Классическое машинное обучение.

Курсовая работа (vo\_PJ)

Отчёт

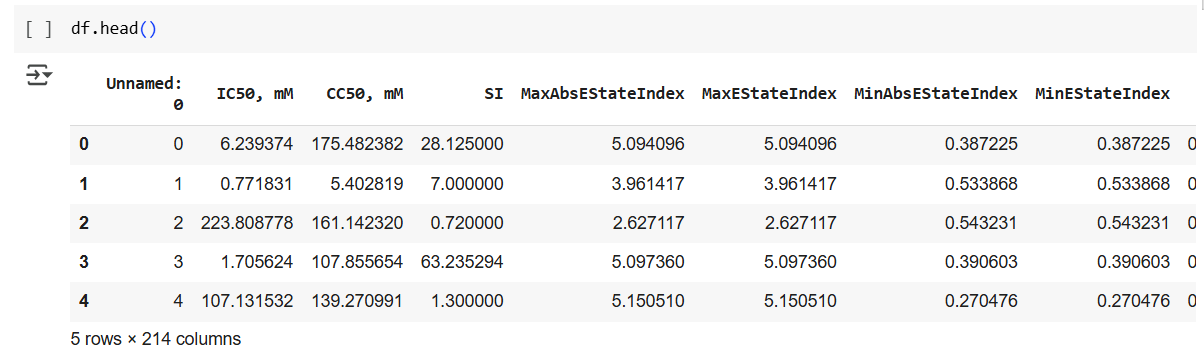
Выполнил: Душкин Василий Алексеевич

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.



**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

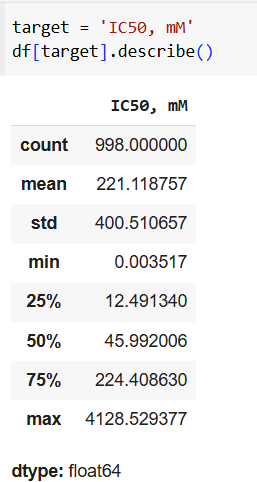
BCUT2D\_LOGPLOW 3

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**



|  |  |
| --- | --- |
|  |  |

**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

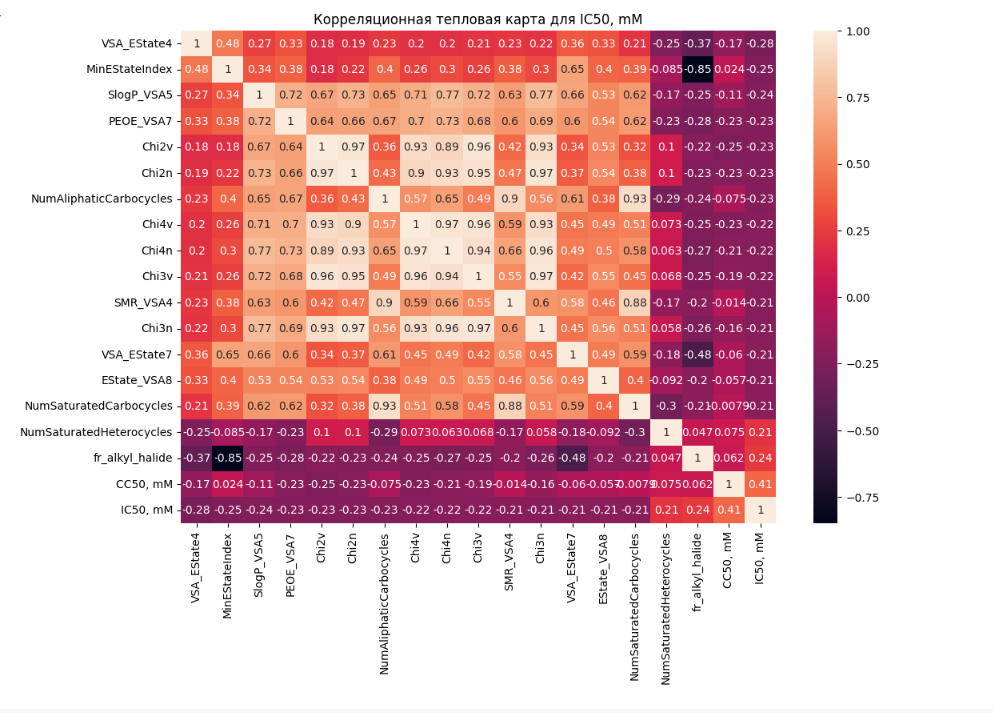
NumSaturatedCarbocycles составляет: -0.2071

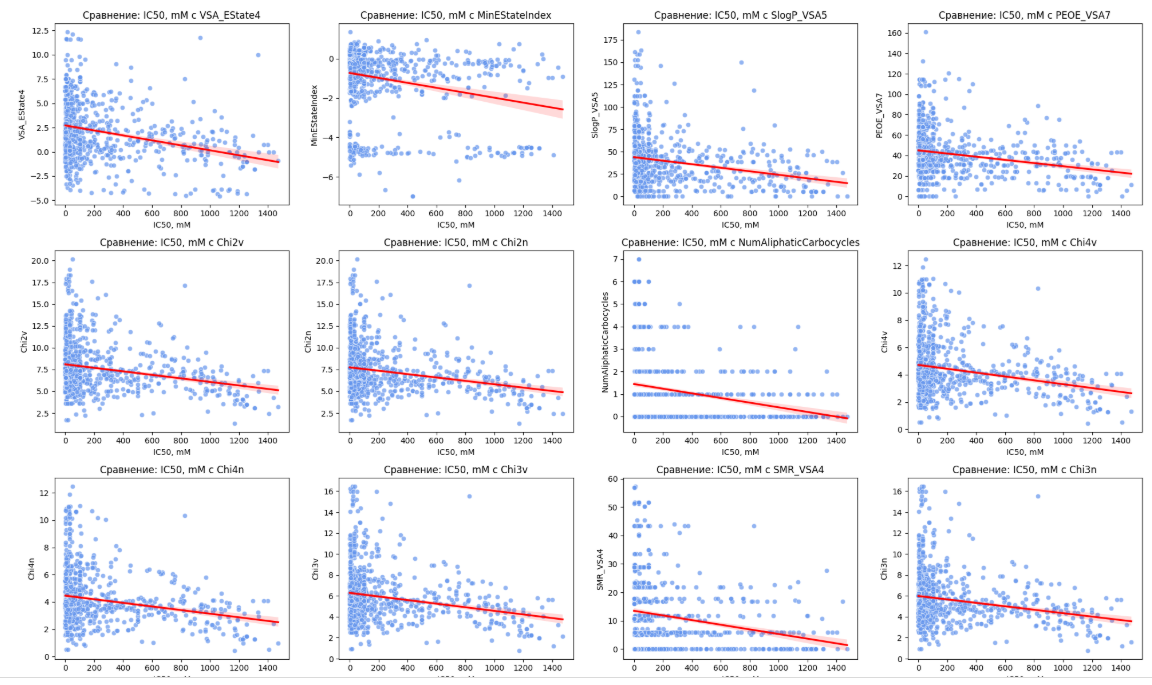
NumSaturatedHeterocycles составляет: 0.2117

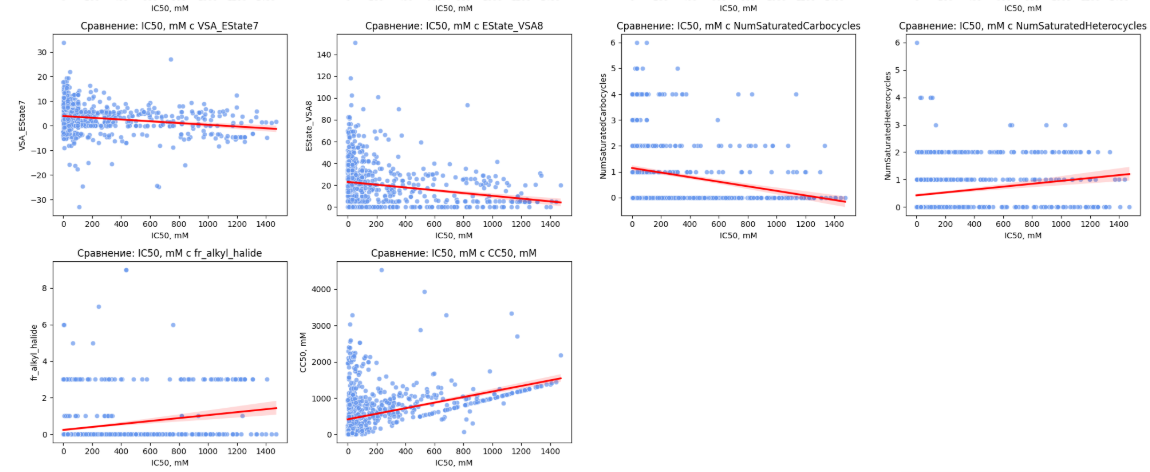
fr\_alkyl\_halide составляет: 0.2363

CC50, mM составляет: 0.4124

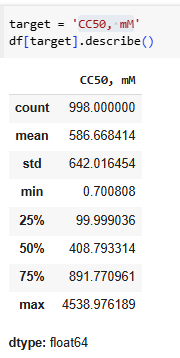
**Тепловая карта**



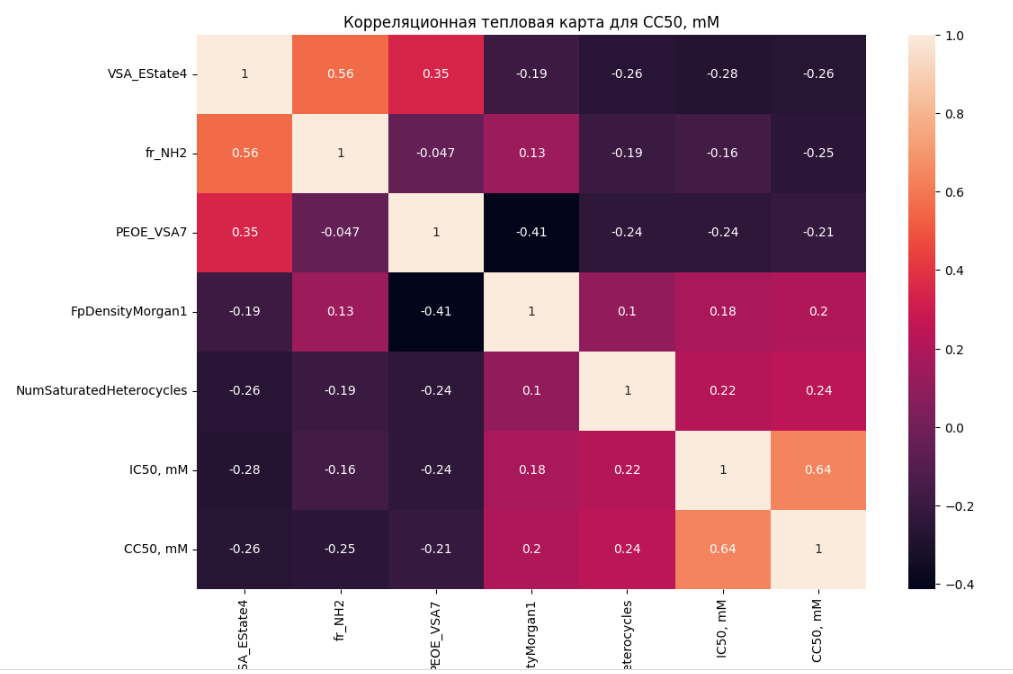


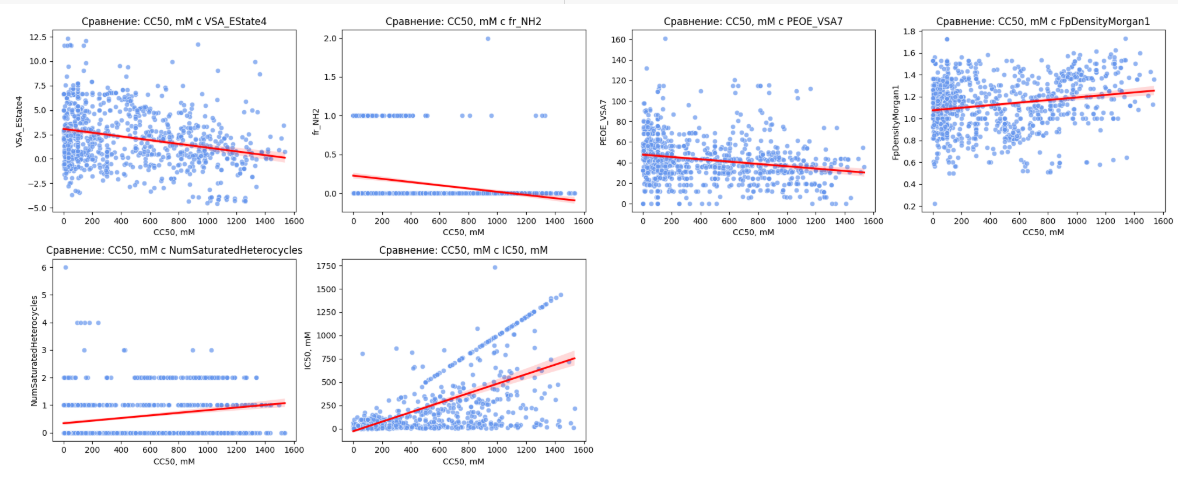


**Анализ CC50, mM**

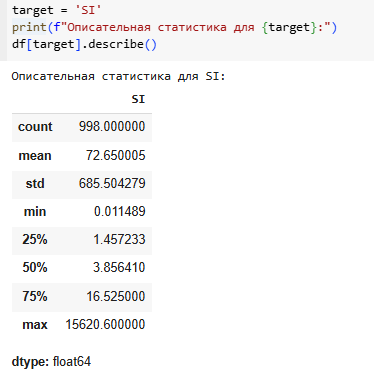


|  |  |
| --- | --- |
|  |  |

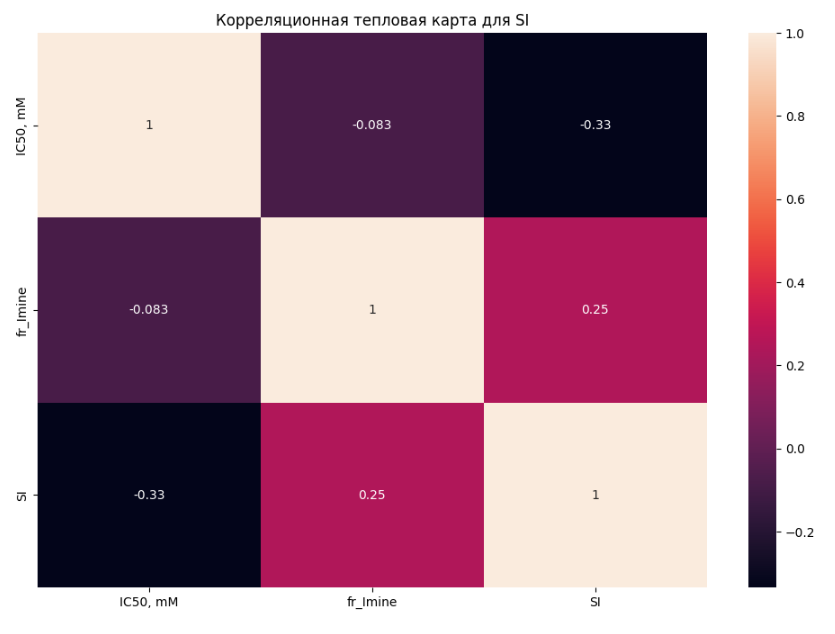


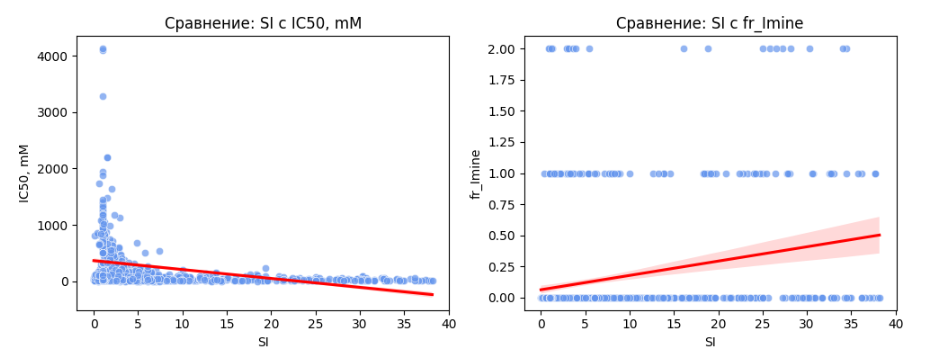


**Анализ SI**



|  |  |
| --- | --- |
|  |  |

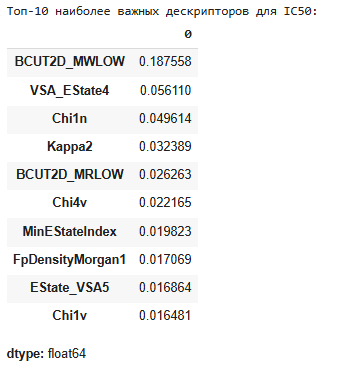


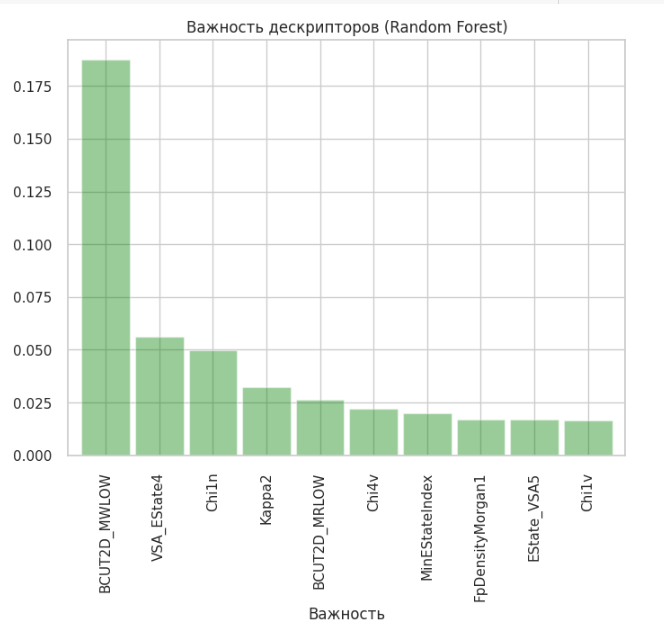


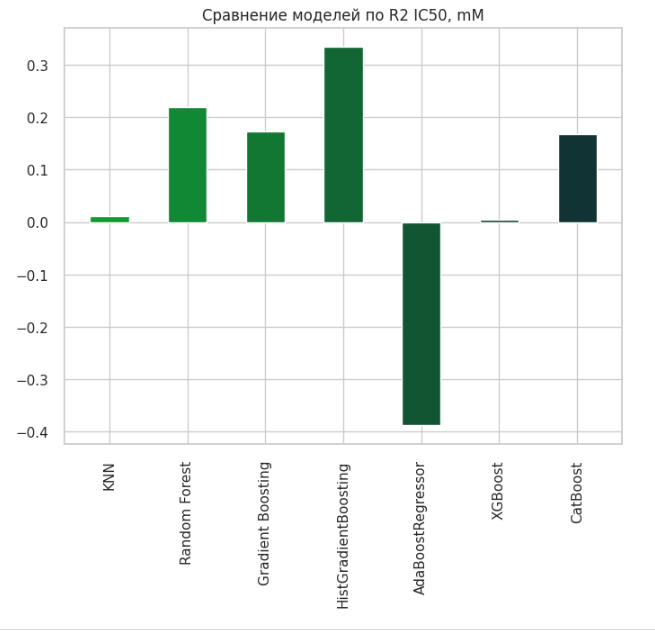
Корреляционный анализ показывает, что признак SI имеет значимую корреляцию только с IC50 и fr\_Imine.

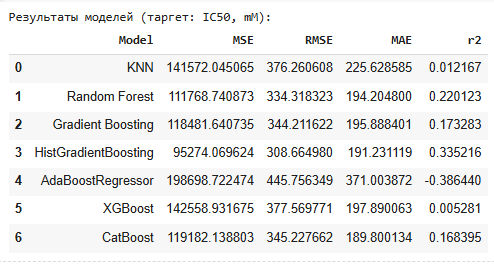
fr\_Imine — это дескриптор, который подсчитывает количество иминных групп в молекуле.

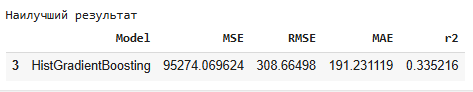
**2. регрессия IC50**





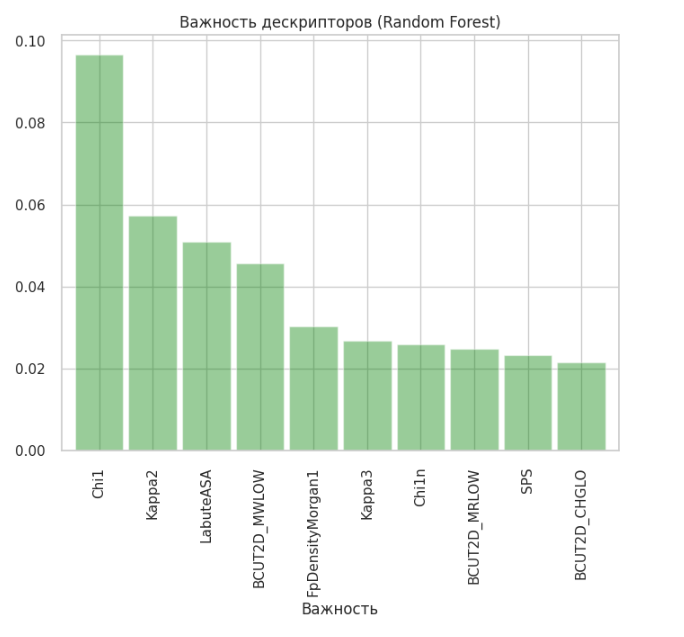


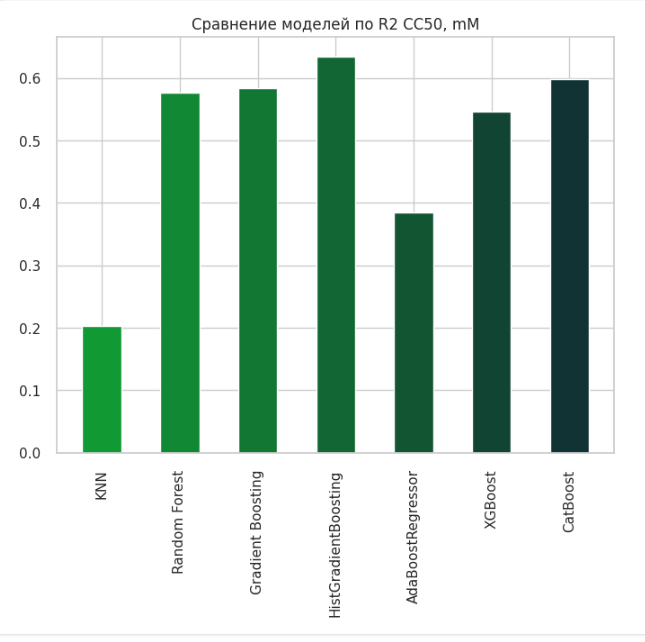


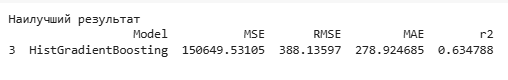
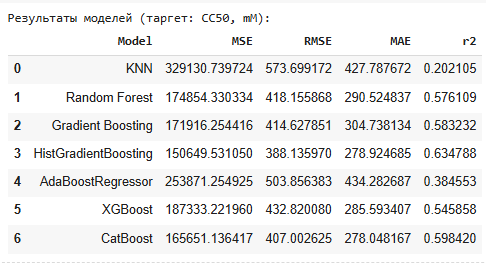


**3. регрессия CC50**



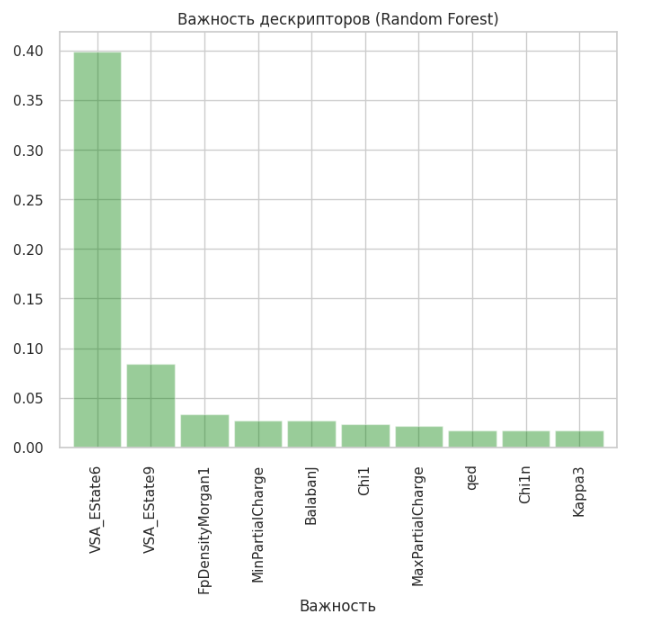


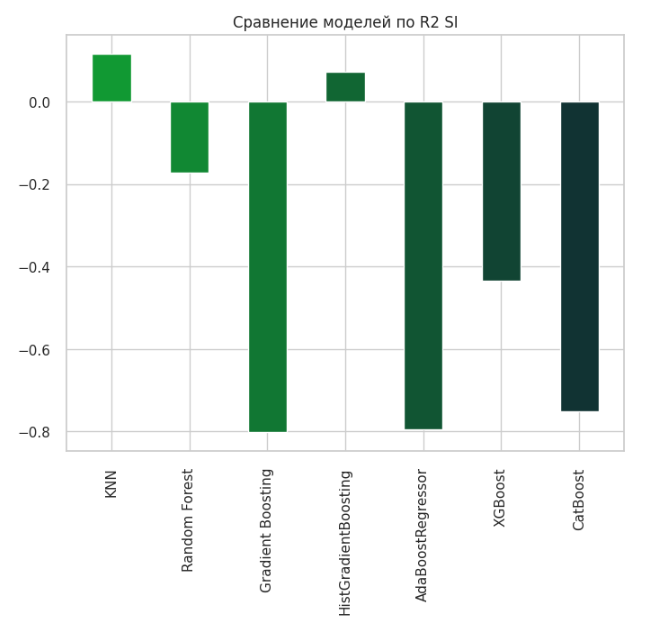




**4. регрессия SI**











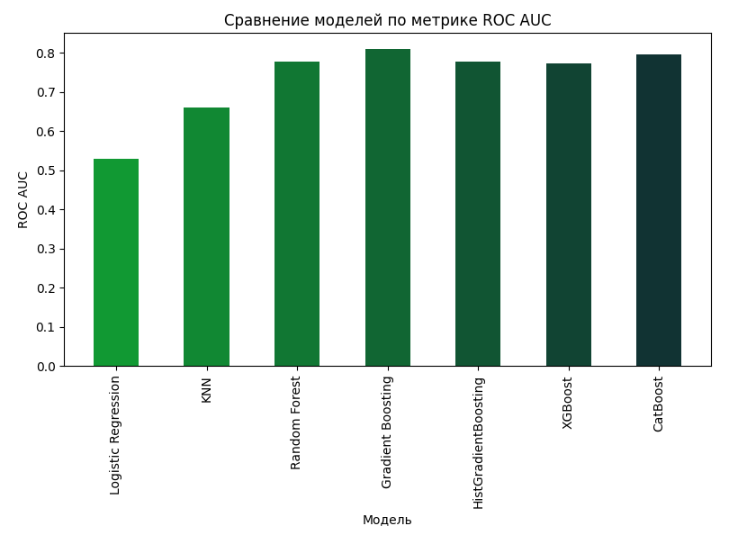
**5 Классификация IC50 медианное значение выборки**

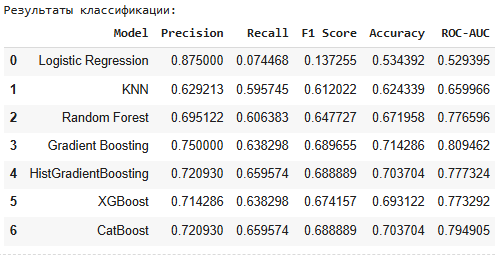
Добавляем новые данные.

а. iC50\_Median

б. 

|  |  |
| --- | --- |
| Logistic Regression | KNN |
|  |  |
| Random Forest | Gradient Boosting |
|  |  |
| HistGradientBoosting | XGBoost |
|  |  |
| CatBoost |  |
|  |  |







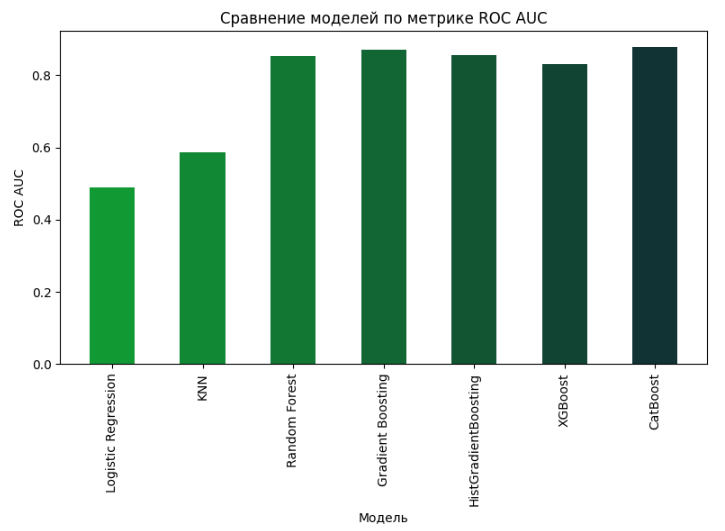
**6 Классификация CC50 медианное значение выборки.**

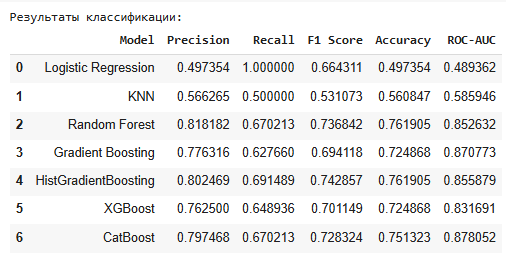
Добавляем новые данные.

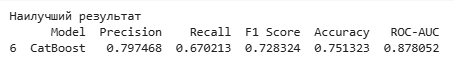
а. CC50\_Median

б. 

|  |  |
| --- | --- |
| Logistic Regression | KNN |
|  |  |
| Random Forest | Gradient Boosting |
|  |  |
| HistGradientBoosting | XGBoost |
|  |  |
| CatBoost |  |
|  |  |







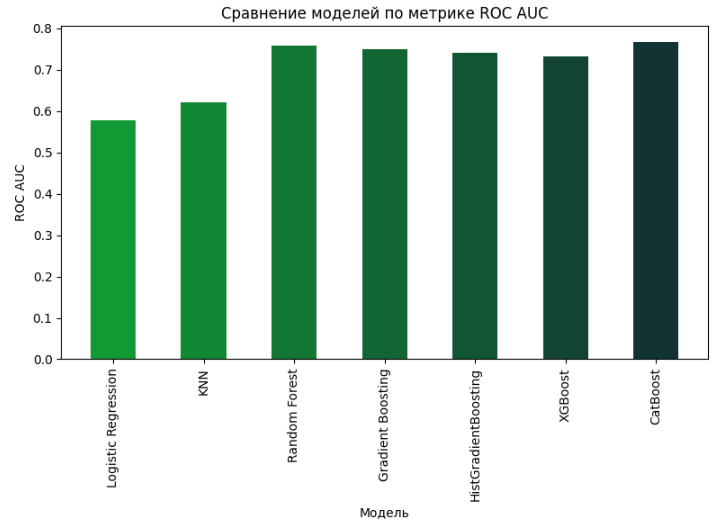
07 Классификация SI медианное значение выборки

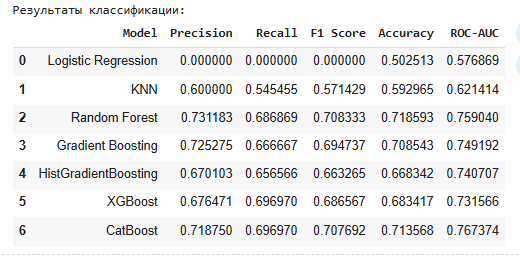
Добавляем новые данные.

а. SI\_Median

б. 

|  |  |
| --- | --- |
| Logistic Regression | KNN |
|  |  |
| Random Forest | Gradient Boosting |
|  |  |
| HistGradientBoosting | XGBoost |
|  |  |
| CatBoost |  |
|  |  |







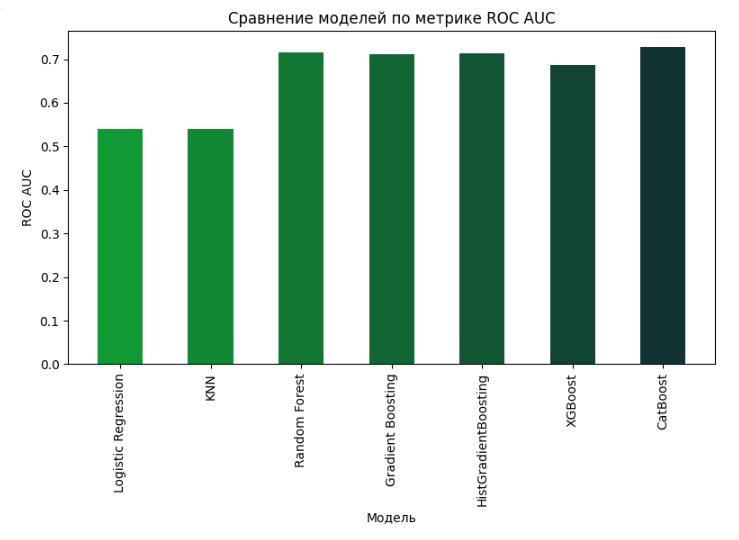
8 Классификация превышает ли значение SI значение 8

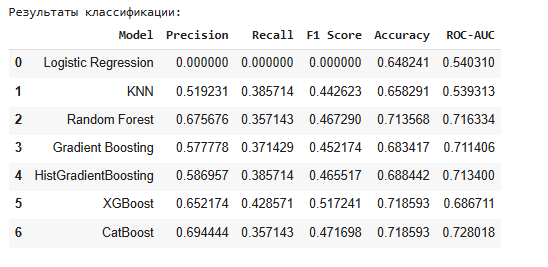
Добавляем новые данные.

а. SI\_8

б. 

|  |  |
| --- | --- |
| Logistic Regression | KNN |
|  |  |
| Random Forest | Gradient Boosting |
|  |  |
| HistGradientBoosting | XGBoost |
|  |  |
| CatBoost |  |
|  |  |

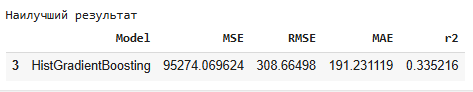




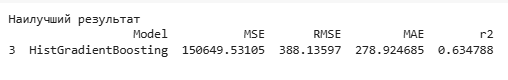


**Выводы:**

У регрессии IC50



У регрессии CC50



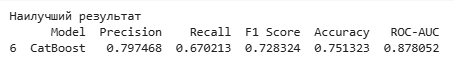
У регрессии SI



У классификации IC50 медианное значение выборки



У классификации CC50 медианное значение выборки.



У классификации SI медианное значение выборки



У классификации превышающее значение SI значение 8

