Аналитический отчет к курсовой работе

1. Цели и задачи курсовой работы.

На основании предоставленных данных о химических соединениях с указанием их эффективности против вируса гриппа необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Для этого требуется выполнить следующие шаги:

1. Проанализировать текущие параметры с использованием различных методов.

2. Создать несколько максимально эффективных моделей для решения следующих задач:

- Регрессия для IC50
- Регрессия для CC50
- Регрессия для SI
- Классификация: превышает ли значение IC50 медианное значение выборки
- Классификация: превышает ли значение CC50 медианное значение выборки
- Классификация: превышает ли значение SI медианное значение выборки
- Классификация: превышает ли значение SI значение 8.

3. Сравнить между собой полученные модели и их результаты, выполнить анализ, обосновать выбор наиболее качественных решений.

2. Исследование данных.

Данные представляют собой информацию о 1001 химическом соединении (строки датасета). Информация о химических соединениях включает:

- 3 колонки с целевыми переменными IC50, CC50, SI
- 210 колонок с дескрипторами (числовыми характеристиками химических соединений).

Целевые переменные **IC50, CC50, SI** показывают эффективность химических соединений:

1. IC50 – минимальная концентрация соединения/экстракта, которая подавляет размножение 50% патогенов или вирусов (мера активности соединения).

2. CC50 – это концентрация соединения, которая приводит к гибели 50% нормальных клеток хозяина (мера токсичности).

3. SI (Selectivity Index) = CC50 / IC50. Чем выше показатель SI, тем лучше. Желательно, чтобы концентрация IC50 была ниже концентрации CC50, а CC50 был как можно выше. Это означает, что вы убиваете патоген до того, как он убьет хозяина.

В датасете содержатся следующие **категории дескрипторов**:

Наименование дескриптора	Описание дескриптора
1. Электронные дескрипторы	
MaxAbsEStateIndex, MaxEStateIndex, MinAbsEStateIndex, MinEStateIndex	количественно определяют электронные характеристики отдельных атомов в молекуле
MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge	показывают значения частичного заряда молекулы
2. Молекулярные дескрипторы	описывают основные свойства молекул, такие как липофильность, молярная рефракция, молекулярный вес, водородные связи, молекулярные объемы и площади поверхностей
qed	оценивает пригодность для использования в качестве лекарства. Чем выше значение, тем лучше молекула подходит для лекарственной химии
SPS (Spacial Score)	отвечает за оценку пространственной сложности молекулы
MolWt, HeavyAtomMolWt, ExactMolWt	показывают молекулярную массу
NumValenceElectrons, NumRadicalElectrons	показывают количество валентных/радикальных электронов в молекуле
TPSA	оценивает площадь поверхности молекулы, участвующей в полярных взаимодействиях, что влияет на ее растворимость и биологическую активность
FractionCSP3	показывают долю sp^3-гибридизованных атомов углерода в соединении
HeavyAtomCount	показывает количество тяжелых атомов в молекуле, исключая атомы водорода
NHOHCount, NOCount	показывают количество азотных и кислородных групп в молекуле
MolLogP (молярный логарифм коэффициента распределения между октанолом и водой)	отражает липофильность (гидрофобность) молекулы
MolMR (молярная рефрактивность)	характеризует размер молекулы и ее поляризуемость
3. Дескрипторы плотности	
FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3	количественная оценка частоты встречаемости определенных подструктур в молекуле на локальном уровне с учетом их непосредственного окружения. Более высокие значения плотности указывают на более высокую плотность уникальных подструктур в молекуле, а более низкие значения указывают на меньшее количество уникальных подструктур или более равномерное распределение подструктур
4. BCUT2D-дескрипторы	
BCUT2D_MWHI, BCUT2D_MWLOW, …	показывают форму, размер, заряд и другие параметры молекулы. Эти дескрипторы используются для описания молекулярной структуры и анализа ее свойств, помогают выделить молекулярные фрагменты, которые повышают или понижают активность соединений
5. Топологические дескрипторы	
AvgIpc, Ipc, BalabanJ, BertzCT	показывают, насколько сложной является структура молекулы, учитывая разветвленность и информационные взаимодействия атомов
Chi0, Chi1, Chi2n, Chi3n, …	показывают, насколько сильно связаны атомы в молекуле
HallKierAlpha	описывает гибкость или жесткость атомов в молекуле

Наименование дескриптора	Описание дескриптора
Kappa1, Kappa2, Kappa3	описывают форму молекулы на основе распределения длин связей и углов
LabuteASA	оценивает доступную для растворителя площадь поверхности молекулы, которая влияет на ее растворимость и проницаемость
6. VSA-дескрипторы	
PEOE_VSA1, …, PEOE_VSA14	описывает вклад атомов в площадь поверхности с учетом их частичного заряда
SMR_VSA1, …, SMR_VSA10	описывает вклад атомов в молярную рефрактивность
SlogP_VSA1, …, SlogP_VSA12	отражает вклад атомов в коэффициент разделения
EState_VSA1, …, EState_VSA11	объединяет электронную и топологическую информацию в пределах определенной площади поверхности
VSA_EState1, …, VSA_EState10	суммирует индексы электротопологического состояния (EState) по интервалам вклада в VSA
7. Структурные количественные дескрипторы	
NumAliphaticCarbocycles, NumAliphaticHeterocycles, NumAromaticCarbocycles, NumAromaticHeterocycles, NumSaturatedCarbocycles, NumSaturatedHeterocycles	показывают количество алифатических, ароматических, насыщенных карбо- и гетероциклов в молекуле
NumAliphaticRings, NumAromaticRings, NumSaturatedRings	показывают количество алифатических, ароматических, насыщенных колец в молекуле
NumHAcceptors	количество групп, принимающих водородную связь
NumHDonors	количество групп, дающих водородную связь
NumHeteroatoms	количество гетероатомов (неуглеродных атомов)
NumRotatableBonds	количество вращающихся связей
RingCount	показывают общее количество колец в структуре молекулы
8. Фрагментные дескрипторы	содержат информацию о присутствии или отсутствии определенных структурных элементов в химическом соединении
fr_Al_OH, fr_phenol	фенольная группа
fr_NH2, fr_amine, fr_aniline	амины
fr_azide, fr_azo, fr_diazo	азо-соединения
fr_halogen, fr_alkyl_halide	галогены, галогеналкилы
fr_barbitur	барбитураты
fr_nitro, fr_nitro_arom	нитро-группы
fr_lactone, fr_lactam	лактон, лактам
fr_benzene, fr_pyridine, fr_thiazole, fr_furan	кольца

Выводы по исследованию и анализу данных:

- **При построении моделей с целевой переменной IC50 или CC50:** переменную SI необходимо удалить, так как SI = CC50 / IC50 и сохранение SI в данных приведет к утечке информации.
- Изучение информации о предикторах, которые включены в датасет, показывает, что все они используются для предсказания активности и токсичности химических соединений. Значит, на текущем этапе исследования нет оснований исключать какие-либо из признаков из анализа.

3. Подготовка данных и разведочный анализ (EDA).

1. Проверка наличия пропусков в данных.

Значения в колонках имеют числовой тип данных. Пропуски содержатся в 12 колонках. В каждой из них по 3 пропуска, что составляет ~ 0.3% от количества значений в колонке.

Значения дескрипторов в соседних строках указывают на отсутствие закономерностей в размещении объектов в датасете, поэтому значения из соседних строк нецелесообразно использовать для заполнения пропусков.

С помощью метода describe определены статистические показатели по каждой колонке с пропусками. По некоторым BCUT2D-дескрипторам наблюдается большая дисперсия в данных, поэтому заполнение пропусков средним значением может исказить результаты анализа.

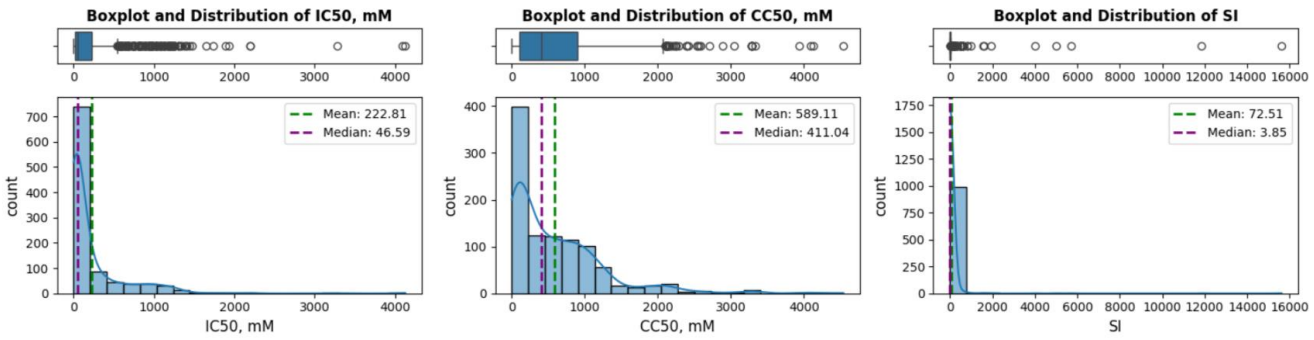
Пропуски содержатся в колонках с электронными дескрипторами и BCUT2D-дескрипторами. Некоторые дескрипторы RDKit возвращают значение NaN, когда сталкиваются с типами атомов, которые не параметризованы. Будем считать, что пропуск в данных означает равенство значения нолю, и заполним пропуски нолями.

2. Проверка наличия пустых колонок.

Анализ показал, что в данных присутствуют 18 пустых колонок, это в основном фрагментные дескрипторы. Исключим эти колонки из дальнейшего анализа. Также исключим колонку Unnamed. В результате количество признаков сократилось до 195.

3. Исследование распределения целевых переменных (histograms, boxplots).

Построим boxplot и гистограмму для каждой целевой переменной для исследования их на предмет выбросов и с целью анализа распределения.



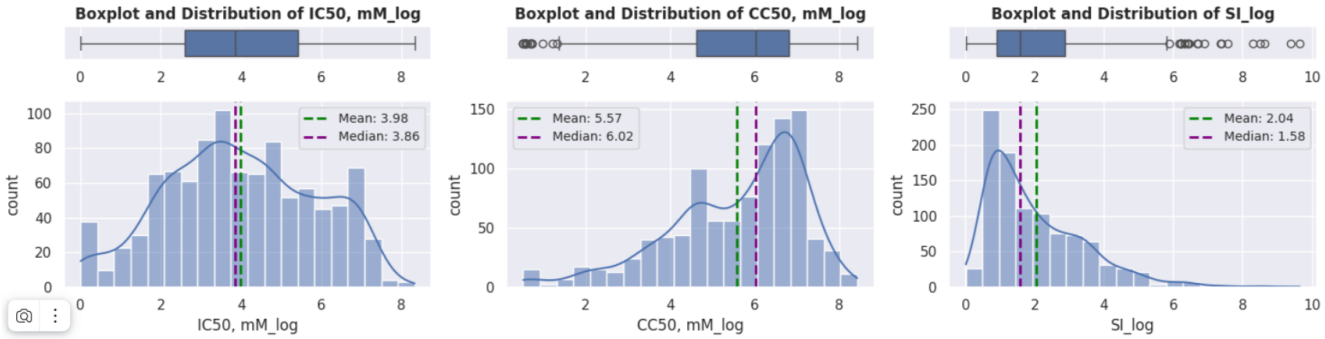
Выводы по диаграммам:

- **Выбросы:** переменные IC50 и CC50 содержат большое количество выбросов. Переменная SI, которая рассчитывается на основании переменных IC50 и CC50, также имеет выбросы, хотя их количество существенно меньше. Некоторые модели машинного обучения (например, линейная регрессия) чувствительны к выбросам. Выбросы в данных могут исказить и ввести в заблуждение процесс обучения моделей, что приводит к увеличению времени обучения, снижению точности моделей и, в конечном итоге, к снижению результатов.

- **Распределение:** распределение всех целевых переменных сильно смещено влево, то есть основная часть данных находится в диапазоне низких значений, с длинным "хвостом" высоких значений.

С целью повышения устойчивости моделей к выбросам и тем самым повышения качества моделей логарифмируем целевые переменные с помощью log(x+1).

Посмотрим, как изменилось распределение целевых переменных IC50, CC50 и SI после логарифмирования.



Выводы по диаграммам:

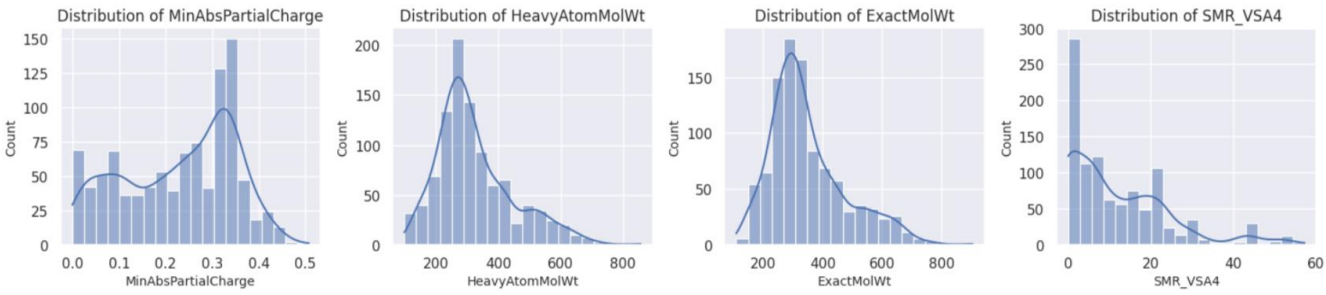
после логарифмирования распределение целевых переменных стало ближе к нормальному. Особенно хорошие результаты логарифмирование показало для переменной IC50.

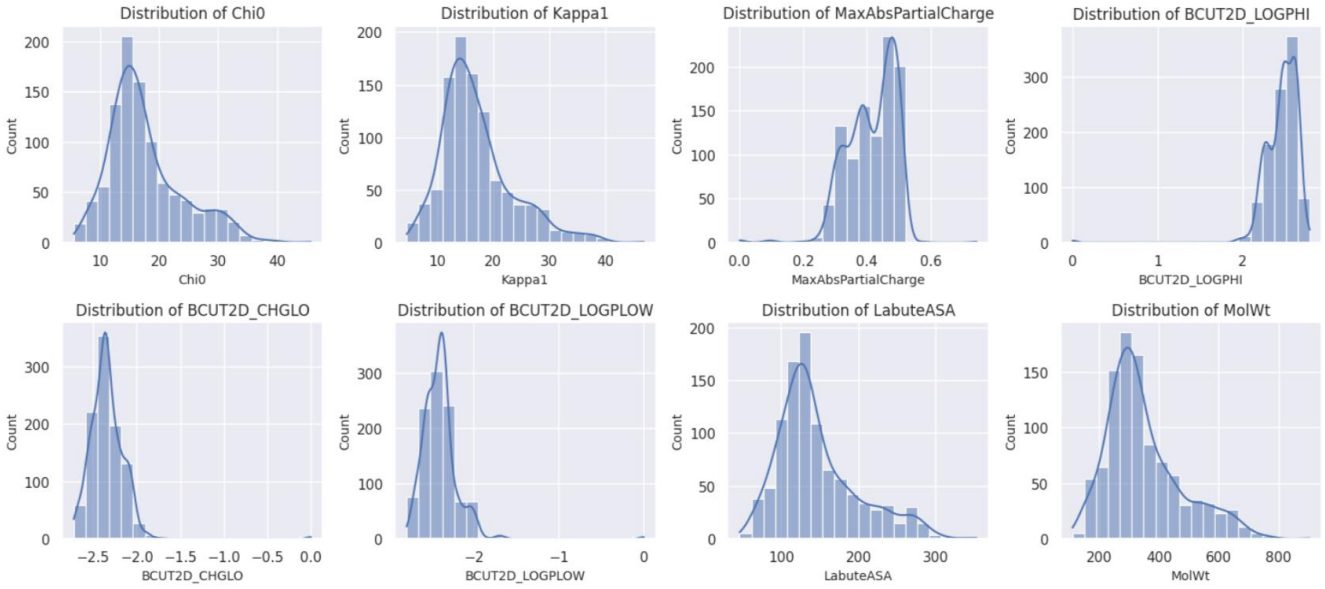
4. Исследование распределения значимых признаков.

С помощью функции SelectKBest определены 5 наиболее значимых признаков для каждой целевой переменной:

	IC50, mM_log	CC50, mM_log	SI_log
0	MolWt	MolWt	MaxAbsPartialCharge
1	ExactMolWt	HeavyAtomMolWt	BCUT2D_CHGLO
2	MinAbsPartialCharge	ExactMolWt	BCUT2D_LOGPHI
3	Chi0	Chi0	BCUT2D_LOGPLOW
4	Kappa1	LabuteASA	SMR_VSA4

Из этих признаков 12 уникальных, посмотрим их распределение:

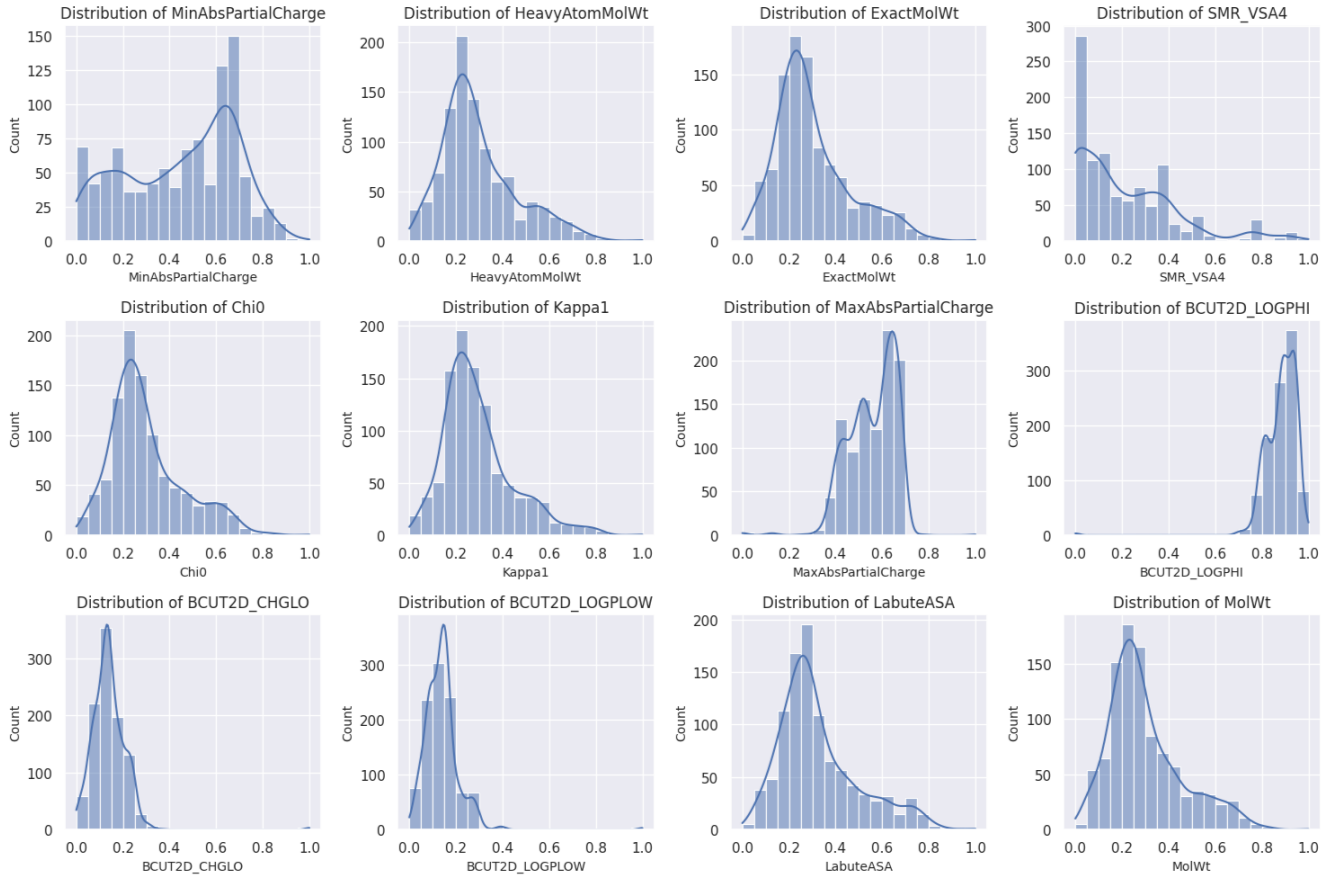




Выводы по диаграммам:

некоторые из признаков распределены несимметрично, наблюдаются выбросы. Кроме того, признаки имеют разный масштаб. Для улучшения качества моделей проведем нормализацию данных с помощью MinMaxScaler.

Посмотрим, как изменилось распределение признаков после нормализации:



5. Корреляционный анализ и отбор признаков.

Корреляция признаков с целевыми переменными: анализ показал, что признаки имеют слабую линейную зависимость с целевыми переменными, всего 4-8 признаков имеют коэффициент корреляции от 0.2 до 0.3.

Корреляция признаков между собой: матрица корреляции показывает сильную корреляцию между собой признаков Chi, Kappa, PEOE_VSA, SMR_VSA, SlogP_VSA, EState_VSA, VSA_EState. Целесообразно сократить признаковое пространство, объединив признаки в группы, что также позволит исключить сильную зависимость признаков.

Создан новый признак Chi путем суммирования значений переменных Chi0, Chi1, Chi2n, Chi3n, … Аналогично созданы признаки Kappa, PEOE_VSA, SMR_VSA, SlogP_VSA, EState_VSA, VSA_EState. Старые колонки исключены. В результате количество признаков сократилось до 129.

6. Выводы по подготовке данных и разведочному анализу (EDA).

С учетом проведенного анализа обработаны данные и сформирован датасет, на котором будем решать задачи регрессии и классификации (датасет сохранен в файл df). Датасет обработан следующим образом:

- **Пропуски:** пропуски в данных заполнены нолями.
- **Колонки, которые исключены из датасета:** удалены колонки, в которых только нулевые значения, и колонка Unnamed, так как не имеют полезной информации для анализа.
- **Логарифмирование:** целевые переменные логарифмированы с помощью log(x+1). Нелогарифмированые целевые переменные также оставим в датасете, они пригодятся для решения задач классификации.
- **Нормализация данных:** проведена нормализация признаков с помощью MinMaxScaler.
- **Объединение взаимосвязанных признаков:** взаимосвязанные признаки объединены в новые переменные. Старые колонки исключены.

4. Задачи регрессии.

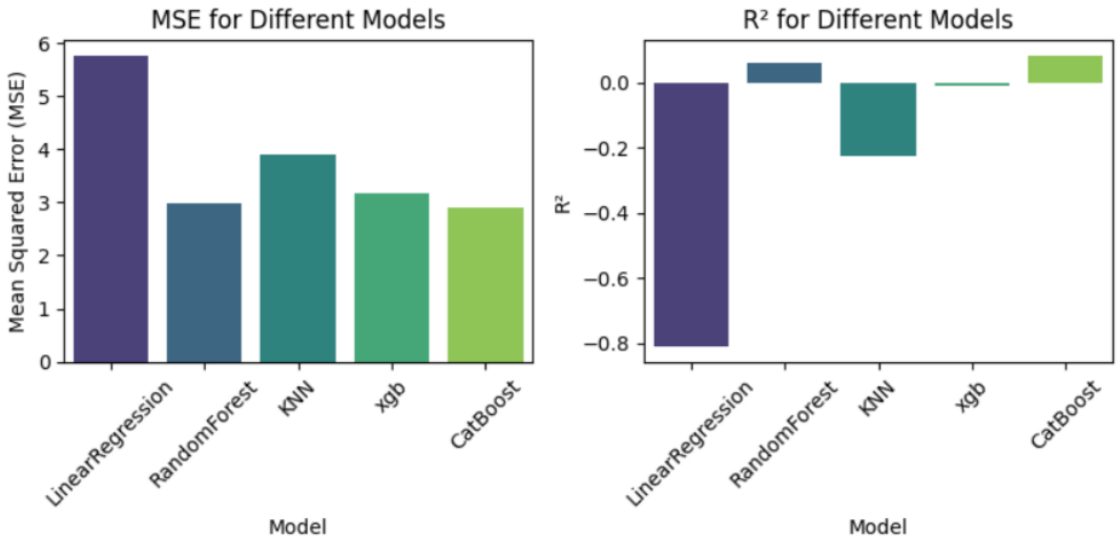
Во всех задачах регрессии использовались следующие модели: LinearRegression, RandomForest, KNN, XGBoost, CatBoost.

1. Регрессия для IC50.

Из датасета выделена логарифмированная целевая переменная IC50_log. Сначала все модели обучены без настройки гиперпараметров. Получены следующие метрики:

	MSE	R^2
LinearRegression	5.758295	-0.811225
RandomForest	2.979500	0.059215
KNN	3.887026	-0.225203
xgb	3.181301	-0.011345
CatBoost	2.893588	0.084480

Выведем полученные результаты в виде диаграмм:



Выводы по результатам обучения без подбора гиперпараметров моделей:

- **Метрики:** полученные метрики показывают низкое качество моделей. Показатель MSE по всем моделям достаточно высокий. R^2 по всем моделям, кроме RandomForest и CatBoost, ниже 0, что говорит о том, что модели хуже, чем просто предсказание по среднему значению.
- **Сравнение результатов моделей:** лучшие метрики показали модели RandomForest и CatBoost. В то же время, R^2 по этим моделям близок к 0, то есть модели плохо описывают вариацию данных.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	R^2	MSE	MAE	Гиперпараметры
LinearRegression	0.1426	3.2458	1.4096	('copy_X', True), ('fit_intercept', False), ('n_jobs', 10), ('positive', True)
RandomForest	0.4114	1.9736	1.1126	('max_depth', 11), ('min_samples_leaf', 5), ('min_samples_split', 6), ('n_estimators', 218)
KNN	0.3486	2.3267	1.2229	('n_neighbors', 6), ('weights', 'uniform')
xgb	0.3746	2.2461	1.1738	('max_depth', 2), ('n_neighbors', 20)
CatBoost	0.3936	2.3783	1.2123	('depth', 4), ('n_estimators', 100)

Выводы по результатам обучения с подбором гиперпараметров моделей:

- **Метрики:** после настройки гиперпараметров метрики по всем моделям существенно улучшились. Показатель MSE снизился, R^2 по всем моделям выше 0.
- **Сравнение результатов моделей:** лучшие метрики показала модель RandomForest. Также хорошие метрики по сравнению с другими моделями получены на моделях CatBoost и XGBoost, что показывает наибольшую эффективность ансамблевых моделей для задачи регрессии IC50. Самые низкие метрики получены на модели линейной регрессии, что говорит о том, что линейная зависимость не подходит для прогнозирования показателя IC50.

2. Регрессия для CC50.

Из датасета выделена логарифмированная целевая переменная CC50_log. При обучении без настройки гиперпараметров получены следующие метрики:

	MSE	R^2
LinearRegression	3.958619	-0.743131
RandomForest	2.579234	-0.134633
KNN	3.180629	-0.381080
xgb	2.718505	-0.204517
CatBoost	2.491514	-0.096805

Выводы по результатам обучения без подбора гиперпараметров моделей:

- **Метрики:** полученные метрики показывают низкое качество моделей. Показатель MSE по всем моделям достаточно высокий. R^2 по всем моделям ниже 0, что говорит о том, что модели хуже, чем просто предсказание по среднему значению.
- **Сравнение результатов моделей:** лучшие метрики показала модель CatBoost. При этом R^2 по модели ниже 0.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	R^2	MSE	MAE	Гиперпараметры
LinearRegression	0.1641	1.5249	0.8958	('fit_intercept', False), ('n_jobs', 10)
RandomForest	0.4188	1.3632	0.8314	('max_depth', 31), ('min_samples_leaf', 5), ('min_samples_split', 4), ('n_estimators', 193)
KNN	0.3210	1.4453	0.7757	('n_neighbors', 7), ('weights', 'distance')
xgb	0.3967	1.3889	0.8415	('max_depth', 2), ('n_neighbors', 60)
CatBoost	0.4276	1.2386	0.7790	('depth', 4), ('n_estimators', 250)

Выводы по результатам обучения с подбором гиперпараметров моделей:

- **Метрики:** после настройки гиперпараметров метрики по всем моделям существенно улучшились. Показатель MSE снизился, R^2 по всем моделям выше 0.
- **Сравнение результатов моделей:** лучшие метрики, как и до настройки гиперпараметров, вновь показала модель CatBoost. Также хорошие метрики по сравнению с другими моделями получены на моделях CatBoost и XGBoost. Самые низкие метрики получены на модели линейной регрессии, что говорит о том, что линейная зависимость не подходит для прогнозирования показателя CC50.

3. Регрессия для SI.

Из датасета выделена логарифмированная целевая переменная SI_log. Сначала все модели обучены без настройки гиперпараметров. Получены следующие метрики:

	MSE	R^2
LinearRegression	1.991972	-0.138863
RandomForest	1.207127	0.380406
KNN	1.420949	0.233926
xgb	1.364933	0.290791
CatBoost	1.188200	0.370472

Выводы по результатам обучения без подбора гиперпараметров моделей:

- **Метрики:** метрики показывают, что модели лучше отработали по сравнению с задачами регрессии IC50 и CC50 (сравниваются метрики обучения до подбора гиперпараметров). Частично это объясняется тем, что при решении задачи регрессии SI из датасета не исключен параметр IC50_log, от которого зависит целевая переменная SI_log. R^2 по всем моделям, кроме LinearRegression, выше 0.
- **Сравнение результатов моделей:** лучшие метрики показали модели RandomForest и CatBoost. При этом R^2 по моделям менее 0.4, попробуем улучшить метрики с помощью подбора гиперпараметров.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	R^2	MSE	MAE	Гиперпараметры
LinearRegression	0.4599	0.8731	0.6982	('copy_X', False), ('fit_intercept', False), ('n_jobs', 1), ('positive', False)
RandomForest	0.6063	0.7101	0.5962	('max_depth', 13), ('min_samples_leaf', 4), ('min_samples_split', 5), ('n_estimators', 94)
KNN	0.5511	0.6545	0.5942	('n_neighbors', 12), ('weights', 'distance')
xgb	0.5884	0.7611	0.6309	('max_depth', 3), ('n_neighbors', 17)
CatBoost	0.6069	0.6446	0.5776	('depth', 7), ('n_estimators', 244)

Выводы по результатам обучения с подбором гиперпараметров моделей:

- **Метрики:** после настройки гиперпараметров метрики по всем моделям существенно улучшились. Показатель MSE снизился, R^2 по всем моделям растет.
- **Сравнение результатов моделей:** лучшие метрики показала модель CatBoost. Также хорошие метрики по сравнению с другими моделями получены на моделях RandomForest и XGBoost, что показывает наибольшую эффективность ансамблевых моделей для задачи регрессии SI. Самые низкие метрики получены на модели линейной регрессии, что говорит о том, что линейная зависимость не подходит для прогнозирования показателя SI.

4. Выводы по задачам регрессии.

По всем задачам регрессии качество моделей существенно улучшилось после подбора гиперпараметров. По всем задачам набольшую эффективность показали модели RandomForest и CatBoost, а также другая ансамблевая модель XGBoost. Они наилучшим образом описывают сложную зависимость между дескрипторами и целевыми переменными IC50, CC50 и SI.

Самые низкие метрики получены на модели линейной регрессии, что говорит о том, что линейная зависимость не подходит для прогнозирования показателей IC50, CC50 и SI.

5. Задачи классификации.

Во всех задачах классификации использовались следующие модели: DecisionTree, RandomForest, KNN, CatBoost.

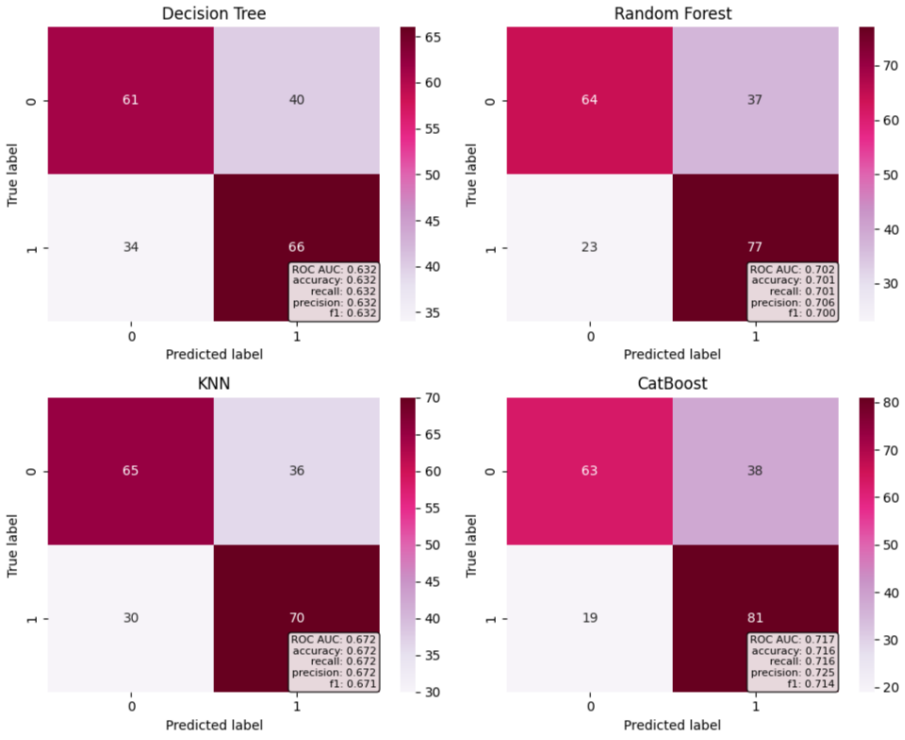
1. Классификация: превышает ли значение IC50 медианное значение выборки.

Из датасета выделена нелогарифмированная целевая переменная IC50. Медиана по колонке IC50 составляет 46.5852. Присвоим значениям из колонки IC50, превышающим значение медианы, класс «1», остальным значениям – класс «0».

Все модели сначала обучены без настройки гиперпараметров. Получены следующие метрики:

	model	roc_auc	accuracy	recall	precision	f1
0	Decision Tree	0.631980	0.631841	0.631841	0.632422	0.631567
1	Random Forest	0.701832	0.701493	0.701493	0.705685	0.700142
2	KNN	0.671782	0.671642	0.671642	0.672353	0.671398
3	CatBoost	0.716881	0.716418	0.716418	0.724700	0.713998

Выведем полученные результаты на матрице ошибок:



Выводы по результатам обучения без подбора гиперпараметров моделей:

по всем моделям получены достаточно высокие метрики, что говорит о том, что модели хорошо описывают зависимости между признаками и целевой переменной. Лучшие метрики показала модель CatBoost.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	ROC AUC	accuracy	recall	precision	f1	Гиперпараметры
DecisionTree	0.7257	0.6368	0.6368	0.6434	0.6330	('max_depth', 7), ('min_samples_split', 8)
RandomForest	0.8123	0.7214	0.7214	0.7274	0.7197	('max_depth', 16), ('min_samples_leaf', 4), ('min_samples_split', 6), ('n_estimators', 207)

Модель	ROC AUC	accuracy	recall	precision	f1	Гиперпараметры
KNN	0.7775	0.6517	0.6517	0.6537	0.6504	('n_neighbors', 6), ('weights', 'uniform')
CatBoost	0.8086	0.7015	0.7015	0.7057	0.7001	('depth', 7), ('n_estimators', 86)

Выводы по результатам обучения с подбором гиперпараметров моделей:

после настройки гиперпараметров метрики по всем моделям улучшились. При этом наилучшие показатели теперь демонстрирует модель RandomForest, хотя по модели CatBoost они тоже по-прежнему высокие.

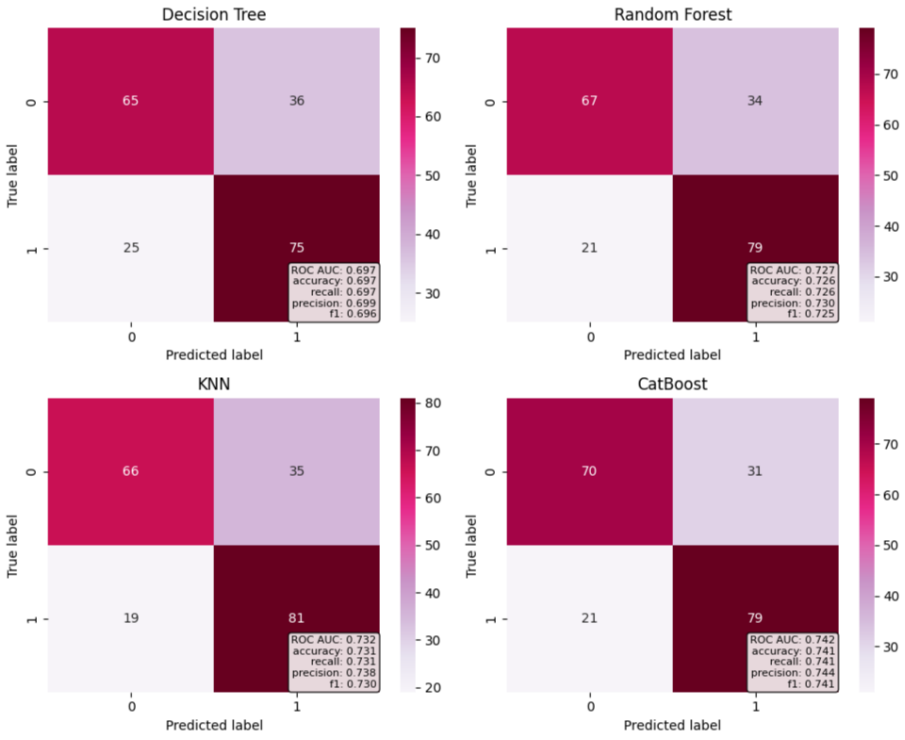
2. Классификация: превышает ли значение CC50 медианное значение выборки.

Из датасета выделена нелогарифмированная целевая переменная CC50. Медиана по колонке CC50 составляет 411.0393. Присвоим значениям из колонки CC50, превышающим значение медианы, класс «1», остальным значениям – класс «0».

Все модели сначала обучены без настройки гиперпараметров. Получены следующие метрики:

	model	roc_auc	accuracy	recall	precision	f1
0	Decision Tree	0.696782	0.696517	0.696517	0.699065	0.695689
1	Random Forest	0.726683	0.726368	0.726368	0.730394	0.725308
2	KNN	0.731733	0.731343	0.731343	0.737568	0.729738
3	CatBoost	0.741535	0.741294	0.741294	0.743833	0.740716

Выведем полученные результаты на матрице ошибок:



Выводы по результатам обучения без подбора гиперпараметров моделей:

по всем моделям получены достаточно высокие метрики, что говорит о том, что модели хорошо описывают зависимости между признаками и целевой переменной. Лучшие метрики показала модель CatBoost.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	ROC AUC	accuracy	recall	precision	f1	Гиперпараметры
DecisionTree	0.7852	0.6866	0.6866	0.6890	0.6857	('max_depth', 12), ('min_samples_split', 25)
RandomForest	0.8500	0.7015	0.7015	0.7084	0.6992	('max_depth', 7), ('min_samples_leaf', 8), ('min_samples_split', 10), ('n_estimators', 345)
KNN	0.8168	0.7114	0.7114	0.7172	0.7097	('n_neighbors', 7), ('weights', 'uniform')
CatBoost	0.8552	0.7363	0.7363	0.7419	0.7349	('depth', 3), ('n_estimators', 270)

Выводы по результатам обучения с подбором гиперпараметров моделей:

после настройки гиперпараметров метрики по всем моделям улучшились. При этом наилучшие метрики снова показала модель CatBoost.

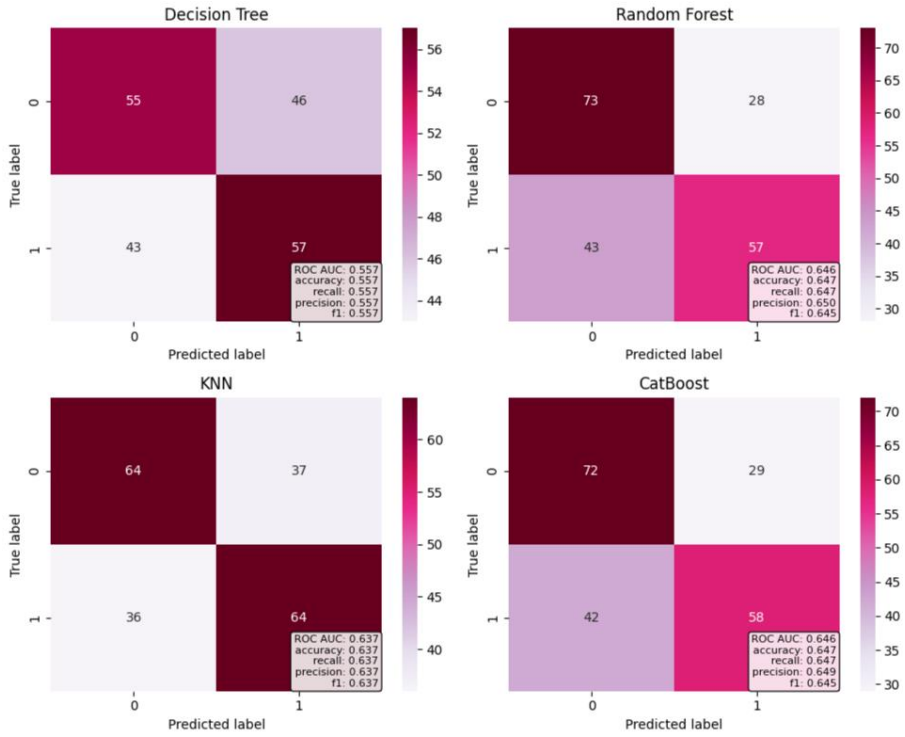
3. Классификация: превышает ли значение SI медианное значение выборки.

Из датасета выделена нелогарифмированная целевая переменная SI. Медиана по колонке SI составляет 3.8462. Присвоим значениям из колонки SI, превышающим значение медианы, класс «1», остальным значениям – класс «0».

Все модели сначала обучены без настройки гиперпараметров. Получены следующие метрики:

	model	roc_auc	accuracy	recall	precision	f1
0	Decision Tree	0.557277	0.557214	0.557214	0.557331	0.557148
1	Random Forest	0.646386	0.646766	0.646766	0.649847	0.644654
2	KNN	0.636832	0.636816	0.636816	0.636847	0.636816
3	CatBoost	0.646436	0.646766	0.646766	0.649036	0.645167

Выведем полученные результаты на матрице ошибок:



Выводы по результатам обучения без подбора гиперпараметров моделей:

по всем моделям получены достаточно высокие метрики, что говорит о том, что модели хорошо описывают зависимости между признаками и целевой переменной. При этом модели RandomForest и CatBoost показали примерно одинаковые метрики.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	ROC AUC	accuracy	recall	precision	f1	Гиперпараметры
DecisionTree	0.6844	0.6119	0.6119	0.6125	0.6113	('max_depth', 14), ('min_samples_split', 23)
RandomForest	0.7401	0.6667	0.6667	0.6713	0.6641	('max_depth', 13), ('min_samples_leaf', 4), ('min_samples_split', 6), ('n_estimators', 322)
KNN	0.7149	0.6418	0.6418	0.6418	0.6418	('n_neighbors', 10), ('weights', 'distance')
CatBoost	0.7358	0.6667	0.6667	0.6693	0.6652	('depth', 6), ('n_estimators', 50)

Выводы по результатам обучения с подбором гиперпараметров моделей:

после настройки гиперпараметров метрики по всем моделям улучшились. При этом наилучшие показатели демонстрирует модель RandomForest.

4. Классификация: превышает ли значение SI значение 8.

Из датасета выделена нелогарифмированная целевая переменная SI. Присвоим значениям из колонки SI, превышающим значение 8, класс «1», остальным значениям – класс «0».

Все модели сначала обучены без настройки гиперпараметров. Получены следующие метрики:

	model	roc_auc	accuracy	recall	precision	f1
0	Decision Tree	0.663114	0.701493	0.701493	0.695201	0.697202
1	Random Forest	0.664729	0.711443	0.711443	0.702558	0.703331
2	KNN	0.638243	0.661692	0.661692	0.666024	0.663599
3	CatBoost	0.691053	0.741294	0.741294	0.734423	0.731251

Выводы по результатам обучения без подбора гиперпараметров моделей:

по всем моделям получены достаточно высокие метрики, что говорит о том, что модели хорошо описывают зависимости между признаками и целевой переменной. Лучшие метрики показала модель CatBoost.

Посмотрим на результаты, полученные при подборе гиперпараметров для каждой модели:

Модель	ROC AUC	accuracy	recall	precision	f1	Гиперпараметры
DecisionTree	0.6841	0.7164	0.7164	0.7101	0.7117	('max_depth', 5), ('min_samples_split', 20)
RandomForest	0.7523	0.7313	0.7313	0.7235	0.7193	('max_depth', 9), ('min_samples_leaf', 1), ('min_samples_split', 4), ('n_estimators', 250)
KNN	0.7304	0.6766	0.6766	0.6703	0.6726	('n_neighbors', 13), ('weights', 'distance')
CatBoost	0.7547	0.7313	0.7313	0.7239	0.7238	('depth', 7), ('n_estimators', 538)

Выводы по результатам обучения с подбором гиперпараметров моделей:

после настройки гиперпараметров метрики по всем моделям улучшились. При этом наилучшие показатели демонстрирует модель CatBoost.

5. Выводы по задачам классификации.

По всем задачам классификации качество моделей улучшилось после подбора гиперпараметров. По всем задачам набольшую эффективность показали модели RandomForest и CatBoost, а также другая ансамблевая модель XGBoost. Они наилучшим образом описывают сложную зависимость между дескрипторами и целевыми переменными IC50, CC50 и SI.

6. Заключение.

В рамках данной курсовой работы были решены следующие задачи:

- Исследованы данные, проведен разведочный анализ и предобработка данных.
- Обучены различные модели для задач регрессии и классификации.
- Подобраны наилучшие гиперпараметры для моделей, проведена оценка качества моделей и их сравнение по метрикам.

Основные полученные выводы:

наибольшую эффективность как на задачах регрессии, так и на задачах классификации показали ансамблевые модели, особенно модели RandomForest и CatBoost. Подбор гиперпараметров является мощным инструментом повышения качества моделей. В целом модели классификации показали более высокие результаты по сравнению с задачами регрессии, что вероятнее всего связано со сложными зависимостями между признаками и целевыми переменными.