РЕЦЕНЗІЯ

на дисертаційну роботу Земляного Олексія Дмитровича на тему «Розроблення методів та програмного забезпечення інтелектуального імпутування пропусків даних», представлену на здобуття ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 121 Інженерія програмного забезпечення

Актуальність теми дисертації

Стрімкий розвиток інформаційних технологій, методів машинного навчання та інтелектуального аналізу даних зумовлює зростання вимог до якості та повноти інформації, що використовується для побудови моделей класифікації, прогнозування та підтримки прийняття рішень. Однією з найбільш поширених проблем сучасних інформаційних систем є наявність пропущених значень у наборах даних, що може бути спричинена особливостями процесів збору, передавання та зберігання інформації. Неповнота даних негативно впливає на достовірність результатів їх аналізу, знижує точність алгоритмів машинного навчання та обмежує можливості застосування сучасних методів обробки даних у критично важливих прикладних галузях Особливої актуальності проблема імпутування пропусків набуває в умовах використання великих обсягів різнорідних даних, що характеризуються складними структурними та статистичними залежностями між ознаками, а також різними механізмами виникнення пропущених значень. Незважаючи на значну кількість існуючих методів імпутування, питання розроблення універсальних, обчислювально-ефективних та адаптивних підходів, здатних враховувати особливості предметної області, типи даних та патерни пропусків, залишається актуальним науково-прикладним завданням. Не менш важливим є забезпечення можливості інтеграції таких методів до сучасних програмних платформ аналізу даних із дотриманням принципів повторного

використання програмних компонентів, модульності та сумісності зі стандартними конвеєрами обробки даних.

Дисертаційна робота Земляного О.Д. є завершеним дослідженням, у якому розроблені методи та програмне забезпечення інтелектуального імпутування пропусків даних.

Актуальність теми дослідження

Проблема імпутування пропущених значень є складною та багатоаспектною, оскільки ефективність її розв'язання значною мірою залежить від механізмів виникнення пропусків, типів даних, наявності кореляційних залежностей між ознаками та особливостей предметної області.

Дисертаційна робота Земляного О.Д. присвячена розробленню методів та програмного забезпечення інтелектуального імпутування пропусків даних, які поєднують підходи на основі класифікації, регресійного аналізу, ентропійних та кореляційних залежностей. Важливою перевагою проведеного дослідження є не лише створення нових та удосконалення існуючих методів імпутування. Запропоновані підходи дозволяють підвищити якість імпутування пропущених значень та ефективність подальшого використання даних у задачах класифікації та прогнозування, що визначає наукове та практичне значення отриманих результатів.

Зв'язок роботи з науковими програмами, планами, темами

Дисертаційна робота виконувалась у відповідності з індивідуальним планом підготовки аспіранта кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара. Дослідження здійснювалось в рамках науково-дослідної роботи № ФПМ-2-22 «Розроблення програмного забезпечення аналізу та кластеризації часових рядів» 2022-2024 рр. номер держреєстрації

0122U001465 та № 58 – ФПМ-2-25 «Розроблення інформаційної технології обробки статистичних даних» 2025-27рр. номер держреєстрації 0125U002280. Дослідження було частково профінансоване 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)», реалізованого Карловим університетом і ДНУ: Project «Strengthening the Standards of Teaching, Research and International Cooperation at Oles Honchar Dnipro National University (DNU)» granted by the Ministry of Foreign Affairs of the Czech Republic 2024, Registration Number 24-PKVV-UM-011, Activity 2.1.4: Providing mini-grants for research and publication activities of DNU researchers, project name – Imputation of gaps in hydrological monitoring data.

Формулювання наукової задачі, нове вирішення якої одержане в дисертації

Наукова задача дисертаційного дослідження полягає у розробленні та вдосконаленні методів інтелектуального імпутування пропусків даних і програмного забезпечення для їх реалізації, здатних забезпечити підвищення точності відновлення пропущених значень у різнорідних наборах даних шляхом урахування механізмів виникнення пропусків, структурних та статистичних залежностей між ознаками, а також особливостей предметної області. Розв'язання поставленої наукової задачі передбачає створення адаптивних, обчислювально-ефективних та сумісних із сучасними програмними платформами методів обробки даних, придатних для інтеграції до стандартних конвеєрів машинного навчання та інтелектуального аналізу даних. *Об'єктом дослідження* є процеси обробки, аналізу та підготовки даних в системах моніторингу. *Предметом дослідження* є методи алгоритми та програмні засоби інтелектуального

імпутування пропущених значень у вибірках для підвищення ефективності обробки даних в системах моніторингу.

Основні завдання дослідження:

1. Провести аналіз існуючих методів імпутування та виявити переваги та обмеження наявних підходів.

2. Розробити нові методи та алгоритми імпутування, а саме запропонувати нові підходи, орієнтовані на різні типи даних і завдань, зокрема методи на основі класифікації та регресії, ентропійному підході та гібридні методи, що поєднують ентропійний і регресійний підходи.

3. Розробити алгоритми та програмне забезпечення методів імпутування на мові програмування Python відповідно до архітектурних принципів бібліотеки scikit-learn.

4. Оцінити ефективність нових методів за точністю імпутування та часом обробки на датасетах медичного та гідрологічного моніторингу.

5. Проаналізувати вплив нових методів імпутування на якість моделей класифікації та прогнозування для даних медичного та гідрологічного моніторингу.

Наукова новизна отриманих автором результатів

Наукова новизна одержаних результатів полягає у розробленні та вдосконаленні методів інтелектуального імпутування пропусків даних і програмного забезпечення для їх реалізації, які забезпечують підвищення точності відновлення пропущених значень, обчислювальної ефективності їх обробки та покращення якості моделей класифікації й прогнозування. Основними науковими результатами, що визначають новизну дисертаційного дослідження, є:

1. Удосконалено метод імпутування NoNa шляхом розроблення методу UnifiedClassRegrImputer, який поєднує використання класифікатора та регресора з урахуванням типів ознак, патернів виникнення пропусків та змінного порядку їх обробки залежно від кількості відсутніх значень, що

дозволило підвищити точність імпутування та ефективність використання даних у задачах класифікації та прогнозування.

2. Вперше розроблено метод імпутування RegrImputer, особливістю якого є застосування регресійного підходу з уточненням значень для категоріальних даних, що забезпечує підвищення швидкодії та точності відновлення пропущених значень порівняно з існуючими підходами.

3. Удосконалено ентропійний метод EntropyImputer шляхом реалізації однокрокової та ітераційної процедур зменшення умовної ентропії ознак із підтримкою послідовного та паралельного виконання алгоритмів, що сприяє підвищенню обчислювальної ефективності та якості імпутування даних.

4. Вперше розроблено гібридний метод імпутування HybridRegrEntropyImputer, який адаптивно поєднує ентропійний та регресійний підходи залежно від типів даних і патернів виникнення пропусків, забезпечуючи підвищення точності імпутування та швидкодії алгоритму.

5. Вперше розроблено метод імпутування CorrelationImputer, заснований на автоматизованому виборі найбільш придатної моделі залежності між ознаками серед лінійних та квазілінійних моделей, що дозволяє підвищити точність відновлення даних та покращити результати прогнозування часових рядів.

6. Вперше запропоновано підхід до імпутування пропущених значень у часових рядах із сезонною складовою на основі кореляційного аналізу ознак із виконанням процедури імпутування окремо для кожного місяця року, що забезпечує підвищення точності відновлення даних та якості моделей прогнозування порівняно з підходами без урахування сезонності.

7. Удосконалено методи перетворення якісних ознак на кількісні шляхом розроблення модифікованих засобів кодування IgnoreNANLabelEncoder та IgnoreNANFrequentEncoder, які забезпечують збереження інформації про пропущені значення, враховують особливості розподілу категоріальних ознак та підтримують можливість виконання зворотного перетворення даних.

Практичне значення одержаних результатів

Практичне значення одержаних результатів полягає у розробленні методів та програмного забезпечення для інтелектуального імпутування пропусків даних, які можуть бути використані в сучасних системах аналізу даних та машинного навчання. Запропоновані алгоритми реалізовано мовою програмування Python відповідно до архітектурних принципів бібліотеки scikit-learn, що забезпечує можливість їх інтеграції до стандартних конвеєрів обробки даних (pipeline) спільно з методами масштабування, класифікації та прогнозування.

Проведені експериментальні дослідження підтверджують, що розроблені методи забезпечують підвищення точності імпутування пропущених значень у даних медичного та гідрологічного моніторингу, а також сприяють покращенню якості моделей класифікації та прогнозування порівняно з базовими підходами до обробки неповних даних. Особливістю запропонованих методів є їх адаптивність до різних типів даних, механізмів виникнення пропусків та наявності кореляційних залежностей між ознаками, що розширює можливості їх практичного застосування.

Розроблене програмне забезпечення характеризується модульністю та обчислювальною ефективністю, що дозволяє використовувати його як складову сучасних програмних засобів інтелектуального аналізу даних. Запропоновані методи можуть бути використані під час розроблення програмних систем підтримки прийняття рішень, інформаційно-

аналітичних систем та засобів обробки даних у прикладних галузях, де критично важливим є забезпечення достовірності результатів аналізу за наявності пропущених значень.

Практична цінність отриманих результатів підтверджується їх впровадженням в освітній процес кафедри інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара під час викладання дисциплін, пов'язаних з аналізом даних, технологіями машинного навчання та оптимізацією програмного коду. Крім того, запропоновані методи та програмне забезпечення мають перспективи впровадження у реальних системах аналізу даних, зокрема в задачах медичного та гідрологічного моніторингу, де дозволяють підвищити якість моделей класифікації та прогнозування.

Обґрунтованість і достовірність наукових положень, висновків і рекомендацій, які захищаються

Дисертація виконана на належному науковому рівні та відповідає вимогам до досліджень третього рівня вищої освіти. Усі отримані результати є науково обґрунтованими, що забезпечується коректністю постановки задач та узгодженістю отриманих результатів з відкритих джерел.

Обґрунтованість наукових положень, висновків та рекомендацій, сформульованих у дисертаційній роботі Земляного Олексія Дмитровича, забезпечується коректною постановкою наукової задачі, використанням сучасних методів інтелектуального аналізу даних, машинного навчання та програмної інженерії, а також комплексним підходом до проведення теоретичних та експериментальних досліджень.

Достовірність одержаних результатів підтверджується застосуванням апробованих методів математичної статистики, регресійного та кореляційного аналізу, використанням сучасних програмних засобів реалізації алгоритмів машинного навчання, а також проведенням широкого

спектра експериментальних досліджень на реальних наборах даних медичного та гідрологічного моніторингу. Експериментальна перевірка запропонованих методів здійснювалася шляхом їх порівняння з існуючими підходами до імпутування пропущених значень за показниками точності відновлення даних, обчислювальної ефективності та впливу на якість моделей класифікації й прогнозування.

Наукові положення дисертаційної роботи є достатньо аргументованими та логічно пов'язаними з поставленою метою дослідження. Запропоновані автором методи досліджені в різних умовах виникнення пропусків та для різних типів даних, що дозволило визначити особливості їх використання та оцінити ефективність запропонованих рішень.

Практичну достовірність отриманих результатів підтверджує їх впровадження в освітній процес Дніпровського національного університету імені Олеся Гончара, а також можливість використання розробленого програмного забезпечення в сучасних інформаційно-аналітичних системах та програмних комплексах обробки даних.

Особистий внесок здобувача

Аналіз літературних джерел, розроблення алгоритмів та програмного забезпечення імпутування пропусків у даних, порівняльний аналіз методів, оптимізація коду на мові програмування Python, обробка отриманих результатів виконані автором самостійно. Постановка мети і завдань дослідження, аналіз і узагальнення отриманих результатів проводились спільно з науковим керівником доктором технічних наук, професором О. Г. Байбузом.

Апробація матеріалів дисертації

Основні положення та результати дисертаційної роботи доповідалися й обговорювалися на XXI, XXII, XXIII Міжнародних науково-практичних конференціях «Математичне та програмне забезпечення інтелектуальних

систем (MPZIS)» (м. Дніпро, 2023, 2024, 2025), XXVII Міжнародній конференції «Автоматика-2024» (м. Дніпро, 2024), Виклики та проблеми сучасної науки (м. Дніпро, 2024), Міжнародній науковопрактичній інтернет-конференції «Тенденції та перспективи розвитку науки і 26 освіти в умовах глобалізації» (м. Переяслав, 2023), підсумкових наукових конференціях Дніпровського національного університету імені Олеся Гончара (м. Дніпро, 2023, 2024, 2025, 2026).

Мова і стиль роботи

Дисертаційна робота характеризується чіткою та логічною структурою викладу матеріалу. Її зміст є послідовним, системним і повною мірою відображає результати проведених наукових досліджень. Автором коректно використано сучасну наукову термінологію, а виклад основних положень роботи здійснено грамотною українською мовою з дотриманням вимог наукового стилю. Теоретичні положення, результати експериментальних досліджень та сформульовані висновки належним чином обґрунтовані та взаємопов'язані між собою. Зміст дисертації відповідає вимогам, що висуваються до дисертаційних робіт на здобуття ступеня доктора філософії, а наведені в ній результати повною мірою розкривають поставлені мету та завдання дослідження й відображають особистий науковий внесок автора.

Публікації

За темою дисертаційної роботи опубліковано у 11 роботах, серед яких 4 статті в періодичних наукових журналах і збірниках наукових праць категорії Б, 2 статті у інших збірниках наукових праць та 5 тез доповідей у збірниках матеріалів наукових конференцій.

Структура та обсяг дисертації

Дисертаційна робота викладена на 169 сторінках, включаючи вступ, 4 розділи, висновки, список використаних джерел охоплює 121 найменування

та додатки. Робота є завершеним науковим дослідженням, яке поєднує теоретичні розробки, алгоритмічне забезпечення та практичну реалізацію.

У *вступі* обґрунтовані актуальність теми, мета та завдання роботи, наведені наукова новизна, практичне значення, особистий внесок дисертанта та апробація роботи. В основу дисертаційної роботи покладені результати, отримані автором під час виконання наукових досліджень, проведених на кафедрі інженерії програмного забезпечення та інформаційних технологій Дніпровського національного університету імені Олеся Гончара.

У *першому розділі* проведено огляд сучасних підходів до обробки пропусків даних. Особлива увага приділена аналізу задач класифікації (на прикладі діагностування ішемічної хвороби серця) та прогнозування часових рядів (на прикладі гідрологічного моніторингу басейну ріки Дніпро). Цей розділ демонструє розуміння проблеми та обґрунтовує необхідність розробки нових методів.

У *другому розділі* автор пропонує опис розроблених методів імпутування (UnifiedClassRegrImputer, RegrImputer, EntropyImputer, HybridRegrEntropyImputer, CorrelationImputer). Особливо цікавим є урахування патернів пропусків та адаптація методів до різних типів даних (категоріальних, кількісних, часових рядів).

У *третьому розділі* автор реалізував розроблені алгоритми у вигляді бібліотеки на мові Python, сумісної з архітектурою scikit-learn. Це дозволяє інтегрувати методи в конвеєри обробки даних разом з іншими етапами, такими як масштабування, класифікація або прогнозування. Особлива увага приділена оптимізації коду (векторізація, паралельне виконання) для підвищення продуктивності.

У *четвертому розділі* наведено експериментальне дослідження ефективності розроблених методів на реальних датасетах (медичних та гідрологічних). Проведено порівняння точності імпутування та швидкодії

рызними методами, проаналізовано вплив імпутування на якість моделей класифікації та прогнозування, проведено аналіз змін ентропії до та після імпутування. Результати свідчать про переваги розроблених методів порівняно з існуючими підходами для розглянутих датасетів.

У *висновках* до дисертації підведено підсумки дослідження та виокремлено основні наукові та практичні здобутки автора при розробленні методів та програмного забезпечення інтелектуального імпутування пропусків даних. Розроблення інтелектуальних методів імпутування є актуальним і перспективним завданням, яке має значний потенціал для подальшого дослідження та розширення в сферах, де важлива обробка неповних даних.

Результати дослідження та наукові положення чітко сформульовано та якісно узагальнено.

Окремі дискусійні питання і зауваження

1. У дисертаційній роботі запропоновано методи імпутування UnifiedClassRegrImputer та RegrImputer, які ґрунтуються на використанні моделей класифікації та регресії. Разом з тим, опис процедури вибору конкретних моделей машинного навчання та критеріїв їх налаштування представлено недостатньо детально. Більш ґрунтовне обґрунтування вибору базових моделей та аналіз їх впливу на результати імпутування дозволили б повніше оцінити універсальність і можливості практичного застосування запропонованих підходів.

2. У підрозділі 2.5.4 (с. 57) автором зазначено про проведення рефакторингу програмного коду та його оптимізації. Водночас більш детальне висвітлення причин проведення рефакторингу, особливостей реалізації внесених змін та їх впливу на швидкодію й масштабованість розробленого програмного забезпечення дозволило б повніше оцінити ефективність запропонованих архітектурних рішень.

3. У підрозділі 2.5 дисертації автор приділяє значну увагу питанням інтеграції розроблених методів до архітектури бібліотеки scikit-learn. Водночас представлення архітектури програмного забезпечення могло б бути доповнене UML-діаграмами класів або компонентів із відображенням взаємозв'язків між реалізованими імп'ютерами, засобами кодування ознак та програмними інтерфейсами бібліотеки scikit-learn. Це сприяло б кращому розумінню особливостей реалізації та повторного використання програмних компонентів.

4. На сторінці 83 (табл. 3.2) наведено результати порівняння ефективності різних реалізацій ентропійного методу імпутування (послідовної, багатопотокової та багатопроцесної). Водночас потребують додаткового пояснення причини відмінностей в отриманих результатах для різних способів реалізації одного й того ж алгоритму. Доцільним було б більш детально проаналізувати вплив особливостей паралельних обчислень, використання генераторів випадкових чисел та програмної реалізації алгоритму на відтворюваність і точність отриманих результатів.

5. Автором виконано оптимізацію програмної реалізації запропонованих методів, зокрема шляхом проведення рефакторингу та векторизації програмного коду. Проте в роботі недостатньо уваги приділено опису проведених змін та кількісній оцінці їх впливу на підвищення продуктивності програмного забезпечення. Наведення більш детального аналізу ефективності застосованих підходів дозволило б повніше оцінити внесок виконаної оптимізації у загальну швидкодію розроблених алгоритмів.

6. У підрозділі 3.3 наведено результати оцінювання якості моделей класифікації та прогнозування після застосування розроблених методів імпутування. Разом із тим робота могла б виграти від більш детального аналізу обчислювальної складності запропонованих алгоритмів та їх масштабованості залежно від розмірності наборів даних. Такий аналіз

дозволив би повніше оцінити перспективи використання розробленого програмного забезпечення при роботі з великими обсягами даних.

7. У розділі 3 наведено результати експериментальних досліджень ефективності запропонованих методів імпутування на даних медичного та гідрологічного моніторингу. Разом із тим було б доцільно більш детально обґрунтувати вибір параметрів експериментів, зокрема частки пропущених значень та механізмів їх генерації. Наведення результатів досліджень для різних сценаріїв виникнення пропусків (MCAR, MAR та MNAR) дозволило б повніше оцінити особливості застосування розроблених методів у практичних задачах.

8. У роботі реалізовано програмне забезпечення відповідно до архітектурних принципів бібліотеки scikit-learn та забезпечено можливість його інтеграції щодо обробки даних. Разом із тим у дисертації практично не розглянуто питання тестування розробленого програмного забезпечення. Наведення результатів модульного та інтеграційного тестування програмних компонентів дозволило б повніше оцінити їх надійність, коректність функціонування та готовність до практичного використання.

Проте, зазначені зауваження не впливають на загальний високий рівень дисертаційної роботи. Автором досягнуто значних наукових і практичних результатів, які мають значення для розвитку методів імпутування пропусків у даних. Це дозволяє оцінити дисертацію як цілісне, завершене та актуальне наукове дослідження.

Загальні висновки

Дисертаційна робота Земляного Олексія Дмитровича «Розроблення методів та програмного забезпечення інтелектуального імпутування пропусків даних» є актуальною за змістом, містить наукову новизну, основні результати та висновки обґрунтовані, мають теоретичне та практичне значення. Оформлення дисертації відповідає вимогам, затвердженим наказом Міністерства освіти і науки України № 40 від 12.01.2017 року.

Методичний рівень, наукова новизна і практичне значення роботи відповідають вимогам до дисертацій на здобуття ступеня доктора філософії згідно з Постановою Кабінету Міністрів України від 12 січня 2022 року № 44 (зі змінами) «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», а здобувач Земляний Олексій Дмитрович заслуговує на присудження ступеня доктора філософії в галузі знань 12 Інформаційні технології за спеціальністю 121 Інженерія програмного забезпечення.

Рецензент

завідувачка кафедри
кібербезпеки і комп'ютерно-інтегрованих
технологій фізико-технічного факультету
Дніпровського національного університету
імені Олеся Гончара,
кандидат технічних наук, доцент Світлана КЛИМЕНКО

Підпис канд. техн. наук, доцента Клименко С.В. засвідчую

Проректор з наукової роботи
Дніпровського національного університету
імені Олеся Гончара,
кандидат біологічних наук, професор Олег МАРЕНКОВ