**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

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

**ОТЧЕТ**

**по лабораторной работе №10**

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

**на тему «Кластеризация данных**»

**Студент гр. 23Б15-пу**

**Бек В.А.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

Оглавление

1. [Цель работы 3](#_Toc198923171)
2. [Описание задачи (формализация задачи) 4](#_Toc198923172)
3. [Теоретическая часть 5](#_Toc198923173)
4. [Основные шаги программы 8](#_Toc198923174)
5. [Блок схема программы 9](#_Toc198923175)
6. [Описание программы 10](#_Toc198923176)
7. [Рекомендации пользователя 12](#_Toc198923177)
8. [Рекомендации программиста 13](#_Toc198923178)
9. [Контрольный пример 14](#_Toc198923179)
10. [Анализ работы 17](#_Toc198923180)
11. [Вывод 19](#_Toc198923181)
12. [Источники 20](#_Toc198923182)

# Цель работы

**Цель работы** — изучение метода кластеризации данных при помощи алгоритма ISODATA, выбора наиболее информативных признаков с использованием метода последовательного добавления, а также применение метрики расстояния Чебышева и такого критерия оценки качества кластеризации как индекс Фоулкса – Мэллова. Результатом работы станет программа позволяющая кластеризовать данные и анализ успешности данной кластеризации.

# Описание задачи (формализация задачи)

В данной лабораторной работе требуется исследовать влияние выбора признаков и обезличивания на качество кластеризации данных. Подробные пункты решения задачи:

1. **Формирование датасета:** Подготовить исходный датасет с не менее чем 15 признаками.
2. **Обычная кластеризация:** Применить алгоритм ISODATA (с расстоянием Чебышева) к полному набору признаков.
3. **Оценка качества:** Вычислить индекс Фоулкса–Мэллова между истинными значениями и полученными.
4. **Отбор признаков:** При помощи метода последовательного добавления (add) выбрать n наиболее информативных признаков.
5. **Кластеризация на отобранных признаках:** Повторить шаг 2 на уменьшенном наборе признаков.
6. **Оценка качества после отбора:** Повторить шаг 3 на уменьшенном наборе признаков.
7. **Сравнение результатов:** Сравнить результаты до и после отбора признаков, сделать выводы.
8. **Обезличивание данных:** Обезличить датасет до приемлемого k-anonymity.
9. **Кластеризация после обезличивания:** Повторить шаг 2 на обезличенном датасете.
10. **Оценка качества после обезличивания:** Повторить шаг 3 для обезличенного датасета.
11. **Итоговое сравнение:** Сравнить результаты для трёх этапов:  
    • обычная кластеризация  
    • после отбора признаков  
    • после обезличивания

# Теоретическая часть

#### **Кластеризация:**

**Кластеризация** — разделение множества объектов на группы (кластеры) таким образом, чтобы объекты внутри одного кластера были максимально похожи между собой, а объекты из разных кластеров — максимально различны.

#### **Алгоритмы ISODATA:**

**ISODATA** (Iterative Self-Organizing Data Analysis Technique Algorithm) — итеративный алгоритм кластеризации, расширяющий возможности алгоритма k-средних. Его ключевая особенность — способность динамически изменять количество кластеров путём их объединения или разбиения в процессе работы.

1. **Инициализация**:
   * Задаётся начальное число кластеров​.
   * Случайным образом выбираются начальные центры кластеров или используются заранее заданные.
2. **Назначение объектов кластерам**:
   * Каждый объект присваивается ближайшему кластеру согласно выбранной метрике расстояния.
3. **Пересчёт центров кластеров**:
   * Центр каждого кластера пересчитывается как среднее значение всех объектов, принадлежащих данному кластеру.
4. **Разбиение кластеров**:
   * Кластеры с большим разбросом объектов разделяются на два новых кластера.
5. **Объединение кластеров**:
   * Если расстояние между центрами двух кластеров малое, эти кластеры объединяются.
6. **Проверка критерия останова**:
   * Алгоритм останавливается, если достигнуто максимальное число итераций, либо центры кластеров больше не изменяются.

#### **Расстояние Чебышева:**

**Расстояние Чебышева** — это метрика расстояния между двумя точками в n-мерном пространстве, определяемая как:

То есть это максимальное из модулей разностей соответствующих координат. Эту метрику часто используют, когда интересует наиболее значимое отклонение по какому-либо одному признаку. Она особенно полезна, когда объекты следует считать "разными", если хотя бы по одному параметру наблюдается большое расхождение.

#### **Метод последовательного добавления признаков (алгоритм add):**

**Метод последовательного добавления признаков (Additive Feature Selection)** — это жадный алгоритм отбора признаков, который на каждом шаге добавляет к текущему подмножеству тот признак, который в сочетании с уже выбранными даёт наилучший результат по заданной метрике.

1. Начать с пустого множества признаков.
2. Последовательно добавлять по одному признаку, каждый раз выбирая тот, при добавлении которого достигается наилучшее качество кластеризации.
3. Повторять до достижения заданного числа признаков или пока не перестанет улучшаться результат.

Метод позволяет обнаружить наиболее информативные признаки для задачи кластеризации, что может повысить точность и интерпретируемость модели.

#### **Индекс Фоулкса – Мэллоуза:**

**Индекс Фоулкса – Мэллоуза (FMI)** — это метрика оценки качества кластеризации при наличии истинных результатов. Он измеряет сходство между двумя разбиениями множества определяется как:

где:

* **TP** — число пар объектов, принадлежащих к одному кластеру как в истинной, так и в предсказанной кластеризации;
* **FP** — число пар, находящихся в одном предсказанном кластере, но в разных истинных классах;
* **FN** — число пар, находящихся в одном истинном классе, но в разных предсказанных кластерах.

Значение FMI лежит в диапазоне от 0 до 1. Чем ближе значение к 1, тем выше качество кластеризации. FMI симметричен и устойчив к изменению числа кластеров, что делает его удобным для сравнительного анализа.

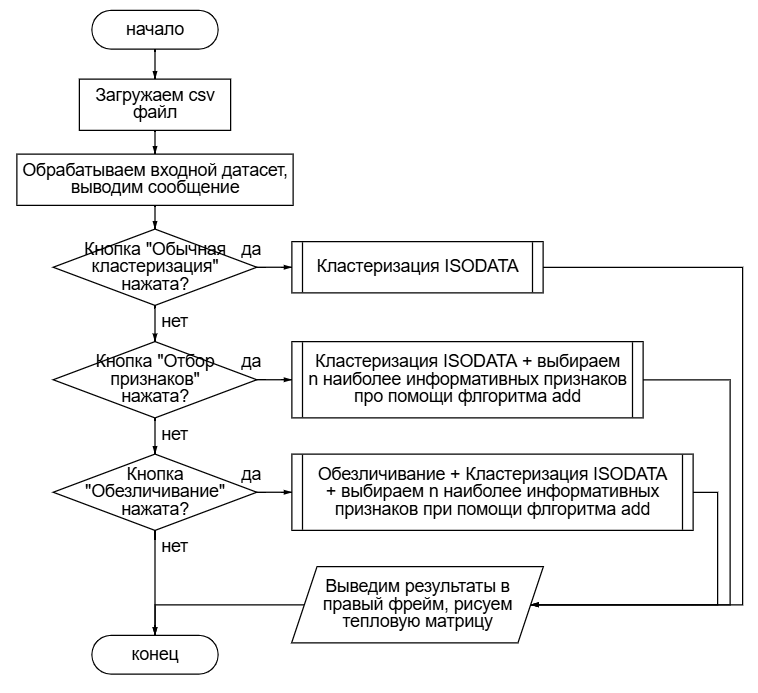
#### **PCA:**

**PCA (Principal Component Analysis)** — это метод, который **уменьшает число признаков в данных**, но при этом старается **сохранить как можно больше информации**. PCA "поворачивает" данные в пространстве так, чтобы найти **главные направления разброса.** Эти главные направления называются **главные компоненты.**

# Основные шаги программы

* **Запуск и отображение графического интерфейса**: С помощью библиотеки tkinter создаётся графическое окно, которое предоставляет пользователю интерфейс для загрузки данных, выполнения кластеризации, отбора признаков и обезличивания.
* **Загрузка и предобработка данных:** Выбираем CSV-файл, который автоматически очищается от нечисловых и неинформативных столбцов (например, Index), заполняются пропущенные числовые значения средними, а категориальные данные кодируются числами.
* **Кластеризация без отбора признаков**: Данные передаются в алгоритм ISODATA с использованием расстояния Чебышева для разделения объектов на два кластера. После этого рассчитывается индекс Фоулкса–Мэллоуза для оценки качества кластеризации и отображается матрица ошибок.
* **Кластеризация с отбором признаков (метод add)**: Программа применяет метод последовательного добавления признаков (алгоритм Add), выбирая наиболее информативные признаки, улучшая качество кластеризации.
* **Обезличивание данных и кластеризация**: Выполняется удаление самых значимых признаков (на основе алгоритма Add), затем производится выбор m признаков. После чего проводится кластеризация и расчёт индекса Фоулкса–Мэллоуза.
* **Оценка качества кластеризации**: Для всех вариантов рассчитывается **индекс Фоулкса–Мэллоуза**, который измеряет согласованность между истинными метками и результатами кластеризации. Также отображается **матрица ошибок** в виде тепловой карты.

# Блок схема программы

Рис. 1 Блок-схема основной программы

# Описание программы

Программная реализация задачи кластеризации данных выполнена на языке **Python 3.12.7** с использованием библиотек **tkinter, pandas, numpy, sklearn, matplotlib, seaborn**. Программа представляет собой графический интерфейс, позволяющий загружать датасет в формате CSV, выполнять кластеризацию данных методом ISODATA, оценивать качество кластеризации по индексу Фоулкса–Мэллоуза, а также проводить отбор признаков и обезличивание данных с учетом k-анонимности.

Таблица 1. clusterization.py

|  |  |  |
| --- | --- | --- |
| **Функция** | **Описание** | **Возвращаемое значение** |
| load\_csv() | Загрузка CSV-файла, очистка, кодировка и обработка данных | None |
| cluster\_and\_evaluate() | Выполнение ISODATA-кластеризации и оценка результата без отбора признаков | None |
| select\_features\_cluster\_evaluate() | Выполнение кластеризации после отбора n наиболее важных признаков | None |
| anonymize\_cluster\_evaluate() | Проведение обезличивания данных с k-анонимностью, затем кластеризация | None |
| evaluate(features, labels) | Вычисление индекса Фоулкса–Мэллоуза и отображение матрицы ошибок | float (значение метрики) |
| isodata(data, k) | Реализация алгоритма ISODATA на основе расстояния Чебышёва | np.ndarray (массив меток кластеров) |
| feature\_selection\_add(X, y, n\_features) | Последовательный отбор признаков по максимизации метрики кластеризации | list (индексы отобранных признаков) |
| show\_confusion\_matrix(true\_labels, pred\_labels) | Визуализация матрицы ошибок | None |

# Рекомендации пользователя

Для корректного выполнения программы следуйте приведённым шагам:

* **Загрузка данных**: Нажмите кнопку «Загрузить CSV» в левой панели управления. Выберите файл формата .csv, содержащий исходные данные.
* **Обычная кластеризация**: Нажмите кнопку «Обычная кластеризация», чтобы запустить алгоритм кластеризации ISODATA на всех признаках без дополнительной обработки.
  + В правом текстовом поле отобразятся:
  + количество объектов и признаков
  + индекс Фоулкса–Мэллоуза
  + время выполнения
  + визуализация матрицы ошибок
* **Кластеризация с отбором признаков:** Установите желаемое количество признаков для отбора в поле «Признаков для отбора» (по умолчанию — 5). Нажмите кнопку «Отбор признаков». Программа выполнит отбор наиболее информативных признаков по методу последовательного добавления и проведёт кластеризацию на отобранных признаках.
  + В правом текстовом поле отобразятся:
  + количество объектов и признаков
  + индекс Фоулкса–Мэллоуза
  + время выполнения
  + визуализация матрицы ошибок
* **Кластеризация после обезличивания:** Установите количество признаков для обезличивания в поле «Признаков для обезличивания» (по умолчанию — 10). Нажмите кнопку «Обезличивание».  
  Программа проведёт кластеризацию на обезличенных данных и отобразит:
  + индекс Фоулкса–Мэллоуза
  + значение минимальной k-анонимности
  + время выполнения
* **Вывод результатов:** Все результаты отображаются в правом текстовом поле. Матрица ошибок появляется в отдельном окне после каждого расчёта.

# Рекомендации программиста

Для корректного функционирования программы рекомендуется выполнить следующие действия:

1. **Установите необходимые библиотеки**:
   * Убедитесь, что у вас установлены библиотеки **tkinter, math, ttk** и **tkinter.simpledialog, sklearn**. **thinker** идет в стандартной поставке Python, но если она отсутствует, вы можете установить ее через пакетный менеджер вашей операционной системы.
2. **Проверьте версию Python**:
   * Рекомендуется использовать **Python** версии **3.1** или выше, чтобы избежать возможных проблем с совместимостью библиотек и функциональностью **tkinter**.
3. **Проверка функций и интерфейса**:
   * Убедитесь, что все элементы интерфейса (кнопки, текстовые поля и таблицы) работают корректно. Попробуйте вводить различные значения и проверить, что результаты рассчитываются и отображаются правильно.

**Код программы:**

**<https://github.com/Kliooo/Algorithms-and-data-structures>**

**Датасет:**

[**https://www.kaggle.com/datasets/vanthanadevi08/water-quality-prediction**](https://www.kaggle.com/datasets/vanthanadevi08/water-quality-prediction)

# Контрольный пример

* **Запуск программы:** Для запуска программы используйте файл clusterization.py. Программа откроет графический интерфейс (Рис. 2).

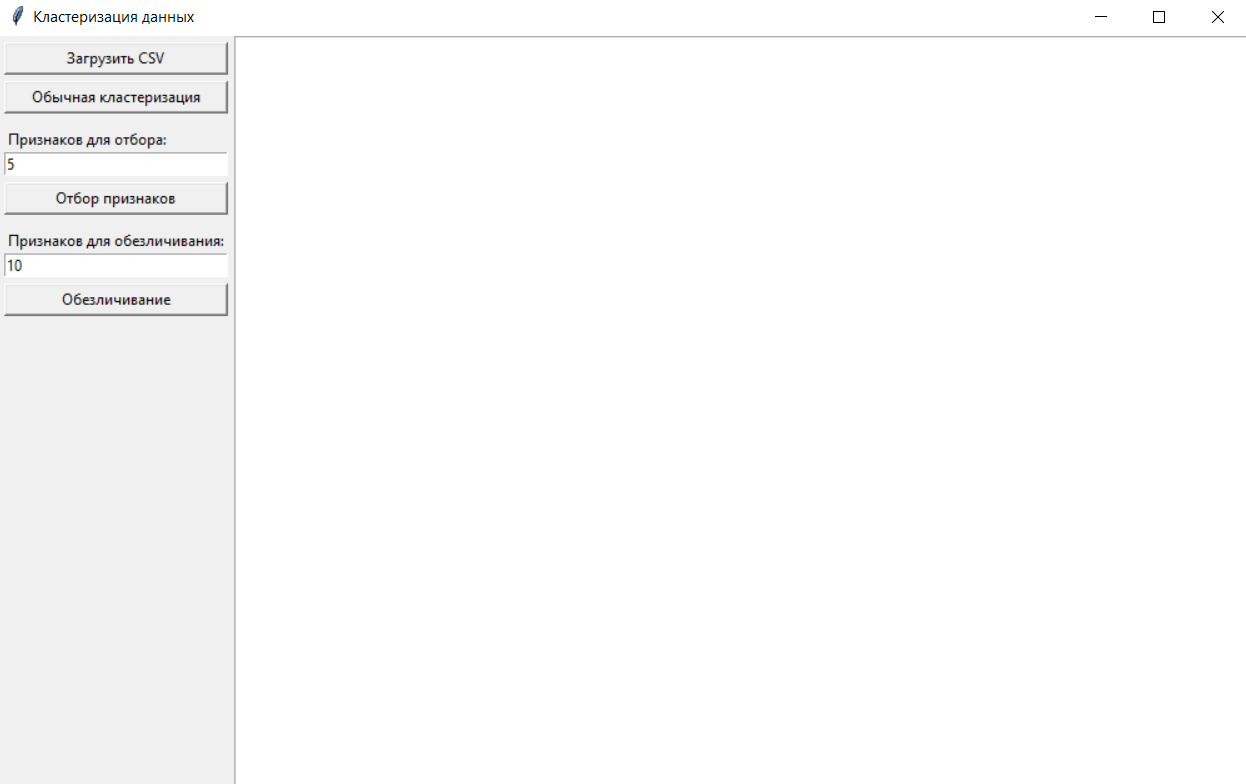


Рис. 2 Графический интерфейс программы

* **Выбор датасета: Нажмите кнопку «Загрузить CSV»** и выберите исходный датасет, программа обработает его и выведет сообщение в правый фрейм. (Рис. 3)

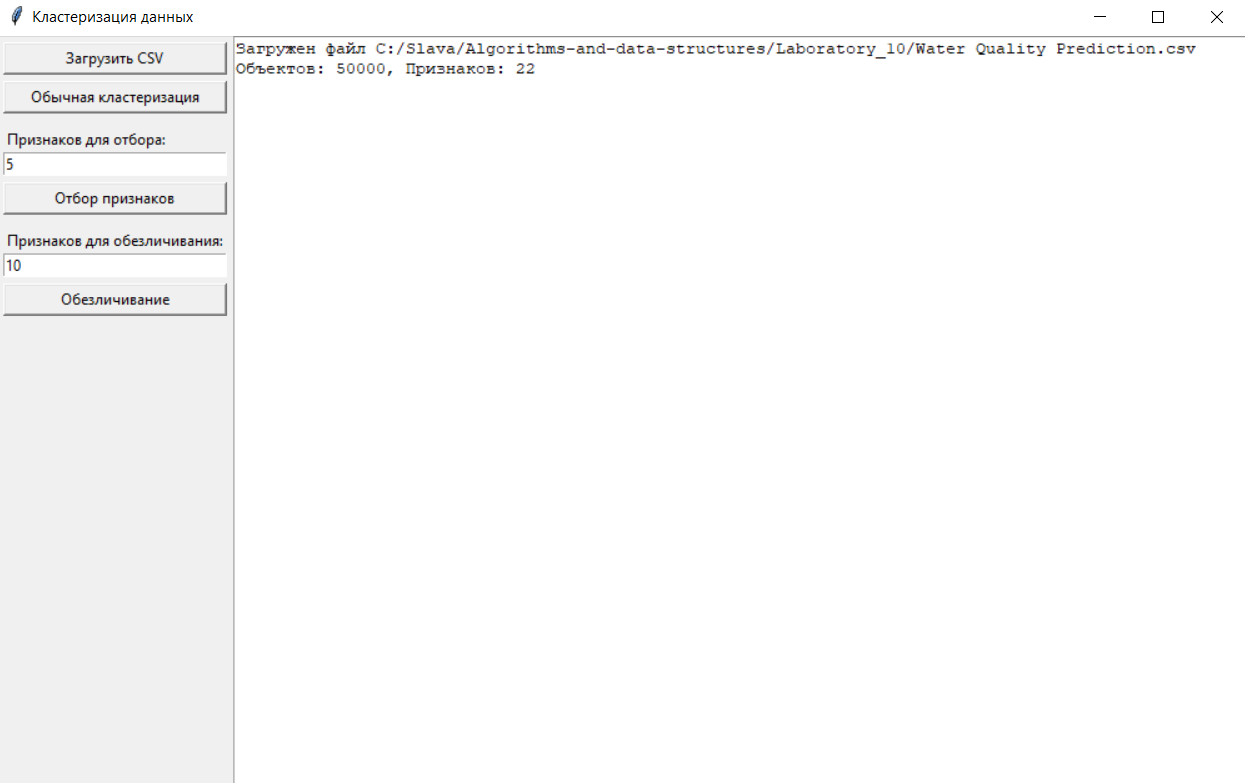


Рис. 3 Выбор датасета

* **Кластеризация:** **Нажмите кнопку «Обычная кластеризация», программа использует ISODATA,** Расстояние Чебышева и Индекс Фоулкса – Мэллоуза для вычисления, итоговый результат появится в правом фрейме и откроется новое окно с матрицей ошибок**. (Рис. 4/ Рис. 5)**

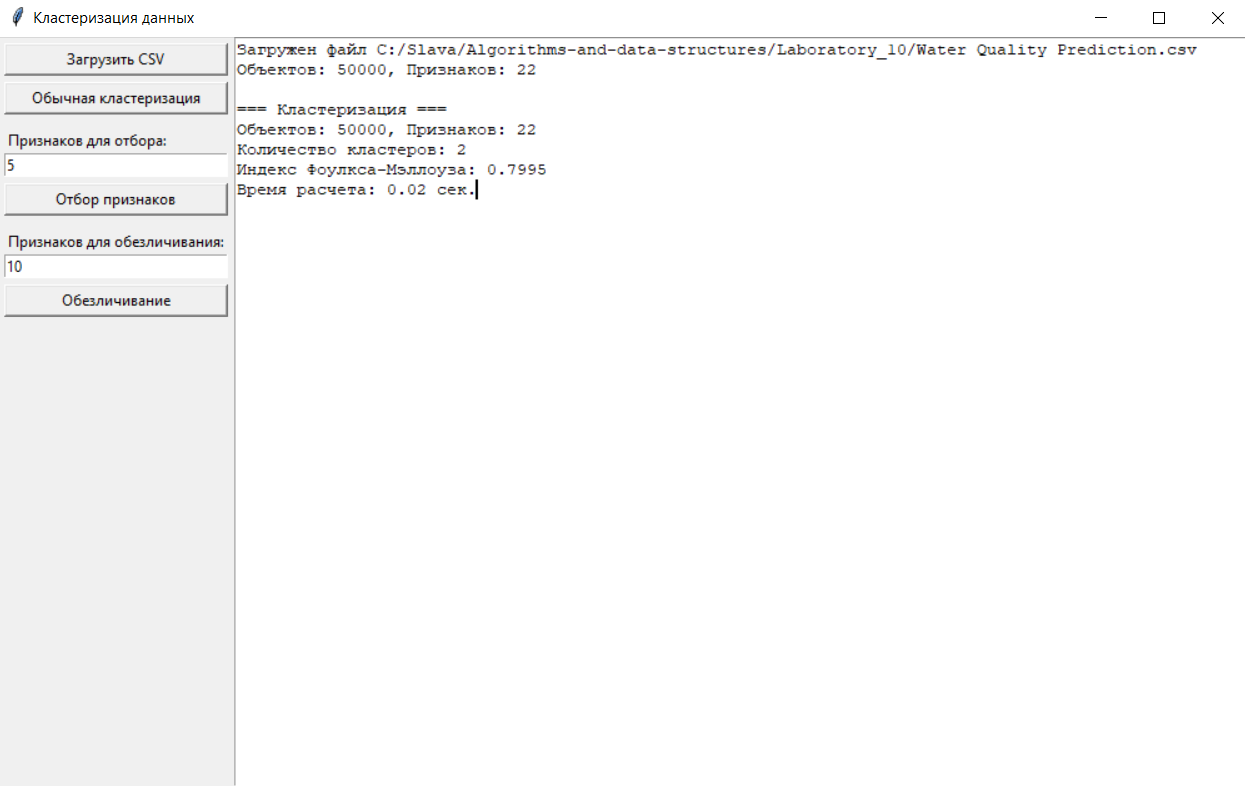


Рис. 4 Обычная кластеризация

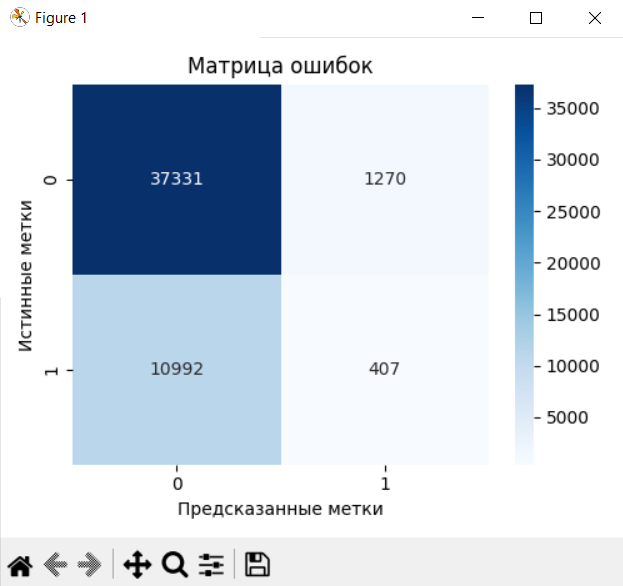


Рис. 5 Матрица ошибок

* **Отбор признаков:** Введите желаемое число признаков для отбора, н**ажмите кнопку «Отбор признаков», программа использует ISODATA,** Расстояние Чебышева, Индекс Фоулкса – Мэллоуза, алгоритм add для вычисления, итоговый результат появится в правом фрейме и откроется новое окно с матрицей ошибок, аналогичное окну с обычной кластеризации**. (Рис. 6)**

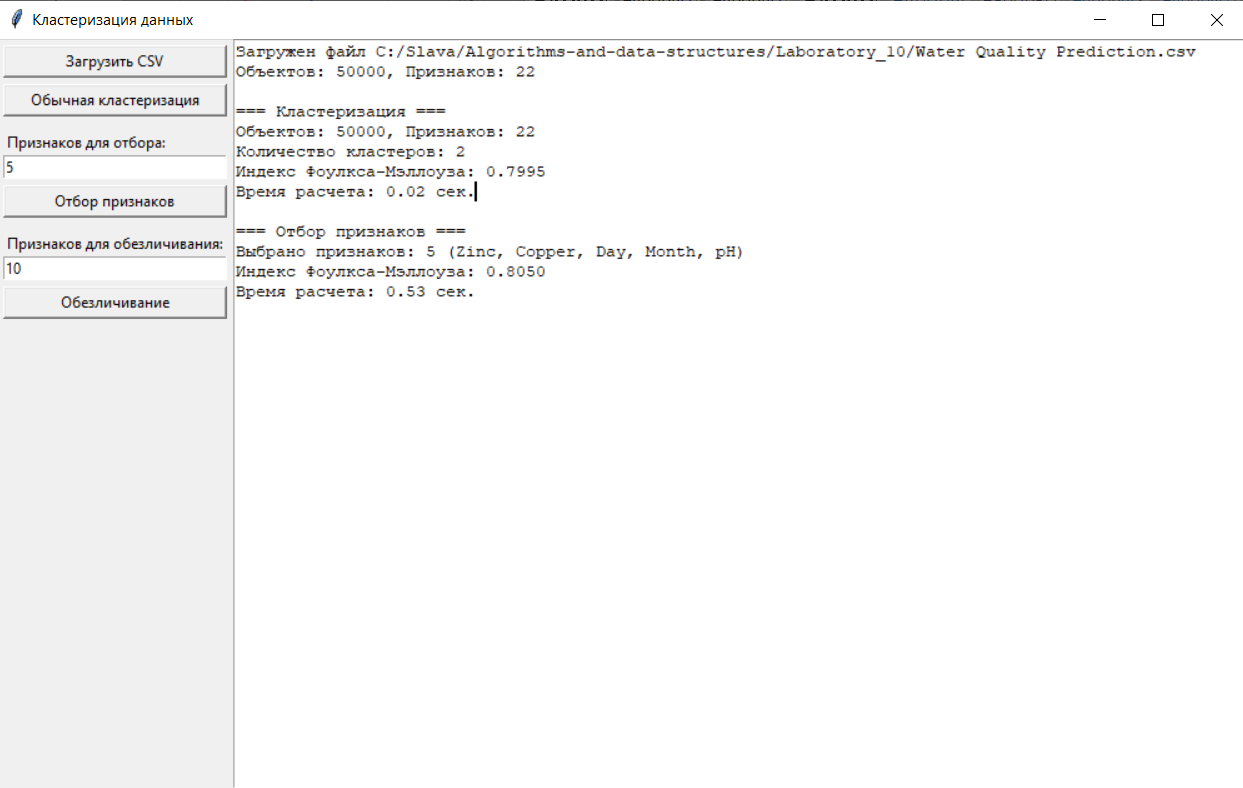


Рис. 6 Отбор признаков

* **Обезличивание:** Введите желаемое число признаков для отбора, н**ажмите кнопку «Обезличивание», программа обезличит датасет, использует ISODATA,** Расстояние Чебышева, Индекс Фоулкса – Мэллоуза, алгоритм add для вычисления, в правом фрейме появится итоговый результат, k-анонимность и откроется новое окно с матрицей ошибок, аналогичное окну с обычной кластеризации**. (Рис. 7)**

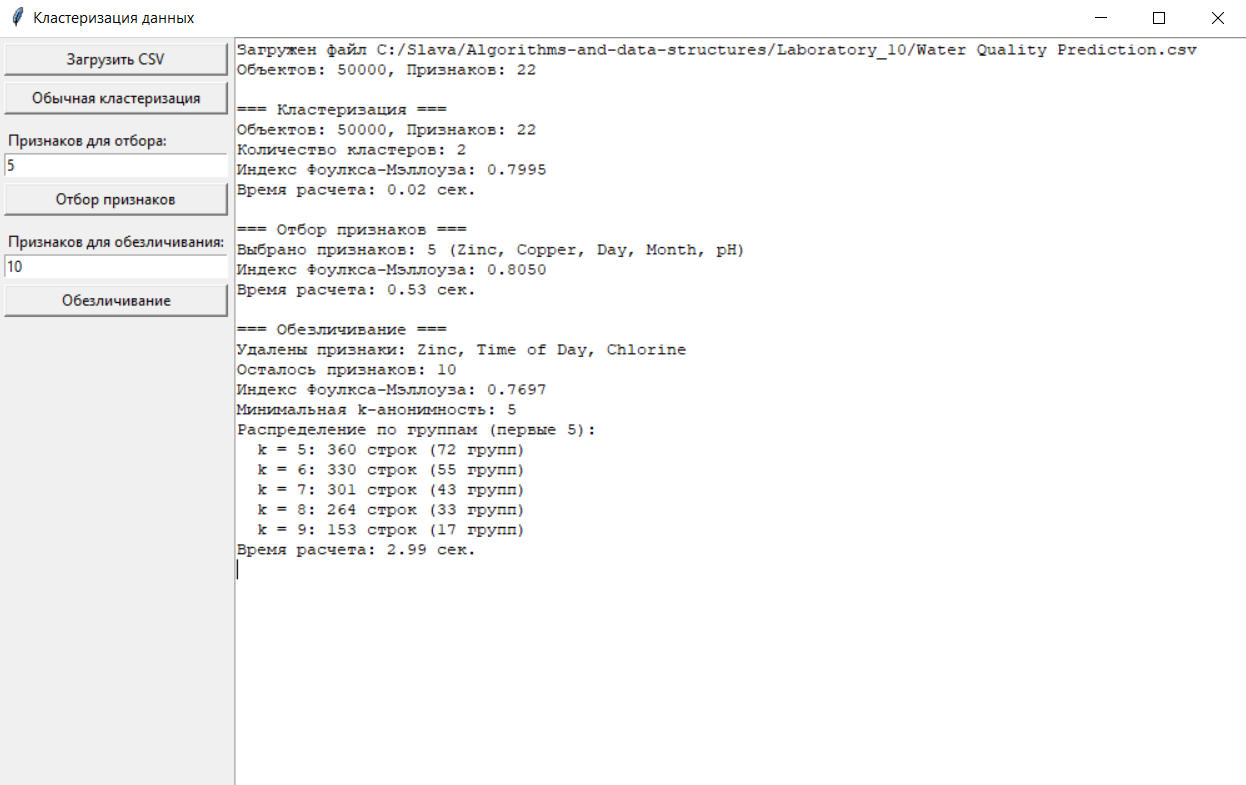


Рис. 7 Обезличивание

# Анализ работы

Для оценки эффективности методов кластеризации были проведены эксперименты на датасете объёмом 50 000 строк и имеющем 24 признака. Сравнивались три подхода: обычная кластеризация, кластеризация после отбора признаков и кластеризация после обезличивания. Результаты представлены в таблице ниже.

Таблица 2. Результаты работы

|  |  |  |  |
| --- | --- | --- | --- |
| Вид | Время работы (c) | Кол-во признаков | FMI |
| Обычная | 0.02 | 22 | 0.7930 |
| Выбор признаков | 2.20 | 20 | 0.8048 |
| Выбор признаков | 2.10 | 18 | 0.8046 |
| Выбор признаков | 1.75 | 15 | 0.8049 |
| Выбор признаков | 1.43 | 12 | 0.8043 |
| Выбор признаков | 0.90 | 8 | 0.8050 |
| Выбор признаков | 0.65 | 5 | 0.8050 |
| Выбор признаков | 0.30 | 3 | 0.8050 |
| Выбор признаков | 0.10 | 1 | 0.8050 |
| Обезличивание | 3.51 | 15 | 0.7308 |
| Обезличивание | 3.24 | 12 | 0.6500 |
| Обезличивание | 2.70 | 8 | 0.5900 |
| Обезличивание | 1.90 | 5 | 0.5900 |
| Обезличивание | 1.80 | 3 | 0.8000 |
| Обезличивание | 1.20 | 1 | 0.8050 |

**Наблюдения:**

* **Последовательный отбор признаков** показывает стабильный результат: FMI держится на уровне **0.8050** даже при использовании всего **одного признака**. Это указывает на высокую информативность ключевых признаков и эффективность данного.
* **Обезличивание признаков** ведёт к значительному снижению качества кластеризации. При уменьшении числа признаков индекс FMI падает с 0.73 до **0.59**, что обусловлено потерей значимой информации при обезличивание данных, но при малом числе выбранных признаков он снова возрастает, это может быть связано с в**ысокой информативностью отдельных признаков** — даже после обобщения они способны хорошо разделять объекты на кластеры.
* **Обычная кластеризация**, несмотря на минимальное время выполнения (0.02 с), даёт меньший индекс (0.7930), чем почти все конфигурации с отбором признаков, даже при значительно меньшем числе признаков.
* **Время выполнения** у метода отбора признаков заметно ниже, чем при обезличивании. При использовании 8 признаков метод с отбором выполняется за 0.90 с, а обезличенный — за 2.70 с. Но при этом оба данных метода уступают классической кластеризации.

Дополнительно стоит отметить, что при анализе тепловых матриц ошибок, построенных после кластеризации, наблюдается постоянная проблема с корректной кластеризацией объектов класса с меткой **"1"**. Это может быть связано с меньшим количеством таких объектов в данных или слабой выраженностью признаков, отличающих класс от остальных. (Рис. 8)

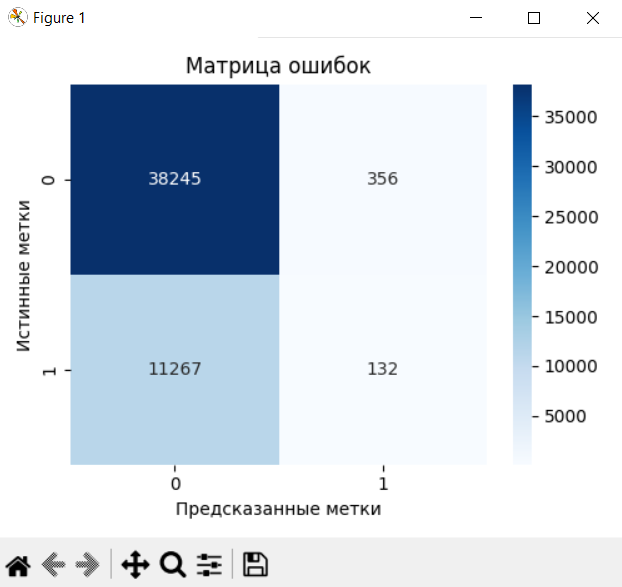


Рис. 7 Тепловая матрица

**Анализ графиков кластеризации:**

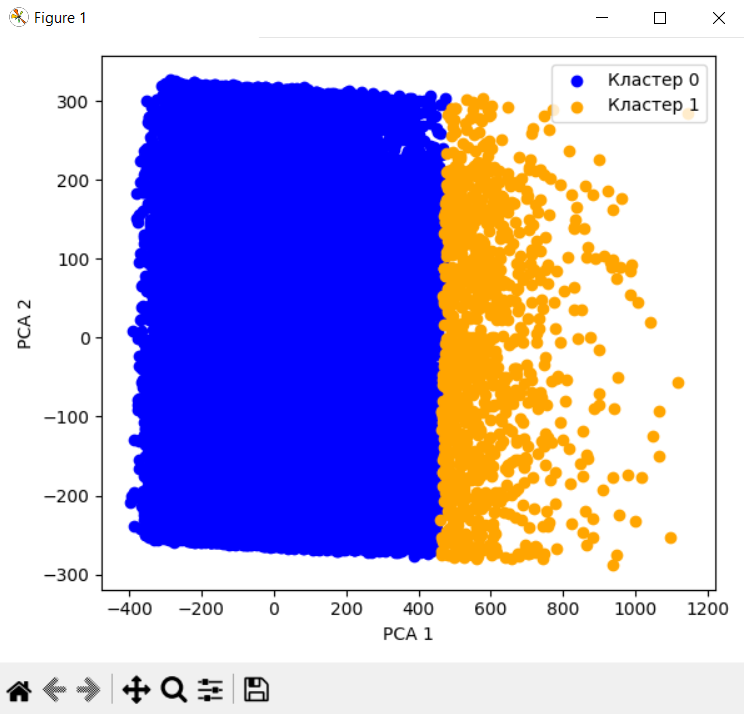
****

Рис. 8 Обычная кластеризация

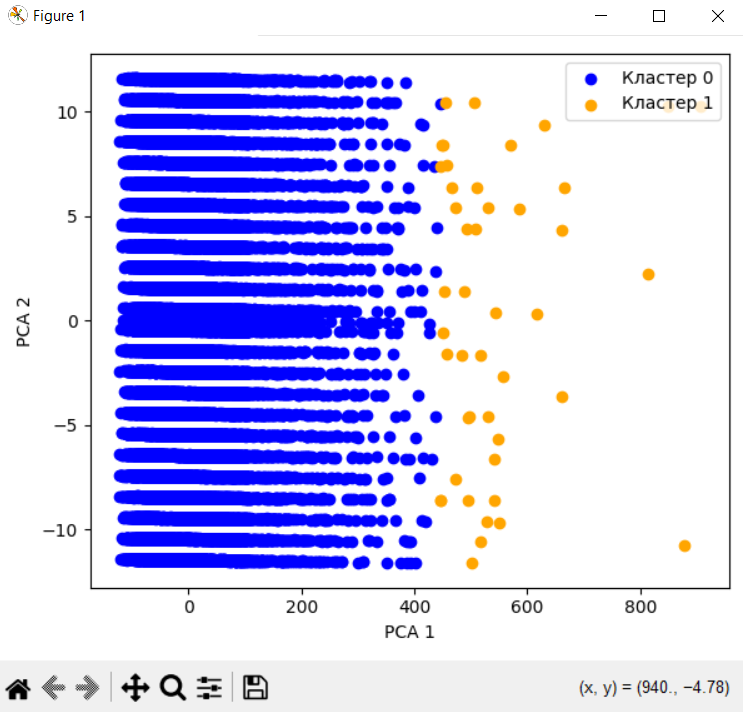


Рис. 9 Отбор признаков, график 1

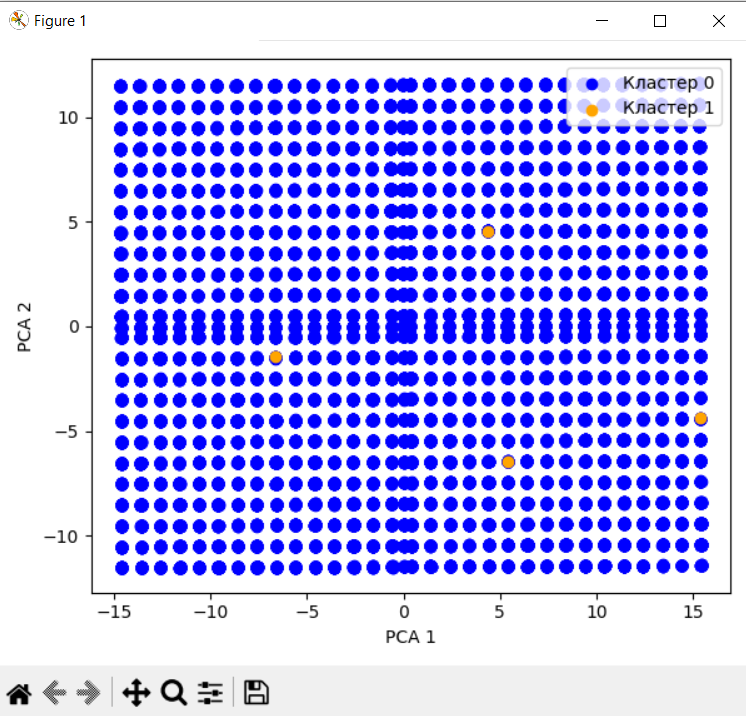


Рис. 10 Отбор признаков, график 2

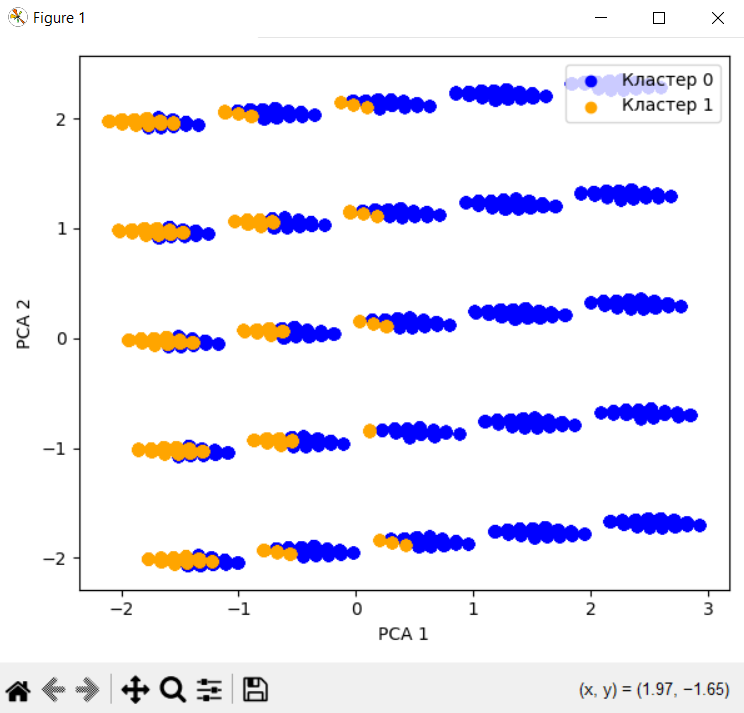


Рис. 11 Обезличивание

Из данных графиков видно, что все варианты дают достаточно странные кластеризации и нельзя однозначно сделать вывод о их качестве.

# Вывод

В рамках лабораторной работы была разработана программа с графическим интерфейсом на языке Python с использованием библиотеки Tkinter. Программа реализует различные подходы к кластеризации и оценке её качества, включая:

* обычную кластеризацию без предварительной обработки признаков;
* кластеризацию с отбором наиболее информативных признаков методом последовательного добавления
* кластеризацию после обезличивания данных с применением обобщения

Проведён сравнительный анализ результатов для датасета (50000 строк, 24 признака). Были измерены индекс Фоулкса–Мэллоуза и время выполнения каждого из подходов. Результаты показали, что:

* метод отбора признаков стабильно улучшает качество кластеризации по сравнению с исходными данными
* метод обезличивания снижает точность кластеризации при большом числе признаков, но способен сохранять высокие значения метрики при работе с малым числом признаков.

Таким образом, предложенная программа позволяет гибко экспериментировать с методами предварительной обработки и оценивать их влияние на результат кластеризации, что полезно как в исследовательских целях, так и при работе с реальными данными.

# Источники

* Редактор блок-схем.

[*https://programforyou.ru/block-diagram-redactor*](https://programforyou.ru/block-diagram-redactor)

*дата обращения: (28.04.2025)*

* tkinter — Библиотека для создания графических интерфейсов в Python.

[*https://docs.python.org/3/library/tkinter.html*](https://docs.python.org/3/library/tkinter.html%20)

*дата обращения: (28.04.2025)*

* math — Библиотека для работы с математическими функциями.

[*https://docs.python.org/3/library/math.html*](https://docs.python.org/3/library/math.html)

*дата обращения: (28.04.2025)*

* openpyxl — Библиотека для записи и чтения файлов Excel.

[*https://openpyxl.readthedocs.io/en/stable/*](https://openpyxl.readthedocs.io/en/stable/)

*дата обращения: (28.04.2025)*