

УДК 004.412:519.25

DOI [https://doi.org/10.15589/znp2025.2\(500\).37](https://doi.org/10.15589/znp2025.2(500).37)**A MATHEMATICAL MODEL FOR CODE METRICS INFORMATION
PROCESSING OF DATA SCIENCE AND MACHINE LEARNING JAVA
APPLICATIONS TO ESTIMATE THEIR SIZE****МАТЕМАТИЧНА МОДЕЛЬ ДЛЯ ОБРОБКИ ІНФОРМАЦІЇ З МЕТРИК
КОДУ DATA SCIENCE ТА MACHINE LEARNING JAVA-ЗАСТОСУНКІВ
ДЛЯ ОЦІНЮВАННЯ ЇХ РОЗМІРУ****Oleksandr S. Oriekhov**

oleksandr.oriekhov@nuos.edu.ua

ORCID: 0000-0002-0001-0140

Tetyana A. Farionova

tetyana.farionova@nuos.edu.ua

ORCID: 0000-0003-3384-4712

О. С. Орехов,

аспірант

Т. А. Фаріонова,

канд. техн. наук, доцент

*Admiral Makarov National University of Shipbuilding, Mykolaiv**Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв*

Abstract. *The aim of the study* is to increase the accuracy and reliability of code metrics information processing to estimate the size of Data Science (DS) and Machine Learning (ML) JAVA applications in the early stages of software project planning using class diagram metrics by building a nonlinear regression model for further use in parametric models of software development effort estimation. Accurate software size estimation is necessary for project planning, resource allocation, and costing in software development. *Methods.* The research methods include analysis of existing regression equations and models for estimating the KLOC of JAVA applications; collection and processing of information from the code metrics of JAVA applications in the DS and ML areas from the training and testing datasets; ; methods for iterative nonlinear regression models constructing, including methods of probability theory, mathematical statistics, multivariate statistical analysis, linear and nonlinear regression analysis; etc. *Results.* A five-factor nonlinear regression model for processing information from code metrics for early KLOC estimation of DS and ML JAVA applications and its prediction interval were constructed. The obtained quality criteria assessments of R^2 , $MMRE$, and $PRED$ (0,25) for the nonlinear regression models on the basis of the initial training sample and the test sample show a high level of accuracy in estimating the KLOC parameter using the obtained five-factor regression compared to the existing regression models and the constructed four-factor regression. The obtained prediction interval is 19.2% shorter than the prediction interval of the built four-factor regression. *Scientific novelty.* Firstly, a five-factor nonlinear regression model was constructed on the basis of multivariate non-Gaussian data from DS and ML JAVA applications. Firstly, the metric of the total quantity of classes and interfaces was decomposed into separate metrics of classes and interfaces. *The practical significance* of the obtained results allows us to recommend the five-factor model for use in practice to process information from code metrics for early size estimation of DS and ML JAVA applications.

Key words: JAVA; Data Science; Machine Learning; non-Gaussian data; nonlinear regression model; normalizing transformation.

Анотація. *Метою* дослідження є підвищення достовірності обробки інформації з метрик коду для оцінювання розміру Data Science (DS) та Machine Learning (ML) JAVA-застосунків на ранніх стадіях розробки програмних проєктів за метриками діаграми класів, шляхом побудови нелінійної регресійної моделі, для подальшого застосування в параметричних моделях оцінювання трудомісткості розробки. Достовірна оцінка розміру програмного забезпечення необхідна для планування проєкту, розподілу ресурсів і витрат у розробці програмного забезпечення. *Методика.* Методи дослідження включають аналіз існуючих регресійних рівнянь та моделей для оцінювання KLOC JAVA-застосунків; збір та обробка інформації з метрик коду JAVA-застосунків за напрямком DS та ML з навчального та тестового набору даних; застосування методики ітеративної побудови нелінійних регресійних моделей, які включають методи теорії ймовірності, математичної статистики, багатовимірного статистичного аналізу, лінійного та нелінійного регресійного аналізу; тощо. *Результати.* Побудовано п'ятифакторну нелінійну регресійну модель для обробки інформації з метрик коду для ранньої оцінки KLOC DS

та ML JAVA-застосунків та її інтервал передбачення. Оцінки критеріїв якості нелінійних регресійних моделей R^2 , $MMRE$ та $PRED$ (0,25) отримані на основі початкової навчальної та тестової вибірок свідчать про високий рівень точності оцінювання параметру KLOC за допомогою побудованої п'ятифакторної моделі у порівнянні з існуючими моделями. Отриманий інтервал передбачення на 19,2 % вужчий ніж інтервал передбачення чотирьохфакторної регресії. *Наукова новизна*. Вперше побудовано п'ятифакторну нелінійну регресійну модель для багатовимірних негаусівських даних DS та ML JAVA-застосунків. Вперше загальну метрику кількості класів та інтерфейсів розділено на окремі метрики класів та інтерфейсів. *Практична значимість* отриманих результатів дозволяє рекомендувати п'ятифакторну модель для використання на практиці для обробки інформації з метрик коду для ранньої оцінки розміру DS та ML JAVA-застосунків.

Ключові слова: JAVA; Data Science; Machine Learning; негаусівські дані; нелінійна регресійна модель; нормалізуюче перетворення.

ПОСТАНОВКА ЗАДАЧІ

Оцінка трудомісткості розробки програмного забезпечення (ПЗ) є одним з ключових показників для бізнесу при формуванні бюджету та плануванні часу та ресурсів, які необхідні для розробки програмного проєкта. Достовірне оцінювання трудомісткості розробки ПЗ дозволить отримати інформацію щодо планування програмного проєкту на ранніх стадіях розробки, що в свою чергу допоможе враховувати ризики та підвищити ефективність процесу розробки [1]. Створення сучасних програмних продуктів неможливо уявити без застосування Data Science (DS) та Machine Learning (ML) технологій, оскільки ринок розробки потребує впровадження цих технологій для отримання конкурентної переваги та пропозиції кращих програмних рішень та сервісів для користувачів. Капіталізація та активне зростання ринку ML свідчить про перспективність розвитку цього напрямку індустрії інформаційних технологій [2], що призводить до активного впровадження технологій та розробок DS і ML в IT інфраструктуру різних галузей, таких як охорона здоров'я, освіта, фінанси та економіка, аерокосмічна галузь, біоінформатика, тощо [3; 4]. Оскільки мова програмування JAVA є однією з найбільш затребуваних [5], тому саме обробка інформації з метрик програмного коду для оцінювання розміру JAVA-застосунків, які використовують або впроваджують технології DS та ML є актуальною задачею, яка невід'ємно пов'язана з життєвим циклом управління програмним проєктом.

Інформація про розмір, а саме кількість рядків коду, JAVA-застосунків для категорії DS та ML проєктів на ранніх етапах розробки дає змогу отримати прогноз трудомісткості розробки ПЗ за допомогою параметричних моделей COCOMO, COCOMO II, COSYSMO, які використовують параметр розміру у вигляді рядків коду для оцінки трудомісткості розробки програмних проєктів [6]. Отримання достовірної оцінки кількості рядків коду є важливим завданням, результат якого впливає на успіх проєкту та зменшує ризики на ранньому етапі розробки. Це потребує застосування відповідних математичних моделей.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

На теперішній час для обробки інформації метрик JAVA-застосунків для оцінювання їх розміру побудовано як лінійні [7; 8] так і нелінійні [9–14] регресійні рівняння та моделі в залежності від метрик концептуальної моделі даних у вигляді діаграми класів. Зокрема в роботі [12] побудовано нелінійну регресійну модель обробки інформації JAVA-застосунків для оцінювання параметру розміру за напрямком DS та ML.

Трьохфакторні лінійні регресійні рівняння [7; 8] не можуть бути використані через застарілу навчальну вибірку. Оскільки мова програмування JAVA за період 20 років зазнала значних змін, метрики коду з програмних застосунків мають негаусівський розподіл та вибірка має гетероскедастичну природу. Нелінійні регресійні моделі: трьохфакторна нелінійна регресійна модель [9] з використанням вибірки інформації з метрик JAVA-застосунків з робіт [7; 8], однофакторна нелінійна регресійна модель для оцінки розміру JAVA веб-застосунків [10], чотирьохфакторна нелінійна регресійна модель на базі метрик CLASS, SMQ LCOM і RFC [11], однофакторні нелінійні регресійні моделі на базі метрики CLASS [13], двофакторна нелінійна регресійна модель на основі метрик CLASS і VMQ [13], трьохфакторна нелінійна регресійна модель на основі метрик CLASS, RFC, aVMQ, та чотирьохфакторна нелінійна регресійна модель на основі метрик CLASS, RFC, aCBO і aVMQ, мають високий рівень достовірності для їх навчальних та тестових вибірок. Але дослідження [13–14] показують, що достовірність оцінювання може відрізнятися в залежності від вибірки, яку було застосовано для навчання моделі, та, у певних випадках, навчальна вибірка може застаріти, у зв'язку із розвитком мови програмування JAVA та зміни концепцій розробки, також для кожної побудованої моделі, навчальна вибірка та нормалізуючі перетворення накладають певні обмеження діапазонів даних, що також впливає на точність оцінювання. Відповідно до [1; 6] категорія програмного застосунку, як фактору середовища, має вплив на достовірність при оцінюванні параметру розміру. Так в роботі [12] проведено порівняння

регресійних моделей [9–11] і рівнянь [7; 8] на вибірці інформації з метрик JAVA-застосунків за напрямком DS та ML, який показав недостатній рівень точності оцінювання параметру KLOC для DS та ML JAVA-застосунків. Моделі [11; 14] мають високі показники точності оцінювання, але вони побудовані на основі метрики RFC, недолік якої полягає в тому, що вона вимагає інформації про унікальні виклики всіх залежних методів незалежно від модифікатору доступу до методу (public, protected, private), на виклик певного методу даного класу. Однак, UML діаграма класів не може надати достатньої деталізації для цієї метрики, оскільки в ній відсутня інформація про внутрішні виклики методів, послідовність викликів і алгоритми, які реалізовані в методах. У порівнянні з public та protected методами, які визначаються як частина інтерфейсу класу, private методи та їхні зв'язки з іншими методами зазвичай визначаються на етапі розробки, та можуть суттєво змінюватись протягом на стадії розробки. Метрику RFC можливо визначити на основі історичних даних з попередніх проєктів, але це створює труднощі для коректної оцінки абсолютного значення цієї метрики, коли інформація про попередні проєкти відсутня недостатньо [15]. В роботі [12] побудовано нелінійну регресійну модель для оцінювання розміру DS та ML JAVA-застосунків із використанням нормалізуючого перетворення у вигляді десятичного логарифму за метриками CLASS, TFQ та VMQ. Модель має високі показники якості, але не була проведена перевірка достовірності за допомогою тестової вибірки.

ВІДОКРЕМЛЕННЯ НЕВИРІШЕНИХ РАНІШЕ ЧАСТИН ЗАГАЛЬНОЇ ПРОБЛЕМИ

Аналіз сучасних досліджень показав, що існуючі моделі та рівняння для обробки інформації з метрик програмного коду JAVA застосунків для оцінювання їх розміру не забезпечують точність оцінювання параметру розміру для DS та ML JAVA-застосунків. Отже, побудова математичних моделей для оцінювання розміру JAVA-застосунків за напрямком DS та ML є актуально науково-практичною задачею, яка потребує вирішення.

МЕТА ДОСЛІДЖЕННЯ

Метою дослідження є підвищення достовірності обробки інформації з метрик коду для оцінювання розміру DS та ML JAVA-застосунків на ранніх стадіях розробки програмних проєктів за метриками діаграми класів, шляхом побудови нелінійної регресійної моделі, для подальшого застосування в параметричних моделях оцінювання трудомісткості розробки.

МЕТОДИ, ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Методи. Дослідження проводиться з використанням наступних методів: пошук, системний аналіз,

синтез, порівняння, абдукція, історичний і логічний методи, теорії ймовірності, математичної статистики, багатовимірного статистичного аналізу, лінійного та нелінійного регресійного аналізу. **Об'єктом дослідження** є процес обробки інформації з метрик коду для оцінювання розміру DS та ML JAVA-застосунків з відкритим кодом. **Предметом дослідження** є математичні моделі обробки інформації з метрик коду для оцінювання розміру (кількості рядків коду) DS та ML JAVA-застосунків.

ОСНОВНИЙ МАТЕРІАЛ

Метрики програмного коду, зазвичай, мають нагусівський розподіл, що обмежує можливість використання лінійних регресійних моделей для достовірного оцінювання параметру розміру JAVA-застосунків. Основними теоретичними умовами для застосування лінійних регресійних моделей є гаусівський розподіл залишків регресії ϵ , розподіл багатовимірних даних повинен бути гаусівським, та вибірка повинна бути гомоскедастичною.

Для побудови нелінійних регресійних моделей було обрано методику, запропоновану у роботах [16; 17], яка ґрунтується на методах статистичного та регресійного аналізу, застосуванні нормалізуючих перетворень з ітеративним видаленням викидів у негаусівських даних, перевірці нормального розподілу залишків регресії ϵ та визначення інтервалу передбачення. Процес складається з наступних шести етапів.

На **першому етапі** вибірки застосовується взаємозворотне нормалізуюче перетворення негаусівського випадкового вектору $P = \{Y, X_1, X_2, \dots, X_k\}^T$ у гаусівський випадковий вектор, $T = \{Z_Y, Z_1, Z_2, \dots, Z_k\}^T$, яке задається як

$$T = \psi(P), \quad (1)$$

де k – кількість незалежних факторів, а зворотне до (1) має вигляд

$$P = \psi^{-1}(T), \quad (2)$$

де ψ – вектор взаємозворотніх функцій нормалізуючого перетворення, $\psi = \{\psi_Y, \psi_1, \psi_2, \dots, \psi_k\}^T$.

Для нормалізації багатовимірних негаусівських даних обрано взаємозворотне багатовимірне нормалізуюче перетворення Бокса-Кокса [18]:

$$Z_j = x(\lambda_j) = \begin{cases} (X_j^{\lambda_j} - 1) / \lambda_j, & \text{якщо } \lambda_j \neq 0 \\ \ln(X_j), & \text{якщо } \lambda_j = 0 \end{cases}, \quad (3)$$

де λ_j – параметр перетворення Бокса-Кокса; j – змінюється від 1 до k ; X_j – негаусівська випадкова величина (ВВ), яка нормалізується; Z_j – нормалізована гаусівська величина.

Для оцінювання параметрів перетворення Бокса-Кокса (3) обрано логарифмічну функцію правдоподібності для багатовимірних даних, яку задано як

$$\hat{\theta} = \arg_{\theta} \max I(P, \theta), \quad (4)$$

де $I(P, \theta)$ – логарифмічна функція правдоподібності; θ – вектор параметрів перетворення, $\theta = \{\lambda_1, \lambda_2, \dots, \lambda_k\}$; k – розмірність багатовимірних даних. Для багатовимірного перетворення Бокса-Кокса логарифмічна функція правдоподібності для багатовимірних даних задано як

$$l(X, \theta) = \sum_{j=1}^k (\lambda_j - 1) \sum_{i=1}^N \ln(X_{ji}) - \frac{N}{2} \ln[\det(S_N)], \quad (5)$$

де N – розмір вибірки, S_N – вибіркова коваріаційна матриця для компонентів вектору Z та $\det(S_N)$ – детермінант вибіркової коваріаційної матриці:

$$S_N = \frac{1}{N} \sum_{i=1}^N (Z_i - \bar{Z})(Z_i - \bar{Z})^T, \quad (6)$$

де Z_i – гаусівський випадковий вектор, $Z_i = \{Z_{i1}, Z_{i2}, \dots, Z_{ik}\}^T$; $\bar{Z}$ – вектор вибірових середніх, $\bar{Z} = \{\bar{Z}_1, \bar{Z}_2, \dots, \bar{Z}_k\}$.

На **другому етапі** виконується перевірка багатовимірних даних на відповідність нормальному розподілу за допомогою критерію Мардія [20], який заснований на вимірюванні багатовимірної асиметрії ($\beta_{1,k}$) і ексцесу ($\beta_{2,k}$) багатовимірної вибірки.

$$\beta_{1,k} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left[(X_i - \bar{X})^T S_N^{-1} (X_j - \bar{X}) \right]^3, \quad (7)$$

$$\beta_{2,k} = \frac{1}{N} \sum_{i=1}^N \left[(X_i - \bar{X})^T S_N^{-1} (X_i - \bar{X}) \right]^2, \quad (8)$$

де X – k -розмірний вектор ВВ, $X = (X_1, X_2, \dots, X_k)$, а S_N – змійнена вибіркова коваріаційна матриця багатовимірної ВВ X , визначена як

$$S_N = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})^T, \quad (9)$$

де $\bar{X}$ – це вектор вибірових середніх, $\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)^T$.

Для $\beta_{1,k}$ застосовується тестова статистика

$$\frac{N}{6} \beta_{1,k} \leq \chi^2, \quad (10)$$

яка має наближений розподіл χ^2 з $k(k+1)(k+2)/6$ ступенями свободи та рівнем значущості α , який становить 0,005.

Для $\beta_{2,k}$ у якості тестової статистики використовується $1-\alpha$ квантиль нормального закону розподілу з математичним сподіванням $\mu = k(k+2)$ та дисперсією $\sigma^2 = 8k(k+2)/N$

$$\beta_{2,k} \leq N_{1-\alpha}(\mu, \sigma^2). \quad (11)$$

Якщо значення багатовимірної асиметрії $N/6\beta_{1,k}$ та ексцесу $\beta_{2,k}$ менші за порогові значення тестових

статистик, то багатовимірний вибірок вважається гаусівською. В іншому випадку визначення та вилучення викидів із багатовимірних нормалізованих даних здійснюється за методом квадрату відстані Махаланобіса [19]. Значення квадрату відстані Махаланобіса це елементи головної діагоналі матриці d^2 , розміром $N \times N$

$$d^2 = (Z_i - \bar{Z})^T S_N^{-1} (Z_i - \bar{Z}), \quad (12)$$

де S_N – змійнена вибіркова коваріаційна матриця (9), Z – ВВ з нормальним розподілом. Суть цього методу полягає в тому, що значення $d_i^2, i=1, 2, \dots, N$, які перевищують значення квантилю розподілу χ^2 для рівня значущості α вважаються викидами і виключаються із вибірки. Зазвичай використовується значення $\alpha = 0,005$ для цього методу.

На **третьому етапі**, для нормалізованих даних, виконується побудова лінійної регресійної моделі, яка має вигляд

$$Z_y = \hat{Z}_y + \varepsilon = \hat{b}_0 + \hat{b}_1 Z_1 + \hat{b}_2 Z_2 + \dots + \hat{b}_k Z_k + \varepsilon, \quad (13)$$

де ε – ВВ розподілена за нормальним законом, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; k – кількість незалежних факторів регресії; $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ – оцінки параметрів лінійної регресійної моделі, які знаходяться за методом найменших квадратів.

На **четвертому етапі** залишки регресії ε побудованої лінійної регресійної моделі (13) перевіряється на відповідність нормальному закону розподілу за допомогою критерію Пірсона χ^2 для рівня значущості $\alpha = 0,05$. Якщо розподіл залишків регресії нормалізованих даних відхиляється від гаусівського, тоді необхідно відкинути точку даних, для якої модуль залишку є максимальним і повернутись до першого етапу.

На п'ятому етапі, після побудови моделі лінійної регресії (13) для нормалізованих даних, застосувавши зворотне перетворення (2), нелінійна регресійна модель має наступний вигляд

$$Y = \Psi_Y^{-1}(\hat{Z}_Y + \varepsilon) = \Psi_Y^{-1}(\hat{b}_0 + \hat{b}_1 \psi_1(X_1) + \hat{b}_2 \psi_2(X_2) + \dots + \hat{b}_k \psi_k(X_k) + \varepsilon). \quad (14)$$

На шостому етапі побудови нелінійної регресійної моделі (14), визначається верхня та нижня границі інтервалу передбачення та проводиться пошук точок значень які знаходяться поза інтервалами передбачення. Якщо такі точки знайдені, то вони вважаються викидами і вилучаються із навчальної вибірки. Границі інтервалів передбачення нелінійної регресійної моделі (14) визначаються за наступною формулою

$$\hat{Y}_{PI} = \Psi^{-1}(\hat{Z}_Y \pm t_{\alpha/2, v} S_{Z_Y} \{1 + \frac{1}{N} + (Z_X^+)^T S_Z^{-1} (Z_X^+)\}^{1/2}), \quad (15)$$

де $t_{\alpha/2, v}$ – квантиль t -розподілу Стюдента з $v = N - k - 1$ ступенями свободи та $\alpha/2$ – рівнем значущості; $S_{Z_Y}^2 = \frac{1}{v} \sum_{i=1}^N (Z_{Y_i} - \hat{Z}_{Y_i})^2$; Z_X^+ – вектор центральних

факторів, який задано $\{Z_{1_i} - \bar{Z}_1, Z_{2_i} - \bar{Z}_2, \dots, Z_{k_i} - \bar{Z}_k\}$;
 $S_Z - k \times k$ матриця

$$S_Z = [S_{Z_q} S_{Z_r}], \quad (16)$$

де $S_{X_q} S_{X_r} = \sum_{n=1}^N (Z_{q_i} - \bar{Z}_q)(Z_{r_i} - \bar{Z}_r)$,
 $q, r = 1, 2, \dots, k$.

Для оцінювання точності регресійних моделей використовуються такі критерії якості, як коефіцієнт детермінації R^2 , значення середньої величини відносної похибки $MMRE$ та значення відсотка передбачення для рівня відносної похибки $0,25 - PRED(0,25)$. Прийнятними вважаються значення $R^2 \geq 0,75$, $MMRE \leq 0,25$ та $PRED(0,25) \geq 0,75$ [21].

Для побудови регресійної моделі зібрано навчальну вибірку, що містить інформації з метрик коду 74 DS та MLJAVA-застосунків з відкритим кодом, розташованих на платформі GitHub (<https://github.com>). Інформація

з метрик коду отримана за допомогою інструменту аналізу коду CK (<https://github.com/mauricioaniche/ck>). Отримана вибірка включає наступні метрики: рядки коду на проєкт (KLOC), загальна кількість класів (CLASS), тип класу класів (TYPE), загальна кількість видимих методів (VMQ), загальна кількість атрибутів класів (TFQ) і загальне значення зв'язності між класами (CBO). Для метрик VMQ, TFQ та CBO розраховано середні значення відносно метрики CLASS для зменшення кореляції між незалежними факторами регресії. Значення кількості класів було розділено на фактичну кількість класів (CLS) та фактичну кількість інтерфейсів (INFC), оскільки інструменту CK під метрикою CLASS збирає інформацію про класи та інтерфейси разом. Ці метрики можна отримати на ранній стадії проєктування з UML діаграми класів на відміну від метрики RFC. Отримана вибірка наведена в Таблиці 1.

Таблиця 1. Навчальна вибірка інформації з метрик JAVA-застосунків

№	Репозиторій	KLOC	CLS	INFC	aVMQ	aTFQ	aCBO	CLASS
1	2	3	4	5	6	7	8	9
1	atlas.git	154,568	1757	89	6,380	3,768	7,008	1846
2	curator.git	49,161	1088	248	3,493	1,171	4,976	1336
3	hama.git	47,105	667	82	5,602	2,525	6,059	749
4	helix.git	156,854	2282	144	4,612	2,355	6,464	2426
5	iceberg.git	258,993	3667	316	5,941	2,607	8,551	3983
6	incubator-celeborn.git	28,49	396	24	5,664	4,062	7,521	420
7	incubator-hivemall.git	76,739	1022	34	5,057	2,134	6,633	1056
8	incubator-myriad.git	8,056	194	6	5,090	2,875	7,070	200
9	incubator-nemo.git	32,999	812	62	3,501	1,944	6,934	874
10	incubator-retired-edgent.git	20,371	423	66	4,456	1,765	5,223	489
11	incubator-seatunnel.git	105,104	2455	210	3,599	3,143	6,302	2665
12	inlong.git	256,154	4469	406	4,356	4,459	7,225	4875
13	mahout.git	95,439	1727	90	4,571	1,980	6,133	1817
14	opennlp.git	64,504	1049	182	3,871	1,822	5,480	1231
15	sqoop.git	69,624	871	45	6,412	2,975	7,144	916
16	submarine.git	112,035	615	74	9,901	3,463	7,131	689
17	mlkit.git	9,181	168	6	3,339	3,569	7,563	174
18	datahub-gma.git	20,01	336	14	5,526	1,826	9,623	350
19	dr-elephant.git	19,825	342	9	3,943	4,385	5,544	351
20	spring-data-elasticsearch.git	51,23	974	124	5,757	2,208	6,714	1098
21	Data_Science_with_Java.git	2,804	87	7	3,277	1,372	5,128	94
22	data-science-in-action.git	9,741	245	1	2,667	1,744	5,004	246
23	angel.git	204,252	1736	145	7,678	2,668	7,547	1881
24	alluxio.git	222,656	3343	240	5,833	3,462	6,716	3583
25	smile.git	98,284	1157	143	7,396	2,265	5,405	1300
26	Alink.git	458,003	6488	1386	4,249	1,936	6,512	7874
27	grobid.git	58,864	524	19	6,805	4,414	5,203	543
28	EasyML.git	29,343	627	50	4,208	1,739	4,449	677
29	oryx.git	14,469	392	22	3,495	1,447	6,845	414
30	seldon-server.git	29,704	601	60	4,885	3,050	4,850	661
31	elasticsearch-learning-to-rank	15,737	280	18	5,101	2,107	7,671	298
32	spatial-framework-for-hadoop	5,754	148	0	2,608	1,541	8,446	148
33	tribuo.git	288,533	1554	278	13,111	3,816	9,846	1832
34	datumbox-framework.git	21,604	392	18	3,944	0,998	5,766	410

Продовження таблиці 1

1	2	3	4	5	6	7	8	9
35	Mallet.git	70,037	766	64	6,712	4,478	6,833	830
36	cv4j.git	7,301	136	4	3,150	1,914	4,329	140
37	kafka-streams-machine-learning-examples	196,02	112	0	1,589	2,554	3,304	112
38	JSAT.git	87,712	1104	71	8,725	2,406	7,207	1175
39	ABAGAIL.git	12,985	293	31	4,336	1,901	5,873	324
40	h2o-3.git	362,876	5155	187	4,335	2,414	5,947	5342
41	moa.git	107,002	1408	108	5,494	3,966	5,455	1516
42	codebuff.git	119,496	2124	101	5,764	1,738	5,742	2225
43	trident-ml.git	4,549	92	8	4,950	1,920	5,240	100
44	Conjecture.git	4,595	63	8	10,451	2,324	4,282	71
45	dagli.git	137,599	2367	605	5,485	1,593	12,590	2972
46	ytk-learn.git	20,453	227	11	4,483	4,181	5,340	238
47	k-NN.git	5,976	132	2	4,716	2,082	7,097	134
48	flink-ml.git	41,079	903	131	3,581	1,871	8,142	1034
49	shifu.git	61,27	765	20	5,485	3,740	5,954	785
50	antimalwareapp.git	1,649	35	0	2,514	2,371	6,514	35
51	entity-fishing.git	29,591	255	1	8,125	3,434	6,965	256
52	FakeImageDetection.git	2,161	44	7	2,157	3,490	5,451	51
53	TextAnalyzer.git	6,885	157	15	4,134	1,860	2,890	172
54	pyramid.git	47,598	605	28	5,384	2,763	6,616	633
55	pignlproc.git	1,937	30	0	4,400	2,400	5,833	30
56	DL-Learner.git	88,46	914	101	7,344	3,613	8,490	1015
57	tsml-java.git	290,27	1894	157	10,954	4,598	7,129	2051
58	Java-Machine-Learning.git	3,539	66	15	5,025	1,889	3,790	81
59	mltk.git	19,314	206	10	5,963	1,870	6,079	216
60	cleartk.git	35,976	908	65	2,888	1,576	5,683	973
61	Foundry.git	160,264	2162	286	7,036	1,223	6,102	2448
62	hudi.git	230,858	3341	181	5,527	3,089	9,024	3522
63	morpheus-core.git	50,934	487	76	8,028	1,940	7,149	563
64	systemds.git	339,938	3664	72	8,563	3,656	8,163	3736
65	shinyproxy.git	1,846	36	0	4,722	2,750	8,000	36
66	vespa.git	624,258	11230	689	5,120	1,990	6,179	11919
67	djl-serving.git	12,676	173	5	4,562	3,017	6,742	178
68	serving.git	224,262	535	112	17,815	3,799	8,495	647
69	toolbox.git	43,157	684	41	5,559	1,981	7,982	725
70	pattern.git	3,146	128	2	3,238	1,008	5,031	130
71	graph-data-science.git	200,338	4536	1083	3,850	1,496	6,372	5619
72	nexus-repository-r.git	2,877	74	4	3,987	3,231	8,936	78
73	ml.git	4,216	134	5	3,403	1,568	5,230	139
74	systemds.git	207,825	2329	12	8,091	3,809	7,713	2341

Для побудови п'ятифакторної нелінійної регресійної моделі були обрані метрик CLS, INFC, aVMQ, aTFQ і aCBO, як незалежні фактори, оскільки розраховані VIF становлять 2,4, 2,5, 1,3, 1,3 і 1,2, що вказує на відсутність мультиколінеарності між ними. Отже вибір середніх значень на CLASS метрик VMQ, TFQ і CBO для побудови нелінійної регресійної моделі є виправданим.

Шестивимірні дані навчальної вибірки перевірено на відповідність гаусівському розподілу даних із застосування багатовимірного тесту Мардія, та виявлено, що розподіл шестивимірних даних CLS, INFC,

aVMQ, aTFQ, aCBO і KLOC навчальної вибірки не є гаусівським, оскільки оцінка багатовимірної асиметрії $N\beta_1 / 6 = 718,94$ перевищує значення квантиля χ^2 86,99 для 56 ступенів свободи та рівня значущості $\alpha = 0,005$, а оцінка багатовимірного ексцесу $\beta_2 = 112,31$ перевищує значення квантилю розподілу Гауса, яке дорівнює 51,89 для математичного сподівання 48 і дисперсії 2,28. Отже перехід до побудови нелінійної регресії є виправданим.

Для підвищення точність та достовірності оцінювання параметру розміру KLOC (Y) JAVA-застосунків за напрямком DS та ML, побудуємо п'ятифакторну

нелінійну регресійну модель із застосування навчальної вибірки розміром 74 проекти, розміщених на Github за метриками CLS (X_1), INFC (X_2), aVMQ (X_3), aTFQ (X_4) і aCBO (X_5) на основі шестивимірного нормалізуючого перетворення Бокса-Кокса. Для застосування багатовимірного перетворення Бокса-Кокса, значення незалежного фактору INFC збільшено на 1, для уникнення нульових значень.

П'ятифакторну нелінійну регресійну модель побудовано за 10 ітерацій, було виявлено і вилучено 9 рядків викидів, а саме:

- за квадратом відстані Махаланобіса видалено проекти під номерами 37, 68 та 44, оскільки значення квадрату відстані Махаланобіса становила 62,60, 23,77 та 20,03 відповідно, що перевищило значення квантилю розподілу χ^2 для рівня значущості $\alpha = 0,05$, яке становить 14,86;

- на четвертому етапі, як викиди, видалено 5 проектів під номерами 16, 26, 23, 40 та 34, оскільки залишки регресії не були розподілені за гаусівським законом розподілу, відповідно до значення критерію Пірсона 13,52, 12,43, 12,24, 13,39 та 10,73, які перевищили порогове допустиме значення 9,49 для рівня значущості $\alpha = 0,05$ та 4 ступенів вільності.

- на шостому етапі видалено проект 8, оскільки значення оцінки KLOC було за межами інтервалу передбачення для рівня значущості $\alpha = 0,05$.

На останній ітерації шестивимірна нормалізована навчальна вибірка ($N = 65$) перевірена на відповідність гаусівському розподілу за допомогою багатовимірного тесту Мардіа. Визначено, що розподіл шестивимірних даних є гаусівським, оскільки оцінка багатовимірної асиметрії $N\beta_1/6 = 71,6$ не перевищує значення квантиля χ^2 86,99 для 56 ступенів свободи та рівня значущості $\alpha = 0,05$, а оцінка багатовимірного ексцесу $\beta_2 = 50,23$ не перевищує значення квантилю розподілу Гауса, яке дорівнює 50,84 для значення математичного сподівання 48 і дисперсії 2,43. Отже, п'ятифакторну нелінійну регресійну модель побудовано. Оцінки параметрів багатовимірного нормалізуючого перетворення Бокса-Кокса з шістьма змінними становлять $\hat{\lambda}_y = 0,104173$, $\hat{\lambda}_{x_1} = 9,151706 \cdot 10^{-2}$, $\hat{\lambda}_{x_2} = 0,103777$, $\hat{\lambda}_{x_3} = 0,513100$, $\hat{\lambda}_{x_4} = -7,194852 \cdot 10^{-3}$ і $\hat{\lambda}_{x_5} = 0,168831$, та оцінки параметрів лінійної регресійної моделі (13) для нормалізованих даних $\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4$, і $\hat{b}_5$ дорівнюють $-4,101551, 0,792901, 2,703106 \cdot 10^{-3}, 0,508769, 0,458755$ та $-3,258332 \cdot 10^{-2}$ відповідно для останньої ітерації. Для отриманої регресійної моделі (14) побудовано інтервал передбачення та довірчий інтервал. Для отриманого інтервалу передбачення розраховані значення нормалізованих вибірових середніх: $\bar{Z}_1 = 8,694431$, $\bar{Z}_2 = 4,485228$, $\bar{Z}_3 = 2,527193$, $\bar{Z}_4 = 0,875256$, та $\bar{Z}_5 = 2,189649$. Для $t_{a/2,v} = 2,001$

рівня значущості $\alpha = 0,05$ та 59 ступенів свободи, $S_{Z_y} = 0,223327$, обернена коваріаційна матриця (16) визначається як

$$S_Z^{-1} = \begin{pmatrix} 1,4491 \cdot 10^{-2} & -1,1869 \cdot 10^{-2} & -2,5827 \cdot 10^{-3} & -8,3694 \cdot 10^{-3} & -8,2195 \cdot 10^{-3} \\ -1,1869 \cdot 10^{-2} & 1,2722 \cdot 10^{-2} & -1,3732 \cdot 10^{-3} & 1,0439 \cdot 10^{-2} & 4,1455 \cdot 10^{-3} \\ -2,5827 \cdot 10^{-3} & -1,3732 \cdot 10^{-3} & 3,4815 \cdot 10^{-2} & -2,6659 \cdot 10^{-2} & -1,68175 \cdot 10^{-2} \\ -8,3694 \cdot 10^{-3} & 1,0439 \cdot 10^{-2} & -2,6659 \cdot 10^{-2} & 1,429 \cdot 10^{-1} & -1,4449 \cdot 10^{-2} \\ -8,2195 \cdot 10^{-3} & 4,1455 \cdot 10^{-3} & -1,6817 \cdot 10^{-2} & -1,4449 \cdot 10^{-2} & 1,8016 \cdot 10^{-2} \end{pmatrix}$$

П'ятифакторна нелінійна регресійна модель оцінювання значення KLOC JAVA-застосунків за напрямком DS та ML має наступні обмеження значень незалежних факторів: $X_1 \in [30; 11230]$, $X_2 \in [1; 1084]$, $X_3 \in [2,1569; 13,1114]$, $X_4 \in [1,0077; 4,5982]$ та $X_5 \in [2,8895; 12,5902]$. На останній ітерації для факторів CLS, INFC, aVMQ, aTFQ і aCBO розраховані VIF дорівнюють 2,5, 2,6, 1,4, 1,3 і 1,2 відповідно, що вказує на відсутність мультиколінеарності між незалежними факторами.

П'ятифакторна регресійна модель перевірена за допомогою критеріїв якості R^2 , $MMRE$, та $PRED(0,25)$ на навчальній вибірці ($N = 65$) без 9 викидів:

$$R^2 = 0,9901, MMRE = 0,1278 \text{ і } PRED(0,25) = 0,8462.$$

Це вказує на високу якість триманої моделі для оцінювання параметру KLOC DS та ML JAVA-застосунків.

Авторами побудовано чотирьохфакторну нелінійну регресійну модель на основі метрик X_1 (CLASS), X_2 (aVMQ), X_3 (aTFQ) та X_4 (aCBO) із застосування багатовимірного перетворення Бокса-Кокса (за даними табл. 1). Метрика CLASS представлена сумою фактичної кількості класів (CLS) та інтерфейсів (INFC).

Оцінки параметрів багатовимірного нормалізуючого перетворення Бокса-Кокса з п'ятьма змінними становлять $\hat{\lambda}_y = 3,612182 \cdot 10^{-2}$, $\hat{\lambda}_{x_1} = 2,638344 \cdot 10^{-2}$, $\hat{\lambda}_{x_2} = 7,268828 \cdot 10^{-2}$ і $\hat{\lambda}_{x_4} = 0,155028$. Оцінки параметрів лінійної регресійної моделі (13) для нормалізованих даних $\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3$, і $\hat{b}_4$ дорівнюють $-4,476251, 0,960536, 0,680573, 0,382057$, та $-3,753264 \cdot 10^{-2}$. Для регресійної моделі (14) визначено інтервали передбачення та довірчі інтервали. Розраховані нормалізовані вибірові середні $\bar{Z}_1 = 7,046423$, $\bar{Z}_2 = 1,700895$, $\bar{Z}_3 = 0,892952$, та $\bar{Z}_4 = 2,151863$, і відповідно, $t_{a/2,v} = 1,9971$ для рівня значущості $\alpha = 0,05$ та 65 ступенів свободи; $S_{Z_y} = 0,214123$.

Обернена коваріаційна матриця (16) має вигляд

$$S_Z^{-1} = \begin{pmatrix} 6,2585 \cdot 10^{-3} & -7,9959 \cdot 10^{-3} & 1,6123 \cdot 10^{-3} & -7,7147 \cdot 10^{-3} \\ -7,9959 \cdot 10^{-3} & 1,2297 \cdot 10^{-1} & -4,2472 \cdot 10^{-2} & -1,9631 \cdot 10^{-2} \\ 1,6123 \cdot 10^{-3} & -4,2472 \cdot 10^{-2} & 1,0845 \cdot 10^{-1} & -2,1411 \cdot 10^{-2} \\ -7,7147 \cdot 10^{-3} & -1,9631 \cdot 10^{-2} & -2,1411 \cdot 10^{-2} & 1,7524 \cdot 10^{-1} \end{pmatrix}$$

Чотирьохфакторна нелінійна регресійна модель перевірена за допомогою критеріїв якості R^2 , $MMRE$, та $PRED(0,25)$:

$$R^2 = 0,9690, MMRE = 0,1531 \text{ і } PRED(0,25) = 0,8428.$$

Це вказує на достатню якість моделі для оцінювання параметру KLOC DS та ML JAVA-застосунків, але гіршу ніж у п'ятифакторній моделі.

Порівняємо побудовані п'ятифакторну та чотирьохфакторну нелінійні регресійні моделі, та трьохфакторну нелінійну регресійну модель для оцінювання параметру розміру DS та ML JAVA-застосунків [12] за критеріями якості регресійних моделей R^2 , $MMRE$, та $PRED(0,25)$ на початкових даних навчальної та тестовій вибірці. Аналогічно до навчальної вибірки, зібрано тестову вибірку (Таблиця 2) інформації з метрик JAVA-застосунків за напрямком DS та ML, проєкти якої відсутні в навчальній вибірці.

Отримані оцінки якості (Табл. 3) регресійних моделей на основі навчальної вибірки показують, що п'ятифакторна нелінійна регресійна модель має

найкращі показники точності за критеріями R^2 та $MMRE$ на основі навчальної та тестової вибірок, але значення $PRED(0,25)$ для навчальної вибірки менше серед трьох моделей, проте трьох- та чотирьохфакторні моделі не досягли достатнього рівня точності за критерієм $PRED(0,25)$ на тестовій вибірці. Отже, побудована п'ятифакторна нелінійна регресійна модель має найкращі оцінки критеріїв якості, та дозволяє отримати найбільш достовірні оцінки параметру розміру JAVA-застосунків за напрямком DS та ML серед розглянутих нелінійних регресійних моделей. Слід зазначити, що раніше побудована трьохфакторна нелінійна регресійна модель [12] має

Таблиця 2. Тестова вибірка інформації з метрик JAVA-застосунків

№	Репозиторій	KLOC	CLS	INFC	aVMQ	aTFQ	aCBO	CLASS
1	mastering-java-data-science.git	11,977	242	2	3,668	1,467	6,701	244
2	bigtop.git	6,637	223	12	2,715	2,209	4,740	235
3	java-data-mining-package.git	30,378	613	36	5,963	1,741	5,709	649
4	h2o-2.git	119,202	1721	20	4,909	2,733	6,049	1741
5	MiA.git	3,847	117	5	2,066	1,123	5,787	122
6	mahout.git	97,394	1754	100	4,420	2,019	6,490	1854
7	OpenRefine.git	72,46	1353	44	4,970	2,361	6,447	1397
8	tablesaw.git	48,446	684	82	7,984	2,364	6,050	766
9	odd-platform.git	23,603	680	227	3,318	1,742	8,894	907
10	DataCleaner.git	106,729	1916	248	4,630	2,373	7,131	2164
11	datavines.git	31,339	826	116	3,718	2,522	5,119	942
12	CQL.git	101,763	1082	60	7,137	8,203	11,935	1142
13	rumble.git	50,978	730	6	6,678	3,057	9,118	736
14	blockchain2graph.git	3,035	49	8	6,526	4,123	5,702	57
15	djl.git	83,615	1326	79	6,628	2,070	6,891	1405
16	serve.git	18,747	237	14	7,243	3,590	5,801	251
17	modeldb.git	54,516	442	34	9,069	3,727	12,363	476
18	conquary.git	44,793	1397	155	3,464	2,001	9,173	1552
19	eskimo.git	30,099	571	41	4,632	1,949	6,172	612
20	jvector.git	18,01	234	31	6,838	2,192	6,513	265
21	AndroidTensorFlowMachineLearningExample.git	0,339	12	1	2,462	2,462	3,154	13
22	qupath.git	158,8	1575	118	6,491	3,409	8,339	1693
23	MLKit.git	0,484	50	0	1,680	0,260	4,820	50
24	jenetics.git	40,895	844	106	5,715	1,493	6,135	950
25	Android-TensorFlow-Lite-Example.git	0,355	13	1	1,929	1,714	3,071	14
26	TonY.git	12,534	174	24	5,152	4,379	6,253	198
27	hms-ml-demo.git	43,232	803	37	3,098	3,614	5,489	840
28	product-category-predict.git	0,424	14	0	2,571	2,929	7,857	14
29	ml-commons.git	109,627	1517	39	5,349	3,828	10,569	1556
30	car-OCR.git	1,498	21	2	5,565	3,174	4,783	23
31	aibolit.git	112,751	1151	42	5,442	3,183	5,588	1193
32	k-means-mapreduce.git	0,394	6	0	5,833	2,667	6,167	6
33	spark-transformers.git	4,328	173	4	2,316	0,876	7,362	177
34	BoostSRL.git	89,489	525	31	10,739	5,664	8,076	556
35	DigitRecognizer.git	0,844	16	0	2,938	3,313	5,000	16
36	DeepLearning.git	1,083	11	0	5,455	6,455	4,545	11
37	onyx.git	10,392	207	11	3,734	1,881	3,624	218
38	opencv-object-detection.git	0,35	7	0	3,571	2,143	5,143	7

Таблиця 3. Порівняння математичних моделей за критеріями якості

Нелінійна регресійна модель	Навчальна вибірка			Тестова вибірка		
	R2	MMRE	PRED (0,25)	R2	MMRE	PRED (0,25)
Трьохфакторна нелінійна регресійна модель на базі перетворення десятичного логарифму [12]	0,9226	0,1853	0,7703	0,9347	0,2073	0,6316
Чотирьохфакторна нелінійна регресійна модель на базі багатовимірного перетворення Бокса-Кокса	0,9197	0,1776	0,7973	0,9306	0,1927	0,7105
П'ятифакторна нелінійна регресійна модель на базі багатовимірного перетворення Бокса-Кокса	0,9227	0,1721	0,7567	0,9456	0,1798	0,7632

високі значення оцінок точності для навчальної і тестової вибірок, але значення критерію $PRED(0,25)$ не досягло граничного значення для цього критерію.

Порівнюємо інтервали передбачення побудованих чотирьохфакторної (14) та п'ятифакторної (14) нелінійних регресійних моделей для початкової навчальної та тестової вибірок за наступною формулою

$$DIFF = \left(1 - \sum_{i=1}^N \frac{UB_{5x_i} - LB_{5x_i}}{UB_{4x_i} - LB_{4x_i}} \right) \cdot 100\%, \quad (17)$$

де UB , LB – верхня та нижня границя інтервалу передбачення, UB_{5x_i}, LB_{5x_i} – i -й інтервал передбачення п'ятифакторної нелінійної регресії (14), та UB_{4x_i}, LB_{4x_i} – i -й інтервал передбачення чотирьохфакторної нелінійної регресії (14).

Отримана оцінка $DIFF$ вказує на те, що інтервал передбачення п'ятифакторної нелінійної регресії на 19,2 % вузький, ніж інтервал передбачення побудованої чотирьохфакторної нелінійної регресії.

Отже побудована п'ятифакторна нелінійна регресійна модель (14) на основі багатовимірного перетворення Бокса-Кокса та метрик $CLS(X_1)$, $INFC(X_2)$, $aVMQ(X_3)$, $aTFQ(X_4)$ і $aCBO(X_5)$ має кращі показники якості обробки інформації з метрик коду для оцінювання параметру розміру JAVA-застосунків за напрямком DS та ML за критеріями R^2 , $MMRE$, та $PRED(0,25)$ та вузький інтервал передбачення у порівнянні з трьохфакторною [12] чотирьохфакторною (14) нелінійною регресійною моделлю.

ОБГОВОРЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Побудована п'ятифакторна нелінійна регресійна модель на основі багатовимірного нормалізуючого перетворення Бокса-Кокса підвищує точність обробки інформації метрик коду для оцінювання параметру розміру JAVA-застосунків за напрямком DS та ML, у порівнянні з існуючими регресійними моделями. П'ятифакторну регресійну модель вперше побудовано з використанням п'яти незалежних факторів – загальної кількості фактичних класів (CLS), загальної

кількості інтерфейсів ($INFC$), та метрик $aVMQ$, $aTFQ$, $aCBO$. Розділення метрики $CLASS$ (сума класів та інтерфейсів) на окремі метрики дозволило підвищити точність оцінювання параметру розміру із застосуванням п'ятифакторної моделі відповідно до критеріїв якості R^2 , $MMRE$, та $PRED(0,25)$ та дозволило отримати вузькі оцінки інтервалів передбачення у порівнянні з чотирьохфакторною моделлю за однаковими незалежними факторами $aVMQ$, $aTFQ$ та $aCBO$, за винятками того, що застосовувалась метрика яка представляє суму класів та інтерфейсів. У порівнянні з існуючими моделями, п'ятифакторна модель побудована без метрики якості RFC , для якої неможливо отримати достовірне значення з UML діаграми класів на ранньому етапі проектування. Застосування метрик коду, таких як $aVMQ$, $aTFQ$ та $aCBO$, які представляють середні значення, дозволило уникнути проблем з мультиколінеарністю.

ВИСНОВКИ

1. Наукова новизна отриманих результатів полягає в тому що вперше побудовано п'ятифакторну нелінійну регресійну модель для багатовимірних негаусівських даних на основі шестивимірного перетворення Бокса-Кокса, з використанням кількісних метрик коду, таких як загальна кількість класів, загальна кількість інтерфейсів, середня кількість видимих методів на клас та інтерфейси, середня кількість полів на клас, та середнє значення зв'язності між класами та інтерфейсами, які добре підходять для обробки інформації з метрик коду для оцінювання параметру розміру KLOC на ранніх етапах проектування DS та ML JAVA-застосунків. Побудована регресійна модель, у порівнянні трьохфакторною та чотирьохфакторною нелінійною регресійною моделями, має вищі оцінки значення R^2 , нижче значення $MMRE$ та вище значення $PRED(0,25)$, та інтервал передбачення п'ятифакторної регресії на 19,2 % вузький у порівнянні з чотирьохфакторною нелінійною регресійною моделлю за метриками $CLASS$, $aVMQ$, $aTFQ$ та $aCBO$.

2. Практична значущість отриманих результатів дозволяє рекомендувати побудовану модель для використання на практиці для обробки інформації з метрик коду для оцінювання параметру розміру JAVA-застосунків за напрямком DS та ML із подальшим отриманих значень в параметричних моделях оцінювання трудомісткості розробки ПЗ, таких як COSOMO II, COSYSMO та ін.

3. Перспективи подальших досліджень можуть включати поділ набору незалежних факторів для оцінки KLOC на метрики коду, пов'язані з інтерфейсом та з класами окремо, з подальшим об'єднанням оцінок KLOC на завершальній стадії процесу оцінювання, використання інших багатовимірних перетворень, використання набору даних який включає не тільки проекти з відкритим кодом.

REFERENCES

- [1] McConnel, S. (2006). *Software Estimation: Demystifying the Black Art*. Redmond, Washington: Microsoft Press.
- [2] Press, G. (2023, Nov 23). Top machine learning statistics to know [2024]. *What's the Big Data?* <https://whatsthebigdata.com/top-machine-learning-statistics/>
- [3] Kumar, N. (2024, Oct 11). Machine learning statistics. *DemandSage*. <https://www.demandsage.com/machine-learning-statistics/>
- [4] Chinthamu, N., & Karukuri, M. (2023). Data science and applications. *Journal of Data Science and Intelligent Systems*, 00. <https://doi.org/10.47852/bonviewJDSIS3202837>
- [5] TIOBE. (2024). TIOBE index. <https://www.tiobe.com/tiobe-index/>
- [6] Munialo, S. W. (2016). A review of agile software effort estimation methods. *International Journal of Computer Applications Technology and Research*, 5, 612–618. <https://doi.org/10.7753/IJCATR0509.1009>
- [7] Tan, H. B. K., Zhao, Y., & Zhang, H. (2006). Estimating LOC for information systems from their conceptual data models. In *Proceedings of the International Conference on Software Engineering*, 321–330. <https://doi.org/10.1145/1134285.1134331>
- [8] Tan, H. B. K., Zhao, Y., & Zhang, H. (2009). Conceptual data model-based software size estimation for information systems. *ACM Transactions on Software Engineering and Methodology*, 19. <https://doi.org/10.1145/1571629.1571630>
- [9] Prykhodko, N. V., & Prykhodko, S. B. (2018). A non-linear regression model for estimation of the size of JAVA enterprise information systems software. *Modeling and Information Technologies*, 85, 81–88. http://nbuv.gov.ua/UJRN/Mtit_2018_85_14
- [10] Makarova, L. M., Prykhodko, N. V., & Kudin, O. O. (2019). Constructing the non-linear regression model for size estimation of web-applications implemented in JAVA. *Herald (Kherson National Technical University)*, 69, 145–153. <http://eir.nuos.edu.ua/handle/123456789/4443>
- [11] Prykhodko, S. B., Prykhodko, N. V., & Smykodub, T. G. (2020). Four-factor nonlinear regression model to estimate the size of open source JAVA-based applications. *Scientific Notes of Taurida National V.I. Vernadsky University. Series: Technical Sciences*, 70, 157–162. <https://doi.org/10.32838/2663-5941/2020.2-1/25>
- [12] Oriekhov, O., & Farionova, T. (2024). Nonlinear regression models for software size estimation of Data Science and Machine Learning JAVA-applications. In *V International Workshop IT Project Management ITPM 2024* (May 22). http://doi.org/10.23939/IW_itpm2024.054
- [13] Oriekhov, O., & Farionova, T. (2024). Mathematical models for the size estimating of JAVA applications. *Herald (Kherson National Technical University)*, 89, 196–203. <https://doi.org/10.35546/kntu2078-4481.2024.2.28>
- [14] Oriekhov, O. (2024). The four-factor nonlinear regression model for early JAVA-applications size estimation. In *ICST-2024: Advances in Information Control Systems and Technologies*, 360–379. <https://doi.org/10.36059/978-966-397-422-4>
- [15] Subramanyam, R., & Krishnan, M. (2003). Empirical analysis of CK metrics for object-oriented design complexity: Implications for software defects. *IEEE Transactions on Software Engineering*, 29, 297–310. <https://doi.org/10.1109/TSE.2003.1191795>
- [16] Prykhodko, S., & Prykhodko, N. (2021). Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations. In S. Shkarlet, A. Morozov, & A. Palagin (Eds.), *Mathematical Modeling and Simulation of Systems (MODS'2020)*, 1265, 166–174. Springer. https://doi.org/10.1007/978-3-030-58124-4_16
- [17] Prykhodko, S., Prykhodko, N., Makarova, L., & Pukhalevych, A. (2020). Outlier detection in nonlinear regression analysis based on the normalizing transformations. In *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, 407–410. <https://doi.org/10.1109/TCSET49122.2020.235464>
- [18] Box, G. E. P. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–252.
- [19] Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. John Wiley & Sons.
- [20] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530. <https://doi.org/10.1093/biomet/57.3.519>
- [21] Port, D., & Korte, M. (2008). Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. In *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 51–60. ACM. <https://doi.org/10.1145/1414004.1414015>

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] McConnel, S. (2006). *Software Estimation: Demystifying the Black Art*. Redmond, Washington: Microsoft Press.

- [2] Press, G. (2023, Nov 23). Top machine learning statistics to know [2024]. *What's the Big Data?* <https://whatsthebigdata.com/top-machine-learning-statistics/>
- [3] Kumar, N. (2024, Oct 11). Machine learning statistics. *DemandSage*. <https://www.demandsage.com/machine-learning-statistics/>
- [4] Chinthamu, N., & Karukuri, M. (2023). Data science and applications. *Journal of Data Science and Intelligent Systems*, 00. <https://doi.org/10.47852/bonviewJDSIS3202837>
- [5] TIOBE. (2024). TIOBE index. <https://www.tiobe.com/tiobe-index/>
- [6] Munialo, S. W. (2016). A review of agile software effort estimation methods. *International Journal of Computer Applications Technology and Research*, 5, 612–618. <https://doi.org/10.7753/IJCATR0509.1009>
- [7] Tan, H. B. K., Zhao, Y., & Zhang, H. (2006). Estimating LOC for information systems from their conceptual data models. In *Proceedings of the International Conference on Software Engineering*, 321–330. <https://doi.org/10.1145/1134285.1134331>
- [8] Tan, H. B. K., Zhao, Y., & Zhang, H. (2009). Conceptual data model-based software size estimation for information systems. *ACM Transactions on Software Engineering and Methodology*, 19. <https://doi.org/10.1145/1571629.1571630>
- [9] Приходько, Н. В., & Приходько С. Б., (2018). Нелінійна регресійна модель для оцінювання розміру програмного забезпечення промислових інформаційних систем на Java. *Моделювання та інформаційні технології*, 85, 81–88. http://nbuv.gov.ua/UJRN/Mtit_2018_85_14
- [10] Макарова, Л. М., Приходько, Н. В., & Кудін, О. О. (2019). Побудова нелінійної регресійної моделі для оцінювання розміру веб-додатків, реалізованих мовою Java. *Вісник Херсонського національного технічного університету*, 69, 145–153. <http://eir.nuos.edu.ua/handle/123456789/4443>
- [11] Приходько, С.Б., Приходько, Н.В., & Смикодуб, Т.Г. (2020). Чотирихфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків з відкритим кодом. *Вчені записки ТНУ імені В.І. Вернадського. Серія: технічні науки*, 1 (2), 157–162. Режим доступу – <https://doi.org/10.32838/2663-5941/2020.2-1/25>
- [12] Oriekhov, O., & Farionova, T. (2024). Nonlinear regression models for software size estimation of Data Science and Machine Learning JAVA-applications. In *V International Workshop IT Project Management ITPM 2024* (May 22). http://doi.org/10.23939/IW_itpm2024.054
- [13] Орехов, О. С., & Фаріонова, Т. А. (2024). Математичні моделі для оцінювання розміру JAVA-застосунків. *Вісник Херсонського національного технічного університету*, 89, 196–203. <https://doi.org/10.35546/kntu2078-4481.2024.2.28>
- [14] [15] Орехов, О. С. (2024). Чотирихфакторна нелінійна регресійна модель для оцінювання розміру JAVA-застосунків на ранніх стадіях розробки. В *ICST-2024: Розвитки інформаційно-керуючих систем та технологій*, 360–379. <https://doi.org/10.36059/978-966-397-422-4>
- [15] Subramanyam, R., & Krishnan, M. (2003). Empirical analysis of CK metrics for object-oriented design complexity: Implications for software defects. *IEEE Transactions on Software Engineering*, 29, 297–310. <https://doi.org/10.1109/TSE.2003.1191795>
- [16] Prykhodko, S., & Prykhodko, N. (2021). Mathematical modeling of non-Gaussian dependent random variables by nonlinear regression models based on the multivariate normalizing transformations. In S. Shkarlet, A. Morozov, & A. Palagin (Eds.), *Mathematical Modeling and Simulation of Systems (MODS'2020)*, 1265, 166–174. Springer. https://doi.org/10.1007/978-3-030-58124-4_16
- [17] Prykhodko, S., Prykhodko, N., Makarova, L., & Pukhalevych, A. (2020). Outlier detection in nonlinear regression analysis based on the normalizing transformations. In *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, 407–410. <https://doi.org/10.1109/TCSET49122.2020.235464>
- [18] Box, G. E. P. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–252.
- [19] Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. John Wiley & Sons.
- [20] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530. <https://doi.org/10.1093/biomet/57.3.519>
- [21] Port, D., & Korte, M. (2008). Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. In *Proceedings of the 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 51–60. ACM. <https://doi.org/10.1145/1414004.1414015>.

© Орехов О. С., Фаріонова Т. А.

Дата надходження статті до редакції: 26.05.2025

Дата затвердження статті до друку: 12.05.2025