

ВІДГУК

офіційного опонента на дисертаційну роботу Земляного Олексія Дмитровича на тему: **«Розроблення методів та програмного забезпечення інтелектуального імпутовання пропусків даних»**, представлену на здобуття ступеня доктора філософії за спеціальністю 121 Інженерія програмного забезпечення

Актуальність теми дисертаційної роботи. Проблема пропусків у даних є однією з найактуальніших у сучасному аналізі даних, машинному навчанні та інтелектуальних системах. Пропущені значення виникають через помилки збору даних, технічні збої, неповну інформацію від респондентів або об'єднання гетерогенних джерел. Вони суттєво впливають на якість моделей, знижуючи точність класифікації, прогнозування та прийняття рішень.

Дисертаційна робота Земляного О. Д. присвячена розробленню інтелектуальних методів імпутовання пропусків, які враховують структуру даних, патерни пропусків, кореляційні та ентропійні зв'язки між ознаками. Такий підхід дозволяє підвищити точність заповнення пропусків і покращити якість подальших аналітичних моделей. Актуальність дослідження не викликає сумнівів, адже сучасні системи аналізу даних вимагають надійних та ефективних рішень для обробки неповних даних.

Зв'язок роботи з науковими програмами та планами. Дисертаційна робота виконана у відповідності до індивідуального плану підготовки аспіранта кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара. Дослідження здійснювалося в рамках науково-дослідних робіт № ФПМ-2-22 «Розроблення програмного забезпечення аналізу та кластеризації часових рядів» (2022–2024 рр., № держреєстрації

0122U001465), № 58-ФПМ-2-25 «Розроблення інформаційної технології обробки статистичних даних» (2025–2027 рр., № держреєстрації 0125U002280). Дослідження було частково профінансовано Міністерством закордонних справ Чеської Республіки в межах проєкту 24-PKVV-UM-011 «Посилення стандартів викладання, досліджень та міжнародного співробітництва в Дніпровському національному університеті імені Олеся Гончара (ДНУ)», що підкреслює міжнародну значущість роботи.

Формулювання наукової задачі, розв’язання якої одержане в дисертації. Дисертаційна робота Земляного О. Д. спрямована на розв’язання актуальної науково-прикладної задачі – розроблення методів, алгоритмів та програмного забезпечення для інтелектуального імпутування пропусків у даних, які забезпечують точність, ефективність та універсальність у задачах моніторингу, класифікації та прогнозування.

Наукова задача сформульована як комплексне дослідження та розробка інтелектуальних підходів до імпутування пропусків, які враховують структуру даних, патерни пропусків, залежності між ознаками.

Конкретні завдання дослідження:

1. Провести аналіз існуючих методів імпутування та виявити їхні обмеження у роботі з різномірними даними.

2. Розробити нові методи імпутування на основі класифікації та регресії, ентропійному підході та гібридні методи, що поєднують ентропійний і регресійний підходи.

3. Реалізувати програмне забезпечення для імпутування пропусків у вигляді класів, сумісних з архітектурою scikit-learn, з можливістю інтеграції в аналітичні конвеєри (pipeline).

4. Оцінити ефективність запропонованих методів за критеріями:

- точність імпутування;

- швидкодія алгоритмів;
- вплив на якість моделей класифікації та прогнозування.

5. Провести експериментальне тестування на реальних датасетах (медичних та гідрологічних) для підтвердження практичної цінності розробок.

Наукова новизна отриманих результатів. Вважаю, що наукова новизна отриманих результатів полягає у наступному:

1. Удосконалено метод NoNa розробленням методу UnifiedClassRegrImputer, який поєднує класифікатор і регресор з урахуванням типу даних ознак, враховує патерни пропусків і надає змінний порядок обходу ознак залежно від кількості пропусків. Це дозволяє підвищити точність імпутування та покращити якість моделей у задачах класифікації та прогнозування.

2. Вперше розроблено метод RegrImputer, де запропоновано застосовувати регресор з уточненням значення для категоріальних даних. Це забезпечує значне підвищення швидкодії та точності імпутування, а також покращує якість моделей у задачах класифікації та прогнозування.

3. Удосконалено ентропійний метод EntropyImputer додаванням двох модифікацій: однокрокової та ітераційної процедури зменшення умовної ентропії кожної ознаки у послідовному та паралельному виконанні. Це дозволяє зменшити умовну ентропію та час виконання алгоритму.

4. Вперше розроблено гібридний метод HybridRegrEntropyImputer, який поєднує ентропійний та регресійний підходи в залежності від типу даних ознаки та патернів пропусків. Це підвищує точність імпутування та швидкодію порівняно з EntropyImputer.

5. Вперше розроблено метод CorrelationImputer на основі виявленого кореляційного зв'язку між ознаками, який автоматизує вибір найкращої

моделі залежності серед лінійних та квазілінійних моделей. Це підвищує точність імпутування, а також точність моделей у задачі прогнозування часових рядів.

6. Вперше розроблено метод імпутування для часових рядів з сезонною складовою, який виконує імпутування для кожного окремого місяця року. Це дозволяє підвищити точність імпутування та якість моделей прогнозування у порівнянні з аналогом без розбиття по місяцям.

7. Удосконалено методи перетворення якісних ознак на кількісні: розроблено дві модифікації IgnoreNaNLabelEncoder та IgnoreNaNFrequentEncoder, які зберігають інформацію про пропуски та дозволяють виконувати зворотне перетворення. Це важливо для автоматизації обробки даних, де імпутація відбувається після перетворення ознак.

Практичне значення одержаних результатів. Практична цінність дослідження виявляється в таких аспектах:

1. Створено алгоритми та програмні рішення для заповнення пропусків у даних, які втілюють запропоновані підходи (UnifiedClassRegrImputer, RegrImputer, EntropyImputer, HybridRegrEntropyImputer, CorrelationImputer, IgnoreNaNLabelEncoder, IgnoreNaNFrequentEncoder). Усі розробки реалізовані на мові Python з дотриманням архітектурних стандартів бібліотеки scikit-learn, що гарантує їхню безпроблемну інтеграцію в існуючі процеси обробки даних. Це відкриває можливість застосовувати методи у складі аналітичних конвеєрів (pipeline) разом з іншими етапами, такими як масштабування, класифікація або прогнозування.

2. Запропоновані методи дозволяють підвищити якість заповнення пропусків у даних медичного та гідрологічного моніторингу. Їхня ефективність залежить від особливостей даних, патернів пропусків та їхньої кількості.

3. Розроблені підходи сприяють покращенню якості моделей класифікації та прогнозування для даних медичного та гідрологічного моніторингу порівняно з базовими моделями без імпутування та найпростішим стандартним методом SimpleImputer.

4. Результати дисертації впроваджено в освітній процес кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара. Розроблені методи та програмне забезпечення використовуються під час викладання дисциплін «Аналіз даних на мові Python», «Оптимізація та підвищення продуктивності програмного коду», «Аналіз та візуалізація даних», «Технології пошуку структури в даних» для студентів першого (бакалаврського) рівня вищої освіти галузі інформаційних технологій.

5. Запропоновані методи та програмне забезпечення можуть бути впроваджені в реальні системи аналізу даних, що дозволить підвищити якість моделей класифікації та прогнозування, зокрема для даних медичного та гідрологічного моніторингу.

Ступінь обґрунтованості наукових положень, висновків та рекомендацій. Наукові положення, висновки та рекомендації, сформульовані в дисертаційній роботі, ґрунтуються на ретельному аналізі сучасних методологій обробки даних, статистичних підходів та методів машинного навчання. Автор використав сучасні архітектурні принципи бібліотеки scikit-learn для реалізації запропонованих методів, що забезпечує їхню уніфіковану інтеграцію в аналітичні конвеєри. Обґрунтованість результатів підтверджується експериментальним тестуванням на реальних датасетах медичного та гідрологічного моніторингу, а також порівняльним аналізом з існуючими методами імпутування, такими як SimpleImputer, IterativeImputer, kNNImputer.

Ефективність розроблених методів оцінювалася за критеріями точності імпутування, швидкодії та впливу на якість моделей класифікації та прогнозування. Результати досліджень підтверджують, що запропоновані підходи забезпечують покращення якості обробки даних порівняно з традиційними методами.

Особистий внесок здобувача. Усі результати дисертаційної роботи, що виносяться на захист, отримані автором особисто та за його безпосередньою участю. Дисертаційна робота є завершеним науковим дослідженням, виконаним автором відповідно до його індивідуального навчального плану. Зокрема, дисертантом проведено пошук та аналіз літературних джерел; побудову математичних моделей; числове розв'язання, програмну реалізацію та ілюстрацію одержаних результатів.

Оцінка структури дисертації, мови та стилю викладення. Дисертаційна робота має класичну структуру, яка включає вступ, 4 розділи, висновки, список використаних джерел та 3 додатки.

У *вступі* обґрунтовано актуальність теми дослідження, визначено мету та завдання дослідження, описано об'єкт і предмет дослідження, а також наукову новизну та практичне значення роботи.

Розділ 1 містить огляд сучасних підходів до обробки даних, зокрема проблеми пропусків, яка є однією з ключових перешкод для ефективного машинного навчання. Розглянуто задачі класифікації на прикладі діагностування ішемічної хвороби серця та прогнозування часових рядів на основі даних гідрологічного моніторингу. Наведено детальний огляд існуючих методів імпутування, їхню класифікацію за механізмами виникнення та стратегії заповнення.

Розділ 2 присвячений опису розроблених методів імпутування, які враховують структуру даних, патерни пропусків, кореляційні та ентропійні

зв'язки між ознаками. Описано алгоритми перетворення якісних ознак на кількісні зі збереженням інформації про пропуски. Наведено UML-діаграми діяльності для візуалізації логіки роботи алгоритмів.

Розділ 3 містить опис реалізації розроблених алгоритмів на мові Python з інтеграцією в екосистему бібліотеки scikit-learn. Описано оптимізацію коду для покращення продуктивності обчислень, зокрема векторизацію обчислень та порівняння багатопоточної та багатопроцесорної реалізації ентропійного методу.

Розділ 4 присвячений тестуванню розроблених методів імпутування. Описано датасети для задач класифікації та прогнозування часових рядів. Проведено 4 типи тестів: порівняння точності імпутування та швидкодії, аналіз впливу імпутування на якість моделей у задачах класифікації та прогнозування, аналіз змін ентропії до та після імпутування. Наведено демонстрацію роботи методів у конвеєрі (pipeline).

Висновки складаються з 10 пунктів. Вони логічно та послідовно сформульовані, повністю відображають отримані здобувачем результати в ході експериментальних досліджень. Чітка та логічна побудова розділу дозволяє характеризувати здобувача як сформованого науковця, здатного самостійно вирішувати поставлені завдання.

Список використаних джерел містить 121 найменування.

Додатки містять перелік публікацій здобувача, код розроблених класів та акт впровадження результатів роботи в освітній процес.

Робота написана українською мовою з використанням сучасної наукової термінології. Матеріал викладено логічно, послідовно та чітко.

Зауваження по роботі

Незважаючи на високий науковий рівень роботи, є декілька зауважень, які не впливають на загальну позитивну оцінку, але можуть бути враховані для подальшого вдосконалення:

1. У першому розділі доцільно було б систематизувати результати аналізу існуючих методів імпутування та виокремити їхні переваги і недоліки. Це дозволило б більш обґрунтовано сформулювати постановку задач дослідження.

2. У другому розділі на сторінках, де описуються алгоритми імпутування, бажано надати більше деталей щодо вибору конкретних моделей машинного навчання (наприклад, чому саме вибрано ті чи інші класифікатори/регресори).

3. У розділі 3 доцільно було б обґрунтувати вибір архітектури scikit-learn для реалізації методів.

4. У розділі 4 доцільно було б надати пояснення, чому саме такі датасети було вибрано для оцінки ефективності запропонованих методів.

5. Автор розглядає кількісні та якісні ознаки, але було б цікаво дізнатися, чи запропоновані методи впораються з бінарними даними, у яких пропуски можуть мати специфічні патерни.

Загальний висновок по дисертаційній роботі

Враховуючи актуальність, новизну, важливість одержаних автором наукових результатів, їх обґрунтованість і достовірність, а також цінність сформульованих положень і висновків, вважаю, що дисертаційна робота Земляного Олексія Дмитровича «Розроблення методів та програмного забезпечення інтелектуального імпутування пропусків даних», представлена на здобуття ступеня доктора філософії за спеціальністю 121 Інженерія програмного забезпечення, відповідає вимогам, що висуваються до дисертацій, які встановлені відповідно наказу МОН України № 40 від 12.01.2017 р. «Про затвердження вимог до оформлення дисертації», «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії» (Постанова Кабінету

Міністрів України № 44 від 12 січня 2022 р. зі змінами від 21.03.2022 р.), а її автор Земляний Олексій Дмитрович заслуговує на присудження ступеня доктора філософії зі спеціальності 121 Інженерія програмного забезпечення.

Офіційний опонент

доцент кафедри інформатики та програмної інженерії факультету інформатики та обчислювальної техніки Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», кандидат технічних наук



**Андрій
ГУБСЬКИЙ**

Підпис Губського А. М. засвідчую.

Т.в.о Вчений секретар КПІ ім. Ігоря Сікорського



**Валерія
ХОЛЯВКО**
Valeria Holivko