**Содержание**

[Список использованных сокращений и условных обозначений 2](#_Toc166717337)

[Термины и определения 3](#_Toc166717338)

[Введение 4](#_Toc166717339)

[1. Обзор и анализ предметной области 6](#_Toc166717340)

[1.1 Методики идентификации растительных масел 6](#_Toc166717341)

[1.2 Машинное обучение в естественных науках 9](#_Toc166717342)

[1.3 Машинное обучение на данных ИК-спектроскопии 11](#_Toc166717343)

[2. Материалы и методы 14](#_Toc166717344)

[2.1 Машинное обучение 14](#_Toc166717345)

[2.2 Метод ИК-спектроскопии с преобразованием Фурье 20](#_Toc166717346)

[2.3 Объект исследования и оборудование 20](#_Toc166717347)

[Список использованных источников 28](#_Toc166717348)

## Список использованных сокращений и условных обозначений

ГОСТ – государственный стандарт

FTIR - инфракрасная спектроскопия с преобразованием Фурье

ИК – инфракрасный диапазон

ЯМР - спектроскопия ядерного магнитного резонанса

УФ спектроскопия - спектроскопия в ультрафиолетовом диапазоне

SMILES – simplified molecular input line entry system

ROC - receiver operating characteristic curve

MAE - средняя абсолютная ошибка

MSE - среднеквадратичная ошибка

МНК - метод наименьших квадратов

PCA - анализ главных компонент

MDS - multidimensional scaling

## Термины и определения

Стандартная библиотека языка программирования – набор готовых программных решений для заданного языка программирования, доступных к использованию по умолчанию, без необходимости дополнительной установки.

Векторная нормализация –

Корректировка базовой линии –

Стохастический градиентный спуск –

Перцептрон – алгоритм машинного обучения, позволяющий осуществлять бинарную классификацию. Является прототипом современных нейронных сетей.

Квантиль – понятие математической статистики, указывающая на значение, которое заданная величина не превышает с фиксированной определенной вероятностью.

## Введение

Растительные пищевые масла являются необходимым элементов питания человека, что делает их неотъемлемой частью пищевой промышленности. В связи с хроническим дефицитом сырья для их производства в индустрии производства масел существует проблема фальсификации, заключающаяся в добавлении посторонних примесей и подмене дорогостоящих высококачественных масел более дешевыми, и даже техническими маслами. Этот факт не только нарушает установленные стандарты производства и реализации пищевой продукции, но и создает угрозу для здоровья потребителей.

В настоящее время для установления случаев подмены масел используются передовые лабораторные методы, такие как спеткроскопия ядерного магнитного резонанса (ЯМР), рамановская спектроскопия, масс-спектроскопия (МС) и инфракрасная спектроскопия с преобразованием Фурье (FTIR). Эти методы позволяют получить данные высокого разрешения в сочетании с хемометрическими методами, позволяющими различать настоящие и фальсифицированные пищевые масла.

Также, возрастает доля научных исследований, посвященных применению методов машинного обучения в естесственно-научных областях при решение прикладных задач. Наличие данных высокого разрешения, содержащих в себе информацию, позволяющую различать образцы фальсифицированного и настоящего масла, позволяет автоматизировать процесс анализа результатов спектроскопических исследований и выявлять новые закономерности в данных, что также может привносить интерес в исследование.

В связи с отсутствием на данный момент методики ГОСТ для определения наличия рапсового масла в подсолнечном, актуальной оказалась разработка такой методики на основе FTIR спектроскопии с применением алгоритмов машинного обучения в качестве хемометрических методов для классификации образцов и выявления тенденций при анализе спектральных данных.

Целью данной работы является разработка стратегии обработки спектральных данных смесей растительных масел и примение методов машинного обучения для классификации смеси подсолнечных и рапсовых масел и количественного определения их компонентного состава. Для достижения поставленной цели были выполнены следующие задачи:

1. Сбор и предобработка спектральных данных различных подсолнечных и рапсовых масел и их попарных смесей.

2. Классификация ИК спектров смесей данных пищевых масел.

3. Поиск алгоритмов, подходящих для количественного определения компонентов в смесях.

4. Анализ и сравнение прогнозных способности предложенных моделей.

## 1. Обзор и анализ предметной области

### 1.1 Методики идентификации растительных масел

Контроль качества растительных масел может осуществляться различными методами. Одним из возможных путей идентификации фальсификаций являются физико-химические показатели масел, такие как йодное число, перекисное число, кислотное число и т.д. [1]. Однако в современной промышленности требуется наличие более точных методов, позволяющих оценить более широкий диапазон параметров растительных масел для определения их качества.

Одним из таких способов является разделение смесей масел при помощи хроматографических методов. Суть данных методов заключается в различном распределении компонентов смеси между фазами установки в зависимости от их физико-химических свойств. Растительные масла являются многокомпонентными смесями, однако для каждого вида масла характерно преобладание остатков определенного набора жирных кислот, и в случае фальсификации масел добавлением сторонних видов, в результате проведения хроматографии можно установить наличие посторонних примесей путем анализа их жирнокислотного состава.

Так, в работе [2] для определения качества растительных масел использовались методы газовой хроматографии с масс-спектроскопическим детектором. В исследовании [3] описывается проблема определения качества облепихового масла методов высокоэффективной жидкостной хроматографии (ВЭЖХ), заключающаяся в проверке гипотезы о том, что по результатам проведения эксперимента можно установить наличие триглицеридов, не характерных для масла облепихи, например, подсолнечного масла. Было показано, что триглицериды, образованные остатками разных жирных кислот элюируются с различной скоростью, что позволяет разделять их по хроматограммам, полученным в результате проведения ВЖЭХ [3].

Еще одним способом получения данных высокого качества при анализе качества растительных пищевых масел является семейство спектральных методов, таких как флюоресцентная спектроскопия, спектроскопия в ультрафиолетовом диапазоне (УФ спектроскопия), флуоресцентная спектроскопия [3], инфракрасная спектроскопия с преобразованием Фурье, ЯМР спектроспия.

Метод ЯМР является высокоэффективным способом анализа качества масел. В работе [1] было показано, что проведение анализа спектров ЯМР значительно упрощает процесс анализа состава растительных масел в связи с отличающимися химическими сдвигами углеродных атомов различных жирных кислот, что результирует в наличие отдельных сигналов в спектре, позволяя определять жирнокислотный состав смесей и по наличию посторонних сигналов в сравнении с эталонным спектром образца рассматриваемого масла, позволяющих указать на наличие постороннего вида масла в составе смеси.

Также находит применение в исследовании качества растительных масел метод спектроскопии в ультрафиолетовом диапазоне (УФ-спектроскопия). Данный способ позволяет получать информацию об изменениях в окислительном профиле масел, позволяя выявлять нарушения в качестве по окислительным коэффициентам, свидетельствующим об окислительном состоянии соединений [5]. При этом можно получить данные о качестве масел, изменениях в их качестве при хранении и в процессе промышленной переработки при производстве. Для каждого вида масла характерен свой жирнокислотный состав, влияющий на уровень насыщения связей и способность к окислению, поэтому информация об окислительном состоянии вещества, полученная методом УФ-спектроскопии, обладает ценностью при проведении контроля качества масел как продуктов питания. Изменения в качестве растительных жиров, результирующие в окислительные процессы, происходящие при нарушениях условий хранения или производства, снижают ценность масел, так как высококачественные дорогостоящие масла обладают полезностью в том числе и из-за того, что могут выступать в роли источника антиоксидантов в питании человека. В свою очередь, изменения в окислительном состоянии масел влияют на эту способность, снижая уровень полезности растительных жиров.

Еще одним спектральным методом, позволяющим оценивать качество растительных масел, является масс-спектроскопия (МС). В работе [4] исследовалась фальсификация оливкового масла холодного отжима соевым маслом. Было показано, что МС может быть использована для решения данной задачи, так как данный способ позволяет выявлять различия в триглицеридном составе. В зависимости от вида масла, для него характерно преобладание того или иного набора жирных кислот, обладающих различной массой и зарядом. Получаемые в МС спектры позволяют оценить представленность компонентов в веществе, и путем сравнения полученных спектров с эталонными может быть установлен случай фальсификации масел путем смешивания более дорогих растительных жиров с более дешевыми.

Исследование [5] описывает методику применения FTIR спектроскопии для анализа смесей ряда растительных масел с подсолнечным. Авторы указывают, что при визуальном анализе полного спектра, полученного при проведении измерений, сложно найти существенные отличия между образцами. Однако при детальном рассмотрении характеристических регионов спектров имеются различия в интенсивности пиков поглощения. Такими регионами являются области 3100-2800 см-1, 1800-1600 см-1 и 1600-1390 см-1 [5], интенсивность поглощения в которых говорит о доле насыщенных и ненасыщенных остатков жирных кислот, а также об общем уровне насыщенности связей в триглицеридах, что позволяет проводить различие между чистыми образцами масел определенного вида, а также смесями различных видов масел, что является признаком недоброкачественности производителей и свидетельствует о наличии акта фальсификации продукции. Было показано, что в зависимости от количественного содержания примеси растительного масла в образцах интенсивность пиков поглощения изменяется [5], что говорит о возможности использования метода FTIR спектроскопии для определения наличия посторонних примесей в образцах реализуемых в продаже растительных масел.

### 1.2 Машинное обучение в естественных науках

В настоящее время машинное обучение находит широкое применение в естественнонаучных областях. Модели машинного обучения используют для предсказания физических и химических свойств материалов [6], моделирования и прогнозирования структуры белков [11], дизайне и разработке лекарств [16, 17], оценке токсичности наноматериалов [15]. Успех в стремительное развитие применения методов машинного обучения в науке обусловлен наличием больших объемов данных, получаемых исследователями в доменных областях, который продолжает пополняться с каждым днем [14]. Машинное обучение дает исследователям широкий набор инструментов, позволяющих выявлять закономерности в данных и разрабатывать предиктивные модели. Это делается как с целью получения новых знаний, так и для автоматизации рутинных задач обработки данных, получаемых в лабораториях.

Задача расчета физико-химических свойств веществ является сложной и требует больших временных и вычислительных ресурсов. Вместе с тем не все свойства веществ могут быть рассчитаны традиционными вычислительными методами. В работе [20] при рассмотрении набора растворителей пластика, производных от глицерина, исследователи столкнулись с проблемой низкой точности предсказания растворимости веществ в воде LogS, логарифмического коэффициента распределения LogP, температуры плавления при помощи метода параметров растворимости Хансена на практике (HSPiP – Hansen Solubility Parameters in Practice) [21] – методе, основанном на оценке параметров дисперсионных, полярных и водородных взаимодействий в молекулах растворяемого вещества и растворителя, и прогнозировании растворимости на основании сходства либо различия данных параметров молекул. В связи с этим были построены полносвязная нейронная сеть, а также двумерная графовая сверточная нейронная сеть. Нейронная сеть для предсказания LogS позволила добиться точности предсказания 86 %, точность предсказания LogP составила 95 %, а сеть для прогнозирования температуры плавления достигла точности 92 %. Во всех случаях точность разработанных алгоритмов оказалась выше, чем у расчетов по методу HSPiP, что говорит об успешности применения методов машинного обучения при решении данной задачи.

Однако, на данный момент существует проблема репрезентации молекул химических веществ для подачи данных в модели машинного обучения [12]. Данная проблема важна, так как производительность и эффективность работы алгоритмов тесно связана с представлением и форматом данных, которые подаются на вход модели. Одним из самых популярных вариантом представления химических соединений является запись формул в формате SMILES [13]. Этот формат удобен для чтения людьми, но вычислительные машины справляются с ним не так хорошо, и несмотря на развитие моделей обработки естественного языка (NLP – natural language processing), эффективность алгоритмов машинного обучения при обработке данных в формате SMILES остается невысокой. В связи с этим при решении большого количества задач исследователи прибегают к использованию базовых физико-химических параметров молекул, а также используют методы векторизации строк формата SMILES при помощи полносвязных нейронных сетей. Описанная проблема вызывает научный интерес, так как разработка унифицированного метода представления молекул может совершить прорыв в области машинного обучения в химии, открывая новые возможности для исследования и позволяя улучшить производительность уже существующих алгоритмов.

Еще одной областью применения алгоритмов машинного обучения является разработка и дизайн лекарств. На данный момент имеются огромные объемы биомедицинских данных, связанных с раковыми заболеваниями. Лекарственным эффектом применимо к данной области потенциально могут обладать молекулы, подавляющие активность определенных веществ, преимущественно белков, в больных тканях. Поиск таких молекул является крайне дорогостоящим из-за необходимости прохождения клинических испытаний, призванных оценить безопасность препарата и его эффективность. Проблема заключается в том, что поиск самой молекулы может занимать долгое время, и в случае успешного нахождения кандидатов и отправки их на клинические испытания, препараты могут быть отвергнуты из-за негативных результатов на одной из стадий испытаний. В результате все затраты временных и материальных ресурсов оказываются бессмысленными. В связи с этим актуальной становится актуальным поиск методов, позволяющих оценивать эффективность и селективность, а также безопасность того или иного препарата по отношению к больным и здоровым клеткам и тканям соответственно. работе [22] были проанализированы транскриптомные данные различных раковых заболеваний для препаратов на третьей стадии клинических испытаний с целью выявления, какие целевые гены могут оказывать влияние на успех клинических испытаний. Ценность исследования заключается в том, что в зависимости от вида ракового заболевания существует разница между уровнем экспрессии тех или иных генов в больных и здоровых тканях, что может являться маркером заболевания и целевым компонентом, против которого должно быть направлено фармакологическое действие искомого препарата. При помощи статистического анализа данных, построения моделей машинного обучения на основе логистической регрессии и случайного леса, а также их итеративного обучения из более чем 150000 параметров было выявлено, что одними из главных маркеров раковых заболеваний являются средняя экспрессия матричной рибонуклеиновой кислоты (мРНК) в тканях и стандартное отклонение экспрессии мРНК среди тканей [22].

### 1.3 Машинное обучение на данных ИК-спектроскопии

Спектральный анализ предоставляет данные высокого разрешения, содержащие в себе определенные закономерности и характеризующие химическую структуру исследуемых веществ и материалов. В сочетании с методами машинного обучения и знаниями в предметной области, исследователи могут проводить анализ данных с целью выявления имеющихся закономерностей, а также строить предиктивные модели машинного обучения, позволяющие автоматизировать процессы, такие как интерпретация данных спектральных исследований, тем самым повышая эффективность исследований и уменьшая время и трудозатраты на выполнение рутинных задач.

Например, в работе [19] было исследовано применение методов машинного обучения для автоматизации обнаружения заболеваний растений маниоки по ИК-спектрам их поверхности. Маниока – одно из самых важных съедобных растений в мире, особенно распространенное в тропических районах [23]. Оно может быть подвержено болезням коричневых пятен (CBSD – cassava brown streak disease) и мозаичной болезни (CMD – cassava mosaic disease), вызываемых вирусами и являющихся одними из основных угроз для производства данного растения [23]. Одной из проблем, с которой можно столкнуться при работе со спектральными данными, является количество параметров в выборке данных, равное длине спектрального диапазона и обычно находящееся в районе двух-трех тысяч параметров. Данная проблема решается либо применением метода главных компонент для снижения размерности данных, либо путем выбора участков спектра, которые являются характеристическими областями при рассмотрении конкретной задачи. В данном случае, было принято решение использовать сокращение размерности, что позволило повысить точность моделей машинного обучения на 10–30 %. В результате исследователями был получен набор обученных математических моделей, способных разделять растения на здоровые, больные CMD или CBSD по результатам снятия ИК-спектров с поверхности листьев маниоки c точность классификации более 90 % для всех классов.

Еще одним примером применения алгоритмов машинного обучения к данным ИК-спектров является исследование [24]. В нем была изучена применимость данного подхода для обнаружения искусственных подсластителей по результатам спектрального анализа в инфракрасном диапазоне. Путем сочетания предобработки спектров и анализа главных компонент авторами в качестве параметров для обучения модели было выделено 131 волновое число. В качестве модели был использован алгоритм опорных векторов, позволяющий решать задачи классификации путем построения нелинейной функции, разделяющей образцы в выборке данных. Точность модели на тестовых, составляющих 25 % всей выборки, для образцов с одним видом подсластителя и для смеси двух подсластителей составила более 95 %, в то время как точность для смеси из четырех и пяти видов происходило снижение точности классификации, в некоторых смесях снижаясь до 70–73 %. Однако, несмотря на снижение точности, авторами было показано, что при практическом использовании обученной модели на образцах чайных напитков с целью обнаружения в них подсластителей, и сравнении прогнозируемых значений с результатами ВЖЭХ, эффективность работы разработанного метода подтвердилась.

Так же, в работе [25] была разработана методика определения волновых чисел, характеризующих качество оливкового масла холодного отжима, при помощи применения методов машинного обучения к данным ИК-спектроскопии отражения с преобразованием Фурье. В результате интерпретации работы моделей на основе регрессии по методу опорных векторов и регрессии по главным компонентам, обученных на всем спектральном диапазоне, было выявлено, что для оливкового масла характеристическими являются диапазоны 3005-3015 см-1, 1734-1752 см-1, а также пики в районе 2900 см-1 и 1100-1200 см-1 [25]. В перспективе, данное исследование может быть применено для разработки программного обеспечения с целью промышленной оценки качества оливковых масел.

Еще одним примером применения машинного обучения к спектральным данным выступает исследование (Kedzierski, M., Falcou-Préfol, M., Kerros, M. E., Henry, M., Pedrotti, M. L., & Bruzaud, S. (2019). A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea. *Chemosphere*, *234*, 242-251.).

## 2. Материалы и методы

### 2.1 Машинное обучение

Машинное обучение - совокупность методов искусственного интеллекта (ИИ), в основе которых лежат математические методы статистики и теории вероятностей.

Одной из основных ценностей машинного обучения является возможность работать с большим данных, которые стремительно накапливаются в современном мире, проводить их анализ, выявлять новые закономерности, которые можно увидеть лишь в масштабах выборки размером в тысячи, десятки и сотни тысяч образцов. Такие объемы данных неподвластны восприятию человеком, однако могут быть обработаны и интерпретированы при помощи современных инструментов машинного обучения.

Основной задачей машинного обучения выступает выявление определенного правила или закономерности на основании больших данных. Такое правило описывается функцией, параметры которой алгоритм подбирает в процессе обучения на выбранном исследователем наборе данных. Методы машинного обучения в общем случае можно классифицировать на три вида: обучение с учителем, обучение без учителя, и обучение с подкреплением [6].

#### 2.1.1 Обучение с учителем

Суть обучения с учителем заключается в том, что алгоритму дается набор входных параметров и набор соответствующих им ожидаемых результатов. При этом задачей является подбор оптимальной функции, позволяющей получать ожидаемые результаты с наименьшей долей ошибки. Ошибкой в методах обучения с учителем является метрика, позволяющая оценить степень различия между предсказанными результатами и ожидаемыми значениями целевой переменной из исходного набора данных.

При решении задач классификации оценить качество работы алгоритма  можно с помощью следующих методов. Первым способом, подходящим для задачи бинарной классификации, то есть классификации, где выбор состоит из двух возможных классов, является построение матрицы ошибок, содержащей информацию о доле ложно-положительных (FP - false positive), ложно-отрицательных (FN - false negative), истинно-отрицательных (TN - true negative) и истинно-положительных (TP - true positive) предсказаний. В группу ложных предсказаний относятся результаты, когда конкретный образец ложно определен к  тому или иному классу. В группе истинных предсказаний относятся случаи верного определения класса, к которому принадлежит образец. Данный метод оценки работы модели машинного обучения особенно важен для некоторых задач, например, для решения задач в медицине. Например, если алгоритм определения возможных заболеваний по результатам медицинских анализов ложно определит больного человека в группу здоровых, то это более критично, чем определение здорового человека в больные и рекомендация пройти более комплексное обследование и  дополнительную консультацию у специалиста. Из вышеописанной метрики оценки истинных и ложных предсказаний вытекает еще один метод оценки точности бинарной классификации - кривая ROC (receiver operating characteristic curve), которая отражает соотношение между специфичность и чувствительностью метода определения класса, что является критичным при проведении анализа, например, в области медицины [9]. Представим ситуацию, при которой перед исследователем стоит задача классификации больных и здоровых людей по результатам медицинских анализов. В таком случае, чувствительность, или true positive rate (TPR), отражает условную вероятность правильного определения заболевшего человека [9], и рассчитывается по формуле:

TPR = TPTP + FN,

где TP - количество истинно-положительных предсказаний, FN - количество ложно отрицательных результатов. Специфичность, или true negative rate (TNR), отражает условную вероятность верного определения здорового человека, и рассчитывается по формуле:

TNR = TNTN + FP,

где TN - количество истинно-отрицательных предсказаний, FP - количество ложно положительных предсказаний. В приведенных формулах знаменатель дроби представляет общее количество образцов, соответствующих либо классу больных, либо классу здоровых. Пример получаемого графика метрики показан на рисунке 1.

A graph of a number of black lines

Description automatically generated with medium confidence

Рис.1 Три эмпирические ROC-кривые [9].

Площадь под графиком (AUC - area under curve) ROC кривой позволяет оценить соотношение между специфичность и чувствительностью метода. показывает, что метод обладает идеальной способностью разделять больных и здоровых людей [9]. Приведенная аналогия определения здоровых и больных людей может быть представлена как пример задачи бинарной классификации. В связи с этим, метод оценки площади под графиком ROC кривой находит применение в оценке работы предсказательных алгоритмов машинного обучения.

Еще одной метрикой является общая точность классификации, выражаемая в долях или процентах и показывающая соотношение образцов, для которых был предсказан класс, к которому они относятся в исходном наборе данных. Этот способ оценки является актуальным как при решении задач бинарной классификации, так и в случае необходимости классификации образцов, которые могут относиться к множеству классов, то есть в таком случае, если набор данных описывает выборку с количеством классов больше двух.

В случае решения задач регрессии для оценки точности алгоритмов машинного обучения также используется несколько метрик:

1. Средняя абсолютная ошибка MAE – mean absolute error
2. Среднеквадратичная ошибка MSE – mean squared error
3. Коэффициент детерминации R2

где

В приведенных формулах *yi* - обозреваемое значение целевой переменной, *fi* - предсказанное значение, *n* - размер выборки. - среднее значение целевой переменной во всей представленной выборке данных.

MAE и MSE позволяют оценить величину численного отклонения прогнозируемой величины переменной от обозреваемой, и поэтому по ним не всегда можно полностью оценить, насколько точно выбранный алгоритм решает регрессионную задачу. Данные величины могут быть использованы в качестве параметров оценки доверительных интервалов, в пределах которых в предсказании присутствует разброс. Чем меньше такой интервал, тем выше точность. В то же время, коэффициент детерминации позволяет оценить, насколько хорошо модель аппроксимирует исходные данные. показывает, что модель осуществляет предсказание с идеальной точностью.

Понимание методов, по которым происходит оценка эффективности и качества работы моделей обучения с учителем, дает возможность рассмотреть некоторые из наиболее распространенных алгоритмов классического машинного обучения, существующих на данный момент.  Первый тип алгоритмов - это линейные алгоритмы, основанные на методе наименьших квадратов (МНК), в которых осуществляется подбор линейного уравнения, позволяющего по набору входных параметров и коэффициентов к ним, подбираемых в процессе обучения, получать результаты, наиболее близкие к значению целевой переменной на обучающем наборе данных. Линейные модели регрессии работают методом непосредственного предсказания целевой переменной. В случае классификации линейным алгоритмом является логистическая регрессия, целью которой является подобрать уравнение такой прямой, у которой сумма квадратов расстояний от каждой обозреваемой точки до нее минимальна, то есть построить прямую, которая стремится наиболее точно разделить два класса.

Примерами линейных алгоритмов классификации и регрессии являются  LinearRegression [27] - простая линейная регрессия по методу МНК, Ridge [29] - регрессия накладывающая ограничения на коэффициенты при построении уравнения, Lasso [30] - метод, основной тенденцией которого является сокращение количества ненулевых коэффициентов, что позволяет в некоторых случаях эффективно производить отбор параметров, влияющих на значение прогнозируемой величины , LogisticRegression [28] - алгоритм логистической регрессии.

К следующей группе относятся алгоритмы, основанные на построении деревьев принятия решения. Дерево принятия решения - это граф, в узлах которого находятся условия, в зависимости от которых так называем процесс принятия решения переходит в тот или другой узел. Условия в узлах формируются на основании обучающего набора  данных. Одним из основных преимуществ деревьев принятия решений является интерпретируемость их работы. Существуют алгоритмы как на базе одного дерева [31], обладающие чувствительностью к выборке данных, так и анасамблевые модели, которые сочетают в себе множество деревьев принятия решения, образуя случайный лес [26]. Стратегия обучения алгоритмов случайного леса заключается в обучении множества деревьев принятия решения на различных подвыборках из общей обучающей выборки, и усреднении их предсказаний для повышения общей точности.

Еще одним распространенным методом является метод опорных векторов (SVM - support vector machines) [18]. Этот тип алгоритмов работает методом построения гиперплоскости - плоскости в многоразмерном пространстве параметров обучающей выборки, которая позволяет разделить образцы и осуществить эффективную классификацию.

#### 2.1.2 Обучение без учителя

Алгоритмы обучения без учителя как правило используются с целью анализа имеющихся данных для выявления тенденций и шаблонов, характерных для образцов в исследуемой выборке. С этой целью используются алгоритмы кластеризации, например, метод K-средних [32], который осуществляет группировку набора данных в количество кластеров K, которое задается исследователем. В каждом кластере имеется центроида -  значение, являющееся средним для всех образцов внутри кластера. Алгоритм подбирает центроиды таким образом, чтобы уменьшить сумму квадратов Евклидовых расстояний внутри кластера.

Однако, оценка Евклидовых расстояний может терять свою эффективность при работе с данными высокой размерности. В данном случае, могут применяться методы снижения размерности, например, метод главных компонент (PCA - principal components analysis).

PCA – метод, позволяющий уменьшить размерность имеющейся выборки данных до набора параметров, называющихся главными компонентами, которые в свою очередь являются линейной комбинацией параметров исходных данных и объясняют наибольшую долю расхождений в исходных данных [33]. Для реализации алгоритма и анализа результатов входные данные проходят следующие преобразования:

* Стандартизацию, то есть приведение всех данных к диапазону со средним значением равным нулю и стандартным отклонением равным единице, чтобы унифицировать вклад каждого параметра в общую дисперсию и устранить влияние единиц измерения каждого параметра на дальнейшие расчеты;
* Снижение размерности за счет декомпозиции матрицы ковариации по собственным числам
* Построение графика (biplot), являющегося отображением параметров входных данных и векторов главных компонент, позволяющем увидеть направление и силу вклада каждого из параметров в эти компоненты.

PCA – мощный инструмент, позволяющий осуществлять анализ данных на предмет выявления закономерностей в имеющейся выборке, а также проводить снижение размерности с целью повышения эффективности и точности работы предиктивных алгоритмов.

Еще одной эффективным алгоритмом является многомерная нормализация данных (MDS – multidimensional scaling), который осуществляет преобразование образцов в выборке в пространство меньшей размерности, чаще двумерное или трехмерное, c наибольшим сохранением расстояний и корреляций между парами образцов в выборке [34]. Векторное представление в пространстве меньшей размерности с сохранением отношений и связей в данных упрощает их визуализацию в процессе анализа, а также может быть решением проблемы повышения точности моделей машинного обучения с учителем.

### 2.2 Метод ИК-спектроскопии с преобразованием Фурье

### 2.3 Объект исследования и оборудование

#### 2.3.1 Объект исследования

Объектом данного исследования является выборка подсолнечных и рапсовых масел, приобретенных в сетях продовольственного питания. В Таблице 1 представлено описание имеющейся выборки масел.

Таблица 1 - Описание выборки исследуемых масел.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Вид масла | Условное обозначение | Описание | Производитель |
| 1 | 2 | 3 | 4 | 5 |
| 1 | Рапсовое | Р1 | Рапсовое масло, рафинированное | Беларусь |
| 2 | Рапсовое | Р2 | Рапсовое масло Первый холодный отжим, 100%, нерафинированное | Organic Altay |
| 3 | Рапсовое | Р3 | Рапсовое масло нерафинированное | Алтайский Лен |
| 4 | Рапсовое | Р4 | Рапсовое масло рафинированное дезодорированное | Одерiха |
| 5 | Рапсовое | Р5 | Рапсовое масло рафинированное | - |
| 6 | Подсолнечное | П1 | Подсолнечное масло раф. | Золотая семечка |
| 7 | Подсолнечное | П2 | Подсолнечное масло раф. | Золотая семечка |
| 8 | Подсолнечное | П3 | Подсолнечное масло нераф. холодный отжим | Индивидуальный предприниматель |
| 9 | Подсолнечное | П4 | Подсолнечное масло нераф. холодный отжим | Маслиница |
| 10 | Подсолнечное | П5 | Подсолнечное масло раф. горячий отжим | Слобода |
| 11 | Кунжутное | Kunj1 | Кунжутное масло, нерафинированное | - |
| 12 | Кунжутное | Kunj2 | Кунжутное масло | Китай |
| 13 | Кунжутное | Kunj3 | Кунжутное масло | Китай |
| 14 | Кунжутное | Kunj3 | Кунжутное масло | Китай |
| 15 | Оливковое | Oliv1 | Оливковое масло нерафинированное | Микос |
| 16 | Оливковое | Oliv2 | Оливковое масло, рафинированное с добавлением нераф. | Окей |

#### 2.3.2 Подготовка образцов

С использованием пипетки объемом 100 мкл масла Р1-Р5 были смешаны с маслами П1-П5 в соотношениях, описанных в Таблице 2.

Таблица 2 – Описание методики пробоподготовки

|  |  |  |
| --- | --- | --- |
| № | Доля подсолнечного масла | Доля рапсового масла |
| 1 | 9 | 1 |
| 2 | 8 | 2 |
| 3 | 7 | 3 |
| 4 | 6 | 4 |
| 5 | 5 | 5 |
| 6 | 4 | 6 |
| 7 | 3 | 7 |
| 8 | 2 | 8 |
| 9 | 1 | 9 |

#### 2.3.3 Оборудование

ИК-спектры были получены на спектрометре в инфракрасном диапазоне с преобразованием Фурье «Bruker Tensor 37».

#### 2.3.4 Программное обеспечение

Для обработки полученных спектров был использован программный пакет для обработки спектральных данных *Opus*.Для анализа данных использован программный пакет *The Unscrambler* [35] – коммерческое программное обеспечение, позволяющие проводить предобработку и анализ данных, а также строить предсказательные алгоритмы машинного обучения. В качестве основного инструмента для разработки предиктивного алгоритма был использован язык программирования *Python* версии 3.11.8 [36]. *Python* – интерпретируемый язык программирования, позволяющий решать самый широкий спектр задач, от разработки графических интерфейсов и небольших веб-сайтов, до высоконагруженных приложений, работы с большими данными и проектировании систем машинного обучения и искусственного интеллекта.

*Python* обладает обширной стандартной библиотекой, а также огромным количеством сторонних пакетов, установка которых возможна при помощи менеджеров пакетов, таких как *PyPi* [38] и *Poetry* [39]. Для изменения формата спектральных данных в формат объектов языка *Python* была использована библиотека *brukeropusreader* [40]. С целью приведения данных к табличному формату, необходимому для обучения моделей машинного обучения, была использована библиотека *Pandas* [41]. В качестве основы для исследования подходящих для решения поставленной задачи моделей машинного обучения была использована библиотека *scikit-learn* [10]. *Scikit-learn* содержит в себе обширный набор методов для обработки данных, инструменты снижения размерности, кластеризации, кодирования данных, а также готовых алгоритмов, опирающихся на математические основы теории классического машинного обучения.

#### 2.3.5 Обработка данных

Была проведена векторная нормализация спектральных данных *Vector normalization*, а также корректировка базовой линии *Baseline correction* в программном пакете *Opus*.

Отсутствующие значения в спектральных данных были заполнены значением 0:

Листинг 1 – Пример заполнения пропущенных значений в выборке данных

df.fillna(0, inplace=True)

где df – объект, представляющий выборку данных, df.fillna() – метод заполнения отсутствующих значений.

#### 2.3.6 Исследовательский анализ моделей машинного обучения

Для поиска лучшей модели машинного обучения, подходящей для решения поставленной задачи классификации, был рассмотрен следующий набор алгоритмов на настройках по умолчанию:

1. Библиотека *scikit-learn*:
   1. sklearn.linear\_model.SDGClassifier – реализует метод линейной классификации с применением техники стохастического градиентного спуска для минимизации ошибки, по умолчанию реализует алгоритм классификации на основе метода опорных векторов;
   2. sklearn.neural\_network.MLPClassifier – модель многослойного перцептрона, является примером полносвязных нейронных сетей, реализует стохастический градиентный спуск для уменьшения ошибки в процессе обучения;
   3. sklearn.neighbours.KNeighboursClassifier – алгоритм классификации по методу k-ближайших соседей. Решение о принадлежности образца к конкретному классу принимается на основании информации о классах k-ближайший по Евклидову расстоянию к данному образцов;
   4. sklearn.ensemble.BaggingClassifier – алгоритм, обучающий множество базовых моделей, таких как, например, дерево принятия решения, на разных подвыборках данных, и принимающий итоговое решение методом усреднения результатов предсказания каждой из базовых моделей либо методом голосования – выбором того класса, к которому образец определило большинство моделей;
   5. sklearn.ensemble.RandomForestClassifier – ансамблевая модель на основании множества деревьев принятия решения, определяющая образец к конкретному классу путем усреднения предсказаний каждой из составляющих его моделей;
   6. sklearn.ensemble.GradientBoostingClassifier – алгоритм на основании деревьев принятия решения, обучение которых осуществляется в направлении спуска по градиенту функции ошибки;
   7. sklearn.svm.SVC – модель классификации на основе метода опорных векторов;
   8. sklearn.svm.LinearSVC – так же как и SVC, основывается на методе опорных векторов. Ключевое отличие заключается в применении линейного ядра.
2. Библиотека *lighgbm*[42]:
   1. lightgbm.LGBMClassifier – модель на основе деревьев принятия решения с применением градиентного спуска в процессе обучения. Отличительной чертой данной реализации является более эффективное использование памяти вычислительной машины, повышенная скорость и оптимизация обучения.

Поскольку была проведена нормализация спектральных данных, была выдвинута гипотеза о том, что этап стандартизации данных, являющийся необходимым при подготовке данных к обучению модели, может быть опущен. Для проверки выдвинутого предположения для каждой из описанных выше моделей была произведена следующая предобработка данных:

1. NoScaler - отсутствие предобработки;
2. sklearn.preprocessing.StandardScaler – преобразует данные в выборке таким образом, что в пределах одного параметра данные имеют среднее значение 0 и стандартное отклонение 1;
3. sklearn.preprocessing.MinMaxScaler – приводит данные в рамках каждого параметра к диапазону от 0 до 1;
4. sklearn.preprocessing.RobustScaler – алгоритм, реализующий стандартизацию каждого параметра на основании его первого и третьего квантилей в выборке;

В таблице 3 представлены результаты десяти лучших сочетаний стратегии предобработки данных и алгоритма машинного обучения.

Таблица 3 – результаты первичного отбора моделей и стратегии предобработки данных

|  |  |  |  |
| --- | --- | --- | --- |
| № | Модель | Стратегия предобработки | Точность |
| 1 | RandomForestClassifier | NoScaler | 0,695652174 |
| 2 | RandomForestClassifier | MinMaxScaler | 0,666666667 |
| 3 | RandomForestClassifier | RobustScaler | 0,666666667 |
| 4 | GradientBoostingClassifier | StandardScaler | 0,637681159 |
| 5 | RandomForestClassifier | StandardScaler | 0,623188406 |
| 6 | LGBMClassifier | RobustScaler | 0,623188406 |
| 7 | BaggingClassifier | StandardScaler | 0,608695652 |
| 8 | LGBMClassifier | NoScaler | 0,594202899 |
| 9 | GradientBoostingClassifier | MinMaxScaler | 0,594202899 |
| 10 | GradientBoostingClassifier | RobustScaler | 0,594202899 |

Приведенная таблица показывает, что лучшей моделью оказался RandomForestClassifier без предобработки данных, точность которого составляет 0,696. Также данная модель занимает три лучшие позиции по точности, из чего можно сделать предположение о том, что дальнейшая оптимизация позволит добиться еще большей точности классификации.

#### 2.3.7 Оптимизация параметров выбранной модели машинного обучения

## Список использованных источников

1. Ильина, Г. Г., Ламоткин, С. А., Колногоров, К. П., & Скаковский, Е. Д. (2014). Идентификация состава растительных масел хроматографическими и спектральными методами. *Труды БГТУ. Серия 2: Химические технологии, биотехнология, геоэкология*, (4 (168)), 207-210.
2. Зверева, И. С., Денисов, И. С., Зубчонок, Н. В., & Ермолаева, Т. Н. (2019). ИЗУЧЕНИЕ КАЧЕСТВА РАСТИТЕЛЬНЫХ МАСЕЛ МЕТОДАМИ ИК-СПЕКТРОМЕТРИИ И ГАЗОВОЙ ХРОМАТОГРАФИИ С МАСС-СПЕКТРОМЕТРИЧЕСКИМ ДЕТЕКТОРОМ. In *XXIX Российская молодежная научная конференция «Проблемы теоретической и экспериментальной химии».—Екатеринбург, 2019* (pp. 94-94). Издательство Уральского университета.
3. Дейнека, В. И., Дейнека, Л. А., & Сорокопудов, В. Н. (2009). Обращенно-фазовая ВЭЖХ в анализе растительных масел. Метод контроля подлинности и установления фальсификации облепихового масла. *Химико-фармацевтический журнал*, *43*(1), 33-36.
4. da Silveira, R., Vágula, J. M., de Lima Figueiredo, I., Claus, T., Galuch, M. B., Junior, O. O. S., & Visentainer, J. V. (2017). Rapid methodology via mass spectrometry to quantify addition of soybean oil in extra virgin olive oil: A comparison with traditional methods adopted by food industry to identify fraud. *Food Research International*, *102*, 43-50.
5. Alexa, E., Dragomirescu, A., Pop, G., Jianu, C., & Dragos, D. (2009). The use of FT-IR spectroscopy in the identification of vegetable oils adulteration. *J. Food Agric. Environ*, *7*(2), 20-24.
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547-555.
7. Kireeva, N. et al. Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Mol. Inform*. 31, 301–312 (2012).
8. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model*. 50, 1189–1204 (2010).
9. Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics*, *48*, 277-287.
10. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python JMLR 12, pp. 2825-2830, 2011.
11. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>
12. Townsend, J., Micucci, C.P., Hymel, J.H. *et al.* Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat Commun* **11**, 3230 (2020). <https://doi.org/10.1038/s41467-020-17035-5>
13. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.
14. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, *18*(6), 463-477.
15. Shirokii, N., Din, Y., Petrov, I., Seregin, Y., Sirotenko, S., Razlivina, J., ... & Vinogradov, V. (2023). Quantitative prediction of inorganic nanomaterial cellular toxicity via machine learning. *Small*, *19*(19), 2207106.
16. Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., & Tylkowski, B. (2022). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(2), e1568.
17. Priya, S., Tripathi, G., Singh, D. B., Jain, P., & Kumar, A. (2022). Machine learning approaches and their applications in drug discovery and design. *Chemical Biology & Drug Design*, *100*(1), 136-153.
18. Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., ... & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome medicine*, *6*, 1-18.
19. Owomugisha, G., Melchert, F., Mwebaze, E., Quinn, J. A., & Biehl, M. (2018). Machine learning for diagnosis of disease in plants using spectral data. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)* (pp. 9-15). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
20. Soyemi, A., & Szilvási, T. (2023). Calculated physicochemical properties of glycerol-derived solvents to drive plastic waste recycling. *Industrial & Engineering Chemistry Research*, *62*(15), 6322-6337.
21. Hansen, C. M. (1967). The three dimensional solubility parameter. *Danish Technical: Copenhagen*, *14*.
22. Rouillard, A. D., Hurle, M. R., & Agarwal, P. (2018). Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets. *PLoS Computational Biology*, *14*(5), e1006142.
23. Uke, A., Tokunaga, H., Utsumi, Y., Vu, N. A., Nhan, P. T., Srean, P., ... & Ugaki, M. (2022). Cassava mosaic disease and its management in Southeast Asia. *Plant Molecular Biology*, 1-11.
24. Wang, Y. T., Li, B., Xu, X. J., Ren, H. B., Yin, J. Y., Zhu, H., & Zhang, Y. H. (2020). FTIR spectroscopy coupled with machine learning approaches as a rapid tool for identification and quantification of artificial sweeteners. *Food chemistry*, *303*, 125404.
25. Scatigno, C., & Festa, G. (2022). FTIR coupled with machine learning to unveil spectroscopic benchmarks in the Italian EVOO. International Journal of Food Science & Technology, 57(7), 4156-4162.
26. Chen, K., Wang, Y., Lang, Y. *et al.* Machine learning models to predict submucosal invasion in early gastric cancer based on endoscopy features and standardized color metrics. *Sci Rep* **14**, 10445 (2024). <https://doi.org/10.1038/s41598-024-61258-1>
27. Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 275-294.
28. Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Topics in biostatistics*, 273-301.
29. McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93-100.
30. Ranstam, J., & Cook, J. A. (2018). LASSO regression. *Journal of British Surgery*, *105*(10), 1348-1348.
31. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
32. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.
33. Greenacre, M., Groenen, P. J., Hastie, T., d’Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100.
34. Pawliczek, P., & Dzwinel, W. (2013). Interactive data mining by using multidimensional scaling. *Procedia Computer Science*, *18*, 40-49.
35. Bruker Opus Software [Электронный ресурс] // URL: <https://www.bruker.com/en/products-and-solutions/infrared-and-raman/opus-spectroscopy-software.html> (дата обращения 14.05.2024)
36. The Unscrambler [Электронный ресурс] // URL: <https://www.aspentech.com/en/products/apm/aspen-unscrambler> (дата обращения 14.05.2024)
37. Python 3.11.8 Documentation [Электронный ресурс] // URL: <https://docs.python.org/release/3.11.8/library/index.html> (дата обращения 14.05.2024)
38. PyPi [Электронный ресурс] // URL: <https://pypi.org/> (дата обращения 14.05.2024)
39. Poetry Documentation [Электронный ресурс] // URL: <https://python-poetry.org/> (дата обращения 14.05.2024)
40. Brukeropusreader [Электронный ресурс] // URL:
41. [https://github.com/qedsoftware/brukeropusreader (дата обращения 14.05.2024)](https://github.com/qedsoftware/brukeropusreader )
42. Pandas Documentation [Электронный ресурс] // URL: <https://pandas.pydata.org/> (дата обращения 14.05.2024)
43. LightGBM Documentaion [Электронный ресурс] // URL: <https://lightgbm.readthedocs.io/en/stable/> (дата обращения 15.05.2024)