Курсовая Работа
по дисциплине
«Классическое машинное обучение»

Аналитический отчёт «Разработка статистических моделей для поиска эффективных лекарственных соединений»

Выполнил:
студент группы М24-525
Тураносов Константин Викторович

Москва 2025

Введение

В последние годы искусственный интеллект (ИИ) активно внедряется в различные сферы, включая фармацевтику. Разработка новых лекарств – это долгий и дорогостоящий процесс, но ИИ может помочь его ускорить, снизить затраты и повысить эффективность исследований. Хотя ИИ не способен полностью заменить ученых, он становится мощным инструментом для анализа данных, поиска новых соединений и оптимизации экспериментов.

Создание лекарств включает множество этапов: от поиска химической формулы до клинических испытаний. Современные методы машинного обучения позволяют быстрее анализировать данные, предсказывать свойства соединений и выбирать наиболее перспективные варианты. Однако для успешной работы важно, чтобы химики и специалисты по ИИ эффективно взаимодействовали, что не всегда просто.

В данном исследовании рассматривается применение ИИ для анализа данных о 1000 химических соединениях, предоставленных химиками. Эти данные содержат параметры эффективности против вируса гриппа (IC_{50}, CC_{50} и SI). Проверка гипотезы, как машинное обучение может помочь в оценке потенциальных лекарственных веществ, даже если у автора нет глубоких знаний в химическом анализе.

Цель работы

Цель работы - разработка программного инструмента на основе методов машинного обучения, который позволит повысить эффективность создания новых лекарственных препаратов за счет:

- прогнозирования ключевых характеристик молекул
- автоматизации анализа химических соединений
- оптимизации процесса отбора перспективных соединений

Основные задачи исследования:

1. Проведение комплексного анализа данных (EDA):
    - Исследование предоставленного набора данных о 1000 соединениях
    - Анализ распределения ключевых параметров (IC50, CC50, SI)
    - Выявление взаимосвязей между характеристиками соединений
2. Разработка моделей регрессии:
    - Создание модели прогнозирования значения IC50
    - Построение модели предсказания CC50
    - Разработка модели расчета индекса SI
3. Создание классификационных моделей:
    - Бинарная классификация по превышению медианного значения IC50
    - Бинарная классификация по превышению медианного значения CC50
    - Бинарная классификация по превышению медианного значения SI
    - Классификация соединений по превышению порога SI=8
4. Анализ результатов:
    - Сравнение качества построенных моделей
    - Выбор наиболее эффективных решений
    - Подготовка рекомендаций по применению
5. Техническая реализация:
    - Оформление EDA в отдельном Jupyter Notebook
    - Создание отдельных Jupyter Notebook для каждой модели

Первичный Анализ Данных.
EDA

Данные представляют собой химические харрактеристики (дескрипторы) описывающие свойства молекулы и ее фрагментов.

1. Целевые переменные: биологическая активность и токсичность

a) IC_{50} (Half Maximal Inhibitory Concentration)

Определение:

Концентрация соединения, необходимая для 50%-ного подавления целевого биологического процесса (например, вирусной репликации).

Интерпретация:

- Чем ниже IC_{50}, тем выше активность (эффект достигается при малых дозах).
- Примеры:
  - IC_{50} = 1 мкМ → высокая активность.
  - IC_{50} = 100 мкМ → низкая активность.

b) CC_{50} (Half Maximal Cytotoxic Concentration)

Определение:

Концентрация вещества, вызывающая гибель 50% клеток (показатель токсичности).

Интерпретация:

- Чем выше CC_{50}, тем безопаснее соединение.
- Примеры:
  - CC_{50} = 10 мМ → низкая токсичность.
  - CC_{50} = 10 мкМ → высокая токсичность.

c) SI (Selectivity Index)

Формула:

$$SI = IC50/\mathrm{IC50}$$

Определение:

Показывает избирательность действия соединения на мишень.

Критерии:

- SI > 10 → перспективный препарат (высокая селективность).
- SI = 1–10 → умеренная токсичность.
- SI < 1 → опасное соединение.

Идеальный препарат:

- Низкий IC_{50} + Высокий CC_{50} + SI ≥ 10.

2. Электронные и стерические свойства

Дескриптор	Описание
MaxAbsEStateIndex	Максимальное абсолютное значение E-State
MinAbsEStateIndex	Минимальное абсолютное значение E-State
MaxEStateIndex	Максимальное значение электронного индекса
MinEStateIndex	Минимальное значение электронного индекса

3. Физико-химические свойства

Дескриптор	Описание
qed	Оценка "лекарственности"
MolWt	Молекулярная масса
ExactMolWt	Точная молекулярная масса
HeavyAtomMolWt	Масса тяжелых атомов (без водородов).
MolLogP	Коэффициент распределения октанол/вода
TPSA	Полярная поверхностная площадь
MolMR (Molar Refractivity)	Молярная рефракция (связана с поляризуемостью).
FractionCSP3	Доля sp^3-гибридизированных атомов углерода (влияет на жесткость молекулы).

4. Электронные и зарядовые характеристики

Дескриптор	Описание
NumValenceElectrons	Число валентных электронов
MaxPartialCharge	Максимальный парциальный заряд
BCUT2D_CHGHI	Дескриптор распределения зарядов

5. Топологические дескрипторы

Дескриптор	Описание
BalabanJ	Индекс сложности молекулярного графа
Chi0, Chi1	Индексы молекулярной ветвистости
Kappa1, Kappa2, Kappa3	Индексы формы молекулы (каппа-индексы).
HallKierAlpha	Индекс, учитывающий гибридизацию и топологию.

6. Поверхностные и объёмные свойства

Дескриптор	Описание
LabuteASA	Оценка молекулярной поверхности.
PEOE_VSA*, SMR_VSA*, SlogP_VSA	Дескрипторы, связывающие заряд, полярность и растворимость с поверхностью.
EState_VSA*, VSA_EStat	Комбинация электронных индексов и поверхностных свойств.

7. Функциональные группы

Дескриптор	Описание
Fr_AL_OH**	Бинарные или количественные признаки наличия функциональных групп
fr_amide, fr_ester, fr_keton	Признаки амидов, сложных эфиров, кетонов

8. Структурные особенности

Дескриптор	Описание
NumHAcceptors, NumHDonors	Число акцепторов и доноров водородных связей.
NumRotatableBonds	Число вращающихся связей (гибкость молекулы).
RingCount	Общее число циклов.

Дескриптор	Описание
NumArsomaticRings, NumAliphaticRing	Число ароматических и алифатических циклов.
HeavyAtomCount	Число тяжелых атомов (не водород).

9. ADME-свойства (Absorption Distribution , Metabolism , Excretion)

Дескриптор	Описание
SPS	Оценка сложности синтеза (1 = легко, 10 = сложно).
NHOHCount, NOCount	Число NH/OH-групп и N/O-атомов.
NumHeteroatoms	Число гетероатомов (N, O, S, P и др.).

Проверка данных на дубликаты и пропуски.

Был удален признак "Unnamed: 0" так как он дублировал индексы строк и мог помешать найти дубликаты.

Были найдены строки с пропусками. Поскольку их количество было небольшим, было принято решение их удалить. Так же были удалены полные дубликаты строк.

Перерасчет значения SI.

Исходя из того, что CC50 и IC50 были собраны экспериментальным путем, а SI рассчитывается исходя из них, был произведен перерасчет значений SI.

Оценка наличия выбросов в целевых переменных.

Когда мы строили коробчатые диаграммы, то заметили, что некоторые значения выглядят как выбросы. Но просто удалить все выбросы - плохая идея, потому что мы можем случайно выкинуть важную информацию. Каждый аналитик должен сам определять границы выбросов, поскольку каждая задача имеет свою специфику, и следовательно индивидуальный подход. Поэтому я решил разобраться, какие значения действительно можно считать ошибочными или нерелевантными.

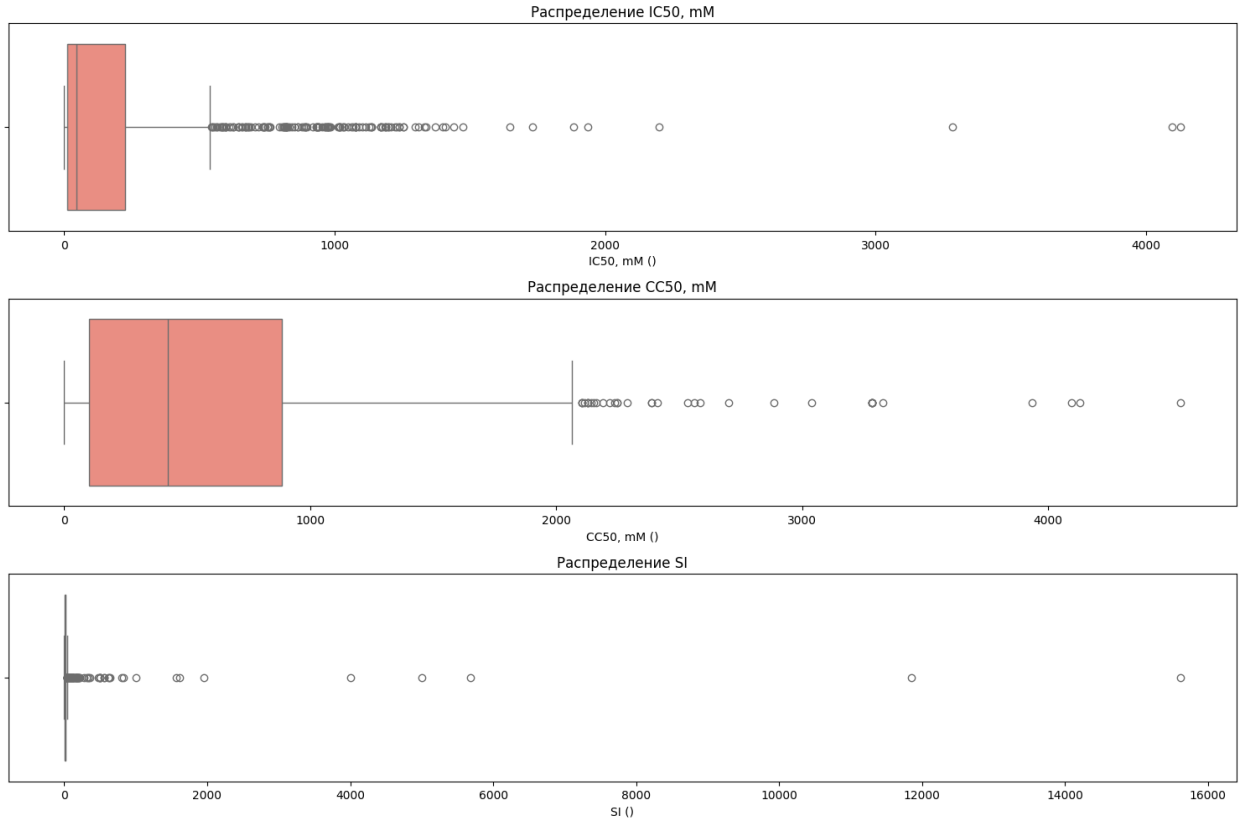

Рисунок 1. Коробчатые диаграммы для целевых переменных

IC50:

- Была найдена информация, что обычно IC50 не превышает 1000 mM (это очень большая концентрация)
- Если вещество имеет IC50 > 1000 mM, значит оно практически не работает (нужна огромная доза для эффекта)
- Поэтому считаем значения IC50 > 1000 mM выбросами

CC50:

- Не смог найти четких критериев, какие значения CC50 считать выбросами
- Токсичность может быть разной, поэтому оставил все значения CC50 без изменений

SI:

- Значения SI > 1000 встречаются крайне редко
- Такие высокие значения означают, что вещество почти нетоксично (CC50 огромное), но при этом слабо активно (IC50 маленькое)
- Скорее всего, это ошибки измерений или очень специфические случаи
- Поэтому SI > 1000 тоже посчитал выбросами

Удаляем:

- IC50 > 1000 mM (слишком неактивные вещества)
- SI > 1000 (маловероятные значения)

Оценка распределения целевых переменных.

Оценка распределения целевой переменной — это ключевой шаг в предобработке данных, так как от его характера зависит выбор модели, корректность выводов и точность прогнозов. Если данные имеют сильный перекос, выбросы или несколько пиков, это может негативно сказаться на работе алгоритмов. Например, методы, предполагающие нормальность распределения, такие как линейная регрессия, могут давать ошибочные результаты при нарушении этого условия.

Понимание особенностей распределения помогает правильно подготовить данные: возможно, потребуется преобразование переменной, устранение выбросов или выбор более подходящей модели, устойчивой к отклонениям от нормальности. Это особенно важно в задачах, где точность прогнозирования критична, например, в финансовой аналитике или медицинских исследованиях. Таким образом, тщательная проверка распределения позволяет избежать ошибок и улучшить качество модели.

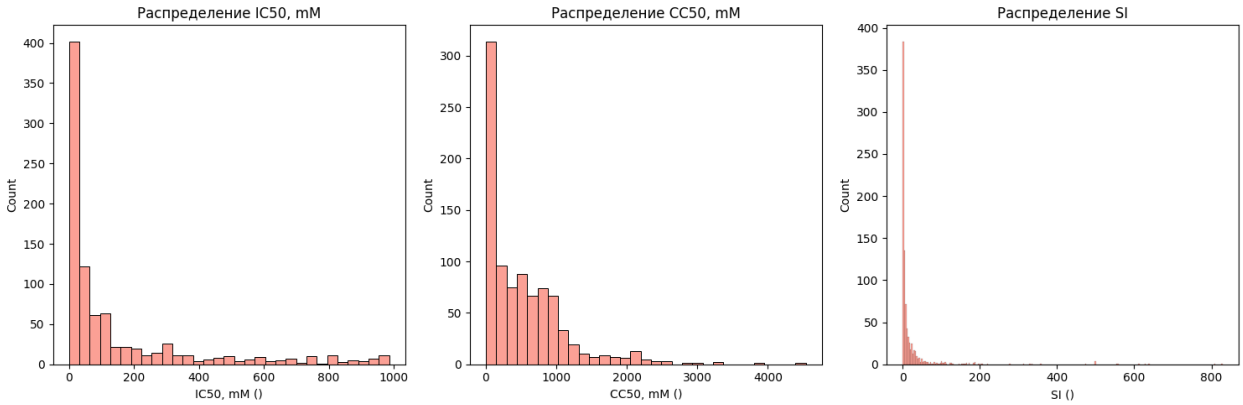

Рисунок 2. Диаграммы распределения для целевых переменных

Распределение данных похоже скорей на геометрическое. Принято решение постараться привести данные к нормальному распределению с помощью логарифмирования.

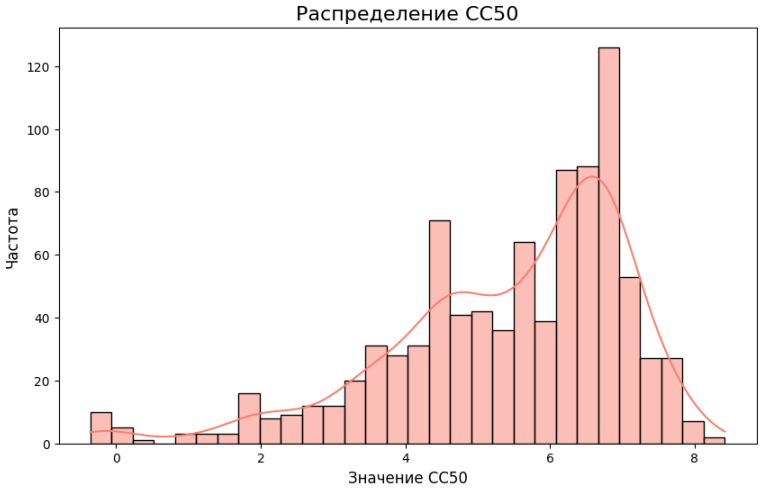

Рисунок 3. Распределение CC50 после логарифмирования.

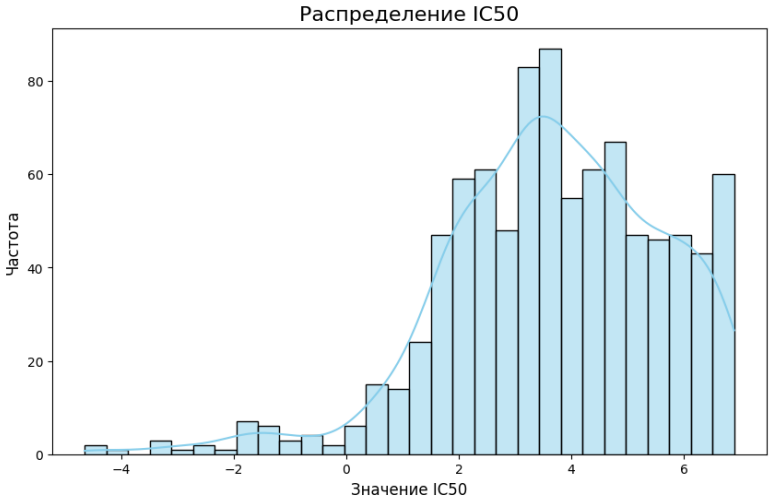

Рисунок 4. Распределение IC50 после логарифмирования.

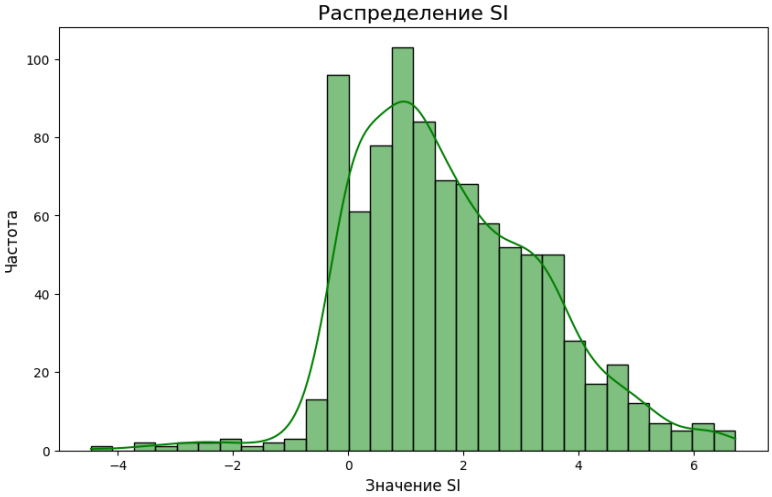

Рисунок 5. Распределение SI после логарифмирования.

Конструирование новых признаков и отбор существующих.

Удаление нулевых признаков.

В первую очередь были удалены признаки, в которых все значения были равны 0. Так как никакой полезной информации для моделей они не несут.

Группировка признаков по липофильности/гидрофильности

Для систематизации признаков, связанных с распределением электронной плотности и полярностью атомов, мы объединили их в три основные группы:

- **Гидрофильные атомы** ($SlogP < 0$):
  SlogP_VSA1, SlogP_VSA2, SlogP_VSA3
- **Умеренно полярные атомы** ($0 \leq SlogP < 0.3$):
  SlogP_VSA4, SlogP_VSA5, SlogP_VSA6
- **Липофильные атомы** ($SlogP \geq 0.3$):
  SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SlogP_VSA10

Также были выделены **граничные значения** (SlogP_VSA11, SlogP_VSA12), требующие дополнительного анализа.

Анализ корреляции и фильтрация признаков

Для исключения мультиколлинеарности были построены **матрицы корреляции** для фрагментных и электронных дескрипторов. На основе этого анализа удалены наиболее скоррелированные признаки, что позволило снизить избыточность данных и улучшить интерпретируемость модели.

Создание новых признаков

Для более детального описания электронных свойств молекул были введены два новых признака:

- **DiffPartialCharge** – разница в парциальных зарядах атомов, отражающая полярность связей.
- **DiffEStateIndex** – разница в индексах электронной плотности, характеризующая распределение электронов.

Эти признаки могут быть полезны для прогнозирования реакционной способности и физико-химических свойств соединений.

Старые признаки при этом не удалялись, поскольку сложно оценить какая комбинация признаков в итоге даст наилучший результат.

Отбор значимых признаков

Для выбора наиболее информативных переменных использовался метод **Mutual Information Regression**, который оценивает степень зависимости между признаками и целевой переменной.

Были созданы отдельные копии данных для каждой целевой переменной, что позволило провести **индивидуальный отбор признаков** в зависимости от задачи (регрессия или классификация). Это повысило релевантность модели и снизило риск переобучения. Для каждой модели была подобрана своя граница.

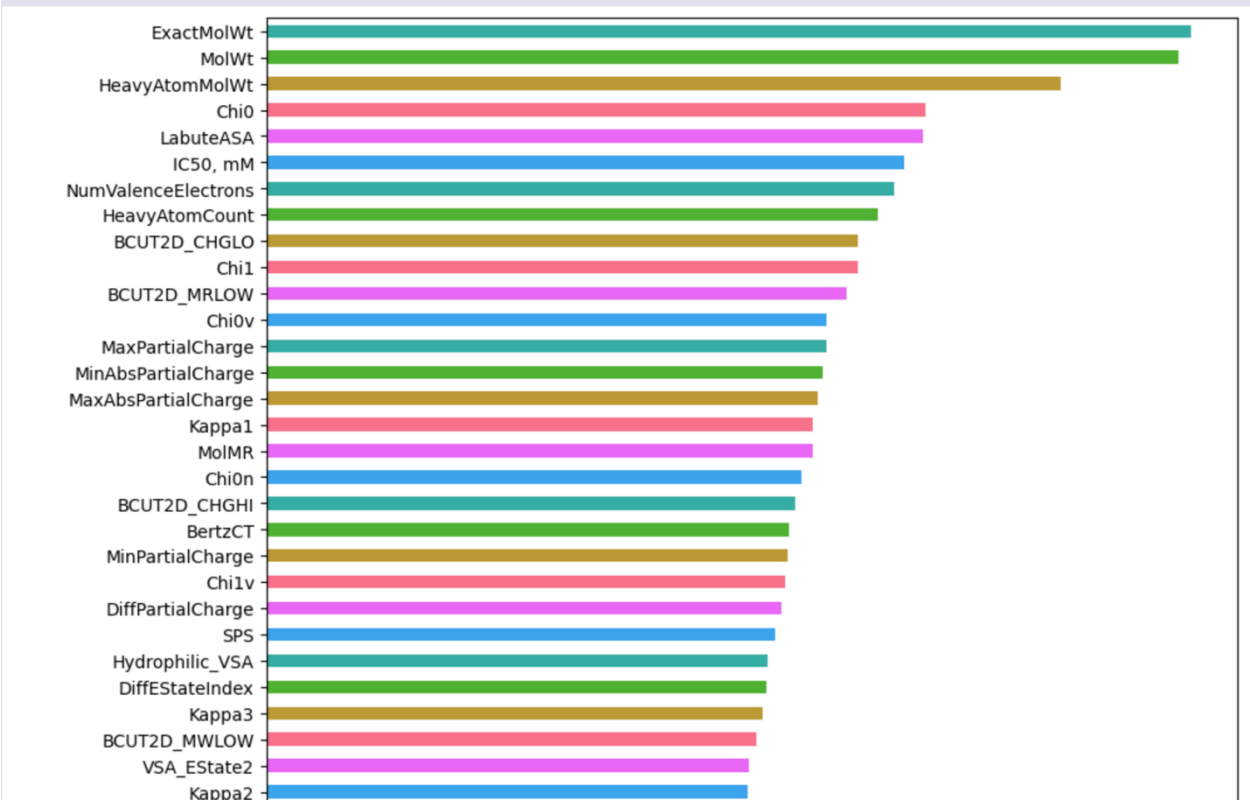

Рисунок 6. Отбор признаков с помощью **Mutual Information Regression**

Модели

Как и в большинстве задач машинного обучения, не существует единственного оптимального алгоритма, гарантирующего наилучший результат. Поэтому для достижения максимальной точности прогнозирования было сделано:

Обработка данных в моделях.

Произведено разделение данных на тестовую и тренировочную выборку, чтобы исключить просачивание данных и как следствие необоснованное улучшение метрик моделей.

Поскольку обработка данных для каждой модели немного отличается, а также данные разделены. Были созданы **конвейеры обработки** тренировочных и тестовых данных, для каждой модели учитывая анализ данных в EDA.

Модели регрессии

Были протестированы различные модели:

Линейная регрессия
Базовый линейный метод, предсказывающий целевую переменную как линейную комбинацию входных признаков

RandomForestRegressor (случайный лес)
Ансамблевый метод, использующий множество решающих деревьев, обученных на случайных подмножествах данных и признаков

SVR (метод опорных векторов для регрессии)
Использует принцип максимизации ε-трубки, в пределах которой допускаются ошибки предсказания

GradientBoostingRegressor (градиентный бустинг)
Ансамблевый метод, последовательно улучшающий предсказания за счет добавления новых деревьев, исправляющих ошибки предыдущих

Произведен подбор гиперпараметров с помощью Байесовской оптимизации.

Произведена оценка метрик по таким параметрам:

- **MAE** (Средняя абсолютная ошибка)
- **MSE** (Средняя квадратичная ошибка)
- **RMSE** (Корень из MSE)
- **R^2** (Коэффициент детерминации)

Подобранные модели регрессии:

Целевая переменная	Выбранная модель	Подобранные гиперпараметры	Метрики
IC_50	GradientBoosting	learning_rate: 0.01 max_depth: 3 max_features: 0.429374458026054 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 1200	MAE: 1.06 MSE:1.93 RMSE: 1.39 R^2: 0.58
CC_50	RandomForest	max_depth: 18 max_features: 0.361554586956777 min_samples_leaf: 1 min_samples_split: 10 n_estimators: 1000	MAE: 0.742 MSE:1.03 RMSE: 1.071 R^2: 0.56
SI	RandomForest	max_depth: 14 max_features: 0.6447047672509045 min_samples_leaf: 1 min_samples_split: 10 n_estimators: 100	MAE: 0.742 MSE:1.12 RMSE: 1.06 R^2: 0.64

Выводы по метрикам:

IC50:

Модель демонстрирует наилучшую предсказательную способность среди тестируемых алгоритмов, однако значительная величина ошибок (RMSE) указывает на существующий разброс в предсказаниях.

CC50:

Качество предсказаний сопоставимо с моделью для IC50, при этом средняя абсолютная ошибка находится на приемлемом уровне.

SI:

Данная модель показала наилучшие результаты, что может быть связано с более выраженными закономерностями в данных для этого параметра.

Все модели демонстрируют умеренно-хорошее качество предсказаний (R^2 в диапазоне 0.56-0.64), однако существует потенциал для улучшения:

- Худшие метрики наблюдаются при предсказании IC50 (наибольшие значения ошибок)
- Наилучшие результаты достигнуты для SI

Направления для улучшения:

- Углубленный анализ признаков с целью исключения выбросов.
- Более детальное составление новых признаков.
- Отбор признаков на основе других алгоритмов, например Boruta и др...

Модели классификации:

Были протестированы различные модели:

Логистическая регрессия
Оценивает вероятность принадлежности к классу с помощью логистической функции (сигмоиды)

KNN
Классифицирует объект по преобладающему классу среди k ближайших соседей

Random Forest (Случайный лес)
Совокупность решающих деревьев, каждое из которых обучается на случайном подмножестве данных и признаков

CatBoost
Последовательно строит деревья, где каждое новое дерево корректирует ошибки предыдущих с учетом категориальных признаков

Была произведена балансировка классов, где это было необходимо, с помощью ADASYN.

Произведен подбор гиперпараметров с помощью Байесовской оптимизации.

Произведена оценка метрик по таким параметрам:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**
- **AUC-ROC**

Целевая переменная	Выбранная модель	Подобранные гиперпараметры	Метрики
IC50_class	KNN	n_neighbors: 45 p: 1	Accuracy: 0.7422 Precision: 0.7563 Recall: 0.7087 F1-score: 0.7317 AUC-ROC: 0.7746
CC50_class	RandomForest	max_depth: 5 max_features: 0.1 min_samples_leaf: 3 min_samples_split: 3 n_estimators: 100	Accuracy: 0.7188 Precision: 0.6909 Recall: 0.8444 F1-score: 0.7600 AUC-ROC: 0.7907
SI_class	RandomForest	max_depth: 10 max_features: 0.1359683207862248 min_samples_leaf: 4 min_samples_split: 2 n_estimators: 736	Accuracy: 0.6162 Precision: 0.6228 Recall: 0.5379 F1-score: 0.5772 AUC-ROC: 0.6579
SI_8	RandomForest	max_depth: 11 max_features: 0.6995058306553149 min_samples_leaf: 1 min_samples_split: 2 n_estimators: 982	Accuracy: 0.7106 Precision: 0.5814 Recall: 0.5376 F1-score: 0.5587 AUC-ROC: 0.7045

Выводы по метрикам:

Модель для IC50_class (KNN)

Модель демонстрирует сбалансированное качество классификации с хорошей точностью (Precision) и приемлемой полнотой (Recall). Значение AUC-ROC (0.775) указывает на удовлетворительную способность различать классы.

Модель для CC50_class (Random Forest)

Алгоритм показывает высокую полноту Recall (0.844), что особенно важно для минимизации ложноотрицательных результатов. При этом сохраняется приемлемая точность предсказаний.

Модель для SI_class (Random Forest)

Качество классификации несколько ниже предыдущих моделей. Низкий Recall (0.538) указывает на трудности с выявлением положительного класса.

Модель для SI_8 (Random Forest):

Приемлемая общая точность. Низкие Precision и Recall свидетельствуют о несбалансированности классов, несмотря на использование ADASYN. AUC-ROC (0.705) указывает на удовлетворительное качество модели

- Для IC50 оптимальным оказался KNN с лучшей точностью (Precision=0.756)
- Для CC50 RandomForest показал наивысшую полноту (Recall=0.844)
- Модели демонстрируют AUC-ROC в диапазоне 0.658-0.791, что указывает на удовлетворительное качество классификации

Направления для улучшения:

- Углубленный анализ признаков с целью исключения выбросов.
- Более детальное составление новых признаков.
- Отбор признаков на основе других алгоритмов, например Boruta и др..
- Подбор других способов балансировки классов.
- Применение других моделей классификации
- Применение нейросетей.

Заключение

В ходе выполнения данной работы была успешно реализована основная цель - разработка моделей машинного обучения для предсказания молекулярных свойств с целью оптимизации поиска перспективных лекарственных соединений. Проведенное исследование включало несколько ключевых этапов:

1. **Предобработка данных**:
   - Проведен анализ признаков
   - Устранены пропущенные значения и дубликаты
   - Выполнена фильтрация выбросов
   - Проверено распределение целевых переменных
     Эти меры позволили значительно повысить качество исходных данных.
2. **Работа с признаками**:
   - Созданы новые информативные характеристики
   - Выполнен отбор наиболее значимых признаков
   - Построены матрицы корреляции для исключения мультиколлинеарности
3. **Разработка и тестирование моделей**:

   - Созданы конвейеры для каждой модели
   - Для регрессионных задач протестированы: линейная регрессия, SVR, GradientBoosting и RandomForest
   - Для задач классификации исследованы: KNN, RandomForest и CatBoost
   - Для каждой модели выполнена оптимизация гиперпараметров
   - Проведена объективная оценка с использованием кросс-валидации

Разработанные модели продемонстрировали удовлетворительные результаты:

- Для регрессионных задач достигнуты значения R^2 в диапазоне 0.56-0.64
- Классификаторы показали accuracy до 0.742 и AUC-ROC до 0.791
- Особенно хорошие результаты получены для моделей CC50 и IC50

Направления для улучшения:

- Применение более сложных архитектур (нейронные сети, ансамбли)
- Дальнейшая оптимизация гиперпараметров
- Улучшение feature engineering:
  - Добавление новых молекулярных дескрипторов
  - Использование методов генерации признаков
- Реализация методов работы с несбалансированными данными

Полученные результаты могут быть использованы для ускорения и оптимизации процесса разработки новых лекарственных препаратов. В тоже время анализ показал, что модели можно попытаться улучшить.