Аналитический отчет к курсовой работе

На тему: «Прогнозирование эффективных параметров для создания лекарственных препаратов»

Выполнил:

Студент группы М24-525

Мартынов А.В.

Оглавление

Введение

В условиях угрозы пандемий разработка новых лекарственных препаратов становится критически важной задачей современной фармакологии. Традиционные методы скрининга химических соединений требуют колоссальных временных и финансовых затрат. В этом контексте методы машинного обучения предлагают революционный подход к ускорению доклинических исследований. Настоящее исследование посвящено построению прогнозных моделей для ключевых параметров эффективности:

- Индекс ингибирующей концентрации IC50 (концентрация соединения, необходимая для подавления активности вируса (или фермента) на 50% [чем больше, тем менее токсично для вируса]);
- Индекс цитотоксической концентрации CC50 (концентрация соединения, при которой гибнет 50% клеток [чем выше, тем менее токсично для обычных клеток]);
- Терапевтический индекс (селективности) SI (показывает, насколько соединение избирательно действует на вирус, а не на клетки: SI = CC50/CI50 [чем выше, тем лучше селективность])

Цели задания:

На основании предоставленных данных химических соединений необходимо:

1. Построить регрессионные модели для предсказания IC50, CC50 и SI;
2. Построить классификационные модели для определения, превышают ли IC50, CC50 и SI определенные пороги (медиану и 8 для SI);
3. Сравнить модели и выбрать лучшие для каждой задачи.

Задачи исследования:

1. Провести исследовательский анализ данных и предобработку данных;
2. Построить модели регрессии и классификации;

3. Оценить модели с помощью адекватных метрик;
4. Интерпретировать результаты и сделать выводы. Выполнить сравнительный анализ качества моделей и выбрать наилучшую для каждой задачи.

Глава 1. EDA и предобработка данных

Разведочный анализ данных и предварительная обработка данных произведены в файле notebooks/1_EDA.ipynb

Описание данных:

Исходный датасет (сохранен под именем data/raw/course_task_data.xlsx) представляет собой таблицу, содержащую 1001 строку-наблюдение (химические соединения), в каждом столбце которой представлены признаки биологической активности и физико-химические свойства.

Всего представлено 214 признаков, из которых 3 (IC50, mM, CC50, mM и SI) – являются будущими целевыми переменными

Основные группы признаков:

1. Базовые молекулярные свойства:
    - MolWt, ExactMolWt, HeavyAtomMolWt
    - HeavyAtomCount, NumHeteroatoms, NOCount
    - FractionCSP3 (степень насыщенности)
    - RingCount (общее число колец)
    - NumRotatableBonds (гибкость молекулы)
    - MolLogP (липофильность)
    - MolMR (поляризуемость)
    - TPSA (полярная поверхность)
2. Электронные свойства и заряды:

- MaxAbsEStateIndex, MinAbsEStateIndex, MaxEStateIndex, MinEStateIndex
- MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge
- NumValenceElectrons, NumRadicalElectrons

3. Топологические индексы (графовые дескрипторы):
    - Chi0-4n/v (индексы связности)
    - BalabanJ, BertzCT
    - Kappa1-3 (индексы формы)
    - HallKierAlpha, Ipc, AvgIpc
4. BCUT-дескрипторы (2D-дескрипторы матрицы связности):
    - BCUT2D_MWHI, BCUT2D_MWLOW
    - BCUT2D_CHGHI, BCUT2D_CHGLO
    - BCUT2D_LOGPHI, BCUT2D_LOGPLOW
    - BCUT2D_MRHI, BCUT2D_MRLOW
5. Дескрипторы поверхности (VSA):
    - PEOE_VSA1-14 (площадь поверхности по зарядам)
    - SMR_VSA1-10 (площадь поверхности по поляризуемости)
    - SlogP_VSA1-12 (площадь поверхности по липофильности)
    - EState_VSA1-11, VSA_EState1-10 (площадь по электронному состоянию)
6. Кольцевые системы:
    - NumAliphaticCarbocycles, NumAliphaticHeterocycles
    - NumAromaticCarbocycles, NumAromaticHeterocycles
    - NumSaturatedCarbocycles, NumSaturatedHeterocycles
    - NumAliphaticRings, NumAromaticRings, NumSaturatedRings
7. Функциональные группы (обозначены как fr_*):
    - Водородные связи: NHOHCount, NumHAcceptors, NumHDonors
    - Кислоты/основания: fr_COO, fr_COO2, fr_Ar_COO, fr_Al_COO

- Азотные группы: fr_NH0-2, fr_N_O, fr_nitro, fr_nitrile
- Кислородные группы: fr_Al_OH, fr_Ar_OH, fr_ether, fr_aldehyde, fr_ketone
- Гетероциклы: fr_furan, fr_imidazole, fr_pyridine, fr_thiazole, fr_thiophene и др.
- Специфичные группы: fr_halogen, fr_sulfide, fr_sulfone, fr_epoxide и т.д.

8. Оценочные метрики:
    - qed (drug-likeness)
    - SPS (сложность синтеза)

Разведочный анализ данных:

1. Загрузка и первичный осмотр датасета:
    a. Доменная обработка: удаление сторонних признаков:
        i. Unnamed: 0 (ID), SPS (предполагаемая сложность синтеза);
    b. Оценка распределения целевых переменных.

Таблица 1. Гистограмма распределения целевых переменных

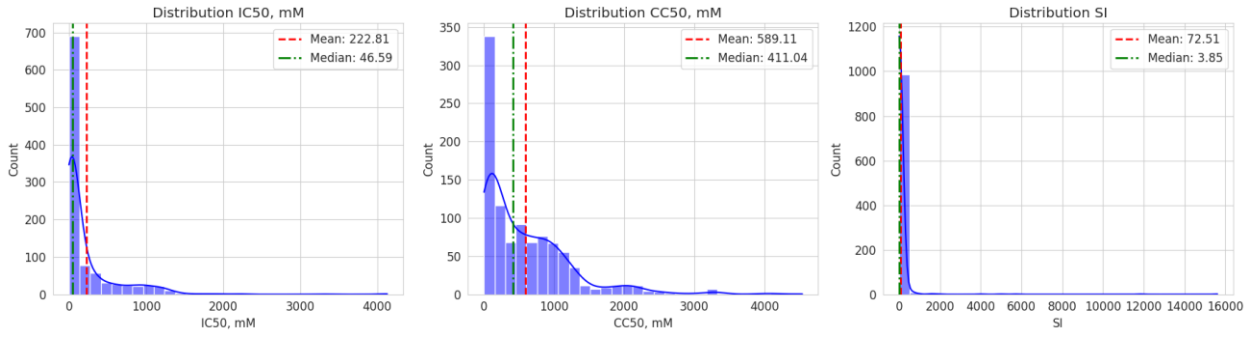
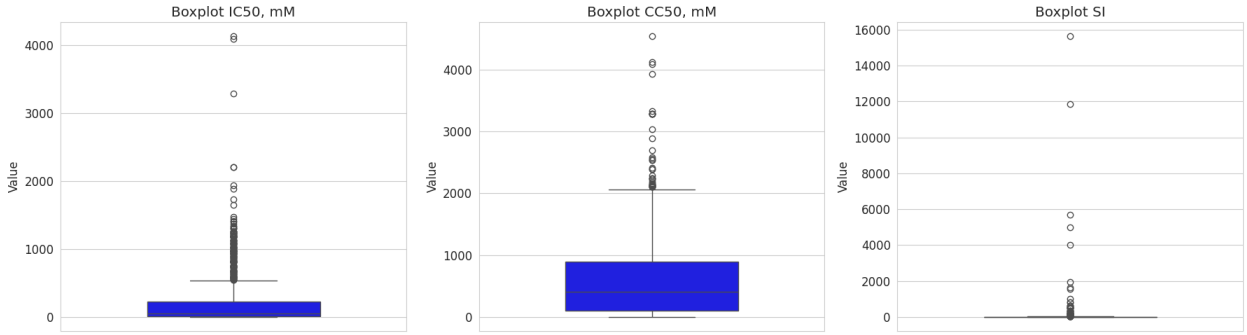
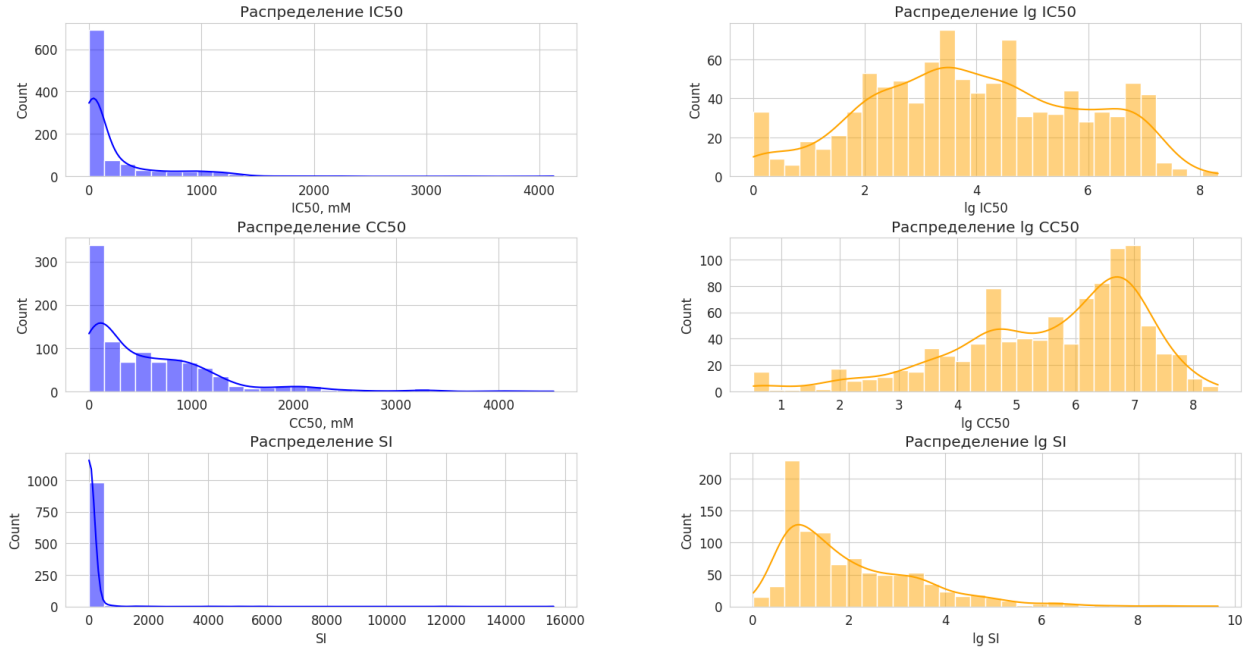


Таблица 2. Ящичковая диаграмма распределения целевых переменных



Для борьбы со скошенностью (нормализация и снижение влияния выбросов) было принято решение логарифмировать целевые переменные.

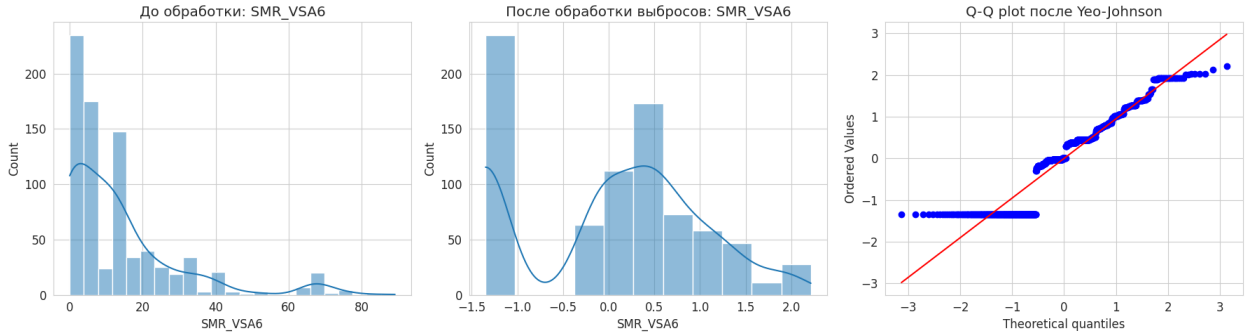
Таблица 3. Сравнение распределения целевых переменных до и после логарифмирования



2. Разделение исходной выборки на train и test;
3. Предобработка данных:
    a. Анализ пропусков показал наличие всего 3 пропусков в 12 переменных (решено применить подход с заменой пустых значений на медиану);

b. Бинаризация фрагментов: для переменных типа fr_* все значения выше 0 приведены к 1;
c. Обработка констант и признаков с низкой вариабельностью: выявлено 19 констант и 52 переменных, в которых у 95% наблюдений сохраняется одно значение – к исключению;
d. Создание новых переменных:
    i. Соотношение акцепторов водорода к донорам HAcceptors_HDonors_ratio (показатель полярности молекулы и способности к образованию водородных связей);
    ii. Масса молекулы на единицу полярной поверхности MolWt_TPSA_ratio (эффективность "полярного покрытия" молекулы (липидная эффективность));
    iii. Общее количество доноров и акцепторов водорода Hydrogen_Total (общая способность к образованию водородных связей);
    iv. Доля вращающихся связей среди всех тяжелых атомов RotatableBonds_HeavyAtomRatio (гибкость молекулярной структуры);
    v. Полярная поверхность на один атом PolarSurfaceArea_per_Atom ("плотность" полярности молекулы, коррелирует с проницаемостью через клеточные мембраны);
    vi. Липофильность на один атом LogP_per_Atom (липофильность на один атом);
e. Оценка распределения прочих переменных. Из-за нехватки памяти, расчет в 1_EDA.ipynb выполнен в виде виджета. Применена обработка с помощью Yeo-Johnson PowerTransformer

Таблица 4. Результат применения Yeo-Johnson PowerTransformer на распределение переменной SMR_VSA6



Несмотря на то, что Q-Q график всё ещё показывает кластеризованность, гистограмма распределения переменной значительно улучшилась.

Таблица 5. Топ-10 максимальных изменений скошенности

	Признак	Skew-ДО	Skew-ПОСЛЕ	Skew, откл.
40	Ipc	27,8332	0,0486	27,7846
50	PEOE_VSA14	10,6775	0,4198	10,2577
98	VSA_EState9	1,7145	-5,3983	7,1128
76	SlogP_VSA7	6,1672	0,0000	6,1672
92	VSA_EState3	4,2099	0,0044	4,2055
110	NumHDonors	3,3922	0,0759	3,3163
45	PEOE_VSA1	3,2500	0,0636	3,1864
46	PEOE_VSA10	3,1511	0,0836	3,0675
79	EState_VSA1	3,0926	0,1460	2,9465
69	SlogP_VSA11	2,8480	0,0000	2,8480

f. Группировка некоторых признаков в смысловые кластеры и формирование на основе кластеров новых переменных по методу PCA (контроль дисперсии 80%);

g. Попарный корреляционный анализ оставшихся переменных и удаление признаков, чья корреляция превышает 90%

Перечисленные преобразования позволяют сократить размерность выборки до ~100 признаков. Все они были преобразованы в функции и классы, упакованы в пайплайн, и сохранены в файле src/preprocessing.py для дальнейшего использования в виде кастомного трансформера.

Глава 2. Построение и оценка моделей

Несмотря на то, что дальнейшие преобразования разделены на отдельные файлы для каждой из задач, часть подхода остаётся общей:

1. Загрузка исходного датафрейма (для создания групп-PCA);
2. Загрузка разделенных выборок train и test;
3. Обучение кастомного трансформера ProcessingPipeline из src/preprocessing.py на основе данных train и трансформация с помощью него train и test выборок. В итоге получились наборы из 97 независимых переменных (что достаточно много для 1000 наблюдений);
4. Для моделей регрессии, как было сказано выше, было решено использовать логарифмированные переменные lg_IC50, lg_CC50, lg_SI. Для моделей классификации на основе исходных переменных были созданы специальные:
    - IC50_above_median;
    - CC50_above_median;
    - SI_above_median;
    - SI_above_8;
5. Далее были подобраны перечни моделей регрессии и классификации:

Таблица 6. Перечень исходных моделей

№	Регрессия	Классификация
1	LinReg	LogReg
2	Ridge	
3	Lasso	
4		NaiveBayes
5	SVR	SVC
6	KNN-Reg	KNN-Clf
7	DecisionTree-Reg	DecisionTree-Clf
8	RandomForest-Reg	RandomForest-Clf
9	XGB-Reg	XGB-Clf
10	GradBoost-Reg	GradBoost-Clf

6. Проведение предварительной кросс-валидации на 5 слоях для получения предварительных результатов (топ-3 модели):
    - Лучшие модели регрессии для всех трёх логарифмированных переменных lg_IC50, lg_CC50, lg_SI оказались одинаковыми: SVR, RandomForest-Reg и GradBoost-Reg
    - Лучшие модели классификации получились следующими:
        - Для IC50_above_median: KNN-Clf, XGB-Clf, RandomForest-Clf
        - Для CC50_above_median: RandomForest-Clf, SVC, LogReg;
        - Для SI_above_median: SVC, KNN-Clf, RandomForest-Clf;
        - Для SI_above_8: GradBoost-Clf, SVC, RandomForest-Clf
7. Далее для каждой переменной методом GridSearchCV были подобраны лучшие модели;
8. Результаты построения моделей сохранены в папку ../models.

Регрессионные модели

Регрессия для lg_IC50:

Target	Model	Best Params	CV R^2	Test R^2	Test RMSE	Test MAE	Search time
lg_IC50	SVR	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}	0,393	0,428	1,483	1,151	45,931
lg_IC50	RandomForest-Reg	{'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 200}	0,398	0,472	1,425	1,142	74,026
lg_IC50	GradBoost-Reg	{'learning_rate': 0,01, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0,8}	0,384	0,458	1,444	1,191	31,987

Хотя все модели показали схожую производительность, RandomForest-Reg демонстрирует наилучшие результаты для прогнозирования lg_IC50: она превосходит другие модели по всем тестовым метрикам, а также показала наибольшее улучшение относительно кросс-валидации (CV), что говорит о высокой обобщающей способности.

Результаты построения моделей сохранены в папку ../models

В качестве дальнейшего улучшения возможны:

- Дополнительная обработка независимых переменных (feature engineering, отбор наиболее значимых признаков (SHAP/LIME анализ), проверка мультиколлинеарности);
- Увеличить n_estimators;
- Применить ансамблирование (Stacking);
- Проверить ошибки на гетероскедастичность.

Регрессия для lg_CC50:

Target	Model	Best Params	CV R^2	Test R^2	Test RMSE	Test MAE	Search time
lg_CC50	SVR	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}	0,360	0,347	1,219	0,842	49,134

Target	Model	Best Params	CV R^2	Test R^2	Test RMSE	Test MAE	Search time
lg_CC50	RandomForest-Reg	{'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}	0,401	0,453	1,116	0,810	72,458
lg_CC50	GradBoost-Reg	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 1.0}	0,394	0,432	1,137	0,847	32,115

Все модели также показывают умеренное качество (R^2: 0.35-0.45), что характерно для биологических данных; но результаты всё равно хуже, чем для IC50. Аналогично IC50 лучшая модель – RandomForest-Reg, которая объясняет 45,3% дисперсии данных со средней абсолютной ошибкой 0.810 log-единиц.

Для дальнейшего улучшения можно также попробовать stacking и анализ ошибок.

Регрессия для lg_SI:

Target	Model	Best Params	CV R^2	Test R^2	Test RMSE	Test MAE	Search time
lg_SI	SVR	{'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}	0,252	0,276	1,324	0,917	50,001
lg_SI	RandomForest-Reg	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}	0,298	0,319	1,284	0,962	77,635
lg_SI	GradBoost-Reg	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}	0,303	0,297	1,304	0,988	31,945

Все модели показывают относительно низкое качество (R^2: 0.25-0.32), что ниже, чем для IC50 и CC50. Это говорит о том, что предсказать селективность сложнее. Скорее всего, это связано с тем, что SI является производной величиной (CC50/IC50), что увеличивает ошибку прогнозирования. Лучшая модель: RandomForest-Reg по тестовому R^2 (0.319) и RMSE (1.284).

Низкие значения R^2 указывают на необходимость фундаментального улучшения подходов к прогнозированию селективности, возможно, через включение дополнительных данных о молекулярной структуре или механизмах действия.

Модели классификации

Классификация по медиане для IC50

Target	Model	Best_Params	Best Score	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC	Test Log loss	Search Time
IC50_above_median	RandomForest-Clf	{'class_weight': None, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}	0,72	0,71	0,73	0,69	0,71	0,80	0,71	25,29
IC50_above_median	LogReg	{'C': 1, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'}	0,68	0,70	0,71	0,70	0,71	0,77	0,60	0,88
IC50_above_median	KNN-Clf	{'n_neighbors': 3, 'p': 2, 'weights': 'uniform'}	0,73	0,71	0,73	0,70	0,72	0,75	4,77	0,12
IC50_above_median	XGB-Clf	{'colsample_bytree': 0.9, 'gamma': 0.1, 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8}	0,73	0,72	0,73	0,72	0,72	0,80	0,55	180,77
IC50_above_median	SVC	{'C': 10, 'class_weight': None, 'gamma': np.float64(0.01), 'kernel': 'poly', 'probability': True}	0,72	0,73	0,74	0,72	0,73	0,78	0,57	9,80
IC50_above_median	GradBoost-Clf	{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 200, 'subsample': 0.9}	0,71	0,74	0,75	0,73	0,74	0,80	0,64	142,39

Все модели показывают умеренное качество (Accuracy: 0.70-0.74). Лучшие результаты показывают ансамблевые методы (GradBoost-Clf, XGB-Clf) и SVC. При этом GradBoost-Clf имеет максимальную точность и F1-score, а также высокую разделительную способность (ROC AUC 0.8)

Классификация по медиане для CC50

Target	Model	Best_Params	Best Score	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC	Test Log loss	Search Time
CC50_above_median	RandomForest-Clf	{'class_weight': None, 'max_depth': None, 'min_samples_leaf': 1,	0,75	0,82	0,86	0,77	0,82	0,88	0,44	25,28

Target	Model	Best_Params	Best Score	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC	Test Log loss	Search Time
		'min_samples_split': 10, 'n_estimators': 50}								
CC50_above_median	LogReg	{'C': 1, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'}	0,74	0,76	0,79	0,75	0,77	0,83	0,52	0,54
CC50_above_median	KNN-Clf	{'n_neighbors': 5, 'p': 1, 'weights': 'uniform'}	0,72	0,77	0,79	0,76	0,78	0,85	1,78	0,11
CC50_above_median	XGB-Clf	{'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 200, 'subsample': 0.9}	0,74	0,76	0,78	0,75	0,76	0,85	0,70	164,15
CC50_above_median	SVC	{'C': 100, 'class_weight': None, 'gamma': 'scale', 'kernel': 'rbf', 'probability': True}	0,75	0,74	0,78	0,70	0,74	0,77	0,58	10,63
CC50_above_median	GradBoost-Clf	{'learning_rate': 0.05, 'max_depth': 3, 'min_samples_split': 5, 'n_estimators': 200, 'subsample': 0.9}	0,74	0,80	0,83	0,77	0,80	0,88	0,44	136,79

Все модели показали хорошие результаты (Accuracy: 0.73-0.81), что значительно лучше результатов для IC50. RandomForest-Clf является оптимальным выбором, обеспечивая высокую точность и надежность прогнозов (precision говорит, что модель верно определяет высокую цитотоксичность в 86.3% случаев). GradBoost-Clf предлагает лучший баланс между точностью и полнотой (F1-score).

Классификация по медиане для SI

Target	Model	Best_Params	Bes Score	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC	Test Log loss	Search Time
SI_above_median	RandomForest-Clf	{'class_weight': None, 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 200}	0,70	0,68	0,68	0,58	0,62	0,74	0,59	25,91
SI_above_median	LogReg	{'C': 100, 'class_weight': None, 'penalty': 'l2', 'solver': 'liblinear'}	0,65	0,62	0,58	0,60	0,59	0,67	0,73	0,71
SI_above_median	KNN-Clf	{'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}	0,67	0,65	0,63	0,59	0,61	0,72	1,12	0,12
SI_above_median	XGB-Clf	{'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50, 'subsample': 0.8}	0,69	0,71	0,71	0,60	0,65	0,73	0,64	182,95
SI_above_median	SVC	{'C': 100, 'class_weight': None, 'gamma': np.float64(0.001), 'kernel': 'rbf', 'probability': True}	0,68	0,65	0,63	0,60	0,61	0,69	0,65	10,37
SI_above_median	GradBoost-Clf	{'learning_rate': 0.01, 'max_depth': 3, 'min_samples_split': 10, 'n_estimators': 200, 'subsample': 0.8}	0,69	0,68	0,68	0,59	0,63	0,73	0,61	141,19

Все модели показали умеренное качество (Accuracy: 0.62-0.71), что ниже результатов для CC50 и IC50. Прогнозирование высокой селективности (SI выше медианы) является наиболее сложной задачей. XGBoost демонстрирует

наилучшее общее качество, но требует дополнительных мер для улучшения низкого Recall. Для практического применения в скрининге рекомендуется использовать LogReg с балансировкой классов, чтобы минимизировать пропуск перспективных соединений. Также для улучшения модели можно добавить признаки, специфичные для селективности.

Классификация по значению 8 для SI

Target	Model	Best_Params	Best Score	Test Accuracy	Test Precision	Test Recall	Test F1	Test ROC AUC	Test Log loss	Search Time
SI_above_8	RandomForest-Clf	{'class_weight': None, 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50}	0,73	0,71	0,70	0,32	0,44	0,73	0,56	26,46
SI_above_8	LogReg	{'C': 1, 'class_weight': None, 'penalty': 'l1', 'solver': 'liblinear'}	0,72	0,71	0,64	0,39	0,49	0,70	0,62	0,91
SI_above_8	KNN-Clf	{'n_neighbors': 3, 'p': 1, 'weights': 'uniform'}	0,72	0,69	0,58	0,48	0,52	0,69	4,45	0,12
SI_above_8	XGB-Clf	{'colsample_bytree': 0.8, 'gamma': 0.1, 'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 100, 'subsample': 0.9}	0,73	0,70	0,66	0,30	0,41	0,69	0,59	181,69
SI_above_8	SVC	{'C': 10, 'class_weight': None, 'gamma': np.float64(0.001), 'kernel': 'rbf', 'probability': True}	0,73	0,69	0,60	0,37	0,46	0,68	0,60	9,87
SI_above_8	GradBoost-Clf	{'learning_rate': 0.01, 'max_depth': 5, 'min_samples_split': 2, 'n_estimators': 100, 'subsample': 0.9}	0,74	0,69	0,61	0,32	0,42	0,67	0,60	139,61

Все модели показали низкое качество, особенно в Recall (0.30-0.48), модели пропускают 52-70% истинно высокоселективных соединений (SI > 8). Для обнаружения потенциально эффективных соединений – это может быть критично в связи с пропуском значительного количества перспекивных кандидатов. Модель KNN показывает лучший Recall, но требует более тонкой настройки. Возможно, ансамблирование улучшило бы ситуацию. Повысить качество модели могло бы также добавление специфических молекулярных дескрипторов.

Заключение

В ходе исследования были разработаны и протестированы модели машинного обучения для прогнозирования ключевых фармакологических параметров: ингибирующей активности (IC50), цитотоксичности (CC50) и индекса селективности (SI). Наилучшие результаты достигнуты для прогнозирования IC50 и CC50, где ансамблевые методы (Random Forest и Gradient Boosting) показали высокую точность, объясняя до 47% дисперсии данных в регрессионных задачах и достигая Accuracy 0.82 в классификации. Для задач классификации цитотоксичности Random Forest продемонстрировал исключительную точность прогнозов с Precision 0.86, что особенно ценно для минимизации ложного срабатывания.

Наибольшие сложности возникли при прогнозировании селективности (SI), особенно для экстремальных значений (SI > 8), где модели показали неприемлемо низкий Recall (<=0.48), что указывает на систематический пропуск перспективных соединений. Эта проблема обусловлена производной природой показателя SI, накапливающего ошибки предсказаний IC50/CC50, и экстремальным дисбалансом классов в данных.