Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Використання технік інтелектуального аналізу даних для визначення рівня цифрової зрілості малих та середніх підприємств

Виконала: студентка VI курсу, групи СНнм-61
спеціальності 122 – Комп'ютерні науки
(шифр і назва спеціальності)

(підпис) Гладьо О.Ю. (прізвище та ініціали)

Керівник (підпис) Струтинська І.В. (прізвище та ініціали)

Нормоконтроль (підпис) Мацюк О.В. (прізвище та ініціали)

Завідувач кафедри (підпис) Боднарчук І.О. (прізвище та ініціали)

Рецензент (підпис) Осухівська Г.М. (прізвище та ініціали)

Тернопіль
2022

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«___» _____ 20___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 122 – Комп'ютерні науки
(шифр і назва спеціальності)

студенту Гладьо Ользі Юріївні
(прізвище, ім'я, по батькові)

1. Тема роботи Використання технік інтелектуального аналізу даних для визначення рівня цифрової зрілості малих та середніх підприємств

Керівник роботи Струтинська І.В., д.е.н., професор кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «28» жовтня 2021 року № 4/7-909

2. Термін подання студентом завершеної роботи 26 травня 2022 року

3. Вихідні дані до роботи Наукові джерела щодо інтелектуального аналізу даних, задачі кластеризації та поняття рівня цифровізації мікро, малих та середніх підприємств

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1 Аналіз даних як інструмент підвищення ефективності бізнесу. 2 Кластеризація як одна із задач видобутку даних. 3 Кластеризація мікро, малих та середніх підприємств Тернопільської області за рівнем їх цифрової зрілості. 4. Охорона праці та безпека в надзвичайних ситуаціях. Перелік використаних джерел. Висновки. Додатки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1 Титульний слайд. 2 Мета, об'єкт та предмет дослідження. 3 Завдання дослідження. 4 Актуальність дослідження. 5 Концепції аналізу даних. 6 Поняття цифровізації та HIT-індексу. 7 Задача кластеризації даних. 8 Поняття міри відстані та якості кластеризації 9 Комп'ютерна програма для проведення кластеризації. 10 Приклад аналізу результатів кластеризації. 11 Висновки щодо загальних тенденцій. 12 Висновки. 13 Завершальний слайд.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Дмитроца Л.П., к.т.н., доц. кафедри КН		
Безпека в надзвичайних ситуаціях	Клепчик В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 27 вересня 2021 року

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	27.09.21-30.09.21	*Виконано*
2.	Підбір наукових джерел щодо аналізу даних, технік	01.10.21-19.10.21	*Виконано*
	інтелектуального видобутку даних і задачі кластеризації		
3.	Переклад, опрацювання та аналіз наукових джерел щодо	20.10.21-21.11.21	*Виконано*
	питань висвітлених у кваліфікаційній роботі		
4.	Обґрунтування виконуваного дослідження	22.11.21-25.11.21	*Виконано*
5.	Виконання практичних завдань магістерського	26.11.21-06.03.22	*Виконано*
	дослідження щодо обчислення індексу цифрової		
	трансформації, кластеризації опитаних та аналізу		
	отриманих результатів		
6.	Оформлення розділу «Аналіз даних як інструмент	07.03.22-16.03.22	*Виконано*
	підвищення ефективності бізнесу»		
7.	Оформлення розділу «Кластеризація як одна із задач	17.03.22-27.03.22	*Виконано*
	видобутку даних»		
8.	Оформлення розділу «Кластеризація мікро, малих та	28.03.22-03.04.22	*Виконано*
	середніх підприємств Тернопільської області за рівнем		
	їх цифрової зрілості»		
9.	Виконання завдання до підрозділу «Охорона праці»	04.04.22-14.04.22	*Виконано*
10.	Виконання завдання до підрозділу «Безпека в	15.04.22-25.04.22	*Виконано*
	надзвичайних ситуаціях»		
11.	Оформлення кваліфікаційної роботи	25.04.22-01.05.22	*Виконано*
12.	Нормоконтроль	02.05.22-04.05.22	*Виконано*
13.	Перевірка на плагіат	09.05.22	*Виконано*
14.	Попередній захист кваліфікаційної роботи	11.05.22	*Виконано*
15.	Захист кваліфікаційної роботи	26.05.22	

Студент _____ Гладьо О. Ю.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Струтинська І. В.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Використання технік інтелектуального аналізу даних для визначення рівня цифрової зрілості малих та середніх підприємств // Кваліфікаційна робота освітнього рівня «Магістр» // Гладьо Ольга Юріївна // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНнм-61 // Тернопіль, 2022 // С. – 124, рис. – 20, табл. – 3, додат. – 6, бібліогр. – 88.

Ключові слова: МАЛІ ТА СЕРЕДНІ ПІДПРИЄМСТВА, КЛАСТЕРИЗАЦІЯ, АНАЛІЗ ДАНИХ, HIT-ІНДЕКС, ЦИФРОВА ЗРІЛІСТЬ

Кваліфікаційну роботу присвячено кластеризації набору даних опитування підприємців Тернопільської області щодо рівня цифровізації бізнес-процесів у їхніх компаніях та аналізу отриманих результатів.

У першому розділі роботи розглянуто загальні поняття аналізу даних і бізнес-аналітики, а також описано різного роду типи задач та методи їх вирішення. Другий розділ роботи присвячено детальному огляду поняття кластеризації та її підходам, методам, поняттям міри відстані та метрикам якості кластерів тощо. У третьому розділі описано розробку програмного забезпечення для виконання поставлених завдань, проведення кластеризації даних декількома способами, візуалізацію та аналіз отриманих результатів.

Окрім цього, у роботі коротко розглянуто поняття індексу цифрової трансформації підприємства HIT. Обґрунтовано вибір методів кластеризації та використовуваних мір відстані на основі наявного набору даних та виокремлено спільні характерні риси у проблематиці цифровізації малих та середніх підприємств.

ANNOTATION

Applying of data mining techniques for estimating of digital maturity level of small and middle companies // Qualification thesis Master Degree // Hlado Olha Yuriivna // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Computer Science Department, group SNnm-61 // Ternopil, 2022 // Pages – 124, Fig. – 20, Tables. – 3, Annexes – 6, References – 88.

Keywords: SMALL AND MEDIUM ENTERPRIZES, CLUSTERING, DATA ANALYSIS, HIT INDEX, DIGITAL MATURITY

The qualification thesis is dedicated to the process of clustering of the data set based on surveying entrepreneurs of the Ternopil region on the question of business processes digitalization in their companies. In addition, this research work is devoted to the analysis of the received results.

The first section of the work describes general concepts of data analysis and business analytics. In addition, various types of tasks and methods used to resolve them are discussed. The second section of the thesis is dedicated to a detailed review of the clustering problem, its methods and approaches, the idea of distance measure, quality metric and more. The development of specialized software aimed at performing tasks set for the work, including clustering of the given data set by a few methods, visualization and analysis of obtained results are described in the third section of qualification work.

Besides this, a short consideration of the digital transformation index (HIT) is presented. The common features in the issue of digitalization of small and medium enterprises are also highlighted.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ І СКОРОЧЕНЬ

DA (з англ. Data Analysis) – аналіз даних.

DM (з англ. Data Mining) – видобуток даних.

BA (з англ. Business Analysis) – бізнес-аналіз.

BI (з англ. Business Intelligence) – бізнес-аналітика.

KDD (з англ. Knowledge in Data Discovery) – виявлення знань у даних.

HIT (з англ. Human, Information, Technology) – індекс цифрової трансформації.

МСП – малі та середні підприємства.

ММСП – мікро, малі та середні підприємства.

ЕМ (з англ. Expectation-Maximization) – алгоритм очікування-максимізації.

СУБД – система управління базами даних.

БД – база даних.

ДСМД – державна служба моніторингу довкілля.

ЄС – Європейський Союз.

ЗМІСТ

ВСТУП

Актуальність обраної теми. У сучасному світі відбувається глобальна цифровізація економіки та впровадження новітніх цифрових інструментів як у менеджерську, так і у виробничу складову діяльності підприємств. Очевидно, що такі зміни мають місце й в Україні, проте вони носять дещо повільніший характер через недостатню кількість відкритих та доступних ресурсів з питань цифровізації та інструментів, які можна використовувати при модернізації власних процесів. Особливу нестачу інформації щодо можливостей змін відзначають представники мікро- та малого бізнесу, адже з одного боку вони можуть бути гнучкими щодо постійних змін, а з іншого – у таких підприємств може не вистачати ресурсів для самостійного вдосконалення. Важливим є розуміння поточного стану цифровізації бізнес-процесів у мікро, малому та середньому бізнесі України, бачення спільних викликів у певних груп підприємств та побудова системи допомоги представникам компаній для пришвидшення цифровізації процесів. Зважаючи на це, першим кроком можна визначити дослідження рівня цифрової зрілості та спільних викликів у представників малого та середнього бізнесу, що є актуальною темою для висвітлення.

Метою кваліфікаційної роботи є виявлення нових цінних ідей щодо розуміння поточного стану використання цифрових технологій для підтримки бізнес-процесів у мікро, малих та середніх підприємствах за допомогою застосування різних методів кластерного аналізу даних.

Для досягнення окресленої мети було поставлено такі **завдання роботи**:

- Розглянути поняття аналізу даних, його основні типи та методи, а також задачі видобутку даних.
- Розглянути та проаналізувати основні підходи і методи кластеризації даних, поняття міри відстані та якості утворених кластерів.

- На основі аналізу здійснити вибір методів кластеризації та її параметрів для застосування у подальших завданнях.
- Опрацювати наявний набір даних опитування щодо поточного стану цифровізації бізнес-структур.
- Розробити допоміжну програму для роботи з набором даних, його кластеризації, отриманні результатів та візуалізації.
- Розрахувати індекс цифрової трансформації та на його основі здійснити кластеризацію респондентів використовуючи обраний перелік методів та мір відстані.
- Проаналізувати отримані результати кластеризації на предмет нових ідей щодо проблем мікро, малого та середнього підприємництва у розрізі цифровізації бізнес-процесів.

Об’єктом дослідження є мікро, малі та середні підприємства Тернопільської області.

Предметом дослідження є кластеризація мікро, малих та середніх підприємств Тернопільської області у розрізі їхнього рівня цифрової зрілості: використання цифрових інструментів, інфраструктура, грамотність працівників, а також пошук внутрішніх груп у вибірці.

Наукова новизна роботи. Вперше використано методи кластерного аналізу даних для розуміння поточного стану цифровізації бізнес-процесів представників малого та середнього бізнесу Тернопільської області. Виявлено спільні риси отриманих груп підприємств, їхні сильні та слабкі місця у розрізі використання цифрових інструментів та цифрової грамотності людського капіталу.

Практичне значення роботи. Отримано результати кластерного аналізу малих та середніх підприємств Тернопільської області та цінні ідеї щодо важливості окремих компонент HIT-індексу. У подальшому стійкі сформовані кластери можуть бути використані для класифікації нових опитаних підприємств та виокремленні значущих атрибутів з найбільшим

впливом на значення цифрової зрілості суб'єкта або ж для формування методології надання рекомендацій щодо покращення рівня цифрової зрілості підприємства.

Апробація результатів магістерської роботи: окремі результати дослідження було обговорено на наступних наукових конференціях:

1. 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T). На тему: «Small and Medium Business Structures Clustering Method Based on Their Digital Maturity».
2. International Workshop on Conflict Management in Global Information Networks (CMiGIN 2019). На тему: «Comparative Analysis of Two Approaches to the Clustering of Respondents (based on Survey Results)».
3. 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops (ICTERI 2020). На тему: «Developing Practical Recommendations for Increasing the Level of Digital Business Transformation Index».
4. 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T). На тему: «Working-Out of Recommendation System to Increase the Digital Maturity Level of Enterprises».
5. 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT). На тему: «Development of Digital Platform to Identify and Monitor the Digital Business Transformation Index».
6. Всеукраїнський конкурс студентських наукових робіт зі спеціалізації «Цифровізація економіки», 2021/2022 н.р. На тему: «Використання технік аналізу даних для кластеризації бізнес-структур за рівнем їх цифрової зрілості».

Публікації щодо теми дослідження. Результати досліджень було опубліковано у збірниках праць, наведених у додатках А-Д. Окрім цього за результатами досліджень було отримано свідоцтво про реєстрацію авторського права на твір (див. додаток Е).

Структура і обсяг кваліфікаційної роботи. Кваліфікаційна робота складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел з 88 найменувань та 6 додатків. Загальний обсяг кваліфікаційної роботи становить 124 сторінки, з них 85 сторінок основного тексту, 20 рисунків та 7 таблиць.

1 АНАЛІЗ ДАНИХ ЯК ІНСТРУМЕНТ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ БІЗНЕСУ

Одним із основних компонентів сучасного світу та постійною темою для обговорень і впровадження нових технологій є дані. З точки зору інформаційних технологій: «Дані – це інформація, представлена в формалізованому вигляді, що забезпечує можливість її зберігання, опрацювання і передавання» [1].

Дані використовуються у чи не всіх сферах, які стосуються як і повсякденного життя, так і роботи організацій, бізнес-структур чи систем забезпечення життєдіяльності. Найчастіше дані застосовують як певне підґрунтя для прийняття рішень – зібрана з різних джерел інформація свідомо чи несвідомо опрацьовується людиною, і на її основі виконуються певні дії. Деякими прикладами можуть слугувати, наприклад, вибір одягу в залежності від погоди, чи рішення про збільшення накладу продукції за даними маркетингового дослідження.

Проте, для того, щоб дані надавали бізнесу значущі ідеї щодо пріоритетного вектору майбутнього рішення, такі дані повинні бути попередньо проаналізовані. Адже сирі – тобто, необроблені – дані найчастіше утворюються, передаються та зберігаються у форматі масивів чисел чи стрічок тексту, і у загальному не становлять практичної цінності для пересічного користувача. Такі дані можуть бути згенеровані користувачами (у вигляді чисел, текстів, зображень, відеопотоку), сенсорами інтернету речей, автоматизованими системами (наприклад, касовими апаратами) чи передаватися як інтернет-трафік використання інтернет-сторінки користувачем тощо.

Завдання виокремлення корисних даних з доступних джерел, їх подальшої обробки, розуміння, пошуку відповідей на поставлені запитання можна назвати основою популярного сьогодні напрямку аналізу даних та

аналітики. Сфера аналізу даних найчастіше використовується у бізнесі для прийняття стратегічних, оперативних чи тактичних рішень, або ж як частина алгоритму пропонованого сервісу. Окрім цього часто потреба у роботі з даними для отримання змістовної інформації і її використання у подальшому, з'являється на перетині декількох суміжних бізнес-сфер, таких як аналіз даних (DA), бізнес-аналіз (BA), бізнес-аналітика (BI) та видобуток даних (DM).

1.1 Ідентифікація концепцій щодо роботи з даними

Аналіз даних (з англійської, Data Analysis, DA) – це процес роботи з даними, що включає в себе стадії вилучення, очищення, перетворення, перевірки, моделювання, вивчення та візуалізації. Під час виконання етапів аналітик отримує цінну інформацію, на основі яких може робити висновки та/або приймати бізнес-рішення. Підходи, методи та аспекти роботи з даними різняться в різних сферах науки або бізнесу, але незмінним є те, що у сучасному світі дані та інформація, яку вони надають, є джерелом істини у прийнятті рішень. [2]

Основними завданнями аналітика є перетворення даних та чисел у зрозумілу форму, щоб бізнес отримав чіткі відповіді на свої запитання. Часто це включає в себе очищення даних, виявлення зв'язків та кореляцій, та пошук шаблонів поведінки з використанням статистичних методів. Окрім цього відбувається аналіз видобутих даних та їх візуалізація у формі, яка найкраще представить знахідки для ключових персон бізнесу. Аналіз даних використовується у дослідженнях ринку, ребрендингу, позиціонуванні компанії, аналізі відгуків клієнтів, дослідження нових та існуючих товарів, аналіз суспільної думки тощо. У загальному, аналіз надає бізнесу уявлення про наявну ситуацію, опираючись на наявні дані, тобто частіше описує ситуацію, якою вона є на даний момент.

Спорідненим до аналізу даних є напрямок, видобутку – або ж, майнингу – даних, тобто процес вилучення корисних даних чи виявлення шаблонів у великих наборах необроблених даних, або ж те що ще називають виявленням знань у даних (KDD, з англійської, Knowledge Discovery in Data). Видобуток даних часто включає в себе методи і підходи, які поєднують математичну статистику, машинне навчання та роботу з базами даних, що дозволяє говорити про міждисциплінарність. Також при видобутку використовується обробка даних, створення моделей, метрики якості, складність алгоритмів, обробка виявлених структур тощо. Більш того, поширеними є моделі машинного навчання, які потім можна перетворити у алгоритми штучного інтелекту [3]. Різниця між аналізом даних та видобутком даних полягає в тому, що аналіз даних використовується для тестування моделей та гіпотез на наборі даних; на відміну від цього, обробка даних використовує машинне навчання та статистичні моделі для розкриття прихованих моделей у великому обсязі даних.

З іншого боку існує менш технічна сторона аналітики, яка стосується вирішення бізнес-проблем, на основі даних, які надаються аналітиками даних. Бізнес-аналіз (з англійської, Business analysis, BA) – це дослідження даних організації для виявлення корисної інформації, яка може стимулювати фінансові результати. Бізнес-аналіз часто виявляє проблеми бізнесу і працює на їх вирішення. Запропоновані шляхи іноді включають потребу розробки програмних систем, але можуть і стосуватися реінжинірингу, удосконалення процесів, організаційних змін, стратегічного планування тощо.

У створенні програмного забезпечення бізнес-аналіз займає ключову роль у формуванні бізнес- та користувацьких вимог, встановленні правил, обмежень та у консультуванні команди розробки у суперечливих питаннях. У загальному випадку ж, бізнес-аналіз приймає на себе роль консультанта, який допомагає у вирішенні певних бізнес-ситуацій, оцінюванні, встановленні

пріоритетності, та власне у вдосконаленні процесів та систем організації через ефективне використання наявної інформації [4].

Дуже спорідненою до бізнес-аналізу галуззю можна визначити бізнес-аналітику (з англійської, Business Intelligence, BI). У деяких джерелах ці терміни визначають як взаємозамінні, проте вони все ж мають незначні відмінності. Бізнес-аналітика також включає в себе аналіз даних та мету пошуку тенденцій і закономірностей у результатах роботи бізнесу. Окрім пошуку стандартних показників, які найчастіше стосуються фінансів, бізнес-аналітика також концентрується на кореляціях продуктів, загальних змінах на поточному ринку, впливі внутрішніх та зовнішніх факторів на прибутки, ефективність управління та маркетингу тощо.

Бізнес-аналітика часто фокусується на стратегічних та операційних цілях, які аналізуються з допомогою історичних, поточних та прогностичних даних, які можуть випливати як результат аналізу та/або видобутку даних. Така аналітика допомагає виявляти та розробляти нові стратегічні можливості, та впливає на довгострокові цілі бізнесу. У випадку аналітики найефективнішим є використання як зовнішніх даних з ринку (продажі, результати промо-кампаній, опитування клієнтів) так і внутрішніх даних організації (фінансові, операційні, управлінські дані). Таке поєднання ідей взятих з великих масивів даних разом з детальним аналізом внутрішніх процесів компанії дозволяє ухвалювати зважені рішення [5, 6].

Ці 4 концепції, також враховуючи машинне навчання та роботу з великими даними, можна назвати спорідненими, адже усі вони стосуються роботи з даними та спрямовуються на отримання доказової бази для прийняття якомога ефективніших бізнес-рішень. Дата аналіз використовує видобуток даних як один із методів аналізу. Бізнес-аналіз та бізнес-аналітика послуговуються ідеями та концепціями, отриманими в результаті аналізу даних для подальшої оцінки та прийняття рішень.

1.2 Загальні підходи до аналізу даних

Як вже було згадано, аналіз даних – це процес збору та впорядкування даних, щоб знайти у них сенс та нові ідеї. Доступні дані можуть надходити як і з зовнішніх джерел (від постачальників даних, з результатів опитувань, державної статистики тощо), так і з внутрішніх, адже часто компанії збирають величезні обсяги даних (транзакції, дані про покупців, логістика, продажі продукту тощо). Справжня цінність аналізу даних полягає в його здатності розпізнавати закономірності в наборі даних, які можуть вказувати на тенденції, ризики чи можливості. Аналіз даних дозволяє компаніям модифікувати свої процеси на основі цих знань для прийняття кращих рішень, адже аргументи було підкріплено доказовою базою: цифрами, візуалізаціями, роботою аналітика тощо.

Окрім цього, все більше організацій працюють над перенесенням власних процесів та внутрішніх додатків у хмару, що дозволяє їм швидше впроваджувати інновації. Хмарні технології створюють швидке інноваційне середовище, де команди аналітиків даних можуть зберігати більше даних і легше отримувати доступ до них та досліджувати їх, що призводить до швидшого оцінювання часу для нових рішень дослідження [8, 9].

Залежно від мети аналізу існує багато різних підходів, які використовують спеціалісти. Проте, всі ці методики та інструменти базуються на двох основних напрямках: кількісні та якісні дослідження. Окрім цього існує декілька основних методів аналізу, серед яких: описовий, дослідницький, діагностичний, прогностичний та настановний аналіз, які будуть згадані нижче.

Описовий аналіз (з англійської, descriptive analytics) розглядає дані у минулому, та працює з запитанням: «Що сталося?». Така аналітика розглядає історичні дані, щоб краще розуміти, яким чином планувати майбутнє. Завдяки різноманітності інформаційних інструментів, бізнес може

отримувати перевагу у використанні великих даних, які описують попередній досвід компанії чи ринку. Описовий аналіз часто є відправною точкою будь-якого аналітичного проєкту, адже стосується обробки, маніпуляцій та інтерпретацій наявних даних з різних джерел. Після чого дані зберігаються в упорядкованому вигляді, готові до інших методів аналізу. Таку аналітику часто використовують для відстеження прогресу компанії за певний період, тому варто зауважити, що сама по собі вона не дозволяє передбачати майбутнє чи спрямовувати дії у певне русло [8, 10, 11].

Дослідницький аналіз (з англійської, exploratory analytics) спрямований на дослідження даних – власне їх набору чи масиву. Дослідницький аналіз часто використовується у видобутку даних для пошуку зв’язку між даними, змінними, формування гіпотез тощо. Проте, дослідницьким також можна назвати й аналіз набору даних у той момент, коли аналітик вперше з ним працює. Такий аналіз включає в себе ознайомлення з схемою даних, їх типами, поверхневий пошук брудних даних, розробка плану стандартизації тощо [11].

Діагностичний аналіз (з англійської, diagnostic analytics) має на меті визначити причину того, чому дані саме такі, тобто відповісти на запитання: «Чому це сталося?». Така аналітика часто використовується у парі з описовим аналізом, і допомагає зрозуміти, чому відбулася позитивна чи негативна зміна даних, яку показав описовий аналіз. У такому випадку бізнес має чітке розуміння контексту та передумов події, що допомагає якомога краще визначити шляхи вирішення проблеми або майбутні кроки для підтримки результатів. Діагностичний аналіз важливий у багатьох сферах, наприклад, у роздрібній торгівлі, адже можна визначити, яка маркетингова кампанія призвела до збільшення продажів у попередній місяць [10, 11] .

Інтелектуальний, або ж прогностичний, аналіз (з англійської, predictive analytics) працює з передбаченням того, що, ймовірно, станеться у майбутньому, і відповідає на питання: «Що буде далі?». У цьому типі

дослідження дані походять з минулого, і на їх основі формуються прогнози про майбутнє. Таким чином, бізнес може виявити наступні тенденції, потенційні проблеми, зв’язки і втрати даних тощо. Оскільки прогностична аналітика – це вищий рівень розуміння даних, він використовує результати раніше згаданих типів аналізу, статистичні методи, машинне навчання та штучний інтелект для передбачення майбутніх подій. Проте, варто розуміти, що такий аналіз не є повністю достовірним, а лише з певною ймовірністю припускає вірогідний результат. Зважаючи на те, що у сьогоденні певні популярні теми з’являються за лічені дні, використання прогностичного аналізу іноді вимагає дуже частої перевірки стану суспільства та трендів, щоб виявити їх на стадії створення. Маючи розуміння, що певний сплеск даних був результатом тренду, можна спроектувати, коли з’явиться схожа можливість у майбутньому [8, 10, 11].

Настановний аналіз (з англійської, prescriptive analytics) поєднує у собі інформацію із попередніх видів аналітики, щоб сформувати план дії для прийняття рішення, керованого даними. Настановна аналітика досліджує можливі дії, вжиті на основі результатів, та їх наслідки. Окрім цього, вона поєднує в собі математичні моделі та бізнес-правила, оптимізуючи прийняття рішень, рекомендуючи різні сценарії, компроміси та можливі варіанти розвитку подій. Часто настановний аналіз використовують у таких бізнес-сферах, як маркетинг, продажі, робота з клієнтами, логістика тощо [8, 10, 11].

1.3 Підходи до видобутку даних

Очевидно, що вищезгадані типи аналітики вимагають використання специфічних технік, методів та інструментів для отримання результатів. Це включає в себе як програмні інструменти, так й теоретичні відомості щодо того чи іншого методу аналізу. У даному підрозділі буде розглянуто кілька

найбільш популярних методів видобутку даних, із зосередженням на суті та прикладній цінності такого аналізу.

Варто згадати поняття кількісних та якісних даних. Кількісні дані включають в себе конкретні величини та цифри. Наприклад, показники продажів, коефіцієнти переходів, відсоток збільшення продаж, трафік веб-сайту тощо. Методи, що використовуються для кількісних даних, зосереджені на статистичному, математичному або чисельному аналізі великих наборів даних, а також на обчислювальних методах та алгоритмах.

З іншого боку, якісні дані часто не можуть бути об’єктивно виміряні і піддаються більш суб’єктивному аналізу. Такими даними можуть бути, наприклад, коментарі в опитуваннях, розшифровки інтерв’ю, публікації у соціальних мережах, відгуки про товари тощо. Аналіз якісних даних також стосується осмисленню даних, які не є структурованими, тому у такому випадку часто використовуються алгоритми штучного інтелекту, наприклад, обробка природної мови [12].

Також важливим кроком у процесі видобутку – або ж, майнингу – даних є їх очищення та підготовка до аналізу, адже при роботі алгоритму з даними, які не було попередньо оброблено, можуть отримувати нерелевантні результати, або помилки роботи програмного коду. Робота з підготовки та очищення даних включає в себе елементи трансформації даних (зміна форматів, додавання атрибутів, заміна порожніх значень, приведення інформації в один інтервал тощо), а також міграцію, інтеграцію, агрегування та процеси перетворення даних.

Дії з очищення даних можна віднести до дослідницького аналізу набору даних, який необхідно виконати, щоб ознайомитися з особливостями та атрибутами даних для їх ефективнішого використання. Компанії повинні довіряти своїм даним, результатам аналітики та дії, створеній на основі цих результатів, і саме для цього початкові набори даних повинні бути надійними та очищеними, що дозволить їм визначатися як якісні [13].

Власне видобуток даних включає в себе різноманіття статистичних та математичних методів, які використовують обчислювальні алгоритми та формули для аналізу даних. Серед цих технік можна виділити [3]: виявлення аномалій, асоціативні правила, мережі Баєса, дерева рішень, факторний аналіз, класифікацію, кластеризацію, генетичні алгоритми, регресію, нейронні мережі, опорні машини векторів, аналіз текстів, часові ряди тощо. Нижче розглянуто основні техніки, що використовуються у видобутку даних.

1.3.1 Задача кластерного аналізу

Кластеризація – це дослідницький прийом спрямований на ідентифікацію раніше невідомих структур у наборі даних. Метою кластерного аналізу є сортування набору даних таким чином, щоб утворилися групи, які є однорідними – подібними – всередині групи, але неоднорідними – достатньо різними – порівнюючи з іншими групами. Тобто, це буде означати, що точки даних у кластері подібні між собою і відрізняються від точок даних в іншому кластері. За допомогою кластеризації можна зрозуміти внутрішню структуру даних, їх подібність чи відмінність між собою.

З точки зору бізнесу реальні застосування кластерного аналізу включають в себе дослідження ринку, розпізнавання образів, аналіз даних тощо. У маркетингу він часто застосовується для групування клієнтів в окремі сегменти, що дозволяє створювати більш спрямовану рекламу та канали комунікації з різними цільовими аудиторіями. Страхові компанії можуть використовувати кластеризацію для пошуку місць з найбільшою кількістю звернень, а у геології кластерний аналіз можна використовувати для оцінки ризику землетрусів чи інших несприятливих природніх явищ [11, 12, 13, 14]

1.3.2 Задача класифікації

Класифікація – це один із найбільш популярних напрямків видобутку даних та машинного навчання. Її технічне завдання полягає у тому щоб розподілити предмети згідно з визначеними ознаками, які вже було задано у наборі даних. Тобто набір даних складається із багатьох точок, які описуються різноманітними атрибутами, і які вже розподілені за якимись категоріями. Завдання полягає у визначенні категорії для нового зразка у наборі. У цьому випадку з'являється поняття цільової змінної – атрибута, який і вказує, до якої категорії відноситься об'єкт. Також класифікація часто використовується для виділення групи нестандартних точок даних – зміни фінансового ринку, негативні медичні показники тощо.

У бізнесі класифікація часто використовується для розподілу нових клієнтів у вже визначені кластери: за покупками, інтересами, звичками тощо. Також класифікація працює у сфері послуг, оцінюючи чи продовжуватиме контракт клієнт на основі його активності за попередній рік, або у банківській сфері для прийняття рішень, наприклад, про надання кредиту [13, 14].

1.3.3 Задача регресійного аналізу

Регресійний аналіз – це метод видобутку даних і машинного навчання, який використовується для оцінки взаємозв'язків у наборі даних. При проведенні аналізу увага звертається на те, чи існує зв'язок між залежною цільовою змінною та іншими незалежними змінними. Регресія використовує історичні дані та одну або кілька незалежних змінних чи їх комбінації для пошуку найкращої моделі. Метою такого пошуку є оцінка того, який вплив мають атрибути набору даних на значення цільової змінної, і чи можна таке відношення описати математичною формулою кореляції.

Такий аналіз особливо цікавий для виявлення тенденцій та закономірностей, а також для прогнозування майбутнього. Типи та підходи до

регресійного аналізу залежать від того, який тип даних використовується – якісний, чи кількісний. У бізнесі він часто використовується для прогнозування кількісних величин – ціні оренди будинку, продажів за наступний квартал, вартості наданого кредиту тощо [11, 12, 13].

1.3.4 Задача на пошук асоціацій

Пошук асоціацій – це технологія аналізу даних, що тісно пов’язана зі статистикою. Під час такого аналізу можна отримати зв’язки та залежності між двома або більше елементами, тобто розпізнавання невідомої раніше закономірності. Результат аналізу вказує на те, що певні дані пов’язані з іншими. Також цікавим є те, що найчастіше асоціативні правила можна описати у вигляді простої конструкції «Якщо-то», яка допомагає знайти зв’язок у, здавалося б, незалежних наборах даних або категоріях. Проте, варто зазначити, що таке правило є радше одностороннім. Тобто, у випадку істинності твердження: «Якщо покупець купує А, то він купує і Б», істинність твердження «Якщо покупець купує Б, то він купує й А» не є автоматично доведеною. Іншими словами, твердження не є еквівалентними при зміні послідовності точок даних.

Правила асоціацій найчастіше називають аналізом споживчого кошика, адже це була перша область застосування методу. Мета такого аналізу полягає у пошуку пар продуктів, які найчастіше зустрічаються в одному кошику, тобто купівля одного продукту з пари може з певною ймовірністю передбачати покупку й іншого. Також асоціації використовують у інтернет-магазинах, на сервісах перегляду фільмів та прослуховування музики, коли застосунок підбирає рекомендації товарів, фільмів, музики чи книг, базуючись на попередньому виборі та даних сотень інших користувачів [13, 14, 15, 16].

1.3.5 Задача виявлення викидів

Викиди визначаються як будь-які аномалії набору даних, тобто такі точки, які не є притаманними середньостатистичній вибірці цього набору. Виявлення викидів є однією із технік, коли аналіз цілеспрямовано здійснюється на пошук елементів даних, які не відповідають очікуваній моделі поведінки. Такий аналіз є важливим, адже коли аномальні точки даних виявлені, то можна знайти та зрозуміти, що їх спричиняє і оптимізувати процеси чи майбутні рішення.

Така техніка часто використовується у сфері виявлення шахрайства – наприклад, незвичних сум покупок, часу активності, поведінки користувача. У такому випадку можливий випадок викрадення даних банківської карти, яку потрібно автоматично заблокувати. Окрім цього, якщо незвичні стрибки у кількості транзакцій чи трафіку на сайт з’являються у певний час доби, це може свідчити про атаку на сайт, щоб заблокувати доступ справжнім покупцям у пікові години [13, 14].

1.3.6 Задача когортного аналізу

Когортний аналіз – це одна із технік, що стосується поведінкової аналітики, і яка використовує історичні дані для вивчення та порівняння визначеного сегменту користувачів. У такій аналітиці вибирається деяка підмножина даних із заданого набору, і алгоритм розбиває її на споріднені групи (когорти) для аналізу. Зазвичай, ці групи мають певні спільні характеристики, які спостерігалися впродовж певного періоду часу. Використовуючи цю техніку аналізу, можна отримати ширше розуміння споживачів, їх потреб, особливостей та відмінностей.

Також можна сказати, що поведінка аудиторії вивчається у контексті «подорожі клієнта», і у результаті отримуються моделі дій користувача на різних етапах подорожі – наприклад, від першого відвідування веб-сайту до покупки. Окрім цього когортний аналіз часто використовують у маркетингу,

адже він допомагає визначити вплив промоційних кампаній на цільові групи клієнтів. Динамічне відслідковування реакції клієнтів на різного роду підходи до комунікації, є ще одним прикладом аналітики [11, 12].

1.3.7 Задача факторного аналізу

Факторний аналіз – це техніка аналізу, що використовується для опису мінливості між корельованими чи спостережуваними змінними з точки зору невеликої кількості інших змінних, які називаються факторами. Окрім цього, факторний аналіз часто використовують для зведення великої кількості змінних до меншої кількості факторів, тобто для визначення тих змінних, які мають найвищий вплив на стан набору даних. Метою такого аналізу є як й зменшення кількості атрибутів, так і виявлення незалежних прихованих змінних чи шаблонів. Існує два типи факторного аналізу – дослідницький (використовується, якщо структура чи кількість вимірів у наборі даних невідома) та підтверджуючий (служить для перевірки гіпотез щодо структури даних) [11, 12].

1.3.8 Задача аналізу часових рядів

Аналіз часових рядів – це статистичний прийом, що використовується для виявлення тенденцій у часі. Часовий ряд – це послідовність точок даних вимірів одної і тої ж змінної в різні моменти часу. Аналізуючи тенденції зміни даних з часом, можна прогнозувати майбутні тренди. При аналізі часових рядів основними закономірностями, на які звертається увага, є тенденція (стабільне лінійне збільшення або зменшення значення змінної), сезонність, циклічність. Аналіз та прогнозування часових рядів використовується в різних галузях, найчастіше для аналізу фондового ринку, економічного прогнозування та прогнозування продажів [12].

Окрім вищезгаданих методів та технік видобутку даних існує також незліченна кількість інших підходів, серед яких можна визначити й аналіз

тексту, дискримінантний аналіз, нейронні мережі, дерева рішень, машини опорних векторів, нечітка логіка, еволюційне програмування, метод Монте-Карло, відслідковування шаблонів та послідовностей, прогнозування, статистичні техніки, візуалізація тощо [11, 12, 13, 14].

1.4. Аналіз наукових праць щодо аналізу даних

Окрім власне практичного застосування у сфері інформаційних технологій та бізнесу, аналіз даних та бізнес-аналітика також є часто обговорюваними темами в академічних колах. Про це свідчить значна кількість наукових статей та виступів на конференціях, пов’язаних зі сферами великих даних, аналізу, машинного навчання, оптимізації бізнес-процесів тощо. Дослідження науковців усього світу стосуються як і загальних теоретичних засад аналізу, так і огляду необхідності та важливості проведення аналітики бізнесу задля підвищення ефективності.

Наукові статті щодо загальних засад аналізу даних почали здобувати популярність на початку 2010-х років. Еволюція бізнес-аналітики, її основні концепції та етапи, а також огляд різних видів аналізу, галузей та можливостей застосування було розглянуто у роботі [17] Чженом, Чжіангом та Сторі. Також ними було виокремлено важливість освіти бізнес-аналітика, основні необхідні навички та можливість створення університетських програм з аналізу даних для бізнесу, адже на їх переконання сфера аналітики повинна була стати однією із найбільш популярних сфер у наступні десятиліття.

Цікавою є стаття науковців з Вашингтонського університету Каніта Воугсуфасавата, Янґа Луї та Джефрі Гіра [18] щодо використання дослідницького аналізу у академічних та бізнес потребах. Також у статті наводяться найчастіші робочі проблеми аналітиків, наприклад щодо неповноти даних, а також підходи та інструменти, які спеціалісти

використовують у своїй щоденній роботі. Власне види та методи аналізу даних, а також техніки збору даних ґрунтовно розглянуто у роботах авторів Джоела Ашірвадама [19] та Кабіра Мухамеда [20].

Оскільки впровадження аналізу даних та бізнес-аналітики в процеси організації вважається одним із кроків на шляху цифровізації підприємства, то вивчення понять цифрових технологій також перетворилося у популярну тему досліджень останніх років. Важливість цифрової освіти, обізнаності та вмінь для підприємництва було розглянуто у роботі американських вчених Раяна Янґа, Люка Волберга, Елайни Девіс та Кавеха Абхарі [21]. В той же час, німецькі автори Пітер Бікан та Александр Брем у [22] провели дослідження цифрових технологій та створення фреймворку «цифровізації» на основі опитування підприємців різних галузей Німеччини.

Приклади аналізу даних у різних сферах та спрямуваннях є найбільш поширеною тематикою наукових досліджень, що стосуються аналітики та великих даних. Такі роботи охоплюють як загальні методи аналізу даних, такі як описовий, прогностичний чи настановний, так і сучасні техніки видобутку даних. Також цікавим є те, що такі дослідження проводяться на різних континентах, включаючи Європу, Азію, Африку та Північну і Південну Америки. Наприклад, у роботі колективу вчених з Китаю, Японії та Південної Кореї [23] було проведено дослідницький аналіз середовища стартапів цих країн у порівнянні з ситуацією у США. Аналіз проводився щодо загальних геополітичних факторів країн, фінансування, культурних відмінностей, а також порівняння агломерацій найбільших центрів розвитку. Дослідницький аналіз компаній роздрібного продажу у Великобританії та використання ними великих даних для роботи з оцінками користувачів було описано Марцелло Маріані та Самуелем Вамбою у [24]. Статистичний аналіз даних опитування щодо важливості підприємницької освіти та основних факторів впливу було проведено у Чехії Беатою Гавуровою та її співавторами у [25].

Оскільки досить популярною сферою аналітики є маркетинг та торгівля, багато статей присвячено аналітиці з цієї тематики. Наприклад, у роботі [26] Андреа Сестіно розглянув використання прогностичного аналізу для створення більш ефективних маркетингових планів та автоматизації процесів через профілювання користувачів. Також у роботі [27] було розроблено аналітичну модель бізнес-аналізу та оцінки споживацької поведінки клієнтів для подальшого прийняття рішень на основі таких даних. Ґрунтовною є монографія американських вчених Янґи Пан та Ґеррі Рассела [28] щодо аналізу даних у роздрібній торгівлі через аналітику споживчого кошика. Науковцями було проведено емпіричний аналіз та розроблено моделі короткострокових та довгострокових споживчих кошиків на основі математичних моделей та відмінностей між імпульсивними та запланованими покупками.

У дослідженні на основі технік видобутку даних значна роль відводиться підходу асоціативних правил на прикладі аналізу ринкових споживчих кошиків. Роботи, пов’язані з таким типом аналізу, найчастіше базуються на розробках нових алгоритмів побудови асоціацій, порівнянні вже існуючих, або їх застосуванні на нових наборах даних для отримання цінних ідей. Прикладами таких статей є [29], [30] та [31], які представляють науковців Індії, Австрії та Індонезії. Окрім цього у торгівлі цікавим є визначення викидів по продажах, що було досліджено Мерісою Голіч, Еміром Жунічем та Дженаною Донко на прикладі мережі магазинів у Боснії та Герцоговині [32].

Прогностичний аналіз щодо використання класифікації, кластеризації та регресії також є популярною тематикою наукових робіт останніх років. У роботі [33] колектив авторів з Південної Кореї розглянув алгоритм класифікації зображень забруднення побережжя через надмірний туризм за допомогою використання нейронних мереж. Бізнес-сектор досліджень та розробки у Румунії було описано у роботі [34] за допомогою факторного

аналізу та логістичної регресії. Логістична регресія, кластеризація та кореляційний аналіз також були використані Еммою Вуд у [35] для аналізу бізнес ефективності малих підприємств у Великобританії. Поєднання кластеризації як техніки видобутку даних з бізнес-моделлю Canvas було запропоновано колективом німецьких авторів у [36] для кластеризації топ-40 авіаперевізників світу. Окрема техніка класифікації даних, а саме, дерева рішень, використовувалася Коеном Ванхуфом та його співавторами у [37] для дослідження лояльності клієнтів до маркетингових кампаній з використанням електронних листів.

Очевидно, що згадані наукові статті становлять надзвичайно малу частку від усіх робіт з тематики аналізу даних чи використання певних технік для проведення аналітики бізнес-процесів або ж пов’язаних задач. Проте, наявність такого різноманіття у дослідженнях, а також постійне зростання кількості робіт та конференцій з тематики свідчить про актуальність та популярність тематики серед науковців усього світу.

1.5 HIT-індекс як методика визначення цифрової зрілості підприємства в Україні

Використання аналізу даних та інших цифрових інструментів у бізнес-процесах чи управлінні є значним кроком у напрямку цифровізації економіки та підприємництва. Зважаючи на досвід країн Європи та США, а також на популярність тематики цифрового суспільства, визначення рівня використання цифрових технологій у підприємництві та його впливу на розвиток бізнес-структури є актуальною темою досліджень.

Цифровізація – це термін, що використовується на позначання цифрової трансформації суспільства та економіки. Він описує перехід від індустріальної епохи до епохи знань і творчості, що характеризується цифровими технологіями та цифровими бізнес-інноваціями. Іншими словами

– це впровадження цифрових технологій в усі сфери життя: взаємодію між людьми, предмети побуту, промислові виробництва, сферу послуг тощо.

Однією із важливих частин цифровізації є перехід до цифрової економіки, де ключовими засобами виробництва є використання цифрових даних як ресурсу, через що можливо підвищити ефективність та продуктивність роботи, а також збільшити цінність наданих послуг та товарів. Згідно з заявою Міністерства цифрової трансформації України, цифрова економіка є одним із векторів розвитку у розрізі економічної стратегії 2030 року. Серед основних напрямків роботи виділятимуть розвиток цифрової інфраструктури та навичок, розбудову сектору інформаційно-комунікативних технологій та цифровізацію сфер життя та секторів економіки [38].

Щодо ситуації в Україні та світі, то на даний час активно відбувається перехід промислової економіки та інформаційного суспільства до понять «цифрової» економіки. Даний процес прийнято називати цифровою трансформацією бізнес-структури – тобто перетворення її бізнес-стратегії, моделей, операцій, цілей, маркетингових підходів тощо у напрямку збільшення використання цифрових технологій та підвищення ефективності. Проте, в Україні такий процес розпочався досить нещодавно, і багато бізнес-структур, особливо із категорій малих та мікропідприємств (у яких працює не більш, ніж 50 співробітників, та з невеликими активами), не володіють достатньою інформацією щодо динамічних процесів, які відбуваються в економіці. Проблемою є відсутність у підприємців необхідних знань щодо застосування інноваційний цифрових технологій Також недостатньо поширеними є й платформи, сервіси чи додатки, які б доступно пояснювали важливість та можливості використання цифрових інструментів у перетворенні бізнесу.

Невисока проінформованість малого та середнього підприємництва щодо можливостей інтеграції технологій у власні бізнес-процеси спричиняє

гальмування розвитку компаній та виникнення труднощів у виході вітчизняного бізнесу на міжнародну арену. Через це особливої уваги заслуговують дослідження щодо розробки систем індикаторів цифрової трансформації бізнесу, забезпечення проведення регулярних оцінювань цифрового розвитку та запровадження регулярних, систематичних статистичних спостережень [39, 40, 41].

У дисертаційному дослідженні [42] було запропоновано створення методики визначення індексу цифрової трансформації бізнесу, що дозволяє оцінювати рівень цифрової зрілості бізнес-структури та враховувати фактори, які потрібно покращувати та розвивати як на рівні окремої бізнес-одиниці, так і національної стратегії. Скорочено даний індекс цифрової трансформації позначатиметься HIT – абревіатура від його основних компонентів: Human, Instruments, Technology (Людина, Інструменти, Технології). Для його безпосереднього розрахунку використовуються такі індикатори, наведені на рисунку 1.1.

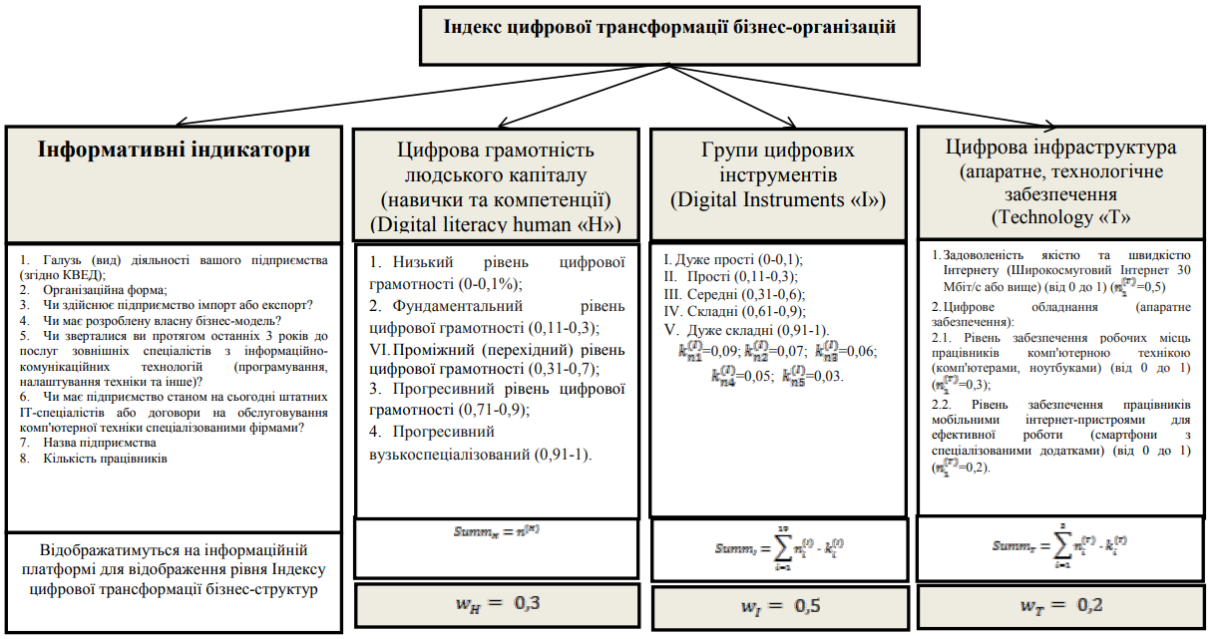


Рисунок 1.1 - Групи індикаторів для визначення Індексу цифрової трансформації бізнес-структур (запозичено з [42])

Розглянемо детальніше основні індикатори, описані у роботах [39-41]:

- Показники цифрової інфраструктури підприємства, що описують рівень її забезпеченості необхідним обладнанням (персональні комп'ютери, ноутбуки, смартфони) та широкосмуговим Інтернетом.
- Використання цифрових інструментів є ключовим показником, що представляє якісні характеристики ефективності технологій у бізнесі. Цей показник включає такі компоненти, як використання управління соціальними медіа, функціонування сайту та пошукової оптимізації, робота зі спеціалізованими системами автоматизації бізнес-процесів тощо;
- Цифрова грамотність (компетентність) людського капіталу, яка визначається як здатність працівника виконувати складні завдання та вимоги, що передбачають як професійні, так і особистісні навички.

На основі зведених структурних показників цифрової трансформації бізнес-організацій було сформовано формулу (1.1) для визначення узагальненого Індексу цифрової трансформації бізнесу [39, 40, 41, 42, 43]:

$$HIT = Sum_H * \omega_H + Sum_I * \omega_I + Sum_T * \omega_T \qquad (1.1)$$

де HIT – індекс цифрової трансформації бізнесу,

Sum_H – зведений показник стану цифрової грамотності людського капіталу організації,

ω_H – ваговий коефіцієнт індикатора цифрової грамотності H,

Sum_I – зведений показник стану функціонування цифрових інструментів інтегрованих в бізнес-процеси організації,

ω_I – ваговий коефіцієнт індикатора цифрових інструментів I,

Sum_T – зведений показник стану функціонування цифрової інфраструктури бізнес-організації,

ω_T – ваговий коефіцієнт індикатора цифрової інфраструктури T.

Причому, сума вагових коефіцієнтів дорівнює одиниці, як наведено у формулі (1.2).

$$\omega_H + \omega_I + \omega_T = 1 \quad (1.2)$$

Також у [42, 43] було визначено, що ваговий коефіцієнт цифрової грамотності становитиме 0.3 ($\omega_H = 0.3$), ваговий коефіцієнт функціонування цифрових інструментів дорівнюватиме 0.5 ($\omega_I = 0.5$), а цифрова інфраструктура підприємства оцінюватиметься у 0.2 ($\omega_T = 0.2$).

Зведений показник за кожним з індикаторів (цифрова грамотність, використання інструментів, інфраструктура) обчислюється за формулою (1.3) і може лежати у інтервалі від 0 до 1:

$$Sum_X = \sum_{i=1}^{m_X} n_i^{(X)} * k_i^{(X)} \quad (1.3)$$

де Sum_X – зведений показник того чи іншого індикатора (H, I, або T),

m_X – кількість елементів індикатора у наборі даних,

$n_i^{(X)}$ – показник рівня функціонування індикатора в організації,

$k_i^{(X)}$– ваговий коефіцієнт індикатора.

Зважаючи на значення вагових коефіцієнтів зведених показників індикаторів, а також на те, що кожен із них належить інтервалу [0;1], можемо зробити висновок, що й значення індексу трансформації належатиме тому ж інтервалу. Окрім цього було визначено рівні цифрової зрілості, де значення HIT-індексу в інтервалі [0; 0.2) вважається дуже низьким рівнем; [0.2; 0.4) – низьким; [0.4; 0.6) – середнім; [0.6; 0.8) – високим; та [0.8; 1] – дуже високим.

1.6 Висновки до першого розділу

У першому розділі кваліфікаційної роботи розглянуто поняття аналізу даних та його споріднених видів. Окреслено загальні підходи до аналізу даних, такі як прогностичний, описовий та діагностичний. Коротко розглянуто основні типи задач видобутку даних, серед яких кластеризація, класифікація, регресійний та факторний аналіз, пошук асоціативних правил тощо.

Проведено аналіз наукових праць дослідників щодо питання цифровізації бізнесу та використання технік аналізу даних для вирішення певних практичних задач у бізнесі на прикладі різних країн світу. Можна зробити висновок, що тематика досліджень є популярною та найчастіше покриває задачі пошуку спорідненості, асоціативних правил, споживчого кошика, кластеризації, класифікації тощо.

Розглянуто поняття цифровізації економіки та одне з українських досліджень щодо визначення рівня цифрової трансформації бізнес-структури як методики спрямованої на розвиток використання цифрових інструментів підприємництвом України. Важливою є допомога у трансформації саме для мікро, малого та середнього бізнесу з огляду на їхні власні ресурси та можливості. Запропонований у дослідженні індекс цифрової трансформації буде використовуватися як основа обчислень у подальших розділах кваліфікаційної роботи.

2 КЛАСТЕРИЗАЦІЯ ЯК ОДНА ІЗ ЗАДАЧ ВИДОБУТКУ ДАНИХ

У сучасному світі речі часто можна віднести до визначених категорій, наприклад, «А» та «не-А» (твердження і його заперечення). Означення групи завжди чітке, проте не у всіх наборах даних такі сегменти визначені від початку. Вибір характеристик груп, які лежать в основі подібності між елементами сегменту та їх відмінностями від інших елементів називається кластеризацією [44].

Іншими словами, кластерний аналіз – це завдання групування набору елементів чи точок даних таким чином, щоб об’єкти в одній групі були більш схожими один на одного, ніж на елементи інших груп. З цього випливає, що кластери формуються так, щоб відстань між точками даних у кластері була мінімальною, тобто точки розташовуються дуже щільно. В той же час, найкращим результатом буде, коли відстань до іншого найближчого кластера буде максимальною, а самі кластери будуть розташовані досить розріджено на площині даних. Кластеризацію також часто називають стисненням даних, неконтрольованою сегментацією, типологічним аналізом, автоматичною класифікацією та виявленням групування [45, 46, 47].

Окрім цього, кластеризація – це один із методів неконтрольованого машинного навчання, також відомого, як навчання «без вчителя». У такому разі алгоритм навчається на наборах даних, які не містять позначок вихідної (результуючої, цільової) змінної. Узагалі, кластерний аналіз не є одним конкретним алгоритмом, який підходить під усі типи задач. Це радше, загальне завдання, яке потрібно вирішити, тому іноді кластеризацію ще називають багатоцільовою проблемною оптимізації. За такого визначення обраний алгоритм вирішення задачі та його параметри (функція відстані, поріг щільності, кількість кластерів та інші) залежать від кожного окремого набору даних та завдання. Тому не можна сказати, що один алгоритм кластеризації з одним і тим же набором параметрів показуватиме однакові

результати на різних даних. З цього можна зробити висновок, що кластеризація є ітеративним процесом, і її алгоритми та параметри можуть та повинні вдосконалюватися з кожним кроком, поки результат не досягне бажаних оцінок якості [47].

Успішне групування визначається двома основними цілями: досягнення подібності між близькими точками даних та досягнення відмінності від тих точок, які розташовані далі. Одним із викликів кластеризації можна визначити завдання масштабування даних. Окрім цього, дані містять атрибути різних видів: категорійні, числові, безперервні тощо, з чого випливає виклик багатовимірності інформації та різного роду її представлення, адже алгоритм кластеризації повинен мати змогу працювати з різними типами даних. Також кластери обмежені геометричною формою, і включення граничних точок до тої чи іншої групи є проблемою, яку потрібно вирішувати. Часто під час роботи виявляється, що дані містять в собі змінні, які є шумом чи завадами і впливають на формування груп даних. У кінці-кінців, кластеризація повинна відповідати бізнес-цілям та потребам, що також може сприйматися як виклик у виборі правильного алгоритму для того чи іншого завдання [47].

Отже, хороший алгоритм кластеризації повинен враховувати можливість масштабування даних, адаптивність до роботи з різними типами атрибутів та з багатовимірними даними, а також стійкість до шуму та можливих викидів даних, які не повинні впливати на якість кластеризації.

2.1 Основні підходи та методи кластеризації даних

Оскільки поняття «кластер» не є чітко визначеним з методологічної точки зору, а також через різноманіття форматів, типів, способів отримання даних існує надзвичайно велика кількість алгоритмів кластеризації. Кожен із них оперує різними підходами до кластеризації, побудови геометричної

форми, визначення центрального елемента, розрахунку відстані, включення чи виключення точок у кластерів та й власне поняттям кластеру. Проте, дослідники виокремили декілька моделей кластеризації, які поверхнево описують основні ідеї, закладені у алгоритмах. Типові моделі кластеризації включають в себе:

- Зв’язні моделі, наприклад, ієрархічні, у яких кластеризація будує моделі на основі відстані елементів один від одного.
- Центроїдні моделі, що представляють кластери вектором або ж точкою середнього значення.
- Моделі на основі розподілу, у яких кластери моделюються за допомогою статистичних розподілів.
- Моделі на основі щільності, які визначають кластери як пов’язані щільні області в просторі даних і часто використовуються у пошуку аномалій.
- Моделі на основі графів, у яких підмножину вузлів графа, з’єднаних ребрами між собою, можна розглядати як прототип кластера.

Окрім цього моделі кластеризації можна розділити на жорсткі та м’які. При жорсткій кластеризації кожен об’єкт однозначно належить або не належить до того чи іншого кластера. При м’якій – або ще, нечіткій – кластеризації, об’єкт належить до кожного кластеру з певною ймовірністю.

2.1.1 Кластеризація на основі зв’язків, ієрархічна кластеризація

Ієрархічна кластеризація – це метод, який визначає кластери на основі близькості точок даних. Дослідник приймає тезу, що наближені точки даних мають характеристики більш подібні між собою. Розподіл даних виконується для злиття точок у кластери на певних відстанях, що дозволяє отримати ієрархію кластерів. Візуально такі методи часто репрезентуються у вигляді дендрограми, у якій коренем є один кластер, що включає в себе усі зразки, а листками – безліч кластерів, що складаються лише з одного елемента. Такі моделі легко візуалізувати та інтерпретувати, але вони досить слабко

масштабуються. Існує два основних підходи до ієрархічної кластеризації – агломеративний (знизу-догори) та дивізивний (зверху-униз) [44, 48]. Розглянемо їх детальніше.

Дивізивний підхід ґрунтується на тому, що спочатку існує єдиний кластер, який вміщує в собі усі точки даних. Після чого алгоритм розбиває цей кластер на дрібніші, і кожен екземпляр вхідних даних присвоюється певному кластеру на основі найближчої міри відстані, визначену як попарну відстань між точками. Такими відстанями можуть бути відстань Уорда, центроїдна відстань, одинарні чи усереднені зв’язки тощо.

Ідеальне застосування дивізивного алгоритму закінчується тим, що кожна точка даних належить власному єдиному кластеру. Проте, часто такий результат не є задовільним для практичних проблем. Тому за допомогою дендрограми чи інших метрик визначення кількості кластерів можна зупинити кластеризацію при досягненні визначеної кількості кластерів або метрики якості [44, 47].

Оскільки на початку роботи алгоритму усі дані знаходяться в одному кластері, то існує $O(2^n)$ способів поділу кластерів. Розглянемо схематичний алгоритм роботи дивізивного підходу [49]:

1. На вхід подається набір даних (d_1, d_2, d_3, ...d_N) розмірності N. Всі точки даних знаходяться в одному кластері.

2. Кластери розбиваються на 2, використовуючи вибраний лінійний алгоритм кластеризації (наприклад, К-середніх).

3. Обирається «найкращий» кластер для наступного розбиття.

4. Кластер розбивається на 2 за лінійний алгоритмом.

5. Послідовність повторюється допоки кожна точка даних не належить окремому кластеру.

На противагу цьому, агломеративна кластеризація базується на підході знизу-догори: тобто, алгоритм розпочинається з прийняття того, що кожна точка даних є окремим кластером. Тобто, якщо набір даних містить N точок

даних, то ми працюватимемо з N кластерами. Після чого кластери, які є найбільш подібними, ітеративно комбінуються по двоє у більший кластер.

Мірою подібності кластерів виступає метрика відстань між ними. Та пара кластерів, у якої найбільша міра подібності – які знаходяться найближче один від одного, зливаються в один кластер [44, 48, 50].

Такі ітерації повторюються до ідеального завершення – коли алгоритм досягне кореня дерева, тобто коли залишиться лише один кластер з усіма точками всередині. Як і у випадку з дивізивним підходом, часто можна обрати кількість кластерів, яку потрібно досягти, щоб зупинити побудову дерева для вирішення практичних завдань. Схематичний алгоритм агломеративного підходу наступний [49]:

1. На вхід подається набір даних (d_1, d_2, d_3, ...d_N) розмірності N. Кожна точка є окремим кластером
2. Обчислюється матриця відстаней між кожною парою точок; Оскільки точки симетричні, можна обчислювати лише верхню або нижню половину матриці.
3. Два кластери з мінімальною відстанню зливаються в один.
4. Матриця відстаней перераховується.
5. Кроки повторюються доки всі точки не належатимуть єдиному кластеру.

Хоча ієрархічна кластеризація не вимагає вказання кількості кластерів розбиття, ми можемо обрати таке число для того, щоб зупинити кластеризацію за досягнення певного критерію. Окрім цього, алгоритм не є чутливим до вибору метрики відстані – більшість відомих метрик добре працюють для такого типу кластеризації [50]. Проте, стандартний алгоритм ієрархічної агломеративної кластеризації має часову складність у $O(n^3)$ і це вимагає розміру пам’яті у $O(n^2)$, що робить такий підхід повільним навіть для середніх наборів даних через те, що на кожному N-1 кроці обчислюється матриця відстані NxN. Для деяких випадків з використанням міри

одинарного або повного зв’язку, часова складність може зменшитися до $O(n^2)$.

Згадуючи про критерії зв’язності, варто зазначити, що вони визначають метрику, яка використовується при злитті кластерів. Іншими словами, критерії зв’язності визначають відстань між набором даних як функцію попарної відстані між точками. Деякими найбільш відомими критеріями зв’язності є [48, 47]:

1. Критерій Уорда, що мінімізує суму квадратів різниць між точками в усьому кластері, тобто $\min\{\|a_i - a_j\|^2 : a_i, a_j \in A\}$.

2. Критерій мінімізації, або ж одинарного зв’язку, який мінімізує функцію відстані між найближчою парою кластерів, тобто $\min\{d(a,b) : a \in A, b \in B\}$. Тобто відстань між кластерами визначається як відстань між найближчими точками усередині.

3. Критерій максимізації, або ж повного зв’язку, який максимізує функцію відстані між парою кластерів, тобто $\max\{d(a,b) : a \in A, b \in B\}$. Відстань визначається між парою точок з різних кластерів, які знаходяться якнайдалі.

4. Критерій середньої зв’язності, що мінімізує середню відстань між усіма парами кластерів, тобто $\min\{\frac{1}{|A|\cdot|B|}\sum_{a\in A}\sum_{b\in B} d(a,b)\}$

До переваг ієрархічних методів кластеризації найчастіше відносять те, що вони не вимагають попередніх відомостей про кількість необхідних кластерів, проте у реальних задачах користувачу все ж потрібно визначити кількість груп для поділу. Також алгоритм є легким для імплементації та підходить для використання різних типів та видів даних, що робить його застосовним для широкого спектру задач [44]. Якщо згадувати сфери застосування ієрархічних алгоритмів то це філогенетика та розшифровка ДНК, кластеризація текстів та виокремлення фейкових новин чи

повідомлень, персоналізація реклами у соцмережах, класифікація мереж та кластеризація вхідного трафіку для пошуку потенційних проблем тощо.

2.1.2 Кластеризація на основі центроїдів. Алгоритм К-середніх

Кластеризація на основі центроїдів вважається одним з найпростіших та найбільш ефективних методів створення кластерів. Її ідея полягає в тому, що кластер характеризується центральним вектором, а точки даних, які знаходяться у безпосередній близькості до такого вектору, відносяться до його кластеру. Іншими словами, центр кластера сформований так, що відстань точок у кластері до його центру – мінімальна. Основним викликом є те, що досліднику потрібно інтуїтивно або науково визначити кількість кластерів, щоб розпочати роботу алгоритму. Тому, зазвичай, оптимальна кількість кластерів для кожного окремого випадку та алгоритму визначається експериментально.

Метод К-середніх (K-means) є одним із найбільш відомих, використовуваних та простих алгоритмів навчання «без вчителя», який використовується для вирішення завдань кластеризації. Алгоритм К-середніх розділяє вхідні дані набору (X) на певну кількість (К) кластерів (С), які не перетинаються та мають приблизну рівну дисперсію всередині. Кожен кластер характеризується середнім значенням своєї вибірки μ_i – тобто, його центроїдом. Важливо зазначити, що центроїди, в загальному, не є точкою вхідного набору даних, хоча й можуть співіснувати у його просторі [48].

Кожна ітерація алгоритму К-середніх прагне мінімізувати певну цільову функцію, у цьому випадку – функцію квадратичної помилки F(V), яку ще називають інерцією і яка наведена у формулі (2.1):

$$F(V) = \sum_{i=1}^{C} \sum_{j=1}^{C_i} (\|X_i - V_j\|)^2 \qquad (2.1)$$

де $\|X_i - V_j\|$ – це відстань між точками даних X_i та Y_j,

C_i – це кількість даних у кластері,

C – кількість центроїдів кластерів.

Оскільки К-середніх є одним із класичних алгоритмів кластеризації, варто розглянути кроки його роботи детальніше [44, 45, 48, 51]:

1. Для початку, визначається К – кількість кластерів чи груп, на які потрібно розділити набір даних. Для оптимального вибору числа добре знати внутрішню структуру або ж виконувати послідовний аналіз різною кількістю кластерів та вимірювати їх якість.

2. Першим кроком власне кластеризації є випадкова ініціалізація місця розташування центроїдів. Очевидно, що центроїд буде вектором тої ж вимірності, що й точки набору даних.

3. Далі кожна вхідна точка класифікується до певного кластеру через обчислення відстані між цією точкою та центром, після чого точка відноситься до того кластеру, до центру якого вона знаходиться найближче.

4. Коли всі точки присвоєні кластерам, алгоритм перераховує центри утворених кластерів з урахуванням точок даних і переміщає центроїди у ці точки.

5. Після визначення нових центроїд, усі точки вхідного набору даних переприсвоюються новим кластерам через перераховування відстаней між точкою і центром, як і у кроці 3.

6. Така послідовність кроків повторюється визначену кількість ітерацій, або до того моменту, поки центри кластері залишаться сталими протягом двох ітерацій.

До переваг методу найчастіше відносять те, що алгоритм може використовуватися для різних типів та видів даних, проте його основною вимогою є те, що дані повинні бути кількісними. Окрім цього алгоритм є швидким для виконання на даних невисокої вимірності, що є важливими фактором у роботі системи, а його часова складність є лінійною, тобто - $O(n)$, що робить алгоритм простим для розуміння та застосування [44, 50, 51]

В той же час, до недоліків алгоритму часто відносять те, що алгоритм К-середніх не підходить для нелінійних даних різної розмірності як у сенсі розкиду значень, так і у сенсі вимірності. Також підхід не може використовуватися для категорійних даних, адже базується на відстані між зразками, яку важко визначити для не числової інформації. Для роботи методу потрібно наперед визначити кількість кластерів, на які розбиватиметься набір даних – коректне визначення цього числа вимагає знання структури даних, або ж використання додаткових математичних та евристичних методів, адже кожне виконання алгоритму надає центри кластерів випадковим точкам, що може впливати на кінцевий результат. Також використання K-Means припускає, що кластери є опуклими (тобто мають правильну геометричну форму) [48].

Окрім алгоритму К-Means визначають також його менш поширені варіації, такі як K-Modes та K-Medians. Принцип роботи у цих алгоритмів аналогічний до вищезгаданого, єдиною різницею є те, що у K-Modes замість центроїд кластерів використовуються їх моди (найпоширеніші значення), а у K-Medians – медіанний вектор кластера [50, 51]. Разом із основним алгоритмом К-Середніх, вони використовуються у кластеризації документів, банківській справі та страхуванні, сегментації зображень або клієнтів.

2.1.3 Кластеризація на основі щільності. Алгоритми DBSCAN та OPTICS

Розглядаючи попередні підходи до кластеризації, можна було помітити, що однією із основних характеристик кластерів та власне алгоритмів є поняття відстані та подібності між точками, адже саме за таким атрибутом і будуються групи об’єктів. На противагу їм, методи кластеризації на основі щільності працюють із поняттям щільності розподілу даних замість відстаней між точками. У таких алгоритмах кластери створюються з максимального набору пов’язаних між собою щільно розташованих точок даних, і вони

розглядаються як найщільніша область даних у просторі, яка відокремлюється від інших менш щільними областями. Розріджені області з меншою кількістю точок даних можуть розглядатися як шум або ж викиди. Алгоритми кластеризації працюють на пошук просторів різної щільності і на основі такого аналізу виділяє різні області щільності, що відповідатимуть окремим кластерам. Зважаючи на це, логічно припустити, що такі кластери можуть мати довільну форму, не обмежуючись кругом чи еліпсом [44, 45, 46].

DBSCAN – алгоритм просторової кластеризації зразків з шумом на основі щільності. DBSCAN розглядає кластери як області високої щільності, розділені між собою областями з низькою щільністю. Групування точок даних відбувається на основі метрики відстані та критерію мінімальної кількості точок даних [44, 45, 48].

Розглянемо алгоритм детальніше. Для початку необхідно визначити два основних параметри: радіус кластера ε та мінімальної кількості точок, що входитимуть до області – m. Епсилон (ε) вказує, наскільки близькими повинні бути точки даних, щоб вважатися сусідами, а значення мінімуму точок (m) повинне бути досягнуто, щоб регіон вважався щільним. Наступні кроки алгоритму такі [44, 50, 51]:

1. DBSCAN розпочинається з довільної точки початкового набору даних. Маючи обране значення радіусу ε, визначається окіл такої точки. Усі інші точки даних, що входять до околу визнаються сусідами.

2. Якщо в такому околі є достатня кількість точок, більша або рівна мінімальному значенню m, то розпочинається процес кластеризації. При чому поточна точка даних стає першою точкою в новому кластері. Якщо кількість точок в околі недостатня, вона позначається як викид, проте у подальшому може увійти до іншого кластера.

3. За прикладом першої точки у кластері, точки в її околі є також стають частиною кластера. Процедура присвоєння всіх точок ε-околу до того ж кластеру повторюється для кожної нової точки, доданої до кластеру. Тобто,

для кожної нової точки визначається ε-окіл, точки якого входять до поточної групи. Цей крок циклічно повторюється до тих пір, поки не залишиться нових точок даних в околах кожної присвоєної точки.

4. З виходом з поточного кластеру, обирається нова довільна «вільна» точка даних, яка утворює ядро нового кластеру, після чого повторюються кроки 2-3.

5. Процес повторюється поки всі точки даних не будуть перевірені і або віднесені до певного кластеру, або визначені як викиди.

Вибір параметрів ε та m дозволяє контролювати щільність, необхідну для формування групи. Мінімальна кількість зразків впливає на те, наскільки алгоритм буде сприйнятливим до шуму та викидів, а радіус контролюватиме сусідство точок. Такі параметри не потрібно залишати за замовчуванням. З цього можна зробити висновок, що більший мінімум зразків або менший радіус вказують на вищу щільність кластеру. Неправильний вибір параметрів може призвести до того, що набір згрупується в єдиний кластер, або ж що всі точки будуть позначені викидами [48].

Перевагами методу є те, що він не потребує вказання кількості груп, легко працює з викидами, а також не залежить від форми кластеру. До недоліків відносять неможливість працювати з групами різної щільності або надто розрідженими даними, а також чутливість до параметрів. Часова складність алгоритму дорівнює $O(n \log n)$. Сферами використання є ідентифікація плагіату, рекомендаційних системах, рентгенівській кристалографії та аналізі кластерів у соціальних мережах [44].

Ще одним алгоритмом є OPTICS – метод, заснований на щільності, який працює за принципом сходим на DBSCAN, але усуває один із його основних недоліків, а саме неможливість роботи з даними довільної щільності. Окрім мінімальної кількості зразків у кластері m та радіуса околу ε, цей алгоритм враховує ще 2 параметри: відстань від ядра та відстань доступності.

Відстань від ядра показує, чи є поточна точка даних ядром кластера, чи ні, шляхом встановлення для неї мінімального значення параметру ε з урахуванням мінімальної кількості точок m. На противагу їй, відстань доступності – це максимум з відстані ядра точки P та значенням метрики відстані, що використовується, між двома точками P та Q. Іншими словами, це відносна досяжність точки Q відносно точки P [51, 52].

2.1.4 Кластеризація на основі розподілу. Алгоритм очікування-максимізації

Окрім попередньо згаданих методів, існує також сімейство підходів до кластеризації, заснованих на метриці ймовірності. У такому разі алгоритм розподілу створює та групує точки даних за аналізом їх ймовірнісної приналежності до розподілу ймовірностей за певною моделлю. Ймовірнісна модель може бути будь-якою статистично відомою, проте найчастіше використовується Гаусовий або біноміальний розподіли. Очевидно, що моделі кластеризації за розподілами тісно пов’язані з статистикою та тим, як набори даних генеруються та впорядковуються, підпорядковуючись принципам випадкової вибірки.

Такий тип кластеризації має значну перевагу перед методами кластеризації за відстанями з точки зору гнучкості формування кластерів та їх фінальної форми. Проте, основною проблемою та викликом є те, що такі методи кластеризації добре працюють на синтетичних та імітованих даних, або таких, про які ми можемо наперед сказати, що вони більш-менш точно описуються якимось відомим розподілом. Якщо ж ні, і статистичний розподіл застосовуватиметься до даних, які розсіяні зовсім не так, результати будуть не надто точними. Тобто, часто існуватиме досить неоднозначне припущення щодо розподілу набору даних, адже для роботи алгоритму необхідна визначена математична модель. Один із популярних прикладів таких методів – алгоритм очікування-максимізації (Expectation-maximization, EM), який

використовує багатовимірні нормальні розподіли [44, 45, 46]. Такий тип кластеризації можна описати наступними кроками:

1. Алгоритм розпочинається з вибору кількості кластерів та випадкової ініціалізації параметрів розподілу Гаусса для кожного кластера.

2. Враховуючи задані розподіли, обчислюється ймовірність того, що кожна точка даних належить певному кластеру – чим ближче точка до центру розподілу, тим більша ймовірність того, що вона належить групі.

3. На основі ймовірностей обчислюється новий набір параметрів Гауссового розподілу таким чином, щоб максимізувати ймовірності точок даних, що входять у кластер. Параметри розраховуються з використанням зваженої суми позицій точок даних, де ваговими коефіцієнтами є ймовірності належності точки до конкретного кластеру.

4. Попередні кроки ітеративно повторюються допоки розподіли не змінюватимуться між ітераціями.

Перевагами алгоритму можна назвати те, що призначення точки кластеру визначається за зрозумілою метрикою ймовірності. До того ж, часто такі моделі припускають членство точки в кількох кластерах з різними значеннями ймовірності. Також кластери можуть набувати форми будь-якого еліпсоїда. Недоліками визначають те, що його часова складність не дозволяє використання на великих даних, а також на вибірках, які є розподіленими за Гауссом. Часова складність алгоритму становить $O(N)$ на кожну ітерацію. Сферами використання є обробка зображень та асоціативні задачі зв’язку між меншими та більшими частинами набору [44, 50].

2.1.5 Нечітка кластеризація

Розглядаючи попередні підходи до кластеризації, ми припускали, що кожна точка даних належить (виключно чи з певною ймовірністю) тільки до одного кластера. У методах нечіткої кластеризації, кожна точка даних

призначається кільком кластерам з певним кількісним ступенем приналежності. Найчастіше – це певний коефіцієнт в інтервалі від 0 до 1.

Нечітка класитеризація найкраще показує себе з даними, у яких можливий високий рівень перекриття точок даних між собою. Часто це використовується у біоінформатиці, розшифровці геномів, диференціації між пікселями зображення чи відтінками кольору [44, 45].

Алгоритм кластеризації нечіткого аналізу – Fuzzy C Means – дотримується принципу роботи, схожого до К-середніх – тобто розподілу кластерів на основі відстані. Проте, його основною відмінністю є те, що кожна точка даних може належати до більш, ніж до одного кластеру. Цей ступінь відношення до тої чи іншої групи можна виразити наступною цільовою функцією, наведеною у формулі (2.2)

$$J(X,V) = \min\left(\sum_{j=1}^{k}\sum_{x_i \in C_j} u_{ij}^{m}(x_i - v_j)^2\right) \qquad (2.2)$$

де u_{ij} – це ступінь приналежності точки даних x_i до кластера c_i,

v_i – це центр кластера j,

m – це аналізатор.

Кроки алгоритму подібні до К-середніх: спочатку визначається кількість кластерів К, після чого проводиться перерахунок відстаней та присвоєння кожної точки до кластерів з різним степенем відношення або ж ймовірності. Ітеративний процес повторюється допоки не досягається максимальне число ітерацій.

Перевагами цього методу вважають його здатність адекватно оцінювати дані, які перетинаються або сильно корелюють між собою. Окрім цього у нечіткої кластеризації вища швидкість збіжності результатів. Часова складність алгоритму дорівнює $O(NT^2)$, де N – кількість зв'язків, а T – кількість перемішувань. Недоліками методу вважають те, що хоч збіжність і завжди гарантована, але її досягнення може займати багато часу та

обчислювальної потужності для багаторозмірних даних. Також, як і більшість алгоритмів, він залежний від шуму та викидів у даних [44].

2.1.6 Кластеризація середніх зсувів

Кластеризація середніх зсувів (mean shift) є формою непараметричного підходу до кластеризації, який не тільки усуває необхідність попередньої специфікації кількості кластерів, але також усуває обмеження кластерів щодо форми та вимірності, що є основними проблемами поширених алгоритмів, таких як К-середніх.

Кластеризація середніх зсувів — це алгоритм на основі ковзного вікна, який знаходить щільні ділянки точок даних. Алгоритм побудований на основі центроїдів, тобто його мета полягає в тому, щоб знайти центральні точки кожної групи. Це працює за принципом оновлення кандидатів на центральні точки таким чином, щоб вони були середнім значенням точок у ковзному вікні. Після чого ці кандидати фільтруються, щоб усунути схожі точки та утворити остаточний набір центроїд та відповідних груп. В кінцевому підсумку кластери зміщуються в область з більшою щільністю [44, 48, 50].

Маючи кандидата у центроїд x_i для ітерації t, він оновлюватиметься відповідно до такого рівняння (2.3):

$$x_i^{t+1} = m(x_i^t) \tag{2.3}$$

де $N(x_i)$ – це округ точок на заданій відстані від x_i,

m – це середній вектор зсуву, який обчислюється для кожного центроїда і вказує на область максимального збільшення щільності точок.

Середній вектор зсуву обчислюється за допомогою наступного рівняння (2.4), фактично оновлюючи центроїд як середнє значення вибірок у його окрузі:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)} \quad (2.4)$$

де K– це центроїд регіону точок.

Якщо коротко розглянути алгоритм, то він є наступним: усе починається із ковзного вікна у формі кола з центром у випадковій точці C та радіусом R. На кожній ітерації ковзне вікно зміщується у напрямку областей двовимірного простору з більш щільним розподілом точок. У цьому разі центр вікна (точка C) зміщується у центроїдну точку вирахувану з-поміж точок набору даних, які потрапили в область дії вікна. Ковзне вікно продовжує «рухатися» допоки існуватиме напрямок, який дозволить вікну покрити більше точок. Таких ковзних вікон може бути різна кількість, що дорівнюватиме числу кластерів – точки, що потрапили під певне ковзне вікно, вважаються членами відповідного кластеру [48, 50].

Перевагами методу вважається те, що він не є параметризованим щодо кількості кластерів, а також те, що форма кластеру не обмежена колом чи овалом. На противагу цьому, недоліком визначають те, що радіус ковзного вікна – це довільна величина, яка не визначається бізнес-логікою, тому її вибір може бути нелегким завданням. Також значення часової складості алгоритму рівне $O(N^2)$. Сферами використання є сегментація зображень, комп’ютерне бачення та аналіз відеоряду [44].

2.1.7 Алгоритм Affinity Propagation – поширення спорідненості

Affinity Propagation – це алгоритм кластеризації, який не потребує попереднього визначення кількості кластерів. Якщо не заглиблюватися у деталі, в Affinity Propagation кожна точка даних «надсилає повідомлення» усім іншим точкам, інформуючи їх про відносну відстань до інших точок. Кожна ціль відповідає всім відправникам та інформує кожного відправника

про його можливість зв’язатися з іншим відправником, враховуючи привабливість повідомлень, які вона отримала від усіх інших відправників. Affinity Propagation алгоритм створює кластери, надсилаючи повідомлення між парами вибірок до зближення. Повідомлення, надіслані між парами, представляють придатність одного зразка бути прикладом іншого, який оновлюється у відповідь на значення з інших пар. Це оновлення відбувається ітераційно до досягнення консенсусу, після чого вибираються остаточні зразки, а отже, надається остаточна кластеризація [53].

Алгоритм починається з обчислення матриці S для знаходження подібності $s(i,k)$ між двома точками x_i та x_k за допомогою обчислення заперечення до значення суми квадрату різниці між цими точками. Далі алгоритм працює у два кроки передачі повідомлення та оновлює дві матриці: відповідальності R та доступності A. На початку обидві матриці містять значення, рівні нулю.

Матриця відповідальності R містить значення $r(i,k)$, які кількісно визначають, наскільки добре точка x_k може служити прикладом для x_i, порівняно з іншими прикладами-кандидатами для x_i. Формальним прикладом оновлення значень у матриці R може слугувати формула (2.5):

$$\mathrm{r}(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \qquad (2.5)$$

де i – це рядок матриці,

k – це стовбець матриці.

Матриця доступності A містить значення $a(i,k)$, які представляють, наскільки «відповідним» було б для x_i вибрати x_k як приклад, беручи до уваги перевагу інших точок щодо x_k як прикладу. Формальним прикладом оновлення значень у матриці A можуть слугувати формули (2.6) та (2.7):

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\right) \quad (2.6)$$

$$a(k,k) \leftarrow \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\} \quad (2.7)$$

Далі обчислюється матриця критеріїв C, де кожне значення є сумою відповідних значень матриці відповідальності та матриці доступності. Найвище значення критерію у кожному рядка визначається зразком. Рядки з одним і тим же значенням зразка знаходяться в одному кластері. Ітерації виконуються до тих пір, поки або межі кластера не залишаться незмінними протягом певної кількості ітерацій, або після певної попередньо визначеної кількості ітерацій [53, 54]. Основним недоліком Affinity Propagation є його складність. Алгоритм має часову складність порядку $O(N^2T)$, де N – кількість вибірок, а T – кількість ітерацій до збіжності. Це робить Affinity Propagation найбільш прийнятним для малих і середніх наборів даних.

2.2 Поняття міри відстані у кластеризації

Оскільки кластеризація тісно заснована на понятті груп, сусідів та зв’язку усередині кластерів, поняття відстані та вимірної метрики подібності є чи не основними концепціями роботи алгоритмів. Вибір найбільш доречної відстані та інших необхідних параметрів є надзвичайно важливим для отримання адекватних, реалістичних та цінних результатів. У даному підрозділі буде розглянуто декілька найбільш поширених метрик відстані, хоча варто зазначити, що таких насправді є сотні.

2.2.1 Евклідова відстань

Відстань Евкліда є основоположною метрикою відстані у математиці, адже саме за нею нормативно визначають відстань між двома точками у геометрії. В такому разі її часто називають довжиною відрізка, який з’єднує

дві точки у просторі, а ще L2 нормою вектору. На рисунку 2.1 наведено схематичне зображення вимірювання відстані Евкліда.

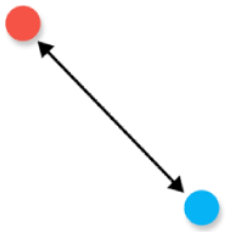


Рисунок 2.1 – Зображення відстані Евкліда у двовимірному просторі

Формула відстані (2.8) є очевидною і обчислюється використовуючи декартові координати точок та теорему Піфагора.

$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.8}$$

де n – кількість вимірів точки,

x_i, y_i – відповідні значення двох окремих точок.

Не зважаючи на те, що це дуже поширена міра відстані, вона не є масштабованою, тобто обчислені відстані можуть відхилятися залежно від одиниць вимірювання і їх поєднання – для уникнення цього дані потрібно нормалізувати та привести в єдиний інтервал. Евклідова відстань чудово працює, коли набір даних має невелику вимірність, а величина векторів може бути виміряна [55, 56].

2.2.2 Косинусна подібність

Косинусна подібність часто використовувалася як спосіб вирішення проблеми з високою розмірністю даних у випадку евклідової відстані, адже косинусна подібність – це звичайний косинус кута між двома векторами, як наведено на рисунку 2.2. Два вектори з повністю однаковою орієнтацією

матимуть значення подібності рівним одиниці. У цьому разі два протилежних вектори матимуть подібність рівною -1. Окрім цього варто зауважити, що така міра відстані не має одиниць вимірювання, що може нівелювати проблему масштабованості також.

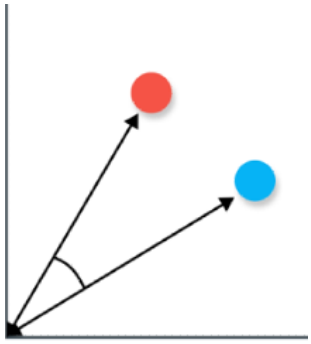


Рисунок 2.2 – Зображення косинусної відстані у двовимірному просторі

Формула косинусної подібності є також очевидною і наведеною у (2.9):

$$D(x,y) = \cos(\theta) = \frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{2.9}$$

де x, y – вектори або ж точки даних,

$\|x\|, \|y\|$ – норми (довжини) цих векторів.

Одним з основних недоліків косинусної подібності є те, що у ній не враховується магнітуда векторів, а лише їх напрямок. На практиці це призводить до того, що відмінності в окремо взятих точках даних не будуть повністю враховані, і таке завдання кластеризація як система рекомендацій не буде враховувати різницю в окремо взятих оцінках між користувачами. Косинусна подібність часто використовується для даних високої розмірності, і коли величини векторів не мають великого значення. Часто це використовується у порівнянні текстів, коли вони представлені кількістю слів у власне тексті [55].

2.2.3 Мангеттенська відстань

Відстань Мангеттена, яку також називають «відстанню таксі» або «відстанню міського кварталу», часто використовується для обчислення відстані між векторами з дійсними значеннями. Ця міра відстань обчислюється як сума абсолютних відстаней між двома точками і також називається L1 нормою вектору. Якщо уявити міську карту, то відстань обчислюється лише за прямими лініями, які перетинаються під прямим кутом, як показано на рисунку 2.3 нижче.

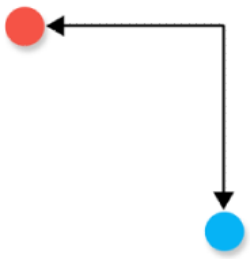


Рисунок 2.3 – Зображення Мангеттенської відстані у двовимірному просторі

Формула для обчислення є сумою модулів різниць між точками – (2.10):

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (2.10)$$

де n – кількість вимірів точки,

x_i, y_i – відповідні значення двох окремих точок.

Мангеттенська відстань може застосовуватися з даними високої розмірності, але її результат може бути менш інтуїтивно зрозумілим. Окрім цього значення відстаней, повернені цією мірою, будуть більшими, що потрібно враховувати при плануванні тієї чи іншої задачі.

Досить схожою до Мангеттенської є Канберрська відстань, яка по суті є зваженим варіантом попередньої, наведеним у формулі (2.11):

$$D(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (2.11)$$

Спершу ця метрика використовувалася як метрика для порівняння рейтингових списків і для виявлення вторгнень у комп'ютерній безпеці. Також її використовували для аналізу мікробіому кишечника при різних хворобливих станах [55, 56].

2.2.4 Відстань Геммінга

Відстань Геммінга – це міра відстані яка обчислює кількість значень, які відрізняються між двома векторами. У більш загальному випадку відстань Геммінга застосовується для порівняння двох рядків однакової довжини будь-яких абеток, що складаються з Q символів, і служить метрикою відмінності об'єктів визначеної вимірності. Очевидно, що порівнюватися відмінності можливо лише у випадку, коли вектори мають однакову довжину. Відстань Геммінга перевіряє кожну пару відповідних вимірів точки даних та визначає, чи відрізняються ці два атрибути, чи ні. Схематичний приклад наведено на рисунку 2.4.

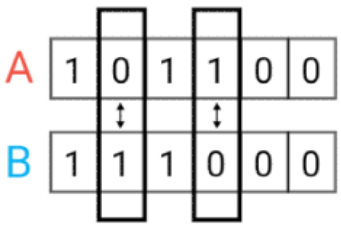


Рисунок 2.4 – Зображення відстані Геммінга для двох векторів

У випадку рівності значень відстань дорівнює 0, інакше – 1, як наведено у формулі (2.12). Загальна відстань між векторами буде сумою відстаней між кожною парою точок, як показано у формулі (2.13):

$$d(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases} \quad (2.12)$$

$$D(x, y) = \sum_{i=0}^{n} d(x_i, y_i) \quad (2.13)$$

де x_i, y_i – значення кожного виміру вектору для точки i.

Наприклад, два рядки “Hello World” та “Hallo Warld” матимуть відстань Геммінга рівною двом (бо існує різниця у два символи: е → а, о → а). Очікувано, що відстань Геммінга не застосовна, коли два вектори не мають однакової довжини. Більше того, цей алгоритм не враховує власне фактчну різницю між значеннями, лише констатує факт, що два значення різні. Тому у випадках коли саме різниця значень є важливою, така відстань не буде корисною. Типові випадки використання такої відстані включають виявлення та виправлення помилок під час передачі даних через комп’ютерні мережі. Крім того, широкого застосування відстань Геммінга набула у вимірювання подібності категорійних змінних [55, 56].

2.2.5 Відстань Чебишева

Відстань Чебишева – це одна із метрик, яка визначається як найбільша різниця між двома векторами вздовж будь-якого координатного виміру. Іншими словами, відстанню Чебишова між n-вимірними числовими векторами називається максимум модуля різниці компонент цих векторів, що обчислюється за формулою (2.14).

$$D(x, y) = \max_i(|x_i - y_i|) \quad (2.14)$$

де x_i, y_i – значення кожного виміру вектору.

Іноді її також називають дистанцією на шаховій дошці, оскільки мінімальна кількість ходів, необхідна королю, щоб перейти від одного поля

до іншого, дорівнює відстані Чебишева. Схематично вона зображена на рисунку 2.5.

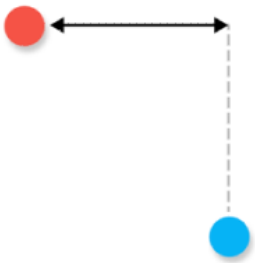


Рисунок 2.5 – Зображення відстані Чебишева (2 ходи)

Відстань Чебишева зазвичай використовується в дуже специфічних випадках використання, що ускладнює використання як універсальної метрики відстані, як-от евклідова відстань або косинусна подібність. На практиці вона часто використовується в складській логістиці [55].

2.2.6 Відстань Міньковського

Відстань Мінковського є дещо складнішою мірою, ніж попередньо розглянуті. Це метрика, яка використовується в нормованому векторному просторі (n-вимірний реальний простір). Тобто, це означає, що така метрика може використовуватися у просторі, де відстані представлені у вигляді вектора, який має певну довжину. Окрім цього відстань Мінковського також називають p-нормою вектора. Її обчислення наведені у формулі (2.15):

$$D(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}} \quad (2.15)$$

де x_i, y_i – значення кожного виміру вектору,

p – параметр відстані.

Найбільш цікавим у цій метриці є використання параметра p. Ми можемо використовувати цей параметр, щоб маніпулювати показниками відстані, щоб вони були схожі на інші, адже відстань Мінковського — це загальна форма евклідової та манхеттенської відстані. Тобто, коли значення p стає 1, це називається відстанню Манхеттена, а коли p приймає значення 2, воно стає евклідовою відстанню. При p, що прямує до нескінченності, можна отримати відстань Чебишева. Схематично це зображено на рисунку 2.6.

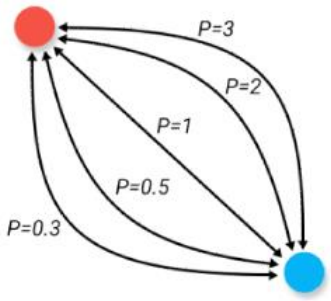


Рисунок 2.5 – Схематичне зображення впливу параметру на відстань Міньковського

Відстань Мінковскького має ті самі недоліки, що й міри відстані, які можуть бути ним представлені. Окрім цього можливість перебору та зміни параметра p є одночасно і цікавою перевагою, так і недоліком. Перебір значень параметра дозволяє знайти ту метрику, яка буде найоптимальнішою для поставленої задачі. З іншого боку, такий перебір значень з різною точністю може бути неефективним з точки зору обчислювальної складності та затраченого часу [55, 57].

2.2.7 Індекс Жаккарда. Індекс Соренсена-Дайса

Індекс Жаккарда, який ще називають перетином над об’єднанням – це показник, який використовується для визначення подібності та різноманітності набору даних. Як видно із назви, індекс є розміром перетину,

поділеним на розмір об’єднання пари наборів даних. Іншими словами, це загальна кількість однакових значень, поділена на загальну кількість значень. З індексу Жаккарда також можна отримати відстань Жаккарда, віднявши значення індексу від одиниці як наведено у формулі (2.16):

$$D(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|} \tag{2.16}$$

де $x \cap y$ – перетин двох векторів x, y,

$y \cup x$ – об’єднання двох векторів x, y .

Основним недоліком індексу Жаккарда є те, що на нього сильно впливає розмір даних. Кількість даних у наборі може стрімко збільшити значення «об’єднання», але «перетин» даних буде залишатися на тому ж місці.

Схожим до індексу Жаккарда є також індекс Соренсена-Дайса, оскільки обидві міри засновані на вимірювання подібності наборів даних. Індекс Соренсена-Дайса можна вважати трохи більш інтуїтивним, адже його результат можна вважати відсотком перекриття між двома наборами даних, що набуває значення від 0 до 1. Формулу розрахунку наведено нижче (2.17):

$$D(x,y) = \frac{2\,|x \cap y|}{|x| + |y|} \tag{2.17}$$

де $|x \cap y|$ – модуль об’єднання векторів x, y,

$|x|, |y|$ – модулі (кількості значень) векторів x, y.

Ці індекси часто використовується в програмах, у яких обробляються двійкові дані, наприклад, глибинне навчання, сегментація зображень тощо. Також їх часто використовують в аналізі текстів для вимірювання їх схожості за використаними словами [55, 56].

2.3 Поняття метрики якості кластеризації

Застосування методів кластеризації для вирішення бізнес-задач у кінцевому результаті дозволить отримати набір даних розділений на наперед визначену кількість груп. Проте, залишатиметься проблема розуміння того, чи сформовані кластери є адекватними і чи дані усередині них подібні. Адже кожен алгоритм кластеризації та кожна відстань використана у ньому, можуть привести до різних результатів. Хороша кластеризація створює кластери з високою схожістю усередині групи та низькою міжкласовою подібністю. Окрім цього алгоритм кластеризації повинен відповідати таким вимогам, як масштабованість, робота з різними типами атрибутів, створення кластерів довільної форми, опрацювання шуму та викидів у даних, нечутливість до порядку точок у наборі даних, підтримка високої розмірності тощо [58].

Існує два основних підходи до вимірювання якості кластеризації: зовнішній та внутрішній. Зовнішній підхід ґрунтується на навчанні «з учителем» – виді алгоритмів, у яких наперед відома «основна істина», тобто експерт наперед знає, до яких кластерів належать чи повинні належати точки і перевіряє точність алгоритму кластеризації. Можна сказати що такі методи оцінки визначають, наскільки виконана кластеризація є близькою до попередньо визначених еталонних класів. До зовнішніх метрик оцінки якості кластеризації відносять чистоту, індекс внутрішньої інформації та індекси Дайса, Калінського-Харабаша, рандомізований індекс та чистоту.

Внутрішній підхід навпаки працює за принципом навчання «без вчителя», тобто істинний розподіл точок по кластерах невідомий. Такий підхід оцінює, наскільки добре та компактно сформовані кластери. Окрім цього внутрішні методи оцінюють результат на основі даних, які і були кластеризовані. Через це вони надають кращу оцінку тим кластерам, у яких висока спорідненість усередині та висока відмінність від інших груп. До

внутрішніх підходів відносять коефіцієнт Силуетта, індекс Данна та індекс Девіеса-Боулдіна [47, 58].

2.3.1 Внутрішній підхід. Коефіцієнт Силуетта

Коефіцієнт Силуетта – це одна з метрик якості, яка визначається для кожного зразка і складається з двох частин: середньої відстані між зразком і всіма іншими точками в тому самому кластері та середньої відстані між зразком і всіма іншими точками в наступному найближчому кластері. Коефіцієнт Силуетта обчислюється за формулою (2.18):

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (2.18)$$

де $a(i)$— середня відстань точки i від усіх інших точок її кластера;

$b(i)$ — найменша середня відстань точки i до всіх точок будь-якого найближчого кластера.

Коефіцієнт Силуетта для набору даних визначається як середнє значення коефіцієнта Силуетта для кожного його зразка. Також варто зазначити, що він відображає міру того, наскільки близько кожна точка в кластері знаходиться до точок у сусідніх кластерах. Коефіцієнт обмежений інтервалом [-1; 1], де значення -1 відповідає неправильній кластеризації, а +1 – дуже щільній та хорошій кластеризації. Оцінка близька 0 вказує на перекриття кластерів. Неважко зробити висновок, що щільні та розділені у просторі кластери матимуть вище значення коефіцієнта, що вказуватиме на краще проведену кластеризацію [59, 60, 61].

2.3.2 Внутрішній підхід. Індекс Девіеса-Боулдіна

Іншою внутрішньою метрикою якості кластеризації є індекс Девіеса-Боулдіна, який визначають як середню міру подібності кластера з іншим

кластером, який найбільш на нього подібний. Таким чином, щільніші скупчення, які розташовуються якнайдалі, призведуть до кращого результату. Індекс Девіеса-Болдіна визначається за формулою (2.19):

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{\,j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \qquad (2.19)$$

де n – кількість кластерів,

σ_i – середня відстань усіх точок кластера i від центроїда кластера c_i.

Індекс Девіеса-Боулдіна ґрунтується на думці, що кластери, які достатньо віддалені один від одного і є щільними, швидше за все є хорошою кластеризацією. У формулі (2.19) багаторазово максимізується значення, де середня точка найбільше віддалена від центроїда, і де центроїди знаходяться найближче. Мінімальне значення індексу дорівнює нулю, і чим значення ближче до нуля, тим кращою вважатиметься кластеризація [60, 61].

2.3.3 Внутрішній підхід. Індекс Данна

Ще одним показником оцінки якості алгоритму кластеризації є Індекс Данна, який дорівнює мінімальній відстані між кластерами, поділеній на максимальний розмір кластера та обчислюється за формулою (2.20):

$$D = \frac{min_{\,1 \le i < j \le n}\, d(i,j)}{max_{\,1 \le k \le n} d'(k)} \qquad (2.20)$$

де i, j, k – індекси кластерів,

d – відстань між кластерами,

d' – відстань між елементами усередині кластеру.

Великі відстані між кластерами вказують на кращий розподіл, менші розміри кластерів – на більшу щільність усередині. Комбінація цих показників призводить до вищого значення індексу, що означає кращу

кластеризацію. Як й інші оцінки якості індекс Данна стає вищим, коли кластери щільніші та розкинуті якнайдалі один від одного [59, 60].

2.3.4 Зовнішній підхід. Індекс Ранда

Часто використовуваним показником є індекс Ранда, який обчислює міру подібності між двома кластерами. Для цього враховуються всі пари точок, і алгоритм обчислює кількість пар, які призначені у правильних кластерах (однакових чи різних), порівнюючи еталонну та реальну кластеризацію. Спрощений вигляд формули наведено у (2.21):

$$RI = \frac{\text{Кількість узгоджених пар}}{\text{Кількість пар}} \tag{2.21}$$

Значення RI може варіюватися від 0 – відсутність узгодження - до 1 – ідеальної кластеризації. Основним недоліком є власне зовнішня природа методу та необхідність попередньо знати справжні мітки кластерів [61].

2.3.5 Зовнішній підхід. Індекс Калінського-Харабаша

Індекс Калінського-Харабаша також відомий як критерій відношення дисперсії є ще однією метрикою зовнішньої оцінки якості. Оцінка визначається як співвідношення між дисперсією всередині кластеру та дисперсією між кластерами. Індекс Калінського-Харабаша – це спосіб оцінити продуктивність алгоритму кластеризації. Окрім цього його застосування не вимагає інформації про еталонні мітки [61]. Для набору даних D з N точок розділеного на K кластерів індекс обчислюється за формулою (2.23) [62]:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|c_k - c\|^2}{K-1}\right] / \left[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|d_i - c_k\|^2}{N-K}\right] \tag{2.23}$$

де n_k – кількість точок у кластері k,

c_k – центроїд кластера k,

c – глобальний центроїд набору даних.

2.3.5 Зовнішній підхід. Чистота

Чистота є оцінкою того, у якій мірі кластери містять один клас даних. Розрахунок такої метрики можна визначити як: для кожного кластера підраховується кількість точок даних із найпоширенішого класу у цьому кластері. Після чого знаходиться сума цього значення для всіх кластерів і ділиться на загальну кількість точок даних. Формально, враховуючи деякий набір кластерів M і деякий набір класів D, які розділяють N точок даних, чистота може бути визначена як (2.24):

$$P = \frac{1}{N} \sum_{m \in M} \max_{d \in D} |m \cap d| \tag{2.24}$$

Оцінка чистоти, рівна 1 можлива, якщо помістити кожну точку даних у власний кластер. Окрім цього така метрика не є корисною для кластерів з великим розкидом розподілу – наприклад, 90% точок знаходяться в одному кластері, і 10% - у іншому [47].

2.3.6 Визначення оптимальної кількості кластерів. Метод Елбоу

Окрім цього існують метрики, які допомагають визначити оптимальну кількість кластерів для того чи іншого набору даних. Одним з методів визначення оптимальної кількості кластерів є метод Елбоу, або у перекладі – метод ліктьового згину.

Метод Елбоу є евристикою, яка використовується для визначення оптимального числа кластерів у наборі даних. Підхід базується на побудові графіка з розглядом дисперсії як функції від кількості кластерів і вибору

«ліктя» кривої як оптимальної кількості груп. Ліктем – або, іноді, коліном – називають точку, після якої припиняється стрімке спадання/зростання значення функції. Використання цього як точки відсікання є поширеною математичною оптимізацією, яка часто застосовується при пошуку місця, у якій зменшення певної величини вже не є вартою подальших зусиль. У кластеризації це визначається як те, що додавання нового кластеру не надаватиме значного покращення у моделі даних. Математично метод Елбоу зазвичай ґрунтується на обчислення суми квадратів різниці усередині кластеру для різних значень кількості кластерів K.

Після обчислення таких відстаней для заданих кількостей кластерів, будується двовимірний графік, на осі абсцис покладені значення кількості кластерів, на осі ординат – значення WCSS. Точки з'єднуються між собою і оптимальною визначається та, у якій з'являється «лікоть» - якій передує максимально стрімке зниження значення. Приклад наведено на рисунку 2.6 (як бачимо, стрімке зростання припиняється у точці «кількість кластерів = 4») [63, 64].

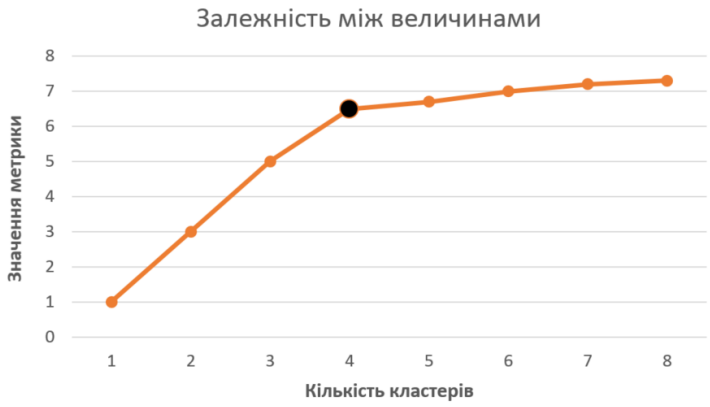


Рисунок 2.6 – Схематичний приклад графіка для методу Елбоу

Варто зазначити, що часто метод Елбоу застосовують й з іншими метриками, не тільки дисперсією. Проте, потрібно звертати увагу, що на практиці при використанні кількох метрик потрібно балансувати між вибором

найкращого значення та «ліктем зламу». У кваліфікаційній роботі метод Елбоу буде застосовано для визначення оптимальної кількості кластерів на основі порівняння значень обраних метрик якості кластеризації.

2.4 Висновки до другого розділу

У другому розділі кваліфікаційної роботи було розглянуто одну з основних задач інтелектуального аналізу та видобутку даних – а саме, кластеризацію. Описано популярні підходи та методи кластеризації: ієрархічний, центроїдний, на основі щільності та розподілу тощо, розглянуто алгоритми, їх переваги, недоліки та сфери застосування. Оскільки задача розподілу бізнес-структур за рівнем їх цифрової зрілості схожа на сегментацію користувачів, найбільш вдалими для застосування видаються ієрархічні та центроїдні алгоритми. Щільність та розподіл даних не є відомими, тому такі алгоритми як DBSCAN, OPTICS, EM можуть не принести хороших результатів.

Також було розглянуто декілька мір відстані: евклідову, косинусну, Геммінга, Чебишева, Міньковського тощо. Відзначимо, що поняття подібності через міру відстані використовується не у всіх розглянутих методах. Зважаючи на знання набору даних, цікавим видається використання косинусної та Мангеттенської відстані, а також відстані Геммінга та індексів Жаккарда і Соренсена-Дайса.

Важливим аспектом кластеризації залишається визначення її якості та адекватності. Як деякі метрики якості утворених кластерів було описано коефіцієнт Силуетта, індекси Данна, Девіеса-Боулдіна, Ранда, Калінського-Харабаша та поняття чистоти групи. Зважаючи на те, що набір даних не містить попередньо визначених цільових міток, необхідним є використання внутрішніх метрик.

3 КЛАСТЕРИЗАЦІЯ МІКРО, МАЛИХ ТА СЕРЕДНІХ ПІДПРИЄМСТВ ТЕРНОПІЛЬСЬКОЇ ОБЛАСТІ ЗА РІВНЕМ ЇХ ЦИФРОВОЇ ЗРІЛОСТІ

У третьому розділі дипломної роботи основну увагу приділено програмному забезпеченню, розробленому для виконання поставлених завдань: кластеризації опитаних підприємств на основі розрахованого рівня цифрової зрілості та аналізу її результатів. У попередніх роботах дослідницької групи [65, 66] оптимальною та опорною кластеризацією було визнано агломеративний підхід з використанням модифікованої метрики відстані Говера про що детально описано у роботі [65]. Проте, згідно з метою кваліфікаційної роботи, важливим є отримання нових цінних ідей щодо можливих спільних рис підприємств у розрізі рівня цифрової трансформації. Тому існує інтерес щодо застосування інших підходів, метрик та комбінацій параметрів кластеризації, ніж попередньо використовувані у дослідженнях.

На основі теоретичного матеріалу щодо задачі кластерного аналізу, розглянутого у розділі 2, як основні методи було обрано наступні:

- Агломеративний – представник ієрархічних методів
    - з використанням повного з’єднання та з’єднання Уорда.
- K-means – представник методів на основі центроїда.
- OPTICS – представник методів на основі щільності.
- Affinity Propagation – представник алгоритмів на основі спорідненості.
- Gaussian Mixture EM – представник алгоритмів на основі розподілу.

Варто зазначити, що поняття міри відстані використовується лише у агломеративному та OPTICS алгоритмах. Як можливі міри відстані було вибрано відстань Евкліда, косинусну відстань, Манхеттенську відстань, відстань Чебишева та відстань Міньковського.

Як формальні показники якості кластеризації було обрано дві внутрішні та одну зовнішню метрики, які не залежать від наявності цільової змінної у наборі даних. Серед них попередньо імплементованими у бібліотеках Python є: індекс Силуетта, індекс Калінського-Харабаша та індекс Девіеса-Боулдіна.

3.1 Огляд набору даних, що використовується у роботі

Важливим є розглянути дані, що використовувався як і у дипломній роботі, так і в окремих дослідженнях [65, 66] – попередніх апробаціях алгоритму та методики. Набір даних представляє собою результати опитування, здійсненого під час вибіркового анкетування підприємців регіону у 2019 році за допомогою сервісу Google Forms. Учасниками опитування стали представники різних сфер діяльності, які зареєстровані у Тернопільській області. На момент проведення кластеризації набір даних містив відповіді 34-ти респондентів. Той самий набір відповідей було визначено базовим набором даних, який використовуватиметься для обчислення індексу цифрової зрілості підприємства (HIT) у цій роботі.

Математично дані представляють собою набір з N респондентів $U = \{\overrightarrow{u_1}, \overrightarrow{u_2}, \dots, \overrightarrow{u_N}\}$ та M запитань $Q = \{q_1, q_2, \dots, q_M\}$. Вважатимемо, що кожен учасник $\overrightarrow{u_i} \in U$ відповів на кожне із запитань $q_k \in Q$. Таким чином було сформовано матрицю відповідей розмірністю $(N \times M)$, у якій кожен опитаний учасник представлений у вигляді наступного кортежу: $\overrightarrow{u_i} = \{u_{i1}, u_{i2}, \dots, u_{ik}, \dots, u_{iM}\}$, де u_{ik} є відповіддю і-го учасника опитування на k-те запитання. Надалі такий кортеж називатимемо точкою. На рисунку 3.1 наведено масив відповідей у матричній формі.

		Questions						
Respondents		q_1	q_2	q_3	...	q_k	...	q_M
	$\overrightarrow{u_1} =$	u_{11}	u_{12}	u_{13}	...	u_{1k}	...	u_{1M}
	$\overrightarrow{u_2} =$	u_{21}	u_{22}	u_{23}	...	u_{2k}	...	u_{2M}
	...	...	...	...	...	...	...	...
	$\overrightarrow{u_N} =$	u_{N1}	u_{N2}	u_{N3}	...	u_{Nk}	...	u_{NM}

Рисунок 3.1 – Отримана матриця відповідей (запозичено з [66])

Щодо специфіки опитування, то учасникам було запропоновано відповісти на 34 запитання, що стосуються двох аспектів функціонування бізнес-структури:

- Загальні дані про бізнес (форма організації, кількість працівників, сфера діяльності, ведення імпорту чи експорту тощо).
- Рівень цифровізації бізнес-діяльності за трьома компонентами HIT-індексу:
  - Використання цифрових інструментів таких як просування сайту, соціальних мереж, спеціалізованих систем фінансової звітності чи логістики тощо.
  - Забезпечення підприємства комп’ютерною та мобільною технікою, а також широкосмуговим швидкісним Інтернет-покриттям.
  - Оцінка цифрової грамотності працівників підприємства.

Частина вищезазначених запитань, особливо у секції загальної інформації про бізнес, дозволяла введення текстової відповіді респондентом, проте більшість питань вимагало вибору одного із запропонованих варіантів відповідей. Варто також зазначити, що варіанти відповідей було попередньо проранжовано за збільшенням ефективності використання, а також закодовано, тобто у підсумку кожному варіанту відповідало певне число. Такий підхід дозволив використати як категорійні, так і числові дані в одному наборі. Для прикладу на рисунку 3.2 наведено частину таблиці вхідного набору даних із вже закодованими відповідями.

ID	Name	organization_ type	import_ export	business_ model	computer_ equipment	mobile_ devices	internet_c onnection	ict_spec_ expernal	ict_spec _internal	web_site	basket_ chain	seo_opti mization	smm_fb_i nst	smm_ other	smm_effe ctiveness	fb_ads	google_ ads
1	Respondent1	1	0	0	2	2	0	0	0	0	0	1	0	0	0	0	0
2	Respondent2	1	0	0	2	2	1	1	0	0	0	0	1	0	2	0	0
3	Respondent3	1	0	0	2	2	1	1	2	0	0	0	2	0	2	0	0
4	Respondent4	1	0	0	2	2	2	1	2	1	1	2	2	0	3	3	0
5	Respondent5	0	0	0	0	0	2	1	2	0	0	0	0	0	0	0	0
6	Respondent6	0	0	0	2	2	2	1	2	0	0	0	0	0	2	0	0
7	Respondent7	1	0	0	2	2	2	1	0	1	0	1	2	1	3	0	0
8	Respondent8	1	0	0	0	0	0	1	0	0	0	0	2	0	3	0	0
9	Respondent9	1	0	0	2	2	2	0	3	1	0	0	2	0	2	0	0

Рисунок 3.2 – Частина закодованої таблиці вхідного набору даних

Під час початкової роботи з даними було проведено їх ручне очищення, що полягало лише у відокремленні (видаленні) запитань, на які було отримано розгорнуті відповіді від респондентів, адже такі не були корисними для обчислення індексу та подальшої кластеризації. Набір даних, що складався із числових відповідей на вибрані запитання було визнано як готовий до використання. Подальша робота з файлом, включаючи очищення та нормалізацію даних, відбувалася у розробленому програмному застосунку.

3.2 Розробка програмного додатку для кластеризації підприємств

Оскільки задача кластеризації є одним із базових завдань зі сфери аналізу та видобутку даних, її реалізація вручну є хоч і можливим, проте неоптимальним способом вирішення. Зважаючи на це, виконання поставлених задач роботи, включаючи попередню нормалізацію даних, обчислення HIT-індексу та проведення кластеризації на його основі було реалізовано у вигляді локальної комп’ютерної програми.

Зважаючи на те, що темою дипломного дослідження є кластеризація та аналіз отриманих результатів, а не власне розробка програмного забезпечення, було прийнято рішення відмовитися від варіантів, що включали

в себе розробку клієнт-серверної архітектури або ж веб-додатку. Вирішено зупинитися на способі рішення, що включало в себе інтерактивне введення команд та отримання результатів у терміналі з можливістю модифікації коду.

Як засіб написання застосунку було обрано мову Python – високорівневу інтерпретовану об'єктно-орієнтовану мову програмування. Її перевагами вважаються чистий синтаксис, зручність для розв'язання математичних питань, велика кількість середовищ для розробки, наявність відкритого коду проєкту, ефективність структур даних та інші. В порівнянні з іншими популярними мовами програмування на кшталт Java, C#, Ruby, Go, Kotlin, JavaScript та іншими, Python відомий своїм поширеним застосуванням у сфері аналізу та обробки даних, а також машинного навчання. Це дозволяє дещо розділити сфери використання різних мов. Звісно, існує можливість застосування вищезгаданих мов програмування у сфері обчислень та аналізу, проте їхня основна ніша припадає на розробку веб-додатків, сайтів, навантажених систем, комп'ютерних програм тощо. В той же час Python також успішно використовується у традиційному програмуванні як back-end мова, але його застосування у науці про дані важко переоцінити. Окрім цього, для аналізу математичних концепцій та обчислень використовуються такі мови програмування, як Scala, Julia та R – проте, у дещо інших задачах, ніж того вимагали завдання, поставлені у дипломній роботі щодо кластеризації суб'єктів малого та середнього бізнесу.

На додачу Python володіє колекцією з безлічі додаткових бібліотек та розширень, які дозволяють використовувати вже наявні функції візуалізації, зберігання, аналізу даних, взаємодії з операційною системою, кластеризації тощо у власних алгоритмах. У роботі серед інших найчастіше використовувалися такі допоміжні бібліотеки:

- scikit-learn – для використання алгоритмів кластеризації та обчислення метрик якості.
- scipy – для обчислення матриць відстані на основі набору даних.

- matplotlib – для візуалізації отриманих даних у вигляді графіків.
- pandas – для зберігання та маніпулювання набором даних у вигляді спеціальної структури – датафрейму (dataframe).
- psycorg2 – для з’єднання з базою даних PostgreSQL.

Також для розробки програми було використано Jupyter Notebook – інтерактивне обчислювальне середовище для Python, яке запускається локально у користувача та дозволяє створювати так звані «блокноти» з кодом. Jupyter поєднує в собі звичне середовище розробки (таке як Microsoft Visual Studio, PyCharm, Eclipse тощо) та інтерактивність, за допомогою якої можна запускати лише окремі блоки коду, отримувати результат в окремих вікнах або терміналі тощо.

На додачу і як вже було згадано вище, у ролі сервера бази даних було використано локально розгорнутий примірник PostgreSQL 14. Як і його аналоги: Microsoft SQL Server, MySQL, Firebird, SQLite – PostgreSQL – це об’єктно-реляційна система керування базами даних з відкритим кодом. Часто її вважають більш гнучкою та легкою системою, яка не прив’язана до розробників із великих корпорацій. Зважаючи на необхідність легкого примірника сервера бази даних для зберігання інформації, вибір було зроблено на користь PostgreSQL.

3.2.1 Функціональні можливості розробленої програми

Зважаючи на поставлені задачі дослідження та розуміння набору даних, з яким потрібно працювати, розробка програми розпочиналася з визначення переліку основних функцій, виконання яких потрібно забезпечити, а також зі встановлення вимог до вхідних даних, алгоритмів та результатів цих функцій. Отже, для проведення кластеризації даних серед інших, було реалізовано такі основні функції, зображені на рисунку 3.3 та розглянуті нижче.

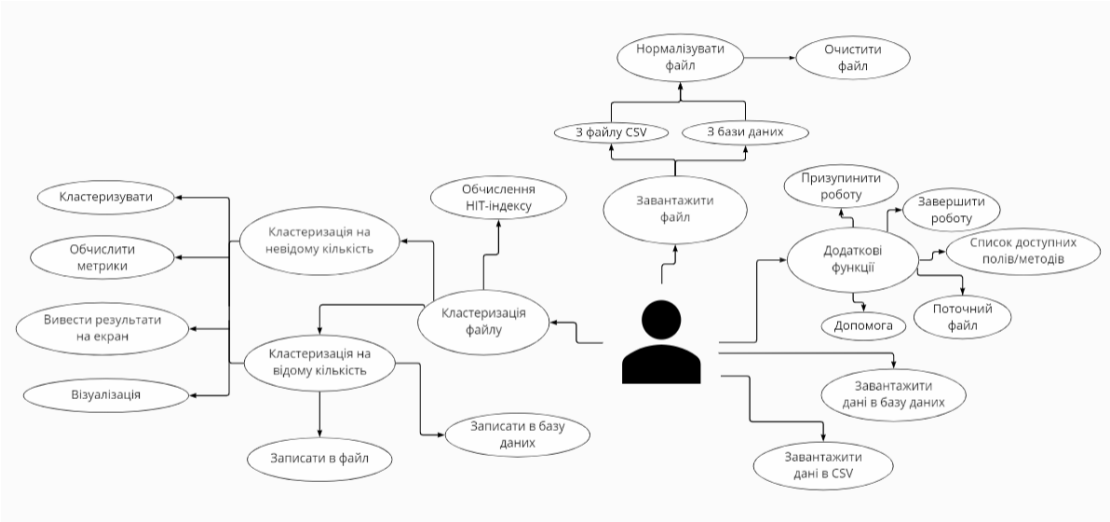


Рисунок 3.3 – Схематична діаграма функцій, доступних користувачу

Розглянемо деякі з цих функцій детальніше:

1. Завантаження даних з файлу у форматі CSV. Варто зазначити, що у такому випадку файл повинен мати визначену структуру зі сталою кількістю стовбців даних, назвами колонок та типами даних у них.

2. Завантаження даних із таблиці бази даних. Одночасно з можливістю завантаження окремих CSV-файлів, ми орієнтуємось на те, що у майбутньому інформація про нові проходження міститиметься в централізованій базі.

3. Очищення та нормалізація даних, що полягає в обробці завантаженого набору даних, пошуку можливих викидів, пропущених даних, та приведення значень у кожному стовбці-запитанні до інтервалу [0;1].

4. Обчислення індексу цифрової трансформації HIT для кожного із респондентів за формулами, наведеними у підрозділі 1.5.

5. Кластеризація за попередньо обчисленим HIT-індексом, що включає в себе:

a. власне кластеризацію даних за обраним методом, відстанню та кількістю кластерів/сусідів;

b. обчислення метрик якості кластеризації таких як: індекси Силуетта, Калінського-Харабаша та Девіеса-Боулдіна;

c. виведення результатів на екран;

d. візуалізацію даних у вигляді точкової та стовпчикової діаграм;

6. Завантаження короткого підсумку кластеризації, виконаних візуалізацій, списку кластерів, а також набору даних з обчисленими індикативними компонентами та власне значенням індексу на комп’ютер користувача у вигляді TXT, CSV та PNG файлів, а також запис таблиці кластеризації у базу даних.

7. Деякі додаткові допоміжні функції, такі як:

a. вивід поточного файлу;

b. вивід переліку полів, методів кластеризації тощо;

c. призупинення роботи;

d. завершення роботи;

e. виклик «Допомоги» з переліком доступних функцій та правил їх виклику.

Звертаючи увагу на останній пункт переліку вище (7.e), варто зазначити, що розроблена комп’ютерна програма містить чималу кількість допоміжних функцій, які забезпечують достатню вкладеність та атомарність операцій. Проте, у цій науковій роботі основна увага буде присвячена саме обчисленню HIT-індексу та проведенню кластеризації на його основі. Оскільки створена програма призначена для використання усередині дослідницької команди, вона не містить графічного інтерфейсу користувача, а взаємодія проводиться через визначений перелік команд. Такі команди складаються з дієслова – опису дії, яку необхідно виконати, а також додаткових параметрів. Після введення команда програмно розподіляється на слова, з яких частина за частиною здійснюється перехід у потрібну функцію програми. Деякими командами є, наприклад, наступні:

- connect raw table raw_file – під’єднання таблиці бази даних з назвою «raw_file», яка містить необроблені дані.

- load /my file – завантаження CSV-файлу з іменем «my file».
- normalize – команда для нормалізації набору даних та приведення його значень до інтервалу [0;1].
- current file – вивід на екран поточного файлу.
- cluster agglomerative cosine 5 hit – проведення кластеризації за агломеративним методом, використовуючи косинусну відстань, HIT-індекс та розподіляючи дані на 5 кластерів.

3.2.2 Збереження інформації у базі даних

Перш, ніж перейти до програмної реалізації поставлених завдань, розглянемо базу даних, яка використовується як сховище даних у роботі. Як вже було згадано вище, у ролі сервера бази даних використовується PostgreSQL 14. Системою управління базами даних (СУБД) у цьому випадку виступає PgAdmin 4, за допомогою якого можна здійснювати моніторинг за станом системи, створювати, модифікувати та видаляти таблиці тощо.

База даних початково складається з двох таблиць: raw_encoded та weights. У першій таблиці зберігається основна база даних з закодованими назвами респондентів, яка використовуватиметься як базовий файл у роботі. Друга таблиця містить сталі вагові коефіцієнти для усіх питань, які задіяні в обчисленні HIT-індексу.

Таблиця raw_encoded містить 36 колонок, а саме ID респондента, його ім’я та відповіді на 34 питання, описані у підрозділі 3.1. Усі стовбці таблиці окрім імені є цілими числами з накладеним обмеженням обов’язковості введення. Оскільки у подальшому дані щодо опитувань користувачів планується отримувати з іншого програмного компоненту, такі обмеження дозволяють проводити перший раунд перевірки, що дані не міститимуть пропусків чи недозволеного формату даних. Очевидно також, що кількість стовбців у такій таблиці є сталою для будь-якого респондента.

Таблиця weights складається з 3-х колонок: назви стовбця, його вагового коефіцієнта та відповідності до індикативного компонента індексу (тобто, до якої категорії відноситься запитання: загальні, цифрова грамотність, використання інструментів, інфраструктура). Розрахунок відповідності запитань та їх вагових коефіцієнтів було попередньо здійснено дослідною групою у рамках роботи над індексом HIT. Очевидним є факт, що кількість рядків цієї таблиці дорівнюватиме 36 – кожен рядок відповідає певному стовбцю таблиці з даними.

Окрім цього, користувач має можливість додавати власні таблиці у базу даних користуючись розробленою програмою, у двох випадках:

- Створення нової сирої / очищеної / нормалізованої таблиці даних після завантаження її у програму у вигляді CSV-файлу.
- Створення нової кластеризованої таблиці даних, що відбувається відразу після кластеризації даних задля збереження історії.

Під час створення таблиці у створеній програмі відбувається перевірка на наявність запитуваних даних, отримання назв стовбців файлу та покрокове генерування SQL-запиту для створення таблиці. Для формування імені створюваної таблиці у програмі використовується: тип таблиці, вказане ім’я або ж комбінація методу+відстані+кількості кластерів та поточні значення дня, місяця, години і хвилини. У подальшому користувач може завантажити цю таблицю як початковий набір даних у програму, або ж переглянути або видалити її, використовуючи PgAdmin СУБД.

3.3 Програмна реалізація основних функцій розробленої програми

Перед використанням програми та аналізом отриманих результатів, коротко розглянемо частини програмної реалізації деяких функцій. Як було зазначено вище, програмний код містить досить велику кількість функцій для забезпечення атомарності операцій та зменшення кількості дубльованого

коду. Розробка програми велася із застосуванням деяких базових практик чистого коду, а саме: зрозумілість імен змінних та функцій, доречна кількість коментарів, вертикальне та горизонтальне вирівнювання, а також функції, що виконують лише одне завдання за раз. Очевидно, що за такого підходу основна функція може послідовно викликати інші для виконання дрібніших задач. На рисунку 3.4 наведено фрагмент діаграм зв'язку між функціями програми.

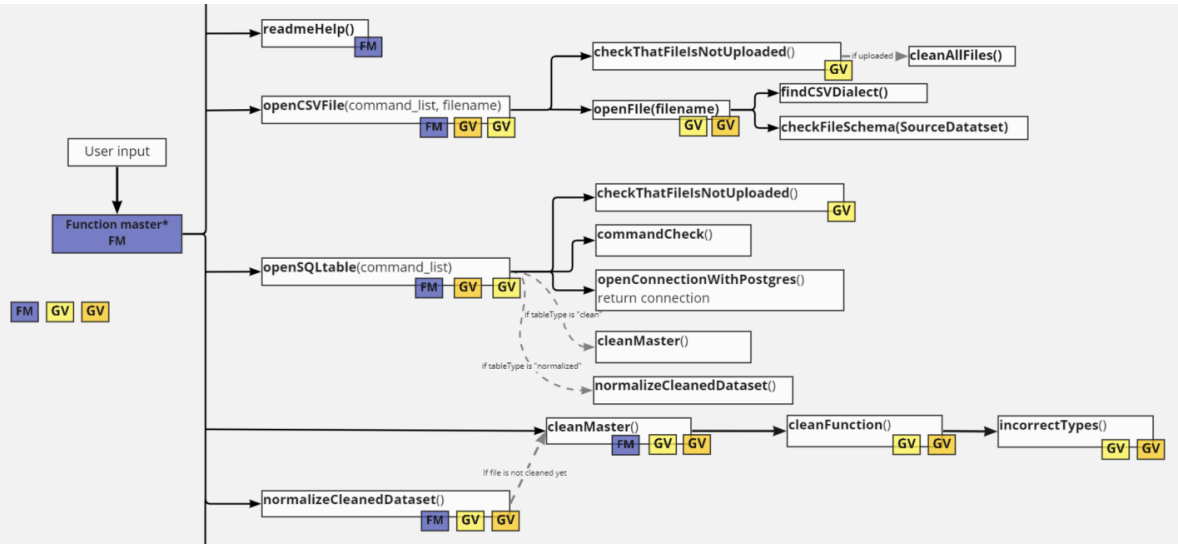


Рисунок 3.4 – Частина діаграми зв'язку між функціями у програмі

Як бачимо з рисунка, функція openCSVFile() послідовно викликає функції нижчого рівня: checkThatFileIsNotUploaded() та openFile(), які у свою чергу викликають ще дрібніші функції cleanAllFiles(), findCSVDialect() та checkFileSchema() відповідно.

Розглянемо вищезгадану функцію зчитування інформації з CSV-файлу. Як було зазначено, такий файл повинен складатися з визначеної кількості стовбців та мати певну погоджену схему. У лістингу 1 наведено частину функції openFile(filename), яка відповідає за відкриття CSV-файлу та зчитування інформації з нього. Певні неважливі частини упущено.

Лістинг 1. Програмний код для відкриття CSV-файлу

```
1 def openFile(filename):
2  filesource = sourceFolder+"\\"+filename+'.csv'# шлях до файлу
3  try:
4    dialect = findCSVDialect(filesource)  #пошук діалекту CSV
5    SourceDataset  =  pd.read_csv(filesource,  sep  =  dialect.
delimiter) # зчитування інформації у датафрейм
6     correctTableSchema     =     checkFileSchema(SourceDataset)
#перевірка чи завантажений файл відповідає схемі
7     if correctTableSchema is False:
8          cleanAllFiles()
9          return
10  except FileNotFoundError as error: #якщо файлу не існує
11      print("Unfortunately,   an   error   was   raised:   \n
{}".format(error.strerror), error.filename)
12       return
```

Як бачимо, у рядку 2 відбувається побудова шляху до файлу, у рядку 4 – визначається діалект CSV, у 5-му рядку файл зчитується у спеціальну структуру Python – dataframe. Опісля відбувається перевірка схеми файлу – кількості стовбців та їх заголовків. Якщо схема неправильна – файл видаляється з пам’яті, користувач повинен завантажити новий. Окрім цього реалізовано перехоплення помилок, наприклад, якщо у папці не існує файлу з вказаною назвою.

Для коректної роботи більшості алгоритмів кластеризації важливим також є використання нормалізованих даних, які не містять різнорідних даних, викидів тощо. Саме тому одним із бізнес-правил було визначено можливість проведення кластеризації лише на нормалізованому наборі даних. У розробленій програмі нормалізація відбувається у два етапи: перевірка файлу на наявність порожніх значень, помилково введених символів, викидів тощо та власне нормалізація – приведення значень всіх числових стовбців в інтервал [0;1]. У лістингу 2 наведено приклад роботи частини функції incorrectTypes(), яка здійснює перевірку, чи у наборі даних є значення з не числовим типом даних (рядки 2-4). Якщо такі значення існують, користувачу виводиться попередження зі списком стовбців, у яких знайдено некоректні типи, і файл видаляється з пам’яті (рядки 5-9).

Лістинг 2. Частина коду щодо пошуку некоректних типів даних

```
1 def incorrectTypes():
2 for column_name in CleanedDataset.columns:
3     if CleanedDataset[column_name].dtype == object:
4          listOfIncorrectColumns.append(column_name)
5 if len(listOfIncorrectColumns) > 0:
6    print(……)
7    for key in listOfIncorrectColumns:
8         print("Errors likely in '{}' column. Unsupported data
type: {}".format(key, "object"))
9          cleanAllFiles()
```

Оскільки набором даних, що використовуватиметься для кластеризації, було визначено значення трьох індикативних компонентів HIT-індексу, важливим видається зупинитися на алгоритмі розрахунку значень індексу для кожного респондента опитування. Як вже було згадано вище, HIT-індекс складається з 3-х компонентів: людського ресурсу, рівня забезпечення необхідною технікою та рівня використання цифрових інструментів на підприємстві. Кожен із індикативних компонентів має свою вагу, яка враховується при розрахунку.

У роботах [42, 43] дослідницькою групою було визначено такі значення ваг компонентів: людського ресурсу – 0.3, технічного забезпечення – 0.2, та використання інструментарію – 0.5. Також власні ваги мають усі запитання, включені в обчислення того чи іншого компонента. Спрощену блок-схему алгоритму розрахунку індексу наведено на рисунку 3.4. Як видно з рисунка, обчислення індексу розпочинається з під’єднання до бази даних та отримання списку вагових коефіцієнтів для кожного з запитань трьох категорій.

Після чого циклічно відбувається обчислення індикаторів H, I та T для кожного респондента зважаючи на його відповіді у анкетуванні за формулою (1.3). Після отримання значень індивідуальних компонентів вираховується загальне значення індексу цифрової трансформації за формулою (1.1). Зрештою, інформація записується у структуру даних для подальшого використання.

Рисунок 3.4 — Схематичний алгоритм обчислення HIT-індексу

У лістингу 3 нижче наведено фрагмент програмного коду що використовується для обчислення.

Лістинг 3. Програмна реалізація обчислення HIT-індексу

```
1  connection = openConnectionWithPostgres()
2  cursor = connection.cursor()
3  dict_h, listOfKeys_h = retrieveData('h', cursor)
4  dict_t, listOfKeys_t = retrieveData('t', cursor)
5  dict_i, listOfKeys_i = retrieveData('i', cursor)
... ... ... ...
6  for j in range(0, numberOfRows):
7      H, I, T = 0, 0, 0
8      for key_h in listOfKeys_h:
9           H += dict_h[key_h]*baseFile.iloc[j][key_h]
10     for key_i in listOfKeys_i:
11          I += dict_i[key_i]*baseFile.iloc[j][key_i]
```

```
12      for key_t in listOfKeys_t:
13           T += dict_t[key_t]*baseFile.iloc[j][key_t]
14      HITSumm = round( (H * weightsForSumm[0] + I *
weightsForSumm[1] + T * weightsForSumm[2]) ,4)
```

Варто також зауважити, що не всі питання, отримані в очищеному наборі даних, були придатними для використання при обчисленні HIT-індексу. Причиною цього є те, що HIT-індекс обчислюється за формулою, що включає в себе математичні дії з числовими величинами. У випадку деяких запитань їх закодовані нормалізовані значення було неможливо проранжувати за ідеєю «краще-гірше» у розрізі цифровізації. Такими запитаннями було визначено ті, які стосувалися кількості працівників, імпорту чи експорту товарів, заключення договорів на обслуговування техніки чи наявність штатних працівників. Ваги таких запитань було визначено як рівними нулю, тож вони не впливали на значення обчисленого індексу.

Після обчислення HIT-індекс використовувався як основний набір даних для проведення кластеризації. Як його аналоги також розглядалися: власне набір відповідей опитування та комбінований набір із відповідей на запитання та значень індексу. Ці варіанти були відкинуті через високу розмірність, залежність між собою та співвідношення близько 1:1 між кількістю опитаних та кількістю вимірів вектору. Деякі пробні запуски алгоритму підтвердили такий вибір, показавши кращі результати на наборі даних значень компонент HIT-індексу.

Програмна реалізація кластеризації, обчислення мір відстані та метрик якості покладалася на використання функцій із сторонніх бібліотек. У загальному кластеризація складається із двох кроків: створення моделі кластеризації та її використання на наборі даних.

При створенні моделі у функцію кластеризації передається певний список параметрів. Список відрізняється між алгоритмами, проте найчастіше включає в себе максимальну кількість ітерацій виконання, кількість

кластерів, міру відстані, яка використовується, мінімальне чи максимальне число сусідів точки тощо.

На наступному кроці модель використовується для передбачення кластерів у наборі даних і найчастіше повертає власне список міток, призначених кожній точці. Наприклад, у лістингу 4 наведено програмний код запуску функції агломеративної кластеризації (у рядках 2-3 відбувається створення матриці відстаней, у рядку 4 – утворення моделі, у рядку 5 – присвоєння моделі до вже утвореної матриці відстаней).

Лістинг 4. Приклад реалізації агломеративної кластеризації

```
1   def   AgglomerativeCluster   (distance,   numberOfClusters,
baseFile):
2        distVector = sp.pdist(baseFile, distance)
3        distMatrix = sp.squareform(distVector, 'tomatrix')
4       model   =   AgglomerativeClustering   (n_clusters   =
numberOfClusters,   affinity   =   "precomputed",   linkage   =
"complete")
5        clusterLabels = list(model.fit_predict(distMatrix))
6        return clusterLabels
```

Розглядаючи створення матриці відстані, можемо помітити, що у функцію передається набір даних, в якому потрібно знайти відстань та закодована назва міри відстані. Наприклад, ми могли б отримати наступний виклик: `distVector = sp.pdist(HITIndex, "cosine")`. Результат функції повертається у вигляді вектору значень, який потрібно перетворити у матрицю за допомогою функції у рядку 5.

Після кластеризації та отримання списку кластерів для кожної точки даних відбувається обчислення метрик якості кластеризації, виведення результатів на екран та їх збереження. Для обчислення значень метрик якості також використовуються функції із бібліотеки Scikit-Learn. Наприклад, значення індексу Калінського-Харабаша знаходиться так: `calinski-harabasz = clusterMetrics.calinski_harabasz_score (baseFile, clusterLabels).`

Цікавою є й візуалізація отриманих результатів. У розробленій програмі результати відображаються на графіках двох типів: точковій та стовпчиковій діаграмах. У лістингу 5 наведено програмний код побудови точкової діаграми для виклику (рядки 2-11 елементи графіка, рядок 12 – побудова, рядки 14-16 – збереження і відображення).

Лістинг 5. Побудова точкової діаграми

```
1 def plotPicture(x, y, colors, xaxis, yaxis, legendLabels):
2      plt.figure(figsize = (8,6))
3      ax = plt.subplot()
4      title = xaxis + "/" + yaxis + " points distribution"
5      plt.xlabel(xaxis)
6      plt.ylabel(yaxis)
7      plt.title(title)
8      ax.set_xlim(-0.1,1.05)
9      ax.set_ylim(-0.1,1.05)
10     plt.xticks(np.arange(0,1.1,0.1))
11     plt.yticks(np.arange(0,1.1,0.1))
12     plt.scatter(x, y, c = colors)
13     ax.legend(handles=legendLabels, loc='lower right')
14     picName     =     destinationFolder     +     "\\"     +
"{}_by_{}.png".format(xaxis, yaxis)
15     plt.savefig(picName, dpi = 100, bbox_inches="tight")
16     plt.show()
```

У стовпчиковій діаграмі за висоту стовпців відповідає значення компонент HIT-індексу або ж його суми. У точковій діаграмі вісь X завжди є значенням HIT-індексу для певного респондента, а вісь Y – значеннями того чи іншого індикативного компонента. Кольори точок та стовпців відповідають кластеру, в якому знаходиться респондент.

3.4 Кластеризація підприємств за рівнем цифрової трансформації

Перейдемо до практичного використання розробленої програми для кластеризації наявного набору даних. Зазначимо, що назви компаній-респондентів зашифровано в цілях приватності. Першим кроком використання програми є під’єднання до бази даних для отримання набору

даних – `connect raw table raw_encoded`. Після цього ми можемо переглянути частину отриманого набору даних за допомогою команди `current raw`, що наведено на рисунку 3.5.

```
In [*]:  1  functionMaster(input())

connect raw table raw_encoded
File was successfully loaded into raw table. Please, enter the next command

current raw

Now you are using the follwing file:  raw_encoded.csv

           name_  organization_type  import_export  business_model  \
0    Respondent1                  1              0               0
1    Respondent2                  1              0               0
2    Respondent3                  1              0               0
3    Respondent4                  1              0               0
4    Respondent5                  0              0               0
5    Respondent6                  0              0               0
6    Respondent7                  1              0               0
7    Respondent8                  1              0               0
8    Respondent9                  1              0               0
9   Respondent10                  0              0               0
```

Рисунок 3.5 – Під’єднання до бази даних та перегляд завантаженої таблиці

Наступним кроком підготовки набору даних до кластеризації є його нормалізація. Оскільки ми завантажили «сирі» дані, нам потрібно очистити та привести їх до одного інтервалу. Для цього скористаємося командою `normalize`.

Як один із додаткових кроків щодо оптимізації аналізу даних щодо кластеризації, було вирішено проводити кластеризацію та аналіз визначивши оптимальну кількість кластерів та міру відстані для алгоритмів, які цього вимагають. Оскільки існує значна кількість комбінацій методу, міри відстані та кількості кластерів, аналіз такого масиву інформації вручну за результатами здійснених кластеризацій може займати значний час.

Іншими словами, наступний крок включає в себе виконання кластеризації даних за визначеним методом та мірою відстані, послідовно використовуючи кількість кластерів від 2-х до 8-ми. Для кожної кластеризації обчислюються метрики якості кластеризації (індекси Силуетта, Калінського-Харабаша та Девіеса-Боулдіна). Варто зазначити, що значення індексів

Силуетта та Калінського Харабаша повинно бути якнайвищим, в той же час значення індексу Девіеса-Боулдіна очікується якнайнижчим. Після чого будуються графіки залежності якості кластеризації від кількості кластерів для кожної з метрик. На основі аналізу цих значень вручну обирається оптимальне значення кількості груп. Практично це забезпечується введенням команди з обраного методу кластеризації, метрики відстані, джерела даних та невідомої кількості кластерів, наприклад: `cluster agglomerative euclidean unknown hit`.

У результаті отримаємо три графіки, збережені у створеній директорії на обраному диску. Нехай, розглянемо графіки отримані для агломеративної кластеризації з використанням відстані Евкліда та зв'язності Уорда, зображені на рисунку 3.6.

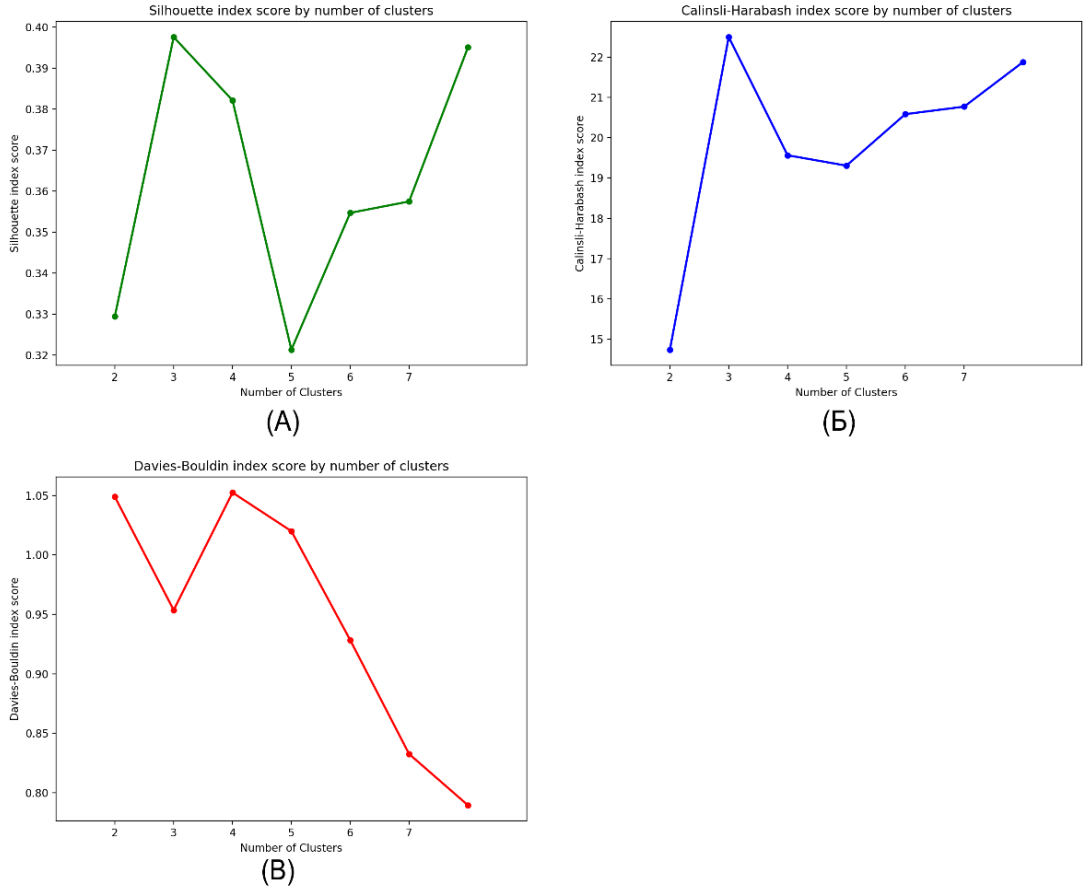


Рисунок 3.6 – Визначення оптимальної кількості кластерів. (А) – Індекс Силуетта, (Б) – Індекс Калінського-Харабаша, (В) – Індекс Девіеса-Боулдіна

Як видно з рисунка вище для згаданої комбінації критеріїв локальні максимуми індексу Силуетта та Калінського-Харабаша досягаються у точках 3 та 8 кластерів. У цих же точках спостерігаються локальні мінімуми для індексу Девіеса-Боулдіна. Значення восьми кластерів видалося занадто великим для набору даних із 34-х точок, тому було обрано значення у 3 кластери.

Оскільки поняття міри відстані використовується лише для двох методів кластеризації: агломеративного та OPTICS, визначення набору критеріїв: відстань + кількість кластерів / сусідів проводилося лише для них. Для кожної можливої міри відстані за принципом описаних вище вручну знаходилась оптимальна кількість кластерів. Після чого серед всіх використаних мір відстані обиралася та, яка показала найкращі результати для поточного методу. Табличний результат такого порівняння для агломеративного методу з використанням повної зв’язності наведено у Таблиці 3.1.

Таблиця 3.1 – Приклад вибору оптимальної міри відстані та кількості кластерів

Міра відстані	Кількість кластерів	Індекс Силуетта	Індекс Девіеса-Боулдіна	Індекс Калінського-Харабаша
Евклідова	3	0.34	1	18
Косинусна	3	0.65	1.4	11
Мангеттенська	7	0.36	0.9	18
Чебишева	4	0.36	1	17
Гаммінга	7	0.13	3	3.5

Така чи подібна оцінка була проведена для кожного використовуваного методу та міри відстані. Як результат, набір оптимальних критеріїв для методів виглядає наступним чином:

- Метод K-means з використанням міри відстані Евкліда та 3 кластерів.

- Агломеративний метод з використанням міри відстані Евкліда, зв’язності Уорда та 3-х кластерів.
- Агломеративний метод з використанням косинусної міри відстані, повної зв’язності та 3-х кластерів.
- Метод OPTICS з використанням міри відстані Чебишева та мінімум 7-ми сусідів для створення кластеру.
- Метод Affinity Propagation – не залежить від кількості кластерів та міри відстані.
- Метод Gaussian Mixture EM – з використанням сферичної коваріації та 3 кластерів.

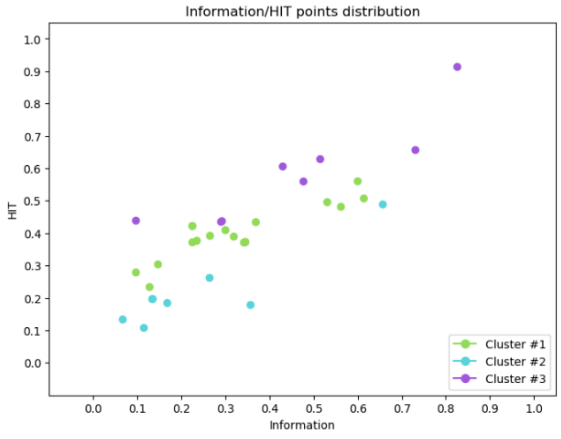
Далі відбувається власне кластеризація обраного набору даних – значень індикативних компонентів HIT-індексу, як вже було зазначено вище. Для цього вводиться команда з вказанням методу кластеризації та за потреби мір відстані і кількості кластерів, наприклад: `cluster kmeans hit 3`. Після цього відбувається обчислення HIT-індексу для завантаженого набору даних, виконання кластеризації, обчислення метрик якості, візуалізація а завантаження результатів на комп’ютер користувача.

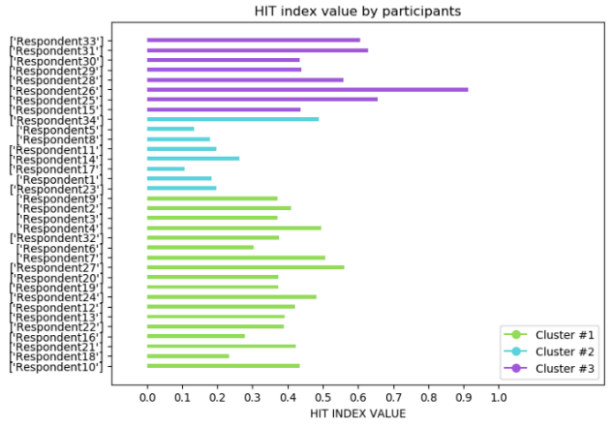
Розглянемо та проаналізуємо результати кластеризації за наборами критеріїв, згаданими вище. Для цього використовуватимемо: мітки утворених кластерів, базовий набір даних для ручного обчислення деяких додаткових параметрів, графік розподілу точок за значеннями HIT-індексу. Для кожного з проведених методів було утворено зведену аналітичну таблицю (знаходиться унизу рисунків 3.8 – 3.13). Також на рисунках 3.8 – 3.13 зобразимо точкову діаграму розподілу точок та кластерів у розрізі відношення значення використання цифрових інструментів до значення HIT-індексу (зліва) та стовпчикову діаграму розподілу членів кластеру за значенням індексу (справа).

1. Кластеризацією за алгоритмом K-means набір даних було розділено на 3 кластери. Як видно з рисунка 3.8, кластери майже не перетинаються між

собою та містять достатньо схожі елементи всередині. Чітко виділяється другий кластер (блакитні точки), який знаходиться у нижній частині графіка значень HIT-індексу до цифрових інструментів. Також цей кластер отримав досить низькі значення власне індексу трансформації. Кластер №1 містить більшість точок, які знаходяться у середині інтервалів як значення HIT-індексу так і використання цифрових інструментів. А кластер №3 відзначається найвищими значеннями індексу трансформації.

Щодо характеристик окремих кластерів, то члени кластеру №1 частково ефективні у використанні соцмереж, проте не використовують власні сайти, інструменти реклами чи аналітики та технічні системи, маючи при цьому достатнє технічне забезпечення; грамотність людського капіталу також досягає лише початкового рівня.





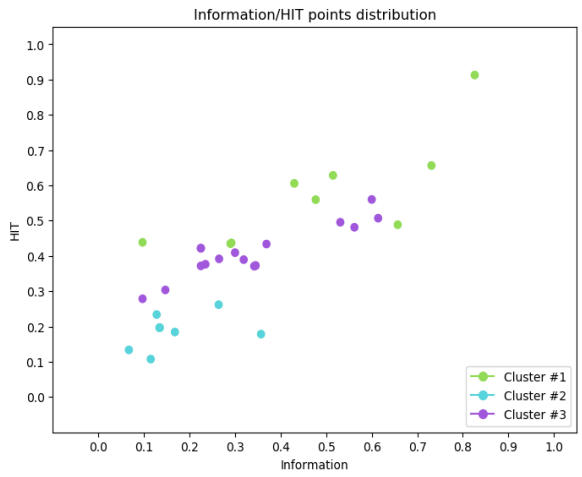
K-means, відстань Евкліда, 3 кластери	Кластер 1 (18)				Кластер 2 (8)				Кластер 3 (8)			
	H [0; 0,364]	I [0.128; 0.614]	T [0,7; 1]	Сума [0,234; 0,56]	H [0; 0,364]	I [0,067; 0,657]	T [0; 0,5]	Сума [0,11; 0,488]	H [0,636; 1]	I [0,29; 0,826]	T [0,5; 1]	Сума [0,44; 0,91]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Не оптимізовано		61.1%		Не оптимізовано		70.0%		Оптимізовано		70.0%	
Наявність та ефективність соціальних мереж	Ефективно		50.0%		Не ефективно		70.8%		Ефективно		70.0%	
Використання онлайн-реклами та аналітики	Не використовується		74.1%		Не використовується		91.6%		Використовується		58.3%	
Використання спеціалізованих систем менеджменту	Не використовується		80.2%		Не використовується		73.2%		Не використовується		71.4%	
Використання спеціалізованих технічних систем	Не використовується		96.4%		Не використовується		79.2%		Не використовується		87.5%	
Рівень технічного забезпечення	Задовільний		98.1%		Не задовільний		62.5%		Задовільний		83.3%	
Рівень цифрової грамотності	Початковий		50.0%		Початковий		62.5%		Середній чи вище		87.5%	
Канали комунікації	З використанням ІКТ		74.7%		З використанням ІКТ		83.3%		З використанням ІКТ		75.0%	
Індекс Силуетта	**0.411**											
Індекс Калінського-Харабаша	**24.105**											
Індекс Девіса-Боулдіна	**0.889**											

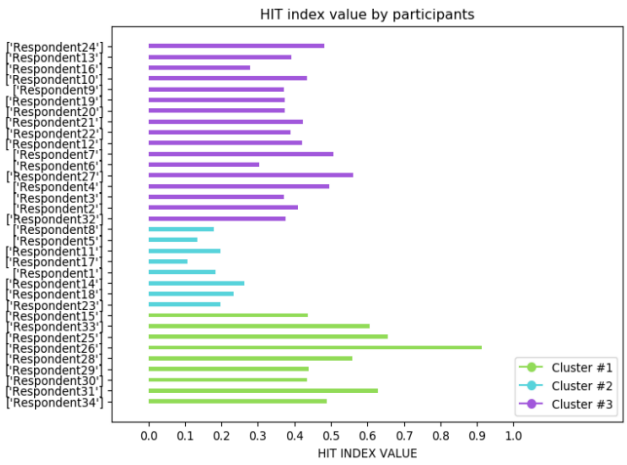
Рисунок 3.8 – Кластеризація K-Means з мірою Евкліда та трьома кластерами

В той же час кластер №2 показує такі ж показники, як і кластер №1 за винятком того, що соціальні мережі не використовують або використовують не ефективно, а також у компаній відсутнє достатнє технічне забезпечення.

На противагу їм, кластер №3 містить респондентів, що більш ефективно використовують необхідні цифрові інструменти: сайт, соцмережі, рекламу, а також мають достатню грамотність людського капіталу.

2. Використання агломеративного методу, міри відстані Евкліда та зв'язності Уорда дозволило отримати досить оптимальний результат для поділу на 3 кластери. На рисунку 3.9 наведена візуалізація розподілів та аналітична таблиця для цього методу. Можна помітити досить хороше відокремлення кластера №2 (блакитні точки), який містить у собі респондентів з найнижчими значеннями індексу. Також кластери №1 та №3 є достатньо рознесеними у просторі, хоча й накладаються у кількох точках.





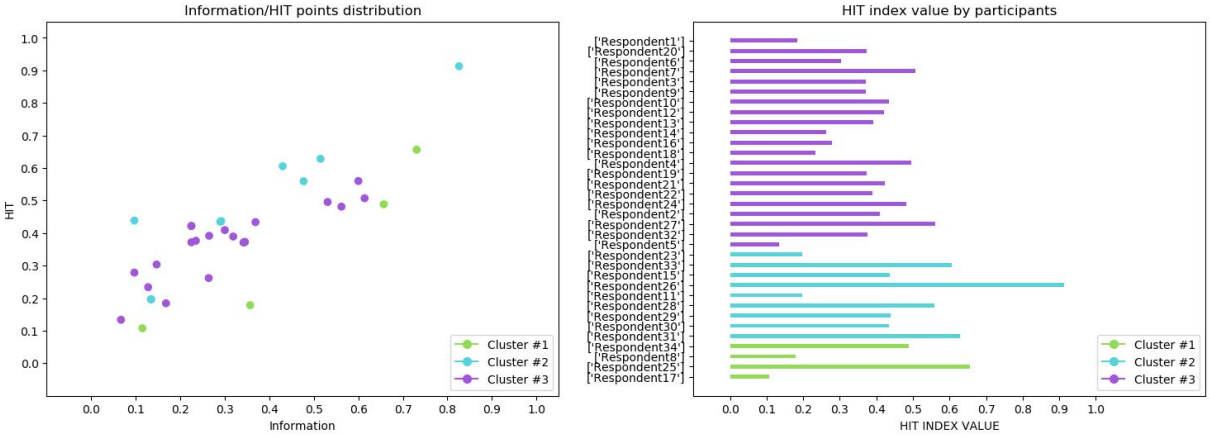
Агломеративна кластеризація, зв'язність Уорда, 3 кластери	Кластер 1 (9)				Кластер 2 (8)				Кластер 3 (17)			
	H [0.2; 1]	I [0.097; 0.826]	T [0.25, 1]	Сума [0.43; 0.91]	H [0; 0.364]	I [0.067; 0.357]	T [0; 0.7]	Сума [0.107; 0.262]	H [0; 0.364]	I [0.097; 0.614]	T [0.75; 1]	Сума [0.2785; 0.56]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Оптимізовано		68.9%		Не оптимізовано		80.0%		Не оптимізовано		60.0%	
Наявність та ефективність соціальних мереж	Ефективно		70.3%		Не ефективно		75.0%		Ефективно		51.0%	
Використання онлайн-реклами та аналітики	Не використовується		55.6%		Не використовується		100.0%		Не використовується		72.5%	
Використання спеціалізованих систем менеджменту	Не використовується		65.1%		Не використовується		82.1%		Не використовується		79.8%	
Використання спеціалізованих технічних систем	Не використовується		88.9%		Не використовується		79.2%		Не використовується		98.0%	
Рівень технічного забезпечення	Задовільне		77.8%		Не задовільне		58.3%		Задовільне		100.0%	
Рівень цифрової грамотності	Середній чи вище		83.3%		Початковий		75.0%		Початковий		70.6%	
Канали комунікації	З використанням ІКТ		77.8%		З використанням ІКТ		75.0%		З використанням ІКТ		76.5%	
Індекс Силуетта	**0.398**											
Індекс Калінського-Харабаша	**22.497**											
Індекс Девіеса-Боулдіна	**0.954**											

Рисунок 3.9 – Агломеративна кластеризація зі зв'язністю Уорда та трьома кластерами

Щодо характеристик трьох кластерів, то члени кластеру №1, який належить до області з найвищими значеннями індикаторів, ефективно використовують сайт та соціальні мережі, а також мають рівень індикатора

цифрової грамотності на середньому чи вище середнього рівня для більш, ніж ¾ респондентів. На противагу йому, кластер №2 характеризується неефективністю використання цифрових інструментів для більшості членів, а також низькою цифровою грамотністю та незадовільним технічним забезпеченням для більш, ніж половини опитаних. Кластер №3 має певну ефективність соцмереж при низьких показниках інших індикаторів включно з початковим рівнем цифрової грамотності працівників.

3. Використання агломеративної кластеризації з повною зв’язністю та косинусною мірою подібності показало найкращий результат на 3х кластерах. До того ж цей варіант видав найвище значення індексу Силуетта – 0.65. Результат кластеризації наведено на рисунку 3.10.



Агломеративна кластеризація, повна зв'язність, косинусна міра, 3 кластери	Кластер 1 (4)				Кластер 2 (9)				Кластер 3 (21)			
	H [0; 0,636]	I [0,115;0,731]	T [0;0,5]	Сума [0,107;0,6562]	H [0,264; 1]	I [0,097; 0,826]	T [0.1; 1]	Сума [0.196; 0.913]	H [0; 0.36]	I [0.067; 0.61]	T [0.5; 1]	Сума [0.133; 0.56]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Оптимізовано		55%		Оптимізовано		51.1%		Не оптимізовано		53.3%	
Наявність та ефективність соціальних мереж	Ефективно		50%		Ефективно		59.2%		Не ефективно		53.9%	
Використання онлайн-реклами та аналітики	Не використовується		58.3%		Не використовується		74%		Не використовується		77.8%	
Використання спеціалізованих систем менеджменту	Використовується		54.2%		Не використовується		75.9%		Не використовується		78.6%	
Використання спеціалізованих технічних систем	Не використовується		83.3%		Не використовується		92.6%		Не використовується		92%	
Рівень технічного забезпечення	Не задовільний		66.7%		Задовільний		77.8%		Задовільний		92.1%	
Рівень цифрової грамотності	Початковий		50%		Середній чи вище		83.3%		Початковий		54.7%	
Канали комунікації	З використанням ІКТ		87.5%		З використанням ІКТ		72.2%		З використанням ІКТ		71.4%	
Індекс Силуетта	**0.65**											
Індекс Калінського-Харабаша	**11.336**											
Індекс Девіеса-Боулдіна	**1.405**											

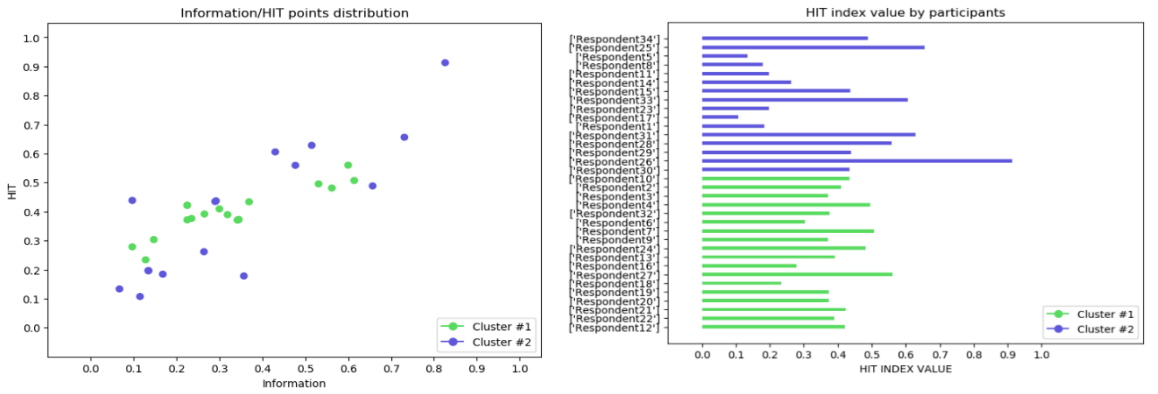
Рисунок 3.10 – Агломеративна кластеризація з повною зв’язністю та коснусною відстанню, 3 кластери

З кластеризації можемо помітити досить однорідний кластер №3 (фіолетові точки), який включає в себе респондентів з середнім значенням

індексу HIT. З іншого боку розрідженим є кластер №2, який охоплює і членів з високим рівнем зрілості, так і кількох з досить низьким. Інший кластер №1 охоплює досить цікавий край графіка, включаючи в себе різнорідні точки.

Щодо характеристик кластерів, то у першому знаходиться 4 члени і зважаючи на їхнє взаємне розташування результати не є в повній мірі адекватними. Респонденти у цьому кластері на близько 50% використовують сайт, соцмережі та спеціалізовані системи менеджменту. При чому рівень технічного забезпечення у них незадовільний, а цифрова грамотність – на початковому рівні. Кластер №2 також характеризується застосуванням соціальних мереж та веб-сайту, проте його члени мають задовільний рівень інформаційної інфраструктури та знань людського капіталу. Третій і найбільший кластер містить 21 респондента, у якому більш, ніж половина учасників не використовує цифрові інструменти та має початковий рівень цифрової грамотності.

4. Ще одним із використаних методів став OPTICS з використанням міри відстані Чебишева та мінімум 7-ми точок для формування кластеру. Не зважаючи на отримане оптимальне значення за метриками якості, власне кластеризація не виявилася успішною з практичної точки зору. На візуаліації рисунку 3.11 помітно, що кластери містять майже однакову кількість членів. Окрім цього, кластери розподілилися як внутрішній та зовнішній, що унеможливлює встановлення принципових відмінностей між ними як видно з аналітичної таблиці. Причиною такого результату є те, що OPTICS належить до алгоритмів на основі щільності, а базовий набір даних не містить щільних областей. У такому разі внутрішній кластер (зелений), виявився штучною областю зі щільними значеннями, в той час як зовнішній було промарковано викидами, тобто значеннями які не несуть цінності.

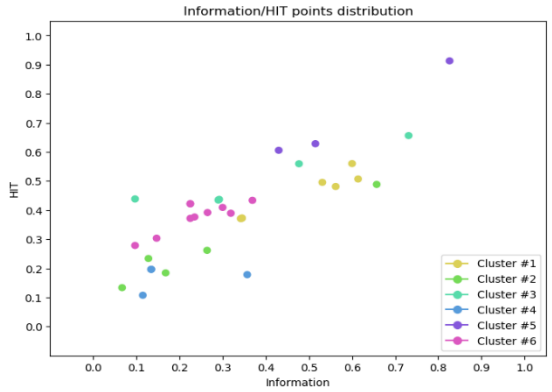


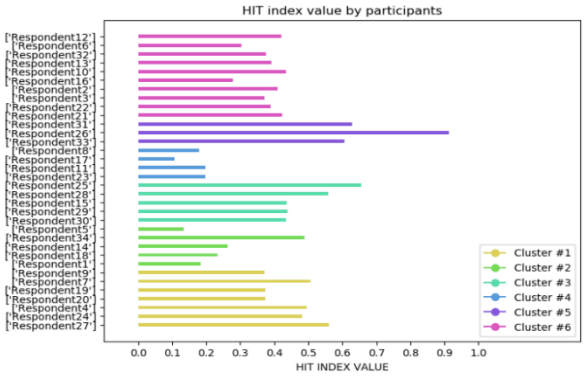
Кластеризація OPTICS, відстань Чебишева, 7 сусідів	Кластер 1 (18)				Кластер 2 (16)			
	H	I	T	Сума	H	I	T	Сума
	[0; 0.364]	[0.128; 0.614]	[0.7; 1]	[0.233; 0.56]	[0; 1]	[0.067;0.826]	[0;1]	[0.13; 0.91]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Не оптимізовано		61.1%		Не оптимізовано		51.3%	
Наявнсть та ефективність соціальних мереж	Ефективно		50.0%		Ефективно		50.0%	
Використання онлайн-реклами та аналітики	Не використовується		74.1%		Не використовується		75.0%	
Використання спеціалізованих систем менеджменту	Не використовується		80.2%		Не використовується		72.3%	
Використання спеціалізованих технічних систем	Не використовується		96.3%		Не використовується		85.4%	
Рівень технічного забезпечення	Задовільний		100.0%		Задовільний		60.4%	
Рівень цифрової грамотності	Початковий		69.4%		Середній чи вище		56.3%	
Канали комунікації	З використанням ІКТ		74.1%		З використанням ІКТ		79.2%	
Індекс Силуетта	**0.327**							
Індекс Калінського-Харабаша	**13.554**							
Індекс Девіеса-Боулдіна	**1.408**							

Рисунок 3.11 — Кластеризація OPTICS з відстанню Чебишева та 7 сусідами

5. Метод Affinity Propagation не залежить від кількості кластерів та міри відстані, тому його результати представляють використання власне структури даних без будь-якого впливу користувача. Як видно з рисунка 3.12 дані було розподілено на 6 кластерів. Деякі із кластерів (наприклад, №№1, 5 та 6) є достатньо відокремленими від інших. У той же час кластери №№2, 3, 4 дещо перекриваються між іншими кластерами. У розподілі респондентів за значенням HIT-індексу чітко виділяється кластер-лідер – №5, а також кластер з найнижчими значеннями – №2 та №4. Кластери № 1, №3 та №6 складаються з респондентів з середнім та вище середнього значеннями індексу.

Щодо характеристик окремих кластерів, то кластери №№ 1, 3 та 5 є достатньо схожими між собою, як можна помітити із таблиці вище. Проте, цікавим є те, що 2/3 участинків кластера №1 ведуть достатню успішну діяльність з використання сайту та соціальних мереж, маючи початковий рівень грамотності людського капіталу.



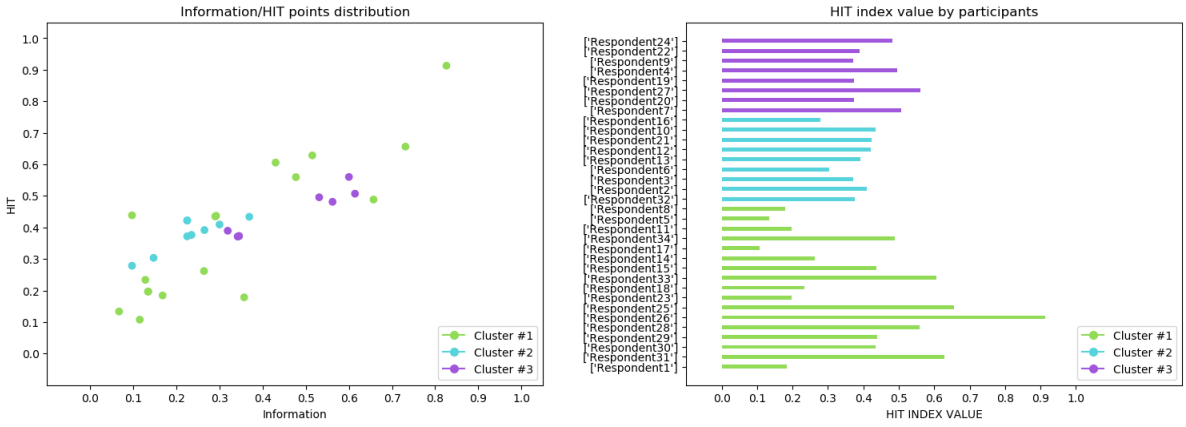


Affinity Propagation метод	Кластер 1 (7)				Кластер 2 (5)				Кластер 3 (5)			
	H	I	T	Сума	H	I	T	Сума	H	I	T	Сума
	[0;0.2]	[0.34;0.61]	[1]	[0.37;0.56]	[0; 0.2]	[0.067; 0.657]	[0.5; 0.7]	[0.13;0.488]	[0.636; 0.8]	[0.097; 0.73]	[0.25; 0.75]	[0.434; 0.656]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Оптимізовано		60.0%		Не оптимізовано		64.0%		Оптимізовано		52.0%	
Наявність та ефективність соціальних мереж	Ефективно		66.7%		Не ефективно		66.6%		Ефективно		53.3%	
Використання онлайн-реклами та аналітики	Не використовується		52.4%		Не використовується		86.7%		Не використовується		60.0%	
Використання спеціалізованих систем менеджменту	Не використовується		81.6%		Не використовується		71.4%		Не використовується		74.2%	
Використання спеціалізованих технічних систем	Не використовується		100.0%		Не використовується		66.6%		Не використовується		93.3%	
Рівень технічного забезпечення	Задовільний		100.0%		Задовільний		53.3%		Задовільний		86.7%	
Рівень цифрової грамотності	Початковий		85.7%		Початковий		70.0%		Середній чи вище		80.0%	
Канали комунікації	З використанням ІКТ		85.7%		З використанням ІКТ		66.6%		З використанням ІКТ		73.3%	
	Кластер 4 (4)				Кластер 5 (3)				Кластер 6 (10)			
	H	I	T	Сума	H	I	T	Сума	H	I	T	Сума
	[0; 0.36]	[0.115; 0.357]	[0; 0.25]	[0.1075; 0.197]	[0.636; 1]	[0.43; 0.826]	[0.9; 1]	[0.606; 0.913]	[0.1; 0.36]	[0.097; 0.369]	[0.75; 1]	[0.279; 0.434]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Не оптимізовано		85.0%		Оптимізовано		93.3%		Не оптимізовано		72.0%	
Наявність та ефективність соціальних мереж	Не ефективно		75.0%		Ефективно		100.0%		Ефективно		60.0%	
Використання онлайн-реклами та аналітики	Не використовується		100.0%		Не використовується		55.5%		Не використовується		86.7%	
Використання спеціалізованих систем менеджменту	Не використовується		78.6%		Не використовується		66.6%		Не використовується		78.6%	
Використання спеціалізованих технічних систем	Не використовується		91.6%		Не використовується		88.9%		Не використовується		96.7%	
Рівень технічного забезпечення	Не задовільний		75.0%		Задовільний		100.0%		Задовільний		100.0%	
Рівень цифрової грамотності	Початковий		75.0%		Середній чи вище		100.0%		Початковий		60.0%	
Канали комунікації	З використанням ІКТ		91.7%		З використанням ІКТ		77.8%		З використанням ІКТ		70.0%	
Індекс Силуетта	**0.351**											
Індекс Калінського-Харабаша	**21.607**											
Індекс Девіса-Боулдіна	**0.931**											

Рисунок 3.12 – Результати кластеризації методом Affinity Propagation

Напротивагу цьому, кластер №4 має високе значення цифрової грамотності, проте лише трохи більше половини учасників успішно використовують цифрові технології (зважаючи на розмір кластера це може входити у статистичну похибку). Кластер №5 є найменшим, проте складається з респондентів з найвищим рівнем використання цифрових інструментів та значенням індексу трансформації. Кластер №2 та кластер №4 характеризуються неефективним використанням цифрових ресурсів. Відмінність між ними полягає у значенні індикатора цифрової грамотності. Цікавим є також кластер №6 у якому було відзначено ефективність роботи соцмереж при низьких показниках інших індикаторів.

6. Останнім з методів кластеризації було застосовано Gaussian Mixture Expectation-Maximization алгоритм, що розділив набір даних на 3 кластери, візуалізацію яких наведено на рисунку 3.13.



Gaussian Mixture EM алгоритм	Кластер 1 (17)				Кластер 2 (9)				Кластер 3 (8)			
	H [0; 1]	I [0,067; 0,826]	T [0; 1]	Сума [0,13; 0,91]	H [0,1; 0,364]	I [0,097; 0,369]	T [0,75; 1]	Сума [0,279; 0,434]	H [0; 0,2]	I [0,319; 0,614]	T [1]	Сума [0,371; 0,56]
	Відповідь		Відсоток випадків		Відповідь		Відсоток випадків		Відповідь		Відсоток випадків	
Наявність, оптимізація та ефективність сайту	Не оптимізовано		54.1%		Не оптимізовано		73.3%		Оптимізовано		55%	
Наявність та ефективність соціальних мереж	Ефективно		50%		Не ефективно		70.3%		Ефективно		62%	
Використання онлайн-реклами та аналітики	Не використовується		76.5%		Не використовується		92.6%		Не використовується		56.2%	
Використання спеціалізованих систем менеджменту	Не використовується		70.5%		Не використовується		77.8%		Не використовується		83.3%	
Використання спеціалізованих технічних систем	Не використовується		86.3%		Не використовується		96.3%		Не використовується		100%	
Рівень технічного забезпечення	Задовільний		60.8%		Задовільний		100%		Задовільний		100%	
Рівень цифрової грамотності	Середній чи вище		67.8%		Початковий		100%		Початковий		81.2%	
Канали комунікації	З використанням ІКТ		74.2%		З використанням ІКТ		61%		З використанням ІКТ		93.8%	
Індекс Силуетта	**0.192**											
Індекс Калінського-Харабаша	**8.578**											
Індекс Девіеса-Боулдіна	**1.352**											

Рисунок 3.13 – Expectation-Maximization кластеризація, 3 групи

Як бачимо, щільним є кластер №2 (блакитні точки), його значення індексу HIT знаходяться у середньому інтервалі, так як і використання цифрових інструментів. Трохи вищі значення можемо спостерігати у кластера №3, який також є досить згрупованим між собою. На противагу їм, найбільший кластер №1 є дуже розрідженим і містить як точки із найнижчими, так і найвищими значеннями компонент індексу HIT. На рисунку можемо побачити його як зовнішнє коло зелених точок. Причиною такого розподілу швидше за все є те, що початковий набір даних є далеким від нормального Гауссового розподілу.

Якщо говорити про характеристики кластерів, то у першому половиною респондентів не використовуються цифрові інструменти, при цьому майже 70% опитаних заявляють про середній чи високий рівень цифрової грамотності. У кластері №2 більшою частиною не використовуються сучасні можливості, при чому у всіх респондентів початковий рівень володіння ними за наявності технічних засобів. Третій кластер виявляє помірний успіх у застосуванні простих інструментів, таких як веб-сайт та соціальні мережі за

умови того, що 80% респондентів вважають рівень цифрових компетенцій своїх співробітників початковим. Ще одним спостереженням є те, що основні характеристики кластерів часто перебувають на рівні 50% – тобто половина респондентів використовує, наприклад, аналітику, а половина – ні, що унеможливлює виокремлення точних відмінних рис між кластерами.

Розглянувши наведені та додаткові візуалізації, а також проаналізувавши таблиці, отримані на основі необроблених даних та HIT-індексу, можна зробити такі висновки щодо тенденцій у наборі даних. По перше, чималий вплив на значення індексу має значення індикатора цифрової грамотності працівників. При початковому рівні грамотності відзначається невикористання сайтів, соцмереж та інших інструментів. З підвищенням умінь персоналу зростає й відсоткове співвідношення використання інструментів та їх ефективність, тож інвестиції у людей видаються важливим внеском в успіх цифровізації. Цікавим є те, що рівень технічного забезпечення не несе значного впливу. Звісно, при повній відсутності техніки рівень використання інструментів є найнижчим. Проте, його підвищення не видається закономірним наслідком покращення технічного стану.

Варто відзначити, що на більшості підприємств не використовуються складні системи менеджменту (такі як системи логістики, управління поставками, життєвим циклом виробництва, планування ресурсів тощо) та спеціалізоване технічне забезпечення (GPS-трекери, 3D-друк). З одного боку це можна пояснити відсутністю такої необхідності через специфіку бізнесу, а з іншого – ефективна робота з такими системами вимагає високого рівня вмінь працівників, а також коштів на закупівлю та підтримку програмного забезпечення. До того ж, часто прості та середньої складності інструменти так, як соцмережі, веб-сайт, реклама чи аналітика приносять більш вимірювані результати.

Щодо результатів кластеризації, то важливим є те, що не завжди високі значення метрик якості свідчать про оптимальний та ефективний поділ на

групи, який можна логічно описати. Наприклад, агломеративна кластеризація повного зв’язку з косинусною мірою відстані показала найкраще значення індексу Силуетта у 0.65, але її результати було досить складно охарактеризувати. Тому з точки зору подібності елементів усередині груп та відмінностей між кластерами, найкращі результати продемонстрували алгоритми Affinity Propagation, агломеративна кластеризація зі зв’язністю Уорда та 3-ма групами, а також K-Means з поділом на 3 кластери.

3.5 Висновки до третього розділу

Третій розділ кваліфікаційної роботи присвячено огляду розробленого програмного забезпечення та аналізу результатів проведеної кластеризації опитаних представників мікро, малого та середнього підприємництва Тернопільської області.

Коротко розглянуто набір даних, що використовується в роботі, його структуру та особливості. Описано функціональні можливості розробленого застосунку, включаючи завантаження файлів, їх очищення та нормалізацію, обчислення та кластеризацію, а також візуалізацію отриманих результатів у вигляді точкової та стовпчикової діаграм. Подано відомості про обрану мову програмування Python, а також її бібліотеки, спеціальне середовище розробки та систему управління базою даних. Також у розділі висвітлено алгоритми та принципи роботи коду окремих функціональних частин програми.

У практичній частині проведено кластеризацію даних на основі індикативних компонентів HIT за визначеними оптимальними наборами критеріїв для шести алгоритмів, візуалізовано та проаналізовано отримані результати, а також визначено найкращі алгоритми серед використаних. Окрім цього отримано певні нові ідеї щодо важливості окремих індикативних компонентів HIT-індексу, а саме цифрової грамотності людського капіталу.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Аналіз регулятивних вимог щодо охорони праці у мікро, малих та середніх підприємствах в Україні та країнах Європи

Оскільки тематика дипломної роботи присвячена використанню технік видобутку даних та подальшому аналізу результатів кластеризації малих та середніх підприємств Тернопільської області, важливим є розглянути регуляції щодо організації охорони праці на малих та середніх підприємствах України та закордону. Обізнаність як працівників, так і роботодавця з практиками та вимогами до організації робочих місць, своїми правами та обов’язками є одним із поки що нерозглянутих аспектів грамотності персоналу. Якщо пов’язувати це з цифровою трансформацією бізнесу, то електронний облік інцидентів, правил, вимог та інструктаж працівників з техніки безпеки у вигляді цифрового тренажеру може стати одним із цікавих предметів подальших досліджень.

Згідно зі статтею 1 Закону України «Про охорону праці» визначається наступне: «Охорона праці — це система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини у процесі трудової діяльності». [67]

У свою чергу, Господарський Кодекс України (стаття 55, частина 3) [68] та Закон України «Про бухгалтерський облік та фінансову звітність в Україні» (стаття 2, частина 2) [69] визначають поняття малого, середнього та великого підприємництва, а також мікропідприємництва у розрізі балансових активів, розміру різних типів прибутку та кількості найманих працівників. До малих та мікропідприємств можуть входити як фізичні, так і юридичні особи, в той час як до середніх та великих – лише юридичні. Зведена таблиця критеріїв входження підприємства до певної групи наведена нижче:

Таблиця 4.1 – Критерії входження підприємства до групи

Група підприємства	Балансова вартість активів	Дохід від будь-якої діяльності	Чистий дохід від реалізації	Середня кількість працівників
Мікропідприємства	до 350 тис. євро	до 2 млн. євро	до 700 тис. євро	до 10 осіб
Малі	до 4 млн. євро	до 10 млн. євро	до 8 млн. євро	до 50 осіб
Середні	до 20 млн. євро	до 50 млн. євро	до 40 млн. євро	до 250 осіб
Великі	понад 20 млн. євро	понад 50 млн євро	понад 40 млн. євро	понад 250 осіб

Окрім цього варто зазначити, що малі та мікропідприємства займають значну частку усіх організацій України. Адже згідно з даними Державної Служби Статистики України у найбільш свіжому звіті за 2020 рік [70] в Україні налічувалося 512 великих підприємств, 17 620 — середніх, та 355 708 малих підприємств (з яких виділяють 307 871 мікропідприємств). У відсотковому співвідношенні частка великих підприємств становить 0.1%, середніх — 4.7%, а малих та мікропідприємств — 95.2%. Зважаючи на це, забезпечення належних умов праці та захист прав працівників такої численної групи є важливим завданням у політиках та стратегіях країни.

4.1.1 Регулювання роботи служб з охорони праці в Україні

Питання охорони праці на підприємствах регулюються на законодавчому рівні, і серед інших нормативних документів, що визначають положення і процедури щодо безпеки працівників, основним є Закон України «Про охорону праці». Окрім загальних політик, врегулювання фінансування, медичних оглядів, забезпечення засобами захисту тощо, закон надає інформацію про службові органи, що функціонують на підприємстві, такі як: служба з охорони праці та комісія з питань охорони праці підприємства. Функції регулятивного органу щодо дотримання законодавства та рекомендацій покладаються на Державну службу України з питань праці.

Згідно зі статтею 15 Закону України «Про охорону праці»: «На підприємстві з кількістю працюючих 50 і більше осіб роботодавець створює

службу охорони праці [….] . На підприємстві з кількістю працюючих менше 50 осіб функції служби охорони праці можуть виконувати в порядку сумісництва особи, які мають відповідну підготовку. На підприємстві з кількістю працюючих менше 20 осіб для виконання функцій служби охорони праці можуть залучатися сторонні спеціалісти на договірних засадах, які мають відповідну підготовку» [67].

Створення окремої комісії з питань охорони праці підприємства регулює стаття 16 того ж Закону України. У ній вказано, що: «[…] з метою забезпечення пропорційної участі працівників у вирішенні будь-яких питань безпеки, гігієни праці та виробничого середовища […] може створюватися комісія з питань охорони праці. Комісія складається з представників роботодавця та професійної спілки, а також уповноваженої найманими працівниками особи, спеціалістів з безпеки, гігієни праці та інших служб підприємства […]» [67].

Необхідна кваліфікація, посадові обов’язки та права таких осіб затверджуються окремими положеннями та документами серед яких можна виділити Типове положення про службу охорони праці.

4.1.2 Регулювання роботи служб з охорони праці в країнах Європи

Щодо країн Європейського Союзу, одним із цікавих спостережень є те, що окрім профільних міністерств та служб, існує окрема інстанція спільна для усіх країн союзу, яка фокусується на модернізації робочого простору країн Єврозони. Європейське агентство з безпеки та гігієни праці (EU-OSHA, European Occupational Safety and Health Administration) — це децентралізована агенція Європейського Союзу, яку було створено у 1994 році. Її завданнями є збір, аналіз та поширення інформації щодо безпеки та здоров’я працівників на робочих місцях [71]. Базові рекомендації агенції стали основою для законодавства щодо умов та охорони праці у країнах-

членах ЄС, а щомісячні аналітичні звіти дозволяють коригувати поточні візії, акти, політики та стратегії щодо гігієни та безпеки праці.

У нещодавньому дослідженні «Покращення безпеки та здоров'я на мікро, малих і середніх підприємствах: огляд ініціатив та механізмів реалізації» від Міжнародної Організації Праці [72] визначаються європейські критерії щодо входження організації у категорію мікро, малого чи середнього підприємства (ММСП), які співпадають з із українськими вимогами, наведеними у таблиці 4.1 вище. Дослідження також відзначає, що ММСП становлять переважну більшість підприємств у всьому світі і, на жаль, часто характеризуються своєю нестабільною природою та поганими умовами безпеки та гігієни праці. Такі соціально-економічні фактори, у поєднанні з низьким рівнем обізнаності та дотриманням стандартів охорони праці, роблять працівників цих підприємств особливо схильними до ризику для їх безпеки та здоров’я.

Оскільки дослідження стосувалося усього світу: як і розвинутих європейських держав, так і країн Середнього Сходу та Південно-Східної Азії, було зроблено висновок, що ММСП дуже часто входять до неформального сектору (тобто, не є зареєстрованими організаціями) і їх працівники менш освічені та проінформовані про свою роботу, що збільшує шанси бути підданим робочим ризикам та незадовільним і шкідливим умовам праці. До вразливих груп найчастіше відносять жінок, дітей та підлітків, людей похилого віку, а також трудових мігрантів [72]. Дослідження визначає такі проблеми ММСП у розрізі виконання регулятивних документів щодо охорони праці:

- Досить часто мікро-, малі та середні підприємства не представлені в національних стратегіях та політиках, через що стандарти охорони праці не є адаптованими до таких форм підприємництва.
- Державні установи та інспекції з питань праці часто не надто ефективні у забезпеченні дотримання норм охорони праці у ММСП.

- ММСП, як правило, обмежені у ресурсах, зосереджені на поточних питаннях забезпечення бізнесу та мають неформальну структуру управління.

Перелік кроків, що можуть посприяти вирішенню проблеми забезпечення відповідних умов праці ММСП у депресивних районах також описано в одному із розділів дослідження [72]. Разом із тим наведено приклади щодо регуляцій, ґрантів та інших способів залучення ММСП в охорону праці на прикладі розвинутих країн Європи, Америки та Середнього Сходу. Деякі із кроків визначаються наступними:

- Адаптація правил охорони праці та спрощення бюрократії для ММСП, тобто відділення від інших видів регулювання.
- Посилення нормативно-правової бази, сприяння та забезпечення її дотримання та виконання приписів.
- Активна участь соціальних партнерів – створення посередників з охорони праці та надання консультаційних послуг.
- Механізми заохочення серед яких економічні, фінансові, страхові, соціальні гарантії, винагороди та визнання тощо.

4.1.3 Порівняння регулятивних вимог щодо служб з охорони праці у деяких країнах Європи

Коротко розглянувши деякі проблеми ММСП щодо охорони праці у менш розвинутих країнах та шляхи їх вирішення, важливим є ознайомлення з законодавством, що визначає наявність особи, відповідальної за дотримання положень щодо охорони праці на підприємстві (аналог до статей 15-16 Закону України «Про охорону праці» [67]) у європейських країнах, які послуговуються актами EU-OSHA. Варто зазначити, що аналогічно з Україною, окремі акти, розділи та параграфи європейського законодавства регулюють питання кваліфікації, навчання та посадових обов’язків таких осіб за їх наявності.

Під час аналізу було розглянуто 11 європейських країн, обраних за принципом різноманітності географічного розташування та економічного розвитку: Австрія, Чеська Республіка, Естонія, Фінляндія, Литва, Румунія, Греція, Німеччина, Ірландія. Інформацію щодо законодавчих актів та принципів забезпечення охорони праці на підприємствах було отримано із законодавства обраних країн, а також з порталу EU-OSHA [73] та дослідження Л. Фалтона, опублікованого на порталі Worker-Participation [74]. Зведені результати аналізу представлені у таблиці 4.2 нижче.

Таблиця 4.2 – Порівняння вимог щодо служб охорони праці в країнах Європи

Країна	**Основні законодавчі акти**	**Відповідальні органи**	**Організація служби охорони праці на виробництві**
Австрія	Закон про безпеку та охорону здоров'я на виробництві [75]	• Федеральне міністерство праці, соціального забезпечення та захисту споживачів • Інспекція праці	• ≥ 5 працівників – створюється робоча рада з представників роботодавця; • ≥ 10 працівників – до ради обираються представники з числа співробітників; • ≥ 100 працівників – створюється спільний комітет з охорони праці
Чеська Республіка	Закон № 262/2006, КЗпП Чехії [76], розділ 3	• Міністерство праці та соціальних справ • Міністерство охорони здоров’я • Інспекція праці	• Представництво працівників з питань охорони праці здійснюється через профспілкову організацію на робочих місцях або виборних представників з питань охорони праці. • Немає законодавчого зобов’язання щодо створення комітету з охорони праці.
Естонія	Закон про гігієну та безпеку праці прийнятий [77], параграф 17	• Міністерство праці та соціальних справ	• < 10 працівників — консультації з охорони праці проводить роботодавець; • ≥ 10 працівників – обирається принаймні один працівник з питань охорони праці; • ≥ 50 працівників АБО за небезпечних умов праці – створюється комітет з охорони праці.

Продовження таблиці 4.2

Фінляндія	Закон про охорону праці та охорону здоров'я та співробітництво з питань охорони праці на робочих місцях № 44/2006 [78], розділ 29	• Міністерство соціальних справ і охорони здоров'я Фінляндії • Фінський інститут гігієни праці	• < 10 працівників – опціонально, консультаціями займається роботодавець; • ≥ 10 працівників – має бути обраний представник з охорони праці плюс два заступники; • ≥ 20 працівників – створюється комісія з охорони праці з 4, 8 або 12 членів
Литва	Закон про охорону здоров'я та безпеку праці №IX-1672 [79], розділ 13	• Міністерство соціального захисту та праці (Держінспекція праці) • Міністерство охорони здоров'я	• < 50 працівників – комітет може бути створений на вимогу більш, ніж половини працівників АБО за небезпечних умов роботи; • ≥ 50 працівників – спільний комітет роботодавця та працівників з охорони праці.
Румунія	Закон № 319 Закон «Про безпеку та охорону праці» [80], розділ 6	• Міністерство праці, сім'ї та соціального захисту	• ≥ 10 працівників – обираються представники з питань охорони праці; • ≥ 50 працівників – разом з роботодавцем створюється спільний комітет з охорони праці
Греція	Закон 3850/10 «Ратифікація Кодексу законів про охорону здоров'я та безпеку працівників» [81], розділ 8	• Міністерство праці та соціального захисту • Головне управління умов праці та охорони здоров'я • Інспекція праці	• < 20 працівників – представник з охорони праці обирається в неформальний спосіб; • ≥ 20 працівників – представник з охорони праці обирається у формальний спосіб; • > 50 працівників – створюється спільний комітет з охорони праці
Німеччина	Закон про безпеку та гігієну праці [82]	• Федеральне міністерство праці та соціального забезпечення	• ≥ 5 працівників – встановлюється внутрішня рада працівників; • ≥ 20 працівників – створюється спільна комісія з охорони праці.
Ірландія	Закон про безпеку, здоров'я та добробут на роботі №10/2005 [83], частина 4	• Орган охорони здоров'я та безпеки	• ≥ 20 працівників – може (опціонально) створюватися спільний комітет з охорони праці

Порівнюючи законодавчі вимоги країн Європи щодо створення комісій, комітетів, служб з охорони праці або вибору окремих представників з таких

питань, можна зробити такі висновки. У більшості країн представники з охорони праці обираються при наявності 5-10 працівників. Найчастіше такі представники входять у спільну з роботодавцем раду, яка забезпечує інтереси обох сторін. Окремі комітети та комісії з питань безпеки праці найчастіше створюються при досягненні 20-50 працівників (залежно від країни). Кількість членів рад, комісій та комітетів регулюється в залежності від розміру підприємства. Окрім цього, у деяких країнах (наприклад, Ірландії та Чеській Республіці) функціонування комітетів не є обов'язковим згідно з чинним законодавством. Отже, українське законодавство та регулятивні органи є відповідними європейським у частині створення служб охорони праці на підприємствах, і за законом українські працівники ММСП є не менш захищеними.

4.2 Державна система моніторингу довкілля як складова частина національної інформаційної інфраструктури, сумісної з аналогічними системами інших країн

Статті 20 та 22 Закону України «Про охорону навколишнього природного середовища» [84] передбачають створення державної системи моніторингу довкілля та проведення спостережень за станом навколишнього природного середовища і рівнем його забруднення.

Основні принципи функціонування державної системи моніторингу довкілля визначені у постанові Кабінету Міністрів України від 30.03.1998 №391 «Про затвердження Положення про державну систему моніторингу довкілля» [85]. Згідно з цим положенням, «Державна система моніторингу довкілля (ДСМД) – це система спостережень, збирання, оброблення, передавання, збереження та аналізу інформації про стан довкілля, прогнозування його змін і розроблення науково-обґрунтованих рекомендацій

для прийняття рішень про запобігання негативним змінам стану довкілля та дотримання вимог екологічної безпеки».

Також визначається, що система моніторингу є відкритою інформаційною системою, складовою частиною національної інфраструктури, сумісної з аналогічними системами інших країн. Пріоритетами такої системи є збереження екосистем, відвернення кризових змін в екологічному стані довкілля та запобігання відповідним надзвичайним ситуаціям. На даний час, у ДСМД функції і задачі спостережень та інформаційного забезпечення виконують такі суб`єкти системи моніторингу [85]:

- Міністерство захисту довкілля та природних ресурсів України.
- Міністерство аграрної політики та продовольства України.
- Міністерство розвитку громад і територій України.
- Державна служба з надзвичайних ситуацій.
- Державна служба геології та надр України.
- Державне агентство України з управління зоною відчуження.
- Державне агентство лісових ресурсів України.
- Державне агентство водних ресурсів України.
- Державна служба України з питань геодезії, картографії та кадастру.
- Державне космічне агентство України.

Окрім цього, виконання таких функцій покладено на інші центральні органи виконавчої влади, які є суб'єктами державної системи моніторингу довкілля, а також на підприємства, установи та організації, діяльність яких призводить або може призвести до погіршення стану довкілля.

В загальному, система моніторингу спрямована на декілька основних цілей, серед яких підвищення рівня вивчення і знань про екологічний стан довкілля, зростання якості інформаційного обслуговування користувачів на всіх рівнях, покращення якості обґрунтування природоохоронних заходів та

ефективності їх здійснення, а також сприяння розвитку міжнародного співробітництва у галузі охорони довкілля, раціонального використання природних ресурсів та екологічної безпеки [85, пункт 6].

Для досягнення поставлених цілей, положення про ДСМД визначає такі завдання: систематичні спостереження за станом довкілля, аналіз стану навколишнього середовища, прогнозування змін довкілля, інформаційна підтримка прийняття рішень, забезпечення органів державної та місцевої влади, населення та партнерів інформацією про актуальний стан довкілля [85, пункт 7]. Також згідно з положенням складовими частинами державного моніторингу навколишнього середовища України є моніторинг атмосферного повітря, води, земель, біологічного різноманіття, лісів, відходів, геологічного середовища, фізичних факторів впливу. Нормативними актами, що регламентують моніторинг таких об’єктів є відповідні постанови Кабінету Міністрів України щодо порядків здійснення моніторингу повітря, вод, земель та ґрунтів.

Система моніторингу ґрунтується на використанні існуючих організаційних структур суб'єктів моніторингу і функціонує на основі єдиного нормативного, організаційного, методологічного і метрологічного забезпечення, об'єднання складових частин та уніфікованих компонентів цієї системи [86]. Зазначимо, що функціонування ДСМД здійснюється на трьох рівнях, які розподіляються за територіальним принципом, а саме: загальнодержавний, регіональний та локальний рівні. Відповідальні суб’єкти здійснюють моніторинг різного роду об’єктів, серед яких [85]:

- Ґрунти на природоохоронних територіях, а також ґрунти сільськогосподарського та лісового фондів.
- Види рослинного і тваринного світу, що перебувають під загрозою зникнення чи під особливою охороною.
- Вміст радіонуклідів в атмосферному повітрі, водах та ґрунтах.
- Наявність та серйозність повеней, паводків, снігових лавин, селів.

- Об’єкти зберігання та захоронення радіоактивних відходів.
- Сільськогосподарські рослини, тварини і продуктів з них, мисливська фауна та лісова рослинність.
- Якість вод водогосподарських систем міжгалузевого та сільськогосподарського водопостачання.
- Зрошувані та осушувані землі у сенсі глибини залягання та мінералізації ґрунтових вод, ступені засоленості та солонцюватості ґрунтів.
- Ґрунти і ландшафти щодо проявів ерозійних та інших екзогенних процесів та просторового забруднення земель об'єктами промислового і сільськогосподарського виробництва.
- Берегові лінії річок, морів, озер, водосховищ, лиманів, заток, гідротехнічних споруд.
- Стічні води міської каналізаційної мережі та очисні споруди, джерела скидання таких вод.
- Зелені насадження у містах і селищах міського типу.

Якщо розглянути моніторинг повітря, то Державною гідрометеорологічною службою здійснюється спостереження за забрудненням атмосферного повітря у містах України. Державна екологічна інспекція здійснює відбір проб на джерелах викидів, а санітарно-епідеміологічна служба координує моніторинг якості атмосферного повітря у житлових зонах. Контроль якості повітря також включає аналіз опадів та снігового покрову. Програма обов`язкового моніторингу якості атмосферного повітря охоплює сім забруднюючих речовин: пил, двоокис азоту, двоокис сірки, оксид вуглецю, формальдегід, свинець та бензапірен.

Спостереження за водами суші на 151 об’єкті проводить Державна гідрометеорологічна служба, що включає в себе оцінку хімічного складу вод, біогенних параметрів, наявність зважених часток та органічних речовин, основних забруднюючих речовин, важких металів та пестицидів. Контроль за водами суші також здійснюють Державна екологічна інспекція, Державний

комітет по водному господарству, Санітарно-епідеміологічна служба та Державна геологічна служба. Дослідження включають моніторинг річок, водосховищ, каналів тощо, контроль хімічних, радіаційних та фізичних показників, а також придатності води до споживання. За схожими параметрами відбувається й моніторинг прибережних вод.

До моніторингу ґрунтів входить вимірювання забруднення ґрунтів пестицидами, агровідходами, токсинами та важкими металами на сільськогосподарських землях та промислових майданчиках. Також досліджується забруднення ґрунту у місцях захоронення відходів. Контроль здійснюється державною гідрометеорологічною службою, Міністерством захисту довкілля та Міністерством аграрної політики.

Моніторинг радіаційного випромінювання включає в себе спостереження за радіоактивним забрудненням атмосфери, поверхневих вод та ґрунтів поблизу атомних електростанцій та у зоні відчуження [86].

Згідно з положенням [85] суб’єкти системи моніторингу інформаційно підтримують рішення в галузі охорони довкілля, безкоштовно обмінюються результатами спостережень на об’єктах та колективно використовують інформаційні ресурси, надаючи всім зацікавленим сторонам відповідні дані.

4.2.1 Системи моніторингу довкілля у країнах Європейського Союзу

За останні 30 років Європейський Союз здійснив значний спектр екологічних заходів, спрямованих на покращення якості довкілля для європейських громадян. Одним із органів влади, який забезпечує дотримання прийнятих стратегій та регуляцій є Європейська Комісія згідно зі стандартами законодавства, підписаного країнами-членами Євросоюзу. Кожна країна приймає та впроваджує екологічне законодавство Співтовариства, а також займається моніторингом поточного стану довкілля власної країни та обміну інформацією з партнерами.

ЄС має одні з найвищих світових екологічних стандартів, розроблених протягом десятиліть і завдяки цьому екологічна політика та законодавство країн Співтовариства захищають природні середовища існування, підтримують повітря та воду в чистоті, забезпечують належну утилізацію відходів, покращують знання про токсичні хімікати та допомагають підприємствам рухатися до сталої економіки [87].

Існують окремі юридичні аспекти проведення моніторингу та звітування про стан довкілля у ЄС. Моніторинг довкілля можна охарактеризувати як програму повторюваних, систематичних досліджень, що виявляє стан навколишнього середовища. Конкретні аспекти навколишнього середовища, що підлягають дослідженню, визначаються екологічними цілями та екологічним законодавством. Метою моніторингу довкілля є оцінка прогресу, досягнутого для досягнення поставлених екологічних цілей, а також виявлення нових екологічних проблем [87].

Існує вичерпний перелік нормативних документів, що встановлюють принципи і правила моніторингу, а також відповідність загальноприйнятому законодавству щодо регулювання якості повітря, хімічного та промислового забруднення, захисту природи і біорізноманіття, шуму, викидів, водного та морського господарства тощо [88].

4.3 Висновки до четвертого розділу

Четвертий розділ кваліфікаційної роботи присвячений питанням охорони праці та безпеки у надзвичайних ситуаціях. У частині охорони праці розглянуто нормативні акти, що регламентують наявність служби та комісії з охорони праці на мікро, малих та середніх підприємствах в Україні та країнах Євросоюзу. Проаналізовано чинне законодавство країн на предмет наявності таких служб, мінімальної кількості працівників для їх утворення та представництва співробітників у спільних радах. В Україні служба з охорони

праці створюється на підприємствах з кількістю працівників більш, ніж 50 осіб. На менших підприємствах – такі функції виконуються за сумісництвом або ж з залученням сторонніх спеціалістів. Щодо Європейського Союзу – у більшості країн окрема служба створюється за наявності 20-50 співробітників, до того – один із працівників може бути представником інтересів колективу.

У частині безпеки в надзвичайних ситуаціях було розглянуто питання державної служби моніторингу довкілля, а саме її основних цілей та задач, суб'єктів та об'єктів моніторингу. В Україні державна служба моніторингу довкілля поєднує різного роду агентства, міністерства та відомства, пов'язані із природнім середовищем, і проводить фіксацію стану повітря, вод, суші, ґрунту, радіаційного фону, природного різноманіття на багатьох об'єктах. Додатково коротко розглянуто систему моніторингу довкілля у країнах Європи як поєднану інформаційну інфраструктуру з широким спектром регулятивних актів для всіх країн-членів ЄС.

ВИСНОВКИ

У результаті виконання кваліфікаційної роботи було досягнуто поставленої мети, а саме отримано цінні ідеї щодо впливу компонентів HIT-індексу на цифрову зрілість мікро, малих та середніх підприємств за допомогою застосування шести алгоритмів кластеризації даних. За результатами проведеного дослідження було отримано такі результати:

- Розглянуто концепції аналітичної роботи з даними та сфери їх застосування. Описано основні напрямки аналізу та типи задач інтелектуального видобутку даних.
- Оглянуто та проаналізовано наукові роботи щодо аналізу даних та застосування методів, підходів і технік для вирішення бізнес-задач.
- Висвітлено поняття цифровізації економіки та важливості трансформації бізнес-структур. Описано метод визначення рівня цифрової зрілості підприємства – HIT-індекс.
- Розглянуто задачу кластеризації інформації. Докладно досліджено, описано та порівняно популярні методи кластеризації даних, поняття міри відстані та метрики якості.
- Розроблено функціональну програму для проведення кластеризації даних на основі обчислених значень індикативних компонент індексу цифрової зрілості HIT. Описано функціонал та програму реалізацію.
- Проведено кластеризацію опитаних респондентів та проаналізовано отримані результати на предмет цінних ідей.

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто питання регулятивних документів та вимог щодо утворення служб та комісії щодо охорони праці на малих та середніх підприємствах в Україні та деяких країнах Європи. Також висвітлено поняття державної служби моніторингу довкілля, її основних цілей, завдань, суб’єктів та об’єктів моніторингу в Україні та Європейському Союзі.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. СЛОВНИК УКРАЇНСЬКОЇ МОВИ ONLINE. ТОМИ 1-11 [Електронний ресурс] – Режим доступу до ресурсу: https://services.ulif.org.ua/expl/Entry/index.

2. Sarangam A. Data Mining vs Data Analysis – An Easy Guide In Just 3 Points [Електронний ресурс] / Ajay Sarangam // Jigsaw. – 2021. – Режим доступу до ресурсу: https://www.jigsawacademy.com/blogs/data-science/data-mining-vs-data-analysis/#Difference-between-data-mining-and-data-analysis.

3. Data mining [Електронний ресурс] // Вікіпедія - вільна енциклопедія – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Data_mining.

4. Business analysis [Електронний ресурс] // Вікіпедія - вільна енциклопедія – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/ Business_analysis

5. Business intelligence [Електронний ресурс] // Вікіпедія - вільна енциклопедія – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/ Business_intelligence.

6. Business Intelligence vs. Business Analytics: What's The Difference? [Електронний ресурс] // Tableau. – 2020. – Режим доступу до ресурсу: https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics.

8. Business Analytics vs. Data Analytics: Which is Better for Your Business? [Електронний ресурс] // Talend – Режим доступу до ресурсу: https://www.talend.com/resources/business-analytics-vs-data-analytics/.

9. What is Data Analysis - Data Analytics Trends and Objectives - Methods of Data Analysis [Електронний ресурс] // The Scientific World. – 2019. – Режим доступу до ресурсу: https://www.scientificworldinfo.com/2019/10/what-is-data-analysis-objectives-of-data-analysis-and-trends.html

10. Data Analysis: What, How, and Why to Do Data Analysis for Your Organization [Електронний ресурс] // Import.io. – 2019. – Режим доступу до ресурсу: https://www.import.io/post/business-data-analysis-what-how-why/.

11. Calzon B. Your Modern Business Guide To Data Analysis Methods And Techniques [Електронний ресурс] / Bernardita Calzon // DataPine. – 2021. – Режим доступу до ресурсу: https://www.datapine.com/blog/data-analysis-methods-and-techniques/.

12. Stevens E. The 7 Most Useful Data Analysis Methods and Techniques [Електронний ресурс] / Emily Stevens // CareerFoundry. – 2021. – Режим доступу до ресурсу: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techni ques/.

13. 16 Data Mining Techniques: The Complete List [Електронний ресурс] // Talend – Режим доступу до ресурсу: https://www.talend.com/resources/data-mining-techniques/.

14. Asrani V. Top 11 Data Mining Techniques of 2021 [Електронний ресурс] / Vivek Asrani // Just Total Tech. – 2021. – Режим доступу до ресурсу: https://justtotaltech.com/data-mining-techniques/.

15. Rai A. An Overview of Association Rule Mining & its Applications [Електронний ресурс] / Abhinav Rai // upGrad blog. – 2019. – Режим доступу до ресурсу: https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/.

16. Remanan S. Association Rule Mining [Електронний ресурс] / Surya Remanan // Towards. Data Science. – 2018. – Режим доступу до ресурсу: https://towardsdatascience.com/association-rule-mining-be4122fc1793.

17. Chen, Hsiu-chin & Chiang, Roger & Storey, Veda. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly. 36. 1165-1188. 10.2307/41703503.

18. Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. (2019). “Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study”. Cornell Univercity. arXiv:1911.00568

19. Joel Ashirwadam. (2014). Communication Research Methods Methods of Data Analysis.

20. Kabir, Syed Muhammad. (2016). METHODS OF DATA COLLECTION.

21. Young, Ryan & Wahlberg, Luke & Davis, Elaina & Abhari, Kaveh. (2020). Towards a Theory of Digital Entrepreneurship Mindset: The Role of Digital Learning Aptitude and Digital Literacy.

22. Bican, Peter M., and Alexander Brem. 2020. «Digital Business Model, Digital Transformation, Digital Entrepreneurship: Is There A Sustainable “Digital”?» *Sustainability* 12, no. 13: 5239. 10.3390/su12135239

23. Hemmert, Martin & Cross, Adam & Cheng, Ying & Kim, Jae-Jin & Kohlbacher, Florian & Kotosaka, Masahiro & Waldenberger, Franz & Zheng, Leven J.. (2019). The distinctiveness and diversity of entrepreneurial ecosystems in China, Japan, and South Korea: an exploratory analysis. Asian Business & Management. 18. 10.1057/s41291-019-00070-6.

24. Marcello M. Mariani & Samuel Fosso Wamba. (2020). Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies. Journal of Business Research. Volume 121, Pages 338-352.

25. Gavurová, Beáta & Belas, Jaroslav & Kotaskova, Anna & Cepel, Martin. (2018). Management of education concepts in the field of entrepreneurship of university students in the Czech Republic. Polish Journal of Management Studies. 17. 52-62. 10.17512/pjms.2018.17.2.05.

26. Sestino, Andrea. (2019). Business Development, Marketing Automation and Predictive Analysis: An Integration Perspective - An Overview Towards New Opportunities for Studying Consumer Behavior and Business Integration.. SSRN Electronic Journal. 10.2139/ssrn.3316759.

27. Zhao, Jia & Xue, Fei & Khan, Shahnawaz & Khatib, Saleh. (2021). Consumer behaviour analysis for business development. Aggression and Violent Behavior. 101591. 10.1016/j.avb.2021.101591.

28. Pan, Yang & Russell, Gary. (2018). CONVENIENCE STORE ANALYTICS: ANALYZING HABITUAL AND SITUATIONAL SHOPPING BEHAVIOR USING CONSUMER BASKET DATA. 10.13140/RG.2.2.16191.41120.

29. Manpreet Kaur & Shivani Kang. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. Procedia Computer Science. Volume 85. Pages 78-85. ISSN 1877-0509

30. Kronberger, Gabriel & Affenzeller, Michael. (2011). Market Basket Analysis of Retail Data: Supervised Learning Approach. 464-471. 10.1007/978-3-642-27549-4_59.

31. Kurniawan, Fachrul & Umayah, Binti & Hammad, Jehad & Nugroho, Supeno & Hariadi, Mohamad. (2017). Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data. Knowledge Engineering and Data Science. 1. 20. 10.17977/um018v1i12018p20-25.

32. Golic, Merisa & Zunic, Emir & Donko, Dzenana. (2019). Outlier detection in distribution companies business using real data set. IEEE EUROCON 2019 - 18th International Conference on Smart Technologies 1-5. Doi: 10.1109 /EUROCON.2019.8861526.

33. Chang, Mona & Yuan, Yuan & Yue, Qi & Mincheol, Han. (2020). A CNN Image Classification Analysis for 'Clean-Coast Detector' as Tourism Service Distribution. Seed Science and Technology. 18. 15-26. 10.15722/jds.17.12.20201.15.

34. Agapie, Alexandru, Cristian Vizitiu, Silvia E. Cristache, Marian Năstase, Liliana Crăciun, and Anca G. Molănescu (2018). «Analysis of Corporate Entrepreneurship in Public R&D Institutions» *Sustainability* 10, no. 7: 2297. https://doi.org/10.3390/su10072297

35. Wood, E.H. (2006), «The internal predictors of business performance in small firms: A logistic regression analysis», Journal of Small Business and Enterprise Development, Vol. 13 No. 3, pp. 441-453. https://doi.org/10.1108/14626000610680299

36. Urban, Marcia & Klemm, Martin & Plötner, Kay & Hornung, Mirko. (2018). Airline categorization by applying the business model canvas and clustering algorithms. Journal of Air Transport Management. 71. 10.1016/j.jairtraman.2018.04.005.

37. Sammour, George & Qabbaah, Hamzah & Vanhoof, Koen. (2019). DECISION TREE ANALYSIS TO IMPROVE E-MAIL MARKETING CAMPAIGNS. 26. 3-36.

38. Михайло Федоров: Цифровізація економіки дозволить досягти мінімум 4% додаткового зростання ВВП на рік [Електронний ресурс] // Прес-офіс Міністерства цифрової трансформації. – 2021. – Режим доступу до ресурсу: https://thedigital.gov.ua/news/mihajlo-fedorov-cifrovizaciya-ekonomiki-dozvolit-dosyagti-minimum-4-dodatkovogo-zrostannya-vvp-na-rik

39. I. Strutynska, L. Dmytrotsa, H. Kozbur, O. Hlado, P. Dudkin and O. Dudkina, «Development of Digital Platform to Identify and Monitor the Digital Business Transformation Index,» *2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT)*, 2020, pp. 171-175, doi: 10.1109/CSIT49958.2020.9322016.

40. Iryna Strutynska, Lesia Dmytrotsa, Halyna Kozbur, Liliya Melnyk, Hlado Olha. «Developing Practical Recommendations for Increasing the Level of Digital Business Transformation Index.» In ICTERI Workshops, pp. 351-362. 2020.

41. I. Strutynska, L. Dmytrotsa, H. Kozbur, O. Hlado and O. Sorokivska, «Working-Out of Recommendation System to Increase the Digital Maturity Level of Enterprises», 2020 IEEE International Conference on Problems of

Infocommunications. Science and Technology (PIC S&T), 2020, pp. 147-151, doi: 10.1109/PICST51311.2020.9467978

42. Струтинська І. В. «Цифрова трансформація як імператив інноваційного розвитку бізнес-структур» : дис. докт. ек. наук : 08.00.04 — Екон / Струтинська Ірина Володимирівна, 2020. – 487 с.

43. Iryna Strutynska, Lesia Dmytrotsa, Halyna Kozbur, Liliya Melnyk., 2020. System-Integrated Methodological Approach Development to Calculating the Digital Transformation Index of Businesses. In ICTERI (pp. 373-379).

44. Prasad S. Different Types of Clustering Methods and Applications [Електронний ресурс] / Sunit Prasad // Analytix Labs. – 2020. – Режим доступу до ресурсу: https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/.

45. Sharma R. What is Clustering and Different Types of Clustering Methods [Електронний ресурс] / Rohit Sharma // upGrad. – 2020. – Режим доступу до ресурсу: https://www.upgrad.com/blog/clustering-and-types-of-clustering-methods/.

46. Different Types of Clustering Algorithm [Електронний ресурс] // GeeksForGeeks. – 2021. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/different-types-clustering-algorithm/.

47. Cluster analysis [Електронний ресурс] // Вікіпедія - вільна енциклопедія. – 2021. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/ Cluster_analysis.

48. Scikit-Learn. Machine learning in Python [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/.

49. ML | Hierarchical clustering (Agglomerative and Divisive clustering) [Електронний ресурс] // GeeksForGeeks. – 2021. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/.

50. Seif G. The 5 Clustering Algorithms Data Scientists Need to Know [Електронний ресурс] / George Seig // Towards Data Science. – 2018. – Режим

доступу до ресурсу: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

51. Srivignesh R. Overview of Clustering Algorithms [Електронний ресурс] / Rajan Srivignesh // Towards Data Science. – 2021. – Режим доступу до ресурсу: https://towardsdatascience.com/overview-of-clustering-algorithms-27e979e3724d.

52. Sinclair C. Clustering Using OPTICS [Електронний ресурс] / Colin Sinclair // Towards Data Science. – 2019. – Режим доступу до ресурсу: https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7.

53. Affinity Propagation in ML [Електронний ресурс] // GeeksForGeeks. – 2019. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/affinity-propagation-in-ml-to-find-the-number-of-clusters/.

54. Malkin C. Affinity Propagation Algorithm Explained [Електронний ресурс] / Cory Malkin // Towards Data Science. – 2019. – Режим доступу до ресурсу: https://towardsdatascience.com/unsupervised-machine-learning-affinity-propagation-algorithm-explained-d1fef85f22c8.

55. Groothenforst M. 9 Distance Measures in Data Science [Електронний ресурс] / Maarten Grootendorst // Towards Data Science. – 2021. – Режим доступу до ресурсу: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa.

56. Subramanian N. Distance/Similarity Measures in Machine Learning [Електронний ресурс] / Niranjan Subramanian // AiAspirant. – 2019. – Режим доступу до ресурсу: https://aiaspirant.com/distance-similarity-measures-in-machine-learning.

57. Kumar A. Different Types of Distance Measures in Machine Learning [Електронний ресурс] / Ajjitesh Kumar // VitaFlux. – 2020. – Режим доступу до ресурсу: https://vitalflux.com/different-types-of-distance-measures-in-machine-learning.

58. Rizk N. Building Skills for Data Science [Електронний ресурс] / Nouhad Rizk // University of Houston. – Режим доступу до ресурсу: https://uhlibraries.pressbooks.pub/buildingskillsfordatascience/chapter/cluster-validity/.

59. Pathak M. Quick Guide to Evaluation Metrics for Supervised and Unsupervised Machine Learning [Електронний ресурс] / Manish Pathak // Analytics Vidhya. – 2020. – Режим доступу до ресурсу: https://www.analyticsvidhya. com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/.

60. Assessment Metrics for Clustering Algorithms. [Електронний ресурс] // Medium. – 2018. – Режим доступу до ресурсу: https://medium.com/ @ODSC/assessment-metrics-for-clustering-algorithms-4a902e00d92d.

61. Zuccarelli E. Performance Metrics in Machine Learning – Part 3: Clustering [Електронний ресурс] / Eugenio Zuccsrelli // Towards Data Science. – 2021. – Режим доступу до ресурсу: https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6.

62. Calinski-Harabasz Index – Cluster Validity indices. Set 3 [Електронний ресурс] // GeeksForGeeks. – 2022. – Режим доступу до ресурсу: https:// www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/.

63. Elbow Method for optimal value of k in KMeans [Електронний ресурс] // GeeksForGeeks. – 2021. – Режим доступу до ресурсу: https:// www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.

64. Elbow Method (clustering) [Електронний ресурс] // Вікіпедія – вільна енциклопедія. – 2022. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Elbow_method_(clustering)

65. I. Strutynska, H. Kozbur, L. Dmytrotsa, I. Bodnarchuk and O. Hlado, «Small and Medium Business Structures Clustering Method Based on Their Digital Maturity», *2019* IEEE *International Scientific-Practical Conference*

Problems of Infocommunications, Science and Technology (PIC S&T), 2019, pp. 278-282, doi: 10.1109/PICST47496.2019.9061464.

66. Iryna Strutynska, Halyna Kozbur, Lesia Dmytrotsa, Olha Hlado, Liliya Melnyk «Comparative Analysis of Two Approaches to the Clustering of Respondents (based on Survey Results).» In CMiGIN, pp. 434-446. 2019.

67. ЗАКОН УКРАЇНИ «Про охорону праці» №2694-12 [Електронний ресурс] // Верховна Рада України. Законодавство України. – 1992. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/2694-12#Text.

68. Господарський кодекс України №436-15 [Електронний ресурс] // Верховна Рада України. Законодавство України. – 2003. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/436-15#Text

69. ЗАКОН УКРАЇНИ «Про бухгалтерський облік та фінансову звітність в Україні» №996-14 [Електронний ресурс] // Верховна Рада України. Законодавство України. – 1999. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/ 996-14#Text.

70. Статистична інформація. Економічна статистика. Діяльність підприємств. Кількість підприємств за видами економічної діяльності з розподілом на великі, середні, малі та мікропідприємства. [Електронний ресурс] // Державна служба статистики України. — 2020. — Режим доступу до ресурсу: http://www.ukrstat.gov.ua/.

71. European Agency for Safety and Health at Work [Електронний ресурс] // Вікіпедія — вільна енциклопедія. – 2021. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/European_Agency_for_Safety_and_Health_at_Work.

72. Improving Safety and Health in Micro-, Small and Medium-Sized Enterprises: An overview of initiatives and delivery mechanisms [Електронний ресурс] // International Labor Organization. – 2020. – Режим доступу до ресурсу: https://www.ilo.org/wcmsp5/groups/public/---ed_dialogue/---lab_admin/documents/publication/wcms_740304.pdf.

73. National Focal Points [Електронний ресурс] // EU-OSHA – Режим доступу до ресурсу: https://osha.europa.eu/en/about-eu-osha/national-focal-points/focal-points-index.

74. Fulton L. Health and safety representation in Europe. Labour Research Department and ETUI (online publication) [Електронний ресурс] / Fulton // Workers' Interest Group of the Advisory Committee for Safety and Health at Work (of the EU Commission). – 2018. – Режим доступу до ресурсу: https://www.worker-participation.eu/National-Industrial-Relations/Countries.

75. Закон про захист працівників №450/1994 [Електронний ресурс]. – 1994. – Режим доступу до ресурсу: https://www.ris.bka.gv.at/NormDokument.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008910&FassungVom=2022-04-04&Artikel=&Paragraf=0&Anlage=&Uebergangsrecht=.

76. Закон про працю № 226/2006 [Електронний ресурс]. – 2006. – Режим доступу до ресурсу: https://www.mpsv.cz/documents/625317/625915/Labour+Code.pdf/b1f02b8f-ece9-c898-cd4b-4d4f448538c3

77. Закон про охорону праці [Електронний ресурс]. – 1999. – Режим доступу до ресурсу: https://www.riigiteataja.ee/en/eli/ee/Riigikogu/act/505052017007

78. Закон про нагляд за безпекою та гігієною праці та співробітництво з охорони праці на робочому місці №20.1.2006/44 [Електронний ресурс]. – 2006. – Режим доступу до ресурсу: https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/73020/97077/F1539695710/ FIN73020.pdf

79. Закон про безпеку та здоров’я працівників №IX-1672 [Електронний ресурс]. – 2003. – Режим доступу до ресурсу: https://e-seimas.lrs.lt/portal/ legalAct/lt/TAD/TAIS.215253

80. Закон про безпеку та гігієну праці №319 [Електронний ресурс]. – 2006. – Режим доступу до ресурсу: https://legislatie.just.ro/Public/DetaliiDocument/73772

81. Кодекс законів про охорону здоров’я та безпеку людини №3850 [Електронний ресурс]. – 2010. – Режим доступу до ресурсу: https://www.e-forosimv.gr/details.asp?ID=5145

82. Закон про вжиття заходів з охорони праці для заохочення вдосконалення безпеки та охорони здоров'я працівників на виробництві [Електронний ресурс]. – 1996. – Режим доступу до ресурсу: https://www.gesetze -im-internet.de/englisch_arbschg/englisch_arbschg.html

83. Закон про безпеку, здоров'я та добробут на роботі №10/2005 [Електронний ресурс]. – 2005. – Режим доступу до ресурсу: https://www.irish statutebook.ie/eli/2005/act/10/enacted/en/print

84. ЗАКОН УКРАЇНИ «Про охорону навколишнього природного середовища» №1264-XII [Електронний ресурс] // Верховна Рада України. Законодавство України. – 1991. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/ show/1264-12.

85. Постанова Кабінету Міністрів України «Про затвердження Положення про державну систему моніторингу довкілля» №391-98-п [Електронний ресурс] // Верховна Рада України. Законодавство України. – 1998. – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/391-98.

86. Екологічний моніторинг довкілля [Електронний ресурс] // Міністерство захисту довкілля та природних ресурсів України. – 2017. – Режим доступу до ресурсу: https://mepr.gov.ua/content/ekologichniy-monitoring-dovkillya.html.

87. Implementation of Community environmental legislation [Електронний ресурс] // European Commission. – Режим доступу до ресурсу: https://ec.europa.eu/ environment/legal/implementation_en.htm

88. Monitoring and reporting of environment legislation [Електронний ресурс] // European Commission. – Режим доступу до ресурсу: https://ec.europa.eu/ environment/legal/reporting/products_en.htm

ДОДАТКИ

Додаток А

2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T 2019)

Kyiv, Ukraine
8 – 11 October 2019

IEEE

IEEE Catalog Number: CFP19PIA-POD
ISBN: 978-1-7281-4185-5

Small and Medium Business Structures Clustering Method Based on Their Digital Maturity

Iryna Strutynska
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
ringtons999@gmail.com

Halyna Kozbur
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
kozbur.galina@gmail.com

Lesia Dmytrotsa
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
dmytrotsa.lesya@gmail.com

Ihor Bodnarchuk
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
bodnarchuk.io@gmail.com

Olha Hlado
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
gladyo.olga@gmail.com

Abstract—Business adaptation and transformation, driven by the challenges of the digital world, is an important point in addressing the challenges of the global market today. To evaluate the effectiveness of digital business transformation, an in-depth analysis of statistical information collected over a given period of time for a particular group of business entities is required. The proposed scientific study concerns the digital transformation of business structures registered in the Ternopil region of Ukraine. The method for researching the level of digital transformation of small and medium-sized enterprises is developed by conducting a survey and further dividing respondents into clusters. Some indicators are offered, the method of preparation and processing of data is described, the analysis of results is made.

Keywords—digital economy; digital technologies; data clustering; survey of respondents; statistic techniques; business-structures of SMEs

I. Introduction

Information technologies allow any company to change its own business model so that to be differentiated from the whole global market. Taking into account some gaps in statistic provision of monitoring of digital economy development and information society building it makes sense to intensify the efforts of the main stakeholders to fulfill the Plan of measures on implementation of the digital economy and society development Conception of Ukraine for the period from 2018 to 2020 [1]. One should pay special attention to the development of indicators system for business digital transformation to provide regular monitoring of digital development and to conduct regular statistic observations. The process involves both the modification of existing statistic forms on Internet use by population and information-communication technologies (ICT) at enterprises and the development of new indicators, methodological and organization basement of new data collecting and analysis.

Small and medium-sized enterprises (SMEs) are of primary concern. Nevertheless, these business-structures don't often have all necessary information on using any innovative digital technologies improving the efficiency of business modeling and conducting [2]. One needs to develop some road maps of digital transformation so that to adapt to the market challenges and to build competitive business models of such enterprises.

The scientific research deals with a method of data collecting which characterize the situation in digital transformation of business structures by surveying the representatives of small and medium-sized enterprises registered in Ternopil region. The job of surveys' findings clustering is a widely spread one due to the rapidly increasing popularity of their collecting by this method and their further processing. Although, the following problems should be solved in the best way: diversity of data given in a questionnaire by a certain respondent; accidental errors while inputting; some empty values occurrence and complexity of answers unification while giving their own answers by respondents; metric selection for the clusters members which are both quantitative and categorical ones with possibility or impossibility of their arrangement; quality factor selection of clustering results and their correct analysis.

The problem of respondents clustering according to the results of answers on the questionnaires has been considered in [3]. The aim of the study was to advise to the respondents regarding their results of their political preferences and to compare the clustering method with other conventional techniques of writing such recommendations. The document also considered other popular methods of processing surveys of this type, such as

978-1-7281-4184-8/19/$31.00 ©2019 IEEE

Додаток Б

Solomiia Fedushko
Sergiy Gnatyuk
Andriy Peleshchyshyn
Zhengbing Hu
Roman Odarchenko
Igor Korobiichuk
(Eds.)

CMiGIN-2019

International Workshop on Conflict Management in Global Information Networks
CMiGIN 2019

Lviv, Ukraine

November 29, 2019

Session 4

Session 5

Comparative Analysis of Two Approaches to the Clustering of Respondents (based on Survey Results)

Iryna Strutynska, Halyna Kozbur, Lesia Dmytrotsa, Olha Hlado, Liliya Melnyk

Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine
{ringtons999, dmytrotsa.lesya, kozbur.galina}@ gmail.com

Abstract. This paper proposes an algorithm for solving the survey respondents' clustering problem, including the steps of collecting, preparing data, summarizing key results, and developing future goals. The research consists of two approaches to clustering: iterative and hierarchical in order to produce consistent and comprehensible results. The iterative method is implemented in MS Excel using the Data Mining add-in, hierarchical one is used with the help of writing code and using Python libraries. Hard clusters with sufficient degree of similarity within the cluster and differences from others were distinguished, the main characteristics of the obtained clusters were described as well. It has been experimentally established that the method of agglomerative hierarchical clustering is more effective for solving the problem of clustering of mixed-type data obtained from the survey of respondents.

Keywords: digital maturity, clustering methods, mixed-type data

1 Introduction

The fastest and the most convenient way to get any information you need today is to directly interview your target audience on a specific topic. With the development of information technology, such questionnaires are increasingly shifting from personal or telephone communication to online questionnaires. This allows you to reach a larger audience in a shorter time span and with fewer human resources. The positive aspects of such surveys are: convenience of expression; partial or complete anonymity of results; the ability to complete a survey in any convenient for the respondent way; no need to communicate with the employees of the survey organization, etc. Online surveys are a particularly effective way of retrieving information if your target audience is the users of the web. Data collection is only part of the complex task of getting the information you need. Further processing and analysis of data with conclusions and recommendations make the data cycle complete. Segmentation or clustering is one of the most important and interesting tasks of data analysis. This paper offers an algorithm for solving the problem of clustering respondents by online survey, including the steps of collecting, preparing data, summarizing key findings, and developing future goals.

Додаток В

Oleksandr Sokolov
Grygoriy Zholtkevych
Vitaliy Yakovyna
Yulia Tarasich
Vyacheslav Kharchenko
Vitaliy Kobets
Olexandr Burov
Serhiy Semerikov
Hennadiy Kravtsov
(Eds.)

ICT in Education, Research and Industrial Applications

Proceedings of the 16th International Conference, ICTERI 2020. Volume II: Workshops

Kharkiv, Ukraine
October 2020

Part III: 8th International Workshop Information Technology in Economic Research (ITER 2020)

Developing Practical Recommendations for Increasing the Level of Digital Business Transformation Index

Iryna Strutynska[1], Lesia Dmytrotsa[1], Halyna Kozbur[1], Liliya Melnyk[1], Hlado Olha[1]

[1]Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine
{ringtons999, dmytrotsa.lesya, kozbur.galina, liliana.mel0512, gladyo.olga}@gmail.com

Abstract. *Research goals and objectives:* to develop practical recommendations for increasing the level of digital business transformation index based on clustering of small and medium enterprises.
Subject of research: development and use of practical recommendations (digital roadmaps of digital transformation) for entrepreneurs.
Research methods used: survey of entrepreneurs, analytical methods for determining the Index, statistical methods of data processing, expert analysis of respondents' answers, cluster analysis of business structures, business analytics.
Results of the research. The list of multilevel recommendations for increasing the Digital Transformation Index were formed as well as the calculation method of the Index were described. Also was displayed gradation of recommendations; the results of clustering of business structures by the level of digital maturity were demonstrated; specific recommendations for raising the HIT for the enterprises of each of the clusters were formed. Such a methodology should take into account the current state of affairs in Ukraine, reflect an in-depth analysis of the level of digital transformation of business structures, while being flexible in order to respond promptly to new phenomena and the emergence of new digital technologies.

Keywords: Digital Transformation, Clustering, Index of Digital Transformation of Business Structures, Digital Transformation Roadmaps.

1 Introduction

At present, the transition of the industrial economy and the information society to the concepts and requirements of the "digital" economy is actively taking place. Such radical transformations require a new approach to understanding the nature and consequences of these processes, as well as the ability to adapt digital technologies to the contemporary demands of society and business. The rapid adaptation and transformation of business structures in the digital sense is one of the key tasks for raising the competitiveness of the domestic economy as a whole and integrating it with the leading global economic system.

The process of digital transformation of a business structure involves the transformation of its business strategy, models, operations, goals, marketing approaches, etc.

Додаток Г

2020 IEEE International Conference on Problems of Infocommunications, Science and Technology (PIC S&T 2020)

Kharkiv, Ukraine
6 – 9 October 2020

IEEE

IEEE Catalog Number: CFP20PIA-POD
ISBN: 978-1-7281-9178-2

Working-out of Recommendation System to Increase the Digital Maturity Level of Enterprises

Iryna Strutynska, Lesia Dmytrotsa, Halyna Kozbur, Olha Hlado, Olena Sorokivska
Computer Science Department
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
ringtons999@gmail.com, dmytrotsa.lesya@gmail.com, kozbur.galina@gmail.com, gladyo.olga@gmail.com, soroka220996@gmail.com

***Abstract*—The current world economy is such that its development is inevitably associated with introducing the latest digital technologies. It includes changing the business strategy and organizing its business processes, creating new goals, and means to achieve them. Undoubtedly, the introduction of digital technologies in business operations makes it possible to significantly increase its competitiveness in the domestic market and expand its activities' geographical scope. Suppose the business enterprise's management is ready for the digital transformation of business processes. In that case, it must find answers to the questions: what are the digital technologies among the many existing on the market that should introduce, which digital tools will be most effective in achieving new business goals. Therefore, one of the important tasks to accelerate the domestic business's digital transformation is to create road maps for business digitalization or development. These maps would contain both clear tools for change and sources of information from which enterprises can get the guidance, help, and further teaching opportunities. This paper proposes a system of practical recommendations for small and medium-sized business enterprises' digital maturity gain. The developed recommendations concern using specific digital tools (like social networks, site optimization, specialized applications in business processes, etc.), improving employees' digital maturity, and creating the vector of development for companies. The authors' Index is proposed as numerically measure the degree of digitalization of the business structure**

Keywords—digital transformation; digital maturity business index; automation of calculating Index; roadmaps for business digitalization; the algorithm for assigning recommendations. Keywords: digital transformation; digital maturity business index; automation of calculating Index; roadmaps for business digitalization; the algorithm for assigning requests.

I. Introduction

Around the world, the share of traditional economy decreases, and digital - is increasing, providing powerful benefits for countries and businesses. As demonstrated in [1], appropriate transformation processes provide changing the business strategy, the emergence of new goals and means of achieving them, adapting the company's vision and mission to the global trends in information technology. But in Ukraine, such processes are proceeding rather slowly. According to [2], one of the problems is the lack of awareness of entrepreneurs about the possibilities and ways of using digital technologies in their processes, especially among representatives of small and medium-sized enterprises (SMEs). The lack of available services, platforms, applications, or portals slows down domestic business innovation speed. This, in turn, does not allow the business to integrate with global trends, which makes access to international business complicated and impedes work in the international economic arena.

That is why research into the lack of necessary information and developing a roadmap with recommendations are important steps in accelerating Ukraine's transition to an information society and a "digital" economy. Given the differences in domestic entrepreneurs' areas of activity, it is important to take an individual approach to create a list of recommendations, specify them, and provide examples of tools for solving a particular problem.

The problem of the development of digital economy and transformational processes taking place in society under the influence of digitization has received a lot of attention among both foreign authors [4] - [7] and Ukrainian researchers [1] - [3], [8] - [10], etc. Despite numerous scientific studies on the development of information and communication technologies and the digital economy, we believe that the impact of digital technologies on business transformation is under-researched.

That is why this study aimed to develop a list of specific recommendations for increasing the level of digital maturity of SMEs and testing it on the example of an existing survey of 34 entrepreneurs of the Ternopil region and using the Digital Maturity Index developed in [3].

II. Methodology And Discussion

A. Digital Maturity Business Index

As a Comprehensive Assessment of SMEs' Digital Usage offered in [3], the Digital Maturity Business Index (HIT) was introduced. This criterion provides information on various aspects of digital development; to investigate their impact on the efficiency of business processes and the enterprise; analyze trends in SMEs' digitalization, and much more.

For the direct calculation of the Index, the last three indicators are used, namely [3]:

- hands of the enterprise digital infrastructure, which describe the level of its provision with the necessary equipment (personal computers, laptops, smartphones) and broadband Internet;

978-x-xxxx-xxxx-x/20/$31.00 ©2020 IEEE

Додаток Д

2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT 2020)

Zbarazh, Ukraine
23-26 September 2020

Volume 1
Pages 1-448

IEEE

IEEE Catalog Number: CFP20D36-POD
ISBN: 978-1-7281-7444-0

Development of Digital Platform to Identify and Monitor the Digital Business Transformation Index

Iryna Strutynska
Department of Computer Science
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
ringtons999@gmail.com

Lesia Dmytrotsa
Department of Computer Science
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
dmytrotsa.lesya@gmail.com

Halyna Kozbur
Department of Computer Science
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
kozbur.galina@gmail.com

Olha Hlado
Department of Computer Science
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
gladyo.olga@gmail.com

Pavlo Dudkin
Department of Innovative activity and Services Management
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
pavlo.dudkin@gmail.com

Olena Dudkina
Department of Management, Public Administration and Personnel
Ternopil National Economic University
Ternopil, Ukraine
olenadudkina65@gmail.com

***Abstract*—Companies around the world are facing a digital transformation that is changing and improving the business processes of an organization and providing opportunities to build a new business model.**

Digital transformation can be planned as a strategic initiative, with a clear vision, new business opportunities and a roadmap for implementation. In this way, the digital transformation roadmap becomes a strategic planning tool.

However, the awareness and digital literacy of small and medium-sized business leaders is not sufficient today. Most of them are not aware of the possibilities of using available services, platforms, digital tools and more. These trends adversely affect the speed of integration of our business and country into the global digital space.

We believe that one of the important tasks for accelerating the digital transformation of Ukrainian business is to create a digital platform that would allow the head of the business structure to assess the level of digital development of the organization and get an individual "roadmap" of business digitization. It is important to keep in mind that consistent and effective transformation requires a development vector, understandable tools for change, and an information source that provides guidance, assistance, and further learning opportunities.

Keywords*—*IT-project, web platform, digital transformation, Index of digital transformation of business structures, digital transformation roadmaps.

I. Stage 1: IT Project Initiating

The transition of industrial production to the concepts of "digital" economy, transformation of enterprises and widespread use of technology in business processes is rapidly happening around the world. It involves changing the business strategy and organization of its business processes, the emergence of new goals and means of achieving them, adapting the vision and mission of the company to the global trends in the impact of information technology [1].

However, in Ukraine such processes are proceeding rather slowly. One of the problems is the lack of awareness of entrepreneurs about the possibilities and ways of using digital technologies in their own processes [2], especially among representatives of small and medium-sized businesses (SMEs). The lack of available services, platforms, applications or portals slows down the speed of innovation in domestic business. This, in turn, does not allow business to integrate with global trends, which makes the access to international business complicated and impedes work in the international economic arena.

That is why the purpose of this study was to specify the stages of IT project management to design and implement a digital platform for small and medium businesses that would automatically determine the Digital Business Transformation Index and provide a roadmap listing specific recommendations for enhancing digital transformation.

The problem of the development of digital economy and transformational processes taking place in society under the influence of digitization has received a lot of attention among both foreign authors [4-8] and Ukrainian researchers [1-2, 9,10], etc. Despite numerous scientific studies on the development of information and communication technologies and the digital economy, we believe that the issues of the impact of digital technologies on business transformation are under-researched.

The main goals of the IT project implementation.

The digital platform will serve as an information base for domestic businesses, foster the creation of an appropriate eco-culture for digital maturity, create healthy competition between businesses, enhance the digital literacy of business owners and, accordingly, the human capital of organizations.

II. Stage 2: Project Planning

Decomposition of Work Breakdown Structure (WBS) for this project, which must be achieved in order to achieve the project objectives (Fig. 1).

Task A. Digital Business Transformation Index

The creation and use of the Digital Business Transformation Index was proposed as one of the methods for assessing the level of digital development of small and medium-sized enterprises (SMEs). If applied as a national methodology, it will be possible to evaluate the digital maturity of business structures and provide recommendations for improving it. This method allows to take into account sufficient indicators of impact on the development of business and the information society as a whole.

Regular calculation of the Index for a specific business structure can be used as a tool for monitoring and evaluating

Додаток Е

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 98279

Комп'ютерна програма "Цифрова трансформація бізнес-структур"

(вид, назва твору)

Автор(и) **Струтинська Ірина Володимирівна, Гладьо Ольга Юріївна, Дмитроца Леся Павлівна, Зозуля Роман Петрович, Козбур Галина Володимирівна**

(повне ім'я, псевдонім (за наявності))

Дата реєстрації 23.06.2020

Заступник Міністра розвитку економіки, торгівлі та сільського господарства України Д. О. Романович