Отчет о работе:

Цель работы — создание нескольких максимально эффективных моделей для решения следующих задач:

- Регрессия для IC50

- Регрессия для CC50

- Регрессия для SI

- Классификация: превышает ли значение IC50 медианное значение выборки

- Классификация: превышает ли значение CC50 медианное значение выборки

- Классификация: превышает ли значение SI медианное значение выборки

Целевые переменные исследования:

- IC50: Более низкие значения указывают на более высокую противовирусную активность
- CC50: Более высокие значения указывают на меньшую токсичность

- SI (индекс селективности = CC50 / IC50): Чем выше значение, тем лучше. SI > 8 считается хорошим показателем для разработки вакцины против гриппа

Результаты работы:

1. Exploratory Data Analysis (EDA) для данных о химических соединениях:

- Обнаружены выбросы, особенно в CC50 и SI

- Некоторые признаки имеют постоянные значения и могут быть удалены

2. Регрессионный анализ для предсказаний IC50, CC50, SI

- Вывод по CC50: - Лучшая модель — XGBoost, так как у нее самый низкий средний RMSE (0.9095).

- Вывод по IC50 – Лучшая модель - Gradient Boosting, так как у нее самый низкий средний RMSE (0.8193)

- Вывод по SI – Лучшая модель - XGBoost, так как у нее самый низкий средний RMSE (0.881).

3. Классификация для IC50, CC50, SI

Создание целевых переменных, где IC50, CC50, SI > медианы и SI > 8

- Удаление константных столбцов и заполнение пропущенных значений медианой - Для обучения были использованы следующие модели:

- LogisticRegression

- Random Forest

- XGBoost

- Вывод по CC50 > медианы: - Лучшая модель — XGBoost с accuracy 0.8111

- Вывод по IC50 > медианы: - Лучшая модель — Random Forest с accuracy 0.7111

- Вывод по SI > медианы: - Лучшая модель — Random Forest с accuracy 0.5444

Заключение

В ходе работы были построены модели для предсказания IC50, CC50 и SI.

Для регрессионного анализа в большинстве случаев лучшей моделью стала XGBoost, а для классификации случайный лес.

Рекомендации по улучшению:

- Объединение связанных между собой признаков

- Провести анализ выбросов с целью определения их важности для дальнейшего исследования