Курсовая работа (vo_PJ)

Исследование лекарственной активности

Аналитический отчёт

Белянини Роман Сергеевич

Цель исследования

На основе предоставленных данных о химических соединениях спрогнозировать их эффективность с целью подбора оптимального состава лекарственного препарата. Основное внимание уделяется прогнозированию ключевых показателей активности (IC50, CC50, SI) и классификации соединений на «сильные» и «слабые» ингибиторы.

Задачи исследования

1. Выполнить исследовательский анализ данных (EDA) и оценить информативность признаков.
2. Построить и обучить модели машинного обучения:
   Регрессионные задачи
   • прогноз IC50;
   • прогноз CC50;
   • прогноз SI.

   Классификационные задачи
   • бинарный прогноз: превышает ли IC50 медианное значение;
   • бинарный прогноз: превышает ли CC50 медианное значение;
   • бинарный прогноз: превышает ли SI медианное значение;
   • бинарный прогноз: превышает ли SI порог 8.
3. Оценить и сравнить качество моделей по соответствующим метрикам (например, RMSE и MAE для регрессии; Accuracy, ROC-AUC, F1-score для классификации) и выбрать лучшие решения.

Целевые переменные

- **IC_{50} (мМ)** — концентрация соединения, необходимая для подавления вирусной активности на 50 %.
- **CC_{50} (мМ)** — концентрация соединения, вызывающая гибель 50 % клеток (цитотоксичность).
- **SI (Selectivity Index)** — индекс селективности, рассчитываемый как отношение CC_{50} к IC_{50}; чем выше значение, тем более селективен препарат.

Исследовательский анализ (EDA)

Описание датасета

Датасет представляет собой таблицу, содержащую данные о **1 001** химическом соединении. Каждая строка соответствует одному веществу, а столбцы — его физико-химическим признакам и показателям биологической активности.

Состав признаков:

- **107 числовых признаков** (*float64*)
- **107 целочисленных признаков** (*int64*)

Эти признаки описывают структурные, физико-химические и молекулярные свойства соединений. Данные будут использованы для построения моделей регрессии и классификации, а также для оценки их селективности.

В датасете присутствуют пропуски в размере примерно 30% от данных

```
MaxPartialCharge       0.2997
MinPartialCharge       0.2997
MaxAbsPartialCharge    0.2997
MinAbsPartialCharge    0.2997
BCUT2D_MWHI            0.2997
BCUT2D_MWLOW           0.2997
BCUT2D_CHGHI           0.2997
BCUT2D_CHGLO           0.2997
BCUT2D_LOGPHI          0.2997
BCUT2D_LOGPLOW         0.2997
BCUT2D_MRHI            0.2997
BCUT2D_MRLOW           0.2997
dtype: float64
```

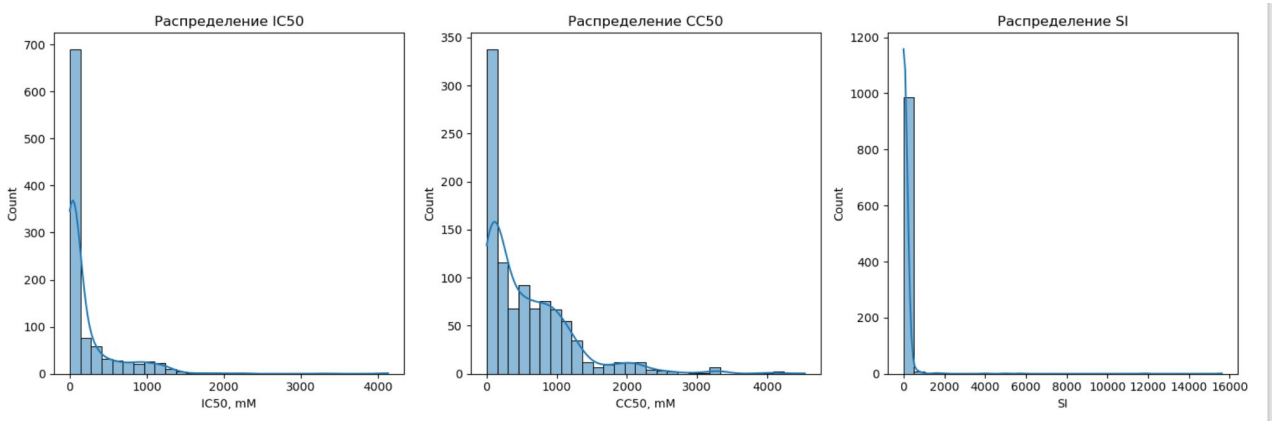
Данные пропуски заполняем медианой

Так же в данных присутствуют данные с 1 уникальным значением

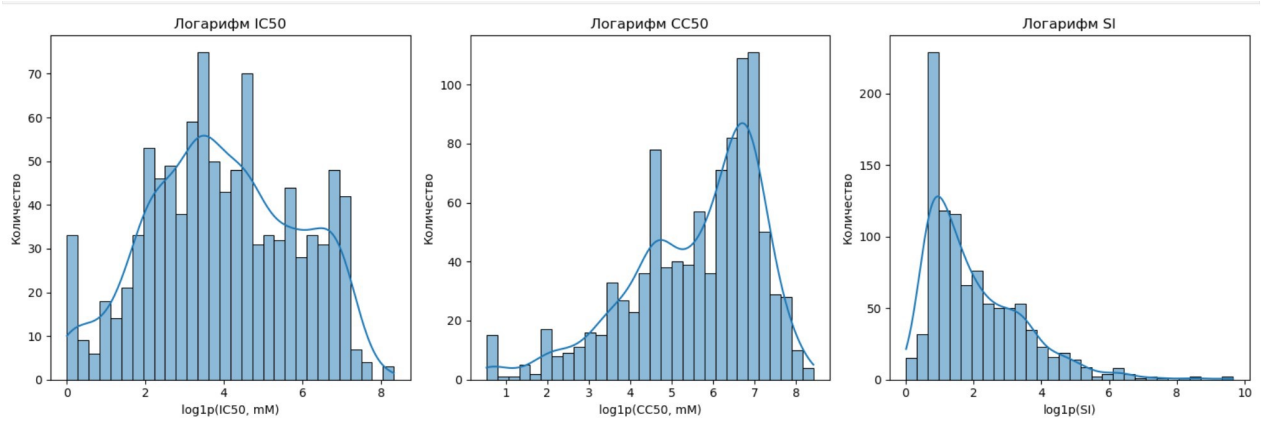
Столбцы заполненые константой ['NumRadicalElectrons', 'SMR_VSA8', 'SlogP_VSA9', 'fr_N_O', 'fr_SH', 'fr_azide', 'fr_barbitur', 'fr_benzodiazepine', 'fr_diazo', 'fr_dihydropyridine', 'fr_isocyan', 'fr_isothiocyan', 'fr_lactam', 'fr_nitroso', 'fr_phos_acid', 'fr_phos_ester', 'fr_prisulfonamd', 'fr_thiocyan']

Удалим их из данных

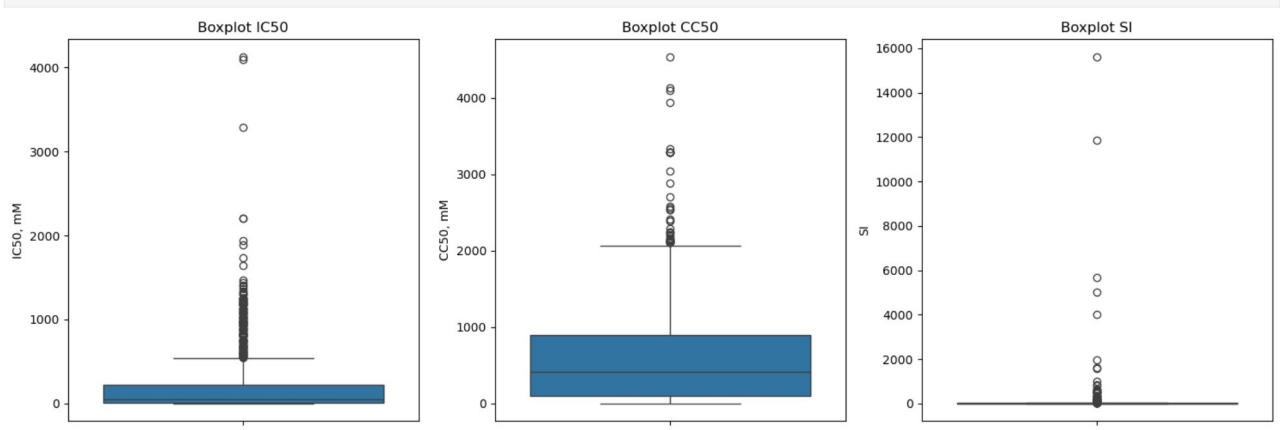
Анализ распределения целевых переменных



Большинство соединений демонстрируют низкие значения активности/цитотоксичности, а высокая активность встречается редко. Для корректного моделирования полезны логарифмические преобразования.



На графиках видны выбросы на всех трех целевых переменных



Обработка выбросов

Для смягчения влияния аномально высоких значений протестированы два подхода:

1. **Ограничение верхних порогов.** Значения выше IC_{50} > **1 200 мМ**, CC_{50} >

2 500 мМ и SI > **250** заменялись пороговыми. Метод сохраняет все строки, быстро устраняя экстремальные пики.

2. **Исключение выбросов при помощи Isolation Forest.** Алгоритм выявил и удалил ≈ **2 %** наблюдений, что привело к среднему росту регрессионных метрик на ≈ **3 %**.

Таким образом, жертвуя 2 % данных, можно добиться заметного улучшения качества модели. В контексте задачи это считается целесообразным компромиссом между полнотой выборки и точностью прогнозов.

Итоговый размер датасета (980, 196)

Регрессия для CC50

Выбранные алгоритмы регрессии:

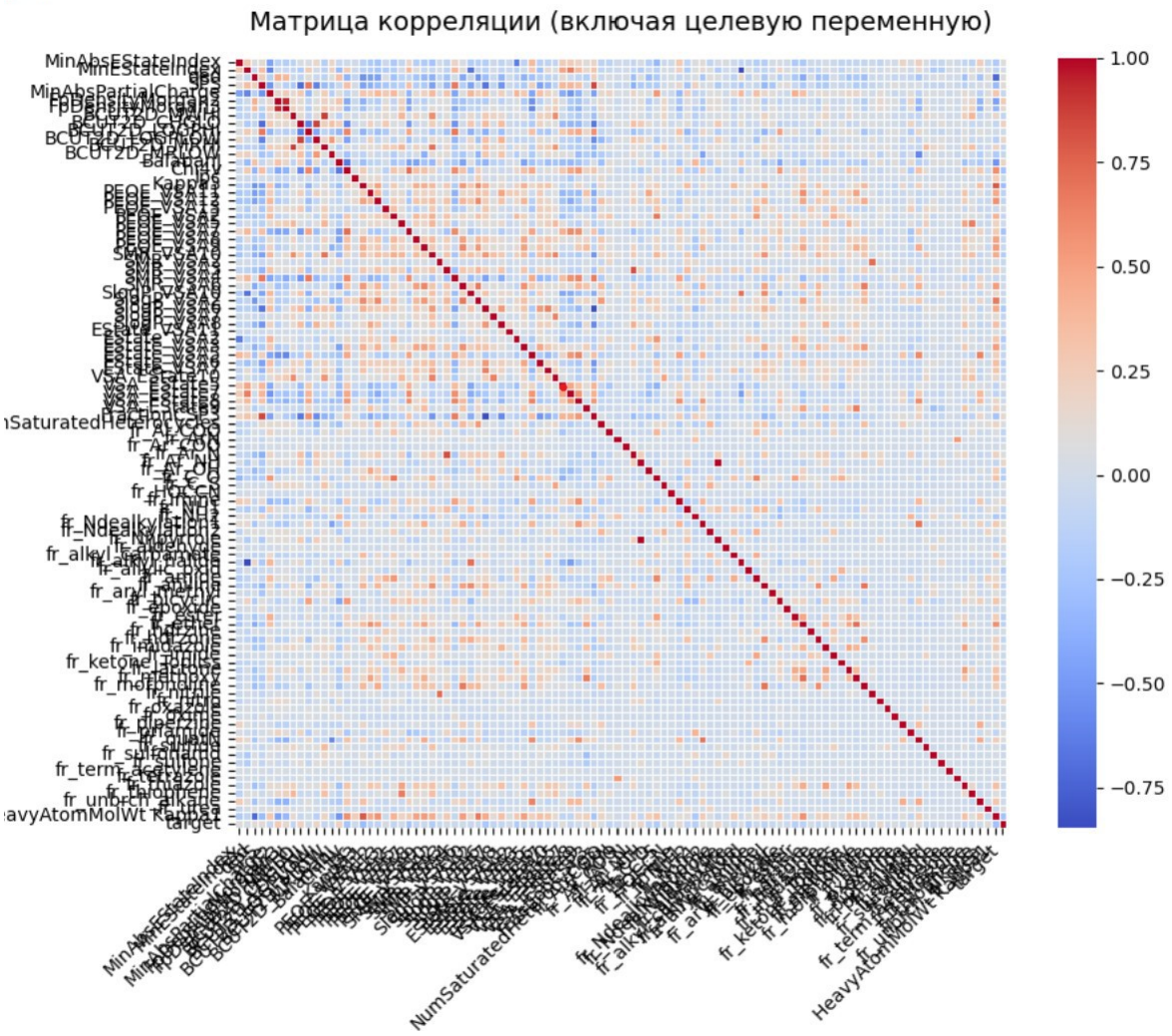
- LinearRegression;
- RandomForestRegressor;
- ExtraTreesRegressor;
- HistGradientBoostingRegressor;
- XGBoostRegressor;
- LightGBMRegressor;
- CatBoostRegressor.

Расширение признакового пространства

1. PolynomialFeatures до второй степени. Из сгенерированного набора выбираются 20 наиболее коррелированных с целевой переменной признаков (|r| по Пирсону).
2. Базовые + отобранные полиномиальные признаки объединяются в единый датафрейм.
3. На объединённом наборе выполняется Lasso (α подбирается перекрёстной проверкой), что окончательно отбирает наиболее информативные переменные и снижает мультиколлинеарность.

Итог: получаем компактный, но информативный набор признаков, на котором обучаются все вышеперечисленные модели.

```
Топ-10 признаков по модулю корреляции с целевой переменной:
MolMR             -0.309959
LabuteASA         -0.309096
MolWt             -0.306069
ExactMolWt        -0.306010
HeavyAtomCount    -0.305167
Chi1              -0.304619
Chi0              -0.304326
HeavyAtomMolWt    -0.302825
Chi1v             -0.301270
Kappa1            -0.301159
Name: target, dtype: float64
Топ-20 новых полиномиальных признаков по корреляции с целевой переменной:
MolMR HeavyAtomMolWt        0.262604
MolMR MolWt                 0.261554
MolMR ExactMolWt            0.261535
MolMR Chi0                  0.261474
LabuteASA Chi0              0.260893
Chi0^2                      0.260609
MolWt Chi0                  0.260415
Chi0 HeavyAtomMolWt         0.260393
ExactMolWt Chi0             0.260386
LabuteASA HeavyAtomMolWt    0.260191
MolMR HeavyAtomCount        0.260164
LabuteASA MolWt             0.260140
LabuteASA ExactMolWt        0.260121
LabuteASA^2                 0.259976
HeavyAtomCount Chi0         0.259675
LabuteASA HeavyAtomCount    0.259485
MolMR LabuteASA             0.259466
MolMR Chi1                  0.259123
HeavyAtomMolWt Kappa1       0.259028
HeavyAtomMolWt Chi1v        0.258978
Name: target, dtype: float64
```



Результаты обучения моделей (R^2, кросс-валидация)

Модель	R^2 (среднее)	R^2 (ст. отклонение)
LinearRegression	-1.27e+21	± 2.54e+21
RandomForest	0.5014	± 0.0996
ExtraTrees	0.4846	± 0.1251
HistGBR	0.5229	± 0.0940
XGBoost	0.4483	± 0.1296
LightGBM	0.5140	± 0.0911
CatBoost	0.5105	± 0.1034

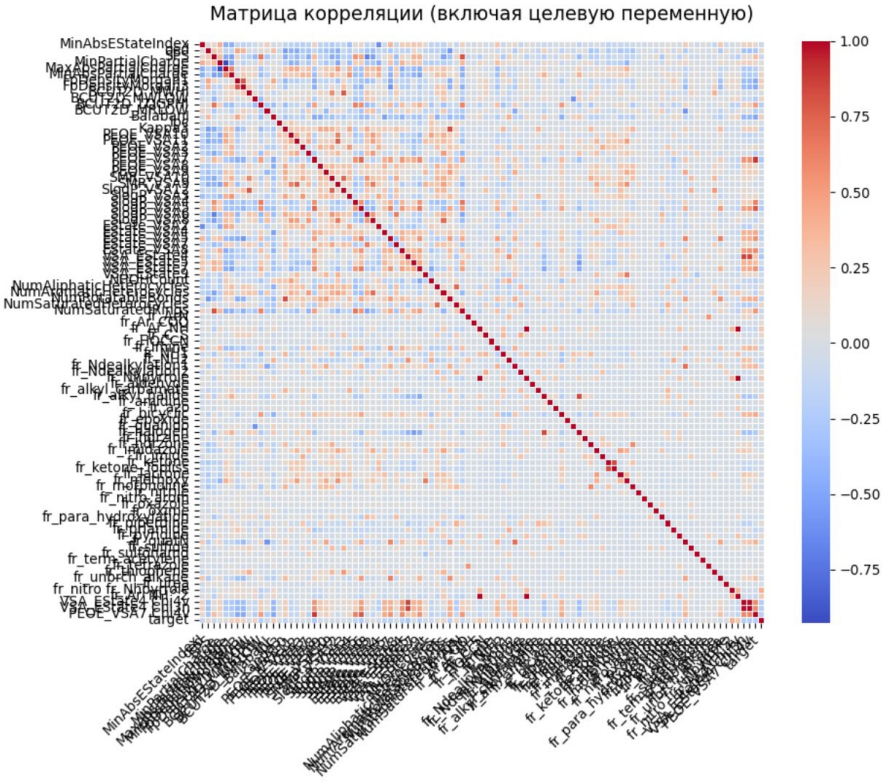
- LinearRegression показывает аномально большое отрицательное значение R^2 — вероятно, из-за несоответствия предпосылок линейной модели (например, сильной мультиколлинеарности или влияния выбросов).
- Деревья и бустинг-алгоритмы демонстрируют R^2 около 0.45–0.52 с допустимой вариацией, что подтверждает их устойчивость и способность объяснять данные лучше среднего прогноза.

Для подбора гиперпараметров была выбрана модель LightGBMRegressor, показавшая хорошее сочетание качества ($R^2 \approx 0.51$) и скорости обучения. В качестве инструмента для автоматизированного поиска оптимальных настроек применялся фреймворк Optuna.

Финальные метрики RMSE: 437.75 MAE : 282.50 R^2 : 0.5559

Регрессия для IC50

Используем такие же алгоритмы регрессии что и для CC50.



```
Топ-10 признаков по модулю корреляции с целевой переменной:
VSA_EState4    -0.277346
fr_nitro        0.261613
Chi2n          -0.259928
PEOE_VSA7      -0.258582
fr_Nhpyrrole    0.251986
fr_Ar_NH        0.251986
Chi2v          -0.250932
Chi4v          -0.247722
Chi4n          -0.247521
Chi3n          -0.242723
Name: target, dtype: float64
Топ-20 новых полиномиальных признаков по корреляции с целевой переменной:
fr_nitro fr_Nhpyrrole    0.414983
fr_nitro fr_Ar_NH        0.414983
VSA_EState4 fr_nitro     0.317588
fr_nitro^2               0.261613
fr_Ar_NH^2               0.251986
fr_Nhpyrrole^2           0.251986
fr_Nhpyrrole fr_Ar_NH    0.251986
VSA_EState4 Chi2v        0.230966
VSA_EState4 Chi2n        0.229980
VSA_EState4 Chi4v        0.225399
VSA_EState4 Chi4n        0.224423
VSA_EState4 Chi3n        0.222733
PEOE_VSA7 Chi2v          0.212331
Chi2n PEOE_VSA7          0.209329
PEOE_VSA7 Chi4v          0.208155
Chi2n Chi2v              0.205365
Chi2v^2                  0.205329
Chi2v Chi4v              0.205162
Chi2n Chi4v              0.203970
PEOE_VSA7 Chi4n          0.203385
Name: target, dtype: float64
```

Модель	R^2 (среднее)	R^2 (ст. отклонение)
LinearRegression	-4.97e+18	± 9.93e+18
RandomForest	0.2467	± 0.1984
ExtraTrees	0.1693	± 0.2883
HistGBR	0.2725	± 0.1584
XGBoost	0.0909	± 0.3575
LightGBM	0.2729	± 0.1543
CatBoost	0.2193	± 0.2517

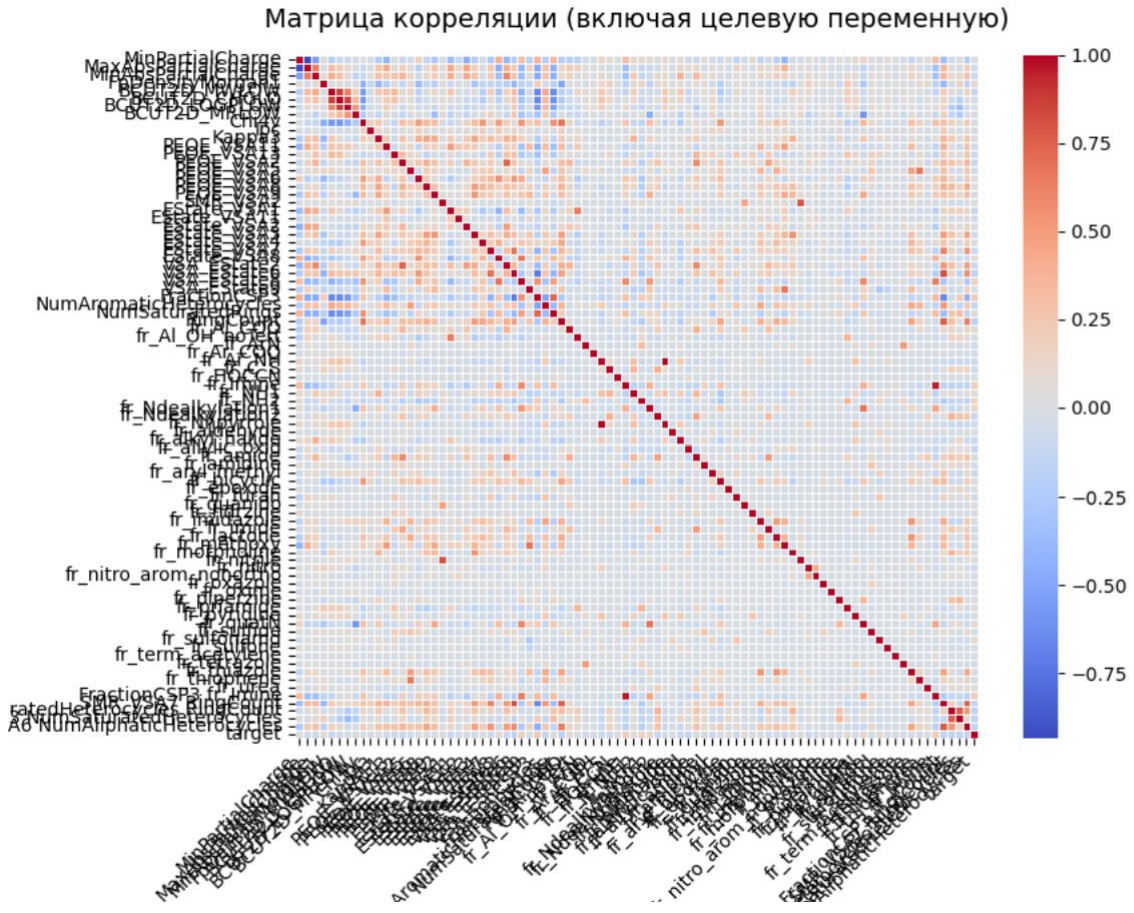
Для подбора гиперпараметров была выбрана модель LightGBMRegressor

Финальные метрики RMSE: 392.33 MAE : 212.70 R^2 : 0.4646

Регрессия для SI

Используем такие же алгоритмы регрессии что и для CC50. Только добавляем логарифмирование целевой переменной что увеличивает R^2 примерно на 0.10

```
Топ-10 признаков по модулю корреляции с целевой переменной:
FractionCSP3                0.224561
SMR_VSA7                   -0.217856
SlogP_VSA6                 -0.206151
NumAliphaticHeterocycles   -0.200207
fr_Imine                    0.193947
NumSaturatedHeterocycles   -0.193116
RingCount                  -0.189473
VSA_EState6                -0.187608
NumAromaticRings           -0.187272
VSA_EState8                 0.184113
Name: target, dtype: float64
Топ-20 новых полиномиальных признаков по корреляции с целевой переменной:
FractionCSP3 fr_Imine                     0.219658
FractionCSP3^2                            0.212699
FractionCSP3 VSA_EState8                  0.199942
SMR_VSA7 SlogP_VSA6                       0.182757
SlogP_VSA6^2                              0.181976
SMR_VSA7^2                                0.176943
FractionCSP3 NumAliphaticHeterocycles     0.175708
SlogP_VSA6 RingCount                      0.174153
SMR_VSA7 RingCount                        0.169456
NumSaturatedHeterocycles RingCount        0.169340
SlogP_VSA6 NumAromaticRings               0.168672
FractionCSP3 NumSaturatedHeterocycles     0.166133
SlogP_VSA6 VSA_EState6                    0.164637
SMR_VSA7 VSA_EState6                      0.163809
fr_Imine^2                                0.163756
SMR_VSA7 NumAromaticRings                 0.160879
VSA_EState6 NumAromaticRings              0.159601
RingCount VSA_EState6                     0.154672
NumAliphaticHeterocycles^2                0.153191
SlogP_VSA6 NumAliphaticHeterocycles       0.152064
Name: target, dtype: float64
```



Модель	R^2 (среднее)	R^2 (ст. отклонение)
LinearRegression	-1.96e+20	± 3.92e+20
RandomForest	0.3162	± 0.0660
ExtraTrees	0.2447	± 0.0629
HistGBR	0.2650	± 0.0714
XGBoost	0.2168	± 0.0613
LightGBM	0.2583	± 0.0759
CatBoost	0.2851	± 0.0728

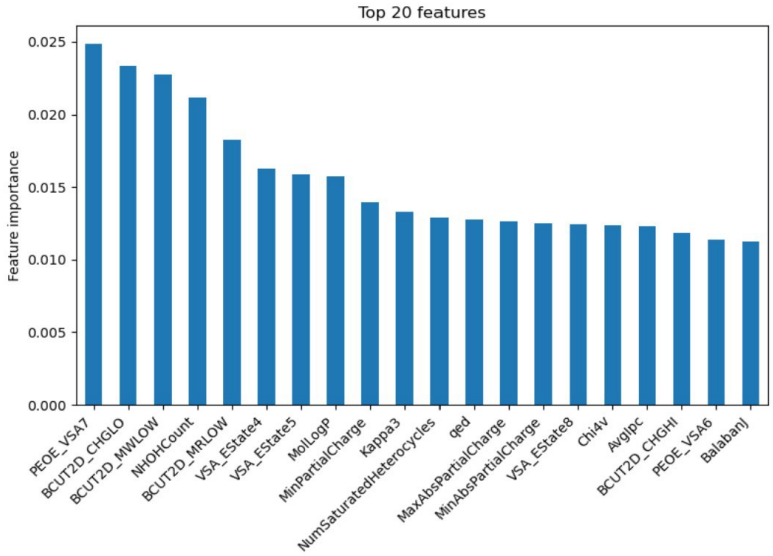
Для подбора гиперпараметров была выбрана модель RandomForestRegressor

Финальные метрики RMSE: 1.05 MAE : 0.83 R^2 : 0.3831

Классификация: превышает ли значение CC50 медианное значение выборки

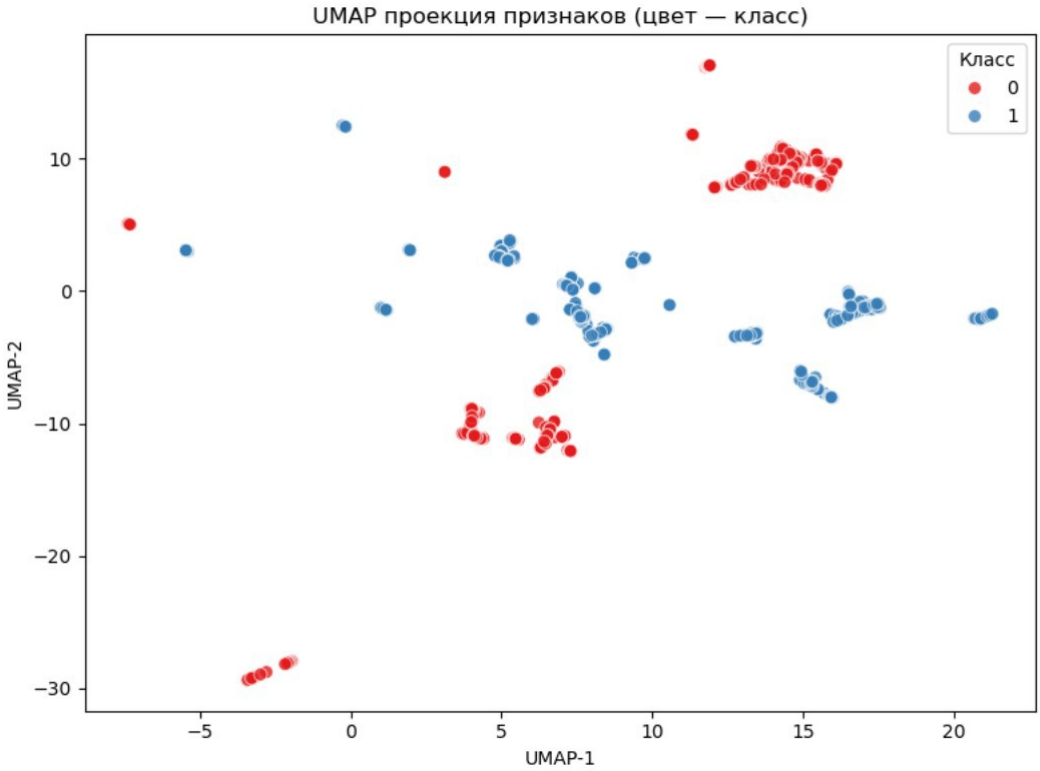
Выбранные алгоритмы классификации:

- RandomForestClassifier;
- ExtraTreesClassifier;
- HistGradientBoostingClassifier;
- XGBoostClassifier;
- LightGBMClassifier;
- CatBoostClassifier.



Метод отбора признаков:

Для сокращения размерности данных используется SelectFromModel с порогом в виде медианы важностей признаков. Автоматически отбираются только признаки с важностью выше медианы.



Красные и синие точки в некоторых областях хорошо разделены (например, плотный красный кластер в правом верхнем углу и плотный синий кластер ближе к центру).

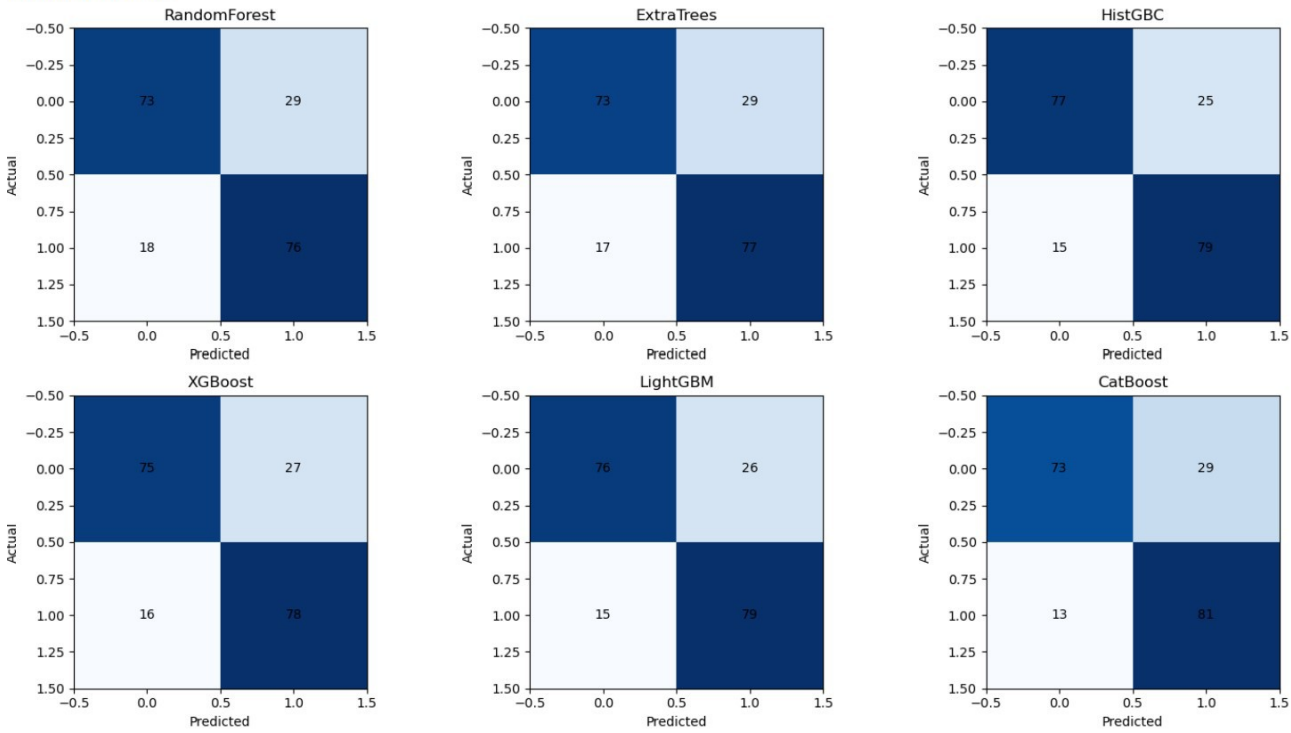
- Однако значительное число объектов разных классов перемешаны между собой (в центральной области проекции).
- Это указывает, что полное разделение классов по выбранным признакам затруднено.

Модель	Accuracy	Precision	Recall	F1 Score	ROC AUC
RandomForestClassifier	0.753 ± 0.033	0.749 ± 0.026	0.768 ± 0.063	0.757 ± 0.038	0.837 ± 0.018
ExtraTreesClassifier	0.758 ± 0.021	0.766 ± 0.028	0.753 ± 0.040	0.758 ± 0.022	0.826 ± 0.025
HistGBC	0.742 ± 0.030	0.742 ± 0.028	0.750 ± 0.037	0.746 ± 0.031	0.836 ± 0.023
XGBoostClassifier	0.746 ± 0.025	0.742 ± 0.012	0.763 ± 0.056	0.751 ± 0.033	0.831 ± 0.024
LightGBMClassifier	0.759 ± 0.027	0.763 ± 0.025	0.758 ± 0.050	0.760 ± 0.032	0.833 ± 0.018
CatBoostClassifier	0.758 ± 0.017	0.751 ± 0.020	0.780 ± 0.053	0.764 ± 0.023	0.842 ± 0.018

Сводная таблица метрик на тестовой выборке:

	accuracy	precision	recall	f1	roc_auc
HistGBC	0.796	0.760	0.840	0.798	0.860
LightGBM	0.791	0.752	0.840	0.794	0.860
CatBoost	0.786	0.736	0.862	0.794	0.874
XGBoost	0.781	0.743	0.830	0.784	0.858
ExtraTrees	0.765	0.726	0.819	0.770	0.841
RandomForest	0.760	0.724	0.809	0.764	0.852

Confusion matrices:

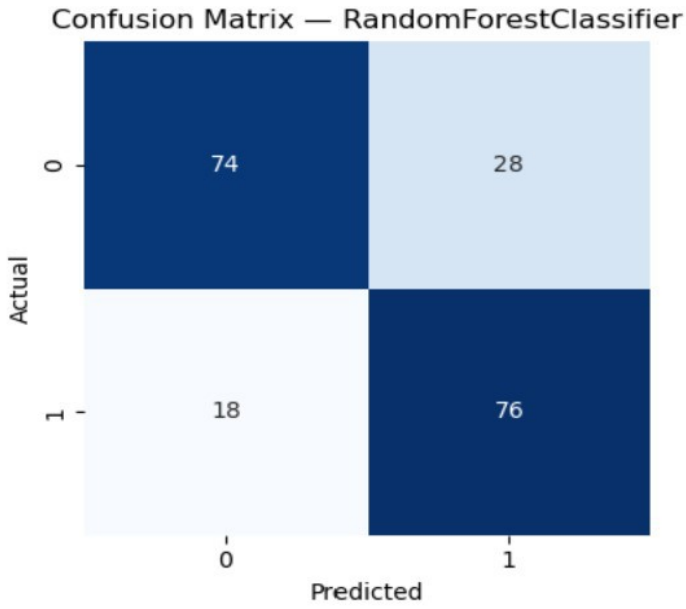


Для подбора гиперпараметров была выбрана модель RandomForestClassifier

Финальные метрики

Метрики на тестовой выборке:

	Accuracy	Precision	Recall	F1-score	ROC AUC
0	0.765	0.731	0.809	0.768	0.88

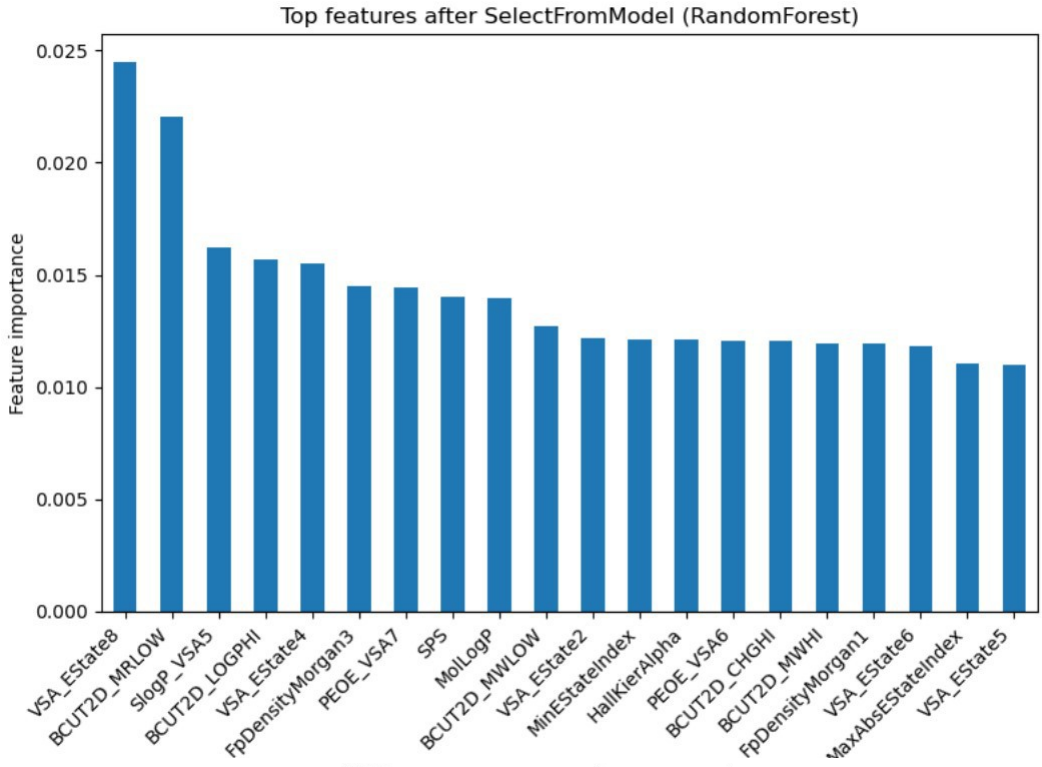


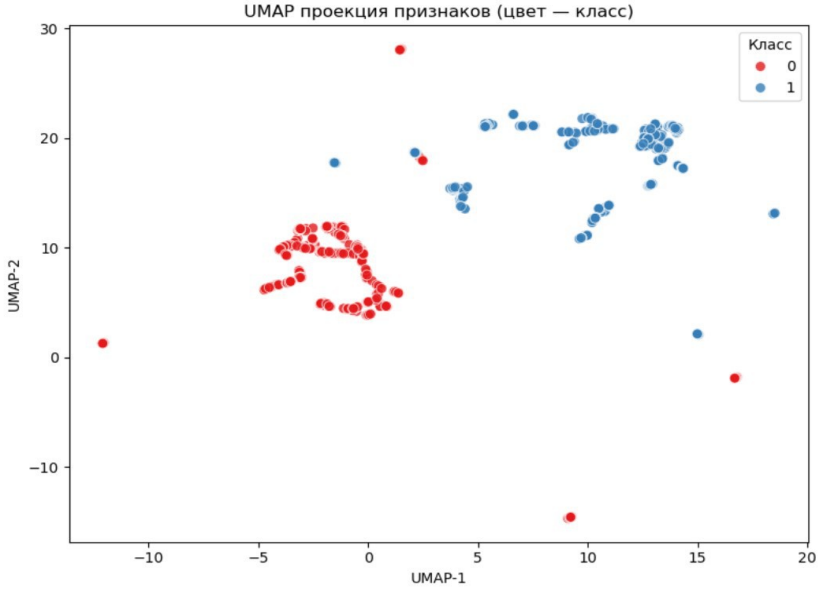
Классификация: превышает ли значение IC50 медианное значение выборки

Выбранные алгоритмы классификации:

- RandomForestClassifier;
- ExtraTreesClassifier;
- HistGradientBoostingClassifier;
- XGBoostClassifier;
- LightGBMClassifier;
- CatBoostClassifier.

Метод отбора признаков: такой же как был в CC50



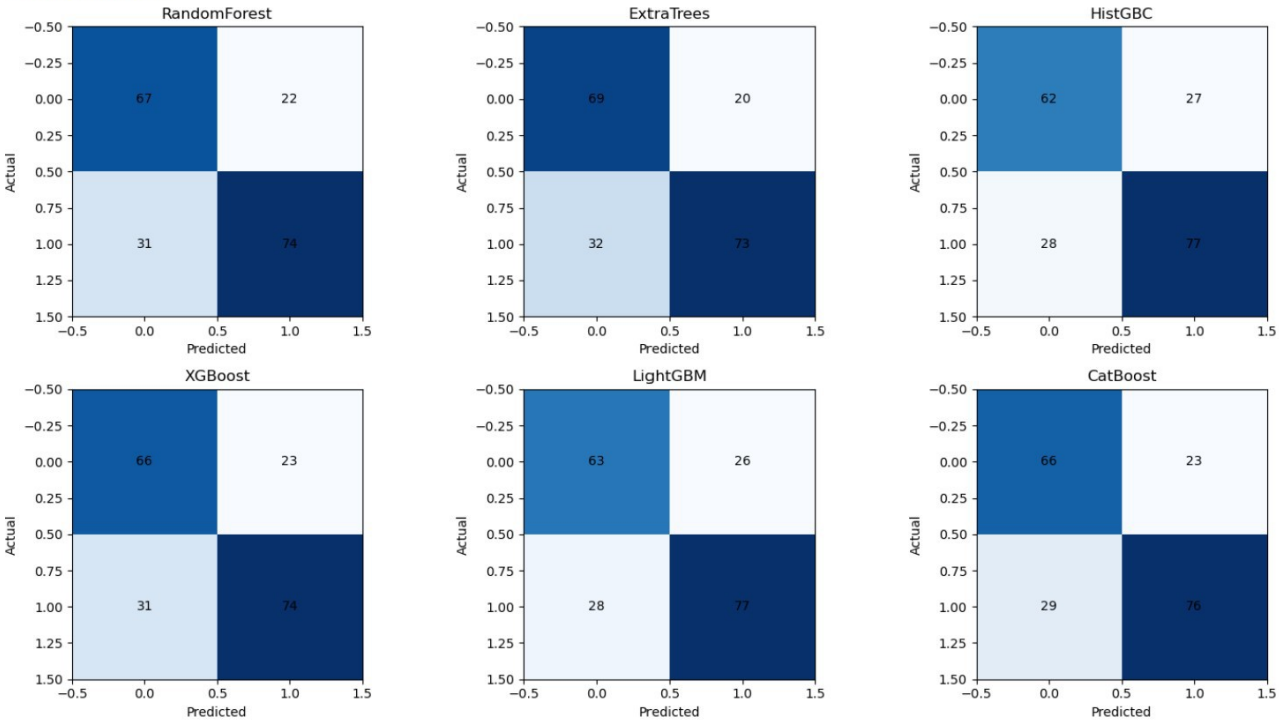


Наблюдается чёткое разделение классов: основная масса точек каждого класса образует свой кластер с минимальным наложением. Количество выбросов невелико.

Сводная таблица метрик на тестовой выборке:

	accuracy	precision	recall	f1	roc_auc
CatBoost	0.732	0.768	0.724	0.745	0.809
LightGBM	0.722	0.748	0.733	0.740	0.778
ExtraTrees	0.732	0.785	0.695	0.737	0.768
HistGBC	0.716	0.740	0.733	0.737	0.788
RandomForest	0.727	0.771	0.705	0.736	0.794
XGBoost	0.722	0.763	0.705	0.733	0.775

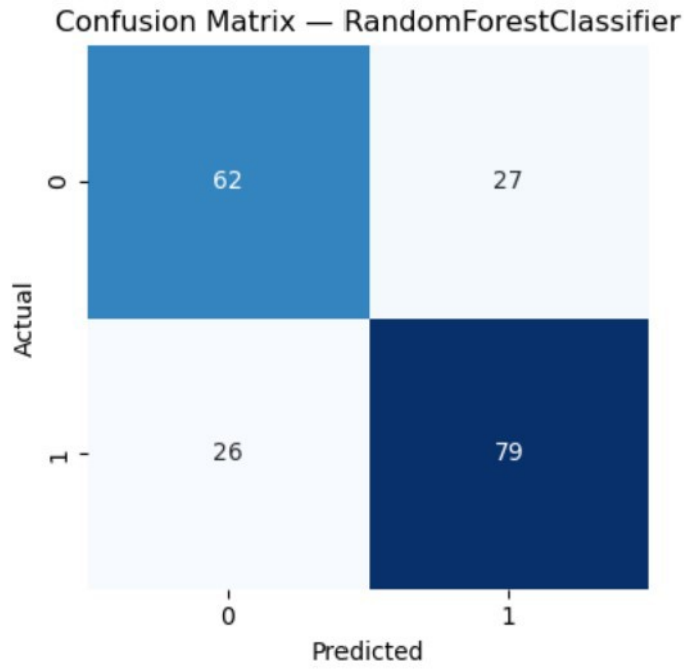
Confusion matrices:



Для подбора гиперпараметров была выбрана модель RandomForestClassifier
Финальные метрики

Метрики на тестовой выборке:

	Accuracy	Precision	Recall	F1-score	ROC AUC
0	0.727	0.745	0.752	0.749	0.802

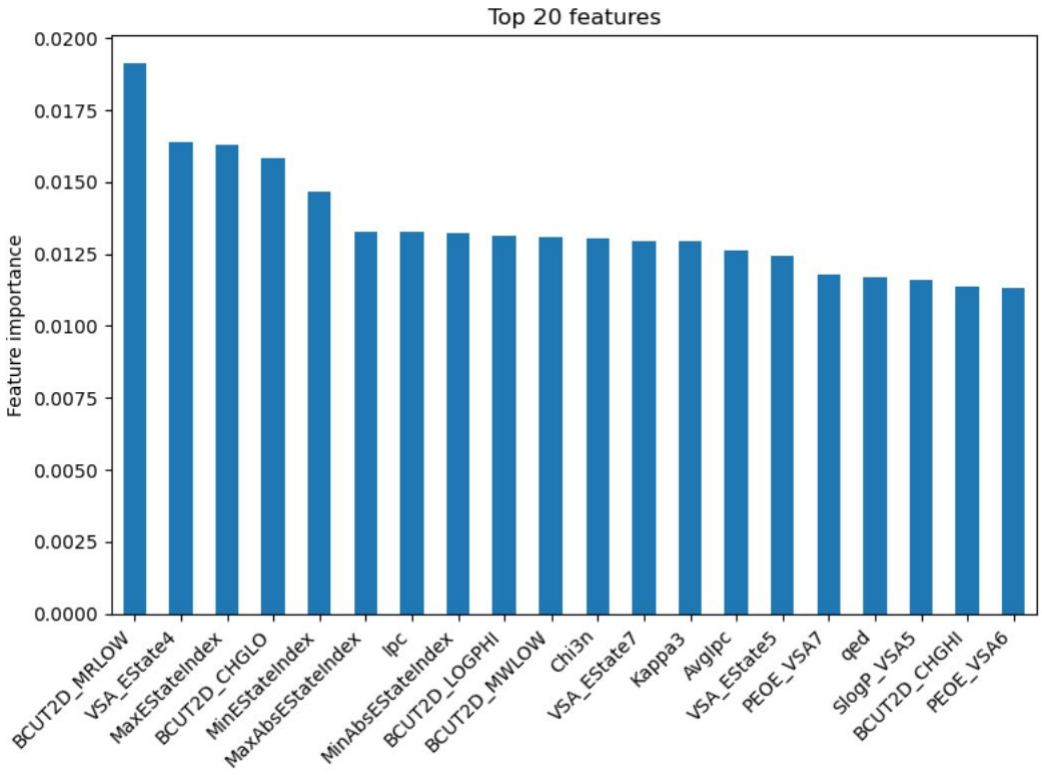


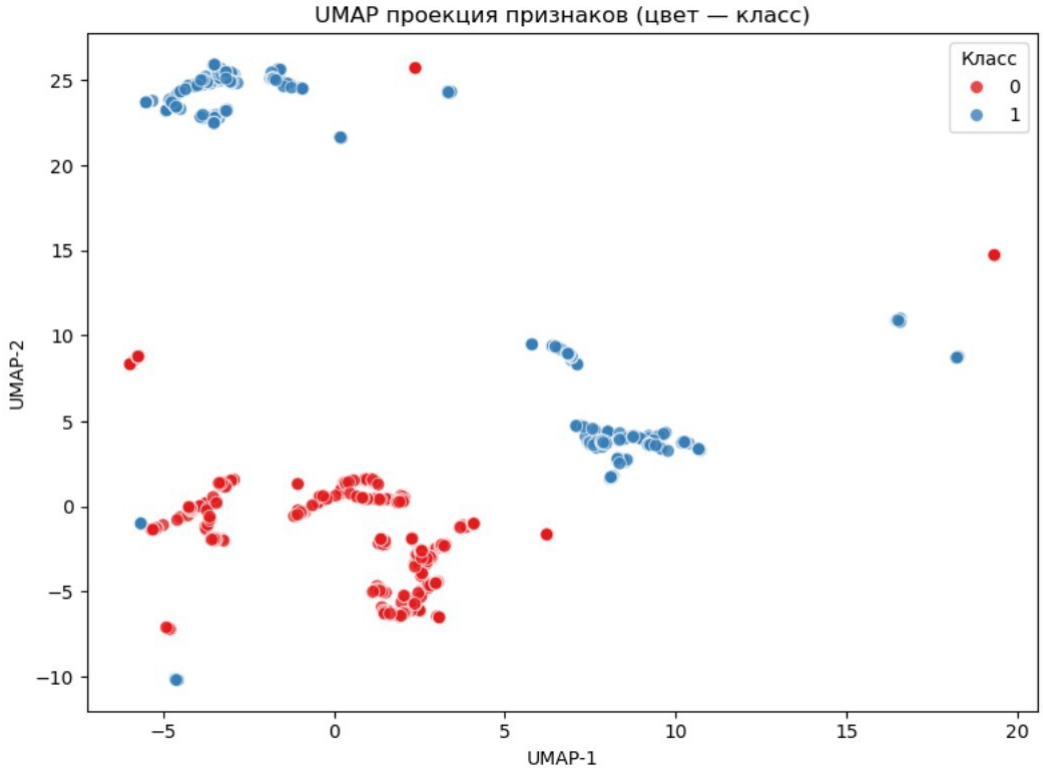
Классификация: превышает ли значение SI медианное значение выборки

Выбранные алгоритмы классификации:

- RandomForestClassifier;
- ExtraTreesClassifier;
- HistGradientBoostingClassifier;
- XGBoostClassifier;
- LightGBMClassifier;
- CatBoostClassifier.

Метод отбора признаков: такой же как был в CC50

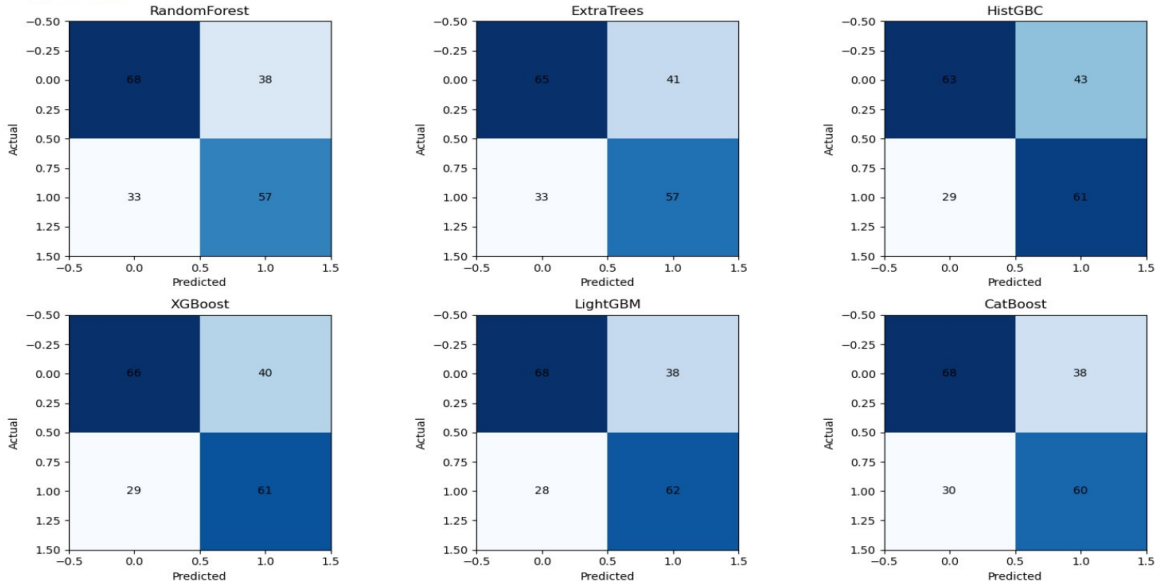

Top 20 features
Feature importance
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
BCUT2D_MRLOW
VSA_EState4
MaxEStateIndex
BCUT2D_CHGLO
MinEStateIndex
MaxAbsEStateIndex
Ipc
MinAbsEStateIndex
BCUT2D_LOGPHI
BCUT2D_MWLOW
Chi3n
VSA_EState7
Kappa3
AvgIpc
VSA_EState5
PEOE_VSA7
qed
SlogP_VSA5
BCUT2D_CHGHI
PEOE_VSA6


UMAP проекция признаков (цвет — класс)
Класс
0
1
UMAP-2
UMAP-1
25
20
15
10
5
0
−5
−10
−5
0
5
10
15
20

Сводная таблица метрик на тестовой выборке:

	accuracy	precision	recall	f1	roc_auc
LightGBM	0.663	0.620	0.689	0.653	0.702
XGBoost	0.648	0.604	0.678	0.639	0.710
CatBoost	0.653	0.612	0.667	0.638	0.713
HistGBC	0.633	0.587	0.678	0.629	0.695
RandomForest	0.638	0.600	0.633	0.616	0.692
ExtraTrees	0.622	0.582	0.633	0.606	0.695

Confusion matrices:

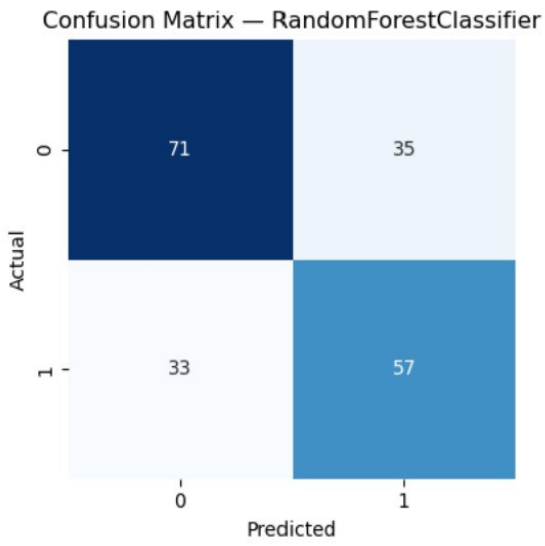


Для подбора гиперпараметров была выбрана модель RandomForestClassifier

Финальные метрики

Метрики на тестовой выборке:

	Accuracy	Precision	Recall	F1-score	ROC AUC
0	0.653	0.62	0.633	0.626	0.706

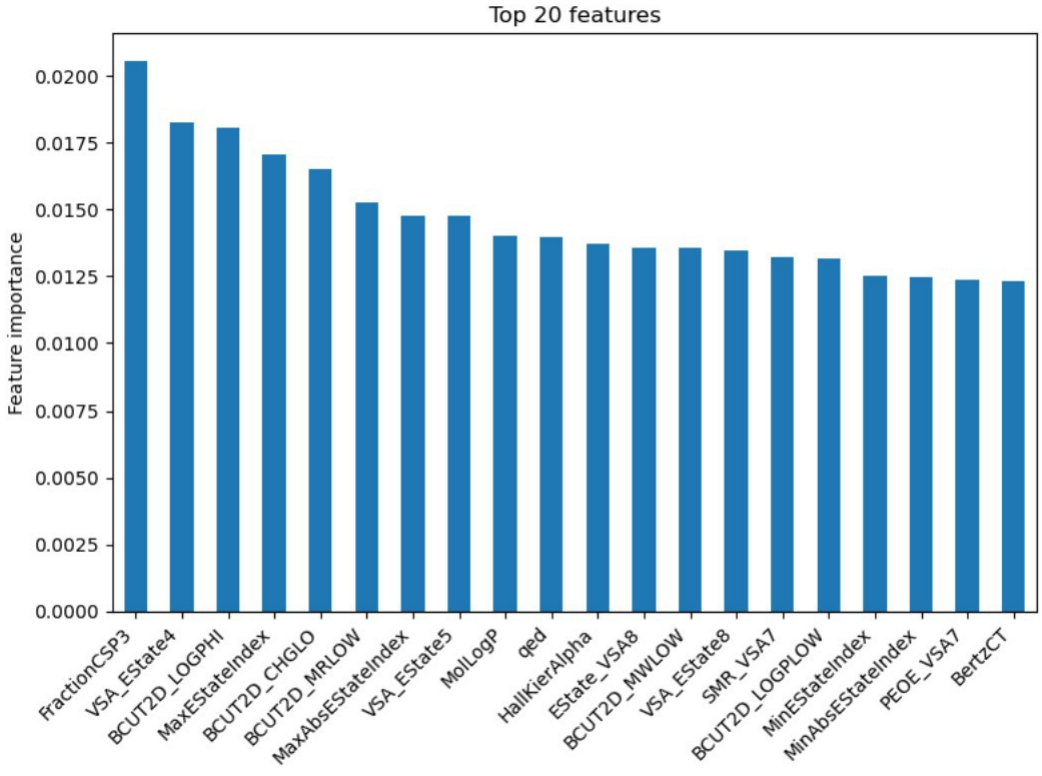


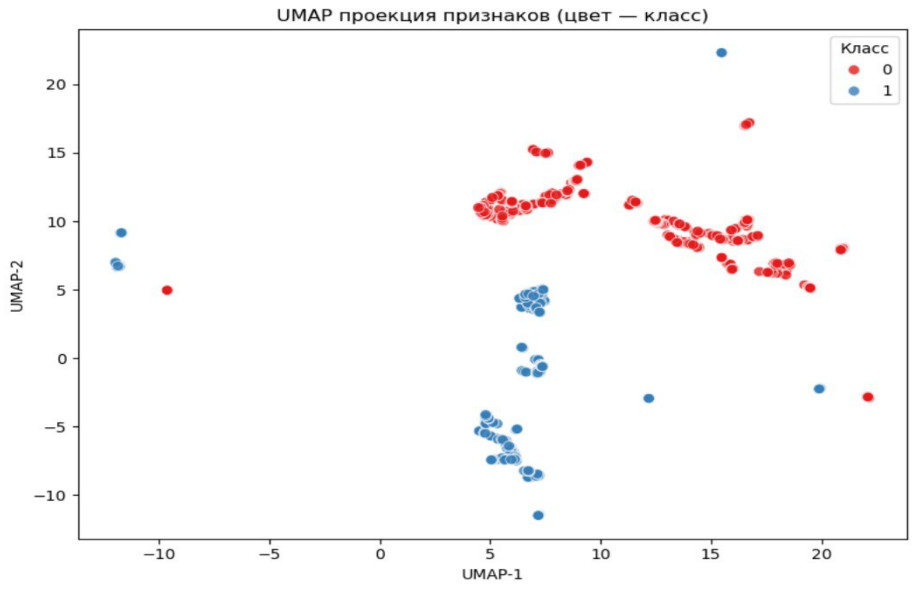
Классификация: превышает ли значение SI значение 8

Выбранные алгоритмы классификации:

- RandomForestClassifier;
- ExtraTreesClassifier;
- HistGradientBoostingClassifier;
- XGBoostClassifier;
- LightGBMClassifier;
- CatBoostClassifier.

Метод отбора признаков: такой же как был в CC50

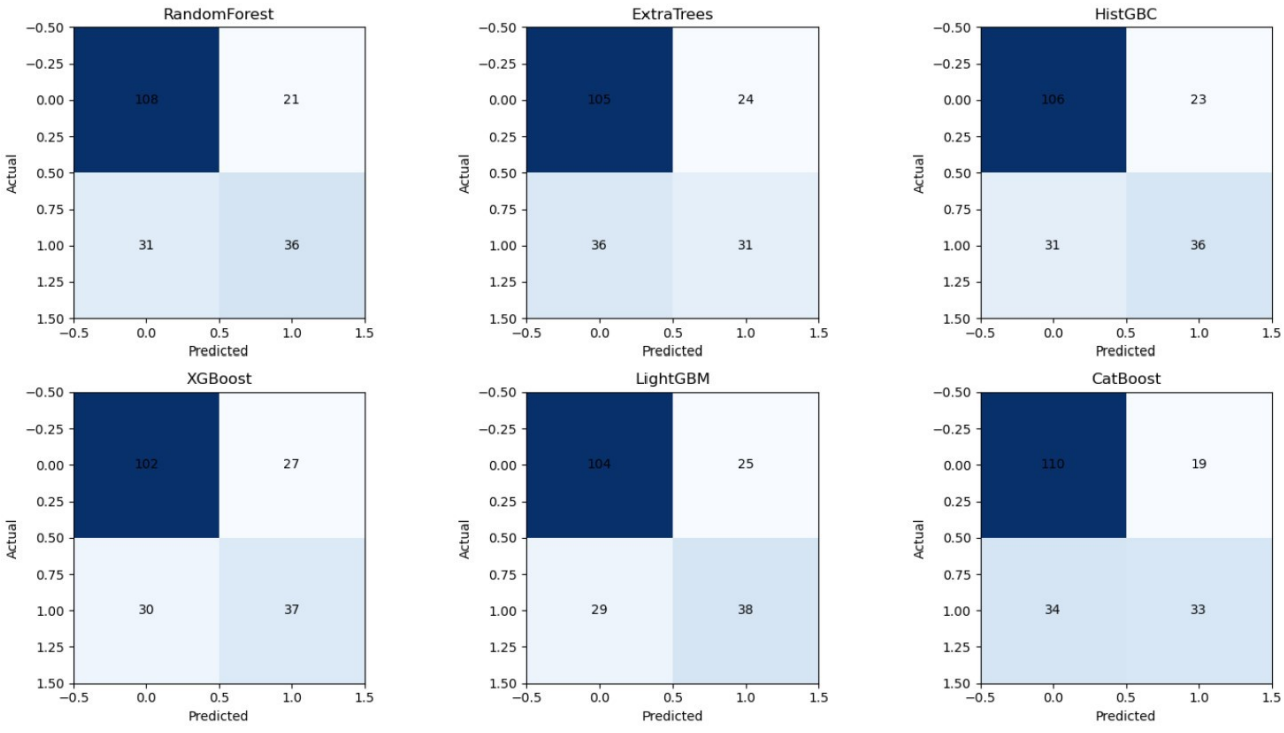




Сводная таблица метрик на тестовой выборке:

	accuracy	precision	recall	f1	roc_auc
LightGBM	0.724	0.603	0.567	0.585	0.722
RandomForest	0.735	0.632	0.537	0.581	0.700
HistGBC	0.724	0.610	0.537	0.571	0.718
XGBoost	0.709	0.578	0.552	0.565	0.709
CatBoost	0.730	0.635	0.493	0.555	0.714
ExtraTrees	0.694	0.564	0.463	0.508	0.696

Confusion matrices:



Для подбора гиперпараметров была выбрана модель RandomForestClassifier

Финальные метрики

```
Test Accuracy : 0.6990
Test ROC-AUC  : 0.7151
Метрики на тестовой выборке:
```

	Accuracy	Precision	Recall	F1-score	ROC AUC
0	0.699	0.562	0.537	0.55	0.715