**ВСТУП**

Вивчення інформаційних потоків та їхньої якості набуває великого значення для бізнесу, науки та суспільства. Розвиток нових технологій та збільшення обсягів даних вимагає вдосконалення методів їх обробки. Одна з найважливіших проблем в аналізі даних є робота з відсутніми значеннями та невизначеністю, особливо у великих наборах даних.

Ця проблема є важливою у багатьох сферах, таких як: бізнес, наукові дослідження, фінанси, медицина.

**Актуальність дослідження роботи**

1. **Збільшення обсягів даних:** у світі зростає обсяг інформації дуже швидко, і щоб отримати точні і надійні результати необхідно розробляти ефективні методи обробки даних.
2. **Підвищення вимог до якості даних:** точність даних має важливе значення для наукових досліджень. Щоб отримати якісні дані, необхідно усунути невизначеності та пропущені значення.
3. **Інтеграція даних з різних джерел:** організації використовують дані з різних джерел і форматів. Необхідно враховувати різні підходи до обробки пропущених значень та невизначеності для ефективної обробки даних.
4. **Необхідність прийняття обґрунтованих рішень:** важливо приймати обґрунтовані рішення, які базуються на точних і надійних даних.
5. **Розвиток штучного інтелекту та машинного навчання:** через швидкий розвиток штучного інтелекту та машинного навчання, вивчення підходів до обробки даних стає важливою передумовою для розробки ефективних алгоритмів та моделей.

Отже, вивчення та вдосконалення способів обробки даних, особливо в роботі з відсутніми даними та невизначеністю, має велике значення в сучасному світі. Для зростаючих обсягів інформації підвищені вимоги до якості даних, а також потребу в об'єднанні даних з різних місць.

Розроблення ефективних методів обробки даних не тільки забезпечує точність і надійність даних, але й сприяє прийняттю впевнених рішень у різних галузях і загальному прогресу суспільства та бізнесу.

**Метою даної курсової роботи** є аналіз і порівняння різних сучасних підходів до обробки пропущених значень та визначення невизначеності у великих наборах даних.

Щоб досягти цієї мети, поділимо завдання на такі основні завдання:

1. Аналіз та порівняння сучасних існуючих методів для обробки пропущених значень.
2. Розробка і вдосконалення алгоритмів, які забезпечують велику швидкість і точність для обробки пропущених значень та визначення невизначеності у великих наборах даних.
3. Тестування та оптимізація розробленої програми.
4. Порівняння та аналізування отриманих результатів з існуючими методами. Виявлення можливостей застосування отриманих результатів у реальних умовах.

**Об’єктом дослідження** є порівняння іпокращення методів, які використовуються для обробки пропущених значень, визначення їх ефективності і можливостей у застосуванні у практичних ситуаціях.

**Предметом дослідження** є методи та алгоритми, які використовуються для обробки пропущених значень.

Використовуються такі **методи дослідження**:

* **Аналіз літературних джерел:** перегляд уже наявних наукових робіт та досліджень, які розглядають методи обробки пропущених даних та визначення невизначеності у великих обсягах інформації.
* **Емпіричні дослідження:** використання фактичних даних для порівняння різних методів обробки пропущених даних та оцінки їх ефективності в реальних умовах.
* **Статистичний аналіз:** використання статистичних методів для порівняння результатів різних методів обробки пропущених значень та визначення невизначеності великих наборів даних.
* **Моделювання:** створення математичних моделей для вивчення впливу різних методів обробки даних на точність та достовірність результатів.

**Практичне значення досліджен**ня різних підходів до обробки пропущених значень та визначення невизначеності у великих наборах даних полягає в наступному:

* **Покращення якості даних:** розробка ефективних методів роботи з відсутніми даними та невизначеністю може допомогти підвищити якість даних, що використовуються в наукових дослідженнях, бізнес-аналітиці, фінансовому аналізі та інших сферах.
* **Зниження ризику помилок:** використання найефективніших методів дозволить уникнути помилок у аналізі даних і ухилення від неправильних рішень на основі неповних або неточних даних.
* **Підвищення ефективності аналізу:** оптимальний вибір методів обробки даних підвищить ефективність аналітичних процесів, зменшить час, необхідний для обробки даних, та покращить точність результатів.
* **Підтримка прийняття раціональних рішень:** надійні та точні дані є ключовими для прийняття раціональних та обґрунтованих управлінських, наукових та стратегічних рішень, що сприяє успішному функціонуванню організацій та проектів.
* **Застосування у реальних умовах:** результати дослідження можуть бути використані у практичних задачах обробки даних у різних сферах діяльності, що сприятиме підвищенню ефективності та конкурентоспроможності.

**Апробація з дослідження** з обробки відсутніх значень та визначення невизначеності, у великих наборах даних містить: тестування розробленого методу на реальних даних.

Робота буде складатися з наступних розділів:

● Вступ

● Загальний розділ

● Аналіз матеріалів та методів

● Експерименти

● Результати експериментів дослідження

● Висновки

1. **ЗАГАЛЬНИЙ РОЗДІЛ**

У розділі буде проаналізовано наступні літературні джерела: аналіз наукових статей, що стосуються теми курсової роботи.

* 1. **Аналіз літературних джерел**

Список літературних джерел, які використовувались в роботі.

1. *Little, Roderick JA, and Donald B. Rubin. "Statistical analysis with missing data." John Wiley & Sons, 2019.* – фокусується на проблемі відсутніх даних та способах її вирішення, а також детально описує методи статистичного аналізу.
2. *Gelman, Andrew, et al. "Bayesian data analysis." CRC Press, 2021*. –розглядається байєсівський підхід до аналізу даних, який є корисним для визначення невизначеності в даних.
3. *Rubin, Donald B. "Multiple imputation for nonresponse in surveys." John Wiley & Sons, 2004*. – присвячена методам множинних підстановок, які зазвичай використовуються в соціології та інших дослідженнях для роботи з пропущеними даними. Книга містить практичні поради та методологічні підходи для роботи з пропущеними значеннями.
4. *Van Buuren, Stef. "Flexible imputation of missing data." CRC Press, 2012*. – описує гнучкі методи відновлення відсутніх даних, які дозволяють ефективно уникнути втрати інформації під час аналізу. Це важливе джерело інформації для дослідників, зацікавлених у розробці методів обробки неповних даних.
5. *Carpenter, James R., and Michael G. Kenward. "Missing data in clinical trials: a practical guide." CRC Press, 2020*. – фокусується на аналізі пропущених даних у клінічних дослідженнях і надає практичні поради щодо їх обробки. Це джерело є корисним для фахівців у медичній галузі, які стикаються з проблемою відсутніх даних у дослідженнях.
6. *Enders, Craig K. "Applied missing data analysis." Guilford Press, 2010.* – зосереджується на аналізі відсутніх даних у клінічних випробуваннях і надає практичні поради, як працювати з відсутніми даними.
7. *Graham, J. W. (2012). Missing data: Analysis and design. Springer.* – містить огляд того, як аналізувати та планувати дослідження з пропущеними даними. Він допомагає зрозуміти принципи роботи з відсутніми даними та їхній вплив на аналіз результатів дослідження.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Літератур-не джерело № | Методи | Тип завдань | Переваги | Недоліки |
| 1 | Множинні підстановки | Соціологічні дослідження, аналіз опитувань | Ефективне врахування пропущених даних | Вимагається вибір правильного моделювання |
| 2 | Байєсівський підхід | Дослідження з великою кількістю параметрів | Враховує невизначе-ність і враховує апріорні знання | Вимагає досить глибоких знань у байєсівській статистиці |
| 3 | Множинні підстановки | Соціологічні дослідження, аналіз опитувань | Практичні поради і методологічні підходи | Необхідний великий набір даних |
| 4 | Гнучкі методи відновлення | Аналіз даних з різноманітними типами пропущених даних | Ефективно уникає втрати інформації під час аналізу | Є обчислюва-льно складними і зазвичай вимагають більшої обробки даних |
| 5 | Аналіз пропущених даних у клінічних дослідженнях | Клінічні дослідження | Практичні поради щодо обробки пропущених даних | Обмежена область застосування |
| 6 | Аналіз пропущених даних у клінічних випробуваннях | Клінічні дослідження | Надає практичні поради для роботи з відсутніми даними | Обмежена область застосування |
| 7 | Аналіз та планування досліджень з пропущеними даними | Дослідження з пропущеними даними | Допомагає зрозуміти принципи роботи з відсутніми даними | Загальне порівняння методів |

У машинному навчанні, існують сучасні, різні методи, які обробляють відсутні дані. Ця література надає конкретні поради та методологічні підходи для кращого розуміння та вирішення проблеми відсутніх даних і невизначеності. Для науковців, аналітиків даних, які працюють з великими обсягами даних і мають справу з відсутніми значеннями та проблемами невизначеності є важливо знати: Байєсівський аналіз, методи множинного відновлення, гнучкі методи підстановки та інші підходи, описані в цих джерелах.

* 1. **Постанова задачі**

*Задача:* вивчення різних підходів до обробки пропущених значень та визначення невизначеності у великих наборах даних.

Для досягнення цієї цілі, потрібно виконати наступні етапи:

* **Постановка завдання:** на цьому етапі визначається основна мета дослідження, яка полягає у вивченні різних підходів до обробки пропущених значень та невизначеності у великих масивах даних. На цьому етапі також визначаються вимоги до якості даних і вибираються параметри для оцінки ефективності методів обробки.
* **Аналіз існуючих методів:** проводиться огляд та аналіз існуючих методів обробки пропущених значень та невизначеності, таких як методи імплікації, статистичні підходи та машинне навчання. Це дозволить визначити найбільш підходящий метод для поставленого завдання.
* **Підготовка та підготовка даних:** на цьому етапі збираються дані, необхідні для дослідження, в тому числі великі масиви даних, що містять пропущені значення. Дані проходять попередню обробку, включаючи видалення шуму і пропусків, для підготовки до подальшого аналізу.
* **Вибір і розробка методів обробки:** на основі аналізу наявних методів і характеристик даних обирається конкретний метод обробки, який найкраще відповідає цілям дослідження. Розробляються алгоритми для реалізації обраних методів обробки.
* **Реалізація та тестування методу:** на цьому етапі реалізується обраний метод обробки даних та проводяться тестові експерименти для оцінки його ефективності та точності. Будь-які виявлені недоліки оптимізуються для покращення результатів.
* **Оцінка ефективності та валідація:** ефективність і точність обраного методу оцінюється за допомогою таких показників, як точність, повнота, F-міра і коефіцієнт Жаккара. Результати тестуються на нових даних для перевірки стабільності та відтворюваності підходу.