**КУРСОВАЯ РАБОТА**

По дисциплине: «Классическое машинное обучение»

На тему: «Построение моделей машинного обучения для задач выявления наиболее подходящих параметров для разработки лекарственных средств»

|  |  |
| --- | --- |
|  | **Выполнил:**  Студент М24-525  (группа)  Блюм Игорь Андреевич  ыфвывфывфывыфвфы вфы(Ф.И.О.) |
|  |  |
|  |  |

Москва, 2025

СОДЕРЖАНИЕ

[**Введение** 3](#_Toc200544180)

[**1.** **Описание данных** 4](#_Toc200544181)

[**1.1.** **Показатели IC50, CC50 и SI** 4](#_Toc200544182)

[**2.** **Описание задач** 5](#_Toc200544183)

[**3.** **Выполнение работы** 6](#_Toc200544184)

[**3.1.** **Анализ представленных данных (EDA)** 6](#_Toc200544185)

[**3.2.** **Анализ распределения целевых переменных** 8](#_Toc200544186)

[**3.3.** **Отбор признаков (Feature Selection)** 11](#_Toc200544187)

[**3.3.1.** **Регрессия** 11](#_Toc200544188)

[**3.3.2** **Классификация** 13](#_Toc200544189)

[**3.4** **Выбор моделей машинного обучения. Выбор метрики, для отбора лучшей модели согласно задаче.** 14](#_Toc200544190)

[**3.4.1.** **Для задач регрессии** 14](#_Toc200544191)

[**3.4.2.** **Для задач классификации** 15](#_Toc200544192)

[**3.5.** **Подбор гипперпараметров** 16](#_Toc200544193)

[**4.** **Создание класса для отбора лучшей модели.** 17](#_Toc200544194)

[**Заключение** 18](#_Toc200544195)

**ВВЕДЕНИЕ**

Процесс создания нового лекарственного препарата является сложным и многоэтапным. Необходимо определить его химическую формулу, синтезировать соединение, провести первичные биологические испытания и организовать тестирование. Все эти этапы требуют значительного времени, однако современные методы машинного обучения способны существенно ускорить данный процесс.

Так, например, с помощью различных моделей можно спрогнозировать эффективность соединений и подобрать наиболее подходящие сочетания параметров для разработки лекарственных средств.

Тем не менее, для достижения качественного результата в подобного рода задачах важно наладить эффективное взаимодействие между химиками и специалистами по машинному обучению, что зачастую оказывается непростой задачей.

Представим следующую ситуацию: химиками были предоставлены конфиденциальные данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность, обозначаются как IC50, CC50 и SI.

Значение SI рассчитывается на основе параметров IC50 и CC50. Все остальные представленные признаки являются числовыми характеристиками химических соединений.

1. **Описание данных**

Данные предоставленные химиками имеют размерность 1001 на 214, из которых 213 столбцов являются признаками. В 213 признаках имеются целевые переменные IC50, CC50 и SI, которые необходимо будет преобразовать перед тем, как приступать к обучению моделей машинного обучения.

* 1. **Показатели IC50, CC50 и SI**

IC50 (Half Maximal Inhibitory Concentration) – это показатель концентрации соединения, которая необходима для подавления активности вируса на 50% в пробирке. Чем меньше значение IC50, тем выше эффективность. Единицы измерения: μM или nM (микромоляр или наномоляр)

CC50 (Half Maximal Cytotoxic Concentration) - это показатель концентрации, при которой гибнет 50% клеток из-за токсичности соединения. Чем выше CC50, тем безопаснее соединение. Единица измерения, как и у IC50.

SI (Selectivity Index) – это индекс характеризующий под собой отношение токсичности образца к его эффективности в борьбе с вирусом. Данный индекс считается по формулу:

Интерпретация:

- SI > 10: Перспективный кандидат.

- SI < 1: Опасная токсичность.

Пример:

* IC50 = 2 μM, CC50 = 100 Μm

SI = 50 – образец является перспективным, так как имеет высокое отношение между показателем токсичности и эффективности

* IC50 = 10 μM, CC50 = 5 μM

SI = 0.5 – образец не является перспективным, хоть и имеет хороший показатель эффективности, но при этом является слишком токсичным, для того чтобы такое соединение использовать в дальнейшем для производства лекарственных средств.

1. **Описание задач**

В рамках курсового проекта, нам необходимо создать 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

Также в процессе работы над отдельно взятой задачей необходимо будет сравнить между собой полученные модели и их результаты, выполнить анализ и обосновать выбор наиболее качественных решений.

1. **Выполнение работы**
   1. **Анализ представленных данных (EDA)**

Перед тем как приступать к процессу обучения моделей машинного обучения необходимо провести анализ представленных данных, избавиться от пропусков в данных и удалить столбцы, которые не представляют никакой ценности, к примеру, колонка с индексами строк.

Как было сказано ранее, размерность источника с данными имеет размерность 1001 на 214. Из 214 признаков 107 имеют тип данных float64, остальные 107 признаков имеют тип данных int64.

Колонка “Unnamed: 0”, которая является колонкой индексов строк и подлежит удалению, так как не несет никакой информационной ценности для наших задач.

В данных также присутствуют пропуски в следующих признаках, которые необходимо обработать:

Таблица 1. Количество пропусков в признаках

|  |  |
| --- | --- |
| Название признака | Количество пропусков |
| MaxPartialCharge | 3 |
| MinPartialCharge | 3 |
| MaxAbsPartialCharge | 3 |
| MinAbsPartialCharge | 3 |
| BCUT2D\_MWHI | 3 |

Продолжение таблицы 1.

|  |  |
| --- | --- |
| BCUT2D\_MWLOW | 3 |
| BCUT2D\_CHGHI | 3 |
| BCUT2D\_CHGLO | 3 |
| BCUT2D\_LOGPHI | 3 |
| BCUT2D\_LOGPLOW | 3 |
| BCUT2D\_MRHI | 3 |
| BCUT2D\_MRLOW | 3 |

Есть несколько способов работы с пустыми значениями в данных.

1. **Удаление строк с пропусками** - простое удаление всех строк (образцов), где есть хотя бы один пропуск. Подходит, если пропусков мало и их удаление не сильно уменьшит выборку.
2. **Удаление признаков с пропусками** - удаление столбцов (признаков), где доля пропусков слишком велика (например, >50%). Используется, если признак не важен для модели.
3. **Замена на константу** – заполнение пропусков фиксированным значением (например, -1, 0, NaN или строкой `"Unknown"`).Подходит для категориальных и числовых признаков.
4. **Замена на среднее/медиану (для числовых признаков)** – пропуски заполняются средним или медианным значением признака. Уменьшает смещение, но может исказить распределение
5. **Замена на моду (для категориальных признаков)** – пропуски заполняются самым частым значением в признаке. Подходит для категориальных данных.
6. **Предсказание пропущенных значений (Imputation)** – использование моделей (KNN, линейная регрессия, Random Forest) для предсказания пропусков на основе других признаков. Более точный, но сложный метод.
7. **Добавление бинарного признака "Был ли пропуск"** – создание нового бинарного признака (`1` — было пропущено, 0 — нет). Полезно, если пропуски несут информацию (например, человек не указал зарплату).
8. **Интерполяция (для временных рядов)** – заполнение пропусков на основе соседних значений (линейная, полиномиальная интерполяция). Используется в данных с временной зависимостью.
9. **Использование алгоритмов, устойчивых к пропускам** – некоторые модели (например, XGBoost, LightGBM) могут обрабатывать пропуски без предварительной обработки.
10. **Множественное заполнение (Multiple Imputation)** – создание нескольких версий датасета с разными заполненными значениями и объединение результатов.Учитывает неопределённость в пропущенных данных.

Для наших признаков выберем способ номер 4 и заменим пропущенные значения медианной соответствующего признака. Была выбрана именно медиана, так как она мене подвержена выбросам.

* 1. **Анализ распределения целевых переменных**

Если посмотреть на числовое распределение данных в целевых переменных, можно будет заметить, что данные имеют большой разброс между min и max значениями, а также при построении графика “ящик с усами”, можно убедиться в том, что в данных присутствуют выбросы.

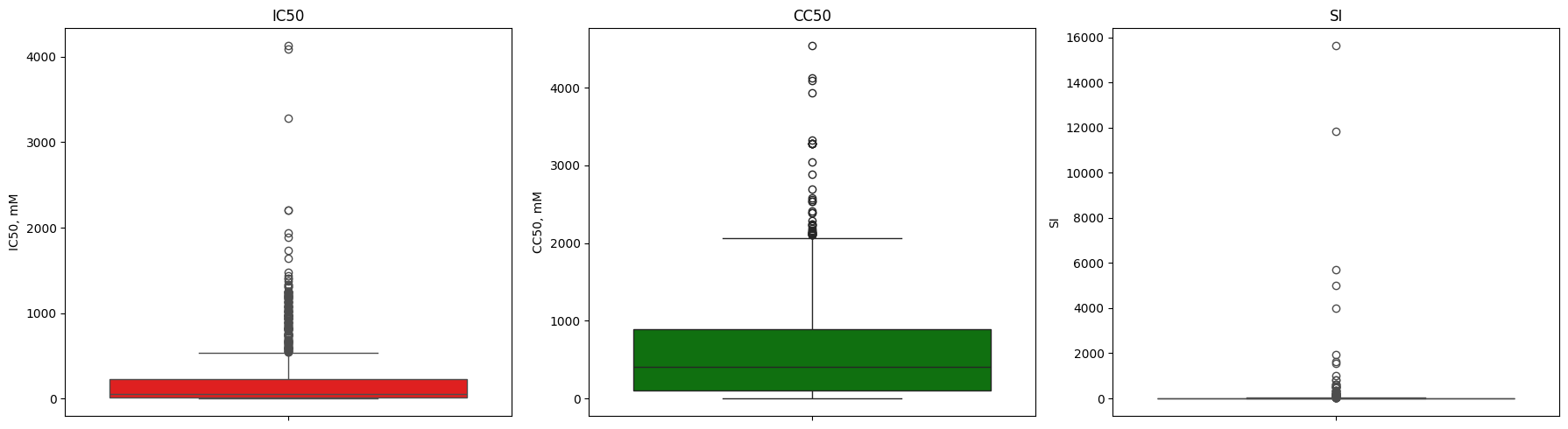


Рисунок 1. График “ящик с усами” распределения целевых переменных (слева - направо: график для IC50, график для CC50, график для SI)

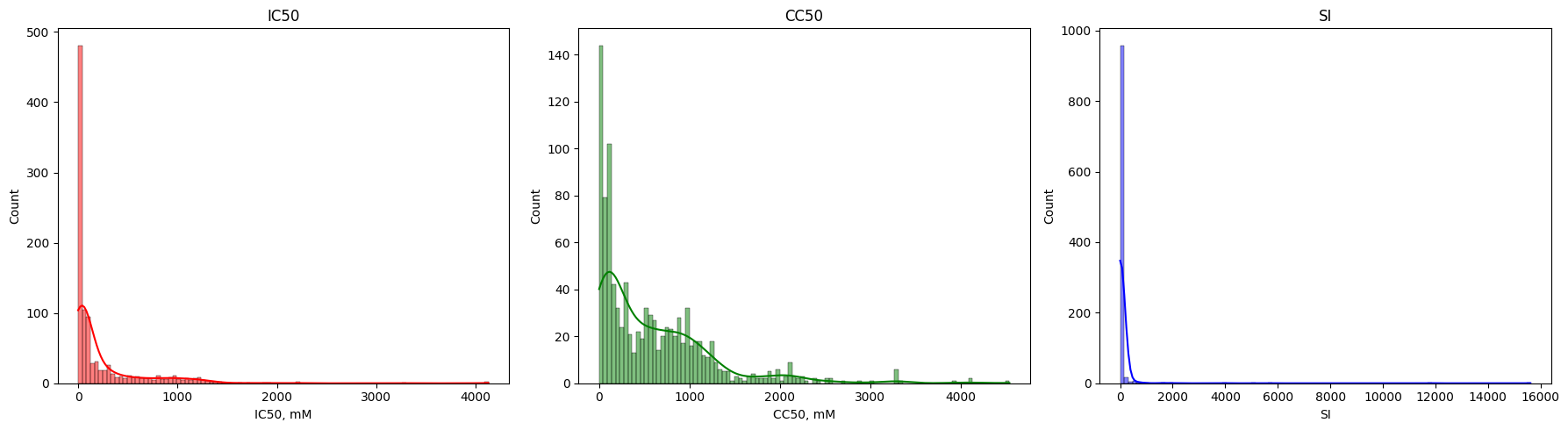


Рисунок 2. Гистограмма распределения целевых переменных (слева - направо: график для IC50, график для CC50, график для SI)

На рисунке 2 можно заметить, что целевые переменные имеют правостороннюю асимметрию - большинство значений сконцентрировано слева, выбросы справа

Для задач регрессии такое распределение будет отрицательно сказываться на качестве моделей, так как в задачах регрессии оптимизируются функции потерь, которые чувствительны к выбросам, такие как:

1. Среднеквадратичная ошибка (MSE)

2. Средняя абсолютная ошибка (MAE)

Для того чтобы нивелировать такой характер распределения данных, применим **логорифмирование** для целевых переменных, что поможет нам:

1. Сжать шкалу (Для SI 15620 → 4.2)

2. Сделает распределение более нормальным, при этом сохраниться биологический смысл

3. Исключит потребность равного распределения выбросов между train, test выборками

Если оставить данные как есть, то это приведет к следующим последствиям:

1. Модели будут оптимизироваться под выбросы

2. Метрики качества (MSE) потеряют смысл

3. Градиентные методы могут расходиться

Для задач классификации, из целевых переменных нам необходимо создать бинарные признаки, которые отвечают условию задач.

Для задачи классификации, где необходимо выяснить превышает ли значение IC50 медианное значение выборки, мы преобразуем признак IC50 в бинарный признак, который отвечает условию “если значение IC50 > медианы от IC50, то записываем 1, иначе 0

Для задачи классификации, где необходимо выяснить превышает ли значение CC50 медианное значение выборки, мы преобразуем признак CC50 в бинарный признак, который отвечает условию “если значение CC50 > медианы от CC50, то записываем 1, иначе 0

Для задачи классификации, где необходимо выяснить превышает ли значение SI медианное значение выборки, мы преобразуем признак SI в бинарный признак, который отвечает условию “если значение SI > медианы от SI, то записываем 1, иначе 0

Для задачи классификации, где необходимо выяснить превышает ли значение SI значение 8, мы преобразуем признак SI в бинарный признак, который отвечает условию “если значение SI > 8, то записываем 1, иначе 0”

* 1. **Отбор признаков (Feature Selection)**
     1. **Регрессия**

Для каждой задачи регрессии нам необходимо отобрать те признаки, которые будут наиболее информативны для задачи предсказания числового признака IC50, CC50, SI. Отбор признаков (feature selection) помогает выбрать наиболее информативные признаки для модели, улучшая её интерпретируемость, скорость обучения и качество.

Так как мы не глубоко не погружены в специфику данных и у нас нет эксперта, который мог бы сказать более детально, какой признак более релевантный для нашей задачи, воспользуемся методов перекрестного отбора признаков.

Данный метод заключается в выборе нескольких методов отбора признаков относительно целевой переменной с последующим сравнением выбранных признаков между собой. В итоговую выборку попадают те признаки, которые при сравнении между собой имеют пересечение, то есть отвечают математическому правилу пересечения множеств: .

Для подбора признаков для задачи регрессии используем методы: mutual\_info\_regression, Lasso, XGBRegressor

**Mutual Info Regression** - метод оценивает взаимную информацию (mutual information) между каждым признаком и целевой переменной. Чем выше значение, тем сильнее зависимость. Суть метода заключается в измерении, насколько уменьшается неопределенность целевой переменной при знании признака.

**Lasso (L1-регуляризация)** - линейная модель регрессии с L1-регуляризацией, которая "обнуляет" неважные коэффициенты признаков. Некоторые коэффициенты становятся нулевыми, что исключает соответствующие признаки из модели.

**XGBRegressor (Feature Importance) –** является алгоритмом машинного обучения, который имеет встроенный метод оценки важности признаков в градиентном бустинге (XGBoost).

Признаки оцениваются на основе:

Weight (частота использования) – сколько раз признак использовался для разбиения.

Gain (улучшение точности) – среднее улучшение метрики при использовании признака.

Cover (охват данных) – сколько образцов затронуто разбиением.

На основе данных методов отбора, строиться диаграмма пересечения признаков и формируется окончательный список признаков, которые в дальнейшем будут использоваться для построения моделей машинного обучения.

Пересечения признаков для задачи регрессии CC50:

**Ядро пересечений:** set()

**Пересечения Mutual Info Regression + Lasso:** {'EState\_VSA9', 'VSA\_EState5', 'FractionCSP3'}

**Пересечения Mutual Info Regression + XGBRegressor:** {'BalabanJ', 'HallKierAlpha', 'Chi0n', 'qed', 'AvgIpc', 'SPS', 'HeavyAtomMolWt', 'MolMR', 'LabuteASA', 'BCUT2D\_CHGHI', 'Kappa1', 'VSA\_EState2', 'Chi2n', 'Chi1v', 'BertzCT', 'Chi1', 'SlogP\_VSA6', 'BCUT2D\_MRHI', 'MolWt', 'Chi1n', 'BCUT2D\_MWLOW', 'BCUT2D\_MRLOW'}

**Пересечения Lasso + XGBRegressor:** {'EState\_VSA3', 'PEOE\_VSA9', 'fr\_ketone\_Topliss'}

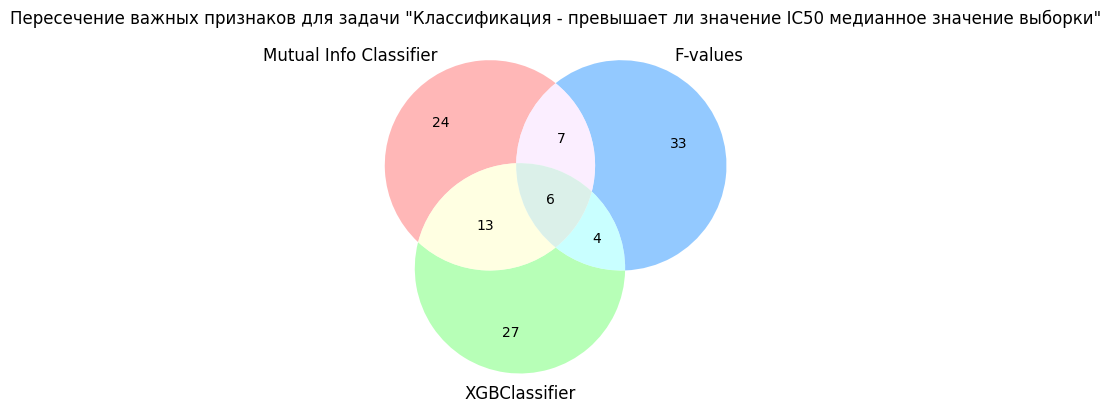


Рисунок 3. Диаграмма пересечение важных признаков для задачи регрессии для CC50.

* + 1. **Классификация**

Для подбора признаков для задачи регрессии используем методы: Mutual Info Classifier, F-values, XGBClassifier.

**Mutual Info Classifier** – является аналогом Mutual Info Regression, но для классификации — оценивает зависимость между признаком и дискретной целевой переменной. Оценивает, насколько признак уменьшает неопределённость класса.

**F-values** (ANOVA) - оценивает значимость признаков с помощью F-статистики (дисперсионный анализ). Сравнивает межклассовую дисперсию с внутриклассовой. Высокое F-значение означает, что признак сильно влияет на целевую переменную.

**XGBClassifier** (Feature Importance) – является аналогом метода XGBRegressor, но для задач классификации.

На основе данных методов отбора, строиться диаграмма пересечения признаков и формируется окончательный список признаков, которые в дальнейшем будут использоваться для построения моделей машинного обучения.

Пересечения признаков для задачи классификации CC50:

**Ядро пересечений:** {'Chi0n', 'Chi0', 'Kappa3', 'PEOE\_VSA6', 'SMR\_VSA10'}

**Пересечения Mutual Info Classifier + F-values:** {'SlogP\_VSA3', 'Kappa2', 'MolMR', 'EState\_VSA8', 'LabuteASA', 'ExactMolWt', 'IC50, mM', 'MolWt'}

**Пересечения Mutual Info Classifier + XGBClassifier:** {'BalabanJ', 'VSA\_EState7', 'HallKierAlpha', 'SMR\_VSA5', 'AvgIpc', 'SI', 'SPS', 'BCUT2D\_LOGPHI', 'SlogP\_VSA4', 'SMR\_VSA7', 'BCUT2D\_CHGHI', 'BCUT2D\_LOGPLOW', 'FpDensityMorgan3', 'PEOE\_VSA9', 'SlogP\_VSA6', 'PEOE\_VSA8', 'TPSA', 'MaxAbsEStateIndex', 'BCUT2D\_MWLOW'}

**Пересечения F-values + XGBClassifier:** {'fr\_allylic\_oxid', 'fr\_NH1', 'HeavyAtomMolWt', 'MolLogP'}

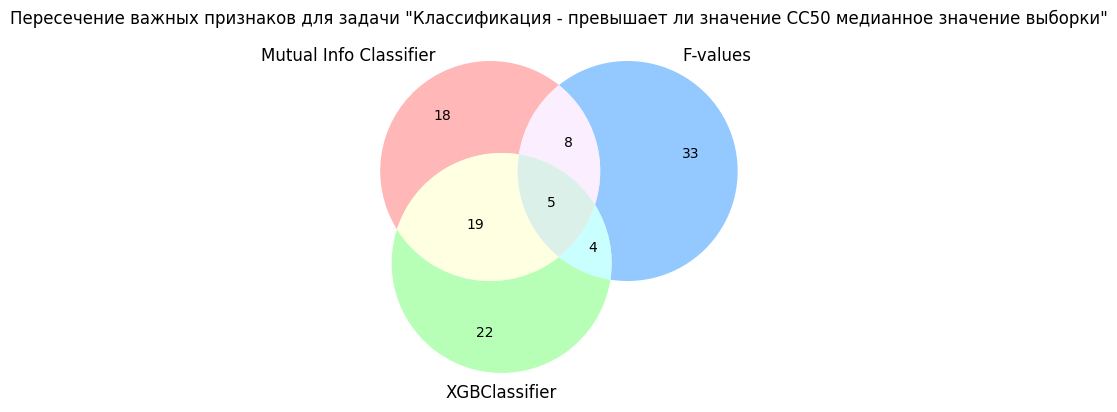


Рисунок 4. Диаграмма пересечение важных признаков для задачи классификации для CC50.

* 1. **Выбор моделей машинного обучения. Выбор метрики, для отбора лучшей модели согласно задаче.**
     1. **Для задач регрессии**

После того как мы преобразовали целевую переменную и отобрали признаки, можно приступать к обучению моделей. По условию задачи, необходимо обучить несколько и сравнить между собой их результаты. На основе полученных результатов мы выберем ту, что показала себя лучше всех, относительно ключевой метрики.

Для задач регрессии были выбраны следующие модели для обучения и дальнейшей оценки их показателей:

**LinearRegression** - Линейная модель, предсказывающая целевую переменную как линейную комбинацию признаков. Оптимизирует квадратичную ошибку (MSE) с помощью метода наименьших квадратов.

**RandomForestRegressor** - ансамбль решающих деревьев, предсказывающий среднее значение их прогнозов. Строит множество деревьев на бутстрап-выборках (bagging). Каждое дерево обучается на случайном подмножестве признаков.

**ElasticNet** - Линейная модель с комбинированной L1- и L2-регуляризацией (Lasso + Ridge). Минимизирует MSE с штрафом.

**SVR** - Метод опорных векторов для регрессии, максимизирующий "допуск" ошибки. Игнорирует ошибки меньше ε, штрафуя за отклонения вне трубки. Может использовать ядра (RBF, полиномиальное) для нелинейности.

В качестве главной метрики оценки эффективности моделей будет использована **R^2 score** - коэффициент детерминации. Коэффициент детерминации – это статистическая мера, которая показывает, насколько хорошо модель линейной регрессии соответствует реальным данным. Он указывает, какую долю дисперсии зависимой переменной объясняет независимая переменная (или переменные) в модели. Чем ближе коэффициент детерминации к 1, тем лучше модель описывает наблюдаемые данные. Для порога отбора моделей, зададим начальный параметр R^2 score равным 0.5.

* + 1. **Для задач классификации**

Для задач классификации были выбраны следующие модели для обучения и дальнейшей оценки их показателей:

**LogisticRegression** - линейная модель для бинарной/многоклассовой классификации через логит-функцию. Оценивает вероятность класса, оптимизирует log-loss (кросс-энтропию).

**KNeighborsClassifier** - метод k ближайших соседей: классифицирует объект по majority vote среди k соседей. Использует расстояния (евклидово, манхэттенское и др.).

**RandomForestClassifier** - ансамбль решающих деревьев, голосующих за класс (аналог RandomForestRegressor). Bagging + случайный выбор признаков для каждого дерева.

**SVC (Support Vector Classification)** - метод опорных векторов для классификации, ищущий максимально широкий разделяющий зазор. Максимизирует margin между классами. Поддерживает ядра (RBF, полиномиальное) для нелинейных границ.

В качестве главной метрики оценки эффективности моделей для задач классификации будет использована **F-мера**. F-мера является гармоническим средним между точностью и полнотой. F1-меру используют, когда важно учитывать как ложные положительные, так и ложные отрицательные срабатывания.

* 1. **Подбор гипперпараметров**

После выбора моделей и метрики, нам необходимо предварительно подобрать лучшие гипперпараметры, для каждой из моделей, чтобы они могли показать лучшие результаты и могли объективно сравнить их эффективность относительно ключевой метрики.

В качестве методов подбора гипперпараметров были использованы методы GridSearchCV и RandomizedSearchCV.

**GridSearchCV** является методом последовательного перебора всех параметров, которые были переданы данному методу с соответствующей моделью машинного обучения.

**RandomizedSearchCV** является методом случайного отбора параметров, где каждая настройка выбирается из распределения возможных значений параметров. Это имеет два основных преимущества перед последовательным поиском: Количество вычислений может быть выбрано независимо от количества параметров и возможных значений.

1. **Создание класса для отбора лучшей модели.**

В ходе работы над курсовым проектом для 2 типов задач, регрессия и классификация, были созданы специальные классы и методы для автоматического отбора лучшей модели машинного обучения согласно ключевой метрике.

Структура классов одинаковая и имеет 4 этапа:

1. Инициализация и передачи в класс DataFrame с признаками (X) и DataFrame с целевой переменной (Y)
2. Подбор лучших гипперпараметров используя метод fit()
3. Отбор лучшей модели с помощью метода get\_best\_model()
4. Формирование предсказания лучшей моделью с помощью метода predict(X), где X - данные с признаками, для предсказания целевой переменной Y

С кодом класса BestRegressionModel и BestClassifierModel можно ознакомиться в файлах ipynb, в которых производилось обучение моделей.

**ЗАКЛЮЧЕНИЕ**

В ходе работы над курсовым проектом были подобраны и обучены различные модели машинного обучения для достижения наилучшего результата для каждой отдельной взятой задачи, которые были поставлены в рамках данного проекта. Были применены различные методы отбора признаков, которые позволили в рамках ограниченной осведомленности и погруженности в отдельно взятый химический параметр, сократить их количество и отобрать топ лучших, используя метод перекрестного отбора.

По итогу работы над курсовым проектом были написаны 2 класса для каждого типа задач, который сравнивает между собой результаты полученных моделей с ключевой метрикой и основе этого делает выбор наиболее качественной из них, которая в дальнейшем может быть использована для формирования предсказаний.

В качестве рекомендации для улучшения показателей моделей можно рассмотреть увеличение количество топ параметров для каждого метода отбора, в работе в каждом методе использовались топ 50 параметров, а также консультация с экспертами в фармацевтической области для утверждения выбранных методом перекрестного отбора.