ВСТУП

Актуальність

Вимоги до алгоритмів моделювання та конкретні їх реалізації вар'юються в залежності від необхідних властивостей моделей, що необхідно отримувати при врахуванні обмежень на існуючі обчислювальні ресурси. Приклади необхідних властивостей – точність, ефективність оцінок, мінімальна чуттєвість до зміни області даних за помилкою 1-го та 2-го родів у шаговій регресійній процедурі і т.д. В залежності від специфіки використання моделей ті, чи інші критерії приймаються за основу при конструюванні конкретного алгоритму моделювання. Однак вирішення питання єдиності отриманої моделі, як правило, залишають за користувачем. У роботі розглядається можливість автоматичної оптимізації на принципах самоорганізації параметрів крокового алгоритму багатомірної регресії на прикладі синтезу моделі багатовимірної лінійної регресії.

Мета і задачі дослідження

Метою даної роботи є досягнення максимальної якості пронозування шляхом автоматизації вибору порогів включення/виключення аргументів для оптимізації алгоритму шагової багатовимірної лінійної регресії.

Задачі дипломної роботи:

1. Огляд існуючих методів регресійного аналізу та пошук можливих шляхів поліпшення точності прогнозування.
2. Формування алгоритму модифікації крокової лінійної регресії на принципах самоорганізації.
3. Формування зовнішнього критерію для перевірки точності прогнозування.

4. Проектування та програмна реалізація модифікованого алгоритму крокової регресії Stepwise на принципах самоорганізації.
5. Перевірка ефективності запропонованого алгоритму на реальних даних

Об’єкт дослідження – процес побудови моделі багатовимірної лінійної регресії.

Предмет дослідження – алгоритм структурно – параметричного синтезу моделі оптимальної структури.

Методи дослідження – лінійна регресія, крокова регресія, методи варіаційної статистики.

Наукова новизна роботи – вперше пропонується одержувати структуру моделі лінійної регресії шаговим алгортмом Stepwise, де його параметри оптимізуються за принципами самоорганізації. У роботі пропонується визначати оптимальні значення порогів на включення та виключення аргументів з моделі за зовнішнім критерієм, що поєднує вимоги до балансу якості прогнозування значень на навчальній та тестовій вибірках.

Проблематика роботи – підвищення якості прогнозування.

Практична значення одержаних результатів – розроблено програмний продукт, який може бути використаний в реальних дослідженнях для вирішення задачі прогнозування. Застосування вказаного алгоритму дозволило підвищити якість прогнозування у порівнянні з класичною версією Stepwise.

РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ З ТЕМИ ДОСЛІДЖЕННЯ

1.1 Розвиток регресійного аналізу

У сучасному суспільстві немає жодної сфери людської діяльності, де б не застосовувалася статистика, будь то економіка, екологія, медицина, природничі науки, політологія, соціологія, психологія і т.д. З допомогу статистики здійснюється наукова обробка, узагальнення та аналіз інформації, що характеризує розвиток економіки країни, охорони здоров'я, політики, культури і рівня життя населення. Статистика дозволяє виявити взаємозв'язок (закономірності), вивчити динаміку розвитку, провести аналіз для отримання обґрунтованих висновків і прийняття правильних рішень, які можуть бути застосовані на практиці.

Великим кроком в розвитку медичної статистичної науки стало застосування математичних методів і широке використання комп'ютерної техніки в аналізі медико-біологічних явищ.

Статистика, як будь-яка наука, вимагає визначення предмета дослідження. Предметом статистики виступають розміри і кількісні співвідношення якісно певних медико-біологічних явищ, закономірності їх взаємозв'язків і розвитку в конкретних умовах місця і часу. свій предмет статистика вивчає методом узагальнюючих показників. Для вивчення предмета статистики розробити й подати застосовуються специфічні прийоми, сукупність яких утворює методологію статистики (методи масових спостережень, угруповань, узагальнюючих показників, динамічних рядів, індексний метод і ін.). Застосування в статистиці конкретних методів зумовлюється поставленими завданнями і залежить від характеру вихідної інформації.

Дослідження зв'язків в умовах масового спостереження і дії випадкових факторів здійснюється, як правило, з допомогою медико-статистичних моделей. В широкому сенсі модель - це аналог, умовний образ (зображення, опис, схема, креслення і т.п.) будь-якого об'єкта, процесу або події, наближено відтворює «оригінал». модель являє собою логічне або математичний опис компонентів і функцій, що відображають істотні властивості модельованого об'єкта або процесу, дає можливість встановити основні закономірності зміни оригіналу. У моделі оперують показниками, обчисленими для якісно однорідних масових явищ (сукупностей). Вираз моделей у вигляді функціональних рівнянь використовують для розрахунку середніх значень модельованого показника по набору заданих величин і для виявлення ступеня впливу на нього окремих факторів. За кількістю включаються факторів моделі можуть бути однофакторний і багатофакторним (два і більше факторів). Залежно від пізнавальної мети статистичні моделі підрозділяються на структурні, динамічні і моделі зв'язку. Найбільш розробленою в теорії статистики є методологія так званої парної кореляції, яка розглядає вплив варіації факторного ознаки X на результативний ознака Y і представляє собою однофакторний кореляційний та регресійний аналіз.

Поняття кореляції в прийнятому нами значенні з'явилося майже в середині XIX століття завдяки роботам сера Френсіса Гальтона (двоюрідного брата Чарльза Дарвіна) і Карла Пірсона. Ф. Гальтон застосував для кореляції наступну форму запису: co-relation, звідки стає зрозумілим значення цього виразу - зв'язок, співвідношення. Спочатку дослідження кореляції проводились в галузі природничих наук, перш за все в біології. Лише пізніше застосування методів кореляційного аналізу поширилося на економіку, де вони привели до вельми корисним результатами [31].

Поняття регресії також сходить до Ф. Гальтону. Після знайомства з книгою Чарльза Дарвіна «Походження видів» в 1859 р Ф. Гальтона стала

займати думка про те, чому люди з покоління в покоління не сильно розрізняються за зовнішнім виглядом і природним здібностям. Це привело його до вивчення спадковості. Зокрема, він зайнявся з'ясуванням залежності зростання дітей від зростання батьків [24]. За логікою діти повинні бути кожен раз дуже схожі на своїх батьків. Високі батьки повинні мати високих дітей, а низькорослі батьки - дітей низького зросту. При такому стані речей через кілька поколінь ми мали б, з одного боку, рід велетнів, а з іншого - рід карликів. Але незабаром в результаті великих статистичних досліджень і дослідів над тваринами Ф. Гальтон переконався, що такої тенденції немає, а, скоріше, навпаки, діти дуже високих або дуже низьких батьків в середньому мають менш високий або відповідно менше низький зріст. Крім того, ухилення зростання дітей не таким значним, як ухилення зростання їх батьків від середнього зросту досліджених осіб. Це рух назад в напрямку до середнього Ф. Гальтон назвав регресією (to regress - рухатися в зворотному напрямку) [27].

У 1885 р була видана відома робота Ф. Гальтона «Регресія в напрямку до загальної середньої розміром при спадкуванні зростання», де він приходить до висновку, що, загалом, ознаки батьків не повністю успадковуються дітьми, і чим віддаленіші предок, тим в меншій мірою позначається його властивості на нащадку. «Закон регресії вагомо свідчить проти повного наслідування якої-небудь ознаки. З великого числа дітей тільки мало хто буде ухилятися від середнього рівня в порівнянні з ухиленням одного з батьків, що відрізняється своїми природними якостями. Чим яскравіше талант одного з батьків, тим рідше батьки мають щастя бачити, що природа також щедро обдарувала їх сина, і ще рідше буває, щоб обдарованість передавалася в наступні покоління. Закон неупереджений і об'єктивний. Він рівномірно розподіляє успадкування хороших і поганих ознак [26]. Він руйнує надмірні ілюзії одного обдарованого батька, який плекає мрію, що його діти успадкують всі його здібності. Закон усуває також

перебільшені побоювання щодо того, що дітям передадуться все слабкості, недоліки і хвороби батьків. Зрозуміло, ці твердження не знаходяться в суперечності з загальною теорією, згідно з якою діти талановитих батьків мають велику ймовірність мати які-небудь дарованнями, ніж діти батьків із середніми здібностями. Наші міркування висловлюють тільки той факт, що найобдарованіший з усіх дітей небагатьох високообдарованих батьківських пар не так буде талановитий, як найобдарованіший з усіх дітей дуже багатьох батьківських пар із середніми здібностями» [30]. Поняття регресії, що застосовується спочатку тільки для процесів з тенденцією зрушуватися в напрямку до середнього, з плином часу все більше узагальнювалося і сьогодні служить для характеристики односторонньої стохастичною залежності [5].

1.2 Цілі регресійного аналізу

1. Визначення ступеня детермінованості варіації критеріальною (залежної) змінної предикторами (незалежними змінними)
2. Передбачення значення залежної змінної з допомогою незалежної (-их)
3. Визначення вкладу окремих незалежних змінних в варіацію залежної

Регресійний аналіз не можна використовувати для визначення наявності зв'язку між змінними, оскільки наявність такого зв'язку і є передумова для застосування аналізу.

1.2.1 Математичне визначення регресії

Строго регресійну залежність можна визначити наступним чином. Нехай $Y, X_{1,} X_{2},..., X_{p}$ - Випадкові величини з заданим спільним розподілом ймовірностей. Якщо для кожного набору значень $X_{1} = x_{1,} X_{2} = x_{2},..., X_{p} = x_{p}$ визначено умовне математичне сподівання $y (x_{1,} x_{2},..., x_{p)} = E (Y | X_{1} = x_{1,} X_{2} = x_{2},..., X_{p} = x_{p)}$ (Рівняння лінійної регресії в загальному вигляді),то функція $y (x_{1,} x_{2},..., x_{p)}$ називається регресією величини Y за

величинами $X_1, X_2,..., X_p$, А її графік – лінією регресії Y по $X_1, X_2,..., X_p$, або рівнянням регресії [23].

Залежність Y від $X_1, X_2,..., X_p$ проявляється в зміні середніх значень Y при зміні $X_1, X_2,..., X_p$. Хоча при кожному фіксованому наборі значень $X_1 = x_1, X_2 = x_2,..., X_p = x_p$ величина Y залишається випадковою величиною з певним розсіюванням [31].

Для з'ясування питання, наскільки точно регресійний аналіз оцінює зміну Y при зміні $X_1, X_2,..., X_p$, Використовується середня величина дисперсії Y при різних наборах значень $X_1, X_2,..., X_p$ (Фактично мова йде про міру розсіяння залежної змінної навколо лінії регресії).

1.2.2 Метод найменших квадратів (розрахунок коефіцієнтів)

На практиці лінія регресії найчастіше шукається у вигляді лінійної функції $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + B_N X_N$ (Лінійна регресія), найкращим чином наближає шукану криву. Робиться це за допомогою методу найменших квадратів [26], коли мінімізується сума квадратів відхилень реально спостережуваних Y від їх оцінок $\hat{Y}$ (Маються на увазі оцінки за допомогою прямої лінії, яка претендує на те, щоб представляти шукану регресійну залежність):

$$\sum_{k=1}^{M}(Y_k - \hat{Y}_k)^2 \to min \tag{1.1}$$

[16] (M - обсяг вибірки). Цей підхід заснований на тому відомому факті, що фігурує в наведеному вираженні сума приймає мінімальне значення саме для того випадку, коли $Y = y(x_1, x_2,... x_N)$ [24].

Для вирішення завдання регресійного аналізу методом найменших квадратів вводиться поняття функції нев'язки [17]:

$$\sigma(\bar{b}) = \frac{1}{2}\sum_{k=1}^{M}(Y_k - \hat{Y}_k)^2 \tag{1.2}$$

Умова мінімуму функції нев'язки:

$$\begin{cases} \frac{d\sigma(\bar{b})}{db_i} = 0 \\ i = 0...N \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{M} y_i = \sum_{i=1}^{M} \sum_{j=1}^{N} b_j x_{i,j} + b_0 M \\ \sum_{i=1}^{M} y_i x_{i,k} = \sum_{i=1}^{M} \sum_{j=1}^{N} b_j x_{i,j} x_{i,k} + M b_0 \sum_{i=1}^{M} x_{i,k} \\ k = 1...N \end{cases} \quad (1.3)$$

Отримана система є системою N + 1 лінійних рівнянь з N + 1 невідомими $b_0 ... b_N$

Якщо уявити вільні члени лівій частині рівнянь матрицею

$$B = \begin{Bmatrix} \sum_{i=1}^{M} y_i \\ \sum_{i=1}^{M} y_i x_{i,1} \\ ... \\ \sum_{i=1}^{M} y_i x_{i,N} \end{Bmatrix} \quad (1.4)$$

а коефіцієнти при невідомих у правій частині матрицею

$$A = \begin{bmatrix} M & \cdots & \sum_{i=1}^{M} x_{i,N} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} x_{i,N} & \cdots & \sum_{i=1}^{M} x_{i,N} x_{i,N} \end{bmatrix} \quad (1.5)$$

То отримуємо матричне рівняння: $A \times X = B$, Яке легко вирішується методом Гаусса. Отримана матриця буде матрицею, яка містить коефіцієнти рівняння лінії регресії [19]:

$$X = \begin{Bmatrix} b_0 \\ b_1 \\ ... \\ b_N \end{Bmatrix} \quad (1.5)$$

Для отримання найкращих оцінок необхідно виконання передумов МНК (умов Гаусса-Маркова). В англомовній літературі такі оцінки називаються BLUE (Best Linear Unbiased Estimators) - найкращі лінійні незміщені оцінки [22].

1.2.3 Інтерпретація параметрів регресії

Параметри b_i є коефіцієнтами кореляції; b_i інтерпретується як частка дисперсії Y, пояснена X_i, При закріпленні впливу інших предикторів, тобто вимірює індивідуальний внесок X_i в пояснення Y. У разі корелюють предикторів виникає проблема невизначеності в оцінках, які стають залежними від порядку включення предикторів в модель. У таких випадках необхідно застосування методів аналізу кореляційного і покрокового регресійного аналізу. Говорячи про нелінійних моделях регресійного аналізу, важливо звертати увагу на те, чи йде мова про нелінійності по незалежним змінним (з формальної точки зору легко зводиться до лінійної регресії), або про нелінійності по оцінюваним параметрами (що викликає серйозні обчислювальні труднощі). При нелінійності першого виду з змістовної точки зору важливо виділяти поява в моделі членів виду $X_1 X_2$, $X_1 X_2 X_3$, Що свідчить про наявність взаємодій між ознаками X_1, X_2 і т. д. [1-4].

1.3 Зовнішні критерії оцінки моделі

Перш, ніж говорити про способи вибору методу навчання, необхідно сформулювати критерії вибору.

Внутрішній критерій - це функціонал Qint (μ, Xℓ)

Характеризує якість методу μ по навчальній вибірці Xℓ. Канонічним прикладом внутрішнього критерію є функціонал помилки навчання (навчальна помилка):

Внутрішні критерії використовуються для налаштування параметрів обраної моделі алгоритмів A. Наприклад, метод мінімізації емпіричного

ризику будує алгоритм, що доставляє мінімальне значення внутрішнього критерію [23]:

Внутрішні критерії не можна використовувати для вибору струткту моделі, так як при цьому буде заохочуватися перенавчання. Зовнішній критерій характеризує якість методу μ за тими даними, які не використовувалися в процесі навчання. Зовнішній критерій перевіряє, чи дійсно отриманий алгоритм добре працює в реальних умовах, коли правильні відповіді Yi заздалегідь не відомі.

Ідея застосування зовнішніх критеріїв для підбору оптимальної структури моделі була запропонована А. Г. Івахненко наприкінці 60-х у методі групового урахування аргументів, МГУА (метод групового обліку даних, GMDH) [5]. Це один з успішних методів, за допомогою якого були вирішені сотні прикладних задач. У частності, досвід МГУА показав, що має сенс використовувати відразу декілька зовнішніх критеріїв, що характеризують якість методу навчання з різних точок зору. Розглянемо найбільш відомі типи зовнішніх критеріїв. Будемо вважати, що всі критерії, будь-то зовнішні або внутрішні, потрібно мінімізувати. Чим менше значення критерію Q (μ), тим вища якість методу μ.

Критерій середньої помилки на контрольних даних

Найпростішим прикладом зовнішнього критерію є функціонал середньої помилки на заданій контрольної вибірці Xk, званий помилкою узагальнення (generalization error). Зрозуміло, на об'єктах XI ∈ Xk також повинні бути відомі правильні відповіді Yi = Y * (XI). Тому будемо вважати, що вихідна повна вибір-ка XL = Xℓ ∪ Xk деяким чином розбита на навчальну і контрольну частини, L = ℓ + K. Зовнішній критерій є функцією методу μ і повної вибірки XL:

У МГУА його прийнято називати критерієм регулярності, в англомовній литературі - помилкою на відкладених даних ((hold-out error).

Вибірки Xℓ і Xk повинні бути не тільки непересічними, але й незалежними. Критерій фактично залишиться внутрішнім, якщо контрольна вибірка Xk буде спеціально складена з об'єктів, що збігаються або незначно відрізняються від об'єктів навчання Xℓ. На практиці наявну вибірку даних XL розбивають на навчання і контроль випадковим чином. У контрольній вибірці, як правило, залишають від чверті до половини об'єктів.

Критерій ковзаючого контролю

Це критерій є узагальненням попереднього. Щоб результат не залежав від способу розбиття, беруть кілька різних разбиений вихідної вибірки XL на навчання і контроль XL = Xℓ N ∪ Xk N, N = 1 ,. , , , N, і середню помилку на контролі усредняют по розбиття. Цей функціонал називається помилкою ковзаючого контролю (cross-validation error, CV):

Зазвичай після вибору методу μ за критерієм ковзаючого контролю кінцевий алгоритм ще раз навчають по повній вибірці: = μ (XL). Але можна також вибрати в якості вирішення кращий з алгоритмів, вже побудованих по подвибор- кам, що не вимагає додаткового застосування методу μ

Залежно від способу формування разбиття розрізняють кілька видів ковзаючого контролю.

Повний ковзний контроль (complete CV)

Будується за всіма N = C K L разбиття. Це число стає занадто великим вже при до> 2, тому повний ковзний контроль використовується або в теоретичних дослідженнях, або в тих рідкісних випадках, коли для нього вдається вивести ефективну обчислювальну формулу. Наприклад, для методу до найближчих сусідів така формула отримана в [6]. На практиці частіше застосовуються інші різновиди ковзаючого контролю.

Контроль за окремими об'єктами (leave-one-out CV)

Окремий випадок повного ковзного контролю при K = 1: Loo (μ, XL) = 1 LXL я = 1 Q μ (XL \ {x}) {Xi}.

Це, мабуть, найпоширеніший варіант ковзаючого контролю. Переваги LOO в тому, що кожен об'єкт рівно один раз бере участь у контролі, а довжина навчальних підвибірок лише на одиницю менше довжини повної вибірки.

Недоліком LOO є велика ресурсомісткість, так як навчатися приходиться L разів. Деякі методи навчання дозволяють досить швидко переналаштовувати внутрішні параметри алгоритму при заміні одного навчального об'єкта іншим. У цих випадках обчислення LOO вдається помітно прискорити.

Контроль за Q блокам (q-fold CV).

Вибірка випадковим чином розбивається на Q непересічних блоків однаковою (або майже однаковою) довжини l1,LQ

Кожен блок по черзі стає контрольної підвибірки, при цьому навчання проводиться по решті Q-1 блокам. Критерій визначається як середня по всім блокам помилка на контролі:

Це компроміс між ЛОО й утримання відмови. З одного боку, навчання про диться тільки Q раз замість L. З іншого боку, довжина навчальних підвибірок L Q-1 Q (з точністю до округлення) не сильно відрізняється від довжини повної вибірки L. Зазвичай вибірку розбивають випадковим чином на 10 або 20 блоків.

Контроль за випадковим підвибірках.

Розбиття п = 1 ,... N вибираються випадково, незалежно і равноймовірно з безлічі всіх C K L розбиття. Позначимо Qn = Q μ (Xℓ п), Хк. Якщо випадкова величина Q = Q μ (Xℓ), Xk має безперервне розподіл, то ймовірність події Q> maxn Qn не перевищує 1 N + 1. Аналогічно можна оцінити і двосторонній довірчий інтервал: ймовірність того, що Q / ∈ [Minn Qn, maxn Qn] не перевищує 2 N + 1. Таким чином, для отримання верх- ній

оцінки з надійністю 0,95 достатньо взяти N = 19, а для двосторонньої оцінки N = 39 розбиттів.

Бутстреп (bootstrap)

Нагадує контроль по випадковим підвибірках. Відмінність в тому, що об'єкти вибираються з поверненням, при цьому довжина навчальних підвибірок завжди дорівнює довжині повної вибірки, зате в них утворюються повтори.

Проблема показності підвибірок

Виникає у всіх критеріях, які обслуговують випадкові розбиття. Навчальні та контрольні підвибірки повинні володіти тими ж статистичними характеристиками, що і повна вибірка XL. В іншому випадку вибір моделі та налаштування її параметрів будуть погано погодити вани один з одним.

У задачах класифікації рекомендується зберігати в кожній підвибірки ті ж пропорції розподілу об'єктів по класах, що і на всій вибірці. Цей прийом називається стратифікацією (стратифікація) вибірки.

Крім того, необхідно забезпечити рівномірний розподіл кожної підвибірки по всьому простору X. На практиці поступають таким чином. Повна вибірка XL упорядковується по деякому спеціально виділеному признаку, і кожен L K -й об'єкт, L = 1 ,. , , , K, заноситься в контрольну частину. Аналогічно здійснюється розподіл вибірки по блоках у разі Q-кратне розділення. Виділена ознака визначається виходячи з особливостей завдання. Це може бути деяка вихідна ознака Fj (XI), лінійна комбінація кількох вихідних при знаків цільової ознака Ю.І., оцінка рівня шуму цільового ознаки Ю.І., відстань від вектора признакового опису об'єкта XI до центру мас вибірки, і т. д.

1.3.1 Критерії несуперечності

Ця група критеріїв заснована на властивості завадостійкості, що також йде від МГУА.

Для зовнішнього критерію визначається властивість завдостійкості, під якою розуміється його здатність вибрати із усієї множини моделей таку модель, що котра достатньо точно відновлює незашумлений вихід об’єкта. Характеристикою шуму є його рівень

$$\theta=\frac{\sigma^2}{c^2}\ , \qquad (1.1)$$

Де σ^2 – дісперсія, с – «потужність шуму»[11].

У найпростішому випадку критерій несуперечності визначається як середня невязка відповідей двох алгоритмів, побудованих за двома випадковим непересікаючихся підвибірках однакової або майже однакової довжини,

Xℓ ∪ Xk = XL: Qext (μ, XL) = 1 LXL я = 1 a1 (XI) - a2 (XI), a1 = μ (X ℓ), a2 = μ (X К). У більш складних варіантах критерію вибірка розбивається декількома різними способами, як при ковзному контролі.

Відмінність двох алгоритмів не обов'язково вимірювати як невязку їх відповідей на вибірці. Якщо α1, α2 - вектори параметрів алгоритмів А1, А2, то зовнішній критерій можна визначити як відстань між цими векторами у відповідній метриці:

ρ: Qext (μ, XL) = ρ (α1, α2). Необхідною умовою для застосування цих критеріїв є надмірність вихідних даних. Половина вибірки повинна бути досить представницької, щоб по ній можна було побудувати алгоритм прийнятної якості. Тому критерії даного типу не рекомендується застосовувати у разі малих вибірок.

1.3.2 Критерії регуляризації

Ці критерії вводяться в тих випадках, коли задача навчання алгоритму по вибірці виявляється нестійкою - багато алгоритми доставляють внутрішньому кри- терію значення, близьке до оптимального, однак далеко не всі з них мають гарну узагальнюючої здатністю.

Ідея регуляризації полягає в тому, щоб накласти обмеження на вектор параметрів алгоритму, або ввести штраф за вихід вектора параметрів з деякої допустимої області. Критерії регуляризації не так універсальні, як попередні - їх вигляд залежить від конкретної моделі алгоритмів.

Наприклад, в лінійних моделях регресії та класифікації різке збільшення норми вектора параметрів kαk в алгоритмі = μ (Xℓ), як правило, свідчить про перенавчання. При цьому модель стає неадекватною, з'являються великі за модулем негативні і позитивні коефіцієнти, які вже не можна інтерпретувати як ступінь важливості відповідної ознаки. Тому в якості зовнішнього критерію регуляризації беруть суму внутрішнього критерію і штрафного доданка. При $\tau \to 0$ рішення нестійка; при $\tau \to \infty$, навпаки, вироджується в константу. Підбір τ позволя- ет знайти компромісне між двома крайностями. З функціоналами такого виду ми вже стикалися при обговоренні проблеми мультиколінеарності в лінійному дискримінант Фішера і багатовимірної лінійної регресії.

Перевага цього критерію, у порівнянні зі ковзаючому контролем, в тому, що немає необхідності багаторазово застосовувати ресурсномісткий метод навчання. Ос новних проблема - необхідність підбирати значення параметра регуляризації τ.

1.4 Критерії, засновані на оцінках узагальнюючої здатності

Теорія Вапніка-Червоненкіса

Дає верхні оцінки частоти помилок на контрольній вибірці, які можна використовувати в якості зовнішнього критерію [7]:

де H - розмірність Вапніка-Червоненкіса (ємність) моделі алгоритмів; η - уровень значущості - ймовірність, з якою дана оцінка має право порушуватися. Функція втрат L (x) у функціоналі Q (, Xℓ), зобов'язана бути двозначною і прий мати тільки значення 0 або 1. Для лінійної моделі класифікації ємність ч збігається з розмірністю простору параметрів і з числом ознак п.

Інформаційний критерій Акаіке

Є оцінкою мат.очікування середньої помилки на незалежних контрольних даних. Він виводиться з припущень, що модель алгоритмів лінійна, розмірність вектора параметрів дорівнює п, і функцио- готівка Q відповідає принципу максимуму правдоподібності:

Це означає, що задана імовірнісна функція втрат L (x) = - LN p (x, (x)), де p (x, y) - щільність імовірнісного розподілу на безлічі X × Y, згідно з яким і отримана навчальна вибірка (Xi, Yi) ℓ = 1. Інформаційний критерій Акаіке (Akaike Information Criterion, AIC) [8]:

AIC (μ, Xℓ) = Q (μ (X ℓ), Xℓ) + 2σ 2 ℓ п, де σ 2 - оцінка дисперсії випадкової величини ξ (x) = y * (x) - * (x), представля- -ющей собою відхилення найкращого в рамках використовуваної моделі алгоритму * від невідомої цільової функції y *. Зокрема, для задач класифікації можна скористатися оцінкою σ.

На практиці критерій AIC часто застосовують і до нелінійних моделям, що не завжди добре обгрунтовано, але в багатьох випадках призводить до вибору моделей цілком прийнятної якості.

Баєсовский інформаційний критерій (Bayesian Information Criterion, BIC)

Так само, як і AIC, випливає з принципу максимуму правдоподібності. Лінійність моделі не передбачається, проте, число параметрів п все одно виникає завдяки використанню апроксимації Лапласа [8]:

BIC (μ, Xℓ) = ℓ σ 2 Q (μ (X ℓ), Xℓ) + σ 2 LN ℓ ℓ п ,

При ℓ> 8 критерій БІК схильний сильніше штрафувати складні моделі, ніж АШС.

Особливістю критерію BIC є те, що він не тільки дозволяє вибрати кращу модель, але і дає оцінку апостеріорної ймовірності кожної моделі.

Якщо вибір проводився з T моделей A1 ,…, AT, то ймовірність пт, що дані Xℓ були породжені моделлю B, дається формулою Байєса.

Вибір методу за сукупністю критеріїв

В МГУА рекомендується використовувати для вибору оптимальної моделі декілька принципово різних зовнішніх критеріїв. Як правило, один зовнішній критерій відбирає кілька кращих методів, якість яких не відрізняється в межах природного похибки критерію. Утворюється додаткова свобода вибору, якою доцільно розпорядитися за допомогою другого зовнішнього критерію. Інший варіант - обчислювати зважену суму декількох критеріїв, але це викликає проблему вибору вагових коефіцієнтів. Агрегований критерій залежить від них істотно, а з яких міркувань їх призначати не є явним [9]. Більш прийнятним представляється двоступеневий відбір. Практична рекомендацію - відібрати деяку кількість кращих методів за критерієм ковзаючого контролю; потім з них вибрати той, для якого критерій регуляризації (або критерій несуперечності) приймає найменше значення.

1.5 Основна відмінність зовнішніх і внутрішніх критеріїв.

У міру збільшення складності моделі | G | внутрішній критерій Qint (G) = Qint (μG, XL), Як правило, монотонно убуває. Зовнішній критерій Qext (G) =Qext (μG, XL) спочатку зменшується, потім проходить через точку мінімуму і далі тільки зростає. Це типова поведінка критеріїв підтверджується як теоретично, так і експериментально.

Корисно побудувати на одному графіку криві внутрішнього критерію і деяких зовнішніх критеріїв. Іноді виявляється, що мінімуми зовнішніх критеріїв досягаються не тільки на різних моделях G, але навіть при різних значенях складності | G |. Всі критерії залежать від даних, отже, мають деякий розкид (дисперсію). Графік дозволяє оцінити рівень шуму візуально і визначити інтервал допустимих значень складності, в якому Qext незначимо

відрізняється від мінімуму. Застосування сукупності зовнішніх критеріїв дозволяє знайти перетин цих інтервалів і з більшою впевніністю визначити оптимальну модель.

У всіх методах будується нижня огинає безлічі точок (| G |, Qext (G)). У разі повного перебору її мінімум відповідає оптимальному набору G опитмальної складності j* = | G* |. Решта алгоритми вирішують задачу пошуку оптимального набору ознак лише приблизно[5].

Висновки до розділу

За результатами аналітичного огляду літературних джерел інформації проаналізовано сучасний стан та тенденцію розвитку методів, алгоритмів регресійного аналізу. Проаналізовано основні зовнішні критерії, що використовуються у МГУА, показана необхідність у створенні оптимізуючих алгоритмів, які використовують зовнішній критерій, що враховує баланс ваги помилки на навчальній та перевірочній вибірках, та вирішують питання перенавчання.

РОЗДІЛ 2 РОЗРОБКА АЛГОРИТМУ ЛІНІЙНОЇ РЕГРЕСІЇ

Вступ

Вимоги до алгоритмів моделювання[10] та конкретні їх реалізації вар'юються в залежності від необхідних властивостей моделей, що необхідно отримувати при врахуванні обмежень на існуючий обчислювальний ресурс. Приклади необхідних властивостей наведені у розділі 1: точність, ефективність оцінок, мінімальна чуттєвість до зміни області даних[11] за помилкою моделі та по дисперсії оцінки параметрів. Значення помилок 1-го та 2-го родів у кроковій регресійній процедурі та ін. В залежності від специфіки використання моделей, ті чи інші критерії[12-14] приймаються за основу при конструюванні конкретного алгоритму моделюівння. У цьому розділі розглядається можливість автоматичної оптимізації алгоритму багатомірної крокової лінійної регресії на принципах самоорганізації на прикладі синтезу лінійної моделі.

2.1 Постановка задачі

Задана матриця вхідних спостережень $x \in R^M$ та вектор залежної змінної $Y \in \{0,1\}$ (рис.2.1)

$$\begin{matrix} x_{11} & \dots \\ x_{21} & \dots \\ \dots & \dots \\ x_{n1} & \dots \end{matrix}$$
$$\begin{pmatrix} x_{1M} & y_1 \\ x_{2M} & y_2 \\ \dots & \dots \\ x_{\text{nM}} & y_n \end{pmatrix},$$

Рисунок 2.1 – матриця вхідних спостережень

де n – кількість спостережень, M – кількість змінних, з яких необхідно обрати m найкращих аргументів, що пояснюють модель.

Необхідно запропонувати алгоритм структурно – параметричного синтезу моделі оптимальної структури

$$Y=\sum_{j=1}^{p} w_j x_i^{j-1}+\varepsilon_i, \quad \text{де} \quad p \text{ — кількість параметрів моделі} \qquad (2.1)$$

2.2 Алгоритм побудови модифікованого алгоритму крокової лінійної регресії

Було запропоновано оптимізувати параметри алгоритму крокової логістичної регресії за алгоритмом (рис.2.2):

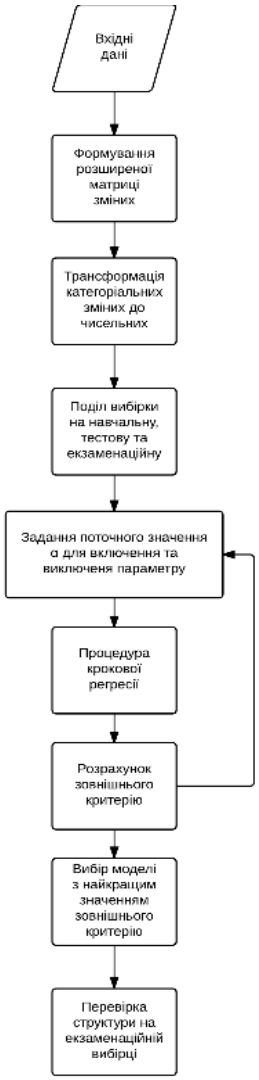

Рисунок 2.2 - Модифікований алгоритм крокової регресії

Алгоритм складається з наступних етапів:

1. Формування розширеної матриці змінних *x*.
2. Трансформування категоріальних змінних до чисельних.
3. Вибірка спостережень поділяється на навчальну, перевірочну та екзаменаційну у заданому співвідношенні довільним чином.
4. Задається поточне значення $\alpha_{\text{вкл}}$ и$\alpha_{\text{искл}}$ сітки оптимізуємих параметрів алгоритму.
5. Для поточного значення параметрів алгоритму проводиться модифікована крокова процедура[14] з визначенням структури, коефіцієнтів лінійної моделі та розрахунком значення зовнішнього критерія $I_{вн}$. Вирішенням крокової процедури вважається модель, для якої отримано мінімум зовнішнього критерія. Структура, коефіцієнти та значення зовнішнього критерія $K_{\text{РН}}$ запам'ятовуються.
6. Якщо всі значення сітк були перебрані, переходимо до п.7, якщо ні - задається наступне значення сітки оптимізуємо параметрів алгоритма, , перехід до п. 4.
7. З отриманих моделей вибирається та, для якої отримано найкраще значення зовнішнього критерія.
8. Кінцевою оцінкою якості отриманої модели вигляду (2.1) вважаємо якість прогнозування на екзаменаційній вибірці.

2.3 Розбиття вибірок

Для роботи алгоритму необхідно розбити вибірку на 2 частини – навчальну та перевірочну довільним чином у заданих пропорціях.

Нехай вибірка позначена W , тоді розбиття буде виглядати наступним чином:

$$W=A+B+C \quad , \qquad (2.5)$$

де A , B , C- навчальна, перевірочна та екзаменаційна вибірки відповідно.

Розрахунок коефіцієнтів лінійних моделей у кроковій процедурі буде відбуватися на навчальній вибірці, значення зовнішнього критерію для вибору структури моделі буде оцінюватися як показник якості на навчальній та перевірочній вибірках, результуюча оцінка – точність прогнозування на екзаменаційній вибірці.

2.3 Кроковий алгоритм

Модифікований кроковий алгоритм для фіксованих значень $\alpha_{\text{вкл}}$ и$\alpha_{\text{искл}}$ складається з декількох етапів:

1 Включення предиктора в модель:

1.1 Проводимо F- тест[15] порівнюючи модель отриману на попередній ітерації з моделью, що включає предиктор x_i для кожного предиктора ще не включенного до моделі:

$$F_i = \frac{\text{SSR}_{\text{prev}+X_i} - \text{SSR}_{\text{prev}}}{\text{MSR}_{\text{prev}+X_i}}, \tag{2.6}$$

$$\text{де} \quad \text{MSR}_{\text{prev}+X_i} = \frac{\text{SSR}_{\text{prev}+X_i}}{\vartheta_d}, \vartheta_d = \text{n} - m - 2, \tag{2.7}$$

$$\text{SSR} = \sum_{\text{i}=1}^{n} \left(Y_i - Y\right), \tag{2.8}$$

$\text{i}=1..k1$, $k1$ – кількість предикторів претендентов, що раніше не були включені до моделі, Y – табличне значення вихідної змінної, Y_i – значення регресійної моделі, n – кількістьспостережень у виборці, m – кількість змінних у моделі, індекс *prev* – позначає модель, отриману на попередній ітерації.

1.2 Обираємо предиктор з найбільшим значенням F. Якщо рівень значимості α, відповідає отриманому значенню критерию F менше фіксованого $\alpha_{\text{вкл}}$, $\alpha < \alpha_{\text{вкл}}$, то приймається гіпотеза про включення предиктора до моделі.

2. Виключення предикторів:

2.1 Проводимо F- тест, порівнюючи поточну модель з моделлю, що не включає предиктор x_i для кожного предиктора з моделі:

$$F_i = \frac{\mathrm{SSR}_{\mathrm{curr}} - \mathrm{SSR}_{\mathrm{curr}-X_i}}{\mathrm{MSR}_{\mathrm{curr}}}, \tag{2.9}$$

$$\mathrm{MSR}_{\mathrm{curr}} = \frac{\mathrm{SSR}_{\mathrm{curr}}}{\vartheta_d}, \tag{2.10}$$

де $\mathrm{i}=1..k2$, $k2$ – кількість предикторів, що не включені в модель, індекс *curr* –позначає модель, отриману на поточній ітерації.

2.2 Якщо рівень значимості α , відповідає отриманому значенню критерія F, більше фіксованого $\alpha_{\text{искл}}$, : $\alpha > \alpha_{\text{искл}}$, то приймається гіпотеза про виключення предиктора з моделі. Серед зміних, що пройшли F- тест для виключення обираємо предиктор з найменшим значенням F.

2.3 Для кожної з порівнюваних вище моделей вирішується задача найменших квадратів на навчальній вибірці даних:

$$\mathrm{b} = (X^{\mathrm{T}} X)^{-1} X^T Y \tag{2.13}$$

Розраховуємо та запам’ятовуємо значення зовнішнього критерія для кожної моделі. Якщо предиктори для включення не вичерпані - переходимо до п.2.1., якщо вичерпані - до п.2.4

2.4.Для фіксованого значення параметрів сітки $\alpha_{\text{вкл}}$ и$\alpha_{\text{искл}}$ запам’ятовуються модель з найкращим значенням зовнішнього критерію $K_{\text{РН}}$.

2.4 Зовнішній критерій

Визначимо далі доцільну форму зовнішнього критерія. Вибір кращих параметрів Stepwise здійснюється відповідно до мінімального значення

критерія селекції (зовнішнього критерію) який є суміш критерію регулярності і критерію балансу. Тоді зовнішній критерій може мати вигляд:

$$K_{\text{РН}}=(\beta-1)\left(\alpha\cdot\Delta_A^A+(1-\alpha)\Delta_B^A\right)+\beta\frac{\left|\Delta_A^A-\Delta_B^A\right|}{\left|\Delta_A^A+\Delta_B^A\right|}\quad , \tag{2.14}$$

Де:

$$\Delta_A^A=\frac{\sum_{i=1}^{N_A}(y_i-y_i^A)^2}{\sum_{i=1}^{N_A}(y_i-y^A)^2} \tag{2.15}$$

- помилка на вибірці А, розрахована на моделі, параметри якої оцінювалися на вибірці А;

$$\Delta_B^A=\frac{\sum_{i=1}^{N_B}(y_i-y_i^A)^2}{\sum_{i=1}^{N_B}(y_i-y^A)^2} \tag{2.16}$$

-помилка моделі на вибірці В, розрахована на моделі, параметри якої оцінювалися на вибірці А,

y_i,y_i -значення виходу об'єкта і моделі відповідно, y_i^A - значення моделі (параметри якої розраховані на вибірці А) в і-тій точці;

y^A , y^W - середнє табличне (об'єктне) на вибірці А і В відповідно;

α и β - коефіцієнти ваги помилки на навчальній вибірці і коефіцієнт ваги критерію балансу в загальному комбінованому зовнішньому критерії

Кінцевим критерієм якості отриманої моделі є якість прогнозування на екзаменаційній вибірці С.

2.5 Результати роботи алгоритму

Для розрахунку моделі багатовимірної лінійної регресії та порівняння якості прогнозування стандартним алгоритмом лінійної регрессії LinearRegression у програмній бібліотеці scikit-learn мови програмування

Python з запропонаваним вище версією алгоритму крокової регресії на принципах самоорганізації з оптимизацією параметрів $\alpha_{\text{викл}}$ и $\alpha_{\text{вкл}}$ були взяті біомедичні дані. Виборка складається із 86 спостережень та 18 змінних. В якості залежної змінної був взятий показник кінцево-діастолічний об'єм лівого шлуночка (КДО).

Мета задачі – отримати прогнозуючу функцію, що дає приблизне значення одного з показників серцево-судиної системи.

Порівняємо результат роботи алгоритму на стандартному алгоритмі лінійної регрессії LinearRegression у програмній бібліотеці scikit-learn з запропонаваним вище версією алгоритму крокової регресії на принципах самоорганізації з оптимизацією параметрів $\alpha_{\text{викл}}$ и $\alpha_{\text{вкл}}$. Вибірка було поділено на навчальну,перевірочну та екзаменаційну у пропорції – 75/20/5. Моделі лінійної регресії порівнюваних алгоритмів рахувались на навчальній та перевірочній виборці, а остаточна оцінка алгоритмів відбувалась на екзаменаційній виборці. В якості показника якості моделі лінійної регресії використовується коефіцієнт детермінації:

$$R=1-\frac{ESS}{TSS} \quad ,$$

де:

$$\text{ESS} = \sum_{t=1}^{n} (y_t - y_t^{ras})^2 \quad ,$$

$$\text{TSS} = \sum_{t=1}^{n} (y_t - y_{mean})^2 \quad ,$$

$$\text{Y}_{\text{mean}} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad ,$$

y_t, y_t^{ras} – фактичні та розрахункові значення залежної змінної,

y_{mean} – середнє значення залежної змінної,

ESS – сумма квадратів залишків регресії,

TSS – загальна сума квадратів.

Після використання стандартного алгоритму критерій детермінації

становив на навчальній та перевірочній вибірках становив 0,79, а на екзаменаційній – 0,69. Після застосування модифікованого алгоритму Stepwise, коефіцієнт детермінації на навчальній та перевірочній вибірках становить 0.87, а на екзаменаційній – 0,82 що демонструє покращення роботи моделі лінійної регресії після застосування нового зовнішнього критерію для підбору оптимальних параметрів алгоритму Stepwise.

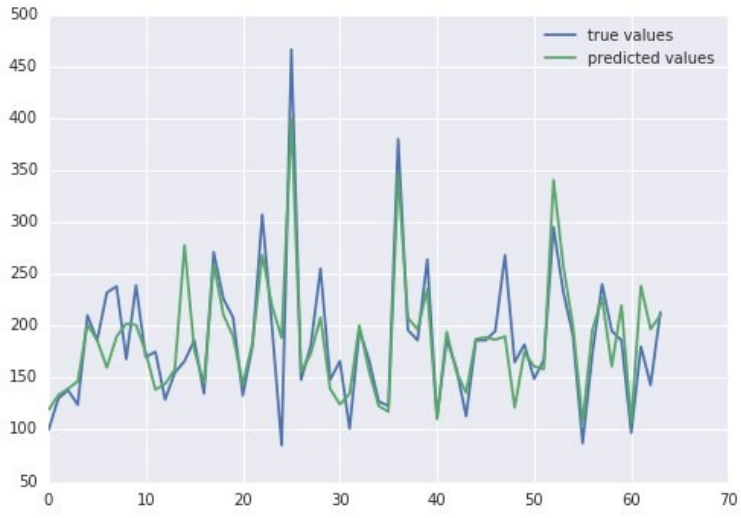

Рисунок 2.3 - Графік роботи моделі на навчальній та перевірочній виборках після використання класичного алгоритму Stepwise

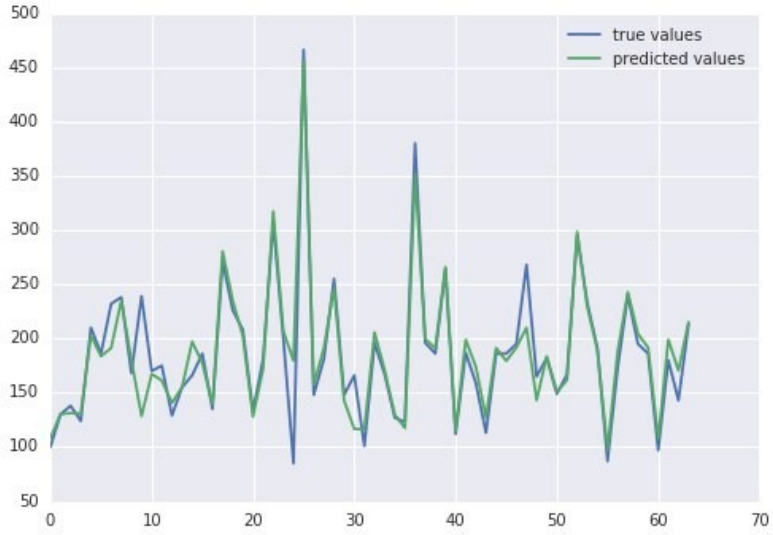

Рисунок 2.4 - Графік роботи моделі на навчальній та перевірочній виборках після використання модифікованого алгоритму Stepwise

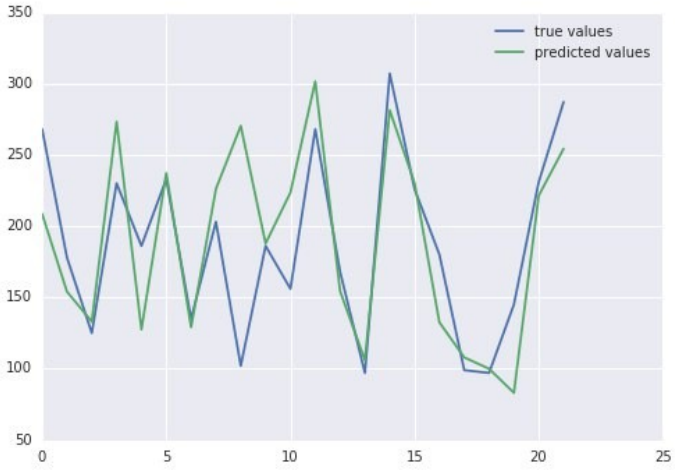

Рисунок 2.5 - Графік роботи моделі на екзаменаційній виборкці після використання класичного алгоритму Stepwise

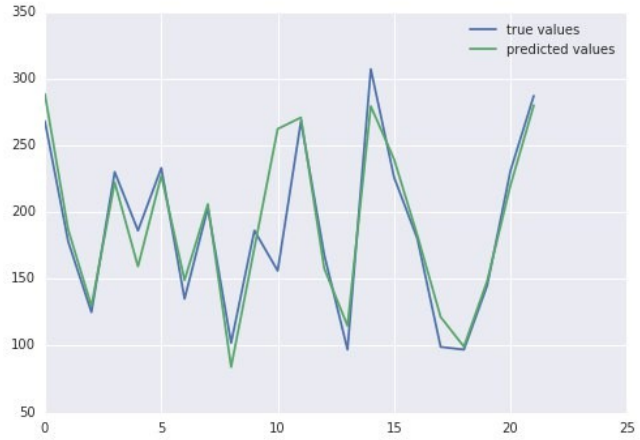

Рисунок 2.6 - Графік роботи моделі на екзаменаційній виборці після використання модифікованого алгоритму Stepwise

Модель, отримана класичним алгоритмом Stepwise:

y = -0.12*Вік + 0,14*Вага + 0,087*ППТ + 14,1*Аритмія – 2,94*ФВ +1,57*(Діаметр ЛП) + 2,2*(Корінь аорти) – 0,4*(Стать)

Модель, отримана модифікованим алгоритмом Stepwise:

$$y = -10,5*(\text{Стать}) + 0,5*(\text{Вік}) – 0,037*(\text{Вага}) + 0,83*(\text{ФВ}) + 0,84*(\text{УО}) + 0,02*(\text{Зріст}) – 1,74*(\text{Арітмія}) + 1,11*(\text{КСО}) – 4*(\text{Корінь аорти}) + 2,05*(\text{Діаметр ЛП}),$$

де ФВ – фракція викиду, діаметр ЛП – діаметр лівого передсердя, УО – ударний об’єм серця, КСО - кінцево-систолічний об'єм лівого шлуночка.

Висновки до розділу

У розділі запропоновано крокoвий алгоритм синтезу лінійної регресії на принципах самоорганіцзцаії. Для оптимізації значень параметрів алгоритму пропоновано зовнішній критерій, що відображає точність прогнозування на навчальній та перевірочній виборках. Критерій є змішаним критерієм селекції, який є суміш`ю критерія регулярності та критерія балансу. Для розглянутого прикладу прогнозування показників серцево-судинної системи у порівнянні стандартного крокового алгоритму Stepwise регресії з запропонованим у роботі алгоритмом показало покращення якості роботи моделі лінійної регресії на екзаменаційній виборці.

РОЗДІЛ 3 ПРОГРАМНА РЕАЛІЗАЦІЯ МОДИФІКОВАНОГО АЛГОРИТМУ КРОКОВОЇ РЕГРЕСІЇ

Вступ

Метою розробки є виконання програмної реалізації модифікованого крокового алгоритму багатовимірної лінійної регресії на принципах самоорганізації. Програмне забезпечення (ПЗ) дозволить визначати оптимальну структуру моделі у сенсі досягнення найкращого значення раціонально обраного зовнішнього критерію на тестовій вибірці даних при усіх можливих значеннях $\alpha_{\text{вкл}}$ та $\alpha_{\text{викл}}$.

Інформаційна технологія складається з наступних етапів:

1. Завантаження даних. Вхідні дані формуються у вигляді таблиці Excel/CSV файл, файл txt формату, розділений комами.

2. Визначення умов моделювання. Підготовка даних для аналізу.

2.1. Задання розширеної матриці змінних.

2.2. Попередні перетворення для номінальних змінних.

2.3. Задання набору предикторів, значення залежної змінної

2.4 Задання кількості ітерацій та кроку α

2.5 Задання пропорцій розбиття на вибірки

3. Рішення задачі моделювання. Проведення розрахунку алгоритму Stepwise з розрахунком зовнішнього критерію.

4. Аналіз результатів моделювання. Аналіз структур з максимальними значеннями зовнішнього критерію та вибір оптимальної для подальшої побудови на екзаменаційній вибірці. Розраховані коефіцієнти на обраній структурі заносяться до файлу Excel

3.1 Проектування програмного продукту

Контекстна діаграма

На контекстній діаграмі (рис. 3.1) зображено процес відображення біологічних об’єктів заданих множинами спостережень у вигляді просторових ліній із його вхідними даними, вихідними, механізмами та умовами здійснення.

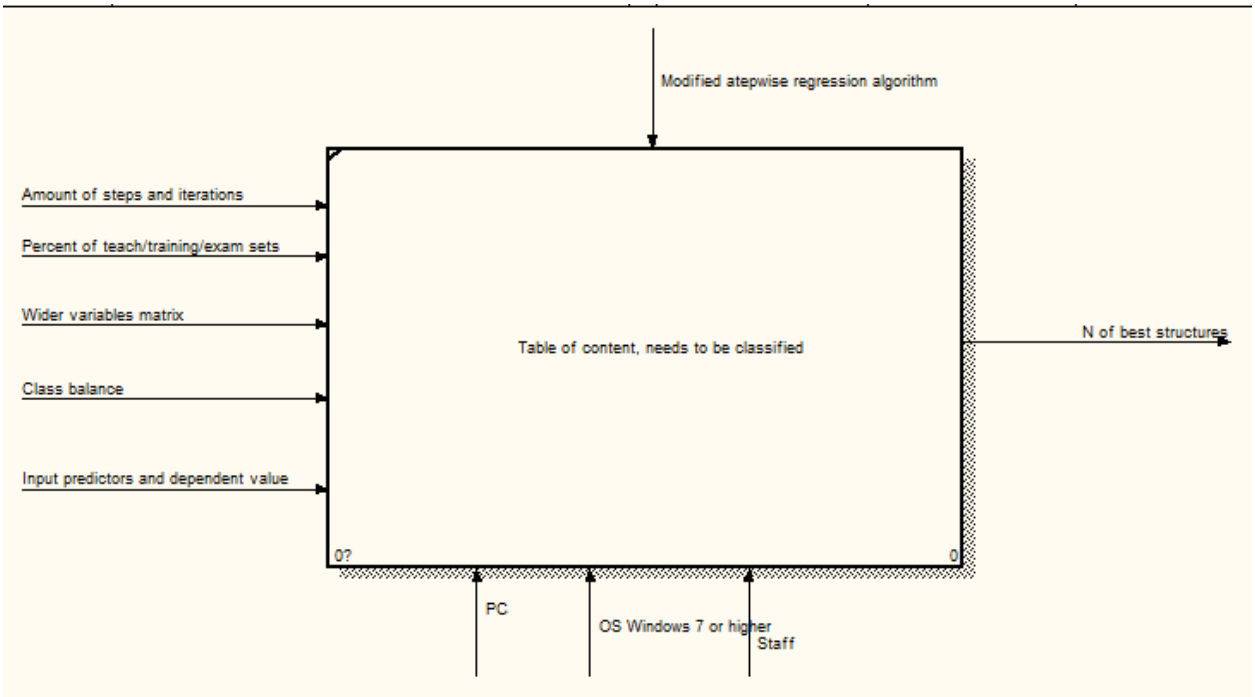

Рисунок 3.1.1 – Контекстна діаграма.

Вхідні дані включають в себе таблицю предикторів та залежної змінної, вхідні, вихідні змінні і всі параметри, що вимагаються від користувача для реалізації задачі. Вихідні дані – шукані структури. Для розв’язання поставленої задачі було обрано модифікований алгоритм крокової регресії на принципах самоорганізації. Умови включають виконання технічних вимог і наявність необхідних працівників.

Діаграма декомпозиції першого рівня (методологія IDEF0)

Для наочного відображення підпроцесів, що включає в себе процес відображення об’єктів заданих множинами спостережень зручно використовувати діаграму декомпозиції (рис. 3.2).

Процес умовно поділено на шість підпроцесів, серед яких: зчитування даних; визначення вхідних і вихідних змінних; формування розширеної матриці змінних, балансування спостережень у класах, задання кількості ітерацій та кроку, поділ на навчальну, перевірочну та екзаменаційну вибірки; виконання модифікованого алгоритму крокової регресії на принципах самоорганізації; виведення і збереження результатів.

Діаграма декомпозиції другого рівня (методологія IDEF0)

Поетапне виконання алгоритму відображено на діаграмі декомпозиції другого рівня (рис. 3.3). Для виконання алгоритму формується сітка альфа включень, виключень і відповідно до неї відбувається розрахунок критерію Фішера для кожної моделі, кожна з яких є ускладненим варіантом попередньої моделі. Для таких моделей розраховується зовнішній критерій.Для кожного значення альфа відбувається розрахунок моделі, розрахунок закінчується, коли перебрано усі можливі комбінації альфа, після цього відбирається модель з кращим значення зовнішнього критерія.

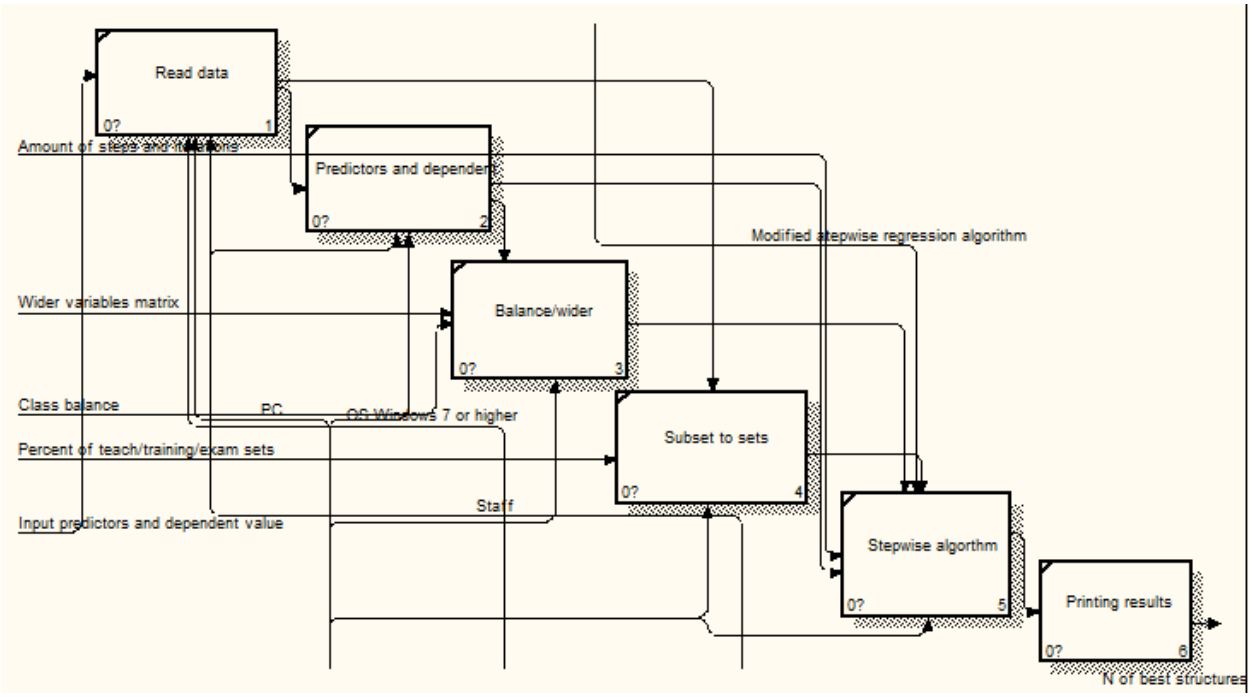

Рисунок 3.2 –Діаграма декомпозиції першого рівня.

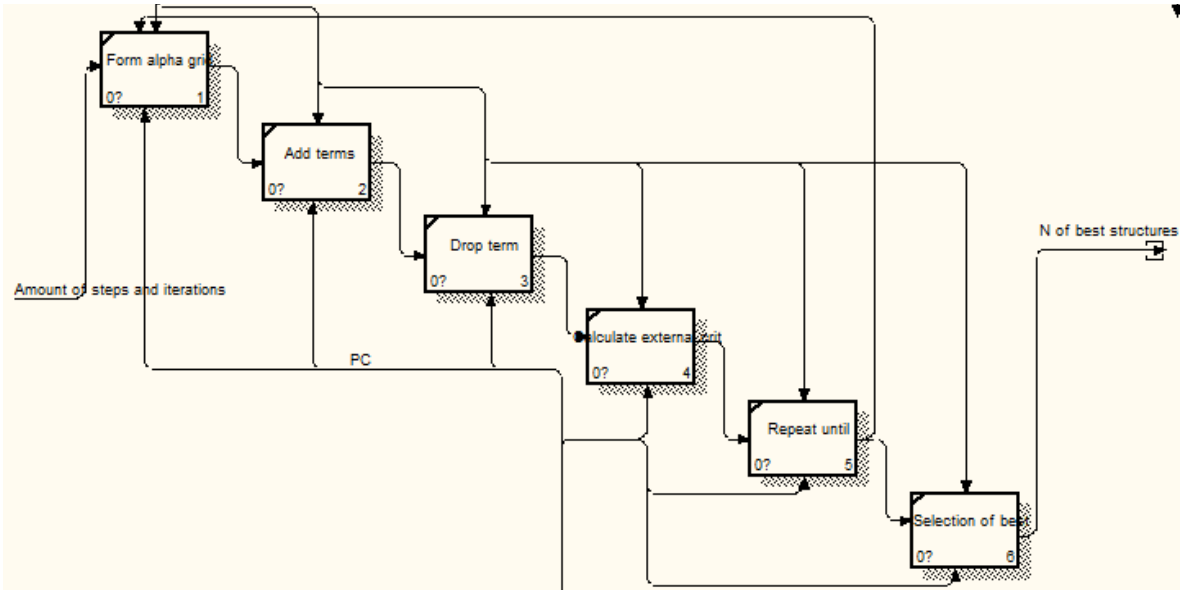

Рисунок 3.3 Діаграма декомпозиції другого рівня

Діаграма декомпозиції (методологія IDEF3)

З діаграми декомпозиції (рис. 3.4) видно послідовність виконання усіх процесів і вхідні дані, що необхідні для виконання цих процесів. Виконання задачі починається зі зчитування даних і визначення предикторів і незалежної змінної. Далі виконується розширення вхідної множини, задання кількості ітерацій, поділ на навчальну та перевірочну вибірки – попередня обробка даних. Необхідну кількість разів виконується основна частина алгоритму: генерація моделей, розрахунок коефіцієнтів моделей, визначення значення зовнішнього критерію і відбір моделей. Останнім етапом у виконанні алгоритму є виведення результатів і збереження їх у файл.

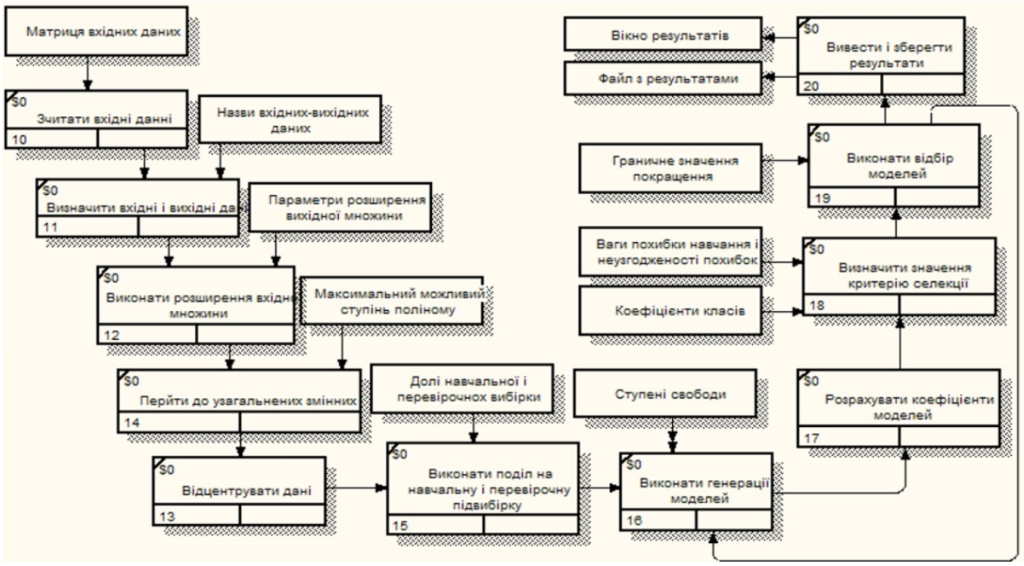

Рисунок 3.4 – Діаграма декомпозиції

Модель варіантів використання (Use Case*)*

Діаграма варіантів використання (Use Case) відображає процеси, що відбуваються при виконанні алгоритму з точки зору акторів (рис. 3.5). Оскільки учасником виконання алгоритму є тільки один актор – користувач, то всі процеси, що відбуваються, прив’язані тільки до нього.

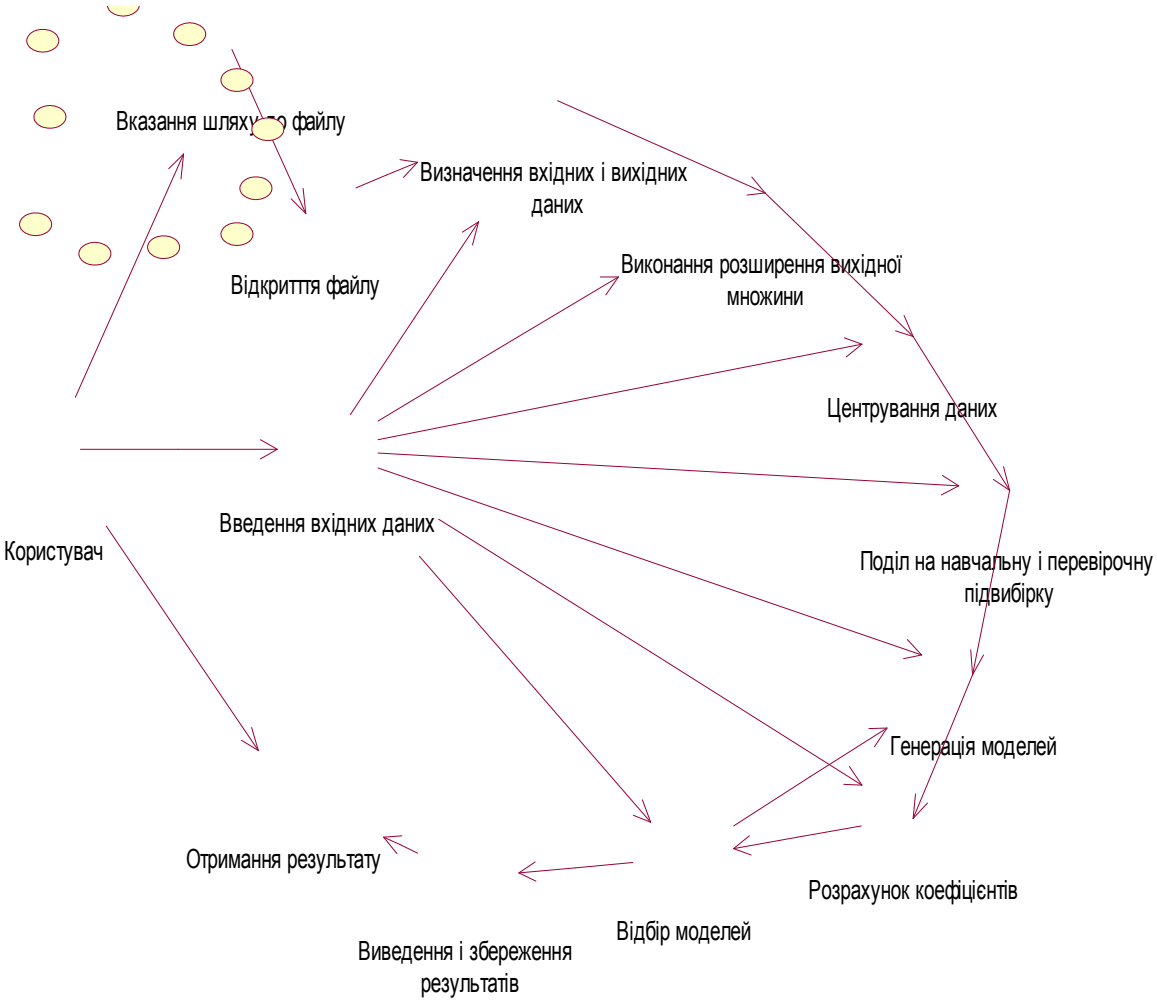

Рисунок 3.5 – Діаграма Use Case.

Діаграма станів

На діаграмі станів (рис.3.6) відображено зв’язки і послідовність проходження всіх станів, від початку роботи алгоритму і до кінця.

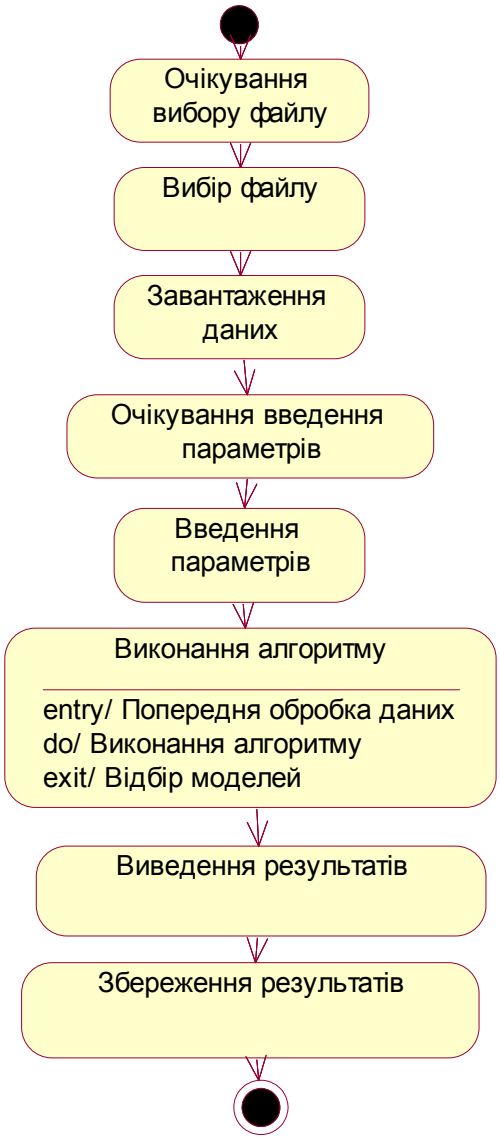

Рисунок 3.6 – Діаграма станів.

Діаграма класів аналізу

Діаграма класів (рис. 3.7) аналізу встановлює залежності між усіма можливими класами (граничними, керуючими і сутностями) і всі можливі види зв’язків між ними – асоціація, агрегація, композиція, узагальнення, залежність. Граничними класами є головне вікно і вікно результатів програми, а також зв’язані із ними зв’язками агрегації кнопки, поля для вводу і таблиці. До класів-сутностей відносяться всі вхідні данні, параметри алгоритму і файл з результатами. До керуючих класів відносяться зв'язок з файлом і усі подальші процеси обробки даних.

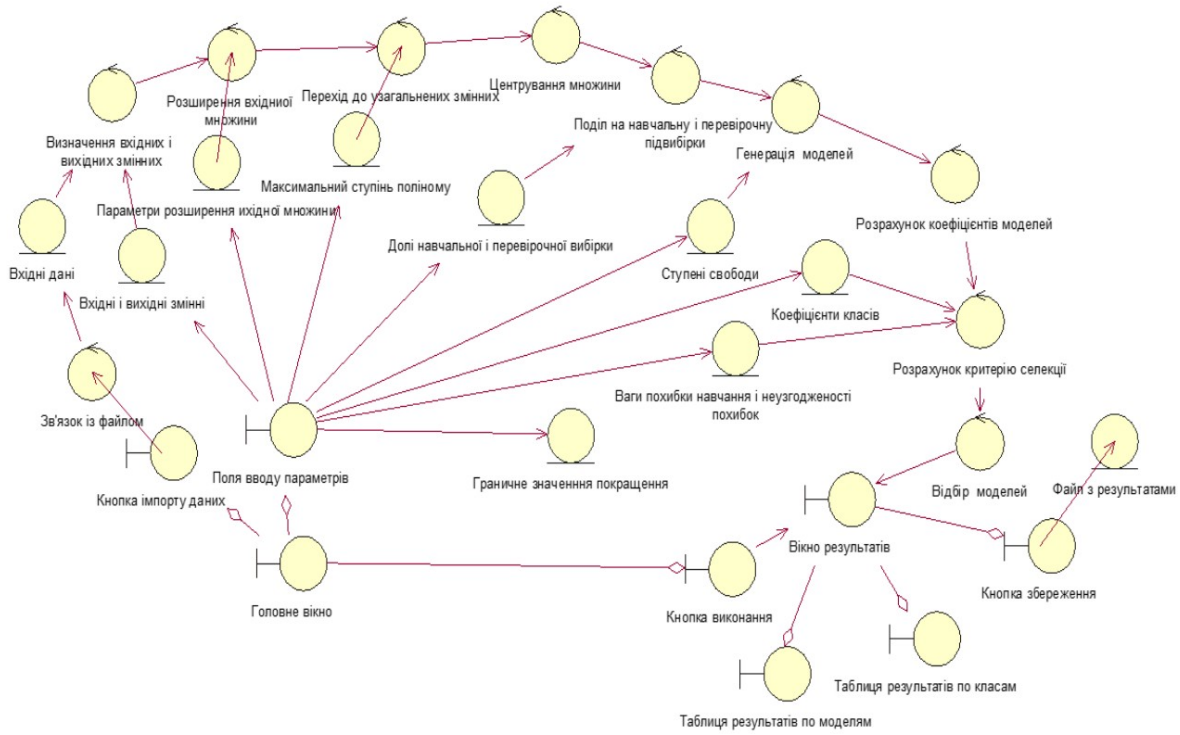

Рисунок 3.7 – Діаграма класів аналізу.

Діаграма кооперації

З діаграми кооперації (рис.3.8) видно, що система взаємодіє з користувачем і з двома файлами – з вхідними і вихідними даними. Користувач запускає програму і вказує шлях до файлу, відбувається імпорт даних і користувач вводить необхідні параметри алгоритму. Після виконання алгоритму система виводить на екран результати роботи та при необхідності зберігає результат в файл.

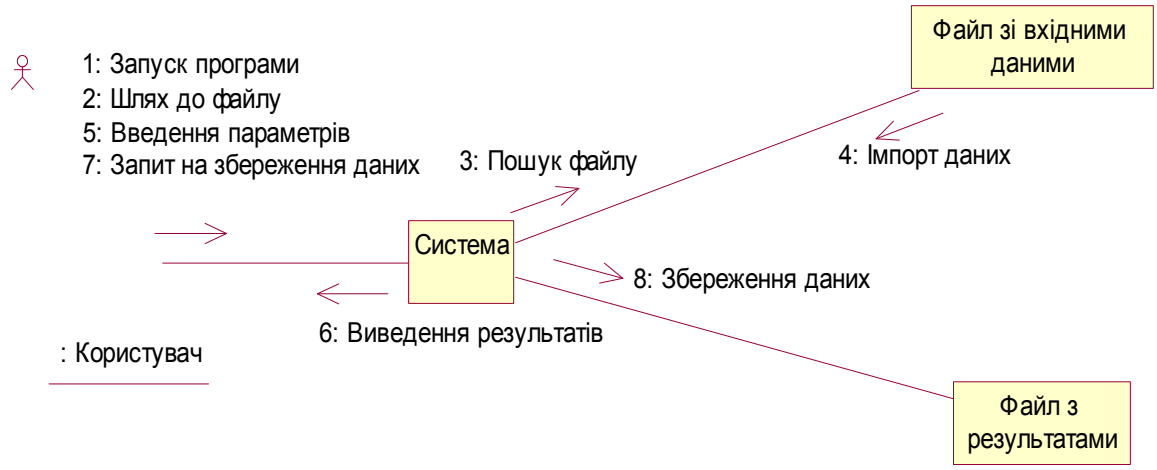

Рисунок 3.8 – Діаграма кооперації.

Діаграма послідовності.

Діаграма послідовності (рис.3.9) показує процеси, що описані в діаграмі кооперації з урахуванням їх послідовності в часі.

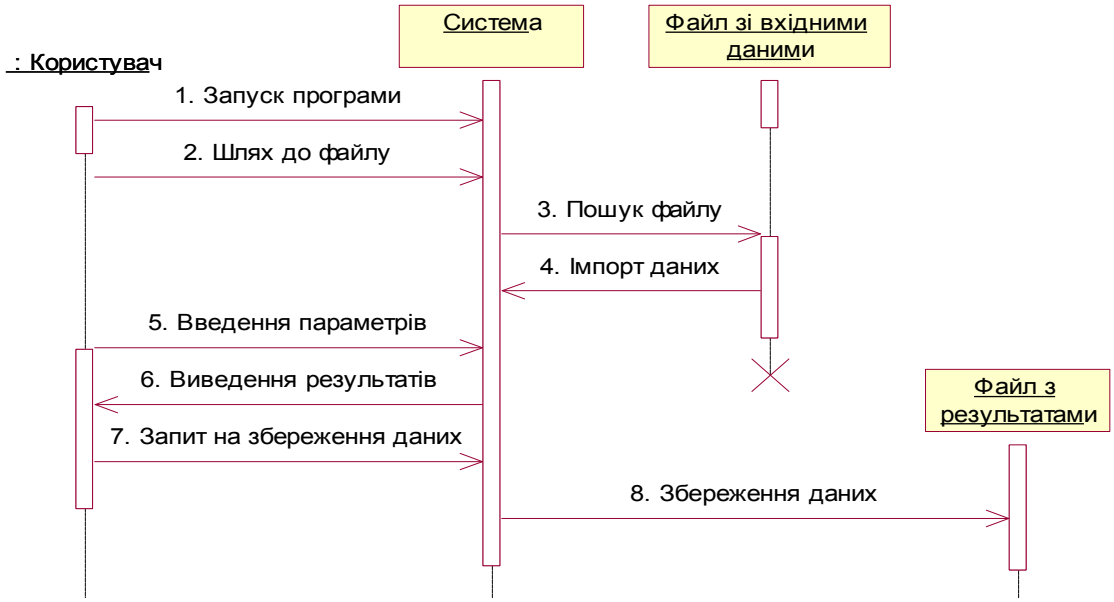

Рисунок 3.9 – Діаграма послідовності.

3.2 Вибір середовища розробки

Для реалізації алгоритму була обрана мова програмування Python: Python — мова програмування, яка є дуже універсальною для вирішеня будь-яких задач -від веб-розробки до обчислення великих об‘ємів даних.

Python розповсюджується безкоштовно за ліцензією Python Software Foundation License у вигляді вільнодоступого вихідного коду або відкомпільованих бінарних версій більшості операційних систем: Linux, FreeBSD, Microsoft Windows, Mac OS X, Solaris.

Python має значні можливості для здійснення статистичних аналізів, включаючи лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. Python легко розбудовується завдяки використанню додаткових функцій і пакетів доступних на сайті PyPl. Існує дуже велика кількість різноманітних бібліотек, які підійдуть для вирішеня будь-яких проблем та задач, які стоять

перед розробниками, і в той же час є дуже легким для освоєння.

Переваги Python:

• безкоштовна загальна ліцензія GNU General Public License

• Python – це потужна мова програмуваня, у якій реалізовані багато бібліотек, дозволяючи застосовувати будь-які способи аналізу даних;

• Отримання даних з різних джерел у пригодному для використання вигляді, також імпорт текстових файлів та табличних даних

• Python являє собою без аналогічну потужну платформу для написання статистичних програм

• Python працює на різних ОС - Windows, Unix и Mac OS X.

Враховуючи вищезазначене Python є дуже гарним вибором для реалізації даного алгоритму.

3.3 Реалізація програмного продукту

Розроблений програмний продукт реалізовано за допомогою мови Python з використанням Spyder.

Розглянемо детально реалізацію деяких методів.

В створеному програмному продукті дані зчитуються з Excel файлу і записуються у елемент управління pandas.DataFrame при натисканні кнопки «Завантажити дані». Програма виконує підрахунок кількості заповнених рядків і стовпчиків в файлі. Далі визначається кількість змінних в базі даних, кількість спостережень. Дані виводяться на екран за допомогою бібліотеки для графічного інтерфейсу Tkinter в зручному для користувача вигляді.

Для реалізації поділу вибірки на навчальну та перевірочну підвибірки було застосовано метод train_test_split бібліотеки scikit-learn. Вхідними аргументами є матриця спостережень та доля навчальної, перевірочної вибірки. Метод визначає кількість точок, для даного об’єкту яка має потрапити до навчальної вибірки. Генерується задана кількість випадкових точок у інтервалі від 1 до кількості спостережень і відбирає по цим точкам з

таблицю спостереження, інші ж спостереження залишаються для перевірочної вибірки.

Для подальшого створення матриці розширенних змінних. Метод degree приймає аргументи: кількість змінних і ступені у які треба піднести, а також перетворює категоріальні та порядкові дані до необхідного формату.

Основна частина модифікованого крокового алгоритму логістичної регресії на принципах самоорганізації описана у методі stepwise_modified. Він повертає рядок типу DataFrame, що містить значення зовнішнього критерія, оптимальну структуру, значення альфа включення/виключення.

На вхід алгоритму подаються наступні параметри: data, test_size, alpha_include, alpha_exclude, A, B.

Параметр data визначає список предикторів, з яких мають бути визначені найвпливовіші.Під час роботи алгоритму кожен з предикторів у списку по черзі додається/видаляється до моделі.

Параметр test_size передає співвідношення навчальної до тестової вибірок.

Параметри alpha_include, alpha_exclude – задають стартові значення альфа включення, альфа виключення, в межах котрих відбувається пошук нових кращих значень. По їх значенням відбувається крокова процедура.

Параметри A, B - відповідають параметрам для розрахунку зовнішнього критерія і відповідають за баланс регулярності та балансу.

На всіх рівнях алгоритму крім першого нові моделі «доформовуються» до вже існуючих, після чого відбувається перерахунок зовнішнього критерію Максимальне значення зовнішнього критерію – показник якості структури.

Вхідні дані при роботі з розробленою програмою – це таблиця змінних-спостережень

Висновки до розділу 3

Програмно реалізовано модифікований кроковий алгоритм Stepwise на принципах самоорганізації з автоматичним підбором кращих параметрів, що

дозволив визначати оптимальну структуру моделі у сенсі досягнення накращого значення раціонально обраного зовнішнього критерію.

4 ОХОРОНА ПРАЦІ

Вступ

У роботі був розглянутий програмно-апаратний комплекс для обробки даних. В якості апаратної частини комплексу виступають комп'ютерний електрокардіограф та персональний комп'ютер з необхідним програмним забезпеченням. Тому на даному етапі роботи розглядаються основні положення охорони праці та безпеки лабораторії функціональних резервів організму людини кафедри фізичного виховання ФБМІ НТУУ «КПІ», в якому даний програмно-апаратний комплекс може бути безпосередньо використаний.

4.1 Характеристика робочого місця та умови експлуатації розробленого програмно-апаратного комплексу

Розглянемо лабораторію функціональних резервів організму людини (рис. 4.1), призначену для зняття антропометричних та функціональних показників тіла. Дані отримуються за допомогою електрокардіографа та у подальшому оброблюються на комп'ютері.

Таблиця 4.2.1 – Характеристики лабораторії функціональних резервів організму людини

Назва	Параметри
Розмір кабінету, $l^x w^x h$	6 х3,35х3 (м)
Площа кабінету	20,1 $м^2$
Об´єм кабінету	60,3 $м^3$
Кількість працюючих	2
Вентиляція	2 вентиляційні шахти
Кондиціонування	Кондиціонер типу "спліт"
Опалення	3 7-и секторні батареї
Підлога	бетон, вкритий лінолеумом
Стіни	Цегляні, силікатна фарба.
Природнє освітлення	1 вікно, що виходять на північ. Розмір: 1×1,5м. Скло віконне листове, подвійне.
Штучне освітлення	2 світильники на 4 лампи Л 201Б 40Вт Розмір 1275×127×675

В таблиці 4.2.2 наведено перелік обладнання, яким оснащена лабораторія

Таблиця 4.2.2 – Характеристики техніки/обладнання та меблів

№	Назва об’єкту	Кількість	Розміри(*a*×*h*×*b (мм)*) та характеристики та об’єкту
8.	Електрокардіограф	1	250×150×65 ЮКАРД-100, USB, Dial-Up модем, GSM модем, 15 Вт, 220 В, 100 Гц
9.	Комп’ютер з монітором	1	629×571×790 Багатопроцесорний комп'ютер с ACPI, Acer P196HQVB, DualCore Intel Pentium E5400, 2700 MHz, ASRock G31M-VS, NVIDIA GeForce GT 220 (1024 Мб), 220-240 В, 50/60 Гц
1.	Стіл лікаря	1	1132×630×750-
2.	Стіл для комп’ютера	1	1132×630×750
3.	Стіл-тумбочка	1	400×440×667
4.	Крісло робоче	2	530×570×850
4.	Кушетка	1	1970×670×520
6.	Стілець напівм’який	2	430×470×770
7.	Стіл для ЕКГ	1	445×780×930
10.	Шафа для докумнтів	1	773×410×1985
11.	Відро для сміття	1	372×315×360
12.	Умивальник	1	550×420×150
13.	Вішалка	1	600×120×100
14.	Світильник	1	1275×127×675

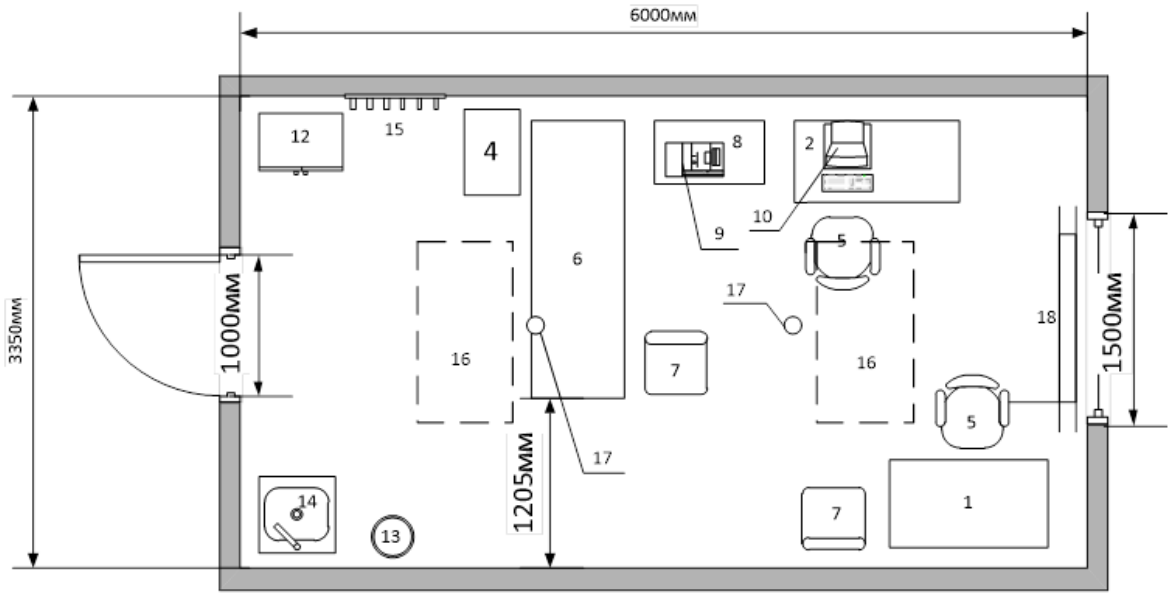

Рисунок 4.1. – План лабораторії функціональних резервів організму людини

Розрахунки:

$$
\begin{gathered}
S_{\text{меб.}}=\sum_{n=1}^{16} S_n \times N_n = 5{,}74\ м^2 \\
S_{\text{віл.}}=S_{к.}-S_{\text{меб.}}=14{,}36\ м^2 \\
S_1=S_{\text{віл.}}/N_{л.}=7{,}18\ м^2 \\
V_{\text{меб.}}=\sum_{n=1}^{16} V_n \times N_n = 4{,}40\ м^3 \\
V_{\text{віл.}}=V_к-V_{\text{меб.}}=55{,}9\ м^3 \\
V_1=V_{\text{віл.}}/N_{л.}=27{,}95\ м^3
\end{gathered}
\qquad (4.1)
$$

де $S_{меб.}$ - площа меблів; $S_{вільн.}$ – вільна площа; S_1 – площа на одну людину; $N_{л.}$ – кількість людей; $V_{меб.}$ – об’єм меблів; $V_{вільн.}$ – вільний об’єм; V_1 – об’єм на одну людину.

Порівняння нормативних параметрів з реальними згідно з нормами, ДСанПіН 3.3.2.007-98 (Табл. 4.2.3) .

Таблиця 4.2.3 – Порівняння розрахованих даних з нормою

Параметр	Нормативні параметри	Реальні параметри
Площа, S/людину, м^2	Не менше 6 м^2	7,18 м^2
Об’єм, V/людину, м3	Не менше 20 м^3	27,95 м^3
Висота приміщення, м	3 – 3,5 м	3 м
Відстань від робочих місць до стін, м	Не менше 1 м	1,2 м
Габарити дверного отвору, м	12.1 м	12 м
Габарити вікна	1.22 м	1.51.8 м

Висновок: Фактичні значення площі та об’єму відповідають встановленим нормам (табл. 4.3.).

4.2 Оцінка небезпечних і шкідливих виробничих факторів

Таблиця 4.2.1 Небезпечні та шкідливі виробничі фактори.

Фактори	Причини виникнення
Фізичні	несприятливі мікрокліматичні умови; шум випромінювання при роботі з обчислювальною технікою.
Хімічні	Спирт метиловий, спирт етиловий
Біологічні	Віруси, бактерії

Небезпека враження ел. струмом	випадковий дотик до частин, що проводять струм та знаходяться під напругою; поява напруги в результаті помилкового вмикання, замикання або наведення напруги сусідніми установками.
Небезпека пожежі	Коротке замикання, іскріння в контактах, тривалі перевантаження.

4.2.1 Мікроклімат

Джерела впливу на параметри мікроклімату та можливі наслідки розглянемо в таблиці 4.2.1.1

Таблиця 4.2.1.1 – Основні джерела впливу на мікрокліматичні умови

Параметри мікроклімату	Джерела	Наслідки
Відносна вологість	- Атмосферне повітря; - Повітря, видихуване людьми	- Проблеми з диханням; - Риніт; - Пересихання шкіри та губ; - Алергічні реакції; - Втома; - Розвиток бактерій, вірусів, грибків;
Температура повітря t	- Люди; - Сонячна радіація; - Система штучного опалення; - Обладнання	- Зниження працездатності та продуктивності праці; - Послаблення організму; - Підвищення небезпеки травмування;
Швидкість повітря	- Вентиляція	- Протяги

Процедура зняття ЕКГ та обробка отриманих даних виконуються сидячи , стоячи, пов’язана з ходінням і не потребують великого фізичного навантаження, тому ці роботи відносяться до категорії «Легкі роботи – 1б».

В таблиці 4.2.1.2 наведені оптимальні і допустимі параметри мікроклімату. Всі ці параметри зазначені для «Легкої – 1б» категорії роботи холодного та теплого періоду року за ДСН 3.3.6.042-99. Для заміру параметрів мікроклімату в кабінеті використовується універсальний термометр, гігрометр TFA 30.3049, анемометр Wigam 8908 для вимірювання швидкості повітря.

Таблиця 4.2.1.2 – Реальні та оптимальні значення мікрокліматичних умов

Параметри мікроклімату	Реальні значення		Нормативні значення	
	Холодний період	Теплий період	Холодний період	Теплий період
Температура повітря t, ºС	21-23	22-24	22-24	23-25
Відносна вологість, %	60-40	60-40	75	40-60
Швидкість повітря, м/с	0,1	0,1	0,2	0,1-0,3

Приміщення регулярно провітрюється за допомогою одного вікна, яке відкривається (1,5 м). Розрахунок природної вентиляції (Кр=4 для кабінету)

$L=V_{пом}*K = 111,8$ ($м^3/ч$), $V_{лаб}=27,95$ $м^3$

Таблиця 4.2.1.3– Повітрообіг лабораторії функціональних резервів

Норма на одну людину	30
Реальне значення на одну людину	55,9

Таблиця 4.2.1.4 – Засоби та захисти, що застосовуються

Опалення	Батерея на 3*7 секцій
Вентиляція	Вентиляція з кратністю обміну 1:3

Виходячи з проведених замірів та порівняння з нормативними значеннями можемо скласти список заходів з охорони праці та безпеки в умовах надзвичайних ситуацій на робочих місцях при експлуатації об'єкту (табл. 4.2.1.5).

Таблиця 4.2.1.5 – Заходи з нормалізації мікроклімату

Заходи		Холодний	Теплий
Технологічні	В обладнанні	Масляний радіатор AEG RA 5520	Кондиціонер типу «спліт» SAMSUNG AQ 12XLN/X;
	В приміщенні	Вентиляція з кратністю обміну 1:3	Вентиляція з кратністю обміну 1:3
Організаційні		Раціоналізації режимів праці й відпочинку, перерви;	Раціоналізації режимів праці й відпочинку, перерви;
		Вологе прибирання після закінчення робочого дня;	Провітрювання; Вологе прибирання після закінчення робочого дня;
		Утеплення вікон, склопакети;	Улаштування жалюзі.

Індивідуальні	Використання спецодягу	Використання спецодягу; Спеціальний питний режим (забезпечення білково-вітамінними напоями, хлібним квасом, підсоленою водою);

Висновок: Всі параметри мікроклімату, що потребує дана категорія робіт дотримуються в повній мірі завдяки системам нормалізації мікроклімату.

4.2.2 Шум

Шум у лабораторії є постійним. Джерела шуму подані у таблиці 4.8.

Таблиця 4.2.2.1 – Реальні та нормативні значення для звуку та шуму

Джерела	Наслідки
ПК (NVIDIA GeForce GT 220)	Головні болі, мігрені, зниження уваги
Зовнішній шум	
Розмова	

В таблиці 4.2.2.2 наведені нормативні і реальні параметри рівня шуму, що визначені встановленою нормою за ДСН 3.3.6.037-99 «Санітарні норми виробничого шуму, ультразвуку та інфразвуку». Для вимірювання рівню шуму встановлено шумомір UT-351.

Таблиця 4.2.2.2 – Реальні та нормативні значення для звуку та шуму

Параметри шуму	Реальні значення	Нормативні значення
ПК (системний блок)	8дБа	
Зовнішній шум	40 дБа	50 дБа
Розмова	30 дБа	

Виходячи з проведених замірів та порівняння з нормативними значеннями можемо скласти список заходів з охорони праці (табл. 4.2.3.3)

Таблиця 4.2.2.3– Засоби та заходи захисту від шумових навантажень

Технічні	В приміщені	Акустична ізоляція (звукопоглинальна піна всередині корпусу)
		Встановлені металопластикові вікна
	В	Відеокарта з пасивним охолодженням;

	обладнанні	
Організаційні	Дотримання правил технічної експлуатації;	
	Проведення планово-попереджувальних оглядів та ремонтів;	
	Попередній та плановий медогляд;	
Засоби індивідуального захисту	Не передбачено	

4.2.3 Випромінювання

У лабораторії присутнє незначне електромагнітне випромінювання, оскільки монітор ПК вироблений на основі рідкокристалічної матриці, що не має сильного електромагнітного випромінювання. Інфрачервоні та ультрафіолетові випромінювання відсутні. У табл. 4.2.4.1 наведені заходи з охорони праці та безпеки та дії в надзвичайних ситуаціях.

Таблиця 4.2.3.1 – Заходи безпеки та дії в надзвичайних ситуаціях

Технологічні	Раціональне розміщення робочих місць
	Монітор на основі рідкокристалічної матриці
Організаційні	Здійснювати контроль випромінювання один раз на 6-12 місяців;
Засоби індивідуального захисту	Не передбачено

4.2.4 Небезпека ураження людини електричним струмом

У приміщенні використовується однофазна мережа змінного струму напругою 220 В, та частотою 50 Гц. Електрокардіограф працює від універсального джерела живлення. Його підключення до комп’ютеру забезпечується за допомогою вузла гальванічної розв’язки, що унеможливлює враження струмом. Джерела небезпеки, параметри споживання та можливі наслідки наведені в таблиці 4.2.4.1

Таблиця 4.2.4.1 – Джерела електричної небезпеки

Джерела	Сила струму	Наслідки
ПК	1.2 А	Легке тремтіння пальців рук
Штучне освітлення	0,3 А	Легке тремтіння пальців рук

Монітор	0,25 А	Відсутні
Кардіограф	0,05 А	Відсутні

Клас небезпеки приміщення відповідно до ДНАОП 0.00-1.21-98 та ДНАОП 1.1.10-1.07-01 описані в таблиці 4.3.4.2

Таблиця 4.2.4.2 – Електричні параметри приміщення

Параметр	Значення
Мережа	Змінний струм напругою 220 В, частота 50 Гц, сила струму від 0.1 до 1,2 А (не допускає перебування людини в дії струму більше 0.1 секунди).
Обладнання	ПК: номінальна напруга 220В
Клас небезпеки приміщення	Електрокардіограф - класс II, тип CF, Acer P196HQVB - класс II

Таблиця 4.2.4.3 – Оцінка критеріїв захисту від ураження електричним струмом [ДСТУ 3798-98. Вироби медичні електричні]

Критерій		Пристрій	ГДР
Струм витоку на землю	При нормальних умовах	0.35 мА	0.5 мА
	В умовах одиничного порушення	0.7 мА	1 мА
Струм витоку на пацієнта	При нормальних умовах	0.075 мА	0.1 мА
	В умовах одиничного порушення	0.3 мА	0.5 мА

Щоб оцінити, чи небезпечний дотик людини до струмоведучих частин, необхідно знати струм, що протікає через людину $I_в$ і порівняти значення з допустимими.

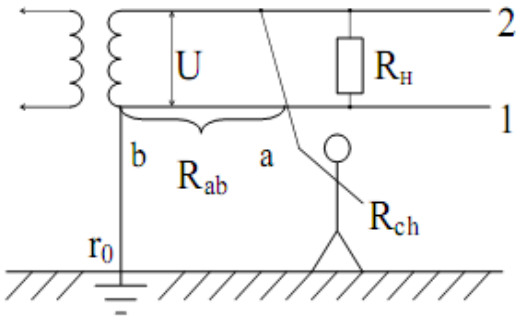

Рисунок 4.2 – Схема включення людини в електричну мережу

Струм що проходить через людину при однополюсному дотику розраховується за формулою:

$$I_в = \frac{U}{2\left(R_ч + R_п + R_{взут}\right) + R_{із}} \quad (4.2)$$

де U – напруга мережі;

$R_ч$ – опір в ланцюзі людини (1 кОм);

$R_п$ – опір підлоги;

$R_{взут}$ – опір взуття;

$R_{із}$ – опір ізоляції відносно землі.

Розрахунок струму, що проходить через людину при двополюсному дотику:

$$I_в = \frac{U}{R_ч} = \frac{220}{10^3} \approx 220\text{мА} \quad (4.3)$$

Даний струм є набагато більше фібриляційного, а це означає, що через людину пройде смертельно небезпечний струм. Щоб уникнути фатальної дії струму передбачені наступні заходи: гумова ізоляція струмоведучих частин, покриття підлоги антистатичним лінолеумом, змінне гумове взуття та інші (таблиця 4.2.4.4).

Таблиця 4.2.4.4 – Заходи по електробезпеці

Технологічні	Нормальний режим	Недопустимість струмовивідних частин;
		Прилади вмикаються в розетку через спеціальні розетки з заземленням;
		Підлога покривається антистатичним лінолеумом (див. розрахунок);
		Мережевий шнур вимірювальної апаратури має гумову ізоляцію, є трижильним і має вилку із заземлюючими контактами (див. розрахунок);
		Прокладання прихованої проводки в поглибленнях підлоги;

		Світильники на висоті 2,8м (не менше 2,5м);
	Аварійний режим	Для заземлення застосовуються спеціальні окремі проводи по всьому контуру стін;
		Пробивний автоматичний запобіжник ВПБ 6-10;
		Занулення однофазної електроустановки;
Організаційні	Інструктаж з електробезпеки;	
	Створення таких умов, щоб працівники, на яких покладено обов'язки з обслуговування електроустановок, відповідно до чинних вимог, своєчасно здійснювали їх огляд та профілактичні дії.	
Засоби індивідуального захисту	Не передбачено	

Вимоги ДНАОП 0.00-1.21-98 та ДНАОП 1.1.10-1.07-01 щодо електробезпеки виконані в повній мірі.

4.2.5 Пожежна безпека в умовах надзвичайної ситуації

У лабораторії функціональних резервів організму людини джерелами займання можуть бути: волокнисті (папір), тверді (столи, стільці, шафи та інше), пластикові (корпуса принтера, монітору, клавіатури, електрокардіографу) матеріали. Згідно нормативного документу ГОСТ 12.1.004-91, дане приміщення відноситься до (табл.4.3.6.1)

Таблиця 4.2.5.1 – Параметри пожежної безпеки

Клас	А (супроводжується тлінням) та Е (горіння електроприладів)
Категорія	В (тверді горючі речовини, які здатні горіти)
Зона	П-IIа (простір у приміщенні, в якому знаходяться тверді горючі речовини)

Щоб запобігти пожежі в лабораторії прийнято такі міри протипожежної безпеки:

Таблиця 4.2.5.2 – Заходи по пожежній безпеці

Технологічні	В обладнанні	Корпуса обладнання виготовлені з важкогорючих матеріалів

	В приміщенні	У загальному коридорі, поруч з кабінетом, знаходиться 2 вогнегасники (маса заряду 4 кг, пожежогасна речовина – порошок, довжина струменю 3,5 м, маса 25 кг)
Технологічні	В приміщенні	Передбачено вільний доступ до мережних рубильників та вимикачів;
		У приміщенні на стелі встановлено два датчика СПД 3,5 розрахованих на 20м2.
Організаційні		Інструктаж з пожежної безпеки;
		Організація навчань з пожежної охорони.
Засоби індивідуального захисту		Не передбачено

В приміщені встановлено вогнегасник типу ВП-5 та сповіщувачі типу ДТЛ (табл. 4.2.6.2)

Таблиця 4.2.5.3 – Характеристика вогнегасника та сповіщувачів

Тип вогнегасника	ВП-5
Категорія приміщення	В
Гранична захищувана площа, м2	400
Клас пожежі	А
Місткість вогнегасника, л	5
Тип сповіщувача	ДТЛ
Темп. спрацювання, C^0	72
Інерційність спрацювання, с	120
Захищувана площа, $м^2$	15
Висота приміщення, м	До 3,5
Площа, що контролюється, м2	До 25
Максимальна відстань між сповіщувачами, м	5,0
Максимальна відстань від сповіщувача до стіни, м	2,5

План евакуації з поверху при ймовірному виникненні пожежі представлений на рисунку 4.3.

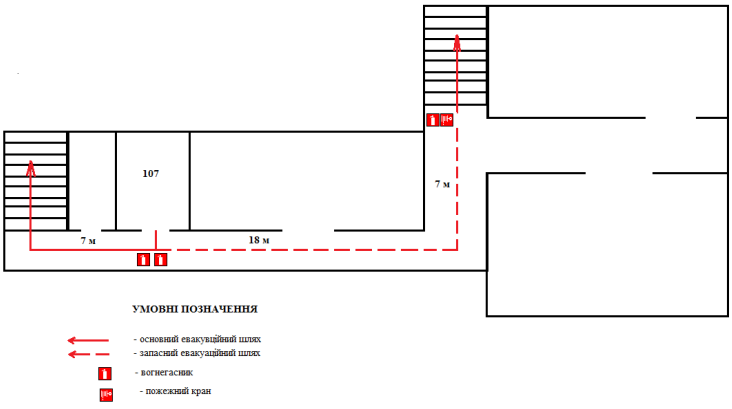

Рисунок 4.3 – План евакуації

Таблиця 4.2.5.4 –Дані розрахунку параметрів евакуації

Найменування параметрів	Значення
Гранична відстань від лабораторії до сходів по основному шляху евакуації	7 м
Гранична відстань від лабораторії до сходів по основному шляху евакуації	25 м
Розрахунковий час евакуації по основному шляху	0,25 хв
Розрахунковий час евакуації по запасному шляху	0,8 хв

Висновок: У приміщенні виконуються усі вимоги з пожежної безпеки відповідно до вимог НАПБ А.0.001-95 “Правила пожежної безпеки в Україні”.

Висновки до розділу 4

У даному розділі були розглянуті норми та заходи з охорони праці й техніки безпеки, які будуть направлені на усунення потенційно шкідливих і небезпечних виробничих факторів. Мікрокліматичні параметри даної лабораторії відповідають встановленим нормам. Для підтримання відповідного рівня показників використовують необхідні засоби та здійснюють заходи їх контролю. Рівень шуму в лабораторії знаходиться в допустимих нормативних межах. Також в даному розділі розглянуто заходи безпеки від ураження електричним струмом. Розглянуті засоби та заходи, щодо захисту працюючого персоналу від пожежі. Всі параметри відповідають нормам.

ЗАГАЛЬНІ ВИСНОВКИ

У роботі було розглянуто та проаналізовано теоретичні відомості щодо математичних методів, що вирішують задачу структурно-параметричного синтезу та зовнішніх критеріїв для оцінювання отриманих структур. Особлива увага приділялася методу групового урахування аргументів та багатовимірній лінійній регресії. Було визначено переваги та недоліки цього методу. Для подальшої роботи було обрано оптимізувати метод крокової регресії Stepwise.

Було запропоновано кроковий алгоритм синтезу лінійної регресії на принципах самоорганізації. Для оптимізації значень параметрів алгоритму запропоновано зовнішній критерій, що відображає точність прогнозування на навчальній та перевірочній виборках. Для прикладу було обрано задачу прогнозування показників серцевосудинної системи. Порівняння стандартного крокового алгоритму Stepwise регресії з запропонованим у роботі продемострувало збільшення коефіцієнта детермінації при використанні останнього на перевірочній виборці.

На основі розробленого алгоритму було розроблено та реалізовано програмне забезпечення, що може бути застосовано для вирішення задачі прогнозування. Програма розроблена у середовищі розробки Spyder на базі мови програмування Python.

Було показано приклади використання створеного програмного забезпечення на біомедичних даних, в яких описуються показники серцевосудинної системи. У подальшому програмне забезпечення буде застосовуватися для вирішення задачі прогнозування будь-яких кількісних показників.

Також були розглянуті норми та заходи охорони праці та безпеки роботи в приміщенні де буде розроблюватись та застосовуватись дане методичне забезпечення. Були визначені всі потенційно шкідливі та

небезпечні фактори, котрі можуть негативно вплинути на організм людини під час роботи, та описані заходи, скеровані на усунення цих факторів або на зменшення їх впливу на працюючих.

ПЕРЕЛІК ПОСИЛАНЬ

1. Орлов А.И. Некоторые неклассические постановки в регрессионном анализе и теории классификации // Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. - М.: Наука, 1987. С.27-40.
2. Айвазян С.А., Бежаева З.И.. Староверов О.В. Классификация многомерных наблюдений. - М.: Статистика, 1974. – 240 с.
3. N. Balakrishnan: Handbook of the Logistic Distribution. Marcel Dekker, Inc., 1991.ISBN 978-0-8247-8587-1.
4. Шехурін Д.Е. Наукове прогнозування засобами інформації С. - Пт.: 1990. 123С.
5. Лекции по методам оценивания и выбора моделей, [Електронний ресурс] / К. В. Воронцов– Режим доступу: www.ccas.ru/voron/download/Modeling.pdf
6. А. Г. Ивахненко. Моделирование сложных систем. — К.: Вища школа, 1987. – 275 с.
7. Розова С.С. Классификационная проблема в современной науке. - Новосибирск: Наука, 1986. – 224 с.
8. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.
9. Орлов А.И. Высокие статистические технологии // Заводская лаборатория. Диагностика материалов. 2003. Т.69. №11. С.55-60.
10. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. - с. 20-23
11. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования.— Киев: «Наук.думка», 1985, - 216 с
12. Akaike, H. A new look at the statistical model identification — IEEE Transactions on Automatic Control. — 1974 Т. 19. — 716—723 с.
13. Schwarz, Gideon E."Estimating the dimension of a model", Annals of Statistics 6 (2) – 1978 , - 461–464p.
14. Mallows, C. L. "Some Comments on CP". Technometrics 15 (4) – 1973, - 661–675

15. Efroymson,M. A. "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York. – 1960
16. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.
17. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. - с. 20-23
18. Schwarz, Gideon E."Estimating the dimension of a model", Annals of Statistics 6 (2) – 1978 , - 461–464p.
19. Mallows, C. L. "Some Comments on CP". Technometrics 15 (4) – 1973, - 661–675
20. Жамбю М. Иерархический кластер-анализ и соответствия. - М: Финансы и статистика, 1988. - 342 с.
21. Орлов А.И. Некоторые вероятностные вопросы кластер-анализа // Общая биология. Новые данные исследований структуры и функций биологических систем. Доклады МОИП, 1985. - М.: Наука, 1987. - С.53-56.
22. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. – 296 с.
23. А. Г. Ивахненко. Моделирование сложных систем. — К.: Вища школа, 1987. – 275 с.
24. Шехурін Д.Е. Наукове прогнозування засобами інформації С. - Пт.: 1990. 123С.
25. Косолапов В.В. Інформаційне прогнозування і забезпечення. До: 1978. 198 с.
26. Орлов А.И. Прикладная статистика. - М.: Экзамен, 2006. - 672 с.
27. Орлов А.И. Эконометрика. – М.: Экзамен, 2004. - 576 с.
28. Шрейдер Ю.А., Шаров А.А. Системы и модели. - М.: Радио и связь, 1982. – 152 с.
29. Орлов А.И. Современная прикладная статистика // Заводская лаборатория. Диагностика материалов. 1998. Т.64. №3. С. 52 - 60.
30. Орлов А.И. Высокие статистические технологии // Заводская лаборатория. Диагностика материалов. 2003. Т.69. №11. С.55-60.

31. Орлов А.И. Некоторые неклассические постановки в регрессионном анализе и теории классификации // Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. - М.: Наука, 1987. С.27-40.