1. Введение и цели исследования

В данном исследовании анализируется набор данных из файла «002 Dataset.xlsx», содержащий информацию о химических свойствах молекул, разработанных для фармацевтических целей. Основное внимание уделяется трем ключевым показателям:

- IC50 (мM) – концентрация вещества, подавляющая активность биологического процесса на 50%. Чем выше значение, тем сильнее ингибирующий эффект.

- CC50 (мM) – концентрация, вызывающая гибель 50% клеток. Высокое значение указывает на низкую токсичность.

- SI (индекс селективности) – отношение CC50 к IC50. Чем выше SI, тем избирательнее действие вещества.

Цели исследования:

1. Построение моделей регрессии и бинарной классификации для IC50, CC50 и SI.

2. Сравнение моделей по метрикам качества.

3. Определение наиболее эффективных алгоритмов для прогнозирования указанных параметров.

2. Анализ и предварительная обработка данных

Исходный датасет содержит 1000 записей и214 признаков (числовые типы: `float64`, `int64`).

Основные этапы обработки:

1. Удаление неинформативных столбцов (например, `Unnamed: 0` – столбец индексов).

2. Замена пропущенных значений (NaN) на 0.

3. Поиск и обработка выбросов\*\* с использованием:

- Графика "ящик с усами" (Boxplot).

- Функции `remove\_all\_outliers()`, которая последовательно удаляет аномалии.

4. Анализ корреляций между признаками:

- Удалены слабокоррелирующие и мультиколлинеарные признаки (порог корреляции > 0.7).

- Нормализация данных.

Результаты предобработки:

- Для IC50 созданы два датасета:

- `dataset\_for\_IC50\_854` (после первичного удаления выбросов).

- `dataset\_for\_IC50\_567` (после полной очистки от аномалий).

- Для CC50лучшим оказался `dataset\_for\_CC50\_950` (без выбросов).

- Для SI подготовлены `dataset\_for\_SI\_876` и `dataset\_for\_SI\_588`.

3. Построение и оценка моделей

3.1. Модели регрессии

Использовались:

- Линейная регрессия (`LinearRegression`).

- Дерево решений (`DecisionTree`).

- Случайный лес (`RandomForest`).

- Градиентный бустинг (`CatBoost`).

- Нейронная сеть (`MLPRegressor`).

- Метод опорных векторов (`SVR`).

Результаты:

- Наилучшие показатели R² (коэффициент детерминации) у `CatBoost` и `RandomForest`.

- Применение PCA и UMAP для снижения размерности не всегда улучшало качество моделей.

3.2. Модели бинарной классификации

Цель: предсказать, превышает ли значение целевой переменной медиану.

Использовались:

- Логистическая регрессия (`LogisticRegression`).

- Дерево решений (`DecisionTree`).

- Случайный лес (`RandomForest`).

- `CatBoost`.

- Нейронная сеть (`MLPClassifier`).

- Метод опорных векторов (`SVC`).

Результаты:

- Наилучшие Accuracy и ROC-AUC у `RandomForest` и `CatBoost`.

- Для SI классификация оказалась сложнее из-за слабой корреляции с признаками.

4. Выводы

1. Лучшие модели:

- Для регрессии – `CatBoost` и `RandomForest`.

- Для классификации – `RandomForest`, `CatBoost` и `MLPClassifier`.

2. Влияние обработки данных:

- Удаление выбросов улучшает качество моделей, даже если сокращает объем данных.

- Снижение размерности (`PCA`, `UMAP`) не всегда оправдано.

3. Рекомендации:

- Использовать несколько моделей для сравнения.

- Оптимизировать гиперпараметры с помощью `GridSearch`.

- Для интерпретации результатов учитывать R² и ROC-AUC.

Итог:

Исследование показало, что даже при слабой зависимости целевых переменных от признаков, комбинация методов машинного обучения позволяет находить эффективные решения. Оптимальный подход – применение ансамблевых алгоритмов (`CatBoost`, `RandomForest`) с тщательной предобработкой данных.