УДК 61:681.3.002
DOI: https://doi.org/10.11603/mie.1996-1960.2018.4.9840

ЗМЕНШЕННЯ РОЗМІРНОСТІ ПРИ ПРОГНОЗУВАННІ УСПІШНОСТІ В ЗАВДАННЯХ МЕДИЧНОЇ ОСВІТИ: ПІДХІД НА ОСНОВІ МЕТОДУ PCA

В. П. Марценюк, Ю. В. Дроняк[1], І. В. Цікорська[1]

Університет Бєльсько-Бяли, Республіка Польща
[1]ДВНЗ «Тернопільський державний медичний університет імені І. Я. Горбачевського МОЗ України»

У роботі запропоновано підхід на основі використання методу PCA з метою виявлення головних чинників (поточної або підсумкової успішностей), що впливають на результати ліцензійного інтегрованого іспиту.

Ключові слова: медична освіта, метод головних компонент, зменшення розмірності, PCA, машинне навчання, R.

REDUCTION OF DIMENSION FOR PREDICTION OF PROGRESS IN PROBLEMS OF MEDICAL EDUCATION: AN APPROACH BASED ON METHOD PCA

V. P. Martsenyuk, Yu. V. Dronyak[1], I. V. Tsikorska[1]

University of Bielsko-Biala, the Republic of Poland
[1]SHEE I. Horbachevsky Ternopil State Medical University of the Ministry of Health of Ukraine

Background. System research in medical science and education is often based on data with a large number of attributes. Therefore, the problem of reducing the dimension of the data is actual, while keeping data with as much variation as possible. At the same time, such problems can be considered as problems of machine learning, which can be solved using artificial intelligence algorithms (for example, artificial neural networks.) The interpretation of the results of the application of classification algorithms without prior reduction of the dimension of such tasks is a difficulty for the decision maker. The results may be simplified by the reduction of the dimension of the task.

Materials and methods. The method proposed and applied in this paper is described in terms of machine learning. Namely, the mathematical description of the tasks of machine learning in research in the field of medical education is based on the following data. We have a set consisting of tuples. Depending on the problem under consideration, certain sets of these tuples will be used for training, testing and prediction. An arbitrary tuple consists of input data (we will call them according to the established terminology of machine learning as attributes) and output data, which are attributes of the class. As a result of reducing the dimension, we obtain a certain numerical matrix, where these data can be used in variety of problems of machine learning. In order to study the possibilities of reducing the dimension of the medical education assessment tasks, the results of the medical students at TSMU were considered (the volume of the general population was 248).

Results. In the work an approach which is based on application of the method PCA with the purpose of determining principle reasons (current and final progresses), which influence on the results of license integrated exam, is offered.

Using the visualization, approach allows us to represent data on a plane in the coordinates of the first two main components. This allows us to identify the main changes in the attributes that cause the data to belong to certain groups. So, using such a visualization to the student data, we can get graphic representations of the «main» performance indicators (current and final) that predict the passing the licensed integrated exam «Step 1».

Conclusions. Reduction of dimension is a very important element in the analysis of statistical data in medical education. Due to reducing the dimensions, we achieve not only that we simply need to analyze less data. The most important is that after the reduction, they often show us more information than before reducing the dimension. Many dependencies in medical education become more readable. It often happens that the amount of data we have is beyond the scope of our perception and then it is very easy to get lost in their analysis.

Key words: medical education, principle component method, dimension reduction, PCA, machine learning, R.

УМЕНЬШЕНИЕ РАЗМЕРНОСТИ ПРИ ПРОГНОЗИРОВАНИИ УСПЕВАЕМОСТИ В ЗАДАЧАХ МЕДИЦИНСКОГО ОБРАЗОВАНИЯ: ПОДХОД НА ОСНОВЕ МЕТОДА PCA

В. П. Марценюк, Ю. В. Дроняк[1], И. В. Цикорская[1]

Университет Бельско-Бялы, Республика Польша
[1]ГВУЗ «Тернопольский государственный медицинский университет имени И. Я. Горбачевского МЗ Украины»

В работе предложен подход на основе использования метода PCA с целью выявления главных причин (поточной либо итоговой успеваемости), влияющих на результаты лицензированного интегрированного экзамена.

Ключевые слова: медицинское образование, метод главных компонент, уменьшение размерности, PCA, машинное обучение, R.

Вступ. Системні дослідження в медичній науці та освіті дуже часто ґрунтуються на даних із значною кількістю ознак (атрибутів) [1-7]. Тому актуальною є проблема зменшення розмірності даних, при цьому зберігаючи дані з якомога більшою варіацією.

Так прийняття рішень у медичній освіті ґрунтується на даних значних обсягів (наприклад, кількість студентів), зважаючи на значну кількість показників (наприклад, результати поточної та підсумкової успішності із значної кількості навчальних дисциплін) [8]. При цьому такі проблеми можуть розглядатися як завдання машинного навчання, що можуть бути вирішені за допомогою алгоритмів штучного інтелекту (наприклад, штучних нейронних мереж [9, 10]). Трактування результатів застосування алгоритмів класифікації без попереднього зменшення розмірності таких завдань становить складність для особи, що приймає рішення. Спростити представлення таких результатів може попереднє застосування зменшення розмірності задачі. Такий підхід ґрунтується на тому, що досить часто лише кілька показників успішності (з певних конкретних навчальних дисциплін) впливають на належність студентів до певної групи успішності (наприклад, щодо складання ліцензійного інтегрованого іспиту).

Під зменшення розмірності зазвичай розуміють процес перетворення багатовимірних даних (із значенні великої кількості атрибутів) у простір з набагато меншим розміром. Із суто практичних причин (можливість легкої візуалізації) дані зазвичай зменшуються до двох або трьох вимірів. На практиці часто виявляється, що багато атрибутів є один із одним досить сильно пов'язані (корельовані). Тому, щоб отримати повну картину описаного явища або помітити певні закономірності в даних, досить розглянути тільки невелику їх частину або на основі оригінальних атрибутів згенерувати зменшений набір на основі нових «прихованих атрибутів». Зменшення розмірності зазвичай передбачає втрату певної кількості інформації. В ідеальному випадку дані повинні бути зменшені до такої міри, що втрата не відбудеться. Тоді ми вважаємо, що існує так звана внутрішня розмірність, що є менша, ніж вихідна розмірність аналізованих даних, яка гарантує, що дані після скорочення будуть нести таку ж інформацію, що й нередуковані дані. В даній роботі ми не розглядаємо питання оцінювання згаданої внутрішньої розмірності.

Слід зазначити, що галузь скорочення розмірності є дуже широкою і тут існує багато різних підходів та методів. Ми зупинимося лише на методі аналізу головних компонент (PCA), як на одному з найбільш поширених.

Важливим етапом, що передує правильному аналізу даних (включаючи зменшення розмірності), є їх попереднє оброблення. Мета такого оброблення — видалення або обмеження різних недосконалостей і певних несприятливих властивостей, що зустрічаються в даних і що можуть мати дуже негативний вплив на отримані результати, в крайніх випадках повністю їх фальсифікувати. Мається на увазі, перш за все, способи подолання відсутності даних шляхом різних перетворень і масштабування даних.

Мета роботи: розробити та запропонувати модифікацію алгоритму головних компонент PCA з метою зменшення розмірності в завданнях прийняття рішень у медичній освіті, продемонструвавши підхід на прикладі проблеми прогнозування складання ліцензійного інтегрованого іспиту «Крок–1».

Постановка завдання. Одним із головних напрямів підвищення ефективності підготовки майбутніх спеціалістів-медиків є впровадження інформаційних, комунікаційних і комп'ютерних технологій, що дає змогу реалізації інноваційних методів організації навчального процесу в медичній освіті [8]. Застосування новітніх технологій вимагає нового підходу й до систем оцінювання майбутніх лікарів. На даний час активно впроваджуються в освітню практику методики комп'ютерного оцінювання знань та розробляються інструментальні засоби комп'ютерних систем тестування. В освітніх установах усе більш масового характеру набуває впровадження систем дистанційного навчання [8]. Основні тенденції наукових досліджень нових ефективних методик контролю знань спрямовані на оцінювання з використанням багатьох методів та конструктивного зворотного зв'язку.

Отже, інноваційні підходи щодо оцінювання студентів-медиків дозволяють зібрати значні обсяги даних про навчальну діяльність. Зазначимо, що, наприклад, в аналізованому далі прикладі лише дані про поточну та підсумкову успішність студентів 4-го курсу з шести навчальних дисциплін містять 58 показників.

Таким чином, універсальний доступ до ефективних і відносно дешевих систем баз даних робить обсяг збережених даних величезним, що постійно зростає. Ці дані часто характеризуються високою розмірністю, що стає істотною проблемою під час їх аналізу. Говорячи про велику розмірність, маємо на увазі окремі кортежі даних (наприклад, записи в реляційних таблицях про успішність студентів), що мають велику кількість атрибутів (змінних). На практиці, однак, часто виявляється, що багато з цих атрибутів тісно пов'язані між собою. Отримати повну картину описаних явищ або повідомлень про певні регулярності в даних можна, розглянувши лише невелику їх частину.

Аналіз головних компонент (PCA) є класичним методом зменшення розмірності. За допомогою певних лінійних перетворень створюється нова система координат, причому така, що проекція вхідних даних на першу вісь має найбільшу можливу дисперсію (варіацію), проекція на другу вісь має другу за величиною можливу дисперсію і так далі.

Тобто, можемо «вирізати» решту вимірів як «недійсні», оскільки вони мають дуже малу дисперсію. В результаті створюється нова множина змінних, найменша можлива підмножина якої здатна достатньо точно відобразити мінливість вхідного набору даних. Новостворені змінні зазвичай не мають розумної фізичної інтерпретації оригінальних змінних. Однак, великою перевагою цього методу є те, що часто невелика підмножина нових змінних (2 або 3) пояснює дуже великий відсоток первинної мінливості та стає можливим зображення багатовимірних наборів на 2D або 3D представленнях.

Матеріали та методи дослідження. Запропонований і застосований у даній роботі метод описується в термінах машинного навчання. А саме, математичний опис завдань машинного навчання в дослідженнях у галузі медичної освіти ґрунтується на таких даних. Маємо множину D, що складається з N кортежів. Залежно від проблеми, що розглядається, певні множини цих кортежів використовуватимуться для навчання, тестування та прогнозування. Довільний і-й кортеж $\left(a_1^i, a_2^i, \text{K}, a_p^i, c^i\right)^T$ складається з вхідних даних $\left(a_1^i, a_2^i, \text{K}, a_p^i\right)^T$ (називатимемо їх відповідно до усталеної термінології машинного навчання атрибутами) та вихідних даних сі, що є атрибутами класу.

Нехай вектор-рядок $a_j = \left(a_j^1, a_j^2, \text{K}, a_j^N\right)$ представляє значення -го атрибуту всіх N кортежів. Атрибути a_1, K, a_p можуть приймати як числові, так і категоріальні значення. Атрибут класу С приймає одне з К дискретних значень: $c \in \{1, \text{K}, K\}$. Метою є спрогнозувати, використовуючи деякий предиктор, значення атрибуту класу С на основі значень атрибутів a_1, K, a_p. При цьому слід максимізувати точність прогнозування атрибуту класу, а саме ймовірність $P\{c = c^*\}$ для довільного $c \in \{1, \text{K}, K\}$.

Найпершою проблемою для вирішення в системних дослідженнях медичної освіти є зменшення розмірності $p \in \text{N}$. Пропонуємо наступну модифікацію методу аналізу головних компонент (PCA) із такими кроками.

Вхідні дані: $A = \left\{\left(a_1^i, a_2^i, \text{K}, a_p^i, c^i\right)^T\right\}_{i=1}^N$.

Вихідні дані: головні компоненти разом із атрибутами:

1. Перетворення всіх категоріальних атрибутів, кодуючи їх як булевські входи, кожен із яких представляє одну категорію зі значеннями 0 або 1[1]. У результаті отримуємо чисельну матрицю $X = \left\{\left(x_1^i, x_2^i, \text{K}, x_{p_1}^i, c^i\right)^T\right\}_{i=1}^N \in R^{p_1+1 \times N}$.

2. Обчислюємо середні значення по рядках: $\overline{x_i} = \frac{1}{N}\sum_{j=1}^{N} x_i^j$, $i = \overline{1, p_1}$.

3. Обчислюємо варіації $Var(x_i)$, $i = \overline{1, p_1}$. Нехай $Var(X) = \sum_{i=1}^{p_1} Var(x_i)$ називається загальною варіацією (сумою варіацій вибірок).

4. Обчислюємо матрицю відхилень: $X' = \left\{x_j^i - \overline{x_i}\right\}_{i=\overline{1,p_1}, j=\overline{1,N}} \in R^{p_1 \times N}$.

5. Обчислюємо коваріаційну матрицю $C = \frac{1}{p_1} X'(X')^T \in R^{p_1 \times p_1}$.

6. Обчислюємо власні значення матриці C: $\lambda_1 \le \lambda_2 \le \text{K} \le \lambda_{p_1}$.

7. Обчислюємо власні вектори матриці C. Розглянемо власні вектори $w_{p_1}, w_{p_1-1} \in R^{p_1}$, що відповідають λ_{p_1} та λ_{p_1-1} відповідно. Покладаємо перших два головних компоненти як $PC1 = X^T w_{p_1}$ та $PC2 = X^T w_{p_1-1}$. Обчислюємо варіації $Var(PC1)$ та $Var(PC2)$. Звідси отримуємо частки поясненої варіації, що відповідають першим двом компонентам. А саме, $ExplainedVar(PC1) := \frac{Var(PC1)}{Var(X)}$ і $ExplainedVar(PC2) := \frac{Var(PC2)}{Var(X)}$ відповідно.

8. Впорядковуємо значення власних векторів w_{p_1} та w_{p_1-1} у спадаючому порядку їх абсолютних значень. З цією метою застосовуємо перестановки $\pi(w_{p_1})$ та $\pi(w_{p_1-1})$. Повертаємо назви перших $ExplainedVar(PC1) \times 100\%$ атрибутів у перестановці $\pi(w_{p_1})$ та перших $ExplainedVar(PC2) \times 100\%$ атрибутів у перестановці $\pi(w_{p_1-1})$.

В результаті зменшення розмірності отримуємо деяку чисельну матрицю $X^{red} = \left\{\left(x_1^i, x_2^i, \text{K}, x_{p_2}^i, c^i\right)^T\right\}_{i=1}^{N} \in R^{p_2+1 \times N}$, де $p_2 \le p_1$. Далі ці дані можуть бути використані в ряді проблем машинного навчання.

Результати та їх обговорення. З метою вивчення можливостей зменшення розмірності завдань оцінювання медичної освіти розглядалися результати успішності студентів 4-го курсу медичного факультету Тернопільського державного медичного університету (ТДМУ) (об'єм генеральної сукупності – 248), використані в роботах [2, 3].

Усуваючи кортежі з відсутніми значеннями певних атрибутів об'єм було зменшено до N = 242. При цьому маємо атрибутів, що включають значення поточної та підсумкової успішності з окремих навчальних модулів. Очевидно, що така кількість атрибутів є занадто велика, щоб встановити ті з них, які найбільше вплинули на складання ліцензійного інтегрованого іспиту «Крок–1». Тому для зменшення розмірності освітньої проблеми використовували модифікований метод PCA.

В якості реалізації методу було використано мову програмування R та середовище RStudio 3.5.1. З метою візуалізації методу PCA використано пакет ggbiplot, індукція дерева рішень реалізована за допомогою пакету C50.

У назвах атрибутів використано такі умовні позначення. Повна назва атрибуту складається з трьох частин: назва дисципліни + номер модуля + вид успішності. При цьому вид успішності позначається таким чином: «c» – поточна успішність, «t» – підсумкова успішність.

Дерево рішень прогнозування успішності складання ліцензійного іспиту «Крок–1» на основі 58 показників поточної та підсумкової успішності представлено на рис. 1.

Повні дані щодо отриманого дерева рішень наведено в дод. 1. Розмір (тобто кількість рівнів) дерева рис. 1 складає 9. Похибка прогнозування на основі дерева рішень складає 0.8 % (2 випадки з 242). При цьому ці 2 випадки стосуються невірного прогнозування для двох студентів, які насправді не склали ліцензійний іспит.

Використовуючи процедуру методу головних компонент на кроці 7 пораховано 58 головних компонент (від PC1 до PC58). З дод. 2 видно, що перших дві головні компоненти разом несуть 52.9 % загальної варіації. Зважаючи на абсолютні значення елементів PC1 (див. дод. 3) та PC2 на кроці 8 було відібрано атрибутів, що були використані для побудови дерева рішень (рис. 2).

При цьому дерево містить 7 рівнів, а величина похибки складає 3.3 % (див. дод. 4). Помилково були класифіковані 8 студентів (із 242). Причому 7 із них насправді не склали ліцензійного іспиту. Звернемо увагу, що таке «зменшене» дерево рішень ґрунтується лише на використанні трьох найінформативніших атрибутів: «histology1c», «pathology2c» та «chemistry5c», а дерево рішень, збудовано на основі всіх даних, використовує 6 атрибутів.

Висновки. Отже, робота показує можливості використання зменшення розмірності та візуалізації багатовимірних даних та пов'язаних із ними завдань на прикладі проблеми медичної освіти.

У роботі запропоновано підхід на основі використання методу PCA з метою виявлення головних

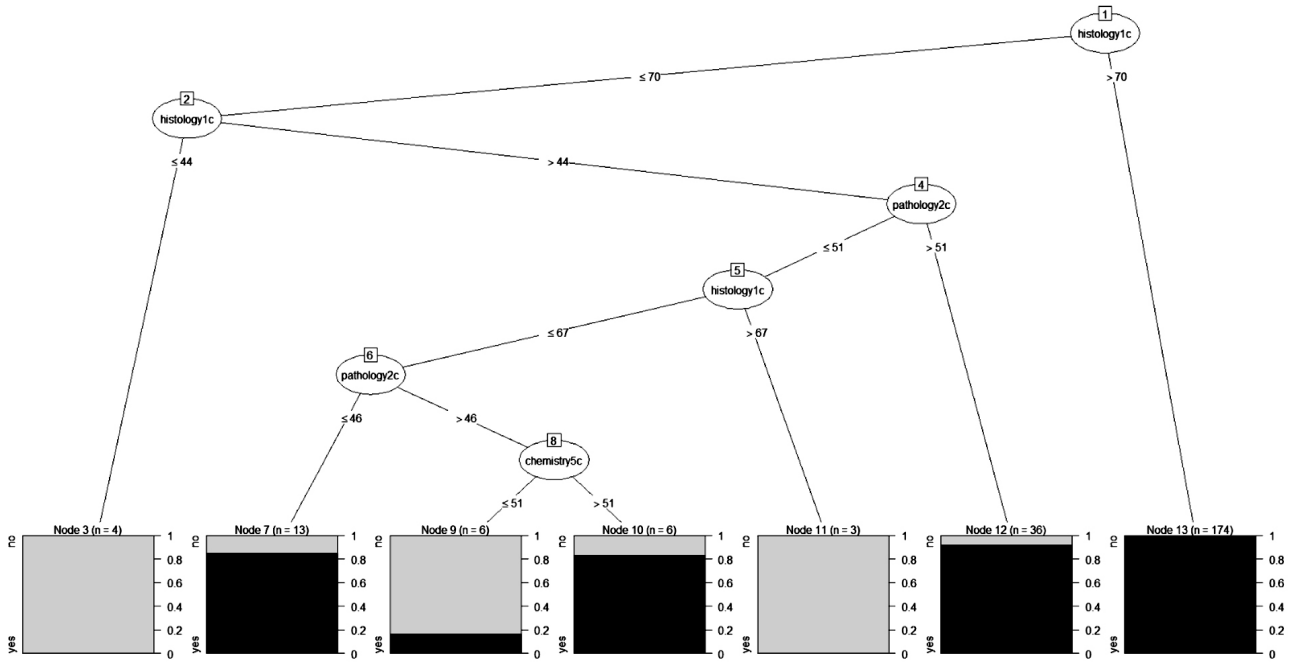

Рис. 1. Дерево рішень прогнозування успішності складання ліцензійного іспиту «Крок–1» на основі 58 показників поточної та підсумкової успішності

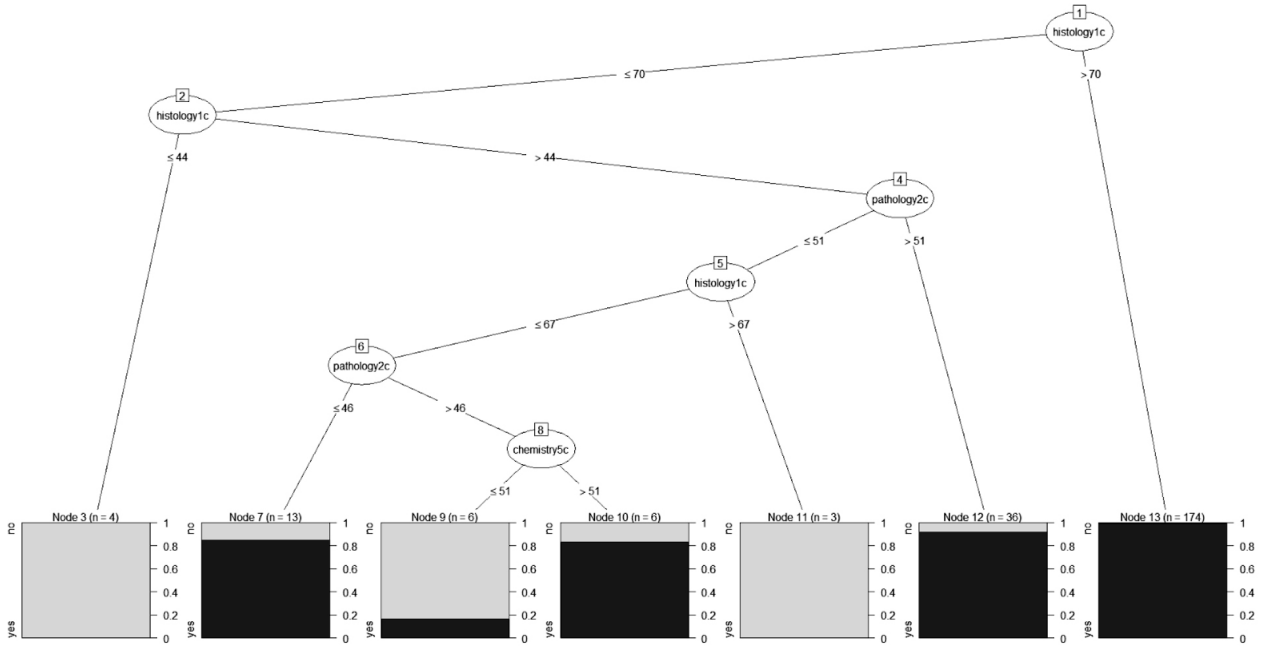

Рис. 2. Дерево рішень прогнозування успішності складання ліцензійного іспиту «Крок–1» на основі показників поточної та підсумкової успішності

чинників (поточної або підсумкової успішностей), що впливають на результати ліцензійного інтегрованого іспиту.

Використовуючи візуалізацію підхід дозволяє представити дані на площині в координатах двох перших головних компонент. Це дозволяє визначити головні зміни в атрибутах, що спричиняють належність даних до певних груп. Так, застосовуючи таку візуалізацію до даних про успішність студентів, можемо отримати графічні представлення «головних» показників успішності (поточної та підсумкової), що прогнозують складання/нескладання ліцензійного інтегрованого іспиту «Крок–1».

Звертаємо увагу, що зменшення розмірності є дуже важливим елементом аналізу статистичних даних, оскільки завдяки зменшенню

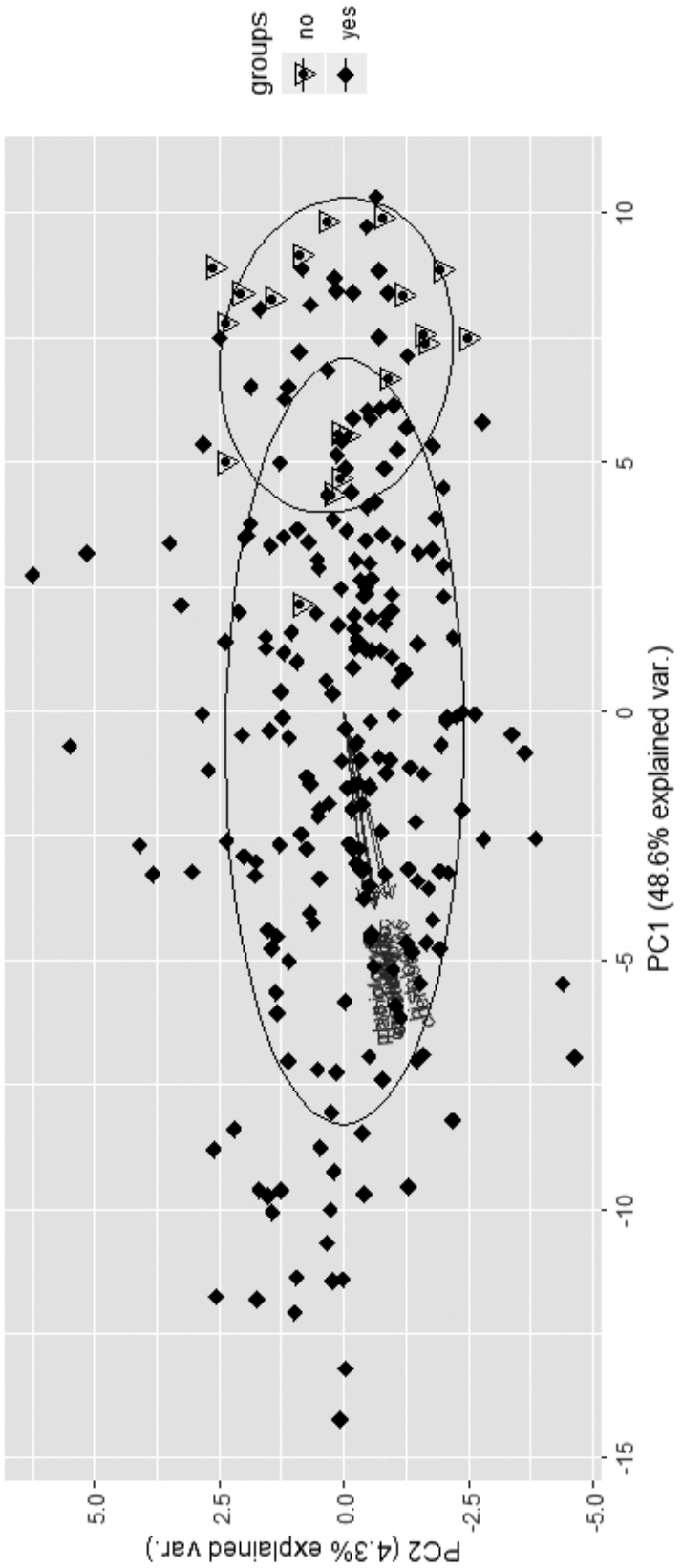

Рис. 3. Візуалізація даних на площині PC1-PC2, що виконана за допомогою пакету ggbiplot. Представлено дві групи студентів: «yes» – які склали ліцензійний іспит, «no» – які не склали ліцензійний іспит «Крок–1». Стрілки вказують на 11 «найбільш варіативних» атрибутів у компонентах PC1 та PC2

розмірності досягається зменшення даних для аналізу. Найважливішим, на думку авторів, зрештою, є те, що дані після скорочення часто показують нам більше інформації, ніж до скорочення розмірності. Багато залежностей стають більш читабельними та знижується ймовірність їх втрати. Дуже часто трапляється так, що кількість наявних даних виходить за межі можливостей нашого сприйняття і тоді дуже легко «заблукати» в їх аналізі. Зазначимо також, що, звичайно, не всі дані можуть бути автоматично усунені зважаючи лише на їх малу варіативність.

Література.

1. Марценюк В. П. Впровадження в навчальний процес комп'ютерних технологій / В. П. Марценюк // Медична освіта. — 2007. — № 2. — С. 40-41.
2. Марценюк В. П. Розробка і впровадження системи електронного навчання в Тернопільському державному медичному університеті імені І. Я. Горбачевського / В. П. Марценюк // Медична освіта. — 2008. — № 2. — С. 74-75.
3. Марценюк В. П. О структуре базы данных информационной системы проверки знаний в медицинском образовании / В. П. Марценюк, А. В. Семенец // Штучний інтелект. — 2009. — № 1. — С. 267-277.
4. Марценюк В. П. Концептуальные подходы к структуре информационной системы проверки знаний в медицинском образовании / В. П. Марценюк, А. В. Семенец // Кибернетика и вычислительная техника. — 2009. — Вып. 156. — С. 18-27.
5. Ковальчук Л. Я. Впровадження в навчальний процес комп'ютерних технологій / Л. Я. Ковальчук, В. П. Марценюк // Медична інформатика та інженерія. — 2008. — № 1. — С. 14-16.
6. Ковальчук Л. Я. Об'єктивізація системи оцінювання знань студентів у Тернопільському державному медичному університеті імені І. Я. Горбачевського / Л. Я. Ковальчук, В. П. Марценюк, П. Р. Сельський // Науковий журнал МОЗ України. — 2012. — № 2. — С. 104-110.
7. Ковальчук Л. Я. Обґрунтування використання інформаційних технологій для підготовки лікарів сімейної медицини та покращення якості медичної допомоги на первинному рівні / Л. Я. Ковальчук, В. П. Марценюк, П. Р. Сельський // Клиническая інформатика и Телемедицина. — 2012. — Т. 8. — Вып. 9. — С. 141-145.
8. Марценюк В. П. Інформаційна система управління якістю підготовки фахівців у вищій медичній освіті / В. П. Марценюк, П. Р. Сельський. — Тернопіль: ТДМУ, 2015. — 312 с.
9. Марценюк В. П. Аналіз результатів семестрових комплексних тестових іспитів в медичній освіті на основі кореляційних показників успішності та багатопараметричної нейромережевої кластеризації / В. П. Марценюк, О. О. Стаханська // Медична інформатика та інженерія. — 2010. — № 1. — С. 53-57.
10. Марценюк В. П. Нейромережеве прогнозування складання студентами-медиками ліцензійного інтегрованого іспиту «Крок 1» на основі результатів поточної успішності та семестрового комплексного тестового іспиту / В. П. Марценюк, А. В. Семенець, О. О. Стаханська // Медична інформатика та інженерія. — 2010. — № 2. — С. 57-62.

References.

1. Martseniuk V. P. (2007). Vprovadzhennia v navchalnyi protses kompiuternykh tekhnolohii [Implementation of the computer technology educational process]. Medychna osvita (Medical education), 2, 40-41. [In Ukrainian].
2. Martseniuk V. P. (2008) Rozrobka i vprovadzhennia systemy elektronnoho navchannia v Ternopilskomu derzhavnomu medychnomu universyteti imeni I. Ia. Horbachevskoho [Development and introduction of e-learning system at Ternopil State Medical University named after I. Ya. Gorbachevsky]. Medychna osvita (Medical education), 2, 74-75. [In Ukrainian].
3. Martseniuk V. P., Semenets A. V. (2009). O strukture bazyi dannyih informatsionnoy sistemyi proverki znaniy v meditsinskom obrazovanii [On the structure of the database of information system for testing knowledge in medical education]. Shtuchnyi intelekt (Artificial Intelligence), 1, 267-277. [In Russian].
4. Martseniuk V. P., Semenets A. V. (2009). Kontseptualnyie podhodyi k strukture informatsionnoy sistemyi proverki znaniy v meditsinskom obrazovanii [Conceptual approaches to the structure of the information system for testing knowledge in medical education]. Kibernetika i vyichislitelnaya tehnika (Cybernetics and Computer Engineering), 156, 18-27. [In Russian].
5. Kovalchuk L. Ia., Martseniuk V. P. (2008). Vprovadzhennia v navchalnyi protses kompiuternykh tekhnolohii [Implementation of the computer technology educational process]. Medichna informatika ta inzheneriya (Medical Informatics & Engineering), 1, 14-16. [In Ukrainian].
6. Kovalchuk L. Ia., Martseniuk V. P., Selskyi P. R. (2012). Obiektyvizatsiia systemy otsiniuvannia znan studentiv u Ternopilskomu derzhavnomu medychnomu universyteti imeni I. Ya. Horbachevskoho [Objectivization of the student assessment system at the I. Ya. Gorbachevsky Ternopil State Medical University]. Naukovyi zhurnal MOZ Ukrainy (Scientific journal of the Ministry of Health of Ukraine), 2, 104-110. [In Ukrainian].
7. Kovalchuk L. Ia., Martseniuk V. P., Selskyi P. R. (2012). Obgruntuvannia vykorystannia informatsiinykh tekhnolohii dlia pidhotovky likariv simeinoi medytsyny ta pokrashchennia yakosti medychnoi dopomohy na pervynnomu rivni [Substantiation of the use of information technologies for the training of family medicine doctors and improvement of the quality of

medical care at the primary level]. Klynycheskaia ynformatyka y Telemedytsyna (Clinical Informatics and Telemedicine), 8 (9), 141–145. [In Ukrainian].
8. Martseniuk V. P. (2015). Informatsiina systema upravlinnia yakistiu pidhotovky fakhivtsiv u vyshchii medychnii osviti [Information system of quality management of specialists training in higher medical education]. Ternopil: TDMU. [In Ukrainian].
9. Martseniuk V. P., Stakhanska O. O. (2010) Analiz rezultativ semestrovykh kompleksnykh testovykh ispytiv v medychnii osviti na osnovi koreliatsiinykh pokaznykiv uspishnosti ta bahatoparametrychnoi neiromerezhevoi klasteryzatsii [Analysis of the results of semester complex test exams in medical education on the basis of the correlation rates of success and multi-parameter neural network clusterization]. Medichna informatika ta inzheneriya (Medical Informatics & Engineering), 1, 53-57. [In Ukrainian].
10. Martseniuk V. P., Semenets A. V., Stakhanska O. O. (2010). Neiromerezheve prohnozuvannia skladannia studentamy-medykamy litsenziinoho intehrovanoho ispytu «Krok 1» na osnovi rezultativ potochnoi uspishnosti ta semestrovoho kompleksnoho testovoho ispytu [Neural network forecasting of the doctoral students of the licensed integrated exam «Step 1» on the basis of the results of the current progress and the semester complex test test]. Medichna informatika ta inzheneriya (Medical Informatics & Engineering), 2, 57-62. [In Ukrainian].

Додаток 1

Вихідні дані на консолі для дерева рішень рис. 1

```
Read 242 cases (58 attributes) from undefined.data
Decision tree:
histology1c > 70: yes (174/1)
histology1c <= 70:
:...histology1c <= 44: no (4)
 histology1c > 44:
 :...microbiology1c > 58: yes (18)
  microbiology1c <= 58:
  :...microbiology1c <= 48:
   :...histology2c > 54: yes (22)
   :  histology2c <= 54:
   :  :...anatomy4t <= 54: yes (3)
   :     anatomy4t > 54: no (3)
   microbiology1c > 48:
   :...histology1t > 66: yes (4)
    histology1t <= 66:
    :...chemistry2t <= 62: no (10)
     chemistry2t > 62: yes (4/1)
```

```
Evaluation on training data (242 cases):

 Decision Tree
 ----------------
 Size   Errors

  9  2( 0.8%)  <<

 (a)  (b)  <-classified as
 ---- ----
  17   2  (a): class no
     223  (b): class yes

Attribute usage:

100.00%          histology1c
 26.45%microbiology1c
 11.57%histology2c
 7.44%  histology1t
 5.79%  chemistry2t
 2.48%  anatomy4t
```

Додаток 2

Варіація головних компонент PC1 та PC2

Варіації	PC1	PC2
Standard deviation	5.3523	1.58444
Proportion of Variance	0.4855	0.04255
Cumulative Proportion	0.4855	0.52810

Додаток 3

Абсолютні значення елементів PC1 у порядку спадання

chemistry5c	chemistry4c	pathology2c	chemistry3c	histology3c	physiology4c
0.16593790	0.16273680	0.16219742	0.16104960	0.15769693	0.15762936
chemistry1c	histology1c	physiology1c	physiology3c	anatomy5c	microbiology1c
0.15696489	0.15645012	0.15588798	0.15560038	0.15529825	0.15476182
pharmacology2c	microbiology2c	anatomy4c	pharmacology1c	pathology1c	anatomy2c
0.15420393	0.15418250	0.15188477	0.14966438	0.14901629	0.14796841
anatomy3c	chemistry2c	histology2c	anatomy1c	physiology2c	microbiology3c
0.14786277	0.14783034	0.14737059	0.14705527	0.14661273	0.14471587
pharmacology3c	Step1	biology3c	biology1c	chemistry4t	biology2c
0.14164811	0.13751593	0.13221943	0.13135423	0.12583112	0.12087724
pharmacology2t	anatomy5t	physiology3t	biiology2t	chemistry2t	pathology2t
0.12063783	0.11947922	0.11893328	0.11750413	0.11742117	0.11741452
anatomy1t	histology2t	biiology1t	microbiology1t	pharmacology1t	chemistry3t
0.11540972	0.11424823	0.11382011	0.11326541	0.11291314	0.11201232
anatomy3t	physiology4t	pharmacology3t	biiology3t	physiology2t	chemistry1t
0.11161992	0.11059491	0.11048702	0.10947247	0.10814984	0.10695451
histology1t	microbiology3t	physiology1t	chemistry5t	histology3t	anatomy2t
0.10591875	0.10163421	0.10148778	0.10148727	0.10077544	0.09875603
pathology1t	microbiology2t	anatomy4t			
0.09788661	0.09605628	0.08627627			

Додаток 4

Вихідні дані на консолі для дерева рішень рис. 2

```
Read 242 cases (11 attributes)

Decision tree:

histology1c > 70: yes (174/1)
histology1c <= 70:
:...histology1c <= 44: no (4)
  histology1c > 44:
  :...pathology2c > 51: yes (36/3)
    pathology2c <= 51:
    :...histology1c > 67: no (3)
      histology1c <= 67:
      :...pathology2c <= 46: yes (13/2)
        pathology2c > 46:
        :...chemistry5c <= 51: no (6/1)
          chemistry5c > 51: yes (6/1)
```

```
Evaluation on training data (242 cases):

 Decision Tree
 ----------------
Size   Errors

  7  8( 3.3%)  <<

 (a)  (b)  <-classified as
---- ----
 12   7  (a): class no
  1  222  (b): class yes

Attribute usage:

100.00%          histology1c
 26.45%pathology2c
 4.96%  chemistry5c
```