DOI: 10.31319/2709-2879.2024iss1(8).306473pp84-91
УДК 330.3+338.5

Дмитрієва В.А., кандидат історичних наук, доцент, доцент кафедри інформаційних систем і технологій
Дніпровський державний аграрно-економічний університет, Дніпро
ORCID ID: 0000-0002-2410-4504
e-mail: dmytriieva.va@dsau.dp.ua

Dmytriieva Viktoriia, PhD of History, docent, Department of Information Systems and Technologies, Dnipro State Agrarian and Economic University, Dnipro

ЗАДАЧА ПІДГОТОВКИ ДАНИХ ДО ЕКОНОМЕТРИЧНОГО АНАЛІЗУ

THE PROBLEM OF DATA PREPARATION FOR ECONOMETRIC ANALYSIS

Економетричний аналіз ґрунтується на релевантних даних, які мають відповідати вимогам актуальності та репрезентативності. Під час формування інформаційної бази дослідник має вирішувати питання, пов'язані з їх повнотою, часовими рамками аналізу, масштабами досліджуваних об'єктів та процесів, неоднорідністю сукупності, різними рівнями агрегації показників та шкалами їх вимірювання, мультиколінеарністю факторів та іншими проблемами. Лише після ретельного вивчення джерел інформації та усунення цих проблем можна приступати до процесу моделювання.

Подібний апріорний аналіз даних є першим важливим етапом в будь-якому дослідженні, яке претендує на отримання обґрунтованих результатів. Неправильно сформована база даних та наявність в ній недоліків призводить до отримання моделей, які можуть бути ідеальними на перший погляд, але не відповідати дійсності, що, як наслідок, призводить до хибних прогнозів та висновків.

В статті висвітлено проблеми, які постають перед дослідником при формуванні бази даних для аналізу процесів в економіці. Розглянуто основні підходи щодо їх вирішення.

***Ключові слова**: репрезентативність даних, валідність результату, адекватність моделі, точність прогнозу, препроцесинг даних.*

Econometric analysis is based on the relevant data that should respond to the reality, representativeness and certain period or moment of time. Thus, such data must be first carefully revised by investigator with all meticulousness before the start of calculations and modelling. During the data base forming and filling the researcher should consider questions concerning to it completeness, time frame of analysis, scale of studied objects and processes, heterogeneity of the data set, different levels of aggregation of indicators and measurement scales, multicollinearity of factors and other problems. It is worth to begin research only after such a thorough study of the sources of information and elimination of these problems.

Such a priori data analysis is the first important stage in any research that claims to obtain substantiated results. An incorrectly formed database and the presence of the listed shortcomings in it provide obtaining models that may be ideal at first glance, but do not correspond to reality, which as a result lead to the false predictions and conclusions. Scientists pay a lot of attention to the methodology for overcoming of such peculiarities of the data base as deficit or redundancy of information, type or format, optimal size of sample, sizes of training and test data sets, number of features for analysis. Such issues are proposed to be considered with appropriate statistical methods. Researchers can use program languages to enhance data preprocessing. Methods of coding, imputation, extra- or interpolation, normalization are applied to fill blanks and transform attributive features or data measured in various ways into quantitative data in certain format or scale. The next step, after overcoming all the shortcomings is the stage of modeling using methods of mathematical analysis for the reconstruction of causal mechanisms and further forecasting. However, after building the model,

investigator deals with another set of problems that need to be solved. For instance, it is necessary to check the accuracy of the results and how model helps to reconstruct reality.

The article highlights the problems are faced by the researcher when creating a database for representative analysis of processes in the economy. The main approaches to their solution are considered.

Key words: *representativeness of the data, validity of the result, adequacy of the model, accuracy of the forecast, data preprocessing.*

JEL Classification: *C10, C18, C43, C50, C51*

Постановка проблеми. В процесах аналізу інформації левова частина часу, як правило, витрачається на підготовку та перед обробкою даних ще до процесу моделювання. Це обумовлено необхідністю сформувати репрезентативну базу, яка буде відповідати критеріям ненадлишковості, повноти та достатності водночас, а також відображатиме реальні процеси та явища, які відбуваються в економічних системах. При цьому досліднику доводиться мати справу з рядом проблем, які пов'язані з недоліками різних джерел інформації через перешкоди при зборі, фіксації та збереженні даних, специфіку їх представлення,структуру джерел та вміст показників. Сьогодні багато ресурсів, розміщених на офіційних сайтах у відкритому доступі, дозволяють вибрати дані в інтерактивному режимі за розділами теми, по роках, за показниками і, навіть, за адміністративно-географічними утвореннями. Різні формати представлення інформації потребують відповідних засобів обробки або перетворення одного формату в інший. Інколи доводиться шукати та вивчати кілька джерел, з яких можна додавати дані. Проте, не завжди вдається усунути перешкоди доступу до даних або їх відсутності з різних причин, через що доводиться мати справу з вибірками. Крім того, залишається ряд проблем, щодо яких виникає потреба у застосуванні методів та прийомів, які допоможуть подолати недоліки інформаційної бази дослідження.

Аналіз останніх досліджень і публікацій. Формування бази даних для економетричного аналізу є важливою складовою процесу дослідження, оскільки на цьому етапі узгоджуються питання репрезентативності та повноти використовуваної інформації. Неправильно дібрані дані спотворюють результати дослідження та призводять до помилкових висновків та рішень.

Питання вибору сутностей та їх характеристик для дослідження цікавить аналітиків давно [1]. Часто для препроцесингу даних використовують спеціальні програмні засоби та алгоритми. Водночас, розробці методології подолання недоліків в інформаційній базі приділено чимало уваги вчених, як вітчизняних так і закордонних. Наприклад, актуальною є проблема пропусків в даних, з чим мають справу абсолютно всі науковці. Вченими пропонуються різні способи для їх заповнення[2; 3], в тому числі метод імпутації. Іншими словами, заповнення пропущених даних відбувається за одним із розроблених алгоритмів, наприклад, з допомогою інтерполяції, методом найближчого сусіда чи іншими спеціальними методами.

Надлишковість та повторюваність інформації часто заважає розв'язувати задачі аналізу, особливо, коли мова йде про великі масиви даних. Розробці підходів для їх обробки присвячено роботу [4], в якій автори підіймають питання, що стосуються швидкості опрацювання інформації в різних економічних сферах. Зокрема, розглядаються алгоритми, які дозволяють виявляти та виключати з аналізу неважливу або повторювану інформацію з допомогою методу сингулярної декомпозиції.

На противагу до думки, представленої вище,з точки зору інших вчених, надлишкова інформація, навпаки, може бути джерелом для відновлення даних, про що йдеться в роботі [5], депріділено увагу резервуванню таких даних для забезпечення ефективної роботи комунікаційних мереж. Дослідники вважають, що надлишкова інформація навпаки допомагає забезпечити надійність інформаційних систем. При цьому, автори розрізняють кілька видів резервування, серед яких визначено структурне, часове, інформаційне, функціональне та навантажувальне. Для оцінки якості резервування передбачається розробка спеціальних моделей.

Враховуючи той факт, що дані сьогодні представлено не лише в числах та тексті, а й в інших форматах, наприклад, у вигляді піктограм чи ідіограм, доводиться шукати спеціальні методи для їх перед та пост обробки в тому числі. Наприклад, це важливо в маркетинговій сфері, де з допомогою методів сентиментного аналізу вивчають емоційні відгуки клієнтів [6]. В цьому випадку доводиться мати справу зі структурованими та неструктурованими даними, які теж мають підлягати належній формалізації та опрацюванню.

Проблему нормалізації даних досліджено в статті [7], де подано способи приведення даних до єдиної безрозмірної шкали та представлено результати оцінки якості нормалізації. Нормалізація є одним із способів, який допомагає усунути проблему, що проявляється в базі з даними, значення яких вимірюються за різними шкалами.

Приклади попередньої обробки даних та підготовки їх для подальшого моделювання актуалізовано в дослідженні [8], де розглянуто способи подоланням проблеми пропущених даних, незбалансованості класів, перетворення категоріальних змінних у бінарні, нормалізації числових значень, зменшення розмірності. Після застосування алгоритмів машинного навчання, в яких наглядно продемонстровано етапи аналізу, автор акцентує увагу на необхідності проведення попереднього опрацювання даних, що допомагає визначати складність поставлених задач та звузити коло методів, якими їх можна вирішити.

Не менш важливо виконувати очищення даних від впливу випадкових чинників. З даною метою можна виділити структурну компоненту, яка відображає основну тенденцію розвитку, як розглянуто в роботі автора [9]. Після цього, за детермінованими рядами даних реконструювання зв’язків між ознаками дозволить отримати більш чіткий портрет поведінки показників.

Вивчення наукових робіт з аналізу даних підтверджує висновок про необхідність їх препроцесингу. Попередня обробка інформації допомагає вирішити ряд проблем, які можуть спричиняти отримання хибних результатів при моделюванні.

Формулювання цілей статті. Метою даної статті є актуалізація питань формування репрезентативного набору даних для проведення економетричного аналізу.

Виклад основного матеріалу дослідження. Вивчення динаміки та реконструкція причинно-наслідкових механізмів економічних процесів потребують використання методів аналізу на основі фактичних даних, отриманих шляхом спостереження, простої фіксації показників або експериментального вимірювання. При цьому перед дослідником постає ряд питань, які можуть стосуватися того, якою за обсягом має бути інформаційна база дослідження, які об’єктипотрібно охопити увагою, який проміжок часу необхідно проаналізувати, або які показники допоможуть розкрити суть проблеми. Звичною практикою є формування бази даних на основі кількох перевірених джерел інформації, кожне з яких може мати свої переваги і недоліки, наприклад, відмінні структуру і формат представлення даних, розпорошеність по окремих файлах або сховищах, подання інформації у занадто деталізованому або агрегованому вигляді. Однак, навіть після остаточного збору необхідних даних виникає потреба у їх апріорній обробці до процесу економетричного моделювання. Розглянемо детальніше задачу підготовки даних, яка включає вирішення ряду проблем.

Проблема браку даних виникає у випадку, коли в ході вивчення об’єкта дослідження на основі проаналізованої інформації виникають нові питання та потреба у перевірці нових гіпотез, для яких слід додатково отримати значення інших характеристик або факторів. Окремі джерела не містять потрібних даних, або вони представлені не в абсолютному вираженні, а лише у відносному, наприклад, лише у вигляді ланцюгових індексів, або лише у вигляді частки від цілого, при чому, інформація про ціле відсутня. Інколи окремі недоліки вдається подолати додатковою обробкою інформації існуючих джерел, наприклад,застосувати методи інтерполяції чи екстраполяції,тобто теоретичного наближення до реальних даних математичними методами. За відсутності значень для окремих показників доводиться приймати рішення про виключення з бази даних цих показників в цілому.

Проблема надлишковості даних. Надлишковість проявляє себе у повторюваності певного контенту, наявності даних, які не несуть змістовного навантаження, не надають досліднику нових знань чи відомостей про об’єкт вивчення, або можуть бути несуттєвими при

проведенні аналізу. Такі джерела є зачумленими зайвою інформацією, від якої потрібно позбавлятися. В такому випадку доводиться застосовувати фільтрацію та очистку даних від шуму, що сьогодні можна робити програмним шляхом. А наприклад, згідно з технологіями адміністрування баз даних, попередньо проводять нормалізацію таблиць, виділяють окремі тематичні структури, застосовують кодування, створюють прості та множинні ключі для зв'язування наборів даних, а також використовують методи групування, зрізів, ротації та злиття.

Проблема сукупності, непідпорядкованої закону нормального розподілу. За численними дослідженнями, проведеними Р. Фішером, В. Госсетом (Стьюдентом) та іншими вченими, якщо з генеральної сукупності, розподіл якої підпорядковується нормальному закону, вилучити вибірку, то розподіл середніх вибіркових буде наближатись до нормального за числа точок більше п'яти. Водночас, якщо вилучити вибірку з генеральної сукупності, яка не підпорядковується нормальному закону, то розподіл середніх вибіркових буде наближатись до нормального за числа об'єктів більше двадцяти п'яти.

До нормального закону розподілу прив'язано більшість методів статистичного аналізу. При цьому, на початковому етапі роботи з джерелом інформації, дослідник апріорі не завжди знає з якою сукупністю він має справу і якого обсягу вибірку йому потрібно сформувати для проведення розрахунків і отримання валідного результату, який можна буде поширити на генеральну сукупність. Напрошується висновок, що, чим більше обсяг вибірки, який буде використано для аналізу даних, тим більш репрезентативними будуть результати, які краще представлятимуть ситуацію по генеральній сукупності. Проте, обсягом вибірки питання не обмежується, оскільки сукупність може бути структурована, а це означає, що набір даних теж має включати структуровані елементи, які слід вилучати групами або поодинці за певними правилами. Отже, для усунення впливу ненормальності на результати дослідження потрібно вирішити задачу формування вибірки, оптимальної за структурою та обсягом.

Проблема не оптимальності типу та обсягу набору даних. Вивчення досліджуваної сукупності та її структури це ключ до розуміння підходу, який потрібно застосовувати при формуванні вибірки. Зокрема, на даному етапі дослідник вирішує, за яким принципом необхідно включати у вибірку дані: за стратами, тобто шарами, чи включати об'єкти групами, які в сукупності організовані за певною подібністю характеристик, або використати по елементний підхід з виключенням можливості потрапляння у вибірку одних і тих самих сутностей, або, навпаки, забезпечити повторність появи низькочастотних елементів у кінцевому наборі даних. При цьому слід розв'язати проблему релевантної кількості досліджуваних об'єктів з дотриманням правил пропорційності та рівної ймовірності їх потрапляння в поле дослідження для проведення неупередженого аналізу та отримання обґрунтованих результатів. У випадку, якщо сукупність є кількісно неоднорідною та в ній мають місце ексцеси, тобто трапляються об'єкти, характеристики яких значною мірою перевищують величину подібних характеристик у переважної частини сутностей, вдаються до розділеного вивчення ексцесів та решти об'єктів. Після чого наступним кроком має бути аналіз рівня агрегації даних.

Проблема зайвої агрегації або деталізації даних. В багатьох офіційних статистичних джерелах дані подано переважно у зведеному вигляді. Наприклад, часто можна побачити рентабельність об'єктів виробництва продукції за видами економічної діяльності чи загалом по галузях, але не рентабельність по кожному окремому підприємству чи закладу. Динаміку кількості тварин та обсягів виробництва продукції тваринництва представляють загалом по Україні, але не за окремими фермами. Внесення різних видів добрив для підвищення врожайності відображають загалом по країні або по областях, але без прив'язки до господарств чи врожайності відповідних видів вирощуваних культур на удобрених землях. Середньомісячні обсяги опадів та середньомісячні температури так само показують загалом по країні, але не за окремими землями, на яких вирощують конкретні види культур та ін. Це позбавляє вченого можливості проводити адекватний аналіз для реконструкції причинно-наслідкових механізмів. Надлишкова агрегація змушує до використання тих даних, які є, що не наближає до досягнення мети дослідження. В противному випадку, найбільш завзяті дослідники витрачають купу часу на отримання інформації з кількох джерел, які можуть бути не верифікованими, що в свою чергу може призвести до отримання невалідного результату.

Якщо зайву деталізацію можна усунути зведенням та отримати по даних, у разі потреби, загальні підсумки, то при надлишковій агрегації без додаткових джерел значення показника по об'єктах розкрити не вдається. Інша проблема полягає у представленні відомостей по об'єктах за показниками, які не несуть інформаційного навантаження без попереднього опрацювання. Наприклад, таблиця ланцюгових індексів ВРП для областей за рік. Жодної корисної інформації, крім короткотермінового річного зростання чи зменшення показника, дослідник не отримає, доки не використає додаткові матеріали, які стосуються попередньої поведінки показника або інших характеристик, від яких можна отримати бажане, або, як альтернатива, створити синтетичні параметри на основі комбінації існуючих.

Проблема часового виміру. Питання тривалості періоду розвитку об'єкта, який має бути охоплений аналізом, займає важливе місце. Поясненням цьому є необхідність побудови адекватного прогнозу на основі динаміки, яка може бути коротко-, середньо- чи довготермінова. Довжина періоду диктує свої умови. Наприклад, короткий термін за 5-6 років, не відобразить тенденції розвитку процесу. Можна зробити висновки про прогрес на основі аналізу даних по 5-6 точках, однак, період за 20-25 років може показати, що в масі значень це був спад показника з незначним підйомом, оскільки тривалий період містить інформацію про більш високі значення показника. В такому разі, зроблений на основі аналізу 5-6 річної тенденції прогноз виявиться хибним, оскільки він не враховує решти динаміки. Крім того, часто короткотермінові тенденції можуть проявляти себе лінійно, але, водночас, більш тривалий період, частиною якого є такі тенденції, може свідчити про нелінійність розвитку. Тоді побудований за лінійною функцією прогноз не буде адекватно відображати зазубреність або хвильовий характер процесу. Це означає, що під час дослідження потрібно чітко визначати часові межі і мету, якої потрібно досягти, та виходити з того, що короткотривала тенденція є лише однією частиною довготривалого процесу.

Проблема взаємозалежності факторних ознак. Фахівцям з аналізу даних доводиться ретельно вивчати чинники впливу на вихідну змінну. При цьому, важливо ідентифікувати та елімінувати фактори, які є колінеарними, тобто кореляційно зв'язаними між собою, або є похідними один від одного. Колінеарність таких ознак може зумовлювати зменшення їх пояснювальної сили, якості побудованої моделі та призвести до некоректного формулювання висновків. При виявленні колінеарності коефіцієнт детермінації може бути досить високим, але при цьому отримані параметри моделі можуть бути незначущими та ненадійними. Якщо при моделюванні динаміки процесів та прогнозу зав'язаність факторів не суттєво впливає на отримані оцінки, то при вирішенні задачі виявлення сили впливу на результативну змінну постає проблема отримання хибних результатів. Виходом із ситуації є проведення апріорного аналізу факторів на наявність між ними колінеарності, і, у випадку підтвердження існування цієї властивості,застосування процедури вилучення ознак, які є причиною аномалії,проведення більш ретельного добору та включення в модель інших характеристик досліджуваних об'єктів.

Проблема вимірності величин за різними шкалами. При побудові інформаційної бази, яку буде використано для моделювання, звертають увагу на шкали вимірювання показників. В такій базі можуть бути значення кількісні чи атрибутивні, вимірюватись в грошових, вагових величинах, поштучно чи за рангами. В такому випадку для атрибутивних ознак виконують цифрове кодування з визначенням унікальних значень. Для кількісних величин можна застосувати нормалізацію статистичними методами.

Проблема розмірності досліджуваних об'єктів. В аналізі даних, особливо при порівнянні економічного розвитку країн, необхідно уникати впливу розміру самих об'єктів спостереження. Зокрема, при порівнянні країн, які суттєво відрізняються між собою за чисельністю населення, географічними розмірами, кількістю підприємств та установ, в результатах аналізу проявляється ефект Гершенкрона, згідно з яким, в найбільших країнах будуть найбільші економічні показники (наприклад, валового внутрішнього продукту або національного доходу), а в найменших країнах будуть найменші значення індикаторів розвитку економіки. При чому, в даному випадку, найвищі значення ще не означають «найкращі», а найменші ще не означають «найгірші» країни за економічним рейтингом. Для відображення реального стану справ потрібно показники перевести у відносні величини. Наприклад,

провести обчислення ВВП або ВНД у розрахунку на одиницю населення або на одиницю економічно-зайнятого населення, яке здатне створювати додаткову вартість. Альтернативним виходом із ситуації є створення синтетичних індикаторів або інтегральних показників, які є комбінацією відносних величин.

Наступним після подолання всіх недоліків зібраних даних технологіями їх перед обробки та кодування є етап моделювання з використанням методів математичного аналізу для реконструкції причинно-наслідкових механізмів та подальшого прогнозу. Проте, після побудови моделі перед дослідником постає інший комплекс задач, які потребують розв'язку. Цього разу необхідно перевірити валідність результатів та наскільки точно отримана модель допомагає реконструювати реальність.

Проблема формування тренувального і тестового наборів даних. Модель потрібна для прогнозу майбутніх значень вихідного показника або визначення приналежності сутності до тієї чи іншої категорії об'єктів. Для оцінки точності побудованої моделі перевіряють рівень відхилень між реальними значеннями вихідного показника та теоретичними значеннями, розрахованими з використанням моделі. Для цього існує ряд методів, що оцінюють наближеність результату до фактичних даних. У підсумку, модель, яка дозволяє отримати найменші відхилення, вважають адекватною. Однак, навіть у випадку, коли модель дуже добре себе показала на вхідних даних, може виникнути питання іншого плану: чи буде отримана модель так само добре працювати на інших значеннях досліджуваних показників та об'єктів, для яких її не застосовували?

Для зведення до мінімуму ризиків побудови неадекватної моделі потрібно вирішити питання добору факторів, а також поділу вхідних даних на вибірки, одна з яких буде використовуватись для побудови моделі (тобто, для тренування), а друга – для перевірки її валідності (тобто, для тестування). Для пошуку оптимального поділу масиву даних на тренувальну та тестову вибірки потрібно неодноразово пройти етапи побудови моделі та її перевірки на різних незалежним і випадковим чином сформованих наборах даних досліджуваної сукупності об'єктів. Усереднення результатів допоможе уникнути таких проблем моделі, як перетренованість (over fitting, коли модель є занадто складною, чи добре проявляє себе на тренувальному наборі, але показує значні відхилення на тестових даних) або недотренованість (under fitting, коли модель занадто проста і погано працює на обох наборах даних: як тренувальному, так і тестовому). Слід зазначити, що крім розмірів тренувальної та тестової вибірок свій вплив на якість моделі чинять відібрані фактори, їх якість та кількість, що може ускладнити або, навпаки, занадто спростити модель.

Проблема незначущості параметрів моделі проявляється через використання в моделі факторів, які не чинять суттєвого впливу на вихідний результат. Після перевірки гіпотези щодо значущості параметрів дослідник знову аналізує вхідні дані та у разі потреби залучає додаткові джерела інформації для включення інших чинників.

Проблема впливу нелінійності на точність прогнозу. Економічний розвиток характеризується нелінійністю, яка може бути описана динамікою плавною (з поступовим зниженням або зростанням), стрибкоподібною (з раптовими підйомами або спадами) або зазубреною (коли відбувається послідовна зміна злетів та падінь показників). Варто зважувати на те, що побудований на фактичних даних нелінійний тренд з використанням степеневих функцій може давати неправильний прогнозна більш тривалий період вперед, оскільки степеневий характер функції спричиняє розвиток стрибкової тенденції зміни вихідного показника. В результаті, побудовані на такому прогнозі стратегії можуть виявитися хибними. Слід з обережністю робити передбачення на подібних моделях на довготривалий період уперед, і для надійності обмежуватися прогнозом на найближчий термін.

Висновки. Якісно підготовлені для аналізу дані обумовлюють побудову моделі, що має з певною точністю стати проекцією реальності. Описані проблеми повинні бути усуненими до основного аналізу для отримання валідного результату прогнозування або класифікації сутностей. Таким чином, задача підготовки даних до економетричного дослідження включає узгодження питань щодо їх повноти і незашумленості, обсягу та структури вибірки, рівня деталізації та нормалізації, обґрунтованого періоду часу, усунення колінеарності факторів,

оптимізації розмірності та ін. Від того, наскільки якісно сформована база даних для аналізу залежить акуратність результатів та значущість параметрів дібраної моделі. Застосування засобів програмування щодо багатьох процесів спрощує вирішення описаних вище задач. Відповіді на більшість питань щодо цього допомагають отримати алгоритми машинного навчання.

Список використаної літератури

1. Sotiris Kotsiantis, & Kanellopoulos, Dimitris & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1, 111-117. URL: https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_Learning (дата звернення: 08.04.2024)
2. Cho E, Chang T-W, Hwang G. (2022). Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process. *Electronics*; 11(3):477. https://doi.org/10.3390/electronics11030477 (дата звернення: 01.03.2024)
3. Ковтун Н. В., Фаталієва А.-Н. Я. (2020). Програмна реалізація відновлення пропущених даних: порівняльний аналіз. *Статистика України*, 4, С. 12–20. DOI:10.31767/su.4(91)2020.04.02. (дата звернення: 20.03.2024).
4. Гордійчук-Бублівська О.В., Бешлей М.І., Кирик М.І., Климаш М.М. (2021). Підвищення ефективності оброблення великих обсягів інформації з використанням методу розподіленого аналізу даних. *Телекомунікаційні та інформаційні технології*, 2 (71), С. 15-23. URL: https://doi.org/10.31673/2412-4338.2021.021523 (дата звернення: 02.04.2024).
5. Кононова І.В., Дубина В.О. (2023). Комплексне використання надлишковості для підвищення надійності комунікаційного обладнання. *Вчені записки ТНУ імені В.І. Вернадського: Технічні науки*, 34 (73), 5, С. 40-45. DOI https://doi.org/10.32782/2663-5941/2023.5/08. (дата звернення: 06.04.2024).
6. Duong, HT., Nguyen-Thi, TA. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Comput Soc Netw* 8, 1. URL: https://doi.org/10.1186/s40649-020-00080-x (дата звернення: 08.04.2024).
7. Ясінська-Дамрі Л.М., Дурняк Б.В. (2016). Модель оцінки якості нормалізації даних на основі застосування критеріїв якості класифікації об’єктів. *Наукові записки*, 1(52), С. 35-44. URL: http://pvs.uad.lviv.ua/static/media/1-81/6.pdf. DOI: 10.32403/0554-4866-2021-1-81-35-44 (дата звернення: 20.03.2024).
8. Піскун О.В. (2019). Застосування методів машинного навчання для побудови моделі рішення задачі класифікації. *Вісник Черкаського національного університету імені Богдана Хмельницького. Серія «Прикладна математика. Інформатика»*, 2019, 1, 42-53. DOI 10.31651/2076-5886-2019-1-42-53. URL: https://eprints.cdu.edu.ua/2585/1/3696-8968-1-SM.pdf (дата звернення: 06.04.2024).
9. Dmytriieva, V., & Sviatets, Y. (2023). Agricultural business in independent Ukraine: thirty-year dynamics of the reorganization process. *Agricultural and Resource Economics: International Scientific E-Journal*, *9*(2), 136-162. URL: https://doi.org/10.51599/are.2023.09.02.06 (дата звернення: 08.04.2024).

References

1. Kotsiantis, Sotiris & Kanellopoulos, Dimitris & Pintelas, P. (2006) Data Preprocessing for Supervised Learning. *International Journal of Computer Science.* No. 1, pp. 111-117. Available at: https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_Learning (accessed 8 April 2024).
2. Cho E, Chang T-W, Hwang G. (2022) Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process. *Electronics*; no. 11(3):477. Available at: https://doi.org/10.3390/electronics11030477 (accessed 1 March 2024).

3. Kovtun N. V., & Fataliieva A.-N. Ya. (2020) Prohramna realizatsiia vidnovlennia propushchenykh danykh: porivnialnyi analiz [Software Implementation of Missing Data Recovery: Comparative Analysis]. *Statystyka Ukrainy – Statistics of Ukraine*, 4, pp. 12–20. Doi: 10.31767/su.4(91)2020.04.02. (accessed 20 March 2024).
4. Hordiichuk-Bublivska O.V., Beshlei M.I., Kyryk M.I., Klymash M.M. (2021) Pidvyshchennia efektyvnosti obroblennia velykykh obsiahiv informatsii z vykorystanniam metodu rozpodilenoho analizu danykh [Increasing the efficiency of processing large volumes of information using the method of distributed data analysis]. *Telekomunikatsiini ta informatsiini tekhnolohii*, no. 2 (71), pp. 15-23. Available at: https://doi.org/10.31673/2412-4338.2021.021523. (accessed 2 April 2024).
5. Kononova I.V., Dubyna V.O. (2023) Integrated use of redundancy to improve the reliability of communications equipment. *Academic notes of TNU named after V.I. Vernadskyi: Technical Sciences,* no. 34 (73), pp. 40-45. DOI: https://doi.org/10.32782/2663-5941/2023.5/08 (accessed 6 April 2024).
6. Duong H.T., Nguyen-Thi T.A. (2021) A review: preprocessing techniques and data augmentation for sentiment analysis. *Comput Soc Netw* 8, 1. Available at: https://doi.org/10.1186/s40649-020-00080-x (accessed 8 April 2024).
7. Yasinska-Damri L.M., Durnyak B.V. (2016) Model of the data normalization quality evaluation on the basis of application of the objects classification quality criteria, *Scientific papers*, no. 1(52), pp. 35-44. http://pvs.uad.lviv.ua/static/media/1-81/6.pdf. DOI: 10.32403/0554-4866-2021-1-81-35-44 (accessed 20 March 2024).
8. Piskun O. (2019) Classification model building using machine learning methods. *Bulletin of the Cherkasy National University named after Bohdan Khmelnytskyi*, Series "Applied mathematics, no. 1, pp. 42-53. DOI 10.31651/2076-5886-2019-1-42-53 (accessed 6 April 2024).
9. Dmytriieva, V., & Sviatets, Y. (2023). Agricultural business in independent Ukraine: thirty-year dynamics of the reorganization process. *Agricultural and Resource Economics: International Scientific E-Journal*, no. *9*(2), pp. 136-162. Available at: https://doi.org/10.51599/are.2023.09.02.06 (accessed 8 April 2024).

Надійшла до редколегії 12.04.2024