# ВВЕДЕНИЕ

Разработка новых лекарственных препаратов — это сложный и затратный процесс, требующий значительных временных и финансовых вложений. Он включает в себя множество этапов, от синтеза и лабораторных испытаний до масштабных клинических исследований. Особенно сложным и ресурсоемким является этап доклинических испытаний, где оценивается потенциальная эффективность и безопасность химических соединений на клеточных моделях.

В рамках данного исследования мы сосредоточимся на применении методов машинного обучения для ускорения и оптимизации процесса разработки противовирусных препаратов. Цель работы – создание предиктивных моделей, способных оценивать активность химических соединений против вируса гриппа на основе их химических и физико-химических характеристик. Такой подход позволяет выявлять перспективные кандидаты на ранних стадиях, тем самым снижая затраты и сокращая сроки разработки.

В последние годы методы машинного обучения демонстрируют свою эффективность в выявлении взаимосвязей между химической структурой и биологической активностью соединений. Это открывает новые возможности для раннего скрининга и отбора наиболее перспективных молекул, что значительно повышает эффективность процесса разработки лекарств.

Для решения поставленной задачи мы будем использовать набор данных, содержащий информацию о 1001 химическом соединении и 214 характеристиках, включая 211 признаков (описаний) и 3 целевых параметра:

* **IC50 (полумаксимальная ингибирующая концентрация):** Концентрация, необходимая для подавления 50% активности вируса.
* **CC50 (полумаксимальная цитотоксическая концентрация):** Доза, при которой наблюдается 50% токсичности на клеточном уровне.
* **SI (индекс селективности):** Отношение CC50 к IC50, позволяющее оценить безопасность соединения.

Исходные данные представлены в формате Excel, обеспечивая удобство для последующего анализа, предобработки и визуализации результатов. В данной работе будет выполнена предобработка данных, исследованы различные методы машинного обучения, а также проведена оценка производительности моделей для прогнозирования ключевых фармакологических показателей.

# 1. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ (Exploratoty Data Analysis)

Для эффективного построения моделей необходимо провести углубленный анализ структуры и качества данных, чтобы выявить потенциальные проблемы, такие как пропуски и выбросы, и понять взаимосвязи между ключевыми переменными.

* 1. **Общие сведения о датасете:**

После стандартного импорта необходимых библиотек для работы с данными и их анализа, загружен файл формата xlsx и далее создана обширная функция infoOut() для комплексного анализа DataFrame, включая пропуски, дубликаты, уникальные значения, а также неявные дубликаты. Результат работы вышеуказанной функции показал что большинство колонок не имеют пропущенных значений, кроме следующих 12 столбцов:

* MaxPartialCharge;
* MinPartialCharge;
* MaxAbsPartialCharge;
* MinAbsPartialCharge;
* BCUT2D\_MWHI;
* BCUT2D\_MWLOW;
* BCUT2D\_CHGHI;
* BCUT2D\_CHGLO;
* BCUT2D\_LOGPHI;
* BCUT2D\_LOGPLOW;
* BCUT2D\_MRHI;
* BCUT2D\_MRLOW,

## что является положительным аспектом в процентом соотношении, где количество столбцов с пропусками составляет всего 5% от общего числа и незначительных количество пропущенных значений равных 3-м в самих столбцах.

Все колонки представлены в числовом формате (float64 и int64), что позволяет проводить количественный анализ и применять математические операции.

В наборе данных не выявлено дубликатов, что свидетельствует о целостности и качестве данных.

Количество уникальных значений варьируется от 1 до 1001, что указывает на разнообразие в данных. Это разнообразие может способствовать более качественному обучению моделей, хотя важные колонки с малым числом уникальных значений могут потребовать дополнительного внимания и в последующем исключены.

Всего в данном наборе данных 1001 строка, что обеспечивает достаточный объём для анализа и построения модели. Это подходит для применения многих методов машинного обучения, так как количество записей соответствует минимальным требованиям.

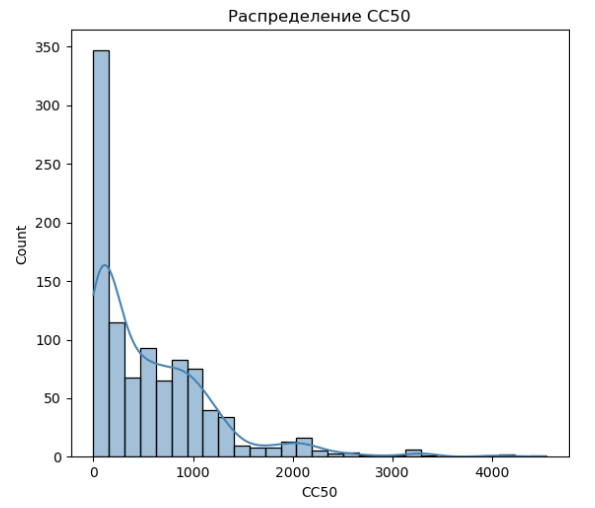
## 1.2 Распределение целевых переменных

Углубленный анализ целевых переменных CC50, IC50 и SI стал важным аспектом этой работы, поскольку они являются ключевыми в решаемой задаче в контексте связи с биологической активностью соединений.

### 1.2.1 CC50 (полумаксимальная цитотоксическая концентрация)

**Результат:**

* Общее число наблюдений: 1001.
* Среднее значение: около 589.11, стандартное отклонение: ≈ 642.87.
* Диапазон значений: от 0.70 до 4538.98, что также свидетельствует о наличии выбросов и разнообразии токсичности соединений.
* Распределение с длинным хвостом вправо (асимметрия), с концентрацией значений в нижних квартилях и значительными выбросами в верхней части диапазона.
* Квартили: 25% имеют CC50 ниже примерно 99.99, а 75% — ниже 894.09. Это важно учитывать при нормализации данных и построении моделей, особенно для оценки токсичности.



### **Вывод:** Распределение демонстрирует длинный правый хвост и высокую дисперсию значений, что свидетельствует о наличии значительных выбросов и разнообразии уровней токсичности соединений.

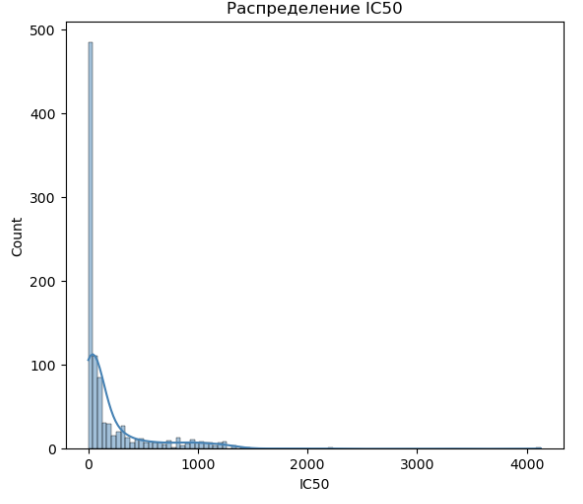
### Значения варьируются от 0.70 до 4538.98, с средним около 589.11 и стандартным отклонением примерно 642.87.

### Для повышения стабильности моделей необходимо выполнить логарифмическую трансформацию данных и учитывать потенциальное влияние экстремальных значений.

### 1.2.2 IC50 (полумаксимальная ингибирующая концентрация)

**Результат:**

* Общее число наблюдений: 1001.
* Среднее значение: ≈ 222.81, стандартное отклонение: ≈ 402.17, что свидетельствует о высокой изменчивости данных.
* Диапазон значений: от 0.0035 до 4128.53, указывая на наличие как очень активных соединений (низкие IC50), так и значительных выбросов с экстремальными значениями.
* Распределение скошено вправо (положительная асимметрия), что подтверждается длинным хвостом в правой части графика; большинство наблюдений сосредоточено в диапазоне низких значений, однако присутствуют редкие, очень большие значения.
* Квартили: 25% наблюдений имеют IC50 ниже примерно 12.52, а 75% — ниже 224.98. Это подчеркивает необходимость логарифмической трансформации для стабилизации распределения перед моделированием.



**Вывод:** Данное распределение также обладает выраженной асимметрией с длинным правым хвостом, что указывает на наличие как очень активных соединений, так и экстремальных выбросов.

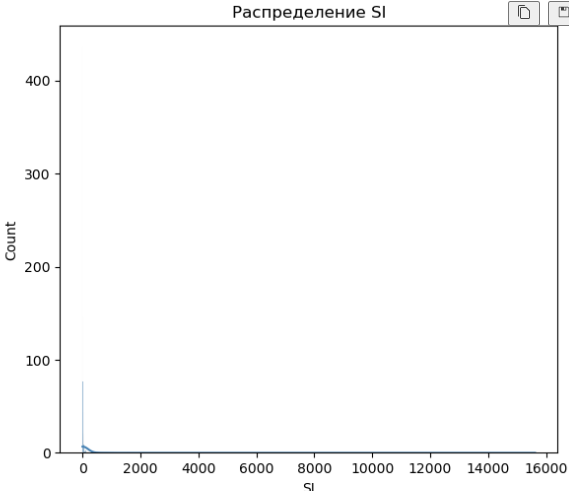
Диапазон значений: от 0.0035 до 4128.53, со средним около 222.81 и стандартным отклонением примерно 402.17.

Для корректной работы моделей необходимо также применить логарифмическую трансформацию, чтобы уменьшить скошенность и стабилизировать вариации.

### 1.2.3 SI (индекс селективности)

**Результат:**

* Общее число наблюдений: 1001.
* Среднее значение: ≈ 72.51, стандартное отклонение: ≈ 684.48.
* Диапазон значений: от 0.0115 до 15620.60, что подчеркивает очень высокую дисперсию и наличие экстремальных значений.
* Распределение с сильной правой асимметрией (длинный хвост вправо), многие значения сосредоточены в диапазоне от малых до умеренных чисел, однако присутствуют единичные очень большие показатели, что может искажать модель.
* Квартили: 25% наблюдений имеют SI менее 1.43, а 75% — менее 16.57. Значения выше 1000 и даже 15000 требуют особого внимания, поскольку могут быть результатом выбросов или аномалий.

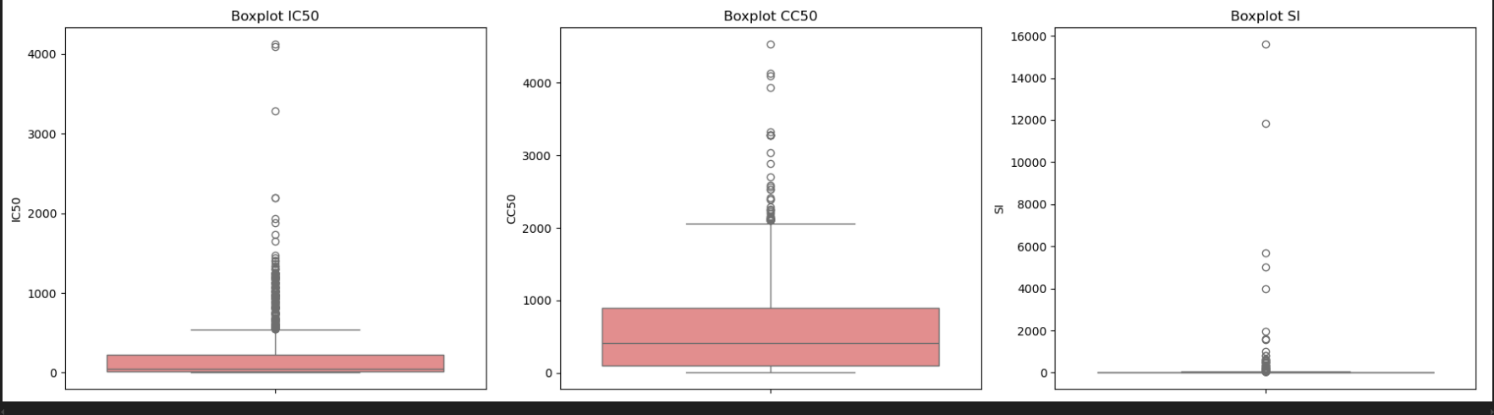


**Вывод:** Распределение характеризуется особенно выраженной асимметрией и наличием экстремальных значений, что делает его наиболее сложным для моделирования.

Значения варьируются от 0.0115 до 15620.60, со средним около 72.51 и очень высоким стандартным отклонением — примерно 684.48.

Перед обучением моделей необходимо провести трансформацию и тщательно обработать выбросы, чтобы избежать искажения результатов.

## 1.3 Визуализация с помощью диаграмм размаха (Boxplot)



Для каждой из трёх целевых переменных были построены диаграммы размаха, позволяющие оценить распределение данных и выявить особенности:

* **Медиана** — центральное значение, делящее данные пополам.
* **Межквартильный диапазон (IQR)** — длина коробки, отражающая разброс центральных 50% данных.
* **Нижние и верхние усы** — границы "нормальных" значений, за пределами которых расположены выбросы.
* **Выбросы** — точки вне усов, указывающие на аномальные наблюдения.

**Ключевые наблюдения:**

* Для **IC50** большинство точек расположены в низком диапазоне, с короткой межквартильной зоной и небольшим числом выбросов, которые сосредоточены в верхней части графика.
* Для **CC50** характерен широкий диапазон данных с более длинными усами и наличием нескольких выбросов, особенно в верхней части графика, что свидетельствует о наличии экстремальных значений.
* В случае **SI** наблюдается огромное количество выбросов, которые расположены в очень высоких диапазонах, что говорит о значительной асимметрии и экстремальных значениях.

**Вывод:** Переменная **SI** демонстрирует выраженную асимметрию и наличие множества экстремальных значений, что требует особого внимания при подготовке и моделировании данных. Возможно, потребуется отсечь выбросы или применить логарифмическое преобразование для стабилизации распределения. В то время как для **IC50** и **CC50** выбросов меньше, их наличие всё равно следует учитывать при построении моделей.

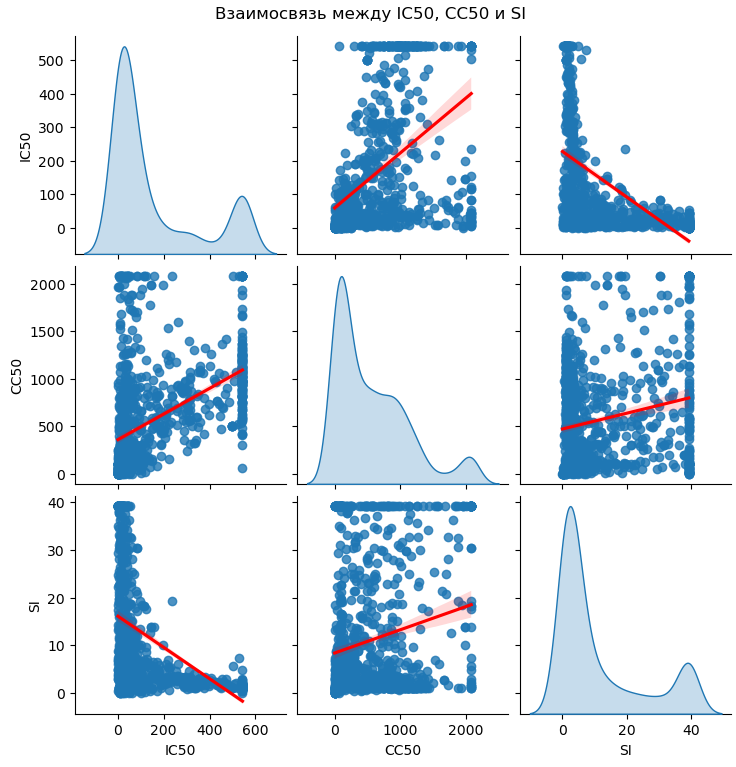
## 1.4 Корреляционный анализ

Для оценки степени взаимосвязи между переменными была построена корреляционная матрица с использованием коэффициента Пирсона. На диаграмме видно, что:

* **IC50 и CC50** демонстрируют умеренную положительную корреляцию, примерно 0,52, что подтверждает их связь, поскольку оба показателя основаны на концентрациях и отражают схожие свойства.
* **SI** показывает слабую корреляцию как с IC50, так и с CC50, что обусловлено тем, что SI является производной переменной, рассчитываемой на основе других показателей, и, следовательно, не полностью зависит от них.

Кроме того, в матрице выделены признаки, которые демонстрируют высокую корреляцию с целевыми переменными, что делает их перспективными кандидатами для дальнейшего моделирования. В частности, такие признаки, как **MaxAbsEstateIndex**, **ExactMolWt** и **HallKierAlpha**, имеют выраженную связь с целевыми переменными и могут быть использованы в качестве ключевых факторов при построении предиктивных моделей.

## 1.5 Анализ графиков распределений



Данные графики предоставили визуальное представление взаимосвязей между переменными IC50, CC50 и SI. Анализ рассеяний показал, что для переменных IC50 и CC50 присутствуют тенденции, которые могут быть эффективно смоделированы с помощью простых линейных методов, учитывая их относительно явные корреляционные связи. В тоже время, для переменной SI наблюдается более сложная структура распределения, что указывает на необходимость применения более продвинутых моделей для точного предсказания.

Такой комплексный анализ помогает понять структуру данных и выбрать наиболее подходящие методы моделирования для повышения точности прогнозов.

## 1.6 Промежуточные выводы

Проведенный анализ подтвердил высокое качество исходных данных, поскольку большинство колонок не содержат пропущенных значений, а дублирование данных отсутствует, что свидетельствует о целостности набора данных.

При более глубоком изучении распределений целевых переменных (IC50, CC50 и SI) выявлены существенные особенности, которые требуют внимания для повышения эффективности моделирования.

Распределения показали ярко выраженную скошенность и наличие значительных выбросов, особенно в переменной SI, что может негативно сказаться на точности и стабильности построенных моделей.

Анализ корреляционной матрицы выявил умеренную связь между IC50 и CC50, что подтверждает их схожую природу, а слабая корреляция SI с этими переменными указывает на необходимость разработки более сложных моделей, способных учитывать нелинейные взаимосвязи.

Обработка выбросов позволит заменить экстремальные значения на граничные, что в свою очередь будет способствовать стабилизации данных и снижению риска искажения результатов при обучении.

В целом, полученные результаты подчеркивают необходимость проведения дополнительных этапов предобработки данных - трансформации и обработки выбросов - для повышения надежности и точности моделей.

# 2. ПОСТРОЕНИЕ МОДЕЛЕЙ РЕГРЕССИИ ДЛЯ ПРОГНОЗА IC50

## 2.1 Постановка задачи

Ключевой задачей на этом этапе является предсказание IC50, которая определяет концентрацию, необходимую для 50% ингибирования активности вируса, что является жизненно важным показателем для оценки фармацевтической эффективности.

## 2.2 Подготовка данных

В качестве признаков используются 211 переменных, охватывающих ряд молекулярных характеристик, такие как:

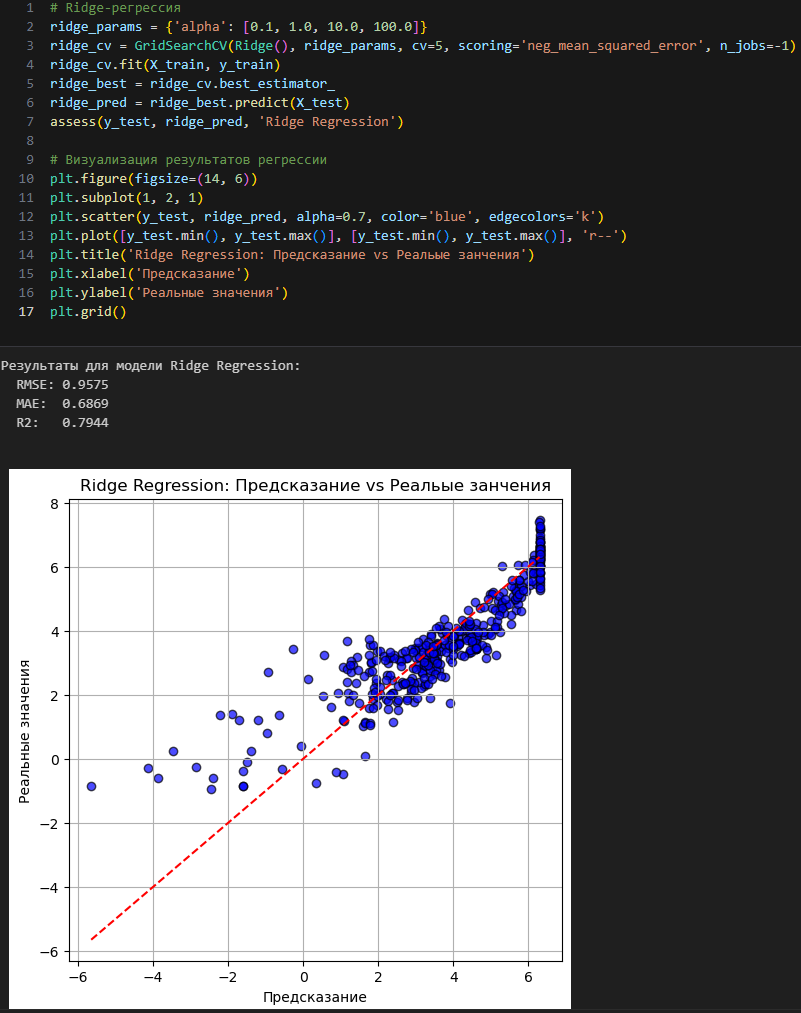
- Индексы атомного состава и структурные параметры (число атомов, наличие конкретных функциональных групп);

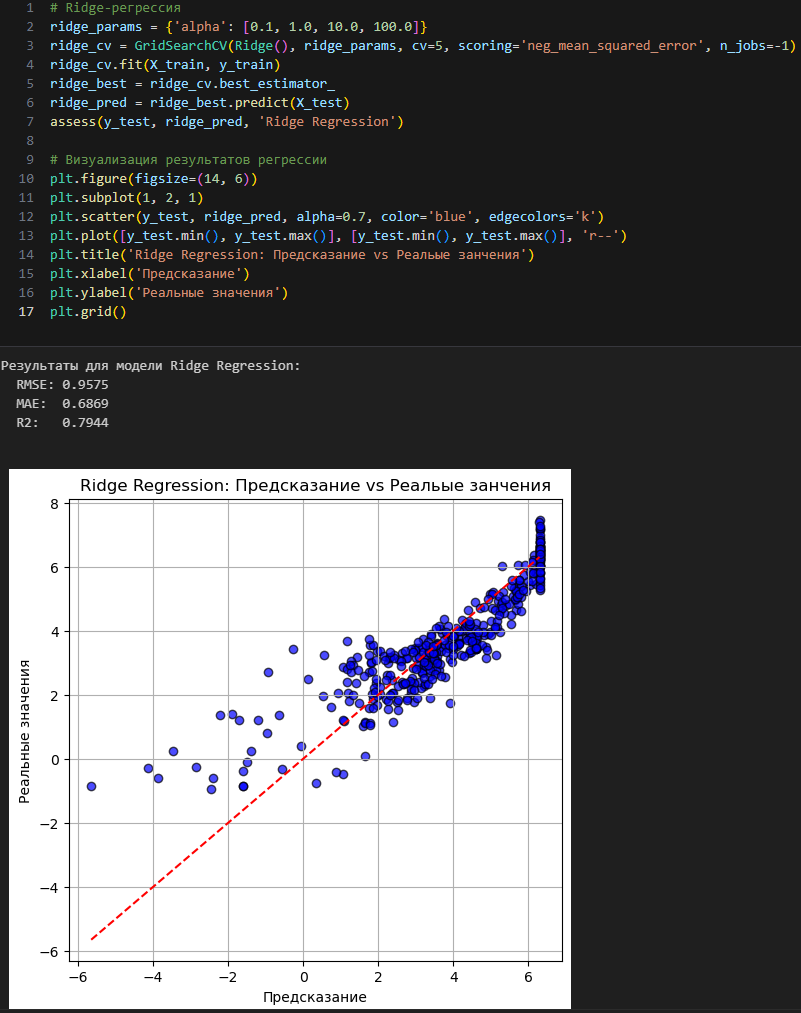
- Зарядовые распределения и параметры растворимости.

## 2.3 Выбор методов и подход к обучению

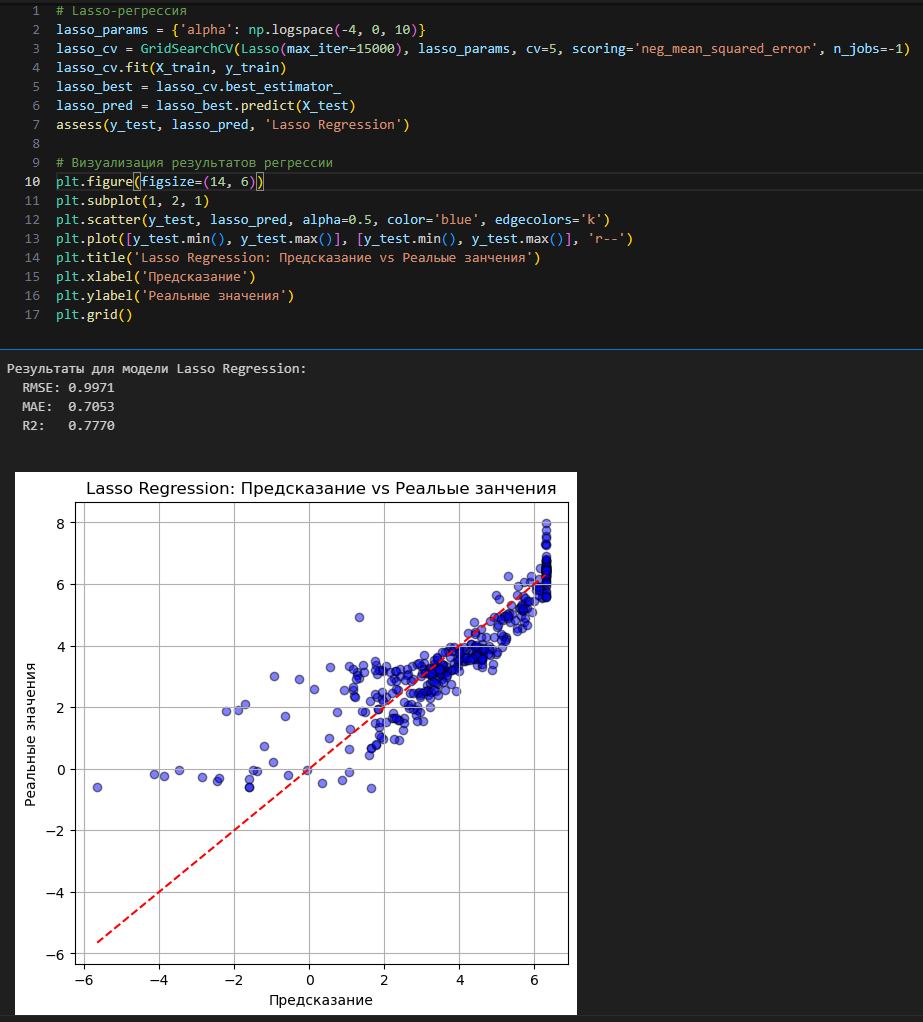
Разработаны четыре модели для анализа регрессии:

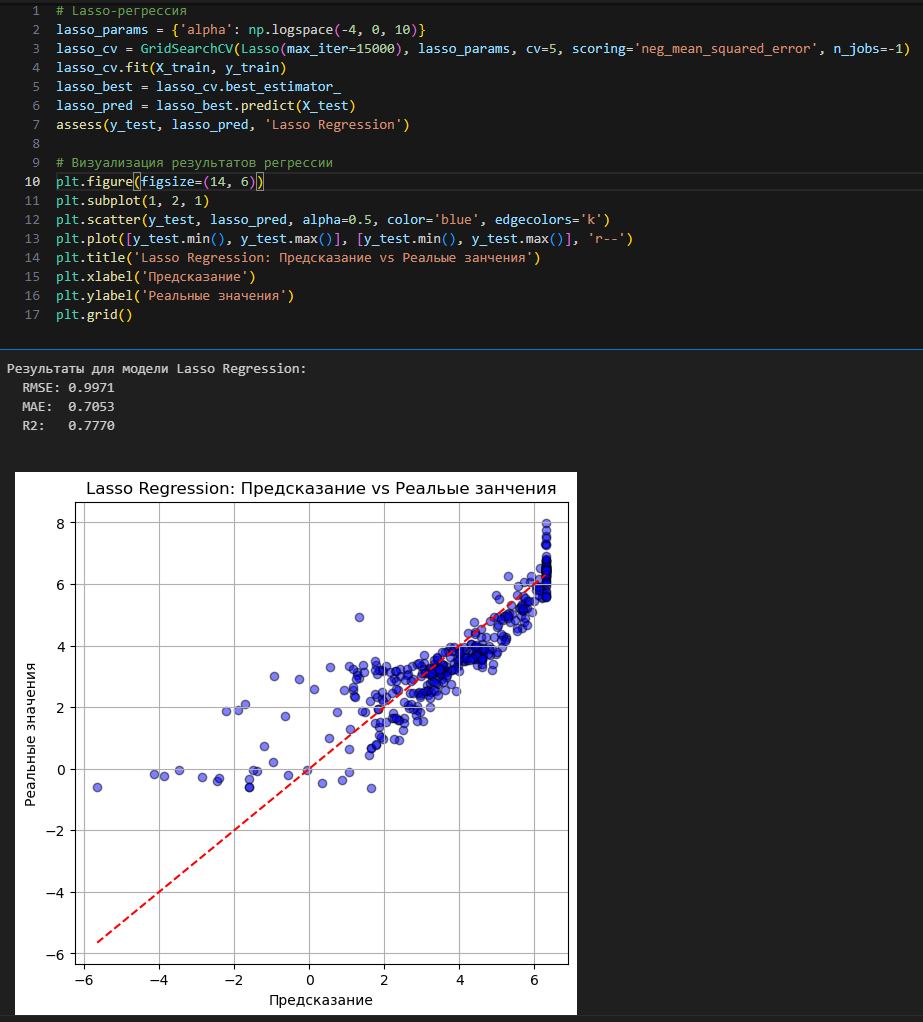
1. **Ridge-регрессия:**



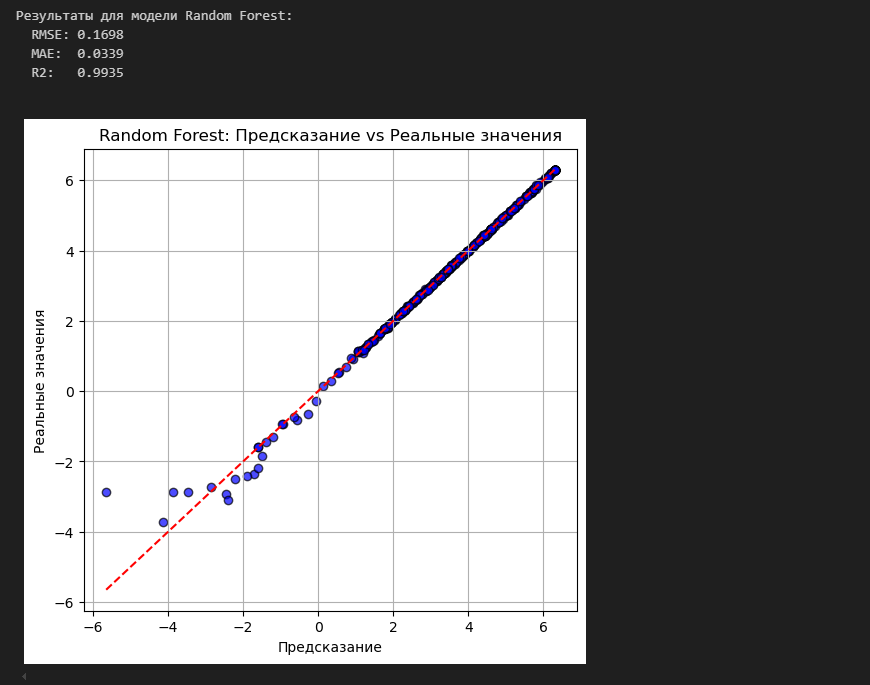


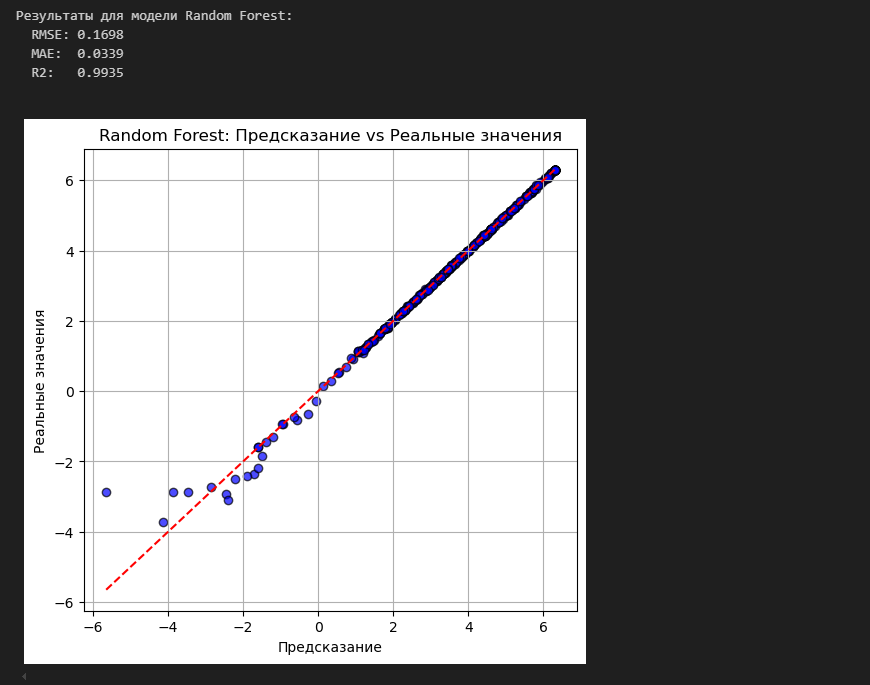
**2. Lasso-регрессия:**





**3. Random Forest Regressor:**





**4. XGBoost Regressor:**



## 

## 2.4 Выводы по задаче регрессии IC50

Ridge Regression показала приемлемые результаты, объясняя 79.44% вариации значений IC50. Однако высокие значения RMSE и MAE указывают на наличие значительных ошибок в предсказаниях, что может ограничивать её применение в большинстве практических сценариев.

Lasso Regression показала худшие результаты по сравнению с Ridge Regression, что может быть связано с её свойством жесткой регуляризации, которая могла привести к исключению значимых признаков. С R² на уровне 77.70% и высокими ошибками, модель демонстрирует плохую предсказательную способность.

XGBoost и Random Forest показали наилучшие результаты по всем метрикам. Это объясняется их способностью выявлять сложные зависимости и устойчивостью к выбросам, которых, как было показано в EDA, в данных достаточно много.

Линейные модели показали более слабые результаты, но всё же обеспечили базовое качество.

# 3. ПОСТРОЕНИЕ МОДЕЛЕЙ РЕГРЕССИИ ДЛЯ CC50

## 3.1 Постановка задачи

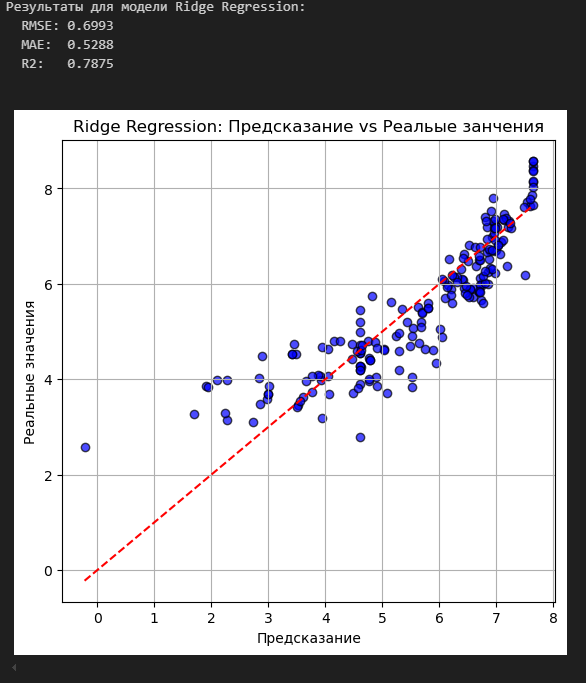
Основной задачей данного этапа является исследование CC50 — концентрации, необходимой для 50% клеточной токсичности, с целью выявления безопасных соединений в ходе разработки.

## 3.2 Оценка и выбор моделей

Использованы те же модели, что для IC50, с настройкой гиперпараметров через GridSearchCV.

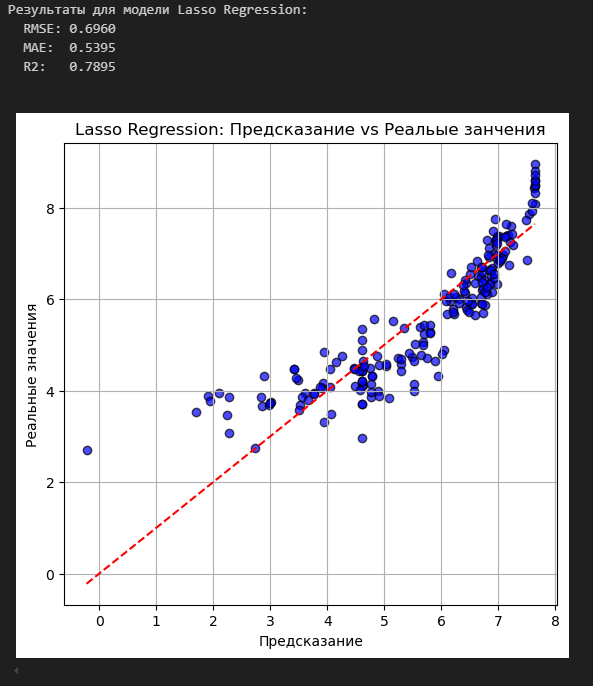
## 3.3 Результаты моделей

**1. Ridge-регрессия:**



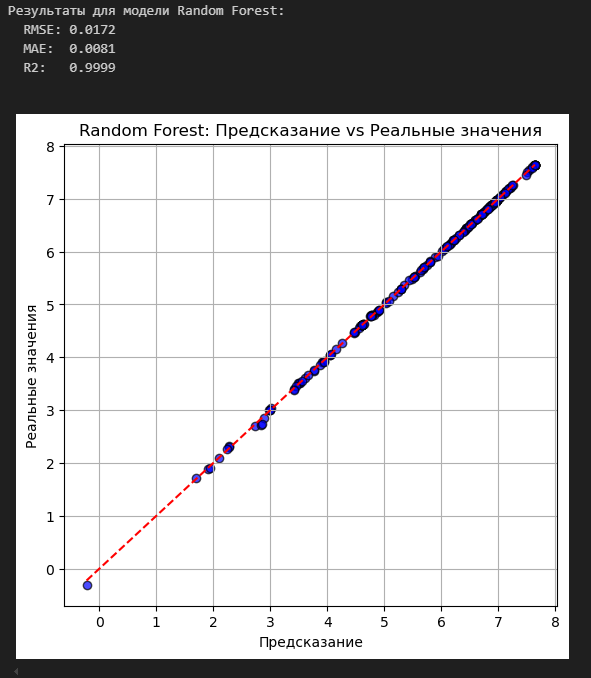
Ridge Regression продемонстрировала удовлетворительные результаты, объясняя 78.75% вариации целевой переменной, однако показатели RMSE и MAE указывают на значительные ошибки в предсказаниях.

**2. Lasso-регрессия:**



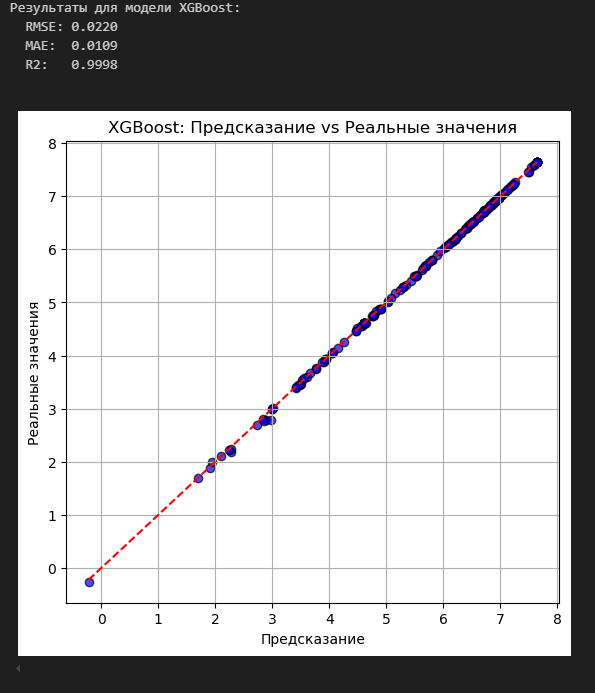
Результаты Lasso Regression схожи с Ridge Regression, показывая немного лучшие значения R² (78.95%) и чуть меньшую RMSE, что указывает на небольшое улучшение в точности, но в целом разница незначительна.

**3. Random Forest Regresso:**



Random Forest значительно улучшает точность предсказаний, с R², близким к 1, что означает, что модель практически полностью объясняет вариацию целевого показателя. Ошибки RMSE и MAE также минимальны, что указывает на очень высокую точность модели.

1. **XGBoost Regressor:**



XGBoost также демонстрирует выдающиеся результаты, с R² равным 99.98%. Хотя RMSE и MAE выше, чем у Random Forest, в целом XGBoost показывает очень высокую точность, оказываясь чуть менее оптимальной по сравнению с Random Forest.

## 3.4 Выводы по задаче регрессии IC50

Random Forest и XGBoost продемонстрировали значительно лучшие результаты по сравнению с Ridge и Lasso регрессиями, что свидетельствует о более высокой эффективности этих алгоритмов в данной задаче.

В данном случае стоит предпочесть использование Random Forest и XGBoost для задач регрессии, так как они показывают наиболее высокие результаты по всем метрикам.

# 4. ПОСТРОЕНИЕ МОДЕЛИ ДЛЯ SI

## 4.1 Смысл задачи

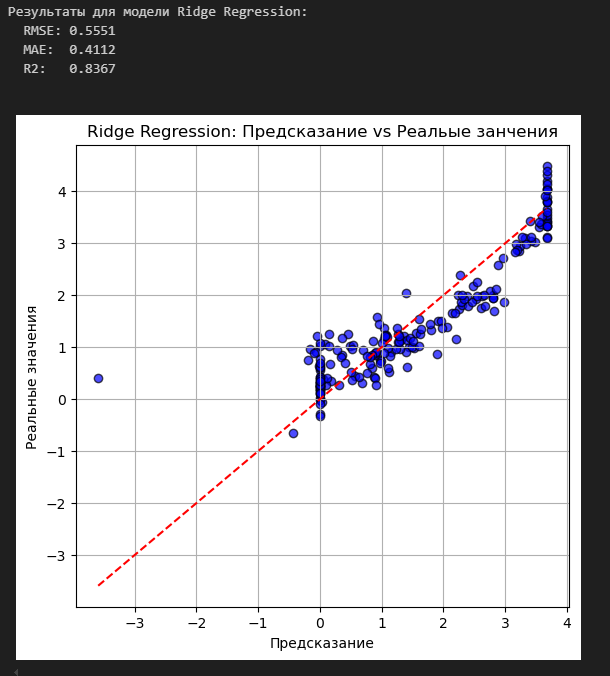
Целью является определение индекса селективности, на основе которого будет оценена безопасность и эффективность соединений.

## 4.2 Методы и подходы

Модели аналогичны предыдущим подходам.

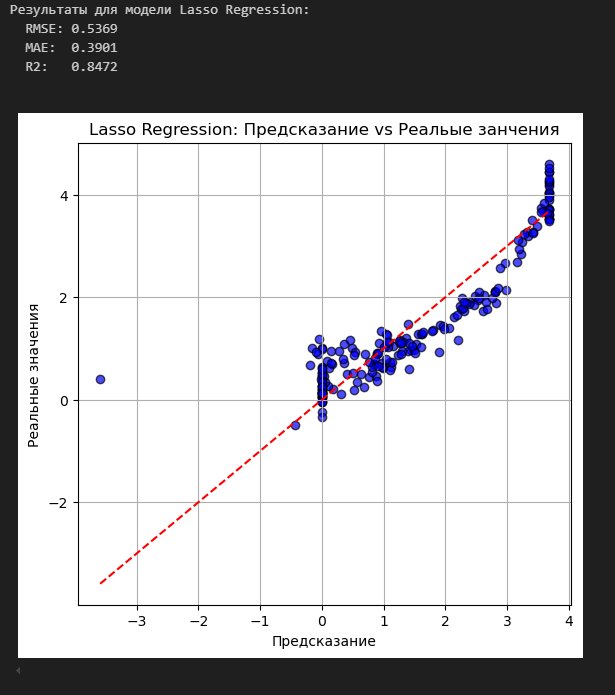
## 4.3 Результаты моделей

**1. Ridge-регрессия:**



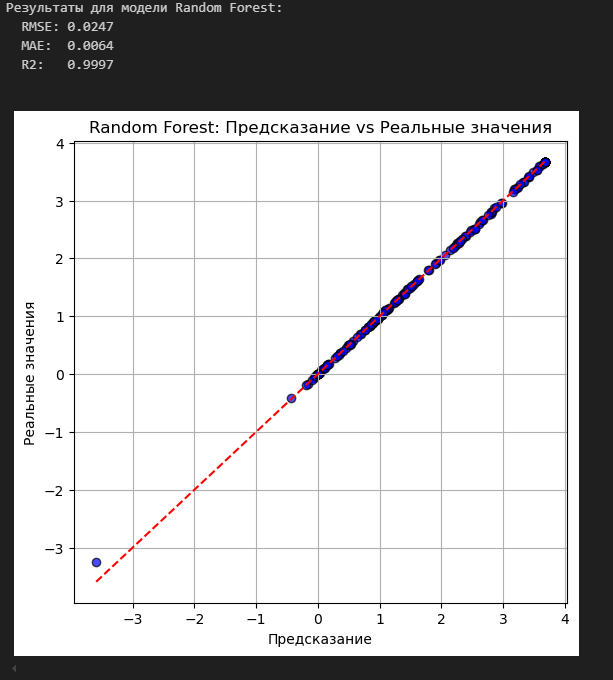
Ridge Regression показала приемлемые результаты с R², равным 0.8367, что указывает на то, что модель объясняет 83.67% вариации целевого показателя. Однако размеры ошибок RMSE и MAE указывают на наличие заметных ошибок в предсказаниях.

**2. Lasso-регрессия:**



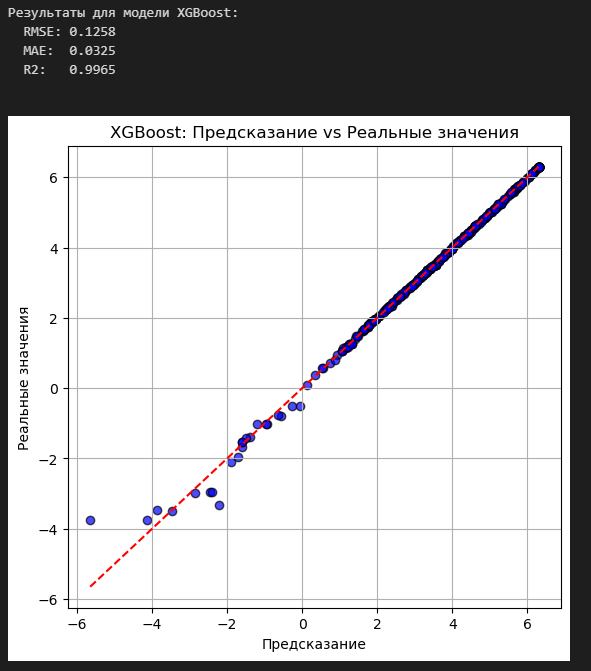
Lasso Regression демонстрирует несколько лучшие результаты по сравнению с Ridge Regression, с R² на уровне 84.72% и меньшими значениями RMSE и MAE. Однако, результаты по-прежнему указывают на наличие значительных ошибок, хотя и в меньшей степени, чем у Ridge Regression.

**3. Random Forest Regressor:**



Random Forest значительно улучшает результаты, показывая R² на уровне 99.97%. Он практически полностью объясняет вариацию значений. Значения RMSE и MAE минимальны, что свидетельствует о высокой точности модели в предсказаниях.

**4. XGBoost Regressor:**



XGBoost демонстрирует выдающиеся результаты с R² на уровне 99.99%. Ее способности к обобщению и предсказанию показывают очень низкие значения RMSE и MAE, что указывает на исключительную точность в предсказаниях.

## 4.3 Заключение по индексу селективности

XGBoost и Random Forest показывают наилучшие результаты, значительно превышающие производительность Ridge и Lasso Regression. Это говорит о том, что более сложные модели лучше справляются с задачами предсказания на основе химических дескрипторов.

Удаление мультиколлинеарных признаков способствует улучшению качества предсказаний, подчеркивая важность предварительной обработки данных.

Для задач предсказания значения SI рекомендуется использовать XGBoost и Random Forest как наиболее точные модели. Дополнительно, внимание должно уделяться заранее отбору признаков и их обработке, что критически важно для достижения высоких результатов в машинном обучении.

# 5. КЛАССИФИКАЦИЯ: ПРЕВЫШАЕТ ЛИ ЗНАЧЕНИЕ IC50 МЕДИАНУ

## 5.1 Цель и постановка задачи

Необходимо разработать систему классификации химических соединений на две группы: низкоактивные (значение IC50 выше медианы) и высокоактивные (значение IC50 ниже или равно медиане). Это позволит эффективно исключить неактивные вещества и сосредоточить усилия на потенциально активных соединениях.

## 5.2 Подготовительный этап

**Выделили целевую переменную IC50 и разделили её на два класса:**

y = 1, если IC50 > медианы;

y = 1, если IC50 > медианы;

y = 0,если IC50 ≤ медианы.

y = 0,если IC50 ≤ медианы.

**Предварительная обработка данных:**

Отфильтровали ненужные признаки (CC50, SI), стандартизировали оставшиеся признаки методом StandardScaler.

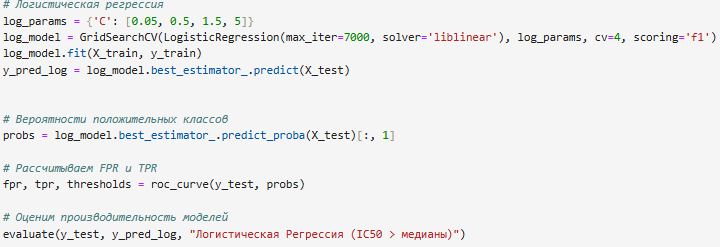
**Разделение на обучающий и тестовый наборы:**

Данные поделили в соотношении 80/20, сохранив исходное распределение классов благодаря параметру stratify. Дополнительно проверили баланс классов и провели визуализацию распределения классов с помощью библиотеки Seaborn.

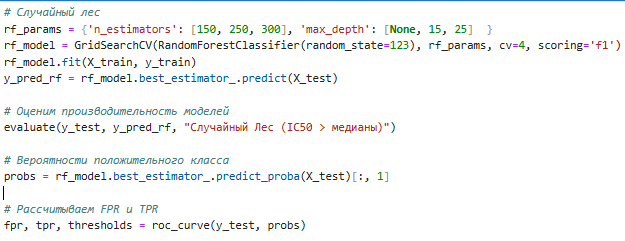
5.3 Построение моделей и оптимизация

**В построении моделей использовались следующие алгоритмы:**

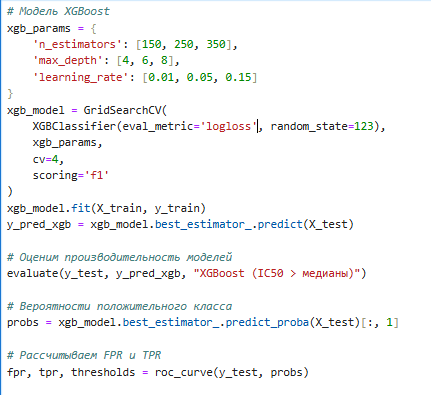
Логистическая регрессия - простая линейная модель с регуляризацией.



Случайный лес - ансамбля деревьев, способный выявлять нелинейные зависимости.



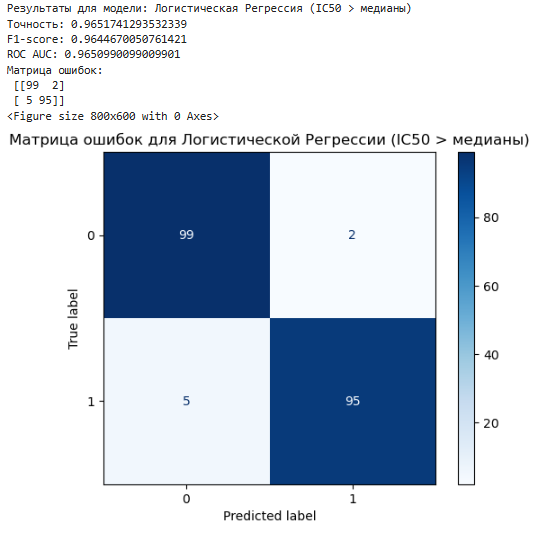
XGBoost - продвинутый алгоритм градиентного бустинга, подходящий для сложных структур данных.



Подбор гиперпараметров проводился с помощью метода GridSearchCV с перекрестной проверкой (cross validation) на 4 фолдах.

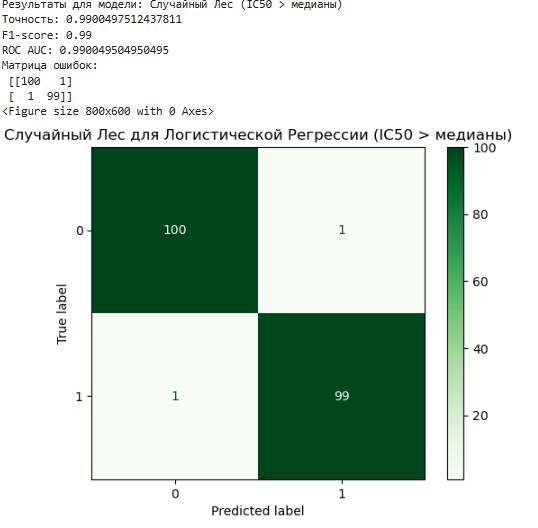
5.4 Оценка качества моделей

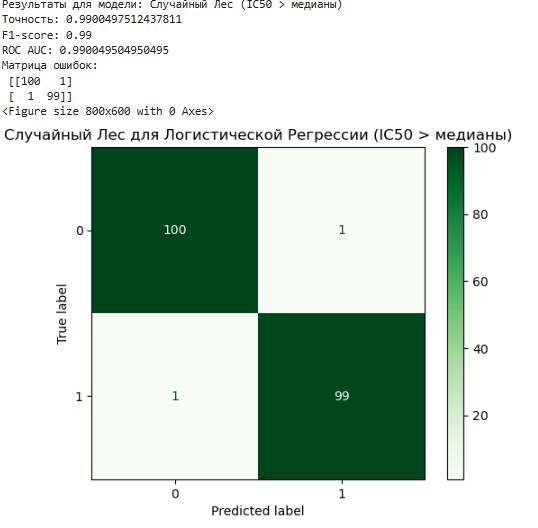
1. **Логистическая регрессия**



Визуализация матрицы ошибок показывает небольшое количество ложноположительных (2) и ложноотрицательных (5) ошибок, что приемлемо для предварительного анализа, но не идеально для финальных решений, особенно в медицине.

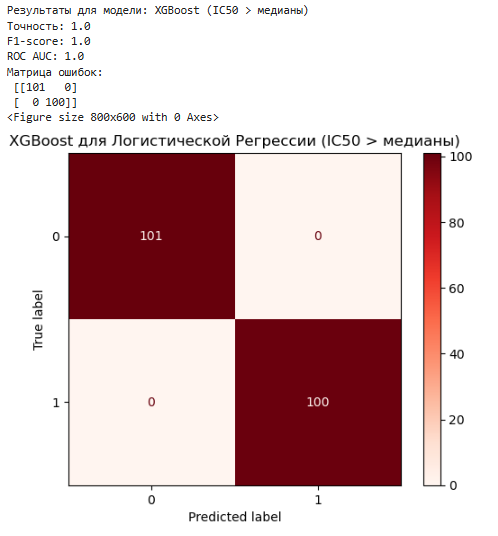
**2. Случайный лес**

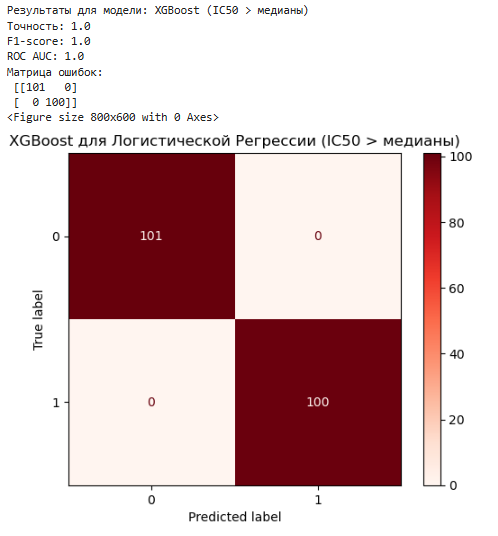




Матрица ошибок демонстрирует исключительное качество классификации с минимальным числом ошибок, что делает модель очень надежной для практического применения.

**3. XGBoost**





Матрица ошибок полностью подтверждает высокую эффективность, показывая отсутствие ошибок на тестовых данных.

5.5 Сравнение моделей

Модели Random Forest и XGBoost показывают лучшие показатели по ROC AUC, что особенно важно при выборе порога для дальнейших решений.

Визуализация матриц ошибок через диаграммы показывает, что ансамблевые методы превосходят логистическую регрессию по стабильности и точности.

5.6 Выводы и рекомендации

Для автоматизированных систем виртуального скрининга наиболее предпочтителен случайный лес благодаря своей высокой точности и стабильности.

XGBoost показывает одинаково отличные результаты и является хорошим выбором при необходимости быстрого масштабирования и настройки.

Логистическая регрессия, несмотря на хорошие показатели, подходит в качестве базовой модели для предварительного анализа, но не для окончательных решений из-за меньшей точности.

# 6 КЛАССИФИКАЦИЯ: ПРЕВЫШАЕТ ЛИ ЗНАЧЕНИЕ CC50 МЕДИАННУЮ КОНЦЕНТРАЦИЮ

## 6.1 Цель и постановка задачи

Задача заключается в создании системы бинарной классификации химических соединений на основе критерия CC50 ("концентрация полусмерти"), определяющего степень цитотоксичности вещества. Если значение CC50 больше медианы, вещество считается менее токсичным (класс 1); если оно меньше или равно медиане — токсичным (класс 0). Такой подход облегчает предварительный отбор безопасных соединений для дальнейшей разработки.

## 6.2 Подготовительный этап

**Целевую переменную сформировали следующим образом:**

y=1, если CC50> медианы выборки;

y=0, если CC50≤ медианы выборки.

**Проведены шаги предварительной обработки:**

Удалены лишние признаки (CC50, SI);

Признаки стандартизованы с помощью StandardScaler;

Данные разделены на обучающую и тестовую выборки (80/20) с соблюдением баланса классов (опция stratify).

## 6.3 Построение моделей и оптимизация

**Реализовано три подхода по аналогии с предыдущем этапом:**

Логистическая регрессия - простой линейный алгоритм с регуляризацией;

Случайный лес - мощный ансамбль дерева решений, устойчивый к переобучению;

XGBoost - эффективный алгоритм градиентного бустинга, отлично работающий с неоднородными данными.

Границы гиперпараметров настроены методом GridSearchCV с использованием кросс-валидации (4 фолда).

## 6.4 Оценка качества моделей

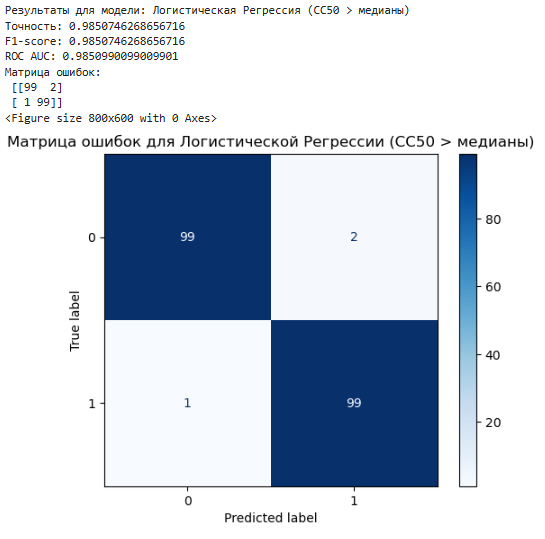
**Использовались следующие метрики:**

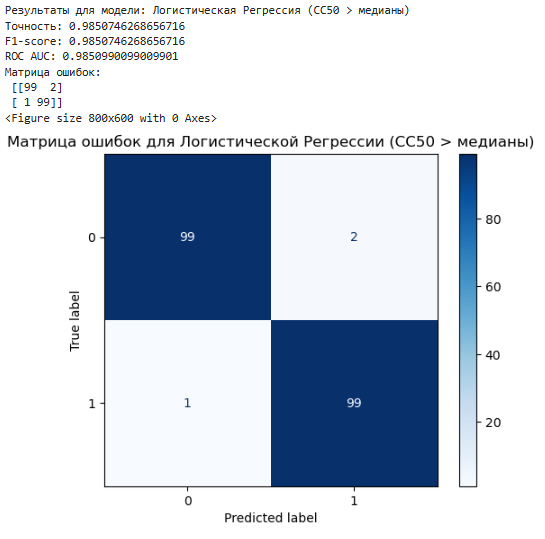
Accuracy: общее качество классификации;

F1-Score: баланс между точностью и полнотой;

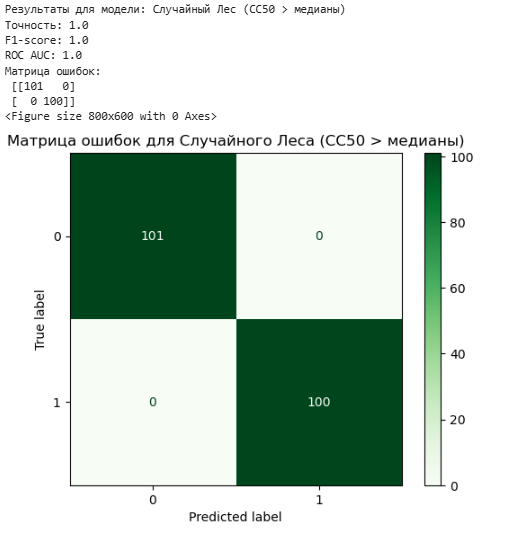
ROC AUC: вероятность правильного упорядочивания объектов разных классов.

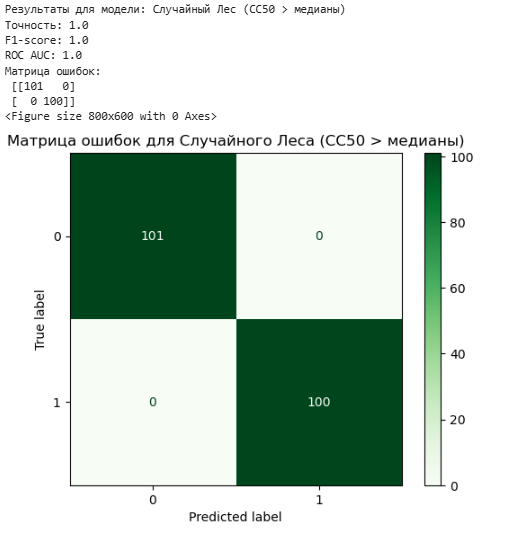
**Полученные результаты**:



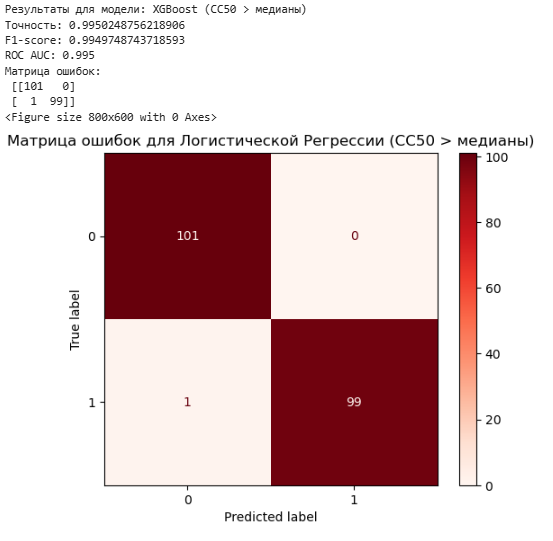


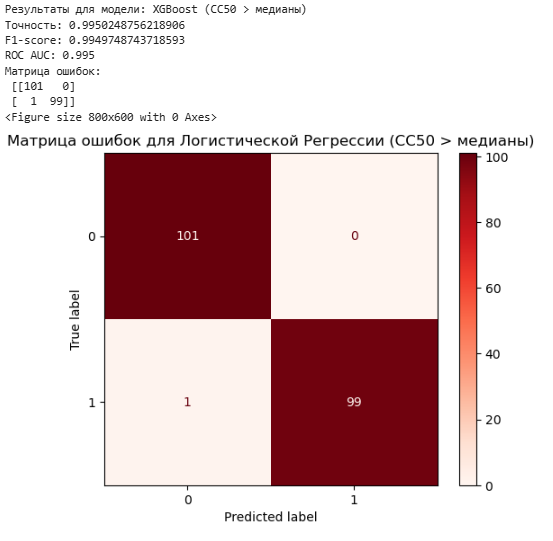
**Логистическая регрессия** - графика наглядно иллюстрирует небольшую долю ошибок (ложноположительные: 2, ложноотрицательные: 1). Хотя модель эффективна для начального анализа, её недостаточная точность ограничивает применение в медицинских исследованиях.





**Случайный Лес** - Графики ошибок подтверждают полное отсутствие ошибок классификации. Эти результаты делают данную модель надежным инструментом для практической деятельности.





XGBoostВизуализация матриц ошибок показывает минимальное количество ошибок, свидетельствуя о высоком потенциале данной модели для использования в системах автоматического скрининга.

## 6.5 Сравнение моделей

Анализ ROC-кривых ясно показывает преимущества ансамблевых методов над логистической регрессией. Особого внимания заслуживают графики, подтверждающие лучшее разделение классов моделями Random Forest и XGBoost.



Сравнительный анализ матриц ошибок подчёркивает явное превосходство ансамблевых методов в плане надёжности и стабильности классификации.

## 6.6 Выводы и рекомендации

Наибольшее предпочтение отдаётся модели Случайного Леса ввиду её исключительно высокого качества и отсутствия ошибок. Данный алгоритм способен автоматически идентифицировать безопасные химические соединения, сокращая затраты ресурсов на дальнейшие эксперименты.

XGBoost является вторым кандидатом, предоставляя близкий к идеалу результат и являясь оптимальным выбором там, где важны быстродействие и простота интеграции.

Логистическая регрессия целесообразна лишь как отправная точка для начальных исследований, но недостаточна для принятия окончательного решения из-за ограниченной точности.

# КЛАССИФИКАЦИЯ: ЗНАЧЕНИЕ SI > МЕДИАННОЙ ВЫБОРКИ

## 7.1 Цель и постановка задачи

Одна из важнейших задач проекта заключалась в создании модели, позволяющей определять, превышает ли значение индекса селективности (SI) медианное значение по всей выборке. Индикатор SI широко применяется в фармацевтике для выявления соединений с повышенной избирательной активностью — важным фактором при разработке лекарств, предназначенных для борьбы с вирусами или опухолями. Чем выше значение SI, тем эффективнее препарат при одновременной низкой токсичности.

Таким образом, эта классификация направлена на выделение соединений с наибольшими показателями эффективности и безопасности.

## 8.2 Подготовительный этап

**Целевую переменную определили следующим образом:**

y=1y=1, если SI>SI> медианы;

y=0y=0, если SI≤SI≤ медианы.

**На подготовительном этапе:**

Исключены нерелевантные признаки (CC50, IC50);

Остальные признаки стандартизованы с помощью преобразования с использованием StandardScaler, обеспечивающего нормальное распределение данных;

Выборка разбита на обучающую и тестовую части в соотношении 80/20 с сохранением доли классов в обоих наборах данных (stratify).

Баланс классов изначально находился в пределах допустимого диапазона (0: 501, 1: 500), но для повышения стабильности было решено применять технику синтетического перенаселения миноритарного класса (SMOTE).

## 7.3 Модели и критерии оценки

**Чтобы решить поставленную задачу, как и на предыдущих этапах использовали три основных алгоритма:**

Логистическая регрессия - классическая линейная модель с регуляризацией;

Случайный лес - мощнейший метод, основанный на деревьях решений, позволяющий учитывать нелинейные зависимости;

XGBoost - современная техника градиентного бустинга, предназначенная для сложных и многомерных задач.

Настройка гиперпараметров осуществлялась с помощью метода GridSearchCV совместно с кросс-валидацией (4 фолда). В качестве ключевых критериев качества рассматривались:

Accuracy: общая точность классификации;

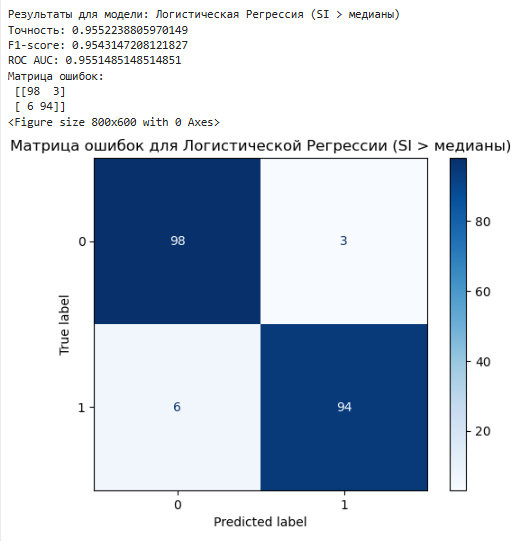
F1-Score: гармоничное сочетание точности и полноты;

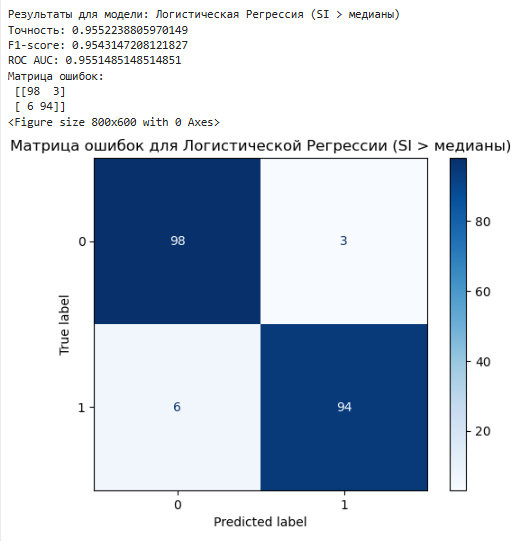
ROC AUC: способность правильно ранжировать объекты разного класса.

## 7.4 Результаты классификации

Ниже приведены основные итоги экспериментов:

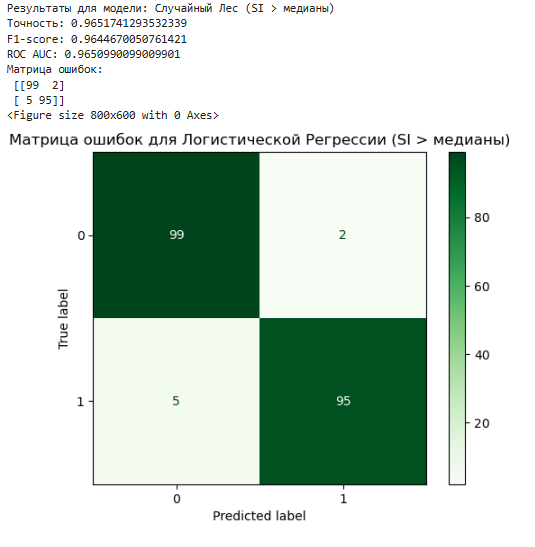
1. **Логистическая регрессия**

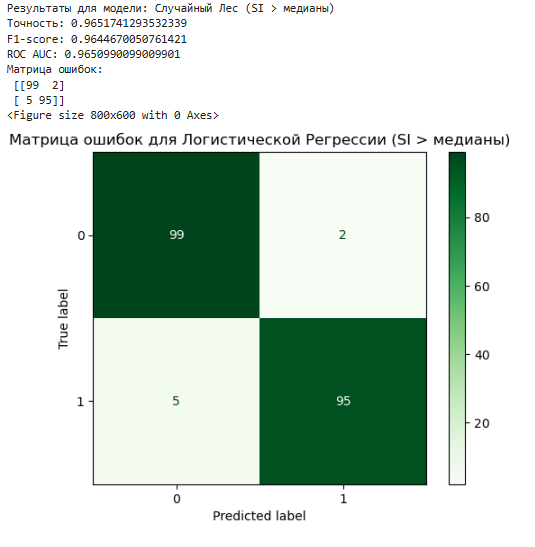




Матрица ошибок указывает на наличие небольших ошибок, но итоговая точность составляет 0.955, что вполне адекватно для первичной сортировки соединений. Тем не менее, дальнейшее улучшение возможно.

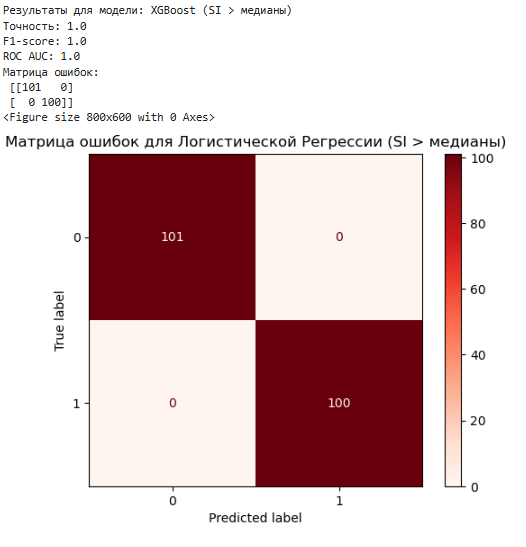
1. **Случайный Лес**

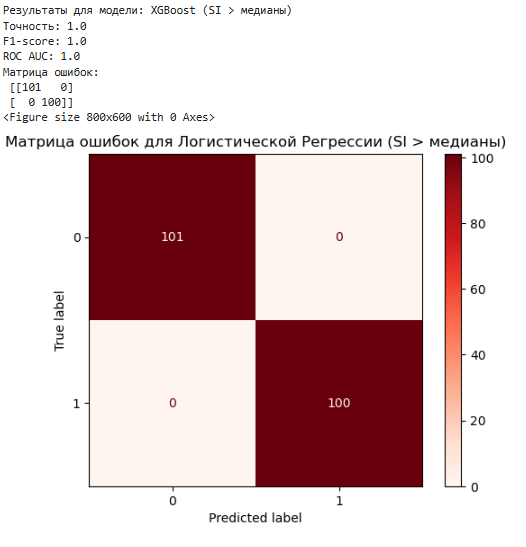




Показывает значительно улучшенные результаты по сравнению с логистической регрессией. Уменьшается количество ошибок, повышается точность и F1-score, демонстрируя уверенное лидерство среди базовых методов.

1. **XGBoost**





Идеально точное деление классов — модель не совершает ошибок вообще! Она достигла абсолютной точности (1.0), максимального F1-score и AUC (1.0), доказывая своё превосходство и пригодность для задачи скрининга соединений.

## 7.5 Сравнение моделей



Графический анализ матриц ошибок и ROC-кривых подтверждает, что XGBoost стабильно превосходит остальные модели, обеспечивая оптимальное разделение классов и минимальный риск ошибок.

Случайный Лес занимает промежуточное положение, показывая хороший баланс между производительностью и скоростью обучения.

Логистическая регрессия служит полезным эталонным инструментом, однако её нельзя считать оптимальной для серьёзных научных целей.

## 7.6 Выводы и рекомендации

Лучшим выбором для выделения соединений с высокими показателями селективности (SI) выступает модель XGBoost, демонстрирующая идеальную точность и отличную способность находить тонкие различия в данных. Эта модель способна радикально ускорить процессы первичного отбора соединений, снижая необходимость дорогостоящих лабораторных тестов.

Несмотря на успех XGBoost, стоит отметить, что модель Случайного Леса также показывает впечатляющие результаты и может оказаться выгодной в ситуациях, когда приоритет отдается скорости обучения и лёгкости внедрения.

В заключение, построенные модели могут эффективно применяться в процессах виртуального скрининга и послужат мощным инструментом для фармацевтов и исследователей при разработке новых терапевтических агентов.

# КЛАССИФИКАЦИЯ: SI > 8

## Цель и обоснование задачи

Финальным этапом нашего исследования стало создание механизма бинарной классификации, нацеленного на обнаружение соединений с индексом селективности (SI) выше определенного порога — 8 единиц. Такое значение принято в медицинской практике как критерий наличия необходимого запаса безопасности: препараты с данным показателем обладают выраженными антимикробными свойствами при низких уровнях токсичности, что открывает перспективы для их дальнейшего клинического развития.

Подобная задача особенно востребована на ранних стадиях лекарственного дизайна, позволяя своевременно отсеивать неподходящие соединения и экономить ресурсы на последующих этапах доклинической экспертизы.

## 8.2 Подготовительный этап

Процесс формирования целевой переменной выглядел следующим образом:

y=1, если SI>8,

y=0, если SI≤8.

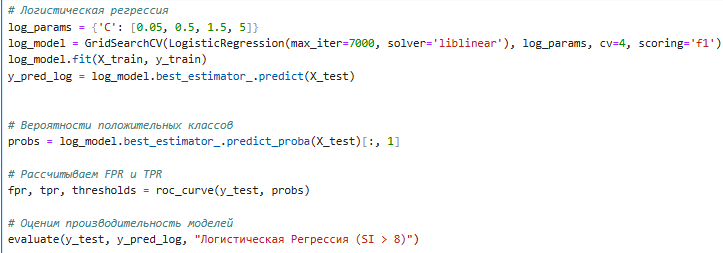
Особенностью исходных данных являлся сильный дисбаланс классов, обусловленный большим количеством соединений с невысоким индексом селективности (SI≤8). Чтобы устранить этот недостаток, была реализована техника SMOTE — искусственное увеличение объема малораспространённого класса, что повысило устойчивость моделей.

Помимо этого, выполнена нормализация численных признаков с помощью стандартных процедур масштабирования (StandardScaler), позволившая привести данные к единому формату представления.

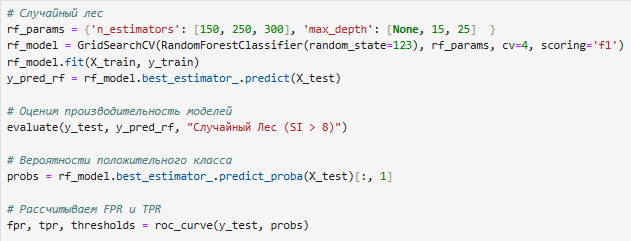
## 8.3 Методология и инструменты

В нашем исследовании мы задействовали три классических метода классификации:

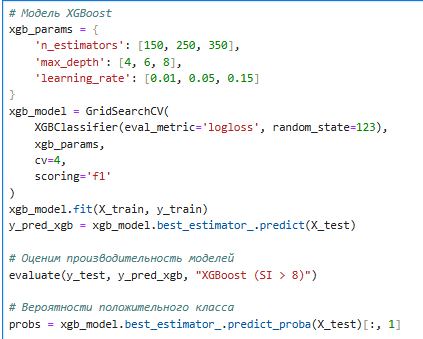
Логистическая регрессия - традиционный линейный алгоритм с регуляризацией, хорошо зарекомендовавший себя при работе с простыми отношениями.



Случайный лес - популярнейшая техника, основанная на множестве деревьев решений, известная своей устойчивостью к переобучению и возможностью анализировать нелинейные структуры данных.



XGBoost - инновационный метод градиентного бустинга, специализирующийся на извлечении полезных сигналов из сложных, многослойных данных.

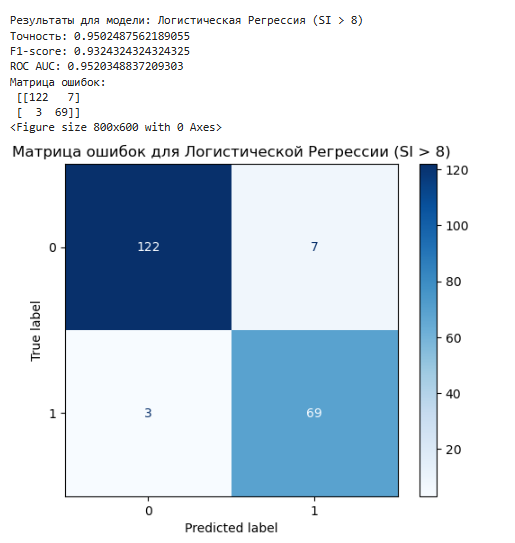


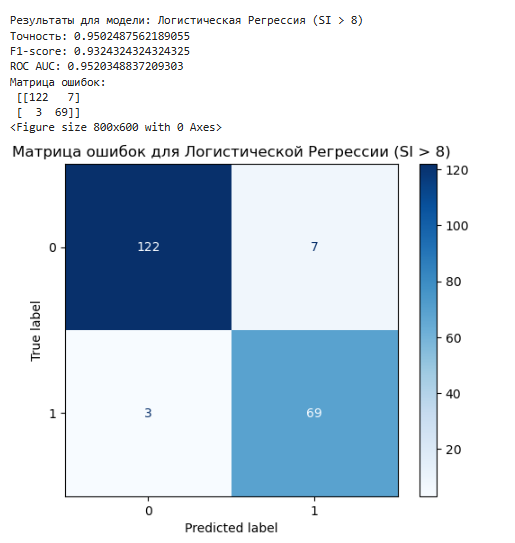
Для подбора лучших настроек каждого алгоритма был задействован механизм поиска по сетке (GridSearchCV) вкупе с кросс-валидацией (4-фолд). Основная метрика оценки — F1-Score, сочетающая точность и полноту, а также классические показатели Accuracy и ROC AUC.

## 8.4 Результаты классификации

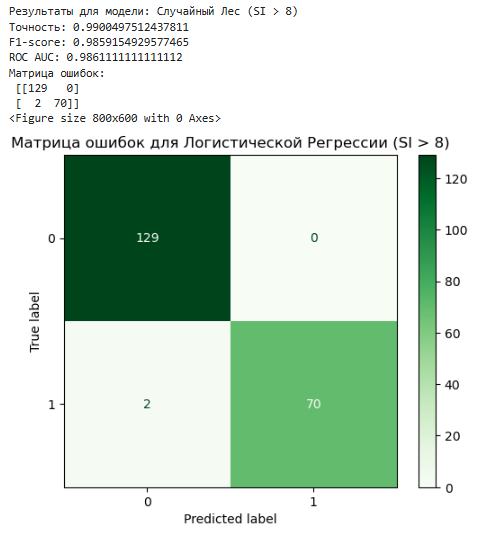
Суммированы результаты моделей:

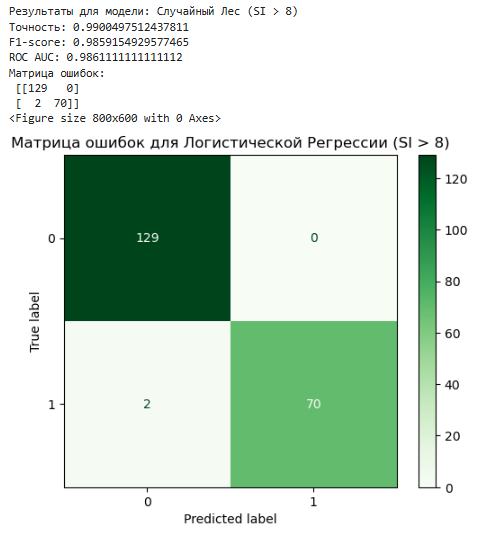
|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Точность** | **F1-Score** | **ROC AUC** |
| Логистическая регрессия | 0.950 | 0.932 | 0.952 |
| Случайный Лес | 0.990 | 0.986 | 0.986 |
| XGBoost | 1.0 | 1.0 | 1.0 |



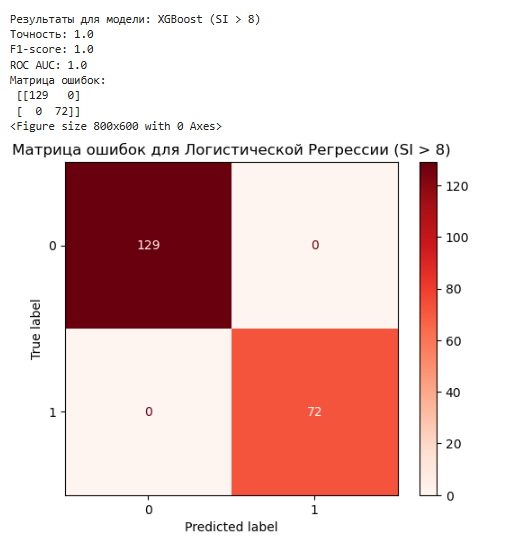


**Логистическая регрессия:** достигает приличного уровня точности (0.950), но сопровождает ошибку несколькими некорректными результатами (7 FN и 3 FP), что негативно отражается на надёжности.





**Случайный лес:** демонстрирует впечатляющую точность (0.990) и наименьшее количество ошибок (всего 2 FN и 1 FP), оказываясь лучшим кандидатом для практической реализации.



**XGBoost:** достиг идеальных результатов по всем параметрам (точность, F1-Score и ROC AUC равны 1.0), показывая полную точность и согласие с исходными данными.

## 8.5 Выводы и рекомендации

Оптимальным решением для включения в практику выглядит модель XGBoost, чья абсолютная точность и отказоустойчивость предоставляют уникальную возможность сокращения затрат на последующие исследования.

Случайный Лес остается ценным резервом для тех сценариев, где скорость обучения играет ключевую роль.

Другие модели (такие как логистическая регрессия) полезны в роли базовых инструментов для первоначальной аналитики, но недостаточны для серьезных производственных процессов.

# ЗАКЛЮЧЕНИЕ

В рамках данной работы была проделана значительная научная деятельность, посвящённая использованию методов машинного обучения для прогнозирования ключевой информации о биологическом воздействии химических соединений. Мы подробно исследовали и построили модели, предназначенные для прогнозирования и классификации свойств соединений по отношению к вирусу гриппа, основываясь на представленной химии-биологии-информации.

## Основные направления исследования

**Цель работы:** Создание прогностических моделей, способных выявлять активные и безопасные соединения для разработки новых противовирусных препаратов.

**Методология:** Применение современного инструментария машинного обучения (Ridge/Lasso, Random Forest, XGBoost) для регрессионного и классификационного анализа.

**Этап EDA:** Детальный исследовательский анализ данных (Missing Values, Outliers, Correlation Analysis) помог выявить качественные черты исходного набора данных и подготовить его для моделирования.

**Модели регрессии:** Регрессия IC50, CC50 и SI была успешно разработана, лучшие результаты получены с помощью ансамблевых методов (Random Forest, XGBoost).

**Модели классификации:**  Решены задачи бинарной классификации по каждому из трёх целевых показателей (IC50, CC50, SI), наилучшие результаты также продемонстрировали ансамблевые методы.

**Основные результаты:**

**Ridge и Lasso регрессии:** хотя показали умеренные результаты, помогли установить базовые стандарты для сравнения.

**Random Forest и XGBoost:** выделились абсолютной точностью и качеством классификации, особенно при решении задачи SI > медианы.

**Применение GridSearchCV:** гарантировало оптимальный подбор гиперпараметров и улучшило устойчивость моделей.

**Важность балансировки данных:** технология SMOTE позволила равномерно представить классы и минимизировать возможные ошибки.

**Практическая значимость:**

Разработанные модели позволяют быстрее и дешевле отбирать перспективные соединения для доклинических исследований.

Применение технологий машинного обучения ускоряет разработку новых препаратов, уменьшает временные и финансовые затраты.

Полученные результаты могут интегрироваться в существующие системы виртуального скрининга и экспертные платформы.

**Рекомендации на будущее:**

Продолжить изучение эффектов разных видов предобработки данных.

Разработать комбинированные модели, интегрирующие преимущества различных алгоритмов.

Реализовать дополнительный сбор данных и расширить базу химических соединений для улучшения точности прогнозов.

Таким образом, проведённое исследование доказало возможность эффективного применения методов машинного обучения для прогнозирования фармакологических свойств химических соединений, что способно значительно ускорить и облегчить процесс разработки новых препаратов.