

**Министерство образования и науки Украины**  
**Донбасская государственная машиностроительная академия**

# **ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ**

## **МЕТОДИЧЕСКИЕ УКАЗАНИЯ**

**к выполнению лабораторных работ  
и самостоятельной работе**

(для студентов специальности  
«Системы и методы принятия решений»)

Утверждено  
на заседании кафедры ИСПР  
Протокол № 2 от 9 сентября 2014г.

**Краматорск 2014**

## **УДК 681.3**

Интеллектуальный анализ данных : методические указания к выполнению лабораторных работ и самостоятельной работе (для студентов специальности «Системы и методы принятия решений» всех форм обучения) / А.Ю. Мельников – Краматорск : ДГМА, 2014. – 28 с.

Приведены задания к интеллектуальному анализу данных, а также вопросы для самоподготовки.

Составитель	Мельников А.Ю., канд. техн. наук, доцент
-------------	--

Отв. за выпуск	Мельников А.Ю., канд. техн. наук, доцент
----------------	--

## СОДЕРЖАНИЕ

Лабораторная работа №1. Задача классификации.....	4
Лабораторная работа №2. Задача поиска ассоциативных правил.....	12
Лабораторная работа №3. Реализация метода интеллектуального анализа.....	21
Лабораторная работа №4. Постановка и решение комплексной задачи.	22
Вопросы для самоподготовки по теоретическому материалу.....	26

## ЛАБОРАТОРНАЯ РАБОТА №1

### Задача классификации

**Цель работы:** получение навыков решения задачи классификации методом деревьев решений в среде аналитического пакета «Deductor».

#### Задание к работе

1. Изучить структуру аналитического пакета «Deductor» и приложения «Deductor Studio Lite».
2. Выяснить назначения составных частей пакета – «Deductor Viewer» и «Deductor Warehouse»
3. Определить перечень задач, которые могут быть решены с помощью данного пакета.
4. В MS-Excel подготовить исходные данные согласно варианту (табл. 1). Возможные характеристики и их значения придумать самостоятельно. Рекомендуемое число записей – не менее 50.
5. Импортировать подготовленные данные в «Deductor».
6. Произвести расчет задачи классификации методом деревьев решений. Представить результат в виде нескольких визуализаторов.

Таблица 1 – Варианты заданий

№	Задание
1	2
1.	Решение о выдаче кредита клиенту банка на основании списка характеристик предыдущих клиентов
2.	Решение о выдаче кредита клиенту банка на основании его кредитной истории
3.	Предположение о том, какой именно кредит попросит клиент банка, на основании списка характеристик предыдущих клиентов
4.	Предположение о том, какой именно кредит попросит клиент банка, на основании его кредитной истории
5.	Решение о том, является ли клиент потенциальным покупателем на основании списка характеристик предыдущих покупателей
6.	Решение о том, является ли клиент потенциальным покупателем на основании списка его предыдущих покупок
7.	Предположение о том, какой именно товар клиент приобретет в магазине, на основании списка характеристик предыдущих покупателей
8.	Предположение о том, какой именно товар клиент приобретет в магазине, на основании списка его предыдущих покупок
9.	Решение о том, приобретет ли потенциальный покупатель автомобиль, на основании списка покупок предыдущих покупателей
10.	Решение о том, приобретет ли потенциальный покупатель автомобиль, на основании списка его предыдущих покупок

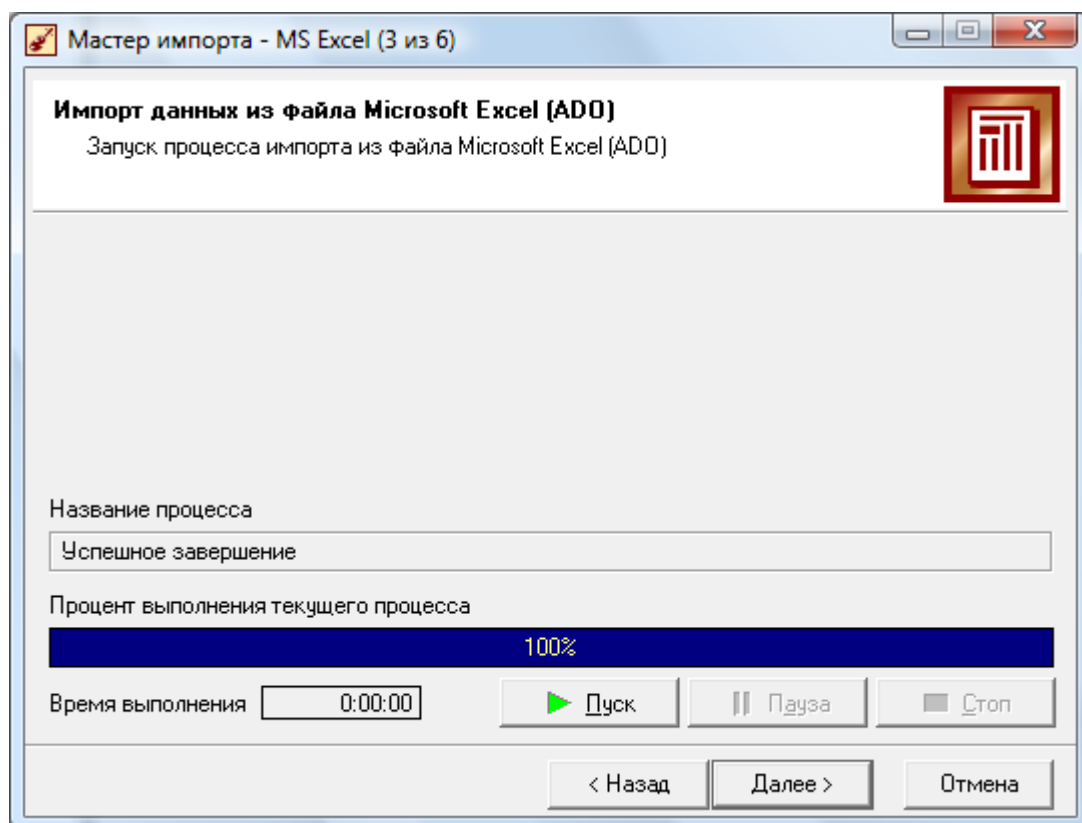
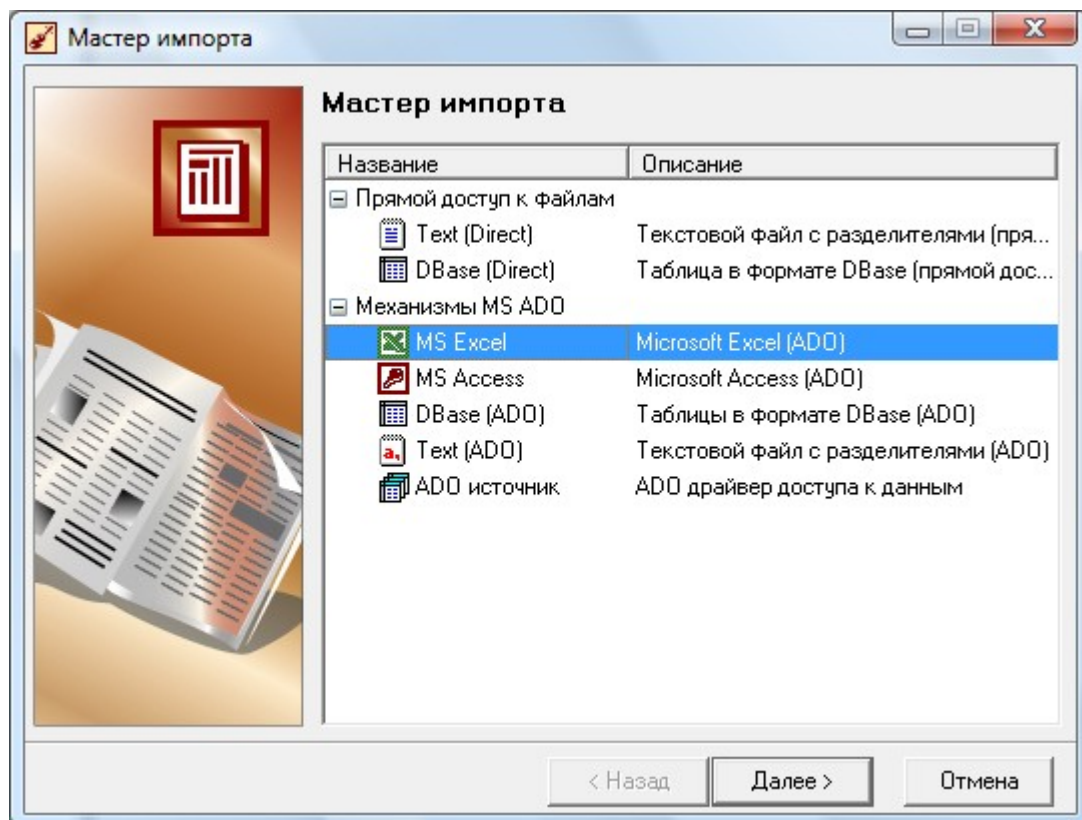
*Продолжение таблицы 1*

1	2
11	Предположение о том, какую именно марку автомобиля приобретет потенциальный покупатель, на основании списка покупок предыдущих покупателей
12	Предположение о том, какую именно марку автомобиля приобретет потенциальный покупатель, на основании списка его предыдущих покупок
13	Решение о страховании клиента на основании списка характеристик предыдущих клиентов
14	Решение о страховании клиента на основании его страховой истории
15	Предположение о том, какой именно договор страхования попросит заключить клиент, на основании списка характеристик предыдущих клиентов
16	Предположение о том, какой именно договор страхования попросит заключить клиент, на основании его страховой истории
17	Решение о предпочтительной категории отдыха («комфортный» или «молодежный») клиента туристического агенства на основании информации о других клиентах
18	Предположение о желаемой стране отдыха клиента туристического агенства на основании информации о других клиентах
19	Решение о том, является ли объект живым существом, на основании информации о характеристиках других живых существ
20	Предположение о том, каким именно существом является объект, на основании информации о характеристиках других живых существ
21	Решение о съедобности гриба на основании имеющихся данных об его внешних характеристиках
22	Предположение о влиянии гриба на здоровье съевшего его человека на основании имеющихся данных о характеристиках гриба

### **Пример выполнения задания**

Задача: решение о выдаче кредита клиенту банка на основании списка характеристик предыдущих клиентов

Доход	Квартира	Члены семьи	Работа	Возраст	Сфера занятости	Кредиты	Решение
2300	нет	3	есть	45	рабочий	нет	выдавать
1600	есть	2	есть	34	рабочий	нет	выдавать
879	есть	3	есть	30	рабочий	нет	отказать
1580	есть	2	есть	41	рабочий	есть	отказать
800	нет	2	нет	50	пенсионер	нет	отказать
760	есть	2	есть	28	рабочий	нет	выдавать
900	есть	4	есть	26	рабочий	нет	выдавать
1500	нет	3	есть	33	рабочий	есть	отказать
2100	есть	3	есть	29	рабочий	есть	выдавать
2470	есть	3	нет	47	рабочий	нет	выдавать
1205	есть	3	есть	37	рабочий	есть	отказать
1347	есть	3	есть	45	рабочий	нет	выдавать
890	есть	3	есть	25	рабочий	нет	отказать
980	есть	3	есть	33	рабочий	нет	отказать
1000	есть	3	есть	44	рабочий	нет	выдавать
450	есть	2	нет	20	студент	есть	отказать
3000	есть	2	есть	43	рабочий	есть	выдавать
2550	есть	2	есть	40	рабочий	нет	выдавать
1630	есть	4	есть	30	рабочий	нет	выдавать
745	есть	4	есть	31	рабочий	нет	отказать
1300	есть	4	есть	28	рабочий	нет	выдавать
2220	есть	3	есть	49	рабочий	есть	выдавать
3800	нет	4	есть	38	рабочий	есть	выдавать
2490	нет	3	есть	46	рабочий	нет	выдавать
480	есть	2	нет	18	студент	нет	отказать
400	нет	2	нет	21	студент	нет	отказать
584	есть	3	есть	19	студент	нет	выдавать
780	нет	4	есть	28	рабочий	нет	выдавать
1900	есть	3	есть	33	рабочий	нет	выдавать
2400	есть	3	есть	39	рабочий	нет	выдавать
3200	нет	4	есть	26	рабочий	есть	выдавать
1500	есть	2	нет	65	пенсионер	есть	отказать
900	есть	4	нет	57	пенсионер	нет	отказать
990	есть	3	нет	37	рабочий	нет	отказать
2000	есть	3	есть	43	рабочий	нет	выдавать
600	есть	2	есть	22	студент	нет	выдавать
4000	нет	2	нет	60	пенсионер	нет	выдавать
750	нет	4	есть	51	рабочий	нет	выдавать
800	есть	2	нет	58	пенсионер	нет	отказать
2400	нет	4	нет	46	рабочий	нет	выдавать



Deductor Studio Lite (Новый) - [MS Excel (База данных: D:\DataMining\IS04zt\Деканенко.xls (Таблица: Лист1\$))]

Файл Правка Вид Сервис Окно ?

Сценарии

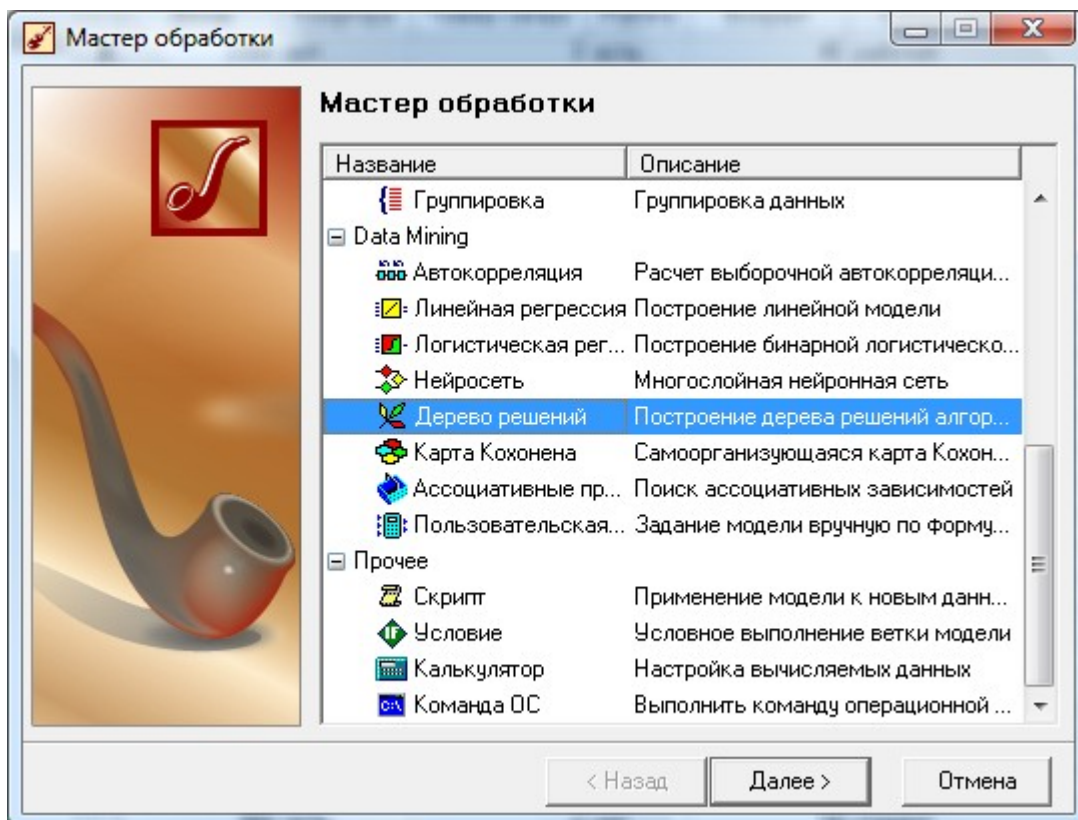
Сценарии

MS Excel (База данных: D:\DataMining\IS04zt\Деканенко.xls (Таблица: Лист1\$))

Таблица

1 / 40

	Доход	Квартира	Члены семьи	Работа	Возраст	Сфера занятости	Кредиты	Решение
	2300	нет	3	есть	45	рабочий	нет	выдавать
	1600	есть	2	есть	34	рабочий	нет	выдавать
	879	есть	3	есть	30	рабочий	нет	отказать
	1580	есть	2	есть	41	рабочий	есть	отказать
	800	нет	2	нет	50	пенсионер	нет	отказать
	760	есть	2	есть	28	рабочий	нет	выдавать
	900	есть	4	есть	26	рабочий	нет	выдавать
	1500	нет	3	есть	33	рабочий	есть	отказать
	2100	есть	3	есть	29	рабочий	есть	выдавать
	2470	есть	3	нет	47	рабочий	нет	выдавать
	1205	есть	3	есть	37	рабочий	есть	отказать
	1347	есть	3	есть	45	рабочий	нет	выдавать
	890	есть	3	есть	25	рабочий	нет	отказать
	980	есть	3	есть	33	рабочий	нет	отказать
	1000	есть	3	есть	44	рабочий	нет	выдавать
	450	есть	2	нет	20	студент	есть	отказать
	3000	есть	2	есть	43	рабочий	есть	выдавать
	2550	есть	2	есть	40	рабочий	нет	выдавать
	1630	есть	4	есть	30	рабочий	нет	выдавать
	745	есть	4	есть	31	рабочий	нет	отказать
	1300	есть	4	есть	28	рабочий	нет	выдавать
	2220	есть	3	есть	49	рабочий	есть	выдавать
	3800	нет	4	есть	38	рабочий	есть	выдавать
	2490	нет	3	есть	46	рабочий	нет	выдавать
	480	есть	2	нет	18	студент	нет	отказать
	400	нет	2	нет	21	студент	нет	отказать
	584	есть	3	есть	19	студент	нет	выдавать
	780	нет	4	есть	28	рабочий	нет	выдавать
	1900	есть	3	есть	33	рабочий	нет	выдавать





**Мастер обработки - Дерево решений (3 из 7)**

**Разбиение исходного набора данных на подмножества**  
 Настройте разбиение исходного множества данных на обучающее, тестовое и валидационное множества

Способ разделения исходного множества данных: Случайно

Множество	Размер		Порядок сортировки
	В процентах	В строках	
<input checked="" type="checkbox"/> Обучающее	95,00	38	По возрастанию
<input checked="" type="checkbox"/> Тестовое	5,00	2	По возрастанию
<b>ИТОГО:</b>	100,00	40	

Количество строк (всего) 40

< Назад Далее > Отмена

**Мастер обработки - Дерево решений (5 из 7)**

**Построение дерева решений**  
 Запустите процесс построения дерева решений

Распределено, шт.

☒ Распознано 34

☒ Нераспознано 4

Распознано, %

Обучающее мн-во 89,47

Тестовое мн-во 100,00

Кол-во узлов 7

Кол-во правил 4

Время обучения 0:00:00

Темп обновления 100

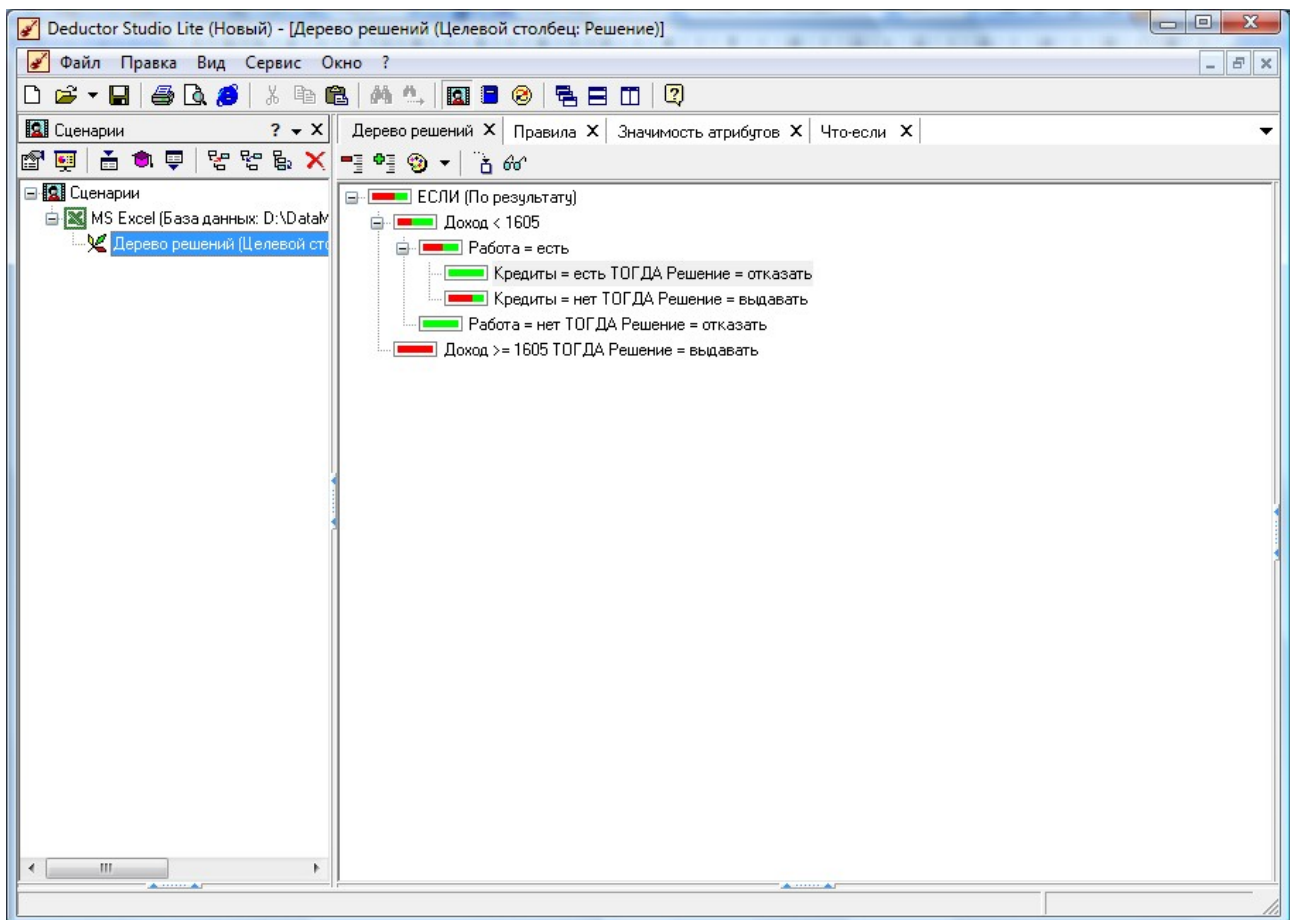
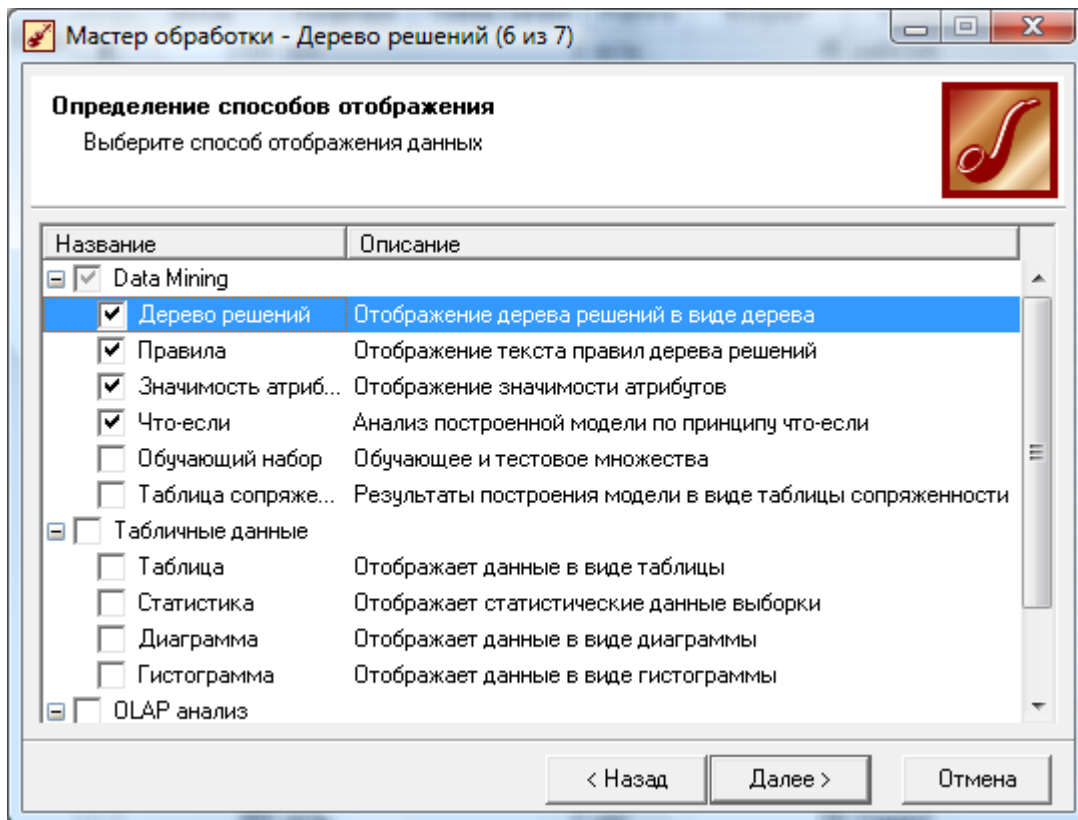
▶ Пуск

|| Пауза

■ Стоп

100%

< Назад Далее > Отмена



Deductor Studio Lite (Новый) - [Дерево решений (Целевой столбец: Решение)]

Файл Правка Вид Сервис Окно ?

Сценарии ? X

Дерево решений X Правила X Значимость атрибутов X Что-если X

Сценарии

- MS Excel (База данных: D:\Data\...
- Дерево решений (Целевой столбец: Решение)

Фильтр: Без фильтрации

Итого правил: 4

№	Условие	Следствие (Решение)	Поддержка		Достоверность	
			%	Кол-во	%	Кол-во
1	Доход < 1605 И Работа = есть И Кредиты = есть	отказывать	7,89	3	100,00	3
2	Доход < 1605 И Работа = есть И Кредиты = нет	выдавать	34,21	13	69,23	9
3	Доход < 1605 И Работа = нет	отказывать	18,42	7	100,00	7
4	Доход >= 1605	выдавать	39,47	15	100,00	15

Deductor Studio Lite (Новый) - [Дерево решений (Целевой столбец: Решение)]

Файл Правка Вид Сервис Окно ?

Сценарии ? X

Дерево решений X Правила X Значимость атрибутов X Что-если X

Сценарии

- MS Excel (База данных: D:\Data\...
- Дерево решений (Целевой столбец: Решение)

Целевой атрибут: Решение

№	Атрибут	Значимость, %	/
1	Доход	56,604	
4	Работа	26,080	
7	Кредиты	17,316	
6	Сфера занятости	0,000	
5	Возраст	0,000	
3	Члены семьи	0,000	
2	Квартира	0,000	

Deductor Studio Lite (Новый) - [Дерево решений (Целевой столбец: Решение)]

Файл Правка Вид Сервис Окно ?

Сценарии ? X

Дерево решений X Правила X Значимость атрибутов X Что-если X

Сценарии

- MS Excel (База данных: D:\Data\...
- Дерево решений (Целевой столбец: Решение)

1 из 40

Поле	Значение
Входные	
9.0 Доход	2300
ab Квартира	нет
9.0 Члены семьи	3
ab Работа	есть
9.0 Возраст	45
ab Сфера занятости	рабочий
ab Кредиты	нет
Выходные	
ab Решение	выдавать
Расчетные	
12 Решение Номер ...	4
9.0 Решение Поддер...	39,4736842105263
9.0 Решение Достов...	100

## ЛАБОРАТОРНАЯ РАБОТА №2

### Задача поиска ассоциативных правил

**Цель работы:** получение навыков решения задачи поиска ассоциативных правил в аналитическом пакете «Deductor».

#### Задание к работе

1. Сформировать список транзакций, используя заданные ограничения согласно варианту. Исходную таблицу сформировать с использованием вспомогательного приложения «Make\_td».
2. Импортировать данные в «Deductor».
3. Произвести расчет задачи поиска ассоциативных правил. Представить результат в виде нескольких визуализаторов.
4. На защите продемонстрировать процесс импорта и работы.

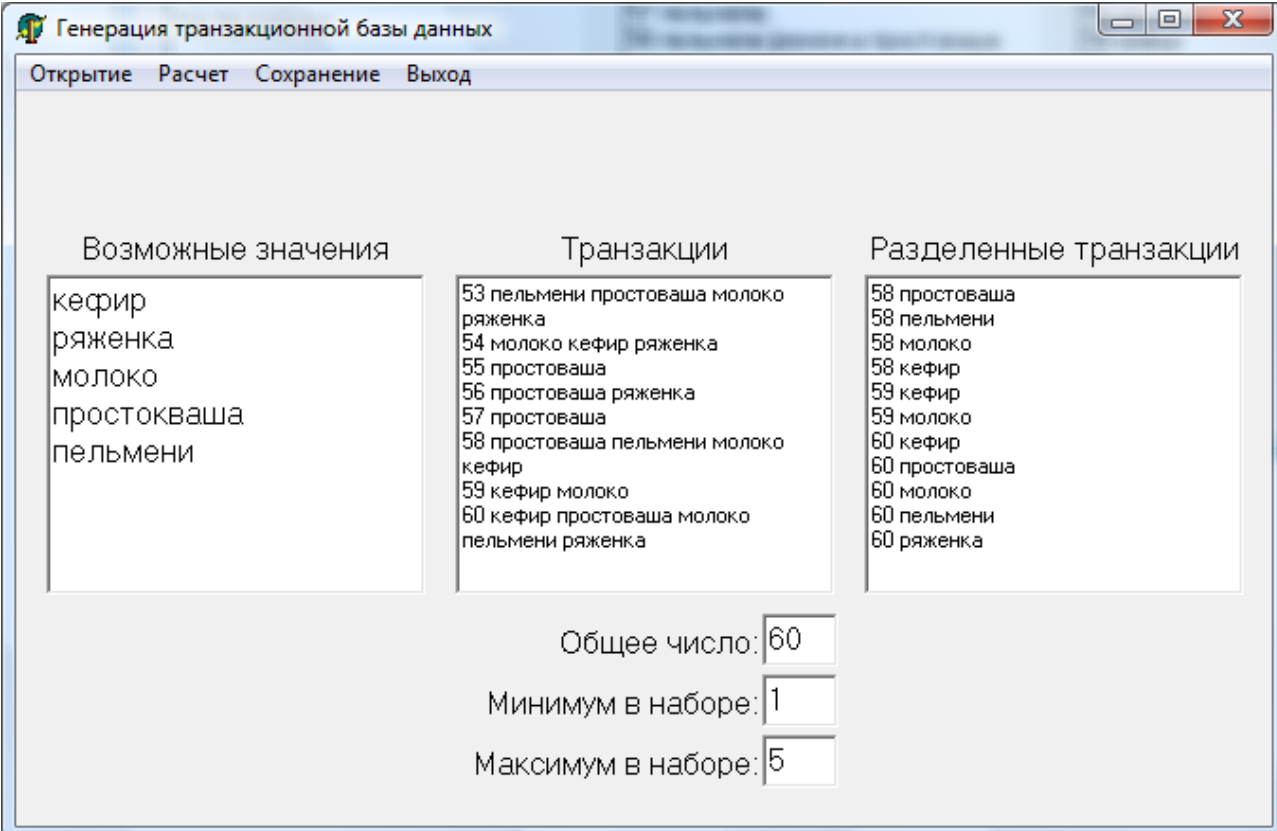
Таблица 2 – Варианты заданий

Вар	Общее число продуктов	Минимальное число продуктов в наборе	Максимальное число продуктов в наборе
1.	4	1	4
2.	5	1	5
3.	6	2	6
4.	7	2	7
5.	4	1	4
6.	5	1	5
7.	6	2	6
8.	7	2	7
9.	4	1	4
10.	5	1	5
11.	6	2	6
12.	7	2	7
13.	4	1	4
14.	5	1	5

#### Пример выполнения задания

Общее число продуктов	Минимальное число продуктов в наборе	Максимальное число продуктов в наборе
5	1	5

## Данные



Генерация транзакционной базы данных

Открытие Расчет Сохранение Выход

Возможные значения

кефир  
ряженка  
молоко  
простокваша  
пельмени

Транзакции

53 пельмени простокваша молоко  
ряженка  
54 молоко кефир ряженка  
55 простокваша  
56 простокваша ряженка  
57 простокваша  
58 простокваша пельмени молоко  
кефир  
59 кефир молоко  
60 кефир простокваша молоко  
пельмени ряженка

Разделенные транзакции

58 простокваша  
58 пельмени  
58 молоко  
58 кефир  
59 кефир  
59 молоко  
60 кефир  
60 простокваша  
60 молоко  
60 пельмени  
60 ряженка

Общее число: 60

Минимум в наборе: 1

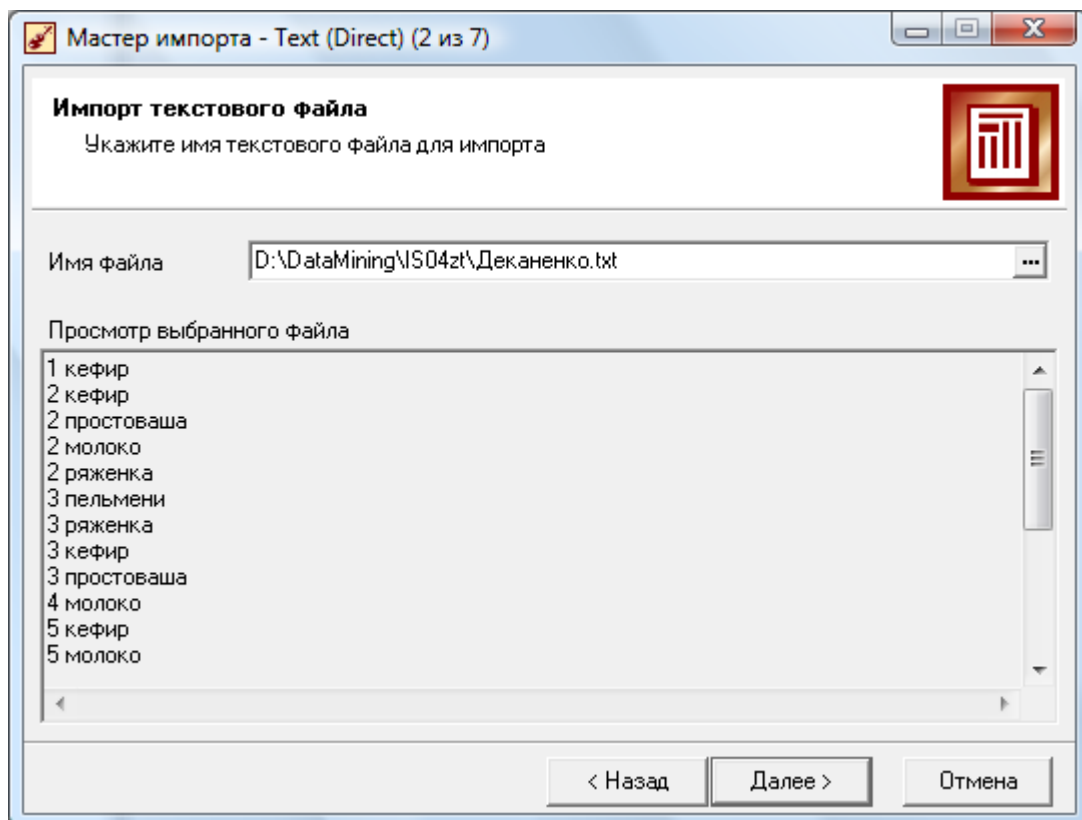
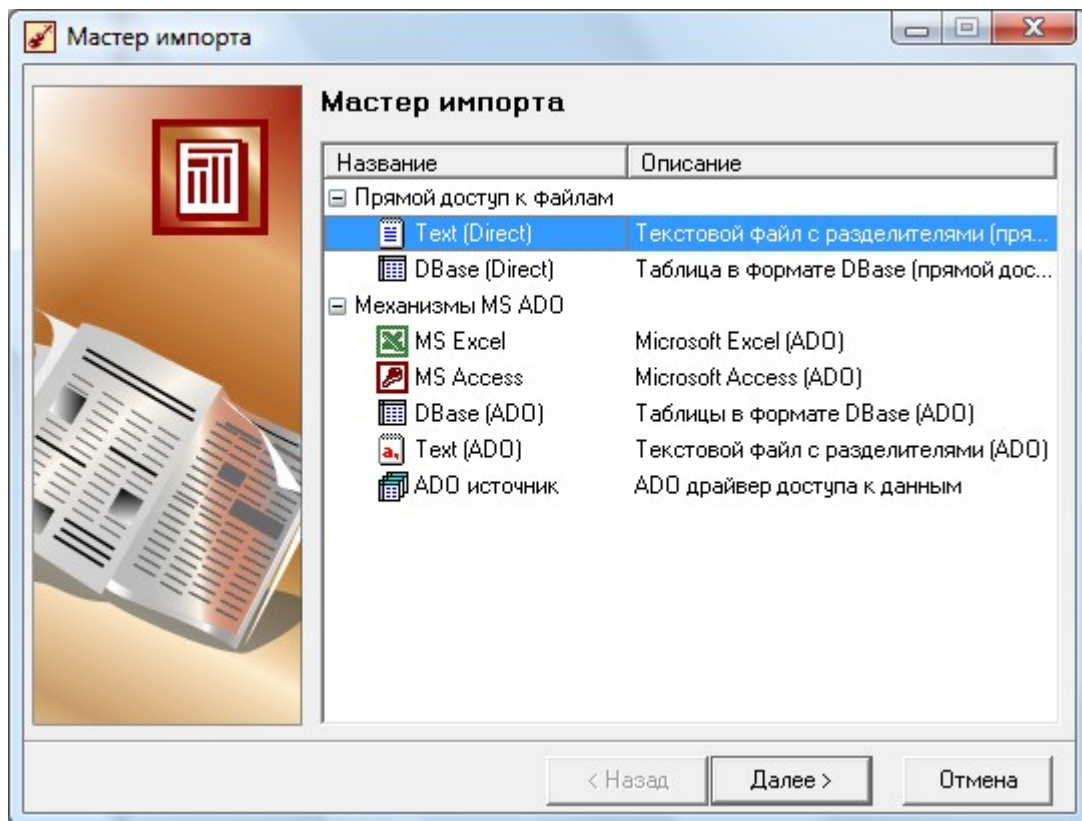
Максимум в наборе: 5

## Файл с транзакциями

- 1 кефир
- 2 кефир простокваша молоко ряженка
- 3 пельмени ряженка кефир простокваша
- 4 молоко
- 5 кефир молоко ряженка пельмени простокваша
- 6 пельмени кефир ряженка молоко простокваша
- 7 ряженка простокваша пельмени
- 8 пельмени
- 9 простокваша молоко
- 10 молоко кефир ряженка пельмени простокваша
- 11 молоко
- 12 пельмени ряженка кефир простокваша молоко
- 13 молоко простокваша кефир пельмени
- 14 простокваша ряженка пельмени молоко кефир
- 15 кефир простокваша пельмени
- 16 молоко пельмени ряженка кефир
- 17 простокваша молоко кефир пельмени
- 18 ряженка простокваша молоко пельмени
- 19 ряженка пельмени
- 20 молоко кефир

- 21 кефир молоко пельмени
- 22 простокваша молоко
- 23 простокваша кефир молоко ряженка пельмени
- 24 ряженка пельмени кефир
- 25 кефир ряженка пельмени
- 26 ряженка простокваша пельмени кефир
- 27 пельмени ряженка
- 28 простокваша
- 29 пельмени
- 30 ряженка
- 31 пельмени простокваша ряженка
- 32 пельмени молоко ряженка простокваша кефир
- 33 молоко пельмени простокваша
- 34 кефир пельмени
- 35 кефир пельмени ряженка молоко
- 36 ряженка простокваша
- 37 простокваша
- 38 ряженка
- 39 ряженка простокваша кефир пельмени молоко
- 40 молоко пельмени ряженка
- 41 молоко кефир ряженка пельмени
- 42 ряженка кефир пельмени простокваша
- 43 пельмени простокваша кефир
- 44 пельмени простокваша ряженка кефир молоко
- 45 простокваша ряженка кефир молоко пельмени
- 46 пельмени молоко простокваша ряженка
- 47 кефир ряженка пельмени молоко простокваша
- 48 простокваша ряженка кефир молоко пельмени
- 49 кефир ряженка простокваша
- 50 пельмени кефир простокваша
- 51 кефир пельмени ряженка простокваша
- 52 кефир пельмени ряженка простокваша
- 53 пельмени простокваша молоко ряженка
- 54 молоко кефир ряженка
- 55 простокваша
- 56 простокваша ряженка
- 57 простокваша
- 58 простокваша пельмени молоко кефир
- 59 кефир молоко
- 60 кефир простокваша молоко пельмени ряженка

## Решение в Дедукторе



Мастер импорта - Text (Direct) (4 из 7)

**Импорт текстового файла**  
Укажите параметры столбцов

Столбец	Имя столбца	Метка столбца	Тип данных	Вид данных	Назначение
12 COL1	COL1	COL1	12 Целый	Непрерывный	ID Транзакция
ab COL2					

< Назад    Далее >    Отмена

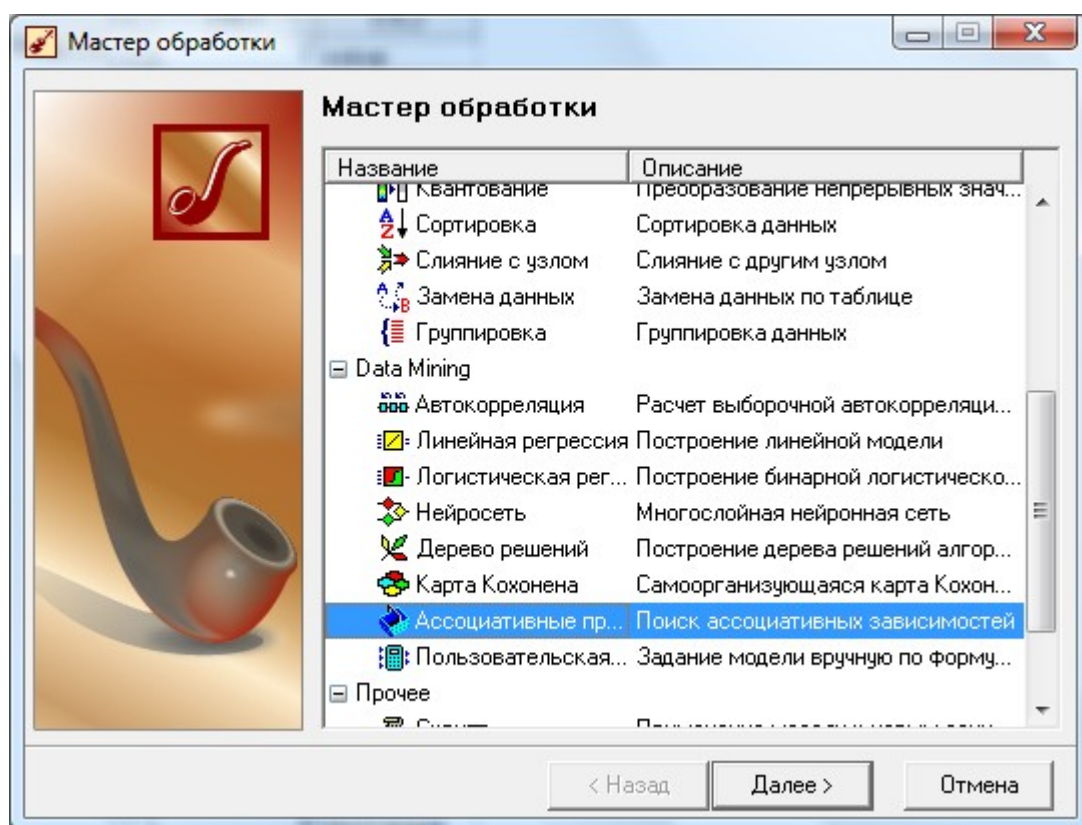
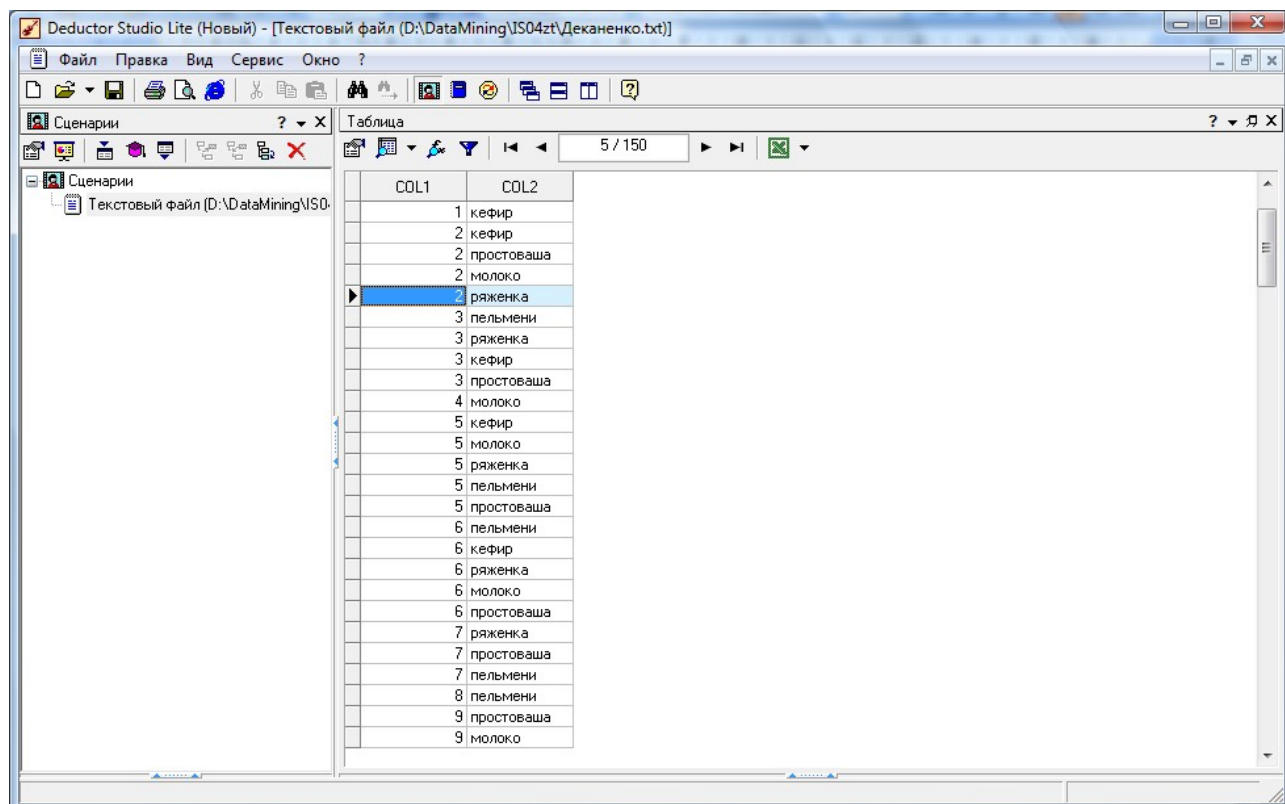
Мастер импорта - Text (Direct) (4 из 7)

**Импорт текстового файла**  
Укажите параметры столбцов

Столбец	Имя столбца	Метка столбца	Тип данных	Вид данных	Назначение
12 COL1					
ab COL2	COL2	COL2	ab Строковый	Дискретный	Входное

< Назад    Далее >    Отмена





Мастер обработки - Ассоциативные правила (2 из 6)

### Настройка назначений столбцов

Задайте назначения исходных столбцов данных

ID COL1

COL2

Имя столбца: COL1

Тип данных: Целый

Назначение: ID Транзакция

Вид данных: Дискретный

Уникальные значения

Кол-во уникальных значений: 48

1  
2  
3  
4  
5  
6  
7

< Назад    Далее >    Отмена

Мастер обработки - Ассоциативные правила (2 из 6)

### Настройка назначений столбцов

Задайте назначения исходных столбцов данных

ID COL1

COL2

Имя столбца: COL2

Тип данных: Строковый

Назначение: Элемент

Вид данных: Дискретный

Уникальные значения

Кол-во уникальных значений: 5

кефир  
молоко  
пельмени  
простоваша  
ряженка

< Назад    Далее >    Отмена

**Мастер обработки - Ассоциативные правила (3 из 6)**

### Настройка параметров построения ассоциативных правил

Укажите значения параметров построения ассоциативных правил

Часто встречающиеся множества

Минимальная поддержка, %

Максимальная поддержка, %

☐ Максимальная мощность искомым часто встречающихся множеств

Ассоциативные правила

Минимальная достоверность, %

Максимальная достоверность, %

< Назад      Далее >      Отмена

**Мастер обработки - Ассоциативные правила (4 из 6)**

### Построение ассоциативных правил

Запуск процесса построения ассоциативных правил

Мощность	Кол-во множеств
1	4
2	6
3	4
4	1

Кол-во множеств

Кол-во правил

Время работы

Текущее состояние процесса построения правил

Процент выполнения текущего процесса

▶ Пуск      || Пауза      ■ Стоп

< Назад      Далее >      Отмена

Deductor Studio Lite (Новый) - [Ассоциативные правила [TID="COL1"; AID="COL2"]]

Файл Правка Вид Сервис Окно ?

Сценарии

Текстовый файл [D:\DataMining\IS042\Деканенко.t  
Ассоциативные правила [TID="COL1"; AID="COL2"]

Правила X Популярныe наборы X Дерево правил X Что-если X

### Ассоциативные правила по элементам транзакций COL2

Фильтр: Без фильтрации

Итого правил: 50

№	Условие	Следствие	Поддержка		Достоверность, %
			%	Кол-во	
1	кефир	молоко	39,58	19	67,86
2	молоко	кефир	39,58	19	70,37
3	кефир	простоваша	39,58	19	67,86
4	простоваша	кефир	39,58	19	63,33
5	кефир	ряженка	41,67	20	71,43
6	ряженка	кефир	41,67	20	66,67
7	молоко	простоваша	39,58	19	70,37
8	простоваша	молоко	39,58	19	63,33
9	молоко	ряженка	37,50	18	66,67
10	ряженка	молоко	37,50	18	60,00
11	простоваша	ряженка	41,67	20	66,67
12	ряженка	простоваша	41,67	20	66,67
13	кефир И молоко	простоваша	29,17	14	73,68
14	кефир И простоваша	молоко	29,17	14	73,68
15	молоко И простоваша	кефир	29,17	14	73,68
16	кефир	молоко И простоваша	29,17	14	50,00
17	молоко	кефир И простоваша	29,17	14	51,85
18	простоваша	кефир И молоко	29,17	14	46,67
19	кефир И молоко	ряженка	31,25	15	78,95
20	кефир И ряженка	молоко	31,25	15	75,00
21	молоко И ряженка	кефир	31,25	15	83,33

Deductor Studio Lite (Новый) - [Ассоциативные правила [TID="COL1"; AID="COL2"]]

Файл Правка Вид Сервис Окно ?

Сценарии

Текстовый файл [D:\DataMining\IS042\Деканенко.t  
Ассоциативные правила [TID="COL1"; AID="COL2"]

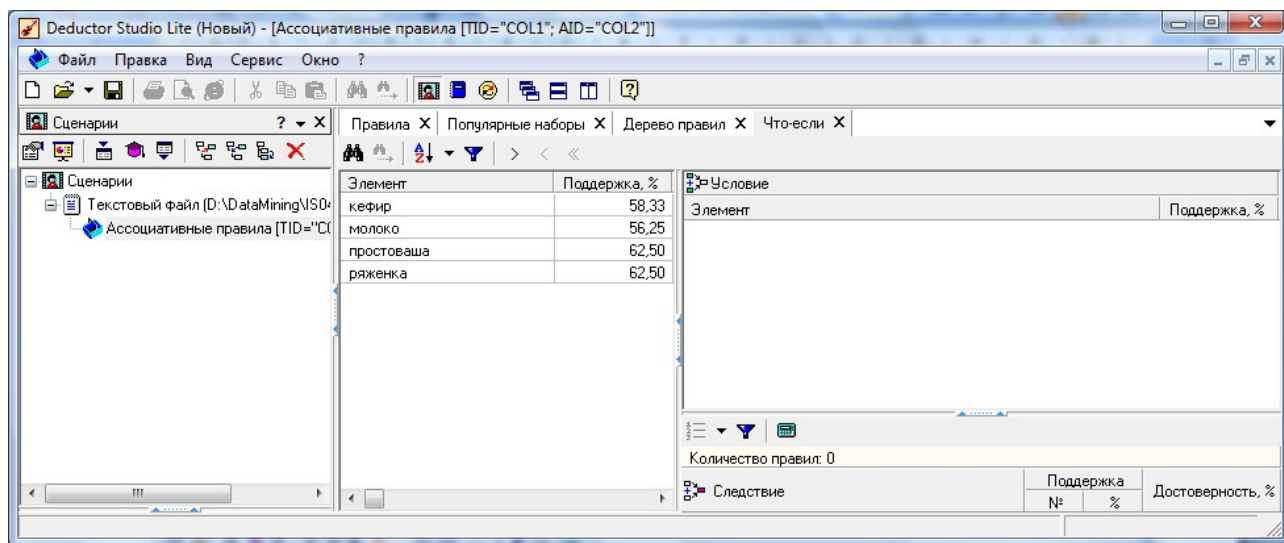
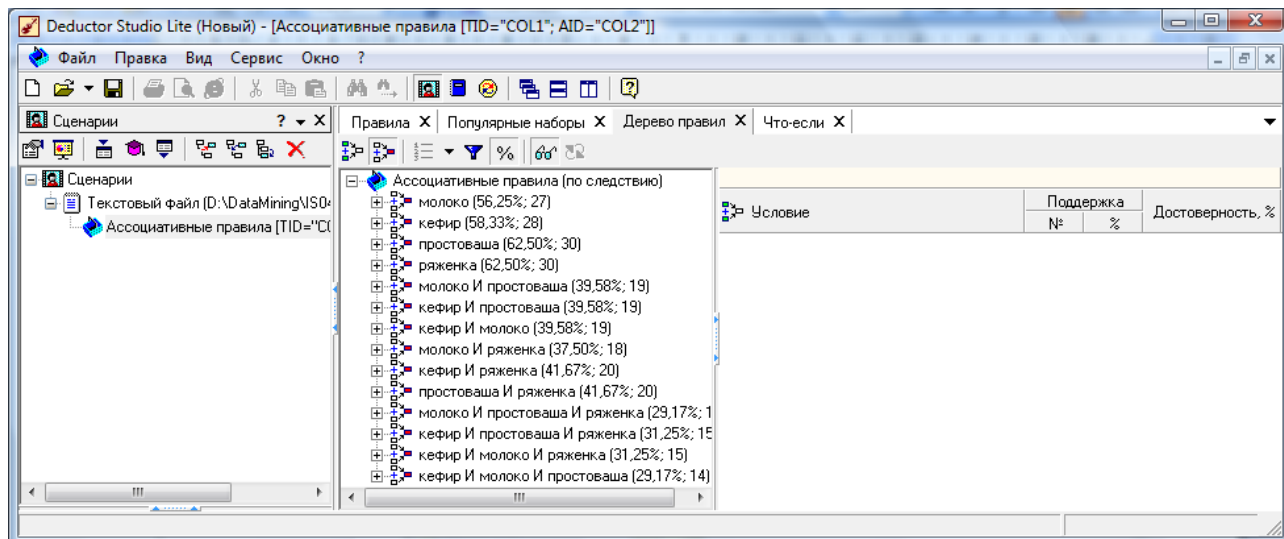
Правила X Популярныe наборы X Дерево правил X Что-если X

### Часто встречающиеся множества по элементам транзакций COL2

Фильтр: Без фильтрации

Итого множеств: 15

№	Множество	Поддержка	
		%	Кол-во
1	кефир	58,33	28
5	кефир И молоко	39,58	19
11	кефир И молоко И простоваша	29,17	14
15	кефир И молоко И простоваша И ряженка	25,00	12
12	кефир И молоко И ряженка	31,25	15
6	кефир И простоваша	39,58	19
13	кефир И простоваша И ряженка	31,25	15
7	кефир И ряженка	41,67	20
2	молоко	56,25	27
8	молоко И простоваша	39,58	19
14	молоко И простоваша И ряженка	29,17	14
9	молоко И ряженка	37,50	18
3	простоваша	62,50	30
10	простоваша И ряженка	41,67	20
4	ряженка	62,50	30



## ЛАБОРАТОРНАЯ РАБОТА №3

### *Реализация метода интеллектуального анализа*

**Цель работы:** получение навыков реализации метода интеллектуального анализа данных в среде программирования.

#### **Задание к работе**

1. Составить программу для реализации выбранного метода (табл. 3).
2. Придумать самостоятельно исходные данные (в качестве хранилища данных использовать MS-Excel, MS-Access или текстовый файл).
3. Применить разработанную программу для обработки данных.
4. Проанализировать полученный результат.

**Примечание:** варианты для этой работы выдаются лектором и могут не совпадать с вариантами по списку, которые используются для остальных лабораторных работ.

Таблица 3 – Варианты заданий

Вар	Методы
I. Методы классификации и прогнозирования:	
1.	метод деревьев решений CART
2.	метод деревьев решений C4.5 (C4.8, C5.0)
3.	метод опорных векторов (support vector machine)
4.	метод ближайшего соседа (nearest neighbor)
5.	метод байесовских сетей (naive-bayes approach)
6.	метод нейронных сетей
II. Методы кластеризации:	
7.	самоорганизующиеся карты Кохонена (self-organizing maps)
8.	иерархический агломеративный метод кластерного анализа (AGNES)
9.	иерархический дивизимный метод кластерного анализа (DIANA)
10.	метод к-средних (k-means)
11.	метод к-медианы (k-medoids, PAM)
12.	метод BIRCH
13.	метод Wave Clusters
14.	метод CLARA
15.	метод Clarans
III. Методы поиска ассоциативных правил:	
16.	метод AIS
17.	метод SETM
18.	метод Apriori
19.	метод AprioriTID
20.	метод AprioriHybrid
21.	метод DHP
22.	метод DIC

## ЛАБОРАТОРНАЯ РАБОТА №4

### *Постановка и решение комплексной задачи*

**Цель работы:** получение навыков постановки задачи применения Data Mining для заданного набора данных, выбора методов для решения поставленной задачи

#### **Задание к работе**

1. Провести этапы процесса Data Mining с 1-го по 6-й (применение и коррекцию модели в данном случае не рассматриваем) на основе индивидуального задания с учетом приведенных далее рекомендаций.
2. Анализ предметной области можно совместить с постановкой задачи.
3. Исходные данные придумать самостоятельно (в качестве хранилища данных использовать MS-Excel, MS-Access или текстовый файл).
4. Этапы работы с моделью проводить в пакете «Deductor».
5. Рассмотреть не менее двух задач.

Отчет должен содержать постановку задачи и решение этой задачи с анализом результатов.

Таблица 4 – Варианты заданий

Вар	Предметная область
1-3	Грибы
4-6	Живые существа
7-9	Туристическое агентство
10-12	Страховая компания
13-15	Приобретение автомобиля
16-18	Кредитование в банке
19-22	Покупка товара в магазине

### Пример постановки и решения задачи

Предметная область: телефонная служба информации.

Объекты: организации и фирмы города, клиенты, журнал регистрации звонков. В данной работе будем рассматривать только последний объект.

Поля журнала регистрации звонков: Дата звонка, Время звонка, Фирма, Товар, Ответ, Реклама, Пол, Диспетчер.

Задачи: классификации (какими товарами и услугами интересуются женщины, а какими мужчины), кластеризации (насколько эффективна работа диспетчеров), поиска ассоциативных правил (перечень наиболее популярных товаров и услуг). Все три задачи решены в пакете «Deductor» с использованием разных алгоритмов.

Время	Фирма	Товар	Ответ	ма нару	Пол	Диспетчер
16:48:53	ЗА3-Сервис КХРП	автосалон	+	-	м	Копыткова
16:49:11	Авто Интернешенел Краматорск ООО	автосалон	+	-	м	Копыткова
14:52:02	Аудит-К ЗАО Аф	аудит	+	+	ж	Копыткова
14:52:04	Имидж-Аудит ООО Аф	аудит	+	+	ж	Копыткова
14:52:21	Натали-Аудит ООО	аудит	+	+	ж	Копыткова
14:52:07	Пионер Аф	аудит, все виды недорого и в короткие ср	+	+	ж	Копыткова
9:07:46	Маг.Охотник-рыболов	бинокль	+	+	ж	Котова
9:07:51	Маг.Охотник-рыболов	бинокль	+	+	ж	Котова
9:07:54	Маг.Охотник-рыболов	бинокль	+	+	ж	Котова
14:41:07	Ивком ТД ООО	быт. химия	+	+	ж	Копыткова
14:41:20	Универсам 1 КП ДП ТД Универсам	быт. химия	+	+	ж	Копыткова
14:41:31	Ситал РСК	быт. химия	+	+	ж	Копыткова
14:42:18	Фиал ООО	быт. химия	+	+	ж	Копыткова
10:41:19	Бытрадиотехника КПКП	быт. техника	+	+	м	Суровцева
10:40:58	Сервис-центр Валдис	быт. техника	+	+	м	Суровцева
10:41:34	Казakov Юрий Александрович Ч/п	быт. техника	+	+	м	Суровцева
11:26:02	-	время	+	-	ж	Суровцева
12:21:55	-	время	+	-	ж	Суровцева
12:48:46	-	время	+	+	м	Суровцева
14:00:25	-	время	+	-	ж	Копыткова
16:47:35	-	время	+	-	ж	Копыткова
17:36:48	-	время	+	-	м	Копыткова
17:38:00	-	время	+	-	м	Копыткова
14:42:27	Фиал ООО	дезинфицирующие средства опт	+	+	ж	Копыткова
9:48:17	-	коды городов	+	+	ж	Суровцева
11:10:19	-	коды городов	+	+	ж	Суровцева
12:03:25	-	коды городов	+	-	ж	Суровцева
12:25:17	-	коды городов	+	+	ж	Суровцева
12:41:17	-	коды городов	+	-	м	Суровцева
14:58:58	-	коды городов	-	-	ж	Копыткова
16:38:49	-	коды городов	-	-	м	Копыткова
16:41:39	Эксофт ООО	компьютеров	+	-	м	Копыткова

Рисунок 1 – Исходные данные

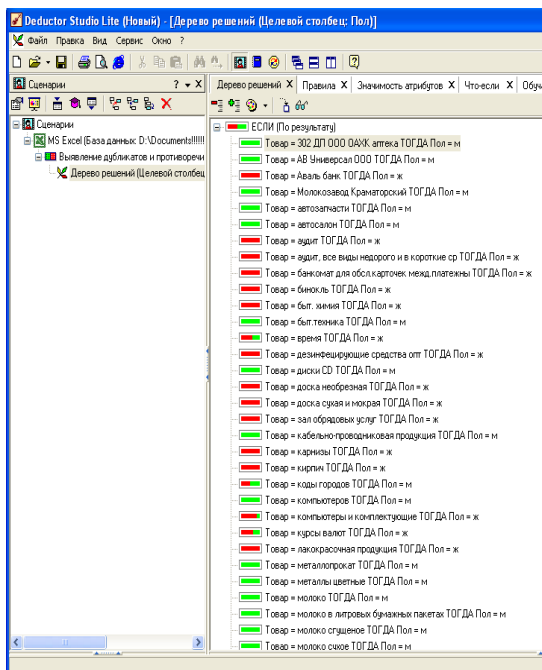


Рисунок 2 – Дерево решений

Дерево решений X Правила X Значимость атрибутов X Что-если X Обучающий набор X Таблица сопряженности

Всё правила

№	Условие	Следствие (Пол)	Поддержка		Достоверность	
			% Кол-во	% Кол-во	% Кол-во	% Кол-во
1	Товар = 302 ДП ООО ОАХК аптека	м	0.67	1	100.00	1
2	Товар = АВ Универсал ООО	м	0.67	1	100.00	1
3	Товар = Аваль банк	ж	0.67	1	100.00	1
4	Товар = Молокозавод Краматорский	м	0.67	1	100.00	1
5	Товар = автозапчасти	м	5.33	8	100.00	8
6	Товар = автосалон	м	1.33	2	100.00	2
7	Товар = аудит	ж	2.00	3	100.00	3
8	Товар = аудит, все виды недорого и в короткие ср	ж	0.67	1	100.00	1
9	Товар = банкомат для обл. карточек межд. платежны	ж	0.67	1	100.00	1
10	Товар = бинокль	ж	2.00	3	100.00	3
11	Товар = быт. химия	ж	2.67	4	100.00	4
12	Товар = быт. техника	м	2.00	3	100.00	3
13	Товар = время	ж	8.67	13	61.54	8
14	Товар = дезинфицирующие средства опт	ж	0.67	1	100.00	1
15	Товар = диски CD	м	2.67	4	100.00	4
16	Товар = доска необрезная	ж	0.67	1	100.00	1
17	Товар = доска сухая и мокрая	ж	0.67	1	100.00	1
18	Товар = зал обрядов услуг	ж	3.33	5	100.00	5
19	Товар = кабельно-проводниковая продукция	м	1.33	2	100.00	2
20	Товар = карнизы	ж	1.33	2	100.00	2
21	Товар = кирпич	ж	1.33	2	100.00	2
22	Товар = коды городов	м	11.33	17	52.94	9
23	Товар = компьютеры	м	0.67	1	100.00	1
24	Товар = компьютеры и комплектующие	ж	11.33	17	82.35	14
25	Товар = курсы валют	ж	5.33	8	62.50	5

Рисунок 3 – Правила

Дерево решений X Правила X Значим

Пол

Фактически	Классифицировано		Итого
	ж	м	
ж	68	8	76
м	13	61	74
Итого	81	69	150

Рисунок 4 – Таблица сопряженности

Таблица

1 / 150

Фирма	Товар	Пол	Противоречие	Дубликат	Группа противоречий	Группа дубликатов	Пол_OUT
Аудит-К ЗАО АФ	аудит	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Имидж-Аудит ООО АФ	аудит	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Натали-Аудит ООО	аудит	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Пионер АФ	аудит, все виды недорого и в короткие ср	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Маг.Охотник-рыболов	бинокль	ж	<input type="checkbox"/>	<input checked="" type="checkbox"/>			9 ж
Маг.Охотник-рыболов	бинокль	ж	<input type="checkbox"/>	<input checked="" type="checkbox"/>			9 ж
Маг.Охотник-рыболов	бинокль	ж	<input type="checkbox"/>	<input checked="" type="checkbox"/>			9 ж
Ивком ТД ООО	быт. химия	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Универсам 1 КП ДП ТД Универсам	быт. химия	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Сигал РСК	быт. химия	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Фиал ООО	быт. химия	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
-	время	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1		1 ж
-	время	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1		1 ж
-	время	м	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1		2 ж
-	время	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1		1 ж
-	время	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1		1 ж
-	время	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>			2 ж
-	время	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>			2 ж
Фиал ООО	дезинфицирующие средства опт	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Маг.Эксперт	компьютеры и комплектующие	м	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	4		10 ж
Салон ВОЗ и Маленькая тележка	компьютеры и комплектующие	м	<input checked="" type="checkbox"/>	<input type="checkbox"/>	5		ж
Маг.Эксперт	компьютеры и комплектующие	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>			10 ж
-	курсы валют	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>	3		6 ж
-	курсы валют	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3		5 ж
-	курсы валют	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>			6 ж
-	курсы валют	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>			6 ж
-	курсы валют	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3		5 ж
-	курсы валют	ж	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3		5 ж
Колос ООО	обувь рабочая	ж	<input type="checkbox"/>	<input checked="" type="checkbox"/>			8 ж
Колос ООО	обувь рабочая	ж	<input type="checkbox"/>	<input checked="" type="checkbox"/>			8 ж
-	погода	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Албо оптовая база	продукты	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж
Каравай оптовая база	продукты	ж	<input type="checkbox"/>	<input type="checkbox"/>			ж

Рисунок 5 – Перечень товаров и услуг, которыми интересуются женщины



Формат: Пол\_OUT

Таблица 127 / 150

Фирма	Товар	Пол	Противоречие	Дубликат	Группа противоречий	Группа дубликатов	Пол_OUT
Карпенко Александр Владимирович Ч.	металлопрокат	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Будкапитал Торговый Дом ООО	металлопрокат	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Сигал РСК	металлопрокат	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>		11	м
Сигал РСК	металлопрокат	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>		11	м
Металлобаза Краматорская ООО	металлопрокат	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Сигал РСК	металлопрокат	м	<input type="checkbox"/>	<input checked="" type="checkbox"/>		11	м
Сигал РСК	металлы цветные	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Молокозавод Краматорский	молоко	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Молокозавод Краматорский	молоко в литровых бумажных пакетах	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Молокозавод Краматорский	молоко сгущенное	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Триал КММП	молоко сухое	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Молокозавод Краматорский	Молокозавод Краматорский	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Лантана ООО	оргтехника	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Салон В03 и Маленькая тележка	оргтехника	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Экософт ООО	оргтехника	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Кописервис (Минолта)	оргтехника	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Бон Тон ООО ДП	отводы стальные	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Сигал РСК	отводы стальные	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Дизайн Плюс ООО	полиграфические	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Офсет 340	полиграфические	м	<input type="checkbox"/>	<input type="checkbox"/>			м
График-Арт	полиграфические	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Контора похоронного обслуживания	ритуальные услуги	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Радио Европа плюс	ролик для радио изготовление	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Бон Тон ООО ДП	рубероид	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Ивком ТД ООО	рубероид	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Центр моды ООО	спецодежда	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Полипром АОЗТ	трансп. КАМАЗ до 14 т, Газон до 3 т по У	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Карпенко Александр Владимирович Ч.	трансп. МАЗ, КАМАЗ, Шкода от 15-20 т, по	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Юнипресс ООО	упаковка	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Бон Тон ООО ДП	шлакокрашенная продукция	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Аптека 302 ДП ООО ОАХК	302 ДП ООО ОАХК аптека	м	<input type="checkbox"/>	<input type="checkbox"/>			м
21 век ООО	АВ Универсал ООО	м	<input type="checkbox"/>	<input type="checkbox"/>			м
Маг.Лада-Резерв ООО	автозапчасти	м	<input type="checkbox"/>	<input type="checkbox"/>			м

Рисунок 6 – Перечень товаров и услуг, которыми интересуются мужчины

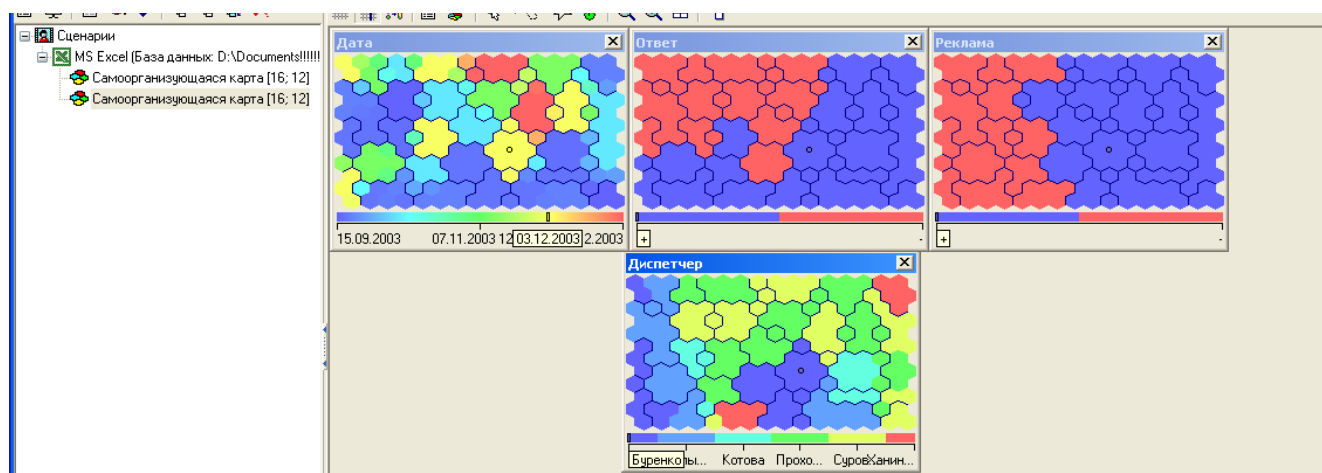


Рисунок 7 – Оценка эффективности работы диспетчеров с помощью самоорганизующейся карты Кохонена

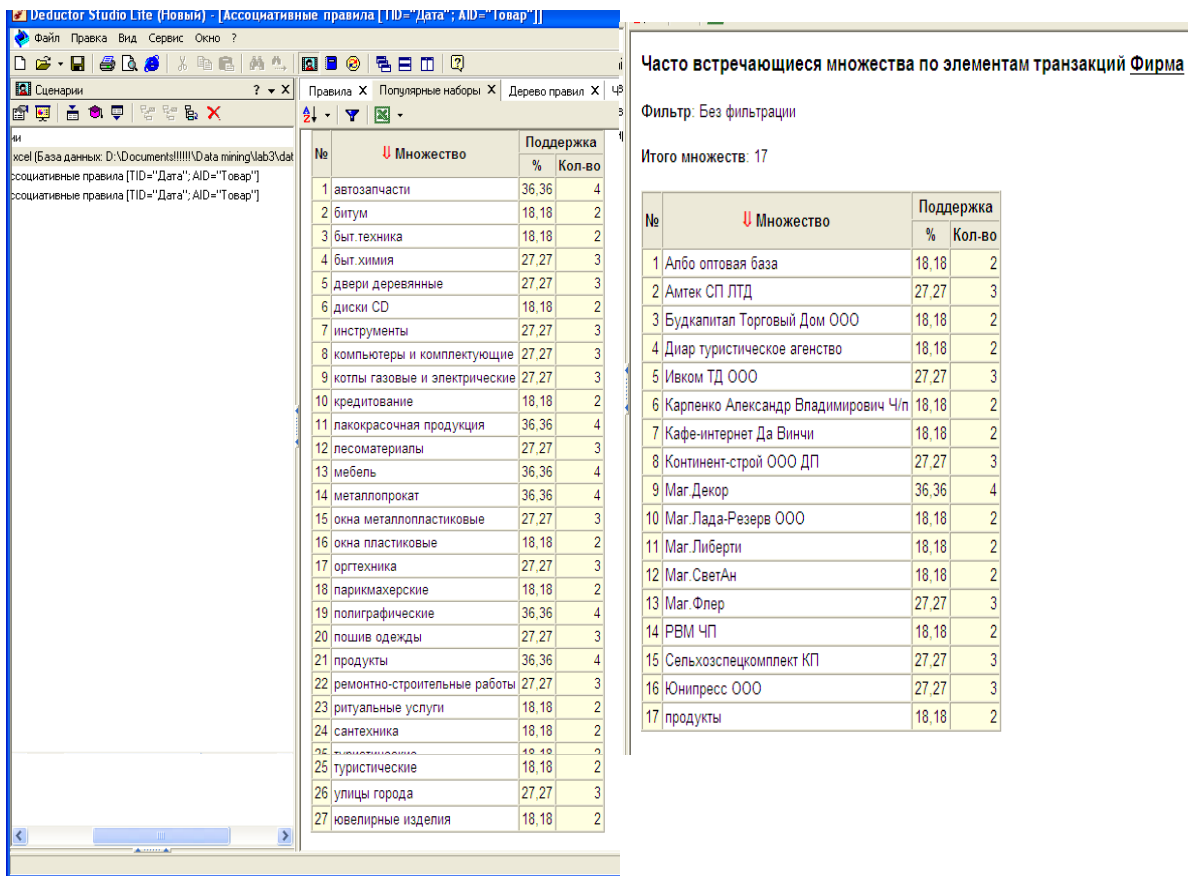


Рисунок 8 – Часто встречающиеся множества по элементам транзакций То-вар и Фирма

## ВОПРОСЫ ДЛЯ САМОПОДГОТОВКИ ПО ТЕОРЕТИЧЕСКОМУ МАТЕРИАЛУ

1. В каком году впервые появилось словосочетание Data Mining?
2. Кто является одним из основателей Data Mining?
3. На основе каких наук возникла и развивается технология Data Mining?
4. В результате чего могут быть получены данные?
5. Для чего потребители данных их используют?
6. Что может быть объектом данных и атрибутом объекта данных?
7. Как называется совокупность изучаемых объектов, интересующая исследователя?
8. Как называется числовые характеристики генеральной совокупности?
9. Как называется числовые характеристики выборки?
10. Назовите стадии Data Mining
11. Перечислите задачи Data Mining
12. Для решения каких задач используются методы ближайшего соседа, байесовских сетей, индукции деревьев решений, нейронных сетей, математической статистики
13. Для решения каких задач используется алгоритм Apriori?
14. Какие методы могут применяться для решения задач классификации, кластеризации и прогнозирования?

15. Что такое краткосрочный, среднесрочный и долгосрочный прогнозы?
16. Что можно отнести к сфере применения Data Mining для решения бизнес-задач, задач государственного уровня, web-задач?
17. Что можно отнести к сфере применения Data Mining для научных исследований?
18. Для решения каких задач промышленного производства, банковской сферы, сферы исследований для правительства может быть применен Data Mining?
19. Какое название носит применение технологии Data Mining для анализа информации, получаемой из сети Интернет?
20. Какая технология анализирует большие и сверхбольшие массивы неструктурированной информации?
21. Какая технология осуществляет автоматический поиск и извлечение информации из интернета?
22. Чем является внутренний узел дерева решений?
23. Чем является конечный узел дерева решений?
24. Что можно отнести к преимуществам деревьев решений?
25. Назовите этапы алгоритма конструирования деревьев решений.
26. Какие существуют правила остановки построения дерева решений?
27. Как называется критерий расщепления, основанный на энтропийном подходе?
28. Как называется критерий расщепления, основанный на расстояниях между распределениями классов и вероятностями наличия данного класса в данном узле?
29. Как называется отношение правильно классифицированных объектов к общему количеству объектов набора данных?
30. Как называется отношение неправильно классифицированных объектов к общему количеству объектов набора данных?
31. Для чего предназначен алгоритм CART?
32. Для чего предназначен алгоритм C4.5?
33. В основе какого метода лежит понятие плоскостей решений?
34. Что можно отнести к недостаткам метода опорных векторов?
35. Что можно отнести к достоинствам метода опорных векторов?
36. В основе какого метода лежит понятие «рассуждения по аналогии»?
37. Какой метод может быть использован для фильтрации получаемой электронной почты?
38. На каком принципе основано решение задач прогнозирования и классификации методом ближайшего соседа?
39. На каком предположении основан метод байесовской классификации?
40. Что можно отнести к преимуществам и недостаткам метода ближайшего соседа?
41. Что такое адаптивный сумматор?
42. Что такое нелинейный преобразователь?
43. Что такое функция активации?
44. Как называется вход (входная связь) нейрона?
45. Как называется выход (выходная связь) нейрона?
46. Как называется группа нейронов, на входы которых подается один и тот же общий сигнал?
47. Каков основной принцип работы сети Кохонена?
48. Когда был введен термин «кластерный анализ»?

49. Как называется схематическое изображение процесса иерархического метода кластерного анализа?
50. Как называется мера сходства, рассчитываемая как квадратный корень из суммы квадратов разностей координат объектов?
51. Как называется мера сходства, рассчитываемая как сумма квадратов разностей координат объектов?
52. Как называется мера сходства, рассчитываемая как среднее координатных разностей объектов?
53. Как называется мера сходства, учитывающая координатную разность объектов по одному измерению?
54. Как называется мера сходства, вычисляемая для категориальных признаков?
55. В чем состоит суть иерархических методов AGNES?
56. В чем состоит суть иерархических методов DIANA?
57. Каким путем осуществляется выбор начальных центроидов в алгоритме «k-means»?
58. Что можно отнести к проблемам кластерного анализа?
59. В чем суть алгоритмов PAM, BIRCH, WaveCluster, CLARA, Clarans?
60. Что называется поддержкой набора, поддержкой правила и достоверностью правила?
61. В чем суть алгоритмов AIS, SETM, Apriori?
62. Какие функциональные компоненты содержит система поддержки принятия решений?
63. Для чего предназначены сервер хранилища данных, инструментарий OLAP, инструментарий Data Mining?
64. Что такое MOLAP, ROLAP, HOLAP?
65. В чем суть подходов Cubing then mining, Mining then cubing, Cubing while mining?
66. Какие существуют виды грязных данных?
67. Если какие-то данные не были собраны, как называется такая проблема?
68. Если набор данных содержит записи с одинаковыми значениями всех атрибутов, как называется такая проблема?
69. Если набор данных содержит объекты или наблюдения, резко отличающиеся от основной массы, как называется такая проблема?
70. Что может быть решениями проблем пропущенных значений, дублирования данных, шумов и выбросов в наборе данных?
71. К какой группе относятся инструменты очистки данных, обрабатывающие данные с целью выявления всех возможных несоответствий и определения очищающих преобразований?
72. К какой группе относятся инструменты очистки данных, предназначенные для исключения дубликатов?
73. К какой группе относятся инструменты очистки данных, позволяющие пользователю определять функциональность очистки при помощи собственных API-функций?
74. Как называется распознавание подстрок в тексте свободного формата и назначение им соответствующих полей?
75. Как называется трансформация полей различных форматов в согласованный набор обозначений?
76. Как называется проверка введенных данных на вхождение в заданный диапазон значений?

77. Как называется введение в набор данных новых фактов, изначально в нем не содержащихся (например, новых полей)?
78. Как называется расстановка приоритетов между полями и контроль очередности сравнения полей?
79. В чем состоят области ответственности специалистов предметной области, в области баз данных, в области анализа данных?
80. В чем суть методологий SEMMA, CRISP-DM?
81. Как называется язык описания моделей?
82. Какие стандарты Data Mining базируются на языках XML, Java, SQL?