8.7. Лабораторна робота №7 «Аналітичні рішення на основі трансформації даних і прогнозування за допомогою лінійної регресії».

Завдання до лабораторної роботи.

1. Відповідно до заданого варіанту підготувати необхідні дані у вигляді таблиць MS Excel і зберегти їх як персональні файли. Для підготовки даних використовувати тематичні сайти Інтернет, результати проходження практик, довідники і каталоги.

2. Здійснити трансформацію даних шляхом:

- розбиття даних на групи відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 1»;
- розбиття дати (по тижнях) відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 2»;
- квантування по інтервалах відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 3»;
- отримання необхідних розрахунків відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 4»;
- групування даних відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 5»;
- перетворення даних до ковзаючого вікна відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 6».

3. Отримати прогноз за допомогою лінійного регресійного аналіза відповідно до методики, представленої в розділі «Порядок виконання роботи. Частина 7».

4. Виконати економічний аналіз різних ситуацій, застосовуючи основні візуалізатора, відповідно до методики, представленої в розділі «Порядок виконання роботи».

Варіанти завдань:

N варіанту	Галузь
1	Продажі продовольчих товарів (шаблон покупок)
2	Продажі промислових товарів (шаблон покупок)
3	Продажі будівельних товарів (шаблон покупок)
4	Продажі комп'ютерної техніки (шаблон покупок)
5	Зручне розміщення товарів на прилавках супермаркетів
6	Зручне розміщення товарів на прилавках промислових магазинів
7	Зручне розміщення товарів на вітринах книжкових електронних магазинів
8	Зручне розміщення товарів на вітринах електронних магазинів загального призначення
9	Зручне розміщення товарів на вітринах комп'ютерних електронних магазинів
10	Стимулювання продажів промислових товарів
11	Стимулювання продажів продукції ЗМІ
12	Стимулювання продажів продовольчих товарів
13	Стимулювання продажів комп'ютерної техніки

Порядок виконання роботи.

ЧАСТИНА 1. 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 проглянути, наприклад інформацію не по всій сукупності даних, а по певних групах (наприклад, яку суму кредиту беруть на ті або інші цілі, або кредитори того або іншого віку). У другому випадку (прогнозування) аналітикові необхідно враховувати той факт, що певні групи (в даному випадку групи кредиторів) поводяться по різному, і що модель прогнозу, побудована на всіх даних не враховуватиме нюансів, що виникають в цих групах. Тобто краще побудувати декілька моделей прогнозу, наприклад,

залежно від сумової групи кредиту, і розробляти прогноз на них, ніж побудувати одну модель прогнозу. Виходячи з цього, в Deductor Studio надається широкий набір інструментів, які дозволяють тим або іншим способом розбивати вихідні дані на групи, групувати будь-яким способом всілякі показники і т. п.

1. Розглянемо розбиття даних на групи на прикладі даних по ризиках кредитування фізичних осіб, підготовлених в MS Excel. Стовпці, що цікавлять нас: «СУМА КРЕДИТУ», «ДАТА КРЕДИТУВАННЯ», «МЕТА КРЕДИТУВАННЯ» і «ВІК» (рис. 8.31).

Сумма кредита	Стоимость кре	Срок кредита	Дата кредитов	Цель кредитования
7000	1400	6	01.01.03	Иное
7500	1500	6	01.01.03	Иное
14500	2900	12	01.01.03	Покупка товара
15000	3000	6	01.01.03	Покупка товара
32000	6400	12	01.01.03	Иное
11500	2300	6	01.01.03	Турпоездки, развлечения и т.п.
5000	1000	6	01.01.03	Покупка и ремонт недвижимости
61500	12300	30	01.01.03	Покупка товара
13500	2700	12	01.01.03	Оплата услуг (мед., юрид. и т.п.)
25000	5000	18	01.01.03	Покупка товара
25500	5100	24	01.01.03	Покупка товара
9500	1900	6	01.01.03	Покупка товара
53000	10600	24	01.01.03	Иное
27500	5500	18	02.01.03	Покупка товара
4000	800	6	02.01.03	Оплата услуг (мед., юрид. и т.п.)

Рис. 8.31. Вхідні дані проекту.

2. Після імпорту даних з табличного файлу найбільш інформативно проглянути дані можна за допомогою візуалізації «*Куб*», вибравши в якості вимірів стовпці «ВІК» і «МЕТА КРЕДИТУВАННЯ», а в якості факту – стовпець «СУМА КРЕДИТУ». Останні стовпці встановити як непридатні (рис. 8.32).

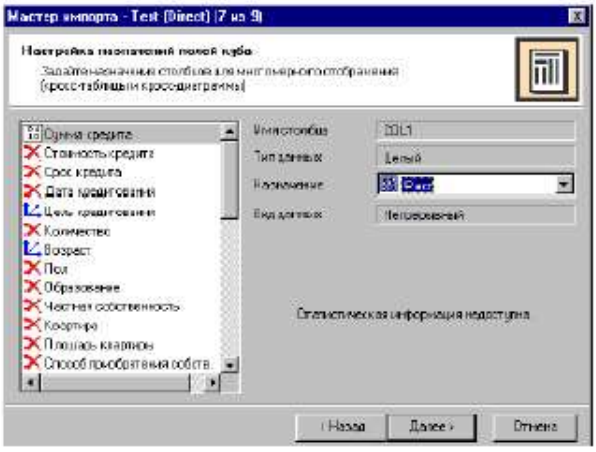

Рис. 8.32. Налаштування візуалізатора «Куб».

3. На наступному кроці налаштування куба слід вказати вимір «МЕТА КРЕДИТУВАННЯ» як вимір в термінах, а вимір «ВІК» як вимір в стовпцях, перетягнувши їх за допомогою миші у відповідні вікна з області доступних вимірів (рис. 8.33).

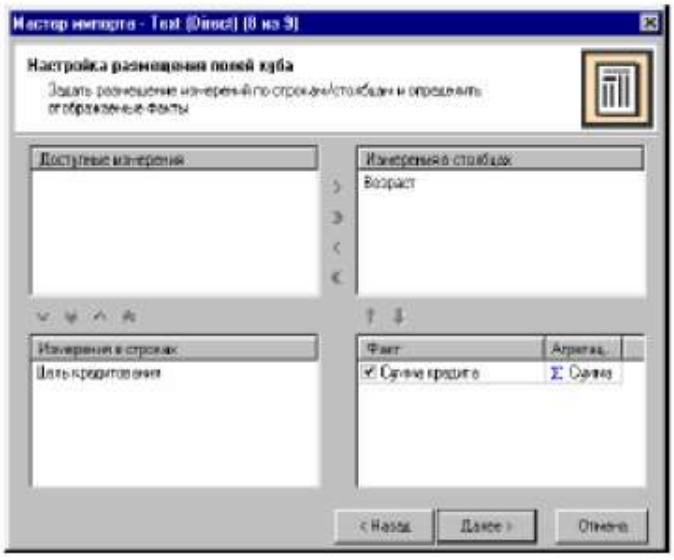

Рис. 8.33. Налаштування полів візуалізатора «Куб».

4. В результаті, на крос-діаграмі (одна із закладок візуалізації «Куб») можна проглянути вихідні дані (рис. 8.34).

	Возраст		
Цель кредитования	19	20	21
Иное	50 000,00	17 000,00	8 500,00
Оплата за образование		17 500,00	29 500,00
Оплата услуг (мед., юрид. и т.п.)			
Покупка и ремонт недвижимости	78 000,00		13 000,00
Покупка товара	46 500,00	73 500,00	76 500,00
Турпоездки, развлечения и т.п.		30 500,00	
Итого	174 500,00	138 500,00	127 500,00

Рис. 8.34. Результати розбиття даних на групи.

ЧАСТИНА 2. Розбиття дати (по тижнях).

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

1. Хай нам необхідно отримати дані по сумах взятих кредитів по тижнях (у раніше створеному файлі міститься інформація за перші два тижні 2010 року). Для цього в майстрові обробки «*Дата и Время*» на другому кроці виберемо як використовуване поле «ДАТА КРЕДИТУВАННЯ», а в таблиці налаштувань, що з'явилася після цього, виберемо значення «*Используемое*» в стовпці «*Строка*» напроти рядка «*Год + Неделя*» (рис. 8.35).

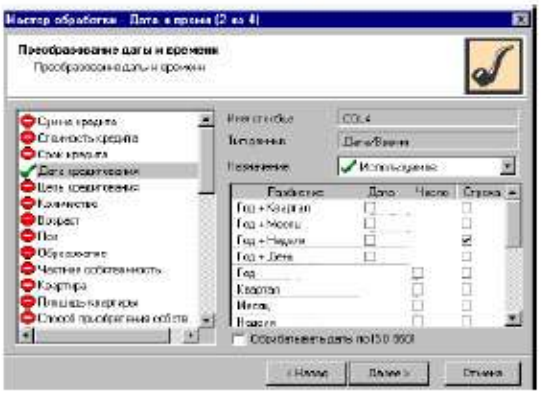

Рис. 8.35. Майстер обробки «Дата и Время».

2. Більше жодні налаштування не знадобляться, тому перейдемо далі до вибору типа візуалізації. Виберемо в якості візуалізації «*Таблиця*» і «*Куб*», поставивши галочок у відповідних позиціях. У майстрові налаштування полів куба виберемо в якості виміру стовпець, що з'явився після обробки, «ДАТА КРЕДИТУВАННЯ_YWStr (Рік + Тиждень)» і стовпець «МЕТА КРЕДИТУВАННЯ», а в якості факту – «СУМА КРЕДИТУ». Останні поля зробимо невживаними.

3. На наступному кроці перенесемо один вимір з області «доступних» в область «*Измерения в строках*», а інше – в область «*Измерения в столбцах*». Таким чином, на крос-діаграмі маємо суми взятих кредитів по тижнях (за перші два тижні року) в розрізі цілей кредитування (рис. 8.36).

	Дата кредитования (Год + Неделя)		
Цель кредитования	2003-W01	2003-W02	Итого
Иное	358 000,00	137 000,0(	495 000,00
Оплата за образование	62 000,00	312 000,0(	374 000,00
Оплата услуг (мед., юрид. и т.п.)	110 500,00	191 000,0(	301 500,00
Покупка и ремонт недвижимости	404 000,00	538 000,0(	942 000,00
Покупка товара	642 000,00	643 500,0(	1 285 500,00
Турпоездки, развлечения и т.п.	35 500,00	113 500,0(	149 000,00
Итого	612 000,00	1 935 000,0(	3 547 000,00

Рис. 8.36. Крос-діаграма результатів.

У таблиці з даними видно, що нове поле - «ДАТА КРЕДИТУВАННЯ_YWStr (Рік + Тиждень)» містить однакові значення (дата початку тижня) для рядків, які потрапляють в один і той же тиждень (дата початку тижня або номер тижня з початку року) (рис. 8.37).

Срок кредита	Дата кредитов	Дата кред	Цель кредитования
24	05.01.03	2003-W01	Покупка товара
12	05.01.03	2003-W01	Покупка товара
30	05.01.03	2003-W01	Иное
36	06.01.03	2003-W02	Покупка и ремонт недвижимости
12	06.01.03	2003-W02	Оплата за образование
18	06.01.03	2003-W02	Иное
6	06.01.03	2003-W02	Покупка товара

Рис. 8.37. Аналіз результатів розрахунків.

ЧАСТИНА 3. Квантування віку кредиторів на 5 інтервалів.

Часто аналітикові необхідно віднести безперервні дані (наприклад, кількість продажів) до якого-небудь кінцевого набору (наприклад, всю сукупність даних про кількість продажів необхідно розбити на 5 інтервалів – від 0 до 100, від 100 до 200 і так далі, і віднести кожен запис вхідного набору до деякого конкретного інтервалу) для аналізу або фільтрації виходячи саме з цих інтервалів. Для цього в Deductor Studio застосовується інструмент *квантування (або дискретизація).*

Квантування призначене для перетворення безперервних даних в дискретні. Перетворення може проходи 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 знаходилася однакова кількість даних). В якості значення результуючого набору даних можуть виступати номер інтервалу, нижній або верхній кордон інтервалу, середина інтервалу, або мітка інтервалу (значення визначені аналітиком).

Прикладом використання даного інструменту може служити 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). Вхідні дані розподіляться по п'яти інтервалах саме так, оскільки, згідно статистиці, мінімальне значення віку кредитора 19, а максимальне 69 років. Це необхідно аналітикові для оцінки кредиторської активності різних вікових груп, з метою ухвалення рішення про стимулювання кредиторів в групах з низькою активністю (наприклад, зменшення вартості кредиту для цих груп) і, мабуть, збільшення прибутку у вікових групах кредиторів з високим ризиком (шляхом збільшення для них вартості кредиту). Причому аналітик бажає бачити дані в розрізі по тижнях (тому продовжимо роботу на останніх отриманих даних попереднього прикладу).

1. Скористаємося майстром квантування (рис. 8.38).

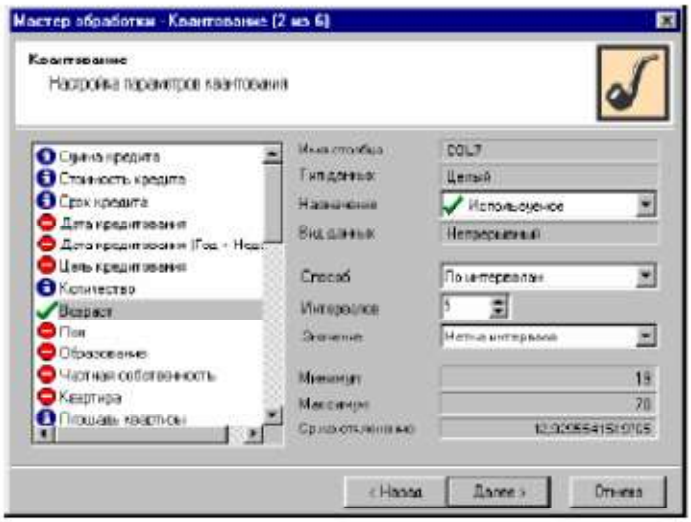

Рис. 8.38. Візуалізатор «Майстер квантування».

В ньому виберемо в якості використовуваного - значення поля «Вік» , вкажемо спосіб розбиття «*По интервалам*», задамо кількість інтервалів рівним 5, в якості значення виберемо «*Метку интервала*».

2. На наступному кроці майстра визначимо мітки віку кредиторів: «до 30 років», «від 30 до 40 років» і так далі (рис. 8.39).

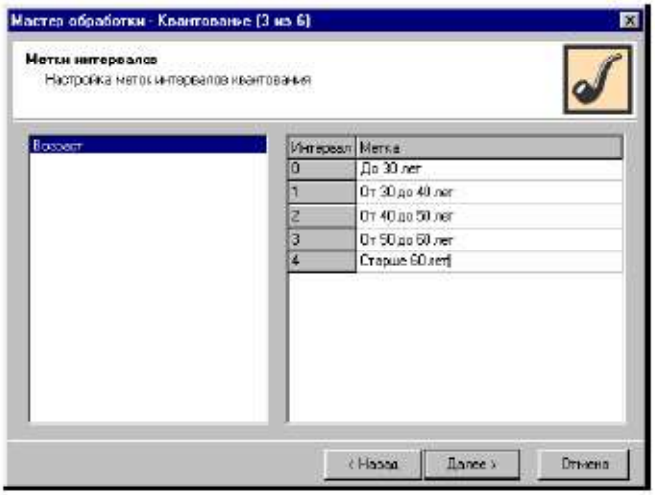

Рис. 8.39. Налаштування міток в «Майстері квантування».

3. Після обробки виберемо в якості способу відображення «*Куб*». У майстрові вкажемо «СУМА КРЕДИТУ» в якості факту, «ВІК» і поле «ДАТА КРЕДИТУВАННЯ (Рік + Тиждень)» в якості виміру, останні поля вкажемо

невживаними. Далі перенесемо «ВІК» з доступних вимірів в «Виміри в рядках», а «ДАТА КРЕДИТУВАННЯ (Рік + Тиждень)» у «Виміри в стовпцях». На крос-діаграмі тепер видно інформацію про те, які суми кредитів беруть кредитори певних вікових груп в розрізі по тижнях (рис. 8.40).

	Дата кредитования (Год + Неделя)		
Возраст	2003-W01	2003-W02	Итого
До 30 лет	798 500,00	808 500,00	1 607 000,00
От 30 до 40 лет	298 000,00	561 000,00	859 000,00
От 40 до 50 лет	195 000,00	295 500,00	490 500,00
От 50 до 60 лет	111 000,00	209 500,00	320 500,00
Старше 60 лет	209 500,00	60 500,00	270 000,00
Итого	1 612 000,00	1 935 000,00	3 547 000,00

Рис. 8.40. Результати квантування.

Тепер аналітик, отримавши такі дані, може дати рекомендації про зниження вартості кредиту для осіб, старше 50 років, або про вживання інших заходів, здатних залучити більшу кількість кредиторів цих груп, або заходів, направлених на те, аби кредитори брали кредит на великі суми.

ЧАСТИНА 4. Обчислювані дані.

Інколи виникає необхідність на деякому етапі обробки даних отримати нові (похідні) дані. Можливо, аналітикові потрібно обчислити процентне відхилення значення одного поля відносно іншого, або підрахувати суму, різницю полів, отримати на основі даних показник і вже його використовувати для подальшої обробки, залежно від значення полів обчислити ті або інші вирази. У Deductor Studio таку можливість надає інструмент «*Вычисляемые данные*». Він дозволяє створювати нові поля, що обчислюють задані аналітиком вирази. Тобто обчислювані дані служать для здобуття похідних даних на основі наявних у вхідному наборі. Майстер надає широкий набір функцій різного напряму. У майстрові представлений список нових виразів, де надаються необхідні аналітикові вираження, список доступних функцій з

коротким описом кожної, список доступних операцій і також список доступних стовпців, які можна задіювати при створенні вираження.

1. Розглянемо застосування цього методу на прикладі даних з файлу, підготовленого засобами MS Excel. У ньому міститься таблиця з полями «АРГУМЕНТ1», «АРГУМЕНТ2», «АРГУМЕНТ3» – набір аргументів. По-перше необхідно імпортувати даний файл в програму. Для перегляду вхідних даних в даному випадку зручніше використовувати візуалізатор «*Таблица*» (рис. 8.41).

	Аргумент1	Аргумент2	Аргумент3
	0	4	4
	0	5	5
	0	6	6
	0	7	7
	0	8	8
	0	9	9

Рис. 8.41. Перегляд даних за допомогою візуалізатора «Таблица».

Допустимо необхідно на основі аргументів розрахувати деякі математичні функції. Хай це будуть дві функції одного аргументу (АРГУМЕНТ3), одна функція від двох аргументів, одна кусково-задана функція і функція, що показує відносне відхилення (АРГУМЕНТ1 + 1 від АРГУМЕНТ2 + 1). Передбачається, що всі ці функції використовуватимуться для подальшої обробки.

2. Функції *F1(АРГУМЕНТ3), F2(АРГУМЕНТ3)*. Розрахуємо значення функцій SIN(АРГУМЕНТ3 * АРГУМЕНТ3) * LN(АРГУМЕНТ3 + 1) * EXP(- АРГУМЕНТ3/10) та 10*SIN(АРГУМЕНТ3 * АРГУМЕНТ3/100) / (АРГУМЕНТ3 + 1) * EXP(- АРГУМЕНТ3/10). Для цього, знаходячись на вузлі імпорту, запустимо майстер обробки. Виберемо в якості обробки - «*Вычисляемые данные*». На другому кроці майстра в списку виразів в першому рядку в графі «*Название выражения*» замість напису «*Выражение*» напишемо

F1(АРГУМЕНТ3). У полі редактора виразів (у верхній частині майстра) напишемо «SIN(COL3*COL3)*LN(COL3+1)*EXP(-COL3/10)» (рис. 8.42).

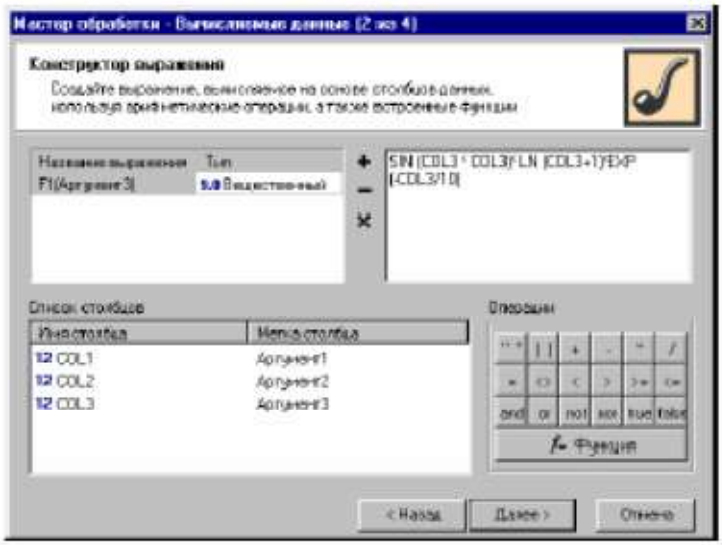

Рис. 8.42. Меню конструктора виразів.

Таким чином, ми створили новий стовпець, задали йому назву «F1(АРГУМЕНТ3)» і також визначили, які значення прийматимуть записи цього поля. На цьому створення обчислюваного значення закінчене, тому переходимо на наступний крок майстра, де пропонується вибрати спосіб відображення даних. Самим інформативним в даному випадку є діаграма, яку і слід вибрати. Якщо вибрати в майстрові налаштувань діаграми в якості відображаємого поля - «F1(АРГУМЕНТ3)» і в якості типа графіка - «Лінії», то можна побачити графік обчисленої функції (рис. 8.43).

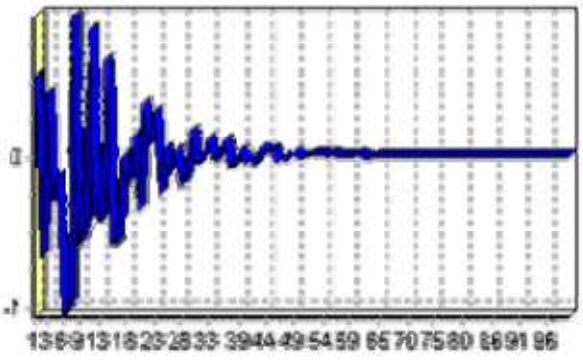

Рис. 8.43. Результати обчислювання даних.

Складна функція F2(АРГУМЕНТ3) відрізняється лише виглядом функції («10*SIN(COL3*COL3/100)/(COL3+1)*EXP(-COL3/10)») (рис. 8.44).

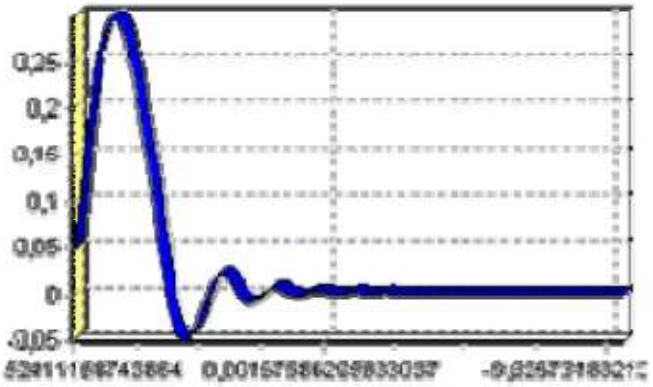

Рис. 8.44. Результати розрахунку функції F2(АРГУМЕНТ3).

3. Функція від двох аргументів *F3(АРГУМЕНТ1; АРГУМЕНТ2)*. Така функція цікава тим, що для її перегляду в трьох вимірах можна використовувати візуалізацію «*Куб*». Задамо назву виразу «F3(АРГУМЕНТ1; АРГУМЕНТ2)», у полі обчислюваного виразу напишемо «COL1*COL1/100 – COL2*COL2/100». Виберемо візуалізацію «*Куб*» і налаштуємо його так, що «АРГУМЕНТ1» і «АРГУМЕНТ2» були б вимірами, «F3(АРГУМЕНТ1; АРГУМЕНТ2)» фактом, а «АРГУМЕНТ3» невживаним. Вибравши «АРГУМЕНТ1» виміром в стовпцях, а «АРГУМЕНТ2» – вимірам в рядках перейдемо до перегляду крос-діаграми. Для наочнішого перегляду встановимо тип діаграми «*Області*». Тепер можна проглянути обчислену функцію в об'ємному вигляді (рис. 8.45).

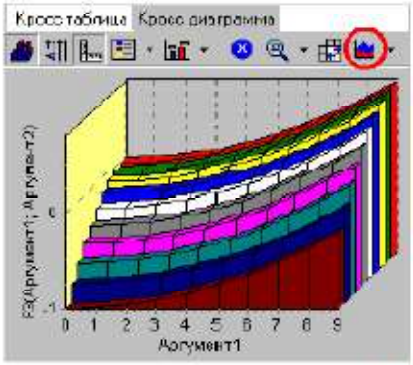

Рис. 8.45. Результати обчислення функцій.

4. Обчислення відхилення «*АРГУМЕНТ1 + 1 від АРГУМЕНТ2 + 1*». Покажемо приклад вживання однієї із вбудованих функцій – обчислення пайового відхилення одного аргументу від іншого (RELDEV). Список всіх вбудованих функцій разом з описом можна поглянути в майстрові в лівому нижньому кутку. Задавши в якості обчислюваного виразу RELDEV(COL1 + 1; COL2 + 1) можна на діаграмі побачити дане відхилення (рис. 8.46).

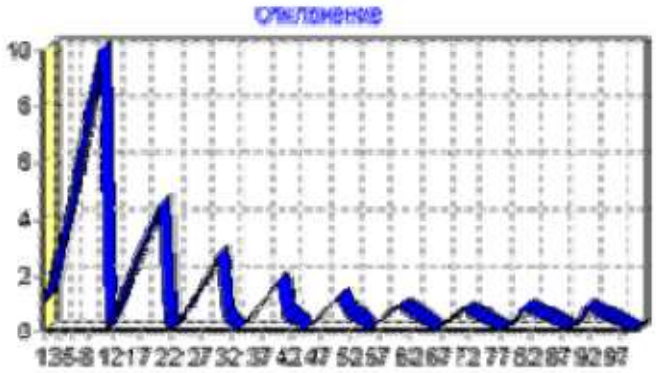

Рис. 8.46. Обчислювання відхилення.

5. Кусково-задана функція. Хай функція приймає значення SQRT(АРГУМЕНТ3/50) при значеннях АРГУМЕНТ3 від 0 до 50 і значеннях АРГУМЕНТ3*АРГУМЕНТ3/2500 - останніх. Для обчислення подібної функції необхідно скористатися функцією *IFF(аргумент1; аргумент2; аргумент3),* яка дозволяє залежно від логічного значення першого аргументу отримати другий або третій аргумент. Згідно прикладу, якщо значення аргументу більше нуля і менше 50 необхідно отримати вираз SQRT(АРГУМЕНТ3/50), інакше – вираз АРГУМЕНТ3*АРГУМЕНТ3/250. Таким чином, в полі побудови виразу необхідно написати «IFF((COL3>0)AND(COL3<50); SQRT(COL3/50); COL3*COL3/2500)». Зробивши це в майстрові обробки «*Вычисляемые данные*», і вибравши далі візуалізацію «*Диаграмма*», і також вибравши в майстрові налаштування діаграми поле із значеннями кусково-заданої функції, можна поглянути на необхідний результат (рис. 8. 47).

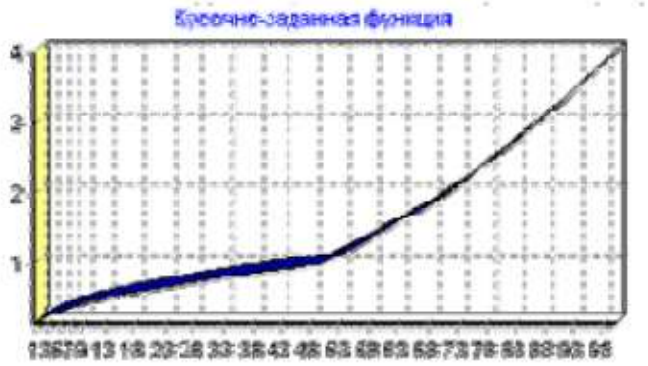

Рис. 8.47. Результати розрахунку кусково-заданої функції.

ЧАСТИНА 5. Групування даних.

Аналітикові для ухвалення рішення завжди необхідна звідна інформація. Сукупні дані набагато більше інформативні, тим більше, якщо їх можна отримати в різних розрізах. У Deductor Studio передбачений інструмент, що реалізовує збір звідної інформації, – «*Группировка*». Групування дозволяє об'єднувати записи по полях - вимірах і агрегувати дані в полях-фактах для подальшого аналізу.

1. Допустимо, що у аналітика є статистика по банках України за певний період. Перед ним стоїть задача виявлення ряду міст, в яких прибуток банків найбільший. Для цього аналітик повинен звернути увагу на наступні поля таблиці з файлу: «БАНК», «ФІЛІЇ», «МІСТО», «ПРИБУТОК». Ясно, що для вирішення поставленої задачі насамперед необхідно знайти сумарний прибуток всіх банків в кожному місті. Для цього і необхідне групування.

По-перше, слід імпортувати дані по банках з табличного файлу. Проглянути вхідну інформацію можна у вигляді куба, де по рядках будуть назви банків, а по стовпцях – міста. За допомогою візуалізації «*Куб*» також можна отримати необхідну інформацію, вибравши в якості виміру поле «МІСТО», а в якості факту - «ПРИБУТОК». Але нам необхідно отримати ці дані для подальшої обробки, отже, необхідно зробити аналогічне групування.

2. Групування по містах. Знаходячись у вузлі імпорту, запустимо майстер обробки. Виберемо в якості обробки – групування даних. На другому кроці майстра встановимо призначення поля «МІСТО» як вимір, а призначення поля «ПРИБУТОК» як факт. В якості функції агрегації поля «ПРИБУТОК» слід вказати «Суму» (рис. 8.48).

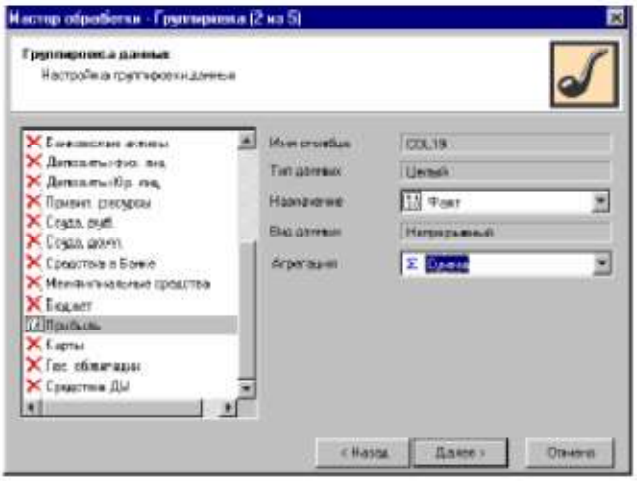

Рис. 8.48. Меню обробки групування даних.

Таким чином, після обробки отримаємо сумарні дані по прибутку всіх банків по кожному місту. Їх можна проглянути, використовуючи таблицю. Тепер аналітикові можна виконувати наступний етап обробки даних (рис. 8.49).

Город	Прибыль
Москва	6076922
Санкт-Петербург	233620
Уфа	370468
Санкт-Петербург	128038
Ханты-Мансийск	30679
Казань	68576
Челябинск	63956

Рис. 8.49. Результат групування даних.

ЧАСТИНА 6. Перетворення даних до ковзаючого вікна.

Коли потрібно спрогнозувати часовий ряд, тим більше, якщо в наявності його періодичність (сезонність), то кращого результату можна

добитися, враховуючи значення чинників не лише в даний момент часу, але і за неділю тому, дві, три, сезон назад, або два. Таку можливість можна дістати після трансформації даних до ковзаючого вікна. Так, наприклад, при сезонності продажів з періодом 12 місяців, для прогнозування кількості продажів на місяць вперед можна як вхідний чинник вказати не лише значення кількості продажів за попередній місяць, але і за 12 місяців назад. Обробка створює нові стовпці шляхом здвигу даних вхідного стовпця вниз і вгору (глибина занурення, горизонт прогнозу).

1. Продемонструємо принцип трансформації даних, використовуючи дані з файлу. У ньому всього 2 поля –«АРГУМЕНТ» - аргумент (час), «ФУНКЦІЯ» – часовий ряд. Імпортуємо дані з файлу (необхідно вказати тип полів – вещественний) і побудуємо діаграму (рис. 8.50).

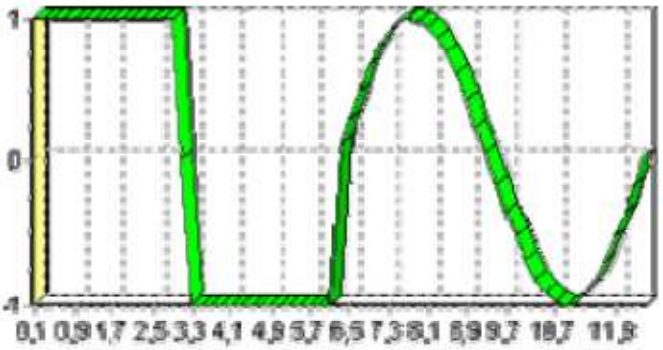

Рис. 8.50. Діаграма вхідних даних.

2. Перетворення ковзаючим вікном. У майстрові перетворення вкажемо призначення стовпця «ФУНКЦІЯ» використовуваним, встановимо для нього глибину занурення 12 і горизонт прогнозу 1.

Після трансформації були отримані нові стовпці – «ФУНКЦІЯ - 11», ... «ФУНКЦИЯ - 1», «ФУНКЦІЯ + 1» на основі стовпця «ФУКЦИЯ». Якщо на діаграмі проглянути декілька таких стовпців, то видно, що дані в них зрушені відносно один одного (рис. 8.51).

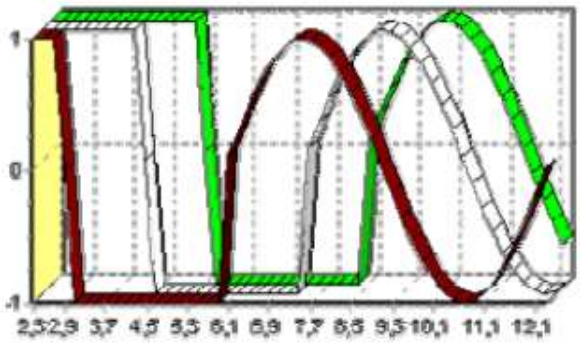

Рис. 8.51. Результати трансформації даних.

ЧАСТИНА 7. Прогнозування за допомогою лінійного регресійного аналізу.

1. Для проведення лінійного регресійного аналізу необхідно запустити майстер обробки і вибрати в якості обробки даних лінійну регресію. На другому кроці майстра налаштуємо поля вхідних даних. Вочевидь, що чинниками будуть спостереження, температура, легковажність і вологість, а результатом – рішення про те, грати в гольф чи ні. Тому необхідно вказати призначення поля «КОД» як інформаційне, поля «РІШЕННЯ» як вихідне, а призначення останніх полів – як вхідні (рис. 8.52).

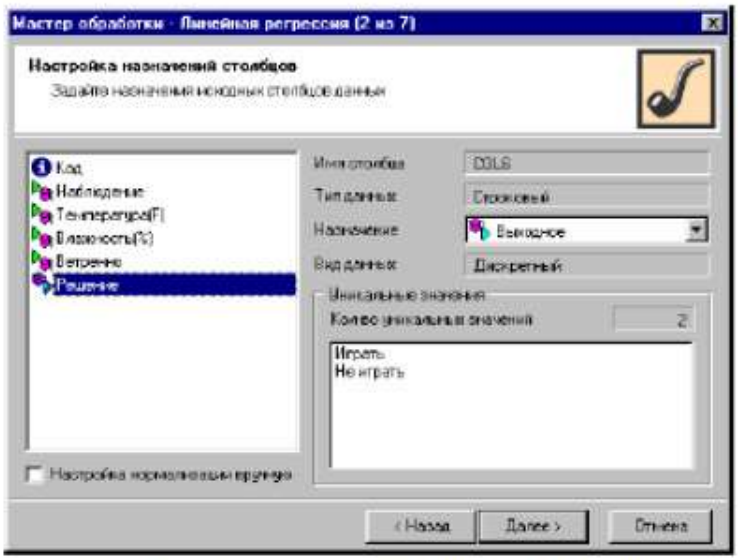

Рис. 8.52. Налаштування поля вхідних даних.

2. На наступному кроці необхідно налаштувати спосіб розділення вхідної множини даних на тестове і навчальне, а також кількість прикладів в тій і іншій множині. Вкажемо, що дані обох множин беруться випадковим чином. Задамо розмір тестової множини рівним двом прикладам шляхом зміни значення стовпця «Розмір в рядках» рядка «Тестова множина».

3. Наступний крок майстра дозволяє виконати обробку, натискуючи на кнопку «Пуск». Під час навчання відображається поточна величина помилки і відсоток розпізнаних прикладів. Після побудови моделі, можна, скориставшись візуалізацією «Що - якщо», зробити прогноз на основі введених метеоумов, а також оцінити якість побудованої моделі, використовуючи таблицю зв'язаності (рис. 8.53).

Фактически	Классифицировано		
	Играть	Не играть	Итого
Играть	8	1	9
Не играть	3	2	5
Итого	11	3	14

	Наблюдение	Темп	Влажн	Ветренно	Решение	Решение_OUT
▶	Солнечно	75	70	Да	Играть	Играть
	Солнечно	69	70	Нет	Играть	Играть
	Пасмурно	83	78	Нет	Играть	Играть

Рис. 8.53. Результат регресійного аналізу.

На таблиці зв'язаності видно, що не всі приклади були класифіковані правильно. Наприклад, ситуація, коли йде дощ, вітер, вологість 80% і температура 65 за Фаренгейтом (18 за Цельсієм) була розцінена як сповна прийнятна для гри в гольф, хоча за таких умов в гольф не грали.

Діаграма «Що – якщо» дозволить провести аналіз впливу одного чинника на результат ухвалення рішення при незмінних останніх (рис. 8.54). З неї, зокрема, видно, що за інших рівних умов (сонячно, температура 75, вітер) на рішення про гру сильно впливає вологість. Якщо вона менше 73%, то варто грати в гольф, якщо більше – то ні.

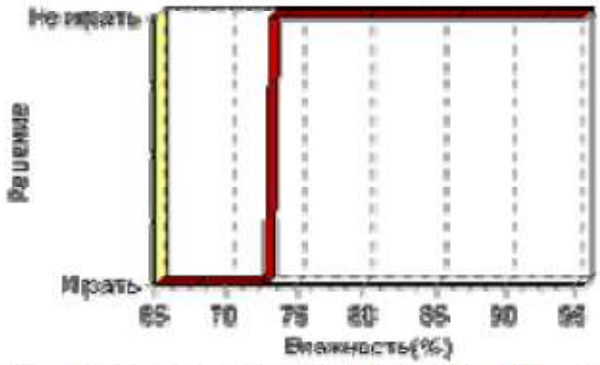

Рис. 8.54. Аналіз за допомогою діаграми «Що – якщо».