1. **Вступ**
2. **Актуальність**
3. **Мета**
4. **Завдання дослідження**
5. **Об’єкт та предмет дослідження**
6. **Опис датасету**
7. **Опис методів машинного навчання**
8. **Опис методів обробки пропущених значень**
9. **Тренувальна та тестувальна вибірка**
10. **Порівняння результатів**
11. **Висновок**

**Вступ**

Пропущені значення та невизначеність є важливими аспектами аналізу даних, особливо у великих наборах даних. Вони можуть суттєво впливати на результати аналізу та моделювання, тому ефективна обробка цих проблем є критичною для отримання точних і надійних висновків.

**Актуальність дослідження**

Актуальність теми обумовлена через зростання обсягів даних, що збираються в різних сферах діяльності, таких як медицина, фінанси, соціальні мережі та інші. Обробка пропущених значень стає критичними завданнями для забезпечення точності аналізу даних і побудови надійних моделей машинного навчання. У майбутньому ця тема залишатиметься актуальною через подальше зростання обсягів даних і ускладнення методів їх аналізу.

Чому це використовують, на це є кілька причин, ефективна обробка пропущених значень дозволяє:

* Підвищити точність моделей машинного навчання.
* Знизити ризики прийняття неправильних рішень на основі неточних даних.
* Забезпечити надійність результатів аналізу.

**Мета**

Метою цієї роботи є визначення та порівняння різних підходів до обробки пропущених значень і визначення невизначеності у великих наборах даних, а також оцінка ефективності цих підходів у машинному навчанні.

**Завдання дослідження**

1. Вивчити існуючі методи обробки пропущених значень: провести детальний огляд сучасних методів, що використовуються для обробки пропущених значень у великих наборах даних.
2. Оцінити ефективність методів: визначити, які методи обробки пропущених значень є найбільш ефективними в різних контекстах і наборах даних.
3. Виявити моделі машинного навчання, що найкраще працюють з великими наборами даних, тобто порівняти ефективність різних моделей машинного навчання.
4. Розробити рекомендації на основі проведеного аналізу для роботи з великими обсягами даних.

**Об’єкт та предмет дослідження**

**Об’єктом дослідження** є порівняння і покращення 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.

**Опис датасету**

Було розглянуто 3 набори даних з медицини, а саме з: інфарктом, ковідом і діабетом. Усі 3 набори даних містять різну кількість рядків і стовпців. У датасеті з інфарктом 1622 людини, у ковіді 1048575, а в діабеті 70692 людини. У кожному з цих датасетів є різна кількість стовпців, а саме 14, 21 і 21. Дані датасети є збалансованими. Вони містять інформацію про стан здоров’я людини, і які фактори можуть впливати на її діагноз.

**Методи машинного навчання**

Для аналізу обробки пропущених значень з цими наборами даних було вибрано 5 моделей машинного навчання: КНН, логістична регресія, дерево рішень, градієнтне прискорення і Наївний Баєс. Щоби перевірити як впливає певний метод обробки пропущених значень на конкретний метод машинного навчання.

**Методи обробки пропущених значень**

У роботі розглядаються такі методи обробки пропущених значень:

* Медіана
* Середнє значення
* Мода
* Видалення пропущених значень
* Заміна на константу

Усі методи обробки використовувалися у всіх методах машинного навчання.

**Тренувальна та тестувальна вибірка**

Перший датасет був роподілений на тренувальну і тестову вибірку на 90 і 10 % відповідно. Інші два датасети, були розподілені на 70 і 30 %.

Коли датасет невеликий, критично важливо забезпечити моделі достатню кількість даних для навчання. Співвідношення 90/10 обирається через:

* Потребу надати моделі якомога більше даних для тренування, щоб вона могла виявити закономірності навіть у обмеженій кількості даних.
* Залишення 10% даних для тестування дозволяє все ще оцінити модель, хоч і з меншою точністю порівняно з великим датасетом.

**Порівняння результатів**

Зліва відображається діаграма, а з справа таблиця результатів. Як видно на діаграмі, найкращі результати дає червоний стовпець – видалення пропущених значень. Найкращим методом являється дерево рішень, порівняно з іншими методами машинного навчання. Наступні найкращі результати є у логістичної регресії і КНН. Цей датасет є невеликим, тому метод КНН швидко обраховувалася і була ефективна у цьому датасеті.

Зі збільшенням даних у 60 разів, помінялися результати суттєво, а саме найкращі результати дає логістична регресія і градієнтне прискорення. Найкращий метод оброблення пропущених значень є медіана, після неї йде мода. Наприклад у Наївному Баєсі усі методи окрім заміни на константу видали приблизно одинаковий результат.

У найбільшому наборі даних найкращими методами являються медіана, середнє значення і мода, також і заміна на константу. Але найкращий метод є середнє значення. Хочу підкреслити, що метод КНН обчислювався врази довше порівняно з іншими методами обробки пропущених значень. Він найбільш затратний по часу, але видав доволі хороші результати. Найкраща модель машинного навчання у цьому наборі даних є градієнтне прискорення.

**Висновок**

Для аналізу великих наборів даних найкращими моделями машинного навчання виявилися градієнтне прискорення та логістична регресія з поєднанням методів обробки пропущених значень медіана і мода. Ці методи забезпечують велику точність. Найменш ефективним методом виявився К-найближчих сусідів, оскільки він є найповільніший серед усіх методів. Найкращими методами обробки пропущених значень виявилися медіана і мода. І найменш точним стає метод при збільшені даних – видалення пропущених значень. На диво, також заміна на константу дала хороші результати. Але на цей метод краще не використовувати, оскільки точність цього методу варіюється від вашої удачі. Я був дуже здивований отримати таку точність від цього методу.

1. **K-найближчих сусідів (KNN)**:

* Нелінійний метод класифікації та регресії, який призначає новий зразок до класу більшості його найближчих сусідів у багатовимірному просторі ознак. Проста і зрозуміла техніка, яка не вимагає навчання моделі, але може бути повільною при великих обсягах даних.

1. **Логістична регресія**:

* Лінійний метод для бінарної класифікації, який оцінює ймовірність належності зразка до одного з двох класів, використовуючи логістичну функцію. Широко використовується через свою простоту, інтерпретованість і ефективність на лінійно роздільних наборах даних.

1. **Дерево рішень**:

* Нелінійний метод, який використовує структуру дерева для прийняття рішень, розділяючи дані на підмножини на основі значень ознак. Простий у розумінні та візуалізації, але схильний до переобучення (overfitting) без належної регуляризації.

1. **Градієнтне прискорення (Gradient Boosting)**:

* Потужний ансамблевий метод, який побудовує модель шляхом послідовного додавання слабких моделей, зазвичай дерев рішень, з метою мінімізувати похибку. Високоефективний для різних задач, але може бути вимогливим до обчислювальних ресурсів та налаштувань гіперпараметрів.

1. **Наївний Баєс (Naive Bayes)**:

* Проста й ефективна імовірнісна модель, яка базується на теоремі Баєса з припущенням про незалежність ознак. Добре працює на текстових даних і в задачах класифікації, хоча припущення про незалежність рідко справджується в реальних даних.