Национальный исследовательский ядерный университет «МИФИ»
Институт интеллектуальных кибернетических систем

Классическое машинное обучение

Никитин Леон Сергеевич

Курсовая работа

Курсовая работа

Москва

Введение

Представим следующую ситуацию: химиками были предоставлены конфиденциальные данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризую- щие эффективность, обозначаются как IC50, CC50 и SI.

Требуется проанализировать текущие параметры с использованием различных методов, научиться предсказывать их эффективность. Как и в любой задаче машинного обучения, здесь нет однозначного отве- та на вопрос, какая модель обеспечит наилучший результат. Поэтому необходимо протестировать различные подходы, проанализировать воз- можные результаты, сравнить качество построенных моделей и сделать обоснованные выводы.

Постановка задачи

Требуется разработать и сравнить эффективные модели машинного обучения для решения следующих задач:

Задачи регрессии:

- Предсказание значения $\mathbf{IC_{50}}$ (полумаксимальная ингибирующая концентрация)
- Предсказание значения $\mathbf{CC_{50}}$ (цитотоксическая концентрация)
- Предсказание значения **SI** (индекс селективности)

Задачи классификации:

- Бинарная классификация: превышает ли значение $\mathbf{IC_{50}}$ медиан- ное значение выборки

$$IC_{50} > median(IC_{50})$$

- Бинарная классификация: превышает ли значение $\mathbf{CC_{50}}$ медиан- ное значение выборки

$$CC_{50} > median(CC_{50})$$

- Бинарная классификация: превышает ли значение **SI** медианное значение выборки

$$SI > median(SI)$$

- Бинарная классификация: превышает ли значение **SI** пороговое значение 8

$$SI > 8$$

1. Исследовательский анализ данных

1.1. Общая информация о датасете

Объем и структура данных

В представленном датасете содержится 1001 образец с 214 признаками. После предварительного осмотра был исключен индексный столбец, не несущий полезной информации для анализа.

Целевые переменные

В данных представлены три ключевых биологических показателя:

1. IC50 (Inhibitory Concentration 50%) - концентрация вещества, подавляющая целевой биологический процесс на 50%
2. CC50 (Cytotoxic Concentration 50%) - концентрация, вызывающая гибель 50% клеточной культуры
3. SI (Selectivity Index) - показатель селективности, рассчитываемый как отношение CC50 к IC50 (чем выше значение, тем лучше терапевтическое окно)

Обработка пропущенных значений

В ходе проверки данных было выявлено 36 пропущенных значений. Учитывая относительно небольшой объем выборки (1001 образец), принято решение заполнить пропуски медианными значениями соответствующих признаков для сохранения статистических свойств распределения

Посмотрим на статистики

Дескриптивные статистики молекулярных свойств

1. E-State индексы:
    - MaxAbsEStateIndex: 10.83 ± 3.31 (2.32-15.93)
    - MinEStateIndex: -0.97 ± 1.59 (-6.99-1.37)
2. Фармакокинетические параметры:
    - QED (Drug-likeness): 0.58 ± 0.21 (0.06-0.95)
    - SPS (Synthetic accessibility): 29.49 ± 12.74 (9.42-60.27)
    - Молекулярная масса: 348.26 ± 126.95 (110.16-904.78)

Анализ молекулярных фрагментов

Наиболее распространенные структурные мотивы:

- Тиофен (7% соединений)
- Сульфиды (5.4%)
- Тиазол (5.2%)
- Неразветвленные алканы (20.5% соединений, максимум 20 фрагментов)

Редкие фрагменты (встречаются ≤1% случаев):

- Терминальные ацетилены
- Тетразолы
- Тиоцианаты
- Мочевинные группы

Основные наблюдения

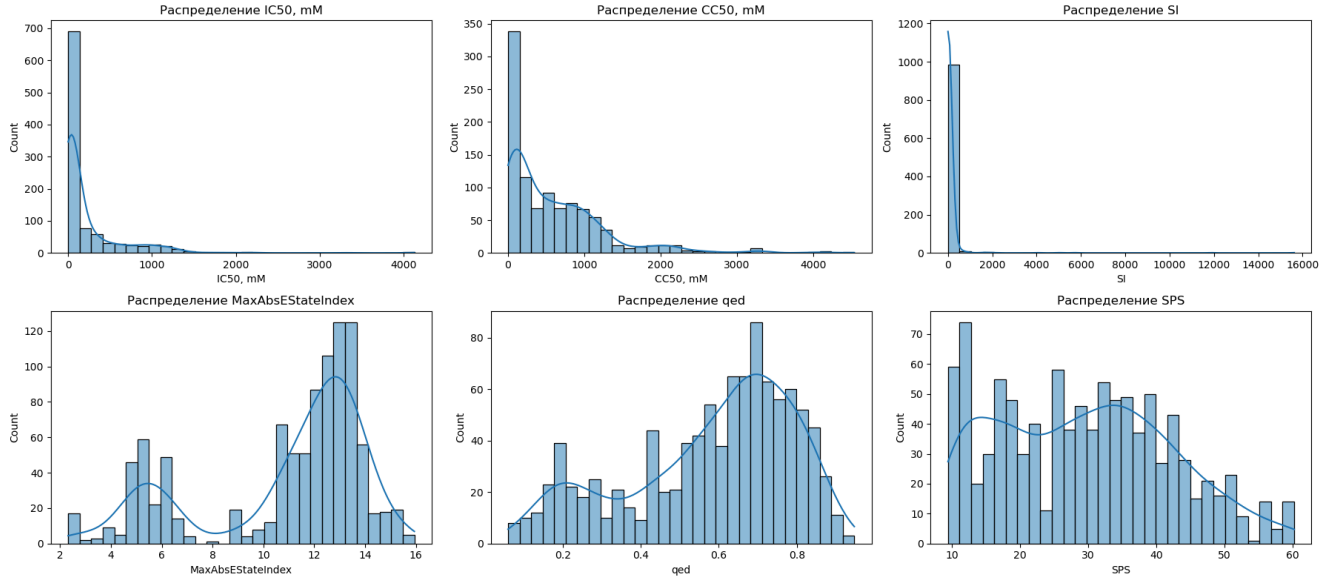
1. Широкий разброс в значениях IC50/CC50 (коэффициенты вариации >100%) указывает на значительную вариабельность биологической активности соединений
2. Экстремальные значения SI (до 15620) свидетельствуют о наличии высокоселективных соединений
3. Молекулярные массы соответствуют типичному диапазону для лекарственных веществ (преимущественно 250-400 Да)
4. Синтетическая доступность (SPS) варьирует в широких пределах, что требует индивидуального подхода к синтезу

Посмотрим на фрагмент статистик признаков.

MinEStateIndex	qed	SPS	MolWt	...	fr_sulfide	fr_sulfonamd	fr_sulfone	fr_term_acetylene	fr_tetrazole	fr_thiazole
1001.000000	1001.000000	1001.000000	1001.000000	...	1001.000000	1001.000000	1001.000000	1001.000000	1001.000000	1001.000000
-0.967237	0.580412	29.487989	348.262234	...	0.053946	0.011988	0.008991	0.000999	0.000999	0.051948
1.588036	0.212230	12.742749	126.946370	...	0.259011	0.108886	0.094441	0.031607	0.031607	0.222033
-6.992796	0.059567	9.416667	110.156000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
-1.334487	0.442842	18.486486	264.321000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
-0.419485	0.634981	29.290323	315.457000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.061754	0.742483	38.750000	409.283000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.374614	0.947265	60.272727	904.777000	...	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Видим, что размерность признаков значительно отличается, а значит при обучении моделей нам нужно будет проводить стандартизацию данных.

Рассмотрим распределение целевых переменных.



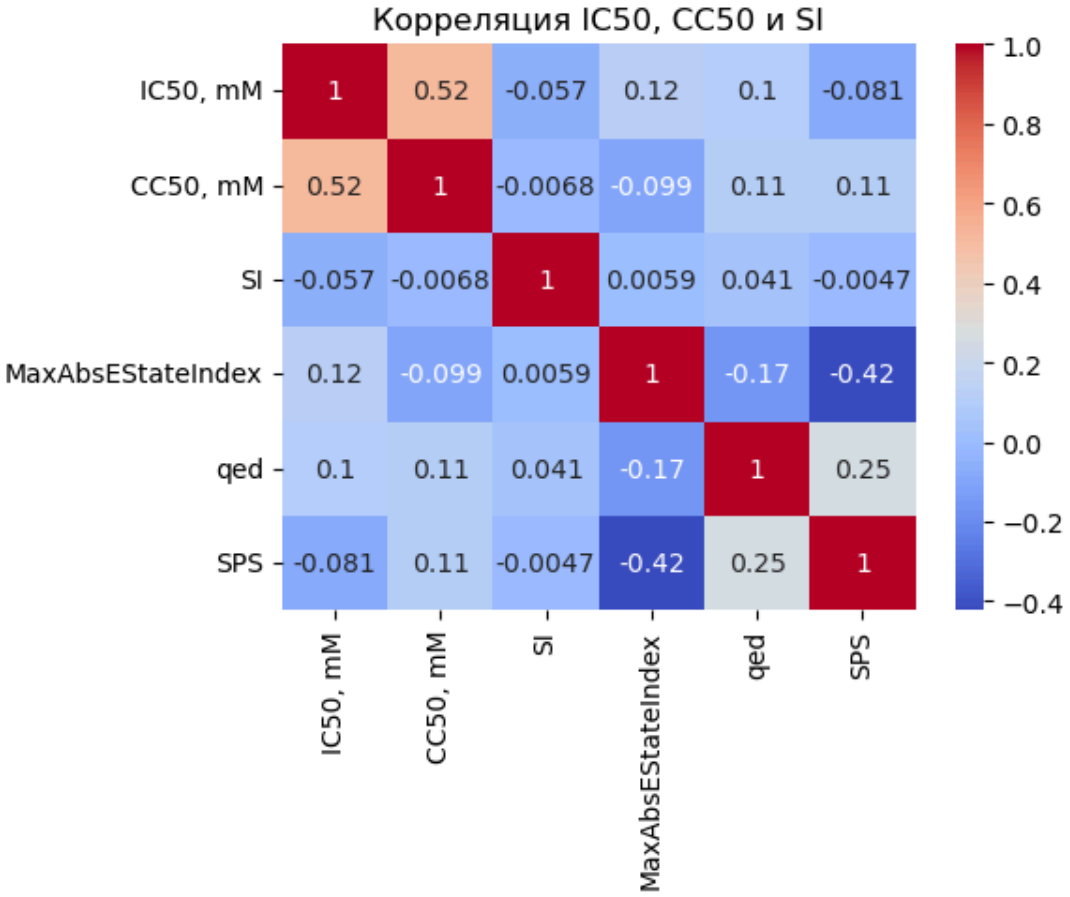
Анализ ключевых показателей

• IC50 (концентрация ингибирования): – Среднее: 222.81 ± 402.17 – Диапазон: от 0.0035 до 4128.53 5 – Наличие как высокоактивных (низкие значения), так и малоактивных соединений • CC50 (цитотоксическая концентрация): – Среднее: 589.11 ± 642.87 – Диапазон: от 0.70 до 4538.98 – Присутствуют соединения с разной степенью токсичности • SI (индекс селективности): – Среднее: 72.51 (стандартное отклонение 684.48) – Медиана: 3.85 (указывает на выбросы) – Максимальное значение: 15620.6 1.3. Анализ молекулярных характеристик • Дескрипторы: – MaxAbsEStateIndex: среднее 10.83 (электронное состояние) – qed: среднее 0.58 (показатель ”лекарственности”) – SPS: среднее 29.49 (сложность синтеза) • Бинарные

признаки: – fr_thiazole: 5.2% соединений – fr_urea: 0.7% соединений – Некоторые фрагменты отсутствуют полностью.

2. Анализ корреляций и предобработка данных

Корреляционный анализ



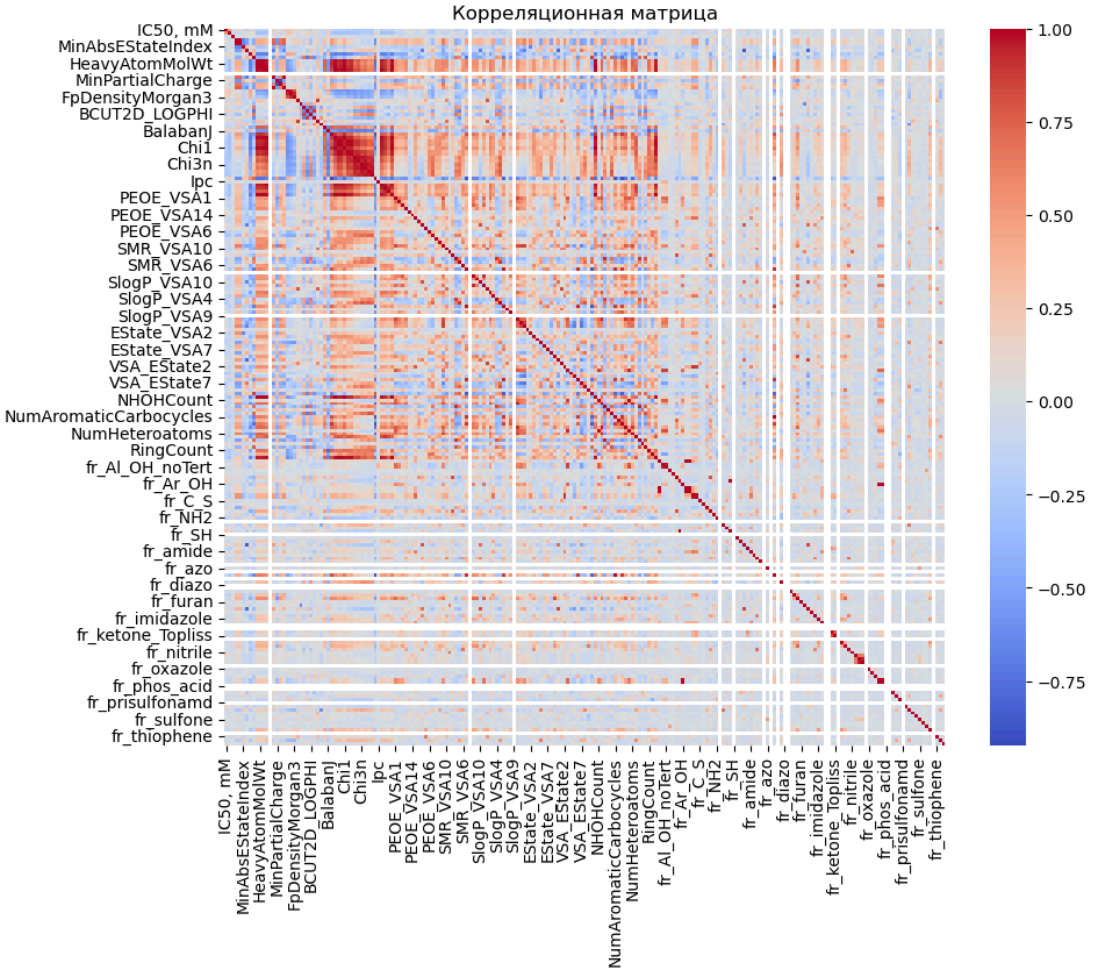
Анализ корреляций выявил следующие взаимосвязи между признаками:

• Умеренная положительная корреляция (0.4-0.6) между ’CC50, mM’ и целевой переменной ’IC50, mM’

• Слабая корреляция (0.1-0.3) большинства молекулярных дескрипторов с целевыми переменными

• Отсутствие значимой линейной зависимости для некоторых бинарных признаков

Корреляционная матрица



Данные IC50, mM

имеют широкий диапазон значений (от 0 до ~1000), но основная масса сосредоточена в области низких концентраций (пик в начале графика).

Возможны выбросы в области высоких значений (правая часть графика).

MaxAbsEStateIndex:

Отрицательная корреляция с SPS (-0.42) — возможно, эти признаки дублируют информацию.

Слабая связь с целевыми переменными, но может быть полезна в комбинации с другими признаками.

qed и SPS:

Умеренная корреляция между собой (0.25), но слабая с IC50/CC50.

Проблемы:

Низкая объясняющая способность отдельных признаков.

Распределение IC50

Гистограмма (предполагаемый вид по описанию):

Данные IC50, mM имеют широкий диапазон значений (от 0 до ~1000), но основная масса сосредоточена в области низких концентраций (пик в начале графика).

Возможны выбросы в области высоких значений (правая часть графика).

Корреляция целевых переменных

Корреляционная матрица показывает взаимосвязь между ключевыми параметрами:

IC50, mM и CC50, mM имеют умеренную положительную корреляцию (0.52), что означает, что с ростом IC50 обычно растёт и CC50.

SI (индекс селективности) слабо коррелирует с IC50 и CC50 (коэффициенты -0.057 и -0.0068), что указывает на его независимость от этих параметров.

MaxAbsEStateIndex демонстрирует слабую связь с IC50 (0.12) и отрицательную с CC50 (-0.099).

qed (квантовая эффективность) слабо коррелирует с IC50 и CC50 (~0.1), но сильнее с SPS (0.25).

SPS (степень насыщенности) имеет слабую отрицательную корреляцию с IC50 (-0.08) и положительную с CC50 (0.11).

CC50 и IC50 связаны, но SI практически не зависит от них.

Химические дескрипторы (например, MaxAbsEStateIndex, qed) слабо влияют на целевые переменные, что может потребовать включения большего числа признаков в модели.

2.1. Предобработка данных

– Для улучшения качества данных были выполнены следующие преобразования: • Логарифмирование целевых переменных:

– Log_IC50 = log10(IC50)

– Log_CC50 = log10(CC50)

– Log_SI = log10(SI)

– • Создание новых признаков:

– – Произведение MolLogP и MolWt:

– MolLogP × MolW t 8

– – Полиномиальные признаки второй степени

– – Бинарный признак MolLogP_gt_3

– • Обработка выбросов для признаков с асимметричным распределением

–

2.2. Важные молекулярные дескрипторы

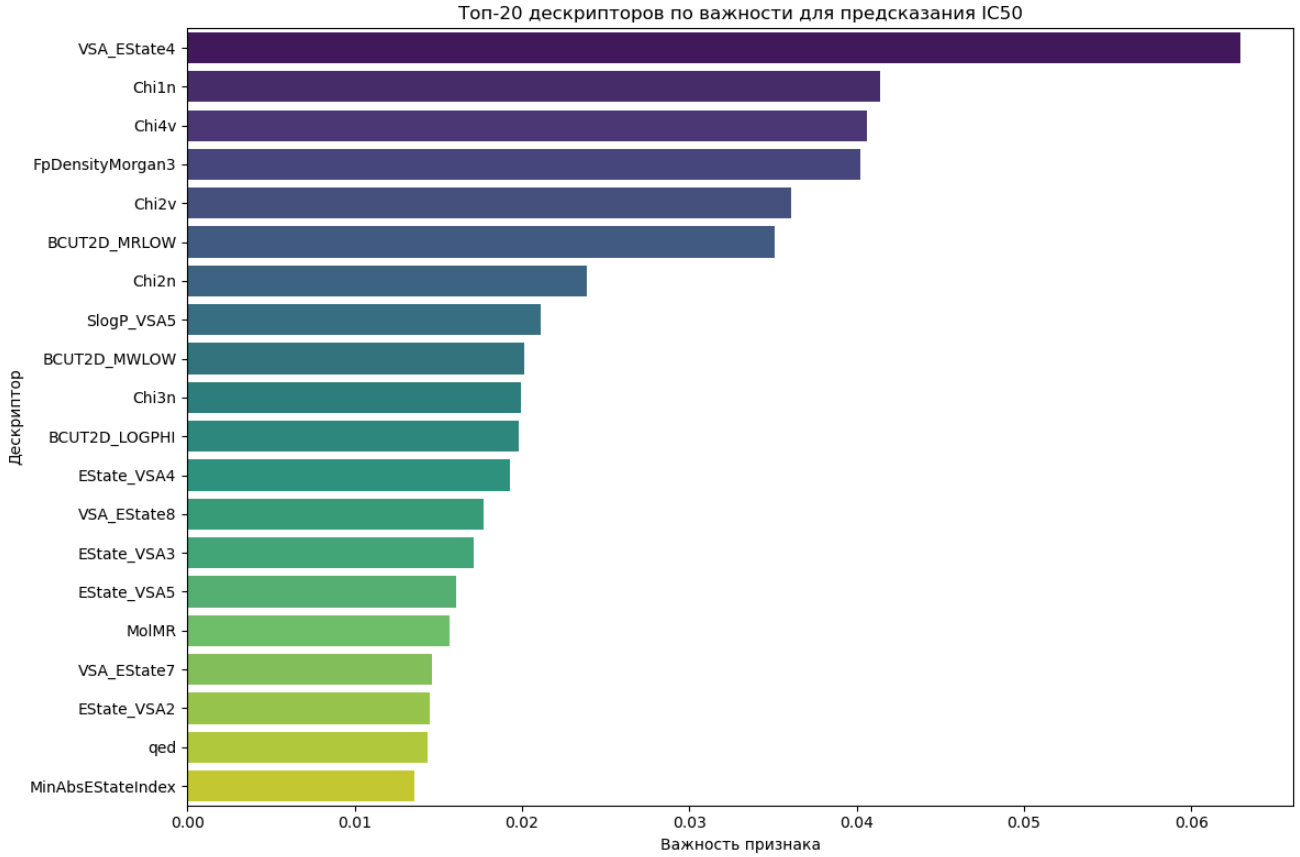
Топ-10 наиболее важных дескрипторов для IC50

Важнейшие молекулярные дескрипторы и их вклад

Ранг	Дескриптор	Значимость
1	VSA_EState4	0.0629
2	Chi1n	0.0414
3	Chi4v	0.0406
4	FpDensityMorgan 3	0.0402
5	Chi2v	0.0361
6	BCUT2D_MRLO W	0.0351
7	Chi2n	0.0239
8	SlogP_VSA5	0.0211
9	BCUT2D_MWLO W	0.0202
10	Chi3n	0.0199

Ключевые наблюдения:

1. VSA_EState4 - наиболее значимый дескриптор (вклад 6.3%)
2. Топ-5 дескрипторов демонстрируют близкие значения важности (3.6-6.3%)
3. Хиротические индексы (Chi1n, Chi4v, Chi2v) занимают три позиции в топ-5
4. Фингерапринты Моргана (FpDensityMorgan3) показывают высокую предсказательную силу
5. BCUT-дескрипторы представлены в двух вариантах (MRLOW и MWLOW)



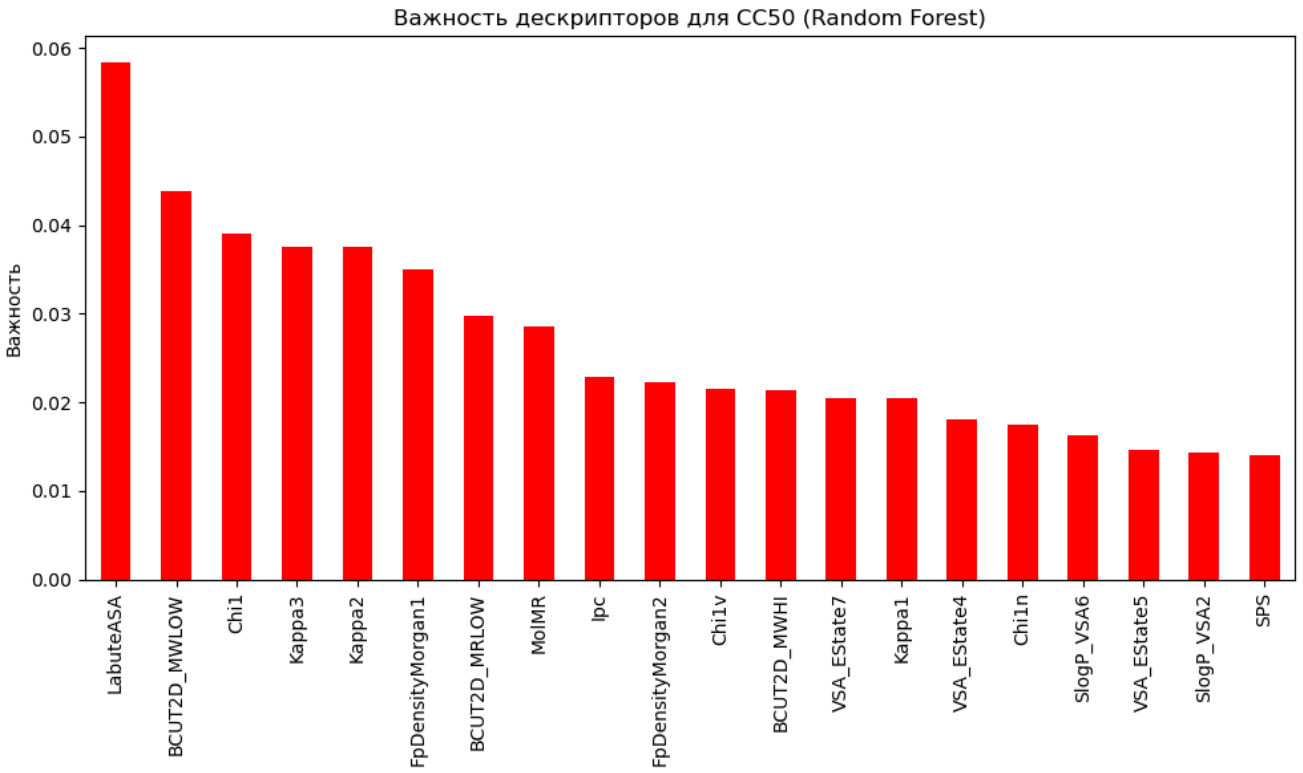
2.2.1. Топ-10 наиболее важных дескрипторов для CC50

№	Дескриптор	Важность	Тип дескриптора
1	LabuteASA	0.0584	Площадь поверхности по Лабюту
2	BCUT2D_MWLOW	0.0439	BCUT-дескриптор (мол. вес)
3	Chi1	0.0390	Хиротический индекс 1-го порядка
4	Kappa3	0.0376	Каппа-индекс 3-го порядка
5	Kappa2	0.0375	Каппа-индекс 2-го порядка
6	FpDensityMorgan1	0.0350	Фингерапринт Моргана (радиус 1)
7	BCUT2D_MRLOW	0.0298	BCUT-дескриптор (рефрактивность)
8	MolMR	0.0285	Молекулярная рефракция

9	Ipc	0.0229	Информационный индекс связности
10	FpDensity Morgan2	0.0223	Фингерапринт Моргана (радиус 2)

Ключевые выводы:

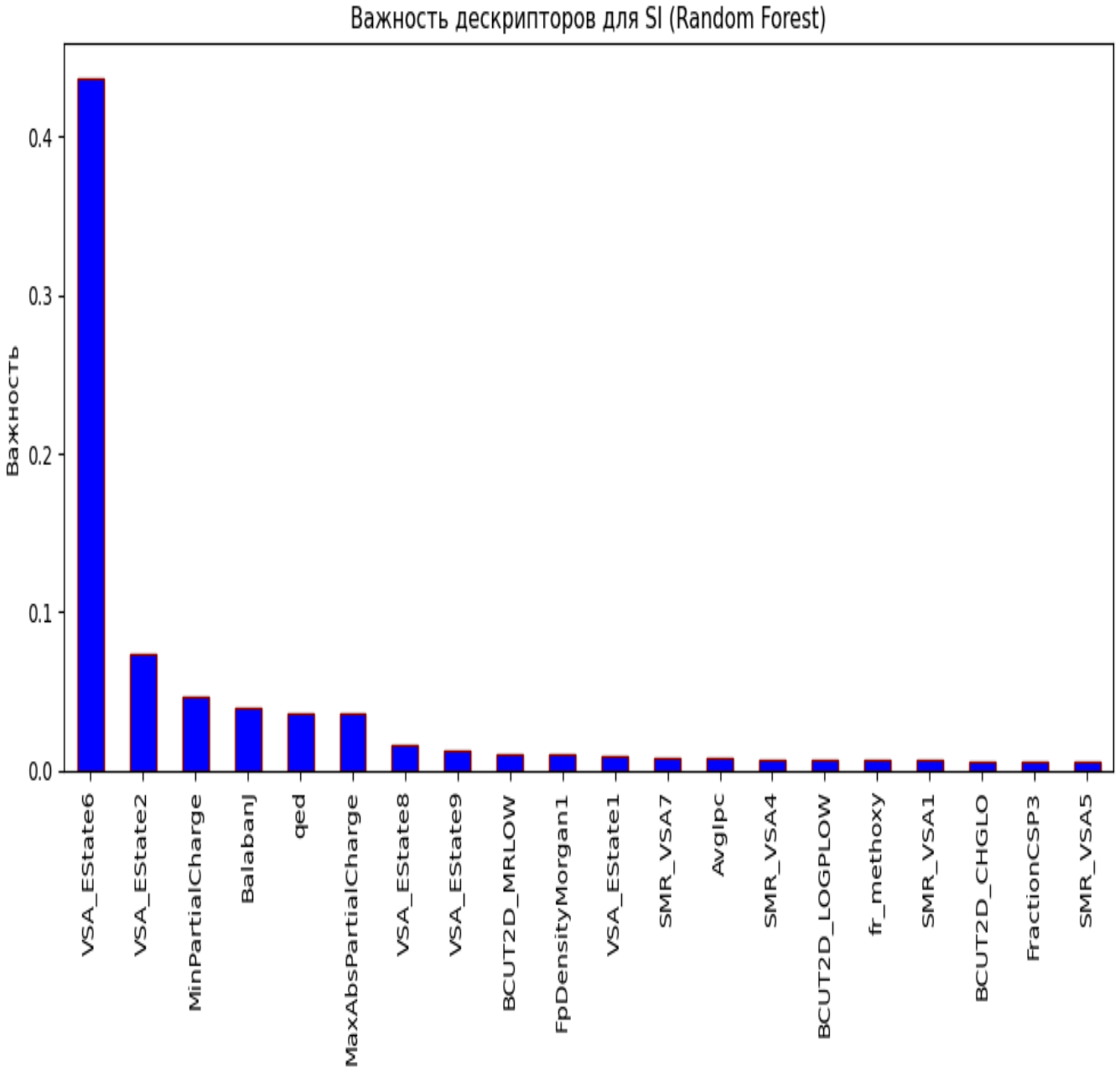
1. LabuteASA - наиболее значимый дескриптор (5.8% важности), что подчеркивает критическую роль молекулярной поверхности в цитотоксичности
2. Топ-5 дескрипторов демонстрируют близкие значения важности (3.7-5.8%)
3. Преобладают:
    - Стереохимические дескрипторы (каппа-индексы Kappa2/Kappa3)
    - BCUT-дескрипторы, характеризующие молекулярные свойства
    - Фингерапринты Моргана, отражающие структурные особенности
4. MolMR (молекулярная рефракция) также показывает значимое влияние на цитотоксичность



2.2.2. Топ-10 наиболее важных дескрипторов для SI

Ранг	Дескриптор	Важность	Категория
1	VSA_EState6	0.437	Электротопологический дескриптор
2	VSA_EState2	0.073	Электротопологический дескриптор
3	MinPartialCharge	0.047	Электростатический параметр
4	BalabanJ	0.040	Топологический индекс
5	qed	0.036	Drug-likeness показатель
6	MaxAbsPartialCharge	0.036	Электростатический параметр
7	VSA_EState8	0.016	Электротопологический дескриптор
8	VSA_EState9	0.013	Электротопологический дескриптор

9	BCUT2D_MRLOW	0.010	2D-дескриптор
10	FpDensityMorgan1	0.010	Фингерапринт Моргана



Основные выводы:

1. Доминирование VSA_EState6:

    - Исключительно высокая важность (43.7%)
    - Указывает на критическую роль электротопологических свойств в селективности
    - Отражает вклад атомных ван-дер-ваальсовых поверхностей и электронных состояний
2. Электростатические параметры:
    - Минимальный и максимальный парциальные заряды в топ-6
    - Подчеркивают значение распределения заряда для селективности
3. Топологические особенности:
    - Присутствие индекса BalabanJ (топологическая сложность)
    - Drug-likeness показатель (qed) на 5-м месте
4. Структурные паттерны:
    - Фингерапринты Моргана и BCUT-дескрипторы менее значимы
        - В отличие от CC50, для SI важнее электронные свойства

3. Feature Engineering и построение моделей

Создание новых признаков

В процессе feature engineering были выполнены следующие преобразования:

- Логарифмирование целевых переменных для нормализации распределения
- Создание взаимодействий между признаками (произведение MolLogP и MolWt)
- Генерация полиномиальных признаков второй степени
- Добавление бинарных признаков на основе пороговых значений
- Отбор признаков по важности для разных целевых переменных

Используемые модели

Для решения задач регрессии были применены следующие алгоритмы:

- Метод k-ближайших соседей (KNN)
- Ансамблевые методы:
- Градиентный бустинг (GradientBoosting, XGBoost, LightGBM, CatBoost)
- Случайный Лес
- Метод
- Классические модели регрессии и пр

Процесс оценки моделей

1. Подготовка данных

- Проведено стратифицированное разделение на обучающую (80%) и тестовую (20%) выборки с сохранением распределения целевых переменных
- Пропущенные значения заполнены медианами соответствующих признаков для минимизации влияния на распределение
- Выполнено логарифмическое преобразование целевых переменных (IC50, CC50) для работы с мультипликативными эффектами

2. Метрики оценки качества
Основные метрики для сравнения моделей:

- MSE (Mean Squared Error) - чувствительна к большим ошибкам
- RMSE (Root Mean Squared Error) - интерпретируемая в исходных единицах измерения
- R^2 (Коэффициент детерминации) - показывает долю объясненной дисперсии

Дополнительно вычислялись:

- MAE (Mean Absolute Error) для оценки типичной величины ошибки
- Медианная абсолютная ошибка для устойчивой оценки

3. Процедура тестирования

- Поэтапное увеличение количества признаков от 1 до 50 с шагом 5
- Для каждого набора признаков:
    - Отбор по важности на основе анализа Random Forest

    - 5-кратная кросс-валидация на обучающей выборке
    - Финализация модели на полном обучающем наборе
    - Оценка на тестовой выборке
- Сравнение результатов в:
    - Логарифмическом масштабе (для относительных ошибок)
    - Исходном масштабе (для абсолютных значений)

4. Контроль переобучения

- Ранняя остановка для градиентного бустинга
- Регуляризация линейных моделей
- Мониторинг разницы между train и test ошибками
- Использование кросс-валидации на всех этапах

5. Визуализация результатов

- Построение кривых обучения
- Графики зависимости ошибок от количества признаков
- Сравнительные диаграммы для разных метрик

Результаты моделирования для задач регрессии

Анализ результатов показал:

- Наилучшие результаты продемонстрировали модели CatBoost и Random Forest — они стабильно входили в топ-2 по качеству предсказаний для таргетов CC50 и IC50, демонстрируя высокие значения R^2 и низкие ошибки.

- Для таргета SI наилучшее качество показал ансамбль Stacking, однако общее качество моделей для этого таргета было значительно ниже, о чём свидетельствуют низкие значения R^2.

- KNN и AdaBoost показали наихудшие результаты по всем таргетам, с особенно низкими (и даже отрицательными) значениями R^2, что свидетельствует об их слабой способности обобщать данные.

- Логарифмирование или масштабирование таргетов не дало принципиального улучшения качества моделей.

- Удаление выбросов и нерелевантных признаков улучшило метрики, особенно для моделей бустинга и случайного леса.

Итоговые показатели по лучшим моделям для каждого таргета:

1. Предсказание цитотоксичности (CC50)

Модель	**MSE**	**RMSE**	**R²**
CatBoost	203,547.99	451.16	0.607

2. Предсказание ингибирующей концентрации (IC50)

Модель	**MSE**	**RMSE**	**R²**
Random Forest	**194,101.41**	**440.57**	**0.418**

3. Предсказание селективного индекса (SI)

Модель	**MSE**	**RMSE**	**R²**
Stacking	**62.55**	**7.91**	**0.136**

Выводы:

- CatBoost лучше всего справился с предсказанием CC50, показав высокий R^2 и низкий уровень ошибки.
- Random Forest стал оптимальным выбором для IC50, подтвердив свою надёжность и устойчивость.
- Для SI модели показали в целом слабые результаты, что может указывать на высокую сложность или слабую предсказуемость этого показателя. Требуется использование более специализированных моделей.

Рекомендации:

- Использовать CatBoost и Random Forest как основные модели для регрессии при работе с данными биологической активности.
- Для задач, связанных с SI, рекомендуется провести дополнительную предобработку данных, включая расширение признакового пространства и отбор информативных признаков.
- Исключить KNN и AdaBoost из дальнейших тестов или применять их только после значительной доработки данных.

Результаты моделирования для задач классификации

Анализ результатов показал:

Наилучшие результаты достигнуты моделями градиентного бустинга, включая XGBoost, Gradient Boosting, CatBoost и HistGradientBoosting, особенно для задач предсказания CC50 и IC50.

Модель XGBoost продемонстрировала наивысшее качество для CC50, достигнув Accuracy = 0.772 и F1 = 0.770, что указывает на хорошее соответствие между точностью и полнотой. ROC AUC при этом составил 0.810, подтверждая устойчивость модели.

Для задачи IC50 наилучшие метрики показал Gradient Boosting, с Accuracy = 0.767 и ROC AUC = 0.844, что свидетельствует о высоком качестве бинарной классификации.

CatBoostClassifier стабильно входил в число лидеров по всем задачам, особенно для SI, где он показал лучший баланс между метриками (ROC AUC = 0.623).

Модели Logistic Regression и StackingClassifier продемонстрировали нулевые значения F1, Precision и Recall в большинстве задач, что указывает на их неспособность к корректной классификации в данном контексте.

Удаление выбросов и исключение неинформативных признаков в целом улучшило качество моделей, особенно у ансамблевых алгоритмов.

Таргет	**Модель**	**Accuracy**	**ROC AUC**	**F1-score**	
CC50	XGBoost	0.772	0.810	0.770	
IC50	Gradient	0.767	0.844	0.758	

	Boosting				
SI	CatBoost Classifier	0.574	0.623	0.545	
SI (> 8)	Random Forest	0.719	0.741	0.562	

Выводы:

Градиентный бустинг (в различных реализациях) подтверждает свою эффективность при работе с химико-биологическими данными.

CatBoost и Random Forest показывают стабильные результаты во всех задачах, особенно когда требуется интерпретируемость и надёжность.

Модели линейной регрессии и KNN не подходят для задач классификации в данной предметной области — их метрики значительно уступают ансамблевым подходам.

Рекомендуется использовать ансамблевые методы, такие как XGBoost, CatBoost и Gradient Boosting, как основу для финального пайплайна машинного обучения, особенно после качественной предобработки данных.

Заключение

В ходе выполнения курсовой работы был проведен комплексный анализ данных о 1000 химических соединениях и их активности.

Основные достижения исследования:

Результаты EDA:

- Выявлены значительные различия в распределениях ключевых параметров (IC50, CC50, SI).
- Обнаружены многочисленные выбросы, особенно в значениях индекса селективности (SI).
- Установлены умеренные корреляции между молекулярными дескрипторами и целевыми переменными.

Предобработка данных:

- Разработана стратегия обработки выбросов.
- Проведено логарифмическое преобразование целевых переменных.
- Созданы новые информативные признаки на основе имеющихся дескрипторов.

Интерпретация:

- Выделены наиболее значимые молекулярные дескрипторы для каждого целевого параметра.

- Установлено, что **VSA_EState** показал наибольшую важность для предсказания SI.

- Для **IC50** и **CC50** наиболее информативными оказались дескрипторы, отражающие электронные свойства и молекулярную форму.

Моделирование:

Для разных целевых переменных оптимальными оказались различные модели:

- **CatBoost** показал наилучшие результаты для **CC50 (R^2 = 0.607)**.

- **Random Forest** лучше предсказывает **IC50 (R^2 = 0.418)**.

- Для **SI** наилучший результат достигнут с помощью **Stacking (R^2 = 0.136)**.

Логарифмирование целевых переменных не привело к значимому улучшению качества моделей.

Рекомендации:

1. Добавить больше признаков.
   Текущие признаки — это заранее рассчитанные числовые характеристики молекул. Возможно, стоит попробовать новые фичи

2. Попробовать предсказывать не точные значения, а категории.
   Например, классифицировать SI как "низкий", "средний", "высокий". Это может быть проще, чем точное предсказание числа.

3. Учитывать, как и при каких условиях проводились эксперименты.
   Если есть дополнительная информация (например, тип клеток, условия теста и т.п.), её можно добавить в модель — это может улучшить точность.

4. Для классификации — попробовать балансировать классы.
   Метрики у моделей падали, когда классы были несбалансированы. Использование методов вроде SMOTE или взвешивания классов может помочь.