**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный университет**

**Факультет прикладной математики-процессов управления**

**Кафедра фундаментальной информатики и информационных технологий**

**отчет**

**по лабораторной работе №2**

**по дисциплине «Алгоритмы и структуры данных»**

**на тему «Обезличивание синтетических данных»**

**Вариант – 4**

|  |  |  |
| --- | --- | --- |
| Студент гр. 22Б15-пу |  | Добренкова Л.С. |
| Преподаватель |  | Дик А.Г. |

**Санкт-Петербург**

**2023 г**

Оглавление

[1. Цель работы 3](#_Toc160105167)

[2. Задачи 3](#_Toc160105168)

[3. Теоретическая часть 3](#_Toc160105169)

[4. Алгоритм метода 6](#_Toc160105170)

[5. Спецификация программы 9](#_Toc160105171)

[6. Рекомендации пользователя 10](#_Toc160105172)

[7. Рекомендации программиста 10](#_Toc160105173)

[8. Контрольный пример 10](#_Toc160105174)

[9. Вывод 11](#_Toc160105175)

[10. Литература 11](#_Toc160105176)

# Цель работы

Цель работы состоит в разработке программного решения для обезличивания датасета с учетом введенных квази-идентификаторов и реализации алгоритма k-анонимизации. Это предполагает создание программы с графическим интерфейсом, который обеспечит удобство использования для пользователя.

# Задачи

1. Изучить методы обезличивания.
2. Проанализировать их преимущества и недостатки для разных столбцов таблицы.
3. Выбрать лучший метод для каждой категории.
4. Реализовать выбранный метод.
5. Протестировать написанный алгоритм.
6. Провести контрольный эксперимент.

# Теоретическая часть

K-анонимность - это свойство релиза данных, при котором для каждой группы значений атрибутов, называемой квази-идентификатором, в наборе данных гарантируется, что каждая комбинация значений квази-идентификаторов встречается не менее k раз, где k - заданный параметр безопасности. Таким образом, это обеспечивает защиту конфиденциальности, предотвращая возможность однозначной идентификации отдельных субъектов данных. Более подробно тема k-анонимности расписана в статье [1].

В контексте датасета поликлиники, который является результатом лабораторной работы №1, необходимо рассмотреть существующие методы обезличивания данных и выбрать подходящие для этого конкретного набора данных.

* 1. Локальное обобщение – В рамках данной техники предполагается уменьшение специфичности атрибута за счет подмены точного значения атрибута его общим значением .
  2. Агрегация – Техника основана на построении обобщений по одному или нескольким атрибутам (групповые операции). Представляет из себя построение обобщенных данных по выделенным атрибутам.
  3. Возмущение –внесение шума в данные, которые перестают быть точными или правдивыми, но сохраняют основные статистические закономерности. Метод применим для бинарных данных (да/нет, например, пола), цифровых BLOBS (фотографии), статистически частых наборов данных. Позволяет сохранить статистическую ценность набора данных при незначительных потерях информации.
  4. Микро-агрегация – Техника предполагает формирование групп записей, для которых вместо точных значений указывается диапазон. Метод рекомендуется к применению для параметров возраста, географических локаций, финансовых транзакций, биометрических данных. Один из самых широко применяемых методов.
  5. Перемешивание – Техника направлена на перемешивание аналогичных записей при сохранении статической значимости всего набора. Наиболее полезна при обезличивании временных интервалов, например, информации об отпусках. Метод применяется, как правило, в совокупности с другими методами обезличивания.
  6. Создание псевдонимов – Для чувствительного атрибута проводится замена идентификаторов на условные обозначение или цифровые ключи. Как правило, проводится для частных данных. Используется в совокупности с другими методами обезличивания.
  7. Маскеризация – В рамках данной техники проводится замена части записи заполнителями по определенному шаблону. Хотя в результате исходной строке сопоставляется строка определенного шаблона, что позволяет отнести этот метод к псевдонимизации, по ряду характеристик этот метод может быть сопоставлен микро-агрегации: в результате применения шаблонов возникают обобщенные группы. Применим для идентификаторов (номера паспортов, телефоны, номера карт)
  8. Локальное подавление – удаление или перекодирование относительно редких записей данных. В отличии от метода удаления атрибутов нацелен на удаление строк или значений для заданных атрибутов в выделенных строках. Наибольшее применение имеет для медицинской и биометрической информации, а также больших наборов статистических данных с аномалиями.
  9. Удаление атрибутов – удаление чувствительного контента без добавления замен. Метод используется для удаления прямых идентификаторов, а также удаления избыточных квази-идентификаторов. Реализован как замена всех значений атрибута на символ “\*”.
  10. Метод декомпозиции – Разделение множества атрибутов на несколько подмножеств с последующим раздельным хранением. Метод не является собственно техникой обезличивания, но позволяет формировать наборы данных, к которым могут применяться методы обезличивания.

Для каждого атрибута набора данных можно применить определенный метод обезличивания:

1. Атрибут "Name" содержат информацию, позволяющую однозначно идентифицировать человека. Для обезличивания данных оптимально использовать метод создания псевдонимов при достижении требуемых показателей k-анонимности по всем выбранным квази-идентификаторам, и метод удаления атрибутов в противном случае. Такой подход позволяет сохранить информацию, не утрачивая при этом уровень анонимизации.

2. "Snils" содержат информацию, позволяющую однозначно идентифицировать человека. Для обезличивания данных оптимально использовать метод маскеризации при достижении требуемых показателей k-анонимности по всем выбранным квази-идентификаторам, и метод удаления атрибутов в противном случае.

3. "Passport": Данный атрибут также содержит информацию, по которой можно однозначно идентифицировать человека. Однако первые две цифры кода региона не относятся к конкретному человеку. Можно применить метод маскеризации, чтобы скрыть оставшуюся информацию. Эти данные могут быть полезны для анализа географии распространения болезней.

4. "Symptoms" и "Analysis": Для этих атрибутов можно применить агрегацию по ключевым симптомам и анализам соответственно. Это поможет сохранить полезную информацию о болезнях и присутствующих симптомах.

5. "Doctor": Информация о докторе может быть обобщена до категории специализации. Таким образом, можно сохранить общую информацию о врачах без раскрытия их личных данных. Это может быть полезно для анализа загруженности различных специализаций врачей.

6. "DateStart": Дату приема можно обобщить до сезона года, что может быть полезно для анализа сезонных заболеваний.

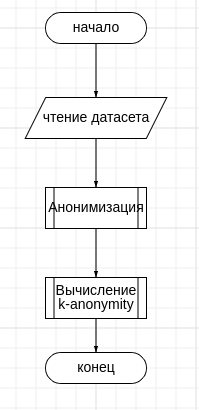
7. "Price": Для атрибута стоимости услуг можно применить микро-агрегацию путем округления вниз до ближайшего кратного 2500. Это позволит сохранить общую структуру данных о ценах на медицинские услуги.

8. "Card": Данные о банковской карте могут быть обобщены, учитывая, что часть цифр относится к платежной системе и банку. Это может быть полезно для анализа популярных платежных систем среди клиентов.

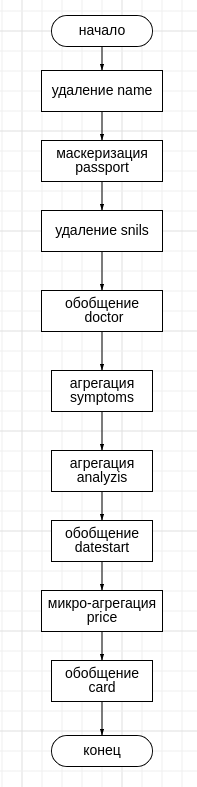
# Алгоритм метода

1. Загрузка датасета из указанного пользователем файла.
2. Выбор квази-идентификаторов, по которым будет высчитываться k-анонимность.
3. Обезличивание датасета.
4. Расчет k-анонимности датасета.
5. Поиск пяти строк с наименьшими значениями коэффициента k-анонимности.
6. Вычисление уникальных строк (показатель k-анонимности = 1).
7. Вывод полученных результатов пользователю.

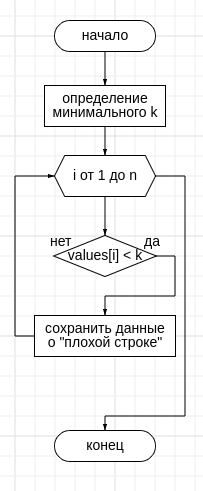
На рисунках 4.1-4.3 представлена блок-схема алгоритма.



*Рис. 4.1 «run\_handler»*



*Рис. 4.2 «anonymize»*



*Рис. 4.3 «k-anonymity»*

# Спецификация программы

Программа реализована на языке python 3.11 с использованием следующих пакетов: pandas, tkinter.

В программе используются 14 функций: 11 связанных с обезличиванием данных и 3 для обработки интерфейса. В таблицах 5.1-5.3 представлено описание всех функций.

*Таблица 5.1. «Описание основных функций»*

|  |  |
| --- | --- |
| Имя функции | Метод обезличивания |
| maskerization\_passport | Маркеризация паспорта |
| generalization\_doctor | Локальное обобщение врачей |
| generalization\_date\_start | Локальное обобщение даты приема |
| generalization\_date\_start | Агрегация |
| aggregation\_price | Микро-агрегация цен |
| generalization\_card | Локальное обобщение карт |
| aggregate | Агрегация |
| anonimyze | Анонимизация датасета |
| calc\_k\_anonymity | Вычисление k-анонимности |

# Рекомендации пользователя

Для успешного запуска программы необходима программная среда, поддерживающая python 3.11.

В поле ввода «Input» необходимо указать путь до файла датасета.

Далее необходимо отметить щелчком левой кнопки мыши все квази идентификаторы, которые б использованы для вычисления k-анонимности.

После этого нужно нажать левой кнопкой мыши на кнопку «Run» в окне программы.

Результат отобразится в поле внизу окна программы.

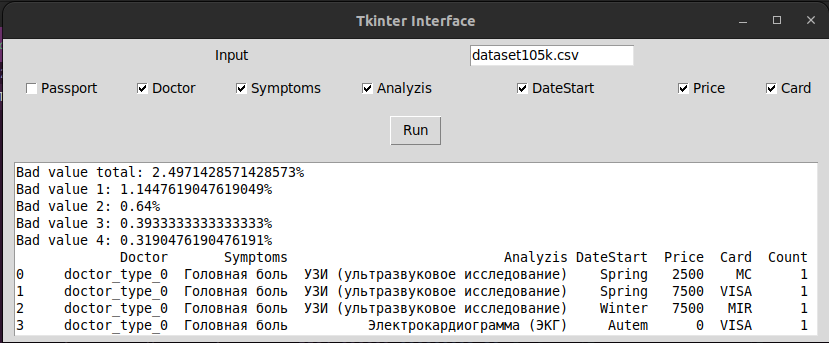
# Рекомендации программиста

Для запуска программы необходима 64-битная операционная система Windows, Linux или macOS. Для работы с кодом необходима среда разработки, совместимая с python 3.10 и библиотеки pandas и tkinter.

Исходный код программы и необходимые текстовые файлы доступны по ссылке: https://github.com/v131v/algo\_labs/tree/main/lab2

# Контрольный пример

В данном разделе представлен контрольный пример, демонстрирующий работу программы.

**

*Рис. 8.1 «Интерфейс программы»*

Датасет в указанном файле был анонимизирован по выбранным квази-идентификаторам. Также были проведены соответствующие вычисления.

# Вывод

В результате выполнения работы разработана и написана программа, позволяющая с помощью интерфейса пользователя выбирать квази-идентификаторы, по которым происходит обезличивание таблицы, и выводится коэффициент анонимности.

# Литература

1. Samarati, P., & Sweeney, L. (1998). Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression. Technical Report SRI-CSL-98-04, SRI International.