**Курсовая работа: Моделирование эффективности лекарств**

# 1. Введение

Цель проекта - построить модели, предсказывающие эффективность лекарственных соединений на основе молекулярных дескрипторов. Рассматриваются как регрессионные, так и классификационные задачи, связанные с метриками IC50, CC50 и SI (индекс селективности).

# 2. Анализ данных (EDA)

Данные содержат 214 признаков и 1000+ наблюдений. Целевые переменные: IC50 (ингибирующая концентрация), CC50 (токсичность) и SI (индекс селективности = CC50 / IC50).

Были построены распределения, найдены медианные значения (IC50 ≈ 46.59, CC50 ≈ 411.04, SI ≈ 3.85). Выбросы особенно выражены в SI. Признаки с большим количеством пропусков удалены, остальные заполнены медианой.

# 3. Построение моделей

## 3.1 Регрессия

• IC50: RMSE = 440.56, MAE = 235.02, R² = 0.418

• CC50: RMSE = 459.65, MAE = 286.38, R² = 0.592

• SI: RMSE = 1353.90, MAE = 191.04, R² = 0.087

## 3.2 Классификация

• IC50 > медианы: Accuracy = 0.75, F1 = 0.75, AUC = 0.78

• CC50 > медианы: Accuracy = 0.81, F1 = 0.81, AUC = 0.88

• SI > медианы: Accuracy = 0.68, F1 = 0.66, AUC = 0.72

• SI > 8: Accuracy = 0.70, F1 = 0.49, AUC = 0.72

# 4. Сравнение моделей

Лучшие результаты в регрессии показала модель для CC50 (R² = 0.592). В классификации лучший результат по всем метрикам достигнут для задачи CC50 > медианы (AUC = 0.88). SI сложен для моделирования из-за нестабильного распределения и выбросов.

# 5. Выводы и рекомендации

• Использование Random Forest даёт хорошие результаты без сложной настройки.  
• Для SI стоит попробовать логарифмическую трансформацию.  
• Наиболее перспективные признаки можно отобрать по важности из модели.  
• Возможно применение градиентного бустинга (XGBoost, CatBoost) для повышения качества.