КУРСОВАЯ РАБОТА

аналитический отчёт

по дисциплине

«Классическое машинное обучение»

на тему

«Разработка моделей машинного обучения для создания лекарственных соединений»

Выполнил:

студент группы М24-525

Шаповал Виктор Николаевич

Москва 2025

ВВЕДЕНИЕ

Создание нового лекарственного препарата – это длительный и ресурсоемкий процесс, включающий несколько ключевых этапов:

- проектирование молекулярной структуры
- химический синтез
- доклинические исследования
- клинические испытания

Каждая стадия требует тщательного анализа и многочисленных экспериментов, что традиционно занимает годы и требует значительных финансовых затрат.
Однако внедрение технологий искусственного интеллекта и машинного обучения открывает новые возможности для ускорения и оптимизации этого процесса. Современные алгоритмы позволяют:

- Предсказывать биологическую активность соединений, сокращая количество необходимых лабораторных тестов
- Оптимизировать молекулярную структуру, подбирая наиболее эффективные и безопасные варианты
- Анализировать большие массивы данных, выявляя новые болезни для лечения
- Снижать риск неудач на поздних стадиях разработки за счет более точного прогнозирования свойств молекул

Таким образом, интеграция машинного обучения в фармацевтические исследования не только ускоряет открытие новых лекарств, но и делает процесс более экономически эффективным, что в перспективе может привести к появлению более доступных и персонализированных методов лечения.

ЦЕЛЬ И ЗАДАЧИ

Цель работы - на основании предоставленных данных от химиков построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Задачи
1) Проанализировать текущие параметры с использованием различных методов
2) Научиться предсказывать эффективность полученных данных

Для решения поставленных задач требуется:

- Провести разведочный анализ – EDA
- Создать модель регрессии для IC50
- Создать модель регрессии для CC50
- Создать модель регрессии для SI
- Создать модель классификации основываясь на превышение медианного значение выборки IC50
- Создать модель классификации основываясь на превышение медианного значение выборки CC50
- Создать модель классификации основываясь на превышение медианного значение выборки SI
- Создать модель классификации основываясь на превышение значение равного 8 выборки SI
- Сравнить между собой полученные модели, выполнить анализ результатов их работы и обосновать выбор наиболее качественных решений

РАЗВЕДОЧНЫЙ АНАЛИЗ – EDA

Разведочный анализ, является основой для построения качественной модели. Таким образом требуется уделить особенное внимание, для наилучшего результата.

После первичного анализа был удалён неинформативный признак 'Unnamed: 0', так как он являлся дублирующим индекс строки.

Затем был проведен построчный анализ на наличие пропусков и дубликатов. Было зафиксировано 3 строки имеющие хотя бы один пропуск и 32 дубликата. Ввиду их небольшого количества – 3,5%, строки были удалены и размерность датасета составила 965x213.

Далее были проанализированы признаки на среднеквадратичное отклонение. Данный анализ позволил выявить столбцы в датасете, имеющие среднеквадратичное отклонение равное нулю. Это говорит о том, что все значения в признаке равны – одинаковы, что является неинформативным показателем для обучения модели. Таким образом были удалены следующие признаки:

- NumRadicalElectrons
- SMR_VSA8
- SlogP_VSA9
- fr_N_O
- fr_SH
- fr_azide
- fr_barbitur
- fr_benzodiazepine
- fr_diazo
- fr_dihydropyridine
- fr_isocyan
- fr_isothiocyan
- fr_lactam
- fr_nitroso
- fr_phos_acid
- fr_phos_ester
- fr_prisulfonamd
- fr_thiocyan

Выбросы в данных были отфильтрованы по превышению 99% квантиля для каждого признака. Данные, попавшие под этот критерий, были заменены на значение перцентиля, то есть на значение 99% квантиля.

После работы с признаками потребовалось оценить целевые переменные на нормальность распределения. Визуальная оценка показала отсутствие нормального распределения, что потребовало применить логарифмирование (рис.1)

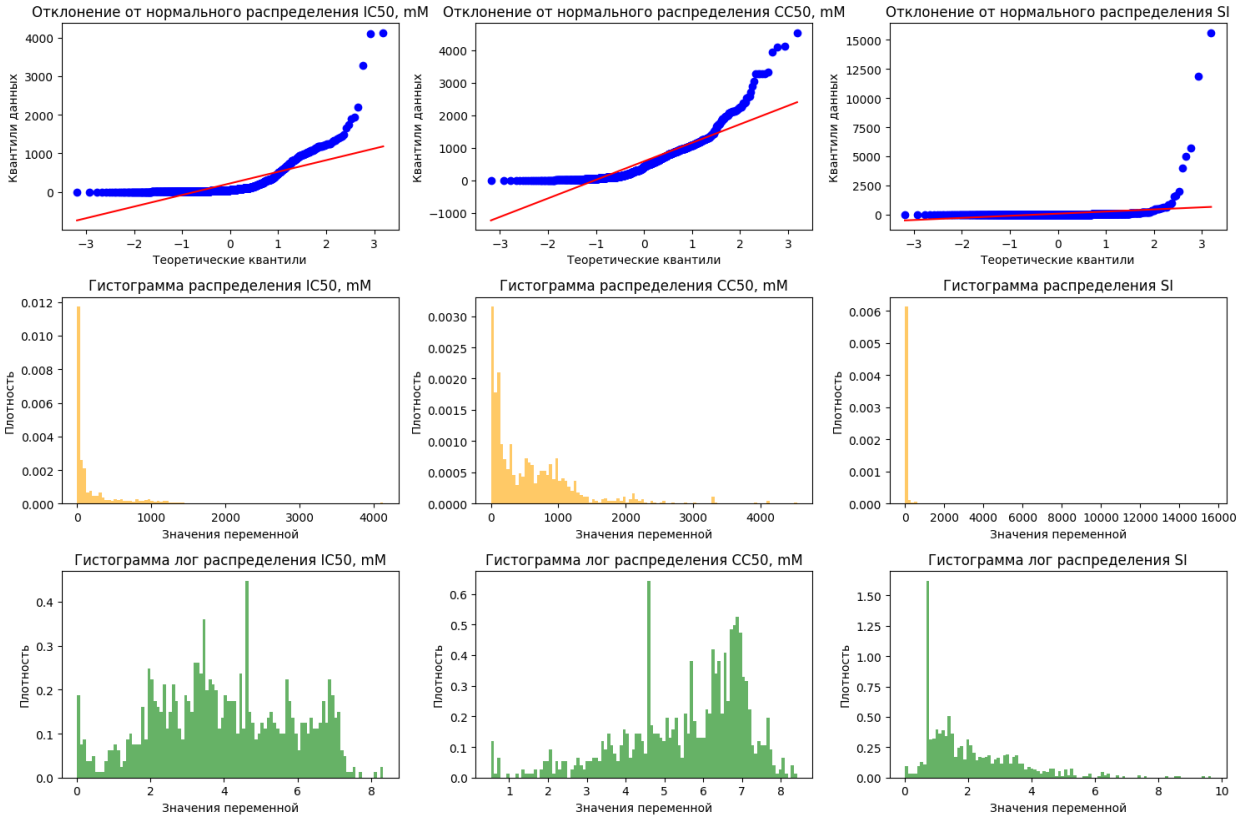

Рисунок 1 – Анализ целевых переменных на нормальное распределение

При помощи метода главных компонент была оценена дисперсия для сохранения 95% данных. Значение составило 50 компонент (рис. 2)

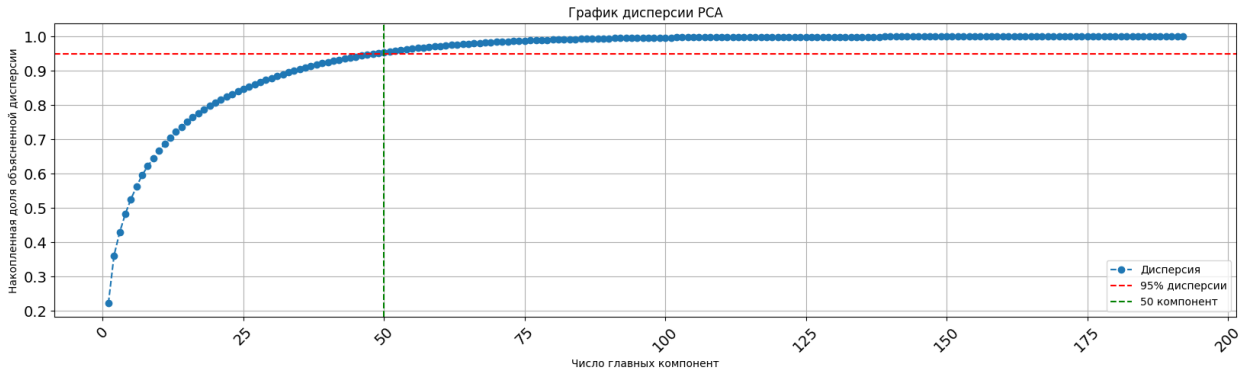

Рисунок 2 – Анализ дисперсии данных

Так же метод Mutual Info позволили оценить зависимость каждой целевой переменной от признаков (рис. 3, рис. 4 и рис. 5)

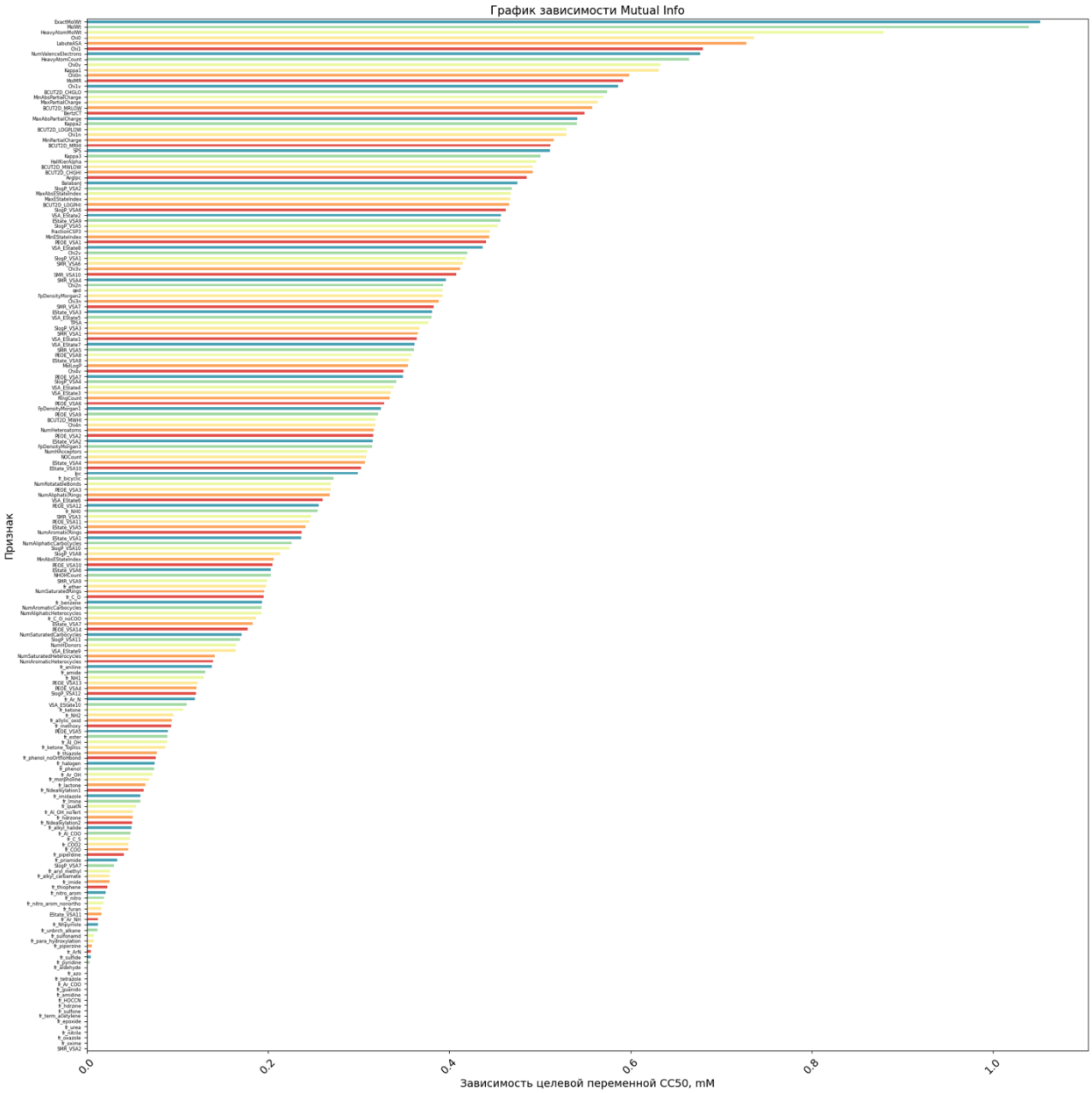

Рисунок 3 – Зависимость целевой переменной CC50, mM

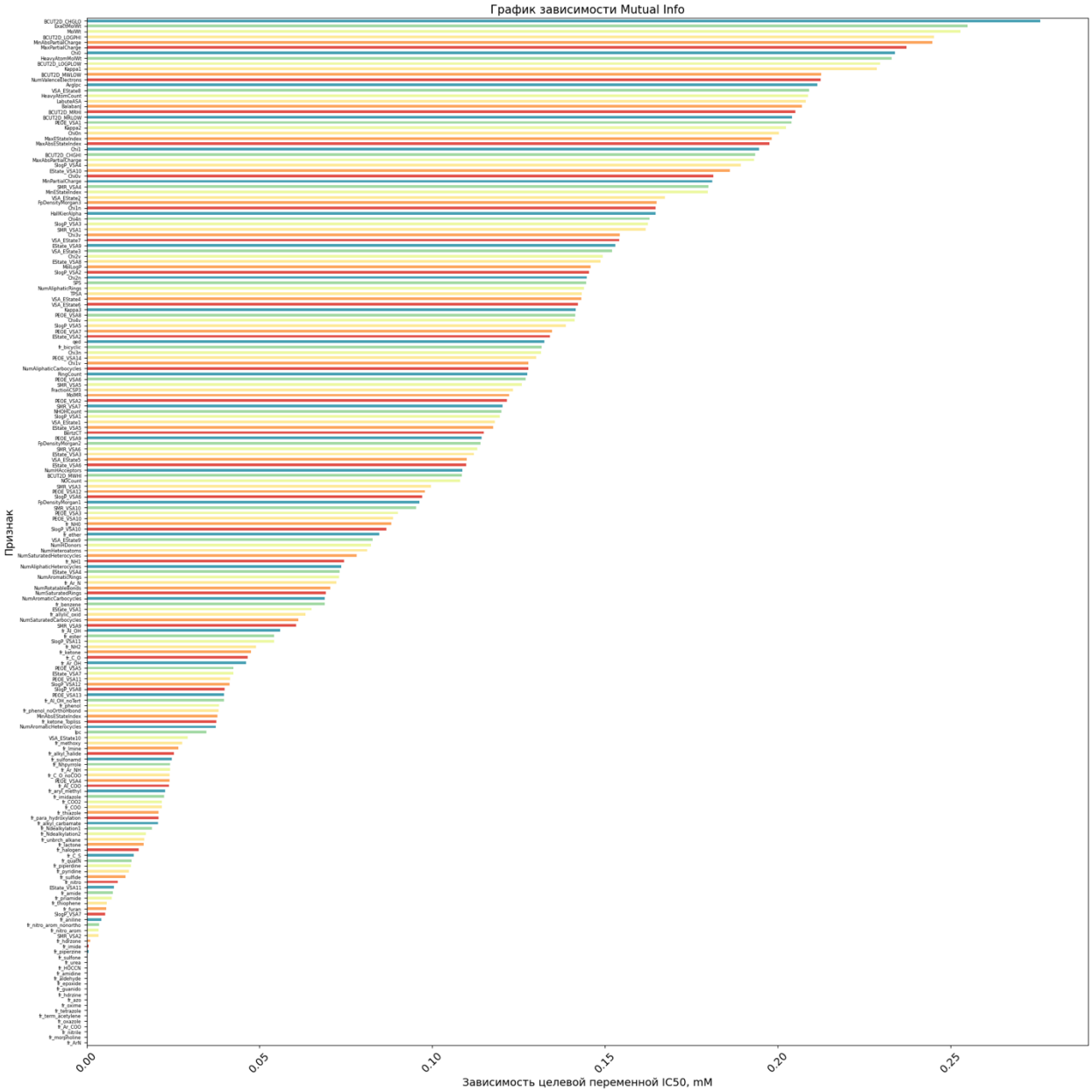

Рисунок 4 – Зависимость целевой переменной IC50, mM

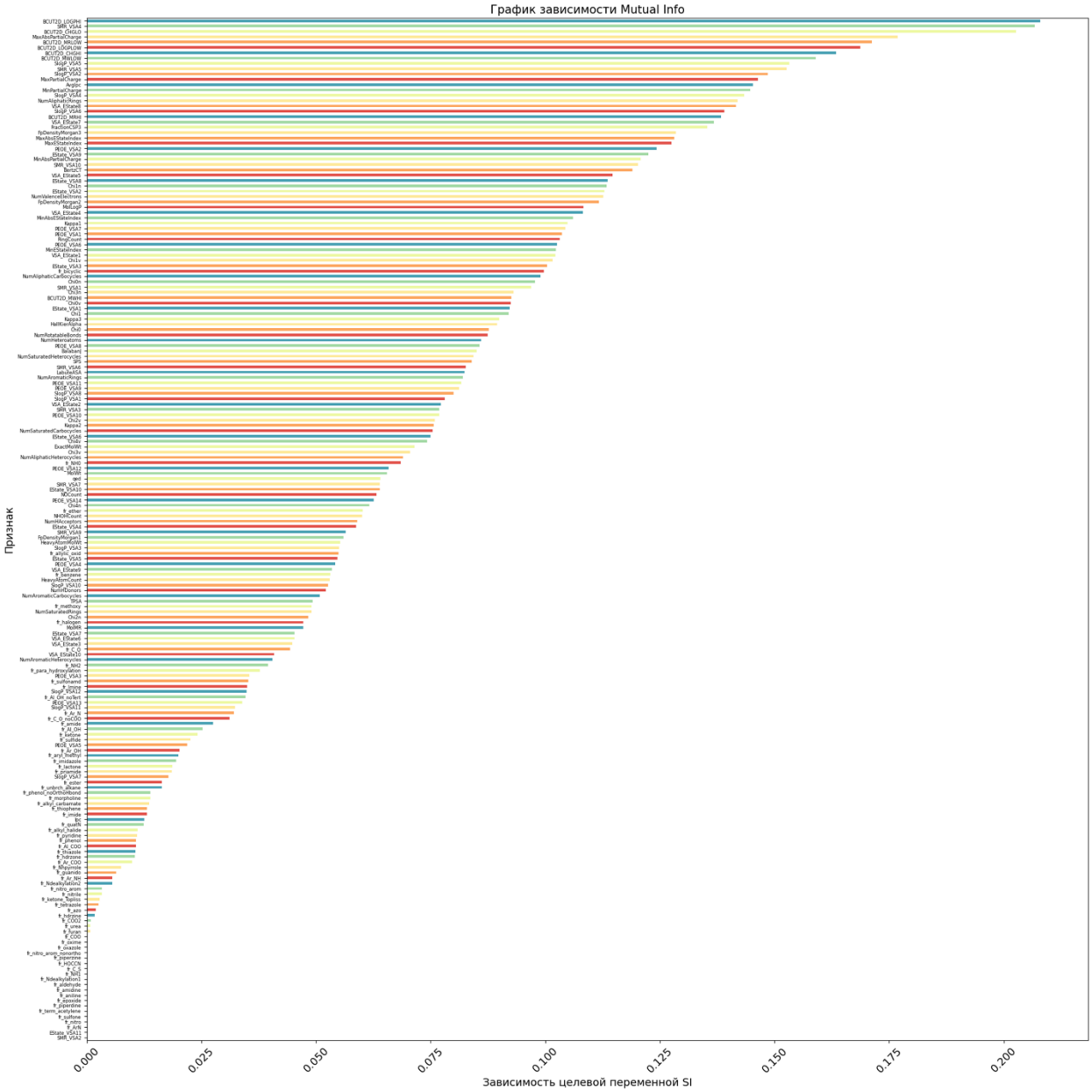

Рисунок 5 – Зависимость целевой переменной SI

Таким образом в дальнейшем при обучении необходимо подходить индивидуально к каждой модели – сокращать количество признаков согласно зависимости каждой переменной, затем делить данные на тестовые и тренировочные, стандартизировать данные, обучать модель и оценивать полученный результат.

Стоит обратить внимание на метрику SI, которая считается на основе метрик IC50 и CC50, следовательно при обучение следует воздержаться от использования метрики SI в качестве признака для целевых переменных IC50 и CC50.

МОДЕЛИ РЕГРЕССИИ

Регрессия — это один из главных методов прогнозного моделирования и работы с данными. Он позволяет установить связь между переменными, чтобы прогнозировать развитие какого-либо явления в будущем.

Для решения задачи регрессии целевой переменной IC50 были дообработаны данные:

- Добавлен дополнительный признак “CC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

1) RandomForestRegressor с параметрами:
   - bootstrap = False
   - max_depth = 47
   - max_features = sqrt
   - min_samples_leaf = 3
   - min_samples_split = 16
   - n_estimators = 218

2) GradientBoostingRegressor с параметрами:
   - max_depth = 1
   - min_samples_leaf = 5
   - min_samples_split = 6
   - n_estimators = 200

3) KNeighborsRegressor с параметрами:
   - algorithm = auto
   - leaf_size = 200
   - n_neighbors = 12

В качестве метрик для оценки работы использовались:

- Среднеквадратичная ошибка (RMSE)
- Коэффициент детерминации (R^2)
- Коэффициент детерминации полученный при обучении

Наилучшие показатели при обучение достигались на модели GradientBoostingRegressor, однако после проверки на тестовых данных, модель RandomForestRegressor показала наилучшие результаты. Данные полученных метрик представлены в табл. 1

Для решения задачи регрессии целевой переменной CC50 были дообработаны данные:

- Добавлен дополнительный признак “IC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

1) RandomForestRegressor с параметрами:
   - bootstrap = False
   - max_depth = 27
   - max_features = sqrt
   - min_samples_leaf = 1
   - min_samples_split = 10
   - n_estimators = 386

2) GradientBoostingRegressor с параметрами:
   - max_depth = 5
   - min_samples_leaf = 6
   - min_samples_split = 2
   - n_estimators = 200

3) KNeighborsRegressor с параметрами:
   - algorithm = brute
   - leaf_size = 200
   - n_neighbors = 10

В качестве метрик для оценки работы использовались:

- Среднеквадратичная ошибка (RMSE
- Коэффициент детерминации (R^2)
- Коэффициент детерминации полученный при обучении

Наилучшие показатели и при обучение и при проверке на тестовых данных удалось достигнуть на модели RandomForestRegressor. Данные полученных метрик представлены в табл. 2

Для решения задачи регрессии целевой переменной SI были дообработаны данные:

- Добавлен дополнительный признак “CC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

1) RandomForestRegressor с параметрами:
   - bootstrap = False
   - max_depth = 27
   - max_features = sqrt
   - min_samples_leaf = 1
   - min_samples_split = 4
   - n_estimators = 311

2) GradientBoostingRegressor с параметрами:
   - max_depth = 19
   - min_samples_leaf = 5
   - min_samples_split = 9
   - n_estimators = 89

3) KNeighborsRegressor с параметрами:
   - algorithm = brute
   - leaf_size = 50
   - n_neighbors = 10

В качестве метрик для оценки работы использовались:

- Среднеквадратичная ошибка (RMSE)
- Коэффициент детерминации (R^2)
- Коэффициент детерминации полученный при обучении

Наилучшие показатели и при обучение и при проверке на тестовых данных удалось достигнуть на модели RandomForestRegressor. Данные полученных метрик представлены в табл. 3

Таблица 1 – Метрики регрессионных моделей IC50

Метрика / Модель	Среднеквадратичная ошибка	Коэффициент детерминации	Коэффициент детерминации при обучении
RandomForestRegressor	1.2439	0.5718	0.5337
GradientBoostingRegressor	1.2530	0.5656	0.5437
KNeighborsRegressor	1.4846	0.3901	0.3834

Таблица 2 – Метрики регрессионных моделей CC50

Метрика / Модель	Среднеквадратичная ошибка	Коэффициент детерминации	Коэффициент детерминации при обучении
RandomForestRegressor	1.0745	0.5964	0.5332
GradientBoostingRegressor	1.1153	0.5651	0.4942
KNeighborsRegressor	1.2756	0.4311	0.3304

Таблица 3 – Метрики регрессионных моделей SI

Метрика / Модель	Среднеквадратичная ошибка	Коэффициент детерминации	Коэффициент детерминации при обучении
RandomForestRegressor	1.2836	0.1940	0.3864
GradientBoostingRegressor	1.2781	0.2009	0.3643
KNeighborsRegressor	1.3795	0.0691	0.201

МОДЕЛИ КЛАССИФИКАЦИИ ПО МЕДИАННОМУ ЗНАЧЕНИЮ ДЛЯ ВЫБОРКИ

Классификация — это один из ключевых методов прогнозного моделирования и работы с данными. Она позволяет разделить данные на категории или классы на основе их признаков, чтобы предсказывать принадлежность новых объектов к определённым группам.

Для решения задачи классификации по целевой переменной IC50 было выполнено преобразование признака IC50, mM по бинарному критерию. Критерий оценки состоял из сравнения медианного значения признака IC50, mM с каждым значением данных построчно. Таким образом данные, превосходившие медианное значение IC50, mM получали значение «True», равные или меньшие получали значение «False».

Также для булевой целевой переменной IC50 были дообработаны данные:

- Добавлен дополнительный признак “CC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

1) RandomForestRegressor с параметрами:
   - bootstrap = False
   - max_depth = 27
   - max_features = sqrt
   - min_samples_leaf = 1
   - min_samples_split = 10
   - n_estimators = 386

2) GradientBoostingRegressor с параметрами:
   - max_depth = 5
   - min_samples_leaf = 6
   - min_samples_split = 2
   - n_estimators = 200

3) CatBoostClassifier с параметрами:
    - depth = 5
    - grow_policy = Depthwise
    - iterations = 103
    - l2_leaf_reg = 1.013135787377226
    - loss_function = Logloss
    - min_data_in_leaf = 15
    - one_hot_max_size = 2

В качестве метрик для оценки работы использовались:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

Наилучшие показатели и при обучение и при проверке на тестовых данных удалось достигнуть на модели CatBoostClassifier. Данные полученных метрик представлены в табл. 4

Для решения задачи классификации по целевой переменной CC50 было выполнено преобразование признака CC50, mM по бинарному критерию. Критерий оценки состоял из сравнения медианного значения признака CC50, mM с каждым значением данных построчно. Таким образом данные, превосходившие медианное значение CC50, mM получали значение «True», равные или меньшие получали значение «False».

Также для булевой целевой переменной CC50 были дообработаны данные:

- Добавлен дополнительный признак “IC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

1) RandomForestRegressor с параметрами:
   - bootstrap = True
   - max_depth =15
   - max_features = sqrt
   - min_samples_leaf = 2
   - min_samples_split = 2
   - n_estimators = 263

2) GradientBoostingRegressor с параметрами:
   - max_depth = 19
   - min_samples_leaf = 5
   - min_samples_split = 9
   - n_estimators = 89

3) CatBoostClassifier с параметрами:
   - depth = 9
   - grow_policy = SymmetricTree
   - iterations = 638
   - l2_leaf_reg = 8.224883030406795
   - loss_function = CrossEntropy
   - min_data_in_leaf = 6
   - one_hot_max_size = 8

В качестве метрик для оценки работы использовались:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

Наилучшие показатели при обучение достигались на модели CatBoostClassifier, однако после проверки на тестовых данных, модель GradientBoostingClassifier показала наилучшие результаты. Данные полученных метрик представлены в табл. 5

Для решения задачи классификации по целевой переменной SI было выполнено преобразование признака SI по бинарному критерию. Критерий оценки состоял из сравнения медианного значения признака SI с каждым значением данных построчно. Таким образом данные, превосходившие медианное значение SI получали значение «True», равные или меньшие получали значение «False».

Также для булевой целевой переменной SI были дообработаны данные:

- Добавлен дополнительный признак “CC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

4) RandomForestRegressor с параметрами:
   - bootstrap = True
   - max_depth =41
   - max_features = log2
   - min_samples_leaf = 3
   - min_samples_split = 12
   - n_estimators = 174

5) GradientBoostingRegressor с параметрами:
   - max_depth = 9
   - min_samples_leaf = 4
   - min_samples_split = 2
   - n_estimators = 198

6) CatBoostClassifier с параметрами:
   - depth = 5
   - grow_policy = SymmetricTree
   - iterations = 408
   - l2_leaf_reg = 5.864048860435419
   - loss_function = Logloss
   - min_data_in_leaf = 21
   - one_hot_max_size = 9

В качестве метрик для оценки работы использовались:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

Наилучшие показатели и при обучение и при проверке на тестовых данных удалось достигнуть на модели CatBoostClassifier. Данные полученных метрик представлены в табл. 6

Для решения задачи классификации по целевой переменной SI было выполнено преобразование признака SI по бинарному критерию. Критерий оценки состоял из сравнения каждого значения данных построчно с 8. Таким образом данные, превосходившие значение 8 получали значение «True», равные или меньшие получали значение «False».

Также для булевой целевой переменной SI были дообработаны данные:

- Добавлен дополнительный признак “CC50, mM” в данные для обучения
- Сокращено количество признаков до 50
- Разделены данные на тестовые и тренировочные
- Проведена стандартизация тренировочных данных и на её основе стандартизированы тестовые данные

Затем были поострены три модели, предварительно, для каждой из которых, были подобраны параметры. Использовались следующие модели:

7) RandomForestRegressor с параметрами:
   - bootstrap = False
   - max_depth = 15
   - max_features = sqrt
   - min_samples_leaf = 1
   - min_samples_split = 2
   - n_estimators = 500

8) GradientBoostingRegressor с параметрами:
   - max_depth = 7
   - min_samples_leaf = 5
   - min_samples_split = 4
   - n_estimators = 200

9) CatBoostClassifier с параметрами:
   - depth = 4
   - grow_policy = Depthwise
   - iterations = 952
   - l2_leaf_reg = 4.978350342428872
   - loss_function = Logloss
   - min_data_in_leaf = 1
   - one_hot_max_size = 8

В качестве метрик для оценки работы использовались:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

Наилучшие показатели при обучение достигались на модели CatBoostClassifier, однако после проверки на тестовых данных, модель GradientBoostingClassifier показала наилучшие результаты. Данные полученных метрик представлены в табл. 7

Таблица 4 – Метрики классификационных моделей IC50

Метрика / Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
RandomForestRegressor	0.8299	0.8571	0.7500	0.8000	0.8266
GradientBoostingRegressor	0.8093	0.8493	0.7045	0.7702	0.8407
KNeighborsRegressor	0.8505	0.8554	0.8068	0.8304	0.8430

Таблица 5 – Метрики классификационных моделей CC50

Метрика / Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
RandomForestRegressor	0.8557	0.8488	0.8295	0.8391	0.8767
GradientBoostingRegressor	0.8505	0.8734	0.7841	0.8263	0.8919
KNeighborsRegressor	0.8660	0.8780	0.8182	0.8471	0.8910

Таблица 6 – Метрики классификационных моделей SI

Метрика / Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
RandomForestRegressor	0.6753	0.6429	0.6923	0.6667	0.7142
GradientBoostingRegressor	0.6804	0.6559	0.6703	0.6630	0.7149
KNeighborsRegressor	0.6959	0.6739	0.6813	0.6776	0.7270

Таблица 7 – Метрики классификационных моделей SI > 8

Метрика / Модель	Accuracy	Precision	Recall	F1-score	ROC AUC
RandomForestRegressor	0.8299	0.8571	0.7500	0.8000	0.8266
GradientBoostingRegressor	0.8093	0.8493	0.7045	0.7702	0.8407
KNeighborsRegressor	0.8505	0.8554	0.8068	0.8304	0.8430

ЗАКЛЮЧЕНИЕ

В рамках данного исследования был разработан программный инструмент, направленный на оптимизацию процесса создания новых фармацевтических препаратов. Ключевой целью работы являлось построение прогноза на основе моделей машинного обучения позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов

Основные этапы исследования:

- Предварительная обработка данных – проведена очистка данных (устранение пропусков, дубликатов и аномальных значений)
- Разработка и тестирование регрессионных моделей – реализован сравнительный анализ различных алгоритмов машинного обучения с подборкой гиперпараметров для каждого алгоритма и целевой переменной
- Разработка и тестирование классификационных моделей – реализован сравнительный анализ различных алгоритмов машинного обучения с подборкой гиперпараметров для каждого алгоритма и целевой переменной

Разработанный инструмент продемонстрировал способность эффективно прогнозировать сочетание параметров для создания лекарственных препаратов, что открывает возможности для сокращения времени и затрат на доклинические исследования.

Перспективные направления для дальнейшего развития проекта:

- Внедрение глубокого обучения и нейронных сетей
- Увеличение масштабов обучающей выборки
- Совместная работа со специалистами предметной области
- Совершенствование методов формирования признаков

Проведённая работа подтвердила достижение поставленных целей. Полученные результаты обладают практической ценностью, однако остаётся потенциал для улучшения прогностической способности моделей за счёт указанных направлений модернизации.