УДК 004.89

М.С. Мазорчук, А.Ю. Пантєлєєва, І.А. Трофимова

# Дослідження факторів впливу на вступ до вищих навчальних закладів на основі методів Data Mining

*Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут»*

Розглянуто методи і моделі виявлення значущих факторів, що впливають на результат вступу абітурієнтів до вищих навчальних закладів (ВНЗ). Як приклад для аналізу використано базу даних за результатами вступу 2009 року. Загальна популяція досліджуваних становила 460710 абітурієнтів. Аналіз виконано за 43 факторними ознаками. У процесі аналізу запропоновано використовувати комплекс методів, що базуються як на статистичній теорії аналізу даних, так і на методах інтелектуального аналізу (Data Mining). Усі розрахунки здійснено з використанням відомого інструментарію для статистичної обробки даних SPSS. Головним результатом дослідження є висновки щодо справедливості вступу абітурієнтів до ВНЗ на основі результатів національного іспиту – зовнішнього незалежного оцінювання (ЗНО).

***Ключові слова*:** аналіз даних, Data Mining, зовнішнє незалежне оцінювання, факторні ознаки, справедливість вступу.

## Вступ

Оцінювання ефективності дії запроваджених форм складання іспитів для вступу у ВНЗ і вплив їх на подальше навчання у вишах й професійне життя абітурієнта є важливою складовою для будь-якої демократичної країни. Зовнішнє незалежне оцінювання (ЗНО), яке запроваджено в Україні як інструмент вступу до вишів, передбачає рівні можливості для всіх абітурієнтів і прозорий механізм проведення вступних іспитів до ВНЗ. Чи дійсно це так? Зараз в Україні не існує єдиного механізму аналізу результатів ЗНО з точки зору надійності, валідності й справедливості. Періодичне обчислювання результатів, обмежений доступ до інформації, незручний формат даних, нестандартизовані класифікатори дисциплін ВНЗ, велика кількість ознак – все це обмежує можливості дослідити множину факторів, які впливають на вступ до ВНЗ.

Впродовж 10 років накопичено багато статистичного матеріалу щодо складання тестів ЗНО й результатів вступу до вишів. Незважаючи на відмінності щодо умов прийому, методів шкалювання й обмеженості доступу до зведених баз даних, спільно з Українським центром оцінювання якості освіти вдається проводити періодичний аналіз результатів ЗНО з точки зору справедливості й валідності [1,2]. Для аналізу в основному застосовувалися класичні статистичні методи, які дозволяють зробити висновки щодо узагальнюючих характеристик тестів й необхідності їх подальшого вдосконалення, модернізації форм їх проведення. Але досі відсутні дані щодо вивчення факторів, що впливають на вступ абітурієнтів. Чи дійсно ЗНО є справедливим інструментом для різних верств населення?

Метою цієї роботи є дослідження факторів впливу на вступ до вищих навчальних закладів. Основне питання – це виявлення факторів, які найчастіше за все вливають на результати вступу до вишів, й обґрунтування методів аналізу пошуку прихованих закономірностей.

## 1. Вихідні дані

Дані для аналізу, де кількість популяції нараховує 460710 осіб, подано у деперсоніфікованій базі даних. Усього 43 факторні ознаки, які відображають результати складання іспитів ЗНО за 2009 рік. Основні характеристик розподілу можна проаналізувати за матеріалами [1].

Усі ознаки (змінні) подані у різних шкалах: категоріальних з великою кількістю категорій (тип населеного пункту, тип школи), номінальних дихотомічних (стать, випускник якого року, чи вступив абітурієнт до ВНЗ), метричних (результати тестів за предметами).

Серед усіх ознак, які досліджувалися, було виділено стать (чоловіча та жіноча), рік випускника (поточний чи минулий рік), тип населеного пункту (визначено лише дві категорії – місто та село) й тип навчального закладу (навчальні заклади об'єднані за категоріями: середні загальноосвітні школи; ліцеї, гімназії, коледжі; професійні заклади освіти та інші; нема інформації). Середній бал ЗНО подано у метричній шкалі 100-200 і за всіма предметами за всією популяцією учнів становив 150,87. На рис. 1 показано розподіл середнього бала ЗНО.

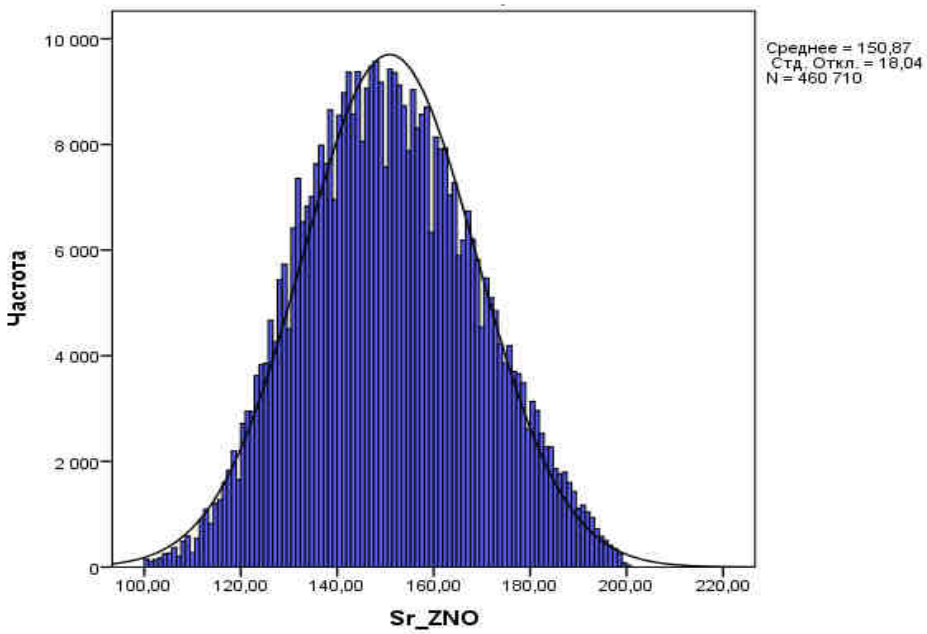


Рис. 1. Розподіл середнього балу ЗНО

Як можна побачити, "пік" графіка розташований у районі 140-160 балів з 200 можливих, тобто саме такі бали отримала порівняно найбільша частина учасників. Однак навіть у районі 160-180 балів гістограма усе ще показує високі значення, тоді як у непрохідній частині (<124) опинилось досить небагато випускників.

Серед загальної кількості абітурієнтів не вступило до ВНЗ 196705, що становить 42,7%. 264005 (57,3%) стали студентами українських вишів. На жаль, яку форму навчання (контрактну чи за кошти держбюджету) обрав абітурієнт – не відома.

Кількість предметів ЗНО відображено у 15 змінних. Крім української мови, усі предмети мали складатися за вибором абітурієнта. При цьому точно не відома,

який предмет абітурієнт використовував як конкурсний – чи з вищим балом, чи з вибраного напрямку. У 2009 році можна було складати до 5 тестів, коли при вступі найчастіше подавалися результати за 2-3 предметами.

З наявних даних неможливо оцінити, йдеться чи про погано складені тести (занадто складний тест з математики, приміром), чи про об'єктивну складність точних наук, чи все ж про серйозні проблеми із викладанням математики і споріднених дисциплін у школі. Але можна оцінити, наскільки такі ознаки, як стать, рік випуску, тип населеного пункту, тип навчального року, впливають на середній бал ЗНО і на результати вступу. Чи є зв'язок між тими предметами, що складав учень, і результатом вступу до ВНЗ?

## 2. Обґрунтування методів аналізу

У процесі обґрунтування вибору методів перш за все були вибрані параметричні та непараметричні методи статистичного аналізу даних. Але ключовий момент полягає в тому, що дані мають досить великий обсяг і велику розмірність, і позбавлені структури та зв'язків. Тому було вирішено, окрім статистичних методів, застосувати технології інтелектуального аналізу даних Data Mining [3-5].

На відміну від статистичних методів, де застосовується технологія перевірки заздалегідь сформульованих гіпотез, на основі Data Mining шукають закономірності у різних підвибірках даних, незважаючи на апріорні припущення щодо структури вибірки і розподілу значень її параметрів. По суті, на основі методів шукають латентні (приховані) зв'язки між даними. Тому окрім відомих непараметричних методів статистики, у Data Mining застосовуються евристичні методи і процедури, які дозволяють виявити асоціативні зв'язки, різні послідовності, визначити ознаки класифікації й кластеризації сукупності даних, які досліджуються. Тому для різних етапів аналізу було обґрунтовано застосування різних методів, які визначалися шкалами оцінювання ознак, видом розподілу, наявністю або відсутністю значень, метою пошуку закономірностей.

Для перевірки зв'язку між середніми балами ЗНО й іншими ознаками: стать, випускник поточного року чи минулих років, із села чи міста абітурієнт, який тип навчального закладу закінчив (середня загальноосвітня школа; ліцей, гімназія, коледж; професійний заклад освіти та ін.) було вибрано дисперсійний аналіз (analysis of variance), тому що *Середнє значення балів ЗНО* є метричною ознакою і розподіл є нормальним за критерієм Колмагорова-Смирнова [6]. Дисперсійний аналіз полягає у виділенні й оцінюванні окремих факторів, що викликають зміну досліджуваної випадкової величини, тобто в нашому випадку – *Середнього значення балів ЗНО*. При цьому проводиться розкладання сумарної вибіркової дисперсії на складові, зумовлені незалежними факторами. Кожна з цих складових є оцінкою дисперсії генеральної сукупності [7]. Щоб дати оцінку дієвості впливу даного фактора, необхідно оцінити значущість відповідної вибіркової дисперсії, порівняно з дисперсією відтворення, зумовленою випадковими факторами. Перевірку значущості оцінок дисперсії проводили за допомогою критерію Фішера.

Для визначення впливу вибраних предметів для вступу до ВНЗ було вибрано один із методів Data Mining, а саме дерева рішень (decision trees) [3-5]. Цей метод базується на створенні ієрархічної структури на основі класифікації правил за типом "ЯКЩО ... ТО ..." (if-then), які мають вигляд дерева. Для прийняття рішення, до якого класу віднести деякий об'єкт або ситуацію (у даному випадку

треба визначити, які саме вибрані предмети частіше є «виграшними» для вступу до ВНЗ), потрібно відповісти на запитання, що стоїть у вузлах цього дерева, починаючи з його кореня. Запитання мають вигляд «значення параметра А більше х». Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативний, – то до лівого вузла; потім знову йде запитання, пов'язане з відповідним вузлом.

Усі розрахунки було зроблено у пакеті IBM SPSS Statistics 23.0, статистичної системи, яка призначена для вирішення дослідницьких і бізнес-задач за допомогою аналізу даних [8].

## 3. Результати аналізу

На рис. 2 показано співвідношення середнього бала ЗНО за факторними ознаками, які найбільше впливають на рівень середнього значення.

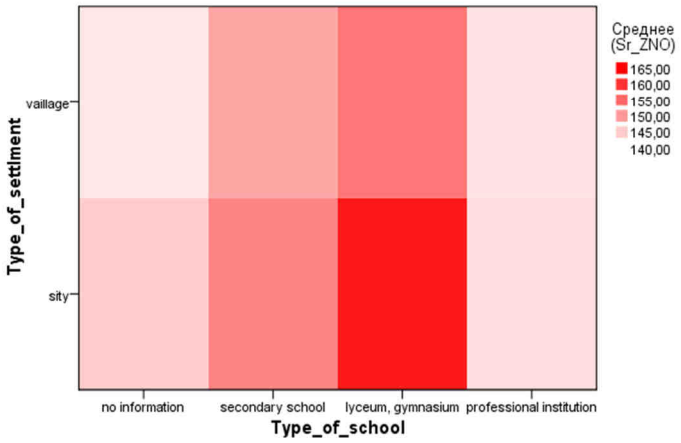


Рис. 2. Діаграма середнього бала ЗНО з типом школи, статтю та типом населеного пункту

Згідно з отриманою діаграмою ми бачимо, що в ліцеях і гімназіях набрано у середньому 160,08 бала, також середні показники абітурієнтів з міста є вищими, ніж показники з села. А для хлопчиків і дівчат значних відмінностей не спостерігається .

Спільний вплив факторних ознак (стать, тип населеного пункту і тип навчального закладу) також є істотним. Не всі випускники “слабких” шкіл мають низькі бали, і не всі випускники з найкращих елітних шкіл мають високі бали. Але школи з вищим рівнем акредитації мають не тільки найкращі результати зі ЗНО, але й найбільш рівномірний розподіл балів серед випускників. Не одиниці, а більшість випускників цих шкіл отримують високі бали. З іншого боку, «звичайні» школи мають гірший середній бал, але не настільки рівномірний розподіл балів: серед їхніх випускників багато таких, хто має погані результати, але є і відмінники.

Протягом дослідження було виявлено, що бали ЗНО значною мірою залежать від того, де вчився учасник тестування. Випускники міських шкіл у середньому мають кращі результати, ніж випускники із сільської місцевості, а учні «елітних» шкіл − гімназій, ліцеїв, колегіумів, спеціалізованих шкіл – з іще більшим

розривом випереджають учнів з більш «звичайних» навчальних закладів. Тип населеного пункту та тип навчального закладу не є незалежними змінними: більшість гімназій та ліцеїв знаходяться у містах. Але навіть якщо порівнювати між собою тільки звичайні школи, перевага міст над селами зберігається, тобто обидва фактори мають значення.

Якими є ці нерівності? За 200-бальною шкалою, середній бал міських випускників з усіх предметів становить 151,58, сільських − 147,45. Нерівність за типом школи і ще більша. «Елітні» школи мають середній бал з усіх предметів 161,38 порівняно з 144,01 у звичайних школах.

Серед значущих факторів можна визначити лише рік випуску – рівень значущості дорівнює 0,008, що свідчить про значущу різницю середніх для випускників різних років, та спільний вплив факторних ознак.

На наступному етапі було проведено аналіз зв’язку результату вступу до ВНЗ із середнім балом ЗНО. Для визначення впливу використовувався метод логістичної регресії й був розрахований точково-бісеріальний коефіцієнт кореляції [9-10].

Коефіцієнт точково-бісеріальної кореляції дорівнює 0,402, що свідчить про низький рівень кореляційного зв’язку. Крім того, ми можемо бачити на рис. 3, як було класифіковано об’єкти. Тут видно дуже великий відсоток невірно класифікованих об’єктів.

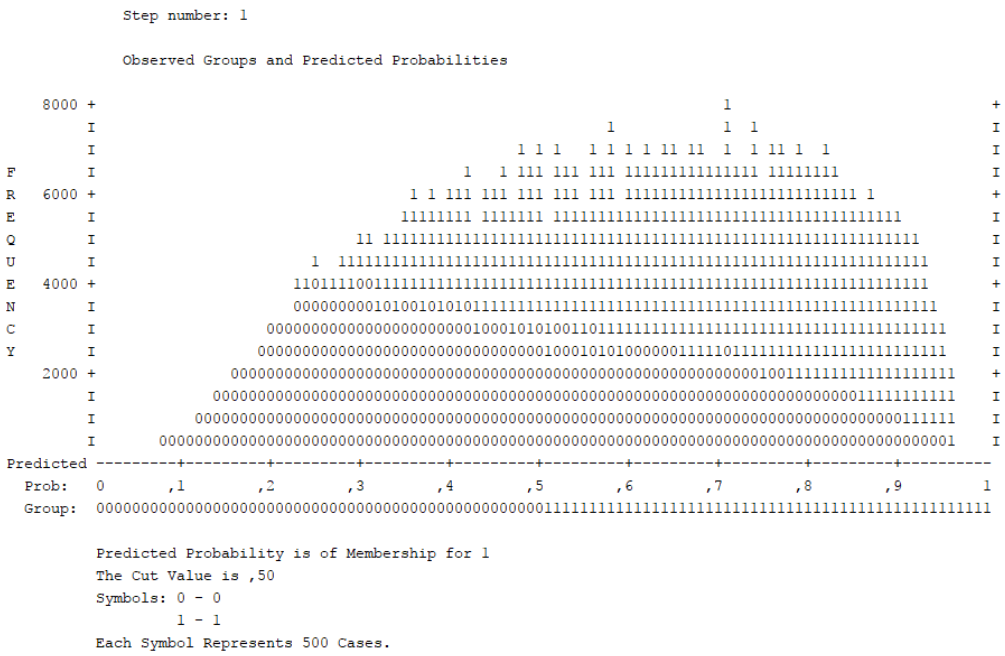


Рис. 3. Класифікація об’єктів за отриманим рівнянням регресії на основі аналізу множини факторних ознак результатів ЗНО 2009 року

Нижче в табл. 1 наведено розраховані ймовірності вступу для кожної групи вступників, які відрізняються середнім балом ЗНО в 10 пунктів.

Як бачимо, ймовірність вступити до ВНЗ у абітурієнтів, які набрали 130 балів, дуже низька, а тих, хто набрав 170 та вище, – дуже висока. При цьому детальний аналіз множини факторів показав, що практично вступники з різних типів шкіл, з різних міст проживання і статі мають шанси вступити до ВНЗ.

Таблиця 1

Ймовірність вступу за моделлю логістичної регресії для різних груп випускників

| Бал | $z$ | $P\{isStudent = 1 \mid Sr\_ZNO\}$ | $P\{isStudent = 0 \mid Sr\_ZNO\}$ |
|---|---|---|---|
| 100 | -2,351 | 0,086986 | 0,913014 |
| 110 | -1,801 | 0,141729 | 0,858271 |
| 120 | -1,251 | 0,222527 | 0,777473 |
| 130 | -0,701 | 0,331591 | 0,668409 |
| 140 | -0,151 | 0,462322 | 0,537678 |
| 150 | 0,399 | 0,598447 | 0,401553 |
| 160 | 0,949 | 0,720914 | 0,279086 |
| 170 | 1,499 | 0,817425 | 0,182575 |
| 180 | 2,049 | 0,885847 | 0,114153 |
| 190 | 2,599 | 0,930797 | 0,069203 |
| 200 | 3,149 | 0,958869 | 0,041131 |

Останнім фактором, що міг би мати вплив, став вибір предметів, результати яких абітурієнт використовував як конкурсі бали. Оскільки не було повних даних відносно того, за результатами яких абітурієнт вступив до ВНЗ, то доцільно було використовувати дерево рішень. Далі прийнято рішення об'єднати за допомогою умов усі предмети, а потім класифікувати у більш вузькі підгрупи.

Необхідно було класифікувати предмети за напрямком. Щоб визначити, на які групи поділити, було використано сайт osvita.ua, за допомогою якого було визначено, які предмети необхідні для вступу до ВНЗ (рис. 4).

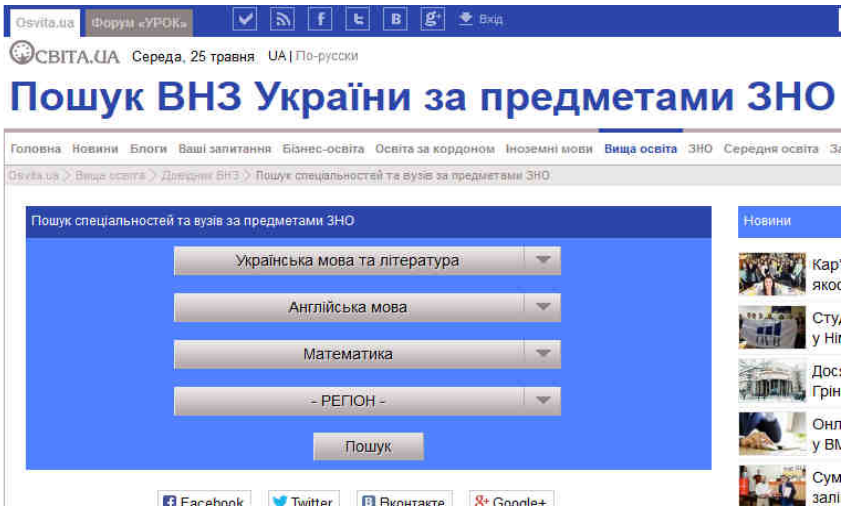


Рис. 4. Сайт для пошуку ВНЗ за предметами ЗНО

Для спрощення розрахунку предмети було об'єднано у меншу кількість категорій, тобто у напрямки. У табл. 2 наведено результати групування предметів за напрямками.

Таким чином, усіх абітурієнтів було поділено на групи за такими напрямками навчання: гуманітарний, технічний, економічний, природничі науки, техніко-економіко-природничий.

Таблиця 2

Групування предметів за напрямками

| ПРЕДМЕТ 1 | ПРЕДМЕТ 2 | НАПРЯМОК |
|---|---|---|
| англійська мова | історія України | Гуманітарний |
| біологія | історія України | |
| англійська мова | географія | Економічний |
| математика | історія України | |
| біологія | географія | Природничі науки |
| біологія | хімія | |
| математика | хімія | |
| математика | фізика | Технічний |
| математика | англійська мова | |
| математика | географія | Технічно-економіко-природничий |

З готовими групами було побудовано дерево рішень для визначення, з якої групи вступило до ВНЗ більше абітурієнтів (рис.5).

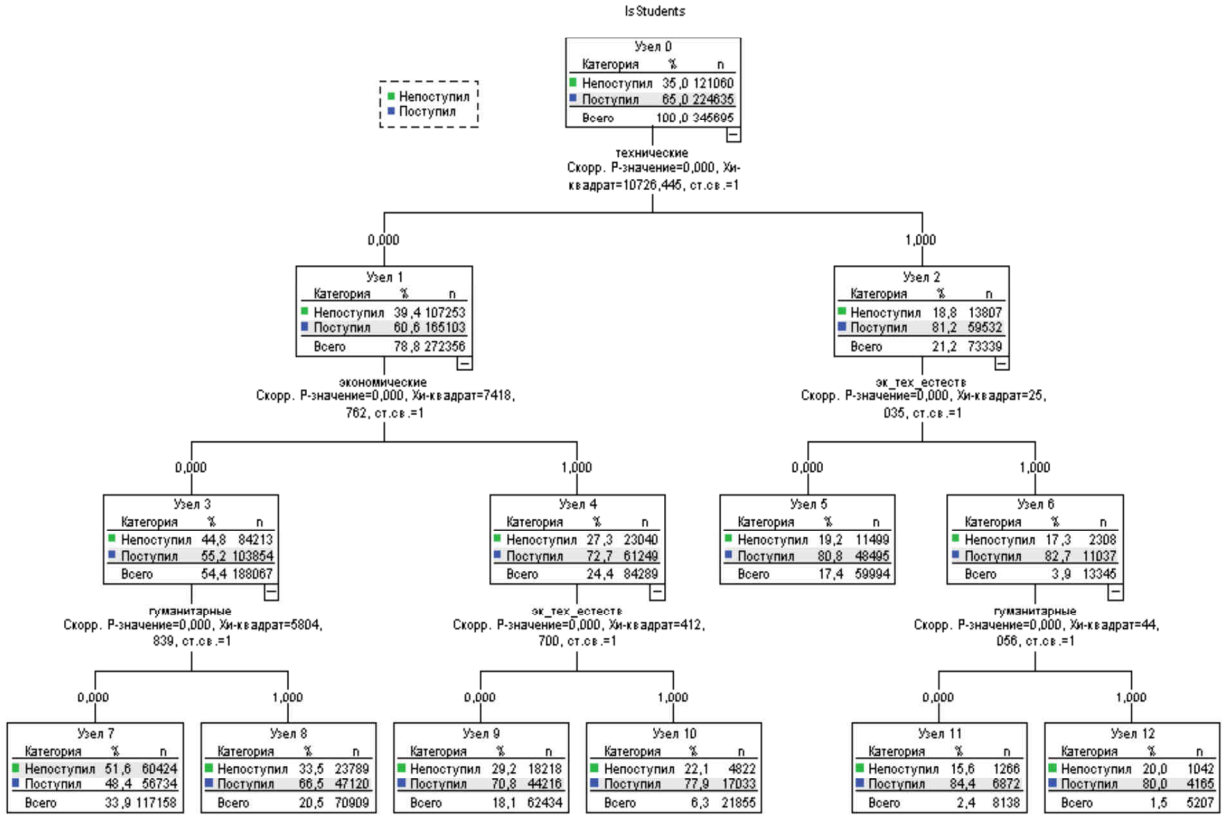


Рис. 5. Дерево класифікації

Для побудови дерева було вибрано алгоритм CHAID, тому що він автоматично виявляє взаємодії за допомогою Хі-квадрат і дає досить стійкі результати незалежно від вибору кореня дерева [11].

На кожному кроці CHAID вибирає незалежну змінну (предиктор), яка має найбільш сильну взаємодію із залежною змінною. Категорії кожного предиктора об'єднуються, якщо вони незначно відрізняються щодо залежної змінної.

Згідно з деревом класифікації 65% із усіх опитаних вступили до ВНЗ, що більше реальної кількості.

Фактором, який розбиває вибірку на дві найбільш сильно різні групи і який має найбільш значиму взаємодію із залежною змінною IsStudents (тобто є студентом чи ні), є технічний напрямок. Іншими словами, алгоритм CHAID автоматично визначив, що вступив чи не вступив абітурієнт до ВНЗ, найбільше залежить від технічного напрямку. З них 81,2% вступили до ВНЗ, а 18,8% − ні.

У вузлі абітурієнтів, які не відносяться до технічного напрямку, вступили 60,6%, а не вступили 40,4%. Таким чином, видно, що у цьому вузлі всього на 5% менше вступили до ВНЗ, ніж у кореневому вузлі. Всього до цього сегмента потрапило 272356 чоловік, що становить 78,8% від усіх абітурієнтів. Далі цей вузол був поділений за змінною економічний напрямок, з них вступили 72,7%. Через те, що ми не мали точних даних, які предмети абітурієнт подавав до ВНЗ, одна людина могла потрапити не в одну групу, а в декілька, тому вузол нетехнічний напрямок був поділений змінною гуманітарний напрямок, в якій 66,5% вступили до ВНЗ, це становить 20,5% від усіх абітурієнтів.

Тобто за допомогою дерева ми змогли, маючи людину, яка потрапила до декількох груп,зрозуміти, які предмети найімовірніше вона подала, щоб вступити до ВНЗ.

У вузлі, що потрапили до технічного напрямку, 81,2% вступили до ВНЗ, а 18,8% − ні, ця група становить усього 21,2% від усіх абітурієнтів, і ми робимо висновок, що технічний напрямок не є фактором, який гарантує вступ до ВНЗ. Далі вузол технічний напрямок був розбитий змінною «економічнотехнічноприродничи» напрямок, тобто ті ж самі абітурієнти здавали такі предмети, що відносяться і до цієї групи. З них вступили 82,3% чоловіка, що становить 3,9% від усіх абітурієнтів. Потім цей вузол був розбитий змінною гуманітарний напрямок, з них вступили до ВНЗ 80%, що становить всього 1,5% від усіх опитаних. Тому можна зробити висновок, що швидше за все ці абітурієнти вибрали технічний напрямок, хоча і здавали інші предмети, і процент абітурієнтів, що вступили до ВНЗ становить 81,2%.

## Висновки

Таким чином, на середній бал ЗНО значно впливають ознаки: тип навчального закладу та тип населеного пункту. Згідно з деревом класифікації 65% із усіх опитаних вступили до ВНЗ. Абітурієнтів, які потрапили до технічного напрямку, вступило 81,2%, але відносно загальної кількості цей напрямок вибрали всього 21,2% абітурієнтів. Тому цей фактор не є значущим. Крім того, розглянувши останні чотири напрямки, ми робимо висновок, що результативність вступу не визначається вибором предметів.

**Список літератури**

1. Дослідження якості конкурсного відбору студентів вищих навчальних закладів за результатами зовнішнього незалежного оцінювання: аналітичні матеріали / за ред. В. В. Ковтунця і С. А. Ракова. – К.: Нора-Друк, 2015. – 160 с.
2. Ковтунець В.В., Дослідження прогностичної валідності критеріїв вступу до ВНЗ (вступ 2012-2014 років) / В.В. Ковтунець, С.А. Раков, М.С. Мазорчук та ін.// Вісник ТІМО. 2017 – №2-3.– С. 4-51.
3. Han, Jiawei. Data mining: concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei, Elsevier Inc., 2005. – 3rd ed. 740 p.
4. Witten, I. H. (Ian H.) Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank, Elsevier Inc., 2012. – 2nd ed. 558 p.
5. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman, Springer Science & Business Media, 2009. – 2nd ed. 745 p.
6. Lilliefors, H.W. On the Kolmogorov- Smirnov test for normality with mean and variance unknown [Text] / H.W. Lilliefors. — J. Am. Statist. Assoc., 1967. — P. 399-402.
7. Шеффе, Г. Дисперсионный анализ / Г. Шеффе. – М.: Наука, 1980. – 512 с.
8. Бююль Ахим, Цёфель Петр. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: пер. с нем. / Ахим Бююль, Петр Цёфель – Спб.: «ДиаСофтЮП», 2005 – 608 с.
9. Hosmer, David W. Applied logistic regression / David W. Hosmer, Jr. Stanly Lemeshov, John Wiley & Sons, Inc. – 2nd ed. 397 p.
10. Гласс, Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стэли. – М.: Прогресс, 1976. – 496 с.
11. Magidson, J. The CHAID approach to segmentation modeling: CHI-squared automatic interaction detection. In R. P. Bagozzi (ed.), Advanced Methods of Marketing Research. - P. 118–159. Blackwell Business, 1994.

Надійшла до редакції 05.12.2017

# Исследование факторов влияния на поступление в высшие учебные заведения на основе методов Data Mining

В работе рассмотрены методы и модели выявления значимых факторов, влияющих на результат поступления абитуриентов в высшие учебные заведения (ВУЗ). В качестве примера для анализа использовано базу данных по результатам поступления 2009 года. Общая популяция исследуемых составила 460 710 абитуриентов. Анализ выполнялся по 43 факторными признакам. В процессе анализа предложено использовать комплекс методов, базирующихся как на статистической теории анализа данных, так и на методах интеллектуального анализа (Data Mining). Все расчеты выполнены с использованием известного инструментария для статистической обработки данных SPSS. Главным

результатом исследования являются выводы относительно справедливости поступления абитуриентов в ВУЗы на основании результатов национального экзамена - внешнего независимого оценивания (ВНО).

***Ключевые слова***: анализ данных, Data Mining, внешнее независимое оценивание, факторные признаки, справедливость поступления.

# Investigation of the Factors of Influence on Entry into Higher Educational Institutions on the Basis of Data Mining Methods

The paper addresses the methods and models for identifying significant factors that influence the outcome of the entrance of entrants to higher education institutions (HEIs). As an example for analysis, the database was used based on the results of the 2009 admission. The total population of the study population was 460,710 entrants. The analysis was performed on 43 factor factors. In the process of analysis, it is proposed to use a set of methods based on both the statistical theory of data analysis and methods of data mining (Data Mining). All calculations are performed using the well-known tool for statistical processing of SPSS data. The main result of the study is the conclusions about the fairness of admittance to higher education institutions on the basis of the results of the national exam - External Independent Evaluation (External Independence).

***Keywords***: Data Analysis, Data Mining, External Independent Evaluation, Factorial Characteristics, Accession Justice.

**Сведения об авторах:**

**Мазорчук Мария Сергеевна** – к.т.н., доцент кафедры информатики Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», Украина; e-mail: mazorchuk_mary@inbox.ru.

**Трофимова Ирина Алексеевна** – старший преподаватель кафедры информатики Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», Украина; e-mail: trophimova_iryna@gmail.com.

**Пантелеева Анна Юрьевна** – студентка 365ам группы Национального аэрокосмического университета им. Н. Е. Жуковского «ХАИ», Украина. e-mail: anncami.94@gmail.com.