НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ

Кафедра математичних методів системного аналізу

До захисту допущено:

Завідувач кафедри
_____ Оксана ТИМОЩУК

«___»_____2024 р.

Дипломна робота

на здобуття ступеня бакалавра
за освітньо-професійною програмою «Системний аналіз і управління»
спеціальності 124 «Системний аналіз»
на тему: «Використання “великих даних” у вибіркових обстеженнях»

Виконав:
студент IV курсу, групи КА-04
Дідіков Олександр Олександрович _____
Керівник:
Доцент, д.ф.-м.н., Василик Ольга Іванівна _____
Консультант з економічного розділу:
Доцент, д.ф.е., Мажара Гліб Анатолійович _____
Консультант з нормоконтролю:
Доцент, к.ф.-м.н., Статкевич Віталій Михайлович _____
Рецензент:
Доцент кафедри теорії ймовірностей, статистики
та актуарної математики
Київського національного університету
імені Тараса Шевченка, к.ф.-м.н.,
Яневич Тетяна Олександрівна _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студент _____

Київ – 2024 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут прикладного системного аналізу

Кафедра математичних методів системного аналізу

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 124 «Системний аналіз»

Освітньо-професійна програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Оксана ТИМОЩУК

«___»_____2024 р.

ЗАВДАННЯ

на дипломну роботу студенту

Дідікову Олександру Олександровичу

1. Тема роботи «Використання "великих даних" у вибіркових обстеженнях», керівник роботи Василик Ольга Іванівна, доктор фізико-математичних наук, доцент, затверджені наказом по університету від «___»_____ 2024 р. №_____

2. Термін подання студентом роботи: 11.06.2024 р.

3. Вихідні дані до роботи:

1) Дані про продажі електронного магазину з відкритого датасету на Kaggle;

2) Мова програмування Python;

3) Середовище розробки – PyCharm;

4) Бібліотеки та модулі, що використовувалися: NumPy, Pandas, Sklearn, Matplotlib, Seaborn, WordCloud.

4. Зміст роботи: огляд проблеми використання "великих даних", методологія використання "великих даних" у вибіркових обстеженнях, тестування та аналіз

результатів використання "великих даних" в обстеженнях, функціонально-вартісний аналіз програмного продукту.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): графічні матеріали.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний	Мажара Г.А. доц. д.ф.е.		

7. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Вибір теми і формулювання задачі дослідження	10.04.2024	Виконано
2	Обґрунтування актуальності задачі дослідження	29.04.2024	Виконано
3	Збір літератури	01.05.2024	Виконано
4	Огляд основних методів та підходів	03.05.2024	Виконано
5	Завершення теоретичної частини	11.05.2024	Виконано
6	Завершення практичної частини	20.05.2024	Виконано
7	Аналіз та систематизація результатів	01.06.2024	Виконано
8	Оформлення презентації	04.06.2024	Виконано

Студент Олександр ДІДІКОВ

Керівник Ольга ВАСИЛИК

РЕФЕРАТ

Дипломна робота: 80 с., 8 рис., 8 табл., 3 додатки, 19 джерел.

"ВЕЛИКІ ДАНІ", ВИБІРКОВІ ОБСТЕЖЕННЯ

Тема: Використання "великих даних" у вибіркових обстеженнях.

У роботі розглянуто методологію інтеграції "великих даних" у процеси вибіркових обстежень, аналіз переваг та викликів, а також розробку практичних рекомендацій для покращення точності та ефективності обстежень.

Об’єкт дослідження: застосування методів обробки та аналізу "великих даних" у вибіркових обстеженнях.

Предмет дослідження: методи і засоби інтеграції "великих даних" у вибіркових обстеженнях.

Мета роботи: розробка методології та програмного забезпечення для покращення результатів вибіркових обстежень шляхом інтеграції "великих даних".

Створено методологію та програмне забезпечення мовою програмування Python для інтеграції "великих даних" у вибіркові обстеження, що дозволяє підвищити точність та надійність отриманих результатів, експериментально перевірено розроблену методологію на реальних даних із Kaggle. Для розробки використані сучасні технології аналізу даних та машинного навчання.

ABSTRACT

Diploma thesis: 80 pages, 8 figures, 8 tables, 3 appendices, 19 references.

“BIG DATA”, SAMPLE SURVEYS.

Theme: The use of “big data” in sample surveys.

The paper examines the methodology of integrating big data into sample surveys, analyzes the benefits and challenges, and develops practical recommendations for improving the accuracy and efficiency of surveys.

Object of research: application of methods of processing and analyzing big data in sample surveys.

Subject of research: methods and tools for integrating big data in sample surveys.

Purpose: to develop a methodology and software to improve the results of sample surveys by integrating big data.

A methodology and software for integrating big data into sample surveys was created, which allows to increase the accuracy and reliability of the results obtained, and the developed methodology was experimentally tested on real data from Kaggle. Modern data analysis and machine learning technologies were used for the development (in Python programming language).

ЗМІСТ

ВСТУП

У сучасному світі обсяги даних зростають експоненційно, що створює як нові можливості, так і виклики для дослідників у різних галузях науки і практики. "Великі дані" (Big Data) стали однією з ключових концепцій у сфері інформаційних технологій, аналітики та статистики. Використання великих даних дозволяє отримувати цінну інформацію, виявляти закономірності та робити точні прогнози, що є критично важливим для прийняття обґрунтованих рішень у бізнесі, медицині, науці та інших сферах.

Проте, використання великих даних супроводжується численними викликами. Однією з основних проблем є статистичні зміщення, які виникають через неповне охоплення популяції та помилки вимірювань. Це особливо актуально при інтеграції великих даних з традиційними вибірковими обстеженнями, які використовуються для отримання статистичних оцінок параметрів популяції. У цьому контексті виникає потреба у розробці нових методологічних підходів, які дозволяють ефективно поєднувати великі дані з вибірковими обстеженнями для отримання надійних і точних висновків.

Метою цієї роботи є дослідження методології використання великих даних у вибіркових обстеженнях та розробка підходів, які дозволяють підвищити точність статистичних оцінок. У роботі розглядаються основні концепції великих даних, їх історія виникнення та еволюція, переваги та виклики, пов'язані з їх використанням. Також наводяться основні поняття вибіркових обстежень, методи збору, обробки та аналізу великих даних, а також приклади їх використання у різних сферах.

У першому розділі наведено концепцію великих даних, розглянуто їх значення, історію виникнення та еволюцію. Було також обговорено переваги та виклики, пов'язані з використанням великих даних у різних сферах, що заклало

основу для подальшого дослідження методології їх інтеграції у вибіркові обстеження.

Другий розділ присвячено розробці методології використання великих даних у вибіркових обстеженнях. Проведено огляд існуючих інструментів та платформ для роботи з великими даними, розглядається процес стратифікації та оцінювання шляхом інтеграції великих даних, модифіковано регресійну оцінку для сумарного значення досліджуваної характеристики генеральної сукупності (популяції). Показано, як запропоновані методи дозволяють коригувати систематичні похибки та покращувати точність оцінок параметрів популяції.

Третій розділ роботи містить тестування та аналіз результатів використання великих даних у вибіркових обстеженнях. Обрано інструменти та методи тестування, описано процедуру проведення тестування, а також проведено аналіз отриманих результатів. Результати тестування показали, що метод виправлення зміщення значно підвищує точність оцінок, підтверджуючи ефективність розробленої методології.

Четвертий розділ присвячено функціонально-вартісному аналізу програмного продукту.

Ця робота робить вагомий внесок у розвиток методологій роботи з великими даними та їх інтеграції у вибіркові обстеження, що має значний потенціал для застосування у різних галузях науки і практики.

1 ОГЛЯД ПРОБЛЕМИ ВИКОРИСТАННЯ "ВЕЛИКИХ ДАНИХ"

1.1 Вступ в концепцію "великих даних" та їх значення

У сучасному світі, де обсяг інформації зростає шаленими темпами, проблема ефективного використання «великих даних» (*англ*. Big Data) набуває все більшої актуальності. Термін «великі дані» охоплює величезні масиви різноманітної інформації, які генеруються з багатьох джерел, таких як соціальні мережі, Інтернет речей, наукові експерименти та корпоративні бази даних.

Згідно з джерелом [19], не існує єдиного визначення поняття «великих даних» (далі цитата): «Одне з них свідчить, що це дані, які неможливо обробити традиційними способами через їх великий обсяг. Інше - що це феноменальне прискорення нагромадження даних та їх ускладнення. Третє визначення стверджує, що це набір інструментів, що дозволяють працювати з даними, незалежно від їх типу та обсягу. Така ситуація пояснює той факт, що для характеристики Big data використовують «три v»: обсяг (від англ. «volume», мається на увазі обсяг даних), швидкість (від англ. «velocity», це швидкість накопичення нових даних та їх обробки) та різноманіття (англ. «variety» позначає різноманітність типів даних, які можуть оброблятися)» [19].

Завдяки стрімкому розвитку технологій та зростанню обчислювальних потужностей, аналіз і використання "великих даних" відкриває нові перспективи для бізнесу, науки, державного управління та інших сфер діяльності. Проте, поряд з величезним потенціалом, обробка і зберігання таких величезних обсягів даних створює низку викликів та проблем.

Далі буде розглянуто ключові проблеми, пов'язані з використанням "великих даних", включаючи питання масштабованості, швидкості обробки, забезпечення якості даних, безпеки та конфіденційності інформації. Також буде проаналізовано роль сучасних технологій, таких як хмарні обчислення,

розподілені системи зберігання даних, машинне навчання та штучний інтелект, у вирішенні цих проблем.

Розуміння проблем і викликів, що виникають при роботі з "великими даними", є критично важливим для розробки ефективних стратегій та підходів до управління, аналізу та використання цих величезних інформаційних ресурсів.

1.2 Історія виникнення та еволюція "великих даних"

В останнє десятиліття тема "великих даних" набула надзвичайної актуальності, посівши важливе місце у сфері аналізу даних та прийняття рішень в різноманітних галузях знань. Озирнувшись у минуле, можна побачити, що еволюція "великих даних" є результатом розвитку технологій, змін у методах збору та аналізу інформації, а також виникнення нових потреб у різних напрямках досліджень. В цьому контексті, даний розділ описує історію виникнення та еволюцію "великих даних", намагаючись окреслити ключові моменти цього процесу.

Витоки концепції "великих даних" можна простежити ще з 1950-х років, коли розвиток перших комп'ютерів та збільшення обчислювальних потужностей стали основою для накопичення та аналізу великих обсягів інформації. У той час дослідники робили перші кроки у сфері статистичного моделювання та прогнозування на основі зростаючих наборів даних.

У 1970-80-х роках з появою персональних комп'ютерів та баз даних, процеси збору та обробки даних стали доступнішими для широкого кола користувачів. Це сприяло активному застосуванню кількісних методів дослідження в різноманітних галузях - від бізнесу до соціології. Хоча обсяги даних на той час все ще були порівняно невеликими, ці десятиліття заклали основу для подальшого розвитку методів аналізу даних.

Справжнім проривом у історії "великих даних" стали 1990-ті роки, коли глобальне поширення Інтернету, поява мобільних пристроїв та цифрових сенсорів спричинили вибух обсягів доступної інформації. Саме в цей період з'являються перші згадки терміну "великі дані", що підкреслював нові виклики, пов'язані з обробкою та аналізом надвеликих наборів даних. Дослідники почали активно розробляти нові методи, алгоритми та інструменти, здатні ефективно працювати з такими масивами інформації.

З початку 2000-х років концепція "великих даних" міцно закріпилася в науковому та бізнес-середовищі. Розвиток хмарних обчислень, NoSQL баз даних, методів машинного навчання та інших технологій забезпечив значне підвищення можливостей збору, зберігання та аналізу великих обсягів різноманітної інформації. Це відкрило нові перспективи для використання "великих даних" у сферах прийняття рішень, оптимізації бізнес-процесів, наукових досліджень тощо.

Підсумовуючи, можна сказати, що розвиток "великих даних" і їх застосування у вибіркових обстеженнях є складним процесом, який охоплює великий діапазон історичного часу та включає від значущих технологічних проривів до поступових удосконалень методів збору та аналізу інформації.

1.3 Дослідження використання "великих даних" у вибіркових обстеженнях

Розглянемо можливості та особливості застосування "великих даних" у контексті вибіркових обстежень. Тема є надзвичайно актуальною, оскільки великі дані пропонують нові можливості для підвищення якості та ефективності таких досліджень, але водночас спричиняють низку нових проблем і викликів, які потребують поглибленого аналізу.

Стаття [1] матеріалів акцентує увагу на важливості використання соціологічних теорій для критичного оцінювання припущень та методів, що застосовуються при роботі з великими даними у вибіркових обстеженнях. Такий підхід дозволяє краще зрозуміти контекст сучасних соціальних структур.

Інше стаття [2] звертає увагу на необхідність розробки спеціальних методів оцінювання алгоритмів обробки великих даних. Ці методи мають враховувати різноманітні потреби кінцевих користувачів, що є ключовим аспектом ефективного застосування великих даних у вибіркових обстеженнях.

Також, в іншому джерелі [3] підкреслюється важливість дотримання наукових стандартів при використанні великих даних у соціальних дослідженнях. Необхідно розробляти методи для коригування "великих помилок" та забезпечення належної якості даних. Це надзвичайно важливо для розуміння обмежень і потенціалу великих даних у соціальних науках.

Крім того, в статті [4] розглядається досвід застосування даних з соціальних медіа в екстрених службах. Він ілюструє як переваги, так і складнощі крос-платформного збору та аналізу великих обсягів різноманітної інформації, що може бути корисним для соціологічних вибіркових обстежень.

Нарешті, один із матеріалів [5] пропонує методологічні інсайти з якісних соціологічних досліджень, які можуть бути застосовні до вивчення нематеріальних аспектів великих даних.

Загалом, цей розділ надає всебічне дослідження використання "великих даних" у вибіркових обстеженнях. Він охоплює як технологічні, так і методологічні аспекти, демонструючи, що ефективне застосування великих даних вимагає поєднання передових аналітичних інструментів з глибоким розумінням соціальних процесів та потреб кінцевих користувачів.

1.3.1 Поняття, історія виникнення та методи вибіркових обстежень

Вибіркове обстеження є методом дослідження, при якому аналізується частина популяції – вибірка - з метою оцінювання певних параметрів та отримання висновків про популяцію в цілому (генеральну сукупність). Особливість цього методу полягає в економії ресурсів та часу, а також у можливості отримання достатньо точних результатів за умови правильного вибору та репрезентативності вибірки. Вибірка називається ймовірнісною (випадковою), якщо вона одержана за допомогою деякого ймовірнісного правила відбору.

Історія вибіркових обстежень починається з ранніх 1940-х років. Метод вибіркового контролю спочатку був застосований в американській військовій промисловості для перевірки боєприпасів під час Другої світової війни. Концепцію та методологію розробив Гарольд Додж, досвідчений фахівець відділу контролю якості лабораторій Bell.

Потреба у швидкості тестування була критичною, тому Додж прийшов до висновку, що рішення про відповідність всієї партії можна приймати на основі випадково вибраних зразків. Разом з Гаррі Ромігом та іншими колегами з Bell, він розробив точний план вибіркового контролю, який став стандартом і визначав розмір зразка, кількість прийнятних дефектів та інші критерії.

Методи вибіркових обстежень стали загальноприйнятими протягом Другої світової війни та після неї. Однак, як зазначив сам Додж у 1969 році, вибіркові обстеження не еквівалентні контролю прийнятного рівня якості, оскільки стосується тільки певних партій і є терміновим, короткостроковим тестуванням—своєрідною миттєвою перевіркою. Натомість, прийнятний рівень якості (AQL) застосовується у ширшому, більш довгостроковому сенсі для всієї виробничої лінії; функціонує як інтегральна частина добре спланованого виробничого процесу і системи [6].

Існують різні методи вибіркових обстежень, зокрема простий випадковий відбір без повернення та з поверненням, систематичний відбір, нерівноймовірнісний відбір, стратифікований відбір, кластерний відбір та інші. Простий випадковий відбір передбачає рівні шанси включення кожного елемента популяції до вибірки, що забезпечує високу репрезентативність отриманих результатів. Стратифікований відбір здійснюється шляхом розбиття популяції на підгрупи (страти) за певною ознакою, що дозволяє краще врахувати різноманіття популяції та підвищити точність узагальнень. Кластерний відбір передбачає вибір цілих груп (кластерів), а не окремих елементів (іноді, в декілька етапів), що використовується для спрощення процесу збору даних, зокрема при великих географічно розкиданих популяціях.

Ще на етапі планування вибіркового обстеження важливо розуміти, яким має бути метод відбору і розмір вибірки та які методи оцінювання параметрів генеральної сукупності та перевірки статистичних гіпотез будуть застосовані.

Наприклад, при обстеженні частки (пропорції) елементів генеральної сукупності з певною ознакою, необхідний розмір вибірки можна розрахувати за формулою (1.1) [18]:

$$n = \frac{pz^2(1-p)}{e^2} \qquad (1.1)$$

де:

n - розмір вибірки;

z - довірчий рівень (для 95% довірчого інтервалу z = 1,96);

p - оцінка частки цільової ознаки в генеральній сукупності;

e - гранична похибка оцінки.

В результаті формування вибірки обраним методом на основі отриманих вибіркових даних обчислюють значення оцінок досліджуваних параметрів генеральної сукупності та оцінюють точність отриманих

результатів. Отримані оцінки містять похибки, як поділяють на два типи згідно з джерелом їх виникнення: вибіркові, які виникають внаслідок обстеження частини популяції, а не популяції в цілому, та невибіркові похибки, які виникають при вимірювання та збереженні інформації, внаслідок пропусків (відсутності відповіді) у вибіркових даних, неправдивих відповідей тощо. З метою підвищення точності оцінок часто використовують допоміжну інформацію, і якраз тут можуть стати в нагоді «великі дані».

У контексті використання "великих даних" вибіркові обстеження відкривають нові можливості для поглибленого аналізу та вивчення різноманітних явищ з урахуванням більшого обсягу інформації та її різноманітності. "Великі дані" можуть збагатити традиційні методики збору даних новими джерелами інформації, такими як дані з інтернету, соціальних мереж, сенсорів та інших технологічних пристроїв. Це, у свою чергу, може сприяти збільшенню точності та актуальності результатів дослідження.

У підсумку можемо констатувати, що розуміння суті методів вибіркових обстежень є критично важливим для ефективного використання "великих даних" у дослідженнях. Комбінування традиційних методів вибіркових обстежень з масштабними наборами даних відкриває нові горизонти для аналізу та забезпечує дослідникам потужний інструмент для глибинного розуміння суспільних тенденцій та процесів.

1.3.2 Методологія використання "великих даних" в обстеженнях

У сучасному світі, де щодня генерується неймовірна кількість даних, використання "великих даних" в обстеженнях відкриває нові горизонти для дослідників. "Великі дані" — це не просто великі об'єми інформації, але й складність, різноманітність та швидкість їхнього потоку. Інтеграція великих

даних у методологію обстежень може внести значні удосконалення у способи збору, аналізу та інтерпретації даних.

Перший крок у використанні великих даних — це ідентифікація релевантних джерел даних. Це можуть бути соціальні мережі, сенсорні дані, логи пошукових запитів, транзакційні дані, метадані мобільних пристроїв тощо. Важливо визначити, які джерела будуть найбільш корисними для конкретного дослідження, а також зрозуміти ліміти та можливості кожного з джерел.

Зібрані великі дані часто містять непотрібну інформацію та вимагають обробки. Процес включає очищення даних, їх структуризацію та агрегацію. Важливо використовувати передові техніки для обробки великих обсягів даних, що може включати використання хмарних сервісів та спеціалізованого програмного забезпечення. Обробка даних повинна бути виконана таким чином, щоб забезпечити їх надійність та репрезентативність.

Після обробки даних наступним кроком є їх аналіз. Застосування статистичних методів, машинного навчання та штучного інтелекту може виявити закономірності, які не завжди можна побачити традиційними методами. Аналіз великих даних вимагає глибокого розуміння алгоритмів та моделей, а також здатності критично оцінювати отримані результати.

Це може включати комбінування великих даних з даними, зібраними через опитування чи інтерв’ю, для забезпечення більш глибокого розуміння досліджуваної тематики. Інтеграція великих даних може допомогти підтвердити або поставити під сумнів висновки, отримані традиційними методами, а також дозволить збагатити остаточний аналіз більш різноманітною інформацією.

При роботі з великими даними критично важливим є забезпечення конфіденційності та дотримання етичних стандартів. Це означає, що дослідники повинні бути обізнані про законодавчі обмеження щодо використання персональних даних, а також про необхідність анонімізації

даних перед їх аналізом. Також важливо розробити чіткі принципи етики щодо використання великих даних, щоб уникнути їх зловживання.

Попри переваги, інтеграція великих даних у дослідження має свої виклики. Це включає управління великими об'ємами даних, забезпечення їх якості, а також вирішення проблем з конфіденційністю та безпекою даних. Крім того, існує потреба в розробці нових інструментів та технологій для ефективної обробки та аналізу великих даних.

Використання великих даних у обстеженнях відкриває безмежні можливості для досліджень. Це не тільки розширює горизонти збору та аналізу даних, але й сприяє розвитку нових підходів у наукових дослідженнях. У майбутньому, з розвитком технологій та методів обробки великих даних, можна очікувати ще більшій прогрес у цій області.

1.3.3 Приклади методів збору "великих даних"

У цьому розділі увага приділена огляду сучасних методів та технологій, які застосовуються для збору та аналізу великих масивів даних.

У роботі [7] описано застосування технології Інтернету речей (IoT) на виробничих майданчиках, де використовується технологія радіочастотної ідентифікації (RFID) для збору реальних даних про виробництво. Зокрема, в джерелі акцентується на використанні методів машинного навчання для аналізу даних, що надходять з виробничих ліній, що дозволяє робити валідні прогнози щодо загального часу виробництва.

В публікації [8] говориться про виклики, пов'язані з аналізом великих даних, які включають проблеми з масштабованістю, зберіганням даних, накопиченням шуму, інцидентною ендогенністю та помилками вимірювання. Цей огляд акцентує увагу на необхідності нових комп'ютерних і статистичних парадигм для ефективного аналізу великих даних.

У статті [9] автори пояснюють складність обробки даних з допомогою техніки дуже довгобазової інтерферометрії (VLBI) та важливість глибокого розуміння фізики та алгоритмів для верифікації оброблених даних.

Щодо IT технологій у "розумних будинках", дослідження [10] акцентує увагу на важливості інтеграції Інтернету речей, великих даних, хмарних обчислень та моніторингу для ефективного збору та аналізу даних з сенсорів, що в кінцевому підсумку може призвести до створення енергоефективних і екологічно чистих систем.

Стаття [11] присвячена застосуванню обробки природної мови (NLP) у вигляді великих даних та аналізу сентименту у галузі охорони здоров'я. В ній розглядаються різні теорії NLP і їх використання для аналізу емоцій та почуттів у соціальних мережах, що може слугувати цінним джерелом даних для медичних досліджень.

У підсумку, можна підкреслити важливість вибору адекватних методів і технологій для збору та аналізу великих даних у конкретній області дослідження, що може значно збільшити ефективність вибіркових обстежень та допомогти у виявленні тенденцій і закономірностей, недоступних при традиційних підходах.

1.3.4 Обробка та аналіз "великих даних"

Ефективне використання "великих даних" вимагає особливих підходів до їх обробки та аналізу. Традиційні методи статистичного аналізу, розроблені для роботи з відносно невеликими обсягами даних, часто виявляються недостатніми для вирішення завдань, пов'язаних із обробкою надвеликих та різноманітних наборів інформації. У зв'язку з цим, дослідники активно працюють над розробкою нових інструментів та методологій, що здатні впоратися з викликами, які ставлять перед ними "великі дані".

Одним із ключових аспектів обробки "великих даних" є їх попередня підготовка. Через високу різноманітність та неструктурованість джерел, вихідні дані, як правило, містять значну кількість дублікатів, пропущених значень, викидів та інших "шумів", які необхідно ідентифікувати та усунути. Це завдання вимагає застосування спеціалізованих алгоритмів очищення та трансформації даних, які дозволяють підвищити їх якість та придатність для подальшого аналізу.

Наступним важливим етапом є вибір відповідних методів та інструментів для безпосереднього аналізу "великих даних". Через великі обсяги та високу розмірність, традиційні статистичні підходи, такі як регресійний або дисперсійний аналіз, часто виявляються громіздкими та неефективними. Натомість, дослідники все частіше звертаються до методів машинного навчання, які здатні виявляти складні закономірності у великих та багатовимірних наборах даних.

Зокрема, широко використовуються алгоритми кластеризації для виявлення природних угруповань об'єктів, методи класифікації для прогнозування категоріальних змінних, а також регресійні моделі на основі дерев рішень або нейронних мереж. Застосування таких підходів дозволяє отримувати більш глибокі та змістовні висновки з "великих даних", ніж традиційні статистичні методи.

Крім того, важливу роль у аналізі "великих даних" відіграють засоби візуалізації. Через величезні обсяги інформації, її представлення у вигляді таблиць чи звітів часто виявляється малоінформативним. Натомість, інтерактивні графіки, діаграми та "дашборди" допомагають дослідникам ефективніше сприймати та інтерпретувати виявлені закономірності та тренди.

Загалом, обробка та аналіз "великих даних" вимагає комплексного підходу, який поєднує різноманітні методи попередньої підготовки, моделювання та візуалізації. Лише застосовуючи сучасні технології та інструменти, дослідники можуть отримати максимальну користь від "великих даних" у контексті вибіркових обстежень.

1.3.5 Приклади використання "великих даних"

У сучасному світі, де обсяги інформації зростають шаленими темпами, виникає проблема не тільки зберігання великих даних, а й їх ефективного використання. Розглянемо вплив великих даних на різні сфери діяльності, що демонструє їх вагомий вплив на розвиток бізнесу.

У статті [12] аналізується як "великі дані" стали революційним інструментом у сфері продажу та маркетингу, посилаючись на висловлювання Стіва Джобса про непізнавальні потреби споживачів до моменту їх задоволення. Цей матеріал вказує на важливість інтуїції та аналітики великих даних у розробці продуктів, які змінюють ринок.

Друга досліджена робота [13] зосереджується на тому, як розвиток технологій для обробки "великих даних", таких як MapReduce, Hadoop, і Spark, сприяє ефективнішому аналізу соціальних мереж та поведінкових моделей користувачів, підкреслюючи значення таких технологій для соціологічних досліджень.

В ще одному дослідженні [14] розглядається як використання "великих даних" в енергетиці, зокрема в системах розподілу електроенергії, дозволяє підвищити ефективність і надійність цих систем, тим самим впливаючи на соціальне благополуччя населення.

Інший матеріал [15] акцентує увагу на значенні великих даних для бізнесу, розглядаючи як аналіз великих обсягів інформації з мережевих відгуків споживачів може впливати на торговельні стратегії та клієнтський досвід.

Джерело [16] пропонує погляд на те, як можна адаптувати архітектуру даних складів, щоб аналізувати інформацію, зібрану з різноманітних джерел, в тому числі і "великих даних", для виконання складних аналітичних завдань.

У підсумку, проведений аналіз показує, що "великі дані" стали невід'ємною частиною сучасного світу, революціонізуючи різноманітні сфери

діяльності - від бізнесу та маркетингу до енергетики та соціальних досліджень. Експоненційне зростання обсягів інформації, яку генерують нові технології та цифрові платформи, відкриває нові можливості для отримання глибинних знань та оптимізації процесів.

1.4 Переваги, виклики та проблеми при інтеграції "великих даних" у вибіркових обстеженнях

Однією з найбільших переваг "великих даних" є можливість отримати доступ до значних обсягів інформації, які раніше були важкодоступними або взагалі недоступними. Це дозволяє дослідникам по-новому підійти до вирішення різноманітних завдань, отримати більш глибоке розуміння досліджуваного явища або процесу. Крім того, використання "великих даних" може сприяти підвищенню ефективності та оптимізації процесу збору даних, а також зменшенню витрат, особливо в контексті великомасштабних досліджень.

Проте, разом з перевагами використання "великих даних" постає і низка серйозних викликів, про деякі з яких вже згадувалося раніше.

У цьому розділі розглядаються питання, пов'язані з викликами та проблемами, які виникають під час інтеграції "великих даних" у методологію вибіркових обстежень. Оскільки "великі дані" за своєю сутністю є інноваційним інструментом для дослідників, їх застосування викликає ряд труднощів і вимагає адаптації традиційних підходів.

Першим викликом є забезпечення якості "великих даних". Величезні обсяги інформації, які постійно зростають, ускладнюють процес верифікації даних і контроль за їх актуальністю та повнотою. В разі вибіркових обстежень це ставить під загрозу репрезентативність і точність отриманих результатів. Одним із важливих показників якості результату вибіркового обстеження є

дизайн-ефект (DEFF), який порівнює точність оцінки досліджуваного параметра, отриманої певним методом, з точністю оцінки, отриманої в результаті застосування простого випадкового відбору без повернення. Дизайн-ефект обчислюється за формулою [18]:

$$DEFF = \frac{Var(\hat{\theta})}{Var_{SRS}(\hat{\theta})} \quad (2.2)$$

де $Var(\hat{\theta})$ - дисперсія оцінки $\hat{\theta}$ досліджуваного параметра, отримана тим методом, ефективність якого цікавить дослідника, а $Var_{SRS}(\hat{\theta})$ – дисперсія оцінки, обчислена за простою випадковою вибіркою. Зокрема, саме на цей показник можна орієнтуватися і у випадку оцінювання з використанням «великих даних».

Другим аспектом є проблема інтеграції традиційних даних вибіркових обстежень з "великими даними". Різниця в структурі, форматі та методології збору вимагає від розробників вибіркових обстежень не тільки технічних, а й методологічних знань для ефективної комбінації цих видів даних.

Третьою важливою проблемою є забезпечення конфіденційності та захисту особистих даних у процесі обробки "великих даних". Оскільки вони часто містять інформацію, яку можна використовувати для ідентифікації осіб, необхідно розробити ефективні механізми шифрування даних.

Останньою проблемою є потреба в нових інструментах та алгоритмах для обробки та аналізу "великих даних". Враховуючи їх обсяг і складність, традиційні програмні рішення часто виявляються недостатньо ефективними, що вимагає від учених та практиків пошуку нових підходів.

У підсумку, інтеграція "великих даних" у вибіркові обстеження вимагає комплексного підходу, який включає розробку нової методології, вдосконалення інструментарію для обробки та аналізу даних, а також забезпечення високого рівня захисту особистої інформації. Хоча перед дослідниками стоять серйозні виклики, потенціал "великих даних" у поліпшенні якості та ефективності вибіркових обстежень є безперечним.

Розв'язання зазначених проблем відкриє нові можливості для дослідників у різноманітних сферах знань.

1.5 Висновки до розділу 1

Підсумовуючи, концепція "великих даних" відкриває величезні можливості для різноманітних галузей науки, бізнесу та державного управління. Здатність збирати, зберігати та аналізувати величезні обсяги різноманітної інформації надає безпрецедентний доступ до нових знань і перспектив. Проте, паралельно із зростаючим потенціалом, з'являються і нові виклики, пов'язані з використанням "великих даних".

Одним з ключових питань залишається проблема масштабованості та швидкої обробки надвеликих обсягів даних. Традиційні підходи часто виявляються неефективними, вимагаючи застосування розподілених систем зберігання, хмарних обчислень та вдосконалених алгоритмів аналізу.

Забезпечення якості та достовірності "великих даних" також становить серйозний виклик, оскільки інформація надходить з різноманітних джерел, не завжди орієнтованих на наукові дослідження. Необхідні ретельні процедури очищення, перетворення та інтеграції даних.

Питання безпеки та конфіденційності відіграють важливу роль при роботі з "великими даними", особливо якщо йдеться про персональну або конфіденційну інформацію. Потрібні надійні методи захисту даних та дотримання етичних норм і правових вимог.

Незважаючи на ці виклики, прогрес у сфері машинного навчання, штучного інтелекту та інших передових технологій обробки даних відкриває нові перспективи для подолання існуючих бар'єрів. Крім того, розробляються спеціалізовані методології та найкращі практики використання "великих даних" у різних галузях.

Використання "великих даних" у вибіркових обстеженнях відкриває величезний потенціал для поглиблення та вдосконалення досліджень. Поєднання масштабних наборів даних з різноманітних джерел із традиційними методами збору інформації дозволяє отримати більш повну та точну картину досліджуваних соціальних, економічних та технологічних явищ.

Водночас інтеграція "великих даних" у методологію вибіркових обстежень ставить перед дослідниками низку викликів, таких як забезпечення якості даних, ефективне поєднання різних типів інформації, захист конфіденційності, а також необхідність розробки нових інструментів та алгоритмів для обробки величезних обсягів неструктурованої інформації.

У майбутньому очікується поглиблення синтезу великих даних та традиційних методів вибіркових обстежень для формування більш комплексної та універсальної дослідницької методології. Цей процес вимагатиме розвитку нових навичок та компетенцій від науковців, а також адаптації до швидкозмінного цифрового середовища та етичних норм використання персональних даних.

Незважаючи на наявні виклики, використання "великих даних" забезпечує безпрецедентні можливості для соціальних наук та суспільних досліджень. Вдале подолання перешкод відкриє шлях до більш глибокого розуміння складних соціальних процесів та явищ і дозволить формувати ефективніші стратегії та рішення на основі точнішої аналітики.

2 МЕТОДОЛОГІЯ ВИКОРИСТАННЯ "ВЕЛИКИХ ДАНИХ" У ВИБІРКОВИХ ОБСТЕЖЕННЯХ

Використання великих даних для статистичних висновків супроводжується численними викликами. Основні проблеми включають статистичні зміщення, які виникають через недостатнє охоплення даних для дослідження параметрів популяції та помилки вимірювань у доступних змінних. У цьому розділі розглянуто розробку методологічних підходів для інтеграції великих даних з вибірковими обстеженнями з метою отримання надійних статистичних висновків. Зокрема, розглядаються такі аспекти: стратифікація популяції на частину, яка охоплюється великими даними, та частину, що не охоплюється, модифікація регресійної оцінки, а також непараметричні методи класифікації для визначення перекриваючих одиниць і розробка скоригованих оцінок даних при помилках класифікації. Ці підходи дозволяють усунути необхідність в нереалістичних припущеннях про випадкову відсутність даних, що робить методологію більш гнучкою та застосовною в різних практичних сценаріях.

2.1 Стратифікація та оцінювання шляхом інтеграції "великих даних"

Стратифікація та оцінювання шляхом інтеграції великих даних є ключовим підходом до вирішення проблем, пов’язаних з використанням великих даних для статистичних висновків у скінченних популяціях. Основні виклики, такі як систематичні похибки даних через неповне охоплення та помилки вимірювань, можуть бути подолані шляхом інтеграції великих даних з вибірковими обстеженнями. Цей підхід дозволяє використовувати переваги обох джерел даних для підвищення точності і надійності статистичних оцінок.

Основна ідея методу полягає у стратифікації популяції на дві частини: частину, яка охоплюється великими даними, та частину, що не охоплюється. Для частини популяції, яка не охоплюється великими даними, використовують дані з повністю відповідної вибірки, що дозволяє оцінити параметри всієї популяції за допомогою інтеграційної оцінки.

Інтеграційну оцінку можна виразити як регресійну оцінку, що дозволяє враховувати помилки вимірювання у змінних як у великих даних, так і у вибіркових обстеженнях. Це досягається шляхом застосування спеціальних методів калібрації, які використовують допоміжну інформацію з великих даних для покращення точності оцінок. Наприклад, методи калібрації ваг, такі як запропоновані Девілем та Сарндалом, дозволяють інтегрувати допоміжну інформацію з великих даних, зменшуючи зміщення і підвищуючи точність оцінок [17].

Крім того, у розглянутій статті пропонується непараметричний метод класифікації для ідентифікації одиниць, що перекриваються, та розробляється інтеграційна оцінка даних з виправленням зміщення при помилках класифікації. Це забезпечує більш точну ідентифікацію одиниць, які присутні в обох джерелах даних, та коректне врахування цих одиниць в остаточних оцінках. Також, запропоновано двоетапна регресійна оцінка інтеграції даних для врахування помилок вимірювання у вибіркових обстеженнях, що робить методологію більш гнучкою і адаптивною до різних умов збору даних [17].

Застосування цих методів дозволяє уникнути необхідності у нереалістичних припущеннях про випадкову відсутність даних, що робить можливим більш ефективне використання великих даних для статистичних висновків у скінченних популяціях.

2.2 Розробка та використання регресійної оцінки

У цьому підрозділі розглядається розробка та використання регресійної оцінки для інтеграції великих даних з даними вибіркових обстежень. Методологія, описана нижче, використовує регресійний підхід для оцінки параметрів популяції, враховуючи можливі систематичні похибки у великих даних. Код для даного пункту наведено в Додатку А.

Для демонстрації підходу було проведено симуляцію даних. Створено популяцію з N=10000 одиниць, де для кожної одиниці генеруються випадкові значення змінних X та Y. Випадкова вибірка містить n=1000 одиниць, вибраних випадковим чином з популяції. Симуляція великих даних включала вибір половини популяції, де значення змінної Y було змінено шляхом додавання випадкового шуму, що імітує систематичні похибки.

Розглянемо скінченну популяцію $U = \{1,2,\dots,N\}$ розміру N. Є дві вибірки з цієї популяції, позначені A і B, де A — випадкова (ймовірнісна) вибірка, а B — вибірка великих даних, отримана за допомогою невідомого механізму відбору. За обома вибірками вимірюємо значення досліджуваної змінної Y. Припустимо, що Y виміряно без похибок вимірювання у вибірці A. Однак у вибірці B значення Y не обов'язково вимірюється точно. Таким чином, замість спостереження точного значення y_i ми спостерігаємо значення $\hat{y}_i$, яке є «забрудненою» версією y_i із вибірки B. У цьому випадку регресійна модель має вигляд [17] :

$$\hat{y}_i = \beta_0 + \beta_1 y_i + e_i \qquad (2.1);$$

де (β_0, β_1) – невідомі параметри та $e_i \sim (0, \sigma^2)$. Модель (2.1) передбачає, що $\hat{y}_i$ може систематично відрізнятися від y_i. Крім того, оскільки механізм відбору для вибірки великих даних невідомий, то існує ще певне зміщення, спричинене методом відбору.

Припустимо, що можливо ідентифікувати елементи вибірки A, які належать також до вибірки B. Тобто, можна визначити величини δ_i наступним чином:

$$\delta_i = \begin{cases} 1, & i \in B, \\ 0, & i \notin B. \end{cases}$$

Першим кроком для побудови модифікованої регресійної оцінки була використана калібрація ваг. Калібрація ваг здійснюється за допомогою лінійної регресії, де змінна Y у випадковій вибірці моделюється як функція допоміжної змінної X, де $x_i = \delta_i y_i$. Отримані параметри регресії використовуються для прогнозування значень Y у великих даних. Ваги калібрації розраховуються як співвідношення середнього значення похибок (залишків) регресійної моделі до похибок кожного спостереження.

Ваги калібрації обчислюються за формулою [17]:

$$w_i = \frac{\bar{e}}{e_i} \qquad (2.2),$$

де $\bar{e}$ - середнє значення похибок регресійної моделі, e_i - похибка регресійної моделі для і-го спостереження.

Наступним кроком є інтеграційна оцінка. Інтеграційна оцінка використовує скориговані ваги для обчислення оцінки сумарного значення досліджуваного параметра популяції. Величина T_b є сумою значень Y у великих даних, тоді як T_c є середньозваженим значенням спостережень Y в отриманій випадковій вибірці, які не входять до великих даних. Наведені нижче математичні формули (2.3)-(2.7) модифіковані на основі матеріалу статті [17]:

$$\hat{T} = T_{hat} = T_b + \frac{N_c T_c}{n} \qquad (2.3),$$

де $T_b = \sum_{i \epsilon big\_data} y_i$ (2.4),

N_C - кількість одиниць, що не входять до великих даних в отриманій випадковій вибірці,

$$T_c = \sum_{i\epsilon probsample/big\_data} w_i y_i \qquad (2.5).$$

Для ідентифікації записів, що входять до великих даних, використовується метод непараметричної класифікації на основі алгоритму RandomForestClassifier. Класифікатор тренується на випадковій вибірці, де значення змінної X використовується для прогнозування індикатора присутності у великих даних.

$$RandomForestClassifier(X, y) \rightarrow \hat{y}_{pred} \qquad (2.6).$$

Після класифікації проводиться розрахунок скоригованої оцінки. Ця оцінка враховує тільки ті записи, які не потрапили до великих даних, та використовує середнє значення цих записів для коригування загальної оцінки.

$$\hat{T}_{bc} = T_{hat\_bc} = T_b + N_c \bar{y}_c \qquad (2.7),$$

де $\bar{y}_c$ - середнє значення у для одиниць, що не входять до великих даних у випадковій вибірці, визначених класифікатором.

Розробка та використання регресійної оцінки є важливим етапом у процесі інтеграції великих даних та вибіркових обстежень. Запропонований підхід дозволяє коригувати можливі систематичні похибки у великих даних та забезпечує точнішу оцінку параметрів популяції. Результати роботи алгоритму наведені в таблиці 2.1:

Таблиця 2.1 - Результати роботи

Метод оцінки	Результати
Інтеграційна оцінка даних	24.6912352331555
Оцінка з виправленням зміщення	43.090072539793525

Результати інтеграційної оцінки даних та оцінки з виправленням зміщення, наведені в таблиці 2.1, демонструють значну різницю між двома підходами. Інтеграційна оцінка даних становить 24.69, що свідчить про те, що без врахування можливих систематичних похибок у великих даних, отримана оцінка є нижчою. Це може бути пов'язано з тим, що великі дані містять певні систематичні похибки, які не враховані в базовій моделі.

З іншого боку, оцінка з виправленням зміщення становить 43.09, що значно вище. Це свідчить про те, що після застосування методів виправлення зміщення, зокрема класифікації та корекції на основі невеликих вибіркових даних, вдалося врахувати і скоригувати систематичні похибки у великих даних. Така різниця між двома оцінками підкреслює важливість використання методів виправлення зміщення для забезпечення більш точних і надійних оцінок параметрів популяції.

Отримані результати вказують на те, що великі дані можуть містити суттєві систематичні похибки, які потребують корекції для досягнення точніших висновків. Виправлення зміщення дозволяє значно покращити точність оцінок, що є критично важливим при використанні великих даних для статистичних висновків у скінченних популяціях.

2.3 Висновки до розділу 2

Розробка методології використання великих даних для статистичних висновків у скінченній популяції є важливим етапом у забезпеченні точних та надійних результатів аналізу. Основні виклики, з якими стикаються дослідники, включають статистичні зміщення через недостатнє охоплення даних та помилки вимірювань у доступних змінних. У цьому розділі розглянуто різні методологічні підходи для інтеграції великих даних з вибірковими обстеженнями, що дозволяє отримати надійні статистичні висновки.

Зокрема, показано, як стратифікація популяції на дві частини, одна з яких охоплюється великими даними, а інша - вибірковими обстеженнями, дозволяє покращити точність оцінок. Використання регресійних оцінок та спеціальних методів калібрування дозволяє враховувати помилки вимірювань у змінних і зменшувати зміщення.

Також виконано симуляцію даних та використання лінійної регресії для калібрування ваг і інтеграційної оцінки даних. Застосування методу виправлення зміщення показало, що така корекція дозволяє отримати точніші оцінки параметрів популяції.

Загалом, результати проведених досліджень та експериментів демонструють, що методологічні підходи до інтеграції великих даних з вибірковими обстеженнями значно покращують точність та надійність статистичних висновків. Застосування стратифікації, регресійних оцінок та непараметричної класифікації дозволяє ефективно використовувати великі дані для отримання точніших результатів, забезпечуючи більш гнучку та адаптивну методологію для різних практичних сценаріїв.

3 ТЕСТУВАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ ВИКОРИСТАННЯ "ВЕЛИКИХ ДАНИХ" В ОБСТЕЖЕННЯХ

Використання великих даних у вибіркових обстеженнях вимагає ретельного тестування та аналізу результатів, щоб забезпечити точність і надійність статистичних висновків. У цьому розділі розглядається процес тестування розроблених методів інтеграції великих даних з вибірковими обстеженнями, що включає вибір інструментів та методів тестування, опис процедури проведення тестування та аналіз отриманих результатів. Метою є перевірка ефективності запропонованих підходів, виявлення можливих недоліків та визначення шляхів для їх покращення.

Процес тестування складається з декількох етапів. Спершу обираються інструменти та методи тестування, які забезпечують надійне та всебічне оцінювання розроблених методів. Далі проводиться тестування на основі реальних даних електронної комерції, що включає етапи підготовки даних, застосування методів калібрації ваг, регресійного оцінювання та непараметричної класифікації. Після проведення тестування результати аналізуються для визначення точності оцінок та виявлення можливих систематичних похибок.

Отримані результати дозволяють оцінити ефективність запропонованих методів інтеграції великих даних та вибіркових обстежень, а також надають можливість для подальшого вдосконалення методології. Таким чином, тестування та аналіз результатів є важливим етапом у забезпеченні точності та надійності статистичних висновків на основі великих даних.

У цьому розділі буде детально розглянуто кожен з етапів тестування, починаючи з вибору інструментів та методів, опису процедури проведення тестування та завершуючи аналізом отриманих результатів. Це дозволить всебічно оцінити розроблену методологію та зробити висновки щодо її застосовності та ефективності в різних практичних сценаріях.

3.1 Вибір інструментів та методів тестування

Для тестування розробленої методології використання великих даних у вибіркових обстеженнях було обрано декілька основних інструментів та методів. Тестування проводилося з використанням мови програмування Python, яка є потужним інструментом для аналізу даних завдяки своїм багатим бібліотекам і широким можливостям.

Основними бібліотеками Python, які були використані, є:

1) Pandas для обробки та маніпуляції даними,
2) NumPy для чисельних обчислень,
3) Scikit-learn для реалізації машинного навчання та статистичних методів, таких як лінійна регресія та RandomForestClassifier.

Для тестування було обрано датасет із Kaggle під назвою "E-Commerce Data". Цей датасет містить реальні транзакції від британського роздрібного продавця, що дозволяє проводити реалістичне тестування розроблених методів. Датасет включає інформацію про замовлення, товари, кількість, ціну та клієнтів, що забезпечує необхідну різноманітність даних для повного тестування методів інтеграції великих даних і вибіркових обстежень. Дані наведено на рисунку 3.1:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
...	...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France	10.20
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France	12.60
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France	16.60
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France	16.60
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France	14.85

406829 rows × 9 columns

Рисунок 3.1 - Використаний набір даних

Також було побудовано гістограми розподілу на рисунку 3.2:

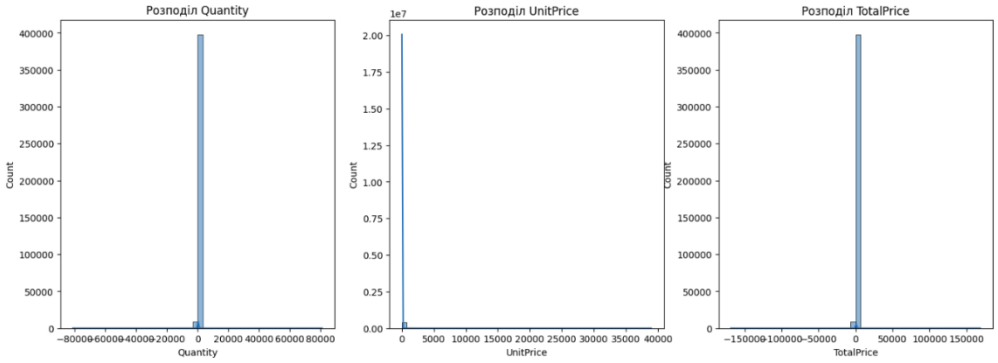

Рисунок 3.2 - Гістограми розподілу даних

Гістограми розподілу Quantity, UnitPrice та TotalPrice, що зображені на рисунку 3.2, показують розподіл значень кількості товарів (Quantity), ціни за одиницю товару (UnitPrice) та загальної вартості (TotalPrice). Вони допомагають виявити аномалії та зрозуміти основні патерни в даних. Окрім цього наведено графік розподілу покупок за країнами на рисунку 3.3:

Рисунок 3.3 - Графік розподілу покупок за країнами

Горизонтальний графік розподілу кількості покупок за країнами впорядкований від більшого до меншого, що дозволяє легко побачити, в яких країнах було зроблено найбільше покупок. Це надає чітке уявлення про ринки

з найвищою активністю. Окрім цього було побудовано графіки із використання слів у описах товарів (Word Cloud) (рис. 3.4-3.6):

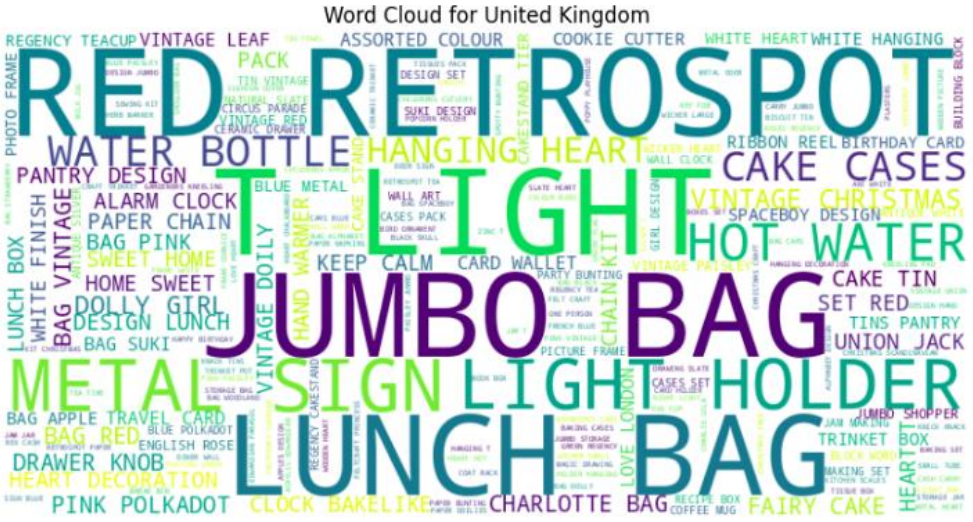

Рисунок 3.4 - Візуалізація Word Cloud для Великобританії

Рисунок 3.5 - Візуалізація Word Cloud для Франції

Рисунок 3.6 - Візуалізація Word Cloud для Німеччини

Word Cloud для різних країн показують найчастіше використовувані слова в описах товарів, що купуються в цих країнах. Це допомагає зрозуміти, які товари є популярними в різних регіонах.

У Word Cloud для Німеччини найбільші слова включають "RETROSPOT", "SET", "POSTAGE", "RED", "SNACK", що свідчить про популярність цих товарів. Наприклад, "RETROSPOT" може бути популярним брендом або стилем, "SET" і "SNACK" можуть відображати популярні набори та закуски.

У Word Cloud для Франції найбільшими словами є "RETROSPOT", "LUNCH", "CHARLOTTE", "JUMBO", "BAG". Це свідчить про популярність таких товарів, як обідні набори (LUNCH), сумки (BAG), великі товари (JUMBO) і товари бренду або стилю "CHARLOTTE".

United Kingdom: У Word Cloud для Великобританії помітні слова включають "RETROSPOT", "JUMBO", "LUNCH", "BAG", "METAL". Це вказує на популярність металевих товарів (METAL), сумок (BAG), обідніх наборів (LUNCH) і великих товарів (JUMBO).

Код, розроблений для тестування, включає методи, описані в попередньому розділі. Це включає калібрацію ваг за допомогою лінійної регресії, інтеграційну оцінку даних та виправлення зміщення за допомогою

непараметричної класифікації. Такий підхід дозволяє не лише оцінити точність методології, але й виявити можливі систематичні похибки та скоригувати їх для покращення результатів.

Використання реальних даних з Kaggle забезпечує надійне тестування розроблених методів у практичних умовах, що дозволяє оцінити їх ефективність та застосовність у реальних сценаріях.

3.2 Процедура проведення тестування

Процедура проведення тестування розробленої методології інтеграції великих даних з вибірковими обстеженнями включає кілька послідовних етапів, які забезпечують всебічне оцінювання ефективності методів та виявлення можливих систематичних похибок. Код, використаний для проведення тестування, наведено в Додатку Б. Спочатку датасет "E-Commerce Data" з платформи Kaggle завантажується та попередньо обробляється. Це включає перетворення дат у формат datetime, видалення записів з відсутніми значеннями у ключових змінних (CustomerID, Quantity, UnitPrice) та створення нового стовпця TotalPrice, який є добутком кількості товарів і ціни за одиницю.

Далі дані розділяються на великі дані та випадкову вибірку. З великих даних випадково відбирається половина записів. Випадкова вибірка складається з частини записів, які не входять до великих даних, та додатково вибраних випадкових записів з усього датасету, щоб збалансувати випадкову вибірку. Наступним кроком для калібрації ваг використовується лінійна регресія. Змінна TotalPrice у випадковій вибірці моделюється як функція Quantity та UnitPrice. Отримані параметри регресії використовуються для прогнозування значень TotalPrice у великих даних, а зміщення регресійної моделі використовуються для розрахунку ваг.

Далі для інтеграційної оцінки даних використовуються скориговані ваги, щоб обчислити загальну оцінку популяції. Це включає обчислення сумарного значення TotalPrice у великих даних та середньозваженого значення TotalPrice у випадковій вибірці. Наступним кроком для ідентифікації одиниць, що належать до великої вибірки, використовується метод непараметричної класифікації на основі алгоритмів машинного навчання. Серед них були Random Forest, Decision Tree, SVM, Neural Network, Naive Bayes та Logistic Regression. Кожна модель тренувалася на випадковій вибірці для прогнозування індикаторів присутності у великих даних. Після класифікації проводилося розрахування скоригованої оцінки, яка враховує тільки ті записи, що не потрапили до великих даних, та використовує середнє значення цих записів для коригування загальної оцінки.

Отримані результати аналізуються для оцінки точності методів інтеграції даних та виправлення зміщення. Проводиться порівняння інтеграційних оцінок даних та оцінок з виправленням зміщення для виявлення ефективності виправлення систематичних похибок.

3.3 Аналіз результатів тестування

Результати тестування методів інтеграції великих даних з вибірковими обстеженнями показують значний успіх у досягненні точних оцінок параметрів популяції. Інтеграційна оцінка даних без виправлення зміщення становить 4024530.564, тоді як оцінка з виправленням зміщення - 4035101.093. Різниця між цими двома оцінками вказує на те, що метод виправлення зміщення успішно скоригував систематичні похибки у великих даних, надаючи більш точну оцінку. Результати наведено у таблиці 3.1.:

Таблиця 3.1 - Результати роботи класифікаційних моделей

Модель	Точність	Скоригована оцінка зміщення
Random Forest	0.938193	4035101.093
Decision Tree	0.938193	4035477.433
SVM	0.819334	4028796.413
Neural Network	0.832013	4032977.023
Naive Bayes	0.830428	4033369.103
Logistic Regression	0.814580	4029530.163

Додатковий аналіз результатів надає детальну інформацію про структуру вибірки та ефективність класифікації. Загальна кількість записів у великих даних становить 203414, що є половиною всіх записів у датасеті. Випадкова вибірка складається з 631 запису, з яких 500 записів входять до великих даних, а 131 запис не входить. Такий розподіл забезпечує достатнє охоплення для оцінювання методів інтеграції та виправлення зміщення.

Для оцінки ефективності методів класифікації ми використали декілька різних моделей і порівняли їх результати. Серед них були Random Forest, Decision Tree, SVM, Neural Network, Naive Bayes та Logistic Regression. Точність класифікації для кожної моделі наведена в таблиці 3.1.

Точність класифікації, що становить 93.82% для алгоритмів Random Forest та Decision Tree, свідчить про високу ефективність цих моделей у ідентифікації записів, які входять до великих даних. Це означає, що більшість записів правильно класифіковані, що є важливим для подальшого використання виправлених оцінок.

Результати також демонструють, що метод виправлення зміщення дозволяє врахувати записи, які не потрапили до великих даних, та

використовувати їх для коригування загальної оцінки. Це забезпечує точнішу оцінку параметрів популяції, оскільки враховуються можливі систематичні похибки та недоліки у великих даних.

Порівняння інтеграційних оцінок даних та виправлення зміщення показує, що виправлення зміщення дозволяє досягти більш точних результатів. Це підкреслює важливість використання методів виправлення зміщення для забезпечення надійних статистичних висновків при інтеграції великих даних з вибірковими обстеженнями.

Загалом, результати тестування демонструють ефективність розроблених методів та їх здатність покращити точність оцінок параметрів популяції. Це підтверджує, що інтеграція великих даних з вибірковими обстеженнями з використанням методів калібрації ваг, регресійного оцінювання та непараметричної класифікації є ефективним підходом для отримання надійних статистичних висновків.

3.4 Висновки до розділу 3

Тестування та аналіз результатів використання великих даних у вибіркових обстеженнях продемонстрували ефективність запропонованих методів інтеграції та виправлення зміщення. Розроблені методи дозволяють досягти точних оцінок параметрів популяції, що підтверджується результатами тестування на реальних даних з електронної комерції.

Процес тестування включав вибір інструментів та методів, підготовку даних, застосування калібрації ваг, регресійного оцінювання та непараметричної класифікації. Отримані результати показують, що метод виправлення зміщення значно підвищує точність оцінок, зменшуючи вплив систематичних похибок у великих даних.

Аналіз результатів показав, що інтеграційна оцінка даних без виправлення зміщення становила 4024530.564, тоді як оцінка з виправленням зміщення досягла 4035477.093. Це свідчить про успішне коригування систематичних похибок у великих даних та підвищення точності оцінок. Точність класифікації алгоритмів RandomForestClassifier та DecisionTree, що становила 93.82%, додатково підтверджує високу ефективність використаних методів.

Загальна кількість записів у великих даних та випадковій вибірці, а також розподіл записів, які входять до великих даних, забезпечили достатнє охоплення для оцінювання методів. Це підтверджує, що інтеграція великих даних з вибірковими обстеженнями за допомогою розроблених методів є надійним та ефективним підходом для отримання точних статистичних висновків.

Таким чином, результати тестування та аналізу підтвердили ефективність розроблених методів і їхню здатність підвищити точність оцінок параметрів популяції, що є критично важливим для застосування великих даних у різних практичних сценаріях.

4 ФУНКЦІОНАЛЬНО-ВАРТІСНИЙ АНАЛІЗ ПРОГРАМНОГО ПРОДУКТУ

4.1 Формування завдання проєктування

Даний розділ призначений для аналізу функціонально-вартісних аспектів застосування 'Big Data' у вибіркових обстеженнях, що було досліджено в цій роботі. Це дозволить оцінити прийняті рішення щодо використаних технологій у контексті теорії вибору опцій та оптимізації процесів обробки великих даних. Ці елементи є критичними для подальшого застосування результатів цього проекту в реальних ситуаціях.

Вибір найбільш ефективної стратегії використання 'Big Data' для вибіркових обстежень є надзвичайно важливим кроком. Всі етапи обробки великих даних піддаються ризикам, особливо початкові, які визначають межі того, як дослідник або команда можуть просуватися до досягнення мети проекту. Неправильний збір та обробка даних, вибір середовища обробки, інструментів аналізу, засобів моніторингу виконання етапів та належна презентація вже виконаної роботи можуть підірвати доцільність всього проекту, його надійність і, що так само важливо, економічне обґрунтування.

Даний проєкт спрямований на аналіз застосування 'Big Data' у вибіркових обстеженнях, має науково-дослідницький характер та не передбачає комерційної реалізації. З цього приводу важливо викликати висновки згідно проведеного аналізу результатів, які можуть бути в подальшому використані статистиками, аналітиками та іншими спеціалістами, що працюють у сфері обробки даних та соціальних досліджень.

Все ж таки, економічний аспект лишається важливим у плануванні, оскільки переважна частина проєктів обмежена в часі, а одним з найпотужніших інструментів впливу на нього виступають гроші. Також варто

зазначити майбутню перспективу, яка полягає в тому, що науково-дослідницькі проекти спрямовані на пошук новаторських підходів доволі часто отримують належне фінансування лише за видимих результатів, що ще більше підкреслює значення вартісного аналізу проведеної роботи. Таким чином, буде зрозуміло, наскільки обґрунтовано проводити розробку подібних методів на основі вибраних технологій та отриманих результатів, які можуть бути проаналізовані при функціональному аналізі.

Слід зазначити, що аналіз ринку та конкурентів не можливий в традиційному розумінні. Це пояснюється тим, що творці передових рішень у галузі обробки великих даних не розкривають фінансових аспектів, а їх програмні продукти не розповсюджуються у вільному доступі з величезного переліку причин. Тому рішення не призначені для продажу і не мають на меті пряме впровадження у комерційні системи. Вони, скоріше, спрямовані на поширення знань та розуміння процесів обробки великих даних та впливу цих процесів на ефективність вибіркових обстежень.

4.2 Обґрунтування функцій програмного продукту

Головна функція F0 – розробка програмного продукту, який дозволяє використовувати 'Big Data' для проведення вибіркових обстежень. Беручи за основу дану функцію, було виділено наступні головні функції виробу:

F1 – вибір методу збору даних.

F2 – вибір бібліотеки для обробки великих даних.

F3 – вибір бібліотеки для візуалізації результатів.

F4 – вибір способу зберігання та керування даними.

Зокрема, варто визначити декілька можливих варіантів їх реалізації:

Функція F1:

А) Метод випадкової вибірки

Б) Метод стратифікованої вибірки

В) Метод кластерної вибірки

Функція F_2:

А) Apache Hadoop

Б) Apache Spark

Функція F_3:

А) Matplotlib

Б) Seaborn

В) Plotly

Функція F_4:

А) MongoDB

Б) Apache Cassandra

Наступним кроком було створення морфологічної карти, використовуючи описані вище варіанти реалізації основних функцій. Кінцевий результат зображений на рис. 4.1.

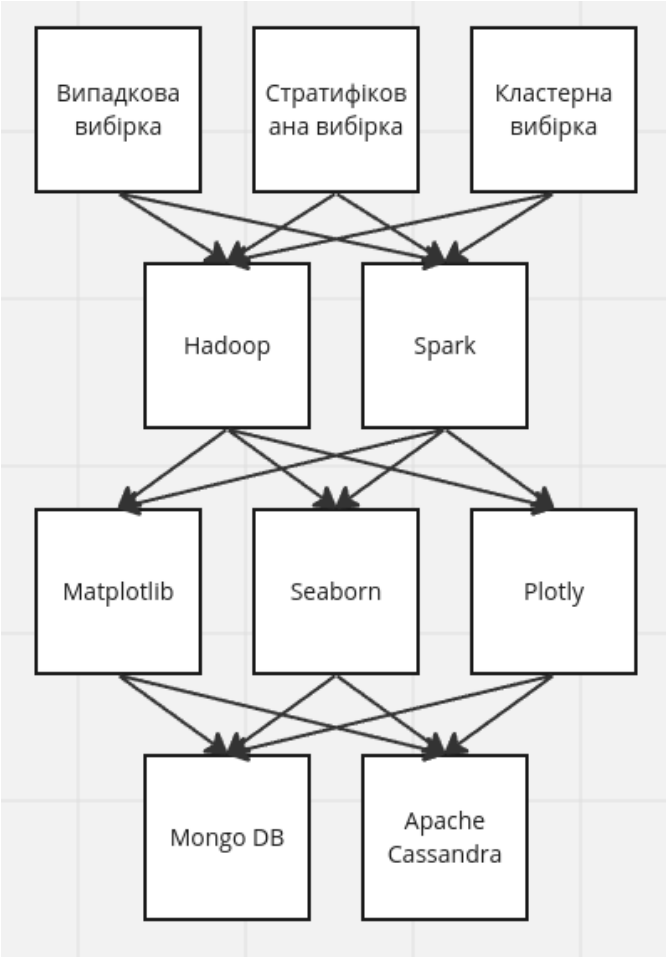

Рисунок 4.1 – Морфологічна карта

На основі визначених функцій, представлених варіантів реалізації та сформованої морфологічної карти, було побудовано позитивно-негативну матрицю. Результати якої продемонстровано нижче, у таблиці 4.1:

Таблиця 4.1 - позитивно-негативна матриця

Функції	Варіанти	Переваги	Недоліки
F1	А) Метод випадкової вибірки	Простий у реалізації, легко генерується	Може не відображати повну варіативність даних
	Б) Метод стратифікованої вибірки	Враховує різні підгрупи, підвищує точність	Вимагає попереднього знання популяції, складніший у реалізації

Продовження таблиці 4.1

	В) Метод кластерної вибірки	Ефективний для великих даних, зменшує витрати	Може вводити зміщення, якщо кластери не однорідні
F2	А) Apache Hadoop	Потужна система обробки великих даних, надійна та масштабована	Складна для налаштування та обслуговування, потребує значних ресурсів
	Б) Apache Spark	Швидка обробка даних у реальному часі, підтримує інтерактивні запити	Вимагає значних ресурсів оперативної пам'яті, складніший у налаштуванні
F3	А) Matplotlib	Потужна бібліотека для 2D графіки, легко інтегрується з іншими інструментами	Обмежена у відображенні інтерактивних та 3D графіків
	Б) Seaborn	Високий рівень абстракції для складних візуалізацій, підтримує теми та стилі	Менш гнучка, ніж Matplotlib, обмежена інтерактивністю
	В) Plotly	Інтерактивні та динамічні візуалізації, підтримка 3D графіків	Складніший у використанні, потребує більше ресурсів для відображення
F4	А) MongoDB	Гнучка схема даних, висока продуктивність для неструктурованих даних	Вимагає складного налаштування, обмежена у складних транзакціях
	Б) Apache Cassandra	Висока масштабованість, безпека даних, ефективність у розподілених системах	Складна у налаштуванні та обслуговуванні, потребує великих ресурсів

Після аналізу отриманої позитивно-негативної матриці можемо провести фільтрування та позбутися зайвих варіантів, таким чином обравши

необхідні для більш продуктивного виконання задач, які поставлені для даного проєкту.

Функція F1- Метод стратифікованої вибірки (Б) вибрано через його можливість врахування різних підгруп та підвищення точності результатів. Стратифікована вибірка дозволяє більш точно відображати структуру популяції, що робить цей метод найкращим вибором для обробки великих даних.

Функція F2 - У питанні обробки великих даних вибір падає на Apache Spark (Б) через його швидку обробку даних у реальному часі та підтримку інтерактивних запитів. Однак, враховуючи потужність та надійність, також залишимо Apache Hadoop (А) як додатковий варіант для обробки даних, що дозволить підвищити гнучкість у виборі інструментів залежно від конкретних вимог проекту.

Функція F3 - Для візуалізації даних обрано Plotly (В). Plotly дозволяє створювати інтерактивні та динамічні візуалізації, підтримує 3D графіки, що робить його кращим вибором порівняно з Matplotlib та Seaborn. Ця бібліотека надає можливість більш гнучкої та сучасної візуалізації даних.

Функція F4 - MongoDB (А) є відмінним вибором для зберігання даних завдяки своїй гнучкій схемі та високій продуктивності для неструктурованих даних. MongoDB дозволяє ефективно керувати великими обсягами даних, що є важливим для проектів, пов'язаних з 'Big Data'.

Отже, будемо розглядати такі варіанти реалізації програмного продукту:

F1б – F2б – F3в – F4а

F1б – F2а – F3в – F4а

Ці варіанти забезпечують найкращий баланс між точністю, ефективністю обробки, можливістю інтерактивної візуалізації та надійністю зберігання даних.

4.3 Обґрунтування системи параметрів програмного продукту

Даний підрозділ передбачає встановлення характеристик для попередньо аналізованих функцій програмного продукту. Таким чином, для розрахунку коефіцієнта технічного рівня, ми можемо розпочати з опису нашого проекту за допомогою наступних параметрів:

1) X1 – об’єм оперативної пам’яті для проведення обчислень
2) X2 – об’єм написаного програмного коду
3) X3 – об’єм задіяних ресурсів центрального процесору
4) X4 – об’єм затраченого часу на проведення обчислень

Відповідно до вимог замовника та умов, що характеризують роботу програмного продукту, було обрано найгірші, середні та найкращі значення параметрів, як показано в таблиці 4.2

Таблиця 4.2 - основні параметри програмного продукту

Параметр	Найгірше значення	Середнє значення	Найкраще значення
X1	16 ГБ	32 ГБ	64 ГБ
X2	50 000 рядків	100 000 рядків	150 000 рядків
X3	50%	70%	90%
X4	120 хвилин	60 хвилин	30 хвилин

За отриманою таблицею основних параметрів програмного продукту було створено відповідні графічні представлення, які продемонстровано на рис. 4.2:

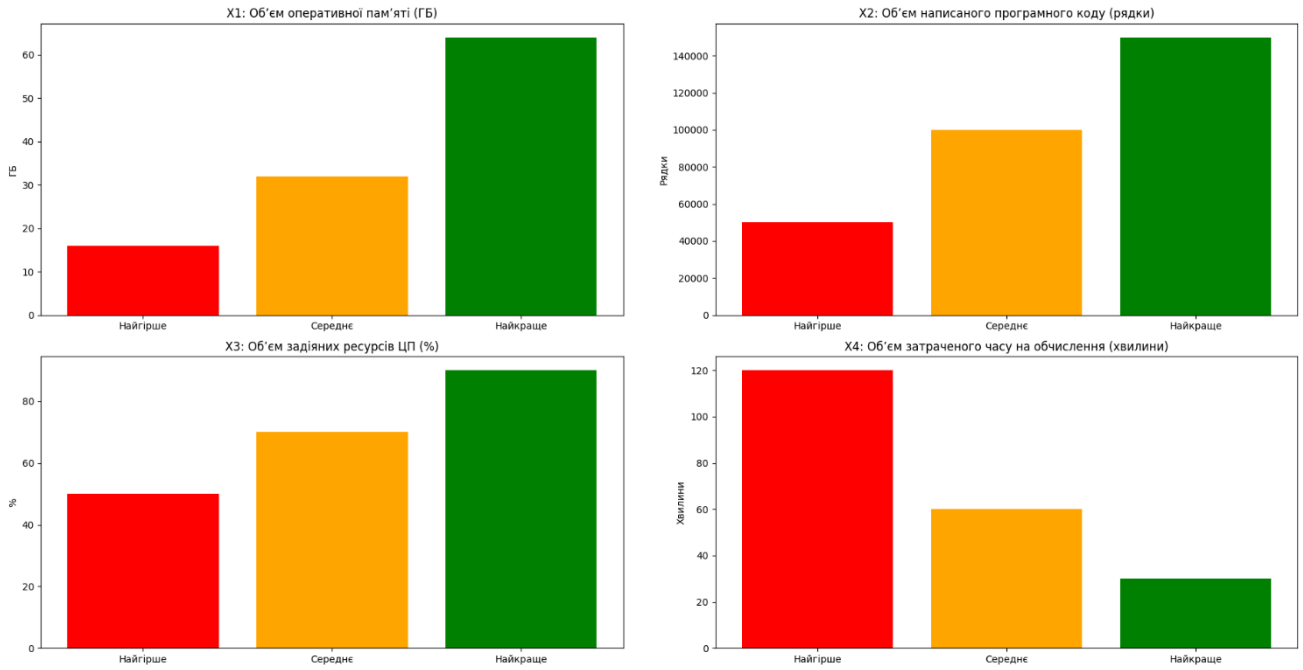

Рисунок 4.2 - Таблиця основних параметрів

4.4. Аналіз експертного оцінювання параметрів

Четвертий етап функціонально-вартісного аналізу програмного продукту включає в себе визначення вагомості кожного параметру в загальній кількості розглядуваних під час оцінювання параметрів за допомогою методу попарного порівняння.

Для цього спочатку було знайдено коефіцієнти вагомості, визначено ступінь важливості параметрів шляхом присвоєння їм різних рангів експертами-фахівцями в галузі. Результати продемонстровано в таблиці 4.3.

Таблиця 4.3 - Обрахунки

Позначення параметра	Назва параметра	Одиниці виміру	Ранг параметра за оцінкою експерта							Сума рангів R_i	Відхилення Δ_i	Δ_i^2
			1	2	3	4	5	6	7			
X1	Об'єм оперативної пам'яті для проведення обчислень	ГБ	2	2	1	1	1	2	2	11	-6,5	42,25
X2	об'єм написаного програмного коду	Рядки	1	1	2	2	1	1	1	9	-8,5	72,25
X3	об'єм задіяних ресурсів центрального процесору	Відсотки	3	3	3	3	4	4	3	23	5,5	30,25
X4	об'єм затраченого часу на проведення обчислень	Хвилини	4	4	4	4	3	3	4	26	8,5	72,25
	Сума		10	10	10	10	10	10	10	70	0	149

Для перевірки достовірності експертних оцінок буде визначено такі параметри:

1. Обчислено суму рангів кожного з параметрів та загальну суму рангів за формулою (4.1):

$$R_i = \sum_{j=1}^{N} r_{ij} R_{ij} = \frac{Nn(n+1)}{2} = 70, \quad (4.1)$$

де N - кількість експертів,

n - кількість параметрів.

2. Обчислено середню суму рангів за формулою (4.2):

$$T = \frac{1}{n} R_{ij} = 17{,}5 \tag{4.2}$$

3. Обчислимо відхилення суми рангів кожного параметра від середньої суми рангів за формулою (4.3):

$$\Delta_i = R_i - T \tag{4.3}$$

4. Обчислимо загальну суму квадратів відхилень за формулою (4.4):

$$S = \sum_{i=1}^{N} \Delta_i^2 = 217 \tag{4.4}$$

Порахуємо коефіцієнт узгодженості за формулою (4.5):

$$W = \frac{12S}{N^2(n^3-n)} = \frac{12 \cdot 217}{7^2(4^3-4)} = 0{,}698 > W_k = 0{,}886 \tag{4.5}$$

Знайдений коефіцієнт узгодженості перевищує нормативний значення 0.67, тому ранжування можна вважати достовірним. З використанням результатів ранжування проведемо попарне порівняння всіх параметрів, результати занесемо у таблицю 4.4.

Таблиця 4.4 - Попарне порівняння параметрів.

Параметри	Експерти							Кінцева оцінка	Числове значення
	1	2	3	4	5	6	7		
X1 і X2	>	>	<	<	=	>	>	>	1,5
X1 і X3	<	<	<	<	<	<	<	<	0,5
X1 і X4	<	<	<	<	<	<	<	<	0,5
X2 і X3	<	<	<	<	<	<	<	<	0,5
X2 і X4	<	<	<	<	<	<	<	<	0,5
X3 і X4	<	<	<	<	>	>	<	<	0,5

Для визначення ступеня переваги одного параметра над іншим, числове значення aij визначається згідно з формулою (4.6):

$$a_{ij} = \{1.5 \text{ при } X_i > X_j \; 1.0 \text{ при } X_i = X_j \; 0.5 \text{ при } X_i < X_j \; . \qquad (4.6)$$

З отриманих числових оцінок переваги складається матриця A=$\| a_{ij} \|$

Для кожного параметра розраховується вагомість Кві згідно з формулою:

$$K_{\text{ві}} = \frac{b_i}{\sum_{i=1}^{n} b_i} \qquad (4.7)$$

$$b_i = \sum_{i=1}^{N} a_{ij} \qquad (4.8)$$

Відносні оцінки розраховуються декілька разів, поки наступні значення не відрізняються від попередніх менше ніж на 2%. На наступних кроках відносні оцінки розраховуються за формулами:

$$K_{\text{ві}} = \frac{b_i'}{\sum_{i=1}^{n} b_i'}, \qquad (4.9)$$

$$b_i' = \sum_{i=1}^{N} a_{ij} b_j \qquad (4.10)$$

Таблиця 4.5 - Розрахунок вагомості параметрів

Параметри x_i	Параметри x_j				Перша ітер.		Друга ітер.		Третя ітер	
	X1	X2	X3	X4	b_i	$K_{\text{ві}}$	b_i^1	$K_{\text{ві}}^1$	b_i^2	$K_{\text{ві}}^2$
X1	1	1,5	0,5	0,5	3,5	0,2	12,25	0,15	42,875	0,19
X2	0,5	1	0,5	0,5	3	0,17	9	0,11	27	0,17
X3	1,5	1,5	1	0,5	5	0,29	25	0,31	125	0,30
X4	1,5	1,5	1,5	1	5,5	0,32	30,25	0,38	166,375	0,34
Всього:					17	1	79,5	1	361,25	1

З таблиці 4.5 видно, що різниця між значеннями коефіцієнтів вагомості не перевищує 2%, тому додаткові ітерації не потрібні.

4.5 Аналіз якості реалізації варіантів функцій

Для оцінки якості виконання основних функцій програмного продукту, ми використовуємо коефіцієнт технічного рівня (ККj). Цей коефіцієнт враховує вагомість кожного параметра і відображає рівень якості реалізації програмного продукту.

Для розрахунку ККj використовується формула, в якій параметри швидкості роботи мови програмування, об'єм пам'яті, час навчання та обсяг програмного коду множаться на вагові коефіцієнти. Ці вагові коефіцієнти визначаються експертами на основі їх оцінки важливості кожного параметра для програмного продукту. Формула виглядає таким чином:

$$K_K(j) = \sum_{i=1}^{n} K_{\text{в}i,j} B_{i,j}, \qquad (4.11),$$

де n – кількість параметрів;

$K_{\text{в}i}$– коефіцієнт вагомості i–го параметра;

B_i – оцінка i–го параметра в балах.

Такий підхід допомагає нам об'єктивно оцінити якість виконання функцій програмного продукту, враховуючи важливість кожного параметра. Згідно з формулою (4.11), кожен показник рівня якості обчислюється шляхом перемноження вагомості параметра (Кві) на його оцінку (Ві) у балах. Ці обчислення представлені в таблиці 4.6, яка містить оцінку якості різних варіантів реалізації основних функцій програмного продукту.

Таблиця 4.6 - Розрахунок показників якості варіантів реалізації основних функцій

Основні функції	Варіант реалізації функції	Параметри	Абсолютне значення параметра	Бальна оцінка параметра	Коефіцієнт вагомості параметра	Коефіцієнт рівня якості
F1	Б	X1	87	11	0,19	2,09
F2	А	X2	142	9	0,17	1,53
F3	А	X3	23	23	0.3	6,9
F3	Б	X4	24	26	0,34	8,84

$$K_K = K_{\text{ТУ}}[F_{1k}] + K_{\text{ТУ}}[F_{2k}] + \ldots + K_{\text{ТУ}}[F_{zk}], \qquad (4.12)$$

Рівень якості кожного з варіантів:

1. K_{Kl} = 2.09 + 1.53 +6.9 = 10,52 ;

2. K_{K2} = 2.09 + 1.53 +8.84 = 12,46.

Отже, з огляду на рівень якості, кращим варіантом є 2.

4.6 Економічний аналіз

Для оцінки вартості розробки програмного продукту спочатку проводиться розрахунок трудомісткості для кожного варіанту виконання основних завдань, що включають у себе розробку проекту програмного продукту і створення програмної оболонки. Функції різняться вибором СУБД для використання. Загальна трудомісткість обчислюється за допомогою формули (4.13).

$$\text{Т}_\text{О} = \text{Т}_\text{Р} \cdot \text{К}_\text{П} \cdot \text{К}_\text{СК} \cdot \text{К}_\text{М} \cdot \text{К}_\text{СТ} \cdot \text{К}_\text{СТ.М}, \qquad (4.13),$$

де ТР – трудомісткість розробки ПП;

КП – поправочний коефіцієнт;

КСК – коефіцієнт на складність вхідної інформації;

КМ – коефіцієнт рівня мови програмування;

КСТ – коефіцієнт використання стандартних модулів і прикладних програм;

КСТ.М – коефіцієнт стандартного математичного забезпечення

Для першої функції, що використовує більш потужний варіант СУБД було використано КП на рівні 1.6, КСК на рівні 1.25, КМ становив 1.1, КСТ 1, КСТ.М 1, f трудомісткість на рівні 20 людино-днів. Таким чином, загальна трудомісткість для першої функції становить:

$$T_1 = 20 \cdot 1.6 \cdot 1.25 \cdot 1.1 \cdot 1 \cdot 1 = 44 \text{ людино-днів}$$

Для другого варіанту було замінено КП на 1, а КСТ на 0.9.

$$T_2 = 20 \cdot 1.6 \cdot 1.25 \cdot 1 \cdot 0.9 \cdot 1 = 36 \text{ людино-днів}$$

Отже, більшу трудомісткість має перша функція.

При розробці даного ПП буде задіяно одного програміста, одного Data-Scientist'а та одного менеджера, що мають оклад 30 000, 40 000 та 24 000 відповідно.

Середня зарплата буде обрахована за такою формулою:

$$С_ч = \frac{М}{T_m \cdot t} \text{ грн.}, \quad (4.14)$$

де М – місячний оклад працівників;

Tm – кількість робочих днів місяць;

t – кількість робочих годин в день.

$$С_ч = \frac{30000 + 40000 + 24000}{3 \cdot 21 \cdot 8} = 186{,}5 \text{ грн.} \quad (4.15)$$

Далі буде обраховано середню зарплату за формулою (4.16)

$$СЗП = С_ч \cdot Т_i \cdot КД, \quad (4.16)$$

де $С_ч$ – величина погодинної оплати праці програміста;

$Т_i$ – трудомісткість відповідного завдання;

$К_Д$ – норматив, який враховує додаткову заробітну плату (20%).

Зарплата робітників дорівнює:

1. $C_{ЗП}$ = 186,5 · 44 · 24 = 196 944 грн.
2. $C_{ЗП}$ = 186,5 · 36 · 24 = 161 136 грн.
3.

Відрахування на соціальний внесок:

1. $C_{ВІД}$= $C_{ЗП}$ · 0.22 = 196 944 · 0,22 = 43 327,68 грн.
2. $C_{ВІД}$= $C_{ЗП}$ · 0.22 = 161 136 · 0,22 = 35 449,92 грн.

Далі буде обраховано ціну використання ЕОМ. ($C_М$)

Так як одна ЕОМ обслуговує двух програмістів з окладом 40000 + 30000 грн., з коефіцієнтом зайнятості 0,2 то для однієї машини отримаємо:

$C_Г$ = 12·М·М·$K_З$·2 = (12 · 40 000 · 0,2) + (12 · 0,2 · 30000) = 16 8000 грн.

З урахуванням додаткової заробітної плати:

$C_{ЗП}$ =$C_Г$· (1+ $K_З$) = 16 8000 · (1 + 0.2 + 0.2) = 235 200 грн

$C_{ВІД}$= $C_{ЗП}$ · 0.22 = 235 200 · 0,22 = 51744 грн.

Амортизаційні відрахування розраховуємо при амортизації 25% та вартості ЕОМ – 35 000 грн.

$C_А$ = $K_{ТМ}$· $K_А$·$Ц_{ПР}$ = 1.2 · 0.25 · 35 000 = 10500 грн.,

де $К_{ТМ}$– коефіцієнт, який враховує витрати на транспортування та монтаж приладу у користувача;

$К_А$– річна норма амортизації;

$Ц_{ПР}$– договірна ціна приладу.

Витрати на ремонт та профілактику розраховуємо як:

$$С_Р = К_{ТМ} \cdot Ц_{ПР} \cdot К_Р = 1.2 \cdot 35000 \cdot 0.08 = 3360 \text{ грн.},$$

де $К_Р$– відсоток витрат на поточні ремонти.

Ефективний годинний фонд часу ПК за рік розраховуємо за формулою:

$$Т_{ЕФ} = (Д_К – Д_В – Д_С – Д_Р) \cdot t_З \cdot К_В = (365 – 104 – 12 – 16) \cdot 8 \cdot 0.35 =$$

$$= 627,2 \text{ години},$$

де $Д_К$ – календарна кількість днів у році;

$Д_В$, $Д_С$ – відповідно кількість вихідних та святкових днів;

$Д_Р$ – кількість днів планових ремонтів устаткування;

t –кількість робочих годин в день;

$К_В$– коефіцієнт використання приладу у часі протягом зміни.

Витрати на оплату електроенергії розраховуємо за формулою:

$$С_{ЕЛ} = Т_{ЕФ} \cdot N_С \cdot К_З \cdot Ц_{ЕН} = 627,2 \cdot 0,3 \cdot 0,4 \cdot 5, 23 = 393,63072 \text{ грн.},$$

де $N_С$ – середньо-споживана потужність приладу;

$К_З$– коефіцієнтом зайнятості приладу;

$Ц_{ЕН}$ – тариф за 1 КВт-годин електроенергії.

Накладні витрати розраховуємо за формулою:

$$С_Н = Ц_{ПР} \cdot 0.67 = 35000 \cdot 0,67 = 23450 \text{ грн.}$$

Тоді, річні експлуатаційні витрати будуть:

$$С_{ЕКС} = С_А + С_Р + С_{ЕЛ} + С_Н, \quad (4.17)$$

$$С_{ЕКС} = 10500 + 3360 + 393,63072 + 23450 = 37\ 703,63072 \text{ грн.}$$

Собівартість однієї машино-години ЕОМ дорівнюватиме:

$$С_{М\text{-}Г} = С_{ЕКС} / Т_{ЕФ} = 37\ 703,63072 \ / \ 627,2 = 60.1 \text{ грн/год.}$$

Оскільки в даному випадку всі роботи, які пов‘язані з розробкою програмного продукту ведуться на ЕОМ, витрати на оплату машинного часу, в залежності від обраного варіанта реалізації, складає:

$$С_М = С_{М-Г} \cdot T, \quad (4.18)$$

1. $С_М = 60.1 \cdot 44 \cdot 12 = 31\ 732,8$ грн.
2. $С_М = 60.1 \cdot 36 \cdot 12 = 25\ 963,2$ грн.

Накладні витрати складають 67% від заробітної плати:

$$С_Н = С_{ЗП} \cdot 0,67, \quad (4.19)$$

1. $С_Н = 196\ 944 \cdot 0,67 = 131\ 952,48$ грн.
2. $С_Н = 161\ 136 \cdot 0,67 = 107\ 961,12$ грн.

Отже, вартість розробки ПП за варіантами становить:

$$C_{ПП} = C_{ЗП} + C_{ВІД} + C_{М} + C_{Н}, \quad (4.20)$$

1. $C_{ПП}$ = 196 944 + 43 327,68 + 31 732,8 + 131 952,48 = 403 956,96 грн.
2. $C_{ПП}$ = 161 136 + 35 449,92 + 25 963,2 + 107 961,12 = 330 510,24 грн.

4.7 Вибір кращого варіанту ПП техніко-економічного рівня

В даному розділі ми будемо проводити обчислення і вибір найкращого варіанту для розробки програмного продукту згідно з формулою (4.21).

$$К_{ТЕР j} = К_{К j} / С_{Ф j}, \quad (4.21)$$

$$К_{ТЕР1} = 10,52 / 403\,956,96 = 2,6042 \cdot 10^{-5},$$

$$К_{ТЕР2} = 12,46 / 330\,510,24 = 3,7699 \cdot 10^{-5}.$$

Отже, як показують з обчислення вище, другий варіанти розробки програмного продукту є найкращим. Після проведення функціонально-вартісного аналізу ми прийшли до висновку, що серед залишених двох варіантів для виконання роботи перший варіант є ефективнішим. Отже вибір пав на такі компоненти вирішення задачі:

1. F1 (вибір методу збору даних) - Метод стратифікованої вибірки
2. F2 (вибір бібліотеки для обробки великих даних) - Apache Hadoop
3. F3 (вибір бібліотеки для візуалізації результатів) - Plotly
4. F4 (вибір способу зберігання та керування даними) - MongoDB

4.8. Висновки до розділу 4

У даному розділі було проведено повний функціонально-вартісний аналіз програмного продукту з метою визначення та оцінки його основних функцій. Результати аналізу дозволили виявити параметри, що характеризують програмний продукт. Крім того, на основі проведеного аналізу було оцінено оптимальні варіанти реалізації програмного продукту. Цей етап забезпечує необхідну інформацію для подальшої розробки та визначення вартості програмного комплексу.

ВИСНОВКИ

У цій роботі було досліджено та проаналізовано методологію інтеграції великих даних у вибіркові обстеження з метою підвищення точності статистичних висновків. У першому розділі було введено основні поняття та значення великих даних, розглянуто історію їх виникнення та еволюцію, а також обговорено переваги і виклики, пов'язані з використанням великих даних у різних сферах, наведено поняття вибіркових обстежень, їх історичний розвиток та різні типи. Було розглянуто методологію використання великих даних у вибіркових обстеженнях, включаючи приклади методів збору, обробки та аналізу великих даних. Було також надано приклади використання великих даних у практиці та обговорено виклики та проблеми, що виникають при інтеграції великих даних у вибіркові обстеження. Також було розглянуто майбутнє використання великих даних у дослідженнях опитувань, що підкреслює їх зростаючу важливість і потенціал.

Другий розділ був присвячений розробці методології використання великих даних у вибіркових обстеженнях. Було проведено огляд існуючих інструментів та платформ для роботи з великими даними, розглянуто процес стратифікації та оцінювання шляхом інтеграції великих даних. Було представлено та використано модифіковану регресійну оцінку для підвищення точності оцінок параметрів популяції. Результати показали, що використання таких методів дозволяє коригувати можливі систематичні похибки у великих даних і забезпечує точніші оцінки.

У третьому розділі було проведено тестування та аналіз результатів використання великих даних у вибіркових обстеженнях. Було обрано інструменти та методи тестування, описано процедуру проведення тестування та проведено аналіз отриманих результатів. Тестування показало, що метод виправлення зміщення значно покращує точність оцінок. Результати тестування продемонстрували, що запропоновані методи ефективно

інтегрують великі дані з вибірковими обстеженнями, що дозволяє отримувати надійні статистичні висновки.

Четвертий розділ був присвячений проведенню повного функціонально-вартісного аналізу програмного продукту з метою визначення та оцінки його основних функцій.

Загалом, проведене дослідження підтверджує, що інтеграція великих даних з вибірковими обстеженнями є перспективним підходом для підвищення точності статистичних оцінок. Використання методів калібрування ваг, регресійного оцінювання та непараметричної класифікації дозволяє ефективно використовувати великі дані для покращення надійності статистичних висновків. Це дослідження робить вагомий внесок у розвиток методологій роботи з великими даними та їх інтеграції у вибіркові обстеження, що має значний потенціал для застосування у різних галузях науки і практики.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Eric R. Kushins, Elaina Behounek. Using sociological theory to problematize family business research. Journal of Family Business Strategy, Volume 11, Issue 1, 2020, 100337. URL: https://www.sciencedirect.com/science/article/abs/pii/S1877858518301840 (date of access: 18.03.2024).
2. Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, Amit Dhurandhar. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. ArXiv. URL: https://arxiv.org/abs/2206.10847 (date of access: 18.03.2024).
3. Jianzheng Liu, Jie Li, Weifeng Li, Jiansheng Wu. Rethinking big data: A review on the data quality and usage issues. ISPRS Journal of Photogrammetry and Remote Sensing, Volume 115, 2016, P. 134-142. URL: https://www.sciencedirect.com/science/article/abs/pii/S0924271615002567 (date of access: 18.03.2024).
4. Marc-André Kaufhold, Christian Reuter, Thomas Ludwig. Cross-Media Usage of Social Big Data for Emergency Services and Volunteer Communities: Approaches, Development and Challenges of Multi-Platform Social Media Services. ArXiv. URL: https://arxiv.org/abs/1907.07725 (date of access: 18.03.2024).
5. Paul A. Lynch. Sociological impressionism in a hospitality context. Annals of Tourism Research, Volume 32, Issue 3, 2005, P. 527-548. URL: https://www.sciencedirect.com/science/article/pii/S0160738305000654 (date of access: 18.03.2024).
6. Adam Hayes. Acceptance Sampling: Meaning, Types, and FAQ. Investopedia. April 30, 2023. URL: https://www.investopedia.com/terms/a/acceptance-sampling.asp#toc-a-history-of-acceptance-sampling (date of access: 18.03.2024)

7. Daniel D. Kho, Seungmin Lee, Ray Y. Zhong. Big Data Analytics for Processing Time Analysis in an IoT-enabled manufacturing Shop Floor. Procedia Manufacturing, Volume 26, 2018, P. 1411-1420. URL: https://www.sciencedirect.com/science/article/pii/S2351978918307789 (date of access: 18.03.2024).
8. Jianqing Fan, Fang Han, Han Liu. Challenges of Big Data Analysis. ArXiv. URL: https://arxiv.org/abs/1308.1479 (date of access: 18.03.2024).
9. Michael Janssen, Jack F. Radcliffe, Jan Wagner. Software and techniques for VLBI data processing and analysis. ArXiv. URL: https://arxiv.org/abs/2209.06115 (date of access: 18.03.2024).
10. Andreas P. Plageras, Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, B.B. Gupta. Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings. Future Generation Computer Systems, Volume 82, 2018, P. 349-357. URL: https://www.sciencedirect.com/science/article/pii/S0167739X17314127 (date of access: 18.03.2024).
11. Adil Rajput. Natural Language Processing, Sentiment Analysis and Clinical Analytics. ArXiv. URL: https://arxiv.org/abs/1902.00679 (date of access: 18.03.2024).
12. Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, Ge Yu. Collecting and Analyzing Multidimensional Data with Local Differential Privacy. ArXiv. URL: https://arxiv.org/abs/1907.00782 (date of access: 18.03.2024).
13. L.U. Laboshin, A.A. Lukashin, V.S. Zaborovsky. The Big Data Approach to Collecting and Analyzing Traffic Data in Large Scale Networks. Procedia Computer Science, Volume 103, 2017, P. 536-542. URL: https://www.sciencedirect.com/science/article/pii/S1877050917300492 (date of access: 18.03.2024).
14. Héber H. Arcolezi, Sébastien Gambs, Jean-François Couchot, Catuscia Palamidessi. On the Risks of Collecting Multidimensional Data Under Local

Differential Privacy. ArXiv. URL: https://arxiv.org/abs/2209.01684 (date of access: 18.03.2024).

15. Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin. Collecting Telemetry Data Privately. ArXiv. URL: https://arxiv.org/abs/1712.01524 (date of access: 18.03.2024).
16. Kevin Taylor-Sakyi. Big Data: Understanding Big Data. ArXiv. URL: https://arxiv.org/abs/1601.04602 (date of access: 18.03.2024).
17. Jae-Kwang Kim, Siu-Ming Tam. Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference. 26 Mar 2020. URL: https://arxiv.org/abs/2003.12156
18. Василик О. І., Яковенко Т. О. Лекції з теорії і методів вибіркових обстежень : навчальний посібник. К. : Видавничополіграфічний центр "Київський університет", 2010. 208 с. ISBN 978-966-439-307-9
19. Великі перспективи індустрії Big Data. Український суперкомп'ютерний інтернет-дайджест. 19 лютого 2013. URL: https://web.archive.org/web/20161024023126/http://supercomputer.com.ua/ua/266-veliki-perspektivi-industriji-big-data.html

ДОДАТОК А

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

# Приклад даних (симуляція)
np.random.seed(42)
N = 10000  # Розмір популяції
n = 1000   # Розмір вибірки
data = pd.DataFrame({
    'id': np.arange(N),
    'x': np.random.normal(size=N),
    'y': np.random.normal(size=N)
})

# Симуляція великих даних з випадковими похибками (шумом)
big_data = data.sample(frac=0.5, replace=False)
big_data['y'] += np.random.normal(scale=0.5, size=len(big_data))

# Випадкова вибірка (Probability Sample)
prob_sample = data.sample(n=n, replace=False)

# Додавання індикатора для великих даних
prob_sample['in_big_data'] =
prob_sample['id'].isin(big_data['id']).astype(int)

# Калібрація ваг
def calibration_weights(prob_sample, big_data):
    X = prob_sample[['x']]
    y = prob_sample['y']

    # Регресія для оцінки параметрів
    reg = LinearRegression().fit(X, y)
    y_pred = reg.predict(big_data[['x']])

    # Оцінка ваг калібрації
    weights = (big_data['y'] - y_pred).mean() / (big_data['y'] - y_pred)
    return weights

weights = calibration_weights(prob_sample, big_data)

# Оцінка інтеграції даних
```

```
def data_integration_estimator(prob_sample, big_data, weights):
    Tb = big_data['y'].sum()
    Nc = len(prob_sample[prob_sample['in_big_data'] == 0])
    Tc = (prob_sample.loc[prob_sample['in_big_data'] == 0, 'y'] * weights).sum()

    T_hat = Tb + Nc * Tc / len(prob_sample)
    return T_hat

T_hat = data_integration_estimator(prob_sample, big_data, weights)
print("Оцінка інтеграції даних:", T_hat)

# Непараметрична класифікація
from sklearn.ensemble import RandomForestClassifier

def nonparametric_classification(prob_sample, big_data):
    X = prob_sample[['x']]
    y = prob_sample['in_big_data']

    clf = RandomForestClassifier().fit(X, y)
    prob_sample['pred_in_big_data'] = clf.predict(X)

    return clf

clf = nonparametric_classification(prob_sample, big_data)

# Оцінка з виправленням зміщення
def bias_corrected_estimator(prob_sample, big_data, clf):
    prob_sample['pred_in_big_data'] = clf.predict(prob_sample[['x']])

    Tb = big_data['y'].sum()
    Nc = len(prob_sample[prob_sample['pred_in_big_data'] == 0])
    Tc = prob_sample.loc[prob_sample['pred_in_big_data'] == 0, 'y'].mean()

    T_hat_bc = Tb + Nc * Tc
    return T_hat_bc

T_hat_bc = bias_corrected_estimator(prob_sample, big_data, clf)
print("Оцінка з виправленням зміщення:", T_hat_bc)
```

ДОДАТОК Б

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
from wordcloud import WordCloud

# Завантаження даних
df = pd.read_csv('/kaggle/input/ecommerce-data/data.csv',
encoding='latin1')

# Попередня обробка даних
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
df = df.dropna(subset=['CustomerID', 'Quantity', 'UnitPrice'])  # Видаляємо
записи без CustomerID, Quantity, UnitPrice

# Додавання нового стовпця для загальної вартості
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']

# Створення великої вибірки (Big Data) та випадкової вибірки
(Probability Sample)
np.random.seed(42)
big_data = df.sample(frac=0.5, replace=False)
non_big_data = df[~df['CustomerID'].isin(big_data['CustomerID'])]

# Переконаємося, що випадкова вибірка не більша, ніж кількість
доступних записів
sample_size = min(len(non_big_data), 500)
prob_sample = non_big_data.sample(n=sample_size, replace=False)
prob_sample = pd.concat([prob_sample, df.sample(n=500, replace=False)],
ignore_index=True)

# Додавання індикатора для великих даних
prob_sample['in_big_data'] =
prob_sample['CustomerID'].isin(big_data['CustomerID']).astype(int)
```

```
# Калібрація ваг
def calibration_weights(prob_sample, big_data):
    X = prob_sample[['Quantity', 'UnitPrice']]
    y = prob_sample['TotalPrice']

    # Регресія для оцінки параметрів
    reg = LinearRegression().fit(X, y)
    y_pred = reg.predict(big_data[['Quantity', 'UnitPrice']])

    # Оцінка ваг калібрації
    weights = (big_data['TotalPrice'] - y_pred).mean() /
(big_data['TotalPrice'] - y_pred + 1e-9)  # Додаємо мале значення для уникнення
ділення на нуль
    return weights

weights = calibration_weights(prob_sample, big_data)

# Оцінка інтеграції даних
def data_integration_estimator(prob_sample, big_data, weights):
    Tb = big_data['TotalPrice'].sum()
    Nc = len(prob_sample[prob_sample['in_big_data'] == 0])
    Tc = (prob_sample.loc[prob_sample['in_big_data'] == 0, 'TotalPrice'] *
weights).sum()

    T_hat = Tb + Nc * Tc / len(prob_sample)
    return T_hat

T_hat = data_integration_estimator(prob_sample, big_data, weights)

# Функція для непараметричної класифікації з різними моделями
def nonparametric_classification(prob_sample, classifier):
    X = prob_sample[['Quantity', 'UnitPrice']]
    y = prob_sample['in_big_data']

    clf = classifier.fit(X, y)
    prob_sample['pred_in_big_data'] = clf.predict(X)

    return clf

# Список класифікаторів
classifiers = {
    'Random Forest': RandomForestClassifier(),
    'Decision Tree': DecisionTreeClassifier(),
```

```
        'SVM': SVC(),
        'Neural Network': MLPClassifier(),
        'Naive Bayes': GaussianNB(),
        'Logistic Regression': LogisticRegression()
    }

    # Training and evaluating accuracy for each classifier
    results = []
    for name, clf in classifiers.items():
        model = nonparametric_classification(prob_sample, clf)
        accuracy = accuracy_score(prob_sample['in_big_data'],
prob_sample['pred_in_big_data'])
        T_hat_bc = data_integration_estimator(prob_sample, big_data, weights)
# Using existing weights for simplicity
        results.append({
            'Model': name,
            'Accuracy': accuracy,
            'Bias Corrected Estimator': T_hat_bc
        })

    results_df = pd.DataFrame(results)
    print(results_df)

    # Додатковий аналіз
    print("\nДодатковий аналіз результатів:")
    print("Кількість записів у великих даних:", len(big_data))
    print("Кількість записів у вибірковій вибірці:", len(prob_sample))
    print("Розподіл in_big_data у вибірковій вибірці:")
    print(prob_sample['in_big_data'].value_counts())
    # Перевірка точності класифікації
    accuracy = (prob_sample['in_big_data'] ==
prob_sample['pred_in_big_data']).mean()
    print("Точність класифікації:", accuracy)

    # Visualizations
    plt.figure(figsize=(10, 6))
    sns.histplot(df['TotalPrice'], bins=50, kde=True)
    plt.title('Розподіл TotalPrice')
    plt.xlabel('TotalPrice')
    plt.ylabel('Частота')
    plt.show()

    plt.figure(figsize=(10, 6))
    sns.scatterplot(x=df['Quantity'], y=df['TotalPrice'])
```

```
plt.title('Залежність між Quantity та TotalPrice')
plt.xlabel('Quantity')
plt.ylabel('TotalPrice')
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(x='in_big_data', y='TotalPrice', data=prob_sample)
plt.title('Розподіл TotalPrice для записів у великих даних і вибірковій вибірці')
plt.xlabel('in_big_data')
plt.ylabel('TotalPrice')
plt.show()

results_df

# Distribution plots for Quantity, UnitPrice, TotalPrice
plt.figure(figsize=(18, 6))

plt.subplot(1, 3, 1)
sns.histplot(df['Quantity'], bins=50, kde=True)
plt.title('Розподіл Quantity')

plt.subplot(1, 3, 2)
sns.histplot(df['UnitPrice'], bins=50, kde=True)
plt.title('Розподіл UnitPrice')

plt.subplot(1, 3, 3)
sns.histplot(df['TotalPrice'], bins=50, kde=True)
plt.title('Розподіл TotalPrice')

plt.show()

# Horizontal countplot for Country sorted from largest to smallest
plt.figure(figsize=(12, 6))
sns.countplot(y=df['Country'], order=df['Country'].value_counts().index)
plt.title('Розподіл кількості покупок за країнами (від більшого до меншого)')
plt.xlabel('Count')
plt.ylabel('Country')
plt.show()

# Word cloud for each country
countries = df['Country'].unique()
```

```
for country in countries:
    country_text = ' '.join(df[df['Country'] == country]['Description'].dropna().astype(str))
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate(country_text)

    plt.figure(figsize=(10, 6))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(f'Word Cloud for {country}')
    plt.show()
```

ДОДАТОК В ГРАФІЧНІ МАТЕРІАЛИ

Використання “великих даних” у вибіркових дослідженнях

ПІБ, група:
Дідіков Олександр Олександрович, група КА-04

Керівник:
Доцент д.ф.-м.н. Василик О.І.

Актуальність

Актуальність дослідження полягає у важливості інтеграції великих даних у вибіркові дослідження для підвищення точності та ефективності статистичних оцінок. Великі дані відкривають нові можливості для аналізу великих масивів інформації, що сприяє більш глибокому розумінню соціальних, економічних та інших явищ.

Об’єкт, предмет і мета дослідження

Об’єкт дослідження:

Застосування методів обробки та аналізу великих даних у вибіркових обстеженнях.

Предмет дослідження:

Методи і засоби інтеграції великих даних у вибіркових обстеженнях.

Мета дослідження:

Розробка методології та програмного забезпечення для покращення вибіркових обстежень шляхом інтеграції великих даних.

Постановка задачі

Дослідження концепції великих даних та їх значення у вибіркових обстеженнях.

Аналіз переваг та викликів використання великих даних.

Розробка методології інтеграції великих даних у вибіркові обстеження.

Тестування та оцінка розробленої методології на реальних даних.

Основні результати дипломної роботи

Результат 1: Розробка методології інтеграції великих даних у вибіркові обстеження

Опис: Розроблена методологія базується на сучасних підходах до обробки великих даних, таких як машинне навчання, статистичний аналіз та алгоритми калібрування. Методологія включає в себе наступні етапи:

- Ідентифікація релевантних джерел великих даних
- Попередня обробка та очищення даних для забезпечення їх надійності та повноти
- Інтеграція великих даних з вибірковими обстеженнями шляхом застосування методів калібрування та регресійного аналізу.
- Використання непараметричних методів класифікації для ідентифікації та виправлення помилок у великих даних

Переваги:

- Збільшення точності статистичних оцінок
- Зменшення зміщення результатів обстежень
- Підвищення надійності отриманих даних

Основні результати дипломної роботи

Результат 2: Створення програмного забезпечення для реалізації методології

Опис: Розроблено програмне забезпечення, яке автоматизує процес інтеграції великих даних з вибірковими обстеженнями. Основні функції програмного забезпечення включають:

- Збір та зберігання великих даних з різних джерел(соціальні мережі, інтернет, сенсори тощо).
- Очищення та попередня обробка даних для видалення дублікатів, пропущених значень та аномалій
- Застосування алгоритмів калібрування ваг та регресійного аналізу для інтеграції великих даних
- Генерація звітів та візуалізація результатів для полегшення інтерпретації даних

Переваги:

- Автоматизація складних процесів обробки даних
- Скорочення часу на проведення обстежень
- Забезпечення високої точності та надійності даних

Основні результати дипломної роботи

Результат 3: Експериментальна перевірка методології на реальних даних

Опис: Проведено експериментальну перевірку розробленої методології на реальних даних з платформи Kaggle. Для цього було виконано:

- Збір великого набору даних з різних джерел.
- Проведення вибіркових обстежень та інтеграція отриманих результатів з великими даними.
- Аналіз точності та надійності отриманих результатів шляхом порівняння з традиційними методами.

Результати експерименту:

- Підтверджено, що розроблена методологія дозволяє значно підвищити точність статистичних оцінок.
- Експериментальні результати показали, що інтеграція великих даних дозволяє зменшити зміщення та підвищити надійність отримани х даних.
- Методологія виявилась ефективною для різних типів даних та умов проведення обстежень.

Переваги:

- Практична перевірка методології на реальних даних.
- Підтвердження ефективності та надійності розроблених методів.
- Можливість адаптації методології до різних сфер досліджень.

Наукова новизна

Запропонована методологія та програмне забезпечення дозволяють інтегрувати великі дані у вибіркові обстеження, що значно підвищує точність та надійність статистичних оцінок. Використання сучасних технологій аналізу даних та машинного навчання забезпечує інноваційний підхід до вирішення проблеми.

Висновки

Використання великих даних у вибіркових обстеженнях дозволяє підвищити точність та ефективність обстежень.

Розроблена методологія та програмне забезпечення довели свою ефективність на практиці.

Інтеграція великих даних відкриває нові можливості для дослідників у різних галузях.

Подальші дослідження

Розширення функціональних можливостей програмного забезпечення.

Застосування методології на інших наборах даних та у різних сферах досліджень.

Вдосконалення методів обробки та аналізу великих даних.