Курсовая работа

Тема:
Применение методов машинного обучения для прогнозирования эффективности, токсичности и селективности химических соединений

Выполнил: Серов Илья Алексеевич

Содержание:

1. Введение

В рамках курсовой работы рассмотрен практический сценарий применения методов машинного обучения для анализа химико-биологических данных — задача предсказания активности, токсичности и селективности соединений (IC50, CC50, SI).

Проведён разведочный анализ данных (EDA), предобработка и очистка признаков, подбор гиперпараметров и сравнение моделей для задач регрессии и классификации.

В результате выбраны оптимальные алгоритмы, обеспечивающие лучшее качество предсказания для каждой задачи.

---

2. Описание данных

Исходные данные — таблица с 1001 химическим соединением и 210 признаками (дескрипторами)
Три целевые переменные:

- IC50 (концентрация для 50% ингибирования — индикатор эффективности)
- CC50 (концентрация для 50% токсичности — индикатор безопасности)
- SI (селективность, рассчитывается как CC50/IC50)

На этапе анализа обнаружено:

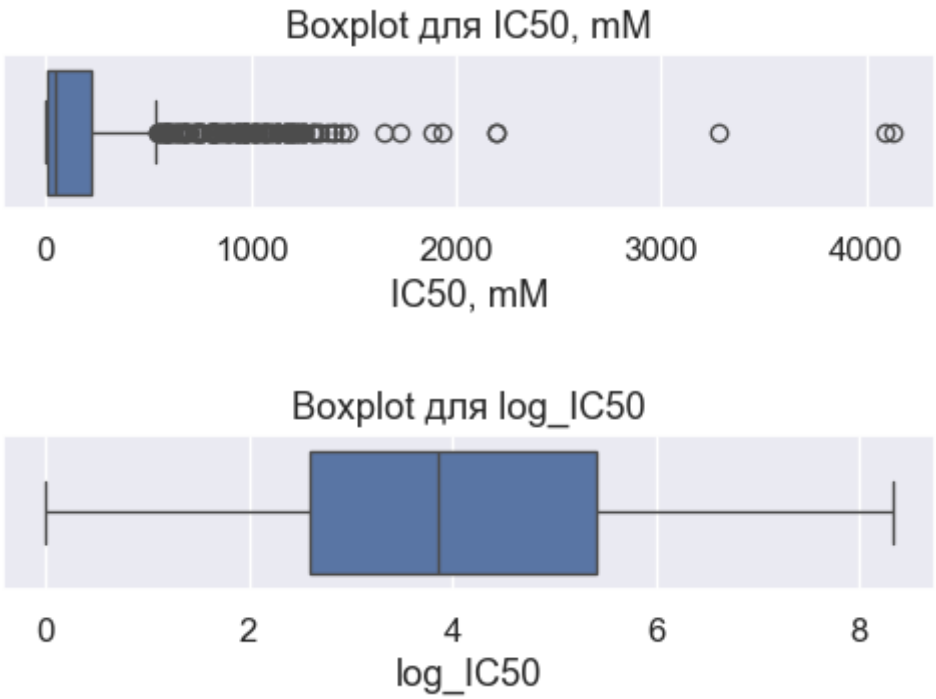
- Пропуски — <2%, заполнены медианой.
- Дубликаты строк (≈3%), удалены для предотвращения утечек.

- 18 признаков — постоянные, удалены.
- Около 90 пар признаков имеют очень высокую корреляцию (r > 0.95).
- Сильная правосторонняя асимметрия у всех таргетов.

---

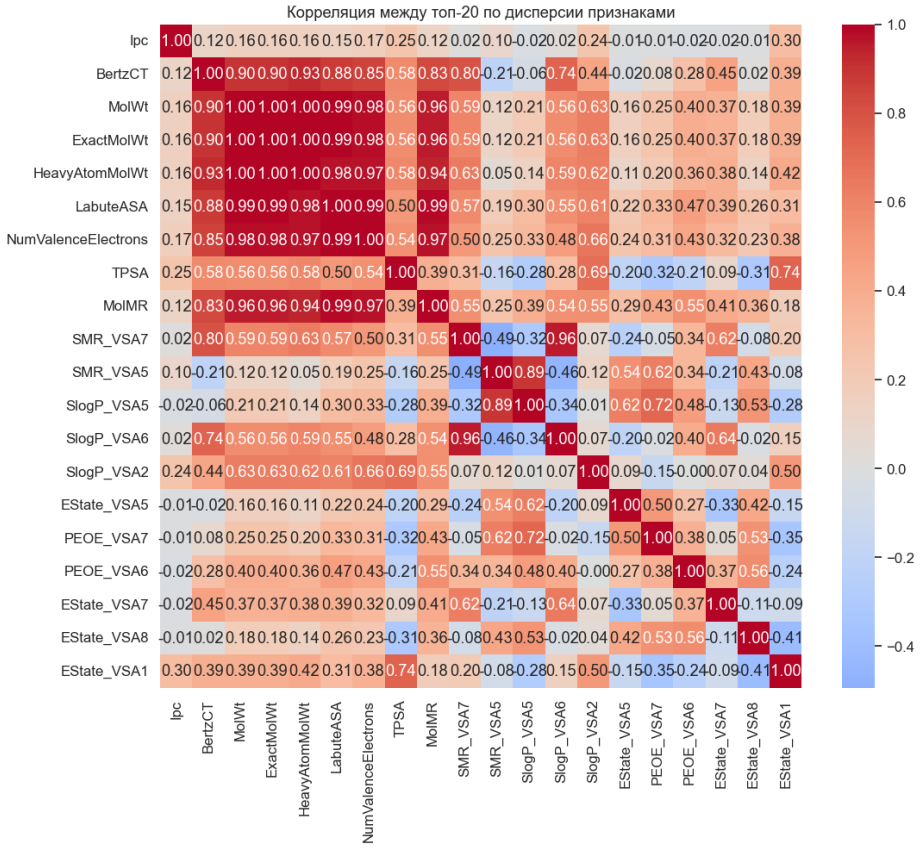
3. EDA

- **Распределения:**
  Все целевые переменные имеют длинные правые хвосты. Логарифмирование улучшает симметрию, приближая данные к нормальному распределению.



Пример – boxplot для IC50 до и после логарифмирования

- **Корреляция:**
  Некоторые дескрипторы существенно коррелируют между собой, но не с таргетами (максимальные ≈ 0.3).
- **Мультиколлинеарность:**
  Для линейных моделей избыточность является проблемой, для деревьев менее критично.
- **Были визуализированы:** распределения, boxplot’ы, матрица корреляций.



Матрица корреляций

---

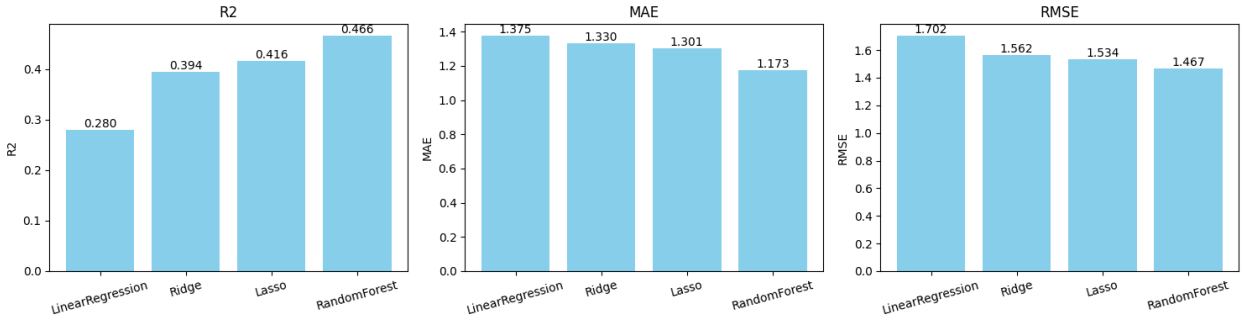
4. Ход работы

- **Предобработка:**
  - Дубликаты — удалены.
  - Пропуски — заполнены медианой.
  - Логарифмирование IC50, CC50, SI.
  - Удаление постоянных и сверхкоррелированных признаков.
- **Разделение на train/test (90/10)**
- **Модели:**
  - Регрессия: LinearRegression, Ridge, Lasso, RandomForestRegressor
  - Классификация: LogisticRegression, RandomForestClassifier, SVC
- **Параметры:**
  Для сложных моделей перебирались гиперпараметры через GridSearchCV.

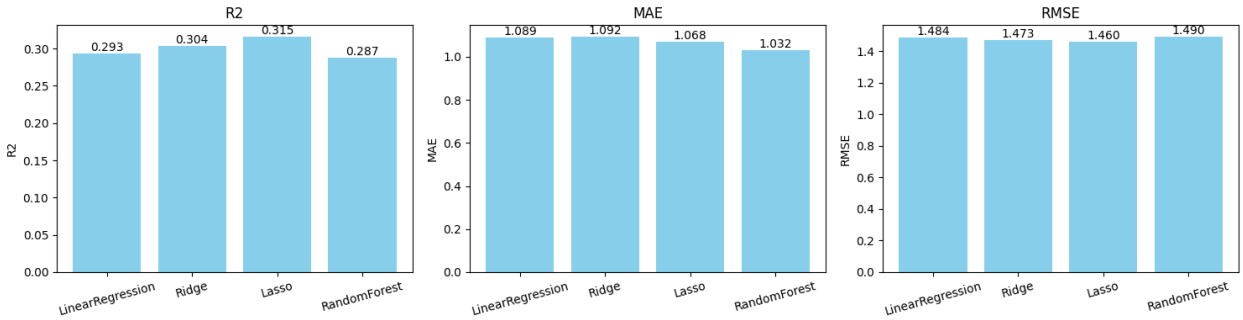
---

5. Результаты

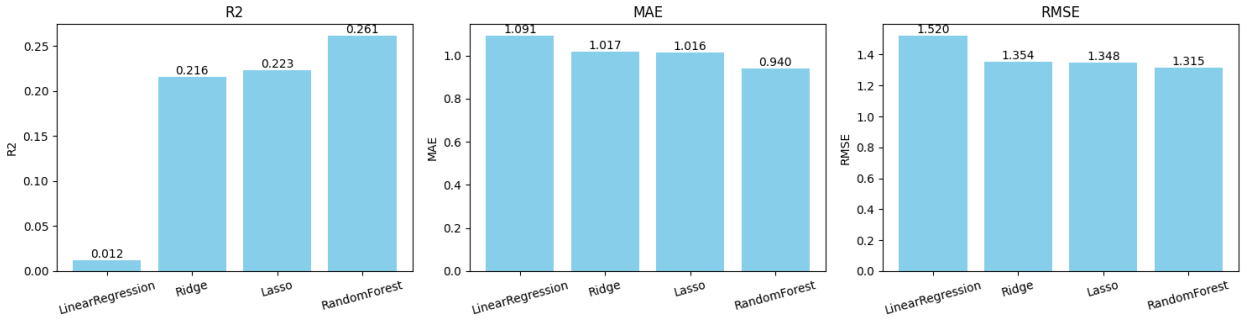
5.1 Регрессия



IC50



CC50



SI

Target	Лучшая модель	RMSE	MAE	R^2
IC50	RandomForest	1.47	1.17	0.47
CC50	Lasso	1.46	1.07	0.32
SI	RandomForest	1.31	0.94	0.26

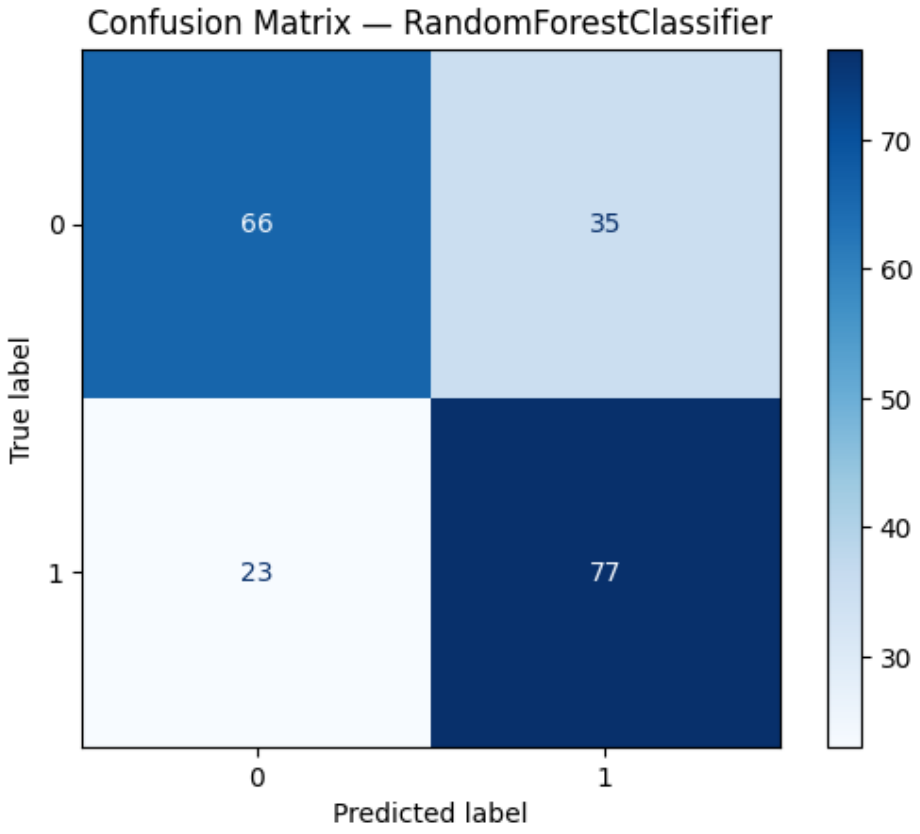
Комментарий по SI:
Также была выполнена прямая регрессия для SI. Для этого были экспортированы лучшие модели для IC50, CC50, т.к. по условию задачи значение SI рассчитывается на основе параметров IC50 и CC50. Она дала провальные метрики ($R^2 < 0$), вероятно, из-за ошибок в

подмоделях для IC50 и CC50. Попытка предсказывать log(SI) через две отдельные модели не привела к улучшению.

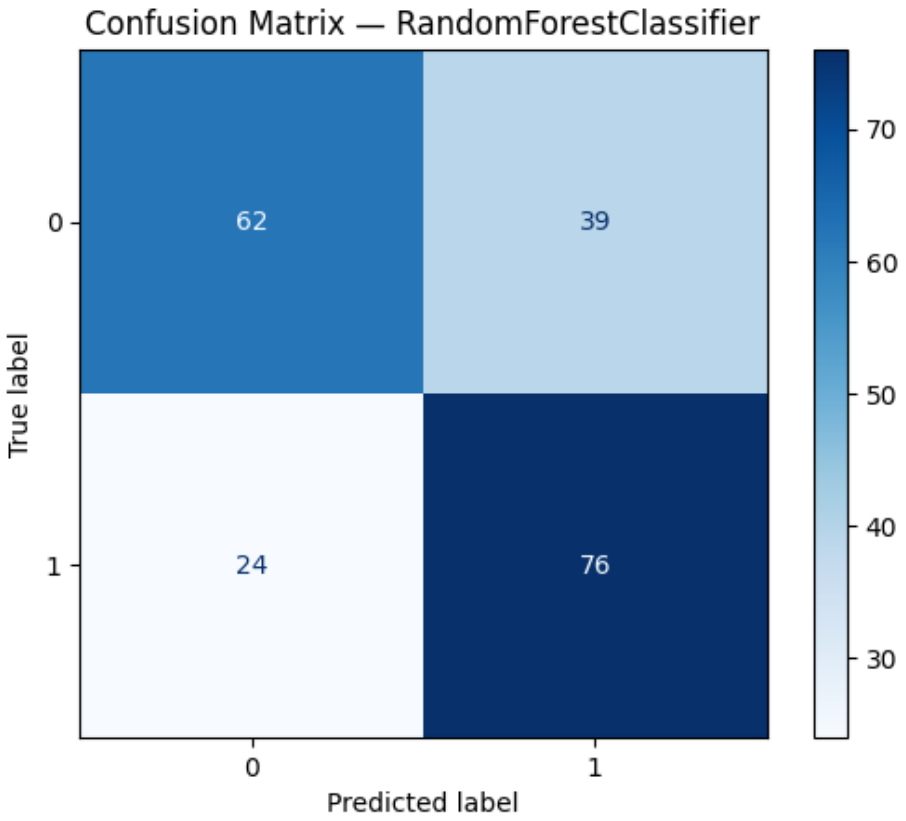
5.2 Классификация (по медианам и по SI>8)

Задача	Лучшая модель	Accuracy	F1	ROC-AUC
IC50 > median	RandomForest	0.71	0.73	0.79
CC50 > median	RandomForest	0.687	0.71	0.832
SI > median	SVC	0.68	0.62	0.73
SI > 8	RandomForest	0.74	0.57	0.75

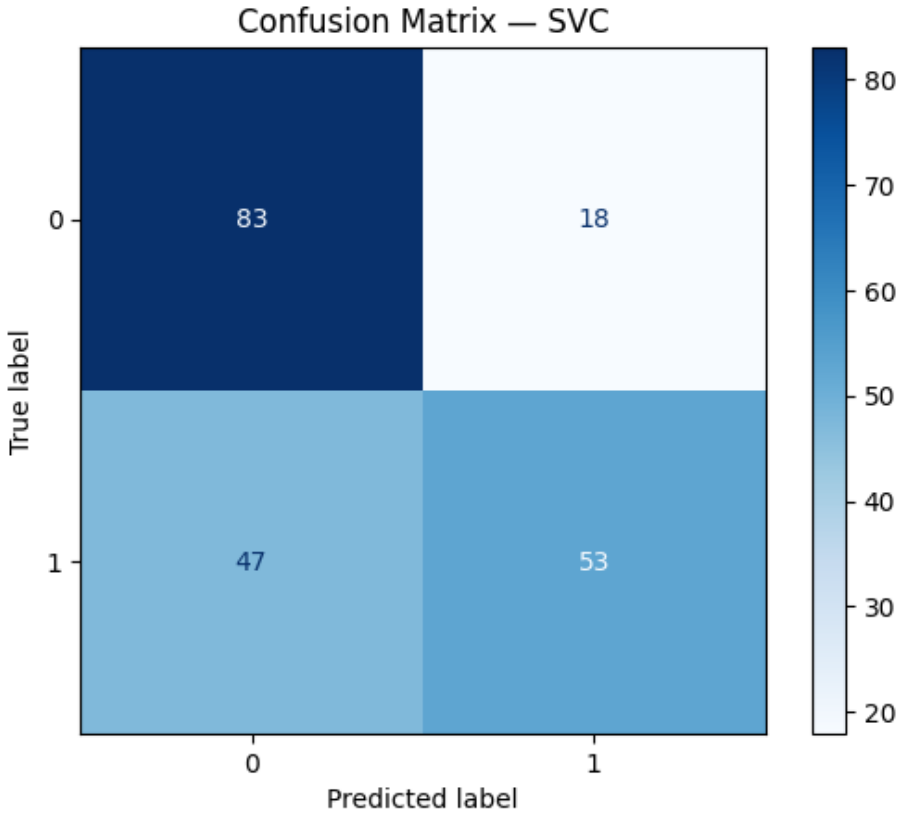
Матрицы ошибок для каждой из выбранных моделей:



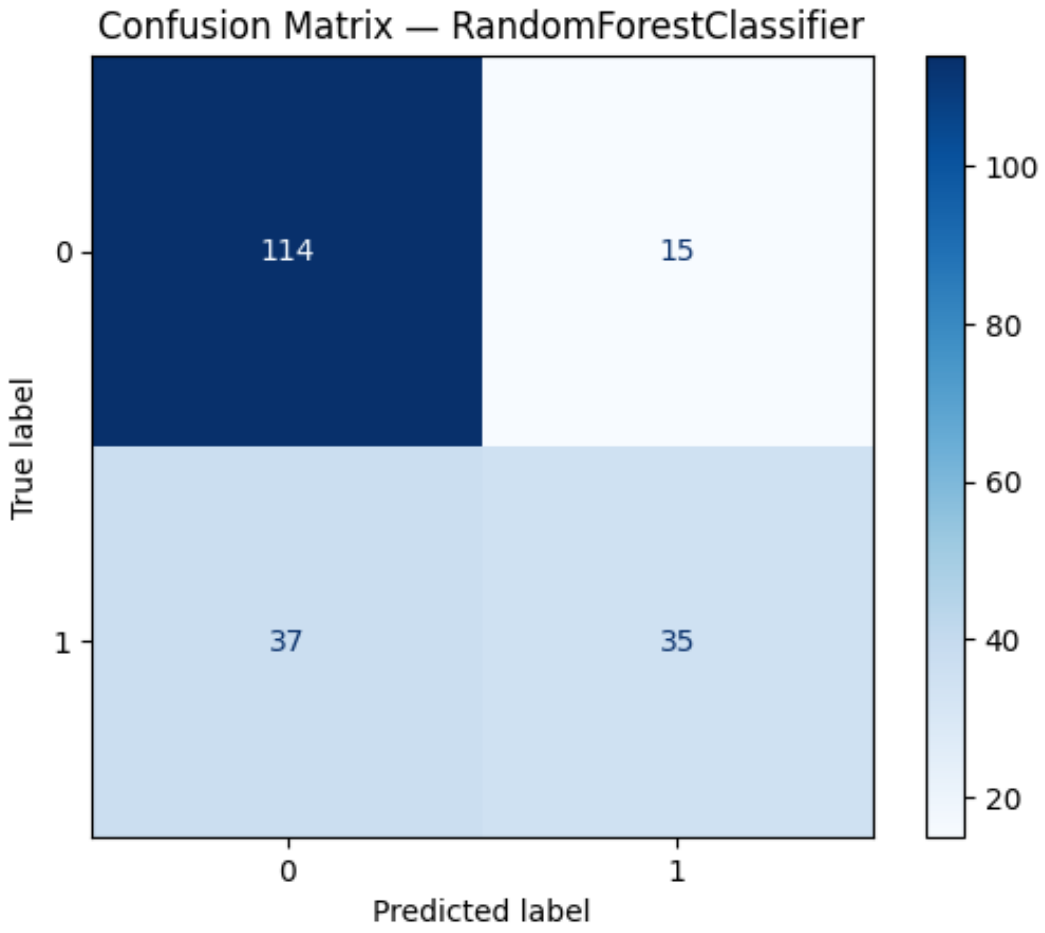
Матрица для IC50



Матрица для CC50



Матрица для SI



Матрица для SI (>8)

- **IC50 > медианы:**
  Модель неплохо различает классы, но иногда путает неактивные соединения с активными. FP заметно, FN — умеренно. Для скрининга лучше снизить FN.
- **CC50 > медианы:**
  Модель склонна ошибаться в сторону «токсичности» (много FP), но основные кандидаты находит. Такой уклон допустим, чтобы не пропустить реально опасные соединения.
- **SI > медианы:**
  Много FN — модель часто не видит селективные соединения. FP почти нет. Не хватает recall для положительного класса.
- **SI > 8:**
  Задача самая сложная — модель пропускает много действительно селективных (много FN), но ложных тревог почти нет.
- **Общий вывод:**
  Модели чаще не добирают важные объекты (много FN). Это безопасно для первичного отбора, но снижает шанс найти реально ценные кандидаты. Нужно пробовать балансировку классов или изменить метрику под задачи с критичным recall.

---

6. Анализ

Регрессия

- **IC50**: Наилучший результат показал Random Forest — на порядок выше по R^2 (0.47), чем любые линейные алгоритмы, что ожидаемо из-за выраженной нелинейности и потенциальных взаимодействий между дескрипторами.
- **CC50**: Неожиданно, лучшей оказалась модель Lasso ($R^2 = 0.32$), что указывает на сильную избыточность и коррелируемость признаков, которую Lasso эффективно отсекает.
- **SI**: Самая сложная задача — даже Random Forest дает скромный $R^2 = 0.26$, что косвенно говорит о высокой вариативности биологических данных по селективности, а также о шумности целевого показателя.
- **Общее наблюдение:**
  Значения R^2 по всем задачам не превышают 0.5, что типично для реальных биомедицинских данных и отражает сложность их структурного описания через классические дескрипторы. SI — самый “шумный” и слабопредсказуемый показатель, а наиболее “обучаемым” оказался IC50.

Классификация

- **IC50/CC50**: Во всех задачах лучшие результаты показал RandomForestClassifier — высокая ROC-AUC (0.79–0.83), точность около 70%. Это указывает на достаточную информативность дескрипторов для дифференциации соединений по активности/токсичности.
- **SI > median**: Лучше всего справился SVC (ядро rbf) — вероятно, за счет гибкой нелинейной границы в пространстве признаков. Тем не менее, точность и F1-метрика (0.68 и 0.62) заметно ниже, что подтверждает сложность задачи.
- **SI > 8**: RandomForest вновь лидирует (Accuracy 0.74, ROC-AUC 0.75), однако F1-score лишь 0.57, что указывает на значительный дисбаланс классов и сложность предсказания "high SI"-соединений.

Выводы по моделям и задачам

- **Ансамблевые методы (Random Forest)** — вне конкуренции для большинства задач.
- **Линейные методы** — полезны как baseline, а для CC50 даже Lasso стал лучшим (снижает шумные признаки).
- **SVM** — оказался актуален только для задачи классификации SI, где “граница” между классами очень неочевидна.
- **Низкое R^2 для SI** — показывает, что селективность хуже всего поддается прогнозу на основе стандартных дескрипторов (это биологически ожидаемо).
- **Дисбаланс классов** — в задачах “SI > 8” влияет на F1-метрику, требует учета при дальнейшем использовании моделей.
- **Применимость:**
  Построенные модели — только первый фильтр, без гарантии высокой точности в реальных лабораторных условиях.

---

7. Выводы и рекомендации

1. Для задач с высокой нелинейностью деревья и ансамбли эффективнее.
2. SI — метрика с низкой предсказуемостью, если не делать отдельную инженерную обработку.
3. Логарифмирование — обязательный шаг для подобных данных.
4. В реальной задаче важно увеличивать объём данных.
5. Внедрение в pipeline (в медико-химический или фармацевтический исследовательский процесс): только для первичного скрининга, не для финального отбора.