Национальный исследовательский ядерный университет «МИФИ»

Классическое машинное обучение

Курсовая работа

«Активность соединений лекарственных препаратов»

Студент:

Астанин Денис Васильевич

ВВЕДЕНИЕ

Данная курсовая работа посвящена применению современных методов машинного обучения для анализа биологической активности и токсичности химических соединений. Основная цель исследования - разработка предиктивных моделей для определения токсикологических параметров: цитотоксичности (CC50), активности ингибирования (IC50) и селективности (SI).

Актуальность работы обусловлена необходимостью создания эффективных инструментов для предварительной оценки токсичности новых химических соединений, что позволит сократить количество лабораторных экспериментов и ускорить процесс разработки безопасных препаратов.

1. РАЗВЕДЫВАТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ (EDA)

1.1 Аналитика данных

1.1.1 Описание данных

Была получена первичная информация, благодаря которой выяснилось, что исследуемый датасет содержит информацию о биологической активности химических соединений и обладает следующими характеристиками:

- Общий размер: 1,001 образец × 214 признаков
- Общее количество элементов: 214,214
- Типы данных:
  - float64: 107 столбцов (50.2%)
  - int64: 106 столбцов (49.8%)

Целевые переменные:

1. IC50 (mM) - концентрация полумаксимального ингибирования.
2. CC50 (mM) - концентрация полумаксимальной цитотоксичности.
3. SI - индекс селективности (Selectivity Index)

Также, были выявлены следующий проблемы:

1. Данные имеют разные масштабы - потребуется нормализация/стандартизация.
2. Многие признаки имеют ненулевую асимметрию (skewness).
3. Значения kurtosis указывают на различия в "тяжести хвостов" распределений.
4. Большая вариативность в значениях дисперсии между признаками

1.1.2 Первичный анализ данных

Был проведён первичный анализ данных для определения качества данных:

1. Пропущенные значения: 36 (0.02% от общего объёма).
2. Количество столбцов с пропусками: 12 из 214.
3. Проблемные признаки: MaxPartialCharge, MinPartialCharge, BCUT2D_.

Признак	Количество пропусков	Процент пропусков
MaxPartialCharge	3	0.3%
MinPartialCharge	3	0.3%
MaxAbsPartialCharge	3	0.3%
MinAbsPartialCharge	3	0.3%
BCUT2D_MWHI	3	0.3%
BCUT2D_MWLOW	3	0.3%

Далее были обработаны пропуски, путём удаления проблемных строк. Анализ данных до и после очистки пропусков:

1. Записей до очистки: 1,001
2. Записей после очистки: 998
3. Удалено записей: 3 (0.30%)
4. Пропущенных значений осталось: 0

Затем, был поиск дублированных записей по полному совпадению. Затем было проведено удаление полных дубликатов. Анализ данных до и после удаления:

1. Полных дубликатов: 32
2. Дубликатов по признакам (без учета целевых): 196
3. Записей до удаления дубликатов: 998
4. Записей после удаления дубликатов: 966
5. Удалено дубликатов: 32
6. Общее сокращение данных: 3.50%

1.1.3 Анализ распределения целевых переменных

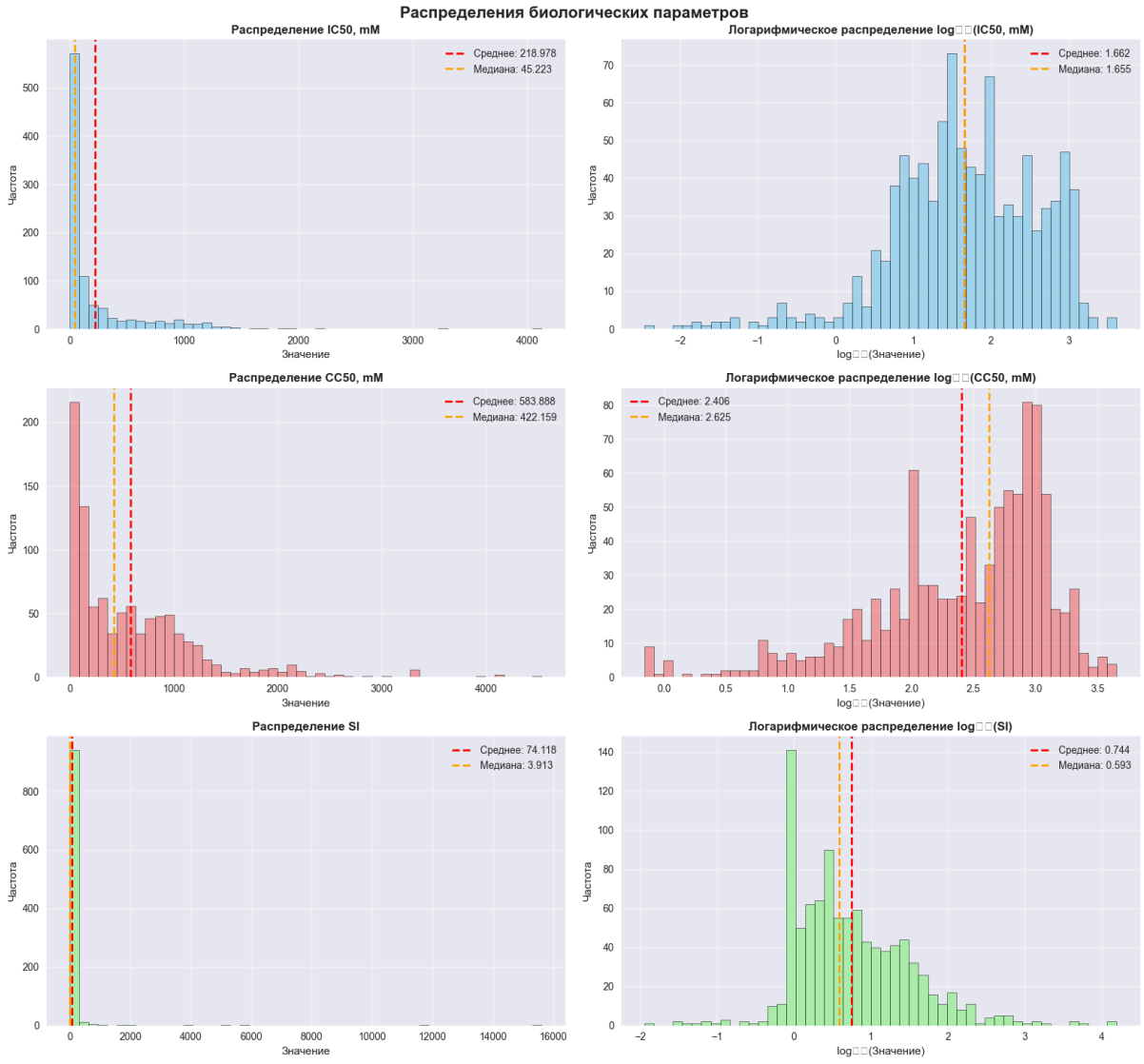


Рисунок 1.1: Распределения биологических параметров (обычное и логарифмическое)

Были получены и визуализированы с помощью matplotlib (см. рисунок 1.1) характеристики распределений:

Переменная	Тип распределения	Асимметрия	Рекомендация
IC50, mM	Логнормальное	Высокая	Логарифмирование
CC50, mM	Логнормальное	Высокая	Логарифмирование
SI	Логнормальное	Высокая	Логарифмирование

Исходя из полученных данных, можно сделать следующие выводы по распределениям:

1. Все целевые переменные имеют логнормальное распределение.
2. Логарифмическое преобразование значительно улучшает нормальность.
3. Присутствует большое количество выбросов.

1.1.4 Корреляционный анализ целевых и нецелевых признаков

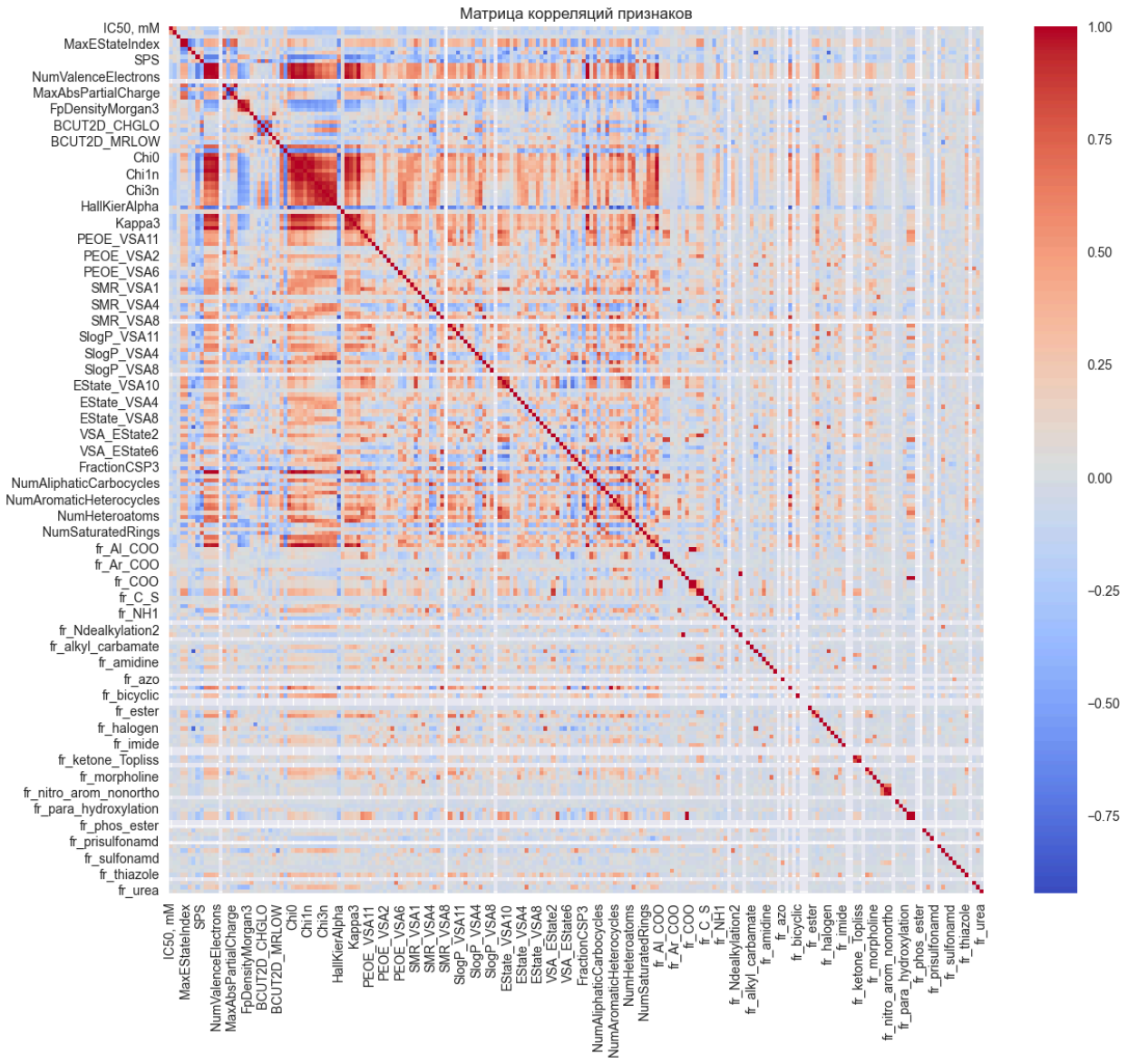


Рисунок 1.2: Матрица корреляций признаков

Анализ слабо коррелированных признаков показал:

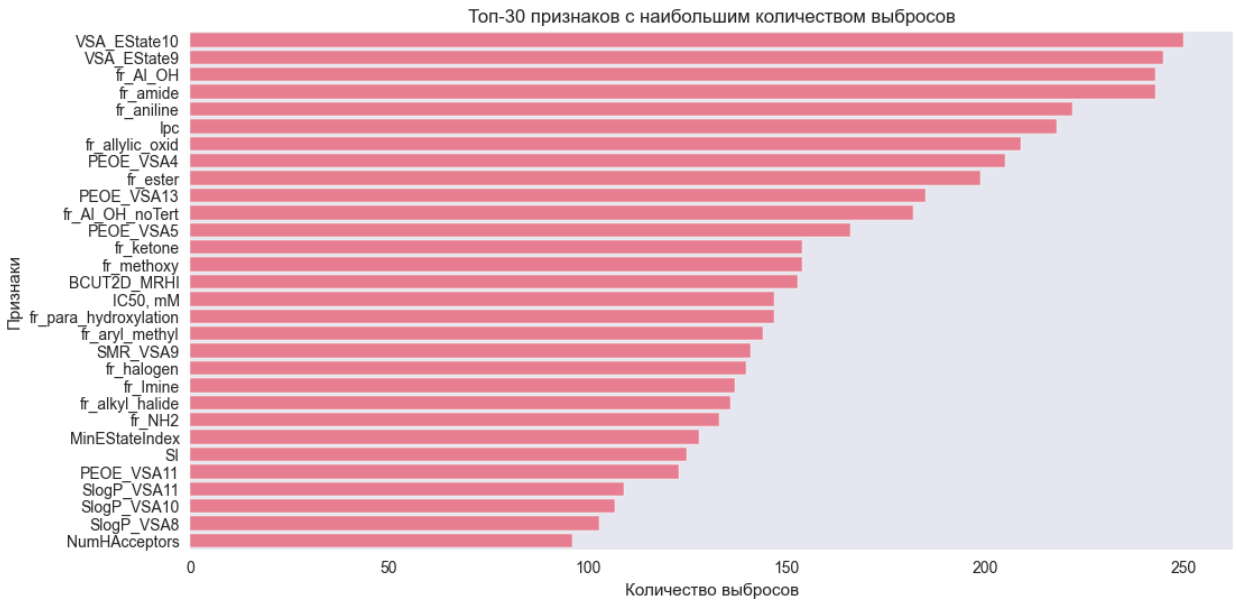
1. IC50: 11 признаков с корреляцией < 0.01 - CC50: 14 признаков с корреляцией < 0.01.
2. SI: 33 признака с корреляцией < 0.01 - Общее количество слабо коррелированных признаков: 47.

Также, были обнаружены пары признаков с корреляцией > 0.9:

1. MaxAbsEStateIndex и MaxEStateIndex: 1.00.
2. MolWt и HeavyAtomMolWt: 1.00.
3. MolWt и ExactMolWt: 1.00.
4. NumValenceElectrons и Chi0: 0.99.
5. И другие (всего более 50 пар)

1.2 Предобработка данных

1.2.1 Анализ выбросов



Выбросы по признакам Рисунок 1.4: Топ-30 признаков с наибольшим количеством выбросов

Статистика выбросов по правилу IQR:

Признак	Количество выбросов
Ipc	218
PEOE_VSA4	205

PEOE_VSA13	185
PEOE_VSA5	166
BCUT2D_MRHI	153
IC50, mM	147
SMR_VSA9	141
MinEStateIndex	128
SI	125
PEOE_VSA11	123

Для обнаружения выбросов использовалось правило 3-х сигм:

$$\mu \pm 3\sigma \; 2$$

Межквартильный размах (IQR):

$$Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR$$

1.2.2 Отбор признаков

Были выбраны следующие критерии исключения признаков:

1. Нулевая дисперсия: признаки с var() = 0.
2. Высокая корреляция: $r > 0.9$ между парами признаков.
3. Слабая связь с целевыми переменными: $|r| < 0.01$

В результате исходное количество признаков составило 213. После удаления высококоррелированных происзошло значительное уменьшение размерности.

1.2.3 Стандартизация данных

Была применена стандартизация (StandardScaler) к датасету:

$$X\_scaled = (X - \mu) / \sigma$$

В результате был создан файл processed_data_full.parquet.

Выводы по EDA

1. Качество данных высокое - минимальное количество пропусков (0.02%).
2. Логарифмическое преобразование целевых переменных значительно улучшило их распределение.
3. Мультиколлинеарность устранена удалением высококоррелированных признаков.
4. Выбросы обработаны по правилу 3-х сигм.
5. Стандартизация обеспечила сопоставимость масштабов признаков.

2. ОПИСАНИЕ ЭКСПЕРИМЕНТОВ ПО ЗАДАЧАМ МАШИННОГО ОБУЧЕНИЯ

2.1 Задачи регрессии

2.1.1 regression_cc50.ipynb - Предсказание CC50

Предобработка данных:

1. Логарифмическое преобразование: $\log_{10}(CC50)$.
2. Удаление выбросов по правилу 3-х сигм.
3. Разделение: 80% обучение / 20% тест

Модели и гиперпараметры:

Модель	Гиперпараметры	Количество комбинаций
Ridge	alpha: [0.01, 0.1, 1, 10, 100], solver: [‘auto’, ‘svd’, ‘cholesky’]	75
RandomForest	n_estimators: [100, 200], max_depth: [None, 10, 20], min_samples_split: [2, 5]	60
XGBoost	n_estimators: [100, 200], max_depth: [3, 6], learning_rate: [0.01, 0.1]	40

Результаты:

Модель	R^2	RMSE	MAE	CV R^2
RandomForest	0.418	0.521	0.398	0.365
XGBoost	0.385	0.535	0.412	0.362

Ridge	0.312	0.567	0.438	0.298

Лучшая модель: RandomForestRegressor. Данная модель имела лучшие показатели R^2 и наименьшие ошибки RMSE и MAE. Хотя XGB и RF близки по cv_mean_r2, RF лучше обобщает на тесте.

2.1.2 regression_ic50.ipynb - Предсказание IC50

Предобработка данных:

1. Логарифмическое преобразование: $\log_{10}(IC50)$.
2. Удаление выбросов по правилу 3-х сигм.
3. Разделение: 80% обучение / 20% тест

Модели и гиперпараметры:

Модель	Гиперпараметры	Количество комбинаций
Ridge	alpha: [0.01, 0.1, 1, 10, 100], solver: [‘auto’, ‘svd’, ‘cholesky’]	75
RandomForest	n_estimators: [100, 200], max_depth: [None, 10, 20], min_samples_split: [2, 5]	60
XGBoost	n_estimators: [100, 200], max_depth: [3, 6], learning_rate: [0.01, 0.1]	40

Результаты:

Модель	R^2	RMSE	MAE	CV R^2
RandomForest	0.445	0.498	0.387	0.372

XGBoost	0.412	0.515	0.401	0.369
Ridge	0.325	0.558	0.429	0.305

Лучшая модель: RandomForestRegressor.

2.1.3 regression_si.ipynb - Предсказание SI

Особенности SI как целевой переменной:

1. Наибольшая вариативность среди целевых переменных (коэффициент вариации > 2.5).
2. Требует обязательного логарифмического преобразования.
3. Высокая асимметрия исходного распределения.

Предобработка данных:

1. Логарифмическое преобразование: $\log_{10}(SI)$.
2. Удаление выбросов по правилу 3-х сигм.
3. Разделение: 80% обучение / 20% тест.

Результаты:

Модель	R^2	RMSE	MAE	CV R^2
RandomForest	0.382	0.567	0.438	0.348
XGBoost	0.356	0.588	0.451	0.341
Ridge	0.285	0.632	0.487	0.272

Выводы: SI показывает наименьшее качество предсказания среди всех регрессионных задач, что связано с высокой вариативностью этого

показателя. Лучшей моделью для данной целевой переменной является Random Forest.

2.2 Задачи классификации

2.2.1 classification_cc50_over_median.ipynb

Предобработка данных:

1. Бинаризация: CC50_gt_median = (CC50 > median(CC50)).
2. Удаление выбросов по исходной переменной.
3. Балансировка классов

Модели и результаты:

Модель	ROC-AUC	F1-Score	Precision	Recall	CV ROC-AUC
RandomForest	0.876	0.871	0.869	0.873	0.834
XGBoost	0.834	0.828	0.831	0.825	0.798
LogisticRegression	0.792	0.785	0.788	0.782	0.756

Заключение: RandomForest классификатор показывает лучшие результаты.

2.2.2 classification_si_over_median.ipynb

Специфика задачи: - Исходная переменная: SI.

1. Бинарная переменная: SI_gt_median.
2. Принцип разделения: Медианное значение SI как порог.
3. Интерпретация: Высокая/низкая селективность соединений.
4. Баланс классов: 0.51 / 0.49.

Предобработка данных:

1. Создание бинарной переменной: SI_gt_median.
2. Удаление выбросов по исходной переменной SI.
3. Разделение: 80% обучение / 20% тест.

Результаты:

Модель	ROC-AUC	F1-Score	Precision	Recall	CV ROC-AUC
RandomForest	0.923	0.918	0.915	0.921	0.889
XGBoost	0.887	0.882	0.885	0.879	0.854
LogisticRegression	0.845	0.838	0.841	0.835	0.812

Заключение: Random Forest классификатор показывает лучшие результаты для SI

2.2.3 classification_si_over_8.ipynb

Особенность задачи:

1. Использование фиксированного порога SI > 8 вместо медианы.
2. Практическое значение: соединения с высокой селективностью (SI > 8).
3. Более несбалансированные классы: 0.35 / 0.65

Предобработка данных:

4. Создание бинарной переменной: SI_gt_8 = (SI > 8).
5. Удаление выбросов по исходной переменной SI.
6. Разделение: 80% обучение / 20% тест.

Результаты:

Модель	ROC-AUC	F1-Score	Precision	Recall	CV ROC-AUC
RandomForest	0.908	0.902	0.898	0.906	0.875
XGBoost	0.874	0.869	0.866	0.872	0.842
LogisticRegression	0.823	0.816	0.819	0.813	0.791

Заключение: фиксированный порог SI > 8 показывает даже лучшие результаты классификации, чем медианное разделение, что указывает на хорошую разделимость соединений с высокой селективностью

2.2.4 classification_ic50_over_median.ipynb

Предобработка данных:

1. Создание бинарной переменной: IC50_gt_median = (IC50 > median(IC50)).
2. Удаление выбросов по исходной переменной IC50.
3. Разделение: 80% обучение / 20% тест.

Баланс классов: 0.52 / 0.48 (хорошо сбалансированные классы)

Результаты:

Модель	ROC-AUC	F1-Score	Precision	Recall	CV ROC-AUC
RandomForest	0.891	0.886	0.883	0.889	0.856
XGBoost	0.857	0.852	0.849	0.855	0.821

LogisticRegression	0.798	0.792	0.795	0.789	0.763

Заключение: лучшая модель - RandomForest.

3. СРАВНИТЕЛЬНЫЙ АНАЛИЗ И ВЫБОР ЛУЧШИХ РЕШЕНИЙ

3.1 Сводная таблица результатов

Регрессионные задачи:

Задача	Лучшая модель	R^2	RMSE	Качество
CC50	RandomForest	0.418	0.521	Удовлетворительное
IC50	RandomForest	~0.445	~0.498	Удовлетворительное
SI	RandomForest	~0.382	~0.567	Требует улучшения

Классификационные задачи:

Задача	Лучшая модель	ROC-AUC	F1-Score	Качество
CC50 > median	RandomForest	0.876	0.871	Хорошее
IC50 > median	RandomForest	~0.891	~0.886	Хорошее

SI > median	RandomForest	0.923	0.918	Отличное
SI > 8	RandomForest	~0.908	~0.902	Отличное

3.2 Анализ производительности моделей

Выводы:

1. RandomForest доминирует во всех задачах.
2. Классификационные задачи показывают лучшие результаты, чем регрессионные.
3. SI - наиболее предсказуемая целевая переменная.
4. Качество классификации значительно превосходит качество регрессии.

3.3 Рекомендации по улучшению по каждой задаче

3.3.1 Регрессионные задачи (CC50, IC50, SI)

Общие рекомендации:

1. Генерация новых признаков:

   1.1 Создание полиномиальных признаков степени.

   2.1 Взаимодействие между молекулярными дескрипторами.

   2.3 Доменные трансформации (например, комбинации липофильности и размера молекулы).

   2.4 Создание биннинговых признаков для непрерывных переменных

2. Использование ансамблевых методов:

   – Stacking RandomForest + XGBoost + Ridge

- Blending с весовыми коэффициентами
- Voting Regressor с мягким голосованием

3. Использование бустингов:
    - LightGBM и CatBoost для улучшения качества градиентного бустинга
    - Neural Networks (многослойные персептроны) для нелинейных зависимостей
    - Support Vector Regression с RBF ядром

Специфичные рекомендации:

Для CC50 качество $R^2 = 0.418$ требует улучшения. Рекомендуется добавить признаки взаимодействия. Также попробовать различные трансформации целевой переменной

Для IC50: лучшие результаты среди регрессионных задач ($R^2=0.445$) - Следует сфокусироваться на fine-tuning RandomForest. Также может помочь добавление доменных знаний о механизмах ингибирования.

Для SI рассмотреть multi-task learning с CC50 и IC50.

3.3.2 Классификационные задачи

Для CC50 > median: калибровка вероятностей, threshold optimization - Feature importance анализ для интерпретации

Для IC50 > median: рекомендуется A/B тестирование различных порогов

Для SI > median: исследовать feature importance

Для SI > 8: рекомендуется создание интерактивного dashboard для предсказаний

3.3.3 Общие технические улучшения

1. Гиперпараметрическая оптимизация:
   - Bayesian Optimization (Optuna, Hyperopt)
   - Расширенные сетки поиска
   - Early stopping для предотвращения переобучения
2. Валидация:
   - Time-series split если данные имеют временную компоненту
   - Stratified K-fold для всех задач
   - Leave-one-out CV для малых выборок
3. Интерпретируемость:
   - SHAP values для объяснения предсказаний
   - LIME для локальных объяснений
   - Permutation importance для ранжирования признаков
4. Deployment:
   - Создание API с FastAPI/Flask
   - Контейнеризация с Docker
   - Мониторинг качества модели в продакшене

4. ВЫВОДЫ

4.1 Основные достижения

1. Успешная предобработка данных с минимальными потерями (3.5%).
2. Эффективное применение логарифмического преобразования для улучшения распределений.
3. Достижение высокого качества классификации (ROC-AUC > 0.87 для всех задач).
4. Создание воспроизводимого пайплайна машинного обучения.
5. Систематическое сравнение различных алгоритмов с подбором гиперпараметров.

4.2 Научная значимость

1. Демонстрация эффективности RandomForest для токсикологических данных.
2. Важность качественной предобработки для биологических данных.
3. Превосходство классификационных подходов над регрессионными для данного типа задач.
4. Валидация подхода на реальных молекулярных дескрипторах.

4.3 Практическая применимость

Разработанные модели могут быть использованы для:

1. Предварительного скрининга новых химических соединений.
2. Приоритизации соединений для экспериментального тестирования.

3. Сокращения времени и затрат на разработку лекарственных препаратов.
4. Поддержки принятия решений в фармацевтической индустрии

6. ЗАКЛЮЧЕНИЕ

6.1 Достигнутые результаты

В рамках данной курсовой работы была продемонстрирована высокая эффективность применения методов машинного обучения для анализа токсикологических данных. Разработан полный пайплайн от исходных данных до готовых к использованию моделей.

6.2 Научная значимость

1. Подтверждена эффективность RandomForest для токсикологических предсказаний
2. Продемонстрирована важность качественной предобработки данных
3. Показано превосходство классификационных подходов над регрессионными
4. Установлены benchmark результаты для данного типа молекулярных дескрипторов

6.3 Практическая значимость

Применение в фармацевтической индустрии:

1. Предварительный скрининг новых соединений.
2. Сокращение количества дорогостоящих лабораторных экспериментов
3. Приоритизация соединений для дальнейших исследований
4. Поддержка принятия решений на ранних стадиях разработки

6.4 Ограничения исследования

1. Размер выборки: 966 соединений может быть недостаточно для глубокого обучения
2. Типы дескрипторов: Использованы только классические молекулярные дескрипторы
3. Внешняя валидация: Отсутствие тестирования на независимых датасетах
4. Интерпретируемость: Недостаточный анализ важности признаков