КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Катедра теорії та технології програмування

Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 122 Комп'ютерні науки

на тему:

ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ ФАКТОРІВ ВПЛИВУ НА РІВЕНЬ ЗАДОВОЛЕНОСТІ ЖИТТЯМ В УКРАЇНІ

Виконала студентка 4 курсу
Євтушенко Олександра Олександрівна

Науковий керівник:
доцент, кандидат фіз.-мат. наук
Лівінська Ганна Володимирівна

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент

Роботу розглянуто й допущено до захисту на засіданні катедри теорії та технології програмування
«01» червня 2022р.
протокол № 10

Завідувач катедри
Нікітченко М.С.

Київ – 2022

РЕФЕРАТ

Обсяг роботи: 48 сторінок, 45 ілюстрацій, 1 таблиця, 15 джерел посилань.

РІВЕНЬ ЗАДОВОЛЕНОСТІ ЖИТТЯМ, СОЦІОЛОГІЧНІ ДАНІ, МАШИННЕ НАВЧАННЯ, РЕГРЕСІЯ, КЛАСТЕРИЗАЦІЯ, LINEAR REGRESSION, K-MEANS, DBSCAN, GAUSSIAN MIXTURE MODEL.

Об’єктом роботи є аналіз соціологічних даних з використанням методів машинного навчання, спрямований на визначення факторів, що впливають на рівень задоволеності життям українців.

Метою роботи є визначення факторів, що впливають на рівень задоволеності життям в Україні за допомогою моделей машинного навчання.

Методи розробки: методи оцінки рівня задоволеності життям суспільства, методи попередньої обробки даних, розробка моделей машинного навчання.

Інструменти розробки: мови програмування Python та R, безкоштовні, вільно поширювані інтегровані середовища розробки Jupyter Notebook та RStudio.

Результати роботи: досліджено теоретичні підходи до визначення факторів задоволеності життям, знайдено дані відкритих джерел про стан українського суспільства, проведено розвідувальний аналіз та підготовку даних для створення моделей, побудовано моделі машинного навчання на загальній вибірці та на кожному з кластерів окремо.

Результати виконаного аналізу можуть бути корисними для спеціалістів з різних галузей: соціологів, психологів, політологів, економістів для будь-якої роботи, спрямованої на покращення рівня задоволеності життям певних осіб.

ЗМІСТ

ВСТУП

Оцінка сучасного стану об'єкта розробки. Протягом останніх років іноземні дослідники активно працюють над вивченням факторів впливу на суб'єктивну оцінку вдоволеності життям громадян. І попит на такі дослідження продовжує зростати. Знання про те, що робить життя людей комфортнішим використовуються соціальними працівниками та психологами, допомагаючи ефективніше працювати з вразливими верствами населення, політиками, задля прийняття обґрунтованих рішень, які краще сприймаються суспільством, економістами, для визначення напрямків розвитку фінансової та соціальної сторони життя.

Актуальність роботи та підстави для її виконання. У результаті згаданих вище досліджень було виявлено, що для кожної країни фактори, які впливають на задоволеність життям відрізняються. Це є зрозумілим, враховуючи унікальні особливості кожної країни у економічному, географічному чи культурному аспекті. Тому виконання такого аналізу спеціально для українського суспільства є актуальним.

Мета й завдання роботи. Метою роботи є визначення факторів, що впливають на рівень задоволеності життям в Україні за допомогою моделей машинного навчання.

Для досягнення мети було поставлено такі завдання:

- пошук та отримання даних відкритих джерел про стан українського суспільства
- проведення розвідувального аналізу та підготовка даних для побудови моделі
- побудова моделей машинного навчання

Об'єкт, методи й засоби дослідження. Об'єктом дослідження є аналіз соціологічних даних за допомогою багатовимірних статистичних методів та методів машинного навчання. Перед початком аналізу було розглянуто теоретичні підходи до визначення показника рівня задоволеності життям. У якості

інструментів розробки було обрано мови програмування Python та R. У якості інтегрованого середовища використано Jupyter Notebook та RStudio відповідно.

Можливі сфери застосування. Результати такого дослідження можуть бути корисними для спеціалістів з різних галузей: соціологів, психологів, політологів, економістів для будь-якої роботи, спрямованої на покращення рівня самопочуття певних осіб та суспільства в цілому.

РОЗДІЛ 1. ЗБІР ДАНИХ

Протягом останніх десятиліть вчені проводять всебічні опитування населення задля виявлення факторів, що можуть суттєво впливати на оцінку свого життя як «добре» чи «погано». Існує досить багато опитувань, які прямо чи опосередковано стосуються цієї теми, але, на жаль, лише в деяких з них брали участь жителі України.

Тому, для виконання цієї роботи було використано дані World Values Survey (WVS) – міжнародного опитування, метою якого є вивчення соціальних, культурних, етнічних, політичних та економічних цінностей людей з більш ніж 120 різних країн. Вперше опитування було проведено в 1981 році за сприяння американського політолога та соціолога Рональда Інглхарта з метою вивчення соціально-культурних відмінностей і вже протягом 30 років результати опитування активно використовуються у соціологічних, економічних, етнографічних та інших дослідженнях. Дані опитування наявні у вільному доступі після попередньої реєстрації на офіційному сайті WVS [1]. Станом на 2022 рік, найсвіжішими є дані сьомої хвилі 2017-2022 років, яка охопила близько 80 країн. Для України опитування сьомої хвилі проводилось у 2020 році.

Рисунок 1.1 – Офіційна сторінка WVS7

Опитування WVS-7 є комплексним дослідницьким інструментом, що складається з 290 питань і вимірює культурні цінності, ставлення та переконання

щодо статі, сім'ї та релігії, ставлення та досвід бідності, освіти, здоров'я та безпеки, соціальну толерантність та довіру, тощо. Крім того, порівняно з минулими хвилями, WVS-7 включає нові теми, такі як правосуддя, етичні принципи, корупція, ризики, імміграція, національна безпека та глобальне управління.

Загалом анкета представлена чотирнадцятьма тематичними розділами, серед яких:

- соціальні цінності, установки та стереотипи (45 питань)
- суб'єктивне вдоволення життям (11 питань)
- соціальний капітал, довіра та членство в організаціях (49 питань)
- економічні цінності (6 питань)
- корупція (9 питань)
- міграція (10 питань)
- постматеріалістичний покажчик (6 питань)
- наука і техніка (6 питань)
- релігійні цінності (12 питань)
- безпека (21 питання)
- етичні цінності та норми (23 питання)
- політичні інтереси та участь у політичному житті країни (36 питань)
- політична культура та політичні режими (25 питань)
- демографія (31 питання) [2].

Повний перелік питань анкети доступний для перегляду та завантаження на офіційному сайті. Основним методом збору інформації було очне інтерв'ювання респондентів у кожному регіоні країни. Стандартний розмір вибірки по країні від 1000 до 5000 чоловік. Вік опитаних – від 18 років.

РОЗДІЛ 2. ПІДГОТОВКА ТА РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ

Наступним кроком після збору даних є розвідувальний аналіз даних та їх попередня обробка. Для дослідження було скомбіновано дані візуальних і невізуальних методів з метою поглиблення розуміння наявних даних і підготовки їх для подальшого використання в моделі.

Різні спеціалісти виділяють різну кількість та послідовність етапів розвідувального аналізу даних. Але для поточної роботи оберемо наступний порядок:

- Опис наявних змінних - необхідний етап, спрямований на детальне вивчення поданих даних, бо без повного розуміння структури даних, їх вигляду та мети, неможливо побудувати якісну модель. Також на цьому етапі вивчаються кожні змінні окремо, розраховуються статистичні показники.
- Попередня обробка даних - необхідний етап, спрямований на зведення даних до потрібної структури, придатної для обробки чи побудови моделі, також на цьому етапі проводиться очистка даних від «шуму»: пропущених, зайвих чи аномальних даних. Важливо коректно визначити способи визначення та усунення зайвих даних. Тобто, що робити з пропущеними даними, чи даними, що занадто виділяються з поточної картини.
- Візуалізація даних - важливий етап для розуміння поданих даних. Людський мозок краще сприймає інформацію, подану у вигляді малюнків, ніж у вигляді набору чисел. До того ж, різні види графіків допомагають швидко отримати інформацію про необхідні показники. Наприклад, з гістограми можна зробити висновок про вигляд розподілу, боксплот (ящик з вусами) ілюструє розмах даних, міри центральної тенденції, наявність викидів, а графік кореляції може допомогти у пошуку неявних залежностей у даних, які можуть не визначатися обрахунком коефіцієнту кореляції.

- Нормалізація даних - опціональний, але не менш важливий етап перед побудовою моделей машинного навчання. Діапазони змінних дуже часто відрізняються, що може впливати на якість моделей, бо змінні з більшим абсолютним значенням можуть сильніше впливати на вагу показника у моделі, не маючи нічого спільного з реальними факторами впливу на залежну змінну.

2.1 Опис наявних змінних

Дані, з якими проводилася робота, доступні для завантаження у форматі .csv. Загалом, датасет містить 76898 записів та 520 параметрів. Оскільки ми будемо займатися аналізом даних жителів України, то одразу відфільтруємо їх за кодом країни. Побачимо, що розмір вибірки – 1289 респондентів.

Усі дані опитування подані у шкалі Лікерта - психометричній шкалі, яка зазвичай використовується в анкетах та анкетних дослідженнях. При використанні шкали респондент оцінює рівень своєї згоди або незгоди з кожним судженням у діапазоні від «повністю згоден» до «повністю не згоден». Сума балів за кожним окремим судженням дозволяє виявити ставлення до того чи іншого питання. Передбачається, що ставлення до предмета, що вивчається, ґрунтується на простих, несуперечливих судженнях і являє собою скінчену множину значень від однієї критичної точки і до протилежної критичної точки, включаючи нейтральний елемент. Наприклад: від гуманізму до мізантропії, від релігії до атеїзму. Більшість питань градуйовані на чотири рівні, де 1 - «повністю згоден», 2 – «частково згоден», 3 – «не зовсім згоден», 4 – «зовсім не згоден» (Рисунок 2.1). Також наявні питання, які пропонують оцінити показник по шкалі від 1 до 10, де 1 – «найгірше», а 10 – «найкраще» та питання які передбачають лише дві можливі ситуації 1 - «був досвід» та 2 - «не було досвіду».

Існує декілька способів підходу до обробки даних такої шкали. Але, зі статистичної точки зору, буде правильним визначити їх як ординальні дані, та відповідно працювати, використовуючи підходи для нечислових даних.

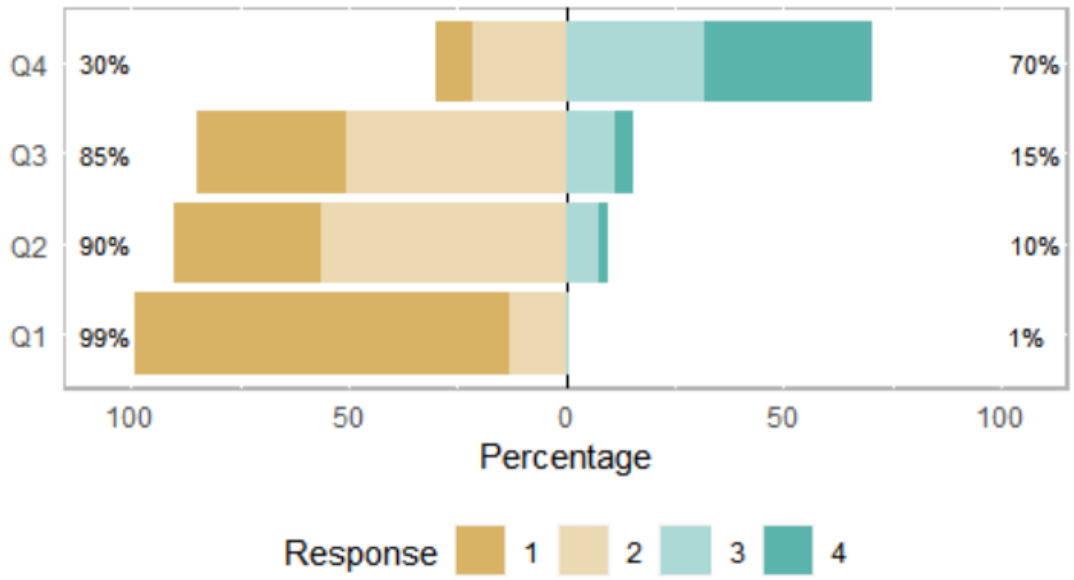

Рисунок 2.1 – Зображення деяких питань на діаграмі Лікерта

Цільовою змінною вибірки оберемо питання №49, яке сформульоване як «Оцініть по шкалі від 1 до 10, наскільки ви вдоволені своїм життям станом на сьогодні, де 1 – «зовсім не вдоволений» та 10 – «Повністю вдоволений».

На наступному етапі за допомогою python-команди info() переконаємось, що усі дані розпізнаються як числові, типу integer. Далі за допомогою python-команди describe отримали основні статистичні показники для кожного стовпчика (Рисунок 2.2):

- count – кількість записів
- mean – середнє значення змінної
- sdt – стандартне відхилення змінної
- min – мінімальне значення змінної
- 25% – 25 процентіль змінної (1 квартиль)
- 50% – 50 процентіль змінної (2 квартиль)
- 75% – 75 процентіль змінної (3 квартиль)
- max – максимальне значення змінної

	count	mean	std	min	25%	50%	75%	max
Q1	1289.0	1.147745	0.388028	1.0	1.000000	1.000000	1.000000	4.0
Q2	1289.0	1.783658	0.676730	1.0	1.000000	2.000000	2.000000	4.0
Q3	1289.0	1.847912	0.762905	1.0	1.000000	2.000000	2.000000	4.0
Q4	1289.0	2.962097	0.946233	1.0	2.000000	3.000000	4.000000	4.0
Q5	1289.0	1.872139	0.909443	1.0	1.000000	2.000000	2.000000	4.0
Q6	1289.0	2.254167	0.943052	1.0	2.000000	2.000000	3.000000	4.0
Q7	1289.0	1.444531	0.497106	1.0	1.000000	1.000000	2.000000	2.0
Q8	1289.0	1.648565	0.477604	1.0	1.000000	2.000000	2.000000	2.0
Q9	1289.0	1.283165	0.450711	1.0	1.000000	1.000000	2.000000	2.0
Q10	1289.0	1.371606	0.483421	1.0	1.000000	1.000000	2.000000	2.0
Q11	1289.0	1.878976	0.326282	1.0	2.000000	2.000000	2.000000	2.0
Q12	1289.0	1.521334	0.499739	1.0	1.000000	2.000000	2.000000	2.0
Q13	1289.0	1.607448	0.488508	1.0	1.000000	2.000000	2.000000	2.0
Q14	1289.0	1.544608	0.498199	1.0	1.000000	2.000000	2.000000	2.0
Q15	1289.0	1.851047	0.356180	1.0	2.000000	2.000000	2.000000	2.0
Q16	1289.0	1.839410	0.367294	1.0	2.000000	2.000000	2.000000	2.0
Q17	1289.0	1.666408	0.471679	1.0	1.000000	2.000000	2.000000	2.0
Q18	1289.0	1.316083	0.459134	1.0	1.000000	1.000000	2.000000	2.0
Q19	1289.0	1.733223	0.428142	1.0	1.733223	2.000000	2.000000	2.0
Q20	1289.0	1.600847	0.468743	1.0	1.000000	2.000000	2.000000	2.0

Рисунок 2.2 – Описові статистики для декількох перших питань

2.2 Попередня обробка

Під час попередньої обробки нашого набору даних було виконано такі кроки:

- Видалено стовпці, що містять технічну інформацію, наприклад дату початку опитування, його тривалість, код інтерв'юера, тощо
- Видалено питання, які з тих чи інших причин не були поставлені респондентам з України. Зазвичай це стовпці, середнє значення яких менше нуля (від'ємні значення в датасеті позначають, що значення пропущене або респондент відмовився надати відповідь)
- Знайдено пропущені значення та значення, які були відмічені як «Не знаю» / «Немає відповіді», та вилучено питання, що містять більше, ніж 20 відсотків пропущених значень, так як вони не містять корисної інформації та можуть погіршити якість моделі (Рисунок 2.3). Для всіх інших стовпчиків, пропущені значення замінені значенням моди по стовпцю. Після цього, пропусків в даних не залишилось.

У датасеті, задля полегшення програмної обробки, формулювання питань замінені відповідними номерами питань. Точні формулювання знаходяться на офіційному сайті за посиланням [1].

Q48	3.49
Q49	1.94
Q50	1.47
Q51	2.09
Q52	1.47
Q53	1.24
Q54	2.09
Q55	2.64
Q56	5.04
Q57	2.56
Q58	1.09
Q59	1.40
Q60	1.32
Q61	5.74
Q62	17.15
Q63	16.52
Q64	4.19
Q65	4.81
Q66	5.59
Q67	3.34
Q68	27.70
Q69	6.05

Рисунок 2.3 – Відсоток пропущених значень для деяких стовпців

- Питання 261 (рік народження респондента) та 262 (вік респондента) подані у вигляді абсолютних значень, що через значно більший розмах може впливати на якість моделі, тому стовпчик 261 було вилучено, а стовпчик 262 представлено у вигляді категоріальної змінної з чотирма значеннями, де:

  а) значення «1» - від 18 до 24 років

  б) значення «2» - від 25 до 54 років

  в) значення «3» - від 55 до 64 років

  г) значення «4» - від 65 років

За біологічною ознакою виділяють лише три вікові групи населення: діти, молодь і дорослі, люди похилого віку. Але експерти соціально-демографічних досліджень ООН зазвичай виділяють п’ять основних вікових груп за працездатністю: діти, особи раннього працездатного віку, особи основного

працездатного віку, особи зрілого працездатного віку та літні люди. Оскільки, в нашій вибірці представлені респонденти віком від 18 років, залишимо чотири вікові групи населення.

2.3 Візуалізація даних

Для подальшої оцінки адекватності відібраних факторів було побудовано візуалізації, що описують залежності рівня задоволеності життям від різних параметрів.

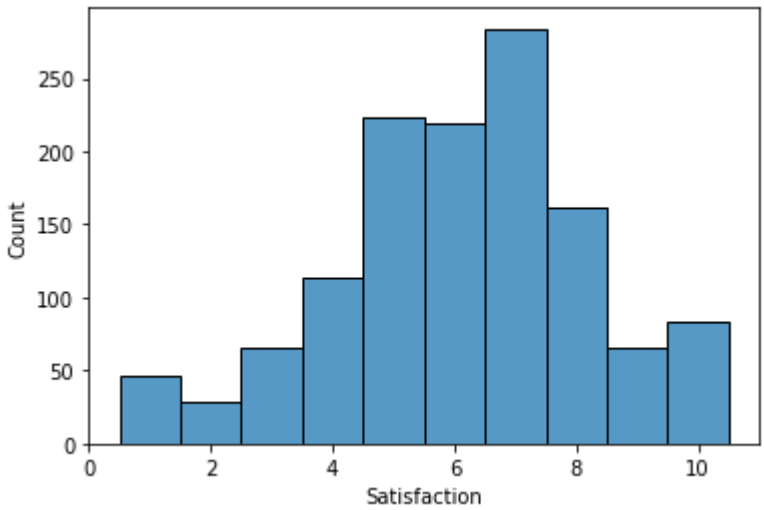

Рисунок 2.4 – Розподіл оцінок рівня задоволеності життям в Україні

Видно, що дані розподілені приблизно нормально. Найчастіша оцінка – «7», більшість значень знаходяться у проміжку від «5» до «8». Середній показник - 6,18.

Більш детально вивчимо структуру опитаного населення. Побудуємо секторну діаграму статевого розподілу опитаних (Рисунок 2.5). Видно, що значна більшість опитаних – жінки. Це варто буде врахувати при подальшому аналізі та обробці висновків.

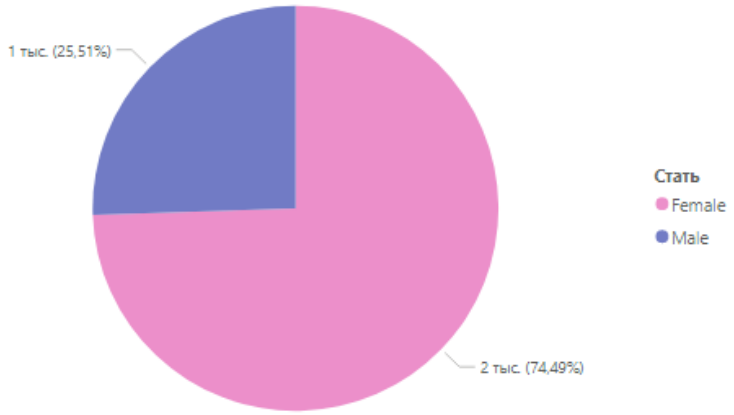

Рисунок 2.5 – Розподіл опитаних залежно від статі

Далі було побудовано гістограму вікового розподілу населення (Рисунок 2.6). В цьому випадку, дані розподілені приблизно нормально.

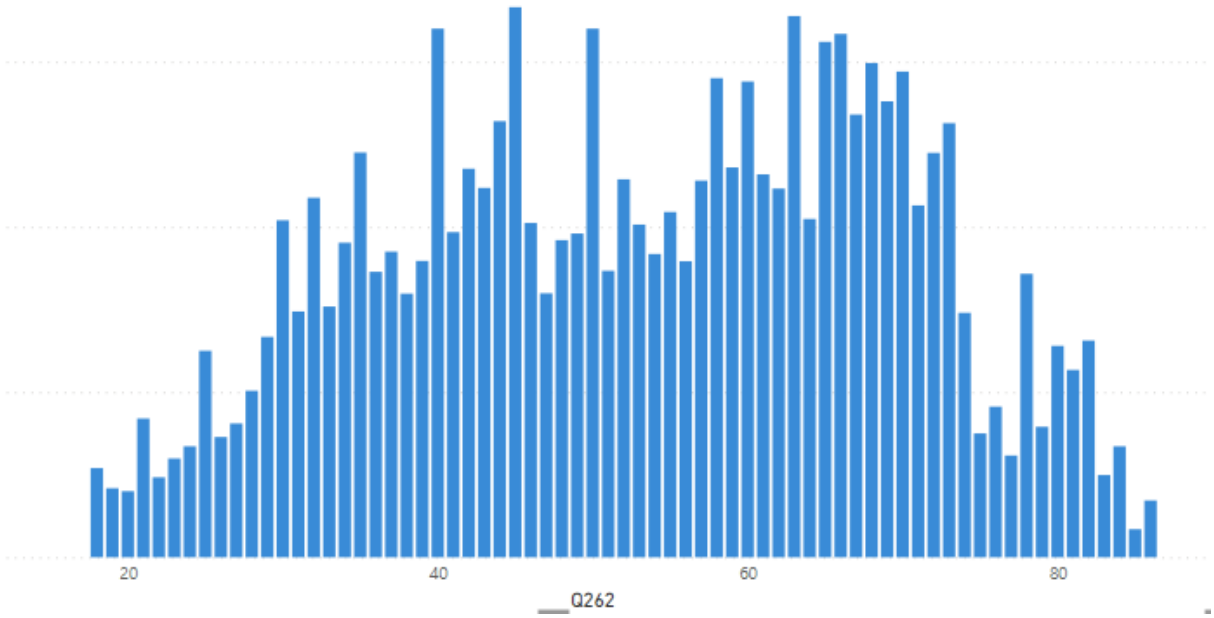

Рисунок 2.6 – Вікова структура опитаного населення

Дослідивши інформацією джерел [3, 4], присвячених вивченню факторів впливу та оцінки суб’єктивного рівня вдоволеності життям побудовано графіки (боксплоти) залежності цієї оцінки від віку (Рисунок 2.7), статі (Рисунок 2.8), економічно-соціального становища (Рисунок 2.9), рівня освіти (Рисунок 2.10),

сімейного стану (Рисунок 2.11).

Варто відмітити, що люди молодого та середнього віку (18-50 років) в середньому на 1-2 одиниці шкали більше задоволені своїм життям, ніж старші люди. Середні значення в групах 1 та 2 приблизно однакові, але респонденти з групи 1 майже не ставили оцінок від одного до трьох, що видно з розмаху даних на графіку.

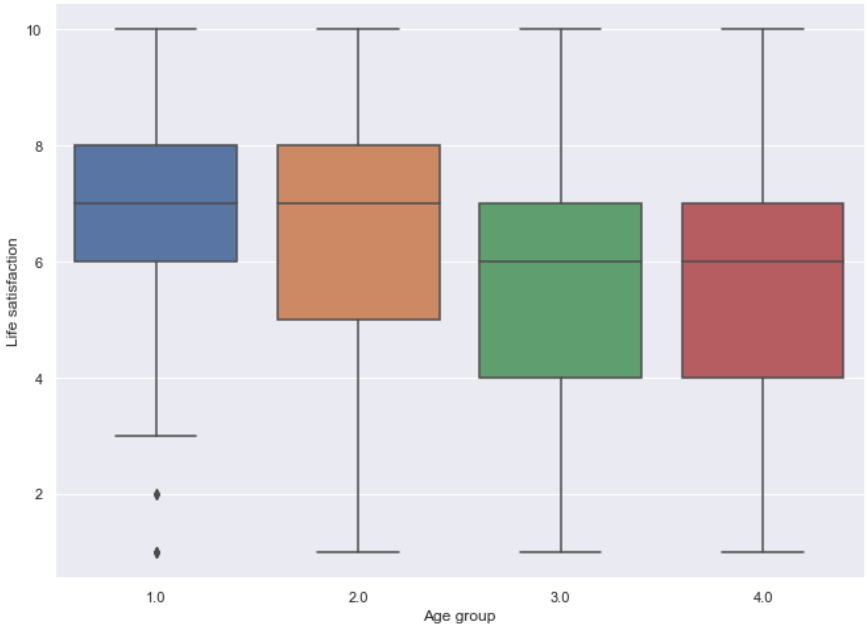

Рисунок 2.7 – Рівень задоволеності життям залежно від вікової групи

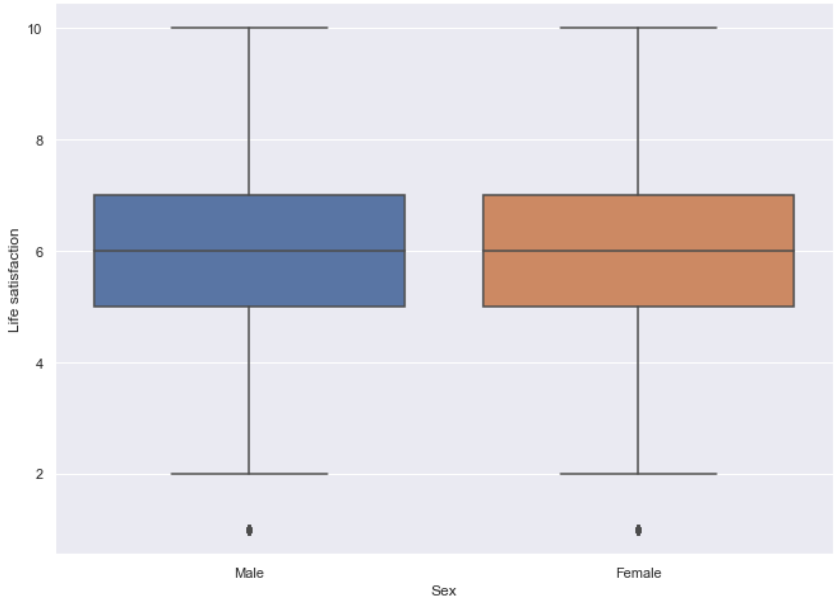

Рисунок 2.8 - Рівень задоволеності життям залежно від статі

На графіку видно, що очевидної залежності між статтю та рівнем щастя немає.

З графіків (Рисунок 2.9) можна помітити, що люди, які постійно потрапляють у скрутні життєві ситуації, в середньому менш задоволені життям, ніж інші групи. Але, в той же час різниці між середніми значеннями у групах з відповідями «Іноді», «Рідко», «Ніколи» майже немає. Є невелика відмінність у питаннях про нестачу їжі та необхідної медичної допомоги.

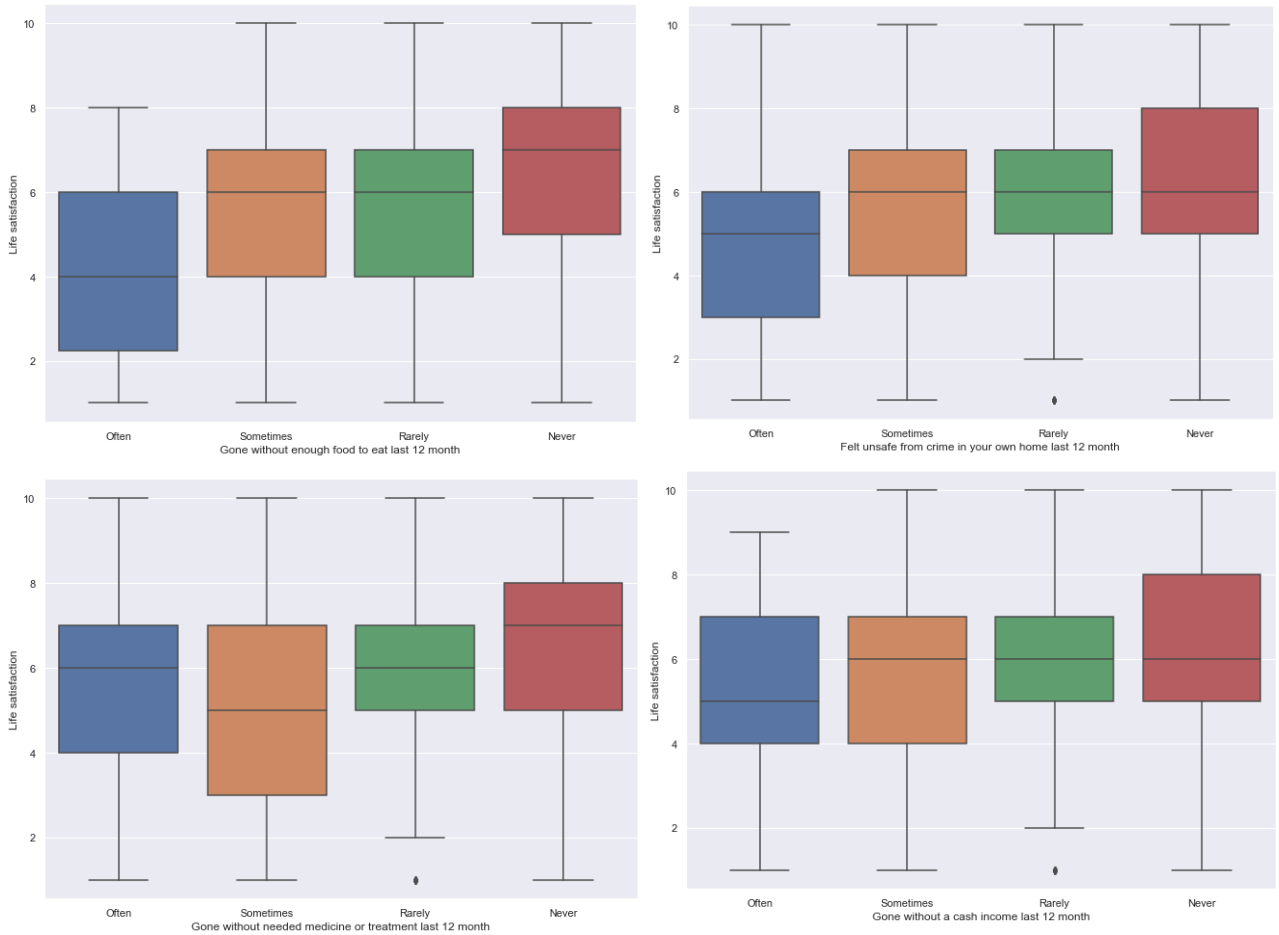

Рисунок 2.9 - Рівень задоволеності життям залежно від соціально-економічних чинників

На Рисунку 2.10 можна побачити, що люди, які отримали тільки початкову освіту значно менше задоволені життям, ніж люди з іншими рівнями освіти. Але серед випускників шкіл, коледжів та ВНЗ вагомої різниці немає. Варто зазначити, що кількісно групи респондентів дуже нерівномірні, через особливості рівнів

освіти в Україні. Дослідивши вікову структуру респондентів, можна відмітити, що середній вік найбільш незадоволеної життям освітньої групи – 59 років, тому в цьому випадку, вік може впливати більше за рівень освіти.

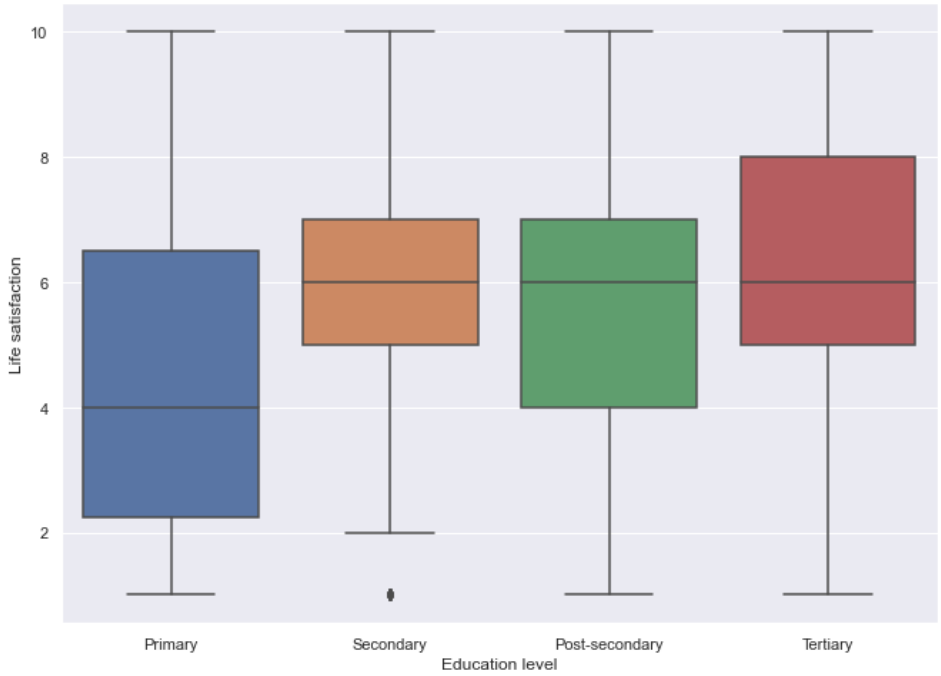

Рисунок 2.10 - Рівень задоволеності життям залежно від рівня освіти

Але на графіку Рисунку 2.11 помітно, що одружені люди або люди, які знаходяться у довготривалих стосунках є трохи більш задоволеними своїм життям, ніж одинаки, розлучені чи вдівці.

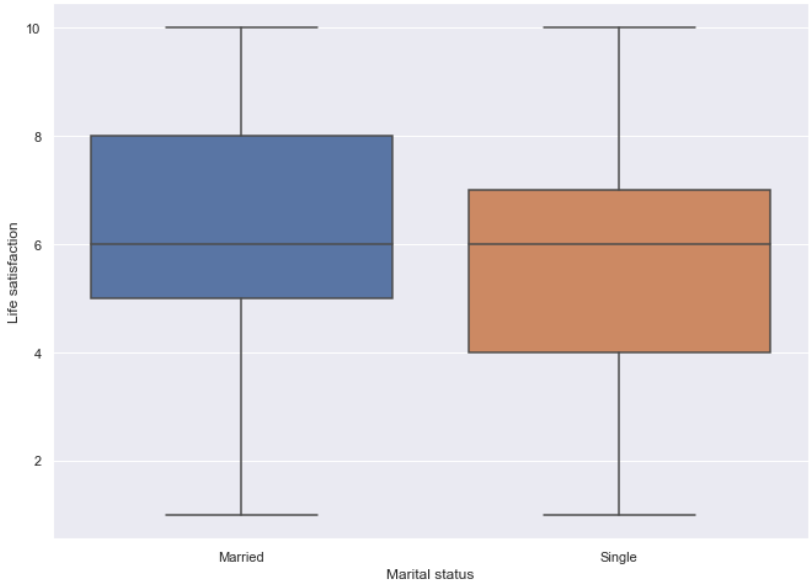

Рисунок 2.11 - Рівень задоволеності життям залежно від сімейного стану

2.4 Нормалізація даних

Зауважимо, що дані опитування подано в шкалах різної кількості градацій, тобто, наприклад, питання про стать має лише два варіанти відповіді, а про рівень освіти – вісім. В такому випадку порушується баланс між впливами вхідних змінних, представлених у різних масштабах, на вихідну змінну. Тобто, ефект обумовлений не реальними взаємодіями, а лише змінами масштабу. Як результат, модель може показувати хибні залежності. Для запобігання такого ефекту, дані було нормалізовано. Нормалізацією називають метод попередньої обробки числових ознак у навчальних наборах даних з метою приведення їх до деякої загальної шкали без втрати інформації про відмінність діапазонів [5]. Для нашого випадку було використано мінімаксну нормалізацію зі зведенням до діапазону [0, 1], так як дані містять невелику обмежену кількість значень та завдяки цьому не містять викидів.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}},$$

де x – значення змінної

x_{min}, x_{max} – найменше та найбільше значення у стовпчику.

РОЗДІЛ 3. ПОБУДОВА МОДЕЛІ РЕГРЕСІЇ

Перший підхід, який було використано для визначення факторів впливу на рівень задоволеності життям в Україні – це модель лінійної регресії. Вона описує взаємозв'язок однієї (залежної) змінної від іншої або кількох інших незалежних змінних (регресорів) з лінійною функцією залежності.

$$y = f(x, b) + \varepsilon,$$

де b – параметри моделі

ε – похибка моделі

Для оцінки якості моделі регресії найчастіше використовують такі метрики:

- Середньоквадратична похибка моделі (англ. Mean Squared Error): середня квадратична різниця між очікуваними значеннями та значеннями, передбаченими моделлю. Чим MSE менший, тим модель краща.
- R-квадрат (коефіцієнт детермінації): показує, яка доля дисперсії моделі пояснюється змінними-предикторами. Для моделей множинної регресії обраховується як квадрат коефіцієнту кореляції між спостережуваними значеннями результату та прогнозованими значеннями моделі. Чим ближче значення R^2 до 1, тим модель краща.
- Середня абсолютна похибка (англ. Mean Absolute Error): середній модуль різниць між фактичними значеннями та значеннями моделі. Чим показник менший, тим модель краща. MAE є більш робастним, ніж MSE.
- Стандартна похибка залишків (англ. Residuals Standard Error): одна з варіацій MSE з поправкою на кількість предикторів моделі. Чим RSE нижче, тим модель краща.

На жаль, велика кількість змінних при обмеженій кількості спостережень спричиняє певні проблеми. Вони стосуються відношення кількості спостережень до кількості змінних. Тобто, в загальному випадку, модель повинна добре знаходити загальні залежності у даних та апроксимувати їх для будь-яких валідних значень предикторів. Але при «перенавчанні» виникає ситуація, коли на тренувальних даних модель показує добрий результат, а на тестових – поганий.

Тобто модель просто «запам’ятовує» велику кількість усіх можливих прикладів, замість того, щоб підмічати загальні особливості.

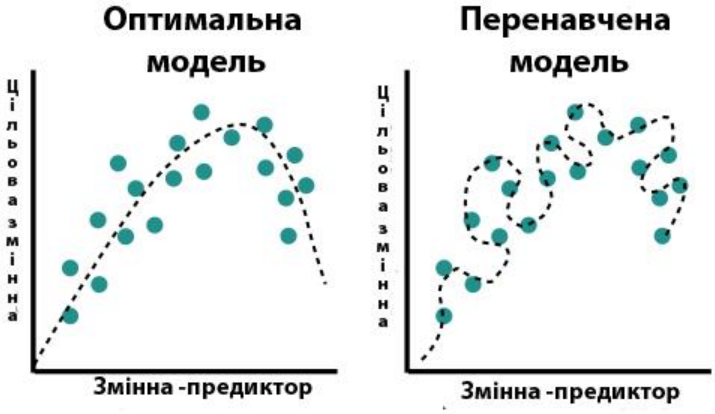

Рисунок 3.1 – Графічне зображення явища перенавчання

Одним з підходів для уникнення перенавчання є відбір лише мінімальної необхідної кількості предикторів, для цього у нашому випадку було використано так звану регресію «Ласо».

Одним із способів вирішення проблем перенавчання регресійної моделі є зміна цільової функції таким чином, щоб включити додаткові «штрафи» на моделі з великими коефіцієнтами. Популярний підхід — це штрафування моделі, використовуючи суму абсолютних значень коефіцієнтів. Це називається регуляризацією L1. Такий підхід мінімізує величини всіх коефіцієнтів і дозволяє будь-якому коефіцієнту зводитись до нуля, ефективно видаляючи вхідні характеристики з моделі.

Отже, після нормалізації даних, можемо почати побудову моделі. Спочатку, було розділено дані на тренувальну та тестову вибірки. Так як загалом даних не дуже багато, було використано поділ у співвідношенні 80/20 відповідно. Для того, щоб попередньо оцінити якість моделі, побудуємо її одразу за всіма змінними датасету, використовуючи метод найменших квадратів.

Метод найменших квадратів (англ. Ordinary Least Squares) - це популярна техніка оцінки невідомих параметрів у моделях лінійної регресії. OLS оцінює

параметри функції за принципом найменших квадратів: мінімізує суму квадратів різниці між спостережуваною залежною змінною (значенням спостережуваної змінної) і прогнозованою змінною, отриманою за допомогою функції. Геометрично це може бути подано як сума квадратів відстаней між кожною точкою даних у наборі та відповідною точкою на поверхні регресії – чим менше різниця, тим краще модель відповідає даним.

У якості міри оцінки якості моделі використаємо середньоквадратичну похибку моделі.

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.941
Model:	OLS	**Adj. R-squared (uncentered):**	0.921
Method:	Least Squares	**F-statistic:**	47.80
Date:	Sat, 29 Jan 2022	**Prob (F-statistic):**	0.00
Time:	23:52:07	**Log-Likelihood:**	494.02
No. Observations:	1031	**AIC:**	-472.0
Df Residuals:	773	**BIC:**	802.0
Df Model:	258		
Covariance Type:	nonrobust		

Рисунок 3.2 – Результати моделі з використанням усіх змінних

Також, для кожної змінної датасету отримали її коефіцієнт у моделі регресії, стандартну похибку, значення t-статистики, p-значення та довірчі інтервали (Рисунок 4.3).

Для поточної та всіх наступних моделей цієї роботи визначимо загальноприйнятий α-рівень значущості як 0.05. Тобто, це означає, що з ймовірністю 5 відсотків результат буде обумовлений випадковим шумом. Тепер вилучимо змінні для яких $p > |t| > 0.05$. У результаті залишилось тільки двадцять сім предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q1	-0.0511	0.054	-0.949	0.343	-0.157	0.055
Q2	-0.0597	0.031	-1.954	0.051	-0.120	0.000
Q3	0.0052	0.029	0.180	0.857	-0.052	0.062
Q4	0.0330	0.024	1.387	0.166	-0.014	0.080
Q5	0.0035	0.027	0.130	0.897	-0.049	0.056
Q6	0.0143	0.028	0.509	0.611	-0.041	0.069
Q7	0.0133	0.014	0.963	0.336	-0.014	0.040

Рисунок 3.3 – Деякі коефіцієнти регресії

Середньоквадратична похибка моделі – 0.036. Це вже доволі гарний результат, але нашою метою було визначити мінімальну кількість найбільш важливих факторів, що впливають на рівень задоволеності життям, тому для цієї моделі було застосовано регуляризацію Ласо. Щоб визначити параметр регуляризації α, скористаємось перехресною перевіркою (кросс-валідацією).

	MSE	Number of Coefficients	Alpha
0	0.036193	26	0.0001
1	0.032969	26	0.0012
2	0.031961	19	0.0023
3	0.031637	16	0.0034
4	0.031793	15	0.0045
5	0.031946	8	0.0056
6	0.032117	4	0.0067
7	0.032116	4	0.0078
8	0.032226	3	0.0089
9	0.032511	3	0.0100

Рисунок 3.4 – Результати перехресної перевірки

За результатами крос-валідації, оберемо значення параметру регуляризації α = 0.0034. При такому значені середньоквадратична похибка найменша, а кількість параметрів моделі значно зменшилась: з двадцяти семи до шістнадцяти.

	Question	Coefficient
0	Q25	-0.007068
1	Q40	-0.017753
2	Q46	-0.175310
3	Q48	0.125100
4	Q50	0.377028
5	Q62	-0.013529
6	Q69	-0.028516
7	Q118	-0.022460
8	Q120	0.027041
9	Q125	0.027520
10	Q147	-0.007888
11	Q176	0.018629
12	Q203	0.009653
13	Q242	-0.019802
14	Q243	0.031703
15	Q262	-0.037617

Рисунок 3.5 – Предиктори моделі після регуляризації

Побудуємо остаточну модель лінійної регресії для цільової змінної Q49, використовуючи отримані шістнадцять змінних-предикторів: Q25, Q40, Q46, Q48, Q50, Q62, Q69, Q118, Q120, Q125, Q147, Q176, Q203, Q242, Q243, Q262. Після цього ще раз відберемо предиктори, p-значення для яких менші за 0.05. У результаті отримали шість незалежних змінних у моделі з наступними коефіцієнтами.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1099	0.030	-3.714	0.0	-0.168	-0.052
Q48	0.2165	0.024	9.060	0.0	0.170	0.263
Q50	0.4438	0.026	17.048	0.0	0.393	0.495
Q53	0.1905	0.020	9.667	0.0	0.152	0.229
Q57	0.1363	0.022	6.265	0.0	0.094	0.179
Q286	0.0778	0.014	5.530	0.0	0.050	0.105

Рисунок 3.6 – Незалежні змінні моделі

Далі дослідимо наявність кореляції між змінними, використавши критерій незалежності Хі-квадрат Пірсона. Цей критерій перевіряє наявність зв'язку між двома змінними. Він є зручним інструментом для перевірки наявності залежності між ординальними та номінальними даними, бо звичні коефіцієнти кореляції не можуть бути коректно використані для категоріальних даних. Нульовою гіпотезою тесту є статистична незалежність поданих даних. Попарно порахуємо рівні взаємодії кожного з шести предикторів між собою. На жаль, ні у бібліотеках мови R, ні у бібліотеках мови Python наразі немає функцій, що можуть автоматично від початку і до кінця виконати Хі-квадрат тест, тому вручну побудуємо таблиці спряженості для кожної пари змінних.

Q48	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	All
Q46											
1.0	10	2	10	15	23	25	46	45	12	44	232
2.0	27	10	31	59	153	125	188	99	40	82	814
3.0	21	4	16	21	49	27	42	13	4	14	211
4.0	5	4	4	0	5	3	7	0	1	3	32
All	63	20	61	95	230	180	283	157	57	143	1289

Рисунок 3.7 – Таблиця суміжності для змінних Q46 та Q48

Далі в середовищі R виконаємо тест незалежності.

```
        Pearson's Chi-squared test

data:  chi_table
X-squared = 121, df = 27, p-value = 0.07000000000006
```

Рисунок 3.8 – Результати Хі-квадрат тесту для змінних Q46 та Q48

За результатами тесту видно, що змінні значуще не впливають одна на одну. Далі тест було виконано для усіх змінних моделі, переважно усі змінні є статистично незалежними.

Отже, отримали модель лінійної регресії наступного вигляду:

$$y = -0.11 * Q46 + 0.22 * Q48 + 0.44 * Q50 + 0.19 * Q53 + 0.14 * Q57$$
$$+0.08 * Q286 + 0.42,$$

де Q46, Q48, Q50, Q53, Q57, Q286 – значення відповідних змінних-предикторів.

Наведемо формулювання цих питань безпосередньо з опитувальника.

Feeling of happiness

Taking all things together, would you say you are:

1.- Very happy
2.- Quite happy
3.- Not very happy
4.- Not at all happy
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Not available

Рисунок 3.9 – Питання Q46

How much freedom of choice and control

Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "none at all" and 10 means "a great deal" to indicate how much freedom of choice and control you feel you have over the way your life turns out.

10.- A great deal
9.- 9
8.- 8
7.- 7
6.- 6
5.- 5
4.- 4
3.- 3
2.- 2
1.- None at all
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Unknown

Рисунок 3.10 – Питання Q48

Satisfaction with financial situation of household

How satisfied are you with the financial situation of your household? If '1' means you are completely dissatisfied on this scale, and '10' means you are completely satisfied, where would you put your satisfaction with your household's financial situation?

10.- Satisfied
9.- 9
8.- 8
7.- 7
6.- 6
5.- 5
4.- 4
3.- 3
2.- 2
1.- Dissatisfied
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Unknown

Рисунок 3.11 – Питання Q50

Frequency you/family (last 12 month): Gone without needed medicine or treatment that you needed

In the last 12 months, how often have you or your family:
Gone without needed medicine or treatment that you needed

1.- Often
2.- Sometimes
3.- Rarely
4.- Never
-1-.- Don´t know
-2-.- No answer

Рисунок 3.12 – Питання Q53

Most people can be trusted

Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?

1.- Most people can be trusted
2.- Need to be very careful
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing

Рисунок 3.13 – Питання Q57

Family savings during past year

During the past year, did your family…

1.- Save money
2.- Just get by
3.- Spent some savings and borrowed money
4.- Spent savings and borrowed money
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Unknown

Рисунок 3.14 – Питання Q286

Фінальна отримана середньоквадратична похибка моделі – 0.03, значення $R^2 = 0.93$.

Отже, можна зробити висновок, що для українців, факторами, які найбільше впливають на рівень задоволеності життям, виявились (за спаданням вагомості у моделі):

- Задоволеність фінансовим становищем (Q50)
- Рівень контролю над власним життям та свобода вибору (Q48)
- Наявність необхідної медичної допомоги (Q53)
- Довіра до інших людей (Q57)
- Відчуття щастя (Q46)
- Фінансові накопичення родини за останній рік (Q286).

РОЗДІЛ 4. ПОБУДОВА МОДЕЛЕЙ КЛАСТЕРИЗАЦІЇ

Наведені вище результати регресії щодо задоволеності життям є усередненими, тому можуть не завжди бути застосовними для усіх громадян України, бо крім індивідуальних відмінностей серед респондентів, різні соціальні групи також можуть знаходитись в різних умовах та, відповідно, мати різні чинники впливу на рівень задоволеності життям. Тому, було вирішено спочатку розділити суспільство за певним ознаками, а згодом побудувати нові моделі регресії для кожної з груп окремо. Такий поділ має назву кластеризація. Кластеризація — це завдання поділу даних на групи таким чином, щоб точки даних в одній групі були більш схожими на інші точки даних у тій же групі, ніж на інші точки даних в інших групах. Простіше кажучи, мета полягає в тому, щоб розділити та згрупувати дані зі схожими характеристиками [8]. Об’єктами кластеризації у нашому випадку є респонденти у кількості 1289 осіб.

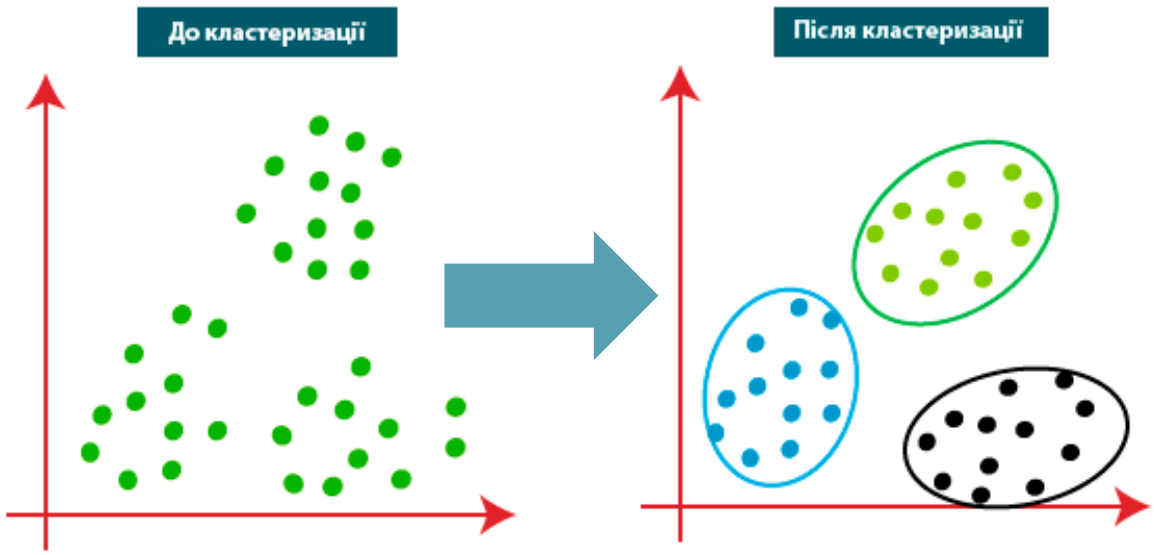

Рисунок 4.1 – Графічне представлення роботи алгоритму кластеризації

Існує декілька методів кластеризації, але враховуючи специфіку даних WVS7 було обрано три методи: метод k-середніх, метод DBSCAN та модель сумішей Гауса (англ. Gaussian Mixture Model). Детальніше розглянемо кожен з них.

4.1 Метод k-середніх

Алгоритм k-середніх — це ітераційний алгоритм, який намагається розділити набір даних на k попередньо визначених кластерів, де кожна точка даних належить лише одному кластеру [10]. Мета алгоритму - зробити точки даних всередині кластерів якомога більш схожими між собою, зберігаючи при цьому максимально можливу різницю (відстань) між кластерами. Для нашого випадку у якості міри міжкластерної відстані використаємо евклідову відстань. Точки призначаються кластеру таким чином, щоб сума квадратів відстаней між точками даних і центроїдами кластера (середнє арифметичне всіх точок даних, що належать до кластера) була зведена до мінімуму. Чим менше ми маємо варіацій у кластері, тим однорідніші є точки даних у ньому. Метою алгоритму є мінімізація цільової функції, яка мінімізує міжкластерну відстань та максимізує відстань між елементами різних кластерів.

Математична модель:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{k} w_{ik} \left\| x^i - \mu_k \right\|^2,$$

де J – цільова функція

x^i - і-те спостереження

μ_k – центроїд

Оскільки цей алгоритм працює лише із заздалегідь визначеним числом кластерів, то скористаємось двома методами для визначення числа кластерів: метод ліктя та метод середнього силуету.

Метод ліктя:

1. Запустимо метод k-середніх для різних $k = \overline{1,12}$
2. Для кожного значення k обчислимо внутрішньокластерну суму квадратів відстаней (англ. Within-Clusters Squared Sum)
3. Побудуємо графік залежності k від WSS.
4. Оберемо таке значення k, при якому WSS починає спадати менш інтенсивно.

Для нашого датасету графік виглядатиме наступним чином:

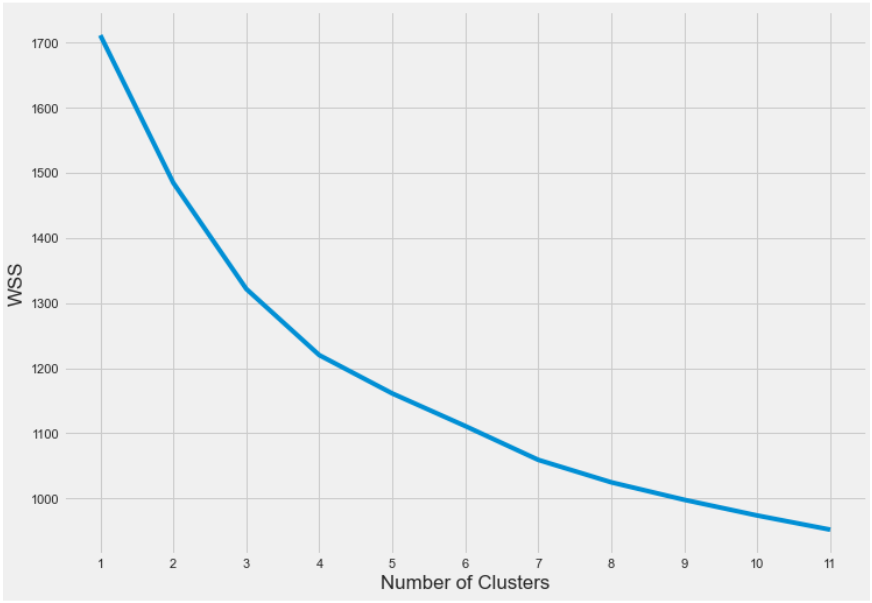

Рисунок 4.2 – Метод ліктя

За методом ліктя, наші дані можна спробувати розділити на чотири кластери.

Метод силуету:

Наступним способом оцінки кількості кластерів є коефіцієнт силуету. Коефіцієнт силуету - це показник, що використовується для визначення якості обраної кластеризації. Значення коефіцієнта знаходяться на проміжку від -1 до 1, де «1» позначає, що кластери чітко розрізняються один від одного та помітно віддалені, «0» позначає, що кластеризація не є чіткою, відстань між кластерами не є значною та «-1» позначає, що є суттєві помилки у присвоєні кластерів даним, кластеризація не є успішною.

Коефіцієнт обраховується за формулою:

$$s = \frac{b-a}{\max(a,b)},$$

де a – середня відстань між точками всередині кластера

b – середня відстань між усіма кластерами

Зобразимо графік залежності коефіцієнта силуету від кількості кластерів для наших даних:

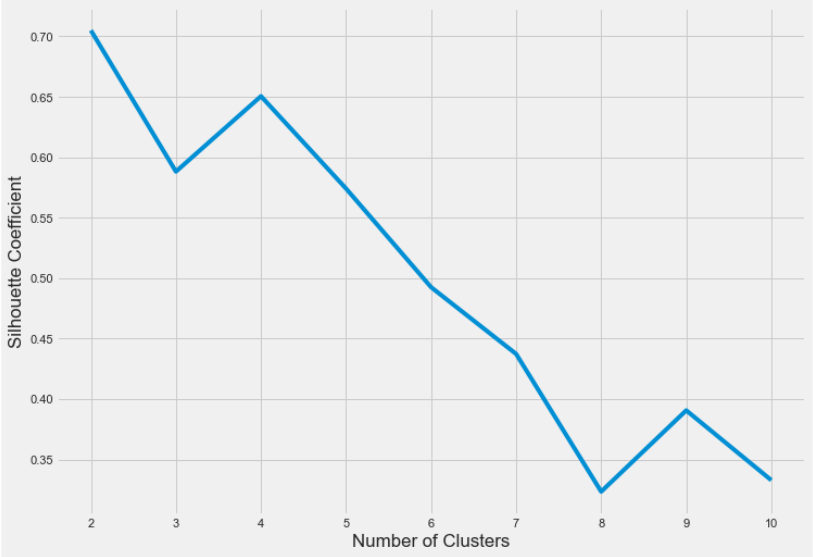

Рисунок 4.3 – Залежність коефіцієнта силуету від кількості кластерів

За цим графіком видно, що при k = 2 та k = 4 отримаємо оптимальне значення коефіцієнта силуетів. Щоб однозначно визначитись з кількістю кластерів, побудуємо графіки силуетів (Рисунок 4.4) та графік кластеризованих даних (Рисунок 4.5 та Рисунок 4.6).

Отже, на графіках видно, що дані вдало можна розділити як на два, так и на чотири кластери. Але, враховуючи результат методу ліктя та беручи до уваги, що при k = 2 розмір кластеру «0» майже в три рази більший за кластер «1», що не є збалансованим поділом, оптимальним рішенням буде обрати чотири кластери.

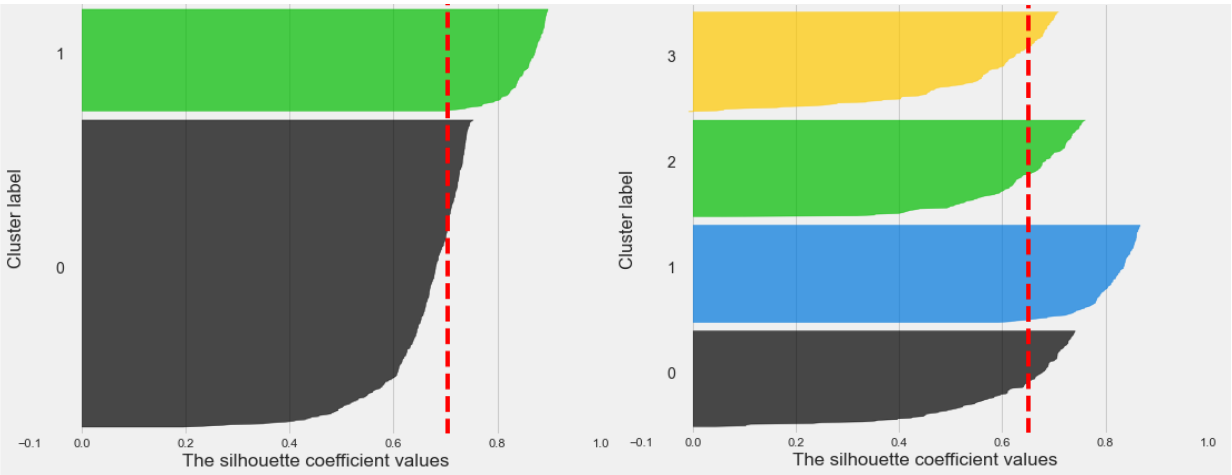

Рисунок 4.4 – Графік силуетів

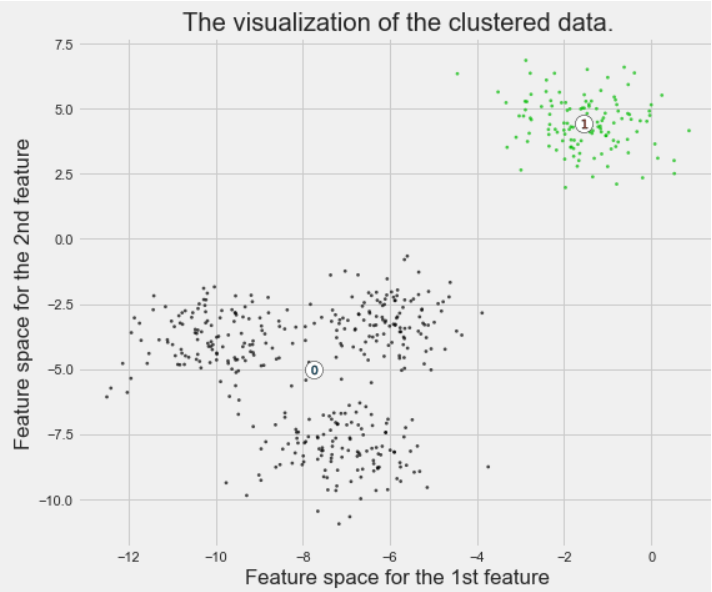

Рисунок 4.5 – Розподіл точок даних при k = 2 у просторі перших головних компонент

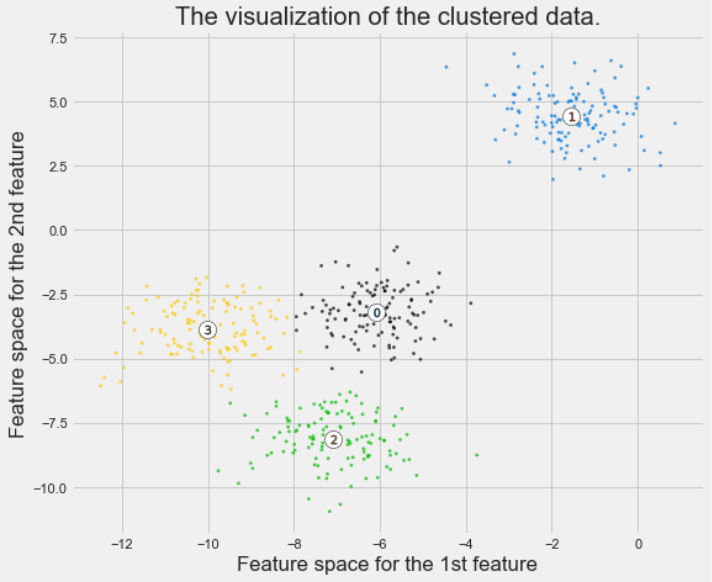

Рисунок 4.6 – Розподіл точок даних при k = 4 у просторі перших головних компонент

Отже, тепер аналогічно до розділу 3, побудуємо регресійні моделі для кожного з чотирьох кластерів та визначимо найбільш вагомі чинники впливу на рівень вдоволеності життям в Україні.

Отримані результати моделей:

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.2049	0.039	2.333	0.000	-0.260	-0.093
Q48	0.3912	0.014	14.01	0.000	0.031	0.420
Q50	0.5123	0.016	13.161	0.000	0.416	0.508
Q286	0.0317	0.011	2.008	0.002	0.023	0.091
Q142	0.1207	0.029	9.827	0.001	0.146	0.225
Q51	0.0901	0.017	6.054	0.000	0.108	0.201

Рисунок 4.7 – Результати для кластеру 1

Модель лінійної регресії для кластера «1»:

$$y = -0.21 * Q46 + 0.39 * Q48 + 0.51 * Q50 + 0.03 * Q286 + 0.12 * Q142 + 0.09 * Q51 + 0.41,$$

де Q46, Q48, Q50, Q286, Q142, Q51 – значення відповідних змінних-предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.2111	0.030	7.027	0.000	-0.270	-0.151
Q57	0.0597	0.014	4.460	0.000	0.033	0.086
Q48	0.4052	0.023	16.986	0.000	0.302	0.450
Q50	0.4617	0.024	19.456	0.000	0.459	0.508
Q95	0.0815	0.015	5.367	0.001	0.052	0.111
Q47	0.0439	0.117	2.007	0.002	0.121	0.867
Q142	0.1509	0.014	8.925	0.000	0.117	0.798

Рисунок 4.8 – Результати для кластеру 2

Модель лінійної регресії для кластера «2»:

$$y = -0.21 * Q46 + 0.06 * Q57 + 0.40 * Q48 + 0.46 * Q50 + 0.08 * Q95$$

$$+ 0.04 * Q47 + 0.15 * Q142 + 0.16,$$

де Q46, Q57, Q48, Q50, Q95, Q47, Q142 – значення відповідних змінних-предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1908	0.030	2.121	0.000	-0.270	-0.086
Q48	0.5113	0.017	15.003	0.000	0.0302	0.447
Q50	0.5028	0.014	12.382	0.000	0.419	0.527
Q286	0.0356	0.288	3.117	0.001	0.063	0.100
Q51	0.0678	0.170	4.056	0.000	0.117	0.304
Q142	0.1390	0.078	10.001	0.001	0.151	0.227

Рисунок 4.9 – Результати для кластеру 3

Модель лінійної регресії для кластера «3»:

$$y = -0.19 * Q46 + 0.51 * Q48 + 0.50 * Q50 + 0.04 * Q286 + 0.07 * Q51$$

$$+ 0.14 * Q142 + 0.23,$$

де Q46, Q48, Q50, Q286, Q142, Q51 – значення відповідних змінних-предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1717	0.041	-4.222	0.000	-0.252	-0.092
Q57	0.0400	0.014	2.814	0.005	0.012	0.068
Q95	0.0502	0.016	3.161	0.002	0.019	0.081
Q48	0.3799	0.030	12.611	0.000	0.321	0.439
Q50	0.4617	0.029	15.827	0.000	0.404	0.519
Q286	0.0520	0.015	3.442	0.001	0.022	0.082
Q285	0.0426	0.015	2.809	0.005	0.013	0.072

Рисунок 4.10 – Результати для кластеру 4

Модель лінійної регресії для кластера «4»:

$$y = -0.18 * Q46 + 0.04 * Q57 + 0.05 * Q95 + 0.38 * Q48 + 0.46 * Q50 + 0.05 * Q286 + 0.04 * Q285 + 0.25,$$

де Q46, Q57, Q95, Q48, Q50, Q286, Q285 – значення відповідних змінних-предикторів.

Зобразимо відмінності між чотирма кластерами у вигляді діаграми, на якій позначимо середні значення кожної змінної-предиктора у певному кластері.

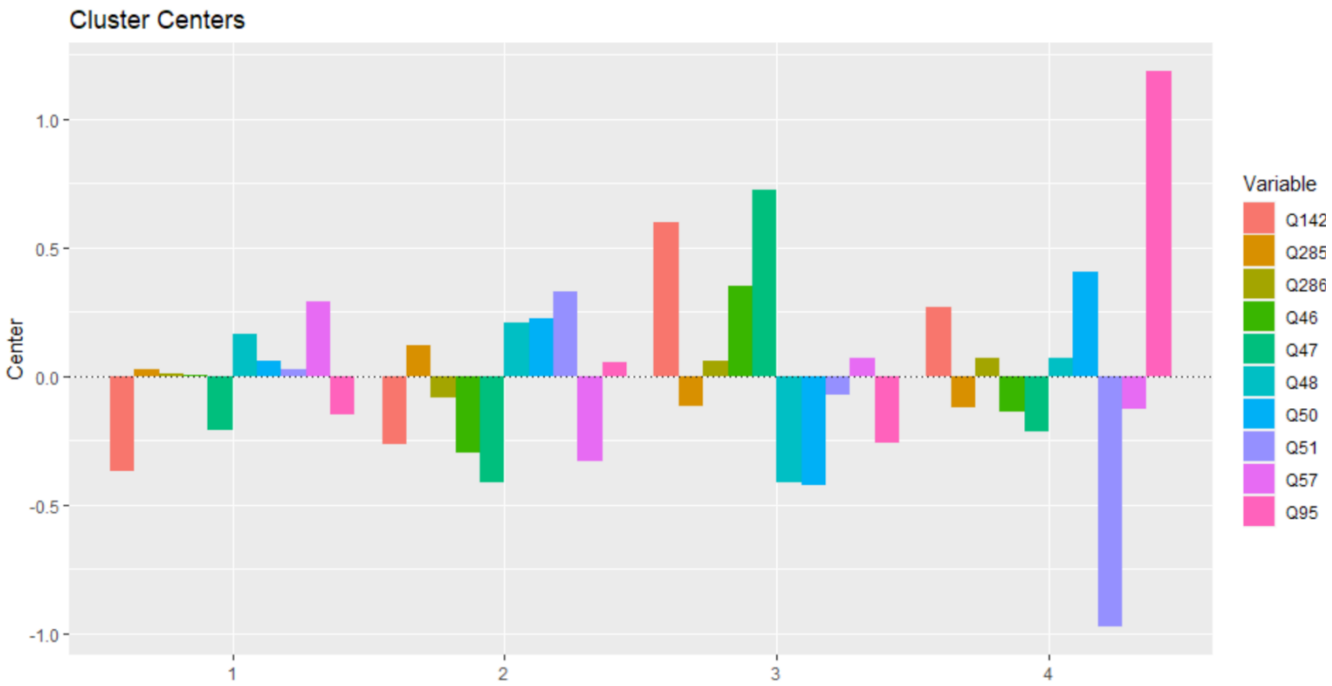

Рисунок 4.11 – Середні значення змінних-предикторів по кожному кластеру

Бачимо, що перший кластер містить осіб, які дають майже середній рівень по всіх ключових питаннях, з дещо вищим від середнього рівнем оцінки питання Q51 (Соціально складні становища) та з дещо нижчими відносно середнього відповідями на питання Q142 (Наявність стабільної роботи), тобто представники цього кластеру частіше переймаються про втрату роботу та Q46 (Відчуття щастя), представники цього кластеру нижче оцінюють свій рівень щастя. Якщо подивитися на вихідний датасет, можна побачити, що це люди, які належать переважно до старшої вікової групи 58 – 72 років, переважно жіночої статі, з вищою освітою, у

шлюбі.

В другому кластері люди дають дещо більш радикальну оцінку: негативну на питання 142, 46, 47, 57, та дещо вище середнього рівня – 48, 50, 51. Об’єкти цього кластеру – це люди, які належать переважно до середньої вікової групи 30 - 49 років, переважно жіночої статі, з вищою освітою, у шлюбі, з однією дитиною у родині.

Об’єкти третього кластеру виражено позитивно відповідають на питання 142, 46, 47, і помітно нижче середнього – на питання 48, 50, 95. Це - люди, які належать переважно до середньої вікової групи 33 – 55 років, переважно чоловічої статі, з вищою освітою, у шлюбі.

Четвертий кластер найбільше відрізняється від трьох попередніх різко позитивними відповідями на питання 95, позитивними – на питання 142 і 50, і з різко негативними відповідями на питання 51. До цього кластеру належать переважно люди середньої вікової групи 31 – 48 років, жіночої статі, з вищою освітою, у шлюбі, з двома та більше дітьми у родині.

У таблиці нижче наведено питання, які статистично значущо впливають на рівень вдоволеності життям для респондентів кожного кластера. Фактори відсортовані за спаданням коефіцієнта регресії.

Кластер 1	Кластер 2	Кластер 3	Кластер 4
Вдоволеність фінансовим становищем (Q50)	Вдоволеність фінансовим становищем (Q50)	Свобода та рівень контролю над власним життям (Q48)	Вдоволеність фінансовим становищем (Q50)
Свобода та рівень контролю над власним життям (Q48)	Свобода та рівень контролю над власним життям (Q48)	Вдоволеність фінансовим становищем (Q50)	Свобода та рівень контролю над власним життям (Q48)

Рівень щастя (Q46)	Рівень щастя (Q46)	Рівень щастя (Q46)	Рівень щастя (Q46)
Наявність стабільної роботи (Q142)	Наявність стабільної роботи (Q142)	Наявність стабільної роботи (Q142)	Фінансові накопичення родини (Q286)
Довіра до людей (Q57)	Спортивна активність (Q95)	Довіра до людей (Q57)	Спортивна активність (Q95)
Фінансові накопичення родини (Q286)	Довіра до людей (Q57)	Фінансові накопичення родини (Q286)	Основний дохід родини (Q285)
	Стан здоров’я (Q47)		Довіра до людей (Q57)

Таблиця 4.1 – Результати алгоритму k-середніх

Отже, за результатами побудованих моделей помітно, що приблизно 80% моделі пояснюється одними и тими ж змінними, тому в цьому випадку їх знову можна відмітити як основні значущі фактори, що впливають на рівень задоволеності життям українців. Але наявна відмінність між вагою кожної зі змінних у певних кластерах, що дає можливість у подальшому працювати з цими групи людей, спрямовуючись на найбільш вагомі чинники.

4.2 Метод DBSCAN

Алгоритм DBSCAN (англ. Density-based spatial clustering of applications with noise) використовує щільність даних та інтуїтивне означення понять «кластер» та «шум». Основна ідея алгоритму полягає в тому, щоб об’єднувати дані, в радіусі яких є не менш ніж деяка задана кількість точок [14]. Цей алгоритм схожий на інші звичайні методи кластеризації, але завдяки тому, що використовує не відстані до деяких точок, а радіус навколо них та кількість об’єктів в околі вказаного радіусу (щільність), він є добре прийнятним для даних з великою кількістю шуму. Більш того, він значно краще визначає більш складні форми кластерів чи, наприклад,

вкладені кластери, тому частіше використовується у сучасних системах машинного навчання (Рисунок 4.12).

Але, завдяки такій гнучкості алгоритм часто схильний визначати занадто велику кількість кластерів, включаючи надто дрібні, з несуттєвими відмінностями. Також, гарним прикладом застосування алгоритму DBSCAN є сегментація зображень та моделювання поведінки користувачів того чи іншого сервісу.

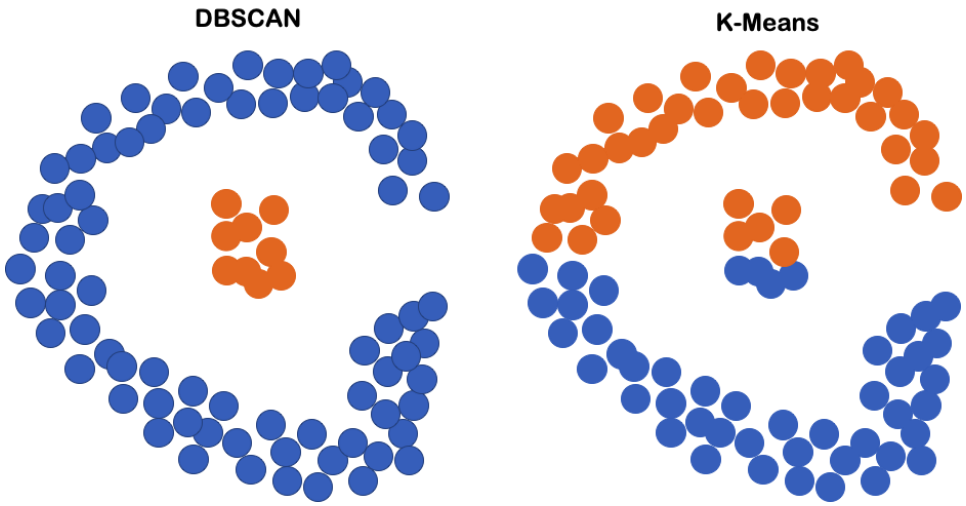

Рисунок 4.12 – Ілюстративне зображення складності форм кластерів для алгоритмів DBSCAN та K-Means

Для побудови моделі DBSCAN потребує два початкових параметри: eps – радіус для кожної точки, в якому відбуватиметься пошук (схоже поняття до епсілон-околу точки в математичному аналізі), та minPts – мінімальна кількість точок, яка має міститись у радіусі, для того, щоб точку не було позначено як викид.

В загальному вигляді алгоритм може бути описано наступним чином:

1. Шукаємо інші точки в епсілон-околі кожної точки, ті, які містять більш ніж minPts сусідів відмічаємо як окремі кластери.
2. Для кожного отриманого кластера шукаємо компоненти зв’язності.
3. Точки, що залишились, приєднуємо до найближчого кластера (якщо відстань не більше за eps), інакше – точка відмічається як викид.

Для вибору параметрів eps та minPts було використано стандартний підхід, детально описаний у роботі [15].

Побудуємо модель з параметрами eps = 0.3 та minPts = 40. Алгоритм виділив два кластера.

Отже, отримані результати моделей.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1696	0.043	-3.961	0.000	-0.268	-0.099
Q48	0.4342	0.028	11.212	0.000	0.18	0.236
Q95	0.0484	0.017	3.066	0.002	0.019	0.079
Q57	0.0413	0.015	2.651	0.005	0.012	0.07
Q50	0.4935	0.028	14.612	0.000	0.413	0.473
Q285	0.0546	0.016	3.439	0.001	0.024	0.089

Рисунок 4.13 – Результати регресії для кластера 1

Модель лінійної регресії для кластера «1»:

$$y = -0.17 * Q46 + 0.43 * Q48 + 0.05 * Q95 + 0.04 * Q57 + 0.49 * Q50 + 0.06 * Q285 + 0.41,$$

де Q46, Q48, Q95, Q57, Q50, Q285 – значення відповідних змінних-предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1753	0.040	-4.279	0.000	-0.265	-0.091
Q48	0.4341	0.028	12.003	0.000	0.164	0.253
Q95	0.0518	0.016	3.195	0.002	0.017	0.079
Q57	0.0390	0.013	2.579	0.005	0.013	0.069
Q50	0.4242	0.027	16.626	0.000	0.382	0.567
Q286	0.0491	0.014	3.456	0.001	0.021	0.089

Рисунок 4.14 – Результати регресії для кластера 2

Модель лінійної регресії для кластера «2»:

$$y = -0.18 * Q46 + 0.43 * Q48 + 0.05 * Q95 + 0.03 * Q57 + 0.42 * Q50 + 0.05 * Q286 + 0.24,$$

де Q46, Q48, Q95, Q57, Q50, Q286 – значення відповідних змінних-предикторів.

Також зобразимо відмінності між двома кластерами на діаграмі.

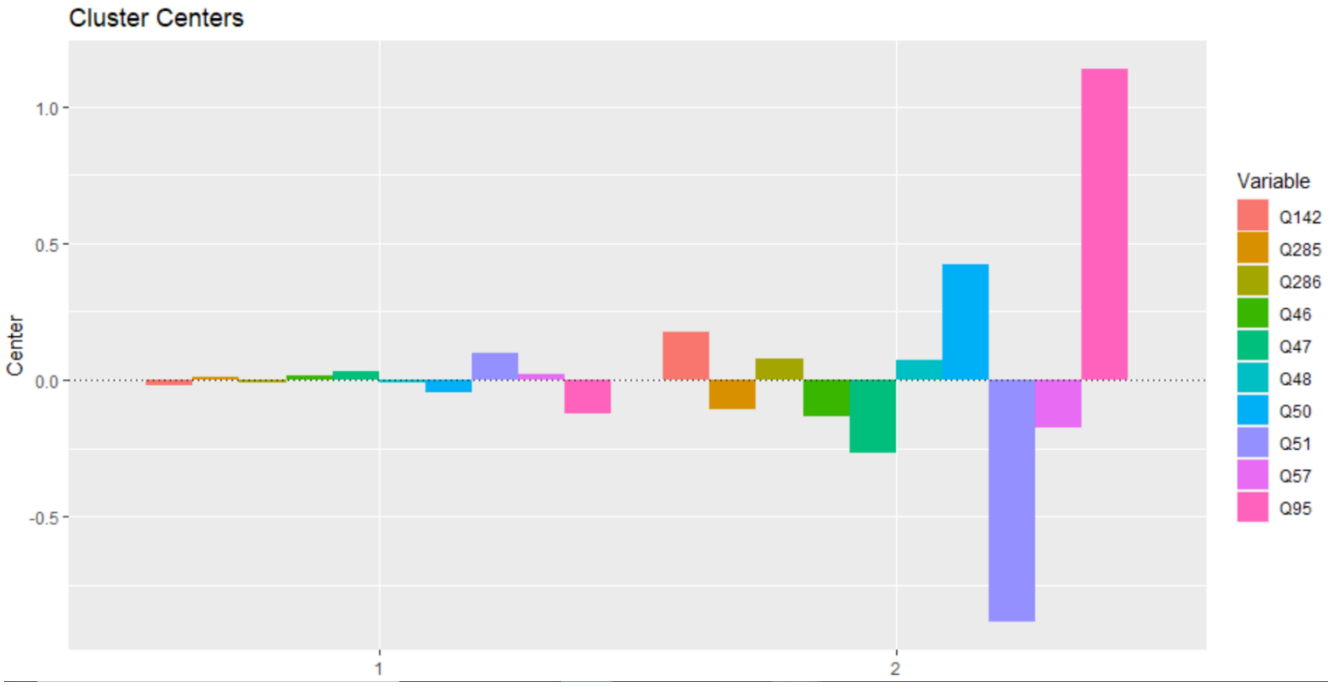

Рисунок 4.15 - Середні значення змінних-предикторів по кожному кластеру

Бачимо, що для такого поділу на кластери в перший кластер попадають об’єкти, які дають середню оцінку всіх ключових питань, вони не є сильно задоволеними чимось, проте не є дуже незадоволеними чимось. Це переважно люди старшого середнього віку 35-61 років, переважно жіночої статі, з вищою освітою, у шлюбі, переважно з двома дітьми, професійно-техічного роду занять, соціального класу вище середнього.

В другому кластері зібрано об’єкти, які дають більш контрастні оцінки, зокрема високі оцінки для питання 95, та негативну оцінку питання 51. Якщо

подивитися на датасет, бачимо, що це представники середнього віку 31-54 років, переважно чоловічої статі, з вищою освітою, у шлюбі, перважно з однією дитиною, професійно-техічного роду занять, соціального класу вище середнього.

4.3 Модель сумішей Гауса

Модель сумішей Гауса – це модель, яка складається з функцій Гауса у кількості k, де k – кількість кластерів у датасеті. Кожен k-й Гаусіан представлений наступними параметрами:

- математичне очікування μ_k
- дисперсія σ_k (або матриця коваріацій Σ_k для багатовимірного простору)

Математична модель:

$$p(x) = \sum_{i=1}^{K} \varphi_i N(x|\mu_i, \sigma_i)$$

$$N(x_i|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\sum_{i=1}^{k} \varphi_i = 1,$$

де φ_i – вага і-го Гаусіана у моделі[7]

Для оцінки якості нашої моделі будемо використовувати інформаційний критерій Баєса (англ. BIC) - індекс, який використовується в баєсівській статистиці для вибору між декількома альтернативними моделями.

$$BIC = -2\,ln(L) + \; k\,ln(n),$$

де L – максимальне значення цільової функції

k – число параметрів моделі

n – число об’єктів вибірки

Для визначення оптимальних параметрів, було використано крос-перевірку: спочатку для визначення оптимальної матриці коваріацій (Рисунок 4.14), потім – для кількості кластерів.

	spherical	tied	diag	full
1	163738.712427	185392.009614	110491.532758	1.853920e+05
2	154829.677096	185228.354880	64074.938495	3.736460e+05
3	148639.825675	187712.425689	54931.730121	5.543535e+05
4	145393.210184	187968.065094	46869.888299	7.125017e+05
5	142168.369769	189046.224799	38493.411866	8.786912e+05
6	140987.965612	189954.306722	34126.815397	1.037315e+06
7	135803.859280	191248.715394	24844.620597	1.198763e+06
8	134713.318442	191392.632848	18283.744451	1.415675e+06
9	134151.323772	192242.908145	16367.925053	1.581961e+06

Рисунок 4.16 – Визначення типу коваріаційних параметрів

Найкращі результати було отримано при використанні діагональної матриці коваріацій для кожної компоненти. Тому, при задані параметрів моделі використаємо covariance_type=’diag’.

Далі було проведено обчислення BIC при різній кількості кластерів: від одного до тридцяти. Мінімальне значення BIC = 17901.90 отримане на одинадцяти кластерах. Тому побудуємо модель для одинадцяти кластерів.

	Observations
1	0.0
2	0.0
3	0.0
4	0.0
5	933.0
6	0.0
7	0.0
8	355.0
9	0.0
10	1.0
11	0.0

Рисунок 4.17 – Розподіл кількості спостережень по кластерах

Після тестування моделі стало помітно, що усі спостереження розподілились

усього по двох кластерах. Навчання було проведено ще декілька разів, але результат не змінився, з чого можна зробити висновок, що незважаючи на значення BIC, поділ на два кластери є оптимальним.

Побудуємо окремі моделі для двох кластерів, використовуючи метод найменших квадратів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1569	0.041	-4.505	0.000	-0.228	-0.097
Q48	0.4593	0.032	12.514	0.000	0.183	0.261
Q95	0.0537	0.016	3.073	0.002	0.021	0.076
Q57	0.0424	0.014	2.953	0.005	0.011	0.074
Q50	0.4822	0.031	14.512	0.000	0.432	0.542
Q286	0.0504	0.015	3.305	0.001	0.024	0.082

Рисунок 4.18 – Результати моделі для кластеру 1

Модель лінійної регресії для кластера «1»:

$$y = -0.16 * Q46 + 0.46 * Q48 + 0.05 * Q95 + 0.04 * Q57 + 0.48 * Q50 + 0.05 * Q286 + 0.23,$$

де Q46, Q48, Q95, Q57, Q50, Q286 – значення відповідних змінних-предикторів.

	coef	std err	t	P>|t|	[0.025	0.975]
Q46	-0.1868	0.039	-4.019	0.000	-0.234	-0.100
Q48	0.4708	0.030	12.793	0.000	0.172	0.230
Q95	0.0508	0.015	3.348	0.002	0.018	0.075
Q57	0.0367	0.014	2.618	0.005	0.013	0.067
Q50	0.4938	0.031	14.596	0.000	0.395	0.472
Q286	0.0529	0.014	3.609	0.001	0.024	0.087

Рисунок 4.19 – Результати моделі для кластеру 2

Модель лінійної регресії для кластера «2»:

$$y = -0.17 * Q46 + 0.47 * Q48 + 0.05 * Q95 + 0.04 * Q57 + 0.49 * Q50 + 0.05 * Q286 + 0.32,$$

де Q46, Q48, Q95, Q57, Q50, Q286 – значення відповідних змінних-предикторів.

Отже, можна зробити висновок, що принципової різниці між факторами, що впливають на рівень задоволеності життям між двома кластерами немає. Для деяких предикторів коефіцієнти збільшились або зменшились, але сам набір предикторів майже не варіюється. Але так само як і в попередніх випадках, можна виділити групи за ваговими коефіцієнтами змінних у моделі.

ВИСНОВКИ

У ході виконання кваліфікаційної роботи було зібрано та досліджено дані World Values Survey 7 про соціальні, культурні, етнічні, політичні та економічні цінності українського суспільства станом на 2020 рік. Попередньо дані було очищено та нормалізовано, згідно загальноприйнятих статистичних практик, та проведено розвідувальний аналіз даних.

На наступному етапі за допомогою моделей машинного навчання було відібрано фактори, що найбільшою мірою впливають на рівень задоволеності життям в Україні. Спочатку моделлю лінійної регресії було отримано усереднені результати для всіх опитаних. Далі, з метою визначення особливостей, що можуть безпосередньо впливати на невеликі групи респондентів, але не бути помітними при розширені моделі на все суспільство, населення було кластеризоване трьома різними алгоритмами кластеризації. Після побудови необхідних моделей машинного навчання було з’ясовано, що самі показники впливу на суб’єктивний рівень задоволеності життям помітно не відрізняються при поділі суспільства на групи, але наявна відмінність у їх вазі для кожної групи респондентів. Серед основних таких факторів було визначено: вдоволеність фінансовим становищем, свобода та рівень контролю над власним життям, відчуття щастя, довіра до людей та блага сучасного світу: робота та медицина.

У ході виконання роботи були отримані знання у галузі роботи з соціологічними даними, поглиблено знання про сучасні алгоритми машинного навчання, розглянуто бібліотеки мов Python та R для статистичних розрахунків, візуалізації та машинного навчання.

Подальші дослідження та перспективи. Одним з основних варіантів розширення цієї роботи є збір свіжих даних про стан українського населення. Враховуючи поточну воєнно-політичну ситуацію в Україні є вагомі підстави вважати, що фактори впливу задоволеності життям українців доволі сильно зміняться під час та після війни. Це дає поштовх для подальших досліджень у цій сфері. Результати можуть бути корисними для спеціалістів з різних галузей:

соціологів, психологів, політологів. Дослідження морального стану суспільства можуть бути використані для допомоги біженцям чи військовим, для вивчення способів їх підтримки та повернення до нормального життя. В цьому плані, Україна зараз отримує унікальний досвід, вартий уваги.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven - Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi.org/10.14281/18241.13
2. Questionnaire and research topics [Електронний ресурс] – Режим доступу до ресурсу: https://www.worldvaluessurvey.org/WVSContents.jsp
3. Ryff C: Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *J Pers Soc Psychol* 1989, 57: 1069–1081.
4. Kashdan TB: The assessment of subjective well-being (issues raised by the Oxford happiness questionnaire). *Pers Indiv Differ* 2004, 36: 1225–1232. 10.1016/S0191-8869(03)00213-7
5. Data normalization [Електронний ресурс] – Режим доступу до ресурсу: https://wiki.loginom.ru/articles/data-normalization.html.
6. Regularization [Електронний ресурс] – Режим доступу до ресурсу: https://ranalytics.github.io/data-mining/042-Regularization.html.
7. Gaussian mixture model [Електронний ресурс] – Режим доступу до ресурсу: https://brilliant.org/wiki/gaussian-mixture-model/
8. An Introduction to Clustering and different methods of clustering [Електронний ресурс] – Режим доступу до ресурсу: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/.
9. Bayesian Information Criterion (BIC) / Schwarz Criterion [Електронний ресурс] – Режим доступу до ресурсу: https://www.statisticshowto.com/bayesian-information-criterion/.
10. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks [Електронний ресурс] – Режим доступу до ресурсу: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.
11. Regression model validation [Електронний ресурс] – Режим доступу до

ресурсу: http://https//www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

12. Lasso regression with Python [Електронний ресурс] – Режим доступу до ресурсу: https://machinelearningmastery.com/lasso-regression-with-python/.
13. Determining the optimal number of clusters [Електронний ресурс] – Режим доступу до ресурсу: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.
14. DBSCAN clustering in machine learning [Електронний ресурс] – Режим доступу до ресурсу: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/.
15. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN.