КУРСОВАЯ РАБОТА

аналитический отчёт

по дисциплине «Классическое машинное обучение»

на тему: «Прогностическая модель для разработки эффективных лекарственных соединений»

Выполнил:

студент группы М24-525

Лапшин Степан Сергеевич

Москва 2025

ВВЕДЕНИЕ

Искусственный интеллект (ИИ) становится все более важным инструментом в разных отраслях, и фармацевтика не является исключением. Разработка новых лекарственных препаратов – это сложный и дорогостоящий процесс, который может занять много лет. Внедрение ИИ в фармацевтику может значительно ускорить этот процесс, снизить затраты и улучшить качество продукции.

Одним из главных преимуществ использования ИИ в фармацевтике является возможность обработки большого количества данных и выявления скрытых закономерностей. Это позволяет ускорить поиск новых лекарственных соединений, а также оптимизировать процессы производства и контроля качества.

Несмотря на это, сам по себе ИИ не может создать лекарство, готовое к выходу на рынок. Инструменты на основе ИИ — одни из множества используемых учеными программ, которые ускоряют процесс, но человеческий контроль остается незаменимым на всех этапах разработки лекарств: проведение экспериментов в физическом мире, проверка данных, интерпретация результатов, проведение клинических испытаний и соблюдение нормативных требований полностью контролируются людьми

ЦЕЛЬ И ЗАДАЧИ

Цель работы - создание инструмента, который повысит эффективность разработки новых лекарственных препаратов за счёт предсказания свойств молекул и оптимизации процесса разработки лекарств.

Задачи данной работы:

- Создание модели для решения задачи регрессии для IC50
- Создание модели для решения задачи регрессии для CC50
- Создание модели для решения задачи регрессии для SI
- Создание модели для решения задачи классификации IC50 (превышение медианного значения)
- Создание модели для решения задачи классификации CC50 (превышение медианного значения)
- Создание модели для решения задачи классификации SI (превышение медианного значения)
- Создание модели для решения задачи классификации SI (превышение 8)

Для решения данных задач требуется проанализировать представленные данные (EDA), оформить все это в отдельном файле блокнота Jupyter Notebook.

Также требуется в отдельных файлах блокнота Jupyter Notebook предоставить код построения каждой из модели и подбора гиперпараметров.

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ - Exploratory Data Analysis (EDA)

Знакомство с данными

В самом начале после считывания данных был сразу же удален неинформативный признак 'Unnamed: 0' (индексы строк).

Был изучены признаки, что они обозначают, статистика каждого признака, а также их тип данных. Краткое описание признаков представлено в табл. 1.

Таблица 1 - Описание признаков

Категория	Признак	Краткое описание
Электротопологические дескрипторы	MaxAbsEStateIndex, MinAbsEStateIndex	Максимальное и минимальное абсолютное значение электротопологического индекса
	MaxEStateIndex, MinEStateIndex	Максимальное и минимальное значения E-State без учёта знака
	EState_VSA1–11	Дескрипторы, связывающие E-State с поверхности (VSA)
	VSA_EState1–9	Дескрипторы, связывающие E-State с поверхности (VSA) с группировкой по диапазонам значений
Физико-химические свойства	qed	Quantitative Estimate of Drug-likeness (оценка "лекарственности" молекулы)
	SPS	Synthetic Accessibility Score (оценка сложности синтеза)
	MolWt	Молекулярная масса

	ExactMolWt	Точная молекулярная масса
	HeavyAtomMolWt	Масса тяжёлых атомов
	MolLogP	Логарифм коэффициента распределения октанол-вода
	MolMR	Молярная рефракция
	TPSA	Топологическая полярная поверхностная площадь
	FractionCSP3	Доля гибридизированных атомов углерода
Электронные и зарядовые характеристики	NumValenceElectrons	Число валентных электронов
	NumRadicalElectrons	Число неспаренных электронов
	Max/MinPartialCharge	Максимальный и минимальный парциальные заряды
	Max/MinAbsPartialCharge	Абсолютные значения парциальных зарядов
	PEOE_VSA1–14	Дескрипторы, связывающие парциальные заряды (PEOE) с площадью поверхности
	BCUT2D_	Дескрипторы, основанные на матрицах связности и зарядов
Топологические дескрипторы	BalabanJ	Индекс Балабана
	BertzCT	Индекс сложности Бертца
	Chi0–4n/v	Индексы Ки (хи-индексы) разных порядков
	Kappa1–3	Каппа-индексы (форма молекулы)
	HallKierAlpha	Альфа-модификация индекса Кьера

	Ipc	Информационная ёмкость молекулы
Поверхностные и объёмные дескрипторы	LabuteASA	Приблизительная площадь поверхности
	SlogP_VSA1–12	Дескрипторы, связывающие SlogP (липофильность) с площадью поверхности
	SMR_VSA1–10	Дескрипторы, связывающие молярную рефракцию (SMR) с площадью поверхности
Функциональные группы и фрагменты	fr_	Префикс для фрагментов
	NHOHCount, NOCount	Число NH/OH-групп и NO-групп
	NumHAcceptors, **NumHDonors**	Число акцепторов и доноров водородных связей
	NumRotatableBonds	Число вращающихся связей
	RingCount	Число циклов в молекуле
	NumAromaticRings	Число ароматических циклов

Столбцы в качестве целевых переменных:

- IC50 (концентрация полумаксимального ингибирования) — показатель эффективности лиганда при ингибирующем биохимическом или биологическом взаимодействии;
- CC50 - Концентрация соединения, которая вызывает гибель 50% нормальных клеток;
- SI — это индекс селективности, который вычисляется как отношение CC50 к IC50. Высокий показатель SI означает, что вещество убивает вирус, а не клетки организма, то есть обладает селективностью в отношении вируса, а не организма.

Работа с пропусками, дубликатами и пустыми признаками

Данные были проверены на наличие пропусков. Их оказалось немного, поэтому было принято решение просто удалить строки с пропусками.

Также было проверено наличие дубликатов, удалены найденные строчки данных.

При анализе были найдены столбцы с одним значением - 0. Данные признаки не несут никакой информации и были удалены.

Анализ и обработка выбросов

Для проверки данных на выбросы, были построены коробчатые диаграммы (рис.1 и рис.2)

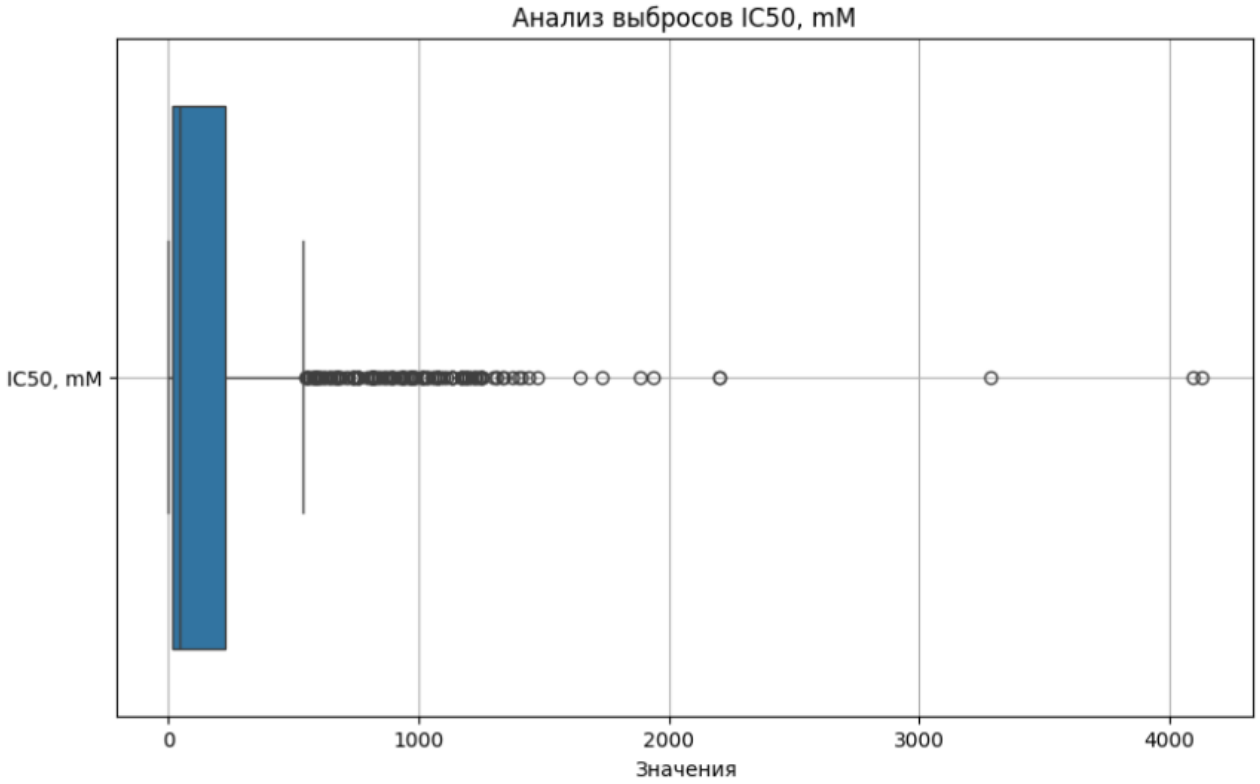

Рисунок 1 - Анализ выбросов IC50

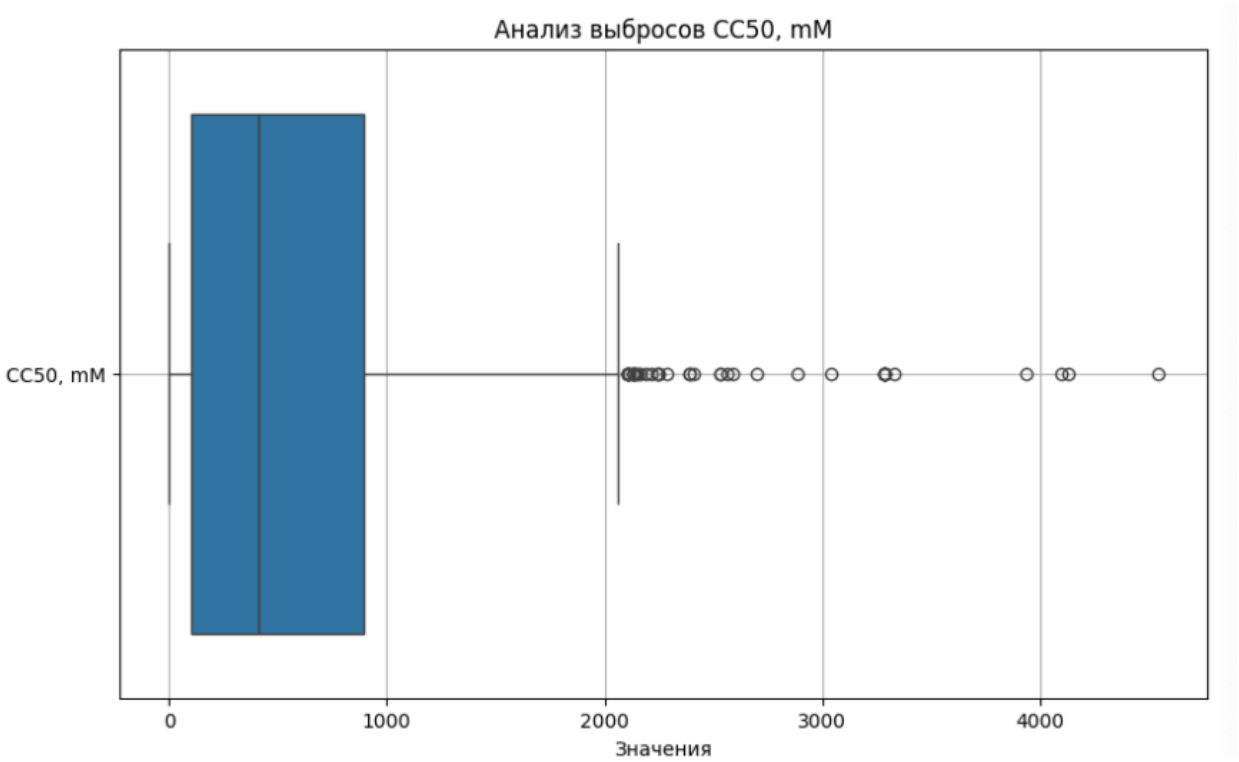

Рисунок 2 - Анализ выбросов CC50

Если считать выбросы исходя из коробчатых диаграмм, то можем потерять полезные данные. Поэтому была изучена дополнительная литература для анализа этих данных с практической точки зрения.

Из дополнительных источников была получена информация, что IC50 может превышать 1000 mM, но при таком повышении вещество становится практически неактивным в биологическом контексте. Поэтому значения свыше 1000 mM в нашем случае считаются выбросами.

В свою очередь CC50 может быть абсолютно любым, нет информации о выбросах.

Значения SI> 1000 скорее всего тоже будут являться выбросами, так как такие значения встречаются крайне редко, так как требуют почти нулевой токсичности.

Проверка распределения целевых переменных

Проверка распределения целевой переменной является важным этапом предварительного анализа данных, поскольку от характера распределения

зависят выбор модели, интерпретация результатов и качество прогнозирования.

Если распределение имеет значительную асимметрию, выбросы или мультимодальность, это может привести к некорректной работе алгоритмов, особенно тех, что чувствительны к предположениям о нормальности данных, таких как линейная регрессия.

Кроме того, анализ распределения помогает выявить проблемы в данных, такие как аномалии или ошибки измерения.

Таким образом, проверка распределения целевой переменной — это не просто формальность, а ключевой шаг, влияющий на весь дальнейший процесс анализа и прогнозирования.

Распределения целевых переменных представлено на рис.3-5.

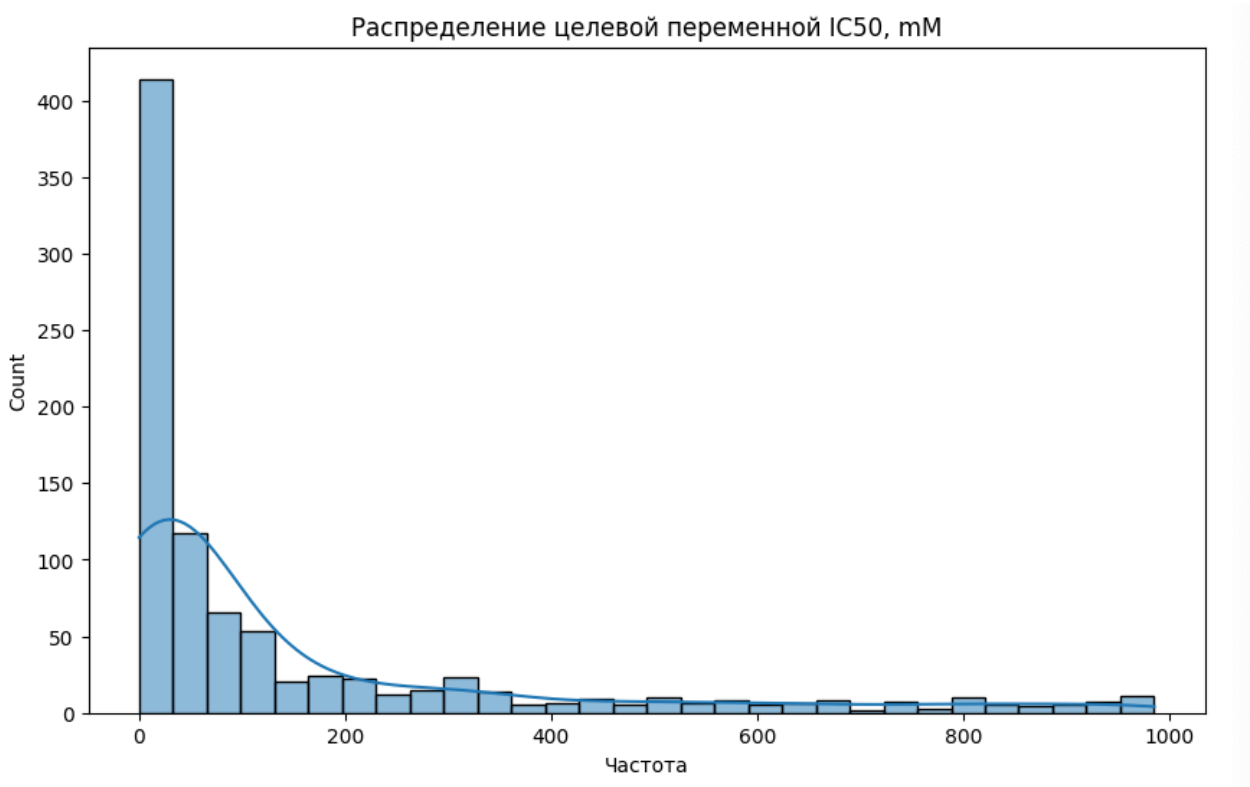

Рисунок 3 - Распределение IC50

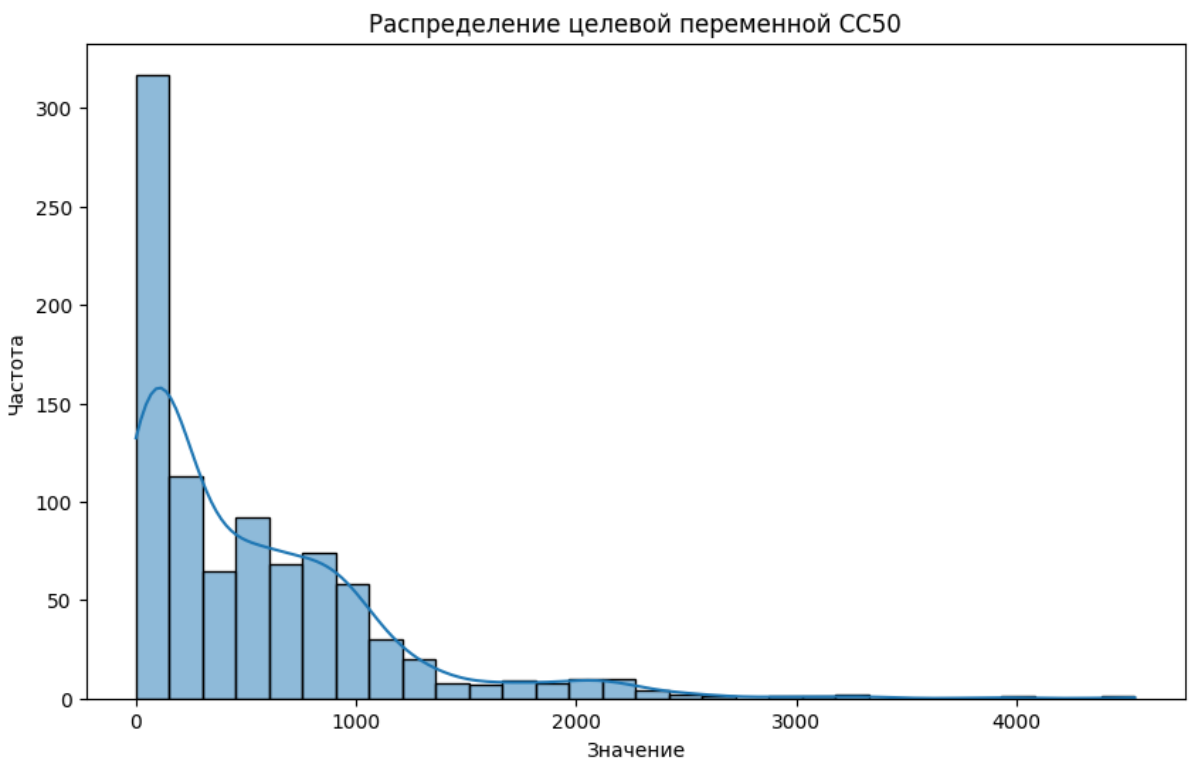

Рисунок 4 - Распределение CC50

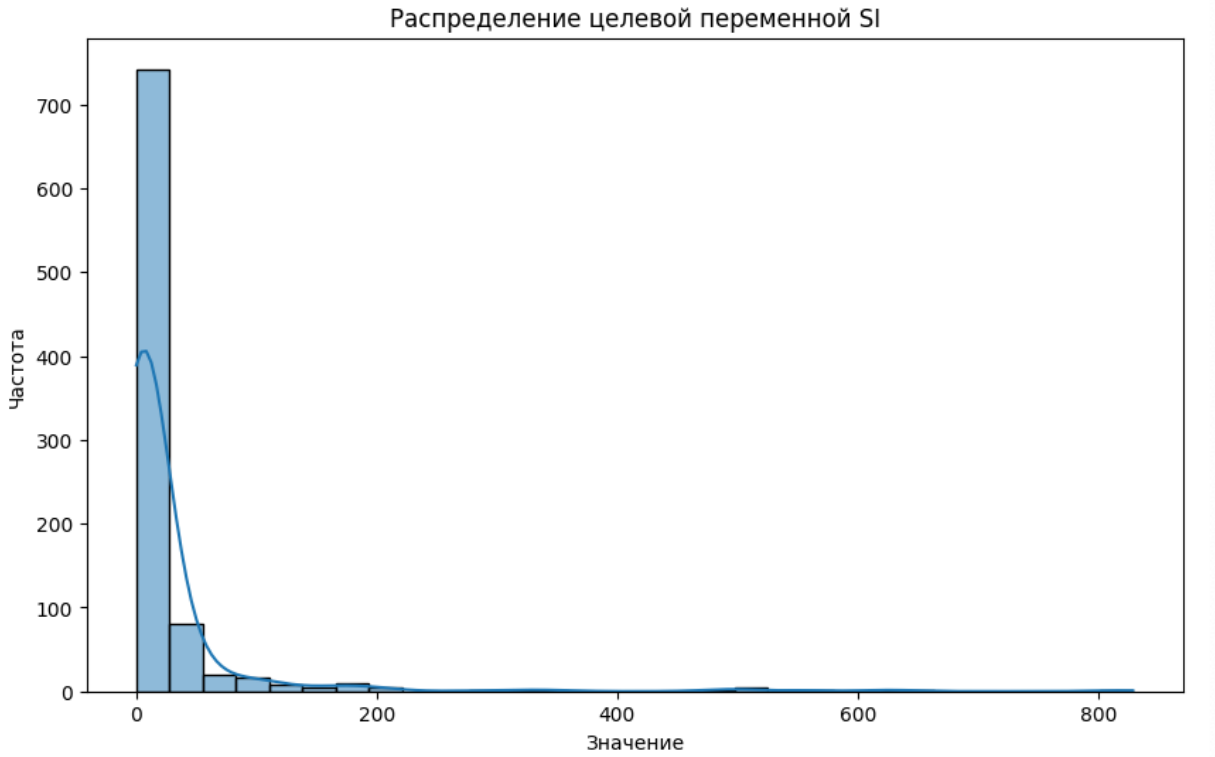

Рисунок 5 - Распределение SI

Для устранение ненормального распределение было принято решение провести логарифмирование. Графики распределений после логарифмирований представлены на рис.6-7.

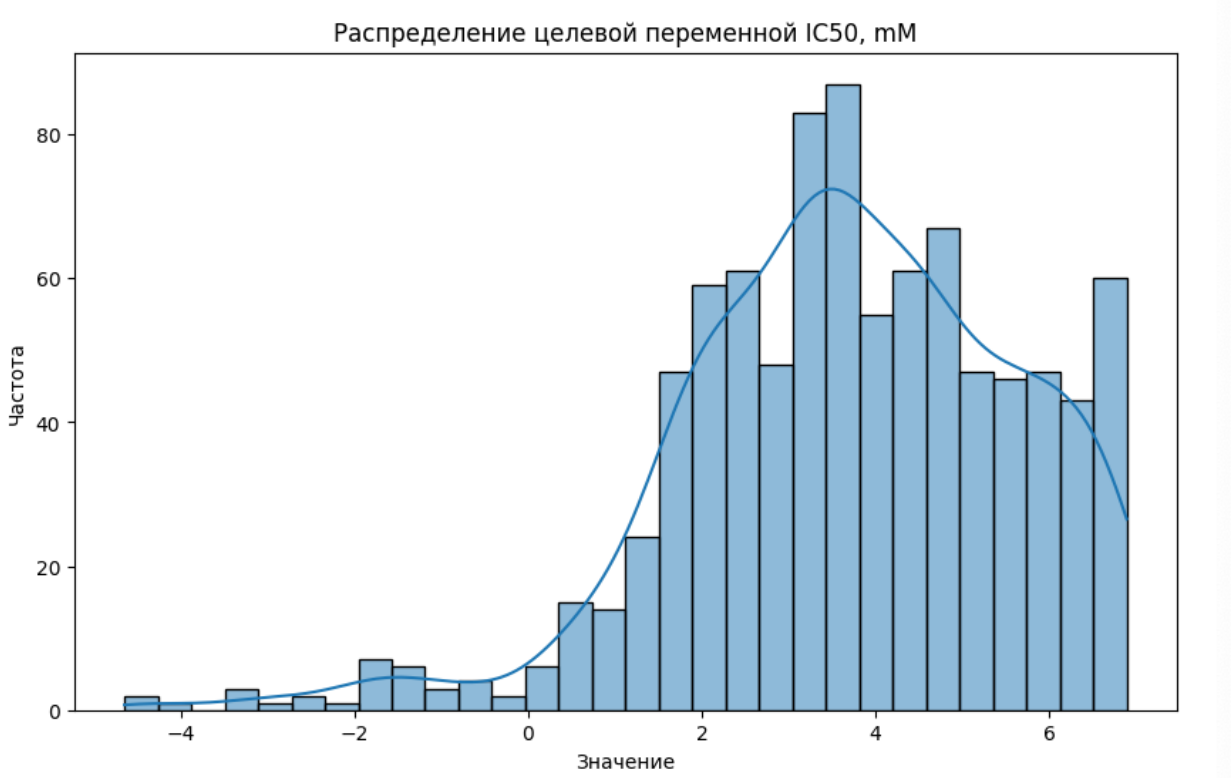

Рисунок 6 - Распределение IC50

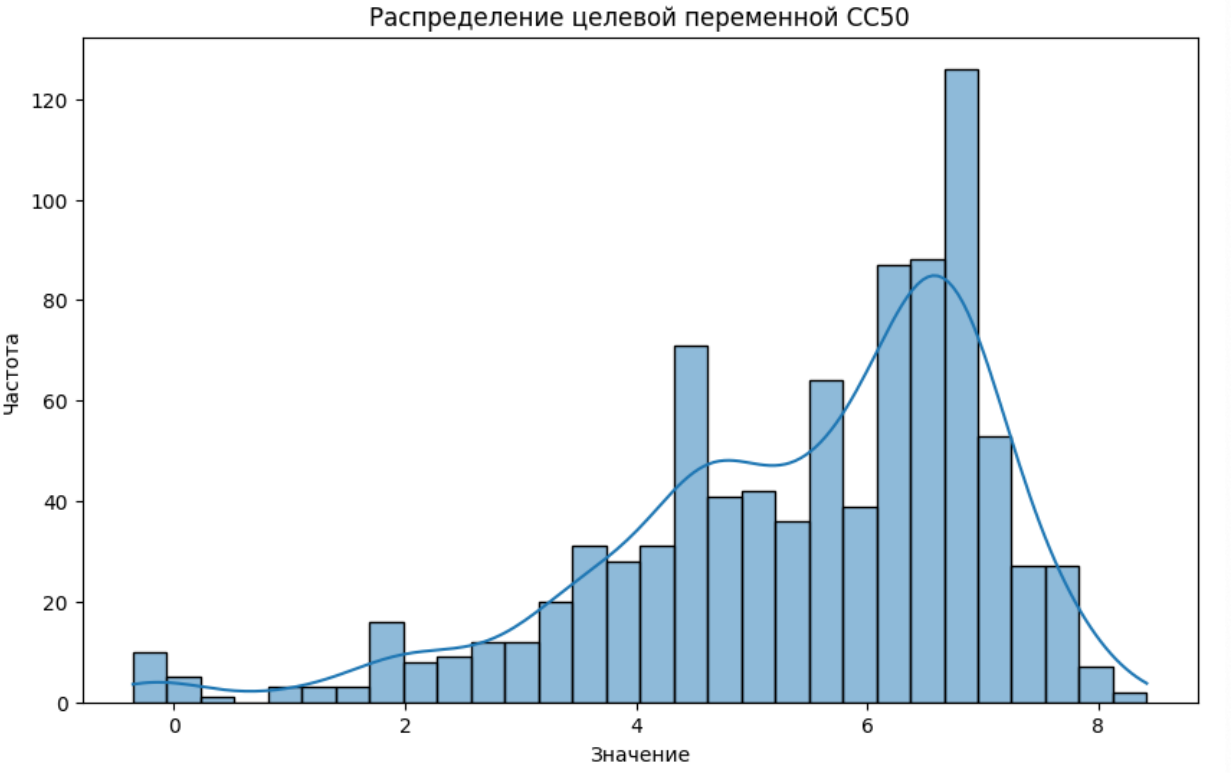

Рисунок 7 - Распределение CC50

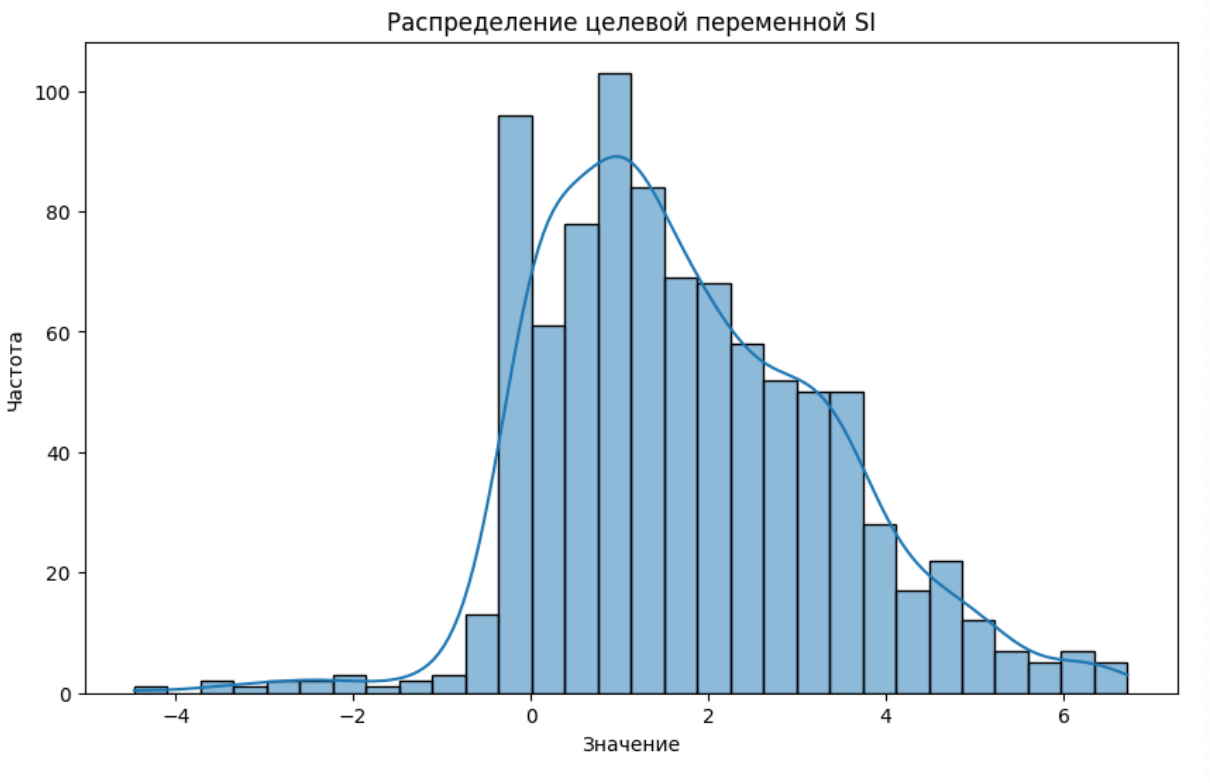

Рисунок 8 - Распределение SI

Модернизация признаков

- **MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex;**

Созданы новые признаки - RangeEState (разброс электронных состояний) и MeanEState (средняя абсолютная величина). После преобразований были удалены используемые признаки.

- **MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge;**

Созданы новые признаки - RangePartical (разброс зарядов) и MeanPartical (средняя абсолютная величина). После преобразований были удалены используемые признаки.

- **BCUT2D;**

Заменен ряд признаков верхней и нижней границы на один (разницу этих признаков). После преобразований были удалены используемые признаки.

- **Chi дескрипторы;**

Проведен анализ матрицы корреляций между Chi дескрипторами, оставлены только те, которые между собой не коррелируют. После преобразований были удалены используемые признаки.

- **PEOE_VSA, SMR_VSA, SlogP_VSA, EState_VSA;**

Объединены в общий признак дескрипторы, связывающие парциальные заряды (PEOE) с площадью поверхности (VSA). После преобразований были удалены используемые признаки.

- **fr_ счётчики определённых функциональных групп.**

Счётчики определённых функциональных групп были сгруппированы в отдельные группы:

1. Кислотные
2. Азотосодержащие
3. Гидроксильные и эфирные
4. Карбонильные
5. Амиды и родственные
6. Галогены
7. Серосодержащие
8. Ароматические системы
9. Гетероциклы
10. Нитро- и азогруппы
11. Ненасыщенные фрагменты

Все что не удалось отнести не к одной группе остаются отдельными признаками. После преобразований были удалены используемые признаки.

По завершению всех преобразований преобразованные данные были сохранены в файл ***data_eda.csv*** для дальнейшего использования моделями.

РАБОТА С МОДЕЛЯМИ

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.

Для выбора всех 7 моделей вначале была произведена оценка зависимости признаков и конкретной целевой переменной, после этого были выбраны самые значимые признаки.

После данные были разделены на тренировочные и тестовые для дальнейшей оценки качества модели на независимых данных.

Был применен метод стандартизации для приведения данных к единому масштабу, чтобы улучишь работу алгоритмов. Обучение происходило на тренировочной выборке, а для тестовой выборки только трансформирование.

Для некоторых моделей также был применен метода PCA (метод главных компонент) для ускорения обучения алгоритмов, где это было критично в связи с имеющими вычислительными мощностями.

Для выбора моделей регрессий тестировались:

1. Модель линейной регрессии

Простая и интерпретируемая модель, предсказывающая целевую переменную как линейную комбинацию признаков

2. Модель KNeighborsRegressor

Предсказывает значение на основе среднего значений k ближайших объектов в обучающей выборке.

3. Модель SVR

Использует метод опорных векторов для регрессии, пытаясь уместить ошибки в заданную границу

4. Модель GradientBoostingRegressor

Ансамблевая модель, строит последовательность деревьев, где каждое новое дерево исправляет ошибки предыдущего

5. Модель RandomForestRegressor

Ансамбль решающих деревьев, где каждое дерево обучается на случайном подмножестве данных и признаков, а итоговый прогноз — среднее предсказаний всех деревьев

Для каждой из модели был произведен подбор гиперпараметров, в качестве инструмента использовался GridSearchCV, инструмент в библиотеке scikit-learn для автоматического подбора оптимальных гиперпараметров модели машинного обучения. Результат выбора модели для каждой из задач регрессии представлен в табл.2.

Таблица 2 - Результат выбора моделей

Целевая переменная	**Выбранная модель**	**Получившиеся метрики**
IC50	RandomForestRegressor	Mean Absolute Error: 1.172 Mean Squared Error: 2.002 Root Mean Squared Error: 1.415 Коэффициент детерминации: 0.484
CC50	SVR	Mean Absolute Error: 0.810 Mean Squared Error: 1.240 Root Mean Squared Error: 1.113 Коэффициент детерминации: 0.540
SI	GradientBoostingRegressor	Mean Absolute Error: 0.920 Mean Squared Error: 1.396 Root Mean Squared Error: 1.182 Коэффициент детерминации: 0.527

Модель для IC50 демонстрирует умеренную предсказательную способность ($R^2 \approx 0.48$), но ошибки (MAE, RMSE) указывают на значительный разброс предсказаний. Возможно, модель недостаточно точно улавливает сложные зависимости в данных и требуется доработка.

Модель для CC50 показала лучшую производительность по сравнению с Random Forest для IC50 (более высокий R^2 и меньшие ошибки). Однако $R^2 = 0.54$ всё ещё указывает на ограниченную объясняющую способность модели.

Модель для SI GradientBoosting демонстрирует сопоставимую с SVR точность ($R^2 \approx 0.53$), но с чуть более высокими ошибками.

Все модели имеют R^2 в диапазоне 0.48–0.54, что указывает на умеренное качество предсказаний. Возможно, стоит рассмотреть:

1. Улучшение feature engineering (добавление новых признаков)
2. Использование ансамблевых методов или нейронных сетей
3. Увеличение объёма данных или балансировку признаков
4. Протестировать другие алгоритмы или попытаться лучше подобрать гиперпараметры

Результат выбора модели для каждой из задач регрессии представлен в табл.3.

Таблица 3- Результат выбора моделей

Целевая переменная	**Выбранная модель**	**Получившиеся метрики**
Class_IC50	CatBoost	Accuracy: 0.768 Precision: 0.779 Recall: 0.744 F1-score: 0.761
Class_CC50	CatBoost	Accuracy: 0.812 Precision: 0.798 Recall: 0.833 F1-score: 0.815
Class_SI_1	LogisticRegression	Accuracy: 0.790 Precision: 0.796 Recall: 0.777 F1-score: 0.787
Class_SI_2	CatBoost	Accuracy: 0.796 Precision: 0.813 Recall: 0.582 F1-score: 0.678

Модель Class_IC50 имеет хорошую сбалансированность между точность и полнотой.

Модель Class_CC50 демонстрирует высокую предсказательную способность, особенно при Recall = 0.833, что важно для минимизации ложноотрицательных прогнозов.

Модель Class_SI_1 сопоставимы с CatBoost других моделей, что может говорить о линейной природе зависимости.

Модель Class_SI_2 несмотря на приемлемую точность (Precision = 0.813), низкий Recall (0.582) указывает на проблему с обнаружением положительного класса. Возможно, требуется дополнительная балансировка данных или настройка порога классификации

CatBoost проявил себя как наиболее универсальный алгоритм, обеспечивающий высокое качество предсказаний для большинства задач.

Для улучшения качества моделей можно рассмотреть несколько направлений:

1. Оптимизировать гиперпараметры моделей
2. Добавить полиномиальные признаки
3. Использовать альтернативные модели

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы был разработан инструмент для повышения эффективности разработки новых лекарственных препаратов.

Основной задачей являлось создание моделей машинного обучения, способной предсказывать свойства молекул и оптимизировать процесс поиска потенциальных лекарственных соединений.

Были выполнены следующие этапы:

1. **Предобработка данных** – проведён анализ признаков, устранены пропуски и дубликаты, а также удалены выбросы, что позволило улучшить качество данных.
2. **Модернизация признаков** – выполнено преобразование и создание новых признаков для повышения информативности данных.
3. **Выбор и обучение моделей** – протестированы различные алгоритмы машинного обучения с подбором гиперпараметров для каждой модели.

В результате проведённого анализа удалось создать инструмент, способный эффективно предсказывать свойства молекул, что может значительно ускорить процесс разработки новых лекарственных препаратов.

Дальнейшее развитие проекта может включать в себя:

- использование более сложных архитектур нейронных сетей;
- увеличение объёма данных для обучения;
- оптимизировать гиперпараметры моделей;
- улучшить feature engineering (добавить или изменить новые признаки)

Таким образом, в рамках курсовой работы удалось достичь поставленных целей. Результаты показали свою эффективность, но остаются перспективы для дальнейшего совершенствования.