МИФИ

Национальный исследовательский ядерный университет

Направление: **Классическое машинное обучение**

Прогнозирование биологической активности химических соединений методами машинного обучения

Курсовая работа

Студента группы **М24-525**

Жадаева Василия Васильевича

1. Постановка исследовательской задачи

Требуется разработать прогностические модели для оценки ключевых параметров лекарственных соединений:

- **Регрессионные задачи:**
  - Предсказание IC_{50} (ингибирующая концентрация)
  - Предсказание CC_{50} (цитотоксическая концентрация)
  - Расчет индекса селективности (SI)
- **Классификационные задачи:**
  - Бинарная классификация превышения медианных значений IC_{50}, CC_{50}, SI
  - Классификация SI > 8 (порог селективности)

Цель: Оптимизация подбора соединений с высокой противовирусной активностью и низкой токсичностью.

2. Методология

2.1. Предобработка данных

- **Feature Engineering:**
  - Логарифмирование целевых переменных (IC_{50}, CC_{50}, SI) для нормализации распределений.
  - Генерация полиномиальных признаков 2-й степени.
  - Создание бинарных признаков на основе пороговых значений.
- **Обработка выбросов:**
  - Удаление аномалий по правилу 1.5×IQR.
  - Замена пропущенных значений медианами.

2.2. Используемые алгоритмы

- **Регрессия:**
  - CatBoost, Random Forest, Gradient Boosting, XGBoost.
- **Классификация:**
  - Stacking (ансамбли CatBoost + Random Forest), HistGradientBoosting.

2.3. Метрики оценки

- **Регрессия:** MSE, RMSE, R^2.
- **Классификация:** Accuracy, ROC AUC, F1-score.

3. Результаты

3.1. Регрессионный анализ

Целевая переменная	Лучшая модель	MSE	R^2
CC_{50}	CatBoost	203548	0.607
IC_{50}	Random Forest	194488	0.417
SI	Stacking	63	0.133

Выводы:

- CatBoost демонстрирует наивысшую точность для CC_{50}
- Random Forest лучше предсказывает IC_{50}, но объясняет лишь 41.7% дисперсии.
- Stacking лучше предсказывает IC_{50}, но объясняет лишь 13.3% дисперсии.

3.2. Классификация

Задача	Лучшая модель	Accuracy	ROC AUC
CC_{50} > медиана	StackingClassifier	0.772	0.841
IC_{50} > медиана	GradientBoostingClassifier	0.767	0.844
SI > медиана	GradientBoostingClassifier	0.603	0.656
SI > 8	XGBClassifier	0.729	0.743

- **Выводы**
- С классификацией SI имеются какие-то фундаментальные проблемы

4. Заключение

В рамках данного проекта был выполнен всесторонний анализ данных, содержащих информацию о тысяче химических соединений и их противовирусной активности. Ключевые результаты представлены ниже.

Основные достижения

1. **Исследовательский анализ данных (EDA)**
    - Обнаружены существенные различия в распределениях ключевых показателей: IC_{50}, CC_{50} и индекса селективности (SI).
    - Выявлено значительное количество выбросов, особенно в значениях SI.
    - Установлены умеренные, но статистически значимые корреляции между молекулярными дескрипторами и целевыми переменными.
2. **Предварительная обработка данных**
    - Разработана и успешно применена методика обработки выбросов.
    - Проведено логарифмическое преобразование целевых переменных.
    - Создан набор новых производных признаков, повышающих информативность данных.
3. **Интерпретация результатов**
    - Определены наиболее значимые молекулярные дескрипторы для каждого исследуемого параметра.
    - Установлено, что показатель VSA_EState обладает наибольшей предсказательной способностью в отношении индекса селективности.
    - Для IC_{50} и CC_{50} наиболее информативными оказались различные группы дескрипторов.
4. **Построение и оценка моделей**

Ключевые наблюдения

- Задача классификации значений SI остается наиболее сложной, что указывает на необходимость дальнейшего совершенствования методов.

Приложения:

- Исходный код: [https://github.com/Turchkas/SkillFactory/tree/main/%D0%9A%D1%83%D1%80%D1%81%D0%BE%D0%B2%D0%B0%D1%8F]