

АНАЛІЗ ВПЛИВУ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ПОКАЗНИКІВ НА РЕЗУЛЬТАТИ ЗНО ТА НМТ

Анотація: Робота зосереджена на комплексному аналізі факторів, що впливають на результати ЗНО/НМТ, з акцентом на взаємозв'язок між соціально-економічними показниками регіонів та індивідуальними характеристиками учнів. Метою дослідження є не лише виявлення статистичних зв'язків, але й надання глибшого розуміння механізмів, за допомогою яких ці фактори можуть впливати на освітні досягнення.

Ключові слова: інтелектуальний аналіз даних, модель кластеризації, модель класифікації.

Вступ

Система освіти є критично важливим компонентом соціально-економічного розвитку будь-якої країни. Результати зовнішнього незалежного оцінювання (ЗНО) та національного мультипредметного тесту (НМТ) слугують не лише індикатором індивідуальних досягнень учнів, але й віддзеркалюють якість освітнього процесу на регіональному та національному рівнях. Розуміння факторів, що впливають на ці результати, є ключовим для формування ефективних освітніх стратегій та політик.

Існує гіпотеза про те, що освітні досягнення не існують у вакуумі, а тісно пов'язані з ширшим соціально-економічним контекстом. Регіональні відмінності в економічному розвитку, рівні доходів населення, зайнятості та соціальній інфраструктурі можуть створювати різні умови для навчання та впливати на успішність учнів. Крім того, індивідуальні соціально-демографічні характеристики, такі як місце проживання та гендер, також можуть відігравати значну роль у формуванні освітніх можливостей та результатів.

Матеріали та методи

Освіта є основою прогресу та конкурентоздатності держави, а якість знань громадян – визначним фактором її розвитку. В Україні основним механізмом оцінювання рівня підготовки випускників шкіл є зовнішнє незалежне оцінювання, а з 2022 року – національний мультипредметний тест. Вони не лише контролюють якість освіти, а й є вирішальним фактором при вступі до закладів вищої освіти [1]. Відтак, результати ЗНО та НМТ мають значний вплив на освітню траєкторію та подальші соціально-економічні перспективи молоді.

Предметом дослідження є сфера оцінювання освітніх досягнень в Україні, представлена у вигляді результатів ЗНО та НМТ [1]. Особливу увагу приділено пошуку та аналізу факторів, що можуть впливати на ці результати. Серед потенційних чинників

соціально-економічні показники, які відображають рівень розвитку регіонів, матеріальний стан населення, наявність ресурсів для навчання та інші соціальні й економічні чинники.

Актуальність дослідження зумовлена наявністю значної різниці в освітніх результатах учнів, які походять з різних соціально-економічних середовищ та різних куточків країни. Визначення та оцінка факторів впливу на результати ЗНО та НМТ можуть стати підґрунтям для розробки дієвих заходів щодо підвищення рівня знань української молоді. Зокрема, було визначено ряд питань, відповіді на які допоможуть дійти висновків:

— Чи впливає рівень соціально-економічного розвитку регіонів на успішність їх мешканців на ЗНО та НМТ?

— Чи є зв'язок між типом місцевості та якістю надання освітніх послуг?

— Чи є значна відмінність в результатах тестування між хлопцями та дівчатами?

Для отримання обґрунтованих відповідей на ці питання проведено комплексний аналіз даних, який включає в себе створення сховища даних, застосування ETL процесів та використання методів інтелектуального аналізу, зокрема кластеризації та класифікації [2].

Одним з головних джерел є Український центр оцінювання якості освіти (УЦОЯО) [3]. Цей вебсайт є офіційним джерелом інформації щодо ЗНО та НМТ. На порталі представлено статистичні звіти, дані про результати тестувань [4]. Дані УЦОЯО є першоджерелом інформації про результати тестувань і забезпечують об'єктивну основу для дослідження. Звідси отримано дані про бали усіх учасників тестування, їхню стат, регіон та тип місцевості, де вони проживають.

Як джерело інформації про соціально-економічні показники регіонів обрано Державну службу статистики України [5]. Вебсайт Укрстату є основним офіційним джерелом статистичної інформації в Україні. Його було обрано для отримання інформації про соціально-економічні показники регіонів, а саме: валовий регіональний продукт у розрахунку на одну особу, кількість населення кожного регіону, а також рівень безробіття в кожному з них. Дані для дослідження взято за період з 2019 по 2022 рік. Це дає можливість оцінити ситуацію за різних обставин. На жаль, інформація про ВРП та рівень безробіття за 2022 рік відсутня у відкритих джерелах, проте було прийнято рішення залишити показники за цей рік у наборі даних. Під час аналізу враховано цей факт і проведено окремо розрахунки для цієї частини показників.

Також для дослідження обрано такий економічний показник, як середня заробітна плата. Цей показник є важливим соціально-економічним індикатором, який може відображати загальний рівень добробуту населення в регіоні та потенційно впливати на доступність освітніх ресурсів для учнів. Дані взято за січень кожного з досліджуваних років з інтернет-ресурсу Minfin.com.ua [6], розділ «Середня заробітна плата в Україні». Тут наявна інформація в розрізі регіонів та видів економічної діяльності.

У моделі сховища за типом сніжинка спроектовано 8 таблиць вимірів та 2 фактові таблиці.

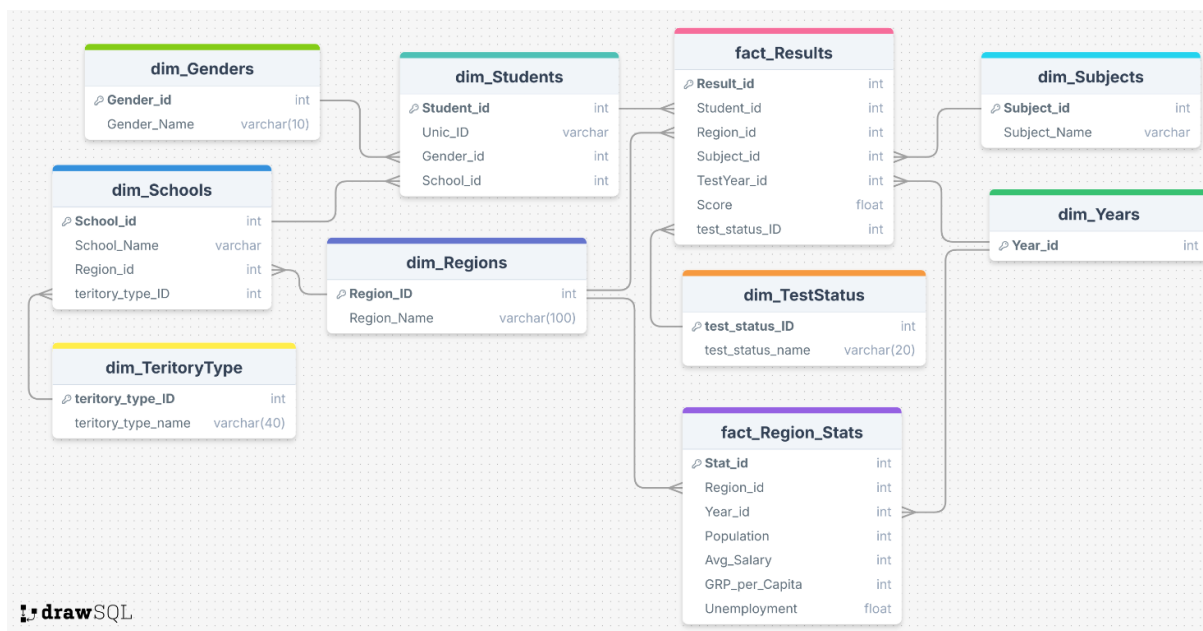


Рисунок 1. Модель сховища даних

Для комплексного дослідження впливу соціально-економічних показників на результати ЗНО та НМТ було обрано кілька різнотипних методів інтелектуального аналізу даних для кожної з поставлених дослідницьких задач.

Результати та обговорення

Аналіз факторів

Першим проведено кореляційний аналіз з використанням коефіцієнта кореляції Пірсона, щоб дізнатися чи є зв'язок між соціально-економічними показниками регіонів та результатами ЗНО і НМТ. Результати тестування мають слабкий або помірний зв'язок з соціально-економічними показниками. Серед них найбільший зв'язок спостерігається з середньою заробітною платою (рис. 2).

Додаткове дослідження з використанням регресійного аналізу показало, що кожна з незалежних змінних (ВРП на душу населення, середня заробітна плата, чисельність населення, рівень безробіття) має статистично значущий вплив на результати тестування ($p < 0.001$). Однак, моделі мають низьку пояснювальну силу, про що свідчать низькі значення R-squared (від 0.001 до 0.032). Це означає, що кожен показник окремо пояснює лише невелику частку варіативності результатів ЗНО/НМТ. Найбільший позитивний вплив спостерігається для середньої заробітної плати, а найбільший негативний - для рівня безробіття. Отже, необхідно провести багатofакторну регресію для оцінки сукупного впливу всіх показників.

З метою зменшення мультиколінеарності, було проведено перетворення змінних 'GRP_per_capita' та 'avg_salary'. Оскільки дані мають регіональну структуру, було вирішено стандартизувати 'GRP_per_capita' та 'avg_salary' в межах кожного регіону,

щоб уникнути спотворення результатів. Після стандартизації було створено узагальнюючий індекс економічного добробуту, що дозволило замінити дві вихідні змінні на одну та спростити модель. Значення VIF для всіх предикторів (індексу економічного добробуту, чисельності населення та рівня безробіття) знаходяться в прийнятних межах. Проте, незважаючи на усунення мультиколінеарності, пояснювальна сила моделі залишилася низькою ($R\text{-squared} = 0.032$), а аналіз залишків все ще виявляв ознаки нелінійності, гетероскедастичності та відхилення від нормального розподілу, подібні до тих, що спостерігалися в початковій багатофакторній регресії. Ймовірно, зв'язок між соціально-економічними показниками та результатами ЗНО/НМТ є більш складним і нелінійним, або ж існують інші важливі фактори, які не були враховані в цій частині дослідження.

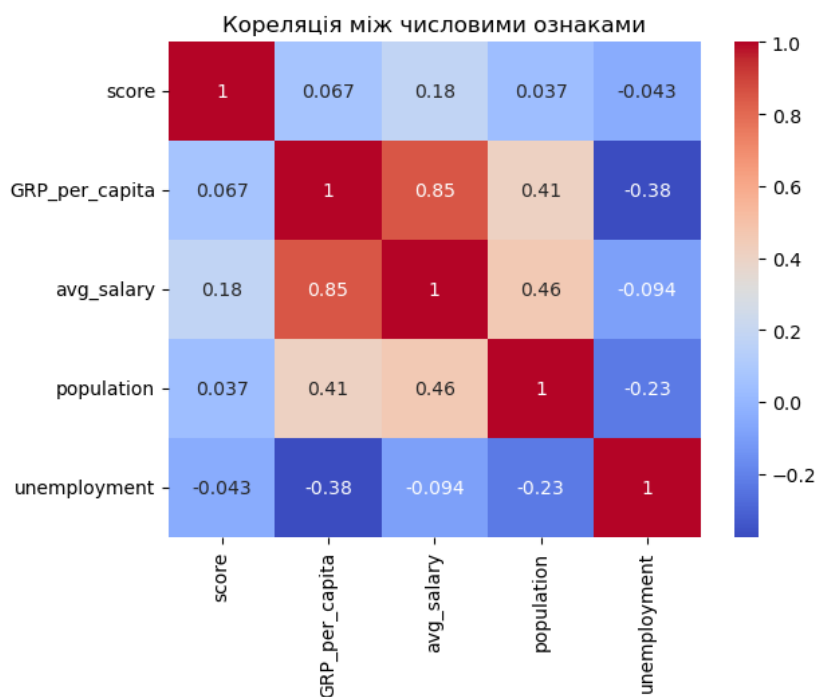


Рисунок 2. Кореляція соціально-економічних показників і результатів ЗНО/НМТ

На наступному етапі було проведено агрегацію загальної оцінки соціально-економічного розвитку регіонів, що дозволило узагальнити інформацію про різні показники для подальшого аналізу їхнього впливу на результати ЗНО/НМТ.

Для агрегації було використано підхід, що враховує силу впливу кожного показника на середній бал ЗНО/НМТ, визначену за допомогою регресійного аналізу. Було розраховано дві агреговані оцінки:

– Зважена оцінка на основі коефіцієнтів регресії. Цей метод використовував безпосередньо коефіцієнти регресійної моделі (з агрегованими на рівні регіону даними) як ваги для об'єднання стандартизованого індексу економічного добробуту, чисельності населення та рівня безробіття.

– Зважена оцінка на основі стандартизованих коефіцієнтів регресії. Цей метод спочатку стандартизував агреговані на рівні регіону показники (індекс економічного добробуту, чисельність населення та рівень безробіття), а потім використовував коефіцієнти регресійної моделі, побудованої на цих стандартизованих даних, як ваги для їх об'єднання.

Аналіз результатів показав помірний позитивний зв'язок обох методів з середнім балом ЗНО/НМТ. Однак, агрегована оцінка, розрахована з використанням стандартизованих коефіцієнтів регресії, продемонструвала дещо вищу кореляцію (0.554) порівняно з оцінкою, зваженою за не стандартизованими коефіцієнтами (0.501). Це свідчить про її кращу відповідність статистичному зв'язку між стандартизованими складовими індексу та освітніми результатами.

Зважаючи на вищу кореляцію, для подальшого аналізу та включення до більш повної моделі, що пояснює варіативність результатів ЗНО/НМТ, буде обрано агреговану оцінку соціально-економічного розвитку регіонів, розраховану з використанням стандартизованих вагових коефіцієнтів регресії.

Наступним кроком проведено дослідження на наявність залежності між типом місцевості та результатами ЗНО і НМТ. Перед початком аналізу було об'єднано категорії «село» та «селище, село» в одну категорію «село». Таким чином залишилося лише 3 типи місцевості: село, місто, селище міського типу.

Для формальної перевірки наявності статистично значущих відмінностей між середніми балами учасників ЗНО/НМТ залежно від типу місцевості було застосовано однофакторний дисперсійний аналіз (ANOVA), результати якого показали: F-статистика = 12216.508, р-значення = 0.000, що вказує на статистично значущі відмінності між середніми балами залежно від типу місцевості. Результати post-hoc та тесту Тьюкі чітко показують, що існують статистично значущі відмінності в середніх балах ЗНО/НМТ між усіма трьома типами місцевості.

Для дослідження впливу гендеру на результати ЗНО/НМТ проведено порівняння середніх балів між чоловіками та жінками за допомогою t-тесту для незалежних вибірок. Аналіз проводився як для всієї вибірки учасників, так і окремо для тих, хто успішно склав тест.

Результати t-тесту для всіх учасників показали статистично значущі відмінності між середніми балами чоловіків та жінок ($t = -202.391$, $p < 0.001$). Середні бали для жінок (133.24) виявилися значно вищими, ніж для чоловіків (119.78). Результати наступного аналізу також виявили статистично значущі відмінності між середніми балами жінок (146.58) та чоловіків (140.20) ($t = -184.860$, $p < 0.001$). Проте, абсолютна різниця в середніх балах між статями серед тих, хто склав тест, є меншою (приблизно 6.38 бала), ніж при розгляді всієї вибірки (приблизно 13.46 бала).

Додатково було проаналізовано частку учасників, які не склали тест, залежно від гендеру, результати свідчать про те, що частка неуспішно складених тестів є вищою серед чоловіків (14.57%), ніж серед жінок (9.10%).

Отримані результати свідчать про наявність статистично значущого впливу гендеру на результати ЗНО/НМТ. Жінки демонструють як вищі середні бали загалом, так і вищі середні бали серед тих, хто успішно склав тест. Крім того, серед чоловіків спостерігається вища частка тих, хто не зміг подолати поріг складання.

Отже, аналізуючи результати уже проведених досліджень, можна однозначно сказати, що рівень соціально-економічного розвитку регіону, стать та тип місцевості впливають на результати ЗНО та НМТ.

Моделі для кластерного аналізу

Наступним етапом виконано кластеризацію регіонів на основі соціально-економічних показників, поданих у форматі агрегованої оцінки, та на основі результатів ЗНО та НМТ. Використано методи: ієрархічна кластеризація, кластеризація методом k-середніх та DBSCAN. Окремо зроблено кластеризації регіонів за соціально-економічними показниками, за результатами складання ЗНО та НМТ, після чого проведено порівняння кластерів однакових рівнів і для визначення конкретного показника подібності пораховано відсоток збігу.

ієрархічна кластеризація	learning_level	Високий	Низький	Середній	Всього
	economic_level				
	Високий	3	0	1	4
	Низький	0	0	4	4
	Середній	0	1	16	17
метод k-середніх	Всього	3	1	21	25
	learning_level_kmeans_defined	Високий	Низький	Середній	Всього
	economic_level_kmeans_defined				
	Високий	3	0	1	4
	Низький	2	0	6	8
DBSCAN	Середній	2	1	10	13
	Всього	7	1	17	25
	learning_level_dbscan	Високий	Низький	Середній	Всього
	economic_level_dbscan				
	Високий	2	0	0	2
	Низький	1	2	17	20
	Середній	1	0	1	2
	Всього	4	2	18	24

Рисунок 3. Таблиця спряженості рівнів кластерів

Застосування трьох методів кластеризації виявило різний рівень збігу між соціально-економічним розвитком регіонів та їхніми результатами ЗНО/НМТ. Ієрархічна кластеризація показала найвищий збіг (76%), вказуючи на помітний зв'язок. Метод k-середніх виявив менший збіг (52%), а DBSCAN – найнижчий (20.83% без викидів), що може бути пов'язано з особливостями алгоритму та виявленням складніших структур.

Загалом, кластеризація підтверджує наявність певного зв'язку між досліджуваними аспектами, проте його вираженість залежить від використаного методу.

Моделі для класифікації

Для вирішення задачі класифікації було побудовано 3 моделі: дерево рішень, випадковий ліс та логістичну регресію. Дані було поділено на навчальні та тренувальні та оцінено якість кожної з моделей. На рисунках 4-6 показано результати аналізу моделей.

Загальна точність усіх трьох моделей є відносно низькою. Найкращі результати показують Дерево рішень та Випадковий ліс з точністю 39%, тоді як Логістична регресія демонструє ще нижчу точність у 32%. Це свідчить про те, що менше ніж 40% прогнозів кожної з моделей співпадають з фактичними рівнями балів у тестовому наборі.

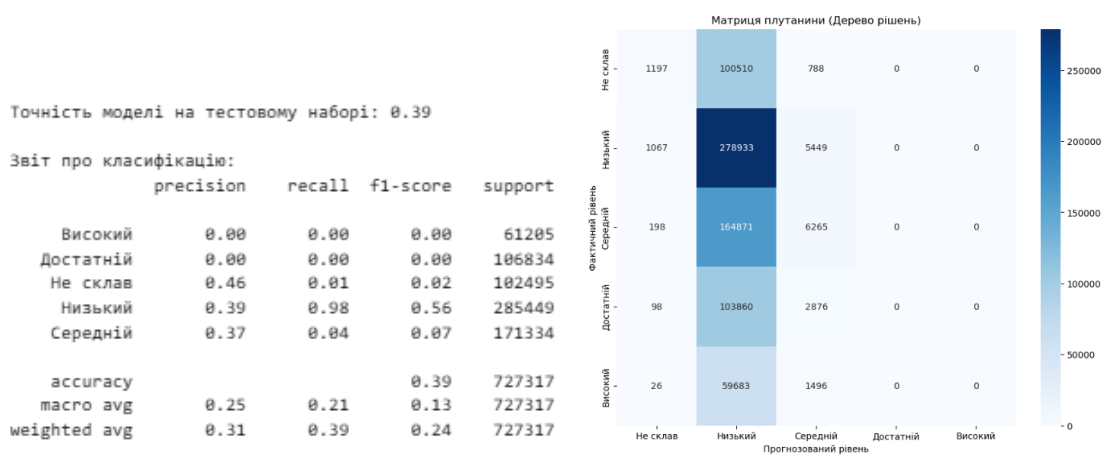


Рисунок 4. Оцінка якості дерева рішень

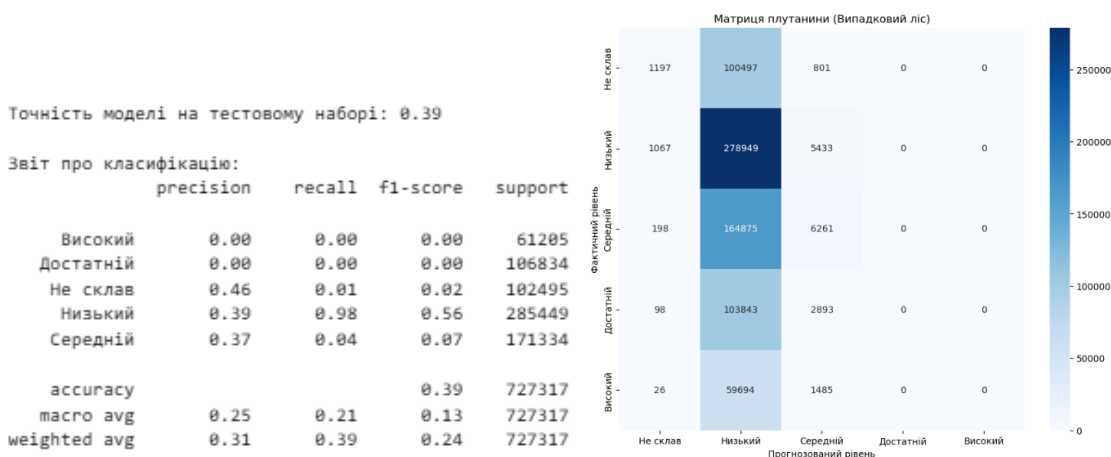


Рисунок 5. Оцінка якості випадкового лісу

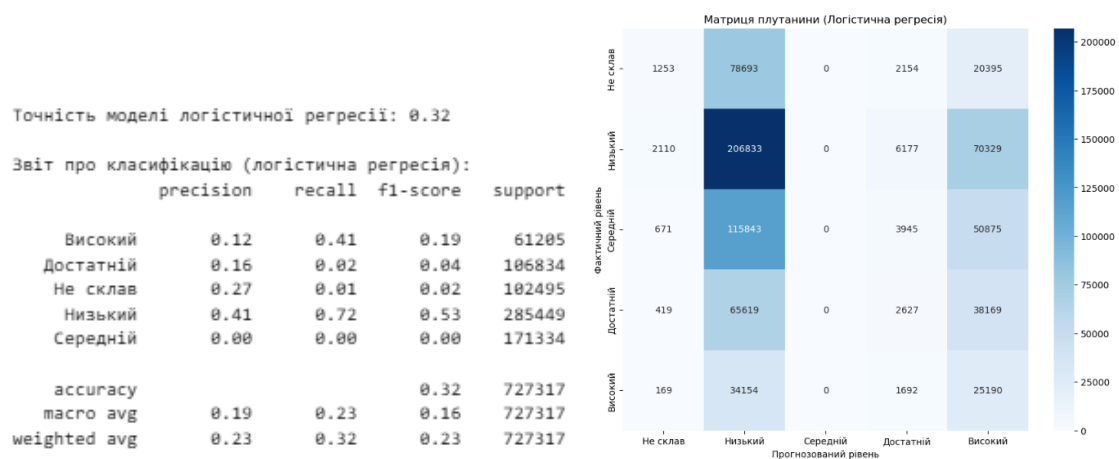


Рисунок 6. Оцінка якості логістичної регресії

Поглиблений аналіз звіту про класифікацію виявляє суттєві проблеми з прогнозуванням окремих класів. Так, для класів "Високий" та "Достатній" Дерево рішень та Випадковий ліс показали нульові значення precision, recall та F1-score, що вказує на повну нездатність моделей ідентифікувати ці рівні. Логістична регресія демонструє деякі передбачення для цих класів, але їх якість залишається вкрай низькою. Для класу "Не склав" моделі характеризуються низьким recall при помірному precision, що означає, що вони пропускають значну кількість фактичних випадків цього рівня. У випадку класу "Низький" спостерігається високий recall, але низький precision, що свідчить про схильність моделей помилково відносити до цього рівня об'єкти інших класів. Прогнозування класу "Середній" також є слабким у всіх трьох моделях.

Низька точність отриманих моделей може бути зумовлена кількома факторами. По-перше, використані ознаки (регіон, тип місцевості, стать) можуть не нести достатньої інформації для точного прогнозування рівня балів. По-друге, значна незбалансованість класів, могла ускладнити навчання моделей для менш представлених рівнів. По-третє, залежності між вхідними ознаками та цільовою змінною можуть бути складними та нелінійними, що обмежує можливості лінійних моделей, таких як логістична регресія, та навіть деревних моделей з базовими параметрами.

Проведене дослідження включало два основні етапи: аналіз зв'язку між соціально-економічними показниками регіонів та результатами ЗНО/НМТ, а також кластеризацію і побудову моделей класифікації рівня балів учасників тестування на основі індивідуальних характеристик.

Аналіз кореляції та регресійний аналіз виявили статистично значущий, але слабкий зв'язок між соціально-економічними показниками (ВРП на душу населення, середня заробітна плата, чисельність населення, рівень безробіття) та середнім балом ЗНО/НМТ на рівні регіонів. Низькі значення R-squared вказують на те, що ці показники окремо або в сукупності пояснюють лише незначну частку варіативності результатів тестування. Виявлені проблеми мультиколінеарності та невідповідність залишків припущенням

лінійної регресії свідчать про складніший характер залежностей або необхідність врахування інших факторів. Агрегація соціально-економічних показників у зважену оцінку показала помірний позитивний зв'язок з середнім балом ЗНО/НМТ, що підтверджує наявність певного впливу регіонального контексту на освітні результати.

Дослідження впливу типу місцевості та гендеру на індивідуальні результати ЗНО/НМТ виявило статистично значущі відмінності. Учасники із селищ міського типу демонстрували вищі середні бали порівняно з учасниками з міст та сіл. Жінки показали вищі середні бали та менший відсоток неуспішного складання тесту порівняно з чоловіками. Ці результати підкреслюють важливість соціально-демографічних факторів у визначенні успішності складання ЗНО/НМТ.

Застосування методів кластеризації (ієрархічної, k-середніх та DBSCAN) для виявлення схожих груп регіонів на основі соціально-економічних показників та результатів ЗНО/НМТ показало різний рівень збігу між цими двома аспектами. Ієрархічна кластеризація виявила найбільший збіг, що свідчить про певну кореспонденцію між рівнями соціально-економічного розвитку та освітніми досягненнями регіонів. Менший збіг, отриманий за допомогою методів k-середніх та DBSCAN, може бути пов'язаний з особливостями алгоритмів та складнішою структурою даних.

Побудовані моделі класифікації рівня балів (Дерево рішень, Випадковий ліс та Логістична регресія) на основі індивідуальних характеристик (регіон, тип місцевості, стать) продемонстрували низьку загальну точність. Усі три моделі мали проблеми з точним прогнозуванням усіх рівнів балів, особливо класів "Високий" та "Достатній". Це може бути зумовлено недостатньою інформативністю обраних ознак, значною незбалансованістю класів цільової змінної, а також складним і, можливо, нелінійним характером залежностей між цими ознаками та рівнем балів. Отримані результати свідчать про необхідність подальшого дослідження з використанням ширшого набору ознак, застосування більш складних методів моделювання та ретельнішою обробкою проблеми незбалансованості класів для покращення якості прогнозування рівня балів ЗНО/НМТ.

Висновки

Проведене дослідження дозволило комплексно оцінити вплив соціально-економічних факторів регіонів, а також типу місцевості та гендеру на результати ЗНО/НМТ.

Аналіз кореляції та регресійний аналіз виявили статистично значущий, але слабкий зв'язок між соціально-економічними показниками регіонів та середніми балами ЗНО/НМТ. Це свідчить про те, що макроекономічні чинники мають певний вплив на освітні результати, проте їх пояснювальна сила обмежена. Виявлені проблеми мультиколінеарності та нелінійності вказують на складність взаємозв'язків та необхідність врахування інших, можливо, не макроекономічних факторів.

Порівняльний аналіз виявив значні відмінності в результатах ЗНО/НМТ залежно від типу місцевості та гендеру. Учасники із селищ міського типу демонструють

в середньому кращі результати, ніж учасники з сільської місцевості та міст. Жінки, в свою чергу, показують вищі середні бали та менший відсоток неуспішного складання тесту порівняно з чоловіками. Ці результати підкреслюють наявність соціальних нерівностей в освітньому процесі.

Кластеризація регіонів за соціально-економічними показниками та результатами ЗНО/НМТ виявила певний рівень збігу між цими двома аспектами, що підтверджує вплив регіонального контексту на освітні досягнення. Однак, ступінь збігу залежить від використаного методу кластеризації, що свідчить про складність структури даних та необхідність обережного вибору методів аналізу.

Побудовані моделі класифікації рівня балів показали низьку точність прогнозування. Це може бути пов'язано з недостатньою інформативністю використаних ознак, значною незбалансованістю класів цільової змінної, а також складністю та нелінійністю залежностей між ознаками та результатами тестування.

Загалом, проведене дослідження підтверджує наявність впливу соціально-економічних факторів, типу місцевості та гендеру на результати ЗНО/НМТ. Отримані результати можуть бути використані для розробки адресних заходів освітньої політики, спрямованих на зменшення нерівностей та підвищення якості освіти в різних регіонах та для різних груп населення. Подальші дослідження можуть бути спрямовані на вивчення інших факторів, що впливають на результати ЗНО/НМТ, а також на розробку більш точних моделей прогнозування.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Освіта.ua [Електронний ресурс]. – Режим доступу: <https://osvita.ua/> (дата звернення: 20.03.2025).
2. Основи інтелектуального аналізу даних: методи інтелектуального аналізу даних / Київський інститут науково-технічної інформації. – Київ, 2011 (дата звернення: 21.03.2025).
3. ЗНО онлайн [Електронний ресурс]. – Режим доступу: <https://zno.testportal.com.ua/> (дата звернення: 21.03.2025).
4. Відкриті дані ЗНО [Електронний ресурс] / Український центр оцінювання якості освіти. – Режим доступу: <https://zno.testportal.com.ua/opendata> (дата звернення: 21.03.2025).
5. Державна служба статистики України [Електронний ресурс]. – Режим доступу: <https://www.ukrstat.gov.ua/> (дата звернення: 22.03.2025).
6. Середня заробітна плата в Україні [Електронний ресурс] / Мінфін. – Режим доступу: <https://index.minfin.com.ua/ua/labour/salary/average/> (дата звернення: 22.03.2025).