Міністерство освіти і науки України
Херсонський державний аграрно-економічний університет

Г. О. Димова, О. В. Ларченко

МОДЕЛІ І МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Навчальний посібник

Херсон
2021

УДК 004.6:004.422:519.688
Д 46

Рекомендовано до друку
Вченою Радою Херсонського державного
аграрно-економічного університету
(протокол № 4 від 06.12.2021)

Рецензенти:

Рудакова Г.В. д.т.н., професор, професор кафедри автоматизації, робототехніки та мехатроніки Херсонського національного технічного університету;

Шарко О.В. д.т.н., професор, професор кафедри транспортних технологій та механічної інженерії Херсонської державної морської академії.

Димова, Г.О., Ларченко, О.В.

Д 46 Моделі і методи інтелектуального аналізу даних: навчальний посібник / Г.О. Димова, О.В. Ларченко – Херсон: Книжкове видавництво ФОП Вишемирський В. С., 2021. 142 с.

ISBN 978-617-7941-60-5

В посібнику розглядаються методи, інструментальні засоби та застосування інтелектуального аналізу даних, спрямованих на вивчення можливостей програми Deductor Studio: імпорт даних, створення сценаріїв обробки, застосування методів візуалізації. Опис кожного способу супроводжується конкретним прикладом його використання.

УДК 004.6:004.422:519.688

ISBN 978-617-7941-60-5

ЗМІСТ

ВСТУП

Deductor – це аналітична платформа, основа для створення закінчених прикладних рішень у сфері аналізу даних. Реалізовані в Deductor технології дозволяють на основі єдиної архітектури пройти всі етапи побудови аналітичної системи: від консолідації даних до побудови моделей та візуалізації отриманих результатів.

До появи аналітичних платформ аналіз даних здійснювався переважно у статистичних пакетах. Їхнє використання вимагало високої кваліфікації користувача. Більшість алгоритмів, реалізованих у статистичних пакетах, не дозволяло ефективно опрацьовувати великі обсяги інформації. Для автоматизації рутинних операцій доводилося використовувати інтегровані мови програмування.

Наприкінці 80-х років відбулося стрімке зростання обсягів інформації, що накопичується на машинних носіях та зросли потреби бізнесу щодо застосування аналізу даних. Відповіддю цьому стала поява нових парадигм у аналізі: сховища даних, машинне навчання, Data Mining, Knowlegde Discovery in Databases. Це дозволило популяризувати аналіз даних, вивести його на промислову основу та розв'язати величезну кількість бізнес-задач з великим економічним ефектом.

Вінцем розвитку аналізу даних стали спеціалізовані програмні системи – аналітичні платформи, які повністю автоматизували всі

етапи аналізу від консолідації даних до експлуатації моделей та інтерпретації результатів.

Перша версія Deductor побачила світ у 2000 р. і з того часу йде безперервний розвиток платформи. У 2007 р. випущено п'яту за ліком версію системи, у 2009 р. – версію 5.2.

Сьогодні Deductor – це яскравий представник як настільної, так і корпоративної системи аналізу даних останнього покоління.

Перший і другий розділи направлені на знайомство з аналітичною платформою Deductor Studio Academic, підготовку та імпортування даних в цю систему.

Третій розділ навчає працювати з майстром обробки даних. В ньому розглядається методи: відновлення пропущених даних, видалення аномалій, спектральна обробка та видалення шумів.

Четвертий розділ присвячений освоєнню навичок застосування факторного та кореляційного аналізу.

П’ятий розділ спрямований на отримання навичок розбиття даних, квантування та фільтрації для трансформації даних.

Розділ шість спрямований на освоєння інструменту, що дозволяє розв’язувати та використовувати математичні функції.

В сьомому розділу робота спрямована на вивчення асоціативних правил та використання візуалізаторів «Популярні набори», «Правила», «Дерево правил», «Що-якщо».

Восьмий розділ присвячений застосуванню методів Data Mining для розв’язання задач прогнозування часових рядів на прикладі побудови моделі.

Усі розділи мають перелік питань для самоконтролю. Наприкінці посібника є список літератури, що рекомендується.

Набуті здобувачами вищої освіти теоретичні знання та практичні навички допоможуть їм успішно освоїти програму курсу «Інтелектуальний аналіз даних» та будуть затребувані у їхній подальшій професійній діяльності.

Розділ 1
ЗНАЙОМСТВО З ПЛАТФОРМОЮ DEDUCTOR STUDIO ACADEMIC. ПІДГОТОВКА ДАНИХ

1.1 Аналітична платформа Deductor Studio Academic

Deductor Studio Academic є платформою, орієнтованою на розв'язання задач аналізу найширшого спектра: від створення систем корпоративної звітності до розв'язування задач Data Mining.

Метою написання посібника є вивчення корпоративної системи звітності та багатостороннього аналізу будь-якого виду діяльності на основі аналітичної платформи Deductor Studio 5 Academic компанії Base Group.

В Deductor Studio використовуються найпотужніші технології, такі як багатовимірний аналіз, нейронні мережі, дерева рішень, карти, що самоорганізуються, спектральний аналіз і безліч інших. При цьому акцент зроблений на самонавчаючи методи і машинне навчання, що дозволяє будувати адаптивні системи, здатні реагувати на зміну ситуації.

Реалізовані в Deductor Studio технології дають можливість на базі єдиної платформи пройти всі етапи побудови аналітичної системи: від створення сховища даних до автоматичного підбору моделей і візуалізації отриманих результатів:

- системи аналітичної звітності;

- багатовимірний аналіз;
- прогнозування;
- пошук закономірностей;
- управління ризиками;
- сегментація клієнтів/товарів/послуг;
- побудова профілів споживачів;
- оцінка ефективності реклами;
- аналіз маркетингових даних.

Deductor Studio – аналітичне ядро платформи Deductor. Deductor Studio містить повний набір механізмів імпорту, обробки, візуалізації і експорту даних для швидкого і ефективного аналізу інформації. У ньому зосереджені найсучасніші методи вилучення, очищення, маніпулювання і візуалізації даних. З ним стають доступні моделювання, прогнозування, кластеризація, пошук закономірностей і багато інших технології виявлення знань (Knowledge Discovery in Databases) і видобутку даних (Data Mining).

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, таблиці, діаграми, дерева рішень і т.д.) і експортувати результати. Вся робота з аналізу даних в 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 доступним способом. Результатом обробки так само є набір даних, який, в свою чергу, знову може бути оброблений. Результати обробки можна переглянути безліччю способів і експортувати в найбільш популярні формати.

Послідовність дій, які необхідно провести для аналізу даних, є сценарієм, який можна автоматично виконувати на будь-яких даних. Deductor Studio підтримує безліч джерел даних - промислові СУБД (Oracle, MS SQL), текстові файли, офісні додатки (Excel, Access), ADO і ODBC джерела, і повністю інтегровано з багатовимірним сховищем даних Deductor Studio Warehouse.

Під обробкою розуміють будь-яку дію, пов'язану з перетворенням даних, наприклад, побудова моделей, очищення від шумів і аномальних значень. При цьому механізми обробки можна комбінувати довільним чином так, щоб досягти найкращого результату.

Візуалізація – це відображення імпортованих і оброблених даних. Візуалізувати можна будь-який об'єкт в сценаріях обробки. Програма самостійно аналізує, яким чином можна відобразити інформацію, користувач повинен лише вибрати потрібний варіант.

1.2 Підготовка даних для імпорту

Для початку роботи з Deductor Studio в програму слід імпортувати дані з будь-якого джерела. Для цього необхідно створити вихідні дані - базу даних.

Дані можуть бути представлені в будь-якому стандартному табличному вигляді: текстові файли і файли DBF, MS Excel, бази даних MS Access, MS SQL, Oracle, InterBase, будь ODBC джерело.

Academic версія призначена тільки *для освітніх цілей*. У ній обмежені можливості інтеграції і автоматичної обробки. Підтримується тільки три джерела і приймача даних: *Deductor Warehouse, Deductor Data File і* ***текстові*** *файли.*

Для проведення аналізу даних за допомогою аналітичної платформи Deductor Studio необхідно імпортувати текстовий файл. Для цього потрібно створити файл *.xls або *.xlsx з сайту, наприклад, https://finance.i.ua, як показано на рисунку 1.1.

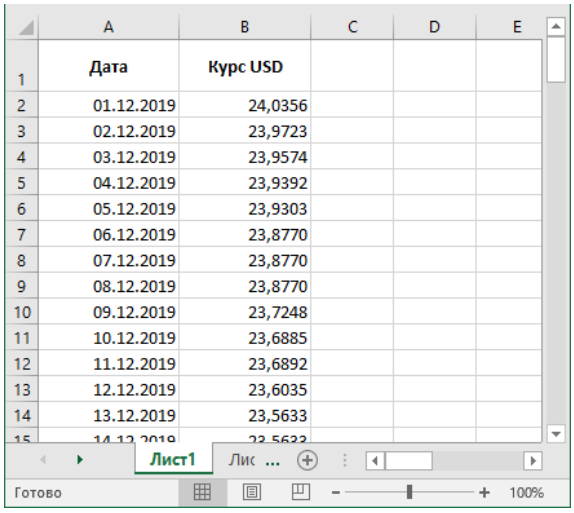


Рис. 1.1 – Приклад формування даних

Експортуємо цей файл в текстовий з роздільниками. Результат цієї дії показаний на рис. 1.2.

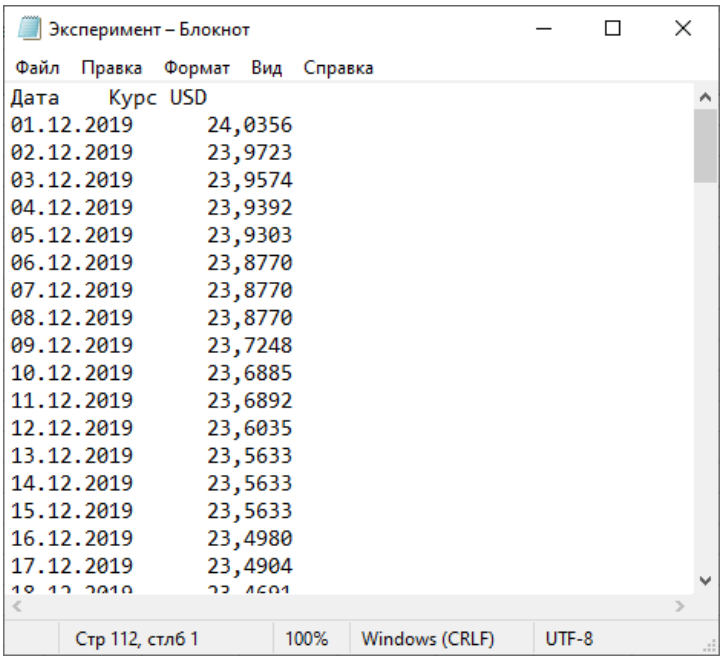


Рис. 1.2 – Приклад формування текстових даних

Дані сформовані. Тепер можна переходити до їх імпортування в програму Deductor Studio Academic.

Розділ 2
РОБОТА З ПЛАТФОРМОЮ DEDUCTOR STUDIO ACADEMIC

2.1 Виконання імпорту в платформі Deductor Studio

Інтерфейс Deductor Studio складається з головного вікна, всередині якого розташовуються панелі сценаріїв, звітів, джерел даних і результати моделювання (таблиці, графіки, крос-діаграми, правила і т.д.).

Для автоматизації отримання даних з будь-якого джерела, передбаченого в системі, слід використовувати майстер імпорту. На першому кроці майстра імпорту відкривається список всіх передбачених в системі типів джерел даних. Число кроків майстра імпорту, а також набір параметрів, що настроюються відрізняється для різних типів джерел.

Імпорт здійснюється шляхом виклику майстра імпорту на панелі «Сценарії» (рис. 2.1, 2.2).

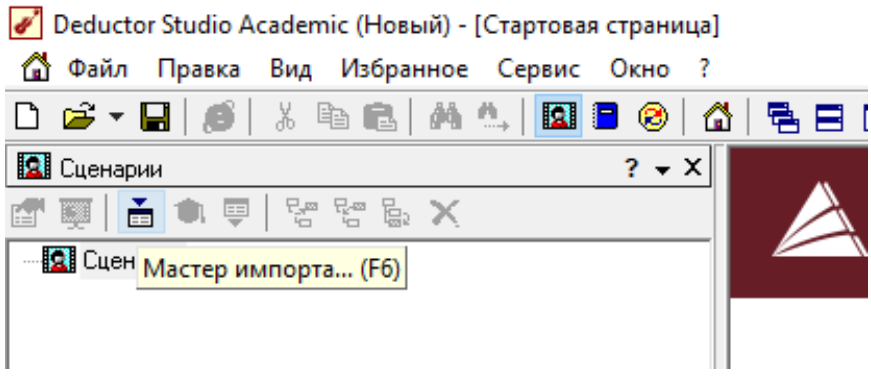


Рис. 2.1 – Імпорт даних

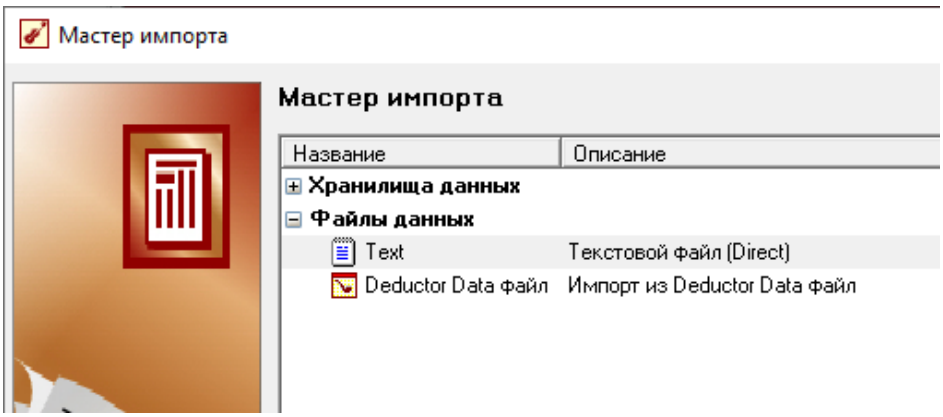


Рис. 2.2 – Імпорт даних (крок 1)

Після запуску майстра імпорту вкажемо тип імпорту «Текстовий файл з роздільниками» і перейдемо до налаштування імпорту. Зазначимо ім'я файлу, з якого необхідно отримати дані (приклад для парціальної обробки). У вікні перегляду обраного файлу можна побачити зміст даного файлу (рис. 2.3).

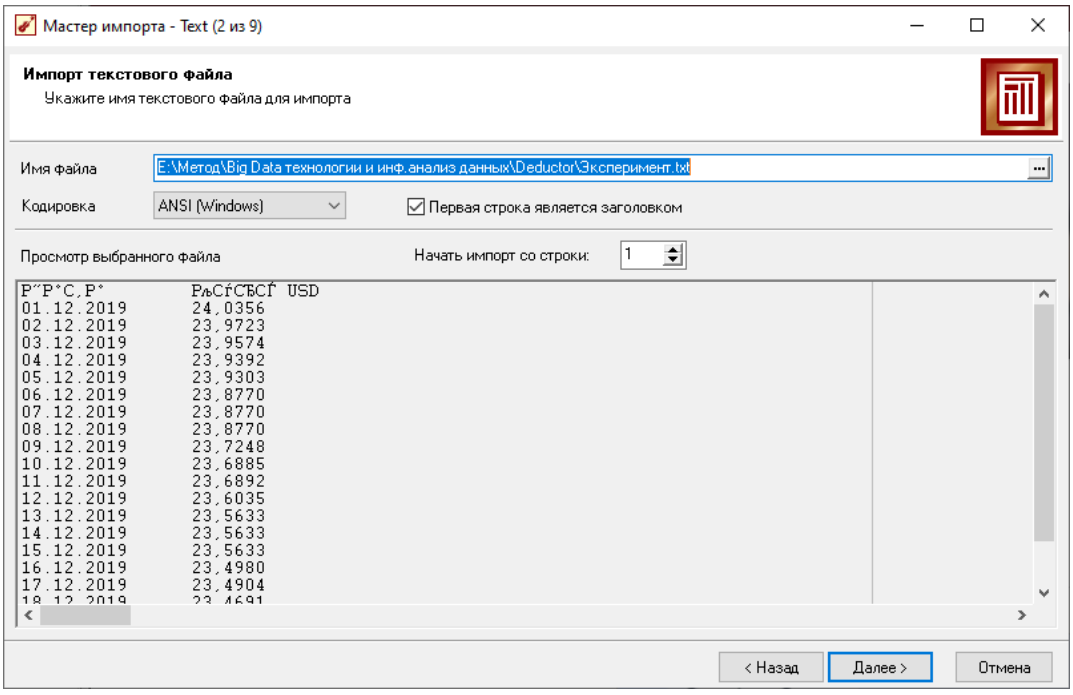


Рис. 2.3 – Імпорт даних (крок 2)

Далі перейдемо до налаштування параметрів імпорту (рис. 2.4, 2.5). На цій сторінці майстра надається можливість вказати, з якого рядка слід почати імпорт, вказати, то, що перший рядок є заголовком, можливість додати первинний ключ. Вказати, що є символом-роздільником стовпців, а також вказати обмежувач рядків, роздільник цілої та дробової частини дійсного числа, роздільник компонентів дати і її формат.

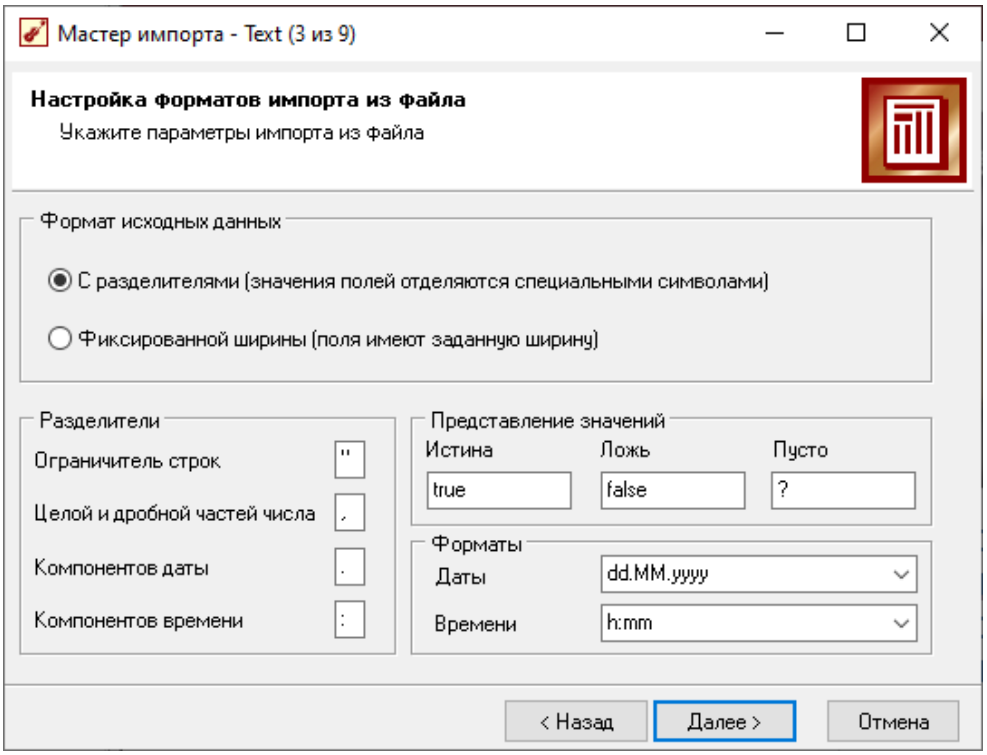


Рис. 2.4 – Імпорт даних (крок 3)

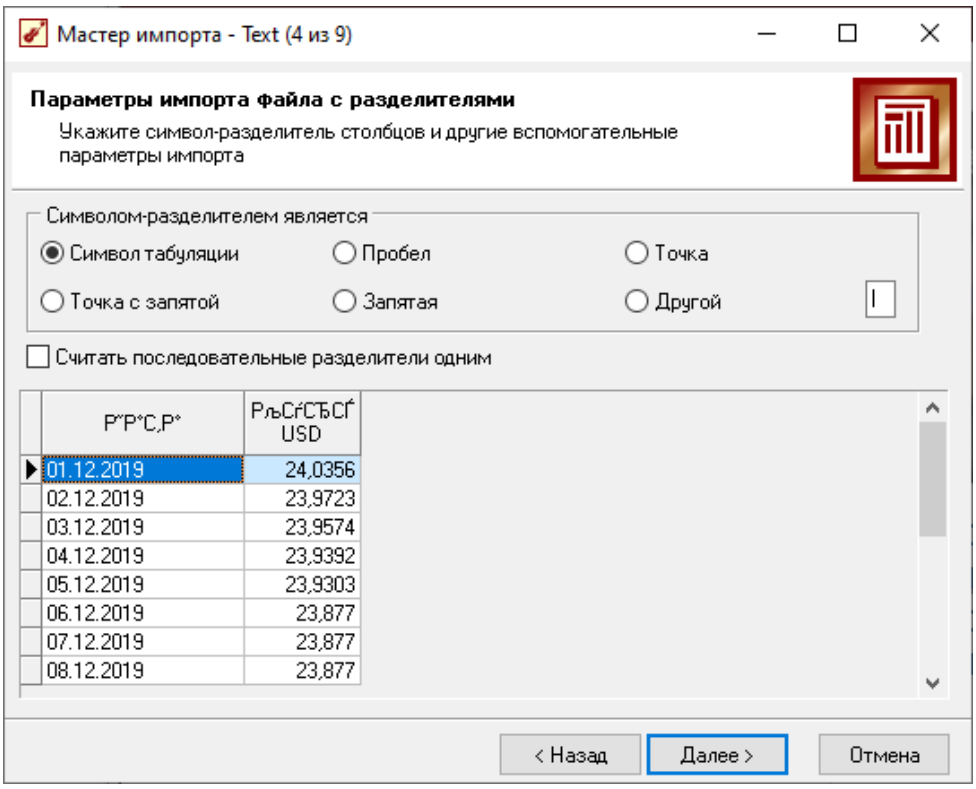


Рис. 2.5 – Імпорт даних (крок 4)

В даному випадку параметри за замовчуванням на цій сторінці майстра встановлені правильно, а саме: почати імпорт з першого рядка, перший рядок є заголовком, роздільником між стовпцями є знак табуляції, роздільником цілої та дробової частин є кома.

На наступному кроці майстра надається можливість налаштувати ім'я, назву (мітку), розмір, тип даних, вид даних і призначення. Деякі властивості (наприклад, тип даних) можна задавати для виділеного набору стовпців (рис. 2.6). Для правильного імпорту даних необхідно змінити тип даних у стовпців.

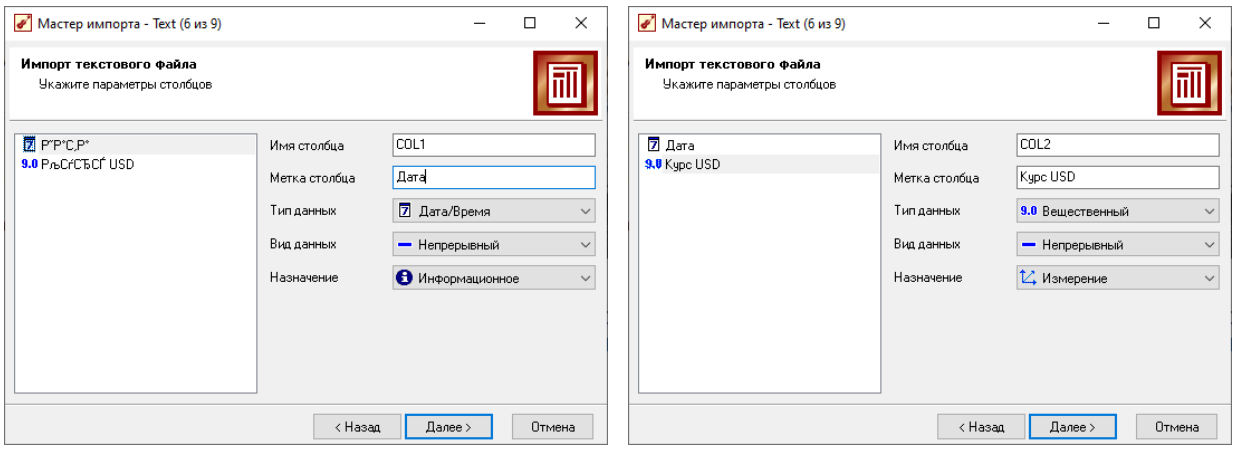


Рис. 2.6 – Налаштування властивостей стовпців

Після імпорту даних на наступному кроці майстра необхідно вибрати спосіб відображення даних (рис. 2.7, 2.8). В даному випадку найбільш інформативним є діаграма, виберемо її.

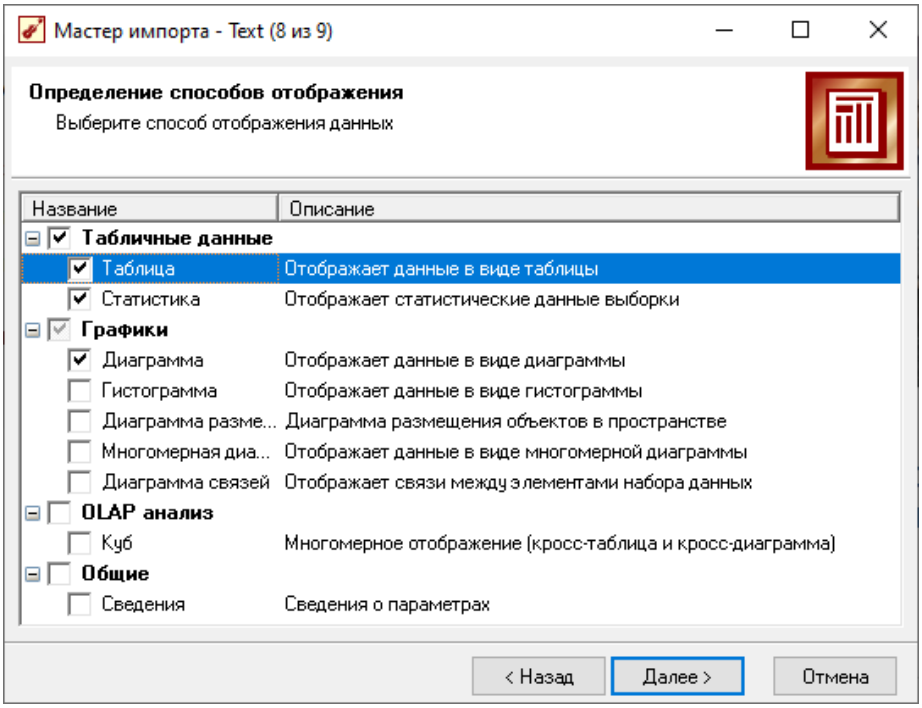


Рис. 2.7 – Спосіб відображення даних

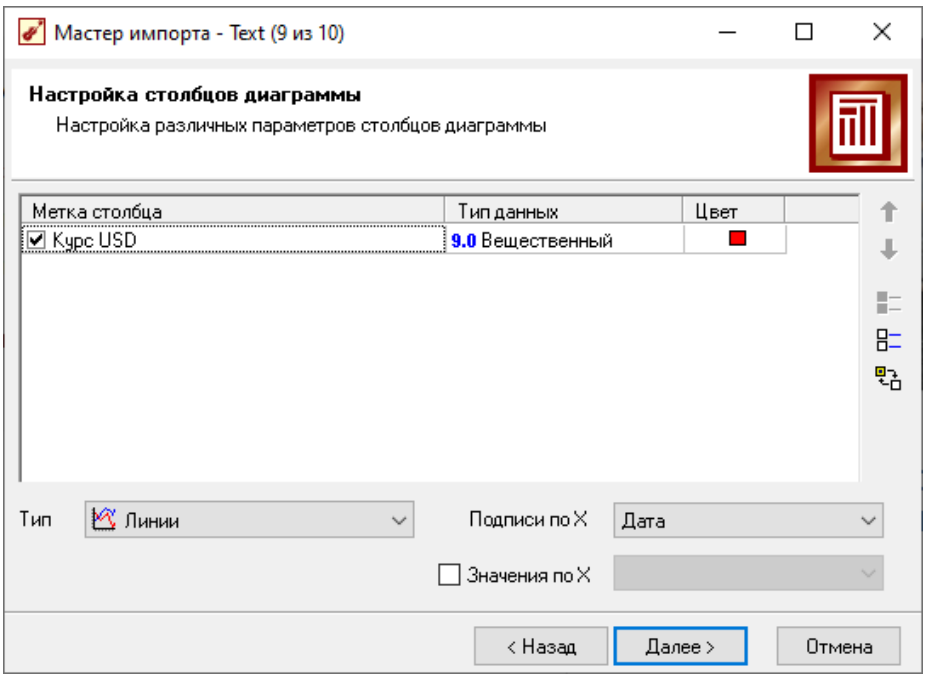


Рис. 2.8 – Налаштування способу відображення

2.2 Результат імпорту даних

По закінченню імпорту отримуємо результати, що були налаштовані для відображення. Результати імпортування даних з використанням платформи Deductor Studio Academic показано на рисунках 2.9 – 2.12.

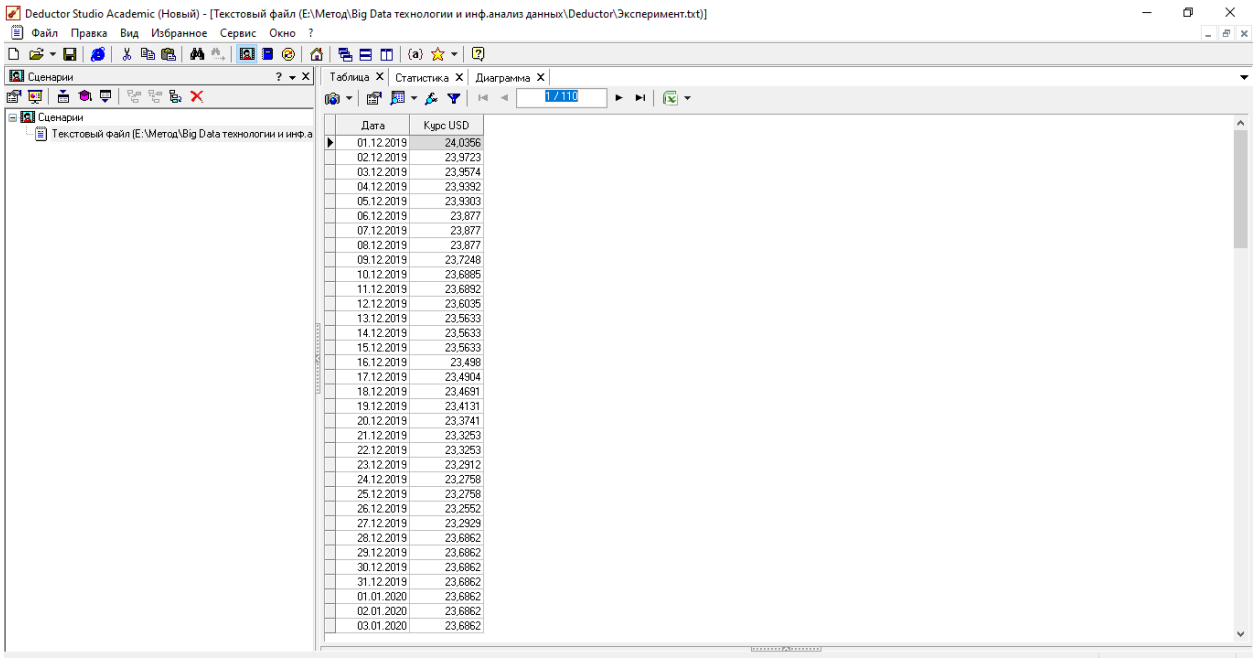


Рис. 2.9 – Вигляд таблиці імпортованих даних

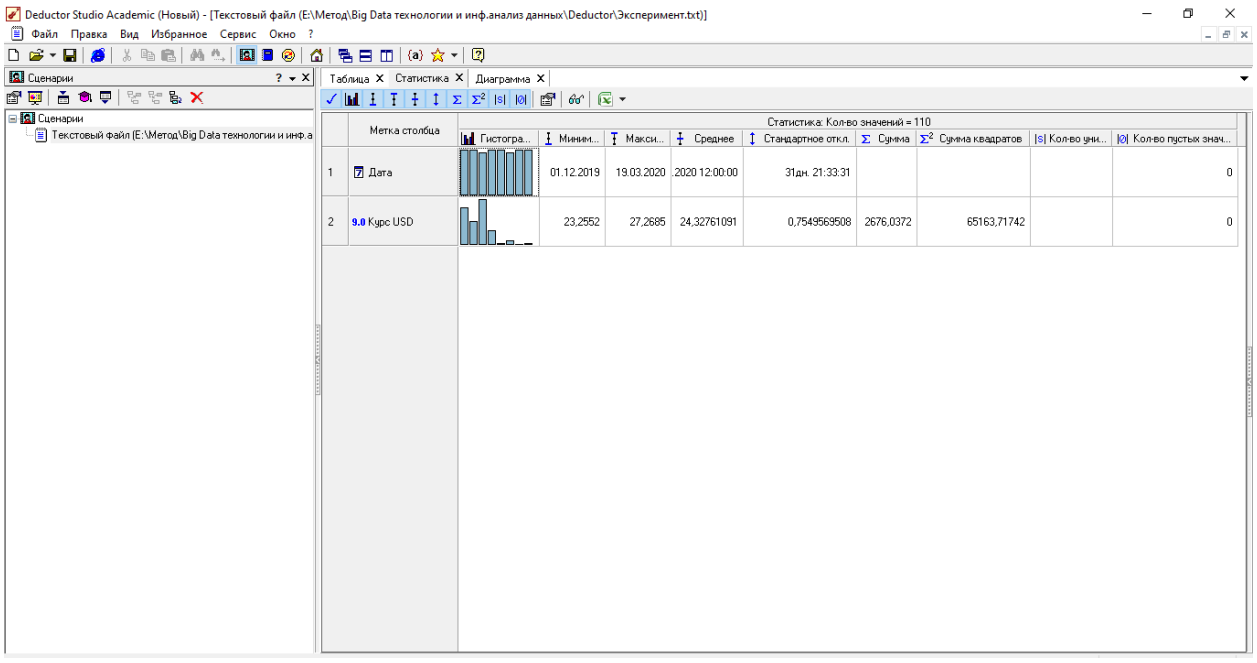


Рис. 2.10 – Вигляд таблиці розрахунку статистики імпортованих даних

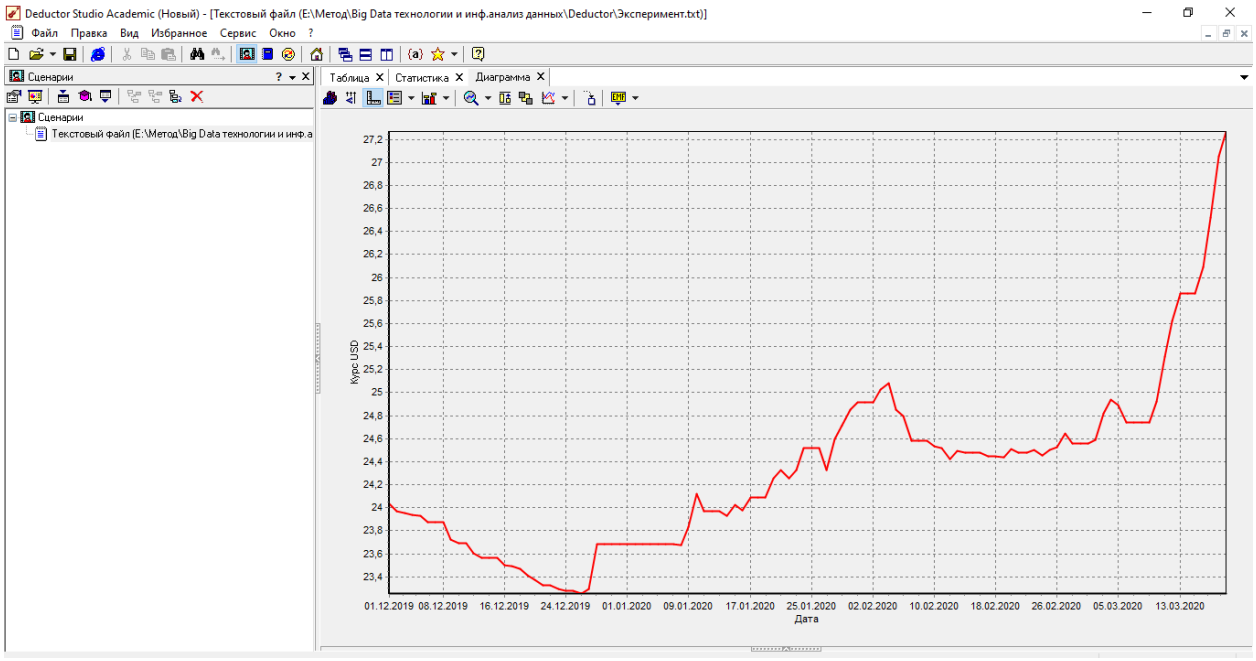


Рис. 2.11 – Вигляд плоскої діаграми імпортованих даних

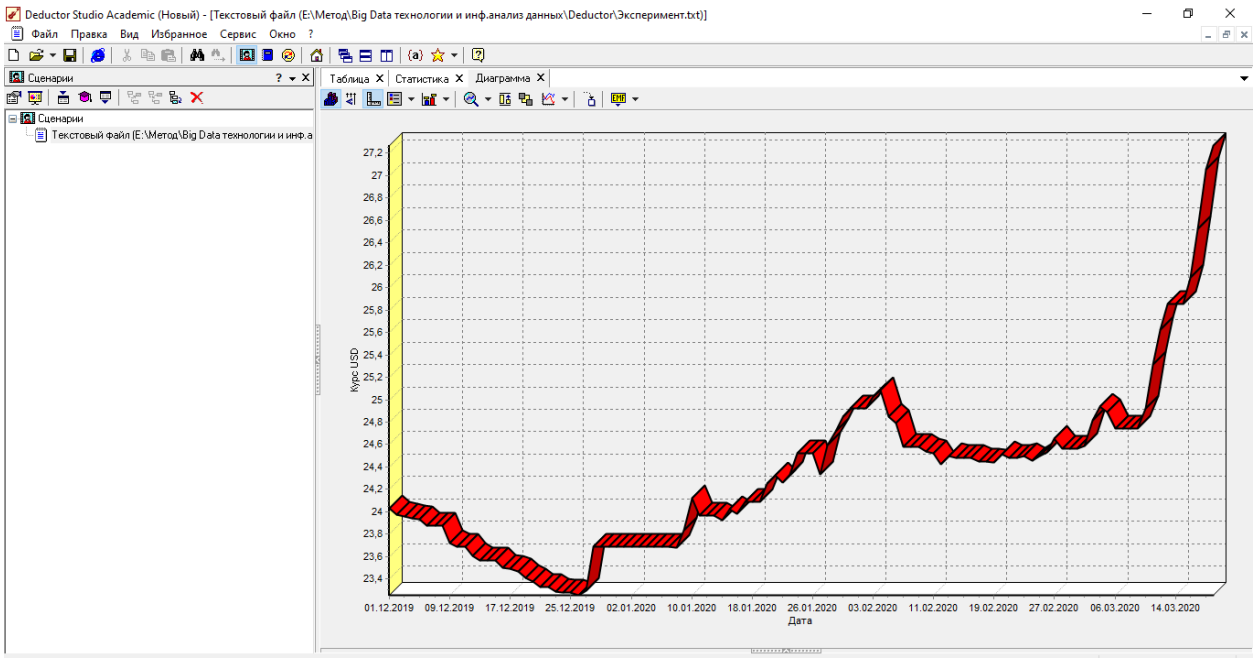


Рис. 2.12 – Вигляд об’ємної діаграми імпортованих даних

На цьому закінчується процес імпорту даних, які в подальшому потрібно обробляти.

Розділ 3
ОБРОБКА ДАНИХ ЗА ДОПОМОГОЮ ПЛАТФОРМИ DEDUCTOR STUDIO ACADEMIC

3.1 Робота з майстром обробки

Deductor Studio – програма, яка 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 спільне 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 ...) і експортувати результати на сторону. Це повністю відповідає концепції вилучення знань з баз даних (KDD).

Майстер обробки призначений для налаштування всіх параметрів обраного алгоритму.

У покроковому режимі вибрати і налаштувати найбільш зручний спосіб представлення даних можна за допомогою майстра відображень. Залежно від обробника, в результаті якого була отримана гілка сценарію, список доступних для нього видів відображень буде різним. Наприклад, після побудови дерев рішень їх можна відобразити за допомогою візуалізаторів «Дерева рішень»

і «Правила». Ці способи відображення не доступні для інших обробників.

Вид даних визначає – чи кінцевий це набір (дискретні) або нескінченний (безперервні). Призначення стовпців визначає характер їх використання в алгоритмах обробки (при імпорті можна залишити значення за замовчуванням).

Часто вихідні дані не є достатньо повними або мають різні шуми і не годяться для аналізу, а якість даних впливає на якість результатів. Так що питання підготовки даних для подальшого аналізу є дуже важливим. Зазвичай «сирі» дані містять в собі різні шуми, за якими важко побачити загальну картину, а також аномалії – вплив подій, що відбувалися випадково або рідко. Очевидно, що вплив цих факторів на загальну модель необхідно мінімізувати, тому що модель, що враховує їх, вийде неадекватною.

3.2 Очищення даних

Майстер обробки складається з трьох частин:

- очищення даних;
- трансформація даних;
- Data Mining;
- інше.

Розглянемо частину очищення даних.

Парціальна передобробка служить для відновлення пропущених даних, редагування аномальних значень і спектральної

обробці даних (наприклад, згладжування даних). Саме цей крок часто проводиться в першу чергу.

Відновлення пропущених даних. Часто буває так, що в стовпці деякі дані відсутні в силу будь-яких причин (дані не відомі, або їх забули внести тощо). Зазвичай через це довелося би прибрати з обробки всі рядки, які містять пропущені дані. Але механізми Deductor Studio дозволяють вирішити цю проблему. Один із кроків парціальної обробки якраз відповідає за відновлення пропущених значень. Якщо дані впорядковані (наприклад, за часом), то рекомендується в якості відновлення пропущених значень використовувати апроксимацію. Алгоритм сам підбере значення, яке має стояти на місці пропущеного значення, ґрунтуючись на прилеглих даних. Якщо ж дані не впорядковані, то слід використовувати режим максимальної правдоподібності, коли алгоритм підставляє замість пропущених даних найбільш ймовірні значення, ґрунтуючись на всій вибірці.

Видалення аномалій.

Аномалії – це відхилення від нормальної поведінки чого-небудь. Це може бути, наприклад, різке відхилення величини від її очікуваного значення.

Автоматичне редагування аномальних значень здійснюється з використанням 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.

По суті аномалії взагалі не повинні чинити жодного впливу на результат. Якщо ж вони присутні при побудові моделі, то надають на неї дуже великий вплив. Тобто попередньо їх необхідно усунути. Також вони псують статистичну картину розподілу даних. Дані з аномаліями, а також гістограма їх розподілу представлені на рис. 3.1

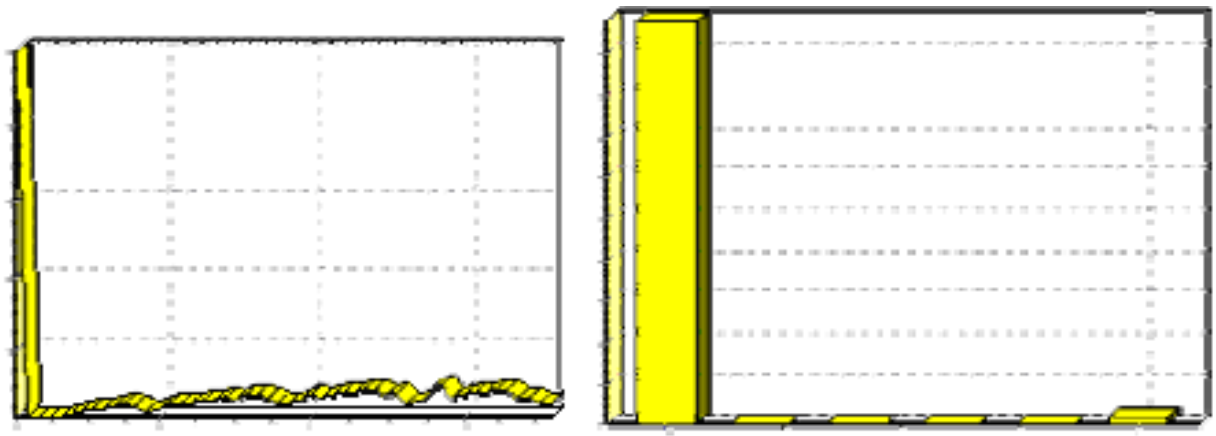


Рис. 3.1 – Дані з аномаліями і гістограма розподілення

Очевидно, що аномалії не дозволяють визначити як характер самих даних, так і статистичну картину. Дані після усунення аномалій представлені на рис. 3.2.

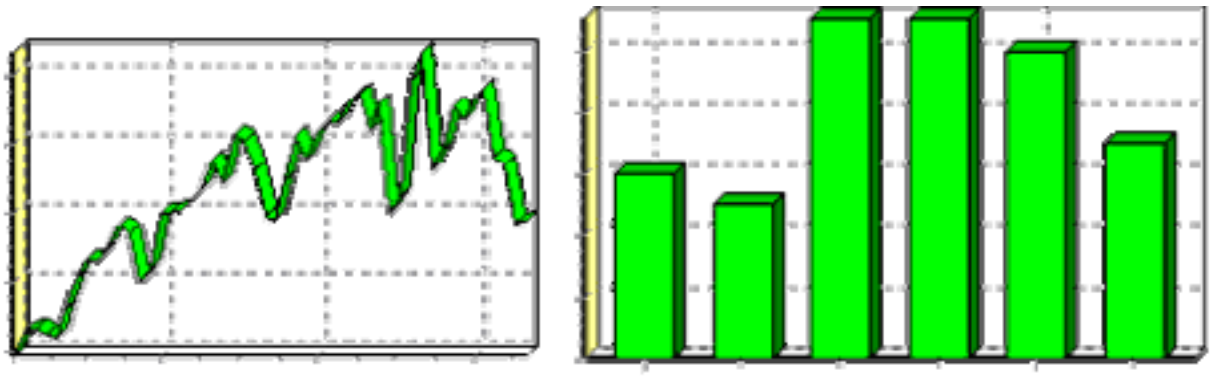


Рис. 3.2 – Результат усунення аномалій

Спектральна обробка. Згладжування даних застосовується для видалення шумів з початкового набору. Платформа Deductor Studio пропонує кілька видів спектральної обробки: згладжування даних шляхом вказівки смуги пропускання, віднімання шуму шляхом вказівки ступеня вирахування шуму і вейвлет перетворення шляхом вказівки глибини розкладання і порядку вейвлета.

Видалення шумів.

Шуми в даних не тільки приховують загальну тенденцію, а й виявляють себе при побудові моделі прогнозу. Через них модель може вийти з поганими узагальнюючими якостями.

Спектральна обробка дозволяє зробити це за допомогою вказівки для цих полів як тип обробки «Віднімання шуму». Налаштування володіють певною гнучкістю. Так, існує велика, середня і мала ступінь вирахування шуму. Аналітик може підібрати ступінь, що влаштовує його.

У деяких випадках непогані результати видалення шумів дає вейвлет перетворення.

3.3 Приклад обробки даних: заповнення пропущених даних та редагування викидів

В якості прикладу, використовуючи табличний редактор MS Excel, створимо таблицю, що містить такі стовбці: Аргумент, Синус, Аномалії, Великі шуми, Середні шуми, Малі шуми.

Стовбцю «Аргумент» присвоюються значення від 0 до 2,96 з кроком 0,02. У стовпці «Синус» приймаються значення синуса аргументу (9 знаків після коми). Значення стовпця «Аномалії» дорівнюють значенням стовпця «Синус». Значення стовпців «Великі шуми», «Середні шуми», «Малі шуми» мають значення близькі до значення синуса аргументу, але мають відхилення в проміжку -1 до 1 для великих шумів, з -0,75 до 0,75 для середніх шумів, з -0,5 до 0,5 для малих шумів (рис. 3.3).

Для будь-яких двадцяти значеннях аргументу, введення даних в значеннях синуса пропустити. Значення стовпця «Аномалії» не мають пропущених даних, однак 10 значень різко відхиляються від істинного значення синуса аргументу. Приклад показаний на рис. 3.4.

Необхідно виконати імпорт даних створеного файлу, обробку даних, відновити пропущені значення синуса, виконати парціальну обробку, видалити аномалії і шуми.

L4 =SIN(A4)+СЛУЧМЕЖДУ(-0,5;0,5)

	G	H	I	J	K	L
1	Аргумент	Синус	Аномалії	Великі шуми	Середні шуми	Малі шуми
2	0	0	0	0	-1	0
3	0,02	0,019998667	0,019998667	0,019998667	0,019998667	0,019998667
4	0,04	0,039989334	0,039989334	-0,960010666	0,039989334	0,039989334
5	0,06	0,059964006	0,059964006	1,059964006	0,059964006	0,059964006
6	0,08	0,079914694	0,079914694	-0,920085306	0,079914694	0,079914694
7	0,1	0,099833417	0,099833417	-0,900166583	1,099833417	0,099833417
8	0,12	0,119712207	0,119712207	-0,880287793	-0,880287793	0,119712207
9	0,14	0,139543115	0,139543115	0,139543115	1,139543115	0,139543115
10	0,16	0,159318207	0,159318207	0,159318207	0,159318207	0,159318207
11	0,18	0,179029573	0,179029573	1,179029573	0,179029573	0,179029573
12	0,2	0,198669331	0,198669331	1,198669331	-0,801330669	0,198669331
13	0,22	0,218229623	0,218229623	0,218229623	1,218229623	0,218229623
14	0,24	0,237702626	0,237702626	1,237702626	-0,762297374	0,237702626
15	0,26	0,257080552	0,257080552	0,257080552	-0,742919448	0,257080552
16	0,28	0,276355649	0,276355649	-0,723644351	-0,723644351	0,276355649
17	0,3	0,295520207	0,295520207	0,295520207	0,295520207	0,295520207
18	0,32	0,314566561	0,314566561	1,314566561	0,314566561	0,314566561
19	0,34	0,333487092	0,333487092	0,333487092	-0,666512908	0,333487092
20	0,36	0,352274233	0,352274233	0,352274233	1,352274233	0,352274233
21	0,38	0,370920469	0,370920469	0,370920469	1,370920469	0,370920469
22	0,4	0,389418342	0,389418342	-0,610581658	0,389418342	0,389418342

Рис. 3.3 – Зразок таблиці формування даних

	A	B	C	D	E	F
1	Аргумент	Синус	Аномалії	Великі шуми	Середні шуми	Малі шуми
2	0	0	0	-1	1	0
3	0,02	0,019998667	0,019998667	0,019998667	0,019998667	0,019998667
4	0,04	0,039989334	0,039989334	0,039989334	-0,960010666	0,039989334
5	0,06	0,059964006	0,059964006	0,059964006	-0,940035994	0,059964006
6	0,08	0,079914694	0,5	-0,920085306	1,079914694	0,079914694
7	0,1		0,099833417	-0,900166583	0,099833417	0,099833417
8	0,12	0,119712207	0,119712207	-0,880287793	0,119712207	0,119712207
9	0,14	0,139543115	0,139543115	1,139543115	-0,860456885	0,139543115
10	0,16		0,159318207	0,159318207	0,159318207	0,159318207
11	0,18	0,179029573	0,179029573	-0,820970427	0,179029573	0,179029573
12	0,2	0,198669331	0,198669331	1,198669331	0,198669331	0,198669331
13	0,22	0,218229623	0,218229623	0,218229623	1,218229623	0,218229623
14	0,24	0,237702626	0,8	0,237702626	0,237702626	0,237702626
15	0,26	0,257080552	0,257080552	1,257080552	-0,742919448	0,257080552
16	0,28	0,276355649	0,276355649	-0,723644351	1,276355649	0,276355649
17	0,3	0,295520207	0,295520207	-0,704479793	1,295520207	0,295520207
18	0,32	0,314566561	0,314566561	1,314566561	1,314566561	0,314566561
19	0,34		0,333487092	0,333487092	0,333487092	0,333487092
20	0,36	0,352274233	0,352274233	1,352274233	-0,647725767	0,352274233
21	0,38	0,370920469	0,370920469	1,370920469	1,370920469	0,370920469
22	0,4	0,389418342	0,389418342	1,389418342	1,389418342	0,389418342
23	0,42	0,407760453	0,407760453	-0,592239547	0,407760453	0,407760453
24	0,44		0,425939465	1,425939465	1,425939465	0,425939465
25	0,46	0,443948107	0,443948107	-0,556051893	1,443948107	0,443948107
26	0,48	0,461779176	0,461779176	-0,538220824	-0,538220824	0,461779176
27	0,5	0,479425539	0,479425539	0,479425539	0,479425539	0,479425539
28	0,52	0,496880138	0,496880138	-0,503119862	1,496880138	0,496880138

Рис. 3.4 – Зразок таблиці з пропущеними даними для синуса та різкими відхиленнями в стовбці «Аномалії»

Робимо експорт сформованих даних (рис. 3.5)

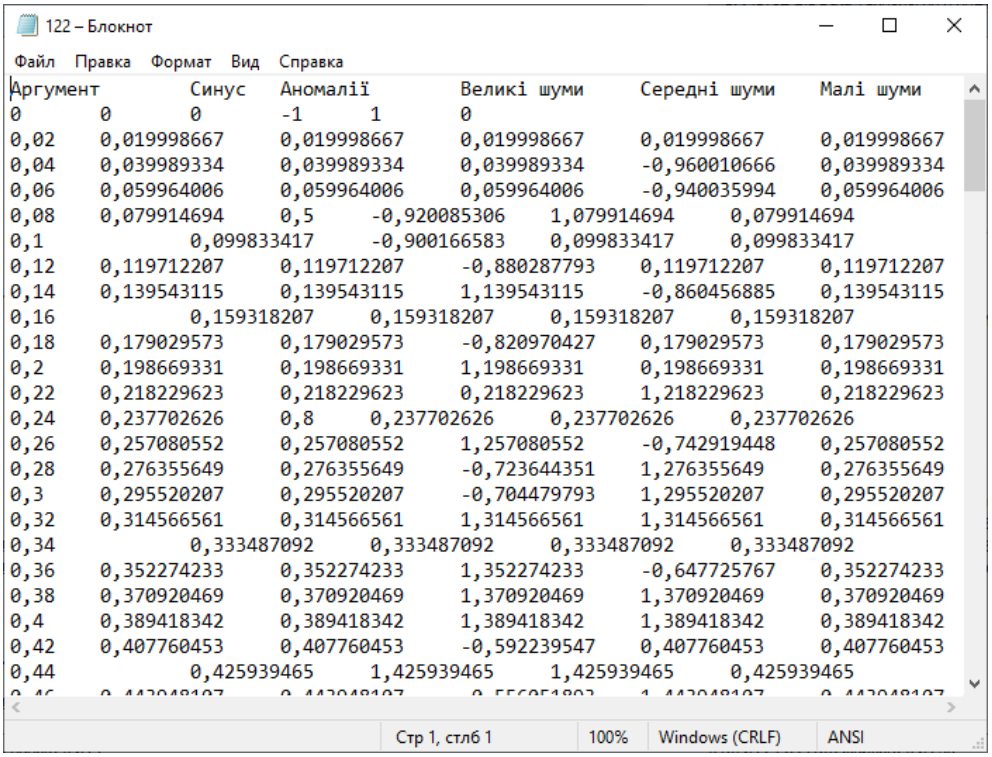


Рис. 3.5 – Експорт даних в «Блокнот»

Здійснюємо імпорт отриманих даних шляхом виклику майстра імпорту на панелі «Сценарії». У вікні перегляду обраного файлу можна побачити зміст даного файлу (рис. 3.6).

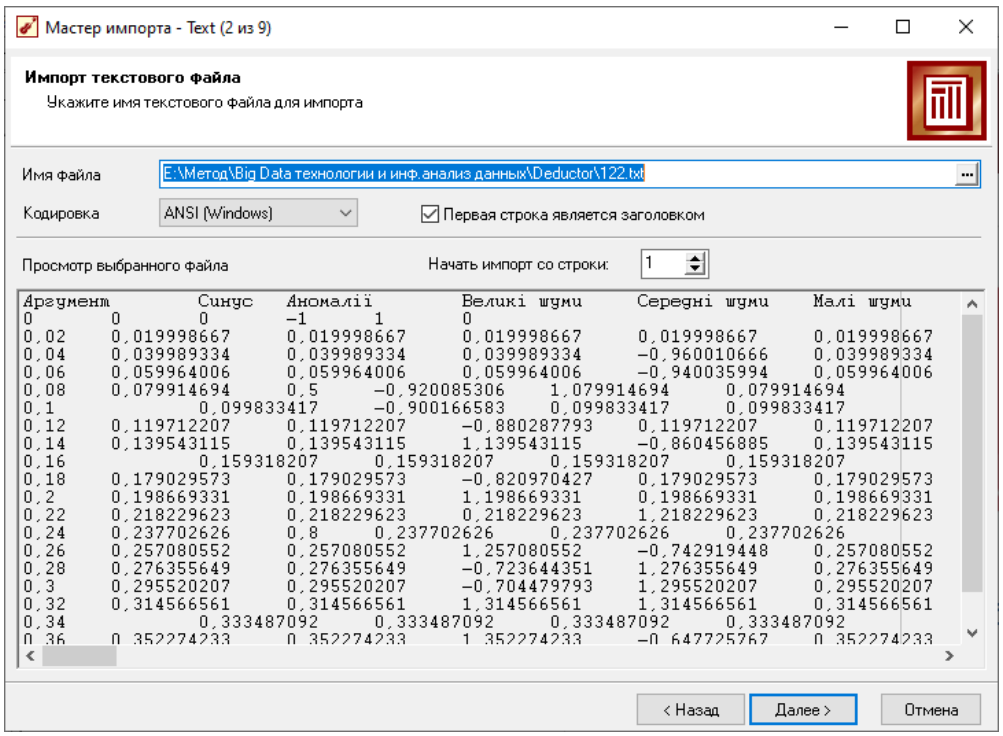


Рис. 3.6 – Імпорт текстового файлу

Налаштування параметрів імпорту (рис. 3.7, 3.8).

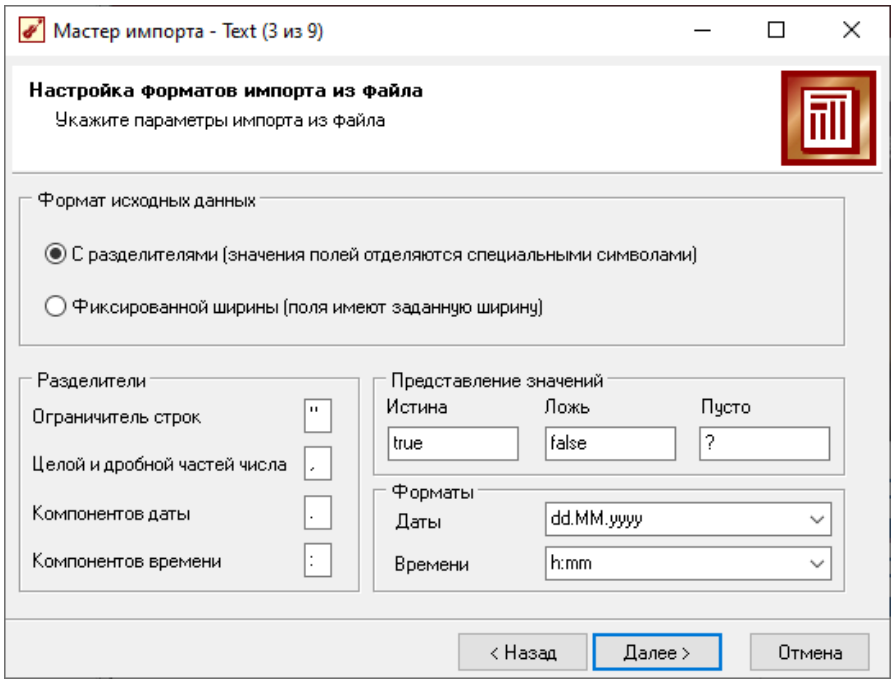


Рис. 3.7 – Налаштування форматів імпорту з файлу

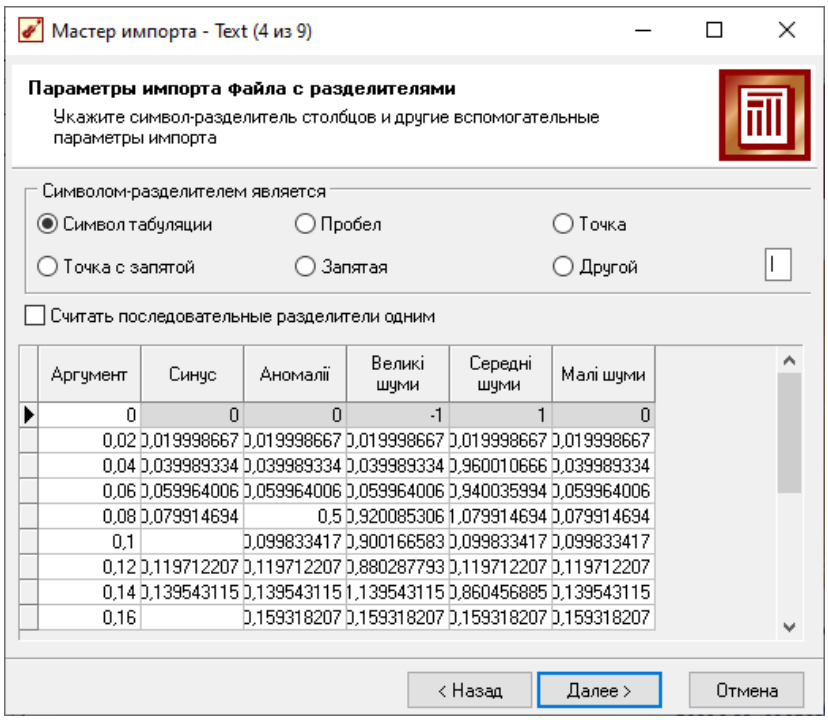


Рис. 3.8 – Параметри імпорту файлу з розділювачами

Налаштувати ім'я, назву (мітку), розмір, тип даних, вид даних і призначення (рис. 3.9).

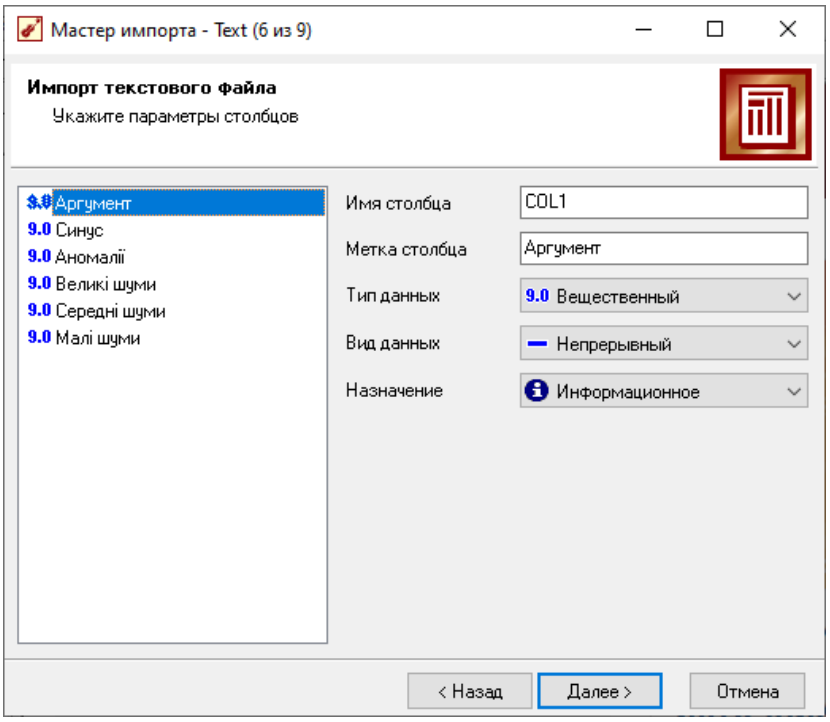


Рис. 3.9 – Імпорт текстового файлу

Далі натискаємо кнопку «Пуск» та імпортуємо дані.

Після імпорту даних необхідно вибрати спосіб відображення даних (рис. 3.10). В даному випадку найбільш інформативним є діаграма, вибираємо її.

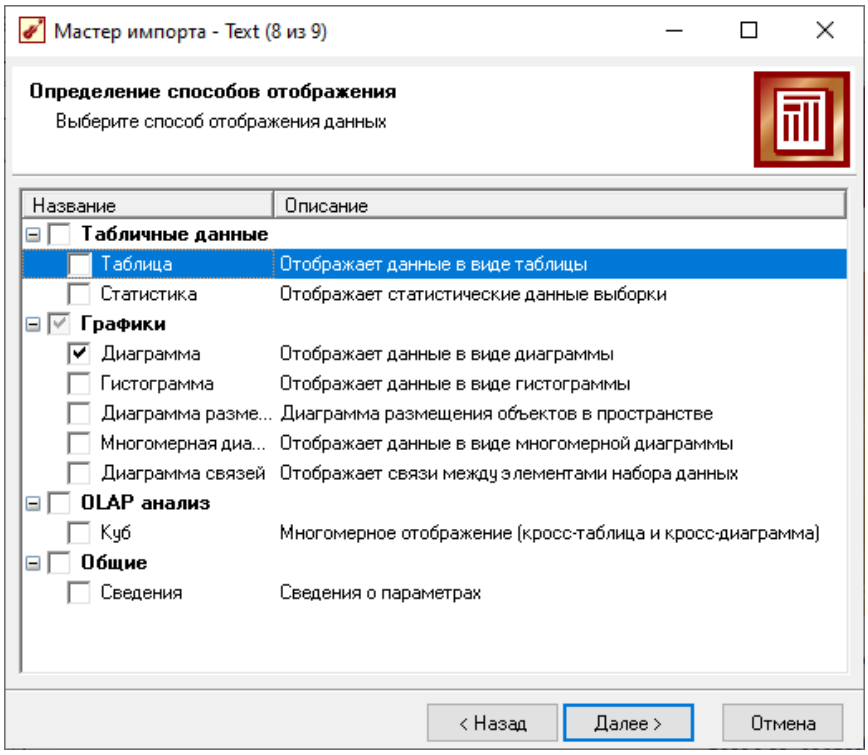


Рис. 3.10 – Визначення способу відображення

Необхідно налаштувати, які стовпці діаграми слід відображати і як саме. Виберемо для відображення поле «СИНУС» (рис. 3.11) і тип діаграми «Лінії».

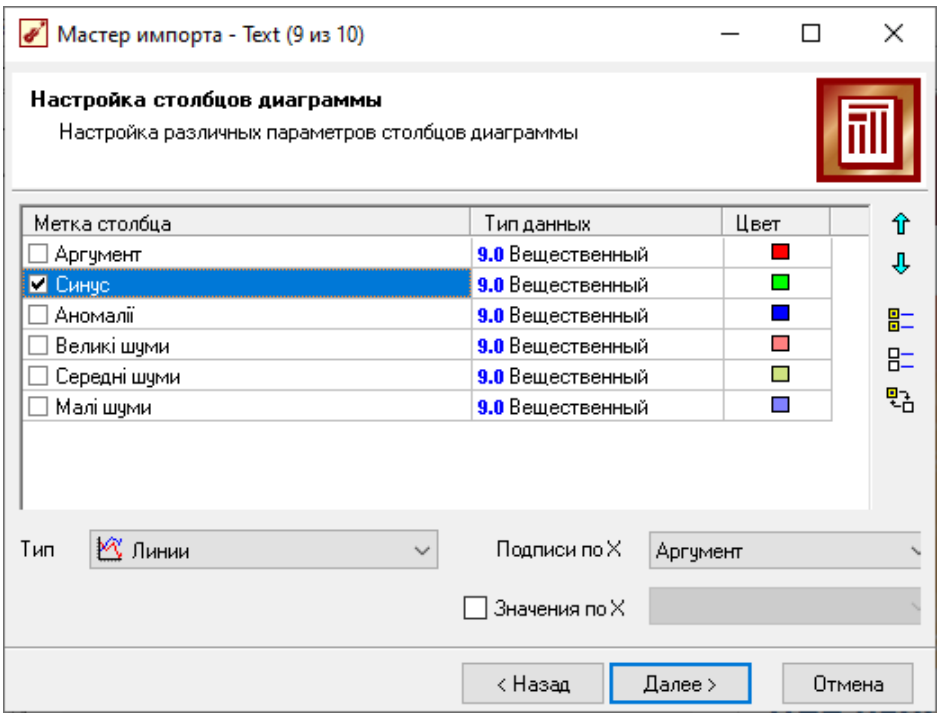


Рис. 3.11 – Налаштування стовбців діаграми

На останньому кроці майстра необхідно вказати назву гілки в дереві сценаріїв. На цьому робота майстра імпорту закінчується. Тепер в дереві сценаріїв з'явиться новий вузол з необхідними даними. У головному вікні програми представлені всі вибрані відображення даних цього вузла. В даному випадку тільки діаграма (рис. 3.12)

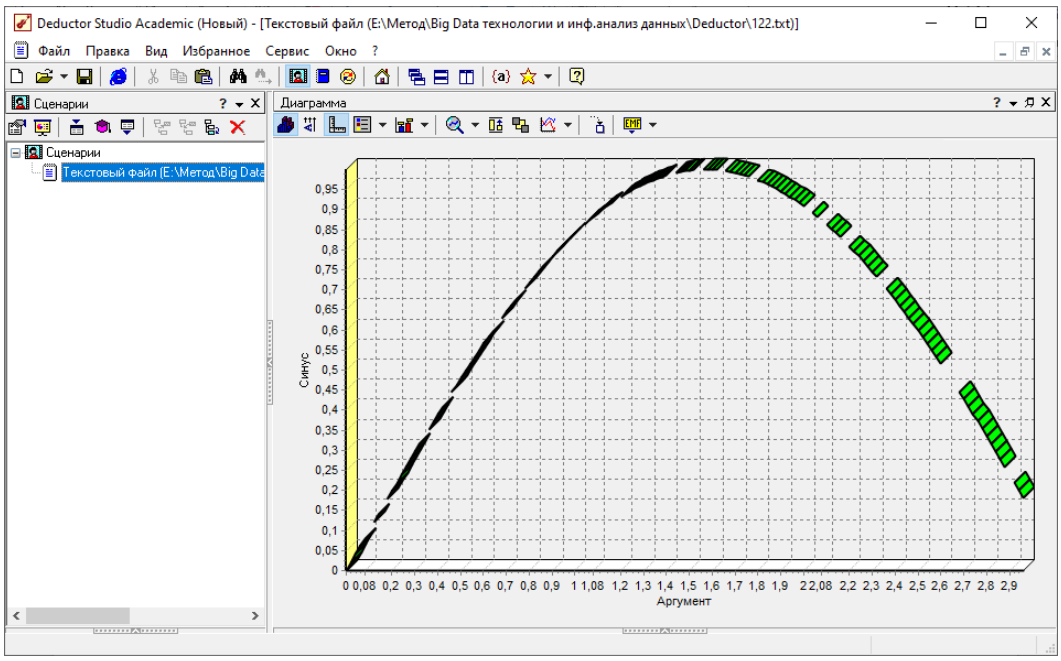
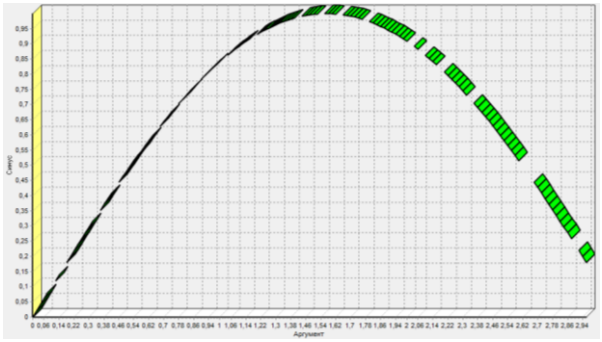


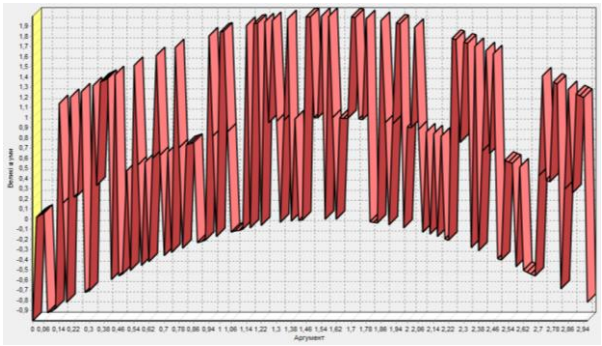
Рис. 3.12 – Результат імпорту даних у вигляді діаграми

Далі зробимо обробку даних з сформованого файлу. Він містить таблицю з наступними полями: «АРГУМЕНТ» – аргумент, «СИНУС» – значення синуса аргументу (деякі значення порожні), «АНОМАЛІЇ» – синус з викидами, «ВЕЛИКІ ШУМИ» – значення синуса з великими шумами, «СЕРЕДНІ ШУМИ» – значення синуса із середніми шумами, «МАЛІ ШУМИ» – значення синуса з малими шумами. Всі дані можна побачити на діаграмі після імпорту з текстового файлу (рис. 3.13).

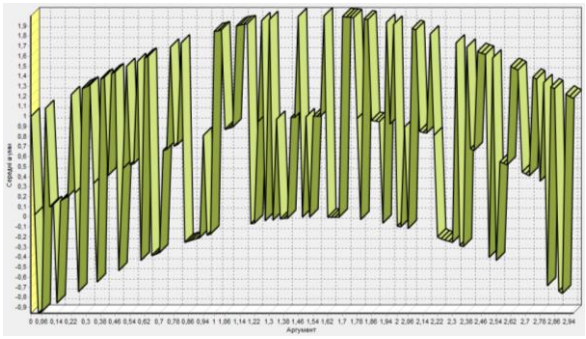


а)

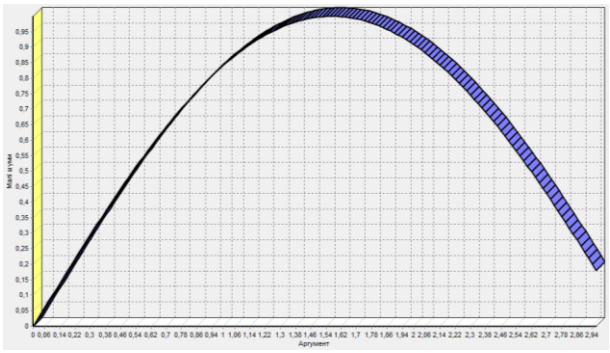
б)



в)



г)



д)

а) стовпець з пропущеними даними;
б) стовпець з аномаліями;
в) стовпець з великими шумами;
г) стовпець з середніми шумами;
д) стовпець з малими шумами

Рис. 3.13 – Діаграми, отримані в результаті імпорту даних

Імпортувавши файл можна побачити, що в стовпці «СИНУС» містяться порожні значення. На діаграмі вище видно, що деякі значення синуса пропущені. Для подальшої обробки необхідно їх відновити. Для цього слід запустити майстер парціальної обробки. У

вікні «Сценарії» обираємо гілку «Текстовий файл» та натискаємо кнопку «Майстер обробки» (рис. 3.14).

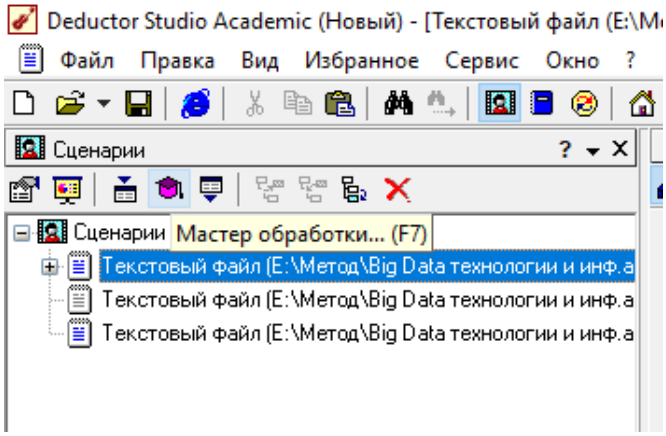


Рис. 3.14 – Запуск «Майстра обробки»

Виконуємо вибір обробки за необхідністю. В даному випадку обираємо «Заповнення пропущених даних» (рис. 3.15, 3.16).

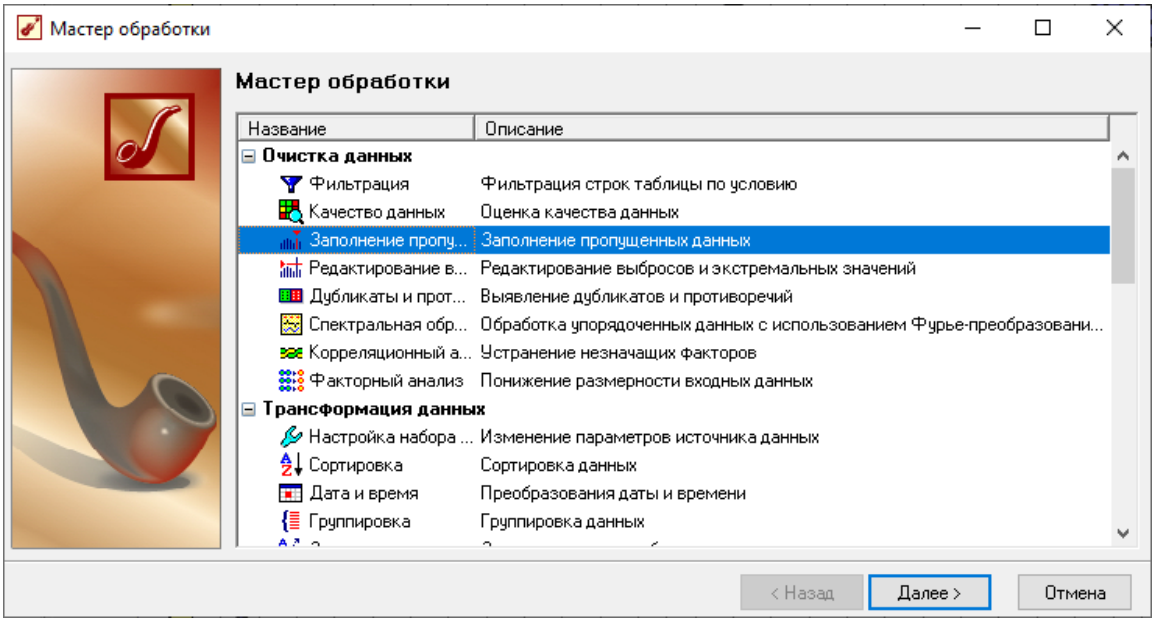


Рис. 3.15 – Вибір методу обробки даних

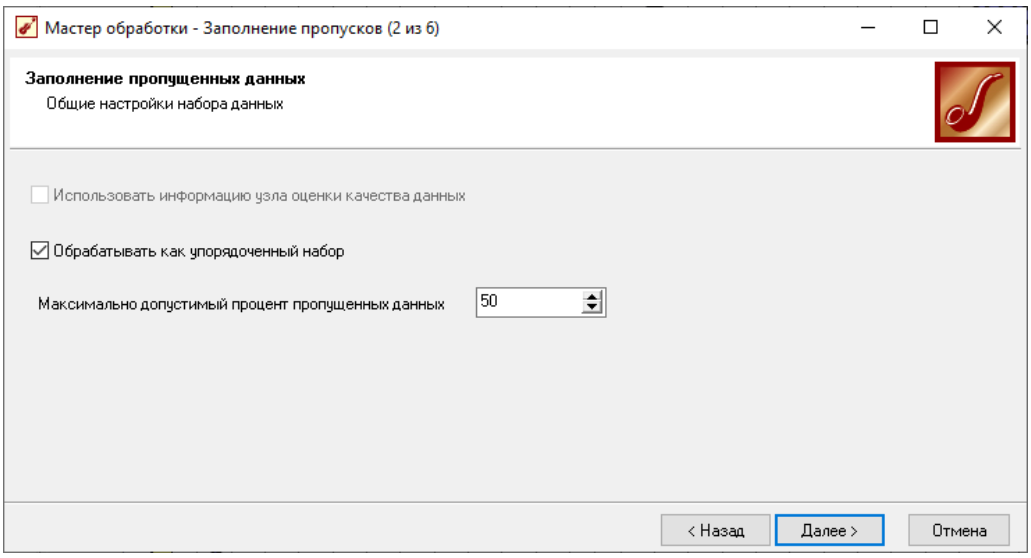


Рис. 3.16 – Загальні налаштування набору даних для обробки

Після виконання процесу обробки, як видно з рисунку 3.17, на діаграмі пропуски в даних зникли, що і було необхідно зробити.

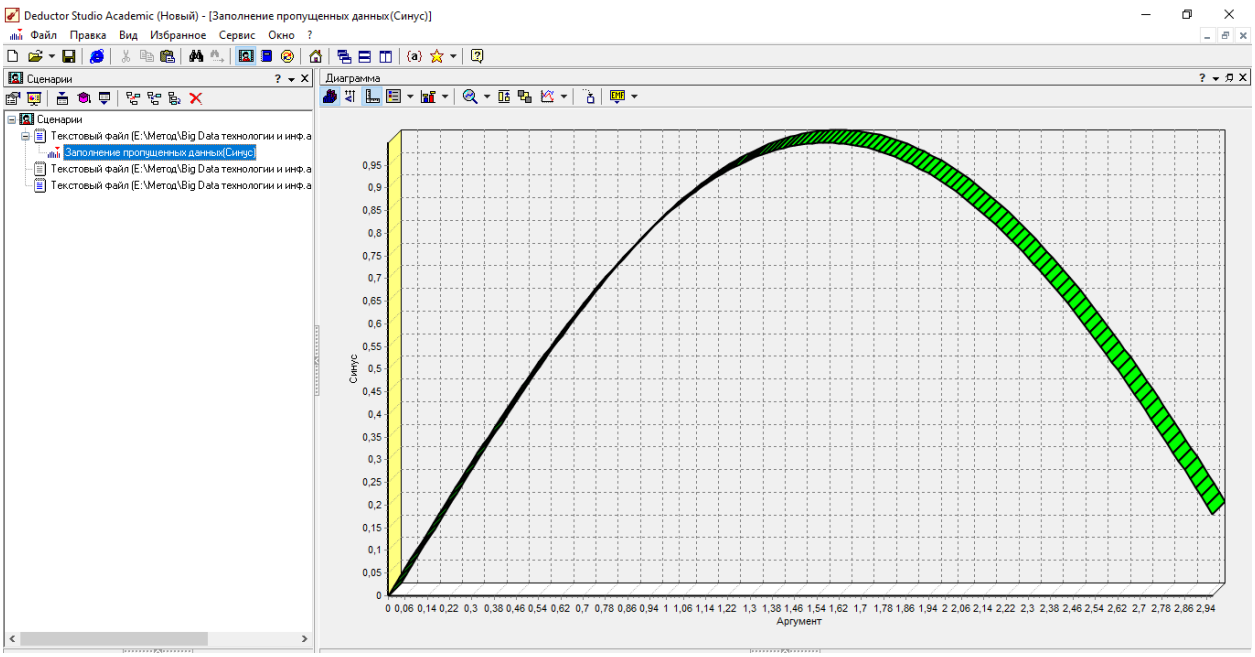


Рис. 3.15 – Результат виконання обробки «Заповнення пропущених даних»

Далі видалимо аномалії з поля «АНОМАЛІЇ» імпортованої таблиці.

У майстрі парціальної передобробки на третьому кроці вибираємо поле «АНОМАЛІЇ» і вказуємо йому тип обробки «Редагування викидів і екстремальних значень», ступінь придушення «Велика» (рис. 3.16 – 3.18).

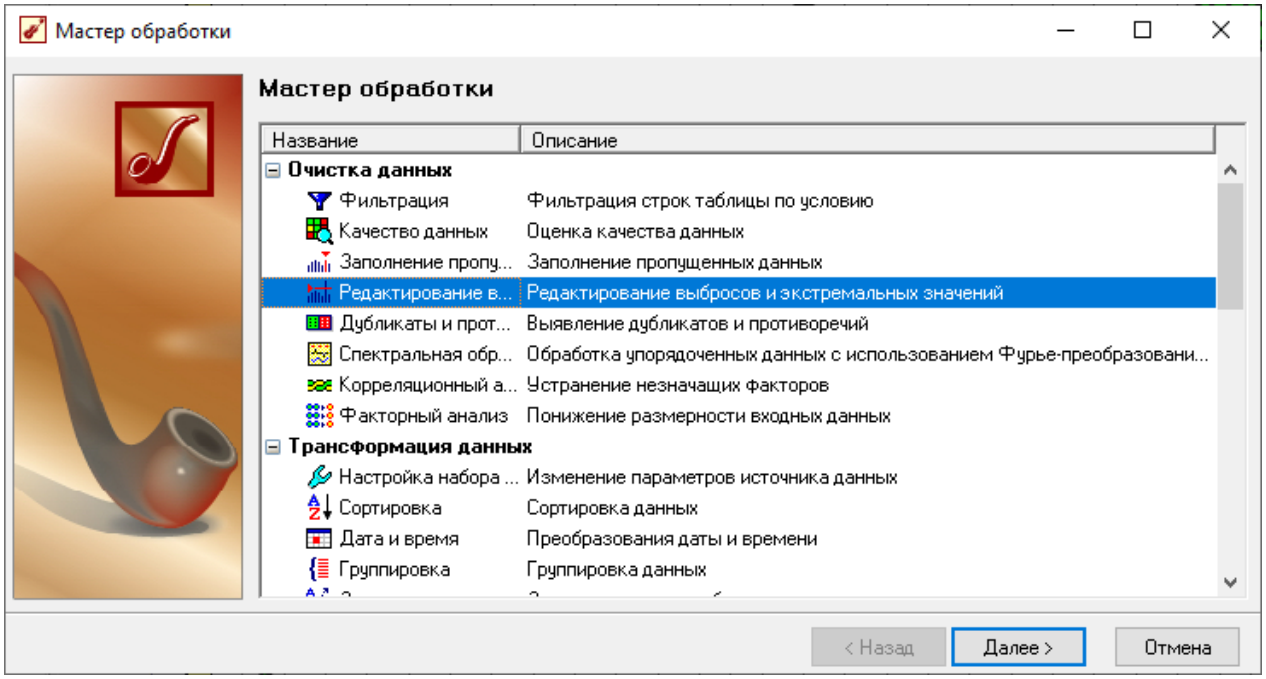


Рис. 3.16 – Вибір типу обробки

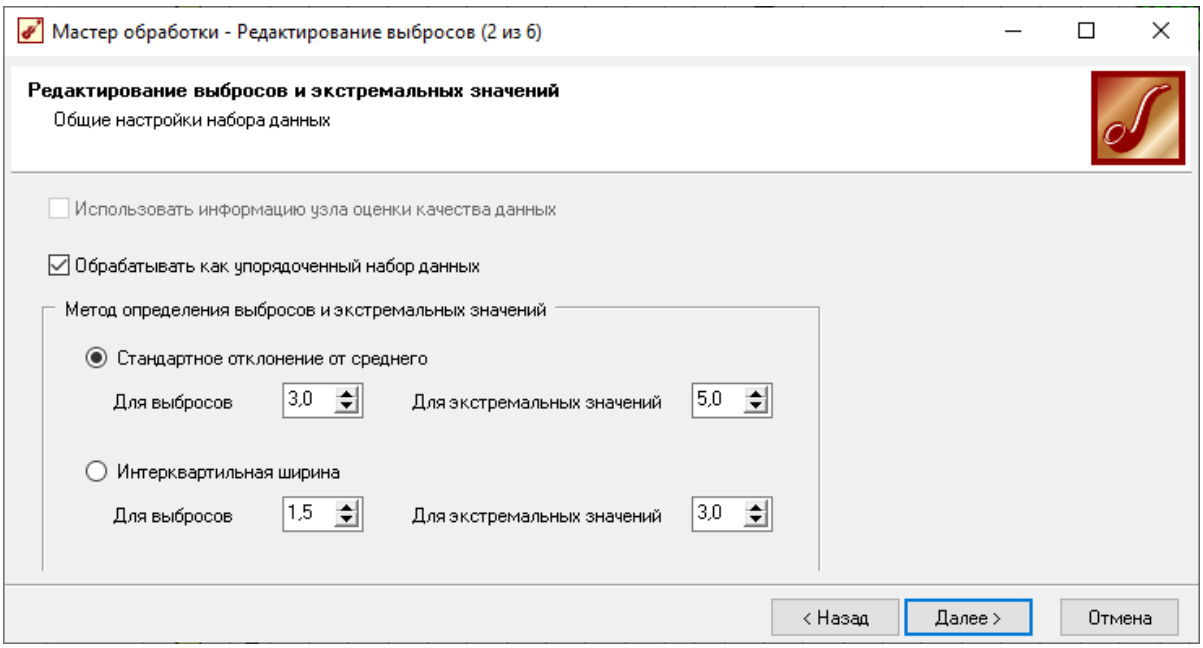


Рис. 3.17 – Вибір методу визначення викидів
та екстремальних значень

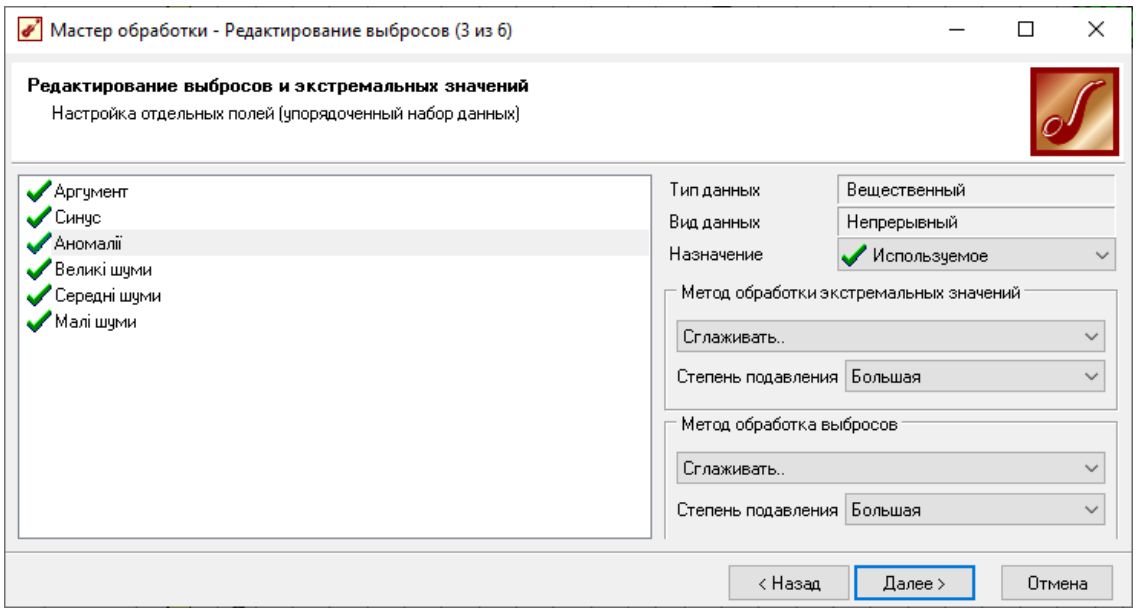


Рис. 3.18 – Налаштування окремих полів

Так як більше ніяких обробок не планувалася, то переходимо на крок запуску процесу обробки і натискаємо «Пуск».

Після виконання процесу обробки на діаграмі видно, що викиди зникли, залишилися лише невеликі обурення, які легко згладити за допомогою спектральної обробки.

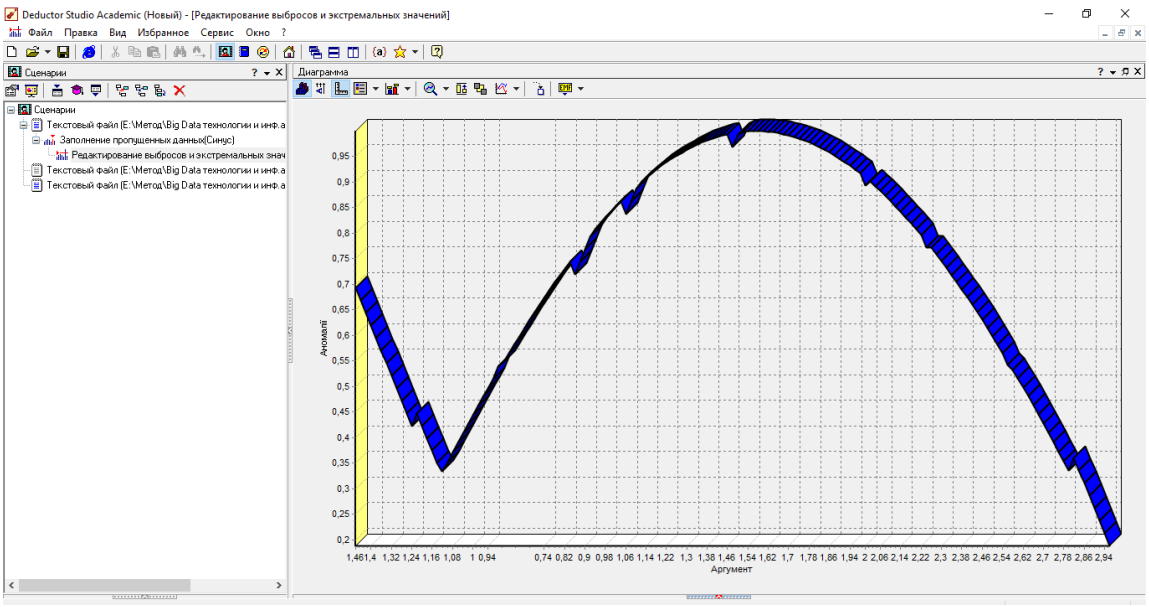


Рис. 3.19 – Результат обробки викидів

Як видно на рисунку 3.19, аномалії були усунені, однак невеликі обурення залишилися. Згладимо їх за допомогою парціальної обробки. Для цього після видалення аномалій знову запустимо майстер парціальної обробки. У ньому на другому кроці виберемо поле «АНОМАЛІЇ» і вкажемо йому тип обробки «Спектральна обробка», що обробляє упорядковані дані з використанням Фур'є-перетворення і вейвлет перетвореня, з параметрами за замовчуванням (глибина розкладання 3, порядок вейвлета 6) (рис. 3.20 – 3.22).

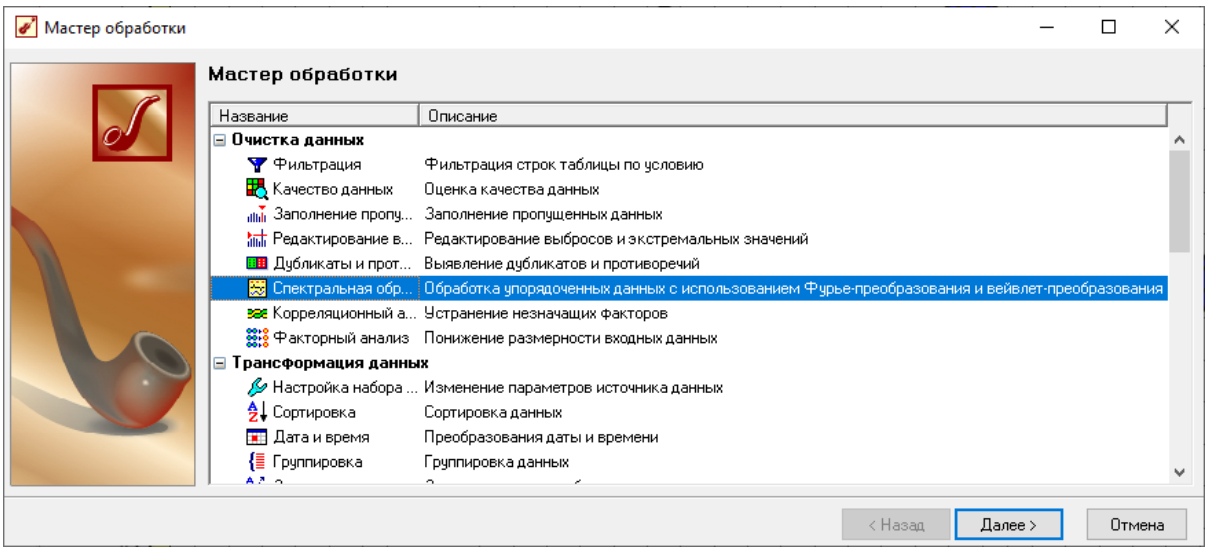


Рис. 3.20 – Вибір методу обробки

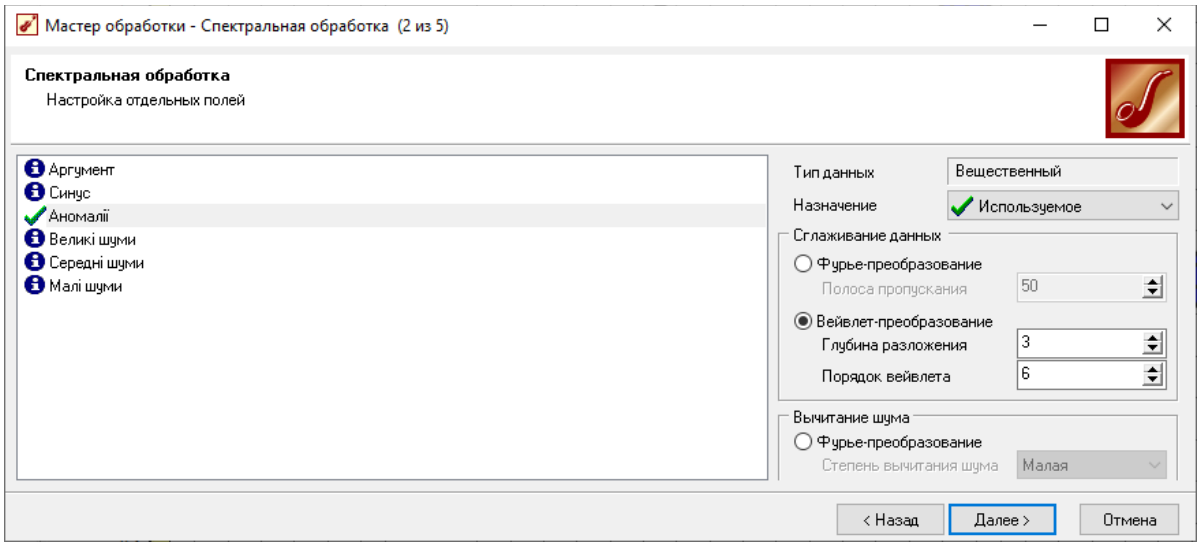


Рис. 3.21 – Вибір і налаштування окремих полів

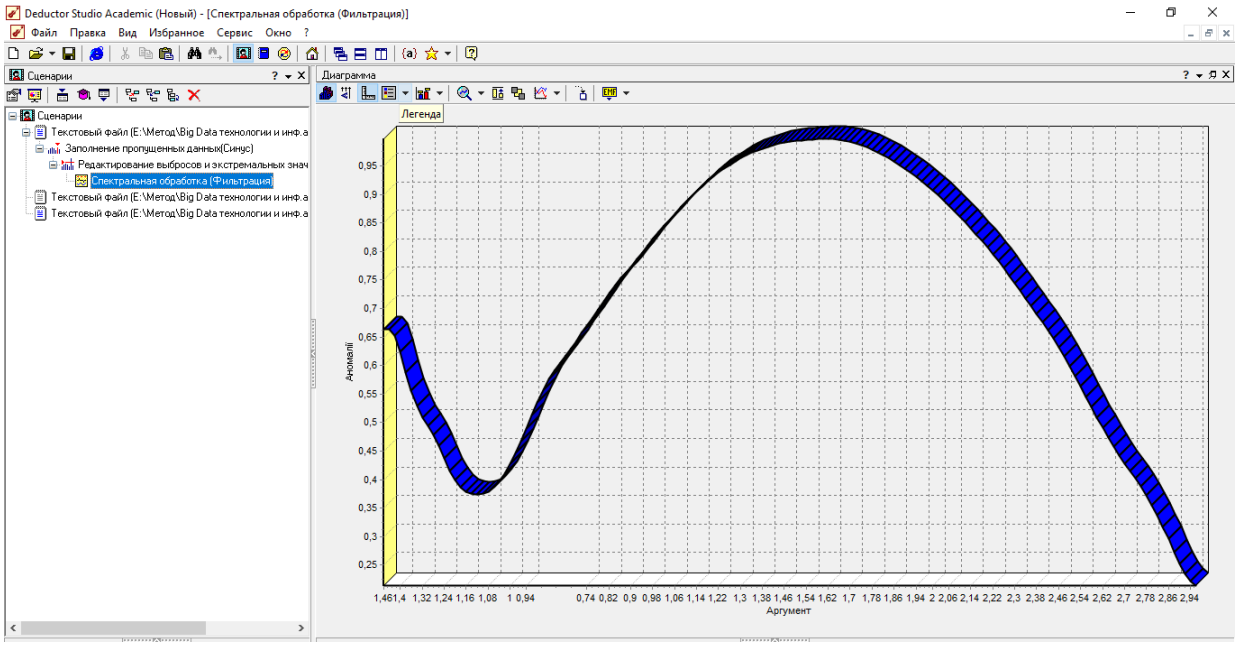


Рис. 3.22 – Результат спектральної обробки поля «Аномалії»

У прикладі з парціальної обробці, як було показано раніше, є 3 стовпці з шумами: «ВЕЛИКІ ШУМИ», «СЕРЕДНІ ШУМИ», і «МАЛІ ШУМИ» - відповідно синус з великими, середніми і малими шумами. Ясно, що для подальшої роботи з даними ці шуми необхідно усунути.

Таким чином, в майстра парціальної обробки на другому кроці виберемо по черзі поля «ВЕЛИКІ ШУМИ», «СЕРЕДНІ ШУМИ» і «МАЛІ ШУМИ», задамо тип обробки «Віднімання шуму» і вкажемо ступінь придушення - «велика», «середня» і « мала »відповідно. Після виконання обробки на діаграмі можна переглянути отримані результати (рис. 3.23).

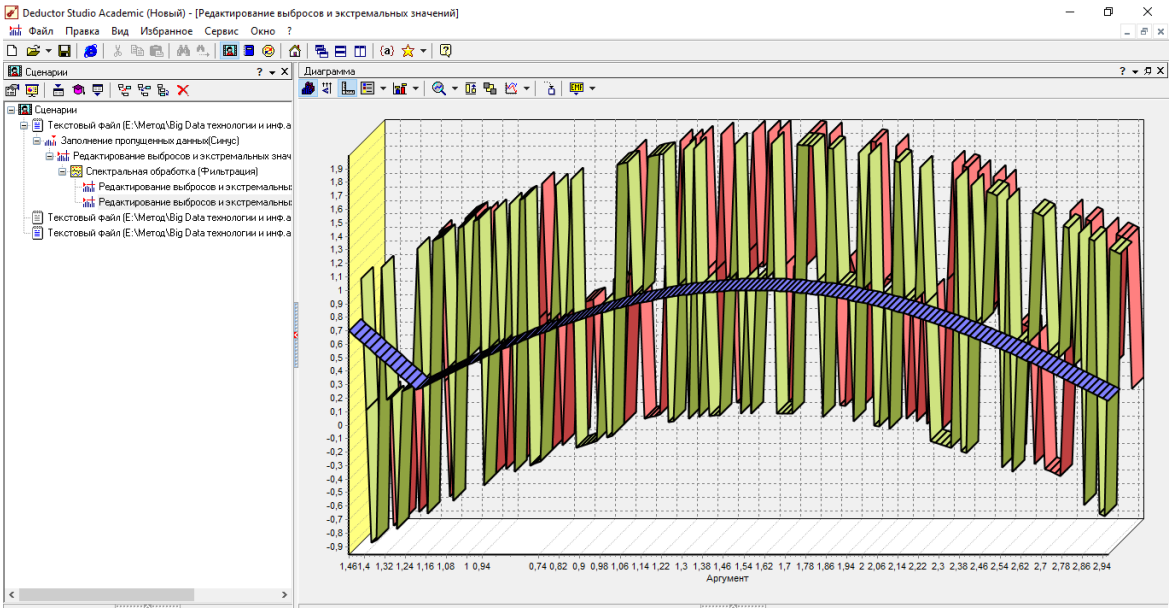


Рис. 3.23 – Результат обробки шумів

Тепер видалимо шуми за допомогою вейвлет перетворення. У майстра парціальної обробки виберемо поля «ВЕЛИКІ ШУМИ», «СЕРЕДНІ ШУМИ» і «МАЛІ ШУМИ», вкажемо тип обробки «Спектральна обробка», залишивши параметри обробки за замовчуванням (глибина розкладання – 3, порядок вейвлета – 6). На діаграмі можна переконатися в тому, що дані згладилися (рис. 3.24, 3.25, 3.26). Підвищити якість згладжування шумів таким способом можна, шляхом підбору задовільних параметрів обробки.

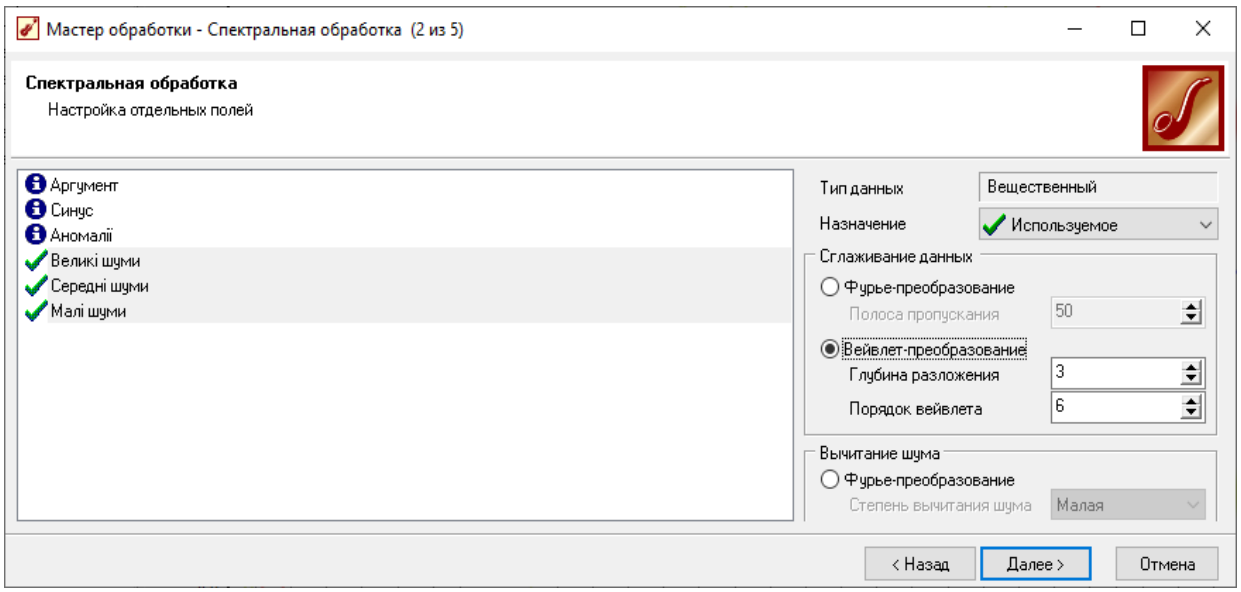


Рис. 3.24 – Вибір і налаштування обробки окремих полів

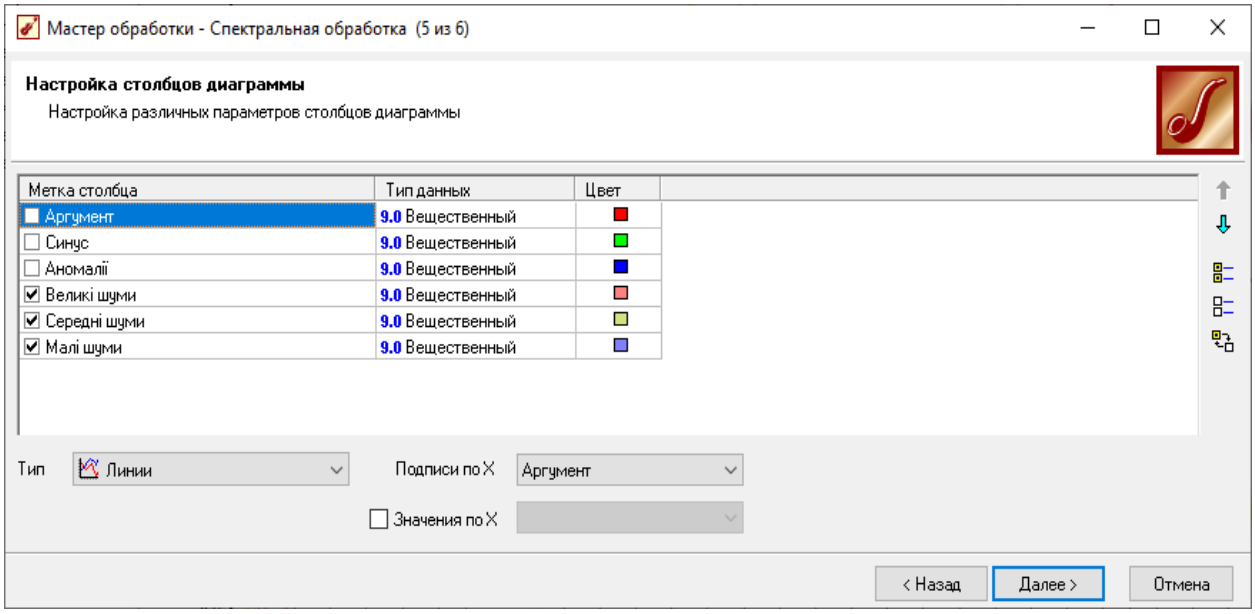


Рис. 3.25 – Налаштування стовбців діаграми

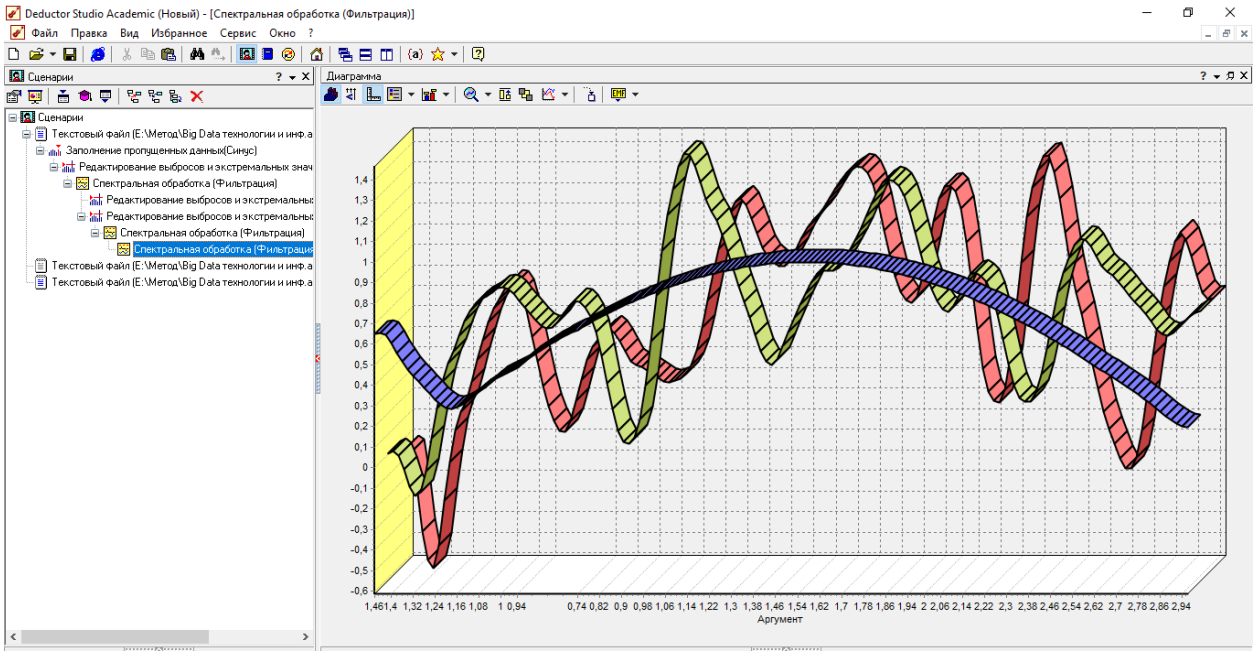


Рис. 3.26 – Результат згладжування шумів

Ознайомилися з методами обробки Deductor Studio Academic, здобули навички парціальної передобробки, відновлення пропущених даних, видалення аномалій, спектральної обробки, видалення шумів.

Питання до розділу 3

1. Для чого призначено майстер імпорту програми Deductor Studio?
2. Для чого призначений майстер обробки програми Deductor Studio?
3. Для чого призначений майстер відображення програми Deductor Studio?
4. Навіщо слід проводити підготовку даних для аналізу?
5. Що таке шуми та аномалії в даних?

6. Якими методами можна забрати шуми в системі Deductor?
7. Якими методами можна прибрати аномалії даних у системі Deductor?
8. Для чого використовується парціальна передобробка?
9. Навіщо використовується спектральна обробка?
10. Які види спектральної обробки є у системі Deductor?

Розділ 4
ОБРОБКА ДАНИХ ПРИ ФАКТОРНОМУ ТА КОРЕЛЯЦІЙНОМУ АНАЛІЗІ

4.1 Факторний і кореляційний аналізи даних

В майстрі обробки платформи Deductor Studio Academic до методів очищення даних відноситься факторний та кореляційний аналізи даних.

Факторний аналіз. Факторний аналіз – група методів багатовимірного статистичного аналізу, які дозволяють представити в компактній формі узагальнену інформацію про структуру зв'язків між ознаками досліджуваного об'єкта, що спостерігаються, на основі виділення деяких факторів, які безпосередньо не спостерігаються. Факторний аналіз служить для зниження розмірності простору вхідних факторів. Обробку можна виконувати як в автоматичному режимі (з зазначенням порога значимості), так і самостійно (ґрунтуючись на значеннях матриці значущості).

Першим етапом факторного аналізу є вибір нових ознак, які є лінійними комбінаціями колишніх і «вбирають» в себе більшу частину загальної мінливості вхідних факторів. Тому вони містять велику частину інформації, що містяться в первинних даних. В обробнику «Факторний аналіз» здійснюється за допомогою методу головних компонент. Цей метод зводиться до вибору нової ортогональної системи координат в просторі спостережень. У якості

першої головної компоненти обирають напрямок, уздовж якого масив даних має найбільший розкид. Вибір кожної наступної головної компоненти відбувається так, щоб розкид даних уздовж неї був максимальним і щоб ця головна компонента була ортогональна іншим головним компонентам, обраним раніше.

Кореляційний аналіз. Кореляційний аналіз – сукупність заснованих на математичній теорії кореляції методів виявлення кореляційної залежності між двома випадковими ознаками або факторами. Кореляційний аналіз застосовується для оцінки залежності вихідних полів даних від вхідних факторів і усунення незначних факторів. Принцип кореляційного аналізу полягає в пошуку таких значень, які в найменшій мірі 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 значимості. Якщо ступінь взаємозалежності між вхідним і вихідним факторами менше порога значимості, то відповідний фактор відкидається як незначний.

4.2 Оцінка якості даних

Для прикладу використовуємо файл з даними, що сформований раніш і містить стовбці «Аргумент», «Синус», «Аномалії», «Великі шуми», «Середні шуми», «Малі шуми» (рис. 4.1). Експортуємо дані в текстовий файл.

Необхідно виконати обробку даних факторним і кореляційним аналізом.

	A	B	C	D	E	F
1	Аргумент	Синус	Аномалії	Великі шуми	Середні шуми	Малі шуми
2	0	0	0	-1	1	0
3	0,02	0,019998667	0,019998667	0,019998667	0,019998667	0,019998667
4	0,04	0,039989334	0,039989334	0,039989334	-0,960010666	0,039989334
5	0,06	0,059964006	0,059964006	0,059964006	-0,940035994	0,059964006
6	0,08	0,079914694	0,5	-0,920085306	1,079914694	0,079914694
7	0,1		0,099833417	-0,900166583	0,099833417	0,099833417
8	0,12	0,119712207	0,119712207	-0,880287793	0,119712207	0,119712207
9	0,14	0,139543115	0,139543115	1,139543115	-0,860456885	0,139543115
10	0,16		0,159318207	0,159318207	0,159318207	0,159318207
11	0,18	0,179029573	0,179029573	-0,820970427	0,179029573	0,179029573
12	0,2	0,198669331	0,198669331	1,198669331	0,198669331	0,198669331
13	0,22	0,218229623	0,218229623	0,218229623	1,218229623	0,218229623
14	0,24	0,237702626	0,8	0,237702626	0,237702626	0,237702626
15	0,26	0,257080552	0,257080552	1,257080552	-0,742919448	0,257080552
16	0,28	0,276355649	0,276355649	-0,723644351	1,276355649	0,276355649
17	0,3	0,295520207	0,295520207	-0,704479793	1,295520207	0,295520207
18	0,32	0,314566561	0,314566561	1,314566561	1,314566561	0,314566561
19	0,34		0,333487092	0,333487092	0,333487092	0,333487092
20	0,36	0,352274233	0,352274233	1,352274233	-0,647725767	0,352274233
21	0,38	0,370920469	0,370920469	1,370920469	1,370920469	0,370920469
22	0,4	0,389418342	0,389418342	1,389418342	1,389418342	0,389418342
23	0,42	0,407760453	0,407760453	-0,592239547	0,407760453	0,407760453
24	0,44		0,425939465	1,425939465	1,425939465	0,425939465
25	0,46	0,443948107	0,443948107	-0,556051893	1,443948107	0,443948107
26	0,48	0,461779176	0,461779176	-0,538220824	-0,538220824	0,461779176
27	0,5	0,479425539	0,479425539	0,479425539	0,479425539	0,479425539
28	0,52	0,496880138	0,496880138	-0,503119862	1,496880138	0,496880138

Рис. 4.1 – Приклад заповнення файлу з даними

Виконаємо обробку даних з текстового файлу за допомогою факторного аналізу. Він містить таблицю з наступними полями: «Аргумент» – інформаційне; «Великі шуми», «Середні шуми», «Малі шуми» – вхідні значення; «Аномалії» – вихідні значення; «Синус» – інформаційне (рис. 4.2).

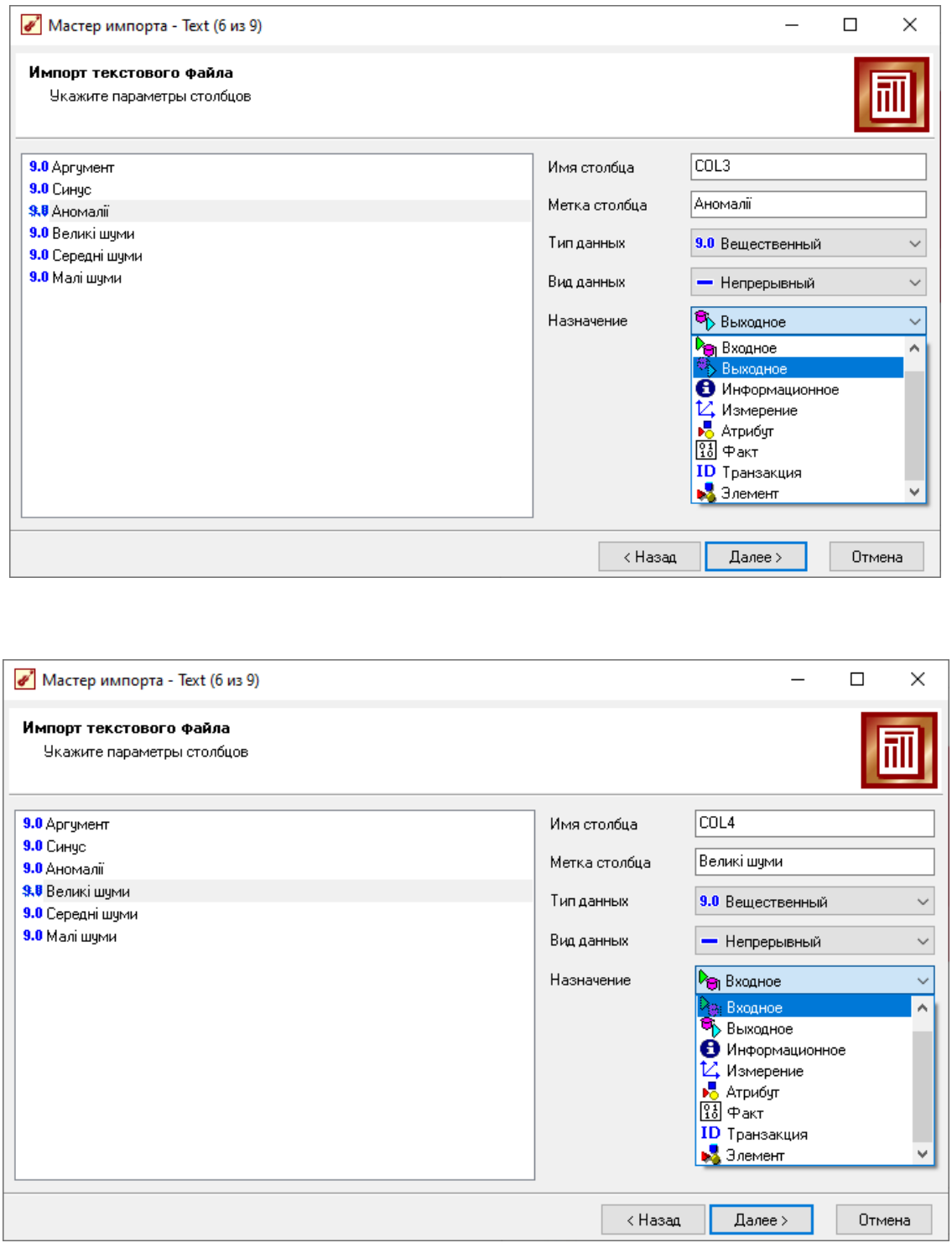
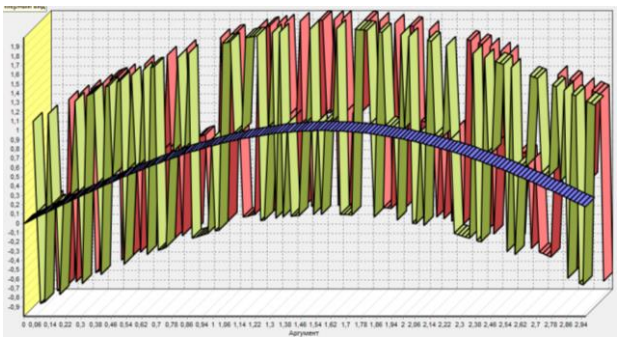
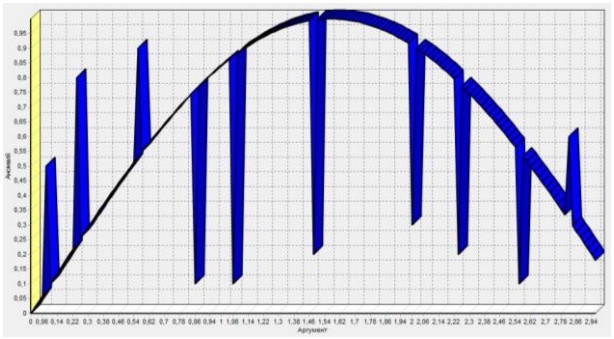


Рис. 4.2 – Призначення параметрів стовбців

На рисунку 4.3 показані вхідні та вихідні дані для обробки.



а) Вхідні дані



б) Виход

Рис. 4.3 – Дані для обробки

В майстрі обробки обираємо «Якість даних», «Аргумент» і «Синус» не використовуємо, та оцінюємо якість даних (рис. 4.4).

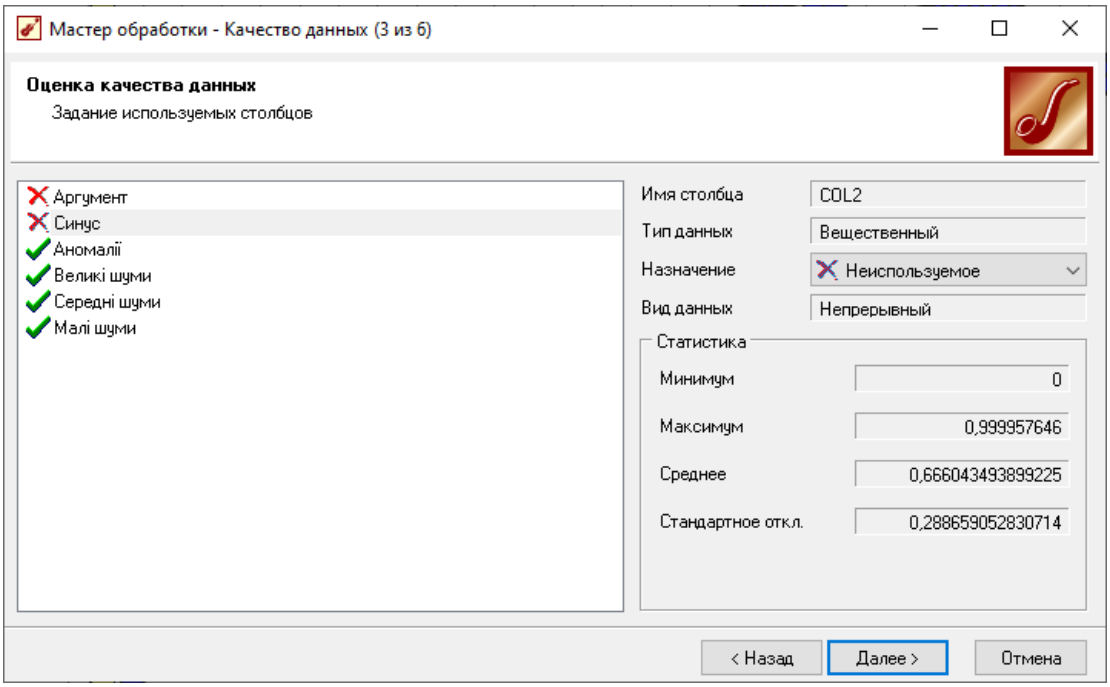


Рис. 4.4 – Завдання використовуваних стовбців

Результат оцінки якості даних для обробки факторним аналізом показаний на рисунку 4.5.

№	Столбец	Тип данных	Вид данных	Пропуски		Выбросы		Экстремальные		Кол-во уникальных	Качество данных	Резюме
				Кол-во	Действие	Кол-во	Действие	Кол-во	Действие			
✔ 1	Аномалії	9.0 Веществе...	Непрерыв...								0,9209	Пригоден
2	Великі шуми	9.0 Веществе...	Непрерыв...								0,9425	Пригоден
3	Середні шуми	9.0 Веществе...	Непрерыв...								0,9366	Пригоден
4	Малі шуми	9.0 Веществе...	Непрерыв...								0,9112	Пригоден

Рис. 4.5 – Результат оцінки якості даних

Дані, що прийнято використовувати для факторного і кореляційного аналізу, придатні для подальшої обробки.

4.3 Обробка даних за допомогою факторного аналізу

В майстрі обробки обираємо «Факторний аналіз» та задамо «Великі шуми», «Середні шуми», «Малі шуми» вхідними полями, «Аномалії» – інформаційне, поле «Аргумент» – невикористовуваним, а поле «Синус» – непридатним (рис. 4.6).

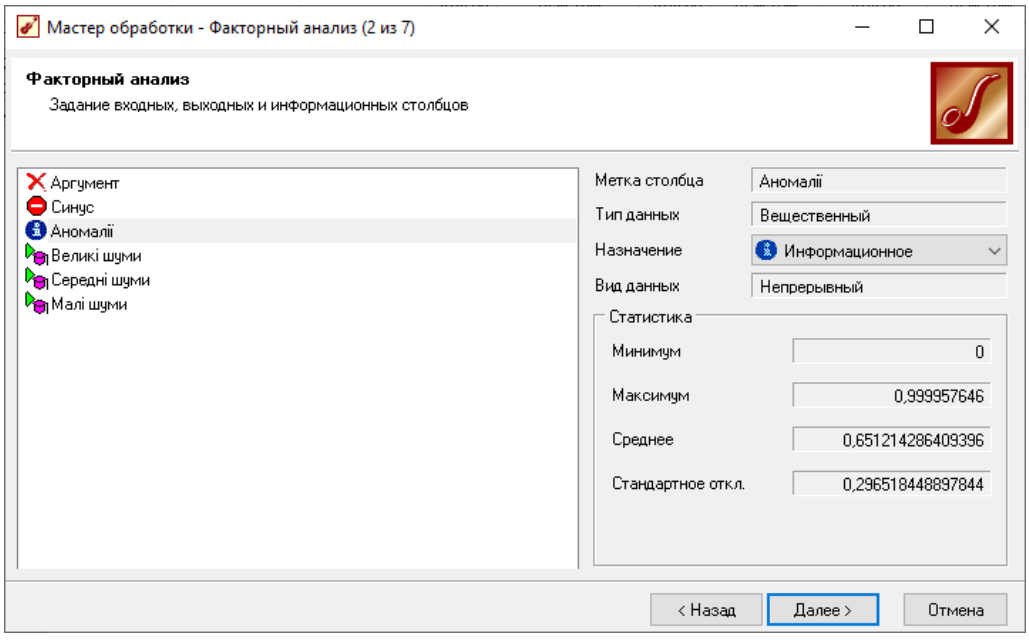


Рис. 4.6 – Завдання призначення стовбців

Наступний крок пропонує запустити процес зниження розмірності простору вхідних факторів. Після завершення процесу на наступному кроці вибираємо, які з отриманих в результаті обробки фактори залишити для подальшої роботи (рис. 4.7). Це робиться шляхом вказівки необхідного порога значимості, який має дорівнювати 90%.

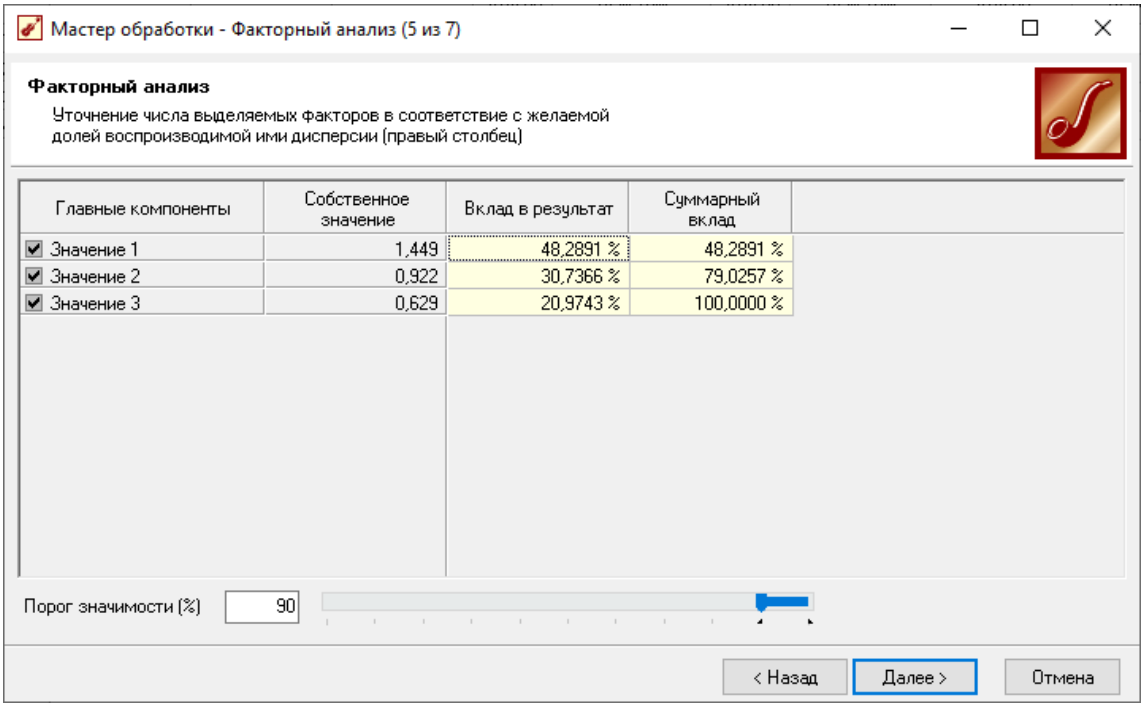


Рис. 4.7 – Вказівка порога значимості

Тепер необхідно перейти на наступний крок і вибрати спосіб візуалізації, вибираємо діаграму.

Далі запускаємо обробку даних факторним аналізом та отримуємо результати (рис. 4.9, 4.10).

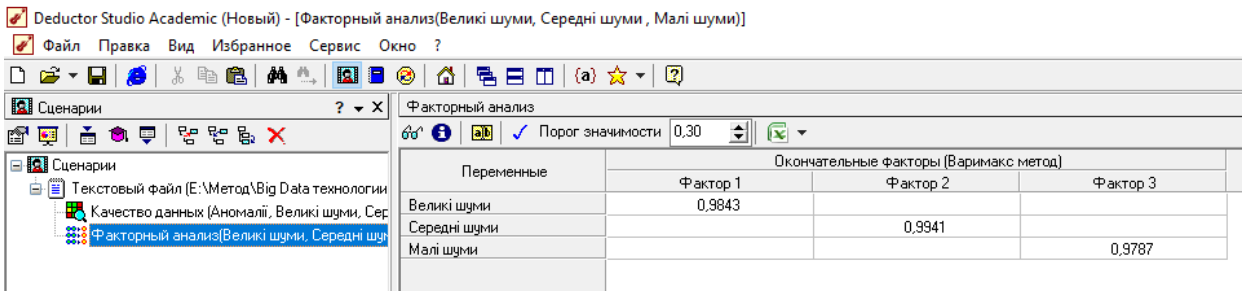


Рис. 4.8 – Результати факторного аналізу

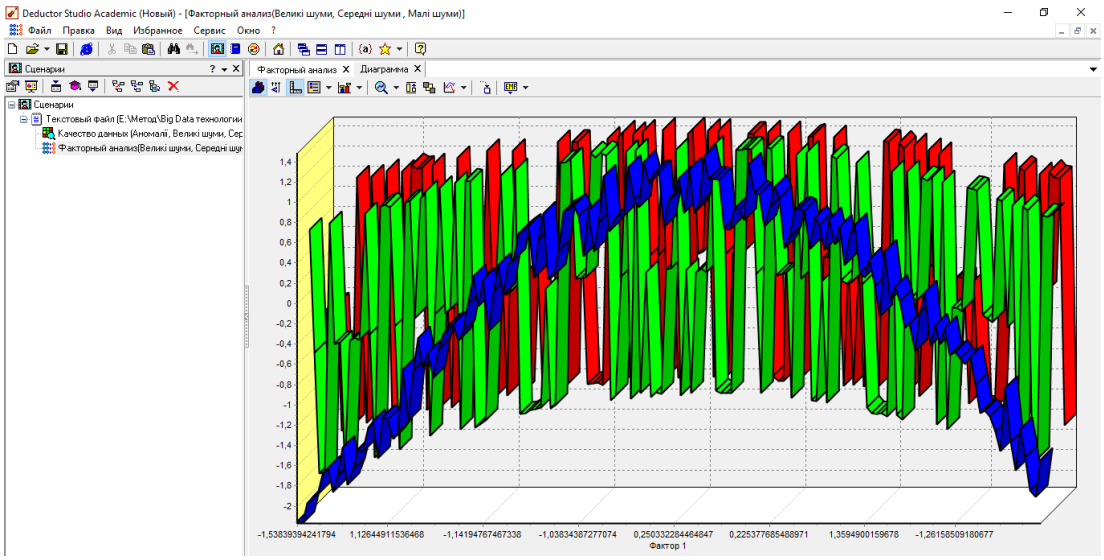


Рис. 4.9 – Результати факторного аналізу у вигляді діаграми

4.4 Обробка даних за допомогою кореляційного аналізу

Далі виконаємо обробку даних файлу за допомогою кореляційного аналізу. В майстрі обробки даних «Кореляційний аналіз» задамо: «Аргумент» – інформаційне, «Великі шуми», «Середні шуми», «Малі шуми» – вхідні значення, «Аномалії» – вихідні значення, «Синус» – непридатне (рис. 4.10).

Визначимо ступінь впливу вхідних факторів на вихід – «Аномалії» і залишимо тільки значущі фактори.

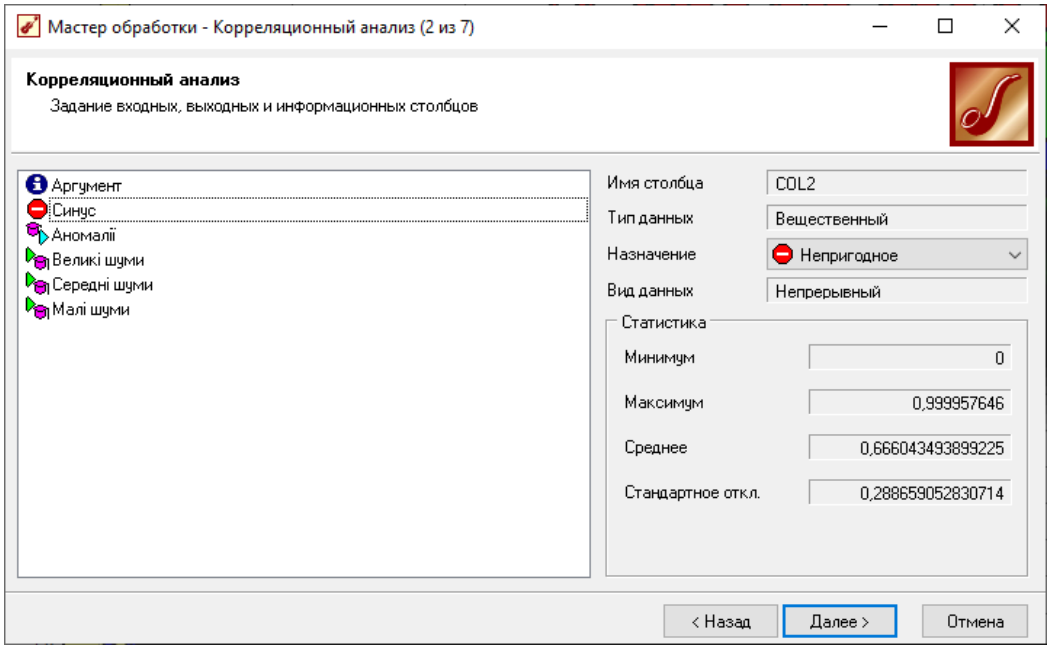


Рис. 4.9 – Дані для обробки

На наступному кроці запускаємо процес кореляційного аналізу. Після завершення процесу вибираємо, які фактори залишити для подальшої роботи. Це робиться або вручну, ґрунтуючись на значеннях матриці коваріації, або шляхом вказівки порога значимості (за замовчуванням поріг значимості дорівнює 0,05). З розрахованої матриці коваріації видно, що вихідне поле «Малі шуми» безпосередньо залежить від поля «Аномалії» (взагалі, значення коефіцієнта, рівне 1,000 говорить про те, що ці поля ідентичні), і в меншій мірі від інших факторів. В даному випадку без втрати корисної інформації можна виключити з подальшого розгляду «Великі шуми» і «Середні шуми» (рис. 4.11).

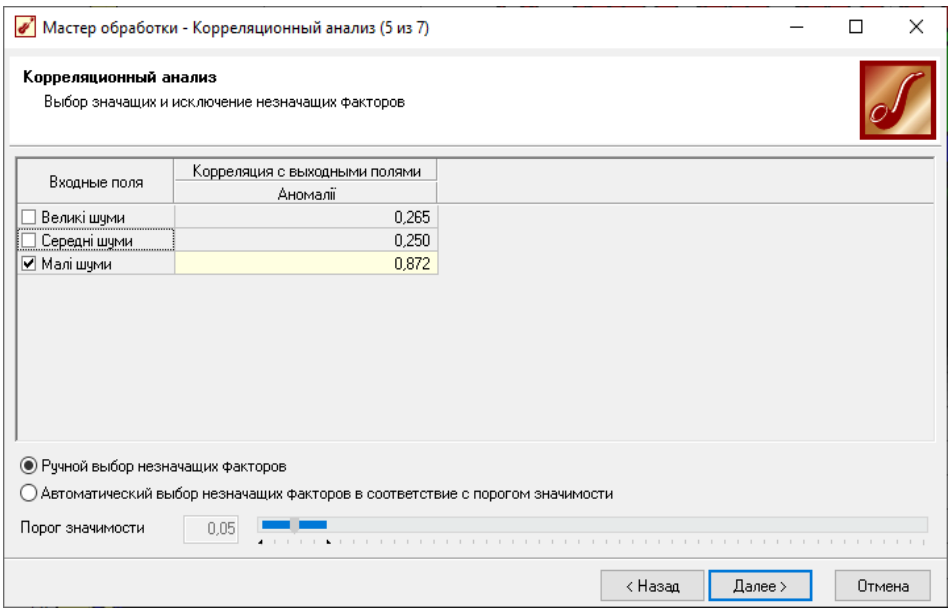


Рис. 4.11 – Вказівка порога значимості

Тепер необхідно перейти на наступний крок і вибрати спосіб візуалізації. Переглянемо результати на діаграмі (наприклад, можна переконатися в ідентичності полів «Малі шуми» і «Аномалії») (рис. 4.12, 4.13).

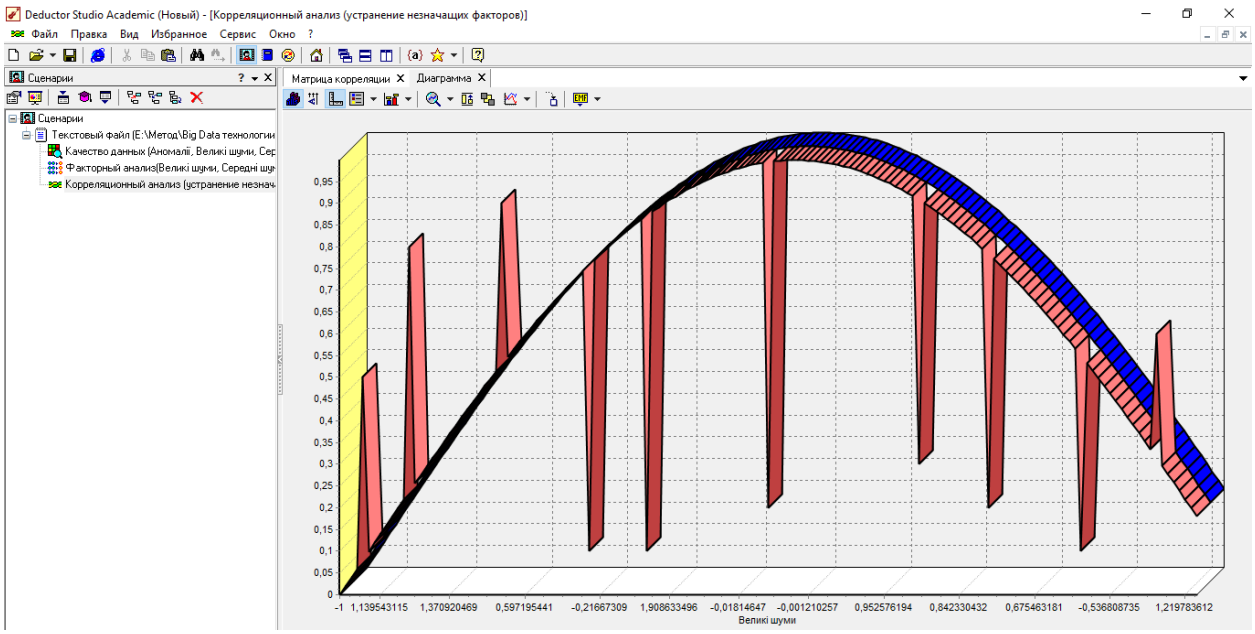


Рис. 4.12 – Результат ідентичності полів «Малі шуми» і «Аномалії»

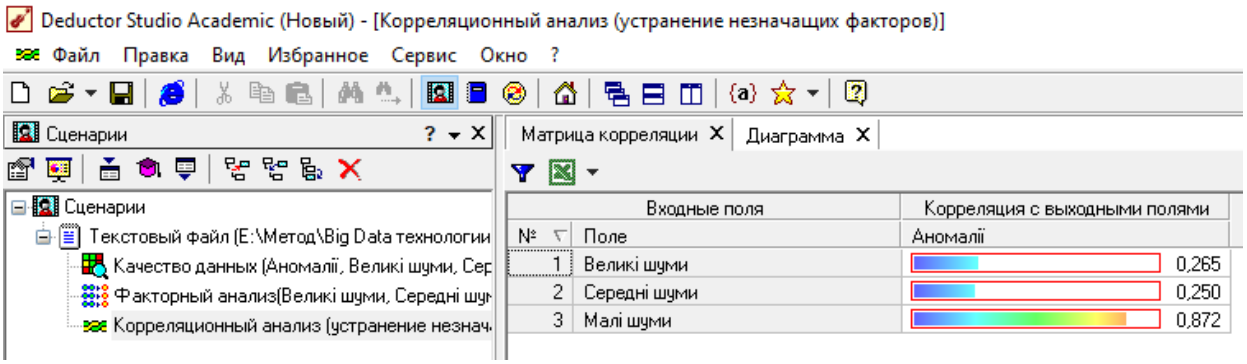


Рис. 4.13 – Результати кореляційного аналізу

Таким чином, кореляційний аналіз дозволив проаналізувати вплив вхідних факторів на результат і виключити незначні фактори з подальшого аналізу.

Питання до розділу 4

1. Що таке факторний аналіз?
2. Для чого використовується факторний аналіз, при обробці та аналізі даних?
3. Що таке кореляційний аналіз?
4. Для чого використовується кореляційний аналіз, при обробці й аналізі даних?
5. Що є критерієм прийняття рішення про виключення фактору при кореляційному аналізі?
6. Які методи статистичного аналізу Ви ще знаєте?
7. Що таке матриця коваріації?

Розділ 5
ТРАНСФОРМАЦІЯ ДАНИХ

5.1 Способи трансформації даних в Deductor Studio

Трансформація даних – перенесення і перетворення проводиться на основі правил трансформації, що зіставляють аналітику двох "облікових зон" і визначають критерії передачі даних.

Розбиття даних на групи. Часто для проведення аналізу або побудови моделі прогнозу доводиться розбивати дані на групи, виходячи з певних критеріїв. У першому випадку така необхідність виникає, якщо аналітик бажає переглянути, наприклад, інформацію не по всій сукупності даних, а за певними групами (наприклад, яку суму кредиту беруть на ті чи інші цілі, або кредитори того чи іншого віку). У другому випадку (прогнозування) аналітику необхідно враховувати той факт, що певні групи (в даному випадку групи кредиторів) поводяться по різному, і що модель прогнозу, побудована на всіх даних не буде враховувати нюансів, що виникають у цих групах. Тобто краще побудувати кілька моделей прогнозу, наприклад, в залежності від сумової групи кредиту і будувати прогноз на них, ніж побудувати одну модель прогнозу. Виходячи з цього і не тільки, в Deductor Studio надається широкий набір інструментів, тим або іншим способом дозволяють розбивати

вихідні дані на групи, групувати будь-яким способом всілякі показники тощо.

Розбиття дати (по тижнях). Розбиття дати служить для аналізу різноманітних 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 заданому 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 інтервалах.

Квантування. Часто аналітику необхідно віднести безперервні дані (наприклад, кількість продажів) до будь-якого кінцевого набору (наприклад, всю сукупність даних про кількість продажів необхідно розбити на 5 інтервалів – від 0 до 100, від 100 до 200 і т.д., і віднести кожен запис вихідного набору до якогось конкретного інтервалу) для аналізу або фільтрації виходячи саме з цих інтервалів. Для цього в Deductor Studio застосовується інструмент квантування (або дискретизації). Квантування призначено для перетворення безперервних даних в дискретні. Перетворення може проходити як по інтервалах (дані розбиваються на задану кількість інтервалів однакової довжини), так і по квантилях (дані розбиваються на інтервали різної довжини так, щоб в кожному інтервалі знаходилося однакову кількість даних). В якості значень результуючого набору даних можуть виступати номер інтервалу, нижня або верхня межа інтервалу, середина інтервалу, або мітка інтервалу (значення визначаються аналітиком).

Фільтрація даних. У більшості випадків вихідний набір даних, або набір даних після обробки аналітику необхідно відфільтрувати. Фільтрація буває необхідна для розбиття даних на будь-які групи (наприклад, товарні групи) для подальшої обробки або аналізу даних вже окремо по кожній групі. Також деякі дані можу не підходити, або навпаки, підходити для подальшого аналізу в силу накладених умов (наприклад, якщо на якомусь етапі обробки даних були виявлені суперечливі записи, то їх необхідно виключити з подальшої обробки). Тут теж виникає необхідність фільтрації. Фільтрація дозволяє з базового набору даних отримати набір даних, що задовольняє певним аналітиком умов. В Deductor Studio механізм побудови умов фільтрації простий для розуміння. У вікні майстра можна визначити кілька елементарних умов фільтрації (<ПОЛЕ> <СТАВЛЕННЯ> <ЗНАЧЕННЯ>), послідовно пов'язаних логічними операціями (І, АБО).

Угруповання даних. Складно робити висновки на основі необробленої первинної інформації. Аналітику для прийняття рішення майже завжди потрібна зведена інформація. Сукупні дані набагато більш інформативні, тим більше, якщо їх можна отримати в різних розрізах. В Deductor Studio передбачений інструмент, який реалізує збір зведеної інформації – «Угруповання». Угруповання дозволяє об'єднувати записи по полях-вимірах і агрегуючи дані в полях-фактах для подальшого аналізу.

5.2 Застосування способів трансформації даних

Для того, щоб навчитися застосовувати розбиття даних, квантування і фільтрацію для трансформації даних створимо файл «Credit.xlsx», що містить дані кредитування. У файлі повинні бути такі стовпці, як «Сума кредиту», «Дата кредитування» (в форматі ДД.ММ.РР), «Мета кредитування», «Вік» (рис.5.1). Експортуємо файл у текстовий з роздільниками («Credit.txt»).

	A	B	C	D
1	Сума кредиту	Дата кредитування	Мета кредитування	Вік
2	7000	01.09.21	Оплата послуг	49
3	14578	01.09.21	Купівля товару	30
4	34567	03.09.21	Ремонт нерухомості	23
5	23567	04.10.21	Турпоїздка	22
6	7500	05.10.21	Оплата послуг	56
7	12345	05.10.21	Купівля нерухомості	78

Рис. 5.1 – Приклад заповнення файлу «Credit.xlsx»

Імпортуємо в систему Deductor Studio текстовий файл «Credit.txt».

Необхідно зробити розбиття даних за ризиками кредитування фізичних осіб; отримати дані за сумами взятих кредитів по тижнях; розбити дані про вік кредиторів на 5 інтервалів (до 30 років, від 30 до 40, від 40 до 50, від 50 до 60, старше 60 років). Причому представити дані в розрізі по тижнях.

Створимо файл «banks.xlsx» та імпортуємо в текстовий («banks.txt») і в систему. Файл повинен містити статистику по банках

України за певний період («Банк», «Філія», «Місто», «Прибуток») (рис.5.2).

	A	B	C	D
1	Банк	Філія	Місто	Прибуток
2	Глобус банк	32	Київ	355197
3	Альфа-банк	1786	Київ	0
4	Монобанк	100	Одеса	4678389
5	Радабанк	24	Суми	0
6	Приватбанк	1821	Харків	356564

Рис. 5.2 – Призначення параметрів стовбців

Необхідно виявити ряд міст, в яких прибуток банків найбільша.

Стовпці, що цікавлять, для виконання першої частини поставленої задачі: «СУМА КРЕДИТУ», «ДАТА КРЕДИТУВАННЯ», «МЕТА КРЕДИТУВАННЯ» і «ВІК». Після імпорту даних з текстового файлу найбільш інформативно переглянути дані можна за допомогою візуалізатора «Куб», вибравши в якості вимірювань стовпці «ВІК» і «МЕТА КРЕДИТУВАННЯ», а в якості факту - стовпець «СУМА КРЕДИТУ». Інші стовпці встановити як непридатні (рис. 5.3).

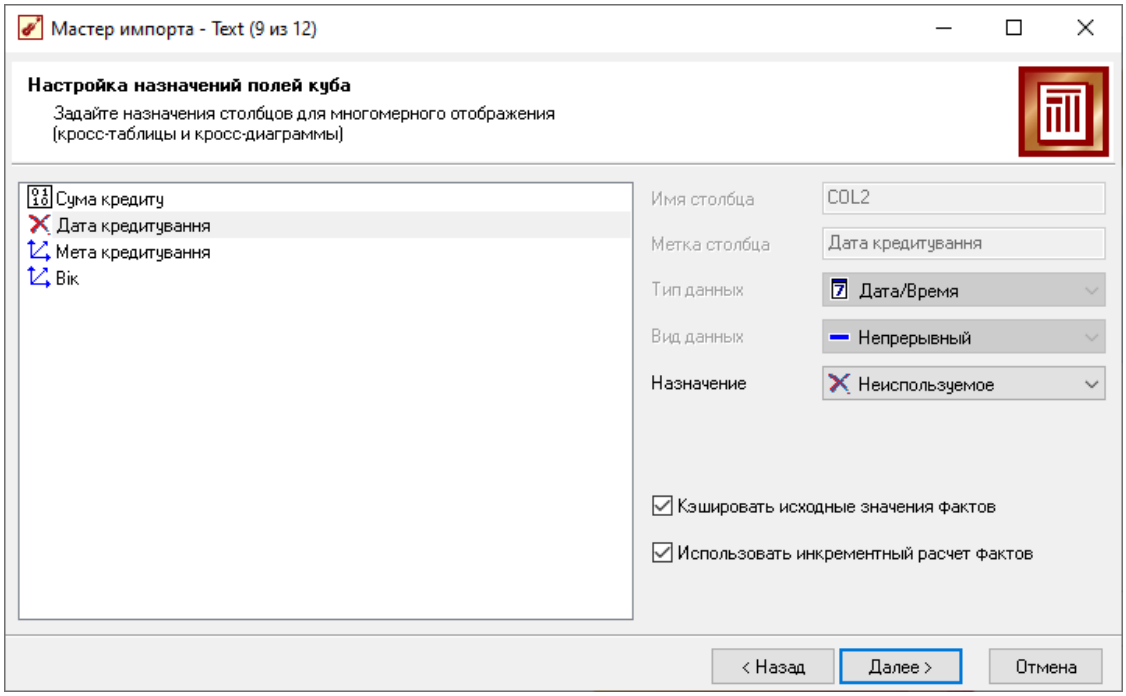


Рис. 5.3 – Налаштування призначення полів «Куб»

На наступному кроці налаштування куба слід вказати вимір «МЕТА КРЕДИТУВАННЯ» як вимір в рядках, а вимір «ВІК» як вимір в стовпцях, натиснувши відповідні кнопки в центральній області вікна, попередньо обравши з області доступних вимірів (рис. 5.4, 5.5).

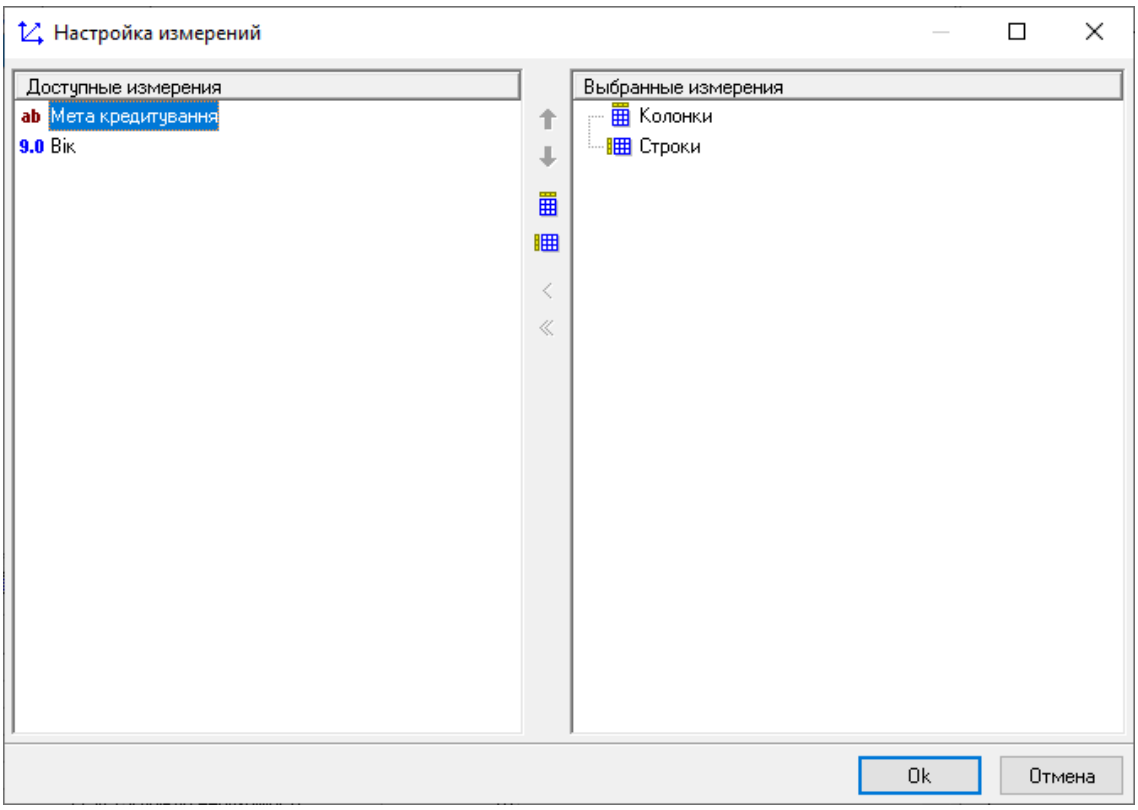


Рис. 5.4 – Вигляд вікна налаштування вимірювань візуалізатора «Куб»

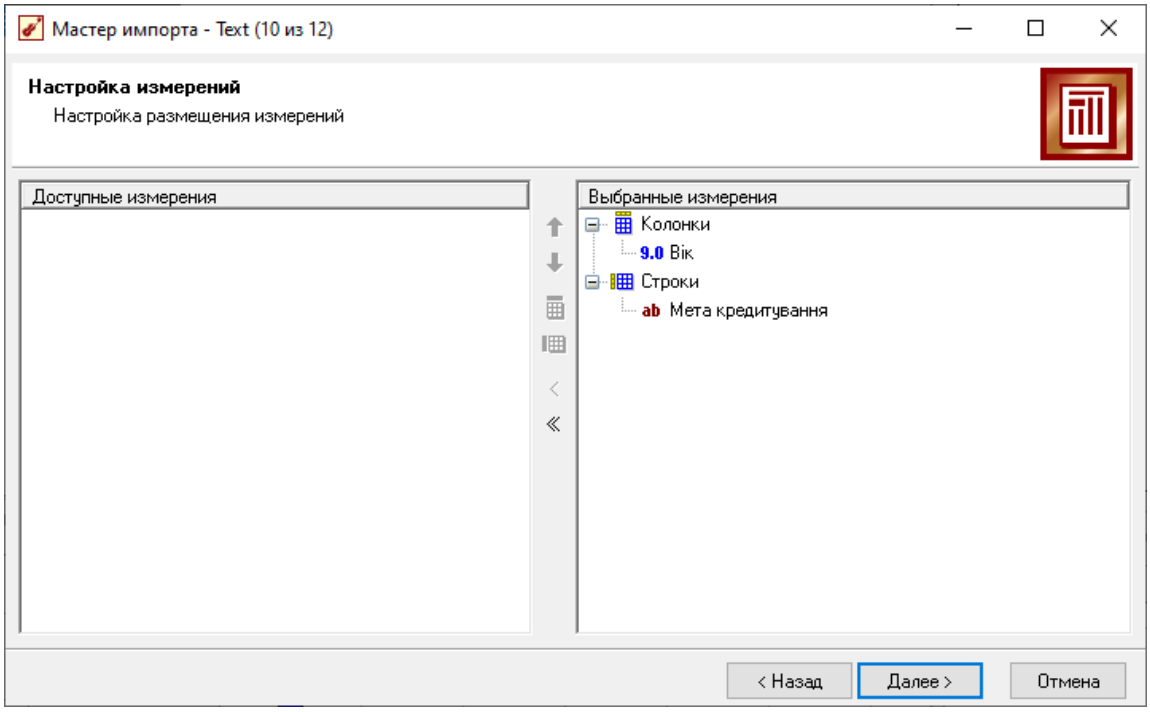
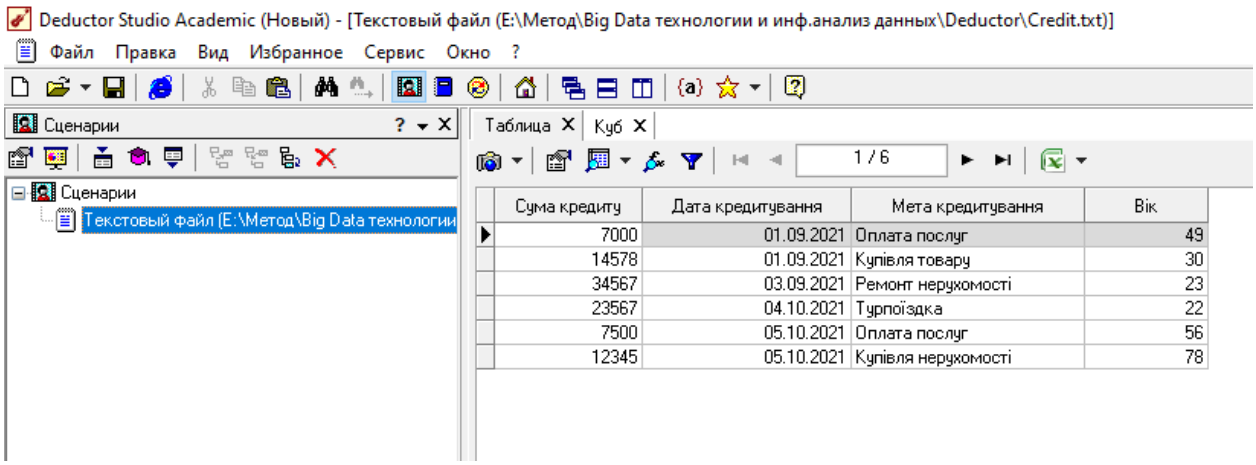


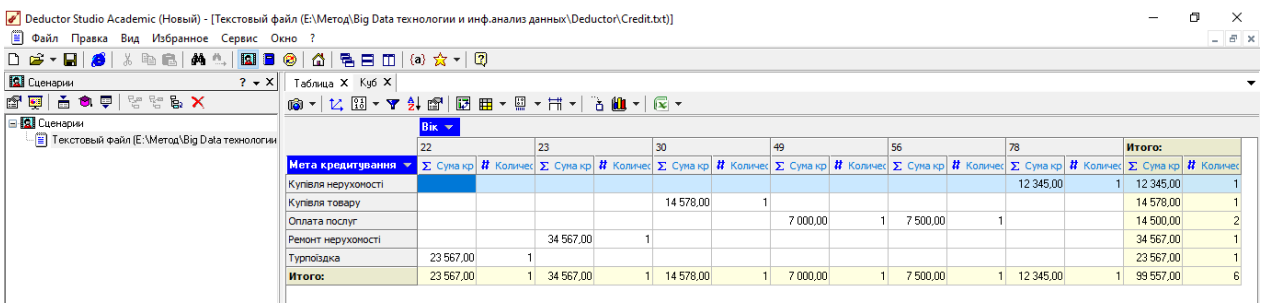
Рис. 5.5 – Налаштування вимірювань візуалізатора «Куб»

Результати візуалізатора «Куб» показані на рисунках 5.6 і 5.7.



Сума кредиту	Дата кредитування	Мета кредитування	Вік
7000	01.09.2021	Оплата послуг	49
14578	01.09.2021	Купівля товару	30
34567	03.09.2021	Ремонт нерухомості	23
23567	04.10.2021	Турпоїздка	22
7500	05.10.2021	Оплата послуг	56
12345	05.10.2021	Купівля нерухомості	78

Рис. 5.6 – Результат імпорту даних файлу «Credit.txt»



Мета кредитування	22		23		30		49		56		78		Итого:	
	Σ Сума кр	# Количес	Σ Сума кр	# Количес	Σ Сума кр	# Количес	Σ Сума кр	# Количес	Σ Сума кр	# Количес	Σ Сума кр	# Количес	Σ Сума кр	# Количес
Купівля нерухомості											12 345,00	1	12 345,00	1
Купівля товару					14 578,00	1							14 578,00	1
Оплата послуг							7 000,00	1	7 500,00	1			14 500,00	2
Ремонт нерухомості			34 567,00	1									34 567,00	1
Турпоїздка	23 567,00	1											23 567,00	1
Итого:	23 567,00	1	34 567,00	1	14 578,00	1	7 000,00	1	7 500,00	1	12 345,00	1	99 557,00	6

Рис. 5.7 – Результат візуалізатора «Куб»

У підсумку, на крос-діаграмі (одна з закладок візуалізатора куб) можна переглянути вихідні дані (рис. 5.8).

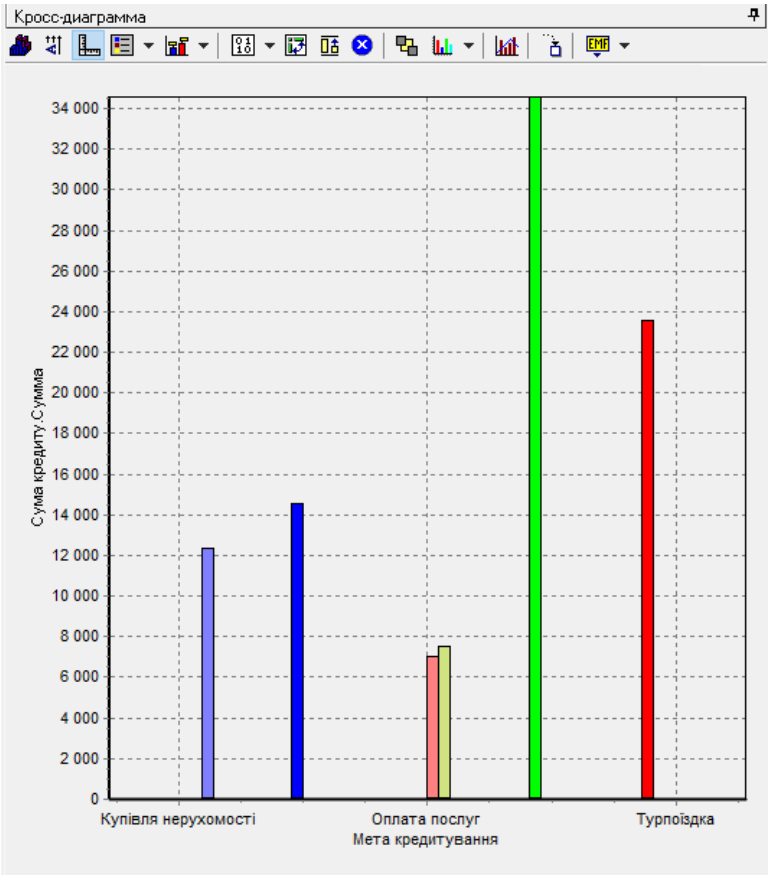


Рис. 5.5 – Крос-діаграма

У майстра обробки «Дата і Час» (рис. 5.6) на другому кроці виберемо поле «ДАТА КРЕДИТУВАННЯ» використовуваним, в таблиці налаштувань, що з'явилася після цього, виберемо призначення «Використовуване» в стовпці «Рядок» навпроти рядка «Рік + Тиждень».

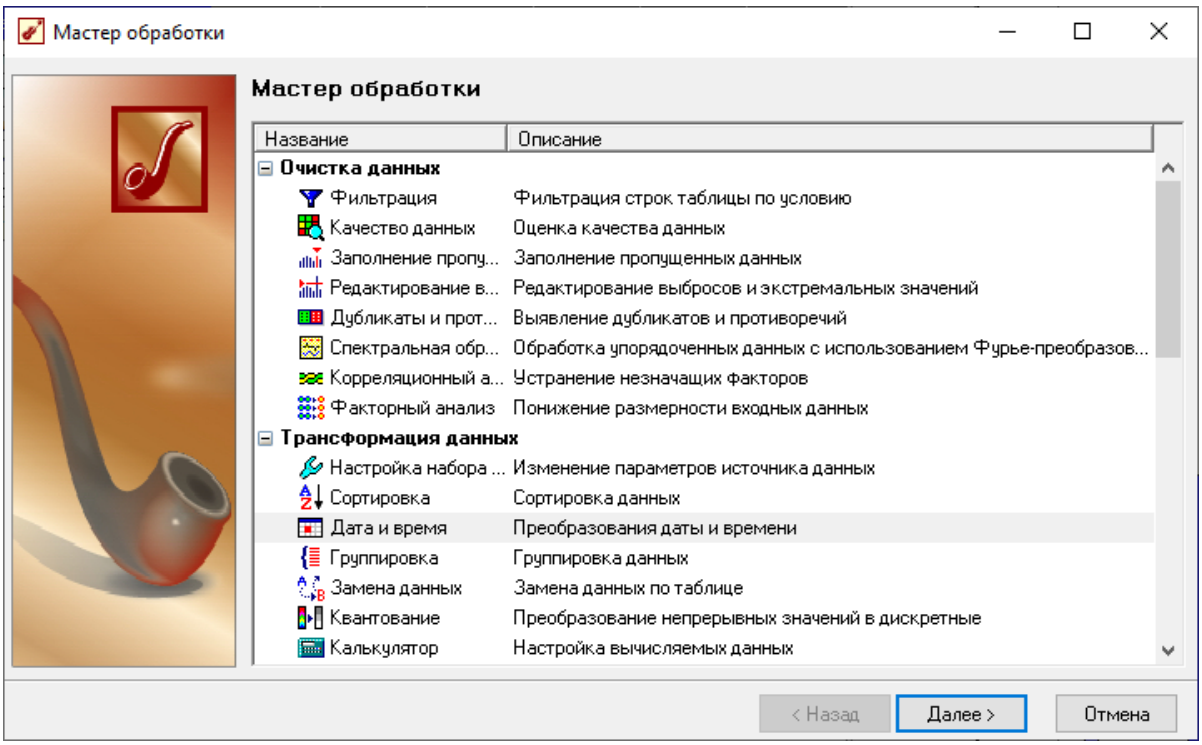


Рис. 5.5 – Вікно майстра обробки «Дата і Час»

Виберемо в якості візуалізаторов «Таблицю» і «Куб», поставивши галочки у відповідних позиціях. У майстра настройки полів куба виберемо в якості вимірювання з'явився після обробки стовпець «ДАТА КРЕДИТУВАННЯ_(Рік + Тиждень)» і стовпець «МЕТА КРЕДИТУВАННЯ», а в якості факту – «СУМА КРЕДИТУ». Решту полів зробимо невикористовуваними (рис. 5.7).

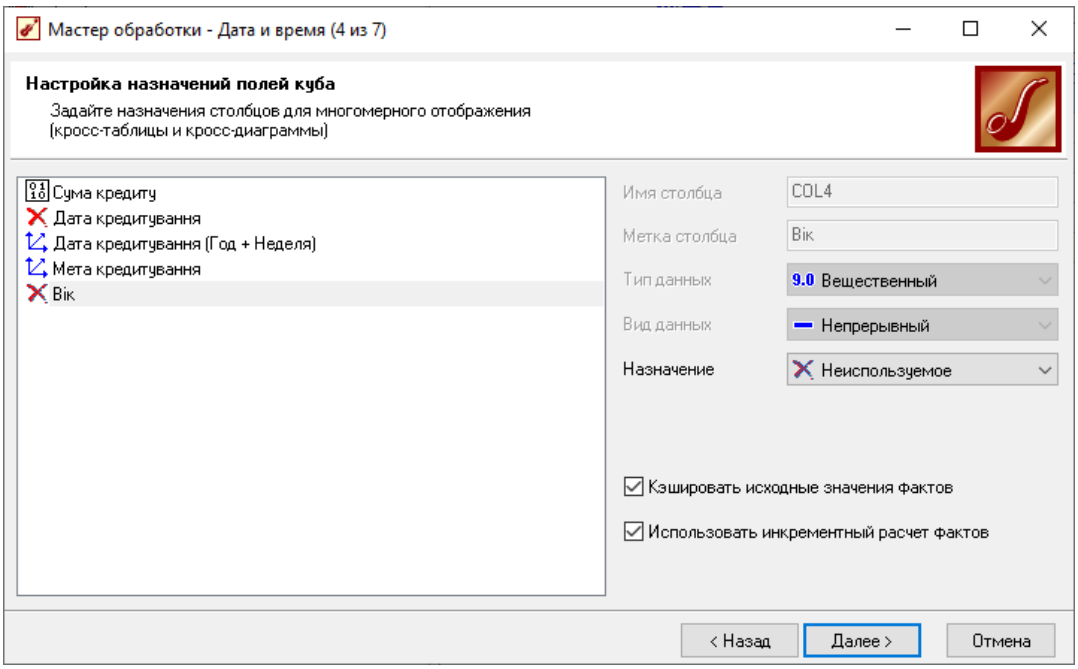


Рис. 5.7 – Налаштування призначення полів кубу

На наступному кроці перенесемо один вимір з області «доступних» в область «Вимірювання в рядках», а інше - в область «Вимірювання в стовпцях» (рис. 5.8, 5.9).

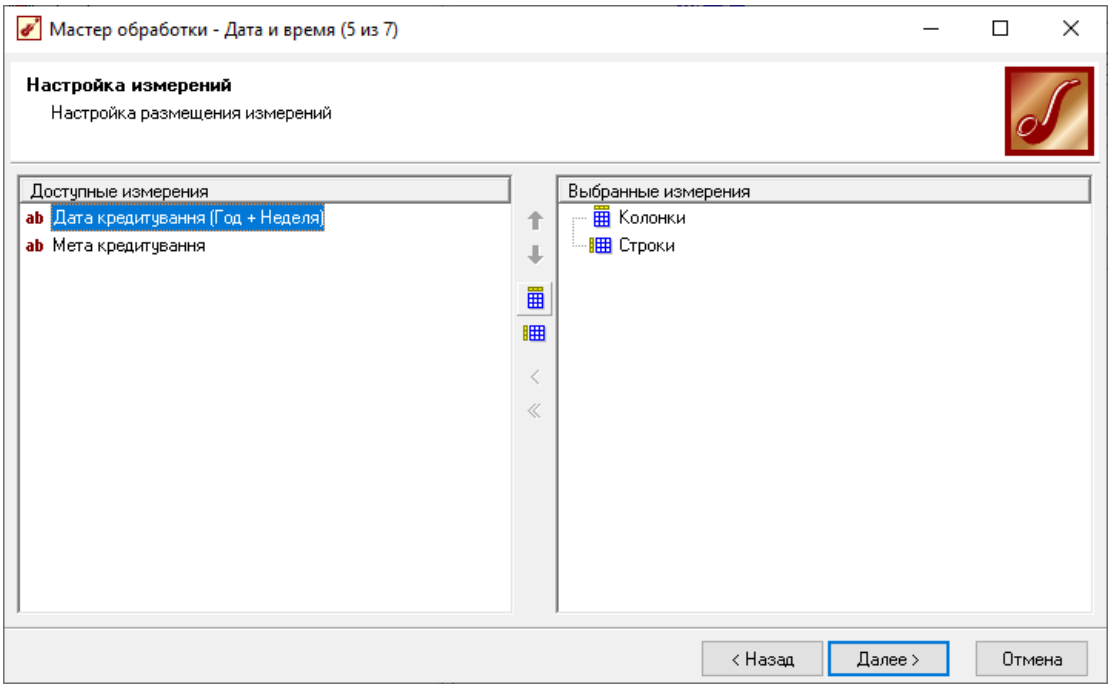


Рис. 5.8 – Вигляд вікна налаштування вимірів

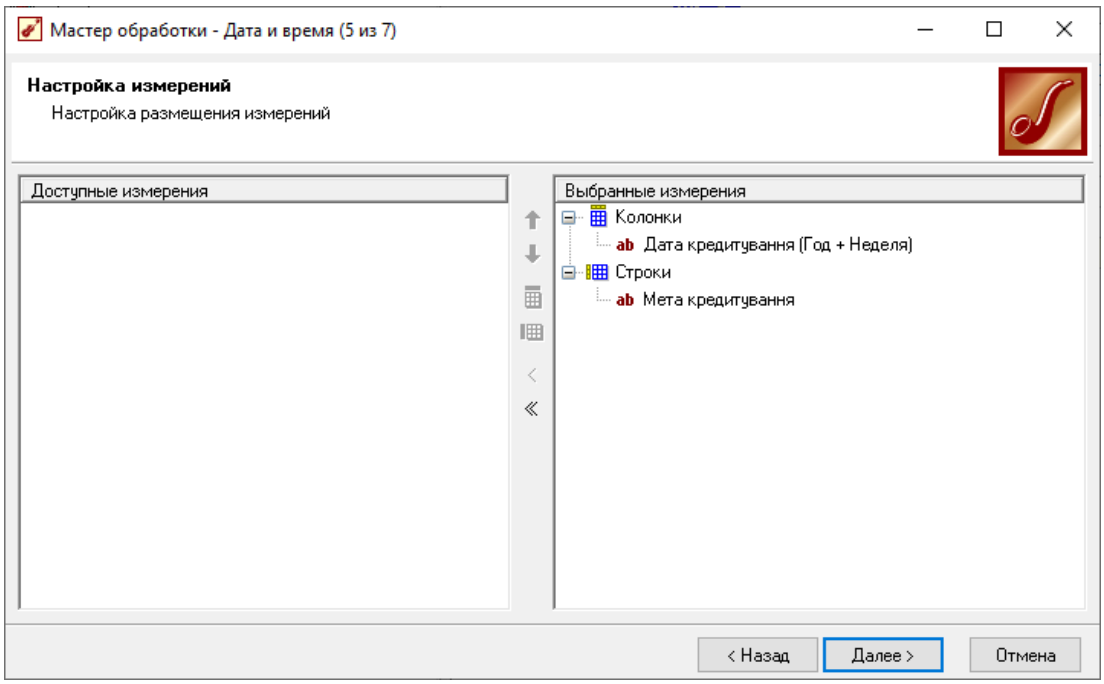


Рис. 5.9 – Налаштування вимірів

Таким чином, на крос-діаграмі маємо суми взятих кредитів по тижнях (за два тижні року) в розрізі цілей кредитування (рис. 5.10, 5.11).

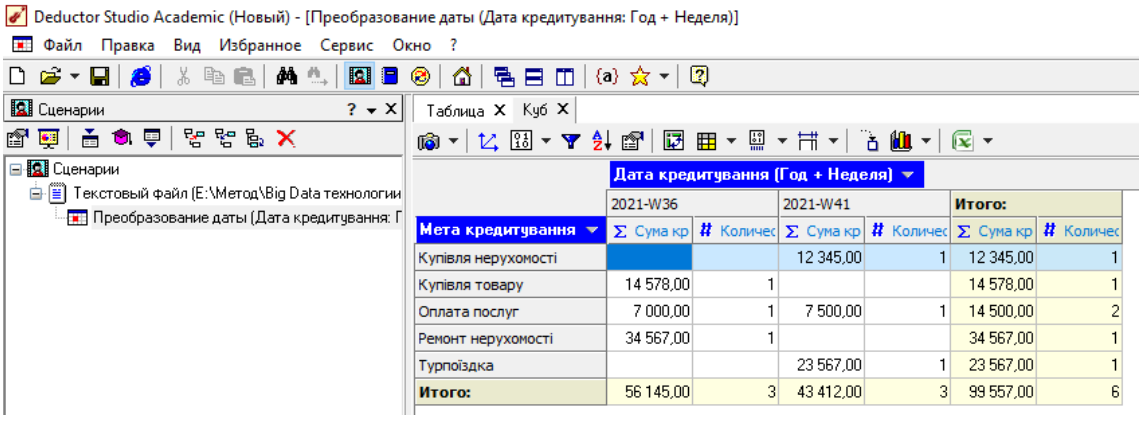


Рис. 5.10 – Результат перетворення дати

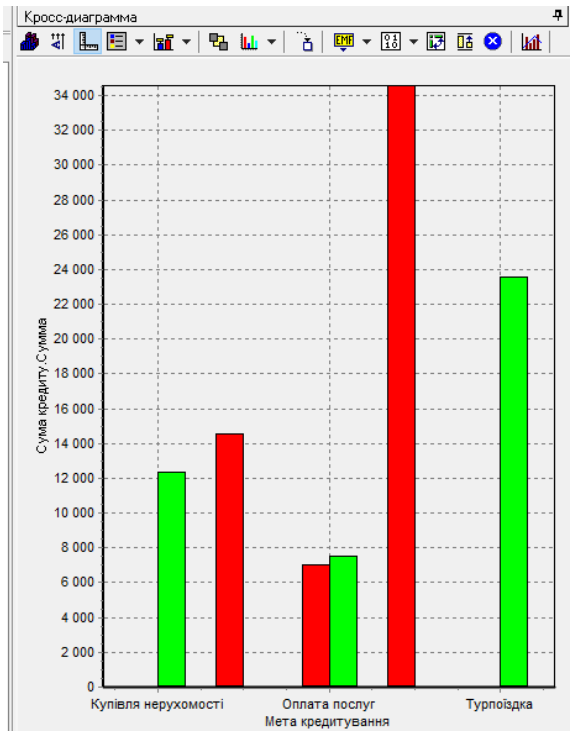


Рис. 5.11 – Крос-діаграма

Далі використовуємо інструмент квантування для розбиття даних про вік кредиторів на 5 інтервалів (до 30 років, від 30 до 40, від 40 до 50, від 50 до 60, старше 60 років). Вихідні дані розподіляться по п'яти інтервалах саме так, оскільки, згідно зі статистикою, мінімальне значення віку кредитора 19, а максимальне 69 років. Це необхідно аналітику для оцінки кредиторської активності різних вікових груп, з метою прийняття рішення про стимулювання кредиторів в групах з низькою активністю (наприклад, зменшення вартості кредиту для цих груп) і, можливо, збільшення прибутку в вікових групах кредиторів з високим ризиком (шляхом пропозиції додаткових платних послуг). Причому аналітик бажає бачити дані в розрізі по тижнях. Скористаємося майстром квантування (рис. 5.12).

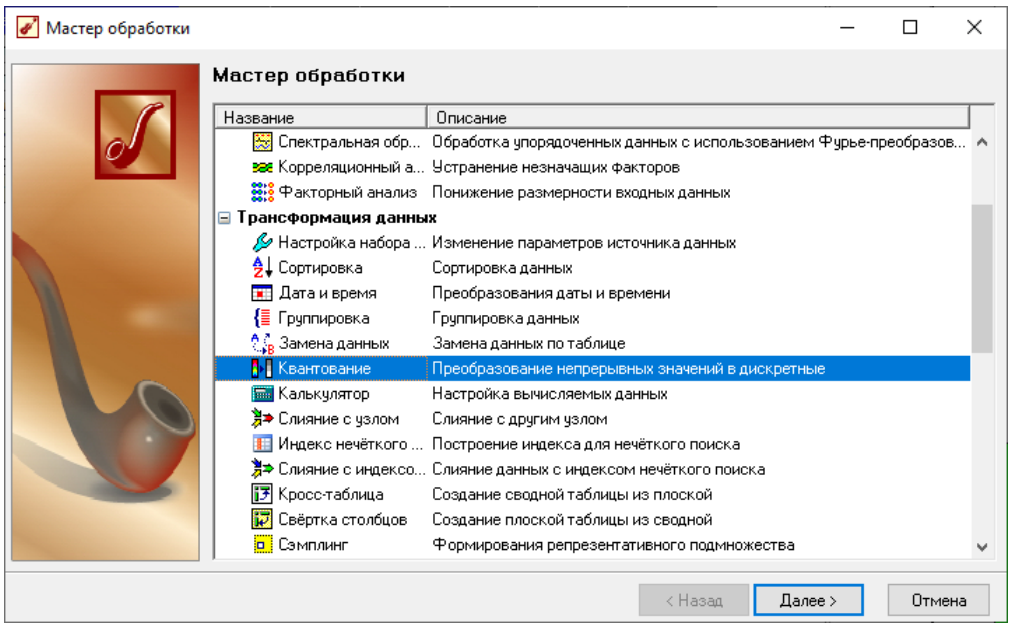


Рис. 5.12 – Майстер обробки «Квантування»

В майстрі обробки "Квантування" виберемо призначення поля «Вік» використовуваним, вкажемо спосіб розбиття «За інтервалами», задамо кількість інтервалів рівний 5, як значення виберемо «мітку інтервалу» (рис. 5.13).

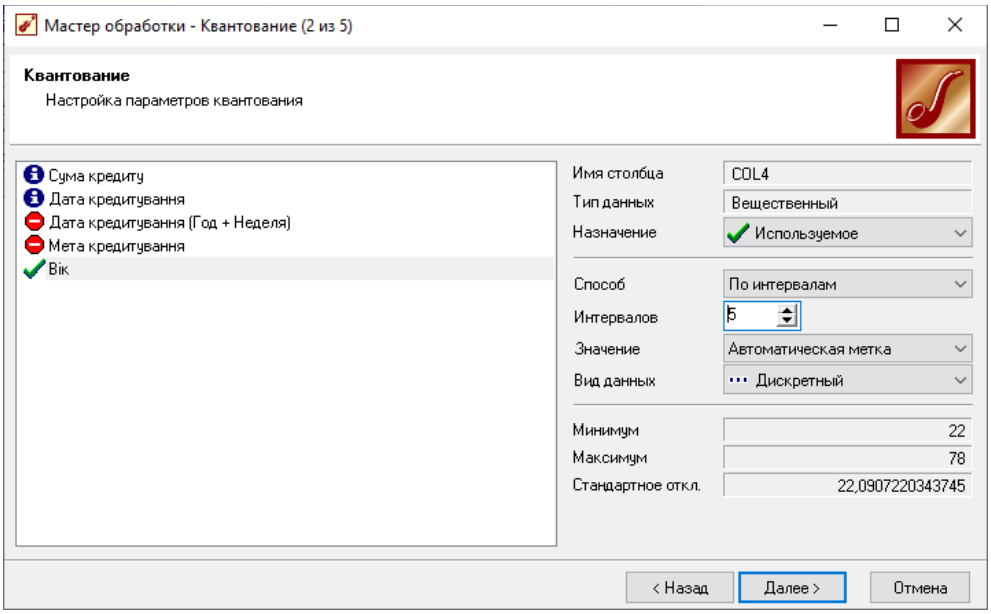


Рис. 5.13 – Майстер квантування

На наступному кроці майстра визначимо самі мітки відповідно віку кредиторів: «до 30 років», «від 30 до 40 років» і т.д. (рис. 5.14)

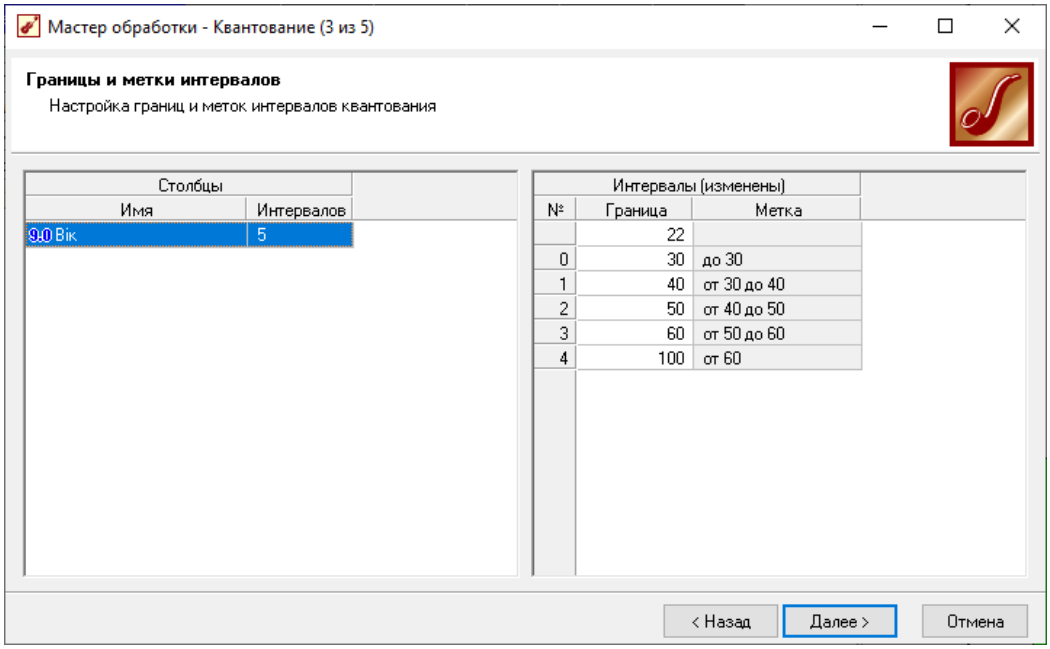


Рис. 5.14 – Визначення міток

Після обробки виберемо в якості способу відображення «Куб». У майстрі вкажемо «СУМА КРЕДИТУ» в якості факту, «ВІК» і поле «ДАТА КРЕДИТУВАННЯ (Рік + Тиждень)» в якості вимірювання, інші поля вкажемо невикористовуваними. Далі перенесемо «ВІК» з доступних вимірів в «Вимірювання в рядках», а «ДАТА КРЕДИТУВАННЯ (Рік + Тиждень)» в «Вимірювання в стовпцях» (рис. 5.15, 5.16).

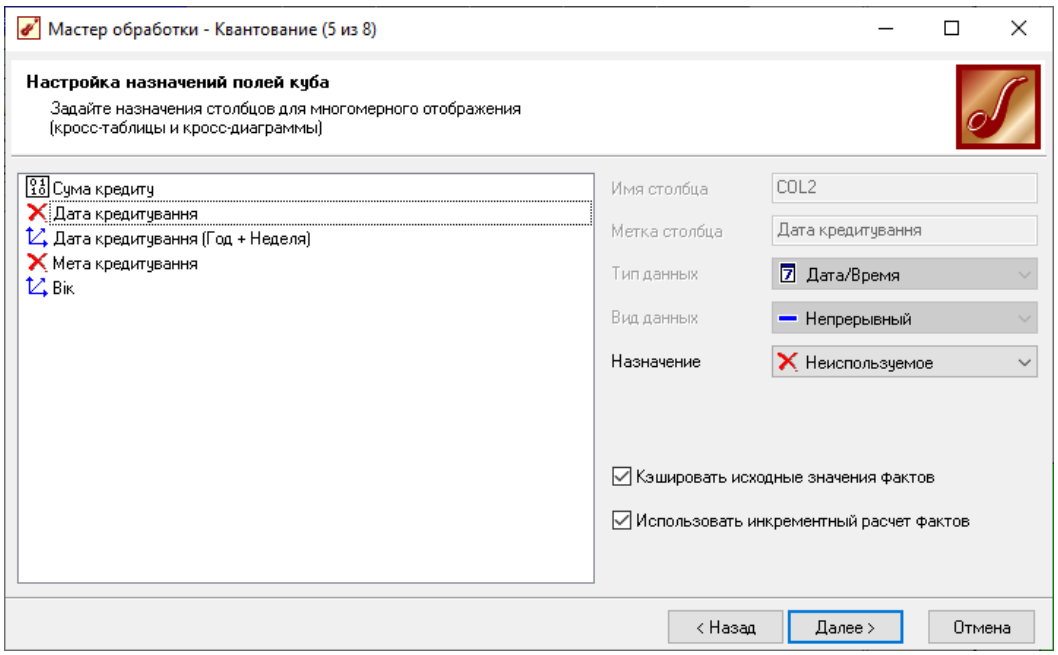


Рис. 5.15 – Налаштування призначень полів кубу

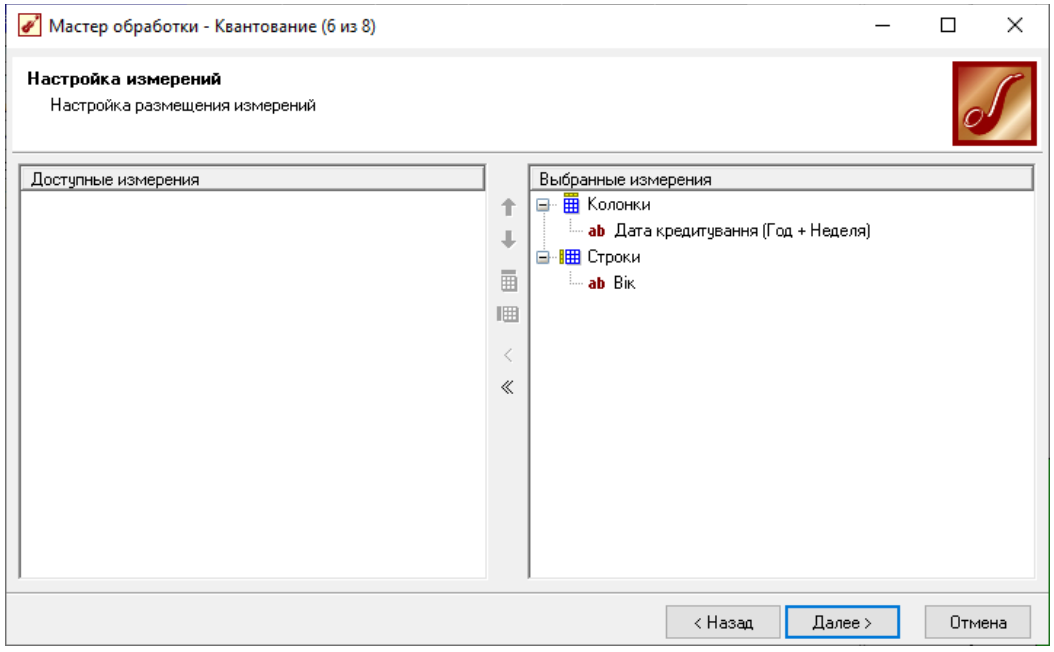
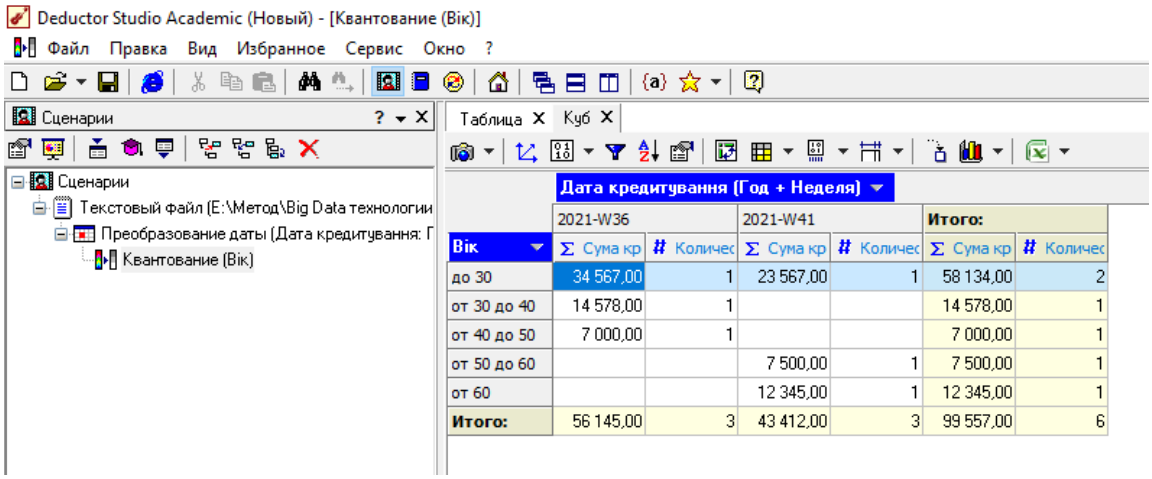


Рис. 5.16 – Налаштування вимірів

Таблиця результатів квантування показаний на рис. 5.17. На крос-діаграмі (рис. 5.18) тепер видна інформація про те, які суми кредитів беруть кредитори певних вікових груп в розрізі по тижнях.



Вік	2021-W36 Σ Сума кр	2021-W36 # Количес	2021-W41 Σ Сума кр	2021-W41 # Количес	Итого: Σ Сума кр	Итого: # Количес
до 30	34 567,00	1	23 567,00	1	58 134,00	2
от 30 до 40	14 578,00	1			14 578,00	1
от 40 до 50	7 000,00	1			7 000,00	1
от 50 до 60			7 500,00	1	7 500,00	1
от 60			12 345,00	1	12 345,00	1
Итого:	56 145,00	3	43 412,00	3	99 557,00	6

Рис. 5.17 – Таблиця результатів квантування

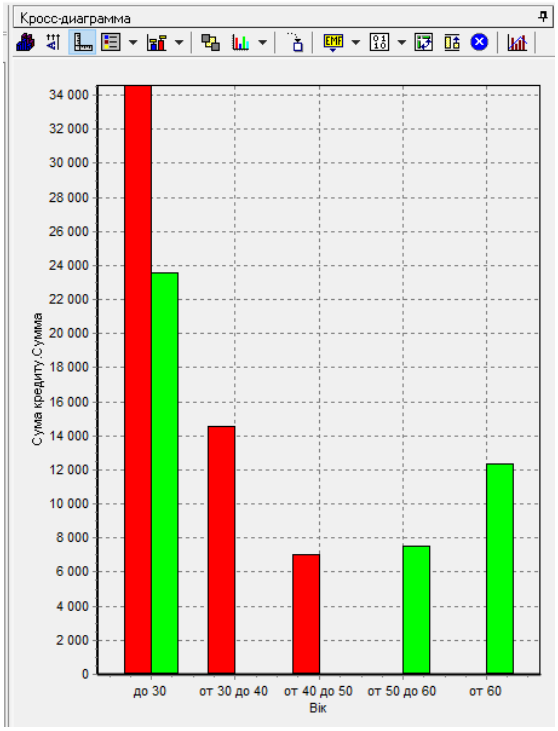


Рис. 5.18 – Крос-діаграма

Тепер аналітик, отримавши такі дані, може дати рекомендації про зниження вартості кредиту для осіб, старше 50 років, або про застосування якихось інших заходів, здатних привернути більшу кількість кредиторів цих груп, яких заходів, спрямованих на те, щоб кредитори брали кредит на великі суми.

Тепер припустимо, що у аналітика є статистика по банках України за певний період. Перед ним стоїть завдання виявлення ряду міст, в яких прибуток банків найбільша для використання цих даних в подальшому. Для цього аналітик повинен звернути увагу на наступні поля таблиці з файлу: «БАНК», «ФІЛІЇ», «МІСТО», «ПРИБУТОК». Тобто інформація про назву банку, місті, в якому він знаходиться, (філії банку можуть перебувати в різних містах - отже, по одному і тому ж банку може бути кілька записів з даними по різних містах) і прибуток банку.

Для розв'язання поставленої задачі в першу чергу необхідно знайти сумарний прибуток всіх банків в кожному місті. Для цього і необхідна угруповання. Для початку слід імпортувати дані по банкам з текстового файлу. Переглянути вихідну інформацію можна в вигляді куба, де по рядках будуть назви банків, а по стовпцях - міста. За допомогою візуалізатора «Куб» також можна отримати необхідну інформацію, вибравши в якості вимірювання поле «МІСТО», а в якості факту «ПРИБУТОК». Але необхідно отримати ці дані для подальшої обробки, отже, необхідно зробити аналогічне групування.

Перебуваючи у вузлі імпорту, запустимо майстер обробки. Виберемо в якості обробки групування даних (рис. 5.19).

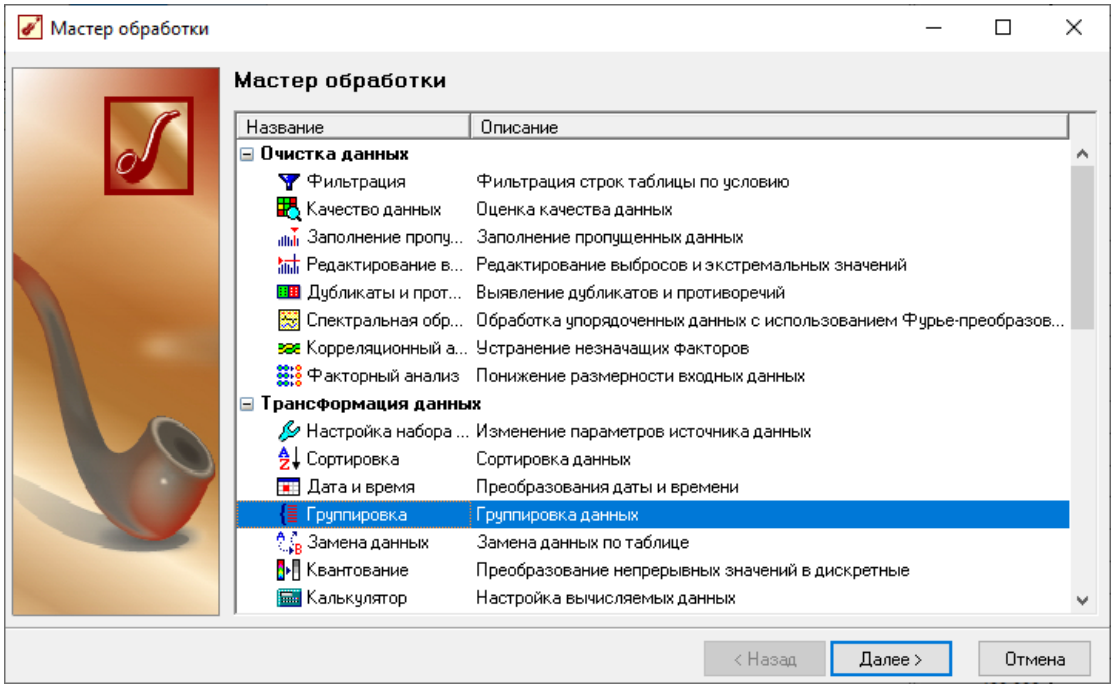


Рис. 5.19 – Запуск «Групування»

На другому кроці майстра встановимо призначення поля «МІСТО» як вимір, а призначення поля «ПРИБУТОК» як факт. В якості опції агрегації у поля «ПРИБУТОК» слід вказати Суму (рис. 5.20).

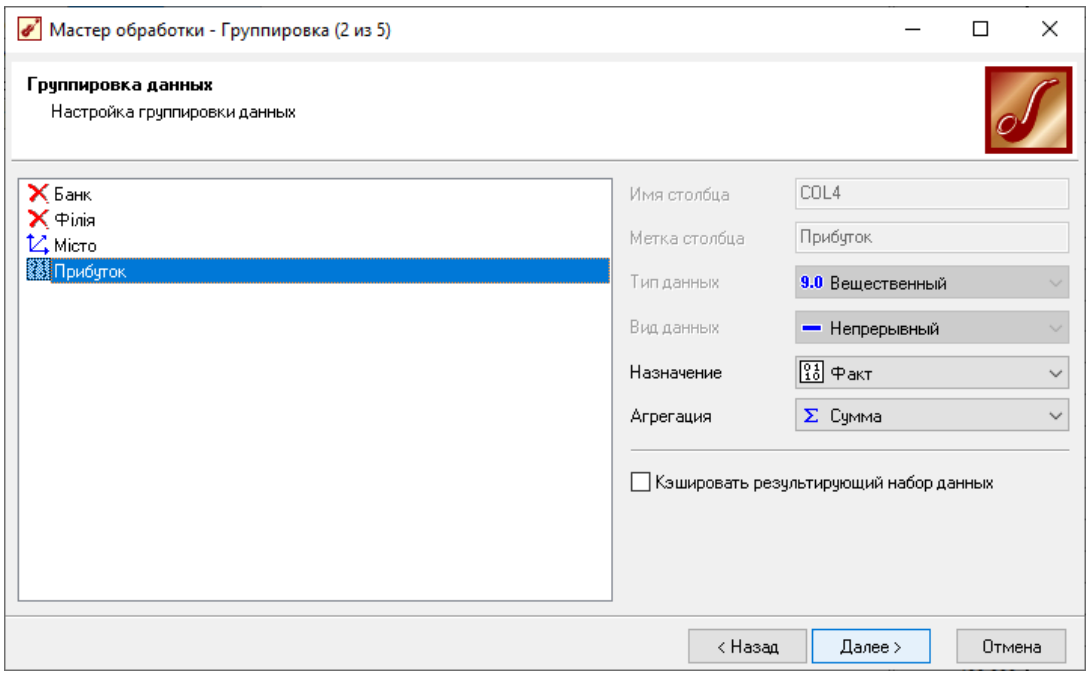


Рис. 5.20 – Налаштування групування даних

Таким чином, після обробки отримаємо сумарні дані по прибутку всіх банків по кожному місту. Їх можна переглянути, використовуючи таблицю (рис. 5.21, 5.22).

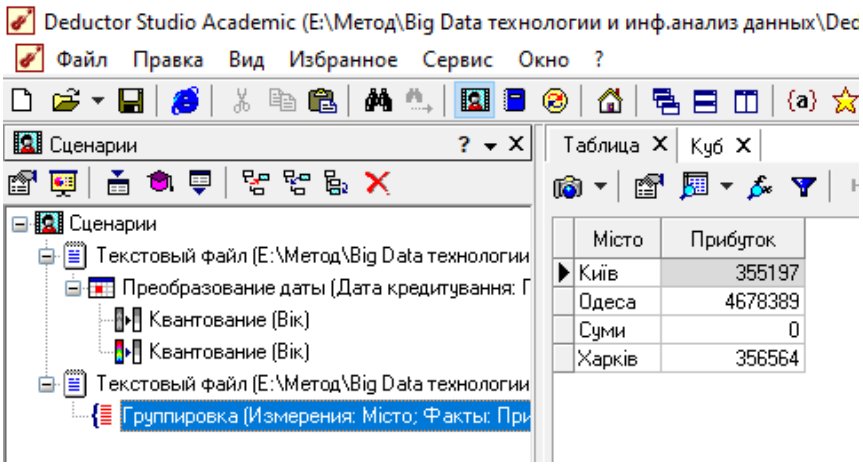


Рис. 5.21 – Результат імпорту даних файлу «banks.txt»

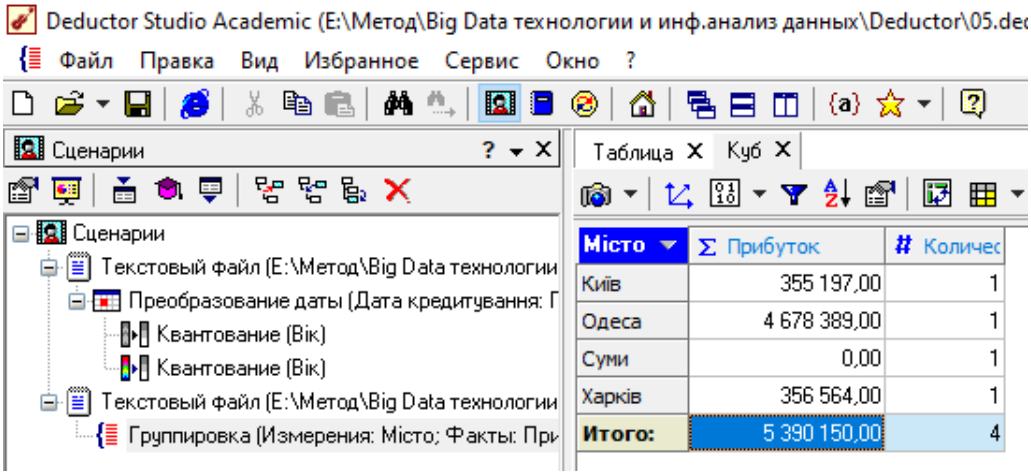


Рис. 5.22 – Таблиці даних

В цій частині було показано застосування розбиття даних, квантування і фільтрація для трансформації даних.

Питання до розділу 5

1. Що таке розбиття даних на групи?
2. Для чого використовується розбиття даних на групи в ході аналізу даних?
3. Для якого аналізу використовується розбиття за датою?
4. Що таке фільтрація даних?
5. Для чого використовується фільтрація даних для аналізу рішень?
6. Як у системі Deductor Studio здійснюється фільтрація даних?
7. Що таке угруповання даних?
8. Для чого використовується угруповання даних в аналізі рішень?

9. Як у системі Deductor Studio здійснюється угруповання даних?
10. Для чого використовується квантування даних в аналізі рішень?
11. Як у системі Deductor Studio здійснюється квантування даних?
12. Що таке крос-діаграма?

Розділ 6
ВИКОРИСТАННЯ СТАНДАРТНИХ МАТЕМАТИЧНИХ ФУНКЦІЙ ПРИ АНАЛІЗІ ТА ФОРМУВАННІ ДАНИХ

6.1 Інструмент «Калькулятор» платформи Deductor Studio

Іноді виникає необхідність на якомусь етапі обробки даних одержати на їх основі нові (похідні) дані. Можливо, аналітику потрібно обчислити відсоткове відхилення значення одного поля щодо іншого, або підрахувати суму, різницю полів, отримати на основі даних показник і вже його використовувати для подальшої обробки, залежно від значення полів обчислити ті чи інші вирази.

У Deductor Studio таку можливість надає інструмент «Калькулятор». Він дозволяє створювати нові поля, що обчислюють задані аналітиком вирази. Тобто калькулятор служить для отримання похідних даних на основі наявних у вихідному наборі. Майстер надає широкий набір функцій різного спрямування. У майстрі представлений список нових виразів, де додаються необхідні аналітику вирази, список доступних функцій з коротким описом кожної, список доступних операцій і список доступних стовпців, які можна використовувати при створенні виразу.

Список усіх вбудованих функцій разом з описом можна переглянути в майстрі, натиснувши кнопку *Функція*.

Реалізований у Deductor Studio конструктор виразів під час побудови використовує не мітки (Сума, Кількість, Ціна...), а імена полів таблиці, задані у джерелі даних (Summ, Count, Price…). При імпорті в деяких випадках (наприклад, з текстового файлу) можна задати як мітки, так і імена полів, що імпортуються.

Зліва у вікні конструктора знаходиться список виразів, що обчислюються. Спочатку воно містить один порожній вираз. Для управління списком обчислених виразів передбачені кнопки, що представлені в таблиці 6.1.

Таблиця 6.1 – Кнопки управління списком обчислених виразів

Кнопка	Призначення
	перемістити поточний вираз на одну позицію вгору за списком
	перемістити поточний вираз на одну позицію вниз за списком
	додає новий вираз з параметрами, що встановлюються за замовчуванням, та порожньою формулою, потім викликає діалог редагування параметрів виразу
	додає новий вираз з типом даних, описом та формулою як у поточного виразу, потім викликає діалог редагування параметрів виразу
	викликає діалог редагування параметрів виразу
	видаляє поточний вираз

Для зміни властивостей виразу використовується конструктор виразів. У ньому задаються такі параметри:

- Ім'я - рядок, який буде ідентифікатором стовпця у процедурах обробки. За бажанням користувач може ввести будь-які імена, які точніше відображають призначення стовпця;
- Мітка - назва, під якою цей стовпець буде видно в таблиці, крос-таблиці або на діаграмі після обробки. Бажано, щоб воно відбивало зміст стовпця;
- Тип даних – тип даних обчислюваного виразу. Тип вибирається зі списку, що відкривається клацанням по кнопці у правій частині поля.

Спочатку при відкритті сторінки конструктора список виразів містить лише один елемент "Вираз". Для нього слід встановити потрібні параметри та при необхідності додати нові рядки. За замовчуванням для нового виразу призначається мітка «Вираз N», де N – номер, що забезпечує унікальність. Імена полів, що формуються в результаті обчислень за цим виразом, призначаються автоматично і мають вигляд: EXPR_N, де N – унікальний номер.

Після налаштування параметрів виразу в полі «Вираз» потрібно ввести формулу, що розраховується. Правила складання виразів відповідають загальноприйнятим, зокрема, кількість дужок, що відкривають, повинна дорівнювати числу закриваючих. Вираз може містити:

- числа в явному виді;
- змінні у виді імен стовбців;
- дужки, що визначають порядок виконання операцій;

- знаки математичних операцій та відношень;
- імена функцій;
- дати у форматі «ДД.ММ.РР», які обов'язково вказуються в лапках. Такий спосіб введення дати, хоч і допускається, але може виявитися непереносимим між різними комп'ютерами. Тому краще використовувати функцію STRTODATE();
- рядкові вирази в лапках - «рядковий вираз»;
- однорядкові та багаторядкові коментарі. Однорядковий коментар починається символами "//" (два слеша) і триває до кінця рядка. Багаторядковим коментарем вважаються всі символи, що містяться між дужками "/*" (слеш-зірочка) і "*/" (зірочка-слеш).

Вираз можна ввести вручну з клавіатури, проте зручніше вибирати функції, змінні та знаки операцій за допомогою миші. У полі «Вираз» завжди відображається той вираз, який наразі виділено у списку. Для його редагування достатньо клацнути у полі мишею, викликавши курсор, а потім редагувати як звичайне текстове поле.

Щоб додати до виразу функцій, натисніть кнопку «Функція» на панелі «Операції». При цьому відкриється вікно вибору функції. Щоб ввести функцію, потрібно в полі «Список функцій» відкрити потрібний вид функцій, клацнувши по значку «+» праворуч від його найменування. В результаті буде розгорнутий список функцій цього виду. Якщо виділити функцію у списку, праворуч з'явиться короткий опис функції.

Щоб ввести функцію у вираз, достатньо двічі клацнути по її імені у списку. Ім'я функції у виразі з'являється разом із дужками,

куди слід ввести аргумент чи аргументи. Аргументами можуть бути числа в явному вигляді, рядки в лапках, дати в лапках, імена функцій, імена полів, а також арифметичні, логічні та строкові вирази. Імена полів зручно вводити за допомогою подвійного клацання у списку полів. Якщо в аргументі кілька полів, їх імена розділяються точкою з комою. Знаки математичних операцій та відносин можна вибирати клацанням миші у секції «Операції».

6.2 Застосування математичних функцій в Deductor Studio

Для того, щоб навчитися застосовувати стандартні математичні функції з використанням платформи Deductor Studio необхідно створити файл «Calculate.xlsx», в якому містяться стовпці «Аргумент1», «Аргумент2», «Аргумент3» – набір аргументів для обробки програмою. У кожному стовпці має бути 100 значень. У стовпці «Аргумент1» кожне значення повторюється десять разів. У стовпці «Аргумент2» вибираються десять значень, які повторюються. Стовпець «Аргумент3» заповнюється значеннями від 1 до 100 (за зростанням) (рис. 6.1). Експортуємо файл у текстовий з роздільниками («Calculate.txt») (рис. 6.2).

	A	B	C
1	Аргумент 1	Аргумент 2	Аргумент 3
2	0	1	1
3	0	2	2
4	0	3	3
5	0	4	4
6	0	5	5
7	0	6	6
8	0	7	7
9	0	8	8
10	0	9	9
11	0	10	10
12	1	1	11
13	1	2	12
14	1	3	13
15	1	4	14
16	1	5	15
17	1	6	16
18	1	7	17
19	1	8	18
20	1	9	19
21	1	10	20
22	2	1	21
23	2	2	22
24	2	3	23
25	2	4	24

Рис. 6.1 – Зразок заповнення файлу «Calculate.xlsx»

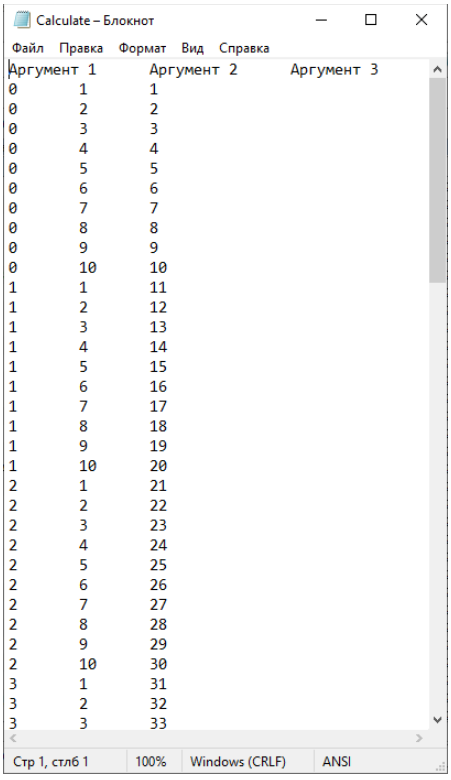


Рис. 6.2 – Зразок заповнення файлу «Calculate.txt»

На основі аргументів розрахувати математичні функції:

- дві функції одного аргументу (Аргумент3);
- sin(Аргумент3*Аргумент3)*ln(Аргумент3+1) – exp (-Аргумент3/10);
- 10*sin(Аргумент3*Аргумент3/100)/(Аргумент3+1)*exp(-Аргумент3/10);
- одну функцію від двох аргументів;
- Аргумент1 * Аргумент1/100 – Аргумент2 * Аргумент2/100;
- функцію, що показує відносне відхилення (Аргумент1+1 від Аргумент2+1).

Імпортувати дані з файлу «Calculate.txt», для перегляду вихідних даних зручніше використовувати візуалізатор «Таблиця» (рис. 6.3)

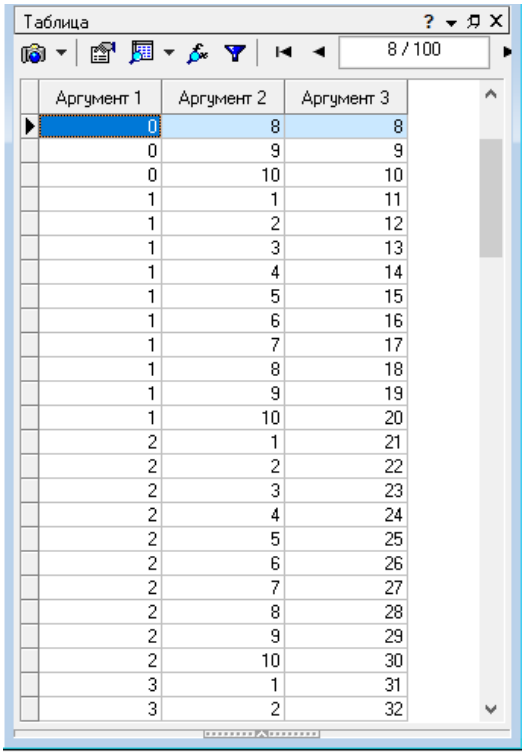


Рис. 6.3 – Перегляд даних

Розрахуємо значення функцій:

1) SIN (Аргумент3*Аргумент3)* LN (Аргумент3+1) * EXP (- Аргумент3/10)

2) 10*SIN(Аргумент3 * Аргумент3/100)/(Аргумент3+1)*EXP(- Аргумент3/10)

Для цього, перебуваючи на сайті імпорту, запустимо майстер обробки. Виберемо як оброблювач калькулятор (рис. 6.4).

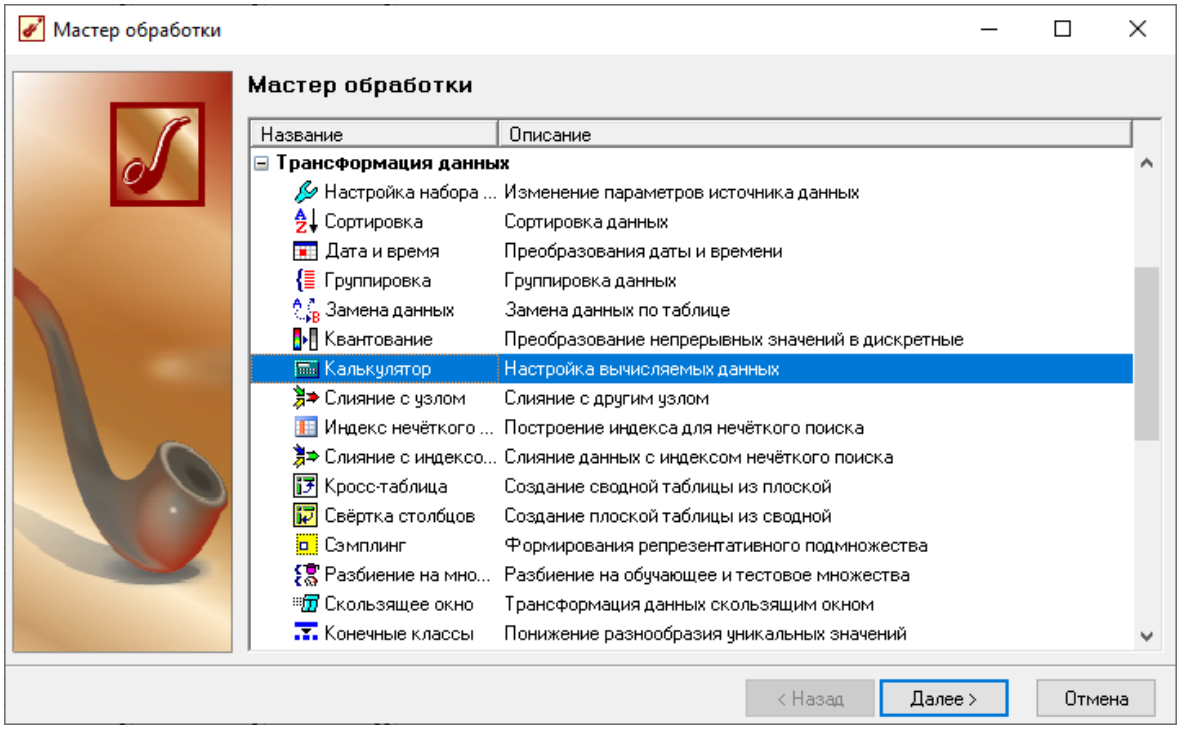


Рис. 6.4 – Вибір оброблювача

На другому кроці майстра у списку виразів у першому рядку у графі «Назва виразу» замість напису «Вираз» напишемо F1(Аргумент3). В полі редактора виразу (у верхній частині майстра) напишемо «SIN(COL3*COL3)*LN(COL3+1)*EXP(-COL3/10)» (рис. 6.5).

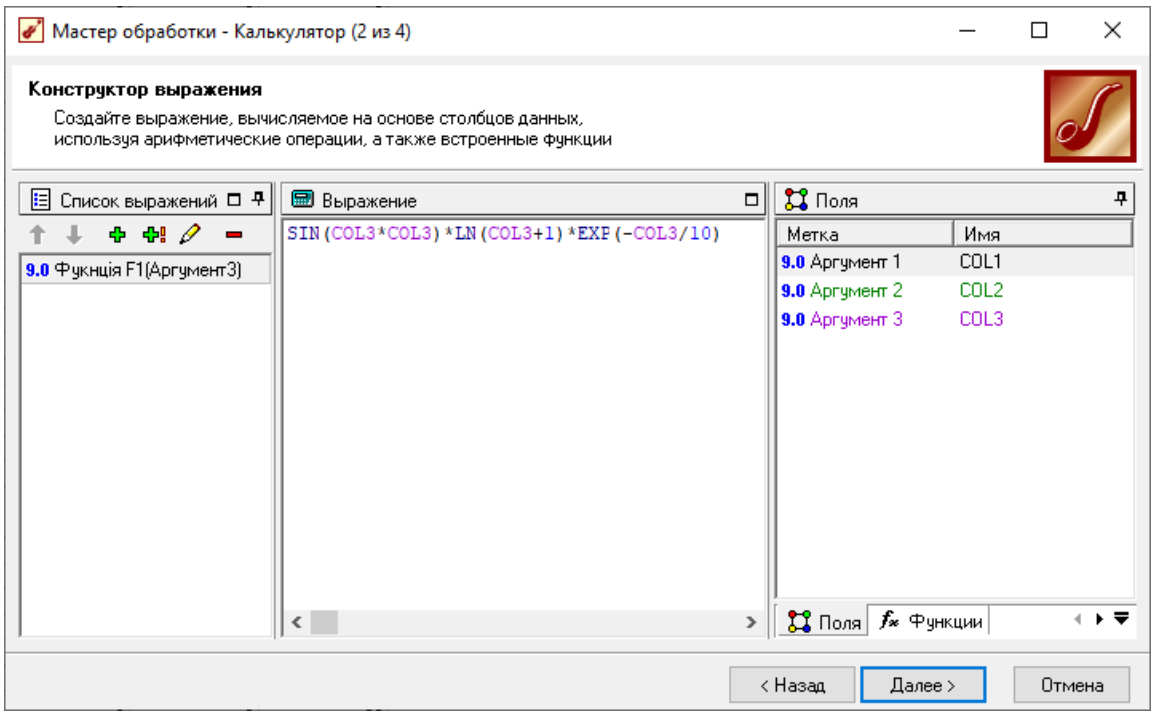


Рис. 6.5 – Конструктор виразу

Таким чином, створено новий стовпець, задали йому назву «F1(Аргумент3)» і також визначили, які значення прийматимуть записи цього поля. У цьому створення обчислюваного значення закінчено, тому переходимо до наступного кроку майстра, де пропонується вибрати спосіб відображення даних. Найінформативнішим у цьому випадку є діаграма, яку і слід вибрати.

Далі, вибравши у майстрі налаштувань діаграми в якості відображуваного поля «F1(Аргумент3)», тип графіка «Лінії» і підписи по осі X значення поля «Аргумент3» (рис. 6.6), можна побачити графік обчисленої функції, представлений на рисунку 6.7.

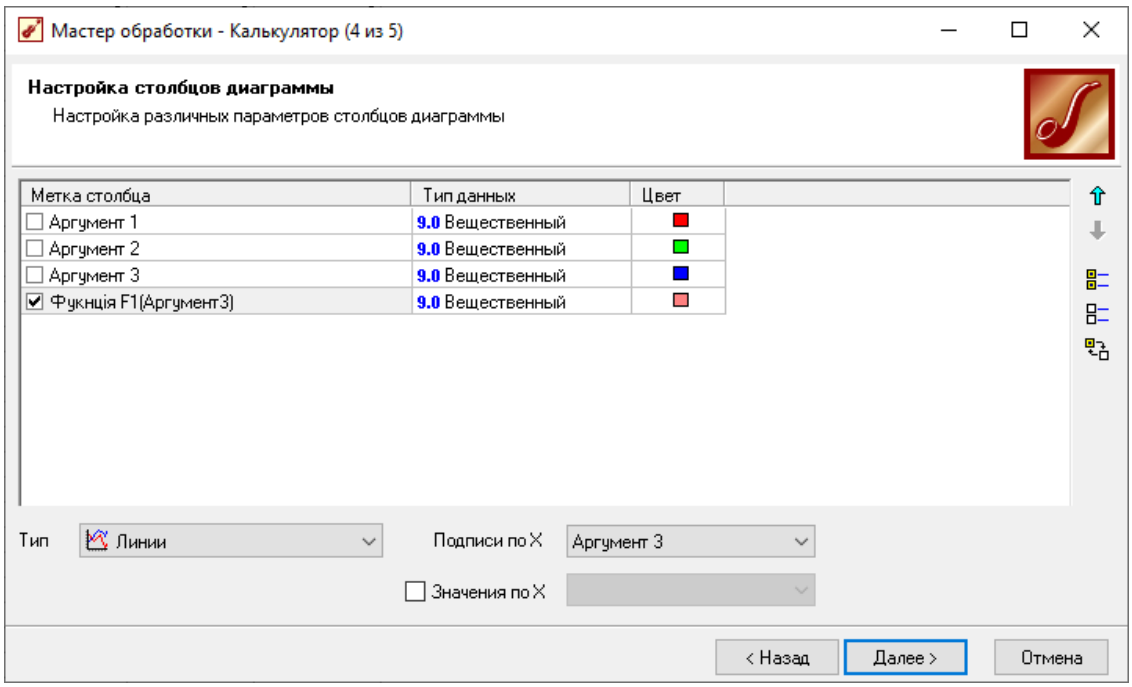


Рис. 6.6 – Налаштування стовбців діаграми

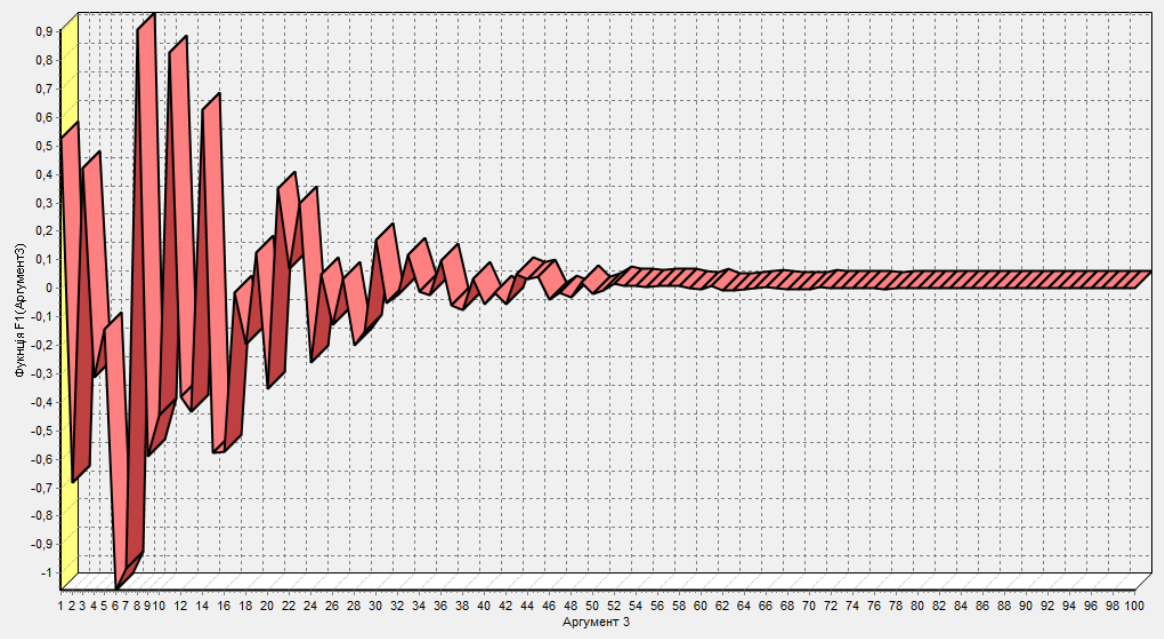


Рис. 6.7 – Графік обчисленої функції

Складна функція F2(Аргумент3) відрізняється тільки видом функції («10*SIN(COL3*COL3/100)/(COL3+1)*EXP(-COL3/10)») (рис. 6.8).

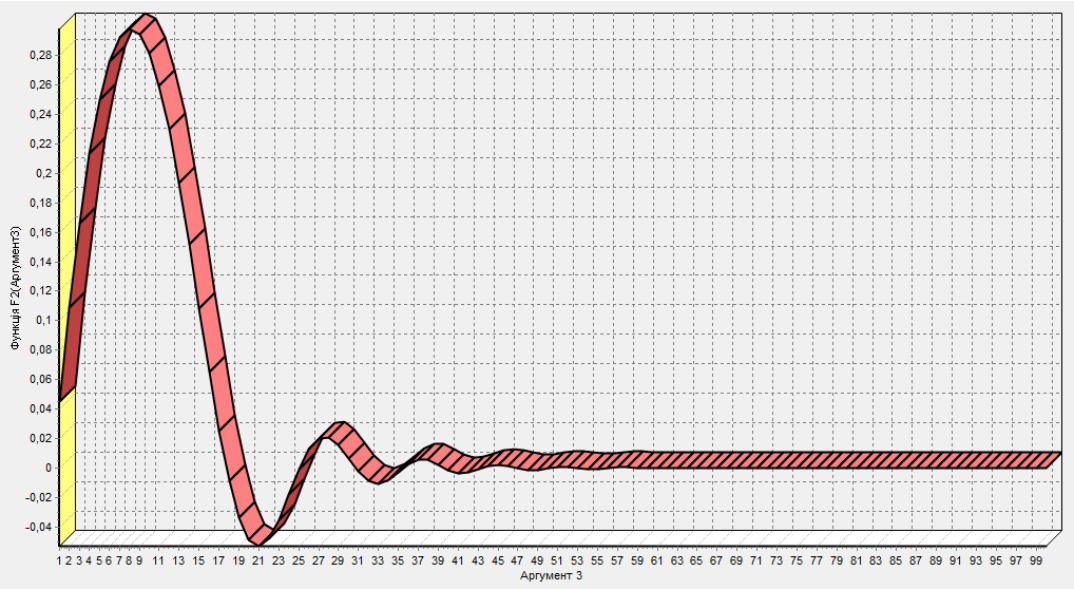


Рис. 6.8 – Графік складної функції

Функція від двох аргументів F3(Аргумент1; Аргумент2)

Дана функція цікава тим, що для перегляду в трьох вимірах можна використовувати візуалізатор «Куб». Задамо назву виразу «F3(Аргумент1;Аргумент2)», у полі обчислюваного виразу напишемо «COL1*COL1/100–COL2*COL2/100». Виберемо візуалізатор "Куб" і налаштуємо його так, що "Аргумент1" і "Аргумент2" були б вимірами, F3 (Аргумент1;Аргумент2) - фактом, а "Аргумент3" – не використовується (рис. 9).

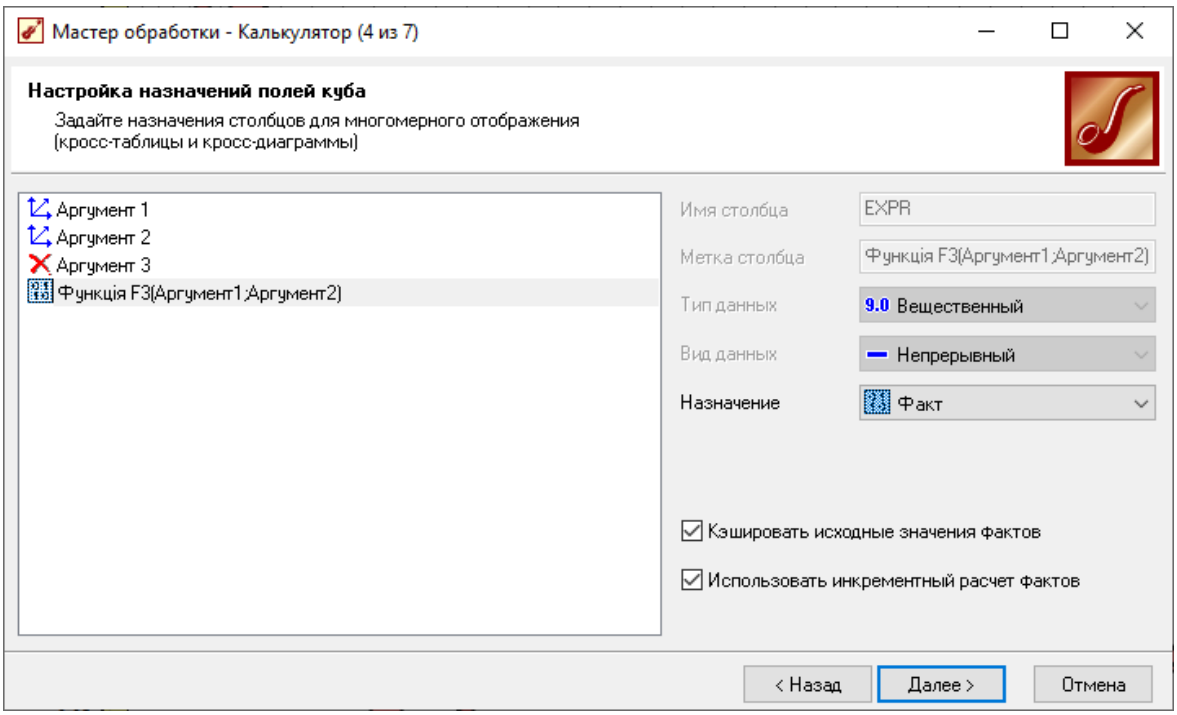


Рис. 6.9 – Налаштування призначень полів куба

Вибравши «Аргумент1» виміром у стовбцях, а «Аргумент2» – виміром у рядках перейдемо до перегляду Крос-діаграми (рис. 6.10). Для наочного перегляду встановимо тип діаграми «області» –  . Тепер можна переглянути обчислену функцію в об'ємному вигляді (якщо діаграма має вигляд не такий, як показано на рисунку, то натисніть кнопку «Транспонування»).

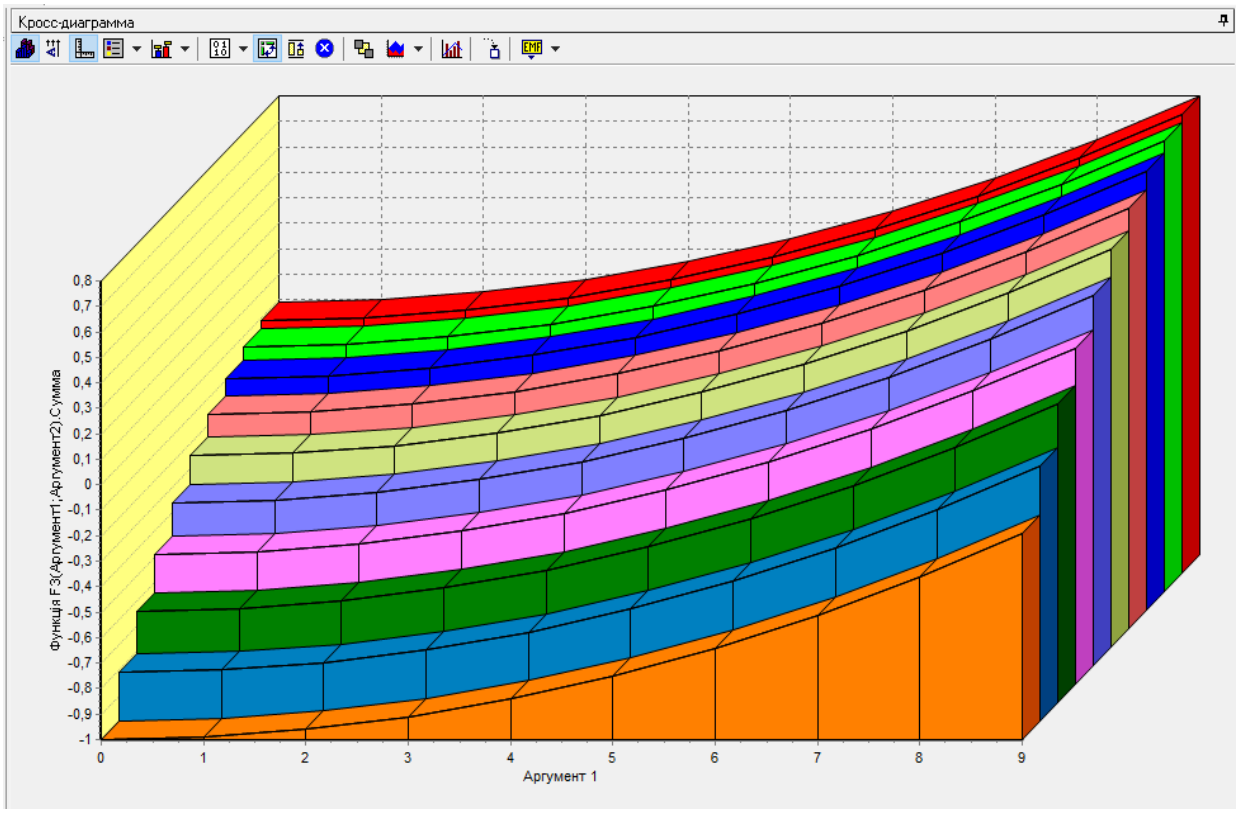


Рис. 6.10 – Крос-діаграма

Тепер обчислимо пайове відхилення Аргумент1+1 від Аргумент2+1 (RELDEV). Задавши в якості виразу, що обчислюється RELDEV(COL1+1;COL2+1) можна на діаграмі побачити дане відхилення (рис. 6.11).

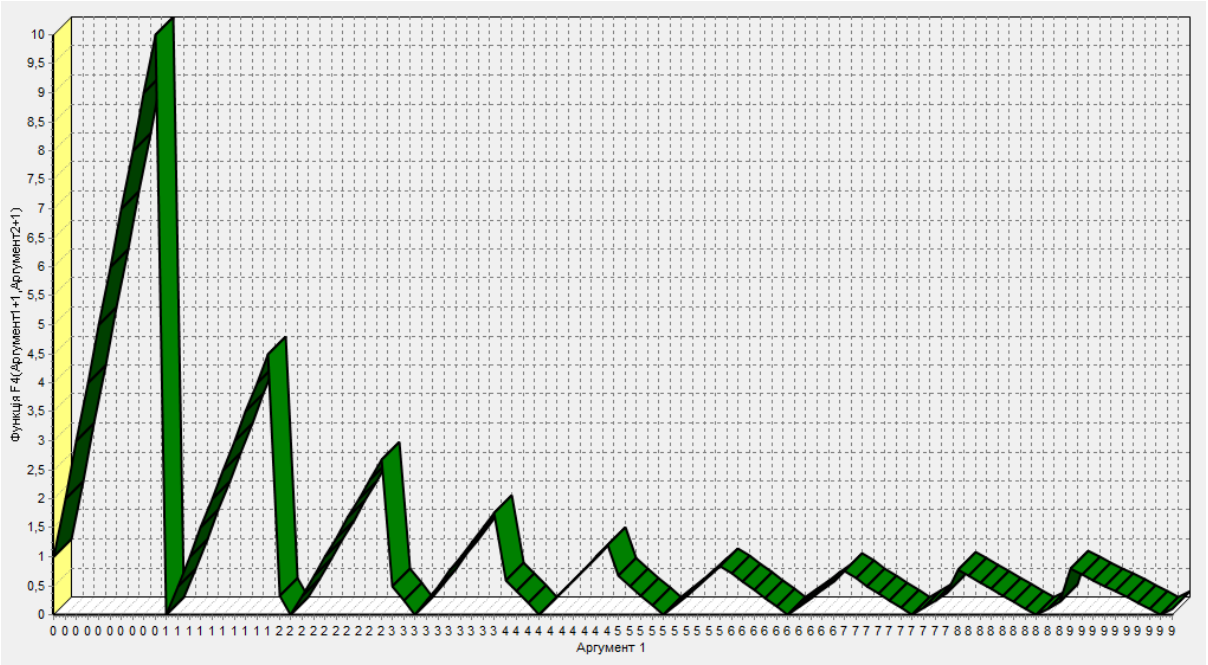


Рис. 6.11 – Діаграма відхилення

Нехай функція приймає значення SQRT(Аргумент3/50) (квадратний корінь) при значеннях Аргумент3 від 0 до 50 і значення Аргумент3* Аргумент3/2500 при інших. Для обчислення подібної функції необхідно скористатися функцією IFF(Аргумент1;Аргумент2;Аргумент3), що є в наявності, яка дозволяє в залежності від логічного значення першого аргументу отримати другий або третій аргумент. Якщо значення аргументу більше нуля і менше 50 необхідно отримати вираз SQRT(Аргумент3/50), в іншому випадку – вираз Аргумент3*Аргумент3/2500.

Таким чином, в полі побудови виразу необхідно написати «IFF((COL3>0)AND(COL3<50);SQRT(COL3/50);COL3*COL3/2500) ». Зробивши це в майстрі обробки «Калькулятор», і обравши далі візуалізатор «Діаграма», і також обравши в майстрі налаштування

діаграми поле зі значеннями кусково-заданої функції, можна отримати необхідний результат (рис. 6.12).

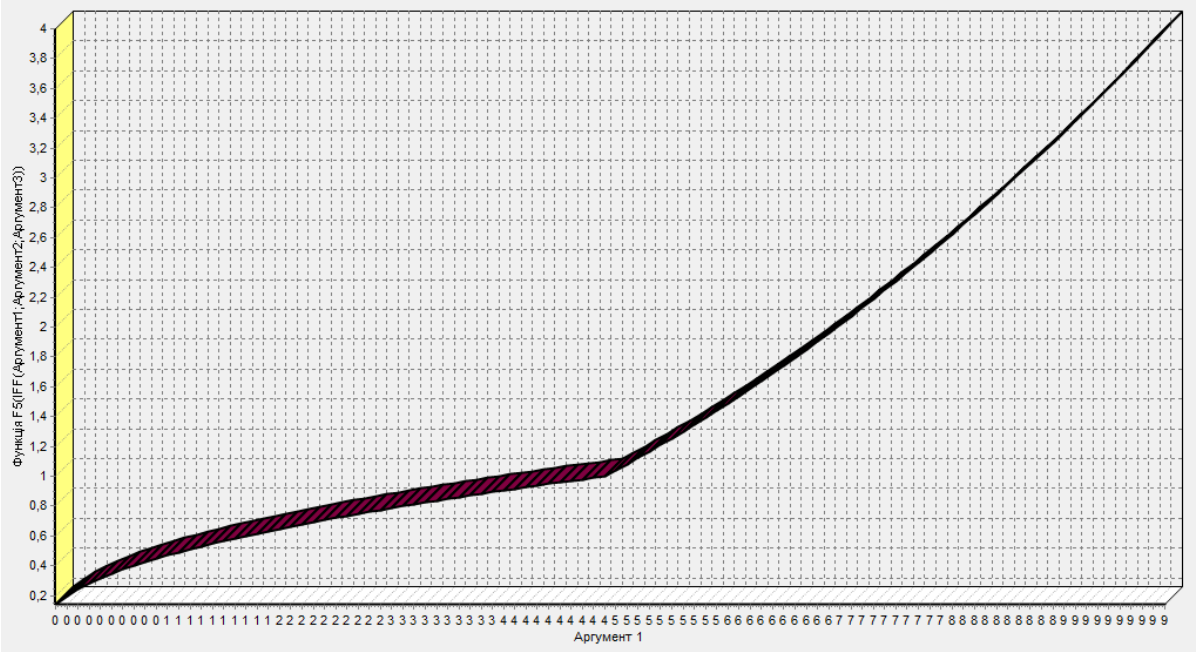


Рис. 6.12 – Діаграма кусково-заданої функції

В 5 частині навчилися застосовувати інструмент «Калькулятор» системи Deductor Studio Academic.

Питання до розділу 6

1. Для чого необхідно формування нових даних з використанням математичних функцій?
2. Який інструмент є в системі Deductor Studio для створення даних за допомогою математичних функцій?
3. Наведіть приклади необхідності використання математичних функцій для проведення аналізу даних.
4. В чому особливості інструмента «Калькулятор»?

Розділ 7
ПОШУК АСОЦІАТИВНИХ ПРАВИЛ ДЛЯ ВСТАНОВЛЕННЯ ЗАЛЕЖНОСТЕЙ МІЖ ПОДІЯМИ

7.1 Асоціативні правила – метод Data Mining

Асоціативні правила дозволяють знаходити закономірності між пов'язаними подіями. Прикладом такого правила є твердження, що покупець, який купує «Хліб», придбає і «Молоко». Вперше це завдання було запропоновано для пошуку асоціативних правил для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізом ринкового кошика (market basket analysis). Нехай є база даних, що складається з купівельних транзакцій. Кожна транзакція – це набір товарів, куплених покупцем за візит. Таку транзакцію ще називають ринковим кошиком.

Метою аналізу є встановлення наступних залежностей: якщо в транзакції зустрівся деякий набір елементів X, то на підставі цього можна зробити висновок про те, що інший набір елементів Y також повинен з'явитися в цій транзакції. Встановлення таких залежностей дає змогу знаходити дуже прості та інтуїтивно зрозумілі правила.

Основними характеристиками таких правил є підтримка та достовірність.

Правило «з X слідує Y» має підтримку s, якщо s% транзакцій з усього набору, містять набори елементів X і Y. Достовірність правила показує, яка ймовірність того, що з X випливає Y.

Правило «з X слідує Y» справедливо з достовірністю c, якщо c% транзакцій з усієї множини, що містять набір елементів X, також містять набір елементів Y.

Покажемо на конкретному прикладі: нехай 75% транзакцій, що містять хліб, також містять молоко, а 3% від загальної кількості всіх транзакцій містять обидва товари. 75% – це достовірність правила, а 3% – це підтримка.

Алгоритми пошуку асоціативних правил призначені для знаходження всіх правил виду «з X слідує Y», причому підтримка і достовірність цих правил повинні перебувати в рамках деяких наперед заданих меж, званих відповідно мінімальною та максимальною підтримкою та мінімальною та максимальною достовірністю.

Межі значень параметрів підтримки та достовірності вибираються таким чином, щоб обмежити кількість знайдених правил. Якщо підтримка має велике значення, то алгоритми будуть знаходити правила, добре відомі аналітикам або настільки очевидні, що немає сенсу проводити такий аналіз. З іншого боку, низьке значення підтримки веде до генерації величезної кількості правил, що потребує суттєвих обчислювальних ресурсів. Проте більшість цікавих правил перебуває саме за низького значення порога підтримки. Хоча занадто низьке значення підтримки веде до створення статистично необґрунтованих правил.

Таким чином, необхідно знайти компроміс, який би, по-перше, забезпечував цікавість правил і, по-друге, їх статистичну обґрунтованість. Тому значення цих меж залежать від характеру аналізованих даних і підбираються індивідуально. Ще одним параметром, що обмежує кількість знайдених правил є максимальна потужність множин, що часто зустрічаються. Якщо цей параметр вказано, то при пошуку правил будуть розглядатися лише множини, кількість елементів яких буде не більшою за цей параметр. Отже, будь-яке знайдене правило складатиметься не більше, ніж з максимальної потужності елементів.

Популярні набори – це множини, що складаються з одного і більше елементів, які найчастіше зустрічаються в транзакціях одночасно. На скільки часто зустрічається множина у вихідному наборі транзакцій, можна судити з підтримки. Цей візуалізатор відображає множини у вигляді списку.

7.2 Візуалізатори відображення асоціативних правил

Візуалізатор «*Правила*» відображає асоціативні правила у вигляді списку правил.

Візуалізатор «*Дерево правил*» – це завжди дворівневе дерево. Воно може бути побудовано або за умовою, або за наслідком. При побудові дерева правил за умовою, на першому (верхньому) рівні перебувають вузли з умовами, а на другому рівні – вузли з наслідком.

Другий варіант дерева правил – дерево, побудоване за наслідком. Тут на першому рівні розташовуються вузли з наслідком.

Праворуч від дерева знаходиться перелік правил, побудований за вибраним вузлом дерева. Для кожного правила відображаються підтримка та достовірність. Якщо дерево побудовано за умовою, то у верхній частині списку відображається умова правила, а список складається з його наслідків. Тоді правила відповідають на запитання, що буде за такої умови. Якщо дерево побудоване за наслідком, то вгорі списку відображається наслідок правила, а список складається з його умов. Ці правила відповідають питанням, що потрібно, щоб було задане слідство. Даний візуалізатор відображає ті самі правила, що й попередній, але в більш зручній для аналізу формі.

Аналіз «*Що-якщо*» в асоціативних правилах дозволяє відповісти на питання що отримаємо як наслідок, якщо виберемо ці умови? Наприклад, які товари купуються разом із вибраними товарами. У вікні зліва розташований список усіх елементів транзакцій. Праворуч від кожного елемента вказана підтримка – скільки разів цей елемент зустрічається у транзакціях.

У верхньому правому куті розташований список елементів, що входять в умову. Це, наприклад, перелік товарів, які придбав покупець. Для них слід знайти наслідок. Наприклад, товари, які купуються разом із ними. Щоб запропонувати людині те, що він, можливо, забув купити.

У правому нижньому кутку розміщено перелік наслідків. Праворуч від елементів списку відображається підтримка та достовірність.

Результати аналізу можна застосувати і для сегментації покупців за поведінкою при покупках, і для аналізу переваг клієнтів, і для планування розташування товарів у супермаркетах, крос-маркетингу. Пропонований набір візуалізаторів дозволяє експерту знайти цікаві, незвичайні закономірності, зрозуміти, чому так відбувається і застосувати їх на практиці.

7.3 Використання методу обробки «Асоціативні правила»

За допомогою MS Excel створити таблицю, в якій представлена інформація щодо покупок продуктів кількох груп. У створеному файлі мають бути два стовпці «Номер чека» та «Товар», у кожному з яких має бути 147 значень. Кожен номер чека повинен мати від двох до п'яти найменувань, всього видів продуктів має бути близько семи. Зберегти файл з ім’ям «Supermarket.xlsx» (рис. 7.1).

	A	B
1	Номер чека	Товар
2	10101	макаронні вироби
3	10101	чай
4	10101	кетчуп, соуси, аджика
5	10211	кофе
6	10211	хлібо-булочні вироби
7	10211	макаронні вироби
8	10247	чай
9	10247	макаронні вироби
10	10247	чай і сири
11	10263	хлібо-булочні вироби
12	10263	макаронні вироби
13	10263	чай
14	10263	молоко і кофе
15	10269	макаронні вироби
16	10269	кетчуп, соуси, аджика
17	10269	чай і сири
18	10269	хлібо-булочні вироби
19	10278	чай і сири
20	10278	кофе
21	10281	макаронні вироби
22	10281	молоко і кофе

Рис. 7.1 – Зразок заповнення файлу «Supermarket.xlsx»

Експортувати «Supermarket.xlsx» в текстовий файл з роздільниками з назвою «Supermarket.txt» (рис. 7.2).

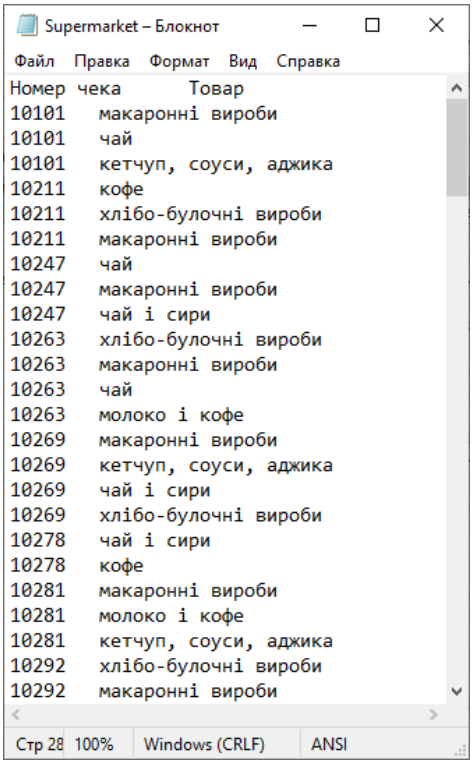


Рис. 7.2 – Зразок заповнення файлу «Supermarket.txt»

Необхідно розвязати задачу аналізу споживчого кошика з метою подальшого застосування результатів для стимулювання продажів.

Імпортуємо дані з файлу «Supermarket.txt». Основні моменти імпорту показані на рисунках 7.3, 7.4 та 7.5.

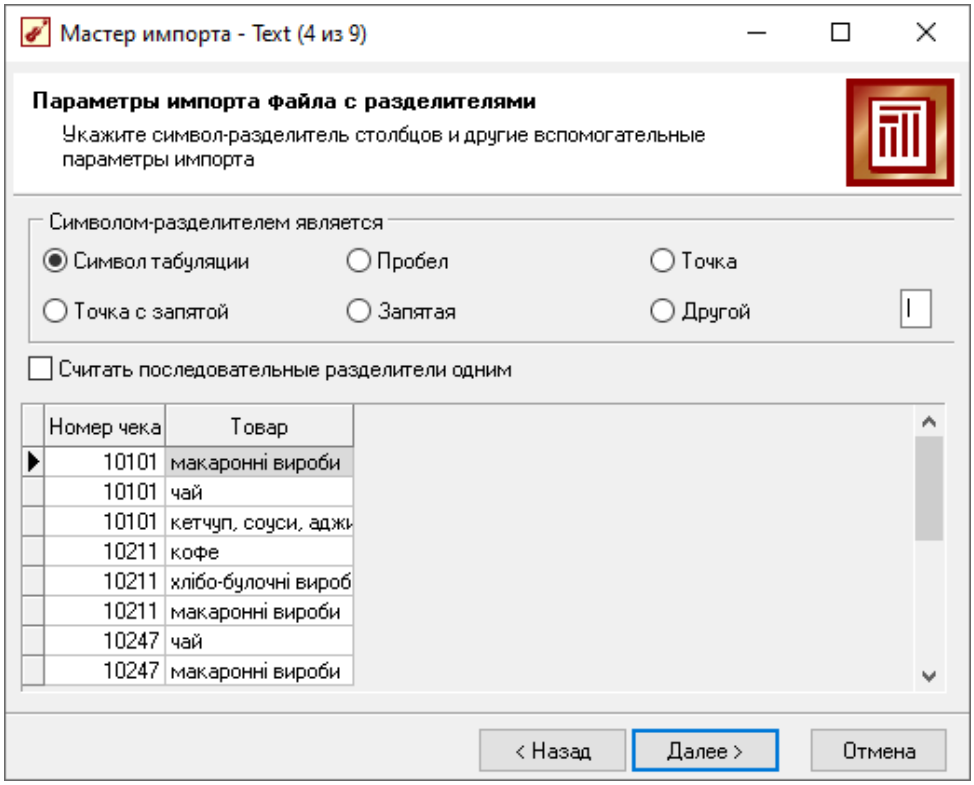


Рис. 7.3 – Параметри імпорту файлу «Supermarket.txt»

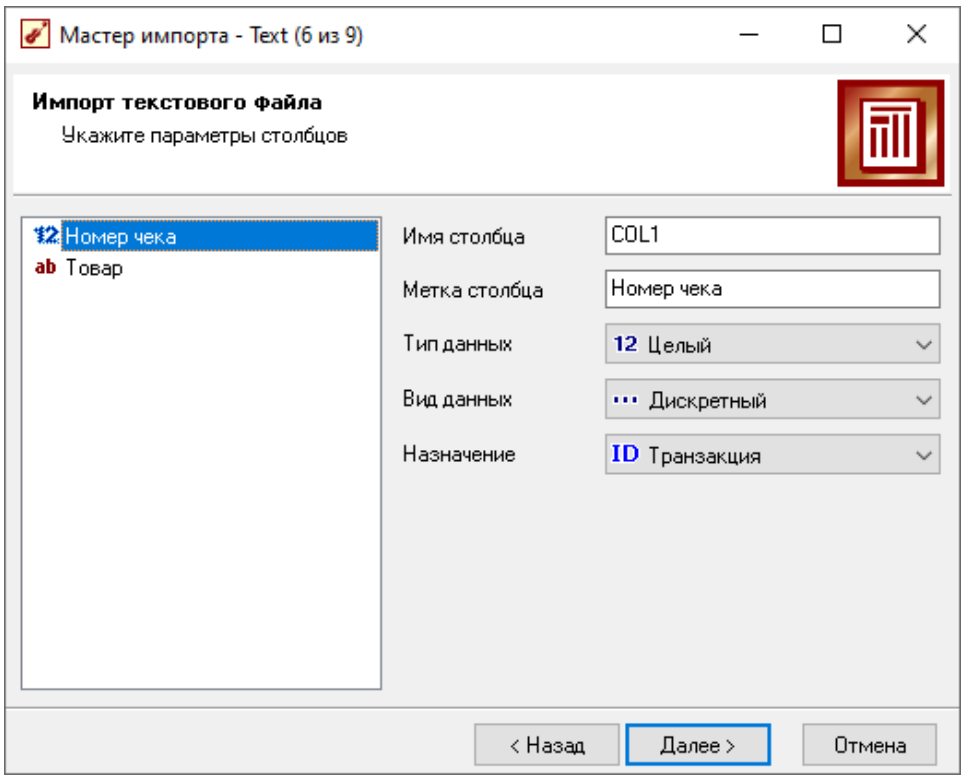


Рис. 7.4 – Параметри стовбця «Номер чека»

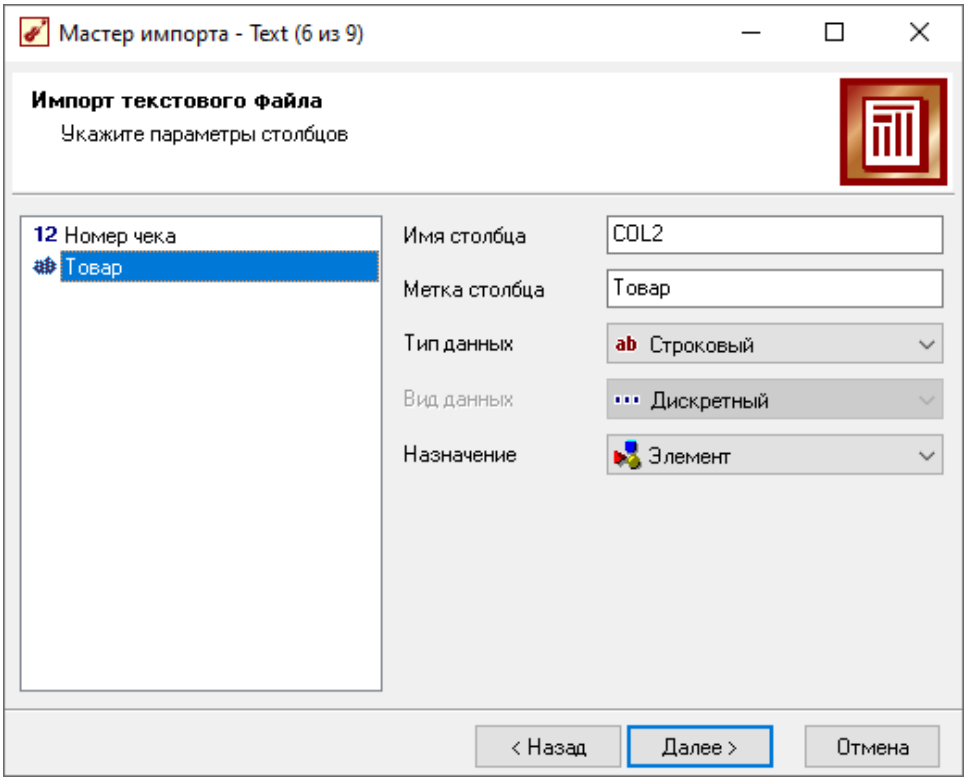


Рис. 7.5 – Параметри стовбця «Товар»

Для пошуку асоціативних правил запустимо майстер обробки (рис. 7.6). У ньому виберемо тип обробки "Асоціативні правила".

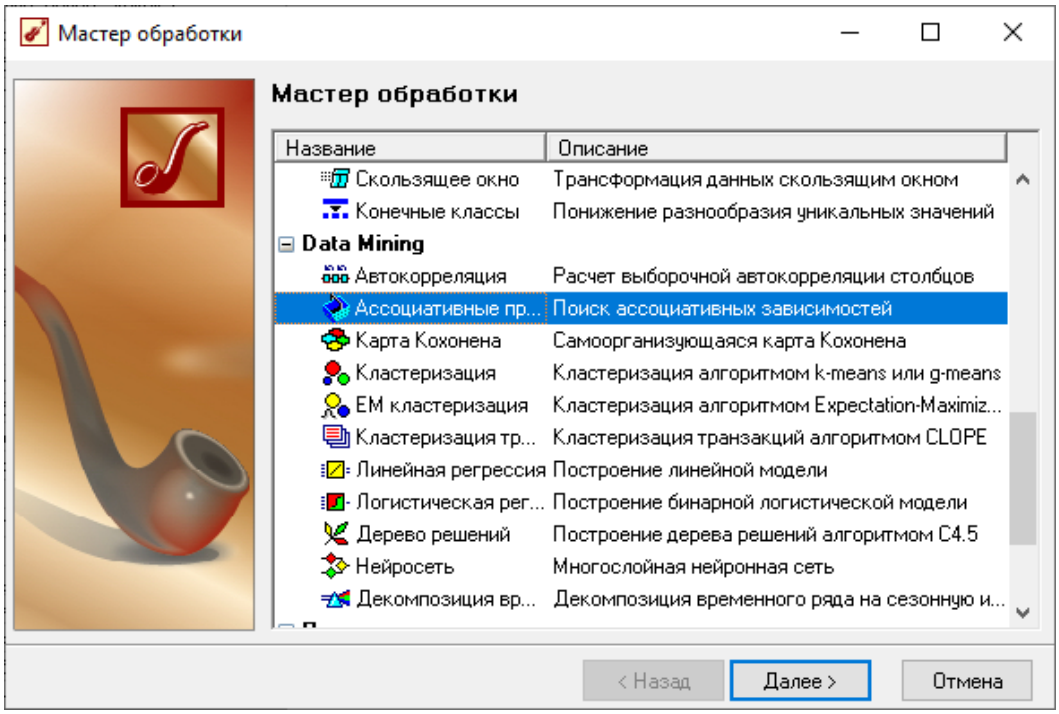


Рис. 7.6 – Вибір типу обробки даних

На другому кроці майстра необхідно вказати, який стовпець є ідентифікатором транзакції (чек), а який елемент транзакції (товар) (рис. 7.7).

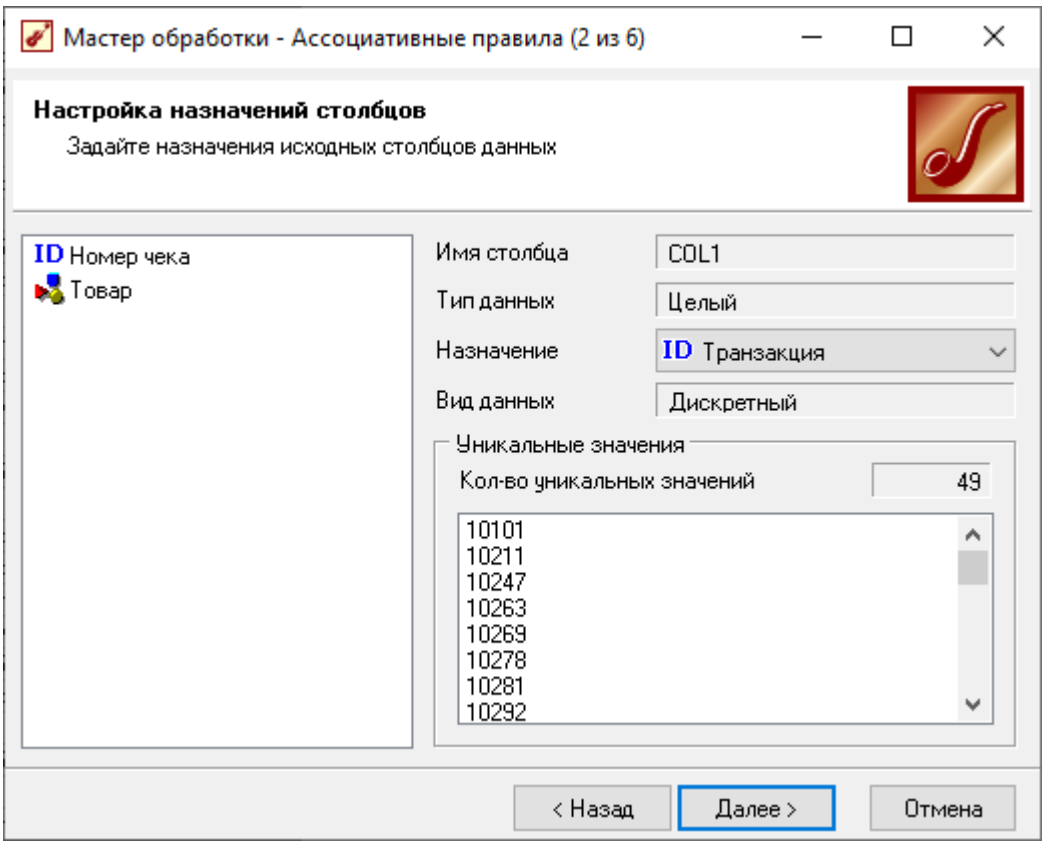


Рис. 7.7 – Налаштування призначення стовбців

Наступний крок дозволяє налаштувати параметри побудови асоціативних правил: мінімальну та максимальну підтримку, мінімальну та максимальну достовірність, а також максимальну потужність множини (рис. 7.8). Виходячи з характеру наявних даних, слід вказати межі підтримки – 13% та 80%, і достовірності 60% та 90%.

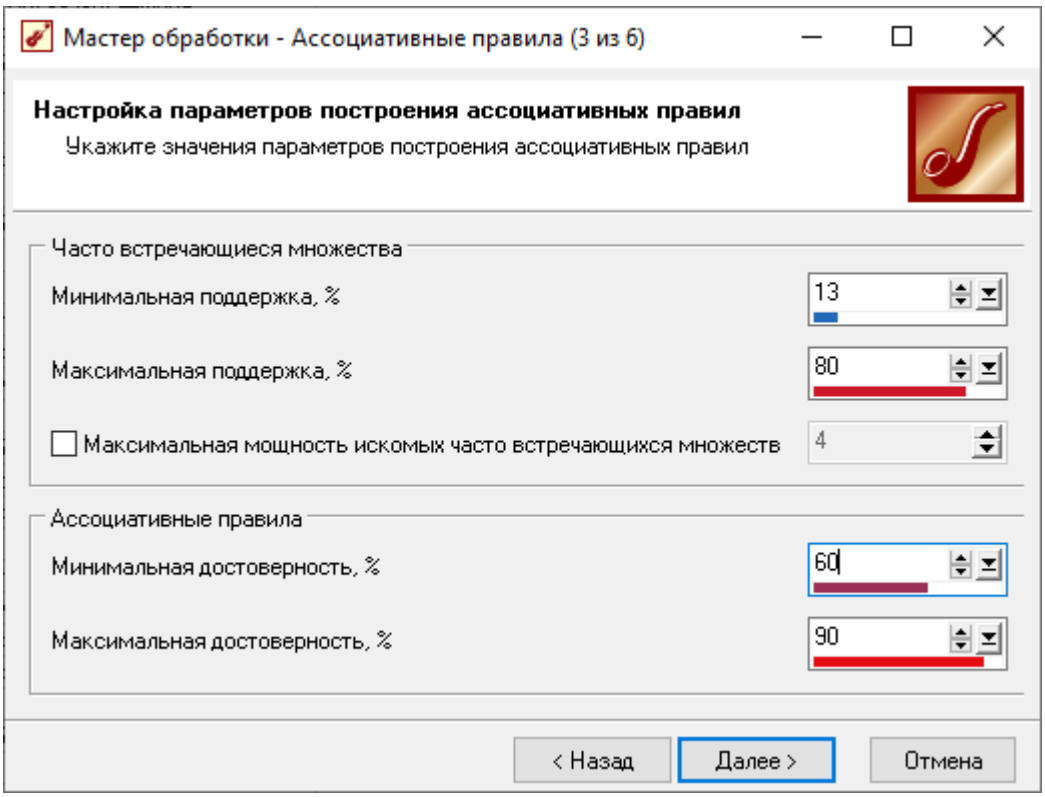


Рис. 7.8 – Налаштування параметрів побудови асоціативних правил

Наступний крок дозволяє запустити пошук асоціативних правил. На екрані відображається інформація про кількість множин, кількість знайдених правил, а також гістограма розподілу знайдених множин, що часто зустрічаються, за потужністю (рис. 7.9).

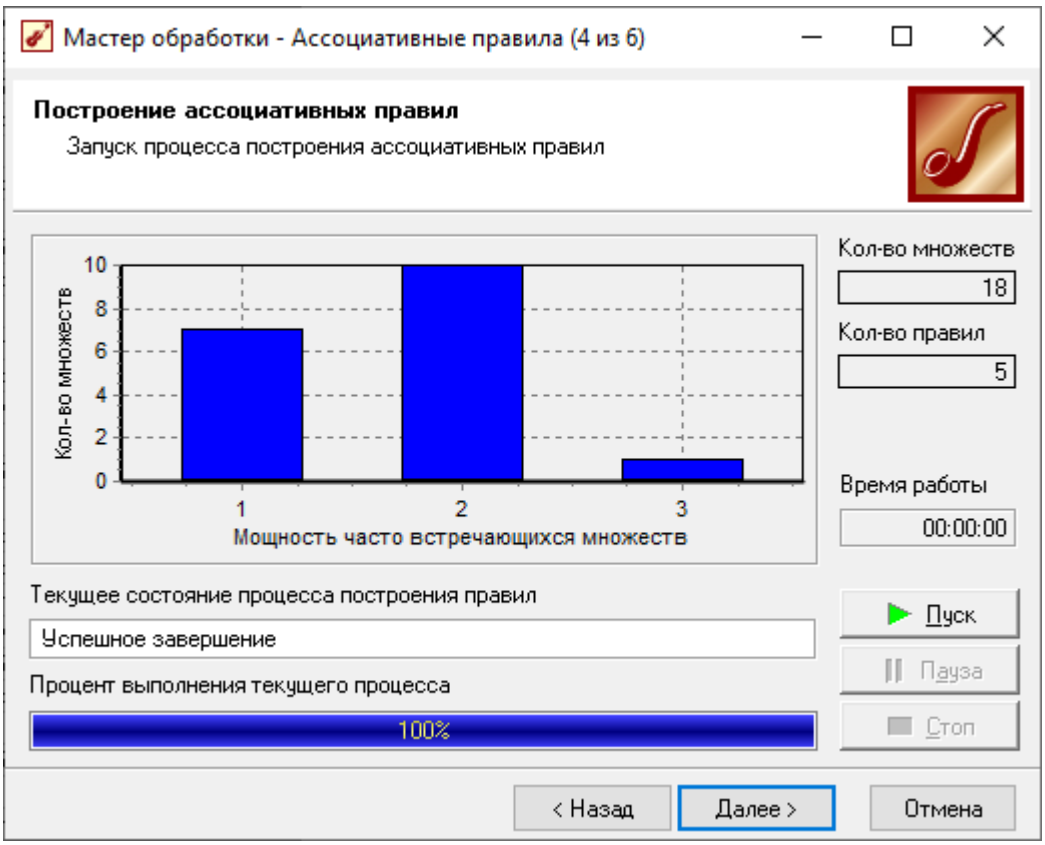


Рис. 7.9 – Результат процесу пошуку

На наступному кроці необхідно визначитися зі способом відображення даних. Для розв'язання задачі асоціативного пошуку вибираємо всі способи відображення даних з типу Data Mining (рис. 7.10).

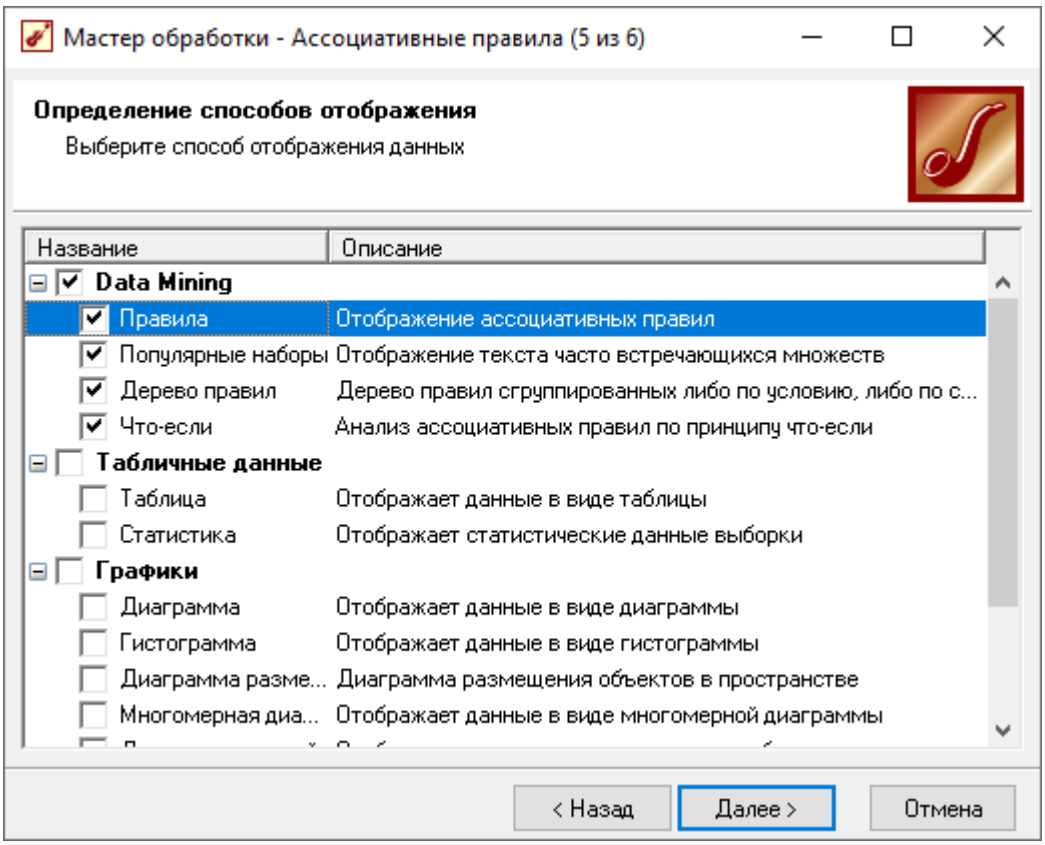
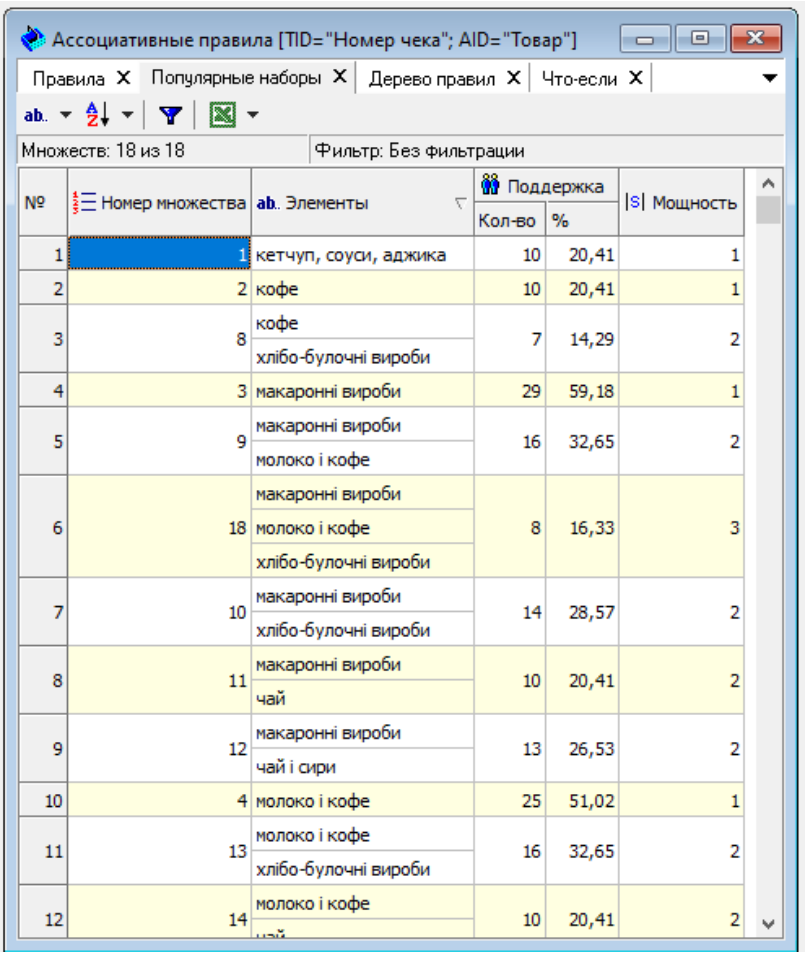


Рис. 7.10 – Вибір способу відображення даних

Після завершення процесу пошуку отримані результати можна подивитися, використовуючи спеціальні візуалізатори «Популярні набори», «Правила», «Дерево правил», «Що-якщо».

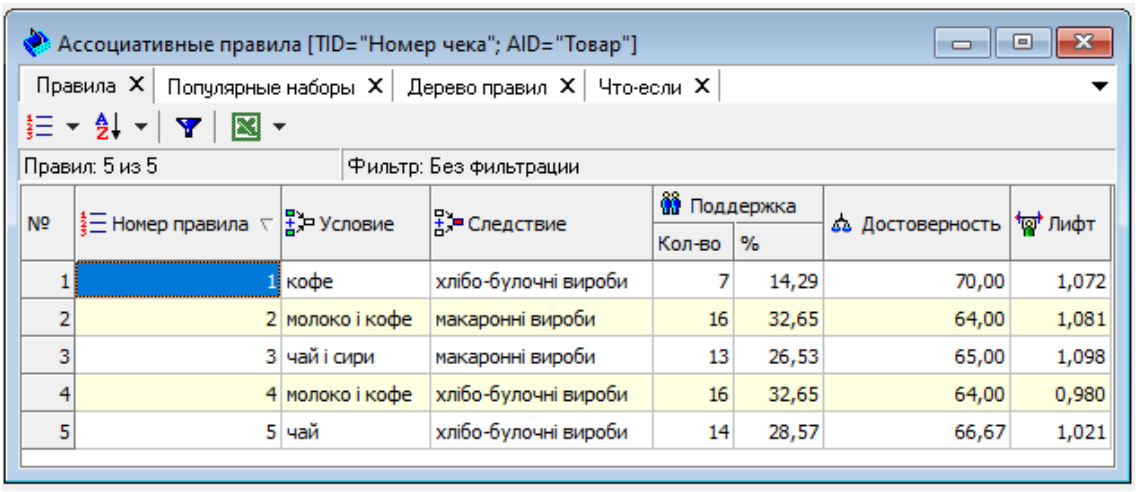
При використанні візуалізатора «Популярні набори», отримані набори товарів найчастіше купують у даній торговій точці, отже можна приймати рішення про постачання товарів, їх розміщення тощо. (рис. 7.11).



№	Номер множества	Элементы	Поддержка		Мощность
			Кол-во	%	
1	1	кетчуп, соуси, аджика	10	20,41	1
2	2	кофе	10	20,41	1
3	8	кофе хлібо-булочні вироби	7	14,29	2
4	3	макаронні вироби	29	59,18	1
5	9	макаронні вироби молоко і кофе	16	32,65	2
6	18	макаронні вироби молоко і кофе хлібо-булочні вироби	8	16,33	3
7	10	макаронні вироби хлібо-булочні вироби	14	28,57	2
8	11	макаронні вироби чай	10	20,41	2
9	12	макаронні вироби чай і сири	13	26,53	2
10	4	молоко і кофе	25	51,02	1
11	13	молоко і кофе хлібо-булочні вироби	16	32,65	2
12	14	молоко і кофе чай	10	20,41	2

Рис. 7.11 – Візуалізатор «Популярні набори»

Візуалізатор «Правила» виводить список асоціативних правил, представлений таблицею зі стовпцями: «номер правила», «умова», «наслідок», «підтримка, %», «підтримка, кількість», «достовірність», «ліфт» (рис. 7.12).



№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	кофе	хліб-булочні вироби	7	14,29	70,00	1,072
2	2	молоко і кофе	макаронні вироби	16	32,65	64,00	1,081
3	3	чай і сири	макаронні вироби	13	26,53	65,00	1,098
4	4	молоко і кофе	хліб-булочні вироби	16	32,65	64,00	0,980
5	5	чай	хліб-булочні вироби	14	28,57	66,67	1,021

Рис. 7.12 – Візуалізатор «Правила»

Таким чином, експерту надається набір правил, що описують поведінку покупців. Наприклад, якщо покупець купив каву, то він із ймовірністю 70% також купить і хліб-булочні вироби.

Використовуючи візуалізатор «Дерево правил» отримаємо наступне (рис. 7.13, 7.14).

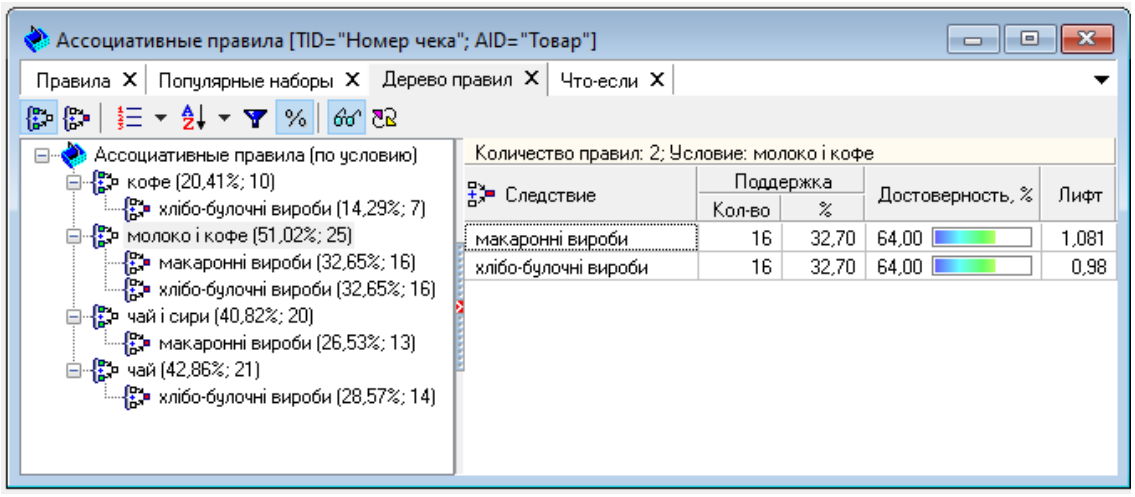


Рис. 7.13 – Візуалізатор «Дерево правил». Відображення за умовою

В даному випадку правила відображені за умовами. Тоді результат, що відображається в даний момент, можна інтерпретувати як 2 правила:

1. Якщо покупець придбав молоко і каву, то він із ймовірністю 64% також придбає макаронні вироби.

2. Якщо покупець придбав молоко і каву, то він із ймовірністю 64% також придбає хлібобулочні вироби.

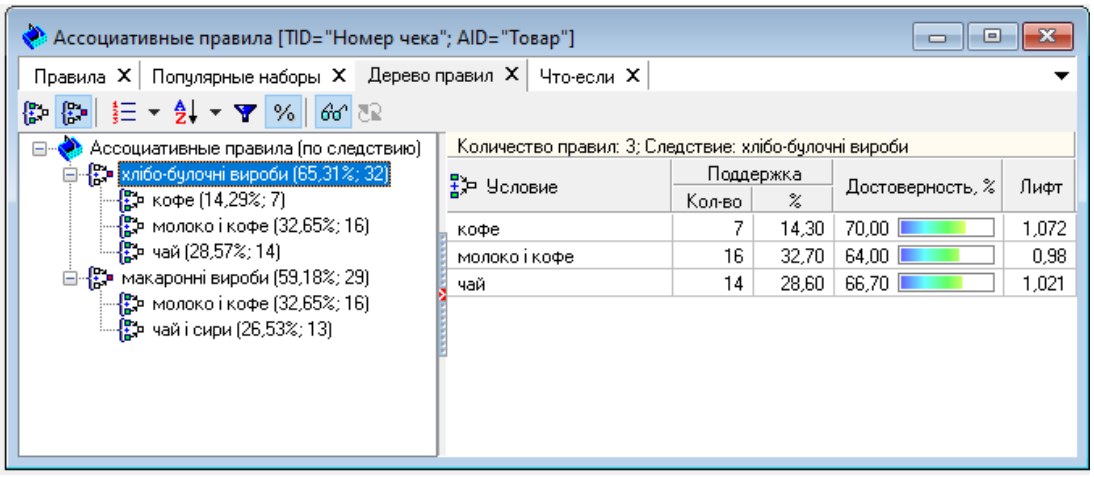


Рис. 7.14 – Візуалізатор «Дерево правил». Відображення за наслідком

При відображенні правил за наслідком, результат, який відображається на рисунку 7.14, можна прокоментувати так: якщо покупець вибирає хлібобулочні вироби, то з імовірністю 70% у цьому кошику вже є кава, з імовірністю 64% – молоко і кава і з імовірністю 66,7% – чай.

Нехай необхідно проаналізувати, що, можливо, забув покупець придбати, якщо він уже взяв хлібобулочні вироби, молоко та каву? Для цього необхідно додати до списку умов ці товари (наприклад, за

допомогою подвійного клацання миші) та натиснути на кнопку «Обчислити правила». При цьому в списку наслідків з'являться товари, які купуються разом з даними. У цьому випадку з'являться "МАКАРОННІ ВИРОБИ". Тобто можливо, покупець забув придбати макаронні вироби (рис. 7.15).

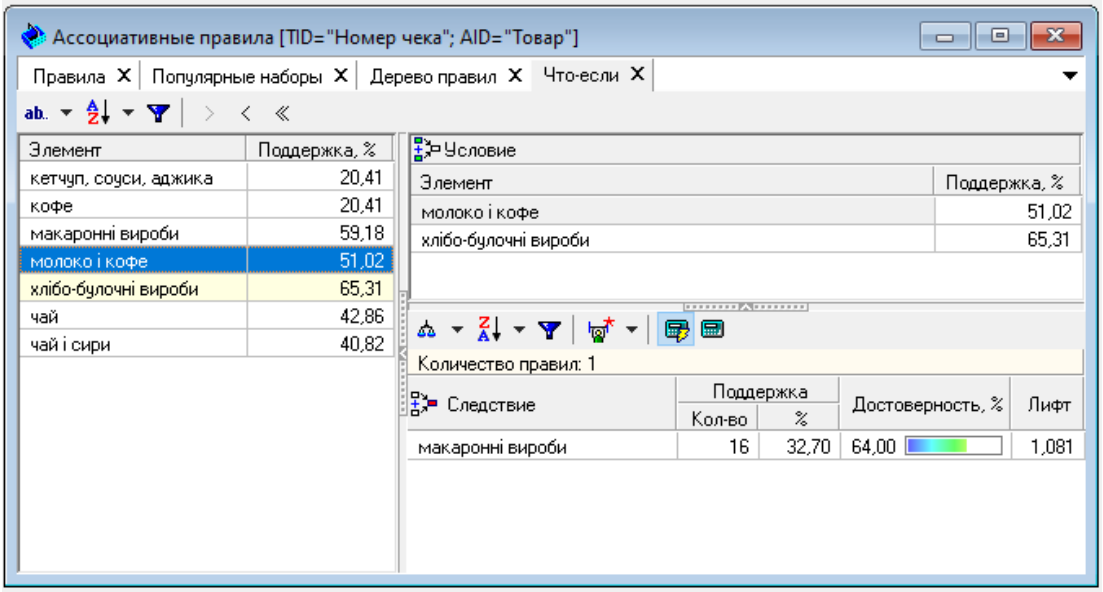


Рис. 7.15 – Візуалізатор «Що-якщо»

В 7 розділі застосовували асоціативні правила та використовували візуалізатори «*Популярні набори*», «*Правила*», «*Дерево правил*», «*Що-якщо*».

Питання до розділу 7

1. Що таке асоціативні правила?
2. Як створюються асоціативні правила?
3. Навіщо використовуються асоціативні правила під час аналізу даних?

4. Що таке достовірність правила?
5. Що таке підтримка правила?
6. Які інструменти для побудови асоціативних правил є у системі Deductor?
7. Що таке дерево правил?
8. Які варіанти створення дерева правил існують у Deductor?
9. Наведіть приклад одержаних результатів аналізу даних за допомогою асоціативних правил.

Розділ 8
ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

8.1 Аналіз часових рядів в Deductor Studio

Важливим чинником для аналізу часового ряду та прогнозу є визначення сезонності. У Deductor Studio таким інструментом є автокореляція.

Метою автокореляційного аналізу є з'ясування степені статистичної залежності між різними значеннями (відліками) довільної послідовності, яку утворює поле вибірки даних. Якщо їхня кореляція дорівнює одиниці, то величини прямо залежні один від одного, якщо нулю – то ні, якщо мінус одиниця, то залежність зворотна. У процесі автокореляційного аналізу розраховуються коефіцієнти кореляції (захід взаємної залежності) для двох значень вибірки, що віддаляються один від одного на певну кількість відліків, звані також лагом.

Стосовно аналізу часових рядів автокореляція дозволяє виділити місячну та річну сезонність у даних. Видно, що пік залежності на даних припадає на 12 місяць, що свідчить про річну сезонність. Тому величину продажів річної давності необхідно обов'язково враховувати при побудові моделі (якщо використовується нейронна мережа, то подавати на вхід).

Лінійна автокореляція шукає залежності між значеннями однієї й тієї ж величини, але в різний час, тому знаходження лінійної

автокореляційної залежності і застосовується для визначення періодичності (сезонності) при обробці часових рядів.

Прогноз часового ряду. Прогнозування результату на певний час вперед, ґрунтуючись на даних за минулий час – завдання, що зустрічається досить часто (наприклад, перед більшістю торгових фірм стоїть завдання оптимізації складських запасів, для розв'язання якої потрібно знати, чого і скільки має бути продано через тиждень, тощо; завдання передбачення вартості акцій якого-небудь підприємства через день тощо та інші подібні питання). Deductor Studio пропонує для цього інструмент «Прогнозування».

Прогнозування з'являється у списку майстра обробки лише після побудови будь-якої моделі прогнозу: нейромережі, лінійної регресії тощо. Прогнозувати на кілька кроків уперед має сенс лише часовий ряд (наприклад, якщо є дані щодо тижневих сум продажу за певний період, можна спрогнозувати суму продажів на два тижні вперед).

Обробник «Нейромережа». Обробник призначений для розв'язання задач регресії та прогнозування. У разі нейромережа будується для прогнозування майбутніх значень часового ряду. Для перевірки узагальнюючої здатності нейромережі рекомендується розбити наявну множину даних на дві частини: навчальне та тестове. Як правило, при прогнозуванні часових рядів частка тестової множини становить не більше 10-20%.

За допомогою візуалізатора «Діаграма» оцінюється здатність побудованої нейромережевої моделі до узагальнення. Для цього в

одному вікні виводяться графіки вихідного та спрогнозованого часових рядів.

8.2 Використання методу обробки «Автокореляція»

За допомогою MS Excel створити файл «Trade.xlsx» з даними (рис. 8.1), що містять історію продажів за певний період. Файл повинен містити два стовпці «Дата (Рік+Місяць)» (формату РРРР-ММ) та «Кількість» (десяткове число). Дані повинні бути зібрані за 5 років. Експортувати файл у додаток блокнот з ім’ям «Trade.txt» (рис. 8.2).

Визначити чи є сезонність, якщо є, то яка.

Яка кількість товару буде продана через тиждень та через два.

	A	B
1	Дата (Рік+Місяць)	Кількість
2	2017-M01	462523,419
3	2017-M02	633208,196
4	2017-M03	660159,299
5	2017-M04	617455,3417
6	2017-M05	597354,4794
7	2017-M06	793517,4512
8	2017-M07	1015944,286
9	2017-M08	1148052,252
10	2017-M09	1156623,172
11	2017-M10	1255021,942
12	2017-M11	1410114,561
13	2017-M12	1357230,339
14	2018-M01	954679,341
15	2018-M02	1034982,213
16	2018-M03	1427864,541
17	2018-M04	1432592,087
18	2018-M05	1643252,194
19	2018-M06	1796543,902
20	2018-M07	1865043,412
21	2018-M08	1984347,435
22	2018-M09	1647706,367

Рис. 8.1 – Приклад заповнення файлу «Trade.xlsx»

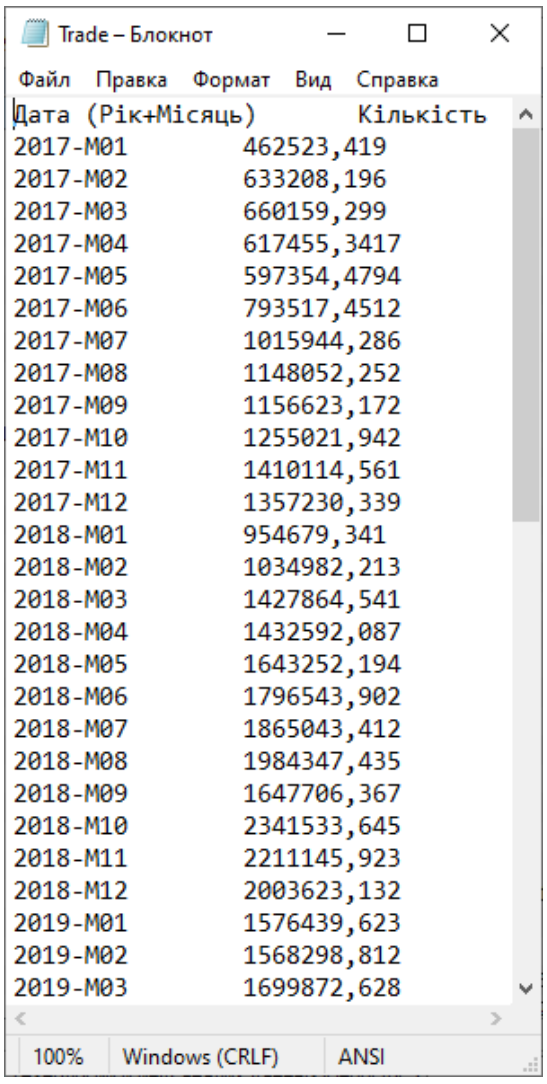


Рис. 8.2 – Вигляд експортованого файлу «Trade.txt»

Імпортуємо дані із текстового файлу. Виберемо як візуалізатор діаграму для перегляду вихідної інформації (рис. 8.3).

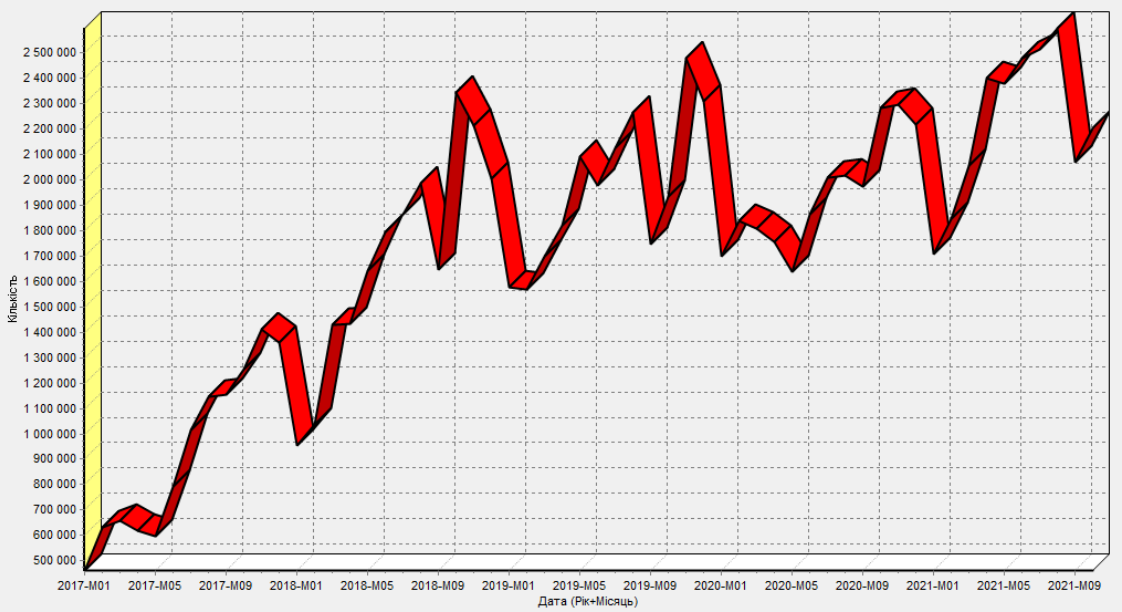


Рис.8.3 – Діаграма даних

Як видно, не кожен аналітик зможе судити про сезонність за цими даними, тому необхідно скористатися автокореляцією. Для цього відкриємо майстер обробки, виберемо як обробку автокореляцію (рис. 8.4) і перейдемо на другий крок майстра.

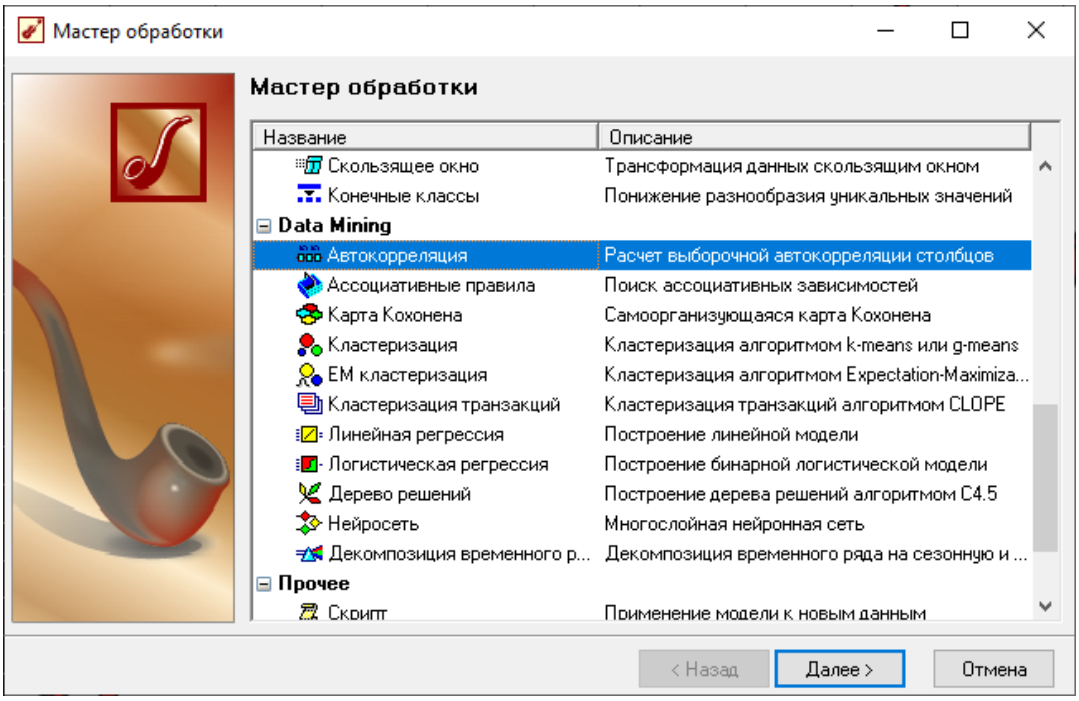


Рис.8.4 – Вибір методу обробки «Автокореляція»

У ньому необхідно налаштувати параметри стовбців (рис. 8.5). Вкажемо поле «Дата (Рік+Місяць)» невикористовуваним, а поле «Кількість» використовуваним (адже необхідно визначити сезонність кількості продажів). Припустимо, що сезонність, якщо вона має місце, не більша за рік. У зв'язку з цим поставимо кількість відліків рівним 15 (тоді шукатиметься залежність від місяця назад, двох, ..., п'ятнадцяти місяців назад). Також має стояти прапорець "Увімкнути поле відліків набір даних". Він необхідний для більш зручнішої інтерпретації автокореляційного аналізу.

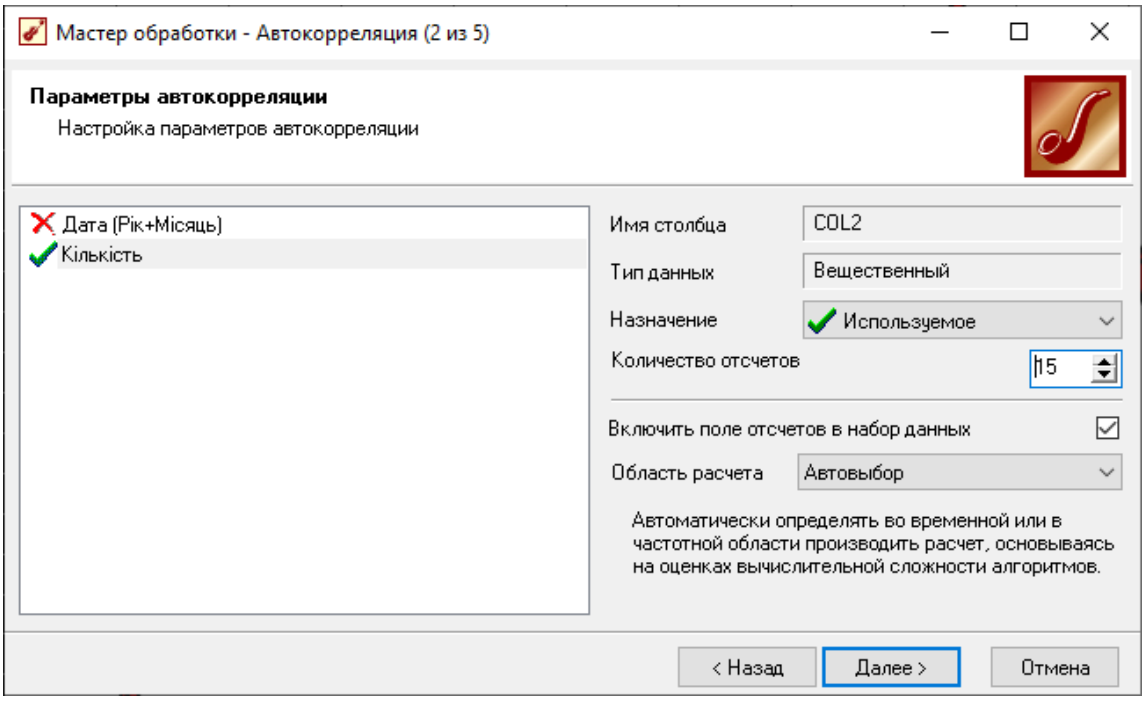


Рис. 8.5 – Майстер обробки

Перейдемо на наступний крок майстра та запустимо процес обробки.

По закінченню результати зручно аналізувати як у вигляді таблиці, так і у вигляді діаграми. Після обробки було отримано два стовпці – «Лаг» (завдяки встановленому прапорцю в майстрі) та «КІЛЬКІСТЬ» – результат автокореляції (рис. 8.6).

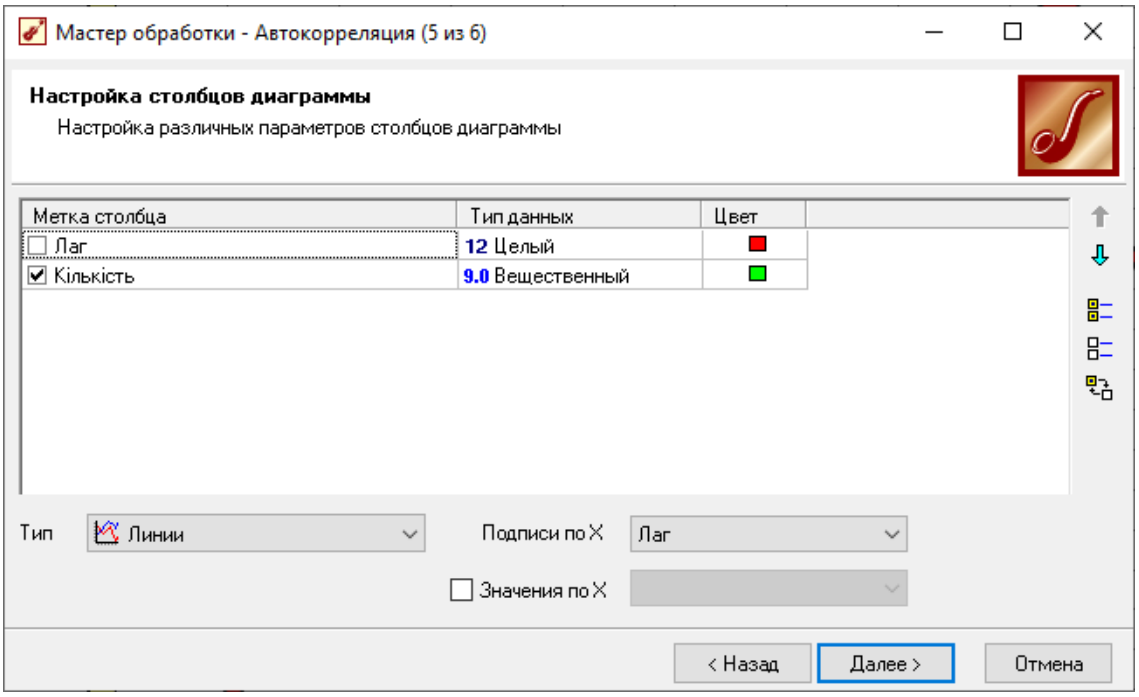


Рис. 8.6 – Налаштування стовбців діаграми

Результат автокореляції у вигляді діаграми показаний на рис. 8.7.

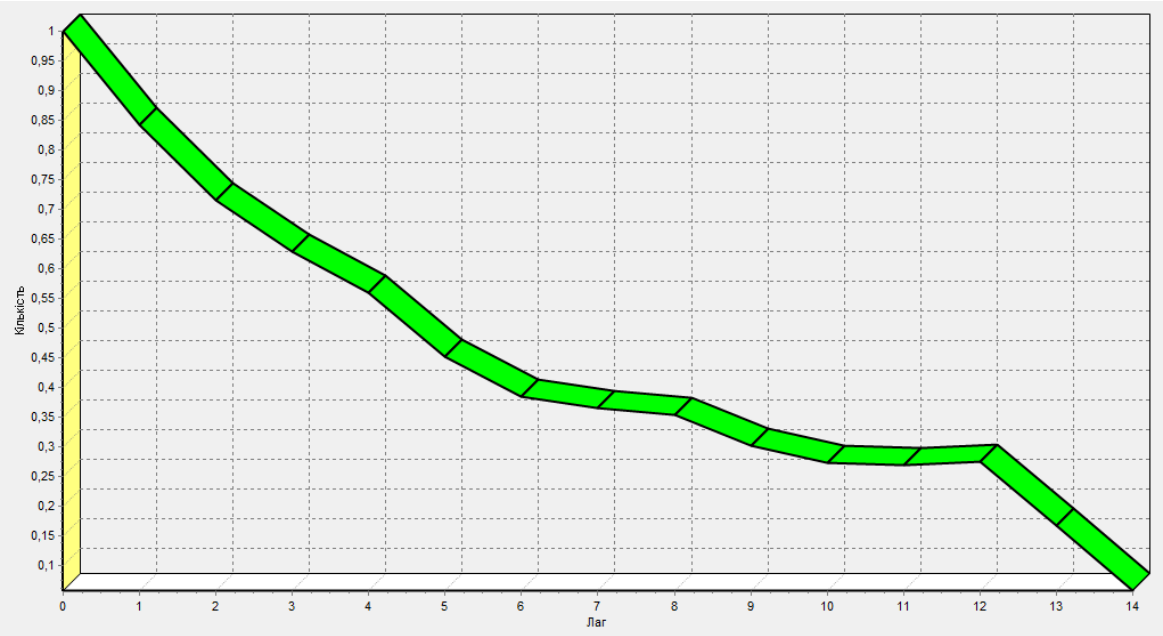


Рис. 8.7 – Підсумкова діаграма

Видно, що спочатку кореляція дорівнює одиниці – це те, як значення залежить саме від себе. Далі залежність зменшується і потім видно невеликий пік залежність від даних 12 місяців тому. Це якраз і говорить про наявність річної сезонності.

8.3 Редагування викидів

Після імпорту даних скористаємося діаграмою для перегляду (рис. 8.8). На ній видно, що дані містять аномалії (викиди) та шуми, за якими важко розглянути тенденцію.

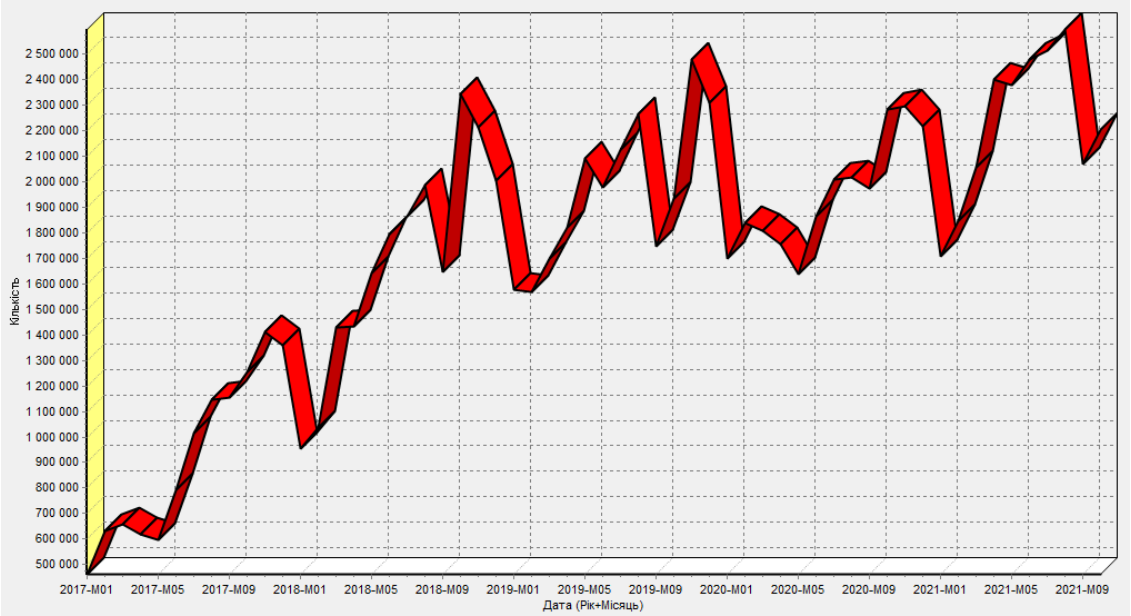


Рис. 8.8 – Діаграма даних

Тому перед прогнозуванням необхідно видалити аномалії і згладити дані. Зробити це можливо за допомогою «Редагування викидів».

Запустимо майстер обробки, виберемо в якості обробки даних «Редагування викидів» и перейдемо на наступний крок майстра (рис. 8.9).

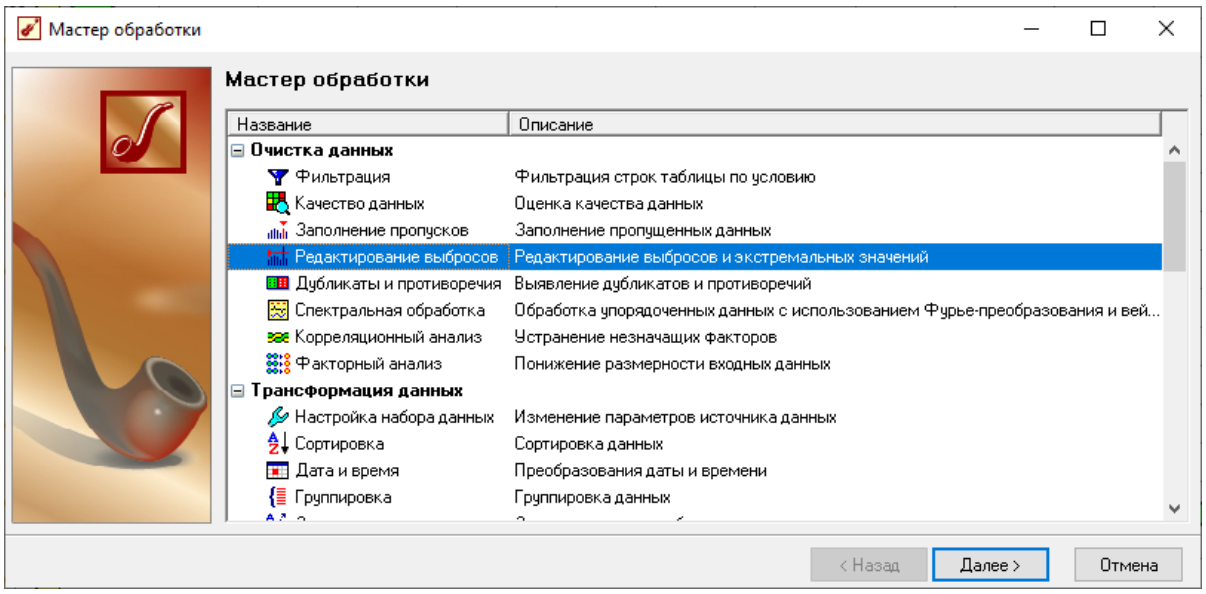


Рис. 8.9 – Вибір обробки даних

Другий крок майстра відповідає за обробку пропущених значень, яких у початкових даних немає, тому тут нічого не налаштовуємо, тільки потрібно поставити мітку для обробки упорядкованого набору даних (рис. 8.10).

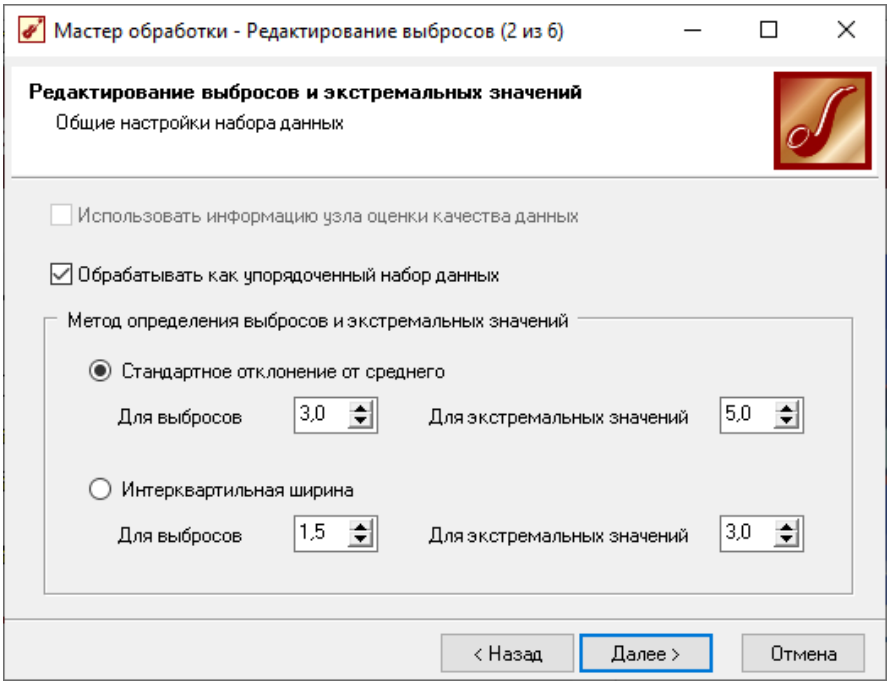


Рис. 8.10 – Загальні налаштування набору даних

Наступний крок відповідає за видалення аномалій з початкового набору.

Виберемо поле для обробки «Кількість» та вкажемо для нього обробку аномальних явищ (ступінь придушення – велика). Для спектральної обробки з початкових даних необхідно виключити шуми, тому вибираємо стовпець «Кількість» і вказуємо спосіб обробки «згладжувати» (ступінь віднімання – велика) (рис. 8.11).

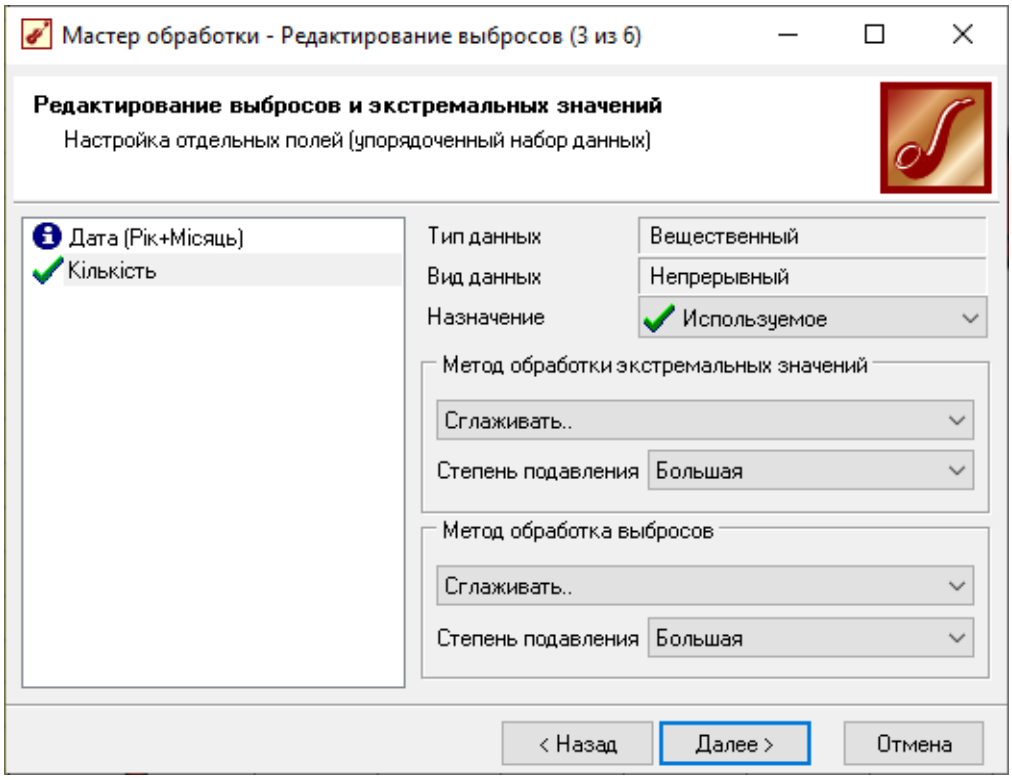


Рис. 8.11 – Налаштування окремих полів

На наступному етапі запустимо обробку, натиснувши на «пуск». Після обробки переглянемо отриманий результат на діаграмі (рис. 8.12).

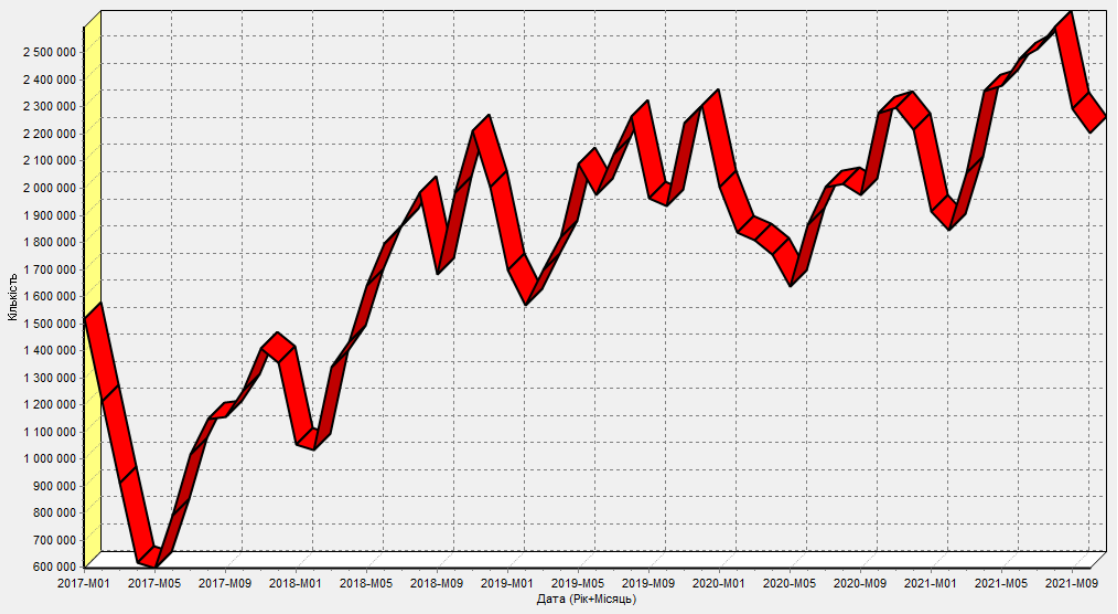


Рис. 8.12 – Діаграма результату

Видно, що дані трохи згладилися, аномалії та великі шуми зникли. Також видно тенденцію. Тепер перед аналітиком постає питання, а як, власне, прогнозувати часовий ряд. У даному випадку стовпець один. Будувати прогноз на майбутнє потрібно, виходячи з даних попередніх періодів, тобто передбачається, що кількість продажів наступного місяця залежить від кількості продажів за попередні місяці. Таким чином, вхідними факторами для моделі може бути продаж за поточний місяць, продаж за місяць раніше і так далі, а результатом повинен бути продаж за наступний місяць. Виходячи з цього, в даному випадку явно необхідно трансформувати дані до ковзного вікна.

8.4 Використання методу обробки даних «Ковзне вікно»

Запустимо майстер обробки (рис. 8.13), виберемо в якості оброблювача ковзне вікно і перейдемо до наступного кроку.

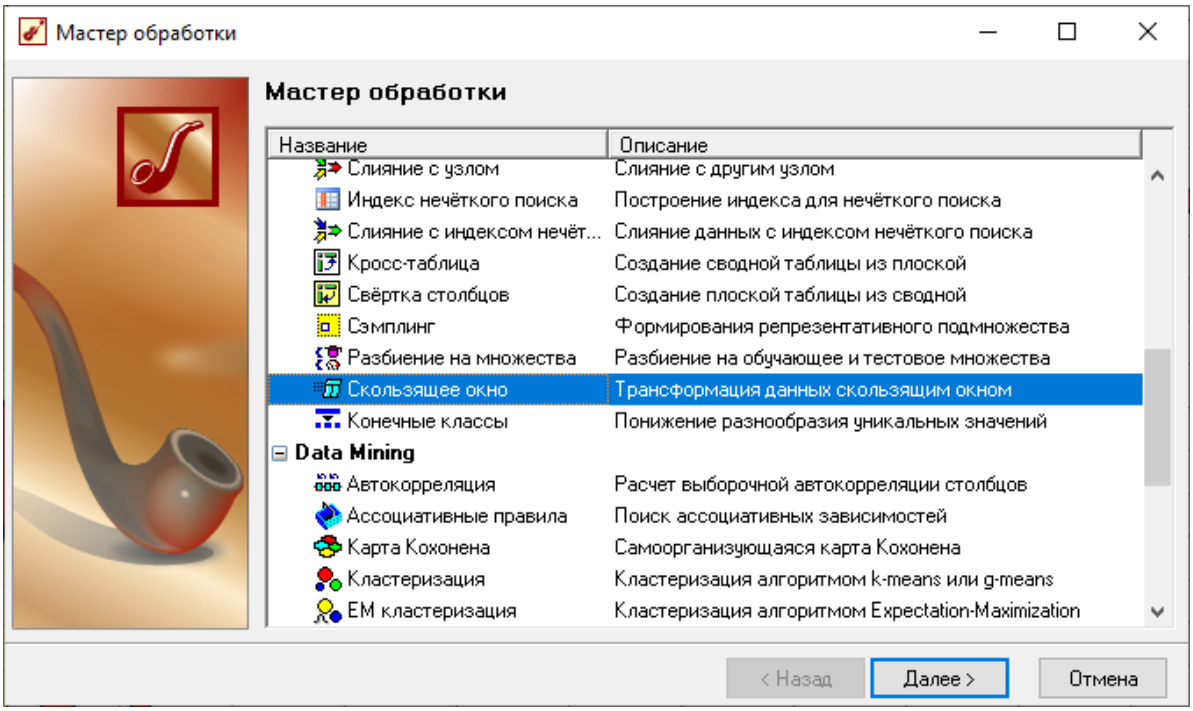


Рис. 8.13 – Вибір майстра обробки «Ковзне вікно»

Аналітик провів також авторегресійний аналіз та з'ясував наявність річної сезонності (див. приклад з авторегресією). У зв'язку з цим було вирішено будувати прогноз на тиждень вперед, ґрунтуючись на даних за 12, 11 місяців тому, два місяці тому та місяць тому. Виходячи з цього необхідно, призначивши поле «Кількість» використовуваним, вибрати глибину занурення 12. Тоді дані транспонуються до ковзного вікна так, що аналітику будуть доступні всі необхідні фактори для побудови прогнозу (рис. 8.14).

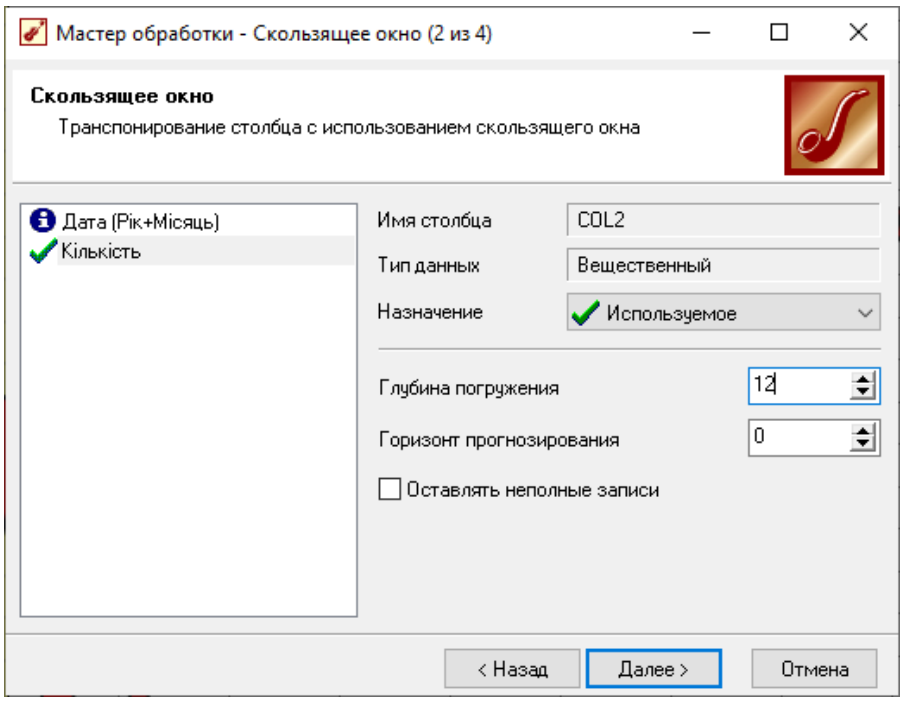
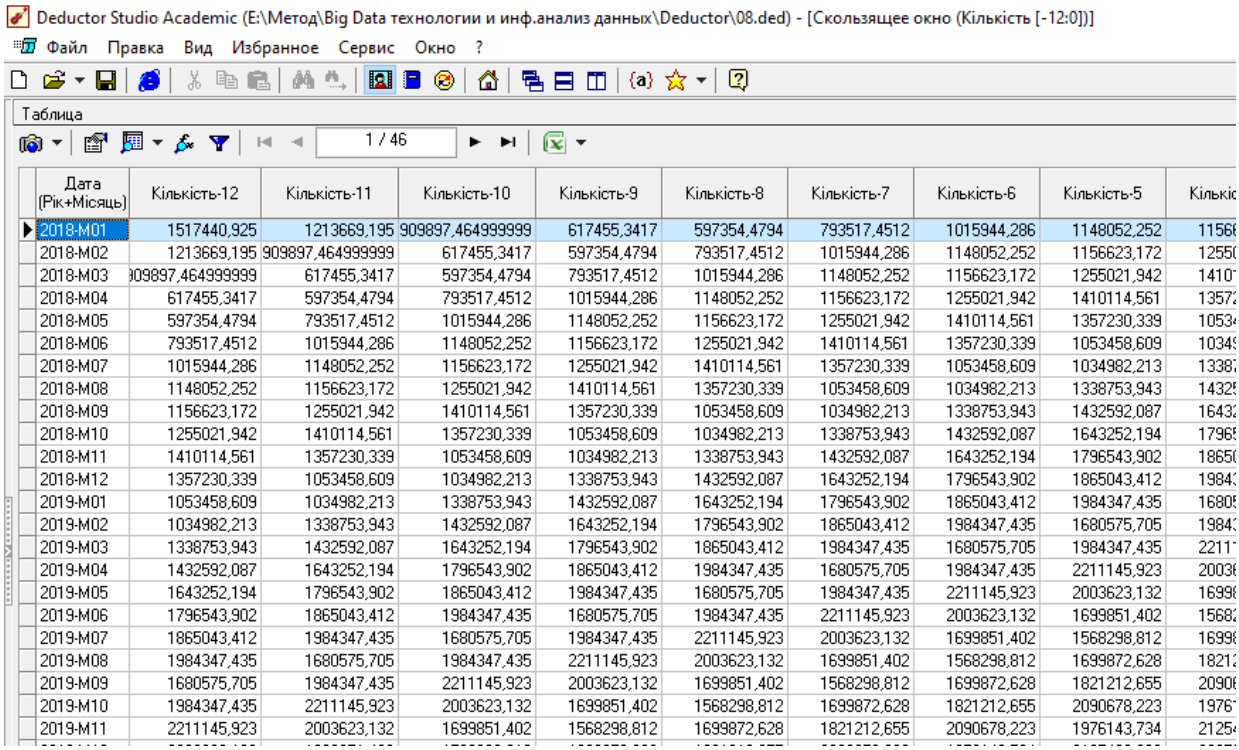


Рис. 8.14 – Транспонування стовбця з використанням ковзного вікна

Переглянути отримані дані можна у вигляді таблиці (рис. 8.15).



Дата (Рік+Місяць)	Кількість-12	Кількість-11	Кількість-10	Кількість-9	Кількість-8	Кількість-7	Кількість-6	Кількість-5	Кількіс
2018-M01	1517440,925	1213669,195	909897,464999999	617455,3417	597354,4794	793517,4512	1015944,286	1148052,252	1156
2018-M02	1213669,195	909897,464999999	617455,3417	597354,4794	793517,4512	1015944,286	1148052,252	1156623,172	1255
2018-M03	909897,464999999	617455,3417	597354,4794	793517,4512	1015944,286	1148052,252	1156623,172	1255021,942	1410
2018-M04	617455,3417	597354,4794	793517,4512	1015944,286	1148052,252	1156623,172	1255021,942	1410114,561	1357
2018-M05	597354,4794	793517,4512	1015944,286	1148052,252	1156623,172	1255021,942	1410114,561	1357230,339	1053
2018-M06	793517,4512	1015944,286	1148052,252	1156623,172	1255021,942	1410114,561	1357230,339	1053458,609	1034
2018-M07	1015944,286	1148052,252	1156623,172	1255021,942	1410114,561	1357230,339	1053458,609	1034982,213	1338
2018-M08	1148052,252	1156623,172	1255021,942	1410114,561	1357230,339	1053458,609	1034982,213	1338753,943	1432
2018-M09	1156623,172	1255021,942	1410114,561	1357230,339	1053458,609	1034982,213	1338753,943	1432592,087	1643
2018-M10	1255021,942	1410114,561	1357230,339	1053458,609	1034982,213	1338753,943	1432592,087	1643252,194	1796
2018-M11	1410114,561	1357230,339	1053458,609	1034982,213	1338753,943	1432592,087	1643252,194	1796543,902	1865
2018-M12	1357230,339	1053458,609	1034982,213	1338753,943	1432592,087	1643252,194	1796543,902	1865043,412	1984
2019-M01	1053458,609	1034982,213	1338753,943	1432592,087	1643252,194	1796543,902	1865043,412	1984347,435	1680
2019-M02	1034982,213	1338753,943	1432592,087	1643252,194	1796543,902	1865043,412	1984347,435	1680575,705	1984
2019-M03	1338753,943	1432592,087	1643252,194	1796543,902	1865043,412	1984347,435	1680575,705	1984347,435	2211
2019-M04	1432592,087	1643252,194	1796543,902	1865043,412	1984347,435	1680575,705	1984347,435	2211145,923	2003
2019-M05	1643252,194	1796543,902	1865043,412	1984347,435	1680575,705	1984347,435	2211145,923	2003623,132	1699
2019-M06	1796543,902	1865043,412	1984347,435	1680575,705	1984347,435	2211145,923	2003623,132	1699851,402	1568
2019-M07	1865043,412	1984347,435	1680575,705	1984347,435	2211145,923	2003623,132	1699851,402	1568298,812	1699
2019-M08	1984347,435	1680575,705	1984347,435	2211145,923	2003623,132	1699851,402	1568298,812	1699872,628	1821
2019-M09	1680575,705	1984347,435	2211145,923	2003623,132	1699851,402	1568298,812	1699872,628	1821212,655	2090
2019-M10	1984347,435	2211145,923	2003623,132	1699851,402	1568298,812	1699872,628	1821212,655	2090678,223	1976
2019-M11	2211145,923	2003623,132	1699851,402	1568298,812	1699872,628	1821212,655	2090678,223	1976143,734	2125

Рис. 8.15 – Таблиця отриманих даних

Як видно, тепер в якості вхідних факторів можна використовувати «Кількість-12», «Кількість-11» – дані за кількістю 12 і 11 місяців тому (відносно прогнозованого місяця) та інші необхідні фактори. Як результат прогнозу буде вказано стовпець «Кількість».

8.5 Побудова нейромережі в Deductor Studio

Перейдемо безпосередньо до самої побудови моделі прогнозу. Відкриємо майстер обробки і виберемо в ньому нейронну мережу (рис. 8.16).

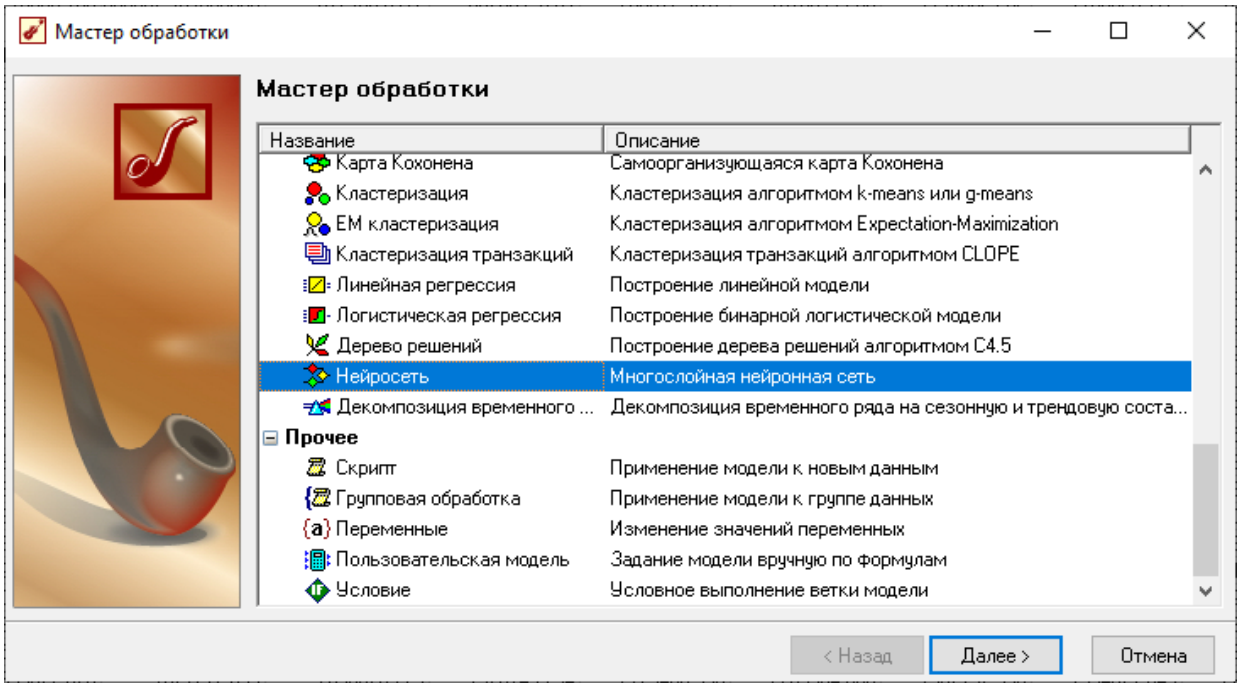


Рис. 8.16 – Вибір майстра обробки «Нейромережа»

На другому кроці майстра, згідно з прийнятим раніше рішенням, встановимо в якості вхідних поля «Кількість-12», «Кількість-11», «Кількість-2» і «Кількість-1», а в якості вихідних – «Кількість» (рис. 8.17). Інші поля зробимо інформаційними.

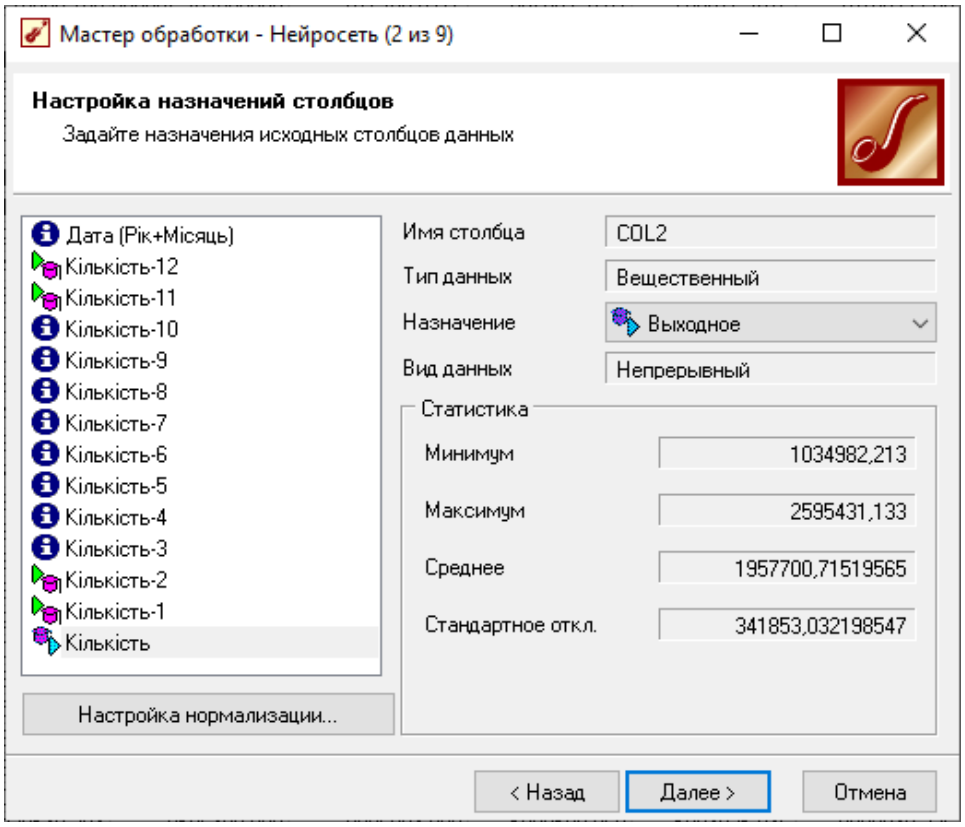


Рис. 8.17 – Налаштування призначень стовбців

Залишивши всі інші параметри побудови моделі за замовчанням, навчимо нейромережу (рис. 8.18 – 8.20).

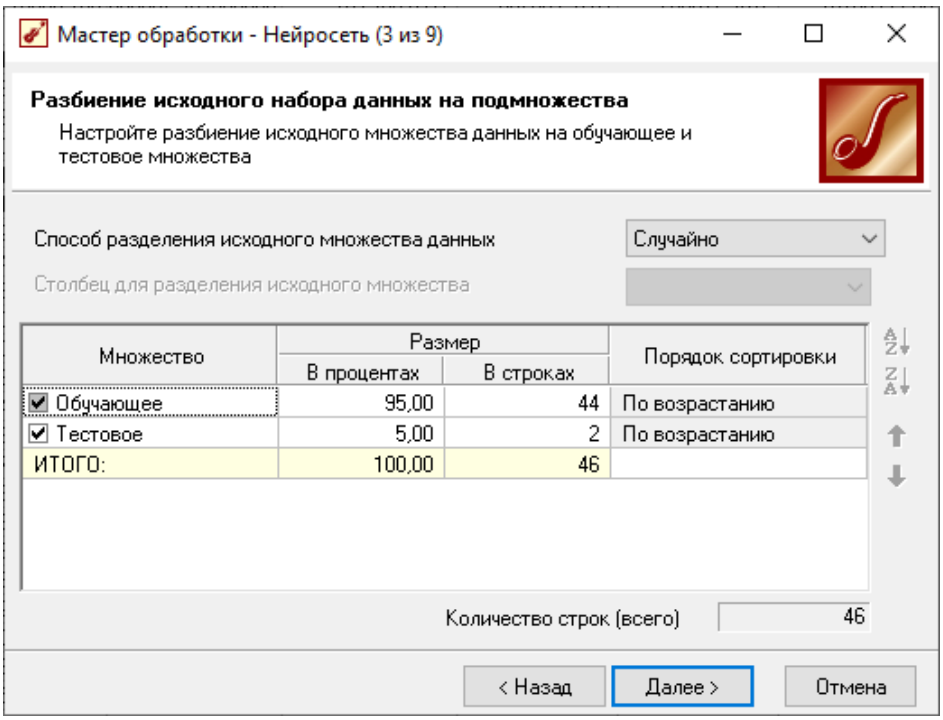


Рис. 8.18 – Розбиття початкового набору даних на підмножини

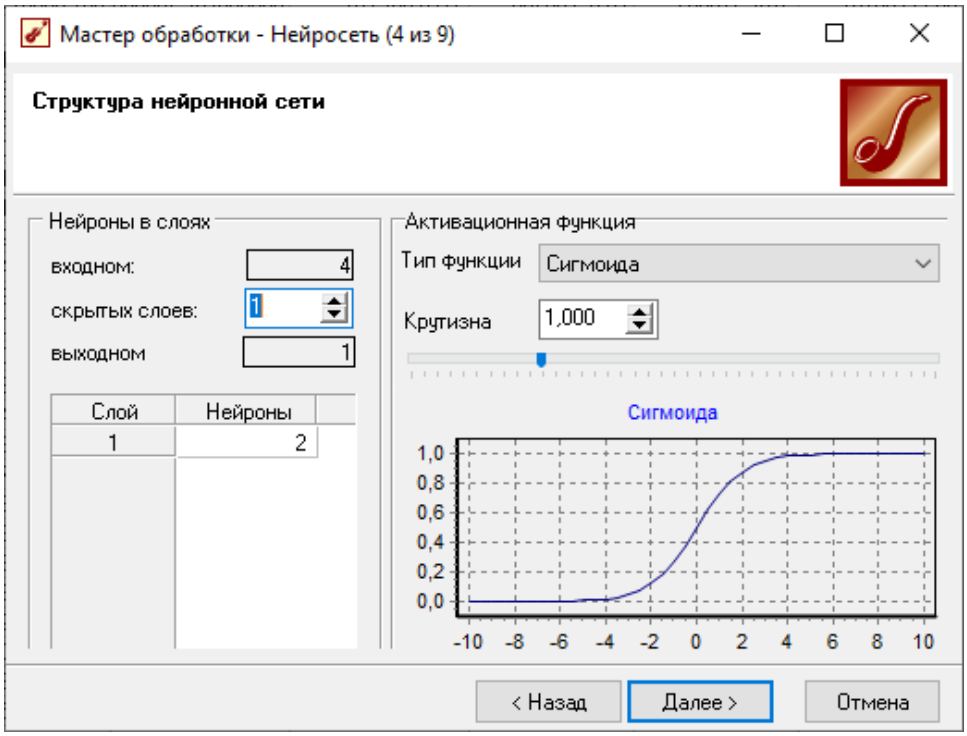


Рис. 8.19 – Побудова нейронної мережі

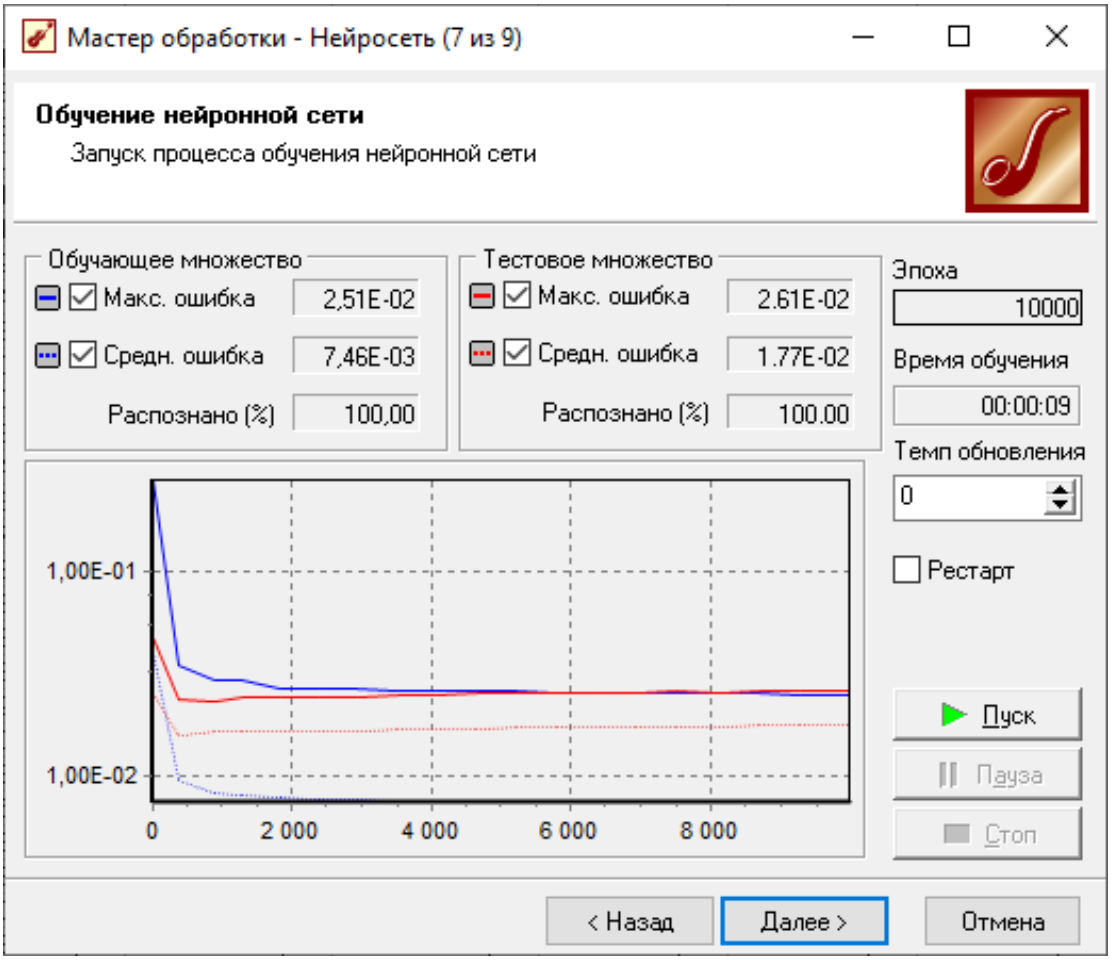


Рис. 8.20 – Навчання нейронної мережі

Після побудови моделі для перегляду якості навчання представимо отримані дані у вигляді діаграми та діаграми розсіювання (рис. 8.21).

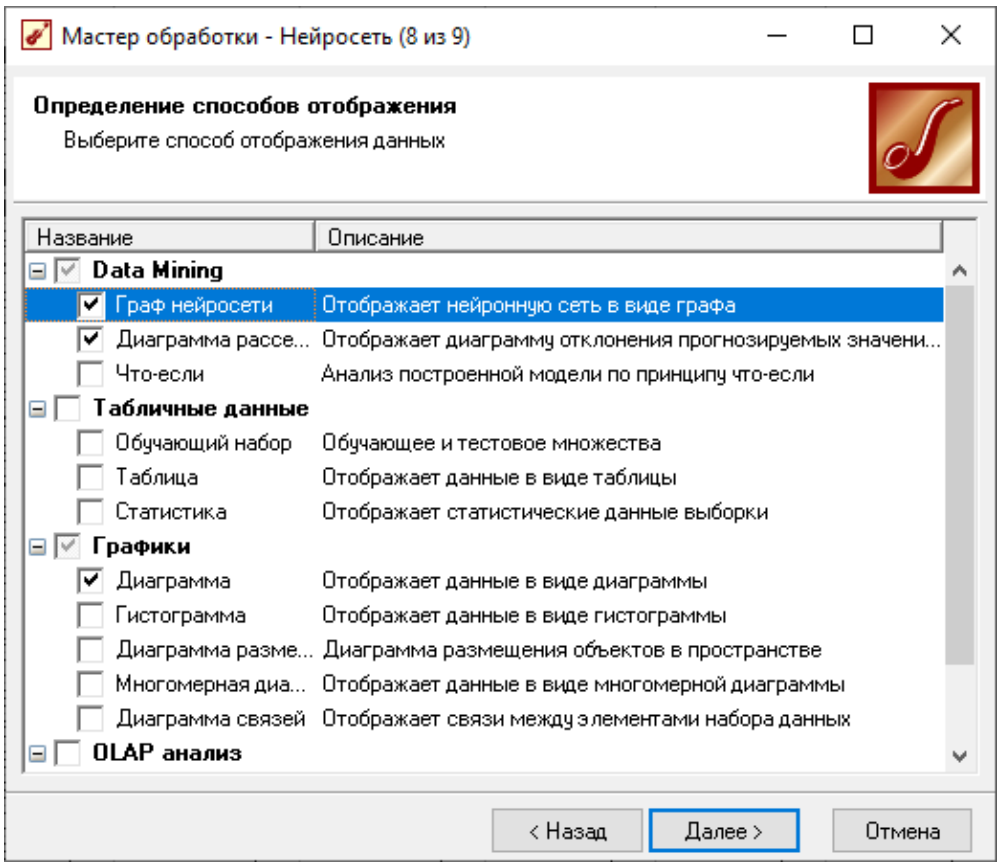


Рис. 8.21 – Визначення способів відображення даних

У майстрі налаштування діаграми (рис. 8.22) виберемо для відображення поля «Кількість» та «Кількість_OUT» – реальне та спрогнозоване значення.

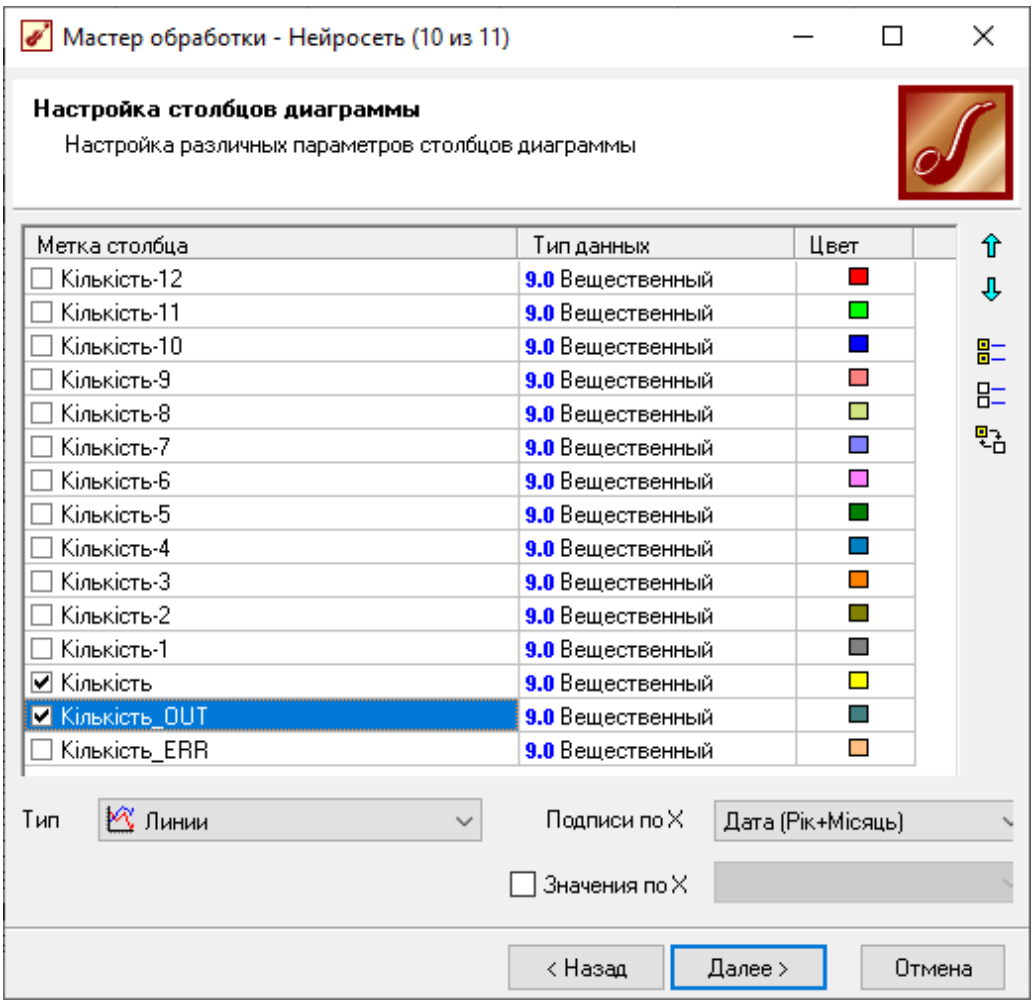


Рис. 8.22 – Вікно майстра налаштування стовбців діаграми

Після закінчення обробки «Нейромережа» отримуємо граф нейромережі, який відображує побудовану нейронну мережу (рис. 8.23).

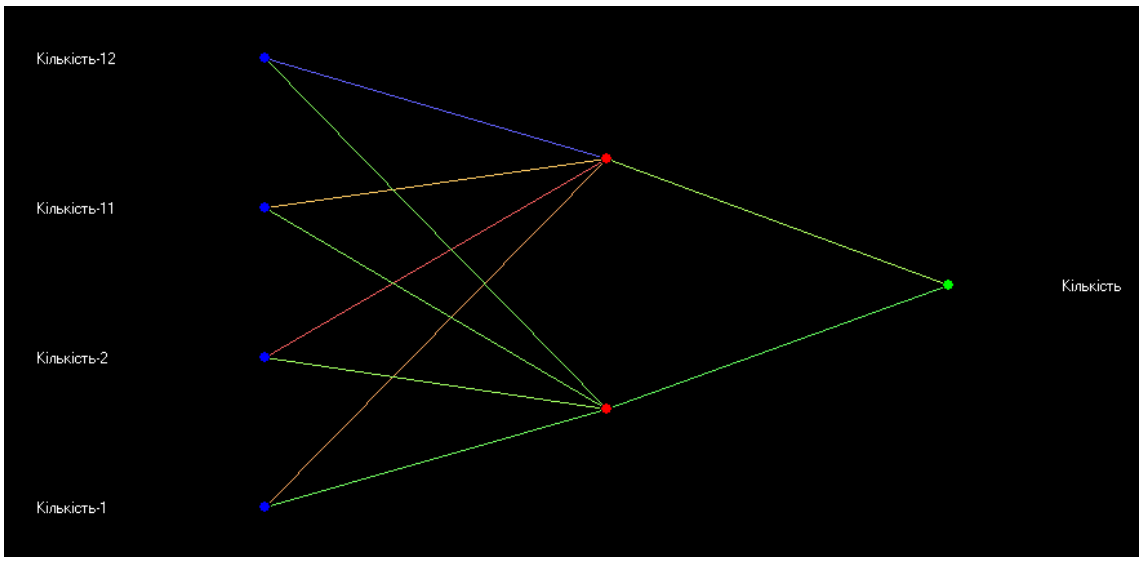


Рис. 8.23 – Граф нейромережі

Також результатом будуть два графіки, зображені на рис. 8.24

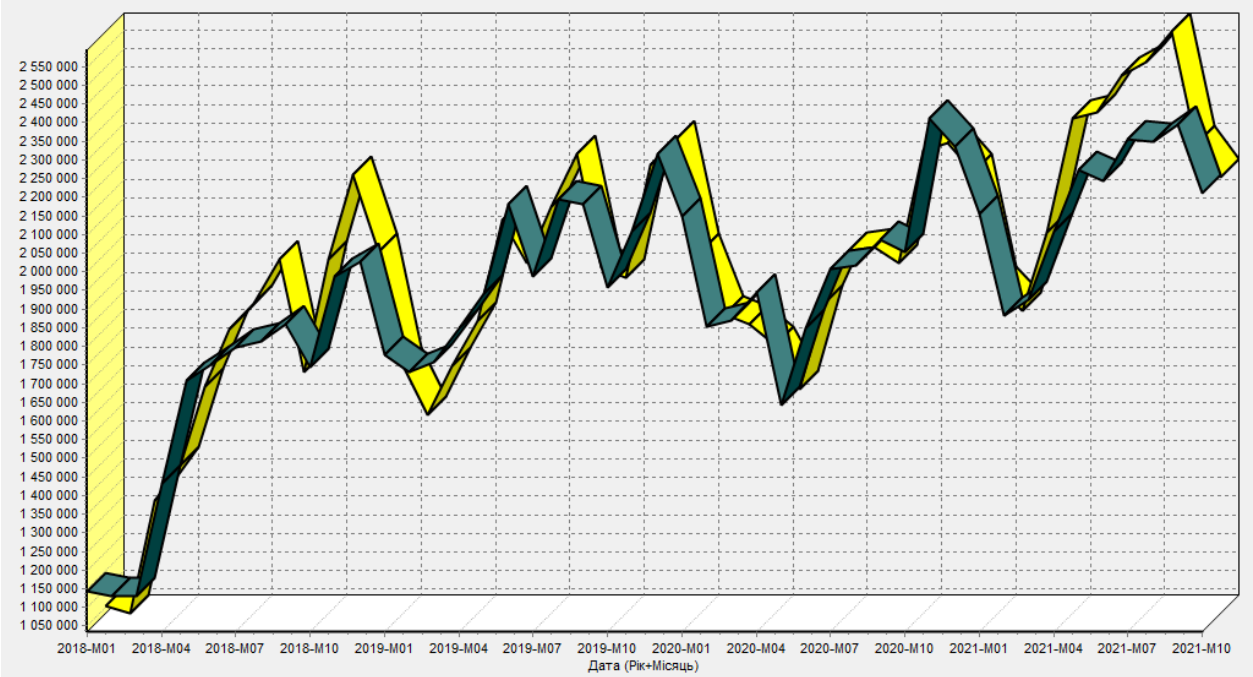


Рис.8.24 – Діаграма якості навчання

Діаграма розсіювання більш наглядно показує якість навчання (рис. 8.25).

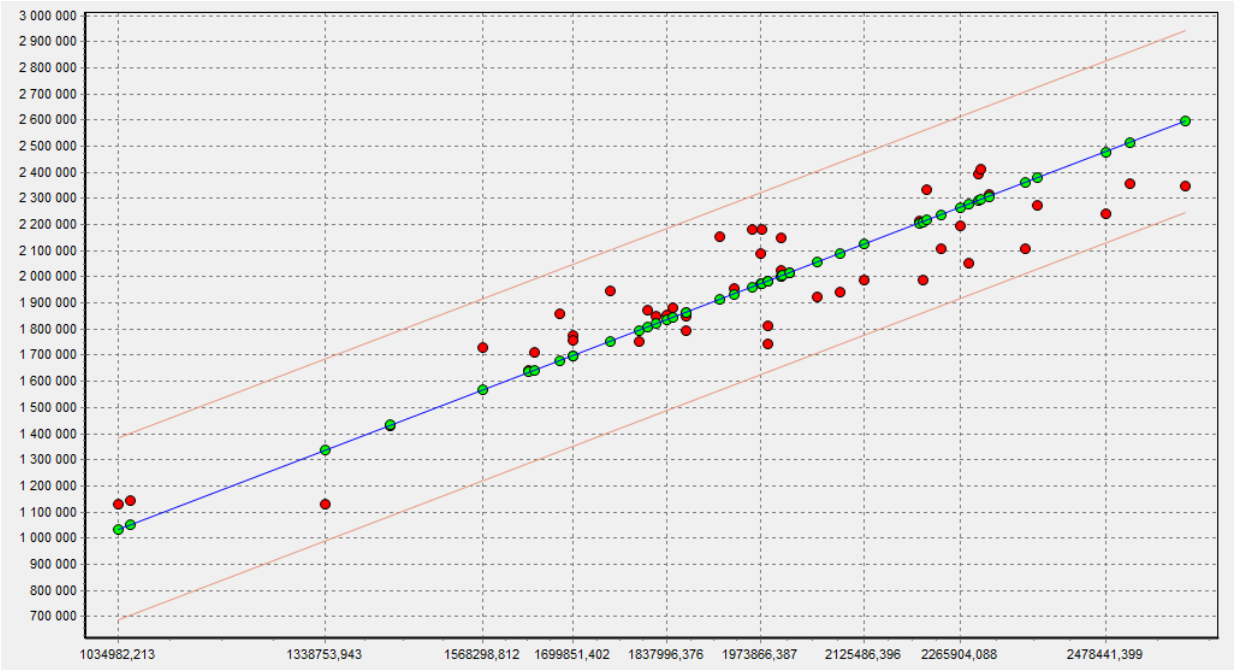


Рис. 8.25 – Діаграма розсіювання

Нейромережа навчена, тепер залишилося найголовніше – побудувати необхідний прогноз.

8.6 Побудова прогнозу в Deductor Studio

Для побудови прогнозу відкриваємо майстер обробки і вибираємо обробник «Прогнозування» (рис. 8.26).

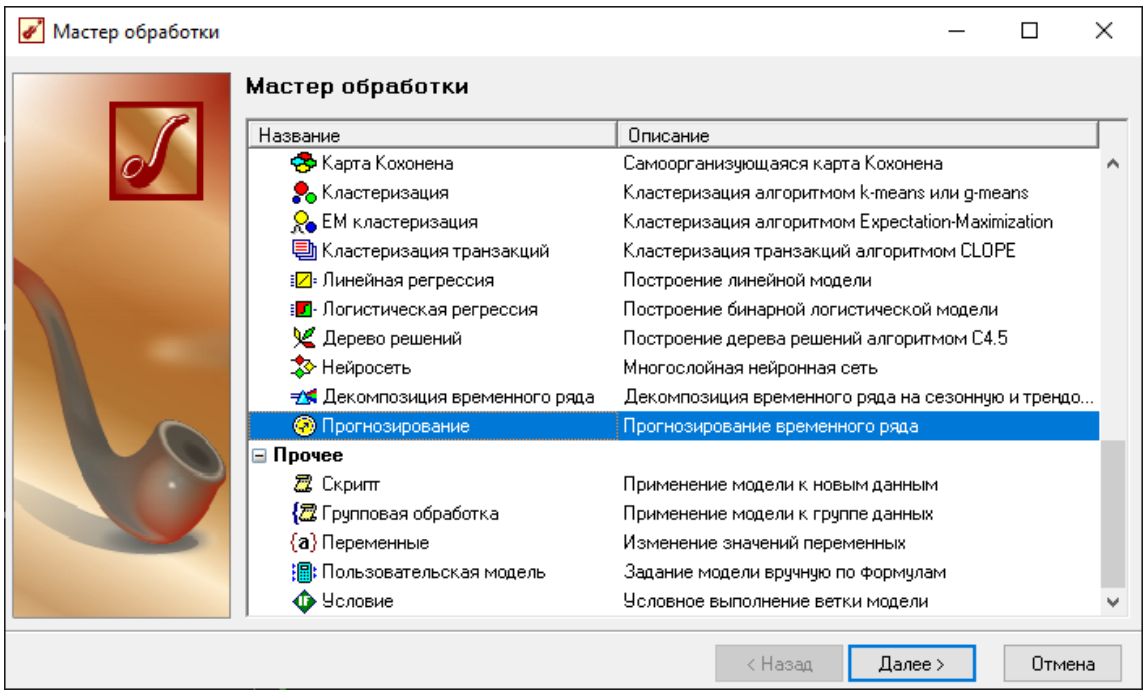


Рис. 8.26 – Вибір майстра обробки «Прогнозування»

На другому етапі майстра пропонується налаштувати зв'язки стовпців для прогнозування часового ряду – звідки брати дані для колонки при черговому етапі прогнозу. Майстер сам правильно налаштував всі переходи, тому залишається тільки вказати горизонт прогнозу (на скільки вперед прогнозуватимемо) рівним трьом, а також, для наочності, необхідно додати до прогнозу вихідні дані, встановивши в майстрі відповідний прапорець (рис. 8.27).

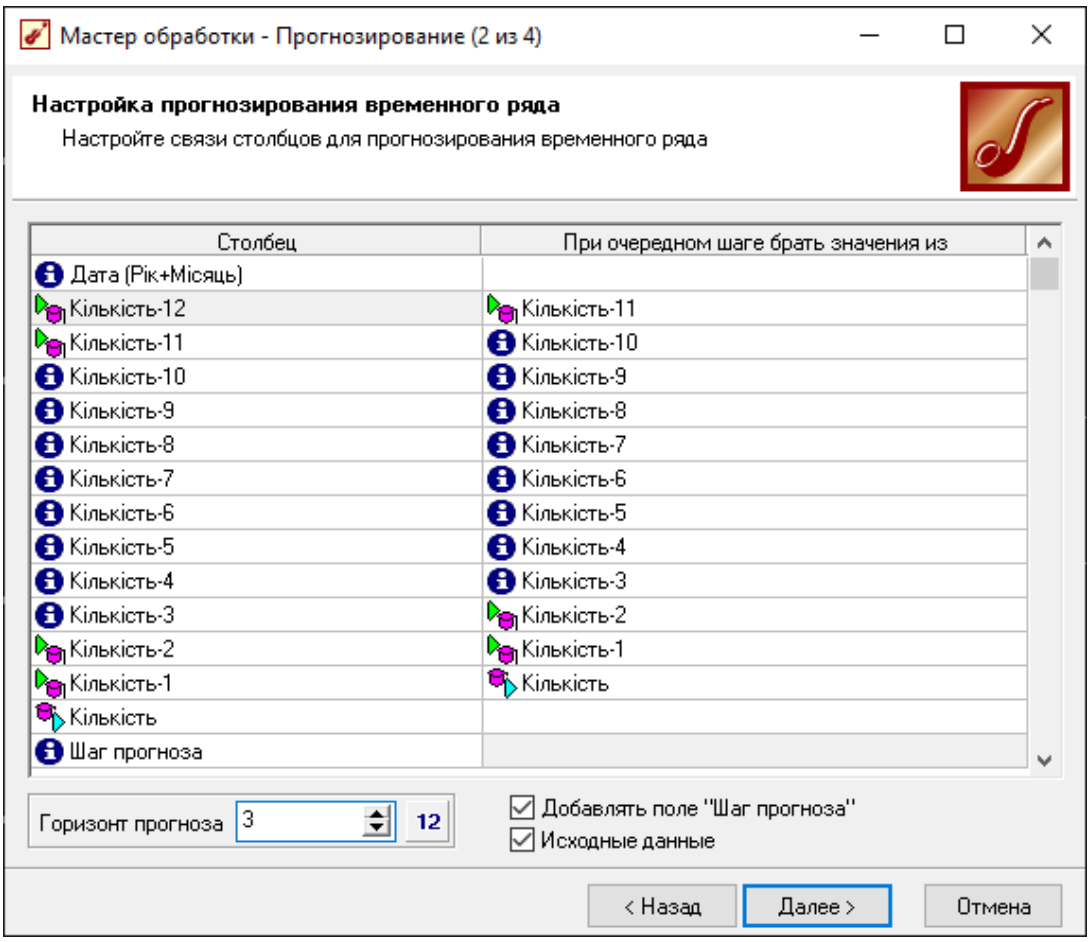


Рис. 8.27 – Налаштування прогнозування часового ряду

Після цього необхідно в якості візуалізатора вибрати діаграму прогнозу, яка з'являється тільки після прогнозування часового ряду (рис. 8.28).

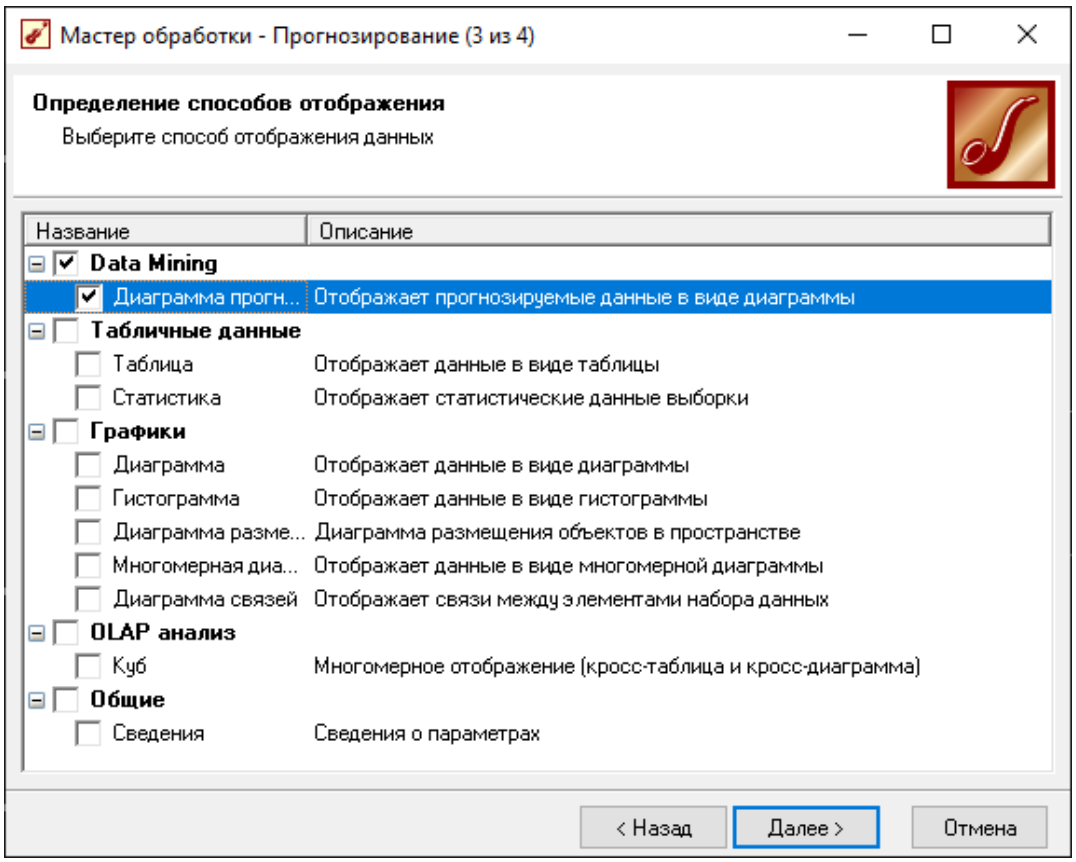


Рис. 8.28 – Визначення способів відображення

У майстрі налаштування стовпців діаграми прогнозу необхідно вказати в якості відображуваного стовпця «Кількість», а в якості підписів по осі X вказати стовпець «Крок прогнозу» (рис. 8.29).

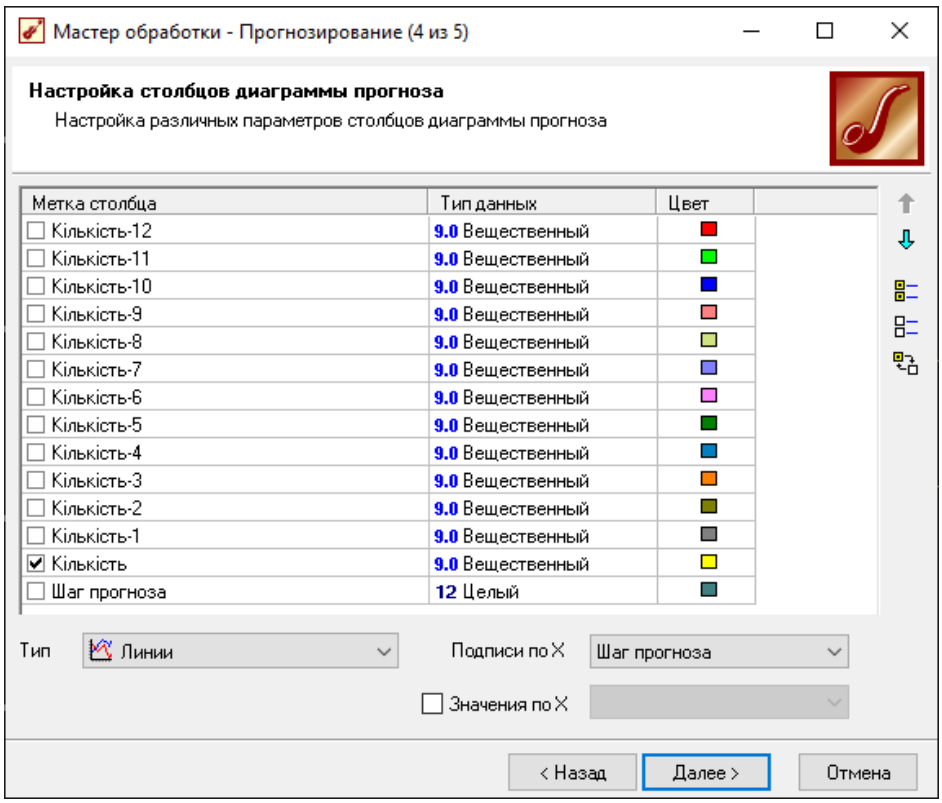


Рис. 8.29 – Налаштування стовбців діаграми прогнозу

Тепер аналітик може дати відповідь на запитання, яку кількість товарів буде продано наступного місяця і навіть через два місяці (рис. 8.30).

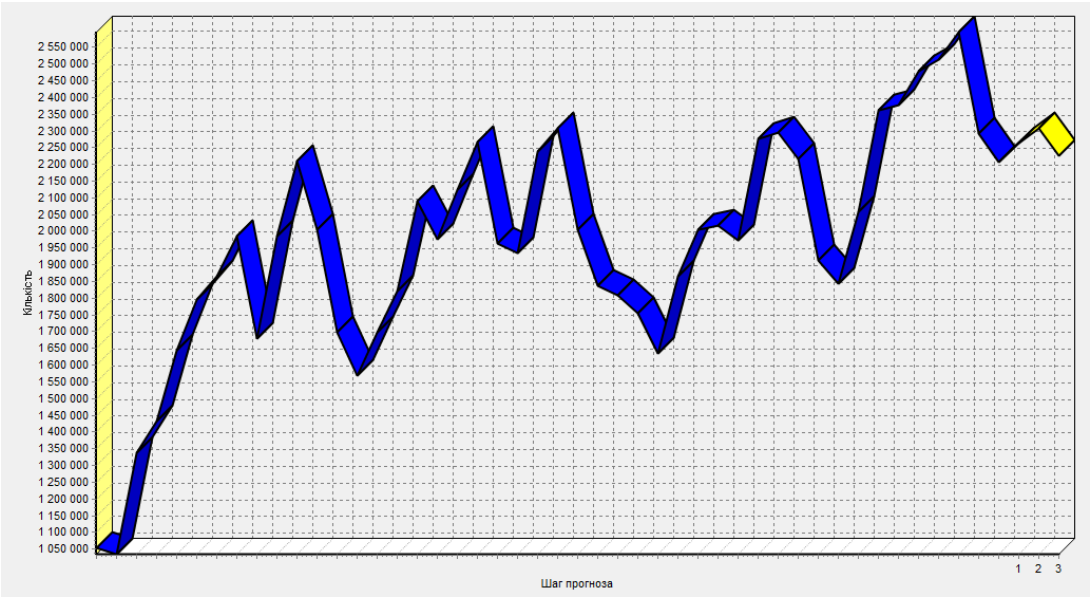


Рис. 8.30 – Діаграма прогнозу

Так, для виконання прогнозування були використані різні методи обробки: автокореляція, редагування викидів, ковзне вікно, нейромережа та прогнозування.

Питання до розділу 8

1. Що таке сезонність?
2. Для чого використовується автокореляційний аналіз?
3. Як визначити чи існує залежність між даними?
4. Для чого потрібен прогноз часового ряду?
5. Який інструмент в системі Deductor Studio використовується для прогнозування часових рядів?
6. Яке призначення обробника «Нейромережа» системи Deductor Studio?
7. Як обробник «Нейромережа» можна використовувати при прогнозуванні?

СПИСОК ЛІТЕРАТУРИ

1. Марченко О.О., Россада Т.В. Актуальні проблеми Data Mining: Навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. Київ. 2017. 150 с.
2. Казаков Ю.М., Тищенко А.А. Модели и методы анализа проектных решений: лабораторный практикум. Брянск: БГТУ, 2009. 82 с.
3. Чубукова І.О. Data Mining. URL: http://kist.ntu.edu.ua/textPhD/Chubukova-Data_Mining.pdf
4. Афанасьева Т.В., Афанасьев А.Н. Введение в проектирование систем интеллектуального анализа данных : учебное пособие. Ульяновск: УлГТУ, 2017. 64 с.
5. Замятин А.В. Интеллектуальный анализ данных: учеб. пособие. Томск : Издательский Дом Томского государственного университета, 2016. 120 с.
6. Афанасьева С.В. Технология интеллектуального анализа данных: учеб.пособие; Нац. исслед. ун-т «Высшая школа экономики», Санкт-Петербургский филиал. М.: Нац. исслед. ун-т «Высшая школа экономики», 2013. 152 с.
7. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. Спб.: БХВ-Петербург, 2004. 336 с.
8. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.

9. Дюк В., Самойленко А. Data mining: учебный курс. СПб: Питер, 2001. 368 с.
10. Ханк Д.Э., Уичерн Д.У., Райтс А.Дж. Бизнес-прогнозирование, 7-е издание.: Пер. с англ. М.: Издатель-ский дом «Вильямс», 2003. 656 с.

Навчальне видання

Димова Ганна Олегівна

Ларченко Оксана Валеріївна

Моделі і методи інтелектуального аналізу даних

Навчальний посібник

ISBN 978-617-7941-60-5

Підписано до друку 07.12.2021 р. Формат 60×90/16. Папір офсетний.
Друк: різографія. Гарнітура Times New Roman.
Ум. друк. арк. 10,5. Обл.-вид. арк. 11,29.
Наклад 300 прим. Замовлення № 3027.

Книжкове видавництво ФОП Вишемирський В. С.
Свідоцтво про внесення до Державного реєстру суб'єктів
видавничої справи: серія ХС № 48 від 14.04.2005 р.
видано Управлінням у справах преси та інформації
73000, Україна, м. Херсон, вул. Соборна, 2,
тел. (050) 133-10-13, e-mail: printvvs@gmail.com, vish_sveta@rambler.ru