Екзаменаційна робота

з дисципліни “Data Science and Big Data”

студента групи ПП-41

Селецького В.Р.

1. Визначення критерію Фішера.

Критерій Фішера - це статистичний критерій, який використовується для з’ясування статистичної значущості між середніми значеннями декількох груп. Тестова статистика критерію Фішера має F-розподіл Фішера при виконанні нульової гіпотези (припущення про те, що між групами не існує ніякого зв'язку). Статистика тесту полягає у відношенні вибіркових дисперсій. Щоб ця статистика мала розподіл Фішера, необхідно, щоб чисельник і знаменник були незалежними величинами, і щоб відповідні суми квадратів мали розподіл Хі квадрат; відповідно, для цього потрібно, щоб дані мали нормальний розподіл.

1. Опишіть відкриті дані, їх формати та засоби обробки.

За визначенням Open Knowledge Foundation, відкриті дані - це “будь-який вміст, інформацію або дані, які люди можуть використовувати і перерозподіляти без будь-яких юридичних, технологічних чи соціальних обмежень”. Загалом відкриті дані надаються установами, які, очевидно, мають доступ до них і мають бажання зробити їх відкритими (наприклад, дані про ресторани Києва, що пройшли перевірку санепідемстанцією, або дані про кількість ввезених до країни автомобілів на польских номерах).

Оскільки ці дані зазвичай попередньо збираються установою внутрішньо, то їх потрібно привести до певного загального стандарту. CSV (comma-separated values) - це тип простого тексту, розділений комами; також використовуються JSON та XML. Ці три типи даних є дуже зручними для об’єднання інших даних в один тип.

Обробка відкритих даних може зайняти набагато більше часу, ніж займе їх аналіз, оскільки зазвичай вони неструктуровані, можуть бути несумісні між собою. Формати дати та часу, різні мови теж можуть заважати коректному об’єднанню. Тому дані мають бути перевірені за допомогою правил перевірки, щоб переконатися, що вони мають бажане значення. Якщо дані не відповідають правилам перевірки, вони можуть бути відхилені, після чого також можуть бути виправлені та підтверджені. Після цього відбувається перетворення перетворення

вихідних даних у необхідний тип даних. Крок

трансформації також вимагає декількох інших завдань. Деякі з цих завдань – це, як було описано вище, об'єднання даних із кількох джерел, їх агрегування, сортування, визначення нових значень, які обчислюються з агрегованих даних та застосування правил

перевірки. Дані (включаючи деякі відхилені дані) можуть пройти через іншу частину етапу перетворення, відомий як “очищення” даних. Частина очищення етапу трансформації додатково забезпечує узгодженість вихідних даних.