Міністерство освіти і науки України

**ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
БУДІВНИЦТВА ТА АРХІТЕКТУРИ**

О.О.Шаповалова

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ З ПРАКТИКУМОМ В DEDUCTOR

з дисципліни «Інтелектуальний аналіз даних»

Рекомендовано науково-методичною радою університету
як навчально-методичний посібник
для студентів спеціальностей:

122 “Комп’ютерні науки”
126 “Інформаційні системи та технології”

Харків 2020

УДК 519
Ш-67
Рецензенти:

А.Ю. Гайдусь, к.т.н. доцент кафедри ФТМДМ Харківського національного технічного університету сільського господарства ім. П. Василенка

О.Г. Ніколаєва, к. фіз.-мат.н. доцент кафедри економічної кібернетики та прикладної економіки Харківського національного університету ім. Н.В.Каразіна

Г.В. Солодовник, к.т.н., доц. кафедри комп'ютерних наук та інформаційних технологій ХНУБА

Рекомендовано кафедрою комп'ютерних наук та інформаційних технологій, протокол № 6 від 25. 05 . 2020р.

Затверджено науково-методичною радою університету,
протокол № 6 від 17.09. 2020р.

Ш-67 О.О. Шаповалова Інтелектуальний аналіз даних з практикумом в Deductor: Навчально-методичний посібник. – Х.: ХНУБА, 2020. – 160 с.

В навчально-методичному посібнику подається теоретичний матеріал та практичні роботи з дисципліни «Інтелектуальний аналіз даних», зокрема з застосуванням аналітичної платформи Deductor. Висвітлено концептуальні основи аналізу даних на підґрунті алгоритмів штучного інтелекту з застосуванням можливостей комп'ютерних наук та інформаційних технологій. Подання матеріалу організовано так, що методи інтелектуального аналізу даних викладаються з використанням елементарних понять та супроводжується практичними прикладами з різних галузей.

Для розв'язання практичних задач залучено аналітичну платформу Deductor, яка має сучасний інтерфейс і розвинену структуру та орієнтована на інтелектуальний аналіз даних з застосуванням методів вилучення, маніпулювання та візуалізації, а також кластеризації даних та нейромережевих технологій. Кожна із лабораторних робіт супроводжується теоретичним матеріалом, тестовими прикладами, питаннями для самоперевірки знань.

Призначено для здобувачів вищої освіти спеціальностей: 122 "Комп'ютерні науки", 126 "Інформаційні системи та технології".

Іл.:192; табл.: 15; бібліогр.: 40 назв.

ВСТУП

Розроблений навчальний посібник входить до циклу методичного забезпечення з дисципліни «Інтелектуальний аналіз даних», яка викладається здобувачам вищої освіти першого (бакалаврського) рівня спеціальностей 126 «Інформаційні системи та технології» та 122 «Комп'ютерні науки» і є логічним продовженням матеріалу, що викладається в курсі «Штучний інтелект».

Розробка посібника обумовлена тим, що за останні десятиріччя в науковому середовищі спостерігається стрімке зростання кількості робіт, присвячених використанню методів та моделей штучного інтелекту в інформаційних системах. До того ж, застосування комп'ютерних технологій при обробці та підготовці великих масивів статистичних даних, які необхідні для застосування алгоритмів інтелектуального аналізу даних, дозволяє скоротити час моделювання та підвищити ефективність процесів отримання нових знань.

При створені баз даних як побічний продукт обробки накопиченого людством та цифрованого матеріалу створюються великі масиви невпорядкованих даних, які в «сирому» вигляді містять інформацію та закономірності, які не завжди наочні для людини, розум якої не пристосований для сприйняття великих масивів різнорідної інформації. Середньо пересічна людина, як правило, не здатна розпізнати більше двох-трьох взаємозв'язків навіть у невеликих вибірках. Математичних апарат традиційної статистики, що довгий час претендувала на роль основного інструмента аналізу даних, також не здатна виявити приховані взаємозв'язки при розв'язанні реальних задач, зокрема, коли характер таких зв'язків наперед невідомий. Вона оперує усередненими характеристиками вибірки, які часто є фіктивними величинами.

Метою використання технології *Data Mining* є вилучення прихованих правил та закономірностей у великих масивах даних, зокрема у випадках, коли характер зв'язку між змінними є наперед невідомим. *Data Mining (DM)* - це сукупність великого числа різних методів вилучення знань. У навчальному посібнику розглядаються такі методи технології *Data Mining* як асоціація, класифікація, кластеризація, прогнозування, нейронні мережі, дерева рішень тощо[1]. Наведені теоретичні положення та практичні приклади використання технології *Data Mining* для розв'язання сучасних завдань розроблені з урахуванням реалій сьогодення та дозволять майбутнім спеціалістам в галузі комп'ютерних наук та інформаційних систем використовувати отримані знання при розробці програмного забезпечення та проектуванні складних інформаційних комплексів.

Як інструментальний засіб для реалізації методики аналізу *Data Mining, Knowldge Discovery in Databases (KDD)* та *OLAP*, обрано аналітичну платформу *Deductor*. Опис *Deductor Academic* – технологічної платформи для створення закінчених аналітичних рішень – базується на технічній

документації, яка знаходиться в відкритому доступі на офіційному сайті компанії BaseGroup Labs (http://www.basegroup.ru) і є безкоштовною для некомерційного навчального використання і освітніх цілей [2].

Навчальний посібник містить опис сучасних інформаційних технологій, які дозволяють вирішувати складні інженерні задачі різного характеру та пов'язані з аналізом даних та прийняттям обґрунтованих управлінських рішень[1]. Посібник містить лабораторні роботи, які дозволяють отримати навички інтелектуального аналізу даних роботи з використанням аналітичної платформи *Deductor* та допомогти здобувачам вищої освіти спеціальностей «Інформаційні системи та технології» та «Комп'ютерні науки» отримати необхідні знання щодо виявлення нових знань на підґрунті великих багатовимірних масивів інформації.

Автор ставив перед собою завдання не тільки надати студентам знання щодо теоретичних основ дисципліни, що вивчається, але й уможливити набуття певних практичних навичок та вмінь щодо використання сучасних інтелектуальних інформаційних технологій для розв'язання навчальних задач за обраною спеціальністю.

І ТЕОРЕТИЧНІ ПОЛОЖЕННЯ

1 ІНТЕЛЕКТУАЛЬНИЙ АНАЛИЗ ДАНИХ

В даний час елементи штучного інтелекту активно впроваджуються в практичну діяльність. На відміну від традиційних систем штучного інтелекту технологія інтелектуального пошуку і аналізу даних або "видобуток даних" (*Data Mining*) не намагається моделювати природний інтелект, а посилює його можливості потужністю сучасних обчислювальних серверів, пошукових систем і сховищ даних. Технологія *Data Mining* є одним з етапів технології «виявлення знань у базах даних» – *Knowledge Discovery in Databases*.

Класичне визначення технології «видобутку даних» (*Data Mining (Discovery-driven Data Mining*) – це виявлення у початкових даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, які необхідні для прийняття рішень в різних сферах людської діяльності.

В основу сучасної технології *Data Mining* покладена концепція шаблонів, що відображають фрагменти багатоаспектних взаємовідносин між даними. Ці шаблони являють собою закономірності, притаманні вибіркам даних, що можуть бути стисло виражені в зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень щодо структури вибірки у вигляді розподілів значень аналізованих показників [3].

1.1 Методи та алгоритми *Data Mining*

Data Mining є мультидисциплінарною галуззю, що виникла і розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних тощо.

У технології *Data Mining* можна виділити наступні класи методів і алгоритмів [4,5].

1.1.1 Статистичні методи часто використовують нескладний статистичний апарат, але максимально враховують сформовану в своїй області специфіку (професійна мова, системи різних індексів тощо) [6].

Недоліком систем цього класу вважають вимогу до спеціальної підготовки користувача. Також відзначають, що потужні сучасні статистичні пакети є занадто «великоваговими» для масового застосування у бізнесі [7]. Є ще більш серйозний принциповий недолік статистичних пакетів, що обмежує їх застосування в *Data Mining*. Більшість методів, що входять до складу пакетів, спираються на статистичну парадигму, згідно з якою головними фігурантами служать усереднені характеристики вибірки. Нажаль, ці характеристики при дослідженні реальних процесів виявляються неефективними [4].

1.1.2 Дерева рішень – це метод представлення правил в ієрархічній, послідовній структурі, де кожному об'єкту відповідає єдиний вузол, що дає

рішення [4]. Під правилом мають на увазі логічну конструкцію, вигляду «якщо ... то ...».

Область застосування дерева рішень у даний час є достатньо широкою, але всі задачі, які вирішуються цим апаратом, можуть бути об'єднані в наступні три класи [8]:

1) опис даних, тобто зберігання інформації про дані в компактній формі, яка містить точний опис об'єктів;

2) класифікація, тобто віднесення об'єктів до одного із наперед відомих класів;

3) регресія, тобто встановлення залежності досліджуваної змінної (цільової функції) від незалежних (вхідних) змінних у разі, коли вона має безперервні значення.

На сьогоднішній день існує значна кількість алгоритмів, що реалізують побудову дерев рішень – CART, C4.5, NewId, ITrule, CHAID, CN2 тощо, – але найбільшого поширення і популярності отримали наступні два:

– алгоритм CART (*Classification and Regression Tree*) – алгоритм побудови бінарного дерева рішень дихотомічної класифікаційної моделі. Кожен вузол дерева при розбитті має тільки двох нащадків. Як видно з назви, алгоритм вирішує завдання класифікації і регресії;

– алгоритм C4.5 – алгоритм побудови дерева рішень, при цьому кількість нащадків вузла необмежена. Цей алгоритм не вміє працювати з безперервним цільовим полем, тому вирішує тільки задачі класифікації [8].

До основних переваг використання дерев рішень відносять:

– швидкий процес навчання;

– генерація правил в галузях, де експерту важко формалізувати свої знання;

– формулювання правил на природній мові;

– інтуїтивно зрозуміла класифікаційна модель;

– висока в порівнянні з іншими методами (статистика, нейронні мережі) точність прогнозу;

– побудова непараметричних моделей.

Дерева рішень є гарним інструментом в системах підтримки прийняття рішень, інтелектуального аналізу даних (*Data Mining*) [8,10].

1.1.3 Ідея прийняття рішень на основі аналогічних випадків (*case based reasoning - CBR)* наступна: для того, щоб зробити прогноз на майбутнє чи обрати найкраще рішення *CBR* системи знаходять у минулому найближчий аналог наявної ситуації і обирають ту ж відповідь, яка вважалася кращою в аналогічному випадку. Цей метод ще називають методом «найближчого сусіда» (*nearest neighbour*). Останнім часом поширення отримав також термін *memory based reasoning*, який акцентує увагу на тому, що рішення приймається на підставі всієї інформації, накопиченої в пам'яті [4].

Головним мінусом систем *CBR* вважають те, що вони взагалі не створюють будь-яких моделей чи правил, узагальнюючих попередній досвід. У виборі рішення ці системи ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на основі яких конкретно факторів *CBR*-системи будують свої відповіді.

1.1.4 Алгоритми обмеженого перебору обчислюють частоти комбінацій простих логічних подій у підгрупах даних [4,11]. Обмеженнями служить довжина комбінації простих логічних подій. На підставі аналізу обчислених частот робиться висновок про корисність тієї чи іншої комбінації для встановлення асоціації в даних, для класифікації, прогнозування і т.п.

1.1.5 Нейронні мережі – це великий клас систем, архітектура яких є аналогічною побудові нервової тканини з нейронів. В одній з найбільш поширених архітектур – багатошаровому персептроні зі зворотним поширенням помилки – імітується робота нейронів у складі ієрархічної мережі, де кожен нейрон більш високого рівня з'єднаний своїми входами з виходами нейронів нижчого шару. На нейрони найнижчого шару подаються значення вхідних параметрів, на основі яких приймається рішення, прогнозується розвиток ситуації і т. д. Ці значення розглядаються як сигнали, що передаються в наступний шар, ослаблюючись чи посилюючись в залежності від числових значень (ваг), які приписані міжнейронним зв'язкам. У результаті на виході нейрона самого верхнього шару виробляється деяке значення, яке розглядається як відповідь – реакція всієї мережі на введені значення вхідних параметрів. Для того, щоб мережу можна було застосовувати в подальшому, її спочатку треба «натренувати» на отриманих раніше даних, для яких відомі як значення вхідних параметрів, так і правильні відповіді на них. Тренування полягає в підборі ваг міжнейронних зв'язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей [12,13].

Основним недоліком нейромережевої парадигми є необхідність мати значний обсяг вибірки, що навчає. Інший суттєвий недолік полягає в тому, що навіть натренована нейронна мережа є так званою «чорною скринькою»., вміст якої невідомий. Знання, зафіксовані як ваги кількох сотень міжнейронних зв'язків, не можуть бути проаналізовані та інтерпретовані людиною [5].

1.1.6 Генетичні алгоритми призначені для вирішення задач оптимізації. Прикладом подібної задачі може служити навчання нейромережі, тобто підбір таких значень ваг, при яких досягається мінімальна помилка. При цьому в основі генетичного алгоритму лежить метод випадкового пошуку. Основним недоліком випадкового пошуку є те, що наперед невідомо, скільки часу знадобиться для вирішення задачі. Щоб уникнути значних витрат часу при розв’язанні задачі, застосовуються методи, що проявилися в біології: методи, відкриті при вивченні еволюції, і походження видів.

Для вирішення задач представлення об'єктів необхідно представити кожну ознаку об'єкта у формі, придатній для використання в генетичному алгоритмі.

Все подальше функціонування механізмів генетичного алгоритму виконується на рівні фенотипу, дозволяючи обійтися без інформації про внутрішню структуру об'єкта, що й обумовлює його широке застосування в різноманітних задачах.

У різновиді генетичного алгоритму, що найбільш часто зустрічається, для представлення фенотипу об'єкта застосовуються бітові рядки. При цьому кожному атрибуту об'єкта в фенотипі відповідає один ген у фенотипі об'єкта. Ген є бітовим рядком, найчастіше фіксованої довжини, який представляє значення цієї ознаки [5, 14].

1.1.7 Еволюційне програмування засноване на формулюванні гіпотези про вид залежності цільової змінної від інших змінних у вигляді програм на деякій внутрішній мові програмування. Процес побудови програм будується як еволюція у світі програм (цим підхід трохи схожий на генетичні алгоритми). При знаходженні системою програми, що більш-менш задовільно виражає шукану залежність, вона починає вносити в неї невеликі модифікації і відбирає серед побудованих дочірніх програм ті, які підвищують точність. Таким чином, система «вирощує» кілька генетичних ліній програм, що конкурують між собою в точності висловлювання шуканої залежності. Спеціальний модуль переводить знайдені залежності з внутрішньої мови системи на мову, зрозумілу користувачу (математичні формули, таблиці тощо) [5].

Інший напрямок еволюційного програмування пов'язаний з пошуком залежності цільових змінних від інших у формі функцій певного виду. Наприклад, в одному з найбільш вдалих алгоритмів цього типу – методі групового урахування аргументів (МГУА) – залежність шукають у формі поліномів.

1.1.8 Засоби для візуалізації багатовимірних даних підтримуються всіма системами *Data Mining*. У подібних системах увага приділяється зручності користувацького інтерфейсу, що дозволяє асоціювати з аналізованими показниками різні параметри діаграми розсіювання об'єктів (записів) бази даних. До таких параметрів відносяться колір, форма, орієнтація щодо власної вісі, розміри та інші властивості графічних елементів зображення. Крім того, системи візуалізації даних забезпечені зручними засобами для масштабування і обертання зображень [4,15-17].

1.2 Задачі, які вирішуються методами *Data Mining*

Усі задачі, які вирішуються методами *Data Mining*, умовно поділяються на п'ять класів [17-20].

1)Класифікація (*Classification*) – встановлення функціональної залежності між вхідними і дискретними вихідними змінними. За допомогою класифікації вирішується задача віднесення об'єктів (спостережень, подій) до

одного з наперед відомих класів. Це робиться за допомогою аналізу вже класифікованих об'єктів і формулювання деякого набору правил. Класифікація використовується у випадку, якщо є можливість виділити класи віднесення об'єктів.

2) Кластеризація (*Clustering*) – це групування об'єктів (спостережень, подій) на основі даних (властивостей), що описують сутність об'єктів. Об'єкти усередині кластера повинні бути «схожими» один на одного і відрізнятися від об'єктів, які увійшли в інші кластери. Чим більш схожі об'єкти усередині кластера і чим більше відмінностей між кластерами, тим точніша кластеризація. Інколи, якщо мова йде про економічні об’єкти, замість кластеризації вживають термін сегментація.

3) Регресія (*Regression)* – встановлення функціональної залежності між вхідними і безперервними вихідними змінними. Прогнозування найчастіше зводиться до вирішення задачі регресії. До цього ж типу задач відноситься прогнозування часового ряду на основі історичних даних.

4) Асоціація (*Associations*) – виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події *X* витікає подія *Y*. Такі правила називаються асоціативними. Вперше ця задача була запропонована для знаходження типових шаблонів покупок, здійснених в супермаркетах, тому іноді її ще називають аналізом ринкової корзини (*market basket analysis*).

5) Послідовні шаблони (*Sequence*) – встановлення закономірностей між пов'язаними в часі подіями. Послідовні шаблони можуть бути використані при плануванні продажів або надання послуг.

2 АНАЛІТИЧНА ПЛАТФОРМА DEDUCTOR

Технологію *Data Mining* можна вважати частиною більш широкого процесу, званого пошуком знань (*Knowledge Discovery in Databases*). *KDD* включає в себе питання підготовки даних, вибору інформативних ознак, очищення даних, застосування методів *Data Mining*, постобробки даних та інтерпретації отриманих результатів.

Так як навчальний посібник призначено для навчальних цілей, то в якості пакета інструментів *Data Mining* обрано аналітичну платформу *Deductor Academic*, яка є безкоштовним програмним продуктом, що надається для використання в навчальних закладах з метою освіти та навчання [2].

Аналітична платформа – спеціалізоване програмне рішення (або набір рішень), які містять всі інструменти для вилучення закономірностей з «сирих» даних: засоби консолідації інформації в єдиному джерелі (сховище даних), вилучення, перетворення, трансформації даних, алгоритми *Data Mining*, засоби візуалізації та поширення результатів серед користувачів, а також можливості конвеєрної обробки даних.

2.1 Задачі, які вирішуються в *Deductor*

Реалізовані в *Deductor* технології можуть використовуватися як в комплексі, так і окремо для вирішення широкого спектру бізнес-проблем [2]:

– системи корпоративної звітності. Готове сховище даних і гнучкі механізми передобробки, очищення, завантаження, візуалізації дозволяють швидко створювати системи звітності в стислі терміни;

– обробка нерегламентованих запитів. Кінцевий користувач може з легкістю отримати відповідь на питання типу "Скільки було продажів товару по групах в Харківській області за минулий рік з розбивкою за місяцями?" і переглянути результати найбільш зручним для нього способом;

– аналіз тенденцій і закономірностей, планування, ранжирування. Простота використання і інтуїтивно зрозуміла модель даних дозволяє проводити аналіз за принципом "що–якщо", співвідносити гіпотези з відомостями, що зберігаються в базі даних, знаходити аномальні значення, оцінювати наслідки прийняття бізнес рішень;

– прогнозування. Якщо побудувати модель на історичних прикладах, можна використовувати її для прогнозування ситуації в майбутньому. При зміні ситуації немає необхідності перебудовувати все, необхідно всього лише довчити модель;

– управління ризиками. Реалізовані в системі алгоритми дозволять достатньо точно визначитися з тим, які характеристики об'єктів і як впливають на ризики, завдяки чому можна прогнозувати настання ризикової події і завчасно вживати необхідні заходи щодо зниження розміру можливих несприятливих наслідків;

– аналіз даних маркетингових і соціологічних досліджень. Аналізуючи відомості про споживачів, можна визначити, хто є клієнтом фірми і чому, як змінюються їх уподобання в залежності від віку, освіти, соціального статусу, матеріального стану та інших показників. Розуміння цього сприятиме правильному позиціонуванню продуктів і стимулюванню продажів;

– діагностика. Механізми аналізу, наявні в системі *Deductor*, з успіхом застосовуються в медичній діагностиці та діагностиці складного обладнання;

– виявлення об'єктів на основі нечітких критеріїв. Часто трапляється ситуація, коли необхідно виявити об'єкт, ґрунтуючись не на чітких критеріях, таких, як вартість, технічні характеристики продукту, а на розмитих формулюваннях, наприклад, знайти продукти, схожі на ваші з точки зору споживача.

Це тільки невеликий список задач, які вирішуються в *Deductor*. Фактично мова йде про будь-які задачі, де потрібно консолідувати дані, відобразити їх різними способами, побудувати моделі і застосувати отримані моделі до нових даних.

2.2 Алгоритми, що використовуються в *Deductor*

Аналітична платформа *Deductor* є яскравим представником *Business Intelligence (BI)*-систем [2]. *BI* – це технологія ведення "інтелектуального" бізнесу. Основні функції *Business Intelligence* систем:

– введення даних (завантаження);
– зберігання даних (сховище даних);
– аналіз даних (*OLAP*, *Data Mining*).

Аналітична платформа *Deductor* є основою для створення закінчених прикладних рішень. Реалізовані в *Deductor* технології дозволяють на базі єдиної архітектури пройти всі етапи побудови аналітичної системи – від створення сховища даних до автоматичного підбору моделей і візуалізації отриманих результатів.

Deductor надає аналітикам інструментальні засоби, необхідні для вирішення найрізноманітніших аналітичних завдань: корпоративна звітність, прогнозування, сегментація, пошук закономірностей та інші.

На рис.1 представлені аналітичні алгоритми, які використані в *Deductor* і які згруповані за призначенням [21].

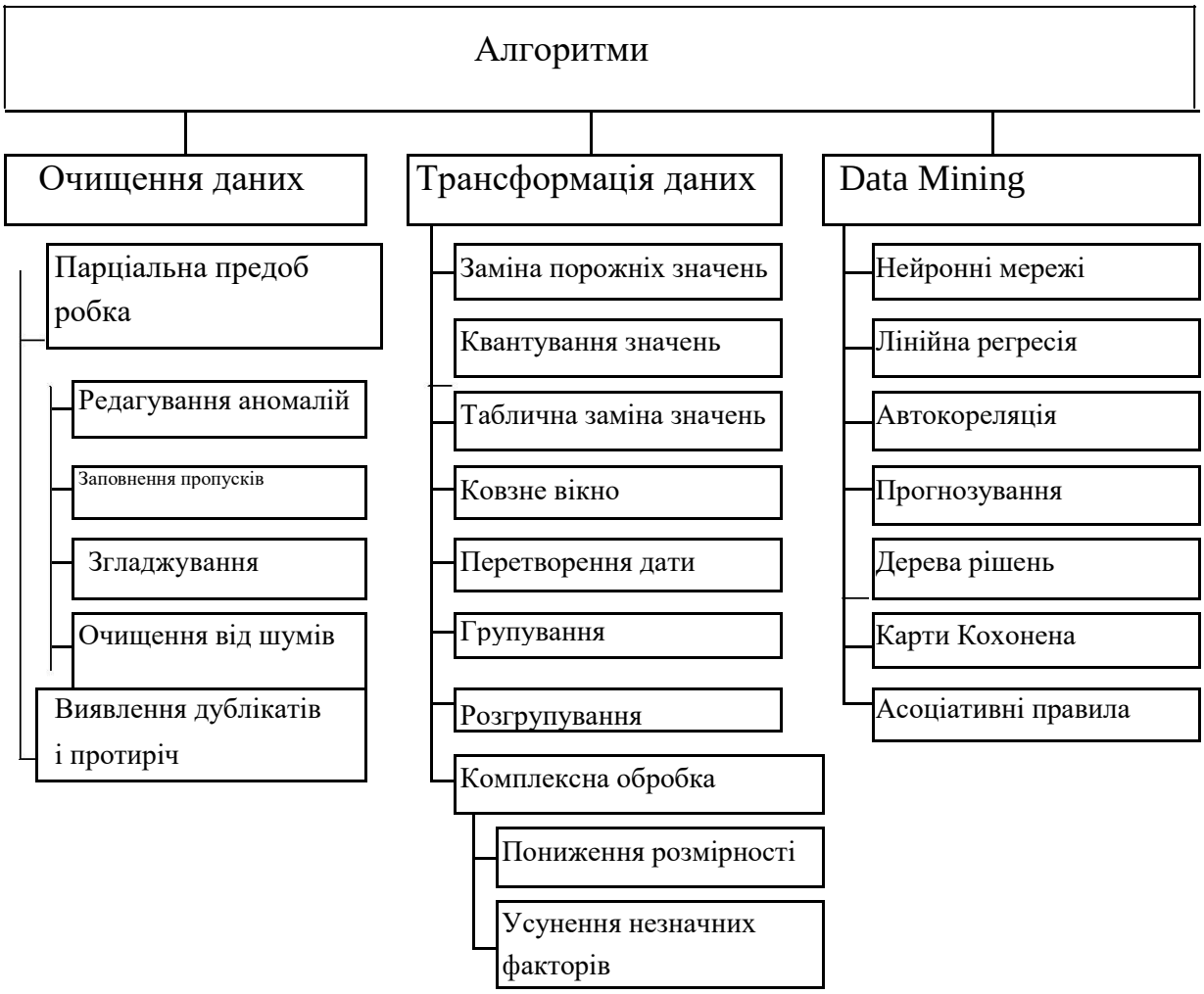


Рисунок 1 – Алгоритми, що використовуються в *Deductor*

На етапі *Очищення даних* проводиться редагування аномалій, заповнення пропусків, згладжування, очищення від шумів, виявлення дублікатів і протиріч.

Автоматичне *редагування аномальних значень* здійснюється із застосуванням методів робастної фільтрації, в основі яких лежить використання робастних статистичних оцінок, таких, наприклад, як медіана. При цьому можна задати емпірично підібраний критерій того, що вважати аномалією. Наприклад, завдання в якості міри придушення аномальних даних значення "слабка" означає найбільш толерантне ставлення до величини допустимих викидів.

У програмі передбачено два способи *заповнення відсутніх даних*:

– апроксимація, тобто відсутні дані відновлюються методом апроксимації;

– максимальна правдоподібність, тобто алгоритм підставляє найбільш ймовірні значення замість відсутніх даних.

Метод апроксимації рекомендується використовувати в рядах, де дані впорядковані. У цьому методі застосовується послідовний рекурентний фільтр другого порядку (фільтр Калмана). Вхідні дані послідовно подаються на вхід фільтра, і якщо чергове значення ряду відсутнє, воно замінюється значенням, яке екстраполюється фільтром.

Метод максимальної правдоподібності рекомендується застосовувати на невпорядкованих даних. При використанні цього методу будується щільність розподілу ймовірностей, і відсутні дані замінюються значенням, відповідним її максимуму.

Для *згладжування рядів даних* у програмі використовуються два алгоритми.

Перший спосіб згладжування – це низькочастотна фільтрація з використанням швидкого перетворення Фур'є. При цьому задається верхнє значення смуги частот, що пропускається. При придушенні шумів на основі аналізу розподілу складових Фур'є спектру на вихід фільтра пропускаються спектральні складові, які перевищують певний поріг, розрахований за емпіричними формулами відповідно із заданим критерієм ступеня віднімання шуму. Чим більше потрібно згладити дані, тим менше повинно бути значення смуги. Однак занадто вузька смуга може призвести до втрати корисної інформації. Слід зауважити, що цей алгоритм найбільш ефективний, якщо аналізовані дані є сумою корисного сигналу й білого шуму.

Другий спосіб згладжування – це вейвлет-перетворення. Якщо обрано цей метод, то необхідно задати глибину розкладання і порядок вейвлета. "Масштаб" відсіяних деталей залежить від глибини розкладання: чим більше ця величина, тим більш "великі" деталі у початкових даних будуть відкинуті. При досить великих значеннях параметра (порядку 7-9) виконується не тільки очищення даних від шуму, але і їх згладжування ("обрізаються" різкі викиди). Використання занадто великих значень глибини розкладання може призвести

до втрати корисної інформації через занадто високого ступеня "огрублення" даних. Порядок вейвлета визначає гладкість відновленого ряду даних: чим менше значення параметра, тим яскравіше будуть виражені "викиди", і навпаки – при великих значення параметра "викиди" будуть згладжені.

При виборі режиму *очищення від шумів* необхідно задати ступінь віднімання шуму: малий, середній чи великий. При використанні віднімання шуму слід дотримуватися обережності, тому реалізований евристичний алгоритм гарантує задовільні результати лише при виконанні двох умов:

– дисперсія шуму значно менше енергії корисного сигналу;
– шум має нормальний розподіл.

Суть обробки в режимі *виявлення дублікатів і протиріч* полягає в тому, що визначаються вхідні і вихідні поля. Алгоритм шукає у всьому наборі записи, для яких однаковим вхідним полям відповідають однакові (дублікати) або різні (суперечності) вихідні поля. На підставі цієї інформації створюються два додаткових логічних поля – *Дублікат* і *Протиріччя*, які приймають значення *Істина* або *Хибно*.

На етапі *Трансформація даних* проводиться заміна порожніх значень, квантування, таблична заміна значень, перетворення до ковзного вікна, зміна формату набору даних.

Інформація, яка аналізується, представлена у вигляді набору даних і має певний формат. Для аналізу різних аспектів інформації може знадобитися зміна її формату, або *трансформація*. Трансформація даних складається з трьох етапів, які виконуються у строгій послідовності (кожен з яких, однак, може бути відсутнім).

При виконанні операції *Квантування значень* здійснюється розбивка діапазону числових значень на вказану кількість інтервалів певним методом і заміна кожного значення, що обробляється, на число, яке пов'язане з інтервалом, до якого воно відноситься, або на мітку інтервалу. Інтервали розбиття включають в себе нижню межу, але не включають верхню, крім останнього інтервалу, який включає в себе обидві межі. Результатом перетворення може бути: номер інтервалу (від нуля до значення, на одиницю меншого кількості інтервалів), значення нижньої або верхньої межі інтервалу розбиття, середнє значення інтервалу розбиття, мітка інтервалу. Квантування може бути здійснено інтервальним або квантільним методом.

Інтервальне квантування здійснює розбивку діапазону значень на вказану кількість значень рівної довжини. Наприклад, якщо значення в полі потрапляють в діапазон від 0 до 10, то при інтервальному квантуванні на 10 інтервалів отримаємо відрізки від 0 до 1, від 1 до 2 і т.д. При цьому 0 буде відноситися до першого інтервалу, 1 – до другого, а 9 і 10 – до десятого.

Квантільне квантування здійснює розбивку діапазону значень на рівноймовірні інтервали, тобто на інтервали, що містять рівну (або, принаймні, приблизно рівну) кількість значень. Порушення рівності можливе тільки тоді, коли значення, що потрапляють на межу інтервалу, зустрічаються

в наборі даних кілька разів. У цьому випадку всі значення відносяться до одного певного інтервалу і можуть викликати "перевагу" у його бік.

В результаті виконання операції *Таблична заміна значень* проводиться заміна значень за таблицею підстановки, яка містить пари, що складаються з початкового і вихідного значення. Наприклад, 0 – "червоний", 1 – "зелений", 2 – "синій". Для кожного значення початкового набору даних шукається відповідність серед вихідних значень таблиці підстановки. Якщо відповідність знайдено, то значення змінюється на відповідне вихідне значення з таблиці підстановки. Якщо значення не знайдено в таблиці, воно може бути або замінено значенням, зазначеним для заміни "за замовчуванням", або залишено без змін (якщо таке значення не вказане).

При вирішенні деяких задач, наприклад, при прогнозуванні часових рядів за допомогою нейромережі, потрібно подавати на вхід аналізатора значення кількох суміжних відліків з початкового набору даних. Такий метод відбору даних називається *Ковзним вікном* (вікно – оскільки виділяється тільки деяка безперервна ділянка даних, ковзне – оскільки це вікно "переміщюється, ковзає" вздовж набору). При цьому ефективність реалізації помітно підвищується, якщо не обирати дані кожного разу з декількох послідовних записів, а послідовно розташувати дані, що відносяться до конкретної позиції вікна, в одному записі.

Перетворення дати та розбиття дати необхідно для аналізу показників за певний період (день, тиждень, місяць, квартал, рік). Суть розбиття полягає в тому, що на основі стовпця з інформацією про дату формується інший стовпець, в якому вказується, до якого заданого інтервалу часу належить рядок даних. Тип інтервалу задається аналітиком, виходячи з того, що він хоче отримати, – дані за рік, квартал, місяць, тиждень, день або відразу по всіх інтервал.

В *Deductor Studio* передбачений інструмент, який реалізує збирання узагальненої інформації – *Угрупування*, який дозволяє об'єднувати записи за полями-вимірами з агрегацією даних в полях-фактах для подальшого аналізу.

Угрупування використовується для об'єднання фактів за вимірами. При цьому під об'єднанням розуміється застосування деякої функції агрегації. Якщо у початковому наборі даних були присутні інші виміри, то втрачається інформація про значення фактів у розрізі цих вимірів. Алгоритм *розгрупування* дозволяє відновити ці факти, але їх значення відновлюються не точно, а пропорційно внеску в згруповані значення.

Термін "передобробка" можна трактувати ширше, а саме як процес попереднього експрес-аналізу даних. Наприклад, як оцінити, є фактор значущим чи ні, чи всі фактори враховані для пояснення поведінки результуючої величини і так далі. Для цих цілей використовуються такі алгоритми як *кореляційний аналіз, факторний аналіз, метод головних компонентів, регресійний аналіз*. Подібний аналіз у *Deductor Studio* називається *комплексною передобробкою*, в рамках якої здійснюється зниження розмірності вхідних даних та / або усунення незначущих факторів.

Зниження розмірності простору факторів необхідно у випадках, коли вхідні фактори корельовані один з одним, тобто взаємозалежні. Є можливість перерахувати їх в іншу систему координат, виділяючи при цьому головні компоненти. Пониження розмірності здійснюється шляхом відкидання компонент, що в найменшій мірі пояснюють дисперсію результуючих значень (при цьому передбачається, що початкові фактори повністю пояснюють дисперсію результуючих факторів).

Потрібно вказати поріг значущості, який задає дисперсію результату. Значення порогу значущості може змінюватися від 0 до 1.

Усунення незначущих факторів засноване на пошуку таких значень, які найменшою мірою корельовані (взаємопов'язані) з вихідним результатом. Такі фактори можуть бути виключені з результуючого набору даних практично без втрати корисної інформації. Критерієм прийняття рішення про виключення є поріг значущості. Якщо кореляція (ступінь взаємозалежності) між вхідним і вихідним факторами менше порога значущості, то відповідний фактор відкидається як незначущий.

На етапі *Data Mining* будуються моделі з використанням нейронних мереж, дерев рішень, карт, що самоорганізуються, асоціативних правил.

2.3 Опис аналітичної платформи *Deductor*

Аналітична платформа Deductor складається з п'яти частин [2]:

1 *Deductor Warehouse* – багатовимірне сховище даних, що акумулює всю необхідну для аналізу предметної області інформацію. Використання єдиного сховища дозволяє забезпечити несуперечність даних, їх централізоване зберігання, і автоматично забезпечує всю необхідну підтримку процесу аналізу даних. *Deductor Warehouse* оптимізований для вирішення саме аналітичних задач, що позитивно позначається на швидкості доступу до даних, орієнтований на вирішення задач консолідації інформації з різнорідних джерел та швидкого вилучення потрібного набору даних.

2 *Deductor Studio* – аналітичний додаток, програма, що реалізує функції імпорту, обробки, візуалізації і експорту даних. *Deductor Studio* може функціонувати і без сховища даних, отримуючи інформацію з будь-яких інших джерел, але найбільш оптимальним є їх сумісне використання. В *Deductor Studio* включений повний набір механізмів, що дозволяє отримати інформацію з довільного джерела даних, провести весь цикл обробки (очищення, трансформацію даних, побудову моделей), відобразити отримані результати найбільш зручним чином (*OLAP*, діаграми, дерева тощо) і експортувати результати. В *Deductor Studio* є панель звітів, що зовні нагадує *Провідник* в *Windows*, на якій аналітик формує ієрархічну структуру папок та в певні папки виносить посилання на вузли сценарію, що потрібні користувачу. Архітектура системи *Deductor* подана на рис.2.

3 *Deductor Viewer* – робоче місце кінцевого користувача, яке дозволяє відокремити процес побудови сценаріїв від використання вже готових моделей. Всі складні операції щодо підготовки сценаріїв обробки виконуються

аналітиками-експертами за допомогою *Deductor Studio*, а *Deductor Viewer* забезпечує користувачам простий спосіб роботи з готовими результатами, приховує від них всі складнощі побудови моделей і не пред'являє високих вимог до кваліфікації співробітників. У ньому відсутні механізми побудови сценаріїв, налаштування джерел даних та інші складності. Робота з програмою спрощена: користувач бачить налаштовану аналітиком панель звітів, обирає потрібний звіт, а програма автоматично виконує всі необхідні дії.

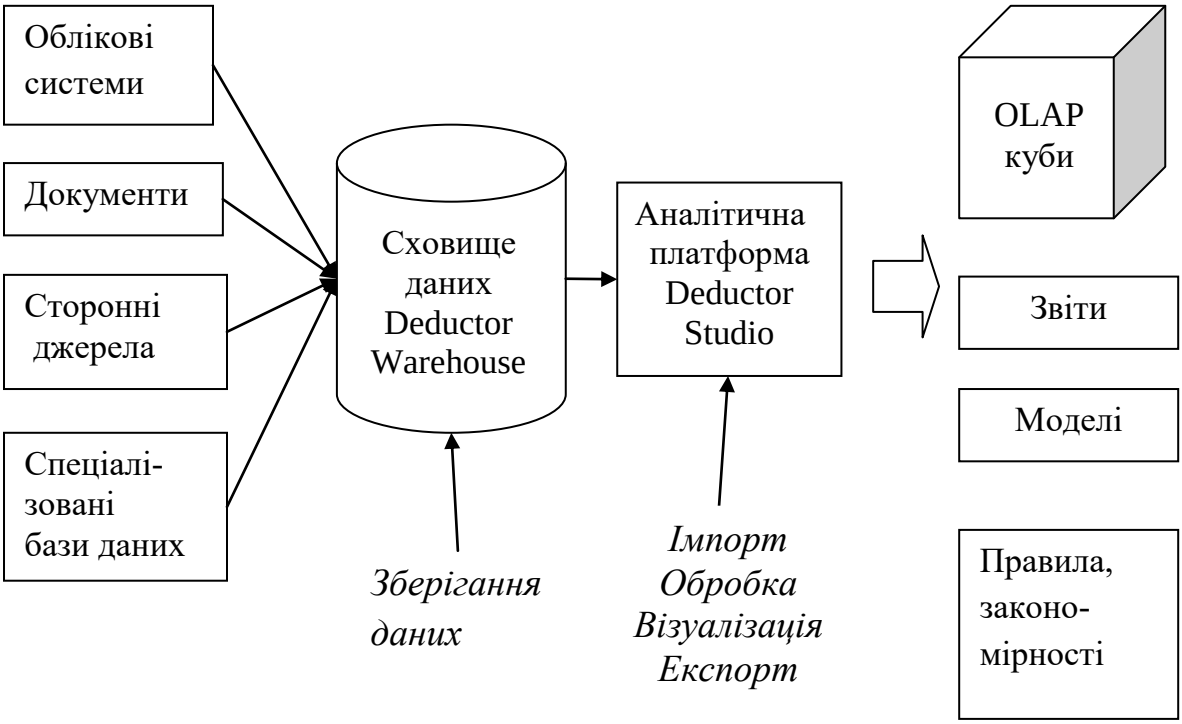


Рисунок 2 – Архітектура системи Deductor

4 *Deductor Server* – служба, що забезпечує віддалену аналітичну обробку даних, дозволяє автоматично обробляти дані і перенавчати моделі на сервері, оптимізує виконання сценаріїв за рахунок кешування проектів і використання багатопотокової обробки.

5 *Deductor Client* – програма доступу клієнта до сервера аналітичної обробки *Deductor Server* із сторонніх додатків і управління його роботою.

Реалізована в *Deductor* архітектура дозволяє досягти максимальної гнучкості при побудові остаточного варіанта рішення. Завдяки цій архітектурі можна зібрати в одному аналітичному додатку всі необхідні інструменти аналізу і реалізувати автоматичне виконання підготовленого сценарію (рис.3).

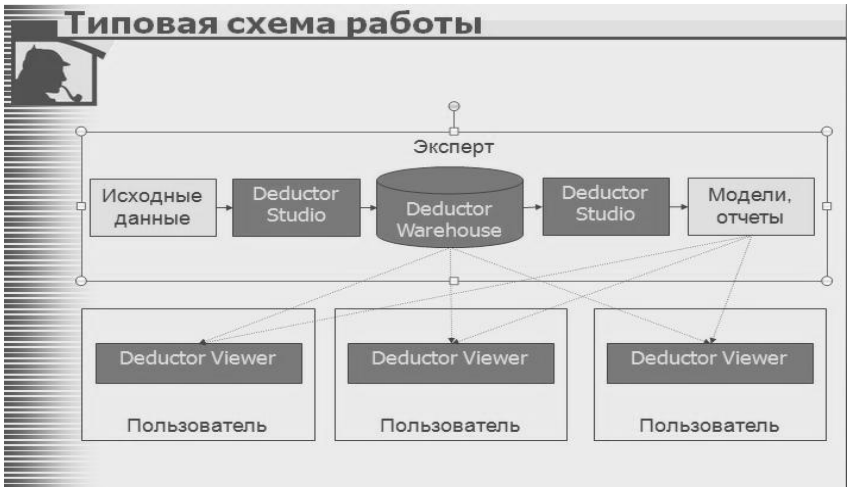


Рисунок 3 – Типова схема роботи з аналітичною платформою

Побудова остаточного варіанта рішення може бути здійснено за достатньо короткий проміжок часу. Досить лише отримати дані, визначити сценарій обробки і задати місце для експорту отриманих результатів. Наявність потужного набору механізмів обробки і візуалізації дозволяє покроково рухатися вперед від самих простих способів аналізу до більш потужних. Таким чином, перші результати користувач отримує практично відразу, але при цьому він має можливість постійно нарощувати потужність рішення і наприкінці досягти бажаного результату.

2.4 Робота з аналітичною платформою *Deductor*

Після запуску додатку головне вікно *Deductor Studio* має вигляд, наведений на рис.4 [1].

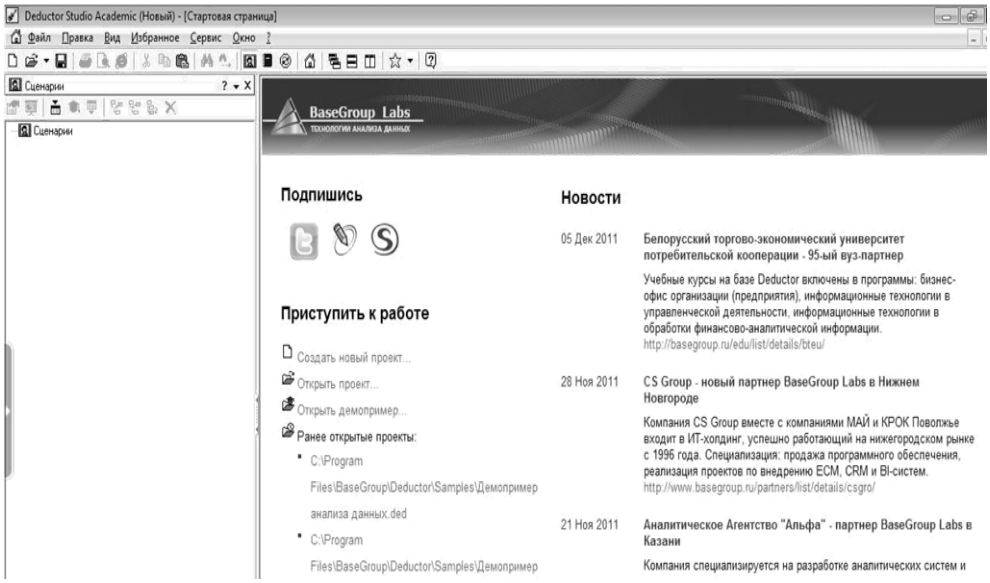


Рисунок 4 – Головне вікно *Deductor Studio*

За замовчуванням панель керування представлена однією вкладкою *Сценарії*. Крім того, досяжні ще дві вкладки: *Звіти* і *Підключення*. Зробити їх активними можна, скориставшись командами *Звіти* та *Підключення* меню *Вид*.

Ключовим поняттям *Deductor Studio* є проект, тобто файл з розширенням **. ded*, який за структурою відповідає стандартному *.xml* файлу і зберігає послідовність обробки даних (сценарії), налаштовані візуалізатори, змінні проекту і службову інформацію.

Кожен проект має відомості про автора, які заповнюються в діалоговому вікні, що відкривається командою *Файл* / *Властивості проекту*… (рис.5).

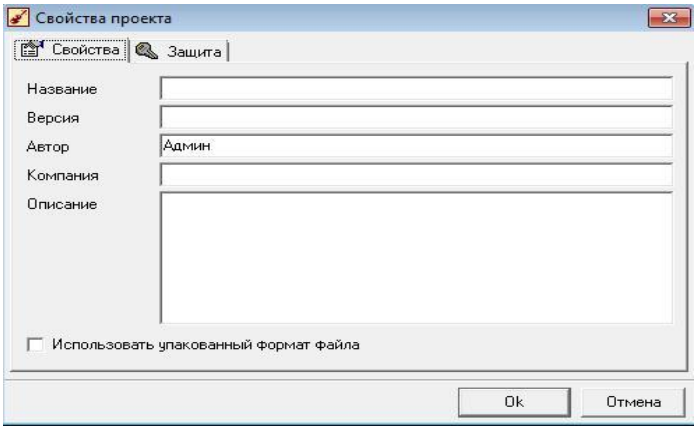


Рисунок 5 – Діалогове вікно *Властивості проекту*

Дії щодо створення нового проекту або збереження існуючого виконуються за допомогою команд меню *Файл*.

2.4.1 У *Deductor Studio* організувати всі види робіт допомагають п'ять майстрів:

- майстер імпорту;
- майстер експорту;
- майстер обробки;
- майстер візуалізації;
- майстер підключень.

За допомогою *Майстрів імпорту, експорту* та *обробки* формується сценарій, який в свою чергу складається з вузлів. *Майстер підключень* призначений для створення налаштувань підключень до різних джерел і приймачів даних. *Майстер візуалізації* налаштовує візуалізатори для конкретного вузла. *Візуалізатором* називається будь-яке представлення набору даних в будь-якому вигляді: табличному, графічному, описовому.

– *Deductor* передбачені наступні способи організації та візуалізації даних:

– *Таблиця,* тобто стандартне табличне представлення з можливістю фільтрації даних;

– *Діаграма,* тобто графік зміни будь-якого показника;

– *Статистика,* тобто статистичні показники набору даних;

– *Діаграма розсіювання,* тобто точковий графік, що ілюструє відхилення прогнозованих за допомогою моделі значень від реальних;

– *Таблиця спряженості,* тобто таблиця, яка призначена для оцінки результатів класифікації незалежно від моделі, що використовується;

– *Аналіз «Що-якщо»,* тобто таблиця і діаграма, які дозволяють спостерігати за динамікою зміни даних, що цікавлять користувача, і оцінювати вплив того чи іншого чинника на результат;

– *Вибірка, що навчає,* тобто набір даних, що використовується для побудови моделі;

– *Граф нейромережі,* тобто візуальне відображення нейромережі, що навчена;

– *Дерево рішень,* тобто графічний образ дерева рішень, отриманого за допомогою відповідного алгоритму;

– *Правила,* тобто текстовий запис правил, які отримані за допомогою алгоритму побудови дерев рішень або пошуку асоціацій;

– *Карта Кохонена,* тобто відображення карт, побудованих за допомогою відповідного алгоритму;

– *OLAP* багатовимірне представлення даних. Будь-які дані, що використовуються в програмі, можна переглянути у вигляді крос-таблиці і крос-діаграми;

– *Відомості,* тобто текстовий опис параметрів імпорту/обробки/ експорту в дереві сценаріїв обробки.

2.4.2 Для аналітика базовим інструментом в *Deductor Studio* є сценарій (рис.6).

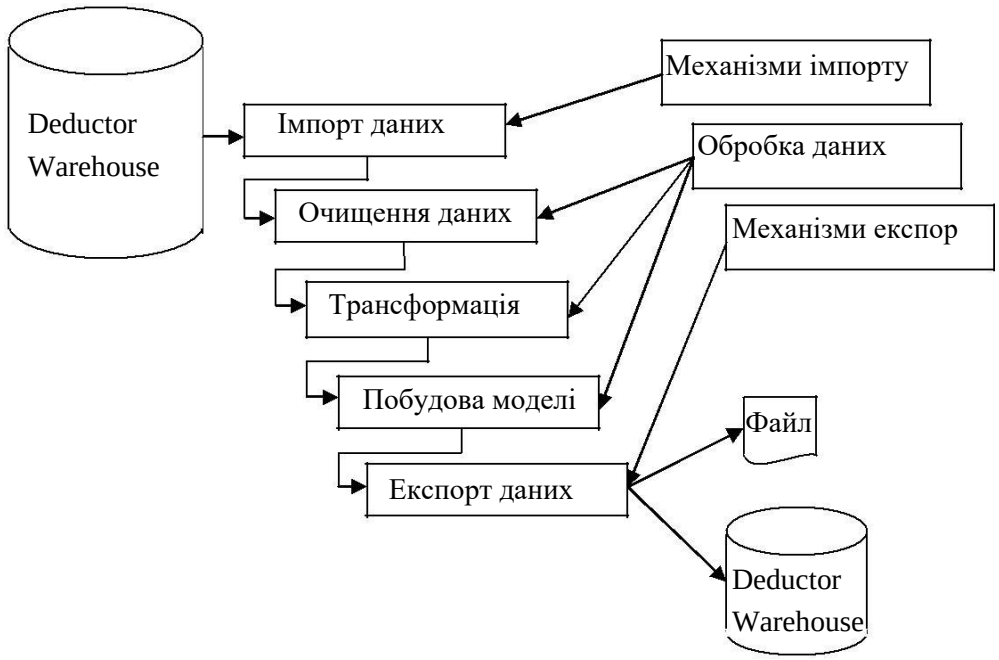


Рисунок 6 – Типовий сценарій *Deductor Studio*

Сценарій є ієрархічною послідовністю обробки і візуалізації наборів даних. Сценарій завжди починається з імпорту набору даних з довільного джерела. Після імпорту може використовуватися довільне число обробників будь-якого ступеня глибини і вкладеності. Кожній операції обробки відповідає окремий вузол дерева або об'єкт сценарію. Будь-який об'єкт можна візуалізувати тим або іншим доступним способом. Набір даних служить механізмом, що сполучає всі об'єкти сценарію. Можна сказати, що сценарій – це найбільш природний з погляду аналітика спосіб представлення етапів побудови моделі, що дозволяє швидко створювати моделі, яким притаманна гнучкість і розширюваність, та порівнювати їх між собою.

Сценарій являє собою послідовність операцій з даними, яка представлена у вигляді ієрархічного дерева. У дереві кожна операція утворює вузол, заголовок якого містить: ім'я джерела даних, найменування застосовуваного методу обробки, використані при цьому поля і т.д. Крім цього, зліва від імені вузла стоїть позначка, яка відповідає типу операції.

Сценарій складається з гілок. *Deductor* не має власних засобів для введення даних, тому сценарій завжди починається вузлом імпорту з відповідного джерела. Знову створюваний вузол імпорту буде знаходитися на верхньому рівні і буде підлеглим по відношенню до головного вузла *Сценарії*.

2.4.2.1 Створення нового вузла імпорту здійснюється за допомогою *Майстра імпорту*. За викликом *Майстра імпорту* відкриється вікно першого кроку *Майстра*, в якому всі джерела даних згруповані за наступними чотирма категоріями:

– сховища даних,
– налаштовані підключення,
– файли даних,
– бізнес-підключення.

В версії *Deductor Academic* присутня тільки категорія *Файли даних* (рис.7).

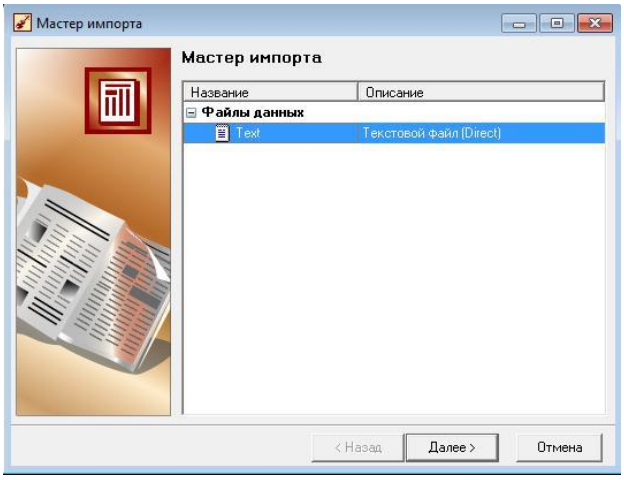


Рисунок 7 Діалогове вікно першого кроку *Майстра імпорту*

До будь-якого вузла імпорту можна додати вузол обробки або вузол експорту, попередньо виділивши вузол імпорту мишею. Новий вузол буде доданий як підлеглий до вузла імпорту.

2.4.2.2 Створення нового вузла обробки здійснюється за допомогою *Майстра обробки.* За викликом *Майстра обробки* відкриється вікно першого кроку *Майстра* (рис.8), в якому всі обробники згруповані за наступними чотирма категоріями:

– очищення даних,
– трансформація даних,
– Data Mining,
– інше.

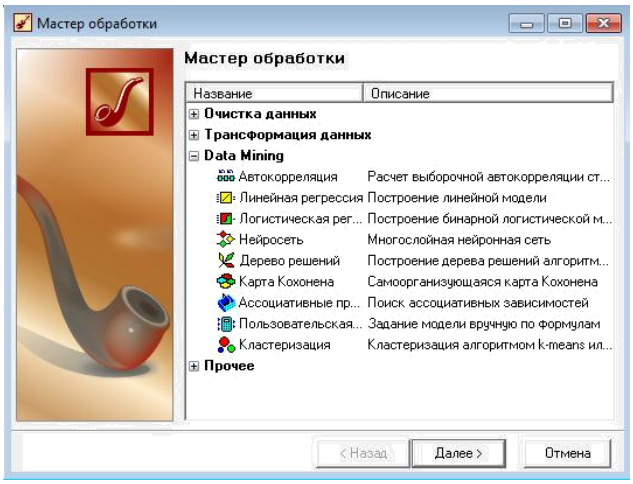


Рисунок 8 – Діалогове вікно першого кроку *Майстра обробки*

2.4.2.3 Створення нового вузла експорту здійснюється за допомогою *Майстра експорту,* у якому всі приймачі даних згруповані за наступними п’ятьма категоріями:

– сховища даних,
– бази даних,
– файли,
– Web-сервери,
– інше.

Після вузла експорту неможливо додати жоден вузол.

В версії *Deductor Academic* присутня тільки категорія *Файли.*

2.4.3 Крім команд виклику *Майстрів*, до кожного вузла можна застосувати базові операції над вузлами сценарію (рис.9).

За поданням команди *Відкрити* вузол запускається на виконання, причому виконуються всі батьківські вузли, а праворуч відкриваються візуалізатори, налаштовані для даного вузла. В інтерактивному режимі для кожного вузла повинен бути налаштований хоча б один візуалізатор, наприклад, *Таблиця* або *Відомості*.

Командою *Налаштувати…* викликається *Майстер імпорту*, *Майстер обробки* або *Майстер експорту*, для зміни параметрів обробки, яка виконується в вузлі. Тип *Майстра,* який викликається визначається типом вузла.

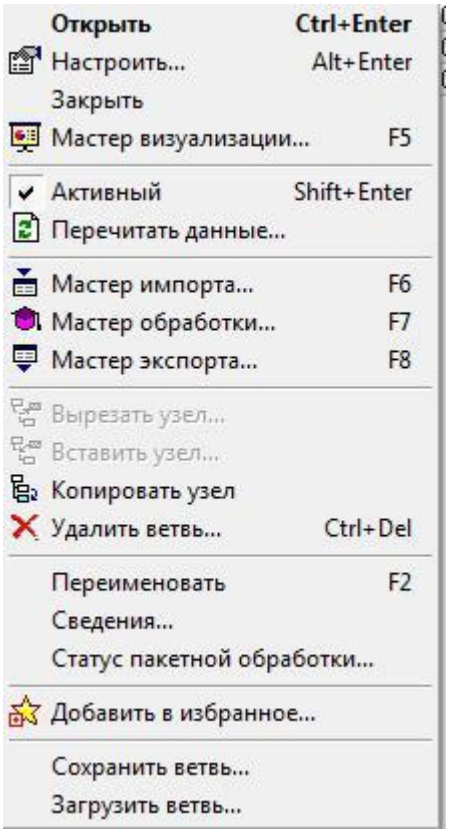


Рисунок 9 – Базові операції над вузлами сценарію

Щоб активувати вузол при роботі з операціями команди *Налаштувати… ,* слід *в*становити прапорець біля операції *Активний*. Якщо вузол неактивний, то, зробивши його активним, можна перейти до виконання сценаріїв, але візуалізатори при цьому відображені не будуть. При неактивному вузлі закриваються всі візуалізатори для нього і для всіх підлеглих вузлів.

При поданні команди *Перечитати дані…* всі вузли до кореневого включно будуть закриті, а потім виконана гілка сценарію від кореневого до поточного вузла.

Командою *Вирізати вузол…* поточний вузол видаляється з сценарію обробки, всі його нащадки при цьому переміщуються на один рівень вгору.

При поданні команди *Вставити вузол…* перед поточним вузлом сценарію вставляється новий вузол і викликається для нього *Майстер обробки*. Вставити вузол перед вузлом імпорту даних не можна.

При поданні команди *Відомості* діалогове вікно *Відомості* для поточного вузла відкривається. В діалоговому вікні редагується ім'я, мітка і опис вузла.

При поданні команди *Статус пакетної обробки* встановлюється статус пакетної обробки для вузла.

При поданні команди *Додати в Обране* поточний вузол додається до списку обраних вузлів.

При поданні команди *Завантажити гілку* викликається стандартний діалог *Відкриття файлу*, в якому можна вказати шлях і ім'я файлу, що зберігає гілку сценарію. Завантажена гілка сценарію стане нащадком поточного вузла. Гілка, що починається з вузла імпорту даних, буде додана в проект як нова коренева гілка. За замовчуванням гілка сценарію має розширення *. *deb*.

2.4.4 В *Deductor* взаємодію вузлів один з одним спроектовано на рівні програмного ядра, тому принцип взаємодії єдиний і не залежить від типу вузла.

Кожен вузол можна представити «чорною скринькою», на вхід якого подається структурований набір даних з полями, а на виході доступний один або кілька оброблених вузлом наборів даних. Може вестися будь-яка обробка від простого сортування до моделювання. Вихідний набір, в свою чергу, можна знову подати на вхід вузла. Це процедура конструювання сценарію. В *Deductor Studio* вузлами, які видають на виході більше одного набору даних, є:

- лінійна регресія,
- логістична регресія,
- асоціативні правила,
- кореляційний аналіз.

2.4.5 Структурований текстовий файл із роздільниками є одним з найпоширеніших форматів зберігання даних. Це звичайний текстовий файл, стовпці даних в якому розділені однотипними символами-роздільниками, наприклад,символами табуляції, пробілу, крапки з комою тощо.

Процес імпорту даних з текстового файлу з роздільниками в *Майстрі імпорту* (категорія Текстовой файл) містить наступні кроки:

- вибір файлу за іменем;
- налаштування параметрів імпорту;
- налаштування імпортованих полів;
- запуск процесу імпорту;
- вибір способу візуалізації;
- завдання відомостей про вузол.

На кроці *Вибір файлу за іменем* необхідно обрати ім'я текстового файлу (розширення *. *txt*, *. *csv*), з якого слід виконати імпорт даних. Після цього в полі ІМ'Я ФАЙЛА вікна *Майстра імпорту* з'явиться ім'я обраного файлу і шлях до нього.

На кроці *Налаштування параметрів імпорту* потрібно налаштувати параметри імпорту даних з текстового файлу, так як існує декілька форматів структурованих текстових файлів.

Наступне вікно *Майстра* залежить від встановленого перемикача в прапорці *Формат початкових даних*. Якщо був обраний формат з роздільниками, то з'явиться вкладка, на якій потрібно явно вказати символ-роздільник (за замовчуванням – *табуляція*).

Передперегляд текстового файлу у вигляді таблиці внизу (завантажуються тільки перші 10 рядків) дозволяє переконатися в коректності вибору налаштувань імпорту, навіть не запускаючи його.

На кроці *Налаштування параметрів стовпців* потрібно налаштувати наступні параметри стовпців імпортованих даних, вказавши відповідні значення в полях (рис.10):

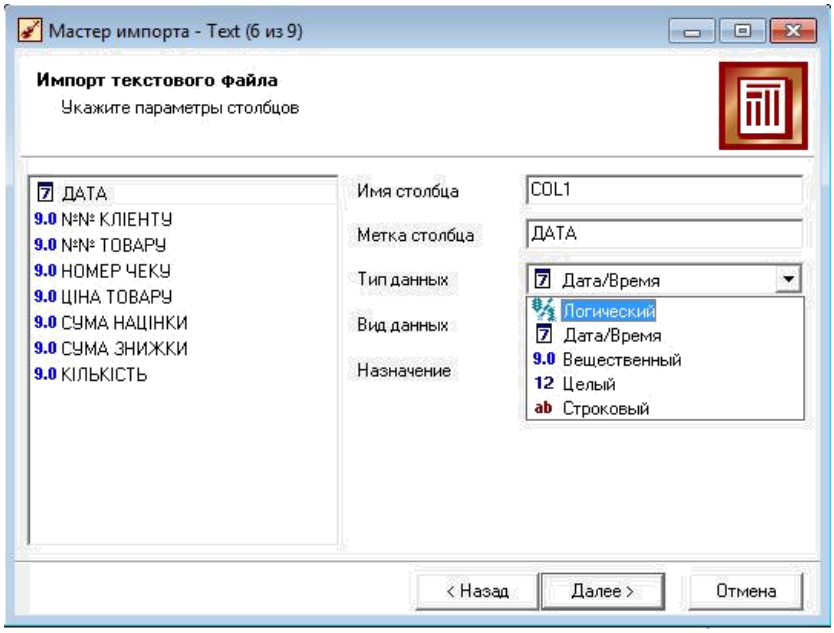


Рисунок 10 – Налаштування параметрів стовпців та типи даних

– *Ім'я стовпця,* тобто ім'я, яке буде виступати ідентифікатором стовпця в наступних вузлах. За замовчуванням пропонується заголовок стовпця з текстового файлу, якщо на попередньому кроці був встановлений прапорець *Перший рядок є заголовком*. Інакше будуть запропоновані імена типу COL1, COL2 і т.д. Можна ввести будь-які імена, які семантично відображають вміст стовпця, однак допускаються тільки латинські символи, і ім'я стовпця повинно бути унікальним в межах всіх стовпців імпортованого файлу;

– *Мітка стовпця,* тобто назва, під якою даний стовпець буде видно в візуалізаторах. Допускаються будь-які символи, унікальність імен не обов'язкова;

– *Тип даних,* тобто тип даних, що містить стовпчик. Тип обирається зі списку, що відкривається клацанням кнопкою в правій частині поля (рис.10).

– *Логічні дані* в полі можуть приймати тільки два значення 0 або 1, поле даних типу *Дата / Час* містить дані типу дата / час. *Дійсний тип* – це числа з плаваючою точкою, *Цілий тип* призначений для визначення цілих чисел, *Строковий тип* – це рядки символів ;

– *Вид даних,* тобто характер даних, що містить стовпчик: безперервні або дискретні. Якщо обрано безперервний вид, то дані в стовпці можуть приймати будь-яке значення в рамках свого типу. Якщо обрано дискретний вид, то дані в стовпці можуть приймати кінцеве число значень. Безперервними можуть бути тільки числові дані. Дискретний характер носять, як правило, строкові дані, але не завжди. Дискретними можуть бути призначені в залежності від контексту задачі, що розв'язується, дані цілого типу, рідше дійсного. Вид даних стовпця впливає на алгоритм розрахунку статистики за стовпчиком і роботу аналітичних алгоритмів;

– *Призначення,* тобто порядок використання поля набору даних, отриманого в результаті імпорту стовпця (поля), при подальшій обробці імпортованих даних (рис.11, табл.1).

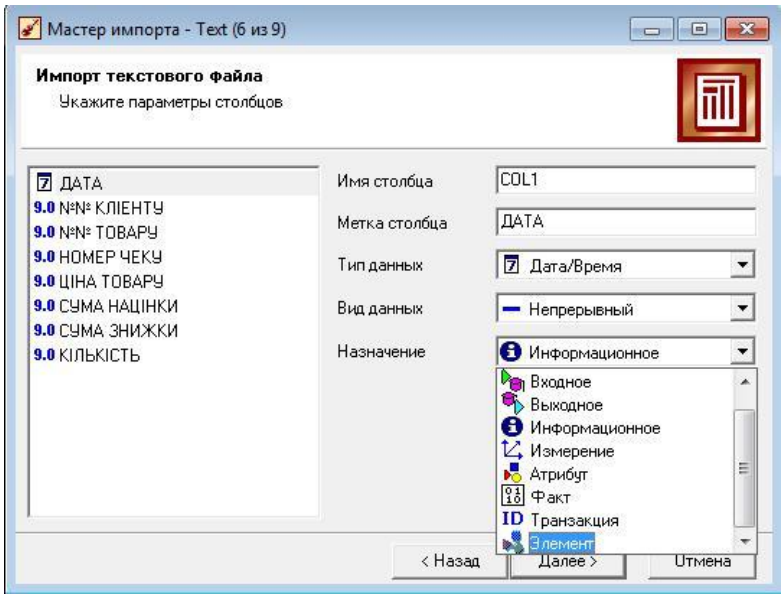


Рисунок 11 – Призначення стовпців

Завдання призначення стовпця набору даних при імпорті не є обов'язковою дією (за замовчуванням призначення встановлено як *Інформаційне*), однак це дозволяє знизити обсяг рутинних дій при подальшому конструюванні сценарію.

На кроці *Запуск процесу імпорту* стартує сам процес імпорту даних з раніше налаштованими параметрами. Хід процесу імпорту відображається за допомогою індикатора. Якщо процес імпорту зупинився, це сигналізує про можливі помилки при читанні даних. У цьому випадку з'являється вікно

з повідомленням про помилку. У разі виникнення помилок невідповідності типів процес імпорту буде продовжений, але після його закінчення буде відображений журнал реєстрації помилок з інформацією про місце і причини їх появи.

Таблиця 1 Призначення стовпців

№	*Призначення*	*Опис*
1	Вхідне	Поле набору даних, що побудовано на основі стовпця, який буде вхідним полем обробника (нейронна мережа дерева рішень та інші)
2	Вихідне	Поле набору даних, що побудовано на основі стовпця, який буде вихідним полем обробника (цільовим полем для навчання нейронної мережі)
3	Інформаційне	Поле вміщує додаткову інформацію, яку часто корисно відображати, але не слід використовувати при обробці
4	Вимір	Поле буде використовуватися в якості виміру в багатовимірній візуалізації
5	Атрибут	Поле вміщує опис властивостей або параметрів деякого об’єкту
6	Факт	Значення поля будуть використовуватися в якості фактів в багатовимірній візуалізації
7	Транзакція	Поле вміщує ідентифікатор подій, що відбуваються одночасно
8	Елемент	Поле, яке вміщує елемент транзакції (подія)

На останніх двох кроках *Майстра імпорту* буде запропоновано обрати візуалізатор набору даних (за замовчуванням пропонується *Таблиця*) і задати відомості про вузол.

2.4.6 Вузол *Налаштування набору даних* (етап *Трансформація даних*) дозволяє (рис. 12):

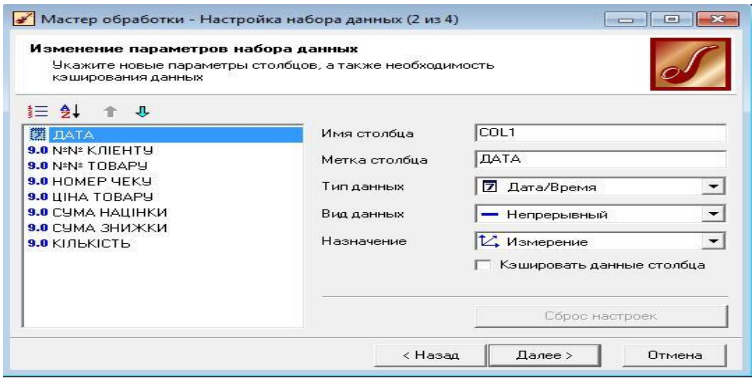


Рисунок 12 – Налаштування набору даних

– змінити ім'я, мітку, тип, вид і призначення полів поточного набору даних;

– змінити порядок проходження стовпців у наборі даних;

– приховати стовпці набору даних;

– задати опцію кешування вихідного набору.

Зміна імені або мітки поля має сенс у випадках, коли імена стовпців можуть змінитися в джерелі даних або при переналаштуванні вузлів верхніх рівнів. Після такої операції зміна імен полів на верхніх рівнях не потребує переналаштування всіх дочірніх вузлів в дереві сценаріїв.

Кнопка *Скидання налаштувань* дозволяє повернутися до первинних параметрів стовпців.

Кешування - це завантаження інформації, що часто використовується, в оперативну пам'ять для швидкого доступу до неї, минаючи багаторазові зчитування з жорсткого диска.

2.4.7 Експорт даних в текстовий файл із роздільниками (категорія *Файли*) виконується за допомогою *Майстра експорту* та містить наступні кроки:

– налаштування параметрів експорту;

– вибір символу роздільника стовпців;

– вибір експортованих полів;

– запуск процесу експорту;

– вибір способу візуалізації;

– введення відомостей про вузол.

На кроці *Налаштування параметрів експорту* задаються параметри експорту даних з текстового файлу аналогічно тому, як це здійснювалось при роботі з *Майстром імпорту*. Експортуватися будуть не всі поля, а тільки ті, які було помічено прапорцями на кроці *Вибір полів, які експортуються.*

2.4.8 Кожному вузлу сценарію, який містить структурований набір даних, завжди пропонується кілька візуалізаторів. *Майстер візуалізації* дозволяє обрати і налаштувати найбільш зручний спосіб представлення даних в інтерактивному режимі. В залежності від обраного способу будуть налаштовуватися різні параметри, а *Майстер*, відповідно, буде пропонувати різне число кроків (рис. 13).

Майстер візуалізації запускається для виділеного вузла сценарію. Крім того, робота цього Майстра завжди є продовженням роботи Майстра обробки і може бути розпочата при створенні (налаштуванні) будь-якого вузла.

Якщо на першому кроці Майстра візуалізації одночасно обрано кілька способів відображення даних, то всі відповідні кроки будуть послідовно включені в загальну процедуру налаштування.

Базовими візуалізаторами в *Deductor* є наступні: *Таблиця, Статистика, Відомості.*

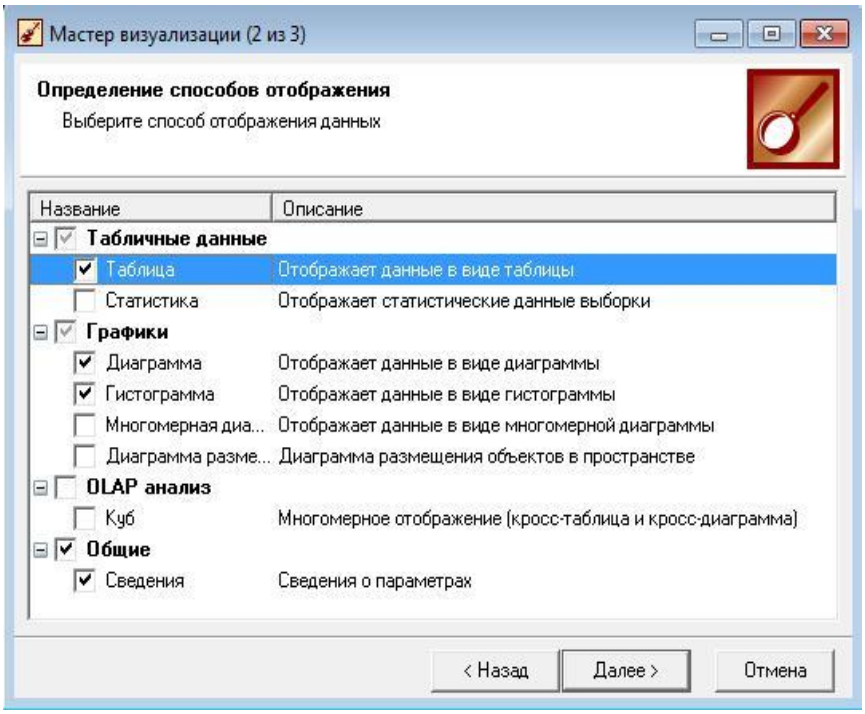
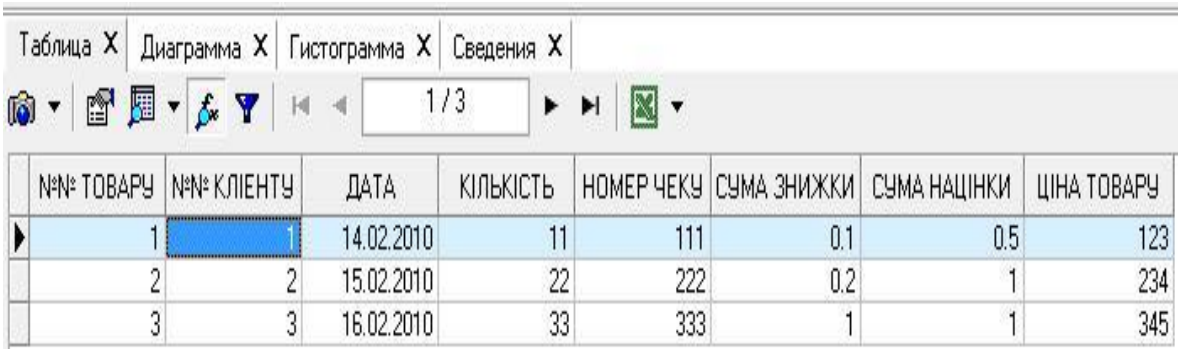


Рисунок 13 – *Майстер візуалізації*

У візуалізаторі *Таблиці* кожне поле набору даних розміщується в окремому стовпці. Стовпці мають мітки полів, або, якщо мітки не були задані, то імена полів. Ширину і послідовність стовпців в таблиці можна змінювати за допомогою миші. У таблиці є можливість налаштування об'єднання заголовків декількох стовпців (рис.14).



№№ ТОВАРУ	№№ КЛІЄНТУ	ДАТА	КІЛЬКІСТЬ	НОМЕР ЧЕКУ	СУМА ЗНИЖКИ	СУМА НАЦІНКИ	ЦІНА ТОВАРУ
1	1	14.02.2010	11	111	0.1	0.5	123
2	2	15.02.2010	22	222	0.2	1	234
3	3	16.02.2010	33	333	1	1	345

Рисунок 14 Візуалізатор *Таблиця*

У верхній частині вікна таблиці представлена панель інструментів, кнопки якої надають можливість скористатися низкою функцій (табл.2).

Команда *Налаштування полів* викликає відповідне діалогове вікно, в якому можна обрати опції індикації чи приховання полів таблиці, визначити спосіб вирівнювання вмісту, ширину поля, а також задати формат відображення числових даних і дат (рис. 15).

Таблиця 2 – Піктограми панелі інструментів візуалізатора *Таблиця*

№	*Призначення*	*Опис*
1	Управління конфігураціями	Збереження та відновлення конфігурацій відображення таблиці
2	Налаштування полів (F11)	Налаштування видимості полів, які відображені в таблиці, а також завдання їх форматів та способів вирівнювання
3	Спосіб відображення (Ctrl+F12	Перемикання між відображенням даних в вигляді таблиці або в вигляді форми
4	Статистика	Перегляд онлайн статистики за поточними даними таблиці
5	Фільтрація (Ctrl+D)	Фільтрація записів в таблиці за заданими умовами
6	Перший запис (Ctrl+PgUp	Перехід до першого запису набору даних
7	Попередній запис (PgUp)	Перехід до попереднього запису набору даних
8	Номер рядку	Індикатор поточного запису
9	Наступний запис (PgDn)	Перехід до наступного запису набору
10	Останій запис (Ctrl+ PgDn	Перехід до останнього запису набору даних
11	Експорт	Виклик вікна вибору файлу для експорту даних з таблиці в одному з прийнятних форматів

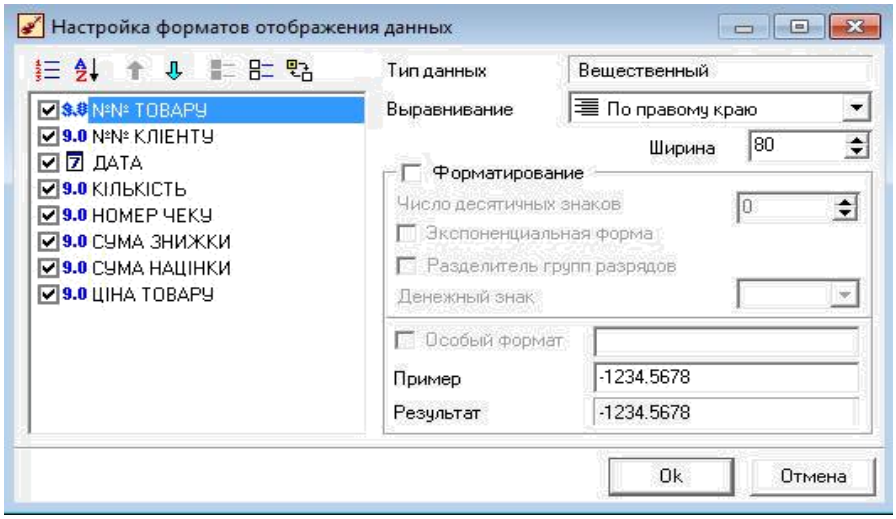


Рисунок 15 – Діалогове вікно *Налаштування форматів відображення даних*

Візуалізатор *Статистика* призначений для відображення основних статистичних характеристик набору даних вузлів. Статистичні характеристики відображаються в таблиці за кожним полем вибірки. У верхній частині вікна статистики відображається загальна кількість записів в наборі даних. Панель інструментів вікна дозволяє керувати відображенням статистичних характеристик (середнє, мінімум, максимум і т.п.) за допомогою групи кнопок (рис. 16).

	Метка столбца	Статистика: Кол-во значений = 2								
		Гистогра...	Мини...	Макс...	Сред...	Стан...	Сумма	Σ^2 Сумм...	|s| Кол-в...	|∅| Кол-в...
1	ДАТА		14.02.2010	15.02.2010	10 12:00:00	16:58:14				0
2	9.0 №№ КЛІЄНТУ		1	2	1.5	0.70711	3	5		0
3	9.0 №№ ТОВАРУ		1	2	1.5	0.70711	3	5		0

Рисунок 16 – Візуалізатор *Статистика*

Для полів дискретного типу, крім інших, завжди розраховуються наступні статистичні показники: кількість унікальних значень, кількість порожніх значень. Для полів безперервного типу в огляді статистики у відповідному стовпці присутня гістограма розподілу частот.

Візуалізатор *Відомості* дозволяє переглянути всі параметри, за якими було виконано відповідний процес перетворення даних при формуванні нової вибірки: імпорт, обробка одним з методів або експорт, а саме

- час і тривалість процесу, що виконується,
- умови зупинки,
- наявність первинного ключа,
- обмежувачі стовпців,
- роздільники цілої і дробової частин чисел,
- елементів дати і т.д.

Передбачено два види подання опису: дерево (за замовчуванням) або текстовий. Візуалізатор *Відомості* призначений для оперативного аналізу поточних налаштувань вузлів і для пошуку можливих помилок та є єдиним досяжним візуалізатором для вузлів експорту.

2.4.9 При виконанні дій з імпортованими даними в *Deductor* використовується *Майстер обробки*, який забезпечує наступні етапи:

- очищення даних,
- трансформація даних,
- Data Mining,
- інші.

2.4.9.1 Група вузлів *Очищення даних* наведена на рис.17.

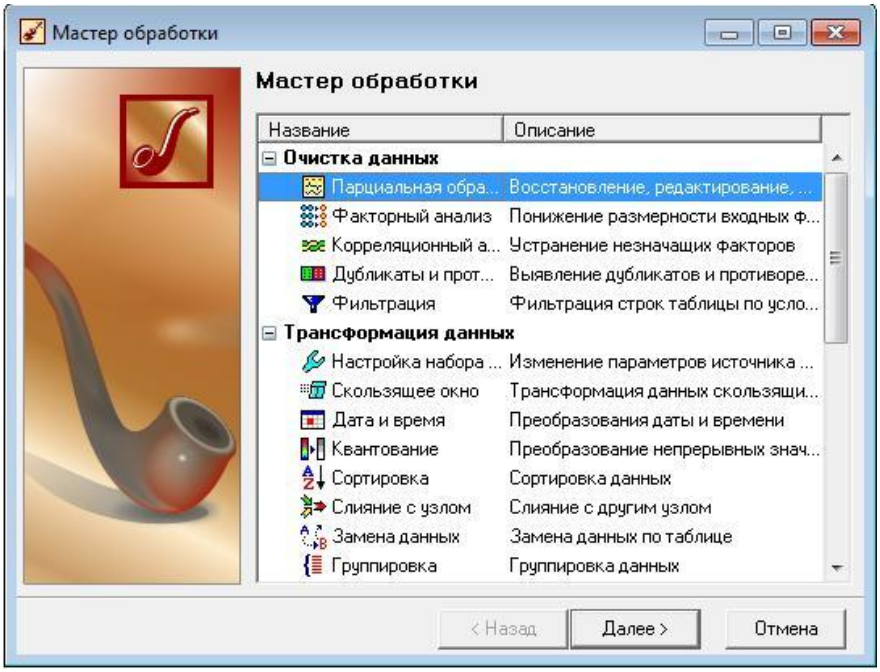


Рисунок 17 – Можливості етапу обробки даних *Очищення даних*

Парціальна обробка служить для відновлення відсутніх даних, редагування аномальних значень і спектральної обробки (рис. 18).

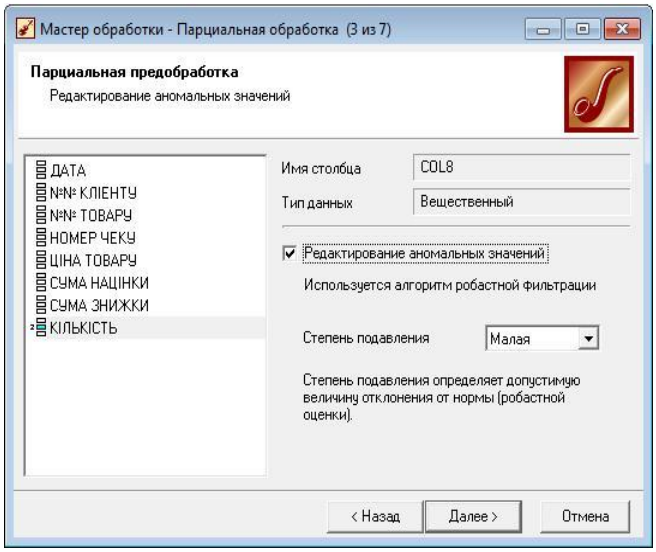


Рисунок 18 – *Парціальна обробка* (редагування аномальних значень)

Одним з методів *Спектральної обробки* є віднімання шуму, в результаті чого дані становляться більш згладженими, і їх можна використовувати для подальшої обробки (рис.19).

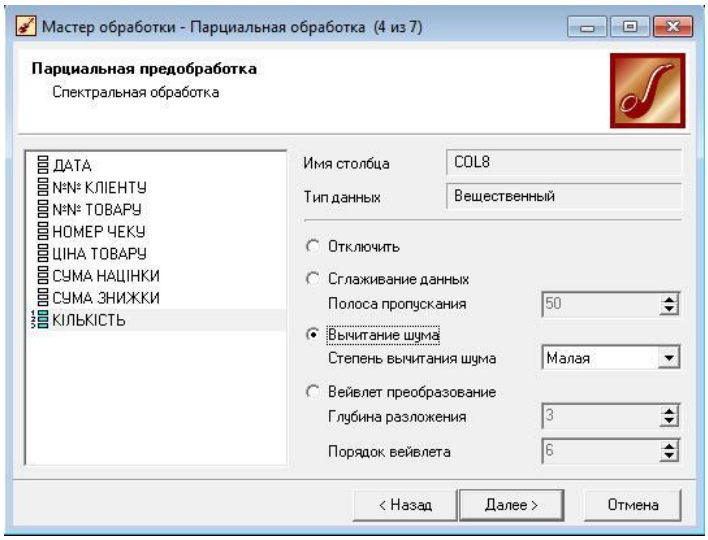


Рисунок 19 – *Спектральна обробка* (віднімання шуму)

Факторний аналіз дозволяє знизити розмірність простору вхідних факторів. Обробку можна виконувати як в автоматичному режимі (із зазначенням порогу значущості), так і вручну (ґрунтуючись на значеннях матриці значущості).

На першому кроці *Майстра обробки* обирається *Факторний аналіз* та задаються вхідні поля. Наступним кроком є процес зниження розмірності простору вхідних факторів, після завершення якого обираються фактори, які будуть залишені для подальшої роботи. Це робиться шляхом завдання необхідного порогу значущості (рис.20).

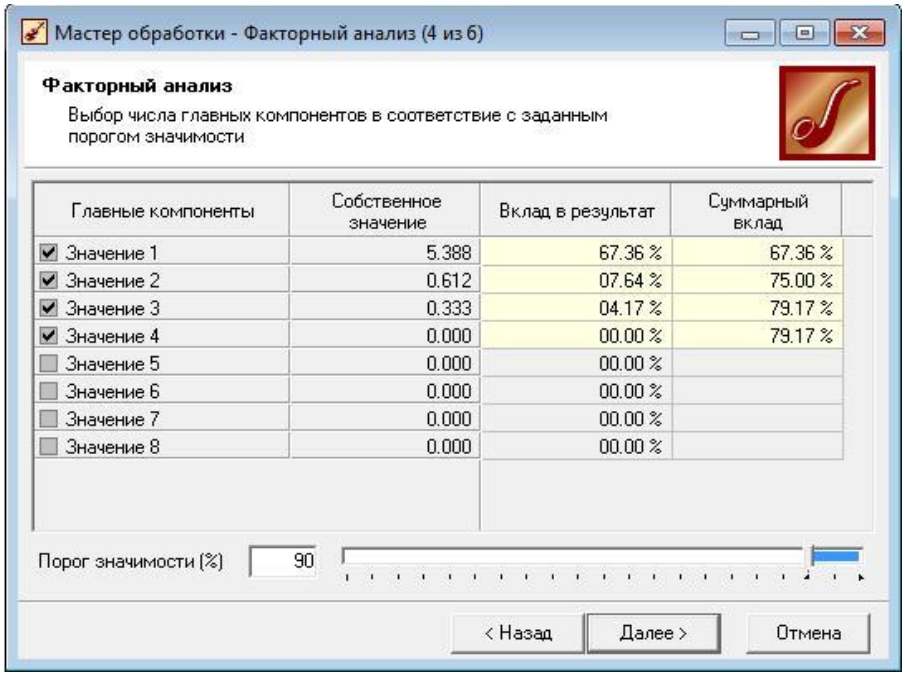


Рисунок 20 – *Факторний аналіз*

Кореляційний аналіз застосовується для усунення незначущих факторів за результатами оцінки ступеня лінійної залежності вихідних полів даних від вхідних факторів. Принцип кореляційного аналізу полягає в пошуку таких значень, які в найменшій мірі корелюють (взаємопов'язані) з вихідним результатом. Такі фактори можуть бути виключені з результуючого набору даних практично без втрати корисної інформації. Критерієм прийняття рішення про виключення є поріг значущості. Якщо кореляція (ступінь лінійної залежності) між вхідним і вихідним факторами менша порога значущості, відповідний фактор відкидається як незначущий. На першому кроці *Майстра обробки* обирається *Кореляційний аналіз* та задаються вхідні та вихідні поля. На наступному кроці обирають метод, на основі якого буде відбуватися розрахунок коефіцієнтів кореляції (рис.21).

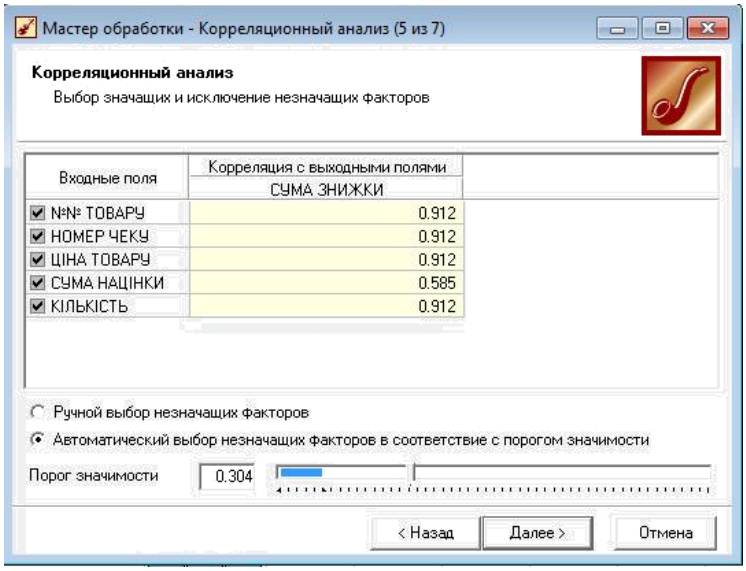


Рисунок 21 – Вікно вибору метода, за яким відбувається розрахунок коефіцієнтів кореляції (метод Пірсона)

Коефіцієнт кореляції Пірсона r є безрозмірним індексом в інтервалі від –1.0 до 1.0 включно, який характеризує ступінь лінійної залежності між двома множинами даних

Показник тісноти зв'язку між двома ознаками визначається за формулою лінійного коефіцієнта кореляції:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}$$

де

x – значення факторного ознаки;

y – значення результативної ознаки;

n – число пар даних.

За результатами кореляційного аналізу обирається, які фактори залишаються для подальшої роботи. Це робиться або вручну, ґрунтуючись на значеннях матриці коваріації, або за величиною порогу значущості (за замовчуванням поріг значущості дорівнює 0.05) (рис.22).

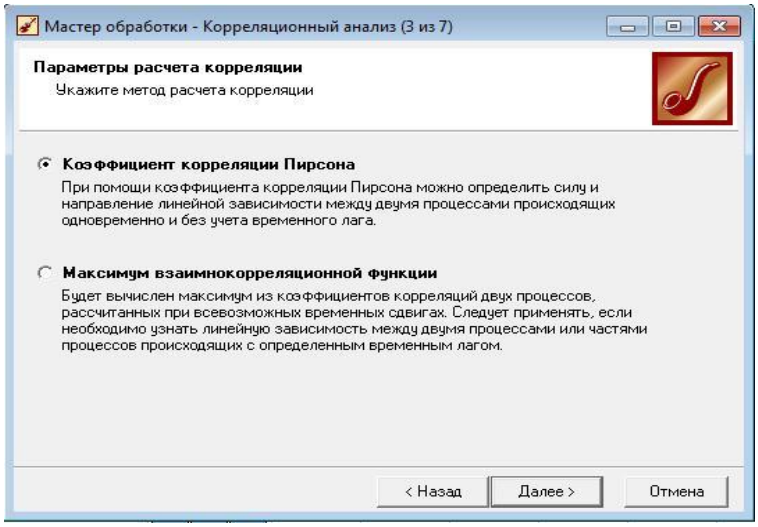


Рисунок 22 – Вибір значущих та виключення незначущих факторів

За отриманою матрицею кореляції можна чисельно оцінити ступінь впливу кожного з чинників на результуючий фактор і не враховувати при побудові моделей ті чинники, які впливають найменше (рис.23).

Матрица корреляции

Входные поля		Корреляция с выходными полями	
№	Поле	Результат1	Результат2
1	Фактор1	-0,538	0,549
2	Фактор2	-1,000	1,000
3	Фактор3	0,676	-0,685

Рисунок 23 – Матриця кореляції

Суперечливими є групи записів, в яких містяться рядки з однаковими вхідними факторами, але різними вихідними, тобто *протиріччя*. Якщо суперечливі дані використовувати для побудови моделі, то вона виявиться неадекватною. Тому суперечливі дані найчастіше краще взагалі виключити з вихідної вибірки.

Також в даних можуть зустрічатися записи з однаковими вхідними чинниками і однаковими вихідними, тобто *дублікати*. Присутність дублікатів в аналізованих даних можна розглядати як спосіб підвищення "значущості" інформації, що дублюється.

Сутність обробки полягає в тому, що визначаються вхідні (чинники) і вихідні (результати) поля. Алгоритм шукає у всьому наборі записи, для яких однаковим вхідним полях відповідають однакові (дублікати) або різні (протиріччя) вихідні поля. На підставі цієї інформації створюються два додаткових логічних поля *Дублікат* і *Протиріччя*, які приймають значення *Істинно* або *Хибно*. У додаткові числові поля *Група дублікатів* і *Група протиріч* записуються номер групи дублікатів та групи протиріч, в які потрапляє даний запис. Якщо запис не є дублікатом або протиріччям, то відповідне поле буде порожнім.

Обробник *Фільтрація* призначений для виключення з набору даних записів, які не задовольняють умовам фільтрації. Параметри фільтрації задаються у вигляді списку умов, який містить наступні стовпці:

– *Операція,* що дозволяє встановити функцію відносин *І* чи *АБО* між полями, для кожного з яких виконується фільтрація. Можлива фільтрація за декількома умовами та декількома полями одночасно. У результаті фільтрації за кожним полем (або умовою) буде отримано окрему множину значень. Функція в полі ОПЕРАЦІЯ встановлює відношення між цими множинами;

– *Ім'я поля,* що дозволяє обрати поле, за значеннями якого виконується фільтрація. Одне і те ж поле може використовуватись в декількох умовах;

– *Умова,* де вказуються параметри умов, за якими виконується фільтрація для даного поля;

– *Значення,* де вказуються значення, за якими виконується фільтрація записів за заданою умовою. Спосіб введення значення буде різним у залежності від типу даних і обраних умов.

2.4.9.2 Група вузлів *Трансформація даних* наведена на рис.17.

При прогнозування часового ряду кращого результату можна досягти, враховуючи значення факторів не тільки в даний момент часу, але, наприклад, за аналогічний період минулого року. Є змога використати таку можливість з застосуванням *Ковзного вікна* групи трансформації даних. В результаті обробки нові стовпці генеруються шляхом зсуву даних вниз і вгору від початкового стовпця (глибина занурення і горизонт прогнозу) (рис. 24).

В *Deductor Studio* існує інструмент *Квантування (або дискретизації),* який призначено для перетворення безперервних даних в дискретні. Перетворення може проходити як за інтервалом (дані розбиваються на задану кількість інтервалів однакової довжини), так і за квантилем (дані розбиваються на інтервали різної довжини так, щоб в кожному інтервалі знаходилася однакова кількість записів). Вихідними значеннями можуть бути номер інтервалу, нижня або верхня межа інтервалу, середина або мітка інтервалу (значення визначаються аналітиком).

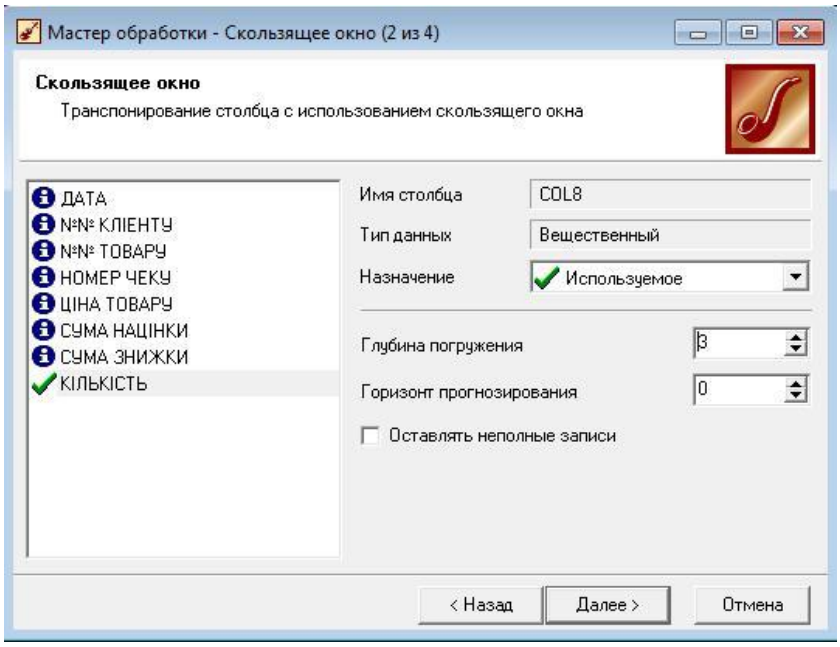


Рисунок 24 – Вікно вибору параметрів трансформації даних з застосуванням *Ковзного вікна*

У *Майстрі квантування* обирають (рис.25):

– призначення поля, наприклад, *Таке, що використовується*,
– спосіб розбиття, наприклад, *За інтервалом*,
– кількість інтервалів,
– значення, наприклад, *Мітка інтервалу*.

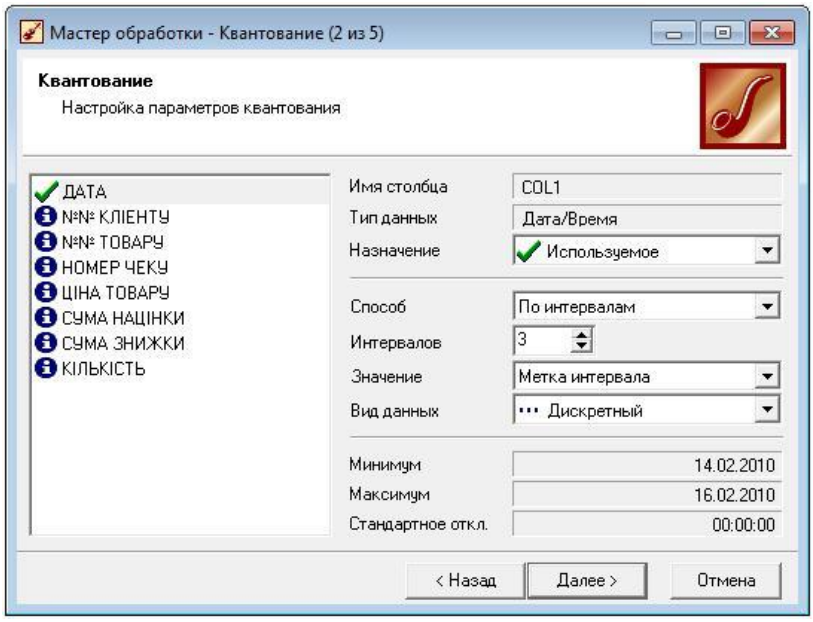


Рисунок 25 – Налаштування параметрів квантування

На наступному кроці визначаються мітки, та як спосіб відображення інформації – *Куб*. Інформація представляється в вигляді кросс-діаграми.

Обробник *Сортування* призначений для зміни порядку проходження записів у наборі даних у відповідності з обраним типом сортування. Результатом виконання сортування є новий набір даних, записи в якому розташовані у відповідності з заданими параметрами сортування.

Обробник *Заміна даних* призначений для заміни значень набору даних за таблицею підстановок, яка містить пари, що складаються з початкового та результуючого значення. Для кожного значення початкового набору даних шукається відповідність серед вихідних значень таблиці підстановки.

Обробник *Кросс-таблиця* призначений для перетворення початкової структури таблиці даних у зручну для форму. При роботі з обробником *Кросс-таблиця* початкову таблицю необхідно відредагувати так, щоб з'явилися додаткові поля. Наступним кроком є налаштування полів, що використовуються, для формування таблиці. Поля, що використовуються для побудови, повинні знаходитися в стовпчиках або рядках. У стовпчиках розміщують поля, на основі значень яких створюються нові; їх значеннями будуть обрані факти. У рядки поміщаються поля, які не потребують зміни (рис.26).

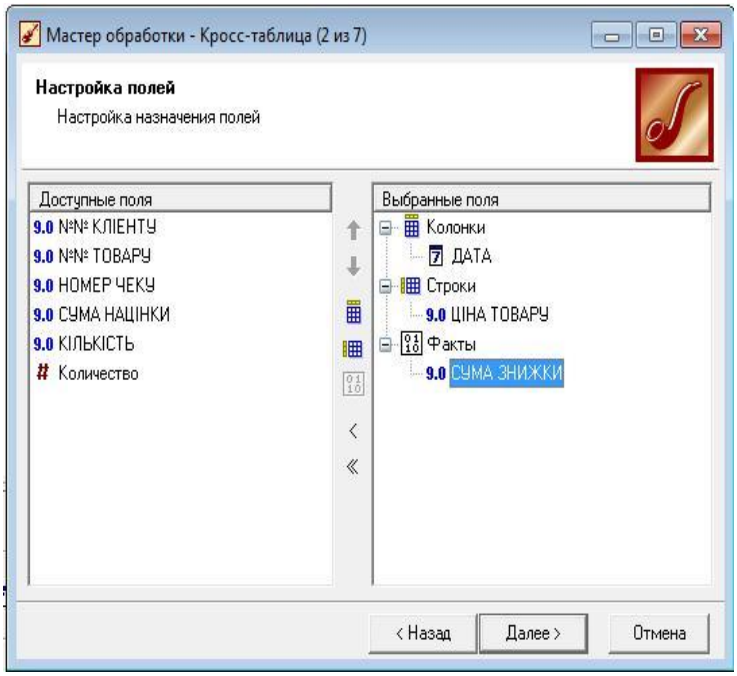


Рисунок 26 – Налаштування призначення полів в *Кросс-таблиці*

Наступним кроком необхідно налаштувати параметри агрегації обраних фактів (рис.27).

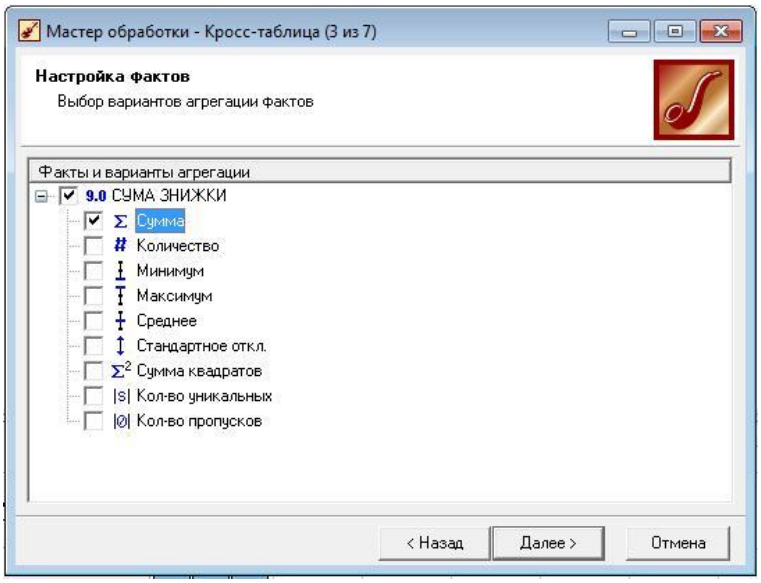


Рисунок 27 – Вибір варіантів агрегації фактів

Після натискання кнопки *Далі* відкривається наступне вікно *Майстра обробки*, в якому обирається налаштування параметрів вимірів у стовпцях і резервуються додаткові поля для можливого внесення змін до значення вихідного поля таблиці, а так само для вимірювань, у назві яких містяться пропуски. Таким чином, після обробки утворюється нова таблиця даних, на основі якої зручно робити необхідні обчислення індексів.

II ПРАКТИКУМ

Лабораторна робота 1
ВИКОРИСТАННЯ ПАРЦІАЛЬНОЇ ОБРОБКИ

Мета роботи – вивчення процесу очищення даних.
Завдання для підготовки до виконання лабораторної роботи.
Виконати парціальну обробку даних, наведених на рис. 28.

	A	B	C
1	Дата	Найменування	Кількість
2	08.01.2013	Товар1	156
3	17.01.2013	Товар2	23
4	22.01.2013	Товар3	67
5	02.02.2013	Товар1	178
6	04.02.2013	Товар3	77
7	06.02.2013	Товар2	34
8	08.02.2013	Товар3	89
9	10.02.2013	Товар1	197
10	12.02.2013	Товар2	12
11	14.02.2013	Товар3	93
12	16.02.2013	Товар1	205
13	28.02.2013	Товар3	64
14	03.03.2013	Товар2	21
15	04.03.2013	Товар1	186
16	05.03.2013	Товар2	41
17	06.03.2013	Товар3	87
18	07.03.2013	Товар2	43
19	12.03.2013	Товар3	92
20	15.03.2013	Товар2	37
21	25.03.2013	Товар3	94

Рисунок 28 – Вихідні дані для лабораторної роботи 1

1Загальні положення

Вся робота з аналізу даних у *Deductor Studio* базується на виконанні наступних дій: імпорт даних, обробка даних, візуалізація, експорт даних.

Відправною точкою для аналізу завжди є процедура імпорту даних. Отриманий набір даних може бути оброблений будь-яким способом з арсеналу *Deductor Studio*. Імпортований набір даних, а також дані, отримані на кожному етапі обробки, можуть бути експортовані. Результати кожної дії можна відобразити різними способами, вибір яких залежать від методу обробки даних.

Послідовність дій, які необхідно виконати для аналізу даних, називається *сценарієм* та може бути автоматично виконана для будь-яких даних.

Початкові дані через погану якість не завжди пригідні для аналізу в «сирому» вигляді, тому питання їх попередньої підготовки для аналізу є дуже важливим. Зазвичай "сирі" дані можуть містити пусті інтервали та шуми, а також аномалії та події, що рідко відбуваються, що заважає побачити загальну картину. Очевидно, що для отримання якісної моделі, вплив випадкових та перешкоджаючих факторів слід мінімізувати, використовуючи стійкі до впливу алгоритми аналізу і спеціалізовані механізми очищення.

В багатьох випадках етапу досліджень передує *парціальна обробка*. Парціальна передобробка має на меті відновлення втрачених даних, редагування аномальних значень і спектральну обробку даних (наприклад, згладжування).

В *Deductor Studio* використовуються алгоритми, де дані полів аналізованого набору обробляються незалежно один від одного, тобто по частинах, тому така обробка отримала назву парціальної. До процедур передобробки даних, реалізованих в *Deductor Studio*, входять

- згладжування,
- видалення шумів,
- редагування аномальних значень,
- заповнення пропусків у рядах даних.

Згладжування даних дозволяє видалити з початкового набору шуми, а також виявити тенденції, які не завжди чітко простежуються у вихідному наборі. Платформа *Deductor Studio* пропонує кілька видів спектральної обробки:

- згладжування даних шляхом визначення смуг пропускання,
- віднімання шуму,
- вейвлет-перетворення шляхом зазначення глибини розкладання і порядку вейвлета.

Вейвлети (від англ. *Wavelet* – сплеск, імпульс) – це математичні функції, що дозволяють аналізувати різні частотні компоненти даних. Вейвлет-коефіцієнти визначаються інтегральним перетворенням сигналу. Отримані вейвлет-спектрограми дають чітку прив'язку спектру різних сигналів до часу.

Вейвлет-перетворення (англ. *wavelet transform*) – це інтегральне перетворення, яке є згорткою вейвлет-функції сигналу або спосіб перетворення функції (або сигналу) в форму, яка або робить деякі величини початкового сигналу такими, що більш піддаються вивченню, або дозволяє стиснути вихідний набір даних. Вейвлет-перетворення сигналів є узагальненням спектрального аналізу [1].

2Порядок виконання лабораторної роботи

1Створити в табличному процесорі файл з полями ДАТА, НАЙМЕНУВАННЯ та КІЛЬКІСТЬ (рис. 28).

2Зберегти файл в форматі *MS Excel або Calc*.

3Зберегти файл в форматі «Текстовий файл (з розподільниками табуляції)».

4Запустити аналітичну платформу *Deductor*.

5 Обрати в аналітичній платформі *Deductor Майстер імпорту* та перейти до налаштування параметрів.

5.1 Обрати ім’я файлу, з якого планується імпортувати дані. В вікні перегляду обраного файлу можна побачити його вміст (другий крок *Майстра імпорту*) (рис. 29). В цьому вікні слід вказати, з якого рядка слід починати імпорт та що перший рядок є заголовком.

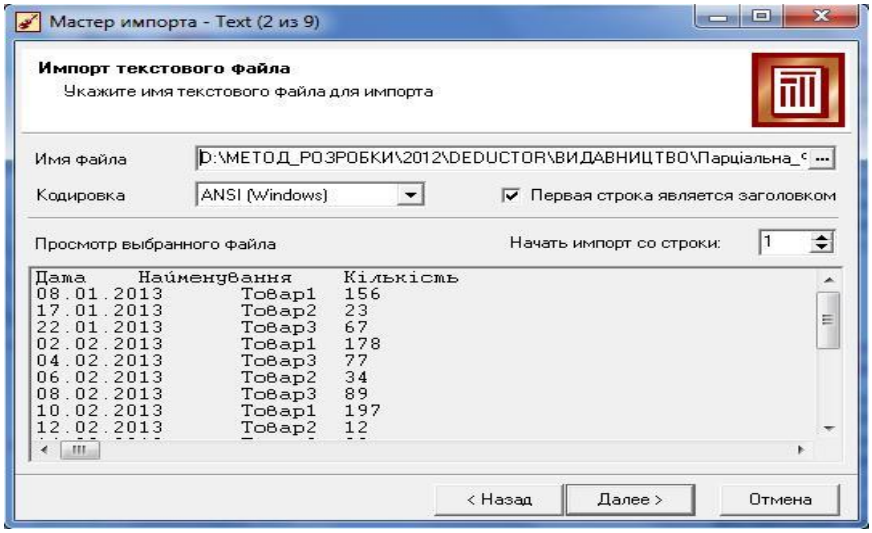


Рисунок 29 – Вибір текстового файлу для імпорту в *Deductor*

5.2 На наступному кроці *Майстра імпорту* визначити символ-розподільник стовпців, обмежувач рядків та розподільник цілої та дробової частини дійсного числа, формат компонентів дати (рис. 30).

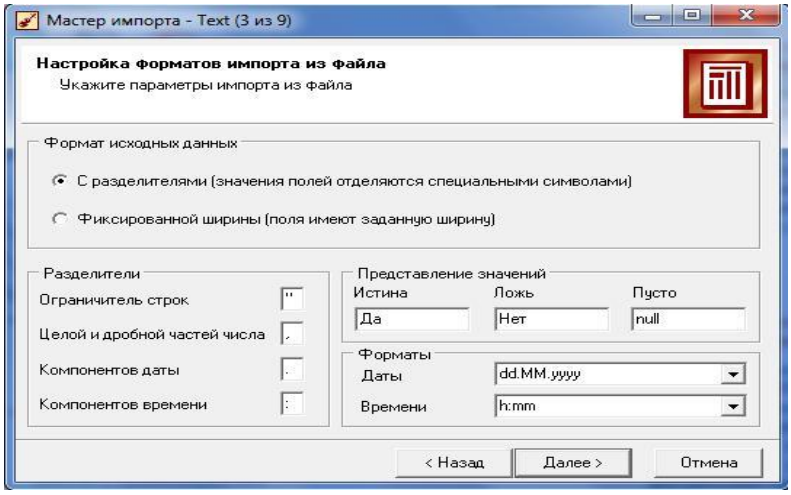


Рисунок 30 – Налаштування форматів імпорту з файлу

6 Налаштувати в вікні *Майстра імпорту (6 з 9)* (рис. 31) параметри кожного поля:

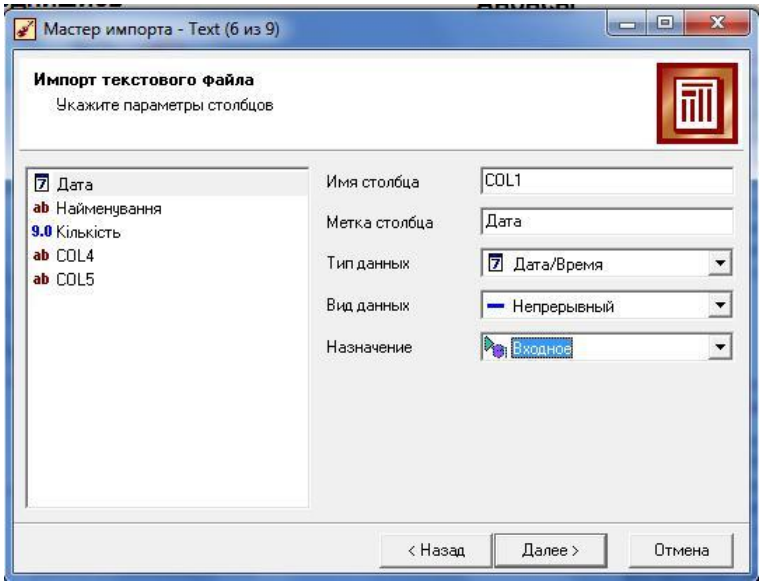


Рисунок 31 – Налаштування параметрів полів

- *ім'я стовпця,*
- *мітка стовпця,*
- *тип даних,*
- *вид даних* (дискретний або безперервний),
- *призначення (* характер їх використання в алгоритмах обробки).

Вказати в вікні ІМПОРТ ТЕКСТОВОГО ФАЙЛУ параметри стовпців як наведено в табл. 3.

Таблиця 3 – Завдання параметрів стовпців

№ №	Параметри стовпця	Ім'я стовпця		
		Дата	Найменування	Кількість
1	Тип даних	Дата/Час	Строковий	Дійсний
2	Вид даних	Безперервний		Безперервний
3	Призначення	Вхідне	Вхідне	Вхідне

7 Запустити процес імпорту, натиснувши кнопку Пуск в вікні *Майстра імпорту* (*7 з 9*) та кнопку Далі.

8 Обрати на наступному кроці спосіб відображення даних: *Таблиця* (рис. 32) та *Діаграма*, та налаштувати параметри діаграми (рис.33), в результаті чого отримаємо тривимірну діаграму (рис.34).

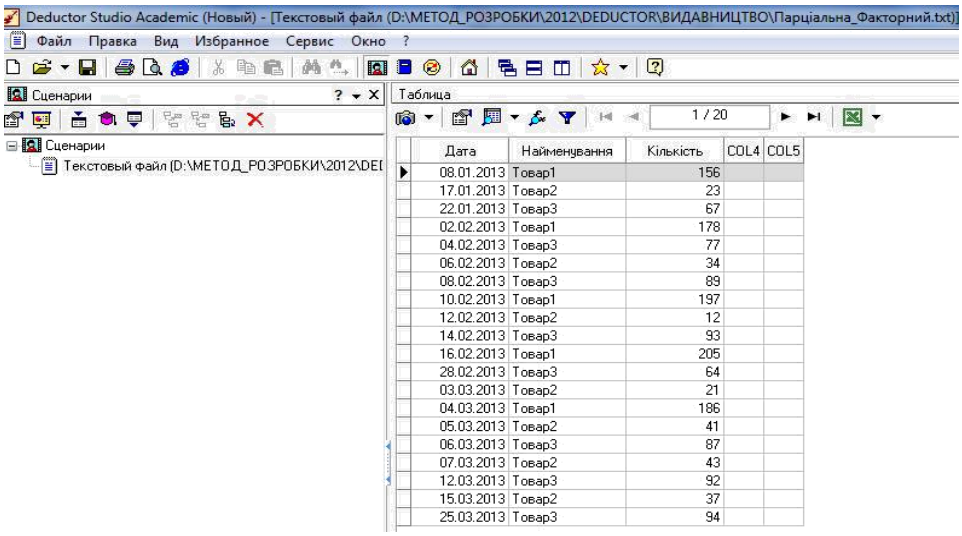


Рисунок 32 – Результати імпорту в вигляді *Таблиця*

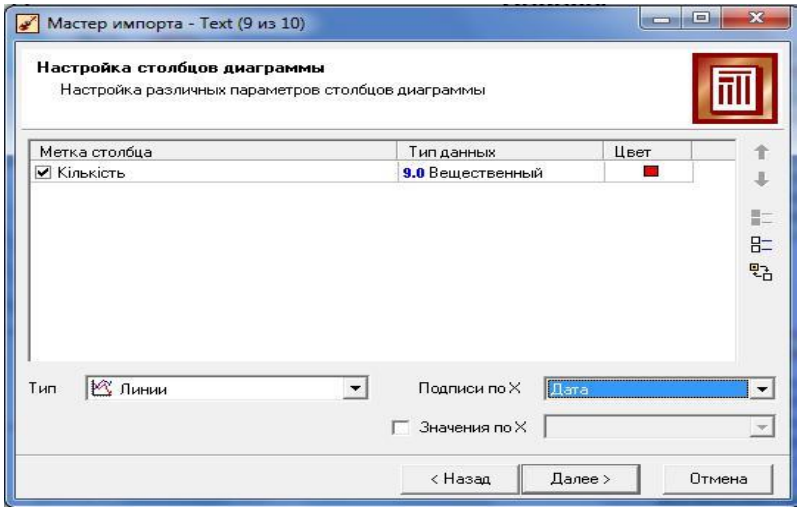


Рисунок 33 – Налаштування стовпців діаграми

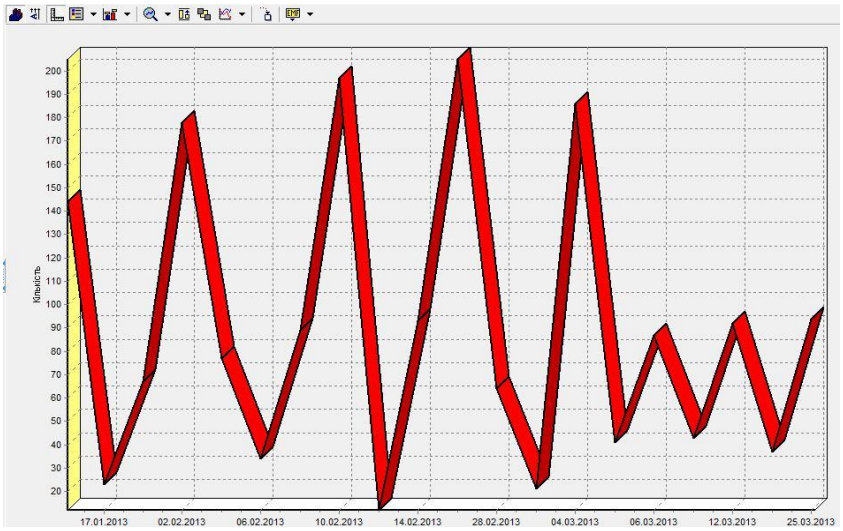


Рисунок 34 – Побудована діаграма

Як видно, викиди погіршують статистичну картину розподілення даних.

9 Запустити процес парціальної обробки *Майстру обробки* (рис. 35).

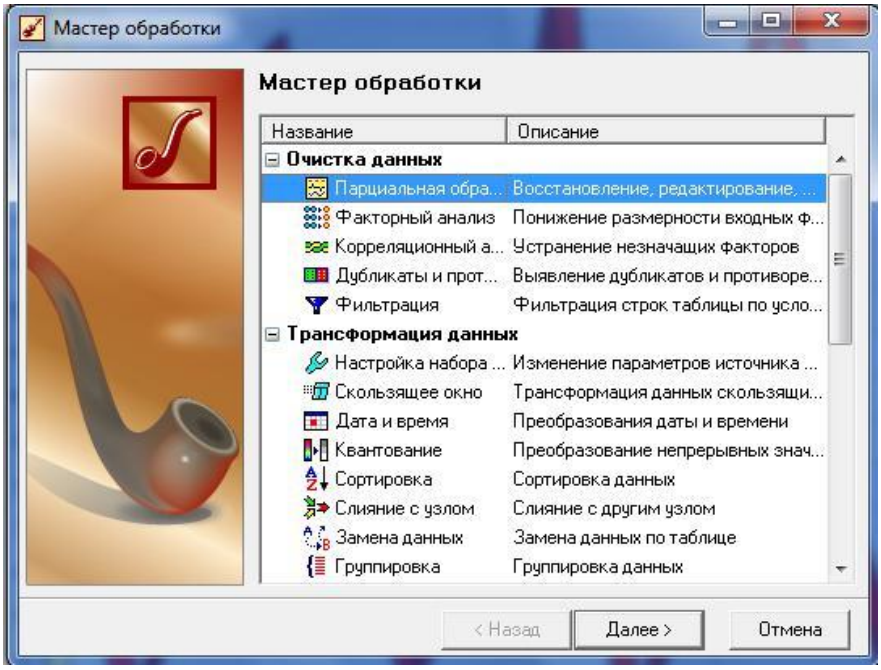


Рисунок 35 – *Парціальна обробка*

9.1 На другому кроці парціальної обробки (відновлення пропущених даних) обрати поле КІЛЬКІСТЬ та встановити перемикач в положення АПРОКСИМАЦІЯ, що доцільно для упорядкованих за часом даних;

9.2 На третьому кроці парціальної обробки (редагування аномальних значень) обрати поле КІЛЬКІСТЬ та вказати тип обробки РЕДАГУВАННЯ АНОМАЛЬНИХ ЗНАЧЕНЬ зі ступенем придушення – великий (рис.36).

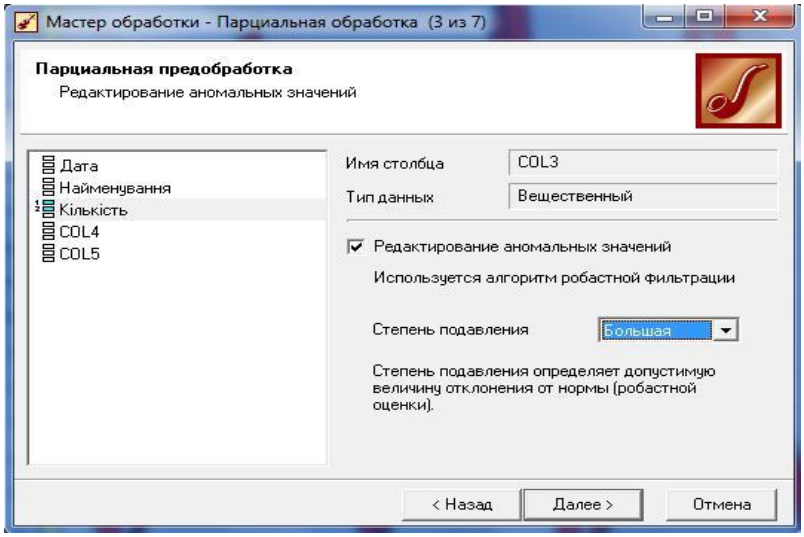


Рисунок 36 Редагування аномальних значень

9.3 На четвертому кроці парціальної обробки (спектральна обробка) обрати поле КІЛЬКІСТЬ та метод спектральної обробки – ВІДНІМАННЯ ШУМУ З ВЕЛИКИМ СТУПЕНЕМ ВІДНІМАННЯ ШУМУ.

9.4 Запустити процес згладжування даних натисканням кнопки Пуск та Далі.

9.5 На наступному кроці парціальної обробки обрати в якості візуалізації діаграму (рис.37) та налаштувати її стовпці (рис.33).

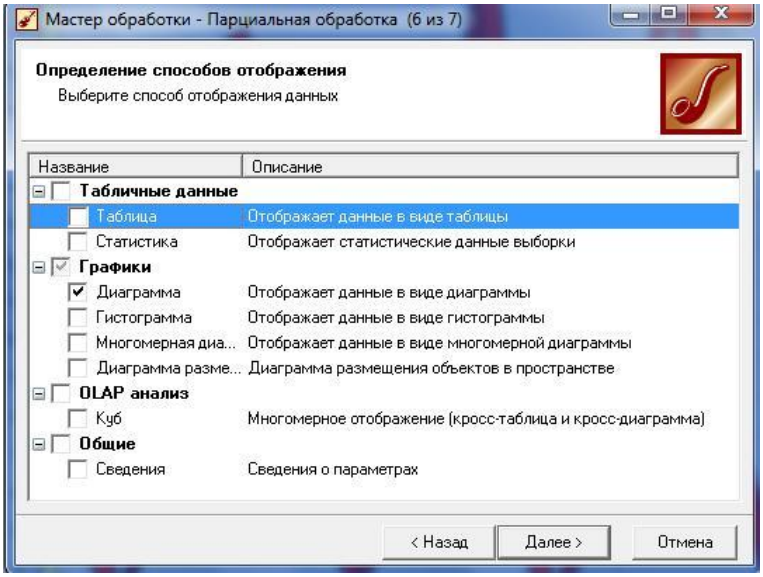


Рисунок 37 – Способи відображення

На рис. 38 наведено діаграму з типом міток – мітка, значення.

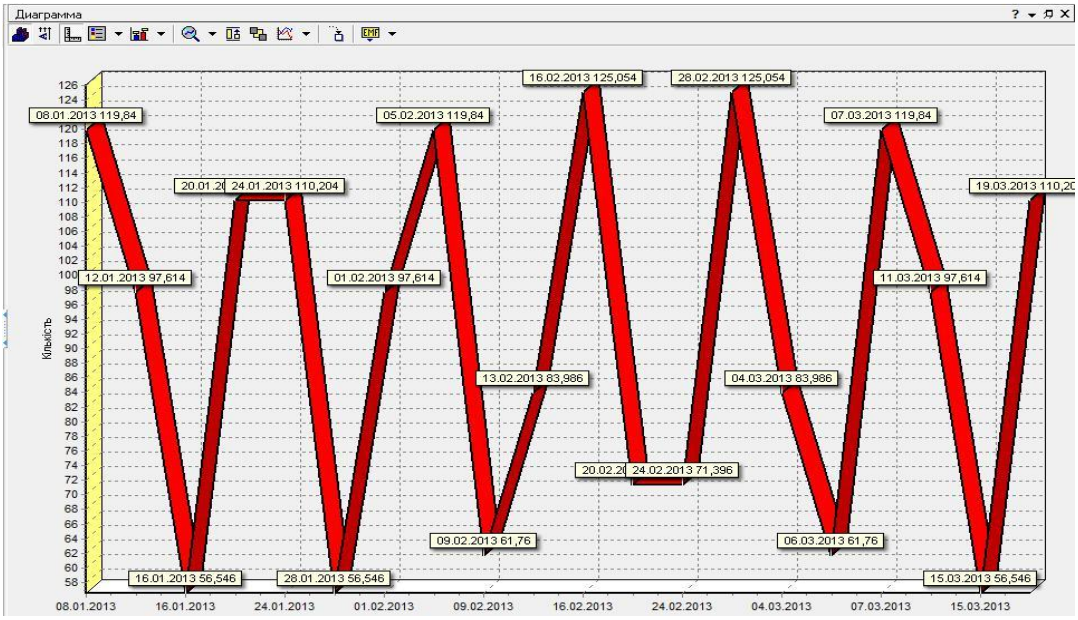


Рисунок 38 – Діаграма з мітками

Отримані дані можна використовувати для подальшої обробки, але на більш великій вибірці даних було б отримано кращий результат.

Цей результат висвітлює процес імпорту даних та процес очищення даних як алгоритм навчального процесу. В залежності від вхідних даних користувач повинен обирати найбільш прийнятні варіанти передобробки даних.

10 Доповнити запропоновану вибірку даних 30 записами та закріпити навички використання парціальної передобробки.

Контрольні питання

1 Описати процес імпорту даних в аналітичну платформу *Deductor*.

2 Для яких цілей використовується парціальна обробка даних?

3 Назвати види спектральної обробки.

4 Дати визначення вейвлет-перетворення.

5 Які параметри в вікні *Майстра імпорту* виявляють характер використання стовпців в алгоритмах обробки?

6 Назвати способи відображення даних в *Майстрі імпорту.*

Лабораторна робота 2
ФАКТОРНИЙ ТА КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Мета роботи – зниження розмірності та усунення незначущих факторів в таблиці початкових даних.

Завдання для підготовки до виконання лабораторної роботи.

Виконати комплексну обробку, зниження розмірності вхідних факторів (факторний аналіз) та усунення незначущих факторів (кореляційний аналіз) на заданій вибірці даних (рис.39).

	A	B	C	D	E	F
1	Аргумент	Фактор1	Фактор2	Фактор3	Результат1	Результат2
2	0	0	1	0,14	0	1
3	0,05	0,05	1	0,09	0,05	1
4	0,1	0,1	0,99	-0,01	0,1	0,99
5	0,15	0,15	0,98	-0,06	0,15	0,98
6	0,2	0,2	0,97	-0,011	0,2	0,97
7	0,25	0,25	0,96	-0,016	0,25	0,96
8	0,3	0,3	0,94	-0,021	0,3	0,94
9	0,35	0,35	0,92	-0,026	0,35	0,92
10	0,4	0,4	0,9	-0,3	0,4	0,9
11	0,45	0,45	0,88	-0,35	0,45	0,88
12	0,5	0,5	0,86	-0,4	0,5	0,86
13	0,55	0,55	0,84	-0,45	0,55	0,84
14	0,6	0,6	0,82	-0,5	0,6	0,82
15	0,65	0,65	0,8	-0,55	0,65	0,8
16	0,7	0,7	0,78	-0,6	0,7	0,78
17	0,75	0,75	0,76	-0,65	0,75	0,76
18	0,8	0,8	0,74	-0,7	0,8	0,74
19	0,85	0,85	0,72	-0,75	0,85	0,72
20	0,9	0,9	0,7	-0,8	0,9	0,7
21	0,95	0,95	0,68	-0,85	0,95	0,68
22	1	1	0,66	-0,9	1	0,66

Рисунок 39 – Початкові дані

1 Загальні положення

Комплексна передобробка служить для зниження розмірності простору вхідних та усунення незначущих факторів за допомогою факторного та кореляційниого аналізу.

Факторний аналіз – метод багатовимірного статистичного аналізу, який на основі експериментального спостереження впливу ознак об'єкта один на одного дозволяє виділити групу змінних, що визначають кореляційний зв'язок між ними.

Факторний аналіз дозволяє знизити розмірність факторного простору за рахунок виділення з множини вимірюваних характеристик об'єкта нових факторів, які найкращим чином відображають властивості об'єкта.

Поля використовуються в факторному аналізі за дотримання умов:

– мають числовий тип даних;

– не містять пропусків;

– мають ненульове стандартне відхилення, тобто в стовпці поля розташовані різні значення.

В іншому випадку, поля будуть автоматично позначені як непридатні. Зниження розмірності факторного простору має сенс за наявністю хоча б двох вхідних полів. За допомогою факторного аналізу кількість незалежних змінних скоротиться, якщо *Deductor* вилучить фактори, значення яких за критерієм значущості виявляться нижче граничного.

Першим етапом факторного аналізу є вибір нових ознак, які є лінійними комбінаціями попередніх і "вбирають" в себе більшу частину загальної мінливості вхідних факторів. У обробнику *Факторний аналіз* це здійснюється за допомогою *методу головних компонент*, який зводиться до вибору нової ортогональної системи координат в просторі спостережень. Як першу головну компоненту обирають напрямок, уздовж якого масив даних має найбільший розкид. Вибір кожної наступної головної компоненти відбувається за умови максимального розкиду даних уздовж неї та її ортогональнності обраним раніш головним компонентам.

Вибір головних компонент при факторному аналізі може здійснюватися напівавтоматично: користувач задає рівень значущості, який в сумі повинні давати головні компоненти. У результуючому наборі залишаються розташовані в порядку убування головні компоненти, сумарний внесок яких не менше заданого користувачем рівня значущості.

Кореляційний аналіз застосовується для оцінки залежності вихідних полів даних від вхідних факторів і усунення незначущих факторів. Принцип кореляційного аналізу полягає в пошуку таких значень, які найменшою мірою корелюють з вихідним результатом. Такі фактори можуть бути виключені з результуючого набору даних практично без втрати корисної інформації. Критерієм прийняття рішення про виключення є поріг значущості. Якщо кореляція між вхідним і вихідним факторами менша порога значущості, то відповідний фактор відкидається як незначущий.

Кореляція може бути позитивною і негативною (можлива також ситуація відсутності статистичного взаємозв'язку, наприклад, для незалежних випадкових величин). Негативна кореляція – це така, за якої збільшення однієї змінної пов'язано зі зменшенням іншої змінної, при цьому коефіцієнт кореляції є від'ємним. Позитивна кореляція – це така, за якої збільшення однієї змінної пов'язано зі збільшенням іншої змінної, при цьому коефіцієнт кореляції є позитивним.

– Поле може бути використане в кореляційному аналізі за виконання декількох умов:

– числовий тип даних та ненульове стандартне відхилення стовпця (в стовпчик містить різні значення), в іншому випадку, поле буде автоматично позначено як непридатне для аналізу;

– великий обсяг вибірки для вивчення, для конкретного виду коефіцієнта кореляції становить від 25 до 100 пар спостережень;

– друге обмеження випливає з гіпотези кореляційного аналізу, в яку закладена лінійна залежність змінних.

Факт кореляційної залежності між змінними не дозволяє виявити, яка змінна є залежною, а яка ні, але є доказом того, що змінні взагалі причино пов'язані між собою, наприклад, через дії третього чинника.

Відкидання незначущих факторів проводиться на підставі розрахованої кореляції. Можливі два варіанти прийняття рішення, що визначаються вибором відповідного пункту в нижній частині вікна:

– в ручному режимі стовпці зі значущими чинниками, які мають бути включеними до вихідного набору, необхідно помітити прапорцями та позначки навпроти тих стовпців, які треба виключити з набору, мають бути відсутніми;

– в автоматичному режимі можна задати необхідний рівень значущості за допомогою повзунка смуги *Поріг значущості* (рекомендовані значення порогу значущості виділено синім кольором). Стовпці, де максимальне розраховане значення кореляції менше порога, будуть виключені з вихідного набору.

У вихідний набір потраплять інформаційні поля, відзначені на цьому кроці стовпці й усі вихідні стовпці. Для усунення незначущих факторів вибірка повинна мати хоча б два вхідні й одне вихідне поле.

При виділенні в списку безперервного (числового) поля в секції *Статистика* для нього буде відображено набір основних статистичних характеристик: мінімальне, максимально і середнє значення, а також стандартне відхилення. При виділенні в списку дискретного поля в секції *Унікальних значень* для нього буде відображено кількість унікальних значень та наведено їх список [1].

Після проведення кореляційного аналізу користувачу надається можливість застосувати обробник *Матриця кореляції* [21].

2Порядок виконання лабораторної роботи

1 Підготовка до комплексної обробки

1.1 Створити файл в табличному процесорі (*MS Excel, Calc)* (рис. 39).

1.2 Зберегти файл в форматі «Текстовий файл (з розподільниками табуляції)».

1.3 Запустити аналітичну платформу *Deductor* та ініціювати *Майстер імпорту* для завантаження сформованого текстового файлу в *Deductor*.

1.4 Налаштувати в вікні *Майстра імпорту (6 з 9)* (рис.31) параметри кожного поля: ім'я та мітка стовпця, тип та вид даних, призначення.

Вказати в вікні ІМПОРТ ТЕКСТОВОГО ФАЙЛУ параметри стовпців як наведено в табл. 4.

Таблиця 4 – Завдання параметрів стовпців

№	*Параметри стовпця*	*Ім'я стовпця*					
		Аргумент	Фактор1	Фактор2	Фактор3	Результат 1	Результат2
1	Тип даних	Строковий	Дійсний	Дійсний	Дійсний	Дійсний	Дійсний
2	Вид даних		Безперервний	Безперервний	Безперервний	Безперервний	Безперервний
3	Призначення	Інформа-ційне	Вхідне	Вхідне	Вхідне	Вихідне	Вихідне

1.5 Запустити процес імпорт, натиснувши кнопку Пуск в вікні *Майстра імпорту* (*7 з 9*) та кнопку Далі.

1.6 Обрати на наступному кроці спосіб відображення даних: *Таблиця* та *Діаграма*, та налаштувати параметри діаграми (рис.33). Встановити прапорці для відображення на діаграмі полів ФАКТОР1, ФАКТОР2, ФАКТОР3 та обрати тип діаграми – ЛІНІЇ (рис.40).

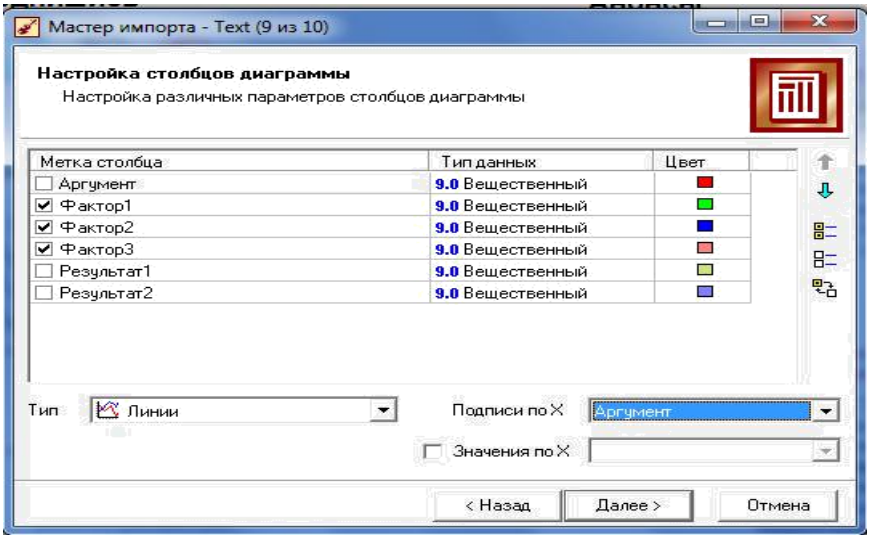


Рисунок 40 – Налаштування стовпців діаграми

В результаті побудована діаграма набути вигляду як на рис.41.

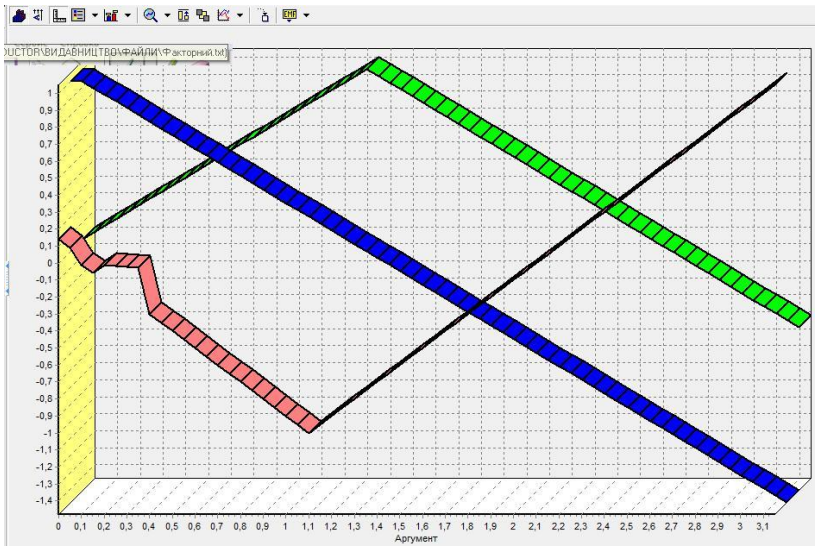


Рисунок 41 – Побудована діаграма

2 Факторний аналіз – зниження розмірності вхідних даних

2.1 Ініціювати *Майстер обробки* та в групі *Очищення даних* обрати *Факторний аналіз.*

2.2 Покроково налаштувати *Майстер обробки Факторний аналіз* так, щоб на 4 кроці отримати результат як на рис.42.

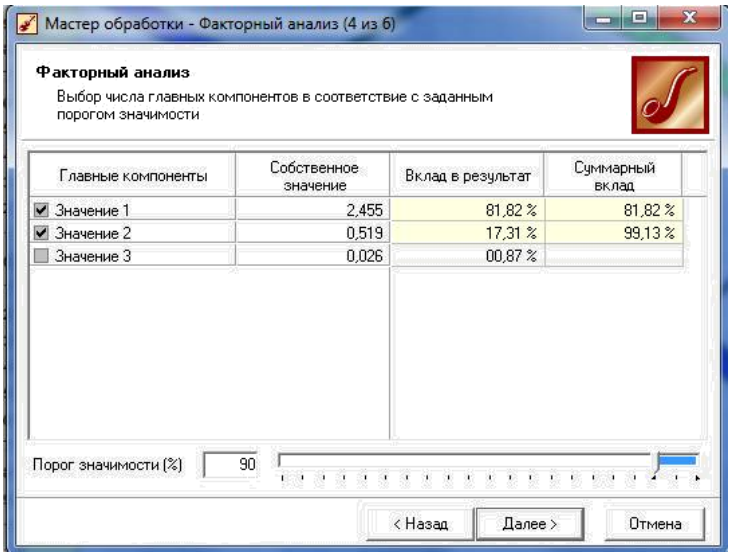


Рисунок 42 – Вікно завдання параметрів *Факторного аналізу*

2.3 Проаналізувати вплив значень вхідних параметрів. Після завершення процесу можна залишити для подальшої роботи два фактори: ФАКТОР1 та ФАКТОР2. Значенням ФАКТОРУ3 можна проігнорувати (рис.42).

2.4 Визначити спосіб відображення даних у вигляді *Діаграми* (крок 5 *Майстра обробки*) (рис.37).

2.5 Задати мітки відображення ФАКТОРА1 та ФАКТОРА2 від АРГУМЕНТУ (крок 6 *Майстра обробки*) (рис.43).

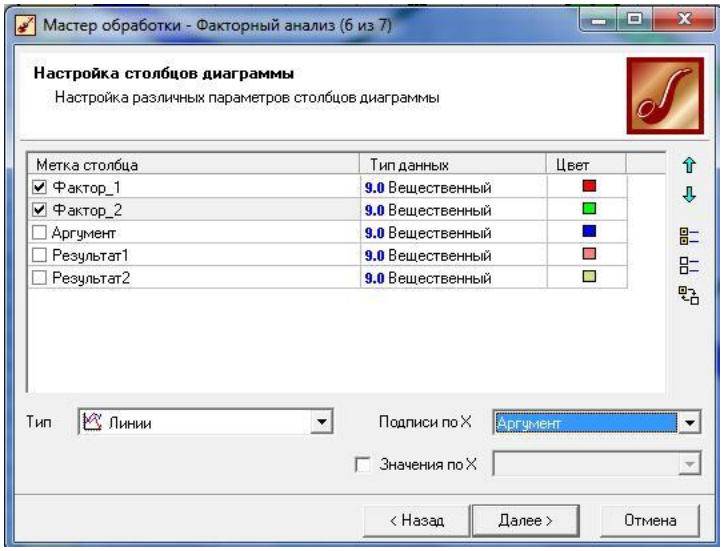


Рисунок 43 – Налаштування стовпців діаграми

2.6 Завершити процес обробки (рис.44- 46).

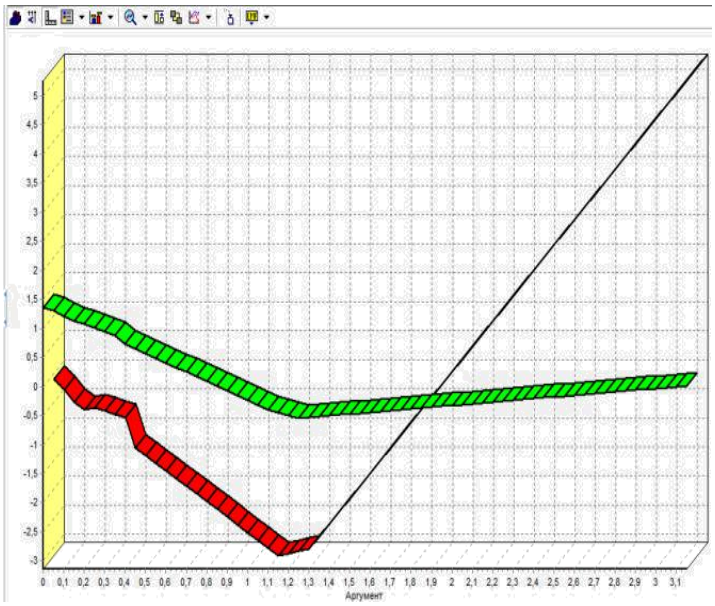
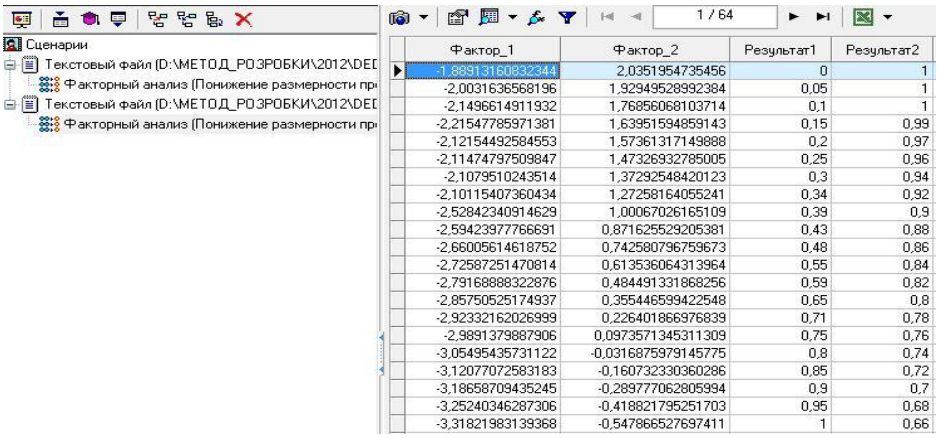
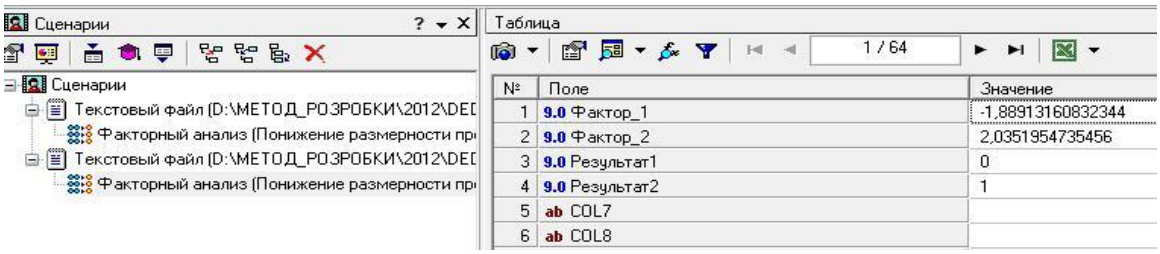


Рисунок 44 – Результат зниження розмірності моделі (діаграма)



Фактор_1	Фактор_2	Результат1	Результат2
-1,88913160832344	2,0351954735456	0	1
-2,0031636568196	1,92949528992384	0,05	1
-2,1496614911932	1,76856068103714	0,1	1
-2,21547785971381	1,63951594859143	0,15	0,99
-2,12154492584553	1,57361317149888	0,2	0,97
-2,11474797509847	1,47326932785005	0,25	0,96
-2,1079510243514	1,37292548420123	0,3	0,94
-2,10115407360434	1,27258164055241	0,34	0,92
-2,52842340914629	1,00067026165109	0,39	0,9
-2,59423977766691	0,871625529205381	0,43	0,88
-2,66005614618752	0,742580796759673	0,48	0,86
-2,72587251470814	0,613536064313964	0,55	0,84
-2,79168888322876	0,484491331868256	0,59	0,82
-2,85750525174937	0,355446599422548	0,65	0,8
-2,92332162026999	0,226401866976839	0,71	0,78
-2,9891379887906	0,097357134531309	0,75	0,76
-3,05495435731122	-0,0316875979145775	0,8	0,74
-3,12077072583183	-0,160732330360286	0,85	0,72
-3,18658709435245	-0,289777062805994	0,9	0,7
-3,25240346287306	-0,418821795251703	0,95	0,68
-3,31821983139368	-0,547866527697411	1	0,66

Рисунок 45 – Результат зниження розмірності моделі (таблиця)



№	Поле	Значение
1	9.0 Фактор_1	-1,88913160832344
2	9.0 Фактор_2	2,0351954735456
3	9.0 Результат1	0
4	9.0 Результат2	1
5	ab COL7	
6	ab COL8	

Рисунок 46 – Фрагмент відображення вхідних параметрів (форма)

3 Кореляційний аналіз – усунення незначущих факторів

3.1 Виділити імпортований файл та зробити його копію. Знищити з отриманої копії гілку *Факторний аналіз.*

3.2 Ініціювати *Майстер обробки* та в групі *Очищення даних* обрати *Кореляційний аналіз.*

3.3 Перевірити, щоб ФАКТОР1, ФАКТОР2, ФАКТОР3 мали призначення *Вхідні*, РЕЗУЛЬТАТ1 та РЕЗУЛЬТАТ2 – *Вихідні*, а АРГУМЕНТ – *Такий, що не використовується* (рис.47).

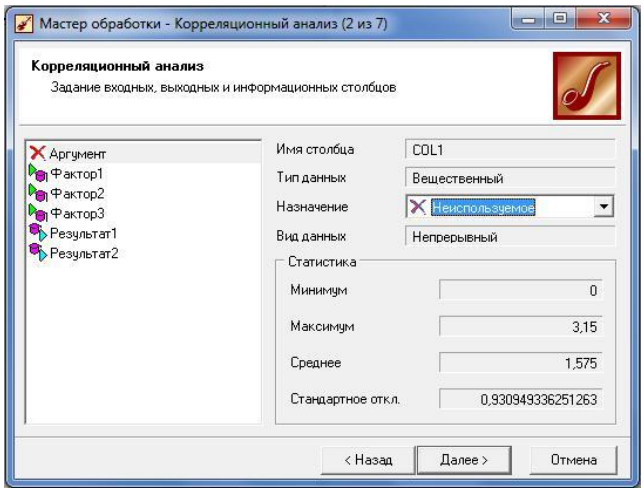


Рисунок 47 – Вікно вибору призначення вхідних та вихідних змінних

3.4 Обрати метод розрахунку кореляції – *Коефіцієнт кореляції Пірсона* (рис.48).

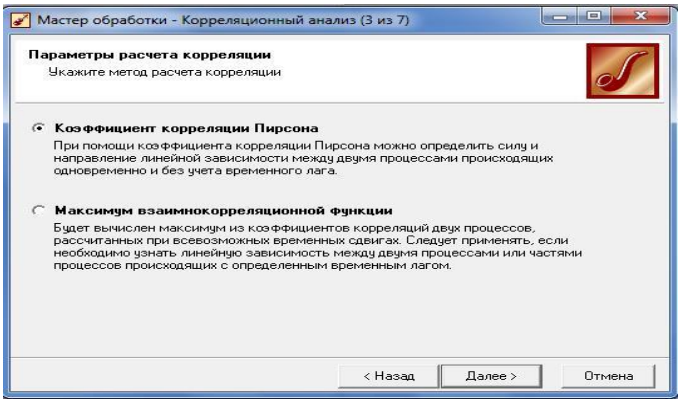


Рисунок 48 – Вікно вибору метода розрахунку кореляції

3.5 Обрати значущі та вилучити незначущі фактори в автоматичному режимі шляхом встановлення порогу значущості таким, що дорівнює 0.05 (рис.49).

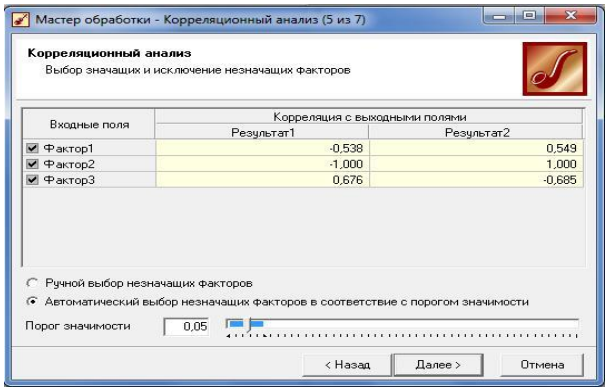


Рисунок 49 – Вікно вибору значущих та вилучення незначущих факторів

На цьому етапі розраховуються коефіцієнти кореляції, обирається поріг значущості, та *Deductor* виключає всі стовпці з коефіцієнтом нижче граничного рівня.

3.6 Задати спосіб відображення даних у вигляді *Матриці кореляції* (рис.50).

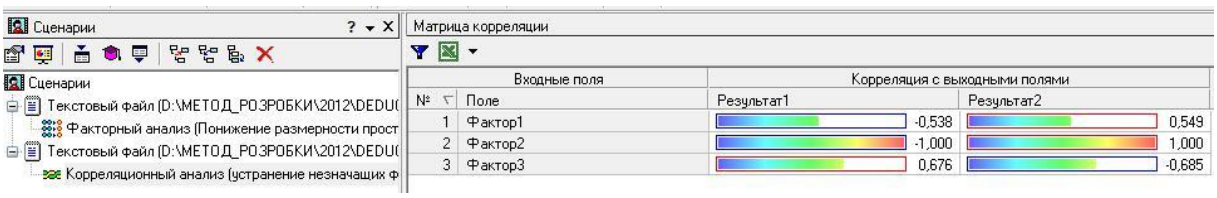


Рисунок 50– Матриця кореляції вхідних полів з вихідними параметрами

За даними *Матриці кореляції* (рис.50) можна дійти висновку, що найбільше на РЕЗУЛЬТАТ1 впливає ФАКТОР3 (коефіцієнт корреляції=0,676), на РЕЗУЛЬТАТ2 – ФАКТОР2 (коефіцієнт кореляції =1). Обидва коефіцієнти додатні, тобто зі зростанням ФАКТОРА3 та ФАКТОРА2, РЕЗУЛЬТАТУ1 та РЕЗУЛЬТАТ2 теж зростатиме.

3.7 Запустити *Майстер візуалізації* та на другому кроці обрати подання результатів у вигляді *Статистики* (рис.51).

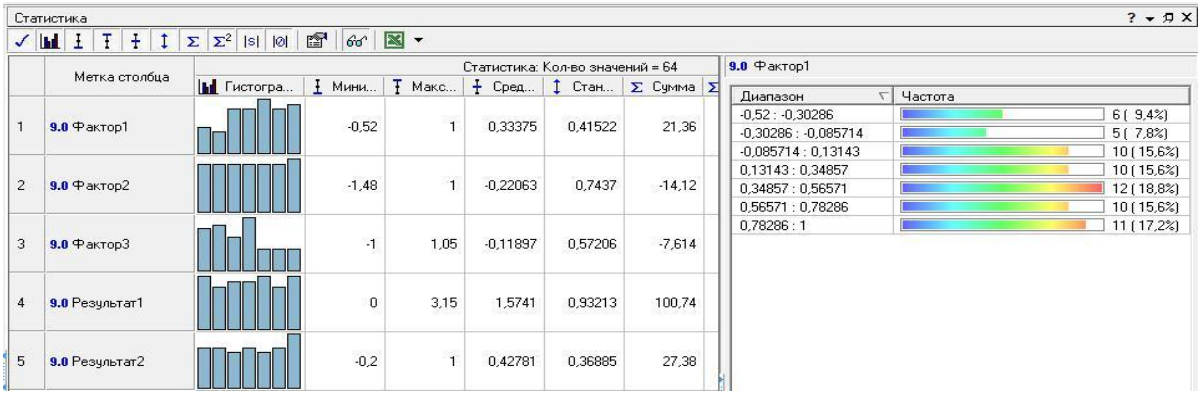


Рисунок 51 – Подання результатів кореляційної обробки в вигляді *Статистики*

3.8 Ознайомитися з можливостями кнопок меню режиму *Статистика* та виконати аналіз статистичних даних.

3.9 Запустити *Майстер візуалізації* та на другому кроці обрати подання результатів у вигляді *Багатовимірної діаграми* (рис.37).

3.10 Задати параметри подання багатовимірної діаграми: вісь *X* – ФАКТОР1, вісь *Y* – ФАКТОР2, вісь *Z* – ФАКТОР3, тип діаграми – ПОВЕРХНЯ (рис. 52).

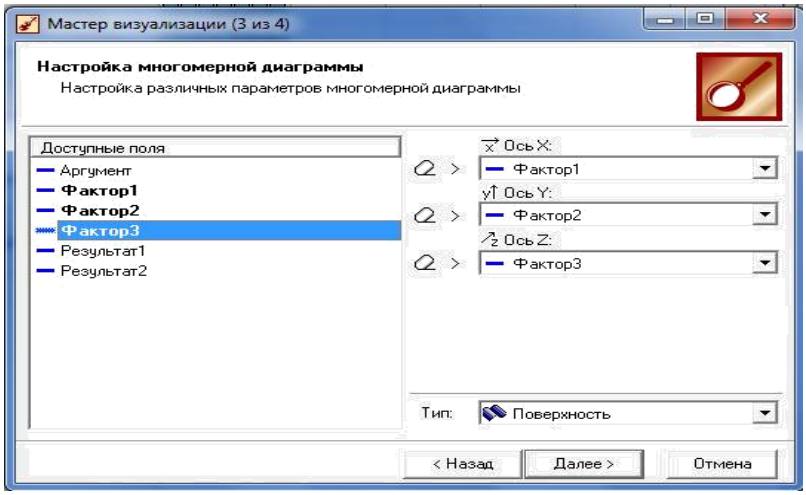


Рисунок 52 – Вікно налаштування багатовимірної діаграми

3.11 Перейти до завершення візуалізації та отримати багатовимірну діаграму (рис.53).

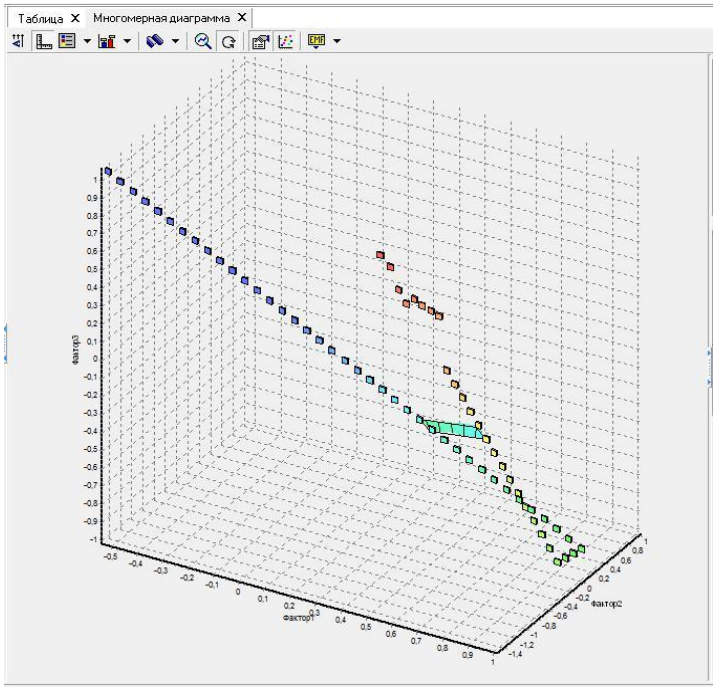


Рисунок 53– Подання результатів кореляційної обробки в вигляді багатовимірної діаграми

3.12 Ознайомитися з можливостями кнопок меню *Багатовимірна діаграма* та проаналізувати отримані результати (рис. 53).

3.13 Запустити *Майстер візуалізації* та на другому кроці обрати подання результатів у вигляді *Гістограми* (рис.37).

3.14 Налаштувати параметри стовпців гістограми як наведено на рис.54.

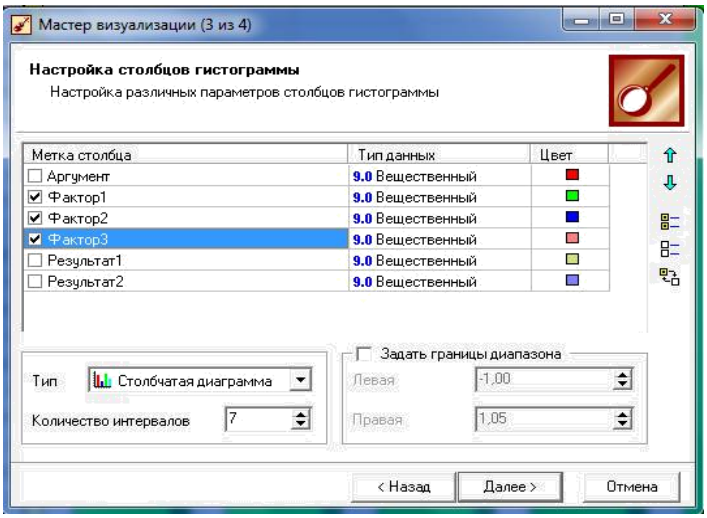


Рисунок 54 – Налаштування параметрів стовпців гістограми

3.15 Перейти до завершення візуалізації та отримати гістограму (рис.55).

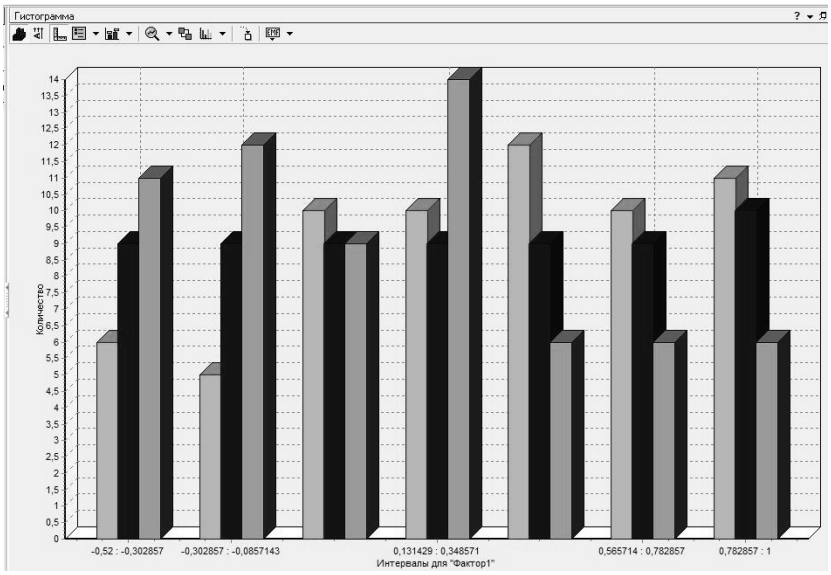


Рисунок 55 – Подання результатів кореляційної обробки в вигляді гістограми

3.16 Ознайомитися з можливостями кнопок меню в режимі *Гістограма* та проаналізувати отримані результати.

4 Самостійно додати до вибірку від 30 до 50 записів та провести факторний та кореляційний аналіз. Порівняти результати.

Контрольні питання

1 З якою метою використовується комплексна передобробка даних?

2 Сформулювати мету факторного аналізу.

3 В чому полягає сутність кореляційного аналізу?

4 Яка характеристика є критерієм прийняття рішення про вилучення деяких факторів з результуючого набору даних?

Лабораторна робота 3
ПРОГНОЗУВАННЯ ЗА ДОПОМОГОЮ ЛІНІЙНОЇ РЕГРЕСІЇ

Мета роботи – побудова прогнозу продажу товару за допомогою лінійної регресії.

Завдання для підготовки до виконання лабораторної роботи. Побудувати лінійну регресію на основі даних, наведених на рис.56.

1 Загальні положення

Лінійну регресію доцільно обирати, коли передбачається лінійна залежність між вхідними факторами та результатом [23]. Перевагою її можна назвати швидкість обробки вхідних даних і простоту інтерпретації отриманих результатів.

	A	B	C	D	E	F	G	H	I	J
1	Дата	Кількість1	Кількість2	Кількість3	Кількість4	Кількість5	Кількість6	Кількість7	Кількість8	Кількість
2	27.03.2013	53365325	527480510,6	137354647	86921976	201349100	265512940	171050704	275040895	1718076097
3	28.03.2013	54227369	198516168,6	181436944	202820528	135672326	139351413	263509423	62400501	1237934672
4	29.03.2013	52942963	4260918031	265503271	153709493	253582727	119835735	19586783	174302033	5300381036
5	30.03.2013	48287549	4019468009	137712196	161988204	251555638	252773717	153428132	269031717	5294245162
6	31.03.2013	25706639	2418370399	237621628	185489251	112478015	64935898	295065872	223639178	3563306880
7	01.04.2013	63508662	2881809040	268517885	285071835	76662968	189444003	480896	63143283	3828638571
8	02.04.2013	3330110,8	1971831714	168237654	158058549	95963935	214572711	129147784	279625100	3020767558
9	03.04.2013	49231371	5178244606	297666540	3018607	233300129	224244468	111717344	139152564	6236575629
10	04.04.2013	21424375	5861947997	152894957	35874849	27612	255494758	108701973	109098606	6545465127
11	05.04.2013	80572976	4422206993	228629911	110907026	20405269	44135714	200046455	53102605	5160006948
12	06.04.2013	10510027	1427879899	42097967	97210213	221841943	123844649	204723635	53270388	2181378721
13	07.04.2013	7158462,7	3539732111	283504508	52192642	192134512	197665718	146213100	137924147	4556525201
14	08.04.2013	38088380	396101299,4	129967847	285108698	68082438	47077583	24861977	39460828	1028749050
15	09.04.2013	31615400	4075147046	87825468	201735264	92761229	133124993	278206310	111672102	5012087812
16	10.04.2013	18278193	2383969015	27844216	191548264	220783857	212655689	40742082	252586683	3348407999
17	11.04.2013	10949518	808148967,1	48370031	252824502	50027156	268956507	22752409	205546405	1667575496
18	12.04.2013	1031298,2	4533447420	299408129	16157837	283956402	92363424	99428607	252170021	5577963139
19	13.04.2013	62713361	2202832713	1710206	88170718	154070282	128009416	84003030	292499123	3014008849

Рисунок 56 – Вихідні дані

Аналітику достатньо вказати вхідні (фактори) та вихідні (результат) стовпці, спосіб розбиття даних на тестову множину та множину, що навчає, і запустити процес навчання. Причому після цього будуть досяжні всі механізми візуалізації та аналізу даних, що дозволять побудувати прогноз, провести експеримент "що-якщо", дослідити залежність результату від значень вхідних факторів, оцінити якість побудованої моделі за діаграмою розсіювання. Також результати роботи цього алгоритму можуть підтвердити або спростувати гіпотезу щодо лінійного характеру залежності.

В результаті роботи даного компонента будується лінійна модель даних за наступним алгоритмом.

Нехай є набір вхідних значень X_i, де $i=1,...n$, тобто $X=\{x_1.x_2,....x_n\}$. Тоді можна вказати таких набір вихідних значень Y_j , де $j=1,....m$, який буде відповідати лінійній комбінації вхідних значень з коефіцієнтами a_i, де $i=1,...n$:

$$[1,.x_2,....x_n][a_0,a_1,a_2,...a_n][y_1,y_2,...y_m].$$

За умови одиничності вихідного значення можна записати:

$$a_0+ a_1+ a_2x_2+....+ a_nx_n=y.$$

Таким чином, задача зводиться до підбору коефіцієнтів a_i, оцінка яких здійснюється за методом найменших квадратів (МНК).

Інструмент *Прогнозування* з'являється в меню *Майстра обробки* тільки після побудови моделі прогнозу: лінійної регресії, нейромережі тощо. Прогнозувати на кілька кроків вперед має сенс тільки часовий ряд (наприклад, якщо є дані щодо тижневих сум продажів за певний період, можна спрогнозувати суму продажів на два тижні вперед).

2Порядок виконання лабораторної роботи

1Створити файл в табличному процесорі (*MS Excel*, *Calc)* (рис.56).

1.2 Зберегти файл в форматі «Текстовий файл (з розподільниками табуляції)»

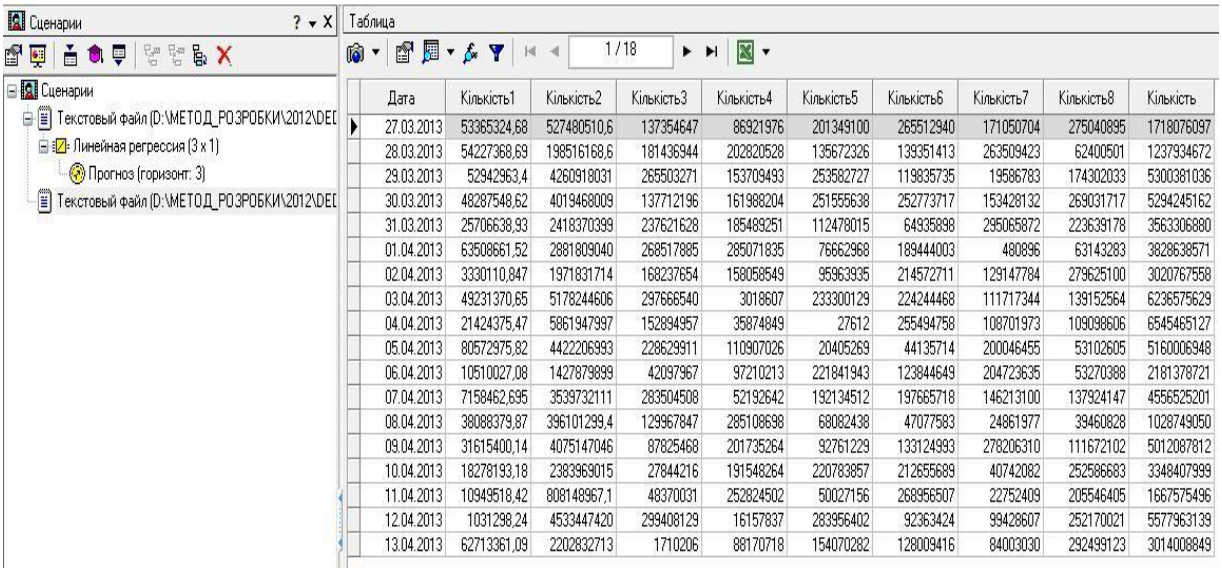
1.3 Запустити аналітичну платформу *Deductor* та ініціювати *Майстер імпорту* з метою завантаження текстового файлу в *Deductor*.

1.4 У вікні *Майстра імпорту (6 з 9)* (рис.31) налаштувати параметри кожного поля: ім'я та мітку стовпця, тип та вид даних, призначення.

Будемо вважати, що майбутні продажі підпорядковуються тому ж закону, що і попередні, отже будемо будувати прогноз, спираючись на інформацію за 3 попередні місяці. Вхідними стовпцями вкажемо поля КІЛЬКІСТЬ1, КІЛЬКІСТЬ2, КІЛЬКІСТЬ3 (дійсні, безперервні), а вихідним полем КІЛЬКІСТЬ.

1.5 Запустити процес імпорту, натиснувши кнопку Пуск в вікні *Майстра імпорту* (*7 з 9*) та кнопку Далі.

1.6 На наступному кроці обрати як спосіб відображення даних *Таблиця* (рис.57).



Дата	Кількість1	Кількість2	Кількість3	Кількість4	Кількість5	Кількість6	Кількість7	Кількість8	Кількість
27.03.2013	53365324,68	527480510,6	137354647	86921976	201349100	265512940	171050704	275040895	1718076097
28.03.2013	54227368,69	198516168,6	181436944	202820528	135672326	139351413	263509423	62400501	1237934672
29.03.2013	52942963,4	4260918031	265503271	153709493	253582727	119835735	19586783	174302033	5300381036
30.03.2013	48287548,62	4019468009	137712196	161988204	251555638	252773717	153428132	269031717	5294245162
31.03.2013	25706638,93	2418370399	237621628	185489251	112478015	64935898	295065872	223639178	3563306880
01.04.2013	63508661,52	2881809040	268517885	285071835	76662968	189444003	480896	63143283	3828638571
02.04.2013	3330110,847	1971831714	168237654	158058549	95963935	214572711	129147784	279625100	3020767558
03.04.2013	49231370,65	5178244606	297666540	3018607	233300129	224244468	111717344	139152564	6236575629
04.04.2013	21424375,47	5861947997	152894957	35874849	27612	255494758	108701973	109098606	6545465127
05.04.2013	80572975,82	4422206993	228629911	110907026	20405269	44135714	200046455	53102605	5160006948
06.04.2013	10510027,08	1427879899	42097967	97210213	221841943	123844649	204723635	53270388	2181378721
07.04.2013	7158462,695	3539732111	283504508	52192642	192134512	197665718	146213100	137924147	4556525201
08.04.2013	38088379,87	396101299,4	129967847	285108698	68082438	47077583	24861977	39460828	1028749050
09.04.2013	31615400,14	4075147046	87825468	201735264	92761229	133124993	278206310	111672102	5012087812
10.04.2013	18278193,18	2383969015	27844216	191548264	220783857	212655689	40742082	252586683	3348407999
11.04.2013	10949518,42	808148967,1	48370031	252824502	50027156	268956507	22752409	205546405	1667575496
12.04.2013	1031298,24	4533447420	299408129	16157837	283956402	92363424	99428607	252170021	5577963139
13.04.2013	62713361,09	2202832713	1710206	88170718	154070282	128009416	84003030	292499123	3014008849

Рисунок 57 – Імпортована таблиця

7 Ініціювати *Майстер обробки* (F7), в групі *Data Mining* обрати *Лінійна регресія* та задати на другому кроці *Майстра обробки* призначення початкових стовпців даних (рис.58).

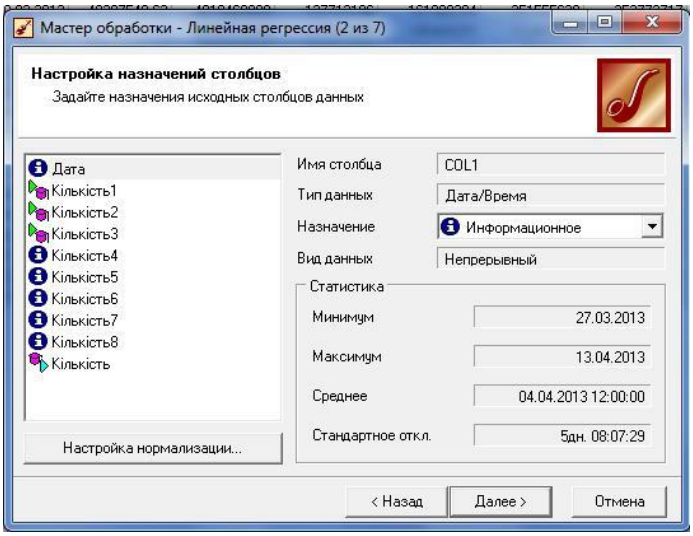


Рисунок 58– Налаштування призначення стовпців

8 На наступному кроці *Майстра обробки* виконати налаштування тестової множин та множини, що навчає та спосіб розбиття початкової множини даних (рис.59).

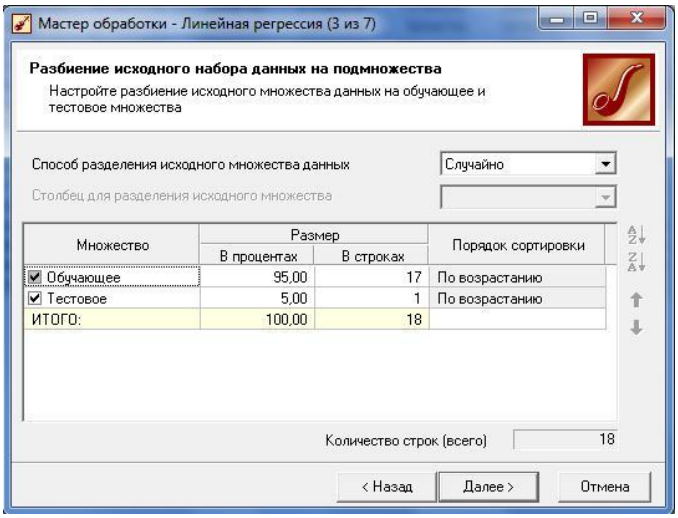


Рисунок 59– Налаштування розбиття початкового
набору даних на підмножини

9 Залишити на третьому кроці обмеження діапазону вхідних значень без змін. При натисканні на кнопці Далі з’являється вікно запуску процесу навчання. В процесі виконання видно, яка частина даних розпізнана на етапах навчання та тестування.

10 Запустити процес побудови лінійної моделі (рис.60).

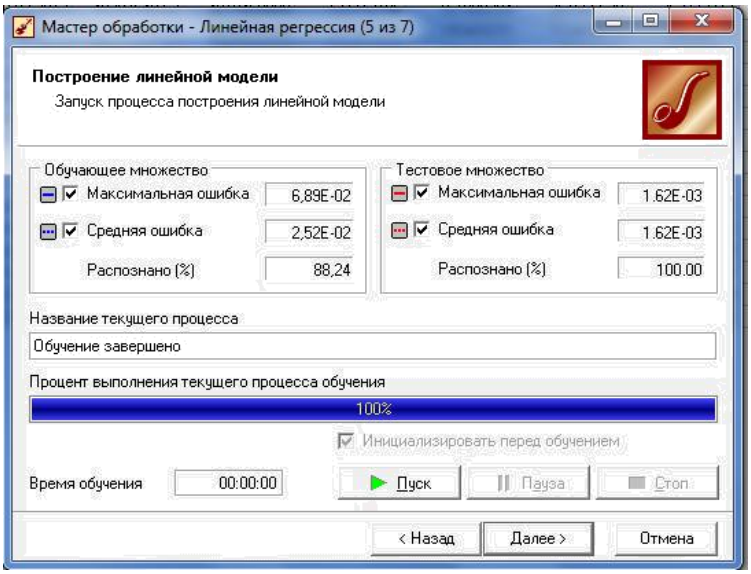


Рисунок 60– Побудова лінійної моделі

11 Обрати як спосіб відображення *Коефіцієнти регресії* та *Діаграму розсіювання* (рис.61). На діаграмі розсіювання поля КІЛЬКІСТЬ та КІЛЬКІСТЬ_OUT – це реальні та прогнозні значення відповідно.

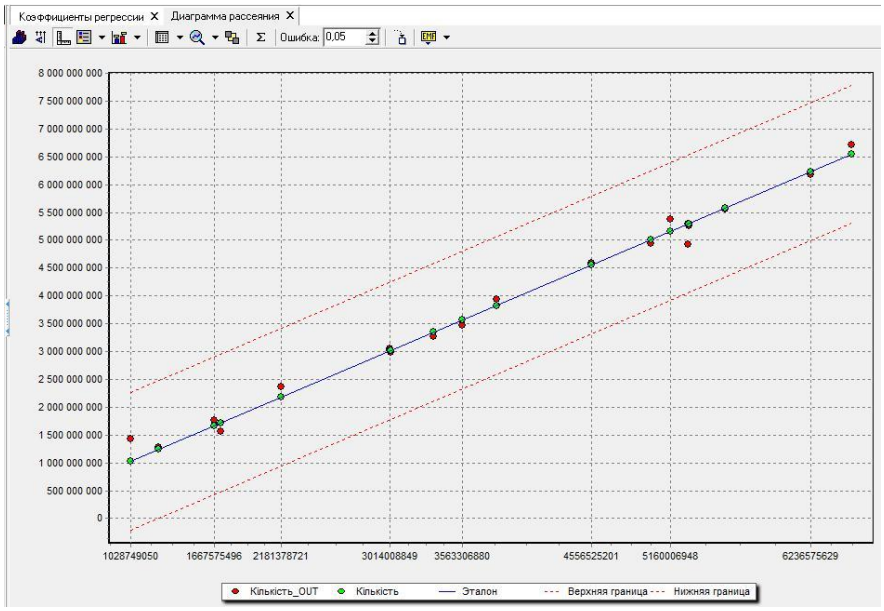


Рисунок 61 – Діаграма розсіювання

12 Запустити *Майстер обробки* та обрати *Прогнозування* (*Прогнозування часового ряду*).

13 Налаштувати на першому кроці обробника зв’язки стовпців для прогнозування (рис.62).

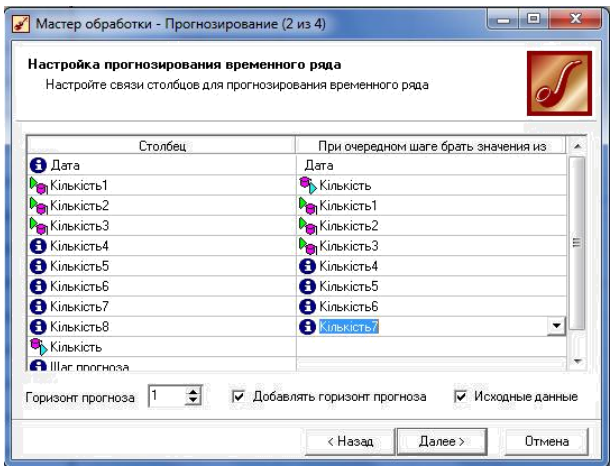


Рисунок 62– Налаштування прогнозування часового ряду

14 На 3 кроці *Майстра обробки – Прогнозування* обрати як спосіб візуалізації *Діаграму прогнозу*.

15 Налаштувати стовпці *Діаграми прогнозу* (рис.63) та побудувати її (рис.64).

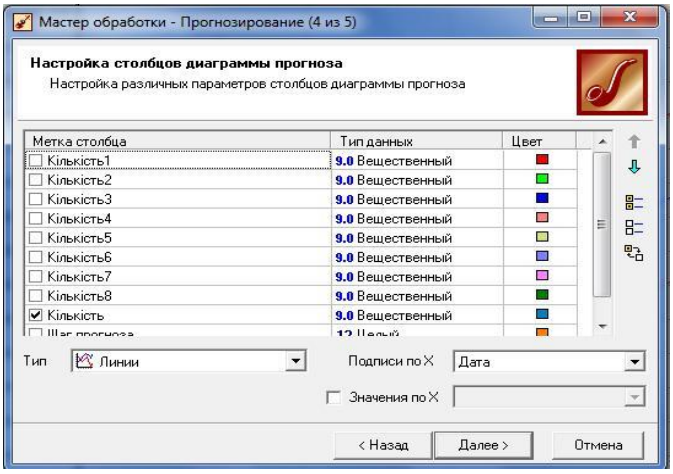


Рисунок 63– Налаштування стовпців діаграми прогнозу

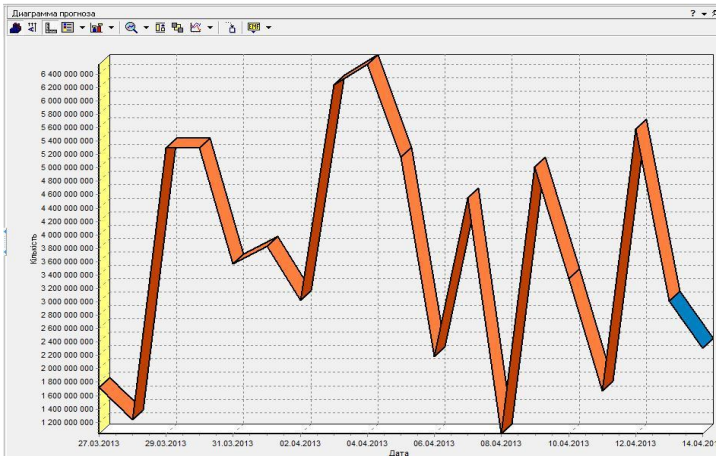


Рисунок 64 – Побудована діаграма прогнозу

Базуючись на моделі, яка побудована за допомогою лінійної регресії, а саме, за діаграмою прогнозу аналітик здатен розробити рекомендації та надати прогноз щодо продажів.

16 Самостійно додати до запропонованої вибірки 30-50 записів та побудувати модель за допомогою лінійної регресії. Порівняти отримані результати.

Контрольні питання

1 Записати вид функції регресії *f(x, b)*.

2 Охарактеризувати коефіцієнти лінійної регресії.

3 Перелічити переваги використання лінійної регресії в аналітичній платформі *Deductor*.

4 Охарактеризувати механізми візуалізації та аналізу даних, які дозволяють побудувати прогноз.

Лабораторна робота 4
ЛОГІСТИЧНА РЕГРЕСІЯ

Мета роботи – обробка даних та прогнозування подій з використанням можливостей логістичної регресії.

Завдання для підготовки до виконання лабораторної роботи.

Провести регресійний аналіз та побудувати *ROC*-криву (рис.65).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	*№№*	*Прізвище*	*Адреса*	*Стать*	*Вік*	*Розмір позики*	*Строк позики*	*Мета позики*	*С/м дохід*	*С/м витрати*	*Наявність нерухом*	*Наявність авто*	*Наявність банк/рахунку*	*Наявність страховки*	*Стаж роботи*	*Давати кредит*
2	1	Мищенко	Харків	ч	50	12000	12	авто	6000	3000	так	так	так	так	30	так
3	2	Бурейко	Чугуїв	ж	38	20000	6	побутова/тех	7000	2000	так	ні	так	ні	20	так
4	3	Плющ	Харків	ч	54	6500	6	будівництво	1000	500	ні	так	ні	ні	37	ні
5	4	Савченко	Зміїв	ч	40	7200	6	побутова/тех	1200	500	так	так	ні	ні	22	ні
6	5	Левченко	Чугуїв	ж	34	23000	12	авто	10000	5000	так	ні	так	так	12	так
7	6	Шевченко	Зміїв	ч	38	7800	6	будівництво	1000	500	ні	так	так	ні	16	ні
8	7	Сахно	Зміїв	ч	49	6000	6	будівництво	2000	500	так	ні	ні	ні	32	так

Рисунок 65 – Вихідні дані

1 Загальні положення

Логістична регресія – це різновид множинної регресії, загальне призначення якої полягає в аналізі зв'язку між кількома незалежними змінними (званими також регресорами або предикторами) і залежною змінною. Бінарна логістична регресія застосовується у разі, коли залежна змінна є бінарною.

За допомогою логістичної регресії можна оцінювати ймовірність того, що подія настане для конкретного випробуваного (хворий/здоровий, повернення кредиту/дефолт і т.д.).

Всі регресійні моделі можуть бути записані у вигляді формули:

$$y=F(x_1,x_2,...x_n)$$

Задача регресії може бути сформульована таким чином: передбачити безперервну змінну зі значеннями на відрізку [0,1] при будь-яких значеннях незалежних змінних. Це досягається застосуванням наступного регресійного рівняння (*логіт-перетворення*):

$$P = \frac{1}{1+e^{-y}},$$

де

P – ймовірність того, що подія відбудеться;

$y=F(x_1,x_2,...x_n)$ – стандартне рівняння регресії.

Перетворення, яке зазвичай називають *логістичним* або логіт-перетворенням, має вигляд:

$$P' = \log_e (P/(1-P))$$

Існує кілька способів знаходження коефіцієнтів логістичної регресії. На практиці часто використовують метод *максимальної правдоподібності*, який застосовується в статистиці для отримання оцінок параметрів генеральної сукупності за даними вибірки.

ROC-крива (Receiver Operator Characteristic) – крива, яка найчастіше використовується для подання результатів бінарної класифікації в машинному навчанні. Оскільки класів два, один з них називається класом з позитивними наслідками, другий – з негативними наслідками. *ROC*-крива показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів. У термінології *ROC*-аналізу перші називаються істинно позитивними, другі – помилково негативними множинами. При цьому передбачається, що у класифікатора є деякий параметр, варіюючи який, можна отримати те чи інше розбиття на два класи. Цей параметр часто називають *порогом* або точкою відсікання. В залежності від його величини отримують різні величини *помилок I і II роду*.

У логістичній регресії поріг відсікання змінюється від 0 до 1, це і є розрахункове значення рівняння регресії, яке називають *рейтингом*.

Розуміння суті помилок I і II роду дає *таблиця спряженості*, яка будується на основі результатів класифікації моделлю і фактичною (об'єктивною) приналежністю випадків до класів (табл.5).

Таблиця 5 *Таблиця спряженості*

	Фактично	
Модель	Позитивно	Негативно
позитивно	TP	FP
негативно	FN	TN

TP (True Positives) – вірно класифіковані позитивні випадки (так звані істинно позитивні випадки);

TN (True Negatives) – вірно класифіковані негативні випадки (істинно негативні випадки);

FN (False Negatives) – позитивні випадки, класифіковані як негативні (помилка I роду). Це так званий "помилковий пропуск", коли подія помилково не виявляється (хибно негативні випадки);

FP (False Positives) – негативні випадки, класифіковані як позитивні (помилка II роду); Це помилкове виявлення, тому що за відсутності події помилково виноситься рішення про її наявність (хибно позитивні випадки).

При аналізі частіше оперують не абсолютними, а відносними показниками – частками, поданими у відсотках.

Частка істинно позитивних випадків розраховується за формулою:

$$TPR = \frac{TP}{TP + FN} \cdot 100\%.$$

Частка хибно позитивних випадків розраховується за формулою:

$$FPR = \frac{FP}{TN + FP} \cdot 100\%.$$

Чутливість (Sensitivity), частка істинно позитивних випадків:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\%.$$

Специфічність (Specificity), частка істинно негативних випадків, які були правильно ідентифіковані моделлю:

$$Sp = \frac{TN}{TN + FP} \cdot 100\%.$$

Чутливістю та специфічністю визначається об'єктивна цінність будь-якого бінарного класифікатора. Треба зауважити, що

$$FPR = 100\% - S_p.$$

Модель з високою чутливістю часто дає істинний результат при наявності позитивного результату (виявляє позитивні випадки). Навпаки, модель з високою специфічністю частіше дає істинний результат при наявності негативного результату (виявляє негативні випадки).

ROC-крива отримують таким чином:

– розраховуються значення чутливості S_e й специфічності S_p для кожного значення порогу відсікання, яке змінюється від 0 до 1 з кроком *dx* (наприклад, 0.01)*;*

– будується графік залежності: по осі *Y* відкладається чутливість S_e, по осі *X – 100% –* S_p (сто відсотків мінус специфічність), або, що те ж саме, *FPR*, частка хибно позитивних випадків. Чисельний показник площі під кривою називається *AUC (Area Under Curve)*, обчислити який можна, наприклад, за допомогою чисельного методу трапецій:

$$AUC = \int f(x)dx = \sum_i \left[\frac{X_{i+1} + X_i}{2}\right] \cdot (Y_{i+1} - Y_i)$$

Можна вважати, що чим більше показник *AUC*, тим кращі прогностичні властивості притаманні моделі. Однак варто пам'ятати, що:

– показник *AUC* призначений для порівняльного аналізу декількох моделей;

– показник *AUC* не містить інформації щодо чутливості й специфічності моделі.

В літературі іноді наводиться така експертна шкала для значень *AUC*, що характеризує якість моделі:

– відмінна якість моделі – 0,9-1,0;

– дуже гарна якість моделі – 0,8-0,9;

– гарна якість моделі – 0,7-0,8;

– середня якість моделі – 0,6-0,7;

– незадовільна якість моделі – 0,5-0,6.

Ідеальна модель має 100% чутливість і специфічність. Однак на практиці досягти цього неможливо, більше того, неможливо одночасно підвищити і чутливість, і специфічність моделі. Компроміс знаходять за допомогою порога відсікання, який впливає на співвідношення S_e і S_p.

Логістична регресія на виході розраховує значення рейтингу, яке можна трактувати як ймовірність того, що подія настане для конкретного випробування. Тому часто бажано вказати, ймовірність якої саме (з двох варіантів вихідного поля) події буде оцінюватися, щоб вона кодувалася істиною [24-26].

2Порядок виконання лабораторної роботи

1 Створити в файл в табличному процесорі (*MS Excel)* (рис.65).

2 Ініціювати *Майстер імпорту*.

3 У вікні *Майстра імпорту (6 з 9)* (рис.31) налаштувати параметри полів: поле ДАВАТИ КРЕДИТ – *Вихідне*, всі числові поля – *Вхідні*, а всі інші поля – *Інформаційні*.

4 Запустити процес імпорту файлу.

5 Ініціювати *Майстер обробки,* в групі *Data Mining* обрати *Логістична регресія* та перевірити на другому кроці *Майстра обробки* призначення

вихідних стовпців даних (рис.66). Провести нормалізацію полів та налаштувати вибірку, що навчає (кнопка Налаштування нормалізації…).

У вікні НАЛАШТУВАННЯ НОРМАЛІЗАЦІЇ ДАНИХ зліва наведено повний список вхідних і вихідних полів (рис.67). При цьому кожне поле позначено міткою, що відповідає виду нормалізації:

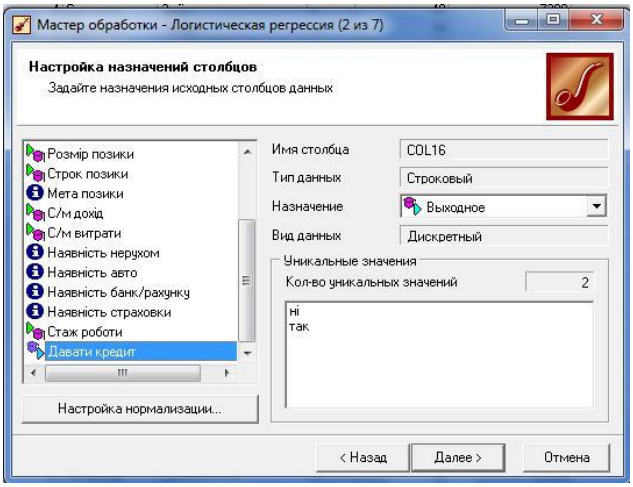


Рисунок 66 – Налаштування призначення стовпців

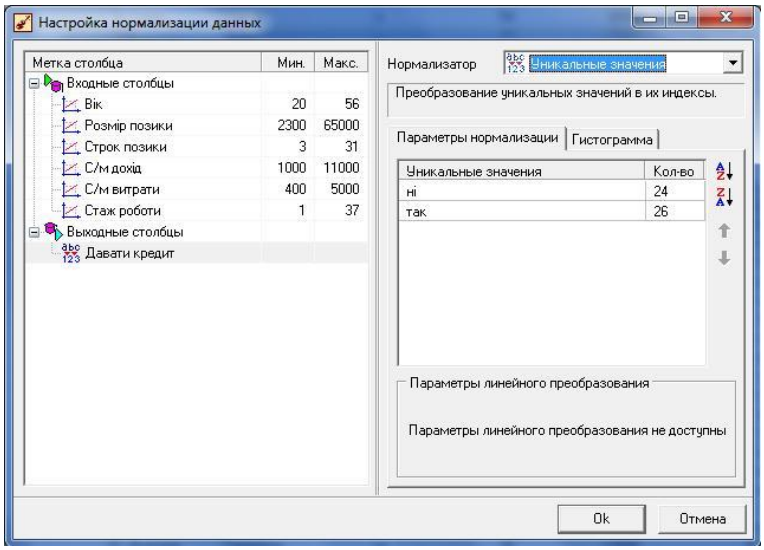


Рисунок 67 – Вікно налаштування нормалізації даних

–лінійна нормалізація вихідних значень;

–унікальні значення, тобто перетворення унікальних значень у їх індекси;

–бітова маска, тобто перетворення дискретних значень в бітову маску.

У правій частині вікна для виділеного поля відображаються параметри нормалізації (рис.67).

Для числових (безперервних) полів з лінійною нормалізацією додаткові параметри недосяжні. У полях МІНІМУМ і МАКСИМУМ можна подивитися мінімальне та максимальне значення цього поля.

Для дискретних полів можуть бути використані два види нормалізації унікальні значення й бітова маска (рис.67).

6Налаштувати вибірку, що навчає, для побудови лінійної моделі (рис.68).

Структура записів множини, що навчає, та тестової однакова: вхідні дані й відповідні вихідні значення, але, якщо перша використовується для навчання моделі, то друга – для перевірки якості навчання.

Примітка: навчання може вважатися успішним з високим ступенем ймовірності, якщо відсоток вірно розпізнаних випадків на тестовій множині достатньо великий (близький до 100%).

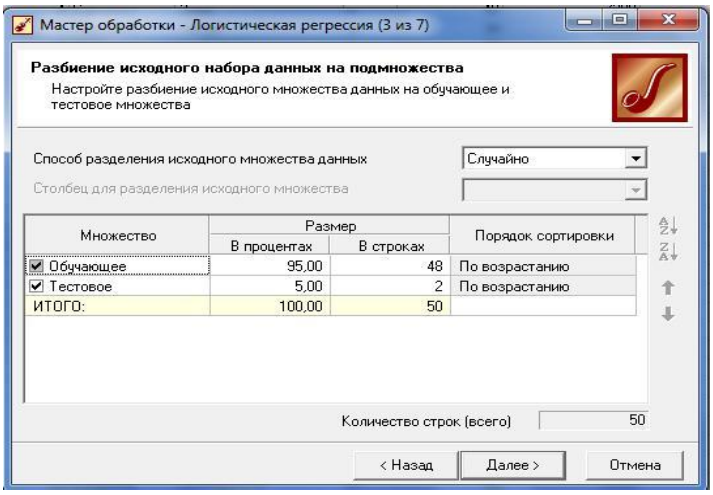


Рисунок 68 – Розбиття початкового набору даних на підмножини

7 На наступному кроці налаштувати параметри завершення навчання, для чого задати максимальну кількість ітерацій (точність), функцію правдоподібності, поріг відсікання та припустиму помилку (рис.69).

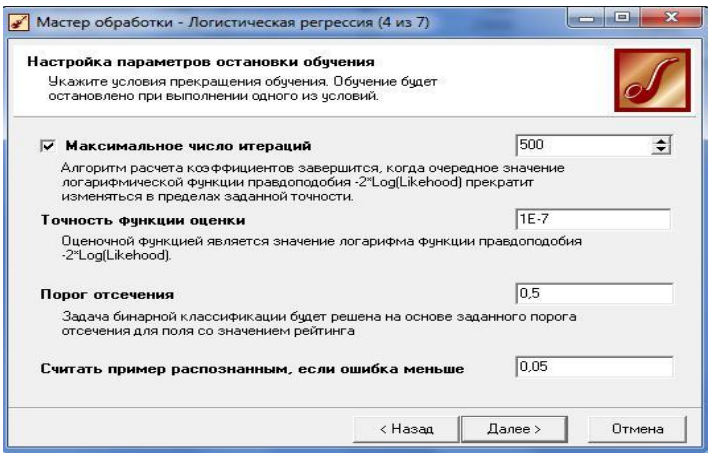


Рисунок 69 – Налаштування параметрів завершення навчання

8 На наступному кроці запустити процес побудови логістичної моделі (рис.70).

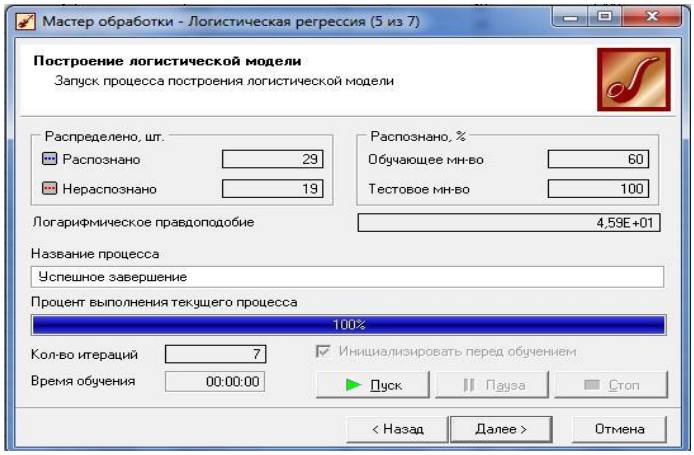
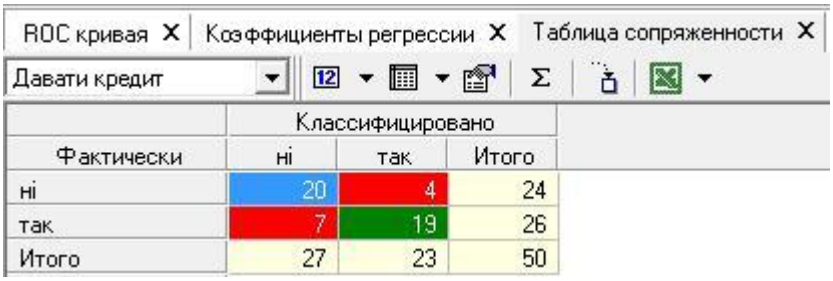


Рисунок 70– Процес побудови логістичної моделі

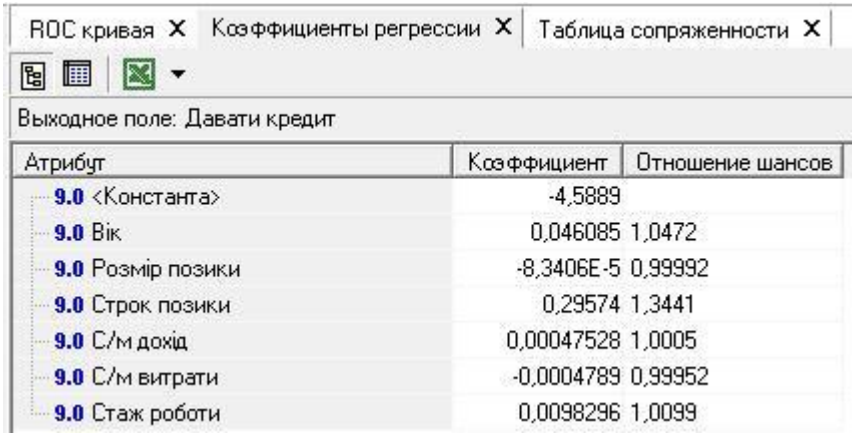
9 Ознайомитись з результатами роботи обробника на вкладках *Таблиця спряженості* (рис.71), *Коефіцієнти регресії* (рис.72) та *ROC*- аналіз. За замовчуванням поріг відсікання дорівнює 0,5.



	Классифицировано		
Фактически	ні	так	Итого
ні	20	4	24
так	7	19	26
Итого	27	23	50

Рисунок 71–Вікно вкладки *Таблиця спряженості*

За чисельними даними комірок таблиці спряженості можна оцінити якість логістичної регресії як класифікатора. Згідно з наведеною на рис. 71 таблицею спряженості за результатами моделювання зафіксовано 7 випадків хибного виявлення та 4 випадки хибного пропуску подій.



Атрибут	Коэффициент	Отношение шансов
9.0 <Константа>	-4,5889	
9.0 Вік	0,046085	1,0472
9.0 Розмір позики	-8,3406E-5	0,99992
9.0 Строк позики	0,29574	1,3441
9.0 С/м дохід	0,00047528	1,0005
9.0 С/м витрати	-0,0004789	0,99952
9.0 Стаж роботи	0,0098296	1,0099

Рисунок 72– Коефіцієнти регресії

Для покрашення показника «доля вірно класифікованих випадків» необхідно знайти оптимальну точку відсікання для віднесення об'єктів до певного класу, що найкраще зробити з застосуванням *ROC*-аналізу, який до того ж дозволяє провести оцінку якості моделі класифікатора та порівняти прогностичні властивості декількох моделей.

За результатами проведення регресійного аналізу буде побудована *ROC*-крива (рис.73).

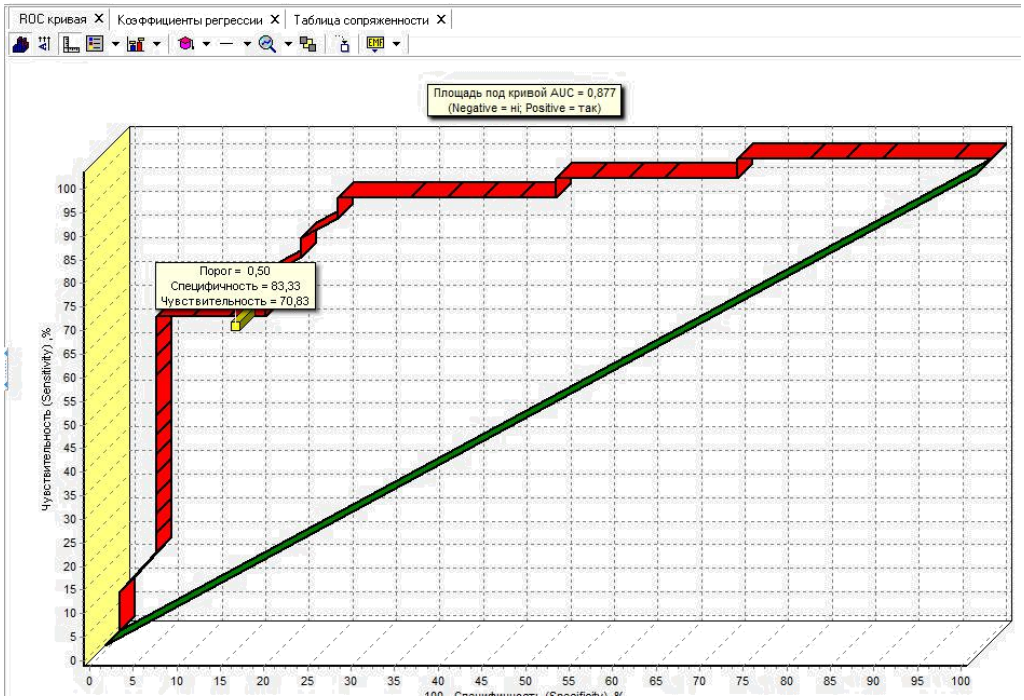


Рисунок 73– Побудована *ROC*-крива

ROC-крива показує залежність кількості вірно класифікованих позитивних випадків від кількості невірно класифікованих негативних випадків. Графік доповнюють прямою y = x.

Для ідеального класифікатора *ROC*-крива проходить крізь верхній лівий кут, де частка істинно позитивних випадків складає 100% або 1,0 (ідеальна чутливість), а частка хибно позитивних випадків дорівнює нулю. Тому чим ближче крива до верхнього лівого кута, тим вище прогнозні якості моделі. Навпаки, чим ближче вона до діагональної прямої, тим модель менш ефективна. Діагональна лінія відповідає "даремному" класифікатору, тобто повній нерозрізненості двох класів.

Треба звернути увагу на оцінку площі під кривою. Теоретично вона змінюється від 0 до 1, але, оскільки модель завжди характеризуються кривою, розташованою вище діагоналі, то зазвичай говорять про зміни від 0,5 ("даремний" класифікатор) до 1,0 ("ідеальна" модель). Ця оцінка може бути отримана безпосередньо обчисленням площі під багатогранником, обмеженим праворуч і знизу вісями координат і зліва вгорі

експериментально отриманими точками (рис.73). Чисельний показник площі під кривою є *AUC*.

В загальних положеннях даної лабораторної роботи наведена експертна шкала для значень *AUC*, за якою оцінюють якість моделі. Побудовану модель (рис.73) можна віднести до «дуже добра якість моделі» (*AUC*=0,877). На *ROC*-кривій наведено також параметри специфічності (83.33) та чутливості (70.83).

10 Додати до запропонованої вибірку 30 -50 записів, провести регресійний аналіз та побудувати *ROC*- криву в такому варіанті. Порівняти отримані результати.

Контрольні питання

1 Дати визначення логістичної регресії.
2 Для розв'язання яких задач призначена логістична регресія?
3 В яких випадках використовується бінарна логістична регресія?
4 Описати способи знаходження коефіцієнтів логістичної регресії.
5 Яку залежність показує *ROC*-крива?
6 Дати визначення істино додатних та хибно від'ємних множин.
7 В яких межах в логістичній регресії знаходиться поріг відсікання?

Лабораторна робота 5
РІШЕННЯ ЗАДАЧІ ПОШУКА АСОЦІАТИВНИХ ПРАВИЛ

Мета роботи – вивчення процесу побудови асоціативних правил.

Завдання для підготовки до виконання лабораторної роботи.

1 Знайти в транзакційній базі даних (рис.74) набори товарів, що найбільш часто зустрічаються, та сформулювати асоціативні правила з визначеними границями значень підтримки та довіри. Знайдені залежності в вигляді правил, використати при розробці рекомендацій щодо раціоналізації збутової політики продуктових товарів та прогнозування попиту на товари.

2 Сформувати асоціативні правила на визначеному часовому інтервалі, спираючись на дані щодо ринкових тенденцій за такими показниками: попит, пропозиція, ціна та положення (рис.75).

Ціна, попит та пропозиція можуть приймати наступні значення: зростати, падати або бути в рівновазі. Положення на ринку може приймати наступні значення: дефіцит, надлишок, рівновага ринку.

Вказати в вікні ІМПОРТ ТЕКСТОВОГО ФАЙЛУ *Майстра імпорту (6-*9) (рис.31) параметри стовпців (табл.6).

	A	B	C
1	№№	Номер транзакції	Товар
2	1	100698	Масло
3	2	100698	Хліб та булки
4	3	100698	Чай
5	4	100747	Хліб та булки
6	5	100747	Соки
7	6	100747	Чай
8	7	101217	Масло
9	8	101217	Хліб та булки
10	9	101217	Молоко
11	10	101243	Масло
12	11	101243	Хліб та булки
13	12	101243	Молоко
14	13	101354	Масло
15	14	101354	Хліб та булки
16	15	101354	Чай
17	16	101567	Хліб та булки
18	17	101567	Кефір
19	18	101567	Йогурт
20	19	101567	Масло
21	20	101567	Молоко

Рисунок 74 – Інформація, створена в *MS Excel*

	A	B	C
1	*№№*	*Дата*	*Ситуація на ринку*
2	1	11012	ціна зростає
3	2	11012	попит падає
4	3	11012	пропозиція зростає
5	4	11012	надлишок
6	5	21012	ціна падає
7	6	21012	попит зростає
8	7	21012	пропозиція падає
9	8	21012	дефіцит
10	9	31009	ціна=210

Рисунок 75 – Фрагмент файлу, який визначає ринкові тенденції

Таблиця 6 – Параметри стовпців

№	*Параметри стовпців*	*Ім'я стовпця*		
		№	*Дата*	*Ситуація на ринку*
1	Тип даних	Інформаційне	Дійсний	Строковий
2	Вид даних		Дискретний	
3	Призначення		ID Транзакція	Елемент

2.1 Додати до вихідних даних 30 – 50 записів.

2.2 Визначити та вказати значення параметрів побудови асоціативних правил: множини, що часто зустрічаються, асоціативні правила.

2.3 Отримати набір асоціативних правил з визначеними границями значень підтримки та довіри.

3 Знайти закономірності при замовленні комплектації електронно-обчислювальної техніки. У магазині електронно-обчислювальної техніки є 4 групи товарів, що пропонуються для комплектації (табл.7).

Таблиця 7 – Групи товарів для комплектації електронно-обчислювальної техніки

Процесор	Системна плата	Чипсет	Пам’ять
Core Duo	Abit AB9	P965	Dual
Duron	Abit SI7	AMD	DDR
P4	Acorp	SiSR	PC
ADM	ALD	KM1	
K62	Amptron	Cx	
Pentium	Aopen	RC	
Celeron	Asrock	nForce	
Xeon	Asus	i820	
	Biostar	KT400	
	Chaintech	Crusoe	
	DFI	KX	
	ECS	MVP	
	Epox	ApolloPro	
	EVGA	Rsint	
	Gigabyte	i440	

За час *t* було продано *n* товарів. Завдання аналітика – визначити закономірність поєднання марок (брендів) комплектуючих електронно-обчислювальної техніки, тобто закономірність поєднання (покупки) елементів груп товарів, що пропонуються. До уваги не береться тип моделі пристрою, враховується лише виробник (бренд) товару. Параметри стовпців подано в табл.8, а фрагмент бази даних в табличному процесорі– на рис. 76.

Таблиця 8 – Параметри стовпців

№№	Параметри стовпця	Ім’я стовпця		
		№№	Зборка номер	Конфігурація ПК
1	Тип даних	Інформаційне	Дійсний	Строковий
2	Вид даних		Дискретний	
3	Призначення		ID Транзакція	Елемент

	A	B
1	*Код*	*Покупка*
2	160698	Core Duo
3	160698	Abit AB9
4	160698	P965
5	160698	Dual
6	160698	ALD
7	160698	ADM
8	160747	Duron
9	160747	Aopen
10	160747	Asrock
11	161217	Asus
12	161217	P965
13	161217	AMD
14	160747	Dual
15	160747	DDR
16	160747	PC
17	161243	Pentium

Рисунок 76 – Фрагмент бази даних в табличному процесорі

3.1 Самостійно додати ще 70 нових записів та провести моделювання з формуванням асоціативних правил.

3.2 Визначити та вказати значення параметрів побудови асоціативних правил: множини, що часто зустрічаються, асоціативні правила.

3.3 Сформулювати асоціативні правила на основі підготовлених статистичних даних про закази, які поступають в магазин.

4 Сформулювати власну економічну задачу та вирішити її з використанням методу *Асоціативні правила.*

1 Загальні положення

Значний обсяг сучасних баз даних викликає стійкий інтерес до нових масштабованих алгоритмів аналізу даних. Одним з популярних методів виявлення знань стали алгоритми пошуку асоціативних правил, метою яких є відшукання закономірностей між пов'язаними подіями.

Асоціативне правило має вигляд: "З події *A* випливає подія *B*". У результаті такого виду аналізу встановлюються закономірності виду: "Якщо в транзакції зустрівся набір товарів (або набір елементів) *A*, то можна дійти висновку, що в цій же транзакції повинен з'явитися набір елементів *B*". Встановлення такого роду закономірностей дозволяє формулювати прості та логічні висловлювання – асоціативні правила.

Основними характеристиками асоціативного правила є *підтримка* і *достовірність*. Розглянемо їх сутність на прикладі.

Припустимо, є транзакційна база даних з наборами товарів, що включають, наприклад, {хліб, молоко, масло}. Виразимо цей набір за допомогою змінних *abc = {a, b, c}.*

Підтримкою називають кількість або відсоток транзакцій, що містять певний набір даних. Якщо цей набір товарів зустрічається в базі даних три рази, то його підтримка дорівнюватиме 3: *SUP (abc) = 3.*

Підтримку іноді також називають *забезпеченням* набору.

Набори називають такими, що часто зустрічаються, якщо їх підтримка вище визначеного користувачем мінімального значення (min support).

Правило має підтримку *s*, якщо *s* процентів транзакцій набору містять одночасно елементи *A* і *B*, тобто обидва товара.

Достовірність правила показує ймовірність того, що з події *A* витікає подія *B*. Правило "З *A* витікає *B*" справедливо з достовірністю *c*, якщо *c* процентів транзакцій множини, що містить набір елементів *A*, також містить набір елементів *B*. За використання алгоритмів пошуку асоціативних правил аналітик здатен отримати множину правил "З *A* витікає *B*" з різними значеннями підтримки та достовірності. На практиці кількість правил обмежують наперед заданими мінімальним і максимальним значеннями підтримки та достовірності.

Занадто велике значення підтримки дозволяє алгоритму виявити лише тривіальні правила, в той час як занадто низьке значення призведе до знаходження величезної кількості правил, які будуть недостатньо обґрунтованими та некорисними для аналітика. Тобто необхідно знайти таку "золоту середину", яка з одного боку дозволить знаходити неочевидні закономірності, а з іншого – забезпечить їх підґрунтям. За умови низького рівня достовірності, цінність правил викликає серйозні сумніви.

На сьогоднішній день найбільш поширеним алгоритмом пошуку асоціативних правил є алгоритм Apriori.

Асоціативні правила знайшли широке застосування в наступних галузях:

а) роздрібна торгівля, а саме визначення товарів, які варто просувати спільно, вибір місця розташування товару в магазині, аналіз споживчого кошика, прогнозування попиту тощо;

б) перехресні продажі, тобто якщо ви маєте інформацію про те, що якісь клієнти придбали продукти *A*, *Б* і *B*, то з певною мірою достовірності можете сказати, що вони куплять продукт *Г*;

в) маркетинг, а саме пошук ринкових сегментів, тенденцій купівельної поведінки;

г) сегментація клієнтів, тобто виявлення загальних характеристик груп клієнтів компанії, виявлення груп покупців;

д) оформлення каталогів з урахуванням аналізу пріоритетів збутових кампаній фірми, визначення послідовностей покупок клієнтів (яка покупка слідує за покупкою товару *A*);

є) аналіз Web-логів [2].

2Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc)* транзакційну базу даних, записи якої містять номери чеків та інформацію про товари, який було придбано (рис.74). Ввести 30 записів з даними, що повторюються [1].

2 Запустити аналітичну платформу *Deductor* та за допомогою *Майстра імпорту* ввести дані з побудованого файлу в аналітичну платформу *Deductor*.

3 Вказати в вікні ІМПОРТ ТЕКСТОВОГО ФАЙЛУ *Майстра імпорту (6 з* 9) (рис.77) параметри стовпців (табл.9).

Таблиця 9 – Параметри стовпців

№	*Параметри стовпців*	*Ім'я стовпця*		
		№№	*Номер тран закції*	*Товар*
1	Тип даних	Інформаційне	Дійсний	Строковий
2	Вид даних		Дискретний	
3	Призначення		ID Транзакція	Елемент

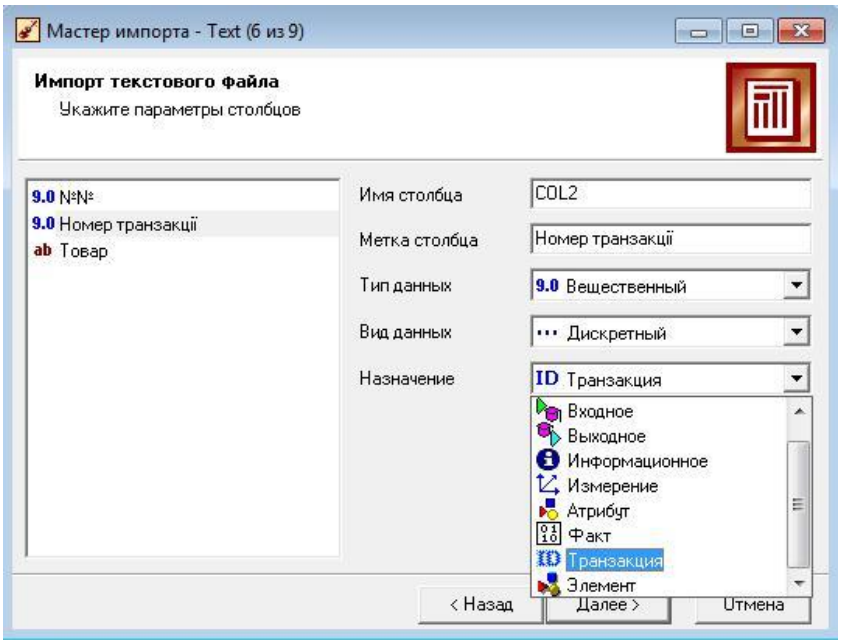


Рисунок 77 – Завдання параметрів стовпців

4 Визначити в вікні *Майстра імпорту (8 з 9*) спосіб відображення даних – *Таблиця* та *Статистика*. Результат імпорту бази даних з файлу табличного процесору (*MS Excel)* в середовище *Deductor* матиме визгяд як на рис.78.

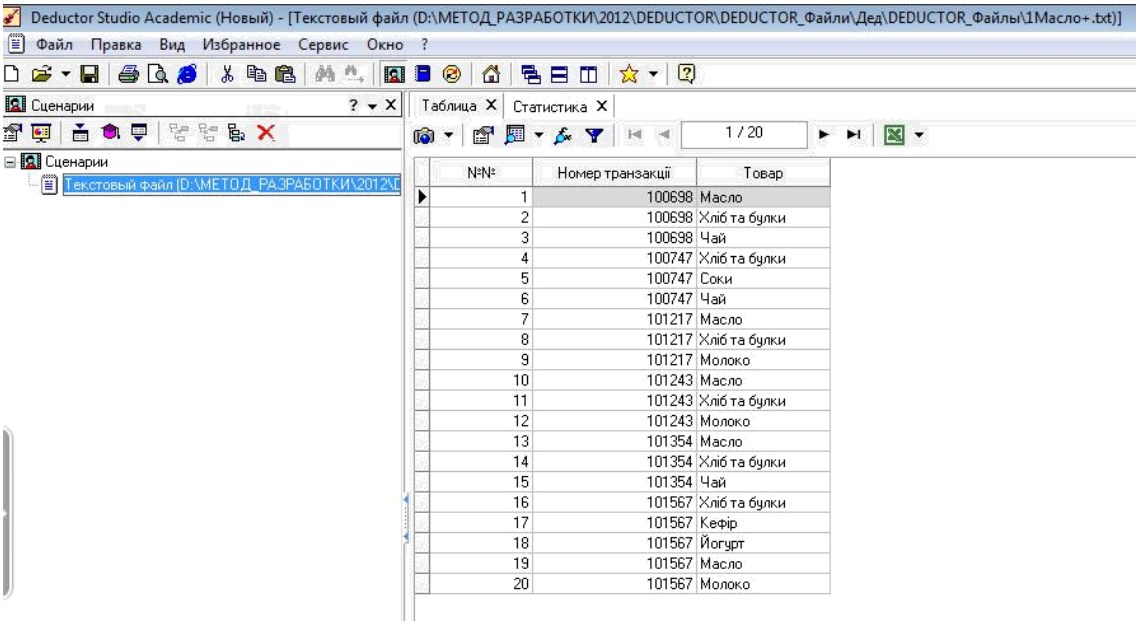


Рисунок 78 – Транзакційна база даних

5 В групі *Data Mining Майстеар обробки* обрати метод *Асоциативні правила* та натиснути кнопку Далі.

6 На другому кроці *Майстра обробки* перевірити призначення вихідних стовпців даних. Вони повинні мати тип *ID Транзакція* та *Елемент*.

7 На третьому кроці налаштувати параметри пошуку правил (рис. 79), тобто встановити мінімальне і максимальне значення підтримки та достовірності. Це найбільш "відповідальний" момент в формуванні набору правил. Вибір здійснюється згідно з наявним досвідом аналізу подібних даних, інтуїції або ж визначається в ході експериментів.

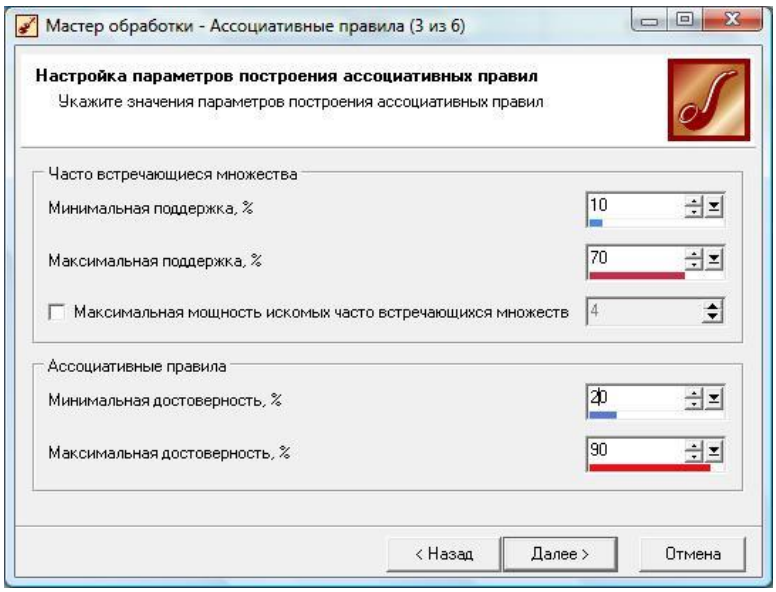


Рисунок 79 – Налаштування параметрів побудови асоціативних правил

Встановити такі границі параметрів пошуку: мінімальний і максимальний рівень підтримки (розділ *Множини, що часто зустрічаються*) 10% та 70% відповідно, мінімальний і максимальний рівень достовірності (розділ *Асоціативні правила*) – 20% та 90% відповідно. Ці значення було отримано експериментальним шляхом і виявились достатньо ефективними (за рівня підтримки від 30% до 50%, набір правил взагалі не формується).

8 Натиснути кнопку Далі. На наступному кроці *Майстра* запускається процес пошуку асоціативних правил. В результаті виводиться інформація про кількість множин і знайдених правил у вигляді *Гістограми розподілу* множин, що часто зустрічаються, за їх потужністю (рис. 80). Отримано 4 правила, кількість множин – 10.

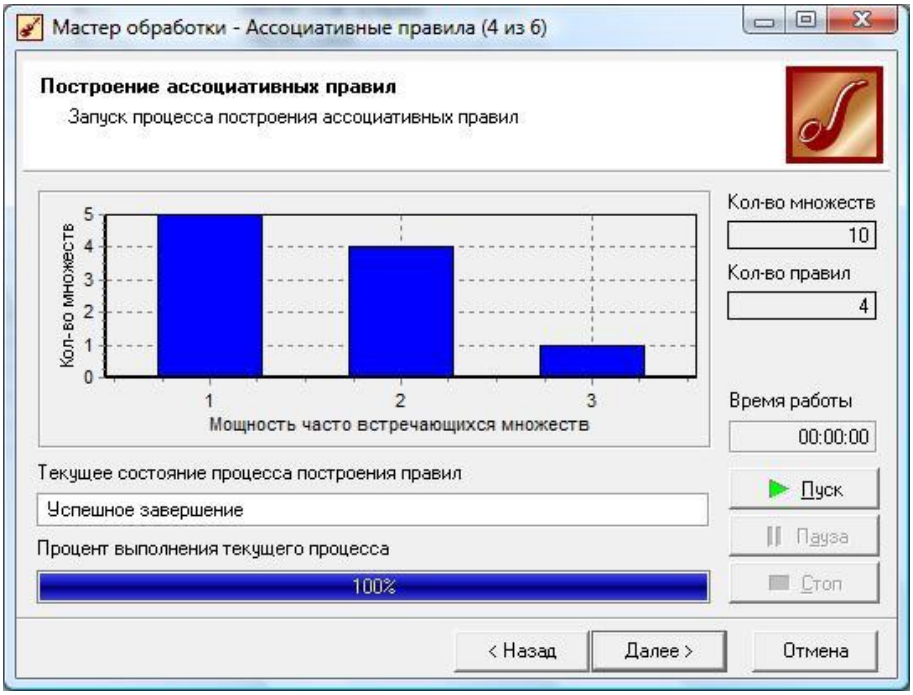


Рисунок 80 – Процес побудови асоціативних правил

9 На наступному кроці для перегляду отриманих результатів обрати візуалізатори *Правила, Популярні набори, Дерево правил, Що-якщо.*

Візуалізатор *Популярні набори* дозволяє виявити набори з одного або декількох товарів, що найчастіше одночасно зустрічаються в транзакціях та характеризуються *підтримкою*.

Для нашого завдання за умови заданих параметрів популярні набори наведені на рис. 81. Для визначення рейтингу популярності товарів і їх наборів доцільно відсортувати їх за рівнем підтримки.

Візуалізатор *Правила* подає інформацію у вигляді списку правил "умова-наслідок" (рис.82), які характеризується значенням підтримки в абсолютному і відсотковому вираженні, а також достовірністю. Таким чином, аналітику надається можливість ознайомитись з моделлю поведінку покупців, поданою у вигляді набору правил.

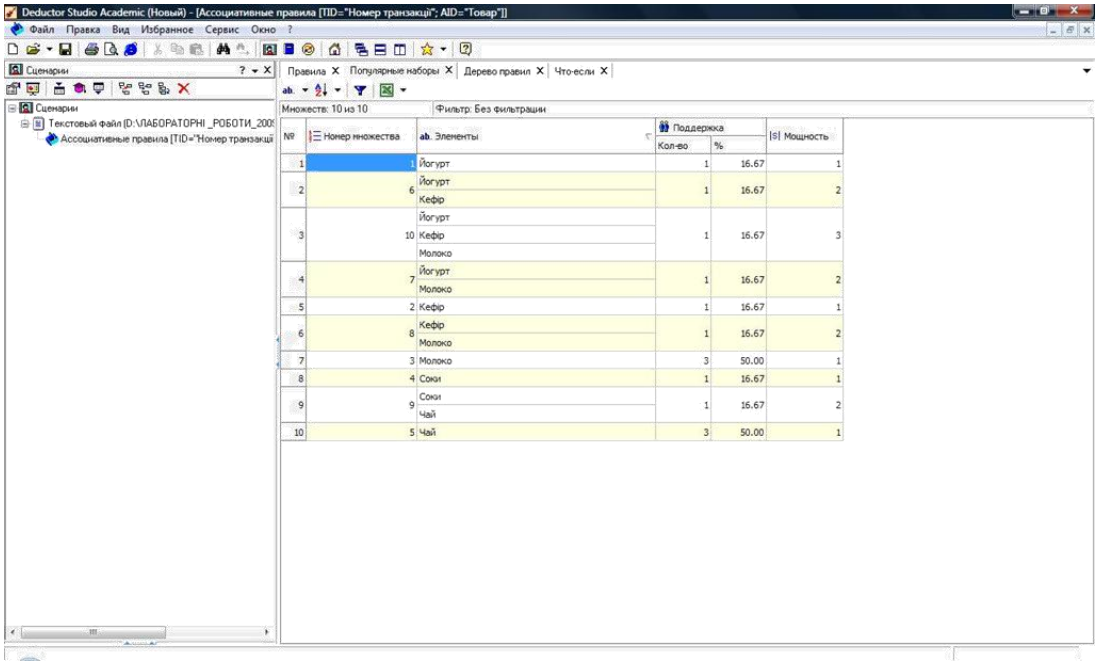


Рисунок 81 – Візуалізатор *Популярні набори*

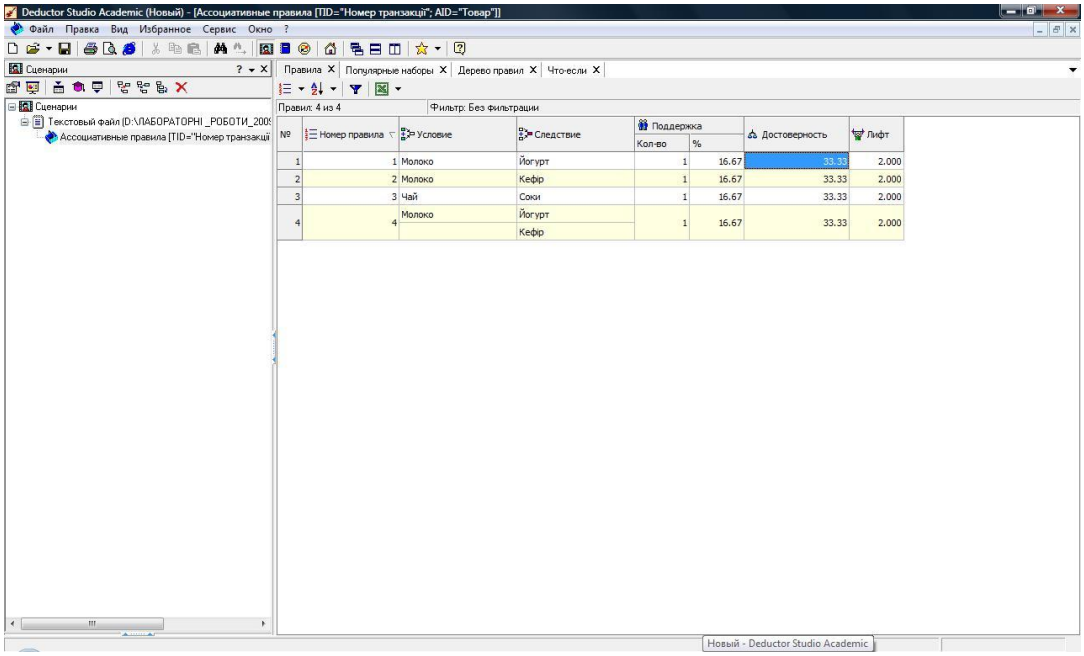


Рисунок 82 – Візуалізатор *Правил*

Ліфт *lift (XY)* – це відношення частоти появи умови у транзакціях, які також містять наслідок, до частоти появи слідства в цілому:

$$lift\ (XY) = conf\ (XY) / supp\ (Y).$$

Якщо значення ліфта перевищує одиницю, це означає, що умова частіше з'являється у транзакціях, які також містять і наслідок.

При великій кількості знайдених правил і широкому асортименті товарів аналізувати отримані правила досить складно. Для зручності аналізу таких наборів правил пропонуються візуалізатори *Дерево правил* і *Що-якщо*.

Візуалізатор *Дерево правил* – це дворівнева ієрархія, яка може бути побудована за двома критеріями: за умовою і за наслідком. Якщо дерево побудоване за умовою, то вгорі списку відображається умова правила, а список, що додається до даної умови, складається з його наслідків. При виборі певної умови, в правій частині візуалізатора відображаються наслідки умови, рівень підтримки та достовірності.

У разі побудови дерева за наслідком, вгорі списку відображається наслідок правила, а список складається з його умов. При виборі певного наслідку, в правій частині візуалізатора користувач бачить умови цього правила із зазначенням рівня підтримки та достовірності (рис.83).

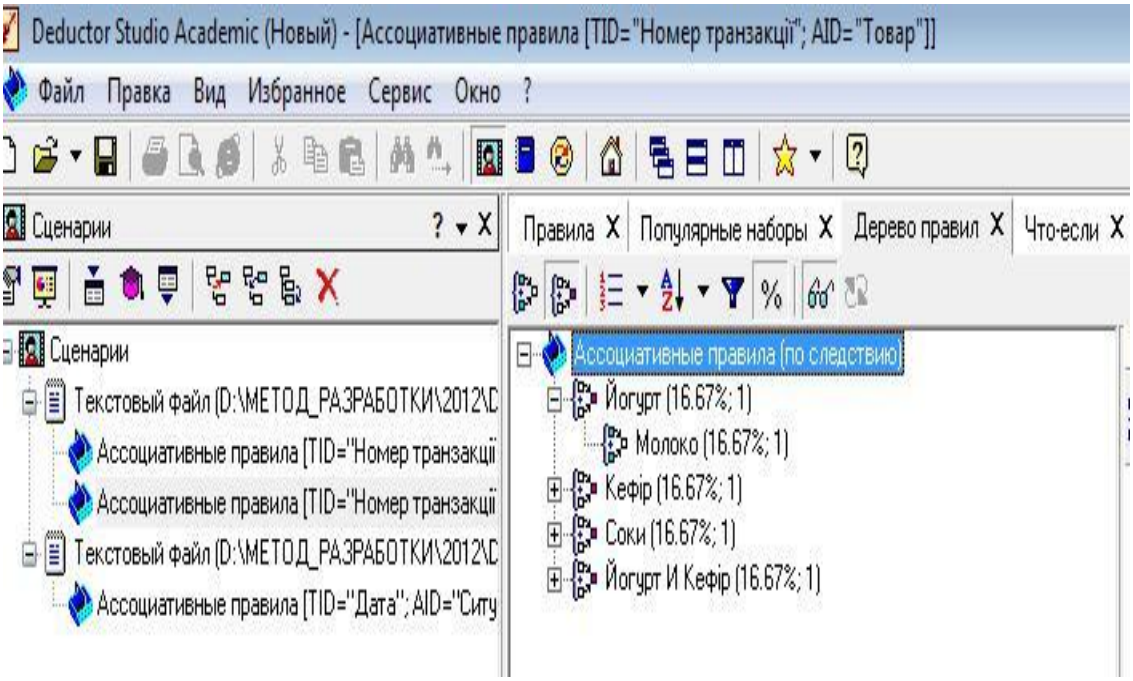


Рисунок 83 – Візуалізатор *Дерево правил*

Набір правил, поданий у візуалізаторі *Дерево правил* для завдання лабораторної роботи наведено на рис. 83. Наприклад, за першим правилом якщо покупець купив молоко, то з ймовірністю 33% він також купить йогурт. Ця інформація корисна, наприклад, при розв’язанні задачі розташування товарів на полиці.

Візуалізатор *Що-якщо* зручний, якщо необхідно дослідити наслідки певних умов (рис.84).

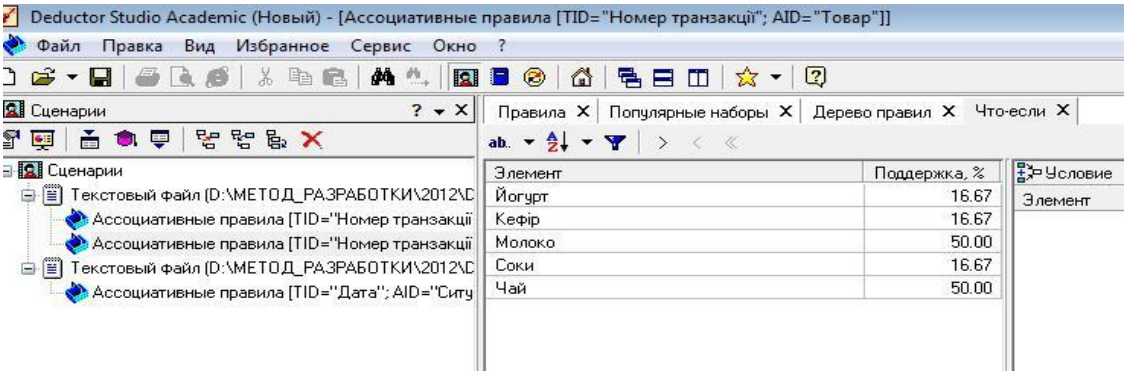


Рисунок 84 Візуализатор *Що-якщо*

10 Додати до бази даних 50 - 60 записів, налаштувати параметри пошуку правил, проаналізувати отримані результати за допомогою візуалізаторів.

Розглянутий приклад пошуку асоціативних правил є типовою ілюстрацією задачі аналізу продуктової корзини. У результаті її розв'язання визначаються набори товарів, які найчастіше зустрічаються, а також такі, що придбані покупцями спільно.

11 Проаналізувати записи бази даних, сформувати асоціативні правила та використати їх для надання рекомендацій щодо раціонального розташування товарів на полицях магазинів.

12 Проаналізувати записи бази даних, сформувати асоціативні правила та використати їх для надання знижок на пари товарів для підвищення обсягу продажів і, отже, прибутку.

13 Отримати асоціативні правила на визначеному часовому інтервалі на підґрунті аналізу ринку з заданими показниками.

14 Визначити закономірність поєднання брендів електронно-обчислювальної техніки, тобто закономірність поєднання компоновки елементів груп запропонованих товарів.

15 Сформулювати власну задачу та розвязати її з використанням методу *Асоціативних правил.*

Контрольні питання

1 Які сучасні методи виявлення потреб покупців Вам відомі?

2 Назвіть стандартні методи інтелектуального аналізу даних, які застосовуються для аналізу поведінки покупців?

3 Охарактеризуйте основні задачі, які вирішуються за допомогою інтелектуального аналізу даних для підприємств роздрібної торгівлі.

4 Дайте поняття підтримки, достовірності та ліфту? Які дані ці поняття характеризують?

5Яку структуру має асоціативне правило?

6Охарактеризувати основні характеристики асоціативних правил.

7Який найпоширішений алгоритм пошуку асоціативних правил, в чому його сутність?

Лабораторна робота 6
ОБРОБКА ДАНИХ З ВИКОРИСТАННЯМ НЕЙРОННОЇ МЕРЕЖІ

Мета роботи – вивчення процесу вибору структури, алгоритму та параметрів навчання нейронної мережі.

Завдання для підготовки до виконання лабораторної роботи.

Побудувати модель, яка зможе дати відповідь, чи входить клієнт, який бажає отримати кредит, в групу ризику неповернення кредиту. Як навчальний набор даних розглядається база даних з інформацією про клієнтів (вихідні дані лабораторної роботи 4, рис.65).

1 Загальні положення

Побудовану модель можна використовувати для прийняття рішень, пояснення причин, оцінки значущості факторів, моделювання різних варіантів розвитку.

Нейронні мережі це моделі, створені за подобою біологічних нейронних мереж. Вчені створюють такі моделі для отримання інтелектуальних властивостей в областях, де традиційні методи обчислень утруднені і малоефективні. Нейронні мережі в більшості своїй використовуються для класифікації, тобто для вибору деяких опцій в залежності від вхідних факторів. Нейронні мережі також можуть використовуватися для зберігання комплексних даних. При пред'явленні частини комплексних даних або зашумлених (зіпсованих) даних вони можуть відновити вихідний набір.

Мережі складаються з елементарних одиниць, клітин, з'єднаних одна з одною таким чином, щоб передавати сигнали. На практиці сигнал, переданий однією клітиною іншій – це число в діапазоні від 0 до 1. Відсутність сигналу кодується 0, його наявність – 1.

Зв'язки між нейронами характеризуються *вагами*. Навчання мережі починається з ініціалізації ваг зв'язків (також званих ваговими коефіцієнтами) випадковими величинами. Мережі пред'являються різні дані, а вагові коефіцієнти налаштовуються згідно обраної математичної схеми. Після навчання мережа може розпізнавати вхідні дані. Інформація про здобутий під час навчання досвід зберігається у вигляді вагових коефіцієнтах зв'язків.

Навчання проходить у кілька кроків і типова схема навчання виглядає так:

– мережі пред'являється перший зразок;

– вагові коефіцієнти модифікуються в незначній мірі, але так, щоб збільшити шанси розпізнавання даного зразка;

– пред'являється другий зразок і повторюється другий крок;

– повторюють попередні кроки для всіх зразків;

– повторюють всі попередні кроки сотні (або навіть тисячі!) разів.

Більшість нейронних мереж навчаються з вчителем. Це означає, що їм пред'являються зразки входів і відповідних виходів, тобто інформація не тільки про вхідні фактори, а й про те, що від мережі очікують отримати на виході.

Існує також навчання без вчителя, коли мережі надаються лише вхідні дані і мережа розбиває їх на групи. Стандартні нейронні мережі такого типу – це карти, що самоорганізуютьсся (*SelfOrganizing Maps SOM*) від Тево Кохонена (університет Хельсінкі). Ці мережі координують зв'язки так, що схожі вхідні образи завжди породжують схожі образи активності.

Зазвичай в мережі є три шари – вхідний, прихований і вихідний. Кожен зв'язок між вузлами характеризується силою зв'язку, що позначає ступінь впливу зв'язку на результат розрахунку, ваговий коефіцієнт. Інформація, що запам'ятовується мережею, зберігається саме в масиві цих коефіцієнтів.

При прямому поширенні сигнал від одного шару до іншого може поширюватися тільки в одному напрямку, тобто можлива передача сигналу з шару 1 в шар 2, з шару 2 в шар 3 і т.д., передача в інших напрямках є неможливою.

Протилежністю прямого поширення є рекурентна мережа, яка має з'єднання зі зворотним зв'язком. Нейронні мережі зворотного поширення – це інструмент пошуку закономірностей, прогнозування, якісного аналізу. Таку назву мережі зворотного поширення (back propagation) отримали через використання алгоритму навчання, в якому помилка поширюється від вихідного шару до вхідного, тобто в напрямку, протилежному напрямку поширення сигналу при нормальному функціонуванні мережі.

Процес навчання нейронної мережі засновано на використанні алгоритму BackPropagation, тобто зворотного поширення помилки, і алгоритму RPROP (Resilient Propagation). Сутність алгоритму зворотного поширення помилки полягає в наступному. У процесі функціонування вихідна помилка мережі, яка обчислюється на кожній ітерації, поширюється нейронною мережею від виходу до входу (тобто в напрямку, зворотному поширенню сигналу) та використовується для розрахунку коригування ваг нейронів кожного прихованого шару мережі:

$$\sum(\omega) = \frac{1}{2}\sum_{i=1}^{N}(Y_i - d_i)^2$$

де

Y_i – очікувані значення на виході мережі,

d_i – отримані значення на виході мережі після навчання.

Алгоритм RPROP використовує так зване «навчання за епохами», коли корекція ваг відбувається після пред'явлення мережі всіх прикладів з вибірки, що навчає:

$$\Delta w_{ij}(k) = -\eta_{ij}(k)\ \mathrm{sgn}(\partial E(\omega(k))/\partial\omega_{ij}),$$

де

η_{ij} – швидкість навчання,

w_{ij} – коефіцієнт зв'язку *i* нейрона шару *n-1* з *j* нейроном шару *n* [27].

Навчання за використання будь-якого алгоритму відбувається доти, поки цільова функція не досягне заданого значення. Основним недоліком нейромережевої парадигми є високі вимоги до обсягу вибірки, що навчає. Інший суттєвий недолік полягає в тому, що навіть натренована нейронна мережа є чорною скринькою, який "ковтає" початкові умови і видає прогноз [28-29].

2Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc)* файл, який містить інформацію про клієнтів (рис.65).

2 Налаштувати параметри в вікні *Майстра імпорту (6 з 9)*: поле ДАВАТИ КРЕДИТ– *Вихідне*, всі числові поля – *Вхідні*, всі інші поля – *Інформаційні* (як в лабораторній роботі *4*) (рис.85).

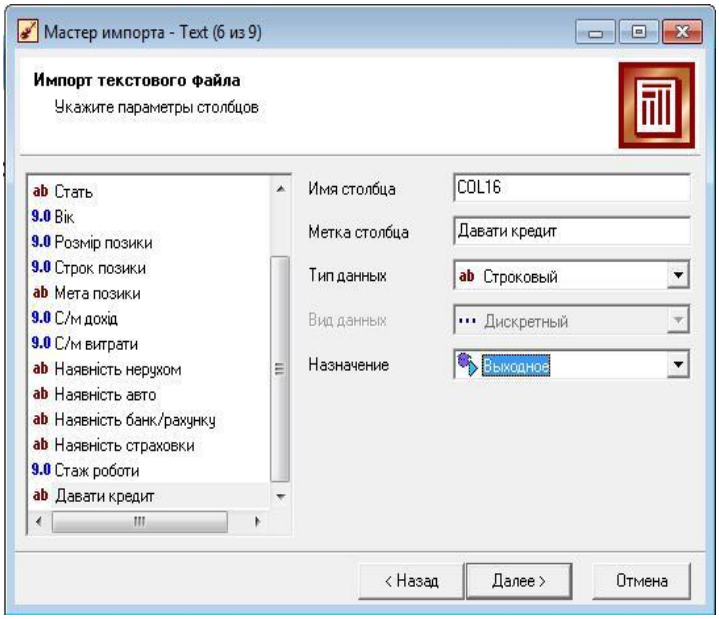


Рисунок 85 – Параметри полів

3 Запустити *Майстер обробки*, де як метод обробки даних обрати *Нейромережа (Багатошарова нейронна мережа)*.

4 Розбити початкову множину даних на тестову та множину, що навчає (рис.86). Як спосіб розбиття вихідної множини даних обрати ВИПАДКОВИЙ.

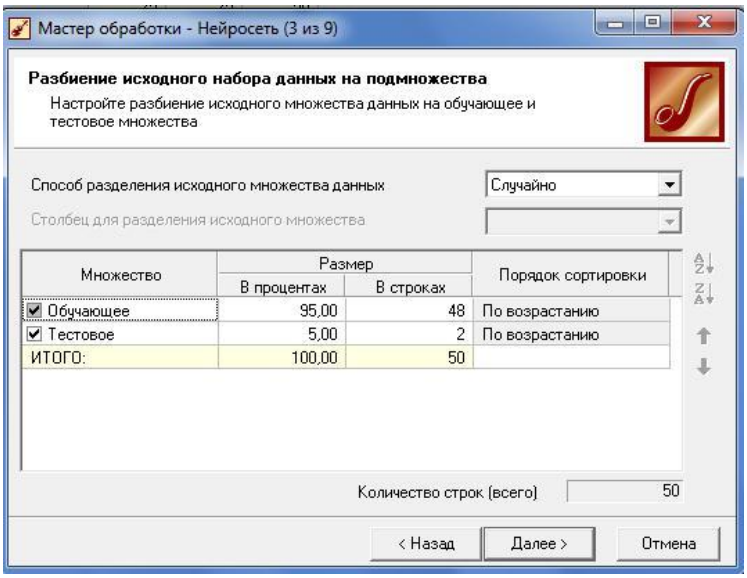


Рисунок 86 – Розбиття вихідного набору даних на підмножини

5 Задати параметри та структуру нейронної мережі (рис.87). Для цього необхідно тобто вказати кількість нейронів у вхідному шарі, що дорівнює кількості вхідних змінних (6 для даного файлу), у скритому шарі (1 для даного файлу) та у вихідному шарі, що дорівнює кількості вихідних змінних (1 для даного файлу), а також тип та параметри активаційної функції.

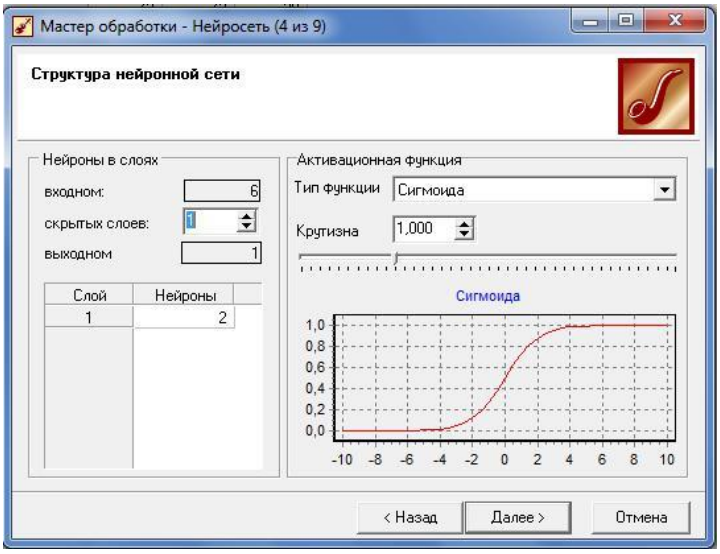


Рисунок 87 – Налаштування структури нейронної мережі

6 Задати алгоритм BackPropagation та параметри навчання нейронної мережі (рис.88).

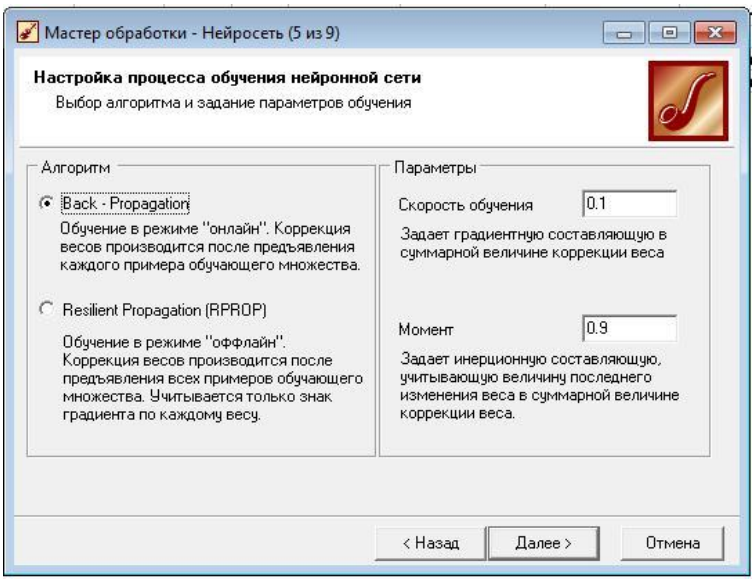


Рисунок 88 – Налаштування процесу навчання нейронної мережі

7 Налаштувати умови завершення навчання – якщо помилка є меншою 0,05, то випадок вважається розпізнаним; якщо кількість епох дорівнює 10000, то навчання можна завершити (рис.89).

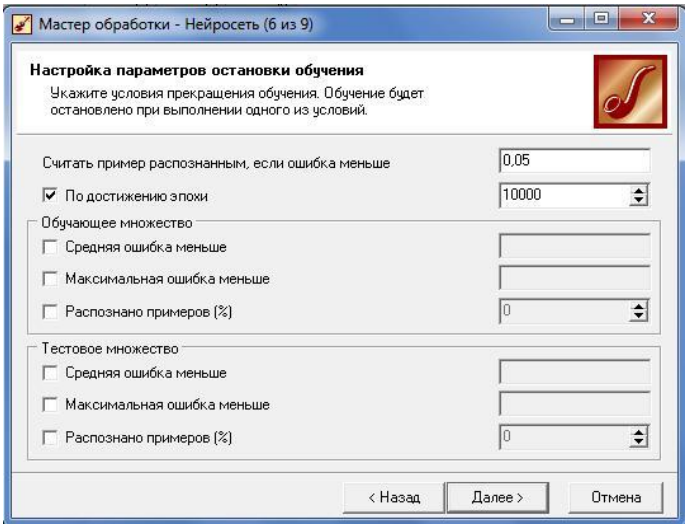


Рисунок 89– Налаштування умов завершення навчання

8 На наступному кроці запустити процес навчання та подивитись, як з часом буде змінюватись величина помилки та відсоток розпізнаних випадків в тестовій множині та множині, що навчає. На рис. 90 видно, що на епосі № 10000 в множині, що навчає, розпізнано 89,58% випадків, а на тестовій множині – 100% випадків.

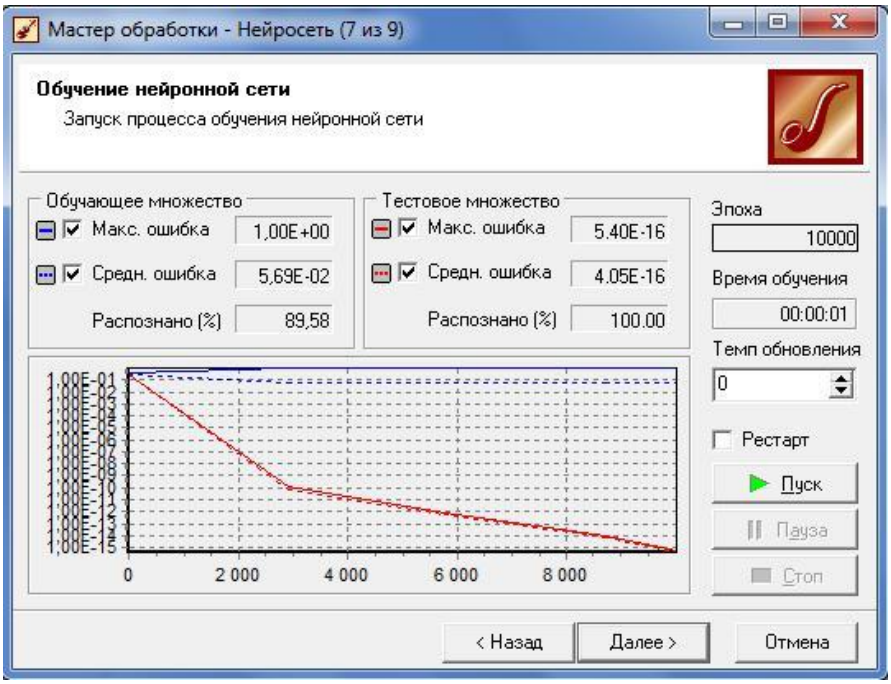


Рисунок 90 – Процес навчання нейронної мережі

9 Після завершення процесу навчання для інтерпретації отриманих результатів обрати візуалізатори *Таблиця спряженості*, *Граф нейромережі*, *Що-якщо,* та з їх допомогою проаналізувати отримані результати, тобто визначитися кому з клієнтів можна давати кредит, а кому – ні.

На рис.91 показана таблиця спряженості, на основній діагоналі якої розташовані випадки, які були розпізнані правильно, тобто 25 клієнтів, яким можна видавати кредит, і 22 клієнта, яким видавати кредит не варто. На додатковій діагоналі надано кількість клієнтів, які були віднесені до іншого класу (1 і 2). Таким чином, з 50 випадків 47 були класифіковані вірно, що складає 94% всіх випадків.

Рисунок 91 – Таблиця спряженості

Візуалізатор *Що-якщо* дозволяє провести віртуальний експеримент в певному діапазоні значень. Дані щодо претендента на кредит вводяться у відповідні поля, і для кожного набору значень модель приймає рішення щодо кредиту (*Так* або *Ні*), тобто розв’язує поставлену задачу (рис.92).

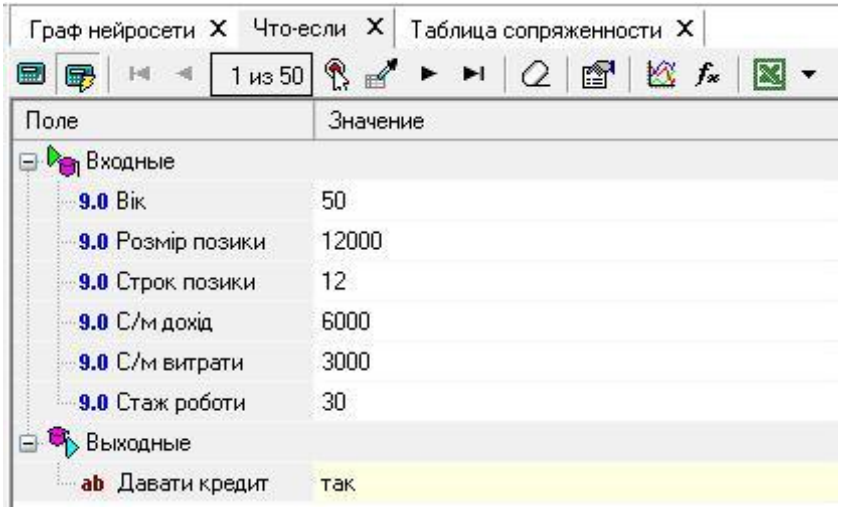


Рисунок 92 – Візуалізато *Що-якщо*

Граф нейромережі наведено на рис. 93.

10 Додати до бази даних 80 записів та побудувати модель, яка зможе дати відповідь, чи входить клієнт, який бажає отримати кредит, в групу ризику неповернення кредиту.

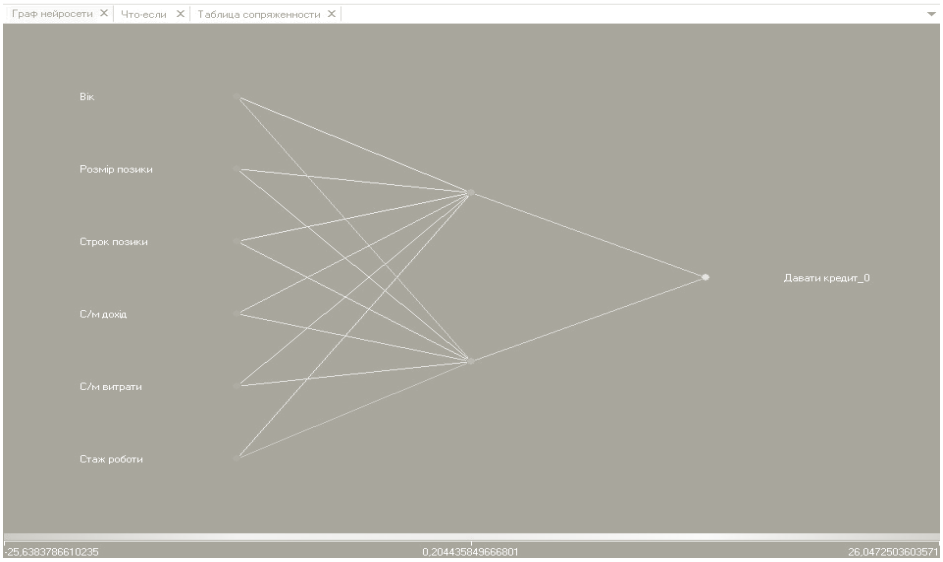


Рисунок 93 –Граф нейромережі

Контрольні питання

1 Описати структуру штучного нейрону.

2 Що визначає вага синапсу?

3 Як протікає процес навчання нейронної мережі?

4 Який метод використовується для розв’язання задачі мінімізації цільової функції помилки нейромережі, що забезпечує обмеження простору пошуку при навчанні?

Лабораторна робота 7
ПРОГНОЗУВАННЯ ЗА ДОПОМОГОЮ НЕЙРОННОЇ МЕРЕЖІ

Мета роботи – прогнозування часового ряду.

Завдання для підготовки до виконання лабораторної роботи

Побудувати модель прогнозу з використанням механізмів очищення даних від шумів та аномалій, що забезпечить високу якість моделі й достовірний результат прогнозування кількості продаж на три місяці вперед.

Розглянути принципи моделювання та прогнозування часового ряду: імпорт, очищення, згладжування, побудова моделі та прогнозу часового ряду.

В якості навчального набору використати базу даних з лабораторної роботи 3 (рис.56) .

1Загальні положення

У сучасному світі часто доводиться приймати рішення в умовах невизначеності, зокрема, будувати прогнози на майбутнє. Для вирішення задач прогнозування поряд з класичними статистичними методами, моделюванням, експертним пошуком рішення, успішно використовуються нейромережеві технології, що представляють перспективний напрямок розвитку штучного інтелекту [30]. Нейронні моделі використовуються при вирішенні широкого класу задач прогнозування, апроксимації, розпізнавання образів тощо.

Розв’язання задачі прогнозування часових фінансових рядів складної динамічної системи, що погано формалізується, не є простим завденням. Щоб коректно спрогнозувати ціну, слід спрогнозувати поведінку учасників ринку. Це можливо, якщо виявити приховані закономірності колективної психології шляхом аналізу даних за минулими торгами. Слід зазначити, що прийняті рішення суб'єктивні і не завжди можуть бути пояснені логічно [31].

Рішення задач виконується в два етапи: проектування мережі та її навчання для конкретних цілей [23].

При вирішенні задачі необхідно застосувати механізми очищення даних від шумів, аномалій, щоб забезпечити якість побудови моделі й достовірність результату прогнозування.

Подібний сценарій є основою будь-якого прогнозування за моделлю часового ряду з тією різницею, що для кожного випадку доводиться обробляти часовий ряд за допомогою інструментів *Deductor Studio* (наприклад, угруповання), обирати параметри очищення даних та параметри моделі прогнозу.

2Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc)* файл як в лабораторній роботі №3(рис.56) .

2 Налаштувати в вікні *Майстра імпорту (6 з 9)* (рис. 31) параметри кожного поля: ім’я стовпця, мітка стовпця, тип даних, вид даних, призначення. Встановити всім полям призначення – *Інформаційне.*

3 Запустити процес імпорт, натиснувши кнопку Пуск в вікні *Майстра імпорту* (*7 з 9*) та кнопку Далі.

4 На наступному кроці обрати спосіб відображення даних – *Таблиця.*

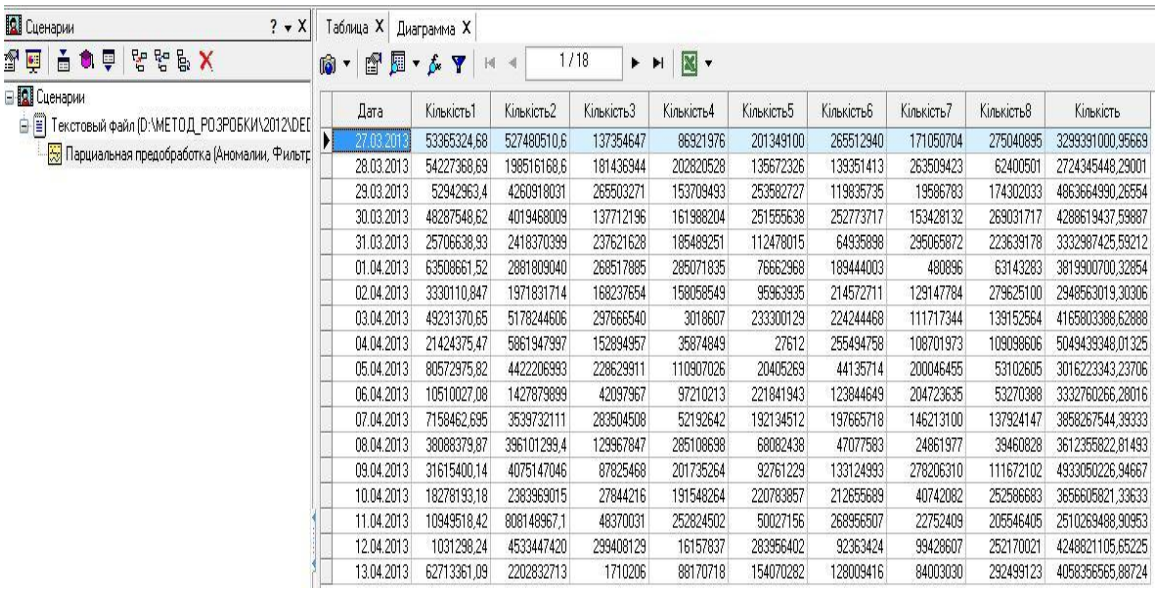
5Ініціювати *Майстер обробки* та в групі *Очищення даних* для виключення аномалій (викидів) та згладження даних обрати *Парціальна обробка.*

6 Пропустити другий крок *Майстра*, на якому виконується обробка відсутніх значень.

7 Виконати редагування аномальних значень на третьому кроці *Майстра*. Для цього обрати поле КІЛЬКІСТЬ та обробку аномальних значень зі малим ступенем придушення.

8 Щоб вилучити шуми з початкових даних на 4 кроці *Майстра* провести спектральну обробку, для чого обрати стовпець КІЛЬКІСТЬ та вказати спосіб обробки – *Віднімання шуму* (ступінь віднімання – мала).

9 Обрати спосіб відображення – *Таблиця* та *Діаграма* (рис.94). Діаграма будується за полями КІЛЬКІСТЬ та ДАТА (рис.95).



Дата	Кількість1	Кількість2	Кількість3	Кількість4	Кількість5	Кількість6	Кількість7	Кількість8	Кількість
27.03.2013	53365324,68	527480510,6	137354647	86921976	201349100	265512940	171050704	275040895	3299391000,95669
28.03.2013	54227368,69	198516168,6	181436944	202820528	135672326	139351413	263509423	62400501	2724345448,29001
29.03.2013	52942963,4	4260918031	265503271	153709493	253582727	119835735	19586783	174302033	4863664990,26554
30.03.2013	48287548,62	4019468009	137712196	161988204	251555638	252773717	153428132	269031717	4288619437,59887
31.03.2013	25706638,93	2418370399	237621628	185489251	112478015	64935898	295065872	223639178	3332987425,59212
01.04.2013	63508661,52	2881809040	268517885	285071835	76662968	189444003	480896	63143283	3819900700,32854
02.04.2013	3330110,847	1971831714	168237654	158058549	95963935	214572711	129147784	279625100	2948563019,30306
03.04.2013	49231370,65	5178244606	297666540	3018607	233300129	224244468	111717344	139152564	4165803388,62888
04.04.2013	21424375,47	5861947997	152894957	35874849	27612	255494758	108701973	109098606	5049439348,01325
05.04.2013	80572975,82	4422206993	228629911	110907026	20405269	44135714	200046455	53102605	3016223343,23706
06.04.2013	10510027,08	1427879899	42097967	97210213	221841943	123844649	204723635	53270388	3332760266,28016
07.04.2013	7158462,695	3539732111	283504508	52192642	192134512	197665718	146213100	137924147	3858267544,39333
08.04.2013	38088379,87	3961012994	129967847	285108698	68082438	47077583	24861977	39460828	3612355822,81493
09.04.2013	31615400,14	4075147046	87825468	201735264	92761229	133124993	278206310	111672102	4933050226,94667
10.04.2013	18278193,18	2383969015	27844216	191548264	220783857	212655689	40742082	252586683	3656605821,33633
11.04.2013	10949518,42	808148967,1	48370031	252824502	50027156	268956507	22752409	205546405	2510269488,90953
12.04.2013	1031298,24	4533447420	299408129	16157837	283956402	92363424	99428607	252170021	4248821105,65225
13.04.2013	62713361,09	2202832713	1710206	88170718	154070282	128009416	84003030	292499123	4058356565,88724

Рисунок 94– *Таблиця (парціальна обробка)*

10 Побудувати модель, базуючись на даних попередніх періодів, тобто припускаючи, що обсяги продажів в майбутньому залежать від обсягів продажів за попередні місяці.

11 Ініціювати *Майстер обробки* та в групі *Data Mining* обрати *Нейромережа*.

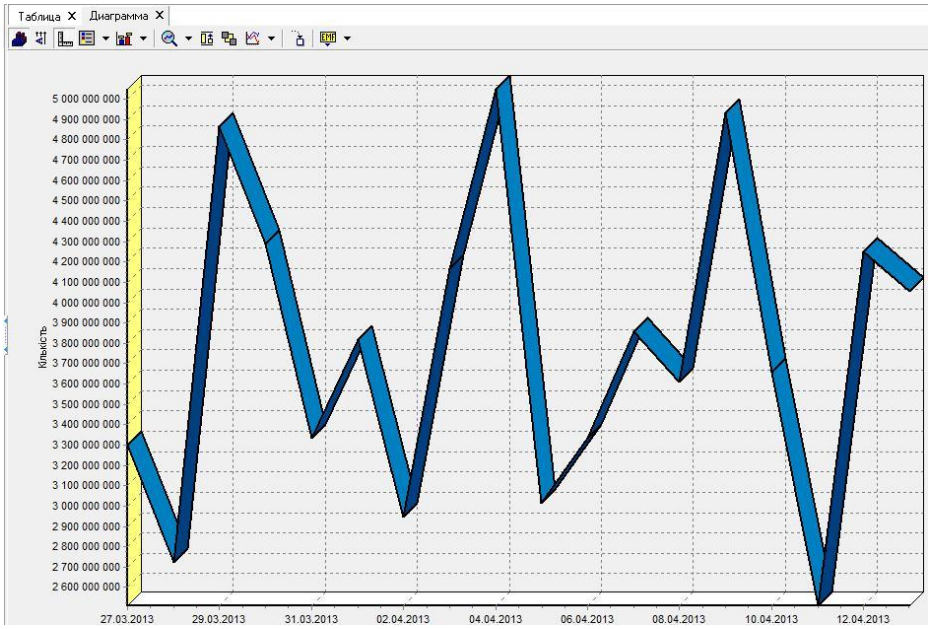


Рисунок 95 – Побудована діаграма*(парціальна обробка)*

12 На другому кроці *Майстра* встановити як *вхідні* поля –КІЛЬКІСТЬ1, КІЛЬКІСТЬ2, КІЛЬКІСТЬ8, як *вихідне* – КІЛЬКІСТЬ. Всі інші поля мають мати призначення – *Інформаційне*.

13 Встановити розбиття даних на множину, що навчає, та тестову.

14 На наступному кроці задати кількість шарів та нейронів в нейромережі, тобто визначити її структуру (рис.96).

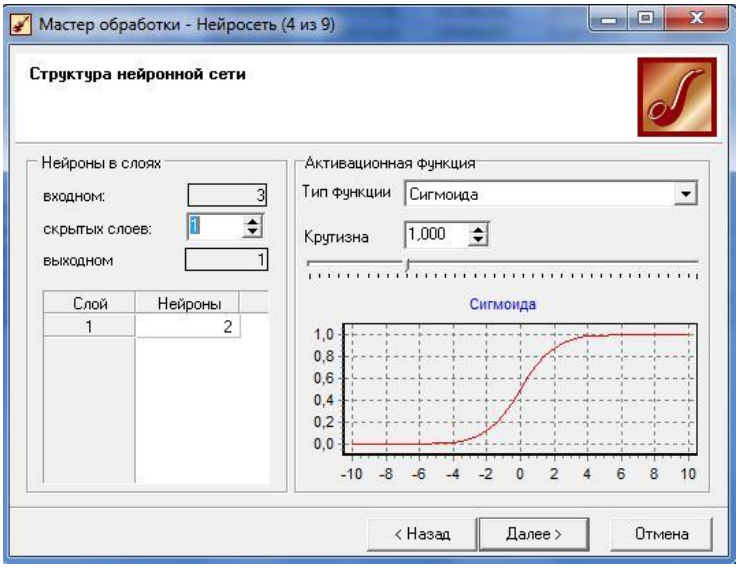


Рисунок 96– Структура нейромережі

Наближення значень вихідних сигналів до очікуваних можна досягти, змінюючи ступінь кривизни сигмоїди.

15 Обрати алгоритм навчання нейромережі – алгоритм RPROP (рис.97).

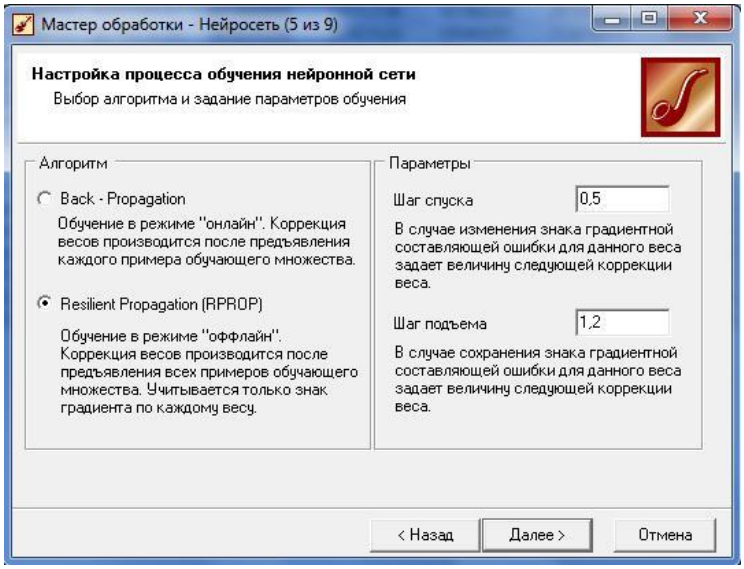


Рисунок 97 – Вибір алгоритму та визначення параметрів навчання

16 Обрати способи візуалізації – *Діаграма*, *Діаграма розсіювання*, *Граф нейромережі, Відомості.* Необхідна інформація знаходиться в гілках *Об'єкт/ Вчитель/ Помилки – Те, що навчає* і *Об'єкт/ Вчитель/ Помилки – Тестове*.

Діаграму побудувати за полями КІЛЬКІСТЬ та КІЛЬКІСТЬ_OUT (рис.98).

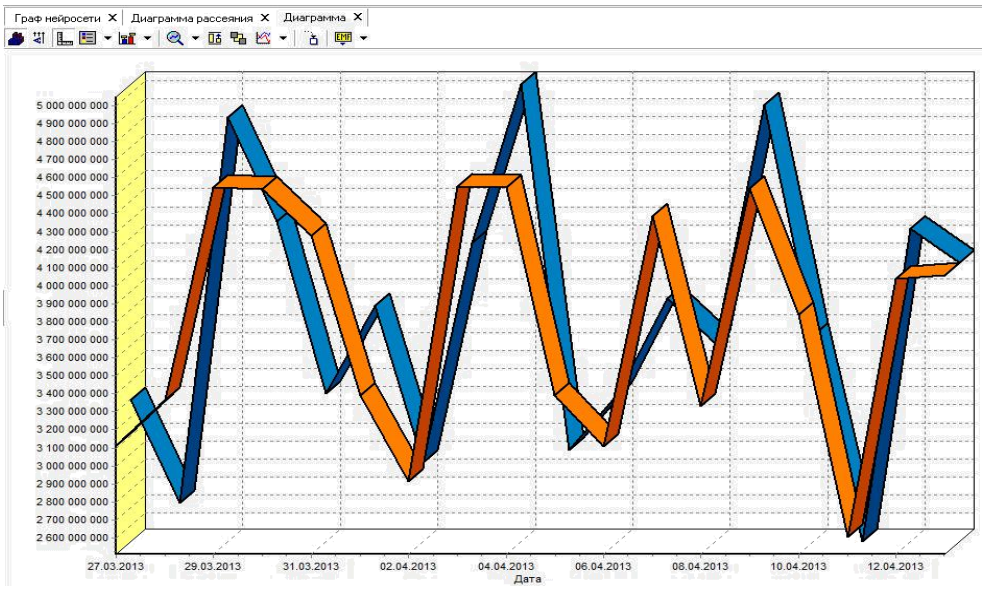


Рисунок 98 – Діаграму за полями КІЛЬКІСТЬ та КІЛЬКІСТЬ_OUT

Діаграма розсіювання показує якість навчання (рис.99).

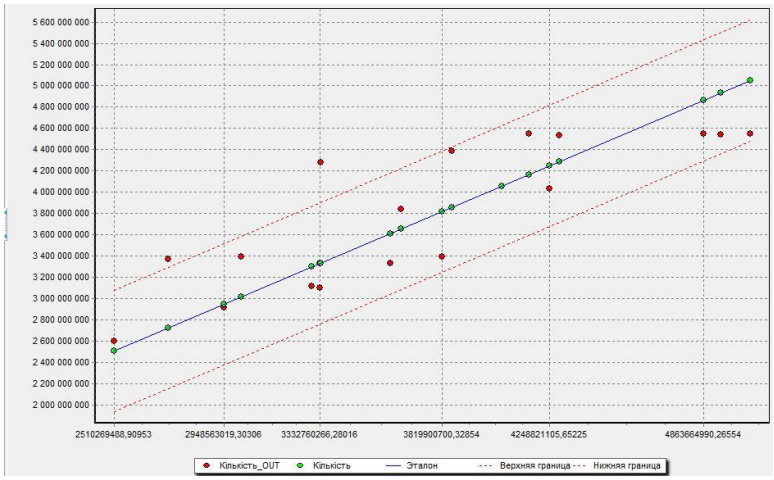


Рисунок 99 –Діаграма розсіювання

Отже, нейромережа навчена та можна перейти до етапу прогнозування. Відкрити *Майстер обробки*, де в групі *Data Mining* з'явився обробник *Прогнозування*.

17 Налаштувати зв'язки стовпців для прогнозування часового ряду (де брати дані для стовпця при наступному кроці прогнозу). Вказати горизонт прогнозу (на скільки вперед виконується прогноз), що дорівнює трьом, а також додати початкові дані, встановивши в *Майстрі* відповідний прапорець (рис.100).

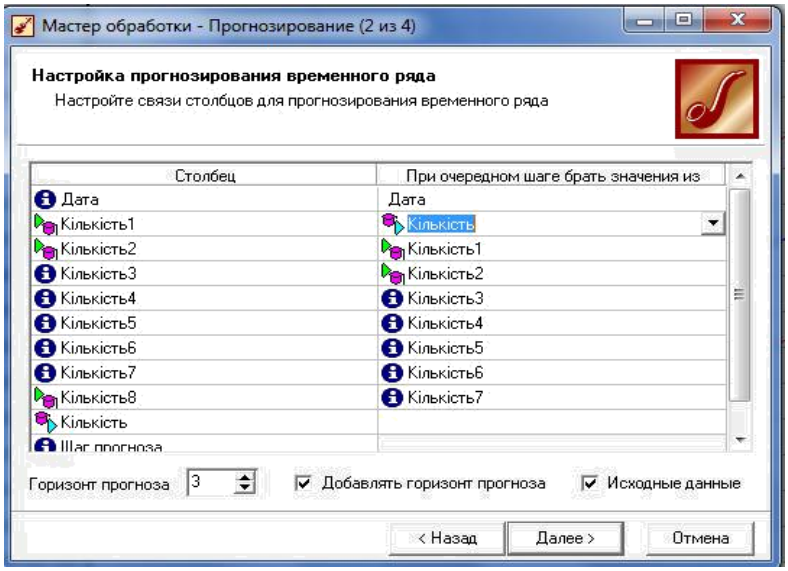


Рисунок 100– Налаштування параметрів прогнозування часового ряду

18 Після здійснення прогнозування часового ряду обрати як візуалізатор *Діаграму прогнозу*. В вікні НАЛАШТУВАННЯ СТОВПЦІВ ДІАГРАМИ ПРОГНОЗУ вказати як стовпець, що відображається, КІЛЬКІСТЬ, а як підпис по вісі *X* – стовпець КРОК ПРОГНОЗУ (рис.101).

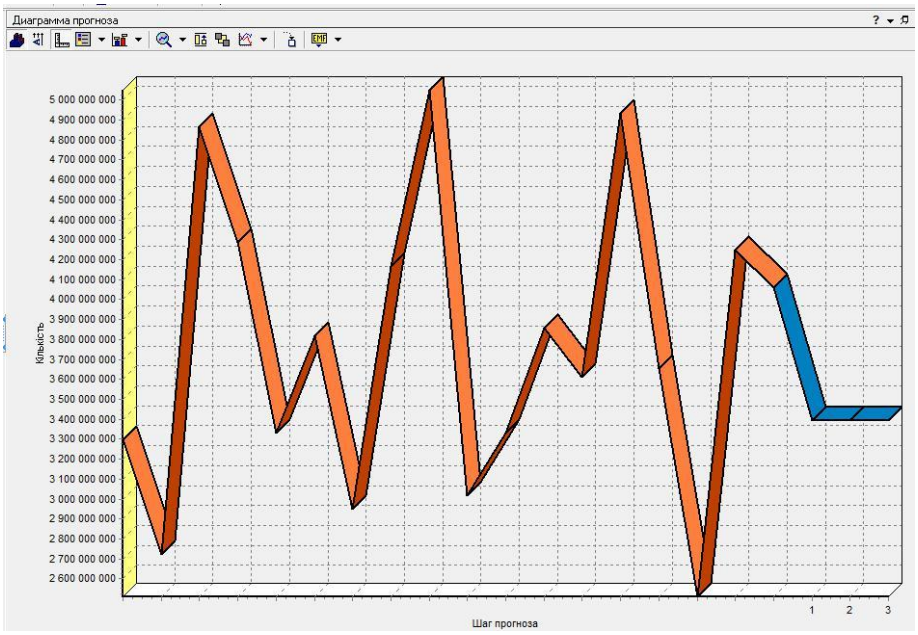


Рисунок 101– Діаграма прогнозу

Базуючись на діаграмі прогнозу, аналітик зможе відповісти на питання щодо кількості товарів, що буде продано протягом наступних 3 місяців.

19 Додати до вибірки 30-50 записів та побудувати модель прогнозу.

Контрольні питання

1 Які методи використовуються для здійснення прогнозу?

2 Для розв’язання яких задач використовуються нейронні моделі?

3 Які механізми передобробки необхідно виконати для забезпечення високої якості моделі прогнозу?

4 Описати алгоритм прогнозування часового ряду.

Лабораторна робота 8
РОБОТА З КАРТАМИ КОХОНЕНА

Мета роботи – навчитися використовувати карти, що самоорганізуються, для аналізу даних з застосуванням можливостей Deductor.

Завдання на виконання лабораторної роботи.

1. Нехай є база даних комерційних банків з показниками діяльності, такими як: депозити юридичних осіб, активи банку, прибутковість активів, прибуток. Доповнити базу даних 30 записами та визначити показники

діяльності комерційних банків. Для цього необхідно провести кластеризацію, тобто виділити однорідні групи банків на основі показників з бази даних (рис.102).

	A	B	C	D	E	F
1	№№	Банк	Депозити юридичних осіб	Активи банку	Прибутковість активів	Прибуток
2	1	Аваль	2900000	500000	160	123000
3	2	Приват	3500000	4000000	120	100000
4	3	Правекс	5000000	5500000	200	250000
5	4	Надра	1100000	900000	10	0
6	5	Родовід	30000	15000	5	0
7	6	Укрсоцбанк	1200000	3000000	190	300000
8	7	Укрсиббанк	30000000	3000000	190	500000
9	8	Хрещатик	400000	3000000	50	20000
10	9	Донбас	12300	1000	0	0
11	10	Слобожанка	700000	800000	50	0
12	11	Захід	90000	90000	0	0
13	12	Схід	60900000	100000000	150	790000
14	13	Північ	109999	200000	15	0
15	14	Південь	3000000	500000	15	0
16	15	Карпати	430000	500000	70	20000

Рисунок 102 – База даних комерційних банків з показниками їх діяльності

2. Нехай є база даних комерційних банків з показниками діяльності, такими як: назва банку, реєстраційний номер, назва міста, номери філіалів банку, кількість робітників, сума активів, власні активи, банківські активи, депозити фізичних осіб, депозити юридичних осіб, позика в млн. гривень, позика в млн. доларів, кошти в банку млрд., міжфілійні кошти, млрд, прибуток, млн. Визначити показники діяльності комерційних банків. Для цього необхідно провести кластеризацію банків (рис.103).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	№	банк	реєстраційний номер	філіали	місто	кіл-сть робітників	сума активів,млн	власні активи	банківські активи	депоз. фіз. осіб	депоз. юр. осіб	позика в млн грн.	позика в млн дол.	кошти в банку млрд	міжфіл. кошти, млн	прибуток млн
2	1	Альянс	300119	20	Київ	100	83,89	53,9	30,5	2,28	8,53	50	30	100	90	7
3	2	Базис	351760	30	Харьков	150	996,7	450	500	399,08	233,37	100	50	150	100	12,19
4	3	ВТБ	321767	35	Київ	150	33678,36	10650,1	1200	6282,7	5282,38	1000	500	200	120	173,19
5	4	Велес	3222799	50	Київ	300	150,82	70	50	10,34	16,82	200	135	238	50	104
6	5	Гратн	351607	15	Харьков	123	668,48	150	149,9	284,17	172,24	234	346	500	129	5
7	6	Инвестбанк	328210	10	Одесса	50	461,8	262	150	198,01	42,89	78	90	74	86	51
8	7	Кредобанк	325912	38	Львов	359	4556,29	2890	678	1932,48	1022,7	90	100	234	23	644
9	8	Мегабанк	351629	80	Харьков	456	4755,81	3456,98	1900	1573,56	1173,7	60	45	500	123	17
10	9	Меркурий	351663	150	Харьков	680	1713,05	590	700	812,67	330,92	85	37	187	34	7
11	10	Надра	320003	67	Київ	890	28695,63	10987,87	456,9	3285,06	7798,25	987	456	678	87	11
12	11	НРБ	320627	90	Київ	1000	3686,32	1500,9	675,98	459,87	826,27	100	65	876	98	24,49
13	12	ОТП	300528	87	Київ	987	20688,46	9000	987	5535,68	3942,35	120	98	543	23	78,52
14	13	Ощадбанк	300465	25	Київ	500	81587,4	25863,4	23698,5	29828,9	8482,15	400	86	165	147	52,83
15	14	Приват	305299	85	Днепропет	1000	174995,98	36985,3	24158,2	85679,34	21244,98	265	258	658	421	403,63
16	15	Премьер	339555	68	Київ	250	852,83	265	348,2	10,61	17,17	259	248	128	63	10,61
17	16	Ренесанс	380010	85	Київ	850	576,66	25	368	50,09	18,41	23	154	895	856	39,62
18	17	Свербанк	339500	75	Київ	456	2550,34	569	15623,5	995,24	875,04	123	125	456	789	7,01
19	18	Сбербанк	320627	79	Київ	900	27425,05	6589,23	5426,58	9537,28	4686,32	65	45	789	652	20,87
20	19	Стандарт	339400	78	Київ	850	2550,44	2654,2	1542,6	996,34	885,04	45	12	62	23	8,01
21	20	Трансбанк	300089	45	Київ	560	301,44	254	123	451,4	233,1	26	35	125	145	10,78
22	21	Укргаз	320478	16	Київ	456	18656,71	2456,32	1589,36	5502,3	2199,65	48	25	620	145	331,1
23	22	Укринбанк	300142	56	Київ	862	5205,92	3256,23	25369,12	2149,11	1411,33	256	126	236	862	28
24	23	Укрсиббанк	351005	84	Київ	654	44022,86	25368,23	12563,25	12276,83	7804,8	145	100	320	120	30
25	24	Укрсоцбанк	300023	56	Київ	256	40568,03	2598,6	1658,2	11753,84	8075,35	184	85	111	10	182
26	25	Универсал	337500	25	Київ	1578	7012,23	1254,2	1542,3	2580,33	616,91	623	123	852	625	272,6
27	26	Финексбанк	380311	100	Одесса	236	271,08	236,54	568,4	47,69	91,75	158	126	1000	20	11
28	27	Финростбанк	328599	42	Одесса	365	1359,67	100,3	105,6	554,05	351,5	236	145	147	158	14
29	28	Хрещатик	300670	154	Київ	956	7762,94	2569,36	2547,3	3003,98	1916,25	652	236	123	156	294
30	29	Експрес	322959	568	Київ	856	2677,25	1236,26	1254,2	1152,61	588,51	236	125	852	42	301
31	30	Експобанк	322294	500	Київ	1500	2522,11	1256,3	1986,3	776,17	644,49	265	123	650	85	210

Рисунок 103 – База даних комерційних банків

Відобразити дані в вигляді *Карт Кохонена, Профілів кластеров*, візуалізатора *Що-якщо, Матриця розташування.* В якості спеціальних відображень карти Кохонена обрати *Матрицю відстані, Матрицю щільності попадання, Кластери, Проекцію Саммона.* В *Матриці розташування* за вісю *X* відкласти поле СУМА АКТИВІВ, за вісю *Y* – поле БАНКИ, кольором об’єкта обрати поле ПРИБУТОК.

1 Загальні положення

Форма навчання на основі самоорганизуючого конкурентного навчання, яке використовується для побудови обчислювальних відображень, отримала назву карт самоорганізації. Карта Кохонена, що самоорганізується, (англ. *Selforganizing map — SOM*) – це нейронна мережа з навчанням без вчителя, яка виконує завдання візуалізації та кластеризації, і є методом проекціювання багатовимірного простору в простір з більш низькою розмірністю (найчастіше двовимірний). Вирішення задачі кластеризації полягає в розподіленні даних за декількома кластерами. Алгоритм визначає розташування кластерів в багатовимірному просторі факторів. Початкові дані будуть віднесені до певного кластеру в залежності від відстані до нього. Графічна форма подання багатомірного простору є складною для сприйняття, в той час як механізм побудови карти Кохонена дозволяє відобразити багатомірний простір в двовимірному, що більш зручно для візуалізації і інтерпретації результатів аналітиком. *SOM* застосовуються також для вирішення завдань моделювання та прогнозування.

Принципи роботи і навчання такої нейромережі були сформульовані фінським вченим Тойво Кохоненом в 1982 році. Основною ідеєю Т. Кохонена є введення в правило навчання нейрона інформації щодо його розташування. За Кохоненом нейромережа має один вхідний шар з числом нейронів, що дорівнює числу входів і єдиний прихований (вихідний) шар нейронів, який утворює одновимірну (лінія) або двомірну (прямокутник) решітку. За аналогією з топографічними картами таку нейромережу також називають картою Кохонена.

За цією парадигмою навчання проводиться без вчителя, тобто в процесі навчання немає порівнювання виходів нейронів з еталонними значеннями. В процесі навчання на вхід такої нейромережі послідовно подаються навчальні набори. Після подачі чергового набору визначається найбільш схожий нейрон, тобто нейрон, у якого різниця між скалярним добутком ваг і поданим на вхід вектором мінімальна. Такий нейрон оголошується переможцем і становиться центром при підлаштуванні ваг у сусідніх нейронів. Правило навчання, запропоноване Кохоненом, припускає змагальне навчання з урахуванням відстані нейронів від "нейрона -переможця".

Алгоритм формування карт ознак вхідних векторів складається з низки кроків.

1) Ініалізація мережі. Ваговим коефіцієнтам мережі привласнюються малі випадкові значення. Загальне число синаптичних ваг дорівнює *NM, N* – кількість вхідних параметрів, *M* – загальна кількість нейронів в мережі.

2) Надходження в мережу нового вхідного сигналу та обчислення відстані Кохонена між вхідними векторами та нейронами.

Відстань d_j від вхідного вектору до *j-го* нейрона визначається за формулою

$$d_j = \sum_{i=o}^{N-1} \left(x_i(t) - \omega_{ij}(t) \right)^2 ,$$

де

$x_i\ (t)$ – значення *i-го* елементу вхідного вектору в момент часу *t*,

$w_{ij}\ (t)$ – вага зв'язку *i-го* елементу з *j –м* нейроном у момент часу *t*; *N* – кількість вхідних параметрів.

3) Вибір *j** нейрона з найменшою відстанню d_j.

4) Налаштування ваг нейрона *j** та його сусідів. Нові значення ваг обчислюються за формулою:

$$w_{ij}\ (t-1) = w_{ij}\ (t) + r(t)\ (x_i\ (t) - w_{ij}\ (t))$$

де *r(t)* – крок навчання, позитивне число менше за одиницю, що зменшується з часом.

5) Повернення до кроку 2.

Для нейрона переможця функція сусідства дорівнює 1. З віддаленням від вона зменшується за експоненціальним законом. Таким чином, в процесі навчання підлаштування ваг відбувається не тільки в нейроні-нейроніпереможці, але і в його оточенні.

Після завершення процесу навчання карта Кохонена групує вхідні набори за подібністю. Вся сукупність нейронів у вихідному шарі точно моделює структуру розподілу навчальних наборів в багатовимірному просторі. Унікальність технології карт, що самоорганізуються, полягає в перетворенні *N-вимірного* простору в двох або одномірний. Але треба пам'ятати про специфічні риси такого перетворення, що можуть призвести до помилок інтерпретації. З того, що дві точки є сусідами на карті Кохонена витікає, що вони сусіди в *N-вимірному* вхідному просторі, але зворотне судження хибне.

У двомірному вигляді карта розфарбовується відповідно до рівня виходу нейрона. Карта виходів є головною картою в аналізі карт Кохонена. Згідно з неї проектується взаємне розташування досліджуваних даних. Подібні вхідні дані утворюють на карті кластери, замкнуті області, що складаються з нейронів з однаковими значеннями виходів. Як правило, добре виражені кластери в даних мають чіткі межі з іншими областями карти.

Після завершення навчання кожен вхідний набір потрапляє в "свій" нейрон. При цьому деякі нейрони не матиме жодного набору, а деякі – кілька наборів.

При аналізі карт Кохонена проводиться оцінка як виходів, так і ваг нейронів. Для кожного входу нейрона малюється своя карта, яка розфарбовується у відповідності зі значенням ваги нейрона. У нейронної мережі, навченою з учителем, ваги нейронів не мають фізичного смислового навантаження і не використовуються в аналізі. При навчанні без вчителя ваги нейронів підлаштовуються під точні значення вхідних змінних і відображають їх внутрішню структуру. Для ідеально навченої нейронної мережі вага нейрона дорівнює відповідній компоненті вхідного набору. Зазвичай одночасно аналізують кілька карт входів. Спочатку на одній карті виділяють області однакового кольору. У цій області групуються вхідні набори, які мають однакове значення відповідного входу. Далі нейрони з цієї області вивчаються на інших картах на предмет колірного розподілу.

При роботі з картами Кохонена важливо розуміти, що всі розглянуті вище карти – це різні розмальовки одних і тих же нейронів. При цьому кожен навчальний набор має одне і те ж розташування на кожній з розглянутих карт.

Унікальна властивість карт Кохонена – відображення міри схожості вхідних наборів. Чим ближче на карті точки один до одного, тим більше схожі між собою набори, які цим точкам відповідають [32-33].

2 Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc)* базу даних показниками діяльності комерційних банків (рис.102).

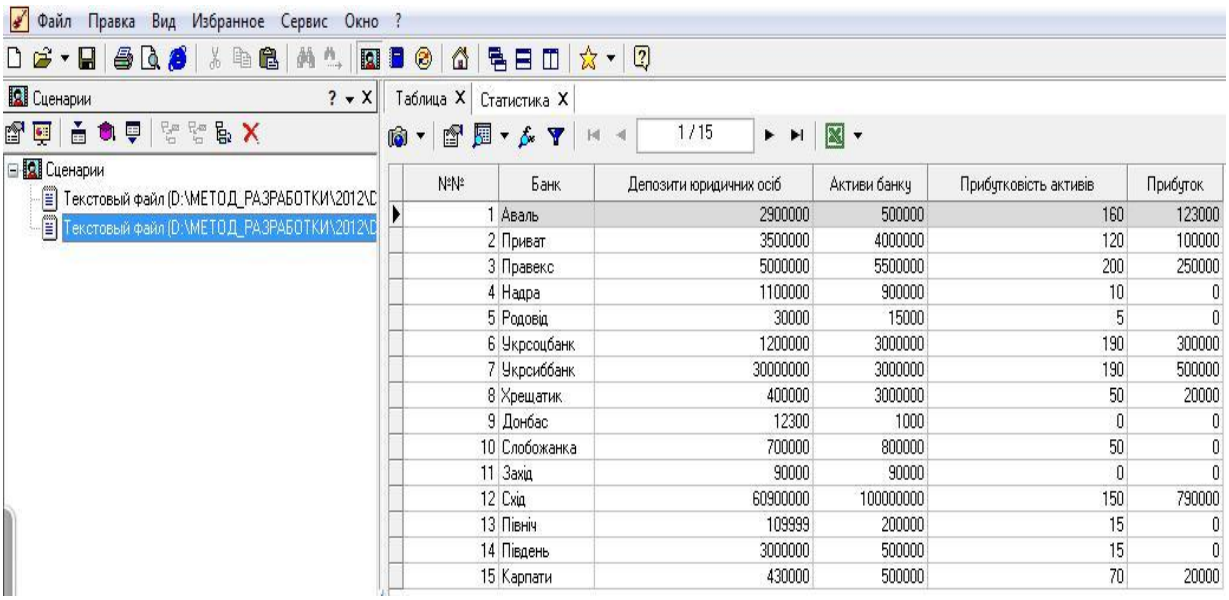
2 Імпортувати дані з файла табличного процесора (*MS Excel, Calc)* в аналітичну платформу *Deductor.*

3 Встановити в вікні ІМПОРТ ТЕКСТОВОГО ФАЙЛУ *Майстра імпорту* (6 з 9) такі параметри стовпців (табл.10):

Таблиця 10 – Параметри стовпців

№	Ім’я стовпця	Параметри стовпця		
		Тип даних	Вид даних	Призначення
1	№№	Інформаційне		
2	Банк	Строковий	Дискретний	Вхідне
3	Депозити юридичних осіб	Дійсний	Безперервний	Вхідне
4	Активи банку	Дійсний	Безперервний	Вхідне
5	Прибутковість активів	Дійсний	Безперервний	Вхідне
6	Прибуток	Дійсний	Безперервний	Вихідне

4 Визначити спосіб відображення – *Таблиця* та *Статистика*. Результат імпорту бази даних з файлу *MS Excel* в середовище *Deductor* наведено на рис. 104.



№№	Банк	Депозити юридичних осіб	Активи банку	Прибутковість активів	Прибуток
1	Аваль	2900000	500000	160	123000
2	Приват	3500000	4000000	120	100000
3	Правекс	5000000	5500000	200	250000
4	Надра	1100000	900000	10	0
5	Родовід	30000	15000	5	0
6	Укрсоцбанк	1200000	3000000	190	300000
7	Укрсиббанк	30000000	3000000	190	500000
8	Хрещатик	400000	3000000	50	20000
9	Донбас	12300	1000	0	0
10	Слобожанка	700000	800000	50	0
11	Захід	90000	90000	0	0
12	Схід	60900000	100000000	150	790000
13	Північ	109999	200000	15	0
14	Південь	3000000	500000	15	0
15	Карпати	430000	500000	70	20000

Рисунок 104 – Фрагмент завантаженої бази даних

5 Запустити *Майстер обробки* та обрати в групі *Data Mining* метод *Карта Кохонена.*

6 На другому кроці *Майстра обробки* перевірити призначення вихідних стовпців даних.

7 Розбити вихідну множину на множину, що навчає, та тестову (валідаційну). За замовчуванням 95% складає множина, що навчає, і 5% тестова.

8 На 4 кроці налаштувати параметри *Майстра*: кількість клітин вісі *X,* вісі *Y* та їх форму (шестикутна чи чотирикутна) (рис.105).

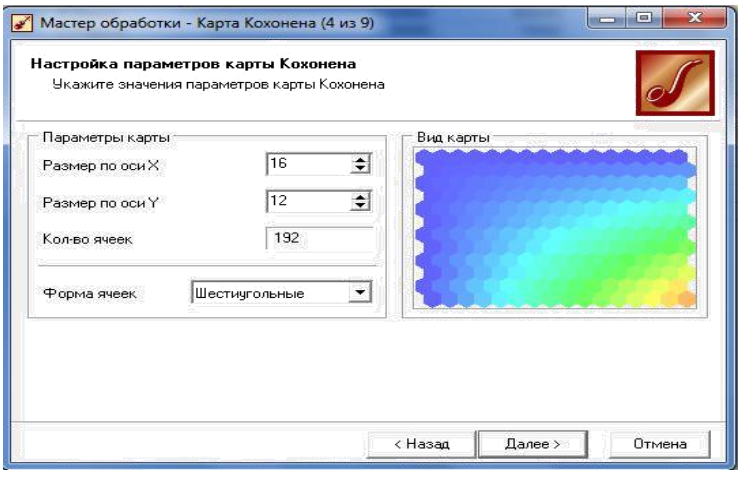


Рисунок 105 – Налаштування параметрів карти Кохонена

9 На п’ятому кроці *Майстра* встановити параметри завершення навчання, які наведено на рис. 106.

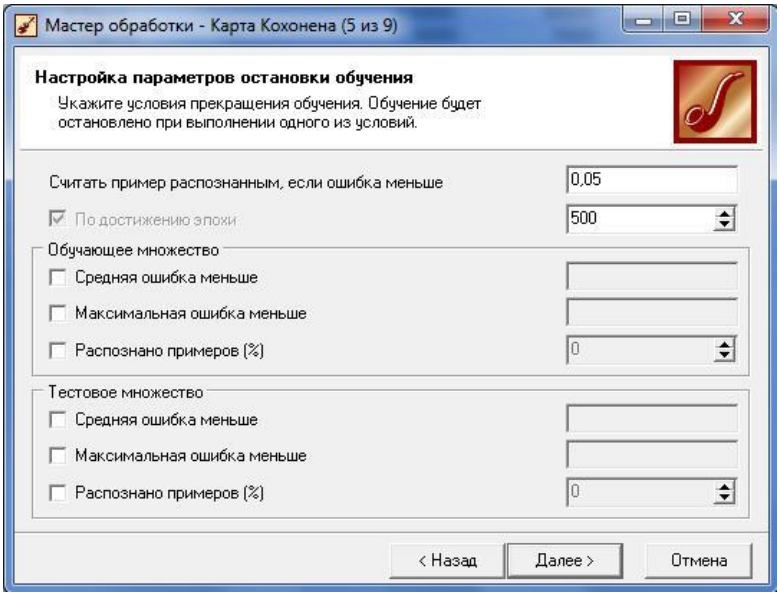


Рисунок 106 – Налаштування параметрів завершення навчання

Встановлених за замовчуванням 500 епох не завжди достатньо для якісного навчання карт Кохонена.

10 На шостому кроці налаштувати параметри навчання: спосіб початкової ініціалізації, тип функції сусідства. Можливі два варіанти кластеризації: автоматичне визначення числа кластерів з відповідним рівнем значущості і фіксована кількість кластерів (визначається користувачем, зазвичай 78 кластерів). Обрати фіксовану кількість кластерів (7 кластерів) (рис. 107).

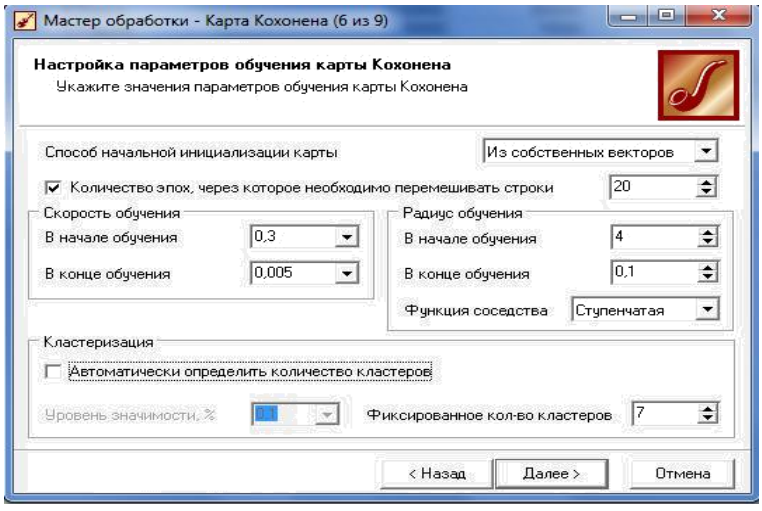


Рисунок 107 – Налаштування параметрів навчання карти Кохонена

На кроці налаштування параметрів навчання можна обрати спосіб початкової ініціалізації карти: випадковим значенням, з множини, що навчає, з власних векторів.

Для прискорення навчання за замовчуванням використовується спосіб ініціалізації – *власних векторів*. При цьому початкові ваги нейронів ініціюються значеннями підмножини гіперплощини, крізь яку проходять два головних власних вектора матриці коваріації вхідних значень вибірки, що навчає.

11 Запусти процес побудови карти Кохонена (процес навчання мережі) на сьомому кроці. Під час навчання можна спостерігати зміну кількості розпізнаних наборів і поточні значення помилок.

Під час навчання треба стежити за зміною середньої та максимальної помилок на графіку навчальної множині. Як правило, на початку навчання графік буде виглядати як лінійна функція, яка монотонно зменшується. Якщо навчання завершується на цьому етапі, то, ймовірно, це не найкращий результат: можна припустити, що помилка і надалі буде зменшуватися. У цьому випадку необхідно збільшити число епох навчання.

Після деякого числа епох можна побачити, що помилка різко зменшується та стабілізується. Графік перетворюється на горизонтальну лінію. Якщо протягом декількох сотень епох змін не відбувається, ймовірно, продовжувати навчання не має сенсу – мережа досягла межі навчання.

12 Після завершення навчання визначити спосіб відображення: візуалізатор *Карта Кохонена* для перегляду результатів Кластеризації, візуалізатор *Що-якщо* для прогнозування прибутковості банків та *Профілі кластерів*.

13 На останньому кроці налаштувати параметри відображення карти Кохонена (рис.108). У списку активних способів відображення карти обрати відображення вхідних та вихідного стовпця, матриці відстаней та кластери.

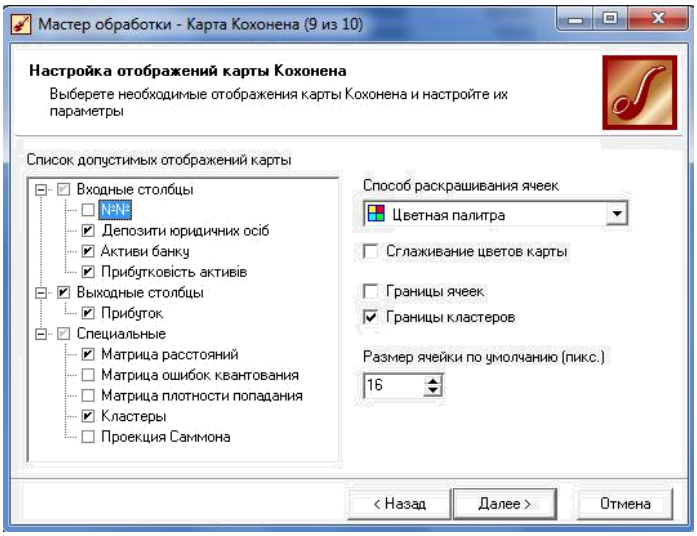


Рисунок 108 – Налаштування відображень карти Кохонена

Вигляд карт Кохонена в вікні *Deductor* наведено на рис. 109.

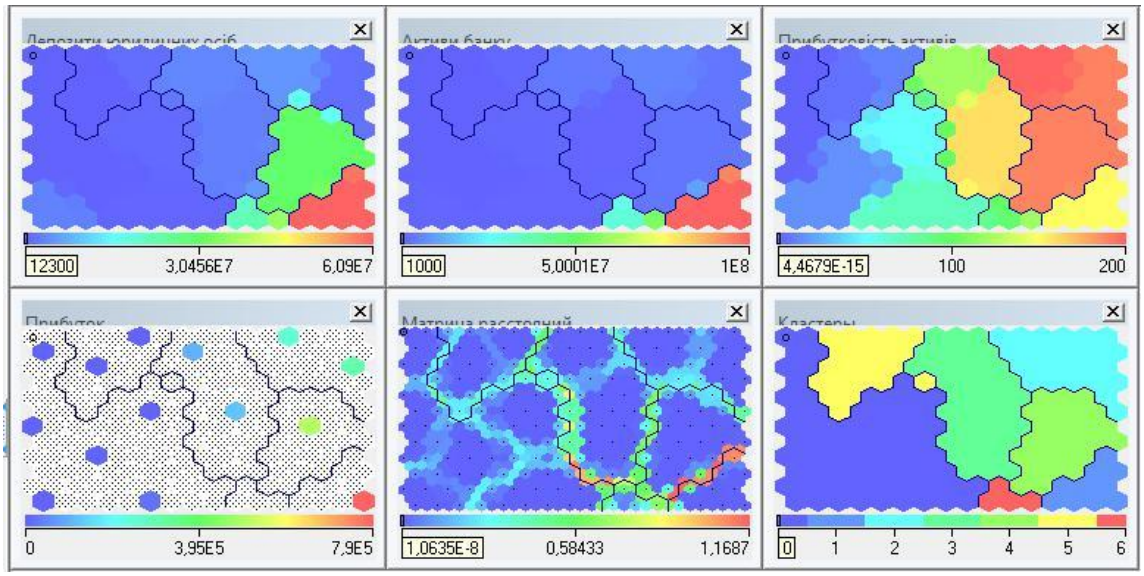


Рисунок 109 – Карти входів та виходів

При аналізі карт входів рекомендується використовувати одразу декілька карт. Розглянемо три карти входів: *Депозити юридичний осіб, Активи банку* та *Прибуток*.

На карті *Депозіти юридичних осіб* (1 карта) виділяємо область з найбільшими значеннями показника депозитів. На цій карті найбільші значення мають об'єкти, розташовані в правому нижньому кутку. Далі має сенс вивчити ці ж нейрони на інших картах.

Розглядаючи одночасно три карти, можна сказати, що ці ж об'єкти мають найбільші значення показника, зображеного на карті *Активи банку* (2 карта) і на карті *Прибуток* (4 карта). Також по розмальовці цих карт можна зробити висновок, що існує взаємозв'язок між цими показниками.

Можна визначити, наприклад, таку характеристику: кластер, розташований в правому верхньому куті на карті *Прібутковість актівів* (3 карта), характеризується високими значеннями показників прибутковості активів, низькими значеннями показників депозитів юридичних осіб і активів банку (порівнювати з правими верхніми кутами карти 1 і 2) і невеликими значеннями показників прибутку (карта 4).

Ця інформація дозволяє охарактеризувати кластер, що знаходиться в правому верхньому кутку наступним чином: це банки з невеликими активами та залученими депозитними коштами від юридичних осіб, але з найбільш прибутковими активами (карта 3).

Це лише фрагмент виводу, який можна зробити, досліджуючи карти.

Шоста карта на рис. 109 – карта кластерів, де кожен кластер виділений окремим кольором. Для знаходження на карті конкретного об'єкта необхідно натиснути правою кнопкою миші на досліджуваному об'єкті, обрати команду *Показати /Скрити дані* (клавіша F4) *й* обрати команду *Знайти клітину на карті*. У результаті виконання команди можна буде побачити як сам об'єкт, так і значення того вимірів, які проглядаються. Таким чином, можна оцінити положення аналізованого об'єкта, а також порівняти його з положенням інших об'єктів.

В результаті застосування карт, що самоорганізуються, багатомірний простір вхідних факторів було подано на площині. Якщо на початку було надано чотиривимірний простір вхідних наборів, то алгоритм трансформував його у двовимірний вигляд, який зручно аналізувати. Також вихідні дані було розподілено за 7 кластерами. Основним візуалізатором є *Карта, що самоорганізується*. На *Матриці відстані* та *Проекції Саммона* наведені відстані між окремими клітинами карти, тобто чіткі границі різних скупчень даних.

Банки були розподілені на групи, для кожної з яких можливо визначення конкретних характеристик, виходячи з розмальовки відповідних показників.

Основна відмінність карт Кохонена від інших моделей полягає в наочності і зручності використання. Вони дозволяють спростити багатовимірну структуру. Їх можна вважати одним з методів проектування багатовимірного простору в простір з більш низькою розмірністю. Інтенсивність кольору в певній точці карти визначається даними, які туди потрапили: клітини з мінімальними значеннями зображуються темно-синім кольором, осередки з максимальними значеннями – червоним.

Друга принципова відмінність карт Кохонена від інших нейромережевих моделей – інший підхід до навчання, а саме некероване або неконтрольоване навчання, який дозволяє оперувати при навчанні лише вхідними наборами змінних. Мережа Кохонена навчається розуміти саму структуру даних і вирішує задачі кластеризації.

У випадку з картами Кохонена, на відміну від деяких інших концепцій нейронних мереж, надлишковий вибір вхідних факторів може призвести до певних наслідків. Незначущі фактори не відкидаються самі собою в ході навчання, а навпаки, істотно впливають на результати кластеризації.

14 Провести кластеризацію банків та їх показників (рис.103). Відобразити дані в вигляді *Карт Кохонена, Профілів кластеров* та візуалізатора *Що-якщо*. Відобразити результати за допомогою візуалізаторів *Матриці відстані, Матриці щільності попадання, Кластери, Проекції Саммона.*

Контрольні питання

1. В чому полягає процес навчання нейромережі?
2. Яка величина характеризує помилку нейромережі?

3 В чому принципова різниця навчання «з вчителем» та «без вчителя»?

4 В якому вигляді представлені нейрони на карті Кохонена?

5 Описати порядок роботи з картой Кохонена.

6 В чому різниця архітектури карт Кохонена та багатошарової нейромережі?

7 Описати візуалізатори карт Кохонена.

Лабораторна робота 9
ОЦІНКА РИЗИКУ КРЕДИТУВАННЯ ФІЗИЧНИХ ОСОБ

Мета роботи – розв'язання практичних задач з використанням методики кредитного скорінгу, методики Дюрана.

Завдання на виконання лабораторної роботи.

1 Є список потенційних позичальників, кожен з яких характеризується певним набором атрибутів (факторів) (рис.110). Оцінити ризик кредитування фізичних осіб з застосуванням технології інтелектуального аналізу даних. Додати до бази даних 30-50 записів.

Для аналізу ринку необхідно в першу чергу оцінити загальну картину: хто та навіщо бере кредити, які є причини відмов у видачі кредитів або неплатоспроможності. Для цього необхідно наочне представлення всіх наявних даних. Таку задачу можна вирішити за допомогою побудови карт, що самоорганізуються.

	A	B	C	D	E	F	G	H
1	№№	Прізвище	Вік	Кіл. утриманців	Середньоміс. дохід	Сума кредиту	Строк кредиту	Мета кредиту
2	1	Лазарев	27	0	3000	80000	12	нерухомість
3	2	Сахно	45	3	2300	60000	12	турпоїздка
4	3	Юхно	23	1	2900	78000	12	нерухомість
5	4	Полищук	20	0	1900	30000	12	оплата за навчання
6	5	Росляков	56	5	2800	10000	12	оплата мед послуг
7	6	Титаренко	60	3	1000	5000	6	оплата мед послуг
8	7	Маляков	67	4	3000	40000	12	ремонт
9	8	Поляков	19	0	1500	10000	12	опата за навчання
10	9	Сидорренко	22	0	1800	10000	12	оплата за навчання
11	10	Крайненко	34	2	2500	50000	12	турпоїздка
12	11	Святненко	39	5	2700	40000	12	турпоїздка
13	12	Лазуренко	26	2	2670	30000	12	ремонт
14	13	Соболенко	25	1	1000	5000	6	розваги
15	14	Дзюбенко	23	1	1000	5000	6	розваги

Рисунок 110 База даних «Кредитування»

2 Оцінити ризик кредитування фізичних осіб (рис.111) за допомогою моделі Дюрана. Додати до бази даних 30-50 записів.

L5 =ЕСЛИ(K5>1,25;"помірний ризик";"відмовити")

	A	B	C	D	E	F	G	H	I	J	K	L
1	№№	Прізвище	Стать	Вік	СтрокПроживання	Професія	НаявністьБанківського рахунку	НаявністьСтраховки	Робота	Зайнятість	Сума балів	ДаватиКредит
2	1	Міщенко	0	0,3	0,06	0,55	0,45	0	0	0,236	1,596	помірний ризик
3	2	Чепенко	0,4	0,3	0,06	55	0	0	0,21	0,23	56,2	помірний ризик
4	3	Сахно	0,4	0,3	0,06	0,55	0	0	0,21	0,23	1,75	помірний ризик
5	4	Волков	0	0,2	0,05	0	0,45	0,19	0	0,12	1,01	відмовити
6	5	Федорченко	0	0,3	0,06	0,55	0	0	0,21	0,17	1,29	помірний ризик
7	6	Гончар	0,4	0,1	0,042	0,55	0	0	0,21	0,059	1,361	помірний ризик
8	7	Гончаренко	0	0,1	0,042	0,55	0	0	0,21	0,059	0,961	відмовити
9	8	Петренко	0	0,2	0,05	0	0,45	0,19	0	0,1	0,99	відмовити
10	9	Фесенко	0,4	0,1	0,042	0,55	0	0	0,21	0,059	1,361	помірний ризик

Рисунок 111 –База даних з використанням коефіцієнтів нарахування балів

Кластеризація показала, що на ринку кредитування фізичних осіб існують не тільки різні напрями (кредитування товарів, освітні кредити), але і різні сегменти позичальників, що користуються одними видами послуг. Отже, для кожної такої групи необхідний свій спосіб класифікації на 'гарних' і 'поганих' позичальників. Модель Дюрана – гарний вибір для розв’язання такої задачі.

1 Загальні положення

Скоринг широко застосовується в банківській сфері. Слово scoring перекладається з англійської як «підрахунок очок» і буквально означає бальну оцінювання чогось. Для банків скоринг – це, в одного боку, метод оцінки ризиків та управління ними на основі прогнозу, коли шукається відповідь на питання щодо ймовірності прострочити платежі за кредитом з боку конкретного позичальника, тобто статистичний метод оцінки його кредитоспроможності. З іншого боку, це процес автоматизації прийняття рішення.

Залежно від задач, які необхідно вирішувати з застосуванням скорингу, він буває декількох видів.

– Application scoring (скоринг заявника) оцінює ймовірність того, що новий клієнт не виплатить кредит;

– Behavioral scoring (поведінковий скоринг) обчислює рівні ризику утворення боргів на основі наявних даних щодо поведінки позичальників;

– Collection scoring (скоринг для роботи з простроченою заборгованістю) визначає, коли і які саме заходи повинні бути запровадженні до неплатників;

– Fraud scoring (скоринг проти шахраїв) оцінює ймовірність того, що новий клієнт є шахраєм;

– Response scoring (скоринг відгуку) оцінює реакцію споживача (відгук) на надані йому пропозиції;

– Attrition scoring (скоринг втрат) оцінює ймовірність використання продукту надалі або перехід до іншого постачальника продукту.

Модель на основі технології *Data Mining* носить назву *data mining Score* і є множиною математичних методів, призначених для виявлення

об'єктивних, неочевидних і в той же час практично корисних закономірностей і взаємозалежностей.

Модель *data mining Score* – це передова технологія побудови скорингу, оскільки, по-перше, вона не передбачає надзнань в точних областях науки, по-друге, є гнучкою системою, яка підлаштовується під економічну ситуацію як країни, так і конкретного банку.

Техніка кредитного скорингу була вперше запропонована американським економістом Д. Дюраном для відбору позичальників за споживчим кредитом. Дюран відмічав, що виведена ним формула може допомогти кредитному робітнику оперативно оцінити якість претендента на позику.

Задача полягає в побудові моделі оцінки (класифікації) потенційних позичальників. В процесі аналізу необхідно отримати достовірну класифікацію, можливість адаптації до будь-яких умов та просту у використанні модель.

На думку експертів, за факторами, які впливають на кредитоспроможність людини, можна оцінити сумарний ризик і віднести потенційного позичальника до однієї з груп – здатних чи нездатних повернути кредит.

Дюран виявив групу факторів, що на його думку, дозволяють з достатньою достовірністю визначити ступінь кредитного ризику при отриманні споживчого кредиту. При нарахуванні балів він рекомендує використовувати наступні коефіцієнти:

– вік: 0,01 за кожний рік більше 20 років (максимум 0,30);

– стать: жіноча 0,40; чоловіча — 0;

– термін проживання в даній місцевості: 0,042 за кожен рік (максимум 0,42).

– професія: 0,55 за професію з низьким ризиком, 0 за професію з високим ризиком, 0,16 – для інших професій;

– робота в галузі: 0,21 підприємства загального користування, державні установи, банки та брокерські фірми;

– зайнятість: 0,059 за кожен рік праці на даному підприємстві (максимум 0,59);

– фінансові показники: 0,45 за наявність банківського рахунку, 0,35 за володіння нерухомістю, 0,19 при наявності полісу страхування життя.

Застосовуючи ці коефіцієнти, Дюран визначив межу, що розподіляє “надійних ” та “ поганих ” позичальників – 1,25 бала. Клієнт, що отримав більше 1,25 бала, може бути віднесений до групи помірного ризику, а той, що отримав менше 1,25 бала, вважається ненадійним. Метод скорингу дозволяє провести експрес аналіз заявки на кредит в присутності клієнта.

При аналізі ділових позик також застосовуються різні прийоми кредитного скорингу від найпростіших формул до складних математичних моделей. Статистичним алгоритмом автоматичного розрахунку балів скорингової карти сьогодні є логістична регресія. Допомогу при перевірці

гіпотез кредитоспроможності фізичних осіб може надати реалізований в *Deductor* факторний аналіз, який виявляє значущість тих чи інших факторів [34].

2 Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc*) базу даних (рис. 110).

2 При імпорті файлу для полів бази даних встановити наступні призначення:

–для поля №№ та ПРІЗВИЩЕ – *Інформаційне*;

–для поля ВІК, КІЛЬКІСТЬ УТРИМАНЦІВ, СЕРЕДНЬОМІСЯЧНИЙ ДОХІД, СУМА КРЕДИТУ, СТРОК КРЕДИТУ – *Вхідне*;

–для поля МЕТА КРЕДИТУ – *Вихідне*.

3 Запустити процес побудови *карти Кохонена* з фіксованою кількістю кластерів.

4 По завершенні навчання визначити спосіб відображення в списку візуализаторів: *Карта Кохонена, Профілі кластерів, Набір, що навчає, Таблиця спряженості* і *Що-якщо*.

5 На останньому кроці *Майстра* обрати *карти Кохонена з* відображення всіх вхідних, вихідних стовпців, кластерів та матриці відстаней.

На рис. 112 наведені карти, які показують розподіл позичальників за характеристиками "Сума кредиту", "Термін кредиту", "Мета кредиту", "Середньомісячний дохід", "Кількість утриманців" і "Вік". Проведемо аналіз представлених даних.

Проаналізуємо кожну характеристику окремо, а потім оцінимо їх загальний зв'язок. На поточному рисунку на картах виділені кластери, які містять молодих людей віком до 27 років, самотніх, з невеликим доходом (1000), але які беруть грошовий кредит в розмірі 5000 на півроку для розваг.

6 Сформулювати інформаційну модель оцінки потенційних позичальників.

7 Створити в табличному процесорі (*MS Excel, Calc*) базу даних (рис.111).

8 Розрахувати в діапазоні клітин K2:K10 загальну суму балів для кожного позичальника.

9 Ввести в діапазон клітин L2:L10 формулу, яка визначає, чи досяг позичальник граничного значення суми балів для надання йому кредиту

=ЕСЛИ (K1>1,25;”помірний ризик “; ”відмовити“).

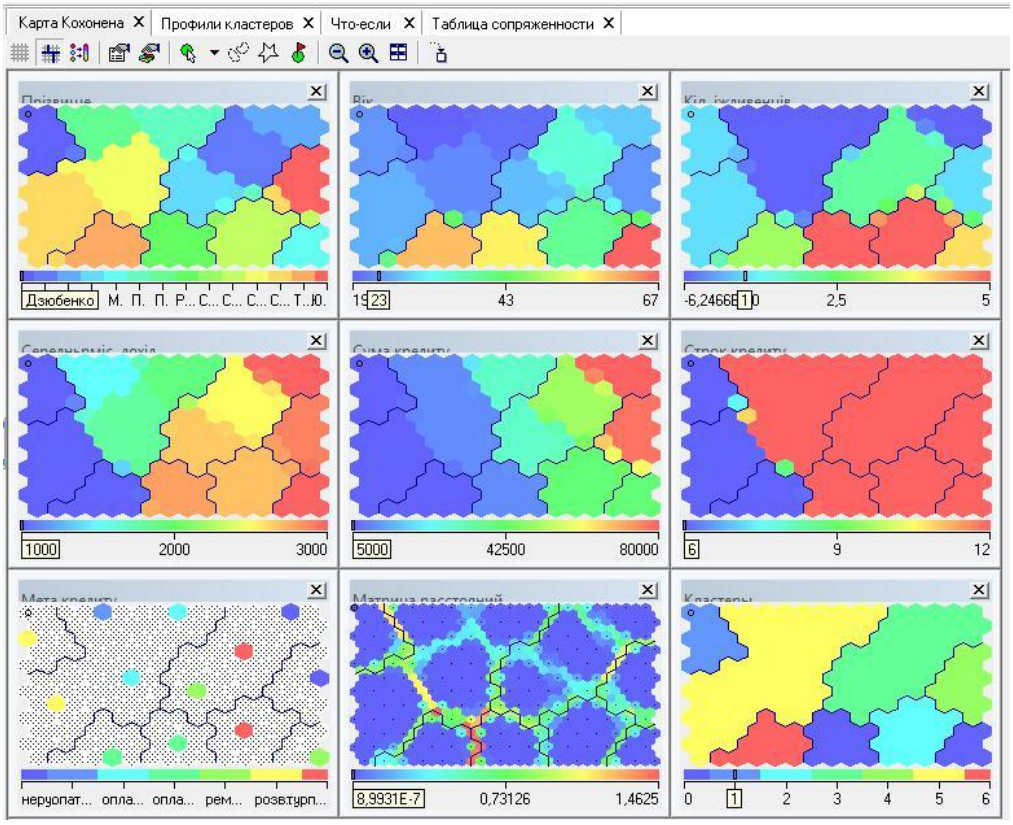


Рисунок 112 – Карти, що показують розподіл позичальників за характеристиками

10 Встановити при імпорті файлу наступні параметри полів:

– №№ – *дійсне, неперервне, Інформаційне*;
– ПРІЗВИЩЕ– *строкове, дискретне*, *Інформаційне*;
– СТАТЬ – дійсне, безперервне, *Інформаційне*;
– ВІК– *дійсне*, *безперервне*, *Вхідне*;
– ТЕРМІН ПРОЖИВАННЯ – *дійсне, безперервне*, *Вхідне*;
– ПРОФЕСІЯ – *дійсне, безперервне*, *Вхідне*;
– НАЯВНІСТЬ БАНКІВСЬКОГО РАХУНКУ– *дійсне, безперервне*, *Вхідне*;
– НАЯВНІСТЬ СТРАХОВКИ – дійсне, безперервне, *Вхідне*;
– СТРОК РОБОТИ – дійсне, безперервне, *Вхідне*;
– ЗАЙНЯТІСТЬ – дійсний, безперервний, *Вхідне*;
– ДАВАТИ КРЕДИТ – дійсний, безперервний, *Вихідне*.

11Запусти процес побудови *карти Кохонена* з фіксованою кількістю кластерів.

12Визначити спосіб відображення карт Кохонена самостійно.

13Проаналізувати отримані результати та сформулювати інформаційну модель оцінки потенційних позичальників.

14 Здійснити постановку власної задачі та розв’язати її з використанням технології *Data Mining.*

Контрольні питання

1 Який метод, реалізований в *Deductor,* може допомогти у перевірці гіпотез кредитоспроможності фізичних осіб?

2 Які фактори на думку експертів впливають на кредитоспроможність людини?

3 Які задачі розв’язуються при побудові моделі оцінки потенційних позичальників?

4 Які задачі в банківській сфері розв’язуються з використанням технології *Data Mining*?

5 Які моделі оцінки ризиків в банківській сфері Вам відомі?

6 Які види скорінгу Вам відомі?

7 Охарактеризувати фактори, що на думку Дюрана, дозволяють з достатньою достовірністю визначити ступінь кредитного ризику при отриманні споживчого кредиту.

8 Назвати статистичний алгоритм автоматичного розрахунку балів скорингової карти.

Лабораторна робота 10
РОБОТА З ДЕРЕВАМИ РІШЕНЬ

Мета роботи – побудова дерева рішень з застосуванням класифікуючих правил.

Завдання на виконання лабораторної роботи.

1Використати механізм побудови дерева рішень для визначення:

а) розподілу менеджерів країни при укладанні договорів за регіонами;

б) ефективності діяльності менеджерів при укладанні договорів в різних сферах діяльності.

Додати до бази даних 30-50 записів (рис.113) та провести моделювання.

2Побудувати ієрархічне дерево класифікуючих правил для визначення прибутковості вкладень, якщо дохід розраховується за формулою

=(E2*C2+E2)*F2,

а вихідні дані (вибірка) наведені на рис. 114.

Додати до бази даних 30-50 записів (рис.114) та провести моделювання.

3 Побудувати ієрархічне дерево класіфікуючих правил для визначення індексу прибутковості грошових надходжень банків. Фрагмент бази даних з вихідними даними (вибірка) наведений на рис 115, причому:

	A	B	C	D	E	F	G	H	I	J
1	№№	Статус клієнту	Менеджер	Регіон	Віддаленість	Сфера діяльності	Кіл. листів	Кіл. факсів	Кіл. тел.розмов	Стан угоди
2	1	юр. особа	Міщенко	Схід	6000	будівництво	5	6	33	успіх
3	2	фіз. особа	Тихонов	Центр	120	будівництво	2	6	12	невдача
4	3	фіз. особа	Сидоров	Південь	3000	послуги	3	5	6	невдача
5	4	фіз. особа	Шевцов	Північ	1000	будівництво	4	6	7	успіх
6	5	юр. особа	Міщенко	Захід	4000	послуги	7	9	28	невдача
7	6	юр. особа	Міщенко	Захід	3400	послуги	6	11	18	невдача
8	7	юр. особа	Сидоров	Південь	2900	послуги	7	7	26	успіх
9	8	фіз. особа	Тихонов	Центр	290	страхування	2	3	11	успіх
10	9	фіз. особа	Тихонов	Центр	50	страхування	3	2	9	успіх
11	10	юр. особа	Міщенко	Захід	1990	страхування	6	14	26	успіх
12	11	фіз. особа	Сидоров	Північ	780	послуги	2	5	7	невдача
13	12	юр. особа	Міщенко	Захід	2100	будівництво	5	14	35	невдача
14	13	юр. особа	Міщенко	Схід	4900	фінансуванння	4	9	14	успіх
15	14	фіз. особа	Шевцов	Схід	2000	фінансуванння	3	2	9	успіх
16	15	фіз. особа	Шевцов	Схід	3900	фінансуванння	2	7	8	успіх

Рисунок 113 – База даних діяльності менеджерів

G2 =(E2*C2+E2)*F2

	A	B	C	D	E	F	G
1	№ операції	Вид вкладення	Прибутковість, в рік	Ризик	Сума	Термін вкладення, років	Дохід
2	1	Банківський вклад	11%	незначний	10000	1	11100
3	2	Металевий внесок	35%	високий	6785	0,75	6869,8125
4	3	Кредитний кооператив	30%	середній	1230	2	3198
5	4	Нерухомість	30%	середній	5000000	1	6500000
6	5	Ювелірні вироби	15%	вище середнього	4000	0,5	2300
7	6	ПІФ	25%	нижче середнього	3400	2	8500
8	7	ОФБУ	40%	вище середнього	67000	1	93800
9	8	Довірче управління	70%	за договором	12899	1	21928,3
10	9	Металевий внесок	35%	високий	4502	3	18233,1
11	10	ПІФ	25%	нижче середнього	459900	5	2874375
12	11	Банківський вклад	11%	незначний	43222	4	191905,68
13	12	Банківський вклад	11%	незначний	55500	2	123210
14	13	Самост. управління	55%	середній	6000	3	27900
15	14	Накопич. Страховка	10%	незначний	40000	2	88000
16	15	Цінні папери	80%	високий	17890	3	96606
17	16	ОФБУ	40%	вище середнього	340	1	476
18	17	Металевий внесок	35%	високий	2300	2	6210
19	18	Кредитний кооператив	30%	середній	7800	3	30420
20	19	Кредитний кооператив	30%	середній	43210	0,75	42129,75
21	20	Цінні папери	80%	високий	6000	0,5	5400
22	21	Банківський вклад	11%	незначний	7000	2	15540
23	22	Ювелірні вироби	15%	вище середнього	56000	1	64400
24	23	Довірче управління	70%	за договором	45550	0,25	19358,75
25	24	Металевий внесок	35%	високий	34	1	45,9
26	25	Ювелірні вироби	15%	вище середнього	9000	2	20700

Рисунок 114 – Фрагмент бази даних для розрахунку прибутковості вкладень

U3 =S3/B3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		Обсяг вкладу, тис.	Множник дисконтування				Ставка дисконту	Грошові надходження, тис. грн.				Дисконтовані грошові надходження, тис. грн.				Множник дисконтування зі зростаючим підсумком				Чистий приведений прибуток	Індекс прибутковості
2	Банк	грн	d1	d2	d3	d4	вання	1	2	3	4	1	2	3	4	1	2	3	4		
3	Приватбанк	1230,3	0,83	0,69	0,58	0,48	0,2	1500	1850	2000	1250	1250,00	1284,72	1157,41	602,82	1250,00	2534,72	3692,13	4294,95	3064,65	3,49
4	Русский стандарт	1590,3	0,77	0,59	0,46	0,35	0,3	1250	1300	1400	1900	961,54	769,23	637,23	665,24	961,54	1730,77	2368,00	3033,24	1442,94	1,91
5	ОТР-банк	1253,3	0,71	0,51	0,36	0,26	0,4	1350	1450	1502	1800	964,29	739,80	547,38	468,55	964,29	1704,08	2251,46	2720,01	1466,71	2,17
6	Укрсоцбанк	1485,2	0,74	0,54	0,40	0,29	0,36	1360	1460	1560	1700	1000,00	789,36	620,17	496,93	1000,00	1789,36	2409,53	2906,45	1421,25	1,96
7	Укрсиббанк	2236,9	0,77	0,59	0,46	0,35	0,3	1420	1480	1592	1750	1092,31	875,74	724,62	612,72	1092,31	1968,05	2692,67	3305,40	1068,50	1,48
8	Аваль	1699,3	0,80	0,64	0,51	0,41	0,25	1460	1500	1650	1800	1168,00	960,00	844,80	737,28	1168,00	2128,00	2972,80	3710,08	2010,78	2,18
9	Ощадбанк	1893,2	0,71	0,51	0,36	0,26	0,4	1700	1760	1890	1950	1214,29	897,96	688,78	507,60	1214,29	2112,24	2801,02	3308,62	1415,42	1,75
10	Мегабанк	5520,3	0,67	0,44	0,30	0,20	0,5	1600	1690	1760	1900	1066,67	751,11	521,48	375,31	1066,67	1817,78	2339,26	2714,57	-2805,73	0,49
11	Юникредит	3582,3	0,82	0,67	0,55	0,45	0,22	1500	1596	1670	1800	1229,51	1072,29	919,68	812,52	1229,51	2301,80	3221,48	4034,00	451,70	1,13
12	Банк Кипру	7569,3	0,78	0,61	0,48	0,37	0,28	1550	1600	1750	2300	1210,94	976,56	834,47	856,82	1210,94	2187,50	3021,97	3878,78	-3690,52	0,51
13	Меркурий	1478,3	0,81	0,66	0,54	0,44	0,23	1600	1700	1800	2390	1300,81	1123,67	967,29	1044,19	1300,81	2424,48	3391,77	4435,96	2957,66	3,00
14	Базис	12501	0,80	0,64	0,51	0,41	0,25	1630	1730	1800	2900	1304,00	1107,20	921,60	1187,84	1304,00	2411,20	3332,80	4520,64	-7980,36	0,36
15	Правекс	9000	0,74	0,55	0,41	0,30	0,35	1670	1770	1890	3000	1237,04	971,19	768,18	903,20	1237,04	2208,23	2976,41	3879,61	-5120,39	0,43
16	ПУМБ	12286	0,77	0,59	0,46	0,35	0,3	1700	1150	1959	2000	1307,69	680,47	891,67	700,26	1307,69	1988,17	2879,84	3580,09	-8705,91	0,29

Рисунок 115 – Визначення індексу прибутковості

–в діапазон клітин A1:B16 введено назви банків та обсяг вкладів в тис. грн.;

– в діапазон клітин G3:G16 введено ставку дисконтування;

– в діапазон клітин H3:K16 введено обсяги грошових надходжень в тис. грн.;

–в діапазоні клітин C3:F16 розраховано множники дисконтування:

для d1 – за формулою =1/(1+G3)^1;

для d2 – за формулою =1/(1+G3)^2 ;

для d3 – за формулою =1/(1+G3)^3;

для d4 – за формулою =1/(1+G3)^4;

–в діапазоні клітин L3:O16 розраховано дисконтовані грошові надходження, тис. грн., а саме:

в діапазоні L3:L16 – за формулою =H3*C3;

в діапазоні M3:M16 – за формулою =I3*D3;

в діапазоні N3:N16 – за формулою =J3*E3;

в діапазоні O3:O16 – за формулою =K3*F3;

– в діапазоні клітин P3:S16 розраховано множник дисконтування зі зростаючим підсумком, а саме:

в діапазоні P3:P16 – за формулою =L3;

в діапазоні Q3:Q16 – за формулою =P3+M3;

в діапазоні R3:R16 – за формулою =Q3+N3;

в діапазоні S3:S16 – за формулою =R3+O3;

– в діапазоні клітин T3:T16 розраховано чистий приведений прибуток за формулою =S3B3;

– в діапазоні клітин U3:U16 розраховано індекс прибутковості за формулою =S3/B3.

Додати до бази даних 30-50 записів (рис 115) та провести моделювання.

1 Загальні положення

Дерево рішень (decision trees) – це графічне подання послідовності взаїмопов’язанних рішень у вигляді ієрархічної структури.

Дерево рішень є одним із способів розбиття множини даних на класи або категорії. Корінь дерева неявно містить усі дані, що мають бути класифіковані, а гілки – остаточне розбиття на класи після завершення процесу класифікації. Проміжні вузли дерева є пунктами прийняття рішення щодо вибору.

Нехай маємо T – множину, що навчає, яка містить об'єкти, кожен з яких характеризується m атрибутами, причому один з них вказує на належність об'єкта до певного класу, причому $\{C_1, C_2, \ldots C_k\}$ – класи. Тоді існують три ситуації:

– множина T містить хоча б один набір, що відносяться до класу C_k. Тоді дерево рішень для T – це гілка, який визначає клас C_k;

– множина T не містить жодного набору, тобто є порожньою. Тоді клас, асоційований з цією гілкою, обирається з множини, відмінної від T;

– множина T містить набори, що відносяться до різних класів. Тоді доцільно розбити множину T на підмножини. При цьому обирається одна з ознак, що має два і більше відмінних один від одного значень $O_1, O_2, ...O_n$.

Множина T розбивається на підмножини $T_1, T_2, ... T_n$, де кожна підмножина T_i містить всі набори, що мають значення O_i для обраної ознаки. Це процедура є рекурсивною і завершується, коли кінцева множина буде містити набори, що відносяться до одного класу.

Наведена процедура є підґрунтям більшості сучасних алгоритмів побудови дерева рішень та відома під назвою методу «поділу і захоплення». Очевидно, що при цьому побудова дерева рішень відбувається зверху вниз [37].

Дерева рішень або дерева вирішальних правил типу «Якщо... То ...» (if then), призначені для рішення задач класифікації та генерують ієрархічну структуру. Для прийняти рішення щодо віднесення деякого об'єкта або ситуації до певного класу, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд «Значення параметра *A* більше *B*?». Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, де слід відповісти на наступне питання, пов'язане з відповідним вузлом і т. д. Наведений приклад ілюструє роботу так званих бінарних дерев рішень, в кожному вузлі яких розгалуження проводиться за двома напрямами (тобто на запитання, поставлене у вузлі, є тільки два варіанти відповідей, наприклад *Так* або *Ні*). Однак, в загальному випадку, відповідей а, отже, гілок, що виходять з вузла, може бути більше.

Якість побудованого дерева після навчання можна оцінити за кількома параметрами. По-перше, це число розпізнаних випадків у навчальному і тестовому наборах даних. Чим воно більше, тим вище якість побудованого дерева. По-друге, це кількість вузлів дерева. Велика кількість утруднює сприйняття інформації та свідчить про слабку залежність вихідного від вхідних полів. Кожне правило характеризується підтримкою і достовірністю.

Підтримка – це загальна кількість випадків, класифікованих даним вузлом дерева.

Достовірність – це кількість правильно класифікованих даним вузлом випадків [24].

Дерева рішень дозволяють проводити аналіз значущих чинників через те, що при визначенні параметра, за яким на кожному рівні ієрархії відбувається розподіл на дочірні вузли, використовується критерій максимального усунення невизначеності. Таким чином, в процесі

класифікації більш значущі фактори знаходяться на ближче до кореня дерева, ніж менш значущі.

Майстри аналітичної платформи *Deductor Studio* пропонують широкі можливості з налаштування процесу побудови дерева рішень. Це і налаштування призначення стовпців, способів нормалізації, джерел даних для вчителя (тестова множина та множина, що навчає), кількості наборів у вузлі й достовірності правил. Після побудови дерева алгоритм сам виявляє ступінь впливу тих чи інших факторів на результат та відсікає несуттєві фактори, описує за допомогою формальних правил спосіб класифікації, а також видає інформацію щодо достовірності та підтримки правил [3, 36].

2 Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc)* базу даних, наведену на рис. 113.

2 Імпортувати створений файл в аналітичну платформу *Deductor.*

3 Встановити для полів наступні призначення:

– №№, СТАТУС КЛІЄНТА, КІЛЬКІСТЬ ЛИСТІВ, КІЛЬКІСТЬ ФАКСІВ, КІЛЬКІСТЬ ТЕЛЕФОНИХ РОЗМОВ – *Інформаційне*;

– МЕНЕДЖЕР, РЕГІОН, ВІДДАЛЕНІСТЬ, СФЕРА ДІЯЛЬНОСТІ – *Вхідне*;

– СТАН УГОДИ – *Вихідне.*

4 Обрати як способи відображення *Таблиця та Статистика.*

5 Обрати в групі *Data Mining Майстра обробки* пункт *Дерево рішень* та перевірити налаштування призначення полів.

6 Налаштувати спосіб розбиття початкової множини даних на множину, що навчає, та тестову. Обрати випадковий спосіб розбиття, при якому дані для тестової множини та множини, що навчає, обираються з початкового набору випадковим чином.

7 Налаштувати параметри процесу навчання, а саме, мінімальну кількість наборів, при якому буде створено новий вузол (нехай вузол створюється, якщо в нього потрапили два і більше набори), а також обрати можливість будувати дерево з більш достовірними правилами (рис.116).

8 Обрати процес побудови дерева рішення в автоматичному або інтерактивному (напівавтоматичний) режимі та побачити інформацію про кількість розпізнаних наборів (рис.117).

Обрати необхідні способи візуалізації отриманих результатів: аналіз *Що-якщо, Дерево рішень, Правила, Значимість атрибутів, Набір, що навчає, Таблиця спряженості.*Механізм віднесення повинен бути таким, щоб можна було вказати, який менеджер, в якому регіони країни та в якій професійній сфері укладав угоду, та наскільки вона була результативна. Такий механізм пропонує візуалізатор *Що-якщо* (рис.118).

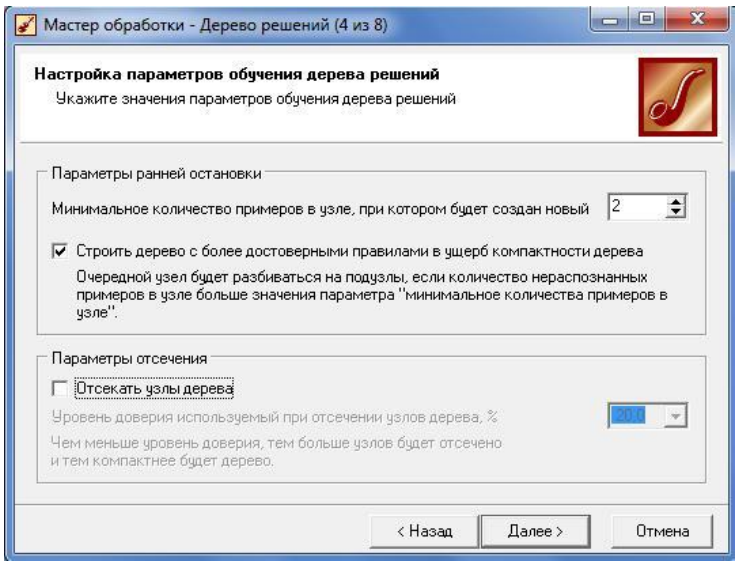


Рисунок 116 – Налаштування параметрів навчання дерева рішень

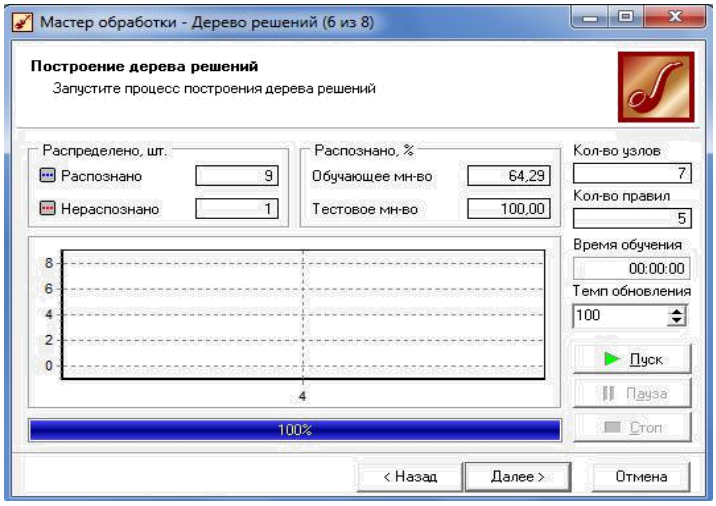


Рисунок 117 – Побудова дерева рішень

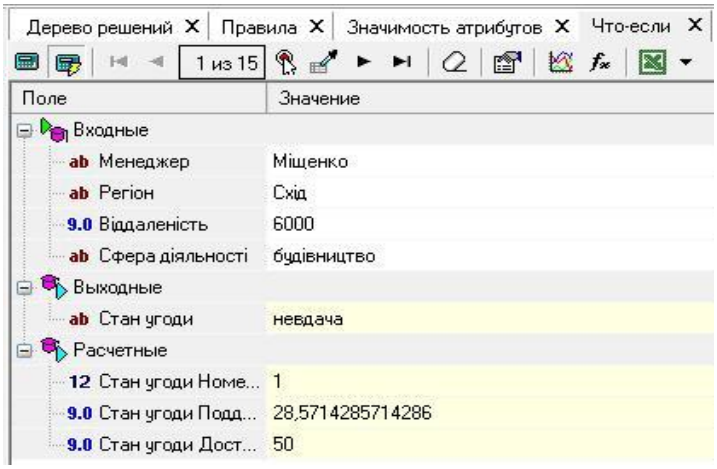
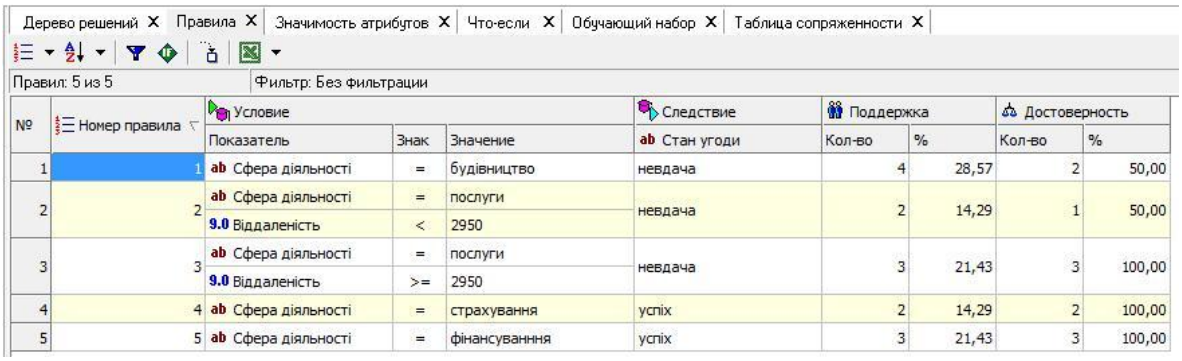


Рисунок 118– Візуалізатор *Що-якщо*

Не менш важливим є перегляд самого дерева рішень (візуалізатор *Дерево рішень*), бо це дозволяє можна визначити, які фактори є важливими (верхні вузли дерева), які – другорядними, а які – взагалі впливають на результат (рис.119). Як критерій розщеплення платформою було обрано СФЕРУ ДІЯЛЬНОСТІ.



№	Номер правила	Условие			Следствие	Поддержка		Достоверность	
		Показатель	Знак	Значение	Стан угоди	Кол-во	%	Кол-во	%
1	1	Сфера діяльності	=	будівництво	невдача	4	28,57	2	50,00
2	2	Сфера діяльності	=	послуги	невдача	2	14,29	1	50,00
		Віддаленість	<	2950					
3	3	Сфера діяльності	=	послуги	невдача	3	21,43	3	100,00
		Віддаленість	>=	2950					
4	4	Сфера діяльності	=	страхування	успіх	2	14,29	2	100,00
5	5	Сфера діяльності	=	фінансуванння	успіх	3	21,43	3	100,00

Рисунок 119 – Візуалізатор *Дерево рішень*

Формалізовані правила класифікації, виражені у формі "Якщо <Умова>, тоді <Стан угоди>", можна побачити у візуалізаторі *Правила* (рис.120), де *Правила* подані у вигляді таблиці, полями якої є номер правила, умова, наслідок, підтримка (кількість і відсоток наборів вихідної вибірки, які відповідають цій умові), достовірність (відношення кількості вірно розпізнаних наборів, що відповідають умові, до загального кількості наборів, що відповідають умові, виражене у відсотках).

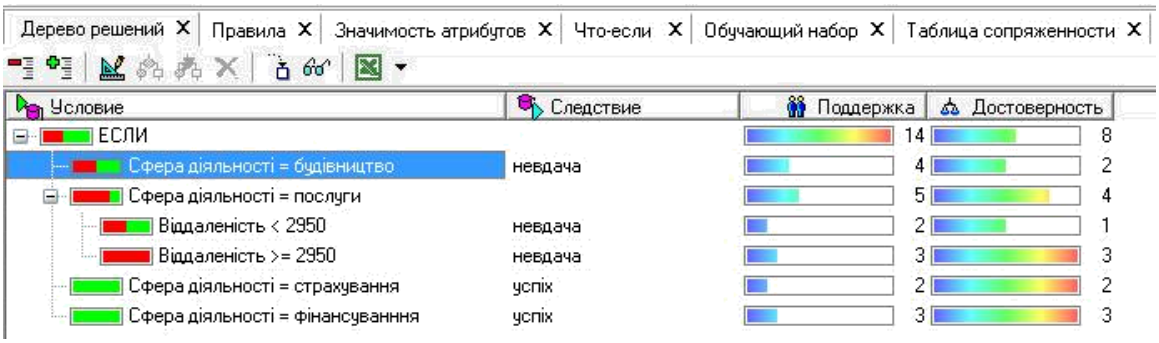


Рисунок 120 – Візуалізатор *Правила*

Даний візуалізатор надає можливість перегляду наборів, які потрапили в той чи інший вузол, а також інформацію про сам вузол. Найбільш значущим фактором в цьому прикладі є СФЕРА ДІЯЛЬНОСТІ.

Часто аналітику стає на користь дізнатися про кількість невірно розпізнаних наборів та які саме набори були віднесені до певного класу

помилково. На це питання дає відповідь візуалізатор *Таблиця спряженості* (рис.121). В цьому випадку дерево правильно класифікувало 80 % прикладів.

	Классифицировано		
Фактически	невдача	успіх	Итого
невдача	6		6
успіх	3	6	9
Итого	9	6	15

Рисунок 121 – Візуалізатор *Таблиця спряженості*

Інформацію щодо ступеня впливу кожного фактора на класифікацію надає візуалізатор *Значимість атрибутів* (рис.122).

Целевой атрибут: Стан угоди			
№	Номер	Атрибут	Значимость, %
1	4	Сфера діяльності	79,346
2	3	Віддаленість	20,654
3	1	Менеджер	0,000
4	2	Регіон	0,000

Рисунок 122 – Візуалізатор *Значущість атрибутів*

За допомогою даного візуалізатора можна визначити, наскільки сильно вихідне поле залежить від кожного з вхідних факторів. Чим більше значимість атрибуту, тим більший внесок він вносить при класифікації. В даному випадку найбільший внесок вносить СФЕРА ДІЯЛЬНОСТІ, як і було вказано вище.

10 Внести деякі коригування, для чого необхідно обрати інтерактивний режим побудови. Для внесення змін використовуються наступні кнопки панелі інструментів *Дерева рішень*: *Вмикання/вимикання інтерактивного режиму; Розбити поточний вузол на підвузли; Побудувати дерево рішень починаючи з поточного вузла* [36].

10.1 Натиснути кнопку *Вмикання/вимикання інтерактивного режиму* для активації режиму;

10.2 Обрати кореневий каталог в дереві рішень, натиснути кнопку *Розбити поточний вузол на підвузли* та у вікні, що з'явиться, обрати МЕНЕДЖЕР (рис.123). Натиснути кнопку Ok.

В результаті отримаємо нове дерево з новими правилами (рис.124).

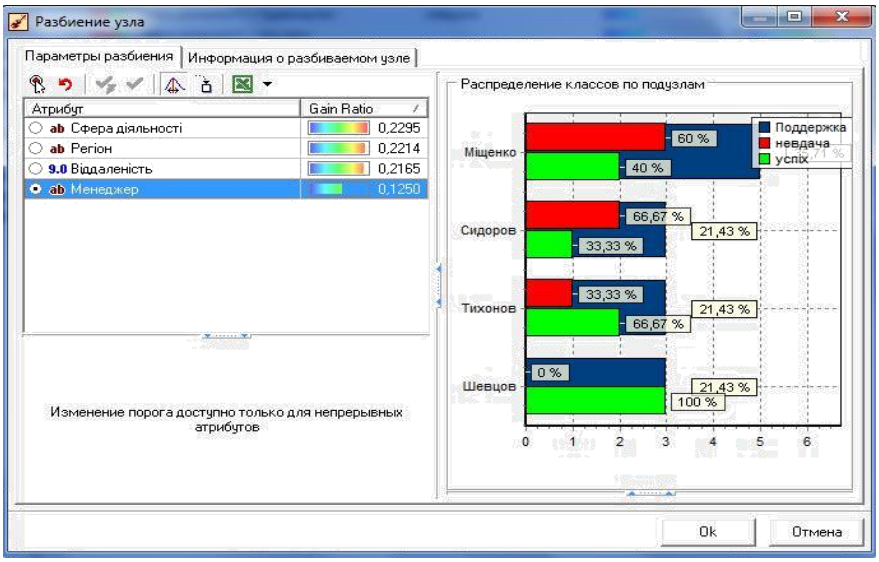


Рисунок 123 – Діалогове вікно РОЗБИТТЯ ВУЗЛА з виділеним атрибутом МЕНЕДЖЕР

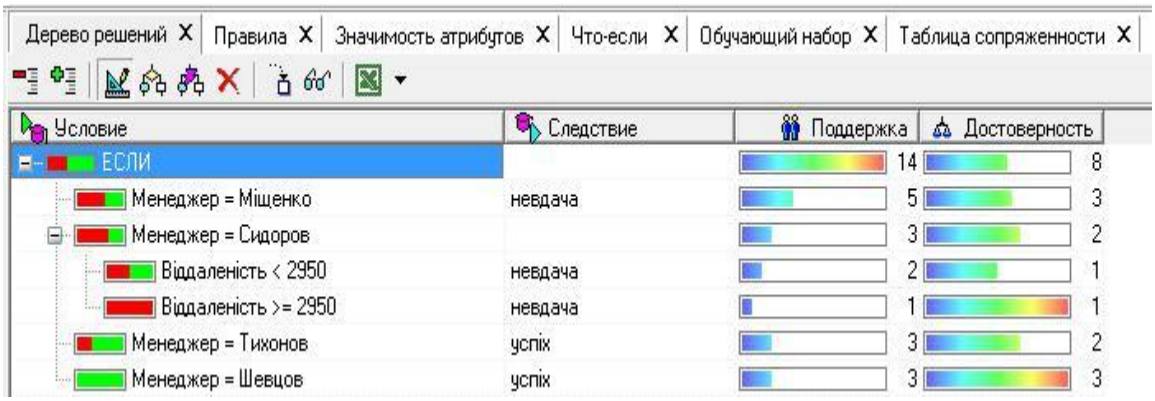


Рисунок 124 – Дерево з новими правилами

11 Визначити ефективності діяльності менеджерів при укладанні угод в різних сферах діяльності. Побудувати модель *Дерево рішень* при вихідному полі МЕНЕДЖЕР.

12 Змінити призначення полів та з’ясувати, як змінюються правила.

13 Побудувати ієрархічне дерево правил, що класифікують, для визначення прибутковості вкладень (рис.114).

14 Побудувати ієрархічне дерево правил, що класифікують, для визначення індексу прибутковості грошових надходжень банків (рис.115).

15 Вирішити власну економічну задачу.

Контрольні питання

1 З яких елементів складається дерево рішень?

2 Для яких задач використовуються дерева рішень?

3 Який вигляд має структура правил, що класифікують?

4 За якими параметрами можна оцінити якість побудованого дерева?

Лабораторна робота 11
СЕГМЕНТАЦІЯ КЛІЄНТІВ ЗА ДОПОМОГОЮ КАРТ КОХОНЕНА ТА ДЕРЕВ РІШЕНЬ

Мета роботи – побудова карти Кохонена та ієрархічного дерева правил, що класифікують.

Завдання на виконання лабораторної роботи.

Використати пасивну рекламу для залучення фірмою нових клієнтів в залежності від їх сфери діяльності (сегментація клієнтів) (рис.125). Додати до бази даних 30 -50 записів та розв’язати задачу в такій постановці.

	A	B	C	D	E	F
1	№№	Статус	Віддаленість району	Сфера діяльності	Середньогодова сума угоди	Частота угоди
2	1	фіз. особа	900	народна освіта	120	3
3	2	фіз. особа	790	будівництво	500	2
4	3	фіз. особа	980	управління	1300	2
5	4	фіз. особа	50	зв'язок	500	3
6	5	фіз. особа	320	будівництво	500	2
7	6	фіз. особа	400	послуги	500	3
8	7	фіз. особа	15	послуги	500	2
9	8	фіз. особа	35	послуги	500	3
10	9	фіз. особа	40	послуги	500	3
11	10	фіз. особа	10	харчова промисловість	500	3
12	11	фіз. особа	5	фінанси, кредит, стахування	500	3
13	12	юрид.особа	600	консультація та аудит	1600000	1
14	13	юрид.особа	650	харчова промисловість	80000	1
15	14	юрид.особа	80	будівництво	3000	1
16	15	юрид.особа	900	фінанси, кредит, стахування	50000	1
17	16	юрид.особа	10	управління	2200	3
18	17	юрид.особа	810	фінанси, кредит, стахування	20000	2
19	18	юрид.особа	720	консультація та аудит	5000000	2
20	19	юрид.особа	10	консультація та аудит	1000	2
21	20	юрид.особа	50	народна освіта	1000	3
22	21	юрид.особа	600	зв'язок	1000	2
23	22	юрид.особа	1000	послуги	90000	2
24	23	фіз. особа	550	консультація та аудит	1200	3

Рисунок 125 – База даних «Сегментація клієнтів»

1 Загальні положення

Найбільш активні фірми для залучення нових клієнтів використовують як, рекламу на телебаченні, радіо та в пресі, тобто пасивну рекламу, так і розсилку з прямими комерційними пропозиціями. Для підвищення ефективності подібних заходів необхідно враховувати інтереси клієнтів, тобто пропонувати їм саме той товар, якому вони віддають перевагу. Для цього необхідно виділяти деякі групи, сегменти, клієнтів і їм пропонувати конкретні категорії товарів.

Виділяти сегменти клієнтів можна за різними принципами. Це може бути угрупування за сферою діяльності або за географічним розташуванням. Після сегментування можна дізнатися, яка саме група є найбільш активною,

хто має найбільший прибуток і найбільш лояльних клієнтів, виділити характерні для них ознаки. Для вирішення цього завдання скористаємося *картами Кохонена та деревами рішень.*

2 Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel, Calc*) базу даних (рис. 125).

2 Встановити для всіх полів бази даних (№№, СТАТУС, ВІДДАЛЕНІСТЬ РАЙОНУ, СФЕРА ДІЯЛЬНОСТІ, СЕРЕДНЬОГОДОВА СУМА УГОДИ, ЧАСТОТА УГОДИ) призначення – *Вхідне.* Для типа даних *Строковий* обрати вид даних – *Дискретний*, а для типа даних *Дійсний* – вид даних – *Безперервний*.

3 Запусти процес побудови *карти Кохонена*, перевірити призначення полів та задати число кластерів таким, що дорівнює 7.

4 Обрати як способи відображення *Карта Кохонена*, *Профілі кластерів*, *Набір, що навчає,* та *Що-якщо*.

5 Вказати на 9 кроці *Майстра відображення карти Кохонена* налаштування як показано на рис.126.

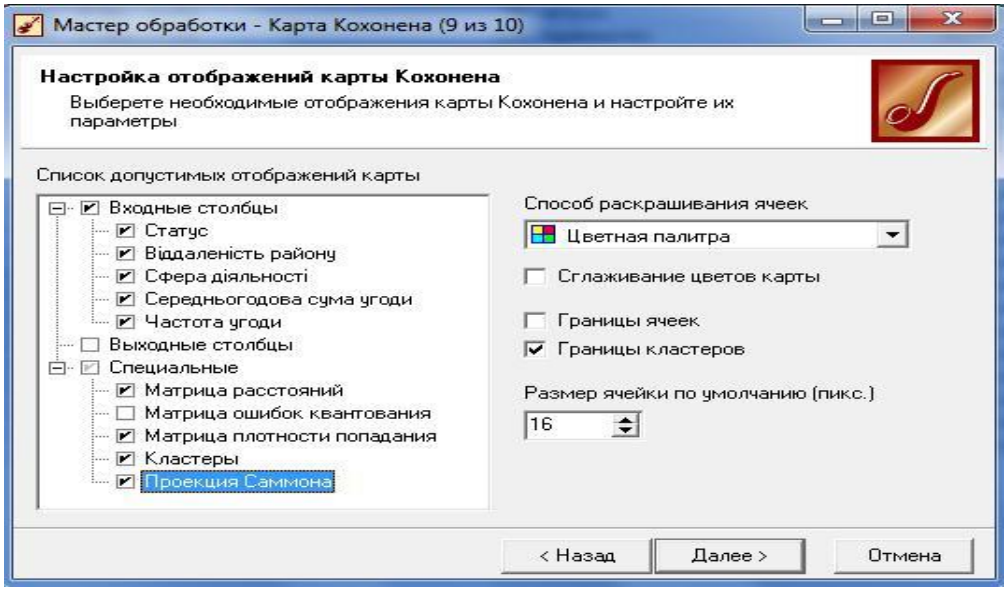


Рисунок 126 – Налаштування *відображення карти Кохонена*

На рис. 127 представлені карти Кохонена, які отримані після кластеризації клієнтів агентства за характеристиками: статус, віддаленість району, сфера діяльності, середньорічні суми угод і частота угод (кількість угод на тиждень).

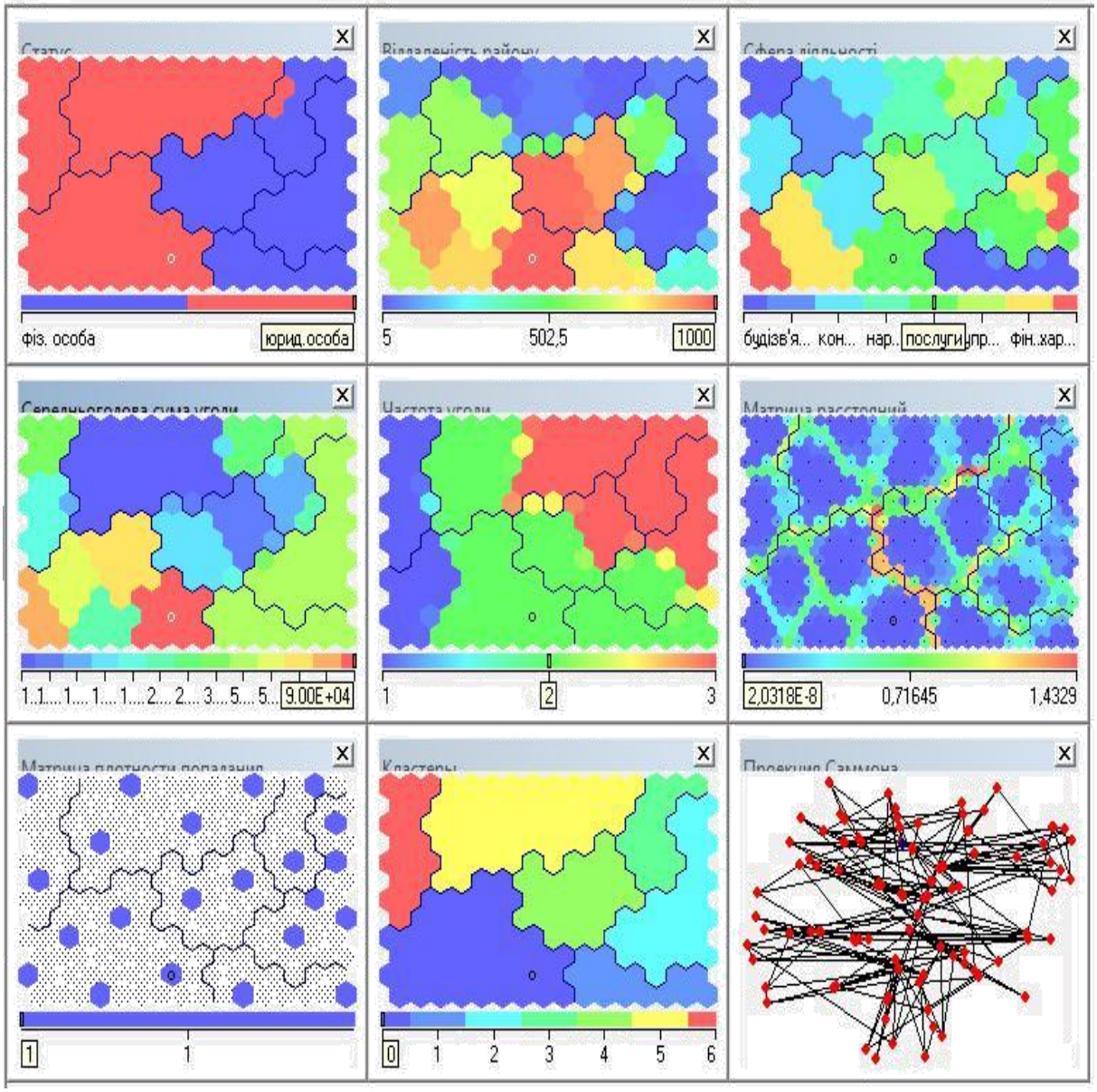


Рисунок 127– *Карти Кохонена*

Отримані кластери можна інтерпретувати в такий спосіб: максимальну середньорічну суму угоди (90000 грн.) надають фірмі юридичні особи, які працюють в сфері послуг, мешкають у радіусі 500 км і звертаються до агентства двічі на тиждень. Згідно висновку, фірмі треба рекламувати свою діяльність серед юридичних осіб, що працюють на теренах надання послуг населенню.

6 Виділити в розділі *Сценарії* гілку *Карта, що самоорганізується,* та обрати в *Майстрі обробки* пункт *Прогнозування*.

7 Обрати як спосіб відображення даних *Діаграму прогнозу*.

8 Побудувати *Діаграму прогнозу* на один тиждень, яка відображає залежність частоти укладання угод від сфери діяльності (рис.128).

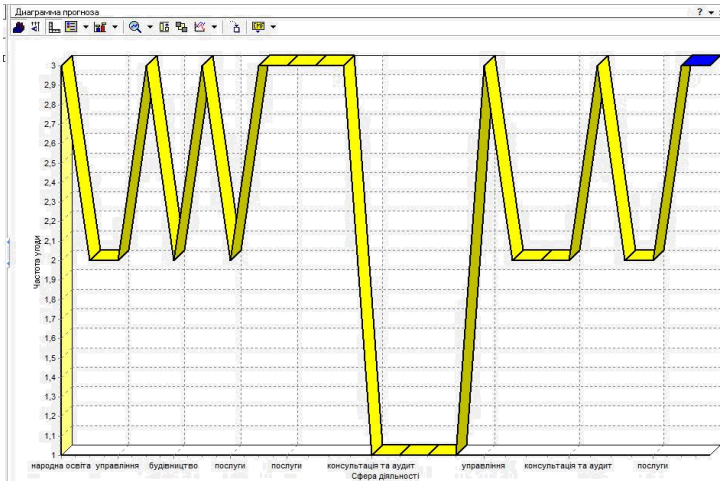


Рисунок 128 – Діаграма прогнозу залежності частоти укладання угод від сфери діяльності

Згідно з цією діаграмою наступного тижня до агентства тричі будуть звертатися працівники сфери послуг.

9 Побудувати *Діаграму прогнозу* на три тижня, для чого на 2 кроці *Майстра обробки – Прогнозування* встановити *Горизонт прогнозу* таким, що дорівнює 3.

10 Виділити в розділі *Сценарії* гілку *Карта, що самоорганізується,* та обрати в *Майстрі обробки* пункт *Дерево рішень.*

11 Налаштувати параметри вхідних (СТАТУС, ВІДДАЛЕНІСТЬ РАЙОНУ, СЕРЕДНЬОГОДОВА СУМА УГОДИ, ЧАСТОТА УГОДИ, СФЕРА ДІЯЛЬНОСТІ) та вихідного поля НОМЕР КЛАСТЕРА (рис.129).

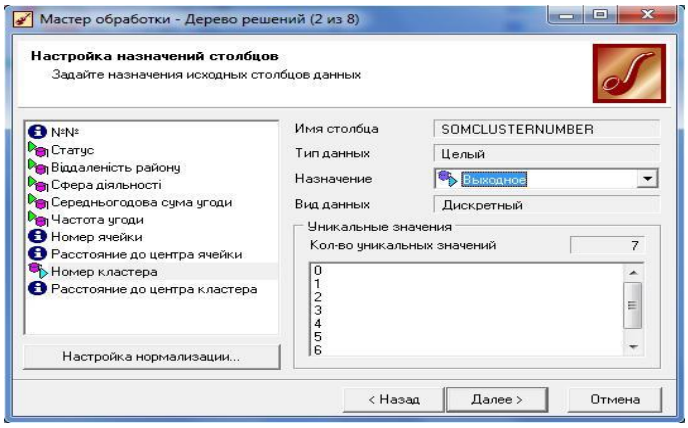


Рисунок 129– Налаштування параметрів полів при побудові *Дерева рішень* на основі *Карти, що самоорганізується*

12 Запустити процес побудови дерева рішення.

13 Обрати способи візуалізації отриманих результатів як аналіз *Що-якщо*, *Дерево рішень, Правила, Значимість атрибутів, Набір, що навчає, Таблиця спряженості.*

Результат у вигляді дерева класифікуючих правил наведено на рис.130.

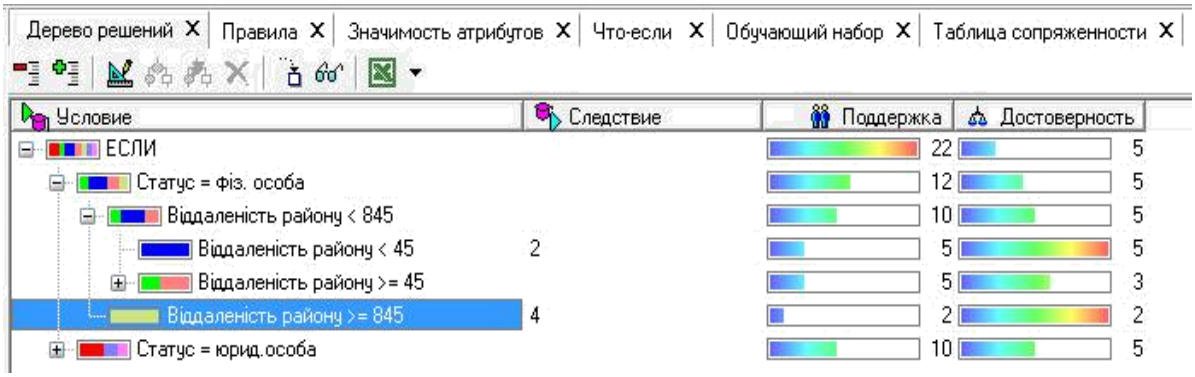


Рисунок 130 – Ієрархічне дерево класифікуючих правил

Отримана модель – це спосіб представлення правил у вигляді в ієрархії, де кожному об'єкту відповідає вузол, що дає рішення. На основі побудованої моделі можна сформулювати таке правило:

якщо статус=фіз.особа та віддаленість району >=845, то кластер 4.

14 Виділити в розділі *Сценарії* гілку *Текстовий файл* та обрати в *Майстрі обробки* пункт *Дерево рішень*.

15 Обрати як вхідні дані поля: СТАТУС, ВІДДАЛЕНІСТЬ РАЙОНУ, СЕРЕДНЬОГОДОВА СУМА УГОДИ, ЧАСТОТА УГОДИ. Поле СФЕРА ДІЯЛЬНОСТІ встановити як *Вихідне*.

16 Побудувати модель залучення фірмою нових клієнтів в залежності від СФЕРИ ДІЯЛЬНОСТІ.

17 Побудувати модель залучення фірмою нових клієнтів в залежності від СЕРЕДНЬОГОДОВОЇ СУМИ УГОДИ.

18 Сформулювати та розв’язати за допомогою технології *Data Mining* власну економічну задачу**.**

Контрольні питання

1 В чому сутність задачі сегментації клієнтів?
2 Які механізми Кластеризації Вам відомі?
3 Описати процес побудови карти Кохонена.
4 Які візуалізатори використані в лабораторній роботі?
5 Описати метод *Data Mining Прогнозування.*

Лабораторна работа 12
МЕХАНІЗМ ВІЗУАЛІЗАЦІЇ ДАНИХ

Мета роботи – використання *OLAP*-аналізу.

Завдання на виконання лабораторної роботи.

Створити та заповнити у табличному процесорі таблицю (не менше 50 записів). Ввести

– в поле ГРУПА ТОВАРУ: Аксесуари, М’які меблі, Офісні меблі, Передпокої, Журнальні столики, Письмові столи, Стелажі, Стінки, Кухні;

- в поле СТАТУС ТОВАРУ: тест, новий, постійний;

- в поле ФІРМА-ВИРОБНИК: Natuzzi, ArredoItalia, Эквмимеблі, MERX, Матис, STALK , BUROTIME, Cilek, Eudata ZRt, INTERSPAN;

- в поле КРАЇНА-ВИРОБНИК: Італія, Україна, Росія, Туреччина, Угорщина(рис.131).

	A	B	C	D	E	F	G	H	I
1	Дата поставки	Найменування товару	Група товару	Кількість (в наявності)	Статус товару	Фірма-виробник	Країна-виробник	Прогноз за сумою (продаж), тис.грн.	Прогноз за кількістю (замовлення), шт.
2	Вер.13	декоративні подушки	Аксесуари	52	тест	Natuzzi	Італія	2,4	42
3	Сер.13	настольні світильники		14	новий	Natuzzi		14,1	9
4	Вер.13	диван	М'які меблі	6	постійний	Arredo-Italia		11,5	3
5	Сер.13	крісло		11	тест	Arredo-Italia		9,9	9
6	Лип.13	кровать		3	тест	Arredo-Italia		4,2	3
7	Сер.13	кухня		4	постійний	Arredo-Italia		12	4
8	Лип.13	комп'ютерний стіл	Офісні меблі	16	постійний	Natuzzi		24,64	14
9	Лип.13	стул для відвідувачів		68	постійний	Natuzzi		5,44	50
10	Вер.13	стул для керівників		21	постійний	Natuzzi		12,6	16
11	Сер.13	стіл для переговорів		7	постійний	Natuzzi		16,1	7

Рисунок 131 – Фрагмент початкової таблиці

Виконати аналіз та зробити аргументовані висновки, використовуючи технологію *OLAP*.

1 Загальні положення

OLAP (*Online Analytical Processing*) – технологія аналітичної обробки та аналізу інформації в режимі реального часу. Це система формування звітів, в якій дії користувача щодо зміни аргументів, призводять до перебудови звіту в режимі реального часу.

В 1993 році ідеолог реляційних баз даних Едгар Кодд ввів поняття технології *OLAP* і сформулював 12 її особливостей:

1) багатовимірне концептуальне подання даних;

2) інтуїтивне маніпулювання даними;

3) доступність і деталізація даних;

4) архітектура «клієнт-сервер» *OLAP*;

5) багатокористувацька підтримка;

6) обробка неформалізованих даних;

7) збереження результатів *OLAP* окремо від початкових даних;

8) вилучення відсутніх значень;

9) гнучкість формування звітів;

10) стандартна продуктивність звітів;

11) автоматичне налаштування фізичного рівня вилучення даних;

12) необмежене число вимірів і рівнів агрегації.

Сьогодні термін *OLAP* характеризує не тільки багатовимірний погляд на дані з боку кінцевого користувача, а й багатовимірне подання даних у

цільовій базі даних (БД). Саме з цим пов'язана поява таких самостійних термінів як «Реляційний *OLAP*» (ROLAP) і «Багатомірний *OLAP*» (MOLAP).

OLAP-сервіс є інструментом аналізу багатовимірних масивів даних. Взаємодія з *OLAP*-системою надає користувачу можливість здійснювати перегляд інформації, одержувати довільні зрізи даних і виконувати аналітичні операції деталізації, згортки, наскрізного розподілу, порівняння в часі одночасно за багатьма параметрами. Вся робота з *OLAP*-системою відбувається в термінах предметної області і дозволяє будувати статистично обґрунтовані моделі ділових ситуацій.

OLAP і *Data Mining* гармонійно доповнюють один одного. На початку процесу дослідження аналітик може використати *OLAP* для кращого розуміння характеру даних і прийняття рішення щодо подальшої технології аналізу. Використання технології *Data Mining* дозволяє виявити найбільш впливові величини та діапазони, що може бути застосовано при розробці бази даних для додатків *OLAP*.

У складі модулів системи *Deductor* є інструменти: багатовимірне сховище (*Deductor Warehouse*) та механізм візуалізації (*OLAP*-куб).

У сховищі даних інформація зберігається у вимірах і процесах. *Вимір* – це об'єкт аналізу, який характеризується притаманними тільки йому властивостями і своїм унікальним ідентифікатором. *Процес* можна розглядати як зірку, у центрі якої зберігаються факти, а промені є вимірами. Сховище даних буде вміщувати наступну інформацію про клієнтів: назва, вид (юридична або фізична особа), сфера діяльності, географічний регіон розташування клієнта, тип клієнта за класифікацією ABC, тип клієнта за класифікацією XYZ, потенційний або реальний клієнт.

Процес відображає угоду з клієнтом, тобто продажу клієнту деякого товару. Фактом процесу буде сума операції і кількість товару, що купується, вимірами – клієнт, менеджер організації, що курирує операцію, дата здійснення операції, стан угоди (відкрита, відмова, успіх), причина відмови у разі неуспішної угоди, джерело інформації про здобуття товару, що купується, товар.

Завдяки аналітичній звітності дані зі сховища подаються у зручному для подальшого аналізу вигляді. Найбільш зручним інструментом для отримання аналітичної звітності є *OLAP*-куби. *OLAP* дає можливість в реальному часі генерувати описові та порівняльні відомості щодо даних і отримувати відповіді на аналітичні запити. *OLAP*-куби є проекціями куба початкових даних на куб даних меншої розмірності. При цьому значення клітин об'єднуються. Такі проекції або зрізи початкового куба подаються на площині у вигляді *кросс-таблиці.*

В аналітичній платформі *Deductor OLAP*-куб – це вбудований візуалізатор, який виконує багатовимірні операції: довільне розміщення вимірів і фактів, фільтрація, сортування, групування, різні способи агрегації і деталізації. Отримані результати відображаються у вигляді *кросс-таблиць* і *кросс-діаграм.*

За допомогою *OLAP*-куба можна проводити розвідувальний та порівняльний аналіз, виявляти тенденції та сезонність, кращі й гірші товарні позиції, розраховувати їх частки в продажах [38].

У *Deductor* в *OLAP*-кубі за допомогою інструменту *Селектор* можна агрегувати факти з якого-небудь виміру, залишаючи тільки ті об'єкти, які відповідають вказаній умові. Наприклад, можна об'єднати суму угод за клієнтами, залишивши тільки тих, які в сумі приносить 50% прибутку.

2 Порядок виконання лабораторної роботи

1 Створити та заповнити таблицю в табличному процесорі (*MS Excel, Calc)* (рис.131).

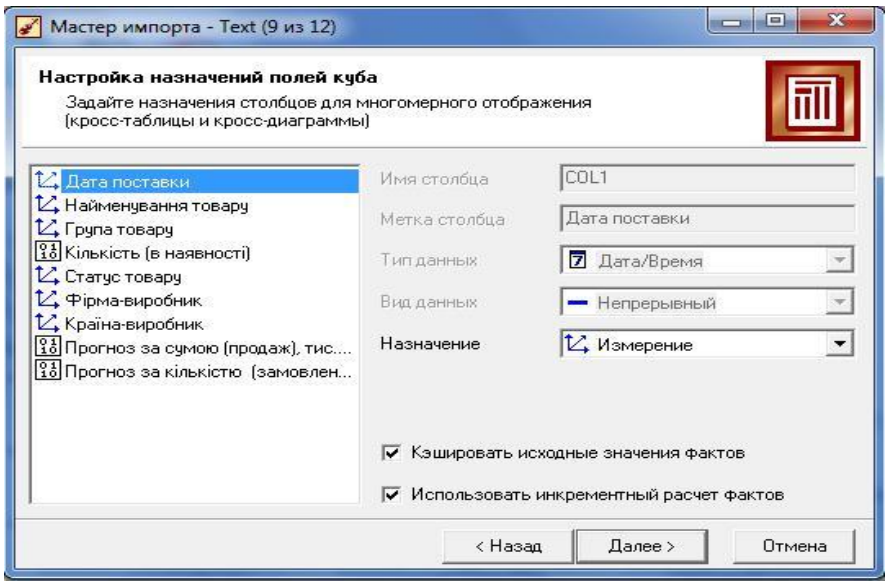


Рисунок 132 – Налаштування призначень полів куба

2 Зберегти файл у форматі Текстовий документ (з роздільниками табуляції) з ім’ям *Прізвище.txt.*

3 Додати в *Deductori* документ *Прізвище.txt* до списку сценаріїв.

4 На 6 кроці *Майстра імпорту* задати для всіх полів призначення – *Інформаційне*. В стовпчику *Дата поставки* записи повинні мати формати Дата/час, значення в стовпчиках НАЗВА ТОВАРУ, ГРУПА ТОВАРУ, СТАТУС ТОВАРУ та ФІРМА-ВИРОБНИК – текстовий формат, а в стовпчиках КІЛЬКІСТЬ (В НАЯВНОСТІ), ПРОГНОЗ ЗА СУМОЮ (ПРОДАЖ*)* та ПРОГНОЗ ЗА КІЛЬКІСТЮ (ЗАМОВЛЕННЯ) – числовий.

Для коректного *OLAP*-аналізу заповнити таблицю не менш ніж 50 записами.

5 Обрати *OLAP*-аналіз в списку способів відображення даних.

6 Залишити на 9 кроці *Майстра імпорту* налаштування призначень полів куба за замовчуванням (рис.132).

7 Ввести на кроці НАЛАШТУВАННЯ ВИМІРІВ в список *Колонки* ДАТА ПОСТАВКИ, а в список *Рядки* – ГРУПА ТОВАРУ (рис.133).

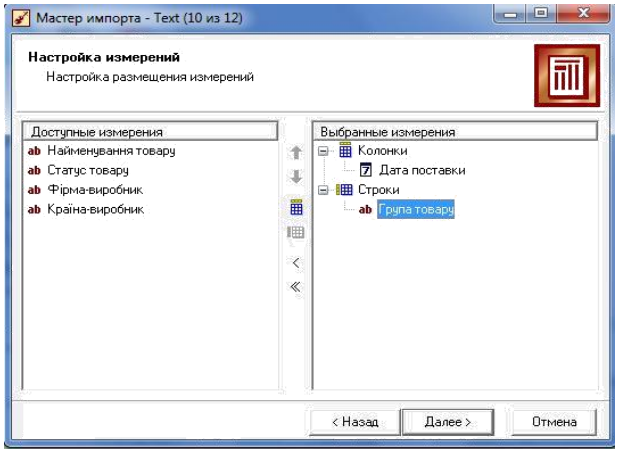


Рисунок 133 – Налаштування вимірів

8 Обрати на кроці НАЛАШТУВАННЯ ФАКТІВ *Прогноз за сумою (продаж)* варіант агрегації *Сума* (рис.134).

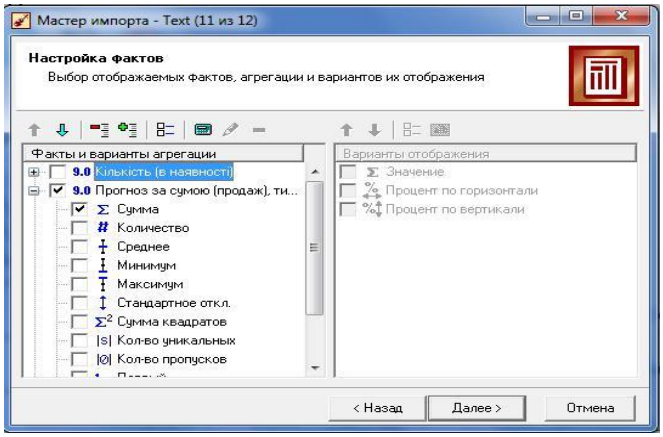


Рисунок 134 – Налаштування фактів

9 Завершити роботу *Майстра імпорту.* Створена *кросс-таблиця* наведена на рис 135.

Найменування товару	Статус товару	Фірма-виробник		
	Дата поставки			
Група товару	13.07.2013	13.08.2013	13.09.2013	Итого:
<...>	215,40	277,59	206,64	699,63
Аксесуари	2,70		4,55	7,25
Журнальні столики	40,50			40,50
Кухні		10,00	7,50	17,50
М'які меблі	10,00		25,50	35,50
Офісні меблі	24,64		21,56	46,20
Передпокої	7,00	7,00	10,50	24,50
Письмові столи	8,40	7,20	8,40	24,00
Стінки	5,80		5,80	11,60
Стелажі	7,20	7,20		14,40
Итого:	321,64	308,99	290,45	921,08

Рисунок 135 Отримана *кросс-таблиця*

10 Виконати маніпулювання вимірами при роботі з кубом. Натиснути на піктограмі *Налаштування розміщення вимірів*  (рис.136), в результаті чого відкриється діалогове вікно НАЛАШТУВАННЯ ВИМІРІВ (рис.133), в якому виконати наступні дії:

– перемістити всі вибрані виміри в область ДОСТУПНІ ВИМІРИ;

– перемістити ГРУПА ТОВАРУ до списку *Колонки*, а ФІРМА-ВИРОБНИК – до списку *Рядки*.

Змінена *кросс-таблиця* наведена на рис. 136 при *Прогнозі за сумою (продаж)* – варіант агрегації *Сума.*

Таблица X Куб X

Дата поставки | Найменування товару | Статус товару | COL7

Група товару

Фірма-виробник	<...>	Аксесуари	Журнальні	Кухні	М'які меблі	Офісні мебл	Передпокої	Письмові ст	Стінки	Стелажі	Итого:
Arredo-Italia	52,91				11,50		7,00				71,41
BUROTIME	56,76			10,00			10,50				77,26
Cilek	98,80							8,40		7,20	114,40
Eudata ZRt	104,27	2,70					7,00		5,80		119,77
INTERSPAN	57,17	2,15		7,50	10,00						76,82
MERX	89,20								5,80	7,20	102,20
Natuzzi	53,19	2,40				24,64					80,23
STALK	54,60					21,56		8,40			84,56
Матисс	89,68		19,80		14,00						123,48
Экми-мебель	43,05		20,70					7,20			70,95
Итого:	699,63	7,25	40,50	17,50	35,50	46,20	24,50	24,00	11,60	14,40	921,08

Рисунок 136 – Змінена *кросс-таблиця*

11Виконати функцію *Транспонування*, яка змінює місцями рядки та колонки. Для виконання цієї функції необхідно натиснути на піктограму  (рис.136). Результат *Транспонування* наведено на рис.137.

Дата поставки | Найменування товару | Статус товару | Країна-виробник

Фірма-виробник

Група товару	Arredo-Italia	BUROTIME	Cilek	Eudata ZRt	INTERSPAN	MERX	Natuzzi	STALK	Матис	Экми-меблі	Итого:
<...>	52,91	56,76	98,80	104,27	57,17	89,20	53,19	54,60	89,68	43,05	699,63
Аксесуари				2,70	2,15		2,40				7,25
Журнальні столики									19,80	20,70	40,50
Кухні		10,00			7,50						17,50
М'які меблі	11,50				10,00				14,00		35,50
Офісні меблі							24,64	21,56			46,20
Передпокої	7,00	10,50		7,00							24,50
Письмові столи			8,40					8,40		7,20	24,00
Стінки				5,80		5,80					11,60
Стелажі			7,20			7,20					14,40
Итого:	71,41	77,26	114,40	119,77	76,82	102,20	80,23	84,56	123,48	70,95	921,08

Рисунок 137 – Результат *Транспонування*

12Виконати маніпулювання вимірами при роботі з кубом за допомогою піктограми *Налаштування розміщення вимірів*  . Перемістити СТАТУС ТОВАРУ до списку *Колонки*, а ФІРМА-ВИРОБНИК – до списку *Рядки*. Трансформована *кросс-таблиця* наведена на рис.138.

Дата поставки	Найменування товару	Група товару	Країна-виробник
	Статус товару		

Фірма-виробник	<...>	новий	постійний	тест	Итого:
Arredo-Italia			50,31	21,10	71,41
BUROTIME		10,50	66,76		77,26
Cilek		38,70	75,70		114,40
Eudata ZRt		16,60	96,27	6,90	119,77
INTERSPAN	12,24	15,95	24,23	24,40	76,82
MERX		5,80	79,60	16,80	102,20
Natuzzi		14,10	58,78	7,35	80,23
STALK		21,56	42,00	21,00	84,56
Матис		38,70	84,78		123,48
Экми-меблі		7,20	63,75		70,95
Итого:	12,24	169,11	642,18	97,55	921,08

Рисунок 138 – Трансформована кросс-таблиця

13 Натиснути піктограму *Управління конфігураціями*  і зберегти поточну конфігурацію як *Конфігурація №1* за допомогою команди *Зберегти конфігурацію…*, яка обирається зі списку, що розкривається (рис.139).

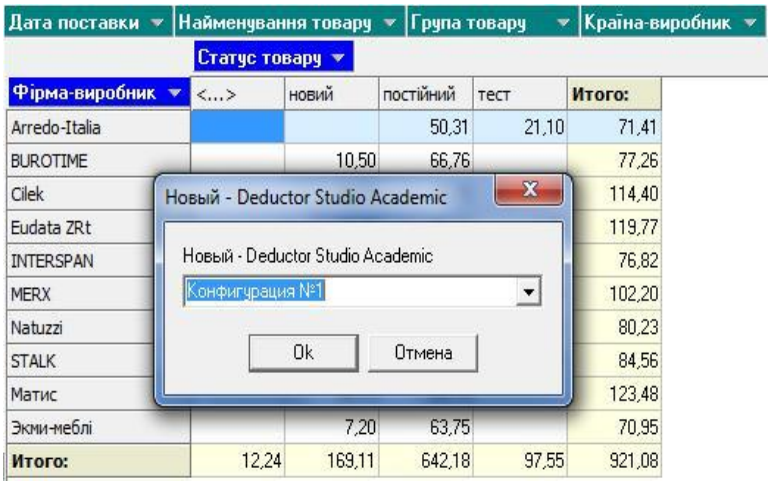


Рисунок 139 – Вікно *Управління конфігураціями*

Аналогічно, виконавши будь-які зміни налаштувань – додавши або змінивши списки рядків та колонок, можна зберегти *Конфігурацію № 2* (рис.140).

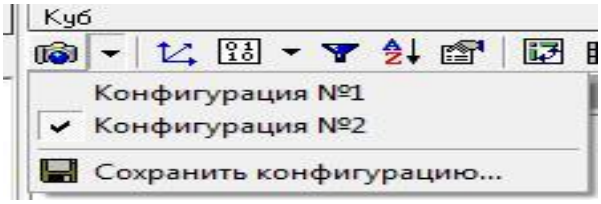


Рисунок 140 – *Конфігурація №2*

14 Натиснути в списку, що розкривається, піктограму *Налаштування фактів* (рис.136), яка має два пункти: НАЛАШТУВАННЯ ФАКТІВ… та ДОДАТИ ФАКТ, ЩО ОБЧИСЛЮЄТЬСЯ…

Обрати пункт НАЛАШТУВАННЯ ФАКТІВ…, в діалоговому вікні якого змінити факт *Прогноз за сумою (продаж)* варіант агрегації *Сума* на *Прогноз за кількістю* варіант агрегації *Кількість*.

15 Побудувати нову *кросс-таблицю* та зберегти утворену конфігурацію (рис.141).

Дата поставки	Найменування товару	Група товару	Країна-виробник
Статус товару			

Фірма-виробник	<...>	новий	постійний	тест	Итого:
Arredo-Italia			4	3	7
BUROTIME		1	7		8
Cilek		3	5		8
Eudata ZRt		2	7	2	11
INTERSPAN	1	2	4	2	9
MERX		1	6	1	8
Natuzzi		1	4	2	7
STALK		1	4	2	7
Матис		2	6		8
Экми-меблі		1	6		7
Итого:	1	14	53	12	80

Рисунок 141 – Побудована *кросс-таблиця*

16 Виконати фільтрацію, для чого послідовно виконати наступні дії:

– натиснути піктограму *Селектор* , після чого відкриється діалогове вікно (рис.142);

– обрати в діалоговому вікні СЕЛЕКТОР зі списку *Вимір та факти* пункт ФАКТИ;

– обрати в діалоговому вікні СЕЛЕКТОР зі списку *Вимір* пункт ГРУПА ТОВАРА*;*

– встановити в діалоговому вікні СЕЛЕКТОР зі списку ФАКТИ ТА ВАРІАНТИ АГРЕГАЦІЇ *Прогноз за кількістю (замовлення)* варіант агрегації *Кількість*;

– обрати в списку УМОВА пункт МЕНЬШЕ, а в поле ЗНАЧЕННЯ ввести число 5 (рис.143).

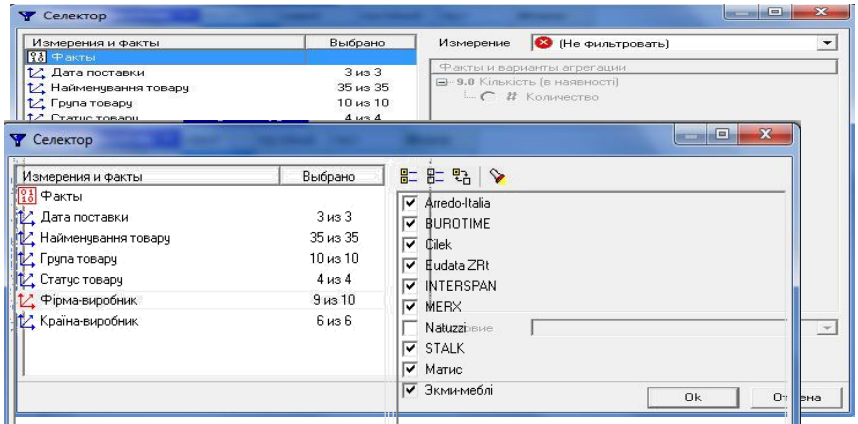


Рисунок 142 – Діалогове вікно СЕЛЕКТОР

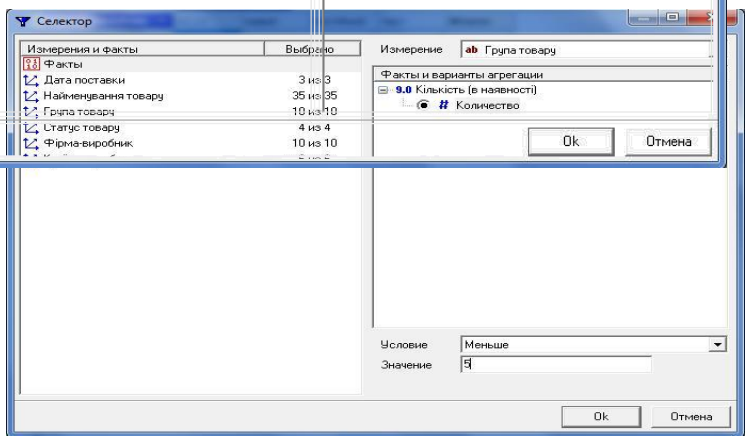


Рисунок 143 – Ввід даних в діалоговому вікні СЕЛЕКТОР

Побудувана *кросс-таблиця* наведена на рис. 144.

Дата поставки	Найменування товару	Група товару		
	Статус товару			
Фірма-виробник	новий	постійний	тест	Итого:
Arredo-Italia		1	1	2
BUROTIME	1	1		2
Cilek	2			2
Eudata ZRt	1	1	1	3
INTERSPAN	1	1	1	3
MERX	1	1		2
Natuzzi		1	1	2
STALK	1		1	2
Матис	1	1		2
Экми-меблі	1	1		2
Итого:	9	8	5	22

Рисунок 144 – *Кросс-таблиця*, побудована з використанням фільтрації

17Відкрити діалогове вікно СЕЛЕКТОР, зі списку *Вимір та факти* обрати пункт ФІРМА-ВИРОБНИК та зняти галочку біля фірми Natuzzi (рис.145).

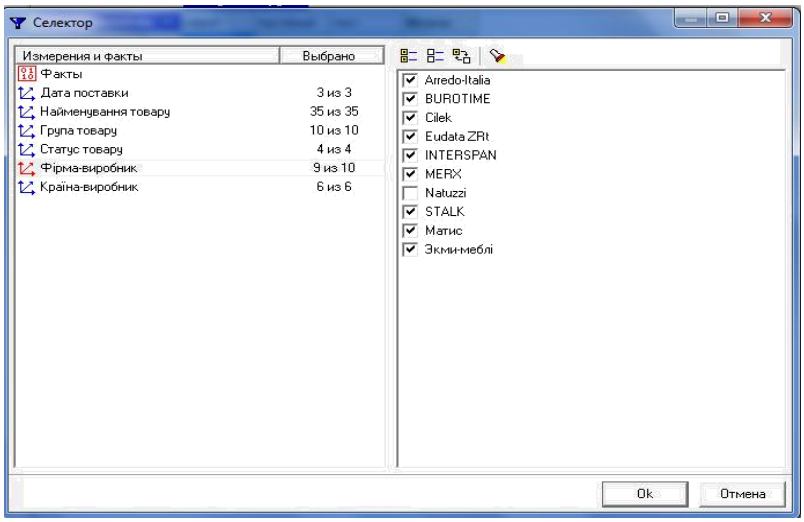


Рисунок 145 – Виключення з *кросс-таблиці* визначенної фірми

18 За зміненими даними побудувати *кросс-таблицю* та *кросс-діаграму* (рис.146).

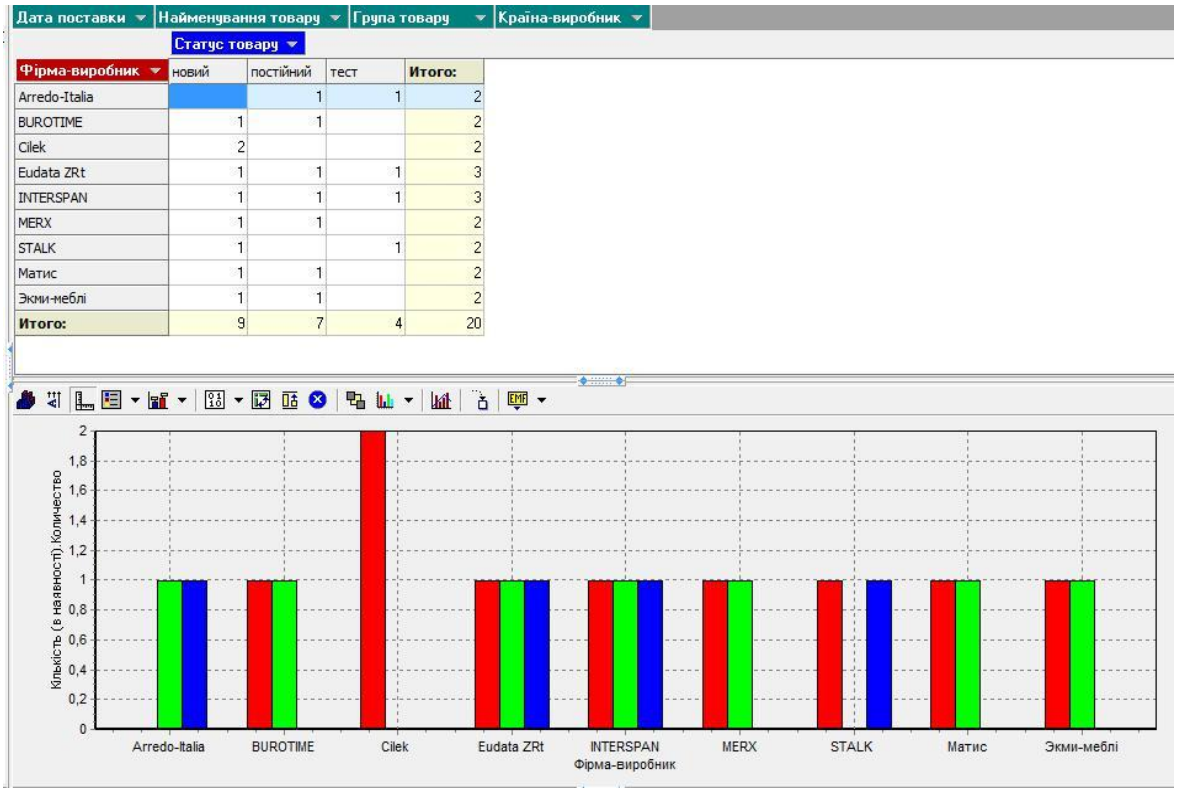


Рисунок 146 – *Кросс-таблиця* та *кросс-діаграма*

19 Змінити тип *кросс-діаграми* на об’ємний та додати на *кросс-діаграму* мітки в вигляді значень.

20 Ознайомитися з іншими функціями маніпулювання даними.

Контрольні питання

1 Дати визначення *OLAP*- технології.

2 Перерахувати особливості *OLAP*- технології.

3 Які багатовимірні операції виконує *OLAP*-куб в аналітичній платформі *Deductor*?

Лабораторна робота 13
ПЕРЕТВОРЕННЯ, ФІЛЬТРАЦІЯ ТА ВІЗУАЛЬНЕ ПОДАННЯ ДАНИХ

Мета роботи – перетворення і фільтрації даних в базі даних щодо ризиків кредитування фізичних осіб.

Завдання на виконання лабораторної роботи.

Реалізувати розбиття на групи і фільтрацію табличних даних і дослідити їх взаємовплив за допомогою візуального подання *Куб* (рис.147).

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	*Прізвище*	*Адреса*	*Стать*	*Вік*	*Розмір позики*	*Строк позики*	*Мета позики*	*Дата кредитування*	*С/м дохід*	*С/м витрати*	*Наявність нерухом*	*Наявність авто*	*Наявність банк/рахунку*	*Наявність страховки*	*Стаж роботи*	*Давати кредит*
2	Мищенко	Харків	ч	50	12000	12	лікування	12.05.2012	6000	3000	так	так	так	так	30	так
3	Бурейко	Чугуїв	ж	38	20000	6	побутова/тех	04.08.2011	7000	2000	так	ні	так	ні	20	так
4	Плющ	Харків	ч	54	6500	6	будівництво	08.11.2010	1000	500	ні	так	ні	ні	37	ні
5	Савченко	Зміїв	ч	40	7200	6	побутова/тех	23.09.2011	1200	500	так	так	ні	ні	22	так
6	Левченко	Чугуїв	ж	34	23000	12	авто	26.05.2012	10000	5000	так	ні	так	так	12	так
7	Шевченко	Зміїв	ч	38	7800	6	туристична поїздка	11.01.2013	1000	500	ні	так	так	ні	16	ні
8	Сахно	Зміїв	ч	49	6000	6	будівництво	06.04.2013	2000	500	так	ні	ні	ні	32	так
9	Новицький	Київ	ч	50	20000	12	авто	08.04.2012	11000	600	так	так	так	так	25	так
10	Зубко	Харків	ж	27	12000	6	будівництво	05.05.2012	2000	900	ні	так	ні	ні	3	так
11	Новіков	Суми	ч	42	5000	9	побутова/тех	08.12.2012	9200	500	так	так	так	так	12	так
12	Браташ	Полтава	ч	36	17000	9	побутова/тех	19.03.2012	7000	800	так	так	так	так	14	так
13	Сердюченк	Одеса	ч	52	65000	9	лікування	20.12.2010	5000	900	так	ні	так	ні	20	ні
14	Кузьмич	Львів	ч	43	25000	6	побутова/тех	06.02.2012	6000	700	так	ні	так	ні	13	так
15	Іванова	Черкаси	ж	28	45000	12	будівництво	04.06.2012	3000	1200	ні	ні	ні	ні	2	ні
16	Маслюков	Чернівці	ч	35	32000	6	туристична поїздка	18.08.2012	2200	450	ні	ні	ні	ні	12	ні
17	Токарєв	Ровни	ч	48	30000	12	будівництво	07.11.2011	5400	570	ні	так	так	так	2	так
18	Гоцка	Ялта	ж	24	12000	9	авто	29.03.2012	6000	900	так	так	так	так	2	так
19	Лізунова	Івано-Франківськ	ж	35	14000	9	авто	05.03.2012	7000	1400	так	так	так	так	10	так
20	Сорокін	Чоп	ч	29	14700	12	будівництво	27.05.2011	5000	3000	ні	так	так	так	9	так
21	Конєва	Уманськ	ж	46	25300	9	побутова/тех	24.08.2012	1000	500	ні	ні	ні	так	16	ні
22	Копиця	Краматорськ	ч	43	15000	12	відпочинок	07.08.2010	3000	500	ні	ні	ні	так	20	так
23	Олешко	Армянськ	ж	51	60000	30	лікування	14.11.2010	4000	500	ні	так	ні	ні	27	так
24	Фіріщак	Харків	ч	40	25000	25	авто	07.10.2012	6000	4000	ні	так	так	так	20	так
25	Павленко	Київ	ч	30	31500	31	відпочинок	29.06.2012	7000	4000	ні	так	так	так	10	так
26	Науменко	Івано-Франківськ	ч	32	32000	12	побутова/тех	09.06.2011	4000	2000	так	ні	так	так	11	ні
27	Ткач	Суми	ч	31	15000	15	будівництво	09.07.2012	3000	1500	так	ні	ні	ні	9	так
28	Шапіро	Армянськ	ч	27	12000	12	відпочинок	16.12.2012	1420	500	так	ні	ні	ні	8	так

Рисунок 147– Початкові табличні дані

1 Загальні положення

За допомогою інструментів передобробки даних у *Deductor Studio* можна досягти рішення ряду проміжних аналітичних завдань очищення даних. Так, при проведенні аналізу або побудові прогнозних моделей часто доводиться розбивати початкові дані на групи за різними критеріями.

У першому випадку аналітика цікавить не узагальнена інформація за всією сукупністю даних, а тенденції, що мають місце у певних групах

(наприклад, яку суму кредиту беруть позичальники на певні цілі, або до якої вікової категорії вони належать).

В другому випадку (при прогнозуванні) аналітику необхідно враховувати той факт, що різні групи (наприклад, категорії позичальників) ведуть себе по-різному, і що узагальнена модель прогнозу не буде враховувати специфіки поведінки в групах. Виходячи з цього, краще будувати кілька прогнозних моделей, наприклад, в залежності від розміру суми кредиту і будувати прогнози окремо для кожної з груп. *Deductor Studio* має широкий набір інструментів, які дозволяють розбивати початкові дані на групи та групувати дані за певними показниками.

Часто аналітику необхідно трансформувати безперервні дані (наприклад, кількість продажів) у якийсь кінцевий набір інтервалів. Наприклад, всю сукупність даних про обсяги продажів необхідно розбити на 5 інтервалів: від 0 до 100, від 100 до 200 і т.д., і віднести кожний запис початкового набору до якогось конкретного інтервалу з подальшим аналізом або фільтрацією, саме цих інтервалів.

Для виконання зазначеної операції в *Deductor Studio* застосовується інструмент *квантування* (або дискретизації). Квантування призначене для перетворення безперервних даних в дискретні. Перетворення даних може проходити як у інтервали (дані розбиваються на задану кількість інтервалів однакової довжини), так і у квантилі (дані розбиваються на інтервали різної довжини так, щоб у кожному інтервалі знаходилася однакова кількість даних). Значенням результуючого набору даних може виступати номер інтервалу, його нижня або верхня межа, середина або мітка (значення визначаються аналітиком).

Квантування дозволяє аналітику оцінити кредиторську активність представників різних вікових груп з метою прийняття рішення про стимулювання позичальників в групах з низькою активністю (наприклад, шляхом зменшення вартості кредиту для представників цих груп) і збільшення прибутку у вікових групах позичальників з високим ризиком (шляхом збільшення для них вартості кредиту) [38].

2 Порядок виконання лабораторної роботи

1 Ініціалізувати *Майстер імпорту* з метою імпортування файлу, наведеного на рис.147.

2 Встановити на 6 кроці *Майстра імпорту* призначення полів текстового файлу: РОЗМІР ПОЗИКИ – *Факт*, МЕТА ПОЗИКИ та ВІК – *Вимір*, всі інші поля – за замовчуванням.

3 Встановити способи відображення результатів імпорту як *Таблиця* та *Куб.*

4 На 9 кроці перевірити призначення полів *Майстра імпорту*: ВІК та МЕТА ПОЗИКИ – *вимір*, РОЗМІР ПОЗИКИ – *факт,* всі інші поля – *Те, що не використовується* (рис.148).

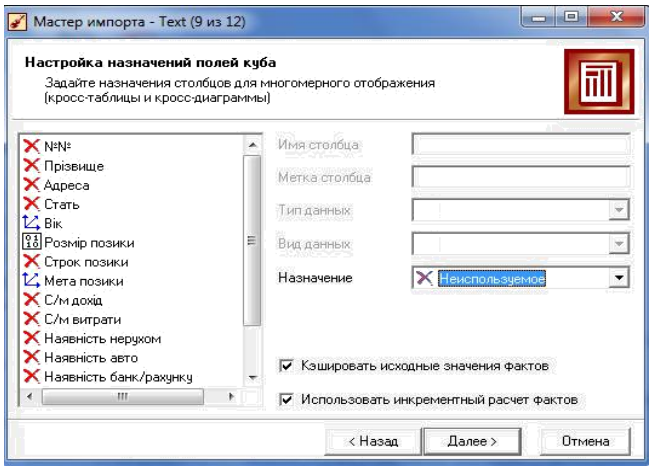


Рисунок 148– Налаштування призначень полів куба

5 На 10 кроці *Майстра імпорту* налаштування вимірів: МЕТА ПОЗИКИ – *Рядки*, ВІК – *Колонки* (рис.149).

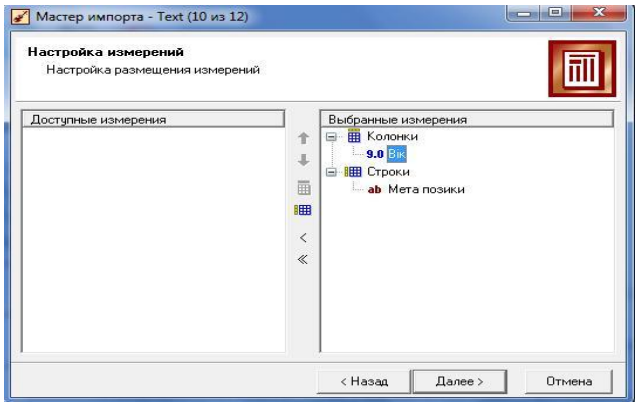


Рисунок 149– Налаштування вимірів

6 Встановити на 11 кроці *Майстра імпорту* для факту РОЗМІР ПОЗИКИ прапорець варіанту агрегації – *Сума* (рис.150).

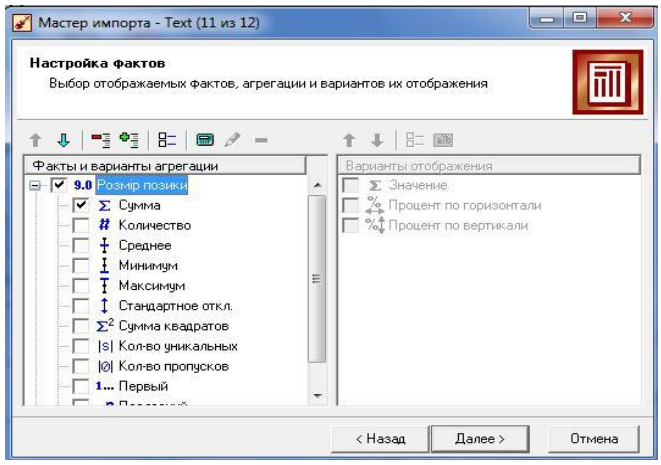


Рисунок 150 – Налаштування фактів

7 Завершити процес імпортування в вигляді *Таблиця, Статистика* та *Куб* (рис.151).

	Вік												
Мета позики	20	24	25	26	27	28	29	30	31	32	34	35	36
авто		36 000,00						3 100,00		24 500,00	23 000,00	14 000,00	
будівництво	4 200,00				12 000,00	45 000,00	14 700,00		15 000,00			23 000,00	
відпочинок		14 000,00			12 000,00			31 500,00				13 000,00	
лікування													
побутова/тех		3 400,00				65 000,00				32 000,00			17 000,00
туристична поїздка			14 000,00	13 000,00								32 000,00	12 500,00
Итого:	4 200,00	53 400,00	14 000,00	13 000,00	24 000,00	110 000,00	14 700,00	34 600,00	15 000,00	56 500,00	23 000,00	82 000,00	29 500,00

Рисунок 151– Результат імпортування в вигляді *Куб*

8Натиснути кнопку Транспонування (Ctrl+T), щоб змінити розташування вимірів (рис.152).

	Мета позики						
Вік	авто	будівництво	відпочинок	лікування	побутова/т	туристична	**Итого:**
20		4 200,00					4 200,00
24	36 000,00		14 000,00		3 400,00		53 400,00
25						14 000,00	14 000,00
26						13 000,00	13 000,00
27		12 000,00	12 000,00				24 000,00
28		45 000,00			65 000,00		110 000,00
29		14 700,00					14 700,00
30	3 100,00		31 500,00				34 600,00
31		15 000,00					15 000,00
32	24 500,00				32 000,00		56 500,00
34	23 000,00						23 000,00
35	14 000,00	23 000,00	13 000,00			32 000,00	82 000,00
36					17 000,00	12 500,00	29 500,00
38					33 600,00	7 800,00	41 400,00
40	25 000,00				7 200,00	6 500,00	38 700,00
42					5 000,00	2 300,00	7 300,00
43			15 000,00		25 000,00		40 000,00
45		6 000,00					6 000,00
46					25 300,00	12 000,00	37 300,00
48		30 000,00	15 000,00		17 600,00		62 600,00
49	31 000,00	6 000,00					37 000,00
50	20 000,00			12 000,00			32 000,00
51				64 000,00			64 000,00
52				65 000,00			65 000,00
54		6 500,00					6 500,00
56				18 000,00			18 000,00
Итого:	176 600,00	162 400,00	100 500,00	159 000,00	231 100,00	100 100,00	929 700,00

Рисунок 152– Транспоноване відображення *Куба*

9 Задати параметри фільтрації для вимірів ВІК та МЕТА ПОЗИКИ: встановити за допомогою піктограми *Селектор* МЕТУ ПОЗИКИ – будівництво та отримати результати фільтрації у вигляді таблиці (рис 153).

	Мета позики	
Вік	будівництв	Итого:
20	4 200,00	4 200,00
27	12 000,00	12 000,00
28	45 000,00	45 000,00
29	14 700,00	14 700,00
31	15 000,00	15 000,00
35	23 000,00	23 000,00
45	6 000,00	6 000,00
48	30 000,00	30 000,00
49	6 000,00	6 000,00
54	6 500,00	6 500,00
Итого:	162 400,00	162 400,00

Рисунок 153– Результати фільтрації

10 Натиснути кнопку Відображати кросс-діаграму, та розташувати її внизу таблиці (рис.154).

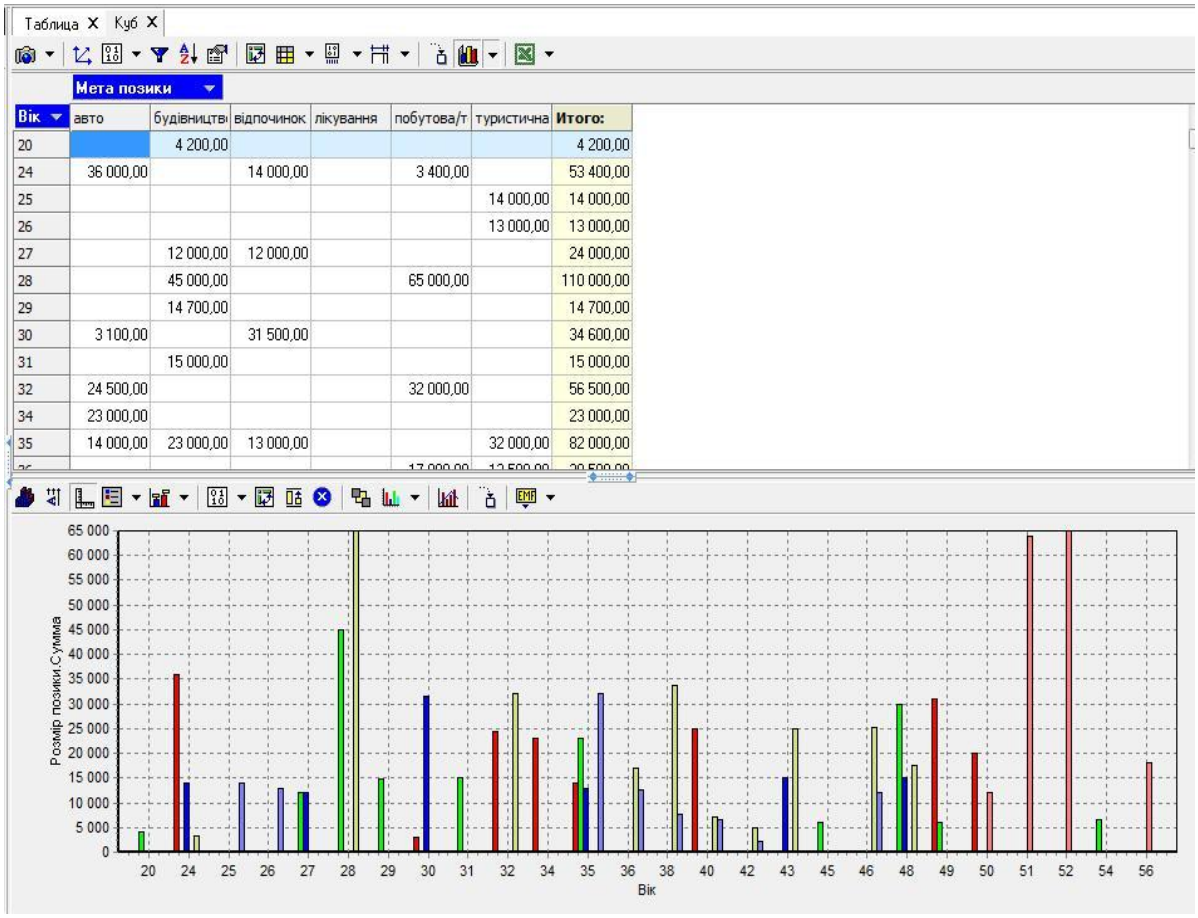


Рисунок 154 – *Кросс-діаграма*

11Змінити тип діаграми.

12 Використати інші інструментальні кнопки в режимі подання *Куба*: Показувати підсумки, Налаштування розміщення вимірів, Налаштування фактів тощо.

13 Натиснути правою кнопкою миші на *кросс-діаграмі* та в контекстному меню обрати команду *Тренд…*

14 Задати подання тренда як вейвлет-перетворення та розташувати легенду внизу діаграми (рис.155).

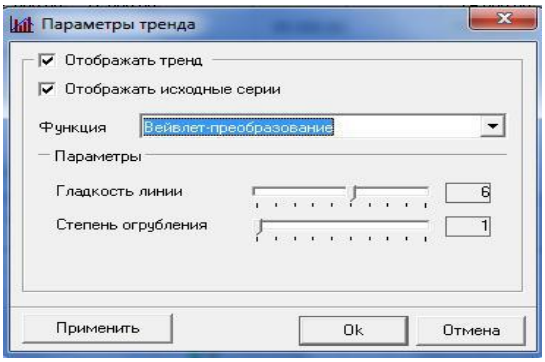


Рисунок 155– Вікно налаштування параметрів тренду

Результати обробки початкових даних з використанням багатовимірного подання (*Куб*), фільтрації і лінії тренду у вигляді вейвлет-перетворення наведено на рис. 156.

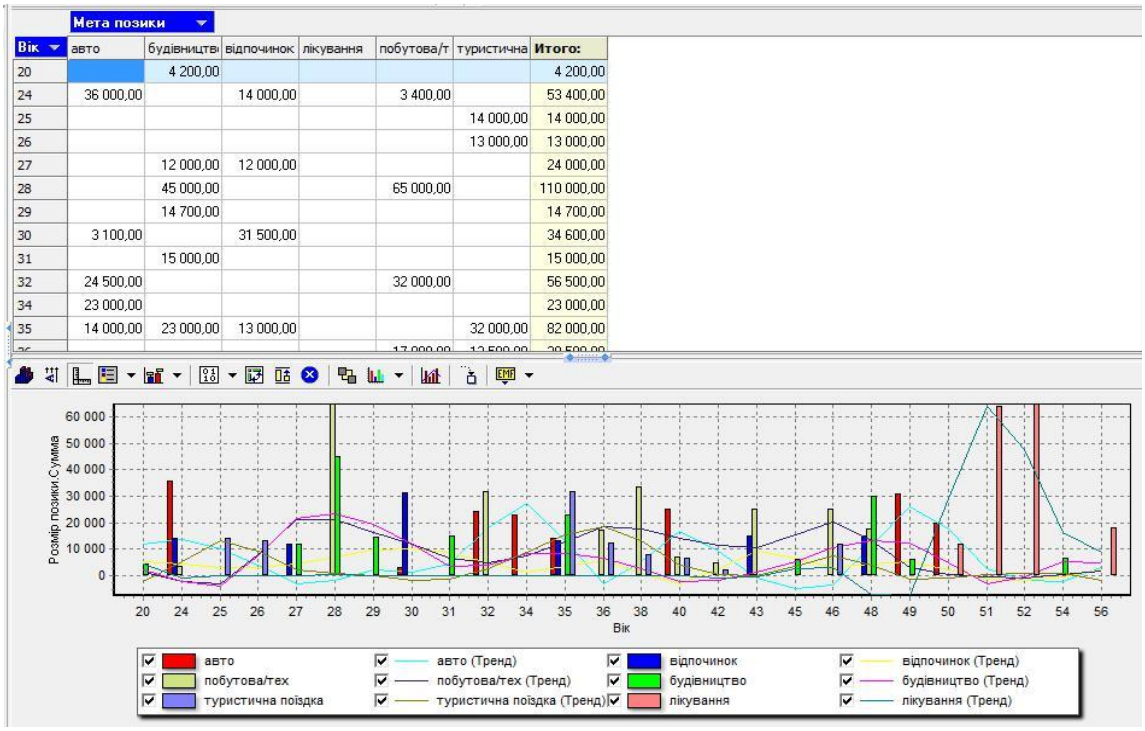
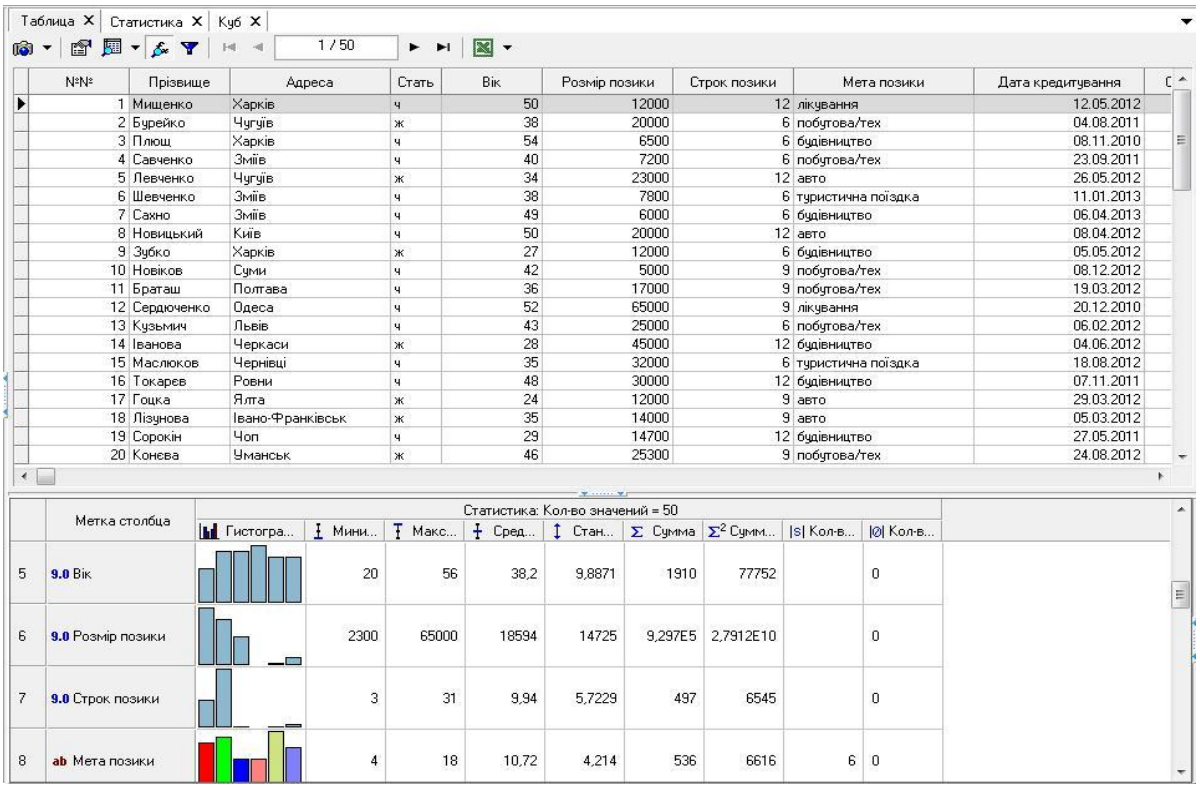


Рисунок 156 – Обробка даних з використанням багатовимірного подання (*Куб*), фільтрації і лінії тренду у вигляді вейвлет-перетворення

15 Виконати збереження конфігурації за допомогою кнопки Управління конфігураціями з ім'ям *Конфігурація №1*.

16 Виконати дії з налаштуванням інших параметрів тренда: лінійний, експоненціальний, ковзне середнє, експоненціальне ковзне середнє.

17 Перейти в режим подання *Таблиця*. Включити відображення статистики (кнопка Показати онлайн статистику) (рис.157). Виконати аналіз.



№№	Прізвище	Адреса	Стать	Вік	Розмір позики	Строк позики	Мета позики	Дата кредитування
1	Мищенко	Харків	ч	50	12000	12	лікування	12.05.2012
2	Бурейко	Чугуїв	ж	38	20000	6	побутова/тех	04.08.2011
3	Плющ	Харків	ч	54	6500	6	будівництво	08.11.2010
4	Савченко	Зміїв	ч	40	7200	6	побутова/тех	23.09.2011
5	Левченко	Чугуїв	ж	34	23000	12	авто	26.05.2012
6	Шевченко	Зміїв	ч	38	7800	6	туристична поїздка	11.01.2013
7	Сахно	Зміїв	ч	49	6000	6	будівництво	06.04.2013
8	Новицький	Київ	ч	50	20000	12	авто	08.04.2012
9	Зубко	Харків	ж	27	12000	6	будівництво	05.05.2012
10	Новіков	Суми	ч	42	5000	9	побутова/тех	08.12.2012
11	Браташ	Полтава	ч	36	17000	9	побутова/тех	19.03.2012
12	Сердюченко	Одеса	ч	52	65000	9	лікування	20.12.2010
13	Кузьмич	Львів	ч	43	25000	6	побутова/тех	06.02.2012
14	Іванова	Черкаси	ж	28	45000	12	будівництво	04.06.2012
15	Маслюков	Чернівці	ч	35	32000	6	туристична поїздка	18.08.2012
16	Токарєв	Ровни	ч	48	30000	12	будівництво	07.11.2011
17	Гоцка	Ялта	ж	24	12000	9	авто	29.03.2012
18	Лізунова	Івано-Франківськ	ж	35	14000	9	авто	05.03.2012
19	Сорокін	Чоп	ч	29	14700	12	будівництво	27.05.2011
20	Конєва	Уманськ	ж	46	25300	9	побутова/тех	24.08.2012

	Метка столбца	Статистика: Кол-во значений = 50								
		Гистогра...	Мини...	Макс...	Сред...	Стан...	Сумма	Σ² Сумм...	|s| Кол-в...	|∅| Кол-в...
5	9.0 Вік		20	56	38,2	9,8871	1910	77752		0
6	9.0 Розмір позики		2300	65000	18594	14725	9,297E5	2,7912E10		0
7	9.0 Строк позики		3	31	9,94	5,7229	497	6545		0
8	ab Мета позики		4	18	10,72	4,214	536	6616	6	0

Рисунок 157– Відображення статистики

18 Задати параметри фільтрації табличних даних: візуалізувати інформацію про кредити на будівництво чоловіків віком від 40 до 50 років (рис .158).

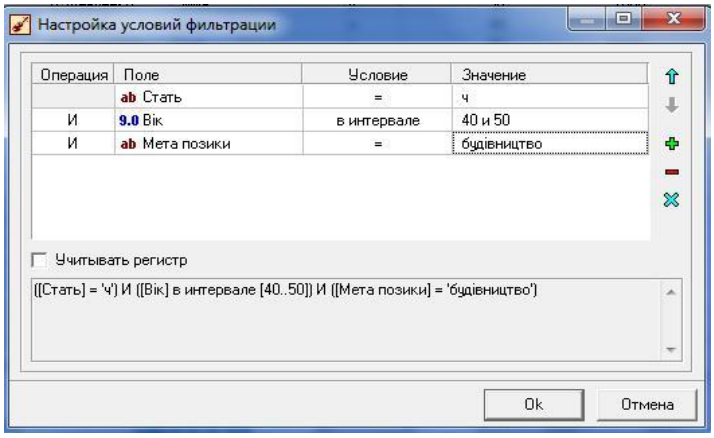
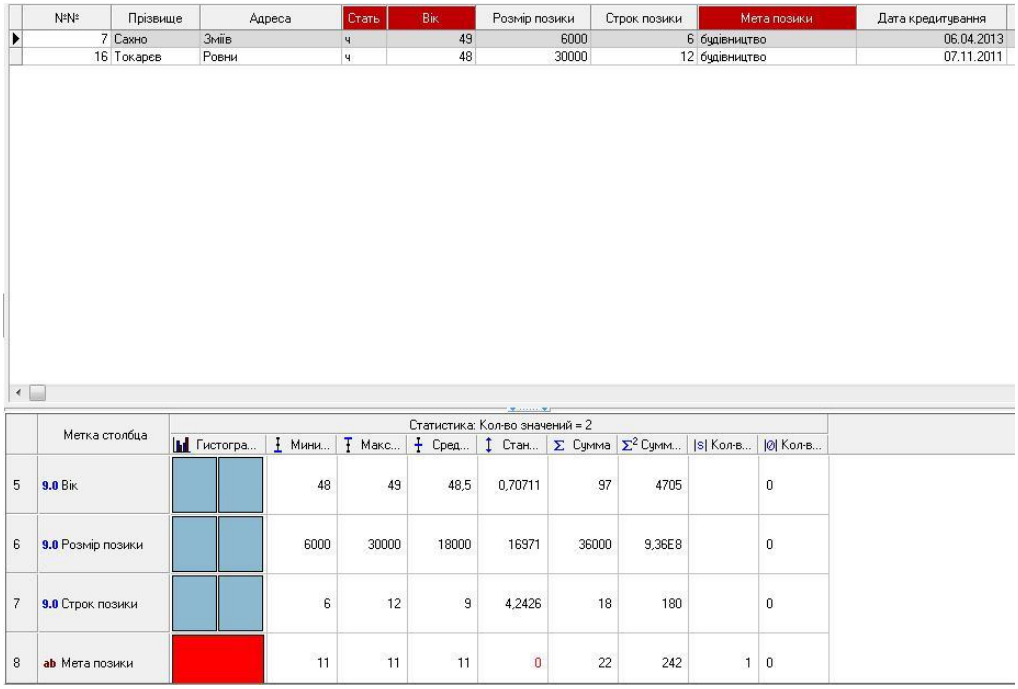


Рисунок 158 – Умова фільтрації

Результати фільтрації наведено на рис. 159.



№№	Прізвище	Адреса	Стать	Вік	Розмір позики	Строк позики	Мета позики	Дата кредитування
7	Сахно	Зміїв	ч	49	6000	6	будівництво	06.04.2013
16	Токарев	Ровни	ч	48	30000	12	будівництво	07.11.2011

Статистика: Кол-во значений = 2

	Метка столбца	Гистогра...	Мини...	Макс...	Сред...	Стан...	Сумма	Σ² Сумм...	|s| Кол-в...	|0| Кол-в...
5	9.0 Вік		48	49	48,5	0,70711	97	4705		0
6	9.0 Розмір позики		6000	30000	18000	16971	36000	9,36E8		0
7	9.0 Строк позики		6	12	9	4,2426	18	180		0
8	ab Мета позики		11	11	11	0	22	242	1	0

Рисунок 159 – Результати фільтрації

19 Зберегти результати з ім’ям *Конфігурація №2*.

20 Задати власні умов фільтрації. Проаналізувати результати.

21 Перейти в режим відображення *Куб*.

22 Задати умови фільтрації за допомогою кнопки Селектор для фактів, для МЕТИ ПОЗИКИ, для ВІКУ: відібрати тільки ті записи, які містять молодих людей до 30 років включно, що беруть кредити на туристичні поїздки (рис.160).

Вік	Мета позики: туристична	Итого:
25	14 000,00	14 000,00
26	13 000,00	13 000,00
Итого:	27 000,00	27 000,00

Рисунок 160 – Результати фільтрації

23 Виконати квантування віку позичальників за інтервалами, тобто віднести перетворити безперервні дані в дискретні.

23.1 Виділити в розділі *Сценарії* імпортований в *Deductor* текстовий файл з розподільниками табуляції, наведений на рис. 147.

23.2 Обрати в *Майстрі обробки* в групі *Трансформація* метод обробки *Квантування*.

23.3 Задати на 2 кроці *Майстра обробки* параметри квантування для поля ВІК: призначення – *Те, що використовується*, спосіб – *За інтервалами*, інтервалів –5, значення – *Мітка інтервала*, вид даних – *Дискретний* (рис.161).

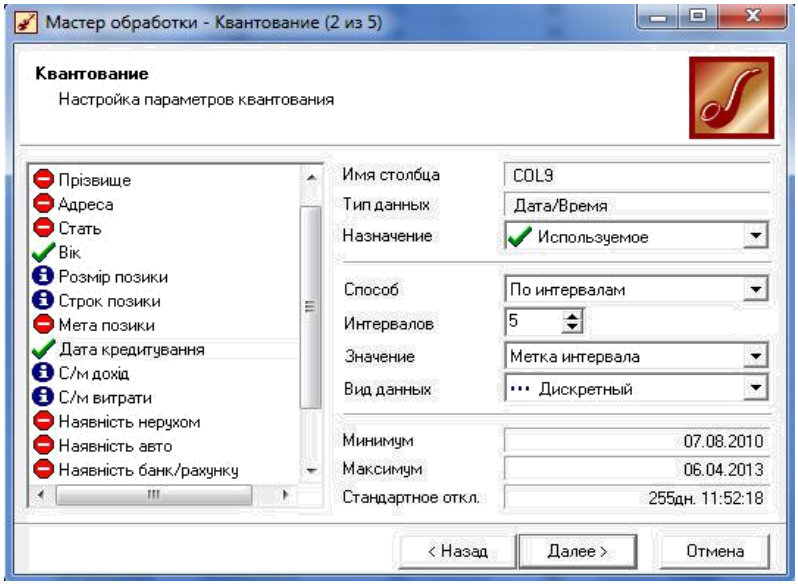


Рисунок 161– Налаштування квантування за інтервалами за віком позичальників

23.4 На 3 кроці *Майстра обробки* задати мітки та границі інтервалам в залежності від віку (рис.162).

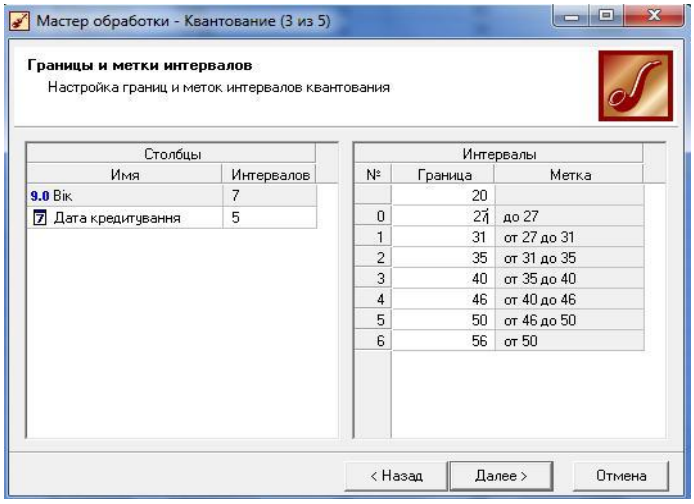


Рисунок 162 – Налаштування міток та границь інтервалів

23.5 Задати спосіб подання даних *Таблиця* та *Куб*.

23.6 На 5 кроці *Майстра обробки* вказати як *Вимір* поля ВІК та ДАТА, а як *Факти* РОЗМІР ПОЗИКИ. Інші поля вказати як *Такі, що не використовуються* (рис.163).

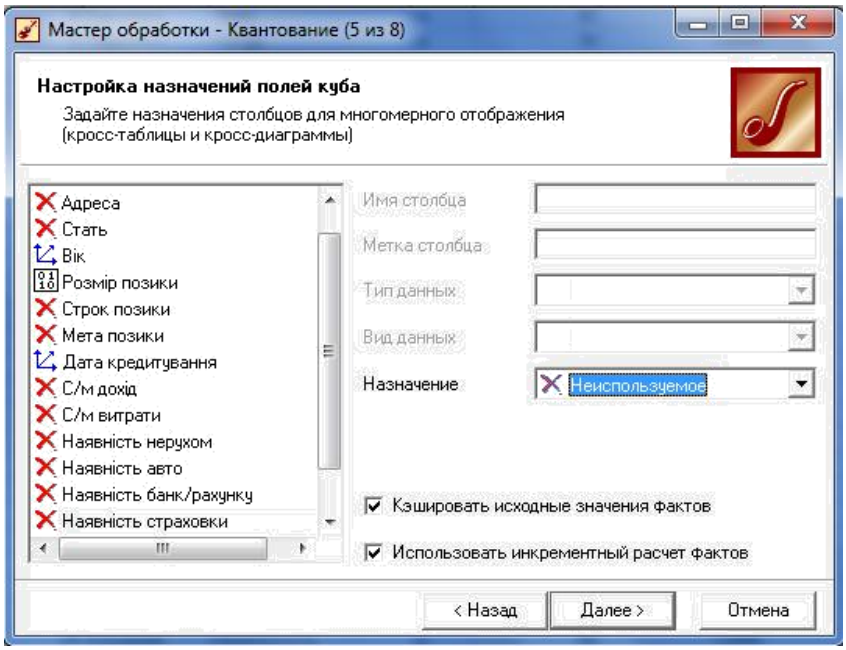


Рисунок 163– Налаштування призначень полів куба при квантуванні за віком

23.7 На кроці 6 *Майстра обробки* вказати розташуватнняполів вимірів: ДАТА – в *Колонки*, ВІК – в *Рядках*.

23.8 На 7 кроці *Майстра обробки* задати параметри налаштування фактів (рис.164).

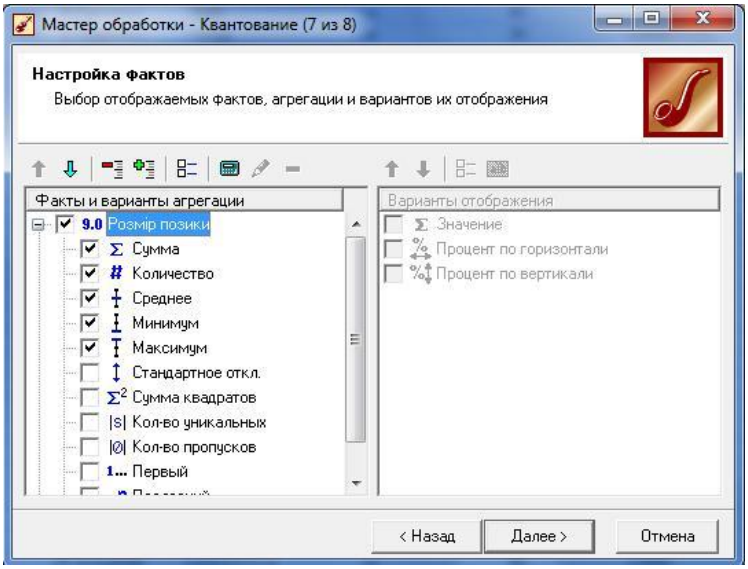


Рисунок 164–Налаштування фактів при квантуванні за віком

24 Завершити роботу *Майстра обробки.*

25 Відобразити результати у вигляді *Куб* (рис.165).

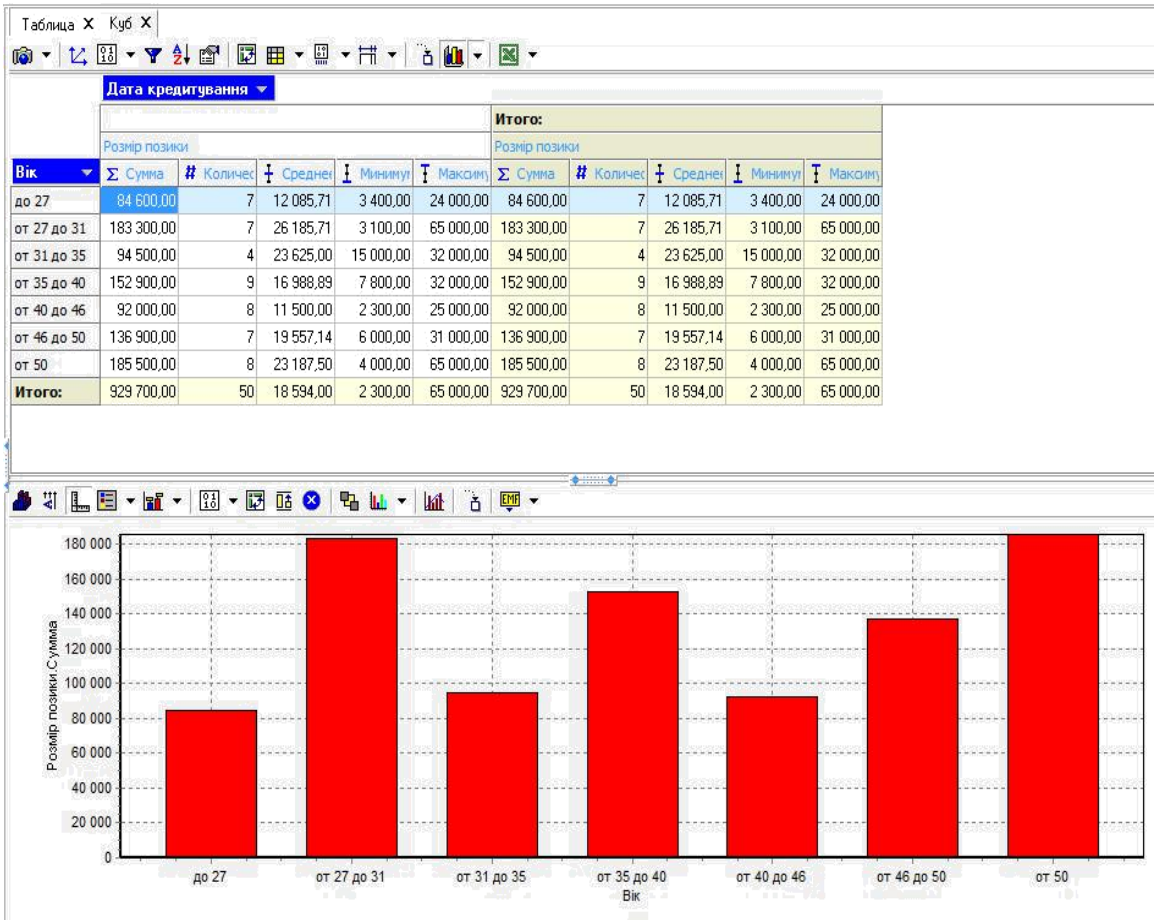


Рисунок 165 – Результати квантування за віком

26 Провести аналіз отриманих результатів.

27 Зберегти конфігурацію з ім'ям *Конфігурація №3.*

Контрольні питання

1 Які інструментів передобробки даних у *Deductor Studio* Вам відомі?

2 Для чого призначений інструмент квантування (або дискретизації)?

3 Який інструмент для отримання аналітичної звітності існує в *Deductor*?

4Що таке *кросс-діаграма* і що вона візуалізує?

Лабораторна робота 14
КЛАСТЕРНИЙ АНАЛІЗ

Мета роботи – використання кластеризації для аналізу даних

Завдання на виконання лабораторної роботи.

Провести кластеризацію банків (розбити їх на кластери) з точки зору привабливості депозитування коштів, спираючись на параметри, подані в базі даних, яка містить наступні поля: Банк, Довгострокова динаміка депозитів

населення, Відповідність капіталу активам, Ліквідність, Рівень підтримки акціонерів, Ефективність діяльності банку, Рівень боргового навантаження, Динаміка власного капіталу за півріччя, Чиста відсоткова маржа, Сумарний рейтинг (числове значення), Сумарний рейтинг (буквений код), Динаміка депозитів населення за 12 міс.,%, Рейтинг на 1.01.13 (буквений код), Рейтинг аналітиків на 15.05.12 (числове значення), Коефіцієнт системності/ проблемності (рис.166).

№	Банк	Довгострокова динаміка депозитів населення	Відповідність капіталу активам	Ліквідність	Рівень підтримки акціонерів	Ефективність діяльності банку	Рівень боргового навантаження	Динаміка власного капіталу за півріччя	Чиста відсоткова маржа	Сумарний рейтинг	Сумарний рейтинг	Динаміка депозитів населення за 12 міс.,%
1	Укрексімбанк	3	4	4	4	3	4	3	3	3,96	A	29,4
2	Ощадбанк	4	4	3	4	3	4	3	3	3,91	A	3
3	Райффайзен банк аваль	2	4	4	3	3	4	3	4	3,74	A	-4,6
4	Форум	3	4	4	4	2	4	4	3	3,55	B	3,
5	Укрсоцбанк	3	4	3	3	3	4	3	4	3,52	B	13,4
6	Правекс-банк	2	4	4	3	3	4	4	4	3,5	B	-28,6
7	Ощадбанк Росії	4	3	3	4	3	4	3	4	3,45	B	85,4
8	Приватбанк	4	3	4	1	3	4	3	3	3,41	B	48,9
9	ОТП банк	3	4	3	3	4	4	2	4	3,4	B	10
10	Укрсіббанк	3	3	4	4	2	3	2	3	3,36	B	-
11	Креді Агріколь	3	3	4	3	3	4	3	3	3,3	B	21,8
12	Альфа-банк	4	3	4	3	3	2	3	3	3,26	B	41,1
13	Кредобанк	2	4	4	4	2	4	2	3	3,25	B	-9
14	Промінвестбанк	2	3	4	4	3	4	3	3	3,17	B	-1,7
15	Сведбанк	1	4	4	3	3	3	3	4	3,15	B	-35,4
16	Унікредит	3	3	4	3	3	3	3	3	3,15	B	25,5
17	VAB банк	4	3	4	2	2	4	4	2	3,15	B	46,6
18	БТА	4	4	4	2	2	4	2	3	3,15	B	123
19	ВТБ	3	3	3	4	3	3	1	3	3,1	B	2
20	Дельта	4	2	4	2	3	3	4	4	3,1	B	92,1
21	Кредит-Дніпро	4	3	4	1	3	4	3	3	3,1	B	51
22	Універсал	2	3	4	2	2	4	4	4	3,05	B	-3,1
23	Ерсте банк	4	3	4	3	2	3	2	3	3	B	56
24	Південний	3	3	4	1	3	4	3	3	3	B	13,5

Рисунок 166 – Фрагмент бази даних

1 Загальні положення

Кластеризацію в контексті інтелектуального аналізу звичайно розуміють як поділ загальної множини на певну кількість підмножин (кластерів) за наперед невідомими ознаками, причому об'єкти всередині кожного з кластерів повинні бути близькі між собою за однією або декількома ознакам, доступним для інтерпретації.

Кластеризація є одним з ключових методів пошуку загальних закономірностей в інтелектуальному аналізі даних. По завершенні Кластеризації в кожному кластері опиняться схожі за своїми властивостями об'єкти. Які в той же час будуть відрізняться від об’єктів, розташованих в інших кластерах. При цьому, чим більш схожі об'єкти всередині кластера і чим сильніше вони відрізняються від об'єктів в інших кластерах, тим краще кластеризація.

Формальна постановка задачі кластеризації виглядає наступним чином. Нехай задана множина об'єктів $X = (x_1,x_2,...,x_n)$ і номерів (імен, міток) кластерів $Y = (y_1, y_2,...y_k)$. Для X визначена деяка функція відстані між об'єктами $D(x, x')$. Крім цього, є кінцева вибірка навчальних наборів $X_m = (x_1,x_2,...,x_m)$ з множини X, яку потрібно розбити на X_m непересічних підмножин (кластерів) так, щоб кожна з них складалася б тільки з елементів,

близьких за метрикою *D*. При цьому кожному об'єкту x_i з множини X_m привласнюється номер кластера y_j.

Іншими словами, задача полягає в пошуку функції *f*, яка будь-якому об'єкту *x* з множини *X* ставить у відповідність номер кластера *y* з множини *Y*.

Кластеризація дозволяє досягти наступного:

– покращує розуміння даних за рахунок виявлення структурних груп. Розбиття вибірки на групи схожих об'єктів дозволяє спростити подальшу обробку даних і прийняття рішень, застосовуючи до кожного кластеру свій метод аналізу;

– дозволяє компактно зберігати дані. Для цього замість зберігання всієї вибірки можна залишити по одному типовому представнику кожного кластеру;

– виявляє нові нетипові об'єкти, які не потрапили в жоден кластер.

В межах концепції інтелектуального аналізу даних було розроблено низку інструментів кластеризації. Так, у складі аналітичної платформи *Deductor Studio* компанії Basegroup Labs представлена кластеризація методами *kmeans* та *gmeans*, а також нейромережевими методами на основі карт Кохонена [34].

Одним з найбільш поширених і простих алгоритмів кластеризації є алгоритм *kmeans*. Цей алгоритм заснований на оптимізації цільової функції, що визначає оптимальне в певному сенсі розбиття множини об'єктів на кластери. В якості цільової функції використовується сума квадратів зважених відхилень координат об'єктів від центрів шуканих кластерів. Кластери шукаються сферичної або еліпсоїдної форми. Алгоритм оптимізації цільової функції носить ітеративний характер, і на кожній ітерації розраховується матриця відстаней між об'єктами. Обчислювальна складність *i*-ої ітерації алгоритму *kmeans* оцінюється як *O(kmn)*, де *k, m, n* – кількість кластерів, атрибутів і об'єктів відповідно [2, 35].

На сьогоднішній день розроблена значна кількість алгоритмів, що дозволяють здійснювати автоматичний вибір кількості кластерів, оптимальний з точки зору того чи іншого критерію. Зазвичай будується кілька моделей для різних значень *k*, а потім обирається найбільш ефективна. Одним з найбільш поширених алгоритмів цього типу є алгоритм *gmeans*, в основі якого лежить припущення про те, що дані, які кластеризуються, підпорядковуються унімодальному закону розподілу, наприклад гаусовському. Тоді центр кластера, який визначається як середнє значення ознак об'єктів, які потрапили в нього, можна розглядати як моду відповідного розподілу.

2Порядок виконання лабораторної роботи

1 Створити в табличному процесорі (*MS Excel,Calc)* базу даних (рис. 166).

2 Встановити при імпорті файлу для полів бази даних наступні призначення: для поля № – *Інформаційне*, всі інші поля – *Вхідне*.

3 Обрати як спосіб відображення даних *Таблиця* та *Статистика.*

4 Обрати в *Майстрі обробки* в групі *Data Mining* пункт *Кластеризація.*

5 Налаштувати призначення стовпців, тобто обрати властивості, за якими буде виконане групування об’єктів (рис.167).

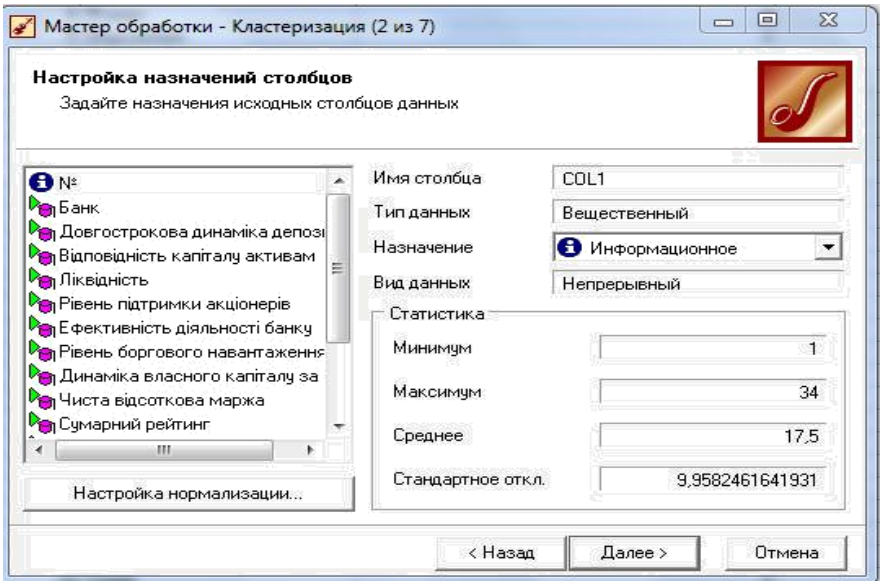


Рисунок 167 – Налаштування призначень стовпців

6 На наступному кроці *Майстра* налаштувати спосіб розподілу початкової множини даних на множину, що навчає, та тестову, а також кількість наборів в цих множинах. Дані для обох множин обрати випадковим чином, і визначити всю множину, як таку, що навчає (рис.168).

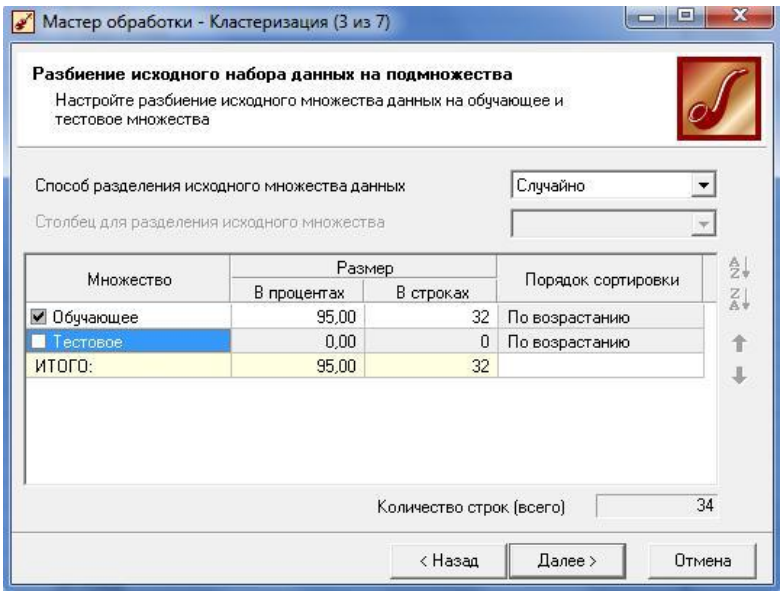


Рисунок 168– Розбиття початкового набору даних на підмножини

7 На наступному кроці *Майстра* налаштувати параметри кластеризації, визначити кількість кластерів. Банки можна поділити за рейтингом, тому обирається фіксована кількість кластерів, яка дорівнює чотирьом (рис.169).

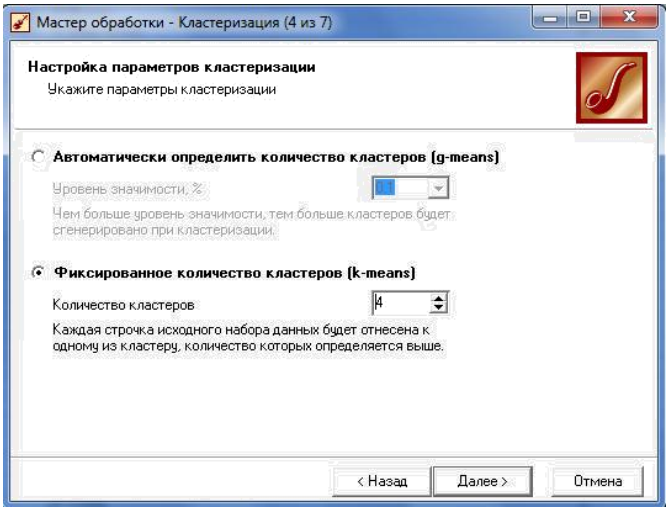


Рисунок 169– Налаштування параметрів Кластеризації

8Запустити процес Кластеризації.

9 Відібрати зі списку візуалізаторів такі способи відображення даних: аналіз *Що-якщо* для вирішення задачі класифікації віднесення банку до одного з кластерів, *Профілі кластерів* для визначення структури формування групи кластерів, *Набір,що навчає, Діаграма розміщення* та *Куб* (*OLAP*–аналіз) для наочного перегляду отриманих результатів.

10 Налаштувати для візуалізатора *Куб* призначення полів. Властивості, які розглядаються, обрати як *Факти*, а БАНК, СУМАРНИЙ РЕЙТИНГ та РЕЙТИНГ НА 1.01.13 як *Вимір* (рис.170).

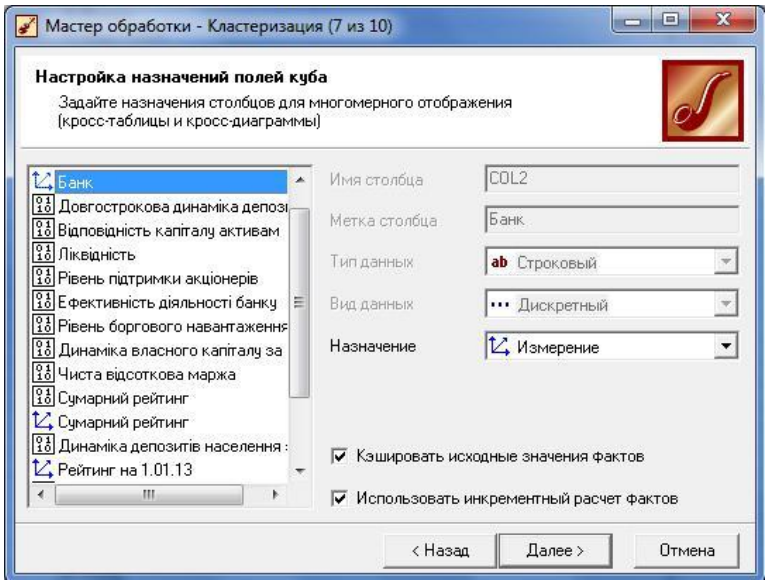


Рисунок 170 – Налаштування призначень полів куба

11 Налаштувати параметри діаграми розташування: по вісі *X* – БАНК, по вісі *Y* – ДІНАМІКА ДЕПОЗИТІВ НАСЕЛЕННЯ ЗА 12 МІС.,%, колір об’єкту – НОМЕР КЛАСТЕРУ (рис.171).

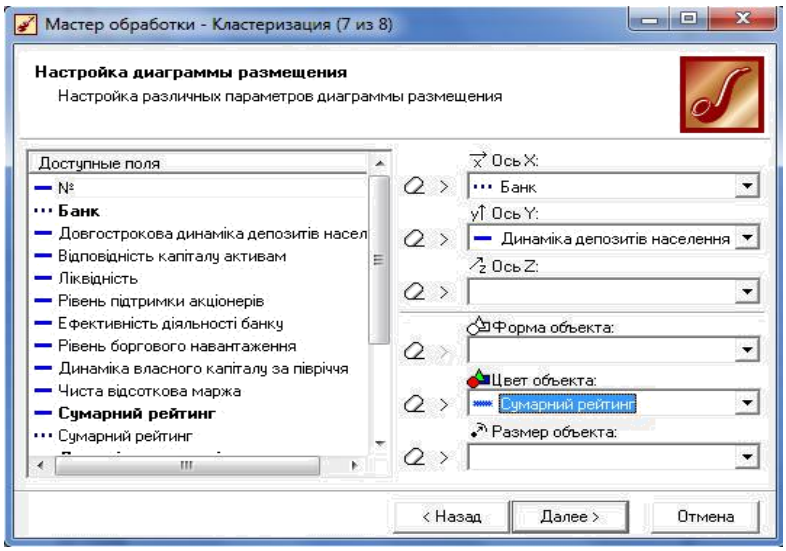


Рисунок 171 – Налаштування параметрів діаграми розташування

12 На 10 кроці *Майстра* виконати налаштування фактів. Так як вирішується задача вибору депозиту, то можна обрати такі факти: ДОВГОСТРОКОВА ДИНАМІКА ДЕПОЗИТІВ НАСЕЛЕННЯ, ЛІКВІДНІСТЬ, ДИНАМІКА ДЕПОЗИТІВ НАСЕЛЕННЯ ЗА 12 МІС.,%, КОЕФІЦІЄНТ СИСТЕМНОСТІ / ПРОБЛЕМНОСТІ. Доцільно в подальших налаштуваннях задати відображення фактів як середнє за групою, що розглядається (рис.172).

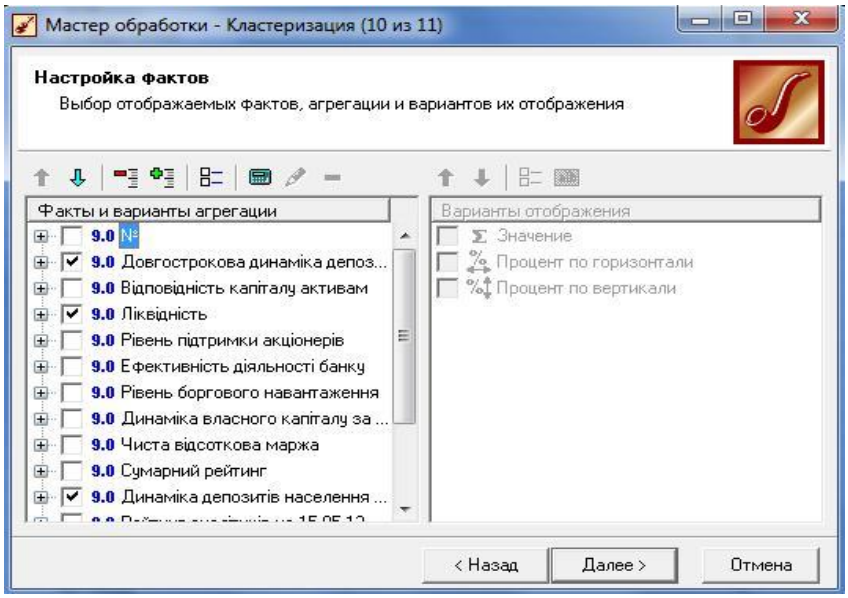


Рисунок 172– Налаштування фактів

13 Проаналізувати отримані результати.

13.1 *Профілі кластерів.* Загальну структуру сформованих алгоритмом кластерів можна переглянути в візуалізаторі *Профілі кластерів*, де подані всі розглянуті властивості разом з характером впливу їх на склад кластера.

Основним фактором, що визначає склад кластера, є *значущість* властивостей, подана у відсотках. Загальна значущість поля визначається варіабельністю його параметрів. Значущість для безперервних і дискретних полів визначається по різному. Для безперервних полів вона встановлюється в залежності від відхилення середнього значення розглянутої групи кластерів від загального середнього всієї вибірки. Чим більше таке відхилення, тим більше значущість поля. Для дискретних полів вона визначається наявністю індивідуальних відмінностей між розглянутими групами. Чим більш виражена відмінність, тим більша значущість поля. Для кожної властивості в кластері обчислюється довірчий інтервал, середнє, стандартне відхилення і стандартна помилка.

Алгоритм автоматично розбив банки на чотири кластери з різною підтримкою і різними відсотками значущості властивостей.

Найбільш значущим параметром для всіх 4 кластерів є РЕЙТИНГ НА 1.1.13 (рис.173).

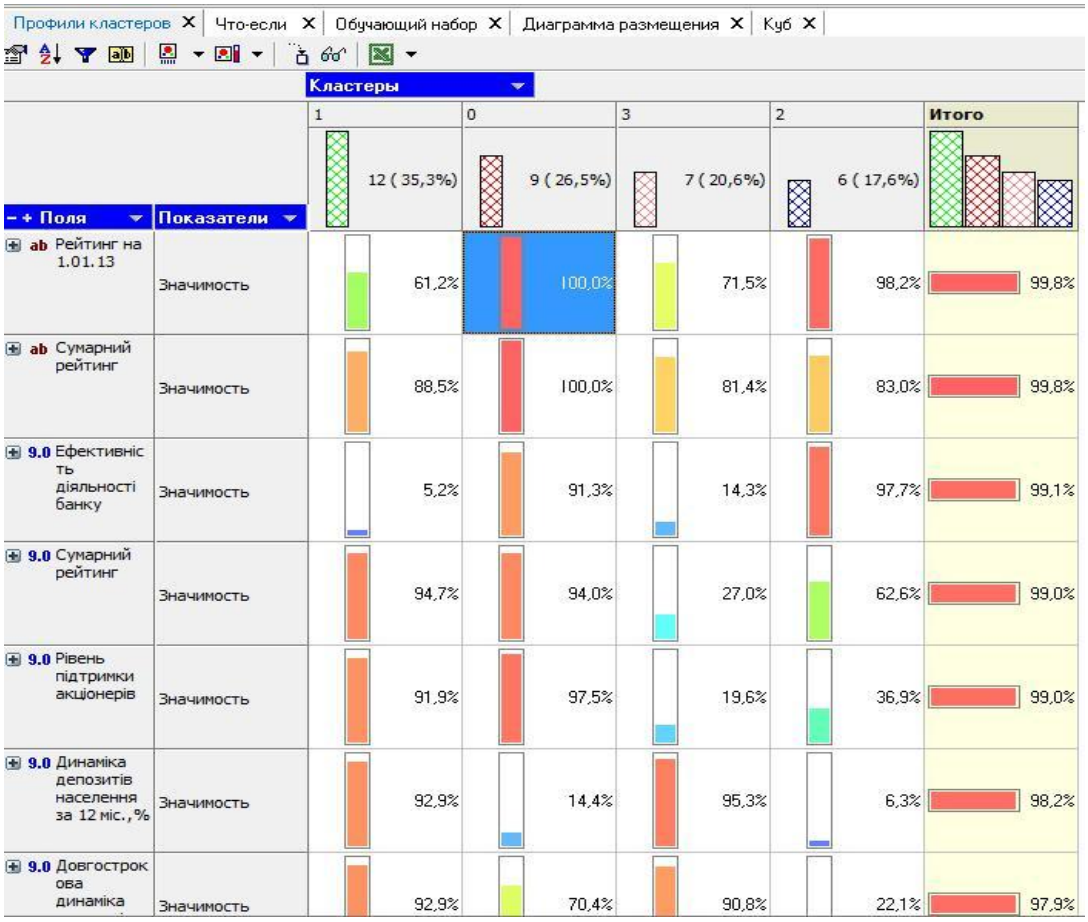


Рисунок 173– *Профілі кластерів*

Розглянемо 0 кластер: рейтинг на 1.1.13 та сумарний рейтинг мають значення 100%, ефективність діяльності банку достатньо висока (91,3%), довгострокова динаміка депозитів населення – велика, активність населення за попередній рік – мала.

Для банків кластеру 0 властивість, пов'язана з відповідністю капіталу активам, є малозначущою (рис.173).

13.2 Визначити кластери, в яких найбільш значущим фактором є ЛІКВІДНІСТЬ. Для цього натиснути кнопку налаштування сортування панелі інструментів та задати параметри сортування. Обрати тип сортування – за значимістю, напрямок – за убуванням й поле, за яким проводиться сортування та інше залишити без зміни (ри.174). Виконати сортування.

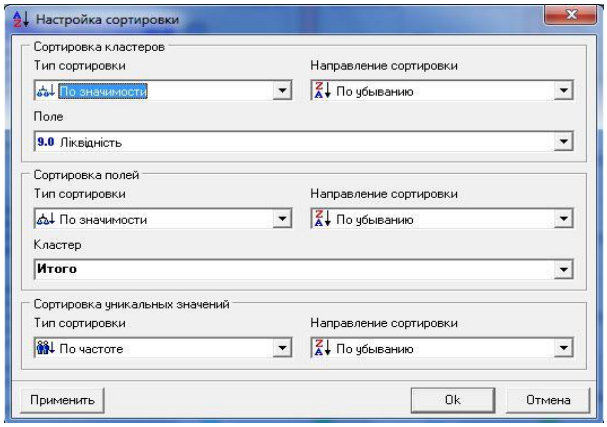


Рисунок 174 – Налаштування сортування

В залежності від значущості ліквідності кластери помінялися місцями в даному наборі. Кластери, що найбільш відрізняються за ліквідністю мають максимальну значущість (рис.175).

Рисунок 175– Результат сортування зі статистичною інформацією

13.3 *Куб.* Результати щодо сформованих кластерів зручно розглядати за допомогою візуалізатора *Куб,* який має вбудовану *кросс-діаграму*, що подає кластери в графічному вигляді.

14 Запустити процес Кластеризації. Для візуалізатора *Куб* призначення полів налаштувати самостійно.

15 Налаштувати призначення стовпців, тобто обрати властивості, за якими буде виконане групування об'єктів. *Вхідні* поля: ЛІКВІДНІСТЬ, ЧИСТА ВІДСОТКОВА МАРЖА, СУМАРНИЙ РЕЙТИНГ (числові значення), КОЕФІЦІЄНТ СИСТЕМНОСТІ/ПРОБЛЕМНОСТІ, всі інші поля – *Інформаційні*. При налаштуванні *Фактів* обрати для ЛІКВІДНІСТЬ – мінімальне значення, ЧИСТА ВІДСОТКОВА МАРЖА – максимальне значення, СУМАРНИЙ РЕЙТИНГ – середнє значення, КОЕФІЦІЄНТ СИСТЕМНОСТІ/ПРОБЛЕМНОСТІ – середнє значення.

16 Спосіб відображення даних обрати самостійно.

17 Проаналізувати отримані результати.

Контрольні питання

1 Дати визначення кластеризації.

2 Чим відрізняються об'єкти, що знаходяться в різних кластерах?

3 Зробити формальну постановку задачі кластеризації.

4 Перелічити задачі кластеризації.

5 Який найбільш поширеніший алгоритм кластеризації, в чому його сутність?

Лабораторна робота 15
РОБОТА ЗІ СХОВИЩЕМ ДАНИХ

Мета роботи – створення нового сховища даних та заповнення його даними.

Завдання на виконання лабораторної роботи.

Розробити проект структури сховища аптечної мережі. Для цього розробити та заповнити в табличному процесорі (*MS Excel, Calc)* чотири таблиці (табл. 11-14). Додати до 100 записів в кожну таблицю.

Таблиця 11 Групи товарів (фрагмент)

Код групи	*Найменування групи*
21	Імуномодулятори
25	Мікро та макроелементи
38	Вітаміни
76	Жовчогонні засоби
….	

Таблиця 12 –Товари (фрагмент)

Код товару	*Найменування товару*	*Код групи*
621	Імунорм	21
835	Ревіт	38
473	Альмагель	1
…		

Таблиця 13– Відділи

Код відділу	*Назва відділу*
1	Аптека 1
2	Аптека 2
3	Аптека 3
…	

Таблиця 14 – Продажі (фрагмент)

Дата	*Код відділу*	*Код товару*	*Час покупки*	*Кількість*	Сума
11.4.20	2	473	11	2	56,34
11.4.20	1	621	12	1	131,14
11.4.20	1	835	15	3	98,22
…					

Табл. 14 ілюструє процес продажів в трьох аптеках. За структурою сховища можна припустити, що унікальність точки в просторі визначається сукупністю вимірів: *Дата, Товар, Код відділу, Час покупки*. Це означає, що якщо в певній аптеці в певний день і час буде здійснено декілька покупок певного препарату, то в сховищі даних це буде відображено як один запис [1, 39].

1 Загальні положення

Аналітична платформа *Deductor* є основою для створення прикладних рішень в галузі аналізу даних. Реалізовані в *Deductor* технології дозволяють реалізувати всі етапи побудови аналітичної системи на базі єдиної архітектури від створення сховища даних до автоматичного підбору моделей і візуалізації отриманих результатів.

В основі концепції сховищ даних лежить інтеграція окремих деталізованих даних, що містяться в архівах традиційних систем транзакційної обробки із зовнішніх джерел, в єдину базу даних.

Б. Інмон охарактеризував сховища даних, як "предметно орієнтовані, інтегровані, незмінні набори даних, що підтримують хронологію та організовані з метою підтримки управління" [40].

Deductor Warehouse – багатовимірне сховище даних, що акумулює необхідну для аналізу предметної області інформацію, яка зберігається в структурах типу «сніжинка», де в центрі розташовані таблиці фактів, а «променями» є виміри, які, в свою чергу, можуть посилатися на інші виміри.

Така архітектура сховища є найбільш прийнятною для задач аналізу даних, в яких аналітик оперує багатовимірними поняттями. Кожна «сніжинка» асоціюється з *процесом* і описує певну дію, наприклад, продаж товару, відвантаження, надходження грошових коштів тощо.

Виміри можуть бути простими списками *(*наприклад, *Клієнт)*, або містити додаткові стовпці, звані *атрибутами вимірів*. Так вимір *Товар* може складатися з *Назви товару* (первинний ключ), *Ваги* та *Об'єму,* атрибутів даного виміру.

Процеси, як і виміри, можуть мати свої атрибути, які так і називаються – *атрибути процесу*, які на відміну від атрибутів виміру не визначають координати в багатовимірному просторі, а є довідковими значеннями, що характеризують процес, наприклад, *№ накладної, Валюта документа* тощо. Значення атрибутів процесу на відміну від атрибутів виміру можу бути невизначеними. Часто складно визначитися, що віднести до атрибутів процесу, а що – до атрибутів виміру.

Фізично *Deductor Warehouse* – це реляційна база даних, що містить таблиці для зберігання інформації й таблиці зв'язків, що забезпечують цілісність відомостей. Крім того, в реляційній базі даних реалізований спеціальний семантичний шар, який перетворює реляційне подання в багатовимірне, що найкращим чином відповідає ідеології аналізу даних. Завдяки цьому шару, користувач оперує не полями і записами таблиць бази даних, а багатовимірними поняттями (вимір, факт), і система автоматично виконує всі маніпуляції, необхідні для роботи з реляційною СУБД.

Сховище даних *Deductor Warehouse* є прозорим для користувача в питаннях проведення операцій щодо підключенню реляційної СУБД та пошуку потрібної інформації. Прозорість роботи сховища забезпечується системою на базі трьох СУБД: *Firebird, Microsoft SQL, Oracle*.

Deductor Warehouse реалізує універсальне багатовимірне зберігання, тобто може містити множину процесів з різною кількістю вимірів і фактів. Налаштування процесів, завдання вимірів, атрибутів і фактів може здійснюватися за допомогою *редактора метаданих*, вбудованого в *Deductor Studio* [2, 39].

2Порядок виконання лабораторної роботи

1 Створення нового порожнього сховища.

1.1 Підключити *Deductor Studio* до *Deductor Warehouse6*. Для цього в меню *Вид* обрати команду *Підключення*.

1.2 Натиснути правою кнопкою миші на піктограмі *Підключення* (меню *Вид*) та в контекстному меню обрати пункт *Майстер підключень*, де вказати джерело даних – сховище даних *Deductor Warehouse 6* (рис.176).

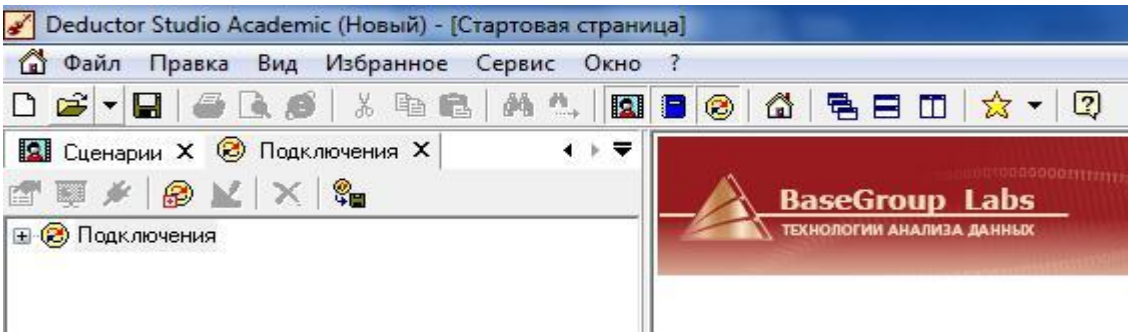


Рисунок 176 Робоче вікно *Deductor*

1.3 Обрати рядок *Deductor Warehouse 6* і натиснути кнопку Далі (рис.177).

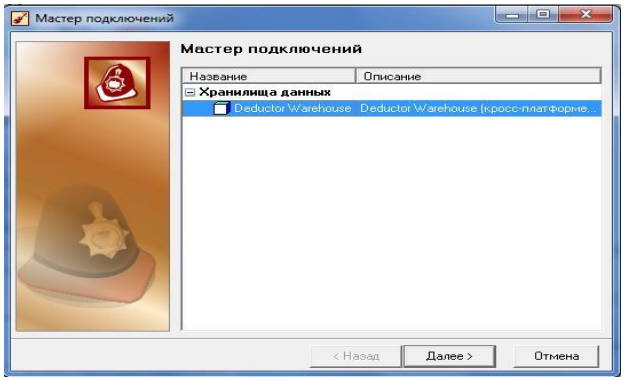


Рисунок 177– Діалогове вікно *Майстра підключень*

1.4 На 3 кроці *Майстра підключень* налаштувати параметри бази даних, де буде створюватись фізична та логічна структура сховища даних (рис.178). Для цього натиснути кнопку з 3 крапками в полі БАЗА ДАНИХ, після чого відкриється форма, яка містить поля для підключення до серверу бази даних (рис.179).

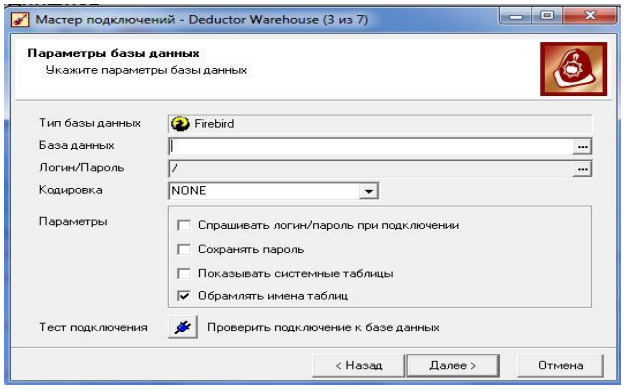


Рисунок 178 – Налаштування параметрів бази даних

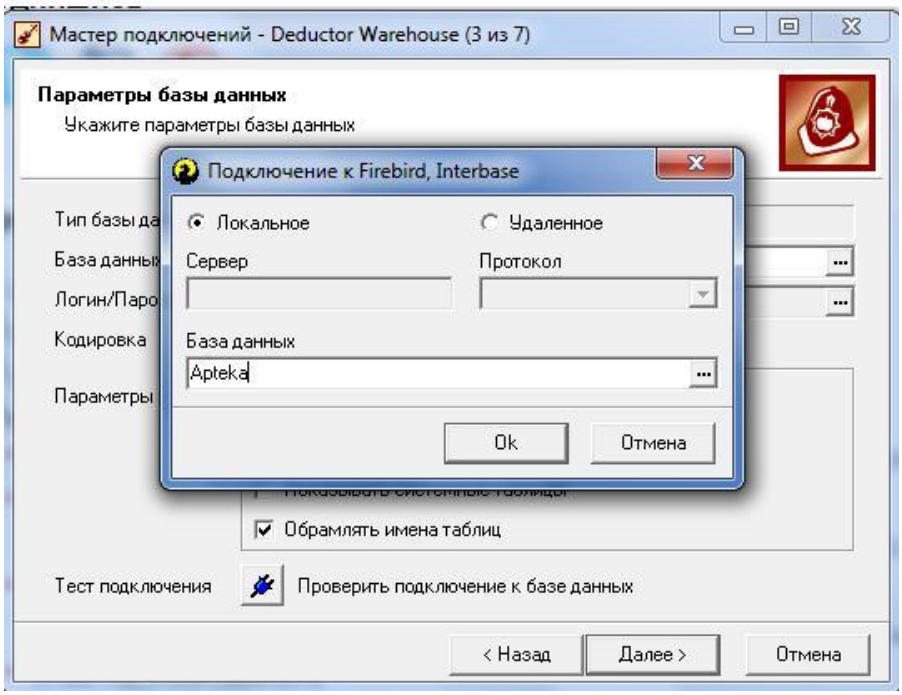


Рисунок 179 – Діалогове вікно підключення до серверу бази даних

1.5 Встановити перемикач в положення *Локальне*, в полі БАЗА ДАНИХ вписати ім’я бази даних – *Apteka* та закрити вікно ПІДКЛЮЧЕННЯ ДО… кнопкою Ok (рис.179).

1.6 На 4 кроці *Майстра підключень* обрати *Deductor Warehouse 6* та натиснути кнопку Далі (рис.180).

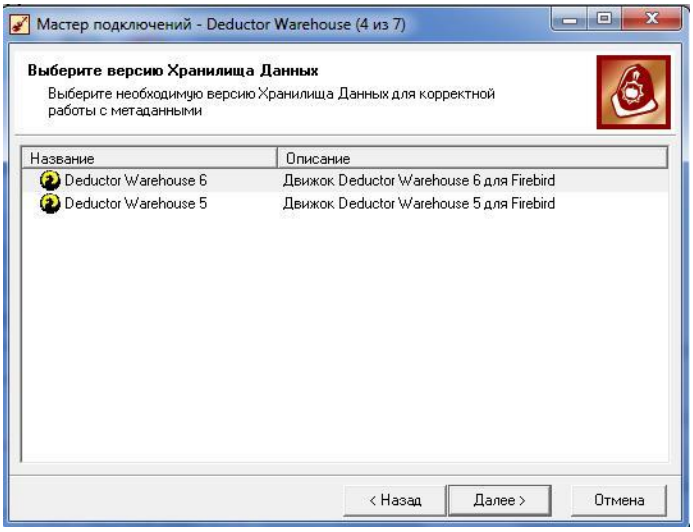


Рисунок 180 – Вибір версії сховища даних

1.7 На 5 кроці *Майстра підключень* (Інструменти роботи зі сховищем даних) натиснути кнопку Створити (рис.181).

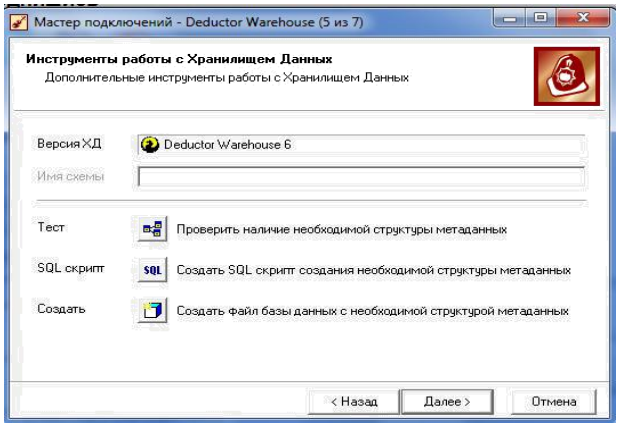


Рисунок 181 – Інструменти роботи зі сховищем даних

1.8 На 6 кроці *Майстра підключень* обрати способи відображення даних – *Відомості* та *Метадані* та натиснути кнопку Далі (рис.182).

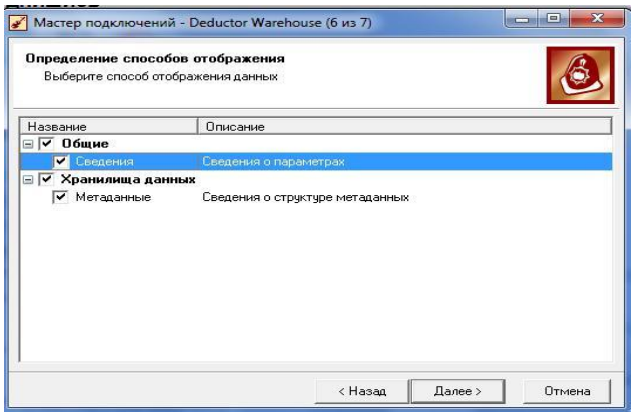


Рисунок 182 – Вікно визначення способів відображення даних

1.9 На 7 кроці *Майстра підключень* залишити за замовчуванням ім’я сховища даних та мітку та натиснути кнопку Готово (рис.183).

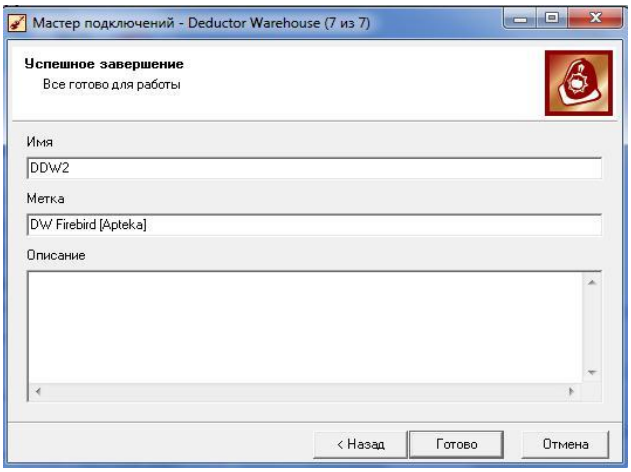


Рисунок 183– Успішне закінчення роботи *Майстра підключень*

На дереві вузлів підключень з’явиться мітка сховища (рис.184).

Рисунок 184 – Дерево вузлів підключень

1.10 Натиснути піктограму *Зберегти налаштування підключень* .

Таким чином, порожнє сховище створено, але в ньому поки що немає жодного об'єкта (процесу, виміру, факту). Потрібно завантажити спроектовану базу даних аптечної мережі в сховище. Скористаємось для цього *Редактором метаданих*.

1.11 Відкрити діалогове вікно *Редактор метаданих* піктограмою (рис.184,185).

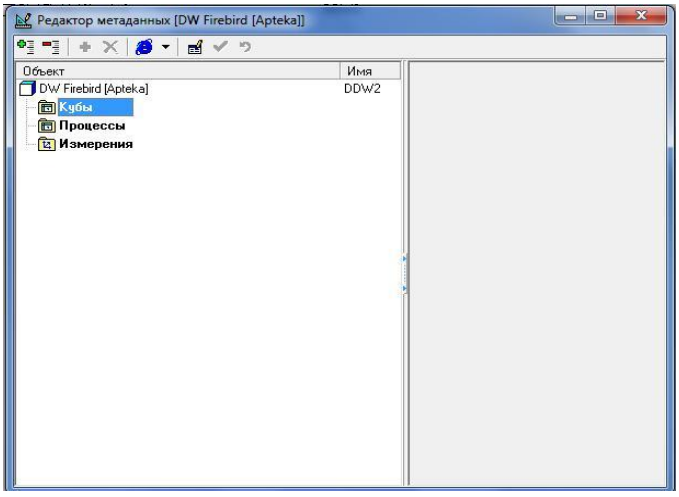


Рисунок 185– Діалогове вікно *Редактору метаданих*

Для переходу до режиму зміни структури сховища необхідно натиснути піктограму *Дозволити редагування.* З'явиться діалогове вікно з попередженням про небезпечність цієї операції.

1.12 Обрати вузол *Вимір* (рис.185) та в контекстному меню – команду *Додати* і створити перший вимір, *Код групи*, з наступними параметрами: *Ім'я*– GR_ID; *Мітка* – Група.Код; *Тип даних* – Цілий (рис.186).

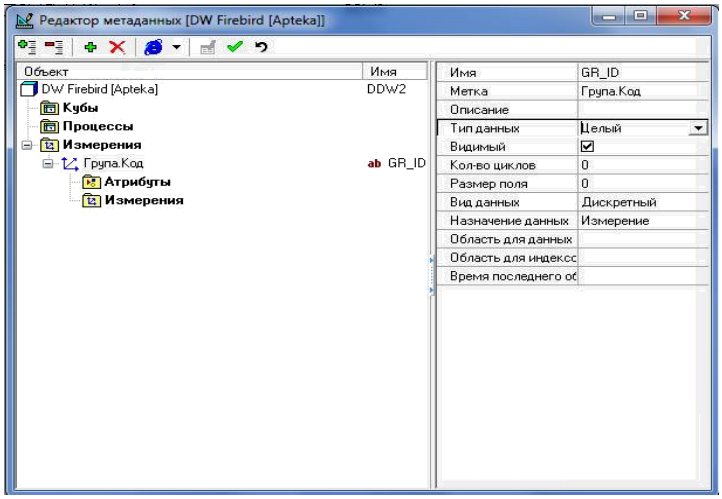


Рисунок 186 – Створення виміру в *Редакторі метаданих*

Мітка – це семантична назва об'єкта сховища даних, що побачить користувач, який працює зі сховищем даних.

1.13 Виконати аналогічні дії для створення всіх інших вимірів, з параметрами з табл.15.

Таблиця 15 – Параметри вимірів

Вимір	*Імя*	*Мітка*	*Тип даних*
Код групи	GR_ID	Група.Код	Цілий
Код товару	TV_ID	Товар.Код	Цілий
Код відділу	PART_ID	Відділ.Код	Цілий
Дата	S_DATE	Дата	Дата/час
Час покупки	S_HOUR	Час	Цілий

В результаті структура метаданих сховища матиме 5 вимірів(рис.187).

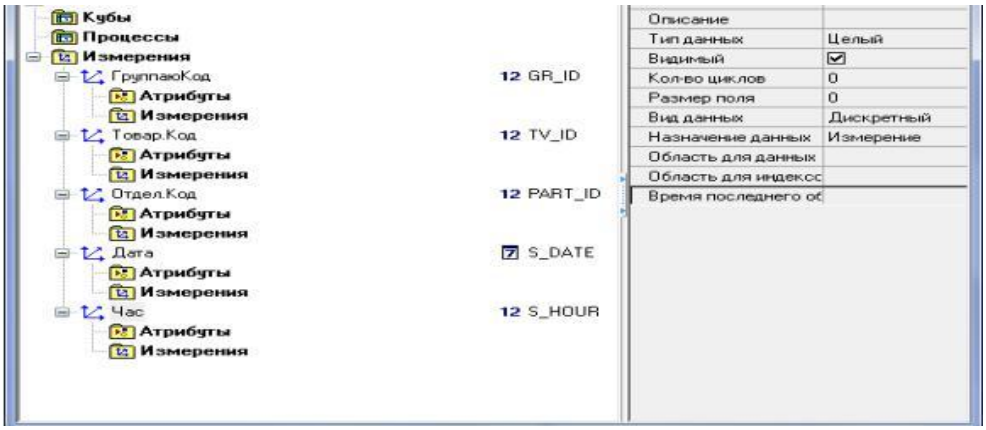


Рисунок 187 – Структура метаданих сховища

1.14 Обрати в контекстному меню вузла *Атрибути* команду *Додати* для додавання до кожного виміру, окрім *Дата* і *Час*, текстового атрибуту. Для виміру *Группа.Код* додати текстовий атрибут *Група.Найменування*, для виміру *Товар.Код – Товар.Найменування*, для виміру *Відділ.Код – Відділ.Найменування* (рис.188).

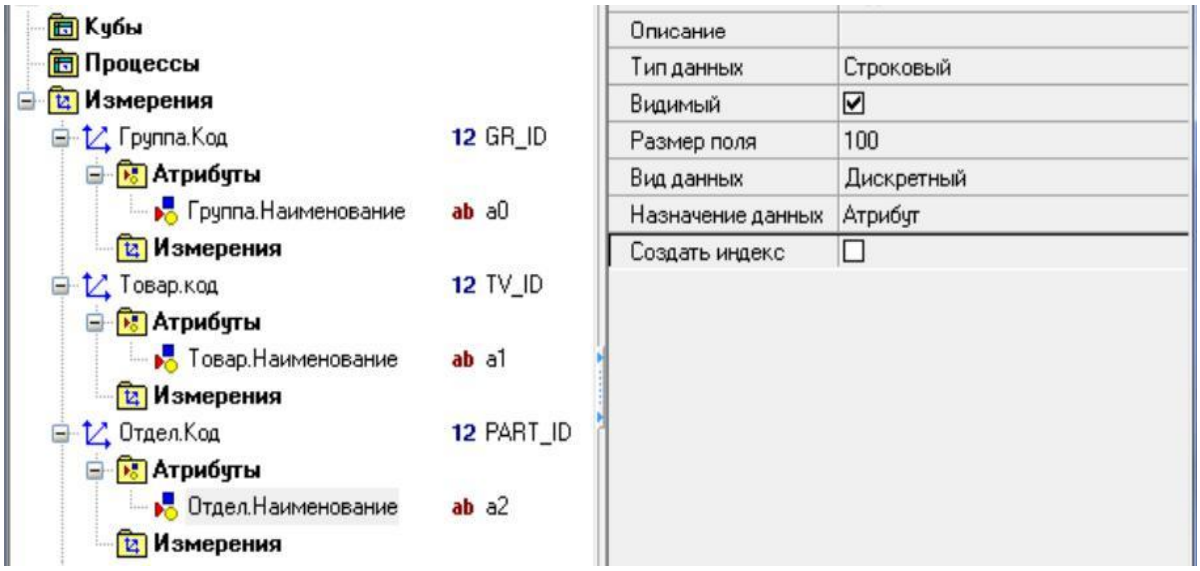


Рисунок 188 – Додавання текстового атрибуту

1.15 Зробити посилання виміру *Товар.Код* на вимір *Група.Код* (реалізація ієрархії вимірів) за допомогою команди *Додати* (рис.189).

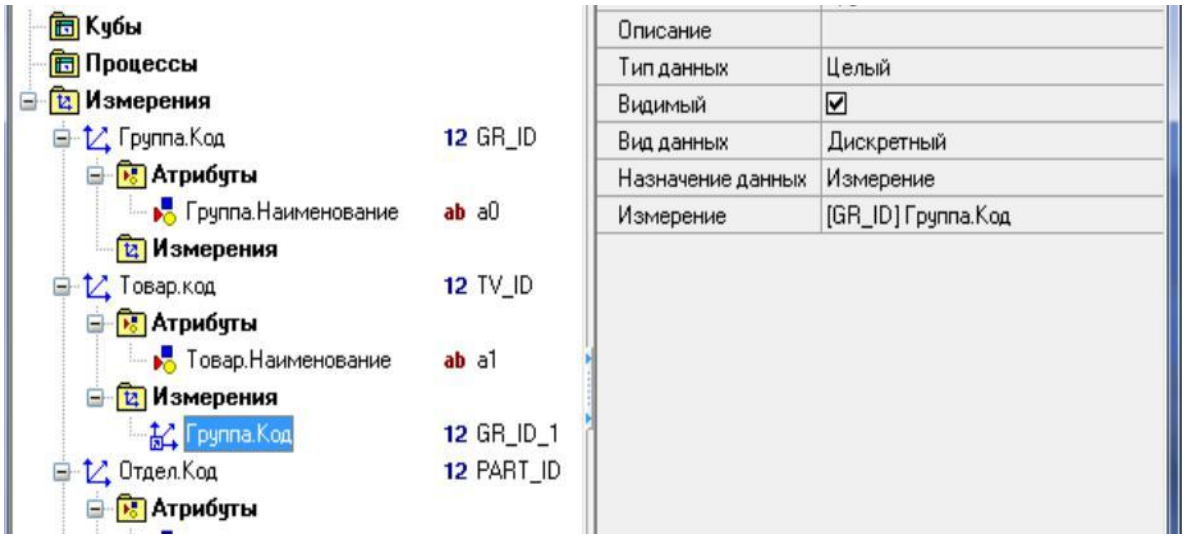


Рисунок 189 – Реалізація ієрархії вимірів

1.16 Після того, як всі виміри та посилання будуть описані виконати процес формування «сніжинки», який назвати *Продаж*. Для цього необхідно додати (рис.190):

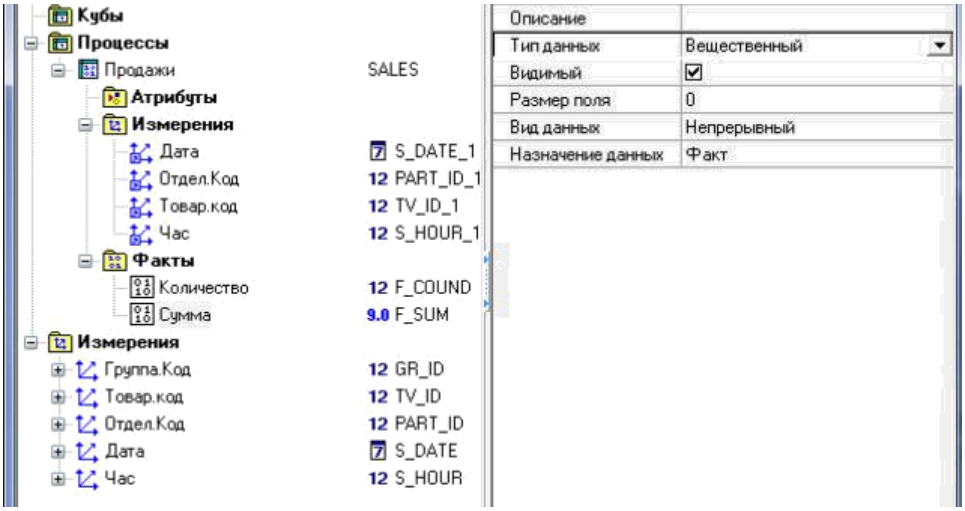


Рисунок 190 – Процес формування «сніжинки»

– чотири існуючих виміра: *Дата, Код відділу (мітка Відділ.Код), Код товару (мітка Товар.Код), Час покупки (мітка – Час)*;

– два факти: *Кількість* і *Сума*.

1.17 Зафіксувати зміни структури сховища даних піктограмою *Прийняти зміни*. На цьому проектування структури і метаданих закінчено.

2 Наповнення сховища даних.

Після створення структури сховище є порожнім файлом з налаштованим семантичним шаром, який повністю підготовлений до завантаження в нього даних із зовнішніх структурованих джерел. Для цього в *Deductor Studio* слід написати відповідний сценарій для виконання низки функцій:

• імпорт в *Deductor Studio* записів бази даних або визначених файлів;

• операційна передобробки даних, наприклад, очищення або перетворення формату;

• завантаження даних в виміри і процеси сховища *Deductor Warehouse6.*

В нашому прикладі вихідними даними для сховища є чотири текстових файли: *Групи Товарів.txt, Товари.txt, Відділи.txt, Продажі.txt* і сценарій повинен бути налаштованим на використання цих файлів як джерел даних.

При створенні сценарію необхідно дотримуватися наступних правил:

• першими завантажуються всі виміри, що мають атрибути. Тільки після завантаження всіх вимірів дані завантажуються в процес;

• виміри потрібно завантажувати, починаючи з верхнього рівня ієрархії. Це вкрай важливо: в іншому випадку ієрархія не буде відтворена;

• припускається не завантажувати окремі виміри, які не мають атрибутів і не перебувають у ієрархії вимірів. Значення таких вимірів можна створювати під час завантаження в процес за допомогою спеціальної операції.

2.1 Імпортувати 4 текстових файли в визначеній послідовності: *Групи Товарів.txt, Товари.txt, Відділи.txt, Продажі.txt* (вкладка *Сценарії, Майстер імпорту*).

2.2 Виділити перший вузол (текстовий файл *Групи Товарів.txt) Сценарію* та активізувати *Майстер експорту.*

2.3 Обрати зі списку типу приймачів *Майстра експорту* пункт *Deductor Warehouse6*, а зі списку сховищ — *Apteka* (рис.191).

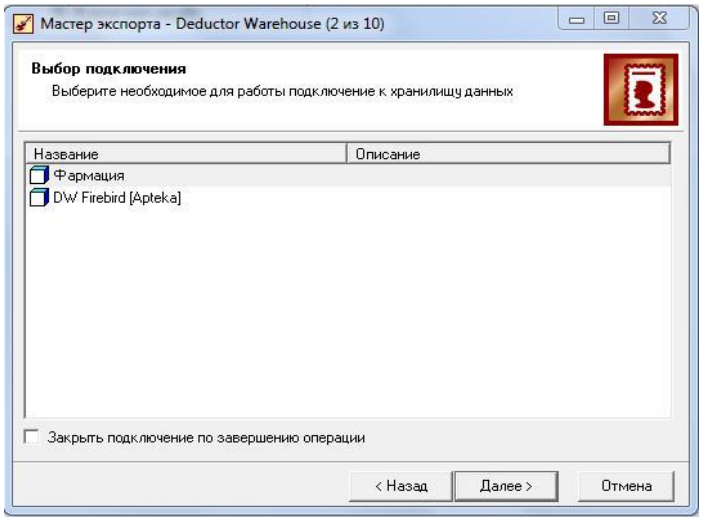


Рисунок 191 – Вибір підключення *Майстра експорту*

2.4 Вказати вимір, куди буде завантажуватися інформація – *Група.Код,* та встановити відповідність між елементами об'єкта сховища даних та полями вхідного джерела даних (таблиці *Групи Товарів.txt*) (рис.192). Якщо імена полів і (або) мітки в семантичному шарі сховища даних збігаються, відповідність буде встановлено автоматично.

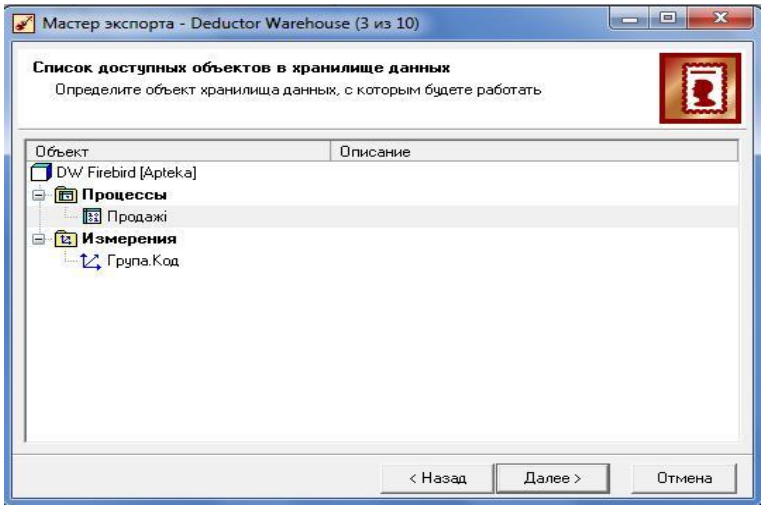


Рисунок 192 – Список доступних об’єктів сховища даних

2.5 Аналогічно завантажити виміри *Товар.Код, Відділ.Код.*

Виміри *Дата* та *Час* не містять атрибутів і не приймають участь в побудові ієрархії, тому їх значення можна завантажити на етапі експорту даних в процес.

2.6 Завантажити дані в процес *Продажі*. На відміну від завантаження вимірів в *Майстрі експорту* з'являться два специфічних кроки. На одному з них потрібно задати параметри контролю несуперечності даних у сховищі та вказати виміри, за якими слід видаляти дані зі сховища. Обирається дія, що виконується, коли в процес завантажується інформація, яка в деяких вимірах збігається за значенням. В такому випадку можливі два варіанти: видалити застарілі дані й завантажити нові або залишити те, що було завантажено раніше.

2.7 Виділити останній текстовий файл *Продажі.txt* і запустити *Майстер експорту*. На третьому кроці виділити процес *Продажі* та натиснуть кнопку Далі. На четвертому кроці встановити відповідність елементів об'єкта сховища полям вхідного джерела даних. На цьому кроці починається завантаження даних в сховище.

2.8 На п'ятому кроці вказати вимір *Дата*, за яким будуть видалятися дані зі сховища.

2.9 На 6 кроці *Майстра експорту* залишити налаштування за замовчуванням. Прапорець АВТОМАТИЧНО ДОДАВАТИ ЗНАЧЕННЯ ВИМІРУ дозволяє "зльоту" додавати нові значення в існуючі виміри. Цю операцію потрібно застосовувати обережно, оскільки можна дуже швидко засмітити сховище непотрібними даними.

2.10 На наступному кроці *Майстра* налаштувати варіант агрегації даних. Як спосіб агрегації обрати *Суму*.

У результаті всіх виконаних дій буде:

• створено і наповнено сховище даних;

• написано сценарій завантаження (поповнення) даних з джерела і сховища даних;

• продумано контроль несуперечності даних у сховищі даних.

Отриманий сценарій завантаження прив'язаний не до даних безпосередньо, а до їх структури, тобто в ньому змодельована послідовність дій, які потрібно виконувати для завантаження даних у сховище даних: імена файлів-джерел, відповідність полів тощо. Одного разу створений сценарій постійно застосовується для поповнення сховища даними. Як правило, ці процедури проводяться за регламентом у неробочий час (наприклад, вночі) з використанням пакетного або серверного режиму.

3 Імпорт даних зі сховища Deductor Warehouse.

3.1 Викликати *Майстер імпорту*, який призначено для автоматизації отримання даних з джерела, передбаченого в списку всіх налаштованих в системі типів джерел даних.

3.2 Обрати на 2 кроці *Майстра імпорту* тип джерела даних – *Сховище даних Deductor Warehouse*, та ім'я сховища –*Apteka* .

3.3 На 3 кроці *Майстра імпорту* зі *Списку доступних в сховищі даних о*брати об'єкт, з яким плануєте працювати, наприклад, процес *Продажі*.

3.4 В діалоговому вікні ІМПОРТ ДАНИХ ЗІ СХОВИЩА (4 крок *Майстра імпорту*) визначити, які факти, виміри та атрибути будуть імпортуватися, наприклад, вимір – *Група.Код* та факти – *Сума* та *Кількість*.

3.5 На 5 кроці *Майстра імпорту* визначити зрізи для обраних вимірів та атрибутів. На етапі визначення зрізів задаються умови відбору даних зі сховища.

3.6 Запустити процес вилучення даних з *Deductor Warehouse*, натиснувши кнопку Пуск на 7 кроці *Майстра імпорту.*

3.7 На 8 кроці *Майстра імпорту* визначити способи відображення даних, наприклад, *Таблиці, Діаграми та OLAP- куб.*

Наступні кроки пов'язані з налаштуванням таблиці, діаграми та *OLAP*– куба.

3.8 Самостійно здійснити налаштування *Таблиці, Діаграми* та *OLAP* – куба. В останньому вікні *Майстра імпорту* визначається ім'я імпортованого набору даних. Після натискання кнопки Готово в робочому вікні з'явиться набір даних.

3.9 Проаналізувати отримані результати.

Контрольні питання

1 Дати визначення поняття «сховище даних» та сформулювати концепцію сховища даних.

2 Навести архітектуру сховища *Deductor Warehouse.*

ЗЯкий інструмент інтелектуального аналізу даних дозволяє перетворити реляційне подання даних в багатовимірне?

4 Дати визначення реляційної бази даних.

5 Дати визначення процесу, виміру та атрибуту.

6 Описати алгоритм створення порожнього сховища.

7 З яких типів джерел даних можна виконувати імпорт даних?

СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ

1 Новожилова М.В., Петрова О.О., Чуб О.І. Лабораторний практикум «Інтелектуальний аналіз даних в аналітичній платформі Deductor»: Навчальний посібник– Х.:ХНУБА, 2014. – 160 с.

2 http://www.basegroup.ru

3 Паклин, Н. Б., Орешков В. И. Бизнесаналитика: от данных к знаниям (+CD) : учеб. пособие 2е изд., перераб. и доп. – СПб. : Питер, 2010. – 704 с.

4 Дюк В.А., Самойленко А.П. Data Mining: учебный курс – СПб: Питер, 2001 – 368 с.

5 Дюк В.А. Осколки знаний// Экспресс Электроника, 2002, № 6, С. 6065.

6 Киселев М., Соломатин Е.. Средства добычи знаний в бизнесе и финансах. Открытые системы, № 4, 1997, с. 4144.

7 Хальд А, Математическая статистика с техническими приложениями. М: ИЛ, 1956.

8 S.Murthy. Automatic construction of decision trees from data: A Multidisciplinary survey.1997.

9 J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.

10W. Buntine. A theory of classification rules. 1992. 10Бонгард М.М. Проблема узнавания. М: Наука, 1967

11 Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. – М., 2004.

12 Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М: Наука, 1979.

13 Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.

14 http://www.intuit.ru/department/itmngt/itmangt/6/7.html

15 http://masters.donntu.edu.ua/2006/fvti/ichernov/diss/index.htm

16 Дюк В.А. Data Mining интеллектуальный анализ данных http://www.iteam.ru/publications/it/section_55/article_1448/print/

17 Дюк *В.А.* Data Data Mining состояние проблемы, новые решения http://www.osp.ru/text/302/177842.html

18 Венкатеш Ганти, Йоханнес Герке, Раджу Рамакришнан Добыча даннях в сверхбольших базах данных Открытые системы №0910/1999 http://www.osp.ru/text/302/177842.html

19 http://www.basegroup.ru/deductor/repl_of_knowledge/

20http://www.intuit.ru/department/database/datamining/26/datamining_26.html #image.26.1

21http://fakit.ru/common/ITBank/PZ3_Complex.pdf Золотарюк А.В. Интеллектуальные компьютерные технологии обработки социологической информации

22 Интеллектуальные модели анализа экономической информации: электронный курс лекций. – BaseGroup Labs, 2005.

23 http://isusibadi.ru/scince/books/detail.php?ID=3705&PAGEN_1=9

24 А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВПетербург, 2004.336 с.: ил.

25 http://www.basegroup.ru/library/analysis/regression/logistic/ Паклин Н

26 Оссовский С. Нейронные сети для обработки информации / Пер. с польского И. Д. Рудинского. – М.: Финансы и статистика, 2004. – 344 с.

27 Чубукова И. А. Data Mining: учеб. пособие. М.: Интернетуниверситет информационных технологий: БИНОМ: Лаборатория знаний, 2006. 382 с.

28 http://www.gotai.net/documents/docnn009.aspx

29 Ширяев В.И. Нейросетевые методы в анализе финансовых рынков // учеб. пособие. М.: КомКнига, 2007

30 http://www.scienceeducation.ru/1068039

31http://ru.wikipedia.org/wiki/%D0%A1%D0%B0%D0%BC%D0%BE%D0%BE%D1%80%D0%B3%D0%B0%D0%BD%D0%B8%D0%B7%D1%83%D1%8E%D1%89%D0%B0%D1%8F%D1%81%D1%8F_%D0%BA%D0%B0%D1%80%D82%D0%B0_%D0%9A%D0%BE%D1%85%D0%BE%D0%BD%D0%B5%D0%BD%D0%B0

32 Демин И.С. Методическое пособие по изучению средств самоорганизующихся карт Кохонена с помощью програмного средства Deductor Studio. – М., 2010, 28 с.

33 http://www.basegroup.ru/library/practice/solvency/

34 http://www.basegroup.ru/glossary/definitions/clustering/

35http://www.stgau.ru/company/personal/user/7684/files/lib/Информационные%20системы%20в%20экономике/Лабораторные%20работы/Лабораторные%20работы%20в%20Deductor%20Studio%205.2.pdf

36 http://isusibadi.ru/scince/books/detail.php?ID=3705&PAGEN_1=8

37Золотарюк А.В. Интеллектуальные компьютерные технологии обработки социологической информации http://fakit.ru/common/ITBank/PZ4_Transfor.pdf

38 http://www.infology.ru/2008/05/04/284/

39 http://liber.hse.perm.ru/files/UMP.pdf Лебедев В.В

40 http://xn7sbaaeuq0a9apc1a6b.xnp1ai/blog/?tag=billinmon

ЗМІСТ

Навчальне видання

Шаповалова Олена Олександрівна

Інтелектуальний аналіз даних з практикумом в Deductor

Роботу до видання рекомендував Сироватський О.А.

За редакцією автора

План 2020, поз.1	Форм. 60х84.1/16.	Папір друк. №2.
Підп. до друку	Обл.вид. арк. 8,0.	
Надруковано на ризографі.	Умов. друк. арк. 7,8.	Безкоштовно.
Тираж 100 прим.	Зам. № 2374.	

ХНУБА, Україна, 61002, Харків, вул. Сумська, 40

Підготовлено та надруковано РВВ Харківського національного університету будівництва та архітектури