**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

**“Большие данные и распределенная цифровая платформа”**

**ОТЧЕТ**

**по лабораторной работе №1**

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

**на тему «Обезличивание данных»**

**Студент гр. 23Б15-пу**

**Абрахин Е.Д.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2024 г.**

Оглавление

1. [Цель работы 3](#_Toc177901232)
2. [Описание задачи (формализация задачи) 3](#_Toc177901233)
3. [Теоретическая часть 4](#_Toc177901234)
4. [Основные шаги программы 6](#_Toc177901235)
5. [Блок схема программы 8](#_Toc177901236)
6. [Описание программы 10](#_Toc177901237)
7. [Рекомендации пользователя 12](#_Toc177901238)
8. [Рекомендации программиста 12](#_Toc177901239)
9. [Исходный код программы 13](#_Toc177901240)
10. [Контрольный пример 14](#_Toc177901241)
11. [Вывод 19](#_Toc177901242)
12. [Источники 19](#_Toc177901243)

# Цель работы

Цель данной лабораторной работы заключается в разработке программы для анонимизации датасетов железнодорожных билетов и анализа их K-анонимности. Программа должна обеспечивать защиту персональных данных пассажиров, применяя различные методы анонимизации, и вычислять показатели K-анонимности, указывающие на возможные уязвимости данных.

# Описание задачи (формализация задачи)

Задача заключается в разработке программы для анонимизации данных и оценки их K-анонимности. Программа должна уметь считывать входные данные, которые могут содержать конфиденциальную информацию, и выполнять их обезличивание, применяя различные методы анонимизации. Пользователь должен иметь возможность выбирать квази-идентификаторы, которые будут использоваться для анонимизации и анализа данных. После анонимизации программа вычисляет показатель K-анонимности, демонстрируя, насколько эффективно данные были обезличены.

# Теоретическая часть

В данной работе реализована программа для обработки и анонимизации датасетов с использованием различных методов защиты данных, направленных на достижение К-анонимности. К-анонимность гарантирует, что каждая запись в анонимизированном наборе данных не может быть однозначно идентифицирована среди группы из K записей. Это минимизирует вероятность утечки конфиденциальной информации при публикации или анализе данных.

Основные компоненты программы

1. Загрузка и обработка данных. Программа работает с данными в формате XML. Она считывает входные файлы, содержащие информацию о пассажирах, таких как ФИО, паспортные данные, точки отправления и прибытия, даты поездок, стоимость билетов и данные платежных карт.
2. Обезличивание данных. Для защиты конфиденциальности применяется ряд методов анонимизации. Пользователь может выбрать набор квази-идентификаторов — атрибутов, которые могут быть использованы для идентификации личности, но не являются уникальными. Основные используемые методы:
   * Локальное обобщение: Детализированные данные заменяются более общими значениями. Например, даты могут быть преобразованы в годы, а стоимость билета — в категорию (низкая, средняя, высокая).
   * Маскеризация: Чувствительные данные, такие как паспортные данные и номера карт, скрываются частично или полностью. Например, паспортные данные заменяются на "XXXX XXXXXX", а данные банковских карт — на название банка.
   * Удаление атрибутов: Некоторые атрибуты могут быть полностью удалены из датасета, чтобы снизить риск утечки информации.
   * Локальное подавление: Строки, которые остаются уникальными или сильно выбиваются из общего числа записей (то есть имеют низкий показатель K), могут быть удалены для соблюдения К-анонимности.
3. Расчет К-анонимности. К-анонимность — это мера того, насколько обезличен набор данных. Для вычисления К-анонимности программа группирует записи на основе выбранных квази-идентификаторов и подсчитывает количество строк в каждой группе. Минимальный размер группы определяет значение К. Программа также выводит пять "плохих" значений K, чтобы продемонстрировать К-анонимность после анонимизации более подробно.
4. Полезность данных. Обезличивание неизбежно приводит к потере некоторой полезной информации, что может снизить точность анализа данных. Программа старается сохранить полезность данных, анонимизируя исходный наборы данных различными методами, удаляя личную информацию и оставляя общую. Это позволяет в последующем оценить, насколько примененные методы анонимизации повлияли на структуру и информативность набора данных.

# Основные шаги программы

1. Запуск программы

Программа запускается с использованием библиотеки tkinter, которая открывает графический интерфейс для работы с данными о железнодорожных билетах и расчетами K-анонимности.

2. Загрузка данных о городах

При запуске программа загружает данные о городах из файла russian\_cities.csv через функцию loadCitiesData. Эти данные используются для категоризации городов в билетах.

3. Расчет K-анонимности

При нажатии на кнопку "Посчитать K-анонимити", программа:

* + Пользователь выбирает файл XML с данными о билетах, используя функцию filedialog.askopenfilename.
  + Загружает данные о билетах из выбранного XML-файла.
  + Преобразует данные в pandas.DataFrame для дальнейшей обработки.
  + Пользователь выбирает квази-идентификаторы (поля, по которым будут рассчитываться группы для K-анонимности), такие как ФИО, паспортные данные, место отправления, дата отправления, стоимость билета и другие.
  + Программа группирует данные по выбранным идентификаторам и вычисляет K-анонимность — минимальный размер группы, в которой каждый элемент неотличим от других.
  + Результаты K-анонимности выводятся на экран, включая информацию о записях, где K-анонимность равна 1 (риск идентификации).

4. Анонимизация данных

При нажатии на кнопку "Обезличить датасет", программа:

* + Пользователь выбирает файл XML с данными о билетах, используя функцию filedialog.askopenfilename.
  + Загружает выбранный XML-файл с билетами.
  + Анонимизирует данные:
    - ФИО заменяются на обозначение пола ("М" для мужчин и "Ж" для женщин), на основе окончания фамилии.
    - Паспортные данные маскируются: вместо реальных данных используется "XXXX XXXXXX".
    - Даты отправления и прибытия категоризируются по годам.
    - Названия городов заменяются на федеральные округа (если данные о городе есть в загруженном файле городов).
    - Номера поездов разделяются на группы (пассажирские, скоростные, высокоскоростные).
    - Места в поездах и вагоны заменяются на "XX".
    - Платежные карты заменяются на банк, к которому они относятся, на основе первых шести цифр BIN-кода (определяется по словарю binCodesBank).
    - Стоимость билетов категоризируется на три группы:
      * Низкая (если цена ниже 2000).
      * Средняя (от 2000 до 4000).
      * Высокая (выше 4000).
  + После того как данные анонимизированы, они собираются обратно в XML-формат с помощью функции formatXmlWithIndentation, которая добавляет отступы для читаемости.
  + Программа предлагает сохранить новый анонимизированный файл через диалоговое окно, используя filedialog.asksaveasfilename. Если пользователь сохраняет файл, он записывается на диск в формате XML.

# Блок схема программы

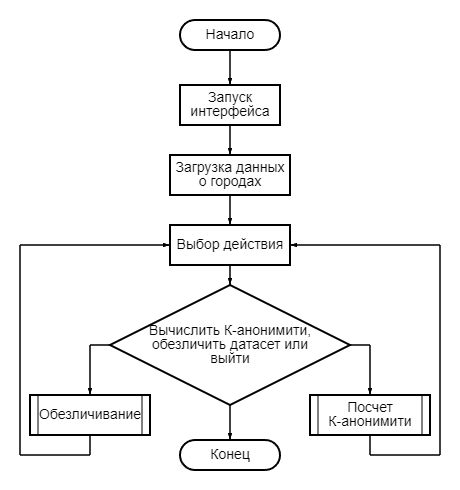


Рис 1. Блок-схема основной программы

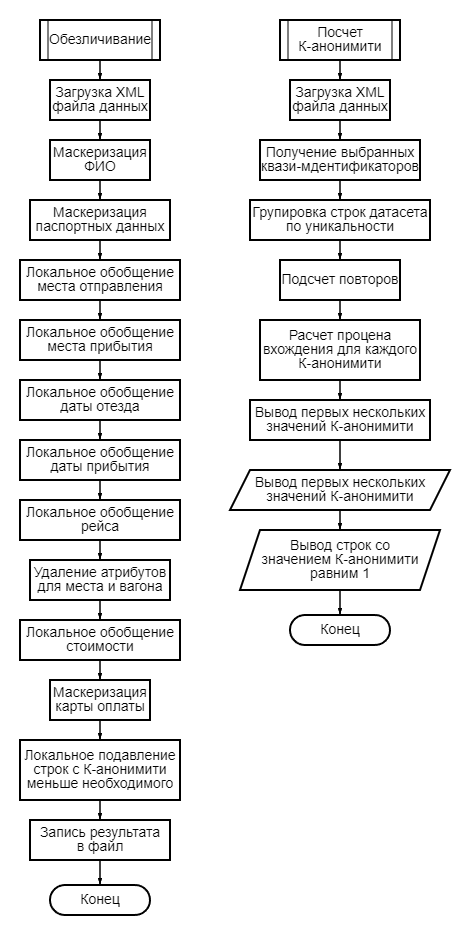


Рис 2. Блок-схема подпрограмм

# Описание программы

Программная реализация написана на языке Python 3.12.6 с использованием следующих библиотек: tkinter [[1]](#_Источники), pandas [[2]](#_Источники), и xml [[3]](#_Источники). Программа организована в единый модуль DatasetAnonymization.py с акцентом на анонимизацию данных о железнодорожных билетах и подсчет их К-анонимити. В процессе разработки программы использовались 11 функций, каждая из которых имеет чётко определённое назначение и 1 основной модуль:

Таблица 1. DatasetAnonymization.py

|  |  |  |
| --- | --- | --- |
| Функция | Описание | Возвращаемое значение |
| anonymizeDataset | Анонимизация датасета на основе правил K-анонимности. Чтение XML файла, обработка данных, замена персональных данных на анонимные. Файл сохраняется. | None |
| calculateKAnonymity | Рассчитывает уровень K-анонимности для загруженного датасета, показывает статистику по группам записей с определённой анонимностью. | None |
| categorizeCity | Категоризация городов на основе данных о федеральных округах. | str |
| categorizeCost | Категоризация стоимости поездки по диапазонам (низкая, средняя, высокая). | str |
| categorizeDate | Преобразование даты в строку с годом, чтобы уменьшить точность данных для анонимизации. | str |
| categorizeTrainNumber | Категоризация номера поезда на основе его диапазона (пассажирский, скоростной, высокоскоростной поезд). | str |
| determineGender | Определение пола пассажира по окончанию отчества (на основе правила о последней букве). | str |
| formatXmlWithIndentation | Форматирование XML с отступами для удобства чтения. | None |
| loadCitiesData | Загрузка данных о городах из CSV-файла для их категоризации по федеральным округам. | None |
| maskPaymentCard | Маскирование номера банковской карты, замена номера на название банка на основе BIN-кода. | str |
| maskPassportInfo | Маскирование паспортных данных (замена на "XXXX XXXXXX"). | str |

# 

# Рекомендации пользователя

Для успешного запуска программы убедитесь, что у вас установлен Python и необходимые библиотеки, такие как tkinter [[1]](#_Источники), pandas [[2]](#_Источники) и xml [[3]](#_Источники). Все файлы программы должны находиться в одной директории для корректного выполнения. Программа позволяет обрабатывать XML-файлы, которые содержат информацию о билетах, и предоставляет возможность анонимизации данных с учетом параметра K-анонимности. Перед началом работы запустите скрипт DatasetAnonymization.py и проверьте, не выводится ли ошибка через графический интерфейс, предоставляемый программой. Для корректной работы убедитесь, что файл russian\_cities.csv, содержащий данные о городах и федеральных округах, доступен в директории с программой. Если файл отсутствует, программа выведет сообщение об ошибке. Чтобы анонимизировать данные, выберите опцию обезличивания, затем XML файл, к которому ее нужно применить, она заменит личные данные в выбранном файле (ФИО, паспортные данные, данные о платёжных картах) и категоризирует другие значения (города, даты, номера поездов и стоимость билетов). Чтобы вывести К-анонимити, выберите опцию подсчета К-анонимити, затем XML файл, в котором его нужно посчитать, и результат выведется в интерфейс. Убедитесь, что структура выбираемого XML файла совпадает с требуемой: FullName, PassportInfo, Departure, Destination, DepartureDate, ArrivalDate, Train, SeatChoice, TotalCost, PaymentCard. Настройка квази-идентификаторов, по которым считается К-анонимити, и отображение данных, осуществляется в интерфейсе программы.

# Рекомендации программиста

Поддерживайте актуальные версии используемых библиотек и Python для обеспечения работоспособности программы на современных системах. Следуйте передовым практикам разработки (best practices), уделяйте внимание четкому именованию переменных и функций. Регулярно проводите тестирование программы на различных входных данных, чтобы убедиться в её надежности и корректности.

# Исходный код программы

<https://github.com/FasterXaos/Algorithms_and_Data_Structures>

# Контрольный пример

1. Запуск программы

Для запуска программы используйте файл **DatasetAnonymization.py.** Программа запустит графический интерфейс (Рис. 1), в котором можно будет выбрать “Обезличить датасет” или “Посчитать К-анонимити”, предварительно выбрав квази-идентификаторы и чекбоксов в столбике слева.

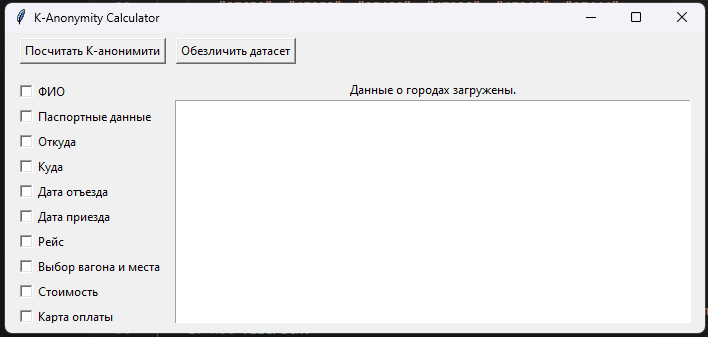


Рис 1. пример графического интерфейса программы

2. Обезличить датасет

После выбора “Обезличить датасет” программа откроет Проводник, где пользователю предложено выбрать XML файл (Рис. 2) и **после процесса анонимизации будет предложено сохранить получившийся датасет в новый или существующий файл** (Рис. 3).

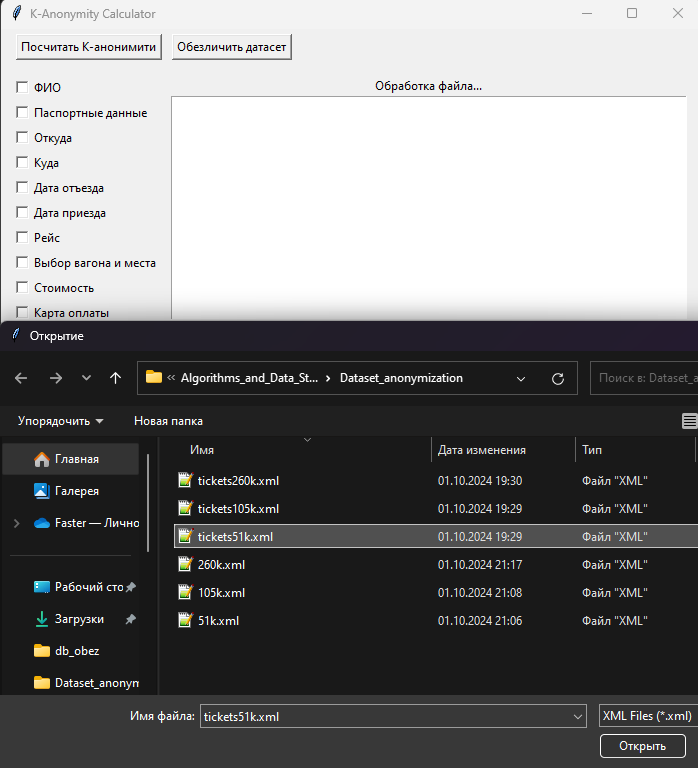


Рис 2. пример выбора файла для обезличивания

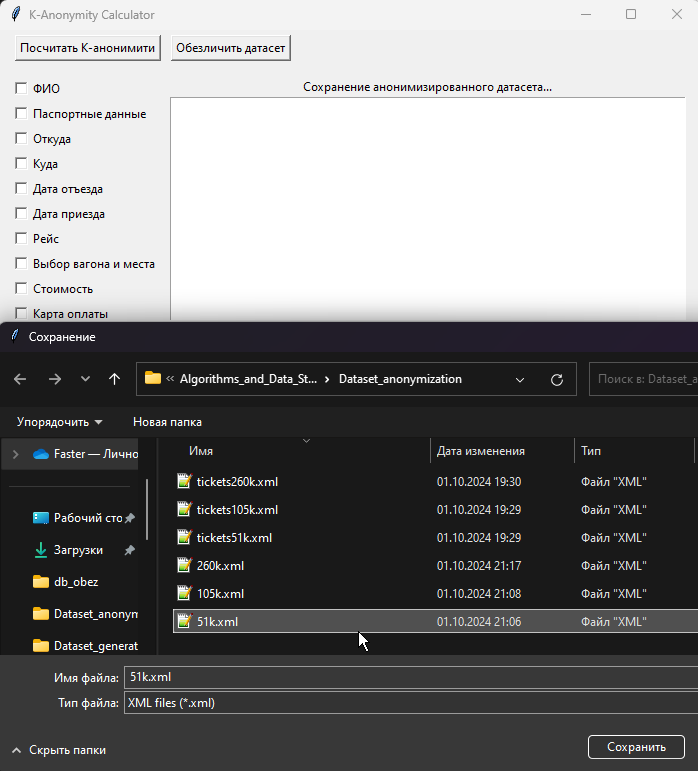


Рис 3. пример сохранения файла после обезличивания

3. Посчитать К-анонимити

После выбора квази-идентификаторов и нажатия на кнопку “Посчитать К-анонимити”, вызовется Проводник, где пользователю предложено выбрать XML файл для подсчета К-анонимити (Рис. 4). После подсчета топ пять худших значений К-анонимити с их процентом в датасете будут выведены в интерфейсе программы(Рис. 5).

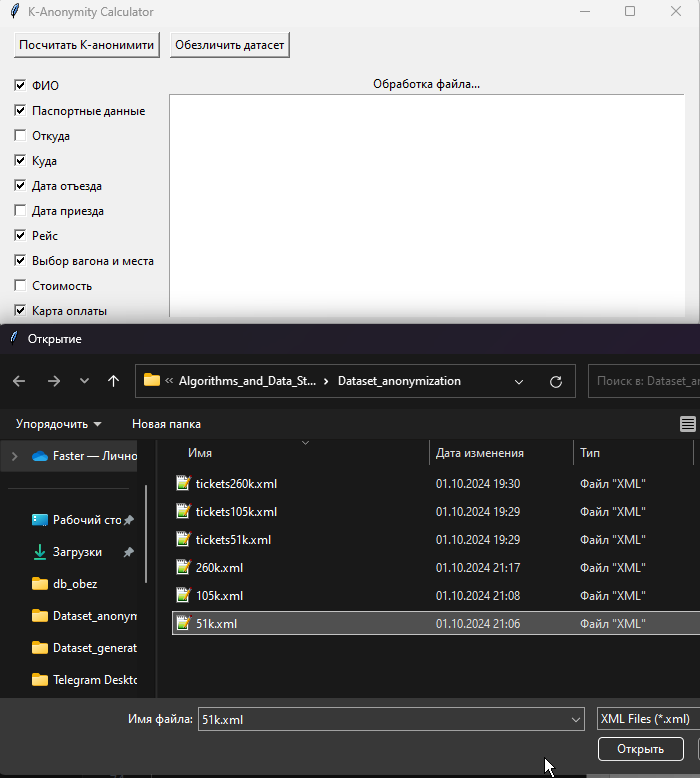


Рис 4. пример выбора файла для подсчета К-анонимити с выбранными квази-идентификаторами

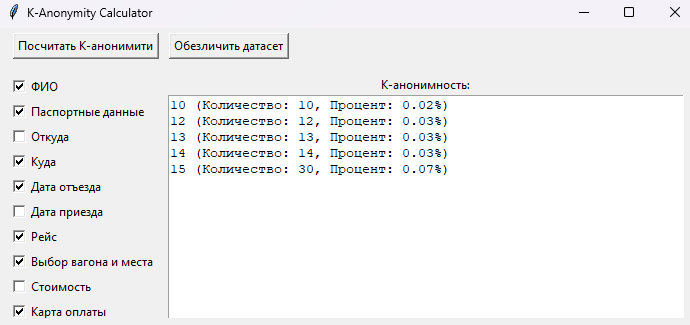


Рис 5. пример ввода после подсчета К-анонимити

# Вывод

В рамках данной лабораторной работы была реализована система для обезличивания данных на примере ж/д билетов. Основной целью было достижение приемлемой K-анонимности данных, позволяющей снизить риск идентификации конкретных пассажиров при анализе информации.

Были рассмотрены такие квази-идентификаторы, как ФИО, паспортные данные, город отправления и прибытия, даты поездки, номер поезда, выбор места, стоимость билета и данные банковских карт. Для обезличивания информации применялись методы маскирования паспортных данных и номеров банковских карт, преобразования городов в федеральные округа, а также классификация поездов и стоимости билетов.

Программа была разработана на языке Python с использованием библиотеки Tkinter для создания графического интерфейса пользователя, который позволяет загружать и сохранять XML файлы с данными, выбирать квази-идентификаторы для анализа, а также рассчитывать уровень K-анонимности.

# Источники

1. tkinter — Python interface to Tcl/Tk // Tkinter URL:  [https://docs.python.org/3/library/tkinter.html/](%20https://docs.python.org/3/library/tkinter.html/) (дата обращения: 1.10.2024).
2. pandas // Pandas URL: <https://pandas.pydata.org/> (дата обращения: 1.10.2024).
3. XML Processing Modules // Python URL: <https://docs.python.org/3/library/xml.html> (дата обращения: 18.09.2024).