Аналитический отчет по курсовой работе Классическое машинное обучение

1 Цель и задачи исследования

Целью данной работы является разработка и сравнительный анализ моделей машинного обучения для решения задач регрессии и классификации, направленных на оценку эффективности химических соединений против вируса гриппа. Основные задачи включают:

- Регрессия для предсказания значений IC50 (концентрация соединения, подавляющая 50% вирусной активности, мМ).
- Регрессия для предсказания значений CC50 (концентрация соединения, вызывающая гибель 50% клеток, мМ).
- Регрессия для предсказания индекса селективности (SI), рассчитываемого как отношение CC50 к IC50.
- Классификация: определение, превышает ли значение IC50 медианное значение выборки.
- Классификация: определение, превышает ли значение CC50 медианное значение выборки.
- Классификация: определение, превышает ли значение SI медианное значение выборки.
- Классификация: определение, превышает ли значение SI порог 8 (индикатор высокой селективности).

2 Исходные данные

Исходный набор данных включает информацию о 1000 химических соединениях, характеризующихся числовыми параметрами, связанными с их молекулярными свойствами, и целевыми показателями эффективности: IC50, CC50 и SI. Значение SI определяется как $\mathrm{SI} = \frac{\mathrm{CC50}}{\mathrm{IC50}}$, где более высокие значения указывают на лучшую селективность препарата. Порог SI > 8 считается индикатором высокой эффективности против вирусной инфекции.

3 Анализ данных (EDA)

3.1 Предобработка данных

Для устранения правосторонней асимметрии распределений целевых переменных (IC50, CC50, SI) применено логарифмическое преобразование, что позволило приблизить распределения к нормальному виду. Пропуски в данных, составляющие незначительную долю, были заполнены медианными значениями. Признаки с нулевой вариативностью (константные значения) удалены из набора данных.

3.2 Анализ выбросов

Выбросы в целевых переменных анализировались с использованием метода межквартильного размаха (IQR) и Z-оценки на логарифмированных данных. Метод IQR выявил большее количество выбросов, чем Z-оценка, что указывает на различия в чувствительности методов к аномалиям. Для дальнейшего анализа использовались данные с минимальной обработкой выбросов, чтобы сохранить информацию.

3.3 Корреляционный анализ

Корреляционный анализ выявил умеренную корреляцию между IC50 и CC50 (коэффициент корреляции 0.521). Корреляция между IC50, CC50 и SI оказалась слабой, что ожидаемо, учитывая производную природу SI. Некоторые признаки, такие как молекулярный вес (MolWt), мера молекулярного объема и поляризуемости (MolMR), площадь доступной растворителю поверхности (LabuteASA) и количество ротируемых связей (NumRotatableBonds), показали высокую корреляцию между собой, что указывает на их взаимосвязь.

4 Методология и результаты

4.1 Регрессия IC50

Для задачи регрессии IC50 применялись модели линейной регрессии, Ridge-регрессии, Lasso-регрессии, случайного леса, XGBoost и LightGBM. Данные подготовлены следующим образом: логарифмическое преобразование целевой переменной, удаление константных признаков, заполнение пропусков медианными значениями.

Наилучший результат показала модель случайного леса (200 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- $R^2 = 0.45$ (объясняет 45% дисперсии данных).
- RMSE = 1.45.

Топ-5 наиболее важных признаков: VSA_EState8, VSA_EState4, VSA_EState6, BCUT2D_MRLOW, PEOE_VSA1.

4.2 Регрессия CC50

Аналогичный подход применялся для регрессии CC50. Наилучший результат достигнут с моделью случайного леса (100 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- $R^2 = 0.43$.
- RMSE = 1.14.

Топ-5 наиболее важных признаков: BCUT2D_MWLOW, NHOHCount, Kappa1, VSA_EState8, VSA_EState6.

4.3 Регрессия SI

Для регрессии SI использовались те же модели и подход к предобработке данных. Наилучший результат показала модель случайного леса (100 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- $R^2 = 0.34$.
- RMSE = 1.26.

Топ-5 наиболее важных признаков: VSA_EState6, VSA_EState8, SMR_VSA7, BCUT2D_CHGLO, BCUT2D_MRLOW.

4.4 Классификация IC50 > медианы

Для классификации IC50 относительно медианы (46.59) создана бинарная целевая переменная (0: $\leq$ медианы, 1: > медианы). Применялись модели логистической регрессии, случайного леса, XGBoost и LightGBM. Данные подготовлены аналогично задачам регрессии, с проверкой баланса классов.

Наилучший результат достигнут с моделью случайного леса (100 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- Точность (accuracy) = 0.73.
- F1-score = 0.73.

Топ-5 наиболее важных признаков: VSA_EState8, BCUT2D_MRLOW, SlogP_VSA5, Kappa3, PEOE_VSA7.

4.5 Классификация CC50 > медианы

Для классификации CC50 относительно медианы (411.04) применялся аналогичный подход. Наилучший результат показала модель случайного леса (100 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- Точность (accuracy) = 0.79.
- F1-score = 0.79.

Топ-5 наиболее важных признаков: NHOHCount, PEOE_VSA7, VSA_EState4, SMR_VSA5, BCUT2D_MWLOW.

4.6 Классификация SI > медианы

Для классификации SI относительно медианы (3.85) использовались те же модели. Наилучший результат достигнут с моделью случайного леса (50 деревьев, максимальная глубина 10, минимальный порог разделения узла 5). Метрики качества:

- Точность (accuracy) = 0.67.
- F1-score = 0.64.

Топ-5 наиболее важных признаков: BCUT2D_MRLOW, BCUT2D_MWLOW, VSA_EState4, BCUT2D_LOGPHI, MinEStateIndex.

4.7 Классификация SI > 8

Для классификации SI относительно порога 8 создана бинарная целевая переменная (0: ≤ 8, 1: > 8). Наилучший результат показала модель XGBoost (50 деревьев, максимальная глубина 5, скорость обучения 0.1). Метрики качества:

- Точность (accuracy) = 0.72.
- F1-score = 0.53.

Топ-5 наиболее важных признаков: SMR_VSA7, fr_allylic_oxid, SMR_VSA4, BCUT2D_CHGLO, fr_ether.

5 Выводы

В рамках исследования были разработаны и протестированы модели машинного обучения для задач регрессии и классификации параметров эффективности химических соединений (IC50, CC50, SI). Модели случайного леса и XGBoost продемонстрировали наилучшие результаты. Оптимальные гиперпараметры подбирались индивидуально для каждой задачи, что позволило достичь приемлемых метрик качества.

6 Рекомендации

Для улучшения результатов и дальнейшего развития исследования предлагаются следующие шаги:

- Разработка новых признаков на основе комбинации коррелированных параметров для выявления наиболее значимых характеристик соединений.
- Применение методов снижения размерности, таких как метод главных компонент (PCA), для уменьшения количества признаков и устранения мультиколлинеарности.
- Проведение более детального анализа выбросов с использованием комбинированных подходов для их обработки.
- Тестирование дополнительных моделей машинного обучения (например, нейронных сетей) с тщательным подбором гиперпараметров.