Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ   
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

імені Ігоря Сікорського»

Кафедра біомедичної кібернетики

**РОЗРАХУНКОВА РОБОТА**

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

**на тему «Прогнозування вірогідності захворювання нирок»**

**за спеціальністю** 163 «біомедична інженерія»

#### Робота захищена з оцінкою Керівник курсової роботи

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

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

(підпис)

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

Виконала студентка групи БМ-73мп

Михайлова П. О. ⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀

(прізвище та ініціали)

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

Київ 2018

|  |  |
| --- | --- |
| ЗМІСТ | Стор. |
| ВСТУП | 3 |
|  |  |
| РОЗДІЛ 1. ВХІДНІ ДАНІ…………………………………..……..… | 5 |
| РОЗДІЛ 2. ЛОГІСТИЧНА РЕГРЕСІЯ……………………….......…. | 7 |
| Висновки до розділу……………………………………….. | 12 |
| РОЗДІЛ 3. ДИСКРИМІНАНТНИЙ АНАЛІЗ……..………….……. | 13 |
| Висновки до розділу……………………………………….. | 17 |
| РОЗДІЛ 4. КЛАСИФІКАЦІЯ ЗА ДОПОМОГОЮ ПАКЕТУ  GMDA SHELL……………………………………………………….. | 18 |
| Висновки до розділу……………………………………….. | 26 |
|  |  |
| ВИСНОВКИ ………………………………………..…………..…... | 27 |
| Список використаної літератури …………………………………... | 29 |

ВСТУП

Актуальність.  До хронічної хвороби нирок може призводити діабет, підвищення кров’яного тиску та інші захворювання. Уражені нирки не здатні підтримувати організм у здоровому стані, оскільки вони припиняють виконувати свої функції. Якщо хвороба нирок прогресує, в організмі значно підвищується вміст шкідливих речовин. Це зумовлює погіршення загального самопочуття, можуть виникати ускладнення, зокрема, підвищення кров’яного тиску, анемія (зменшення кількості кров’яних тілець), патологія кісток, втрата апетиту і схуднення та ураження нервів. [1]

Крім того, хвороба нирок підвищує ризик виникнення захворювань серця та кров’яних судин. Ці захворювання можуть розвиватися повільно впродовж тривалого часу. Тому актуальною є задача діагностування наявності даних захворювань на початкових стадіях. Якщо вчасно діагностувати хронічну хворобу нирок і розпочати лікування, можна запобігти її прогресуванню, що може призводити до ниркової недостатності. У випадку чого для збереження життя хворого вже необхідна пересадка нирки або діаліз.[1]

Об’єкт дослідження: база даних «Early stage of Indians Chronic Kidney Disease(CKD)».

Предмет дослідження: класифікаційна модель розподілення пацієнтів на хворих хронічним захворюванням нирок та здорових.

Мета роботи: побудувати найбільш просту і коректну модель класифікації, що забезпечить швидкість розрахунку вірогідності захворювання пацієнта на хронічну хворобу нирок за заданими параметрами з високою точністю.

У відповідності з метою ставляться такі завдання:

1) ввести додаткові нелінійні значення параметрів до моделі;

2) розрахувати логістичну регресію на вхідних даних та додатково з штучно введеними змінними;

3) за допомогою дискримінантного аналізу розрахувати класифікаційну модель на вхідних даних та додатково з штучно введеними змінними;

4) за допомогою програмного пакету GMDH Shell for Data Science розрахувати декілька класифікаційних моделей змінюючи параметри МГУА та обрати найбільш коректну;

5) порівняти отримані результати.

РОЗДІЛ 1. ВХІДНІ ДАНІ

В даній розрахунковій роботі згідно з поставленими задачами необхідно побудувати найбільш відповідну модель класифікації для визначення на ранніх стадіях наявності у пацієнтів захворювання нирок, чи відсутність даної патології.

Обрана база даних містить 400 досліджених пацієнтів та 25 параметрів, з яких один є класифікаційна змінна (class), що розділяє всіх пацієнтів на 2 класи [2]:

* ckd – ті, що мають захворювання нирок (250 досліджуваних);
* notckd – ті, що не мають захворювання нирок (150 досліджуваних).

Предикторами при класифікації виступають параметри, що вказані в таблиці 1.

Таблиця 1 – Перелік предикторів

|  |  |  |  |
| --- | --- | --- | --- |
| Назва змінної | Значення змінної | Тип даних | Значення |
| age | вік | кількісні | в роках |
| bp | кров’яний тиск | кількісні | mm/Hg |
| sg | питома вага | категоріальні | 1.005,1.010,1.015,1.020,1.025 |
| al | рівень альбуміну | категоріальні | 0,1,2,3,4,5 |
| su | рівень цукру | категоріальні | 0,1,2,3,4,5 |
| rbc | червоні кров'яні тільця | номінальні | normal, abnormal |
| pc | лейкоцити | номінальні | normal, abnormal |
| pcc | згустки лейкоцитів | номінальні | present, notpresent |
| ba | бактерії | номінальні | present, notpresent |
| bgr | рівень глюкози в крові | кількісні | mgs/dl |
| bu | сечовина крові | кількісні | mgs/dl |
| sc | сироватковий креатинін | кількісні | mgs/dl |
| sod | натрій | кількісні | mEq/L |
| pot | калій | кількісні | mEq/L |
| hemo | гемоглобін | кількісні | gms |
| pcv | гематокрит | кількісні |  |
| wc | кількість білих кров'яних тілець | кількісні | cells/cumm |
| rc | кількість червоних кров'яних тілець | кількісні | millions/cmm |
| htn | гіпертонія | номінальні | yes, no |
| dm | цукровий діабет | номінальні | yes, no |
| cad | захворювання коронарних артерій | номінальні | yes, no |
| appet | апетит | номінальні | good, poor |
| pe | pedal edema | номінальні | yes, no |
| ane | анемія | номінальні | yes, no |

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

РОЗДІЛ 2. ЛОГІСТИЧНА РЕГРЕСІЯ

Для задачі прогнозування було використано метод логістичної регресії – це різновид множинної регресії, загальне призначення якої полягає в аналізі зв'язку між декількома незалежними змінними (званими також регресорів або предикторами) і залежною змінною.

Залежною змінною була змінна class, що є бінарною, і розділяє всіх пацієнтів на тих, що мають захворювання нирок і тих, що не мають.

Задачею було розрахувати коефіцієнти логістичної регресії, для оцінки вірогідності захворювання нирок по заданим параметрам.

Це досягається застосуванням наступного регресійного рівняння:

|  |  |
| --- | --- |
|  | (1.1) |

де *р* – вірогідність того, що відбудеться подія, що цікавить; у - стандартне рівняння регресії: *y=a+b1x1+b2x2+…+bnxn.*

Предикторами виступають параметри, що вказані в таблиці 1.

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

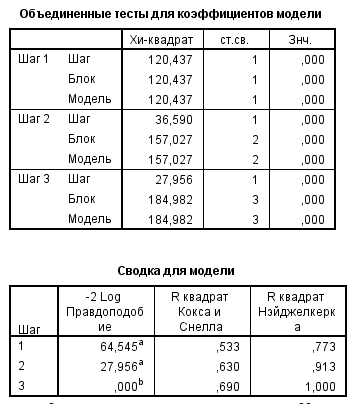


Рис. 1 – Результати побудови моделі

Якість наближення регресійній моделі оцінюється за допомогою функції подібності. Мірою правдоподібності служить негативне подвоєне значення логарифма цієї функції (-2LL). В якості початкового значення для -2LL застосовується значення, яке виходить для регресійній моделі, що містить тільки константи. На другому кроку це значення зменшилося майже до нуля. Подібне зниження величини означає істотне поліпшення моделі; різницю позначається як величина хі-квадрат і є дуже значущою.

Два інших виведених показника, названі іменами Кокса і Снелла і Нейджелкерка, є мірами визначеності. Вони також вказують на ту частину дисперсії, яку можна пояснити за допомогою логістичної регресії. Частина дисперсії, зрозумілої за допомогою логістичної регресії, в даному прикладі становить на першому кроці 77,3%, на другому – 91,3% і на третьому100 %.

Далі наводиться класифікаційна таблиця, в якій спостерігаються показники приналежності до групи (1 = хворий, 2 = здоровий) протиставляються передбаченим на основі розрахованої моделі.



Рис. 2 – Таблиця класифікації

З таблиці можна зробити висновок про те, що із загального числа хворих, що було проаналізовано, рівного 43, тестом були визнані такими тільки 40 (в медичній діагностиці в таких випадках говорять про "строго позитивних" результатах). Інші 3 визнано являються хибно негативними, Із загальної кількості здорових, рівного 115, тестом були визнані такими всі 115 ( "строго негативні").

У висновку виводяться результати про розрахованих коефіцієнтах і перевірці їх значимості (рис. 3).

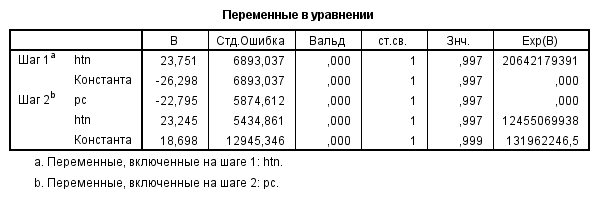


Рис. 3 – Розраховані коефіцієнти регресії

На другому кроці до рівняння регресії увійшли змінні pc та htn. І цього достатньо, щоб побудувати логістичну регресію з високою точністю.

Перевірка значущості відмінності коефіцієнтів від нуля, проводиться за допомогою статистики Вальда, що використовує розподіл хі-квадрат, який представляє собою квадрат відношення відповідного коефіцієнта до його стандартної помилки.

Вийшли значимі коефіцієнти:

*а = 18,698*

*b1 = -22,795*

*b2 = 23,245*

За допомогою цих двох значень коефіцієнтів можна для кожного значення Т-типізації розрахувати ймовірність р:

Для підвищення точності побудованої моделі, штучно в аналіз були введені додаткові змінні *x2,**x3* та *1/x* для кожного предиктора *х*.

Результати після введення додаткових змінних *x2* представлені на рисунках 4-6.

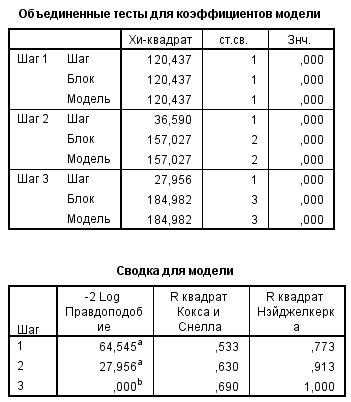


Рис. 4 – Результати побудови моделі (2)

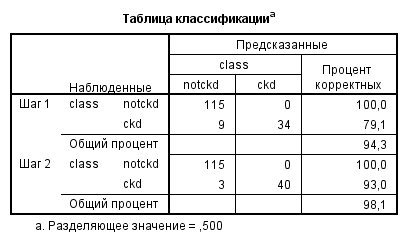


Рис. 5 – Таблиця класифікації (2)

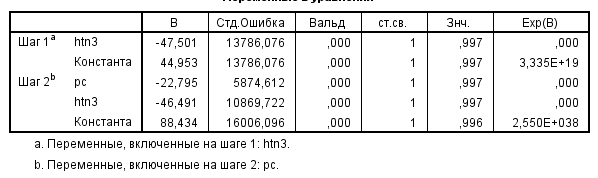


Рис. 6 – Розраховані коефіцієнти регресії (2)

На другому і останньому кроці до рівняння регресії увійшли змінні pc та htn3. І цього достатньо, щоб побудувати логістичну регресію з високою точністю.

Вийшли значимі коефіцієнти:

*а = 88,434*

*b1 = -22,795*

*b2 = -46,491*

За допомогою цих двох значень коефіцієнтів можна для кожного значення Т-типізації розрахувати ймовірність р:

Висновки до розділу

Було розраховано модель логістичної регресії, що з високою точністю класифікую на базі двох параметрів пацієнтів на тих, що мають захворювання нирок та тих, які не мають даного недугу.

Придатна модель класифікації була побудована на двох змінних: pc – кількість лейкоцитів та htn – наявність гіпертонії. Введення додаткових нелінійних змінних не вплинуло на остаточний результат і якість моделі прогнозування.

РОЗДІЛ 3. ДИСКРИМІНАНТНИЙ АНАЛІЗ

За допомогою дискримінантного аналізу на підставі деяких ознак (незалежних змінних) індивідуум може бути зарахований до однієї з двох (або до однієї з кількох) заданих заздалегідь груп.

Така постановка завдання, особливо в разі двох заздалегідь заданих груп, дуже сильно нагадує постановку задачі для методу логістичної регресії. Ядром дискримінантного аналізу є побудова так званої дискримінантної функції. [3, 4]

|  |  |
| --- | --- |
|  | (2.1) |

де *x1 … хn* – значення змінних, відповідних даним випадкам,

константи *b1 … bn* і *a* – коефіцієнти, які і належить оцінити за допомогою дискримінантного аналізу.

Метою є визначення таких коефіцієнтів, щоб за значеннями дискримінантної функції можна було з максимальною точнустб провести поділ по групах. [3]

Результати дискримінантного аналізу для початкових змінних без введення нелінійних значень параметрів представлено на рис. 7.

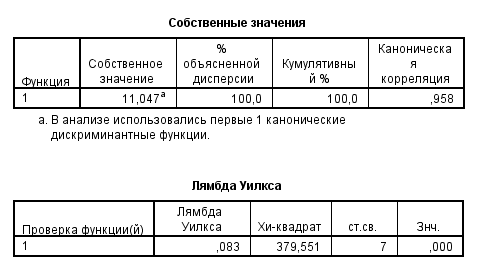


Рис. 7 – Результати Дискримінантного аналізу (1)

Проведено аналіз коефіцієнтів дискримінантної функції. Значення цієї функції повинні якомога точніше розділяти обидві групи. Мірою вдалості цього поділу служить кореляційний коефіцієнт між розрахованими значеннями дискримінантної функції і показником приналежності до групи.

Судячи по значенню коефіцієнта, рівному 0,958, кореляція не задовільна. За допомогою Лямбда Уїлкса проводиться тест на те, значимо чи в обох групах відрізняються один від одного середні значення дискримінантної функції; значення р <0,001, вказує на дуже значущу відмінність.

Великі власні значення (11,047) вказують на вдало підібрані коефіцієнти дискримінантні функції.

Класифікаційні результати представлені на рис. 8.

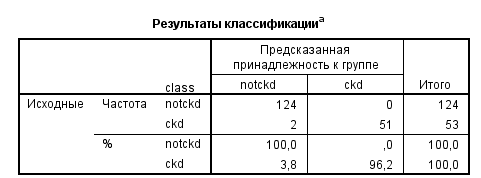


Рис. 8 – Класифікаційні результати дискримінантного аналізу (1)

Судячи з результатів, було визначено правильно лише 96,2 % пацієнтів, що мають захворювання нирок: з 53 пацієнтів 2 було класифіковано як хибнонегативні.

Коефіцієнти дискримінантної функції представлені на рис. 9.



Рис. 9 – Нормовані коефіцієнти канонічної дискримінантної функції (1)

Таким чином, дискримінантна функція для прогнозування вірогідності захворювання нирок на ранній стадії має вигляд:

Для підвищення якості моделі, як і для логістичної регресії до аналізу штучно були введені нелінійні змінні *x2,**x3* та *1/x* для кожного предиктора.

Результати дискримінантного аналізу при додаванні нелінійних змінних представлено на рис. 10.



Рис. 10 – Результати Дискримінантного аналізу (2)

За допомогою Лямбда Уїлкса проводиться тест на те, значимо чи в обох групах відрізняються один від одного середні значення дискримінантної функції; значення р <0,001, вказує на дуже значущу відмінність.

Підвищення власного значення на 7,476 вказує на підвищення точності дискримінантної функції.

Класифікаційні результати представлені на рис. 11.

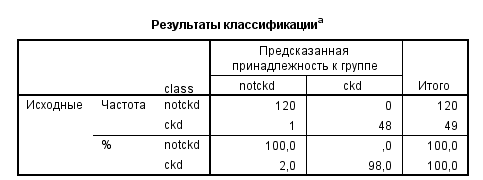


Рис. 11 – Класифікаційні результати дискримінантного аналізу (2)

На відміну від минулої моделі, де було правильно розпізнано 96,2 % хворих, дана модель правильно розкласифікувала 98 % хворих, що говорить про підвищення її точності.

Коефіцієнти дискримінантної функції представлені на рис. 12.

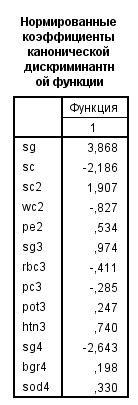


Рис. 12 – Нормовані коефіцієнти канонічної дискримінантної функції (2)

Таким чином, дискримінантна функція для прогнозування вірогідності захворювання нирок на ранній стадії має вигляд:

Висновки до розділу

За допомогою дискримінантного аналізубуло отримано дискримінантну функцію прогнозування, що з високою точністю (98 %) класифікує пацієнтів на тих, що мають захворювання нирок та тих, які не мають даного недугу. При чому краща модель була отримана при штучному введенні в аналіз нелінійних змінних *x2,**x3* та *1/x* для кожного предиктора.

РОЗДІЛ 4. 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. [5]

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

На початковому етапі були задані параметри класифікації за замовчуванням (див. рис. 13).

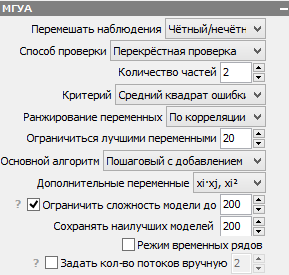


Рис. 13 – Параметри МГУА (1)

Класифікаційна модель (1):

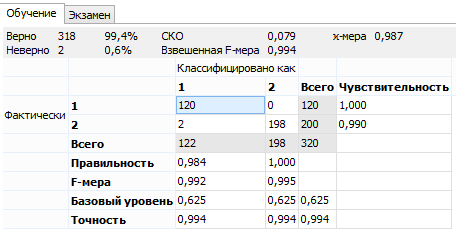


Рис. 14 – Таблиця точності класифікації МГУА при навчанні (1)

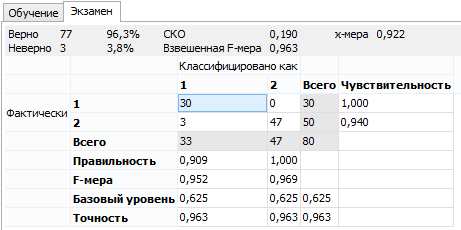


Рис. 15 – Таблиця точності класифікації МГУА при екзамені (1)

Для наглядності результатів можна використовувати ROC-криву, яка показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів.

Для ідеального класифікатора графік ROC-кривої проходить через верхній лівий кут, де частка істинно позитивних випадків становить 100% або 1.0 (ідеальна чутливість), а частка хибно позитивних прикладів дорівнює нулю. Тому чим ближче крива до верхнього лівого кута, тим вище прогнозувальна здатність моделі. При візуальній оцінці ROC-кривих розташування їх відносно один одного вказує на їх порівняльну ефективність. Крива, що розташовується вище і лівіше, свідчить про більшу прогнозувальну здатнысть моделі. [6]

Візуальне порівняння кривих ROC не завжди дозволяє виявити найбільш ефективну модель. Своєрідним методом порівняння ROC-кривих є оцінка площі під кривими. Теоретично вона змінюється від 0 до 1.0, але, оскільки модель завжди характеризуються кривою, розташованої вище позитивної діагоналі, то зазвичай говорять про зміни від 0.5 ( "даремний" класифікатор) до 1.0 ( "ідеальна" модель). [6]

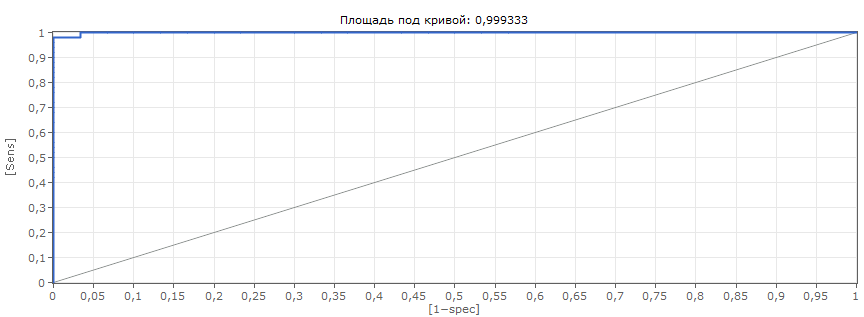


Рис. 16 – ROC-крива при екзамені (1)

Отримані результати мають достатній рівень чутливості, але будуються на дуже складній моделі. Для покращення результатів і знаходження моделі меншої складності з найбільш точним результатом було проведено декілька послідовних класифікацій зі зміною параметрів МГУА, що представлено надалі.

Не ранжуючи змінні по кореляції і обмежуючи складність моделі до 20 при заданих параметрах (див. рис. 17) отримано наступні результати.

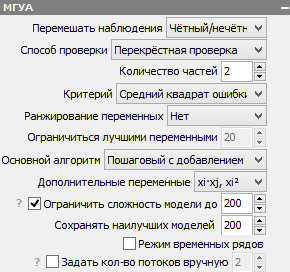


Рис. 17 – Параметри МГУА (2)

Класифікаційна модель (2):

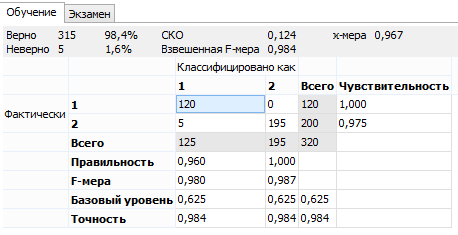


Рис. 18 – Таблиця точності класифікації МГУА при навчанні (2)

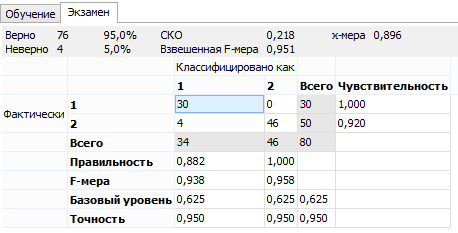


Рис. 19 – Таблиця точності класифікації МГУА при екзамені (2)

Візуально результати представлені у вигляді ROC-кривої (див. рис. 20).

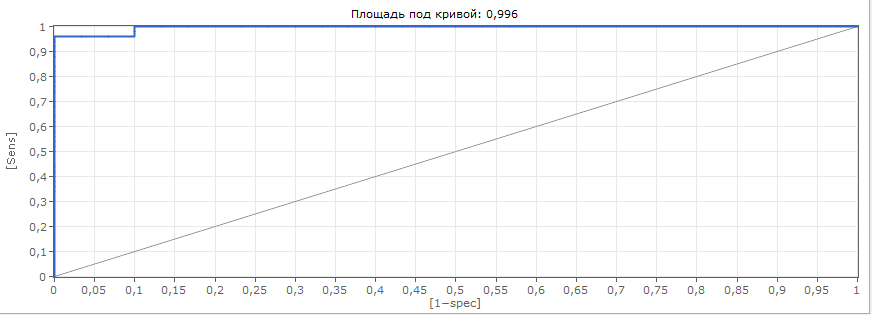


Рис. 21 – ROC-крива при екзамені (2)

З результатів видно, що при екзамені значно знизилась специфічність моделі, що є негативною зміною, але модель стала простішою, тому було зменшено кількість додаткових змінних і обмежено складність моделі до 10 але залишено ранжування по кореляції (див. рис. 22).

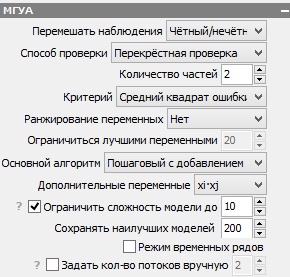


Рис. 22 – Параметри МГУА (3)

Класифікаційна модель (3):

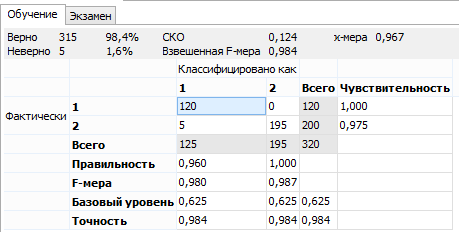


Рис. 23 – Таблиця точності класифікації МГУА при навчанні (3)

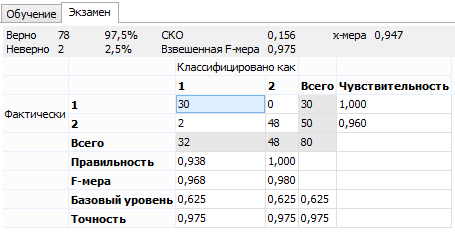


Рис. 24 – Таблиця точності класифікації МГУА при екзамені (3)

Візуально результати представлені у вигляді ROC-кривої на екзамені (див. рис. 25).

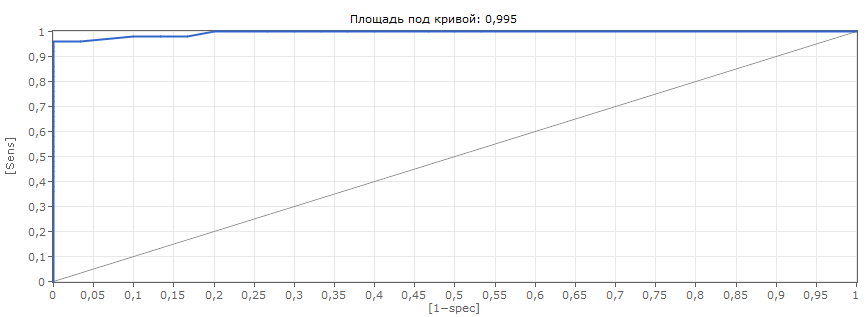


Рис. 25 – ROC-крива при екзамені (3)

За таких параметрів модель має високу чутливість і специфічність та низьку складність. У ході експериментів підвищення кількості додаткових змінних не дало позитивних результатів. Тому було вирішено прийняти за оптимальну останню отриману модель.

Висновки до розділу

При будові моделі прогнозування за допомогою програмного пакету GMDA Shell було поставлено задачу отримати найбільш точну модель прогнозування з найменшою складністю, тому було проведено декілька експериментів зі зміною параметрів МГУА. За результатами визначено:

1) при ранжуванні моделі за кореляцією специфічність моделі підвищується;

2) при виключенні додаткових змінних чутливість моделі значно зменшується, як на навчанні, так і на екзамені;

3) підвищення кількості додаткових змінних не дуже в сильній мірі впливає на чутливість моделі;

4) при обмеженні складності моделі до 10, модель має найбільшу специфічність.

При прийнятті рішення єдиної правильної моделі більше значення мали результат отримані при екзаменах.

ВИСНОВКИ

В даній розрахунковій роботі для вирішення задачи класифікації пацієнтів на групу здорових і хворих було використано три основні способи класифікації: логістичну регресію, дискримінантний аналіз та програмний пакет GMDA Shell, що працює за методом групового урахування аргументів. Основною задачею було поставлено отримати максимально просту і точну модель.

Спектр можливостей застосування логістичної регресії вужчий, ніж для дискримінантного аналізу, тому використання дискримінантного аналізу в якості універсального методу краще. Це пов'язано, зокрема, з тим, що при застосуванні методів логістичної регресії важко чітко уявляти, який тип мають залежна і незалежні змінні і, виходячи з цього, вибирати одну з трьох можливих процедур логістичної регресії: бінарну, поліноміальний або порядкову. При дискримінантному аналізі 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. Таким чином, дискримінантний аналіз є більш універсальною методикою.

При проведених експериментальних розрахунках отримано наступні результати:

1) Отримано логістичну модель на вхідних параметрах із штучно доданими нелінійними змінними *x2,**x3* та *1/x* для кожного предиктора, що з точністю 93% класифікувала хворих пацієнтів;

2) Отримано дискримінантну модель на вхідних параметрах із штучно доданими нелінійними змінними *x2,**x3* та *1/x* для кожного предиктора, що з точністю 98% класифікувала хворих пацієнтів;

3) Отримано оптимально просту та точну модель класифікації за допомогою МГУА, що на екзамені мала чутливість 0,96 і класифікувала вірно 97,5% пацієнтів.

Отже, так як змінна класифікації бінарна, то можна використовувати без додаткових розрахунків логістичну регресію. Введення нелінійних змінних в модель значно підвищує її точність. Тому оптимальним було обрано метод МГУА, що забезпечує швидкий пошук параметрів, при яких забезпечується достатня простота і точність моделі класифікації. При пошуку найбільш точної моделі більше приділялось значення чутливості моделі при екзамені, а для наглядності високої результативності даного методу було отримано ROC-криві.

**Список використаної літератури**

1. Хронічна хвороба нирок: посібник для пацієнтів [Електронний ресурс] // National Kidney Foundation. – 2007. – Режим доступу до ресурсу: https://www.kidney.org/sites/default/files/docs/11-50-1901\_lah\_patbro\_aboutckd\_pharmanet\_ukranian\_dec07.pdf.

2. L.Jerlin Rubini. Chronic Kidney Disease Data Set [Електронний ресурс] / L.Jerlin Rubini // The UCI Machine Learning Repository. – 2015. – Режим доступу до ресурсу: https://archive.ics.uci.edu/ml/datasets/Chronic\_Kidney\_Disease.

3. Буреева Н. Н. Многомерный статистический анализ с использованием ППП «STATISTICA». Н. Новгород: ННГУ, 2007. — 112 с.

4. Иллюстрированный самоучитель по SPSS [Електронний ресурс] – Режим доступу до ресурсу: http://www.datuapstrade.lv/rus/spss/.

5. GMDA Shell Documantation [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: http://d.gmdhshell.com/docs.

6. Паклин Н. Логистическая регрессия и ROC-анализ — математический апарат – Режим доступу до ресурсу: https://basegroup.ru/community/articles/logistic.

7. Мороз О. В. Порівняльний аналіз генераторів структур моделей у перебірних алгоритмах МГУА / О. В. Мороз, В. С. Степашко. // Індуктивне моделювання складних систем. – 2016. – №8. – С. 133–148.

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.