**Національний технічний університет України**

**«Київський політехнічний інститут**

**імені Ігоря сікорського»**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_факультет біомедичної інженерії

(повна назва інституту/факультету)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_кафедра біомедичної кібернетики

(повна назва кафедри)

«До захисту допущено»

Завідувач кафедри БМК

\_\_\_\_\_\_\_\_\_\_ \_\_Є.А. Настенко\_

(підпис) (ініціали, прізвище)

“\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_\_2017 р.

**Дипломна робота**

**на здобуття ступеня бакалавра**

|  |  |  |  |
| --- | --- | --- | --- |
| З напряму підготовки | | | 6.050101 «Комп’ютерні науки» |
| зі спеціальності | |  | |
| (код і назва) | | | |
| на тему: | Оцінка функціональних реакцій на тестове навантаження у студентів | | |
| 1-2 курсу. Чоловіки. | | | |
|  | | | |

Виконав: студент \_4\_ курсу, групи \_БС-31\_\_

(шифр групи)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ВОЙНИК БОГДАН ОЛЕКСІЙОВИЧ** | | | | | |  |  |
| (прізвище, ім’я, по батькові) | | | | | |  | (підпис) |
| Керівник | зав каф БМК, д.б.н., к.т.н., с.н.с, Настенко Є.А. | | | | |  |  |
| (посада, науковий ступінь, вчене звання, прізвище та ініціали) | | | | | |  | (підпис) |
| Консультант з розділів ДР | | | ст.викл. Носовець О.К. | | |  |  |
| (назва розділу) ( посада, вчене звання, науковий ступінь, прізвище, ініціали) | | | | | |  | (підпис) |
| Консультант з охорони праці | | | | | доцент, доцент, к.т.н. Демчук Г.В. |  |  |
| (назва розділу) | | | | ( посада, вчене звання, науковий ступінь, прізвище, ініціали) | |  | (підпис) |
| Рецензент | | ас. каф. БМІ Захарчук Н.В. | | | |  |  |
| (посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) | | | | | |  | (підпис) |

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_

(підпис)

Київ – 2017 року

**Національний технічний університет України**

**«Київський політехнічний інститут**

**імені Ігоря сікорського»**

|  |  |  |  |
| --- | --- | --- | --- |
| Інститут (факультет) | | | Біомедичної інженерії |
| (повна назва) | | | |
| Кафедра | Біомедичної кібернетики | | |
| (повна назва) | | | |
| Рівень вищої освіти – перший (бакалаврський) | | | |
| Напрям підготовки | | 6.050101 «Комп’ютерні науки» | |
| (код і назва) | | | |

ЗАТВЕРДЖУЮ

Завідувач кафедри БМК

\_\_\_\_\_\_\_\_\_\_ \_Є.А. Настенко\_

(підпис) (ініціали, прізвище)

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2017 р.

**ЗАВДАННЯ**

**на дипломну роботу студенту**

|  |  |  |
| --- | --- | --- |
| **ВОЙНИКА БОГДАНА ОЛЕКСІЙОВИЧА** | | |
| (прізвище, ім’я, по батькові) | | |
| 1. Тема роботи | **Оцінка функціональних реакцій на тестове навантаження** | |
| **у студентів 1-2 курсу. Чоловіки.** | | |
|  | | |
| керівник роботи | | зав каф БМК, д.б.н., к.т.н., с.н.с,Настенко Євгеній |
|  | | Арнольдович |
|  | | (прізвище, ім’я, по батькові, науковий ступінь, вчене звання) |
| затверджені наказом по університету від «\_\_\_»\_\_\_\_\_\_\_\_\_ 2017 р. №\_\_\_\_\_ | | |

|  |  |  |  |
| --- | --- | --- | --- |
| 2. Термін подання студентом роботи | | | ***12-14 червня 2017 року*** |
| 3. Вихідні дані до роботи | | **База даних студентів НТУУ «КПІ ім. Ігоря** | |
| **Сікорського»** | | | |
|  | | | |
| 4. Зміст роботи | **Проведення аналізів, виведення таблиць і графіків для** | | |
| **чоловіків. Реалізація алгоритмів розрахунку мінімальної відстані та** | | | |
| **автоматичного створення бази даних.** | | | |
|  | | | |
| 5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо) | | | |
| **Робота містить 53 ілюстрації, 2 графіки, 1 блок-схему і план приміщення** | | | |
|  | | | |
|  | | | |

6. Консультанти розділів роботи

|  |  |  |  |
| --- | --- | --- | --- |
| Розділ | Прізвище, ініціали та посада  консультанта | Підпис, дата | |
| завдання  видав | завдання прийняв |
| Дипломної роботи | Носовець О.К., ст.викл. |  |  |
| Охорони праці | Демчук Г.В., доц., к.т.н. |  |  |
|  |  |  |  |

7. Дата видачі завдання ***12 травня 2017 р.***

Календарний план

|  |  |  |  |
| --- | --- | --- | --- |
| № з/п | Назва етапів виконання  дипломної роботи | Термін виконання  етапів роботи | Примітка |
| 1 | Розділ ДР з «Безпеки життєдіяльності та охорони здоров’я» | 10 травня 2017р. |  |
| 2 | Отримати завдання на ДР | 12 травня 2017р. |  |
| 3 | Робота з літературою та інтернет ресурсами по темі дипломної роботи | 13-15 травня 2017р. |  |
| 4 | Підготовка розділів дипломної роботи на основі отриманих даних | 16 травня 2017р. |  |
| 5 | Виведення описових статистик та проведення дисперсійного аналізу для чоловіків | 17 травня 2017р. |  |
| 6 | Побудова результуючих таблиць для чоловіків | 18 травня 2017р. |  |
| 7 | Побудова графіків тиску та пульсу | 19 травня 2017р. |  |
| 8 | Надання характеристик кластерів у чоловіків | 20-21 травня 2017р |  |
| 9 | Реалізація алгоритму розрахунку мінімальної відстані | 22-24 травня 2017р |  |
| 10 | Реалізація алгоритмів автоматичного створення бази даних та збереження результатів | 25-26 травня 2017р |  |
| 11 | Розробка вікон для інструкції користувача | 27 травня 2017р. |  |
| 12 | Оформлення ДР | 27-28 травня 2017р. |  |
| 13 | Проходження нормоконтролю по оформленню ДР | 29 травня 2017р -  7 червня 2017р |  |
| 14 | Предзахист ДР та допуск до захисту ДР | 7-9 червня 2017р. |  |
| 15 | Подання ДР рецензенту. Отримання рецензії. | 12-14 червня 2017р |  |
| 16 | Подання в електронному вигляді ДР та анотації до неї на сайт кафедри. | 14 червня 2017р |  |
| 17 | Подання пакету документів по ДР до захисту в ЕК | 14 червня 2017р. |  |
| 18 | Захист ДР в ЕК | 19-24 червня 2017р |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент |  |  |  | Б.О. Войник |
|  |  | (підпис) |  | (ініціали, прізвище) |
| Керівник роботи |  |  |  | Є.А. Настенко |
|  |  | (підпис) |  | (ініціали, прізвище) |

АНОТАЦІЯ

Обсяг пояснювальної записки становить 105 сторінок, 53 ілюстрації, 27 таблиць та 47 джерел за переліком посилань.

Метою дослідження є розробка програмного забезпечення для визначення мінімальної відстані до кластеру.

Завдання:

* Провести аналіз наявних моделей, методів та інформаційних систем, призначених для оцінки реакцій на тестове навантаження;
* Вивести описові статистики, провести дисперсійний аналіз та проаналізувати міри відстані;
* Побудова характеристик кластерів;
* Побудова результуючих таблиць;
* Проектування та реалізація програмного продукту для визначення мінімальної відстані до кластеру.

Розробка була здійснена засобами мови програмування C# в середовищі розробки Microsoft Visual Studio 2013 та за допомогою статистичного пакету IBM SPSS Statistics 20.

Результати дипломної роботи описані в статті: Г.Л. Бойко, Б.О. Войник, М.О. Федчишин, Є.А. Настенко, О.К. Носовець, В.А. Павлов Оцінка умов коронарного кровопостачання за модифікованою пробою Мартіне у студентів молодших курсів // Вісник Університету «Україна» – серія: Інформатика, обчислювальна техніка та кібернетика. – №1 (20). – липень 2017.

Проект розроблено на замовлення кафедри «Фізичного виховання» НТУУ «КПІ ім. Ігоря Сікорського» та буде впроваджений в навчальну програму в 2017-2018 н.р. (акт впровадження від 9 червня 2017 р.)

Ключові слова: кластер, дисперсійний аналіз, характеристики кластерів, програмний продукт.

ABSTRACT

This explanatory note contains a total of 105 pages, 53 illustrations, 27 tables and 47 sources for references.

The purpose of research is to develop a program product to find minimum distance to the cluster

Tasks:

* Analyze existing models, methods and information systems designed to assess reactions to the test load;
* Display descriptive statistics, analysis of variance to conduct and analyze measure distances;
* Construction of the characteristics of clusters;
* Construction of the resulting tables;
* Design and implementation of software to determine the minimum distance to the cluster.

Development was carried out by means of the С# programming language in Microsoft Visual Studio 2013 development environment and statistical package IBM SPSS Statistics 20.

The results of the diploma work are described in the article: Г.Л. Бойко, Б.О. Войник, М.О. Федчишин, Є.А. Настенко, О.К. Носовець, В.А. Павлов Оцінка умов коронарного кровопостачання за модифікованою пробою Мартіне у студентів молодших курсів // Вісник Університету «Україна» – серія: Інформатика, обчислювальна техніка та кібернетика. – №1 (20). – July 2017.

The project developed for the department "Physical Education" NTUU “Igor Sikorsky KPI” and will be introduced in the curriculum in 2017-2018 academic year (implementation Act of June 9, 2017.)

Key words: cluster, ANOVA, cluster characteristic, program product.

ЗМІСТ

[ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 7](#_Toc529958185)

[ВСТУП 8](#_Toc529958186)

[РОЗДІЛ 1 ЗАДАЧА З МЕДИЧНОЇ ТОЧКИ ЗОРУ 9](#_Toc529958187)

[1.1 Алгоритм квадрату евклідової відстані 9](#_Toc529958188)

[1.2 Опис бази даних студентів 9](#_Toc529958189)

[Висновки до розділу 1 10](#_Toc529958190)

[РОЗДІЛ 2 МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕННЯ 11](#_Toc529958191)

[2.1 Дисперсійний аналіз 11](#_Toc529958192)

[2.2 Регресійний аналіз 11](#_Toc529958193)

[2.2.1 Бінарна логістична регресія 11](#_Toc529958194)

[2.2.2 Лінійна регресія і метод найменших квадратів 11](#_Toc529958195)

[РОЗДІЛ 3 ПРОВЕДЕННЯ АНАЛІЗІВ ДЛЯ РОЗРОБКИ КОМП’ЮТЕРНОЇ СИСТЕМИ 12](#_Toc529958196)

[Висновки до розділу 3 21](#_Toc529958197)

[РОЗДІЛ 4 ПРОГРАМНИЙ ПРОДУКТ ДЛЯ ВИЗНАЧЕННЯ ФУНКЦІОНАЛЬНОГО СТАНУЧИСТЕМИ КРОВООБІГУ 21](#_Toc529958198)

[4.1 Проектування програмного продукту 21](#_Toc529958199)

[4.2 Розробка програмного продукту 30](#_Toc529958200)

[4.3 Робота з програмним продуктом 30](#_Toc529958201)

[Висновок до розділу 4 41](#_Toc529958202)

[ЗАГАЛЬНІ ВИСНОВКИ 42](#_Toc529958203)

[ПЕРЕЛІК ПОСИЛАНЬ 43](#_Toc529958204)

[ДОДАТОК А 46](#_Toc529958205)

[ДОДАТОК Б 49](#_Toc529958206)

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

АТС – систолічний артеріальний тиск

АТД – діастолічний артеріальний тиск

ЧСС – частота сердечний скорочень (пульс)

БД – база даних

ПП – програмний продукт

ІМТ – індекс маси тіла

ВНЗ – вищий навчальний заклад

ССС – серцево-судинна система

ВНС – вегетативна нервова система

ЛФК – лікувальна фізична культура

EM-алгоритм - Expectation-maximization (EM) algorithm

ANOVA – Analysis Of VAriance

GUI – Graphical user interface

# ВСТУП

Актуальність.

Зв'язок роботи з науковими програмами, планами, темами (за необхідності)

Мета роботи.

У відповідності з метою ставлять такі завдання:

Об’єкт дослідження.

Предмет дослідження.

Методи дослідження.

Наукова новизна одержаних результатів

Практичне значення одержаних результатів

Особистий внесок здобувача

Апробація результатів дисертації

Публікації

Структура дисертації

# РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ ЗА ТЕМОЮ І ВИБІР НАПРЯМІВ ДОСЛІДЖЕНЬ

## 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 квадра 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.

Дана функціональна проба дозволяє оцінити пристосування організму до фізичних навантажень різноманітного характеру і різноманітної інтенсивності.

Виділяють 5 основних типів реакції серцево-судинної системи: нормотонічний, гіпотонічний, гіпертонічний, дистонічний та східчастий.



Рисунок 1.1 – Типи реакцій серцево-судинної системи

Проба Мартіне використовується при масових обстеженнях осіб, які займаються фізичною культурою і спортсменів нижчих розрядів. У клінічній практиці вона може використовуватися для вивчення функціональних можливостей серцево-судинної системи осіб різних вікових категорій. Практичний досвід показав, що особам до 40 років без виражених відхилень у стані здоров'я можна давати 20 присідань за 30 сек, до 50 років - 15 присідань за 22 сек, більше 50 років-10 присідань за 15 сек. Функціональні особливості серцево-судинної системи вважаються як задовільні, якщо при оцінці проби її результати укладаються в нормотонічний тип реакції.

Можна використовувати пробу Мартіні з діагностичною метою: для визначення причини тахікардії в стані спокою. Якщо після проби показники вкладаються в несприятливий тип реакції, то тахікардія зумовлена захворювань серцево-судинної системи. Іноді до навантаження пульс лабільний і відновлення його йде хвилеподібно, може виникнути негативна фаза пульсу, а нерідко - пульс після навантаження стабілізується на показниках нижчих, ніж до навантаження. Це дає можливість припустити, що тахікардія в стані спокою зумовлена порушенням функціонального стану нервової системи. Якщо ж до навантаження показники ЧСС вище норми, після проби всі показники вкладаються в нормотонічний тип реакції, але пульс відновлюється до вихідних цифр (як до навантаження, підвищений) - можна припустити, що тахікардія в спокої зумовлена гіперфункцією щитовидної залози. Наступні цілеспрямовані поглиблені обстеження дадуть можливість виключити, 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 статні 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.

## 2.1 Дисперсійний аналіз

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

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

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

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

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

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

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

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

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

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

— загальна середня величина ознаки X;

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

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

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

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

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

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

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

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

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

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

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

Однофакторний дисперсійний аналіз

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

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

|  |  |
| --- | --- |
|  | (2.14) |

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

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

|  |  |
| --- | --- |
|  | (2.15) |

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

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

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

|  |  |
| --- | --- |
|  | (2.16) |

|  |  |
| --- | --- |
|  | (2.17) |

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

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

|  |  |
| --- | --- |
|  | (2.18) |

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

|  |  |
| --- | --- |
|  | (2.19) |

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

|  |  |
| --- | --- |
|  | (2.20) |

Справджується рівність

|  |  |
| --- | --- |
|  | (2.21) |

**В.** Визначаємо оцінки дисперсій:

|  |  |
| --- | --- |
|  | (2.22) |

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

|  |  |
| --- | --- |
|  | (2.23) |

Для заданого рівня значущості α знаходимо критичні значення статистики .

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

Таблиця 2.1 – Результати однофакторного дисперсійного аналізу

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

## 2.2 Регресійний аналіз

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

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

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

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

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

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

де - залежна змінна (вона завжди одна);

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

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

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

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

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

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

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

Де  ,

І коваріаційною матрицею:

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

Де – дисперсія *у,*

Рівняння регресії (умовне математичне очікування) матиме вигляд

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

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

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

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

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

Вирішується задача відшукання оцінки  вектора . Отримана регресія називається середньоквадратичною.

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

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

Отримана регресія називається середньоабсолютною (медіанною).

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

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

Отримана при цьому регресія називається мінмаксною.

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

Найбільш часто зустрічаються такі види рівнянь регресії:

* лінійне множинне;
* поліноміальний;
* гіперболічне.

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

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

де *y* - число спостережень;

розрахункове значення результатного фактора.

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

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

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

**Мета регресійного аналізу:**

* визначення ступеня детермінованості варіації критеріальної (залежної) змінної предикторами (незалежними змінними).
* пророкування значення залежної змінної за допомогою незалежної.
* визначення внеску окремих незалежних змінних у варіацію залежної. Регресійний аналіз не можна використовувати для визначення наявності зв'язку між змінними, оскільки наявність такого зв'язку і є передумова для застосування аналізу.
* побудова рівняння регресії, тобто знаходження виду залежності між результатними показником і незалежними факторами .
* оцінка значимості отриманого рівняння, тобто визначення того, наскільки вибрані факторні ознаки пояснюють варіацію ознаки *у*.

**Алгоритм регресійного аналізу**

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

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

де — задані функції.

У цьому випадку матрицю  називається регресійною матрицею.

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

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

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

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

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

У цьому випадку легко знайти оцінку

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

Звідси випливає, що обчислення  не залежить від обчислення інших

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

— поліном заданого степеня;

— відмінний від нуля в околиці точки .

## 2.2.1 Бінарна логістична регресія

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

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

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

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

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

де

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

де *X1* - значення незалежних змінних;

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

*а* - деяка константа.

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

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

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

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

## 2.2.2 Лінійна регресія і метод найменших квадратів

**Лінії тренду**

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

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

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

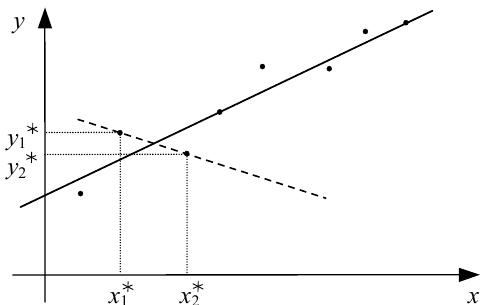
Після побудови ліній тренду на базі теоретично придатних залежностей, кожний результат пропонується оцінити шляхом ранжування за кількома критеріями, які характеризують достовірність, відповідність, надійність та інші параметри прогнозу.

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

**Метод найменших квадратів**

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

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

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

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

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

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

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

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

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

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

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

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

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

Від цієї процедури і назва методу — метод найменших квадратів.

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

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

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

Із (2.5) після низки нескладних перетворень отримаємо:

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

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

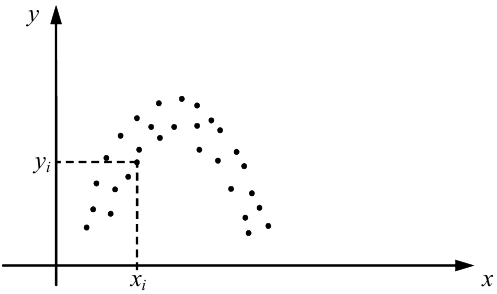
Рівняння, що входять у систему (2.6), називають нормальними рівняннями Гаусса. Коефіцієнтами у них є суми, які «згладжують» дію похибок вимірювань величин *x, y* і зменшують їх вплив на оцінки параметрів *b, a.* Завдяки цьому підвищується точність їх визначення.

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

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

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

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

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

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

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

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

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

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

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

2.3 Дискримінантний аналіз

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

* інтерпретація;
* класифікація.

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

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

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

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

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

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

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

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

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

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

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

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

Критерій значення Уілкса обчислюють за формулою (1.16):

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

де *К –* кількість значень;

*k –* число вже обчислених дискримінаційних функцій.

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

Значення ксі-квадрат розраховують за формулою (1.17):

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

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

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

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

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

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

uj — шукані коефіцієнти дискримінантної функції;

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

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

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

## 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 опис бази даних для попередніх досліджень і теоретичні відомості щодо досліджень, які будуть проведені для покращення або підтвердження істинності попередніх досліджень Також нами було вияснено 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 результуючій таблиці для кожної статі окремо було визначено за допомогою кластерного аналізу, що був проведений Настенко Є.А. та Носовець О.К.

Аналіз кластеризації – це статистична процедура, що є багатовимірною, і направлена на збір даних, вміст яких несе інформацію про об’єкти, а зокрема про їх вибірку, з подальшим впорядкуванням об’єктів в порівняно однакові класи, які ще називають групами чи кластерами. Кластеризація є одним із методів, що відносять до класу навчання без учителя [12].

Суттєвою особливістю аналізу кластеризації є те, що це зовсім незвичайний метод статистичного аналізу, так як частіше за все до нього неможливо застосувати алгоритми перевірки статистичної значимості, а його результат надає найбільш можливі вагомі значення. Маючи на вході масив даних, але не маючи ніяких апріорних гіпотез щодо класу цих даних, дуже часто дослідник спирається на описану вище особливість і застосовує алгоритм кластеризації.

Кластеризація має на меті організувати отримані дані у наглядні структури. Фактично, кластерний аналіз – це набір різних алгоритмів, направлених на класифікацію. Згідно з авторами підручника Statsoft – кластеризувати можна більшість симптомів захворювань, а також види їх лікування, крім цього це дає змогу досить цікаво їх класифікувати.

При класифікації великих масивів даних на групи кластерний аналіз є інструментом, який неможливо замінити, оскільки 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 застосовується також 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.

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

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

В результаті було отримано стовпчик з номерами кластерів, що характеризують групу ризику. База даних, що було використано для побудови результуючої таблиці для використання алгоритму визначення оптимальних характеристик студента, вже містила відповідний стовпчик з групами та представляла собою файл формату \*.xls (рис.1)

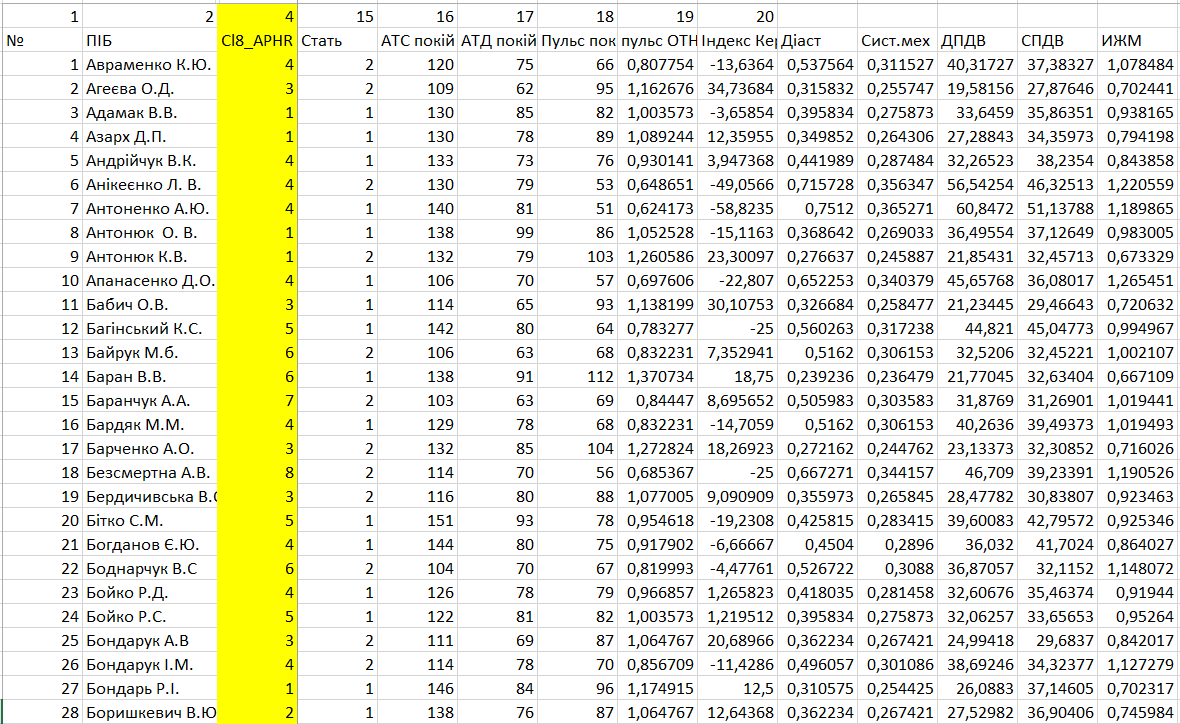


Рисунок 1.2 – Приклад бази даних студентів

База даних також містила показники дихальної системи (ємність легень, проба генче, проба штанге, частота дихання), нервової системи (проста зорова-моторна реакція, складна зорово-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 хвилинах після навантаження.

Вхідну таблицю було перетворено до прийнятного вигляд, яка без проблем сприймається програмою для обробки даних SPSS. Після цього проводиться дисперсійний аналіз для визначення середьного значення для кожної змінної в кожному кластері.

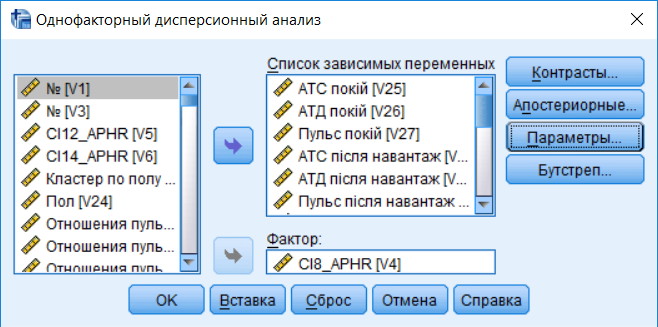


Рисунок 3.5 – Вибір необхідних параметрів для аналізу

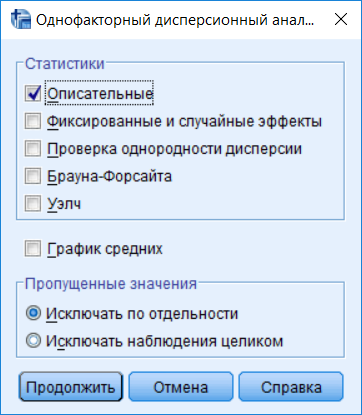


Рисунок 3.7 – Вибір потрібних статистик

Результатом проведення дисперсійного є наступна таблиця:

Таблиця 3.2 – Описові статистики по кластерам

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | N | Среднее | Стд. отклонение | Стд. Ошибка |
|
| АТС покій | 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 |
| Итого | 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 |
| Итого | 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 |
| Итого | 794 | 81,45 | 14,171 | ,503 |

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

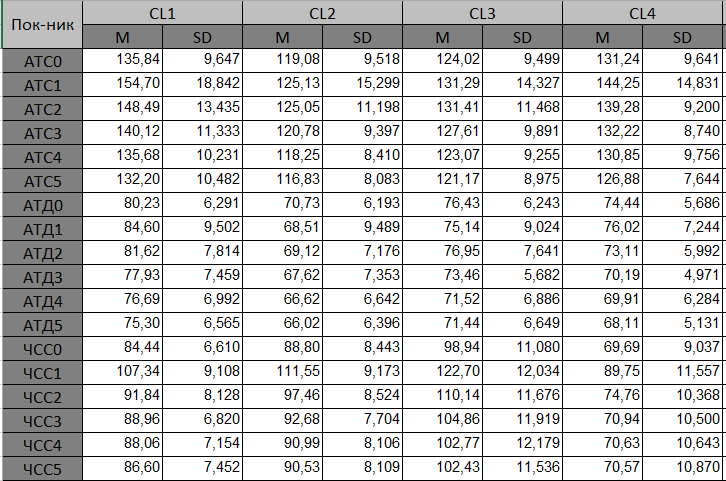


Рисунок 3.8 – Згрупована таблиця результатів

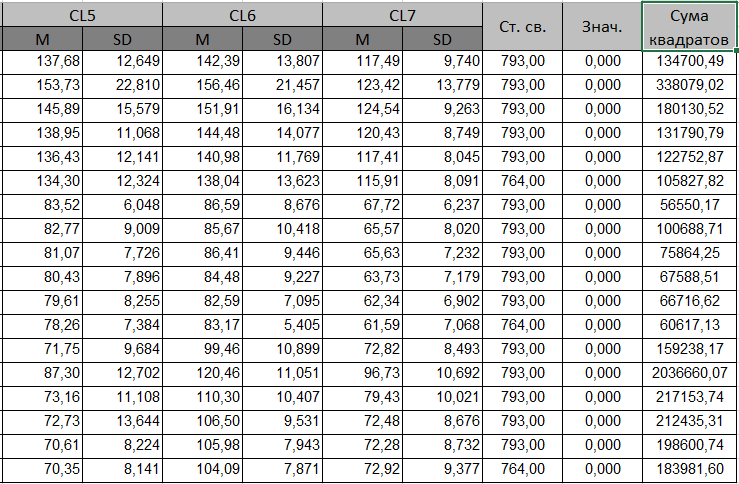


Рисунок 3.9 – Згрупована таблиця результатів

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 чоловічою, а 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 у нас не містить значень групи ризику в повному обсязі, тому ми не можемо охарактеризувати всіх студентів. Також не відомо чи відповідає колонка з кластерами дійсності, оскільки на цей раз у нас не було у наявності листа з розшифруванням та методами отримання кластерів.

Але ми можемо застосувати програмний продукт для визначення групи ризику і заодно переконатися в справній роботі алгоритму квадрата евклідової відстані. Задача перевірки ефективності алгоритму випливає з того, що попередні дослідження показали високий відсоток збіжності (80%) між алгоритмом кластеризації методом k-середніх та реалізованим у програмі алгоритмом знаходження мінімальної відстані.

# РОЗДІЛ 3 ПРОВЕДЕННЯ АНАЛІЗІВ ДЛЯ РОЗРОБКИ КОМП’ЮТЕРНОЇ СИСТЕМИ ДЛЯ ВИЗНАЧЕННЯ РЕАКЦІЙ НА ТЕСТОВЕ НАВАНТАЖЕННЯ

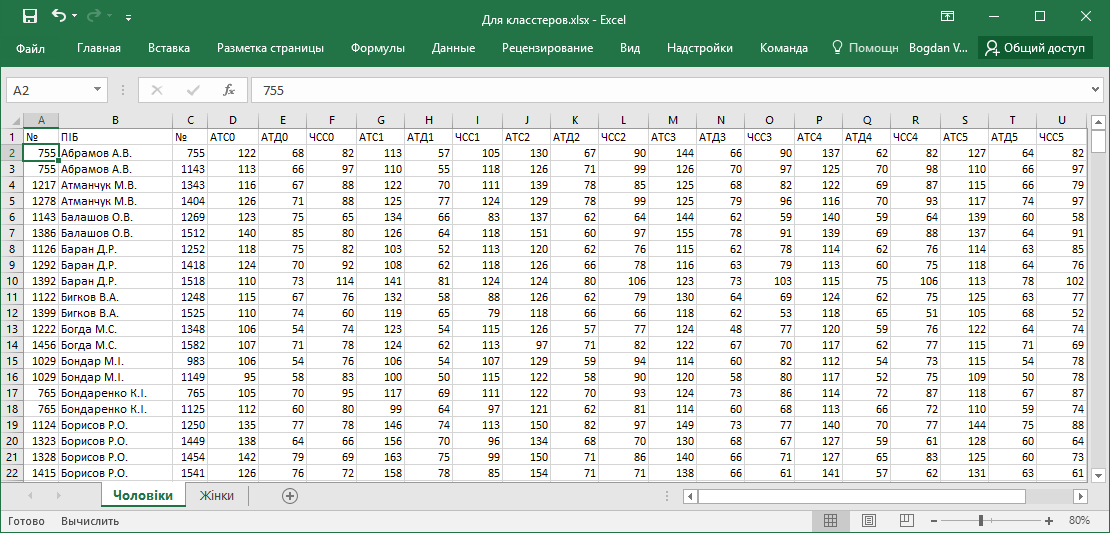
Для перевірки коректності алгоритму квадрата евклідової відстані необхідно розщепити базу даних по статі та провести кластеризацію для кожної половини окремо. Проведення кластеризації для кожного пацієнта окремо передбачає вдосконалення програмного продукту та розробки нового модулю кластеризації, який 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, що 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:

1. Приведення бази даних до прийнятного для SPSS виду
2. Розщеплення бази даних по статі
3. Проведення глобальної кластеризації
4. Проведення логістичної регресії
5. Проведення дискримінантного аналізу
6. Дослідження графіків, побудованих на базі результуючих таблиць
7. Розробка модулю для дослідження подібності кластерів (реалізація побудови ліній тренду)
8. Побудова нових результуючих таблиць
9. Розробка модулю універсальної кластеризації
10. Проведення дисперсійного аналізу
11. Вдосконалення автоматизації програмного продукту «Clusterbox».

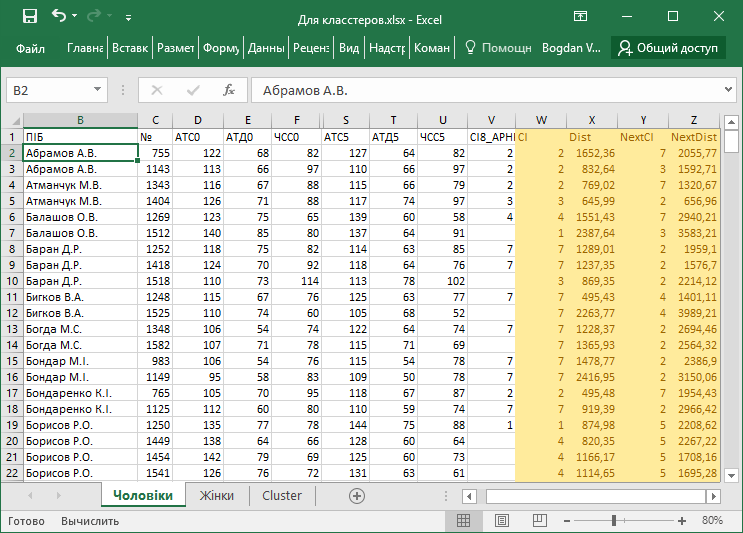
У ході проведення досліджень вихідну базу даних було приведено до вигляду, наведеному на рисунку:



Рисунок

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:



Вона містить чотири додаткові поля: Cl – характеризує кластер (група ризику), до якого відноситься студент, Dist – мінімальна відстань до цього кластеру, NextCl – 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, тому було прийнято рішення розбити пацієнтів на групи методом «один проти всіх» та вирівняти дані в групах, де були отримані показники класифікації занадто асиметричні. Результати дослідження представлені для третього кластеру, але аналогічна процедура проводилася для кожного кластеру окремо. Результуючу таблицю для всіх кластерів зображено у таблиці 1.

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

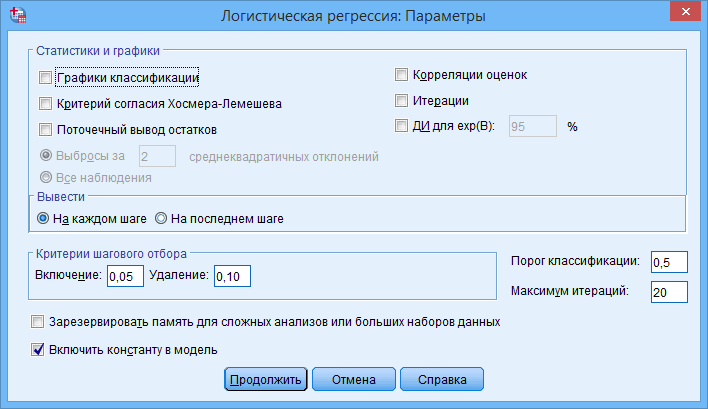


Рисунок 4 – Зображення параметрів

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

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

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

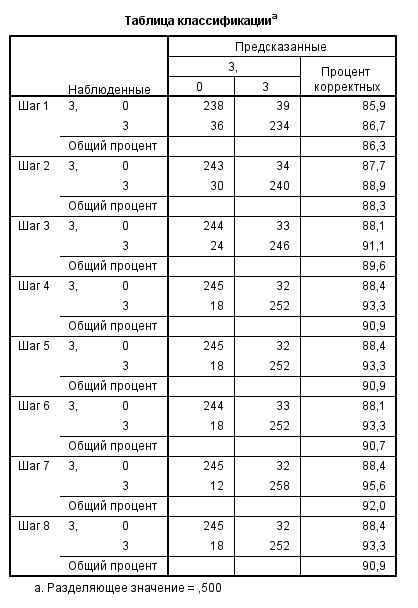


Рисунок 5 – Результати класифікації ЛР

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

Для побудови рівняння регресії було використано наступну таблицю:

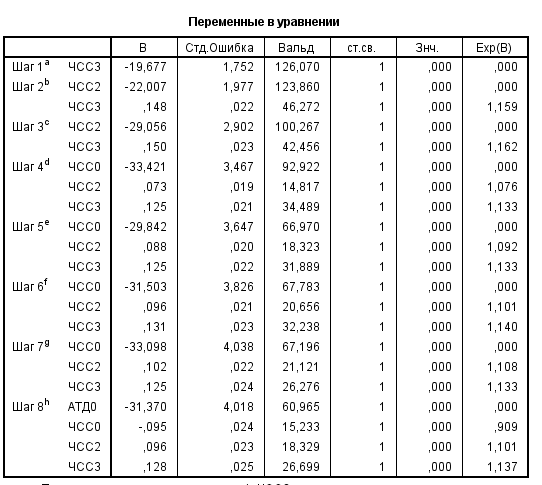


Рисунок 6 – Змінні для рівняння регресії

Таким чином, рівняння регресії набуває вигляду:



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

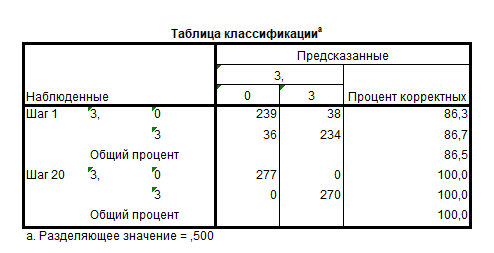


Рисунок 7 – Результати класифікації ЛР з нелінійними перетвореннями

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

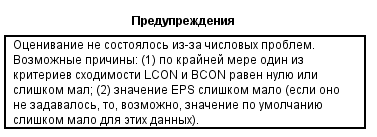


Рисунок 8 – Попередження

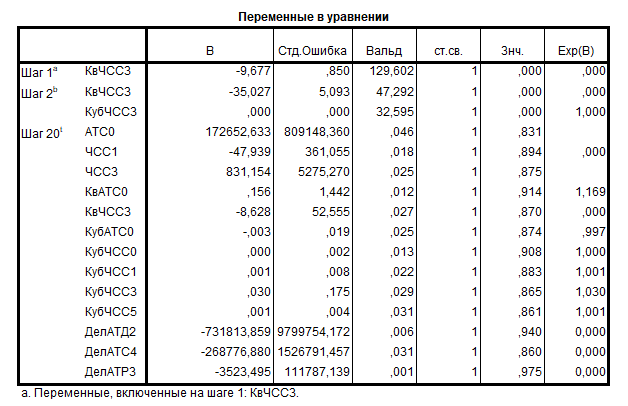


Рисунок 9 – Змінні для рівняння регресії з нелінійними перетвореннями

Таким чином, рівняння регресії має вигляд:



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

Таблиця 1 – Порівняльна характеристика

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

Розшифрування міток в таблиці 1:

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

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

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

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 показаний вибір відстані та критерій розпізнавання.

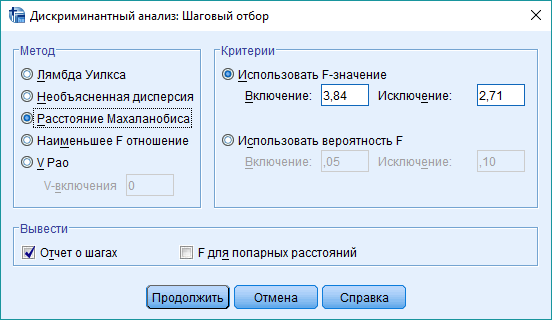


Рисунок 10 – Вибір методів

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

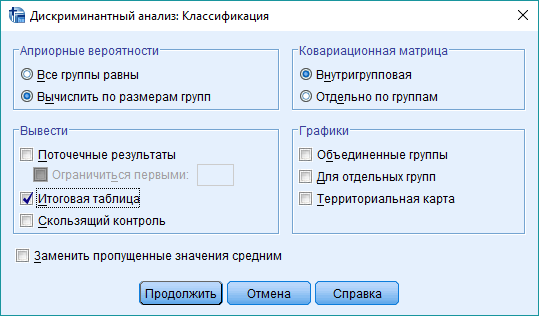


Рисунок 11 – Вибір класифікації

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

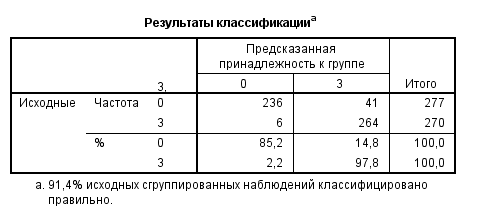


Рисунок 12 – Результати класифікації

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

Для побудови дискримінантної функції застосовуються дані з наступної таблиці:



Рисунок 13 – Коефіцієнти для дискримінантної функції

Таким чином, дискримінантна функція набуває вигляду:

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

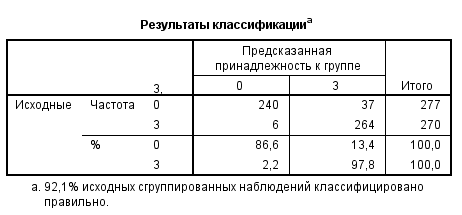


Рисунок 14 – Результати класифікації з нелінійними перетвореннями

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



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

Таким чином, дискримінантна функція має вигляд:



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

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

Таблиця 2 – Порівняльна характеристика

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

З таблиці 2 видно, що аналіз, проведений на базі нелінійних перетворень дає складніші рівняння моделі, але їх складність, беручи до уваги всі змінні, менша. Слід також зауважити, що у 7 тестах із 7 відсоток коректно спрогнозованих даних збільшився. За допомогою дискримінантного аналізу було побудовано функції прогнозування для кожного кластера, при чому результати дослідження показують, що введення нелінійних змінних (для кожної вхідної змінної) до аналізу покращую результат класифікації.

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

Дослідження результатів проведення логістичної регресії та дискримінантного аналізу показують, що в середньому відсоток класифікації становить близько в середньому відсоток класифікації становить 92, але це з урахуванням аналізу на істинних даних. Таким чином, високий відстоток 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 риснуку… ,працює коректно та ефективно.

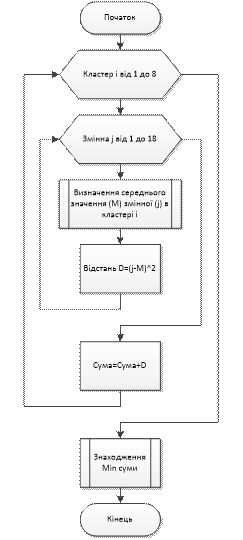


Рисунок 3.14 – Блок-схема алгоритму знаходження мінімальної відстані до кластеру [32].

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 1495спостережень). Таким чином за даними середніх значень (рисунок…) ми можемо побудувати графіки середніх значень АТД і АТС по кластерам

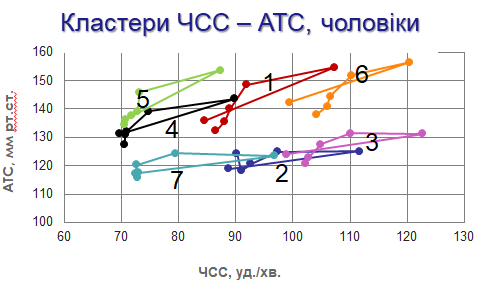


Рисунок 3.10 – Графіки ЧСС, АТС по кластерам

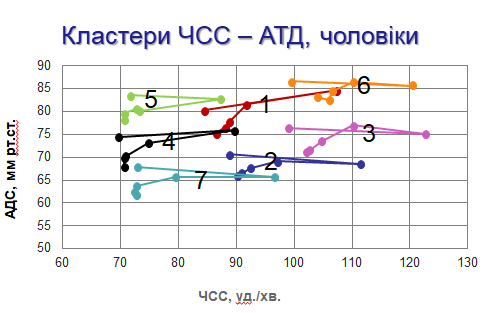


Рисунок 3.11 – Графік ЧСС, АТД по кластерам

З графіків чітко видно динаміку артеріального тиску та пульсу при виконанні проби Мартіне у кожному кластері. Ми бачимо, що деякі кластери можуть перетинатися, але за своїми властивостями вони різні.

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



Рисунок 3.12 – Характеристики кластерів

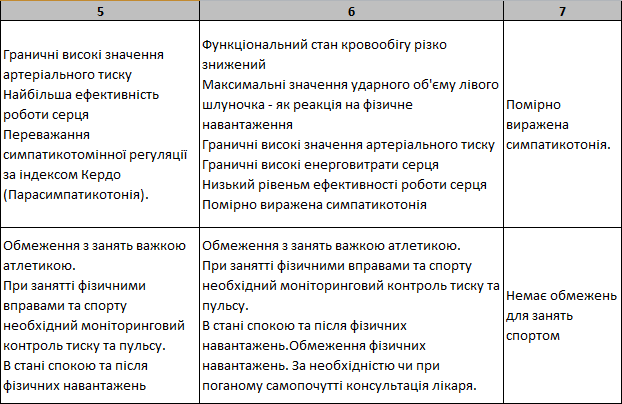


Рисунок 3.13 – Характеристики кластерів

# Як ми бачимо, за АТС кластери 2 і 3, 7 і 2 перетинаються, а за АТД перетинаються кластери під номерами 1 і 4, 2 і 7, 1 і 6. Таким чином висуваємо гіпотезу про те, що вони мають спільну підгрупу. Також ми можемо побачити, що за АТС кластери 4 і 5 знаходяться приблизно в одному діапазоні як за значеннями тиску, так і за значеннями пульсу.

# 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.

# РОЗДІЛ 4 ПРОГРАМНИЙ ПРОДУКТ ДЛЯ ВИЗНАЧЕННЯ ФУНКЦІОНАЛЬНОГО СТАНУ СИСТЕМИ КРОВООБІГУ

## 4.1 Проектування програмного продукту

Контекстна діаграма

На контекстній діаграмі зображено процес використання програмного забезпечення для знаходження мінімальної відстані до кластеру та виведення попереднього діагнозу, що включає в себе рекомендації про подальше заняття спортом.

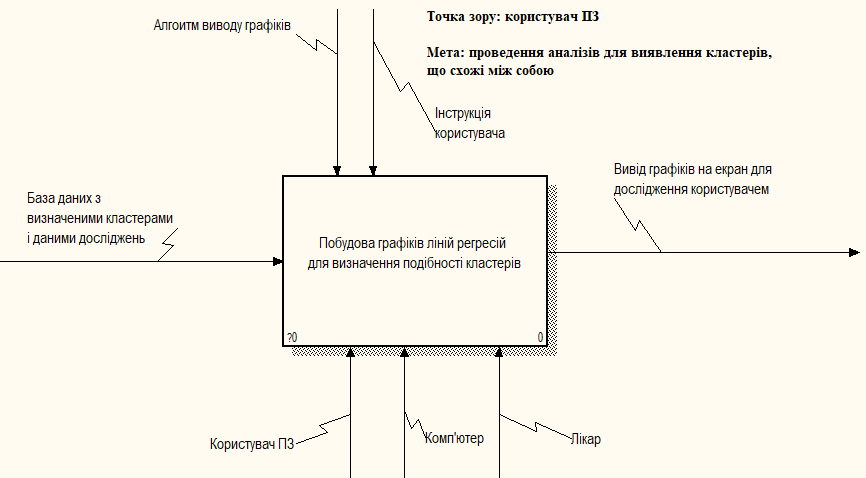


Рисунок 4.1 – Контекстна діаграма [33].

Вхідними даними є параметри систолічного артеріального тиску, диастолічного артеріального тиску, пульсу в стані спокою та ці ж параметри на 1-5 хвилинах після навантаження. Вихідні дані – це діагноз із протипоказаннями, що залежить від кластеру в який потрапив студент.

Діаграма декомпозиції першого рівня

Для поетапної демонстрації кожного етапу використання програмного продукту для знаходження мінімальної відстані до кластеру було створено діаграму декомпозиції. Така діаграма деталізує окремі етапи виконуваної роботи.

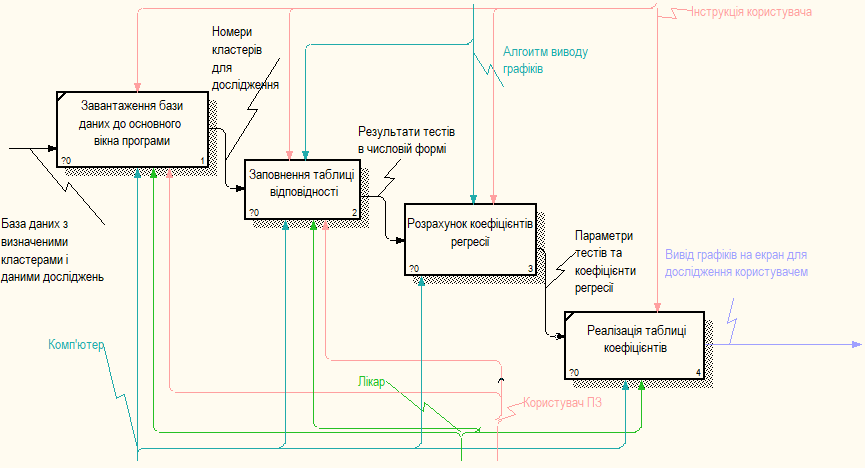


Рисунок 4.2 – Діаграма декомпозиції першого рівня [34].

Процес умовно розділений на 4 підпроцеси, серед яких: реєстрація введених даних в головній базі даних програми, обробка введених параметрів, реалізація таблиці результатів та встановлення попереднього діагнозу.

Діаграма декомпозиції другого рівня

Для наочного відображення пункту «Реалізація таблиці результатів» була розроблена діаграма декомпозиції другого рівня.

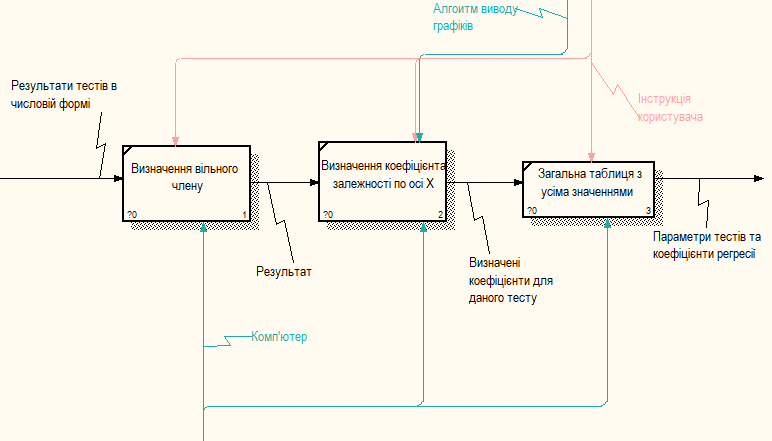


Рисунок 4.3 – Діаграма декомпозиції другого рівня [33, 34].

Цей процес складається з 3 підпроцесів та включає в себе: таблицю початкових, проміжних та кінцевих значень, таблицю параметрів авторизації та загальну таблицю з усіма значеннями. Вхідні дані це результати тестів в числовій формі, а на виході ми отримуємо таблицю результатів.

Діаграма дерева вузлів

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

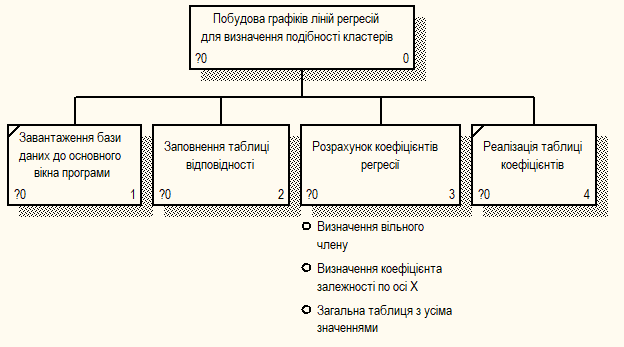


Рисунок 4.4 – Діаграма дерева вузлів [35].

Use Case діаграма

Діаграма Use Case описує процеси, котрі відбуваються при виконанні алгоритму, з точки зору акторів, що приймають участь у використанні програмного продукту. Учасниками виконання алгоритму є лікар-адміністратор та користувач. Варіант використання слугує для того, щоб описати сервіси, що система представляє акторам. Окремі варіанти використання на діаграмі позначаються еліпсами, а актор позначається фігуркою людини. Одні варіанти використання можуть бути частиною інших варіантів, розширювати їх або бути більш узагальненою версією.

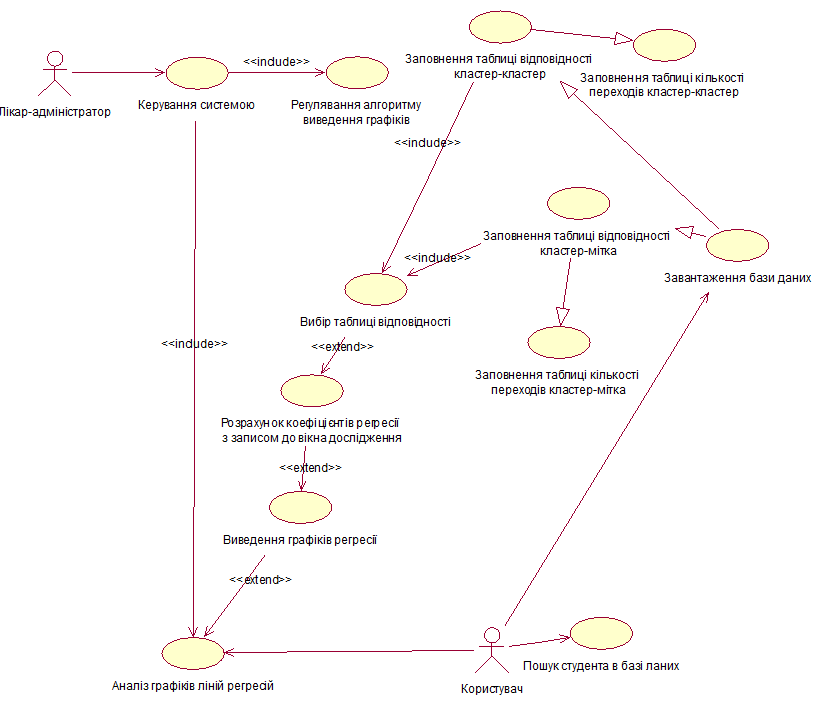


Рисунок 4.5 – Діаграма Use Case [36].

Схема демонстує нам, що лікар займається тільки адмініструванням системи і аналізом результатів програми, тоді як користувач сам заносить свої результати в програмний продукт.

Діаграма послідовності

Діаграма послідовності дає змогу спостерігати за виконанням актором визначеного варіанту використання, враховуючи їх послідовність в часі, а також показує взаємодію логічних елементів між собою.



Рисунок 4.6 – Діаграма послідовності [37]

Діаграма відображає послідовність в якій використовує систему користувач. Починається робота з введення ідентифікаційних даних та перевірки їх на правильність. Далі в відповідні комірки записуються параметри тиску та пульсу в стані спокою та після навантаження, перевірка коректності введених даних та реєстрація цих параметрів. Наступним кроком є застосування алгоритму знаходження мінімальної відстані та алгоритму знаходження субмінімальної відстані. Далі виходячи з введених даних користувача відносять до визначеного кластеру та генерується таблиця результатів з попереднім діагнозом. Останнім кроком є аналіз отриманих результатів користувачем [37].

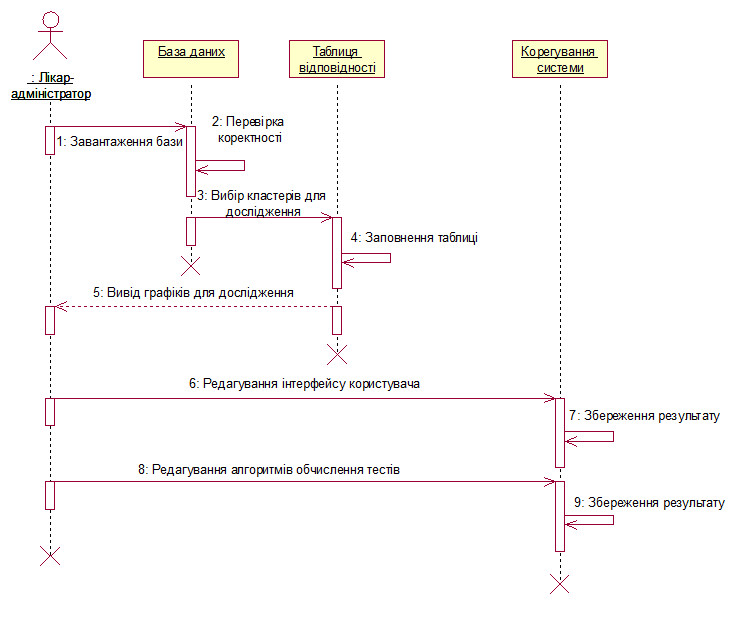


Рисунок 4.7 – Діаграма послідовності розрахунку мінімальної відстані до кластеру

Діаграма послідовності роботи алгоритму обчислення мінімальної відстані до кластеру регулюється лікарем та виглядає так: вводяться значення АТС, АТД та ЧСС, далі відбувається пошук різниці між середніми значеннями, результати підносяться до квадрату та додаються для кожного кластеру, мінімальне з цих значень стає відстанню до найближчого кластеру [37].

Діаграма кооперації

Діаграма кооперації показує взаємодію користувача з основними елементами програмного продукту.



Рисунок 4.8 – Діаграма кооперації

Діаграма кооперації для лікаря програмного продукту дає змогу відслідкувати його взаємодію з елементами інтерфейсу та побачити збереження введених ним даних та результатів до бази.



Рисунок 4.9 – Діаграма кооперації знаходження мінімальної відстані до кластеру

Дана діаграма кооперації відображає агоритм заходження мінімальної відстані та радіусу кластера. Цей процес відбувається послідовно та після знаходження радіусу та квадрату евклідової відстані генерує звіт та рекомендації для користувача, які лікар може регулювати в залежності від індивідуальних особливостей пацієнта [38].

Діаграма діяльності

Діаграма діяльності відображає послідовність дій, що виконується в процесі реалізації певного варіанта використання або функціонування системи в цілому. Діаграми діяльності є аналогом блок-схеми будь-якого алгоритму. Вони, як і діаграми станів та переходів, відображаються у вигляді орієнтованого графу, вершинами якого є дії, а ребрами – переходи між діями.

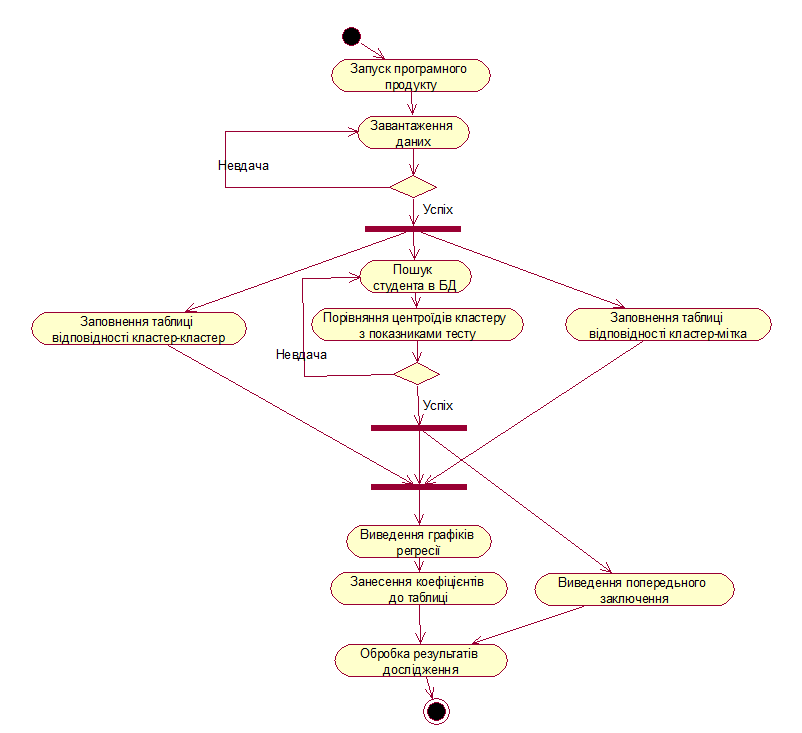


Рисунок 4.10 – Діаграма діяльності [39].

Діаграма станів

Діаграма станів демонструє всі зв'язки та послідовність роботи агоритму від початку і до кінця враховуючи успіх або невдачу, а також пропуски в проходжені визначених кроків роботи програмного продукту.

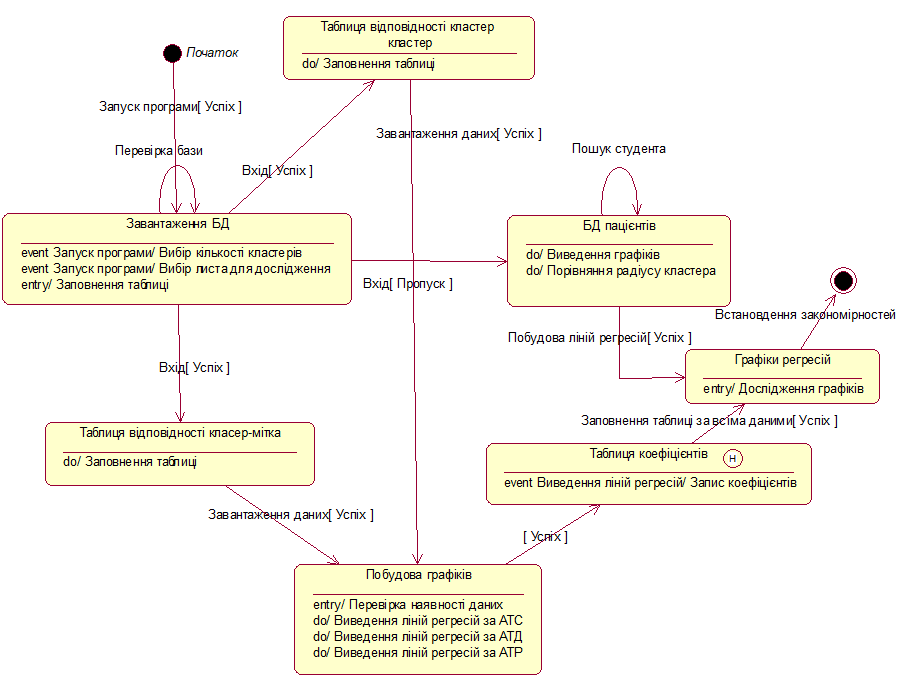


Рисунок 4.11 – Діаграма станів [40].

## 4.2 Розробка програмного продукту

Програмний продукт розроблено в середовищі Microsoft Visual Studio 2017 Community за допомогою мови програмування С# з використанням елементів WinForms та фреймворку .Net Framework. Для демонстрації результатів кластеризації був розроблений багатофункціональний інтерфейс користувача. Призначення даного доданку є дослідження подібності кластерів шляхом проведення лінійної регресії та виводу ліній трендів. У разі, якщо лінії трендів схожі між собою та їх коефіцієнти близькі за значення, можна буде зменшити кількість кластерів та побудувати нові результуючі таблиці.

Оскільки при кластеризації ми визначаємо субоптимальну групу (мітку), то даний додаток дає змогу дослідити переходи між кластерами, 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 завантаженій базі даних для того, щоб подивитись чи перевищує радіус студента радіус кластера. Також розраховується загальна кількість зміни кластерів в базі даних за одним студентом. Дані дослідження показують динаміку роботи алгоритму та можуть бути використані для моделювання системи виводу характеристик кластеру, до якого відноситься студент, та рекомендацій мітки, якщо радіус студента перевищує радіус кластера.

Вхідними даними для роботи програми є діастолічний артеріальний тиск, систолічний артеріальний тиск і частота серцевих скорочень в на першій, другій та третій хвилині 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).

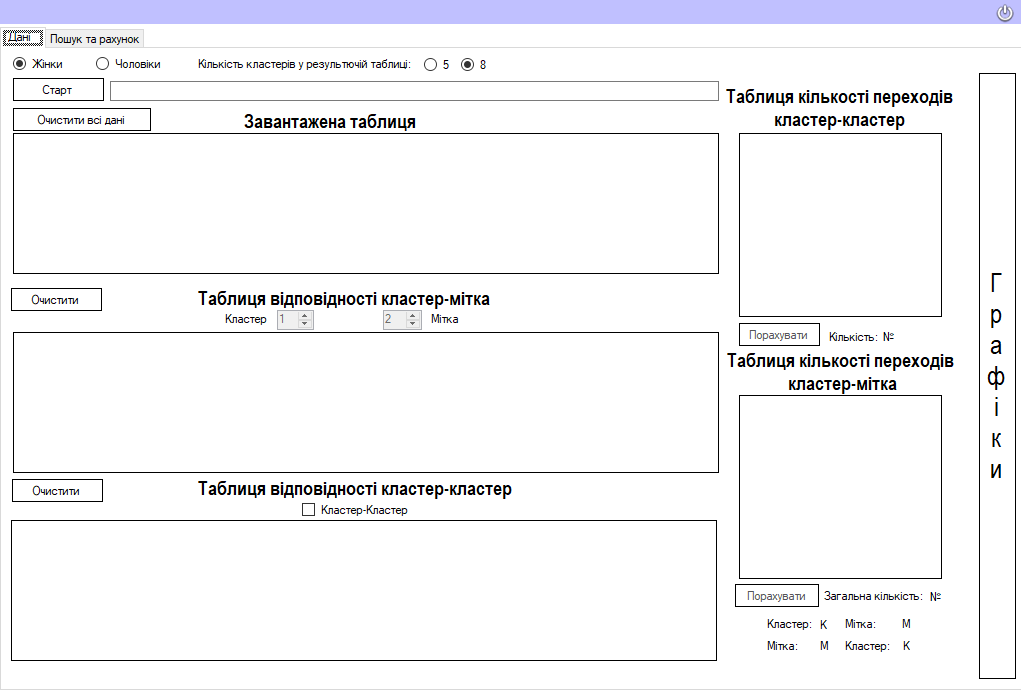
Вихідними даними є виведення графіків в залежності від методу дослідження: при порівнянні кластерів – графіки ліній трендів кластер-кластер, при дослідженні зміни кластеру – кластер-мітка (мітка-кластер).

Основні можливості системи, що повинні бути 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 направлення студента з подальшим порівнянням та виведенням відповідного графіку за необхідністю.

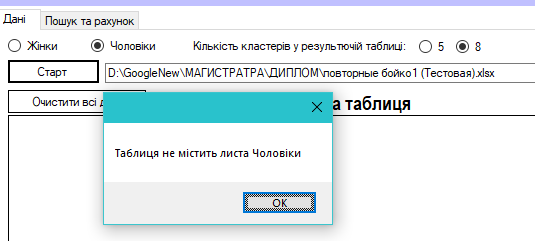
## 4.3 Робота з програмним додатком

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



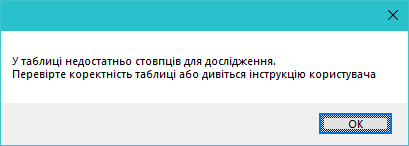
Для цього необхідно вибрати, для якої частини бази даних буде проводитися дослідження, та натиснути кнопку «Старт». Наступним кроком є вибір бази даних, що мітить прізвище студента, значення артеріального тиску та пульсу в стані спокою та на кожній хвилині після навантаження, включно до п’ятої хвилини, 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:

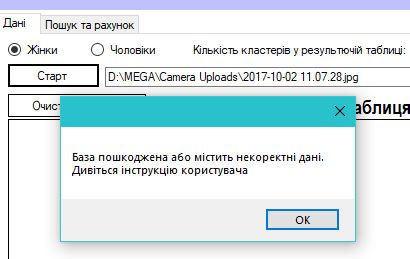


Для того, щоб завантажити іншу базу для дослідження, необхідно щоб поле з шляхом до бази було пустим. В іншому випадку за цим шляхом буде йти пошук бази даних.

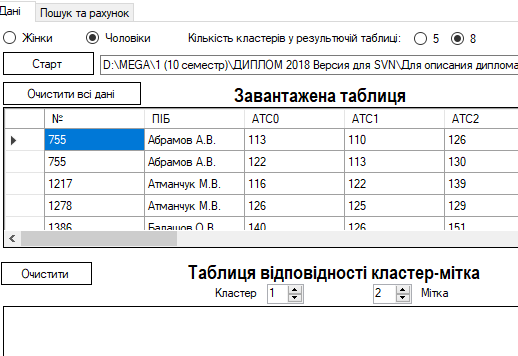
Якщо у таблиці є даний лист, але на ньому відсутні хоча б одне з основних полів, які мають бути завантажені, програмою буде оброблено виключення з виведенням інформації на екран користувача.



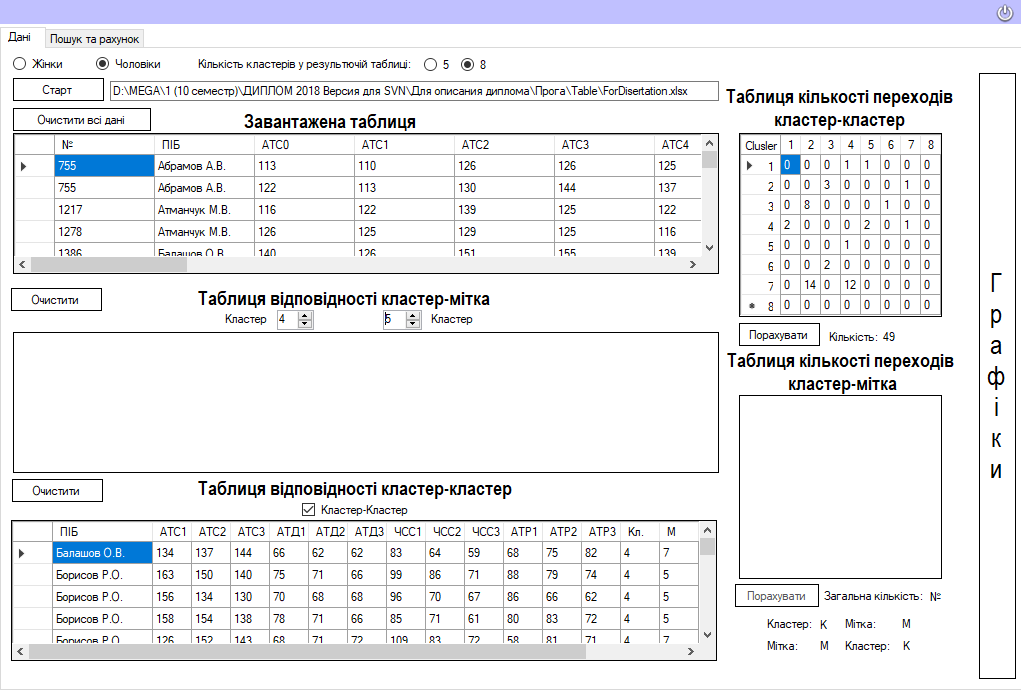
Слід також зазначити, що база даних для завантаження повинна бути у форматі таблиці Misrosoft Excel та бути у розширенні \*.xls або \*.xlsx. Якщо спробувати завантажити базу іншого формату або взагалі інший файл, програмою буде видано попередження:



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

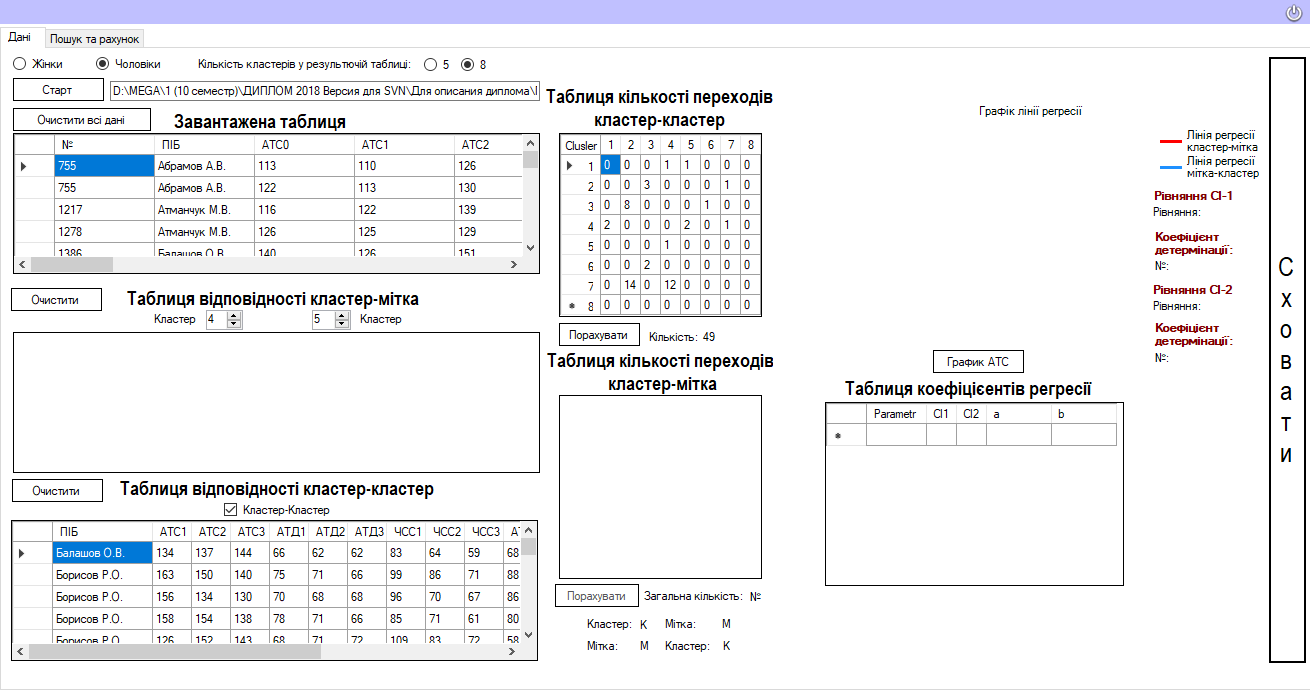


Оскільки ми хочемо дослідити подібність кластерів, необхідно встановити галочку навпроти напису «Кластер-Кластер» у таблиці відповідності кластер-кластер. Якщо галочка буде ввімкненою, тоді ми зможемо обрати два кластери, які будуть порівнюватися між собою. При цьому вікно таблиці «кластер-кластер» буде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

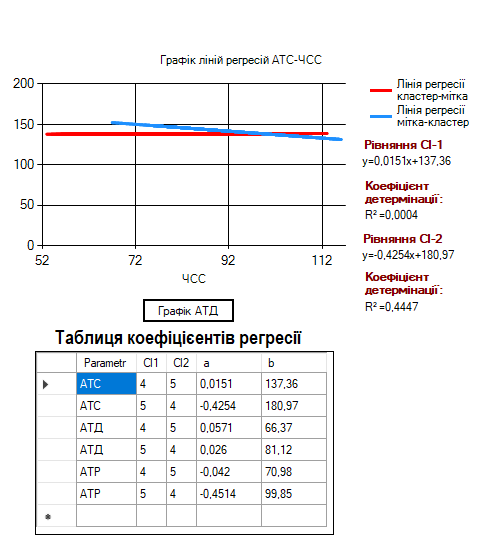


Також на рисунку … ми можемо побачити заповнену таблицю кількості переходів кластер-кластер. Вона показує скільки змін кластерів є у базі даних. Наприклад, студент Іванов проходив дослідження тричі. Перший раз він був у дургому кластері, а мітка показувала на сьомий. При повторному обстеженні – змінив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

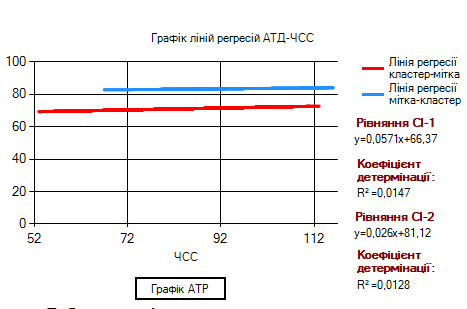
Оскільки таблиця відповідності кластер-кластер заповнена, можемо побудувати відповідні графіки. Для цього необхідно натиснути на клавішу «Графіки».



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».



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.



За проведеними дослідженнями за кластерами 4 і 5 ми бачимо, що за АТС лініє 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 3.10). 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 2.24.

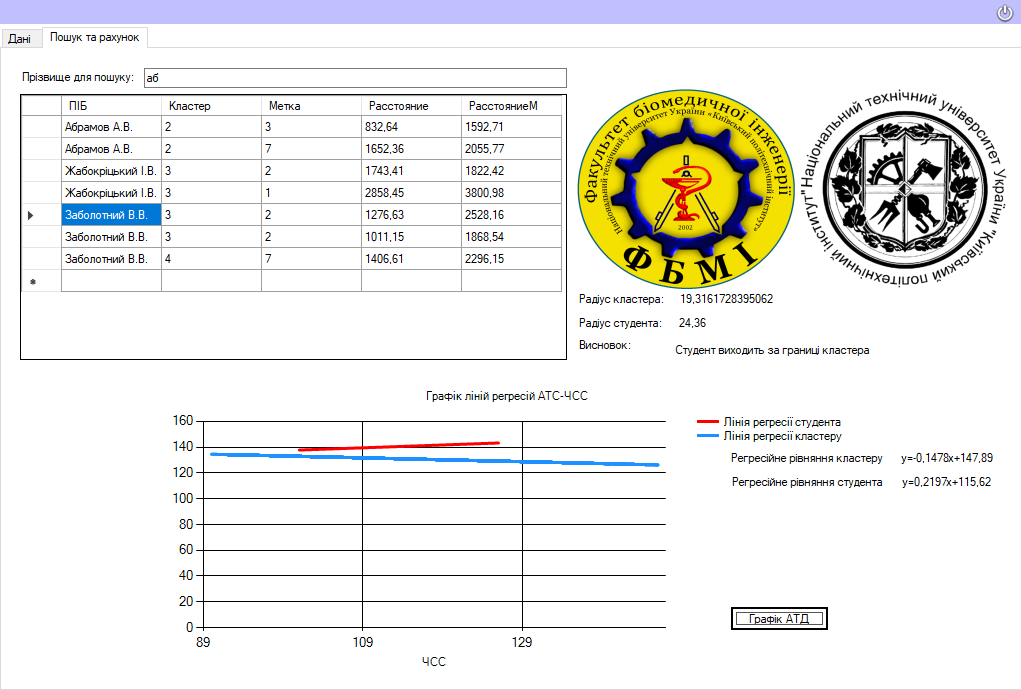


Рисунок 2.24 – Пошук студента в БД з виводом графіків

## Висновок до розділу 4

В даному розділі проведено проектування програмного продукту, що зображено на контекстній діаграмі, декомпозиції, послідовності, станів та кооперації. Описано реалізацію програмного продукту та наведено приклад його використання в системі.

Розділ 5. Аналіз зменшення кількості кластерів з вдосконаленням програмного продукту

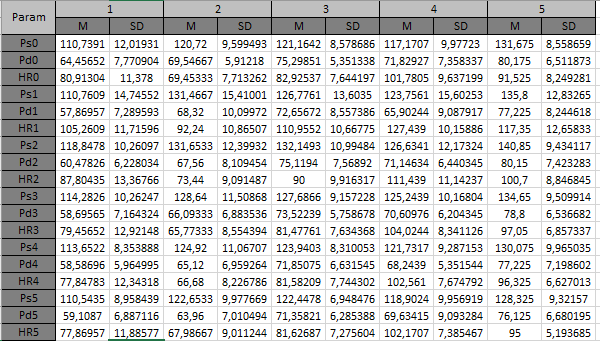
Наші дослідження показали, що зменшення кількості кластерів є необхідним кроком для поліпшення роботи алгоритму класифікації на основі квадрату евклідової відстані.

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 цього 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 отримано і для всіх інших значень тиску та пульсу.

  
Рисунок 2.12 – Описові статистики

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

  
Рисунок 2.13 - Результуюча таблиця

Результуюча таблиця виступає базою даних для побудови нових графіків.

Рисунок 2.14 - Графіки АТД, ЧСС для 5 кластеру

Рисунок 2.15 - Графіки АТС, ЧСС для 5 кластеру

Аналогічні аналізи при провели для чотирьох і трьох кластерів, отримали відповідні результуючі таблиці та побудували наступні графіки:

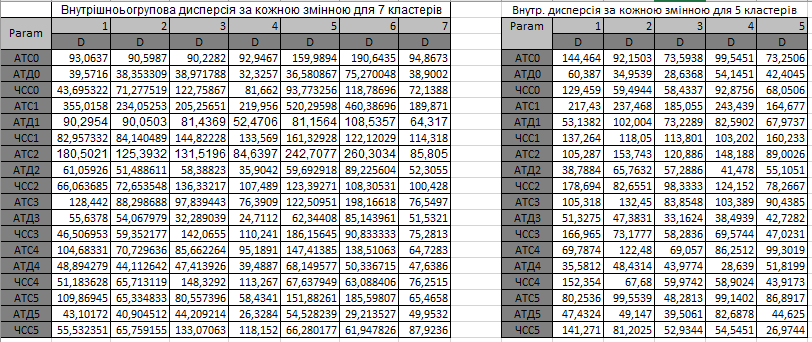
Рисунок 2.16 - Графіки АТД, ЧСС для 4 кластеру

Рисунок 2.17 - Графіки АТС, ЧСС для 4 кластеру

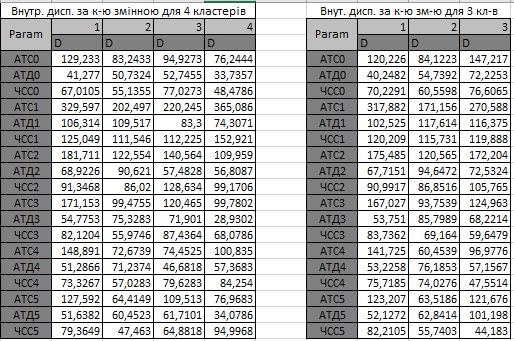
Рисунок 2.18 - Графіки АТД, ЧСС для 3 кластеру

Рисунок 2.19 - Графіки АТС, ЧСС для 3 кластеру

Дивлячись на графіки ми бачимо, що графіки на базі трьох кластерів є найменш інформативними, оскільки ми втрачаємо групу з даними, що показують високі показники систолічного тиску. Графіки на базі чотирьох кластерів показують, що група з низькими показниками систолічного тиску зникає, але в цілому інші графіки показують, що кластери досить різні. Тому виходячи з графіків ми можемо сказати, що найбільш інформативними є графіки на базі п’яти кластерів. Також ми можемо подивитися як змінюється дисперсія зі зменшенням кількості кластерів



Таблиця 2.3 – Зміна дисперсії



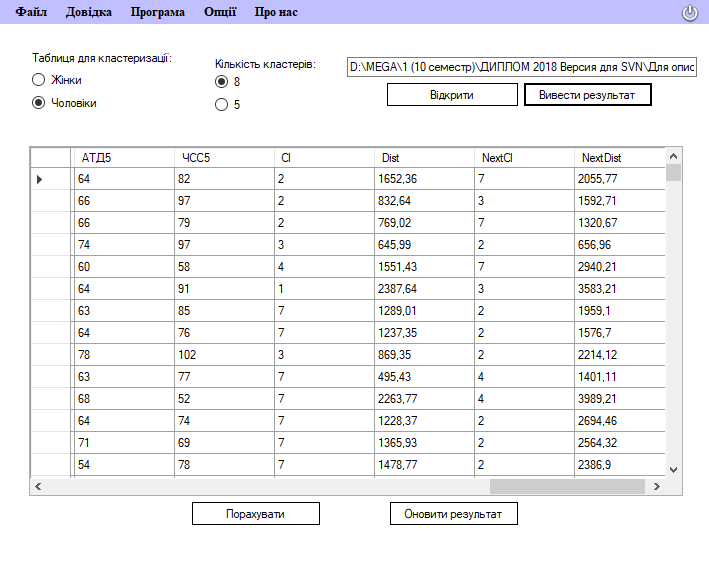
З порівняльної таблиці ми бачимо, що дисперсія досить сильно змінюється при зменшенні кластерів на 2 і 3 позиції, але при зменшенні на 4, дисперсія майже не змінилася, тому вибирати найоптимальнішу кількість виходячи з аналітичної частини треба серед таблиць на 5 і 4 кластери.

Виходячи з графіків та дисперсійного аналізу чітко видно, що таблиці на 5 кластерів дають найоптимальніший результат, тому нами було вирішено розширити функціонал програмного продукту Clusterbox і додати до нього можливість вибору результуючої таблиці на 5 кластерів.

Таким чином 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.

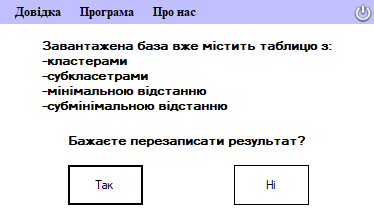


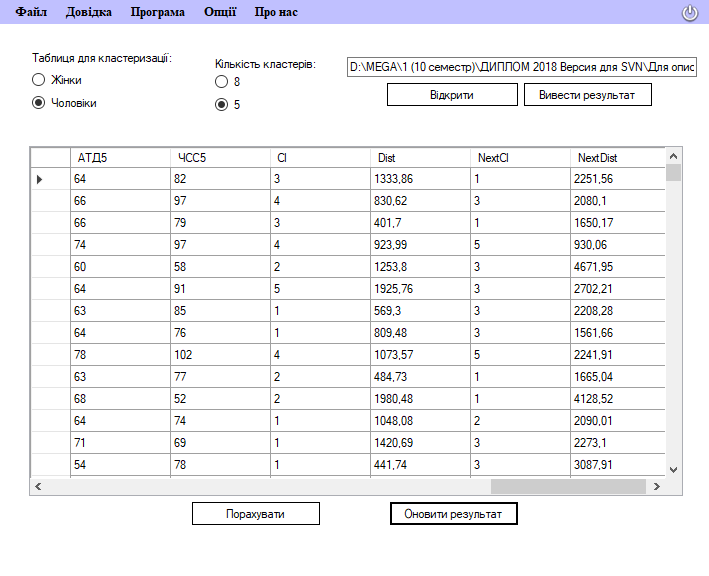
При натисненні на клавішу «Глобальний режим» відкривається вікно для проведення глобальної кластеризації.



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.

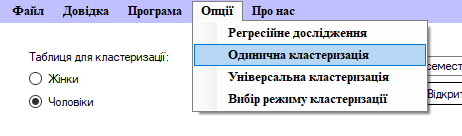
Слід зазначити, що якщо база даних містить значення кластеру, то при натисненні клавіші «Порахувати» випливе попередження: якщо натиснути так, тоді дані будуть перезаписані. Для прикладу проведемо кластеризацію з використанням результуючої таблиці на 5 кластерів та перезаписом результатів попередніх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





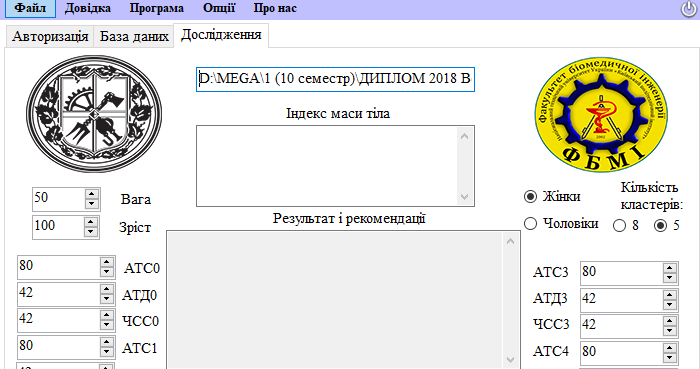
При оновленні результату ми бачимо, що дані були успішно перезаписані.

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



Також ми можемо перейти то 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.



Висновок до розділу 5

………………………………………………..

# ЗАГАЛЬНІ ВИСНОВКИ

У результаті роботи над дипломною роботою було опрацьовано багато наукових джерел інформації стосовно теми диплому, набуто професійних вмінь та навичок у роботі з сучасними інформаційними технологіями.

В ході виконання проекту було проаналізовано запропоновану базу даних студентів 1-2 курсів НТУУ «КПІ ім. Сікорського». Розщеплено базу даних по статі та виведені описові статистики по чоловікам.

Проведено дисперсійний аналіз, що дав змогу вирахувати середні показники значення, стандартне відхилення, степені свободи, значимість та суму квадратів.

Результати дисперсійного аналізу згруповано по чоловікам в результуючі таблиці, по яких побудовано графіки.

Проведено аналіз графіків та охарактеризовано кластери чоловіків.

Розроблено додаткові вікна для програмного продукту, а також реалізовано методи визначення мінімальної відстані та автоматичного створення бази даних Access.

Результатом дипломної роботи став програмний продукт, що здатний визначати мінімальну відстань до кластеру спираючись на дані дисперсійного аналізу. Програмним продуктом також передбачено збереження результатів дослідження до локальної бази даних.

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

Дипломну роботу виконано згідно з «Положенням про державну атестацію студентів НТУУ «КПІ»».

# ПЕРЕЛІК ПОСИЛАНЬ

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

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

1. Ланда Б.Х. Методика комплексной оценки физического развития и физической подготовленности: учеб. пособие / Б.Х. Ланда. – М.: Советский спорт, 2011. – 348 с.
2. Оценка функционального состояния и адаптивных возможностей организма у студентов вуза в процессе занятий физической культурой: методические указания для преподавателей физической культуры и студентов / сост. С.Е. Бебинов. – Омск: Изд-во СибАДИ, 2004. – 16 c.
3. Пономарёв В.В. Интегративный контроль физкультурного образования школьников на Крайнем Севере / В.В. Пономарёв В.Е. Лыков. – М.: Теория и практика физической культуры, 2007. – 130 с.
4. Холодов Ж.К. Теория и методика физической культуры и спорта / Ж.К. Холодов. – 10-е изд. – М., 2012. – 480 с.
5. Брехман И.И. Валеология – наука о здоровье. – М.: ФиС, 1990. – 208с.
6. Войтенко В.П. Здоровье здоровых. Введение в санологию. – К.: Здоров’я, 1991. – 248с.
7. Дембо А.Г. Врачебный контроль в спорте. – М.: Медицина, 1988.–288 с.
8. Казначеев В.П., Баевский Р.М., Берсенева А.П. Донозологическая диагностика в практике массовых обследований населения. – Л.: Медицина, 1980. – 210с.
9. Карпман В.Л., Белоцерковский Э.Б., Гудков И.А. Тестирование в спортивной медицине. – М.: ФиС, 1988. – 208с.
10. Оценка функциональных способностей организма человека с помощью функциональных проб [Електронний ресурс] – Режим доступу до ресурсу: http://intranet.tdmu.edu.ua
11. Кластерний аналіз [Електронний ресурс] – Режим доступу до ресурсу: http://www.nickart.spb.ru/analysis/cluster.php
12. Задачі кластерного аналізу [Електронний ресурс] – Режим доступу до ресурсу: https://sites.google.com/site/ne4itkalogika/necitka-klasterizacia/zadaci-klasternogo-analizu
13. Кластерний аналіз [Електронний ресурс] – Режим доступу до ресурсу: http://studopedia.su/7\_34114\_klasterniy-analiz.html
14. Індекс маси тіла [Електронний ресурс] – Режим доступу до ресурсу: http://meduk.net.ua/archives/13904
15. Розрахунок ідеальної ваги [Електронний ресурс] – Режим доступу до ресурсу: http://malibu-sport.com.ua/uk/kak-rasschitat-svoj-idealnyj-ves/
16. Індекс маси тіла [Електронний ресурс] – Режим доступу до ресурсу: http://a-yak.com/indeks-masi-tila-chi-vse-u-vas-v-poryadku-z-vagoyu/
17. Основні завдання та методи математичної статистики [Електронний ресурс] – Режим доступу до ресурсу: http://pidruchniki.com/14170120/statistika/predmet\_matematichnoyi\_statistiki
18. Завдання описової статистики [Електронний ресурс] – Режим доступу до ресурсу: http://biblio.royalwebhosting.net/zadachi-opisatelnoy-statistiki.html
19. Статистика і обробка даних [Електронний ресурс] – Режим доступу до ресурсу: http://psyfactor.org/lib/stat1.htm
20. Дисперсійний аналіз [Електронний ресурс] – Режим доступу до ресурсу: http://um.co.ua/4/4-7/4-70248.html
21. Дисперсійний аналіз [Електронний ресурс] – Режим доступу до ресурсу: http://ukrefs.com.ua/print:page,1,133443-Dispersionnyiy-analiz.html
22. Дисперсійний аналіз даних [Електронний ресурс] – Режим доступу до ресурсу: http://wiki.tntu.edu.ua
23. Основні заходи відстаней [Електронний ресурс] – Режим доступу до ресурсу: http://stud.com.ua/75021/statistika/osnovni\_zahodi\_vidstaney
24. Відстані між об’єктами і міра близькості [Електронний ресурс] – Режим доступу до ресурсу: http://bibliograph.com.ua/economicheskaya-statistika-2/11.htm
25. Кластерний аналіз. Загальне поняття, його математичні основи і завдання [Електронний ресурс] – Режим доступу до ресурсу: http://westudents.com.ua/glavy/88806--27-klasterniy-analz-zagalne-ponyattya-yogo-matematichn-osnovi-ta-zavdannya.html
26. Ієрархічна кластеризація [Електронний ресурс] – Режим доступу до ресурсу: https://studme.com.ua/1435012010869/marketing/ierarhicheskaya\_klasterizatsiya.htm
27. IBM SPSS Statistic Base [Електронний ресурс] – Режим доступу до ресурсу: http://www-03.ibm.com/software/products/ru/spss-stats-base
28. Программное обеспечение IBM SPSS [Електронний ресурс] – Режим доступу до ресурсу: http://www.predictivesolutions.ru/software/
29. Однофакторний дисперсійний аналіз [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/support/knowledgecenter/ru/SSLVMB\_22.0.0/com.ibm.spss.statistics.help/spss/base/idh\_onew.htm
30. Описательные статистики [Електронний ресурс] – Режим доступу до ресурсу: https://www.ibm.com/support/knowledgecenter/ru/SSLVMB\_22.0.0/com.ibm.spss.statistics.help/spss/base/idh\_desc.htm
31. Создание простой блок-схемы [Електронний ресурс] – Режим доступу до ресурсу: https://support.office.com/en-US/article/Basic-tasks-in-Visio-B58073C2-12C8-4981-AD7E-235066FA910D
32. IDEF0 методология. Нотация, принципы моделирования [Електронний ресурс] – Режим доступу до ресурсу: http://www.nazametku.com/dlia-raboty/idef0-metodologiya-notaciya-principy-model/
33. Моделирование деловых процессов [Електронний ресурсу] – Режим доступу до ресурсу: http://www.info-system.ru/designing/methodology/sadt/sadt\_for\_bp.html
34. Діаграма дерева вузлів [Електронний ресурсу] – Режим доступу до ресурсу: http://studopedia.com.ua/1\_162876\_diagrami-dereva-vuzliv-i-FEO.html
35. Диаграмма сценариев использования в процессе разработки ПО [Електронний ресурс] – Режим доступу до ресурсу: https://habrahabr.ru/company/luxoft/blog/312188/
36. UML-схемы последовательностей: справочные материалы [Електронний ресурс] – Режим доступу до ресурсу: https://msdn.microsoft.com/ru-ru/library/dd409377.aspx
37. Диаграмма кооперации [Електронний ресурс] – Режим доступу до ресурсу: http://itteach.ru/rational-rose/diagrammi-vzaimodeystviya
38. Диаграмма деятельности UML [Електронний ресурс] – Режим доступу до ресурсу: http://www.planerka.info/item/Diagrammy-deyatelnosti-UML
39. Автоматическая генерация тестов для графического пользовательского интерфейса по UML диаграммам действий [Електронний ресурс] – Режим доступу до ресурсу: http://citforum.ru/SE/testing/generation\_uml/
40. Інструкція з охорони праці №97. Для працівників, які використовують персональні комп'ютери. – Львів, 2010, 5 с.
41. ГОСТ 12.0.003-74. Система стандартов безопасности труда. Опасные и вредные производственные факторы. Классификация. М.: Изд-во стандартов, 1988. - 79 с.
42. Санітарні норми мікроклімату виробничих приміщень ДСН 3.3.6.042-99 [Електронний ресурс]. – Режим доступу до ресурсу: http://mozdocs.kiev.ua/view.php?id=1972
43. ДСанПіН 3.3.2.007-98 Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ЕОМ [Електронний ресурс]. – Режим доступу до ресурсу: http://document.ua/derzhavni-sanitarni-pravila-i-norm.. – Document.ua
44. ДСН 3.3.6.037-99. Санітарні норми виробничого шуму, ультразвуку та інфразвуку [Електронний ресурс]. – Режим доступу до ресурсу: http://mozdocs.kiev.ua/view.php?id=1789
45. СНиП 2.09.02—85. Строительные нормы и правила. Производственные здания. М: ГОССТРОЙ СССР, 1991. – 35 с.
46. Методичні вказівки до виконання розділу «Охорона праці» в дипломних роботах для студентів Факультету біомедичної інженерії / Уклад.: Демчук Г.В., Качинська Н.Ф., Чикунова Н.П. К.: НТУУ «КПІ», -15 с

# ДОДАТОК А

**Технічне завдання**

1. Найменування програмного продукту

Найменування ПП: «Програма для оцінки функціональних реакцій на тестове навантаження у студентів»

1. Призначення і область застосування

ПП призначений для:

− Визначення групи студента

− Виведення характеристик групи

− Виведення рекомендацій та застережень

Програмний продукт має власний, приємний GUI-інтерфейс і може використовуватися в:

− Медичних закладах для оцінки серцево-судинної системи

− Учбових закладах для оцінки функціональних реакцій організму

1. Вимоги до ПП
   1. Вимоги до функціональних характеристик

ПП повинен забезпечувати можливість виконання перерахованих нижче функцій:

− Завантаження результуючої таблиці до програми

− Розрахунок мінімальної відстані

− Розрахунок субмінімальної відстані

− Визначення найближчого кластеру

− Перевірка на попадання об’єкта в радіус кластеру

− Виведення характеристик та рекомендацій для кластеру

− Збереження результатів до локальної бази даних

* 1. Вимоги до забезпечення надійного функціонування ПП

Надійне (стійке) функціонування ПП має бути забезпечене виконанням Замовником сукупності організаційно-технічних заходів, перелік який наведено нижче:

− Організацією безперебійного живлення технічних засобів

− Використанням ліцензійного програмного забезпечення

− Відсутністю сторонніх або шкідливих програм, що можуть привезти до непрацездатності даної програми

* 1. Час відновлення після відмови

Оскільки в роботі використовується відлагоджений продукт, то час відновлення після відмови, викликаного збоєм електроживлення технічних засобів (іншими зовнішніми чинниками), не фатальним збоєм (не крахом) операційної системи, не повинно перевищувати 5-ми хвилин за умови дотримання умов експлуатації технічних і програмних засобів.

Аналогічно час відновлення після відмови, викликаного несправністю технічних засобів, фатальним збоєм ОС, не повинно перевищувати часу, необхідного на усунення несправностей технічних засобів і переустановлення програмних засобів.

* 1. Відмови через некоректні дії користувача системи

Відмови програмного продукту внаслідок некоректних дій користувача при взаємодії з ПП через GUI-інтерфейс недопустимі.

1. Умови експлуатації
   1. Кліматичні умови експлуатації

Кліматичні умови експлуатації, при яких повинні забезпечуватися задані характеристики, повинні задовольняти вимогам, що пред’являються до технічних засобів в частині умов їх експлуатації.

* 1. Вимоги до складу і параметрів технічних засобів

До складу технічних засобів повинен входити IBM-сумісний персональний комп’ютер (ПЕОМ), що включає в себе

− Процесор Intel ® Pentium G3240 ™, не менше

− ОЗУ об’ємом, 2 Гігабайти, не менше

− HDD, 32 Гігабайти, не менше

− Операційну систему Windows XP або більш пізню версію

* 1. Вимоги до програмних засобів, які використовуються ПП

Системні програмні засоби, що використовуються ПП, повинні бути представлені ліцензійною локалізованою версією операційної системи Windows XP (або пізнішої версії). Програмним забезпеченням: Microsoft Vusial Studio 2013 та Microsoft Access 2010.

* 1. Спеціальні вимоги

ПП повинен забезпечувати роботу користувача за допомогою

GUI-інтерфейсу. При необхідності встановити додаток AccessDatabaseEngine.

1. Стадії та етапи розробки
   1. Стадії розробки

* Аналіз вимог
* Проектування
* Реалізація
* Тестування і налагодження роботи ПП
* Розробка документації
* Впровадження та супровід

1. Порядок прийому та контролю
   1. Види випробувань

* Перевірка алгоритмів знаходження мінімальної та субмінімальної відстані за допомогою програми Microsoft Excel 2010.
* Перевірка роботи ПП на різних комп’ютерах
* Перевірка роботи ПП в різних версіях ОС Windows
  1. Загальні вимоги до приймання роботи

Робота виконана згідно вимог методичних вказівок.

# ДОДАТОК Б

**Лістинг методу розрахунку мінімальної відстані з подальшим визначення кластеру**

private void btnCulculate\_Click(object sender, EventArgs e)

{

if (tbPathToFile.Visible == true)

{

try

{

double[] DataEnter = new double[18];

double dist = 0.0;

Data();

DataEnter = Data();

btnStart\_Click();

double dataExcelFormat;

List<double> SummClasters = new List<double>();

for (int i = 1; i < dataGridViewExcel.Columns.Count; ++i)

{

double summ = 0;

for (int j = 1; j < dataGridViewExcel.Rows.Count; ++j)

{

double.TryParse(dataGridViewExcel[i, j].Value.ToString(), out dataExcelFormat);

summ += Math.Pow((DataEnter[j - 1] - dataExcelFormat), 2);

}

SummClasters.Add(summ);

}

List<double> SummClastersCopy = new List<double>();

for (int i = 0; i < SummClasters.Count; i++)

{

SummClastersCopy.Add(SummClasters[i]);

}

SummClastersCopy.Sort();

for (int i = 0; i < DataEnter.Length;i++ )

{

dist += DataEnter[i];

}

double middle = (dist / 18);

double min = SummClastersCopy[0];

double nextMin = SummClastersCopy[1];

string read = "";

double middlePoint=0.0;

for (int i = 0; i < DataEnter.Length;i++ )

{

middlePoint += Math.Abs(DataEnter[i] - middle);

}

tbResult.Clear();

tbRecomend.Clear();

tbIndexMass.Clear();

for (int i = 0; i < SummClasters.Count; i++)

{

if (min == SummClasters[i])

{

switch (i)

{

case 0: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " +Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 1: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 2: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 3: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 4: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 5: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 6: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

case 7: tbResult.Text += "Студент відноситься до кластеру \*" + (i + 1) + "\*" + Environment.NewLine +

"Мінімальна відстань до кластеру становить: " + Math.Round(SummClasters[i], 2) + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); cluster = i + 1; break;

}

}

}

if ((Math.Round((middlePoint/18),2)> radiusCl[cluster - 1]) || (nextMin < ((min \* 0.16) + min))) {

for (int i = 0; i < SummClasters.Count; i++)

{

if (nextMin == SummClasters[i])

{

switch (i){

case 0: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 1: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 2: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 3: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 4: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 5: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 6: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

case 7: tbRecomend.Text += "Наступна відстань після мінімальної становить \*" + Math.Round(SummClasters[i], 2) + "\*"

+ Environment.NewLine + "Це кластер \*" + (i + 1) + "\*" + Environment.NewLine + Environment.NewLine +

"Кластер має наступні характристики: " + ReadFile(read, i + 1); break;

}

}

}

}

BodyMassIndex();

}

catch (FormatException)

{

MessageBox.Show("Перевірте корректність введених даних\nДив.інструкцію користувача\"Програма->Інструкція користувача\"");

}

catch (InvalidOperationException)

{

MessageBox.Show("Встановіть додаток \"AccessDatabaseEngine\"\n з папки \"Додаток\"");

}

catch (OleDbException)

{

if (File.Exists(tbPathToFile.Text)){

MessageBox.Show("Невірні поля у завантаженому файлі\nДив.інструкцію користувача\"Програма->Інструкція користувача\"");

}

else if (!File.Exists(tbPathToFile.Text)){

tbPathToFile.Visible = true;

MessageBox.Show("Файл не знайдено\nВиберіть інший файл!");

}

}

}

else {

MessageBox.Show("Виберіть файл\nНатисніть Файл->Відкрити");

}

}