ВСТУП

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

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

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

У відповідності з метою ставлять такі завдання:

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

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

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

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

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

Практичне значення одержаних результатів. Програмний продукт може бути використаний в медичних клініках та учбових закладах для оцінки функціональних реакцій організму. Насамперед використання програми передбачено на факультеті фізичного виховання Національного технічного університету України «Київського політехнічного університету ім. Ігоря Сікорського».

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

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

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

РОЗДІЛ 1
ОГЛЯД ЛІТЕРАТУРИ ЗА ТЕМОЮ І ВИБІР НАПРЯМІВ ДОСЛІДЖЕНЬ

1.1. Алгоритм квадрату евклідової відстані

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

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

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

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

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

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

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

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

для осіб більше 50 років без виражених відхилень – 10 присідань на 15 секунд.

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

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

Після проведення проби Мартіне для нашого дослідження буде отримано 18 значень, що характеризують динаміку зміни артеріального тиску. За даними значеннями можна побудувати графіки ЧСС-АТС та ЧСС-АТД і побачити як відновлюється організм після фізичних вправ. Якщо значення до навантаження не близькі до значень після навантаження, тоді це свідчить про погану відновлюваність організму та потребує внесення певних змін до способу життя пацієнта.

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

Оскільки у наявності програмного продукту є 7 груп чоловічої 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.

1.2. Дисперсійний аналіз

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

Мета дисперсійного аналізу - дослідження наявності або відсутності істотного впливу будь-якого якісного або кількісного фактору на зміни досліджуваної результативної ознаки. Для цього фактор, який імовірно має або не має істотного впливу, поділяють на класи (інакше кажучи, групи) і з'ясовують, чи однаковий вплив фактору шляхом дослідження значущості між середніми в наборах даних, у відповідних градаціях фактору.

При дисперсійному аналізі визначають питому вагу сумарного впливу одного або декількох факторів. Істотність впливу фактору визначається шляхом перевірки гіпотез:

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

Якщо вплив фактору не суттєвий, то несуттєва і різниця між класами градації цього фактору і в ході дисперсійного аналізу нульова гіпотеза Н0 не відкидається. Якщо вплив фактору істотний, то нульова гіпотеза Н0 відхиляється: не всі класи градації мають одне і теж середнє значення, тобто серед можливих різниць між класами градації одна або кілька є суттєвими.

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

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

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

1.2.1.Однофакторний дисперсійний аналіз

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

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

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

$$F = \frac{MS_{BG^4}}{MS_{WG}}, \text{ де} \qquad (1.1)$$

$$MS_{BG} = \frac{SSBG}{\nu BG}, MS_{WG} = \frac{SSWG}{\nu WG}$$

Виходячи з цього значення F має бути близько до 1 в разі, якщо статистично значущих відмінностей все-таки немає. Критичне значення F визначається рівнем значущості (зазвичай 0,05 або 0,01) і

внутрішньогруповим і міжгруповим числом ступенів свободи (ν). Воно досить складне для обчислення, тому частіше використовуються табличні значення із зазначенням α, νBG, νWG.

Міжгруповое число ступенів свободи: νBG = m - 1. m – число груп.