St Petersburg University
www.spbu.ru

Влияние уровня обезличивания на устойчивость кластеров датасета в больших данных

Богданов Александр Владимирович
Щеголева Надежда Львовна
Дик Геннадий Давидович
Дик Александр Геннадьевич
Киямов Жасур Уткирович
Савков Егор Кириллович

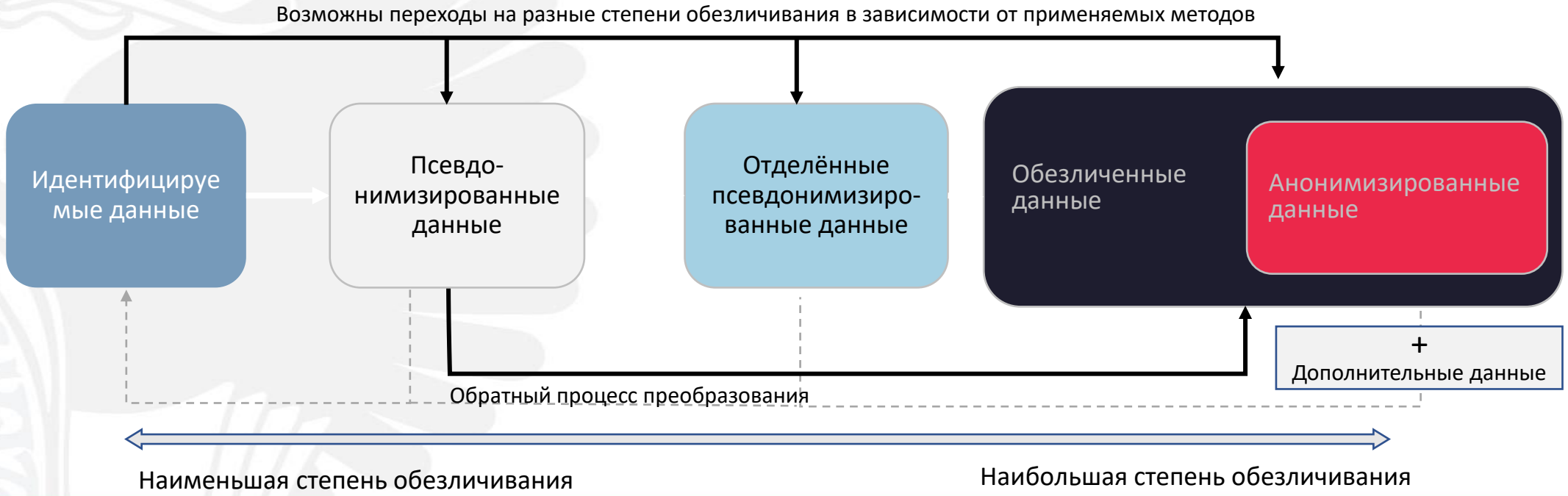
2025 г.

Защита персональных данных

В последнее время Большие Данные все больше превращаются из модного направления исследований в инструмент практического использования для многих разделов практической деятельности.

Появление новых источников информации делает возможным их сопоставление с опубликованными (доступными) ранее и неизбежно ведет к появлению рисков повторной идентификации ранее обезличенных данных.

Это в свою очередь заставляет отказаться от концепции гарантировано анонимизированных данных, вводя определенные границы работы с рисками и выстраивая непрерывный процесс оценки угроз.



Виды техник обезличивания

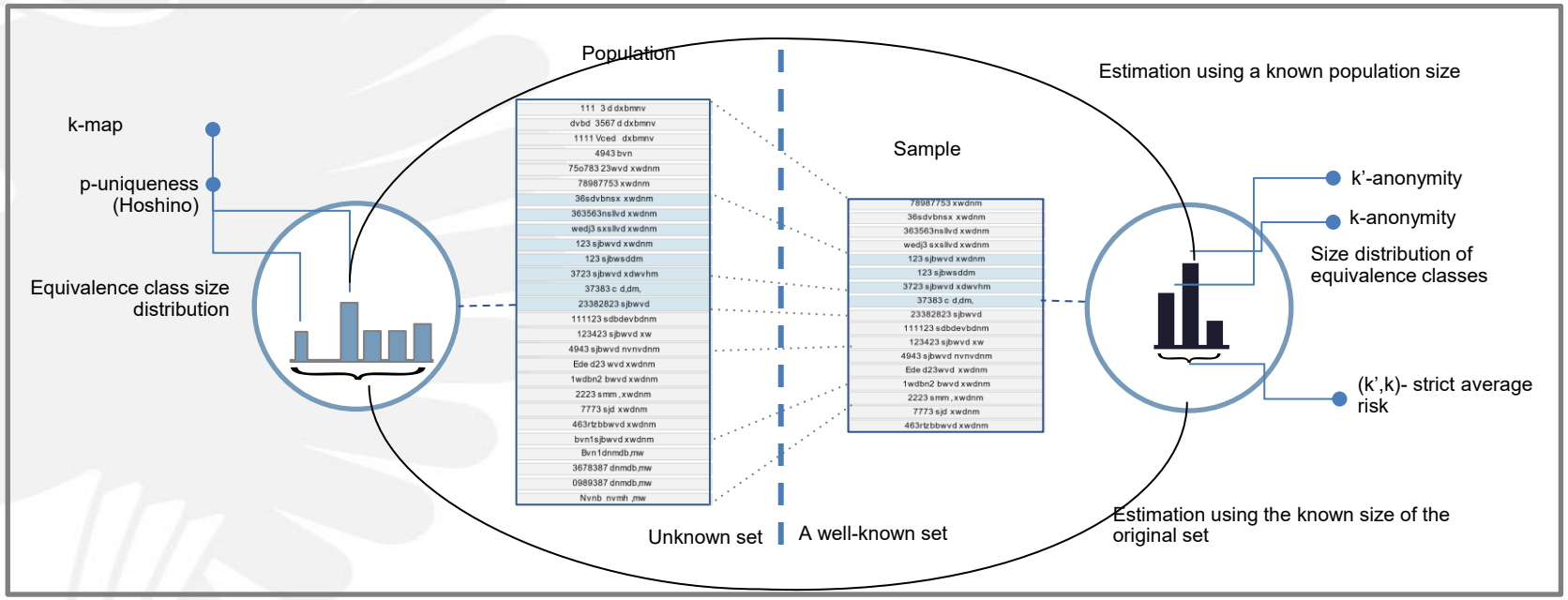
- Обобщение
  - Локальное обобщение
  - Агрегация
- Техники рандомизации
  - Возмущение
  - Микро-агрегация
  - Перемешивание
- Псевдонимизация
  - Создание псевдонимов
  - Маскеризация
- Шифрование
  - Детерминированное шифрование
  - Гомоморфное шифрование
- Подавление
  - Локальное подавление
  - Удаление атрибутов
- Другие методы
  - Семплинг
  - Синтетические данные
  - Метод декомпозиции

Метрики анонимизации

1. **k - anonymity** - Мера степени анонимности в наборе данных. Набор данных считается k-анонимным, если каждая запись в наборе данных неотличима от по меньшей мере k-1 других записей в наборе данных.

2. **ℓ - diversity** Мера разнообразия чувствительных значений атрибутов в группе k-анонимных записей. Он измеряет вероятность того, что злоумышленник сможет повторно идентифицировать человека на основе комбинации квази-идентификаторов и конфиденциальных атрибутов..

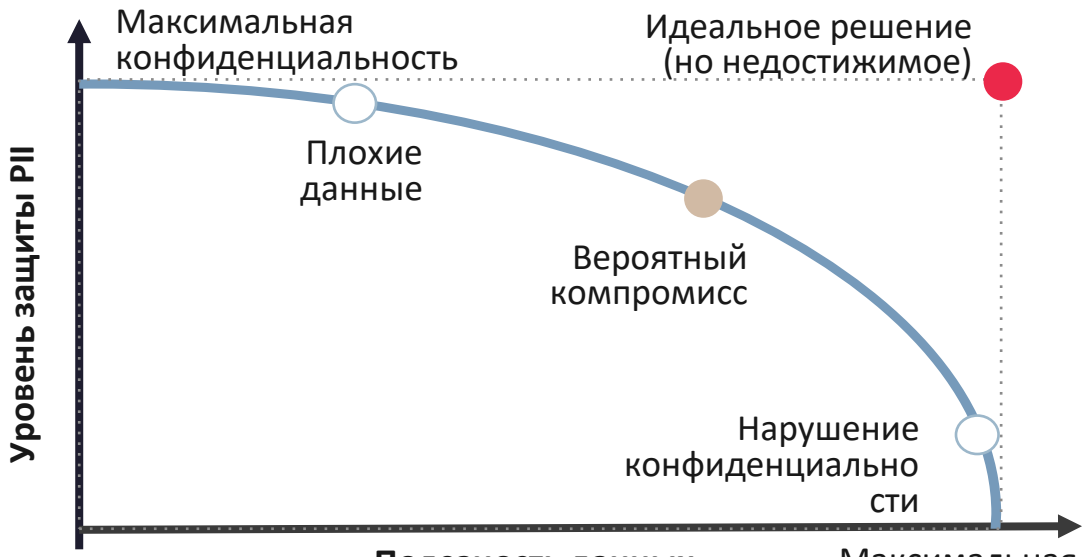
3. **t - closeness** Мера сходства между распределением значений чувствительных атрибутов в исходном наборе данных и в анонимном наборе данных. Он измеряет степень, в которой анонимизированный набор данных сохраняет статистические свойства исходного набора данных.

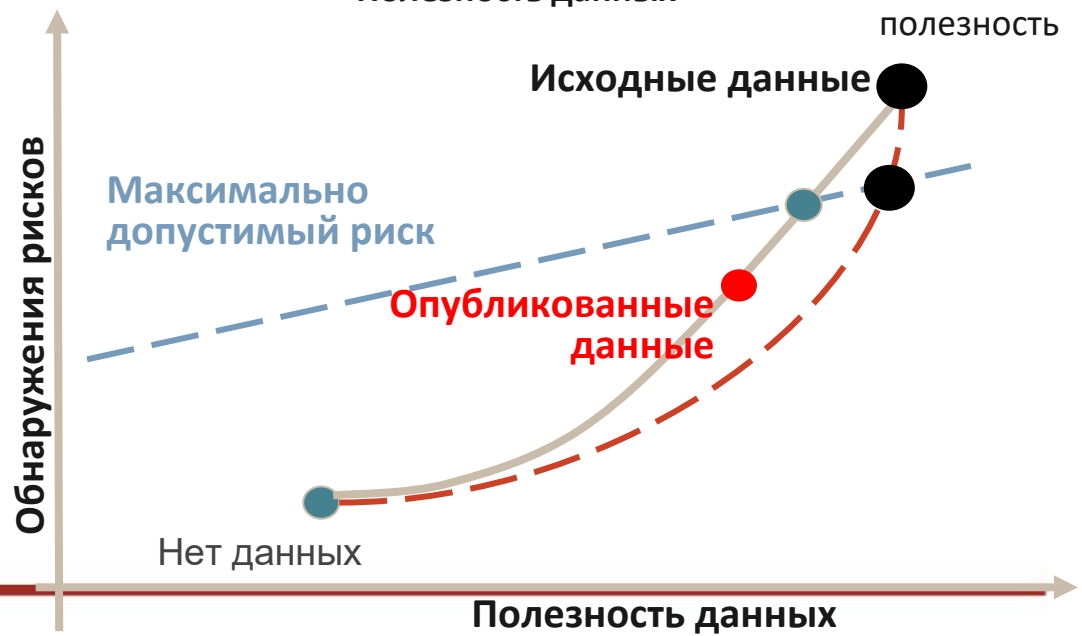


Полезность и приемлемый порог риска

Обезличивание **требует баланса** между **полезностью** получаемых в результате обезличивания данных и **приемлемой величиной риска**. Приемлемый порог риска повторной идентификации персональных данных зависит от различных факторов, включая конфиденциальность данных, контекст их использования и применимые законы и правила.

Как правило, допустимый порог риска повторной идентификации считается очень низким, поскольку любая повторная идентификация персональных данных может привести к нарушению конфиденциальности и безопасности.





Обезличенный датасет

Идентификатор | Квази - идентификатор | Конфиденциальные атрибуты

ФИО	Возраст	Семейное положение	Почтовый индекс	Доход	Сексуальная ориентация
Сергей Петров	32	⚭	198000	450 000 ₽	⚥
Дмитрий Иванов	34	◯	198023	350 000 ₽	⚥
Татьяна Сергеева	33	⚭	198101	150 000 ₽	⚢
Джек Румор	43	◯	103126	700 000 ₽	⚧
Пелагея Ильина	47	⚭	103173	550 000 ₽	⚥
Николай Борисов	45	⚭	103139	600 000 ₽	⚥

Обобщенные квази-идентификаторы | Конфиденциальные атрибуты

Возраст	Семейное положение	Почтовый индекс	Доход	Сексуальная ориентация
30-35	0.33	198***	450 000 ₽	⚥
30-35	0.33	198***	350 000 ₽	⚥
30-35	0.33	198***	150 000 ₽	⚢
40-45	0.67	1031**	700 000 ₽	⚧
40-45	0.67	1031**	550 000 ₽	⚥
40-45	0.67	1031**	600 000 ₽	⚥

Класс эквивалентности

Информация, известная злоумышленнику

ФИО | Адрес | Пол | Дата Рождения | ..

Повторная идентификация

Пол | Возраст | .. | Доход | .. | ..

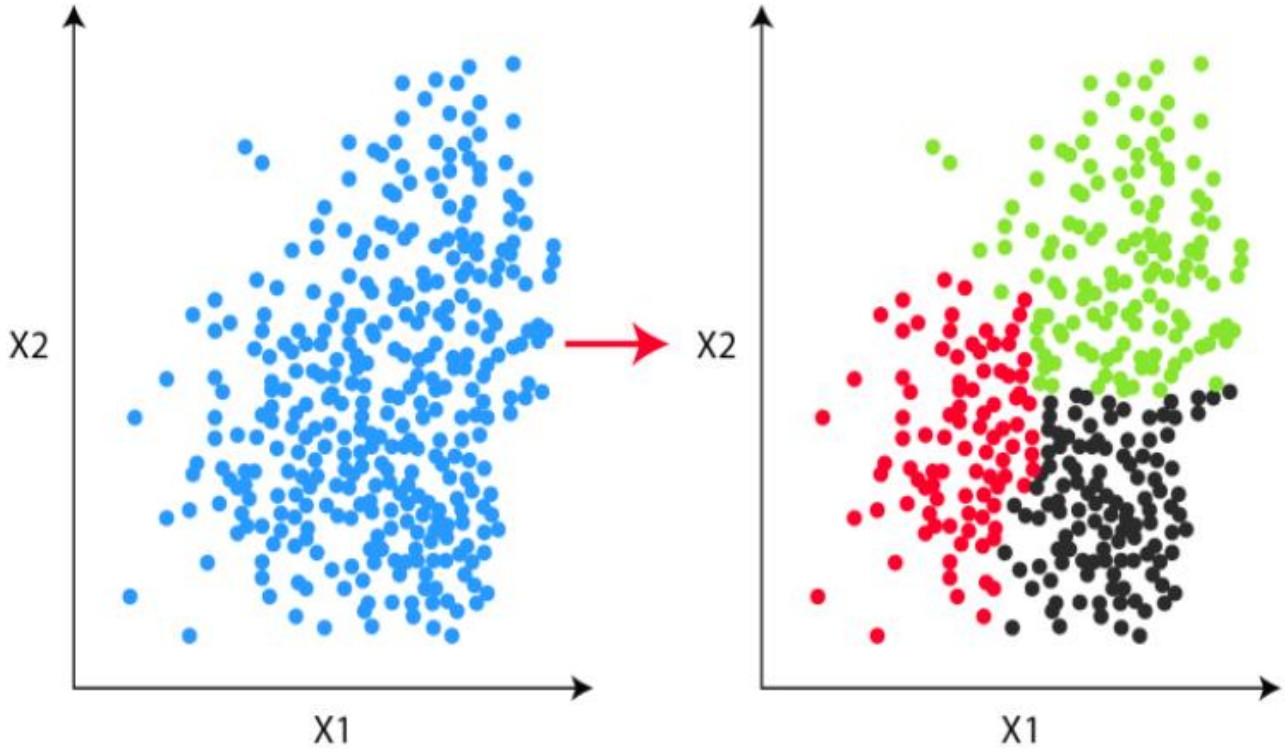
Обезличенный набор данных

Кластеризация

Кластеризация или кластерный анализ — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы

Для кластеризации используются следующие шаги:

1) Выбор метода кластеризации
2) Выбор метода наиболее информативных признаков
3) Выбор метода измерения расстояния
4) Выбор метода оценки качества кластеризации

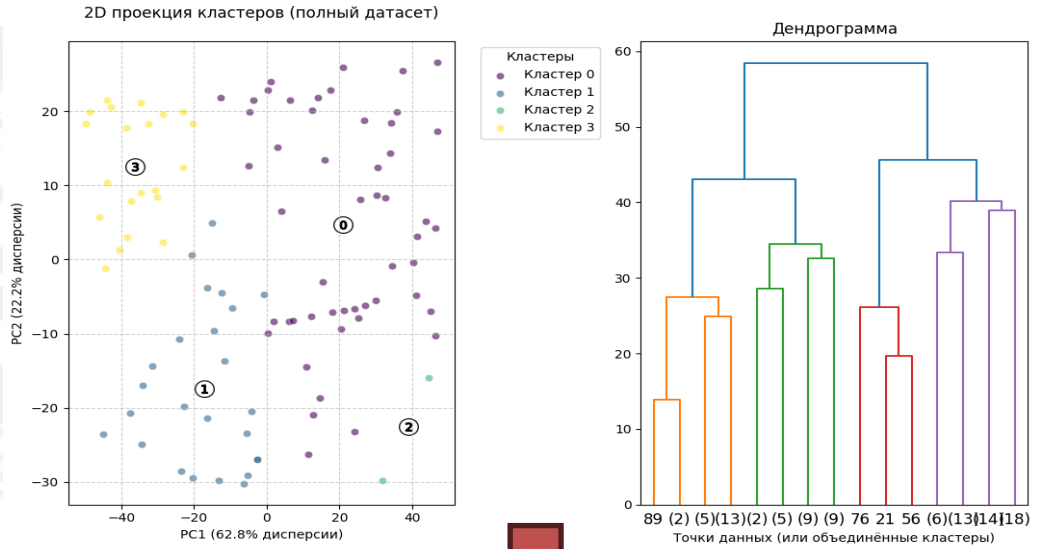


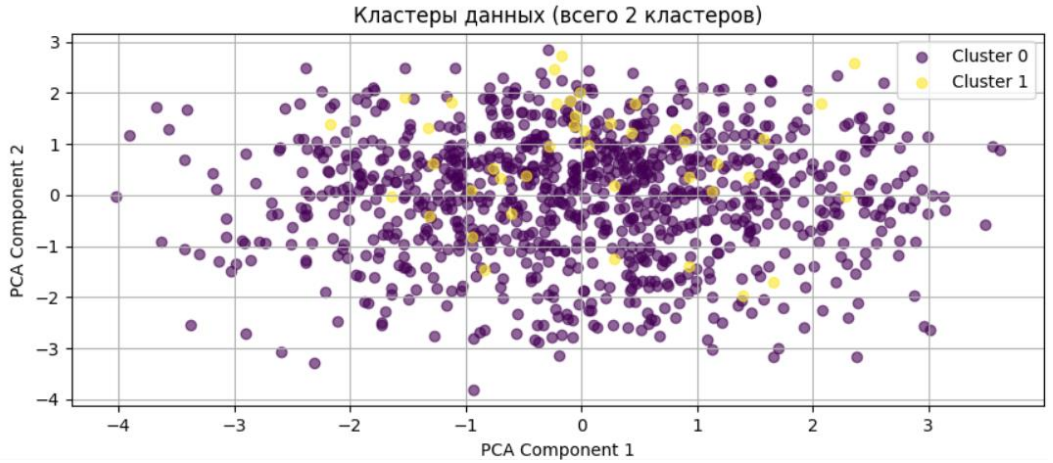
Методы для задачи кластеризации

Методы кластеризации	Методы выбора наиболее информативных признаков	Методы способа измерения растояния	Методы оценки качества кластеризации
Иерархическая	Компактность (плотность)	Евклидово расстояние	Индекс Rand
Алгоритм максиминного расстояния	Разнесенность образов в пространстве характеристик	Квадрат евклидового расстояния	Индекс Жаккара
Алгоритм ISODATA	Метод последовательного сокращения (алгоритм Del)	Корреляция Пирсона	Индекс Фоулкса-Меллова
Алгоритм CURE	Метод последовательного добавления признаков (алгоритм Add)	Расстояние Чебышева (Chebychev)	Индекс Phi
Алгоритм FOREL	Метод случайного поиска с адаптацией (алгоритм СПА)	Расстояние Минковского (Minkowski)	Компактность кластеров
			Отделимость кластеров

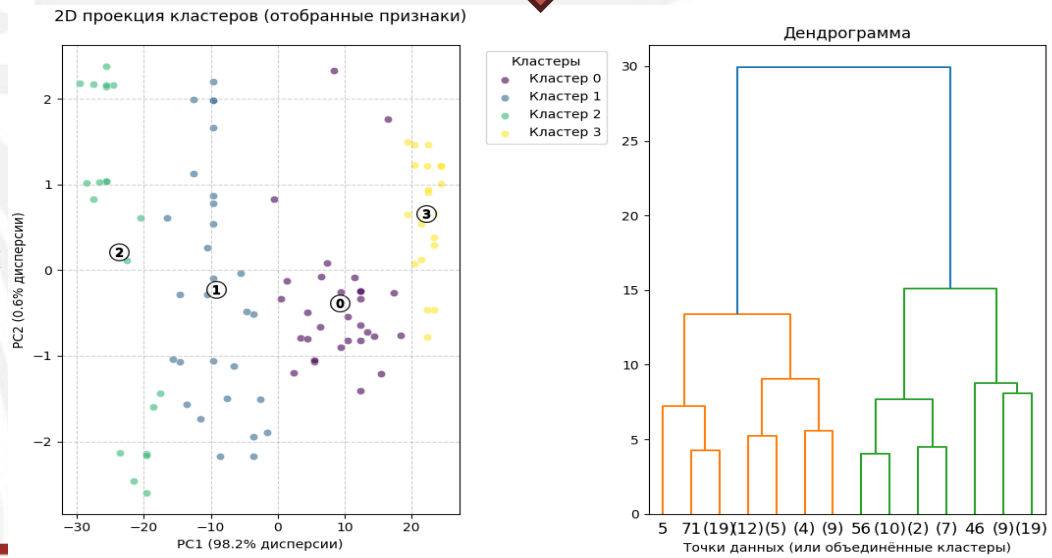
Результаты

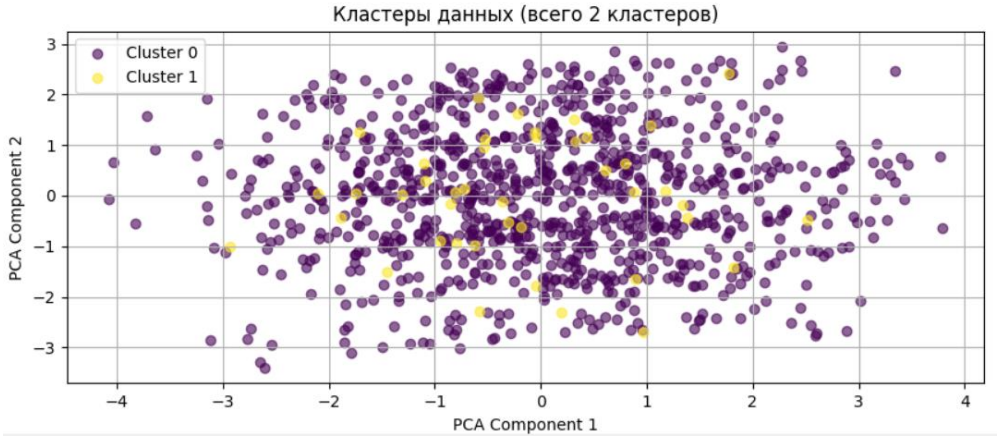
До обезличивания





После обезличивания



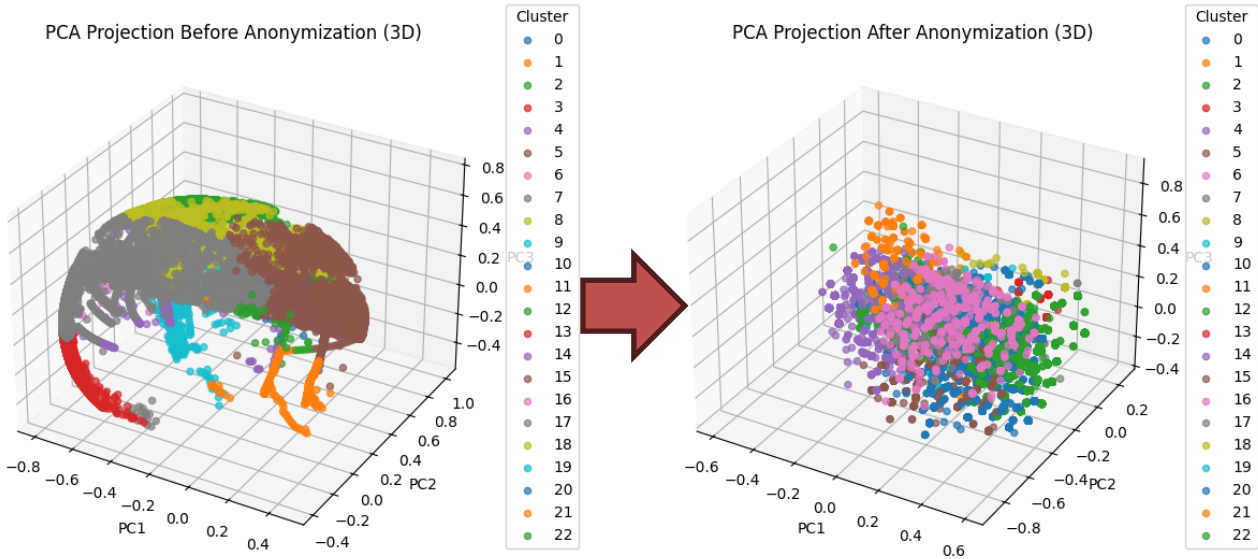


Выводы

1) Выбор самого метода кластеризации не сильно влияет на сравнение обезличенного и не обезличенного датасета

2) Количество числа признаков сильно влияет на оценку качества кластеризации, особенно в обезличенном датасете, так как биннинг и группировка квази-идентификаторов устраняют «шум», а затем отбор самых разнесённых признаков восстанавливает качество кластеризации почти до первоначального уровня.

3) Если обезличивание идет в сторону защиты, а не полезности – то есть используются методы, такие как - … то обезличенный датасет сильно отличается от не обезличенного



До обезличивания

После обезличивания

Спасибо за внимание!

St Petersburg University
spbu.ru