# Аналитический отчёт по курсовой работе

## 1. Введение

Цель проекта – применение классических методов машинного обучения для прогноза активности (IC50), токсичности (CC50) и избирательности (SI) химических соединений против вируса гриппа на основе их дескрипторов.

## 2. Разведочный анализ данных (EDA)

• Данные: 998 соединений, 195 дескрипторов, три целевые переменные (IC50, CC50, SI).  
• Удалены технические и константные признаки; пропуски удалены.  
• Применено логарифмическое преобразование распределений.  
• Spearman корреляции: IC50↔CC50 ≈ 0.62, IC50↔SI ≈ -0.64, CC50↔SI ≈ 0.53.  
• Ни один дескриптор не имеет сильной корреляции (|ρ|<0.3) с целевыми переменными.

## 3. Регрессия

Задачи регрессии:  
• IC50: XGBoost R²=0.468, MAE≈158  
• CC50: XGBoost R²=0.468, MAE≈293  
• SI: лучшая модель R²≈0.09  
Вывод: регрессия IC50 и CC50 возможна с помощью XGBoost/LightGBM; прямая регрессия SI неэффективна.

## 4. Классификация

Задачи классификации:  
• IC50 > медианы: RandomForest Acc=0.74, ROC AUC=0.794  
• CC50 > медианы: XGBoost Acc=0.73, ROC AUC=0.800  
• SI > медианы: LightGBM Acc=0.635, ROC AUC=0.675  
• SI > 8: LightGBM Acc=0.73, F1=0.614, ROC AUC=0.724  
Вывод: градиентный бустинг и случайный лес — лучшие методы для задач классификации.

## 5. Выводы и рекомендации

1. XGBoost и LightGBM демонстрируют наилучшие результаты во всех задачах.  
2. Для отбора высокоселективных соединений рекомендована классификация SI > 8.  
3. Для улучшения моделей можно использовать feature engineering, расширенный подбор гиперпараметров, сбор новых признаков.  
4. Подход позволяет ускорить приоритизацию кандидатов в противовирусной разработке.