

# Аналитический отчёт

**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
- Классификация: превышает ли значение SI значение 8

**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

**Этапы работы:**

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

- Логарифмическое преобразование целевых переменных для нормализации, так как IC50, CC50 и SI имеют правостороннее распределение
- Обнаружены выбросы, особенно в CC50 и SI
- SI имеет сильную корреляцию с IC50 и CC50 (что ожидаемо, так как  $SI = CC50/IC50$ )
- Некоторые признаки имеют высокую корреляцию между собой так как обозначают разные, но связанные понятия в химии. Например, «MolWt» (молярный вес) и «MolMr» (относительная молекулярная масса)
- Некоторые признаки имеют постоянные значения и могут быть удалены

## 2. Регрессионный анализ для предсказаний IC50, CC50, SI

- На основании EDA произведено логарифмическое преобразование целевых переменных, удалены константные столбцы и пропущенные значения заполнены медианой

- Для обучения были использованы следующие модели:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM
- CatBoost

**- Вывод по IC50:**

- Лучшая модель — LightGBM. Модель объясняет около 47% дисперсии данных ( $R^2 = 0.47$ )
- Наиболее важными признаками для предсказания IC50 являются: PEOE\_VSA6, VSA\_EState4, BCUT2D\_MRLOW, EState\_VSA5, VSA\_EState6, EState\_VSA4, qed, FpDensityMorgan

**- Вывод по CC50:**

- Лучшая модель — XGBoost. Модель объясняет около 47% дисперсии данных ( $R^2 = 0.47$ )
- Наиболее важными признаками для предсказания CC50 являются: fr\_phenol, Kappa1, fr\_C\_O, MolMR, NumHDonors, VSA\_EState8, fr\_Ar\_OH, BCUT2D\_MWLOW

**- Вывод по SI:**

- Лучшая модель — XGBoost. Модель объясняет около 36% дисперсии данных ( $R^2 = 0.36$ )
- Наиболее важными признаками для предсказания SI являются: SMR\_VSA7, NumHDonors, Chi2n, MolMR, BCUT2D\_CHGLO

### **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
- CatBoost
  
- **Вывод по CC50 > медианы:**
  - Лучшая модель — XGBoost с accuracy 0.8060
  - Наиболее важные признаки для классификации: NHOHCount, BCUT2D\_MRLOW, NumAromaticHeterocycles, BCUT2D\_MWLOW, fr\_imide
  
- **Вывод по IC50 > медианы:**
  - Лучшая модель — XGBoost с accuracy 0.7114
  - Наиболее важные признаки для классификации: VSA\_EState8, NumHDonors, PEOE\_VSA1, SlogP\_VSA1, fr\_ketone\_Topliiss
  
- **Вывод по SI > медианы:**
  - Лучшая модель — XGBoost с accuracy 0.6667
  - Наиболее важные признаки для классификации: NumSaturatedCarbocycles, SMR\_VSA7, BCUT2D\_MRLOW, NHOHCount, FractionCSP3
  
- **Вывод по SI > 8:**
  - Лучшая модель — XGBoost с accuracy 0.7264
  - Наиболее важные признаки для классификации: SMR\_VSA7, BCUT2D\_CHGLO, fr\_bicyclic, Chi1n, BertzCT

## Заключение

В ходе работы были построены модели для предсказания IC50, CC50 и SI. Лучшие результаты показала модель XGBoost — как для задач регрессии, так и для задач классификации

Рекомендации по улучшению:

- Объединение связанных между собой признаков
- Провести анализ выбросов с целью определения их важности для дальнейшего исследования
- Попробовать ансамбли моделей, например, стекинг