Аналитический отчет по курсовому проекту

"Прогноз эффективности противовирусных соединений с использованием методов машинного обучения"

Проблематика

В современной фармацевтической индустрии разработка новых препаратов – сложный, многоступенчатый процесс, который требует значительных временных, интеллектуальных и финансовых ресурсов. Дополнительным вызовом является необходимость комплексного тестирования препаратов на различных этапах разработки.

Временные затраты - классический процесс разработки лекарственного препарата может занимать от 10 до 15 лет, что существенно замедляет появление жизненно важных медикаментов на рынке.

Высокая стоимость исследований - экспериментальное тестирование тысяч химических соединений требует колоссальных финансовых вложений, при этом большинство кандидатов оказываются неэффективными.

Неопределенность прогнозирования – сложность предсказания эффективности и безопасности соединений в краткосрочной и долгосрочной перспективе без проведения дорогостоящих экспериментов.

Учитывая данную проблематику, применение методов машинного обучения на различных этапах разработки лекарственных соединений может существенно дополнить традиционные методы и улучшить результаты. Это достигается за счет синергии подходов, замены части дорогостоящих реальных исследований математическим прогнозированием, эффективной обработки и систематизации больших объемов информации, а также минимизации влияния субъективного восприятия исследователя на процесс отбора перспективных соединений.

Постановка задачи

Для анализа предоставлены данные о 1000 химических соединениях с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность, обозначаются как IC50 (концентрация полумаксимального ингибирования), CC50 (концентрация полумаксимальной цитотоксичности) и SI (индекс селективности).

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Основные задачи:

1. Провести исследовательский анализ данных (EDA)
2. Построить модели регрессии для прогнозирования IC50, CC50 и SI
3. Построить модели классификации для определения превышения медианных значений IC50, CC50 и SI
4. Построить модель классификации для определения SI > 8

Ожидаемые результаты:

EDA – провести и описать комплексный анализ данных, выявить значимые признаки и закономерности, на их основе подготовить данные для эффективного решения поставленных задач.

Решение задач регрессии и классификации – сравнить различные модели с настройкой гиперпараметров, сделать выводы о применимости методов, дать рекомендации по улучшению качества прогнозирования и практическому применению результатов.

Предварительная обработка данных

Датасет (размерностью 1001x214) содержит данные по 1001 химическому соединению, каждая строка соответствует определенному веществу, 3 столбца – целевые характеристики вещества, остальные столбцы – его характеристики (структурные, химические, физические и молекулярные).

	Unnamed: 0	IC50, mM	CC50, mM	SI	MaxAbsEStateIndex	MaxEStateIndex	MinAbsEStateIndex	MinEStateIndex	qed	SPS	...	fr_sulfide	fr_sulfonamd	fr_sulfone
0	0	6.239374	175.482382	28.125000	5.094096	5.094096	0.387225	0.387225	0.417362	42.928571	...	0	0	0
1	1	0.771831	5.402819	7.000000	3.961417	3.961417	0.533868	0.533868	0.462473	45.214286	...	0	0	0
2	2	223.808778	161.142320	0.720000	2.627117	2.627117	0.543231	0.543231	0.260923	42.187500	...	0	0	0
3	3	1.705624	107.855654	63.235294	5.097360	5.097360	0.390603	0.390603	0.377846	41.862069	...	0	0	0
4	4	107.131532	139.270991	1.300000	5.150510	5.150510	0.270476	0.270476	0.429038	36.514286	...	0	0	0

5 rows × 214 columns

Выполнена проверка данных на наличие пропусков

Первоначальна форма: (1001, 214)
Форма после удаления пропущенных данных: (998, 214)

Анализ целевых переменных IC50, CC50, SI

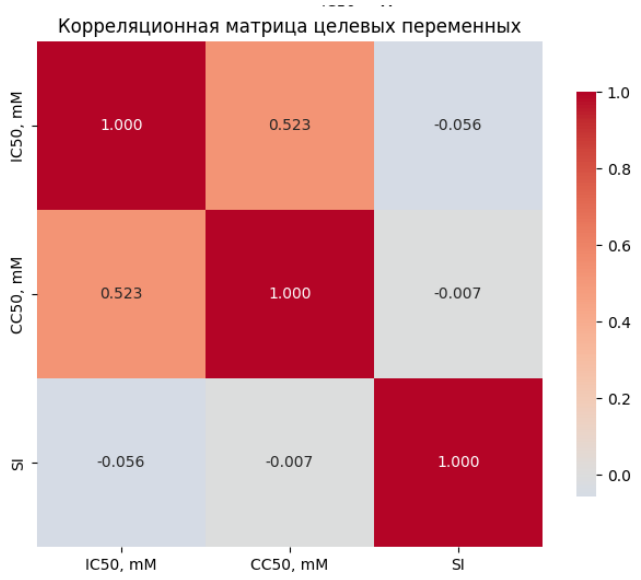
IC50 — это концентрация вещества, при которой оно ингибирует 50% активности целевого белка/фермента/процесса. Чем меньше IC50, тем выше активность соединения.

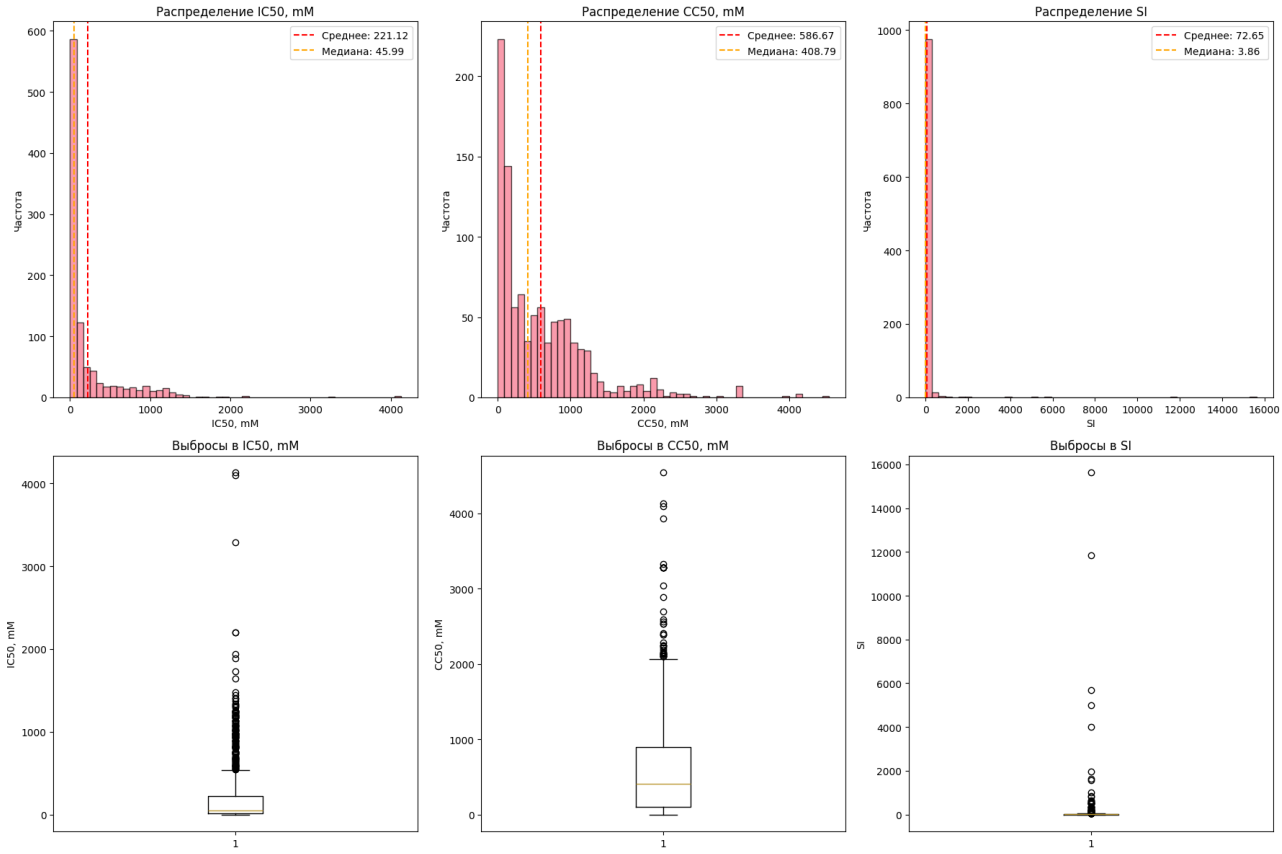
CC50 – это 50% Cytotoxic Concentration — это концентрация вещества, при которой оно вызывает гибель 50% клеток в клеточной культуре. Этот показатель используется для оценки цитотоксичности (токсичности для клеток) химического соединения или лекарства.

SI - Индекс селективности

Для целевых переменны датасета выполняется соотношение IS = CC50 / IC50

	IC50, mM	CC50, mM	SI
count	998.000000	998.000000	998.000000
mean	221.118757	586.668414	72.650005
std	400.510657	642.016454	685.504279
min	0.003517	0.700808	0.011489
25%	12.491340	99.999036	1.457233
50%	45.992006	408.793314	3.856410
75%	224.408630	891.770961	16.525000
max	4128.529377	4538.976189	15620.600000





Анализ выбросов данных:

```
IC50, mM:
  Q1: 12.491, Q3: 224.409, IQR: 211.917
  Границы выбросов: [-305.385, 542.285]
  Количество выбросов: 145 (14.5%)
  Максимальное значение: 4128.529
  Минимальное значение: 0.004

CC50, mM:
  Q1: 99.999, Q3: 891.771, IQR: 791.772
  Границы выбросов: [-1087.659, 2079.429]
  Количество выбросов: 39 (3.9%)
  Максимальное значение: 4538.976
  Минимальное значение: 0.701

SI:
  Q1: 1.457, Q3: 16.525, IQR: 15.068
  Границы выбросов: [-21.144, 39.127]
  Количество выбросов: 124 (12.4%)
  Максимальное значение: 15620.600
  Минимальное значение: 0.011
```

Выполнен **анализ вариативности признаков** - диапазона изменений признака, что поможет исключить константы и почти стабильные признаки. Если некоторые компоненты или характеристики химических соединений практически не меняются, то это не поможет нам найти различия в целевых переменных (хотя возможна, например,ситуация, когда химическое вещество работает как катализатор)

```
Общее количество признаков: 211
Числовые признаки: 211
Признаки с нулевой вариативностью: 18
Признаки с низкой вариативностью: 0
```

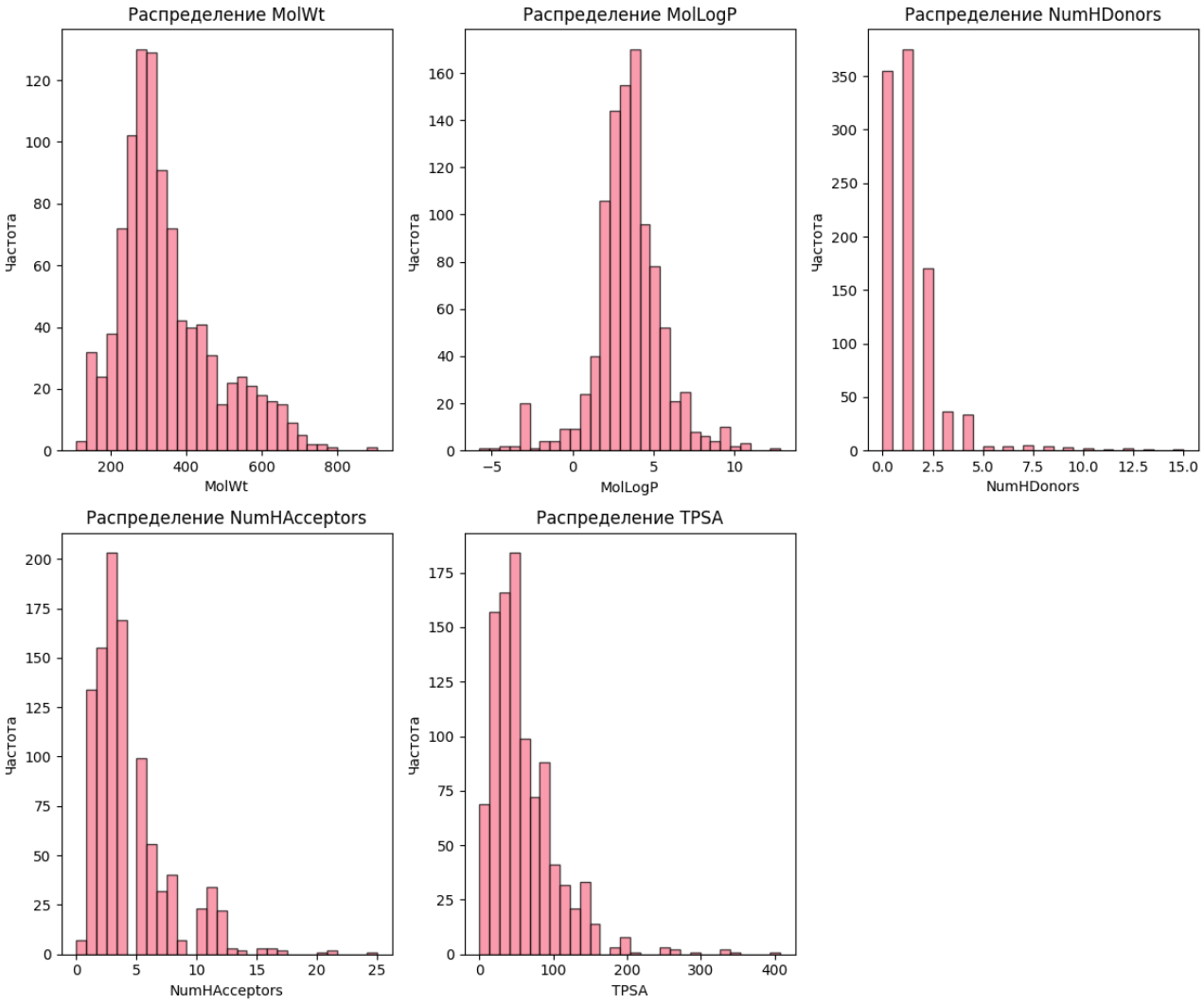
Проанализированы распределения некоторых признаков, например:

MolWt Молекулярная масса (Molecular Weight) - Слишком большие молекулы хуже проникают через клеточные мембраны и плохо абсорбируются

MolLogP - Логарифм распределения между октанолом и водой (гидрофобность) - растворимость и проницаемость вещества.

NumHDonors Количество доноров водородных связей и NumHAcceptors Количество акцепторов водородных связей влияют на биодоступность вещества

TPSA Полярная поверхность (Topological Polar Surface Area) - Связана с способностью к абсорбции и проникновению через биомембраны (включая гематоэнцефалический барьер)



Вывод: Из-за асимметричного распределения и наличия выбросов рекомендуется использовать модели, устойчивые к шуму и выбросам (например, градиентный бустинг, деревья решений) и применять масштабирование и/или преобразование признаков (логарифмирование)

Корреляционный анализ

Топ-10 признаков, наиболее коррелирующих с IC50, mM:

VSA_EState4	0.271743
Chi2n	0.252705
PEOE_VSA7	0.250772
fr_Ar_NH	0.247728
fr_Nhpyrrole	0.247728
Chi2v	0.246602
Chi4v	0.240485

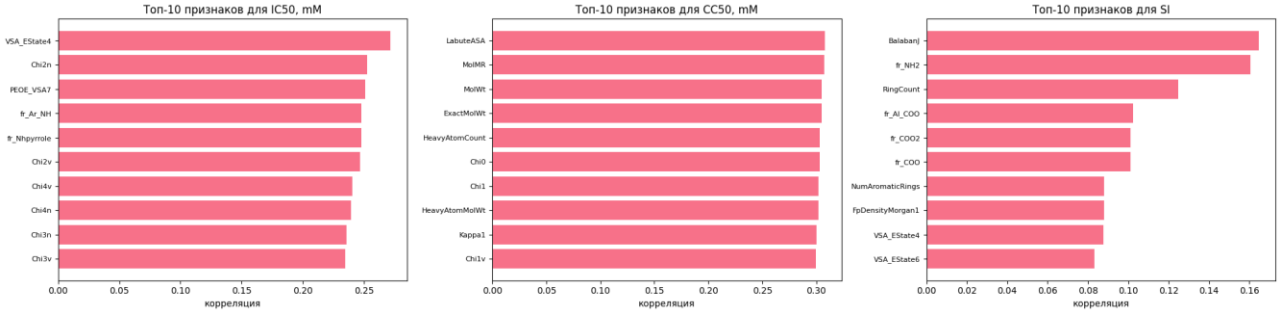
```
Chi4n           0.239209
Chi3n           0.235491
Chi3v           0.234816
```

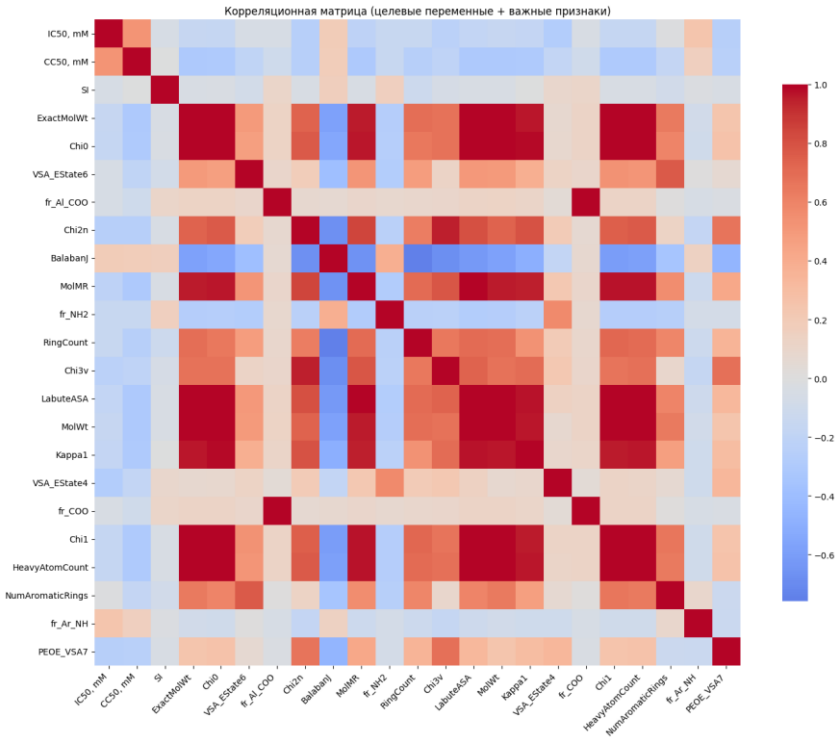
Топ-10 признаков, наиболее коррелирующих с CC50, mM:

```
LabuteASA       0.307774
MolMR           0.307252
MolWt           0.304940
ExactMolWt      0.304883
HeavyAtomCount  0.303065
Chi0            0.302859
Chi1            0.302238
HeavyAtomMolWt  0.302112
Kappa1          0.300393
Chi1v           0.299828
```

Топ-10 признаков, наиболее коррелирующих с SI:

```
BalabanJ          0.164715
fr_NH2            0.160428
RingCount         0.124835
fr_Al_COO         0.102374
fr_COO2           0.101075
fr_COO            0.101075
NumAromaticRings  0.088006
FpDensityMorgan1  0.087894
VSA_EState4       0.087770
VSA_EState6       0.083298
```





По итогам EDA выполнен отбор признаков для каждой целевой переменной:

1. убраны признаки с нулевой и низкой вариативностью и низкой корреляцией с целевой переменной
2. удалены мультиколлинеарные признаки
3. отобраны признаки с помощью модели Random Forest.

Признаки, отобранные для всех целевых переменных

	IC50, mM	CC50, mM	SI
BCUT2D_LOGPLOW	1	1	1
BalabanJ	1	1	1
FpDensityMorgan1	1	1	1
MolLogP	1	1	1
VSA_EState4	1	1	1

На выходе получим датасет для каждой из 3-х целевых переменных с отобранными признаками

```
IC50, mM - selected_features_IC50, mM.csv
CC50, mM - selected_features_CC50, mM.csv
SI -       selected_features_SI.csv
```

Вывод

- Исходный датасет очищен и сокращён до релевантных признаков.
- Были удалены шумовые и дублирующие признаки, учтена мультиколлинеарность.
- Подготовлены 3 оптимизированных датасета (selected_features_SI.csv, selected_features_IC50_mM.csv, selected_features_CC50_mM.csv) по каждому таргету для построения моделей