**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**Факультет прикладной математики – процессов управления**

**отчет**

**по лабораторной работе**

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

**на тему «Обезличивание данных»**

**Вариант 2**

|  |  |  |
| --- | --- | --- |
| Студент гр. 23.Б16 |  | Барафанов А.А. |
| Преподаватель |  | Дик А.Г. |

**Санкт-Петербург**

**2024 г.**

**Оглавление**

[**1.** **Цель работы** 3](#_Toc179501287)

[**2.** **Задача (формализация)** 3](#_Toc179501288)

[**3.** **Теоретическая часть** 3](#_Toc179501289)

[**4.** **Описание алгоритма** 6](#_Toc179501290)

[**5.** **Описание программы** 9](#_Toc179501291)

[**5.1** **Описание функций** 9](#_Toc179501292)

[**6.** **Рекомендации пользователю** 11](#_Toc179501293)

[**7.** **Рекомендации программисту** 11](#_Toc179501294)

[**8.** **Контрольный пример** 12](#_Toc179501295)

[**9.** **Вывод** 13](#_Toc179501296)

# **Цель работы**

Написать алгоритм для вычисления K-anonymity датасета, а также провести его обезличивание различными методами обезличивания.

# **Задача (формализация)**

Разработка делится по функционалу на две части: написание функции, высчитывающей K-anonymity, и обезличивание датасета предлагаемыми на выбор методами:

* + Локальное обобщение;
  + Агрегация;
  + Возмущение;
  + Микро-агрегация;
  + Перемешивание;
  + Создание псевдонимов;
  + Маскеризация;
  + Локальное подавление;
  + Удаление атрибутов;
  + Метод декомпозиции.

# **Теоретическая часть**

*Обезличивание данных* — это процесс обработки персональных данных с целью удаления или замены всех прямых или косвенных идентификационных характеристик, которые могут связать данные с конкретным человеком.

*Методы обезличивания* — техники, применяемый к персональным данным с целью снижения риска идентификации субъекта персональных данных. В результате применения соответствующего метода из набора данных удаляется идентифицирующая личность человека информация. Оценка риска обезличенных данных производится для каждого набора данных и каждого метода. Все методы являются параметризуемыми, то есть содержат настраиваемые характеристики (такие как набор квази-идентификаторов), определяющие конкретный алгоритм в отношении конкретного набора данных в заданных условиях (сценариях и контексте).

*K-анонимность (k-anonymity)* — концепция в области обезличивания данных, представляющая собой методологию, при которой каждая запись в наборе данных неотличима от по крайней мере k-1 других записей по определенному набору атрибутов. Этот подход направлен на устранение уникальных идентификационных признаков и сведение к минимуму риска нежелательного раскрытия личной информации при анализе данных.

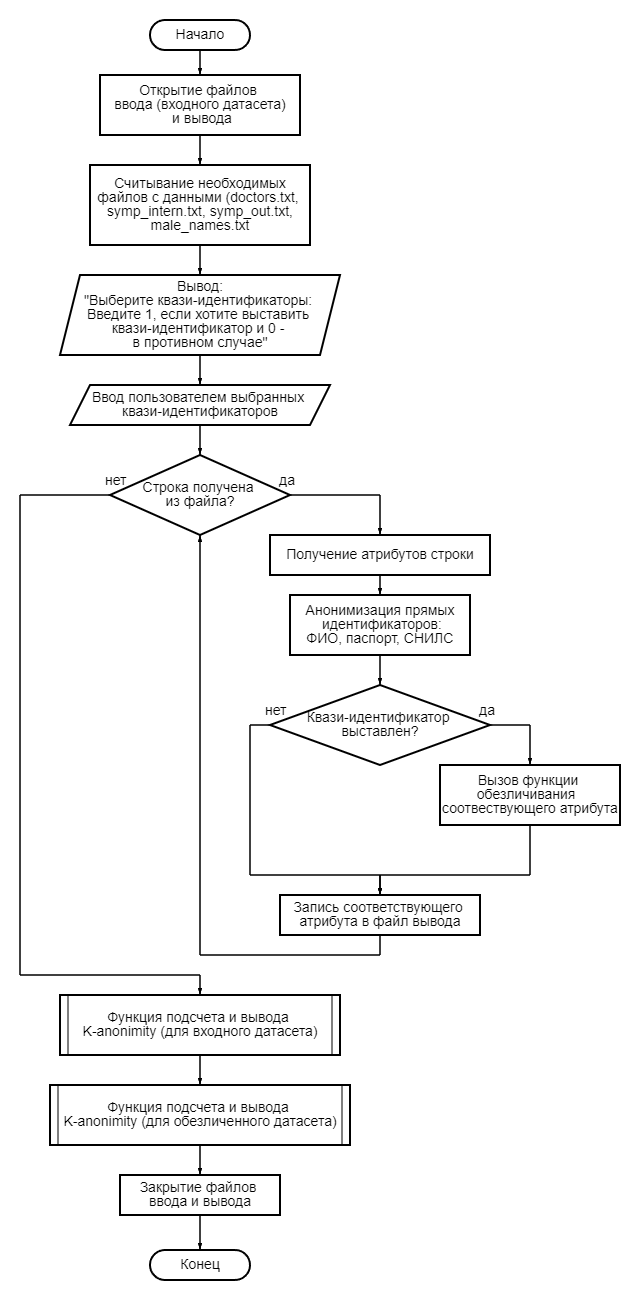
Для наиболее эффективного обезличивания данных по критерию k-anonymity и сохранению их полезности в работе выбраны следующие методы к атрибутам входящих записей:

* ФИО – «локальное обобщение». Разделение пациентов на группы: мужчина/женщина позволяет полностью убрать информацию о ФИО пациента (являющееся прямым идентификатором личности), оставив при этом имеющую значение для мед. учреждения информацию о поле пациента.
* Паспорт – «локальное обобщение». Деление пациентов на группы по гражданству позволяет исключить информацию о серии и номере паспорта пациента.
* СНИЛС – «маскеризация». Полное уничтожение информации о номере СНИЛСА пациента (XXX-XXX-XXX XX)
* Симптомы – «локальное обобщение». Симптомы делятся на две группы: «внешние признаки» и «боль во внутренних органах». Это позволяет сохранить медицинскую специфику данных этого атрибута, уменьшив при этом его детализацию.
* Врач – «локальное обобщение». Врачи делятся на три группы: "Внутренняя медицина и смежные специальности", "Хирургические и интервенционные специальности" и "Специализация в области диагностики, профилактики и оказания вспомогательной помощи". Это сохраняет полезность данных в анализе специфики работы каждой отдельной группы врачей (например, стоимости анализов или характере симптомов), при этом исключая информацию о посещении конкретного врача.
* Дата посещения – «локальное обобщение». Год посещения заменяется диапазоном лет работы поликлиники. Месяц заменяется временами года: «Зима/Лето» или «Осень/Весна», что позволяет проследить за изменениями в работе поликлинике в зависимости от времени года (за частотой посещения определенной группы врачей, например).
* Анализы – «локальное обобщение». Анализы заменяются их количеством. Это позволяет исключить информацию о редко назначающихся анализах, из-за которых можно идентифицировать пациента. При этом сохраняется информация, с помощью которой можно примерно рассчитывать стоимость одного анализа у разных групп врачей.
* Дата анализов – «локальное обобщение». Конкретная дата заменяется длительностью ожидания анализов относительно даты посещения. Это позволяет клинике следить за эффективностью работы подразделений и не хранить излишне детализированную информацию о точном времени получения.
* Стоимость – «локальное обобщение». Заменяется диапазонами «2000-6000 руб» и «6000-10000 руб».
* Банковская карта – «локальное обобщение». Полностью исключается конфиденциальная информация о номере карты. Остается информация о платежной системе и банке (то есть остается ясна общая картина финансов пациентов).

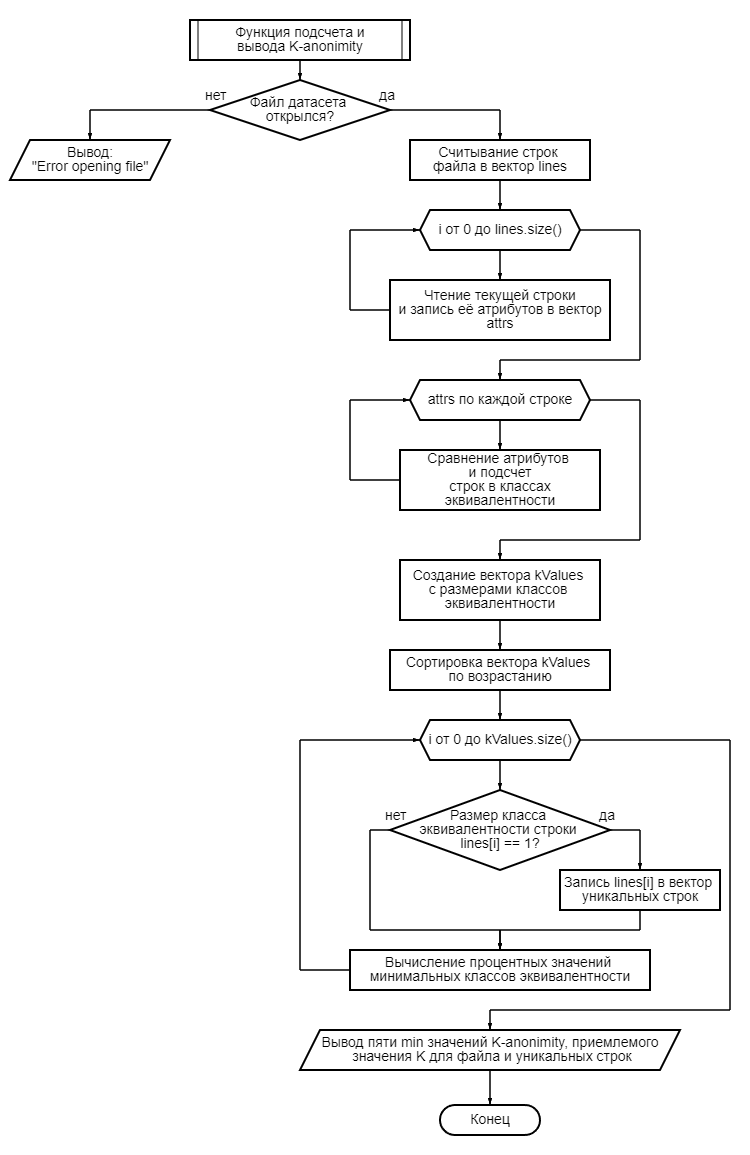
# **Описание алгоритма**

1. Открытие файлов ввода и вывода;
2. Запись необходимых данных из вспомогательных файлов (группы симптомов, врачей, мужские имена);
3. Ввод пользователем квази-идентификаторов;
4. Анонимизация прямых идентификаторов: ФИО, паспорт, СНИЛС;
5. Обезличивание других атрибутов согласно указанным квази-идентификаторам;
6. Вычисление и вывод k-anonymity двух датасетов: исходного и обезличенного. Производится в несколько этапов: считывание строк, сравнение атрибутов, подсчет размеров классов эквивалентности, вывод 5-ти min значений k-anonymity и их процента от общего числа строк датасета, вывод приемлемого значения k-anonymity для файлов определенного размера для анализа эффективности полученных результатов обезличивания;
7. Закрытие файлов ввода и вывода.

*Блок-схемы, иллюстрирующие шаги алгоритма показаны на рисунках 4.1-4.2.*



*Рисунок 4.1. Блок-схема основной программы.*



*Рисунок 4.2. Блок-схема подпрограмм.*

# **Описание программы**

## **Описание функций**

В программе используется 14 функций. В таблице 5.1 представлено их описание.

*Таблица 5.1. Описание функций*

|  |  |  |
| --- | --- | --- |
| Имя функции | Входные данные | Описание функции |
| GetRandomInt | Максимальное и минимальное число | Получение случайного числа в указанном диапазоне |
| GetVectorFromFile | Файл, из которого берутся данные;  Вектор, в который записываются данные | Запись данных из файла в вектор (по строкам) |
| calc\_K\_anonimity | Имя файла, для которого рассчитывается k-anonymity | Вычисление k-anonymity и вывод его пяти минимальных значений (с процентами) |
| extract\_first\_name | ФИО пациента | Получение имени пациента |
| anon\_fio | ФИО пациента, название файла с мужскими именами | Обезличивание ФИО пациента |
| anon\_pass | Паспорт пациента | Обезличивание паспорта пациента |
| anon\_snils | СНИЛС пациента | Обезличивание СНИЛСа пациента |
| anon\_symp | Симптомы пациента, вектор «внешних симптомов», вектор «внутренних симптомов» | Обезличивание симптомов пациента |
| anon\_doctor | Лечащий врач, вектор врачей клиники, векторы с индексами 1-ой, 2-ой и 3-ей групп врачей | Обезличивание врача |
| anon\_date | Дата посещения врача | Обезличивание даты посещения |
| anon\_an | Назначенные анализы | Обезличивание анализов |
| anon\_date\_an | Дата посещения, дата получения анализов | Обезличивание даты получения анализов |
| anon\_price | Стоимость анализов | Обезличивание цены анализов |
| anon\_card | Банковская карта пациента | Обезличивание карты |

Описание переменных функции main() представлено в таблице 5.2.

*Таблица 5.2. Описание переменных функции main()*

|  |  |  |
| --- | --- | --- |
| Тип | Имя | Описание |
| std::ifstream | file | Файл ввода |
| std::ofstream | fout | Файл вывода |
| std::vector<std::string> | doctors\_vector | Вектор докторов |
| std::vector<std::string> | symp\_intern\_vector | Вектор «внутренних симптомов» |
| std::vector<std::string> | symp\_out\_vector | Вектор «внешних симптомов» |
| std::vector<std::string> | vec\_male\_names | Вектор мужских имён |
| std::vector<int> | intern | Вектор индексов врачей группы «Внутренняя медицина и смежные специальности» |
| std::vector<int> | surg | Вектор индексов врачей группы «Хирургические и интервенционные специальности» |
| std::vector<int> | diag | Вектор индексов врачей группы «Специализация в области диагностики, профилактики и оказания вспомогательной помощи» |
| std::string | line | Переменная, хранящая текущую строку при проходе файла |
| std::string | fio, pass, snils, symp, doctor, date, an, date\_an, price, card | Переменные, хранящие атрибуты пациента |
| int | kv\_symp, kv\_doctor, kv\_date, kv\_an, kv\_date\_an, kv\_price, kv\_card | Переменные, хранящие данные о выставленных квази-идентификаторах |

# **Рекомендации пользователю**

Файл с исходным датасетом должен называться outout.csv

Убедитесь, что файлы с необходимыми исходными данными находятся в одной директории с проектом.

Обратите внимание, что при установлении квази-идентификаторов можно указывать только цифры 0 и 1 (если будет указано иное значение атрибут по умолчанию не будет обезличиваться). После ввода нужного значения нажимайте клавишу *Enter*.

Датасет будет сохранен в файл depersonal.csv в директории проекта. Если у вас возникают неполадки с кодировкой при открытии файла csv в Excel, настройте кодировку utf-8 в Excel. Для этого зайдите в *Данные->Получить данные->Из файла->Из текстового/csv файла*. Выберите файл outout.csv нажмите *Импорт* и настройте кодировку *utf-8*. Разделительный знак по умолчанию «;».

1. **Рекомендации программисту**

Для работы с программой должен быть установлен компилятор языка С++, а также сборщик проектов (CMake) или же интегрированная среда разработки (Visual Studio). Старайтесь поддерживать актуальные версии компиляторов и сред разработки. Поддерживайте актуальный стандарт языка (необходим стандарт не ниже С++11). Для запуска программы необходима 64-битная операционная система Windows. Все используемые в коде библиотеки входят в стандартную библиотеку С++, поэтому нет необходимости в дополнительной установке библиотек.

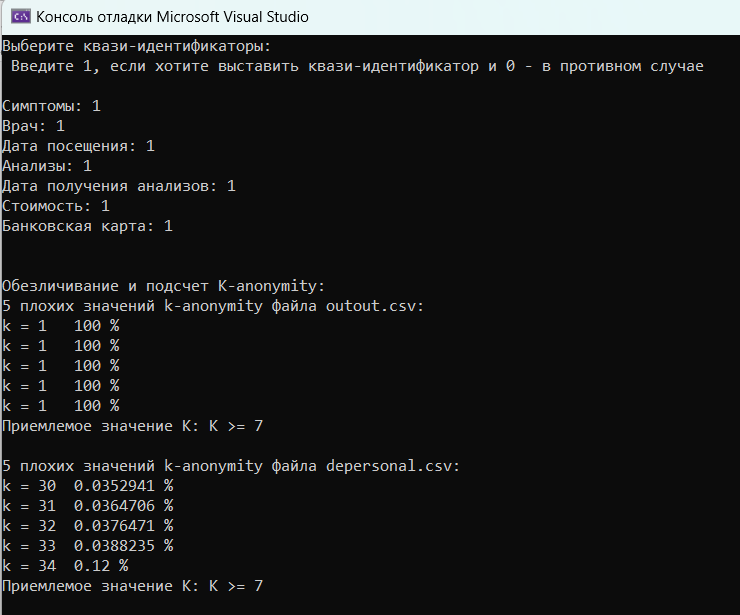
1. **Контрольный пример**

В этом разделе продемонстрирован контрольный пример, показывающий вид консоли и файла depersonal.csv при указании всех атрибутов в качестве квази-идентификаторов.

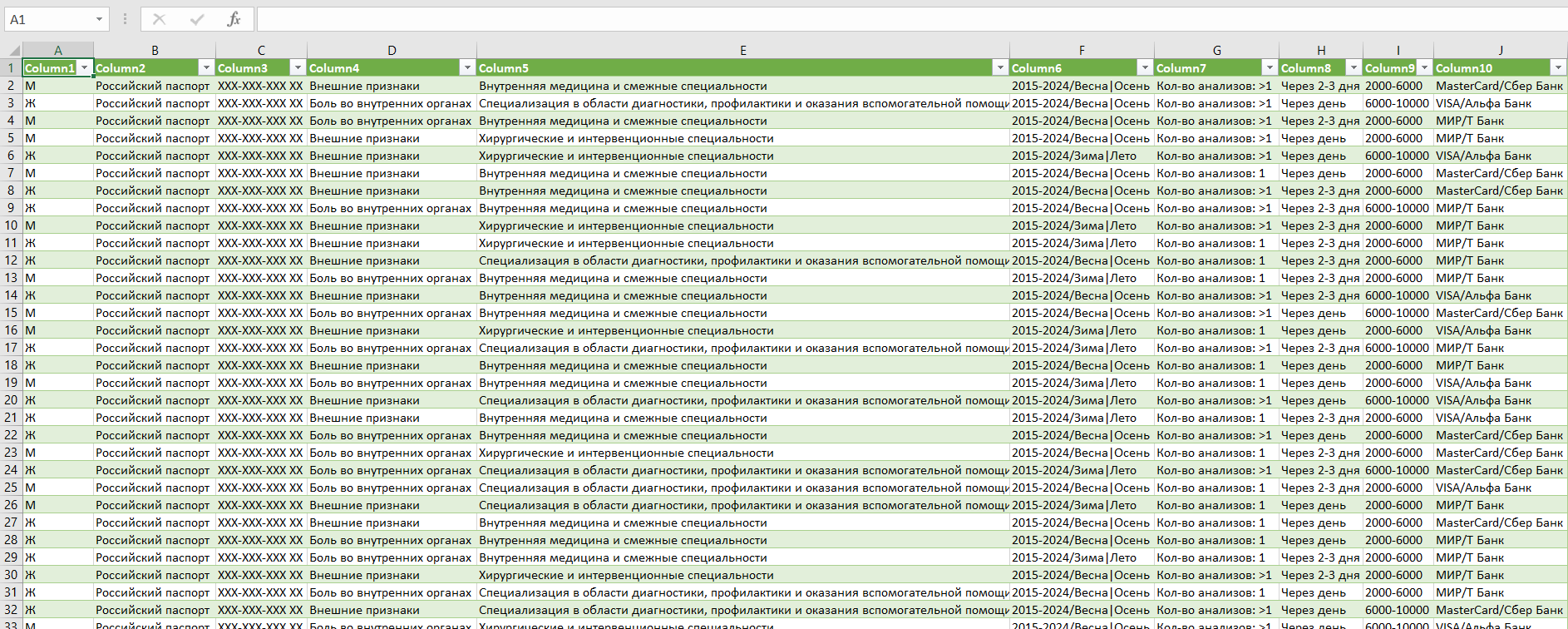
После ввода нужного значения для квази-идентификаторов нужно нажать клавишу *Enter*. После ввода последнего значения появится надпись «Обезличивание и подсчет K-anonymity» и останется только дождаться результата работы программы.

*Пример вида консоли после завершения программы показан на рисунке 8.1.*

*Пример обезличенного датасета показан на рисунке 8.2.*



*Рисунок 8.1 Вид консоли*



*Рисунок 8.2 Обезличенный датасет*

1. **Вывод**

В ходе работы были приобретены базовые навыки в методологии обезличивания данных, реализована функция вычисления k-anonymity. Кроме этого, была проанализирована степень эффективности обезличивания по критерию k-anonymity и оценка полезности данных файла depersonal.csv.