DOI: https://doi.org/10.36910/6775-2524-0560-2024-57-08
УДК: 004.67
Коваль Ігор Михайлович, аспірант
https://orcid.org/0009–0001–2083–1747
Головня Сергій Анатолійович, аспірант
https://orcid.org/0009-0005-2997-9202
Луцький національний технічний університет, м. Луцьк, Україна

МЕТОДИ ЛІНІЙНОЇ РЕГРЕСІЇ ТА K-MEANS ДЛЯ ПРОГНОЗУВАННЯ І КЛАСТЕРИЗАЦІЇ ВИРОБНИЧИХ ПОКАЗНИКІВ У ORANGE DATA MINING

Коваль І. М., Головня С. А. Методи лінійної регресії та k-means для прогнозування і кластеризації виробничих показників у Orange Data Mining. У роботі представлено специфіку роботи програмної системи Orange Data Mining у сфері дата аналітики, а саме практичне застосування для прогнозування та кластеризації виробничих показників підприємств. Розроблено та протестовано модель лінійної регресії, яка має універсальні властивості та може використовуватися підприємствами для прогнозування й коригування даних. Додатково модель доповнено алгоритмом кластеризації K-Means, що дозволяє отримати точні кластери та провести аналітику результатів. Отримані результати візуалізовано за допомогою внутрішніх інструментів програмного забезпечення. Запропоновано проміжні та загальні рекомендації щодо застосування моделі з різними типами даних. Результати експерименту свідчать, що програмна система Orange Data Mining успішно може використовуватись для прогнозування та кластеризації виробничих показників.
Ключові слова: лінійна регресія, Clustering, Orange Data Mining, дата аналітика, підприємство, K-means, передбачення, машинне навчання.

Koval I., Holovnia S. Linear regression and k-means methods for forecasting and clustering of production indicators in Orange Data Mining. The paper presents the specifics of the Orange Data Mining software system in the field of data analytics, namely, its practical application for forecasting and clustering of production indicators of enterprises. A linear regression model has been developed and tested, which has universal properties and can be used by enterprises for forecasting and adjusting data. Additionally, the model is supplemented with the K-Means clustering algorithm, which allows obtaining accurate clusters and analysing the results. The obtained results are visualised using internal software tools. Intermediate and general recommendations for applying the model with different types of data are proposed. The experimental results show that the Orange Data Mining software system can be successfully used for forecasting and clustering production indicators.
Keywords: Linear regression, Clustering, Orange Data Mining, Data mining, Enterprise, K-means, Prediction, Machine learning.

Постановка проблеми. Ефективне управління виробничою діяльністю підприємств вимагає аналізу великих обсягів даних для прийняття обґрунтованих рішень. Однією з актуальних задач є прогнозування та кластеризація виробничих показників, що відображають продуктивність, фінансовий стан та інші аспекти діяльності підприємства. Окреслені задачі набувають особливої актуальності в умовах динамічних змін ринкових відносин, що вимагають від підприємств гнучкості та адаптивності.

Сьогодення пропонує широкий спектр інструментів для аналізу даних, серед яких гідне місце посідає Orange Data Mining – платформа для візуального програмування та машинного навчання. Вона дає змогу швидко синтезувати, порівнювати та аналізувати дані, виявляти закономірності та вибудовувати прогностичну оцінку, використовуючи інтуїтивний інтерфейс. Водночас застосування Orange Data Mining для задач прогнозування та кластеризації виробничих показників вимагає:

- Розробки методології аналізу, що включає підготовку та нормалізацію даних.
- Обрання ефективних моделей для прогнозування (лінійна регресія, дерева рішень, нейронні мережі).
- Побудови кластерних моделей для виявлення груп підприємств із подібними характеристиками, що може складати основу для управлінських рішень.

Проблема дослідження полягає в тому, як адаптувати інструменти Orange Data Mining для аналізу реальних виробничих показників, забезпечуючи точність прогнозування та інтерпретацію результатів кластеризації. Слід визначити алгоритми, які найбільш валідні для вирішення цих задач і встановити критерії оцінки їх ефективності. У роботі застосовані різні підходи для аналізу даних. Оптимізовано параметри моделі для отримання точного результату, а також проведено кластеризацію отриманих показників. Отримані результати сприяють підвищенню ефективності управління виробництвом завдяки автоматизації аналізу даних і розробці адаптивних стратегій.

Формулювання мети дослідження. Метою даного дослідження є окреслення можливостей використання методів лінійної регресії та K-means для прогнозування та кластеризації виробничих

показників підприємств за допомогою Orange Data Mining, а також формування висновків щодо доцільності використання цієї програмної системи.

Аналіз останніх досліджень і публікацій. Прогнозування та кластеризація виробничих показників – важливий напрям досліджень у сфері економіки. У ряді досліджень методи кластеризації та прогнозування показують свою дієвість у різного роду виробничих системах, репрезентують багатовимірні статистичні методи, що застосовуються для аналізу виробничих даних, включаючи кластеризацію та факторний аналіз [1; 2]. Цікавими для аналізу вважаємо дослідження [3] у контексті введення до статистичного навчання з акцентом на практичне застосування в R. Воно охоплює методи регресії, класифікації та кластеризації. Хоч напрацювання [3] і стосуються переважно практичного навчання в середовищі програмування R, основні методики та алгоритми є корисними для нашого дослідження.

Orange Data Mining – відкрита платформа [4] для аналізу даних і машинного навчання, розроблена для дослідників та аналітиків. Вона базується на мові програмування Python і має графічний інтерфейс, що дає змогу працювати з даними без потреби у програмуванні. Orange пропонує інтерактивний підхід до аналізу даних через створення робочих процесів із використанням модулів (віджетів), які можна комбінувати та налаштовувати.

За допомогою Orange Data Mining можна виконувати широкий спектр завдань із аналізу даних, серед яких: візуалізація даних, попередня обробка даних, кластеризація, прогнозування, зменшення розмірності даних, оцінка моделей та перевірка їхньої якості, а також побудова звітів і візуальних інтерфейсів для представлення результатів.

Orange Data Mining вдало підходить для прогнозування виробничих показників підприємств завдяки наявності інструментів для роботи із значними обсягами даних і інтеграції різноманітних алгоритмів машинного навчання. Його можливості дають змогу:

- Автоматизувати процес аналізу даних.
- Використовувати адаптивні алгоритми для точного прогнозування.
- Легко інтерпретувати результати завдяки інтерактивним графікам.

Проте для роботи з досить великими або специфічними наборами даних може виникнути потреба в інтеграції з іншими платформами чи бібліотеками.

У Orange Data Mining для прогнозування можна використовувати такі основні елементи (віджети):

Data Preprocessing **(**очищення та підготовка), Regression (моделі регресії**),** Classification (прогнозування класів), Neural Networks (глибоке навчання), Evaluate Models (оцінка точності та порівняння), Test and Score (перевірка моделей на тестових наборах), Predictions (передбачення).

Orange Data Mining підтримує різноманітні методи аналізу та прогнозування, зокрема:

- Методи класифікації.
- Методи регресії.
- Методи кластеризації.
- Методи зменшення розмірності:
    1. Головні компоненти (PCA).
    2. t-SNE.
- Моделі глибокого навчання:
    1. Просторові нейронні мережі.
- Оцінка моделей:
    1. Крос-валідація.
    2. ROC-криві, метрики точності, чутливість тощо.

У роботі пропонується сфокусувати увагу на лінійній регресії та кластеризації методом К-середніх, як базових функціях моделей передбачення

Лінійна регресія – це метод моделювання [5; 6] залежності між скаляром (γ) та однією або кількома незалежними змінними ($x_{(1,)}$ x_2, … x_n), що дозволяє визначити лінійну функцію, яка найкраще описує залежність між змінними, щоб використовувати її для прогнозування або пояснення (1):

$$\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \qquad (1)$$

- γ: залежна змінна (прогнозована).
- x_1: незалежні змінні (фактори).
- β_0,β_i: параметри моделі.

- ϵ: залишковий шум.

У разі, якщо змінна x також є скаляром, регресію називають простою. При використанні лінійної регресії взаємозв'язок між даними моделюється за допомогою лінійних функцій, а невідомі параметри моделі оцінюються за вхідними даними.

Лінійна регресія широко застосовується для прогнозування, зокрема для вирішення простих або лінійно залежних задач. Основна ідея – використати тренд, визначений на основі наявних даних, щоб передбачити майбутні значення. Метод є одним із перших і найважливіших інструментів у прогнозуванні, але для складніших задач часто використовуються інші методи (наприклад, логістична регресія, дерева рішень чи нейронні мережі).

Лінійну регресію в Orange слід використовувати для прогнозування, аналізу зв'язків між змінними та оцінки впливу різних факторів [7]. Середовище дає змогу прогнозувати, наприклад, обсяг продукції підприємства залежно від витрат, кількості співробітників, розміру підприємства чи інших факторів. Процес прогнозування відбувається за таким планом:

- Підготовка даних: завантаження даних про виробничі показники.
- Вибір регресії: у Orange використовується віджет "Linear Regression".
- Навчання моделі: передати дані через віджет для навчання регресійної моделі.
- Прогнозування: використати модель для передбачення обсягів продукції на основі нових даних.

K-means – це метод кластеризації [8; 9], який використовується для групування об'єктів у k кластерів на основі схожості їх характеристик. Алгоритм визначає k центрів кластерів (центроїдів), які оновлюються ітеративно, щоб мінімізувати різницю між об'єктами одного кластера (2):

$$J = \Sigma_{i=1}^{k} \Sigma_{x \in C_i} \|x - \mu_i\|^2 \qquad (2)$$

- J: функція вартості (сума квадратів відстаней між об'єктами та їх центроїдами).
- C_i: кластер i.
- x: об'єкти даних.
- μ_i: центр (центроїд) кластеру i.
- $\|.\|$: відстань (зазвичай евклідова).

Алгоритм K-means у середовищі Orange можна використовувати для групування об'єктів у кластери на основі їхніх характеристик. Це дає змогу виявляти закономірності, сегментувати дані та отримувати нову інформацію про структуру набору даних. Наприклад, K-means можна застосовувати для кластеризації підприємств за продуктивністю, кількістю працівників [10], витратами або іншими ключовими показниками.

Процес кластеризації в Orange за допомогою K-means складається з таких кроків:

- Підготовка даних: Завантажте набір даних про виробничі показники підприємств у середовище Orange. Наприклад, це можуть бути обсяги продукції, чисельність співробітників або рівень витрат.
- Вибір методу кластеризації: У середовищі Orange використовується віджет "K-Means".
- Навчання моделі: Передайте дані через віджет "K-Means", який автоматично виконає кластеризацію, призначаючи кожен об'єкт до одного з кластерів.
- Аналіз результатів: Використовуйте віджети візуалізації, такі як "Scatter Plot" (Діаграма розсіювання) або "Silhouette Plot" (Графік силуетів), щоб оцінити якість кластеризації та переглянути, як об'єкти розподілені між кластерами.

Окреслений підхід дає змогу виявляти схожі групи об'єктів (наприклад, підприємства з подібною продуктивністю) та використовувати цю інформацію для оптимізації процесів чи прийняття стратегічних рішень.

Виклад основного матеріалу.

У дослідженні використано базу даних з порталу "Дія", зокрема розділ "Державна служба статистики України". Назва бази даних – "Частка продажу підприємствами роздрібної торгівлі товарів, що вироблені на території України, за товарними групами", яка містить інформацію про кількість продажу товарів підприємствами по кожному регіону [15]. Файл має розширення ".xlsx", містить у собі 189 рядків, 8 стовпчиків даних та включає розділену по регіонах інформацію про частку продажів підприємствами різного розміру за 2017 – 2023 роки.

Зважаючи на характеристики досліджуваних даних було вирішено, що для вирішення задачі прогнозування обсягів виробництва підприємства та моделювання залежностей виробленої продукції в залежності від періоду виробництва доцільно використовувати лінійну регресію. Серед

найпоширеніших специфікацій лінійних моделей є класична модель [3] лінійної регресії та узагальнена модель лінійної регресії [11].

У ході експерименту використано базу даних, що містить показники виробництва у 26 регіонах та загалом по Україні в проміжку за 2017 по 2023 роки. Дані також розподілені за розмірами підприємств від великих (large enterprises) до малих (small – sale enterprises).

Опишемо створення та тестування моделі прогнозування виробничих показників підприємств у Orange Data Mining. Для початку роботи з базою даних відповідний файл необхідно завантажити в Orange за допомогою віджету "File", (рис. 1):

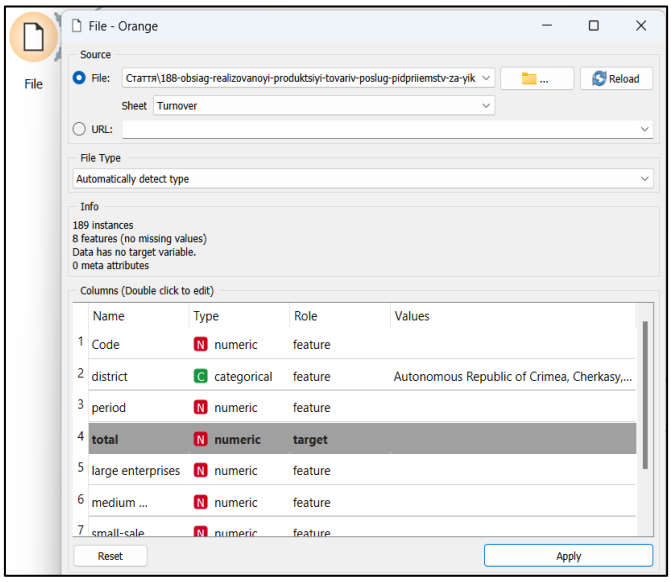


Рис. 1. Віджет "File"

Для детального ознайомлення з даними використовуються віджети: 1) "Data Info", (рис. 2), віджет не є надто інформативним, але здатен відобразити загальну базову інформацію, що міститься у файлі та 2) "Data Table" (рис. 3) – віджет, який відображає наповнення тестового файлу, в якому можна наочно переглянути дані.

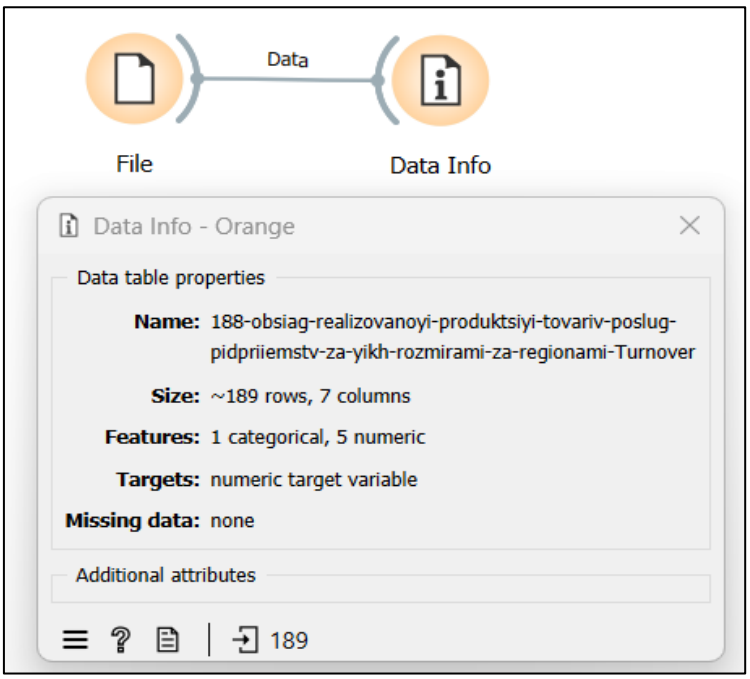
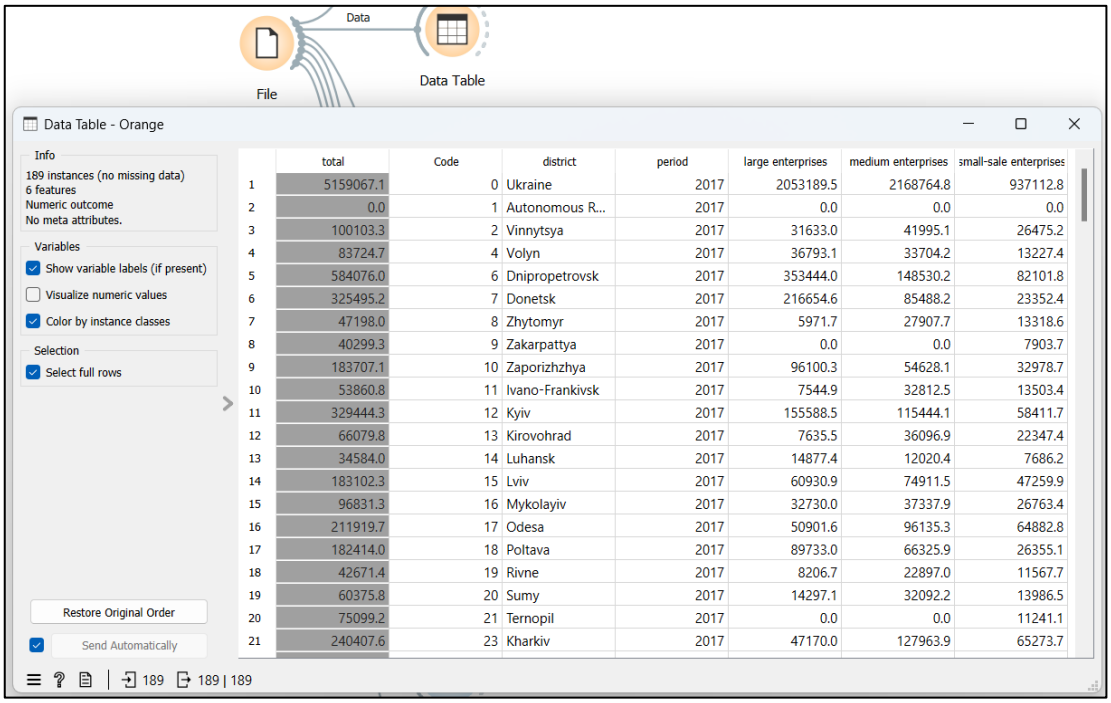


Рис. 2. Віджет "Data Info"



	total	Code	district	period	large enterprises	medium enterprises	small-sale enterprises
1	5159067.1	0	Ukraine	2017	2053189.5	2168764.8	937112.8
2	0.0	1	Autonomous R...	2017	0.0	0.0	0.0
3	100103.3	2	Vinnytsya	2017	31633.0	41995.1	26475.2
4	83724.7	4	Volyn	2017	36793.1	33704.2	13227.4
5	584076.0	6	Dnipropetrovsk	2017	353444.0	148530.2	82101.8
6	325495.2	7	Donetsk	2017	216654.6	85488.2	23352.4
7	47198.0	8	Zhytomyr	2017	5971.7	27907.7	13318.6
8	40299.3	9	Zakarpattya	2017	0.0	0.0	7903.7
9	183707.1	10	Zaporizhzhya	2017	96100.3	54628.1	32978.7
10	53860.8	11	Ivano-Frankivsk	2017	7544.9	32812.5	13503.4
11	329444.3	12	Kyiv	2017	155588.5	115444.1	58411.7
12	66079.8	13	Kirovohrad	2017	7635.5	36096.9	22347.4
13	34584.0	14	Luhansk	2017	14877.4	12020.4	7686.2
14	183102.3	15	Lviv	2017	60930.9	74911.5	47259.9
15	96831.3	16	Mykolayiv	2017	32730.0	37337.9	26763.4
16	211919.7	17	Odesa	2017	50901.6	96135.3	64882.8
17	182414.0	18	Poltava	2017	89733.0	66325.9	26355.1
18	42671.4	19	Rivne	2017	8206.7	22897.0	11567.7
19	60375.8	20	Sumy	2017	14297.1	32092.2	13986.5
20	75099.2	21	Ternopil	2017	0.0	0.0	11241.1
21	240407.6	23	Kharkiv	2017	47170.0	127963.9	65273.7

Рис. 3. Віджет "Data Table"

Для перегляду різного роду залежностей даних у графічному форматі доцільно використовувати віджет "Scatter Plot" (рис. 4). Цей графік дає змогу модифікувати змінні на осях X/Y та дійти певних висновків, наприклад, що в столичному регіоні "City of Kyiv" найбільша кількість виготовленої продукції щороку:

Рис. 4. Віджет "Scatter Plot"

За допомогою віджету "Correlation" перевіряємо, чи присутні кореляції в тестових даних (рис. 5):

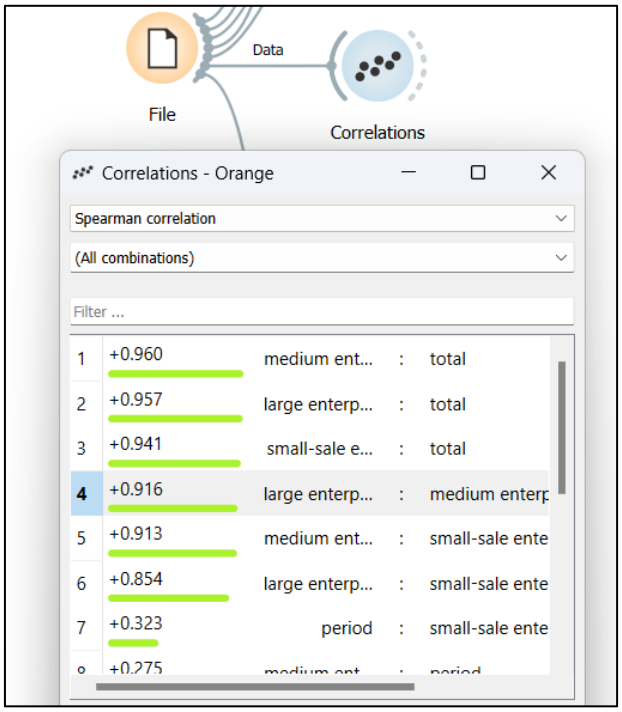


Рис. 5. Віджет "Correlation"

Згідно з отриманими результатами можна дійти висновку, що в наборі даних присутні декілька значень з високою кореляцією. Для запобігання ситуації множинної кореляції необхідно відфільтрувати дані та видалити з них ті, які в подальшому не відіграють ключової ролі. Для цього використаємо віджет "Select Columns".

Після фільтрації даних та приведення їх до необхідного вигляду слід переходити до створення лінійної регресії та передбачення обсягів виробництва товарів. Для цього використаємо віджет "Linear Regression" в парі з віджетом "Test and Score", (рис. 6). Слід зазначити, що при додаванні віджету та з'єднування його з віджетом "Data Table" автоматично створюється зв'язок типу "Selected Data" --> "Data", який необхідно виправити на "Data" --> "Data". При використанні віджета Test and Score слід приділити особливу увагу якості та повноті даних, також обов'язково представити результати у декількох метриках, задля повноти оцінки моделі. Якщо класи у класифікаційних задачах незбалансовані, метрики, такі як точність (Accuracy), можуть відображати неправильний результат. У такому випадку застосовується F1-score, Precision або Recall.

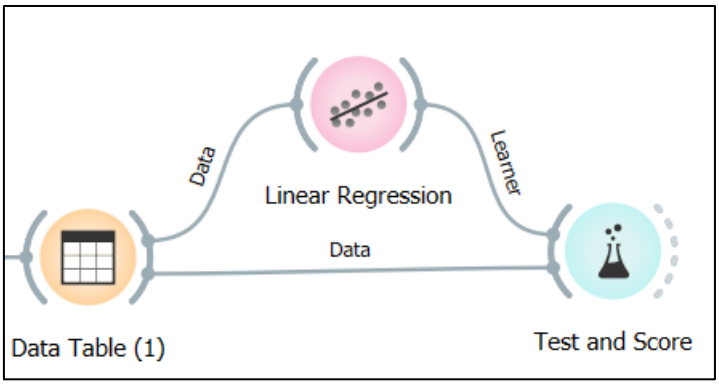


Рис. 6. Набір віджетів для лінійної регресії

В результаті використання моделі лінійної регресії та відформатованого набору даних отримано такі оцінки (рис. 7):

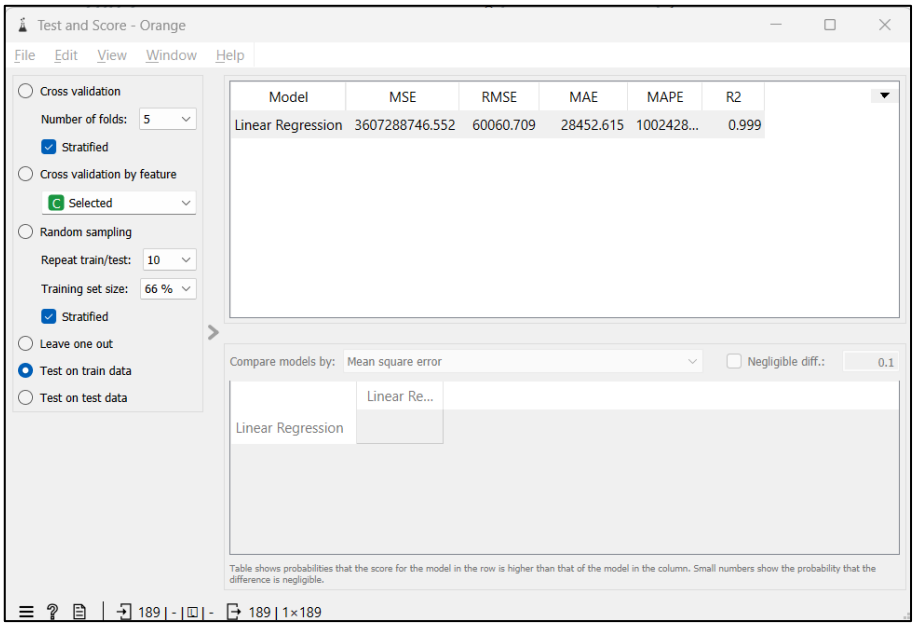


Рис. 7. Оцінка Лінійної регресії

В отриманому результаті моделі лінійної регресії присутні показники, які є метриками для оцінки точності передбачень моделі: MSE, RMSE, MAE та MAPE. Кожна з них аналізує різницю між передбаченими ($\hat{y}_i$) і фактичними значеннями (y_i) залежної змінної:

MSE (Mean Squared Error) – Середньоквадратична помилка. Вимірює середній квадрат відхилення між передбаченнями моделі і фактичними значеннями. Одиницями виміру є квадрат одиниць залежної змінної. Чутливий до великих помилок через піднесення до квадрату. Дає більшу вагу великим відхиленням.

RMSE (Root Mean Squared Error) – Корінь середньоквадратичної помилки. Показує середню різницю між передбаченими та фактичними значеннями, вимірює корінь квадратного середнього відхилення. Одиницями виміру є одиниці залежної змінної. Легше інтерпретується, оскільки має ту ж одиницю виміру, що й залежна змінна. Також чутливий до великих помилок.

MAE (Mean Absolute Error) – Середня абсолютна помилка. Вимірює середнє абсолютне відхилення між передбаченими та фактичними значеннями. Одиницями виміру є одиниці залежної змінної. Менш чутливий до великих помилок у порівнянні з MSE та RMSE. Дає рівну вагу всім помилкам.

MAPE (Mean Absolute Percentage Error) — Середня абсолютна відносна помилка. Вимірює середню абсолютну відсоткову помилку передбачення. Одиниця виміру є відсотки (%). Зручна для порівняння моделей у різних масштабах. Проблематична при малих або нульових значеннях y_i, оскільки може давати дуже великі значення або бути нерозрахованою.

R^2 (коефіцієнт детермінації): показує частку варіації залежної змінної, яка пояснюється незалежними змінними. Значення варіюється від 0 до 1; чим ближче до 1, тим краще модель.

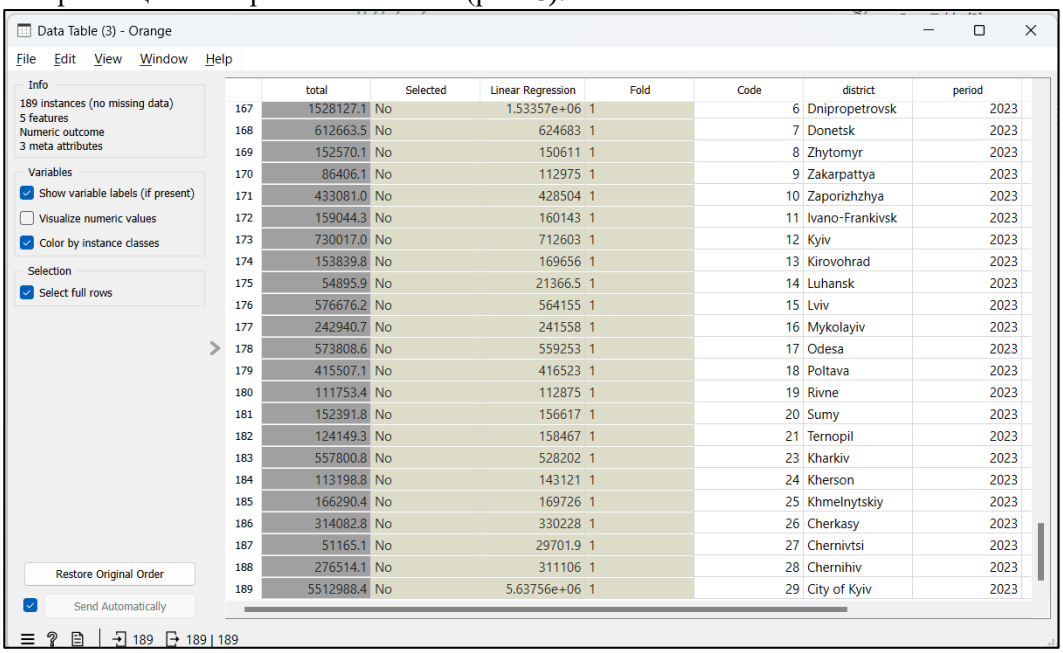
- Значення $R^2 = 1$: модель ідеально пояснює дані (всі точки лежать на лінії регресії).
- Значення $R^2 = 0$: модель не пояснює варіацію даних
- Значення $R^2 < 0$: модель погано підходить до даних

Простий для інтерпретації. Добре підходить для оцінки ступеня відповідності моделі даним.
Перебільшення якості моделі: R^2 може зростати із збільшенням кількості незалежних змінних, навіть якщо ці змінні не покращують модель. Для цього використовується скоригований R^2.
Невідповідність задачам прогнозування: високе значення R^2 не гарантує точних прогнозів (можливий overfitting).

R^2 є важливим інструментом [12] для оцінки пояснювальної сили моделі, але це не єдина метрика. Для повної оцінки моделі варто враховувати інші показники (наприклад, MAE, RMSE, MAPE).

Таким чином, зважаючи на отримані результати $R^2 = 0{,}9$ можна дійти висновку, що модель лінійної регресії чудово справляється з поставленим завданням.

Тепер перевіримо, як розподілені прогнозовані обсяги виробництва товарів у відношенні до істинних. Для цього за допомогою віджета "Data Table" відобразимо таблицю з прогнозованими обсягами виробництва товарів та істинними (рис. 8):



	total	Selected	Linear Regression	Fold	Code	district	period
167	1528127.1	No	1.53357e+06	1	6	Dnipropetrovsk	2023
168	612663.5	No	624683	1	7	Donetsk	2023
169	152570.1	No	150611	1	8	Zhytomyr	2023
170	86406.1	No	112975	1	9	Zakarpattya	2023
171	433081.0	No	428504	1	10	Zaporizhzhya	2023
172	159044.3	No	160143	1	11	Ivano-Frankivsk	2023
173	730017.0	No	712603	1	12	Kyiv	2023
174	153839.8	No	169656	1	13	Kirovohrad	2023
175	54895.9	No	21366.5	1	14	Luhansk	2023
176	576676.2	No	564155	1	15	Lviv	2023
177	242940.7	No	241558	1	16	Mykolayiv	2023
178	573808.6	No	559253	1	17	Odesa	2023
179	415507.1	No	416523	1	18	Poltava	2023
180	111753.4	No	112875	1	19	Rivne	2023
181	152391.8	No	156617	1	20	Sumy	2023
182	124149.3	No	158467	1	21	Ternopil	2023
183	557800.8	No	528202	1	23	Kharkiv	2023
184	113198.8	No	143121	1	24	Kherson	2023
185	166290.4	No	169726	1	25	Khmelnytskiy	2023
186	314082.8	No	330228	1	26	Cherkasy	2023
187	51165.1	No	29701.9	1	27	Chernivtsi	2023
188	276514.1	No	311106	1	28	Chernihiv	2023
189	5512988.4	No	5.63756e+06	1	29	City of Kyiv	2023

Рис. 8. Порівняльна таблиця результатів

За допомогою діаграми розсіювання графічно відобразимо залежність між прогнозованими обсягами виробництва товарів та істинними (рис. 9):

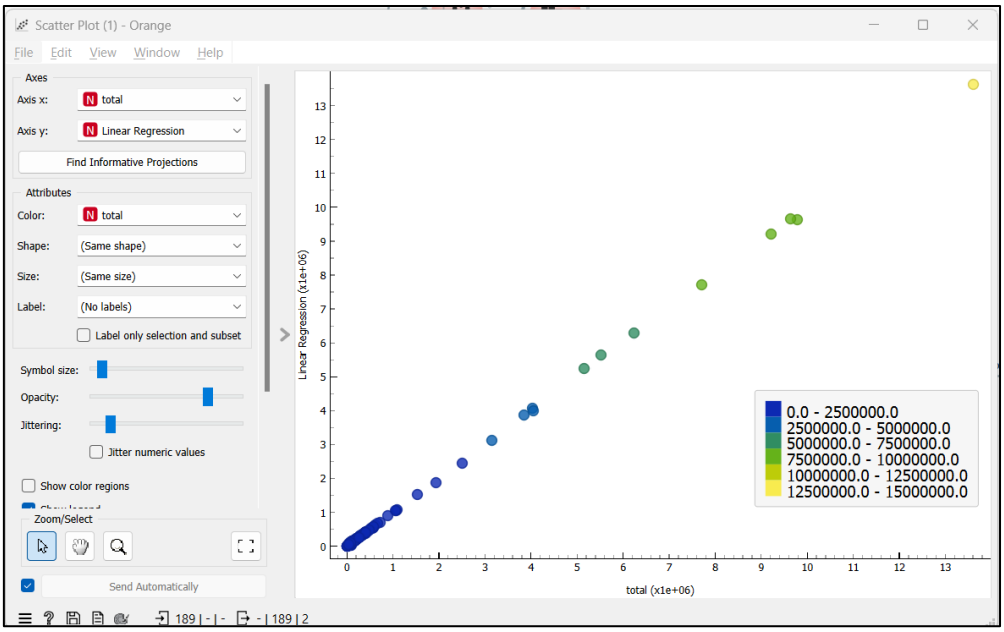


Рис. 9. Діаграма розсіювання

Зважаючи на те, що значення загальних обсягів виробництва та моделі лінійної регресії на діаграмі розсіювання розташовані рівномірно вздовж прямої лінії регресії, можна дійти висновку, що модель працює та справляється з поставленою задачею.

Після побудови, застосування та візуалізації моделі на основі лінійної регресії експеримент продовжується в контексті кластеризації. Підготовка даних [13] є важливим етапом для отримання точного результату. Перед початком застосування нових віджетів перевіряємо, чи коректно відображаються усі дані підприємств за допомогою "Data Table".

За допомогою віджету K-Means та t-SNE виконаємо кластеризацію даних підприємств (рис. 10):

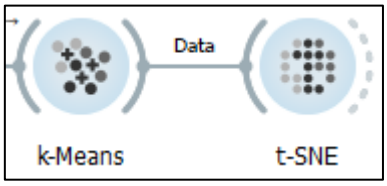


Рис.10. Віджети кластеризації

Метрика Silhouette Scores (рис. 11) оцінює якість кластеризації; зазвичай обирається значення, максимально наближене до 1.

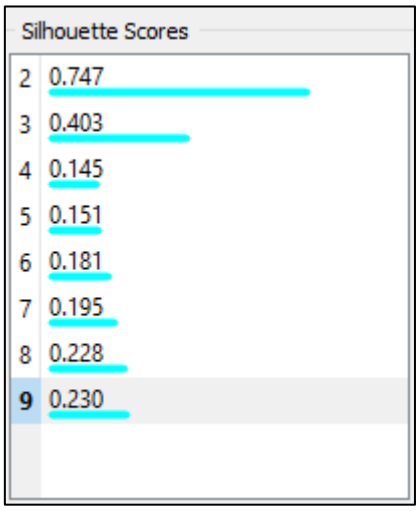


Рис.11. Метрика Silhouette Scores

У випадку з тестованими даними k=2, Silhouette Score = 0.747 на виході програмна система пропонує до отримання два кластери. Така оцінка сформована метрикою оскільки регіон "City of Kyiv" кожного року має набагато більшу частку підприємств, ніж інші регіони. Тому для продовження та кращої візуалізації експерименту обираємо значення 0.403, яке відповідає k=3 (три кластери).

Після завершення роботи алгоритму K-Means проводиться візуалізація отриманих кластерів за допомогою віджету t-SNE (t-Distributed Stochastic Neighbor Embedding) [14], який дозволяє представити дані у двовимірному або тривимірному просторі та допомагає дослідити структури даних, кластери та зв'язки між об'єктами. Результат кластеризації представлений на (рис. 12):

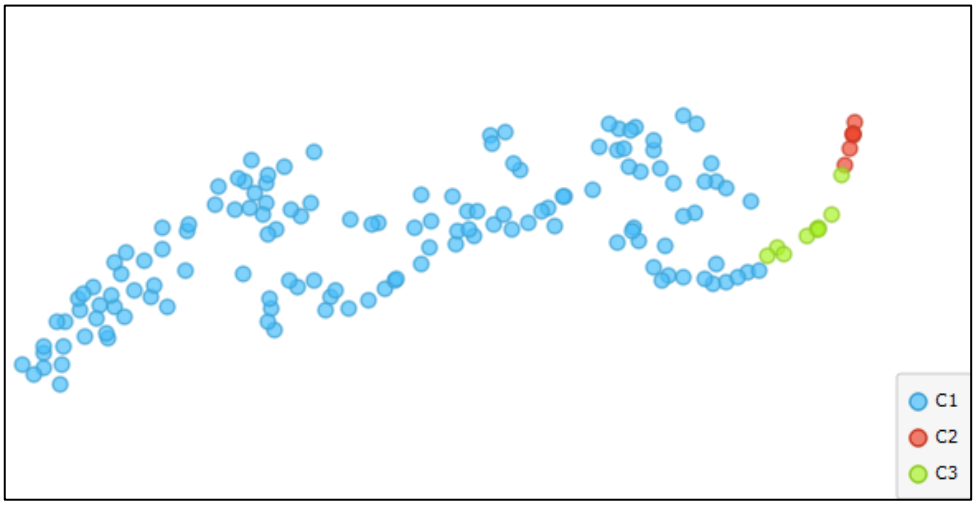


Рис.12. Результати кластеризації

Orange Data Mining дає змогу дослідити кластери C1, C2 та C3 безпосередньо за допомогою внутрішніх інструментів віджета t-SNE. Алгоритм K-Means розподілив регіони по кластерам (рис. 13), а саме:

- City of Kiyv – C2 кластер (як регіон у якому зосереджено найбільше великих та середніх підприємств).
- Dnipro district, Donetsk district – C3 кластер (як регіони у яких зосереджено велику кількість підприємств).
- Other districts – C1 кластер (усі інші регіони країни у яких зосереджено значну кількість підприємств).

Cluster	Silhouette	t-SNE-x	t-SNE-y	Group	district	large enterprises	medium enterprises
C3	0.588143	10.7497	-3.01687	G1	Dnipropetrovsk	353444.0	148530.2
C3	0.583012	13.4582	-0.768659	G1	City of Kyiv	714871.0	926459.0
C3	0.611475	11.0768	-2.82654	G1	Dnipropetrovsk	384632.1	182795.2
C2	0.509628	13.6143	-0.473494	G1	City of Kyiv	963188.0	1147367.8
C3	0.66002	12.0601	-2.63713	G1	Dnipropetrovsk	510565.6	252293.8
C2	0.649382	13.8295	0.0282214	G1	City of Kyiv	1241848.4	1414202.9
C3	0.67563	12.4402	-2.47011	G1	Dnipropetrovsk	607426.5	282202.6
C2	0.698327	13.94	0.438731	G1	City of Kyiv	1478592.1	1774826.9
C3	0.675912	12.5016	-2.43375	G1	Dnipropetrovsk	631962.1	287513.5
C2	0.700147	13.9564	0.508937	G1	City of Kyiv	1541477.3	1870319.8
C3	0.675761	12.4614	-2.47271	G1	Dnipropetrovsk	621738.3	269465.1

rnover: 153 instances, 12 variables

Cluster	Silhouette	t-SNE-x	t-SNE-y	Selected (1)	district	large enterprises	medium enterprises
C1	0.633384	-11.8546	3.36178	No	Ivano-Frankivsk	7544.9	32812.5
C1	0.603394	7.37552	-2.23117	No	Kyiv	155588.5	115444.1
C1	0.631337	-10.9766	3.43811	No	Kirovohrad	7635.5	36096.9
C1	0.631129	-14.2469	2.57139	No	Luhansk	14877.4	12020.4
C1	0.624891	-0.0197928	-1.6847	No	Lviv	60930.9	74911.5
C1	0.630601	-4.64868	-3.38636	No	Mykolayiv	32730.0	37337.9
C1	0.625503	1.34201	-0.556342	No	Odesa	50901.6	96135.3
C1	0.621099	2.32826	1.47409	No	Poltava	89733.0	66325.9
C1	0.633662	-13.1679	3.11239	No	Rivne	8206.7	22897.0
C1	0.63278	-11.9327	2.19557	No	Sumy	14297.1	32092.2

Рис.13. Кластери C1-3

Кластеризація завершена успішно, про що свідчать чітко розподілені кластери, які відповідають вхідним даним. Особливу увагу було приділено очищенню даних. Слід зазначити, що шум та деякі неточності потребували коригування, аби отримати чіткіший результат.

Кінцева модель експерименту включає підготовку та очищення даних, кластеризацію, лінійну регресію, а також візуалізацію отриманих результатів (рис. 14).

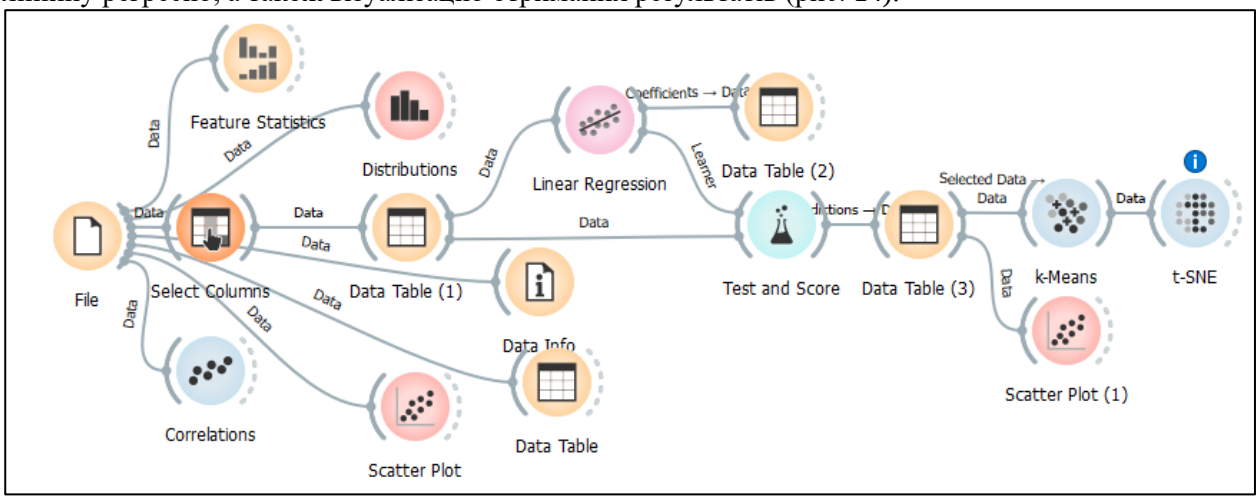


Рис.14. Кінцева модель

Висновки та перспективи подальшого дослідження. У проведеному дослідженні детально проаналізовано можливості використання Orange Data Mining для вирішення задач прогнозування та кластеризації виробничих показників підприємств.

Розроблена модель лінійної регресії продемонструвала високу точність у прогнозуванні обсягів виробництва, що підтверджується значенням коефіцієнта детермінації $R^2 = 0,9$. Це свідчить про те, що модель здатна пояснювати більшу частину варіацій залежної змінної, що робить її ефективним інструментом для управління та планування діяльності підприємств.

Алгоритм K-means дозволив провести кластеризацію підприємств за ключовими характеристиками, такими як продуктивність, витрати та кількість працівників. Визначення трьох кластерів із використанням метрики Silhouette Score (0,403 для k=3) підтвердило адекватність обраної моделі та дозволило сегментувати підприємства для подальшого аналізу.

Проте, для повноти оцінки ефективності використаних методів, доцільно було б провести порівняльний аналіз точності та продуктивності з іншими інструментами для обробки даних, що може стати перспективою для майбутніх досліджень.

Таким чином, отримані результати підтверджують доцільність застосування Orange Data Mining для аналізу виробничих показників підприємств. Застосування лінійної регресії та алгоритму K-means дозволяє вирішувати задачі прогнозування та кластеризації, сприяючи прийняттю обґрунтованих управлінських рішень.

Список бібліографічного опису

1. Saumya Singh, Smriti Srivastava. Review of Clustering Techniques in Control System: Review of Clustering Techniques in Control System, 2020 URL: https://doi.org/10.1016/j.procs.2020.06.032 (дата звернення: 10.11.2024).
2. Richard A. Johnson, Dean W. Wichern. Applied Multivariate Statistical Analysis, 2007 URL: https://archive.org/details/appliedmultivari04edjohn (дата звернення: 12.11.2024).
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning: With Applications in R, 2013 URL: https://bit.ly/4fSMOrF (дата звернення: 15.11.2024).
4. Orange Data Mining. URL: https://orangedatamining.com/ (дата звернення :15.11.2024).
5. Montgomery D. C., Peck E. A., Vining G. G. Introduction to Linear Regression Analysis, 2012 URL: https://archive.org/details/introduction-to-linear-regression-analys (дата звернення: 20.11.2024).
6. Alvin C. Rencher, G. Bruce Schaalje. Linear Models and Statistics, 2008 URL: http://surl.li/atpixs (дата звернення: 20.11.2024)
7. Gourav Kalbalia, Vivek Tambi. Forecasting GDP: A Linear Regression Model, 2016 URL: http://surl.li/lsjzen (дата звернення: 21.11.2024).
8. University of Iowa. K-means Algorithm. 2012 URL: http://surl.li/rbieda (дата звернення: 21.11.2024).
9. Tapas Kanungo, Nathan S. Netanyahu, Angela Y. An Efficient k-Means Clustering Algorithm: Analysis and Implementation, 2002 URL: http://surl.li/wasazf (дата звернення: 21.11.2024).
10. Ananya Sarker, S.M. Shamim, Dr. Md. Shahiduz Zama. Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm, 2018 URL: http://surl.li/iopaca (дата звернення: 23.11.2024).
11. P. McCullagh, J.A. Nedler. Generalized Linear Models, 1998 URL: https://archive.org/details/generalizedlinea0000mccu/ (дата звернення: 23.11.2024).
12. Y. Babich, L. Hlazunova, T. Kalinina, Y. Petrovych. R2 METRIC DYNAMICS FOR K-NEAREST NEIGHBORS REGRESSION MODEL TRAINED ON SERIES OF DIFFERENT SIZES, 2024 URL: https://doi.org/10.23939/ictee2024.02.010 (дата звернення: 24.11.2024).
13. Alberto Amato, Vincenzo Di Lecce. Data preprocessing impact on machine learning algorithm performance, 2023 URL: https://doi.org/10.1515/comp-2022-0278 (дата звернення: 24.11.2024).
14. Sanjeev Arora, Wei Hu Pravesh, K. Kothari. An Analysis of the t-SNE Algorithm for Data Visualization, 2018 URL: http://surl.li/qkytkl (дата звернення: 25.11.2024).
15. Дія. Частка продажу підприємствами роздрібної торгівлі товарів, що вироблені на території України, за товарними групами, 2021 URL: https://data.gov.ua/dataset/4060036b-9868-4cde-b5bc-a6c865757f25. (дата звернення: 22.11.2024).

References

1. Saumya Singh, Smriti Srivastava. Review of Clustering Techniques in Control System: Review of Clustering Techniques in Control System, 2020 URL: https://doi.org/10.1016/j.procs.2020.06.032 (access date: 10.11.2024).
2. Richard A. Johnson, Dean W. Wichern. Applied Multivariate Statistical Analysis, 2007 URL: https://archive.org/details/appliedmultivari04edjohn (access date: 12.11.2024).
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning: With Applications in R, 2013 URL: https://bit.ly/4fSMOrF (access date: 15.11.2024).
4. Orange Data Mining. URL: https://orangedatamining.com/ (access date:15.11.2024).
5. Montgomery D. C., Peck E. A., Vining G. G. Introduction to Linear Regression Analysis, 2012 URL: https://archive.org/details/introduction-to-linear-regression-analys (access date: 20.11.2024).
6. Alvin C. Rencher, G. Bruce Schaalje. Linear Models and Statistics, 2008 URL: http://surl.li/atpixs (access date: 20.11.2024)
7. Gourav Kalbalia, Vivek Tambi. Forecasting GDP: A Linear Regression Model, 2016 URL: http://surl.li/lsjzen (access date: 21.11.2024).

8. University of Iowa. K-means Algorithm. 2012 URL: http://surl.li/rbieda (access date: 21.11.2024).
9. Tapas Kanungo, Nathan S. Netanyahu, Angela Y. An Efficient k-Means Clustering Algorithm: Analysis and Implementation, 2002 URL: http://surl.li/wasazf (access date: 21.11.2024).
10. Ananya Sarker, S.M. Shamim, Dr. Md. Shahiduz Zama. Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm, 2018 URL: http://surl.li/iopaca (access date: 23.11.2024).
11. P. McCullagh, J.A. Nedler. Generalized Linear Models, 1998 URL: https://archive.org/details/generalizedlinea0000mccu/ (access date: 23.11.2024).
12. Y. Babich, L. Hlazunova, T. Kalinina, Y. Petrovych. R2 METRIC DYNAMICS FOR K-NEAREST NEIGHBORS REGRESSION MODEL TRAINED ON SERIES OF DIFFERENT SIZES, 2024 URL: https://doi.org/10.23939/ictee2024.02.010 (access date: 24.11.2024).
13. Alberto Amato, Vincenzo Di Lecce. Data preprocessing impact on machine learning algorithm performance, 2023 URL: https://doi.org/10.1515/comp-2022-0278 (access date: 24.11.2024).
14. Sanjeev Arora, Wei Hu Pravesh, K. Kothari. An Analysis of the t-SNE Algorithm for Data Visualization, 2018 URL: http://surl.li/qkytkl (access date: 25.11.2024).
15. Diia. Share of sales by retailers of goods produced in Ukraine by product group, 2021 URL: https://data.gov.ua/dataset/4060036b-9868-4cde-b5bc-a6c865757f25. (access date: 22.11.2024).