Курсовую работу выполнил: Кабанов Алексей

**ОТЧЕТ**

1. **Описание данных**

Первый этап разработки медикаментов включает в себя определение безопасных и потенциально активных молекул на ранних стадиях, прежде чем переходить к дорогостоящим in vitro и in vivo исследованиям. В этом процессе важную роль играют методы химоинформатики и машинного обучения, которые помогают прогнозировать фармакологические свойства соединений на основе их молекулярной структуры.

В данной работе рассматривается задача предсказания трех ключевых биологических показателей:

* **CC₅₀,(полумаксимальная цитотоксическая концентрация)** определяющий токсичность соединения для клеток, то есть концентрацию, при которой погибает 50% клеток;
* **IC₅₀, (полумаксимальная ингибирующая концентрация)** отражающий эффективность соединения как ингибитора определенного биологического процесса;
* **SI (индекс селективности**), который рассчитывается как отношение CC₅₀ к IC₅₀ и показывает соотношение активности и токсичности. Более высокое значение SI свидетельствует о большей безопасности и избирательности соединения.

Целью курсовой работы является создание и оценка моделей машинного обучения (регрессии и классификации), способных предсказывать значения указанных показателей на основе структурных молекулярных дескрипторов с использованием обширного набора признаков:

1. **Общемолекулярные дескрипторы**

* **MolWt:** молекулярная масса.
* **HeavyAtomCount:** количество тяжёлых атомов (не включая водород).
* **NumValenceElectrons:** общее количество валентных электронов.
* **NumRadicalElectrons:** количество неспаренных (радикальных) электронов.
* **FractionCSP3:** доля spі3-гибридизованных атомов углерода. TPSA: топологическая полярная поверхность, оценивает проницаемость через мембраны.
* **MolLogP:** логарифм коэффициента распределения (гидрофобность).
* **MolMR:** молекулярная рефрактивность (показатель поляризуемости). QED представляет собой комплексную числовую характеристику "лекарственной способности" молекулы, которая учитывает её химические и физические свойства. Этот показатель позволяет оценить потенциал соединения для создания медикаментов. На практике расчет QED осуществляется с применением специализированного программного обеспечения и библиотек в области химической информатики, таких как RDKit.

1. **Топологические дескрипторы**

* **HallKierAlpha:** эмпирический дескриптор стерической насыщенности т.е. насколько молекула "заполнена" в

пространстве.

* **Kappa1, Kappa2, Kappa3:** индексы Кьера. которые характеризуют форму молекулы, её компактность и степень
* разветвлённости. **Chi0, Chi1, ..., Chi4v:** индексы связности Чи. отражающие молекулярную топологию и связанные с числом
* связей, типом атомов и степенью разветвления. **BalabanJ:** индекс связности Балабана (BalabanJ) принимает во внимание длину путей и наличие циклов в молекуле, что даёт возможность оценить уровень разветвлённости и топологическую сложность структуры.
* **Ipc, AvgIpc, BertzCT:** индексы, отражающие информационные и сложностные характеристики, основанные на графовом анализе молекулы. Эти индексы служат для количественной оценки разнообразия и запутанности структуры.

1. **Электронные дескрипторы**

* **MaxPartialCharge / MinPartialCharge / MaxAbsPartialCharge / MinAbsPartialCharge**: максимальные значения
* частичных зарядов. **PEOE\_VSA**: набор дескрипторов, описывающих распределение электрических зарядов, полученных с помощью метода выравнивания орбитальной электроотрицательности (PEOE).
* **EState\_VSA**: сумма данных о состоянии заряда и их пространственной конфигурации в молекуле.
* **MaxEStateIndex / MinEStateIndex / MaxAbsEStateIndex /** **MinAbsEStateIndex:** максимальные значения индекса

электротопологического состояния.

1. **VSA-дескрипторы**

* **VSA**-дескрипторы (Van der Waals Surface Area) — представляют собой параметры, которые иллюстрируют распределение физико-химических свойств по поверхности молекулы. Эти характеристики основаны на площади ван-дер-ваальсовых поверхностей атомов.
* **SMR\_VSA1–10:** связаны с молекулярной рефрактивностью.
* **SlogP\_VSA1–12**: связь с гидрофобностью (logP).
* **EState\_VSA1–10**: электротопология по поверхности.
* **PEOE\_VSA1–14**: связь с частичными зарядами. 5

1. **BCUT-дескрипторы**

Метод визуализации молекулярной структуры через числовой вектор, который отражает как физико-химические характеристики, так и структурные детали. Для этого используется взвешенная матрица смежности, созданная на основе молекулярной структуры, с последующим получением собственных значений этой матрицы.

* **BCUT2D\_MWHI / MWLOW:** с учётом молекулярной массы.
* **BCUT2D\_CHGHI / CHGLOW:** по заряду.
* **BCUT2D\_MRHI / MRLOW:** по молекулярной рефрактивности.

1. **Фрагментные дескрипторы**

Это числовые характеристики, показывающие, присутствуют ли в молекуле определенные химические группы или структурные элементы, а также указывающие на их количество. Они могут принимать бинарные значения (присутствует/отсутствует) или быть счетными (количество повторений).

* **Фенолы:** fr\_phenol, fr\_Ar\_OH.
* **Амины:** fr\_NH2, fr\_amine, fr\_aniline.
* **Азосоединения:** fr\_azide, fr\_azo, fr\_diazo.
* **Галогены:** fr\_halogen, fr\_alkyl\_halide.
* **Барбитураты:** fr\_barbitur.
* **Нитро-соединения**: fr\_nitro, fr\_nitro\_arom.
* **Лактон/лактам:** fr\_lactone, fr\_lactam.
* **Кольца:** fr\_benzene, fr\_pyridine, fr\_furan, fr\_thiazole.

1. **Структурные количественные дескрипторы**

Характеризуют ключевые аспекты молекулярной структуры, которые отображают возможные свойства молекул, оказывающие влияние на их фармакокинетику, реакционную активность и эффективность в качестве лекарств.

* **NumHAcceptors / NumHDonors**: акцепторы/доноры водородных связей.
* **NumRotatableBonds:** количество вращающихся связей.
* **NumAromaticRings / NumAliphaticRings / NumSaturatedRings:** кольцевые структуры.
* **NumHeteroatoms:** количество гетероатомов.
* **RingCount:** общее количество колец.

1. **Задача посторения моделй**
   1. **Задача классификации**

Модели классификации будут применяться для распределения соединений по категориям, таким как активные и неактивные (на основе порога IC₅₀), токсичные и нетоксичные (по порогу CC₅₀), а также селективные и неселективные (по порогу SI).

Цели, которые необходимо достигнуть:

* важно установить, можно ли надежно делить соединения по заранее установленным биологическим критериям; определить,
* какие характеристики имеют наибольшее значение для разделения классов; а также оценить возможность использования модели для исключения неэффективных или опасных соединений на ранних этапах исследования.

Выводы, которые можно сделать после создания моделей:

* выявить, какие структурные черты чаще всего встречаются у активных и безопасных соединений;
* оценить, насколько эффективны модели для задач высокопроизводительного скрининга (HTS);
* определить, в какой мере модели могут быть использованы для первоочередного выбора кандидатов на синтез.
  1. **Задача регрессии**

Необходимо продемонстрировать следующее:

* какова степень точности в предсказании значений активности, токсичности и селективности соединений, основываясь на их структурных особенностях;
* какие дескрипторы (структурные характеристики) оказывают наибольшее воздействие на каждую из метрик;
* как можно сопоставлять соединения друг с другом и оценивать их по предсказанным значениям активности и безопасности.

Из построенных моделей можно попробовать сделать такие выводы:

* какие структурные параметры в наибольшей степени связаны с высокой активностью (низким IC₅₀) или значительной токсичностью (низким CC₅₀);
* возможно ли разработать «идеальное» соединение с высоким SI (высокой селективностью);
* насколько точно дескрипторы могут быть применены для количественного предсказания (оценка возможности использования модели в химическом дизайне).

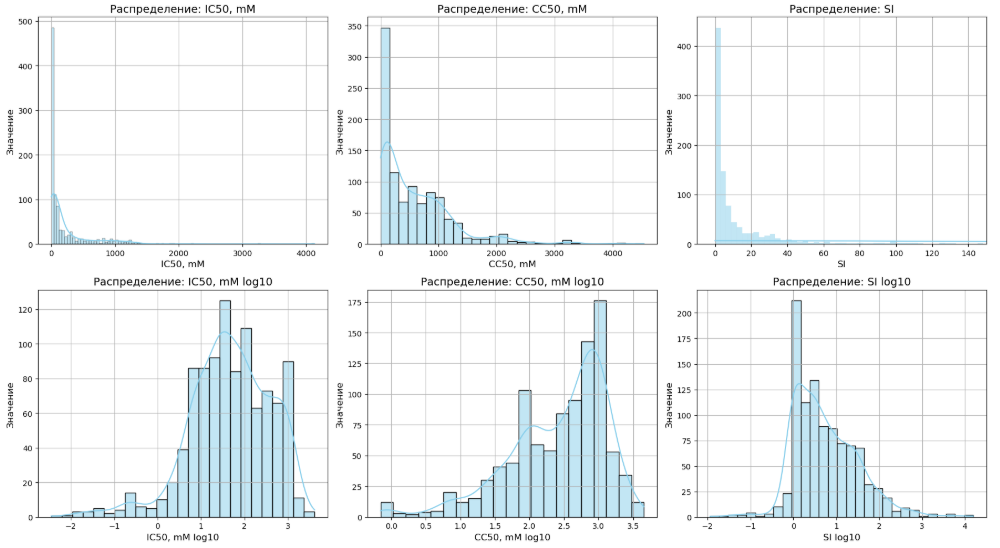
1. **Анализ и подготовка данных**
   1. **Обработка целевых переменных**

На первом этапе исследуем целевые признаки:

* анализ пропусков в данных;
* замена пропусков в даных на нули;
* удаление строк с пропусками, отрицательными значениями или нулями в целевых переменных.
  1. **Обработка данных:**
* замена пропусков в данных на нули;
* бинарная трансформация столбцов, где пропусков в данных > 75% – бинарная трансформация полезна для столбцов с высокой долей пропусков (или заменённых нулями значениями), потому что она позволяет явно выделить наличие или отсутствие исходных данных. Если более 75% значений в столбце отсутствуют, преобразование такого столбца в бинарный (например, 1 — признак был заполнен, 0 — был пропуск/ноль) помогает модели лучше различать важные структурные особенности данных и уменьшить влияние шума, вызванного массовыми пропусками. Такой подход также может улучшить качество предсказаний моделей, поскольку они будут оперировать более информативным признаком, отражающим сам факт пропуска информации, а не случайное или подменное значение;
* заполнение пропусков в данных медианой – замена пропусков в данных полезна при анализе датасета, потому что позволяет сохранить максимальное количество информации, не теряя значимых наблюдений из-за отсутствующих значений. Такой подход помогает предотвратить снижение объёма выборки и потенциальное искажение результатов, как это бывает при удалении строк с пропусками. Замена может быть предпочтительной, потому что корректно проведённый процесс замены поддерживает целостность и корректность анализа, снижая влияние пропусков на итоговые выводы и повышая устойчивость моделей машинного обучения;
* логарифмирование данных – логарифмирование данных в датасете полезно тем, что позволяет уменьшить влияние выбросов, нормализовать распределение и сделать его ближе к нормальному, а также улучшить масштабируемость признаков, что положительно сказывается на работе многих статистических моделей и алгоритмов машинного обучения. Однако логарифмическая функция определена только для положительных значений, поэтому важно отбросить или обработать отрицательные значения в данных заранее, чтобы избежать ошибок и некорректных результатов при трансформации;
* логарифмирование с сохранением знака – для данных, содержащих отрицательные значения, стандартное логарифмирование невозможно, поскольку логарифм определён только для положительных чисел. Однако можно воспользоваться логарифмом с сохранением знака, чтобы обойти это ограничение. На выходе мы получим более нормализованное распределение, но с сохранением знака.

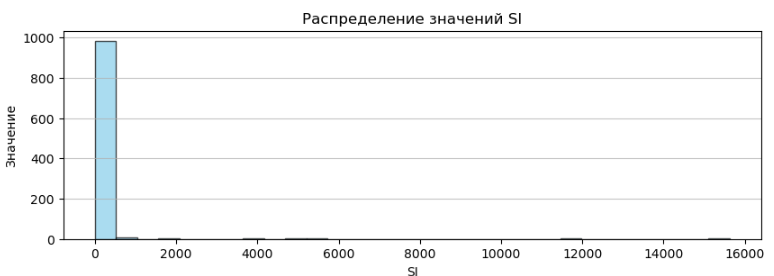
1. **Исследование датасета**
   1. **Изучение распределения целевых переменных**

На рисунке 1 показано распределение целевых переменных до и после логарифмирования.

Рисунок 1 – Распределения целевых показателей

Как видно из представленного выше изображения, исходные данные демонстрируют значительные искажения, причем значительная часть значений сосредоточена вблизи нуля. У переменной "SI" настолько большой правый хвост, что пришлось его искусственно ограничивать, т.к. тот "ломал" график и последний становился неинформативным. Для улучшения наглядности график распределения "SI" будет представлен отдельно ниже.

Логарифмирование целевых переменных способствовало улучшению симметрии распределения и его приближению к нормальному. В дальнейшем, в процессе обучения моделей, мы будем опираться на логарифмированные целевые переменные, так как большинство моделей демонстрируют свои худшие результаты при работе с данными, обладающими выраженной асимметрией. На рисунке 2 показано распределение параметра «SI».

Рисунок 2 – Распределение параметра «SI»

Как видно, "SI" имеет длинный "хвост". При этом, практически все значения находятся в диапазоне от 0 до 1000.

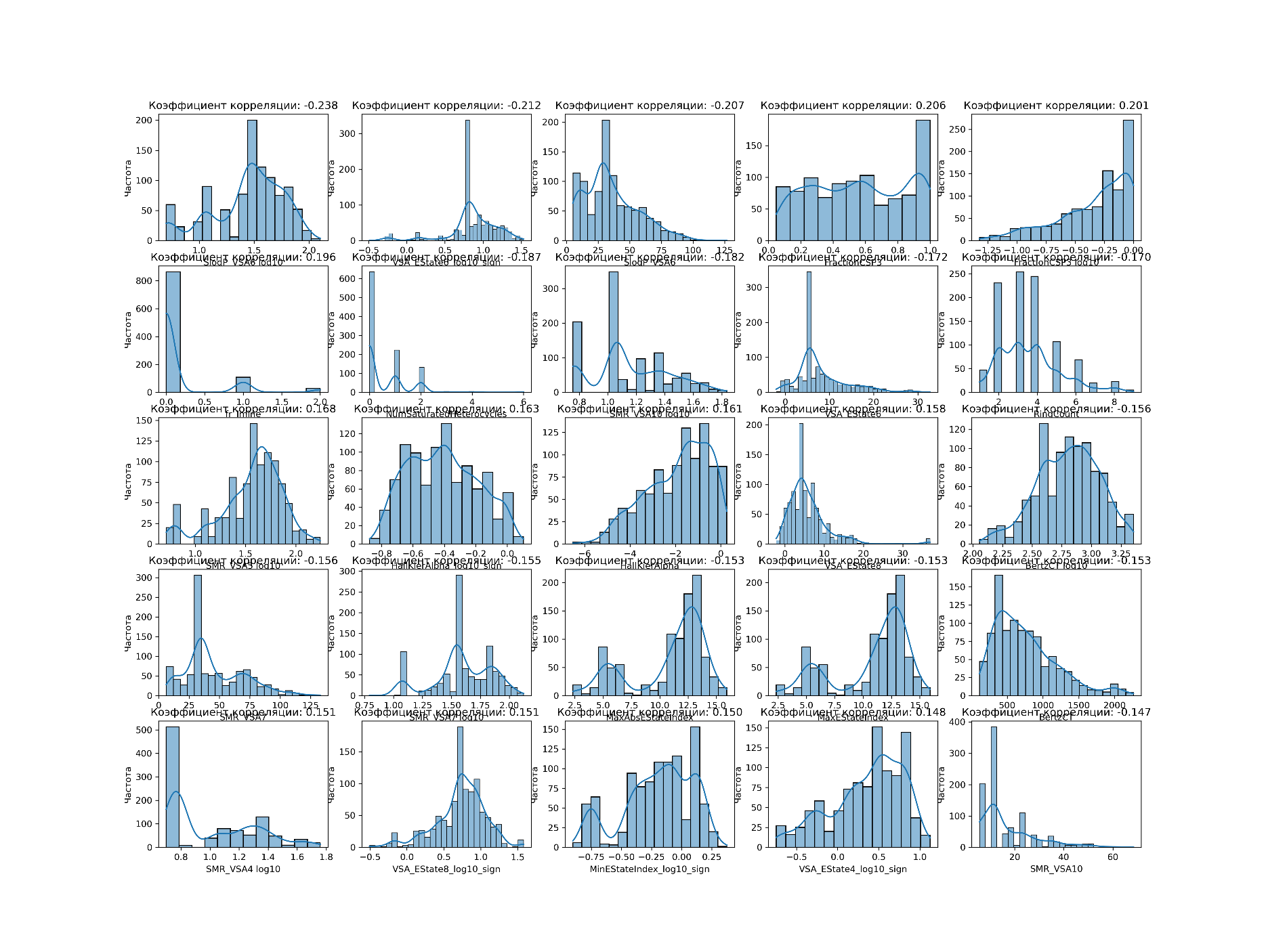
* 1. **Исследование корреляций**

На следующем этапе необходимо провести исследование взаимосвязей между признаками с целью выявления тех, которые оказывают наибольшее влияние на анализируемый параметр. Для этого в библиотеке Pandas будет использован встроенный метод corr(), который позволяет рассчитывать коэффициенты корреляции между парами переменных. Исходя из полученных значений, можно отобрать параметры с наибольшими по абсолютной величине корреляциями. Для лучшего восприятия результаты будут визуализированы.

В качестве примера представлены графики частот для наиболее коррелирующих параметров с "SI log10" на рисунке 3.

Выводы:

* коэффициенты корреляции целевых переменных до логарифмирования оказываются выше, чем после него. Вероятно, это связано с тем, что после логарифмирования снизилась роль выбросов, которые "тянули" ковариацию вверх, т.к. логарифм стремится уменьшить влияние крупных значений и выбросов, выравнивая распределение по масштабу.  
  Также коэффициент корреляции показывает только линейную связь. При логарифмировании меняется характер связи;
* после логарифмирования распределение многих показателей стало ближе к нормальному.

Рисунок 3 – Графики частот для наиболее коррелирующих параметров с "SI log10"

* 1. **Очистка от признаков с наименьшей корреляцией и сохранение датасетов**

С целью повышения эффективности обучения существующий датасет был разбит на три части и очищен от наименее коррелирующих признаков для каждого целевого признака. Поскольку при обучении моделей использовались прологарифмированные целевые переменные, признаки SI, CC50 и IC50 не учитывались — использовались только SI log10, CC50 log10 и IC50 log10.

1. **Обучение моделей. Задачи классификации и регрессии**
   1. **Задачи классификации**

Для каждого целевого признака проводилось отдельно исследование. Для обучения были выбраны четыре модели: логистическая регрессия, дерево решений, случайный лес и градиентный бустинг.

Каждая модель имеет свои преимущества, например логистическая регрессия проста в интерпретации и хорошо работает на линейных данных, что позволяет легко понять влияние отдельных признаков на результат. Дерево решений является интуитивно понятным и наглядным, что делает его удобным для визуализации процессов принятия решений. Случайный лес, в свою очередь, обладает высокой устойчивостью к переобучению благодаря созданию множества деревьев и объединению их результатов, что значительно улучшает точность предсказаний. Градиентный бустинг, имея возможность уменьшать ошибку модели, использует комбинацию слабых моделей и может достичь высокой точности на сложных данных, однако требует более тщательной настройки гиперпараметров.

Каждая из этих моделей была протестирована на наборе данных, чтобы определить наилучший подход для достижения максимальной производительности в задаче.

Далее будет показано исследование на примере признака CC50.

После применения четырех моделей были получены следующие результаты. Они показаны в таблице 1.

Таблица 1 – Результаты четырех моделей для CC50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Результаты | Модели | | | |
| Логистическая регрессия | Дерево решений | Случайный лес | Градиентный бустинг |
| Accuracy | 0,7662 | 0,6965 | 0,7861 | 0,8060 |
| Precision | 0,8085 | 0,7340 | 0,8298 | 0,8438 |
| Recall | 0,7238 | 0,6571 | 0,7429 | 0,7714 |
| F1-score | 0,7638 | 0,6935 | 0,7839 | 0,8060 |

Логистическая регрессия показала устойчивое качество: точность и F1-меру чуть выше 0.76. Модель хорошо сбалансирована, имеет высокое значение precision, но recall немного ниже, что может указывать на некоторое количество пропущенных положительных примеров.

Дерево решений продемонстрировало худшие показатели среди рассмотренных моделей: точность и полнота здесь ниже остальных. Такая модель проще, склонна к переобучению на небольших данных и хуже обобщает информацию.

Случайный лес дал лучшие результаты по сравнению с предыдущими двумя моделями. Эта ансамблевая модель оказалась наиболее сбалансированной после градиентного бустинга, улучшила как точность, так и F1-меру.

Градиентный бустинг стал лидером эксперимента: модель показала наилучшее качество по всем основным метрикам, в том числе по точности и F1-мере. Это говорит о высокой эффективности и умении хорошо выявлять целевой класс на подготовленном датасете.

В целом, по совокупности метрик, градиентный бустинг показал себя лучше остальных моделей.

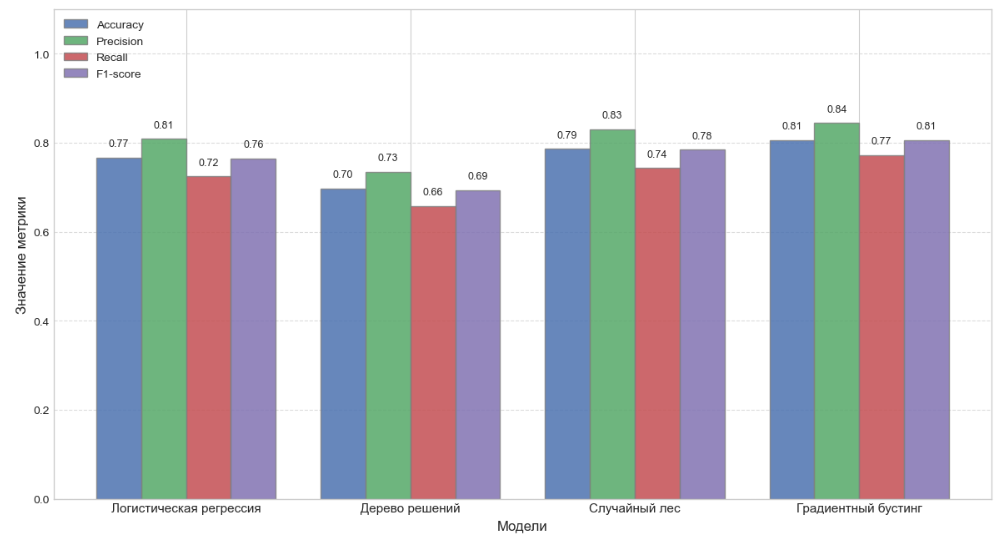
На рисунке 4 представлено сравнение моделей.

Рисунок 4 – Сравнение показателей моделей

**Выводы:**

В целом, все четыре модели неплохо себя показали. Как видно, логистическая регрессия, случайный лес и градиентный бустинг показали себя лучше всего, и их результаты отличаются незначительно. Лидер с небольшим преимуществом — градиентный бустинг. Дерево решений — явный аутсайдер, хоть и отстает совсем немного.

Стоит отметить, что даже простые модели, такие как логистическая регрессия, обеспечивают достаточно высокие значения метрик качества, что указывает на хорошую информативность признаков и адекватный выбор целевого признака. Наилучшее качество на тестовой выборке показал градиентный бустинг, что ожидаемо, учитывая его способность эффективно справляться с небольшими выборками и сложными взаимосвязями между признаками. Случайный лес выступил немного слабее градиентного бустинга, но заметно обошёл дерево решений, что подтверждает необходимость использования ансамблевых методов при работе с подобными данными.

В заключение, градиентный бустинг и случайный лес показали наиболее стабильные и высокие результаты на имеющемся датасете.

* 1. **Задачи регрессии**

Для каждого целевого признака также как и в задачах классификации проводилось отдельное исследование. Для обучения были выбраны пять моделей: полиномиальная регрессия, градиентный бустинг, регрессия лесом решений, линейная регрессия, CatBoostRegressor.

**CatBoostRegressor** чаще всего используется, когда у есть категориальные признаки, которые нужно эффективно обрабатывать. Эта модель прекрасно справляется с высокоразмерными данными и часто демонстрирует высокую точность, минимизируя вероятность переобучения. К сильным сторонам CatBoost можно отнести его автоматическую обработку категориальных переменных и высокую скорость обучения, в то время как среди слабых сторон стоит отметить необходимость выбора параметров для достижения оптимального результата и достаточно большие требования к памяти.

**LinearRegression** применяется в ситуациях, когда данные хорошо поддаются линейному моделированию. Эта модель проста в интерпретации и очень быстрая в обучении. Сильные стороны включают простоту и возможность понимания модели, а среди недостатков – невозможность эффективно работать с нелинейными зависимостями и чувствительность к выбросам.

**DecisionTreeRegressor** полезна для работы с непрерывными и категориальными данными и не требует предварительной обработки (например, нормализации). Сильными сторонами данной модели являются легкость в интерпретации, а также способность захватывать сложные, нелинейные зависимости. Однако дерево решения склонно к переобучению, что является его слабой стороной, особенно при недостатке данных.

**Gradient Boosting** часто используется в ситуациях с большим количеством данных, где важна высокая точность. Эта модель помогает уменьшить ошибку за счет последовательного обучения. К сильным сторонам можно отнести гибкость и возможность обрабатывать сложные зависимости в данных. Тем не менее, среди её слабостей – высокая вычислительная сложность и возможность переобучения.

Результаты работы моделей для целевого параметра CC50 показаны на рисунке 5.

**Выводы:**

В результате анализа для прогнозирования значения целевой переменной CC50 наилучшие результаты продемонстрировали модели полиномиальной регрессии и регрессия с использованием метода леса решений. Тем не менее, существует вероятность, что успех модели леса решений может быть обусловлен переобучением, так как размер датасета составляет всего 1001 строчку и 349 столбцов, что создает условия для избыточной адаптации модели к обучающим данным.

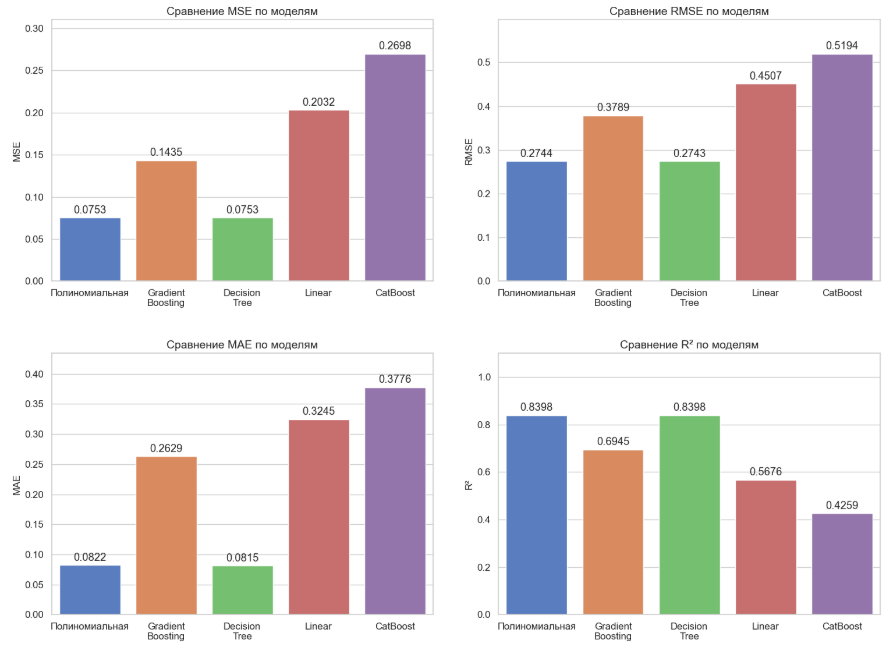
Полиномиальная регрессия показала такие впечатляющие результаты, вероятно, благодаря логарифмизации значений целевой переменной, что способствовало улучшению предсказательной способности этой модели. Полиномиальная регрессия представляет собой расширение линейной регрессии, которое позволяет моделировать сложные зависимости между переменными. Основное преимущество полиномиальной регрессии заключается в ее способности эффективно справляться с не линейными отношениями. В отличие от простой линейной модели, которая описывает зависимость в виде прямой линии, полиномиальная регрессия может использовать полиномы для более точного описания кривых и сложных паттернов в данных.

Рисунок 5 – Сравнение моделей

Одним из ключевых аспектов является возможность подстраивания под характер данных. Благодаря включению полиномиальных терминов, таких как квадратичные и кубические, модель может лучше соответствовать форме распределения данных. Это особенно полезно, когда изучаемая зависимость имеет завитки или изменяющуюся скорость изменения.

В конечном итоге, хотя обе модели продемонстрировали хорошую эффективность, полиномиальная регрессия, похоже, оказалась более устойчивой и предпочтительной в данной задаче.