**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**

**імені ІГОРЯ СІКОРСЬКОГО”**

**Факультет біомедичної інженерії**

**Кафедра біомедичної кібернетики**

ЗВІТ

**з переддипломної практики**

освітньо-кваліфікаційного рівня “**магістр**”

з напряму підготовки – 6.050101 «Комп’ютерні науки»

|  |  |
| --- | --- |
| **На тему** | **Ідентифікація осіб із зниженими регуляторними** |
| **резервами** | |

(тема індивідуального завдання на початок практики)

**Студент 6 –го курсу гр. БС-71мп Войник Богдан Олексійович \_\_\_\_\_\_\_\_**

(підпис)

**Керівник практики від кафедри БМК**

**\_\_\_\_\_*ст.викл. Аверьянова О.А*.\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_**

(вчені ступінь та звання, прізвище,ініціали)(підпис)

**\_\_\_\_\_\_*ас. Корнієнко Г.А.*\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_**

(вчені ступінь та звання, прізвище,ініціали)(підпис

**Керівник ДР \_\_\_\_\_*зав. каф. Настенко Є.А*.\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_**

(вчені ступінь та звання, прізвище,ініціали)(підпис)

#### Київ – 2018 р.

Зміст

[ВСТУП 3](#_Toc481710203)

[АНОТАЦІЯ 4](#_Toc481710204)

[АННОТАЦИЯ 5](#_Toc481710205)

[ANNOTATION 6](#_Toc481710206)

[1 ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ З ТЕМИ індивідуального завдання 7](#_Toc481710207)

[1.1. Кластерний аналіз 7](#_Toc481710208)

[1.2. Метод кластеризації k-середніх 8](#_Toc481710209)

[1.3. Міри відстані 8](#_Toc481710210)

[2 ОСНОВНА ЧАСТИНА ПЕРЕДДИПЛОМНОЇ ПРАКТИКИ 10](#_Toc481710211)

[2.1. Теоретична частина 10](#_Toc481710212)

[2.2. Аналітична частина 12](#_Toc481710213)

[2.3. Програмна частина 20](#_Toc481710214)

[ВИСНОВОК 23](#_Toc481710215)

[СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 24](#_Toc481710216)

[ПРЕЗЕНТАЦІЯ ДО ЗАХИСТУ ПЕРЕДДИПЛОМНОЇ ПРАКТИКИ 25](#_Toc481710217)

# ВСТУП

**Актуальність:**

**Об’єкт дослідження:**

**Предмет дослідження:**

**Мета роботи:**

У відповідності з метою ставлять такі завдання:

**Методи розробки проекту:**

# АНОТАЦІЯ

1 «Переддипломна практика» (практика) є частиною циклу нормативних дисциплін ООП *бакалавра* за напрямом підготовки *6.050101 «Комп'ютерні науки».*

2. Загальна трудомісткість практики становить 4,5 кредитів (ЕКТС), 135 годин. Період проведення практики с 18.04.2017р. по 12.05.2017р. (12.05.2017р.).

3. Практика реалізувалась **Войником Богданом Олексійовичем** студентом *4-го курсу, гр. БС-31* кафедри *Біомедичної кібернетики* факультету *Біомедичної інженерії НТУУ «КПІ»* .

4. Тема практики: «Реалізація програмного продукту для віднесення студентів 1-2 курсу за функціональними реакціями на тестове навантаження до найближчого кластеру».

5. Ціль та задачі практики.

Ціль – розробити програмний продукт для знаходження мінімальної відстані до кластеру.

Задачі:

1. Аналіз бази даних студентів;

2. Аналіз методів кластеризації та визначення оптимального;

3. Розробка програмного продукту;

4. Перевірка роботи алгоритму на тестовій вибірці.

6. Результати по темі практики:

Отримано навички роботи в середовищі розробки Microsoft Visual Studio 2013 Express на мові C#;

Реалізовано практичне застосування набутих знань шляхом створення програмного продукту для знаходження мінімальної відстані до кластеру.

7. Зміст звіту по практиці:

Титульний аркуш, зміст, вступ, анотацію, огляд літературних джерел з теми індивідуального завдання, основна частина, висновки, план публікацій по темі дипломної роботи, список використаних джерел, презентація до захисту з практики.

8. За практикою надані документи контролю проходження практики:

- щоденник практики;

- індивідуальне завдання;

- звіт на \_29\_\_ листах, додаток до звіту на \_\_\_\_\_\_ листах;

- презентація на \_\_10\_\_\_ слайдах;

- відгук керівника ДР;

- план публікацій чи впровадження;

- гарантійний лист.

9. Сформована тема дипломної роботи до наказу: Оцінка функціональної реакції на тестове навантаження у студентів 1-2 курсу. Жінки. Чоловіки. Комплексний проект.

10. Проміжна атестація у формі *диференціального заліку*.

11. Ключові слова: кластер, кластеризація, k-середніх, мінімальна відстань.

# АННОТАЦИЯ

1 «Преддипломная практика» (практика) является частью цикла нормативных дисциплин ООП *бакалавра* по направлению подготовки *6.050101 «Компьютерные науки»*.

2. Общая трудоемкость практики составляет 4,5 кредитов (ЕКТС), 135 часов. Период проведения практики с 18.04.2017г по 12.05.2017г. (12.05.2017г.).

3. Практика реализовывалась **Войником Богданом Алексеевичем** студентом *4-го курса, гр. БС-31* кафедры *Биомедицинской кибернетики* факультета *Биомедицинской инженерии НТУУ «КПИ»* .

4. Тема практики: «Реализация программного продукта для отнесения студентов 1-2 курса за функциональными реакциями на тестовое нагрузки к ближайшему кластеру».

5. Цель и задачи практики.

Цель – разработать программный продукт для нахождения минимального расстояния к ближайшему кластеру.

Задачи:

1. Анализ базы данных студентов

2. Анализ методов кластеризации и определение самого оптимального

3. Разработка программного продукта

4. Проверка работы алгоритма на тестовой выборке.

6. Результаты по теме практики:

Получены навыки работы в среде разработки Microsoft Visual Studio 2013 Express на языке C#;

Реализовано практическое применение приобретенных знаний путем создания программного продукта для нахождения минимального пути к кластеру.

7. Содержание отчета по практике:

Титульный лист, содержание, введение, аннотация, обзор литературных источников по теме индивидуального задания, основная часть, выводы, план публикаций по теме дипломной работы, список использованных источников, презентация к защите на практике.

8. По практике предоставлены документы контроля прохождения практики:

- дневник практики;

- индивидуальное задание

- отчет на \_29\_\_ листах, приложение к отчету на \_\_\_\_\_\_ листах.

- презентация на \_\_10\_\_\_ слайдах;

- отзыв руководителя ДР;

- план публикаций или внедрения;

- гарантийное письмо.

9. Сформированная тема дипломной работы в приказ: Оценка функциональных реакций на тестовые нагрузки у студентов 1-2 курса. Женщины. Мужчины. Комплексный проект.

10. Промежуточная аттестация в форме *дифференциального зачета*.

11.Ключевые слова: кластер, кластерный анализ, k-средних, минимальное расстояние.

# ANNOTATION

1 «Undergraduate practice» (practice) is a part of cycle regulatory disciplines PLO undergraduate field of study *6.050101 «Computer Science»*.

2. Total labor practices is 4.5 credits (ECTS), 135 hours. Practice period from 18.04.2017 to 12.05.2017. (12.05.2017.).

3. The practice was implemented by **Voinyk Bohdan Oleksyiovich** 4th course student, group BS-31, *Biomedical cybernetics* department of *Biomedical engineering faculty*, NTUU «KPI».

4. Practice theme: « Implementation of the software to assign students 1-2 courses for functional responses to a test load to the closest cluster ».

5. Goal and tasks:

Goal – To develop software for finding minimum distance to the cluster.

Tasks:

1. Analysis of students database

2. Clustering analysis methods and finding the most optimal

3. Software product development

4. Checking the algorithm using a test sample.

6. Practice results:

Acquired skills in development environment Microsoft Visual Studio 2013 Express language C#;

Implemented practical usage of knowledge by creating a software for finding minimum distance to the cluster.

7. Contents of the practice report:

The title page, contents, introduction, annotation, review of the literature on the subject of individual assignment, the main part, the conclusions, the plan of publications on the topic of the theses, a list of sources used, the presentations.

8. The control documents in practice:

- practice diary;

- individual task;

- report on \_29\_\_ pages, attachment to report on \_\_\_\_\_\_ pages;

- presentation on \_\_\_10\_\_ slides;

- scientific director’s recall;

- plan of publishing or implementation;

- letter of indemnity.

9. Formed theme of the graduate work in order: Assessment of functional response to stress test the students 1-2 course. Women. Men. Complex project.

10. Interim certification in the form of a differential offset.

11. Key words: k-means, clustering, minimum distance.

# РОЗДІЛ 1. ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ З ТЕМИ ІНДИВІДУАЛЬНОГО ЗАВДАННЯ

# 1.1. РЕГРЕСІЙНИЙ АНАЛІЗ

Термін "регресія" введений англійським психологом і антропологом Ф. Гальтон і пов'язаний з одним з його перших прикладів, в якому Гальтон, обробляючи статистичні дані, пов'язані з питанням про спадковість зростання, знайшов, що якщо зростання батьків відхиляється від середнього зросту всіх батьків на *х* дюймів, то зростання їх синів відхиляється від середнього зросту всіх синів менше, ніж на *x* дюймів. Виявлена тенденція була названа регресією до середнього стану.

Після проведення кореляційного аналізу, коли виявлено наявність статистично значущих зв'язків між змінними і оцінена ступінь їх тісноти, зазвичай переходять до математичного опису виду залежностей з використанням методів регресійного аналізу. З цією метою підбирають клас функцій, що зв'язує результативний показник  і аргументи , обчислюють оцінки параметрів рівняння зв'язку і аналізують точність отриманого рівняння [3, 13].

На відміну від кореляційного аналізу, який тільки відповідає на питання, чи існує зв'язок між аналізованими ознаками, регресійний аналіз дає і її формалізоване вираження. Крім того, якщо кореляційний аналіз вивчає будь-яку взаємозв'язок факторів, то регресійний - односторонню залежність, тобто зв'язок, що показує, яким чином зміна факторних ознак впливає на ознака результативний.

Регресійний аналіз - один з найбільш розроблених методів математичної статистики. Строго кажучи, для реалізації регресійного аналізу необхідно виконання ряду спеціальних вимог (зокрема, повинні бути незалежними, нормально розподіленими випадковими величинами з постійними дисперсіями). У реальному житті суворе відповідність вимогам регресійного і кореляційного аналізу зустрічається дуже рідко.

Регресійний аналіз - це метод встановлення аналітичного виразу стохастичної залежності між досліджуваними ознаками. Рівняння регресії показує, як у середньому змінюється *у* при зміні будь-якого з і має вигляд:

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

де - залежна змінна (вона завжди одна);

- незалежні змінні (фактори) (їх може бути декілька).

Якщо незалежна змінна одна - це простий регресійний аналіз. Якщо ж їх декілька, то такий аналіз називається багатофакторним.

На відміну від кореляційного аналізу не з'ясовує чи істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії. Регресійний аналіз використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія. На жаль, більшість реальних моделей не вкладаються в рамки лінійної регресії. Таким чином, необхідні комплексні методи для передбачення майбутніх значень. Функція , що описує залежність умовного середнього значення результативної ознаки *у* від заданих значень аргументів, називається функцією (рівнянням) регресії.

Для точного опису рівняння регресії необхідно знати умовний закон розподілу результативного показника *у.* У статистичній практиці таку інформацію отримати зазвичай не вдається, тому обмежуються пошуком підходящих апроксимацій для функції заснованих на вихідних статистичних даних. У рамках окремих модельних припущень про тип розподілу вектора показників може бути отриманий загальний вигляд рівняння регресії

|  |  |
| --- | --- |
|  | (1.2) |

Наприклад, в припущенні про те, що досліджувана сукупність показників підпорядковується ( -мірному нормальному закону розподілу з вектором математичних сподівань:

|  |  |
| --- | --- |
|  | (1.3) |

Де  ,

І коваріаційною матрицею:

|  |  |
| --- | --- |
|  | (1.4) |

Де – дисперсія *у,*

Рівняння регресії (умовне математичне очікування) матиме вигляд

|  |  |
| --- | --- |
|  | (1.5) |

Однак у статистичній практиці зазвичай доводиться обмежуватися пошуком підходящих апроксимацій для невідомої істинної функції регресії *f (x),* так як дослідник не має точним знанням умовного закону розподілу ймовірностей аналізованого результативного показника *у* при заданих значеннях аргументів *х.*

З метою найкращого відновлення по вихідним статистичними даними умовного значення результативного показника  і невідомої функції регресії  найбільш часто використовують такі критерії адекватності функції втрат [29].

* Метод найменших квадратів, згідно з яким мінімізується квадрат відхилення спостережуваних значень результативного показника  , від модельних значень , де коефіцієнти рівняння регресії;  значення вектора аргументів на "-М спостереженні:

|  |  |
| --- | --- |
|  | (1.6) |

Вирішується задача відшукання оцінки  вектора . Отримана регресія називається середньоквадратичною.

* Метод найменших модулів , згідно з яким мінімізується сума абсолютних відхилень спостережуваних значень результативного показника від модульних значень , тобто

|  |  |
| --- | --- |
|  | (1.7) |

Отримана регресія називається середньоабсолютною (медіанною).

* Метод мінмакса зводиться до мінімізації максимуму модуля відхилення спостережуваного значення результативного показника у, від модельного значення , тобто

|  |  |
| --- | --- |
|  | (1.8) |

Отримана при цьому регресія називається мінмаксною.

У регресійному методі вид рівняння регресії вибирають виходячи з аналізу фізичної суті досліджуваного явища і результатів спостереження.

Найбільш часто зустрічаються такі види рівнянь регресії:

* лінійне множинне;
* поліноміальний;
* гіперболічне.

Побудова рівняння регресії здійснюється, як правило, методом найменших квадратів, суть якого полягає в мінімізації суми квадратів відхилень фактичних значень результатного ознаки від його розрахункових значень, тобто:

|  |  |
| --- | --- |
|  | (1.9) |

де *y* - число спостережень;

розрахункове значення результатного фактора.

Коефіцієнти регресії рекомендується визначати за допомогою аналітичних пакетів для персонального комп'ютера або спеціального фінансового калькулятора. У найбільш простому випадку коефіцієнти регресії однофакторного лінійного рівняння регресії виду можна знайти за формулами:

|  |  |
| --- | --- |
|  | (1.9) |

|  |  |
| --- | --- |
|  | (1.10) |

**Мета регресійного аналізу:**

* визначення ступеня детермінованості варіації критеріальної (залежної) змінної предикторами (незалежними змінними).
* пророкування значення залежної змінної за допомогою незалежної.
* визначення внеску окремих незалежних змінних у 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.
* побудова рівняння регресії, тобто знаходження виду залежності між результатними показником і незалежними факторами .
* оцінка значимості отриманого рівняння, тобто визначення того, наскільки вибрані факторні ознаки пояснюють варіацію ознаки *у*.

**Алгоритм регресійного аналізу**

Нехай у точках  незалежної змінної *x* отримані виміри . Потрібно знайти залежність середнього значення величини  від величини *х*, тобто  де *a* — вектор невідомих параметрів  Функцію  називають функцією регресії. Звичайно припускають, що  є лінійною функцією параметрів *а*, тобто має вигляд:

|  |  |
| --- | --- |
|  | (1.11) |

де — задані функції.

У цьому випадку матрицю  називається регресійною матрицею.

Для визначення параметрів  звичайно використовують [метод найменших квадратів](https://wiki.tntu.edu.ua/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B9%D0%BC%D0%B5%D0%BD%D1%88%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D1%96%D0%B2&action=edit&redlink=1), тобто оцінки визначають із умови мінімуму [функціонала](https://wiki.tntu.edu.ua/index.php?title=%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D1%96%D0%BE%D0%BD%D0%B0%D0%BB&action=edit&redlink=1):

|  |  |
| --- | --- |
|  | (1.12) |

і з мінімуму функціонала: для корельованих вимірів з кореляційною матрицею *R*.

У якості функцій  при невеликих звичайно служать [степеневі функції](https://wiki.tntu.edu.ua/index.php?title=%D0%A1%D1%82%D0%B5%D0%BF%D0%B5%D0%BD%D0%B5%D0%B2%D0%B0_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D1%96%D1%8F&action=edit&redlink=1)  Часто використовують [ортогональні](https://wiki.tntu.edu.ua/index.php?title=%D0%9E%D1%80%D1%82%D0%BE%D0%B3%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%96_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D1%87%D0%BB%D0%B5%D0%BD%D0%B8&action=edit&redlink=1) й нормовані поліноми на множині :

|  |  |
| --- | --- |
|  | (1.13) |

У цьому випадку легко знайти оцінку

|  |  |
| --- | --- |
|  | (1.14) |

Звідси випливає, що обчислення  не залежить від обчислення інших

Популярне використання в якості  [сплайнів](https://wiki.tntu.edu.ua/index.php?title=%D0%A1%D0%BF%D0%BB%D0%B0%D0%B9%D0%BD&action=edit&redlink=1) , які мають дві основні властивості:

— поліном заданого степеня;

— відмінний від нуля в околиці точки .

# 1.2. МЕТОД ДИСКРИМІНАНТНОГО АНАЛІЗУ

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

* інтерпретація;
* класифікація.

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

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

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

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

У ролі дискримінантного аналізу найчастіше береться лінійна функція записана у вигляді формули (1.15):

|  |  |
| --- | --- |
| , | (1.15) |

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

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

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

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

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

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

Критерій значення Уілкса обчислюють за формулою (1.16):

|  |  |
| --- | --- |
| , | (1.16) |

де *К –* кількість значень;

*k –* число вже обчислених дискримінаційних функцій.

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

Значення ксі-квадрат розраховують за формулою (1.17):

|  |  |
| --- | --- |
| , | (1.17) |

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

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

Канонічна дискримінантна функція для загального випадку k класів записана у формулі (1.18):

|  |  |
| --- | --- |
| , | (1.18) |

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

uj — шукані коефіцієнти дискримінантної функції;

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

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

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

# 1.3. БІНАРНА ЛОГІСТИЧНА РЕГРЕСІЯ

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

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

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

Ймовірність настання події для деякого випадку розраховується за формулою (1.19) та формулою (1.20):

|  |  |
| --- | --- |
| , | (1.19) |

де

|  |  |
| --- | --- |
| *z=b1\*X1 + b2\*Х2 + ...+ bn\*Xn + a*, | (1.20) |

де *X1* - значення незалежних змінних;

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

*а* - деяка константа.

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

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

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

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

# РОЗДІЛ 2. ОСНОВНА ЧАСТИНА ПЕРЕДДИПЛОМНОЇ ПРАКТИКИ

# 2.1. ТЕОРЕТИЧНА ЧАСТИНА

**Лінії тренду**

Підпрогнозуванняммається на увазі науково обгрунтоване передбачення ймовірнісних шляхів розвитку явищ і процесів для більш-менш віддаленого майбутнього.  Задачі прогнозування в економіці й управлінні дуже популярні, оскільки можуть бути використані, наприклад, для завбачення майбутніх прибутків або тенденцій продажу.

Прогнозування засноване на зберіганні загальної тенденції розвитку явищ у часі, тому на практиці процес прогнозування зводиться до добору на підставі даних минулих періодів аналітичних залежностей досліджуваного параметра від чинників, що впливають, і екстраполяції цих залежностей на майбутнє.  Прогноз показника одержують підстановкою необхідного значення чинника в отримане регресійне рівняння.  Таким чином, прогнозне значення є точковою оцінкою середнього значення показника при даних рівнях чинників [1].

Одним з найбільш поширених засобів прогнозування є побудова ліній тренда засобами MS Excel на основі точкової діаграми за фактичними даними минулих періодів. Ті чи інші якісні властивості розвитку виражають різні рівняння трендів. MS Excel пропонує різноманітні типи апроксимуючої залежності: лінійна, логарифмічна, поліноміальна, степенева, експонентна, лінійна фільтрація. Однак, спроба реалізувати запропоновані моделі призводить до значно різних чисельних результатів, які до того ж, часто мають різний напрямок розвитку. Це призводить нас до завдання оцінки можливості використання тієї чи іншої моделі прогнозування.

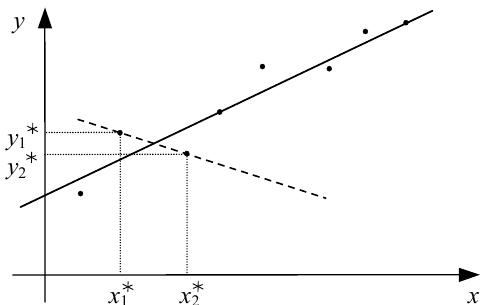
Після побудови ліній тренду на базі теоретично придатних залежностей, кожний результат пропонується оцінити шляхом ранжування за кількома критеріями, які характеризують достовірність, відповідність, надійність та інші параметри прогнозу.

Для визначення достовірності прогнозу можна використати значення похибки апроксимації (*R2*). Чим ближе значення (*R2*) до одиниці, тим точніше обрана модель відбивае тенденцію розвитку, тобто, тим більше можна довіряти результатам прогнозування. При ранжуванні за цим критеріем моделі з максимальним значенням похибки апроксимації присвоюється мінімальний ранг і т.д.

**Метод найменших квадратів**

Метод найменших квадратів (МНК), завдяки широкій сфері застосування, посідає виняткове місце серед методів математичної статистики. Задачею МНК є оцінка закономірностей, які спостерігаються на тлі випадкових коливань, та її використання для подальших розрахунків, зокрема, для прогнозів. Особливу роль відіграють МНК у геофізиці, визначаючи концепцію й методологію розв‘язання оберненої задачі геофізики. Задача МНК розв’язується шляхом параметричної оцінки функції регресії, що описує залежність однієї величини *Y*, значення якої () спостерігають з випадковими похибками (), від групи невипадкових величин

Нехай відомо, що вихідний параметр процесу, який вивчається, позначимо його *y*, лінійно залежить від вхідного параметра *x* (суцільна пряма лінія на рисунку 2.1).

  
Рисунок 2.1 — Графічна інтерпретація причин, які обумовлюють необхідність використання МНК

Тобто припустимо, що статична характеристика цього процесу може бути подана у вигляді

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

де *a* і *b* — коефіцієнти, для визначення числових значень яких необхідно, як мінімум, задати два значення *x*1, *x*2 вхідній величині x і заміряти відповідні їм значення *y*1, *y*2 вихідної величини *y*, оскільки лише під час виконання цих умов для моделі (2.1) можна скласти систему двох алгебраїчних рівнянь із двома невідомими *a* і *b*

|  |  |
| --- | --- |
|  | (2.2) |

Але результати будь-яких експериментальних вимірювань несуть у собі похибки, обумовлені класом точності вимірювальних засобів, дією різноманітних завад, неточністю зчитування показів приладів, округленням під час приведення даних до однакових умов обробки інформації — список умов виникнення похибок можна продовжити, але для обґрунтування МНК цього досить.

Тож через наявність цих похибок в експериментальних значеннях *x*1, *x*2, *y*1, *y*2 безпосередній розв’язок системи рівнянь (2.2) відносно *a* та *b* може нести в собі похибку в 10, 100, 1000 і більше відсотків.

Наприклад, якщо використати лише значення  (рис. 2.1) для розв’язання системи рівнянь (2.2), то похибка буде вже не у відсотках, а у характері функціональної залежності (пунктирна лінія на рис. 2.1).

У свій час Гаусс запропонував інший спосіб визначення коефіцієнтів *a*, *b* моделі (2.1). Він запропонував сформувати суму квадратів різниць  між теоретично заданими за допомогою рівняння (2.1) значеннями вихідної координати *y* при значеннях аргументу   та її експериментальними значеннями :

|  |  |
| --- | --- |
|  | (2.3) |

а потім знайти такі значення коефіцієнтів a , b рівняння (2.1), котрі мінімізують вираз (2.3).

Від цієї процедури і назва методу — метод найменших квадратів.

З курсу математичного аналізу відомо, що для знаходження мінімуму якоїсь функції необхідно взяти від неї похідну, прирівняти цю похідну до нуля і розв’язати отримане рівняння — його корінь задає значення аргументу, за якого функція досягає мінімуму, а само значення функції у цій точці, якщо вона опукла донизу, задає її мінімальне значення.

Згідно з цією ідеєю, підставимо у вираз (2.3) замість  його значення з (2.1) і візьмемо від отриманого виразу частинні похідні за *b* та *a*, які прирівняємо до нуля, тобто

|  |  |
| --- | --- |
|  | (2.4) |
|  | (2.5) |

Із (2.5) після низки нескладних перетворень отримаємо:

|  |  |
| --- | --- |
|  | (2.6) |

Розв’язавши систему рівнянь (2.6) відносно *b* і *a* , отримаємо такі їх значення, які мінімізують суму квадратів відхилень експериментально виміряних значень величин від теоретично заданих згідно з вибраною функціональною залежністю.

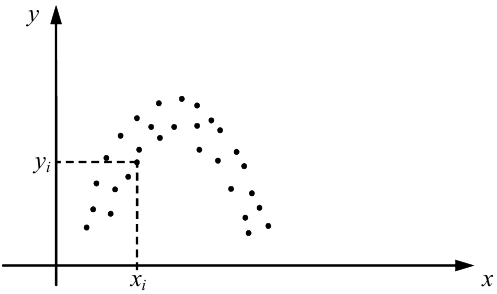
Рівняння, що входять у систему (2.6), називають нормальними рівняннями Гаусса. Коефіцієнтами у них є суми, які «згладжують» дію похибок вимірювань величин *x, y* і зменшують їх вплив на оцінки параметрів *b, a.* Завдяки цьому підвищується точність їх визначення.

А тепер припустимо, що поле точок експериментально визначених величин x , y має такий вигляд, як це показано на рис. 2.2.

Із цього рисунка видно, що середньою лінією цього поля, яка віддзеркалює «в середньому» функціональну залежність y від x , є парабола

|  |  |
| --- | --- |
|  | (2.7) |

параметри якої *a*, *b*, *c* також доцільно визначати за допомогою МНК.

  
Рисунок 2.2 — Поле точок експериментально визначених величин *x*, *y*

Для отримання нормальних рівнянь Гаусса у цьому випадку підставимо (2.7) у (2.3), що дасть вираз

|  |  |
| --- | --- |
|  | (2.8) |

Далі від цього виразу візьмемо частинні похідні за *c , b, a* та прирівняємо їх нулю, що дасть систему рівнянь:

|  |  |
| --- | --- |
|  | (2.9) |

Після спрощень в системі (2.9) матимемо систему нормальних рівнянь Гаусса (2.10), розв’язавши яку отримаємо оптимальні за критерієм мінімуму відхилень експерименту від теорії значення параметрів *c*, *b*, *a* математичної моделі (2.7) функціональної залежності величини *y* від величини *x*, заданих експериментальним полем точок , яке зображено на рис. 3.2:

|  |  |
| --- | --- |
|  | (2.10) |

# 

**Дисперсійний аналіз**

Дисперсійний аналіз (від латинського Dispersio – розсіяння) – статистичний метод, що дозволяє аналізувати вплив різних чинників на досліджувану змінну. Метод був розроблений біологом Р. Фішером в 1925 році і застосовувався спочатку для оцінки експериментів в рослинництві. Надалі з'ясувалася загальнонаукова значимість дисперсійного аналізу для експериментів у психології, педагогіки, медицини та ін .

Метою дисперсійного аналізу є перевірка значущості відмінності між середніми за допомогою порівняння дисперсій. Дисперсію вимірюваної ознаки розкладають на незалежні складові, кожна з яких характеризує вплив того чи іншого фактора або їх взаємодії. Наступне порівняння таких доданків дозволяє оцінити значимість кожного досліджуваного фактору, а також їх комбінації [21].

При істинності нульової гіпотези (про рівність середніх в декількох групах спостережень, вибраних з генеральної сукупності), оцінка дисперсії, пов'язаної з внутрішньогруповою мінливістю, повинна бути близькою до оцінки міжгрупової дисперсії.

При проведенні дослідження ринку часто постає питання про порівнянності результатів. Наприклад, проводячи опитування з приводу споживання якого-небудь товару в різних регіонах країни, необхідно зробити висновки, на скільки дані опитування відрізняються чи не відрізняються один від одного. Зіставляти окремі показники не має сенсу і тому процедура порівняння і подальшої оцінки проводиться за деякими усередненими значеннями і відхиленням від цієї усередненої оцінки. Вивчається варіація ознаки. За міру варіації може бути прийнята дисперсія. Дисперсія – міра варіації, що визначається як середня з відхилень ознаки, зведених в квадрат.

На практиці часто виникають задачі більш загального характеру – завдання перевірки суттєвості відмінностей середніх вибіркових кількох сукупностей. Наприклад, потрібно оцінити вплив різної сировини на якість виробленої продукції, вирішити задачу про вплив кількості добрив на врожайність продукції [22].

Іноді дисперсійний аналіз застосовується, щоб встановити однорідність кількох сукупностей (дисперсії цих сукупностей однакові за припущенням, якщо дисперсійний аналіз покаже, що і математичні сподівання однакові, то в цьому сенсі сукупності однорідні). Однорідні ж сукупності можна об'єднати в одну і тим самим отримати про неї більш повну інформацію, отже, і більш надійні висновки.

Сутність цього аналізу полягає в тому, що загальну дисперсію досліджуваної ознаки розділяють на окремі компоненти, які обумовлені впливом певних конкретних чинників. Істотність їх впливу на цю ознаку здійснюється методом дисперсійного аналізу. Відповідно до дисперсійного аналізу будь-який його результат можна подати у вигляді суми певної кількості компонент. Так, наприклад, якщо досліджується вплив певного чинника на результат експерименту, то модель, що описує структуру останнього, можна подати так:

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

Де — значення ознаки X, одержане при i-му експерименті на j-му рівні фактору.

Під рівнем фактору розуміють певну його міру. Наприклад, якщо фактором є добрива, які вносяться в землю з метою збільшення врожайності сільськогосподарської культури, то рівнем фактора в цьому разі є кількість добрива, що вноситься в ґрунт;

— загальна середня величина ознаки X;

— ефект впливу фактора на значення ознаки X на j-му рівні;

— випадкова компонента, що впливає на значення ознаки X в i-му експерименті на j-му рівні.

При цьому  і , як випадкові величини мають закон розподілу ймовірностей   і між собою незалежні  [21].

Складнішою моделлю аналізу є вивчення впливу на результати експерименту кількох факторів. Зокрема при аналізі впливу двох факторів структура моделі набуває такого вигляду:

|  |  |
| --- | --- |
|  | (2.2) |

де  – значення ознаки Х в *i*-му експерименті на *j*-му рівні впливу фактору *A* і на *k*-му рівні впливу фактора *В*;  — загальна середня величина ознаки X;   — ефект впливу фактору *А* на *i*-му рівні,  — ефект впливу фактора *В* на *j*-му рівні;  — ефект одночасного впливу факторів *A* і *В*;  — випадкова компонента. У разі проведення дисперсійного аналізу досліджуваний масив даних, одержаних під час експерименту, поділяють на певні групи, які різняться дією на результати експерименту певних рівнів факторів.

Попередні методи статистичного аналізу даних використовують для порівняння двох об’єктів. Але на практиці часто виникають завдання, що стосуються групи об’єктів (наборів спостережуваних даних). Одним з методів для таких завдань є дисперсійний аналіз – статистичний метод виявлення на досліджувану випадкову величину (параметр) одночасної дії одного або декількох факторів. Дія деякого фактора на складну систему спричинює мінливість його властивостей. Фактор може бути відомий або невідомий, природного або штучного походження, як от: умови експерименту, методика вимірювань і опрацювання тощо [22].

За кількістю оцінюваних факторів дисперсійний аналіз поділяють на одно-, дво- та багатофакторний. Кожен фактор може бути дискретною чи неперервною випадковою змінною, яку розділяють на декілька сталих рівнів (градацій, інтервалів). Якщо кількість вимірювань на всіх рівнях кожного з факторів однакова, то дисперсійний аналіз називають рівномірним, інакше – нерівномірним.

В основі дисперсійного аналізу є такий принцип (факт з математичної статистики): якщо на випадкову величину діють взаємно незалежні фактори *A*, *B*, то загальна дисперсія дорівнює сумі дисперсій, зумовлених дією окремо кожного з факторів:

|  |  |
| --- | --- |
|  | (2.3) |

Цей метод ґрунтується на розділенні загальної дисперсії  на складові, що відповідають впливу різних джерел мінливості (дисперсія , зумовлена дією факторів, і залишкова дисперсія , ), а застосовувані критерії дають змогу одночасно вивчати відмінності як у середніх значеннях, так і в дисперсіях [21, 22].

Однофакторний дисперсійний аналіз

Для простоти розглянемо спочатку рівномірний дисперсійний аналіз (одну з можливих моделей), а потім наведемо необхідні модифікації для виконання нерівномірного аналізу.

Результати вимірювань запишемо у вигляді матриці з *n* рядків та *p* стовпців:

|  |  |
| --- | --- |
|  | (2.4) |

Кожен стовпець (градацію фактора) треба розглядати як вибірку нормально розподілених випадкових величин    з параметрами ,   для всіх *j=1,…,p* (дисперсії однакові).

Отже, для кожної градації фактора (стовпця таблиці даних) маємо фіксоване середнє значення, що є сталим у межах експерименту. Гіпотезу для перевірки сформулюємо так:

|  |  |
| --- | --- |
|  | (2.5) |

Отже, дисперсія випадкової величини , зумовлена дією фактора на всіх рівнях, , і вся мінливість буде спричинена неврахованими факторами:   або

У математичній статистиці розроблено формальну процедуру дисперсійного аналізу (ANOVA, ANalysis Of VAriance). Схема перевірки нульової гіпотези така.

**А.** Обчислюємо генеральне середнє  і вибіркові середні  для рівномірного однофакторного аналізу:

|  |  |
| --- | --- |
|  | (2.6) |

|  |  |
| --- | --- |
|  | (2.7) |

**Б.** Знаходимо суми квадратів відхилень від відповідних середніх значень:

* сума, що характеризує мінливість, зумовлену досліджуваним фактором (факторна сума),

|  |  |
| --- | --- |
|  | (2.8) |

* сума, що характеризує мінливість у межах кожної градації фактором (залишкова сума),

|  |  |
| --- | --- |
|  | (2.9) |

* сума, що характеризує загальну мінливість (загальна або тотальна сума),

|  |  |
| --- | --- |
|  | (2.10) |

Справджується рівність

|  |  |
| --- | --- |
|  | (2.11) |

**В.** Визначаємо оцінки дисперсій:

|  |  |
| --- | --- |
|  | (2.12) |

**Г.** Критерій Фішера для перевірки гіпотези  має вигляд

|  |  |
| --- | --- |
|  | (2.13) |

Для заданого рівня значущості α знаходимо критичні значення статистики .

Обчислені значення записуємо у вигляді таблиці (табл. 2.1), (ANOVA) [21, 22, 23].

Таблиця 2.1 – Результати однофакторного дисперсійного аналізу

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Різновид дисперсії** | **Сума квадратів відхилень** | **Кількість степенів свободи** | **Середній квадрат (оцінка дисперсії)** | **F – критерій** |
| **Факторна (між вибірками)** |  |  |  |  |
| **Залишкова (у вибірці)** |  |  |  |  |
| **Загальна** |  |  |  |  |

# 2.2. АНАЛІТИЧНА ЧАСТИНА

Дослідження передбачають проведення глобальної кластеризації на базі даних студентів та викладачів, що пройшли пробу Мартіне декілька разів. Для кластеризації студентів і визначення їх субгрупи ризику ми використали програмний продукт «ClusterBox», попередньо модифікувавши його, додавши функцію глобального режиму, здатну розставити кластери та мітки (субгрупи ризику) не для одного студента, а для всіх одразу, що присутні в базі даних. На даному етапі дослідження дані зберігаються до текстових файлів. Вікно програми показано на рисунку 2.3.



Рисунок 2.3 – Вікно дослідження

Таким чином, якщо стоїть галочка на пункті «Глобальний режим», то викликається меню з вибором бази даних для глобальної кластеризації, а потім запускається алгоритм знаходження групи ризику.

Наступним кроком у дослідженні є аналіз графіків кластерів, побудованих на базі результуючих таблиць. Слід зазначити, що перед дослідженнями база даних була розщеплена по статі, оскільки попередні дослідження показують відмінність чоловічої групи від жіночої. Всі дослідження показані на прикладі чоловічої групи

Результуючі таблиці були побудовані на базі студентів НТУУ «КПІ ім. І. Сікорського» 1-2 курсу, за допомогою проведення дисперсійного аналізу. Приклад результуючої таблиці показано на рисунку 2.4 – 2.5.

Також дослідження показали, що кластеризація початкової бази даних по чоловікам збігається з методом кластеризації k-середніх на 79,49%. Отже алгоритм показує хороші результати.

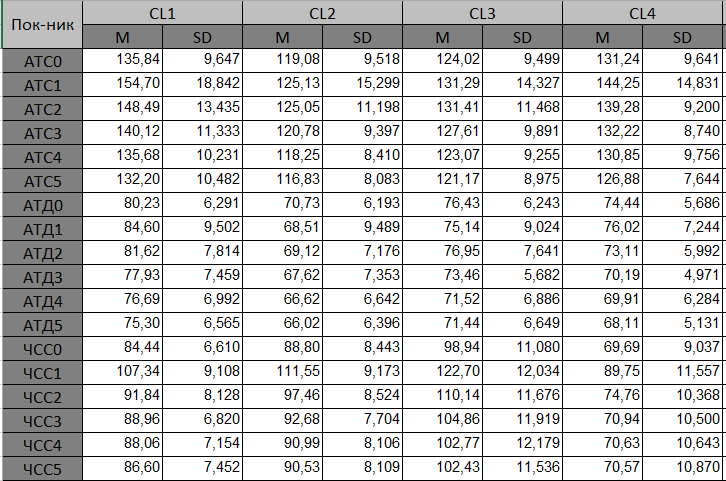


Рисунок 2.4 – Згрупована таблиця результатів

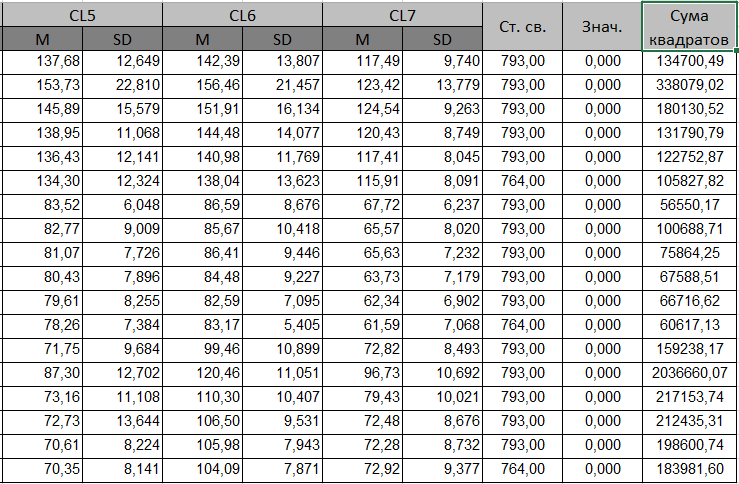


Рисунок 2.5 – Згрупована таблиця результатів

За даними середніх значень із результуючих таблиць ми побудували графіки середніх значень АТД і АТС по кластерам

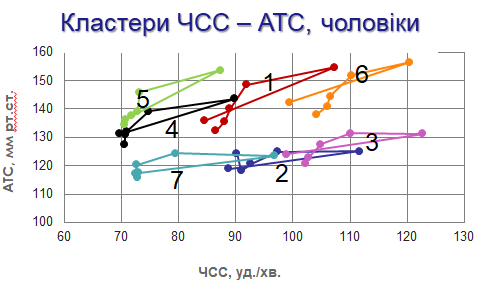


Рисунок 2.6 – Графіки ЧСС, АТС по кластерам

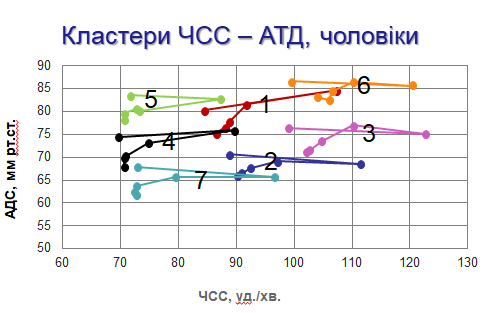


Рисунок 2.7 – Графік ЧСС, АТД по кластерам

Дані графіки показують динаміку артеріального тиску та пульсу при виконанні проби Мартіне у кожному кластері. Як ми бачимо, за АТС кластери 2 і 3, 7 і 2 перетинаються, а за АТД перетинаються кластери під номерами 1 і 4, 2 і 7, 1 і 6. Таким чином висуваємо гіпотезу про те, що вони мають спільну підгрупу. Також ми можемо побачити, що за АТС кластери 4 і 5 знаходяться приблизно в одному діапазоні як за значеннями тиску, так і за значеннями пульсу, тому ми можемо зменшити кількість кластерів, щоб покращити наші результати.

Також слід перевірити чи дійсно алгоритм квадрату евклідової відстані дає найоптимальніший результат. Для цього необхідним було провести логістичну регресію та дискримінантний аналіз для побудови моделей з їх подальшою оцінкою та порівнянням із алгоритмом знаходження мінімальної дистанції за квадратом евклідової відстані.

Для проведення даних аналізів було застосовано статистичний пакет для обробки даних «Statistical Package for the Social Sciences » або просто SPSS.

Для початку проведемо логістичну регресію. Оскільки класифікатор студентів є не бінарною змінною, тому було прийнято рішення розбити пацієнтів на групи методом «один проти всіх» та вирівняти дані в групах, де були отримані показники класифікації занадто асиметричні. Результати дослідження представлені для третього кластеру, але аналогічна процедура проводилася для кожного кластеру окремо. Результуючу таблицю для всіх кластерів зображено у таблиці 2.1.

На рисунку 2.8 зображено встановлення параметрів, які вводилися для отримання класифікації даних.

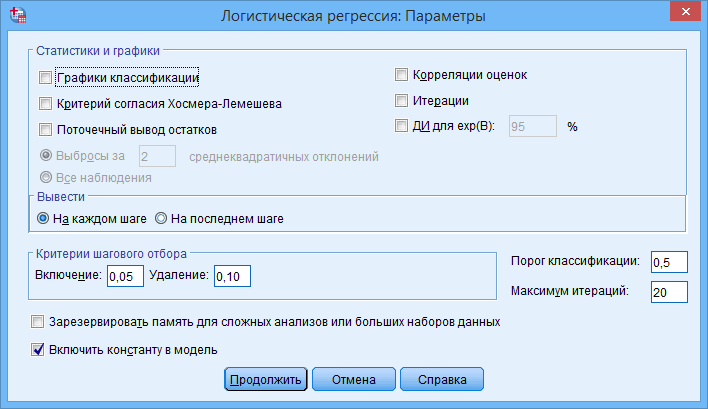


Рисунок 2.8 – Зображення параметрів

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

На рисунку 2.9 наведені результати класифікації при об’єднанні у групи: перша група – студенти, що знаходяться в 3 кластері, друга – студенти в інших кластерах (3 проти 1, 2, 4, 5, 6, 7).

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

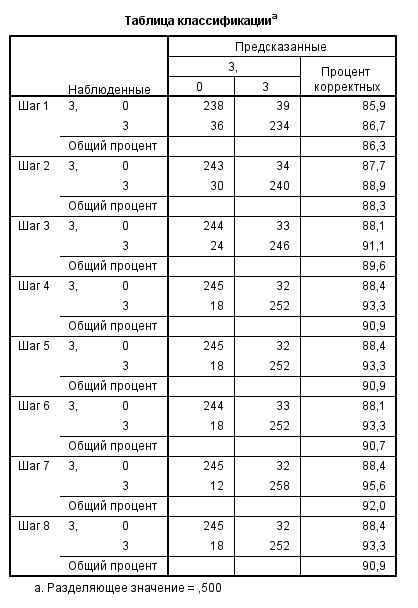


Рисунок 2.9 – Результати класифікації ЛР

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

Для побудови рівняння регресії було використано наступну таблицю:

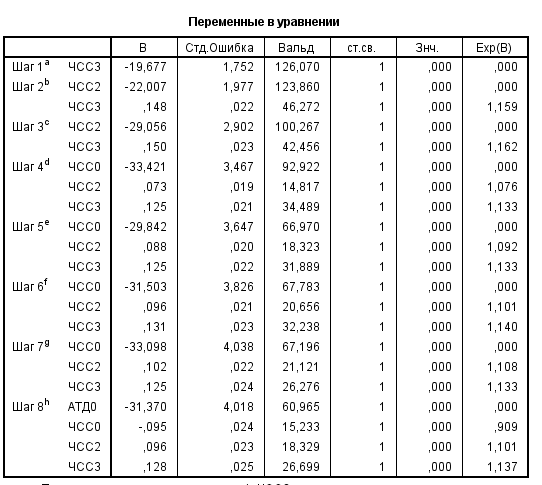


Рисунок 2.10 – Змінні для рівняння регресії

Таким чином, рівняння регресії набуває вигляду:

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

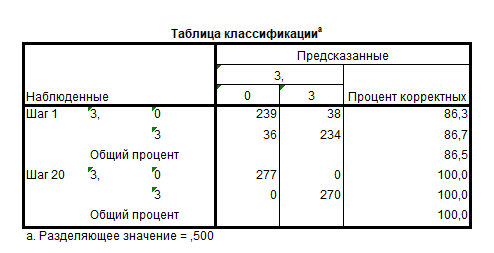


Рисунок 2.11 – Результати класифікації ЛР з нелінійними перетвореннями

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

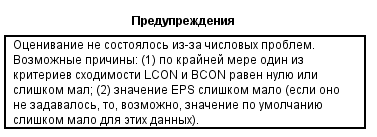


Рисунок 2.12 – Попередження

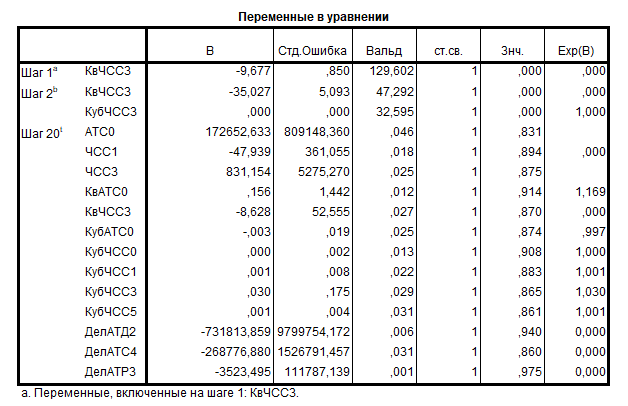


Рисунок 2.13 – Змінні для рівняння регресії з нелінійними перетвореннями

Таким чином, рівняння регресії має вигляд:

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

Таблиця 2.1 – Порівняльна характеристика

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

Розшифрування міток в таблиці 1:

Заг.Кор,% - загальний відсоток коректно спрогнозованих даних

Ск-ть – складність моделі

Ск,% - відсоток складності моделі (чим нижче, тим краще)

21 – кількість істинних змінних

84 – кількість істинних змінних разом з нелійнійними перетвореннями

1- виграш ЛР на істинних даних

2 – виграш ЛР з нелінійними перетвореннями

\* - неповність моделі

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

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

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

Всі дані було класифіковано методом ДА, використовуючи покроковий відбір. На рис. 2.14 показаний вибір відстані та критерій розпізнавання.

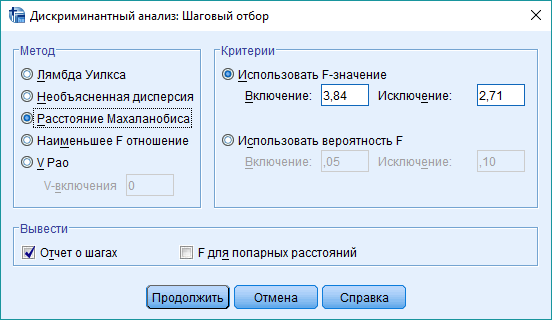


Рисунок 2.14 – Вибір методів

У вікні вибору класифікації, що зображено на рис 2.15 показаний вибір розрахунку класифікації.

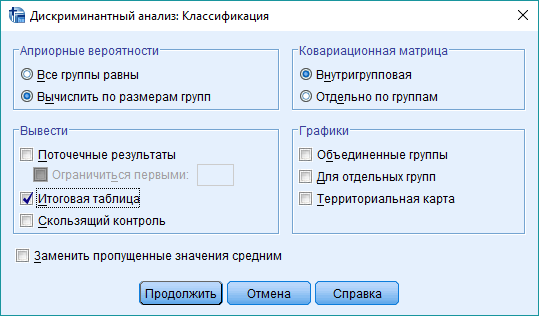


Рисунок 2.15 – Вибір класифікації

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

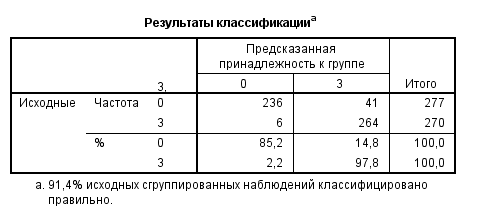


Рисунок 2.16 – Результати класифікації

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

Для побудови дискримінантної функції застосовуються дані з наступної таблиці:



Рисунок 2.17 – Коефіцієнти для дискримінантної функції

Таким чином, дискримінантна функція набуває вигляду:

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

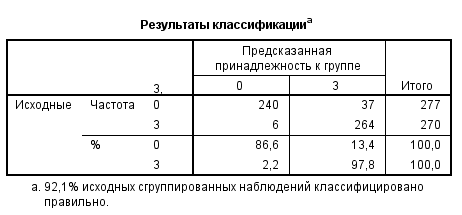


Рисунок 2.18 – Результати класифікації з нелінійними перетвореннями

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



Рисунок 2.19 – Коефіцієнти для дискримінантної функції з нелінійними перетвореннями [6,7].

Таким чином, дискримінантна функція має вигляд:

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

Аналогічно до таблиці 2.1 було побудовано таблицю порівняльної характеристики для дискримінантного аналізу щоб оцінити результати дослідження.

Таблиця 2.2 – Порівняльна характеристика

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

З таблиці 2.2 видно, що аналіз, проведений на базі нелінійних перетворень дає складніші рівняння моделі, але їх складність, беручи до уваги всі змінні, менша. Слід також зауважити, що у 7 тестах із 7 відсоток коректно спрогнозованих даних збільшився.

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

Виходячи з результатів досліджень, ми можемо сказати, що при порівнянні алгоритмів, у 2 випадках із 7 (виключаючи неповну модель при логістичній регресії) найбільш точну і просту модель будує дискримінантний аналіз, ще в 2 із 7 – метод логістичної регресії. Ще у одному випадку логістична регресія та дискримінантний аналіз дають ідентичний результат. При цьому дискримінантний аналіз є більш простим та універсальним у використанні, оскільки при його застосуванні ми завжди маємо справу тільки з однією статистичною процедурою, в якій беруть участь одна категоріальна залежна змінна і кілька незалежних змінних з будь-яким типом шкали. Також з досліджень видно, що логістична регресія не завжди може побудувати повну модель. Тому ми ввели нелінійні змінні для методів логістичної регресії та дискримінантного аналізу

При введені нелінійних змінних ми бачимо, що логістична регресія дає високі результати, але при цьому зростає складність моделі. Таким чином, найбільш оптимальних результатів можна досягти за допомогою логістичної регресії з введенням нелінійних змінних , та .

Якщо подивитись на алгоритм знаходження мінімальної відстані, можна побачити, що модель є трохи складнішою, але його застосування є набагато простішим. Він не вимагає побудови моделей для кожного кластеру з попереднім визначенням коефіцієнтів. Для дослідження необхідно знати лише показники артеріального тиску та пульсу і мати в наявності результуючу таблицю (поставляється разом з програмним продуктом). Високий відсоток коректності в методах ДА і ЛР свідчать про те, що алгоритм квадрату евклідової відстані розставляє кластери таким чином, що одну групу можна легко відрізнити від усіх інших.

Слід також зазначити, що при зменшенні кількості кластерів робота алгоритму не зміниться, а побудова нової результуючої таблиці є набагато простішим у порівнянні з виведенням моделей методами логістичної регресії чи дискримінантного аналізу. Також серед переваг алгоритму є високий відсоток збіжності з методом кластеризації k-середніх, простота в обчисленнях, одна модель (алгоритм), що застосовуються для визначення мінімальної відстані до кожного кластеру і можливість швидкої кластеризації як всіх даних в базі даних, так і одного пацієнта.

# 2.3. ПРОГРАМНА ЧАСТИНА

Дослідження кластерів було вирішено проводити не на початковій базі даних, що містить 1500 досліджень і виступає в ролі бази для побудови результуючих таблиць, а на базі студентів та викладачів НТУУ «КПІ ім. І. Сікорського», що пройшли пробу Мартіне більше одного разу, оскільки на початковій базі даних результати відповідатимуть графікам кластерів, так як вони побудовані на основі даних із цієї бази даних.

Програма для визначення закономірностей і подібності кластерів передбачає побудову графіків «кластер-кластер» та графіків переходу «кластер-мітка». Ми припускаємо, що студенти, вектор направлення яких є графіком «кластер-мітка» може відповідати групі студентам, вектор направлення яких «мітка-кластер». Наприклад, студенти, що знаходяться в кластері 1, а субоптимальним кластером (міткою) є 3, то таким чином вектор направлення студента описується рухом студента з кластеру один до кластера три. Аналогічно є студенти, які прямують із кластера 3 в кластер 1. Ідея полягає в наступному: чи зв’язані ці переходи між собою. Але оскільки кількість кластерів є досить великою і деякі з кластерів, виходячи з графіків на рисунку 2.6, схожі, тому дане дослідження доцільно проводити після зменшення кількості кластерів. Але сам програмний продукт може аналітично показати чи дійсно необхідним є зменшення кількості кластерів.

Для перевірки гіпотези достатньо провести звичайну лінійну регресію та побудувати лінії тренду у відповідних групах. Для цього розроблено спеціальний модуль «RegressMod», який на даному етапі виступає окремим програмним продуктом.

Програмний продукт розроблено в середовищі Microsoft Visual Studio 2017 Community за допомогою мови програмування С# з використанням елементів WinForms та фреймворку .Net Framework. Для демонстрації результатів кластеризації був розроблений багатофункціональний інтерфейс користувача.

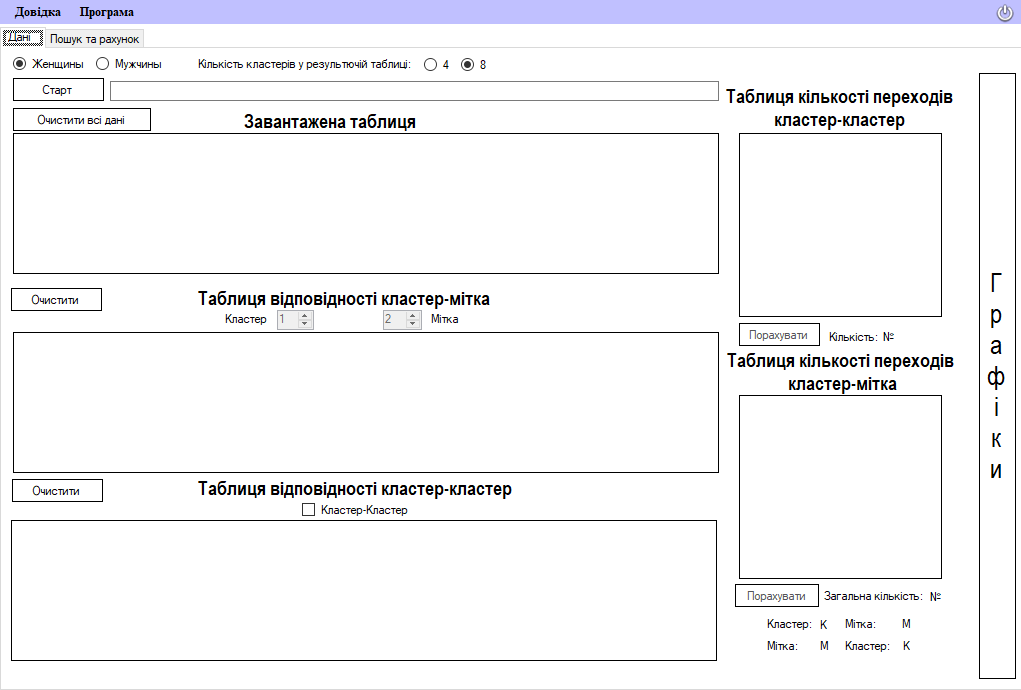


Рисунок 2.20 – Вікно проведення регресійного дослідження

Для того, щоб розпочати дослідження необхідно натиснути клавішу «Старт» і вибрати базу даних, що на одному листі містить артеріальний тиск та пульс, кластер, мітку, мінімальну та субмінімальну відстань. Слід зазначити, що для дослідження необхідно попередньо розрахувати АТР1, АТР2, АТР3, що показують різницю між систолічним та діастолічним тиском за першу, другу і третю хвилині після навантаження.

Для побудови графіків необхідно заповнити таблицю відповідності кластер-мітка. Вона показую всі переходи між вибраним кластером і міткою, а також між міткою та кластером. Також після заповнення таблиці стає можливим розрахунок всіх пар «кластер-мітка» з виводом результатів до вікна «Таблиця кількості переходів кластер-мітка», де по вертикалі показано кластери, а по горизонталі – мітки. Також при натисненні на комірку в таблиці можна побачити загальну кількість переходів між обраними кластером і міткою.

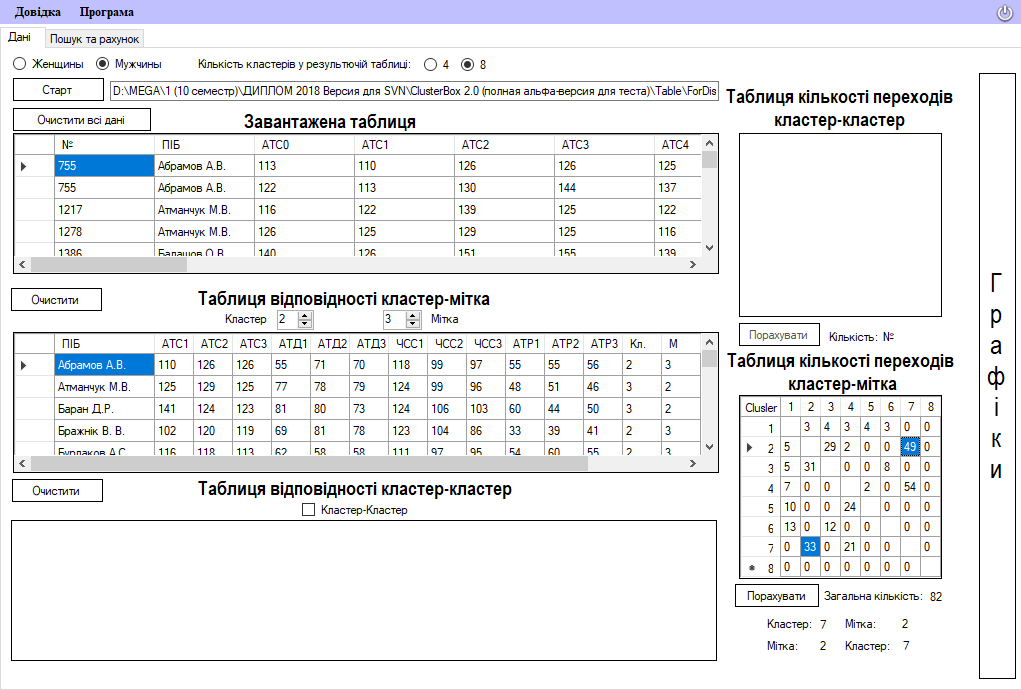


Рисунок 2.21 – Заповнені таблиці для дослідження

Якщо натиснути на клавішу «Графіки», то побачимо меню, в якому є таблиця коефіцієнтів регресії і клавіша «Графік АТС». При натисненні на «Графік АТС» побудуються лінії трендів «кластер-мітка» та «мітка-кластер», дані яких завантажені до таблиці відповідності кластер-мітка. Відповідні коефіцієнти автоматично розрахуються та будуть занесені до таблиці коефіцієнтів регресії для того, щоб користувач міг оцінити їх подібність.

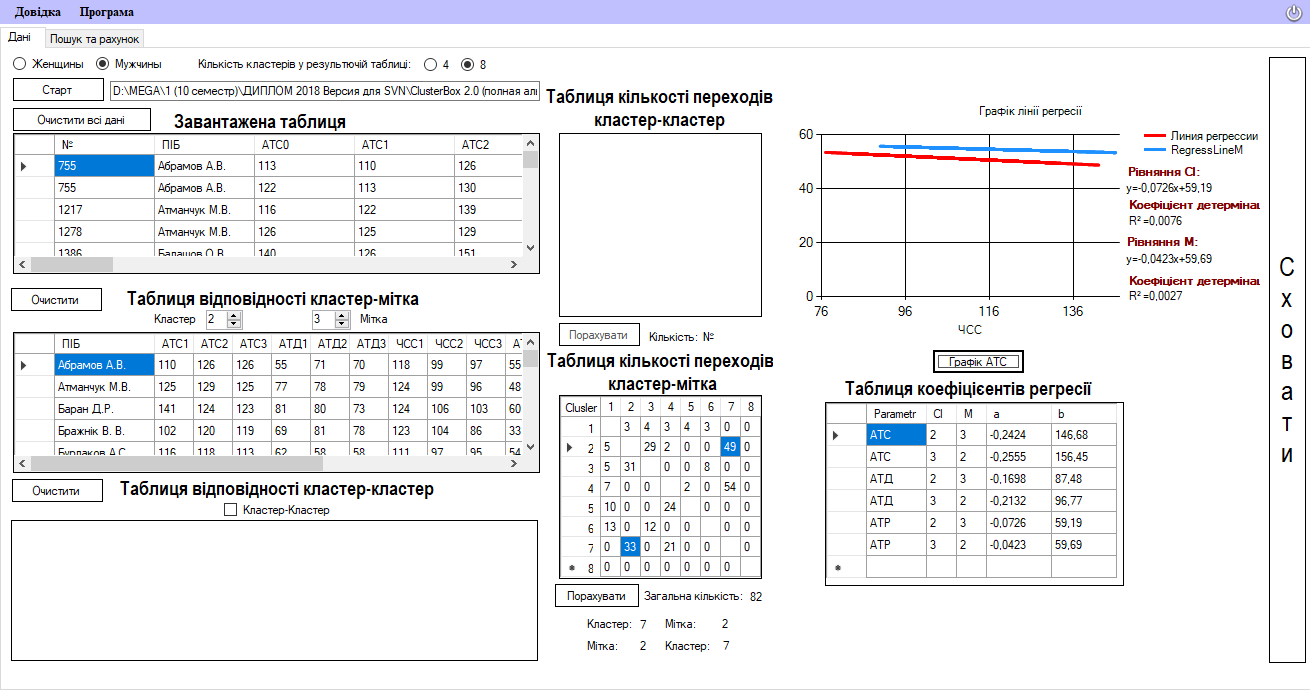


Рисунок 2.22 – виведення графіків і таблиці коефіцієнтів регресії

За проведеними дослідженнями за кластером 2 і міткою 3 ми бачимо, що за АТС коефіцієнт регресії по кластеру збігається з коефіцієнтом регресії по мітці на 94,87%, а коефіцієнт – на 93,75%; за АТД коефіцієнт регресії по кластеру збігається з коефіцієнтом регресії по мітці на 79,64%, а коефіцієнт – на 90,39%; за АТР коефіцієнт регресії по кластеру збігається з коефіцієнтом регресії по мітці на 58,26%, а коефіцієнт – на 99,16%;

Таким чином ми бачимо, що лінії регресії достатньо схожі між собою і можемо припустити, що кластери 2 і 3, які, як видно з графіків на рисунках 2.6 перетинаються, мають спільну підгрупу, а, отже, це є ще одним фактором зменшення кількості кластерів та перерахунку результуючих таблиць вже на новій базі даних

Для підтвердження даної гіпотези проведемо дослідження між кластерами 2 і 3. Для цього в програмі поставимо галочку «кластер-кластер».

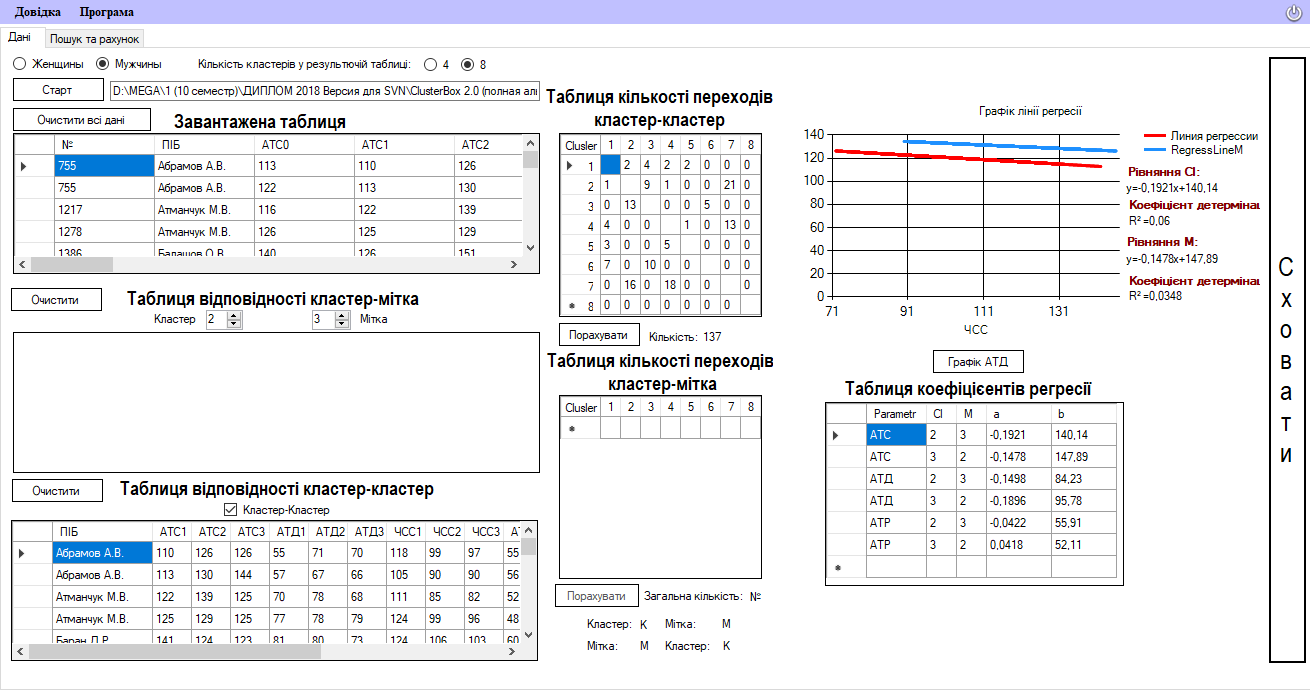


Рисунок 2.23 – Дослідження кластер-кластер

Оскільки кластери 2 і 3 перетинаються на графіках АТС-ЧСС, то доцільним буде дослідити подібність ліній тренду в цій області. Ми бачимо, що коефіцієнт регресії по кластеру 2 збігається з коефіцієнтом регресії по кластеру 3 на 76,93%, а коефіцієнт – на 94,75%

Дані показники є досить високими, щоб стверджувати, що кластери 2 і 3 мають спільну підгрупу.

Також дослідження алгоритму передбачає пошук студента в завантаженій базі даних для того, щоб подивитись чи перевищує радіус студента радіус кластера. Також розраховується загальна кількість зміни кластерів в базі даних за одним студентом. Дані дослідження показують динаміку роботи алгоритму та можуть бути використані для моделювання системи виводу характеристик кластеру, до якого відноситься студент, та рекомендацій мітки, якщо радіус студента перевищує радіус кластера. Робота алгоритму наведена на рисунку 2.24.



Рисунок 2.24 – Пошук студента в БД з виводом графіків

Якщо радіус студента перевищує радіус кластера, то автоматично розраховуються коефіцієнти ліній регресії і виводяться відповідні графіки.

# 2.4. АНАЛІТИЧНА ЧАСТИНА ЗА ВИСНОВКАМИ ПРОГРАМНОЇ ЧАСТИНИ

Наші дослідження показали, що зменшення кількості кластерів є необхідним кроком для поліпшення роботи алгоритму класифікації на основі квадрату евклідової відстані.

Таким чином ми повторно провели кластеризацію алгоритмом k-середніх. Спочатку було вирішено зменшити кількість кластерів на дві позиції, тому кількість кластерів становить 5.

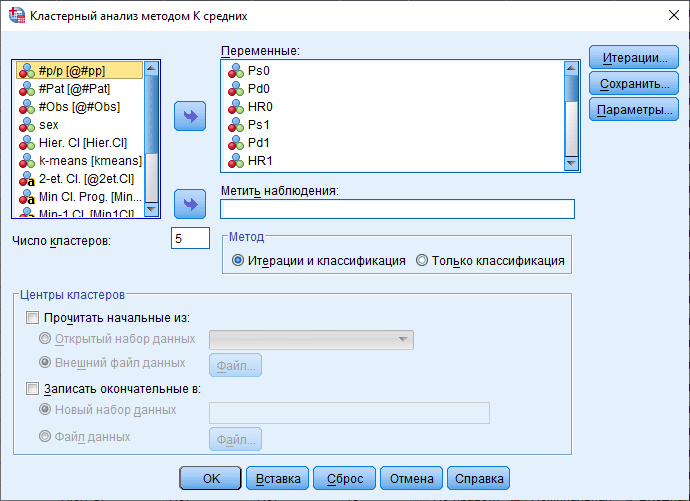
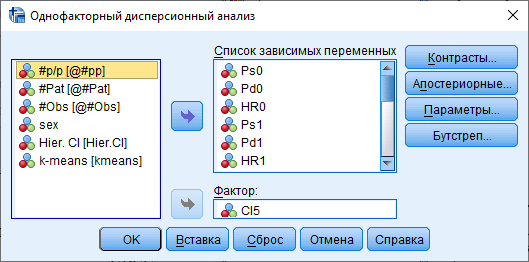
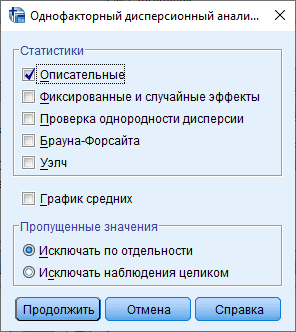
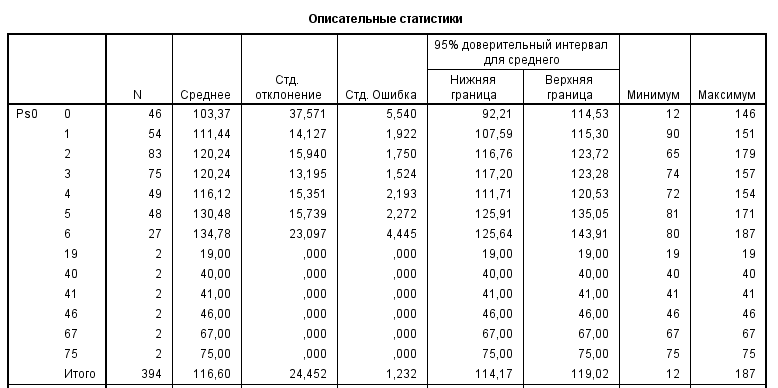


Рисунок – Кластерний аналіз

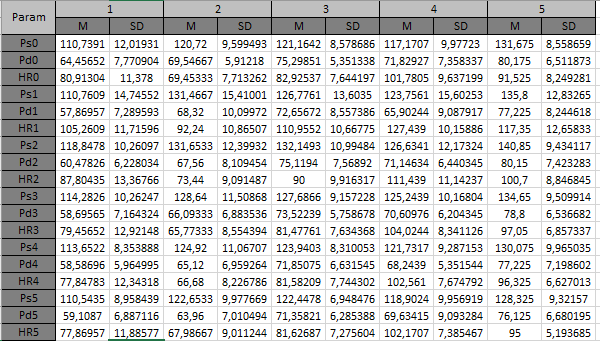
Після кластеризації ми занесли результати до окремої таблиці Excel. Після цього завантажили її до SPSS та провели дисперсійний аналіз.





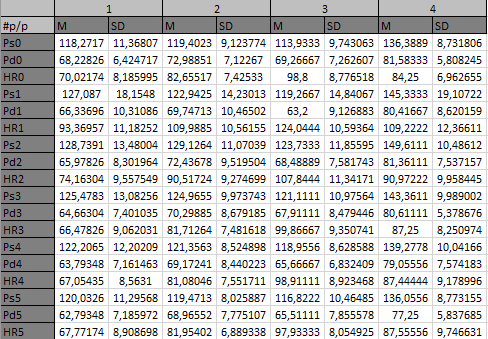


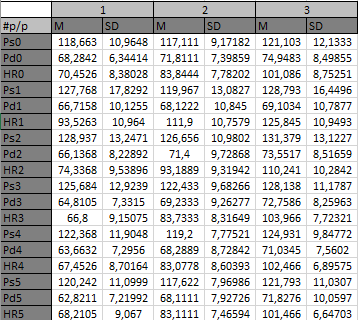
За результатами дисперсійного аналізу було побудовано результуючу таблицю, яка містить показники стандартного відхилення та середніх значень за кожною змінною для кожного кластера.



Результуюча таблиця виступає базою даних для побудови нових графіків.

Аналогічні аналізи при провели для чотирьох і трьох кластерів та отримали наступні таблиці з графіками.





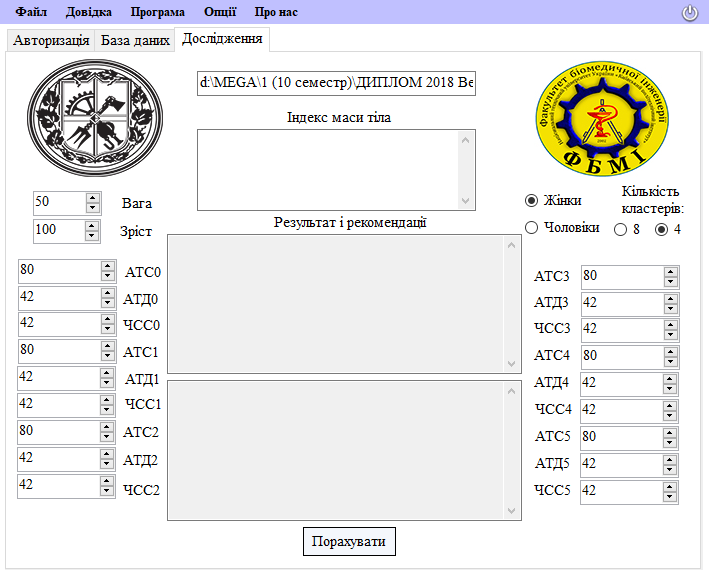
Дивлячись на графіки ми бачимо, що графіки на базі трьох кластерів є найменш інформативними, оскільки ми втрачаємо групу з даними, що показують високі показники систолічного тиску. Графіки на базі чотирьох кластерів показують, що група з низькими показниками систолічного тиску зникає, але в цілому інші графіки показують, що кластери досить різні.

Тому виходячи з графіків ми можемо сказати, що найбільш інформативними є графіки на базі п’яти кластерів. Також ми можемо подивитися як змінюється дисперсія зі зменшенням кількості кластерів

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Параметр | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Внутригруп. | 80,00887 | 78,13669 | 79,88485 | 73,29032 | 41,71111 | 96,80202 | 81,71799 |
| Междугрупп. | 75,93598 |  |  |  |  |  |  |
| Внутригруп. | 106,4002 | 92,65487 | 72,12717 | 89,51328 | 74,038 |  |  |
| Междугрупп. | 86,9467 |  |  |  |  |  |  |
| Внутригруп. | 110,0338 | 84,21743 | 93,54562 | 92,33333 |  |  |  |
| Междугрупп. | 95,03256 |  |  |  |  |  |  |
| Внутригруп. | 107,6799 | 85,97063 | 104,1654 |  |  |  |  |
| Междугрупп. | 99,27197 |  |  |  |  |  |  |

З порівняльної таблиці ми бачимо, що дисперсія досить сильно змінюється при зменшенні кластерів на 2 і 3 позиції, але при зменшенні на 4, дисперсія майже не змінилася, тому вибирати найоптимальнішу кількість виходячи з аналітичної частини треба серед таблиць на 5 і 4 кластери.

Виходячи з графіків та дисперсійного аналізу чітко видно, що таблиці на 5 кластерів дають найоптимальніший результат, тому нами було вирішено розширити функціонал програмного продукту Clusterbox і додати до нього можливість вибору результуючої таблиці на 5 кластерів. Таким чином ми надаємо вибір автоматичного застосування різних таблиць, просто вибираючи їх у меню програми.



# ВИСНОВОК

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. A.K. Jain, M.N. Murty, P.J. Flynn – “Data Clustering: A Review”

[Електронний ресурс] – Режим доступу: http://www.csee.umbc.edu/ nicholas/clustering/p264-jain.pdf

1. J. Kogan, C. Nicholas, M. Teboulle – “Clustering Large and High Dimensional data”

[Електронний ресурс] – Режим доступу: http://www.csee.umbc.edu/ nicholas/clustering/tutorial.pdf

1. Кластеризация данных – Автор: Александр Котов, Николай Красильников, - 2 октября 2006 г.

[Електронний ресурс] – Режим доступу: http://yury.name/internet/02ia-seminar-note.pdf

1. Алгоритмы кластеризации на службе Data Mining

[Електронний ресурс] – Режим доступу: https://basegroup.ru/community/articles/datamining

1. Кластерный анализ

[Електронний ресурс] – Режим доступу: http://statlab.kubsu.ru/sites/project\_bank/claster.pdf

1. Кластерный анализ

[Електронний ресурс] – Режим доступу: http://statsoft.ru/home/textbook/modules/stcluan.html

1. K-means

[Електронний ресурс] – Режим доступу: https://ru.wikipedia.org/wiki/K-means

# ПРЕЗЕНТАЦІЯ ДО ЗАХИСТУ ПЕРЕДДИПЛОМНОЇ ПРАКТИКИ