|  |
| --- |
| НИУ ВШЭ |
| “Методы кластеризации медицинских данных” |
| Дипломная работа |
|  |
| **Воробьёва Мария Николаевна** |
| **23.06.2023** |

|  |
| --- |
|  |

# Введение

Заболевания щитовидной железы являются одними из наиболее распространенных эндокринных нарушений, влияющих на здоровье и качество жизни миллионов людей по всему миру. Щитовидная железа играет ключевую роль в регуляции обмена веществ, энергетического баланса и функционирования организма в целом. Недостаточная или избыточная активность щитовидной железы может привести к различным заболеваниям, включая гипотиреоз, гипертиреоз и различные формы узлового образования.

Для диагностики и обнаружения заболеваний щитовидной железы врачи обычно применяют различные методы, включая клинический осмотр, анализы крови на уровень гормонов щитовидной железы, ультразвуковое и радионуклидное исследование. Однако эти методы имеют свои ограничения, и иногда может быть сложно достичь высокой точности диагностики, особенно при обнаружении ранних стадий заболевания или при оценке сложных патологических процессов.

В последние годы машинное обучение и искусственный интеллект значительно прогрессировали, предоставляя новые возможности для развития инновационных подходов к диагностике и обнаружению заболеваний. Методы машинного обучения позволяют компьютерным системам автоматически извлекать полезную информацию из множества медицинских данных, включая клинические показатели, результаты обследований и изображения. Это открывает новые перспективы для повышения эффективности и точности диагностики заболеваний щитовидной железы.

Целью данной научной работы является исследование и оценка применения методов кластеризации и классификации для обнаружения заболевания щитовидной железы. Оценка качества работы моделей будет выполнена на основе данных Института медицинских исследований Гарвана. Кроме того, мы проведем обзор существующих исследований и работ, посвященных этой теме, и проанализируем их результаты и достижения.

Ожидается, что результаты данной работы смогут способствовать развитию более точных и эффективных инструментов для обнаружения заболевания щитовидной железы, что в свою очередь позволит раннюю диагностику и своевременное лечение пациентов, улучшая их прогноз и качество жизни.

# 

# Обзор литературы

В данном разделе представлен обзор актуальной литературы, посвященной проблеме детекции аномалий в клинических данных. Исследования, проведенные в этой области, направлены на разработку методов и алгоритмов, способных автоматически обнаруживать аномалии в медицинских данных, таких как данные о пациентах, результаты лабораторных анализов, изображения и прочее.

В литературе существует множество методов, применяемых для детекции аномалий в клинических данных, среди которых:

* **Статистические методы**: Эти методы основаны на моделях распределения данных и статистических метриках. Например, одним из таких методов является использование Z-оценки или стандартного отклонения для определения аномалий. Это классические и относительно простые методы, однако несмотря на это некоторые современные работы показывают преимущество этих методов, например, в работе [7] показано как применение статистических методов удаления аномалий улучшает качество классификации.
* **Методы машинного обучения**: Множество алгоритмов машинного обучения применяются для детекции аномалий, таких как метод опорных векторов (Support Vector Machines), случайный лес (Random Forest), нейронные сети и другие. Эти методы основаны на обучении моделей на нормальных данных и последующем обнаружении отклонений. В работе [6] показано, что после обучения случайного леса можно применять его для классификации новых данных на признаку аномальности. Для этого новые данные подвергаются оценке каждым деревом случайного леса. Если данные отклоняются от нормального распределения и попадают в область, которая слабо представлена в обучающем наборе данных, модель может рассматривать их как аномалии.
* **Глубокое обучение**: Современные алгоритмы глубокого обучения, такие как сверточные нейронные сети и рекуррентные нейронные сети, демонстрируют высокую эффективность в области детекции аномалий. Они способны автоматически извлекать сложные признаки из данных и обнаруживать аномалии на основе этих признаков. Например, в работе [5] изучается применение генеративно-состязательной сети и состязательного автоэнкодера для детекции аномальных значений. После обучения такую сеть можно использовать для генерации новых данных. При этом аномальные данные, которые сильно отличаются от обучающих данных, могут быть классифицированы как аномалии, поскольку дискриминатор будет иметь трудности в их классификации как реальные данные.

# Содержание работы

## Изучение и подготовка данных

Для получения качественных моделей методы машинного обучения необходимо требуют большой набор данных. В качестве набора данных мы используем датасет [1], который содержит 6 медицинских баз данных. В сумме датасет содержит 2800 обучающих и 972 тестовых примеров, каждый пример описан 29 признаками, среди которых есть как категориальные, так и числовые. Датасет содержит множество пропусков и аномальных значений. При использовании методов машинного обучения важно не только размер датасета, но и его качество. Таким образом, первый этап в при построении модели машинного обучения это подготовка данных, которая включает в себя следующие шаги:

### Валидация типов данных

Валидация типов данных (англ. data type validation) относится к процессу проверки соответствия значений данных определенным типам данных или форматам. Она выполняется для обнаружения и исправления ошибок, связанных с неправильным типом данных, которые могут возникнуть при сборе, вводе или обработке данных.

Каждый тип данных имеет определенные правила и ограничения, которые определяют, какие значения могут быть представлены в этом типе. Например, целочисленный тип данных может принимать только целочисленные значения, а строковый тип данных может содержать символы и текст. Валидация типов данных позволяет проверить, что значения данных соответствуют ожидаемому типу и наличествующим ограничениям.

### Удаление дубликатов

Удаление дубликатов из датасета является важным шагом предобработки данных в машинном обучении. Вот несколько причин, почему это важно:

* **Повышение точности моделей**: Наличие дубликатов в наборе данных может исказить процесс обучения и привести к завышенной точности моделей. Если одно и то же наблюдение или точка данных присутствует несколько раз, модель может считать, что эта точка имеет большее значение или вес, и, таким образом, производить неправильные или завышенные прогнозы. Удаление дубликатов позволяет избежать искажений и получить более точные результаты.
* **Повышение надежности и обобщаемости моделей**: Дубликаты могут привести к переобучению моделей, когда модель "запоминает" или "подстраивается" под конкретные наблюдения вместо извлечения общих закономерностей из данных. Удаление дубликатов помогает модели обобщить данные, улучшая ее способность делать точные прогнозы на новых, ранее не виденных данных.
* **Экономия вычислительных ресурсов**: Наличие дубликатов увеличивает объем данных, которые нужно обработать и анализировать. Это может замедлить процесс обучения модели и увеличить потребление вычислительных ресурсов. Удаление дубликатов позволяет сократить объем данных и сэкономить ресурсы.
* **Предотвращение ошибок и искажений**: Дубликаты могут привести к неправильным статистическим оценкам и выводам, а также искажать анализ данных. Удаление дубликатов помогает сохранить точность и достоверность данных, предотвращая ошибки и искажения в результатах исследования.

В целом, удаление дубликатов из датасета способствует повышению качества и достоверности данных, точности моделей и их способности обобщать новые данные. Это помогает избежать искажений, ошибок и проблем, связанных с дублированными наблюдениями, и обеспечивает более надежный и эффективный процесс машинного обучения.

### Удаление аномалий

Обнаружения и удаления аномальных или необычных точек данных из набора данных. Аномалии представляют собой наблюдения или значения, которые существенно отличаются от ожидаемого поведения или нормы в данном контексте. В рамках данной работы было изучено и реализовано три метода удаления аномалий: Метод IQR, метод 3 сигм, метод локального уровня выброса.

**Метод IQR** (Interquartile Range, межквартильный размах) — это статистический метод, используемый для определения наличия выбросов в наборе данных. IQR представляет собой мера разброса данных и вычисляется как разность между третьим квартилем (Q3) и первым квартилем (Q1) в упорядоченном наборе данных. Точнее:

### где - межквартильных размах.

**Метод 3 сигм**, также известен как правило 68-95-99,7 или эмпирическое правило, представляет собой статистический метод обнаружения аномалий, предполагающий нормальное распределение данных. Он идентифицирует аномалии как точки данных, выходящие за пределы трех стандартных отклонений от среднего значения. Точнее:

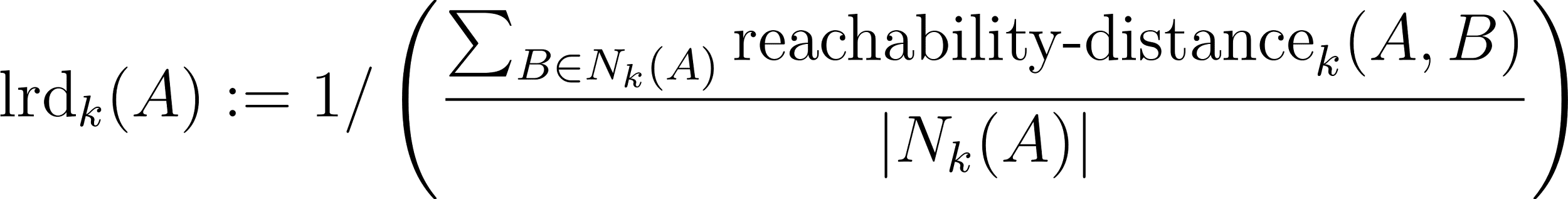
где и - средние и стандратное отклонение распределения случайной величины .

**Метод локального уровеня выброса** — алгоритм нахождения аномальных точек данных путём измерения локального отклонения данной точки с учётом её соседей

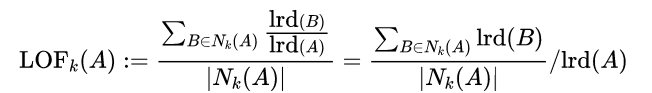
Локальный уровень выброса основывается на концепции локальной плотности, где локальность задаётся k -ближайшими соседями, расстояния до которых используются для оценки плотности. Путём сравнения локальной плотности объекта с локальной плотностью его соседей можно выделить области с аналогичной плотностью и точки, которые имеют существенно меньшую плотность, чем её соседи. Эти точки считаются выбросами. Формально: пусть k-distance(*A*) является расстоянием от объекта *A* до *k*-ого ближайшего соседа. Мы обозначаем множество *k* ближайших соседей как . Определим оператор достижимого расстояния (англ. reachability-distance):



Тогда локальная плотность достижимости объекта A определяется как обратное значение среднему расстоянию достижимости объекта *A* из его соседей:

[](https://www.codecogs.com/eqnedit.php?latex=%7B%5Cmbox%7Blrd%7D%7D_%7Bk%7D(A)%3A%3D1%2F%5Cleft(%7B%5Cfrac%20%7B%5Csum%20_%7BB%5Cin%20N_%7Bk%7D(A)%7D%7B%5Cmbox%7Breachability-distance%7D%7D_%7Bk%7D(A%2CB)%7D%7B%7CN_%7Bk%7D(A)%7C%7D%7D%5Cright)#0)

Локальные плотности достижимости затем сравниваются с локальными плотностями достижимости соседей:



Значение, примерно равное 1 , означает, что объект сравним с его соседями (а тогда он не является выбросом). Значение меньше 1 означает плотную область (которая может быть внутренностью), а значения, существенно большие 1 , свидетельствуют о выбросах.

В результате эксперимента было определено, что наилучшие решение достигается при использовании IQR метода.

### Расчет описательных статистик и визуализация.

Расчет описательных статистик позволяет сформировать первичное представлении о данных, а визуализация дополнительно облегчает восприятие описательных статистик и дает возможность увидеть закономерности среди признаков.

### Балансировка данных

На практике не всегда удается представить все классы примеров в равном соотношении. Это ведет к дисбалансу данных, который негативно сказывается на качестве обучения модели. Oversampling состоит в увеличении числа образцов минорного класса (менее представленного класса) путем создания дубликатов или генерации новых образцов.

### Стандартизация и декоррелирование данных

Cтандартизация и декоррелирование данных в машинном обучении - это две важные техники предобработки данных, которые применяются для улучшения процесса обучения моделей и повышения их эффективности.

Стандартизация (англ. Standardization), также известная как нормализация или масштабирование, относится к процессу приведения значений признаков в наборе данных к одному и тому же масштабу или диапазону. Это осуществляется путем вычитания среднего значения признака и деления на его стандартное отклонение. Результатом стандартизации являются признаки, которые имеют среднее значение равное 0 и стандартное отклонение равное 1. Это позволяет моделированию и алгоритмам машинного обучения более эффективно работать с данными, особенно когда признаки имеют различные шкалы или единицы измерения. Стандартизация также помогает предотвратить проблему смещения весов модели, когда некоторые признаки имеют более широкий диапазон значений, чем другие.

Декоррелирование данных (англ. Decorrelation) относится к процессу устранения или уменьшения линейной зависимости между признаками в наборе данных. Корреляция между признаками может быть проблемой при обучении моделей, особенно для алгоритмов, которые предполагают некоррелированные признаки или могут быть чувствительны к мультиколлинеарности. Декоррелирование помогает устранить избыточность информации и улучшить стабильность и точность моделей. Одним из популярных методов декоррелирования данных является применение метода главных компонент (PCA), который выполняет преобразование признакового пространства, преобразуя их в новые некоррелированные признаки, называемые главными компонентами.

## Решение задачи кластеризации

Одним из ключевых методов анализа данных является кластеризация, которая позволяет группировать объекты схожих характеристик внутри больших наборов данных. Алгоритмы кластеризации находят широкое применение в различных областях, таких как биология, медицина, экономика, социальные науки и многие другие. В нашей работе будет осуществлен анализ клинических данных пациентов, таких как пол, возраст, результаты анализа крови, прием медикаментов и других релевантных показателей, с целью выявления скрытых закономерностей и типов заболеваний.

Для достижения поставленной цели будут применены различные алгоритмы кластеризации – метод k-средних, иерархическая кластеризация, алгоритм сдвига среднего значения и алгоритм DBSCAN. Каждый алгоритм будет подвержен поиску оптимальных гиперпараметров и аналзу преимуществ и недостатков применимости для данного медицинского контекста.

### Оценка качества решения задачи кластеризации

В качестве метрики качества кластеризации будет использован коэффициент силуэта. Коэффициент силуэта (англ. silhouette score) является одной из метрик, используемых для оценки качества кластеризации данных. Он предоставляет числовую оценку насколько объекты внутри одного кластера схожи между собой и насколько они отличаются от объектов в других кластерах. Чем выше значение Silhouette score, тем более качественная и однородная кластеризация.

Для вычисления Silhouette score для каждого объекта из набора данных выполняются следующие шаги:

1. Рассчитывается среднее расстояние между объектом и всеми другими объектами в том же кластере. Это называется "a" (average intra-cluster distance).
2. Для каждого кластера рассчитывается среднее расстояние между объектом и всеми объектами из других кластеров. Это называется "b" (average inter-cluster distance).
3. Вычисляется Silhouette score для данного объекта как (b - a) / max(a, b).
4. Повторяются шаги 1-3 для всех объектов в наборе данных.

Значение Silhouette score находится в диапазоне от -1 до 1. Значение ближе к 1 указывает на хорошую кластеризацию, где объекты внутри кластеров очень похожи между собой, а объекты из разных кластеров значительно отличаются. Значение ближе к -1 указывает на плохую кластеризацию, когда объекты внутри кластеров похожи на объекты из других кластеров, что может указывать на перекрытие или неправильное определение границ кластеров.

### Метод k-средних

Основная идея метода k-средних заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

### Метод иерархической кластеризации

Метод иерархической кластеризации с агломеративным подходом [4] является одним из наиболее распространенных методов в области анализа данных и кластеризации. Этот метод основывается на идее постепенного объединения близких объектов для создания иерархической структуры кластеров.

Агломеративный метод начинает с того, что каждый объект в данных рассматривается как отдельный кластер. Затем на каждой итерации наиболее близкие кластеры объединяются, пока не будет получен окончательный иерархический набор кластеров. В ходе этого процесса обычно используется мера расстояния, такая как евклидово расстояние или мера близости, для определения того, насколько близки или различны объекты или кластеры.

Агломеративный метод может быть реализован двумя основными стратегиями:

* Стратегия объединения ближайших объектов (Single Linkage): На каждой итерации выбираются два наиболее близких объекта или кластера и объединяются в один кластер.
* Стратегия объединения наиболее удаленных объектов (Complete Linkage): На каждой итерации выбираются два наиболее удаленных объекта или кластера и объединяются в один кластер.

При использовании агломеративного метода создается дендрограмма, которая представляет иерархическую структуру кластеров. Дендрограмма позволяет визуализировать процесс объединения кластеров и определить оптимальное число кластеров путем анализа длины разветвлений на графике.

Преимуществом метода иерархической кластеризации с агломеративным подходом является его способность выявлять иерархическую структуру данных, что позволяет исследователям анализировать данные на различных уровнях детализации. Однако этот метод может быть вычислительно сложным для больших наборов данных, так как требуется вычисление всех возможных расстояний между объектами или кластерами.

### Алгоритм сдвига среднего значения

Алгоритм сдвига среднего значения (Shifted Mean Algorithm) [3] является одним из алгоритмов кластеризации, который используется для выделения компактных и однородных кластеров в данных.

Основная идея алгоритма заключается в следующем:

1. **Инициализация**: Каждый объект данных считается отдельным кластером. Изначально центр каждого кластера устанавливается в среднее значение его объектов.
2. **Шаг сдвига**: Повторяются следующие шаги до сходимости:
   1. Для каждого объекта данных вычисляется расстояние до центра каждого кластера.
   2. Объект присваивается к кластеру с наименьшим расстоянием.
   3. Для каждого кластера пересчитывается центр, который устанавливается в новое среднее значение всех объектов этого кластера.
3. **Критерий сходимости**: Алгоритм продолжает итерации до тех пор, пока изменения центров кластеров становятся незначительными или достигнут заданный предел итераций.

Алгоритм сдвига среднего значения эффективен при кластеризации данных, особенно в случаях, когда кластеры имеют форму сферы или окружности, и когда объекты внутри кластеров плотно сгруппированы вокруг центра. Этот алгоритм также устойчив к выбросам и шуму в данных, так как сдвиг среднего значения помогает смягчить их влияние.

Однако алгоритм сдвига среднего значения имеет некоторые ограничения, включая зависимость от начальных значений центров кластеров и чувствительность к выбору метрики расстояния. Также он может неэффективно работать с большими объемами данных, поскольку требует вычисления расстояний для каждого объекта на каждой итерации. Поэтому выбор подходящей метрики расстояния и тщательное подбор начальных значений центров кластеров имеют важное значение для успешного применения этого алгоритма.

### Алгоритм DBSCAN

Алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [2] является одним из популярных алгоритмов класстеризации, который определяет кластеры на основе плотности данных в пространстве.

Основная идея алгоритма DBSCAN заключается в следующем:

1. **Определение ядерных объектов**: Алгоритм начинает с выбора случайного нерассмотренного объекта данных. Если этот объект имеет достаточное количество соседей в заданном радиусе (eps), то он считается ядерным объектом.
2. **Формирование кластеров**: Для каждого ядерного объекта создается кластер. Затем алгоритм расширяет этот кластер, добавляя к нему все достижимые объекты, которые находятся в радиусе (eps) от ядерного объекта. Также объекты, которые являются достижимыми для других объектов, присоединяются к соответствующим кластерам.
3. **Обработка шума**: Объекты, которые не являются ядерными и недостижимы для других объектов, считаются шумом и не присоединяются ни к одному кластеру.

Алгоритм DBSCAN основывается на двух основных параметрах:

* **Epsilon (eps)**: Радиус окрестности, в пределах которого объекты считаются соседями.
* **MinPts**: Минимальное количество соседей, которые должны находиться в радиусе (eps), чтобы объект считался ядерным.

Преимущества алгоритма DBSCAN включают его способность обнаруживать кластеры произвольной формы и обрабатывать шум в данных. Он также не требует заранее заданного числа кластеров. Кроме того, алгоритм DBSCAN может эффективно работать с большими объемами данных.

Однако алгоритм DBSCAN имеет некоторые ограничения, включая зависимость от настройки параметров eps и MinPts, что может потребовать некоторой экспертизы в выборе этих значений. Также алгоритм может столкнуться с проблемами в случае, когда в данных присутствуют кластеры различной плотности или когда кластеры имеют значительные различия в плотности.

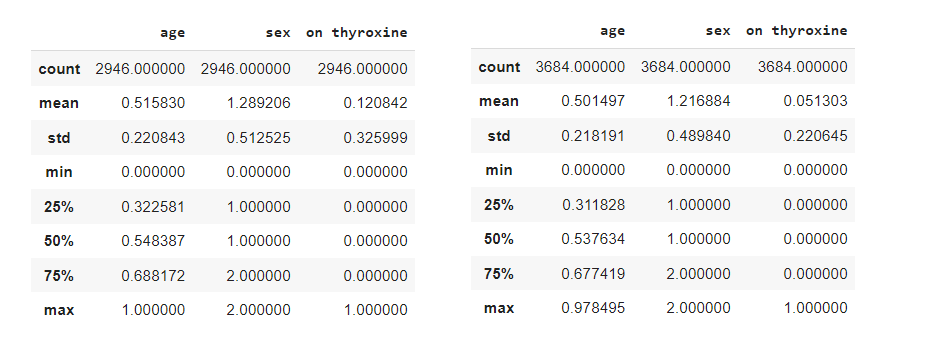
### Результаты экспериментов

После подбора гиперпараметров для каждого алгоритма были получены следующие результаты:

|  |  |
| --- | --- |
| Название метода | Значение коэффициента силуэта |
| k-средних | 0.43 |
| Иерархическая кластеризация | 0.42 |
| Сдвига среднего значения | 0.41 |
| Алгоритм DBSCAN | 0.14 |

### Интерпретация кластеров

Рассмотрим описательные статистики некоторых признаков, которые можно интерпретировать, для кластеров, полученные лучшим алгоритмом.



Слева представлены описательные статистики для примеров, который метод k-средних определил первым кластером, справа – вторым кластером. Как можно видеть средние значение признака “on thyroxine” более чем в два раза больше у первого кластера, чем у второго. Т.е. прием тироксина (thyroxine) оказывает значительное влияние на кластеризацию.

Основной функцией тироксина является активация процессов метаболизма, которая осуществляется через стимуляцию синтеза РНК и соответствующих белков. Тироксин влияет на обмен веществ, повышает температуру тела, контролирует рост и развитие организма, увеличивает синтез белков и чувствительность к катехоламинам, увеличивает частоту сердечных сокращений, утолщает слизистую оболочку матки. Усиливает окислительные процессы в клетках всего организма, в частности и клетках мозга. Тироксин важен для надлежащего развития и дифференцировки всех клеток человеческого тела, также может стимулировать метаболизм витаминов.

Таким образом, метод k-средних разделяет примеры на кластеры, которые можно проинтерпретировать с медицинской точки зрения.

# Список литературы

[1] Quinlan,Ross. (1987). Thyroid Disease. UCI Machine Learning Repository. <https://doi.org/10.24432/C5D010>.

[2] Ester, M., H. P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996

[3] Dorin Comaniciu and Peter Meer, “Mean Shift: A robust approach toward feature space analysis”. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. pp. 603-619.

[4] Nielsen, Frank (2016). "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.

[5] Lim, Swee Kiat et al. “DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN.” 2018 IEEE International Conference on Data Mining (ICDM) (2018): 1122-1127.

[6] R. Primartha and B. A. Tama, "Anomaly detection using random forest: A performance revisited," 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Indonesia, 2017, pp. 1-6, doi: 10.1109/ICODSE.2017.8285847.

[7] S. Garg, K. Kaur, N. Kumar, S. Batra and M. S. Obaidat, "HyClass: Hybrid Classification Model for Anomaly Detection in Cloud Environment," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 2018, pp. 1-7, doi: 10.1109/ICC.2018.8422481.