Санкт-Петербургский государственный университет

Факультет прикладной математики – процессов управления

Лабораторная работа №2

**Отчет**

по дисциплине «Алгоритмы и структуры данных»

Обезличивание датасета

Вариант 4

Автор работы: Добренкова Л.С.

Группа: 22.Б15-пу

Преподаватель: Дик А. Г.

Санкт-Петербург, 2023

Оглавление

[1. Цель работы 3](#__RefHeading___Toc399_3468946893)

[2. Задачи 3](#__RefHeading___Toc401_3468946893)

[3. Теоретическая часть 3](#__RefHeading___Toc403_3468946893)

[4. Алгоритм метода 5](#__RefHeading___Toc405_3468946893)

[5. Спецификация программы 8](#__RefHeading___Toc407_3468946893)

[6. Рекомендации пользователя 9](#__RefHeading___Toc409_3468946893)

[7. Рекомендации программиста 9](#__RefHeading___Toc411_3468946893)

[8. Контрольный пример 9](#__RefHeading___Toc413_3468946893)

[9. Вывод 10](#__RefHeading___Toc415_3468946893)

# Цель работы

Разработать и написать программу, обезличивающую датасет с учетом введенных квази идентификаторов. Программа должна вычислять k анонимизацию. Так же у программы должен быть графический интерфейс.

# 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

К-anonymity (k-анонимность) - это концепция в области защиты личной информации, применяемая к датасетам. Идея заключается в том, чтобы обеспечить такой уровень анонимности, при котором каждая запись в датасете не может быть однозначно идентифицирована на основе определенного набора атрибутов.

Более формально, датасет считается k-анонимным, если для каждой уникальной комбинации значений в определенных атрибутах существует по крайней мере k записей в датасете. Это означает, что для каждой конкретной группы данных в определенных столбцах (атрибутах) существует как минимум k записей с одинаковыми значениями этих атрибутов, что делает невозможным точное идентифицирование отдельных индивидов.

Рассмотрим существующие методы обезличивания и выберем подходящие для датасета поликлиники, являющегося результатом лабораторной работы №1.

* 1. Локальное обобщение – замена конкретных данных общими или абстрактными значениями для сокрытия чувствительных деталей.
  2. Агрегация – объединение данных для создания обобщенных или анонимных групп данных.
  3. Возмущение – изменение данных путем введения небольших искажений или шумов.
  4. Микро-агрегация – аналогично агрегации, но в более мелком масштабе, с группировкой данных на более низком уровне.
  5. Перемешивание – перестановка данных или замена их местами.
  6. Создание псевдонимов – использование вымышленных или абстрактных идентификаторов вместо реальных.
  7. Маскеризация – замена чувствительных данных символами или общими значениями (масками).
  8. Локальное подавление – скрытие чувствительных данных на уровне локальных копий, но с сохранением общей структуры.
  9. Удаление атрибутов – удаление определенных атрибутов данных.
  10. Метод декомпозиции – разбиение данных на более мелкие агрегаты.

Выберем метод на каждый столбец таблицы и оценим полезность обезличенного датасета.

«Name», «Snils» - содержиат информацию информацию, по которой можно однозначно идентифицировать человека. Один из оптимальных способов обезличивания – удаление атрибута.

«Passport» - так же содержит информацию, по которой можно однозначно идентифицировать человека, однако первые 2 цифры отвечают за код региона, а не конкретного человека, поэтому мы можем применить маскеризацию. Данные могут быть полезны, так как дают представление о географии распространения болезней.

«Symptoms, «Analyzis» содержат симптомы, с которыми пациенты обратились и анализы, которые им назначили. В большинстве случаев болезни соответствуют какие-то преимущественные симптомы, по которым и определяют болезнь, поэтому данные можно агрегировать по ключевым симптомам. С анализами можно поступить похожим образом, так как они зависят от симптомов.

«Doctor» - информация о докторе, к которому пациент обратился. Для профессий врачей можно сделать классификацию и оставлять только класс, к которому относится доктор. То есть провести обобщение. Может быть полезно, например для статистики о наиболее загруженных классах врачей.

«DateStart» - дата приема. Так же можно обобщить до сезона года. Это может быть полезно для выявления сезонных заболеваний.

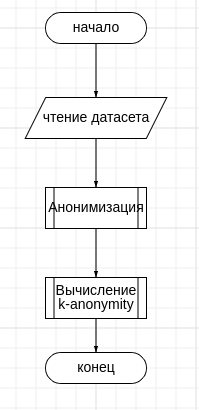
«Price»- стоимость услуг. Применима микро-агрегация - округление вниз до ближайшего кратного 2500. Может помочь при исследованиях рынка мединцинских услуг.

«Card» - данные банковской карты. В большинстве случаев банковские карты привязаны к конкретному клиенту, поэтому по ним можно идентифицировать человека, однако как и в случае с паспортом, часть цифр отвечает за принадлежность к платежной системе и банку. Поэтому можем обобщить. Полезно для определения популярных у клиентов платежных систем.

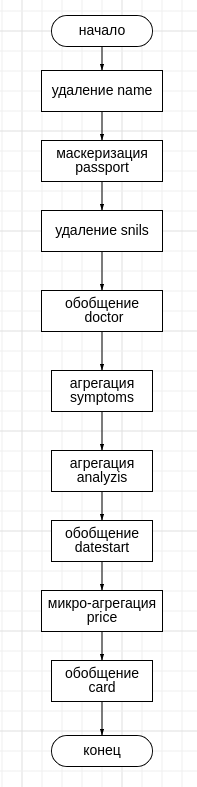
# Алгоритм метода

1. Загрузить датасет, указав файл.
2. Выбрать квази-идентификаторы, по которым будет высчитываться k-анонимности.
3. Происходит обезличивание датасета.
4. По размерам датасета вычисляется минимальное значение k.
5. Высчитываются первые 5 «плохих» значений коэффициента анонимности.
6. Вычисляются уникальные строки (k=1).
7. Полученные результаты выводятся пользователю.

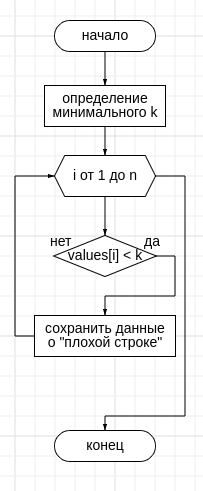
На рисунках 4.1-4.3 представлена блок-схема алгоритма.



*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 среда, поддерживающая python 3.11.

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

Далее необходимо отметить все квази идентификаторы, которые будут использованы для вычисления k-анонимности.

После этого нужно нажать «Run».

Результат отобразится ниже.

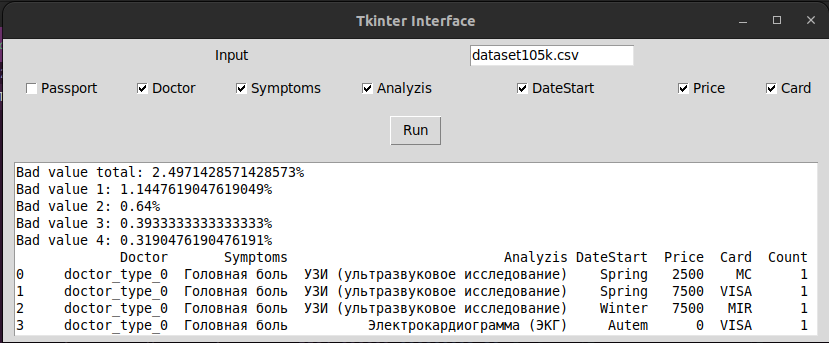
# Рекомендации программиста

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»*

После указания файладатасета и выбора квази-идентификаторов, датасет был анонимизирован и 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.