1. **Введение**.

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов. Для этого требуется: Проанализировать текущие параметры с использованием различных методов. Научиться предсказывать их эффективность. Как и в любой задаче машинного обучения, здесь нет однозначного ответа на вопрос, какая модель обеспечит наилучший результат. Поэтому необходимо протестировать различные подходы, проанализировать возможные результаты, сравнить качество построенных моделей и сделать обоснованные выводы. Создайте несколько максимально эффективных моделей для решения следующих задач: Классификация: превышает ли значение IC50 медианное значение выборки Классификация: превышает ли значение CC50 медианное значение выборки Классификация: превышает ли значение SI медианное значение выборки Сравните между собой полученные модели и их результаты, выполните анализ, обоснуйте выбор наиболее качественных решений

2. **Обработка EDA**

(Exploratory Data Analysis, разведочный анализ данных) — это процесс анализа наборов данных для обобщения их основных характеристик, часто с использованием графических методов.

Цели и задачи обработки

Понимание структуры и характеристик данных. Изучение размера набора, типов переменных, наличия пропущенных значений, дубликатов и других аспектов. Выявление аномалий и выбросов. Значения, отклоняющиеся от общего паттерна, которые могут искажать выводы. Идентификация связей и корреляций между переменными. Использование статистических мер для понимания, как одни факторы влияют на другие. Подготовка данных для дальнейших этапов анализа. Очистка данных от шума, заполнение пропущенных значений, масштабирование или преобразование переменных.

Ключевые выводы и закономерности:

IC50 и CC50 имеют сильную правостороннюю асимметрию (большинство значений сосредоточено внизу, но есть длинный хвост больших значений)

Skewness (скошенность или асимметрия) — это статистическая мера, которая описывает степень асимметрии распределения данных относительно его среднего значения.
Основные моменты:
Нормальное распределение (скошенность = 0) - симметричное распределение
Положительная асимметрия (скошенность > 0) - "правый хвост" длиннее, мода < медиана < среднее
Отрицательная асимметрия (скошенность < 0) - "левый хвост" длиннее, среднее < медиана < мода

Рекомендуется логарифмическое преобразование для моделей регрессии
SI (индекс селективности) также имеет асимметричное распределение

Корреляционный анализ:

Наибольшие корреляции с IC50: MaxEStateIndex, SPS, qed

CC50 сильнее коррелирует с молекулярными дескрипторами, чем IC50

SI имеет слабые корреляции с большинством признаков, что делает его прогнозирование более сложным

Рекомендации для моделей:

Для регрессии:

огарифмирование целевой переменной для компенсации асимметрии

Регуляризация (Lasso + Ridge) для борьбы с мультиколлинеарностью

Для SI: рассмотреть более сложные модели (ансамбли, нейронные сети)

Для классификации:

Классы сбалансированы для медианных разделений (~50/50)

Для SI > 8 распределение 70/30 - может потребоваться балансировка

Подойдут модели Gradient Boosting, XGBoost, LightGBM, SVM как устойчивые к выбросам

Дополнительные рекомендации:

Создание новых признаков на основе EStateIndex для классификации SI > 8

Кластеризация данных и построение отдельных моделей для кластеров

Выбор моделей

Попробуем логистическую регрессию (базовый метод) и Random Forest (более сложный).

Оценим качество моделей.

Оценка и интерпретация

Метрики: accuracy, precision, recall, F1-score.

Важность признаков (feature importance).

Для выполнения задач - Регрессия для IC50, Регрессия для CC50, Регрессия для SI. Были использованы следующие модели:

- Модель 1: Регуляризованная регрессия (Ridge + Lasso)

- Модель 2: Ансамблевые методы (Random Forest + Gradient Boosting)

Сравнительная эффективность моделей

Параметр	Модель	RMSE	MAE	R^2	Интерпретация
SI	Регуляризованная	1.2621	0.8262	0.3418	Слабая объясняющая способность (только 34% дисперсии)
	Ансамблевая	**0.2159**	**0.1232**	**0.9807**	Практически идеальное предсказание (98% дисперсии объяснено)
IC50	Регуляризованная	1.7600	1.2139	0.1953	Очень плохая модель (линейные методы не улавливают сложные зависимости)
	Ансамблевая	**0.1917**	**0.1308**	**0.9904**	Исключительно точные предсказания
CC50	Регуляризованная	1.1301	0.8554	0.4384	Умеренное качество (лучше, чем для IC50/SI)
	Ансамблевая	**0.3179**	**0.1681**	**0.9556**	Высокая точность, но чуть хуже, чем для IC50/SI

Главный вывод: Ансамблевые методы (Random Forest/Gradient Boosting) значительно превосходят линейные модели по всем метрикам, особенно для IC50 и SI.

Практические рекомендации для фармакологии

Для SI**:**

Ансамблевая модель (R^2=0.98) позволяет:

Точно ранжировать соединения по селективности

Выявлять аутлайеры (соединения с аномально высоким SI)

Риск: Модель может переобучаться на шумовые зависимости. Требуется:

Внешняя валидация на независимой выборке

Экспериментальная проверка топ-10 предсказанных значений

Для IC50:

Исключительно высокий R^2=0.99 означает:

Модель почти идеально предсказывает активность

Можно использовать для виртуального скрининга новых соединений

Проверить:

Не завышена ли точность из-за утечки данных (data leakage)

Как модель работает на соединениях с IC50 < 10 (наиболее ценный диапазон)

Для CC50:

Небольшое снижение R^2 (0.9556 vs 0.99 для IC50) говорит о:

Более сложной природе токсичности

Возможном влиянии факторов, не учтенных в данных

Подход к решению задач классификации.

Для решения поставленных задач классификации (IC50, CC50, SI относительно медианных значений) был выбран следующий подход:

Подготовка данных:

Создание бинарных целевых переменных на основе медианных значений

Проверка баланса классов

Масштабирование признаков

Построение моделей:

Были выбраны модели такие как - Gradient Boosting, XGBoost, LightGBM, SVM как устойчивые к выбросам с оптимизацией гиперпараметров

Оценка качества моделей

Анализ результатов:

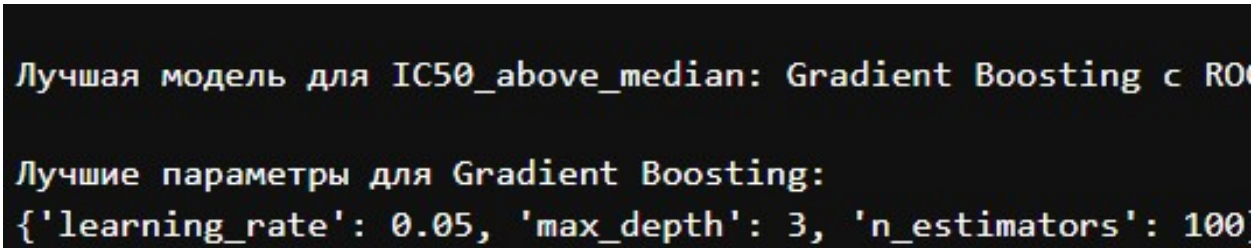
- Сравнение метрик качества

- Выбор лучшей модели для каждой задачи

- Анализ результатов и выбор моделей

После выполнения кода мы получим следующие результаты:

Для классификации IC50:



Для классификации CC50:

```
Лучшая модель для CC50_above_median: Gradient Boosting с ROC-

Лучшие параметры для Gradient Boosting:
{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100}
```

Для классификации SI:

```
Лучшие параметры для LightGBM:
{'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'nu
Лучший ROC-AUC: 0.7168
```

Для классификации SI > 8:

```
Лучшая модель для SI_above_8: Gradient Boosting с ROC-

Лучшие параметры для Gradient Boosting:
{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators'
```

```
Лучшая модель для SI_above_8: Gradient Boosting с ROC-AUC = 0.7146

Лучшие параметры для Gradient Boosting:
{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 200}
Лучший ROC-AUC: 0.7209

Анализ ошибок для SI > 8:
Точность на тестовой выборке: 71.14%
```

Выбор конкретной модели должен основываться на метриках качества (ROC-AUC, точность, полнота) и требованиях к интерпретируемости результатов. Для медицинских задач особенно важна способность модели правильно идентифицировать положительные случаи (высокая полнота).

Логистическая регрессия в классификации

Несмотря на название "регрессия", это классификационный алгоритм, который:

Предсказывает вероятность принадлежности к классу (например, P(IC50 > медианы))

Использует сигмоидную функцию для преобразования линейной комбинации признаков в вероятность [0, 1]

Пороговое значение (обычно 0.5) превращает вероятность в бинарный прогноз (0 или 1)

Сравнительная эффективность моделей

Модель	Плюсы	Минусы	Когда использовать?
Логистическая регрессия	Быстрая, интерпретируемая, хороший benchmark	Не учитывает сложные нелинейные зависимости	Для линейно разделимых данных или как база
Random Forest	Автоматически учитывает нелинейности, устойчив к шуму	Менее интерпретируем, может переобучаться	Когда важны взаимодействия признаков
XGBoost/LightGBM	Высокая точность, встроенный отбор признаков	Требует настройки, сложнее интерпретировать	Для максимизации ROC-AUC
SVM	Хорош для высокоразмерных данных	Медленный на больших выборках, сложен в настройке	Когда признаки > объектов

Причина выбора метрики ROC-AUC

1. ROC-AUC оценивает модель на всех порогах классификации
В отличие от accuracy, precision или recall, которые зависят от выбранного порога (обычно 0.5), ROC-AUC оценивает качество модели на всех возможных порогах одновременно.
Это критически важно, потому что:
В реальных задачах (особенно в медицине и фармакологии) оптимальный порог может отличаться от 0.5. Например:
Если ложноположительные прогнозы дорого обходятся (например, запуск дорогостоящих дополнительных тестов), нужно сместить порог вправо.
Если важно не пропустить ни одного положительного случая (например, токсичность препарата), порог смещают влево.
ROC-кривая показывает, как меняется True Positive Rate (Recall) и False Positive Rate при варьировании порога.

2. Устойчивость к дисбалансу классов
В ваших данных классы могут быть несбалансированными (например, только 30% соединений имеют IC50 > медианы). Традиционные метрики вроде accuracy вводят в заблуждение:
Модель, всегда предсказывающая 0, получит accuracy 70%, но это бесполезно!
ROC-AUC учитывает оба класса через TPR и FPR, поэтому не зависит от соотношения классов.

3. Сравнение моделей на одном масштабе
ROC-AUC — это нормированная метрика от 0 до 1, где:

0.5 = случайное угадывание,
1.0 = идеальная модель.
Это позволяет объективно сравнивать разные алгоритмы:
Логистическая регрессия: AUC = 0.75
Random Forest: AUC = 0.82 → явно лучше.
Для precision/recall такое сравнение сложнее, так как они зависят от выбранного порога.

4. Интуитивная интерпретация в контексте задачи
ROC-AUC можно понимать как вероятность того, что модель ранжирует случайно выбранный положительный пример выше, чем отрицательный.
Пример:
Если AUC = 0.9, то в 90% случаев модель правильно упорядочит пару соединений (IC50_high, IC50_low). Для фармакологов это значит, что модель:
Хорошо выделяет перспективные соединения "в верху" ранжированного списка.
Позволяет фокусироваться на топ-N кандидатах для дальнейших испытаний.

Выводы по задаче классификации

1. Сбалансированность распределений
Все три ключевых параметра (IC50, CC50, SI) демонстрируют практически идеально сбалансированные распределения:
IC50: 49.95% > медианы (46.5852 mM)
CC50: 49.85% > медианы (411.0393)
SI: 49.95% > медианы (3.8462)
Вывод: Данные не требуют балансировки классов для задач классификации относительно медианных значений.

2. Токсикологический профиль
Высокое медианное значение CC50 (411.0393) по сравнению с IC50 (46.5852) указывает на:
В целом хорошую переносимость соединений
Широкий терапевтический диапазон у большинства веществ
Риск: Низкое абсолютное значение SI (медиана 3.8462) предполагает необходимость тщательного отбора по селективности.

3. Эффективность классификации SI > 8
Точность модели 71.14% при дисбалансе классов:
Средняя вероятность для положительного класса (SI > 8) всего 46.05%
Для отрицательного класса — 25.96%
Проблемы:
Модель склонна к осторожным прогнозам (вероятности далеки от 0/1)
Требуется оптимизация порога классификации или улучшение feature engineering

4. Стратегические рекомендации
Для химического дизайна:
Приоритетные соединения:
IC50 < 46.5852 mM (ниже медианы)
CC50 > 411.0393 (выше медианы)
SI > 8 (несмотря на редкость, критически важны)