ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ
НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
Кафедра інженерії програмного забезпечення

Пояснювальна записка

до магістерської роботи
на ступінь вищої освіти магістр
на тему: **«ЕКСПЕРТНА СИСТЕМА ДЛЯ МЕДИЧНОГО СКРИНІНГУ НА ОСНОВІ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ»**

Виконала: студентка 6 курсу, групи ПДМ–61
спеціальності 121 Інженерія програмного забезпечення
(шифр і назва спеціальності)
Куцук В.А.
(прізвище та ініціали)
Керівник Шевченко С.М.
(прізвище та ініціали)
Рецензент_____
(прізвище та ініціали)

Київ – 2022

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра Інженерії програмного забезпечення

Ступінь вищої освіти - «Магістр»

Спеціальність - 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного
забезпечення
_____О.В. Негоденко

“ ____”____2022 року

З А В Д А Н Н Я

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Куцук Валерії Андріївні

(прізвище, ім‘я, по батькові)

1. Тема роботи: «Експертна система для медичного скринінгу на основі методів кластерного аналізу»

Керівник роботи к. пед. н., доцент Шевченко С.М.,

(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)

затвердженні наказом вищого навчального закладу від 12.10.2022 №122.

2. Строк подання студентом роботи 31.12.2022.
3. Вихідні дані до роботи:
   3.1. Науково-технічна, науково-медична література та експериментальні дані у медицині.

3.2. Експертні системи у медичній галузі.

3.3. Кластерний аналіз.

3.4. Про захист персональних даних. Закон України від 01.06.2010 № 2297-VI

4. Зміст розрахунково – пояснювальної записки (перелік запитань, які потрібно розробити):

4.1. Порівняльний аналіз методів машинного навчання у експертних системах для медичного скринінгу.

4.2. Модель експертної системи на основі методу ближнього сусіда і k-means та її застосування.

4.3. Дослідження шляхів забезпечення захисту персональних даних у медичних експертних системах.

5. Перелік графічного матеріалу:

5.1. Класифікація експертних систем у медичній галузі.

5.2. Структура експертної системи для медичного скринінгу.

5.3. Модель експертної системи для медичного скринінгу на основі методу ближнього сусіда.

5.4. Модель експертної системи для медичного скринінгу на основі методу k-means

5.5. Шляхи забезпечення захисту медичних персональних даних у експертній системі

5.6.

6. Дата видачі завдання 14.10.2022.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів магістерської роботи	Строк виконання етапів роботи	Примітка

1	Підбір науково-технічної та науково-медичної літератури	14.10.2022	Виконано
2	Дослідження структури експертних систем	21.09.2022	Виконано
3	Аналіз методів машинного навчання у експертних системах	28.09.2022	Виконано
4	Розробка моделі експертної системи на основі кластерного аналізу	02.10.2022	Виконано
5	Дослідження шляхів забезпечення захисту медичних персональних даних	17.10.2022	Виконано
6	Оформлення роботи, розробка демонстраційних матеріалів	27.12.2022	Виконано
7	Здача роботи	31.12.2022	Виконано

Студентка  Куцук В.А.

Керівник роботи _____ Шевченко С.М.

РЕФЕРАТ

Текстова частина магістерської роботи 69 сторінок, 29 рисунків, 56 джерел, 4 таблиці.

ШТУЧНИЙ ІНТЕЛЕКТ; МЕТОДИ МАШИННОГО НАВЧАННЯ; ЕКСПЕРТНА СИСТЕМА У МЕДИЧНІЙ ГАЛУЗІ; КЛАСТЕРНИЙ АНАЛІЗ; МЕТОД БЛИЖНЬОГО СУСІДА; МЕТОД K-MEANS; МЕДИЧНИЙ СКРИНІНГ.

Об'єкт дослідження – процес функціювання експертної системи у медичній галузі.

Предмет дослідження – методи кластерного аналізу.

Мета роботи – поліпшення ранньої діагностики можливих захворювань людини через впровадження медичного скринінгу на основі методів кластерного аналізу.

Методи дослідження – системно-структурний, порівняльний, методи кластерного аналізу.

Дане дослідження присвячене проблемі моделювання експертних систем для медичного скринінгу на основі методів кластерного аналізу. З метою визначення структури експертних систем у роботі проаналізовані існуюча класифікація та методи машинного навчання, на основі яких здійснюються різноманітні клінічні процеси. Здійснено обґрунтування створення експериментальної системи на основі методів ближнього сусіда та k-means, виконано їх порівняльний аналіз. Представлено експериментальне впровадження застосування даних методів для медичного скринінгу раку шийки матки. Наведені шляхи забезпечення захисту медичних даних у експертних системах.

Результати дослідження можуть бути використані для розробки програмного забезпечення експертної системи для медичного скринінгу та впроваджені у навчальний процес медичних закладів.

ЗМІСТ

ВСТУП

Штучний інтелект стрімко входить у всі сфери та галузі нашого буття. Обробка великої бази даних, оптимізація процесів, покращення і швидкодія послуг, навчання та прийняття рішення – ці та багато інших функцій та можливостей належить машинному навчанню. Такий стан розвитку інформаційних технологій дозволяє застосувати алгоритми та методи, які притаманні людському розуму, у медичній сфері, зокрема для медичного скринінгу, з метою виявлення факторів ризику, генетичних схильностей і ранніх проявів захворювання. Актуальність даного дослідження також підтверджується зростанням кількості різних захворювань, що вимагає постійного оновлення лікарської інформації та її опанування. Тому ефективним рішенням для реалізації задач автоматизації в медицині на допомогу медичним працівникам є створення та впровадження експертних систем – інтелектуальних комп'ютерних програм, які можуть консультувати, проводити аналіз, ставити діагноз на рівні фахівців-експертів у деякій вузькій предметній області. Цим і підтверджується актуальність даного дослідження.

Мета роботи: поліпшення ранньої діагностики можливих захворювань людини через впровадження медичного скринінгу на основі методів кластерного аналізу.

Для досягнення цієї мети в роботі необхідно вирішити такі **завдання:**

1. Проаналізувати наукові праці з досліджуваної проблеми і обґрунтувати застосування машинного навчання для медичного скринінгу.
2. Дослідити класифікацію експертних систем у медичній галузі.
3. З’ясувати структуру експертної системи для медичного скринінгу.
4. Здійснити аналіз методів кластерного аналізу: метод ближнього сусіда та метод k-means.
5. Розробити та теоретично обґрунтувати моделі експертної системи для медичного скринінгу на основі методу ближнього сусіда та методу k-means.

6. Дослідити шляхи захисту персональної інформації у медичних експертних системах.

Виходячи з цього, **об'єктом** дослідження є процес функціювання експертної системи у медичній галузі, а **предметом** дослідження – методи кластерного аналізу.

Методи дослідження: системно-структурний, порівняльний, методи кластерного аналізу.

Наукова новизна одержаних результатів: розроблено алгоритми функціювання експертної системи для медичного скринінгу на основі методу ближнього сусіда та методу k-means з врахуванням блоку забезпечення захисту персональних даних.

Практичний результат: основні положення та результати магістерської роботи можуть бути використані для розробки програмного забезпечення експертної системи для медичного скринінгу та впроваджені у навчальний процес медичних закладів.

Апробація результатів дослідження:

1. Куцук В.А. Модель експертної системи для медичного скринінгу на основі методів кластерного аналізу / Шевченко С.М., Жданова Ю.Д., Негоденко О.В., Куцук В.А. // Moderní aspekty vědy: XXVII. Díl mezinárodní kolektivní monografie / Mezinárodní Ekonomický Institut s.r.o.. Česká republika: Mezinárodní Ekonomický Institut s.r.o., 2023. – С. 478 – 494 [1].
2. Куцук В.А. Експертна система для медичного скринінгу на основі методів кластерного аналізу // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» – Київ: ДУТ, 2022 [2].

1 ШТУЧНИЙ ІНТЕЛЕКТ У МЕДИЦИНІ

Штучний інтелект науковцями описується як наука та інженерія створення розумних машин. Його впровадження у медичну галузь має широкі діапазони: починаючи від організації і управління лікарською установою та закінчуючи медичною робототехнікою у хірургії. Згідно дослідження міжнародного аналітичного агентства Global Market Insights, по 2024 р. очікується щорічне зростання використання штучного інтелекту у сфері охорони здоров'я аж 40%. За прогнозами консалтингової компанії Precedence Research, розмір світового ринку штучного інтелекту в ультразвуковій візуалізації оцінювався в 863,59 мільйонів доларів США в 2022 році та, за прогнозами, сягне приблизно 1691,2 мільйонів доларів США до 2030 року, зростаючи на 8,76% протягом прогнозованого періоду з 2022 по 2030 рік [3].

У медичній сфері системи, які використовують штучний інтелект, називають експертними. Розглянемо їх суть, структуру, кваліфікацію та деякі методи машинного навчання, які застосовуються в таких системах.

1.1. Аналіз експертних систем для медичного скринінгу

Експертна система — це комп’ютерна система, яка імітує здатність людини-експерта приймати рішення. Тай в усіх відношеннях діє як експерт-людина. Вона використовує експертні (людські) знання для вирішення проблем, які вимагають людського інтелекту. У сфері штучного інтелекту існує багато додатків, які намагаються допомогти експертам-людям, пропонуючи рішення проблеми. Експертні системи знаходять застосування в різних областях медицини. Медичні експертні системи спочатку були розроблені для академічних сфер, а пізніше також для клінічних застосувань. Системи охорони здоров’я виробляють величезну кількість інформації (пацієнти, демографічні, клінічні та платіжні дані), які піддаються аналізу за допомогою інтелектуального програмного забезпечення та потребують нових методів для отримання нових знань. Доступні різноманітні

інструменти медичних експертних систем, які можуть функціонувати як інтелектуальні помічники лікарів, допомагаючи в діагностичних процесах, лабораторному аналізі, протоколі лікування та навчанні студентів-медиків і резидентів. Експертні системи (ES) були запроваджені дослідниками Стендфордського проекту евристичного програмування, включаючи «батька експертних систем» Едварса Фейгенбаума, з системами DENDRAL і MYCIN. Основними учасниками технології були Брюс Бьюкенен, Едвард Шортлайф, Рендалл Девіс, Вільям ВенМелле, Кралі Скотт та інші зі Стенфорда.

Сутність експертної системи (від лат. *Expertus* — досвідчений) важко пояснити одним визначенням, оскільки точного єдиного у наукових колах не існує.

Так, розробник медичної експертної системи MYCIN Е. Фейгенбаум вважає, що експертна система – це інтелектуальна комп'ютерна програма, у якій використовуються знання і процедури логічного виводу для вирішення завдань, досить важких для того, щоб вимагати для свого вирішення значного обсягу експертних знань людини [4].

У галузі інженерії знань [5] визначають експертну систему трьома характеристиками:

1. Інтелектуальна система, орієнтована на тиражування досвіду висококваліфікованих спеціалістів в областях, де якість прийняття рішень традиційно залежить від рівня експертизи.
2. Система обробки даних і знань, яка забезпечує експертне рішення проблем в заданій області.
3. Інтелектуальна система, призначена для надання консультативної допомоги спеціалістам, які працюють в деякій предметній області.

Фахівці програмування окреслюють експертну систему як комплекс комп'ютерного програмного забезпечення, що допомагає людині приймати обґрунтовані рішення; використовує інформацію, отриману заздалегідь від експертів − людей, які в якій-небудь області є найкращими фахівцями; зберігає

знання про певну предметну область; має комплекс логічних засобів для виведення нових знань, виявлення закономірностей, виявлення протиріч і ін. [6].

На сучасному етапі існують аргументи для того, щоб називати будь-яку систему підтримки прийняття рішень експертною системою, якщо вона призначена для надання проблем експертного рівня специфічні поради, навіть якщо методи програмування та аналізу, що лежать в основі, відрізняються від методів, заснованих на знаннях, розроблених дослідниками штучного інтелекту [7].

Дана трактовка підтверджується у роботі [5]. На думку авторів розрізняють два типи експертних систем. Системи першого типу призначені для спеціалістів, чий професійний рівень не дуже високий. В базах знань таких систем зберігаються знання, отримані від висококваліфікованих спеціалістів. Системи другого типу покликані допомогти спеціалістам високої кваліфікації, виконуючи для них значну частину рутинних операцій і перегляд великих масивів інформації. Особливістю експертних систем є наявність в них системи пояснень, яка підвищує їх консультативну силу.

Дані підходи щодо поняття «експертна система» не суперечать один одному, бо розглядаються в тісному зв'язку і доповнюють суттєві якості цього поняття. Фахівці у роботі [8] пропонують вкладати у процесі визначення експертної системи саме її предметну область, яка і дозволяє здійснити наступну класифікацію (рис. 1.1) :

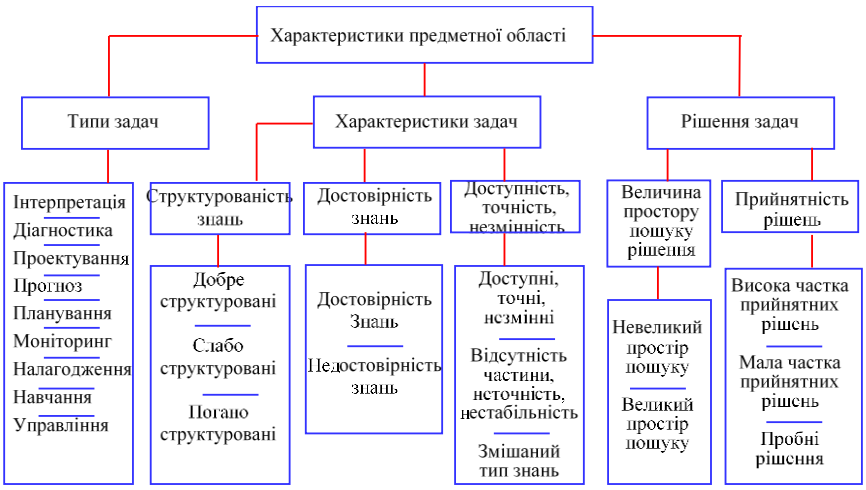


Рис.1.1 Класифікація експертних систем на основі предметної області

У дослідженні [9] класифікація експертної системи здійснена відповідно до принципу роботи: експертна система на основі правил; експертна система на основі фреймів; експертна система на основі нечіткої логіки; експертна система на основі нейронної мережі.

У даній роботі експертні системи для медичного скринінгу можуть розглядатись у контексті тестових даних для жінок різного віку зі степенем ймовірності раку шийки матки. Приклад тестових даних зображений на рисунку 1.2 :

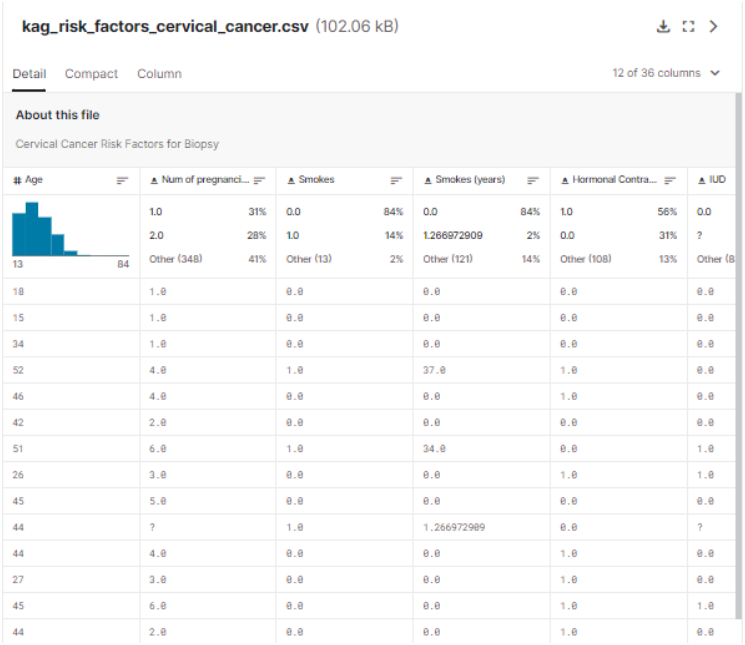


Рисунок 1.2 – Приклад тестових даних для експертної системи медичного скринінгу

Тестові даних жінок включають багато індивідуальних даних, які вказують на проблематику та рівень розвитку хвороби. Приклад класифікації даних зображено на рисунку 1.3 :

AG	AH
1.	Age
2.	Num of pregnancies
3.	Smokes
4.	Hormonal Contraceptives
5.	Hormonal Contraceptives (years)
6.	IUD
7.	IUD (years)
8.	STDs
9.	STDs (number)
10.	STDs:condylomatosis
11.	STDs:cervical condylomatosis
12.	STDs:vaginal condylomatosis
13.	STDs:vulvo-perineal condylomatosis
14.	STDs:syphilis
15.	STDs:pelvic inflammatory disease
16.	STDs:genital herpes
17.	STDs:molluscum contagiosum
18.	STDs:AIDS
19.	STDs:HIV
20.	STDs:Hepatitis B
21.	STDs:HPV
22.	STDs: Number of diagnosis
23.	STDs: Time since first diagnosis
24.	STDs: Time since last diagnosis
25.	Dx:Cancer
26.	Dx:CIN
27.	Dx:HPV
28.	Dx
29.	Hinselmann
30.	Schiller
31.	Citology
32.	Biopsy

Рисунок 1.3 – Приклад класифікації даних для експертної системи медичного скринінгу

Експертна система зазвичай складається з чотирьох основних компонентів [1, 8 – 11]:

1. База знань: це знання в експертній системі, закодовані у формі, яку система може використовувати. Він розроблений деякою комбінацією людей (наприклад, інженера знань) і автоматизованої системи навчання (наприклад, такої, яка може навчатися шляхом аналізу хороших прикладів роботи експерта).
2. Вирішувач проблем: це поєднання алгоритмів і евристик, призначених для використання Бази знань у спробі розв’язати проблеми в певній галузі.
3. Комунікатор: призначений для сприяння належній взаємодії як з розробниками експертної системи, так і з користувачами експертної системи.
4. Пояснення та допомога: це призначено для надання допомоги користувачеві, а також для надання детальних пояснень «що і чому» діяльності експертних систем, оскільки вони працюють для вирішення проблеми.

Структура представлена на рисунку 1.4 :

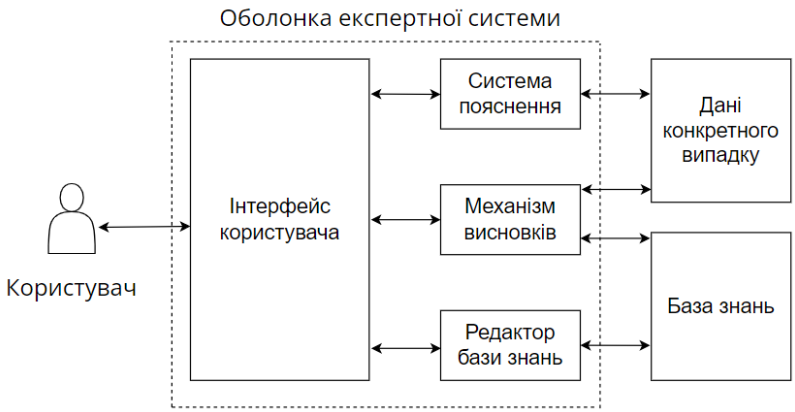


Рисунок 1.4 – Структура експертної системи

На сьогодні існує достатня кількість розробок експертних систем [11], удосконалюючи процес для ефективного рішення у медичній галузі, наприклад, рисунок 1.5 :

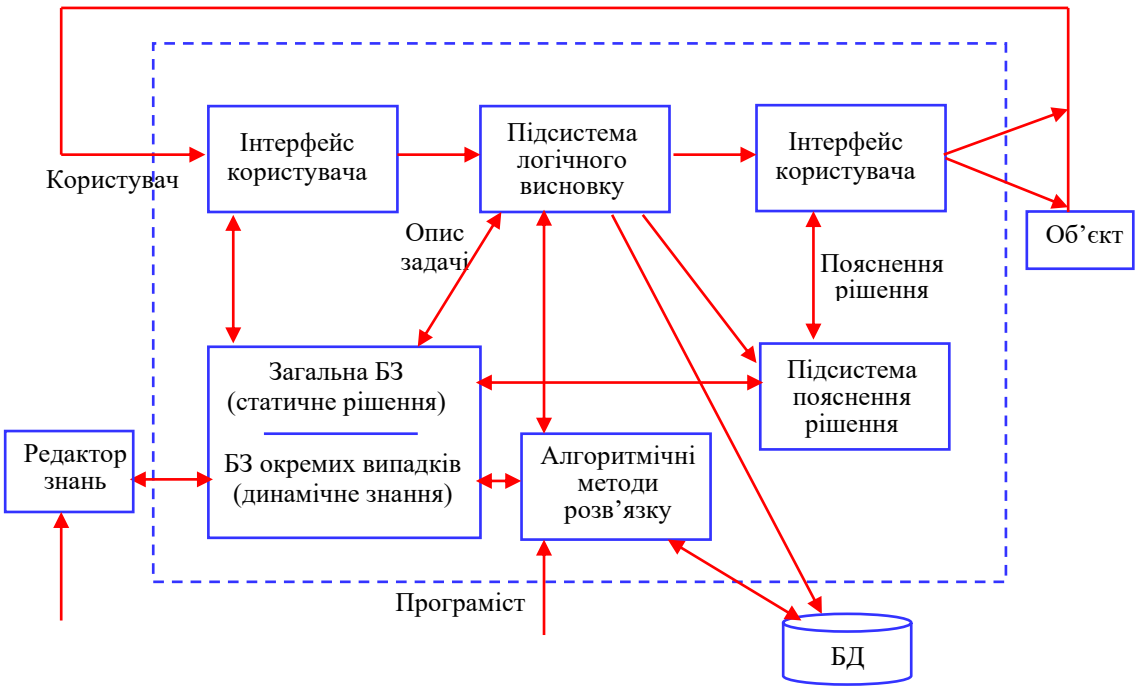


Рисунок 1.5 – Удосконалена структура експертної системи

Дослідження використання штучного інтелекту в медицині почалися на початку 1970-х років і дали низку експериментальних систем. До теперішнього

часу розроблено багато експертних систем для діагностики різних видів захворювань. Експертні системи для діагностики та лікування були розроблені для використання в різних медичних контекстах:

1. Практикуючі лікарі – лікарі лікарень, медсестри, лікарі загальної практики, консультанти, відділення швидкої та невідкладної допомоги, операційні, а також персонал будинків для престарілих, іноді батьки, самі пацієнти.
2. Основні завдання – діагностика, прогноз, лікування, моніторинг, ранні системи штучного інтелекту/підтримки прийняття рішень.

Медична діагностика - це складний когнітивний клінічний процес, що вимагає високого рівня знань. Клініцист використовує кілька джерел даних через серію алгоритмів, щоб справити діагностичне враження. Вся мета медичної діагностики полягає в тому, щоб прийти до рішення про лікування, яке, в свою чергу, призводить до хорошого прогнозу для конкретної недуги або захворювання. Тому будь-яка неправильна діагностика призведе до неправильного лікування і шляхом продовження доповнення до вартості медичної допомоги.

Це дослідження має на меті розглянути експертну систему в медичній діагностиці ESMD. ESMD є союзником, розробленим для того, щоб клініцисти могли описувати методи лікування, які повинні бути використані з урахуванням можливостей користувача. The Language Integrated Production System (CLIPS) - це інструмент, який в основному використовується для проектування ESMD. В системі ряд випадків пацієнтів буде відібрано в якості прототипів і збережено в окремій базі даних. Знання набуті в огляді літератури та експертах-людях конкретної галузі і використовуються як база для аналізу, діагностики та досліджень. Знання представляються за допомогою внутрішньодержавного формалізму, який поєднує виробничі правила та нейронну мережу. Це призводить до кращої репрезентації та полегшує отримання та підтримку знань. Запропонована система буде експериментуватися за різними сценаріями з метою оцінки її результативності.

Пацієнти звертаються за медичною допомогою для визначення (або діагностики) і лікування різних проблем зі здоров'ям. Іноді досить поєднання пацієнта і клінічного огляду у лікаря, щоб поставити діагнози і вирішити, чи потрібне медикаментозне лікування, і яке лікування слід призначити. Однак часто лабораторні дослідження або діагностичні процедури візуалізації потрібні для підтвердження клінічно підозрюваного діагнозу або для отримання більш точної інформації.

Медичний діагноз або власне процес постановки діагнозу – це пізнавальний процес. Це стрижнева пізнавальна діяльність практикуючого лікаря.

Як правило, чотири ключові фактори впливають на прийняття клінічних рішень після медичного діагнозу. Це якість, вартість, етика і правові проблеми. Показниками фактора якості в основному є процес і результат медичного втручання. Однак вони залежать від доступності медичного лікування та догляду. Як фактор, платники за медичну допомогу турбуються про витрати на медичну допомогу, включаючи витрати на діагностичне та лабораторне дослідження, тривалість перебування в стаціонарі, витрати на послуги лікаря та економічну ефективність різних профілактичних діагностичних або схеми лікування.

Оскільки вся мета медичної діагностики полягає в тому, щоб прийти до відповідного рішення про лікування, варто буде час від часу надавати заявки, які покращують медичний діагноз, щоб також покращити загальну якість медичної допомоги. Крім того, вартість охорони здоров'я в усьому світі висока. Таким чином, залучення інтелектуальної інформації про охорону здоров'я для вирішення подвійних проблем скорочення витрат на інфраструктури. Однак цей виклик не є нездоланним .

Більш того, медична діагностика – це складний процес, що вимагає високого рівня знань. Тому будь-яка спроба розробити інтелектуальну інформатику охорони здоров'я або експертну систему для медичної діагностики повинна бути готова протистояти викликам. Ці виклики включають в себе те, що інформатика охорони здоров'я відповідає обстановці, в якій вона застосовується. Саме тому проектування експертної системи для медичної діагностики тут має бути ретельно

реалізовано з поставленою метою.

Мета систем, заснованих на знаннях, полягає в тому, щоб зробити критичну інформацію, необхідну для роботи системи, явною, а не неявною. У традиційній комп'ютерній програмі логіка вбудована в код, який зазвичай може бути переглянутий лише IT-спеціалістом. За допомогою експертної системи метою було вказати правила у форматі, який був інтуїтивно зрозумілим і легко зрозумілим, переглядався та навіть редагувався експертами з доменів, а не IT-експертами.

Найбільш поширеним недоліком, який наводиться для експертних систем в академічній літературі, є задача визначення придбання знань. Знайти експертів для будь-якого програмного застосування завжди складно, але для експертних систем це було особливо складно, оскільки експерти за визначенням високо цінувалися і користувалися постійним попитом з боку організації. У результаті цієї проблеми велика кількість досліджень в останні роки роботи експертних систем була зосереджена на інструментах отримання знань, які допомагають автоматизувати процес проектування, налагодження та підтримки правил, визначених експертами. Однак при розгляді життєвого циклу експертних систем в реальному використанні інші проблеми здаються принаймні такими ж критичними, як і отримання знань. Ці проблеми були по суті такими ж, як і у будь-якої іншої великої системи: інтеграція, доступ до великих баз даних.

Проектування системи допоможе зменшити частину проблем, що виникають при аналізі старої системи, деякі з них :

1. Лікарня повинна мати можливість брати на себе будь-який ризик, який може виникнути при реалізації в разі подальшого навчання персоналу.
2. Необхідна мова програмування призведе до того, що реалізація набуде нового виміру.
3. Валідація, яка визначає, що тільки правильні дані повинні вноситися в програму користувачами.
4. Послуги медичного обслуговування набудуть нового виміру.

5. Нарешті, керівництво лікарні повинно мати можливість забезпечити достатньо коштів для реалізації.

Надалі розглянемо методи машинного навчання, які застосовуються в експертних системах для медичного скринінгу.

1.2 Методи машинного навчання для медичного скринінгу

Машинне навчання — це галузь штучного інтелекту, яка досліджує моделювання та розробку обчислювальних алгоритмів на основі навчання даних, а не за попередньо запрограмованими інструкціями. Основна мета моделі машинного навчання полягає у створенні інформаційної системи, яка вивчає заздалегідь визначену базу даних і, зрештою, генерує модель для передбачення, класифікації або виявлення [12]. Діджиталізація медичних записів, лабораторних тестів і візуалізації спонукала до зростання кількості баз даних, які у своєю чергу, слугують джерелами для застосування методів машинного навчання з метою профілактики, ранньої діагностики та лікування захворювань у медичній сфері.

Автори дослідження [13] представили процес розробки алгоритму машинного навчання (рис. 1.2.1) :

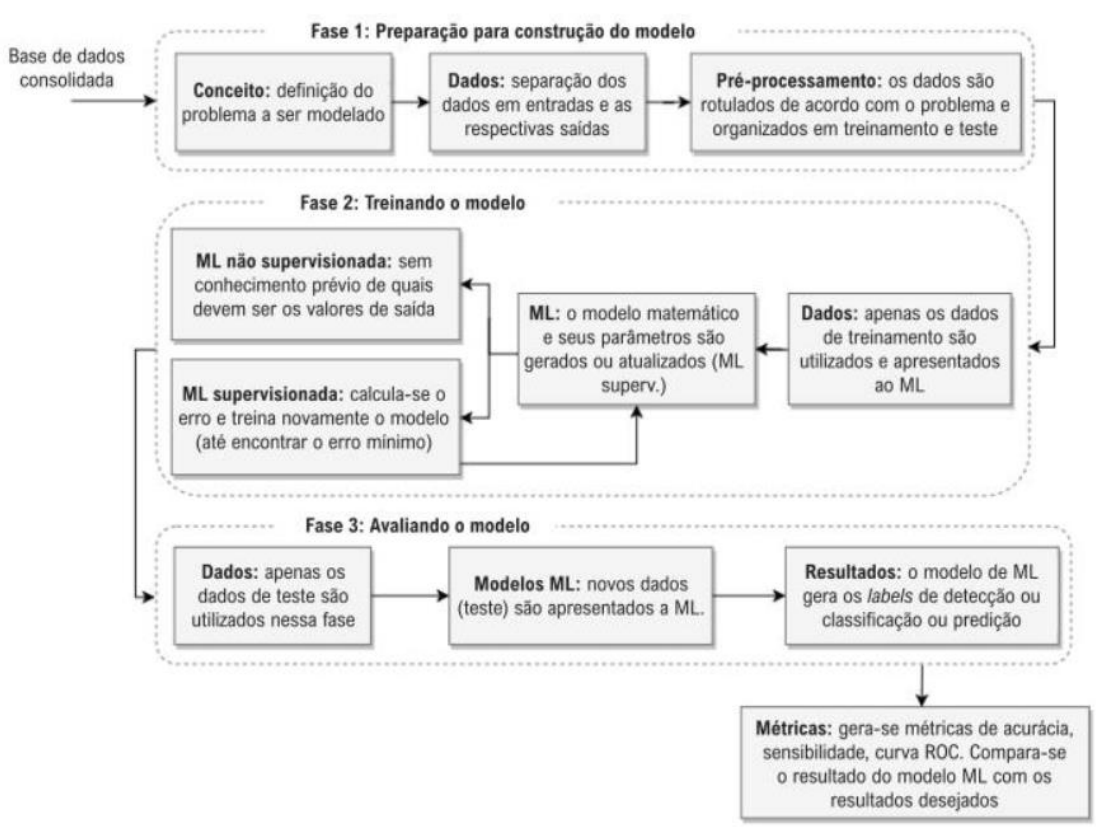


Рисунок 1.2.1 – Етапи розробки алгоритмів машинного навчання

Алгоритм складається з трьох етапів: попередня обробка, навчання та тестування моделі. Перший етап складається з організації банку даних, визначення питання дослідження та розподілу даних на навчання та тестування. Другий етап – навчання може здійснювати як контрольоване, так і не контрольоване (табл. 1.2.1). На етапі тестування модель порівнюється з тестовими даними та генеруються результати. Таким чином, алгоритми навчаються за допомогою повторних спостережень і встановлюють шаблон відображення, щоб позначити дані та створити модель, яка узагальнює інформацію, щоб нові дані (які ніколи не аналізувалися алгоритмом) могли бути точними та надійними.

Таблиця 1.2.1 Порівняльний аналіз контрольованого і неконтрольованого навчання (навчання з учителем і навчання без учителя – інші назви)

Властивості	Контрольоване навчання (Supervised learning)	Неконтрольоване навчання (Unsupervised learning)

Визначення	Алгоритм, який відносить елемента до певного класу із заздалегідь відомими параметрами, отриманими на етапі навчання**.** Кількість класів при класифікації – строго обмежена	Алгоритм розбиття даних на кластери, де у кожному є схожі об’єкти, а об’єкти різних кластерів мають бути якомога більш відмінні. Кількість кластерів – невідома, визначається у процесі алгоритму.
Приклади	Регресія, дерево рішень, штучні нейронні мережі	Факторний аналіз, кластерний аналіз, дискримінантний аналіз
Функції	Розробляють моделі з навчальних даних, і ці моделі можна використовувати для класифікації інших немаркованих даних.	Здатні самостійно визначити кількість кластерів, на які потрібно розбити дані, а також виділити характеристики цих кластерів без участі людини, тільки за допомогою використовуваного алгоритму.
Переваги	Допомагає оптимізувати критерії ефективності, використовуючи досвід; швидке і автоматичне обчислення для великих даних; точні і надійні алгоритми.	Знаходить види невідомих шаблонів у даних і нові функції, які можуть бути корисними при кластеризації
Недоліки	Потребує часу та технічних знань від команди висококваліфікованих спеціалістів із обробки даних.	Менш точний і надійний алгоритм; складні обчислювальні процеси.

Перейдемо до розгляду алгоритмів машинного навчання, де

використовується контрольоване навчання.

1.2.1 Логістична регресія

Логістична регресія – це вид ймовірнісної статистичної моделі, що розрахована на аналіз між декількома незалежними змінними. Також можна розглядати цю регресію як керований алгоритм навчання. Цільовий клас, який потрібно передбачити – це залежна зміна, а незалежні змінні – це атрибути чи функції, що необхідні для прогнозування цільового класу. Зазвичай використовується у тому випадку, коли є тільки два класи, наприклад, так чи ні, чи 0 та 1, чоловіки та жінки, рослини та тварини, вік. При необхідності можна описати та проаналізувати зв’язок між бінарною залежною змінною та декількома порядковими незалежними змінними. Визначити результат потрібно за допомогою логістичної функції, яка оцінює ймовірність, а потім визначає найближчий клас – він може бути що позитивний, що негативний, до отриманого значення ймовірності [14].

Може бути записана проста формула для моделі лінійної регресії (1.2.1.1) [14]:

$$y = F(x_1, x_2, \dots, x_n) \qquad (1.2.1.1)$$

У лінійній регресії зв’язок між залежними та незалежними змінними дотримується лінії, як зв’язок між ними. Лінія регресії зазвичай представляється рівнянням: Y = a * X + b (рис. 1.2.1.1) [15] :

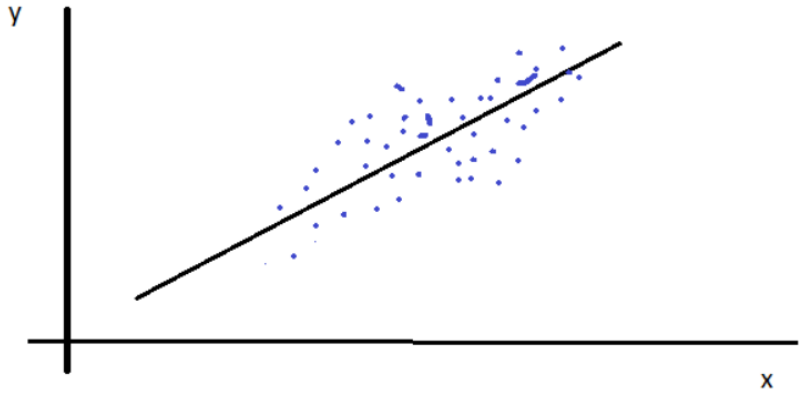


Рисунок 1.2.1.1 – Лінія лінійної регресії

Коли в алгоритмі існує більше, ніж одна зміна, тоді потрібно припускати таку модель – це називається множинна лінійна регресія (1.2.1.2) [16]:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad (1.2.1.2)$$

Зазвичай, рівняння показує зв’язок між результатом та декількома незалежними змінними.

Метод, що використовується для того, щоб виявити вплив незалежних змінних на залежну змінну називається множинним регресійним аналізом. Залежна змінна може змінюватись відносно незалежних змінних – таким чином можна передбачити наслідки змін у аналізі.

Логістична регресія використовує модель, що спрямована на математичний розбір біноміального результату з декількома змінними, що пояснюють наслідок.

Також негативні чи позитивні класи вказують на те, яка ймовірність результату за допомогою статистичних методів [15].

Приклад побудови логістичної регресії, де відбувається співвідношення шансів [17] чи подія буде успішною чи негативною зображено на рисунку 1.2.1.2 (1.2.1.3):

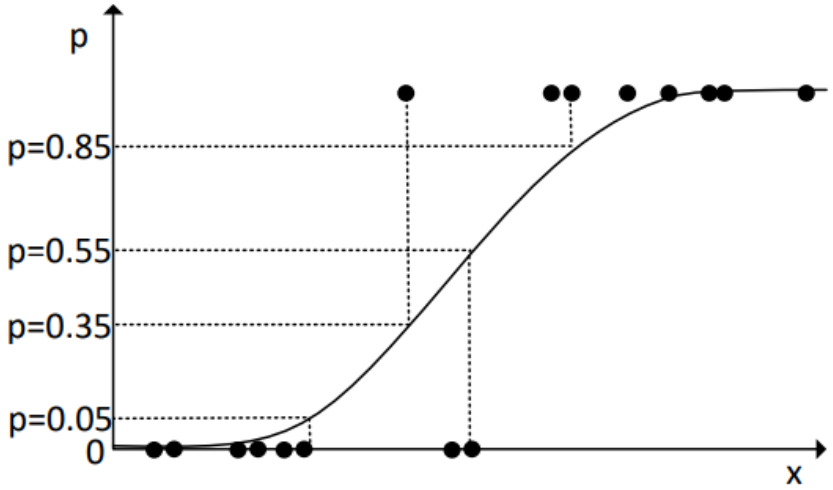


Рисунок 1.2.1.2 – Співвідношення OR(odds ratio)

$$OR = p / (1 - p) \qquad (1.2.1.3)$$

Де p –ймовірність успіху, log (OR) = y .

Для того, щоб оцінити коефіцієнти застосовують метод максимальної правдоподібності, основою є функція правдоподібності, яка виражає точність ймовірності за результатами вибірки.

Навчальна вибірка може бути виключно дискретного типу. Для успішного результату потрібно, щоб кількість прикладів була в кілька разів більшою, аніж кількість вхідних ознак. Якщо вхідних даних мало необхідно штучно зменшувати структуру регресійної моделі, залишаючи тільки основні ознаки. Для залежної змінної варто визначити, що є негативною чи позитивною подією. Наприклад, якщо вираховувати алгоритм наявності захворювання, тоді позитивним буде «Ознак раку шийки матки немає», негативним – «Дані показники вказують на наявність захворювання раку шийки матки».

Існує чотири варіанти класифікації [18]:

1. Істино позитивні випадки (True Positives).
2. Істино негативні випадки (False Negatives).
3. False Negatives – приклади, що являються негативними, але є помилкою другого типу, коли подія, що цікавить, не виявляється.

4. False Positives – приклади, що являються позитивними, але є помилкою першого типу, помилкове виявлення, що помилково ухвалюється про її наявність.

Приклад чотирьох варіантів класифікації зображено на рисунку 1.2.1.3 [18] :

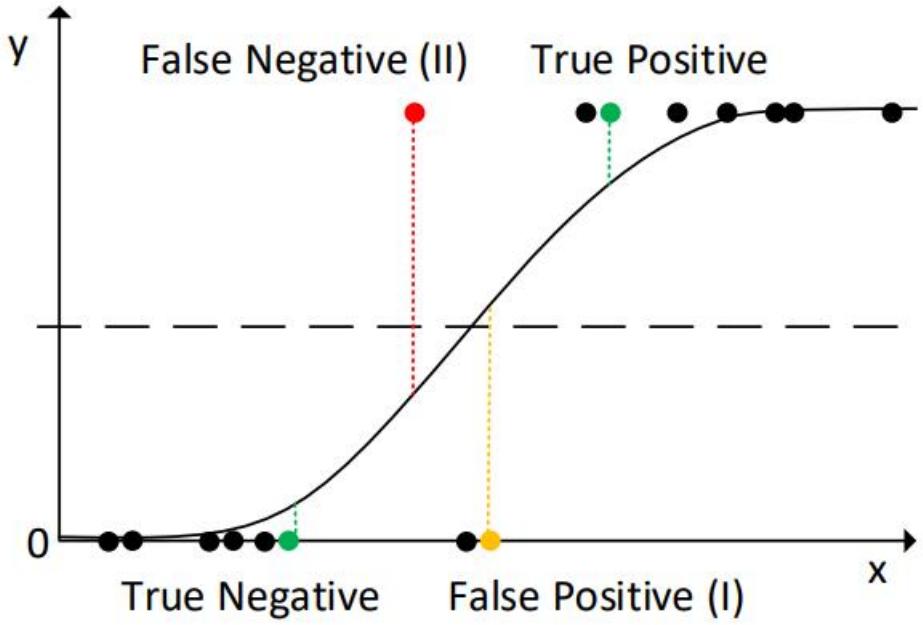


Рисунок 1.2.1.3 – Приклад чотирьох класифікацій для логістичної регресії

Приклад таблиці для оцінки результатів зображено на рисунку 1.4 [19]:

		Модель	
		Негативно	Позитивно
Фактично	Негативно	TN	FP (I)
	Позитивно	FN (II)	TP

Рисунок 1.4 – Приклад таблиці для оцінки результатів класифікацій

На основі таблиці оцінюють [19] :

1. Точність моделі (1.2.1.4) :

$$Error\ Rate = \frac{FP + FN}{Total} \quad (1.2.1.4)$$

2. Помилки (1.2.1.5) :

$$Accuracy\ Rate = \frac{TP + TN}{Total} \quad (1.2.1.5)$$

3. Частина істино позитивних випадків, які правильно ідентифіковані моделлю – чутливі (1.2.1.6):

$$Recall = \frac{TP}{TP+FN} \quad (1.2.1.6)$$

4. Специфічність – частина істино негативних випадків, які були правильно ідентифіковані моделлю (1.2.1.7):

$$Sp = \frac{TN}{TN + FP} \quad (1.2.1.7)$$

ROC аналіз використовується з використанням графіків ROC кривих (Receiver Operator Characteristic) для оцінки моделей. Також називають її як крива похибок. Використовується для оцінки якості бінарної класифікації, що відображаються спільний зв’язок між частиною загальної кількості об’єктів до правильно класифікованих об’єктів, помилково класифікованих, ті, що мають ознаку [20].

Будувати ROC криву потрібно так: чим ближче до лівого кута координат розташована крива – тим точніший результат алгоритму та якість системи

відображення даних. Графік залежності будується по осі Y.

Якщо крива лежить ближче до діагоналі, то інформативність моделі низька.

Приклад зображено на рисунку 1.2.1.4 [20] :

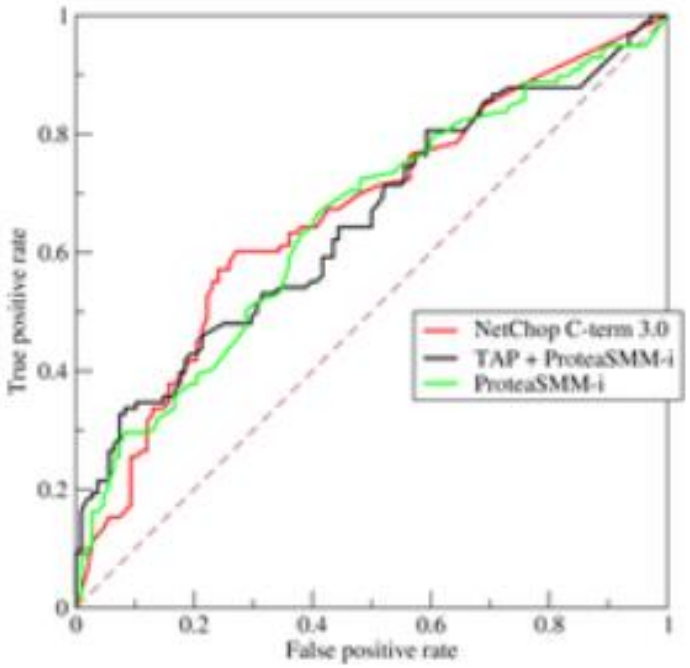


Рисунок 1.2.1.4 – Приклад ROC кривої

За найбільш точний результат відповідає критерій «Чутливість» для істино позитивних рішень, а для помилково позитивних – критерій «Специфічність». Але досягнути ці два показники в одній моделі неможливо, оскільки одночасно підвищити специфічність та чутливість на практиці практично виключається. Поріг відсікання допомагає досягти посередніх значень для обох класифікаторів – співвідношення Se і Sp, де Se – це значення чутливості, специфічності – Sp. Вісь Y – Se(чутливість), по осі X – Sp(специфічність).

1.2.2 Дерево рішень

Дерево рішень – це діаграма або механічна схема презентації класифікації вибору. Деревом називають цей метод рішення, тому що це лінія, яка розгалужується на варіанти рішень.

Компоненти моделі дерева рішень :

1. Три поля. Поле дій – всі можливі шляхи до розв’язку задачі. Поле можливих подій – ймовірності реалізації кожної альтернативи виникнення ситуацій. Поле можливих наслідків або поле очікуваних результатів – опис кожного результату в залежності від обраної ситуації.
2. Три компоненти або три типи вузлів. Перша точка прийняття рішення – на моделі зображена у вигляді квадрата та означає місце, де потрібно прийняти остаточне рішення. Точка можливостей – зображена у вигляді кола вказує на очікувані результати можливих подій.
3. Гілки дерева – зображені у вигляді ліній, які йдуть від першої точки прийняття рішень до результату реалізації кожної ситуації, якщо вузол замикаючий, то зображений у вигляді трикутника [21].

На даному рисунку 1.2.2.1 представлена структура дерева рішень :

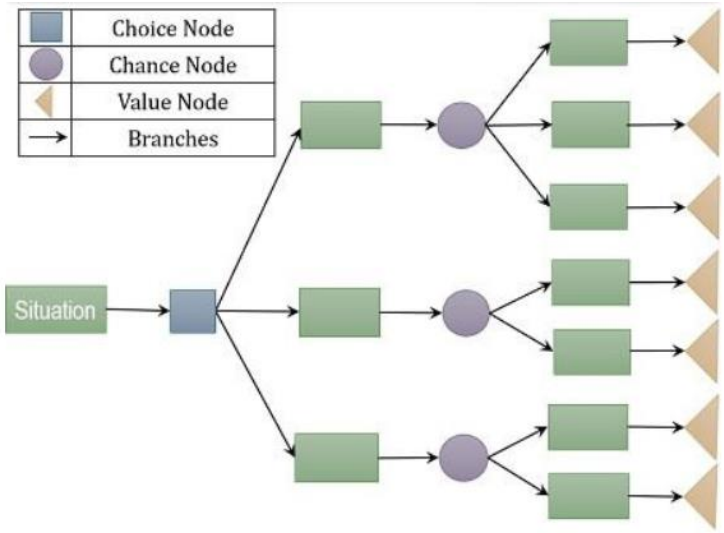


Рисунок 1.2.2.1 – Структура дерева рішень

Цей метод використовується найчастіше для прогнозування та вибору класифікації. Також дерево рішень називають, як деревами класифікації, деревами регресії, деревами вирішальних правил. Якщо цільова змінна приймає дискретне

значення – розв’язується задача класифікації, якщо безперервне значення – прогнозування. Отже, дерево рішень – це структура варіації запитань відповіді на які «Так» чи «Ні». Кожен варіант запитання називається вузлом. Розв’язанням являється кінцевий вузол дерева [22].

Недоліки дерев рішень [23]:

1. Обмежене число розв’язку. Тобто, при побудові дерева рішень потрібно будувати його неперевантаженим, оскільки це зменшує можливість до узагальнення класу розв’язку, відповідно розв’язок може бути неточним.
2. Розв’язок дерева рішень може бути неоптимальним.
3. Якщо мітка класу домінує над рішенням – розв’язок може бути упередженим.
4. Дерево рішень може бути побудованим надто перенавантаженим, що призводить до неповних даних.

Переваги дерев рішень [24]:

1. Інтуїтивність зображення класифікаційної моделі дерева рішень допомагає просто використовувати та розуміти розв’язувану задачу. Дерево рішень дозволяє зрозуміти чому об’єкт відноситься саме до цього класу.
2. За допомогою гнучкості дерева рішень дослідник може використовувати дані з бази даних за допомогою рядків умов.
3. Дерево рішень допомагає скласти модель для тієї сфери, яка складно формулює умову позначень.
4. Побудова моделі дерева рішень не вимагає від користувача вибору незалежних змінних(вхідних атрибутів). Тобто, на початку роботи з алгоритмом потрібно вписати всі існуючі дані, алгоритм сам вибере найбільш важливі на буде використовувати їх для побудови дерева.
5. Точність моделі є більш якіснішою, аніж, для прикладу, нейронні мережі чи статистичні методи.
6. Масштабованість алгоритму допомагає обробляти великі бази даних. У

цьому допоможе алгоритм SLIQ або SPRINT.

7. На навчання та побудову класифікаційних моделей за допомогою алгоритмів спорудження дерев рішень потрібно значно менше часу, аніж, для прикладу, нейронним мережам.
8. Дерево рішень працює як і з числовими, так і з категоріальними типами даних.

Дерева рішень часто використовуються в медицині та охороні здоров'я вже більше 20 років. Автори дослідження [25] здійснили огляд більше сорока літературних джерел у цьому напрямі і довели, що дерева рішень — це надійна та ефективна техніка прийняття рішень, яка забезпечує високу точність класифікації з простим представленням зібраних знань. При використанні дерев рішень сам процес прийняття рішень може бути легко перевірений експертом. Через ці причини дерева рішень особливо підходять для підтримки процесу прийняття рішень у медицині.

1.2.3 Штучні нейронні мережі

Нейронні мережі, також відомі як штучні нейронні мережі (ANN) або імітовані нейронні мережі (SNN), є підмножиною машинного навчання та є основою алгоритмів глибокого навчання. Їх назва та структура навіяні людським мозком, імітуючи спосіб, яким біологічні нейрони передають сигнали один одному.

Штучні нейронні мережі — це математичні системи, які використовуються як перспективний інструмент для надійного, гнучкого та швидкого оцінювання. Вони демонструють високу потужність в оцінці багатофакторних даних, засвоєнні інформації з багатьох джерел і виявленні тонких і складних закономірностей.

Приклад нейронної мережі зображено на рисунку 1.2.3.1 :

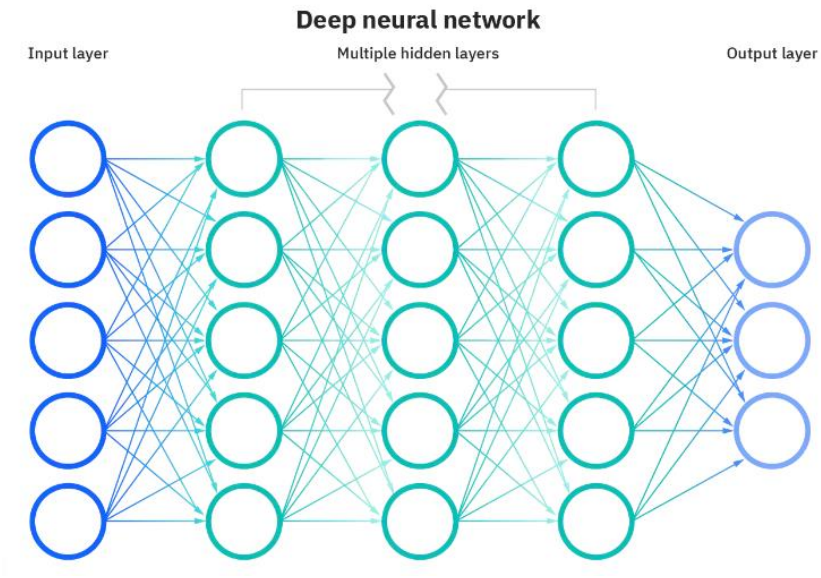


Рисунок 1.2.3.1 – Приклад нейронної мережі

Штучні нейронні мережі побудовані з взаємопов’язаних нейронів. Кожен нейрон підсумовує зважені вхідні дані і передає результат нелінійної функції, зазвичай називається сигмоїдною функцією, щоб створити вихід. Декілька нейронів з’єднані один з одним таким чином, що вихід одного є входом для іншого чи інших, таким чином утворюючи зв’язок, подібний до зв’язку людського мозку. Загальна модель взаємозв’язку не вимагає спеціальної структури.

Популярні моделі взаємозв’язків визначають багаторівневий підхід, який включає нейрони з входами/стимулами із зовнішнього світу, що забезпечує вихідні дані системи. Нейрони з’єднані зв’язками, і кожна зв’язок має числовий показник ваги. Нейронна мережа «навчається» шляхом повторних коригувань цих ваг.

Однією з важливих характеристик штучних нейронних мереж є те, що вони можуть вчитися на своєму досвіді в навчальному середовищі. Штучний нейрон має у своєму складі суматор та функціональний перетворювач [26]. На рисунку 1.2.3.2 представлена структура штучного нейрона :

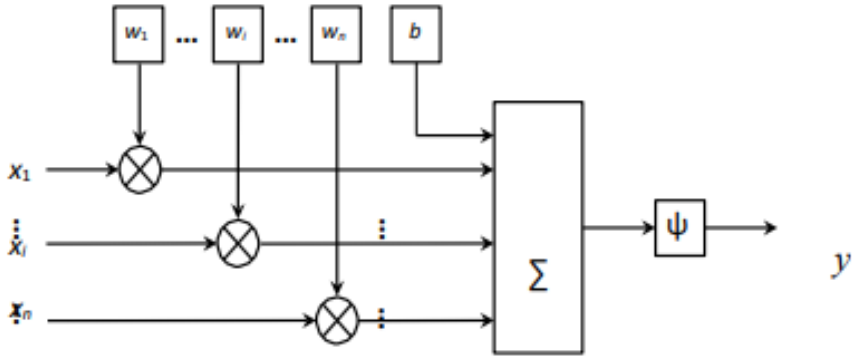
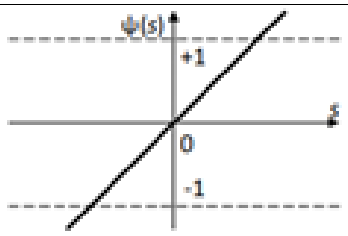
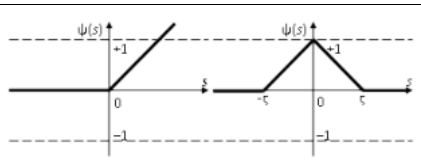
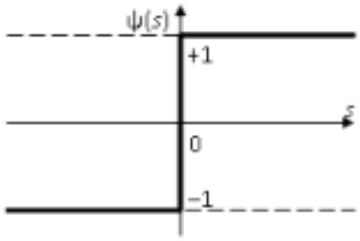


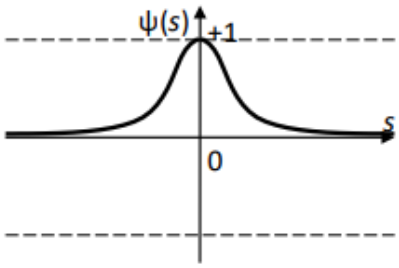
Рисунок 1.2.3.2 – Структура штучного нейрона

У – вихідний сигнал нейрона; n – вхідний номер нейрона; x_i – і-й вхідний сигнал; w_i – вага і-го входу нейрона; b – параметр зміщення суматора; ψ (·) – характерна функція (функція активації) нейрона. Функції активації є фундаментально важливим будівельним блоком нейронів. З його участю кожен нейрон здатний підсилювати або послаблювати сигнал, що надходить на його вхід (подібно до природних нейронів, які досягають збудження або гальмування нервових імпульсів). Здатність підсилювати або гасити імпульси регулюється сигналами, що проходять через нейрони. Як природні, так і штучні нейрони можуть навчатися на основі активності процесів, що в них відбуваються. Крім того, в результаті навчання змінюються ваги зв'язків між нейронами, що впливає на поведінку відповідних нейронів. Найпоширеніший тип функції [26] використовується як функціональний трансформатор для нейронів представлені у таблиці 1.2.3.1:

Таблиця 1.2.3.1 – Найпоширеніші види функцій для нейронної мережі

Назва і вид	Графік	Застосування
Лінійна функція активації (linear function)		Для нейронів вихідного шару персептрона, якщо результуюча змінна не має обмежень і може

		набувати будь-яких значень.
Кусочно-лінійна функція активації (piecewise-linear function)		Коли моделюється деяка величина, що не повинна виходити за встановлені обмеження
Пороговий тип функції активації (threshold function)		Коли результуюча змінна може набувати тільки двох значень – – 1 і +1 (сигнатурна або сигнум-функція) чи 0 і +1 (функція Хевісайда або одиничного стрибка), наприклад, у задачах класифікації або на рекурентному шарі в мережах асоціативної пам’яті
Сигмоїдна функція активації (sigmoid function).		Можливість диференціювання дозволяє використовувати градієнтні методи для оптимізації параметрів моделі (зокрема, метод зворотного поширення помилки).

Радіально-базисна функція активації (radial-basis function) т		Для розв'язання задач, де значення змінних розподілені за нормальним законом, або в радіально-базисних нейронних мережах

Розглянемо основні етапи побудови та функціонування самонавчаючих алгоритмів, які складають основу нейромережі [26, 27]:

Типовий цикл створення корисної системи штучної нейронної мережі складається з кроків, зображених на рисунку 1.2.3.3. Перший крок включає збір відповідної кількості даних, розмір яких пов'язаний із природою проблеми. Під час другого кроку дані попередньо обробляються з точки зору очищення будь-яких узгоджених записів, застосування математичних перетворень для відображення буквено-цифрових даних у числа, придатні для моделі штучної нейронної мережі, і поділ доступних даних на 2 набори: навчальний набір, який буде використовуватися для формування штучної нейронної мережі, і тестовий набір, який використовується для оцінки його продуктивності. Третій крок включає вибір відповідної моделі штучної нейронної мережі, характеристики параметрів (наприклад, швидкість навчання, кількість ітерацій або сигмоїдна функція) і, у подальшому, навчання штучної нейронної мережі. На останньому етапі відбувається оцінка продуктивності за допомогою тестового набору або повного набору; якщо результати задовільні, систему можна розмістити у робочому середовищі для звичайного використання, якщо ні, усі кроки слід повторно оцінити з самого початку. Побудова штучної нейронної мережі призводить до розробки алгоритмів, спрямованих на вирішення певного питання або проблеми. Різниця між штучною нейронною мережею і звичайною обчислювальною системою полягає в тому, що за своєю здатністю до навчання через навчання останній також багато в чому нагадує функцію навчання мозку. Завдяки цьому процесу він самостійно

адаптується до вправи, змінюючи свої структурні характеристики на основі зовнішньої або внутрішньої інформації, яка протікає через мережу.

Наступний рисунок 1.2.3.3 представляє типову процедуру створення штучної нейронної мережі [28] :

Рисунок 1.2.3.3 – Процедура створення штучної нейронної мережі

Переваги та недоліки штучних нейронних мереж представлені у таблиці 1.2.3.2 :

Таблиця 1.2.3.2 – Переваги та недоілки штучних нейронних мереж

Переваги	Недоліки
1.На основі циклічної обробки навчальних вибірок створюється власний алгоритм розв’язування штучної нейронної мережі	1.Вимагає досвідчених інструкторів системи
2.Вирішується проблема класифікації на основі якісних і кількісних відмінностей	2.Тривалий ресурс часу
3.Необхідна менша формальна статистична підготовка	3.Перенавчання
4.Можна використовувати, коли змінні, що впливають на певну подію, точно не відомі	4.Обмежена здатність ідентифікувати можливі причинно-наслідкові зв’язки, спосіб, час і тип параметрів, які підлягають навчанню
5. Може виявляти складні нелінійні зв’язки між залежними та незалежними	

змінними та всі можливі взаємодії між змінними-прогнозами	

Найпоширенішими застосуваннями нейронних мереж у медичній галузі [29] є:

1. Класифікація – розподіл даних за параметрами. Наприклад, на вхід дається набір людей і потрібно вирішити задачу розподілення потоку пацієнтів, для проходження щорічних профілактичних оглядів. Цю роботу може зробити нейронна мережа, аналізуючи таку інформацію як: стать, вік, тип зайнятості та інші дані, що містяться в електронній медичній картці.
2. Передбачення – можливість прогнозування. Наприклад, прогнозування зростання або падіння пацієнтопотоку, ґрунтуючись на даних звернення до фахівців, епідемічної ситуації в регіоні тощо.
3. Розпізнавання – на сьогодні найширше застосування нейронних мереж. Використовується в багатьох медичних системах, для розпізнавання даних зі зображеннями обстежень. Наприклад, за даними IBM Watson аналізу результатів флюорографічних обстежень, площа кривої помилок алгоритму становила 0,788.

2 РОЗРОБКА ЕКСПЕРТНОЇ СИСТЕМИ ДЛЯ МЕДИЧНОГО СКРИНІНГУ НА ОСНОВІ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ

2.1 Обґрунтування методів вибору кластерного аналізу для медичного скринінгу

На основі огляду наукової літератури були виділені основні вимоги до експертної системи у медицині:

1. Копіювати поведінку лікаря для пошуку клінічного діагнозу.
2. Швидко та структуровано адаптуватись до переіначення старих чи нових висновків для медичних знань.
3. Представляти розв’язки у такому вигляді, щоб вони були зрозумілі лікарю та пацієнту.

Розробка експертної системи є нелегкою працею. Для цього потрібно мати великі людські і матеріальні ресурси, що під силу великим компаніям. Впровадження готових експертних систем теж вимагає, по-перше, інженерів, які будуть підтримувати роботу експертних систем, по-друге, навчених медичних працівників, які б могли застосовувати їх у своїй діяльності.

У зв’язку з цим, ми здійснили спробу використати методи кластерного аналізу для медичного скринінгу. Обґрунтовується це тим, що кластерний аналіз не вимагає окремого програмного забезпечення, його можна здійснити за допомогою пакету Statistica.

Наукові дослідження у галузі медицини [30 – 36] сприяли також вибору методів кластерного аналізу для медичного скринінгу пацієнтів. Так, у роботі [34] підкреслюється, що протягом багатьох років класифікація відігравала важливу роль у медицині: пацієнтів можна класифікувати на основі деяких так званих змінних ознак, що стосуються клінічно оцінених симптомів і лабораторних вимірювань. Наприклад, у медичній діагностиці остаточну класифікацію пацієнта

часто можна зробити лише після вичерпного фізичного та клінічного обстеження або, можливо, хірургічного втручання. У деяких випадках справжню класифікацію можна зробити лише на основі доказів, які з’являються через деякий час, наприклад, розтин. Тому часто використовуються діагностичні тести. Там, де це можливо, тести базуються на клінічних і лабораторних спостереженнях, які можна зробити без зайвих незручностей для пацієнта. Класифікація також дуже корисна в медичному прогнозі, де групи в схемі класифікації відповідають можливим результатам захворювання. Проте, здебільшого бувають ситуації, коли відсутні апріорні знання про базову групову структуру. Хоча в деяких ситуаціях можуть бути обмежені знання про групову структуру, в тому, що кількість груп може бути відома, однак у цьому випадку немає навчальних даних відомого походження для оцінки групово-умовних розподілів. Також у деяких випадках корисної інформації щодо можливих причин захворювання може бути мало або взагалі не бути. В інших випадках можуть бути деякі відомі причини захворювання, яке досліджується, і в цьому випадку однією з цілей кластерного аналізу може бути об’єднання пацієнтів у групи відповідно до відомих причин захворювання. Друга мета може полягати в тому, щоб з'ясувати, чи існують ще невиявлені причини захворювання. Попереднє кластеризування пацієнтів у передбачувану кількість груп може бути отримано спочатку, відповідно до клінічного діагнозу пацієнтів з використанням однієї або, можливо, деяких змінних ознак. Якщо використовуються всі змінні функції, то це, як правило, обмежено та досить випадково. Наприклад, пацієнт може бути діагностований як такий, що належить до певної групи захворювань, якщо будь-яка з виміряних змінних ознак потрапляє в діапазон, який традиційно асоціюється з цією групою. Тому цікаво обчислити статистичний діагноз, отриманий за допомогою кластерного аналізу, і порівняти його з клінічним діагнозом.

Таку ж думку висловлюють науковці у розробці [32]. З появою сучасних методів збору наукових даних велика кількість даних біомедичної та медичної інформатики накопичується в різних базах. Щоб отримати з них корисну інформацію, необхідні ефективні методи аналізу даних. Кластерний аналіз є одним

із основних методів інтелектуального аналізу даних, який допомагає ідентифікувати природні групи та цікаві моделі з величезних банків даних.

Кластерний аналіз є видом статистичного групування для того, щоб зробити дані у кожному кластері максимально схожими один на одного та різними по відношенню до інших кластерів. Цей метод дає можливість відносити об'єкти до однієї групи не за одним показником, а за декількома водночас. Також допомагає віднайти структуру даних, що неможливо зробити з боку експерта чи зовнішньому аналізі [37, 38].

Нехай множина A= $\{a_1, a_2, \dots a_n\}$ – множина об'єктів, B – множина номерів кластерів. Вибирається метрика (найчастіше формула відстані) $d_{ij}(x_{ik}, x_{jk})$. Необхідно розбити множину A на підмножини (кластери), які не перетинаються, і кожен кластер містив об'єкти близькі за метрикою $d_{ij}(x_{ik}, x_{jk})$, а об'єкти різних класів істотно відрізнялися. Кожному об'єкту a_i приписується номер кластера B_i.

Множина A може складатися із об'єктів, які мають різні одиниці вимірювання або різний діапазон представлених значень, тому потрібно здійснити нормування вхідних даних. При кластерному аналізі є два основні способи нормалізації даних: MinMax-нормалізація та Z-нормалізація.

MinMax-нормалізація здійснюється наступним чином (2.1.1) :

$$x' = \frac{x - \min[X]}{\max[X] - \min[\mathrm{X}]} \tag{2.1.1}$$

У разі всі значення будуть у діапазоні від 0 до 1; дискретні бінарні значення визначаються як 0 та 1.

Z-нормалізація (2.1.2) :

$$x' = \frac{x - \mathrm{M}[\mathrm{X}]}{\sigma[X]}, \tag{2.1.2}$$

Де M[X] – математичне сподівання, $\sigma[X]$ – середньоквадратичне відхилення.

Основою практичного застосування кластерного аналізу завжди є деяка аксіоматика перетворень і введення метрики, яка служить для оцінки досліджуваних об'єктів і відстаней між ними. Як правило, відстань між двома об’єктами представлена невід’ємною функцією близькості, яка є вводиться для будь-яких об’єктів кластерного аналізу. Якщо розглядати реальні умови контролю, то ефективніше порівнювати об'єкти за інтегральними характеристиками. На жаль, цей спосіб далеко не завжди застосовний через неможливість узгодження всіх одиниць виміру з урахуванням різниці метричних полів.

У кластерному аналізі можуть використовуватися міри подібності: коефіцієнти кореляції, міри відстані, коефіцієнти асоціативності, ймовірнісні коефіцієнти подібності.

У нашому дослідження ми застосуємо міри відстані, які представлені у таблиці 2.1.1 :

Таблиця 2.1.1 – Відстані для кластерного аналізу

Назва відстані	Формула
Евклідова відстань	$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$
Зважена Евклідова відстань	$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k \left(x_{ik} - x_{jk}\right)^2}$
Метрика Мінковського	$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \left\lvert x_{ik} - x_{jk} \right\rvert^r}$

Хеммінгова відстань	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$

Отже, основними кроками кластерного аналізу є наступні 5 кроків.

Крок 0. Підготовка емпіричної інформації для кластеризації.

Крок 1. Визначення множини змінних-ознак, які описують дані об’єкти і за якими будуть оцінюватися ці об'єкти.

Крок 2. Вибір міри подібності між об'єктами відповідно до обраної метрики (відстані) і обчислення цієї відстані.

Крок 3. Вибір алгоритму кластерного аналізу і його застосування для групування об'єктів у кластери.

Крок 4. Перевірка достовірності результатів кластерного аналізу.

На сьогодні існує величезна кількість алгоритмів кластерного аналізу та їх модифікацій. Найбільш загальні представлені на рисунку 1.2.1 :

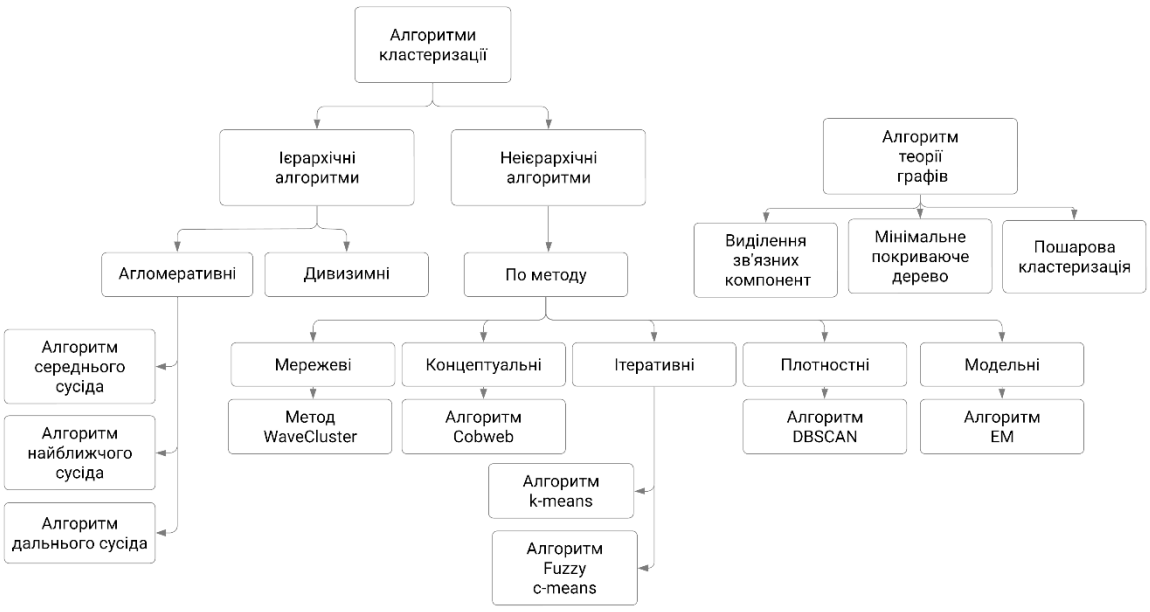


Рисунок 1.2.1 – Алгоритми та методи кластеризації

Вибір алгоритму для конкретної задачі більшою мірою визначається математичною культурою, наявністю програмно-технічного забезпечення з

доступом дослідника, тип і потужність інформаційного поля, а також залежить від багатьох інших факторів [39].

У нашому дослідженні будемо застосовувати ієрархічний агломеративний алгоритм найближчого сусіда та неієрархічний ітеративний алгоритм k-means.

2.2 Математичне моделювання експертної системи на основі ієрархічного агломеративного методу ближнього сусіда

Ієрархічні (агломерація та розщеплення) методи в основному використовуються для побудови ієрархічних дерев відносно невеликих агрегатів. Порівняно з іншими методами перевага ієрархічного методу полягає в тому, що він дозволяє більш повно і тонко проаналізувати структуру досліджуваної сукупності, а також чітко представити результати кластеризації. Їх головними недоліками є обчислювальна громіздкість, пов’язана з повторним обчисленням усієї матриці відстані на кожному кроці, і «обмежена неоптимальність» гранично оптимальних алгоритмів у багатьох випадках [40].

Метод найближчого сусіда є найпростішим для розуміння ієрархічним агломеративним методом для кластерного аналізу. У цьому випадку процес класифікації починається з пошуку та об’єднання двох найближчих один до одного об’єктів у матриці подібності. На наступному етапі знаходять наступні два найближчі об'єкти і так до тих пір, поки матриця подібності не буде повністю вичерпана. Як правило, робота алгоритму закінчується, коли всі спостереження об'єднуються в один клас. Щоб розрізнити кластери, після завершення кластеризації встановлюється поріг подібності, досягнувши цього порогу, можна розрізнити кілька кластерів. На рисунку 2.2.1 представлено процес виконання алгоритму ближнього сусіда :

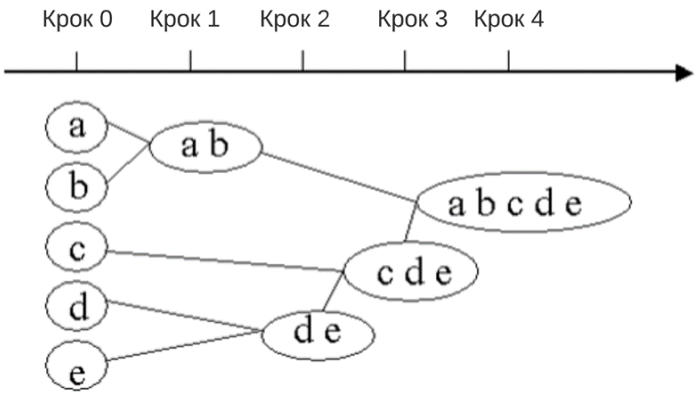


Рисунок 2.2.1 – Процес виконання алгоритму ближнього сусіда

На наступному рисунку 2.2.2 представлено код виконання алгоритму ближнього сусіда :

```
float tx = width_source/width_dst;
float ty = height_source/ height_dst;

for(i=0; i<height_dst; i++)
 for(j=0; j<width_dst; j++)
  {
      x = ceil(j*tx);
      y = ceil(i*ty);
      U(i,j) = P(y,x);
 }
```

Рисунок 2.2.2 – Код виконання алгоритму ближнього сусіда

У медичній галузі [30] групування ґрунтується на різних кількісних показниках подібності, так що симптоми в одному кластері більш схожі один на одного, ніж на симптоми в іншому кластері. Наприклад, якщо шаблон відповідей на два симптоми подібний для більшості людей, симптоми групуються, тоді як різні шаблони відповідей припускають, що симптоми оцінюються більш незалежно. Групи послідовно приєднуються до найближчого угруповання, доки не об’єднаються всі групи. Ця ієрархічна структура представлена на дендрограмі, що показує змінні в кожній групі та відстань між групами. Рівень, на якому об’єднання інтерпретуються як значущі категоризації, довільно вибирається дослідником, щоб

відобразити контекст. Симптоми які недостатньо представлені, можуть не групуватися з іншими симптомами, тому можуть бути визначені як викиди.

Отже, після проведення кластеризації рекомендується візуалізувати результати шляхом побудови дендрограми, яка дає можливість отримати уявлення про загальну конфігурацію об’єктів (рисунок 2.2.3) :

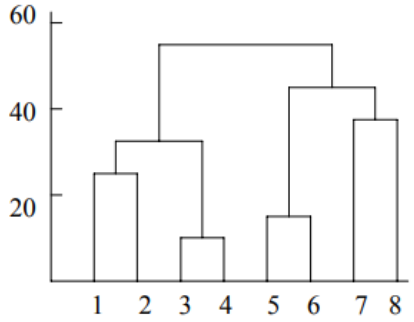


Рисунок 2.2.3 – Зображення дендрограми

Процес алгоритму ближнього сусіда.

Крок 0. Нормування даних.

Крок 1. Побудова матриці відстаней.

Крок 2. Вибір початкової пари, найближчої одна до одної, їх об’єднання в один кластер і побудова нової матриці відстаней.

Крок 3. Повторюємо дану операцію до тих пір, поки не будуть задіяні всі елементи. При цьому кожний вибір залишених елементів має здійснюватися за принципом найменшої відстані.

Крок 4. Побудова дендрограми.

Узагальнюючи результати дослідження, представляємо алгоритм процесу кластеризації методом ближнього сусіда (рисунок 2.2.4) :

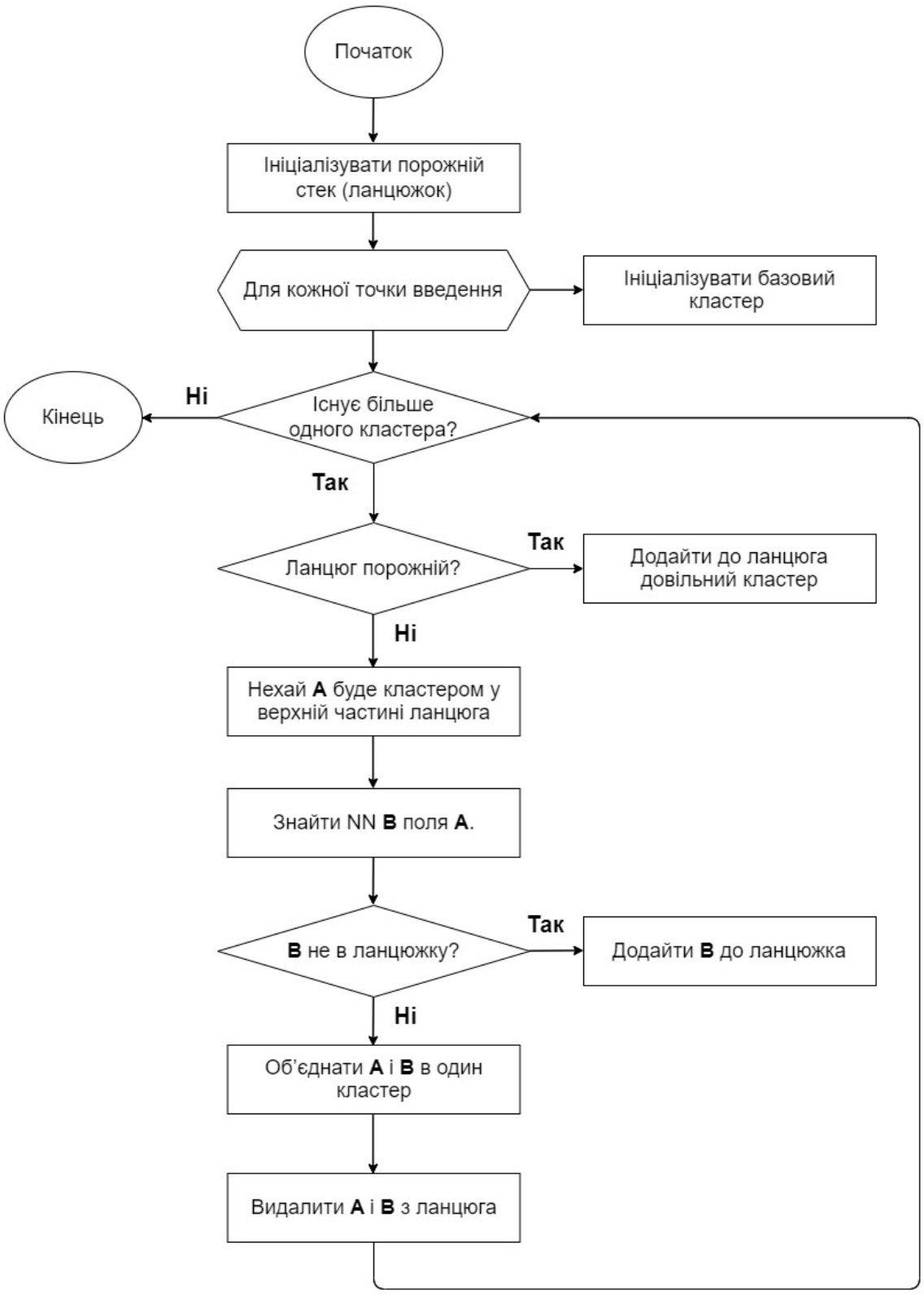


Рисунок 2.2.4 – Алгоритм процесу кластеризації методом ближнього сусіда

Основна ідея алгоритму найближчого сусіда полягає в підтримці стека, який називається ланцюгом кластерів. Перший кластер довільний. Ланцюг завжди

розширюється найближчим сусідом поточного кластера на вершині ланцюга. Отже, відстані між парами послідовних кластерів у ланцюжку постійно зменшуються. Таким чином, не може виникнути повторних кластерів, і ланцюг залишається ациклічним. Зрештою ланцюг досягає пари, скажімо, A і B. У цей момент A і B об'єднуються та видаляються з ланцюга. Важливо, що після злиття решта ланцюжка не відкидається. Нехай C — один із кластерів, які залишаються в ланцюжку, крім останнього. Через можливість зведення новий кластер A∪B, який виникає в результаті злиття A і B, не є найближчим сусідом для C. Це тому, що зведення стверджує, що новий кластер не ближчий до C, ніж один з A і B, і ні A та B були найближчими для C. Таким чином, після злиття ланцюжок все ще залишається ланцюгом найближчих сусідів: за кожним кластером, який залишився в ланцюзі, слідує його найближчий сусід, крім останнього, за яким ніхто не слідує. Процес продовжується з нової вершини ланцюга.

Перевага алгоритму найближчого сусіда, що відноситься до ієрархічного алгоритму кластеризації полягає у тому, що є наочний результат, простота реалізації, використовується до широкого кола сфер, являється пошуком найкращого рішень із можливих.

Недоліки найближчого сусіда: зберігає всю вибірку об'єктів, що провокує до витрат пам'яті, якщо серед об'єктів існує викид (тобто, об'єкт розташований всередині чужого класу), то всі об'єкти, які будуть знаходитись найближче до всіх інших, будуть класифікуватись неправильно.

2.3 Математичне моделювання експертної системи на основі неієрархічного ітеративного методу k-means

Кластеризація *k*-means середніх є широко використовуваним алгоритмом для багатьох практичних застосувань.

Термін «k-means» вперше використав Джеймс Маккуенін у 1967 році, хоча ідея сформувалася до 1957. Алгоритм *k*-means - це ітеративний метод, який мінімізує суму квадратів для заданої кількості кластерів.

Алгоритм процесу кластеризації на основі методу *k*-means [30 – 34]:

1. Виберіть k точок як початкові центроїди.
2. Повторіть.
3. З K кластерів, приписуючи кожній точці свій найближчий центроїд.
4. Повторно обчисліть центроїд кожного кластера.
5. Поки центроїди не зміняться k-means досягає стану, в якому немає точок перехід від одного кластера до іншого, напр. повторюючи до тих пір, поки лише 1% точок змінюють кластери.

Демонстрація алгоритму представлена на рисунку 2.3.1 :

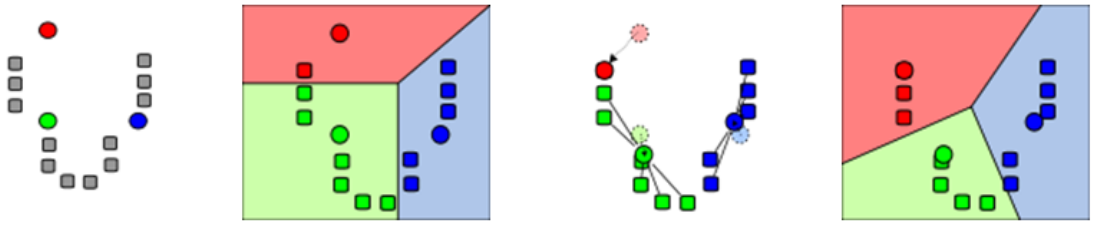


Рисунок 2.3.1 – Алгоритм процесу кластеризації на основі методу k – means

1. K початкових «середніх випадково згенеровано у межах даних (три кольорові крапки)/
2. Створено k кластерів, асоціюючи кожне спостереження з найближчим середнім.
3. Центроїд кожного з k кластерів стає новим середнім.
4. Кроки 2 і 3 повторюються до досягнення збіжності.

Узагальнюючи роботи з дослідження алгоритму k-means [31 – 33; 40; 41] представляємо процес виконання даного алгоритму у вигляді блок-схеми.

Алгоритм знаходження початкових центроїдів.

Вхідні дані: D = {d1, d2,.....,dn} // набір з n елементів даних

k // Кількість бажаних кластерів

Результат: набір із k початкових центроїдів.

Кроки:

Крок 0. Встановіть m = 1;

Крок 1. Обчисліть відстань між кожною точкою даних і всіма інші точки даних у наборі D;

Крок 2. Знайдіть найближчу пару точок даних із множини D і сформуйте набір точок даних Am (1<= m <= k), який містить ці дві точки даних; видаліть ці дві точки даних із множина D;

Крок 3. Знайдіть точку даних у D, яка є найближчою до набору точок даних Am, додайте її до Am і видаліть з D;

Крок 4. Повторюйте крок 3, доки кількість точок даних в Am досягає 0,75*(n/k);

Крок 5. Якщо m<k, то m = m+1, знайдіть іншу пару точок даних від D, між якими відстань найкоротша, інша точка даних встановлює Am і видаляйте їх із D, йти до кроку 4;

Крок 6. Для кожного набору точок даних Am (1<=m<=k) знайдіть середнє арифметичне векторів точок даних в Am,

Ці елементи будуть початковими центроїдами.

Результат алгоритму представлено у блок-схемі (рисунок 2.3.2) :

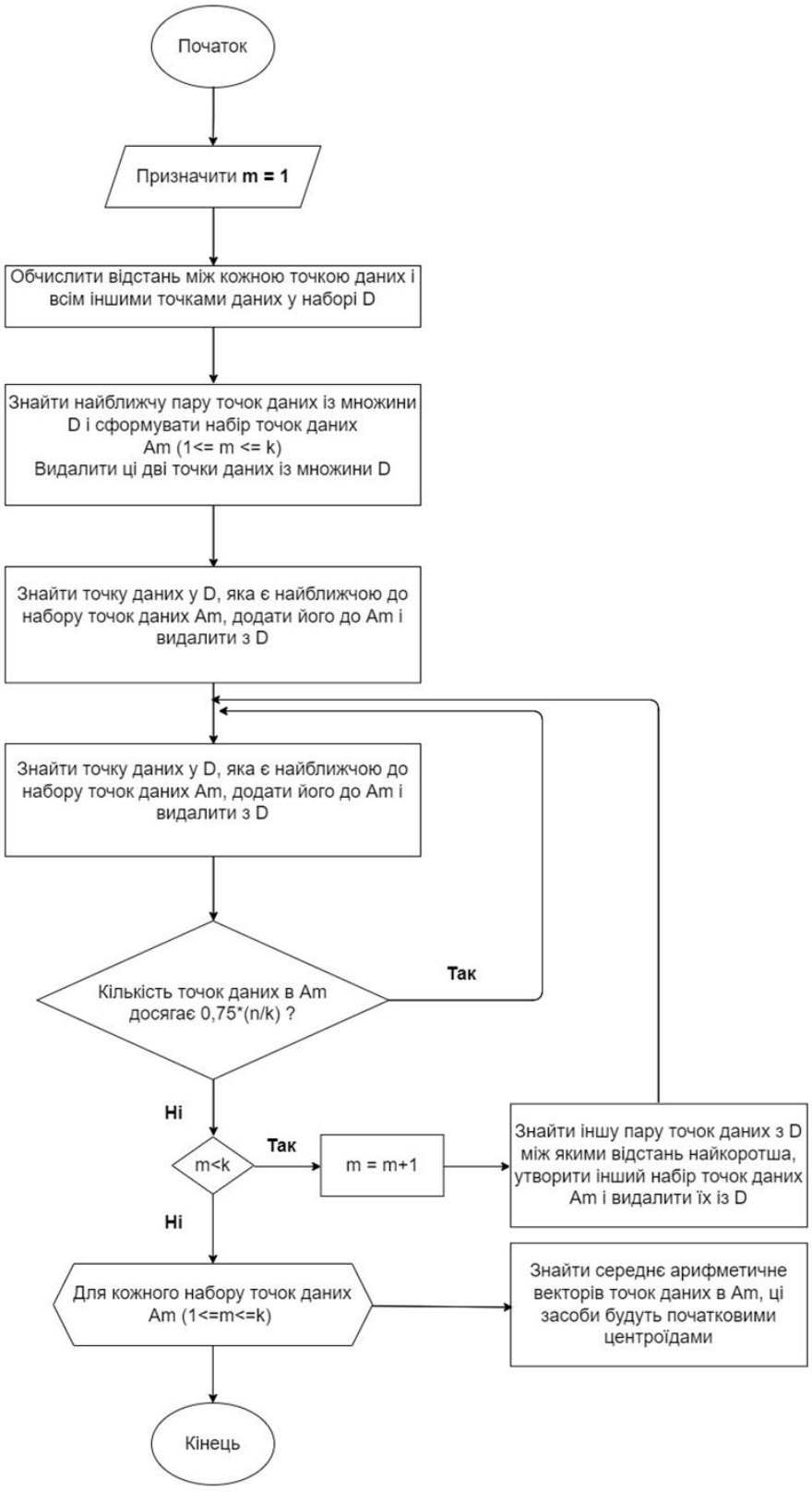


Рисунок 2.3.2 – Алгоритм кластеризації методом k-means

Вхідні дані:

D = {d1, d2,.....,dn} // набір із n точок даних.

C = {c1, c2,.....,ck} // набір з k центроїдів

Результат:

Набір з k кластерів

Кроки:

1. Обчисліть відстань від кожної точки даних di (1<=i<=n) до всіх центроїдів cj (1<=j<=k) як d(di, cj).
2. Для кожної точки даних di знайдіть найближчий центроїд cj і призначте di кластеру j.
3. Встановити ClusterId[i]=j; // j:Id найближчого кластера.
4. Встановити Nearest_Dist[i]= d(di, cj).
5. Для кожного кластера j (1<=j<=k) перерахуйте центроїди.
6. Повторіть.
7. Для кожної точки даних di:

   7.1. Обчисліть його відстань від центроїда поточного найближчого кластера.

   7.2. Якщо ця відстань менша або дорівнює поточній найближчій відстані, точка даних залишається в кластері.

   Інакше

   7.2.1. Для кожного центроїда cj (1<=j<=k). Обчисліть відстань d(di, cj).

   7.2.2. Призначте точку даних di кластеру з найближчий центроїд cj.

   7.2.3. Встановити ClusterId[i]=j.

   7.2.4. Встановити Nearest_Dist[i]= d(di, cj).

8. Для кожного кластера j (1<=j<=k) перерахуйте центроїди. Поки не буде виконано критерій конвергенції.

Даний алгоритм представлено у вигляді блок-схеми (рисунок 2.3.3) :

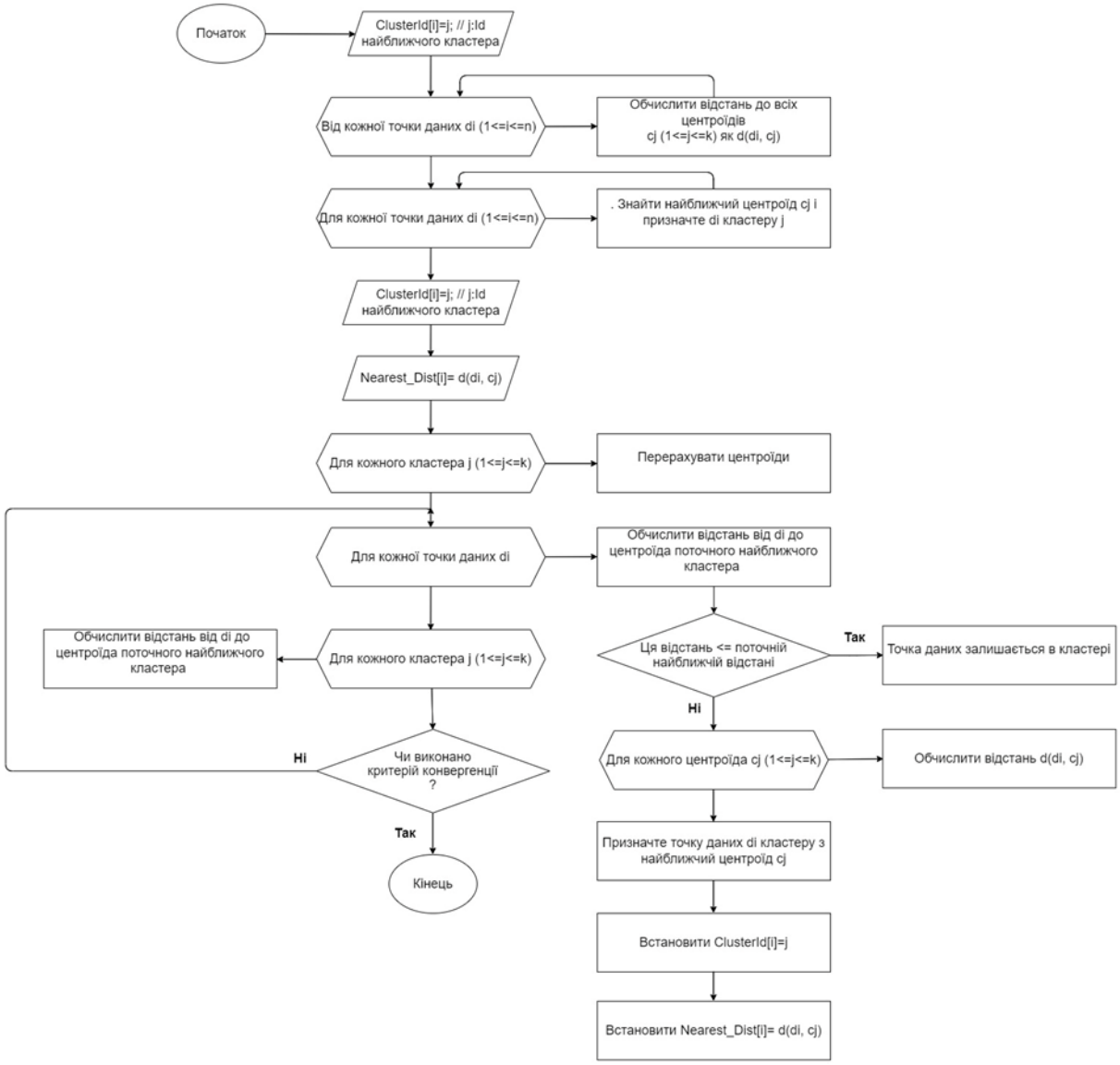


Рисунок 2.3.3 – Блок-схема алгоритму кластеризації методом k-means

Аналіз досліджень з теми Кластерний аналіз дозволив виділити переваги та недоліки алгоритмів k-means та найближчого сусіда зображено на рисунку 2.3.4 :

	Метод ближнього сусіда	Метод k середніх
Переваги	1. Алгоритм простий та легко реалізується. 2. Нечутливий до викидів. 3. Не потрібно будувати модель. 4. Універсальний, оскільки використовується для завдань класифікації та регресії.	1. Гнучкість, швидкість та простота використання. 2. Зрозумілий. 3. Перевірка статистичної значимості відмінностей між виділеними кластерами.
Недоліки	1. Алгоритм працює повільніше, якщо збільшити обсяг вибірки, предикторів чи незалежних змінних. 2. Обчислювальні витрати під час виконання та обробки великих даних. 3. Не створює жодних моделей або правил, які узагальнюють попередній досвід.	1. Результати кластеризації залежать від вибору початкової конфігурації центроїдів. 2. Заздалегідь визначити кількість кластерів. 3. Повільний при кластеризації велику обсягу даних. 4. Чутливий до викидів.

Рисунок 2.3.4 – Переваги та недоліки алгоритмів k середніх та найближчого сусіда

Визнання того, що хворі на рак зазвичай відчувають численні симптоми, пов’язані з хворобою та лікуванням, спонукало до дослідження кластерів симптомів.

Коли будується алгоритм ієрархічного алгоритму, у даному випадку – алгоритм найближчого сусіда, потрібно визначити вибірку даних для визначення кількості кластерів. Параметри, що використовувались для алгоритму найближчого сусіда :

1. Вік – оскільки, це індивідуальний параметр, що впливає на якому етапі знаходиться розвиток ракових пухлин.
2. Кількість разів вагітності впливає на репродуктивну систему жінки.
3. Те, чи людина палить вливає на розвиток ракових пухлин.
4. Якщо людина вживає гормональні контрацептиви – кількість позитивних та негативних показників визначає норму гормонів в організмі.

5. Те, скільки років людина вживає гормональні контрацептиви впливає на стабілізацію цих гормонів до, після та під час вживання.
6. Внутрішньоматкова спіраль впливає на те, які ризики було враховано під час та після носіння її як контрацептиву.
7. Скільки років власник носив внутрішньоматкову спіраль впливає на розвиток хвороб та ризик захворювань раку.
8. Захворювання, що передаються статевим шляхом впливають на весь організм, показники крові та гормони, оскільки сильно вражають здорові клітини за рахунок запалення.
9. Кількість захворювань статевим шляхом впливає на якість репродуктивної системи, самопочуття.
10. Оскільки, кондиломатоз – це захворювання, що викликане статевим шляхом через вірус папіломи людини, потрібно дослідити коли хвороба з’явилась та що встигла уразити, як визначити лікування.
11. Кондиломатоз шийки матки – це захворювання, що включено до факторів розвитку ризику раку шийки матки.
12. Генітальні бородавки – інфекція, що передається статевим шляхом.
13. Інфекція вірусу вульвопромежини папіломи людини.
14. Сифіліс передається статевим шляхом та впливає на якість роботи всього організму.
15. Загальне захворювання органів малого тазу впливає на розвиток раку, інфекцій та інших статевих захворювань.
16. Генітальний герпес.
17. Контагіозний молюск захворювання, що передається при контакті шкіра до шкіри.
18. Синдром набутого імунодефіциту (СНІД) впливає на весь організм та має погані наслідки для людини, якщо не лікуватись.
19. Гепатит В – це інфекція печінки та може спричинити рак для печінки.
20. Вірус папіломи людини має поширену групу вірусів. Деякі з них можуть викликати рак.

21. Зразки сечі дозволяють побачити всі показники відхилень та норми.
22. Час з першого дослідження аналізу вказує на історію хвороби.
23. Час, що прийшов з останньої діагностики вказує на історію хвороби.
24. Ступінь раку вказує на що звернути увагу чи розробити, змінити план лікування.
25. Аномальні клітини виявляються на поверхні шийки матки. Діагностика ураження робиться для визначення степені раку.
26. Скринінговий тест на наявність раку шийки матки вказує на те, який висновок розробили дослідники.
27. Колькоскопія – вид медичного тестування, який робиться щоразу при відвідуванні лікаря.
28. Цитологія – дослідження типу клітин, що знаходяться у рідинах.
29. Біопсія передбачає взяття зразка клітини на наявність відхилення.

Оскільки, всі показники можуть вказувати на різні захворювання ступінь ураження, машинне навчання допоможе визначити на які показники потрібно звернути увагу в першу чергу при консультації лікарем.

Вигляд дендограми для кластерного аналізу, де використовувались дані вище зображено на рисунку 2.3.5 :

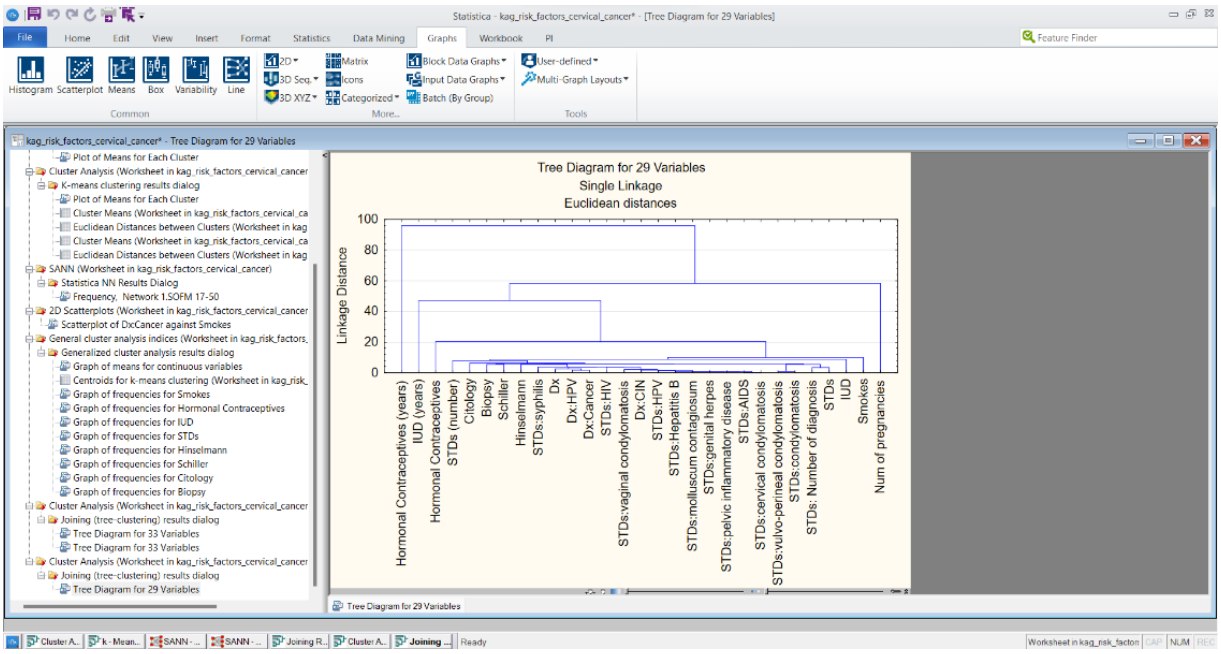


Рисунок 2.3.5 – Дендограма для кластерного аналізу

Для неієрархічного алгоритму кластеризації – у даному випадку алгоритм k середніх, також потрібно обрати вибірку даних для обробки.

Вигляд дендограми для алгоритму k середніх, де використовувалась визначна вибірка даних зображено на рисунку 2.3.6 :

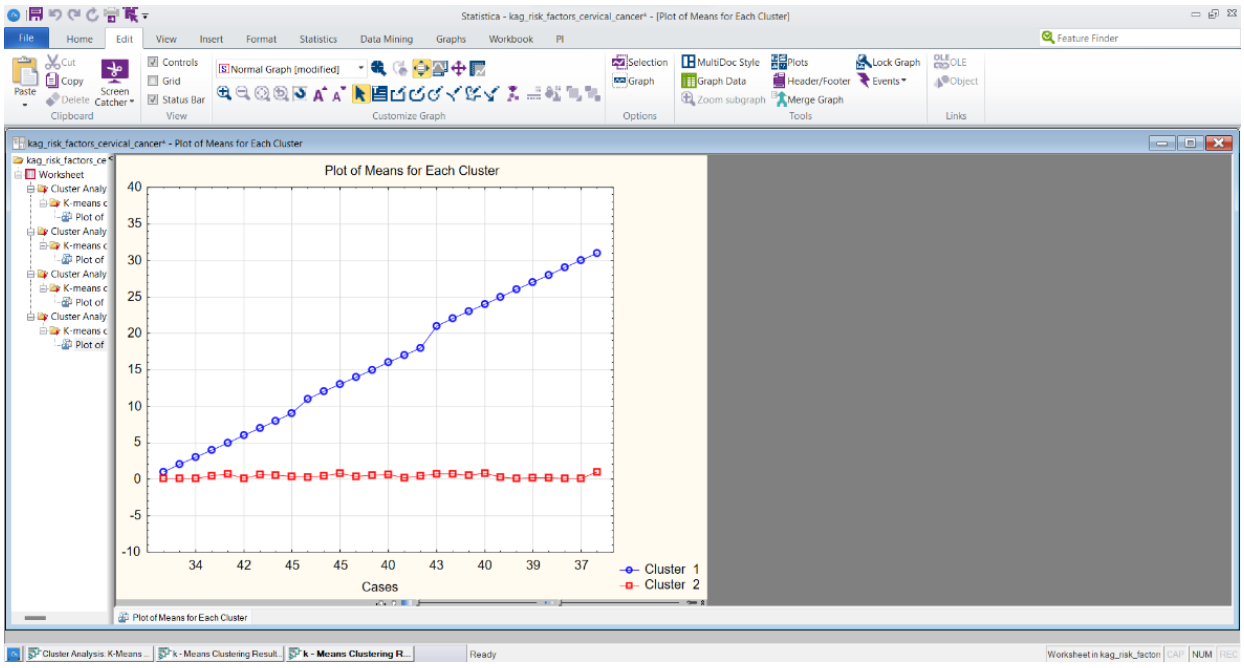


Рисунок 2.3.6 – Дендограма для алгоритму k середніх

3 ЗАБЕЗПЕЧЕННЯ ЗАХИСТУ МЕДИЧНИХ ДАНИХ У ЕКСПЕРТНИХ СИСТЕМАХ

3.1 Об’єкти захисту у медичній експертній системі

Штучний інтелект у медичній галузі має значний потенціал для покращення результатів лікування пацієнтів, отримання цінної інформації з метою запобігання епідеміям та хворобам та розв’язання інших клінічних проблем. Проте є очевидним, що використовувати інформаційні технології можна лише за умови вирішення питань безпеки та конфіденційності. Наукові розробки, зокрема [42 - 51] підтверджують тезу щодо забезпечення захисту інформації у медичних системах. Медична інформація належить до конфіденційної і є об’єктом захисту в Україні на законодавчому рівні у відповідності до Закону «Про захист персональних даних» від 01.06.2010р., № 2297-VI.

Як відомо, у медичних інформаційних системах об’єктами захисту є:

1. Інформація в базах даних систем керування базами даних.
2. Ресурси файлового сервера лікувально-профілактичного закладу.
3. Резервні копії бази даних систем керування базами даних і архівні копії ресурсів файлового сервера.
4. Керуюча інформація операційної системи, системи керування базами даних, автоматизоване робоче місце адміністратора медичної інформаційної системи та адміністратора інформаційної безпеки.
5. Технологічний процес збору, обробки, зберігання та передачі інформації в медичній інформаційній системі.
6. Апаратно-програмний комплекс, що забезпечує роботу медичної інформаційної системи.

Засоби та методи безпеки у медичній системі мають бути спрямовані на наступні аспекти:

1. Забезпечення цілісності даних: захист від несанкціонованих маніпуляцій.
2. Забезпечення конфіденційності: захист від несанкціонованого доступу.
3. Захист автентичності та невідмовності.

3.2 Засоби і методи забезпечення захисту інформації у медичній експертній системі

Не існує незалежних основ і систем загальної інформаційної безпеки в медицині.

Авторизація — це процес надання комусь доступу до ресурсу. Хорошим прикладом є власність на будинок. Власник має повні права доступу до власності (ресурсу), але може надати право доступу іншим людям. Власник надає людям доступ до нього. Наприклад, доступ до будинку – це дозвіл, тобто дія, яку можна виконати на ресурсі. Інші дозволи на будинок можуть бути меблями, прибиранням, ремонтом тощо. Іноді авторизація певною мірою пов’язана з ідентифікацією. У літаку є посадковий талон, у якому зазначено, що існує право літати цим літаком. Однак цього недостатньо, щоб агент з воріт дозволив потрапити на борт. Також потрібен паспорт із зазначенням особи. У цьому випадку агент з виходу на посадку порівнює ім’я в паспорті з іменем у посадковому талоні та пропускає людину, якщо вони збігаються. У контексті авторизації ім’я є атрибутом особистості. Іншими атрибутами є вік, мова, кредитна картка та будь-що інше, що стосується певного сценарію. Людина, яка читає ваше ім’я у вашому паспорті, може бути впевнена у вашому імені, тому що вона довіряє уряду, який видав ваш паспорт. Посадковий талон разом із документом, що посвідчує особу споживача, є своєрідним «токеном доступу», який надає право доступу на посадку в літак. У сценаріях, описаних вище, можна побачити, що акт авторизації дозволяє об’єктам виконувати завдання, які іншим об’єктам заборонено виконувати. Комп’ютерні системи, які використовують авторизацію, працюють подібним чином.

Багатофакторна автентифікація (MFA) — це метод автентифікації, який

вимагає від користувача надати два або більше факторів перевірки, щоб отримати доступ до ресурсу, наприклад програми, онлайн-облікового запису або VPN. MFA є основним компонентом надійної політики керування ідентифікацією та доступом (IAM). Замість простого запиту імені користувача та пароля MFA вимагає один або кілька додаткових факторів перевірки, що зменшує ймовірність успішної кібератаки. Види автентифікації : ресурс автентифікації певного фактора, наприклад, сканування відбитків пальців. Отже, якщо перша автентифікація вдається, вхід повинен бути за допомогою іншого ресурсу автентифікації того самого фактора, наприклад, сканування обличчя або оболонки ока. Іншим прикладом багатоетапної автентифікації є схема з двома паролями, прийнята деякими банками по всьому світу для виконання онлайн-операцій. У цьому випадку спочатку довгий і складніший пароль для доступу до онлайн-системи банку. Крім того, крім довгого пароля, банк також надає короткий і цифровий пароль для підтвердження кожної операції, виконаної користувачем в системі.

Різниця між багаторівневою та багатофакторною автентифікацією полягає у тому, що у багаторівневій – користувач повинен створити дві або більше форми ідентифікації. Наприклад, на державному сайті студентських позик, якщо пароль було забуто, буде запропоновано три незрозумілі питання безпеки, які було обрано. Оскільки, все це підпадає під щось, що користувач знає», у багатофакторній – стосується використання кількох форм автентифікації, таких як пароль і сканування сітківки ока. Існує два різні фактори, які використовуються для автентифікації. Якщо хакер вкраде пароль, для отримання доступу все одно потрібна зовсім інша форма автентифікації (сканування сітківки ока). Якщо використовуються лише два фактори, називається це двофакторною автентифікацією (2FA). Однак технічно багатофакторна означає два або більше факторів, тому люди часто використовують терміни багатофакторна автентифікація та двофакторна автентифікація як синоніми.

У випадку з Face ID і паролем від Apple є два типи факторів. Те, що є у користувача (обличчя), і те, що користувач знає (пароль). Таким чином, ця форма автентифікації технічно є 2FA, але називати її MFA також буде доцільно.

Також можна використовувати систему захист брутфорс (Brute Force) – це пробивання захисту, для прикладу, телефону декілька разів. Тобто, у програму можна внести кількість спроб для вводу паролю, що ні зловмисник, ні інша людина не змогла ввійти в систему.

Прості атаки брутфорсу – хакери намагаються логічно вгадати облікові дані — абсолютно без допомоги програмних засобів чи інших засобів. Вони можуть виявити надзвичайно прості паролі та PIN-коди. Наприклад, пароль guest12345. Атаки за словником: у стандартній атаці хакер вибирає ціль і запускає можливі паролі для цього імені користувача. Вони відомі як атаки за словником. Атаки за словником є основним інструментом атак грубою силою. Хоча самі по собі не обов'язково є атаками грубої сили, вони часто використовуються як важливий компонент для злому паролів. Деякі хакери переглядають нескорочені словники та доповнюють слова спеціальними символами та цифрами або використовують спеціальні словники слів, але цей тип послідовної атаки є громіздким. Гібридні атаки грубої сили: ці хакери поєднують зовнішні засоби зі своїми логічними припущеннями, щоб спробувати зламати систему. Гібридна атака зазвичай поєднує атаки за словником і грубою силою. Ці атаки використовуються для визначення комбінованих паролів, які змішують загальні слова з випадковими символами. Приклад грубої атаки такого характеру включатиме такі паролі, як NewYork1993 або Spike1234. Зворотні атаки підбору: як випливає з назви, зворотна атака підбору скасовує стратегію атаки, починаючи з відомого пароля. Потім хакери шукають мільйони імен користувачів, поки не знайдуть збіг. Наповнення обліковими даними: якщо хакер має комбінацію імені користувача та пароля, яка працює на одному веб-сайті, він також спробує її на багатьох інших. Оскільки відомо, що користувачі повторно використовують дані для входу на багатьох веб-сайтах, вони є винятковими цілями такої атаки.

На рисунку 3.2.1 узагальнено системи ідентифікації та автентифікації у експертних системах :

Рисунок 3.2.1 – Системи ідентифікації та автентифікації у експертних системах

Шифрування — це спосіб зробити дані нечитабельними, забезпечуючи доступ до цих даних лише авторизованій особі. Шифрування використовує складні алгоритми для шифрування даних і розшифровує ці дані за допомогою ключа, наданого відправником повідомлення. Шифрування гарантує, що інформація залишається приватною та конфіденційною, незалежно від того, зберігається вона чи передається. Терміни шифрування:

1. Алгоритм – це правила або інструкції для процесу шифрування. Довжина ключа, функціональність і особливості використовуваної системи шифрування визначають ефективність шифрування.
2. Дешифрування - це процес перетворення нерозбірливого зашифрованого тексту в читабельну інформацію.
3. Ключ шифрування – кожен ключ унікальний, а довші ключі важче зламати.

Існує два види систем криптографічних ключів: симетричні та асиметричні:

1. У системі симетричного ключа кожен, хто отримує доступ до даних, має

однаковий ключ. Ключі, які шифрують і розшифровують повідомлення, також повинні залишатися секретними для забезпечення конфіденційності.

2. Асиметрична система ключів, також відома як система відкритих/приватних ключів, використовує два ключі. Один ключ залишається в таємниці — приватний ключ — тоді як інший ключ стає широко доступним для всіх, хто його потребує. Цей ключ називається відкритим ключем. Приватний і відкритий ключі математично пов'язані між собою, тому відповідний закритий ключ може розшифрувати лише ту інформацію, зашифровану за допомогою відкритого ключа.

Загальні алгоритми шифрування :

1. Потрійний DES був розроблений, щоб замінити оригінальний алгоритм стандарту шифрування даних (DES), який хакери врешті-решт навчилися перемагати з відносною легкістю. Потрійний DES використовує три окремі ключі по 56 біт кожен. Загальна довжина ключа становить 168 біт.
2. Розширений стандарт шифрування (AES) — це алгоритм, якому уряд США та численні організації довіряють як стандарт. Незважаючи на високу ефективність у 128-бітній формі, AES також використовує ключі 192 та 256 бітів для важкого шифрування. AES переважно вважається несприйнятливим до всіх атак, за винятком грубої сили, яка намагається розшифрувати повідомлення за допомогою всіх можливих комбінацій у 128, 192 або 256-бітному шифрі.
3. RSA — це алгоритм шифрування з відкритим ключем і стандарт для шифрування даних, що надсилаються через Інтернет. Це також один із методів, який використовується в програмах PGP і GPG. На відміну від Triple DES, RSA вважається асиметричним алгоритмом через використання пари ключів.
4. Blowfish — ще один алгоритм, призначений для заміни DES. Цей симетричний шифр розбиває повідомлення на блоки по 64 біти та шифрує їх

окремо. Blowfish відомий своєю надзвичайною швидкістю та загальною ефективністю. Тим часом постачальники повною мірою скористалися його безкоштовною доступністю у відкритому доступі.

Оновлення програмного забезпечення (також відоме як патч) — це набір змін до програмного забезпечення для його оновлення, виправлення або покращення. Зміни в програмному забезпеченні зазвичай виправляють помилки, усувають уразливості безпеки, надають нові функції або покращують продуктивність і зручність використання. Нечасто патчі також можуть використовуватися для обмеження функціональності, видалення або вимкнення функцій. Залежно від програмного забезпечення оновлення можна встановлювати вручну або автоматично, якщо пристрій підключено до Інтернету та має відповідні можливості. Оновлення програмного забезпечення особливо важливі, коли застосовуються до операційної системи, оскільки інше програмне забезпечення (наприклад, програми чи драйвери) залежить від неї. Наприклад, основний випуск операційної системи, такої як Android або iOS, може зробити ряд програм застарілими, якщо всі версії, випущені після оновлення, несумісні з попередньою версією ОС. З точки зору безпеки, оновлення програмного забезпечення мають важливі наслідки. Коли оновлення включає виправлення вразливостей безпеки, будь-який пристрій, на якому працює застаріла версія програмного забезпечення, стає особливо вразливим. Це дозволяє зловмисникам знати, які вразливості існують у даній системі, і, отже, піддає більшому ризику пристрої, на яких запущено цю програму (версію). Наприклад, використання застарілої версії Android означає, що всі вразливості безпеки, помічені та виправлені в наступних версіях, все ще існують на будь-якому пристрої. Відсутність оновлення програмного забезпечення також може негативно вплинути на функціональні можливості пристрою, наприклад, зробити деякі його функції. Це також може означати, що виявлені помилки та проблеми можуть ніколи не бути виправлені (наприклад, поганий акумулятор). Сучасна ринкова практика не вимагає мінімальної підтримки програмного забезпечення для пристрою або версії програмного забезпечення під час випуску,

тобто пристрій можна виробляти, випускати та продавати з вбудованою застарілою операційною системою або без регулярних оновлень програмного забезпечення. Це принципово дозволяє виробникам продавати пристрої, які можуть стати застарілими та вразливими протягом кількох місяців після випуску. Це регулярна практика, яка ставить під загрозу безпеку та конфіденційність користувачів. Також оновлення будь – яких браузерів, програмного забезпечення, мобільних застосунків впливає на тестування продукту та використання при поточних умовах.

Дуже часто при релізі нових версій продукту важливо мати бекап файли та можливість відновлення, оскільки під час з'єднування всіх частин коду, переносу даних у бази даних іншого середовище може щось піти не так, тому можливість повернутись на минулу версію відіграє велику роль для того, щоб попрацювати над помилками.

Основні збої даних можуть бути наслідком збою апаратного чи програмного забезпечення, пошкодження даних або події, спричиненої людиною, як-от зловмисна атака (вірус або зловмисне програмне забезпечення) або випадкове видалення даних. Резервні копії дозволяють відновити дані з попереднього моменту, щоб допомогти бізнесу відновитися після незапланованої події. Зберігання копії даних на окремому носії є критичним для захисту від втрати або пошкодження первинних даних. Цей додатковий носій може бути таким простим, як зовнішній диск чи USB-накопичувач, або чимось більш значним, таким як дискова система зберігання, хмарний контейнер для зберігання чи стрічковий накопичувач. Альтернативний носій може бути в тому самому місці, що й основні дані, або у віддаленому місці. Можливість погодних явищ може виправдати наявність копій даних у віддалених місцях. Для досягнення найкращих результатів резервні копії створюються на послідовній регулярній основі, щоб мінімізувати кількість даних, втрачених між резервними копіями. Що більше часу проходить між резервними копіями, то більша ймовірність втрати даних під час відновлення з резервної копії. Зберігання кількох копій даних забезпечує страхування та гнучкість для відновлення до моменту часу, на який не вплинуло пошкодження даних або зловмисні атаки.

Параметри резервного копіювання даних :

1. Простим варіантом є створення резервних копій файлів на знімних носіях, таких як компакт-диски, DVD-диски або USB-накопичувачі. Це може бути практичним для менших середовищ, але для великих обсягів даних потрібно буде створювати резервні копії на кількох дисках, що може ускладнити відновлення.
2. Можна налаштувати додатковий жорсткий диск, який є копією диска чутливої системи в певний момент часу, або всю резервну систему. Наприклад, інший сервер електронної пошти, який знаходиться в режимі очікування, створюючи резервну копію основного сервера електронної пошти. Резервування є потужною технікою, але нею складно керувати.
3. Можна розгорнути зовнішній жорсткий диск великого обсягу у своїй мережі та використовувати програмне забезпечення для архівування, щоб зберегти зміни локальних файлів на цьому жорсткому диску.
4. Багато постачальників надають повні пристрої для резервного копіювання, які зазвичай розгортаються як 19-дюймові пристрої для встановлення в стійку. Пристрої для резервного копіювання мають великий обсяг пам’яті та попередньо інтегроване програмне забезпечення для резервного копіювання.
5. Програмні рішення для резервного копіювання складніші для розгортання та налаштування, ніж апаратні пристрої, але пропонують більшу гнучкість.
6. Багато постачальників і хмарних провайдерів пропонують рішення Backup as a Service (BaaS), за допомогою яких можна надсилати локальні дані в загальнодоступну або приватну хмару, а в разі аварії відновлювати дані з хмари. Рішення BaaS прості у використанні та мають велику перевагу в тому, що дані зберігаються у віддаленому місці.

Цілісність даних означає точність і послідовність (валідність) даних протягом їх життєвого циклу. Зрештою, скомпрометовані дані мало корисні для підприємств, не кажучи вже про небезпеку, яку становить втрата конфіденційних

даних. З цієї причини підтримка цілісності даних є основною метою багатьох корпоративних рішень безпеки. Термін цілісність даних також призводить до плутанини, оскільки він може стосуватися або стану, або процесу. Цілісність даних як стан визначає набір даних, який є дійсним і точним. З іншого боку, цілісність даних як процес описує заходи, які використовуються для забезпечення дійсності та точності набору даних або всіх даних, що містяться в базі даних чи іншій конструкції. Наприклад, методи перевірки помилок і підтвердження можуть називатися процесами цілісності даних.

Підтримка цілісності даних важлива з кількох причин. По-перше, цілісність даних забезпечує можливість відновлення та пошуку, відстеження (до джерела) та підключення. Захист достовірності та точності даних також підвищує стабільність і продуктивність, одночасно покращуючи можливість повторного використання та обслуговування. Цілісність даних може бути порушена різними способами, що робить методи забезпечення цілісності даних важливим компонентом ефективних протоколів безпеки підприємства. Цілісність даних може бути порушена через:

1. Помилки передачі, включаючи ненавмисні зміни або компрометацію даних під час передачі з одного пристрою на інший.
2. Помилки, віруси/зловмисне програмне забезпечення, хакерство та інші кіберзагрози.
3. Пошкоджене апаратне забезпечення, наприклад збій пристрою або диска.
4. Фізичний компроміс пристроїв.

Оскільки лише деякі з цих компромісів можна належним чином запобігти за допомогою захисту даних, резервне копіювання та дублювання даних стає критичним для забезпечення цілісності даних.

Таким чином, у медичній інформаційній системі реалізовано ряд заходів безпеки, які мають проводитися системно на всіх етапах її діяльності: від проектування і розробки до впровадження і експлуатації, перекривати всі відомі загрози безпеки, орієнтовані на тактичне випередження загроз, при цьому повинні

відповідати нормам законодавства і відомчим актам системи охорони здоров'я.

Узагальнення представлено на рисунку 3.2.2 :

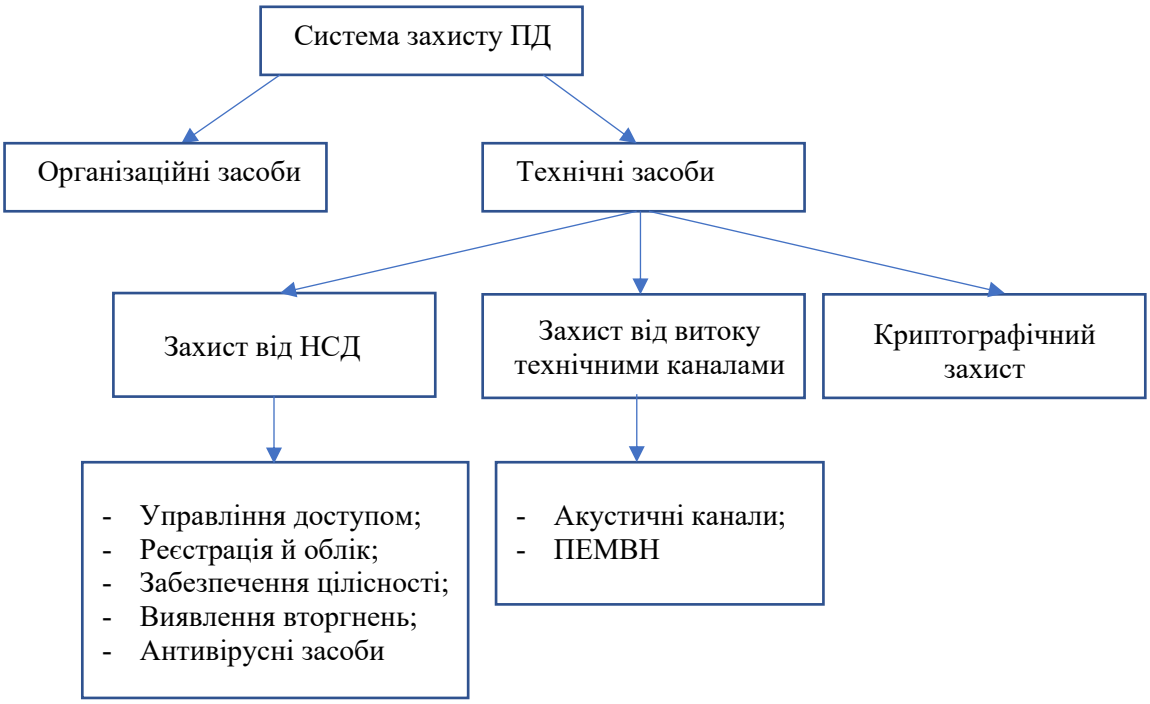


Рисунок 3.2.2 – Безпека медично системи

3.3 Моделювання експертної системи з врахуванням блоку інформаційної безпеки

У попередніх розділах представлено алгоритми виконання кластеризації методом ближнього сусіда та методом k-means для медичного скринінгу пацієнтів. Тому на всіх етапах життєвого циклу медичної інформації пацієнта необхідно забезпечити зберігання даних, цілісність даних і контроль доступу до них. Представляємо модель і алгоритм роботи частини експертної системи щодо забезпечення захисту персональних даних (рисунок 3.3.1) :

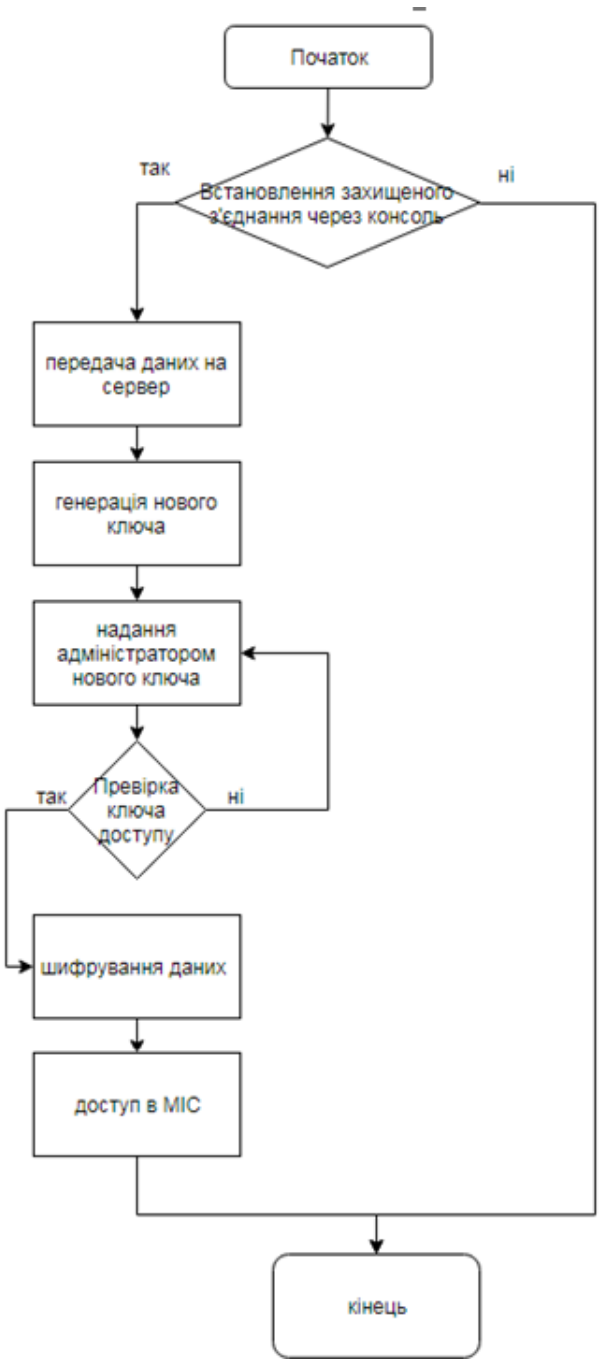


Рисунок 3.3.1 – Блок-схема захисту даних для експертної системи

IT – аудит або аудит інформаційних технологій — це дослідження та оцінка IT – систем, інфраструктури, політики та операцій. За допомогою IT – аудитів компанія може визначити, чи існуючі засоби IT – контролю захищають корпоративні активи, забезпечують цілісність даних і відповідають системі контролю за бізнесом і фінансами організації. Як неупереджений спостерігач IT – аудитор переконується, що ці засоби контролю встановлено належним чином і

ефективно, тому компанія менш вразлива до витоку даних та інших ризиків безпеки. IT – аудитор розробляє, впроваджує, тестує та оцінює всі процедури перевірки IT – аудиту в компанії, яка покладається на технології. Ці процедури аудиту можуть поширюватися на мережі, програмні додатки, системи зв'язку та безпеки, а також будь – які інші системи, які є частиною технологічної інфраструктури організації.

Обов'язки IT – аудитора :

1. Розробка та планування планів тестування аудиту.
2. Визначення обсягу та цілей аудиту.
3. Координація та виконання аудиторської діяльності.
4. Ведення та оновлення документації IT – аудиту.
5. Повідомлення результатів аудиту та рекомендацій.
6. Забезпечення виконання попередніх рекомендацій.

Навички IT – аудитора:

1. Формальна кваліфікація: вона може не вимагатися в усіх компаніях, але може допомогти IT – аудиторам застосовувати систематичний підхід до своєї роботи.
2. Практичний досвід: попередній досвід роботи в галузі безпеки даних та IT – аудиту завжди є плюсом.
3. Розуміння основних бізнес-процесів: це допомагає IT – аудитору пов'язувати IT – системи з цінністю, яку вони приносять для бізнесу.
4. Розуміння ключових IT – процесів – це дозволяє IT – аудитору визначити пріоритетність IT – ризиків.
5. Сильна аналітична та логічна здатність міркувати: IT – аудитори повинні вміти використовувати аналіз даних та інструменти візуалізації.
6. Сильні комунікативні навички: ця здатність необхідна для пояснення складних питань безпеки нетехнічним командам управління.

ВИСНОВКИ

Методи штучного інтелекту мають великий потенціал для застосування майже в кожній галузі медицини: діагностика, лікування та прогнозування результатів у багатьох клінічних ситуаціях залежать від складної взаємодії багатьох клінічних, біологічних і патологічних змінних, тому зростає потреба у потужних алгоритмах машинного навчання, які можуть використовувати складні взаємозв'язки між цими змінними. Тому моделювання та розробка експертних систем є актуальним та важливим дослідженням.

У дослідженні здійснено аналіз експертних систем, основою яких слугують алгоритми машинного навчання. Експертні системи на практиці використовують контрольовані алгоритми машинного навчання. Застосування неконтрольованих алгоритмів машинного навчання у медичних дослідженнях є частковим, найчастіше у теоретичних розробках. Виявлення кластерів симптомів дозволяє визначити пріоритети симптомів для оцінки та розвитку нових напрямків у лікуванні цих симптомів.

Аналіз методів кластерного аналізу дозволив визначити їх впровадження у медичну експертну систему. Була спроба застосувати метод ближнього сусіда та метод k-means у моделюванні експертних систем та здійснити їх порівняльний аналіз. Визначено кроки у кожному алгоритмі і представлено обґрунтування. Доведено, що метод ближнього сусіда застосовувати краще для невеликої вибірки даних, а метод k-means є потужним алгоритмом лише, коли внести значення k.

Разом із застосуванням інформаційно-комунікаційних технологій у медичних системах виникають різні ризики та загрози інформаційній безпеці, що спонукало до моделювання процесів захисту персональних даних у експертних системах.

Результати дослідження можуть бути використані для розробки програмного забезпечення експертної системи для медичного скринінгу та впроваджені у навчальний процес медичних закладів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Куцук В.А. Модель експертної системи для медичного скринінгу на основі методів кластерного аналізу / Шевченко С.М., Жданова Ю.Д., Негоденко О.В., Куцук В.А. // Moderní aspekty vědy: XXVII. Díl mezinárodní kolektivní monografie / Mezinárodní Ekonomický Institut s.r.o.. Česká republika: Mezinárodní Ekonomický Institut s.r.o., 2023. – С. 478 – 494.
2. Куцук В.А. Експертна система для медичного скринінгу на основі методів кластерного аналізу // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» – Київ: ДУТ, 2022.
3. Jarratano D., Riley G. Expert systems. Principles of programming development, Ed. Williams. 2006. P. 2104.
4. Палагін О.В., Петренко М.Г. Тлумачний онтографічний словник з інженерії знань. Київ: ТОВ «НВП Інтерсервіс», 2017. 478 с.
5. Шевчук І.Б. Тлумачний словник основних понять і термінів програмування / І. Б. Шевчук. – ЛДФА, Львів: Видавництво ВТЗНВ, 2013. 45 с. URL: https://financi al.lnu.edu.ua/wp-ntent/uploads/2015.
6. Експертні системи в медицині: Навчальний посібник/Продеус А.М., Синєкоп Ю.С., Швець Є.Я., Кісельов Є.М., Баран М.М. Запоріжжя: Видавництво ЗДІА, 2014. 332 с.
7. Вступ до експертних систем: Навч. посіб. / Кравець В.О., Хавіна І.П. та ін. Харків: НТУ “ХПІ”, 2006. 232 с.
8. Інтелектуальні системи управління: Експертні системи – основи проектування та застосування в системах автоматизації [Електронний ресурс] : навч. посіб. для студ. спеціальності 151 «Автоматизація та комп’ютерно-інтегровані технології» / КПІ ім. Ігоря Сікорського; уклад.: Л. Д. Ярощук. – Електронні текстові дані (1 файл: 2,56 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2019. – 136с.
9. Paixão GMM, Santos BC, Araujo RM, Ribeiro MH, Moraes JL, Ribeiro AL. Machine Learning in Medicine: Review and Applicability. Arq Bras Cardiol. 2022

Jan;118(1):95-102. English, Portuguese. doi: 10.36660/abc.20200596. PMID: 35195215; PMCID: PMC8959062.

10. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380(14):1347-58. doi: 10.1056/NEJMra1814259.
11. Калініна І. О., Гожий О. П. Дослідження ефективності методів класифікації при прогнозуванні в задачах машинного навчання. Управління розвитком складних систем. Київ, 2021. № 46. С. 173 – 180.
12. Alonso-Rodríguez, A. (2001). Logistic regression and world income distribution. International Advances in Economic Research, 7(2), 231-242.
13. Data Mining: A Heuristic Approach / [H. A. Abbass, R. A. Sarker, C. S. Newton та ін.]. – GB: Idea Group Publishing, 2002. – 310 p.
14. Alan. Agresti: Categorical Data Analysis. Wiley-Interscience, Nowy Jork, 2002.
15. T. Amemiya: Advanced Econometrics. Harvard University Press, 1985.
16. Hosmer, David W., Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed.. New York; Chichester, Wiley.
17. William H. Green: Econometric Analysis, fifth edition. Prentice Hall, 2003.
18. Breiman L., Friedman J., Olsen R. and Stone C. Classification and Regression Trees. Monterey, CA: Wadsworth, 1984.
19. Heath D, Kasif S, Salzberg S. Committees of decision trees. In: Gorayska B, Mey J, editors. Cognitive Technology: In Search of a Human Interface. Amsterdam, The Netherlands: Elsevier Science; 1996. pp. 305–317.
20. Chen X-W, Liu W. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics. 2005;21:4394–4400.
21. Quinlan JR, Rivest RL. Inferring decision trees using the minimum Description Length Principle. Inf. Comput. 1989;80:227–248.
22. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002 Oct;26(5):445-63. doi: 10.1023/a:1016409317640. PMID: 12182209.

23. Козак Ю. Г. Математичні методи та моделі для магістрів з економіки. Практичні застосування. [текст] Навч. посіб. / Ю. Г. Козак, В. М. Мацкул. – К.: Центр учбової літератури, 2017. – 254 с.
24. Назірова Т.О., Костенко О.Б. Нейрономережева інформаційна технологія опрацювання медичних даних // Науковий вісник НЛТУ України, 2018, т. 28, № 8. – С. 141 – 145.
25. Abdul Nazeer, K. A., Sebastian M. P., and. Madhu Kumar S.D., (2011) A Heuristic k-Means Algorithm with Better Accuracy and Efficiency for Clustering Health Informatics Data, Journal of Medical Imaging and Health Informatics, Vol. 1, 66.
26. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.
27. Jiawei Han and Micheline Kamber, (2006). Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition.
28. Ямненко Т. М., Літвінова І. Ф. Захист персональних даних у сфері охорони здоров'я (кримінально-правові аспекти)// Юридичний вісник, 1 (50), 2019. – с. 185 – 191.
29. "Data-driven healthcare organizations use big data analytics for big gains" IBM white paper February. 2013.
30. Yazan A, Yong W, Raj Kumar N. Big data life cycle: threats and security model. In: 21st Americas conference on information systems. 2015.
31. Langley, P. and Sage, S., Oblivious decision trees and abstract cases. in Working Notes of the AAAI-94 Workshop on Case-Based Reasoning, pp. 113–117, Seattle, WA: AAAI Press, 1994.
32. Last, M., Maimon, O. and Minkov, E., Improving Stability of Decision Trees, International Journal of Pattern Recognition and Artificial Intelligence, 16: 2, 145–159, 2002.
33. Lopez de Mantras R., A distance-based attribute selection measure for decision tree induction, Machine Learning 6:81–92, 1991.

34. Martin J. K., An exact probability metric for decision tree splitting and stopping. An Exact Probability Metric for Decision Tree Splitting and Stopping, Machine Learning, 28,2–3):257–291, 1997.
35. Mehta M., Rissanen J., Agrawal R., MDL-Based Decision Tree Pruning. KDD 1995: pp. 216–221, 1995.
36. Mingers J., An empirical comparison of pruning methods for decision tree induction. Machine Learning, 4(2):227–243, 1989.
37. Muller W., and Wysotzki F, Automatic construction of decision trees for classification. Annals of Operations Research, 52:231–247, 1994.
38. Murthy S. K., Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery, 2(4):345–389, 1998.
39. Naumov G.E., NP-completeness of problems of construction of optimal decision trees. Soviet Physics: Doklady, 36(4):270–271, 1991.
40. Niblett T. and Bratko I., Learning Decision Rules in Noisy Domains, Proc. Expert Systems 86, Cambridge: Cambridge University Press, 1986.
41. Olaru C, Wehenkel L., A complete fuzzy decision tree technique, Fuzzy Sets and Systems, 138(2):221–254, 2003.
42. Quinlan, J.R., Induction of decision trees, Machine Learning 1, 81–106, 1986.
43. Quinlan, J.R., Simplifying decision trees, International Journal of Man-Machine Studies, 27, 221–234, 1987.
44. Quinlan, J.R., Decision Trees and Multivalued Attributes, J. Richards, ed., Machine Intelligence, V. 11, Oxford, England, Oxford Univ. Press, pp. 305–318, 1988.
45. Quinlan, J. R. and Rivest, R. L., Inferring Decision Trees Using The Minimum Description Length Principle. Information and Computation, 80:227–248, 1989.
46. Rastogi, R., and Shim, K., PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning, Data Mining and Knowledge Discovery, 4(4):315–344, 2000.
47. Rounds, E., A combined non-parametric approach to feature selection and binary decision tree design, Pattern Recognition 12, 313–317, 1980.

48. Sethi, K., and Yoo, J. H., Design of multicategory, multifeature split decision trees using perceptron learning. Pattern Recognition, 27(7):939–947, 1994.
49. Utgoff, P. E., Incremental induction of decision trees. Machine Learning, 4: 161–186, 1989.
50. Utgoff, P. E., Decision tree induction based on efficient tree restructuring, Machine Learning 29,1):5–44, 1997.
51. Utgoff, P. E., and Clouse, J. A., A Kolmogorov-Smirnoff Metric for Decision Tree Induction, Technical Report 96-3, University of Massachusetts, Department of Computer Science, Amherst, MA, 1996.
52. Wallace, C. S., and Patrick J., Coding decision trees, Machine Learning 11: 7–22, 1993.
53. Zantema, H., and Bodlaender H. L., Finding Small Equivalent Decision Trees is Hard, International Journal of Foundations of Computer Science, 11(2): 343–354, 2000.
54. Kearns M. and Mansour Y., A fast, bottom-up decision tree pruning algorithm with near-optimal generalization, in J. Shavlik, ed., 'Machine Learning: Proceedings of the Fifteenth International Conference', Morgan Kaufmann Publishers, Inc., pp. 269–277, 1998.
55. Kearns M. and Mansour Y, On the boosting ability of top-down decision tree learning algorithms. Journal of Computer and Systems Sciences, 58(1): 109–128, 1999.
56. Kohavi R. and Sommerfield D., Targeting business users with decision table classifiers, in R. Agrawal, P. Stolorz & G. Piatetsky-Shapiro, eds, 'Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining', AAAI Press, pp. 249–253, 1998.

ДОДАТОК А
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

МАГІСТЕРСЬКА РОБОТА

«ЕКСПЕРТНА СИСТЕМА ДЛЯ МЕДИЧНОГО СКРИНІНГУ НА ОСНОВІ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ»

Виконала: студентка групи ПДМ-61 Кущук Валерія Андріївна

Керівник: доцент кафедри

кандидат педагогічних наук, доцент ІПЗ Шевченко С.М

Київ - 2023

АКТУАЛЬНІСТЬ

1. Первинний медичний скринінг дозволяє виявити ймовірність розвитку ракових пухлин.
2. Машинне навчання надає швидко обробляти великі дані, що пришвидшує процес діагностики.

АНАЛОГИ

1. Doc.ua (Україна) – розробляє штучний інтелект на базі медичного помічника, який допомагає визначити хворобу.
2. Sumitomo Dainippon Pharma (Японія) – розробляє нову формулу для медичних препаратів від обсесивно – компульсивного розладу.

1. Intel Corporation (США) та Elite Care (США) проектують систему штучного інтелекту, яка піклується про стан пацієнтів із хворобою Альцгеймера та підвищувати якість їхнього життя.
2. Medtronic (США) – додаток, що передбачає критичне зниження цукру за три години до події.

МЕТА, ОБ’ЄКТ, ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: поліпшення ранньої діагностики можливих захворювань людини через впровадження медичного скринінгу на основі методів кластерного аналізу.

Об’єкт дослідження: процес функціювання експертної системи у медичній галузі.

Предмет дослідження: методи кластерного аналізу.

ЗАВДАННЯ

1. Проаналізувати наукові праці з досліджуваної проблеми і обґрунтовати застосування машинного навчання для медичного скринінгу.
2. З'ясувати особливості експертної системи для медичного скринінгу на основі методів машинного навчання.
3. Здійснити порівняльний аналіз методів кластерного аналізу: метод ближнього сусіда та метод к-середніх.
4. Розробити та теоретично обґрунтовати модель експертної системи для медичного скринінгу на основі методів кластерного аналізу.
5. Дослідити шляхи захисту персональної інформації у медичних експертних системах.

МЕТОДИ ДОСЛІДЖЕННЯ

1. Системно – структурний.
2. Порівняльний.
3. Методи кластерного аналізу.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Науковий результат: розроблено алгоритм функціювання експертної системи у медичній сфері на основі методів кластерного аналізу.

Практичний результат: основні положення та результати магістерської роботи можуть бути використані для розробки програмного забезпечення експертної системи для медичного скринінгу та впроваджені у навчальний процес медичних закладів.

МЕДИЧНА ЕКСПЕРТНА СИСТЕМА

Експертна система – це система штучного інтелекту, яка на основі знань застосовує одержання висновків для вирішення задачі.

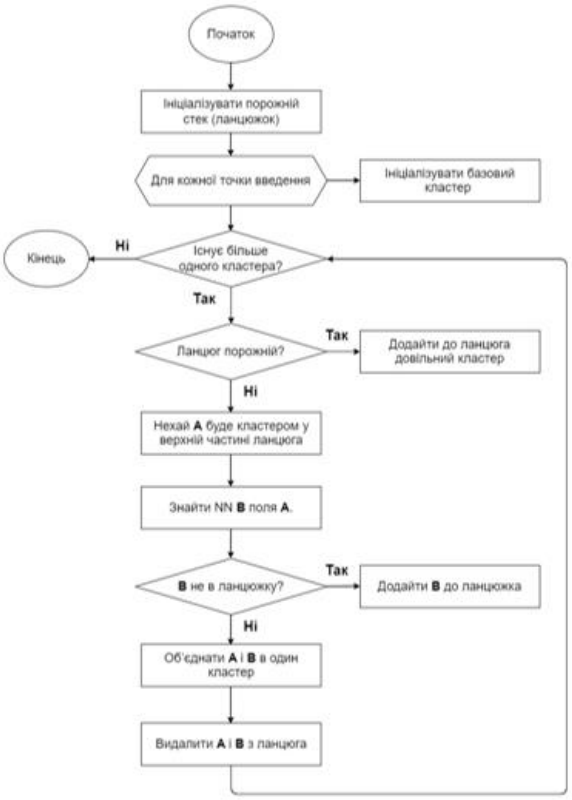
Структура експертної системи

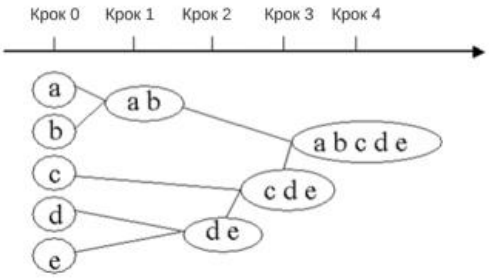
Алгоритми контрольованого машинного навчання	Алгоритми неконтрольованого машинного навчання
Логістична регресія	Факторний аналіз
Дерево рішень	Дискримінантний аналіз
Штучні нейронні мережі	Кластерний аналіз

МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ

Назва відстані	Формула
Евклідова відстань	$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
Зважена Евклідова відстань	$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k (x_{ik} - x_{jk})^2}$
Метрика Мінковського	$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^r}$
Хеммінгова відстань	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$

МЕТОД БЛИЖНЬОГО СУСІДА





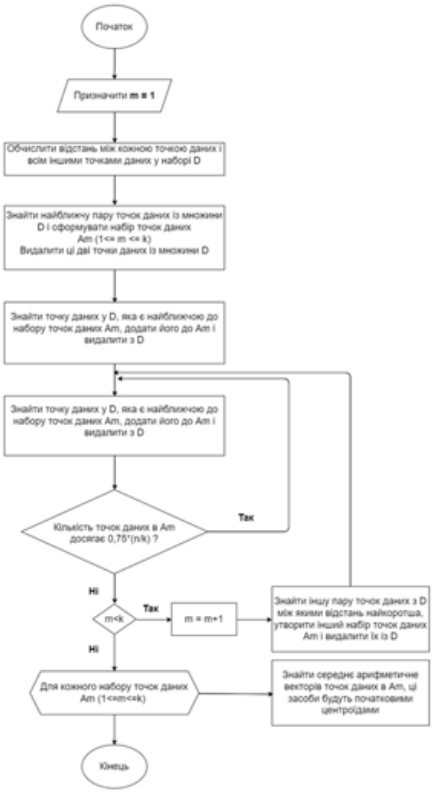

```
float tx = width_source/width_dst;
float ty = height_source/ height_dst;

for(i=0; i<height_dst; i++)
 for(j=0; j<width_dst; j++)
  {
      x = ceil(j*tx);
      y = ceil(i*ty);
      U(i,j) = P(y,x);
  }
```

8

МЕТОД K-MEANS

1. Виберіть k точок як початкові центроїди.
2. Повторіть
3. З К кластерів, приписуючи кожній точці свій найближчий центроїд.
4. Повторно обчисліть центроїд кожного кластера.
5. Поки центроїди не зміняться k-means досягає стану, в якому немає точок перехід від одного кластера до іншого, напр. повторюючи до тих пір, поки лише 1% точок змінюють кластери.

МЕТОД K-MEANS
1. Знаходження початкових центроїдів



Алгоритм знаходження початкових центроїдів.

Вхідні дані: D = {d1, d2,.....,dn} // набір з n елементів даних

k // Кількість бажаних кластерів

Результат: набір із k початкових центроїдів.

Кроки:

Крок 0. Встановіть m = 1;

Крок 1. Обчисліть відстань між кожною точкою даних і всіма інші точки даних у наборі D;

Крок 2. Знайдіть найближчу пару точок даних із множини D і сформуйте набір точок даних Am (1<= m <= k), який містить ці дві точки даних; видаліть ці дві точки даних із множина D;

Крок 3. Знайдіть точку даних у D, яка є найближчою до набору точок даних Am, додайте її до Am і видаліть з D;

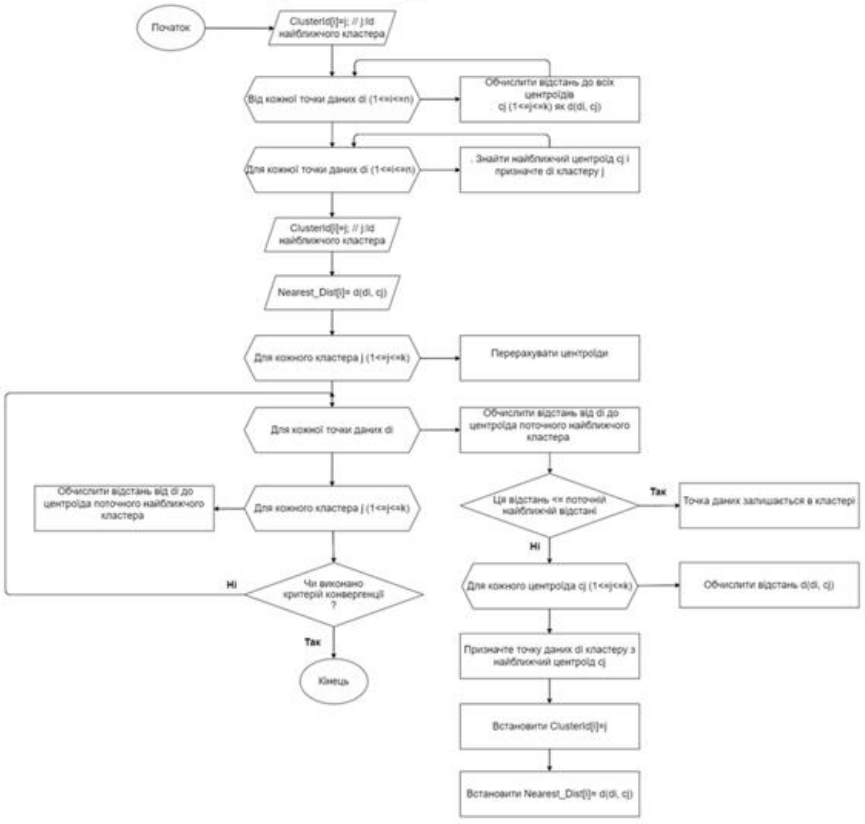
Крок 4. Повторюйте крок 3, доки кількість точок даних в Am досягає 0,75*(n/k);

Крок 5. Якщо m<k, то m = m+1, знайдіть іншу пару точок даних від D, між якими відстань найкоротша, інша точка даних встановлює Am і видаляйте їх із D, йти до кроку 4;

Крок 6. Для кожного набору точок даних Am (1<=m<=k) знайдіть середнє арифметичне векторів точок даних в Am,

Ці елементи будуть початковими центроїдами

МЕТОД K-MEANS
2. Алгоритм узагальнений

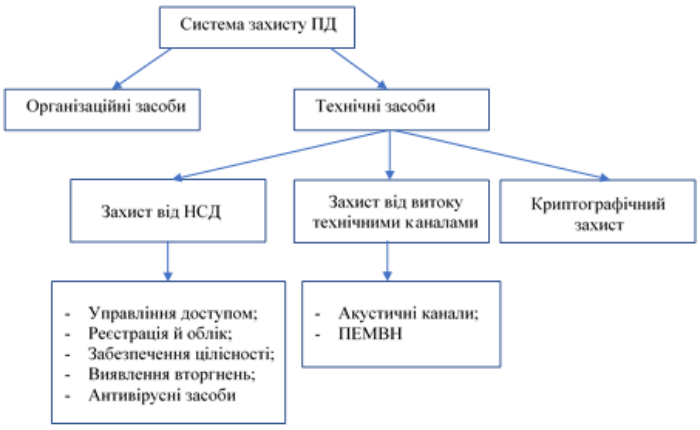


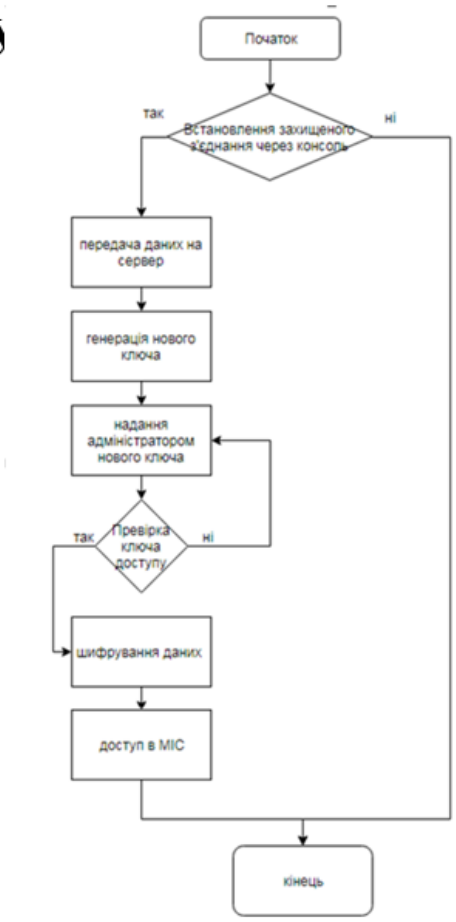
ПЕРЕВАГИ ТА НЕДОЛІКИ МЕТОДІВ НАЙБЛИЖЧНОГО СУСІДА ТА К СЕРЕДНІХ

	Метод ближнього сусіда	Метод k середніх
Переваги	1. Алгоритм простий та легко реалізується. 2. Нечутливий до викидів. 3. Не потрібно будувати модель. 4. Універсальний, оскільки використовується для завдань класифікації та регресії.	1. Гнучкість, швидкість та простота використання. 2. Зрозумілий. 3. Перевірка статистичної значимості відмінностей між виділеними кластерами.
Недоліки	1. Алгоритм працює повільніше, якщо збільшити обсяг вибірки, предикторів чи незалежних змінних. 2. Обчислювальні витрати під час виконання та обробки великих даних. 3. Не створює жодних моделей або правил, які узагальнюють попередній досвід.	1. Результати кластеризації залежать від вибору початкової конфігурації центроїдів. 2. Заздалегідь визначити кількість кластерів. 3. Повільний при кластеризації велику обсягу даних. 4. Чутливий до викидів.

12

ЗАБЕЗПЕЧЕННЯ ЗАХИСТУ МЕДИЧНИХ ДАНИХ У ЕКСПЕРТНИХ СИСТЕМАХ





13

ВИСНОВКИ

У ході проведеного дослідження були вирішені всі поставлені завдання і відповідно до мети отримані наступні **результати**:

1. Проаналізовано наукові праці з досліджуваної проблеми, обґрунтовано застосування машинного навчання для медичного скринінгу.
2. З'ясовано особливості експертної системи для медичного скринінгу на основі методів машинного навчання.
3. Здійснено порівняльний аналіз методів кластерного аналізу: метод ближнього сусіда та метод к-середніх.
4. Розроблено і теоретично обґрунтовано модель експертної системи для медичного скринінгу на основі методів кластерного аналізу.
5. Досліджені шляхи захисту персональної інформації у медичних експертних системах.

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Статті:

1. Куцук В.А. Модель експертної системи для медичного скринінгу на основі методів кластерного аналізу / Шевченко С.М., Жданова Ю.Д., Негоденко О.В., Куцук В.А. // Moderní aspekty vědy: XXVII. Díl mezinárodní kolektivní monografie / Mezinárodní Ekonomický Institut s.r.o.. Česká republika: Mezinárodní Ekonomický Institut s.r.o., 2023. – С. 478 – 494

Тези доповідей:

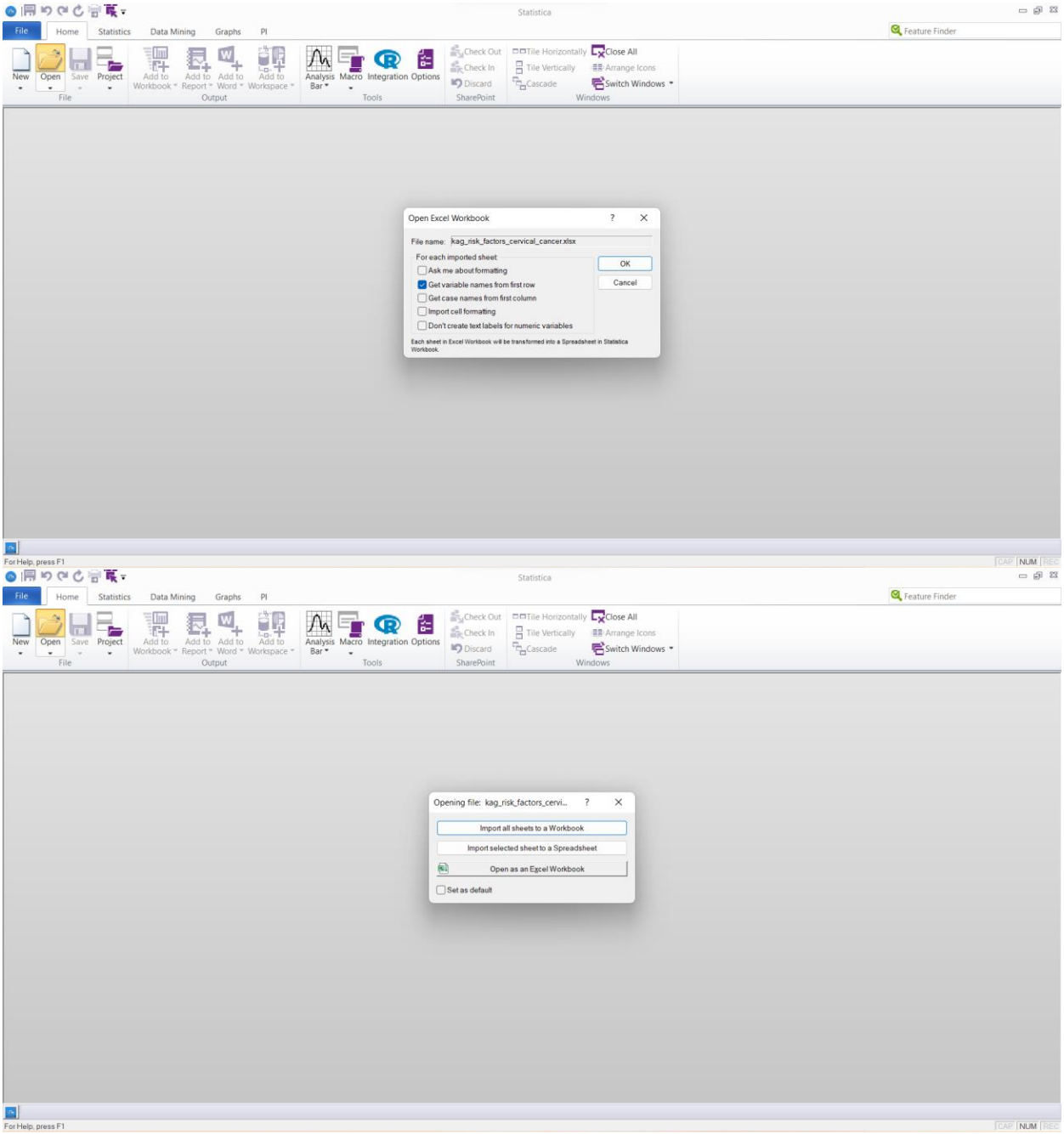
1. Куцук В.А. Експертна система для медичного скринінгу на основі методів кластерного аналізу // XV Науково-технічна конференція «Сучасні інфокомунікаційні технології» – Київ: ДУТ, 2022 – С. 107 – 108

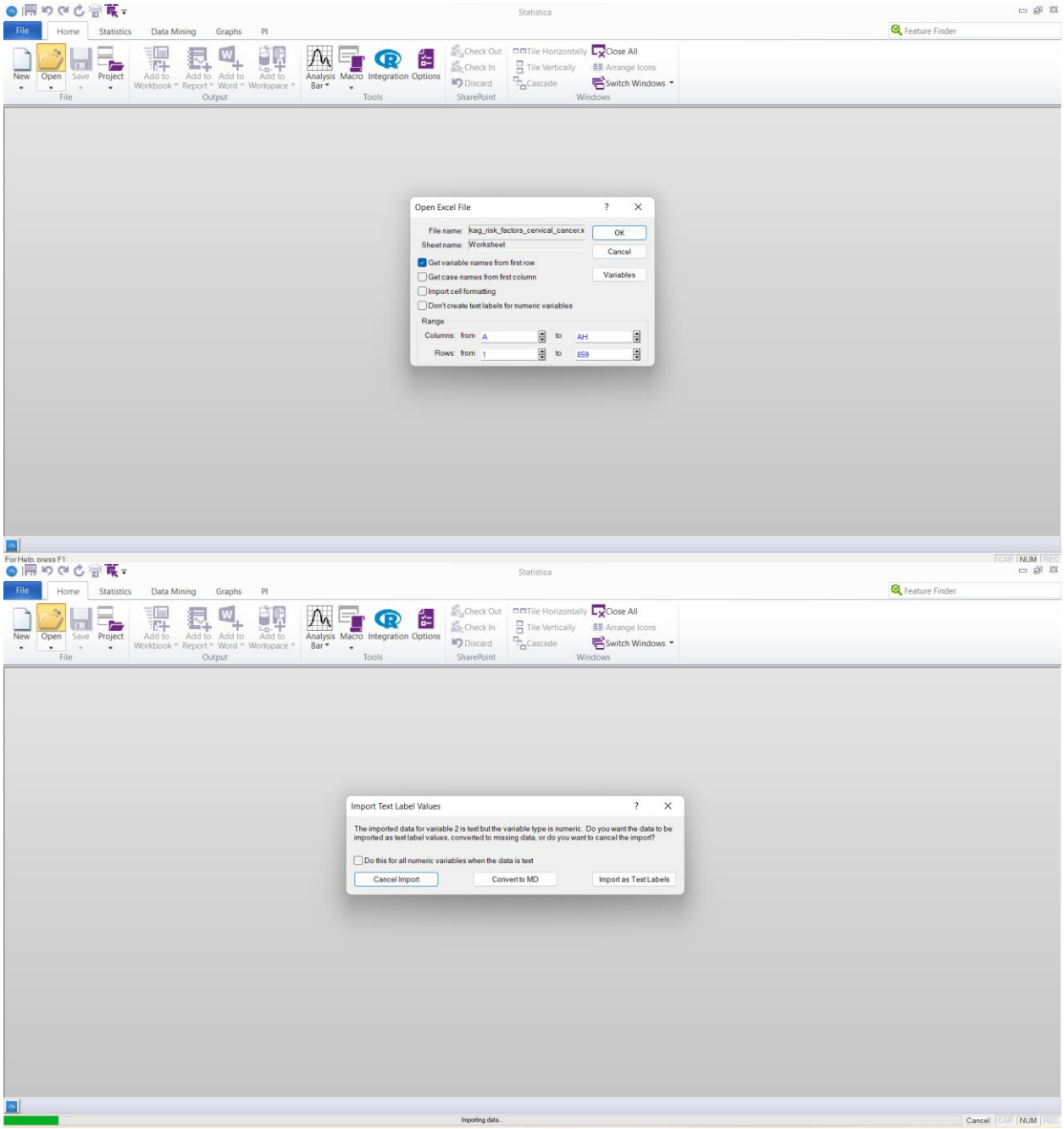
ДЯКУЮ ЗА УВАГУ!

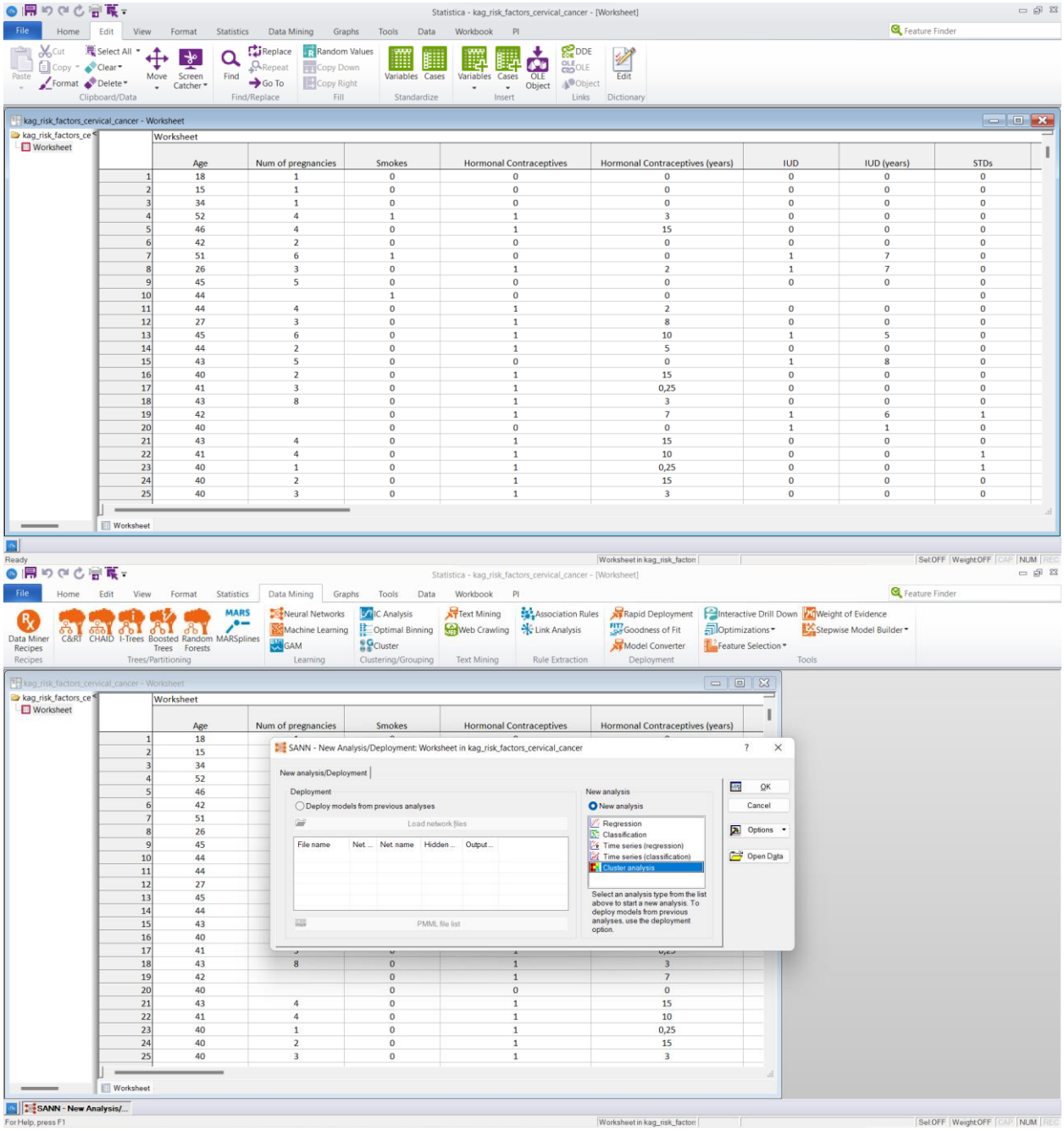
ДОДАТОК Б

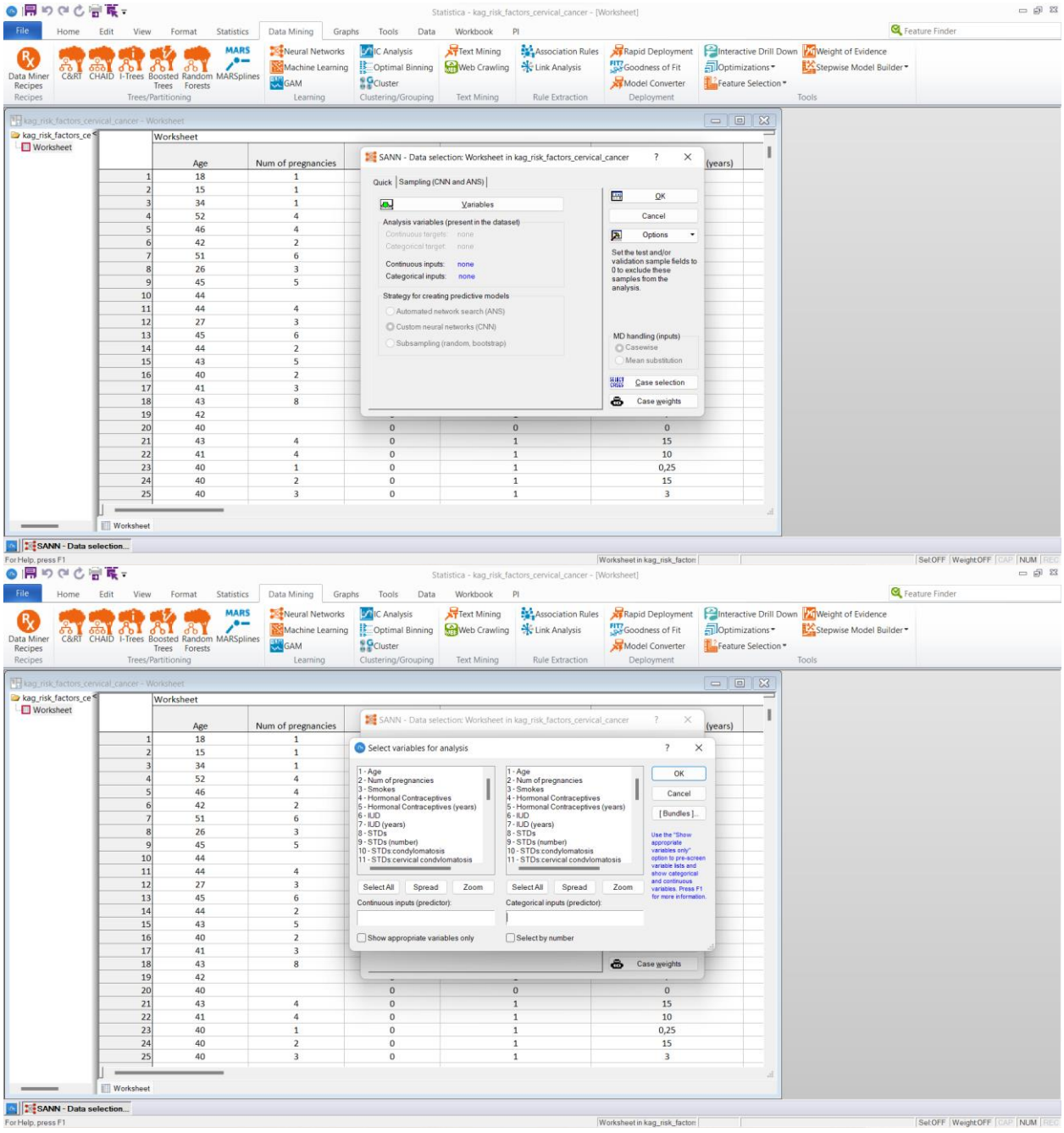
ПРОЦЕС КЛАСТЕРИЗАЦІЇ ЗА ДОПОМОГОЮ «STATISTICA»

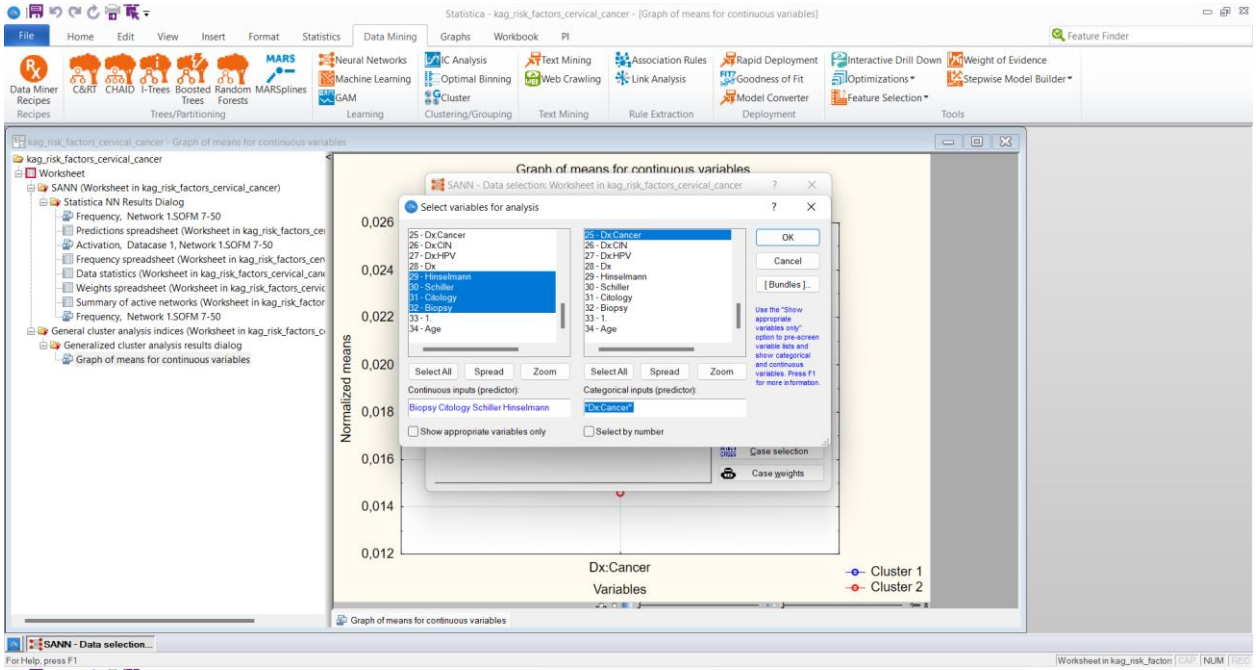
КЛАСТЕРНИЙ АНАЛІЗ

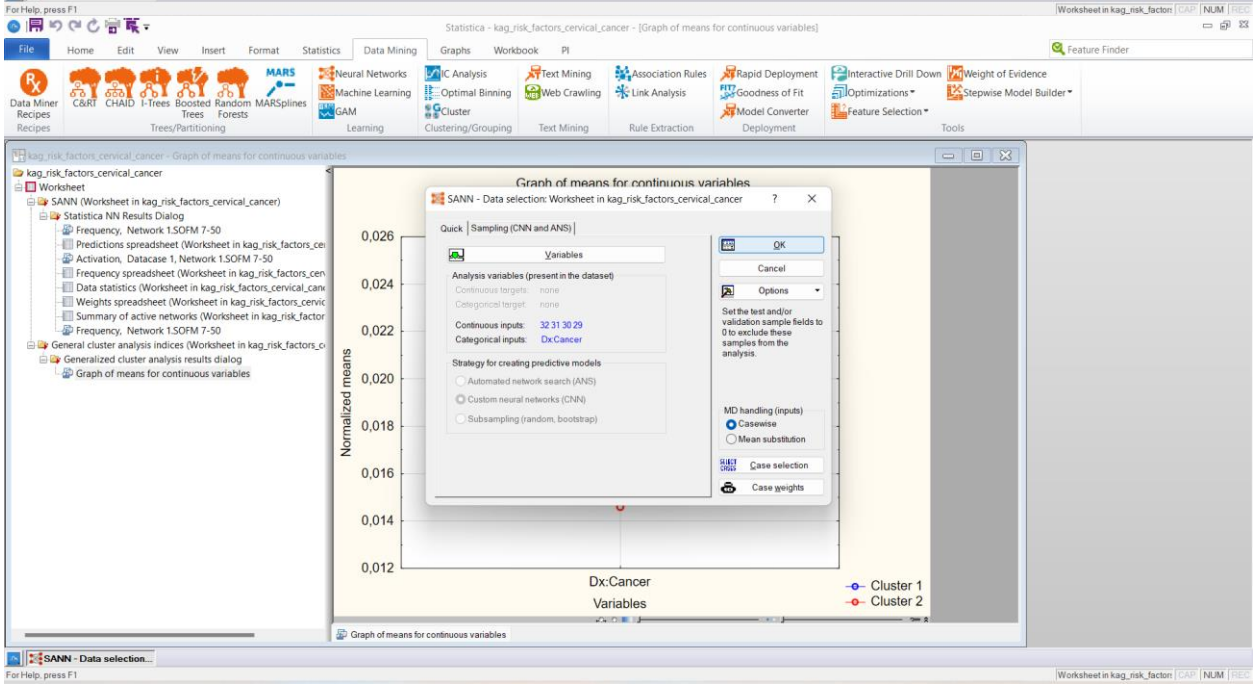


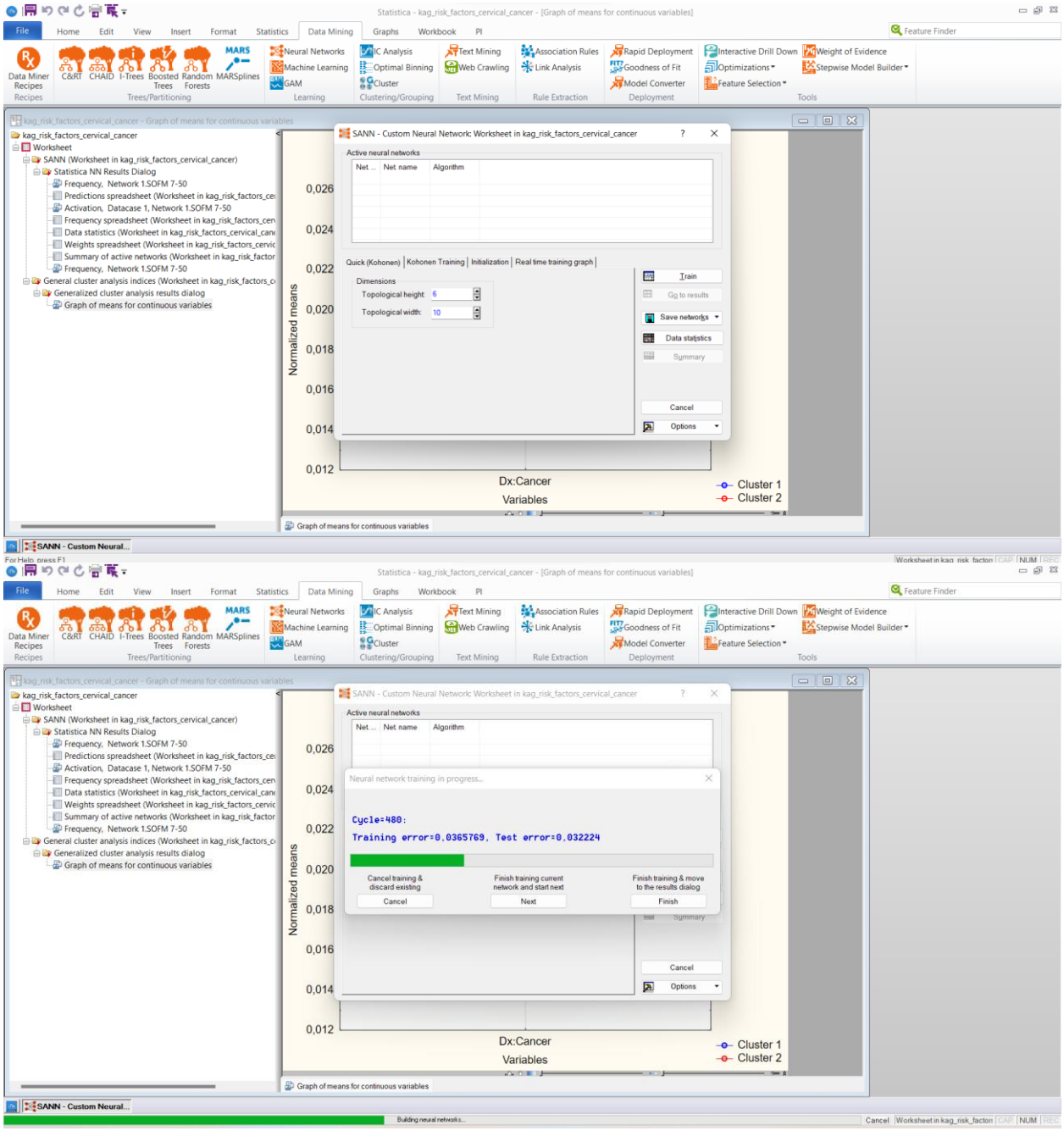

Statistica
File
Home
Statistics
Data Mining
Graphs
PI
Feature Finder
New
Open
Save
Project
Add to Workbook
Add to Report
Add to Word
Add to Workspace
Analysis Bar
Macro
Integration
Options
Check Out
Check In
Discard
Tile Horizontally
Tile Vertically
Cascade
Close All
Arrange Icons
Switch Windows
File
Output
Tools
SharePoint
Windows
Open Excel File
File name: kag_risk_factors_cervical_cancer.x
Sheet name: Worksheet
OK
Cancel
Variables
Get variable names from first row
Get case names from first column
Import cell formatting
Don't create text labels for numeric variables
Range
Columns: from A to AH
Rows: from 1 to 859
For Help, press F1
CAP NUM REC
Import Text Label Values
The imported data for variable 2 is text but the variable type is numeric. Do you want the data to be imported as text label values, converted to missing data, or do you want to cancel the import?
Do this for all numeric variables when the data is text
Cancel Import
Convert to MD
Import as Text Labels
Importing data...
Cancel

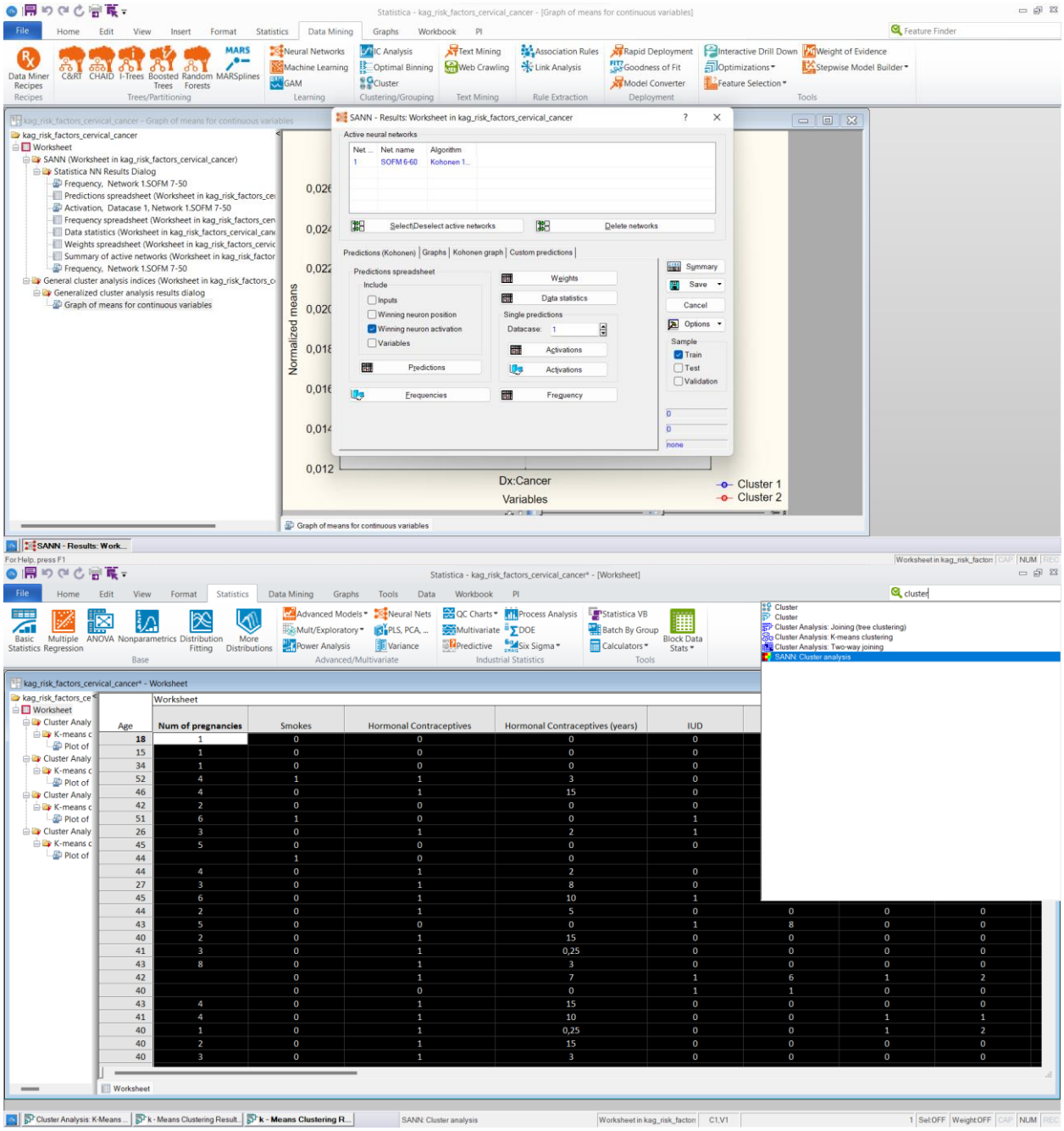

Statistica - kag_risk_factors_cervical_cancer - [Worksheet]
kag_risk_factors_cervical_cancer - Worksheet
Worksheet
Age
Num of pregnancies
Smokes
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
SANN - New Analysis/Deployment: Worksheet in kag_risk_factors_cervical_cancer
New analysis/Deployment
Deployment
Deploy models from previous analyses
Load network files
File name
Net ...
Net name
Hidden ...
Output ...
PMML file list
New analysis
New analysis
Regression
Classification
Time series (regression)
Time series (classification)
Cluster analysis
Select an analysis type from the list above to start a new analysis. To deploy models from previous analyses, use the deployment option.
OK
Cancel
Options
Open Data

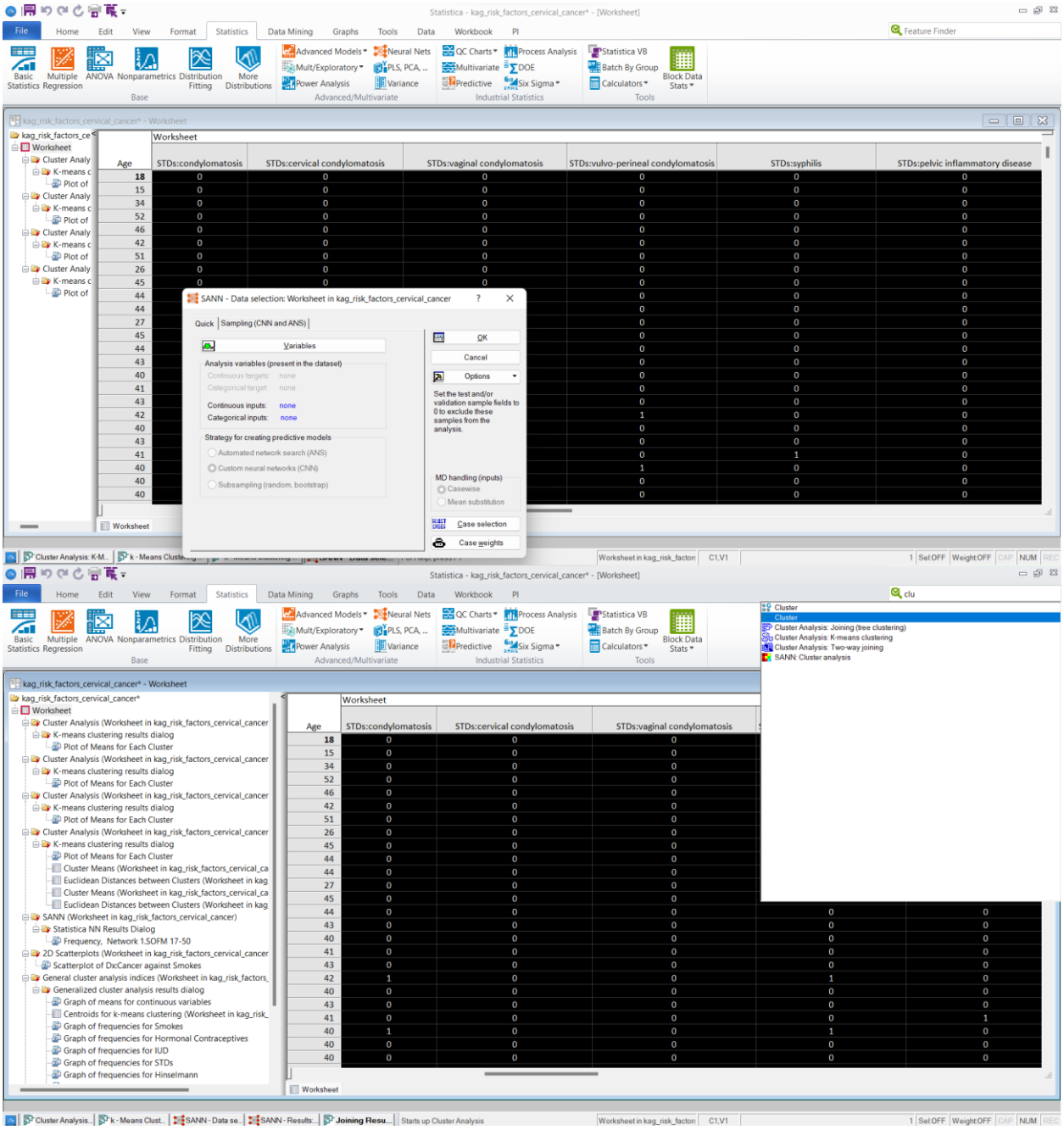

Statistica - kag_risk_factors_cervical_cancer - [Worksheet]
SANN - Data selection: Worksheet in kag_risk_factors_cervical_cancer
Select variables for analysis
Worksheet
Age
Num of pregnancies
(years)
OK
Cancel
Case weights
SANN - Data selection...

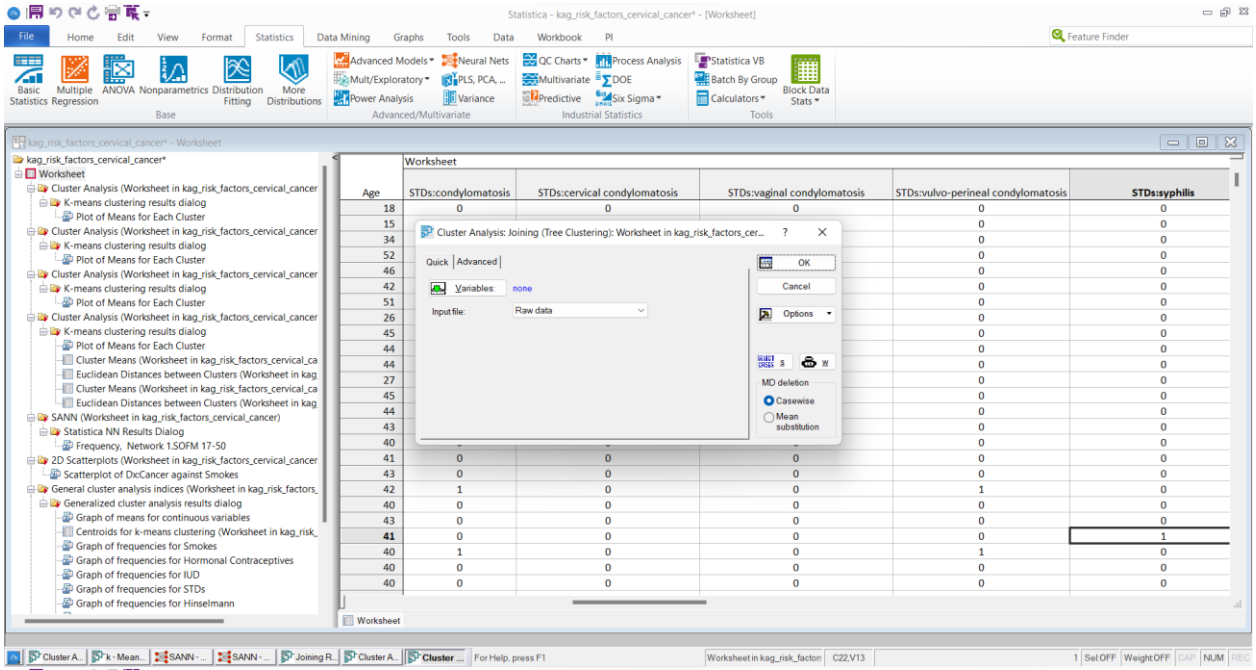

Graph of means for continuous variables
Normalized means
Dx:Cancer
Variables
Cluster 1
Cluster 2
Select variables for analysis
OK
Cancel
Continuous inputs (predictor):
Categorical inputs (predictor):

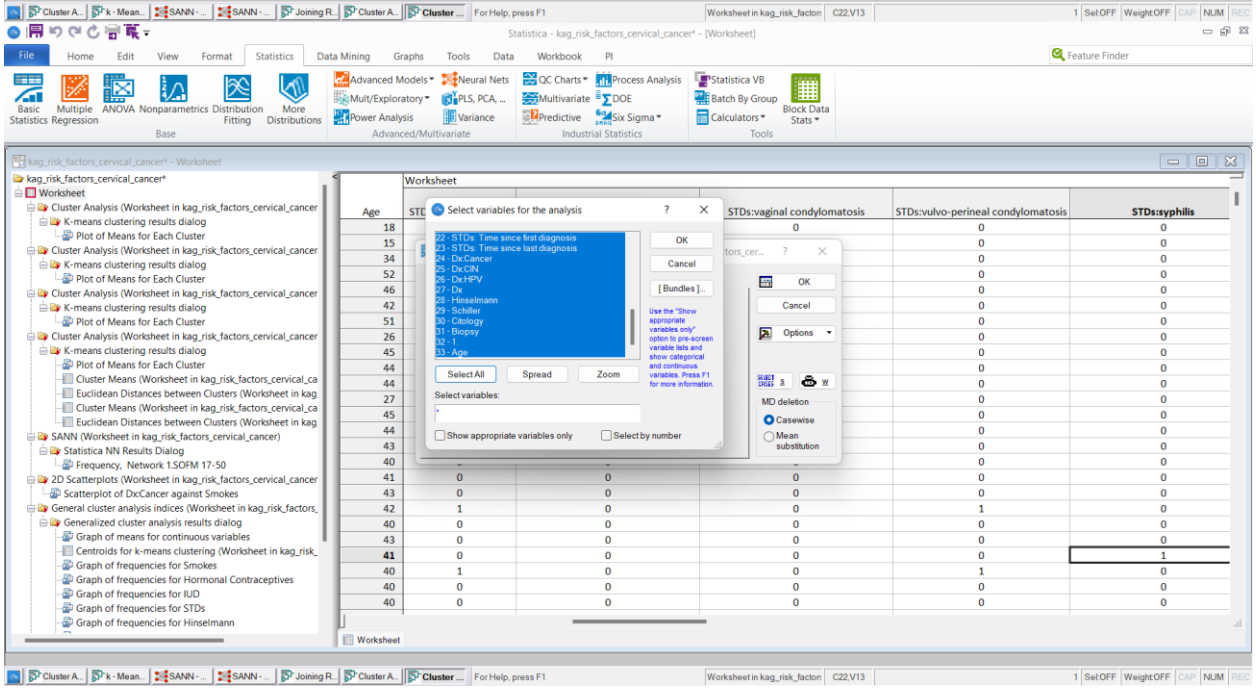

Graph of means for continuous variables
Normalized means
Dx:Cancer
Variables
Cluster 1
Cluster 2
SANN - Data selection: Worksheet in kag_risk_factors_cervical_cancer
Quick
Sampling (CNN and ANS)
Variables
OK
Cancel
Options

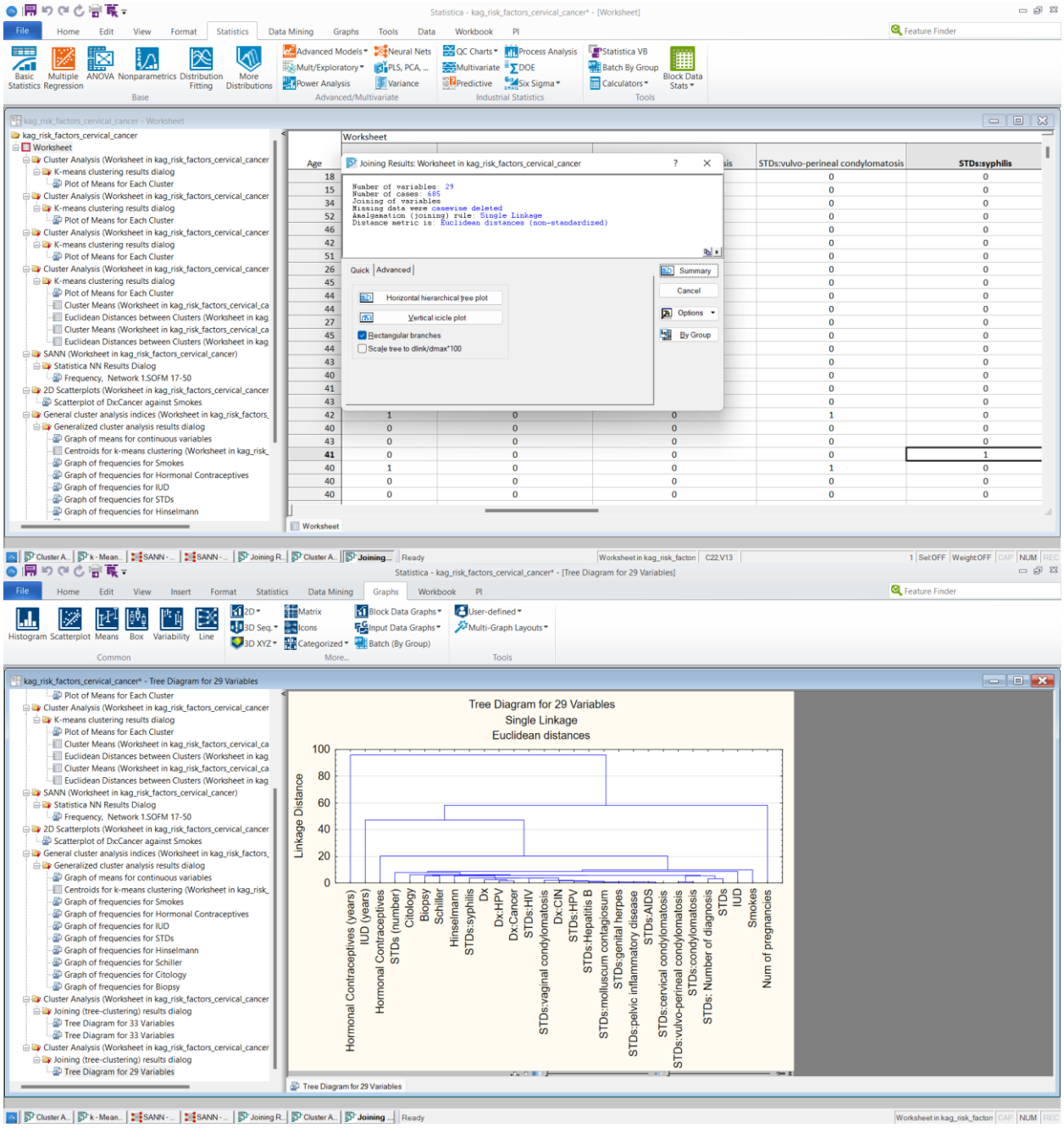

Statistica - kag_risk_factors_cervical_cancer - [Graph of means for continuous variables]
SANN - Custom Neural Network: Worksheet in kag_risk_factors_cervical_cancer
Active neural networks
Quick (Kohonen) | Kohonen Training | Initialization | Real time training graph
Dimensions
Topological height: 6
Topological width: 10
Train
Go to results
Save networks
Data statistics
Summary
Cancel
Options
Normalized means
0,026
0,024
0,022
0,020
0,018
0,016
0,014
0,012
Dx:Cancer
Variables
Cluster 1
Cluster 2
Graph of means for continuous variables
Neural network training in progress...
Cycle=480:
Training error=0,0365769, Test error=0,032224
Cancel training & discard existing
Cancel
Finish training current network and start next
Next
Finish training & move to the results dialog
Finish


Statistica - kag_risk_factors_cervical_cancer - [Graph of means for continuous variables]
SANN - Results: Worksheet in kag_risk_factors_cervical_cancer
Normalized means
Dx:Cancer
Variables
Cluster 1
Cluster 2
Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
kag_risk_factors_cervical_cancer* - Worksheet
Worksheet
Age
Num of pregnancies
Smokes
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD

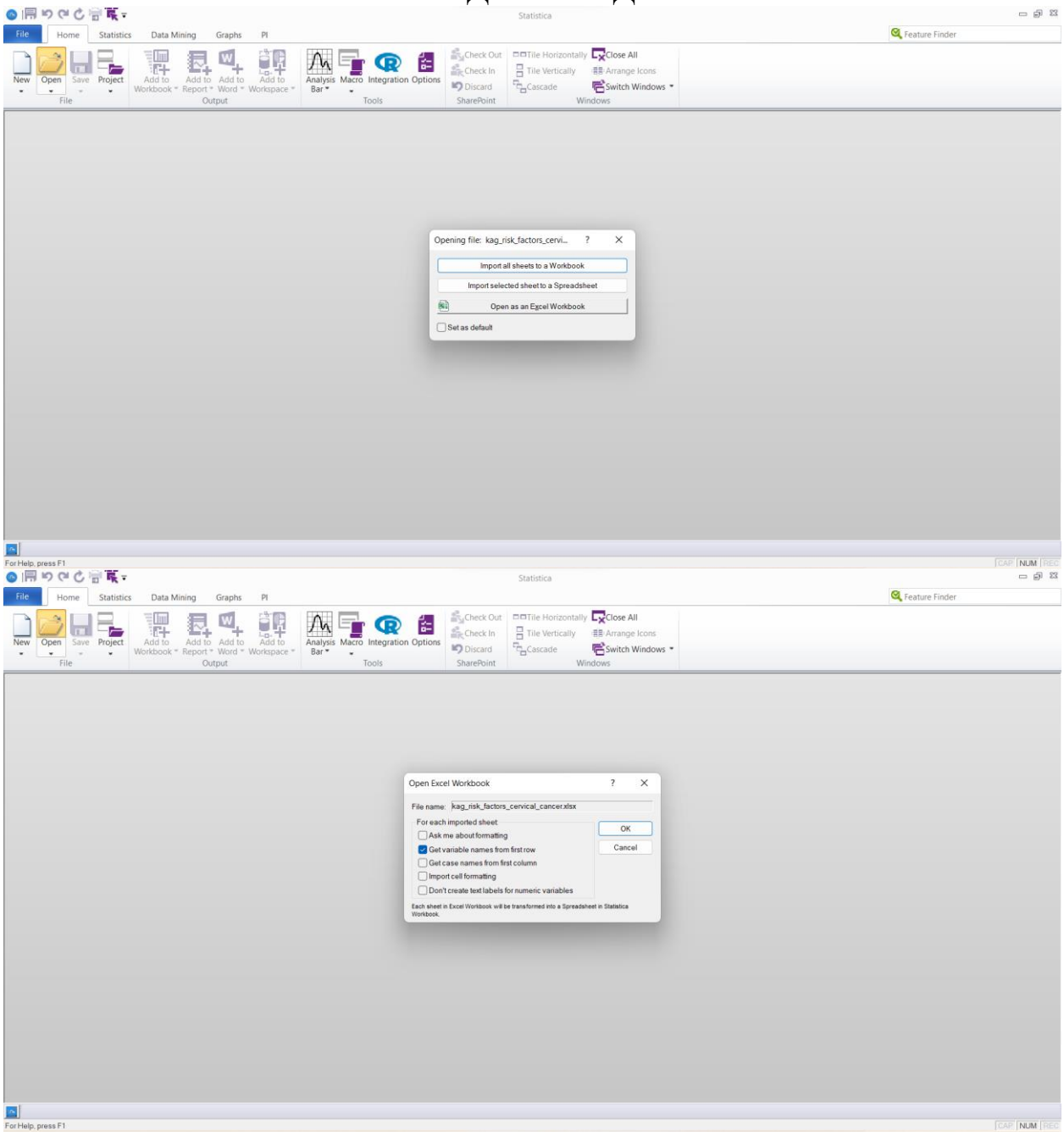

Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
File Home Edit View Format Statistics Data Mining Graphs Tools Data Workbook PI
Feature Finder
Basic Statistics
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
More Distributions
Advanced Models
Mult/Exploratory
Power Analysis
Neural Nets
PLS, PCA, ...
Variance
QC Charts
Multivariate
Predictive
Process Analysis
DOE
Six Sigma
Statistica VB
Batch By Group
Calculators
Block Data Stats
Base
Advanced/Multivariate
Industrial Statistics
Tools
kag_risk_factors_cervical_cancer* - Worksheet
Worksheet
Age
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
SANN - Data selection: Worksheet in kag_risk_factors_cervical_cancer
Quick
Sampling (CNN and ANS)
Variables
Analysis variables (present in the dataset)
Continuous targets: none
Categorical target: none
Continuous inputs: none
Categorical inputs: none
Strategy for creating predictive models
Automated network search (ANS)
Custom neural networks (CNN)
Subsampling (random, bootstrap)
OK
Cancel
Options
Set the test and/or validation sample fields to 0 to exclude these samples from the analysis.
MD handling (inputs)
Casewise
Mean substitution
Case selection
Case weights
Worksheet
Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
clu
Cluster
Cluster
Cluster Analysis: Joining (tree clustering)
Cluster Analysis: K-means clustering
Cluster Analysis: Two-way joining
SANN: Cluster analysis
kag_risk_factors_cervical_cancer* - Worksheet
kag_risk_factors_cervical_cancer*
Worksheet
Cluster Analysis (Worksheet in kag_risk_factors_cervical_cancer
K-means clustering results dialog
Plot of Means for Each Cluster
Cluster Means (Worksheet in kag_risk_factors_cervical_ca
Euclidean Distances between Clusters (Worksheet in kag
SANN (Worksheet in kag_risk_factors_cervical_cancer)
Statistica NN Results Dialog
Frequency, Network 1:SOFM 17-50
2D Scatterplots (Worksheet in kag_risk_factors_cervical_cancer
Scatterplot of DxCancer against Smokes
General cluster analysis indices (Worksheet in kag_risk_factors_
Generalized cluster analysis results dialog
Graph of means for continuous variables
Centroids for k-means clustering (Worksheet in kag_risk_
Graph of frequencies for Smokes
Graph of frequencies for Hormonal Contraceptives
Graph of frequencies for IUD
Graph of frequencies for STDs
Graph of frequencies for Hinselmann
Worksheet
Age
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
Worksheet
Cluster Analysis...
k - Means Clust...
SANN - Data se...
SANN - Results...
Joining Resu...
Starts up Cluster Analysis
Worksheet in kag_risk_factor
C1,V1
Set:OFF
Weight:OFF
NUM

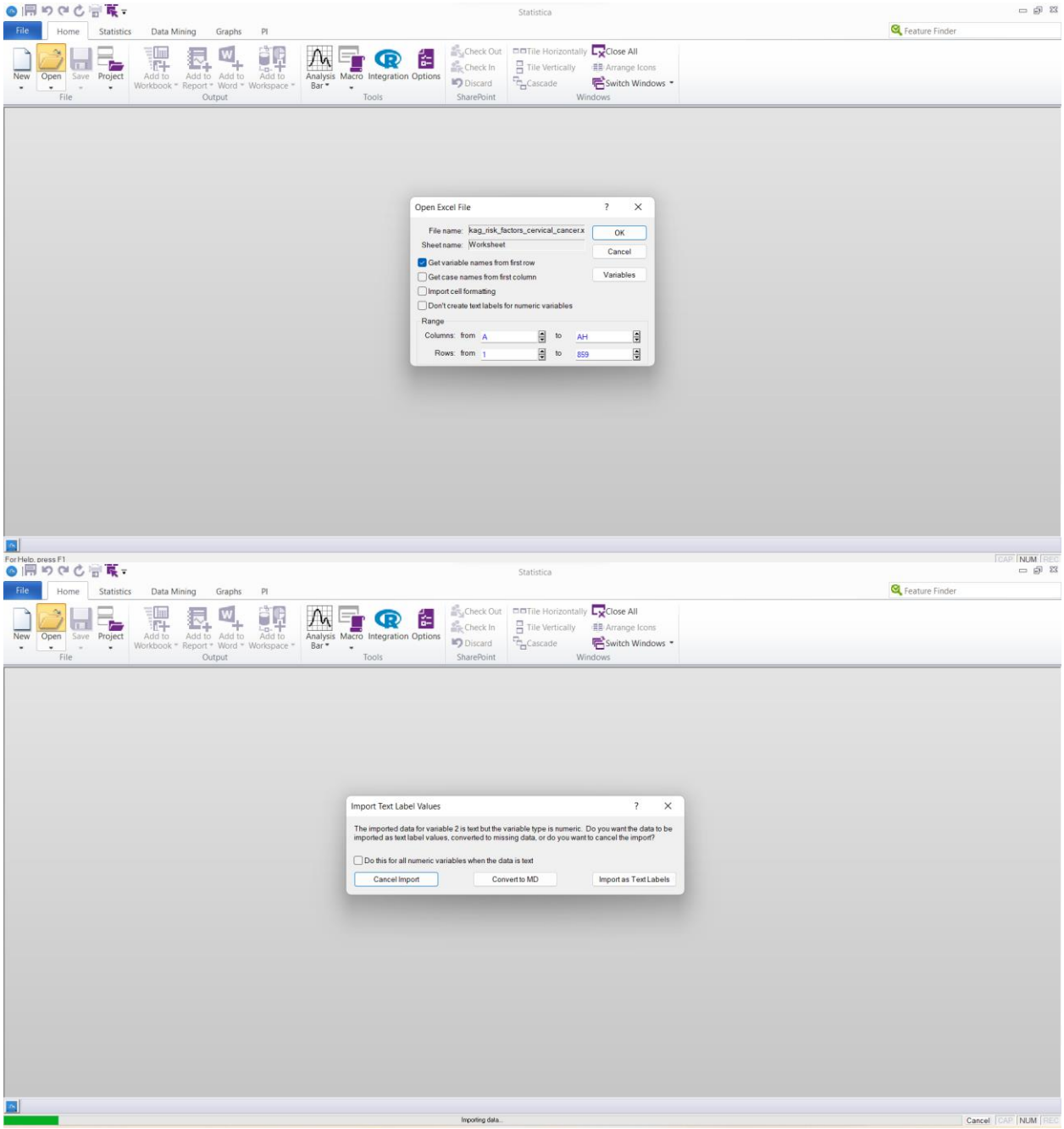

Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
Cluster Analysis: Joining (Tree Clustering): Worksheet in kag_risk_factors_cer...
Worksheet
Age
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis


Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
Select variables for the analysis
22 - STDs: Time since first diagnosis
23 - STDs: Time since last diagnosis
24 - Dx:Cancer
25 - Dx:CIN
26 - Dx:HPV
27 - Dx
28 - Hinselmann
29 - Schiller
30 - Citology
31 - Biopsy
32 - 1
33 - Age
Worksheet
Age
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis

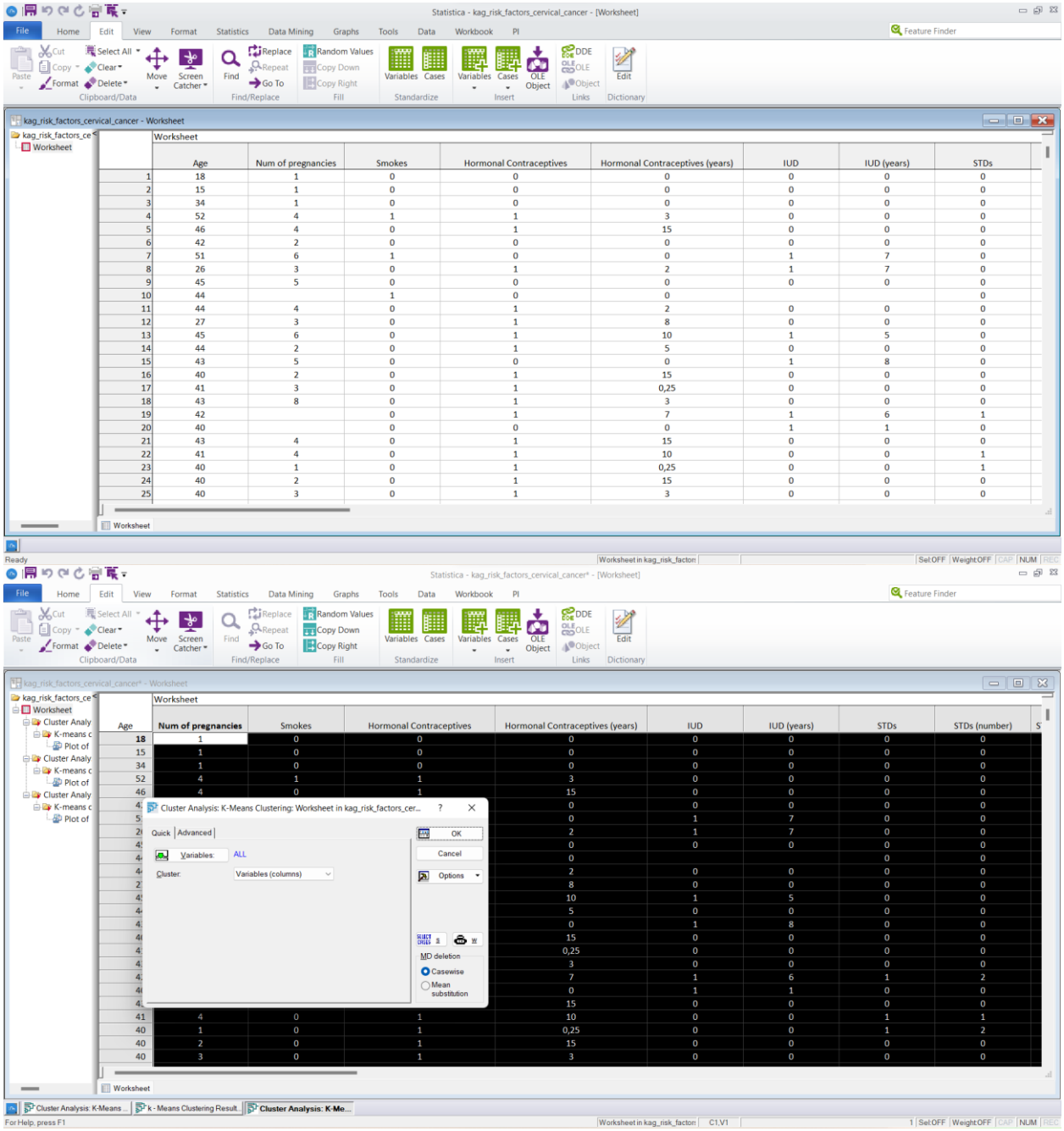

Joining Results: Worksheet in kag_risk_factors_cervical_cancer
Number of variables: 29
Number of cases: 685
Joining of variables
Missing data were casewise deleted
Amalgamation (joining) rule: Single Linkage
Distance metric is: Euclidean distances (non-standardized)
Horizontal hierarchical tree plot
Vertical icicle plot
Rectangular branches
Scale tree to dlink/dmax*100
Summary
Cancel
Options
By Group
Tree Diagram for 29 Variables
Single Linkage
Euclidean distances
Linkage Distance
Hormonal Contraceptives (years)
IUD (years)
Hormonal Contraceptives
STDs (number)
Citology
Biopsy
Schiller
Hinselmann
STDs:syphilis
Dx
Dx:HPV
Dx:Cancer
STDs:HIV
STDs:vaginal condylomatosis
Dx:CIN
STDs:HPV
STDs:Hepatitis B
STDs:molluscum contagiosum
STDs:genital herpes
STDs:pelvic inflammatory disease
STDs:AIDS
STDs:cervical condylomatosis
STDs:condylomatosis
STDs:vulvo-perineal condylomatosis
STDs: Number of diagnosis
STDs
IUD
Smokes
Num of pregnancies

МЕТОД К СЕРЕДНІХ

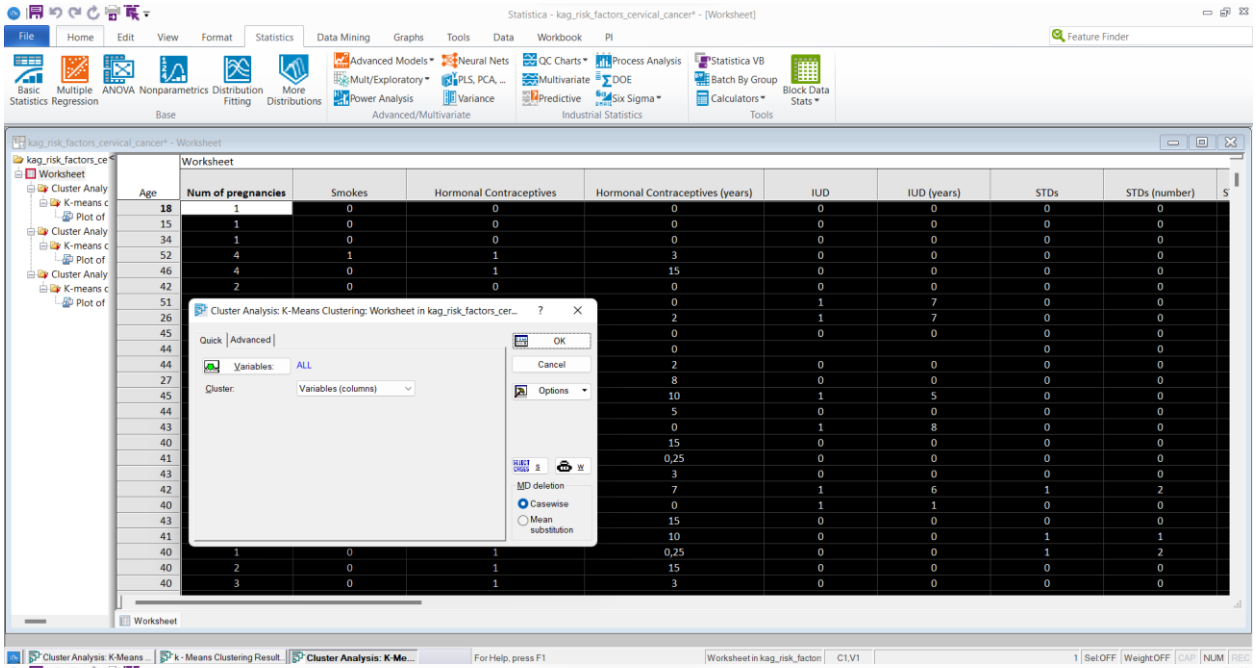


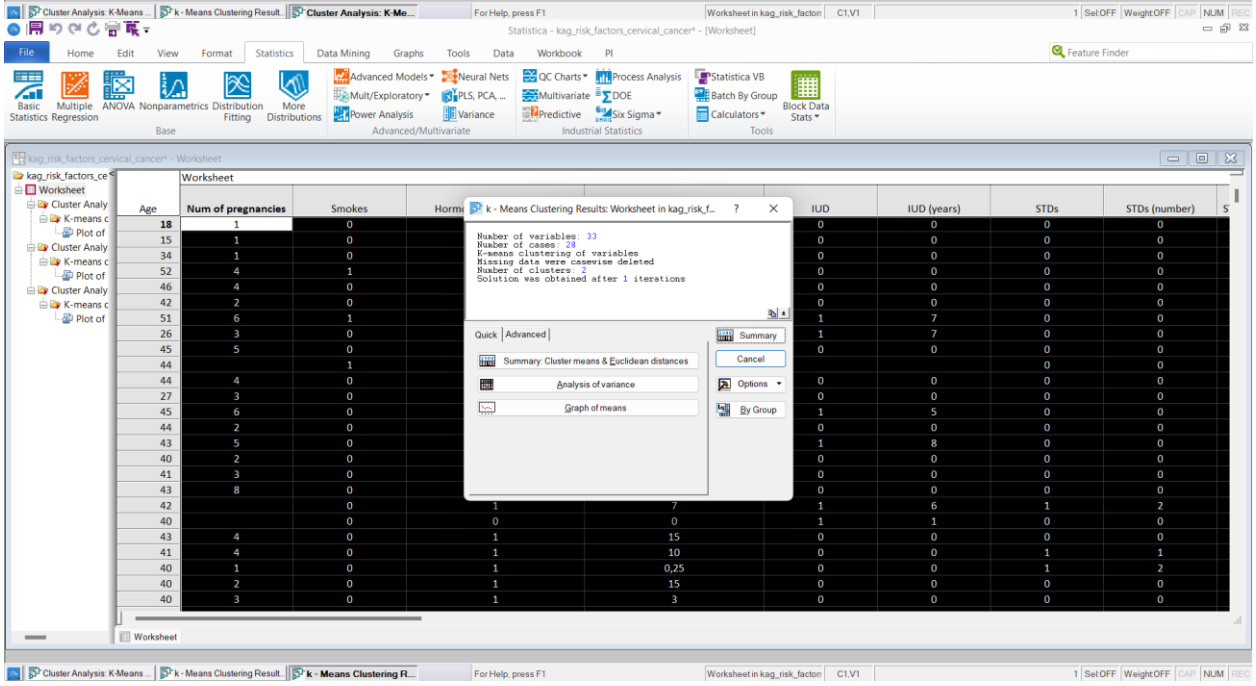

Statistica
File
Home
Statistics
Data Mining
Graphs
PI
Feature Finder
New
Open
Save
Project
Add to Workbook
Add to Report
Add to Word
Add to Workspace
Analysis Bar
Macro
Integration
Options
Check Out
Check In
Discard
Tile Horizontally
Tile Vertically
Cascade
Close All
Arrange Icons
Switch Windows
File
Output
Tools
SharePoint
Windows
Open Excel File
File name: kag_risk_factors_cervical_cancer.x
Sheet name: Worksheet
OK
Cancel
Variables
Get variable names from first row
Get case names from first column
Import cell formatting
Don't create text labels for numeric variables
Range
Columns: from A to AH
Rows: from 1 to 859

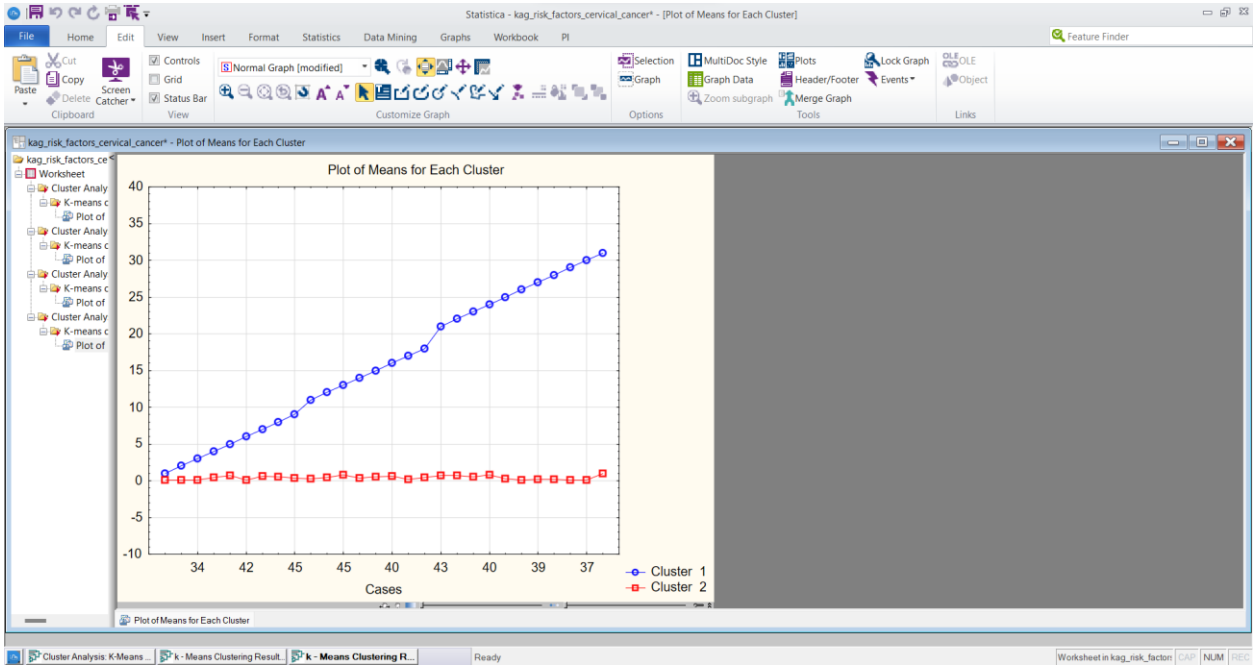
For Help, press F1
Statistica
File
Home
Statistics
Data Mining
Graphs
PI
Feature Finder
Import Text Label Values
The imported data for variable 2 is text but the variable type is numeric. Do you want the data to be imported as text label values, converted to missing data, or do you want to cancel the import?
Do this for all numeric variables when the data is text
Cancel Import
Convert to MD
Import as Text Labels
Importing data...
Cancel



	Age	Num of pregnancies	Smokes	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs
1	18	1	0	0	0	0	0	0
2	15	1	0	0	0	0	0	0
3	34	1	0	0	0	0	0	0
4	52	4	1	1	3	0	0	0
5	46	4	0	1	15	0	0	0
6	42	2	0	0	0	0	0	0
7	51	6	1	0	0	1	7	0
8	26	3	0	1	2	1	7	0
9	45	5	0	0	0	0	0	0
10	44		1	0	0			0
11	44	4	0	1	2	0	0	0
12	27	3	0	1	8	0	0	0
13	45	6	0	1	10	1	5	0
14	44	2	0	1	5	0	0	0
15	43	5	0	0	0	1	8	0
16	40	2	0	1	15	0	0	0
17	41	3	0	1	0,25	0	0	0
18	43	8	0	1	3	0	0	0
19	42		0	1	7	1	6	1
20	40		0	0	0	1	1	0
21	43	4	0	1	15	0	0	0
22	41	4	0	1	10	0	0	1
23	40	1	0	1	0,25	0	0	1
24	40	2	0	1	15	0	0	0
25	40	3	0	1	3	0	0	0


Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
Cluster Analysis: K-Means Clustering: Worksheet in kag_risk_factors_cer...
Quick
Advanced
Variables:
ALL
Cluster:
Variables (columns)
OK
Cancel
Options
MD deletion
Casewise
Mean substitution


Statistica - kag_risk_factors_cervical_cancer* - [Worksheet]
k - Means Clustering Results: Worksheet in kag_risk_f...
Number of variables: 33
Number of cases: 29
K-means clustering of variables
Missing data were casewise deleted
Number of clusters: 2
Solution was obtained after 1 iterations
Quick
Advanced
Summary
Summary: Cluster means & Euclidean distances
Analysis of variance
Graph of means
Cancel
Options
By Group


Statistica - kag_risk_factors_cervical_cancer* - [Plot of Means for Each Cluster]
kag_risk_factors_cervical_cancer* - Plot of Means for Each Cluster
Plot of Means for Each Cluster
40
35
30
25
20
15
10
5
0
-5
-10
34
42
45
45
40
43
40
39
37
Cases
Cluster 1
Cluster 2
Plot of Means for Each Cluster
Ready