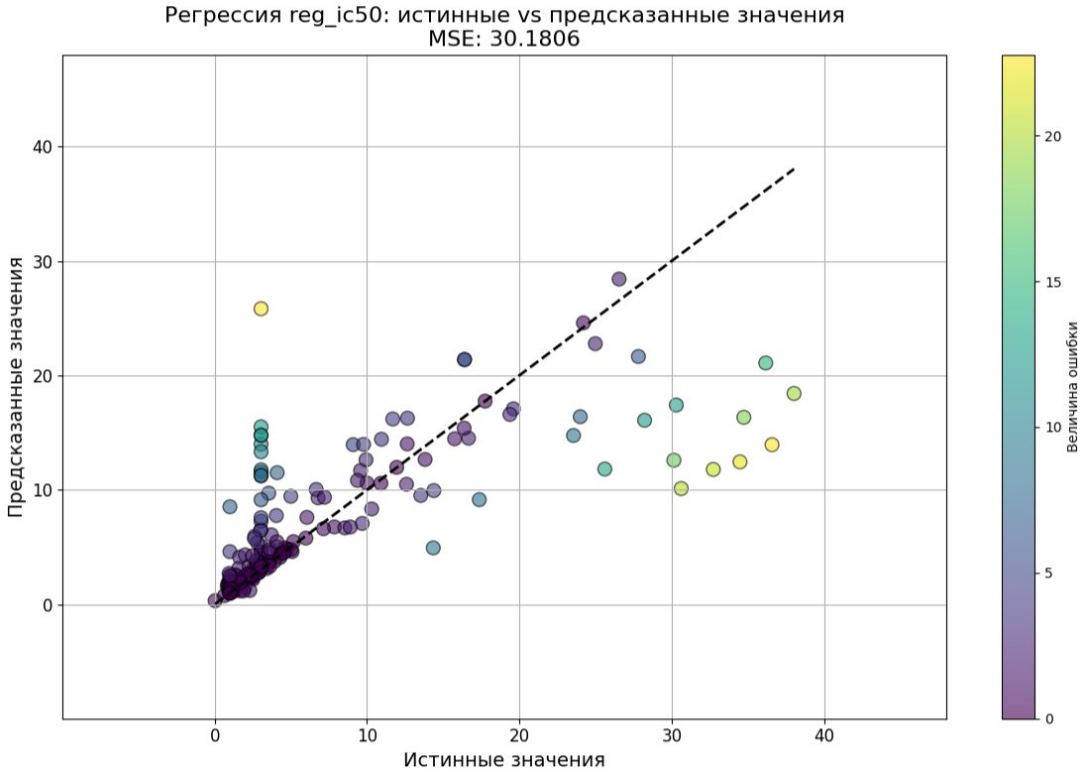
Результаты регрессии:
Linear Regression:
Средняя MSE: 66.8991, стандартное отклонение: 20.3326
Средняя MAE: 5.8754, стандартное отклонение: 0.7600
Средний R^2: 0.0433
Random Forest:
Средняя MSE: 29.6829, стандартное отклонение: 7.1324
Средняя MAE: 3.0058, стандартное отклонение: 0.3914
Средний R^2: 0.5653
SVR:
Средняя MSE: 80.8694, стандартное отклонение: 29.1552
Средняя MAE: 4.7996, стандартное отклонение: 0.9534
Средний R^2: -0.1368
KNN:
Средняя MSE: 76.3261, стандартное отклонение: 24.3282
Средняя MAE: 5.9892, стандартное отклонение: 0.9785
Средний R^2: -0.0838
XGBoost:
Средняя MSE: 29.7081, стандартное отклонение: 10.1954
Средняя MAE: 2.8605, стандартное отклонение: 0.4914
Средний R^2: 0.5709

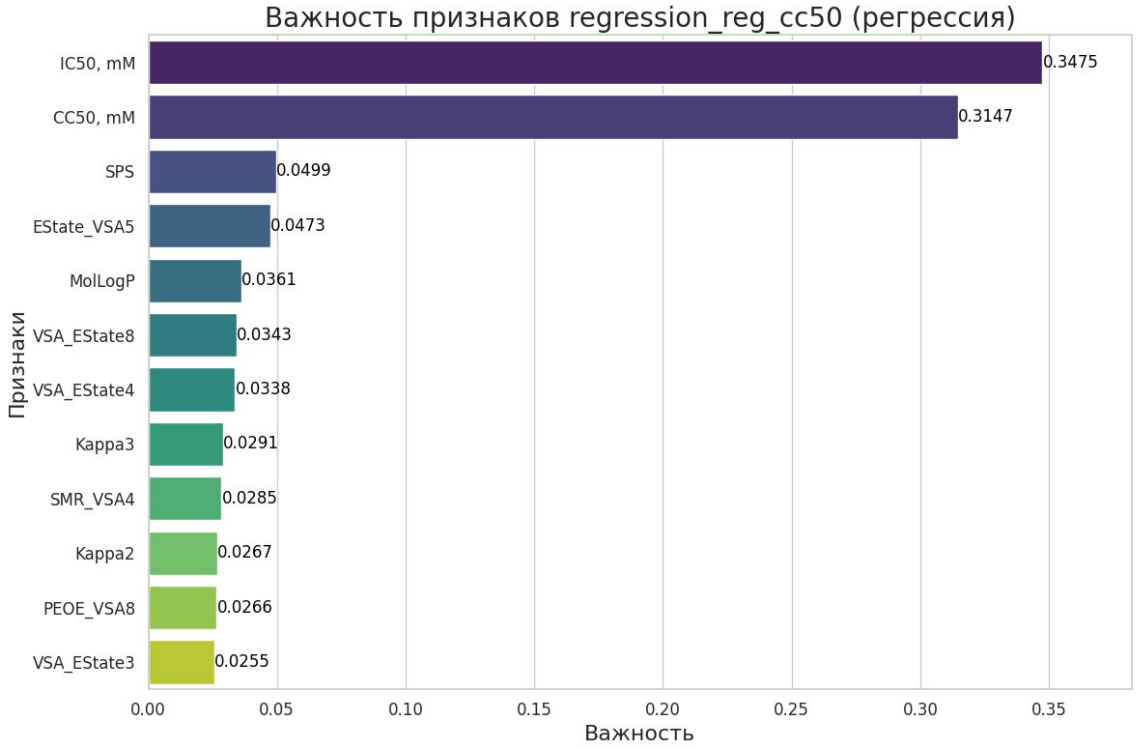
Вывод

Random Forest показывает значительно лучшие результаты, с средней MSE (29.6829) и MAE (3.0058), что указывает на более точные предсказания. Значение R^2 (0.5653) говорит о том, что эта модель способна объяснить 56.53% вариации в данных, что является заметным улучшением по сравнению с линейной регрессией.

Регрессия SI50

методы и средства





Регрессия SI50

методы и средства

Выводы о важности признаков

IC50, mM: Важность = 0.3475
CC50, mM: Важность = 0.3147
SPS: Важность = 0.0499
EState_VSA5: Важность = 0.0473
MolLogP: Важность = 0.0361
VSA_EState8: Важность = 0.0343
VSA_EState4: Важность = 0.0338
Kappa3: Важность = 0.0291
SMR_VSA4: Важность = 0.0285
Kappa2: Важность = 0.0267
PEOE_VSA8: Важность = 0.0266
VSA_EState3: Важность = 0.0255

Наиболее важный признак: IC50, mM с важностью 0.3475
Наименее важный признак: VSA_EState3 с важностью 0.0255

Классификация IC50

методы и средства

Результаты:

Logistic Regression:
Средняя точность: 0.5725, стандартное отклонение: 0.0295
Средняя F1-меры: 0.5579, стандартное отклонение: 0.0321
Средняя полнота: 0.5725, стандартное отклонение: 0.0295
Средняя точность (precision): 0.5602, стандартное отклонение: 0.0333

Random Forest:
Средняя точность: 0.8475, стандартное отклонение: 0.0408
Средняя F1-меры: 0.8470, стандартное отклонение: 0.0410
Средняя полнота: 0.8475, стандартное отклонение: 0.0408
Средняя точность (precision): 0.8478, стандартное отклонение: 0.0412

SVC:
Средняя точность: 0.6950, стандартное отклонение: 0.0165
Средняя F1-меры: 0.6916, стандартное отклонение: 0.0178
Средняя полнота: 0.6950, стандартное отклонение: 0.0165
Средняя точность (precision): 0.6927, стандартное отклонение: 0.0172

KNN:
Средняя точность: 0.6725, стандартное отклонение: 0.0380
Средняя F1-меры: 0.6720, стандартное отклонение: 0.0386
Средняя полнота: 0.6725, стандартное отклонение: 0.0380
Средняя точность (precision): 0.6719, стандартное отклонение: 0.0389

XGBoost:
Средняя точность: 0.9138, стандартное отклонение: 0.0278
Средняя F1-меры: 0.9138, стандартное отклонение: 0.0277
Средняя полнота: 0.9138, стандартное отклонение: 0.0278
Средняя точность (precision): 0.9144, стандартное отклонение: 0.0274

Вывод

- Лучшие результаты показала модель XGBoost, которая имеет самые высокие значения по всем метрикам (точность, F1-мера, полнота и precision).

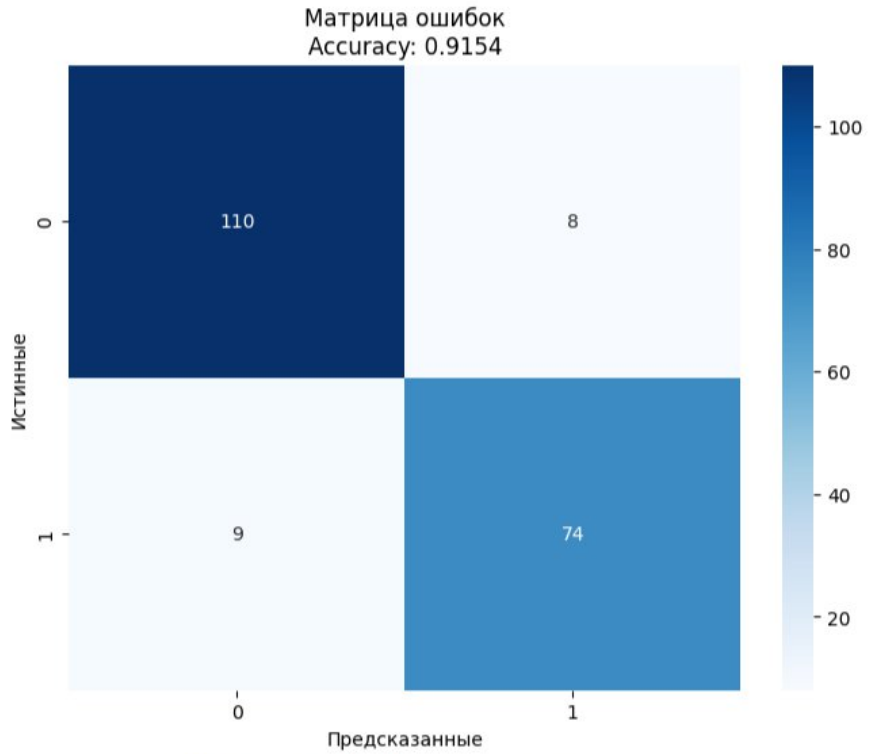
- Random Forest также показал хорошие результаты и находится на втором месте по всем метрикам.

- Logistic Regression и KNN имеют значительно более низкие показатели по сравнению с Random Forest и XGBoost.

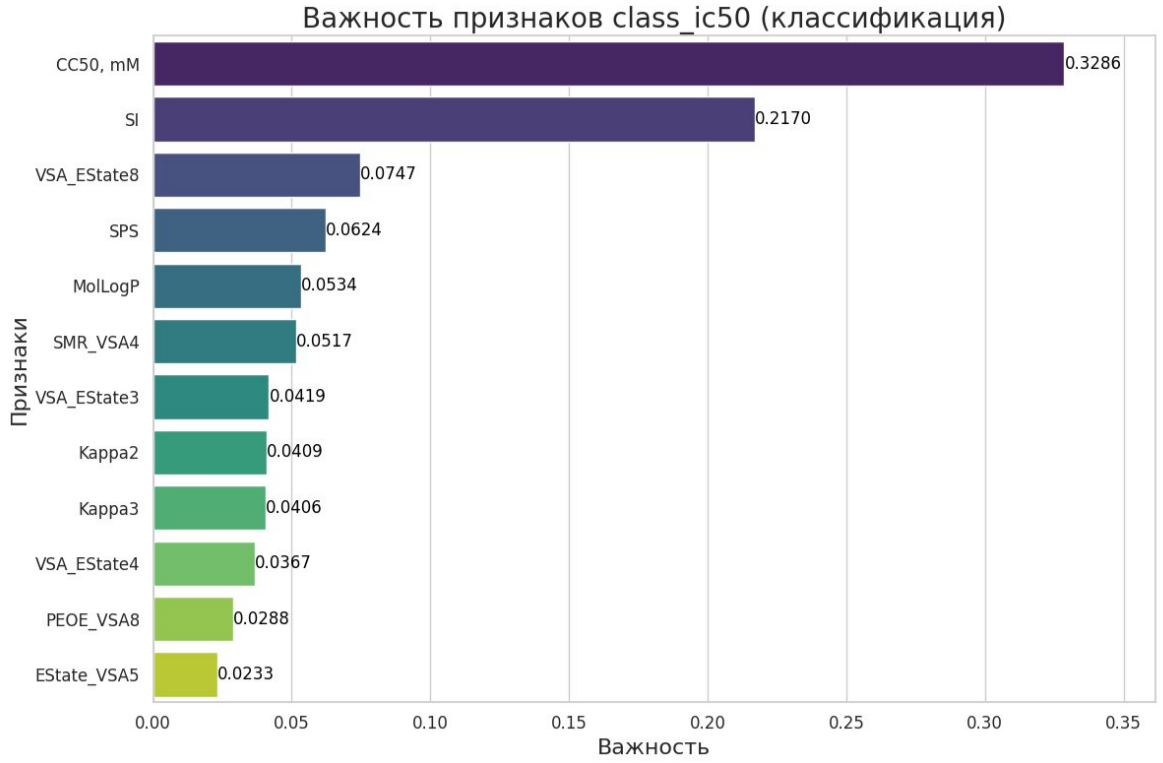
- Модель SVC демонстрирует средние результаты среди всех протестированных моделей.

Классификация IC50

методы и средства



	precision	recall	f1-score	support
0	0.92	0.93	0.93	118
1	0.90	0.89	0.90	83
accuracy			0.92	201
macro avg	0.91	0.91	0.91	201
weighted avg	0.92	0.92	0.92	201



Классификация IC50

методы и средства

Выводы о важности признаков

CC50, mM: Важность = 0.3286
SI: Важность = 0.2170
VSA_EState8: Важность = 0.0747
SPS: Важность = 0.0624
MolLogP: Важность = 0.0534
SMR_VSA4: Важность = 0.0517
VSA_EState3: Важность = 0.0419
Kappa2: Важность = 0.0409
Kappa3: Важность = 0.0406
VSA_EState4: Важность = 0.0367
PEOE_VSA8: Важность = 0.0288
EState_VSA5: Важность = 0.0233

Наиболее важный признак:
CC50, mM с важностью 0.3286

Наименее важный признак:
EState_VSA5 с важностью 0.0233

Классификация CC50

методы и средства

Результаты
Logistic Regression: Средняя точность: 0.6925, стандартное отклонение: 0.0390 Средняя F1-меры: 0.6909, стандартное отклонение: 0.0407 Средняя полнота: 0.6925, стандартное отклонение: 0.0390 Средняя точность (precision): 0.6928, стандартное отклонение: 0.0389

Random Forest: Средняя точность: 0.8287, стандартное отклонение: 0.0393 Средняя F1-меры: 0.8283, стандартное отклонение: 0.0395 Средняя полнота: 0.8287, стандартное отклонение: 0.0393 Средняя точность (precision): 0.8300, стандартное отклонение: 0.0397

SVC: Средняя точность: 0.7275, стандартное отклонение: 0.0380 Средняя F1-меры: 0.7260, стандартное отклонение: 0.0385 Средняя полнота: 0.7275, стандартное отклонение: 0.0380 Средняя точность (precision): 0.7291, стандартное отклонение: 0.0382

KNN: Средняя точность: 0.7112, стандартное отклонение: 0.0341 Средняя F1-меры: 0.7109, стандартное отклонение: 0.0341 Средняя полнота: 0.7112, стандартное отклонение: 0.0341 Средняя точность (precision): 0.7114, стандартное отклонение: 0.0340

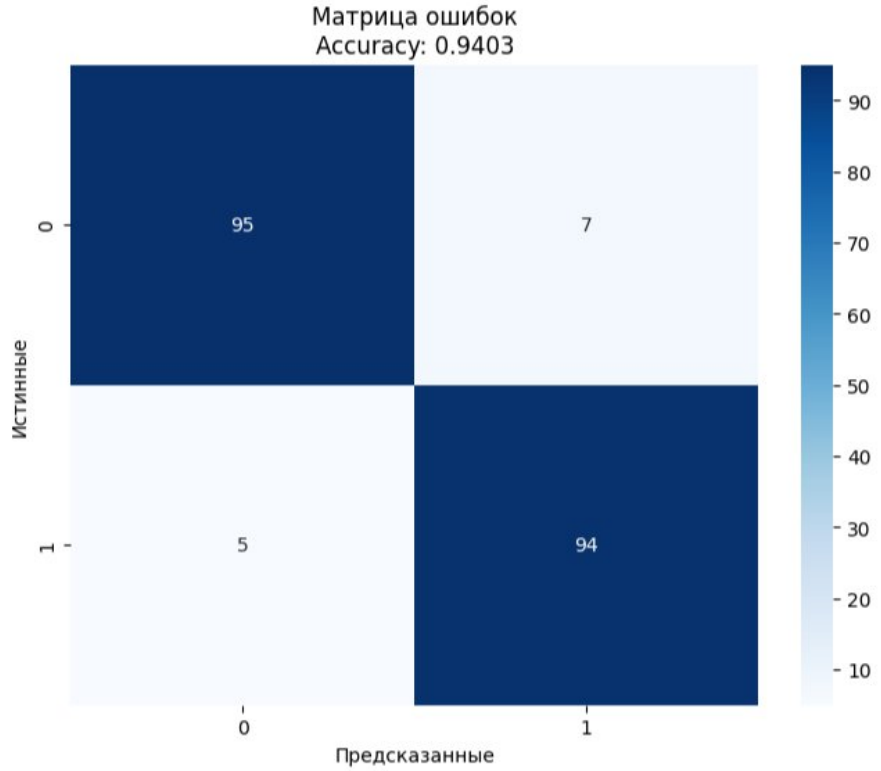
XGBoost: Средняя точность: 0.9213, стандартное отклонение: 0.0341 Средняя F1-меры: 0.9212, стандартное отклонение: 0.0341 Средняя полнота: 0.9213, стандартное отклонение: 0.0341 Средняя точность (precision): 0.9231, стандартное отклонение: 0.0326

Вывод

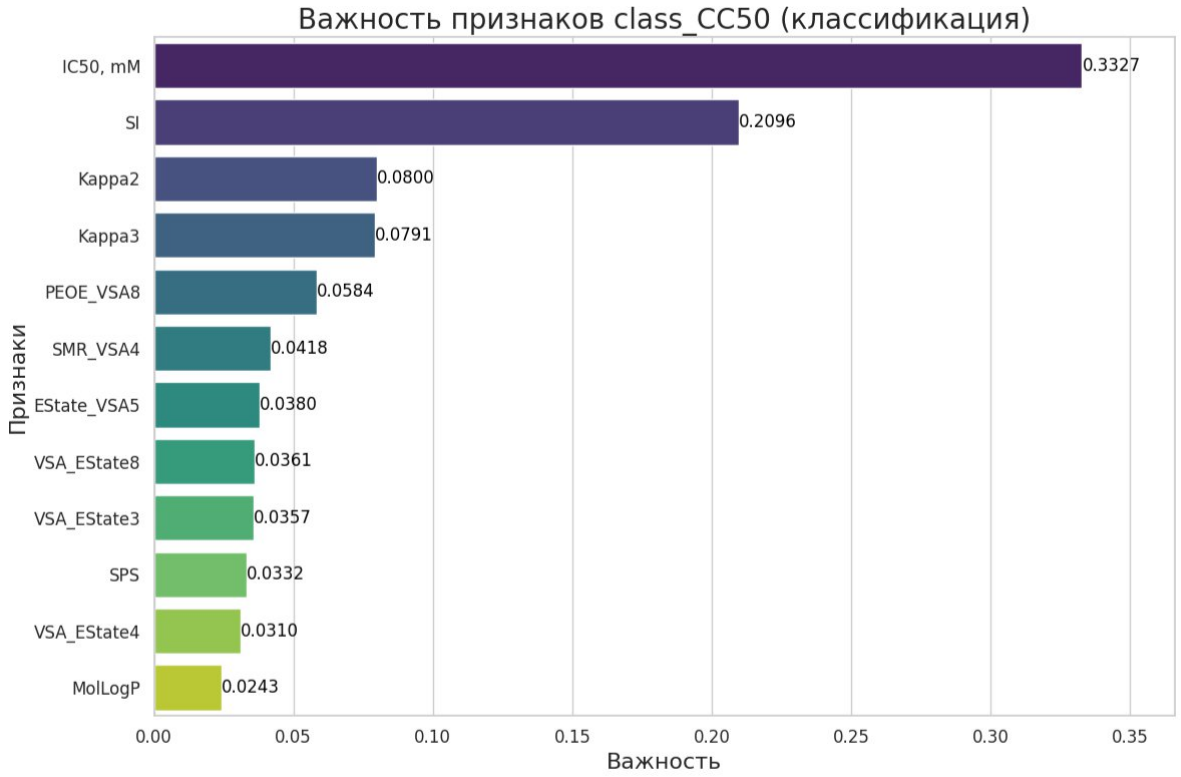
Наилучшие результаты продемонстрировал алгоритм XGBoost, показавший среднюю точность 0.9213 с низким стандартным отклонением (0.0341), что свидетельствует о высокой стабильности модели. Также XGBoost достиг максимальных значений по F1-мере, полноте (recall) и точности (precision)

Классификация CC50

методы и средства



	precision	recall	f1-score	support
0	0.95	0.93	0.94	102
1	0.93	0.95	0.94	99
accuracy			0.94	201
macro avg	0.94	0.94	0.94	201
weighted avg	0.94	0.94	0.94	201



Классификация CC50

методы и средства

Выводы о важности признаков

IC50, mM: Важность = 0.3327 SI: Важность = 0.2096;
Kappa2: Важность = 0.0800;
Kappa3: Важность = 0.0791;
PEOE_VSA8: Важность = 0.0584;
SMR_VSA4: Важность = 0.0418;
EState_VSA5: Важность = 0.0380;
VSA_EState8: Важность = 0.0361;
VSA_EState3: Важность = 0.0357;
SPS: Важность = 0.0332;
VSA_EState4: Важность = 0.0310;
MolLogP: Важность = 0.0243;

Наиболее важный признак: IC50, mM с важностью 0.3327;

Наименее важный признак: MolLogP с важностью 0.0243

Классификация SI50

методы и средства

Результаты:

Logistic Regression: Средняя точность: 0.5875, стандартное отклонение: 0.0259 Средняя F1-меры: 0.5793, стандартное отклонение: 0.0238 Средняя полнота: 0.5875, стандартное отклонение: 0.0259 Средняя точность (precision): 0.5835, стандартное отклонение: 0.0268

Random Forest: Средняя точность: 0.7850, стандартное отклонение: 0.0161 Средняя F1-меры: 0.7837, стандартное отклонение: 0.0148 Средняя полнота: 0.7850, стандартное отклонение: 0.0161 Средняя точность (precision): 0.7877, стандартное отклонение: 0.0187

SVC: Средняя точность: 0.6038, стандартное отклонение: 0.0320 Средняя F1-меры: 0.6023, стандартное отклонение: 0.0334 Средняя полнота: 0.6038, стандартное отклонение: 0.0320 Средняя точность (precision): 0.6059, стандартное отклонение: 0.0349

KNN: Средняя точность: 0.6137, стандартное отклонение: 0.0593 Средняя F1-меры: 0.6129, стандартное отклонение: 0.0596 Средняя полнота: 0.6137, стандартное отклонение: 0.0593 Средняя точность (precision): 0.6130, стандартное отклонение: 0.0601

XGBoost: Средняя точность: 0.8838, стандартное отклонение: 0.0135 Средняя F1-меры: 0.8834, стандартное отклонение: 0.0133 Средняя полнота: 0.8838, стандартное отклонение: 0.0135 Средняя точность (precision): 0.8872, стандартное отклонение: 0.0129

Вывод

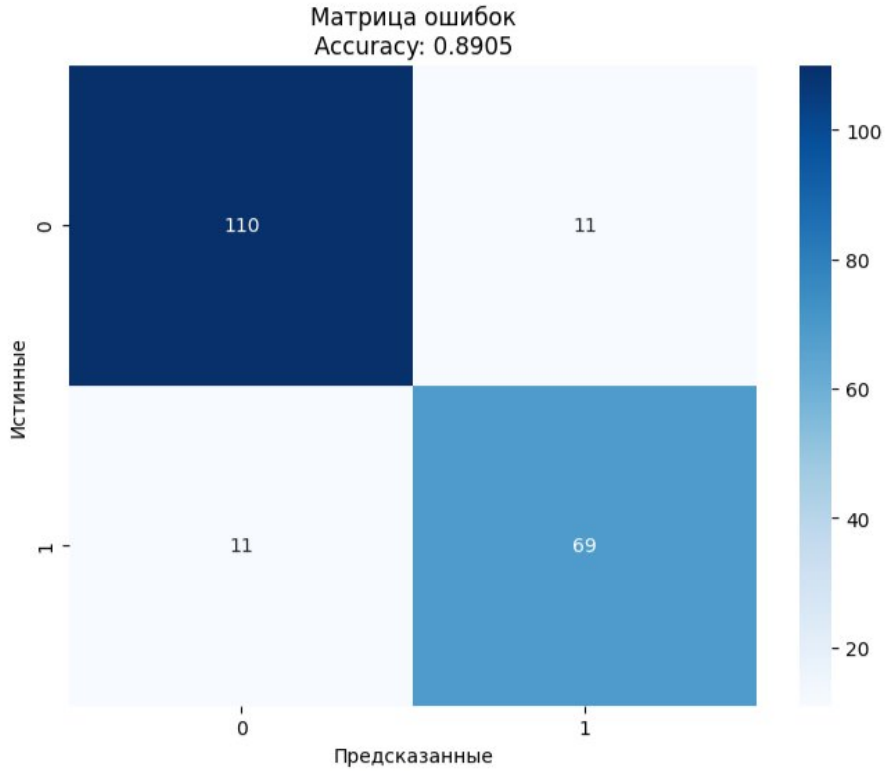
XGBoost продемонстрировал наилучшие результаты по всем метрикам:

- Средняя точность (0.8838) значительно выше, чем у других моделей.

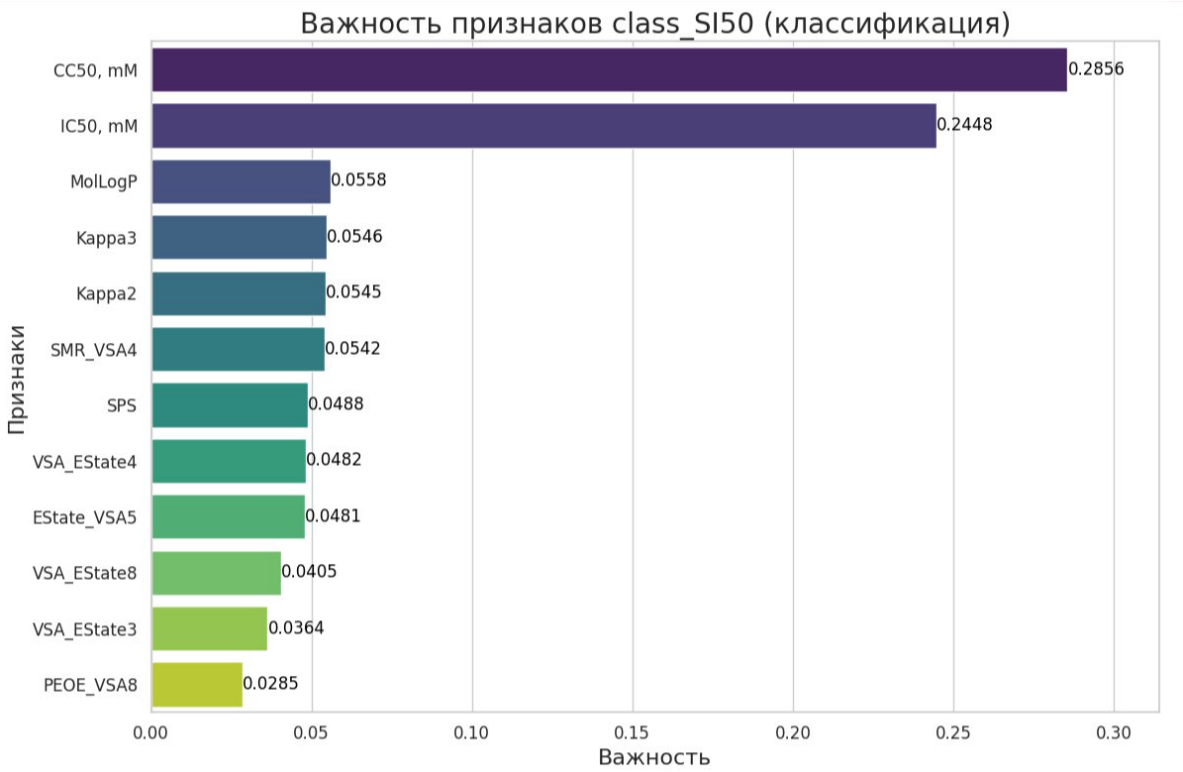
- Высокие значения F1-меры (0.8834), полноты (0.8838) и precision (0.8872) при низком стандартном отклонении (~0.013) указывают на стабильность и надежность модели

Классификация SI50

методы и средства



	precision	recall	f1-score	support
0	0.91	0.91	0.91	121
1	0.86	0.86	0.86	80
accuracy			0.89	201
macro avg	0.89	0.89	0.89	201
weighted avg	0.89	0.89	0.89	201



Классификация SI50

методы и средства

Выводы о важности признаков

CC50, mM: Важность = 0.2856
IC50, mM: Важность = 0.2448
MolLogP: Важность = 0.0558
Kappa3: Важность = 0.0546
Kappa2: Важность = 0.0545
SMR_VSA4: Важность = 0.0542
SPS: Важность = 0.0488
VSA_EState4: Важность = 0.0482
EState_VSA5: Важность = 0.0481
VSA_EState8: Важность = 0.0405
VSA_EState3: Важность = 0.0364 PEOE_
VSA8: Важность = 0.0285

Наиболее важный признак: CC50, mM с важностью 0.2856
Наименее важный признак: PEOE_VSA8 с важностью 0.0285

ГИПОТЕЗА

исследования

Основная гипотеза:

С помощью методов QSAR-моделирования и современных алгоритмов машинного обучения (ML) можно построить статистически значимые модели, которые на основе молекулярных дескрипторов предсказывают противогриппозную активность соединений (IC50), цитотоксичность (CC50) и селективность (SI) с точностью, не уступающей экспериментальной погрешности in-vitro тестов

H0

Соотношение SI = CC50/IC50 не поддаётся адекватному прогнозу из дескрипторов лучше, чем случайная модель

H1

ML-модели способны воспроизводить SI с прескажущей способностью, превышающей случайный уровень (например, $R^2 > 0.5$ или MAE ниже выбранного порога)

MSE: 30.18058481594498

MAE: 2.7801558801062187

R^2: 0.571972666119189

H1: ML-модели способны воспроизводить SI с предсказательной способностью, превышающей случайный уровень.

Отвергаем H0, принмсаем гипотезу H1: "ML-модели способны воспроизводить SI с предсказательной способностью, превышающей случайный уровень." Это значит, что модель действительно может делать предсказания, которые лучше случайных.

Практическое значение

исследования

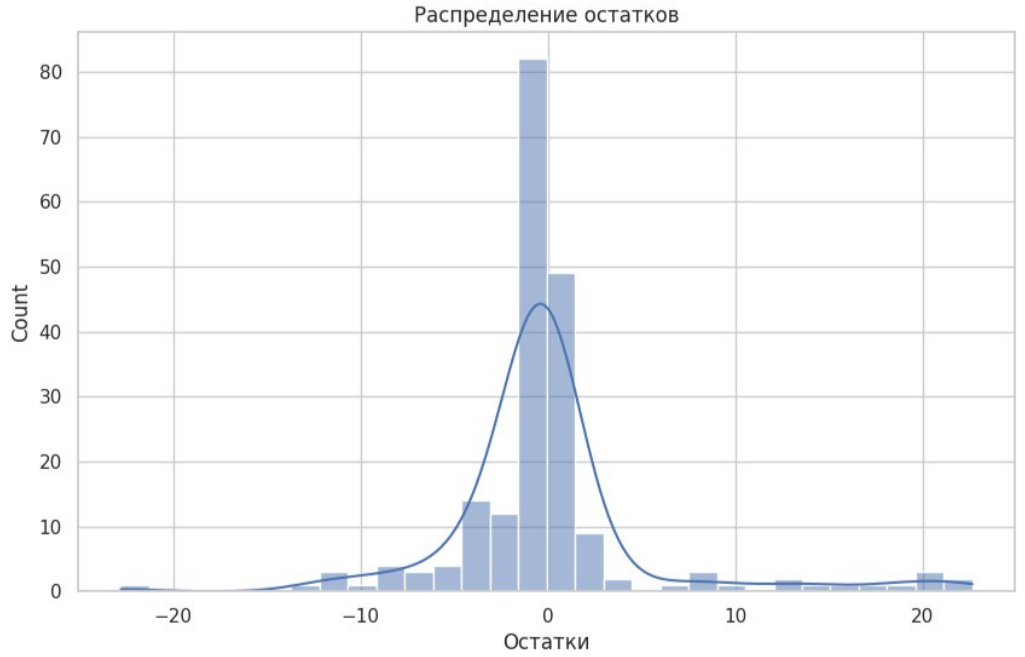
Теоретическая модель процесса разработки лекарств с использованием ML

1. Описание модели
Модель описывает процесс разработки нового лекарственного препарата с применением методов QSAR (Quantitative Structure-Activity Relationship) и машинного обучения (ML). Основная цель – предсказание биологической активности (IC_{50}, CC_{50}, SI) на основе молекулярных дескрипторов.

Ключевые этапы:
1. Сбор данных (1000 соединений с дескрипторами и активностью);
2. Предобработка данных (удаление шумов, нормализация, feature selection);
3. Построение ML-моделей (регрессия, ансамбли, нейросети);
4. Валидация (кросс-валидация, метрики: R^2, MAE, RMSE);
5. Интерпретация (SHAP, feature importance);
6. Экспериментальная проверка (синтез лучших кандидатов)

Вывод

Предложенная модель позволяет систематизировать процесс разработки лекарств с использованием ML, ускоряя поиск перспективных соединений. Визуализация и интерпретация помогают наладить взаимодействие между химиками и data scientists.



Гистограмма показывает распределение остатков (разности между истинными и предсказанными значениями). Идеальное распределение остатков должно быть нормальным с центром в нуле.

Если гистограмма симметрична и сосредоточена вокруг нуля, это говорит о том, что модель не имеет систематических ошибок (например, не переоценивает или недооценивает значения). Если наблюдаются смещения или асимметрия в распределении остатков, это может указывать на проблемы с моделью, такие как необходимость в преобразовании данных или добавлении новых признаков.

Библиография

перечень источников литературы

1. Курс лекци: “Искусственный интеллект в химии и материаловедении” https://teach-in.ru/course/ai-in-chemistry-and-materials-science/material

Работы исследовательской группы, в которую входят А. А. Митрофанов, Е. В. Матазова, Б. В. Егорова и соавт., охватывают два на первый взгляд разнесённых направления: (i) разработку макроциклических хелаторов для терапевтических радиоизотопов висмута-212/213 и (ii) применение методов машинного обучения (ML) для описания и прогнозирования свойств гибридных кристаллических материалов. Несмотря на тематические различия, оба направления решают общую задачу ускоренного дизайна функциональных соединений, опираясь на рациональный синтез, вычислительную химию и экспериментальную валидацию.

2. E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin, A. B. Tarasov
“Database of 2D Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps and Atomic Partial Charges Predicted by Machine Learning.”
Chemistry of Materials, 2020.