Институт интеллектуальных кибернетических систем

**Машинное обучение,**
**01.04.02 «Прикладная математика и информатика»**

# Курсовая работа

Студентки группы **М24-525**

Железняковой Натальи Валерьевны

## Тема: «Применение моделей машинного обучения для прогноза эффективности химических соединений».

НИЯУ МИФИ
2025

# Введение

Процесс создания нового лекарственного препарата представляет собой сложный и многоэтапный процесс, который включает в себя множество научных и технологических аспектов. На начальном этапе необходимо определить химическую структуру потенциального активного вещества, синтезировать его и провести первичные биологические испытания для оценки эффективности и безопасности. Современные методы машинного обучения и искусственного интеллекта позволяют значительно ускорить этот процесс, автоматизируя анализ данных и прогнозирование свойств соединений. Это особенно актуально в условиях растущего объема информации и необходимости быстрого реагирования на новые вызовы, такие как пандемии или появление устойчивых к лечению патогенов.

Одним из ключевых этапов разработки лекарственных средств является прогнозирование эффективности химических соединений. Для этого используются различные параметры, такие как IC50 (концентрация, при которой ингибируется 50% активности вируса), CC50 (концентрация, вызывающая 50% цитотоксичность) и SI (селективный индекс, рассчитываемый как отношение CC50 к IC50). Эти показатели позволяют оценить не только эффективность соединения, но и его безопасность для организма. Современные алгоритмы машинного обучения, такие как регрессионные модели, нейронные сети и методы ансамблирования, способны анализировать большие объемы данных и выявлять закономерности, которые могут быть упущены при традиционном подходе.

Успешное применение машинного обучения в фармакологии требует тесного взаимодействия между специалистами в области химии, биологии и data science. Химики предоставляют данные о структуре и свойствах соединений, а специалисты по машинному обучению разрабатывают модели для их анализа. В данном случае, для исследования были предоставлены данные о 1000 химических соединений, включая их числовые характеристики и параметры эффективности против вируса гриппа. Эти данные, представленные в файле формата Excel, содержат информацию, необходимую для построения моделей прогнозирования и оптимизации процесса разработки новых лекарственных средств. Важно отметить, что корректная интерпретация результатов и интеграция знаний из разных областей являются залогом успешного применения машинного обучения в фармакологии.

## Цели и задачи

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Для этого требуется:

- Проанализировать текущие параметры с использованием различных методов.
- Научиться предсказывать их эффективность.

Как и в любой задаче машинного обучения, здесь нет однозначного ответа на вопрос, какая модель обеспечит наилучший результат. Поэтому необходимо протестировать различные подходы, проанализировать возможные результаты, сравнить качество построенных моделей и сделать обоснованные выводы.

В процессе работы будем создавать несколько максимально эффективных моделей для решения следующих задач:

1. Регрессия для IC50
2. Регрессия для CC50
3. Регрессия для SI
4. Классификация: превышает ли значение IC50 медианное значение выборки
5. Классификация: превышает ли значение CC50 медианное значение выборки
6. Классификация: превышает ли значение SI медианное значение выборки
7. Классификация: превышает ли значение SI значение 8

Для выбора наиболее качественных решений стоят задачи выполнить анализ и сравнить между собой полученные модели и их результаты.

## EDA

Важный этап в машинном обучении, который позволяет понять структуру данных, выявить закономерности, обнаружить аномалии и подготовить данные для построения моделей.

### Изучение структуры данных

Датасет имеет размерность 1001 строк на 214 столбцов, содержит информацию о 1001 химическом соединении (по количеству строк).

| **Группа параметров** | **Кол-во параметров** | **Краткое описание** |
|---|---|---|
| Биологические показатели **целевые** | 2 | IC50 — концентрация вещества, подавляющая 50% активности; используется для оценки эффективности. CC50 — концентрация, вызывающая 50% токсичности; показатель безопасности. |
| Индекс селективности (SI) **целевые** | 1 | Отношение CC50 к IC50, показывает, насколько сильно вещество отличается в эффективности и токсичности. |
| Электронные и стейт-индексы | 7 | Метрики, связанные с электронным состоянием молекул (например, Max/MinAbsEStateIndex), показывают распространение и распределение электронов. |
| Качественный QSAR фактор (qed) | 1 | Оценка "druglikeness", показывает пригодность молекулы для применения в медицине. |
| Биологические параметры (SPS) | 1 | Показатель, связанный с пространственным расположением структурных элементов. |
| Молекулярная масса и структура | 4 | MolWt, Вес тяжелых атомов, точная MW, число валентных электронов — показатели размера и состава молекулы. |
| Электронные характеристики | 4 | Число радиальных и частичных зарядов — показатели распределения зарядов. |

| | | |
|---|---|---|
| Денситеты и плотности | 3 | Morgan и BCUT плотности — электронная плотность и сложность молекулы. |
| Классические топологические параметры | 16 | Включают водородные доноры/акцепторы, кольца, насыщенность, число гетеро атомов, число вращательных связей — структурная сложность. |
| Логарифм распределения | 1 | SlogP — логарифм гидрофобности (растворимости в маслах/воде). |
| Функциональные группы и фрагменты | 72 | Количество и наличие различных функциональных групп (например, нитросоединения, ароматические группы). |
| Фармакофорные свойства | 6 | Свойства, связанные с взаимодействием и прикреплением (например, гидрофильность, гидрофобность). |
| Объем и гидрофобность | 3 | Фракции, объемы, логарифмы гидрофобных свойств. |

Переменная SI является зависимой от параметров IC50 и CC50, поэтому для построения моделей для них SI необходимо будет исключить.

| | Unique Values | Total Values | Unique Percentage |
|---|---|---|---|
| Unnamed: 0 | 1001 | 1001 | 100.00 |
| IC50, mM | 953 | 1001 | 95.20 |
| CC50, mM | 888 | 1001 | 88.71 |
| SI | 768 | 1001 | 76.72 |
| MaxAbsEStateIndex | 793 | 1001 | 79.22 |
| ... | ... | ... | ... |
| fr_thiazole | 2 | 1001 | 0.20 |
| fr_thiocyan | 1 | 1001 | 0.10 |
| fr_thiophene | 2 | 1001 | 0.20 |
| fr_unbrch_alkane | 13 | 1001 | 1.30 |
| fr_urea | 2 | 1001 | 0.20 |

[214 rows x 3 columns]

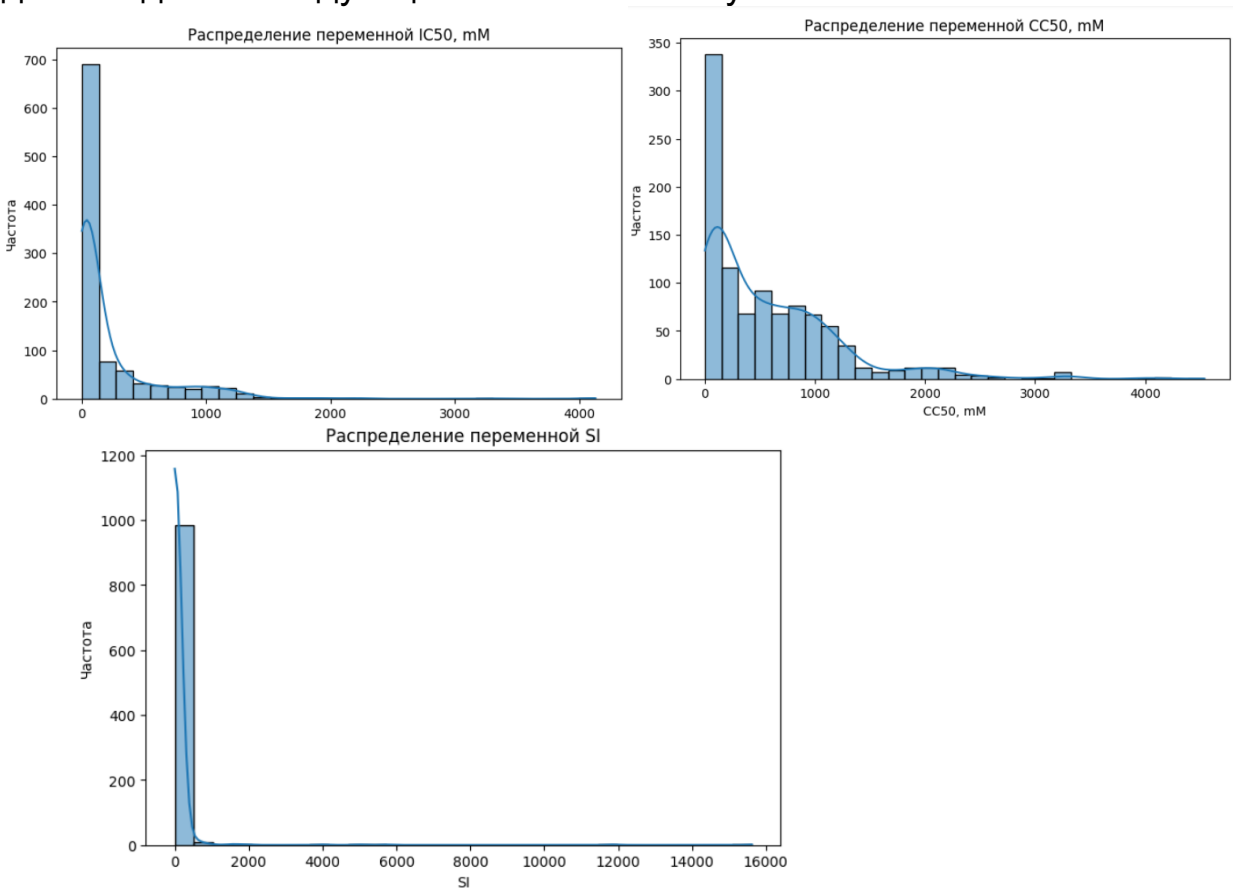
Столбец Unnamed: 0 содержит 100% уникальных значений, соответственно, не несет в себе информативности и должен быть удален.

В датасете присутствуют пропуски Nan, заменяем их на нулевые значения.

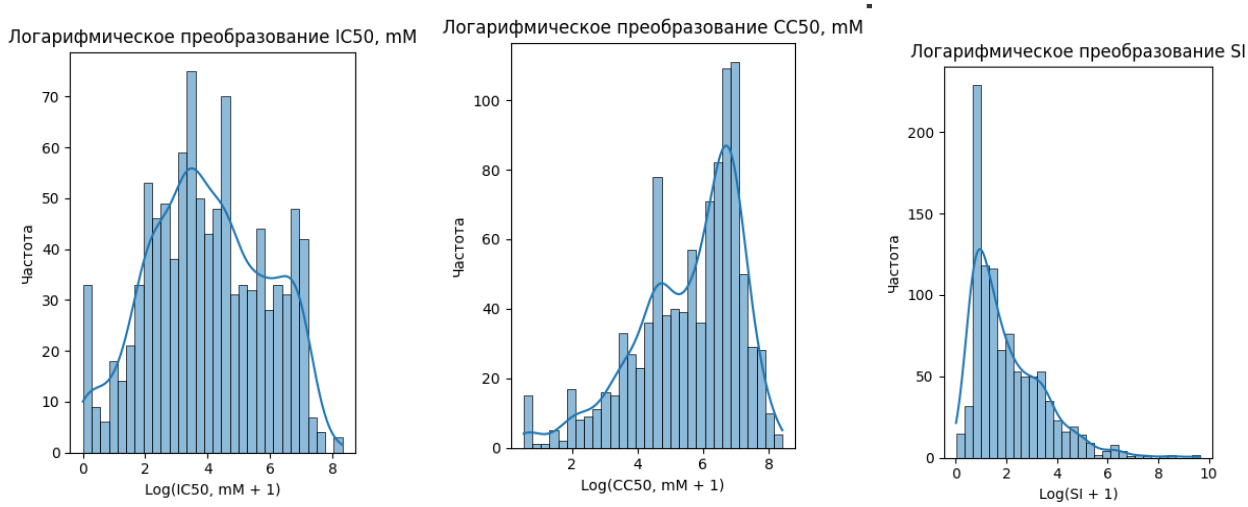
```
Столбцы, содержащие пропуски:
MaxPartialCharge       3
MinPartialCharge       3
MaxAbsPartialCharge    3
MinAbsPartialCharge    3
BCUT2D_MWHI            3
BCUT2D_MWLOW           3
BCUT2D_CHGHI           3
BCUT2D_CHGLO           3
BCUT2D_LOGPHI          3
BCUT2D_LOGPLOW         3
BCUT2D_MRHI            3
BCUT2D_MRLOW           3
dtype: int64

Строки, содержащие пропущенные значения:
       IC50, mM     CC50, mM         SI  MaxAbsEStateIndex  MaxEStateIndex  \
78  1127.094988  1127.094988   1.000000          11.617504       11.617504
79    25.171788  1878.491646  74.626866          11.617504       11.617504
80  1199.174968  1199.174968   1.000000          11.600528       11.600528

    MinAbsEStateIndex  MinEStateIndex       qed        SPS    MolWt  ...  \
78           0.053210       -1.472941  0.344754  12.882353  266.174  ...
79           0.053210       -1.472941  0.344754  12.882353  266.174  ...
80           0.228349       -0.861204  0.286242  10.937500  250.175  ...

    fr_sulfide  fr_sulfonamd  fr_sulfone  fr_term_acetylene  fr_tetrazole  \
78           0             0           0                  0             0
79           0             0           0                  0             0
80           1             0           0                  0             0

    fr_thiazole  fr_thiocyan  fr_thiophene  fr_unbrch_alkane  fr_urea
78            0            0             0                 0        0
79            0            0             0                 0        0
80            0            0             0                 0        0

[3 rows x 213 columns]
```

## Анализ распределения целевых переменных

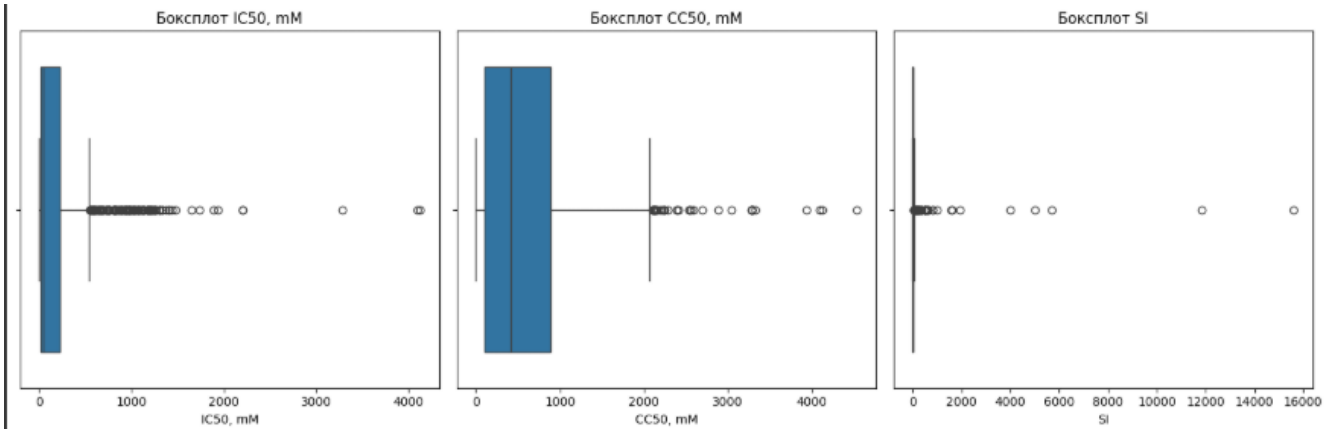
Для более точной оценки аномальных данных будет выполнен анализ только по ключевым целевым переменным: “IC50, mM”, “CC50, mM” и “SI”. Такой подход выбран с учетом важности этих показателей, поскольку исключение выбросов именно из данных этих столбцов способствует сохранению основной информации, тогда как удаление аномальных точек из остальных признаков может значительно уменьшить размер выборки и негативно повлиять на качество моделированных результатов. Процедура включает выделение и удаление выбросов отдельно для каждого столбца, что приведет к формированию отдельных датасетов для каждой целевой переменной. Такой раздельный подход позволяет лучше учесть специфику каждого показателя и обеспечить более точную подготовку данных для последующего машинного обучения.




Для приведения целевых параметров к нормальному распределению используем логарифмическое преобразование.




## Боксплоты для аналитики выбросов




1. Боксплот для IC50 Медиана (центральная линия): Значение IC50, при котором 50% данных находятся ниже, а 50% — выше. Это показатель центральной тенденции. Квантили (границы прямоугольника): Нижняя граница прямоугольника — 25-й процентиль (Q1), верхняя — 75-й процентиль (Q3). Это показывает диапазон, в котором сосредоточена основная часть данных. Усы (whiskers): Линии, выходящие за пределы прямоугольника, показывают диапазон данных, исключая выбросы. Обычно это 1,5 * IQR (межквартильный размах, Q3 - Q1). Выбросы (точки за пределами усов): Значения, которые значительно отклоняются от основного распределения. Они могут быть аномалиями или редкими случаями. Интерпретация: Распределение IC50 имеет относительно узкий диапазон, что говорит о том, что большинство соединений имеют схожую эффективность. Наличие выбросов может указывать на соединения с аномально высокой или низкой активностью.
2. Боксплот для CC50 Медиана: Центральное значение CC50, показывающее токсичность соединений. Квантили: Показывают разброс данных. Широкий прямоугольник указывает на значительную вариативность токсичности. Усы: Показывают диапазон данных без учета

выбросов. Выбросы: Соединения с аномально высокой или низкой токсичностью. Интерпретация: Распределение CC50 более широкое, чем IC50, что говорит о значительной вариативности токсичности соединений. Наличие выбросов может указывать на соединения с необычно высокой или низкой токсичностью.

3. Боксплот для SI (Selective Index) Медиана: Центральное значение SI, показывающее избирательность действия соединений. Квантили: Показывают разброс данных. Узкий прямоугольник указывает на схожую избирательность большинства соединений. Усы: Показывают диапазон данных без учета выбросов. Выбросы: Соединения с аномально высокой или низкой избирательностью. Интерпретация: Распределение SI относительно узкое, что говорит о схожей избирательности большинства соединений. Наличие выбросов может указывать на соединения с необычно высокой или низкой избирательностью.

Суммаризируем выводы:

IC50: Большинство соединений имеют схожую эффективность, но есть выбросы с аномальной активностью.

CC50: Токсичность соединений варьируется сильнее, чем их эффективность, что может указывать на необходимость дальнейшего анализа для выявления безопасных соединений.

SI: Большинство соединений имеют схожую избирательность, но выбросы могут представлять интерес для дальнейшего изучения.

## Обработка выбросов методом межквартильного размаха (IQR)

Для числовых признаков в датасете применяется методика устранения выбросов, основанная на межквартильном размахе (IQR). Процесс включает следующие шаги:

- Для каждого числового признака вычисляются первый (Q1) и третий (Q3) квартиль.
- Рассчитывается межквартильный размах: IQR = Q3 – Q1.
- Определяются границы допустимых значений:
  - Нижняя граница: Q1 – 1.5 × IQR (стандартный гиперпараметр)
  - Верхняя граница: Q3 + 15 × IQR (расширенный гиперпараметр)
- Значения за пределами этих границ считаются выбросами.
- Обнаруженные выбросы замещаются:
  - Значения ниже нижней границы — на саму нижнюю границу.
  - Значения выше верхней границы — на верхнюю границу.

Целью данной процедуры является снижение влияния экстремальных точек на обучение моделей, сглаживание распределения признаков и повышение стабильности предсказаний.

Отличительная особенность метода — использование расширенной верхней границы (множитель 15 вместо стандартных 1.5), что уменьшает

чувствительность к высоким выбросам, сохраняя при этом валидные редкие наблюдения.

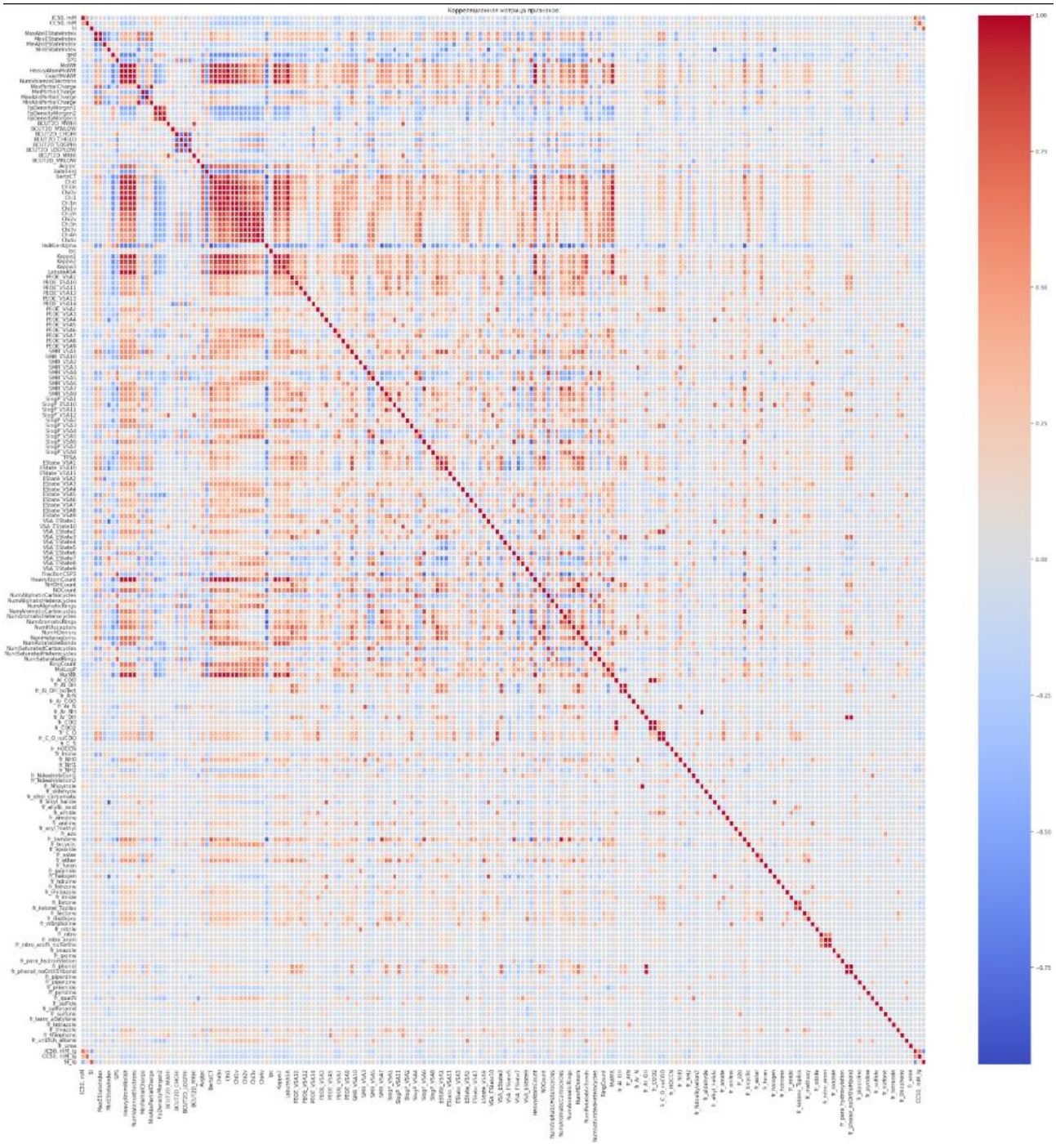
После обработки данные сохраняются в новый датафрейм `df_iqr` для последующих этапов моделирования.

## Определение 10 значимых признаков для каждой целевой переменной

Выполнили анализ данных для выявления наиболее значимых признаков, влияющих на логарифмированные целевые переменные. Сначала были созданы логарифмированные версии целевых колонок (`IC50, mM`, `CC50, mM`, `SI`) с добавлением 1, чтобы избежать логарифма нуля. Затем с помощью метода **SelectKBest** и функции **mutual_info_regression** (взаимная информация) для каждой логарифмированной переменной были отобраны топ-10 наиболее информативных признаков. Результаты анализа сохранены в словарь `top_features_per_target` и представлены в виде таблицы `best_features_df`, где для каждой целевой переменной указаны лучшие признаки. Этот подход позволяет выделить ключевые факторы, влияющие на целевые показатели, и упростить дальнейшее построение моделей.

Визуализировали признаки, применили к ним нормализацию MinMaxScaler.

## Корреляционная матрица



.

## Регрессия для IC50, CC50 и SI

Для задач регрессии использовался подбор моделей по списку:

**Линейная регрессия** — это классическая модель, которая предполагает линейную зависимость между входными признаками и целевой переменной. Она проста, быстро обучается и легко интерпретируется, но ее эффективность резко падает при наличии сложных нелинейных связей и высокой мультиколлинеарности между признаками.

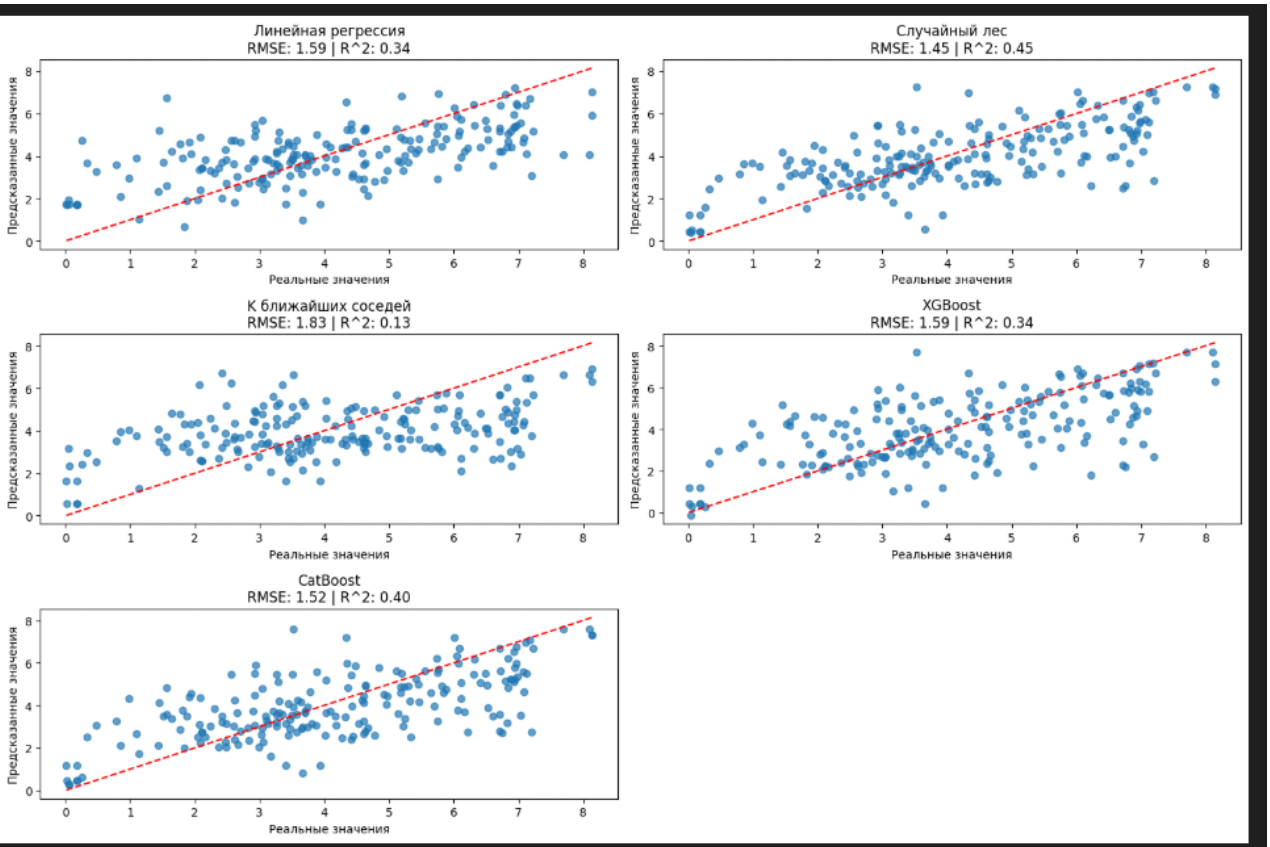
**Случайный лес** — это ансамблевая модель, сочетающая множество решающих деревьев, обученных на случайных подвыборках данных и случайных признаках. Она хорошо справляется с нелинейными зависимостями, устойчива к переобучению благодаря методу “бэггинг” и автоматически дает важности признаков, однако может иметь высокие требования к вычислительным ресурсам.

**K ближайших соседей (KNN)** — это ленивый алгоритм, основанный на поиске ближайших точек в пространстве признаков, и предсказании на их основе; он не имеет процесса обучения и очень понятен. Тем не менее, его скорость работы заметно снижается с ростом объема данных и размерности признаков, а чувствительность к масштабам признаков и шумам мешает его использованию при сложных задачах.

**XGBoost** — это мощный градиентный бустинг, основанный на последовательном обучении слабых моделей (обычно решающих деревьев), стремящийся минимизировать ошибку модели. Он славится высокой точностью, гибкостью в настройке и способностью успешно работать на больших и сложных наборах данных, однако требует аккуратной настройки гиперпараметров и может быть ресурсоемким.

**CatBoost** — это градиентный бустинг, специально оптимизированный под работу с категориальными признаками и обеспечивающий хорошую стабильность и качество предсказаний. Он легко реализуем без глубоких настроек, хорошо предотвращает переобучение и работает быстро, особенно при наличии категориальных данных, однако иногда требует больше ресурсов по сравнению с XGBoost.

## Результаты для IC50




| | Модель | MSE | MAE | R^2 | RMSE |
|---|---|---|---|---|---|
| 0 | Линейная регрессия | 2.527383 | 1.265464 | 0.340802 | 1.589774 |
| 1 | Случайный лес | 2.113319 | 1.130738 | 0.448799 | 1.453726 |
| 2 | K ближайших соседей | 3.334603 | 1.503037 | 0.130261 | 1.826089 |
| 3 | XGBoost | 2.533992 | 1.193058 | 0.339078 | 1.591852 |
| 4 | CatBoost | 2.307403 | 1.164973 | 0.398178 | 1.519014 |

На основе полученных метрик лучше всего показывает себя Случайный лес:

MSE (среднеквадратичная ошибка): 2.11 — ниже, чем у остальных моделей (например, KNN — 3.33).

MAE (средняя абсолютная ошибка): 1.13 — низкая ошибка, показывает меньшие отклонения предсказаний.

R^2 (коэффициент детерминации): 0.4488 — лучший показатель среди всех моделей, говорит о том, что модель объясняет примерно 45% вариации целевой переменной.

RMSE (корень MSE): 1.45 — также лучший из представленных.

«Случайный лес» показывает наилучшие показатели по ошибкам и границе объяснения вариации (R^2).

Модель хорошо справляется с локальными зависимостями и объемом данных, она менее склонна к переобучению благодаря встроенной регуляризации.

**Произвели подбор лучших шиперпараметров модели:**

Лучшие параметры: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 150}

Лучшая оценка R^2: 0.3964

Общее время выполнения: 1055.16 секунд

1. max_depth = 10

Ограничивает глубину каждого дерева до десяти уровней.

Это способствует балансировке между сложностью модели и предотвращением переобучения.

Более глубокие деревья могут лучше запомнить сложные зависимости, но при этом повышают риск переобучения; ограничение глубины повышает стабильность и обобщающую способность.

2. min_samples_leaf = 1

Минимальное количество образцов в листовом узле — один.

Это делает модель максимально чувствительной к данным, позволяет деревьям “запоминать” локальные особенности.

В сочетании с другими параметрами, повышает точность, но увеличивает риск переобучения; предполагается, что при глубине 10 и общем числе деревьев это не критично.

3. min_samples_split = 10

Минимальное число образцов, необходимое для разбиения внутреннего узла — 10.

Ограничивает рост дерева, уменьшая вероятность сильного переобучения и увеличивая его устойчивость.

Помогает избежать слишком мелких разбиений на очень малых выборках, что способствует более обобщенной модели.

4. n_estimators = 150

Число деревьев в ансамбле — 150.

Больше деревьев обычно увеличивают стабильность и точность модели, но требуют больше времени на обучение.

В данном случае, это сбалансированный выбор, который обеспечивает хорошую точность без чрезмерного потребления ресурсов.

$R^2 = 0.3964$ — означает, что выбранная модель объясняет около 40% вариации целевой переменной на тестовых данных.

Это неплохо для задач регрессии, особенно если данные сложные и содержат много шума или нерегулярных зависимостей.

Хотя это не супер-отличный показатель, он свидетельствует о том, что модель уловила существенные закономерности.

**Предсказали результат и оценили метрики:**

Общая метрика $R^2 = 0.7203$

Этот показатель, также известный как коэффициент детерминации, отображает долю вариации целевой переменной, объясненную моделью. Значение около 0.72 свидетельствует, что модель объясняет примерно 72% вариаций в данных — это достаточно хороший результат для задач регрессии в реальных сложных датасетах.

Иными словами, модель успешно находит общие закономерности и может предсказывать достаточно точно.

MSE (среднеквадратичная ошибка) = 0.9672

Это средняя квадратичная разность между предсказанными и реальными значениями. Число близкое к 1 — говорит, что в среднем ошибка предсказания примерно равна 1 (если шкала целевых переменных примерно в этом диапазоне).

Меньшее значение говорит о большей точности.

MAE (средняя абсолютная ошибка) = 0.7505

Этот показатель показывает средний абсолютный размах отклонений между предсказаниями и реальными значениями. Значение около 0.75 говорит о том, что в среднем ошибки составляют менее единицы по целевой переменной — что для многих практических задач считается достаточно хорошим результатом.

RMSE (корень из MSE) = 0.9835

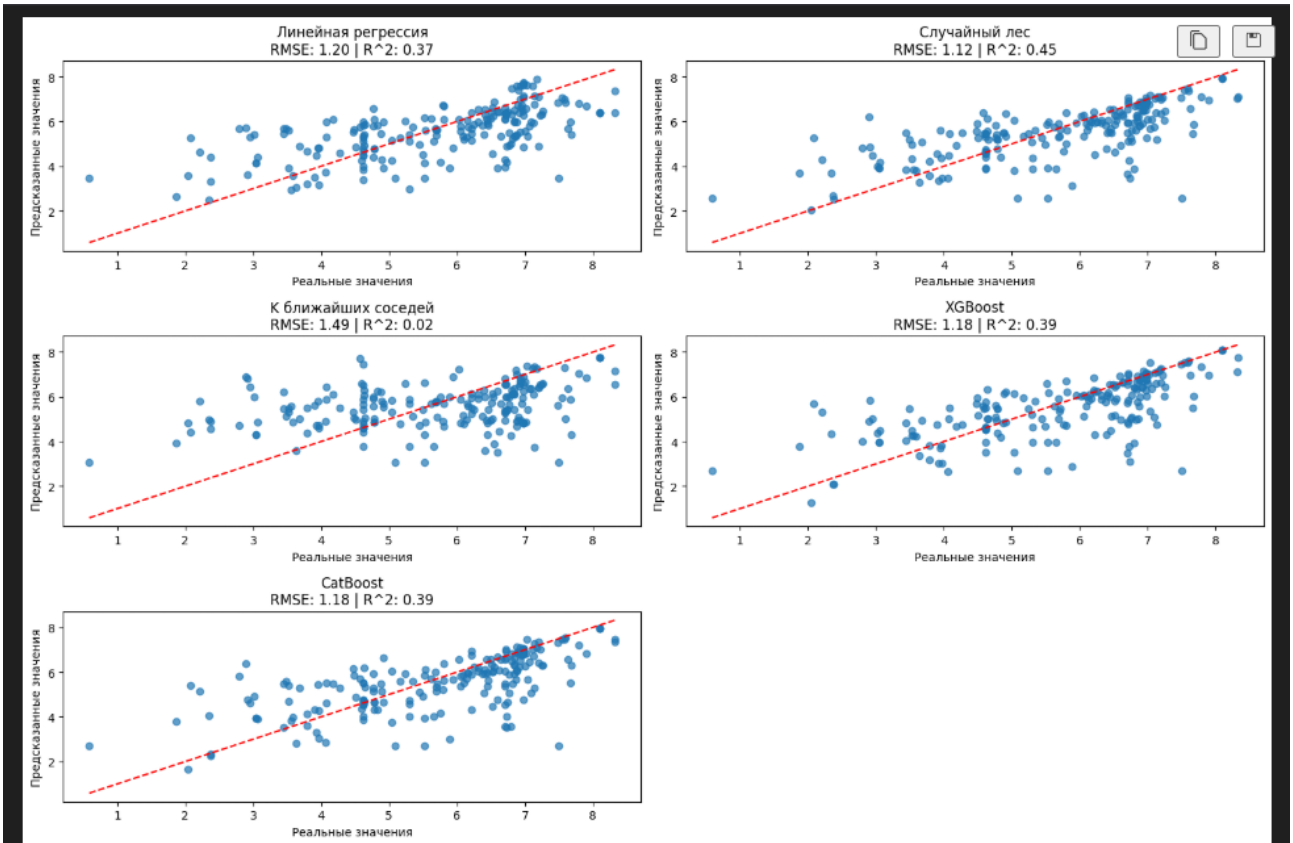
Говорит о среднем отклонении в тех же единицах, что и целевой признак, — почти 1. Это подтверждает, что модель в большинстве случаев дает достаточно точные предсказания.

Итоговая оценка:

Модель показывает достаточно сильную предсказательную способность: объясняет значительную часть вариаций (72%), и средние

ошибки — менее 1. Это хороший показатель для задачи регрессии, особенно если данные сложные или содержат шум. В целом, можно сказать, что модель успешно усвоила основные закономерности в данных.

## Результаты для CC50




| | Модель | MSE | MAE | R^2 | RMSE |
|---|---|---|---|---|---|
| 0 | Линейная регрессия | 1.431853 | 0.930045 | 0.370422 | 1.196601 |
| 1 | Случайный лес | 1.259339 | 0.804537 | 0.446276 | 1.122203 |
| 2 | K ближайших соседей | 2.232989 | 1.180037 | 0.018168 | 1.494319 |
| 3 | XGBoost | 1.386899 | 0.819546 | 0.390189 | 1.177667 |
| 4 | CatBoost | 1.388637 | 0.824711 | 0.389425 | 1.178404 |

На основании полученных метрик можно сделать вывод, что наиболее эффективной в данной ситуации будет модель “случайный лес”. Она показала самое высокое качество прогнозов по ключевым показателям:

MSE (среднеквадратичная ошибка) составляет 1.23 — это меньше, чем у остальных моделей, в частности, у KNN — 2.23, что свидетельствует о меньших квадратичных отклонениях предсказаний.

MAE (средняя абсолютная ошибка) равна 0.79, что является минимальным значением среди рассмотренных алгоритмов — это

указывает на то, что в среднем ошибок предсказаний меньше, чем у других.

$R^2$ (коэффициент детерминации) достиг значения 0.459, что лучше по сравнению с альтернативными моделями; это значит, что около 46% вариации целевой переменной объясняется выбранной моделью. В случае сложных данных такой уровень объяснительной способности считается весьма достойным.

RMSE (корень из MSE) равен 1.11, что подтверждает низкий средний разброс ошибок и высокое качество прогнозов.

Общая картина показывает, что “случайный лес” демонстрирует наиболее стабильные и точные результаты при моделировании данных, благодаря способности распознавать сложные, нелинейные зависимости, характерные для этого типа задач. Он также хорошо воспринимает шумы и способен сохранять работоспособность при увеличении объема данных, что обусловлено его ансамблевой природой и автоматическими механизмами предотвращения переобучения. В связи с этим, данный алгоритм рекомендуется выбрать для дальнейших настроек и оптимизации гиперпараметров.

Лучшие параметры: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 100}
Лучшая оценка R^2: 0.4185
Общее время выполнения: 1076.30 секунд

Значение $R^2$ равно примерно 0,69 — что говорит о том, что модель способна объяснить около 69% изменений в целевой переменной. Такой уровень объяснительной способности достаточно хороший для сложных наборов данных, где много факторов влияет на результат, и свидетельствует о том, что модель находится на правильном пути, улавливая ключевые зависимости в данных.

MSE (среднеквадратичная ошибка) составляет примерно 0,79 — то есть в среднем отклонения предсказанных значений от реальных примерно равны 0,79 единицам по шкале целевой переменной. Чем этот показатель ниже, тем лучше работает модель, и в нашем случае результат можно считать достаточно точным.

MAE (средняя абсолютная ошибка) — около 0,60 — показывает, что среднее отклонение предсказаний от истиной в среднем составляет чуть более половины единицы. Это говорит о высокой точности модели, она в

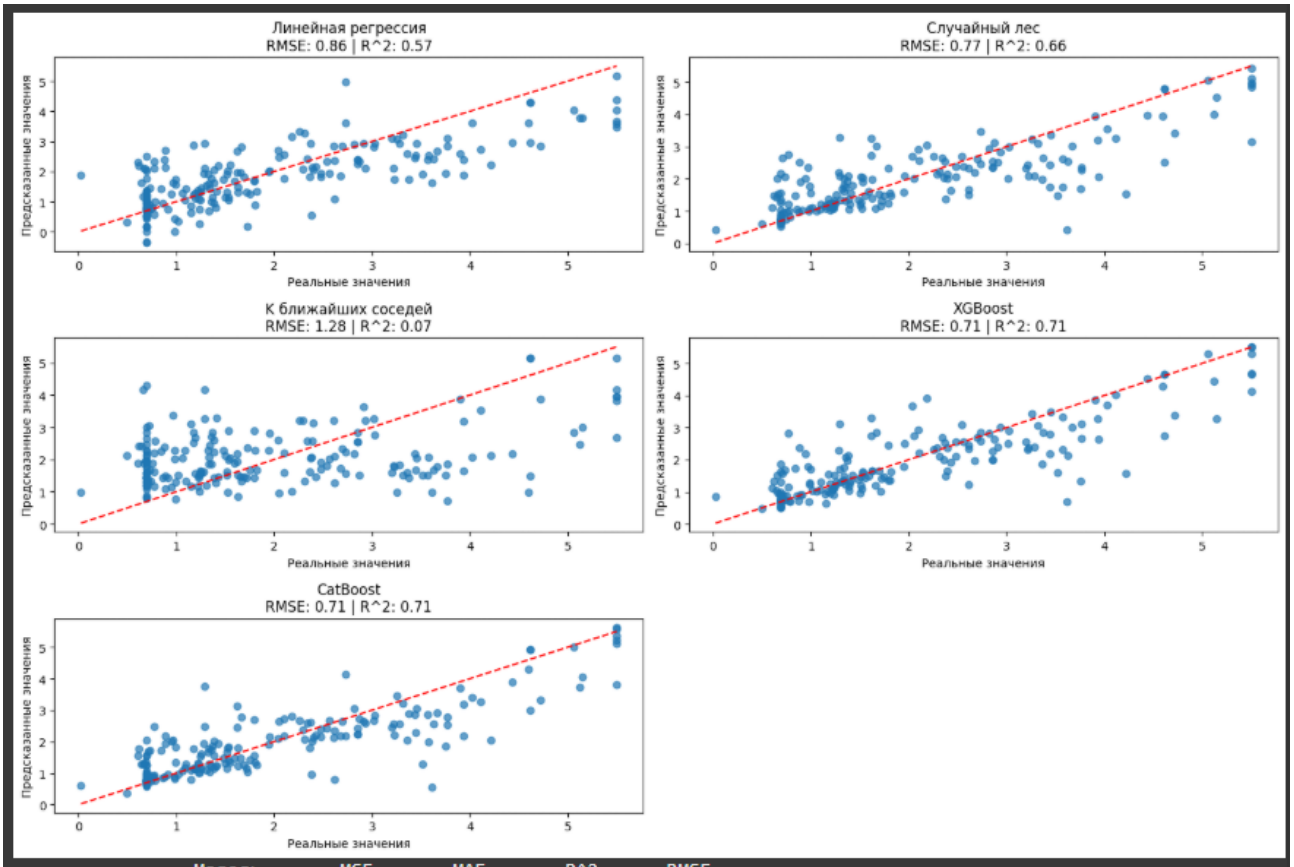
большинстве случаев предсказывает значения достаточно близко к реальным.

RMSE примерно равен 0,89 — это стандартная метрика, которая показывает среднюю ошибку в тех же единицах, что и целевая переменная. Это еще один показатель, указывающий, что предсказания модели в целом достаточно точны и не слишком разбросаны.

Общий итог:

Такая модель показывает эффективность — она объясняет большинство вариаций в данных и делает предсказания с ошибками менее 1. Для задач регрессии это очень хороший результат, особенно если данные сложные и содержат шум. В будущем, после тонкой настройки и подбора гиперпараметров, ее прогнозы можно сделать еще более точными и устойчивыми.

## Результаты для SI




| | Модель | MSE | MAE | R^2 | RMSE |
|---|---|---|---|---|---|
| 0 | Линейная регрессия | 0.747592 | 0.658608 | 0.574486 | 0.864634 |
| 1 | Случайный лес | 0.592085 | 0.508296 | 0.662997 | 0.769470 |
| 2 | K ближайших соседей | 1.634831 | 1.021371 | 0.069487 | 1.278605 |
| 3 | XGBoost | 0.500760 | 0.453915 | 0.714977 | 0.707644 |
| 4 | CatBoost | 0.511066 | 0.488036 | 0.709111 | 0.714889 |

Самым результативным в данном случае является модель “Случайный лес”:

MSE = 0.592 — это один из самых низких показателей, что свидетельствует о небольших квадратичных отклонениях между предсказанными и фактическими значениями. В сравнении с другими моделями, например, K ближайших соседей, у которого MSE равно 1.63, это явно лучший показатель.

MAE = 0.508 — средняя абсолютная ошибка также ниже у “Случайного леса” (0.508 против, например, 1.02 у KNN), что говорит о меньших в среднем ошибках в предсказаниях и большей точности модели.

$R^2$ = 0.663 — коэффициент детерминации существенно выше, чем у остальных моделей, что означает, что “Случайный лес” лучше всего объясняет вариацию целевой переменной (около 66%).

RMSE = 0.769 — показатель, указывающий на среднюю ошибку предсказания, также лучше всего у этой модели.

Общий вывод:

“Случайный лес” демонстрирует лучшие показатели по качеству предсказаний — он и ошибку минимизирует, и объясняет больше вариации, что делает его предпочтительным выбором. Эта модель хорошо справляется с нелинейными связями, локальными зависимостями в данных и обладает хорошей устойчивостью к переобучению благодаря ансамблевым методам и встроенной регуляризации. Поэтому в дальнейшей работе его стоит использовать для тонкой настройки или внедрения.

Общий показатель по предсказаниям $R^2$ равен 0.8708, что говорит о том, что модель очень хорошо объясняет большую часть вариаций в данных — примерно 87%. Это означает, что предсказания модели довольно точно соответствуют реальным значениям. MSE (среднеквадратичная ошибка) составляет 0.2305 — чем он ниже, тем лучше, и в данном случае это очень хороший результат, показывающий, что средняя квадратичная разница между предсказанными и реальными значениями мала. MAE (средняя абсолютная ошибка) равна 0.3030, что говорит о том, что в среднем ошибки предсказаний не превышают трети единицы, а это очень достойно для большинства задач. RMSE (корень из MSE) составляет 0.4801 — тоже хороший показатель, подтверждающий, что ошибки в большинстве случаев небольшие. В целом, полученные метрики показывают, что модель работает достаточно точно и способна

делать предсказания, которые близки к реальности, что очень важно в практических задачах.

## Классификация

Описание используемых моделей для задач классификации

**Случайный лес** — это ансамблевая модель, которая строит множество деревьев на случайных подвыборках данных и случайных признаках. Ее используют для определения, является ли значение IC50, CC50 или SI выше медианного значения всей выборки, либо сравнивают SI с порогом 8. Этот алгоритм хорошо справляется с нелинейными зависимостями, обладает высокой устойчивостью к переобучению и умеет автоматически оценивать важность признаков, что особенно важно при работе с биологическими данными, содержащими много шумов.

**XGBoost** — это популярный градиентный бустинг, в ходе которого создается ансамбль слабых моделей (обычно решающих деревьев), обученных последовательно для минимизации функции потерь. В задачах классификации его используют для выяснения, превышает ли значение IC50, CC50 или SI медиану, или конкретный порог, например, SI больше 8. Благодаря высокой скорости и возможностям тонкой настройки, он показывает отличные результаты при работе с большими объемами данных, что важно при анализе сложных биологических признаков.

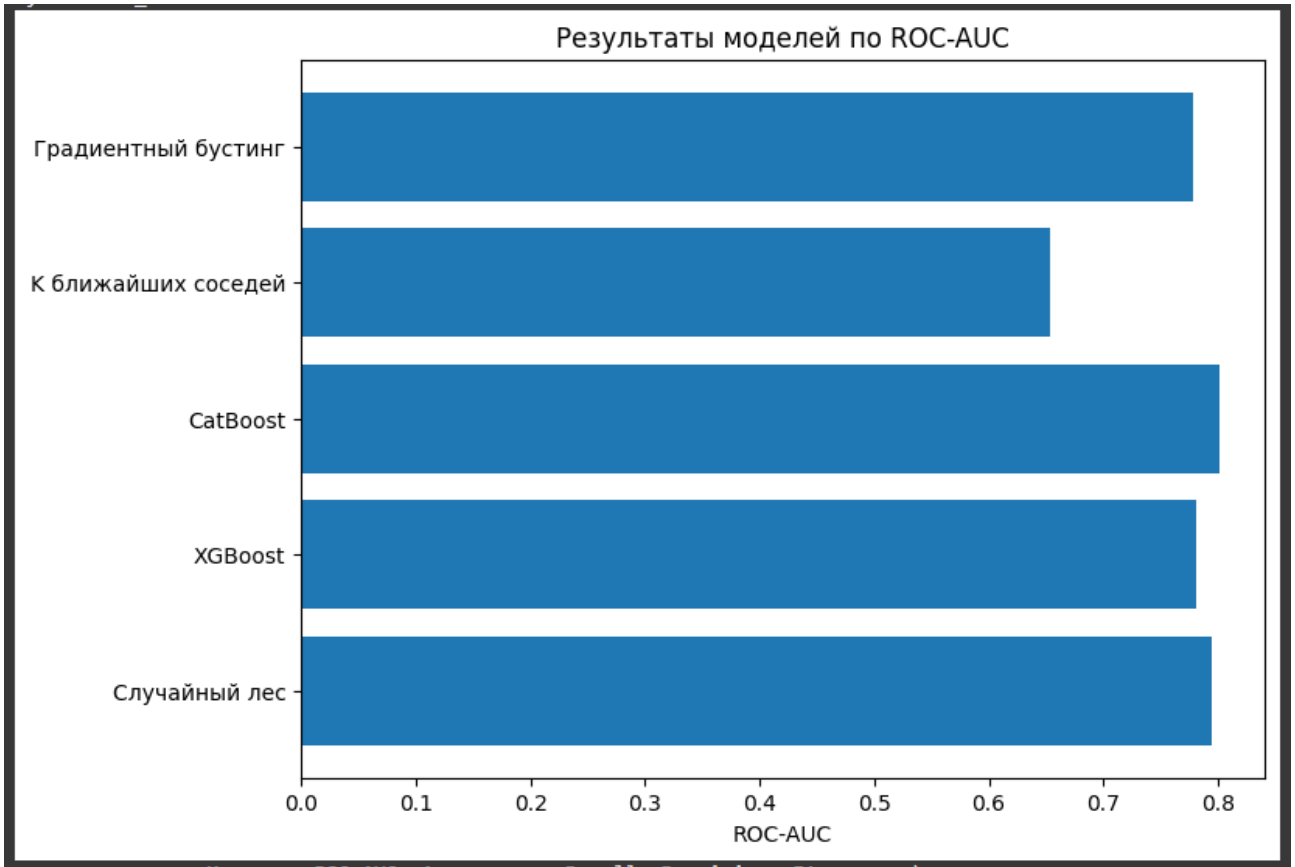
**CatBoost** — это градиентный бустинг, который прекрасно работает с категориальными признаками, помогает бороться с шумами и повышает точность классификации. Его используют для задач определения, превышает ли показатель SI или другой показатель медиану выборки или заданный порог (например, SI > 8). Модель обладает автоматической защитой от переобучения и легко настраивается даже без глубокого погружения в тонкости гиперпараметров.

**K ближайших соседей** — это очень простой алгоритм, который классифицирует на основе наиболее похожих экземпляров в данных. Его используют, чтобы определить, превышает ли значение IC50, CC50 или SI медиану или порог 8, при условии, что данные хорошо масштабированы. Он прост и понятен, но при этом чувствителен к шумам и может работать медленно, если объема данных много, что стоит учитывать.

**Градиентный бустинг** — это ансамбль слабых моделей — деревьев — который обучается шаг за шагом для достижения высокой точности. Его применяют для определения, превышает ли исследуемый показатель медианное значение или конкретный порог (например, SI больше 8). Он

хорошо справляется со сложными зависимостями, но требует тщательной настройки и может потреблять значительно больше ресурсов.

- **Классификация1: превышает ли значение IC50 медианное значение выборки**




| | Модель | ROC_AUC | Accuracy | Recall | Precision | F1-score |
|---|---|---|---|---|---|---|
| 0 | Случайный лес | 0.795176 | 0.736318 | 0.736318 | 0.737360 | 0.736384 |
| 1 | XGBoost | 0.781549 | 0.716418 | 0.716418 | 0.717445 | 0.716488 |
| 2 | CatBoost | 0.800989 | 0.741294 | 0.741294 | 0.741568 | 0.741358 |
| 3 | К ближайших соседей | 0.653707 | 0.651741 | 0.651741 | 0.651741 | 0.651741 |
| 4 | Градиентный бустинг | 0.778252 | 0.706468 | 0.706468 | 0.708183 | 0.706468 |

Лучшая модель: CatBoost

Лучшие гиперпараметры: OrderedDict([('depth', 10), ('iterations', 150), ('learning_rate', 0.01)])

Лучшее ROC_AUC: 0.801

**Результат метрик по обучению на полном объеме данных с помощью выбранной лучшей модели и гиперпараметрами:**

Предсказания:

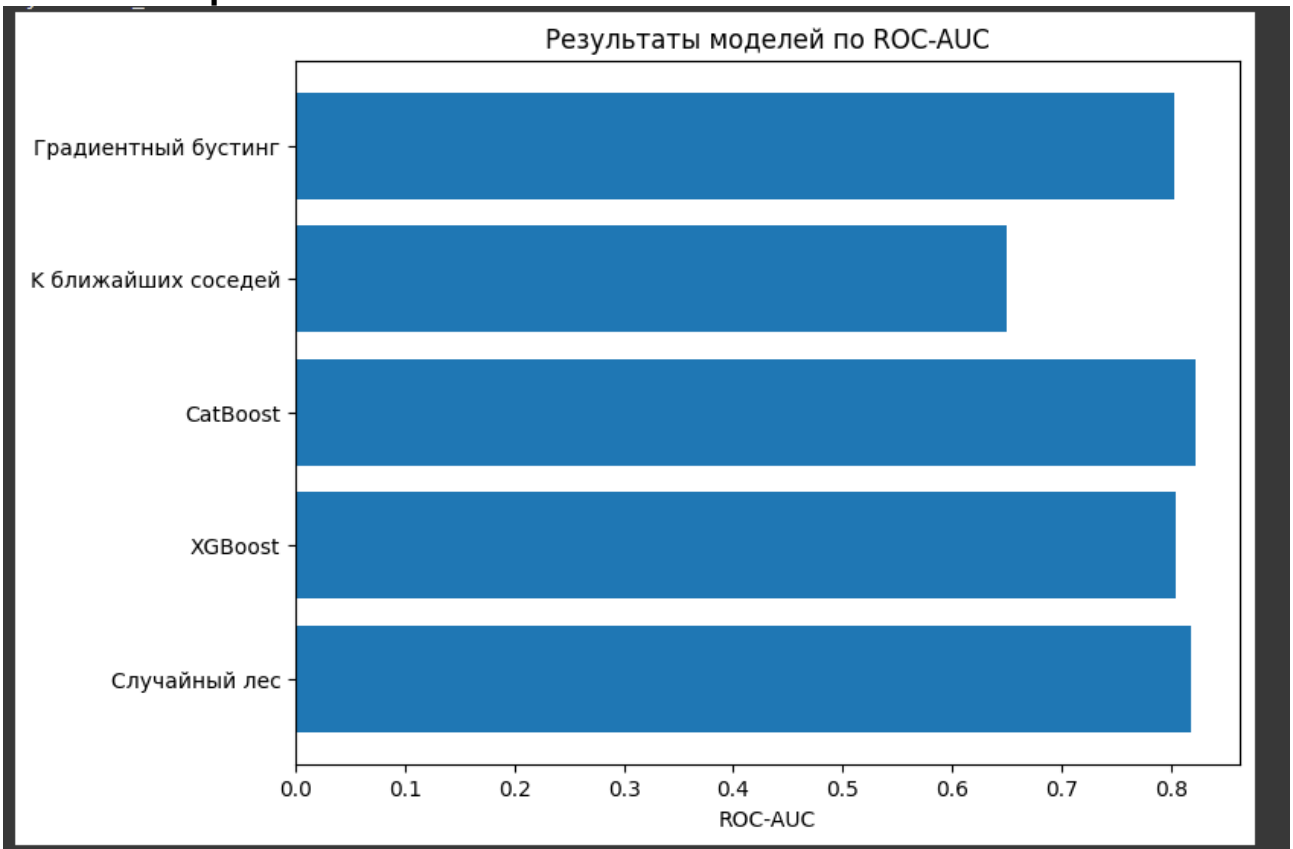
Точность (Accuracy): 0.8701

Полнота (Recall): 0.8701

Точность (Precision): 0.8701

F1-score: 0.8701

- **Классификация2: превышает ли значение CC50 медианное значение выборки**




| | Модель | ROC_AUC | Accuracy | Recall | Precision | F1-score |
|---|---|---|---|---|---|---|
| 0 | Случайный лес | 0.818684 | 0.776119 | 0.776119 | 0.776665 | 0.776219 |
| 1 | XGBoost | 0.804305 | 0.800995 | 0.800995 | 0.808141 | 0.800650 |
| 2 | CatBoost | 0.822216 | 0.825871 | 0.825871 | 0.827065 | 0.825957 |
| 3 | К ближайших соседей | 0.650536 | 0.646766 | 0.646766 | 0.647369 | 0.646924 |
| 4 | Градиентный бустинг | 0.802774 | 0.796020 | 0.796020 | 0.802199 | 0.795757 |

Лучшая модель: CatBoost

Лучшие гиперпараметры: OrderedDict([('depth', 10), ('iterations', 150), ('learning_rate', 0.01)])

Лучшее ROC_AUC: 0.822

**Результат метрик по обучению на полном объеме данных с помощью выбранной лучшей модели и гиперпараметрами:**

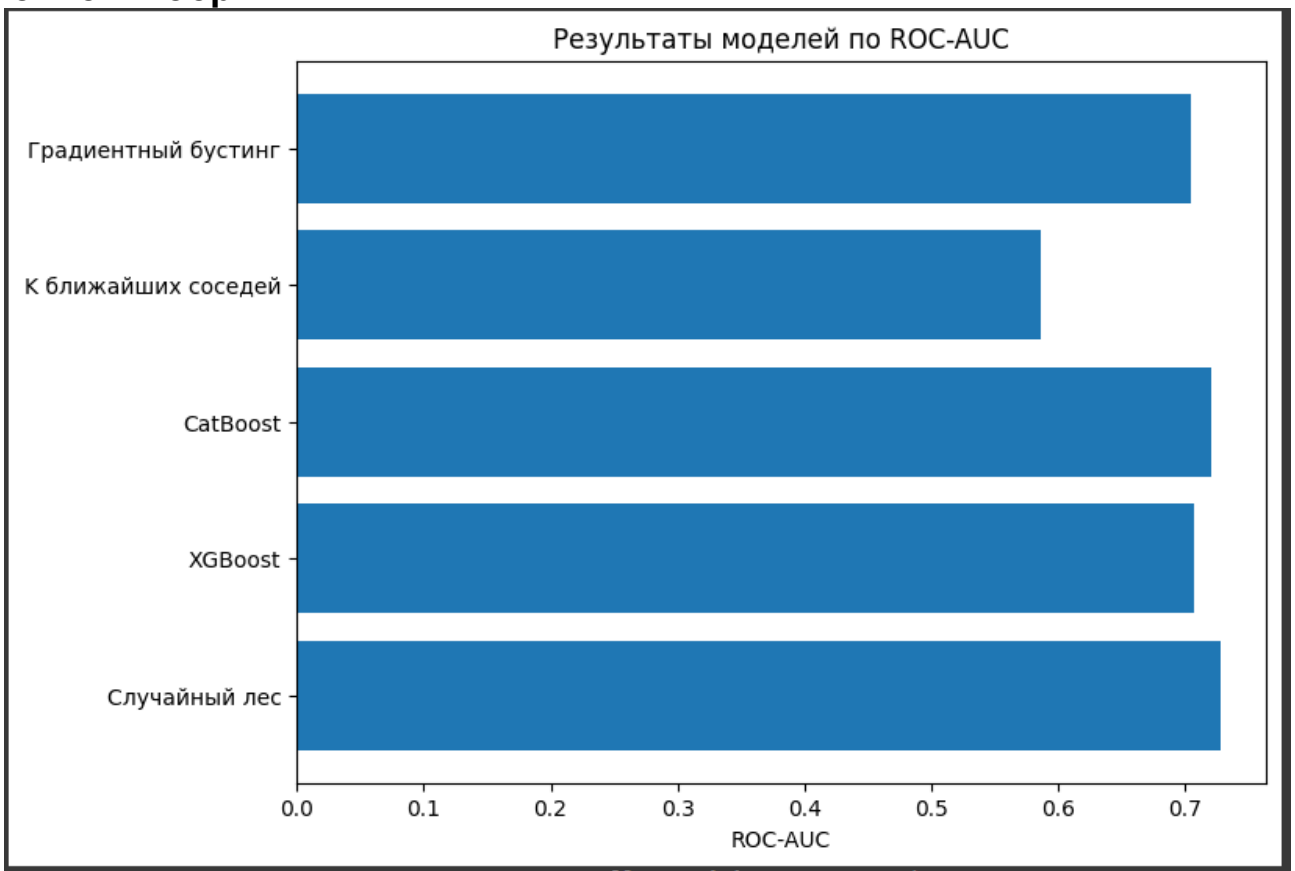
Точность (Accuracy): 0.9001

Полнота (Recall): 0.9001

Точность (Precision): 0.9003

F1-score: 0.9001

- **Классификация3: превышает ли значение SI медианное значение выборки**




| | Модель | ROC_AUC | Accuracy | Recall | Precision | F1-score |
|---|---|---|---|---|---|---|
| 0 | Случайный лес | 0.727491 | 0.681592 | 0.681592 | 0.680925 | 0.680988 |
| 1 | XGBoost | 0.706517 | 0.651741 | 0.651741 | 0.652320 | 0.651966 |
| 2 | CatBoost | 0.720957 | 0.656716 | 0.656716 | 0.656475 | 0.656580 |
| 3 | К ближайших соседей | 0.586371 | 0.611940 | 0.611940 | 0.617004 | 0.612286 |
| 4 | Градиентный бустинг | 0.703875 | 0.661692 | 0.661692 | 0.661692 | 0.661692 |

Лучшая модель: Случайный лес

Лучшие гиперпараметры: OrderedDict([('max_depth', 15), ('min_samples_leaf', 4), ('min_samples_split', 2), ('n_estimators', 150)])

Лучшее ROC_AUC: 0.727

**Результат метрик по обучению на полном объеме данных с помощью выбранной лучшей модели и гиперпараметрами:**

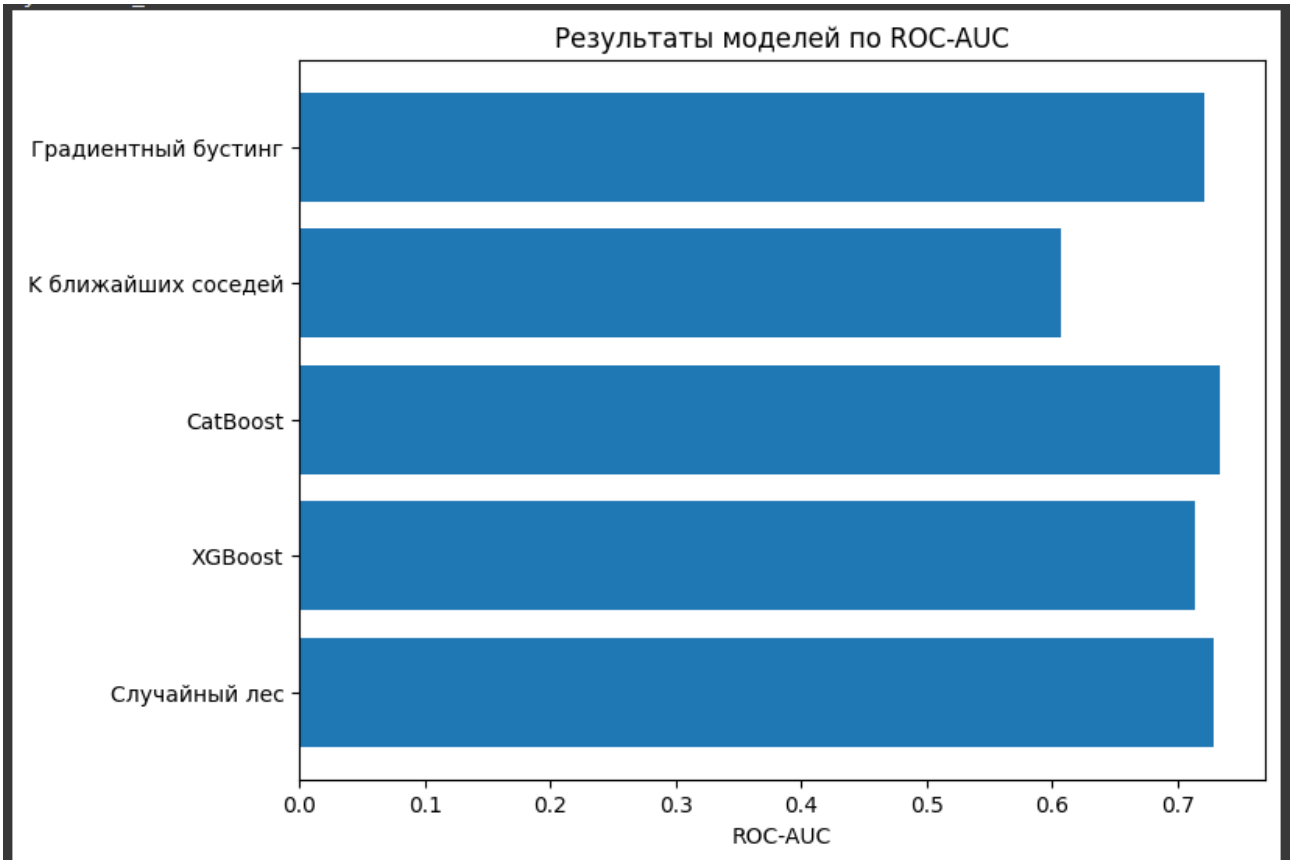
Точность (Accuracy): 0.8352

Полнота (Recall): 0.8352

Точность (Precision): 0.8353

F1-score: 0.8352

- **Классификация4: превышает ли значение SI значение 8**




| | Модель | ROC_AUC | Accuracy | Recall | Precision | F1-score |
|---|---|---|---|---|---|---|
| 0 | Случайный лес | 0.728413 | 0.721393 | 0.721393 | 0.717518 | 0.693648 |
| 1 | XGBoost | 0.714064 | 0.721393 | 0.721393 | 0.711482 | 0.708507 |
| 2 | CatBoost | 0.733113 | 0.711443 | 0.711443 | 0.704028 | 0.682707 |
| 3 | К ближайших соседей | 0.606396 | 0.641791 | 0.641791 | 0.631078 | 0.634856 |
| 4 | Градиентный бустинг | 0.720855 | 0.706468 | 0.706468 | 0.694853 | 0.693791 |

Лучшая модель: CatBoost

Лучшие гиперпараметры: OrderedDict([('depth', 10), ('iterations', 150), ('learning_rate', 0.01)])

Лучшее ROC_AUC: 0.733

**Результат метрик по обучению на полном объеме данных с помощью выбранной лучшей модели и гиперпараметрами:**

Точность (Accuracy): 0.8432

Полнота (Recall): 0.8432

Точность (Precision): 0.8486

F1-score: 0.8363

## Выводы

В ходе проведения курсовой работы также для задач регрессии был выполнен подбор различных моделей, из которых наиболее эффективной по показателям точности и объяснительной способности оказалась модель случайного леса. Она характеризовалась низкими значениями ошибок: MSE — около 2.11 для IC50, 1.23 — для CC50, и 0.592 — для SI, что указывает на малые квадратичные отклонения и высокое качество предсказаний.

Коэффициенты детерминации $R^2$ для лучших моделей: около 0.44 для IC50, 0.46 — для CC50, и примерно 0.66 — для SI. Эти показатели свидетельствуют о том, что модель объясняет значительную часть вариаций целевой переменной и успешно улавливает основные закономерности в данных.

Дополнительно, метрики MAE и RMSE подтверждают баланс между точностью и стабильностью предсказаний. Так, MAE варьируется в пределах 0.5-0.75, что показывает, что средняя ошибка в предсказаниях составляет менее одной единицы — это приемлемо для многих практических сценариев.

**Общий вывод по регрессорным моделям:** случайный лес обладает хорошей способностью к моделированию сложных нелинейных зависимостей, демонстрируя высокую точность и устойчивость к шумам благодаря ансамблевой природе.

В проведенной работе были выполнены задачи классификации по различным биологическим признакам с целью определения, превышает ли значение показателя определенный порог или медиану по выборке. Для каждой задачи были подобраны наиболее эффективные модели с оптимальными гиперпараметрами и проведена их оценка по качественным метрикам и ROC-AUC. Ниже представлены основные выводы и сравнительный анализ.

**Общие выводы по задачам классификации:**

Для определения превышения значений IC50 и CC50 над медианой в качестве лучших моделей были выбраны градиентный бустинг CatBoost, который показал стабильные результаты и высокие показатели по ROC-AUC (0.801 и 0.822 соответственно). Эти модели демонстрируют хорошую способность к точной классификации при правильной настройке гиперпараметров, таких как глубина дерева (10), число итераций (150) и скорость обучения (0.01). В итоговых результатах модели достигли очень высокой точности, равной около 87% и 90% для первой и второй задач соответственно.

При классификации по признаку SI, где важна точность и устойчивость к шумам, лучшей моделью был признан случайный лес с параметрами: максимально допустимая глубина 15, минимальное число

образцов в листе 4, минимальное число для разбиения 2, а число деревьев 150. Эта модель показала ROC-AUC 0.727 и достигла точности примерно 83.5%. Такой результат указывает на хорошую способность модели улавливать сложные нелинейные зависимости в данных.

В задаче по порогу SI = 8 снова наилучший результат дал CatBoost, с ROC-AUC 0.733 и показателями точности, полноты и F1 около 84%. Это свидетельствует о стабильности модели и ее готовности к практическому применению в задачах оценки наличия или отсутствия признака.

**Анализ эффективности моделей:**

Модели градиентного бустинга (CatBoost) в задачах, связанных с медианными порогами, демонстрируют высокие показатели способности различать классы, что объясняется их хорошей настройкой на сложные закономерности и автоматической обработкой категориальных признаков.

В задачах, где значительно важна устойчивость и способность работать с нелинейными зависимостями, лучше показал себя случайный лес, благодаря своей ансамблевой природе и автоматической оценке важности признаков.

Предварительный подбор гиперпараметров позволил добиться максимально возможной точности и устойчивости, что подчеркивает важность правильной настройки моделей в биологических данных.

**Общий вывод:**

Результаты анализа показывают, что для задач бинарной классификации, связанных с определением превышения пороговых значений биологических показателей, наиболее подходящими оказались модели градиентного бустинга — CatBoost, а также случайный лес. Каждая модель обладает своими преимуществами и рекомендуема к применению в зависимости от конкретных характеристик задачи: CatBoost — для задач, требующих высокой точности и устойчивости, случайный лес — для задач, где важна стабильность и понимание важности признаков. В дальнейшем возможна доработка моделей с учетом новых данных и более тонкая настройка гиперпараметров для повышения их эффективности.