Курсовая работа. Классическое машинное обучение. Отчёт.

Мусатов Матвей Геннадьевич

03.06.2025

Содержание

1 Введение

Первым этапом разработки лекарственных средств является выявление потенциально активных и безопасных молекул на ранней стадии, до этапа дорогостоящих in vitro и in vivo испытаний. В этой связи особую роль играют методы химоинформатики и машинного обучения, позволяющие прогнозировать фармакологические свойства соединений на основе их молекулярной структуры.

В данной работе рассматривается задача прогнозирования трех ключевых биологических показателей:

- IC_{50} (полумаксимальная ингибирующая концентрация) — характеризует эффективность соединения как ингибитора определённого биологического процесса или мишени.
- CC_{50} (полумаксимальная цитотоксическая концентрация) — отражает токсичность соединения для клеток, то есть концентрацию, при которой оно убивает 50% клеток.
- SI (Selectivity Index) — индекс селективности, рассчитывается как отношение IC_{50}/CC_{50} и отражает баланс между активностью и токсичностью. Чем выше SI, тем безопаснее и избирательнее считается соединение.

Целью курсовой работы является разработка и оценка моделей машинного обучения (регрессии и классификации), способных предсказывать значения этих показателей на основе структурных молекулярных дескрипторов. Для этого используется обширный набор признаков, включающий:

- Общие молекулярные дескрипторы, такие как молекулярная масса (MolWt), количество тяжёлых атомов (HeavyAtomCount), доля sp^3-гибридизованных атомов углерода (FractionCSP3), топологическая полярная поверхность (TPSA), и др.
- Электронные дескрипторы, такие как экстремальные значения частичных зарядов (MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge), и др.
- Топологические дескрипторы, такие как индексы связности Чи (Chi0, Chi1, Chi2, ..., Chi4v), индексы Кьера (Kappa1, Kappa2, Kappa3), и др.
- Фрагментные дескрипторы, отражающие наличие или количество определённых химических групп или структурных фрагментов в молекуле.
- Структурные количественные дескрипторы, такие как количество акцепторов/доноров водородных связей (NumHAcceptors, NumHDonors), количество вращающихся связей (NumRotatableBonds), и др.

Для достижения поставленной цели были выполнены следующие шаги:

1. Проведен исследовательский анализ данных (EDA), включающий предварительный просмотр данных, проверку структуры данных, оценку пропущенных значений и выбросов, а также анализ распределений признаков.

2. Выполнена предобработка данных, включающая удаление признаков с единственным уникальным значением, заполнение пропущенных значений медианой, проверку и фильтрацию отрицательных значений целевых переменных, логарифмирование целевых переменных для улучшения нормальности распределения, обработку выбросов методом межквартильного размаха (IQR), бинаризацию редких признаков, корреляционный и дисперсионный анализ для удаления избыточных признаков.

3. Построены и оценены модели регрессии для количественного предсказания значений IC_{50}, CC_{50} и SI.

4. Построены и оценены модели классификации для разделения соединений на классы, например, активные/неактивные (по порогу IC_{50}), токсичные/нетоксичные (по порогу CC_{50}), селективные/неселективные (по порогуSI).

Цель построения моделей регрессии:

- Насколько точно можно предсказать конкретные значения активности, токсичности и селективности соединений на основе их структуры.
- Какие структурные характеристики наиболее связаны с высокой активностью (низким IC_{50}) или высокой токсичностью (низким CC_{50}).
- Возможность получения «идеального» соединения с высоким SI (высокой селективностью).
- Насколько точно можно использовать дескрипторы для количественного предсказания (оценка применимости модели в химическом дизайне).

Цель построения моделей классификации:

- Можно ли надежно классифицировать соединения по заранее заданным биологическим критериям.
- Какие признаки особенно важны для различения классов.
- Возможность использования модели для отсева неэффективных или опасных соединений на ранних этапах.

Выводы по построению моделей:

- Какие структурные особенности чаще встречаются у активных/безопасных соединений.
- Эффективность моделей для задач высокопроизводительного скрининга (HTS).
- Применимость моделей для приоритизации кандидатов на синтез.

2 Описание данных

В данной работе рассматривается задача прогнозирования трех ключевых биологических показателей:

- IC50 (полумаксимальная ингибирующая концентрация) — характеризует эффективность соединения как ингибитора определённого биологического процесса или мишени.
- CC50 (полумаксимальная цитотоксическая концентрация) — отражает токсичность соединения для клеток, то есть концентрацию, при которой оно убивает 50% клеток.
- SI (Selectivity Index) — индекс селективности, рассчитываемый как отношение CC50/IC50 и отражающий баланс между активностью и токсичностью. Чем выше SI, тем безопаснее и избирательнее считается соединение.

Целью курсовой работы является разработка и оценка моделей машинного обучения (регрессии и классификации), способных предсказывать значения этих показателей на основе структурных молекулярных дескрипторов. Для этого используется обширный набор признаков, включающий:

2.1 Общие молекулярные дескрипторы

- **MolWt**: молекулярная масса.
- **HeavyAtomCount**: количество тяжёлых атомов (не включая водород).
- **NumValenceElectrons**: общее количество валентных электронов.
- **NumRadicalElectrons**: количество неспаренных (радикальных) электронов.
- **FractionCSP3**: доля sp^3-гибридизованных атомов углерода.
- **TPSA**: топологическая полярная поверхность, оценивает проницаемость через мембраны.
- **LabuteASA**: аппроксимация доступной поверхности (ASA) по методу Labute. Этот метод используется для оценки площади молекулярной поверхности, доступной для взаимодействия с растворителем или другими молекулами.
- **QED**: комплексная числовая оценка «лекарственности» молекулы, основанная на совокупности её химических и физических свойств. Помогает предсказать, насколько соединение подходит для разработки лекарственных препаратов.
- **SPS**: предполагаемая сложность/стоимость синтеза (если доступна). Она помогает предсказать, насколько трудоёмким и дорогим будет получение молекулы в лаборатории или на производстве.
- **MolLogP**: логарифм коэффициента распределения (гидрофобность).
- **MolMR**: молекулярная рефрактивность (показатель поляризуемости).

Примечание: Дескриптор SPS можно исключить, так как он не соответствует поставленной задаче и не несёт дополнительной информативности в рамках целевой постановки.

2.2 Электронные дескрипторы

- **MaxPartialCharge/MinPartialCharge/MaxAbsPartialCharge/MinAbsPartialCharge** экстремальные значения частичных зарядов.
- **PEOE_VSA**: группа дескрипторов, отражающих распределение зарядов, вычисленных методом уравнивания орбитальной электроотрицательности (PEOE).
- **EState_VSA**: объединение информации о зарядовом состоянии и их топологическом расположении в молекуле.
- **MaxEStateIndex/MinEStateIndex/MaxAbsEStateIndex/MinAbsEStateIndex**: экстремальные значения индекса электротопологического состояния.

2.3 Топологические дескрипторы

- **Chi0, Chi1, Chi2, ..., Chi4v**: индексы связности Чи, отражающие молекулярную топологию и связанные с числом связей, типом атомов и степенью разветвления.
- **Kappa1, Kappa2, Kappa3**: индексы Кьера, которые характеризуют форму молекулы, её компактность и степень разветвлённости.
- **HallKierAlpha**: эмпирический дескриптор стерической насыщенности, т.е. насколько молекула "заполнена"в пространстве.
- **BalabanJ**: индекс связности, учитывающий длину путей и цикличность в молекуле. Позволяет оценить степень разветвленности и топологическую сложность структуры.
- **Ipc, AvgIpc, BertzCT**: информационные и сложностные индексы, характеризующие структурную сложность молекулы на основе анализа её графа.

2.4 BCUT-дескрипторы

- **BCUT2D_MWHI/ MWLOW**: с учётом молекулярной массы.
- **BCUT2D_CHGHI/ CHGLOW**: по заряду.
- **BCUT2D_LOGPHI/ LOGPLOW**: по logP.
- **BCUT2D_MRHI/ MRLOW**: по молекулярной рефрактивности.

2.5 VSA-дескрипторы

- **SMR_VSA1–10**: связаны с молекулярной рефрактивностью.
- **SlogP_VSA1–12**: связь с гидрофобностью (logP).
- **EState_VSA1–10**: электротопология по поверхности.
- **PEOE_VSA1–14**: связь с частичными зарядами.

2.6 Отпечатки (Morgan fingerprints)

- **Morgan fingerprints**: векторные представления молекул, которые кодируют их структурные фрагменты (окружения атомов) с помощью алгоритма, аналогичного распространению по графу. Часто используются для сравнения и поиска похожих соединений в химических базах данных.
- **FpDensityMorgan1, 2, 3**: плотность битов при радиусах 1, 2 и 3 (нормализовано на число атомов).

2.7 Фрагментные дескрипторы

- **Фенолы**: fr_phenol, fr_Ar_OH.
- **Амины**: fr_NH2, fr_amine, fr_aniline.
- **Азосоединения**: fr_azide, fr_azo, fr_diazo.
- **Галогены**: fr_halogen, fr_alkyl_halide.
- **Барбитураты**: fr_barbitur.
- **Нитро-соединения**: fr_nitro, fr_nitro_arom.
- **Лактон/лактам**: fr_lactone, fr_lactam.
- **Кольца**: fr_benzene, fr_pyridine, fr_furan, fr_thiazole.

2.8 Структурные количественные дескрипторы

- **NumHAcceptors/ NumHDonors**: акцепторы/доноры водородных связей.
- **NumRotatableBonds**: количество вращающихся связей.
- **NumAromaticRings/ NumAliphaticRings/ NumSaturatedRings**: кольцевые структуры.
- **NumHeteroatoms**: количество гетероатомов.
- **RingCount**: общее количество колец.

3 Методология

3.1 Исследовательский анализ данных (EDA)

Для начала был проведен исследовательский анализ данных, целью которого являлось предварительное знакомство с данными, проверка их структуры, наличие пропусков и выбросов, а также анализ распределений признаков. Были выполнены следующие шаги:

- Просмотр первых строк данных, информация о типах данных и статистические описательные показатели.

- Удаление признаков с единственным уникальным значением.
- Заполнение пропущенных значений медианой.
- Проверка и фильтрация отрицательных значений целевых переменных (IC_{50}, CC_{50}, SI).
- Логарифмирование целевых переменных для улучшения нормальности распределения.
- Обработка выбросов методом межквартильного размаха (IQR).
- Бинаризация редких признаков (признаки с более 90% нулевых значений).
- Корреляционный и дисперсионный анализ для удаления избыточных признаков.

3.2 Построение моделей регрессии

Для предсказания значений IC_{50}, CC_{50} и SI были построены следующие модели регрессии:

- Линейная регрессия.
- Деревья решений.
- Случайный лес.
- Градиентный бустинг.

Для каждой модели были определены лучшие гиперпараметры с использованием GridSearchCV и кросс-валидации. Качество моделей оценивалось по метрикам RMSE и R^2.

3.3 Построение моделей классификации

Для задач классификации на основе параметров IC_{50}, CC_{50} и SI были построены следующие модели:

- Логистическая регрессия.
- Деревья решений.
- Случайный лес.
- Градиентный бустинг.

Качество моделей оценивалось по метрикам Accuracy, Precision, Recall, F1-score и ROC-AUC.

4 Результаты

4.1 Регрессия

4.1.1 IC_{50}

Лучшей моделью для предсказания IC_{50} оказался градиентный бустинг. Он показал наименьшее значение RMSE (0.42966) и наибольшее значение R^2 (0.52810). Случайный лес также показал хорошие результаты, но немного уступил градиентному бустингу.

Таблица 1: Результаты моделей регрессии для IC_{50}

Модель	RMSE	R^2
Линейная регрессия	0.60805	0.33218
Деревья решений	0.82810	0.09050
Случайный лес	0.44415	0.51220
Градиентный бустинг	**0.42966**	**0.52810**

Таблица 1: Результаты моделей регрессии для IC_{50}

4.1.2 CC_{50}

Лучшей моделью для предсказания CC_{50} также оказался градиентный бустинг. Он показал наименьшее значение RMSE (0.21086) и наибольшее значение R^2 (0.51256).

Таблица 2: Результаты моделей регрессии для CC_{50}

Модель	RMSE	R^2
Линейная регрессия	0.29940	0.30790
Деревья решений	0.43975	-0.01654
Случайный лес	0.21630	0.49999
Градиентный бустинг	**0.21086**	**0.51256**

Таблица 2: Результаты моделей регрессии для CC_{50}

4.1.3 SI

Лучшей моделью для предсказания SI оказался случайный лес. Он показал наименьшее значение RMSE (0.38045) и наибольшее значение R^2 (0.37365).

Таблица 3: Результаты моделей регрессии для SI

Модель	RMSE	R^2
Линейная регрессия	0.56310	0.07295
Деревья решений	0.65495	-0.07826
Случайный лес	**0.38045**	**0.37365**
Градиентный бустинг	0.39808	0.34463

Таблица 3: Результаты моделей регрессии для SI

4.2 Классификация

4.2.1 IC_{50} > медиана

Лучшей моделью для классификации IC_{50} > медиана оказался градиентный бустинг. Он показал наивысшие значения всех метрик (Accuracy = 0.7462, Precision = 0.7667, Recall = 0.7307, F1-score = 0.7487, ROC-AUC = 0.7999).

Таблица 4: Результаты моделей классификации для IC_{50} > медиана

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
Логистическая регрессия	0.6716	0.6792	0.6923	0.6857	0.7186
Деревья решений	0.7114	0.7347	0.6923	0.7129	0.7121
Случайный лес	0.7462	0.7667	0.7307	0.7488	0.7956
Градиентный бустинг	**0.7462**	**0.7667**	**0.7307**	**0.7488**	**0.7999**

Таблица 4: Результаты моделей классификации для IC_{50} > медиана

4.2.2 CC_{50} > медиана

Лучшей моделью для классификации CC_{50} > медиана оказался случайный лес. Он показал наивысшие значения всех метрик (Accuracy = 0.7960, Precision = 0.8333, Recall = 0.7619, F1-score = 0.7960, ROC-AUC = 0.8956).

Таблица 5: Результаты моделей классификации для CC_{50} > медиана

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
Логистическая регрессия	0.7065	0.7169	0.7238	0.7204	0.7874
Деревья решений	0.7313	0.7802	0.6762	0.7245	0.7339
Случайный лес	**0.7960**	**0.8333**	**0.7619**	**0.7960**	**0.8956**
Градиентный бустинг	0.7960	0.8077	0.8000	0.8038	0.8945

Таблица 5: Результаты моделей классификации для CC_{50} > медиана

4.2.3 SI > медиана

Лучшей моделью для классификации SI > медиана оказался случайный лес. Он показал наивысшие значения всех метрик (Accuracy = 0.6866, Precision = 0.6344, Recall = 0.6519, F1-score = 0.6519, ROC-AUC = 0.7363).

Таблица 6: Результаты моделей классификации для SI > медиана

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
Логистическая регрессия	0.6418	0.6129	0.6129	0.6129	0.6509
Деревья решений	0.6666	0.6400	0.6882	0.6632	0.6774
Случайный лес	**0.6866**	**0.6344**	**0.6519**	**0.6519**	**0.7363**
Градиентный бустинг	0.6666	0.6707	0.5914	0.6286	0.7204

Таблица 6: Результаты моделей классификации для SI > медиана

4.2.4 $SI > 8$

Лучшей моделью для классификации $SI > 8$ оказался случайный лес. Он показал наивысшие значения всех метрик (Accuracy = 0.7164, Precision = 0.6400, Recall = 0.4507, F1-score = 0.5289, ROC-AUC = 0.7154).

Таблица 7: Результаты моделей классификации для $SI > 8$

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
Логистическая регрессия	0.6716	0.5556	0.3521	0.4310	0.5974
Деревья решений	0.6667	0.5286	0.5211	0.5248	0.6336
Случайный лес	**0.7164**	**0.6400**	**0.4507**	**0.5289**	**0.7154**
Градиентный бустинг	0.7164	0.8182	0.2535	0.3871	0.7040

Таблица 7: Результаты моделей классификации для $SI > 8$

5 Выводы и рекомендации

5.1 Выводы

- 1. Градиентный бустинг показал себя как наиболее эффективная модель для предсказания значений IC_{50} и CC_{50}.
- 2. Случайный лес оказался наиболее эффективной моделью для предсказания значений SI.
- 3. Градиентный бустинг и случайный лес показали наивысшие значения всех метрик для задач классификации.
- 4. Логистическая регрессия и деревья решений показали средние результаты и могут быть полезны для базового сравнения или в случаях, когда требуется простая и понятная модель.

5.2 Рекомендации

- 1. Для дальнейшего исследования можно рассмотреть использование других методов предобработки данных, таких как нормализация или стандартизация.
- 2. Можно попробовать использовать более сложные модели, такие как нейронные сети или модели с поддержкой векторов опорных точек.
- 3. Для улучшения интерпретируемости моделей можно использовать методы важности признаков и попытаться найти более значимые признаки для предсказания биоактивности лекарственных препаратов.
- 4. Можно провести дополнительный анализ выбросов и аномалий в данных для улучшения качества моделей.
- 5. Для задач классификации можно рассмотреть использование дополнительных метрик, таких как Matthews correlation coefficient (MCC) или Cohen's kappa.

- 6. Можно рассмотреть возможность объединения нескольких моделей в ансамбль для улучшения общего качества предсказаний.
- 7. Для улучшения стабильности моделей можно использовать методы регуляризации, такие как Lasso или Ridge регрессия.

6 Заключение

В данной курсовой работе были выполнены все этапы исследовательского анализа данных, предобработки данных, построения и оценки моделей машинного обучения. Были получены качественные результаты для предсказания значений IC_{50}, CC_{50} и SI, а также для задач классификации на основе этих параметров. Полученные результаты могут быть использованы для дальнейшего исследования и разработки новых методов предсказания биоактивности лекарственных препаратов.