**Вывод по курсовой работе**

**Цель исследования**: построить прогностические модели, которые помогут при разработке лекарственных препаратов, на основе предоставленных данных различных химических соединений и показателей IC₅₀, CC₅₀ и SI – эффективности и токсичности веществ.

**Актуальность исследования**: способы упростить и ускорить процесс создания лекарств востребованы в настоящее время, так как часто специалистам приходится разрабатывать новые препараты в короткие сроки. Также подобные исследования важны из-за того, что помогают сфокусироваться на более перспективных данных с помощью отсеивания нерелевантных и неэффективных признаков.

**Задачи исследования**: подготовить предварительный анализ данных (EDA), обработать признаки должным образом, применить Feature Selection для выбора наиболее коррелирующих признаков, протестировать несколько алгоритмов машинного обучения для поиска наилучшего, применить выбранный алгоритм для выполнения задач по регрессии (IC₅₀, CC₅₀ и SI) и классификации (IC₅₀ > median, CC₅₀ > median и SI > median и SI > 8), выбрать самые важные признаки для успешного проектирования лекарств.

**Предварительный анализ данны**х: размерность данных составляет (1001, 214), где 214 – число признаков, 1001 – число наблюдений. Целевые переменные сильно разрежены и содержат выбросы и пропущенные значения, поэтому после удаления выбросов и заполнения пропусков средним по столбцу было проведено логарифмирование IC₅₀, CC₅₀ и SI. До логарифмирования задавалось значение SI > 8 в отдельную переменную, а после задавались переменные для медиан по каждой из целевых переменных. Затем с помощью SelectKbest, mutual\_info\_regression и VarianceThreshold выбирались 50 наиболее важных для каждой целевой переменной признаков – X был задан отдельно для каждой из них. После этого X\_new\_ic, X\_new\_cc, X\_new\_si стандартизировались и создавались тренировочные и тестовые выборки по каждой из целевых переменных для задач регрессии и классификации (всего 7 пар X\_train и X\_test).

**Подбор алгоритмов МО для решения поставленных задач**: для задач регрессии сначала было протестировано использование исключительно алгоритмов машинного обучения с учителем (линейная регрессия, случайный лес и градиентный бустинг) на IC₅₀, однако метрика R2-score была низкой - 0.07293, 0.19389 и 0.14596 соответственно. Поэтому было принято использовать кластеризацию на каждом X-датасете по целевой переменной. С помощью кластеризации удалось выявить, что данные хорошо разделяются на несколько кластеров, так как представленные в исходном датасете химические соединения представляют собой разные физико-химические классы с разными свойствами, поэтому использование их всех вместе давало плохие метрики. Разделение данных на три кластера дало хорошие результаты: метрики значительно улучшились. Сочетание градиентного бустинга и кластеризации дало самые лучшие метрики на целевой переменной IC₅₀, поэтому было принято решение использовать их для всех задач регрессии на CC₅₀ и SI. Кластер 0 в задачах регрессии несет в себе самые низкие значения, однако кластеры 1 и 2 весьма информативны:

Таблица . R2-score по задачам регрессии с применением кластеризации и градиентного бустинга

|  |  |  |  |
| --- | --- | --- | --- |
|  | IC₅₀ | CC₅₀ | SI |
| Кластер 0 | 0.730 | 0.687 | 0.519 |
| Кластер 1 | 0.918 | 0.964 | 0.913 |
| Кластер 2 | 0.946 | 0.978 | 0.893 |

По проведенному анализу был сделан вывод о том, что соединения, принадлежащие кластеру 1 и 2 в IC₅₀ и CC₅₀, больше всего влияют на кластеры 1 и 2 в SI – а так как данная целевая переменная является их производной, то можно сделать вывод о том, что соединения, принадлежащие данным кластерам 1 и 2, обладают необходимыми физико-химическими свойствами для создания более эффективных лекарств.

В задачах классификации были также протестированы на IC₅₀ различные алгоритмы – логистическая регрессия, метод k ближайших соседей, случайный лес. В этот раз метрика accuracy была изначально не слишком плохой – 0.625, 0.625 и 0.64 соответственно, однако их также можно было улучшить. Была также применена кластеризация X по каждой из целевых переменных – лучшего всего в сочетании с кластеризацией отработал алгоритм случайного леса, поэтому данное сочетание применялось для оставшихся задач классификации. Метрики по каждой из них приведены в таблице ниже:

Таблица . Accuracy для задач классификации с применением кластеризации и случайного леса

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IC₅₀ > median | CC₅₀ > median | SI > median | SI > 8 |
| Кластер 0 | 0.92409 | 0.96700 | 0.94719 | 1.00000 |
| Кластер 1 | 0.97260 | 0.97260 | 0.96164 | 1.00000 |
| Кластер 2 | 0.94231 | 0.96154 | 0.94231 | 1.00000 |

В задачах классификации с участием медианы наблюдается одна общая тенденция – несмотря на то, что все кластеры информативны, наиболее важным является кластер 1. Можно сделать вывод, что минимум в 0.92 случаях значения IC₅₀, CC₅₀ и SI превышают свою медиану. В задаче SI > 8 значение достигает своего максимума – следовательно, во всех представленных видах соединений SI превышает данное значение. Данные выводы могут упростить процесс разработки лекарств с опорой на данные показатели.

**Общий вывод по проделанной работе и визуализация признаков**: для иллюстрации сделанных выводов выведем визуализацию кластеров для каждой из целевых переменных и выведем список признаков по компонентам и их вклад во влияние на каждую из компонент (использовался метод главных компонент PCA).

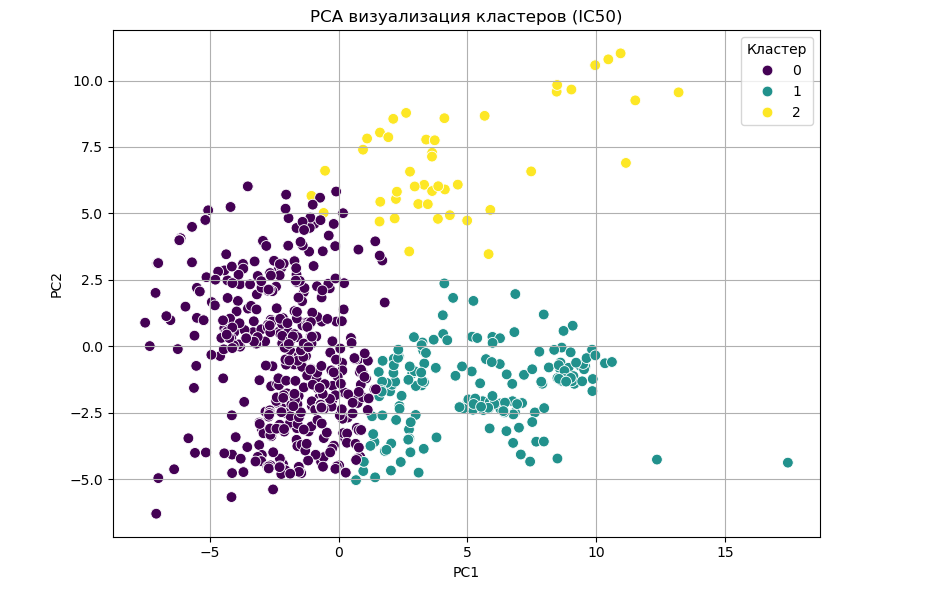


Рисунок 1. Визуализация кластеров для IC50

Список признаков для IC₅₀:

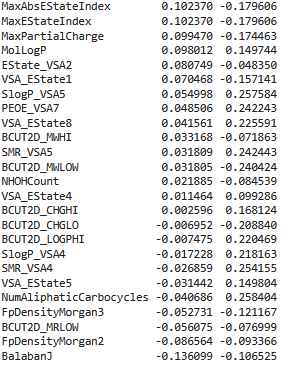
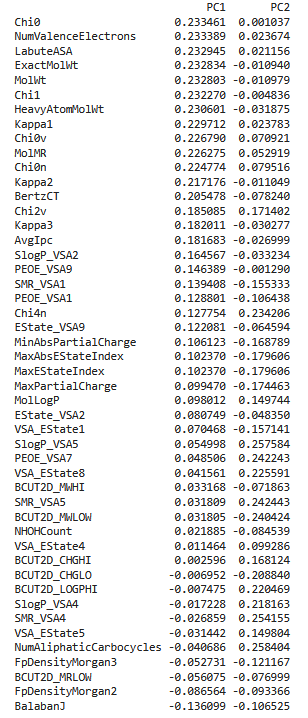


Рисунок 2. Признаки и их вклад в компоненты IC50

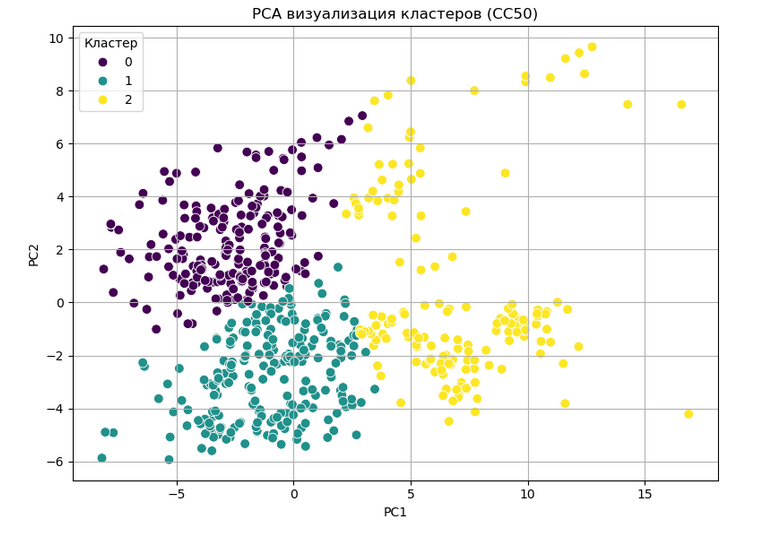


Рисунок 3. Визуализация кластеров для CC50

Список признаков для CC₅₀:

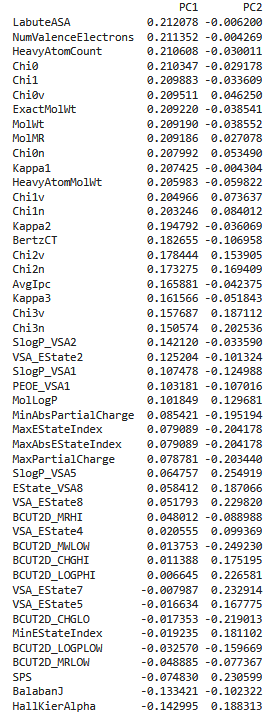
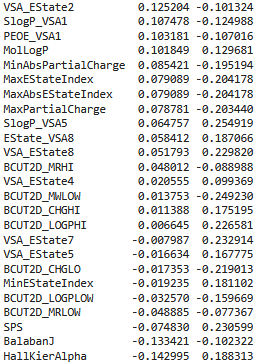
 

Рисунок 4. Признаки и их вклад в компоненты CC50

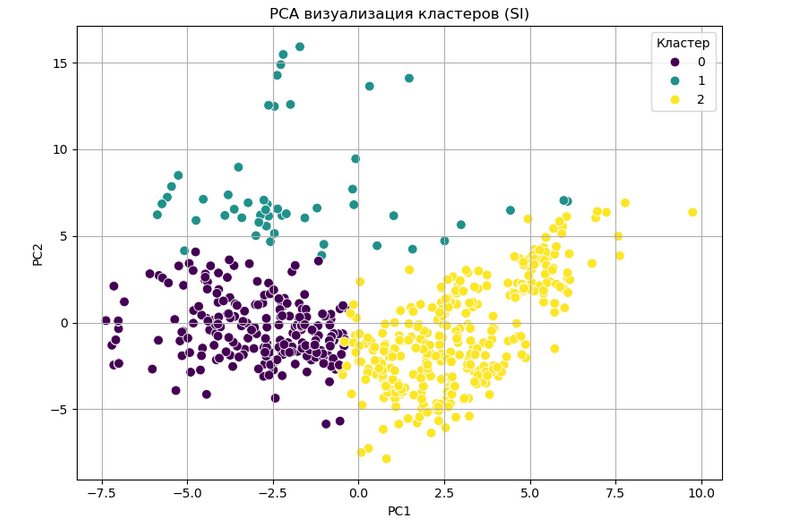


Рисунок 5. Визуализация кластеров для SI

Список признаков для SI:

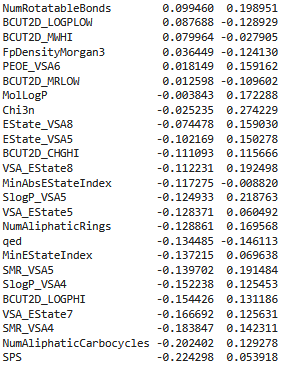
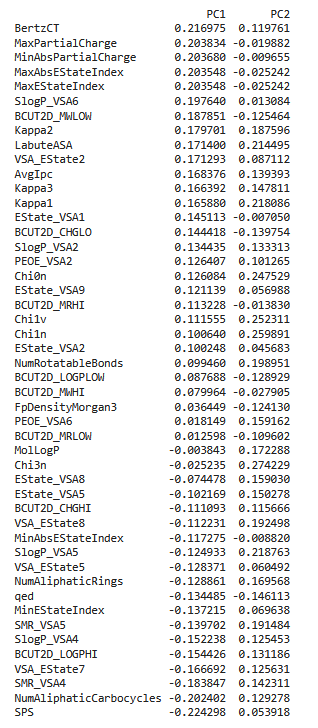


Рисунок 6. Признаки и их вклад в компоненты SI

Признаки, указанные на рисунках для каждой компоненты, наиболее важны для кластеров. Таким образом, на них можно в значительной степени опираться при разработке новых лекарств с ориентиром на один из целевых показателей - IC₅₀, CC₅₀ и SI. Важность кластерного анализа заключается в том, что она помогает учитывать внутреннюю структуру и физико-химические свойства химических соединений, что позволяет соблюдать необходимые тонкости при создании новых препаратов. Примененные модели регрессии позволяют проанализировать эффективность и токсичность соединений еще до их синтеза, а классификация позволяет оценить, пригодны ли созданные лекарства для реализации на потребителей. Применение методов машинного обучения и кластеризации позволяет ускорить анализ соединений, сделать его более эффективным с помощью концентрации на более «успешных» соединениях и кластерах, а также понять внутреннюю структуру химических соединений, что также упрощает работу исследователям и позволяет концентрироваться на отдельных классах веществ.