1. Цель и актуальность задания

Цель задания: разработать и проанализировать модели машинного обучения, способные предсказывать эффективность химических соединений, используемых при создании лекарственных препаратов. В частности, необходимо построить регрессионные и классификационные модели, которые позволяют прогнозировать ключевые параметры эффективности: IC50, CC50 и SI, а также классифицировать соединения по отношению к заданным порогам этих параметров.

Актуальность задания обусловлена высокой значимостью быстрого и точного прогнозирования биологической активности химических соединений на ранних этапах разработки лекарств. Эффективное применение методов машинного обучения в данной области позволяет существенно сократить временные и финансовые затраты на лабораторные исследования и повысить вероятность успешного создания действенного и безопасного лекарственного средства.

2. Анализ данных (EDA)

На начальном этапе исследования была проведена предварительная обработка данных, направленная на обеспечение качества и полноты исходной информации, а также создание условий для корректного построения моделей машинного обучения.

2.1. Обработка пропущенных значений

В предоставленном наборе данных были обнаружены пропущенные значения — по 3 пропуска в 12 числовых столбцах.

Пропуски были обработаны методом замены на медианные значения соответствующих столбцов. Выбор медианы обусловлен ее устойчивостью к выбросам и искажениям распределения, что особенно важно при наличии скошенных данных. Это позволило сохранить распределение признаков без внесения существенных искажений.

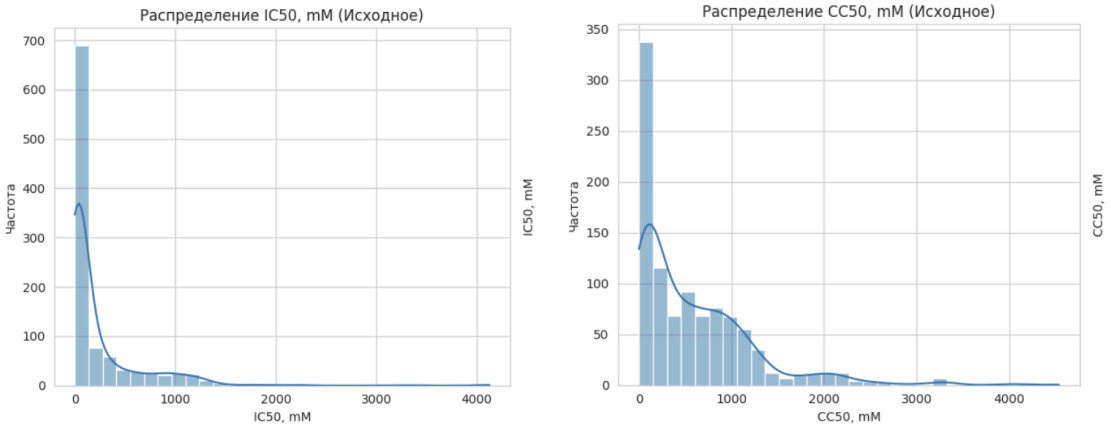
2.2. Удаление неинформативных признаков

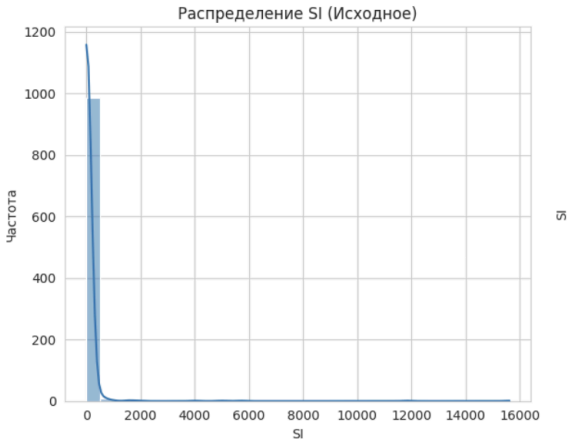
В датасете присутствовал столбец 'Unnamed: 0', который фактически представлял собой индекс, не нес информационной нагрузки и не мог способствовать обучению моделей. Данный столбец был удален во избежание его случайного использования как признака.

2.3. Работа с целевыми переменными

Основными целевыми переменными в задаче являются:

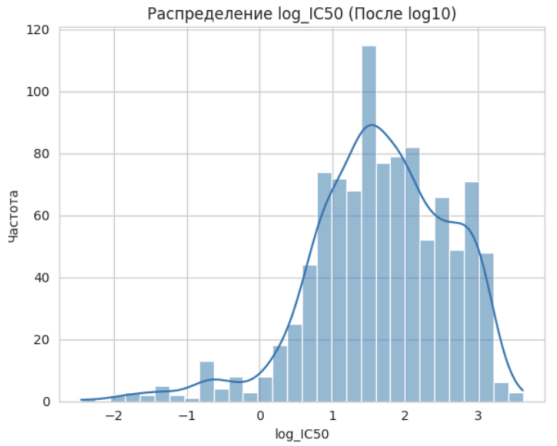
- 'IC50, mM' — концентрация вещества, при которой достигается 50% ингибирование активности,
- 'CC50, mM' — концентрация, вызывающая 50% цитотоксичность,
- 'SI' (selectivity index) — показатель избирательности, рассчитываемый как отношение CC50 к IC50.

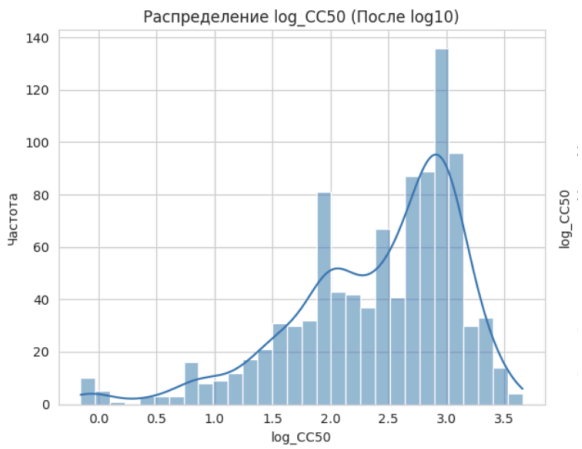


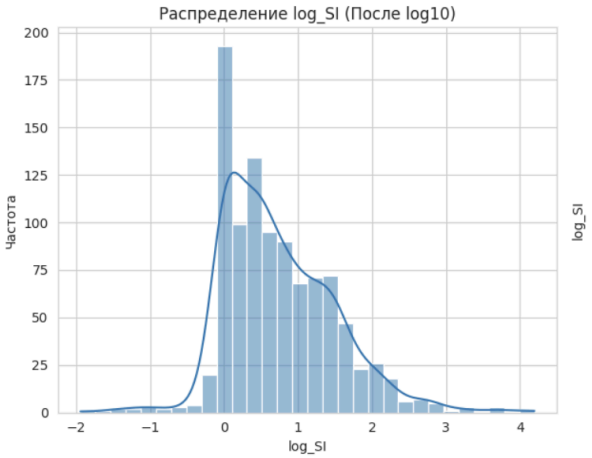


Первичный анализ показал, что все три переменные имеют правостороннее распределение с выраженной положительной асимметрией (скошенность). Для устранения скошенности и стабилизации дисперсии было применено логарифмическое преобразование, в результате чего были получены новые переменные:

- 'log_IC50',
- 'log_CC50',
- 'log_SI'.

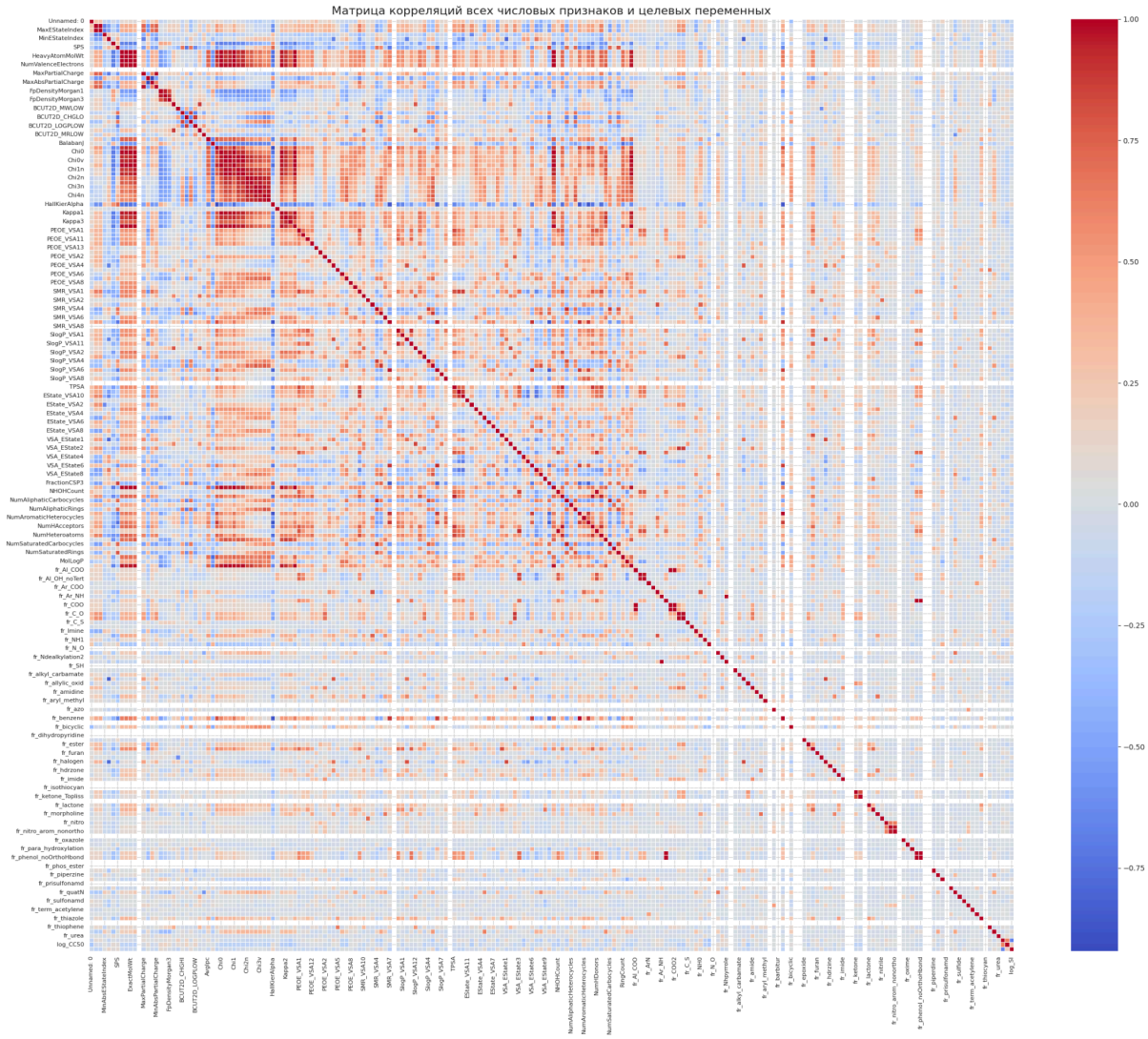






2.4. Корреляционный анализ признаков

Для изучения взаимосвязей между признаками и целевыми переменными была построена тепловая карта корреляций на основе коэффициента Пирсона.



Это позволило устранить признаки с высокой взаимной корреляцией между собой (мультиколлинеарность). Для этого была применена автоматическая процедура удаления одной из двух переменных в каждой паре, где коэффициент корреляции превышал заранее установленный порог (например, 0.9). Это способствует упрощению модели, предотвращению переобучения и повышению устойчивости предсказаний.

Таким образом, данные прошли подготовку и теперь подходят для последующего построения и анализа моделей машинного обучения.

3. Решение задач

Для построения моделей использовались 163 молекулярных дескриптора, рассчитанных для химических структур. Были исключены все признаки, напрямую связанные с IC50, CC50, SI и их логарифмами. Исследование включало как регрессионные задачи (предсказание числовых значений), так и задачи бинарной классификации (определение принадлежности к определенному классу по пороговому значению).

Были рассмотрены четыре типа моделей:

- Линейная регрессия / логистическая регрессия — базовая модель, обеспечивающая интерпретируемость, но ограниченная в способности моделировать нелинейные зависимости.
- Random Forest — ансамблевый метод, основанный на построении набора решающих деревьев, устойчивый к шуму и мультиколлинеарности.
- Gradient Boosting — мощная модель, последовательно обучающая слабые модели и позволяющая улавливать сложные зависимости в данных.
- XGBoost — оптимизированная версия градиентного бустинга, отличающаяся высокой скоростью и гибкой настройкой гиперпараметров.

Во всех случаях для повышения обобщающей способности моделей производилась стандартизация признаков и настройка гиперпараметров с использованием GridSearchCV с 5-кратной кросс-валидацией. Результаты были проанализированы по метрикам, релевантным каждому типу задачи.

3.1. Задача регрессии для IC50

Целью первой регрессионной задачи было построение модели, способной предсказывать логарифм IC50 по молекулярным дескрипторам. Этот показатель важен для оценки потенциала соединения как лекарственного средства.

Модель	MSE	RMSE	MAE	R^2
Random Forest	0.488	0.699	0.544	**0.510**
Gradient Boosting	0.503	0.709	0.558	0.495
XGBoost	0.516	0.718	0.558	0.482
Linear Regression	0.788	0.888	0.673	0.209

Наилучшие результаты были достигнуты с использованием Random Forest, который продемонстрировал наивысшее значение $R^2 = 0.510$, что указывает на относительно хорошее приближение модели к наблюдаемым значениям. Все ансамблевые методы показали сопоставимые и приемлемые значения метрик ошибки (RMSE, MAE), тогда как линейная регрессия существенно уступила, что указывает на наличие нелинейных взаимосвязей между признаками и целевой переменной.

3.2. Задача регрессии для CC50

Во второй задаче оценивалась токсичность соединений по параметру log(CC50). Правильное предсказание этого параметра критически важно для фильтрации соединений с потенциальной цитотоксичностью.

Модель	MSE	RMSE	MAE	R^2
Random Forest	0.254	0.504	0.360	**0.432**

XGBoost	0.260	0.510	0.359	0.419
Gradient Boosting	0.268	0.518	0.367	0.401
Linear Regression	0.353	0.594	0.446	0.212

Random Forest вновь продемонстрировал наивысшее значение $R^2 = 0.432$, что делает его наиболее подходящим выбором для прогнозирования CC50. Примечательно, что XGBoost достиг минимального значения MAE (0.359), показывая высокую точность при меньших отклонениях от реальных значений. Все бустинговые подходы значительно опережали по точности линейную регрессию, которая не смогла адекватно аппроксимировать зависимость.

3.3. Задача регрессии для SI

Прогноз SI — более сложная задача, поскольку SI представляет собой отношение двух параметров и может обладать высокой дисперсией.

Модель	MSE	RMSE	MAE	R^2
Random Forest	0.442	0.665	0.491	**0.275**
Gradient Boosting	0.452	0.672	0.496	0.260
XGBoost	0.465	0.682	0.518	0.237
Linear Regression	0.618	0.786	0.590	-0.014

Несмотря на сравнительно низкие значения R^2 по сравнению с предыдущими задачами, Random Forest вновь продемонстрировал лучшие результаты, обеспечивая наибольшую долю объяснённой дисперсии. Линейная модель, напротив, имела отрицательное значение R^2, что указывает на то, что она хуже среднего приближения и не применима для данной задачи.

3.4. Классификация: превышает ли значение IC50 медианное значение выборки

В данной работе была проведена бинарная классификация химических соединений по признаку активности, выраженной через значение log(IC50). Целевая переменная была сформирована на основе медианного значения log(IC50), равного 1.6682: соединения с log(IC50) выше медианы классифицировались как менее активные (класс 1), а с меньшим значением — как более активные (класс 0).
Для классификации были использованы 163 молекулярных дескриптора, за исключением тех, которые напрямую связаны с IC50, CC50 или SI, чтобы избежать утечек информации.

Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
XGBoost	0.711	0.691	0.76	0.724	**0.792**
Gradient Boosting	0.731	0.695	0.82	0.752	0.781
Random Forest	0.697	0.664	0.79	0.721	0.779
Logistic Regression	0.687	0.664	0.75	0.704	0.753614

Для определения активности соединений по порогу медианы log(IC50), градиентный бустинг и XGBoost показали наиболее высокие значения F1-меры и ROC AUC, демонстрируя отличную способность к разделению классов. Особенно важно, что при высокой полноте (recall), модели сохраняли и высокую точность (precision), что критически важно в биомедицинском контексте.

3.5. Классификация: превышает ли значение CC50 медианное значение выборки

Применен аналогичный подход, только порогом для бинарной классификации выступало медианное значение равное 2.6139.

Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
Gradient Boosting	0.736	0.712	0.79	0.749	**0.851**
Random Fores	0.687	0.658	0.77	0.710	0.839
Logistic Regression	0.716	0.684	0.80	0.737	0.818
XGBoost	0.697	0.676	0.75	0.711	0.817

Наиболее стабильные и высокие результаты по всем метрикам показал градиентный бустинг, особенно по ROC AUC = 0.851, что свидетельствует о высокой чувствительности модели к различиям между классами.

3.6. Классификация. Превышает ли SI медианное значение выборки

Применен аналогичный подход.

Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest	0.637	0.652	0.58	0.614	**0.685**

Gradient Boosting	0.652	0.679	0.57	0.620	0.673
XGBoost	0.627	0.624	0.63	0.627	0.662
Logistic Regression	0.632	0.635	0.61	0.622	0.657

Наиболее точной моделью оказался Random Forest, продемонстрировавший сбалансированность между точностью и полнотой и наивысшее значение ROC AUC.

3.7. Классификация – SI больше 8.

Для создания целевой переменной использовалась логарифмированная форма селективного индекса (log_SI).

В классификации участвовали четыре модели: Logistic Regression, Random Forest, Gradient Boosting, XGBoost.

Для каждой модели была проведена настройка гиперпараметров через GridSearchCV с 5-кратной кросс-валидацией, оптимизация шла по метрике ROC AUC. Отдельно подбирались параметры регуляризации, глубины деревьев, количества слабых моделей, скорости обучения и веса классов.

Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
Gradient Boosting	0.726	0.660	0.486	0.560	**0.748**
Random Fores	0.721	0.643	0.500	0.563	0.738
XGBoost	0.687	0.571	0.500	0.533	0.707
Logistic Regression	0.652	0.516	0.444	0.478	0.664

Наилучший результат показала модель Gradient Boosting, которая достигла ROC AUC = 0.748. Это означает, что модель способна достаточно хорошо отличать высоко-селективные соединения от остальных. При этом она демонстрирует хорошую сбалансированность между полнотой (recall = 0.486) и точностью (precision = 0.660), что особенно важно для выявления редких, но перспективных соединений с высоким SI.
Для моделей Random Forest и XGBoost также были получены хорошие результаты (ROC AUC > 0.70), что подтверждает устойчивость задачи к различным ансамблевым подходам. Logistic Regression выступила слабее, особенно по метрике recall, что говорит о ее ограниченной способности выявлять класс с SI > 8 в данной задаче.

4. Вывод

4.1. Общие выводы по данным, целевым переменным и моделям

Анализ трёх ключевых целевых переменных — log(IC50), log(CC50) и log(SI) — показал, что между молекулярными дескрипторами и биологической активностью существует значимая, хотя и не полностью линейная зависимость.

log(IC50): наиболее устойчиво предсказываемая переменная. Высокие значения R^2 и метрик классификации (Accuracy, ROC AUC) свидетельствуют о чёткой связи между признаками и активностью соединений.

log(CC50): также демонстрирует хорошие результаты, однако показывает больший разброс предсказаний и повышенную чувствительность к параметрам модели.

log(SI): наименее предсказуемая переменная, что объясняется её составной природой (отношение CC50 к IC50). Это приводит к усиленной дисперсии и нестабильности. При классификации по порогу log(SI) > 8 возникает сильный дисбаланс классов, что дополнительно усложняет обучение.

Сравнение четырёх моделей выявило чёткую закономерность:

Ансамблевые методы (Random Forest, Gradient Boosting, XGBoost) значительно превосходят базовые алгоритмы по обобщающей способности, особенно при наличии сложных нелинейных зависимостей.

Random Forest стабильно показал наивысшие значения R^2 в регрессионных задачах, особенно при прогнозировании log(IC50) и log(CC50).

Gradient Boosting оказался наиболее точным в бинарной классификации, особенно при использовании медианных порогов.

XGBoost показал высокую производительность, особенно при грамотной настройке гиперпараметров, которая проводилась с использованием интеллектуальных подходов вместо обычного Grid Search.

В то же время линейная и логистическая регрессия оказались наименее эффективными, так как не способны моделировать сложные взаимосвязи между признаками. Однако их можно использовать в качестве базовой модели (baseline) для оценки эффективности более продвинутых алгоритмов.

4.2. Пути улучшения качества моделей

Для повышения эффективности моделей, используемых в задачах предсказания IC50, CC50, SI и соответствующих классификациях, можно предпринять следующие шаги:

1. Создание новых признаков. Разработка дополнительных, более информативных признаков позволяет точнее описывать структуру и свойства молекул. Это особенно важно при моделировании сложных химических взаимодействий, где стандартных характеристик может быть недостаточно.

Такие признаки помогают моделям глубже понимать внутренние закономерности данных.

2. Балансировка классов в выборке. В задачах классификации часто возникает проблема несбалансированных классов: один класс может значительно преобладать над другими. Это приводит к смещению модели в сторону "большинства", снижая точность предсказаний на малочисленных классах. Методы балансировки, такие как oversampling, undersampling или генерация синтетических примеров, позволяют устранить этот перекос и повысить объективность модели.
3. Расширение обучающей выборки. Для повышения устойчивости и обобщающей способности модели полезно увеличить объем данных. Это можно сделать с помощью методов генерации новых примеров на основе уже имеющихся данных. Такие подходы особенно эффективны, когда получение новых реальных данных затруднено или ресурсоёмко.
4. Интеллектуальная настройка гиперпараметров. Вместо традиционного перебора параметров по сетке (Grid Search), целесообразно использовать более продвинутые методы оптимизации — например, случайный поиск (Random Search), байесовскую оптимизацию или алгоритмы на основе обучения с подкреплением. Это позволяет находить более качественные настройки модели с меньшими затратами ресурсов.