Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 10.

СНИЖЕНИЕ РАЗМЕРНОСТИ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

**1.Цели и задачи**

Цель лабораторной работы: научиться использовать обучаемые алгоритмы снижения размерности. Основные задачи: – получение навыков рефакторинга кода в проектах машинного обучения; – получение навыков по использованию обучаемых алгоритмов снимжения размерности; – получение навыков по аргументрированному выбору и анализу алгоритмов снижения размерности.

1. **Теоретическое обоснование**
   1. Необходимость снижения размерности Считается, что модель стоит обучать на как можно большем объеме разнообразных данных. Но иногда, если информации становится слишком много, она тормозит или останавливает процесс обучения. Чтобы модель работала нормально, нужно уменьшить количество признаков, выбрав самые эффективные. Для этого применяют методы снижения размерности. Уменьшение размерности данных – это подход упрощения сложных наборов данных для облегчения их обработки. По мере того как данные растут и становятся более сложными, извлекать информацию становится все труднее, а визуализация становится более накладной. Методы уменьшения размерности данных решают эту проблему, предоставляя меньшее количество измерений (столбцов) при сохранении наиболее важной информации. Мы можем потерять некоторые детали, но получить более простое представление данных, которое легче обрабатывать и сравнивать. Уменьшение размерности широко используется в области машинного обучения и анализа данных. Его цель состоит в том, чтобы упростить обработку данных за счет уменьшения количества объектов в наборе при сохранении ключевой информации. Когда мы сталкиваемся с данными большой размерности, ее уменьшение может помочь нам снизить вычислительную сложность, повысить производительность и результативность модели.

2.2 Термин «проклятие размерности» Обучая модели, нужно предоставить им на входе переменные (например, цвет и стоимость товара). Их количество называют размерностью. Термин «проклятие размерности» в 1961 году ввел американский математик Ричард Беллман. Он описал трудности при обработке датасетов с большим количеством параметров: возрастающую сложность вычислений, необходимость большого объема памяти для хранения датасетов, рост шума и проблму переобучения искусственных нейронных сетей. Беллман объяснил рост размерности системы на примере единичного интервала [0,1]. 100 точек способны заполнить этот интервал при частоте от 0,01. Но если мы создадим модель 10-мерного куба – нам потребуется уже 1020 точек, то есть в 1018 раз больше. Беллман видел решение этой проблемы в снижении размерности пространства и переносе данных на пространства с меньшей размерностью. При кластеризации большое количество переменных приводит к тому, что точки данных выглядят равноудаленными друг от друга. Это еще одно негативное следствие проклятия размерности. Снижение размерности – это метод подготовки данных перед обучением модели. Оно может быть выполнено после очистки и масштабирования данных. Модель, обученная на данных с широким набором признаков, подвергается риску переобучения. Это приводит к снижению точности. 85 Борьба с переобучением – основная цель уменьшения размерности. Чем меньше предположений делает модель, тем она проще. 2.3 Методы снижения размерности Для снижения размерности используют методы выбора и проектирования переменных. Выбор переменных – это самый простой способ уменьшения размерности. Проектирование – создание новых переменных на основе существующих путем их преобразования. Например, нам нужна модель, которая прогнозирует средний чек. У нас есть массив данных, который подробно описывает каждого покупателя. Столбец с цветом глаз не помог бы предсказать чек. В отличие от информации о доходе покупателей. Метод порога отклонения (Variance Threshold) – простой способ выбора переменных. Он отбрасывает все признаки, в которых дисперсия не превышает заданного порогового значения. Порог выбирается с учетом данных. Маленькая дисперсия часто встречается у признаков, в которых значение многих строк совпадает (например, признак «миллиардер», скорее всего, будет бесполезен). Метод одномерного выбора признаков (Univariate Feature Selection) применяет статистические тесты. Он называется одномерным, потому что анализирует переменные по очереди, сравнивая их с целевым показателем. Признаки, которые слабо коррелируют с ним, отбрасываются. Самые распространенные методы проектирования переменных используют линейные преобразования. Например, PCA (Principal Component Analysis). Его применяют для уменьшения размерности континуальных данных (температура, вес, уровень гемоглобина в крови). Он позволяет снизить потерю информации при уменьшении размерности. PCA применяется, если некоторые признаки коррелируют друг с другом. Он создает из них независимые линейные комбинации. Часто дисперсии получившихся признаков сильно отличаются. В таком случае признаки с 86 низкой дисперсией можно отбросить, потому что они менее информативны. Этот метод решает две задачи: создает более информативные признаки и уменьшает шум в данных. Факторный анализ (Factor analysis), как и PCA, создает новые признаки, комбинируя существующие. Его отличие в том, что FA предполагает наличие реальных скрытых факторов, которые определяют наблюдаемые. Кроме того, он объясняет часть дисперсии наблюдаемых параметров ошибками измерения, а часть – реальной дисперсией скрытых факторов. Данные должны быть количественными, измеряться по интегральной шкале или шкале отношений. Категориальные данные (религия, место рождения, пол) не подходят для этого метода. Например, когда человек выбирает мебель, его решение основано на ряде факторов: его возрасте, доходе, площади квартиры. Это скрытые факторы (их не знают владельцы магазина), и задача анализа — выявить их. Используйте этот метод, если считаете, что такие факторы существуют. LDA (Linear Discriminant Analysis) используется для задач классификации. Этот метод часто применяют для уменьшения размерности при предварительной обработке данных. Цель – спроецировать набор данных на пространство меньшей размерности с хорошей разделимостью кластеров. Метод делает предположения о нормально распределенных классах и равных ковариациях классов. Его можно применять даже в случаях, когда классов больше двух. Методы нелинейного преобразования используются, когда данные не лежат в линейном подпространстве. Если линейное подпространство трехмерного пространства можно представить как плоский лист бумаги, то примером нелинейного многообразия будет свернутый лист. К нелинейным методам относится, например, многомерное масштабирование (MDS), изометрическое отображение объектов (Isomap), локально-линейное вложение (LLE), собственное отображение 87 Гессе, t-SNE (стохастическое вложение соседей с t-распределением), UMAP (Uniform Manifold Approximation and Projection) и другие. Популярный метод уменьшения размерности, который дает впечатляющие результаты, – автокодировщик. Это тип искусственной нейросети, который обучается копировать свои входные данные в выходные, имея уменьшенный, по сравнению с входным и выходным, промежуточный слой. Автокодировщик сжимает входные данные в значения весов промежуточного слоя, а затем реконструирует выходные данные из этих значений.

1. **Методика и порядок выполнения работы**
   1. Учебная задача Задание. На основе предоставленного набора данных выполните снижение размерности

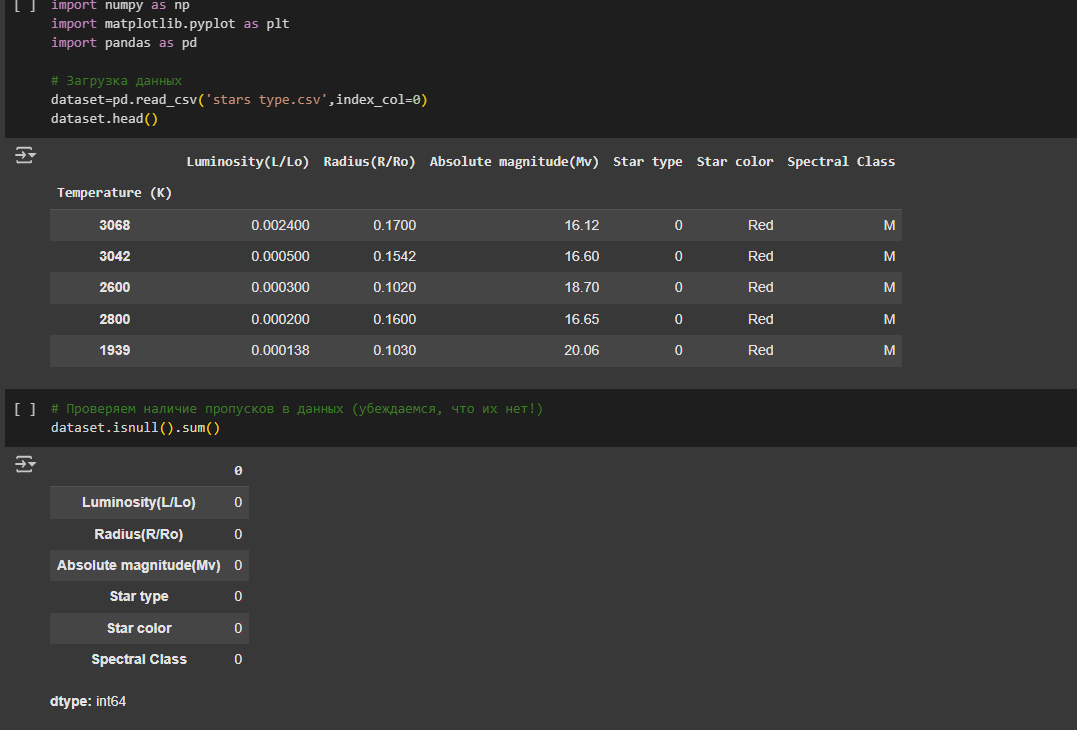


Рисунок 1 - Подключение библиотек и датасета.

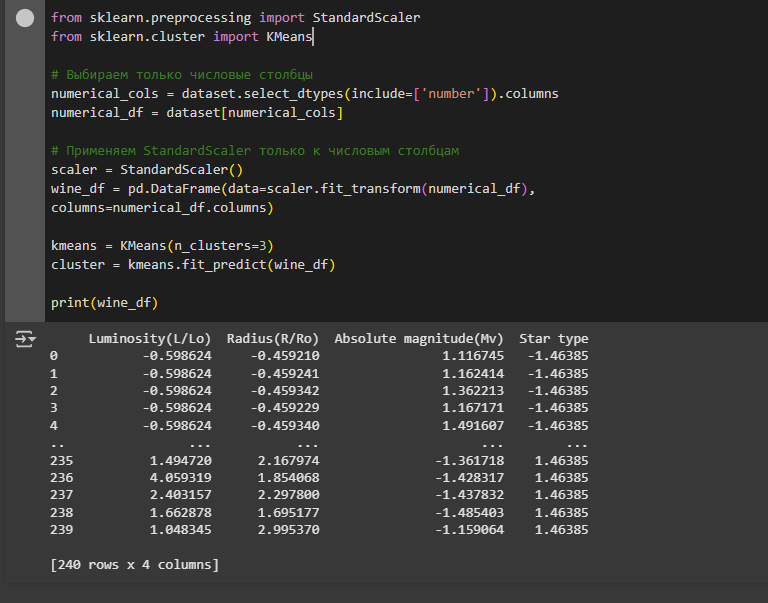
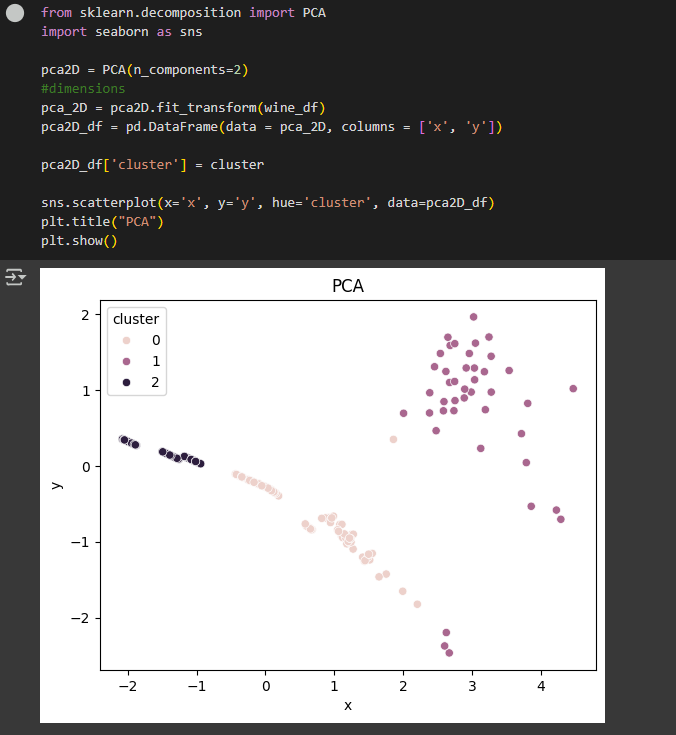


Рисунок 2 - Что-то делаем



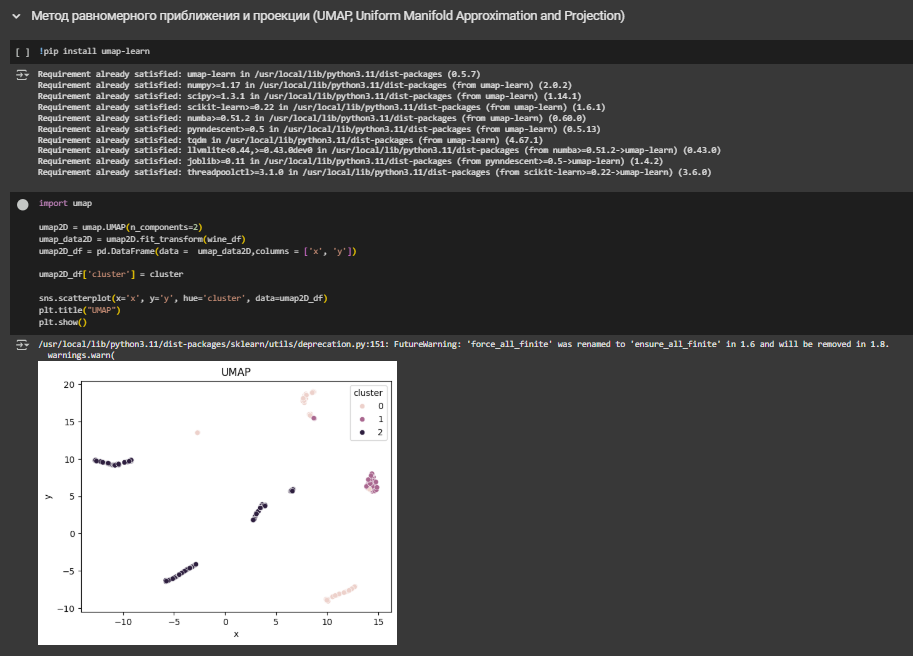
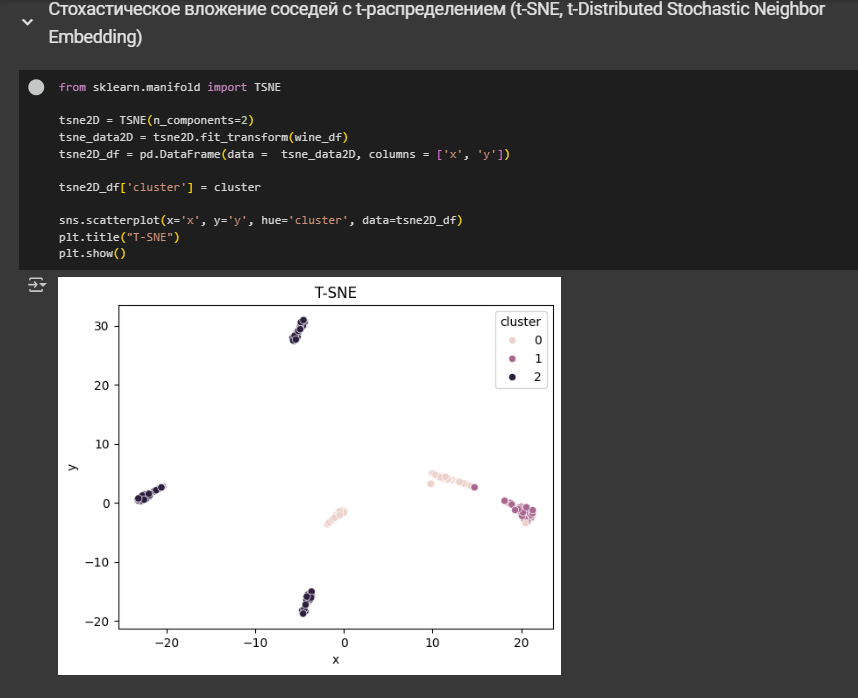
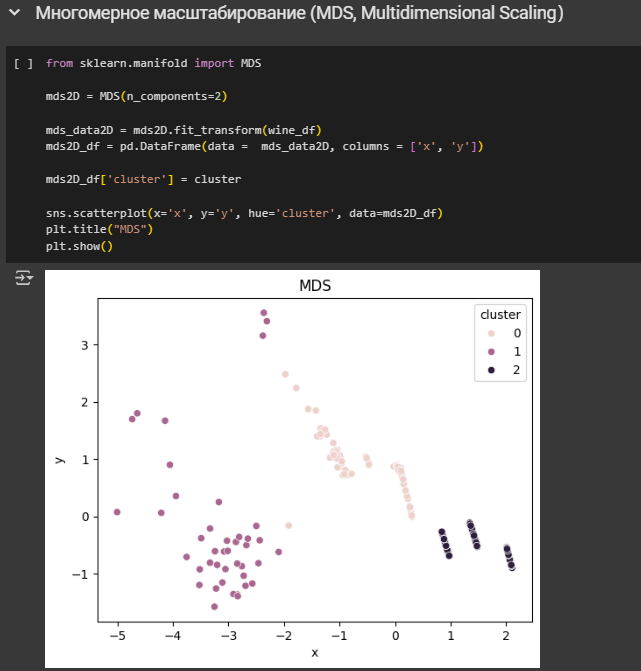
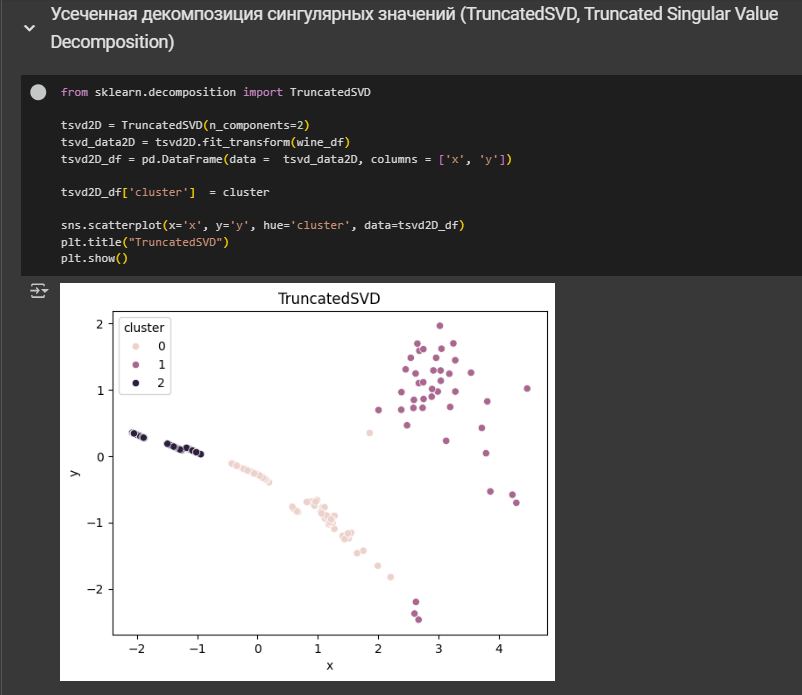
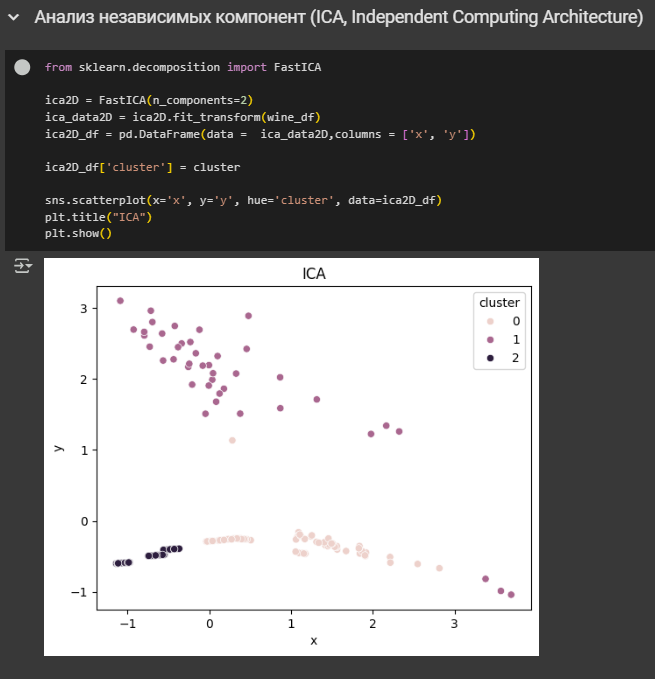


Рисунок 3 - Рисовашки

1. **Контрольные вопросы**

1)Что такое проклятие размерности?

"Проклятие размерности" (Curse of Dimensionality) — это явление, которое возникает при работе с данными высокой размерности. Оно заключается в том, что с увеличением количества признаков (размерности) объем пространства данных экспоненциально увеличивается, что приводит к ряду проблем:

Разреженность данных: Данные становятся более разреженными, то есть расстояния между объектами увеличиваются, и объекты становятся менее похожими друг на друга.

Переобучение: Модели машинного обучения становятся более склонными к переобучению, так как они могут запомнить шум в данных.

Вычислительная сложность: Вычислительная сложность алгоритмов увеличивается экспоненциально с ростом размерности.

Визуализация: Визуализация данных становится невозможной при размерности больше 3.

2)Перечислите основные алгоритмы понижения размерности.

Существует много разных алгоритмов понижения размерности, вот некоторые из основных:

PCA (Principal Component Analysis): Метод главных компонент.

t-SNE (t-distributed Stochastic Neighbor Embedding): t-распределенное стохастическое вложение соседей.

LDA (Linear Discriminant Analysis): Линейный дискриминантный анализ.

NMF (Non-negative Matrix Factorization): Неотрицательная матричная факторизация.

Autoencoders: Автокодировщики (нейронные сети).

ICA (Independent Component Analysis): Независимый компонентный анализ.

3)Опишите алгоритм и основную идею метода PCA.

Основная идея PCA: PCA — это метод, который используется для уменьшения размерности данных путем нахождения главных компонент. Главные компоненты — это новые признаки, которые являются линейными комбинациями исходных признаков и объясняют наибольшую дисперсию в данных. Алгоритм PCA:

Стандартизация данных: Исходные данные стандартизируются, чтобы иметь нулевое среднее и единичную дисперсию.

Вычисление ковариационной матрицы: Вычисляется ковариационная матрица для стандартизированных данных.

Вычисление собственных векторов и собственных значений: Вычисляются собственные векторы и собственные значения ковариационной матрицы.

Сортировка собственных значений: Собственные значения сортируются в порядке убывания.

Выбор главных компонент: Выбираются k собственных векторов, соответствующих k наибольшим собственным значениям. Эти собственные векторы называются главными компонентами.

Проецирование данных на главные компоненты: Исходные данные проецируются на главные компоненты, чтобы получить данные в новом пространстве меньшей размерности.

4)Опишите концепцию метода t-SNE.

Концепция t-SNE: t-SNE — это метод понижения размерности, который особенно хорошо подходит для визуализации данных высокой размерности.

Сохранение локальной структуры: t-SNE пытается сохранить локальную структуру данных, то есть объекты, которые близки друг к другу в исходном пространстве, должны быть близки друг к другу и в пространстве меньшей размерности.

Вероятностное моделирование: t-SNE использует вероятностное моделирование для определения схожести между объектами.

t-распределение: t-SNE использует t-распределение Стьюдента для моделирования расстояний между объектами в пространстве меньшей размерности. Это позволяет уменьшить эффект "скученности" объектов в центре пространства.

В отличие от PCA, t-SNE является нелинейным методом, что позволяет ему обнаруживать более сложные структуры в данных. Однако, t-SNE является вычислительно сложным алгоритмом и может потребовать много времени для обучения на больших наборах данных.

**Вывод:**  научились использовать обучаемые алгоритмы снижения размерности.