МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ   
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

КАФЕДРА БІОМЕДИЧНОЇ КІБЕРНЕТИКИ

**КУРСОВА РОБОТА**

**з дисципліни «Біомедична кібернетика-2. Моделювання складних систем»**

**на тему** «Визначення групи ризику студента за допомогою моделювання систем»

**за спеціальністю** 122 «Комп'ютерні науки»

Робота захищена з оцінкою Керівник курсової роботи

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ доц. Носовець О.К.⠀ ⠀⠀⠀⠀ ⠀⠀⠀⠀ ⠀ ⠀ ⠀ ⠀ (кількість балів) (ПІБ)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

“\_\_\_\_” \_\_\_\_\_\_\_\_\_\_ 201\_ р.

Виконав студент групи БС-71мп

Войник Б. О. ⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀

(прізвище та ініціали)

“\_\_\_\_” \_\_\_\_\_\_\_\_\_\_ 201\_ р.

Київ 2018

ЗМІСТ

[ВСТУП 3](#_Toc515231503)

[РОЗДІЛ 1. ВХІДНІ ДАНІ 4](#_Toc515231504)

[РОЗДІЛ 2. ТЕОРЕТИЧНІ ВІДОМОСТІ 6](#_Toc515231505)

[РОЗДІЛ 3. ЛОГІСТИЧНА РЕГРЕСІЯ 11](#_Toc515231506)

[ВИСНОВОК ДО РОЗДІЛУ 3 16](#_Toc515231507)

[РОЗДІЛ 4. ДИСКРИМІНАНТНИЙ АНАЛІЗ 17](#_Toc515231508)

[ВИСНОВКИ ДО РОЗДІЛУ 4 20](#_Toc515231509)

[РОЗДІЛ 5. КЛАСИФІКАЦІЯ ЗА ДОПОМОГОЮ ПАКЕТУ GMDA SHELL 21](#_Toc515231510)

[ВИСНОВКИ ДО РОЗДІЛУ 5 31](#_Toc515231511)

[ВИСНОВКИ 32](#_Toc515231512)

[СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ 33](#_Toc515231513)

# ВСТУП

Існуючі методи визначення реакцій на тестове навантаження в наш час більше не передбачають аналіз та проведення неприємних медичних процедур. Важливим інструментом, що забезпечує здоров’язберігаючу спрямованість занять людини фізичними вправами, є самоконтроль. Самоконтроль людини – це регулярне використання ним низки простих прийомів для самостійного спостереження за зміною стану свого здоров'я і фізичного розвитку під впливом занять фізичними вправами [1]. З об'єктивних ознак при самоконтролі реєструються частота пульсу, вага, потовиділення, дані спірометрії, дінамометрії, визначається частота дихання. Для оцінки реакцій простіше всього скористатись функціональними пробами. Найзручнішою з них вважається проба Мартіне [2].

Актуальність.При проходженні тестів пов’язаних з реакцією на фізичне навантаження студент хоче дізнатись свої результати та проблеми зі здоров’ям. Для цього необхідно побудувати коректну класифікаційну модель, що дозволить оцінити визначені групи в базі даних на відмінність. Це дасть змогу класифікувати студентів та відносити їх до однієї з визначених груп здоров’я.

Об’єкт дослідження. База даних студентів 1-2 курсу Національного Університету України “Київського політехнічного університету ім. Ігоря Сікорського”.

Предмет дослідження. Класифікаційна модель, побудована на базі логістичної регресії, дискримінантного аналізу та пакету GMDA Shell.

Мета роботи. Побудова та порівняння моделей на кластеризованій базі показників артеріального тиску та пульсу, отриманих в результаті проби Мартіне, з подальшим визначенням найбільш простої та коректної моделі класифікації. У відповідності з метою ставлять такі завдання: проведення логістичної регресії та дискримінантного аналізу, проведення аналізу в програмі GMDH Shell DS, застосування методів поліпшення якості класифікаторів

# РОЗДІЛ 1. ВХІДНІ ДАНІ

Для дослідження було обрано базу даних спостережень студентів, що пройшли пробу Мартіне більше одного разу. Дослідження проводилися лише для чоловічої частини бази, оскільки характеристики кластерів чоловіків відрізняються від характеристик кластерів жінок. Тому важливим є проведення досліджень окремо, щоб виявити біль точні закономірності в кластерах чоловіків та порівняти результати з результатами, отриманими після проведення досліджень над жіночою частиною бази [3].

Згідно з поставленими задачами необхідно побудувати модель, яка б найкращим способом класифікувала студентів за тестом Мартіне, виявляючи кластер, що характеризує групу здоров’я, до якого відноситься студент

Обрана база даних містить 599 досліджень, серед яких 276 досліджень серед жінок та 323 – серед чоловіків. В базі присутні 360 змінних, з яких було вибрано лише значення артеріального тиску та пульсу для визначення реакції серцево-судинної системи студента на навантаження. Частину бази даних з обраними змінними зображено на рисунку 1.

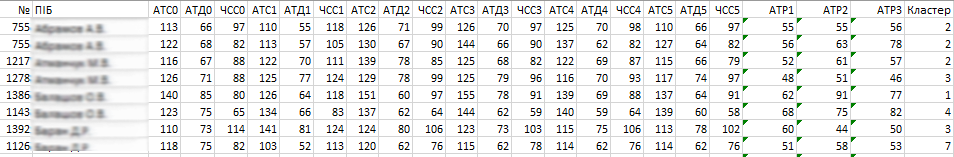


Рисунок 1 – приклад бази даних

Слід зазначити, що до бази даних було додано змінні «АТР1, АТР2, АТР3», що характеризують значення різниці між систолічним та діастолічним тиском на першій, другій та третій хвилинах відповідно.

Також для аналізу було вибрано класифікаційну змінну «Кластер», що розділяє чоловіків на 7 груп. Характеристики кожної з них наведено нижче (рисунок 2):



Рисунок 2 – Характеристики кластерів

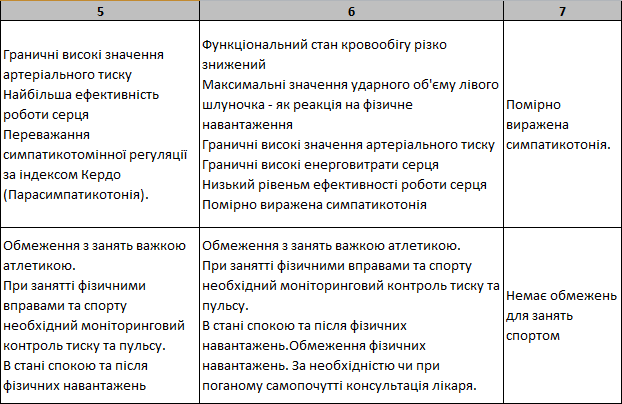


Рисунок 3 – Характеристики кластерів

Зауважимо, що деякі значення в базі пропущені і не будуть братися до уваги при аналізі.

# РОЗДІЛ 2. ТЕОРЕТИЧНІ ВІДОМОСТІ

**Логістична регресія**

Регресійний аналіз один із розділів математичної статистики та являє собою метод моделювання вимірюваних даних і дослідження їх властивостей. Дані складаються з пар значень залежної змінної (змінної відгуку) і незалежної змінної (пояснюватиме змінної). Регресійна модель являє собою функцію незалежної змінної і параметрів із додаванням випадкової змінної. Параметри моделі налаштовуються таким чином, щоб модель найкращим чином наближала дані. Критерієм якості наближення (цільовою функцією) зазвичай є середньоквадратична помилка: сума квадратів різниці значень моделі і залежною змінною для всіх значень незалежної змінної в якості аргументу. Залежна змінна є сумою значень деякої моделі і випадкової величини. Щодо характеру розподілу цієї величини робляться припущення, які називаються гіпотезами породження даних. Для підтвердження або спростування цієї гіпотези виконуються статистичні тести, так званий аналіз залишків. При цьому передбачається, що незалежна змінна не містить помилок. Регресійний аналіз використовується для прогнозу, аналізу часових рядів, тестування гіпотез і виявлення прихованих взаємозв'язків між даними. Бінарна логістична регресія є однією з різновидів регресійного аналізу.

За допомогою методу бінарної логістичної регресії можна дослідити залежність дихотомічних змінних від незалежних змінних, які мають будь-який вид шкали.

Як правило, у випадку з дихотомічними змінними йдеться про деяку подію, яка може відбутися або не відбутися; бінарна логістична регресія в такому випадку розраховує ймовірність настання події в залежності від значень незалежних змінних.

Ймовірність настання події для деякого випадку розраховується за формулою (1) та формулою (2):

|  |  |
| --- | --- |
| , | (1) |

де

|  |  |
| --- | --- |
| *z=b1\*X1 + b2\*Х2 + ...+ bn\*Xn + a*, | (2) |

де *X1* - значення незалежних змінних;

*b1* - коефіцієнти, розрахунок яких є завданням бінарної логістичної регресії;

*а* - деяка константа.

Якщо для *р* вийде значення менше 0,5, то можна припустити, що подія не настане; в іншому випадку передбачається настання події. Розрахована ймовірність *р* завжди вказує на виконання пророкування, яке відповідає більшій з двох кодувань залежних змінних.

Через залучення до аналізу великої кількості змінних комп'ютер повинен вирішити, які з них в кінцевому випадку будуть відібрані для використання в рівнянні ймовірності. Тому потрібно вибирати не метод вкладення, який включає в розрахунок всі змінні, а один з покрокових методів.

Метод прямої селекції починається з використання одних лише констант на стартовому етапі, а потім послідовно підключаються змінні, які демонструють сильну кореляцію з залежними змінними. Далі знову слідує перевірка того, які змінні повинні бути виключені, причому як критерій перевірки вибирається або статистика Вальдовского (Wald), або функція правдоподібності, або один з варіантів, які називаються умовною статистикою. Метод зворотної селекції спочатку бере до уваги всі змінні, а потім у зворотному порядку відбувається виключення малозначущих змінних.

Кількість утворюваних "фіктивних" дихотомічних змінних має бути завжди на 1 менше, ніж число кількість заданих категорій. Категорія, яка виявилася зайвою, називається еталонної категорією і, відповідно до попередніх установками, є останньою категорією. За допомогою поля контрастів (Contrast) можна управляти особливостями залучення в аналіз освічених фіктивних змінних; при контрасті рівному Deviation (Відхилення) всі категорії, крім еталонної, будуть перевіряються щодо сумарного ефекту [4].

**Дискримінантний аналіз**

Дискримінаційний чи дискримінантний аналіз використовується в тому випадку, якщо є дані, класифіковані на кілька груп, і необхідно знайти одну або більше функцій кількісних вимірів, які допоможуть віднести спостереження до однієї з цих груп. В дискримінантному аналізі розрізняють дві мети:

1. інтерпретація;
2. класифікація.

Метою інтерпретації є визначення кількості, значущості дискримінантних функцій і їх значень для пояснення відмінностей між класами. Метою класифікації є визначення класу, до якого належить новий об'єкт – це і є мета даної роботи.

В дискримінаційному аналізі, на відміну від кластерного, є навчальна вибірка, в якій відомо до яких класів відносяться об'єкти. За навчальною вибіркою необхідно отримати правила, які в подальшому дозволять визначити, до якого класу відносяться нові об'єкти.

Найбільш загальне застосування дискримінантного аналізу є включення у дослідження багатьох змінних з метою визначення тих з них, які найкращим чином поділяють сукупності між собою. Наприклад, дослідник в галузі освіти, який цікавиться прогнозом вибору, який зроблять випускники середньої школи щодо своєї подальшої освіти, зробить з метою одержання найбільш точних прогнозів реєстрацію більшої кількості параметрів учнів.

Дискримінантний аналіз використовують, якщо дослідник хоче побудувати модель, яка дозволить краще всього передбачити, до якої сукупності належатиме той чи інший показник. У наступному міркуванні термін модель буде використовуватися для того, щоб позначати змінні, використовувані в пророкуванні приналежності до сукупності, а про невикористовуванні змінні говорять, що вони знаходяться поза межами моделі.

У ролі дискримінантного аналізу найчастіше береться лінійна функція записана у вигляді формули (3):

|  |  |
| --- | --- |
| , | (3) |

де *Х1,Х2,…,Хm*– значення ознак у даного об'єкта;

*С1,С2,…,Сm* – дискримінанті множники.

За допомогою дискримінантних множників виконуємо перехід від m-мірного простору первинних показників до одновимірного простору.

Лінійну функцію можна розглядати як проекцію даного об'єкта на деяку (одновимірну) дискримінантну вісь.

У процедурі дискримінантного аналізу дискримінантні множники визначаються таким чином, щоб забезпечити найбільшу відмінність між проекціями першої та другої вибірок на дискримінантну вісь.

Дискримінантний аналіз потрібно проводити з використанням мінімальної кількості функцій. Їхня кількість залежить від конфігурації класів в багатовимірному просторі дискримінантних змінних. Щоб визначити, скільки функцій необхідно, використовують перевірку функцій на значимість. Для оцінки значущості використовують або А-статистику Уілкса або ксі – квадрат [5].

Критерій значення Уілкса обчислюють за формулою (4):

|  |  |
| --- | --- |
| , | (4) |

де *К –* кількість значень;

*k –* число вже обчислених дискримінаційних функцій.

Чим ближче значення критерію *К*, тим краща відмінності класів, а чим ближче до 1, тим відмінність гірша.

Значення ксі-квадрат розраховують за формулою (5):

|  |  |
| --- | --- |
| , | (5) |

де *р –* кількість членів у дискримінатної функції, виключаючи вільний член функції.

Якщо це значення більше критичного із заданим рівнем значущості і числом ступенів свободи *(р-к) (К-k-1)*, то значимість підтверджується.

Канонічна дискримінантна функція для загального випадку k класів записана у формулі (6):

|  |  |
| --- | --- |
| , | (6) |

де *fki* — значення канонічної дискримінантної функції для 1-го об'єкта в *k*-му класі;

uj — шукані коефіцієнти дискримінантної функції;

*Хjki* — значення дискримінантної змінної *Хj* для i-го об'єкта в класі *k*.

Функцію будують таким чином, щоб її середні значення для різних класів якомога більше розрізнялися. При цьому сукупність функцій повинна утворювати ортогональний простір, тобто функції - незалежні один від одного. З цього випливає, що кількість функцій нe може бути більше кількості класів мінус 1 або числа дискримінантних змінних (в залежності від того, яка з цих величин менше).

Розраховані значення канонічної дискримінантної функції fki, розглядають як точки в деякому просторі. Для кожної групи можна розрахувати центр групування. Тому в цій новій системі координат для нового об'єкта розраховують відстань від нього до кожної точки групування. Зазвичай для цього використовують квадрат відстані Махаланобіса.

# РОЗДІЛ 3. ЛОГІСТИЧНА РЕГРЕСІЯ

Оскільки класифікатор студентів є не бінарною змінною, тому було прийнято рішення розбити пацієнтів на групи методом «один проти всіх» та вирівняти дані в групах, де були отримані показники класифікації занадто асиметричні. Результати дослідження представлені для третього кластеру, але аналогічна процедура проводилася для кожного кластеру окремо. Результуючу таблицю для всіх кластерів зображено у таблиці 1.

На рисунку 4 зображено встановлення параметрів, які водилися для отримання класифікації даних.

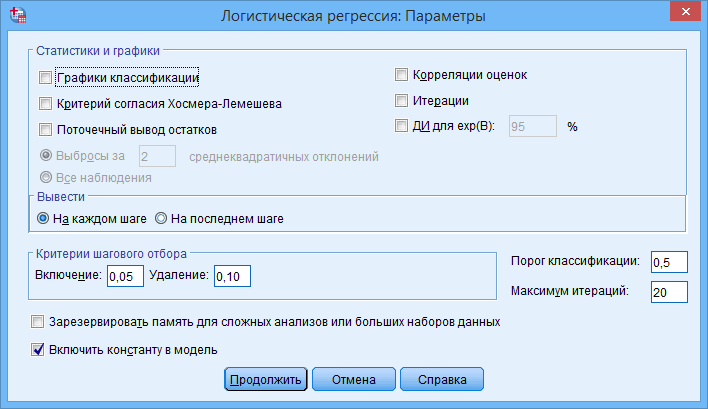


Рисунок 4 – Зображення параметрів

**Розбиття на класи «один проти всіх»**

На рисунку 5 наведені результати класифікації при об’єднанні у групи: перша група – студенти, що знаходяться в 3 кластері, друга – студенти в інших кластерах (3 проти 1, 2, 4, 5, 6, 7)

Для побудови логістичної регресії для великої кількості предикторів використано метод включення (условне).

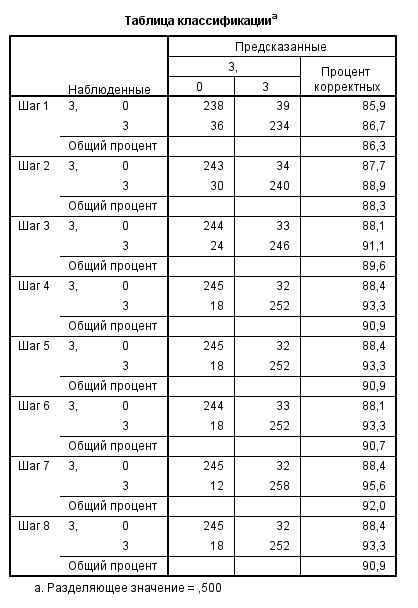


Рисунок 5 – Результати класифікації ЛР

За результатами дослідження ми бачимо, що загальний відсоток коректно спрогнозованих даних складає 90,9%. При цьому з таблиці можна зробити висновок про те, що із загального числа студентів, які знаходяться в 3 кластері, рівного 270, тестом вірно були визнані 252. Інші 18 є хибно негативними. Таким чином, відсоток коректності склав 88,4%. Із загальної кількості спостережень, що відносяться до інших кластерів, рівного 277, тестом були визнані такими 245. В цьому випадку відсоток класифікації склав 88,4%.

Для побудови рівняння регресії було використано наступну таблицю:

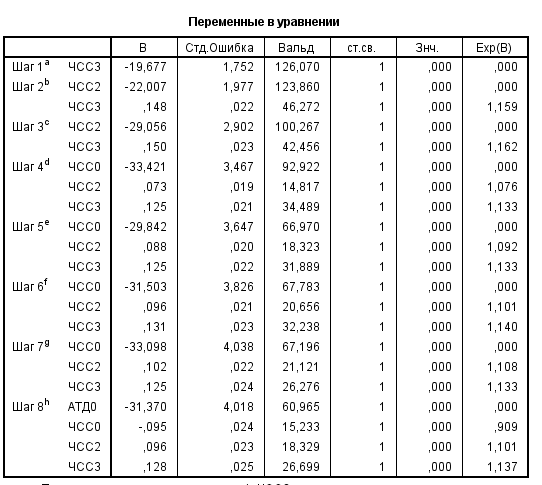


Рисунок 6 – Змінні для рівняння регресії

Таким чином, рівняння регресії набуває вигляду:

Для поліпшення якості класифікаторів було вирішено розширити матрицю змінних за допомогою нелінійних перетворень. Тому необхідним кроком було встановлення в параметрах аналізу покрокового режиму для відбору змінних

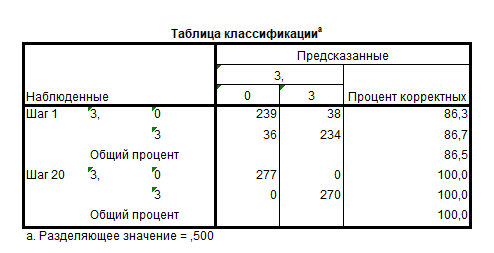


Рисунок 7 – Результати класифікації ЛР з нелінійними перетвореннями

За результатами дослідження ми бачимо, що відсоток коректно спрогнозованих даних складає 100,0%. Але слід зазначити, що це при умові неповної моделі, оскільки при обробці даних SPSS видав наступне попередження:

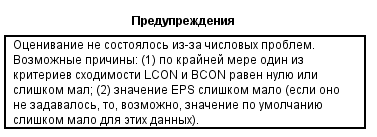


Рисунок 8 – Попередження

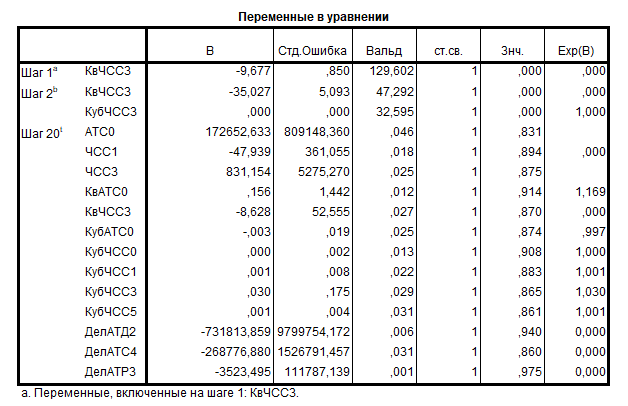


Рисунок 9 – Змінні для рівняння регресії з нелінійними перетвореннями

Таким чином, рівняння регресії має вигляд:

Порівнюючи дві моделі ми можемо дійти висновку, що модель, яка побудована на базі даних з додатковими змінними нелінійних перетворень є більш складною, хоча й не повною, але водночас дає вищий результат в порівнянні з моделлю, що включають лише істинні змінні.

Таблиця 1 – Порівняльна характеристика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Класифікатор** | **Позначення** | **ЛР на істинних даних** | **ЛР з нелінійними перетвореннями** | **Результат** | |
| 1-Всіх | Заг. Кор., % | 90,4 | 100 | 2 |  |
| Ск-ть | 3/21 | 13/84 |  |  |
| Ск,% | 14,29 | 15,48 |  | 1 |
| 2-Всіх | Заг. Кор., % | 77,4 | 91 | 2 |  |
| Ск-ть | 4/21 | 7/84 |  |  |
| Ск,% | 19,05 | 8,33 |  | 2 |
| 3-Всіх | Заг. Кор., % | 90,9 | 100\* | 2 |  |
| Ск-ть | 4/21 | 13/84 |  |  |
| Ск,% | 19,05 | 15,48 |  | 2 |
| 4-Всіх | Заг. Кор., % | 97,5 | 95,8 | 1 |  |
| Ск-ть | 5/21 | 13/84 |  |  |
| Ск,% | 23,81 | 15,48 |  | 2 |
| 5-Всіх | Заг. Кор., % | 99,7\* | 100 | 1 |  |
| Ск-ть | 3/21 | 6/84 |  |  |
| Ск,% | 14,29 | 7,14 |  | 2 |
| 6-Всіх | Заг. Кор., % | 100\* | 100 | 1 |  |
| Ск-ть | 4/21 | 5/84 |  |  |
| Ск,% | 19,05 | 5,95 |  | 2 |
| 7-Всіх | Заг. Кор., % | 88 | 100 | 2 |  |
| Ск-ть | 4/21 | 15/84 |  |  |
| Ск,% | 19,05 | 17,86 |  | 2 |

Розшифрування міток в таблиці 1:

1. Заг.Кор,% - загальний відсоток коректно спрогнозованих даних
2. Ск-ть – складність моделі
3. Ск,% - відсоток складності моделі (чим нижче, тим краще)
4. 21 – кількість істинних змінних
5. 84 – кількість істинних змінних разом з нелійнійними перетвореннями
6. 1- виграш ЛР на істинних даних
7. 2 – виграш ЛР з нелінійними перетвореннями
8. \* - неповність моделі

З таблиці 1 видно, що аналіз, проведений на базі нелінійних перетворень дає складніші рівняння моделі, але їх складність, беручи до уваги всі змінні, в переважній кількості менша. Слід також зауважити, що у 4 тестах із 7 відсоток коректно спрогнозованих даних збільшився

# ВИСНОВОК ДО РОЗДІЛУ 3

Беручи до уваги таблицю 1 ми можемо зробити висновок, що проведення логістичної регресії, включаючи нелінійні перетворення, а саме: операція взяття квадрату та кубу, і операція взяття оберненої змінної, будує складніші рівняння регресії, але самі моделі є простішими за складністю і дають більш високий відсоток коректності.

# РОЗДІЛ 4. ДИСКРИМІНАНТНИЙ АНАЛІЗ

Аналогічно до розділу 3 результати дослідження представлені для третього кластеру, але дана процедура проводилася для кожного кластеру окремо.

Всі дані було класифіковано методом дискримінантного аналізу, використовуючи покроковий відбір. На рис. 3 показаний вибір відстані та критерій розпізнавання.

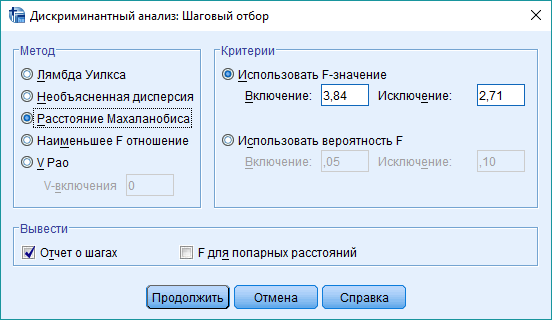


Рисунок 10 – Вибір методів

У вікні вибору класифікації, що зображено на рис 4. показаний вибір розрахунку класифікації.

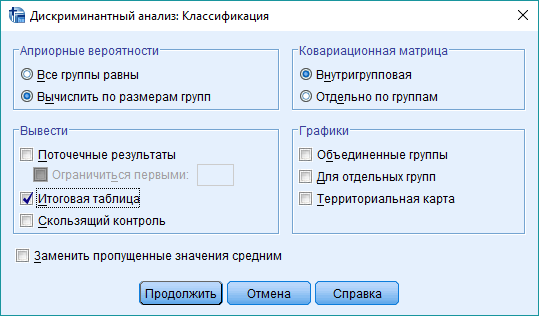


Рисунок 11 – Вибір класифікації

**Результати аналізу для груп «3 проти 1,2,4,5,6,7»**

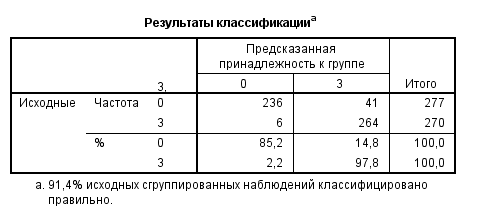


Рисунок 12 – Результати класифікації

За результатами дискримінантного аналізу ми бачимо, що 91,4% вихідних згрупованих спостережень класифіковано правильно.

Для побудови дискримінантної функції застосовуються дані з наступної таблиці:



Рисунок 13 – Коефіцієнти для дискримінантної функції

Таким чином, дискримінантна функція набуває вигляду:

Для підвищення якості моделі, як і для логістичної регресії до аналізу штучно були введені нелінійні змінні , та для кожної змінної.

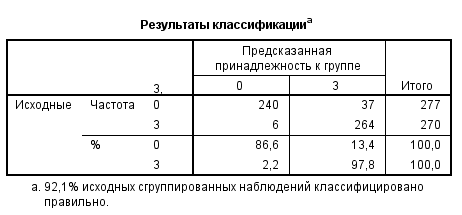


Рисунок 14 – Результати класифікації з нелінійними перетвореннями

З рисунку 14 ми бачимо, що відсоток правильно класифікованих спостережень складає 92,1%



Рисунок 15 – Коефіцієнти для дискримінантної функції з нелінійними перетвореннями [6,7].

Таким чином, дискримінантна функція має вигляд:

Результати дослідження показують, що модель, побудована з додатковими змінними дає біль високий результат класифікації даних. Слід зазначити, що приріст є незначним і складає 0,7%, а модель при цьому стала складнішою в 2,25 рази

Аналогічно до розділу 2 було побудовано таблицю порівняльної характеристики для дискримінантного аналізу щоб оцінити результати дослідження.

Таблиця 2 – Порівняльна характеристика

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Класифікатор** | **Позначення** | **ДА на істинних даних** | **ДА з нелінійними перетвореннями** | **Результат** | |
| 1-Всіх | Заг. Кор., % | 87,2 | 94 | 2 |  |
| Ск-ть | 3/21 | 9/84 |  |  |
| Ск,% | 14,29 | 10,71 |  | 2 |
| 2-Всіх | Заг. Кор., % | 77,2 | 84,6 | 2 |  |
| Ск-ть | 4/21 | 4/84 |  |  |
| Ск,% | 19,05 | 4,76 |  | 2 |
| 3-Всіх | Заг. Кор., % | 91,4 | 92,1 | 2 |  |
| Ск-ть | 4/21 | 9/84 |  |  |
| Ск,% | 19,05 | 10,71 |  | 2 |
| 4-Всіх | Заг. Кор., % | 85,4 | 88,9 | 2 |  |
| Ск-ть | 4/21 | 7/84 |  |  |
| Ск,% | 19,05 | 8,33 |  | 2 |
| 5-Всіх | Заг. Кор., % | 94,3 | 97,6 | 2 |  |
| Ск-ть | 5/21 | 12/84 |  |  |
| Ск,% | 23,81 | 14,29 |  | 2 |
| 6-Всіх | Заг. Кор., % | 93,9 | 97,1 | 2 |  |
| Ск-ть | 4/21 | 12/84 |  |  |
| Ск,% | 19,05 | 14,29 |  | 2 |
| 7-Всіх | Заг. Кор., % | 88 | 88,5 | 2 |  |
| Ск-ть | 3/21 | 3/84 |  |  |
| Ск,% | 14,29 | 3,57 |  | 2 |

З таблиці 2 видно, що аналіз, проведений на базі нелінійних перетворень дає складніші рівняння моделі, але їх складність, беручи до уваги всі змінні, менша. Слід також зауважити, що у 7 тестах із 7 відсоток коректно спрогнозованих даних збільшився

# ВИСНОВКИ ДО РОЗДІЛУ 4

За допомогою дискримінантного аналізу було побудовано функції прогнозування для кожного кластера, при чому результати дослідження показують, що введення нелінійних змінних (для кожної вхідної змінної) до аналізу покращую результат класифікації

# РОЗДІЛ 5. КЛАСИФІКАЦІЯ ЗА ДОПОМОГОЮ ПАКЕТУ GMDA SHELL

У сучасній теорії прогнозованого моделювання добре відомо, що модель повинна забезпечувати компроміс між простотою та точністю. GMDH Shell відповідає цій меті за допомогою алгоритмів навчання методу групового урахування аргументів (МГУА). Ідея всіх алгоритмів типу GMDH полягає в тому, щоб застосувати генератор поступово ускладнюючих моделей і вибрати набір моделей, що показують високу точність прогнозування на раніше невидимі дані. Ці дані зазвичай називають частиною валідації чи тестування, а модель найвищої категорії, як стверджується, є оптимально складною.

Від класичної множинної регресії МГУА відрізняється лише використанням специфічних квадратичних критеріїв зовнішнього або внутрішнього типу, а також багаторядними ітераційними процедурами знаходження оптимального рішення задачі [8].

Слід заначити, що розрахунки були проведені за принципом «один проти всіх» для кожної групи окремо. Параметри в вікні МГУА були встановлені за замовчування – рисунок 16.

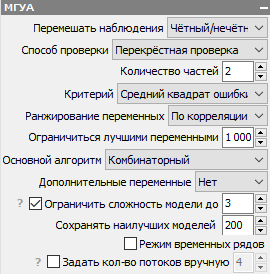


Рисунок 16 – Параметри МГУА

**Результати для групи «1 проти 2,3,4,5,6,7»**

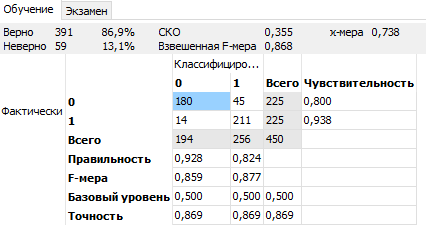


Рисунок 17 – Таблиця точності класифікації МГУА при навчанні

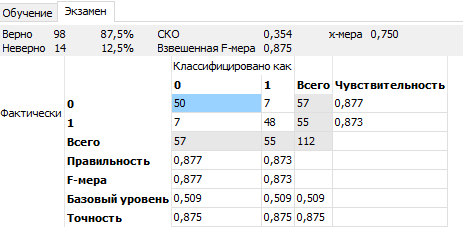


Рисунок 18 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 8/21 (38%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 3 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 80% | 93,8% | 87,7% | 87,3% |

З таблиці 3 видно, що на екзамені рівень чутливості підвищився на 7,7%, але при цьому специфічність знизилась на 6,5%., що є негативною змінною.

**Результати для групи «2 проти 1,3,4,5,6,7»**

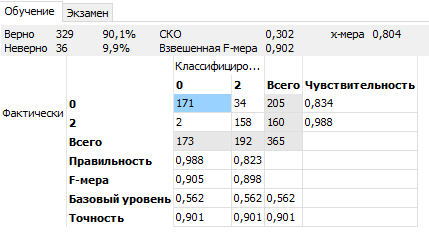


Рисунок 19 – Таблиця точності класифікації МГУА при навчанні

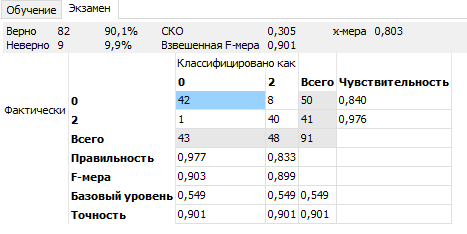


Рисунок 20 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 25/231 (10,8%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 4 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 83,4% | 98,8% | 84% | 97,6% |

З таблиці 4 видно, що на екзамені рівень чутливості підвищився на 0,6%, але при цьому специфічність знизилась на 1,2%.

**Результати для групи «3 проти 1,2,4,5,6,7»**

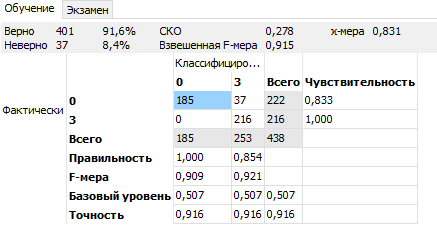


Рисунок 21 – Таблиця точності класифікації МГУА при навчанні

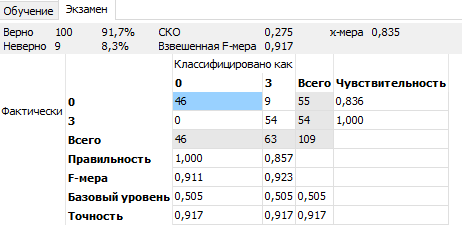


Рисунок 22 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 4/22 (18,18%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 5 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 83,3% | 100% | 83,6% | 100% |

З таблиці 1 видно, що на екзамені рівень чутливості підвищився на 0,3%, рівень специфічності не змінився.

**Результати для групи «4 проти 1,2,3,5,6,7»**

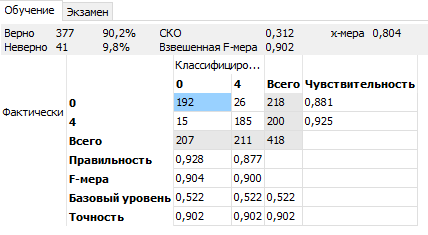


Рисунок 23 – Таблиця точності класифікації МГУА при навчанні

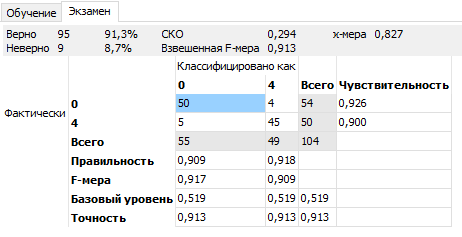


Рисунок 24 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 20/231 (8,65%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 6 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 88,1% | 92,5% | 92,6% | 90% |

З таблиці 6 видно, що на екзамені рівень чутливості підвищився на 4,5%, рівень специфічності знизився на 2,5%.

**Результати для групи «5 проти 1,2,3,4,6,7»**



Рисунок 25 – Таблиця точності класифікації МГУА при навчанні

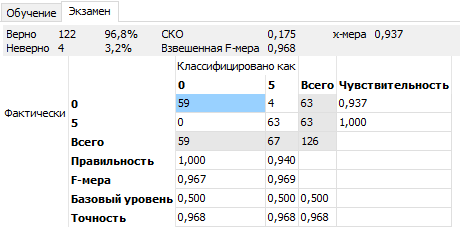


Рисунок 26 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 33/231 (14,28%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 7 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 97,6% | 100% | 93,7% | 100% |

З таблиці 7 видно, що на екзамені рівень чутливості знизився на 2,9%, що є негативною змінною, а рівень специфічності не змінився

**Результати для групи «6 проти 1,2,3,4,5,7»**

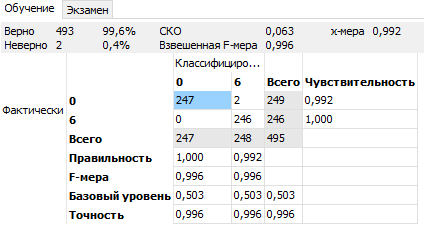


Рисунок 27 – Таблиця точності класифікації МГУА при навчанні

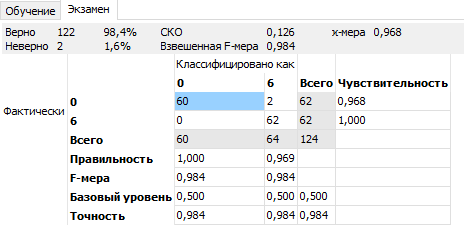


Рисунок 28 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 49/231 (21,21%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 8 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 99,2% | 100% | 96,8% | 100% |

З таблиці 8 видно, що на екзамені рівень чутливості знизився на 2,4%, що є негативною змінною, а рівень специфічності не змінився

**Результати для групи «7 проти 1,2,3,4,5,6»**

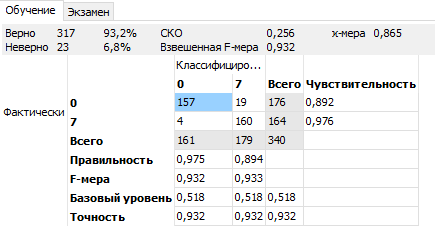


Рисунок 29 – Таблиця точності класифікації МГУА при навчанні

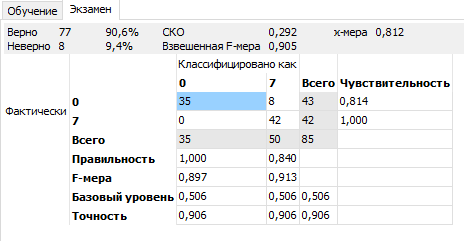


Рисунок 30 – Таблиця точності класифікації МГУА при екзамені

При дослідженнях виявлено, що складність моделі становить 23/231 (9,96%), а рівняння класифікаційної моделі набуває вигляду:

Таблиця 9 – Порівняльна характеристика

|  |  |  |  |
| --- | --- | --- | --- |
| **Навчання** | | **Екзамен** | |
| **Рівень чутливості** | **Специфічність** | **Рівень чутливості** | **Специфічність** |
| 89,2% | 97,6% | 81,4% | 100% |

З таблиці 9 видно, що на екзамені рівень чутливості знизився на 7,8%, що є негативною змінною, а рівень специфічності збільшився на 2,4%.

# ВИСНОВКИ ДО РОЗДІЛУ 5

Побудова моделі прогнозування за допомогою пакету GMDA Shell при параметрах, встановлених за замовчуванням, дає наступні результати: в 4 випадках із 7 рівень чутливості підвищився, в решті випадків він знизився в середньому на 4,4%; в 1 випадку із 7 рівень специфічності підвищився на 2,4%, в 3 випадках із 7 рівень специфічності знизився і в усіх інших випадках – не змінився

# ВИСНОВКИ

В даній курсовій роботі було розглянуто наступні статистичні методи для вирішення задачі класифікації студентів за тестовим навантаженням: логістична регресія та дискримінантний аналіз. За даними досліджень було побудовано необхідні класифікаційні рівняння. Також було застосовано програмний продукт GMDH Shell для побудови рівнянь прогнозування. Виходячи з результатів досліджень, ми можемо сказати, що при порівнянні алгоритмів, без врахування результатів із нелінійними змінними та виключаючи із аналізу GMDH специфічність, у 2 випадках із 7 (виключаючи неповну модель при логістичній регресії) найбільш точну і просту модель будує дискримінантний аналіз, ще в 2 із 7 – метод логістичної регресії, GMDH також в 2 випадках із 7 будує найбільш точну модель. Ще у одному випадку логістична регресія та дискримінантний аналіз дають ідентичний результат. При цьому дискримінантний аналіз є більш простим та універсальним у використанні, оскільки при його застосуванні ми завжди маємо справу тільки з однією статистичною процедурою, в якій беруть участь одна категоріальна залежна змінна і кілька незалежних змінних з будь-яким типом шкали. Також з досліджень видно, що логістична регресія не завжди може побудувати повну модель, а пакет GMDH не завжди будує прості рівняння класифікації. Тому ми ввели нелінійні змінні для методів логістичної регресії та дискримінантного аналізу

При введені нелінійних змінних ми бачимо, що логістична регресія дає високі результати, але при цьому зростає складність моделі

Отже, нами було побудовано класифікаційні моделі, застосовано методи поліпшення якості класифікаторів та виявлено найбільш оптимальний метод для побудови моделей високої точності. Таким чином, найбільш оптимальних результатів можна досягти за допомогою логістичної регресії з введенням нелінійних змінних , та .

# СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ильинич В.И. Физическая культура студента и жизнь: учебник /

В.И. Ильинич. – М.: Гардарики, 2010. – 366 с.

1. Ланда Б.Х. Методика комплексной оценки физического развития и физической подготовленности: учеб. пособие / Б.Х. Ланда. – М.: Советский спорт, 2011. – 348 с.
2. Оценка функционального состояния и адаптивных возможностей организма у студентов вуза в процессе занятий физической культурой: методические указания для преподавателей физической культуры и студентов / сост. С.Е. Бебинов. – Омск: Изд-во СибАДИ, 2004. – 16 c.
3. Райе Дж. Матричные вычисления и математическое обеспечение: Пер. с англ. О.Б. Арушаняна / Под ред. В.В. Воеводина. - М.: Мир, 1984. - 264 с.
4. Факторный, дискриминантный и кластерный анализ: Пер. с англ. — М.: Финансы и статистика, 1989. — 215 с.
5. Иллюстрированный самоучитель по SPSS [Електронний ресурс] – Режим доступу до ресурсу: http://www.datuapstrade.lv/rus/spss/
6. Дискриминантный анализ [Електронний ресурс] – Режим доступу до ресурсу: https://studfiles.net/preview/5249013/page:24/
7. GMDA Shell Documantation [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: http://d.gmdhshell.com/docs
8. Павлов А. А. Метод группового учёта аргументов и анализа иерархий (МГУАиАИ) в задачах принятия решений / А. А. Павлов, А. А. Иванова, Р. А. Зигура // Вісник Національного технічного університету України "Київський політехнічний інститут". Сер. : Інформатика, управління та обчислювальна техніка. - 2007. - Вип. 47. - С. 207-216. - Режим доступу: http://nbuv.gov.ua/UJRN/Vkpi\_iuot\_2007\_47\_23