

ВСТУП

Систематичний самоконтроль стану здоров'я студента дозволяє оптимізувати роботу факультету фізичного виховання НТУУ «КПІ ім. Ігоря Сікорського» та дозволяє підібрати найбільш оптимальний вид спорту, що сприятиме поліпшенню функціонуванню системи організму. Одним із найпоширеніших методів спостереження за фізичним станом організму є функціональна проба Мартіне, яка дає можливість відслідковувати динаміку зміни артеріального тиску та частоти серцевих скорочень між станом спокою та на кожній хвилині після навантаження, включно до п'ятої. У більшості випадків характеристика зміни пульсу і тиску під час функціональної проби досить точно відображає стан серцево-судинної системи у студента. Визначення характерних закономірностей станів системи кровообігу дозволить розробити механізм порівняння цих станів а також вдосконалити алгоритм для оцінки функціонального стану системи кровообігу.

Актуальним є моніторинг фізіологічного стану протягом усього періоду фізичних і спортивних тренувань з періодичним визначення регуляторних реакцій на тестове навантаження. Це потребує розробки нових модулів для досліджень та вдосконалення програмного продукту для визначенню функціонального стану. Результатом використання програми є пришвидшення роботи лікарів та визначення діагнозу без довготривалої затримки.

Мета роботи: вдосконалення системи реєстрації змін функціонального стану системи кровообігу шляхом порівняння параметрів тесту із заздалегідь відомими функціональними схемами кровообігу з подальшим розробленням системи порівняння функціональних патернів.

У відповідності з метою ставлять такі завдання:

- Дослідження існуючої проблеми.
- Проведення глобальної кластеризації.

- Проведення дискримінантного аналізу та логістичної регресії для визначення коректності алгоритму евклідової відстані.
- Поєднання та автоматизація обраних методів.
- Реалізація програмного додатку.
- Зменшення кількості кластерів.
- Побудова нових результуючих таблиць з виведенням графіків.
- Визначення внутрішньогрупової дисперсії.
- Дослідження та вибір оптимальної кількості патернів.
- Вдосконалення програмного продукту.

Об'єкт дослідження. База даних студентів та викладачів Національного Університету України «Київського політехнічного університету ім. Ігоря Сікорського».

Предмет дослідження. Алгоритм розрахунку коефіцієнтів рівняння регресії з виведення графічного матеріалу на екран користувача.

Методи дослідження. Програмний продукт «Clusterbox» та відповідні результуючі таблиці. Для реалізації програмного продукту буде застосовано середовище програмування Microsoft Visual Studio 2017 Community, зокрема використано мову програмування C# та фреймворк .Net Framework 4.5 з використанням елементів Windows Form Application.

Наукова новизна одержаних результатів. Реалізація програмного комплексу для визначення подібності функціональних груп. Розробка методу автоматичної глобальної кластеризації на базі квадрату евклідової відстані з вибором кількості кластерів. Розробка модулю універсальної кластеризації, що передбачає задати будь-яку кількість змінних та кластерів для дослідження.

Практичне значення одержаних результатів. Програмний продукт може бути використаний в медичних клініках та учбових закладах для оцінки функціональних реакцій організму. Насамперед використання програми передбачено на факультеті фізичного виховання Національного

технічного університету України «Київського політехнічного університету ім. Ігоря Сікорського».

Апробація результатів дисертації. Результати досліджень були оприлюднені у доповіді «Оцінка функціональних реакцій на тестове навантаження у студентів 1-2 курсу. Жінки. Чоловіки» на конференції «Презентація наукових розробок студентів і аспірантів» за участі корейської делегації від 28 вересня 2017, що проходила в корпусі №1, кімнаті № 155 Національного технічного університету України «Київського політехнічного університету ім. Ігоря Сікорського».

Публікації. Результати магістерської дисертації описані в 4 статтях: «Estimation of Algorithms Efficiency in the Task of Biological Objects Clustering» та опубліковані в журналі Innovative biosystems and bioengineering, vol. 2 · no. 2; «Automated Assessment of a Students Circulatory System Functional State Using Martine's Test» та опубліковані в журналі Innovative biosystems and bioengineering, vol. 2 · no. 3; «Застосування алгоритму знаходження мінімальної відстані для визначення групи ризику студента» та опубліковані в журналі The scientific heritage № 23 (23), 2018; «Порівняння систем прогнозування та алгоритму знаходження мінімальної відстані для визначення групи ризику студента» та опубліковані в журналі The scientific heritage № 23 (23), 2018.

Структура дисертації. Дисертація побудована за класичним типом та викладена на **208** сторінках машинописного тексту. Складається з вступу, **4** розділів, висновків, списку використаних літературних джерел, який містить **187** найменувань, **6** – на кирилиці, **181** – на латиниці. У роботі представлено **15** рисунків і **45** таблиць.

РОЗДІЛ 1

ОГЛЯД ЛІТЕРАТУРИ ЗА ТЕМОЮ І ВИБІР НАПРЯМІВ ДОСЛІДЖЕНЬ

1.1. Алгоритм квадрату евклідової відстані

Квадрат евклідової відстані є однією із мір відстані, що використовуються для встановлення подібності або відмінності об'єктів класифікації. Зазвичай використовують просту евклідову відстань, що в багатовимірному просторі є геометричною. Але, якщо ознаки досліджуваних об'єктів були виміряні в різних одиницях, то евклідова відстань може втратити сенс. Тому для ефективного використання даного алгоритму доцільно проводити нормування ознак кожного об'єкту дослідження.

Використання квадрату евклідової відстані виправдане в тих випадках, коли надання більшого значення більш віддаленим об'єктам один від одного підвищує якість класифікації об'єктів. У тому випадку, коли слід дослідити відстань між об'єктами, які за однаковим набором змінних є різними, то доцільно використовувати квадрат евклідової відстані.

Прикладом використання квадрату евклідової відстані є різноманітні алгоритми кластеризації. Зазвичай комп'ютерні програми використовують евклідову відстань за замовчуванням в стратегіях об'єднання або в методі деревоподібної кластеризації. Невід'ємною частиною використання даного алгоритму є статистика, зокрема пакет для обробки даних – IBM SPSS. Також важливу роль квадрата евклідової відстані грає в алгоритмах оптимізації та машинному навчанні.

Ще одним прикладом використання алгоритму є реалізація експертних систем прийняття рішень, серед яких є навіть система для вибору дипломного керівника для студента. Критерій вибору будується по характеристикам студентів з вказанням їх значень, але самі характеристики

студенти надають індивідуально та самостійно. За наданими характеристиками будується об'єкт класу «Образ». Задача пошуку наукового керівника полягає у визначенні найбільш близького екземпляру класу «Викладач», що був побудований за наданими характеристиками викладача, для сформованого об'єкта «Образ». Якщо кожен екземпляр класу «Викладач» розглядати як окремий кластер, тоді задача зведеться до кластеризації наступного типу: заданий об'єкт класу «Образ» необхідно класифікувати в один із кластерів виходячи з деякої міри близькості. Подібність між об'єктом класифікації та кластерами визначається у залежності від метричної відстані між ними. Таким чином задача класифікації зводиться до задачі визначення функції близькості між об'єктами даних класів – вибір міри відстані між об'єктами. Оскільки відстань Махаланобіса доцільно використовувати у випадку, коли кореляція між змінними є ненульовою, то ефективним для даної задачі є використання квадрату евклідової відстані.

Відстань між двома об'єктами складається із суми різниці значення ознак двох об'єктів. Так як одна ознака може характеризуватися декількома значеннями, то кожний доданок може бути представлений у виді суми різниці декількох значень однієї ознаки або кожне значення (кожна характеристика) однієї ознаки може бути розглянуто як значення окремої ознаки.

Таким чином для реалізації критерію вибору наукового керівника було побудовано експертну систему, що в своїй сутності використовує квадрат евклідової відстані. База даних експертної системи включає характеристики викладачів. Дерево рішень представляє собою набір правил для обрахунку функції близькості між об'єктами, наведеними в базі даних, і об'єктом класу «Образ», характеристики якого визначає студент (користувач експертної системи). Найдені значення функції близькості задають рейтинг для кожного об'єкта класу «Викладач».

Кожне рішення експертної системи складається зі списку об'єктів класу «Викладач» із вказанням їх характеристик, а також рейтингу даного об'єкта. Рейтинг об'єкта визначається за формулою евклідової відстані між цим об'єктом і об'єктом класу «Образ».

Також яскравим прикладом використання алгоритму квадрату евклідової відстані є задачі з області біології та медицини. Наприклад, розроблена система «Clbsterbox» дозволяє дослідити функціональний резерв організму людини та визначити якість його відновлення після фізичних навантажень. Зокрема програма дозволяє оцінити стан системи кровообігу за пробою Мартіне. Сама проба полягає у наступному: необхідно виміряти значення артеріального тиску до навантаження і записати дані до окремої таблиці. Після цього зробити 20 присідань за 30 секунд – стандартне фізичне навантаження. Наступним кроком є замір значень артеріального тиску та пульсу на кожній хвилині після навантаження. Дані заміри необхідно робити протягом п'яти хвилин та заносити значення до таблиці.

Таким чином оцінюється пристосування організму до різних видів фізичних навантажень з різною інтенсивністю. Серед всіх реакцій серцево-судинної системи виділяють 5 основних типів, які представлені на рис. 1.1.

Тип реакції	Збудженість пульсу	Час відновлення пульсу	Зміна АТ		Час відновлення АТ
			Систолічний	Диастоліний	
Нормотонічний	до 80 %	до 3 хв	до +40	0,-5,-10	до 3 хв
Гіпертонічний	більше, ніж 80 %	Більше 3 хв	значно підвищується	значно підвищується	Більше, ніж 3 хв
Дистонічний	більше, ніж 80 %	Більше 3 хв	значно знижується	Значительно знижується	Більше, ніж 3 хв
Астенічний	Більше, ніж 80 %	Більше 3 хв	не значно змінюється	не значно змінюється	Більше, ніж 3 хв
Пороговий	більше, ніж 80 %	Більше 3 хв	підвищується на 2 хв	підвищується на 2 хв	Більше, ніж 3 хв

Рисунок 1.1. Основні типи реакції серцево-судинної системи на фізичні навантаження

Зазвичай проба Мартіне використовується у спортивних сферах, для визнання фізичних можливостей осіб, швидкості відновлення організму

після та підчас фізичних навантажень різних видів. В клінічній практиці цей метод дозволяє вивчати функціональні можливості серцево-судинної системи враховуючи вікову категорію піддослідного. Виходячи з практичного досвіду, були встановлені оптимальні норми необхідної фізичної активності для певних вікових категорій при використанні проби Мартіне. Таким чином:

- для осіб до 40 років без виражених відхилень у стані здоров'я – 20 присідань на 30 секунд;
- для осіб до 50 років – 15 присідань на 22 секунд;
- для осіб більше 50 років без виражених відхилень – 10 присідань на 15 секунд.

Якщо результати проби вкладаються в нормологічний тип реакції то й функціональний стан серцево-судинної системи вважається задовільним.

Існує можливість використання проби Мартіне і в діагностичних цілях. Наприклад – якщо показники проби потрапляють до несприятливого типу реакції, це є основою вважати що тахікардія в стані спокою є показником захворювань серцево-судинної системи. У випадку, коли до навантажень пульс є стабільним, а відновлення відбувається хвилеподібно, може виникнути негативна фаза пульсу. Нерідко виникає ситуація коли пульс нормалізується на показниках нижчих, ніж ті, що були до навантаження – це є підставою вважати що тахікардія підчас стану спокою зумовлена порушеннями нервової системи. Якщо до навантаження ЧСС вище норми, а після проби всі показники входять до нормологічного типу реакції, але, при цьому, пульс є підвищеним чи таким же як і до навантаження - припускають, що тахікардія в спокої зумовлена гіперфункцією щитовидної залози. Подальші обстеження можуть уточнити, а частіше підтвердити результати функціональних проб.

Після проведення проби Мартіне для нашого дослідження буде отримано 18 значень, що характеризують динаміку зміни артеріального тиску. За даними значеннями можна побудувати графіки ЧСС-АТС та ЧСС-

АТД і побачити як відновлюється організм після фізичних вправ. Якщо значення до навантаження не близькі до значень після навантаження, тоді це свідчить про погану відновлюваність організму та потребує внесення певних змін до способу життя пацієнта.

Для дослідження групи ризику студента доцільно використовувати програмний продукт «Clusterbox». В його наявності є 7 груп, що характеризують стан відновлення кровоносної системи для чоловіків та 8 – для жінок. Кожна група містить свої значення артеріального тиску до навантаження та значення на кожній хвилині після навантаження, включно до п'ятої хвилини. Всі вони занесені до спеціальної таблиці, що називається результуючою таблицею, яка поставляється разом з програмним продуктом. Також кожна група містить свої рекомендації щодо покращення стану системи кровообігу та відповідні характеристики.

Оскільки у наявності програмного продукту є 7 груп чоловічої статі та 8 жіночої, тоді задача полягає у визначення до якої з груп відноситься пацієнт зі своїми значеннями тесту. З цього випливає звичайна задача класифікації об'єкта (розглянуто з боку чоловіків): у наявності 7 класифікаційних груп та один об'єкт, що необхідно класифікувати. Для цього треба визначити на якій відстані знаходиться об'єкт до кожної класифікаційної групи та знайти мінімальну відстань. Алгоритм знаходження мінімальної відстані побудований на основі алгоритму квадрату евклідової відстані і працює наступним чином: від показника студента, що досліджується віднімаємо середнє значення в групі, а результат підносимо до квадрату. Дану процедуру повторюємо 18 разів, оскільки в нас 18 показників. Після проведення цієї процедури всі результати додаються і отримане значення стає відстанню до центру групи (квадрат евклідової відстані). Такі обчислення проводяться для кожної групи окремо, а відстань, яка буде мінімальною, характеризуватиме групу, до якої відноситься студент.

Результатом роботи алгоритму та програми в цілому є виведення знайденої групи і її характеристик на екран користувача. Таким чином можна буде дослідити стан фізичного здоров'я окремого пацієнта.

Проблемою даного програмного продукту є неможливість визначити групу ризику для декількох студентів або цілої бази даних і збереження результатів до окремої таблиці. Також проблематика полягає у неможливості порівняти групи між собою. Ми можемо побудувати графіки зміни артеріального тиску і побачити розташування даних груп у просторі, але це не дає змоги оцінити їх подібність або відмінність у повному обсязі. Для цього достатньо було б реалізувати алгоритм лінійної регресії для додаткової інтерпретації класифікаційних груп. Ще одним недоліком програми є те, що вона працює лише з однією таблицею, яка йде в комплекті з нею, а також немає можливості вибору кількості змінних для дослідження. Наприклад, ми б хотіли дослідити відновлення організму на шести хвилинах, але програмним продуктом просто не передбачено такого функціоналу.

Таким чином ми бачимо, що програмний продукт є досить корисним, але його функціонал дуже обмежений. Він не є універсальним та не може бути використаний для інших досліджень, тому необхідним є розроблення додаткових модулів для дослідження, розширення функціоналу та можливостей продукту.

1.2. Дисперсійний аналіз

Дисперсійний аналіз - аналіз мінливості ознаки під впливом будь-яких контрольованих змінних факторів (ANOVA - «Analysis of Variance»). Основною метою дисперсійного аналізу є дослідження значущості відмінності між середніми.

Мета дисперсійного аналізу - дослідження наявності або відсутності істотного впливу будь-якого якісного або кількісного фактору на зміни

досліджуваної результативної ознаки. Для цього фактор, який імовірно має або не має істотного впливу, поділяють на класи (інакше кажучи, групи) і з'ясовують, чи однаковий вплив фактору шляхом дослідження значущості між середніми в наборах даних, у відповідних градаціях фактору. При дисперсійному аналізі визначають питому вагу сумарного впливу одного або декількох факторів. Істотність впливу фактору визначається шляхом перевірки гіпотез:

$H_0 : \mu_1 = \mu_2 = \dots = \mu_a$, де a – число класів градації – всі класи градації мають одне значення середніх, H_1 : не всі μ_i рівні – не всі класи градації мають одне значення середніх.

Якщо вплив фактору не суттєвий, то несуттєва і різниця між класами градації цього фактору і в ході дисперсійного аналізу нульова гіпотеза H_0 не відкидається. Якщо вплив фактору істотний, то нульова гіпотеза H_0 відхиляється: не всі класи градації мають одне і теж середнє значення, тобто серед можливих різниць між класами градації одна або кілька є суттєвими.

Схематично дисперсійний аналіз поділяється на певні категорії, які визначаються у залежності від кількості факторів, які беруть участь у дослідженні, кількості змінних, на які може впливати фактор та від співвідношення вибірок значень між собою. Якщо при аналізі наявний лише один фактор, вплив якого досліджується, тоді такий аналіз називається однофакторним та поділяється на два види:

- Аналіз вибірок, що не пов'язані між собою (аналіз різних, незв'язних вибірок)
- Аналіз вибірок, що пов'язані між собою. Наприклад, коли на одній і тій самій групі проводять декілька вимірів, але в різних умовах

Якщо при аналізі необхідно дослідити одночасний вплив двох і більше факторів, тоді такий аналіз називається багатофакторним, який в свою чергу ділиться на категорії за типом вибірки.

1.2.1. Однофакторний дисперсійний аналіз

У разі однофакторного дисперсійного аналізу мається на увазі, що середні генеральних сукупностей, з яких були вилучені вибірки, - рівні, іншими словами, всі вони відносяться до однієї генеральної сукупності і відмінності носять випадковий характер. Для перевірки теорій в разі дисперсійного аналізу використовується F-розподіл. F-статистика приймає тільки позитивні або нульові значення.

Процедура дисперсійного аналізу полягає у визначенні співвідношення систематичної (групової) дисперсії до випадкової (внутрішньогрупової) дисперсії в вимірюваних даних. Як показник мінливості використовується сума квадратів відхилення значень параметра від середнього: SS (Sum of Squares). Загальна сума квадратів SSTotal розкладається на міжгрупову суму квадратів SSBG і внутрішньогрупову суму квадратів SSWG : $SSTotal = SSBG + SSWG$

У разі якщо вірна H_0 , то як внутрішньогрупова, так і міжгрупова дисперсії є оцінками однієї і тієї ж дисперсії і повинні бути приблизно рівні.

$$F = \frac{MS_{BG}}{MS_{WG}}, \text{ де} \quad (1.1)$$
$$MS_{BG} = \frac{SSBG}{\nu_{BG}}, MS_{WG} = \frac{SSWG}{\nu_{WG}}$$

Виходячи з цього значення F має бути близько до 1 в разі, якщо статистично значущих відмінностей все-таки немає. Критичне значення F визначається рівнем значущості (зазвичай 0,05 або 0,01) і внутрішньогруповим і міжгруповим числом ступенів свободи (ν). Воно досить складне для обчислення, тому частіше використовуються табличні значення із зазначенням α , ν_{BG} , ν_{WG} .

Міжгрупове число ступенів свободи: $\nu_{BG} = m - 1$.

m – число груп.

Внутрішньогрупове число ступенів свободи: $\nu_{WG} = n - m$.

n – кількість спостережень в кожній з груп.

1.3. Регресійний аналіз

Основною метою регресійного аналізу є прогнозування значення залежної змінної використовуючи змінні, які є незалежними. Також метод регресійного аналізу використовують для побудови моделі, яка описує залежність між змінними, з подальшою оцінкою значимості отриманого рівняння.

До використання методів регресійного аналізу зазвичай переходять, якщо зв'язки між досліджуваними параметрами мають статистичну значимість. Для цього спеціально шукають такий набір функцій, який пов'язує аргументи x_1, x_2, \dots, x_n з результативним показником. Після цього будується рівняння регресії, розраховується коефіцієнт детермінації і аналізується його точність.

Метод регресійного аналізу потребує, щоб аргументи були незалежними та мали нормальний розподіл з константними значеннями дисперсій. Аналіз побудови регресієвих рівнянь має на меті пошук залежності між ознаками, які досліджуються. Знайдене рівняння буде показувати, як у середньому змінюється y при зміні будь-якого з x_n і матиме вигляд:

$$y = f(x_1, x_2, \dots, x_n) \quad (1.2)$$

де y - залежна змінна (вона завжди одна);

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

Аналогічно до дисперсійного аналізу, регресійний називається простим (однофакторним), якщо у дослідженні є одна незалежна змінна. В іншому випадку він є багатофакторним.

У порівнянні з кореляційним аналізом, який показує наскільки істотним є зв'язок, регресійний передбачає, що зв'язок є і шукає модель цього зв'язку, яка описується рівнянням регресії.

У випадку, якщо між змінними можна виразити кількісне відношення у вигляді комбінації обраних змінних, тоді доцільним є використання методу регресії. У такому випадку комбінація змінних буде використовуватися для прогнозування значення, що може приймати залежна змінна, яка обчислюється на значеннях незалежних змінних у певному наборі, який був заданий дослідником. Для найпростішої побудови моделі на базі отриманого рівняння регресії використовують метод лінійної регресії, але, на жаль, він дає значну похибку і тому більшість моделей не можна якісно побудувати за допомогою цього методу.

Для детальнішого опису рівняння, отримане за допомогою регресійного аналізу, необхідно знати умовний закон розподілу результативного показника y . На практиці не завжди вдається отримати дану інформації, тому у більшості випадків використовують пошук апроксимацій для функції $y = f(x_1, x_2, \dots, x_n)$ заснованих на вихідних статистичних даних. У рамках окремих модельних припущень про тип розподілу вектора показників (x_1, x_2, \dots, x_n) може бути отриманий загальний вигляд рівняння регресії

$$f(x) = M(y/x)x = (x_1, x_2, \dots, x_n)^T \quad (1.3)$$

Для відновлення за вихідними даними, що дасть найкращий результат, найчастіше використовують наступні критерії адекватності:[29].

Метод найменших квадратів, згідно з яким мінімізується квадрат відхилення спостережуваних значень результативного

показника $y_i, i = 1, 2, \dots, n$, від модельних значень $\hat{y}_i = f(x_i, \beta)$, де $\beta = (\beta_0, \beta_1, \dots, \beta_n)^T$ коефіцієнти рівняння регресії; x_i - значення вектора аргументів на M спостереженні:

$$\sum_{i=1}^n (y_i - f(x_i, \beta))^2 \rightarrow \min_{\beta} \quad (1.4)$$

Метод найменших модулів, згідно з яким мінімізується сума абсолютних відхилень спостережуваних значень результативного показника від модульних значень $\hat{y}_i = f(x_i, \beta)$, тобто:

$$\sum_{i=1}^n |y_i - f(x_i, \beta)| \rightarrow \min_{\beta} \quad (1.5)$$

Метод мінмакса зводиться до мінімізації максимуму модуля відхилення спостережуваного значення результативного показника y , від модельного значення $f(x_i, \beta)$, тобто:

$$\max |y_i - f(x_i, \beta)| \rightarrow \min_{\beta} \quad (1.6)$$

Найчастіше використовується метод найменших квадратів для побудови моделі регресії, що передбачає мінімізацію суми квадратів відхилень фактичних значень результатного ознаки від його розрахункових значень, тобто:

$$S = \sum_{j=1}^m (y_j - \hat{y}^j) \rightarrow \min \quad (1.7)$$

де y - число спостережень;

$$j = a + b_1 x_{1j} + b_2 x_{2j} + \dots + b_n x_{nj} - \text{розрахункове значення}$$

результатного фактору.

Задачу МНК розв'язують за допомогою оцінки функції регресії, побудованої параметрично, шляхом аналізу залежності однієї величини Y , значення якої (y_i) від класу не випадкових величин X_1, X_2, \dots, X_k .

Припустимо, що нам відомо, що досліджуваний процес y лінійно залежить від параметрів входу x (вихідний параметр процесу, який вивчається від вхідного параметра).

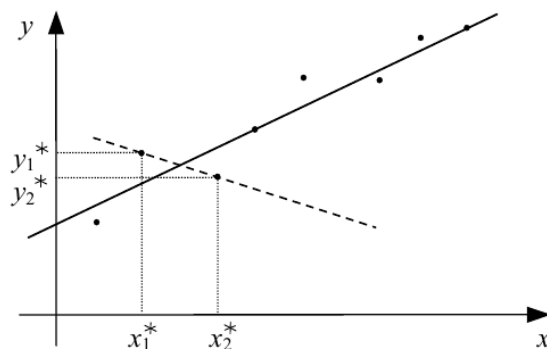


Рисунок 1.2. Графічна інтерпретація причин, які обумовлюють необхідність використання МНК

Таким чином можемо сказати, що даний процес може бути описаний у вигляді наступного рівняння:

$$y = ax + b \quad (1.8)$$

де a і b — коефіцієнти, значення яких обчислюються за допомогою відомих значень, що передаються величині x для визначення вихідної величини y .

Таким чином можна скласти систему з двох алгебраїчних рівнянь:

$$\begin{cases} y_1 = ax_1 + b \\ y_2 = ax_2 + b \end{cases} \quad (1.9)$$

Оскільки будь-які експериментальні вимірювання дають похибку, що обумовлена різними факторами, то для оцінки і зменшення її можна використовувати МНК. Таким чином через наявність похибки в експериментальних даних, розв'язок системи рівнянь буде також складати похибку.

Наприклад, якщо використати лише значення $x_1^*, y_1^*; x_2^*, y_2^*$ (рис. 1.2) для розв'язання системи рівнянь (1.9), то похибка буде вже не у відсотках, а у характері функціональної залежності (пунктирна лінія на рис. 1.2).

Таким чином німецький математик Фрідріх Гаусс запропонував розв'язок системи рівнянь, що направлений на пошук коефіцієнтів a, b моделі (1.9), де необхідним було сформулювати суму квадратів різниць \sum^N між теоретично заданими за допомогою рівняння (1.8) значеннями вихідної координати y при значеннях аргументу $x_i, i = \overline{1, N}$ та її експериментальними значеннями: y_i .

$$\sum^N = \sum_{i=1}^N (y(x_i) - (y_i))^2 \quad (1.10)$$

Наступним кроком був пошук таких значень коефіцієнтів, які мінімізували б рівняння (1.10).

Від цієї процедури і назва методу — метод найменших квадратів.

З курсу вищої математики нам відомо, що мінімум функції знаходиться за допомогою взяття похідної від цієї функції та прирівняти її до нуля.

Згідно з цією ідеєю, підставимо у вираз (1.10) замість $y(x_i)$ його значення з (1.8) і візьмемо від отриманого виразу частинні похідні за b та a , які прирівняємо до нуля, тобто

$$\begin{cases} \frac{\partial \sum_{i=1}^N (ax_i + b - y_i)^2}{\partial b} = \sum_{i=1}^N 2(ax_i + b - y_i) = 0 \\ \frac{\partial \sum_{i=1}^N (ax_i + b - y_i)^2}{\partial a} = \sum_{i=1}^N 2(ax_i + b - y_i)x_i = 0 \end{cases} \quad (1.11)$$

$$\begin{cases} \frac{\partial \sum_{i=1}^N (ax_i + b - y_i)^2}{\partial b} = \sum_{i=1}^N 2(ax_i + b - y_i) = 0 \\ \frac{\partial \sum_{i=1}^N (ax_i + b - y_i)^2}{\partial a} = \sum_{i=1}^N 2(ax_i + b - y_i)x_i = 0 \end{cases} \quad (1.12)$$

За допомогою математичних перетворень можемо отримати систему нормальних рівнянь Гаусса (1.13). Розв'язок даної системи дозволить знайти такі значення параметрів, які будуть мінімізувати систему рівнянь.

$$\begin{cases} cN + b \sum_{i=1}^N x_i + a \sum_{i=1}^N x_i^2 = \sum_{i=1}^N y_i \\ c \sum_{i=1}^N x_i + b \sum_{i=1}^N x_i^2 + a \sum_{i=1}^N x_i^3 = \sum_{i=1}^N y_i x_i \\ c \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i^3 + a \sum_{i=1}^N x_i^4 = \sum_{i=1}^N y_i x_i^2 \end{cases} \quad (1.13)$$

1.4. Дискримінантний аналіз

За своєю природою дискримінантний аналіз схожий на метод регресійного аналізу. Його часто використовують у тому випадку, коли є набір класифікованих даних і необхідно знайти хоча б одну функцію, яка виступатиме у ролі моделі, здатну віднести певне дослідження до однієї з визначених груп. Таким чином основною метою дискримінантного аналізу є класифікації даних - визначення класу, до якого належить новий об'єкт.

Дискримінантний аналіз передбачає наявність навчальної вибірки, де заздалегідь відомо який об'єкт до якого класу відноситься. На навчальній вибірці будується модель, яка в майбутньому дозволяє класифікувати нові об'єкти.

У ролі дискримінантного аналізу найчастіше береться лінійна функція записана у вигляді формули (1.14):

$$Z = C_1 X_1 + C_2 X_2 + \dots + C_m X_m, \quad (1.14)$$

де X_1, X_2, \dots, X_m – значення ознак у даного об'єкта;

C_1, C_2, \dots, C_m – дискримінантні 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 A -статистику Уїлкса або ks^2 – квадрат [5].

Критерій значення Уїлкса обчислюють за формулою (1.15):

$$\Lambda_k = \prod_{i=k+1}^K \frac{1}{1 + \lambda_i}, \quad (1.15)$$

де K – кількість значень;

k – число вже обчислених дискримінаційних функцій.

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

Значення ks^2 -квадрат розраховують за формулою (1.16):

$$\chi^2 = - \left[n - \frac{p+K}{2} \right] \ln \Lambda_k, \quad (1.16)$$

де p – кількість членів у дискримінантній функції, виключаючи вільний член функції.

Якщо це значення більше критичного із заданим рівнем значущості і числом ступенів свободи $(p-k)$ $(K-k-1)$, то значимість підтверджується.

Канонічна дискримінантна функція для загального випадку k класів записана у формулі (1.17):

$$f_{ki} = u_0 + \sum_{j=1}^p u_j X_{jk}, \quad (1.17)$$

де f_{ki} — значення канонічної дискримінантної функції для 1-го об'єкта в k -му класі;

u_j — шукані коефіцієнти дискримінантної функції;

X_{jki} — значення дискримінантної змінної X_j 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 не 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 f_{ki} , 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.

Висновки до розділу 1

Таким чином в даному розділі було розглянуто застосування алгоритму квадрату евклідової відстані в різних областях науки. Наведено приклад застосування в задачах біології та медицини, а також надано оцінку та характеристику програмному продукту для визначення групи ризику студента. Окрім даної системи аналогів знайдено не було. Ми визначили вектор направлення нашої наукової діяльності, розглянули теоретичні відомості щодо досліджень, які будуть проведені для покращення та підтвердження істинності попередніх досліджень Також нами було виявлено необхідні проблеми, які треба розв'язати шляхом розробки нових модулів для програмного продукту «Clusterbox».

РОЗДІЛ 2

МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕННЯ

2.1. Результуюча таблиця програмного продукту «Cluserbox».

Результуюча таблиця, що використовується за замовчуванням програмним продуктом «Clusterbox», побудована на базі даних студентів молодших курсів Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського», що містить 1495 досліджень. У її наявності є 669 обстежень жіночої статі та 826 спостережень чоловіків, що показують стан системи кровообігу. Кількість груп в результативній таблиці для кожної статі окремо було визначено за допомогою кластерного аналізу, що був проведений Настенко Є.А. та Носовець О.К.

Аналіз кластеризації – це багатомірна процедура з статистики, яка направлена на збір даних, що несуть інформацію про об'єкти, а точніше, про їх вибірку, з подальшим впорядкуванням об'єктів в порівняно-однакові класи, групи чи кластери. Кластеризація є одним із методів, що відносять до класу навчання без учителя [12].

Зазвичай до аналізу кластеризації до нього неможливо застосувати алгоритми перевірки статистичної значимості, проте його результат надає найбільш можливі вагомні значення. Не маючи апріорних гіпотез щодо класу даних. Дослідник може використати цю особливість та застосувати алгоритм кластеризації.

В свою чергу, кластеризація ставить за мету структурування отриманих даних. Тобто їх організації в певну структуру. Сам кластерний аналіз включає в себе набір декількох різних алгоритмів, які направлені на виконання задачі класифікації.

При класифікації великих масивів даних на групи кластерний аналіз є інструментом, який неможливо замінити, оскільки він дає змогу зробити це без особливих проблем [13].

Алгоритм кластеризації, що був застосований для визначення кількості груп носить назву – метод кластеризації к-середніх, і являє собою версію ЕМ-алгоритму, котрий також застосовується для розділення суміші гаусової функції. Він розбиває множину елементів векторного простору на завчасно відоме число кластерів k . Робота алгоритму зводиться до мінімізації середньоквадратичного відхилення на точках кожного кластеру. Основна ідея – це те, що на кожній ітерації перераховується центр кластерів для кожного кластеру, котрий був отриманий на попередньому кроці.

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (2.1)$$

Після цього вектори розбиваються на кластери знову, згідно з тим, який з обраних центрів виявився ближче до метрики [14].

В результаті отримано стовпчик з номерами кластерів, що характеризують групу ризику. База даних, що була використана для побудови результуючої таблиці для використання алгоритму визначення оптимальних характеристик студента, вже містила відповідний стовпчик з групами та представляла собою файл формату *.xls (рис.2.1)

1	2	4	15	16	17	18	19	20					
№	ПІБ	СІ8_АРНР	Стать	АТС покій	АТД покій	Пульс пок	пульс ОТН	Індекс Кер	Діаст	Сист.мех	ДПДВ	СПДВ	ИЖМ
1	Авраменко К.Ю.	4	2	120	75	66	0,807754	-13,6364	0,537564	0,311527	40,31727	37,38327	1,078484
2	Агеева О.Д.	3	2	109	62	95	1,162676	34,73684	0,315832	0,255747	19,58156	27,87646	0,702441
3	Адамак В.В.	1	1	130	85	82	1,003573	-3,65854	0,395834	0,275873	33,6459	35,86351	0,938165
4	Азарх Д.П.	1	1	130	78	89	1,089244	12,35955	0,349852	0,264306	27,28843	34,35973	0,794198
5	Андрійчук В.К.	4	1	133	73	76	0,930141	3,947368	0,441989	0,287484	32,26523	38,2354	0,843858
6	Анікеєнко Л. В.	4	2	130	79	53	0,648651	-49,0566	0,715728	0,356347	56,54254	46,32513	1,220559
7	Антоненко А.Ю.	4	1	140	81	51	0,624173	-58,8235	0,7512	0,365271	60,8472	51,13788	1,189865
8	Антонюк О. В.	1	1	138	99	86	1,052528	-15,1163	0,368642	0,269033	36,49554	37,12649	0,983005
9	Антонюк К.В.	1	2	132	79	103	1,260586	23,30097	0,276637	0,245887	21,85431	32,45713	0,673329
10	Апанасенко Д.О.	4	1	106	70	57	0,697606	-22,807	0,652253	0,340379	45,65768	36,08017	1,265451
11	Бабич О.В.	3	1	114	65	93	1,138199	30,10753	0,326684	0,258477	21,23445	29,46643	0,720632
12	Багінський К.С.	5	1	142	80	64	0,783277	-25	0,560263	0,317238	44,821	45,04773	0,994967
13	Байрук М.б.	6	2	106	63	68	0,832231	7,352941	0,5162	0,306153	32,5206	32,45221	1,002107
14	Баран В.В.	6	1	138	91	112	1,370734	18,75	0,239236	0,236479	21,77045	32,63404	0,667109
15	Баранчук А.А.	7	2	103	63	69	0,84447	8,695652	0,505983	0,303583	31,8769	31,26901	1,019441
16	Бардяк М.М.	4	1	129	78	68	0,832231	-14,7059	0,5162	0,306153	40,2636	39,49373	1,019493
17	Барченко А.О.	3	2	132	85	104	1,272824	18,26923	0,272162	0,244762	23,13373	32,30852	0,716026
18	Безсмертна А.В.	8	2	114	70	56	0,685367	-25	0,667271	0,344157	46,709	39,23391	1,190526
19	Бердичівська В.С.	3	2	116	80	88	1,077005	9,090909	0,355973	0,265845	28,47782	30,83807	0,923463
20	Бітко С.М.	5	1	151	93	78	0,954618	-19,2308	0,425815	0,283415	39,60083	42,79572	0,925346
21	Богданов Є.Ю.	4	1	144	80	75	0,917902	-6,66667	0,4504	0,2896	36,032	41,7024	0,864027
22	Боднарчук В.С	6	2	104	70	67	0,819993	-4,47761	0,526722	0,3088	36,87057	32,1152	1,148072
23	Бойко Р.Д.	4	1	126	78	79	0,966857	1,265823	0,418035	0,281458	32,60676	35,46374	0,91944
24	Бойко Р.С.	5	1	122	81	82	1,003573	1,219512	0,395834	0,275873	32,06257	33,65653	0,95264
25	Бондарук А.В	3	2	111	69	87	1,064767	20,68966	0,362234	0,267421	24,99418	29,6837	0,842017
26	Бондарук І.М.	4	2	114	78	70	0,856709	-11,4286	0,496057	0,301086	38,69246	34,32377	1,127279
27	Бондарь Р.І.	1	1	146	84	96	1,174915	12,5	0,310575	0,254425	26,0883	37,14605	0,702317
28	Боришкевич В.Ю	2	1	138	76	87	1,064767	12,64368	0,362234	0,267421	27,52982	36,90406	0,745984

Рисунок 2.1. Приклад бази даних студентів

База даних також містила показники дихальної системи (ємність легень, проба Генче, проба Штанге, частота дихання), нервової системи (проста зорова-моторна реакція, складна зорова-моторна реакція, режим нав'язаного ритму та ін.), кровоносної системи (артеріальний тиск, пульс, індекс Кердо та ін.) та показники фізичного стану (вік, зріст, вага, індекс маси тіла та ін.). Але для дослідження було обрано показники кровоносної системи, зокрема, артеріальний тиск та пульс у стані спокою, а також на 1-5 хвилих після навантаження.

Для дослідження бази даних Excel необхідно привести її до прийнятного виду, після чого можемо запустити базу в SPSS та провести наступні операції перед початком аналізу БД:

- Встановлення імені змінних
- Встановлення мітки відповідності змінних
- Встановлення числовим змінним шкалу «кількісну» [28].

	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропущенн...	Ширина ...	Выравнивание	Шкала	Роль
1	V1	Числовая	12	0	№	Нет	Нет	12	По право...	Количество...	Входная
2	V2	Текстовая	41	0	ПИБ	Нет	Нет	41	По левом...	Номинальная	Входная
3	V3	Числовая	12	0	№	Нет	Нет	12	По право...	Количество...	Входная
4	V4	Числовая	12	0	C18_APHR	Нет	Нет	12	По право...	Количество...	Входная
5	V5	Числовая	12	0	C112_APHR	Нет	Нет	12	По право...	Количество...	Входная
6	V6	Числовая	12	0	C114_APHR	Нет	Нет	12	По право...	Количество...	Входная
7	V7	Числовая	12	0	Кластер по полу	Нет	Нет	12	По право...	Количество...	Входная
8	V24	Числовая	12	0	Пол	{1, Мужско...	Нет	12	По право...	Количество...	Входная
9	pulse_1min...	Числовая	8	4	Отношения пу...	Нет	Нет	8	По право...	Количество...	Входная
10	pulse_2min...	Числовая	8	4	Отношения пу...	Нет	Нет	8	По право...	Количество...	Входная
11	pulse_3min...	Числовая	8	4	Отношения пу...	Нет	Нет	8	По право...	Количество...	Входная
12	pulse_4min...	Числовая	8	4	Отношения пу...	Нет	Нет	8	По право...	Количество...	Входная
13	pulse_5min...	Числовая	8	4	Отношения пу...	Нет	Нет	8	По право...	Количество...	Входная
14	V23	Числовая	12	0	Вік	Нет	Нет	12	По право...	Количество...	Входная
15	V25	Числовая	12	0	АТС покій	Нет	Нет	12	По право...	Количество...	Входная

Рисунок 2.2. База даних готова до аналізу

Наступним кроком є виведення описових статистик за допомогою програми IBM SPSS Statistics 20. Для цього необхідно вибрати вкладку «Аналіз», де зі списку обрати «Описові статистики»-«Частоти».

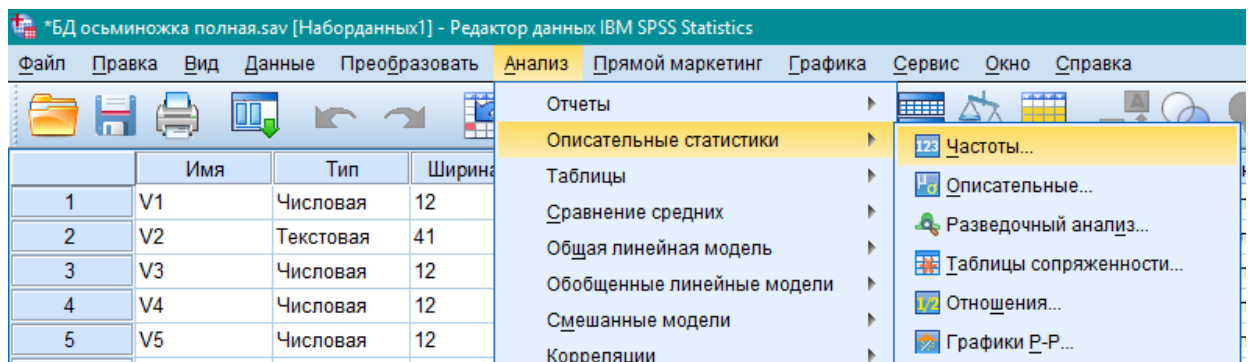


Рисунок 2.3. Описові статистики

У вікні «Частоти» в нашому випадку необхідно обрати лише пункти «Максимум», «Мінімум», «Середнє значення» та «Стандартне відхилення» для того, щоб визначити верхню та нижню границі для систолічного артеріального тиску, діастолічного артеріального тиску та частоти серцевих скорочень. У якості змінних візьмем лише АТС, АТД та ЧСС у стані спокою та ці ж змінні за кожну хвилину після навантаження [29].

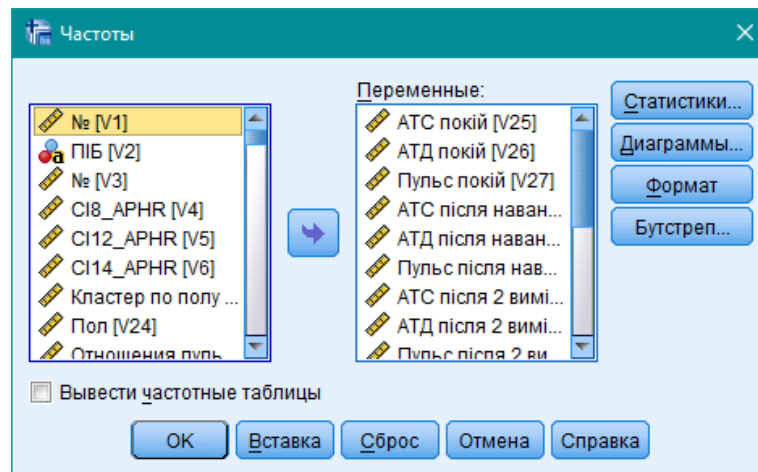


Рисунок 2.4. Вибір необхідних параметрів

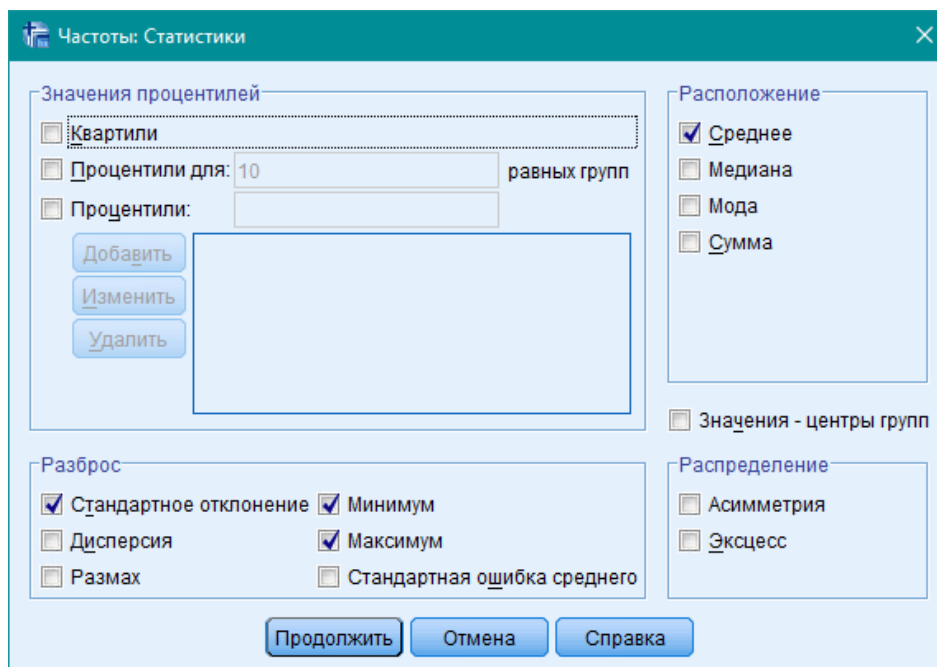


Рисунок 2.5 – Вибір необхідних статистик [28, 29].

Проведений аналіз дав змогу встановити, що максимальне значення АТС приймає відмітку в 219 мм рт. ст., АДТ приймає значення в 129 мм рт. ст., а значення пульсу – 155 ударам за хвилину. Мінімальне значення АТС рівне 80 мм рт. ст., АДТ - 42 мм рт. ст.. Мінімальне значення ЧСС збігається зі значенням АДТ і також дорівнює 42 [29, 30]. Ці значення використовуються для задання діапазону при введенні даних до відповідних полів в програмному продукту «Clusterbox».

Оскільки таблицю було перетворено до прийняттого вигляду, який без проблем сприймається програмою для обробки даних в SPSS, то на БД в 1495 спостереженнями було проведено дисперсійний аналіз для визначення середнього значення для кожної змінної в кожному кластері.

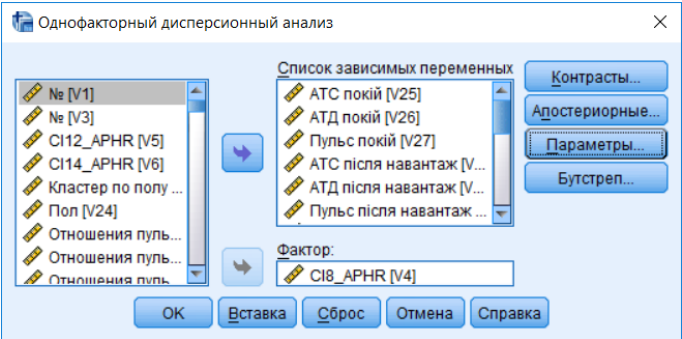


Рисунок 2.6. Вибір необхідних параметрів для аналізу

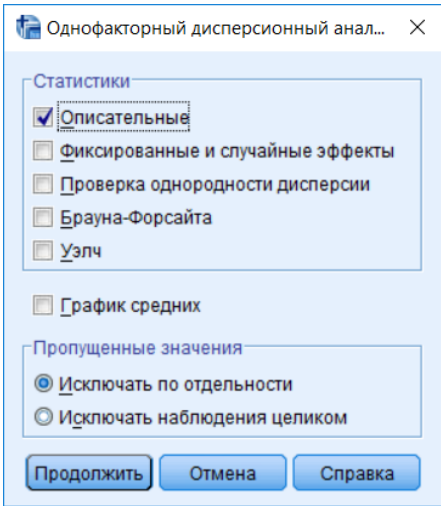


Рисунок 2.7. Вибір потрібних статистик

Результатом проведення дисперсійного є наступна таблиця:

Таблица 2.1

Описові статистики по кластерам

		N	Середнє	Стд. відхилення	Стд. Похибка
АТС покой	1	113	135,84	9,647	,908
	2	130	119,08	9,518	,835
	3	106	124,02	9,499	,923
	4	156	131,24	9,641	,772

	5	44	137,68	12,649	1,907
	6	46	142,39	13,807	2,036
	7	199	117,49	9,740	,690
	Всього	794	126,50	13,033	,463
АТД покій	1	113	80,23	6,291	,592
	2	130	70,73	6,193	,543
	3	106	76,43	6,243	,606
	4	156	74,44	5,686	,455
	5	44	83,52	6,048	,912
	6	46	86,59	8,676	1,279
	7	199	67,72	6,237	,442
	Всього	794	74,45	8,445	,300
Пульс покій	1	113	84,44	6,610	,622
	2	130	88,80	8,443	,740
	3	106	98,94	11,080	1,076
	4	156	69,69	9,037	,724
	5	44	71,75	9,684	1,460
	6	46	99,46	10,899	1,607
	7	199	72,82	8,493	,602
	Всього	794	81,45	14,171	,503

Після цього середнє значення та стандартне відхилення додаються до окремої компактної таблиці.

Пок-ник	CL1		CL2		CL3		CL4	
	М	SD	М	SD	М	SD	М	SD
АТС0	135,84	9,647	119,08	9,518	124,02	9,499	131,24	9,641
АТС1	154,70	18,842	125,13	15,299	131,29	14,327	144,25	14,831
АТС2	148,49	13,435	125,05	11,198	131,41	11,468	139,28	9,200
АТС3	140,12	11,333	120,78	9,397	127,61	9,891	132,22	8,740
АТС4	135,68	10,231	118,25	8,410	123,07	9,255	130,85	9,756
АТС5	132,20	10,482	116,83	8,083	121,17	8,975	126,88	7,644
АТД0	80,23	6,291	70,73	6,193	76,43	6,243	74,44	5,686
АТД1	84,60	9,502	68,51	9,489	75,14	9,024	76,02	7,244
АТД2	81,62	7,814	69,12	7,176	76,95	7,641	73,11	5,992
АТД3	77,93	7,459	67,62	7,353	73,46	5,682	70,19	4,971
АТД4	76,69	6,992	66,62	6,642	71,52	6,886	69,91	6,284
АТД5	75,30	6,565	66,02	6,396	71,44	6,649	68,11	5,131
ЧСС0	84,44	6,610	88,80	8,443	98,94	11,080	69,69	9,037
ЧСС1	107,34	9,108	111,55	9,173	122,70	12,034	89,75	11,557
ЧСС2	91,84	8,128	97,46	8,524	110,14	11,676	74,76	10,368
ЧСС3	88,96	6,820	92,68	7,704	104,86	11,919	70,94	10,500
ЧСС4	88,06	7,154	90,99	8,106	102,77	12,179	70,63	10,643
ЧСС5	86,60	7,452	90,53	8,109	102,43	11,536	70,57	10,870

Рисунок 2.8. Згрупована таблиця результатів

CL5		CL6		CL7		Ст. св.	Знач.	Сума квадратов
M	SD	M	SD	M	SD			
137,68	12,649	142,39	13,807	117,49	9,740	793,00	0,000	134700,49
153,73	22,810	156,46	21,457	123,42	13,779	793,00	0,000	338079,02
145,89	15,579	151,91	16,134	124,54	9,263	793,00	0,000	180130,52
138,95	11,068	144,48	14,077	120,43	8,749	793,00	0,000	131790,79
136,43	12,141	140,98	11,769	117,41	8,045	793,00	0,000	122752,87
134,30	12,324	138,04	13,623	115,91	8,091	764,00	0,000	105827,82
83,52	6,048	86,59	8,676	67,72	6,237	793,00	0,000	56550,17
82,77	9,009	85,67	10,418	65,57	8,020	793,00	0,000	100688,71
81,07	7,726	86,41	9,446	65,63	7,232	793,00	0,000	75864,25
80,43	7,896	84,48	9,227	63,73	7,179	793,00	0,000	67588,51
79,61	8,255	82,59	7,095	62,34	6,902	793,00	0,000	66716,62
78,26	7,384	83,17	5,405	61,59	7,068	764,00	0,000	60617,13
71,75	9,684	99,46	10,899	72,82	8,493	793,00	0,000	159238,17
87,30	12,702	120,46	11,051	96,73	10,692	793,00	0,000	2036660,07
73,16	11,108	110,30	10,407	79,43	10,021	793,00	0,000	217153,74
72,73	13,644	106,50	9,531	72,48	8,676	793,00	0,000	212435,31
70,61	8,224	105,98	7,943	72,28	8,732	793,00	0,000	198600,74
70,35	8,141	104,09	7,871	72,92	9,377	764,00	0,000	183981,60

Рисунок 2.9. Згрупована таблиця результатів

Побудовані таблиці є результуючими та поставляються разом з програмним продуктом.

Для наступних досліджень нами було використано базу студентів та викладачів НТУУ «КПІ ім. Ігоря Сікорського», що пройшли пробу Мартіне декілька разів. Вона містить 599 досліджень, серед яких 323 чоловічої та 276 жіночої. База даних є аналогічною до тієї, за даними якої вже побудовано результуючу таблицю, тому аналогічно її було зведено до прийнятного вигляду для роботи в статистичному пакеті SPSS.

На відміну від попередньої бази, наявна не містить значень групи ризику в повному обсязі, тому ми не можемо охарактеризувати всіх студентів. Також невідомо чи відповідає колонка з кластерами дійсності, оскільки на цей раз не наявні розшифрування та методи отримання кластерів.

Але можливе застосування програмного продукту для визначення групи ризику і заодно є можливість переконатись в справній роботі алгоритму квадрата евклідової відстані. Задача перевірки ефективності алгоритму впливає з того, що попередні дослідження показали високий

відсоток збіжності (80%) між алгоритмом кластеризації методом k-середніх та реалізованим у програмі алгоритмом знаходження мінімальної відстані.

Висновки до розділу 2

У даному розділі нами було розглянуто опис бази даних для попередніх досліджень, на базі якої побудовано результуючу таблицю. Наведено характеристику даної таблиці, метод її побудови та використання. Також було описано вхідну таблицю для майбутніх досліджень та доведено необхідність проведення тестів для підтвердження ефективності алгоритму евклідової відстані.

РОЗДІЛ 3

ПРОВЕДЕННЯ АНАЛІЗІВ ДЛЯ РОЗРОБКИ КОМП'ЮТЕРНОЇ СИСТЕМИ ДЛЯ ВИЗНАЧЕННЯ РЕАКЦІЙ НА ТЕСТОВЕ НАВАНТАЖЕННЯ

Для перевірки коректності алгоритму квадрата евклідової відстані необхідно розщепити базу даних по статі та провести кластеризацію для кожної половини окремо. Проведення кластеризації для кожного пацієнта окремо передбачає вдосконалення програмного продукту та розробку нового модулю кластеризації, який би міг класифікувати одразу всіх пацієнтів в базі даних.

Наступним кроком для дослідження ефективності алгоритму є проведення бінарної логістичної регресії та дискримінантного аналізу. Якщо результат даних тестів буде складати 80% і більше, тоді можна буде стверджувати, що алгоритм квадрату евклідової відстані працює стабільно та ефективно. Це дасть змогу провести наступні дослідження, які будуть слугувати для порівняння наявних кластерів між собою. Також доцільно буде дослідити і графіки середніх значень, отримані на базі результуючих таблиць з попередніх досліджень. У разі збіжності результату алгоритму порівняння кластерів на новій базі даних з графіками, отриманими на навчальній (попередній) базі даних можна буде приступити до побудови нових результуючих таблиць, що в майбутньому будуть використані для дослідження.

Таким чином можемо встановити порядок дій для досягнення поставленої мети:

- Приведення бази даних до прийнятного для SPSS виду;
- Розщеплення бази даних по статі;
- Проведення глобальної кластеризації;
- Проведення логістичної регресії;
- Проведення дискримінантного аналізу;

- Дослідження графіків, побудованих на базі результуючих таблиць;
- Розробка модулю для дослідження подібності кластерів (реалізація побудови ліній тренду);
- Побудова нових результуючих таблиць;
- Розробка модулю універсальної кластеризації;
- Проведення дисперсійного аналізу;
- Вдосконалення автоматизації програмного продукту «Clusterbox».

У ході проведення досліджень вихідну базу даних було приведено до вигляду, наведеному на рисунку 3.1.:

Для кластеров.xlsx - Excel

Файл Главная Вставка Разметка страницы Формулы Данные Рецензирование Вид Настройки Команда ? Помощь Bogdan V... Общий доступ

A2

<

Рисунок 3.1. База даних для дослідження

На рисунку ми можемо побачити, що для дослідження були обрані дані артеріального систолічного тиску, артеріального діастолічного тиску та значення частоти серцевих скорочень. Вони розташовані у порядку зростання: від стану спокою до п'ятої хвилини включно після навантаження. Також ми можемо побачити, що база даних вже розщеплена по статі, тому можемо перейти до визначення групи ризику для чоловіків (наступні дослідження будуть проведені для чоловічої частини бази даних).

Для визначення групи ризику скористаємося програмним продуктом «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. Після цього вони переносяться до таблиці з дослідженнями. Таким чином наша база набуває наступного вигляду:

	В	С	Д	Е	Ф	С	Т	U	V	W	X	Y	Z
1	ПІБ	№	ATC0	ATD0	ЧСС0	ATC5	ATD5	ЧСС5	CI8_APH	CI	Dist	NextCI	NextDist
2	Абрамов А.В.	755	122	68	82	127	64	82	2	2	1652,36	7	2055,77
3	Абрамов А.В.	1143	113	66	97	110	66	97	2	2	832,64	3	1592,71
4	Атманчук М.В.	1343	116	67	88	115	66	79	2	2	769,02	7	1320,67
5	Атманчук М.В.	1404	126	71	88	117	74	97	3	3	645,99	2	656,96
6	Балашов О.В.	1269	123	75	65	139	60	58	4	4	1551,43	7	2940,21
7	Балашов О.В.	1512	140	85	80	137	64	91		1	2387,64	3	3583,21
8	Баран Д.Р.	1252	118	75	82	114	63	85	7	7	1289,01	2	1959,1
9	Баран Д.Р.	1418	124	70	92	118	64	76	7	7	1237,35	2	1576,7
10	Баран Д.Р.	1518	110	73	114	113	78	102		3	869,35	2	2214,12
11	Бигков В.А.	1248	115	67	76	125	63	77	7	7	495,43	4	1401,11
12	Бигков В.А.	1525	110	74	60	105	68	52		7	2263,77	4	3989,21
13	Богда М.С.	1348	106	54	74	122	64	74	7	7	1228,37	2	2694,46
14	Богда М.С.	1582	107	71	78	115	71	69		7	1365,93	2	2564,32
15	Бондар М.І.	983	106	54	76	115	54	78	7	7	1478,77	2	2386,9
16	Бондар М.І.	1149	95	58	83	109	50	78	7	7	2416,95	2	3150,06
17	Бондаренко К.І.	765	105	70	95	118	67	87	2	2	495,48	7	1954,43
18	Бондаренко К.І.	1125	112	60	80	110	59	74	7	7	919,39	2	2966,42
19	Борисов Р.О.	1250	135	77	78	144	75	88	1	1	874,98	5	2208,62
20	Борисов Р.О.	1449	138	64	66	128	60	64		4	820,35	5	2267,22
21	Борисов Р.О.	1454	142	79	69	125	60	73		4	1166,17	5	1708,16
22	Борисов Р.О.	1541	126	76	72	131	63	61		4	1114,65	5	1695,28

Рисунок 3.2. База даних для дослідження

Вона містить чотири додаткові поля: CI – характеризує кластер (група ризику), до якого відноситься студент, Dist – мінімальна відстань до цього кластеру, NextCI – субоптимальний кластер (група ризику, що йде наступною після визначеної), NextDist – відстань до субоптимального кластеру. Маючи всі необхідні дані для перевірки ефективності алгоритму квадрату евклідової відстані можемо провести наступні тести: бінарна логістична регресія та дискримінантний аналіз.

Оскільки класифікатор студентів (група ризику) є не бінарною змінною, було прийнято рішення розбити пацієнтів на групи методом «один проти всіх» та вирівняти дані в групах, де були отримані занадто асиметричні показники класифікації. Результати дослідження представлені для третього кластеру, але аналогічна процедура проводилася для кожного кластеру окремо. Результуючу таблицю для всіх кластерів зображено у таблиці 3.1.

На рисунку 3.3 зображено встановлення параметрів, які вводилися для отримання класифікації даних.

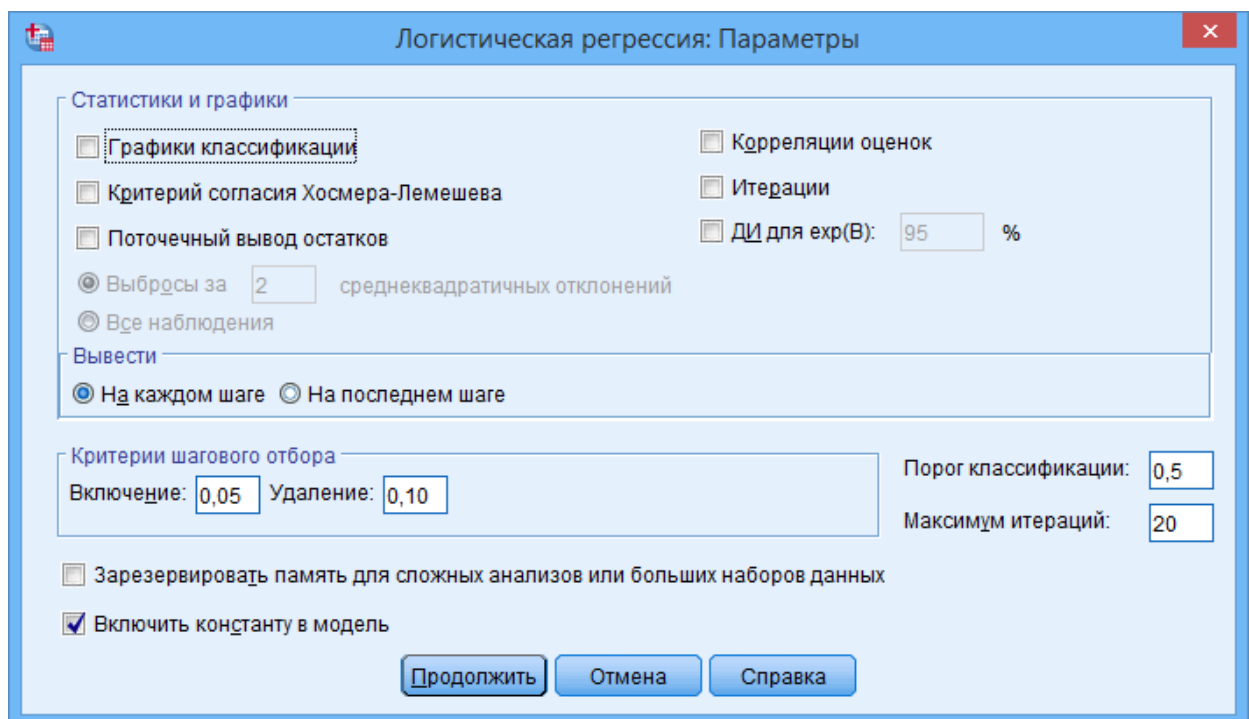


Рисунок 3.3. Зображення параметрів

Розбиття на класи «один проти всіх»

На рисунку 3.4 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 умовного включення.

Таблица классификации^а

Наблюдаемые		Предсказанные		
		3,		Процент корректных
		0	3	
Шаг 1	3, 0	238	39	85,9
	3	36	234	86,7
	Общий процент			86,3
Шаг 2	3, 0	243	34	87,7
	3	30	240	88,9
	Общий процент			88,3
Шаг 3	3, 0	244	33	88,1
	3	24	246	91,1
	Общий процент			89,6
Шаг 4	3, 0	245	32	88,4
	3	18	252	93,3
	Общий процент			90,9
Шаг 5	3, 0	245	32	88,4
	3	18	252	93,3
	Общий процент			90,9
Шаг 6	3, 0	244	33	88,1
	3	18	252	93,3
	Общий процент			90,7
Шаг 7	3, 0	245	32	88,4
	3	12	258	95,6
	Общий процент			92,0
Шаг 8	3, 0	245	32	88,4
	3	18	252	93,3
	Общий процент			90,9

а. Разделяющее значение = ,500

Рисунок 3.4. 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 (277), що відносяться до інших кластерів, коректно були визнані тестом 245. В цьому випадку відсоток класифікації склав 88,4%.

Для побудови рівняння регресії було використано наступну таблицю:

Переменные в уравнении

		В	Стд.Ошибка	Вальд	ст.св.	Знач.	Ехр(В)
Шаг 1 ^a	ЧСС3	-19,677	1,752	126,070	1	,000	,000
Шаг 2 ^b	ЧСС2	-22,007	1,977	123,860	1	,000	,000
	ЧСС3	,148	,022	46,272	1	,000	1,159
Шаг 3 ^c	ЧСС2	-29,056	2,902	100,267	1	,000	,000
	ЧСС3	,150	,023	42,456	1	,000	1,162
Шаг 4 ^d	ЧСС0	-33,421	3,467	92,922	1	,000	,000
	ЧСС2	,073	,019	14,817	1	,000	1,076
	ЧСС3	,125	,021	34,489	1	,000	1,133
Шаг 5 ^e	ЧСС0	-29,842	3,647	66,970	1	,000	,000
	ЧСС2	,088	,020	18,323	1	,000	1,092
	ЧСС3	,125	,022	31,889	1	,000	1,133
Шаг 6 ^f	ЧСС0	-31,503	3,826	67,783	1	,000	,000
	ЧСС2	,096	,021	20,656	1	,000	1,101
	ЧСС3	,131	,023	32,238	1	,000	1,140
Шаг 7 ^g	ЧСС0	-33,098	4,038	67,196	1	,000	,000
	ЧСС2	,102	,022	21,121	1	,000	1,108
	ЧСС3	,125	,024	26,276	1	,000	1,133
Шаг 8 ^h	АТД0	-31,370	4,018	60,965	1	,000	,000
	ЧСС0	-,095	,024	15,233	1	,000	,909
	ЧСС2	,096	,023	18,329	1	,000	1,101
	ЧСС3	,128	,025	26,699	1	,000	1,137

Рисунок 3.5. Змінні для рівняння регресії

Таким чином, рівняння регресії набуває вигляду:

$$Y1 = -31,370 \cdot \text{АТД0} - 0,095 \cdot \text{ЧСС0} + 0,096 \cdot \text{ЧСС2} + 0,128 \cdot \text{ЧСС3}$$

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

Таблица классификации^a

		Предсказанные		Процент корректных
		0	3	
Наблюдаемые	Шаг 1	239	38	86,3
	Общий процент	36	234	86,7
	Шаг 20	277	0	86,5
	Общий процент	0	270	100,0
	Общий процент			100,0

a. Разделяющее значение = ,500

Рисунок 3.6. 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:

Предупреждения

Оценивание не состоялось из-за числовых проблем.
Возможные причины: (1) по крайней мере один из критериев сходимости LCON и BCON равен нулю или слишком мал; (2) значение EPS слишком мало (если оно не задавалось, то, возможно, значение по умолчанию слишком мало для этих данных).

Рисунок 3.7 . Попередження

		Переменные в уравнении					
		В	Стд.Ошибка	Вальд	ст.св.	Знач.	Exp(B)
Шаг 1 ^a	КвЧСС3	-9,677	,850	129,602	1	,000	,000
Шаг 2 ^b	КвЧСС3	-35,027	5,093	47,292	1	,000	,000
	КубЧСС3	,000	,000	32,595	1	,000	1,000
Шаг 20 ⁱ	АТС0	172652,633	809148,360	,046	1	,831	
	ЧСС1	-47,939	361,055	,018	1	,894	,000
	ЧСС3	831,154	5275,270	,025	1	,875	
	КвАТС0	,156	1,442	,012	1	,914	1,169
	КвЧСС3	-8,628	52,555	,027	1	,870	,000
	КубАТС0	-,003	,019	,025	1	,874	,997
	КубЧСС0	,000	,002	,013	1	,908	1,000
	КубЧСС1	,001	,008	,022	1	,883	1,001
	КубЧСС3	,030	,175	,029	1	,865	1,030
	КубЧСС5	,001	,004	,031	1	,861	1,001
	ДелАТД2	-731813,859	9799754,172	,006	1	,940	0,000
	ДелАТС4	-268776,880	1526791,457	,031	1	,860	0,000
	ДелАТР3	-3523,495	111787,139	,001	1	,975	0,000

a. Переменные, включенные на шаге 1: КвЧСС3.

Рисунок 3.8. Змінні для рівняння регресії з нелінійними перетвореннями

Таким чином, рівняння регресії має вигляд:

$$Y_1 = 172652,633 \cdot \text{АТС0} - 47,939 \cdot \text{ЧСС1} + 831,154 \cdot \text{ЧСС3} + 0,156 \cdot \text{КвАТС0} - 8,628 \cdot \text{КвЧСС3} - 0,003 \cdot \text{КубАТС0} + 0,000 \cdot \text{КубЧСС0} + 0,001 \cdot \text{КубЧСС1} + 0,030 \cdot \text{КубЧСС3} + 0,001 \cdot \text{КубЧСС5} - 731813,859 \cdot \text{ДелАТД2} - 268776,880 \cdot \text{ДелАТС4} - 3523,495 \cdot \text{ДелАТР3}$$

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 3.1

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 3.1:

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 нелінійними перетвореннями

1- виграш ЛР на істинних даних

2 – виграш ЛР з нелінійними перетвореннями

* - неповність моделі

З таблиці 3.1 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 3.1 ми можемо сказати, що проведення логістичної регресії, включаючи нелінійні перетворення, а саме: операція взяття квадрату та кубу, і операція взяття оберненої змінної, буде складніші рівняння регресії, але самі моделі є простішими за складністю і дають більш високий відсоток коректності.

Наступним кроком дослідження є проведення дискримінантного аналізу, оскільки метод бінарної регресії дає похибку при дослідженні 3 кластеру на даних з нелінійними перетвореннями та при дослідженні 5,6 кластерів на істинних даних. Аналогічно до розділу 3 результати дослідження представлені для третього кластеру, але дана процедура проводилася для кожного кластеру окремо.

Всі дані було класифіковано методом дискримінантного аналізу, використовуючи покроковий відбір. На рис. 3.9 показаний вибір відстані та критерій розпізнавання.

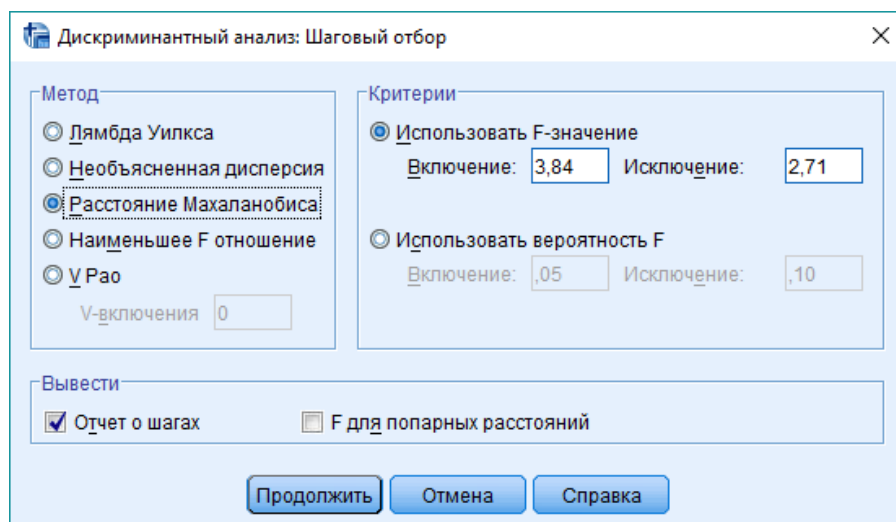


Рисунок 3.9. Вибір методів

У вікні вибору класифікації, що зображено на рис 3.10, показаний вибір розрахунку класифікації.

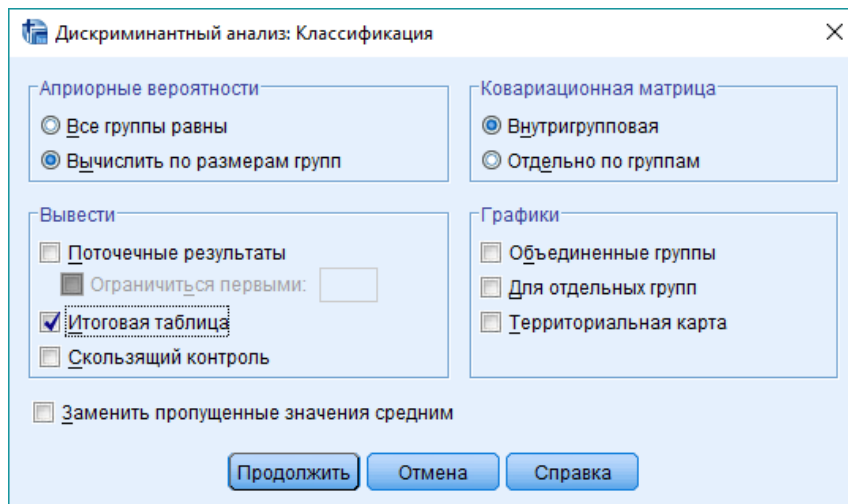


Рисунок 3.10. Вибір класифікації

Результаты анализа для групп «3 проти 1,2,4,5,6,7»

Результаты классификации^а

		Предсказанная принадлежность к группе		Итого
		0	3	
Исходные	Частота			
	0	236	41	277
	3	6	264	270
	%			
	0	85,2	14,8	100,0
	3	2,2	97,8	100,0

а. 91,4% исходных сгруппированных наблюдений классифицировано правильно.

Рисунок 3.11. Результаты класифікації

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

Для побудови дискримінантної функції застосовуються дані з наступної таблиці:

Нормированные
коэффициенты
канонической
дискриминантной
функции

	Функция
	1
АТД0	-,192
ЧСС0	,316
ЧСС1	,255
ЧСС3	,593

Рисунок 3.12. Коефіцієнти для дискримінантної функції

Таким чином, дискримінантна функція набуває вигляду:

$$Y_3 = -0,192 \cdot \text{АТД0} + 0,316 \cdot \text{ЧСС0} + 0,255 \cdot \text{ЧСС1} + 0,593 \cdot \text{ЧСС3}$$

Для підвищення якості моделі, як і для логістичної регресії до аналізу штучно були введені нелінійні змінні x^2 , x^3 та $1/x$ для кожної змінної.

Результаты классификации^а

		Предсказанная принадлежность к группе		Итого
		0	3	
Исходные	Частота	0	3	
		240	37	277
		6	264	270
	%	0	3	
		86,6	13,4	100,0
		2,2	97,8	100,0

а. 92,1% исходных сгруппированных наблюдений классифицировано правильно.

Рисунок 3.13. Результати класифікації з нелінійними перетвореннями

З рисунку 3.13 ми бачимо, що відсоток правильно класифікованих спостережень складає 92,1%

Нормированные коэффициенты канонической дискриминантной функции

	Функция
	1
АТД1	,914
КвАТС0	6,436
КвЧСС1	,260
КвЧСС3	12,217
КубАТС0	-,953
КубЧСС0	,240
КубЧСС3	-9,108
ДелАТС3	-,203
ДелЧСС3	2,967

Рисунок 3.14. Коефіцієнти для дискримінантної функції з нелінійними перетвореннями [6,7].

Таким чином, дискримінантна функція має вигляд:

$$Y_1 = 0,914 \cdot \text{АТД1} + 6,436 \cdot \text{КвАТС0} + 0,26 \cdot \text{КвЧСС1} + 12,217 \cdot \text{КвЧСС3} - 0,0953 \cdot \text{КубАТС0} + 0,24 \cdot \text{КубЧСС0} - 9,108 \cdot \text{КубЧСС3} - 0,203 \cdot \text{ДелАТС3} + 2,967 \cdot \text{ДелЧСС3}$$

Результати дослідження показують, що модель, побудована з додатковими змінними дає більш 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 пункту регресійного аналізу було побудовано таблицю порівняльної характеристики для дискримінантного аналізу щоб оцінити результати дослідження.

Таблиця 3.2

Порівняльна характеристика

Класифікатор	Позначення	ДА на істинних даних	ДА нелінійними перетворенням ³	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 3.2 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 покращують 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 випадках – метод логістичної регресії. Ще у одному випадку логістична регресія та дискримінантний аналіз дають ідентичні результати. 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.

Дослідження результатів проведення логістичної регресії та дискримінантного аналізу показують, що в середньому відсоток класифікації становить близько 92, але це з урахуванням аналізу на істинних даних. Таким чином, високий відсоток класифікації методом логістичної регресії показує, що наш алгоритм розставляє кластери досить ефективно, оскільки колонка з еталонними значеннями легко сприймається і розшифровується аналізом логістичної регресії в SPSS. Відсоток класифікації дискримінантним аналізом на істинних даних складає в середньому 88.2, що також є досить високим показником.

Отже, можемо з впевненістю сказати, що алгоритм знаходження мінімальної відстані з подальшим визначенням групи ризику, робота якого зображена на рисунку 3.15, працює коректно та ефективно.

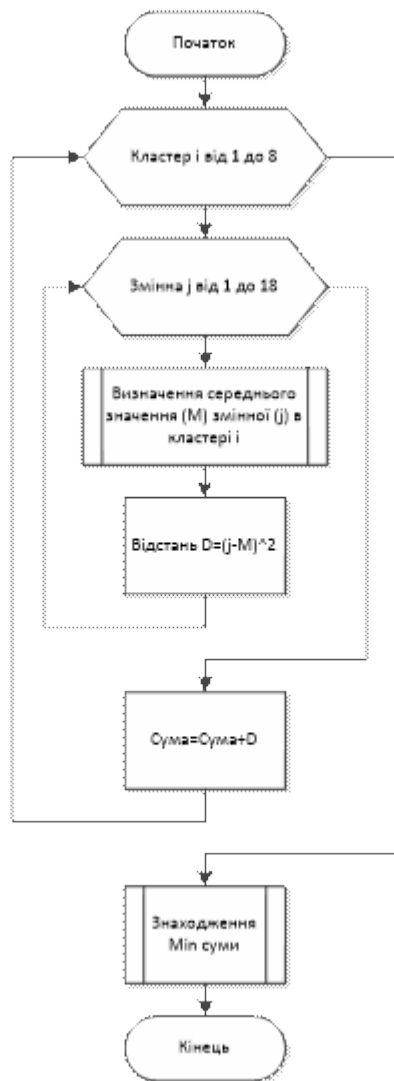


Рисунок 3.15. Блок-схема алгоритму знаходження мінімальної відстані до кластеру [32].

Тому ми можемо перейти до дослідження графіків, отриманих на базі результуючих таблиць, що побудовані на навчальній базі даних (база студентів молодших курсів НТУУ «КПІ ім. Ігоря Сікорського», що містить 1495 спостережень). Таким чином середні значення АТС і АТД (рисунок 2.4 та рисунок 2.5) дають змогу побудувати відповідні графіки по кластерам.

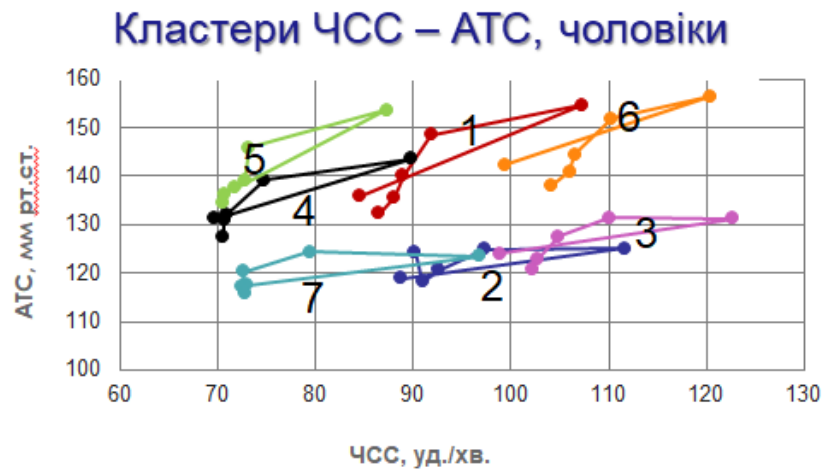


Рисунок 3.16. Графіки ЧСС-АТС по кластерам

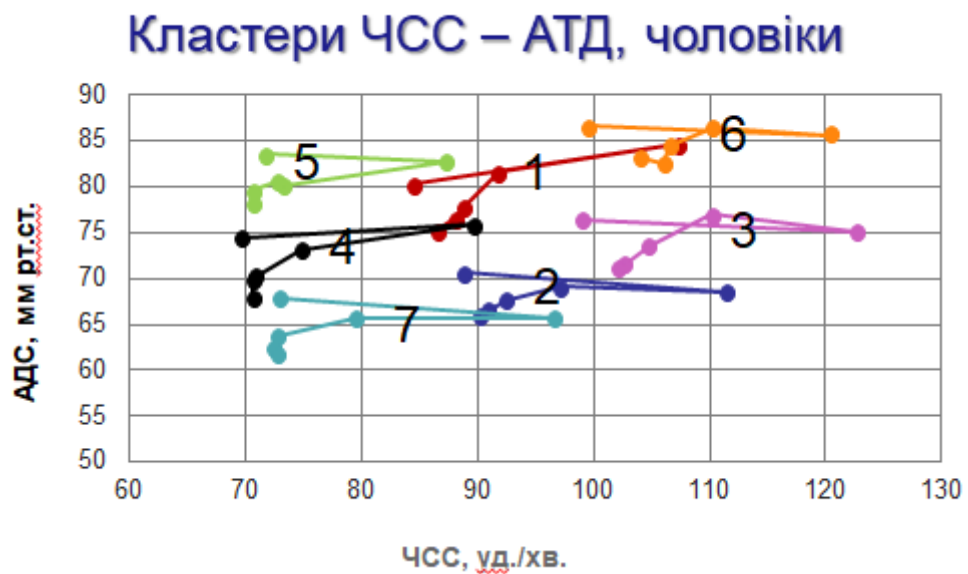


Рисунок 3.17.Графік ЧСС-АДД по кластерам

Графіки зображають динаміку систолічного та діастолічного тисків та зміни пульсу підчас проведення проби Мартіне. Виходячи з побудованих графіків можна сказати, що деякі кластери перетинаються, але, при цьому, їх властивості різні.

За даними попередніх досліджень були сформовані таблиці з характеристиками та рекомендаціями для кожного кластеру.

1	2	3	4
<p>Граничні високі значення артеріального тиску. Помірно виражена симпатикотонія.</p> <p>*Рекомендації: Обмеження з занять важкою атлетикою При занятті фізичними вправами та спорту необхідний моніторинговий контроль тиску та пульсу. В стані спокою та після фізичних навантажень</p>	<p>Виражена симпатикотонія.</p> <p>*Рекомендації: Немає обмежень для занять спортом</p>	<p>Функціональний стан кровообігу різко знижений. Максимальні значення ударного об'єму лівого шлуночка - як реакція на фізичне навантаження. Низький рівень ефективності роботи серця. Виражена симпатикотонія.</p> <p>*Рекомендації: Обмеження фізичних навантажень. За необхідністю чи при поганому самопочутті консультація лікаря.</p>	<p>Нормотонія за індексом Кердо.</p> <p>*Рекомендації: Немає обмежень для занять спортом</p>

Рисунок 3.18. Опис кластерів 1-4 та рекомендації до них

5	6	7
<p>Граничні високі значення артеріального тиску. Найбільша ефективність роботи серця. Переважання симпатикотонічної регуляції за індексом Кердо (Парасимпатикотонія).</p> <p>*Рекомендації: Обмеження з занять важкою атлетикою. При занятті фізичними вправами та спорту необхідний моніторинговий контроль тиску та пульсу. В стані спокою та після фізичних навантажень</p>	<p>Функціональний стан кровообігу різко знижений. Максимальні значення ударного об'єму лівого шлуночка - як реакція на фізичне навантаження. Граничні високі значення артеріального тиску. Граничні високі енерговитрати серця. Низький рівень ефективності роботи серця. Помірно виражена симпатикотонія.</p> <p>*Рекомендації: Обмеження фізичних навантажень. За необхідністю чи при поганому самопочутті консультація лікаря. Обмеження з занять важкою атлетикою При занятті фізичними вправами та спорту необхідний моніторинговий контроль тиску та пульсу. В стані спокою та</p>	<p>Помірно виражена симпатикотонія.</p> <p>*Рекомендації: Немає обмежень для занять спортом</p>

Рисунок 3.19. Опис кластерів 5-7 та рекомендації до них

Як ми бачимо, за показником АТС кластери 2 і 3, 7 і 2 перетинаються, а за показником 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.

Оскільки ефективність алгоритму була підтверджена, а деякі кластери знаходяться поруч, то можемо розробити модуль програми, який буде порівнювати кластери між собою на подібність. Для цього ми використовуємо метод побудови ліній трендів.

Висновки до розділу 3

В даному розділі було проведено глобальну кластеризацію за допомогою програмного продукту «Clusterbox». Результат кластеризації вхідної бази даних використано для методів логістичної регресії та дискримінантного аналізу. Проведені тести даними методами підтверджують ефективність алгоритму квадрата евклідової відстані, що дає шлях до дослідження графіків результуючих таблиць. Дослідження показує, що кластери 4 та 5 в графічному вигляді схожі між собою, тому ми можемо приступити до розробки та реалізації програмного додатку для оцінки подібності кластерів.

РОЗДІЛ 4

ПРОГРАМНИЙ ПРОДУКТ ДЛЯ ВИЗНАЧЕННЯ ФУНКЦІОНАЛЬНОГО СТАНУ СИСТЕМИ КРОВООБІГУ

4.1. Проектування програмного продукту

4.1.1. Контекстна діаграма

Дана діаграма показує процес використання додатку для побудови ліній регресій, що дає змогу оцінити подібність вибраних кластерів. Якщо обрані кластери є подібними, тоді можна буде зменшити початкову кількість кластерів для підвищення ефективності оцінки стану системи кровообігу.

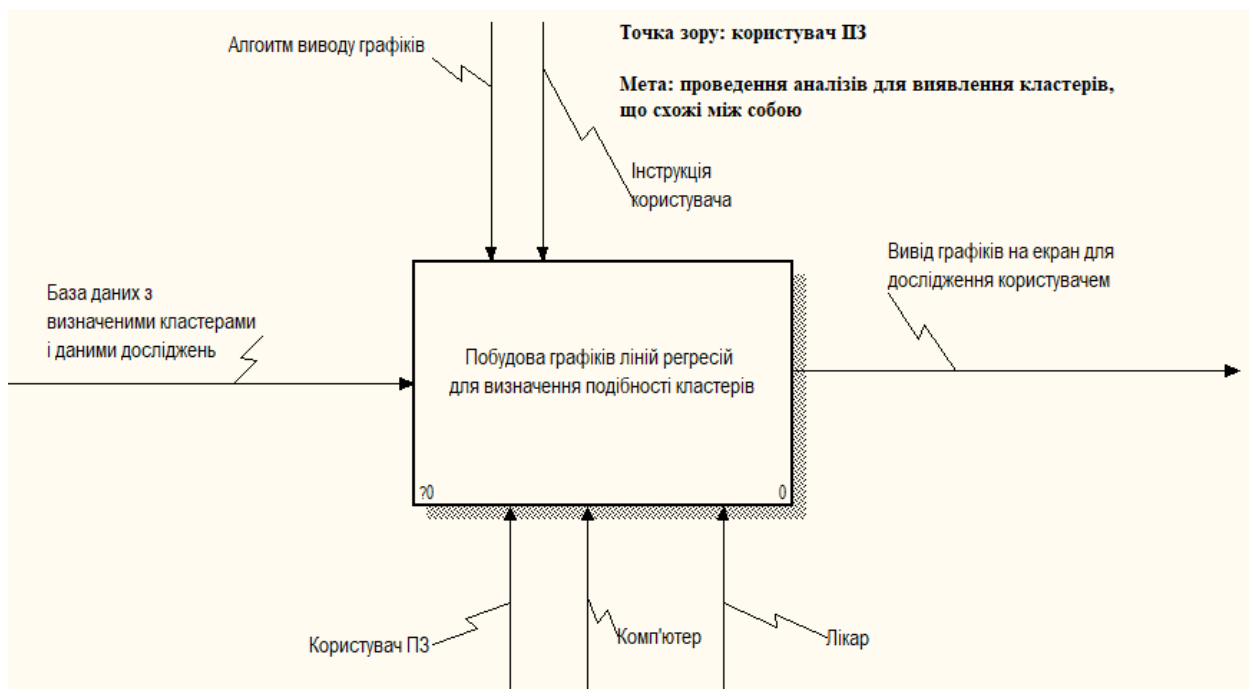


Рисунок 4.1. Контекстна діаграма [33].

Вхідними даними для роботи програмного додатку є розщеплена по статі база даних зі значеннями артеріального тиску та пульсу у стані спокої та на 1-5 хвилинах після проведення тесту на навантаження, значення різниці систолічного та діастолічного тиску на 3-5 хвилинах після навантаження та визначені кластер, субкластер (мітка), мінімальна та

субмінімальна відстані. Вихідними даними є вивід ліній трендів (регресій) за обраними кластерами на екран користувача.

4.1.2. Діаграма декомпозиції першого рівня

Діаграма декомпозиції першого рівня для додатку, розробленого для побудови ліній регресій, показує нам процес роботи програмного продукту на окремих етапах застосування.

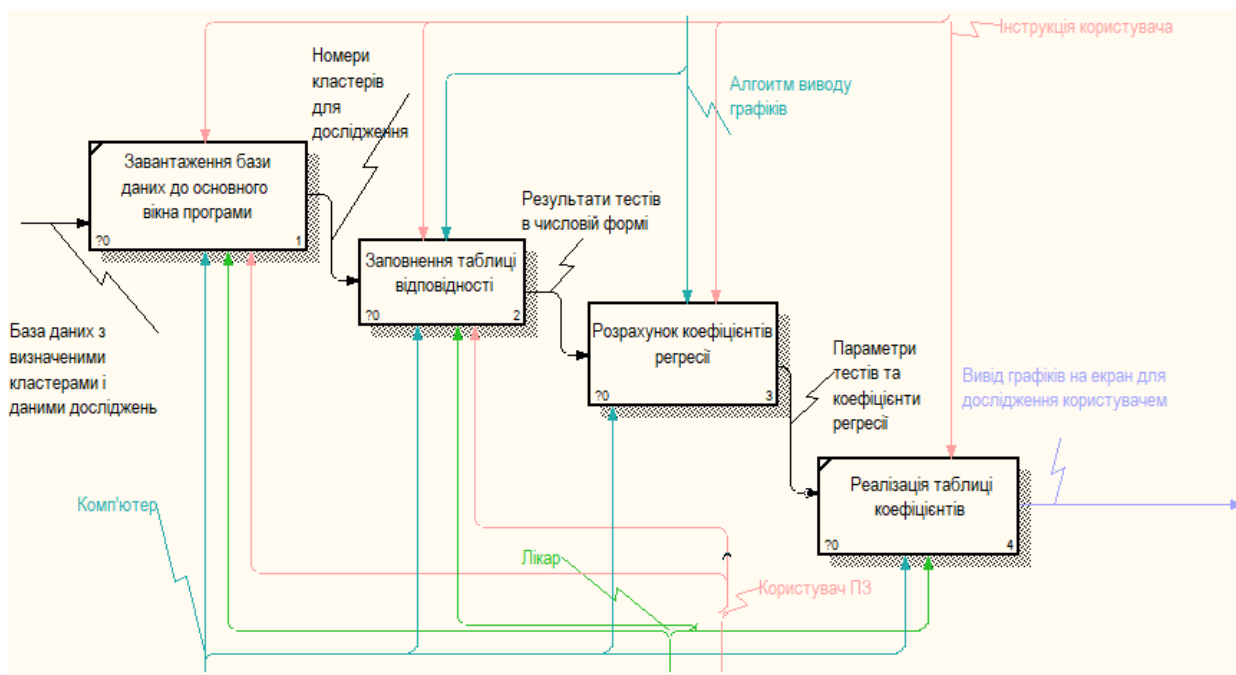


Рисунок 4.2. Діаграма декомпозиції першого рівня [34].

З діаграми видно, що весь процес роботи програми складається з чотирьох етапів: завантаження бази даних до основного вікна програми, заповнення таблиці відповідності, розрахунок коефіцієнтів регресії, реалізація таблиці коефіцієнтів. Всі ці етапи необхідні для побудови та виводу графіків на екран

4.1.3. Діаграма декомпозиції другого рівня

Для більш детального опису процесу розрахунку коефіцієнтів регресії було побудовано діаграму декомпозиції другого рівня.

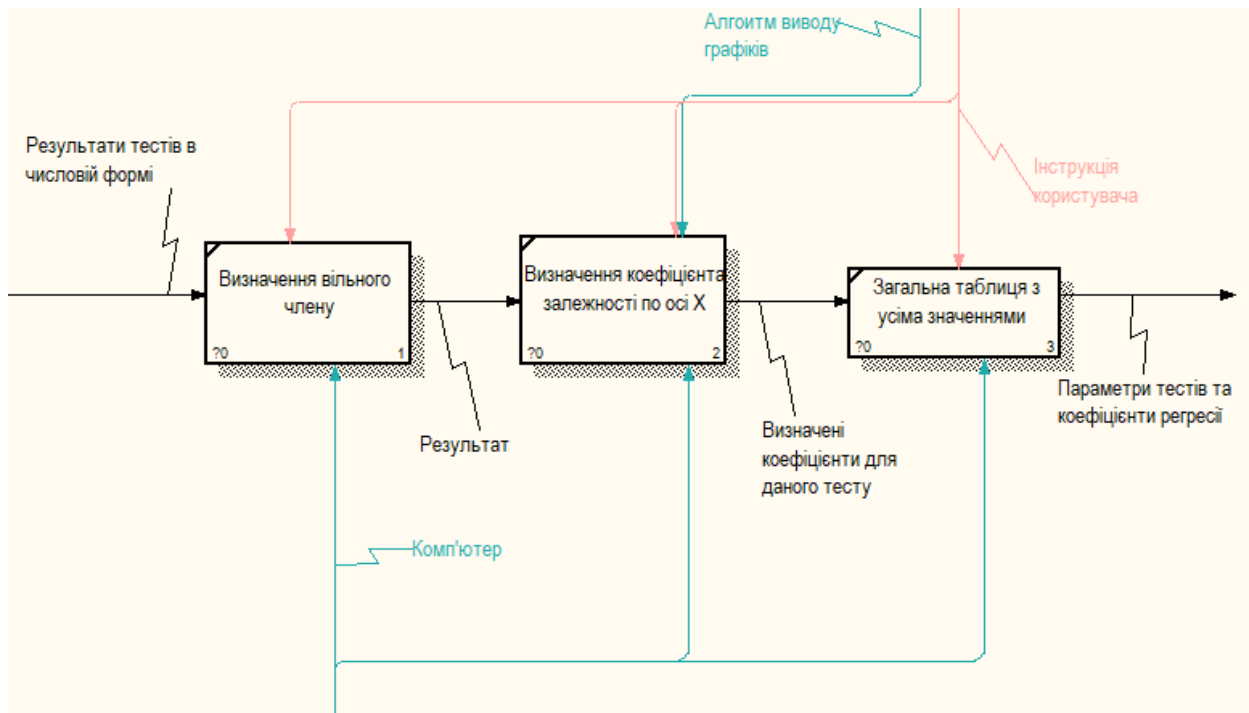


Рисунок 4.3. Діаграма декомпозиції другого рівня [33, 34].

Даний процес включає в себе три блоки: визначення вільного члену, визначення коефіцієнта залежності по осі X та загальну таблицю з усіма значеннями, яка має бути заповнена. Вхідними даними для розрахункового блоку є результати тестів в числовій формі, а вихідними - параметри тестів та коефіцієнти регресії.

4.1.4. Діаграма дерева вузлів

Процеси та підпроцеси, що були розглянуті вище зручно скомпонувати в діаграму дерева вузлів. Вона дозволяє оцінити послідовність основних процесів в цілому.

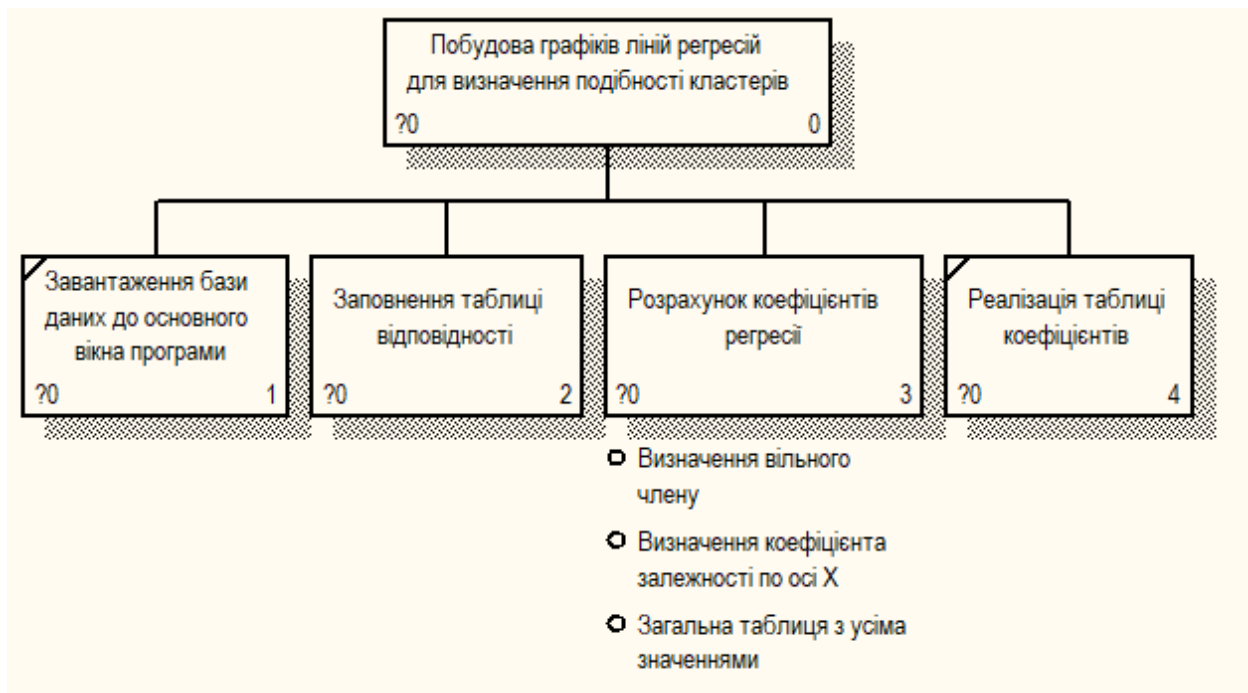


Рисунок 4.4. Діаграма дерева вузлів [35].

4.1.5. Use Case діаграма

Дана діаграма спрямована на демонстрацію процесів та їх виконання в ході алгоритму виходячи з точки зору самих учасників, що безпосередньо беруть участь у використанні програмного продукту. Учасники цієї системи: лікар-адміністратор та користувач, які на діаграмі позначені фігурою людини. Можливі дії представлені еліпсами та зображають доступні комбінації розвитку сценарію використання і роботи програмного продукту. Одні варіанти використання можуть бути частиною інших варіантів, розширювати їх або бути більш узагальненою версією.

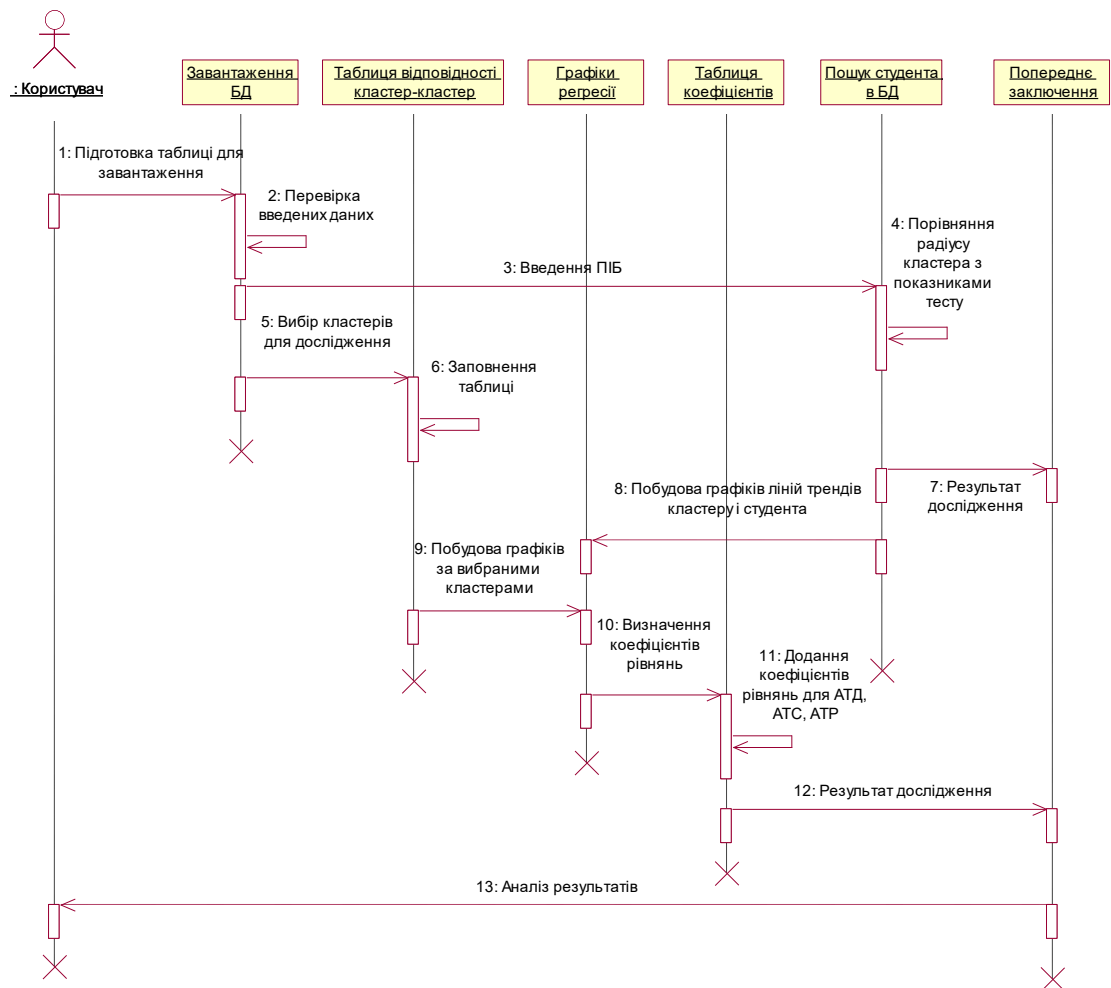


Рисунок 4.6. Діаграма послідовності [37]

Діаграма відображає послідовність, в якій використовує систему користувач. Починається робота з підготовки таблиці для дослідження. Наступним кроком є завантаження її до програми, де дані з таблиці будуть перевірені на коректність. Далі у відповідних комірках вибираються кластери для дослідження, після чого автоматично заповнюється таблиця відповідності кластер-кластер. Заповнена таблиця слугує базою для побудови графіків. На кожному етапі виведення графіків розраховуються відповідні коефіцієнти регресії. Дані коефіцієнти реєструються у відповідній таблиці. Після побудови всіх графіків (АТС-ЧСС, АТД-ЧСС, АТР-ЧСС) даний результат необхідно проаналізувати, чим на останньому кроці і займається користувач. Також на діаграмі ми можемо побачити, що

функціонал додатку передбачає пошук студента в завантаженій базі даних після заповнення таблиці відповідності. Пошук студента дає можливість подивитись чи потрапляють значення параметрів тесту студента у радіус кластера, до якого відноситься студент. Якщо значення виходять за радіус, тоді на екран користувача виводяться графіки регресій для подальшого дослідження[37].

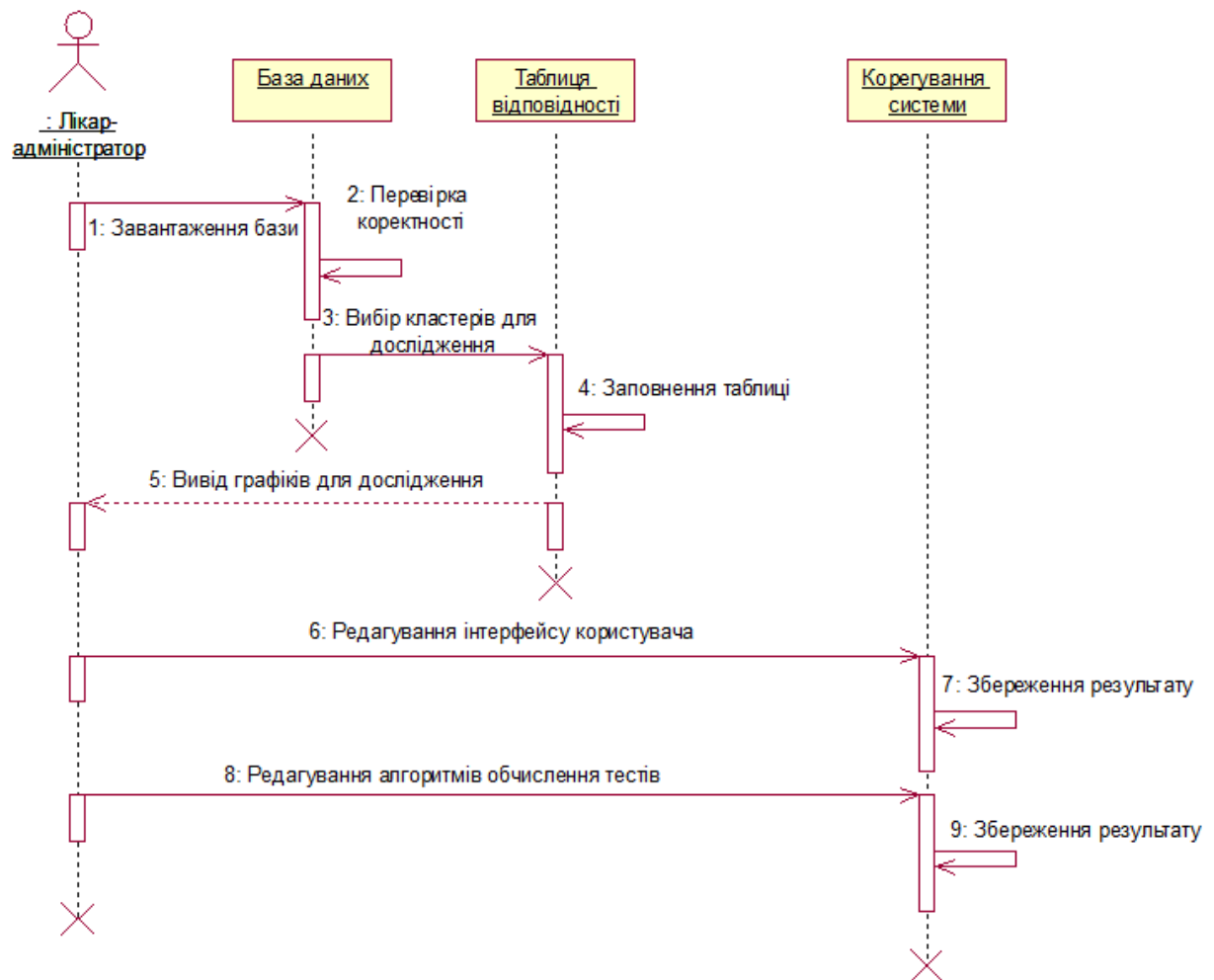


Рисунок 4.7. Діаграма послідовності роботи зі сторони лікаря-адміністратора

Діаграма послідовності роботи зі сторони лікаря-адміністратора аналогічна за роботою до діаграми користувача. Але на відміну від неї лікар-адміністратор має право вносити зміни до інтерфейсу програми та редагувати алгоритми обчислення тестів [37].

4.1.7. Діаграма кооперації

Діаграма кооперації створена для візуалізації взаємодії користувача з основними елементами програмного продукту. Основні етапи позначені прямокутником. Проміжні дії розташовуються між основними блоками. Стрілками вказаний напрямок руху, тобто до чого призведе та чи інша дія.

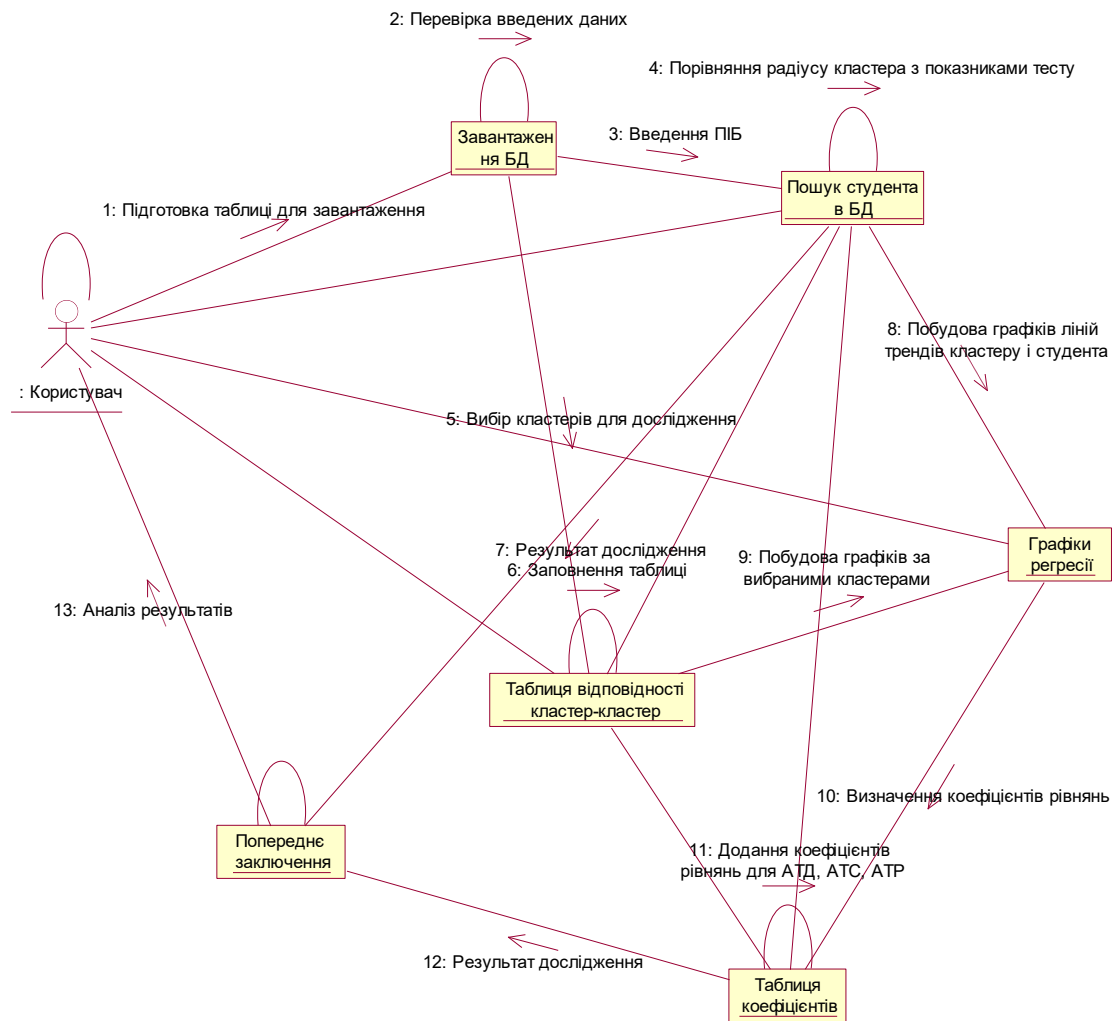


Рисунок 4.8. Діаграма кооперації

Діаграма кооперації для користувача програмного продукту дає змогу відслідкувати його взаємодію з елементами інтерфейсу та побачити побудову і виведення графіків на екран.



Рисунок 4.9. Діаграма кооперації роботи програмного додатку лікарем-адміністратором

Зображена діаграма кооперації роботи лікаря-адміністратора з програмним додатком демонструє процес побудови графіків ліній трендів з виведенням результату для аналізу. Також ми бачимо, що лікар може регулювати алгоритми обчислення тим самим вдосконалюючи роботу програмного додатку [38].

4.1.8. Діаграма діяльності

Дана діаграма передбачає всі варіанти можливого використання програмного продукту. Тобто кожен блок відображає дію, а стрілки, що їх поєднують – вказують на послідовність виконання. Таким чином вибудовується декілька варіантів використання або функціонування системи вцілому. Діаграма діяльності (так само, як і діаграми станів та переходів) зображаються у вигляді орієнтовного графу, де вершини – дії, а ребра – переходи між діями.

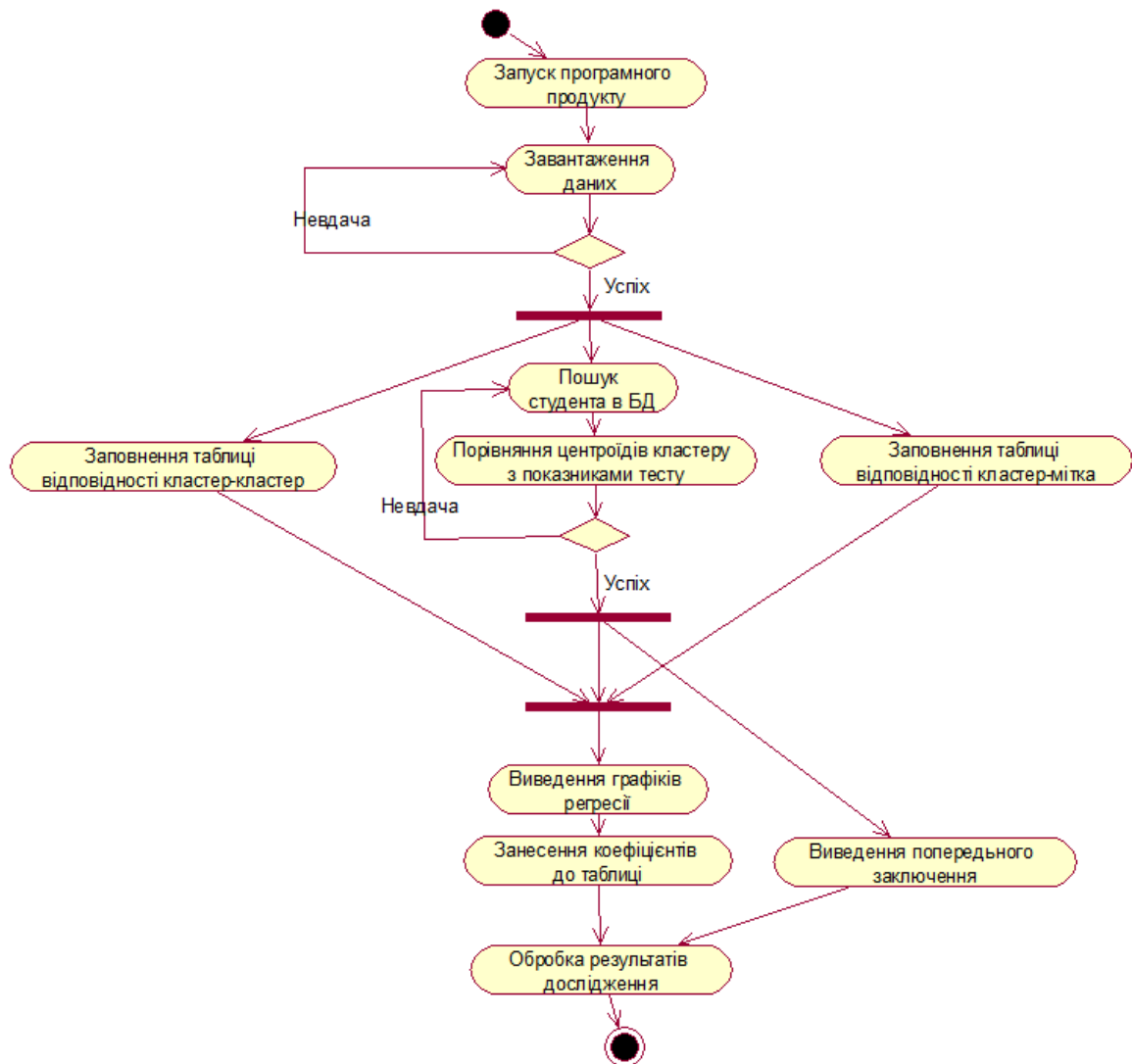


Рисунок 4.10. Діаграма діяльності [39].

4.1.9. Діаграма станів

Діаграма станів є безпосереднім схематичним зображенням стану системи на тому чи іншому етапі. Таким чином вона демонструє всі можливі послідовності і зв'язки роботи алгоритму починаючи від початку і до його кінця, при цьому враховані можливі успіх, невдача та пропуски на певних кроках при проходженні алгоритму.

будуючи у відповідному вікні графіки ліній трендів для кластер-мітка та мітка-кластер.

Додаток працює у сукупності з програмним продуктом «Cluserbox» та має на меті покращення роботи програми для дослідження стану кровоносної системи.

Інтерфейс користувача програми є мінімалістичним та інтуїтивно зрозумілим, що полегшує роботу з програмним продуктом.

Також програмним додатком передбачено пошук студента в завантаженої бази даних для того, щоб подивитись чи перевищує радіус студента радіус кластера. Додатково 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.

Вхідними даними для роботи програми є діастолічний та артеріальний тиски, серцевий тиск і частота серцевих скорочень на першій, другій та третій хвилинах після навантаження. Саме ці показники найкраще характеризують поведінку організму при проведенні тесту Мартіне. Також дослідження передбачає включення до вхідних даних значень АТР1, АТР2, АТР3 (різниця артеріального тиску між діастолічним та систолічним тиском на першій, другій та третій хвилинах після навантаження).

Вихідними даними є виведення графіків в залежності від методу дослідження: при порівнянні кластерів – графіки ліній трендів кластер-кластер, при дослідженні зміни кластеру – кластер-мітка (мітка-кластер).

Основні можливості системи, що розроблені:

- Перевірка коректності завантаженої бази даних;
- Заповнення таблиць спостереженнями, що відносяться до вибраного кластеру;
- Побудова графіків ліній тренду кластер-кластер;

- Можливість вибору таблиці для побудови графіків зміни кластеру (графіки ліній трендів кластер-мітка та мітка-кластер);
- Розрахунок кількості переходів кластер-кластер;
- Розрахунок кількості переходів кластер-мітка;
- Пошук студента у завантаженій базі з визначенням радіусу кластеру та вектору напрямлення студента з подальшим порівнянням та виведенням відповідного графіку за необхідністю.

4.3. Робота з програмним додатком

При запуску програмного додатку ми бачимо відповідне вікно для досліджень, що передбачає заповнення відповідних полів.

Рисунок 4.12. Основне вікно програмного додатку

Для цього необхідно вибрати, для якої частини бази даних буде проводитися дослідження, та натиснути кнопку «Старт». Наступним кроком

є вибір бази даних, що містить прізвище студента, значення артеріального тиску та пульсу в стані спокою та на кожній хвилині після навантаження, включно до п'ятої хвилини, а також значення різниці діастолічного та систолічного тиску на першій, другій і третій хвилині після навантаження.

Якщо вибраного листа не буде знайдено у базі даних буде видано наступне попередження:

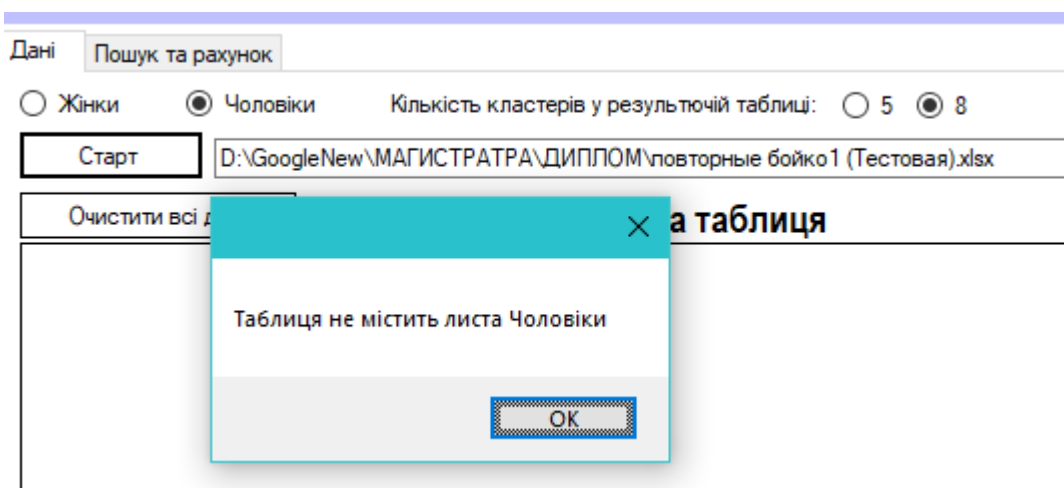


Рисунок 4.13 – Попередження про відсутність листа для дослідження

Для того, щоб завантажити іншу базу для дослідження, необхідно щоб поле з шляхом до бази було пустим. В іншому випадку за цим шляхом буде йти пошук бази даних.

Якщо у таблиці є даний лист, але на ньому відсутнє хоча б одне з основних полів, які мають бути завантажені, програмою буде оброблено виключення з виведенням інформації на екран користувача.

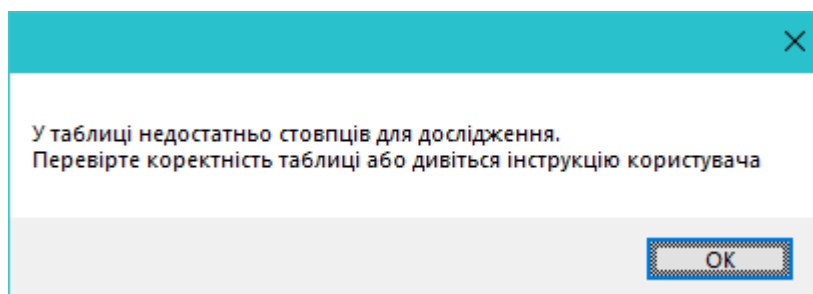


Рисунок 4.14. Попередження про відсутність стовпців для дослідження

Слід також зазначити, що база даних для завантаження повинна бути у форматі таблиці Microsoft Excel та бути у розширенні *.xls або *.xlsx. Якщо спробувати завантажити базу іншого формату або взагалі інший файл, програмою буде видано попередження:

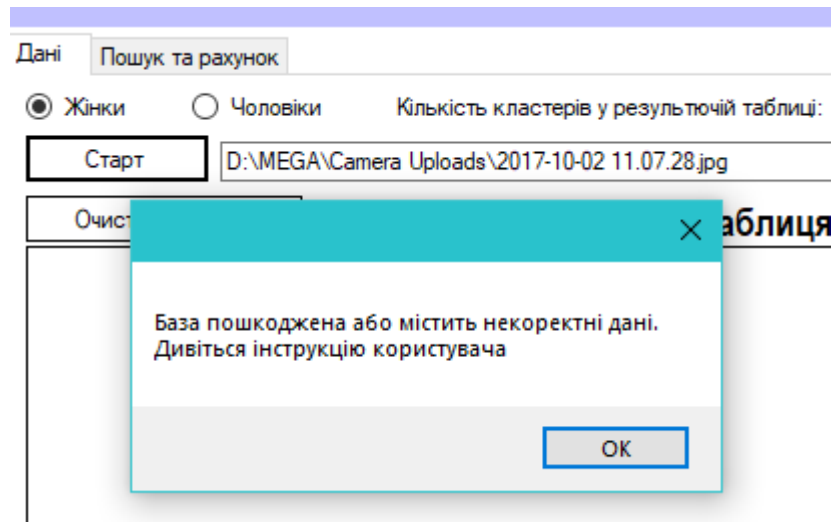


Рисунок 4.15. Попередження про пошкодження бази даних

При успішному завантаженні бази даних до програми вікно «Завантажена таблиця» буде заповненим і користувачу буде доступний пункт з вибором відповідного кластера та мітки для дослідження.

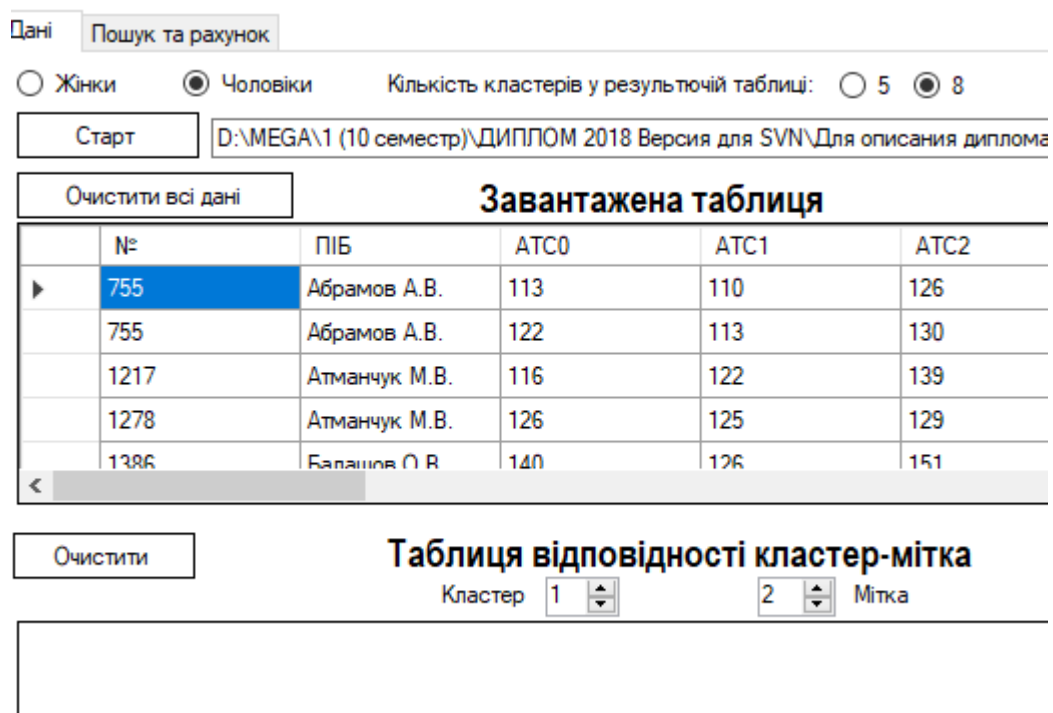


Рисунок 4.16. Завантажена база даних

Оскільки ми хочемо дослідити подібність кластерів, необхідно встановити галочку навпроти напису «Кластер-Кластер» у таблиці відповідності кластер-кластер. Якщо галочка буде ввімкненою, тоді ми зможемо обрати два кластери, які будуть порівнюватися між собою. При цьому, вікно таблиці «кластер-кластер» буде автоматично заповнюватися. Дані у таблиці будуть показувати всі дані в обраних кластерах. Оскільки на графіках (рисунок 3.16 та рисунок 3.17) кластери 4 та 5 знаходяться поруч, тому для дослідження було обрано саме їх.

Дані

Пошук та рахунок

☐ Жінки
 ☒ Чоловіки
 Кількість кластерів у результативній таблиці: ☐ 5 ☒ 8

Старт

D:\MEGA\1 (10 семестр)\ДИПЛОМ 2018 Версія для SVN\Для описання диплома\Прора\Table\ForDisertation.xlsx

Очистити всі дані

Завантажена таблиця

	№	ПІБ	ATC0	ATC1	ATC2	ATC3	ATC4
▶	755	Абрамов А.В.	113	110	126	126	125
	755	Абрамов А.В.	122	113	130	144	137
	1217	Атманчук М.В.	116	122	139	125	122
	1278	Атманчук М.В.	126	125	129	125	116
<	1386	Балашов О.В.	140	126	151	155	139

Очистити

Таблиця відповідності кластер-мітка

Кластер

 Кластер

Очистити

Таблиця відповідності кластер-кластер

☒ Кластер-Кластер

	ПІБ	ATC1	ATC2	ATC3	ATD1	ATD2	ATD3	ЧСС1	ЧСС2	ЧСС3	ATP1	ATP2	ATP3	Кл.	М
▶	Балашов О.В.	134	137	144	66	62	62	83	64	59	68	75	82	4	7
	Борисов Р.О.	163	150	140	75	71	66	99	86	71	88	79	74	4	5
	Борисов Р.О.	156	134	130	70	68	68	96	70	67	86	66	62	4	5
	Борисов Р.О.	158	154	138	78	71	66	85	71	61	80	83	72	4	5
<	Борисов Р.О.	126	152	143	68	71	72	109	83	72	58	81	71	4	7

Таблиця кількості переходів кластер-кластер

Cluster	1	2	3	4	5	6	7	8
▶ 1	0	0	0	1	1	0	0	0
2	0	0	3	0	0	0	1	0
3	0	8	0	0	0	1	0	0
4	2	0	0	0	2	0	1	0
5	0	0	0	1	0	0	0	0
6	0	0	2	0	0	0	0	0
7	0	14	0	12	0	0	0	0
* 8	0	0	0	0	0	0	0	0

Порахувати

Кількість: 49

Таблиця кількості переходів кластер-мітка

Порахувати

Загальна кількість: №

Кластер: К

Мітка: М

Мітка: М

Кластер: К

Г

р

а

ф

і

к

и

Рисунок 4.17. Заповнена таблиця відповідності кластер-кластер

Також на рисунку 4.17 ми можемо побачити заповнену таблицю кількості переходів кластер-кластер. Вона показує скільки змін кластерів є у базі даних. Наприклад, студент Іванов проходив дослідження тричі. Перший раз він був у дургому кластері, а мітка показувала на сьомий. При повторному обстеженні – кластер змінився на сьомий. Студентів, які змінили свій кластер з другого на сьомий всього 14. Загальна кількість зміни кластеру у відповідності з визначеною міткою – 49.

Оскільки таблиця відповідності кластер-кластер заповнена, можемо побудувати відповідні графіки. Для цього необхідно натиснути на клавішу «Графіки».

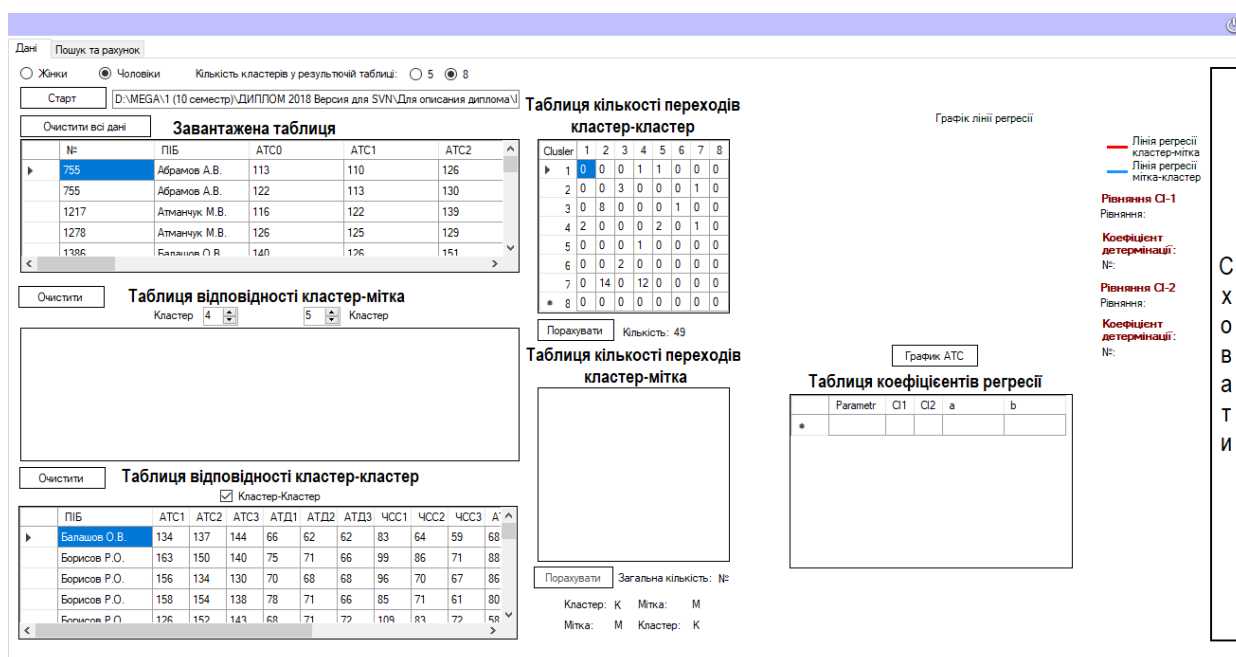
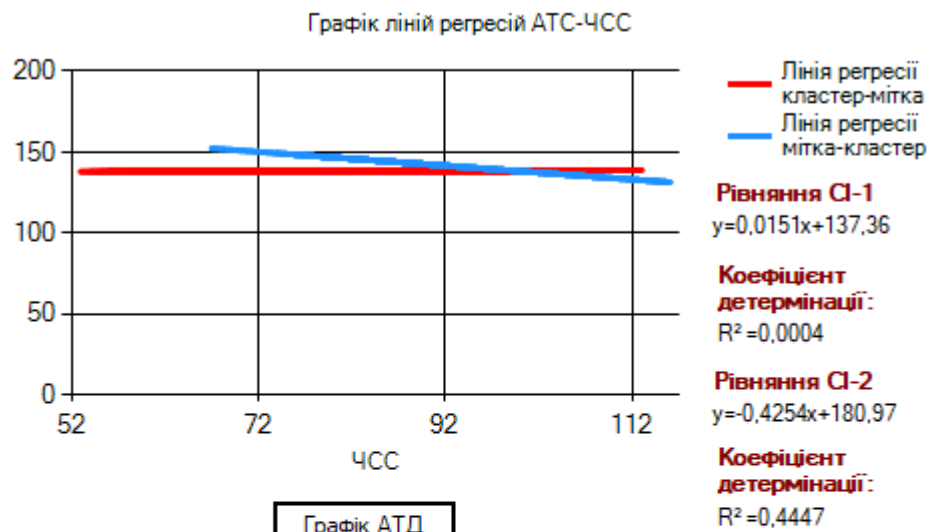


Рисунок 4.18. Вікно з виведенням графіків

Ми бачимо, що у нас з'явилася додаткова область для виводу графіків. При натисненні на «Графік АТС» побудуються лінії трендів «кластер 4» та «кластер 5», дані яких завантажені до таблиці відповідності кластер-кластер. Слід зазначити, що спочатку на екран виведуться лінії трендів за значеннями АТС, про що свідчить заголовок над графіком, а напис на клавіші зміниться на «Графік АТД».



Таблиця коефіцієнтів регресії

	Parametr	CI1	CI2	a	b
►	АТС	4	5	0,0151	137,36
	АТС	5	4	-0,4254	180,97
	АТД	4	5	0,0571	66,37
	АТД	5	4	0,026	81,12
	АТР	4	5	-0,042	70,98
	АТР	5	4	-0,4514	99,85
*					

Рисунок 4.19. Графік ліній регресій АТС-ЧСС із таблицею коефіцієнтів

Також ми можемо побачити, що коефіцієнти ліній регресій записуються до таблиці коефіцієнтів. Таким чином користувач може оцінити не тільки візуальну складову графіків, але й аналітичну.

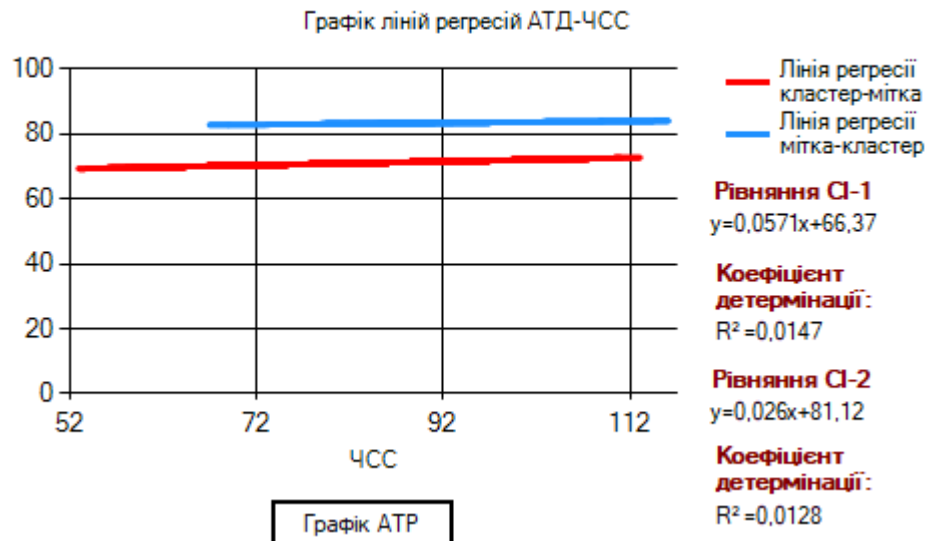


Рисунок 4.20. Графік ліній регресій АТД-ЧСС

За проведеними дослідженнями за кластерами 4 і 5 ми бачимо, що за АТС лінії регресії досить схожі між собою. Аналогічні висновки ми можемо зробити з графіків, побудованих на значеннях АТД. З нього чітко видно, що лінії трендів ідуть паралельно одна одній, а значення АТД відрізняються приблизно на 8-10 позицій. Аналогічну картину ми можемо побачити і на графіках, отриманих на базі результуючих таблиць (рисунок 3.16 та рисунок 3.17). Таким чином ми можемо сказати, що кластери 4 та 5 мають схожі властивості, що підтверджують наші дослідження, тому кількість кластерів може бути зменшеною.

Також модуль дослідження передбачає пошук студента в базі даних з визначенням розташування показників тесту відносно центроїдів кластеру. Якщо показники тесту більші, ніж радіус кластеру, до якого відноситься студент, то виводиться відповідний графік для порівняння результатів. 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 4.21.

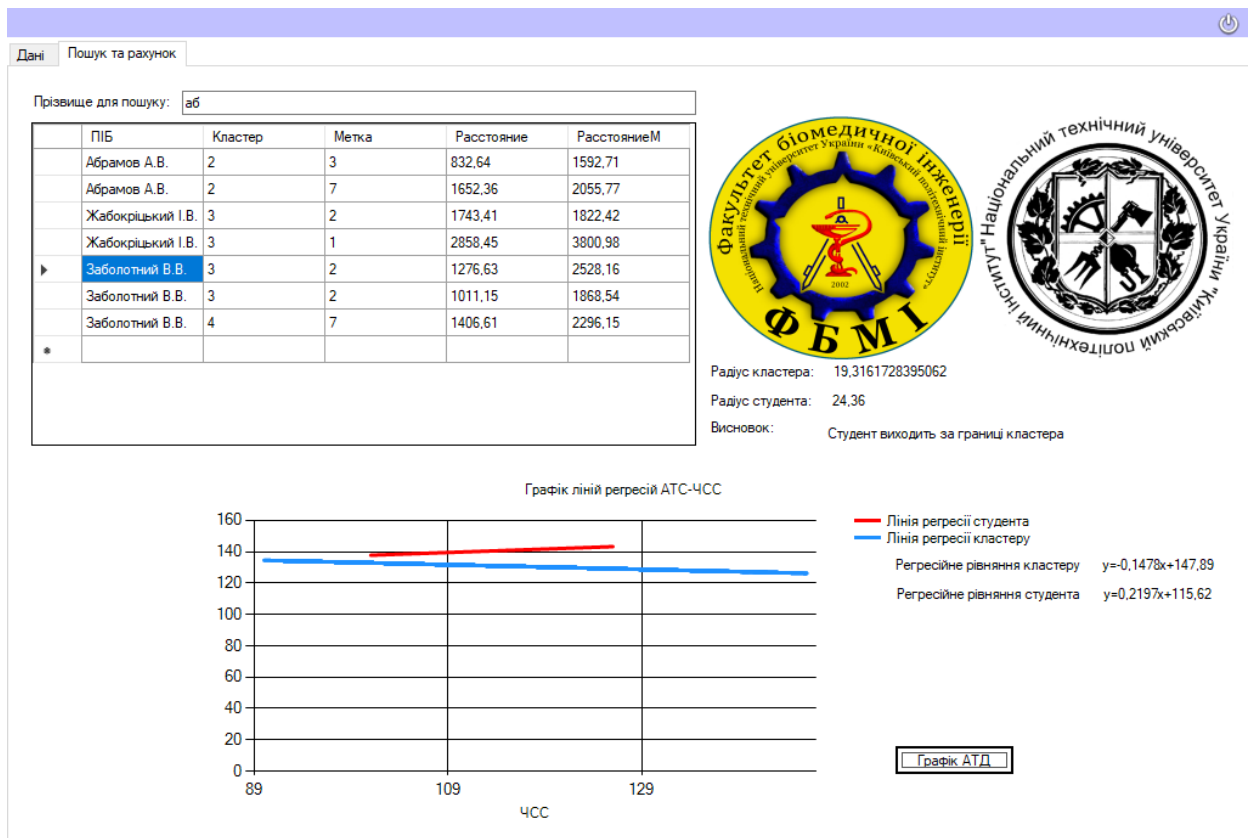


Рисунок 4.21. Пошук студента в БД з виводом графіків

Висновок до розділу 4

В цьому розділі проведено проектування програмного продукту яке представлено у діаграмах та схемах декомпозиції, послідовності, станів та кооперації. Така візуалізація надає змогу чітко відобразити алгоритм виконання, його послідовність, структурованість, подати представлення дій користувачів програмного продукту в залежності від їх функціональної ролі, надає повну картину можливостей програмного продукту та сценаріїв його використання.

Далі детально розглянутій інтерфейс програмного продукту, надані приклади його використання та результати роботи. Також були розглянуть вся можливі випадки та умови невиконання програмою певних задач в залежності від конкретних умов роботи з додатком.

РОЗДІЛ 5

АНАЛІЗ ЗМЕНШЕННЯ КІЛЬКОСТІ КЛАСТЕРІВ З ВДОСКОНАЛЕННЯМ ПРОГРАМНОГО ПРОДУКТУ

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.

Після кластеризації ми занесли результати до окремої таблиці Excel. Після чого завантажили її до SPSS та провели дисперсійний аналіз з виведенням описових статистик. Приклад проведення аналізу наведено для значення систолічного тиску до навантаження, але аналогічні таблиці було отримано і для всіх інших значень тиску та пульсу.

Описательные статистики									
		N	Среднее	Стд. отклонение	Стд. Ошибка	95% доверительный интервал для среднего		Минимум	Максимум
						Нижняя граница	Верхняя граница		
Ps0	0	34	124,088235	12,2458840	2,1001518	119,815444	128,361026	94,0000	146,0000
	1	46	110,739130	12,0193081	1,7721503	107,169837	114,308424	90,0000	151,0000
	2	75	120,720000	9,5994932	1,1084540	118,511357	122,928643	98,0000	152,0000
	3	67	121,164179	8,5786857	1,0480533	119,071673	123,256685	98,0000	139,0000
	4	41	117,170732	9,9772302	1,5581816	114,021529	120,319934	100,0000	146,0000
	5	40	131,675000	8,5586588	1,3532428	128,937808	134,412192	114,0000	149,0000
	6	19	136,578947	9,9460533	2,2817811	131,785103	141,372792	118,0000	155,0000
	Итого	322	121,586957	12,0461296	,6713046	120,266244	122,907669	90,0000	155,0000

Рисунок 5.1. Описові статистики

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

Param	1		2		3		4		5	
	M	SD	M	SD	M	SD	M	SD	M	SD
Ps0	110,7391	12,01931	120,72	9,599493	121,1642	8,578686	117,1707	9,97723	131,675	8,558659
Pd0	64,45652	7,770904	69,54667	5,91218	75,29851	5,351338	71,82927	7,358337	80,175	6,511873
HR0	80,91304	11,378	69,45333	7,713262	82,92537	7,644197	101,7805	9,637199	91,525	8,249281
Ps1	110,7609	14,74552	131,4667	15,41001	126,7761	13,6035	123,7561	15,60253	135,8	12,83265
Pd1	57,86957	7,289593	68,32	10,09972	72,65672	8,557386	65,90244	9,087917	77,225	8,244618
HR1	105,2609	11,71596	92,24	10,86507	110,9552	10,66775	127,439	10,15886	117,35	12,65833
Ps2	118,8478	10,26097	131,6533	12,39932	132,1493	10,99484	126,6341	12,17324	140,85	9,434117
Pd2	60,47826	6,228034	67,56	8,109454	75,1194	7,56892	71,14634	6,440345	80,15	7,423283
HR2	87,80435	13,36766	73,44	9,091487	90	9,916317	111,439	11,14237	100,7	8,846845
Ps3	114,2826	10,26247	128,64	11,50868	127,6866	9,157228	125,2439	10,16804	134,65	9,509914
Pd3	58,69565	7,164324	66,09333	6,883536	73,52239	5,758678	70,60976	6,204345	78,8	6,536682
HR3	79,45652	12,92148	65,77333	8,554394	81,47761	7,634368	104,0244	8,341126	97,05	6,857337
Ps4	113,6522	8,353888	124,92	11,06707	123,9403	8,310053	121,7317	9,287153	130,075	9,965035
Pd4	58,58696	5,964995	65,12	6,959264	71,85075	6,631545	68,2439	5,351544	77,225	7,198602
HR4	77,84783	12,34318	66,68	8,226786	81,58209	7,744302	102,561	7,674792	96,325	6,627013
Ps5	110,5435	8,958439	122,6533	9,977669	122,4478	6,948476	118,9024	9,956919	128,325	9,32157
Pd5	59,1087	6,887116	63,96	7,010494	71,35821	6,285388	69,63415	9,093284	76,125	6,680195
HR5	77,86957	11,88577	67,98667	9,011244	81,62687	7,275604	102,1707	7,385467	95	5,193685

Рисунок 5.2. Резульуюча таблиця

Резульуюча таблиця виступає базою даних для побудови нових графіків.

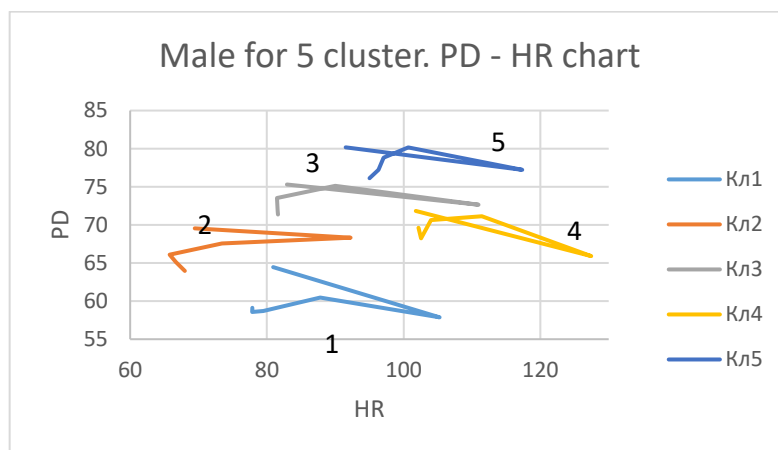


Рисунок 5.3 - Графіки АТД, ЧСС для 5 кластеру

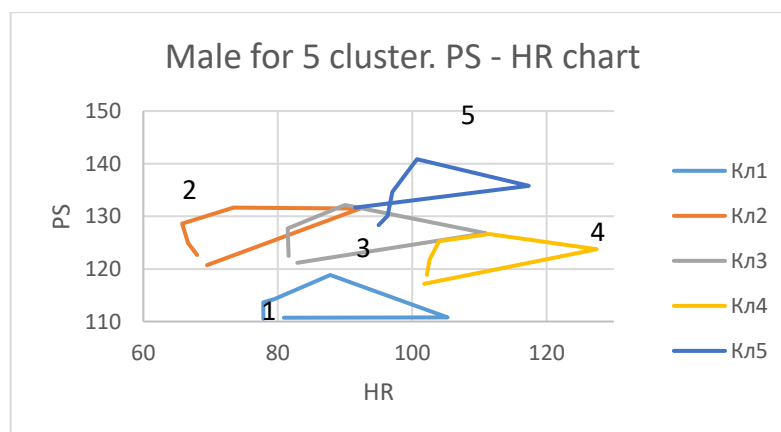


Рисунок 5.4. Графіки АТС, ЧСС для 5 кластеру

Аналогічні аналізи були проведені для чотирьох і трьох кластерів, в ході чого було отримано відповідні результуючі таблиці та побудовано наступні графіки:

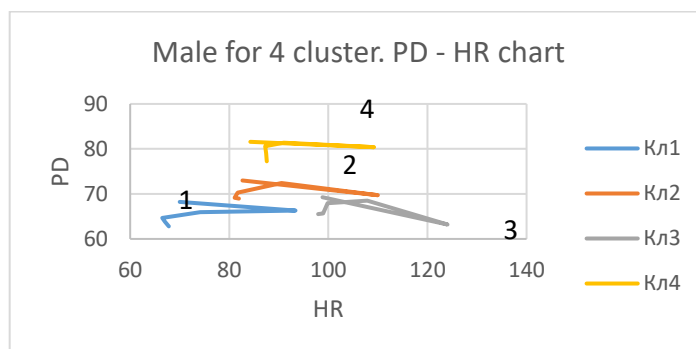


Рисунок 5.5. Графіки АД, ЧСС для 4 кластеру

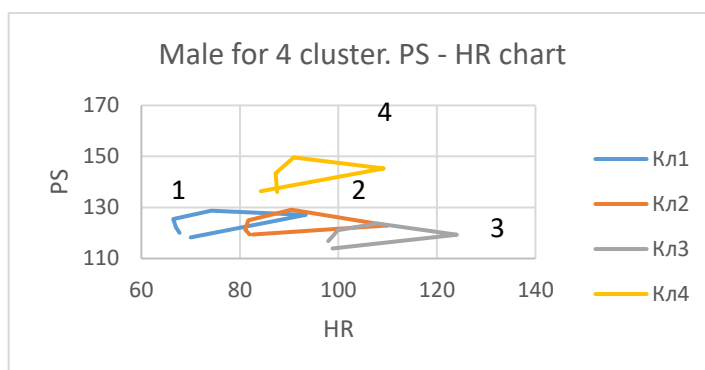


Рисунок 5.6. Графіки АТС, ЧСС для 4 кластеру

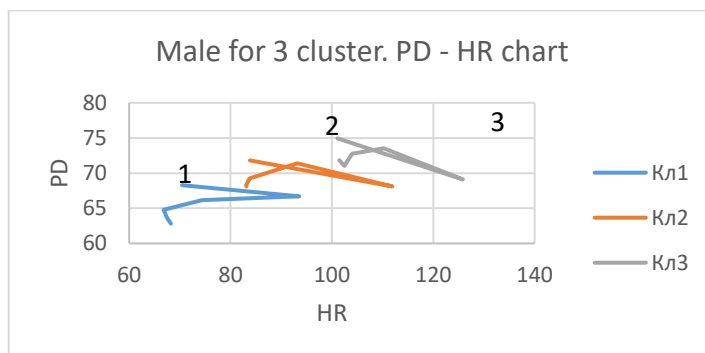


Рисунок 5.7. Графіки АД, ЧСС для 3 кластеру

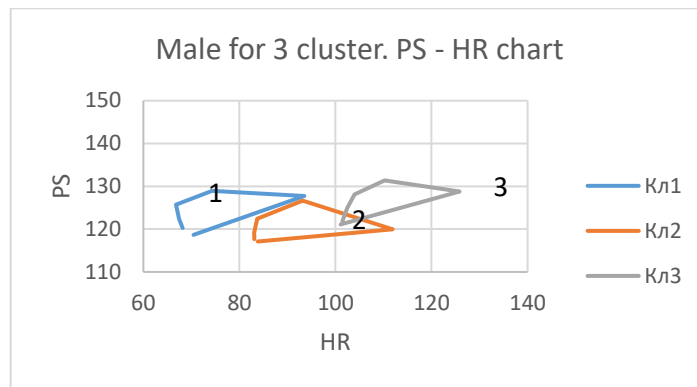


Рисунок 5.8. Графіки АТС, ЧСС для 3 кластеру

Графіки, на базі трьох кластерів, є найменш інформативними, оскільки втрачається група 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 можемо зробити висновок, що найбільш інформативними є графіки на базі п'яти кластерів. Також можемо спостерігати як змінюється дисперсія зі зменшенням кількості кластерів.

Внутрішньогрупова дисперсія за кожною змінною для 7 кластерів								Внутр. дисперсія за кожною змінною для 5 кластерів					
Param	1	2	3	4	5	6	7	Param	1	2	3	4	5
D	D	D	D	D	D	D	D	D	D	D	D	D	D
ATC0	93,0637	90,5987	90,2282	92,9467	159,9894	190,6435	94,8673	ATC0	144,464	92,1503	73,5938	99,5451	73,2506
ATD0	39,5716	38,353309	38,971788	32,3257	36,580867	75,270048	38,9002	ATD0	60,387	34,9539	28,6368	54,1451	42,4045
ЧСС0	43,695322	71,277519	122,75867	81,662	93,773256	118,78696	72,1388	ЧСС0	129,459	59,4944	58,4337	92,8756	68,0506
ATC1	355,0158	234,05253	205,25651	219,956	520,29598	460,38696	189,871	ATC1	217,43	237,468	185,055	243,439	164,677
ATD1	90,2954	90,0503	81,4369	52,4706	81,1564	108,5357	64,317	ATD1	53,1382	102,004	73,2289	82,5902	67,9737
ЧСС1	82,957332	84,140489	144,82228	133,569	161,32928	122,12029	114,318	ЧСС1	137,264	118,05	113,801	103,202	160,233
ATC2	180,5021	125,3932	131,5196	84,6397	242,7077	260,3034	85,805	ATC2	105,287	153,743	120,886	148,188	89,0026
ATD2	61,05926	51,488611	58,38823	35,9042	59,692918	89,225604	52,3055	ATD2	38,7884	65,7632	57,2886	41,478	55,1051
ЧСС2	66,063685	72,653548	136,33217	107,489	123,39271	108,30531	100,428	ЧСС2	178,694	82,6551	98,3333	124,152	78,2667
ATC3	128,442	88,298688	97,839443	76,3909	122,50951	198,16618	76,5497	ATC3	105,318	132,45	83,8548	103,389	90,4385
ATD3	55,6378	54,067979	32,289039	24,7112	62,34408	85,143961	51,5321	ATD3	51,3275	47,3831	33,1624	38,4939	42,7282
ЧСС3	46,506953	59,352177	142,0655	110,241	186,15645	90,833333	75,2813	ЧСС3	166,965	73,1777	58,2836	69,5744	47,0231
ATC4	104,68331	70,729636	85,662264	95,1891	147,41385	138,51063	64,7283	ATC4	69,7874	122,48	69,057	86,2512	99,3019
ATD4	48,894279	44,112642	47,413926	39,4887	68,149577	50,336715	47,6386	ATD4	35,5812	48,4314	43,9774	28,639	51,8199
ЧСС4	51,183628	65,713119	148,3292	113,267	67,637949	63,088406	76,2515	ЧСС4	152,354	67,68	59,9742	58,9024	43,9173
ATC5	109,86945	65,334833	80,557396	58,4341	151,88261	185,59807	65,4658	ATC5	80,2536	99,5539	48,2813	99,1402	86,8917
ATD5	43,10172	40,904512	44,209214	26,3284	54,528239	29,213527	49,9532	ATD5	47,4324	49,147	39,5061	82,6878	44,625
ЧСС5	55,532351	65,759155	133,07063	118,152	66,280177	61,947826	87,9236	ЧСС5	141,271	81,2025	52,9344	54,5451	26,9744

Рисунок 5.9. Зміна дисперсій для 7 і 5 кластерів

Внутр. дисп. за к-ю змінною для 4 кластерів					Внутр. дисп. за к-ю зм-ю для 3 кл-в			
Param	1	2	3	4	Param	1	2	3
D	D	D	D	D	D	D	D	D
АТС0	129,233	83,2433	94,9273	76,2444	АТС0	120,226	84,1223	147,217
АТД0	41,277	50,7324	52,7455	33,7357	АТД0	40,2482	54,7392	72,2253
ЧСС0	67,0105	55,1355	77,0273	48,4786	ЧСС0	70,2291	60,5598	76,6065
АТС1	329,597	202,497	220,245	365,086	АТС1	317,882	171,156	270,588
АТД1	106,314	109,517	83,3	74,3071	АТД1	102,525	117,614	116,375
ЧСС1	125,049	111,546	112,225	152,921	ЧСС1	120,209	115,731	119,888
АТС2	181,711	122,554	140,564	109,959	АТС2	175,485	120,565	172,204
АТД2	68,9226	90,621	57,4828	56,8087	АТД2	67,7151	94,6472	72,5324
ЧСС2	91,3468	86,02	128,634	99,1706	ЧСС2	90,9917	86,8516	105,765
АТС3	171,153	99,4755	120,465	99,7802	АТС3	167,027	93,7539	124,963
АТД3	54,7753	75,3283	71,901	28,9302	АТД3	53,751	85,7989	68,2214
ЧСС3	82,1204	55,9746	87,4364	68,0786	ЧСС3	83,7362	69,164	59,6479
АТС4	148,891	72,6739	74,4525	100,835	АТС4	141,725	60,4539	96,9776
АТД4	51,2866	71,2374	46,6818	57,3683	АТД4	53,2258	76,1853	57,1567
ЧСС4	73,3267	57,0283	79,6283	84,254	ЧСС4	75,7185	74,0276	47,5514
АТС5	127,592	64,4149	109,513	76,9683	АТС5	123,207	63,5186	121,676
АТД5	51,6382	60,4523	61,7101	34,0786	АТД5	52,1272	62,8414	101,198
ЧСС5	79,3649	47,463	64,8818	94,9968	ЧСС5	82,2105	55,7403	44,183

Рисунок 5.10. Зміна дисперсій для 4 і 3 кластерів

З порівняльної таблиці бачимо, що дисперсія досить сильно змінюється при зменшенні кластерів на 2 і 3 позиціях, 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 було вирішено розширити функціонал програмного продукту Clusterbox і додати до нього можливість вибору результуючої таблиці на 5 кластерів.

Таким чином, при запуску основного програмного продукту, виводиться вікно з вибором режиму кластеризації. Воно було розроблене спеціально, щоб у користувача була можливість класифікувати в автоматичному режимі за наявними в програмі результуючими таблицями всю базу даних.



Рисунок 5.11. Вибір режиму кластеризації

При натисненні на клавішу «Глобальний режим» відкривається вікно для проведення глобальної кластеризації.

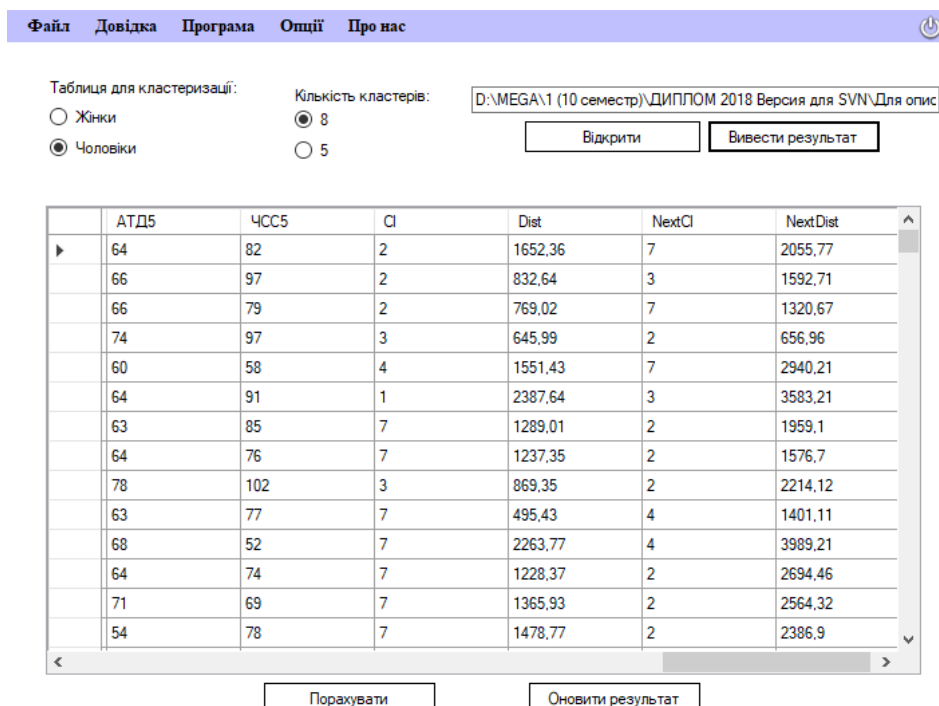


Рисунок 5.12 – Глобальна кластеризація

Для проведення дослідження необхідно обрати з пункту «Кількість кластерів», яку результуючу таблицю слід використовувати для

кластеризації, а також визначити для якої частини бази буде проведена кластеризація. Наступним кроком є завантаження таблиці для дослідження. Якщо вона вже містить значення кластеру, мітки (субкластеру), мінімальної та субмінімальної відстані, то можна вивести дані на екран. В іншому випадку необхідно буде провести дослідження, а потім оновити результат.

Слід зазначити, що якщо база даних містить значення кластеру, то при натисненні клавіші «Порахувати» впливе попередження: якщо натиснути так, тоді дані будуть перезаписані. Для прикладу проведемо кластеризацію з використанням результуючої таблиці на 5 кластерів та перезаписом результатів попередніх досліджень.

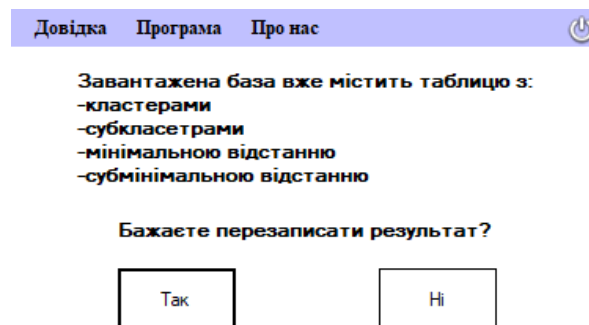


Рисунок 5.13. Вікно з попередженням

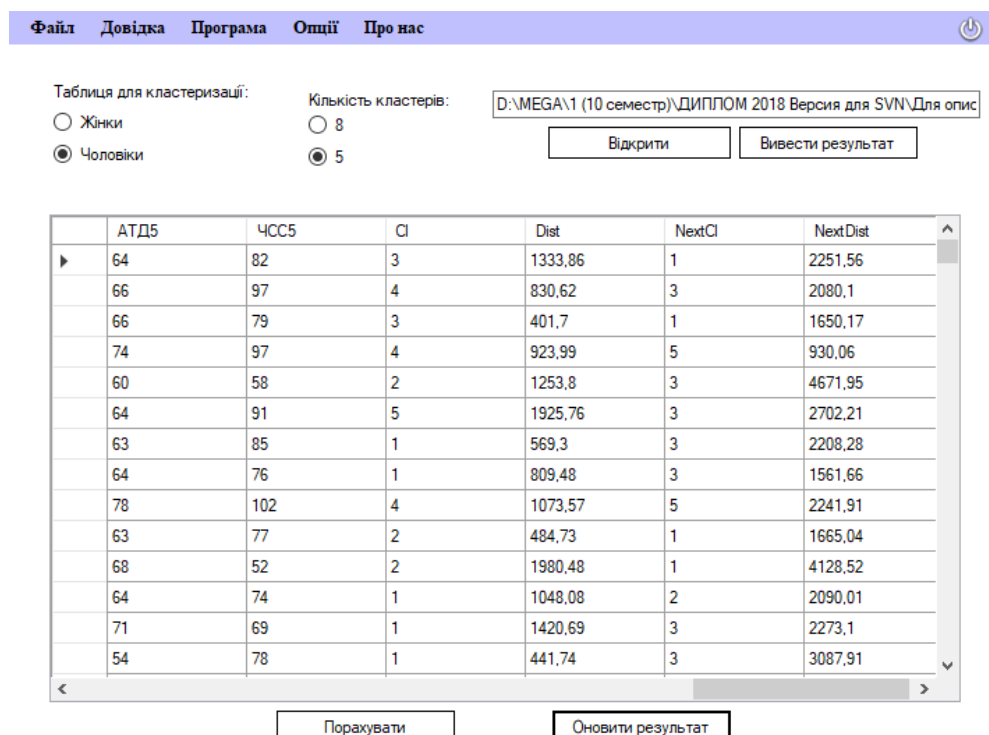


Рисунок 5.14. Оновлене вікно глобальної кластеризації

При оновленні результату ми бачимо, що дані були успішно перезаписані.

З вікна «Глобальна кластеризація» тож можемо легко перейти до вікна «Одинична кластеризація» за допомогою контекстного меню.

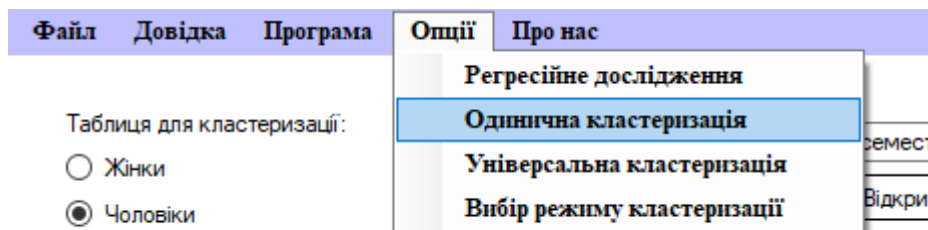


Рисунок 5.15. Вибір одиничної кластеризації

Також можна перейти до модулю проведення регресійного аналізу, що було розглянуто у розділі 4.

Вікно для одиничної кластеризації було модифіковано таким чином, що ми можемо вибрати результуючу таблицю, яка буде використана для розрахунку мінімальної відстані до кластеру.

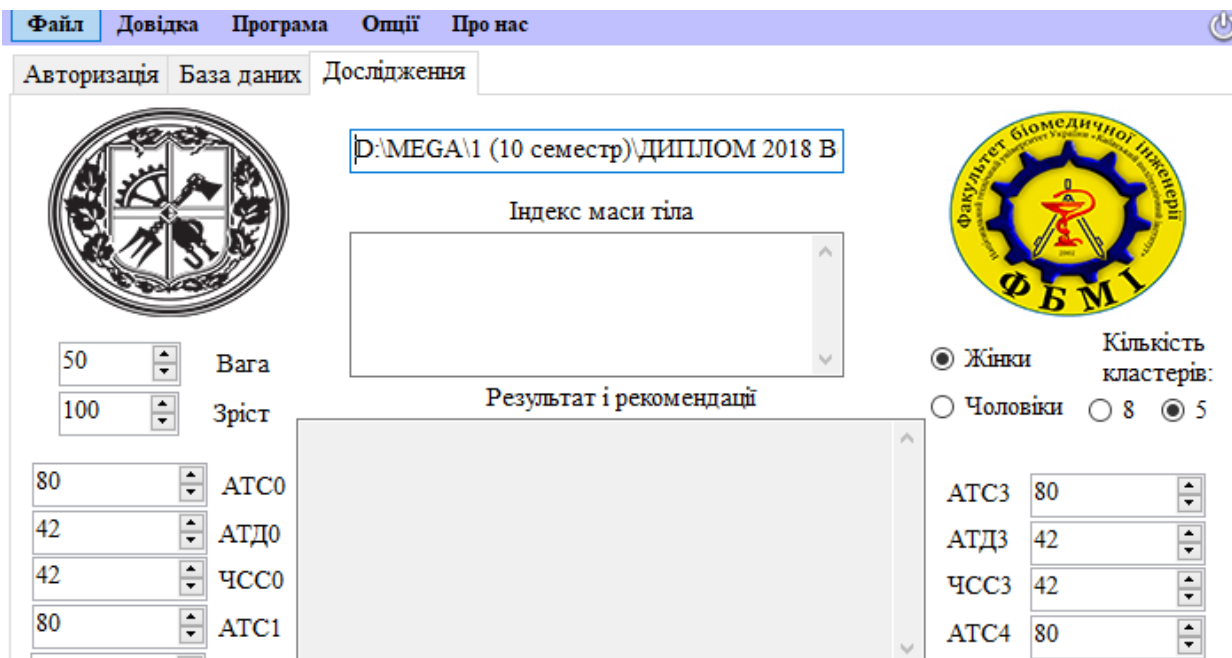


Рисунок 5.16. Вікно одиничної кластеризація

Для роботи в цьому режимі необхідно власноруч заповнити дані артеріального тиску та пульсу в стані спокої і на кожній хвилині після

навантаження, включно до п'ятої. Якщо дані було введено невірно, тоді виникає попередження, що зображено на рис. 4.18.

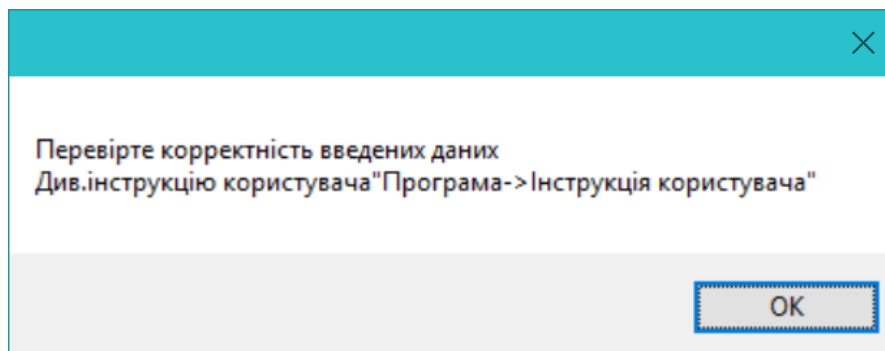


Рисунок 5.17. Попередження про невірність введення даних

Також на цьому вікні ми бачимо модуль для розрахунку індексу маси тіла (ІМТ), що використовується для оцінки загальної кількості жиру в організмі. Для розрахунку ІМТ, потрібно поділити вагу в кілограмах на ріст в метрах, піднесений до квадрату. Програмним продуктом передбачено верхні та нижні межі для полів введення даних, тому при значеннях зросту і ваги, що стоять за замовчуванням, виникає наступне попередження, зображене на рис. 5.18.

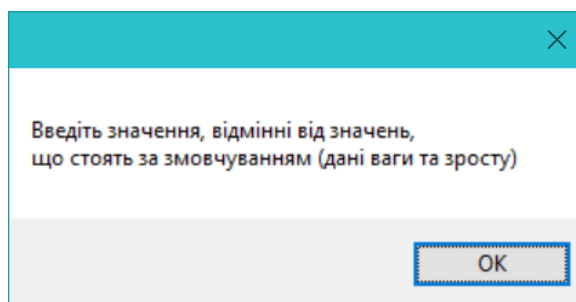



Рисунок 5.18. Попередження введення даних відмінних від даних за замовчуванням

При даному попередженні розрахунок ІМТ не відбувається

При введенні значень, відмінних від стандартних, відбувається розрахунок ІМТ і вивід необхідних характеристик. Приклад розрахунку зображено на рис. 5.19.




d:\MEGA\1 (10 семестр)\ДИПЛОМ 2018 Вє

Індекс маси тіла

Індекс маси тіла складає: 25,76

*Характеристика:
Надлишкога вага тіла (передожиріння)



Вага
 Зріст

☐ Жінки

Кількість кластерів:
☒ Чоловіки ☒ 8 ☐ 5

Рисунок 5.19. Вікно розрахунку індексу маси тіла

Розрахунок мінімальної відстані до кластеру відбувається незалежно від визначення індексу маси тіла. При коректному введенні даних АТС, АТД, ЧСС запускається алгоритм розрахунку квадрату евклідової відстані з подальшим знаходження мінімальної відстані. Коли мінімальна відстань знайдена, відбувається визначення кластера, до якого відноситься студент. Визначивши кластер, виводиться результат розрахунку, відповідні характеристики кластеру та рекомендації. Приклад роботи алгоритму з виведенням результатів зображено на рис. 5.20.

Файл Довідка Програма Опції Про нас

Авторизація База даних Дослідження



Вага
 Зріст

<input type="text" value="153"/>	АТС0
<input type="text" value="80"/>	АТД0
<input type="text" value="74"/>	ЧСС0
<input type="text" value="165"/>	АТС1
<input type="text" value="66"/>	АТД1
<input type="text" value="91"/>	ЧСС1
<input type="text" value="150"/>	АТС2
<input type="text" value="75"/>	АТД2
<input type="text" value="79"/>	ЧСС2

d:\MEGA\1 (10 семестр)\ДИПЛОМ 2018 Вє

Індекс маси тіла

Індекс маси тіла складає: 25,76

*Характеристика:
Надлишкога вага тіла (передожиріння)

Результат і рекомендації

Студент відноситься до кластеру *5*

Мінімальна відстань до кластеру становить:
1343,76

Кластер має наступні характеристики. Граничні високі значення артеріального тиску. Найбільша ефективність роботи серця. Переважання симпатикотоміної регуляції за індексом Кердо



☐ Жінки

Кількість кластерів:
☒ Чоловіки ☒ 8 ☐ 5

АТС3	<input type="text" value="147"/>
АТД3	<input type="text" value="78"/>
ЧСС3	<input type="text" value="77"/>
АТС4	<input type="text" value="146"/>
АТД4	<input type="text" value="85"/>
ЧСС4	<input type="text" value="81"/>
АТС5	<input type="text" value="146"/>
АТД5	<input type="text" value="82"/>
ЧСС5	<input type="text" value="81"/>

Рисунок 5.20 – Розрахунок мінімальної відстані до кластеру

Оскільки нам відомі всі значення АТС, АТД та ЧСС для кожного кластеру, ми можемо визначити приблизний радіус кластеру. Дана процедура необхідна для того, щоб порівнювати значення радіусу з нашим об'єктом. У тому випадку, коли радіус не охоплює дані об'єкта, необхідно вивести субмінімальну відстань та характеристики наступного кластеру. Для реалізації радіусу кластера необхідно визначити центр кластеру, що характеризується середнім значенням всіх змінних кластеру. Від кожного об'єкта кластера віднімаємо середнє значення, а результат беремо по модулю. Далі необхідно знайти суму результатів та розділити її на кількість елементів у кластері. Розрахована відстань і буде характеризувати наш радіус.

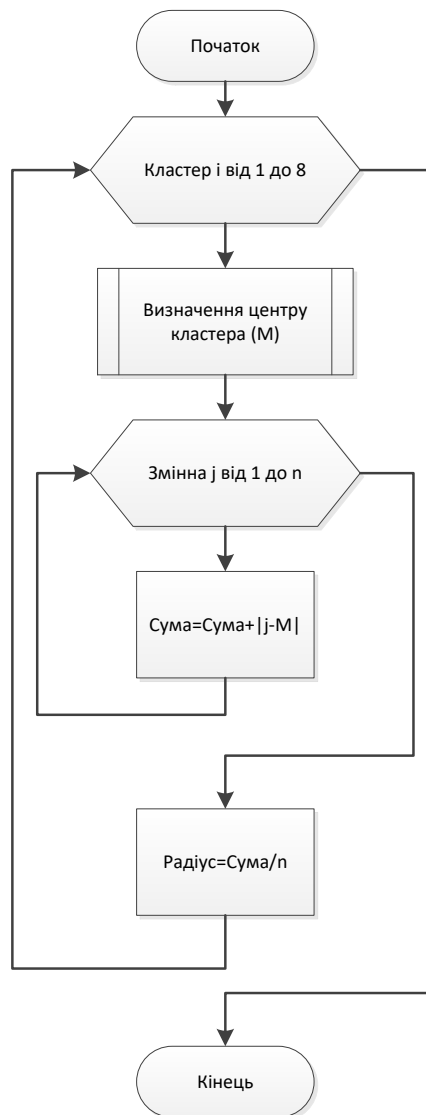


Рисунок 5.21. Блок-схема алгоритму знаходження радіусу кластера

Програмним продуктом передбачено порівняння мінімальної відстані до кластеру з радіусом кластеру. Якщо мінімальна відстань перевищує радіус, відбувається розрахунок субмінімальної відстані до кластеру і визначення номера кластеру з подальшим виводом його характеристик до відповідного поля. (рис.5.22).

Файл Довідка Програма Опції Про нас

Авторизація База даних Дослідження

Індекс маси тіла

Індекс маси тіла складає: 25,76
*Характеристика:
Надлишкового вага тіла (передожиріння)

Вага: 78
Зріст: 174

Результат і рекомендації

Студент відноситься до кластеру *1*
Мінімальна відстань до кластеру становить: 8139,3

Кластер має наступні характеристики: Граничні високі значення артеріального тиску. Помірно виражена симпатикотонія. *Рекомендації: Обмеження з занять важкою атлетикою. При занятті фізичними вправами та спорту необхідний

Наступна відстань після мінімальної становить *8555,75*
Це кластер *6*

Кластер має наступні характеристики: Функціональний стан кровообігу різко знижений. Максимальні значення ударного об'єму лівого шлуночка - як реакція на фізичне навантаження.

Порахувати

Кількість кластерів: 8

Жінки Чоловіки

АТС3 147
АТД3 78
ЧСС3 77
АТС4 146
АТД4 85
ЧСС4 81
АТС5 165
АТД5 94
ЧСС5 97

Рисунок 5.22 – Робота програмного продукту

Програмним продуктом також передбачено виведення результуючої таблиці до відповідного вікна програми «База даних». Приклад вікна наведено на рис. 5.23

Файл Довідка Програма Про нас

Авторизація База даних Дослідження

Parametr	C1	C2	C3	C4	C5	C6
Parametr	M	M	M	M	M	M
АТС0	115,61	130,40	116,63	122,07	127,03	1
АТД0	75,20	85,84	74,21	78,68	84,48	6
ЧСС0	82,82	83,84	92,99	68,51	99,08	7
АТС1	127,36	146,95	125,70	139,34	139,30	1
АТД1	77,26	86,81	77,91	80,76	88,85	7
ЧСС1	107,05	106,95	119,28	91,10	127,10	1
АТС2	123,44	140,33	124,56	133,10	136,33	1
АТД2	76,19	84,26	75,43	78,00	86,28	6
ЧСС2	88,94	90,19	105,92	73,32	113,63	7
АТС3	117,79	132,42	118,13	125,39	131,38	1
АТД3	73,88	81,09	72,07	74,98	82,90	6
ЧСС3	83,21	86,19	101,09	67,88	108,00	7
АТС4	115,49	127,88	114,69	122,80	126,05	1
АТД4	71,77	79,91	70,59	73,29	81,80	6
ЧСС4	82,97	85,63	98,40	68,29	108,58	7
АТС5	113,29	125,51	113,05	118,78	124,40	1
АТД5	71,49	79,23	69,87	72,88	81,30	6
ЧСС5	83,23	83,37	96,58	67,45	104,93	7

Рисунок 5.23. Вікно бази даних

Після проведення необхідних аналізів програмою передбачено збереження. При збереженні файлу перевіряється наявність бази даних з назвою «Students.mdb», відповідні таблиці і поля для реєстрації даних. При існуванні БД дані зберігаються до неї. Якщо БД не існує, відбувається створення бази даних Access, куди записуються дані авторизації, а також показники тиску та пульсу, що були введені при розрахунках і номер кластеру до якого відноситься студент. Приклад БД зображено на рис. 5.24

АТД2 ▾	АТД3 ▾	АТД4 ▾	АТД5 ▾	ЧСС0 ▾	ЧСС1 ▾	ЧСС2 ▾	ЧСС3 ▾	ЧСС4 ▾	ЧСС5 ▾	СІ ▾
42	42	42	42	42	42	42	42	42	42	7
129	129	129	129	155	155	155	155	155	155	5
73	72	70	70	56	87	59	72	68	70	8
71	75	71	63	87	100	88	88	84	85	2
75	65	75	72	66	106	79	70	68	64	4
68	62	61	62	95	108	105	102	96	100	3
69	64	61	60	68	101	70	68	65	63	6
86	93	77	67	82	102	84	86	73	82	1
61	63	59	60	69	102	87	87	83	81	7
75	77	85	82	64	81	69	67	71	71	5

Рисунок 5.24 – Приклад бази даних Access, в яку зберігаються результати тестування

Висновок до розділу 5

Даний розділ розкриває сутність роботи. В ньому розкрито і обґрунтовано зменшення кількості кластерів. Розглянуто умови та процес зміни кількості кластерів в залежності від їх порівняння на побудованих графіках. Нами було отримано декілька нових результуючих таблиць і визначено оптимальну кількість кластерів для чоловіків, що становить 5 інформативних груп. Наведено додаткові модулі для дослідження, серед яких є глобальна кластеризація. Також проведено дисперсійний аналіз для порівняння зміни дисперсій із зменшенням кількості кластерів.

Поданий детальний опис алгоритму роботи з програмним додатком у розрізі проведення певного дослідження в базі даних. Надана інструкція користування програмним додатком для отримання результатів.

ЗАГАЛЬНІ ВИСНОВКИ

У результаті роботи над магістерською дисертацією було опрацьовано багато наукових джерел інформації стосовно теми дисертації, проаналізовано основні методи реалізації алгоритму квадрату евклідової відстані та побудови регресійних рівнянь, набуто професійних вмінь та навичок у роботі з сучасними інформаційними технологіями.

У ході виконання проекту нами було проаналізовано результуючі таблиці та графіки, отримані на основі бази даних молодших курсів НТУУ «КПІ ім. Сікорського», що містить 1495 спостережень, проаналізовано нову базу даних, що містить 599 спостережень, розщеплено її по статі та проведено глобальну кластеризацію чоловічої групи. Нами було проведено логістичну регресію та дискримінантний аналіз та підтверджено ефективність алгоритму квадрата евклідової відстані. На основі цього нами було розроблено програмний додаток для визначення подібності функціональних патернів, визначено кількість оптимальних кластерів та побудовано нові результуючі таблиці для автоматизації процесу визначення регуляторних реакцій на тестове навантаження.

Результатом магістерської дисертації стала комп'ютерна система для визначення стану системи кровообігу з додатковими модулями глобальної кластеризації, універсальної кластеризації та модулем знаходження подібних кластерів. Програмним продуктом також передбачено збереження результатів кластеризації до завантаженої бази даних шляхом створення нової таблиці з відповідними стовпчиками.

На даному етапі програмний продукт повністю готовий до використання. Програма може бути вдосконалена та доповнена новими функціями.