**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики-процессов управления**

**Программа бакалавриата**

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

**ОТЧЁТ**

**По лабораторной работе № 1**

**По дисциплине “Алгоритмы и структуры данных”**

**На тему “Создание датасета”**

**Вариант 2**

**Студент гр. 24Б16-пу**

**Бублик И.О.**

**Преподаватель**

**Дик А.Г.**

**Санкт-Петербург**

**2025 г.**

Оглавление

**Цель** ……………………………………………………………………………3

[**Задачи** 4](#_Toc210290735)

[**Теоретическая часть** 4](#_Toc210290736)

[**Основные шаги программы** 5](#_Toc210290737)

[**Блок схема программы** 6](#_Toc210290738)

[**Рекомендации пользователя** 10](#_Toc210290767)

[**Рекомендации программиста** 10](#_Toc210290768)

[**Контрольный пример** 11](#_Toc210290769)

[**Вывод** 11](#_Toc210290771)

**Цель –** Разработка программы для генерации синтетического датасета,

имитирующего реальные данные.

**Задачи:**

1) Спроектировать структуру датасета.

2) Разработать алгоритм генерации (Создать правила и зависимости между

признаками и целевой переменной, чтобы данные были правдоподобными).

3) Реализовать программу генератор.

4) Протестировать результат (в результате должен сгенерирован быть файл расширения .xlsx).

**Теоретическая часть**

Задача состоит в создании датасета клиентов медицинской организации

1.1. Синтетические данные и их назначение

Синтетические данные — это искусственно созданные данные, имитирующие реальные. Их основное назначение заключается в возможности тестирования, обучения и анализа алгоритмов без необходимости использовать персональные или конфиденциальные сведения. В современном мире, где защита персональных данных имеет первостепенное значение, синтетические данные становятся востребованным инструментом для исследователей, аналитиков и разработчиков.

Главное преимущество синтетических данных состоит в том, что они позволяют сохранять закономерности, характерные для реальных выборок, при этом исключая риск утечки личной информации. Такие данные могут использоваться для тестирования аналитических моделей, обучения систем машинного обучения и разработки прототипов информационных систем.

1.2. Принципы генерации синтетических данных

Алгоритм генерации синтетического датасета основывается на ряде принципов:

Реалистичность распределений.

Согласованность признаков. Значения различных характеристик должны быть логически согласованы (например, мужчина не может иметь диагноз «беременность»).

Случайность и воспроизводимость. Генерация данных должна учитывать случайность, но при этом быть воспроизводимой при одинаковых параметрах генерации.

Вариативность. Датасет должен включать разнообразие наблюдений, чтобы отражать реальную структуру клиентской базы.

**Основные шаги программы**

1. Запуск программы (main.py)
2. Пользователь вводит процентные соотношения банков и платежных систем.
3. Выполняется файл main.py.
4. Генерируется список клиентов
5. Генерируется список симптомов
6. По списку симптомов выбирается врач и генерируются анализы
7. Добавляются остальные данные, такие как паспорт, снилс, карта, дата посещения и дата получения анализов
8. Данные собираются и записываются в файл dataset.xlsx

**Блок схема программы**

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Содержимое, созданное искусственным интеллектом, может быть неверным.

Рис. 1 Блок-схема основной программы

**Описание программы**

Программа реализована на языке python 3.13, с использованием библиотек: random, pandas, datetime.

В таблице 1 представлено описание функций, используемых в программе.

Таблица 1. Описание функций

|  |  |  |
| --- | --- | --- |
| Имя функции | Тип возвращаемого значения | Описание функции |
| gen\_russion\_passport | str | Генерирует русский паспорт |
| gen\_kazakh\_passport | str | Генерирует казахский паспорт |
| gen\_belarus\_passport | str | Генерирует белорусский паспорт |
| gen\_snils | str | Генирирует снилс |
| random\_work\_datetime | datetime | Генерирует дату посещения в рабочее время |
| choose\_card | dict | Генерирует номера банковских карт с учетом вероятностей |
| generate\_dataset | DataFrame | Добавляет новые строки в датасет, следя за выполнениями всех условий |
| Random\_analisys\_datetime | datetime | Генерирует дату получения анализа с учетом ограничений |

В таблице 2 представлено описание структур данных, используемых в программе.

Таблица 2.Описание структур

|  |  |  |  |
| --- | --- | --- | --- |
| Имя структуры | Тип структуры | Тип хранимых данных | Описание |
| surnames\_man | list | str | Хранит фамилии, имена, отчества мужчин |
| first\_names\_man |
| patronymics\_man |
| surnames\_women | list | str | Хранит фамилии, имена, отчества женщин |
| first\_name\_women |
| patronymics\_women |
| base\_sympton\_to\_doctors | dict | str,str | Ключ это симптом, а значения это возможные доктора |
| test\_prises | dict | str,int | Ключ это анализ, а значение цена |
| symptom\_to\_tests | dict | str,str | Ключ это симптомы, а значение возможные анализы |
| bank\_card\_bins | dict | str,str | У каждого банка и у каждой платежной системы хранится свой бин(первые 6 цифр карты) |
| data | DataFrame | str | Итоговый датасет |

# **Рекомендации пользователя**

В консоли через пробел введите 4 числа, вероятность банков, потом enter, через пробел 3 числа, вероятность платежных систем, enter, ожидайте генерации датасета, сравните с контрольным примером сгенерированный датасет.

# **Рекомендации программиста**

Для корректного функционирования программы рекомендуется выполнить следующие действия:

* Убедитесь, что у вас установлена библиотека pandas. Если отсутствует, вы можете установить ее через пакетный менеджер вашей операционной системы.
* Рекомендуется использовать python версии 3.13 или выше, чтобы избежать возможных проблем с совместимостью библиотек и функциональностью.

# **Контрольный пример**

Запуск программы: Для запуска программы используйте файл main.py. Через пробел введите 4 числа, вероятность банков, потом enter, через пробел 3 числа, вероятность платежных систем, enter, ожидайте генерации датасета

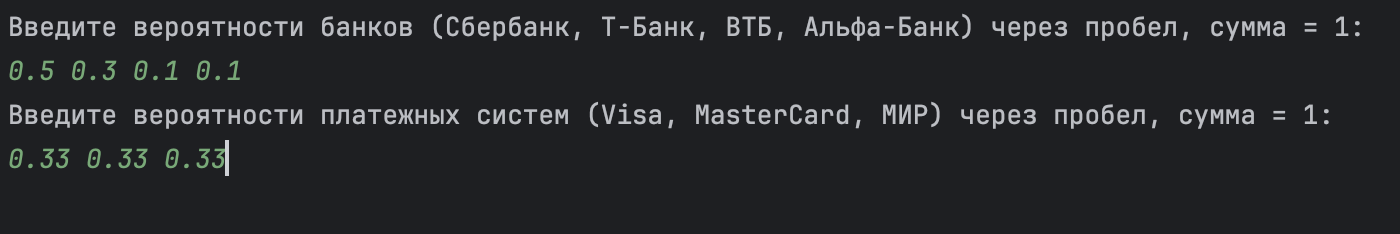


Рисунок 1. Запуск программы

# 

Рисунок 2. Результат работы программы

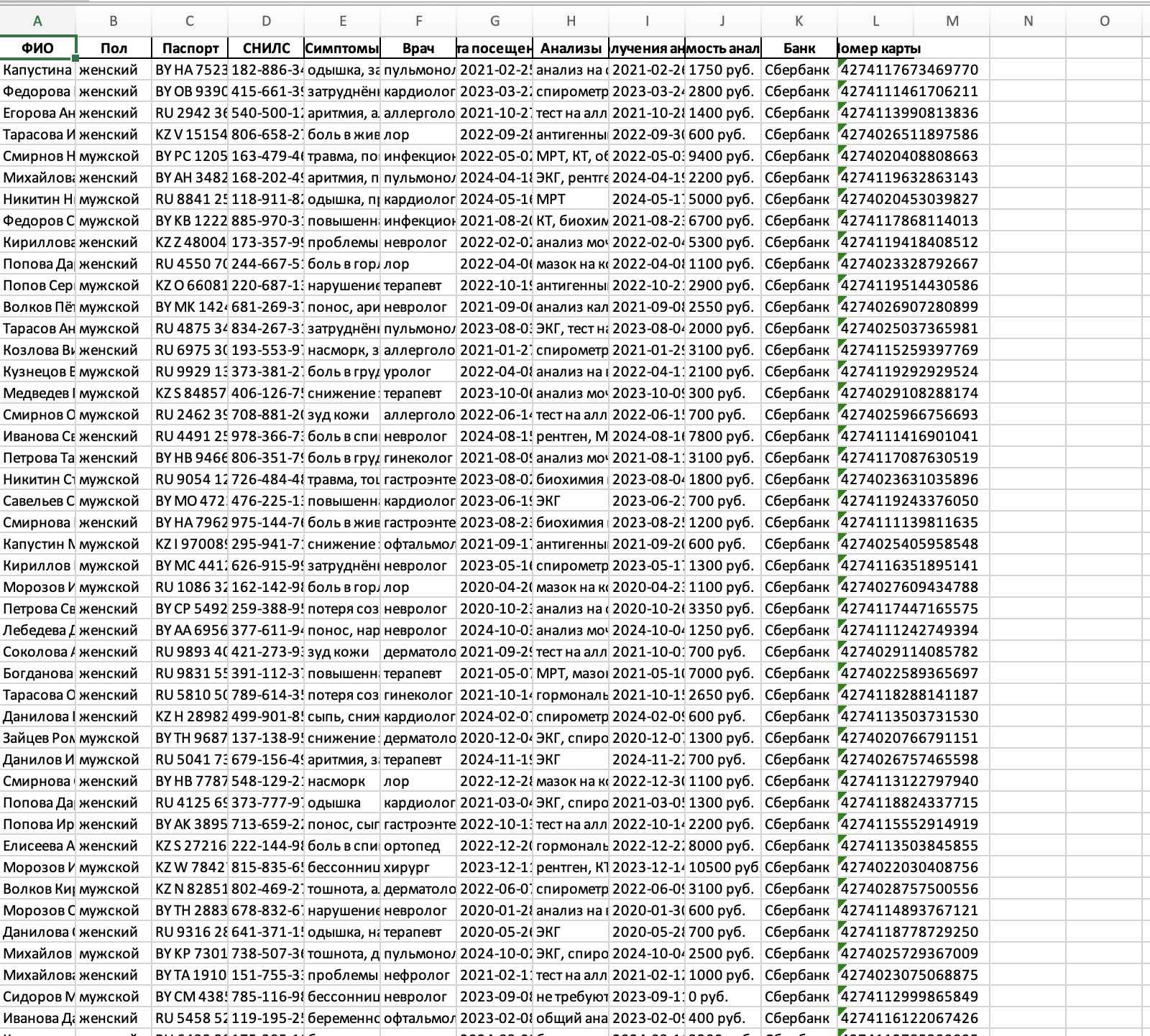


Рисунок 3. Пример корректного итогового датасета

Полученный файл следует открывать в приложение MS Exel.

# **Вывод**

В ходе выполнения данной работы были изучены принципы генерации синтетических данных, на примере медицинских данных, разработан алгоритм генерации синтетических данных данного типа и написана программа, которая при вводе определенных параметров генерирует датасет клиентов поликлиники.