Аналитический отчет

1. EDA

Предоставленный датасет «Данные_для_курсовои_Классическое_МО.xlsx» содержит 1001 строку и 214 числовых признаков типа float и int.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1001 entries, 0 to 1000
Columns: 214 entries, Unnamed: 0 to fr_urea
dtypes: float64(107), int64(107)
memory usage: 1.6 MB
```

Рис. 1. Информация о датасете

	IC50, mM	CC50, mM	SI	MaxAbsEStateIndex	MaxEStateIndex	MinAbsEStateIndex	MinEStateIndex	qed	SPS	MolWt	...	fr_quatN	fr_sulfide	f
0	6.239374	175.482382	28.125000	5.094096	5.094096	0.387225	0.387225	0.417362	42.928571	384.652	...	0	0	
1	0.771831	5.402819	7.000000	3.961417	3.961417	0.533868	0.533868	0.462473	45.214286	388.684	...	0	0	
2	223.808778	161.142320	0.720000	2.627117	2.627117	0.543231	0.543231	0.260923	42.187500	446.808	...	2	0	
3	1.705624	107.855654	63.235294	5.097360	5.097360	0.390603	0.390603	0.377846	41.862069	398.679	...	0	0	
4	107.131532	139.270991	1.300000	5.150510	5.150510	0.270476	0.270476	0.429038	36.514286	466.713	...	0	0	

5 rows × 195 columns

Рис. 2. Структура данных

Было удалено 18 признаков с константным значением и признак «Unnamed: 0» так как он дублирует индексы и не несет полезной информации. Значения NaN затрагивали только 3 строки, поэтому было принято решение заменить их на 0.

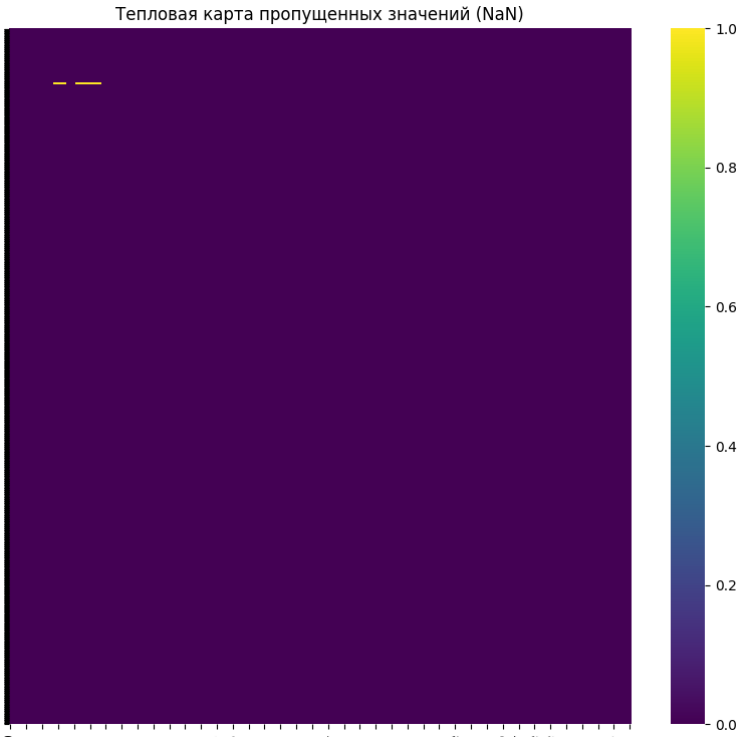

Рис. 3. Тепловая карта пропущенных значений

1.1 Подготовка датафреймов для задач регрессии

Первым шагом стало удаление выбросов для целевых переменных и сохранение отдельных датафреймов с каждым из трех таргетов.

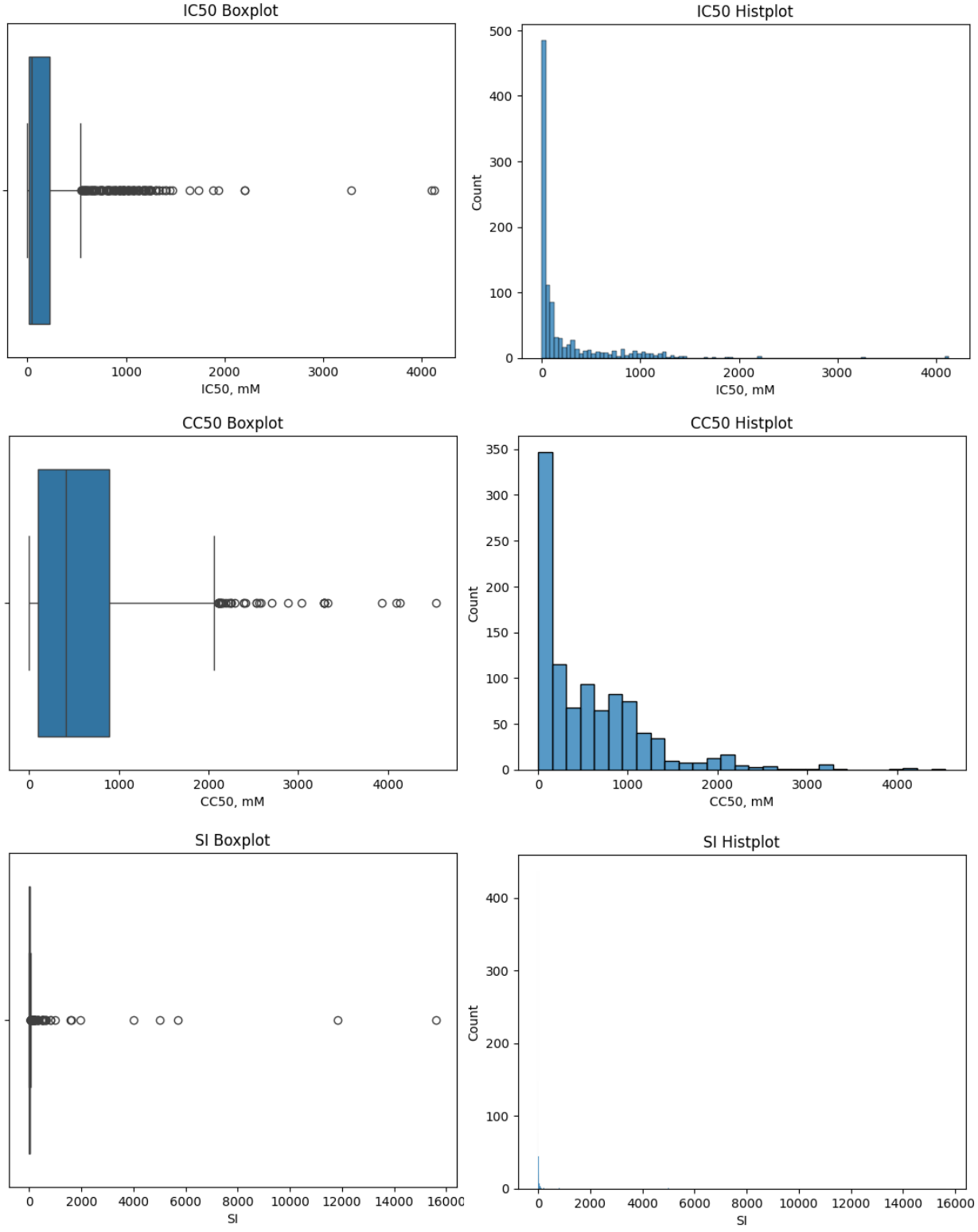

Рис. 4. Распределения целевых переменных

Для удаления выбросов использовался межквартальный размах с удалением данных больше третьего квартиля.

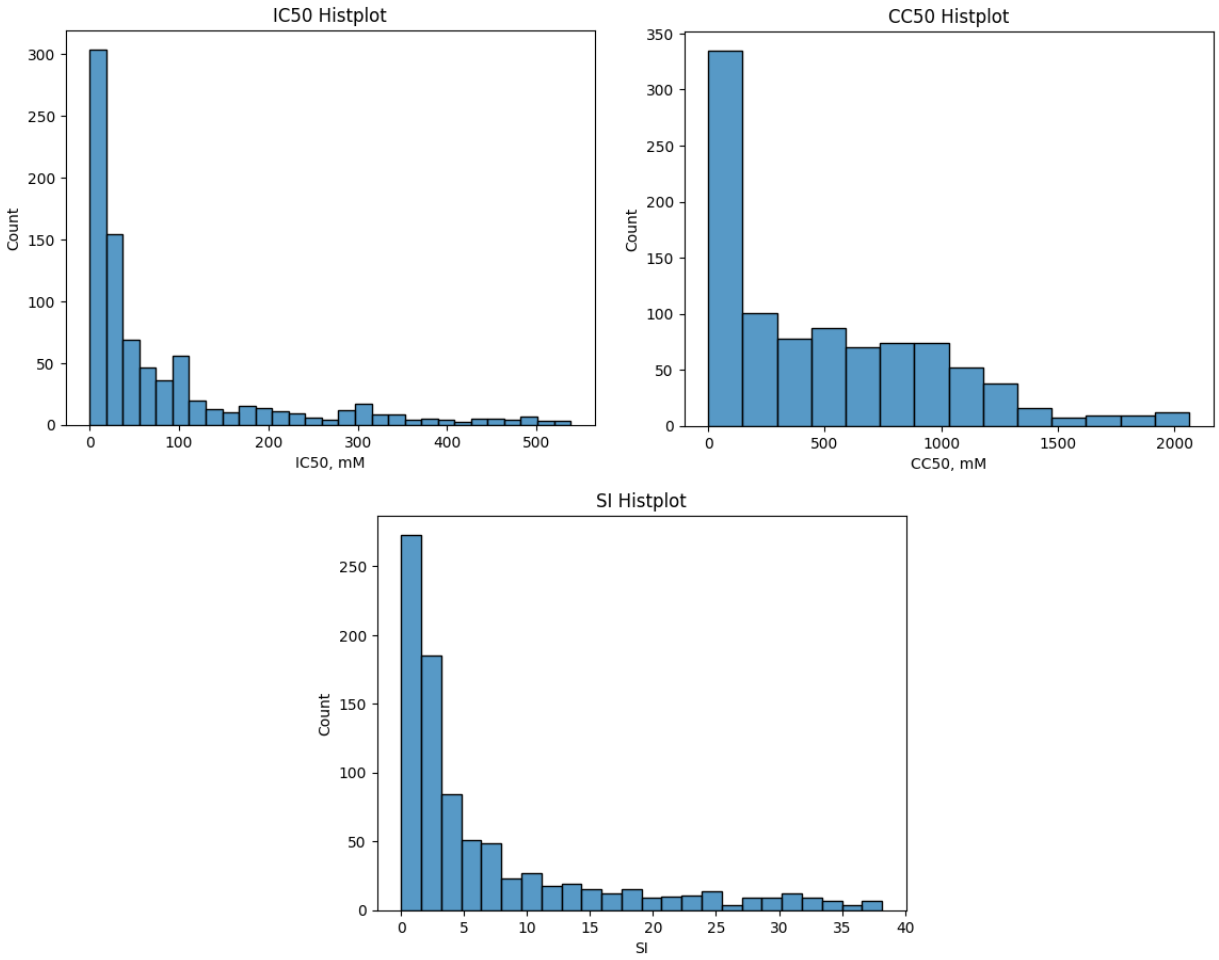

Рис. 5. Полученные распределения

Вторым шагом стал анализ корреляционных матриц для определения, не коррелирующих с таргетом, признаков и устранения мультиколлинеарности. В следствие этого были удалены признаки fr_Ar_COO, fr_tetrazole для IC50 так как имели корреляцию NaN.

Для устранения мультиколлинеарности были удалены коррелирующие с другими признаки при коэффициенте больше 0,7. В результате для IC50 количество признаков (кроме целевой переменной) составило 101, для CC50 – 102, SI – 104.

Для нивелирования действия масштабов данных была применена MinMax нормализация.

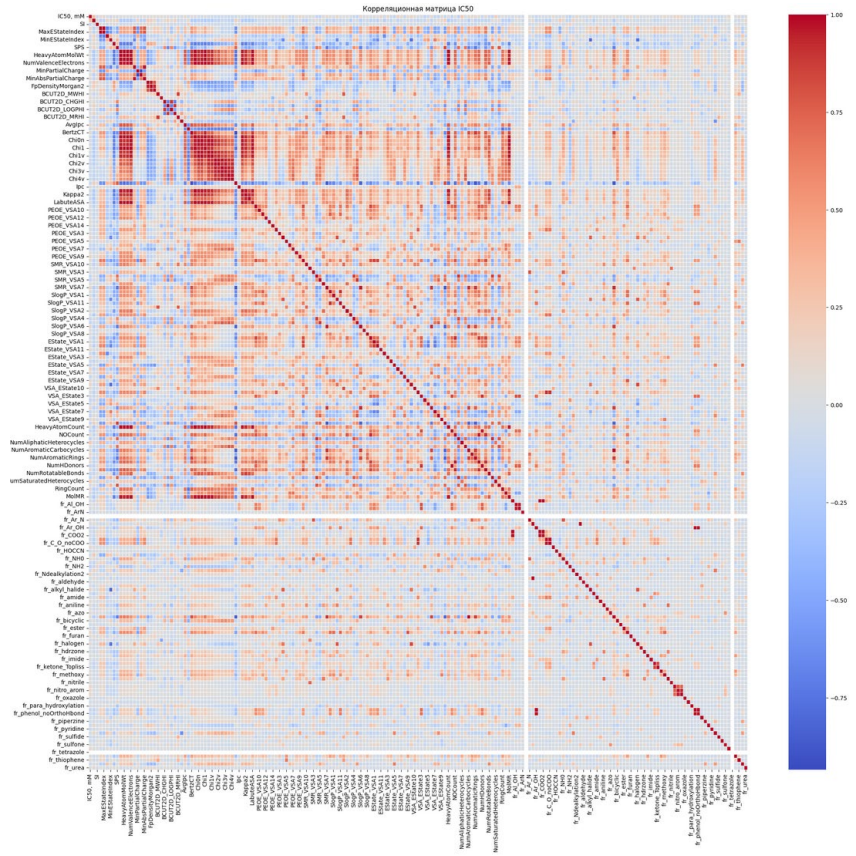

Рис. 6. Полная корреляционная матрица для IC50 регрессии

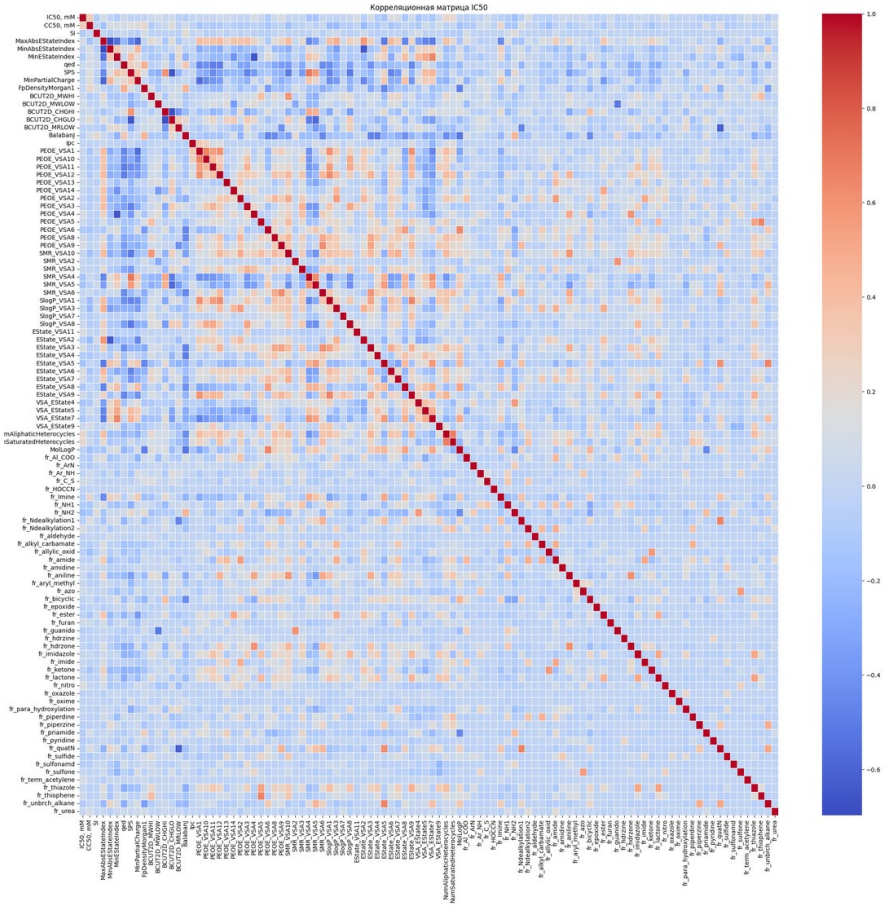

Рис. 7. Итоговая корреляционная матрица для IC50 регрессии

1.2 Подготовка датафреймов для задач классификации

Для задач классификации были созданы 4 датафрейма с заменой целевых переменных на 1 если выполнено условие больше медианы/константы, в иных случаях – 0 с названиями greater_median и greater_const соответственно. Целевые переменные созданы искусственно и исследование на выбросы не имеет значимости.

Для устранения мультиколлинеарности также были удалены коррелирующие с другими признаки при коэффициенте больше 0,7. В результате для IC50 количество признаков (кроме целевой переменной) составило 102, для CC50 – 102, SI – 102.

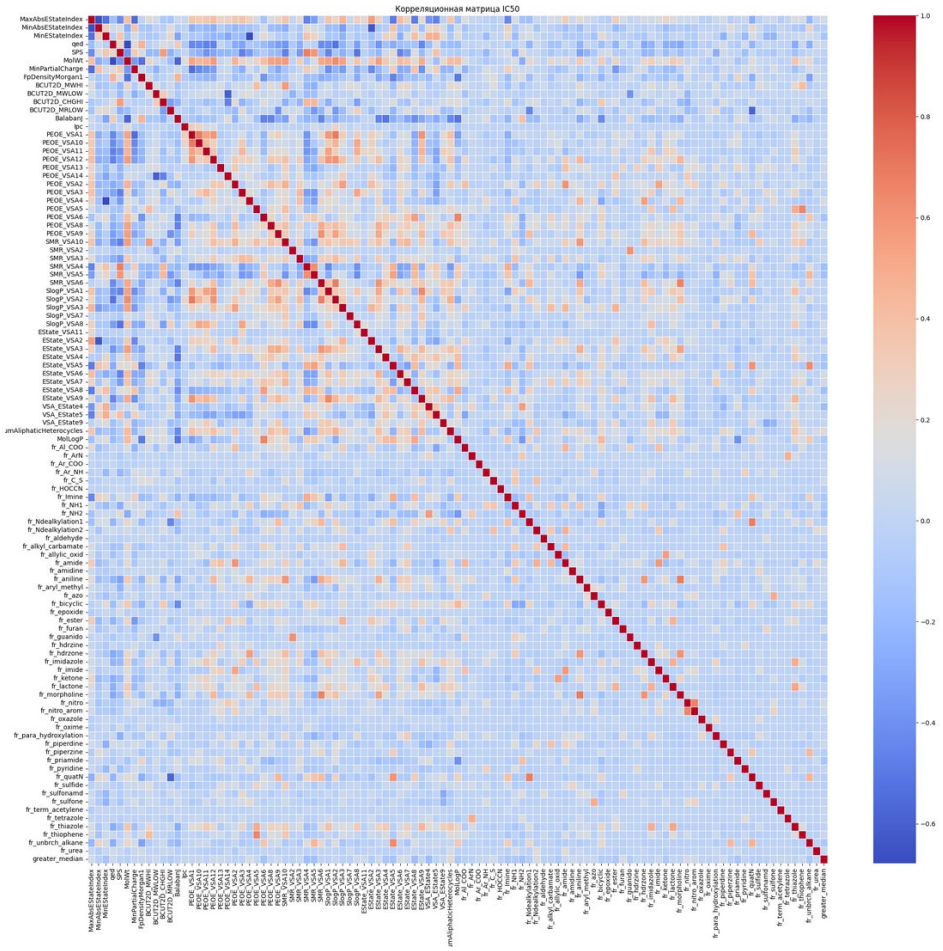

Рис. 8. Итоговая корреляционная матрица для IC50 классификации

Для нивелирования действия масштабов данных была применения MinMax нормализация.

В результате были сформированы 7 датафреймов для каждой задачи регрессии/классификации: df_IC50.csv, df_CC50.csv, df_SI.csv, df_IC50_median.csv, df_CC50_median.csv, df_SI_median.csv, df_SI_const.csv.

2. Создание моделей

Для решения задач регрессии были выбраны:

- LinearRegression;
- DecisionTree;
- RandomForest;
- CatBoost;
- SVR.

Для решения задач классификации были выбраны:

- LogisticRegression;
- DecisionTree;
- RandomForest;
- CatBoost;
- SVC.

Для подбора параметров использовался GridSearch, доля тестовой выборки– 20%.

Результаты подбора параметров показаны на рисунках 9-15:

	Model	Best Params	MAE	RMSE	R2
0	LinearRegression	{}	82.287533	20088.054455	-0.581880
1	DecisionTree	{'max_depth': 7}	20.548324	2439.613438	0.807887
2	RandomForest	{'max_depth': None, 'n_estimators': 100}	13.427617	1082.685854	0.914741
3	CatBoost	{'depth': 4, 'learning_rate': 0.1}	12.844061	849.017691	0.933142
4	SVR	{'C': 10, 'kernel': 'linear'}	65.735781	12709.625625	-0.000849

Рис. 9. Таблица результатов подбора параметров для IC50 регрессии

	Model	Best Params	MAE	RMSE	R2
0	LinearRegression	{}	239.912211	97366.778593	0.517987
1	DecisionTree	{'max_depth': 10}	133.119754	66335.863917	0.671605
2	RandomForest	{'max_depth': 10, 'n_estimators': 100}	81.955090	22370.627781	0.889255
3	CatBoost	{'depth': 4, 'learning_rate': 0.1}	73.398199	15510.893019	0.923214
4	SVR	{'C': 10, 'kernel': 'linear'}	254.356527	113576.619522	0.437741

Рис. 10. Таблица результатов подбора параметров для CC50 регрессии

	Model	Best Params	MAE	RMSE	R2
0	LinearRegression	{}	6.205717	145.034423	-0.728224
1	DecisionTree	{'max_depth': None}	1.797453	15.149786	0.819476
2	RandomForest	{'max_depth': 10, 'n_estimators': 100}	1.248469	7.973520	0.904988
3	CatBoost	{'depth': 4, 'learning_rate': 0.1}	1.000802	4.232411	0.949567
4	SVR	{'C': 10, 'kernel': 'linear'}	4.965624	68.694177	0.181443

Рис. 11. Таблица результатов подбора параметров для SI регрессии

	Model	Best Params	accuracy	f1_score
0	LogisticRegression	{'C': 10}	0.646766	0.646731
1	DecisionTree	{'max_depth': None}	0.582090	0.581996
2	RandomForest	{'max_depth': None, 'n_estimators': 100}	0.626866	0.626829
3	CatBoost	{'depth': 4, 'learning_rate': 0.01}	0.686567	0.685446
4	SVC	{'C': 10, 'kernel': 'rbf'}	0.681592	0.680636

Рис. 12. Таблица результатов подбора параметров для IC50 классификации

	Model	Best Params	accuracy	f1_score
0	LogisticRegression	{'C': 10}	0.711443	0.711378
1	DecisionTree	{'max_depth': 7}	0.666667	0.666634
2	RandomForest	{'max_depth': None, 'n_estimators': 100}	0.726368	0.726124
3	CatBoost	{'depth': 4, 'learning_rate': 0.01}	0.746269	0.746244
4	SVC	{'C': 10, 'kernel': 'rbf'}	0.716418	0.716390

Рис. 13. Таблица результатов подбора параметров для CC50 классификации

	Model	Best Params	accuracy	f1_score
0	LogisticRegression	{'C': 1}	0.601990	0.601980
1	DecisionTree	{'max_depth': 7}	0.587065	0.586696
2	RandomForest	{'max_depth': 5, 'n_estimators': 100}	0.651741	0.651318
3	CatBoost	{'depth': 6, 'learning_rate': 0.01}	0.641791	0.641569
4	SVC	{'C': 10, 'kernel': 'rbf'}	0.641791	0.641569

Рис. 14. Таблица результатов подбора параметров для SI классификации с медианой

	Model	Best Params	accuracy	f1_score
0	LogisticRegression	{'C': 1}	0.691542	0.624367
1	DecisionTree	{'max_depth': 7}	0.606965	0.565637
2	RandomForest	{'max_depth': 10, 'n_estimators': 50}	0.706468	0.651216
3	CatBoost	{'depth': 6, 'learning_rate': 0.01}	0.691542	0.631738
4	SVC	{'C': 10, 'kernel': 'rbf'}	0.641791	0.605453

Рис. 15. Таблица результатов подбора параметров для SI классификации с константой

3. Выводы

Анализируя полученные результаты, можно отметить, что:

- с задачами регрессии лучше всех справились CatBoost и RandomForest которым удалось достигнуть высоких коэффициентов детерминации и наименьших ошибок;
- с задачами классификации лидером также стал RandomForest, но его метрики незначительно превосходят другие модели.

Улучшить результаты могут:

- Использование балансировки классов для классификации
- Использование дополнительных химических дескрипторов
- Использование более сложных архитектур таких как нейросети (MLP или GNN)