**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**Кафедра биомедицинской информатики**

**ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ БИОТОКСИЧНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ ДЛЯ ИДЕНТИФИКАЦИИ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ BCR-ABL ТИРОЗИНКИНАЗЫ**

Курсовая работа

Мищенко Марии Михайловны

студентки 3 курса

специальности «Информатика»

Научный руководитель:

А.Д.Карпенко

Минск, 2024

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 2](#_Toc166243767)

[ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ, ТЕРМИНОВ 4](#_Toc166243768)

[ГЛАВА 1. СОВРЕМЕННЫЕ ПОДХОДЫ IN SILICO ПРИ СОЗДАНИИ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ 5](#_Toc166243769)

[ГЛАВА 2. МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ БИОТОКСИЧНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ 8](#_Toc166243770)

[2.1. ОБЩИЕ ПОДХОДЫ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ БИОТОКСИЧНОСТИ 8](#_Toc166243771)

[2.2. ТИПЫ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ 10](#_Toc166243772)

[2.3. МОДЕЛИ, ОСНОВАННЫЕ НА СХОДСТВЕ 11](#_Toc166243773)

[2.4. ЛИНЕЙНЫЕ МОДЕЛИ 12](#_Toc166243774)

[2.5. МОДЕЛИ НА ОСНОВЕ ЯДРА 13](#_Toc166243775)

[2.6. БАЙЕСОВСКИЕ МОДЕЛИ 14](#_Toc166243776)

[2.7. ДРЕВОВИДНЫЕ МОДЕЛИ 15](#_Toc166243777)

[ГЛАВА 3. ПРОБЛЕМЫ ПРЕДСКАЗАНИЯ БИОТОКСИЧНОСТИ IN SILICO 17](#_Toc166243778)

[ГЛАВА 4. АНАЛИЗ БИОТОКСИЧНОСТИ ИНГИБИТОРОВ BCR-ABL ТИРОЗИНКИНАЗЫ 20](#_Toc166243779)

[ЗАКЛЮЧЕНИЕ 21](#_Toc166243780)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 22](#_Toc166243781)

# ВВЕДЕНИЕ

Токсичность и побочные эффекты являются основными причинами неэффективности потенциальных фармацевтических препаратов. По причинам токсичности количество новых лекарств, поступающих на рынок в рамках традиционного процесса разработки лекарств, с годами сократилось. В настоящее время одним из актуальных аспектов создания новых лекарственных препаратов и материалов является оценка биотоксичности низкомолекулярных соединений.

Это позволяет исключить потенциально опасные вещества на ранних этапах разработки и сосредоточить усилия на наиболее перспективных соединениях.

Традиционные методы оценки риска, а именно эксперименты *in vivo* и клинические испытания, хоть и являются наиболее точным методом выявления токсических эффектов, но выполняются только после разработки продукта, являются дорогостоящими и достаточно затратными по времени, что делает их неэффективными.

В современной химической науке значительное внимание уделяется разработке методов компьютерного прогнозирования и оценки свойств молекулярных соединений, в том числе биотоксичности в фазе разработки (Research&Development-фаза).

В настоящее время с появлением больших данных и искусственного интеллекта подходы *in silico* к прогнозной токсикологии сосредоточены на использовании методов машинного обучения. Данные методы стали потенциальным решением для прогнозирования токсичности и обеспечения эффективной разработки лекарств и химической безопасности.

Модели машинного обучения для прогнозирования токсичности могут сократить затраты и время экспериментов, одновременно решая этические проблемы за счет резкого снижения потребности в животных и клинических испытаниях, что делает их наиболее привлекательными для использования в современных исследованиях.

**Цель исследования** – ознакомиться с возможными методами машинного обучения и применить один из методов для оценки биотоксичности низкомолекулярных соединений для идентификации ингибиторов BCR-ABL тирозинкиназы.

**Задачи исследования**:

1. проанализировать современные подходы *in silico* в области создания лекарственных препаратов;
2. ознакомиться с различными методами машинного обучения для предсказания биотоксичности низкомолекулярных соединений;
3. оценить биотоксичость ингибиторов BCR-ABL тирозинкиназы с помощью модели Случайный лес (Random Forest, RF);
4. изучить проблемы подходов *in silico* для предсказания биотоксичности низкомолекулярных соединений.

# ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ, ТЕРМИНОВ

|  |  |
| --- | --- |
| **ОБОЗНАЧЕНИЕ/ ТЕРМИН** | **ОПРЕДЕЛЕНИЕ** |
| ***in silico*** | компьютерное моделирование (симуляцию) эксперимента |
| ***in vivo*** | проведение экспериментов на (или внутри) живой ткани при живом организме |
| ***in vitro*** | проведение эксперимента «в пробирке» — в искусственных условиях, вне организма или естественной среды |
| **Лиганд** | химическое вещество, связывающееся с белком, то есть соединением, специфическим образом, взаимодействующим с активным сайтом рецептора |
| **Докинг** | процесс расчета взаимодействия между белком и лигандом |
| **Лекарство** | химическое соединение (как правило, низкомолекулярное), специфически взаимодействующее с мишенью и тем или иным образом модифицирующее клеточный ответ, создаваемый мишенью |
| **Свойства ADMET** | абсорбация, распределение, метаболизм, выведение, токсичность |
| **Дескрипторы** | числовые представления химических соединений, которые отражают различные физико-химические, топологические и электронные свойства |
| **QSAR** | Quantitative Structure – Activity Relationship, количественное отношение «структура-активность» |
| **МО** | Машинное обучением |
| **RF** | Random Forest, Случайный лес |
| **ET** | Extremely Randomize Trees, Чрезвычайно рандомизированные деревья |

# ГЛАВА 1. СОВРЕМЕННЫЕ ПОДХОДЫ IN SILICO ПРИ СОЗДАНИИ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

Процесс создания новых лекарственных средств включает три основные фазы – фаза разработки (Research&Development-фаза), доклинические исследования, клинические испытания.

При этом клинические испытания представляют собой весьма длительную стадию, и в силу ряда объективных причин ускорить ее сложно. Поэтому основной вклад в оптимизацию процесса разработки лекарственных средств можно внести именно в фазе доклинических исследований.

В настоящее время благодаря прорывам в молекулярной биологии, геномике и компьютерным технологиям становится возможным рациональное конструирование эффективных лекарственных средств. Компьютерные технологии играют ведущую роль в конструировании лекарств, так как с их помощью можно ускорить процесс исследования, минимизировать ошибки и повысить результативность.

Однако современный уровень развития компьютерных методов не позволяет разработать новый лекарственный препарат, используя только компьютерные технологии.

Наличие у вещества рецепторной активности далеко не во всех случаях приводит к его дальнейшему использованию в качестве лекарственного средства. Многие из веществ-кандидатов, успешно прошедших доклинические исследования, отсеиваются на этапе дорогостоящих клинических испытаний.

Данные клинических испытаний с 2010 по 2017 год показывают, что токсичность является причиной 30% клинических неудач при разработке лекарств.  Традиционный процесс выявления токсичности включает методы скрининга in vivo и in vitro. Хотя эти методы доказали свою эффективность, они, как правило, отнимают много времени, неэффективны и очень дороги. Использование испытаний на животных в клинических испытаниях было под вопросом из-за этических соображений и точности прогнозирования токсичности для человека.

Это привело к появлению in silico подходов к прогнозированию токсичности, которые могут использовать присущие молекулам свойства. Данные подходы имеют преимущество в прогнозировании токсичности еще до того, как соединение будет синтезировано, что экономит время и деньги по сравнению с традиционными методами прогнозирования токсичности.

Подход in silico основан на концепции «принципа молекулярного подобия», который предполагает, что молекулы со схожей структурой будут иметь схожую биологическую активность.  Эти методы используют структурные и молекулярные свойства этих молекул для получения наиболее полезной информации для прогнозирования токсичности.

На пером этапе оценки безопасности лекарства нужно знать только то, является ли соединение токсичным или нетоксичным, высокотоксичным или слаботоксичным, а не точное значение его токсичности, чтобы можно было использовать модели классификации. Для небольшого количества химических аналогов можно получить модели количественной зависимости структура-токсичность (QSTR) для прогнозирования значений токсичности. Для этих уникальных соединений перекрестный анализ также является возможным подходом для вывода определенной конечной точки токсичности на основе их сходных структур с экспериментальными значениями токсичности. Эти модели имеют высокую точность, особенно в локальном химическом пространстве, и иногда они могут заменить анализы *in vitro* или *in vivo* для определенных конечных точек.

Для прогнозирования токсичности было разработано несколько методов  in silico, таких как структурные предупреждения, перекрестное чтение (анализ), машинное обучение (МО) и количественные отношения структура-активность (QSAR). 

Структурная предупреждения (SA) — это химическая субструктура, связанная с токсичностью, она может представлять собой атом или совокупность атомов. Их можно определить с помощью специализированного человеческого опыта или статистического анализа фрагментированных наборов данных. SA обычно связывают наличие структурной опасности или ее комбинации с токсичностью.

Другой метод — перекрестное чтение (анализ), которое использует информацию о соединениях с известными свойствами для вывода информации о других подобных не охарактеризованных соединениях той же химической категории.  Сходство двух химических веществ можно рассчитать статистически, используя различные метрики расстояний. Хотя метод перекрестного чтения прозрачен и прост в реализации, при перекрестном чтении используется небольшой набор данных по сравнению с другими методами из-за небольшого количества аналогов соединения. Интерпретация результатов может быть сложной в зависимости от выбора сходства и его степени. Методы QSAR используют несколько типов алгоритмов машинного обучения для установления связи между химической структурой и токсичностью. В отличие от других методов, QSAR может количественно оценить взаимосвязь между структурой и токсичностью на основе их физико-химических свойств.

Машинное обучение — один из наиболее широко используемых методов для создания таких прогнозов. Используемые алгоритмы машинного обучения могут отображать как линейные, так и нелинейные зависимости.

В дальнейшем роль компьютерного моделирования на разных этапах создания лекарственных средств будет неуклонно возрастать, а совершенствование *in silico* и *in vitro* методов позволит минимизировать *in vivo* исследования и клинические испытания, сделав дизайн лекарств эффективнее, быстрее и дешевле.

# ГЛАВА 2. МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ БИОТОКСИЧНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ

## 2.1. ОБЩИЕ ПОДХОДЫ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ БИОТОКСИЧНОСТИ

В связи с недостатками, возникающими при использовании методов QSAR, популярность методов машинного обучения для исследования биотоксичности возрастает. В этих методах используется статистическая техника, позволяющая делать прогнозы на основе модели. Одним из главных преимуществ машинного обучения является то, что он позволяет моделировать и прогнозировать сложные проблемы, хотя для этого, как правило, требуются вычислительные мощности и большие объемы данных для обучения.

В рамках машинного обучения можно выделить две большие группы: обучение с учителем и обучение без учителя. Первые автоматически сопоставляют набор входных данных с набором выходных на основе аннотированных данных, а вторые позволяют изучать глубинные взаимосвязи непосредственно из заданного набора данных. В контексте оценки токсичности соединений обычно используется обучение с учителем, поскольку оно позволяет анализировать входные характеристики, связанные с соединениями, для получения конкретных результатов, таких как биологическая активность или токсические конечные точки. Наиболее часто используемыми для исследования токсичности моделями являются случайный лес (Random Forest), k-ближайших соседей (kNN), алгоритмы, основанные на нейронных сетях и методы глубоком обучении (рис.1.). Последние представляют собой один из наиболее распространённых методов машинного обучения, основанных на многослойных нейронных сетях.

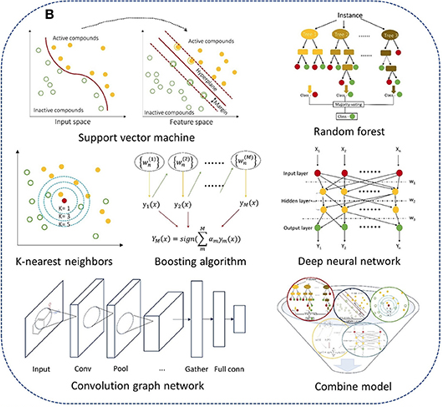


Рисунок 1. Современные алгоритмы машинного обучения [4].

Выбор того или иного метода зависит от конкретной задачи, а также от выбранного молекулярного представления. Построение модели обычно включает в себя:

1. сбор и подготовку данных;

2. описание данных (определение молекулярного представления);

3. построение модели путем обучения и валидации на части набора данных;

4. оценка модели, т.е. проведение тестирования на данных, ранее не встречавшихся модели (тестовый набор).

На каждом этапе существуют свои требования, гарантирующие надежность и точность моделей.

*Сбор данных*: качество экспериментальных данных является наиболее важным при построении модели. В настоящее время в Интернете доступно множество четко определенных данных, что значительно облегчает построение вычислительных моделей методами машинного обучения.

Ниже перечислены некоторые широко используемые базы данных, в том числе те, которые связывают химические структуры с показателями безопасности, белковыми мишенями и/или биологическими путями.

TOXNET — комплексный источник, объединяющий несколько баз данных по токсичности, таких как ToxLine и ChemIDplus.

ACToR — база данных, объединяющая данные из тысяч общедоступных источников ( Judson et al., 2008 ).

DSSTox, подмножество ACToR, предоставляет высококачественный ресурс для прогнозирования токсичности, включая данные ToxCast и Tox21.

Помимо данных о фенотипе, которые имеют непосредственное отношение к токсичности, для прогнозирования токсичности также важны базы данных о биологической активности, пути распространения и побочных эффектах. Несколько баз данных о биоактивности доступны бесплатно, например, PubChem, ChEMBL и BindingDB.

*Описание данных*: существует два способа представления химических структур в виде числовых характеристик, которые можно обрабатывать методами машинного обучения. Один из способов — использовать молекулярные дескрипторы, которые можно рассчитать на основе химических структур, физико-химических или топологических свойств. В настоящее время тысячи непрерывных и дискретных молекулярных дескрипторов могут быть получены с помощью наборов инструментов хемоинформатики, таких как PaDEL-Descriptor, OpenBabel, CDKit, RDKit или веб-серверы, такие как E-Dragon, ChemBCPP и ChemDes. Использование числовых функций может привести к переобучению, когда размер обучающего набора мал. Следовательно, выбор функций должен быть сделан до построения модели, чтобы снизить риск переобучения и повысить производительность модели.

Другой способ — использовать молекулярные отпечатки пальцев, которые представляют молекулу в виде двоичной строки, например MACCS, PubChemFP и KRFP. В молекулярном отпечатке пальца заранее определены списки субструктур или другие виды шаблонов. Если в молекуле присутствует указанный шаблон, соответствующий бит в двоичной строке устанавливается на «1», в противном случае он будет установлен на «0». По сравнению с молекулярными дескрипторами эти бинарные функции более интерпретируемы, поскольку каждый бит соответствует определенной субструктуре. Помимо обычных «отпечатков пальцев», для повышения предсказуемости моделей также можно использовать специальные шаблоны.

Различные *модели машинного обучения* описаны ниже в этой главе.

*Оценка модели*: для регрессионных моделей для оценки эффективности моделей часто используются три показателя оценки, а именно коэффициент корреляции момента произведения Пирсона (R2), средняя абсолютная ошибка (MAE) и среднеквадратическая ошибка (RMSE).

Для традиционных моделей бинарной или множественной классификации с одной меткой большинство показателей производительности рассчитываются на основе количества истинно положительных результатов (TP), истинно отрицательных результатов (TN), ложных положительных результатов (FP) и ложных отрицательных результатов (FN).

## 2.2. ТИПЫ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Различные типы моделей машинного обучения с разной степенью сложности могут предсказывать [молекулярные свойства](https://www.sciencedirect.com/topics/chemistry/molecular-property), такие как модели на основе сходства, линейные модели, модели на основе ядра, байесовские модели, древовидные модели и нейронные сети[8].

Эти типы моделей обычно можно разделить на модели, не относящиеся к нейронным сетям, и более сложные модели нейронных сетей. Среди моделей, не относящихся к нейронным сетям, можно дополнительно выделить модели, которые не соответствуют какой-либо конкретной функциональной форме, например модели, основанные на сходстве, и модели, которые подходят к ней на основе обучающих данных. Модели, основанные на сходстве, пытаются генерировать прогнозы для предоставленных соединений на основе известных молекулярных свойств подобных соединений. Среди моделей, которые соответствуют функциональной форме обучающим данным, можно выделить различные функциональные формы[9]. Первая функциональная форма — это линейные функции, которые используются в линейных моделях и моделях на основе ядра. Линейные модели используют линейные функциональные формы для изучения линейных связей между представлениями входных молекулярных структур и их молекулярными свойствами. Модели на основе ядра преобразуют входные данные с помощью функций ядра, чтобы улучшить прогнозирование данных. Более гибкие подходы включают байесовские модели, в которых точная функциональная форма может быть свободно определена, так что можно подогнать функцию к нескольким функциональным формам. Байесовские модели используют логику байесовского вывода, где они подгоняют к данным функцию свободной формы, чтобы минимизировать разницу между полученным распределением правдоподобия предсказаний модели и истинным апостериорным распределением данных. Эту свободную функциональную форму также можно определить более точно, используя конкретную функциональную форму, при которой подгонка затем будет выполняться только по гиперпараметрам функции. Другой функциональной формой, которую можно использовать, являются деревья решений, которые характеризуются узлами, представляющими изученные правила принятия решений, ветвями, которые являются результатами этих правил принятия решений, и которые могут в дальнейшем разветвляться на более глубокие подветви и конечные узлы, которые представляют результаты при следовании по определенному пути дерева. В процессе обучения решающие деревья изучают правила принятия решений, необходимые для разделения обучающих данных на классы или диапазоны значений. Древовидные модели могут использовать одно или несколько деревьев решений, построенных на основе набора функций из входных структур.

## 2.3. МОДЕЛИ, ОСНОВАННЫЕ НА СХОДСТВЕ

Первая группа описывает простой и понятный метод, который представляет собой метод k-ближайших соседей (kNN). Этот метод основан на сходстве между точками данных и использует принцип, согласно которому схожие точки данных также дадут аналогичный результат. Данный метод выполняет прогнозирование новых точек данных, применяя средневзвешенное значение между k-ближайшими соседями этой точки данных, где k может быть определяемым пользователем целочисленным значением. Для вычисления сходства можно использовать различные метрики. Тесно связанный с ним метод кластеризации  вместо этого использует ближайшее расстояние от центра установленных кластеров. Поскольку последний метод использует расстояния до центральных точек кластера, а не отдельные наиболее похожие точки данных, он больше подходит для задач классификации, а не регрессии.

Простота этих методов и отсутствие обучаемой функциональной формы позволяют легко использовать их без предварительного обучения работе с данными. Используя методы уменьшения размерности для многомерных данных, можно также получить визуализацию определенных кластеров и расстояний между точками данных, что может помочь легко объяснить прогнозы для новых точек данных. С другой стороны, простота отсутствия функциональной формы приводит к тому, что метод для прогнозирования полагается исключительно на сходство между точками данных. Это означает, что модель не может изучить сложные корреляции между точками данных, что может привести к недостаточной подгонке и плохой производительности на невидимых данных.

## 2.4. ЛИНЕЙНЫЕ МОДЕЛИ

Существуют различные типы линейных моделей, которые в основном различаются способами соответствия линейных функций входным данным. Линейные модели и многомерные линейные модели в случае нескольких переменных работают путем подгонки линейной функции к данным, чтобы либо установить функцию регрессии, например многомерную линейную регрессию (MLR), либо границу линейной классификации. Сами данные представлены в виде вектора признаков, и результирующая подобранная линейная функция будет иметь ту же размерность, что и вектор признаков входных данных. Гибкий дискриминантный анализ  можно рассматривать как расширение линейных моделей для задач многоклассовой классификации. Он использует непараметрическую регрессию в отличие от классического линейного дискриминантного анализа для поиска групп данных и применяет дополнительный классический линейный дискриминантный анализ для максимального разделения данных между найденными группами. Сплайны многомерной адаптивной регрессии (MARS) — это еще одно расширение линейных моделей, в котором полный набор данных разбивается на фрагменты и для каждого фрагмента строится несколько отдельных линейных моделей.

При построении вычислительных моделей необходимо учитывать возможность переобучения модели. Модели, которые переобучаются, становятся слишком приспособленными к обучающим данным, что снижает их универсальность и производительность на новых невидимых данных. Существует несколько расширений линейных моделей, которые могут решить эту проблему.

Регуляризация Лассо является одним из таких методов. Она использует L1 регуляризацию, добавляя штрафной коэффициент пропорциональный абсолютному значению весов модели. L1 устанавливает коэффициенты менее значимых параметров равными нулю, и добавляет штрафы для больших коэффициентов. В итоге получается простая модель с меньшей вероятностью. Напротив, регуляризация гребня, реализует L2 регуляризацию, добавляя штрафной коэффициент, пропорциональный квадрату весов, к функции потерь. Таким образом уменьшает веса до нуля, в результате чего получается более плавная модель с меньшими, но ненулевыми весами для всех объектов.

Регрессия частичных наименьших квадратов (PLS) — это еще один линейный метод решения задач регрессии, который особенно полезен, когда количество предикторов превышает количество наблюдений и когда между входными переменными существует мультиколлинеарность. PLS пытается сократить количество входных переменных до подмножества, которое максимально способно объяснить корреляцию с наблюдаемыми значениями.

В то время как отсутствие обучаемой функциональной формы в моделях, основанных на сходстве, не позволяет им изучать более сложные корреляции в данных, введение такой функциональной формы в линейных моделях позволяет им использовать изученные линейные зависимости в последующих прогнозах. Как и в случае со всеми обучаемыми функциями, необходимо помнить о переоснащении, которое для этих моделей можно устранить с помощью таких методов, как лассо и гребень. Однако простота линейной функциональной формы может привести к несоответствию этих моделей данным, где желаемые прогнозируемые свойства зависят от более сложных нелинейных взаимосвязей, которые часто встречаются среди биомедицинских и молекулярных данных.

## 2.5. МОДЕЛИ НА ОСНОВЕ ЯДРА

Одним из очевидных упомянутых недостатков линейных моделей является их неспособность подогнать к данным нелинейные функции, что может снизить их производительность при работе с данными, состоящими из сложных нелинейных взаимосвязей между входными характеристиками и прогнозируемыми молекулярными свойствами. Существует несколько моделей, в которых используются приемы ядра, преобразующие входные данные таким образом, что это может улучшить соответствие линейных моделей этим преобразованным точкам данных. Метод опорных векторов (SVM)  делает это путем преобразования входных данных в более высокие измерения, где они становятся линейно разделимыми. Это преобразование достигается с помощью функций ядра, таких как радиальные базисные функции. Первоначально SVM был определен для задач классификации,  но позже был расширен и для задач регрессии. Регрессия гребня ядра  применяет к входным данным тот же трюк с ядром, что и SVM, но использует регрессию гребня для построения линейной модели. Хотя обе модели кажутся очень похожими, разница заключается в построении линейной модели. В случае гребневой регрессии построение осуществляется путем подгонки данных к линейной функции, тогда как в SVM оно основано на использовании опорных векторов, которые представляют собой минимальные точки в данных, позволяющие описать оптимальную функцию-разделитель данных. Другой простой способ подогнать данные к выходным значениям — использовать радиальные базисные функции  для аппроксимации неизвестной функции.

Использование трюка с ядром позволяет использовать простые линейные функции, которые легче интерпретировать и которые имеют меньшую вероятность переобучения из-за ограниченного числа изучаемых параметров и линейной функциональной формы для данных, где интересующие молекулярные свойства определяются нелинейными связями со своими молекулярными дескрипторами. При этом тип используемой функции ядра имеет решающее значение, поскольку он определяет, насколько легко преобразованные входные данные можно смоделировать с помощью линейных функций. Кроме того, тип и сложность функции ядра также будут влиять на то, насколько легко интерпретировать модель, поскольку это сложнее для более сложных нелинейных функций ядра.

## 2.6. БАЙЕСОВСКИЕ МОДЕЛИ

Другим типом метода, позволяющим моделировать нелинейные зависимости и не задействующим более сложные модели нейронных сетей, являются гауссовские процессы. В этих моделях применяются принципы байесовского вывода и предполагаются априорное распределение вероятностей значений неизвестной функции, которую хочет смоделировать гауссовский процесс. Это распределение обновляется, чтобы полученная вероятность соответствовала апостериорному распределению. В отличие от линейных моделей и моделей на основе ядра, где набор случайных параметров соответствует фиксированной форме функции, гауссовские процессы позволяют моделировать априорное распределение вероятностей модельной функции по всем возможным функциональным формам и их параметрам. Случай байесовских моделей, в которых функциональная форма фиксирована, - это оценка максимального правдоподобия, которая использует байесовский вывод для оценки только параметров выбранного распределения вероятностей фиксированной формы.

Одним из важных преимуществ этих моделей является то, что по своей конструкции они способны давать оценки неопределенности сгенерированных прогнозов. Хотя такие оценки можно получить с помощью других моделей МО с помощью различных методов,  таких как частотные методы или некоторые байесовские подходы, такие как, например, ансамбли обучающих моделей, в чистых байесовских моделях МО такие оценки ошибок изначально включены. Они очень важны для моделей МО, используемых при разработке лекарств, по нескольким причинам. Новые лиганды часто могут находиться вне распределения данных, используемых для обучения моделей, где обобщаемость модели на данных вне распределения может быть плохой. Модели МО также используются для принятия решений на различных этапах открытия лекарств и могут быть включены в циклы активного обучения. Однако при использовании байесовских моделей необходимо уделять внимание предварительному выбору, поскольку это может повлиять на производительность модели. Особенно при использовании информативных априорных данных важно, чтобы они оставались применимыми к данным.

## 2.7. ДРЕВОВИДНЫЕ МОДЕЛИ

Случайный лес (RF) и чрезвычайно рандомизированные деревья (ET) [10], как примеры древовидных моделей.

**Случайный лес и чрезвычайно рандомизированные деревья относятся к классу алгоритмов, известных как алгоритмы ансамблевого обучения**. Алгоритмы ансамблевого обучения используют возможности многих алгоритмов обучения для выполнения задачи. Например, в задаче классификации алгоритм ансамблевого обучения может агрегировать прогнозы от нескольких разных классификаторов, чтобы сделать окончательный прогноз.

Эта концепция основана на представлении о том, что использование нескольких алгоритмов обучения может привести к лучшему окончательному прогнозу.

**Случайный лес (Random forest, или RF)** создает несколько деревьев решений (лес) во время обучения на разных подмножествах обучающих данных. Аналогично, идея остается той же: результаты нескольких объединенных деревьев, вероятно, будут лучше, чем результаты одного дерева.

Учитывая набор данных с несколькими функциями, RF-алгоритм будет выбирать подмножества наблюдений с различными функциями из набора данных. На основе этого подмножества строится дерево решений. Этот процесс выборки подмножеств с заменами известен как начальная загрузка.

При построении дерева решений RF-алгоритм будет выбирать наиболее оптимальное разделение в каждом узле. Затем этот процесс повторяется для другого подмножества данных с различными характеристиками, пока не будет построено указанное количество деревьев.

После получения результатов со всех деревьев окончательный прогноз будет получен путем голосования большинства за классификацию или усреднения для регрессии.

RF можно применять практически к любой задаче классификации или регрессии.

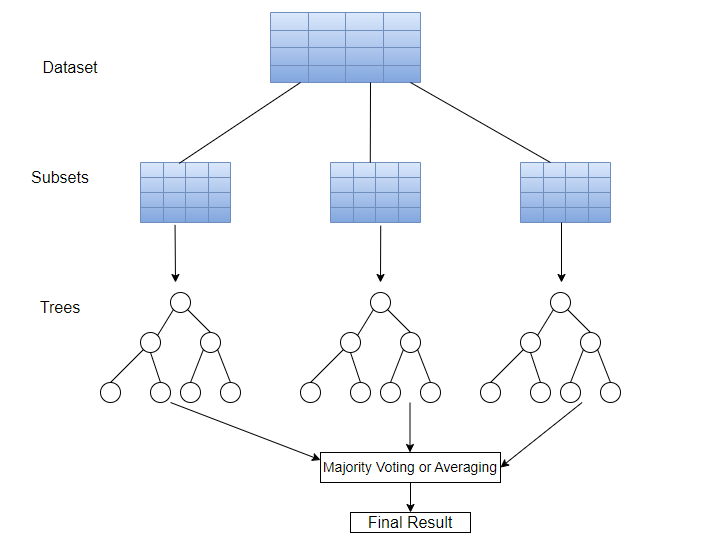


Рис. 2 RF-алгоритм [10]

### Преимущества: с**лучайные леса надежны и хорошо справляются как с задачами регрессии, так и с классификацией** . Кроме того, RF-алгоритмы хорошо работают с большими наборами данных и различными типами данных, такими как числовые, двоичные и категориальные.

Недостатки: при большом количестве деревьев сложность и время вычислений относительно высоки, что приводит к увеличению времени обучения. Кроме того, выборка подмножеств может внести некоторую предвзятость.

**Чрезвычайно рандомизированные деревья (ExtraTrees, или ET), также известные как дополнительные деревья, строят несколько деревьев во время обучения по всему набору данных**. Во время обучения ET будет строить деревья по каждому наблюдению в наборе данных, но с разными подмножествами функций.

Важно отметить, что хотя начальная загрузка не реализована в исходной структуре ET, можно ее добавить в некоторые реализации. Более того, при построении каждого дерева решений алгоритм ET случайным образом разбивает узлы.

Преимущества: является уменьшение смещения. Это с точки зрения выборки из всего набора данных во время построения деревьев. Различные подмножества данных могут вносить разные отклонения в полученные результаты, поэтому дополнительные деревья предотвращают это, отбирая весь набор данных.

Еще одним преимуществом дополнительных деревьев является то, что они уменьшают дисперсию. Это результат рандомизированного разделения узлов в деревьях решений, поэтому на алгоритм не сильно влияют определенные функции или закономерности в наборе данных.

RF и ET похожи в том, что они оба создают несколько деревьев решений для использования для решения поставленной задачи, будь то классификация или регрессия. Однако между ними существуют тонкие различия.

|  |  |
| --- | --- |
| RF | ET |
| Предварительно загруженные подмножества данных | Выборка всего набора данных |
| Узлы разделены с учетом наилучшего разделения | Рандомизированное разделение узлов |
| Средняя дисперсия | Низкая дисперсия |
| Требуется время, чтобы найти лучший узел для разделения | Быстрее, поскольку разделение узлов является случайным |

Самое главное, что выбор того, какой из них использовать, всегда зависит от доступного набора данных и поставленной задачи.

Данные древовидные методы показали хорошие результаты для различных задач прогнозирования молекулярных свойств ADMET. Важным аспектом, который также применим и к другим методам MО, является определение характеристик входных данных. Поскольку это напрямую влияет на производительность моделей MО и может позволить аппроксимировать сложные трехмерные входные данные более простыми моделями, без необходимости прибегать к более сложным моделям нейронных сетей на основе графов. Это может быть хорошим решением, когда доступно мало обучающих данных для соответствия модели без переобучения.

# ГЛАВА 3. ПРОБЛЕМЫ ПРЕДСКАЗАНИЯ БИОТОКСИЧНОСТИ IN SILICO

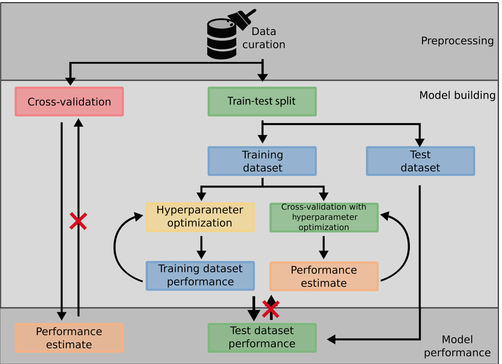
Хотя в прогнозировании химической токсичности *in silico* за последние годы достигнут хороший прогресс, все еще существуют некоторые проблемы и ограничения, требующие улучшения.

Качество данных по-прежнему остается большой проблемой. В настоящее время многие данные о токсичности получены в результате высокопроизводительных анализов *in vitro* или испытаний *in vivo* на животных. Например, Tox21 и ToxCast предоставляют данные об активности тысяч химических веществ на основе сотен анализов. Крайне востребованы дополнительные данные клинических испытаний и клинического применения лекарств.

Все методы, описанные выше, имеют общую черту: все они требуют наборов данных, которые можно проанализировать или использовать для обучения модели. Машинное обучение требует больших объемов данных, чтобы предотвратить переобучение. Однако если наборы данных для традиционного машинного обучения невелики, их применимость также становится очень ограниченной. Для перекрестного считывания требуется наименьшее количество данных, при этом для экстраполяции активности достаточно лишь нескольких соединений с высоким сходством. Тем не менее, найти этот конкретный тип данных может быть еще труднее, особенно для новых структурных классов. Поскольку все методы основаны на данных, крайне важно убедиться, что эти данные содержат как можно меньше ошибок, поскольку они оказывают непосредственное влияние на результат моделирования. Поэтому курирование данных имеет важное значение. Курирование данных, с одной стороны, представляет собой удаление неопределенных или неубедительных значений активности, а с другой стороны, оно предполагает структурное курирование. Удаление точек данных всегда представляет собой балансировку между наличием очень небольшого количества надежных точек данных или большим набором данных с большим количеством (менее надежными) точками данных. Его всегда необходимо тщательно обдумывать и адаптировать к методу.

Для машинного обучения оценка модели также имеет решающее значение. Необходимо разработать больше вычислительных методов для повышения точности прогнозных моделей. В любом случае следует избегать переобучения и тщательно оценивать модель. Модели могут быть проверены путем (а) перекрестной проверки, (б) разделения набора данных на тестовую и обучающую выборки или (в) набора данных внешней оценки.

Перекрестная проверка является самой простой в использовании и может использоваться в случаях, когда доступность данных очень ограничена. Однако она дает только оценку производительности модели и может быть смещена из-за очень похожих кратностей перекрестной проверки. Таким образом, это может не отражать производительность в реальных сценариях. Разделение на тестовую и обучающую выборки обычно используется, когда размер набора данных достаточно велик, чтобы можно было сохранить данные исключительно для тестирования. С помощью этого метода можно гораздо лучше оценить производительность «реального мира», однако часть доступных данных не используется для обучения. Это означает, что модель не извлекает выгоду из всей доступной информации. Таким образом, лучшим методом обучения является оценка с использованием истинного внешнего набора тестов, который был составлен независимо от набора обучающих данных. Естественно, внешний набор тестов должен находиться в области применимости модели.



**Рис. 3** Схематическое изображение возможных способов обучения модели машинного обучения [5].

На рисунке 3 левый путь подчеркивает, что перекрестная проверка возможна без разделения набора данных; однако полученные характеристики являются лишь оценочными, что разделение обучения и тестирования необходимо, например, если необходимо настроить гиперпараметры (в этом случае можно использовать вторую внутреннюю перекрестную проверку) или, в общем, если должна быть реализована ситуация «реального мира». Важно отметить, что после того, как набор тестов был использован для оценки производительности модели, никогда не следует возвращаться к корректировке параметров с помощью обучающего набора, чтобы получить лучшую производительность на тестовом наборе, поскольку это может привести к смещению модели.

Особую осторожность следует проявлять, когда обучение проводится с несбалансированными наборами данных. Из-за обучения с помощью жадных алгоритмов многие алгоритмы машинного обучения (включая нейронные сети) имеют тенденцию игнорировать класс меньшинства, поскольку ошибка становится довольно небольшой, когда все метки прогнозируются как класс большинства. Следовательно, процедуру обучения, возможно, придется изменить за счет использования экономически чувствительного обучения или методов избыточной или недостаточной выборки.

Хотя некоторые методы машинного обучения могут определять важность признаков, ни один метод пока не может дать человеческое объяснение прогноза. Это одна из причин, почему токсикология *in silico* часто используется в качестве индикатора, но окончательное решение всегда будет приниматься экспертной оценкой, исследованиями *in vitro* или *in vivo*.

Важно знать, как была построена модель и особенно для какой цели. Модель прогнозирования токсичности на ранней стадии не должна быть слишком чувствительной. На этом этапе опасности должны быть конкретными, чтобы можно было принять решение о необходимости дальнейшего использования соединения.

Все вычислительные методы прогнозирования имеют свои преимущества и недостатки. Разработчикам моделей машинного обучения и пользователям важно указать их и помнить об ограничениях, особенно во время интерпретации результатов.

Также разработчики моделей должны следить за тем, чтобы пользователь был проинформирован о процессе и предполагаемом использовании модели. Это приведет к большему доверию к моделям и может преодолеть часто слишком строгую оценку вычислительных моделей.

# ГЛАВА 4. АНАЛИЗ БИОТОКСИЧНОСТИ ИНГИБИТОРОВ BCR-ABL ТИРОЗИНКИНАЗЫ

В данной работе была использована одна из моделей машинного обучения, относящихся к древовидным структурам, а именно Случайный лес (Random forest, RF).

В качестве соединений для контроля качества были использованы ингибиторы BCR-ABL тирозинкиназы с уже известными для них классами токсичности.

В общем случае вещества по своей токсичности могут быть разделены на 6 классов:

Класс 1 (A): смертельно при проглатывании (LD50 ≤ 5);

Класс 2 (B): смертельно при проглатывании (5 < LD50 ≤ 50);

Класс 3 (C): токсичен при проглатывании (50 < LD50 ≤ 300);

Класс 4 (D): вреден при проглатывании (300 < LD50 ≤ 2000);

Класс 5 (E): может быть вреден при проглатывании (2000 < LD50 ≤ 5000);

Класс 6 (F): нетоксичный (LD50 > 5000).

Smiles каждого ингибитора и соответствующий ему класс токсичности представлены ниже в качестве фрагмента кода:

Smiles:

nilotinib\_smiles = "CC1=C(C=C(C=C1)C(=O)NC2=CC(=CC(=C2)C(F)(F)F)N3C=C(N=C3)C)NC4=NC=CC(=N4)C5=CN=CC=C5"

imatinib\_smiles = "CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5"

dasatinib\_smiles = "CC1=C(C(=CC=C1)Cl)NC(=O)C2=CN=C(S2)NC3=CC(=NC(=N3)C)N4CCN(CC4)CCO"

bosutinib\_smiles = "CN1CCN(CC1)CCCOC2=C(C=C3C(=C2)N=CC(=C3NC4=CC(=C(C=C4Cl)Cl)OC)C#N)OC"

ponatinib\_smiles = "CC1=C(C=C(C=C1)C(=O)NC2=CC(=C(C=C2)CN3CCN(CC3)C)C(F)(F)F)C#CC4=CN=C5N4N=CC=C5"

asciminib\_smiles = "C1CN(CC1O)C2=C(C=C(C=N2)C(=O)NC3=CC=C(C=C3)OC(F)(F)Cl)C4=CC=NN4"

Классы токсичности:

nilotinib\_real\_tox = 4

imatinib\_real\_tox = 2

dasatinib\_real\_tox = 4

bosutinib\_real\_tox = 4

ponatinib\_real\_tox = 5

asciminib\_real\_tox = 5

В качестве данных для анализа был использован датасет, содержащий около ста сорока тысяч smiles соединений, заранее полученных из базы данных. С целью применений модели Случайного леса для предсказания класса токсичности выбранных соединений были произведены следующие действия с целью подготовки данных:

1. разделение данных на обучающую и тестовую выборки (train-test split).

В качестве него было выбрано часто используемое для модели RF разделение 70-30, где 70% всего набора данных составляет обучающая выборка и 30% - тестовая.

1. приведение данных в вид, подходящий для модели RF

Для дальнейшей работы с соединениями данные были переведены из вида smiles в вид молекулярных отпечатков (строку, состоящую из нулей и единиц). Для этого были использованы модули Chem и AllChem библиотеки rdkit.

Ниже представлен фрагмент кода с преобразованием smiles одного из ингибиторов.

nilotinib\_mol = Chem.MolFromSmiles(nilotinib\_smiles)

nilotinib\_fingerprint = AllChem.GetMorganFingerprintAsBitVect(nilotinib\_mol, *radius*=2, *nBits*=1024)

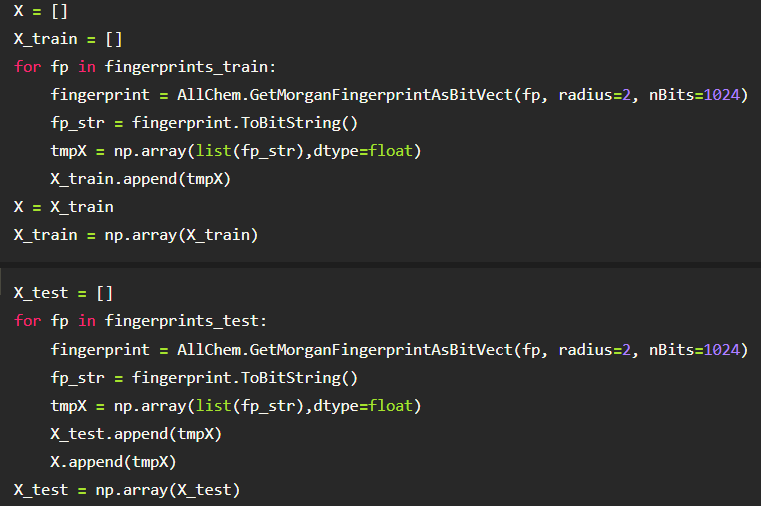
fp\_str = nilotinib\_fingerprint.ToBitString()

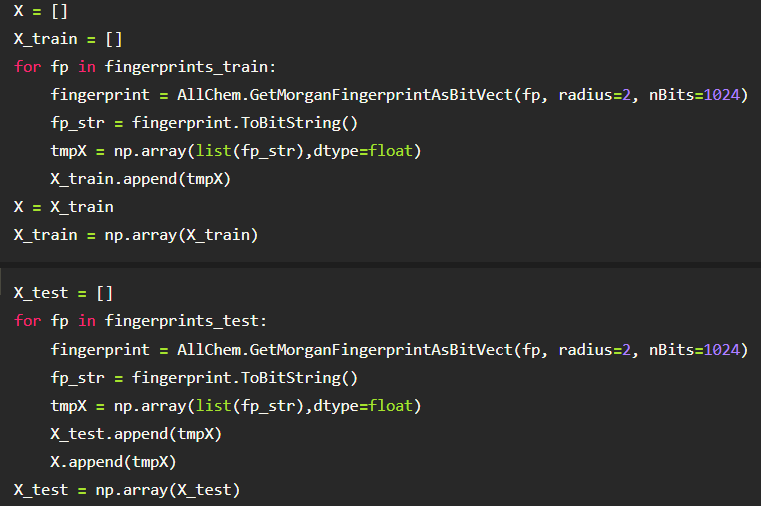
Smiles всех остальных ингибиторов были обработаны и преобразованы в молекулярные отпечатки аналогичным образом.

1. создание списков молекулярных отпечатков исследуемых данных.

Данный шаг выполнен аналогично предыдущему.

В итоге получены следующие списки отпечатков для train и test наборов:



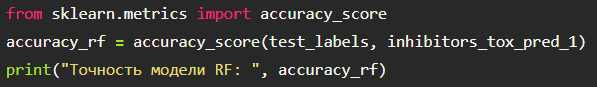


За выполнением действий, описанных выше, следует непосредственно обучение выбранной модели и предсказание классов токсичности для соединений тестовой выборки.

Последний этап – это контроль качества модели или оценка её точности.

Для этого с помощью ранее обученной модели предсказывается класс токсичности для ингибиторов BCR-ABL тирозинкиназы, а затем, используя метрикy, оценивается точность модели.

В данной работе была использована метрика accuracy, отражающая общее число правильных предсказаний. Для её определения была использована функция accuracy\_score, импортированная из библиотеки sklearn, а так же посчитана loss function(функция ошибки), в качестве которой в данной работе была использована функция потерь Multiclass Cross-Entropy. Импортирование функции accuracy\_score, а так же использование вышеупомянутой метрики accuracy представлено в фрагменте кода ниже:



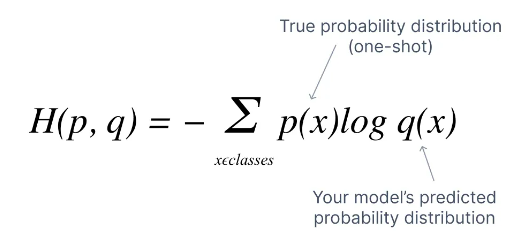
Результат вывода:



Где **test\_labels** – реальные значения классов токсичности ингибиторов;

**inhibitors\_tox\_pred\_1** – предсказанные классификатором значения.

Вычисление функции ошибки Multiclass Cross-Entropy проводилось с помощью следующей формулы:



[11]

Где p(x) – реальное значение класса токсичности соединения, к которому была применена OneHotEncoding кодировка (т.е. преобразование значения класса в список из нулей и единиц); q(x) – вероятность принадлежности соединения каждому из классов, представленная так же в виде списка значений.

Значения Cross-Entropy, полученные для всех исследуемых ингибиторов, представлены ниже:



# ЗАКЛЮЧЕНИЕ

Таким образом, на основании вышеизложенного можно сформулировать следующие выводы:

1. применение различных методов *in silico* перед клиническими и доклиническими испытаниями позволяет выявлять потенциально опасные соединения на ранних этапах разработки лекарств, что ускоряет и удешевляет процесс создания лекарственных средств.
2. актуальной задачей предсказания токсичности для ингибиторов BCR-ABL тирозинкиназы является комбинирование различных моделей машинного обучения с целью повышения точности прогнозирования.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Методология поиска новых биологически активных фармакологических веществ с рецепторной активностью [Электронный ресурс] / // Международный журнал прикладных и фундаментальных исследований. – 2016. – Режим доступа: https://applied-research.ru/ru/article/view?id=10116 – Дата доступа: 25.01.2024
2. Современные методы поиска новых лекарственных средств [Электронный ресурс] /Ю.С.Головко, О.А.Ивашкевич, А.С.Головко //Вестник БГУ. Сер.2.2012 №1 – 2012. – Режим доступа: https://core.ac.uk/download/pdf/290240656.pdf – Дата доступа: 25.03.2024
3. Использование методов QSAR-моделирования для предсказания токсичности координационных соединений / Н.И. Баранов, А.В, Бурякина, Л.С.Гредягина, А.В, Москвин, Е.В. Федлрова// Биомедицина №1. – 2016.
4. *In Silico* Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts [Электронный ресурс] / Hongbin Yang // Frontier in Chemistry. – 2018. – Режим доступа:  [https://frontiersin.org/articles/10.3389/fchem.2018.00030/full](%20https://frontiersin.org/articles/10.3389/fchem.2018.00030/full) – Дата доступа: 19.04.2024.
5. In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways [Электронный ресурс] / Jennifer Hemmerich, Gerhard F. Ecker// WIREs. – 2020. – Режим доступа: https://www.wires.onlinelibrary.wiley.com/doi/10.1002/wcms1475 – Дата доступа: 29.01.2024.
6. An ensemble model of QSAR tools for regulatory risk assessment [Электронный ресурс] / Prachi Pradeep, Richard J.Povinelli, Shannon White, Stephen J.Merrill// Juornal of Cheminformatics – 2016. – Режим доступа: / <https://www.jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0164-0> – Дата доступа: 18.02.2024.
7. In silico prediction of toxicity and its applications for chemicals at work [Электронный ресурс] /Kyung-Taek Rim// National Center for Biotechnology Information. – 2020. – Режим доступа: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7223298/ – Дата доступа: 26.03.2024.
8. Machine learning small molecule properties in drug discovery [Электронный ресурс] / Nikolai Schain, Maciej Majewski // Artificial Intelligence Chemistry. – 2023. – Режим доступа: <https://www.sciencedirect.com/science/article/pii/S2949747723000209> – 24.04.2024
9. MolToxPred: small molecule toxicity prediction using machine learning approach [Электронный ресурс]/ Anjali Setiya, Vinod Jani, Uddhavesh Sonavane, Rajendra Joshi // National Center of Medicine. – 2024. – Режим доступа:

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10826801/ – Дата доступа:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10826801/%20–%20%20Дата%20доступа:/) 17.02.2024

10. Random Forest vs Extremely Randomized Trees [Электронный ресурс]/ Emmanuela Budu// Baeldung. – 2023. – Режим доступа: <https://www.baeldung.com/cs/random-forest-vs-extremely-randomized-trees> – Дата доступа: 12.05.2024.

11. Cross Entropy loss: Intro, Application [Электронный ресурс] – 2024. – Режим доступа: <https://www.v7labs.com/blog/cross-entropy-loss-guide> - Дата доступа: 16.05.2024