Issue 11, 2022

https://doi.org/10.23939/sisn2022.11.023

УДК 004.415

ПРОЦЕСИ ТА СКЛАДОВІ ЕЛЕМЕНТИ АНАЛІЗУ ВЕЛИКИХ ДАНИХ У СИСТЕМАХ ДИСТАНЦІЙНОГО НАВЧАННЯ

Андрій Пришляк[1], Наталія Кунанець[2], Володимир Пасічник[3]

[1] Національний університет "Львівська політехніка", кафедра інформаційних систем та мереж,
[1] e-mail: andrii.a.pryshliak@lpnu.ua; ORCID: 0000-0003-1681-5178,
[2] e-mail: nek.lviv@gmail.com, ORCID: 0000-0003-3007-2462,
[3] e-mail: vpasichnyk@gmail.com, ORCID: 0000-0002-5231-6395

Проаналізовано вплив пандемії на освітні процеси в Україні. Розглянуто проблеми, які спостерігаються під час дистанційного навчання, позитивні та негативні фактори онлайн-освіти. Висвітлено чинники, які можуть призвести до конфліктних ситуацій в освітньому процесі та ускладнювати збирання й аналіз інформації. Запропоновано використовувати методи машинного навчання для аналізу великих даних у системах дистанційного навчання. Розглянуто метод аналізу головних компонент для зменшення розмірності обсягу вибірки та описано основні кроки, які потрібно виконати задля спрощення. Можливість аналізу забезпечується коректним функціонуванням системи дистанційного навчання, регламентованої ЗВО, взаємодією із усіма учасниками навчального процесу, а також своєчасним виконанням обов'язків, покладених на них.

Ключові слова: великі дані; системи дистанційного навчання; онлайн-освіта; Data mining; Machine Learning; метод PCA; розмірність даних.

Вступ

Пройшло немало часу, відколи почалась епідемія і принесла багато змін у різні сфери суспільного життя, особливо освітні системи. Здобувачі освіти здебільшого отримують освітні послуги дистанційно, а деякі за змішаною формою, що поставило нові цілі у формуванні освітніх процесів та їх аналізі. Змінилися вимоги до ведення справ, компетентності учасників освітнього процесу, зокрема викладачів та вчителів, спостереження реальної картини навчального процесу.

Дослідження проблемних елементів освітнього процесу ніколи не втрачає актуальності, оскільки постійно стикається із новими викликами, які часто неможливо спрогнозувати та підготуватись заздалегідь. Сьогодні потрібно шукати нові підходи до аналізу освітнього процесу на зовсім іншому рівні. Зокрема, потрібно надавати пріоритет інтелектуальним системам для збирання актуальної інформації та подальшого її опрацювання. Опрацювання ж передбачає різні етапи, залежно від контексту застосування.

Коли ми говоримо про збирання інформації із застосуванням штучного інтелекту, то зрозуміло, що йдеться про засоби Data mining та Machine Learning. Ці підходи особливо корисні в разі використання концепту великих даних, їх застосовують для створення наборів даних, їх опрацювання, аналізу та навчання.

Постановка проблеми

З одного боку, сам процес надання освітніх послуг став простішим, за рахунок меншого фізичного навантаження, а з іншого – потребує більше контролю за виконанням та дотриманням

обов'язків усіх учасників за допомогою систем дистанційного навчання. Спостерігаючи за освітнім процесом та аналізуючи реальну картину, ми зможемо виявляти та вирішувати проблеми, пов'язані із ними.

На початку карантину заклади вищої освіти (ЗВО) не були повністю підготовлені до онлайн-освіти, використовували лише окремі її елементи. А методи онлайн-навчання потребували розвитку та чіткого регулювання. У травні–червні 2020 р. більшість ЗВО відзначали, що до карантину їхні викладачі використовували певні елементи онлайн-освіти та багато в чому покладалися на систему Moodle. Крім того, онлайн-освіта розвивалася вже до пандемії, особливо у зимовий час, і потребувала більшої уваги, передусім щодо наповнення матеріальної бази, і була відповіддю на вимоги часу чи обставин [1].

Аналіз останніх досліджень

Дистанційна форма навчання призвела до виявлення безлічі проблем, що так чи інакше стосувалися різних аспектів суспільного життя. Серед таких проблем – брак істотного досвіду викладання за допомогою онлайн-систем у деяких викладачів та частка тих, які не хочуть розвивати навички у цьому напрямі. Крім того, діти та підлітки більше часу проводять в ізоляції, менше часу взаємодіють із друзями, колегами та перебувають на вулиці, унаслідок чого відчувають погіршення психологічного стану. Поряд з цим, під час дистанційного навчання мотивація здобувачів освіти до навчального процесу порівняно низька.

В українській освіті все ще дається взнаки неготовність до різкого переходу на дистанційне навчання: слабка матеріально-технічна база, недостатнє освоєння практичних навичок роботи у педагогічного складу, а також проблеми із підготовкою відповідних методичних рекомендацій про організацію освітнього процесу. Поза тим, бракує даних, аби критично оцінити стан готовності освітньої системи до дистанційного формату надання освітніх послуг та їх якість.

Забезпечення якості навчального процесу потребує постійної та стійкої взаємодії між його учасниками. Відповідно до деяких досліджень, доступність навчання прямо впливає на його проходження. Свою лепту вносить також і постійне навантаження на системи дистанційного навчання, що не може не позначитись на їх функціонуванні. Компоненти таких систем зазнають впливу поступового зростання навантаження і зовнішніх чинників [2, 3].

Формулювання цілі статті

Аналізуючи пропоновані на ринку системи дистанційного навчання, варто визначити їхні основні переваги, аспекти, які потрібно досліджувати і ті, якими можна знехтувати, а також вибрати, що саме буде основною метою такого аналізу. В результаті можемо отримати дані про реальний стан речей, оцінити потенціал здобувачів освіти та відстежити проблемні моменти онлайн-освіти.

Особливої уваги потребує проблема браку даних та їх актуальності, моніторингу даних на різних етапах навчального процесу. Умовно розподілимо їх на окремі складові:

- Активність учасників під час навчального процесу. Цей показник істотно впливає на потенційну успішність здобувача освіти. Без активності на різних заняттях взаємодія між учасниками навчального процесу порушується, а отже, можуть виникати непорозуміння між викладачем та студентом. Також під час пандемії є ризик, що здобувачі не вважатимуть за потрібне попередити про хворобу, що згодом також впливає на оцінку їхніх знань.
- Похідною проблемою від активності є дотримання дедлайнів. Зазвичай 20–65 % студентів мають проблеми із оцінками, які пов'язані з кінцевими термінами здавання робіт, серед них показник відсотка таких робіт може сягати 100 %. Також можна зробити припущення про прямий зв'язок між недотриманням дедлайнів та рівнем знань студентів.
- Низька мотивація до виконання завдань. Цей чинник має здебільшого психологічне підґрунтя й істотно впливає на якість виконаних завдань та дотримання дедлайнів, а також студентський соціум загалом [4].

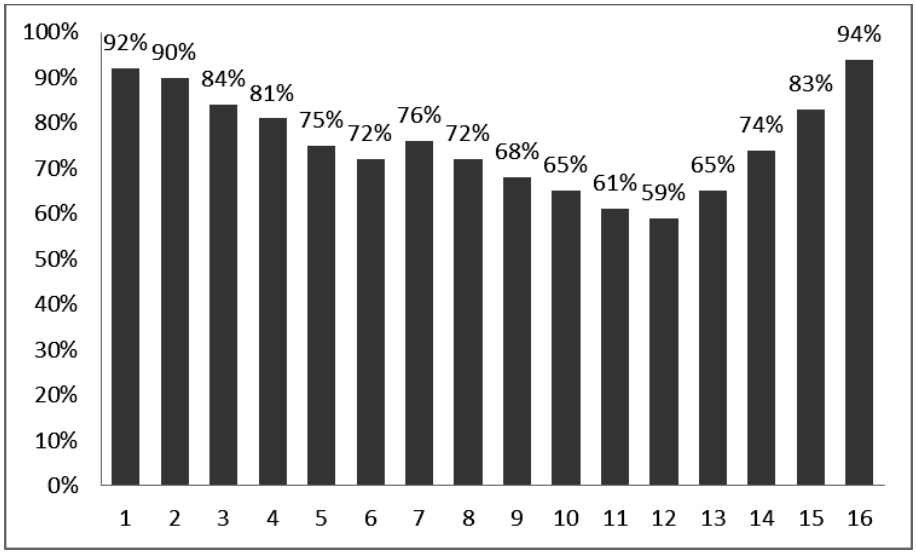

Рис. 1. Середня оцінка активності студентів упродовж кожного тижня навчання

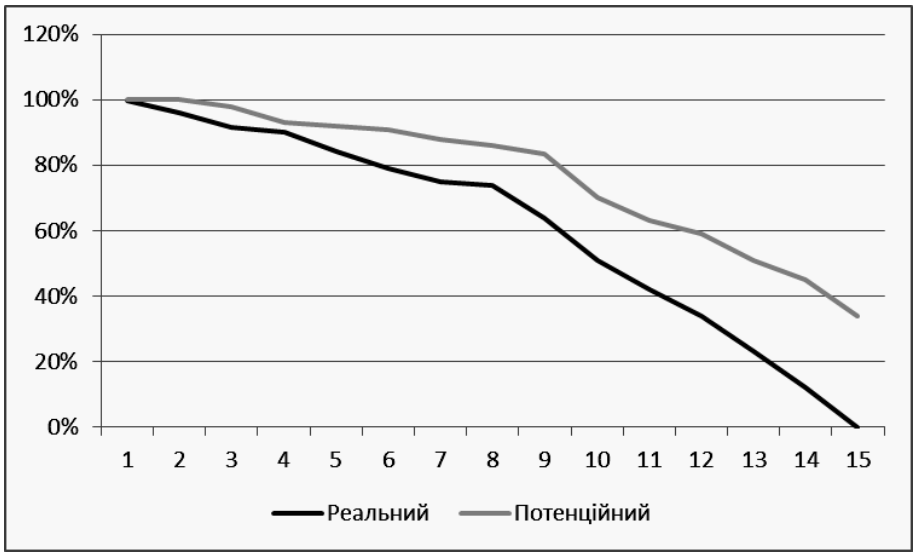

Рис. 2. Відображення впливу дедлайнів на результат поточного оцінювання

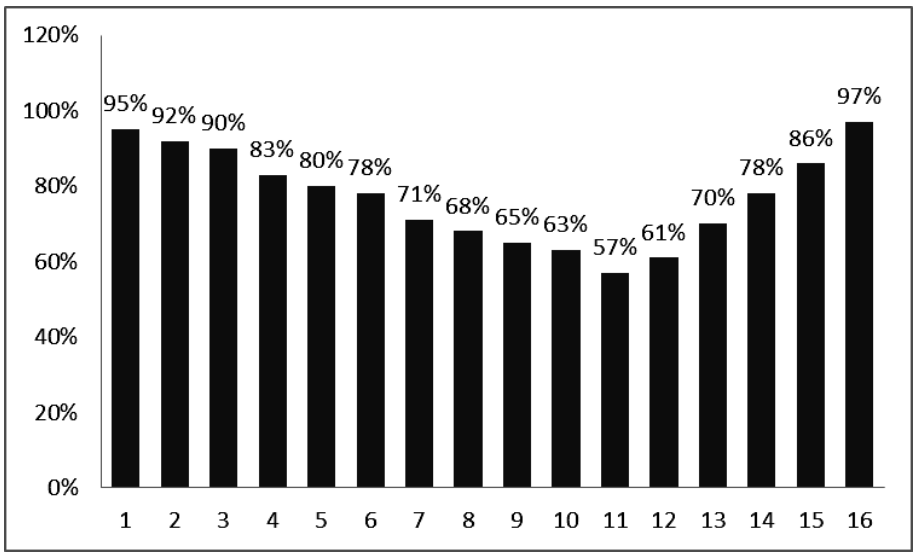

Рис. 3. Відображення рівня мотивації до виконання завдань протягом семестру

Спостерігається кореляція між мотивацією та активністю на заняттях упродовж семестру.

Під час організації навчального процесу в дистанційній формі виникає низка чинників, які потребують детального дослідження із використанням технологій опрацювання Великих даних. Серед них назвемо:

- Перенесення даних із допоміжних засобів дистанційного навчання. Багато викладачів використовують різні способи проведення та контролю навчального процесу, зокрема допоміжні платформи дистанційного навчання, електронну пошту, месенджери чи соціальні мережі. Інколи виникає плутанина із перенесенням даних та відстеженням оцінок чи навіть вищезгаданої активності, крім того, керівництво ЗВО потребує даних про моніторинг активності учасників освітнього процесу. Якщо у регламентованому ЗВО середовищі немає інформації, то виникатимуть претензії як до викладача, так і до студента. Це стосується як оцінювання, так і відвідуваності занять (лекцій, лабораторних, практичних, семінарів), які фіксують декількома способами: викладач і староста через перекличку під час відеозапису, внесенням відомостей у електронний журнал.
- Нехтування заняттями. Внаслідок дистанційного навчання спостерігається зниження відвідуваності занять серед студентства, а також подекуди і серед викладацького складу.
- Обмежений чи ненадійний доступ до мережі. Ця складова створює проблеми для 10–15 % студентів, і якщо під час навчання це не особливо завдає шкоди, то під час проміжних чи семестрових заходів контролю знань може стати відчутною.
- Проблеми із платформою дистанційного навчання. Через велике навантаження робота платформи уповільнюється, що особливо помітно під час складання сесії. З випадками збою системи стикаються 15–35 % студентів, право вирішення ситуації переважно залишається за викладачем.
- З двох останніх чинників випливає проблема об’єктивності як здобувача освіти, так і викладача. Часто таку інформацію неможливо перевірити.

Рис. 4. Кроки аналізу великих даних

Зазначені чинники особливо впливають на кінцеві вибірки даних, які аналізуватимуться. Також можливі коливання у важливості цих чинників залежно від типу ЗВО, кількості здобувачів освіти і підходів до організації дистанційного навчання.

Наступний етап – збирання даних, а оскільки навчальний процес генерує великі обсяги інформації, то потрібно застосовувати концепт великих даних та методи їх аналізу. Якщо необхідно працювати з великими даними, процес аналізу може бути доволі складним, залежно від їх обсягу, різноманітності, структурованості та якості, що особливо відчувається у контексті організації дистанційного навчання [5].

Для опрацювання таких даних доцільно використовувати алгоритми машинного навчання, хоча їх застосування теж супроводжується деякими проблемами.

Послідовність застосування методів опрацювання великих даних [6] наведено на рис. 4.

Виклад основного матеріалу

У статті проаналізуємо перші два кроки – очищення даних та опрацювання даних, що мають на меті спростити вибірку та знизити розмірність вибірки. В разі машинного навчання обсяг розглядається через два варіанти масштабування:

- Горизонтальний – атрибути, які характеризують інформацію. У нашому випадку це будуть основні характеристики навчального процесу, а саме: відстеження активності (можна розподілити на декілька окремих атрибутів), різноманіття дисциплін, кваліфікація викладачів, якість та актуальність наповнення бази даних освітніми матеріалами тощо.
- Вертикальний – кількість записів у наборі даних, на основі яких відбувається навчання.

Інформація про учасника освітнього процесу постійно накопичуватиметься з моменту внесення даних особи у систему дистанційного навчання. За рахунок збільшення обсягу даних у вертикальному чи горизонтальному напрямах час опрацювання даних зростає, а отже, і робота алгоритму сповільнюється. Для подолання цієї проблеми вважаємо за доцільне зменшувати розмірність вибірки.

Для зменшення розмірності використаємо метод PCA (Аналіз головних компонент). PCA – це статистична процедура, яка використовує ортогональне перетворення, переформовуючи групу корельованих змінних у групу некорельованих змінних [7].

Такий підхід має на меті зменшити кількість змінних у наборі даних, зберігаючи якомога більший обсяг інформації. PCA складається із таких етапів [8, 9]:

1. Приведення до єдиної системи усіх змінних і складових аналізу:

$$v = \frac{x - x_B}{\sigma}, \tag{1}$$

де v – змінна у єдиній системі, x – початкове значення змінної, x_B – середнє значення, σ – стандартне відхилення. Оскільки PCA надто чутливий до дисперсій початкових змінних, у разі приведення до єдиної системи кожне значення вибірки, незалежно від початкового джерела, буде давати однаковий внесок у розрахунки.

2. Обчислення коваріаційної матриці – квадратна матриця виду $n \times n$, де n – кількість вимірів і містить коваріації усіх можливих пар вхідних змінних:

$$\begin{matrix} Cov(x,x) & Cov(x,y) & \dots & Cov(x,n) \\ Cov(y,x) & Cov(y,y) & \dots & Cov(y,n) \\ \dots & \dots & \dots & \dots \\ Cov(n,x) & Cov(n,y) & \dots & Cov(n,n) \end{matrix}, \tag{2}$$

Якщо значення коваріації позитивне, то дві змінні зменшуються або збільшуються разом, у протилежному випадку одна змінна збільшується, а інша зменшується. Отже, можна буде спостерігати силу впливу чинників на академічну успішність.

3. Власні значення та вектори обчислюємо із коваріаційної матриці, щоб визначити головні компоненти даних. Головні компоненти у цьому випадку – нові змінні, що будуються як комбінації вихідних змінних, так, що нові змінні не корельовані, й більшість інформації стискається у початкових змінних, а решта заповнює наступні, доки усі виміри не будуть заповнені.

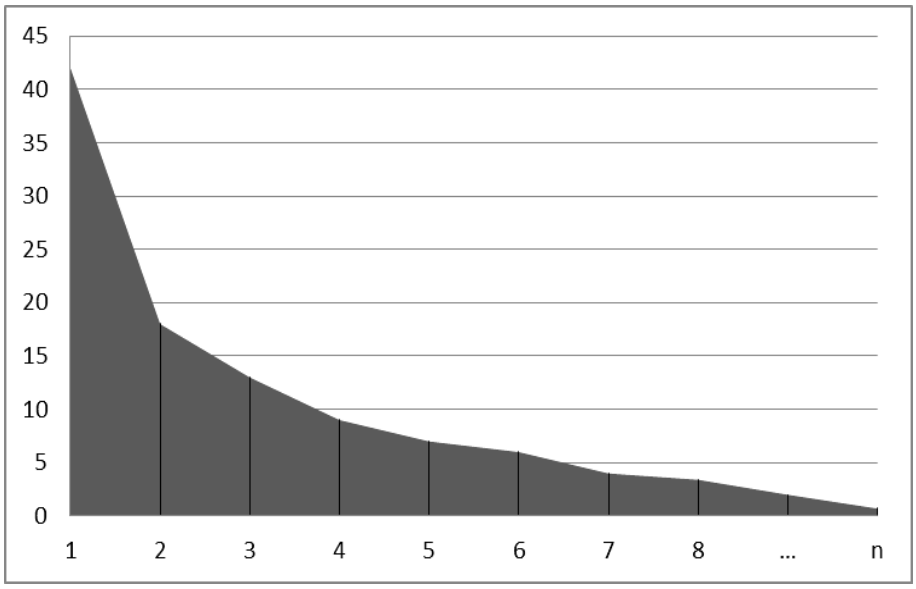

Рис. 5. Головні компоненти для n-вимірних даних

Як видно із графіка (рис. 5), перша компонента зберігає найбільшу кількість даних і враховуватиме найбільшу можливу дисперсію вибірки.

На цьому етапі може виникати проблема пошуку оптимальної кількості головних компонент. Під час роботи із потоковими даними оптимальна кількість компонент динамічна і нестабільна, що вимагає оновлення як основних компонент, так і їх розмірності на кожному кроці часового проміжку. Тому варто звернути увагу на дослідження, спрямовані на покращення взаємодії нейронних мереж та змінних PCA [10].

4. Створюємо вектор ознак на основі значущості компонент. На цьому етапі, залежно від власних значень, можна відкинути найменш значущі для аналізу компоненти. Із компонент, які беремо до уваги, складемо матрицю, що буде вектором ознак, стовпцями виступатимуть значення власних векторів. Відкидання кожного власного вектора призведе до зменшення розмірності даних.

Деякі з описаних вище чинників спричиняють проблем із моніторингом освітнього процесу, зокрема збій роботи платформи дистанційного навчання, тимчасові проблеми з доступом до мережі, недотримання дедлайнів, і впливають на вибірки із даними про успішність і активність. Тож питання нехтування деякими чинниками і їхнім впливом поки залишається відкритим і буде розглядатись під час подальших досліджень.

5. Формуємо кінцевий набір даних, сполучаючи початковий набір даних із вектором ознак. Початкові дані не відкидаються після формування вектора ознак, а проєктуватимуться на нього, тобто на головні компоненти.

Висновки

Нині все ще спостерігаються проблеми в організації навчального процесу із переходом на дистанційну форму. Для подолання таких ситуацій потрібен постійний моніторинг освітніх процесів та збирання даних із платформ дистанційного навчання. Оскільки запровадити цей моніторинг на всіх етапах освітнього процесу достатньо складно, особливо зважаючи на наявність безлічі проблемних моментів, доведено доцільність застосування методів аналізу великих даних, а також алгоритмів машинного навчання. Для роботи із початковими наборами даних запропоновано використовувати метод аналізу головних компонент (PCA), зокрема для зменшення розмірності даних.

Список літератури

1. Stukalo, N., Simakhova, A. (2020). COVID-19 Impact on Ukrainian Higher Education. *Universal Journal of Educational Research*, 8(8), 3673–3678.

2. Chen, T., Peng, L., Yin, X., Rong, J., Yang, J., Cong, G. (2020). Analysis of User Satisfaction with Online Education Platforms in China during the COVID-19 Pandemic. Healthcare. Basel. 2020 Jul 7;8(3):200. DOI: 10.3390/ healthcare8030200. PMID: 32645911; PMCID: PMC7551570.

3. Barabash, O., Laptiev, O., Sobchuk, V., Salanda, I., Melnychuk, Y., Lishchyna, V. (2021). Comprehensive Methods of Evaluation of Distance Learning System Functioning. *International Journal of Computer Network and Information Security*, 13(3):62–71. DOI: 10.5815/ijcnis.

4. CEDOS – аналітичний центр. Освіта в умовах пандемії у 2020/2021 році: аналіз проблем і наслідків. https://cedos.org.ua/researches/osvita-v-umovah-pandemiyi-analiz-problem-i-naslidkiv/

5. Пришляк А., Кунанець Н., Пасічник В., (2020) Інтелектуальна система формування персональних освітніх траєкторій у галузі ІТ. *Вісник Нац. ун-ту "Львівська політехніка" Інформаційні системи та мережі*, 7, 42–50.

6. Терещенко В., Бугайов А. (2018) Алгоритми машинного навчання у контексті великих даних. *Штучний інтелект*, 3(81), 80–86.

7. Maddikunta, P., Lakshmanna, K., Rajput, D., Srivastava, G., Baker, T., Gadekallu, T. & Kaluri, R. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, 8:54776–54788. DOI: 10.1109/ ACCESS.2020.2980942.

8. Nguen, T. (2020) Principal Component Analysis of Education-Related Data Sets. URL: http://resolver.tudelft.nl/uuid:11a166e3-cd94-45e8-91ed-660a0cfe8b9e.

9. Salih Hasan, B. M., Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, 2(1), 20–30. URL: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032

10. Migenda, N., Möller, R., Schenck, W. (2021). Adaptive dimensionality reduction for neural network-based online principal component analysis. *PLOS ONE*, 16(3): e0248896. https://doi.org/10.1371/ journal.pone.0248896.

References

1. Stukalo, N., Simakhova, A. (2020) COVID-19 Impact on Ukrainian Higher Education. *Universal Journal of Educational Research*, 8(8), 3673–3678.

2. Chen, T., Peng, L., Yin, X., Rong, J., Yang, J., Cong, G. (2020). Analysis of User Satisfaction with Online Education Platforms in China during the COVID-19 Pandemic. Healthcare. Basel. 2020 Jul 7;8(3):200. DOI: 10.3390/ healthcare8030200. PMID: 32645911; PMCID: PMC7551570.

3. Barabash, O., Laptiev, O., Sobchuk, V., Salanda, I., Melnychuk, Y., Lishchyna, V. (2021). Comprehensive Methods of Evaluation of Distance Learning System Functioning. International *Journal of Computer Network and Information Security*, 13(3):62–71. DOI: 10.5815/ijcnis.

4. CEDOS – analytical center. Education in the pandemic in 2020/2021: analysis of problems and consequences. URL: https://cedos.org.ua/researches/osvita-v-umovah-pandemiyi-analiz-problem-i-naslidkiv/

5. Pryshliak, A., Kunanets, N., Pasichnyk, V. (2020). Intellectual system of formation of personal educational trajectories in IT. *The Journal of Lviv Polytechnic National University "Information Systems and Networks"* 7, 42–50.

6. Tereshchenko, V., Bugayov, A., (2018) Algorithms of machine learning in the context of big data. *Artificial Intelligence*, 3(81), 80–86.

7. Maddikunta, P., Lakshmanna, K., Rajput, D., Srivastava, G., Baker, T., Gadekallu, T. & Kaluri, R. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, 8:54776–54788. DOI: 10.1109/ACCESS.2020.2980942

8. Nguen, T. (2020) Principal Component Analysis of Education-Related Data Sets. URL: http://resolver.tudelft.nl/uuid:11a166e3-cd94-45e8-91ed-660a0cfe8b9e.

9. Salih Hasan, B. M., Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, 2(1), 20–30. Retrieved from https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032.

10. Migenda, N., Möller, R., Schenck, W. (2021). Adaptive dimensionality reduction for neural network-based online principal component analysis. *PLOS ONE*, 16(3): e0248896. URL: https://doi.org/10.1371/ journal.pone.0248896.

PROCESSES AND ELEMENTS OF BIG DATA ANALISYS OF DISTANCE LEARNING SYSTEMS

A. Pryshiak[1], N. Kunanets[2], V. Pasichnyk[3]

Lviv Polytechnic National University, Department of Information Systems and Networks,
[1] e-mail: andrii.a.pryshliak@lpnu.ua; ORCID: 0000-0003-1681-5178,
[2] e-mail: nek.lviv@gmail.com, ORCID: 0000-0003-3007-2462,
[3] e-mail: vpasichnyk@gmail.com, ORCID: 0000-0002-5231-6395

The impact of the pandemic on educational processes in Ukraine is analyzed. The problematic moments observed during distance learning, positive and negative factors of online education are considered. Factors that can lead to conflict situations in the educational process and complicate the process of collecting and analyzing information are presented. The use of machine learning methods for big data analysis in distance learning systems is proposed. The method of analysis of the main components to reduce the dimensionality of the sample size is considered and the main steps that need to be implemented on the way to simplification are described. The possibility of analysis is ensured by the proper functioning of the distance learning system of the regulated university, interaction with all participants in the educational process, as well as the timely performance of the duties assigned to them.

Key words: Big data; distance learning systems; online education; machine learning; PCA method; data dimension.