Национальный исследовательский ядерный университет «МИФИ»
Институт интеллектуальных кибернетических систем

Классическое машинное обучение

Камерзан Игорь Николаевич

Курсовая работа

Курсовая работа

Научный руководитель:

Москва
2025

Оглавление

Введение

Представим следующую ситуацию: химиками были предоставлены конфиденциальные данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие 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 ответа на вопрос, какая модель обеспечит наилучший результат. Поэтому необходимо протестировать различные подходы, проанализировать возможные 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 $\mathbf{IC_{50}}$ медианное значение выборки

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

- Бинарная классификация: превышает ли значение $\mathbf{CC_{50}}$ медианное 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

- Размерность данных: 1001 строка × 214 столбцов
- Пропущенные значения отсутствуют во всех столбцах
- Большинство признаков числовые, есть бинарные категориальные признаки (fr_thiazole, fr_urea)

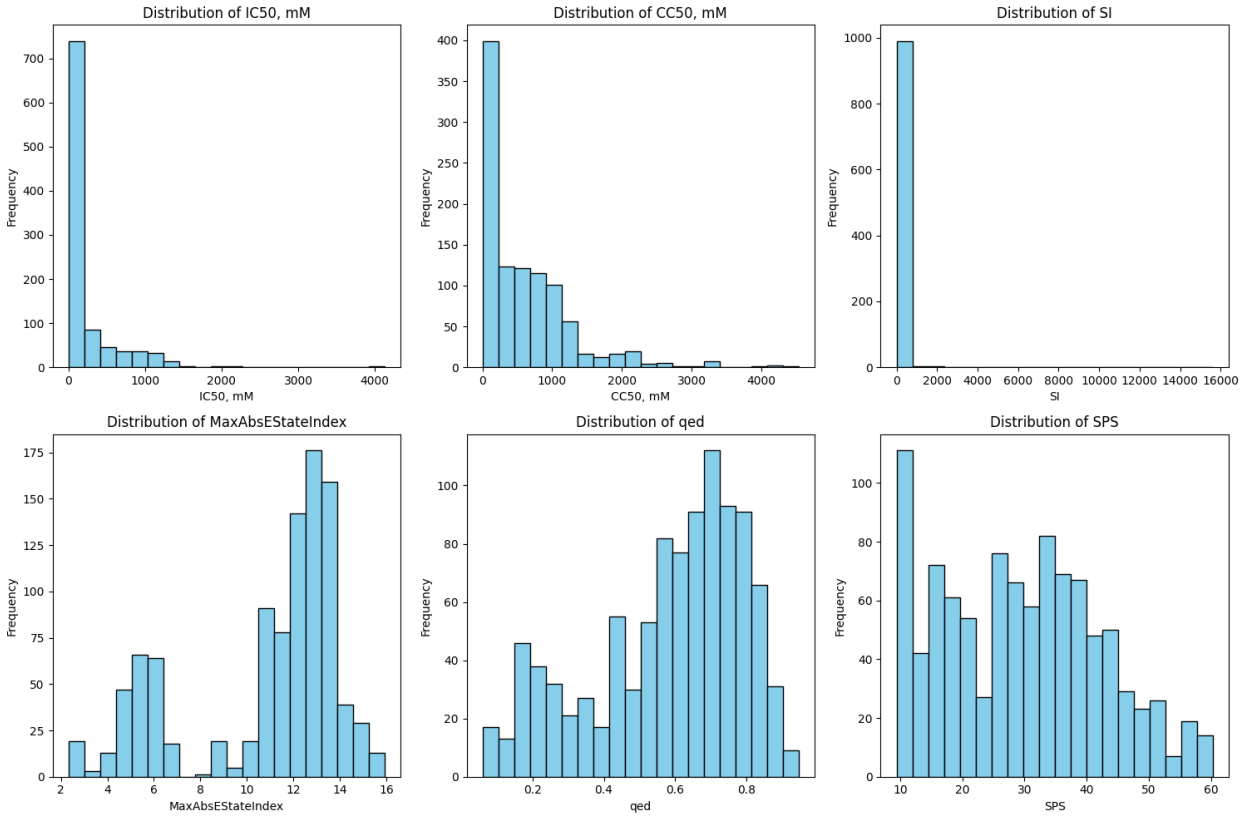

Рис. 1: Распределения целевых признаков

1.2. Анализ ключевых показателей

- **IC50 (концентрация ингибирования)**:
  - Среднее: 222.81 ± 402.17
  - Диапазон: от 0.0035 до 4128.53

  - 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

2.1. 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

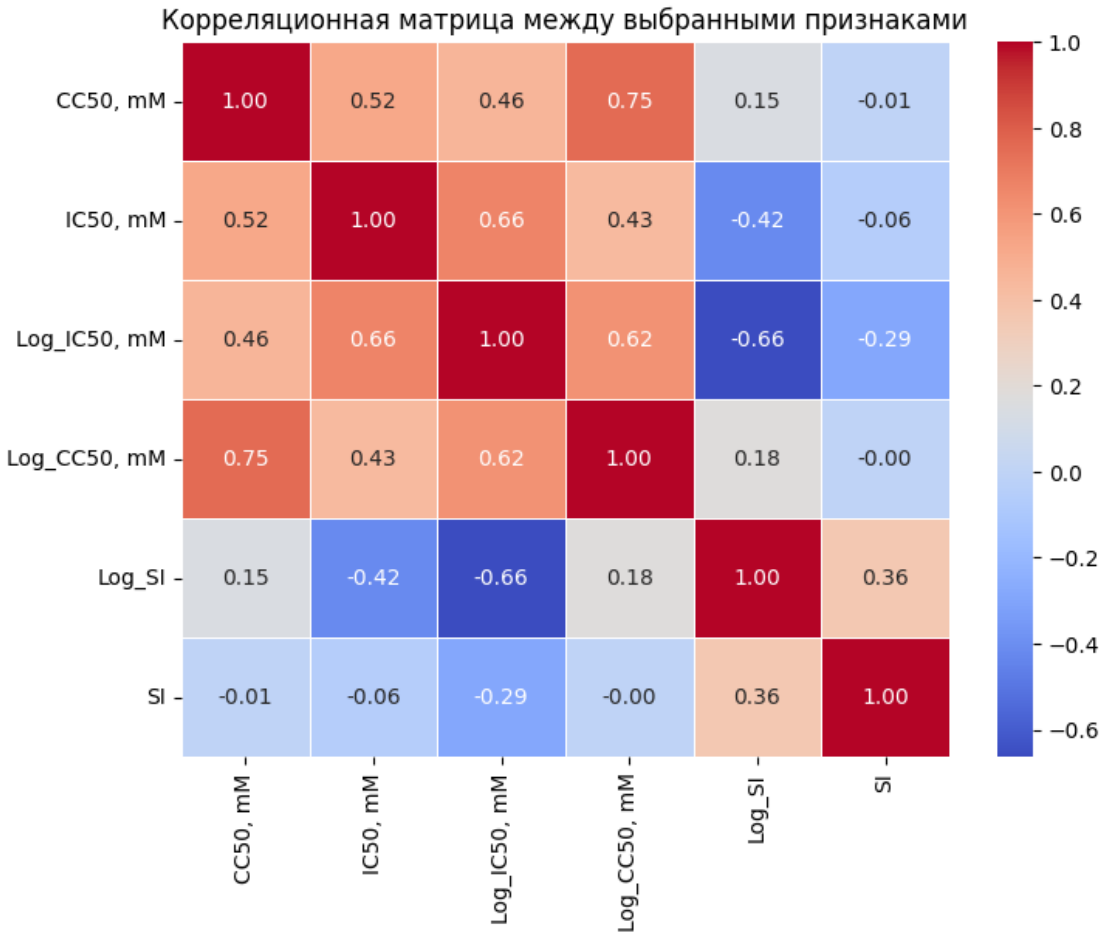

Рис. 2: Матрица корреляций между ключевыми признаками

2.2. Предобработка данных

Для улучшения качества данных были выполнены следующие преобразования:

- Логарифмирование целевых переменных:

$$Log\_IC50 = \log_{10}(IC50)$$

$$Log\_CC50 = \log_{10}(CC50)$$

$$Log\_SI = \log_{10}(SI)$$

- Создание новых признаков:
  - Произведение MolLogP и MolWt: $MolLogP \times MolWt$

  - Полиномиальные признаки второй степени
  - Бинарный признак $MolLogP\_gt\_3$

- Обработка выбросов для признаков с асимметричным распределением

2.3. Важные молекулярные дескрипторы

2.3.1. Топ-10 наиболее важных дескрипторов для IC50

Дескриптор	Важность
VSA_EState4	0.062637
Chi1n	0.041884
FpDensityMorgan3	0.040449
Chi4v	0.039201
Chi2v	0.038437
BCUT2D_MRLOW	0.035050
Chi2n	0.024153
SlogP_VSA5	0.020892
Chi3n	0.020464
EState_VSA4	0.020019

Таблица 1: Важные дескрипторы для IC50

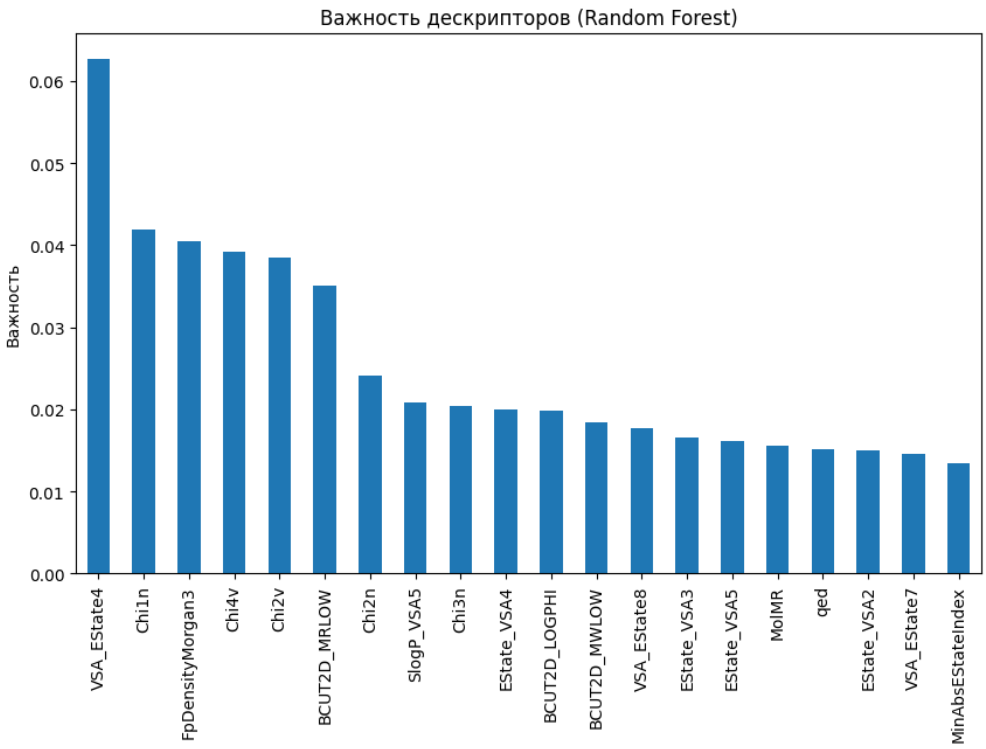

Рис. 3: График важности дескрипторов для IC50

2.3.2. Топ-10 наиболее важных дескрипторов для CC50

Дескриптор	Важность
LabuteASA	0.057715
BCUT2D_MWLOW	0.042473
Chi1	0.038697
Kappa3	0.038478
Kappa2	0.037475
FpDensityMorgan1	0.034774
BCUT2D_MRLOW	0.030914
MolMR	0.026832
VSA_EState7	0.023108
Ipc	0.022703

Таблица 2: Важные дескрипторы для CC50

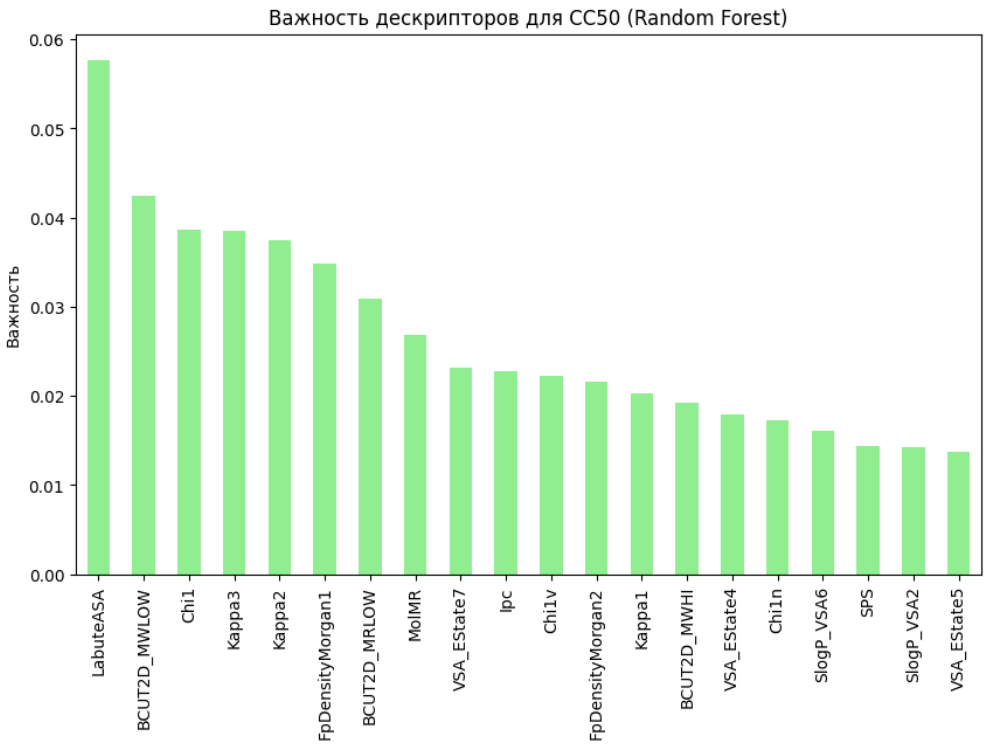

Рис. 4: График важности дескрипторов для CC50

2.3.3. Топ-10 наиболее важных дескрипторов для SI

Дескриптор	Важность
VSA_EState6	0.436867
VSA_EState2	0.073586
MinPartialCharge	0.048444
BalabanJ	0.041674
qed	0.039508
MaxAbsPartialCharge	0.031512
VSA_EState9	0.014044
VSA_EState8	0.013034
SMR_VSA10	0.012428
FractionCSP3	0.012006

Таблица 3: Важные дескрипторы для SI

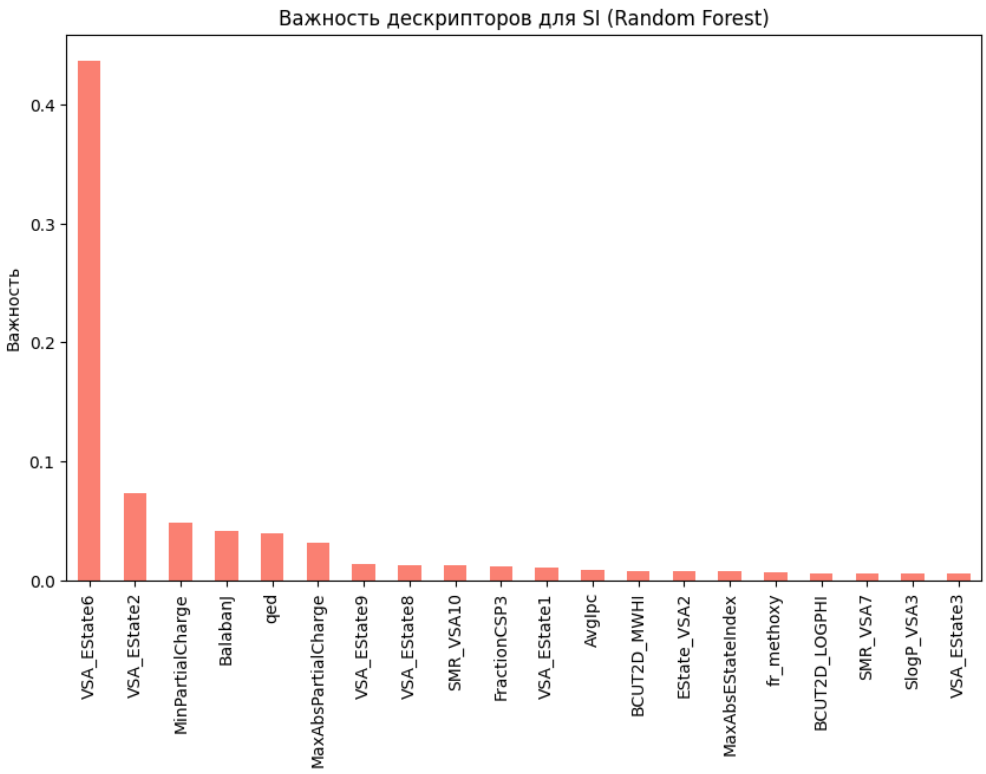

Рис. 5: График важности дескрипторов для SI

3. Feature Engineering и построение моделей

3.1. 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

3.2. 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.

3.3. Процесс оценки моделей

Методика оценки включала:

- Разбиение данных на обучающую и тестовую выборки (80%/20%)
- Заполнение пропущенных значений медианой
- Оценку качества в логарифмическом и исходном масштабе
- Использование метрик:
  - Среднеквадратичная ошибка (MSE)
  - Корень из среднеквадратичной ошибки (RMSE)
  - Коэффициент детерминации (R2)
- Тестирование с разным количеством признаков (от 1 до 50)

3.4. Результаты моделирования для задач Регрессии

Анализ результатов показал:

- Наилучшие результаты показали модели CatBoost и RandomForest.
- Логарифмирование целевых переменных принципиально качество моделей не улучшило.
- Удаление выбросов и лишних колонок в целом дало + к значению метрик.

Таргет	Модель	MSE	RMSE	R^2
CC50 (мМ)	CatBoost	203,547.99	451.16	0.607
IC50 (мМ)	Random Forest	194,487.81	441.01	0.417
SI (безразм.)	CatBoost	2,200.43	46.91	0.370

3.5. Результаты моделирования для задач Классификации

Анализ результатов показал:

- Все модели показывают относительно скромные результаты (Accuracy 50-66, что указывает на сложность задачи классификации для данного набора данных.
- Наилучшие значения у моделей градиентного бустинга: HistGradientBoos CatBoostClassifier, Stacking
- Удаление выбросов и лишних колонок в целом дало + к значению метрик.

Таргет	Модель	Accuracy	ROC AUC	F1
CC50	StackingClassifier	0.77	0.84	0.77
IC50	GradBoostClassifier	0.76	0.84	0.75
SI	CatBoostClassifier	0.76	0.78	0.72
SI (> 8)	HGBClassifier	0.69	0.74	0.54

Заключение

В ходе выполнения курсовой работы был проведен комплексный анализ данных о 1000 химических соединениях и их активности против вируса гриппа. 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 различные группы дескрипторов

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=0.417)

Для SI наилучший результат CatBoost (R^2=0.370)

Логарифмирование целевых переменных не привело к значимому улучшению качества моделей

Для разных целевых переменных лучшими стали:

CC50 → Stacking (Accuracy 0.772, ROC AUC 0.841)

IC50 → GradientBoosting (Accuracy 0.767, ROC AUC 0.844)

SI → CatBoost (Precision 0.857)

SI>8 → HistGradientBoosting (Accuracy 0.698)

Как и в регрессии:

Бустинг-модели (CatBoost, GBM) лидируют

Stacking эффективен для сложных случаев

SI (>8) предсказывается хуже всего (F1=0.543)