Отчет по курсовой работе Ситёва Р.Р.

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

**Задачи**: Разработать несколько максимально эффективных моделей для решения следующих задач:

* Регрессия для IC50
* Регрессия для CC50
* Регрессия для SI
* Классификация: превышает ли значение IC50 медианное значение выборки
* Классификация: превышает ли значение CC50 медианное значение выборки
* Классификация: превышает ли значение SI медианное значение выборки
* Классификация: превышает ли значение SI значение 8

**Решение:**

* 1. **Разведочный анализ данных (EDA)**

Первым шагом был импорт необходимых библиотек и загрузка исходных данных.

Далее был проведен первичный анализ данных, который показал, что датасет состоит из 1001 наблюдения и 214 признаков. Типы данных в основном числовые. Был обнаружен неинформативный столбец Unnamed: 0, который был удален из данных.

Далее была проведена проверка на наличие полных дубликатов строк. Было выявлено 32 дублирования, которые были удалены, чтобы модель не переобучалась на этих данных.

После удаления дубликатов была выполнена проверка на наличие пропущенных значений. Общее количество пропусков составило 36. Анализ показал, что пропуски находятся в 12 столбцах и затрагивают 3 строки.

Учитывая малое количество строк с пропусками, и чтобы не вносить шум в данные, было принято решение удалить эти строки.

Таким образом, после первичной очистки размер датасета составил 966 наблюдений и 213 признаков.

Далее были проанализированы целевые переменные IC50, CC50 и SI. Результаты показали большой разброс значений и наличие потенциальных выбросов. Для проверки корректности расчета SI был пересчитан на основе CC50 и IC50.

Визуализация распределений целевых переменных подтвердила наличие сильной скошенности и выбросов.

Учитывая наличие значительных выбросов в переменной SI, было принято решение ограничить датасет значениями SI < 1000 для уменьшения влияния экстремальных выбросов.

Были проведены эксперименты с логарифмированием целевых переменных, не приведшие к положительному результату при моделировании. Поэтому было принято решение работать с целевыми признаками как есть.

* 1. **Анализ и отбор признаков (общий)**

Был проведен анализ вариативности всех признаков. Были найдены константные признаки, не несущие полезной информации для моделей. Таких признаков оказалось 18, и они были удалены из набора данных.

Для дальнейшего отбора признаков и борьбы с мультиколлинеарностью признаки были разделены на группы (общие физико-химические, электронные, топологические, поверхностно-объемные, количественные структурные, фрагментные). Внутри каждой группы проводился корреляционный анализ с целью выявления и удаления сильно коррелирующих признаков (с коэффициентом корреляции > 0.9).

**Выводы:**

Проведенный анализ данных позволил:

1. Очистить данные от неинформативных столбцов, дубликатов и пропущенных значений.
2. Проанализировать распределения целевых переменных, выявить их скошенность и наличие выбросов.
3. Cократить признаковое пространство за счет удаления константных и сильно коррелирующих дескрипторов, что помогло уменьшить мультиколлинеарность.
   1. **Регрессия для IC50**

Для построения моделей использовался датасет, полученный после анализа данных и первичного отбора признаков.

Данные были разделены на обучающую и тестовую выборки в соотношении 80:20 с использованием функции train\_test\_split.

Для отбора наиболее информативных признаков для прогнозирования IC50 был использован метод, основанный на взаимной информации между признаками и целевой переменной, реализованный в mutual\_info\_regression.

На основе анализа значений взаимной информации и проведенных экспериментов было принято решение оставить 31 лучший признак для дальнейшего моделирования. Отбор был выполнен с помощью SelectKBest.

Отобранные признаки были масштабированы с использованием StandardScaler для приведения их к единой шкале.

После отбора и масштабирования была проведена повторная проверка на наличие сильно коррелирующих признаков в обучающей выборке. Для уменьшения мультиколлинеарности признаки с коэффициентом корреляции > 0.9 были удалены из обучающей и тестовой выборок.

Помимо mutual\_info\_regression были проведены эксперименты с PCA для сокращения размерности данных и логарифмирование целевой переменной. Однако данные эксперименты не привели к успеху, поэтому было принято решение остановиться на текущей реализации, описанной выше.

Были рассмотрены следующие модели машинного обучения: дерево решений; случайный лес; метод k ближайших соседей; линейная регрессия; метод опорных векторов; градиентный бустинг XGBoost; градиентный бустинг LightGBM.

Подбор гиперпараметров осуществлялся с помощью байесовской оптимизации. В качестве метрики для оптимизации использовалась отрицательная среднеквадратичная ошибка.

После обучения и подбора гиперпараметров, каждая из лучших моделей была оценена на отложенной тестовой выборке. В качестве метрик использовались среднеквадратичная ошибка и коэффициент детерминации.

Лучшие результаты на тестовой выборке показала модель LightGBM Regressor, достигнув наименьшего значения MSE – 108630,17 и наибольшего значения R² - 0,4693.

Модель KNN также показала хорошие результаты R² - 0,3975.

Линейная регрессия и дерево решений показали более скромные результаты, что говорит о наличии нелинейных зависимостей в данных, которые эти модели не смогли уловить.

Наименее эффективной оказалась модель SVR с R² - 0,0039, что говорит о том, что модель практически не смогла объяснить вариативность IC50. Возможно, это связано с отсутствием логарифмирования целевой переменной, т.к. SVR был практически единственным методом, у которого наблюдалось увеличение метрик после логарифмирования, однако этого было недостаточно.

**Выводы**:

1. Наиболее эффективной моделью для прогнозирования IC50 является LightGBM Regressor с подобранными гиперпараметрами.
2. Модели, основанные на градиентном бустинге и методе k-ближайших соседей, показали себя лучше линейных моделей и деревьев решений, что говорит о сложности и нелинейности взаимосвязей между признаками.
3. Коэффициент детерминации на уровне 0,47 для лучшей модели указывает на то, что модель способна улавливать часть закономерностей, однако значительная часть дисперсии остается необъясненной.
4. В качестве дальнейшего развития планируется попробовать провести доп эксперименты с другими методами отбора признаков, а попытаться провести FeatureEngeneering и выделить дополнительные признаки.
   1. **Регрессия для CC50**

Задача регрессии для CC50 решалась схожим образом, но в результате экспериментов было отобрано 15 наиболее информативных признаков, часть из которых была отфильтрована корреляционным анализом.

Лучшие результаты на тестовой выборке показала модель XGBoost Regressor, достигнув наименьшего значения MSE – 144696,45 и наибольшего значения R² - 0,5340.

Также показали хорошие результаты модели LightGBM Regressor и Случайный лес R² - 0,5246 и R² - 0,5082 соответственно.

Линейная регрессия и дерево решений показали более скромные результаты, что говорит о наличии нелинейных зависимостей в данных, которые эти модели не смогли уловить.

Наименее эффективной оказалась модель SVR с R² - 0,14, что говорит о том, что модель практически не смогла объяснить вариативность IC50. Возможно, это связано с отсутствием логарифмирования целевой переменной, т.к. SVR был практически единственным методом, у которого наблюдалось увеличение метрик после логарифмирования, однако этого было недостаточно.

Выводы и рекомендации для CC50 в целом аналогичны IC50.

* 1. **Регрессия для SI**

Задача регрессии для SI решалась схожим образом, но в результате экспериментов было отобрано 30 наиболее информативных признаков, часть из которых была отфильтрована корреляционным анализом.

Лучшие результаты на тестовой выборке показали модели LightGBM Regressor и XGBoost Regressor, наибольшего значения R² в районе 0,23.

Сопоставимые результаты показал случайный лес.

Остальные методы показали более скромные результаты, что говорит о наличии нелинейных зависимостей в данных, которые эти модели не смогли уловить.

**Вывод:**

В целом результаты по SI неудовлетворительные, для дальнейшего улучшения результатов необходимо провести более детальную работы с выбросами, а также отбором признаков. Возможно стоит обучить две модели для групп с низким SI и высоким SI.

* 1. **Классификация для IC50**

Для задачи классификации предварительно был выделен столбец с классификацией, показывающий превышает ли значение IC50 медианное значение в выборке.

Далее для отбора признаков использовался mutual\_info\_classif, эксперименты с которым позволили выделить 25 наиболее информативных признаков.

В целом дальнейший ход решения соответствует задачам регрессии, за исключением того, что вместо моделей классификации используются модели регрессии.

Лучшие результаты на тестовой выборке показали модели случайный лес и XGBoost Classifier c Accuracy равным 0,6979.

Остальные модели показали сопоставимые результаты.

**Выводы**:

1. Все модели показали примерно одинаковый результат, что, вероятно, свидетельствует о том, что при текущем распределении признаков был достигнут предел информативности и поэтому алгоритмы сходятся примерно одинакого.
2. В качестве дальнейшего развития планируется попробовать провести доп эксперименты с другими методами отбора признаков и попытаться провести FeatureEngeneering и выделить дополнительные признаки.
   1. **Классификация для СC50**

В целом ход решения данной задачи аналогичен предыдущей.

Лучшие результаты на тестовой выборке показал LightGBM Classifier c Accuracy равным 0,7447.

Остальные модели показали сопоставимые результаты за исключением дерева решений и логистической регрессии. Последние две модели, вероятно, являются слишком простыми для данной задачи и поэтому показывают худшие метрики

Выводы по данной задаче аналогичны классификации IC50.

* 1. **Классификация для SI**

В целом ход решения данной задачи аналогичен предыдущей.

Лучшие результаты на тестовой выборке показал SVC c Accuracy равным 0,708. Однако стоит отметить, что для данной задаче наблюдается сильный дисбаланс между precision и recall, что может говорить о сложности разделяющей поверхности или о проблемах с качеством и информативностью признаков.

Остальные модели показали сопоставимые худшие, но сопоставимые между собой результаты. Стоит отметить метод ближайших соседей, у которого наблюдается наименьший разброс между presicion и recall.

**Выводы**:

1. В целом модели показали примерно одинаковый результат, что, вероятно, свидетельствует о том, что при текущем распределении признаков был достигнут предел информативности и поэтому алгоритмы сходятся примерно одинакого.
2. В качестве дальнейшего развития планируется попробовать провести доп эксперименты с другими методами отбора признаков и попытаться провести FeatureEngeneering и выделить дополнительные признаки, кроме того, необходимо провести работу с выбросами.
   1. **Классификация для SI**

В целом ход решения данной задачи аналогичен предыдущей.

Но с учетом дисбаланса классов (примерно 1:2) при разделении на тренировочную и тестовую выборки было необходимо применить стратификацию по целевому признаку. Также при кроссвалидации в процессе подбора гиперпараметров было необходимо применить стратифицированное разделение на фолды.

Кроме того, для заполнения миноритарного класса был применен подход SMOTE, позволяющий заполнить миноритарный класс до размера мажоритарного с учетом характеристик текущего признакового пространства.

Лучшие результаты на тестовой выборке показал SVC c Accuracy равным 0,672. Однако стоит отметить, что не смотря на применение SMOTE для некоторых моделей сохраняется дисбаланс между precision и recall, может говорить о сложности разделяющей поверхности или о проблемах с качеством и информативностью признаков.

Остальные модели показали худшие, но сопоставимые между собой результаты.

**Выводы**:

1. В целом модели показали примерно одинаковый результат, что, вероятно, свидетельствует о том, что при текущем распределении признаков был достигнут предел информативности и поэтому алгоритмы сходятся примерно одинакого.
2. В качестве дальнейшего развития планируется попробовать провести доп эксперименты с другими методами отбора признаков и попытаться провести FeatureEngeneering и выделить дополнительные признаки, кроме того, необходимо провести работу с выбросами.

**Общие выводы по результатам проделанной работы:**

Несмотря на достигнутые результаты, остается значительный потенциал для повышения качества прогнозных моделей. На основе проведенного исследования были определены следующие ключевые направления для улучшения:

* + 1. Углубленная предобработка целевых признаков с выявлением и последующим удалением действительных выбросов.
    2. Проведение фича инжиниринга для выявления новых, информативных признаков.
    3. Рассмотрение возможности применения нейронных сетей для решения данной задачи
    4. Проведение дополнительных экспериментов с подбором оптимальных гиперпараметров моделей.