# Институт интеллектуальных кибернетических систем НИЯУ МИФИ

# Группа: М24-525

# Студент: Колесников Владислав Вячеславович

# Курсовая работа: классическое машинное обучение для прогнозирования фармакологических показателей

## 

## Введение

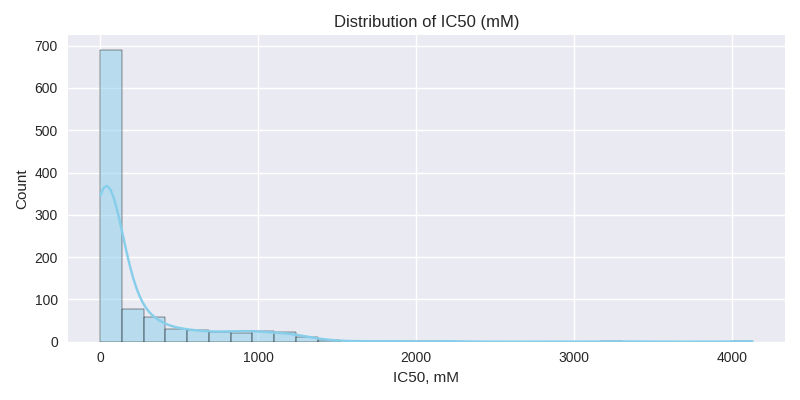
Фармакологические тесты оценивают эффективность и токсичность веществ несколькими ключевыми показателями. **IC50** (half‑maximal inhibitory concentration) — это концентрация вещества, при которой наблюдается 50 % ингибирование активности целевого фермента или вируса. **CC50** (50 % cytotoxic concentration) определяет концентрацию, при которой количество жизнеспособных клеток уменьшается наполовину по сравнению с контролем. **Selectivity Index (SI)** — отношение CC50 к IC50; чем выше SI, тем безопаснее соединение для клеток, поскольку оно обеспечивает желаемый эффект при минимальной токсичности. Цель работы — на основании химических дескрипторов из файла data.xlsx построить модели, которые могут предсказывать эти показатели и выполнять их двоичную классификацию по пороговым значениям.

Файл data.xlsx содержит 1001 строку и 213 столбцов, включая три целевых переменных. Столбец Unnamed: 0 оказался индексом и был удалён. Пропуски обнаружены в 12 дескрипторах, но число пропусков не превышает трёх в каждом, поэтому они заполнялись медианой.

## Исследовательский анализ данных (EDA)

### Обзор данных

После удаления индексного столбца в данных осталось 212 дескрипторов. Значения IC50 и CC50 распределены крайне неравномерно (рис. 1); большинство соединений обладают низкой активностью, а доля высокоактивных/высокотоксичных мала. SI варьирует на четыре порядка величины и имеет сильную правостороннюю асимметрию. В регрессионных задачах использовалась десятичная логарифмизация целей, что уменьшает разброс и облегчает обучение моделей.



*Рис. 1. Распределения IC50, CC50 и SI. По часовой стрелке: IC50, CC50, SI.*

### Корреляционный анализ

Для каждой цели были рассчитаны коэффициенты корреляции Пирсона между целью и дескрипторами. Выявленные связи невелики: наибольшая корреляция для IC50 составляет ~0,27, для CC50 — ~0,31, для SI — ~0,16. Это говорит об отсутствии сильно линейно связанных дескрипторов и оправдывает использование нелинейных моделей.

### Топ‑5 дескрипторов для IC50:

| Дескриптор | Коэфф. Пирсона |
| --- | --- |
| VSA\_EState4 | ≈ 0,274 |
| Chi2n | ≈ 0,257 |
| PEOE\_VSA7 | ≈ 0,256 |
| Chi2v | ≈ 0,249 |
| fr\_Ar\_NH | ≈ 0,246 |

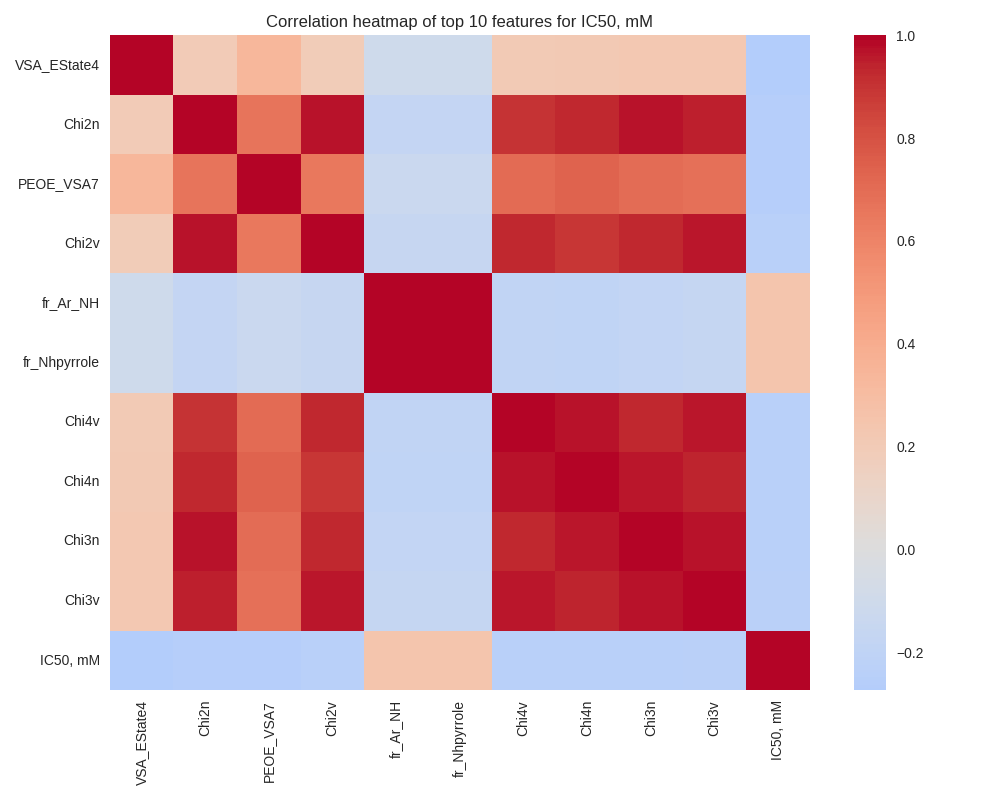
### Топ‑5 дескрипторов для CC50:

| Дескриптор | Коэфф. Пирсона |
| --- | --- |
| MolMR | ≈ 0,310 |
| LabuteASA | ≈ 0,309 |
| MolWt | ≈ 0,306 |
| ExactMolWt | ≈ 0,306 |
| HeavyAtomCount | ≈ 0,305 |

### Топ‑5 дескрипторов для SI:

| Дескриптор | Коэфф. Пирсона |
| --- | --- |
| BalabanJ | ≈ 0,163 |
| fr\_NH2 | ≈ 0,160 |
| RingCount | ≈ 0,124 |
| fr\_Al\_COO | ≈ 0,102 |
| fr\_COO2 | ≈ 0,101 |

Тепловая карта десяти наиболее коррелированных признаков с IC50 (рис. 2) подтверждает, что корреляции между признаками и целью невелики и не образуют простых линейных закономерностей.



*Рис. 2. Корреляции между IC50 и десятью наиболее связанными дескрипторами.*

## Постановка задач и методология

Задание включает три задачи регрессии (предсказание IC50, CC50 и SI) и четыре задачи классификации (определить, превышает ли каждая цель медиану выборки и превышает ли SI порог 8). Входными признаками были все дескрипторы, кроме целевых переменных. Пропуски заполнялись медианой. Для линейных моделей применялось стандартизирование, для ансамблевых деревьев — только импутация.

**Модели регрессии.** Были протестированы гребневая регрессия (Ridge), Lasso, Random Forest, Gradient Boosting и XGBoost. Гиперпараметры подбирались с помощью GridSearchCV (5‑кратная кросс‑валидация). Модели оценивались по среднеквадратичной ошибке (RMSE), средней абсолютной ошибке (MAE) и коэффициенту детерминации R².

**Модели классификации.** Для бинарных задач использовались логистическая регрессия, Random Forest, Gradient Boosting и XGBoost. Качество оценивалось по Accuracy, Precision, Recall, F1‑мере и площади под ROC‑кривой (ROC‑AUC). Порог медианы рассчитывался по всей выборке; для SI > 8 использовался фиксированный порог 8.

## Результаты регрессионных моделей

### Прогнозирование IC50

| Модель | RMSE (ср.) | RMSE (std) | MAE (ср.) | MAE (std) | R² (ср.) | R² (std) |
| --- | --- | --- | --- | --- | --- | --- |
| **Gradient Boosting** | **0,8651** | 0,1955 | **0,6748** | 0,1325 | **0,0534** | 0,2121 |
| XGBoost | 0,8656 | 0,1783 | 0,6766 | 0,1297 | 0,0427 | 0,2161 |
| Random Forest | 0,8695 | 0,2032 | 0,6794 | 0,1395 | 0,0421 | 0,2407 |
| Lasso | 0,9122 | 0,1523 | 0,7197 | 0,0887 | –0,0577 | 0,1203 |
| Ridge | 1,0738 | 0,1972 | 0,8282 | 0,1275 | –0,4690 | 0,2455 |

Наиболее точной оказалась модель градиентного бустинга: её RMSE немного меньше, чем у XGBoost и Random Forest, а R² ≈ 0,053 является единственным положительным среди методов. Линейные модели (Ridge и Lasso) предсказывают хуже и имеют отрицательные значения R².

### Прогнозирование CC50

| Модель | RMSE (ср.) | RMSE (std) | MAE (ср.) | MAE (std) | R² (ср.) | R² (std) |
| --- | --- | --- | --- | --- | --- | --- |
| **XGBoost** | **0,6986** | 0,1284 | **0,5440** | 0,0838 | –0,0898 | 0,2380 |
| Random Forest | 0,7063 | 0,1252 | 0,5601 | 0,0836 | –0,1036 | 0,1895 |
| Lasso | 0,7122 | 0,1255 | 0,5848 | 0,0884 | –0,1183 | 0,1849 |
| Gradient Boosting | 0,7215 | 0,1212 | 0,5683 | 0,0864 | –0,1637 | 0,2378 |
| Ridge | 0,8417 | 0,2242 | 0,6213 | 0,1201 | –0,6592 | 0,7771 |

Все R² оказались отрицательными, что говорит о слабой предсказуемости CC50. Лучшими по RMSE и MAE являются XGBoost и Random Forest. В то же время Ridge демонстрирует существенно худшие результаты, подтверждая, что линейные модели не подходят для этой задачи.

### Прогнозирование SI

| Модель | RMSE (ср.) | RMSE (std) | MAE (ср.) | MAE (std) | R² (ср.) | R² (std) |
| --- | --- | --- | --- | --- | --- | --- |
| **Random Forest** | **0,7522** | 0,1387 | **0,5900** | 0,1083 | –0,1073 | 0,1152 |
| Lasso | 0,7555 | 0,1386 | 0,6046 | 0,0950 | –0,1201 | 0,1445 |
| XGBoost | 0,7613 | 0,1269 | 0,5951 | 0,0970 | –0,1441 | 0,1411 |
| Gradient Boosting | 0,7723 | 0,1253 | 0,5989 | 0,0989 | –0,1805 | 0,1498 |
| Ridge | 0,9070 | 0,1472 | 0,6948 | 0,0848 | –0,6767 | 0,4488 |

Для прогноза SI все модели показывают отрицательные R²; лучшее качество по RMSE и MAE даёт Random Forest. Однако точность далека от удовлетворительной, поэтому для SI разумнее рассматривать задачу классификации по порогам.

## Результаты классификационных моделей

### Превышает ли IC50 медиану?

| Модель | Accuracy (ср.) | Precision (ср.) | Recall (ср.) | F1 (ср.) | ROC‑AUC (ср.) |
| --- | --- | --- | --- | --- | --- |
| **Gradient Boosting** | **0,5893** | **0,5922** | **0,5980** | **0,5849** | **0,5721** |
| XGBoost | 0,5653 | 0,5601 | 0,5840 | 0,5589 | 0,5621 |
| Random Forest | 0,5664 | 0,5579 | 0,5980 | 0,5652 | 0,5537 |
| Logistic Regression | 0,5333 | 0,5568 | 0,4820 | 0,5118 | 0,5231 |

Градиентный бустинг лидирует по всем показателям, но AUC всего ~0,57, что немного выше случайного угадывания. Это отражает слабую информативность дескрипторов для точного разделения IC50 относительно медианы.

### 

### Превышает ли CC50 медиану?

| Модель | Accuracy (ср.) | Precision (ср.) | Recall (ср.) | F1 (ср.) | ROC‑AUC (ср.) |
| --- | --- | --- | --- | --- | --- |
| **Gradient Boosting** | **0,5733** | 0,5727 | **0,6496** | **0,6050** | **0,5960** |
| Logistic Regression | 0,5644 | 0,5683 | 0,6175 | 0,5864 | 0,5939 |
| Random Forest | 0,5593 | 0,5608 | 0,6335 | 0,5919 | 0,5719 |
| XGBoost | 0,5673 | 0,5671 | 0,6635 | 0,6070 | 0,5776 |

Лучшие результаты по большинству метрик показывает градиентный бустинг (AUC≈0,60), хотя отрыв от логистической регрессии невелик. Случайный лес и XGBoost дают сопоставимые показатели.

### 

### Превышает ли SI медиану?

| Модель | Accuracy (ср.) | Precision (ср.) | Recall (ср.) | F1 (ср.) | ROC‑AUC (ср.) |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **0,5234** | **0,5426** | 0,5200 | 0,5034 | **0,5294** |
| XGBoost | 0,5054 | 0,5065 | 0,5040 | 0,4901 | 0,5137 |
| Gradient Boosting | 0,5134 | 0,5214 | 0,5080 | 0,4991 | 0,5046 |
| Logistic Regression | 0,4923 | 0,4805 | **0,5340** | **0,5026** | 0,4924 |

Разделение выборки по медиане SI оказалось самой трудной задачей: AUC едва превышает 0,53. Случайный лес показывает наибольший ROC‑AUC, хотя по F1 он уступает логистической регрессии.

### Превышает ли SI порог 8?

| Модель | Accuracy (ср.) | Precision (ср.) | Recall (ср.) | F1 (ср.) | ROC‑AUC (ср.) |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **0,6194** | **0,4525** | 0,2675 | 0,2849 | 0,5656 |
| Gradient Boosting | 0,6013 | 0,3492 | 0,2983 | 0,2996 | **0,5768** |
| XGBoost | 0,6043 | 0,3519 | 0,2761 | 0,2889 | 0,5675 |
| Logistic Regression | 0,6003 | 0,3696 | **0,3880** | **0,3644** | 0,5714 |

Задача SI > 8 предсказуема лучше: AUC приближается к 0,58. Случайный лес обеспечивает максимальную точность и наибольшую точность среди моделей, в то время как градиентный бустинг даёт чуть лучший ROC‑AUC. Выбор модели зависит от приоритетов (баланс между положительной предсказательной ценностью и чувствительностью).

## Обсуждение и выводы

1. **Сложность задач.** Входные дескрипторы демонстрируют слабые линейные связи с IC50, CC50 и SI; соответственно линейные методы (Ridge, Lasso) оказываются наименее эффективными. Ансамбли деревьев и бустинг‑модели дают лучшие результаты, но даже они имеют низкие R² и AUC, что указывает на высокую сложность задачи и необходимость дополнительных признаков.
2. **Оптимальные модели.** Для IC50 лучшим оказался градиентный бустинг, обеспечивающий минимальный RMSE и положительный R². В случае CC50 XGBoost немного опережает другие модели по ошибкам, но все R² отрицательны. Для SI минимальную ошибку обеспечивает Random Forest, хотя и с отрицательным R².
3. **Классификация.** Задачи классификации показали умеренные результаты. Для IC50 > медианы лучшим оказался градиентный бустинг. Для CC50 > медианы он также лидирует, хотя логистическая регрессия лишь немного уступает. В задачах с SI медиана Random Forest обеспечивает наибольший ROC‑AUC, а при пороге 8 наибольшую точность показывает тоже Random Forest, тогда как градиентный бустинг немного выигрывает по AUC.
4. **Рекомендации по улучшению.** Для повышения качества моделей следует:
5. провести детальный отбор и генерацию признаков, включая нелинейные комбинации и доменные химические индексы;
6. расширить датасет за счёт дополнительных соединений;
7. изучить методы отбора признаков (Recursive Feature Elimination, Permutation Importance) и снизить размерность до наиболее информативных дескрипторов;
8. протестировать альтернативные алгоритмы, такие как CatBoost и LightGBM, а также нейронные сети, которые могут лучше раскрыть сложные зависимости.

## Заключение

Работа продемонстрировала полный цикл решения задач классического машинного обучения: от исследовательского анализа химических данных до построения и оценки нескольких регрессионных и классификационных моделей. Полученные результаты показывают, что предсказание точных количественных значений IC50, CC50 и SI трудно из‑за слабых корреляций и ограниченного объёма данных. Тем не менее градиентный бустинг и Random Forest позволяют ранжировать соединения и давать приближённые оценки. Для практического использования необходимо продолжить исследование, улучшая признаки и расширяя набор данных. Наиболее перспективными направлениями являются применение более современных алгоритмов (LightGBM, CatBoost), а также создание доменно‑специфических дескрипторов, отражающих биологическую активность соединений.