МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп’ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

ЧИЖОВСЬКА Зоряна Ігорівна

Метод визначення споживчого кошика на основі машинного навчання / Machine Learning Based Method for Determining the Consumer Basket

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп’ютерні науки

Кваліфікаційна робота

Виконала:
студентка групи КНм-21
З.І. Чижовська
_____

Науковий керівник:
д.т.н., професор,
А.О. Саченко
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
Завідувач кафедри
_____ М.П. Комар

ТЕРНОПІЛЬ – 2022

Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Ступінь вищої освіти «магістр»
спеціальність: 122 "Комп'ютерні науки"
освітньо-професійна програма - Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____
“____” _____20___року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

_____Чижовській Зоряні Ігорівні_____

1. Тема роботи «Метод визначення споживчого кошика на основі машинного навчання»
керівник роботи_____д.т.н., професор, Саченко Анатолій Олексійович_____ ,
затверджені наказом по університету від 31 грудня 2021 року № 606.

2. Строк подання студентом закінченої кваліфікаційної роботи 16 листопада 2022 року.

3. Вихідні дані до роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

- Визначити суть та особливості поняття «споживчий кошик» та провести аналіз існуючих підходів визначення споживчого кошика;
- Визначити підходи до попередньої обробки даних;
- Визначити моделі кластеризації на основі машинного навчання;
- Розробити інтелектуальний метод формування споживчого кошику;
- Проаналізувати інформаційне забезпечення розробленого методу;
- Провести попередню обробку та дослідження даних;
- Провести програмну реалізацію та експериментальні дослідження на основі розробленого методу.

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

- Структура інтелектуального методу формування споживчого кошику;
- Графіки реалізації інтелектуального методу формування споживчого кошику.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 11 жовтня 2021 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Сучасний стан предметної області і постановка задачі дослідження	12.2021 р. – 03.2022 р.	
2	Інтелектуальний метод формування споживчого кошику	03.2022 р. – 05.2022 р.	
3	Програмна реалізація розробленого методу	05.2022 р. – 11.2022 р.	
4	Повне завершення та представлення кваліфікаційної роботи на кафедру	16.11.2022 р.	

Студент _____ З.І. Чижовська
підпис

Керівник роботи _____ д.т.н., професор, А.О. Саченко
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему «Метод визначення споживчого кошика на основі машинного навчання» на здобуття ступеня вищої освіти «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньо-професійної програми «Комп'ютерні науки» написана обсягом 107 сторінок і містить 17 ілюстрацій, 3 таблиць, 3 додатки та 54 використаних джерел.

Метою кваліфікаційної роботи є розробка інтелектуального методу формування споживчого кошику.

Методи досліджень: у роботі використані методи математичного програмування, системного аналізу, машинного навчання.

Результати дослідження: розроблено метод визначення споживчого кошика на основі машинного навчання по даних мереж супермаркетів, що дозволить формувати прожитковий мінімум та запобігти можливим гуманітарно-економічним катастрофам при зростанні бідності.

Результати роботи можуть бути використані для оцінки соціально-економічних показників територіальних громад.

Ключові слова: СПОЖИВЧИЙ КОШИК, МАШИННЕ НАВЧАННЯ, КЛАСТЕРИЗАЦІЯ, К-СЕРЕДНІ, НАБІР ДАНИХ.

RESUME

The qualification work on the topic «Machine Learning Based Method for Determining the Consumer Basket» for the Master's degree in specialty 122 «Computer Science» educational and professional program «Computer Science» is prepared in 107 page volume and contains 17 illustrations, 3 tables, 3 applications and 54 sources for references.

The purpose of the qualification work is to develop an intelligent method of forming a consumer basket.

Research methods: the work uses methods of mathematical programming, system analysis, and machine learning.

Research results: a method of determining the consumer basket based on machine learning based on the data of supermarket chains has been developed, which will enable to form a living wage and prevent possible humanitarian and economic disasters in conditions of poverty growing.

The results of the work can be applied for assessing the socio-economic indicators of territorial communities.

Keywords: CONSUMER BASKET, MACHINE LEARNING, CLUSTERIZATION, K-MEANS, DATASET.

ЗМІСТ

ВСТУП

На сьогодні методика обчислення споживчого кошика є надзвичайно актуальним питанням, оскільки на основі споживчого кошика формується прожитковий мінімум, згідно з яким розраховують пенсії, соціальні пільги та виплати. В Україні формуванням споживчого кошика безпосередньо займається Кабінет Міністрів України. На основі даних споживчого кошика та споживчого бюджету розраховуються такі показники, як мінімальна заробітна плата та мінімальна пенсія.

Вартість споживчого кошика залежить від рівня роздрібних цін на товари та тарифів на платні послуги (наприклад, комунальні платежі). Така практика відома у всьому цивілізованому світі. Із кожного виду потреб до розрахунку включають придбання відносно дешевих товарів, як правило, за державними фіксованими цінами. Якщо, наприклад, на ринку даний продукт або послуга продається за більш низькими цінами, то за основу береться найнижчий рівень.

Поняття споживчого кошика існує у багатьох країнах світу. Його ціна і національні особливості в кожній країні різні. Наприклад, споживчий кошик американця нараховує 350 продуктів і послуг, француза – 507, англійця – 350, німця – 475. Український споживчий кошик нещодавно було розширено до 297 найменувань.

Купівельна спроможність людей може різко змінюватись відносно кризових ситуацій, таких як війна, пандемії. Тому, оцінка споживчого кошика і його змінна повинна бути швидкою для того, щоб не допустити гуманітарних катастроф та збільшення бідності.

У зв’язку з цим, можна вважати, що розробка методу визначення споживчого кошика на основі машинного навчання по даних мереж супермаркетів є актуальною та дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що

дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшенню бідності.

Метою роботи є розробка інтелектуального методу формування споживчого кошика.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Визначити суть та особливості поняття «споживчий кошик» та провести аналіз існуючих підходів визначення споживчого кошика;
2. Визначити підходи до попередньої обробки даних;
3. Визначити моделі кластеризації на основі машинного навчання;
4. Розробити інтелектуальний метод формування споживчого кошику;
5. Проаналізувати інформаційне забезпечення розробленого методу;
6. Провести попередню обробку та дослідження даних;
7. Провести програмну реалізацію та експериментальні дослідження на основі розробленого методу.

Об’єктом дослідження є процес впровадження машинного навчання для визначення споживчого кошика.

Предметом дослідження є сукупність теоретико-методологічних та прикладних засад по визначенні споживчого кошика.

Методи дослідження: у роботі використані методи математичного програмування, системного аналізу, машинного навчання.

Наукова новизна отриманих результатів: розроблено метод визначення споживчого кошика на основі машинного навчання по даних мереж супермаркетів, що дозволить формувати прожитковий мінімум та запобігти можливим гуманітарно-економічним катастрофам зі збільшенням бідності.

Практичне значення одержаних результатів полягає у визначенні споживчого кошика на основі машинного навчання по даних мереж супермаркетів.

Структура та обсяг роботи. Дипломна робота складається із вступу, трьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг роботи становить 107 сторінок комп'ютерного тексту, який включає 17 рисунків та 3 таблиці. Список використаних джерел із 54 найменувань викладено на 9 сторінках.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися на:

– The 28th International Conference "Information and Software Technologies (ICIST 2022)", 13-15 October, 2022, Kaunas, Lithuania, та опубліковано в Communications in Computer and Information Science, Springer Cham з індексацією в Scopus;

– XIX Міжнародній науково-практичній конференції молодих вчених, 13 травня 2022 року, Тернопіль, Україна.

1 СУЧАСНИЙ СТАН ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Суть та особливості поняття «споживчий кошик»

Споживчий кошик - одна з найважливіших категорій, яка існує практично в кожній країні світу [31]. Застосовується для визначення розміру прожиткового мінімуму, використовується в якості основи для порівняння розрахункових і фактичних рівнів споживання.

Сьогодні тема актуальна, адже на основі даних про склад і розмір кошика формується розмір прожиткового мінімуму, пенсій, соціальних допомог та інших виплат. Питання про склад споживчого кошика хвилює не тільки вчених, а й простих громадян нашої країни вже багато років.

Споживчий кошик - це приблизний набір розрахунків, асортимент товарів і послуг, які необхідні кожному громадянину країни за певний проміжок часу (рік), який повинен зберегти здоров'я і забезпечити його життєдіяльність. Споживчий кошик включає в себе наступні групи товарів і послуг [32]:

1. Харчові продукти. На них припадає основна частина вартості кошика. Для переважної більшості громадян України та інших країн значна частина сімейного бюджету витрачається на харчування.

2. Непродовольчі товари. У цій групі переважають медикаменти, предмети побуту, одяг і взуття. Ці категорії незамінні і життєво необхідні населенню. Далі йдуть витрати на меблі, домашніх тварин, обладнання та інше.

3. Послуги. Такі, як оплата навчання, дозвілля та відпочинку. Сюди ж відносяться витрати на транспорт, страхування і комунальні платежі.

Склад кошика змінюється з плином часу і істотно різниться в різних країнах. Вартість кошика враховується при розрахунку мінімальної

заробітної плати, купівельної спроможності населення, розміру різних видів соціальних виплат.

Споживчий кошик в Україні містить 296 різних товарів і послуг. Його склад розраховується відповідно до постанови Кабінету Міністрів України від 11 жовтня 2016 року 780 "Про затвердження комплектів харчових продуктів, наборів непродовольчих товарів і комплексів послуг для основних соціальних і демографічних груп населення". На виконання цієї постанови Мінсоцполітики видало наказ від 3 лютого 2017 року г178/147/31 «Про затвердження Методики визначення прожиткового мінімуму», яким наводиться не тільки методика, а й умовні приклади розрахунку вартості прожиткового мінімуму [25].

Прожитковий мінімум - це величина того, що достатньо для забезпечення нормального функціонування людського організму, мінімального набору товарів і мінімального набору послуг, необхідних для задоволення основних потреб індивіда [26, 27].

На відміну від України, в Польщі існує дві концепції споживчого кошика. Minimum egzystencji - це прожитковий мінімум, споживчий кошик, в який входять тільки ті елементи, які необхідні для підтримки життєдіяльності людського організму, а також психофізичного здоров'я. Його середня вартість становить приблизно 484 злотих, тобто 3 630 гривень. Але це межа крайньої бідності.

Друге поняття - Minimum socjalne. Це соціальний мінімум, який показує вартість такого споживчого кошика, в який входять витрати на харчування, житло, одяг, охорону праці, гігієну, витрати на спілкування, освіту, сімейні та дружні контакти, невелику участь в культурному житті. Це модель задоволення своїх потреб на рівні мінімального добробуту - аналог нашого споживчого кошика. У Польщі його вартість в середньому становить 914,5 злотих, що дорівнює 7061 гривні [28].

Різниця полягає не тільки у вартості споживчих кошиків, але і в способах її формування. У Польщі уряд аналізує, що і в якій кількості купив середньостатистичний поляк, а в Україні начебто диктує, що і за скільки купувати. Також істотно відрізняється і кількість товарів і послуг, які можна придбати за прожитковий мінімум, тобто кількість товарів і послуг, що входять в споживчі кошики Польщі та України.

Наприклад, українець повинен з'їсти 107 г житнього хліба, 170 г пшениці, 260 г картоплі, 175 г різних плодів, 33 г курки, 83 г м'яса, 14 г вершкового масла. У той час як добова норма поляка - 120 г житнього хліба, 125 г пшеничного хліба, 600 г картоплі, 190 г фруктів, 90 г курки, 165 г м'яса, 25 г вершкового масла.

Показники також відрізняються в категорії «Побутова техніка». Холодильник повинен прослужити українцям 15 років, телевізор - 10, праска - 9, пральна машина - 14, настільний годинник - 6 років. Для поляків термін служби побутової техніки набагато коротший. Наприклад, поляк може купити холодильник на 1 рік, телевізор на 7 років, праску і столовий годинник на 1 рік, пральну машину на 8 років.

Також разючою є різниця між кількістю інформаційних, культурних та рекреаційних послуг, що входять до споживчих кошиків України та Польщі. Наприклад, українська сім'я може придбати 52 газети і журнали. на рік, тоді як польська - 150 шт. В рік українець може відвідувати театр, музей, зоопарк або кіно 6 разів на рік, а поляк - 12.

Аналізуючи споживчі кошики України та Польщі, можна помітити, що показники в категорії «Меблі (на сім'ю)» істотно відрізняються. Наприклад, українець повинен купити гардероб на 25 років, а поляк - один на 17, дзеркало у ванній для українця має прослужити 20 років, для поляка - всього 3 роки, письмовий стіл для українця повинен прослужити 25 років, а для поляка - один на 12 років. Також є різні дані щодо категорії «Посуд». 2 тарілки на 3 роки, 2 сковорідки на 9 років, 1 чайник на 8 років, 2 ложки,

2 виделки і 1 ніж на 10 років для українця. Тоді як 9 тарілок на 3 роки, 2 каструлі на 6 років, 1 чайник для заварювання на 4 роки, 4 ложки, 4 виделки на 8 років. Показники в категорії «Гігієна» практично однакові [29].

В ході аналізу було визначено, що споживчий кошик - це мінімальний набір товарів і послуг, який повинен забезпечувати повсякденне життя людини; прожитковий мінімум - вартість цього мінімального набору товарів і послуг; визначено, що кошик включає 3 категорії - продовольчі товари, непродовольчі товари та послуги; Пояснюється методика розрахунку прожиткового мінімуму. Вважається, що прожитковий мінімум в Україні занижений, тому його потрібно переглядати. Наразі, склад споживчого кошика не може повністю задовольнити базові соціальні та культурні потреби українців. Він являє собою сукупність товарів і послуг, які можуть сприяти тільки фізичному виживанню [30]. Необхідно проаналізувати, що і в якій кількості споживає середньостатистичний українець і на основі цих даних сформувати склад і розмір споживчого кошика, наблизити його до сьогоднішніх реалій та переглянути зміст споживчого кошика і норми споживання.

1.2 Аналіз існуючих підходів визначення споживчого кошика

Дослідження [1] визначило інструменти формування споживчого кошика в різних сферах і рівнях, включаючи національні закони, урядову політику та ініціативи місцевого самоврядування, та проаналізовано їх з різними формами коротких ланцюгів поставок, наявними на території Франції, які є ключовими компонентами місцевої харчової системи, такі як прямий маркетинг, магазини виробників, міське сільське господарство та поставки в громадське харчування. У роботі [5] досліджено еволюцію зв’язків коливання-передачі між класами індексу споживчих цін та побудували зважені мережі причинно-наслідкових зв’язків Грейнджера

(WGCN) для кожної країни G7. У статті [6] моделюється індекс споживчих цін у Сполученому Королівстві як складну мережу та застосовуються методи кластеризації та оптимізації для вивчення еволюції мережі в часі. У статті [7] аналізується подібність моделей споживання споживачів у десяти азіатських країнах, використовуючи три добре відомі загальносистемні моделі попиту, а саме Роттердам, CBS та AIDS.

У дослідженні [2] зроблена спроба передбачити сегмент середньо-атлантичного ринку вина – на основі купівельної поведінки, ставлення та соціально-демографічних ознак. Кластерний аналіз використовується для сегментації середньо-атлантичного винного ринку. У дослідженні [3] проведено аналіз ринкового кошика за допомогою апріорного алгоритму для пошуку моделей споживачів у купівлі товарів за даними транзакцій. У роботі [4] пропонується дослідження інтеграції різнорідних джерел даних з продуктового супермаркету на основі методів аналізу ринкового кошика. У статті [8] йдеться про дослідження та впровадження механізму рекомендацій на відомій португальській платформі електронної комерції, що спеціалізується на звичайному та спортивному одязі, з метою збільшення залучення клієнтів, забезпечуючи персоналізований досвід із різними типами рекомендацій на платформі. У дослідженні [10] були зроблені спроби заповнити пробіл у дослідженні, використовуючи аналітику великих даних для аналізу приблизно 44 000 записів про транзакції в точках продажу для 26 000 клієнтів тайванського роздрібного магазину, щоб зрозуміти, як риси особистості споживача пов’язані з країною походження (COO), риси (індивідуальність бренду) марок пива, а також для прогнозування життєвої цінності потенційного клієнта (CLV).

У дослідженні [12] оцінюється ефективність різних підходів кластеризації даних для пошуку вигідних сегментів споживачів у галузі гостинності у Великобританії. Дослідження [13, 14] засноване на моделі RFM (Recency, Frequency and Monetary) та використовує принципи

сегментації набору даних за допомогою алгоритму K-Means. Отримані таким чином результати щодо операцій з продажу порівнюються з різними параметрами, такими як нещодавній продаж, частота продажів та обсяг продажів. У рукописі [15] представляється модель прогнозування, засновану на поведінці кожного клієнта з використанням методів аналізу даних. Запропонована модель використовує базу даних супермаркету та додаткову базу даних від Amazon, які містять інформацію про покупки клієнтів.

У статті [22] розглядаються можливості, як великі дані випливають на роздрібну торгівлю, зокрема за п’ятьма основними вимірами даних — даних, що стосуються клієнтів, продуктів, часу, (геопросторового) розташування та каналу. З метою [23] вивчення використання розумних технологій з точки зору маркетингу та роздрібної торгівлі, це дослідження надає нові підходи до дисциплін, які спонукаються безперервним удосконаленням технології. Його основна увага зосереджена на розумних технологіях поведінки споживачів через більш комплексну та сучасну перспективу, яка поєднує маркетинг і роздрібну торгівлю з іншими дисциплінами, такими як психологія, медіа-дослідження та соціологія.

Дослідження [21] зосереджено на розумних споживачах, які добровільно беруть участь у діяльності зі створення цінності, щоб концептуалізувати розумне спільне створення досвіду (SEC) і розумну систему обслуговування.

У дослідженні [9] був застосований інтуїтивний алгоритм нечіткої кластеризації до даних покупців у супермаркеті відповідно до суми, витраченої на деякі групи продуктів.

У дослідженні [11] запропонували модифікації алгоритму CDFCM через виявлені в ньому певні недоліки, в результаті яких утворився модифікований динамічний нечіткий c-алгоритм середніх (MDFCM).

У дослідженні [16] проведена сегментація клієнтів за допомогою кластеризації K-середніх у неконтрольованому машинному навчанні. Проведено класифікацію цільових клієнтів за допомогою кластеризації на основі демографічної інформації (наприклад, стать, вік, дохід тощо), географічної інформації, психографії та поведінкових даних.

Згадані вище роботи здебільшого аналізують клієнта зі сторони маркетингу супермаркетів. Проте немає робіт, які б оцінювали клієнта, як споживача та формували його соціальний портрет на основі, якого можна сформувати споживчий кошик та відповідно прожитковий мінімум.

У зв’язку із цим, метою даної роботи є розробка інтелектуального методу формування споживчого кошику.

На відміну від аналогів [9, 11, 16] розроблений інтелектуальний метод формування споживчого кошику, дозволить на основі даних мереж супермаркетів формувати набір товарів у споживчому кошику та встановлювати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшенню бідності, особливо у кризовий період економіки (війна, пандемія).

1.3 Аналіз методів машинного навчання

Машинне навчання (ML) — це дисципліна штучного інтелекту (ШІ), яка надає машинам здатність автоматично навчатися на основі даних і минулого досвіду, одночасно визначаючи шаблони, щоб робити прогнози з мінімальним втручанням людини [33].

Методи машинного навчання дозволяють комп’ютерам працювати автономно без явного програмування. Програми ML отримують нові дані, і вони можуть самостійно навчатися, рости, розвиватися та адаптуватися.

Машинне навчання отримує глибоку інформацію з великих обсягів даних, використовуючи алгоритми для визначення шаблонів і навчання в

ітераційному процесі. Алгоритми ML використовують обчислювальні методи, щоб навчатися безпосередньо з даних замість того, щоб покладатися на будь-яке заздалегідь визначене рівняння, яке може слугувати моделлю.

Продуктивність алгоритмів ML адаптивно покращується зі збільшенням кількості доступних зразків під час процесів «навчання». Наприклад, глибоке навчання — це субдомен машинного навчання, який навчає комп’ютери імітувати природні людські риси, як-от навчання на прикладах. Він пропонує кращі параметри продуктивності, ніж звичайні алгоритми ML.

Незважаючи на те, що машинне навчання не є новою концепцією, починаючи з часів Другої світової війни, коли використовувалася машина Enigma, здатність автоматично застосовувати складні математичні обчислення до зростаючих обсягів і різновидів доступних даних є відносно нещодавньою розробкою.

Сьогодні, з розвитком великих даних, Інтернету речей і повсюдних обчислень, машинне навчання стало необхідним для вирішення проблем у багатьох сферах, таких як [33]:

- Обчислювальне фінансування (кредитний скоринг, алгоритмічна торгівля)
- Комп'ютерний зір (розпізнавання обличчя, відстеження руху, виявлення об'єктів)
- Обчислювальна біологія (секвенування ДНК, виявлення пухлин головного мозку, відкриття ліків)
- Автомобільна, аерокосмічна та промислова промисловість (прогнозне технічне обслуговування)
- Обробка природної мови (розпізнавання голосу)

Розглянемо детальніше, як працює машинне навчання (Рис.1.1).

Алгоритми машинного навчання формуються на основі навчального набору даних для створення моделі. Коли нові вхідні дані вводяться в навчений алгоритм ML, він використовує розроблену модель для прогнозування.

Далі прогноз перевіряється на точність. Базуючись на його точності, алгоритм ML або розгортається, або постійно навчається за допомогою розширеного навчального набору даних, доки не буде досягнуто бажаної точності.

Алгоритми машинного навчання можна навчити багатьма способами, у кожного з яких є свої плюси та мінуси. Виходячи з цих методів і способів навчання, машинне навчання поділяється на чотири основні типи (Рис.1.2) [33]:

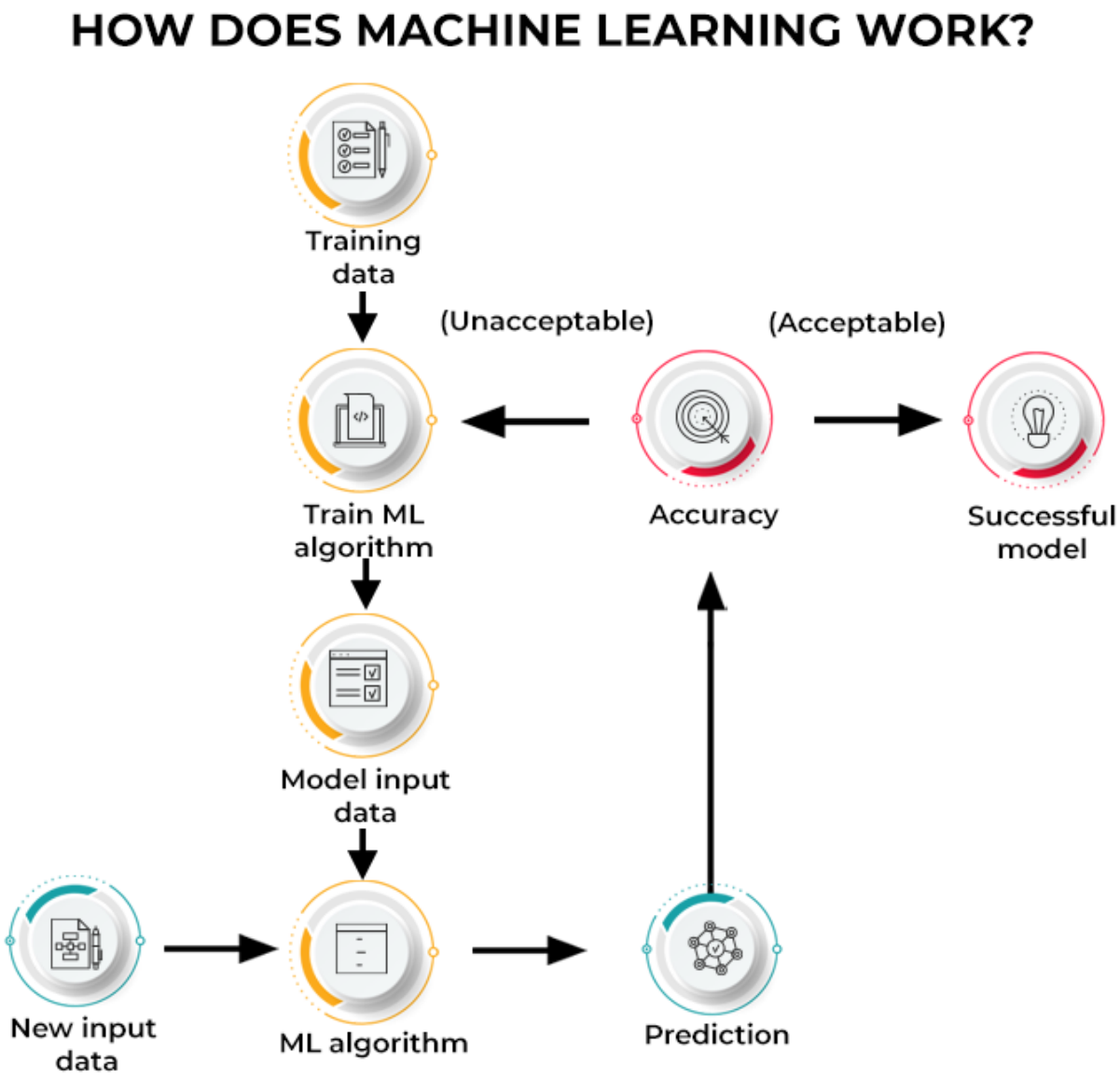


Рис.1.1 - Як працює машинне навчання

1. Контрольоване машинне навчання

Цей тип ML передбачає нагляд, коли машини навчаються на позначених наборах даних і можуть передбачати результати на основі наданого навчання. Позначений набір даних вказує на те, що деякі вхідні та вихідні параметри вже зіставлено. Таким чином, машина навчається з введенням і відповідним виходом. Пристрій створено для прогнозування результату за допомогою тестового набору даних на наступних етапах.

Наприклад, розглянемо вхідний набір даних зображень папуги та ворони. Спочатку машину навчають розуміти зображення, включаючи

колір, очі, форму та розмір папуги та ворони. Після навчання надається вхідне зображення папуги, і очікується, що машина ідентифікує об’єкт і передбачить результат. Навчена машина перевіряє різні характеристики об’єкта, такі як колір, очі, форма тощо, у вхідному зображенні, щоб зробити остаточний прогноз. Це процес ідентифікації об’єктів у керованому машинному навчанні.

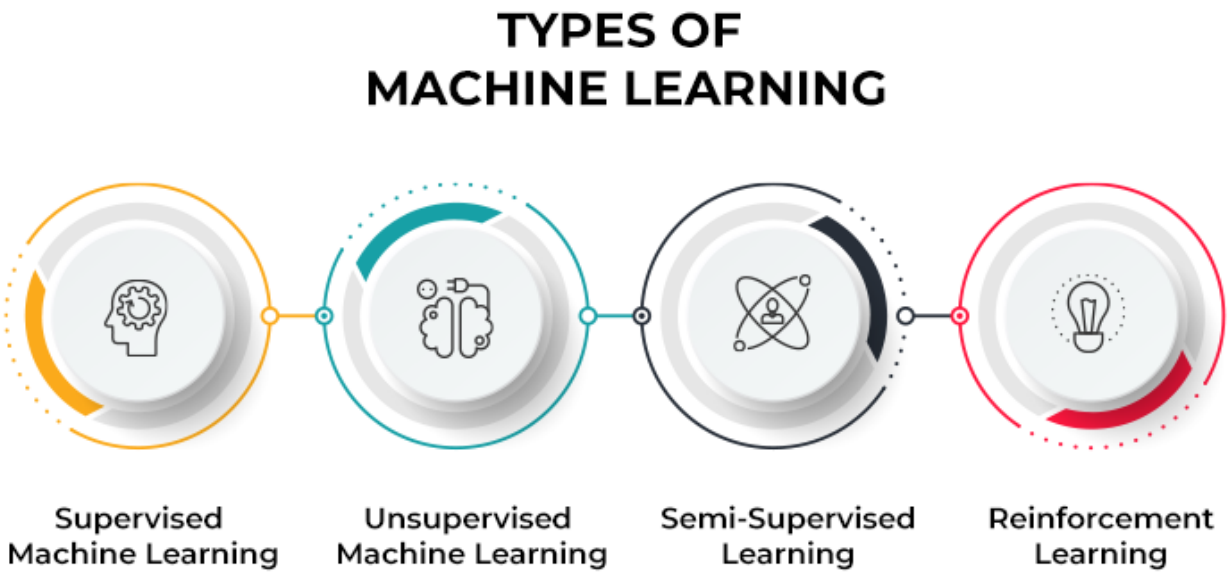


Рис.1.2 - Типи машинного навчання

Основна мета методики навчання під керівництвом полягає в тому, щоб зіставити вхідну змінну (a) з вихідною змінною (b). Кероване машинне навчання далі класифікується на дві великі категорії:

- Класифікація: вона стосується алгоритмів, які вирішують проблеми класифікації, де вихідна змінна є категоричною, наприклад, так чи ні, правда чи хибність, чоловік чи жінка, тощо. Реальні програми цієї категорії очевидні у виявленні спаму та фільтрації електронної пошти. Деякі відомі алгоритми класифікації включають алгоритм

випадкового лісу, алгоритм дерева рішень, алгоритм логістичної регресії та алгоритм опорних векторних машин.

- Регресія: алгоритми регресії обробляють проблеми регресії, коли вхідні та вихідні змінні мають лінійний зв'язок. Відомо, що вони передбачають безперервні вихідні змінні. Приклади включають прогноз погоди, аналіз ринкових тенденцій, тощо. Популярні алгоритми регресії включають алгоритм простої лінійної регресії, алгоритм багатовимірної регресії, алгоритм дерева рішень і регресію ласо.

2. Машинне навчання без нагляду

Навчання без нагляду стосується техніки навчання, яка позбавлена нагляду. Тут машина навчається за допомогою немаркованого набору даних і може передбачати результат без будь-якого нагляду. Алгоритм неконтрольованого навчання має на меті згрупувати несортований набір даних на основі подібностей, відмінностей і шаблонів вхідних даних.

Наприклад, розглянемо вхідний набір даних зображень контейнера, наповненого фруктами. Тут зображення не відомі моделі машинного навчання. Коли ми вводимо набір даних у модель ML, завдання моделі полягає в тому, щоб ідентифікувати шаблон об'єктів, наприклад, колір, форму чи відмінності, які видно на вхідних зображеннях, і класифікувати їх. Після категоризації машина прогнозує результат, коли його перевіряють за допомогою тестового набору даних.

Неконтрольоване машинне навчання далі класифікується на два типи:

- Кластеризація: техніка кластеризації стосується групування об'єктів у кластери на основі таких параметрів, як подібність або відмінність між об'єктами. Наприклад, групування клієнтів за продуктами, які вони купують.
  Деякі відомі алгоритми кластеризації включають алгоритм кластеризації K-Means, алгоритм середнього зсуву, алгоритм

DBSCAN, аналіз основних компонентів і аналіз незалежних компонентів.

- Асоціація: навчання асоціаціям стосується визначення типових зв’язків між змінними великого набору даних. Воно визначає залежність різних елементів даних і відображає відповідні змінні. Типові програми включають аналіз використання Інтернету та аналіз ринкових даних. Популярні алгоритми, які підкоряються правилам асоціації, включають алгоритм Apriori, алгоритм Eclat і алгоритм FP-Growth.

3. Напівконтрольоване навчання

Напівконтрольоване навчання містить характеристики як контрольованого, так і неконтрольованого машинного навчання. Воно використовує комбінацію позначених і не позначених наборів даних для навчання своїх алгоритмів. Використовуючи обидва типи наборів даних, напівконтрольоване навчання долає недоліки згаданих вище варіантів.

Розглянемо приклад студента коледжу. Студент, який вивчає концепцію під наглядом викладача в коледжі, називається навчанням під наглядом. Під час навчання без нагляду учень самостійно вивчає ту саму концепцію вдома без керівництва вчителя. Тим часом перегляд концепції студентом після навчання під керівництвом викладача в коледжі є напівконтрольованою формою навчання.

4. Навчання з підкріпленням

Навчання з підкріпленням – це процес, заснований на зворотньому зв’язку. Тут компонент штучного інтелекту автоматично аналізує навколишнє середовище за допомогою методу «хіт і проб», вживає заходів, навчається на досвіді та покращує продуктивність. Компонент винагороджується за кожну хорошу дію та штрафується за кожен неправильний рух. Таким чином, компонент навчання з підкріпленням спрямований на максимізацію винагород за виконання хороших дій.

На відміну від навчання під наглядом, у навчанні з підкріпленням відсутні позначені дані, і агенти навчаються лише через досвід. Розглянемо відеоігри. Тут гра визначає середовище, а кожен рух агента підкріплення визначає його стан. Агент має право отримувати відгуки за допомогою покарань і винагород, що впливає на загальний рахунок гри. Кінцева мета агента – отримати високий бал.

Навчання з підкріпленням застосовується в різних областях, таких як теорія ігор, теорія інформації та багатоагентні системи. Навчання з підкріпленням далі поділяється на два типи методів або алгоритмів:

- Навчання з позитивним підкріпленням: це стосується додавання підкріплюючого стимулу після певної поведінки агента, що підвищує ймовірність повторення такої поведінки в майбутньому, наприклад, додавання винагороди після поведінки.
- Навчання з негативним підкріпленням: стосується зміцнення певної поведінки, яка дозволяє уникнути негативного результату.

Отже, за своєю суттю машинне навчання є формою штучного інтелекту (ШІ), спрямованого на створення машин, які можна навчати. Моделі машинного навчання виконують завдання, вивчаючи дані замість того, щоб їх явно програмувати. Підкатегорія штучного інтелекту, машинне навчання використовує статистичні методи для отримання інформації з терабайтів даних, а при контакті з новими даними ці програми навчаються та розвиваються самі. Іншими словами, програми ML навчаються на основі попередніх обчислень і використовують розпізнавання шаблонів для покращення та отримання обґрунтованих і надійних результатів.

1.4 Постановка задачі

На сучасному етапі децентралізації проблеми рівня життя стали не лише центральними, а й регіональними. В Україні, внаслідок економічного падіння, все частіше для визначення рівня життя почали вживати категорії, що відображають його мінімальне значення: «споживчий кошик», «прожитковий мінімум» та поняття «бідність».

Оскільки Україна обрала шлях європейської інтеграції, обґрунтованою є необхідність докорінного перегляду складу споживчого кошика та підходів до формування прожиткового мінімуму, до якого прив'язані всі виплати та соціальні стандарти держави. Запропоновані зміни щодо методики розрахунку прожиткового мінімуму сприятимуть зниженню рівня бідності населення, підвищенню рівня здоров'я, поліпшенню трудового потенціалу, підвищенню продуктивності праці і, відповідно, підвищенню якості життя населення та сталому розвитку економічної системи держави.

У зв’язку з цим, можна вважати, що розробка інтелектуального методу формування споживчого кошику на основі даних мереж супермаркетів є актуальною та дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

Відповідно, метою роботи є розробка інтелектуального методу формування споживчого кошику.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Визначити суть та особливості поняття «споживчий кошик» та провести аналіз існуючих підходів визначення споживчого кошика;

2. Визначити підходи до попередньої обробки даних;
3. Визначити моделі кластеризації на основі машинного навчання;
4. Розробити інтелектуальний метод формування споживчого кошику;
5. Проаналізувати інформаційне забезпечення розробленого методу;
6. Провести попередню обробку та дослідження даних;
7. Провести програмну реалізацію та експериментальні дослідження на основі розробленого методу.

Завдання 1 розглянуто в параграфах 1.1, 1.2 та 1.3 відповідно, а виконання завдань 4-7 представлені у параграфах 2.1-3.3.

Висновки до розділу 1

1. У розділі дослідження визначено, що споживчий кошик - це мінімальний набір товарів і послуг, який повинен забезпечувати повсякденне життя людини; прожитковий мінімум - вартість цього мінімального набору товарів і послуг; визначено, що кошик включає 3 категорії - продовольчі товари, непродовольчі товари та послуги; Пояснюється методика розрахунку прожиткового мінімуму.

2. Проведено аналіз існуючих підходів та визначено, що на відміну від аналогів [9, 11, 16] розроблений інтелектуальний метод формування споживчого кошику, дозволить на основі даних мереж супермаркетів формувати набір товарів у споживчому кошику та встановлювати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшенню бідності, особливо у кризовий період економіки (війна, пандемія).

3. Визначено, що машинне навчання є формою штучного інтелекту (ШІ), спрямованого на створення машин, які можна навчати. Моделі машинного навчання виконують завдання, вивчаючи дані замість того, щоб їх явно програмувати. Іншими словами, програми ML навчаються на основі попередніх обчислень і використовують розпізнавання шаблонів для покращення та отримання обґрунтованих і надійних результатів.

4. На основі проведеного дослідження сформовано постановку задачі, а саме розробка інтелектуального методу формування споживчого кошика на основі даних мереж супермаркетів.

2 ВИЗНАЧЕННЯ СПОЖИВЧОГО КОШИКА НА ОСНОВІ МАШИННОГО НАВЧАННЯ

2.1 Попередня обробка даних

Попередня обробка даних — це етап у процесі інтелектуального аналізу даних, на якому необроблені дані перетворюються у формат, який можна зрозуміти й проаналізувати комп'ютерами та машинним навчанням [34].

Необроблені реальні дані у вигляді тексту, зображень, відео тощо є сирими. Вони не тільки можуть містити помилки та невідповідності, але й часто є неповними і не мають регулярного, єдиного дизайну.

Машини, переважно, обробляють акуратну інформацію – вони зчитують дані як 1 і 0. Тому обчислювати структуровані дані, наприклад, цілі числа та відсотки, легко. Однак неструктуровані дані у вигляді тексту та зображень потрібно спочатку очистити та відформатувати перед аналізом.

Використовуючи набори даних для навчання моделей машинного навчання, існує така фраза «сміття входить, сміття виходить». Це означає, що якщо використовувати погані або «брудні» дані для навчання моделі, отримується погана, неправильно навчена модель, яка насправді не матиме відношення до вашого аналізу.

Добрі, попередньо оброблені дані, є важливими не менш, ніж потужні алгоритми, тому що моделі машинного навчання, навчені сирими даними, насправді можуть зашкодити проведеному аналізу, – приносячи «сміттєві» результати.

Залежно від методів збору даних і джерел, можна отримати дані, які виходять за межі діапазону або містять неправильну характеристику, як-от дохід сім'ї нижче нуля або зображення з набору «тварин зоопарку», яке

насправді є деревом. У нашому наборі можуть бути відсутні значення або поля. Або, наприклад, текстові дані часто містять слова з орфографічними помилками та нерелевантні символи, URL-адреси, тощо.

Якщо належним чином обробити та очистити свої дані, можна налаштуватись на більш точні подальші процеси. Важливість «прийняття рішень на основі даних» ґрунтується на використанні правильних даних, щоб уникнути поганих рішень.

Існують наступні кроки, які потрібно буде виконати, щоб переконатися, що дані успішно оброблені [34]:

1. Оцінка якості даних;
2. Очищення даних;
3. Перетворення даних;
4. Скорочення даних;

Оцінка якості даних (DQA) — це процес наукової та статистичної оцінки даних, щоб визначити, чи відповідають вони якості, необхідній для проектів або бізнес-процесів, і чи належать вони до потрібного типу та кількості, щоб фактично підтримувати цільове використання [35]. Його можна вважати набором рекомендацій і методів, які використовуються для опису даних, враховуючи контекст програми, і для застосування процесів для оцінки та покращення якості даних.

Оцінка якості даних (DQA) виявляє проблеми з технічними та бізнес-даними, які дозволяють організації правильно планувати стратегії очищення та збагачення даних. Зазвичай це робиться для підтримки цілісності систем, стандартів забезпечення якості та проблем відповідності. Як правило, проблеми з технічною якістю, такі як неузгоджена структура та стандарти, відсутні дані або дані за замовчуванням, а також помилки в полях даних легко виявити та виправити, але до більш складних проблем слід підходити за допомогою більш визначених процесів.

DQA зазвичай виконується для усунення суб’єктивних проблем, пов’язаних з бізнес-процесами, наприклад, створення точних звітів, а також для забезпечення належної роботи процесів, що керуються та залежать від даних.

Процеси DQA узгоджені з найкращими практиками та набором попередніх умов, а також з п’ятьма вимірами якості даних:

– Точність і надійність
– Ремонтопридатність
– Доступність
– Методична обґрунтованість
– Гарантії чесності

Очищення даних — це процес виправлення або видалення неправильних, пошкоджених, неправильно відформатованих, дублікатів або неповних даних у наборі даних [36]. При поєднанні кількох джерел даних існує багато можливостей для дублювання даних або неправильного позначення. Якщо дані неправильні, результати та алгоритми ненадійні, навіть якщо вони можуть виглядати правильними. Немає єдиного абсолютного способу прописати точні кроки в процесі очищення даних, оскільки процеси відрізняються від набору даних до набору даних. Але вкрай важливо створити шаблон для процесу очищення даних, щоб ви завжди знали, що робите це правильно.

Хоча методи, які використовуються для очищення даних, можуть відрізнятися залежно від типів даних, які зберігаються, можна виконати наступні кроки, щоб скласти структуру для очистки даних [36].

Крок 1. Видалення дублікатів або нерелевантних спостережень. Видалення небажаного спостереження зі набору даних, у тому числі повторювані або нерелевантні спостереження. Дубльовані спостереження трапляються найчастіше під час збору даних. Коли об’єднуються набори даних із кількох місць, збираються або отримуються дані від клієнтів чи

кількох відділів, є можливості для створення дублікатів даних. Дедублікація є однією з найбільших областей, які слід враховувати в цьому процесі. Нерелевантні спостереження — це коли є спостереження, які не вписуються в конкретну проблему, яку потрібно проаналізувати.

Крок 2. Виправлення структурних помилок. Структурні помилки виникають, коли вимірюються або передаються дані та помічаються дивні угоди про найменування, опечатки чи неправильне використання великих літер. Ці невідповідності можуть спричинити неправильне позначення категорій або класів.

Крок 3. Фільтрування небажаних викидів. Часто трапляються одноразові спостереження, які, на перший погляд, не відповідають даним, які аналізуються. Якщо є законна причина видалити викид, як-от неправильне введення даних, це покращить ефективність даних, з якими ведеться робота. Однак, іноді саме поява викиду підтверджує теорію, над якою може працювати дослідник.

Крок 4. Обробка відсутніх даних. Можна ігнорувати відсутні дані, оскільки багато алгоритмів не сприймають відсутні значення. Є кілька способів усунути відсутні дані. Жоден з них не є оптимальним, але можна розглянути кожен.

1. Як перший варіант, можна видалити спостереження, які мають відсутні значення, але це призведе до видалення або втрати інформації, тому потрібно пам’ятати про це, перш ніж видаляти їх.
2. Як другий варіант, можна ввести відсутні значення на основі інших спостережень. Знову ж таки, є можливість втратити цілісність даних, оскільки можна керуватися припущеннями, а не реальними спостереженнями.
3. Як третій варіант, можна змінити спосіб використання даних для ефективної навігації нульовими значеннями.

Крок 5: Перевірка та контроль якості. Наприкінці процесу очищення даних можна відповісти на ці запитання в рамках базової перевірки:

- Чи дані мають сенс?
- Чи відповідають дані відповідним правилам для свого поля?
- Це підтверджує чи спростовує вашу робочу теорію, чи відкриває якусь інформацію?
- Чи можете ви знайти тенденції в даних, які допоможуть вам сформувати наступну теорію?
- Якщо ні, чи це через проблему з якістю даних?

Помилкові висновки через неправильні або «брудні» дані можуть стати причиною неправильної бізнес-стратегії та прийняття рішень. Хибні висновки можуть призвести до незручного моменту на звітній зустрічі, коли ви зрозумієте, що ваші дані не витримують перевірки. Перш ніж потрапити туди, важливо створити культуру якісних даних у організації. Для цього слід задокументувати інструменти, які ви можете використовувати для створення цієї культури, і те, що для вас означає якість даних.

Перетворення даних — це процес зміни формату, структури або значень даних. Для проектів аналітики даних дані можуть бути перетворені на двох етапах конвеєра даних. Організації, які використовують локальні сховища даних, зазвичай використовують процес ETL (вилучення, перетворення, завантаження), у якому перетворення даних є середнім кроком. Сьогодні більшість організацій використовують хмарні сховища даних, які можуть масштабувати обчислювальні ресурси та ресурси зберігання з затримкою, яка вимірюється в секундах або хвилинах. Масштабованість хмарної платформи дозволяє організаціям пропускати попередні завантаження перетворень і завантажувати необроблені дані в сховище даних, а потім перетворювати їх під час запиту — модель під назвою ELT (вилучення, завантаження, перетворення).

Такі процеси, як інтеграція даних, міграція даних, сховище даних і суперечка даних, можуть включати перетворення даних.

Перетворення даних може бути конструктивним (додавання, копіювання та реплікація даних), деструктивним (видалення полів і записів), естетичним (стандартизація привітань або назв вулиць) або структурним (перейменування, переміщення та об'єднання стовпців у базі даних).

Трансформація даних може підвищити ефективність аналітичних і бізнес-процесів і забезпечити краще прийняття рішень на основі даних. Перший етап перетворення даних має включати такі речі, як перетворення типів даних і зведення ієрархічних даних. Ці операції формують дані для підвищення сумісності з аналітичними системами. Аналітики даних і дослідники даних можуть впроваджувати подальші адитивні перетворення, якщо це необхідно, як окремі рівні обробки. Кожен рівень обробки повинен бути розроблений для виконання певного набору завдань, які відповідають відомим діловим або технічним вимогам.

Трансформація даних часто полягає в тому, щоб скоротити дані та зробити їх більш керованими. Дані можна консолідувати шляхом фільтрації непотрібних полів, стовпців і записів. Пропущені дані можуть включати числові індекси в даних, призначених для графіків і інформаційних панелей, або записи з бізнес-регіонів, які не представляють інтересу для конкретного дослідження.

Дані також можуть бути агреговані або узагальнені шляхом, наприклад, перетворення часового ряду транзакцій клієнтів на погодинні або щоденні підрахунки продажів.

Інструменти BI можуть виконувати цю фільтрацію та агрегацію, але ефективніше виконувати перетворення до того, як інструмент звітування отримає доступ до даних.

Метод скорочення даних може досягти стислого опису вихідних даних, який є набагато меншим за кількістю, але зберігає якість вихідних даних.

Методи зменшення даних:

1. Агрегація кубів даних: ця техніка використовується для агрегування даних у простішій формі. Наприклад, уявімо, що інформація, яка є зібраною для аналізу за 2012–2014 роки, включає дохід компанії за кожні три місяці. Вони включають інформацію за річний обсяг продажів, а не в середньому за квартал. Тож ми можемо узагальнити дані таким чином, щоб отримані дані підсумовували загальний обсяг продажів за рік, а не за квартал. Агрегація узагальнює дані.

2. Зменшення розмірності: щоразу, коли ми стикаємося з не надто важливими даними, ми використовуємо атрибут, необхідний для нашого аналізу. Це зменшує розмір даних, оскільки усуває застарілі або зайві функції.

Поетапний прямий вибір – вибір починається з порожнього набору атрибутів, а потім вибираються найкращі з оригінальних атрибутів у наборі на основі їх відповідності іншим атрибутам. Ми знаємо це як p-значення в статистиці.

Поетапний зворотний вибір – цей вибір починається з набору повних атрибутів у вихідних даних і в кожній точці усуває найгірший атрибут, що залишився в наборі.

Поєднання переадресації та зворотного вибору – це дозволяє нам видалити найгірші та вибрати найкращі атрибути, заощаджуючи час і пришвидшуючи процес.

3. Стиснення даних: Техніка стиснення даних зменшує розмір файлів за допомогою різних механізмів кодування (кодування Хаффмана та кодування довжини). Ми можемо розділити його на два типи на основі їхніх методів стиснення.

Стиснення без втрат – методи кодування (кодування довжини серії) дозволяють просте та мінімальне зменшення розміру даних. Стиснення даних без втрат використовує алгоритми для відновлення точних вихідних даних зі стиснутих даних.

Стиснення з втратами – прикладами цього стиснення є такі методи, як техніка дискретного вейвлет-перетворення, PCA (аналіз головних компонентів). Наприклад, формат зображення JPEG є стисненням із втратами, але ми можемо знайти значення, еквівалентне оригінальному зображенню. При стисненні даних із втратами даних розпаковані дані можуть відрізнятися від вихідних даних, але є достатньо корисними для отримання інформації з них.

4. Зменшення кількості: у цій техніці зменшення фактичні дані замінюються математичними моделями або меншим представленням даних замість фактичних даних, важливо зберігати лише параметр моделі. Або непараметричний метод, такий як кластеризація, гістограма, вибірка.

5. Дискретизація та операція ієрархії концепції: Техніки дискретизації даних використовуються для поділу атрибутів безперервної природи на дані з інтервалами. Ми замінюємо багато постійних значень атрибутів мітками з малими інтервалами. Це означає, що результати відображаються в стислій та зрозумілій формі.

Низхідна дискретизація – якщо ви спочатку розглядаєте одну або кілька точок (так звані точки розриву або точки розділення), щоб розділити весь набір атрибутів і повторюєте цей метод до кінця, тоді процес відомий як низхідний. дискретизація, також відома як розщеплення.

Дискретизація знизу вгору. Якщо спочатку розглядати всі постійні значення як точки розділення, деякі відкидаються через комбінацію сусідніх значень в інтервалі, цей процес називається дискретизацією знизу вгору.

Ієрархії понять: він зменшує розмір даних шляхом збору та заміни понять низького рівня (наприклад, 43 для віку) на поняття високого рівня (категоріальні змінні, такі як середній вік або старший).

Для числових даних можна дотримуватися наступних методів:

Бнінг – це процес зміни числових змінних на категоричні відповідники. Кількість категорійних аналогів залежить від кількості бункерів, вказаних користувачем.

Аналіз гістограми. Подібно до процесу групування, гістограма використовується для поділу значення атрибута X на непересічні діапазони, які називаються дужками. Існує кілька правил поділу:

Рівночастотний розподіл: розподіл значень на основі їх кількості входжень у наборі даних.

Поділ рівної ширини: Розбиття значень у фіксований проміжок на основі кількості бункерів, тобто набору значень у діапазоні від 0 до 20.

Кластеризація: групування схожих даних.

Отже, попередня обробка даних у машинному навчанні — це важливий крок, який допомагає підвищити якість даних, щоб сприяти вилученню з них значущої думки. Попередня обробка даних у машинному навчанні стосується техніки підготовки (очищення та організації) необроблених даних, щоб зробити їх придатними для побудови та навчання моделей машинного навчання. Простими словами, попередня обробка даних у машинному навчанні — це техніка інтелектуального аналізу даних, яка перетворює необроблені дані в зрозумілий і читабельний формат.

2.2 Моделі кластеризації на основі машинного навчання

Кластеризація є одним із найпопулярніших методів у науці про дані та є неконтрольованою технікою машинного навчання, яка дозволяє нам знаходити структури в наших даних, не намагаючись отримати конкретну інформацію. У кластерному аналізі ми розбиваємо набір даних на групи, які мають подібні атрибути.

Кластеризація - це завдання навчитися маркувати дані за допомогою набору даних без тегів [37]. Оскільки набір даних не містить міток, оцінити оптимальність отриманої моделі набагато складніше, ніж при контрольованому навчанні.

Існує безліч алгоритмів кластеризації, і на жаль, складно сказати, який з них буде найкращим для того чи іншого набору даних. Зазвичай якість кожного алгоритму залежить від невідомих властивостей розподілу, з якого був виведений набір даних. Розглянемо найбільш корисні і широко використовувані алгоритми кластеризації.

Метою кластерного аналізу (також відомого як класифікація) є побудова груп (або класів чи кластерів), забезпечуючи при цьому наступну властивість: у межах групи спостереження мають бути якомога подібнішими, тоді як спостереження, що належать до різних груп, мають бути максимально різними.

Усі методи кластеризації спрямовані на групування спостережень у підгрупи/кластери таким чином, щоб [38]:

- Кожен кластер містить спостереження, подібні одне до одного. Кластер повинен бути однорідним за певними ознаками кластеризації.
- Спостереження одного кластера відрізняються від спостережень інших кластерів. Кожен кластер відрізняється від інших груп такими ж характеристиками.

До загальних кроків кластерного аналізу та аспектів, які важливо враховувати незалежно від того, який метод обрано, потрібно віднести:

1. Визначення міри подібності/відстані і масштабування;
2. Визначення типу методу кластеризації: ієрархічний або неієрархічний;
3. Визначення кількості кластерів.

Міра відстані/подібності, а також масштабування даних мають великий вплив на результат. Отже, залежно від мети дослідження та типу даних, нам потрібно розглянути, як виміряти подібність і як масштабувати дані.

- Евклідова відстань: квадратний корінь із суми квадратів різниці, це найпоширеніше. Manhattan, Maximum і Minkowski – це інші показники, які можна розглянути.
- Відстань на основі кореляції: вимірювання кореляції підкреслить форму спостережень, а не їх величину. Два спостереження, характеристики яких сильно корельовані, вважатимуться подібними, навіть якщо вони знаходяться на великій відстані одне від одного, виміряної евклідовою відстанню.
- Масштабування змінних: потрібно вирішити, чи стандартне значення/дисперсія має бути 1 чи ні. Вибір залежить від того, за яким шаблоном/подібністю спостережень потрібно робити групування. Якщо необхідно надати кожній функції однакову важливість, то слід зробити std/var рівним 1. Можна масштабувати дані, щоб мати std/var 1, якщо функції вимірюються в різних масштабах.

Визначення подібності має бути визначено для конкретного завдання кластеризації та, як правило, вимагає певних предметних знань про дані. Те, що означає подібність двох спостережень, також залежать від мети дослідження.

Існує два основних види класифікації:

– k-means кластеризація

– Ієрархічна кластеризація

Перший вид зазвичай використовується, коли кількість класів фіксована заздалегідь, тоді як другий зазвичай використовується для невідомої кількості класів і допомагає визначити цю оптимальну кількість. Обидва методи проілюстровано нижче за прикладними програмами вручну та в R. Потрібно зауважити, що для ієрархічної кластеризації у даному випадку представлено лише висхідну класифікацію.

Ієрархічна кластеризація — це стратегія, спрямована на побудову ієрархії кластерів із встановленим порядком зверху вниз [39]. K-means не підпадає під цю категорію, оскільки він не виводить кластери в ієрархії, тож давайте отримаємо уявлення про те, що потрібно отримати в кінцевому результаті під час запуску одного з цих алгоритмів. «Ієрархія кластерів» зазвичай представляється дендрограмою, показаною нижче (рис. 2.1).

Рис.2.1 - Приклад дендрограми

Наприклад, дендрограма, яка описує масштаби географічного розташування, може мати назву країни вгорі, потім вона може вказувати на її

регіони, які потім вказуватимуть на їхні штати/провінції, потім округи чи райони тощо.

Щоб створити дендрограму, потрібно обчислити подібності між атрибутами. Для обчислення подібності двох ознак зазвичай використовуємо Манхеттенську відстань або Евклідову відстань. Ці відстані будуть записані в так звану матрицю близькості, яка містить відстані між кожною точкою. Далі використовують ці комірки, щоб знайти пари точок із найменшою відстанню та почати зв’язувати їх разом, щоб створити дендрограму.

Неієрархічна кластеризація вимагає, щоб початковий розділ/кількість кластерів були відомі апріорі. Потрібно розділити точки даних на k кластерів, щоб варіація всередині кластера для всіх кластерів була якомога меншою. Існують різні неієрархічні методи кластеризації, більшість з яких відрізнятиметься головним чином за:

- Метод, який використовується для встановлення початкових центроїдів кластера
- Правило для повторного призначення спостережень у кожній ітерації

Найпоширенішим методом неієрархічній кластеризації є k-means. До кроків кластеризації k-means належать [39]:

Крок 1. Визначення k кількості кластерів.

Крок 2. Визначення ініціалізації шляхом випадкового призначення кожного спостереження одному з k кластерів;

Крок 3. Проведення повторювання наступні два кроки, доки призначення кластера не перестане змінюватися:

Крок. 3.1. Обчислення центроїд для кожного кластера

Крок. 3.2. Призначення/повторне призначення кожного спостереження кластеру з найближчим центроїдом.

Крок 4. Вивід результату розподілу даних за кластерами (рис.2.2).

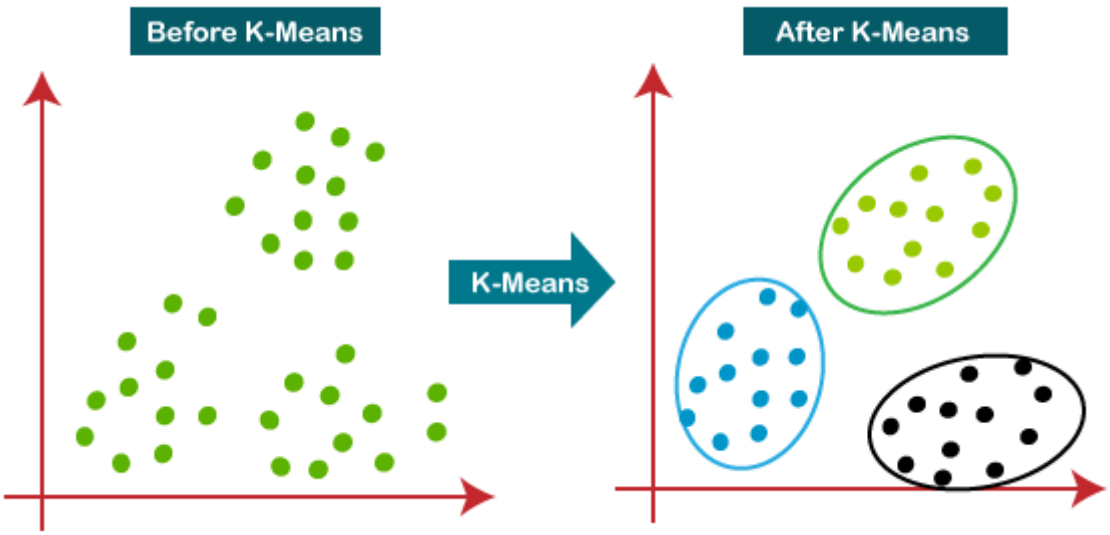


Рис.2.2 – Приклад застосування k-means

Щоб вирішити, скільки кластерів використовувати, можна використати наступні два методи.

– Ліктьова діаграма: побудова в межах суми квадратів кожного кластера для різної кількості кластерів. Квадрати внутрішньої суми будуть зменшуватися з кожним додатковим кластером, але ми будемо шукати лікоть, де квадрати внутрішньої суми зменшуються найбільше порівняно з попередньою кількістю кластерів, а потім вирівнюються для додаткових кластерів.
– Аналіз силуету: показує, наскільки добре кожне спостереження віднесено до кластера.

Далі наведемо математичний опис кластеризації методом k-means.

Цільова функція:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

де $w_{ik}=1$ для точки даних xi, якщо вона належить кластеру *k*; інакше $w_{ik}=0$. Крім того, μ_k є центроїдом кластера x_i.

Це проблема мінімізації з двох частин. Спочатку ми мінімізуємо J відносно w_{ik} і розглядаємо μ_k як фіксований. Потім ми мінімізуємо J відносно μ_k і розглядаємо w_{ik} як фіксований. Технічно кажучи, ми спочатку розрізняємо J відносно w_{ik} і оновлюємо призначення кластерів. Потім ми диференціюємо J відносно μ_k і повторно обчислюємо центроїди після призначення кластеру з попереднього кроку. Отже:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Іншими словами, призначте точку даних x_i найближчому кластеру за сумою квадратів відстані від центроїда кластера.

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}} \quad (3)$$

Що означає переобчислення центроїда кожного кластера для відображення нових призначень.

Оскільки алгоритми кластеризації, включаючи k-means, використовують вимірювання на основі відстані для визначення подібності між точками даних, рекомендується стандартизувати дані, щоб мати середнє значення нуль і стандартне відхилення одиниці, оскільки майже завжди об’єкти в будь-якому наборі даних матимуть різні одиниці вимірювання, наприклад, вік проти доходу.

Враховуючи ітераційну природу k-means і випадкову ініціалізацію центроїдів на початку алгоритму, різні ініціалізації можуть призвести до різних кластерів, оскільки алгоритм k-means може застрягти в локальному оптимумі та не збігатися з глобальним оптимумом. Тому рекомендується

запускати алгоритм, використовуючи різні ініціалізації центроїдів, і вибирати результати виконання, які давали б меншу суму квадратів відстані.

Призначення прикладів не змінюється – це те саме, що відсутність змін у варіації всередині кластера:

$$\frac{1}{m_k}\sum_{i=1}^{m_k}\|x^i - \mu_{c^k}\|^2 \qquad (4)$$

Існує також різниця між цими двома методами кластеризації. Однакова кількість кластерів з використанням різних методів, ієрархічних і k-середніх, дає дуже різні результати. Наприклад, чотири кластери з k-середніми дуже відрізняються від чотирьох кластерів, які використовують ієрархічну кластеризацію.

У обох є переваги та недоліки, і вибір методу залежить від мети дослідження. Можна мати на увазі, що їх можна розглядати як додаткові методи, де ієрархічну кластеризацію можна використовувати в дослідницькому сенсі перед тим, як продовжувати до неієрархічної кластеризації. Для задачі поставленої в параграфі 1.4 краще обрати метод неієрархічної кластеризації, оскільки даний тип кластеризації, використовують вимірювання на основі відстані для визначення подібності між точками даних, що краще різнотипних даних.

2.3 Інтелектуальний метод формування споживчого кошику

Для оперативної зміни набору товарів у споживчому кошику розроблено інтелектуальний метод формування споживчого кошику. Запропонований метод ілюструється схематично (Fig.1) та представлений наступними кроками:

Крок 1. Збір даних з супермаркетів (Блок 1), а саме інформацію про покупця (ID Покупця, Вік, Доходи) та про його покупки (ID Транзакції, Дата, Перелік товарів, Вартість товару). При зборі даних, можна розширити базу даними з аптек, продовольчих та непродовольчих товарів.

Крок 2. Кластеризація даних.

Крок 2.1. Вибір об'єктів для кластеризації (Блок 2).

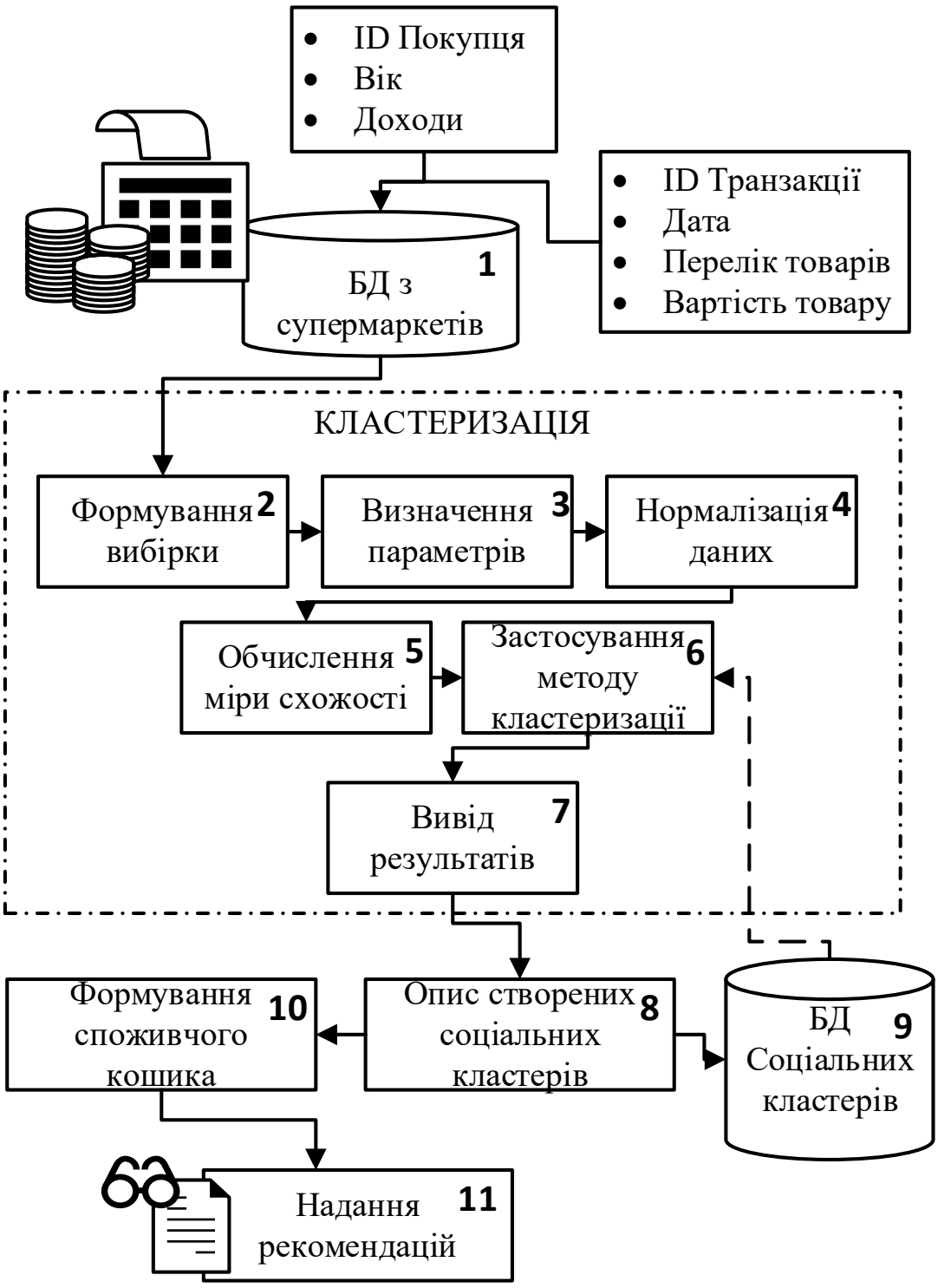
Крок 2.2. Визначення безлічі змінних (Блок 3), якими будуть оцінюватися об'єкти у вибірці.

Крок 2.3. Нормалізація (Блок 4) значень змінних.

Крок 2.4. Обчислення значень міри схожості між об'єктами (Блок 5).

Крок 2.5. Застосування методу кластерного аналізу (Блок 6) створення груп подібних об'єктів (кластерів).

Крок 2.6. Подання результатів аналізу (Блок 7).



Фіг.1. Структура інтелектуального методу формування споживчого кошику

Крок 3. Проведення опису отриманих кластерів, а саме визначення соціального кластеру (Блок 8). Також, передача сформованих соціальних

кластерів в окрему базу даних (Блок 9), що дасть можливість в подальшому максимально автоматизувати процес обробки та формування кластерів. Також, в подальших наукових дослідженнях на основі отриманих соціальних кластерів та методів машинного навчання буде розроблено метод передбачення змін у формуванні споживчого кошика.

Крок 4. Формування споживчого кошика (Блок 10) та надання рекомендацій (Блок 11).

Розроблений метод визначення споживчого кошику на основі машинного навчання та даних мереж супермаркетів дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

2.4 Інформаційне забезпечення розробленого методу

У якості вхідних даних використано дані Customer Personality Analysis з платформи Kaggle [18]. В цьому наборі даних представлено аналіз особистості клієнта. Дані представлені, як суми витрачених Дол.США за 2 роки.

Структура даних [18]:

- ID: унікальний ідентифікатор клієнта
- Year_Birth: рік народження клієнта
- Education: Рівень освіти замовника
- Marital_Status: сімейний стан клієнта
- Income: річний дохід домогосподарства клієнта
- Kidhome: кількість дітей у родині клієнта
- Teenhome: кількість підлітків у родині клієнта
- Dt_Customer: Дата реєстрації клієнта в компанії

- Recency: кількість днів з моменту останньої покупки клієнта
- Complain: 1, якщо клієнт скаржився протягом останніх 2 років, 0 в іншому випадку
- MntWines: сума, витрачена на вино за останні 2 роки
- MntFruits: сума, витрачена на фрукти за останні 2 роки
- MntMeatProducts: сума, витрачена на м’ясо за останні 2 роки
- MntFishProducts: сума, витрачена на рибу за останні 2 роки
- MntSweetProducts: сума, витрачена на солодощі за останні 2 роки
- MntGoldProds: сума, витрачена на золото за останні 2 роки
- NumDealsPurchases: кількість покупок зі знижкою
- AcceptedCmp1: 1, якщо клієнт прийняв пропозицію в 1-й кампанії, 0 в іншому випадку
- AcceptedCmp2: 1, якщо клієнт прийняв пропозицію в 2-й кампанії, 0 в іншому випадку
- AcceptedCmp3: 1, якщо клієнт прийняв пропозицію в 3-й кампанії, 0 в іншому випадку
- AcceptedCmp4: 1, якщо клієнт прийняв пропозицію в 4-й кампанії, 0 в іншому випадку
- AcceptedCmp5: 1, якщо клієнт прийняв пропозицію в 5-й кампанії, 0 в іншому випадку
- Response: 1, якщо клієнт прийняв пропозицію в останній кампанії, 0 інакше
- NumWebPurchases: кількість покупок, зроблених через веб-сайт компанії
- NumCatalogPurchases: кількість покупок, зроблених за допомогою каталогу

- NumStorePurchases: кількість покупок, зроблених безпосередньо в магазинах
- NumWebVisitsMonth: кількість відвідувань веб-сайту компанії за останній місяць

Структура даних, які було обрано з набору даних представлена у таблиці 1. Цих даних не достатньо для повноцінного розв’язання задачі. Отримання повного потрібного Dataset є важким процесом, адже тут потрібно дозвіл клієнтів на використання персональних даних. Вважливо, мати повний список куплених товарів, покупцем, щоб можна було сформувати більш детальніший список. Але, для реалізації розробленого методу, достатньо Dataset Customer Personality Analysis.

Таблиця 2.1 - Структура Dataset

Параметр	**count**	**mean**	**min**	**max**
Income	2240.0	52247.25	1730.0	666666.0
Kidhome	2240.0	0.44	0.0	2.0
Teenhome	2240.0	0.51	0.0	2.0
Recency	2240.0	49.11	0.0	99.0
Wines	2240.0	303.94	0.0	1493.0
Fruits	2240.0	26.30	0.0	199.0
Meat	2240.0	166.95	0.0	1725.0
Fish	2240.0	37.52	0.0	259.0
Age	2240.0	52.19	25.0	128.0
Total_Spend	2240.0	602.25	5.0	2525.0

Вибрані дані Customer Personality Analysis з платформи Kaggle представляють аналіз особистості клієнта. Цей набір даних дозволить провести практичну реалізацію розробленого методу, проте його не достатньо для повноцінного розв’язання задачі.

Висновки до розділу 2

1. Визначено, що попередня обробка даних у машинному навчанні — це техніка інтелектуального аналізу даних, яка перетворює необроблені дані в зрозумілий і читабельний формат.

2. Визначено переваги та недоліки ієрархічної кластеризації та неієрархічної кластеризації. Для задачі поставленої в параграфі 1.4 краще обрати метод неієрархічної кластеризаці, оскільки даний тип кластеризації, використовують вимірювання на основі відстані для визначення подібності між точками даних, що краще різнотипних даних.

3. Розроблено метод визначення споживчого кошику на основі машинного навчання та даних мереж супермаркетів, який дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

4. Вибрано дані Customer Personality Analysis з платформи Kaggle, які представляють аналіз особистості клієнта. Цей набір даних дозволить провести практичну реалізацію розробленого методу, проте його не достатньо для повноцінного розв’язання задачі.

3 РЕАЛІЗАЦІЯ ІНТЕЛЕКТУАЛЬНОГО МЕТОДУ ФОРМУВАННЯ СПОЖИВЧОГО КОШИКУ

3.1 Попередня обробка та дослідження даних

Спершу, щоб зрозуміти з яким набором даних працюємо, потрібно оцінити його структуру:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   2240 non-null   int64  
 1   Year_Birth           2240 non-null   int64  
 2   Education            2240 non-null   object 
 3   Marital_Status       2240 non-null   object 
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64  
 6   Teenhome             2240 non-null   int64  
 7   Dt_Customer          2240 non-null   object 
 8   Recency              2240 non-null   int64  
 9   MntWines             2240 non-null   int64  
 10  MntFruits            2240 non-null   int64  
 11  MntMeatProducts      2240 non-null   int64  
 12  MntFishProducts      2240 non-null   int64  
 13  MntSweetProducts     2240 non-null   int64  
 14  MntGoldProds         2240 non-null   int64  
 15  NumDealsPurchases    2240 non-null   int64  
 16  NumWebPurchases      2240 non-null   int64  
 17  NumCatalogPurchases  2240 non-null   int64  
 18  NumStorePurchases    2240 non-null   int64  
 19  NumWebVisitsMonth    2240 non-null   int64  
 20  AcceptedCmp3         2240 non-null   int64  
 21  AcceptedCmp4         2240 non-null   int64  
 22  AcceptedCmp5         2240 non-null   int64  
 23  AcceptedCmp1         2240 non-null   int64  
 24  AcceptedCmp2         2240 non-null   int64  
 25  Complain             2240 non-null   int64  
 26  Z_CostContact        2240 non-null   int64  
 27  Z_Revenue            2240 non-null   int64  
 28  Response             2240 non-null   int64  
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

Далі перевіряємо, чи в наборі даних є null-значення. У значені " Income " є 24 значення NaN (див. Додаток А). Це може негативно вплинути на процеси аналізу та кластеризації клієнтів за їхніми доходами. Щоб уникнути цього, ми повинні обробляти значення NaN і заповнювати їх середнім значенням доходів.

Потім слід з'ясувати і видалити всі дублікати рядків в наборі даних. Можна припустити, що рядки, які містять однакові значення ознак "year_birth", "освіта", "marital_status", "дохід", "kidhome", "teenhome", "dt_customer", є дублікатами (див. Додаток Б).

Спробуємо з'ясувати винятки в особливостях набору даних, які можуть погано позначитися на результаті аналізу. Є кілька значень, які суттєво відрізняються від інших значень. Це може вплинути на результати. Тому ці рядки слід видалити (рис.3.1).

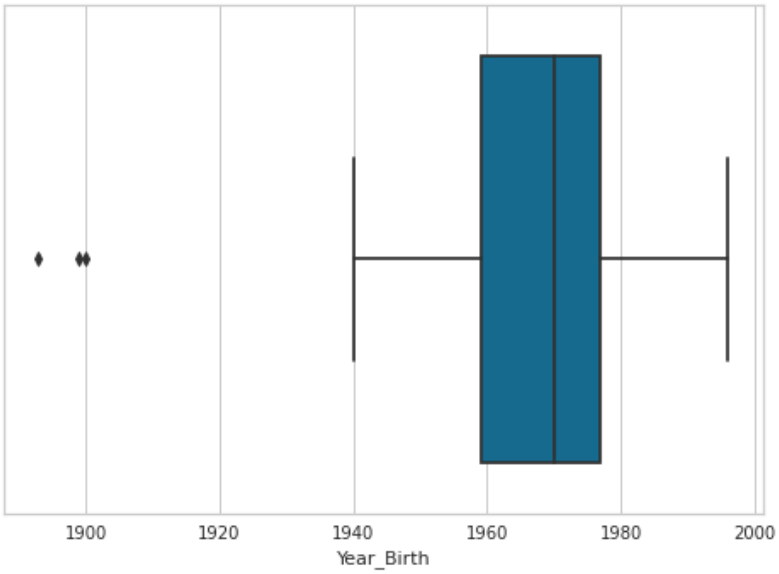


Рис.3.1 - Боксплот діаграма

Далі видаляємо значення "YOLO", "Absurd" та "Alone" як вихідні дані. Слід видалити Revenue outliers - рядки з сумою більше 130 000. З результатів видно (Додаток Б), що функції "Z_CostContact" і "Z_Revenue" мають однакові значення для всіх рядків в наборі даних. Таким чином, ми могли б позначити його як зайві функції та відкинути.

Далі проведемо попередню обробку даних. Спочатку слід перетворити категоріальні ознаки в числові.

```
columns_to_transform = ['Education','Marital_Status']

LE=LabelEncoder()
for i in columns_to_transform:
    df[i]=df[[i]].apply(LE.fit_transform)
```

Потім потрібно перетворити функцію «Dt_Customer» на кількість днів, протягом яких людина є клієнтом магазину.

```
df["Customer_For"] = pd.to_datetime("now") - df["Dt_Customer"]
df['Customer_For'] = pd.to_numeric(df['Customer_For'].dt.days, downcast='integer
')
df = df.drop("Dt_Customer", axis=1)
```

Щоб підрахувати всю суму, яку витратив кожен клієнт, і кількість покупок за весь час створимо функції «Total_Spent» і «Total_Number_Purchases».

```
df['Total_Spent'] = df['MntWines'] + df['MntFruits'] + df['MntMeatProducts'] + d
f['MntFishProducts'] + df['MntSweetProducts'] + df['MntGoldProds']
df['Total_Number_Purchases'] = df['NumWebPurchases'] + df['NumStorePurchases']
```

Оскільки вони відіграють значну роль у витратах, проведемо перевірку вікових та прибуткових стовпців (Рис.3.2). Більшість наших клієнтів 35+ люди з дітьми мають більш високі доходи і витрати. У нас є винятки за віком і доходами, і тому треба це проаналізувати перед кластеризацією.

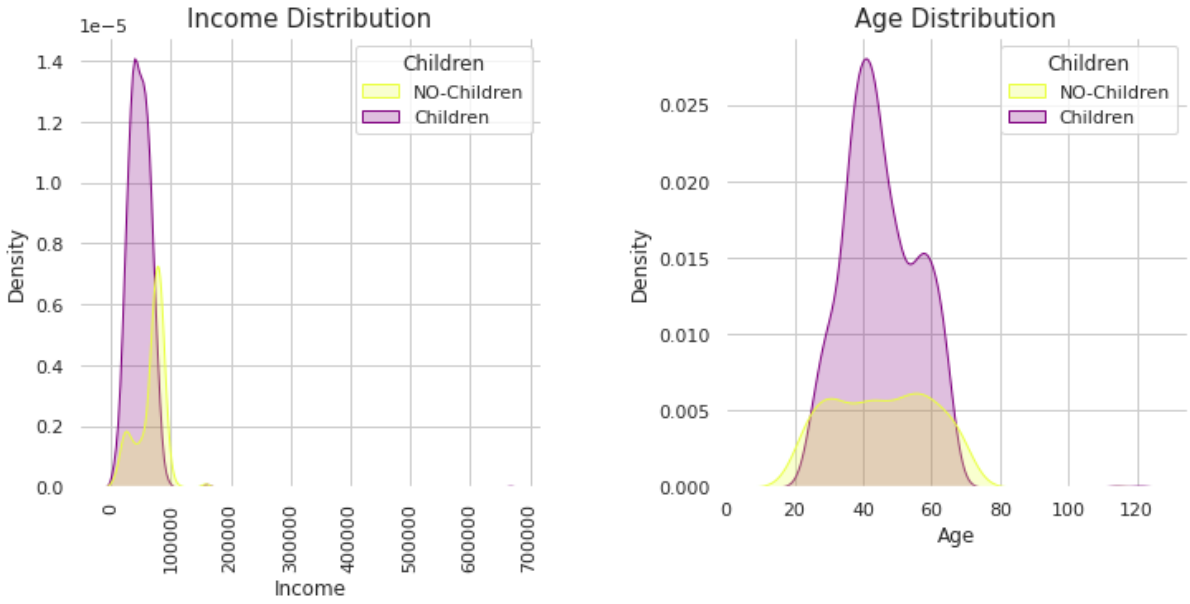


Рис.3.2 - Перевірка вікових та прибуткових стовпців

З рисунка 3.3 видно, що хоча кількість клієнтів, які мають дітей, набагато більша, більшість грошей надходить від клієнтів, які не мають дітей.

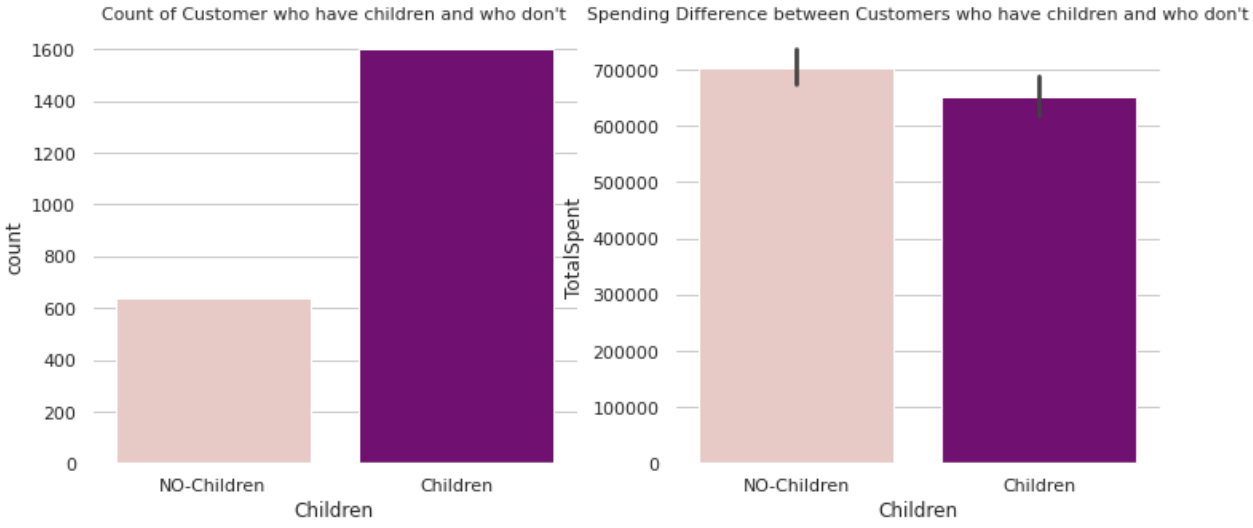


Рис.3.3 - Різниця у витратах між клієнтами, які мають дітей, і такими, які їх не мають

Середні витрати та вартість витрат вищі для клієнтів (Рис.3.4), які мають ступінь доктора філософії або магістра та не мають дітей, а витрати

клієнтів, які не мають дітей, є доволі значними. Парадоксально, що чим більше у людини дітей, тим менша вірогідність витрачання коштів. Дохід і діти мають негативну кореляцію: людина з меншою кількістю дітей, як правило, заробляє більше грошей, і навпаки.

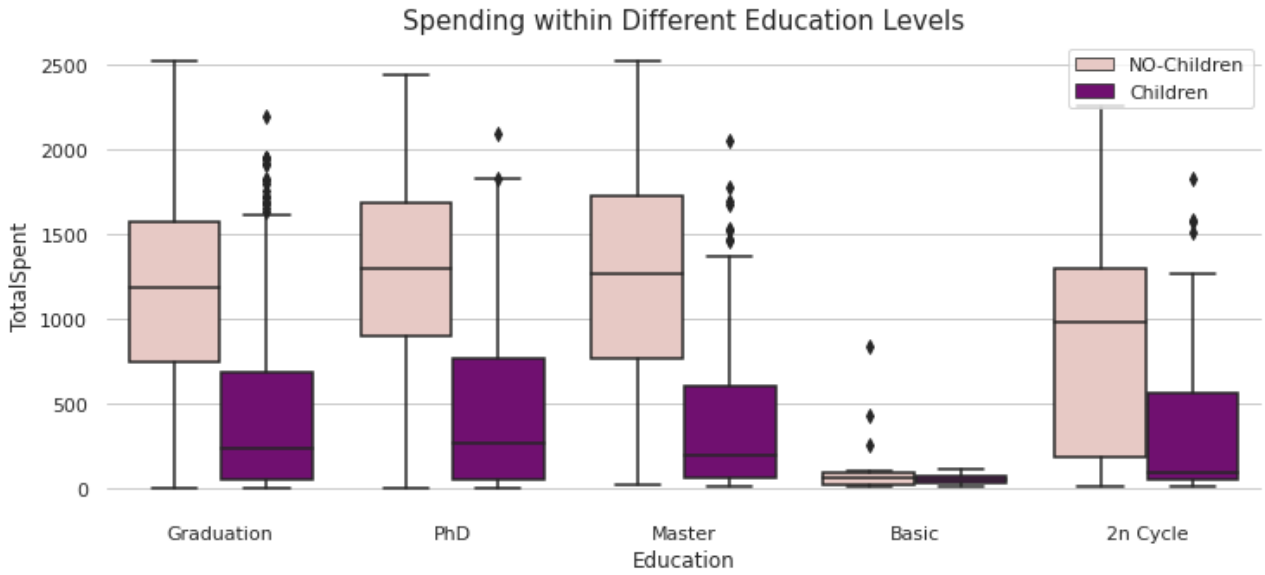


Рис.3.4 - Витрати на різних рівнях освіти

У вина найвищий (Рис.3.5) відсоток продажів (майже половина), за ним йдуть м'ясні продукти, і тоді відсотки продажів по решті продукції практично однакові.

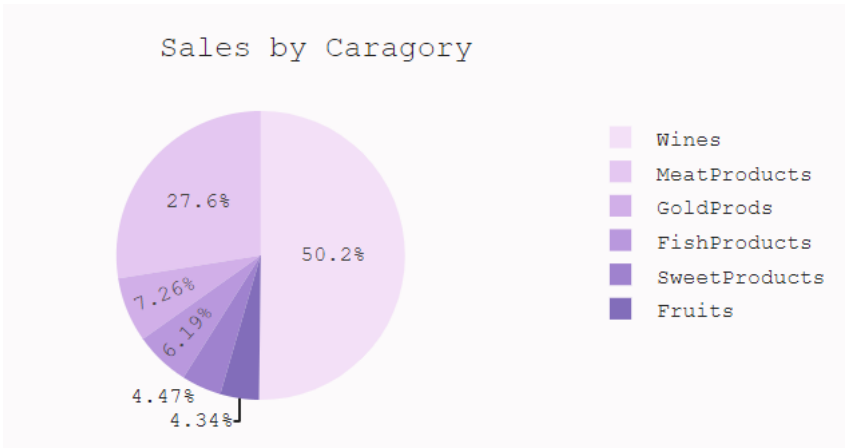


Рис.3.5 – Продаж по категоріях

Більшість продажів надходить з магазину (Рис.3.6), тому потрібно вдосконалювати використання веб-додатку.

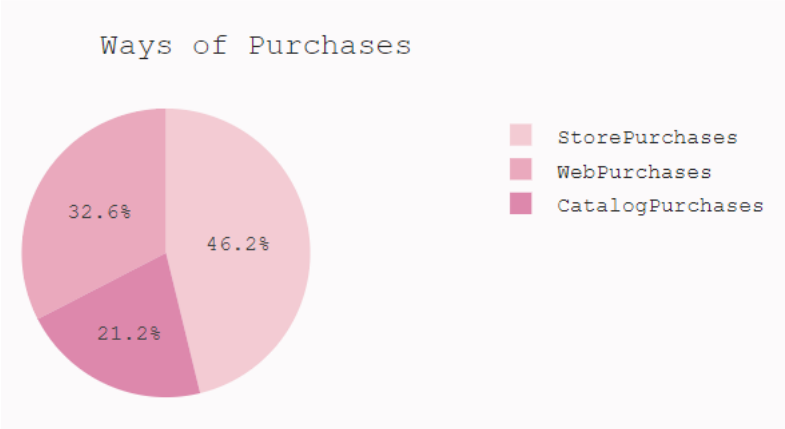


Рис.3.6 – Способи купівлі

16% клієнтів цієї компанії заробляють менше 30 000 дол.США. Середня заробітна плата становить 51 373 дол.США. Давайте подивимося, скільки вони витрачають в залежності від зарплати.

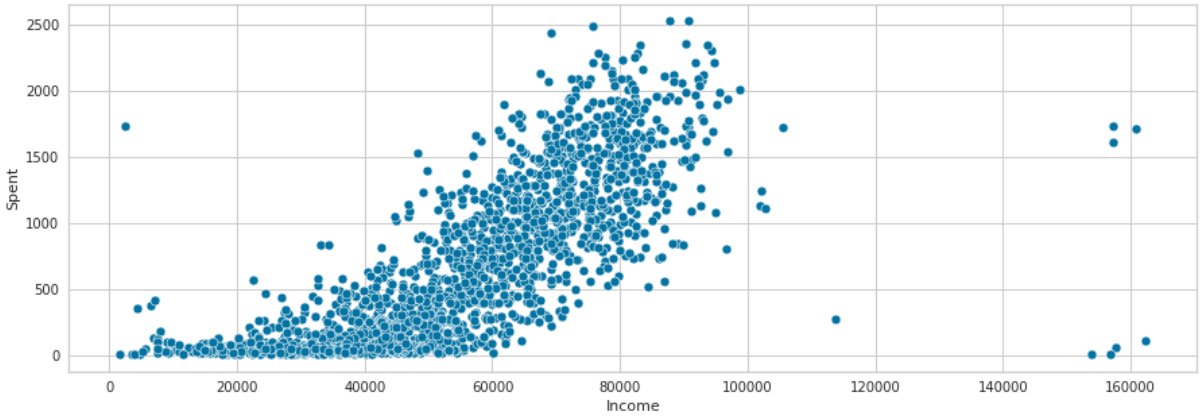


Рис.3.7 – Витрати відносно зарплати

Щоб побачити кореляцію між різними ознаками в наборі даних, створимо теплову карту.

```
plt.figure(figsize=(18,12))
dataplot = sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)
plt.show()
```

З теплової карти видно (Рис.3.8), що функції доходу, Total_Spent та Total_Number_Purchases суттєво впливають один на одного. Також є кореляція між обсягами різних видів продукції та кількістю покупок. Ще один цікавий факт, який могли помітити, полягає в тому, що кількість відвідувань веб-сайтів і кількість веб-покупок взагалі не корелюють. Це означає, що ефективність онлайн-продажів може бути на низькому рівні.

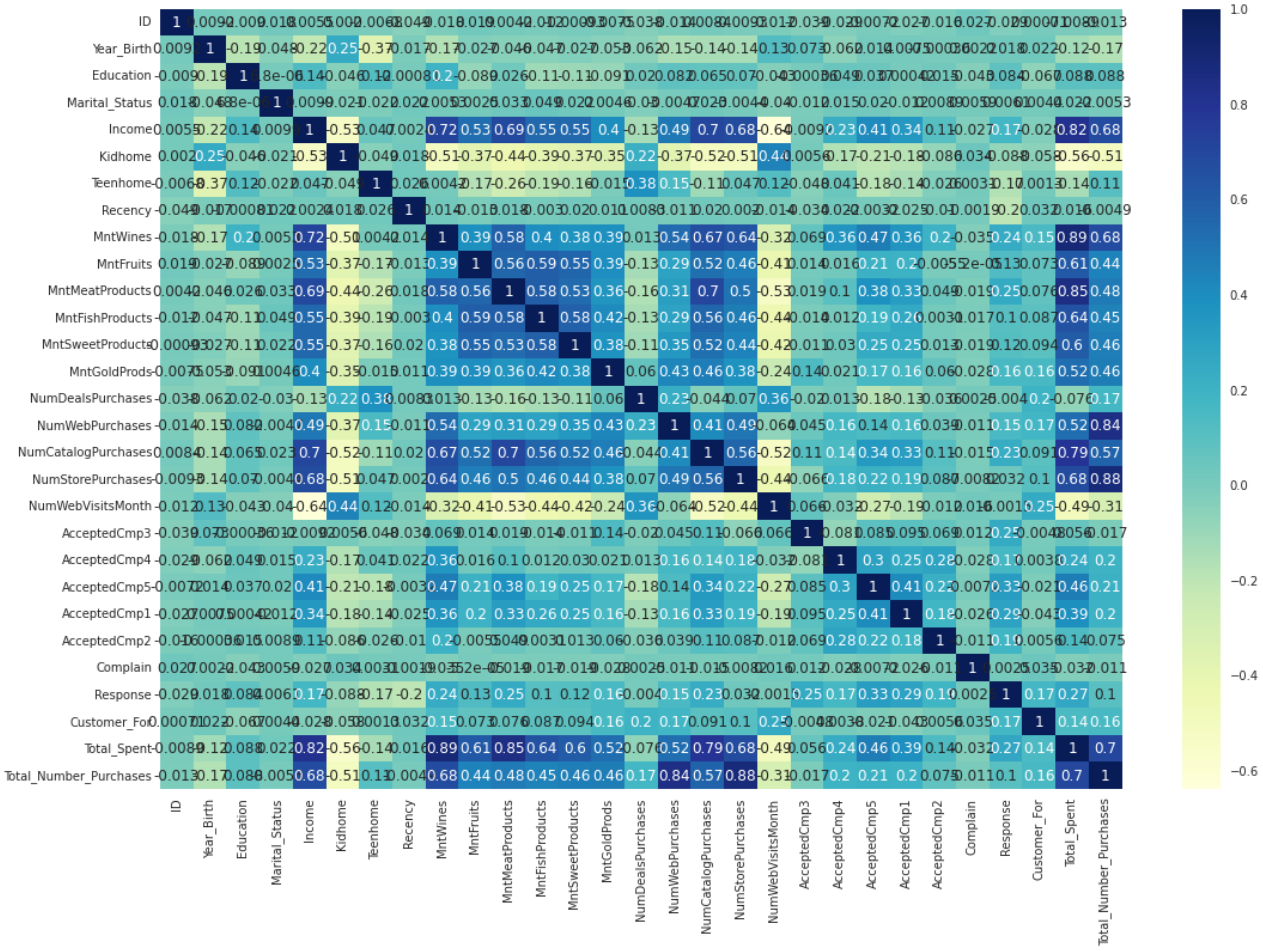


Рис.3.8 - Теплова кореляційна карта

Перш ніж проводити кластеризації, потрібно розробити нормалізацію. Також нам цікаво знати витрати в місяць, тому всі значення по витратах розділимо на 24 місяці. Це дозволить отримати значення готові для формування споживчого кошика.

Алгоритми кластеризації складно розраховуються, коли дані мають багато функцій або коли функції мають багатолінійність. Аналіз основних компонентів (PCA) який дозволяє виправити обидві ці проблеми дуже просто. Для створення одиновимірного масиву для кожної функції, а потім перевірити, які з них пояснюють коваріацію до 90%. Можна вибрати і інший поріг, але було вибрано цей після численних експериментів, тому що результати на етапах кластеризації задовольняли.

Виходячи з результатів PCA (рис.3.9), будемо використовувати 13 основних компонентів (табл.3.1) для кластеризації та 3 для візуалізації результатів (табл.3.2). Таким чином, вдалося позбутися 7 вимірів.

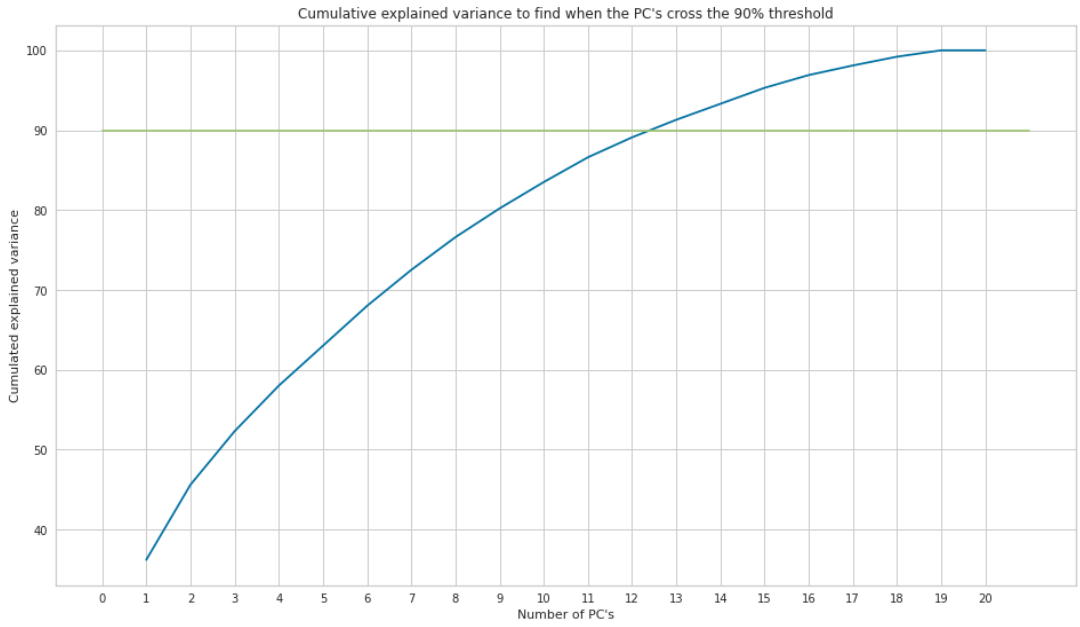


Рис.3.9 - Результати PCA

Таблиця 3.1 - 13 основних компонентів для кластеризації

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
0	4.153752	0.988888	1.806138	-1.165783	0.838833	1.151661	0.348525	-1.173495	0.162600	1.177080	-0.487191	2.191703	0.023004
1	-2.578687	-0.752018	-1.852939	-0.512348	0.294830	1.507821	0.579700	0.372401	0.189037	0.831787	-0.242696	-0.095576	0.003208
2	1.842248	-0.188573	-0.103544	-0.496129	-1.211876	-0.238788	0.438730	-0.547337	-0.707681	-1.690523	-0.088198	0.117368	-0.398350
3	-2.586428	-1.201413	-0.173890	0.160926	-1.227816	-0.430787	-0.886290	0.730247	-0.149161	-0.360461	0.104352	0.367782	-0.100707
4	-0.260048	0.113205	-0.416280	-0.868658	0.652997	-1.375168	-1.033978	1.796776					

Таблиця 3.2 - 3 основних компонентів для візуалізації результатів

	PC1	PC2	PC3
0	4.153752	0.988888	1.806138
1	-2.578687	-0.752018	-1.852939
2	1.842248	-0.188573	-0.103544
3	-2.586428	-1.201413	-0.173890
4	-0.260048	0.113205	-0.416280

3.2 Програмна реалізація інтелектуального методу формування споживчого кошику

Далі проведено кластеризацію. Для реалізації поставленої задачі обрано метод кластеризації k-means [15, 19], адже він є найпопулярнішим та найпростішим підходом кластеризації. У наступних наукових дослідженнях використаємо інші методи кластеризації.

Для реалізації розробленого методу після проведення попередньої обробки даних потрібно підключити бібліотеку, яка дозволить реалізацію кластеризації. Для виклику цієї бібліотеки використовується наступний код [40]:

```
from sklearn.cluster import KMeans
```

клас sklearn.cluster. **KMeans** (*n_clusters = 8 , * , init = 'k-means++' , n_init = 10 , max_iter = 300 , tol = 0,0001 , verbose = 0 , random _state = None , copy_x = True , algorithm = 'lloyd'*)

Параметри що важливі при використанні функцій даної бібліотеки:

- n_clusters *int, за замовчуванням=8*. Кількість кластерів для формування, а також кількість центроїдів для створення.
- init *{'k-means++', 'random'}, що викликається або має форму масиву (n_clusters, n_features), за замовчуванням='k-means++'*

- Спосіб ініціалізації:
  1) 'k-means++': вибирає початкові центроїди кластера за допомогою вибірки на основі емпіричного розподілу ймовірностей внеску точок у загальну інерцію. Ця техніка прискорює конвергенцію, і це теоретично доведено O(log k)-оптимальний.
  2) 'випадковий': вибір n_clusters спостережень (рядків) випадковим чином із даних для початкових центроїдів.
  3) Якщо передається масив, він повинен мати форму (n_clusters, n_features) і давати початкові центри. Якщо викликаний передається, він повинен приймати аргументи X, n_clusters і випадковий стан і повертати ініціалізацію.
- n_init *int, за замовчуванням=10*. Кількість разів, коли алгоритм k-середніх виконуватиметься з різними початковими числами центроїда. Остаточними результатами будуть найкращі результати послідовних прогонів n_init з точки зору інерції.
- max_iter *int, за замовчуванням=300*. Максимальна кількість ітерацій алгоритму k-середніх для одного запуску.
- tol *float, за замовчуванням=1e-4*. Відносна толерантність щодо норми Фробеніуса різниці в центрах кластерів двох послідовних ітерацій для оголошення конвергенції.
- verbose *int, за замовчуванням=0*. Режим багатослівності.
- random_state *int, екземпляр RandomState або None, за замовчуванням=None* Визначає генерацію випадкових чисел для ініціалізації центроїда. Використовується int, щоб зробити випадковість детермінованою.
- copy_x *bool, за замовчуванням = True*. Під час попереднього обчислення відстаней спочатку варто точніше чисельно

центрувати дані. Якщо copy_x має значення True (за замовчуванням), вихідні дані не змінюються. Якщо False, вихідні дані змінюються та повертаються до повернення функції, але невеликі числові відмінності можуть бути внесені шляхом віднімання та додавання середнього значення даних. Якщо вихідні дані не є C-суміжними, буде створено копію, навіть якщо copy_x має значення False. Якщо вихідні дані розріджені, але не у форматі CSR, буде створено копію, навіть якщо copy_x має значення False.

- алгоритм *{“lloyd”, “elkan”, “auto”, “full”}, default=”lloyd”*. K-означає алгоритм для використання. Класичний алгоритм у стилі EM є "lloyd". Варіація "elkan" може бути більш ефективною для деяких наборів даних із чітко визначеними кластерами за допомогою нерівності трикутника. Однак він потребує більше пам’яті через виділення додаткового масиву shape. (n_samples, n_clusters)
- "auto"і "full" є застарілими, і їх буде видалено в Scikit-Learn 1.3. Вони обидва є псевдонімами для "lloyd".
- cluster_centers_ *ndarray форми (n_clusters, n_features)*. Координати центрів кластерів. Якщо алгоритм зупиняється до повної збіжності (див. toli max_iter), вони не будуть сумісні з labels_.
- labels_ *ndarray форми (n_зразків)*. Мітки кожної точки.
- n_iter_ *внутр.* Кількість ітерацій.
- n_features_in_ *int*. Кількість функцій, які видно під час підгонки .
- feature_names_in_ *ndarray форми (n_features_in_,).* Назви функцій, які можна побачити під час підгонки. Визначається лише тоді, коли X всі імена функцій є рядками.

Далі потрібно визначити оптимальну кількість кластерів, це можна зробити методом ліктя. Методом ліктя розраховується квадратна сума відстаней точок від центру кластера відповідно до кожного значення K. За цими значеннями проводиться графік для кожного значення K. Точка ліктя на графіку, де різниця між підсумками починає зменшуватися, визначається як найбільш відповідне значення K.

KelbowVisualizer [41] реалізує метод «ліктя», щоб допомогти спеціалістам із обробки даних вибрати оптимальну кількість кластерів, налаштовуючи модель із діапазоном значень. Якщо лінійна діаграма нагадує руку, то «лікоть» (точка перегину кривої) є хорошим показником того, що базова модель найкраще підходить до цієї точки. У візуалізаторі «лікоть» буде позначено пунктирною лінією.

Щоб продемонструвати, у наступному прикладі KelbowVisualizer підходить K-means модель для діапазону значення від 4 до 11 на зразковому двовимірному наборі даних із 8 випадковими кластерами точок. Коли модель складається з 8 кластерів, ми можемо побачити лінію, що позначає «лікоть» на графіку, що в цьому випадку, як ми знаємо, є оптимальним числом.

Для реалізації цього підходу використовується наступний код:

```
elbow_method = KElbowVisualizer(KMeans(), k=6)
elbow_method.fit(scaled_df)
elbow_method.show()
```

За замовчуванням для параметра оцінки metric встановлено значення distortion, що обчислює суму квадратів відстаней від кожної точки до її призначеного центру. Однак дві інші метрики також можна використовувати з KElbowVisualizer– silhouettei calinski_harabasz. Оцінка silhouette обчислює середній коефіцієнт силуету всіх зразків, тоді як

calinski_harabaszоцінка обчислює співвідношення дисперсії між кластерами та всередині них.

Також KElbowVisualizer відображається кількість часу для навчання моделі кластеризації на K у вигляді пунктирної зеленої лінії, але її можна приховати, встановивши timings=False. У наступному прикладі ми використаємо calinski_harabasz оцінку та приховаємо час, щоб відповідати моделі.

Для покращення результатів кластеризації використаємо GaussianMixture [42] реалізує алгоритм максимізації очікування (EM) для підгонки моделей суміші Гауса. Він також може малювати еліпсоїди довіри для багатовимірних моделей і обчислювати байєсівський інформаційний критерій для оцінки кількості кластерів у даних. Надається GaussianMixture.fit метод, який вивчає модель суміші Гауса з даних поїзда. Враховуючи тестові дані, він може призначити кожній вибірці гауссівську функцію, до якої вона, ймовірно, належить, використовуючи цей GaussianMixture.predict метод. Поставляється GaussianMixture з різними параметрами для обмеження коваріації оцінених класів різниці: сферична, діагональна, пов’язана або повна коваріація.

У нашому випадку даний підхід реалізується наступним кодом:

```
gmm = GaussianMixture(n_components=3, covariance_type=
'spherical', max_iter=2000, random_state=42).fit(scale
d_df)
    labels = gmm.predict(scaled_df)
```

де,

n_components *int, за замовчуванням=1* - Кількість компонентів суміші.

covariance_type *{'full', 'tied', 'diag', 'spherical'}, default='full'* - *р*ядок, що описує тип параметрів коваріації для використання. Має бути одним із:

- «повний»: кожен компонент має власну загальну коваріаційну матрицю.
- «зв’язаний»: усі компоненти мають однакову загальну коваріаційну матрицю.
- «diag»: кожен компонент має власну діагональну коваріаційну матрицю.
- «сферичний»: кожен компонент має свою окрему дисперсію.

max_iter *int, за замовчуванням=100* - кількість EM-ітерацій, які потрібно виконати.

random_state *int, екземпляр RandomState або None, за замовчуванням=None* - контролює випадкове початкове число, надане методу, вибраному для ініціалізації параметрів (див init_params). Крім того, він контролює генерацію випадкових вибірок із підігнаного розподілу (див. метод sample). Потрібно передати int для відтворюваного виведення через кілька викликів функцій.

Далі потрібно створити модель навчання без нагляду. Таким чином, немає позначеної функції, яку можна використовувати для оцінки даної моделі. Оцінка в алгоритмах кластеризації полягає в тому, щоб вивчити закономірності, які виникають. Далі можна сформувати отримані групи, які утворюють кластери.

Вивести можна за допомогою діаграм:

```
cl = ['#FF0000', '#00FF00', '#0000FF']
plt.figure(figsize=(14,8))
sns.countplot(x=data['Clusters'], palette=cl)
```

Для створення нової фігури або активації наявної фігури, використовується функція plt.figure. Для реалізації розробленого методу використано наступні параметри:

- num *int or str or Figure or SubFigure, optional.* Унікальний ідентифікатор фігури. Якщо фігура з цим ідентифікатором вже існує, ця цифра стає активною та повертається. Ціле число відноситься до атрибута Figure.number, рядок - до мітки рисунка. Якщо немає цифри з ідентифікатором або *номер* не наводиться, створюється нова цифра, робиться активною і повертається. Якщо *num* є int, він буде використовуватися для атрибута Figure.number, в іншому випадку використовується автоматично згенероване ціле значення (починаючи з 1 і збільшуючись для кожної нової цифри). Якщо *num* – це рядок, для цього значення буде встановлено підпис рисунка та заголовок вікна. Якщо num є SubFigure, активується його батьківська фігура.
- figsize*(float,float),*
  default: rcParams["figure.figsize"] (default: [6.4, 4.8]). Ширина, висота в дюймах.

```
plt.figure(figsize=(14,8))
sns.jointplot(x=data["Total_Spend"], y=data["Income"],
 hue=data["Clusters"], palette=cl);
```

`sns.jointplot` функція забезпечує зручний інтерфейс для JointGrid класу з кількома стандартними видами сюжетів. Це має бути досить легка обгортка; якщо потрібна більша гнучкість, можна використовувати JointGrid безпосередньо. При виводі результатів даної функції було обрано наступні параметри:

- data *pandas.DataFrame, numpy.ndarray, mapping, or sequence.* Структура вхідних даних, або довга колекція векторів, які можуть бути призначені іменованим змінним, або широкоформатний набір даних, який буде внутрішньо змінений.

- x, y *vectors or keys in data*. Змінні, які визначають положення на осях x та y.
- hue *vector or key in data*. Семантична змінна, яка зіставляється для визначення кольору сюжетних елементів. Семантична змінна, яка зіставляється для визначення кольору сюжетних елементів.
- palette *string, list, dict, or matplotlib.colors.Colormap*. Метод вибору кольорів використовувати при відображенні смислового відтінку. Значення рядків передаються до color_palette(). Значення списку або дикту мають на увазі категоричне відображення, тоді як об'єкт кольорової карти передбачає числове відображення.

Далі проведемо експертне дослідження отриманих результатів.

3.3 Експериментальні дослідження розробленого методу

Для вибору кількості кластерів, обрано Elbow Method [20]. Даний метод обчислюється, як квадратна сума відстаней точок від центру кластера відповідно до кожного значення k. Відповідно до цих значень для кожного значення k складається графік. Точка ліктя на графіку, де різниця між підсумками починає зменшуватися, визначається як найбільш відповідне значення k. Для нашого прикладу оптимальною кількістю кластерів є три (Рис.3.10).

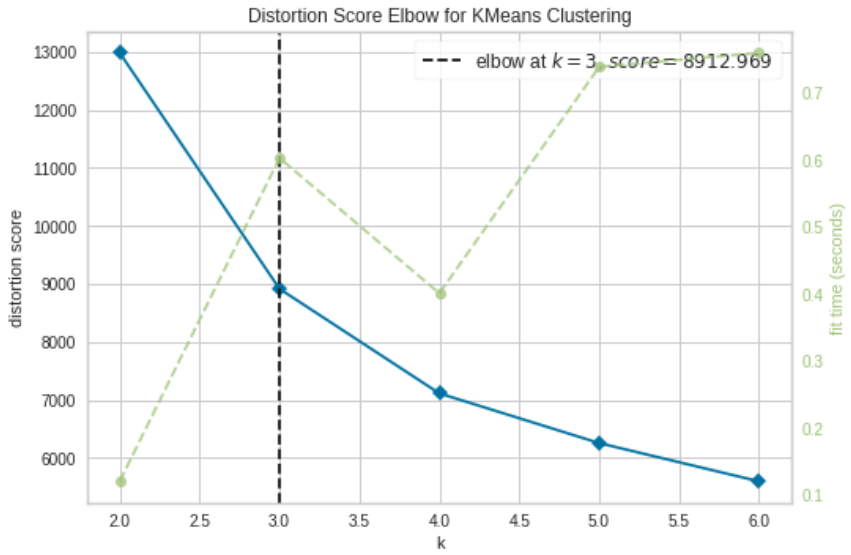


Рис.3.10 - Вибір кількості кластерів методом коліна

Розглянемо візуалізацію отриманих трьох кластерів. Як видно (Рис.3.11a) в кластері під номером нуль найбільше клієнтів. Проведемо інтерпретацію (Рис.3.11b) отриманих кластерів відповідно до характеристик доходів і загальних витрат:

Кластер 2: високі витрати - середній дохід;

Кластер 1: великі витрати - високі доходи;

Кластер 0: низькі витрати - низькі доходи.

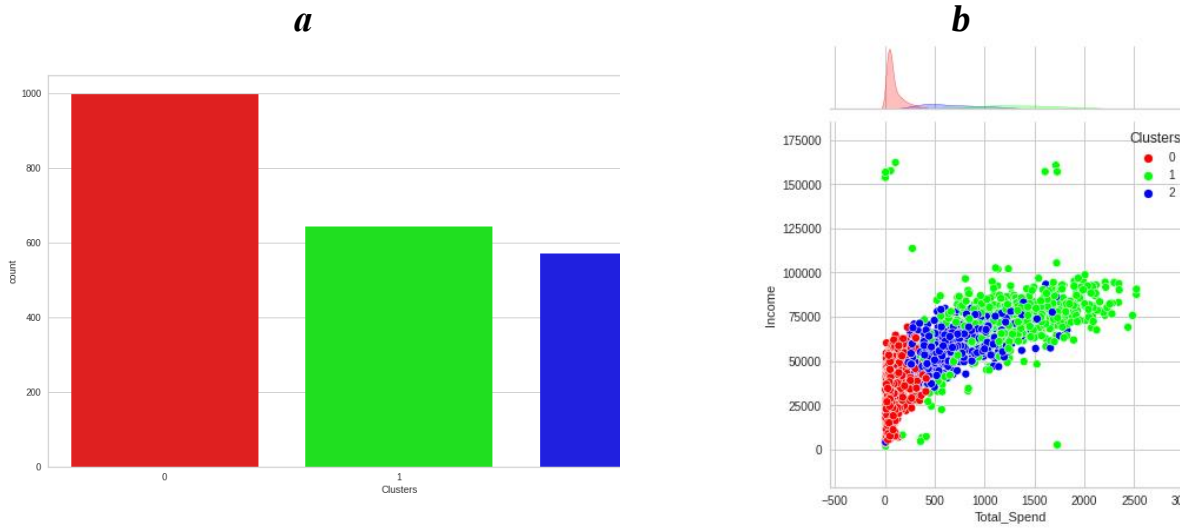


Рис.3.11 - Візуалізація кластерів. (a) Кількість клієнтів в кластерах. (b) розподіл доходів та витрат клієнтів.

Якщо розглядати кластери, як соціальні, то вважливим, є саме Кластер 2 та Кластер 0. Тому варто орієнтуватись у формуванні споживчого кошика на ці два кластера. А саме формувати споживчий кошик на основі Кластера 2 для економічно стабільної ситуації, а в кризових ситуаціях на основі Кластера 0.

При розгляді вікової категорії покупців (Рис.3.12), видно, що розподіл кластерів достатньо рівномірний. Тому на цих даних можна вважати, що вік не впливає на купівельну спроможність.

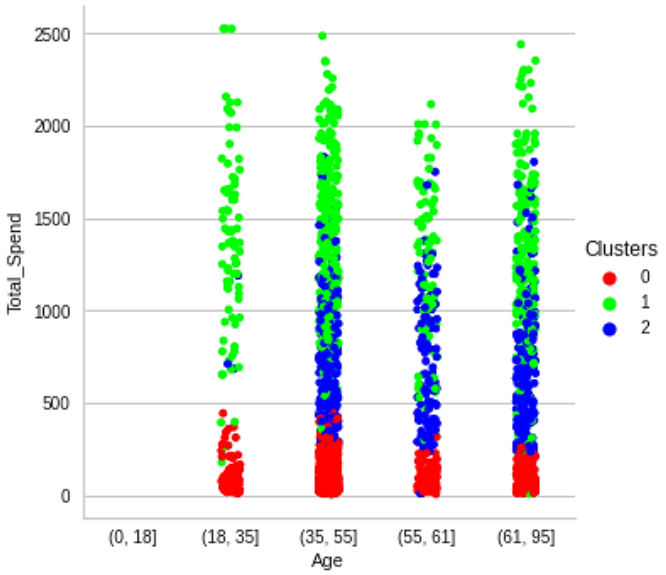
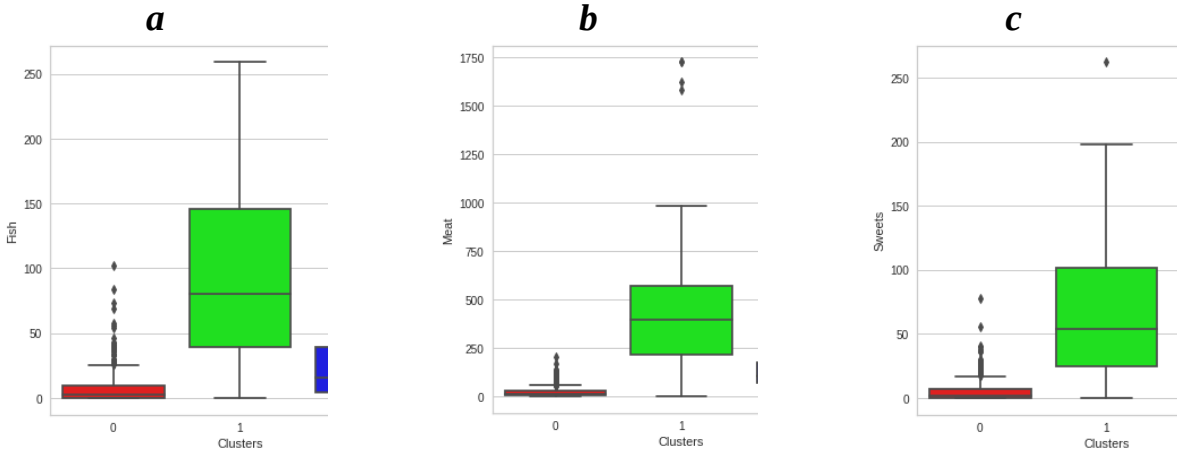


Рис.3.13 - Візуалізація кластерів. Віковий розподіл

Далі детальніше розглянемо кластери відносно суми (дол.США) куплених товарів (Рис.3.13) за 2 роки. Діаграми boxplot, найкраще показують точку визначення середньої кількості придбаних товарів у кластері. З результатів видно, що Кластер 1 витрачають дуже мало коштів на продукти риба, м'ясо, солодощі, вино. Проте Кластер 2 демонструє більші суми витрачені на ці товари. Якщо говорити про технологічні товари, то всі покупці мають приблизно однакові витрати (Рис.3.14e). Отже, по продажах товарів кластер 2 має ті значення, які повинні бути при формуванні споживчого кошика. Опишемо, кластер 2 детальніше, на основі, якого дамо рекомендації стосовно формування споживчого кошика.



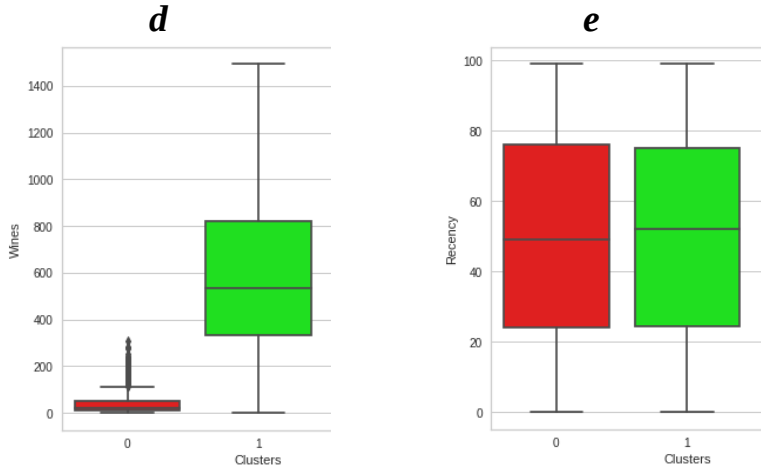


Рис.3.14 - Візуалізація кластерів. Розподіл куплених товарів: (a) риба; (b) м'ясо; (c) солодощі; (d) вино; (e) технологічні товари.

Споживчий кошик (Кластер 2): до нього належить 27% клієнтів різної вікової категорії з середнім доходом; рибної продукції придбано за місяць на суму в 25 дол.США; м’ясної продукції придбано на 100 дол.США/місяць; солодощів на суму 20 дол.США/місяць; винної продукції на 400 дол.США/місяць; техніки на суму 50 дол.США/місяць.

Варто теж розуміти, що ці результати не можуть повністю відповідати реальній ситуації, адже даних є замало.

Отже, інтелектуальний метод формування споживчого кошику, дозволяє визначити набір споживчого кошику на основі даних супермаркетів. Повноцінний збір даних з мережевих супермаркетів можуть дати більш точні результати по кластерах. Це дозволить проводити моніторинг купівельної спроможності людей в реальному часі, що дозволить оперативно сформувати споживчий кошик.

На відміну від аналогів [9, 11, 16] розроблений інтелектуальний метод формування споживчого кошику, дозволить на основі даних мереж супермаркетів сформувати соціальний кластер, на основі якого сформовано споживчий кошик, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

До кластера належить 27% клієнтів різної вікової категорії з середнім доходом; рибної продукції придбано за місяць на суму в 25 дол.США; м’ясної продукції придбано на 100 дол.США/місяць; солодощів на суму 20 дол.США/місяць; винної продукції на 400 дол.США/місяць; техніки на суму 50 дол.США/місяць.

Висновки до розділу 3

1. Проведено попередню обробку даних. Визначено показники, які найбільше між собою корелюють. Проведено зменшення розмірності даних та видалення не потрібних параметрів.

2. Для реалізації поставленої задачі обрано метод кластеризації k-means, адже він є найпопулярнішим та найпростішим підходом кластеризації.

3. З результатів кластеризації можна зробити висновок, що формувати споживчий кошик потрібно на основі Кластера 2, а в кризових ситуаціях на основі Кластера 0. Отже, споживчий кошик (Кластер 2), до нього належить 27% споживачів різної вікової категорії з середнім доходом. На рибну продукцію, має бути виділено не менше 25 дол.США/міс; на м’ясну не менше 100 дол.США/місяць; на солодощі 20 дол.США/місяць; на винну продукцію 400 дол.США/місяць; на техніку 50 дол.США/місяць.

ВИСНОВКИ

Розробка інтелектуального методу формування споживчого кошику на основі даних мереж супермаркетів є актуальною та дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

Вхідними даними використано набір Customer Personality Analysis з платформи Kaggle, в якому представлено аналіз особистості клієнта. Цих даних не цілком достатньо для повноцінного розв’язання задачі, проте для реалізації розроблено методу було вдосталь.

Для реалізації поставленої задачі обрано метод кластеризації k-means, адже він є найпопулярнішим та найпростішим підходом кластеризації. З результатів кластеризації можна зробити висновок, що формувати споживчий кошик потрібно на основі Кластера 2, що відповідає середньому доходу, а в кризових ситуаціях на основі Кластера 0, який містить інформацію про витрати людей із низькими доходами. Споживчий кошик, що визначається Кластером 2, містить інформацію щодо 27% споживачів різної вікової категорії з середніми доходами, які витрачають на рибну продукцію не менше 25 дол.США/міс, на м’ясну не менше 100 дол.США/місяць, на солодощі 20 дол.США/місяць, на винну продукцію 400 дол.США/місяць, на техніку 50 дол.США/місяць.

Варто теж розуміти, що ці результати не можуть повністю відповідати реальній ситуації, адже даних є замало, вони можуть відрізнятись залежно від магазинів, сезону та місцевості. Повноцінний збір даних з мережевих супермаркетів може дати більш точні результати по кластерах, що дозволить проводити моніторинг купівельної спроможності людей в реальному часі та оперативно формувати споживчий кошик.

На відміну від аналогів [9, 11, 16] розроблений інтелектуальний метод формування споживчого кошику, дозволить на основі даних мереж супермаркетів сформувати соціальний кластер, на основі якого можна формувати споживчий кошик.

Майбутні дослідження даної області можливі та необхідні, адже для формування споживчого кошику у певній країні необхідно мати відповідні дані. Наразі повних та достатніх наборів даних відносно України немає, тому в першу чергу варто звернути увагу на методи збору інформації і також їхню відповідність, щоб визначити реальну ситуацію. Також, в наступних наукових дослідженнях на основі отриманих соціальних кластерів та методів машинного навчання можна розробити метод передбачення змін у формуванні споживчого кошика. Корисним буде проведення перехресного кластерного аналізу із різними підходами, для визначення найкращого із них.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kapała AM. Legal Instruments to Support Short Food Supply Chains and Local Food Systems in France. Laws. 2022; 11(2):21. https://doi.org/10.3390/laws11020021
2. Govindasamy, Ramu. "Cluster Analysis of Wine Market Segmentation – A Consumer Based Study in the Mid-Atlantic USA." Economic Affairs, vol. 63, no. 1, 4 Mar. 2018, doi:10.30954/0424-2513.2018.00150.19.
3. Qisman, M., Rosadi, R., & Abdullah, A. S. (2021). Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores). In Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012020). IOP Publishing.
4. Kutuzova Tatiana, Melnik Mikhail, Market basket analysis of heterogeneous data sources for recommendation system improvement, Procedia Computer Science, Elsevier B.V., Volume 136, 2018, Pages 246-254, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.08.263.
5. Sun, Q., Gao, X., Wang, Z. et al. (2020), Quantifying the risk of price fluctuations based on weighted Granger causality networks of consumer price indices: evidence from G7 countries. J Econ Interact Coord, Springer, 15, 821–844. https://doi.org/10.1007/s11403-019-00273-2
6. Sarantitis, G.A., Papadimitriou, T. & Gogas, P. (2018) A Network Analysis of the United Kingdom’s Consumer Price Index. Comput Econ, Springer, 51, 173–193. https://doi.org/10.1007/s10614-016-9625-9

7. Rathnayaka, S. D., Selvanathan, E. A., & Selvanathan, S. (2022). Modelling the consumption patterns in the Asian countries. Economic Analysis and Policy. Elsevier. Volume 74, June 2022, Pages 277-296
8. Peixoto, V., Peixoto, H., Machado, J. (2020). Integrating a Data Mining Engine into Recommender Systems. In: Analide, C., Novais, P., Camacho, D., Yin, H. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2020. IDEAL 2020. Lecture Notes in Computer Science(), vol 12489. Springer, Cham. https://doi.org/10.1007/978-3-030-62362-3_19
9. Dogan, O., Hiziroglu, A., Seymen, O.F. (2021). Segmentation of Retail Consumers with Soft Clustering Approach. In: Kahraman, C., Cevik Onar, S., Oztaysi, B., Sari, I., Cebi, S., Tolga, A. (eds) Intelligent and Fuzzy Techniques: Smart and Innovative Solutions. INFUS 2020. Advances in Intelligent Systems and Computing, vol 1197. Springer, Cham. https://doi.org/10.1007/978-3-030-51156-2_6
10. Chiang, L. L. L., & Yang, C. S. (2018). Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach. Technological Forecasting and Social Change, Elsevier, 130, 177-187.
11. Munusamy, S., Murugesan, P. Modified dynamic fuzzy c-means clustering algorithm – Application in dynamic customer segmentation. Appl Intell, Springer, 50, 1922–1942 (2020). https://doi.org/10.1007/s10489-019-01626-x
12. Arunachalam, Deepak; Kumar, Niraj (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. Expert Systems with Applications, Elsevier, 111, 11-34.doi:10.1016/j.eswa.2018.03.007
13. Anitha, P.; Patil, Malini M. (2019). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud

University - Computer and Information Sciences, Elsevier, 1319-1578. doi:10.1016/j.jksuci.2019.12.011

14. Lipyanina-Goncharenko, H., Brych, V., Sachenko, S., Lendyuk, T., Bykovyy, P., & Zahorodnia, D. (2021). Method of forming a training sample for segmentation of tender organizers on machine learning basis. In CEUR Workshop Proceedings (pp. 1843-1852).
15. Kanavos A, Iakovou SA, Sioutas S, Tampakas V. Large Scale Product Recommendation of Supermarket Ware Based on Customer Behaviour Analysis. Big Data and Cognitive Computing. 2018; 2(2):11. https://doi.org/10.3390/bdcc2020011
16. Alam, M. F., Singh, R., & Katiyar, S. (2021, December). Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 94-98). IEEE.
17. K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
18. Customer Personality Analysis. (б. д.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis
19. Fränti, P., Sieranoja, S. K-means properties on six clustering benchmark datasets. Appl Intell, Springer, 48, 4743–4759 (2018). https://doi.org/10.1007/s10489-018-1238-7
20. D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," 2018 International Seminar on

Application for Technology of Information and Communication, IEEE, 2018, pp. 533-538, doi: 10.1109/ISEMANTIC.2018.8549751.

21. Sanjit K. Roy, Gaganpreet Singh, Megan Hope, Bang Nguyen & Paul Harrigan (2019) The rise of smart consumers: role of smart servicescape and smart consumer experience co-creation, Journal of Marketing Management, 35:15-16, 1480-1513, DOI: 10.1080/0267257X.2019.1680569
22. Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. Journal of Retailing, Elsevier, 93(1), 79-95.
23. Eleonora Pantano (2019) The role of smart technologies in decision making: developing, supporting and training smart consumers, Journal of Marketing Management, 35:15-16, 1367-1369, DOI: 10.1080/0267257X.2019.1688927
24. Мазур Н.О., Ткачук А.В. Прожитковий мінімум у системі соціальних стандартів життя: вітчизняний і зарубіжний досвід. [Електронний ресурс] / Н. О. Мазур // Соціально-трудові відносини – теорія та практика: збірник наукових праць. - 2016. - ғ2(12)2016. - С. 219./ Режим доступу: http://stvua.com/?wpfb_dl=25
25. Про затвердження Методики визначення прожиткового мінімуму: Наказ Міністерства соціальної політики від 03.02.2017 ғ178/147/31 [Електронний ресурс]// База даних «Законодавство України»/ Верховна Рада України/ Режим доступу: http://zakon5.rada.gov.ua/laws/show/z0281-17
26. Про прожитковий мінімум: Закон України від 15.07.1999 ғ 966-XIV в редакції від 20.01.2018 [Електронний ресурс]// База даних «Законодавство України»/ Верховна Рада України/ Режим доступу: http://zakon5.rada.gov.ua/laws/show/966-14 Експертні дослідження товарів і послуг 141

27. Про Державний бюджет України: Закон України від 07.12.2017 ғ 2246-VIII [Електронний ресурс]// База даних «Законодавство України»/ Верховна Рада України/ Режим доступу: http://zakon3.rada.gov.ua/laws/show/2246-19
28. Прожиточный минимум и минимальная зарплата в Польше. Часть 1: «Сухая» статистика. [Електронний ресурс]// Escape to Poland. 2015./ Режим доступу: http://escape-pl.com/zhizn-v-polshe/prozhitochnyiy-minimumminimalnaya-zarplata-v-polshe-chast-1-suhaya-statistika/
29. Про затвердження наборів продуктів харчування, наборів непродовольчих товарів та наборів послуг для основних соціальних і демографічних груп населення: Постанова Кабінету Міністрів України від 11.10.2016 №780 [Електронний ресурс]// База даних «Законодавство України»/ Верховна Рада України/ Режим доступу: http://zakon2.rada.gov.ua/laws/show/780- 2016-%D0%BF/page
30. Ковязіна К. О. Щодо удосконалення методики визначення споживчого кошику. [Електронний ресурс]// Національний інститут стратегічних досліджень. 2014./ Режим доступу: http://www.niss.gov.ua/articles/1233
31. Голуб, Н. С. (2021). Дослідження вартості споживчого кошика в Україні. SCIENCE FOUNDATIONS OF MODERN SCIENCE AND PRACTICE, 10, 99.
32. ЩЕРБА, А. (2021). Споживчий кошик та рівень життя населення в умовах карантину. «MANAGEMENT, ADMINISTRATION AND LAW: PROBLEMS, TRENDS, ACHIEVEMENTS» № 4, 2021, 230.
33. Zhou, Z. H. (2021). Machine learning. Springer Nature.

34. Maharana, K., Mondal, S., & Nemade, B. (2022). A Review: Data Pre-Processing and Data Augmentation Techniques. Global Transitions Proceedings.
35. Kapsner, L. A., Mang, J. M., Mate, S., Seuchter, S. A., Vengadeswaran, A., Bathelt, F., ... & Prokosch, H. U. (2021). Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. Applied Clinical Informatics, 12(04), 826-835.
36. Ilyas, I. F., & Chu, X. (2019). Data cleaning. Morgan & Claypool.
37. Wu, X., Cheng, C., Zurita-Milla, R., & Song, C. (2020). An overview of clustering methods for geo-referenced time series: From one-way clustering to co-and tri-clustering. International journal of geographical information science, 34(9), 1822-1848.
38. Nasraoui, O., & N'Cir, C. E. B. (2019). Clustering methods for big data analytics. Techniques, Toolboxes and Applications, 1, 91-113.
39. Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric pollution research, 11(1), 40-56.
40. sklearn.cluster.KMeans. scikit-learn. URL: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (date of access: 28.10.2022).
41. Elbow Method – Yellowbrick v1.5 documentation. Yellowbrick: Machine Learning Visualization – Yellowbrick v1.5 documentation. URL: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html (date of access: 28.10.2022).
42. sklearn.mixture.GaussianMixture. scikit-learn. URL: http://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html (date of access: 28.10.2022).

43. Gramyak Roman, Hrystyna Lipyanina-Goncharenko, Anatoliy Sachenko, Taras Lendyuk, and Diana Zahorodnia. "Intelligent Method of a Competitive Product Choosing based on the Emotional Feedbacks Coloring." In IntelITSIS, pp. 246-257. 2021.
44. Oleksandr Osolinskyi, Lubomyr Kolodiychyk, Hrystyna Lipyanina-Goncharenko, Anatoliy Sachenko, Lukasz Kopania, Volodymyr Kochan. Conceptual model of IoT-based Laboratory for study the Electrical Engineering and Electronics. Proceedings of The Fourth International Workshop on Computer Modeling and Intelligent Systems (CMIS-2021). CEUR Workshop Proceedings, 2021, 2864, pp. 344–355
45. Hrystyna Lipyanina-Goncharenko, Vasyl Brych, Svitlana Sachenko, Taras Lendyuk, Pavlo Bykovyya and Diana Zahorodnia. Method of Forming a Training Sample for Segmentation of Tender Organizers on Machine Learning Basis. Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021). Volume I: Main Conference Lviv, Ukraine, April 22-23, 2021. CEUR Workshop Proceedings, 2021, 2870, pp. 1843–1852.
46. Andriy Krysovatyy, Hrystyna Lipyanina-Goncharenko, Svitlana Sachenko and Oksana Desyatnyuk. Economic Crime Detection Using Support Vector Machine Classification. Modern Machine Learning Technologies and Data Science Workshop. Proc. 3rd International Workshop (MoMLeT&DS 2021). Volume I: Main Conference. Lviv-Shatsk, Ukraine, June 5-6, 2021. pp 830-840.
47. Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого «бакалаврського» і другого «магістерського» рівнів / За ред. доц. М.І. Шинкарика. Тернопіль: ТНЕУ, 2018. 67с.

48. Комар М.П., Саченко А.О., Васильків Н.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2021. 32с.
49. Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого «бакалаврського» і другого «магістерського» рівнів / За ред. доц. М.І. Шинкарика. Тернопіль: ТНЕУ, 2018. 67с.
50. Комар М.П., Саченко А.О., Васильків Н.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2021. 32с.
51. Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого «бакалаврського» і другого «магістерського» рівнів / За ред. доц. М.І. Шинкарика. Тернопіль: ТНЕУ, 2018. 67с.
52. Комар М.П., Саченко А.О., Васильків Н.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп’ютерні науки» спеціальності 122 «Комп’ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2021. 32с.
53. Ліп'яніна-Гончаренко Х. ІНТЕЛЕКТУАЛЬНИЙ МЕТОД ФОРМУВАННЯ СПОЖИВЧОГО КОШИКУ / Христина Ліп'яніна-Гончаренко, Зоряна Чижовська // ЕКОНОМІЧНИЙ І СОЦІАЛЬНИЙ РОЗВИТОК УКРАЇНИ В XXI СТОЛІТТІ: НАЦІОНАЛЬНА ВІЗІЯ ТА ВИКЛИКИ ГЛОБАЛІЗАЦІЇ : XIX

Міжнар. науково-практ. конф. молодих вчен., Тернопіль, 13 трав. 2022 р. – [Б. м.]. – С. 102–104.

54. Lipianina-Honcharenko, K., Wolff, C., Chyzhovska, Z., Sachenko, A., Lendiuk, T., Grodskyi, S. (2022). Intelligent Method for Forming the Consumer Basket. In: Lopata, A., Gudonienė, D., Butkienė, R. (eds) Information and Software Technologies. ICIST 2022. Communications in Computer and Information Science, vol 1665. Springer, Cham. https://doi.org/10.1007/978-3-031-16302-9_17

ДОДАТОК А

СТРУКТУРА ДАНИХ

```
df[df.isnull().any(axis=1)]
Out[4]:
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
10	1994	1983	Graduation	Married	NaN	1	0	15-11-2013	11	5	...	7	0	0	0	0	0	0	3	11	0
27	5255	1986	Graduation	Single	NaN	1	0	20-02-2013	19	5	...	1	0	0	0	0	0	0	3	11	0
43	7281	1959	PhD	Single	NaN	0	0	05-11-2013	80	81	...	2	0	0	0	0	0	0	3	11	0
48	7244	1951	Graduation	Single	NaN	2	1	01-01-2014	96	48	...	6	0	0	0	0	0	0	3	11	0
58	8557	1982	Graduation	Single	NaN	1	0	17-06-2013	57	11	...	6	0	0	0	0	0	0	3	11	0
71	10629	1973	2n Cycle	Married	NaN	1	0	14-09-2012	25	25	...	8	0	0	0	0	0	0	3	11	0
90	8996	1957	PhD	Married	NaN	2	1	19-11-2012	4	230	...	9	0	0	0	0	0	0	3	11	0
91	9235	1957	Graduation	Single	NaN	1	1	27-05-2014	45	7	...	7	0	0	0	0	0	0	3	11	0
92	5798	1973	Master	Together	NaN	0	0	23-11-2013	87	445	...	1	0	0	0	0	0	0	3	11	0
128	8268	1961	PhD	Married	NaN	0	1	11-07-2013	23	352	...	6	0	0	0	0	0	0	3	11	0
133	1295	1963	Graduation	Married	NaN	0	1	11-08-2013	96	231	...	4	0	0	0	0	0	0	3	11	0
312	2437	1989	Graduation	Married	NaN	0	0	03-06-2013	69	861	...	3	0	1	0	1	0	0	3	11	0
319	2863	1970	Graduation	Single	NaN	1	2	23-08-2013	67	738	...	7	0	1	0	1	0	0	3	11	0
1379	10475	1970	Master	Together	NaN	0	1	01-04-2013	39	187	...	5	0	0	0	0	0	0	3	11	0
1382	2902	1958	Graduation	Together	NaN	1	1	03-09-2012	87	19	...	5	0	0	0	0	0	0	3	11	0

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
1383	4345	1964	2n Cycle	Single	NaN	1	1	12-01-2014	49	5	...	7	0	0	0	0	0	0	3	11	0
1386	3769	1972	PhD	Together	NaN	1	0	02-03-2014	17	25	...	7	0	0	0	0	0	0	3	11	0
2059	7187	1969	Master	Together	NaN	1	1	18-05-2013	52	375	...	3	0	0	0	0	0	0	3	11	0
2061	1612	1981	PhD	Single	NaN	1	0	31-05-2013	82	23	...	6	0	0	0	0	0	0	3	11	0
2078	5079	1971	Graduation	Married	NaN	1	1	03-03-2013	82	71	...	8	0	0	0	0	0	0	3	11	0
2079	10339	1954	Master	Together	NaN	0	1	23-06-2013	83	161	...	6	0	0	0	0	0	0	3	11	0
2081	3117	1955	Graduation	Single	NaN	0	1	18-10-2013	95	264	...	7	0	0	0	0	0	0	3	11	0
2084	5250	1943	Master	Widow	NaN	0	0	30-10-2013	75	532	...	1	0	0	1	0	0	0	3	11	1
2228	8720	1978	2n Cycle	Together	NaN	0	0	12-08-2012	53	32	...	0	0	1	0	0	0	0	3	11	0

ДОДАТОК Б

КОД ПОПЕРЕДНЬОГО АНАЛІЗУ ДАНИХ

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px

from matplotlib import rcParams
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

In [3]:

```
df=pd.read_csv("customer-personality-analysis/marketing_campaign.csv", sep="\t")
```

In [4]:

```
df
```

In [5]:

```
df.info()
```

In [6]:

```
df["Dt_Customer"]=pd.to_datetime(df["Dt_Customer"]) #changing datatype for date
column
```

In [7]:

```
df.isna().sum()
```

In [8]:

```
df["Income"]= df["Income"].fillna(df["Income"].mean()) # fill nan with mean valu
e.
```

In [9]:

```
df=df.drop_duplicates(keep=False, inplace=False)
```

In [10]:

```
df["Marital_Status"].value_counts()
```

In [11]:

```
df["Marital_Status"]=df["Marital_Status"].replace({"Married":"In Relationship",
"Together":"In Relationship",
                                                   "Absurd":"Single", "Widow":"S
ingle", "YOLO":"Single", "Divorced":"Single","Alone":"Single"})

df["Marital_Status"].value_counts()
```

Out[11]:

```
In Relationship    1444
Single              796
Name: Marital_Status, dtype: int64
```

In [12]:

```
df["TotalSpent"]=df['MntMeatProducts']+df['MntFishProducts']+df['MntSweetProduct
s']+df['MntGoldProds']+df['MntFruits']+df['MntWines']
```

In [13]:

```
df["TotalChildren"]=df['Kidhome']+df['Teenhome']

df["Children"] = df["TotalChildren"].apply(lambda x: 'NO-Children' if x == 0 els
e 'Children')
```

In [14]:

```
df["Total_cmp_acceptance"] = df["AcceptedCmp1"]+ df["AcceptedCmp2"]+ df["Accepte
dCmp3"]+ df["AcceptedCmp4"]+ df["AcceptedCmp5"]
```

In [15]:

```
df["Age"]=2014-df['Year_Birth']
```

In [16]:

```
df=df.drop(['Z_CostContact', 'Z_Revenue'],axis=1)
```

In [18]:

```
sns.set_theme(style="white")
sns.despine(fig=None, ax=None, top=False, right=False, left=False, bottom=False,
offset=None, trim=False)
sns.set(rc={'figure.figsize':(10,5)})
sns.set_style({'axes.facecolor':'white', 'grid.color': '.8', 'font.family':'Time
s New Roman'})
rcParams['figure.figsize'] = 12,5
<Figure size 432x288 with 0 Axes>
```

In [19]:

```
plt.subplots_adjust(wspace=0.4)
fig, axes = plt.subplots(1,2)
plt.subplots_adjust(wspace=0.4)
sns.kdeplot(df["Income"],hue=df["Children"],shade=True,ax=axes[0],palette=["#EBF
F41", "#800080"])
sns.kdeplot(data=df,x="Age",hue="Children",shade=True,ax=axes[1],palette=["#EBFF
41", "#800080"])
axes[0].set_title('Income Distribution',fontsize=15)
axes[0].tick_params(axis='x', rotation=90)

axes[1].set_title('Age Distribution',fontsize=15)
fig.show()
<Figure size 864x360 with 0 Axes>
```

In [20]:

```
df=df[df["Income"]<600000]
df=df[df["Age"]< 80]
```

In [21]:

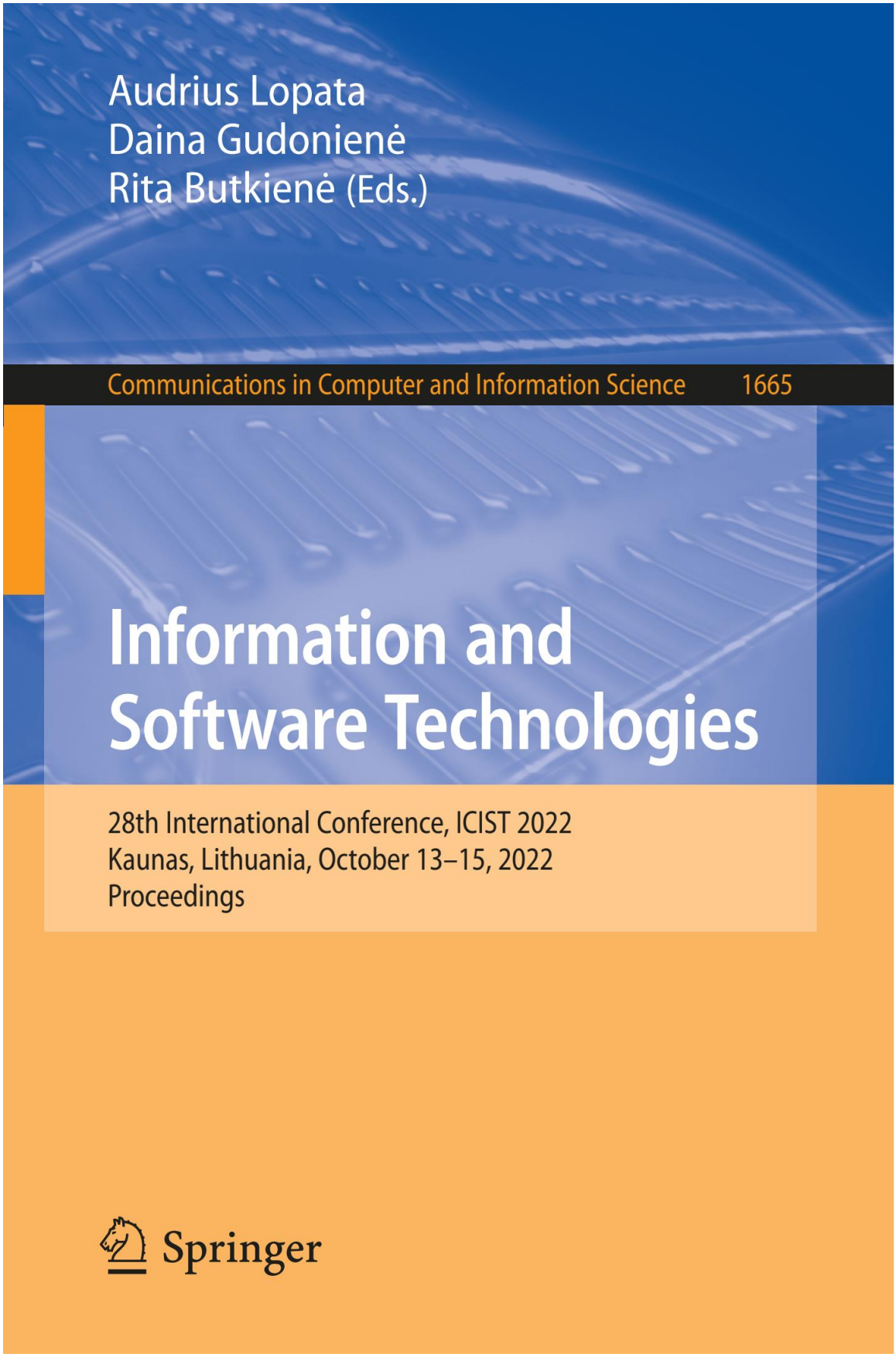
```
plt.subplots_adjust(wspace=0.7)
fig, axes = plt.subplots(ncols=2)
sns.countplot(df["Children"],ax=axes[0],palette=["#ECC5C0", "#800080"])
sns.barplot(df["Children"],df["TotalSpent"],estimator=sum,ax=axes[1],palette=["#
ECC5C0", "#800080"])
axes[0].set_title("Count of Customer who have children and who don't",fontsize=1
1)
axes[1].set_title("Spending Difference between Customers who have children and w
ho don't",fontsize=11)
fig.show()
<Figure size 864x360 with 0 Axes>
```

In [22]:

```
sns.boxplot(df["Education"],df["TotalSpent"],hue=df["Children"],palette=["#ECC5C
0", "#800080"])
plt.legend(loc='upper right')
plt.title("Spending within Different Education Levels",fontsize=15)
fig.show()
```

```
      In [23]:
plot_df = pd.DataFrame(df[['MntWines', 'MntFruits','MntMeatProducts', 'MntFishPr
oducts', 'MntSweetProducts', 'MntGoldProds']].sum()).reset_index()
plot_df.columns=["catagory","Spent_per_catagory"]
plot_df['catagory'] = plot_df['catagory'].str.replace('Mnt', '')
      In [24]:
fig = px.pie(plot_df, values='Spent_per_catagory', names='catagory',color_discre
te_sequence=px.colors.sequential.Purp)
fig.update_layout(
    autosize=False,
    width=600,
    height=300,


    template="simple_white",
    font_family='Courier New',
    paper_bgcolor="#FCFAFB",
    title_text = 'Sales by Caragory',title_x=0.5,

    showlegend=True)



fig.show()


      In [25]:
df3=df[[ 'NumWebPurchases','NumCatalogPurchases', 'NumStorePurchases' ]].sum().r
eset_index()
df3.columns=["purchases_way","pur_num"]
df3['purchases_way'] = df3['purchases_way'].map(lambda x: x.lstrip('Num'))
      In [26]:
      linkcode
fig = px.pie(df3, values='pur_num', names='purchases_way',color_discrete_sequenc
e=px.colors.sequential.Magenta)
fig.update_layout(
    autosize=False,
    width=600,
    height=300,

    template="simple_white",
    font_family='Courier New',
    paper_bgcolor="#FCFAFB",
    title_text = 'Ways of Purchases ',title_x=0.5,

    showlegend=True)



fig.show()
```

ДОДАТОК В
АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ



Editors
Audrius Lopata
Kaunas University of Technology
Kaunas, Lithuania

Daina Gudonienė
Kaunas University of Technology
Kaunas, Lithuania

Rita Butkienė
Kaunas University of Technology
Kaunas, Lithuania

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-16301-2 ISBN 978-3-031-16302-9 (eBook)
https://doi.org/10.1007/978-3-031-16302-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Business Intelligence for Information and Software Systems - Special Session on Intelligent Methods for Data Analysis and Computer Aided Software Engineering

Intelligent Method for Forming the Consumer Basket

Khrystyna Lipianina-Honcharenko[1], Carsten Wolff[2], Zoriana Chyzhovska[1], Anatoliy Sachenko[1], Taras Lendiuk[1(✉)], and Sergii Grodskyi[1]

[1] West Ukrainian National University, Lvivska Str., 11, Ternopil 46000, Ukraine
{as,tl}@wunu.edu.ua
[2] Fachhochschule Dortmund, Otto-Hahn-Str. 23, Dortmund, Germany
carsten.wolff@fh-dortmund.de

Abstract. Authors developed an intelligent method of forming a consumer basket based on data from supermarket chains, which allows modifying the set of goods in the consumer basket and defining a living wage. The consumer basket is forming on a base of k-means clustering approach. The algorithmic structure of the proposed method is described. Experimental research is carried out using the Customer Personality Analysis dataset from the Kaggle platform. After data normalization and clustering, the clusters relative to the amount (USD) of purchased goods for 2 years were analyzed. As a result, the cluster (consumer basket) was selected which includes 27% of middle-aged customers of various ages and counts such goods as fish, meat, sweets, wine and equipment. The novelty of the paper is the automated and intelligent forming the set of goods in the consumer basket, which may promote survival during humanitarian and economic disasters, especially in times of economic crisis (war, pandemic).

Keywords: Consumer basket · Clustering · k-means · Dataset

1 Introduction

Nowadays, the method of defining the consumer basket is an extremely important issue, because based on the consumer basket is formed a living wage, according to which pensions, social benefits and payments are calculated. For example, in Ukraine, the Cabinet of Ministers of Ukraine is directly involved in the formation of the consumer basket. Based on the data of the consumer basket and the consumer budget, such indicators as the minimum wage and the minimum pension are calculated.

The value of the consumer basket depends on the level of retail prices for goods and tariffs for paid services (for example, utility bills). This practice is common throughout the civilized world. For each type of needs, the calculation includes the purchase of relatively cheap goods, usually at government fixed prices. If, for example, in the market this product or service is sold at lower prices, the lowest level is taken as a basis.

The concept of the consumer basket exists in many countries around the world. Its price and national characteristics vary from country to country. For example, the

A. Lopata et al. (Eds.): ICIST 2022, CCIS 1665, pp. 221–231, 2022.
https://doi.org/10.1007/978-3-031-16302-9_17

American consumer basket includes 350 products and services, the French – 507, the Englishman – 350, the German – 475. The Ukrainian consumer basket has recently been expanded to 297 items.

People's purchasing power can change dramatically with respect to crises such as war and pandemics. Therefore, the assessment of the consumer basket and its variable must be rapid, in order to prevent humanitarian catastrophes and increasing poverty.

Hence, a development of an intelligent method for forming a consumer basket based on data from supermarket networks – a subject of this paper, is relevant.

The novelty of the paper is the automated and intelligent formation of a set of goods in the consumer basket, which may promote survival during humanitarian and economic disasters, especially in times of economic crisis (war, pandemic).

Based on the proposed method the analysis of clusters, relative to the amount (USD) of purchased goods for 2 years is made, and the cluster (consumer basket) was defined. It includes the 27% of middle-aged customers of various ages and counts such goods as fish, meat, sweets, wine and equipment.

The paper is distributed as follows. Section 2 describes the analysis of related work. Section 3 presents an intelligent method of the consumer basket forming. Section 4 considers the case study and Sect. 5 contains a discussion. Section 6 summarizes the outcomes of the study.

2 Related Work

The paper [1] identified tools for a consumer basket forming in different areas and levels, including national laws, government policies and local government initiatives, and analyzed them with the various forms of short supply chains available in France that are key components of the local food system such as direct marketing, producer shops, urban agriculture and catering. In [5] the evolution of fluctuations between consumer price index classes is investigated, and a weighted Granger Causal Network (WGCN) for each G7 country is constructed. In [6] the consumer price index in the United Kingdom as a complex network is modelled and clustering and optimization techniques are applied to study the evolution of the network over time. The paper [7] analyzes the similarity of consumer patterns in 10 Asian countries, using three well-known system-wide demand models, namely Rotterdam, CBS and AIDS. None of the papers above consider the intelligent methods for classifying customers.

In paper [9], an intuitive fuzzy clustering algorithm was applied to supermarket customer data according to the amount spent on some product groups. The analysis of the general consumer basket is not carried out in this paper.

The paper [2] attempted to predict a segment of the mid-Atlantic wine market – based on purchasing behavior, attitudes and socio-demographic characteristics. Cluster analysis is used to segment the mid-Atlantic wine market. The paper [3] analyzed the market basket using an a priori algorithm to find consumer models in the goods purchase by transaction data. In [4] a study of the integration of disparate data sources from the grocery store based on methods of market basket analysis is proposed. The paper [8] examines and implements a recommendation mechanism on a well-known Portuguese e-commerce platform specializing in clothing and sportswear to increase

customer engagement by providing a personalized experience with different types of recommendations on the platform. The paper [10] attempted to fill a gap in the study by using big data analytics to analyze approximately 44,000 point-of-sale transaction records for 26,000 Taiwanese retailer customers to understand how the consumer's personality traits are related to country of origin (COO), traits (brand identity) of beer brands, as well as to predict the Customer Lifetime Value (CLV). In those papers above, the possibility of determining the consumer basket is not considered, they deals with the segmentation of customers by individual products only.

The paper [11] proposed the modifications of the CDFCM algorithm due to certain shortcomings identified in it, which resulted in the formation of a modified dynamic fuzzy c-algorithm means (MDFCM). In this paper, the detailed analysis of cluster analysis approaches is absent.

The paper [12] evaluates the effectiveness of different data clustering approaches for finding profitable consumer segments in the UK hospitality industry. The papers [13, 28] are based on the RFM (Review, Frequency and Monetary) model and uses the principles of data set segmentation using the K-Means algorithm.

The sales results obtained in this way are compared with various parameters, such as recent sales, sales frequency and sales amount. In the paper [14, 15] a forecasting model is presented, it is based on the behavior of each customer using data analysis methods. The proposed model uses a supermarket database and an additional database from Amazon, which contains information about customer purchases. In paper [16, 27], customer segmentation was performed by clustering K-means in uncontrolled machine learning. Target customers are classified using clustering based on demographic information (e.g., gender, age, income, etc.), geographic information, psychography, and behavioral data. A drawback is, that customer segmentation is done from the marketing side only.

Paper [22] discusses the possibilities of how big data affects retailers, including the five main dimensions of data – customer, product, time, (geospatial) location and channel. In order to study the use of smart technologies in terms of marketing and retail, this paper provides the approach to disciplines that are driven by continuous improvement of technology [23]. Its focus is aimed on smart technologies for consumer behavior through a more integrated and modern perspective that combines marketing and retail with other disciplines such as psychology, media research and sociology. The authors carried out the analysis of possible technologies that can be used to determine the consumer basket in real time, but the implementation is absent.

The paper [21] focuses on intelligent consumers who voluntarily participate in value creation activities to conceptualize joint Smart Experience Creation (SEC) and intelligent service systems.

The above-mentioned papers mostly analyze the customer in terms of supermarket marketing. There are no papers where the customer is evaluated as a consumer, the consumer basket is defined, and the living wage is found.

Therefore, the purpose of this paper is to fill the existing gap and develop the intelligent method of a consumer basket forming.

3 Proposed Method

The proposed method is illustrated schematically (Fig. 1) and represented by the following steps:

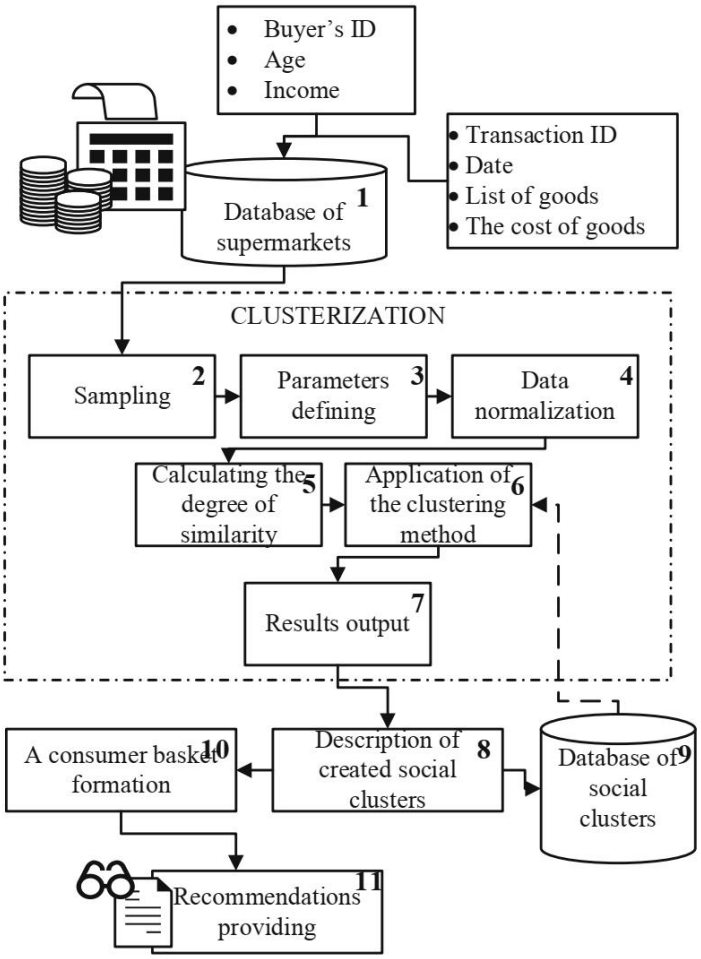


Fig. 1. Algorithmic structure of intelligent method for forming the consumer basket.

Step 1. Collection of data from supermarkets (Block 1), namely information about the customer (Customer ID, Age, Revenue) and customer's purchases (Transaction ID, Date, List of goods, Product price). When collecting data, it is possible to expand the database of pharmacies, food and non-food products.
Step 2. Data clustering.

Step 2.1. Selection of objects for clustering (Block 2).
Step 2.2. Defining the set of variables (Block 3), which will be evaluated objects in the sample.

Step 2.3. Normalization (Block 4) of variable values.
Step 2.4. Calculating the values of similarity degree between objects (Block 5).
Step 2.5. Application of clustering method (Block 6) for creating the groups of similar objects (clusters).
Step 2.6. Presentation of results analysis (Block 7).

Step 3. Description of obtained clusters, namely the definition of the social cluster (Block 8), and the transfer of existing social clusters in a separate database (Block 9) for automatization of processing and clusters formation.
Step 4. Forming the consumer basket (Block 10) and providing recommendations (Block 11).

4 Experimental Results

The Python language was selected to implement the proposed method of forming the consumer basket. Customer Personality Analysis data from the Kaggle platform was used as input [18]. This dataset presents the analysis of customer's personality as the US dollars amount spent over 2 years (Table 1).

Table 1. Dataset structure.

Parameter	count	mean	min	max
Income	2240.0	52247.25	1730.0	666666.0
Kidhome	2240.0	0.44	0.0	2.0
Teenhome	2240.0	0.51	0.0	2.0
Recency	2240.0	49.11	0.0	99.0
Wines	2240.0	303.94	0.0	1493.0
Fruits	2240.0	26.30	0.0	199.0
Meat	2240.0	166.95	0.0	1725.0
Fish	2240.0	37.52	0.0	259.0
Age	2240.0	52.19	25.0	68.0
Total_Spend	2240.0	602.25	5.0	2525.0

Normalization must be run before clustering. We are also interested in knowing the costs per month, so we divide all the cos t values by 24 months. This will get the values ready to form a consumer basket.

Several methods of clustering [9, 11, 13, 15–17, 19] were analyzed and method of k-means clustering was selected [15, 19] as a common and simplest one.

To define the number of clusters, the Elbow method is used [20]. This method is calculated as the square sum of distances of the points from the center of the cluster according to each value of k – a number of clusters. Then a graph is plotted per each

value of k (Fig. 2). The point of the elbow on the graph, where the difference between the results begins to decrease, is defined as the most appropriate value of k. For our example, the optimal number of clusters is three (see Fig. 2).

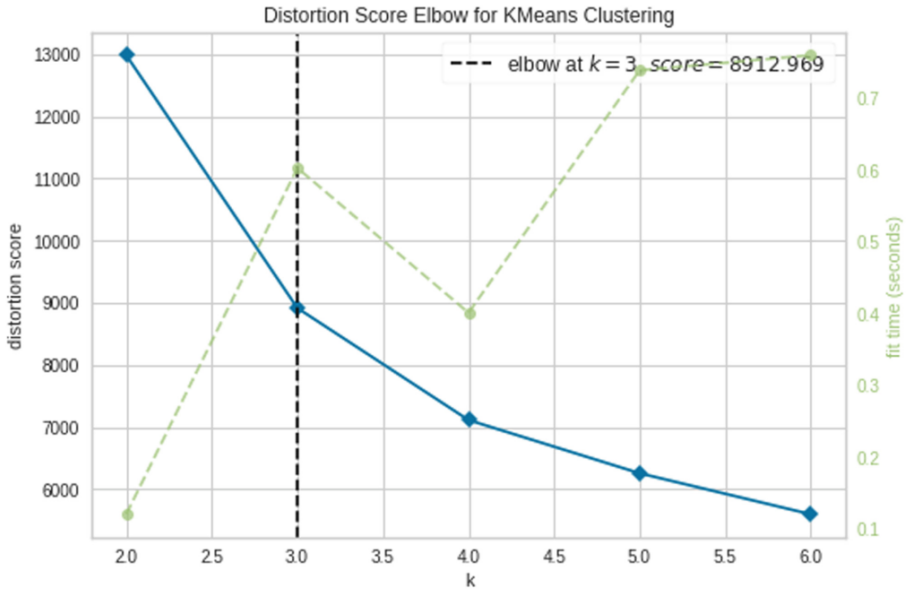


Fig. 2. Selection of clusters number by the Elbow method.

Next, let's consider the visualization of the obtained three clusters. As can be seen (Fig. 3a) the cluster number zero counts the most customers. The interpretation of the obtained clusters (Fig. 3b) is made in accordance with the characteristics of income and total costs:

- Cluster 2: high costs – average income;
- Cluster 1: high costs – high income;
- Cluster 0: low costs – low income.

If we consider clusters as social, then it is important that Cluster 2 and Cluster 0 are taken into account first. So, let's focus on the of the consumer basket formation, using these two clusters. Namely, to form a consumer basket based on Cluster 2 for an economically stable situation, and the consumer basket based on Cluster 0 for crises.

When considering the customers age category of (Fig. 4), it is seen that the distribution of clusters is uniform. Therefore, the authors believe that age does not affect purchasing power.

Next, let us consider in more detail the clusters relation to the amount (USD) of purchased goods (Fig. 5) for 2 years. Boxplot charts show the best the point of determining the average number of purchased goods in the cluster. As it is seen from Fig. 5 the Cluster 1 spends just a little money on fish, meat, sweets, wine, and the Cluster 2 shows larger amounts spent on these products. If we consider technological products,

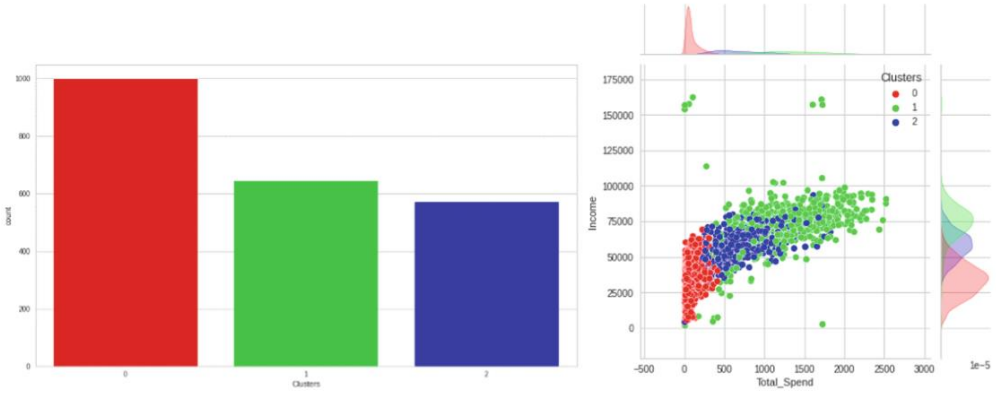


Fig. 3. Visualization of clusters: (a) number of customers in clusters; (b) distribution of customer income and expenses.

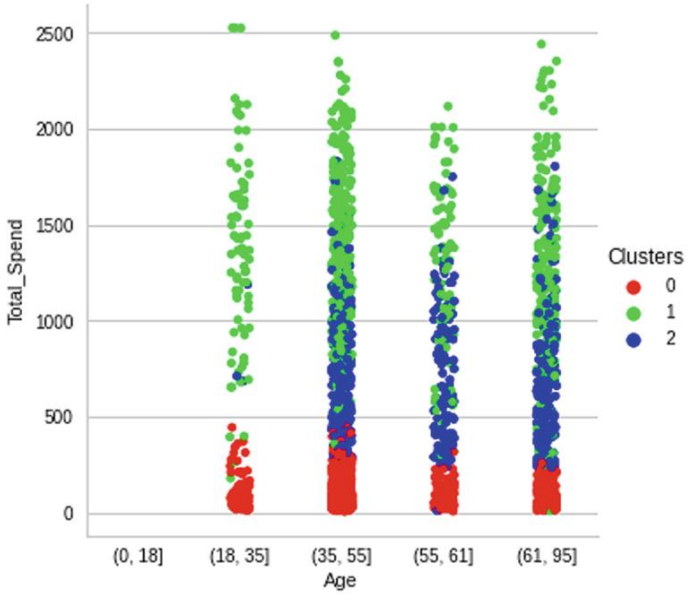


Fig. 4. Clusters visualization: age distribution.

then all customers spend approximately the same costs (Fig. 5e). Hence, in terms of goods sales, the authors believe that Cluster 2 has the optimal values for the consumer basket formation.

Therefore, let us describe Cluster 2 in more details, and use it as the base of recommendations for the consumer basket formation. The consumer basket (Cluster 2) includes 27% of middle-income customers with the different age and counts following goods purchased for the month: fish in the amount of 25 USD; meat in the amount of 100 USD; sweets in the amount of 20 USD; wine in the amount of 400 USD; equipment in the amount of 50 USD.

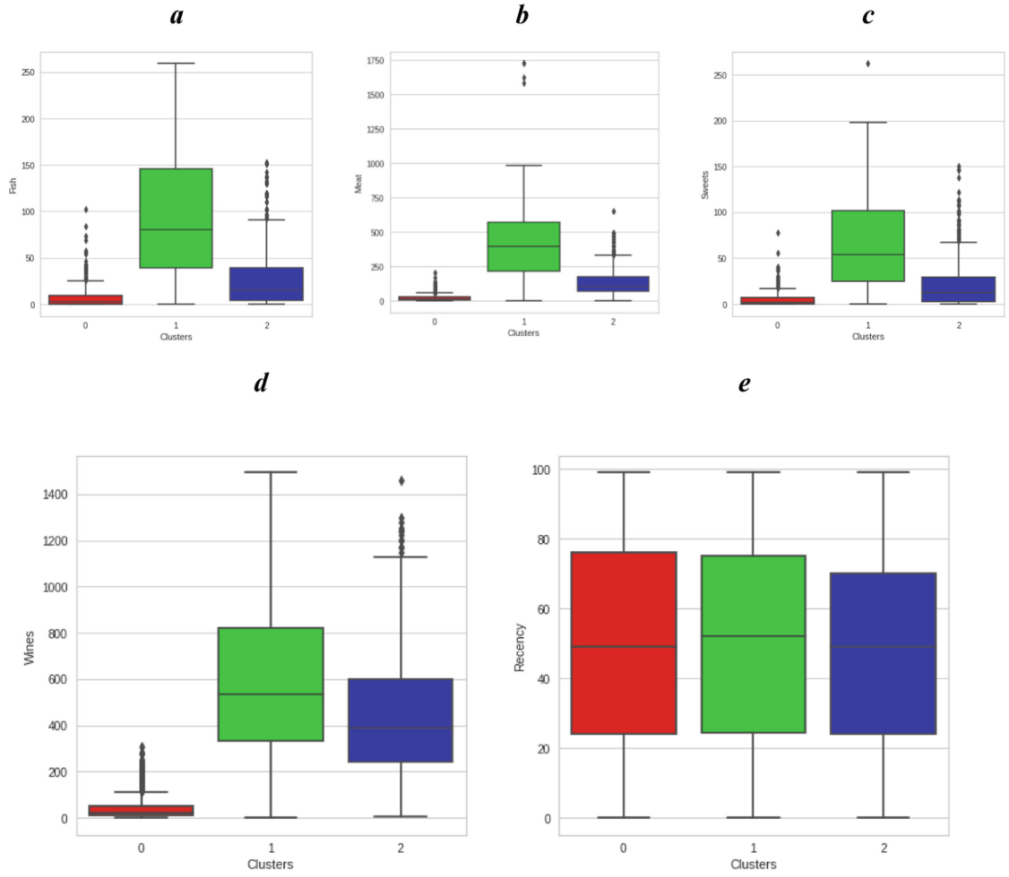


Fig. 5. Clusters visualization by distributing the purchased goods: (a) fish; (b) meat; (c) sweets; (d) wine; (e) equipment.

5 Discussion

To test the proposed method, authors have used training dataset Customer Personality Analysis, which was created in September 2021. In comparison with analogues [9, 11, 16] the developed method allows to form the set of goods in the consumer basket and define the living wage. The last includes 27% of middle-aged customers of various ages, and can count such goods (purchased for the month) as fish in the amount of 25 USD; meat in the amount of 100 USD; sweets in the amount of 20 USD; wine in the amount of 400 USD; equipment in the amount of 50 USD.

Authors realize that the proposed method allows working in real time, therefore changes of market situation will influence on result correspondingly. Authors believe that a full data collection from online supermarkets can give results that are more accurate. This can provide the monitoring of people's purchasing power in real time. Therefore, the authors are going to investigate the method of predicting the changes in the consumer basket formation in the future research.

6 Conclusions

The proposed intelligent method of forming the consumer basket is based on data from supermarket networks enabling to change quickly the set of goods in the consumer basket and define the living wage.

As the input, the Customer Personality Analysis from the Kaggle platform was used presenting the analysis of the customer's personality. To implement a posed task, the method of clustering k-means was selected. The analysis of clusters, relative to the amount (USD) of purchased goods for 2 years, was made, and the cluster (consumer basket) was defined. It includes the 27% of middle-aged customers of various ages and counts such goods as fish, meat, sweets, wine and equipment.

The proposed method allows working in real time, therefore changes of market situation will influence on result correspondingly.

Unlike analogues, the developed method allows to form the set of goods in the consumer basket and define the living wage, which may promote survival during humanitarian and economic disasters, especially in times of economic crisis (war, pandemic).

In the future research, we plan to investigate the method of predicting the changes in the consumer basket formation. In addition, the cross-cluster analysis will be performed with different approaches [24–26] to determine the best one.

References

1. Kapała, A.M.: Legal instruments to support short food supply chains and local food systems in France. Laws **11**(2), 21 (2022). https://doi.org/10.3390/laws11020021
2. Govindasamy, R.: Cluster analysis of wine market segmentation – a consumer based study in the mid-Atlantic USA. Econ. Aff. **63**(1), 151–157 (2018)
3. Qisman, M., Rosadi, R., Abdullah, A.S.: Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores). In: Journal of Physics: Conference Series. IOP Publishing, vol. 1722 no. 1, p. 012020 (2021). https://doi.org/10.1088/1742-6596/1722/1/012020
4. Tatiana, K., Mikhail, M.: Market basket analysis of heterogeneous data sources for recommendation system improvement. Procedia Comput. Sci. **136**, 246–254 (2018). https://doi.org/10.1016/j.procs.2018.08.263
5. Sun, Q., Gao, X., Wang, Z., Liu, S., Guo, S., Li, Y.: Quantifying the risk of price fluctuations based on weighted granger causality networks of consumer price indices: evidence from G7 countries. J. Econ. Interac. Coord. **15**(4), 821–844 (2019). https://doi.org/10.1007/s11403-019-00273-2
6. Sarantitis, G.A., Papadimitriou, T., Gogas, P.: A network analysis of the United Kingdom's consumer price index. Comput. Econ. **51**(2), 173–193 (2016). https://doi.org/10.1007/s10614-016-9625-9
7. Rathnayaka, S.D., Selvanathan, E.A., Selvanathan, S.: Modelling the consumption patterns in the Asian countries. Econ. Anal. Policy **74**, 277–296 (2022). https://doi.org/10.1016/j.eap.2022.02.004
8. Peixoto, V., Peixoto, H., Machado, J.: Integrating a data mining engine into recommender systems. In: Analide, C., Novais, P., Camacho, D., Yin, H. (eds.) IDEAL 2020. LNCS, vol. 12489, pp. 209–220. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62362-3_19

9. Dogan, O., Hiziroglu, A., Seymen, O.F.: Segmentation of retail consumers with soft clustering approach. In: Kahraman, C., Cevik Onar, S., Oztaysi, B., Sari, I.U., Cebi, S., Tolga, A.C. (eds.) INFUS 2020. AISC, vol. 1197, pp. 39–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-51156-2_6
10. Chiang, L.L.L., Yang, C.S.: Does country-of-origin brand personality generate retail customer lifetime value? A big data analytics approach. Technol. Forecast. Soc. Change **130**, 177–187 (2018). https://doi.org/10.1016/j.techfore.2017.06.034
11. Munusamy, S., Murugesan, P.: Modified dynamic fuzzy c-means clustering algorithm – application in dynamic customer segmentation. Appl. Intell. **50**(6), 1922–1942 (2020). https://doi.org/10.1007/s10489-019-01626-x
12. Arunachalam, D., Kumar, N.: Benefit-based consumer segmentation and performance evaluation of clustering approaches: an evidence of data-driven decision-making. Expert Syst. Appl. **111**, 11–34 (2018). https://doi.org/10.1016/j.eswa.2018.03.007
13. Anitha, P., Patil, M.M.: RFM model for customer purchase behavior using k-means algorithm. J. King Saud Uni. – Comput. Inform. Sci. **34**(5), 1785–1792 (2022). https://doi.org/10.1016/j.jksuci.2019.12.011
14. Lipyanina, H., Sachenko, A., Lendyuk, T., Nadvynychny, S., & Grodskyi, S.: Decision tree based targeting model of customer interaction with business page. In: Proceedings of the third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), CEUR Workshop Proceedings, (2608), pp. 1001–1012 (2020). Electronic copy at: http://ceur-ws.org/Vol-2608/paper75.pdf
15. Kanavos, A., Iakovou, S.A., Sioutas, S., Tampakas, V.: Large scale product recommendation of supermarket ware based on customer behaviour analysis. Big Data Cogn. Comput. **2**(2), 11 (2018). https://doi.org/10.3390/bdcc2020011
16. Alam, M.F., Singh, R., Katiyar, S.: Customer segmentation using k-means clustering in unsupervised machine learning. In: Proceedings of the 2021 3rd IEEE International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 94–98 (2021). https://doi.org/10.1109/ICAC3N53548.2021.9725644
17. Sinaga, K.P., Yang, M.: Unsupervised k-means clustering algorithm. IEEE Access **8**, 80716–80727 (2020). https://doi.org/10.1109/ACCESS.2020.2988796
18. Customer Personality Analysis. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis. Accessed 10 Apr 2022
19. Fränti, P., Sieranoja, S.: K-means properties on six clustering benchmark datasets. Appl. Intell. **48**(12), 4743–4759 (2018). https://doi.org/10.1007/s10489-018-1238-7
20. Marutho, D., Hendra Handaka, S., Wijaya, E., Muljono: The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In: Proceedings of the 2018 IEEE International Seminar on Application for Technology of Information and Communication, pp. 533–538 (2018). https://doi.org/10.1109/ISEMANTIC.2018.8549751
21. Roy, S.K., Singh, G., Hope, M., Nguyen, B., Harrigan, P.: The rise of smart consumers: role of smart servicescape and smart consumer experience co-creation. J. Mark. Manag. **35**(15–16), 1480–1513 (2019). https://doi.org/10.1080/0267257X.2019.1680569
22. Bradlow, E.T., Gangwar, M., Kopalle, P., Voleti, S.: The role of big data and predictive analytics in retailing. J. Retail. **93**(1), 79–95 (2017). https://doi.org/10.1016/j.jretai.2016.12.004
23. Pantano, E.: The role of smart technologies in decision making: developing, supporting and training smart consumers. J. Mark. Manag. **35**(15–16), 1367–1369 (2019). https://doi.org/10.1080/0267257X.2019.1688927
24. Komar, M., Golovko, V., Sachenko, A., Bezobrazov, S.: Development of neural network immune detectors for computer attacks recognition and classification. In: Proceedings of the 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced

Computing Systems (IDAACS), pp. 665–668 (2013). https://doi.org/10.1109/IDAACS.2013.6663008

25. Hu, Z., Bodyanskiy, Y.V., Kulishova, N.Y., Tyshchenko, O.K.: A multidimensional extended neo-fuzzy neuron for facial expression recognition. Int. J. Intell. Syst. Appl. (IJISA) **9**(9), 29–36 (2017). https://doi.org/10.5815/ijisa.2017.09.04
26. Turchenko, V., Chalmers, E., Luczak, A.: A deep convolutional auto-encoder with pooling – unpooling layers in Caffe. Int. J. Comput. (18), 8--31 (2019). https://doi.org/10.47839/ijc.18.1.1270.
27. Lipyanina, H., Sachenko, S., Lendyuk, T., Brych, V., Yatskiv, V., Osolinskiy, O.: Method of detecting a fictitious company on the machine learning base. In: Hu, Z., Petoukhov, S., Dychka, I., He, M. (eds.) ICCSEEA 2021. LNDECT, vol. 83, pp. 138–146. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80472-5_12
28. Lipyanina-Goncharenko, H., Brych, V., Sachenko, S., Lendyuk, T., Bykovyy, P., Zahorodnia, D.: Method of forming a training sample for segmentation of tender organizers on machine learning basis. In: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), Volume I: Main Conference, Lviv, Ukraine, 22–23 April 2021, CEUR Workshop Proceedings, 2870, pp. 1843–1852 (2021). http://ceur-ws.org/Vol-2870/paper134.pdf

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДА МОЛОДИХ ВЧЕНИХ

Збірник тез доповідей
XIX Міжнародної
науково-практичної
конференції молодих вчених

ЕКОНОМІЧНИЙ І СОЦІАЛЬНИЙ РОЗВИТОК УКРАЇНИ В XXI СТОЛІТТІ: НАЦІОНАЛЬНА ВІЗІЯ ТА ВИКЛИКИ ГЛОБАЛІЗАЦІЇ

Тернопіль
ЗУНУ
2022

ЕКОНОМІЧНИЙ І СОЦІАЛЬНИЙ РОЗВИТОК УКРАЇНИ В XXI СТОЛІТТІ: НАЦІОНАЛЬНА ВІЗІЯ ТА ВИКЛИКИ ГЛОБАЛІЗАЦІЇ

Збірник тез доповідей XIX Міжнародної науково-практичної конференції молодих вчених (Тернопіль, 13 травня 2022 року)

Збірник тез доповідей укладено за матеріалами XIX Міжнародної науково-практичної конференції молодих вчених «Економічний і соціальний розвиток України в XXI столітті: національна візія та виклики глобалізації», яка відбулася на базі Західноукраїнського національного економічного університету.

За зміст наукових праць та достовірність наведених фактологічних і статистичних матеріалів відповідальність несуть автори.

Секція 5

Міжнародні стандарти та проблеми уніфікації обліку, аналізу і аудиту.

ЛІТЕРАТУРА

1. МВФ оприлюднив тексти Листа про наміри та Меморандуму за результатами завершення першого перегляду Програми «Стенд-бай» для України. *Єдиний веб-портал органів виконавчої влади України.* URL: https://www.kmu.gov.ua/news/mvf-oprilyudniv-teksti-lista-pro-namiri-ta-memorandumu-za-rezultatami-zavershennya-pershogo-pereglyadu-programi-stend-baj-dlya-ukrayini (дата звернення: 11.04.2022).

2. Про Бюро економічної безпеки України: Закон України за редакцією від 01.01.2022р. №1089-IX. URL: https://zakon.rada.gov.ua/laws/show/1150-20#Text (дата звернення: 28.04.2022).

3. БЕБ змінило назву і розширило функціонал свого телеграм боту . *Офіційний веб-портал Бюро економічної безпеки України.* URL: https://esbu.gov.ua/news/beb-zminilo-nazvu-i-rozshirilo-funkcional-svogo-telegram-botu (дата звернення: 09.04.2022).

4. Офіційний веб-портал Бюро економічної безпеки України. URL: https://esbu.gov.ua/ (дата звернення: 28. 04. 2022).

5. Tesliuk S., Demchuk I., Zvirko N. Efficiency of the Bureau of economic security of Ukraine under martial law. Multidisciplinární mezinárodní vědecký magazín "Věda a perspektivy" je registrován v České republice. Státní registrační číslo u Ministerstva kultury ČR: E 24142. № 4(11) 2022. Str. 33-42

УДК 004.891

Чижовська З. І.
магістр спеціальності «Комп'ютерні науки»
Західноукраїнського національного університету
Ліп'яніна-Гончаренко Х.В.
к.т.н., доцент кафедри
інформаційно-обчислювальних систем і управління
Західноукраїнського національного університету
Саченко А.О.
д.т.н., професор кафедри
інформаційно-обчислювальних систем і управління
Західноукраїнського національного університету

ІНТЕЛЕКТУАЛЬНИЙ МЕТОД ФОРМУВАННЯ СПОЖИВЧОГО КОШИКУ

На сьогодні методика обчислення споживчого кошика є надзвичайно актуальним питанням, оскільки на основі споживчого кошика формується прожитковий мінімум, згідно з яким розраховують пенсії, соціальні пільги та виплати. В Україні формуванням споживчого кошика безпосередньо займається Кабінет Міністрів України. На основі даних споживчого кошика та споживчого бюджету розраховуються такі показники, як мінімальна заробітна плата та мінімальна пенсія.

Вартість споживчого кошика залежить від рівня роздрібних цін на товари та тарифів на платні послуги (наприклад, комунальні платежі). Така практика відома у всьому цивілізованому світі. Із кожного виду потреб до розрахунку включають придбання відносно дешевих товарів, як правило, за державними фіксованими цінами. Якщо, наприклад, на ринку даний продукт або послуга продається за більш низькими цінами, то за основу береться найнижчий рівень.

Поняття споживчого кошика існує у багатьох країнах світу. Його ціна і національні особливості в кожній країні різні. Наприклад, споживчий кошик американця нараховує 350 продуктів і послуг, француза – 507, англійця – 350, німця – 475. Український споживчий кошик нещодавно було розширено до 297 найменувань.

Купівельна спроможність людей може різко змінюватись відносно кризових ситуацій, таких як війна, пандемії. Тому оцінка споживчого кошика і його зміна повинна бути швидкою, для того, щоб не допустити гуманітарних катастроф та збільшення бідності.

У зв’язку з цим, можна вважати, що розробка інтелектуального методу формування споживчого кошику на основі даних мереж супермаркетів є актуальною та дає можливість оперативно змінювати набір товарів у споживчому кошику та формувати прожитковий мінімум, що дозволить запобігти можливим гуманітарно-економічним катастрофам та збільшення бідності.

У дослідженні [1] був застосований інтуїтивний алгоритм нечіткої кластеризації до даних покупців у супермаркеті відповідно до суми, витраченої на деякі групи продуктів. У дослідженні [2] запропонували модифікації алгоритму CDFCM через виявлені в ньому певні недоліки, в результаті яких утворився модифікований динамічний нечіткий с-алгоритм середніх (MDFCM). У дослідженні [3] проведена сегментація клієнтів за допомогою кластеризації K-середніх у неконтрольованому машинному навчанні. Проведено класифікацію цільових клієнтів за допомогою кластеризації на основі демографічної інформації (наприклад, стать, вік, дохід тощо), географічної інформації, психографії та поведінкових даних.

Згадані вище роботи здебільшого аналізують клієнта зі сторони маркетингу супермаркетів. Проте, немає робіт, які б оцінювали клієнта як споживача та формували його соціальний портрет, на основі якого можна сформувати споживчий кошик та відповідно прожитковий мінімум.

На відміну від аналогів [1, 2, 3] розроблений інтелектуальний метод формування споживчого кошику дозволить на основі даних мереж супермаркетів формувати набір товарів у споживчому кошику та встановлювати прожитковий мінімум, що допоможе запобігти можливим гуманітарно-економічним катастрофам та збільшенню бідності, особливо у кризовий період економіки (війна, пандемія).

Для оперативної зміни набору товарів у споживчому кошику авторами розроблено інтелектуальний метод формування споживчого кошику. Запропонований метод ілюструється схематично (Рис.1) та представлений наступними кроками:

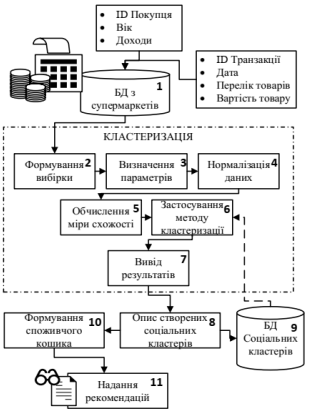


Рис.1. Структура інтелектуального методу формування споживчого кошику

Крок 1. Збір даних з супермаркетів (Блок 1), а саме інформацію про покупця (ID Покупця, Вік, Доходи) та про його покупки (ID Транзакції, Дата, Перелік товарів, Вартість товару). При зборі даних, можна розширити базу даними з аптек, продовольчих та непродовольчих товарів.

Крок 2. Кластеризація даних.

Крок 2.1. Вибір об'єктів для кластеризації (Блок 2).

Крок 2.2. Визначення безлічі змінних (Блок 3), якими будуть оцінюватися об'єкти у вибірці.

Крок 2.3. Нормалізація (Блок 4) значень змінних.

Крок 2.4. Обчислення значень міри схожості між об'єктами (Блок 5).

Крок 2.5. Застосування методу кластерного аналізу (Блок 6) створення груп подібних об'єктів (кластерів).

Крок 2.6. Подання результатів аналізу (Блок 7).

Крок 3. Проведення опису отриманих кластерів, а саме визначення соціального кластеру (Блок 8). Також, передача сформованих соціальних кластерів в окрему базу даних (Блок 9), що дасть можливість в подальшому максимально автоматизувати процес обробки та формування кластерів. Також, в подальших наукових дослідженнях на основі отриманих соціальних кластерів та методів машинного навчання буде розроблено метод передбачення змін у формуванні споживчого кошика.

Крок 4. Формування споживчого кошика (Блок 10) та надання рекомендацій (Блок 11).

ЛІТЕРАТУРА

1. Dogan, O., Hiziroglu, A., Seymen, O.F. (2021). Segmentation of Retail Consumers with Soft Clustering Approach. In: Kahraman, C., Cevik Onar, S., Oztaysi, B., Sari, I., Cebi, S., Tolga, A. (eds) Intelligent and Fuzzy Techniques: Smart and Innovative Solutions. INFUS 2020. Advances in Intelligent Systems and Computing, vol 1197. Springer, Cham. https://doi.org/10.1007/978-3-030-51156-2_6