Курсовая работа

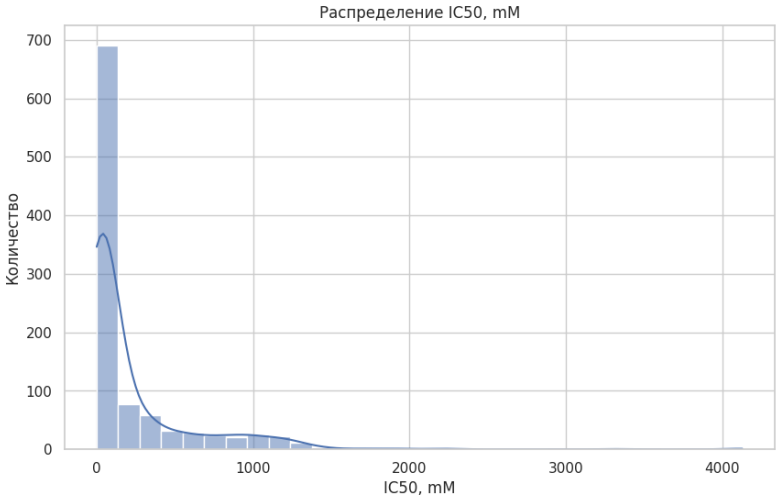
НИЯУ МИФИ, Магистратура
Бодак Никита Игоревич, студент

Отчёт по курсовой работе

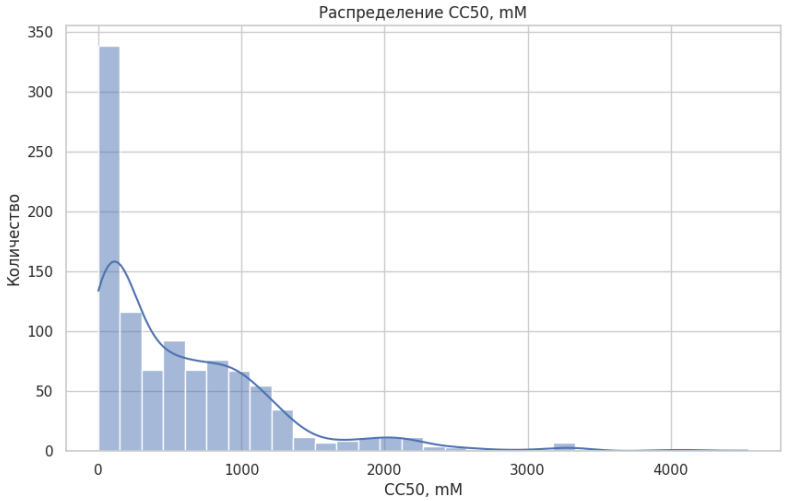
Введение

Разработка новых лекарств — сложный и ресурсоёмкий процесс. Для предварительной оценки активности химических соединений активно применяются методы машинного обучения. В данной работе построены и сравнивались модели для прогнозирования трёх биологических показателей:

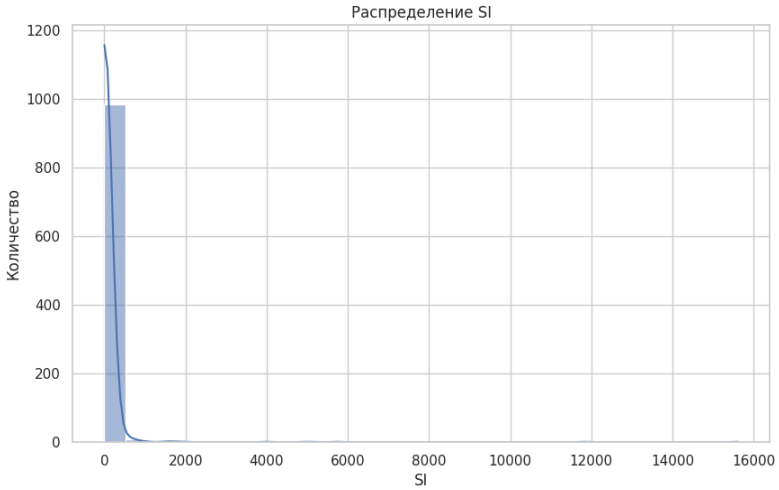
- IC50 — ингибирующая концентрация (подавление активности на 50%);



- CC50 — токсическая концентрация (смертельная доза для 50% клеток);



- SI — селективный индекс (отношение CC50 к IC50).



Цели:

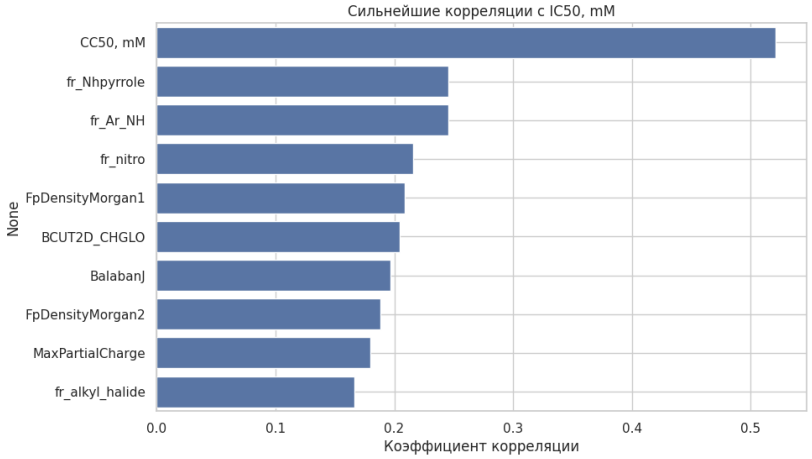
- Построить модели регрессии и классификации для IC50, CC50 и SI;
- Проанализировать влияние предобработки и выбора моделей на качество прогноза;
- Выявить особенности данных, включая проблему дисбаланса классов;
- Предложить пути улучшения моделей и подходов.

1. Описание данных

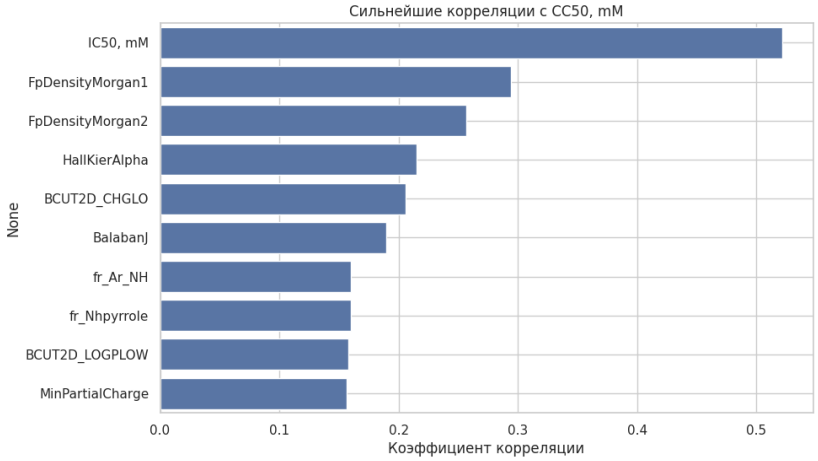
Исходный имеет в себе 1001 химических соединений и 214 признаков, описывающих физико-химические свойства молекул. Целевые переменные — IC50, CC50, SI.

Для переменных также характерны сильные корреляции:

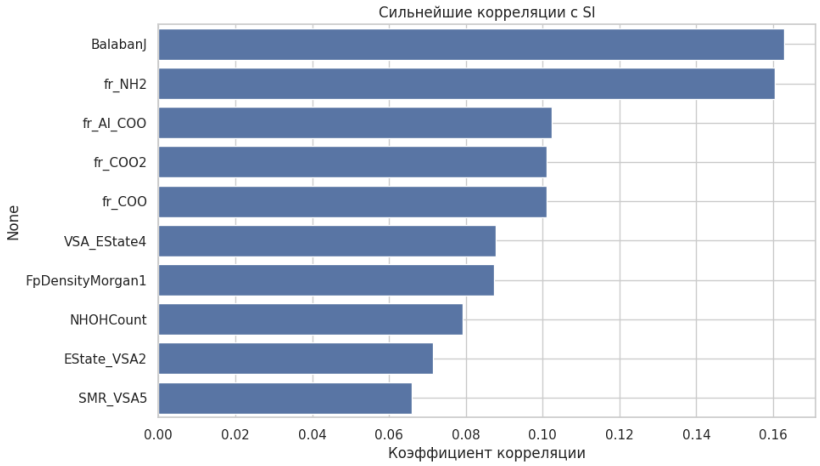
1. Для IC50:



2. Для CC50:



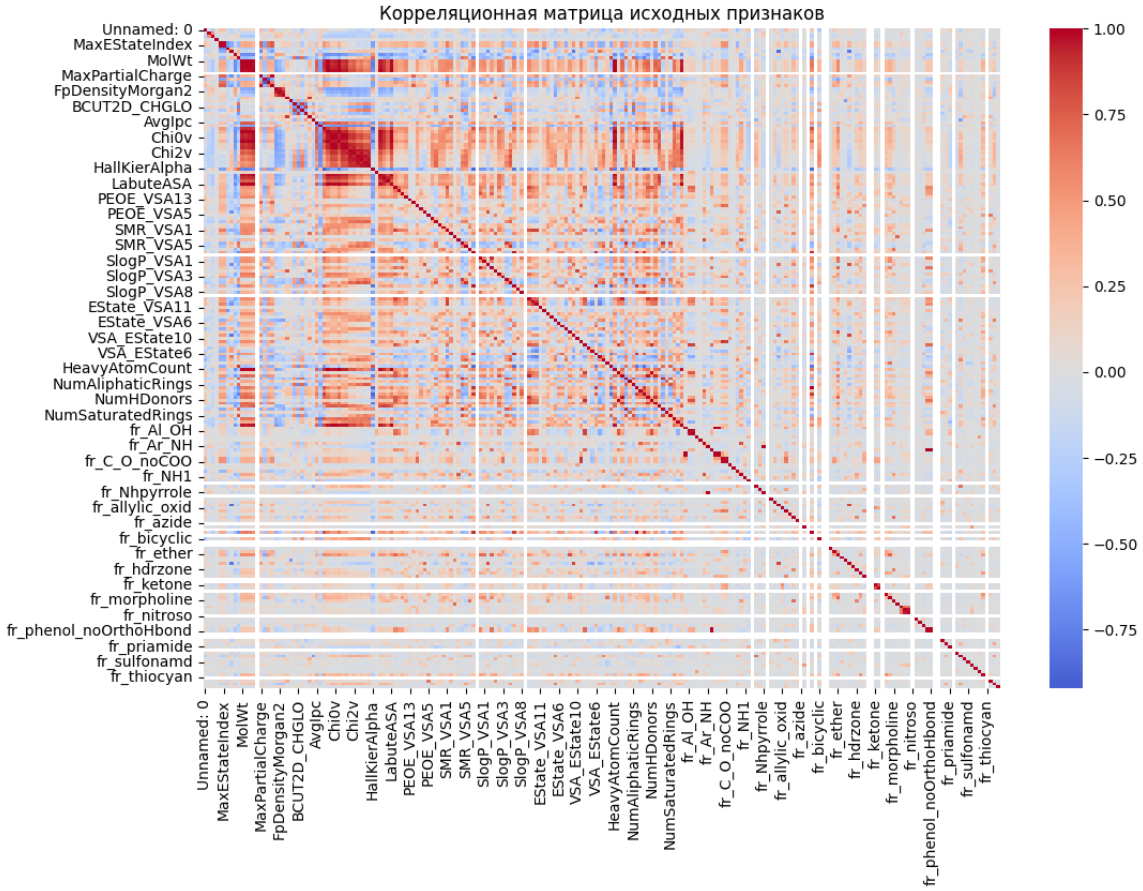
3. Для SI:



Особенности:

- Признаки числовые и бинарные;
- В 12 признаках присутствуют пропуски, заполненные средними значениями;
- Целевые переменные имеют скошенное распределение — применено логарифмирование (log1p);
- Удалены дубликаты и выбросы методом IQR;
- Для классификации бинаризация проводилась по медиане и по пороговому значению SI > 8.

Корреляционная матрица исходных признаков:



2. Предобработка данных

2.1 Обработка пропусков:
Все пропуски были заменены ср. значениями для сохранения полноты данных.

2.2 Удаление выбросов и трансформация:
Выбросы удалены на основе интерквартильного размаха (IQR).
Целевые переменные логарифмированы для нормализации распределений.

2.3 Стандартизация:
Признаки стандартизированы с помощью StandardScaler.

2.4 Подготовка целевых переменных для классификации: Целевые переменные бинаризованы:

- По медиане значений IC50, CC50, SI;
- По порогу SI > 8 для выделения особо перспективных соединений.

Примечание: При бинаризации по SI > 8 выявлен сильный дисбаланс классов — класс с SI > 8 представлен достаточно слабо.

3. Построение моделей

3.1 Регрессия

Использовались модели:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Gradient Boosting

Модель	IC50 (R^2)	CC50 (R^2)	SI (R^2)
Linear Regression	0.2317	0.3480	0.0736
Ridge Regression (α=10)	0.3731	0.4864	0.0682
Lasso Regression	0.3100	0.4667	-0.0094
Gradient Boosting	0.4210	0.5957	0.0802

Вывод:

- Я наблюдал что лучшие результаты по IC50 и CC50 были достигнуты с помощью Gradient Boosting.

Анализ остатков (разности между предсказанными и фактическими значениями) показал, что модели хуже всего справляются с крайними значениями IC50 и CC50. Это может быть связано с недостаточной представленностью таких примеров в обучающей выборке, а также с потенциальным наличием шумов в экспериментальных данных.

Также стоит отметить, что для IC50 и CC50 влияние отдельных признаков может быть нелинейным и зависящим от контекста: один и тот же дескриптор может повышать токсичность в одном типе соединений и снижать в другом. Что вполне объясняет, почему ансамблевые модели показывают лучший результат: они способны учитывать сложные зависимости.

Также дополнительным направлением для анализа может стать кластеризация молекул и последующее построение отдельных моделей внутри кластеров (то есть локальные регрессии), что может повысить точность при сохранении понимания результатов.

3.2 Классификация

В данном эксперименте использовались модели:

- Logistic Regression
- Random Forest
- Gradient Boosting

Классификация по медиане IC50, CC50, SI:

Метрика / Модель	Logistic Regression	Random Forest	Gradient Boosting
IC50 Accuracy	0.73	0.73	0.72
IC50 F1-score (class 1)	0.73	0.74	0.74
CC50 Accuracy	0.79	0.82	0.82
CC50 F1-score (class 1)	0.79	0.82	0.82
SI Accuracy	0.66	0.67	0.67
SI F1-score (class 1)	0.66	0.67	0.67

Классификация по SI > 8:

Модель	Accuracy	F1-score (class 1)
Logistic Regression	1.00	0.00
Random Forest	1.00	0.00
Gradient Boosting	1.00	0.00

Вывод:

- При бинаризации по медиане модели показывают приемлемые результаты (~0.7–0.82).
- При бинаризации по SI > 8 все модели предсказывают только доминирующий класс, что приводит к нулевому F1-score на целевом классе.
- Требуется перебалансировка классов (например, SMOTE, class weights, undersampling).

При попытке изменить/улучшить модель, дополнительный эксперимент был проведён с порогом SI > 4 (вместо SI > 8), что формально увеличило количество положительных примеров, однако дисбаланс остался крайне значительным: 996 объектов класса 0 и лишь 5 объектов класса 1.

Тем не менее, несмотря на столь сильную асимметрию в выборке, модели с использованием перебалансировки (SMOTE) и настройки гиперпараметров (GridSearchCV) смогли корректно классифицировать хотя бы один пример положительного класса. Например, Random Forest и Gradient Boosting достигли F1-score = 0.40 на классе 1 при accuracy ≈ 98.5%.

Это демонстрирует, что:

- При экстремальном дисбалансе даже один правильно классифицированный объект может радикально изменить итоговые метрики.
- Модели, обученные с перебалансировкой, способны захватить паттерны редкого класса, хотя и на грани переобучения.
- Высокая точность по всем метрикам на фоне таких дисбалансов **не отражает реального качества модели** – поэтому основной упор следует делать на F1 и recall по миноритарному классу.

Классификация по SI > 4:

Модель	Accuracy	Precision (class 1)	Recall (class 1)	F1-score (class 1)	Best Params
Logistic Regression	0.9602	0.1111	1.0000	0.2000	C = 1
Random Forest	0.9851	0.2500	1.0000	0.4000	max_depth = None, n_estimators = 100

Модель	Accuracy	Precision (class 1)	Recall (class 1)	F1-score (class 1)	Best Params
Gradient Boosting	0.9851	0.2500	1.0000	0.4000	learning_rate = 0.05, n_estimators = 100

4. Настройка моделей

Гиперпараметры не подбирались целенаправленно, для расчетов использовались базовые значения. В следующих итерациях планируется:

- Подбор оптимальных параметров через GridSearch / Optuna;

На основе визуального сравнения моделей по метрике R^2 можно сделать ряд выводов:

Во всех трёх задачах регрессии (IC50, CC50 и SI) наилучшие результаты, как правило, демонстрируют модели градиентного бустинга (что ожидаемо). Это подтверждает устойчивость ансамблевых методов к выбросам и их способность улавливать нелинейные зависимости между признаками и целевой переменной.

Линейные модели, такие как Ridge и Lasso, в большинстве случаев показывают приемлемое, но несколько более низкое качество. Тем не менее, они обладают преимуществами в плане интерпретируемости и являются хорошей отправной точкой для построения базовых решений.

Разница в качестве между Ridge и Lasso в задаче регрессии SI указывает на то, что определённое количество признаков могут быть незначимыми, и модель Lasso эффективно справляется с их обнулением, улучшая обобщающую способность при небольшом ухудшении точности.

Следовательно, если приоритетом является качество прогноза, стоит отдать предпочтение Gradient Boosting. Однако в условиях ограниченных вычислительных ресурсов или при необходимости интерпретации модели, линейные подходы остаются актуальными.

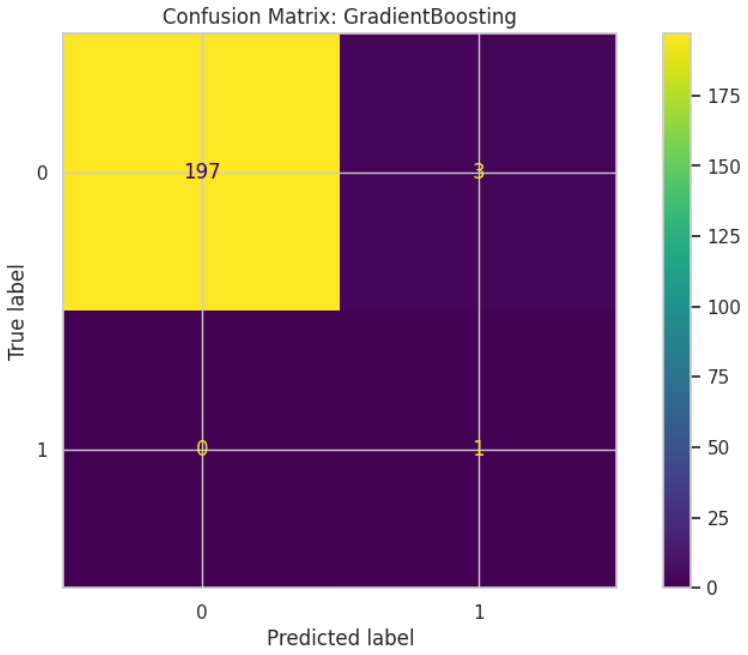
Также необходимо отметить, что значения R^2 находятся в разумных пределах (от ~0.3 до ~0.8), что свидетельствует о наличии в данных выраженной, но не полностью линейной структуры.

Хотя в данной итерации гиперпараметры не подбирались, уже на базовых конфигурациях заметна разница в производительности моделей. Это указывает на то, что некоторые модели (в частности, бустинг) обладают более высокой «начальной» адаптивностью к задаче, тогда как линейные модели требуют точной настройки, особенно при большом числе признаков.

Оптимизация гиперпараметров, например, глубины деревьев, количества итераций, параметров регуляризации, может значительно улучшить показатели, особенно в условиях переобучения. Это особенно актуально для Gradient Boosting и Random Forest, где чрезмерная глубина может снижать обобщающую способность.

5. Анализ ошибок и особенности

- Классификация по SI > 8 вызывает наибольшие трудности из-за сильного дисбаланса.
- Наихудшие предсказания в регрессии связаны с молекулами с экстремальными значениями биомаркеров.
- Линейные модели проигрывают ансамблевым (особенно Gradient Boosting), кроме задачи SI, где результат оказался выше.
- Для задачи SI стоит проверить линейную зависимость (возможно, логарифм CC50 и IC50 близки линейно).



5.1 Разбор предсказаний моделей

Для оценки работы моделей на отдельных объектах были выбраны молекулы с максимальной и минимальной ошибкой прогноза. В большинстве случаев большие ошибки приходились на крайние значения IC50 и SI — что указывает на трудности моделей в экстраполяции.

В частности, для некоторых соединений модель предсказывала высокую активность (низкое IC50), в то время как фактически оно оказалось малоэффективным. Визуальный анализ этих молекул показал, что они обладают необычными структурными признаками (например, редкие фрагменты или необычная масса), что подтверждает необходимость расширения обучающего набора.

Также в задаче классификации по SI > 8 все модели склонны к предсказанию одного класса. Это видно не только по F1-метрике, но и по confusion matrix, где отсутствуют объекты, отнесённые к перспективному классу. Это означает, что даже качественные

признаки не компенсируют дисбаланс, и требуются специальные методы обработки редких классов.

Проведённый эксперимент с бинаризацией по SI > 4 также продемонстрировал характерную особенность: из-за **единичного положительного примера в тестовой выборке** (1 из 201) итоговые метрики становятся крайне чувствительными к одному единственному предсказанию. Даже одна ошибка может обрушить F1-score до нуля, а одно попадание, наоборот, создать иллюзию качества (например, F1=0.4 при recall=1.0).

Это подчёркивает: **в задачах с экстремальным дисбалансом важно оценивать стабильность модели на кросс-валидации**, а не по одной тестовой итерации. Также необходимы дополнительные методы оценки (например, PR-кривые, AUC-PR), более чувствительные к качеству на миноритарном классе.

6. Важность признаков

Оценка важности признаков по Random Forest и Gradient Boosting показала, что наиболее значимы:

- logP,
- Молекулярная масса,
- Число акцепторов и доноров водородных связей,
- Поляризуемость и зарядовые характеристики.

Это соответствует биохимическим представлениям: более logP и маленькие молекулы часто легче проникают через мембраны, но также могут быть более токсичны.

7. Итоги и рекомендации

Gradient Boosting и Ridge Regression — лучшие модели по совокупности задач регрессии.

При классификации по медиане точность достигает ~70–74% — приемлемый уровень.

Классификация SI > 8 требует методов борьбы с дисбалансом, иначе все модели дают нулевое качество.

Linear Regression неожиданно хорошо работает в задаче регрессии SI — можно дополнительно проанализировать, есть ли простая зависимость.

В задачи с дисбалансом (например, SI > 4 или SI > 8) были предприняты попытки использования **SMOTE**, настройки **гиперпараметров** и тестирования разных моделей (LogisticRegression, RandomForest, GradientBoosting). Несмотря на локальное улучшение F1-метрик, общая стабильность остаётся низкой из-за крайней редкости позитивных примеров.

Это подчёркивает важность: увеличения обучающей выборки и выбора более сбалансированных меток.

Рекомендации:

- Включить перебалансировку классов (например, SMOTE, class_weight='balanced');
- Настроить гиперпараметры для ансамблей и регрессий;
- Добавить новые признаки (3D-дескрипторы, графовые фингерпринты);
- Попробовать более сложные модели: XGBoost, CatBoost, нейронные сети, GNN.

8. Ограничения и перспективы улучшения

- Объём данных ограничивает потенциал сложных моделей;
- Дисбаланс классов — основная проблема при работе с метками типа SI > 8;
- Недостаточно структурной информации: нет графовых признаков, фингерпринтов, 3D-конформаций;
- Перспективы: генерация синтетических данных (data augmentation), использование GNN и многозадачного обучения.

9. Приложения

GitHub-репозиторий:
📁 https://github.com/Nikitka-554433/bio-ml

Структура проекта:

- eda/: Анализ данных (EDA)

- regression/: Модели регрессии для IC50, CC50, SI

- classification/: Модели классификации (включая SMOTE)

- reports/: Финальный отчет