Классическое машинное обучение.

Курсовая работа (vo_PJ)

Отчёт

Выполнил: Душкин Василий Алексеевич

г. Москва - 2025 г.

Введение:

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Датасет включает в себя 1001 запись, содержащую 214 признаков.

```
[ ] df.head()
```

	Unnamed: 0	IC50, mM	CC50, mM	SI	MaxAbsEStateIndex	MaxEStateIndex	MinAbsEStateIndex	MinEStateIndex	
0	0	6.239374	175.482382	28.125000	5.094096	5.094096	0.387225	0.387225	0
1	1	0.771831	5.402819	7.000000	3.961417	3.961417	0.533868	0.533868	0
2	2	223.808778	161.142320	0.720000	2.627117	2.627117	0.543231	0.543231	0
3	3	1.705624	107.855654	63.235294	5.097360	5.097360	0.390603	0.390603	0
4	4	107.131532	139.270991	1.300000	5.150510	5.150510	0.270476	0.270476	0

5 rows × 214 columns

Описание данных:

Некоторые из ключевых характеристик включают:

IC50, мМ — концентрация вещества, при которой наблюдается 50% ингибирование активности (измеряет эффективность лекарства), выраженная в миллимолях.

CC50, мМ — концентрация, вызывающая гибель 50% клеток (измеряет токсичность), также в миллимолях.

SI (селективный индекс) — определяется как отношение CC50 к IC50 и отражает терапевтическое окно.

1. EDA:

Делаем первичный анализ

<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

Датасет включает в себя 1001 запись, содержащую 214 признаков.
В качестве первого шага мы вычислим медианные значения для всех трех признаков.

Медианы столбцов
IC50, mM : 46.58518345980803
CC50, mM : 411.0393423370522
SI : 3.846153846153846

Количество пропущенных значений:
MaxPartialCharge 3
MinPartialCharge 3
MaxAbsPartialCharge 3
MinAbsPartialCharge 3
BCUT2D_MWHI 3
BCUT2D_MWLOW 3
BCUT2D_CHGHI 3
BCUT2D_CHGLO 3
BCUT2D_LOGPHI 3
BCUT2D_LOGPLOW 3
BCUT2D_MRHI 3
BCUT2D_MRLOW 3

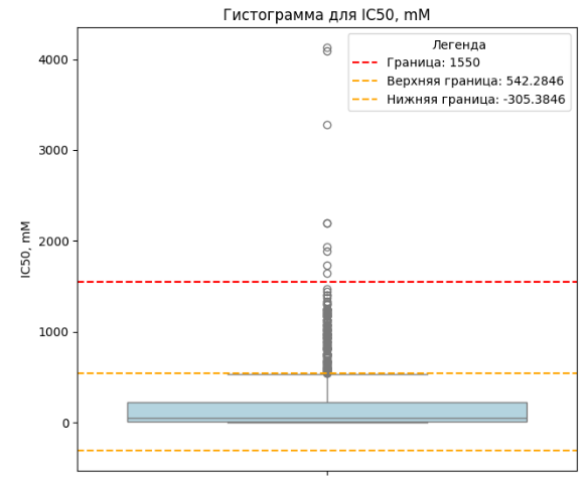
В общей сложности данные с пропусками составляют лишь 3 записи, что эквивалентно 0,3% от всего объема датасета. Таким образом, эти записи можно безопасно удалить. Также мы исключим первый столбец, поскольку он не содержит значимой информации.

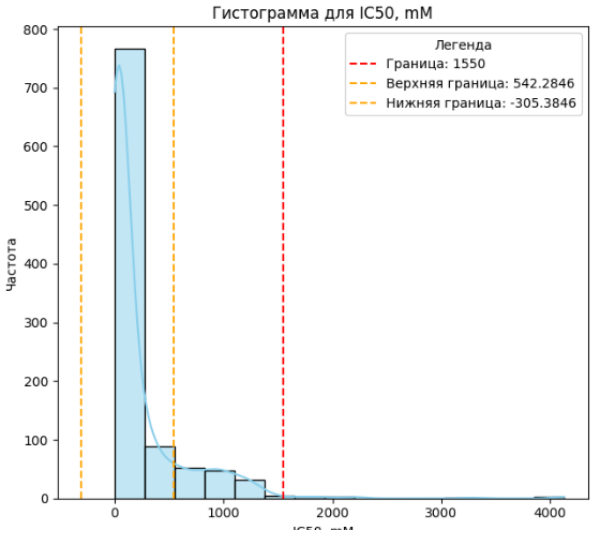
Анализ IC50, mM

```
target = 'IC50, mM'
df[target].describe()
```

	IC50, mM
count	998.000000
mean	221.118757
std	400.510657
min	0.003517
25%	12.491340
50%	45.992006
75%	224.408630
max	4128.529377

dtype: float64

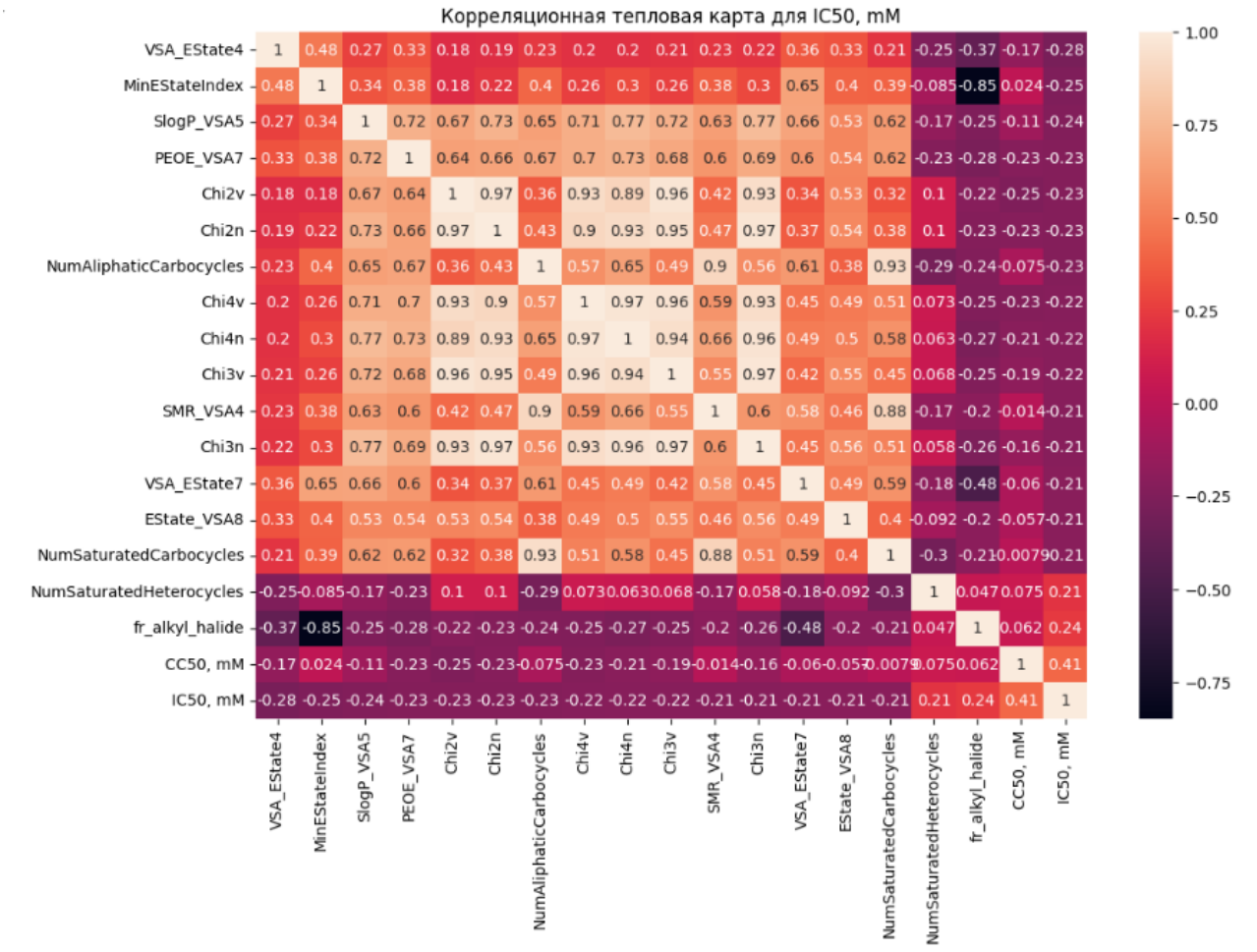


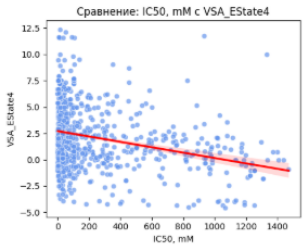


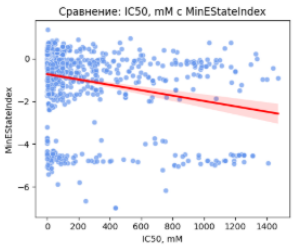
Корреляции между IC50, mM и

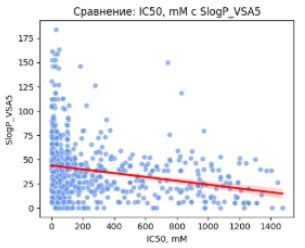
VSA_EState4 составляет: -0.2820
MinEStateIndex составляет: -0.2532
SlogP_VSA5 составляет: -0.2383
PEOE_VSA7 составляет: -0.2303
Chi2v составляет: -0.2292
Chi2n составляет: -0.2280
NumAliphaticCarbocycles составляет: -0.2279
Chi4v составляет: -0.2249
Chi4n составляет: -0.2203
Chi3v составляет: -0.2180
SMR_VSA4 составляет: -0.2149
Chi3n составляет: -0.2148
VSA_EState7 составляет: -0.2131
EState_VSA8 составляет: -0.2113
NumSaturatedCarbocycles составляет: -0.2071
NumSaturatedHeterocycles составляет: 0.2117
fr_alkyl_halide составляет: 0.2363
CC50, mM составляет: 0.4124

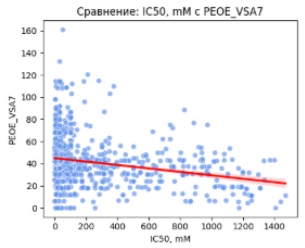
Тепловая карта

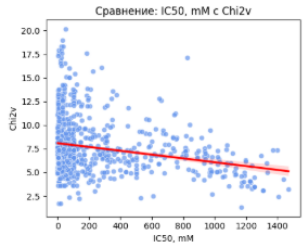


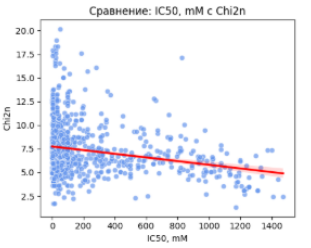

Сравнение: IC50, mM с VSA_EState4
VSA_EState4
IC50, mM

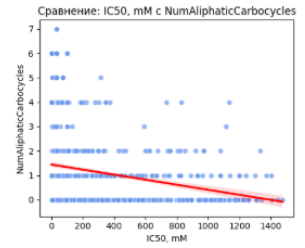

Сравнение: IC50, mM с MinEStateIndex
MinEStateIndex
IC50, mM

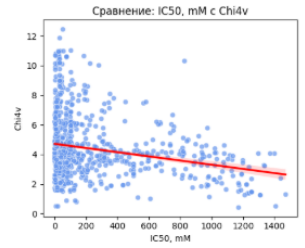

Сравнение: IC50, mM с SlogP_VSA5
SlogP_VSA5
IC50, mM

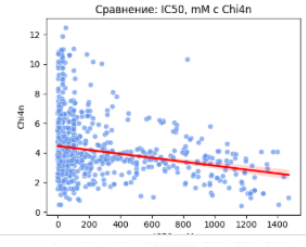

Сравнение: IC50, mM с PEOE_VSA7
PEOE_VSA7
IC50, mM

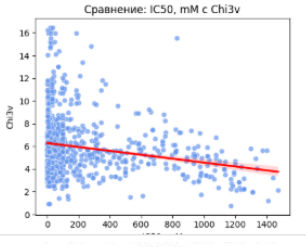

Сравнение: IC50, mM с Chi2v
Chi2v
IC50, mM

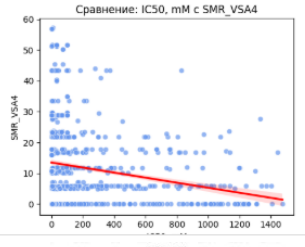

Сравнение: IC50, mM с Chi2n
Chi2n
IC50, mM

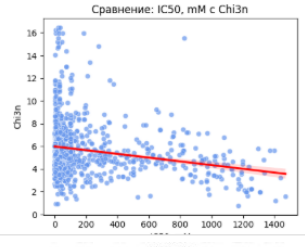

Сравнение: IC50, mM с NumAliphaticCarbocycles
NumAliphaticCarbocycles
IC50, mM

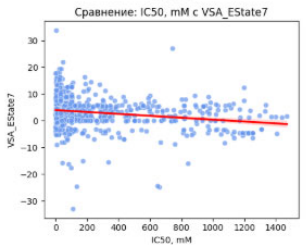

Сравнение: IC50, mM с Chi4v
Chi4v
IC50, mM

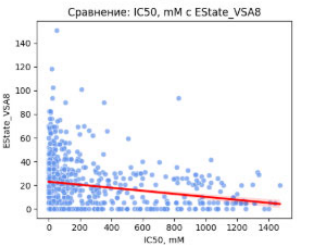

Сравнение: IC50, mM с Chi4n
Chi4n
IC50, mM

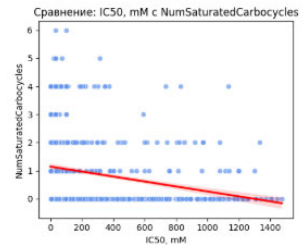

Сравнение: IC50, mM с Chi3v
Chi3v
IC50, mM

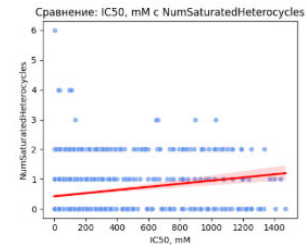

Сравнение: IC50, mM с SMR_VSA4
SMR_VSA4
IC50, mM

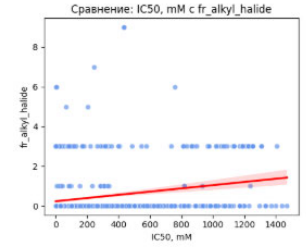

Сравнение: IC50, mM с Chi3n
Chi3n
IC50, mM

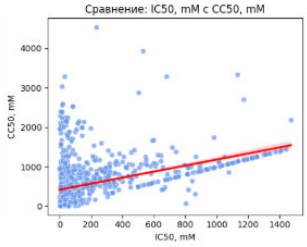

Сравнение: IC50, mM с VSA_EState7
VSA_EState7
IC50, mM


Сравнение: IC50, mM с EState_VSA8
EState_VSA8
IC50, mM


Сравнение: IC50, mM с NumSaturatedCarbocycles
NumSaturatedCarbocycles
IC50, mM


Сравнение: IC50, mM с NumSaturatedHeterocycles
NumSaturatedHeterocycles
IC50, mM


Сравнение: IC50, mM с fr_alkyl_halide
fr_alkyl_halide
IC50, mM

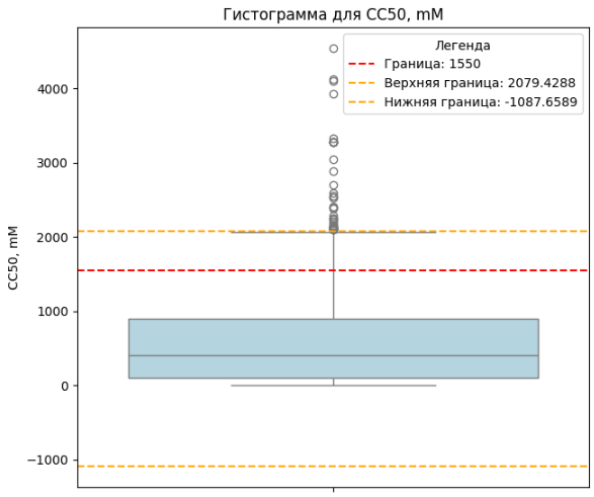

Сравнение: IC50, mM с CC50, mM
CC50, mM
IC50, mM

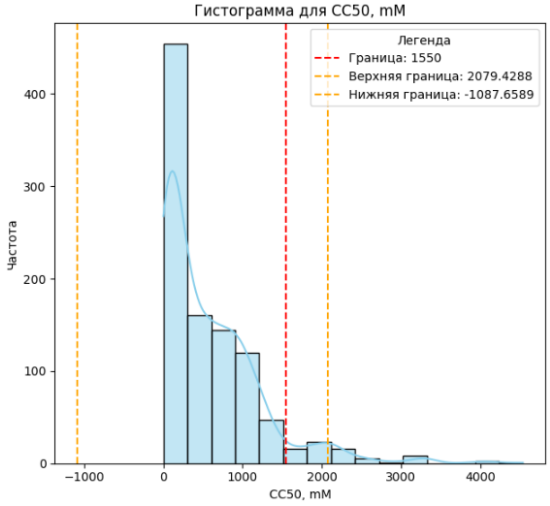
Анализ CC50, mM

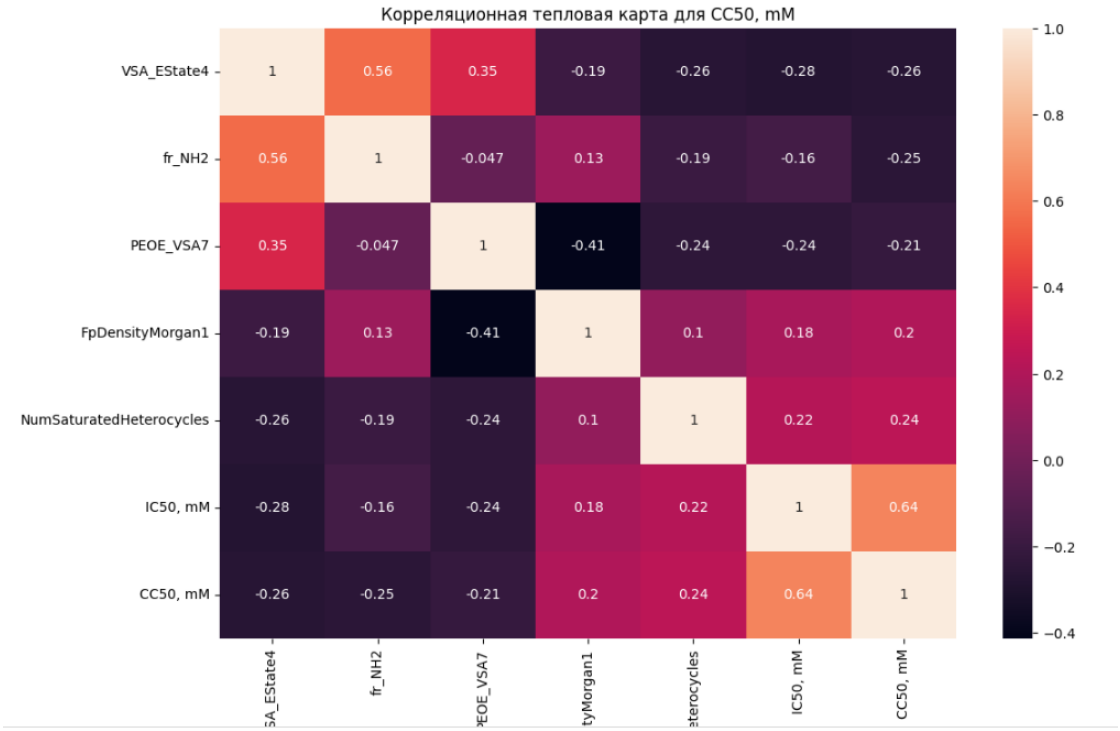
```
target = 'CC50, mM'
df[target].describe()
```

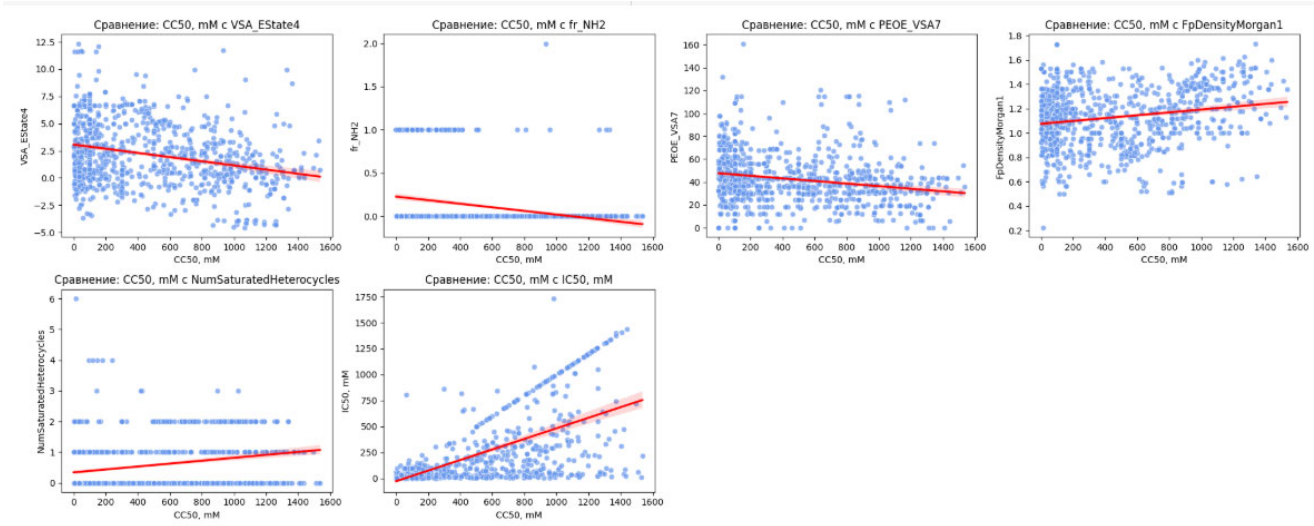
	CC50, mM
count	998.000000
mean	586.668414
std	642.016454
min	0.700808
25%	99.999036
50%	408.793314
75%	891.770961
max	4538.976189

dtype: float64






Корреляционная тепловая карта для CC50, mM
VSA_EState4
fr_NH2
PEOE_VSA7
FpDensityMorgan1
NumSaturatedHeterocycles
IC50, mM
CC50, mM
1 0.56 0.35 -0.19 -0.26 -0.28 -0.26
0.56 1 -0.047 0.13 -0.19 -0.16 -0.25
0.35 -0.047 1 -0.41 -0.24 -0.24 -0.21
-0.19 0.13 -0.41 1 0.1 0.18 0.2
-0.26 -0.19 -0.24 0.1 1 0.22 0.24
-0.28 -0.16 -0.24 0.18 0.22 1 0.64
-0.26 -0.25 -0.21 0.2 0.24 0.64 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Сравнение: CC50, mM с VSA_EState4
Сравнение: CC50, mM с fr_NH2
Сравнение: CC50, mM с PEOE_VSA7
Сравнение: CC50, mM с FpDensityMorgan1
Сравнение: CC50, mM с NumSaturatedHeterocycles
Сравнение: CC50, mM с IC50, mM
CC50, mM

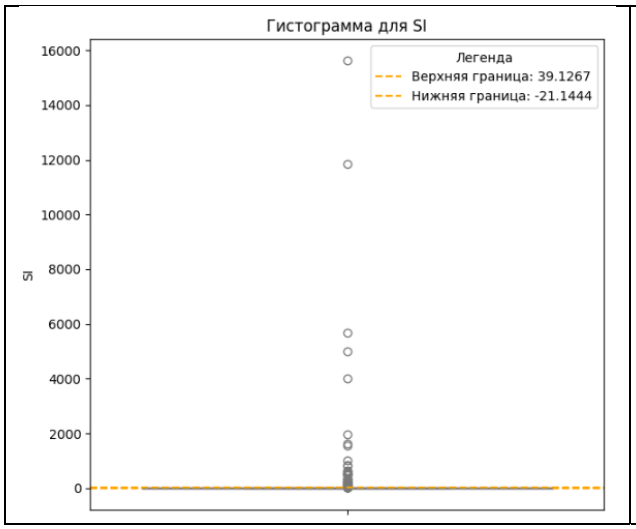
Анализ SI

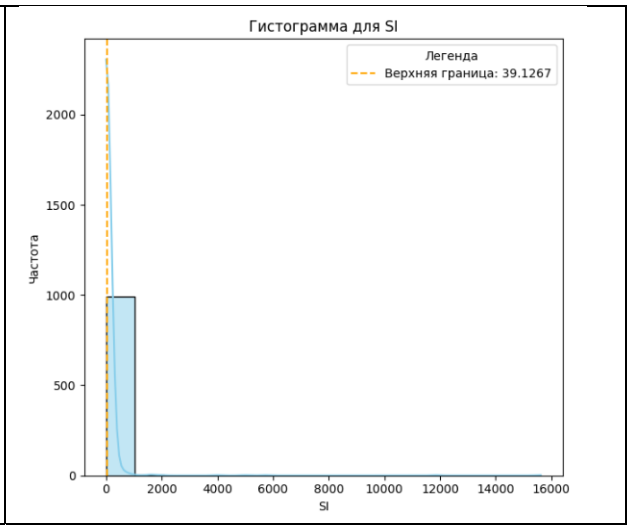
```
target = 'SI'
print(f"Описательная статистика для {target}:")
df[target].describe()
```

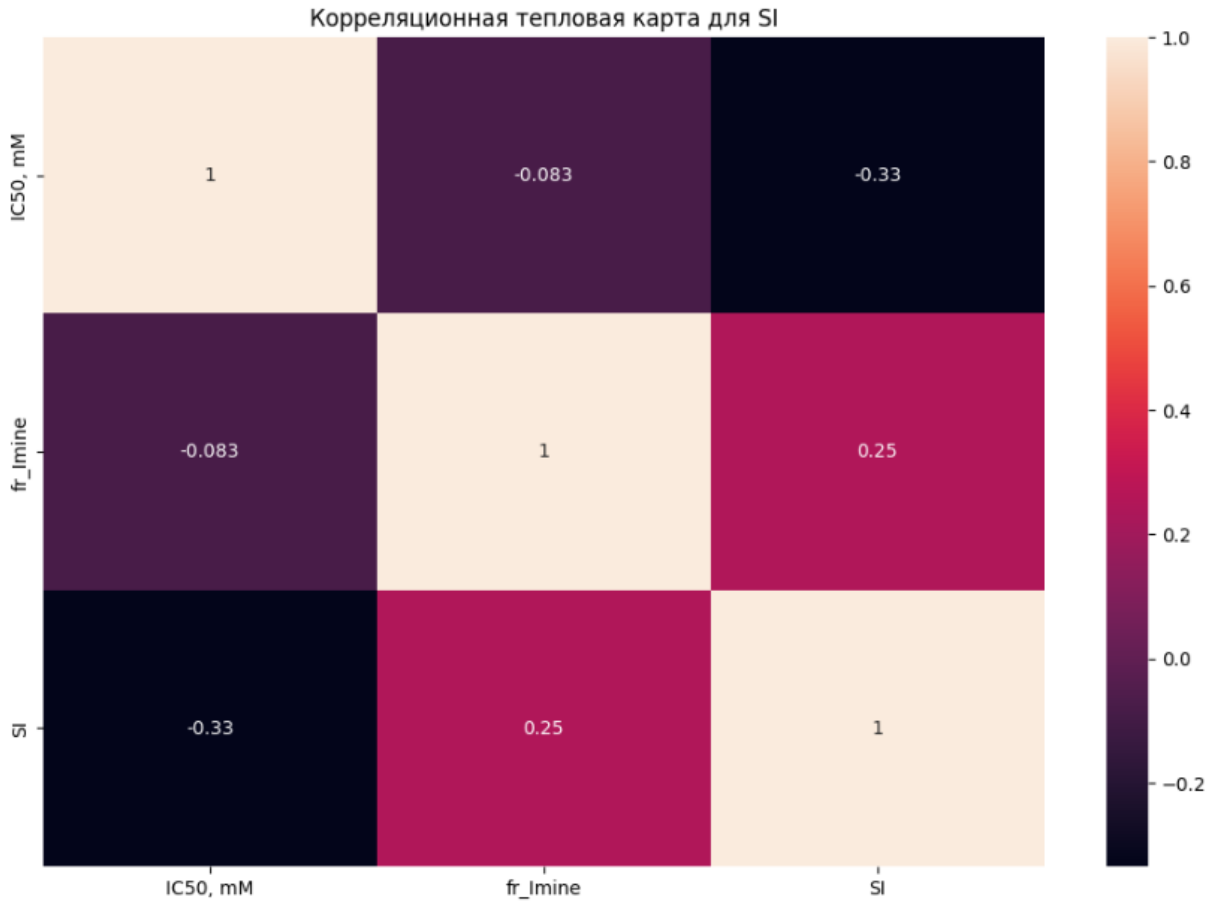
```
Описательная статистика для SI:
```

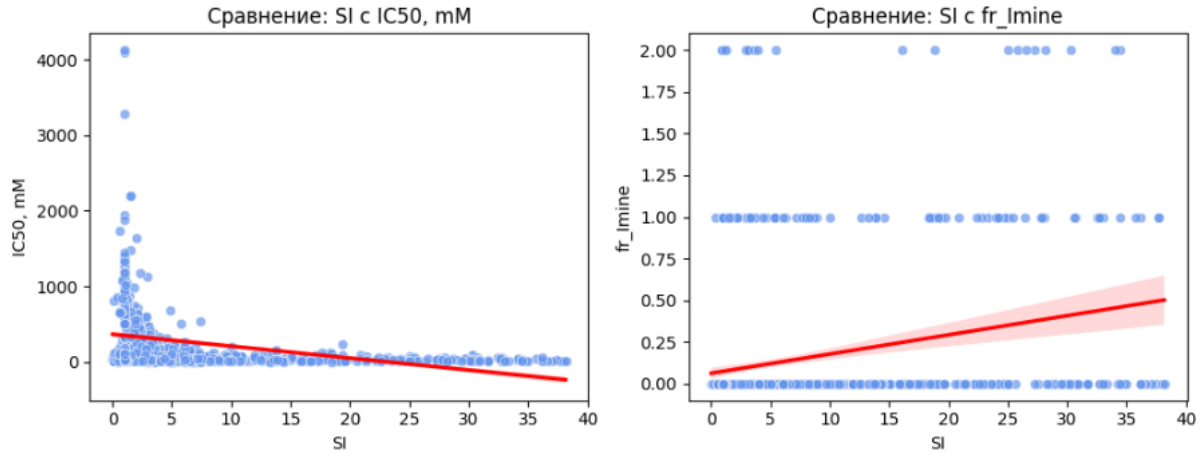
	SI
count	998.000000
mean	72.650005
std	685.504279
min	0.011489
25%	1.457233
50%	3.856410
75%	16.525000
max	15620.600000

dtype: float64









Корреляционный анализ показывает, что признак SI имеет значимую корреляцию только с IC50 и fr_Imine.

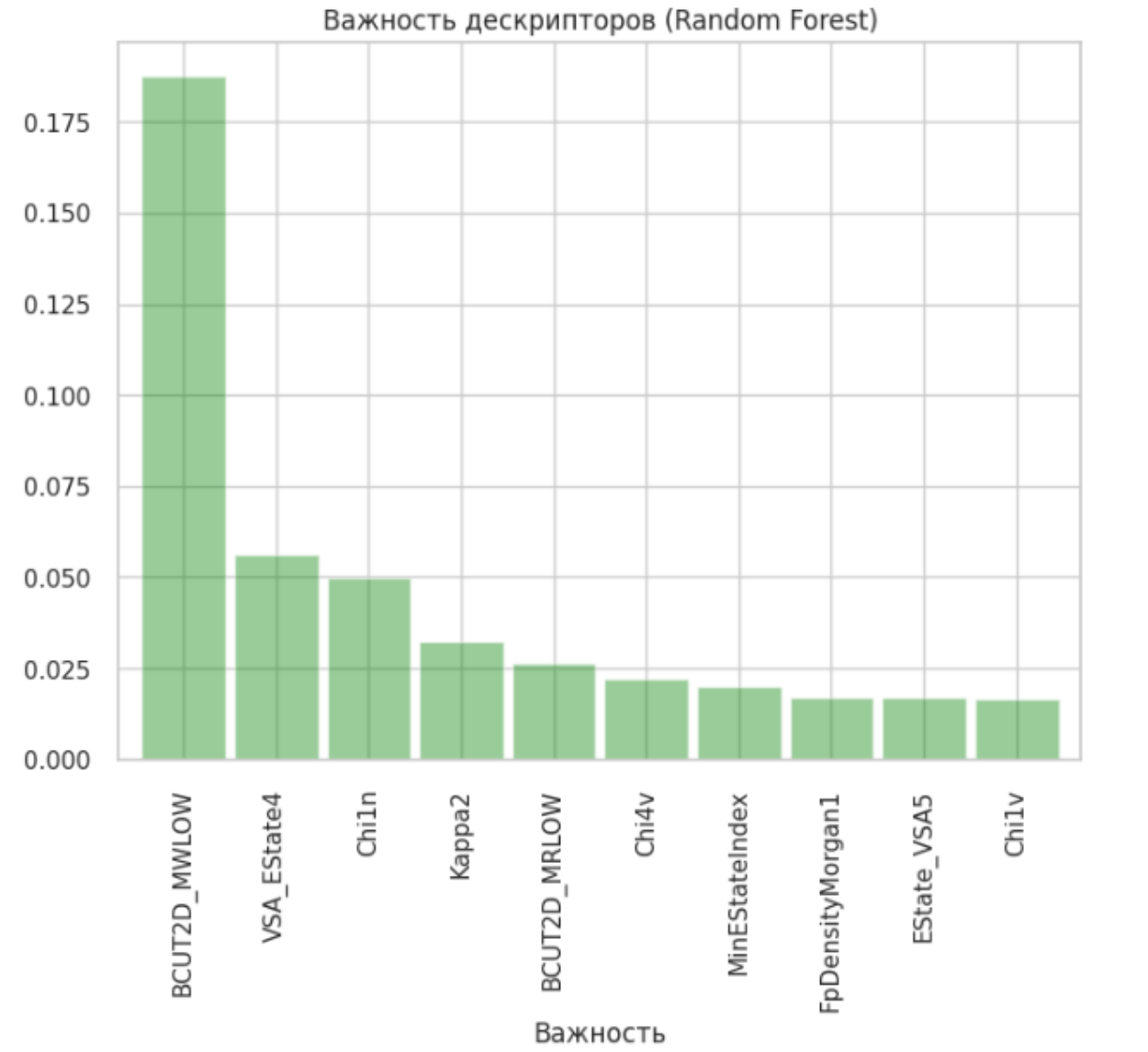
fr_Imine — это дескриптор, который подсчитывает количество иминных групп в молекуле.

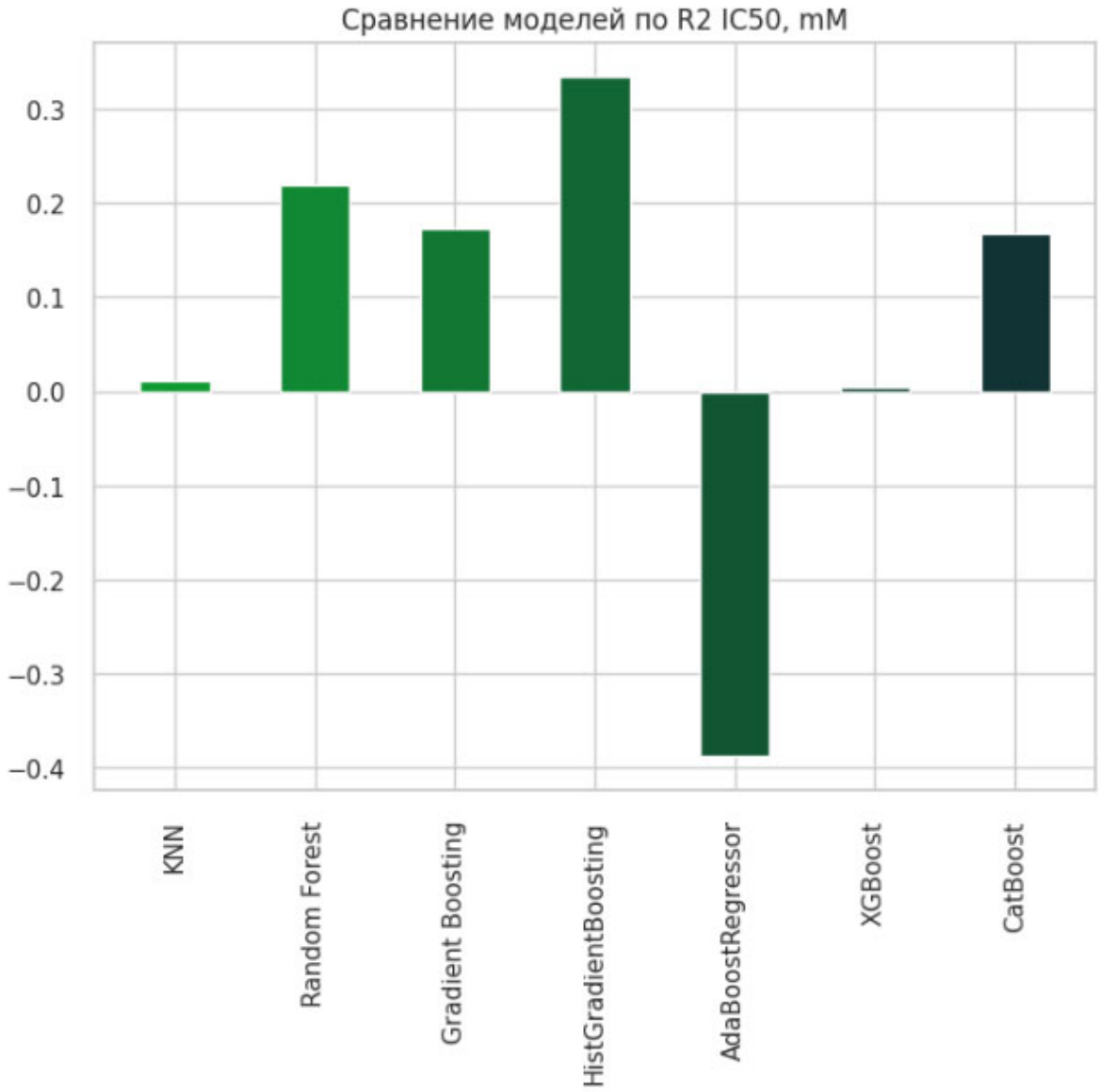
2. регрессия IC50

Топ-10 наиболее важных дескрипторов для IC50:

	0
BCUT2D_MWLOW	0.187558
VSA_EState4	0.056110
Chi1n	0.049614
Kappa2	0.032389
BCUT2D_MRLOW	0.026263
Chi4v	0.022165
MinEStateIndex	0.019823
FpDensityMorgan1	0.017069
EState_VSA5	0.016864
Chi1v	0.016481

dtype: float64





Результаты моделей (таргет: IC50, mM):

	Model	MSE	RMSE	MAE	r2
0	KNN	141572.045065	376.260608	225.628585	0.012167
1	Random Forest	111768.740873	334.318323	194.204800	0.220123
2	Gradient Boosting	118481.640735	344.211622	195.888401	0.173283
3	HistGradientBoosting	95274.069624	308.664980	191.231119	0.335216
4	AdaBoostRegressor	198698.722474	445.756349	371.003872	-0.386440
5	XGBoost	142558.931675	377.569771	197.890063	0.005281
6	CatBoost	119182.138803	345.227662	189.800134	0.168395

Наилучший результат

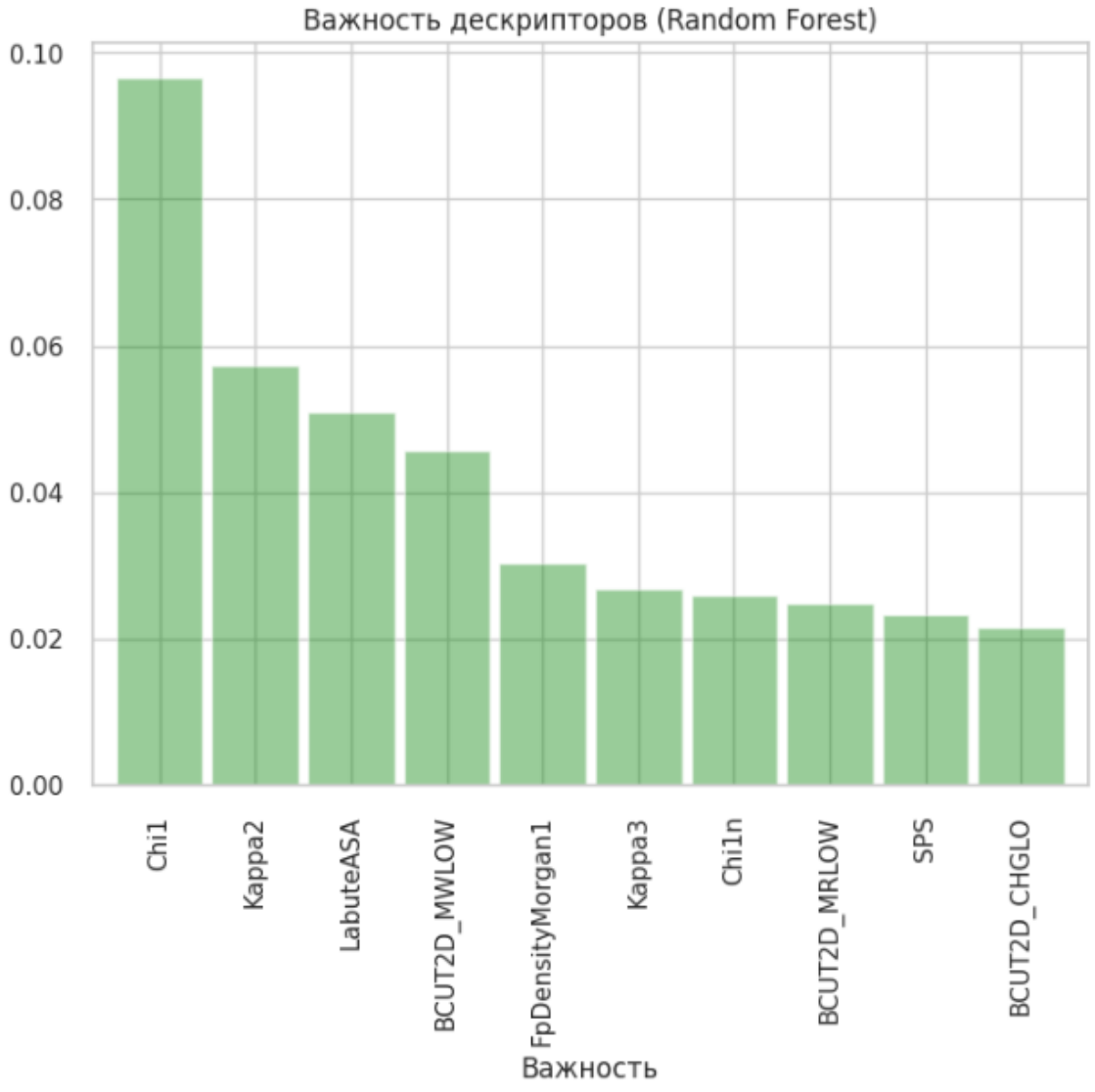
	Model	MSE	RMSE	MAE	r2
3	HistGradientBoosting	95274.069624	308.66498	191.231119	0.335216

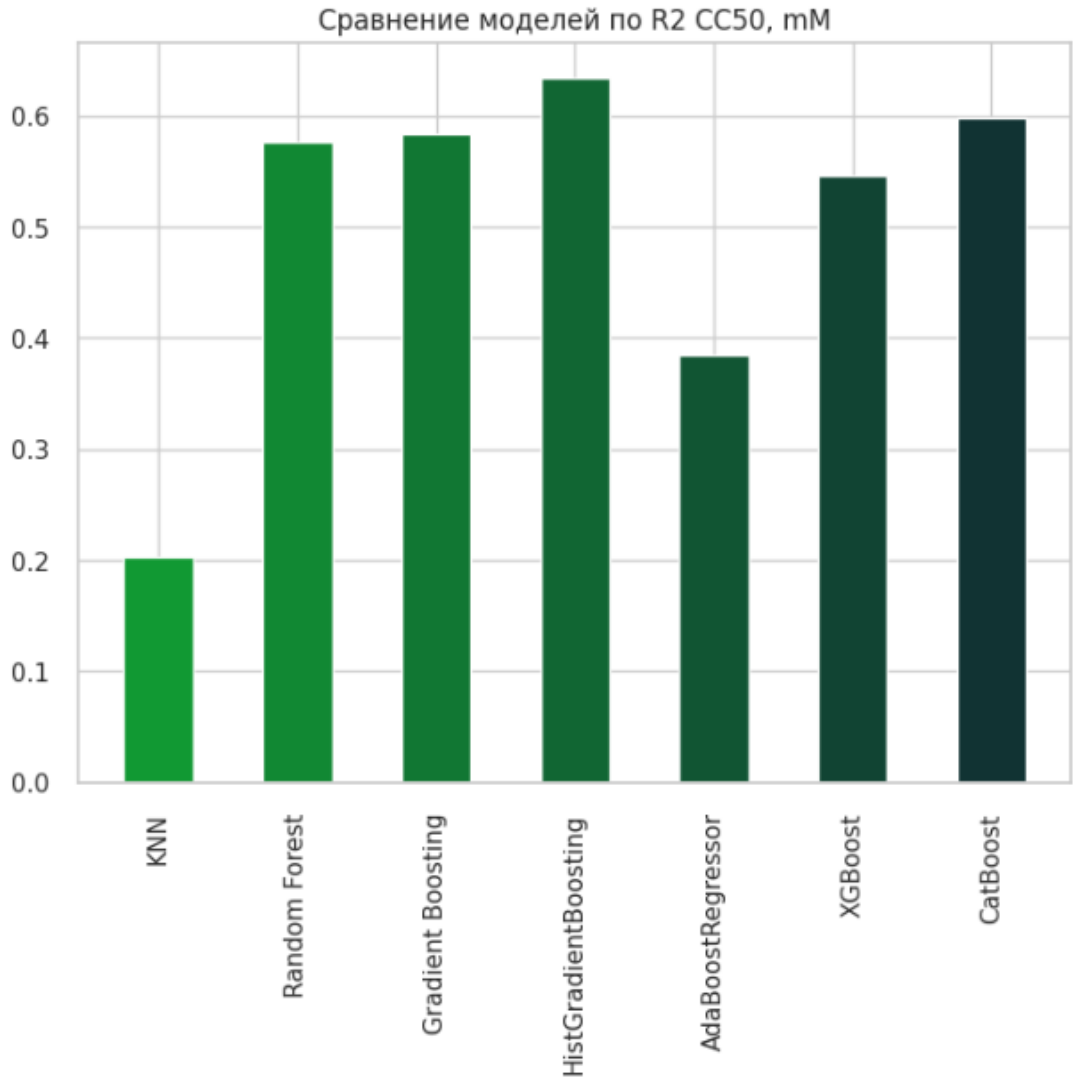
3. регрессия CC50

Топ-10 наиболее важных дескрипторов для IC50:

	0
Chi1	0.096628
Kappa2	0.057268
LabuteASA	0.050829
BCUT2D_MWLOW	0.045647
FpDensityMorgan1	0.030224
Kappa3	0.026649
Chi1n	0.025900
BCUT2D_MRLOW	0.024828
SPS	0.023210
BCUT2D_CHGLO	0.021536

dtype: float64





Результаты моделей (таргет: CC50, mM):

	Model	MSE	RMSE	MAE	r2
0	KNN	329130.739724	573.699172	427.787672	0.202105
1	Random Forest	174854.330334	418.155868	290.524837	0.576109
2	Gradient Boosting	171916.254416	414.627851	304.738134	0.583232
3	HistGradientBoosting	150649.531050	388.135970	278.924685	0.634788
4	AdaBoostRegressor	253871.254925	503.856383	434.282687	0.384553
5	XGBoost	187333.221960	432.820080	285.593407	0.545858
6	CatBoost	165651.136417	407.002625	278.048167	0.598420

Наилучший результат

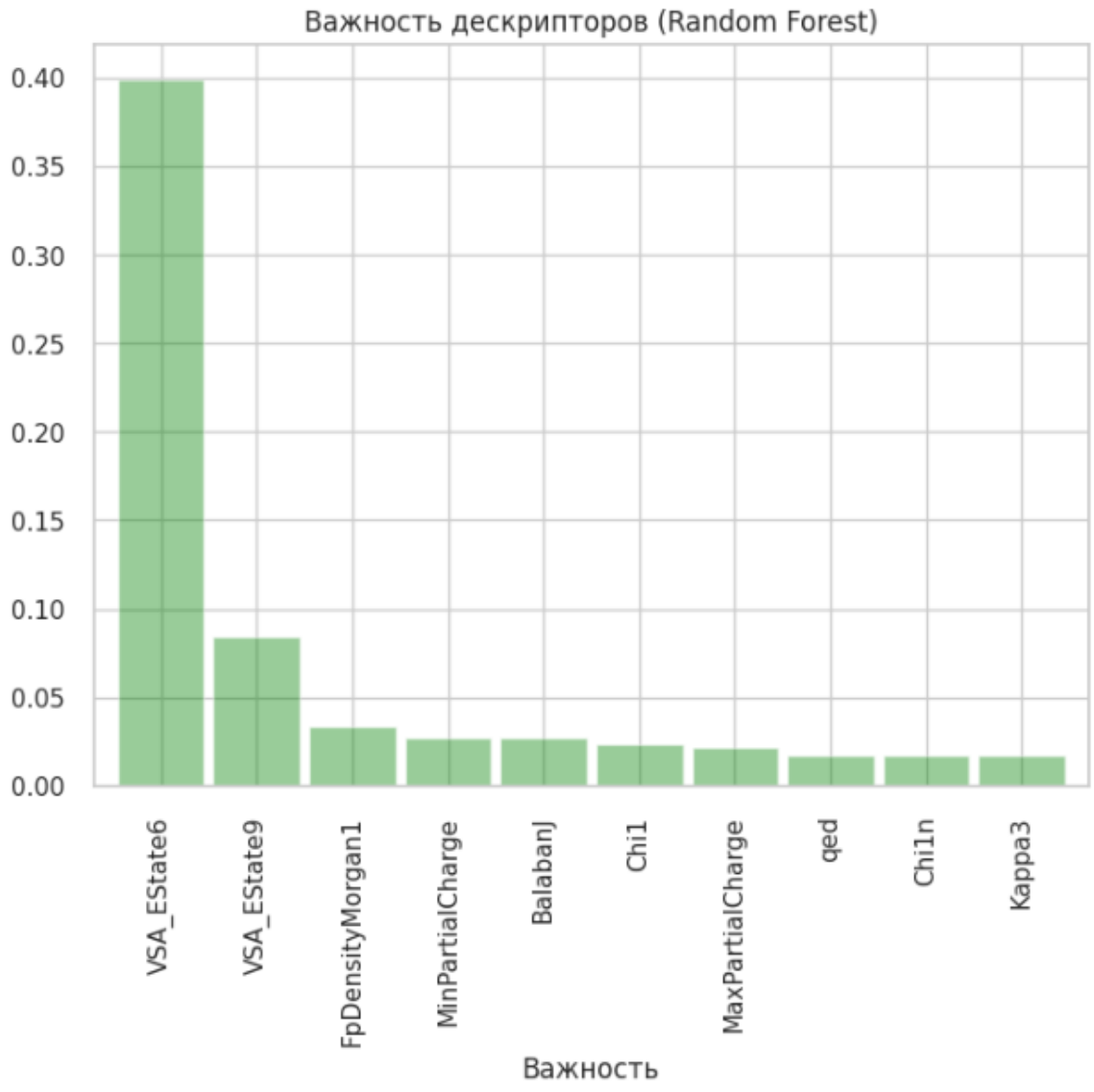
	Model	MSE	RMSE	MAE	r2
3	HistGradientBoosting	150649.53105	388.13597	278.924685	0.634788

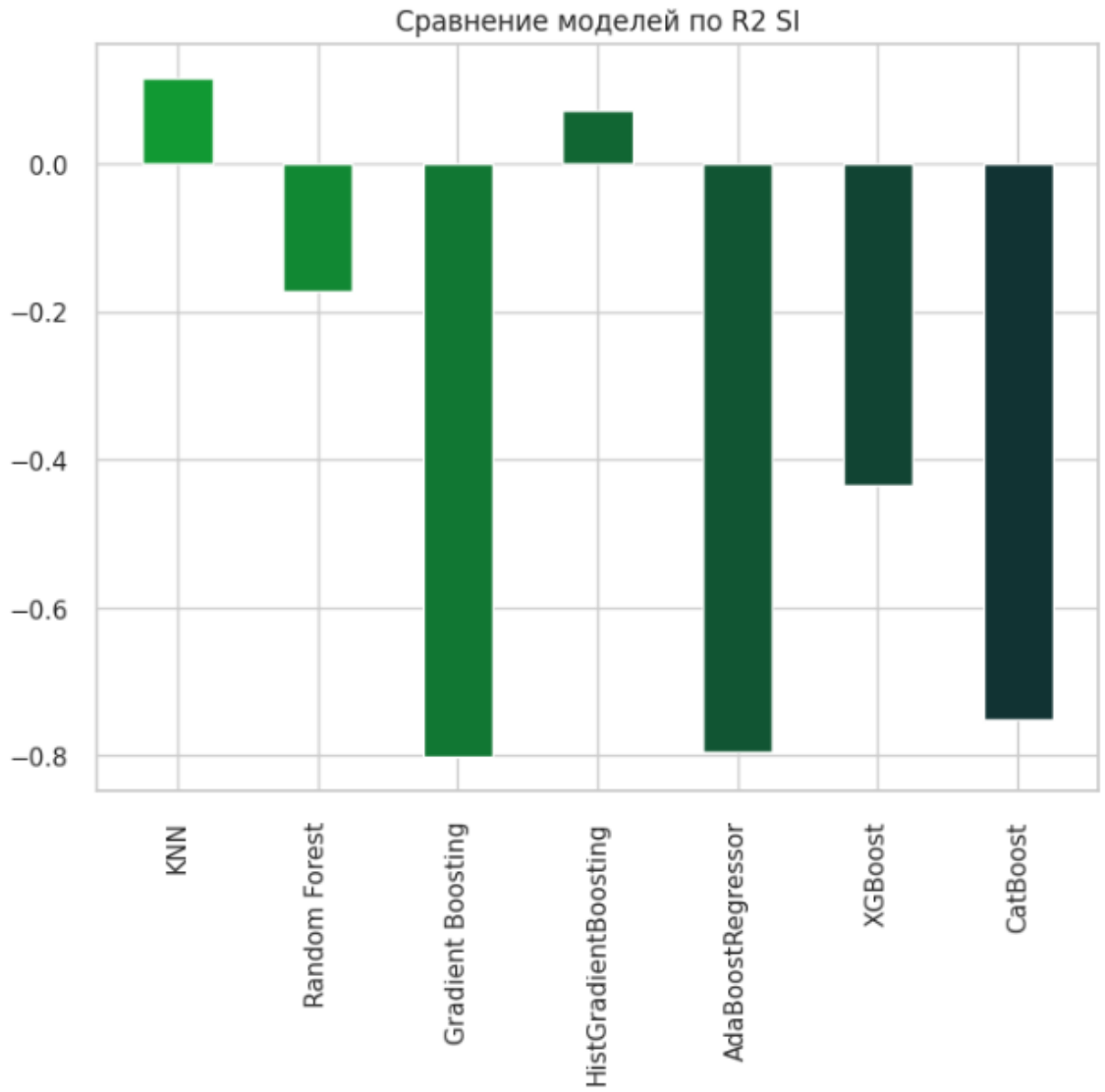
4. регрессия SI

Топ-10 наиболее важных дескрипторов для IC50:

	0
VSA_EState6	0.399126
VSA_EState9	0.083731
FpDensityMorgan1	0.033675
MinPartialCharge	0.026929
BalabanJ	0.026566
Chi1	0.023823
MaxPartialCharge	0.021419
qed	0.017184
Chi1n	0.016737
Kappa3	0.016715

dtype: float64





Результаты моделей (таргет: SI):

	Model	MSE	RMSE	MAE	r2
0	KNN	1.221340e+06	1105.142410	181.916912	0.116976
1	Random Forest	1.620292e+06	1272.906985	213.540525	-0.171466
2	Gradient Boosting	2.489946e+06	1577.956395	257.932775	-0.800223
3	HistGradientBoosting	1.284243e+06	1133.244456	189.409971	0.071497
4	AdaBoostRegressor	2.480675e+06	1575.016015	285.854219	-0.793520
5	XGBoost	1.984235e+06	1408.628915	202.759415	-0.434595
6	CatBoost	2.423124e+06	1556.638701	253.662441	-0.751910

Наилучший результат

	Model	MSE	RMSE	MAE	r2
0	KNN	1.221340e+06	1105.14241	181.916912	0.116976

5 Классификация IC50 медианное значение выборки

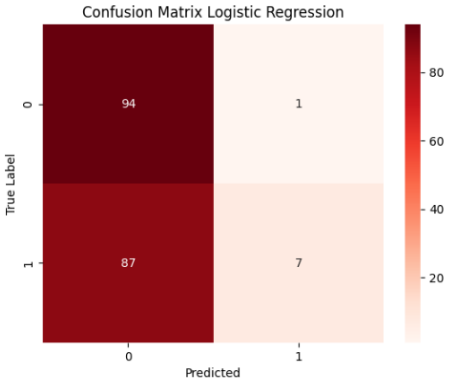
Добавляем новые данные.

а. iC50_Median

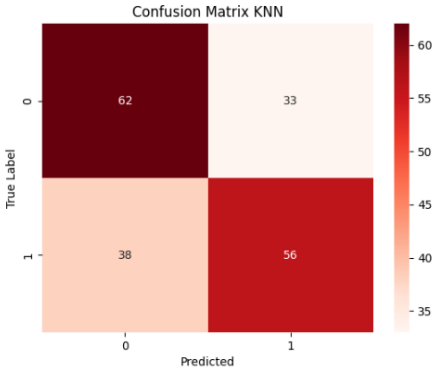
```
Получили новые признаки: ['MolLogP_MolWt', 'MolLogP^2', 'MolLogP MolWt', 'MolWt^2', 'MolLogP_gt_3']
```

б.

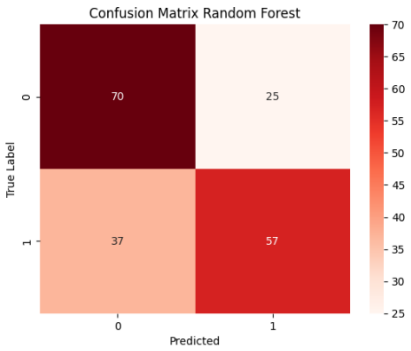
Logistic Regression



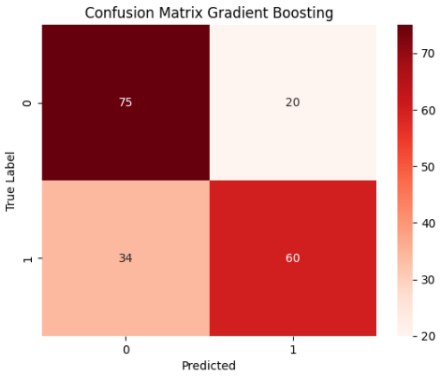
KNN



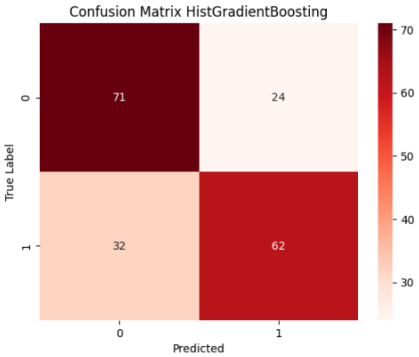
Random Forest



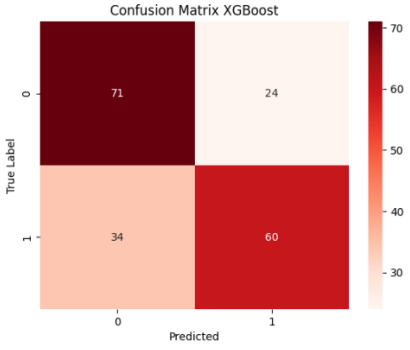
Gradient Boosting



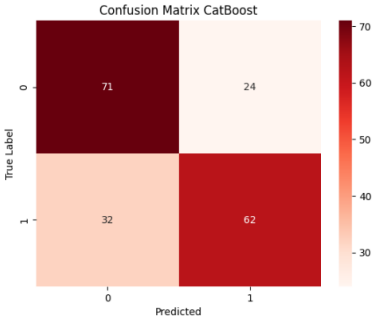
HistGradientBoosting

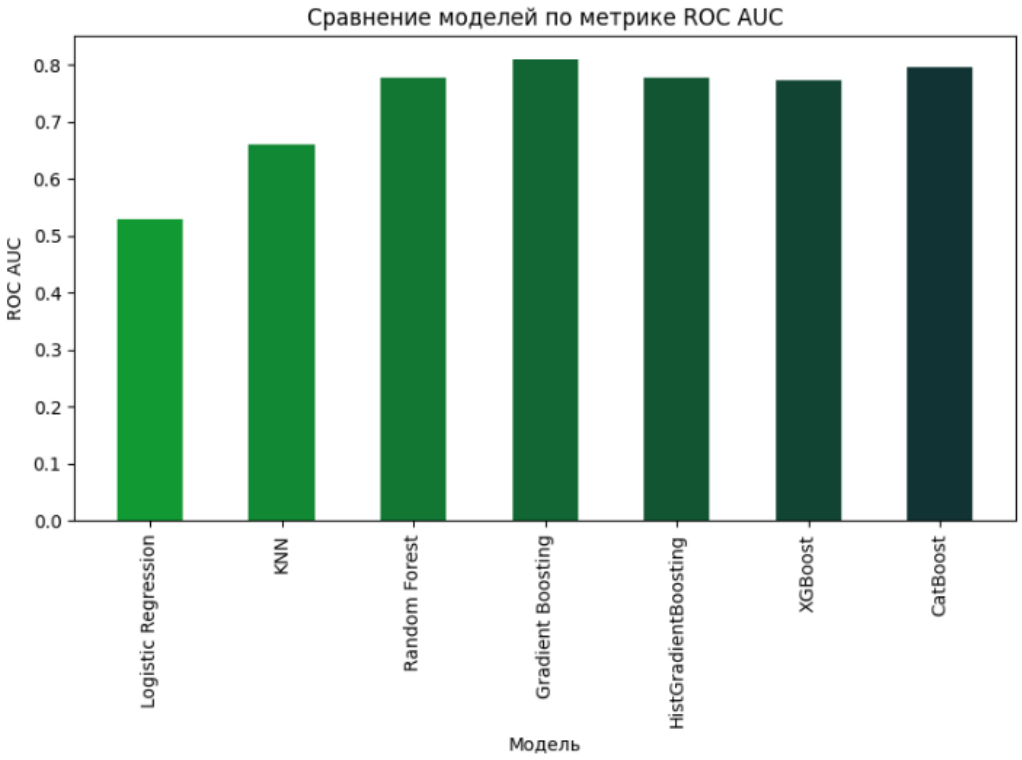


XGBoost



CatBoost





Результаты классификации:

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
0	Logistic Regression	0.875000	0.074468	0.137255	0.534392	0.529395
1	KNN	0.629213	0.595745	0.612022	0.624339	0.659966
2	Random Forest	0.695122	0.606383	0.647727	0.671958	0.776596
3	Gradient Boosting	0.750000	0.638298	0.689655	0.714286	0.809462
4	HistGradientBoosting	0.720930	0.659574	0.688889	0.703704	0.777324
5	XGBoost	0.714286	0.638298	0.674157	0.693122	0.773292
6	CatBoost	0.720930	0.659574	0.688889	0.703704	0.794905

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
3	Gradient Boosting	0.75	0.638298	0.689655	0.714286	0.809462

6 Классификация CC50 медианное значение выборки.

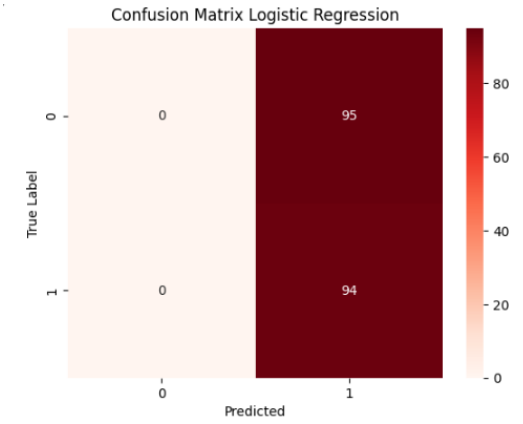
Добавляем новые данные.

а. CC50_Median

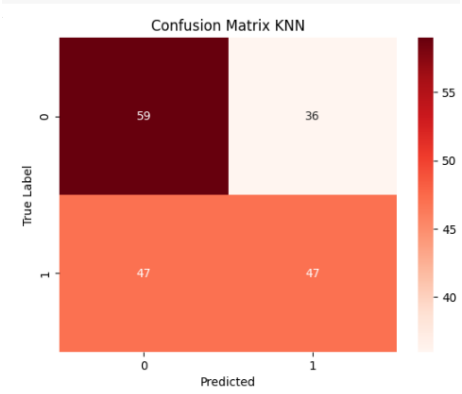
```
Получили новые признаки: ['MolLogP_MolWt', 'MolLogP^2', 'MolLogP MolWt', 'MolWt^2', 'MolLogP_gt_3']
```

б.

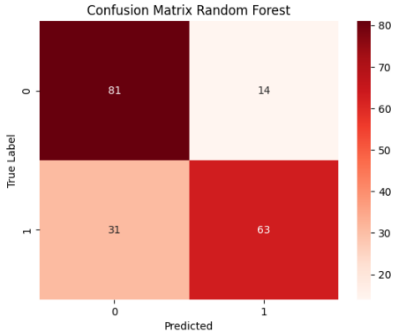
Logistic Regression



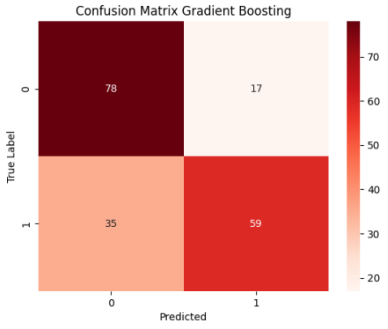
KNN



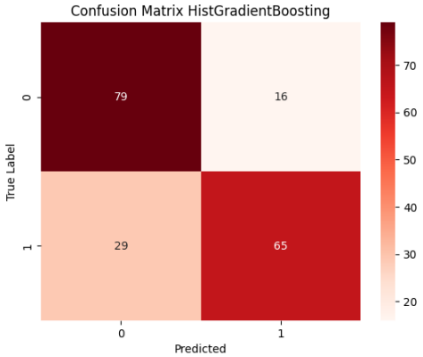
Random Forest



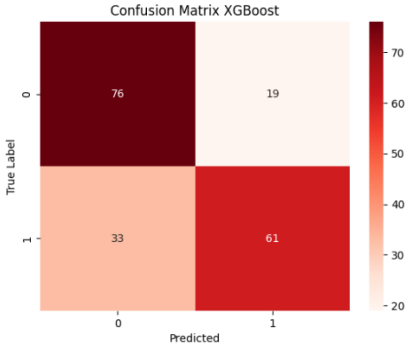
Gradient Boosting



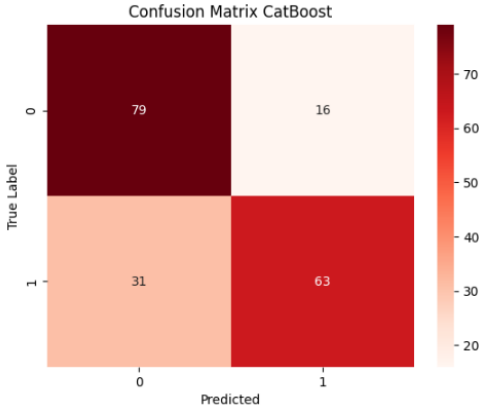
HistGradientBoosting

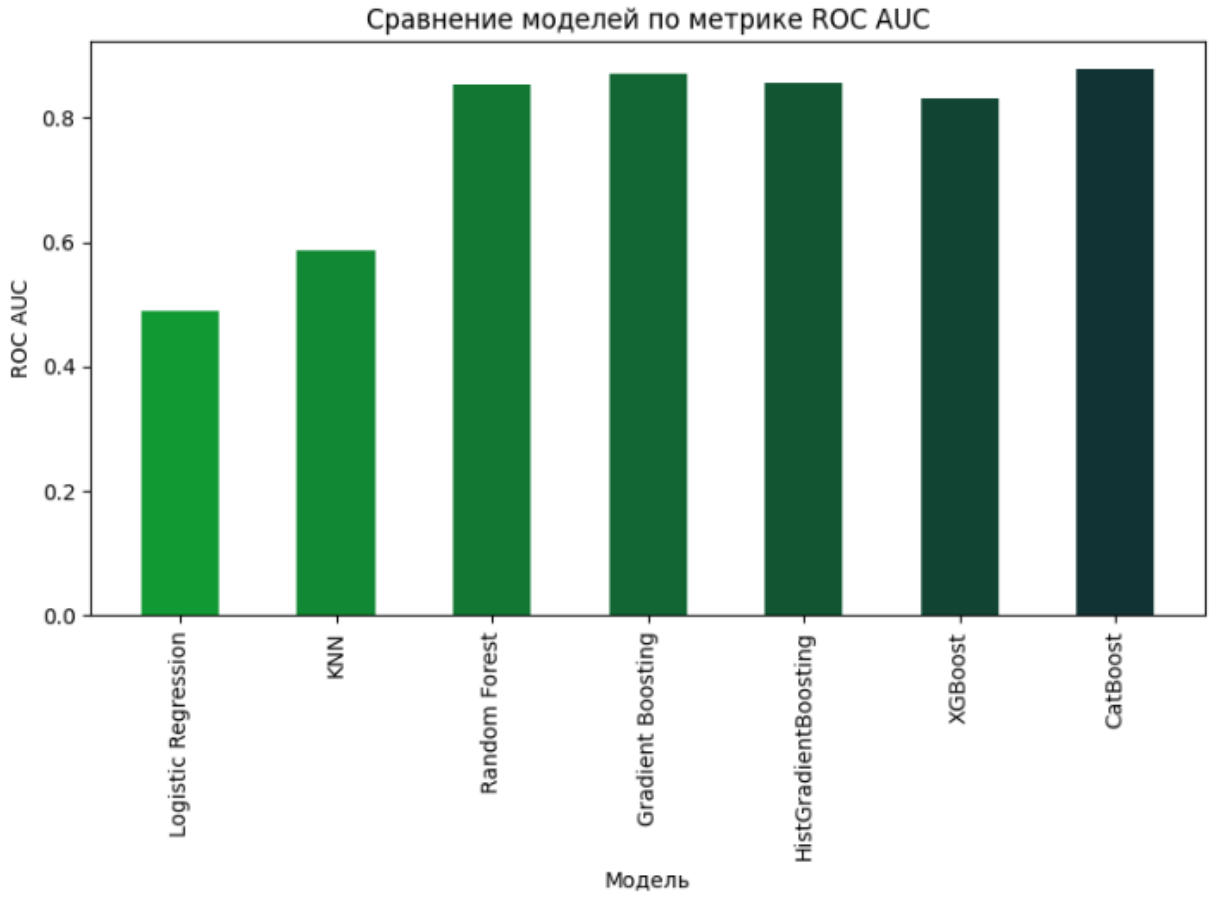


XGBoost



CatBoost





Результаты классификации:

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
0	Logistic Regression	0.497354	1.000000	0.664311	0.497354	0.489362
1	KNN	0.566265	0.500000	0.531073	0.560847	0.585946
2	Random Forest	0.818182	0.670213	0.736842	0.761905	0.852632
3	Gradient Boosting	0.776316	0.627660	0.694118	0.724868	0.870773
4	HistGradientBoosting	0.802469	0.691489	0.742857	0.761905	0.855879
5	XGBoost	0.762500	0.648936	0.701149	0.724868	0.831691
6	CatBoost	0.797468	0.670213	0.728324	0.751323	0.878052

Наилучший результат

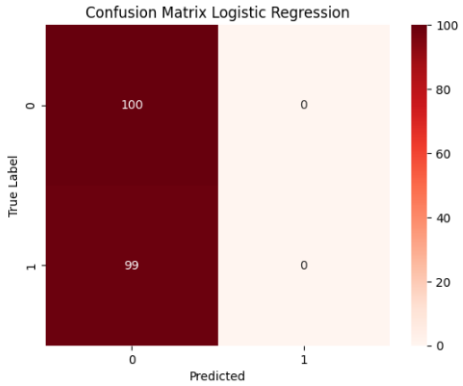
	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.797468	0.670213	0.728324	0.751323	0.878052

07 Классификация SI медианное значение выборки
Добавляем новые данные.
а. SI_Median

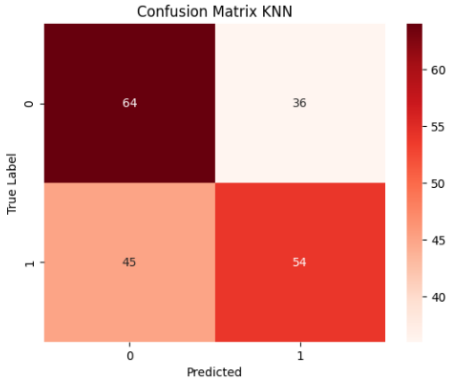
```
Получили новые признаки: ['MolLogP_MolWt', 'MolLogP^2', 'MolLogP MolWt', 'MolWt^2', 'MolLogP_gt_3']
```

б.

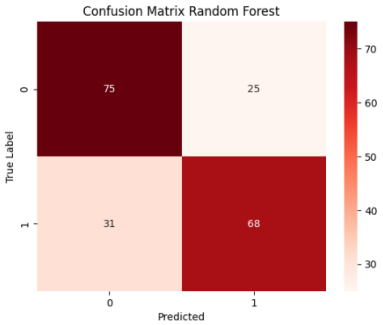
Logistic Regression



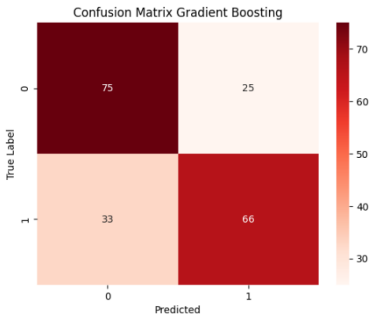
KNN



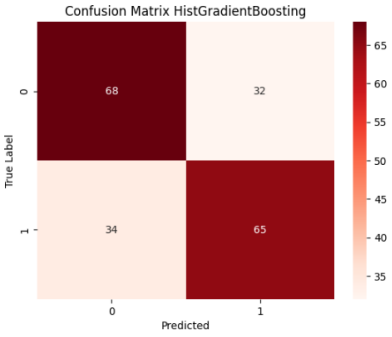
Random Forest



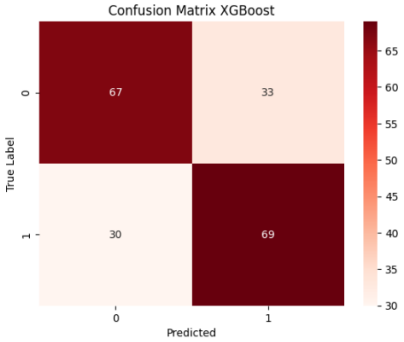
Gradient Boosting



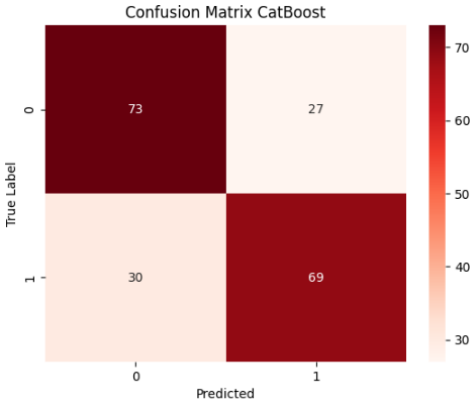
HistGradientBoosting

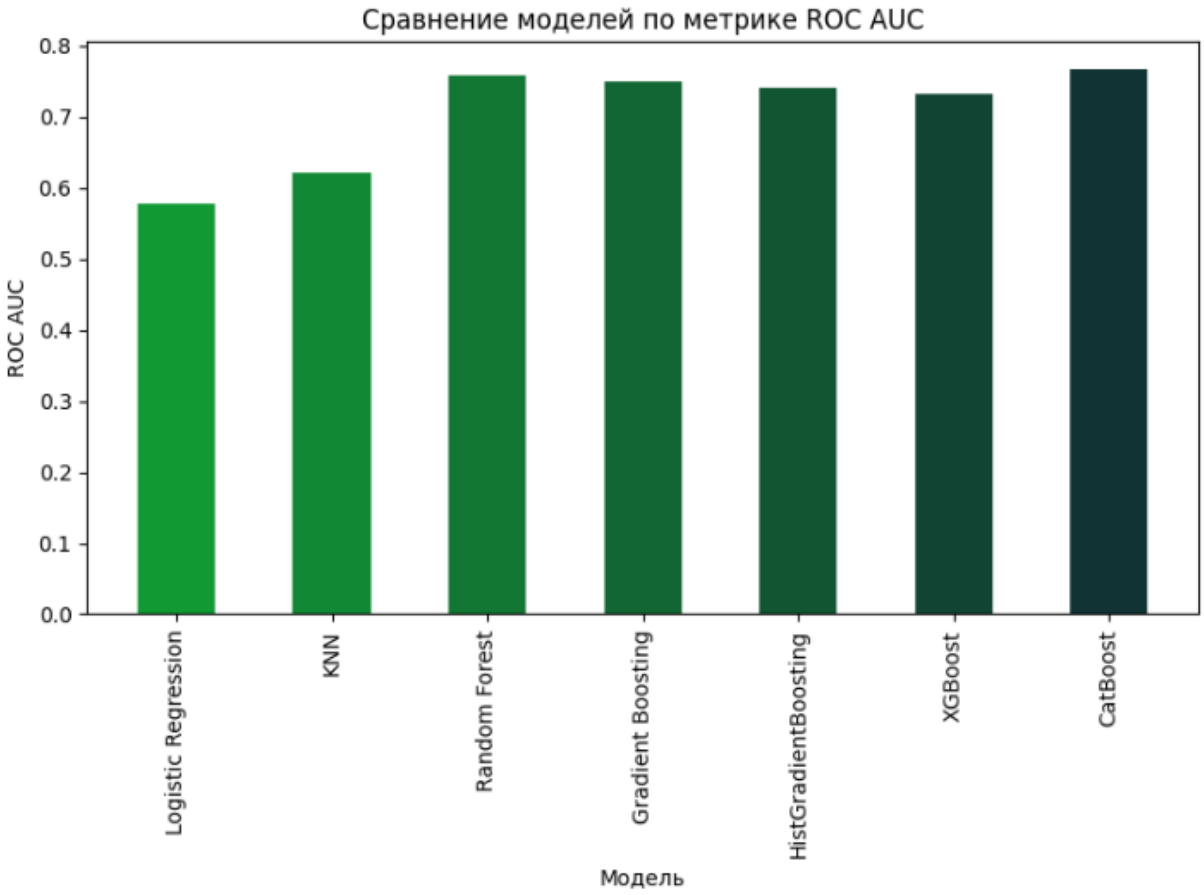


XGBoost



CatBoost





Результаты классификации:

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
0	Logistic Regression	0.000000	0.000000	0.000000	0.502513	0.576869
1	KNN	0.600000	0.545455	0.571429	0.592965	0.621414
2	Random Forest	0.731183	0.686869	0.708333	0.718593	0.759040
3	Gradient Boosting	0.725275	0.666667	0.694737	0.708543	0.749192
4	HistGradientBoosting	0.670103	0.656566	0.663265	0.668342	0.740707
5	XGBoost	0.676471	0.696970	0.686567	0.683417	0.731566
6	CatBoost	0.718750	0.696970	0.707692	0.713568	0.767374

Наилучший результат

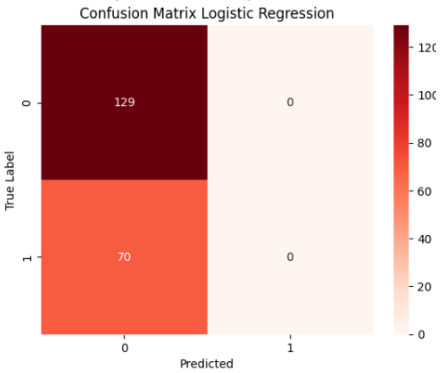
	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.71875	0.69697	0.707692	0.713568	0.767374

8 Классификация превышает ли значение SI значение 8
Добавляем новые данные.
а. SI_8

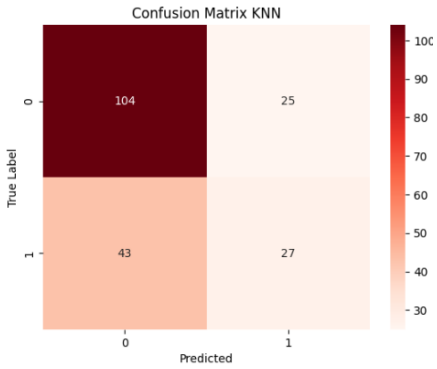
```
Получили новые признаки: ['MolLogP_MolWt', 'MolLogP^2', 'MolLogP MolWt', 'MolWt^2', 'MolLogP_gt_3']
```

б.

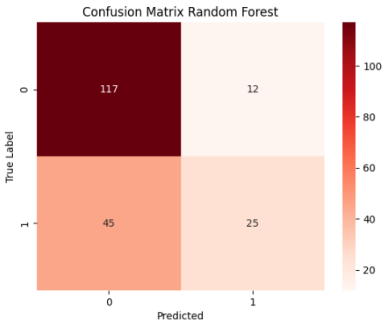
Logistic Regression



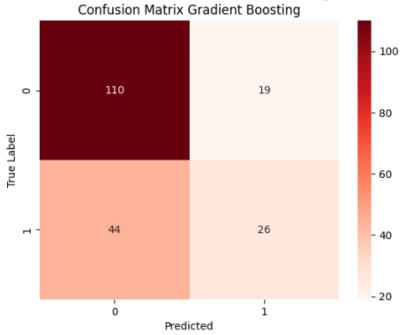
KNN



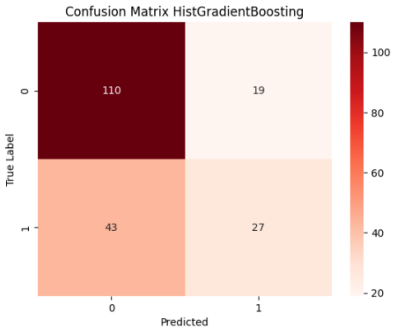
Random Forest



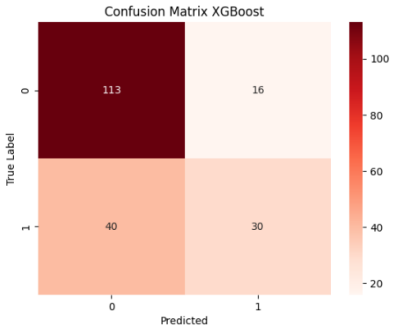
Gradient Boosting



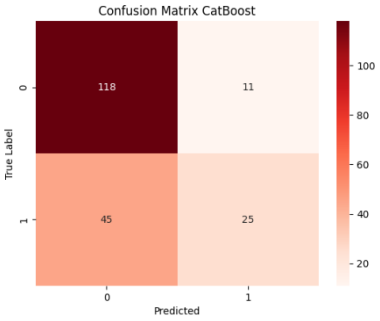
HistGradientBoosting

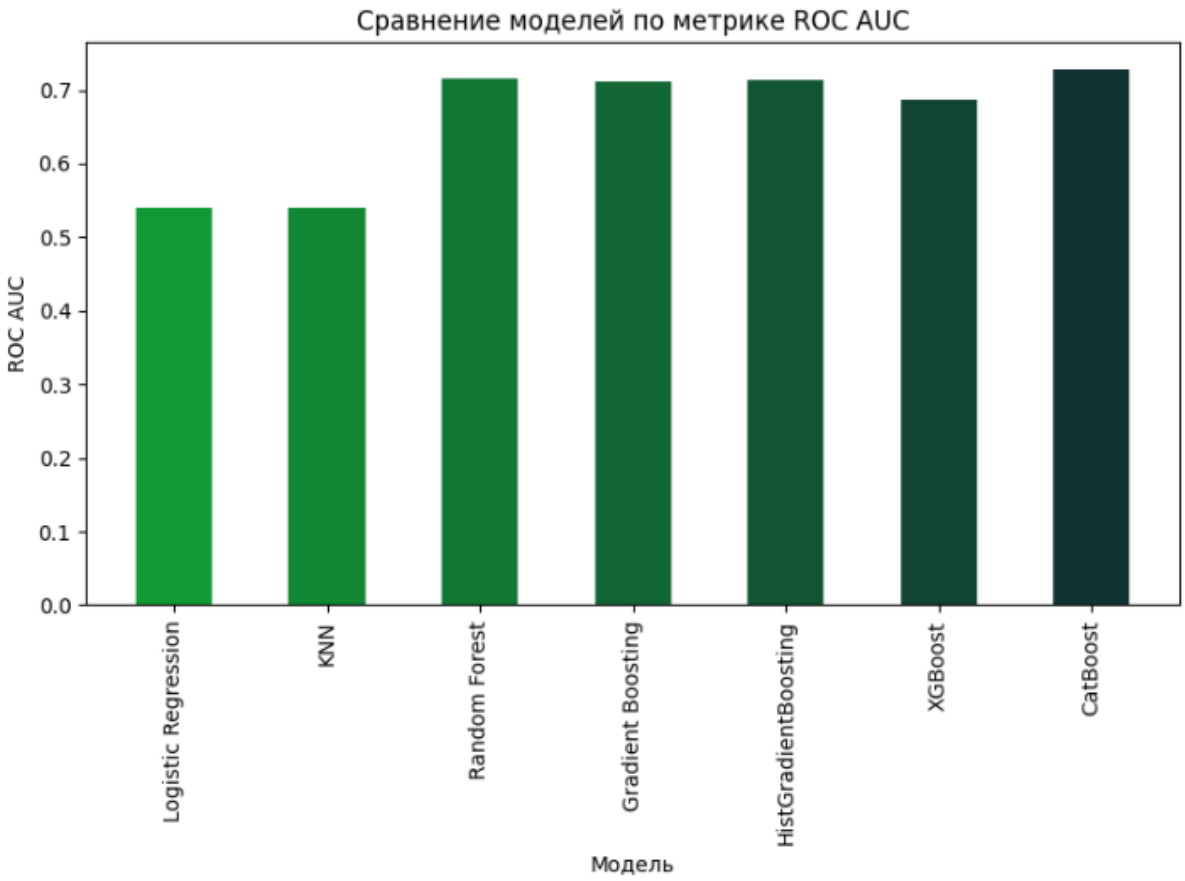


XGBoost



CatBoost





Результаты классификации:

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
0	Logistic Regression	0.000000	0.000000	0.000000	0.648241	0.540310
1	KNN	0.519231	0.385714	0.442623	0.658291	0.539313
2	Random Forest	0.675676	0.357143	0.467290	0.713568	0.716334
3	Gradient Boosting	0.577778	0.371429	0.452174	0.683417	0.711406
4	HistGradientBoosting	0.586957	0.385714	0.465517	0.688442	0.713400
5	XGBoost	0.652174	0.428571	0.517241	0.718593	0.686711
6	CatBoost	0.694444	0.357143	0.471698	0.718593	0.728018

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.694444	0.357143	0.471698	0.718593	0.728018

Выводы:

У регрессии IC50

Наилучший результат

	Model	MSE	RMSE	MAE	r2
3	HistGradientBoosting	95274.069624	308.66498	191.231119	0.335216

У регрессии CC50

Наилучший результат

	Model	MSE	RMSE	MAE	r2
3	HistGradientBoosting	150649.53105	388.13597	278.924685	0.634788

У регрессии SI

Наилучший результат

	Model	MSE	RMSE	MAE	r2
0	KNN	1.221340e+06	1105.14241	181.916912	0.116976

У классификации IC50 медианное значение выборки

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
3	Gradient Boosting	0.75	0.638298	0.689655	0.714286	0.809462

У классификации CC50 медианное значение выборки.

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.797468	0.670213	0.728324	0.751323	0.878052

У классификации SI медианное значение выборки

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.71875	0.69697	0.707692	0.713568	0.767374

У классификации превышающее значение SI значение 8

Наилучший результат

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC
6	CatBoost	0.694444	0.357143	0.471698	0.718593	0.728018