НИЯУ МИФИ

ПРЕДСКАЗАНИЕ ПРОТИВОВИРУСНОЙ АКТИВНОСТИ СОЕДИНЕНИЙ

ИФТЭБ НИЯУ МИФИ
ИНСТИТУТ ФИНАНСОВЫХ ТЕХНОЛОГИЙ И ЭКОНОМИЧЕСКОЙ БЕЗОПАСНОСТИ

Команда: IFTEBeer
Студенты 3-го курса ИФТЭБ НИЯУ МИФИ:
Логинова Е. Г.
Дорофеев Н. А.
Тарасов С. В.
Харченко Т. В.

02.10.2024 г.

Цели и задачи

- **Цель:**
Построить модель машинного обучения на основе алгоритма “Extreme Gradient Boosting” (XGBoost), которая позволит предсказывать индекс селективности SI против вируса SARS-CoV-2, основываясь на введенной SMILES формуле вещества.

- **Задачи:**

1) Собрать данные о ранее изученных противовирусных соединениях, активных против вируса SARS-CoV-2.
2) Провести обработку данных, отобрав только те вещества, которые подходят для дальнейшего анализа.
3) На основе алгоритма градиентного бустинга построить модель машинного обучения, предсказывающую активность введенного соединения в зависимости от его молекулярной структуры.
4) Использовать эту модель для предсказания индекса селективности (SI).
5) Проверить эффективность модели на тестовых данных.

Актуальность

Поиск лекарственных препаратов с высокой селективностью — то есть таких, которые воздействуют исключительно на целевой фермент или белок, не затрагивая другие — представляет собой важнейшую задачу в фармацевтической промышленности. Высокая селективность обеспечивает эффективное лечение с минимальными побочными эффектами.

В целом, использование методов машинного обучения (ML) для выявления соединений с высоким индексом селективности (SI) является актуальной задачей, так как:

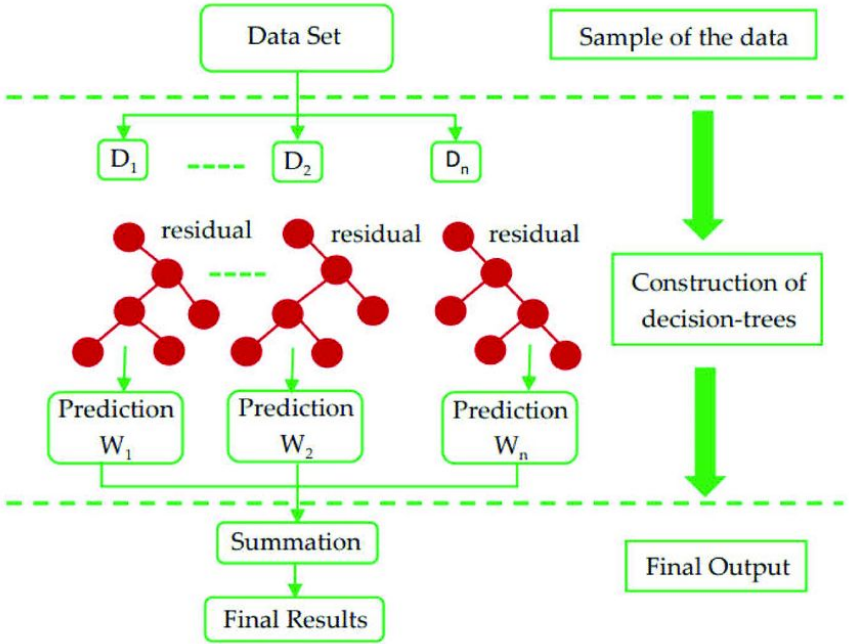
- Увеличивает эффективность и скорость поиска новых лекарственных препаратов.
- Снижает затраты на исследования и разработки.
- Открывает новые возможности для создания более безопасных и эффективных медицинских средств.

Применение машинного обучения в этой области является перспективным направлением, способным существенно повлиять на развитие фармацевтической промышленности и улучшение лечения различных заболеваний.

Материалы и методы

Машинное обучение (machine learning, ML) – совокупность методов искусственного интеллекта, позволяющих строить алгоритмы (модели), которые способны обучаться на каких-либо данных.

В поставленной задаче использовался метод **градиентного бустинга** (Extreme Gradient Boosting, XGBoost)



Материалы и методы

Обработка данных производилась на языке **Python** с помощью библиотеки **Pandas**.

Для обработки формул типа **SMILES** была использована библиотека **RDKit.**

Построение модели машинного обучение осуществлялось с помощью **XGBoostRegressor** из программной библиотеки **XGBoost**

Также для выполнения различных задач, таких как разбивка данных на обучающую и тестовую выборки и настройка гиперпараметров с использованием **RandomizedSearchCV**, была применена библиотека **scikit-learn**, которая предоставляет полезные инструменты для работы с машинным обучением.

Материалы и методы

Для отбора соединений использовались такие базы данных, как Acta Pharmacologica Sinica, ScienceDirect, BindingDB

Для дальнейшей работы были отобраны лиганды со следующими свойствами:

- Тип измеряемой активности: SI
- Анализируемый организм: Homo sapiens
- Целевой организм: Homo sapiens

Материалы и методы

Для сбора данных о неингибирующих соединениях был использован следующий подход:

искались вещества с высоким значением **IC50** относительно **SARS-CoV-2**. Поскольку **IC50** находится в знаменателе, это приводит к тому, что дробь **SI** стремится к нулю, а вместе с этим и ингибирующая способность соединения.

$$SI = \frac{CC_{50}}{IC_{50}}$$

Материалы и методы

Фингерпринты – представление молекул в виде битовой строки, где каждый бит соответствует наличию (1) либо отсутствию (0) в молекуле какой-то определенной структуры. В данной работе каждый лиганд был закодирован в 2048-битную строку.

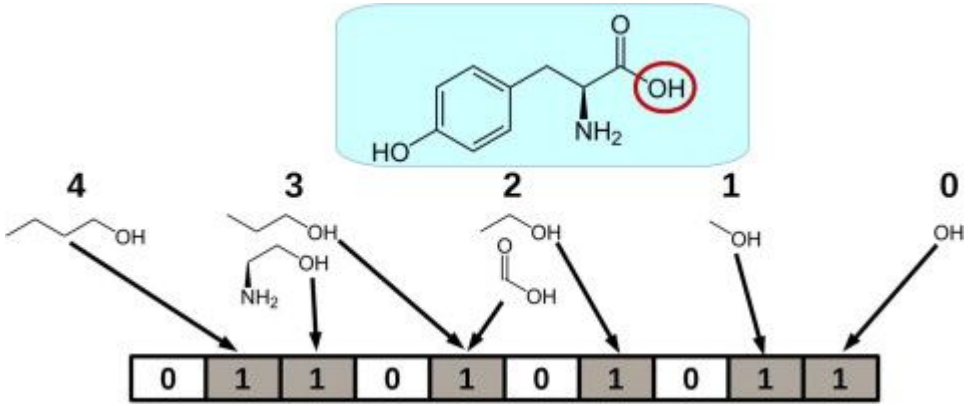


Рис. 1 – Схематичное представление принципа генерации фингерпринтов

Результаты и их обсуждение

Данные были проверены по критерию **MAE** (средняя абсолютная ошибка)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

где N — число примеров обучающей выборки, y_i — целевое значение i-го примера, $\hat{y}_i$ — предсказанное моделью значение.

Было получено значение 2.231

Выводы

Основные результаты:

1) В процессе работы была разработана модель, предсказывающая значение индекса селективности (SI) на основе форматов SMILES. С кодом и моделью можно ознакомиться на платформе GitHub: (https://github.com/Roxasmeei/SARS-CoV-2-_SI_predictive_system)
2) Модель имеет достаточно низкий **MAE** (показатель среднего отклонения) - 2.231
3) Потенциальное применение:

Полученные результаты могут способствовать экономии времени и финансовых ресурсов в лаборатории хемоинформатики НИЯУ МИФИ.

Спасибо за внимание!

НИЯУ МИФИ

Будем рады ответить на все вопросы!

Команда IFTEBeer