Прогнозирование эффективности химических соединений с использованием методов классического машинного обучения

Выполнил: Вячеслав Дмитриев

Дисциплина: Классическое машинное обучение

Дата: 16 августа 2025 г.

1. Введение

Разработка новых лекарственных препаратов — сложный, длительный и дорогостоящий процесс, включающий синтез соединений, биологические испытания и клинические исследования. Современные методы машинного обучения позволяют существенно ускорить начальные этапы, особенно отбор перспективных соединений на основе их химических свойств.

В рамках данной работы была поставлена задача построить модели машинного обучения для анализа 1000 химических соединений, оцененных по трём ключевым показателям:
1) **IC50 — эффективность** (концентрация, подавляющая 50% активности вируса);
2) **CC50 — токсичность** (концентрация, вызывающая гибель 50% клеток);
3) **SI (Selectivity Index) — индекс селективности:** SI = CC50 / IC50, отражающий терапевтическое окно.

Цель работы — построить максимально эффективные модели для решения задач регрессии и классификации, проанализировать их качество и дать рекомендации по использованию в дальнейших исследованиях.

2. Описание данных

Исходный датасет содержит 1001 запись и 214 признаков, включая: 3 целевые переменные: IC50, mM, CC50, mM, SI, 211 числовых и бинарных химических дескрипторов (например, MaxAbsEStateIndex, qed, SPS, fr_urea и др.)

После очистки данных:

- Удалены 3 строки с пропущенными значениями в 12 признаках (менее 0.3% от общего объёма);
- Удалён служебный столбец Unnamed: 0;
- Итоговый размер: 998 строк × 213 столбцов.

Данные сохранены в файл data/cleaned_data.csv для дальнейшего использования.

3. Исследовательский анализ данных (EDA)

3.1. Распределение целевых переменных

IC50 (медиана: 46.59 mM) — распределение с длинным правым хвостом. Низкие значения указывают на высокую активность.

CC50 (медиана: 411.04 mM) — также с перекосом. Высокие значения означают низкую токсичность.

SI (медиана: 3.85) — индекс селективности. Важно, что SI > 8 считается порогом перспективных соединений.

Все три переменные были логарифмированы перед моделированием для нормализации распределения.

3.2. Проверка формулы SI

Было подтверждено, что: SI ≈ CC50 / IC50, со средней ошибкой менее 1e-6. Это означает, что SI нельзя использовать как признак при предсказании IC50 или CC50 — это приведёт к утечке данных.

3.3. Профили соединений

Соединения были классифицированы по биологическому профилю:

Перспективные (2.4%) — низкий IC50 + высокий SI (>8)

Селективные (12.1%) — SI > 8

Активные (48.6%) — IC50 < медианы

Безопасные (49.3%) — CC50 > медианы

Слабые (38.5%) — не соответствуют критериям

Только небольшая часть соединений одновременно эффективна и безопасна.

4. Методология моделирования

Для каждой из 7 задач были протестированы следующие модели:

- Линейная регрессия / Логистическая регрессия;

- Random Forest;

- Gradient Boosting;

- SVM.

Все модели проходили: кросс-валидацию (5 фолдов), подбор гиперпараметров через GridSearchCV, оценку по релевантным метрикам. Данные разделялись на обучающую (80%) и тестовую (20%) выборки с учётом стратификации (для классификации).

5. Результаты моделирования

5.1. Регрессия

Задача	Лучшая модель	R2 test	rmse
IC50	Gradient Boosting	0.87	0.31
CC50	Gradient Boosting	0.91	0.23
SI	Random Forest	0.79	0.41

Выводы:

CC50 предсказывается точнее всего — токсичность зависит от устойчивых химических свойств. SI — сложнее всего, так как это производная величина, чувствительная к шуму. Все модели выиграли от логарифмического преобразования.

5.2. Классификация

Задача	Лучшая модель	AUC	Accuracy
C50 > медианы	Gradient Boosting	0.89	0.82
CC50 > медианы	Gradient Boosting	0.90	0.83
SI > медианы	Random Forest	0.86	0.79
SI > 8	Gradient Boosting	0.92	0.85

Выводы:

Все модели показали высокое качество (AUC > 0.85). Gradient Boosting — лидер по AUC в 3 из 4 задач. Задача SI > 8 — наиболее важная с биологической точки зрения.

6. Сравнение моделей и выбор оптимальных

Модель	Сильные стороны	Слабые стороны	Рекомендации
Gradient	Высокая точность,	Долгое обучение,	Основная модель

Boosting	хорошая AUC	сложнее интерпретировать	для всех задач
Random Forest	Устойчив к переобучению, хорош с нелинейностями	Меньше AUC, чем GB	Альтернатива, особенно для SI
Logistic Regression	Интерпретируема, быстрая	Уступает по качеству	Только для базового анализа
SVM	Хорош в сложных границах	Долгий, чувствителен к масштабу	Не рекомендуется без нормализации

Выводы:

Ансамбли деревьев (Gradient Boosting и Random Forest) показали наилучшие результаты во всех задачах. CC50 предсказывается точнее, чем IC50, так как токсичность зависит от более стабильных молекулярных свойств. SI > 8 — ключевой критерий отбора, и его можно надёжно предсказывать с AUC = 0.92. Логарифмирование целевых переменных значительно улучшает качество моделей. Утечка данных через SI — критическая ошибка, которую необходимо избегать.

Для химиков и фармацевтов:

Следует использовать модель Gradient Boosting для SI > 8 как фильтр при скрининге новых соединений. Необходимо комбинировать предсказания IC50 и CC50 для расчёта SI — это безопаснее, чем предсказывать SI напрямую.

Заключение

В ходе работы были построены и протестированы множество моделей машинного обучения для прогнозирования эффективности и токсичности химических соединений. Все поставленные задачи решены с высоким качеством. Наиболее перспективные модели — Gradient Boosting и Random Forest — могут быть использованы для ускорения разработки новых препаратов против вируса гриппа.Работа демонстрирует сильный потенциал взаимодействия между химией и машинным обучением, позволяя эффективно отбирать соединения для дальнейших испытаний.