Лабораторная работа №1 – Создание датасета.

Цель – Разработка программы для генерации синтетического датасета, имитирующего реальные данные.

Задачи:

1) Спроектировать структуру датаеста.
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).
5) Написать отчет (файл расширения .doc или .docx)

Содержание отчета:

1. Цель работы.
2. Описание задачи (формализация задачи)
3. Теоретическая часть. (Краткое описание создание датасета и его правил с ограничениями)
4. Представить описание блок-схемы пошагового выполнения алгоритма.
5. Описание программы. (Представление программы на выбранном вами языке, спецификация программы (описание классов и тп)).
6. Рекомендации пользователя.
7. Рекомендации программиста.
8. Описание контрольного примера. Продемонстрировать, как можно редактировать различные критерии, для изменения входных параметров для датасета.
9. Вывод по работе.

I. Вариант (Покупка в магазине)

A. Сгенерировать датасет в котором будут следующие наборы свойств:

1. Название магазина - М.Видео
2. Координаты (2 столбика - дата и время + долгота и широта) - 2020-01-22T08:30+03:00 и 59.881653, 29.830170
3. Категорий - ноутбук
4. Бренд - lenovo
5. Стоимость товара - 50 000 руб.
6. Номер карточки - “1234 5678 1234 5678”
7. Количество товаров в чеке - 2 шт.
8. Кассовый чек - № 2967
9. Общая стоимость в чеке - 120 000 руб.

B. Дополнительная информация по каждому свойству (Санкт-Петербург):

1. Название магазина - ”Словарь” по которому будет генерироваться магазины на данной территории.
2. Координаты и время (2 столбика - дата и время + долгота и широта) - согласна реальному местоположению магазина на данной территории.
3. Категорий - возможность настраивать датасет, а категории, должны соответствовать тематики магазина.
4. Бренд - возможность настраивать датасет, а бренд, должен соответствовать тематики категории в магазине.
5. Стоимость товара – стоимость зависит от бренда и магазина.
6. Номер карточки - возможность настраивать сам датасет и вероятность какого банка (Сбербанк, Газпромбанк и тд), через какую платежную систему (Visa, MasterCard и тд) производится оплата. Оплачивать могут несколько раз с одной карты.
7. Количество товаров - свободный вариант генерации данных для 1 чека
8. Кассовый чек – свободная генерация номера чека. В чеке могут быть и другие позиции из магазина.
9. Общая стоимость в чеке - согласна средней стоимости всех товаров в чеке.

C. Ограничения датасета:

1. Всего строк в датасете - минимум 50 000.
2. Название магазина - ”Словарь” должен состоять минимум из 30 магазинов
3. Координаты (2 столбика - дата и время + долгота и широта) - округлять до 10 в -8 степени значения в координатах, а дату и время - актуальную и реального посещения магазина, то есть, если магазин работает с 10:00 до 22:00, то только в этот промежуток, возможно сделать покупку.
4. Категорий - ”Словарь” должен состоять минимум из 50 категорий
5. Бренд - ”Словарь” должен состоять минимум из 500 брендов
6. Номер карточки – максимальное количество повторов - 5 раз
7. Количество товаров - минимум 2 в чеке.
8. Кассовый чек – не может повторяться в одном магазине.

II. Вариант (Платная поликлиника)
A. Сгенерировать датасет в котором будут следующие наборы свойств:

1. ФИО - Иванов Иван Иванович
2. Паспортные данные - 1234 123456
3. СНИЛС - 123-456-789 12
4. Симптомы - боль в горле
5. Выбор врача - лор
6. Дата посещения врача- 2020-01-22T08:30+03:00
7. Анализы - мазок на ковид
8. Дата получения анализов - 2020-01-24T09:30+03:00
9. Стоимость анализов - 2000 руб.
10. Карта оплаты - “1234 5678 1234 5678”

B. Дополнительная информация по каждому свойству:

1. ФИО - словарь по ФИО.
2. Паспортные данные - уникальные значения
3. СНИЛС - уникальные значения
4. Симптомы - ”Словарь” по возможным симптомам
5. Выбор врача - ”Словарь” по возможным специальностям которые могут работать в поликлинике
6. Дата посещения врача- определенный вариант генерации данных.
7. Анализы - ”Словарь” по возможным анализам
8. Дата получения анализов - после посещения врача
9. Стоимость анализов - свободный вариант генерации данных.
10. Карта оплаты - возможность настраивать датасет и вероятность какого банка (Сбербанк и тд), через какую платежную систему (Visa и тд) производится оплата. Оплачивать могут несколько раз с одной карты.

C. Ограничения датасета:

1. Всего строк в датасете - минимум 50 000.
2. ФИО - словарь должен состоять только из славянских ФИО
3. Паспортные данные - только русские, белорусские и казахские паспорта должны быть.
4. СНИЛС - уникальный, но привязан к клиенту (ФИО и паспортные данные), которые могут повторяться при повторном посещение.
5. Симптомы - ”Словарь” должен состоять минимум из 5000 симптомов. То есть можем быть комбинация итоговых симптомов (не более 10 штук)
6. Выбор врача - ”Словарь” должен состоять минимум из 50 врачей.
7. Дата посещения врача - В рабочие время и дни недели. Повторное посещение может быть к врачу минимально через 24 часа после получения анализов.

8. Анализы - ”Словарь” должен состоять минимум из 250 анализов. То есть можем быть комбинация итоговых симптомов (не более 5 штук)

9. Дата получения анализов - В рабочие время и дни (через 24-72 часа).

10. Стоимость анализов - только в рублях.

11. Карта оплаты - максимальное количество повторов - 5 раз

III. Вариант (Покупка жд билета)
A. Сгенерировать датасет в котором будут следующие наборы свойств:

1. ФИО - Иванов Иван Иванович
2. Паспортные данные - 1234 123456
3. Откуда - Санкт-Петербург
4. Куда - Москва
5. Дата отъезда - 2020-01-22T08:30
6. Дата приезда - 2020-01-22T22:30
7. Рейс - 723A
8. Выбор вагона и места - 3-12 (3 вагон, 12 место)
9. Стоимость - 2460 руб.
10. Карта оплаты - “1234 5678 1234 5678”

B. Дополнительная информация по каждому свойству:

1. ФИО - словарь по ФИО
2. Паспортные данные -уникальные значения
3. Откуда - по территории РФ
4. Куда - по территории РФ
5. Дата отъезда - свободный вариант генерации данных. Учитывая рейсы поездов (сапсан, ласточка и т.п.)
6. Дата приезда - свободный вариант генерации данных. Учитывая рейсы поездов (сапсан, ласточка и т.п.)
7. Рейс - все модели поездов дальнего маршрута
8. Выбор вагона и места - нужно учитывать классы вагонов
9. Стоимость - согласно длине маршрута и класса
10. Карта оплаты - “возможность настраивать датасет и вероятность какого банка (Сбербанк и тд), через какую платежную систему (Visa и тд) производится оплата. Оплачивать могут несколько раз по одной карте.

C. Ограничения датасета:

1. Всего строк в датасете - минимум 50 000.
2. ФИО - словарь должен состоять только из славянских ФИО
3. Паспортные данные - только русские
4. Откуда - по территории РФ но не совпадает с “Куда”
5. Куда - по территории РФ, но не совпадает с “Откуда”
6. Дата отъезда - нет ограничений.
7. Дата приезда - нет ограничений.
8. Рейс - идут по следующим кейсам, которые бывают только 1 раз до полного возвращения.
    а) 001–150 — скорые поезда, которые есть в расписании круглый год

b) 151–298 — скорые поезда сезонного или разового назначения (пускают в праздничные дни, или только летом, или только зимой и т.п.)

c) 301–450 — пассажирские круглогодичные поезда

d) 451–598 — пассажирские поезда сезонного или разового назначения

e) 701–750 — скоростные поезда (средняя маршрутная скорость с учётом стоянок —91 км/ч и более; например, «Ласточки»)

f) 751–788 — высокоскоростные поезда (от 161 км/ч; например, «Сапсан», «Аллегро», «Невский экспресс»)

9. Выбор вагона и места идут по следующим кейсам
   a) Поезда «Сапсан»
      (1) 1Р — купе-переговорная, продаётся только целиком.
      (2) 1В — просто места в вагоне 1 класса, без переговорной.
      (3) 1С — вагон бизнес-класса.
      (4) 2С — сидячий вагон эконом-класса.
      (5) 2В — класс «Экономический+» (вагон № 10 и № 20).
      (6) 2Е — места в вагоне-бистро, в стоимость билета включено питание на сумму 2000 рублей по меню. Через интернет можно купить в день отправления и на следующий день.
   b) Поезда «Стриж»
      (1) 1Е — СВ (VIP). Продаётся купе целиком, в нём могут ехать 1 или 2 пассажира.
      (2) 1Р — сидячие вагоны 1 класса.
      (3) 2С — сидячие вагоны 2 класса.
   c) Сидячий вагон
      (1) 1С - обычные сидячие места
      (2) 1Р — в двухэтажном сидячем вагоне так маркируются места в купе
      (3) 1В — вагон с индивидуальным размещением, то есть выкупаются все места.
      (4) 2Р — вагон повышенной комфортности
      (5) 2Е — сидячий вагон
   d) Плацкартные вагоны
      (1) 3Э — плацкартный вагон
   e) Купе
      (1) 2Э — кондиционируемый вагон повышенной комфортности с 4-местными купе.
   f) Люкс (СВ)
      (1) 1Б — бизнес-класс.
      (2) 1Л — вагон СВ.
   g) Мягкий вагон
      (1) 1А — вагон состоит из 4 купе и салона-бара.
      (2) 1И — аналогично 1А, единственное отличие — нет бара.
10. Стоимость - зависит от типа вагона

11. Карта оплаты - максимальное количество повторов - 5