



Лекция 12 кластерный анализ. Аналитическая система Deductor.



КЛАСТЕРИЗАЦИЯ

- Кластеризация — инструмент исследования данных: **обнаружение классов образцов (например, новых подтипов болезни)**; проверка ожидаемого результата (например, что мышь и человек окажутся в разных кластерах).
- Используется в биологии, в анализе изображений, распознавании образов и т.д.
- **Пример.** Кластеризация клиентов страховой компании, застрахованных от несчастных случаев, выявила небольшой по объему кластер, в котором фигурировали одни и те же фамилии врачей; суммы страховых выплат тоже варьировались незначительно. Проверка показала, что в 90 % таких случаев имел место сговор с врачом.



КЛАСТЕРИЗАЦИЯ

- Разбиение объектов на кластеры, т.е. **группы схожих элементов, причем объекты разных кластеров существенно отличаются;**
- Кластеризация пациентов со схожей историей болезни, особенностями восстановления после болезни.
- Анализ спроса на медицинские услуги в зависимости от комбинации входных показателей.
- Обнаружение аномальных отклонений.



ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

Атрибут Возраст представлен следующими двадцатью значениями:

*{3, 56, 23, 39, **156**, 52, 41, 22, 9, 28, **139**, 31, 55, 20, **-67**, 37, 11, 55, 45, 37}*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 39.6$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 45.89,$$

$$T = \bar{x} \pm 2\sigma$$

Потенциальные аномалии: 156, 139 и -67 (ошибки ввода).



ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

- В основе метода лежит оценка мер расстояния между всеми наблюдениями в n -мерном пространстве данных
- Значение S_i множества данных S является аномальным, если хотя бы **часть значений p** из множества S **расположена на большем расстоянии, чем d** , от остальных значений.
- **Пример**

S - множество двумерных наблюдений, где требованием для аномальности является значение порогов $p \geq 4$ и $d \geq 3$.

$$S = \{S_1, S_2, S_3, S_4, S_5, S_6, S_7\} = \{(2, 4), (3, 2), (1, 1), (4, 3), (1, 6), (5, 3), (4, 2)\}.$$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ЗНАЧЕНИЙ

$p \geq 4$ и $d \geq 3$.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1		2,236	3,162	2,236	2,236	3,162	2,828
S_2			2,236	1,414	4,472	2,236	1,000
S_3				3,605	5,000	4,472	3,162
S_4					4,242	1,000	1,000
S_5						5,000	5,000
S_6							1,414

Значение	S_1	S_2	S_3	S_4	S_5	S_6
Параметр p	2	1	5	2	5	3

S_3 и S_5 - кандидаты в аномальные, для них значение $p = 5$ превышает заданный порог $p \geq 4$.



КЛАСТЕРИЗАЦИЯ

- В Data Mining распространенной мерой оценки близости между объектами является метрика или способ задания расстояния.
- Мера расстояния – мера различие/сходства между двумя объектами. Наиболее популярные метрики — *евклидово расстояние* и *расстояние Манхэттена*, корреляция.
- Алгоритм кластеризации – процедура минимизации расстояний внутри кластера и/или увеличения межкластерных различий.
Объект – то, что мы хотим кластеризовать (например, пациентов).
- Признаки – информация по каждому объекту, которую мы используем для кластеризации (например, клинические характеристики пациента).



МЕТРИКИ В КЛАСТЕРНОМ АНАЛИЗЕ

- ▣ **Меры, основанные на расстоянии:** Евклидово (Euclidian), Манхэттена и Канберра (Manhattan & Canberra), Махаланобиса (Mahalanobis)
- ▣ **Меры, основанные на корреляции:** Коэффициент корреляции Пирсона (Pearson product-moment correlation), Коэффициент ранговой корреляции Спирмена (Spearman rank correlation)
- ▣ **Информационно-теоретические:** Расстояние Хэмминга (Hamming distance) для категориальных данных



МЕТРИКИ В КЛАСТЕРНОМ АНАЛИЗЕ

Евклидово расстояние

По умолчанию R измеряет расстояния Евклида между всеми строками объектов. Для двух образцов X и Y с количеством в N i-х признаков:

$$d_{Euc}(X, Y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

Расстояние Манхэттена

Расстояние Манхэттена (или «расстояние городских кварталов») похоже на перемещение по координатной сетке:

$$d_{Manh}(X, Y) = \sum_{i=1}^N |x_i - y_i|$$

Расстояние Канберра

Расстояние Канберра – это расстояние Манхэттена, масштабированное по величинам X и Y (используется не часто, не входит в пакет *bioDist*):

$$d_{Cunb}(X, Y) = \sum_{i=1}^N \frac{|x_i - y_i|}{|x_i| + |y_i|}$$



МЕТРИКИ В КЛАСТЕРНОМ АНАЛИЗЕ

Расстояние Минковского

Расстояние Минковского – это p -мерное расстояние, обобщенная метрика расстояний:

$$d_p(X, Y) = \sqrt[p]{\sum_{i=1}^N (x_i - y_i)^p}$$

$p = 1$ ---> расстояние Манхэттена;

$p = 2$ ---> расстояние Евклида;

$\lim_{p \rightarrow \infty}$ ---> расстояние максимумов:

$$d_p(X, Y) = \max_{1 \leq i \leq N} |x_i - y_i|$$

Расстояние Махаланобиса

Расстояние Махаланобиса берет в расчет ковариационную структуру данных:

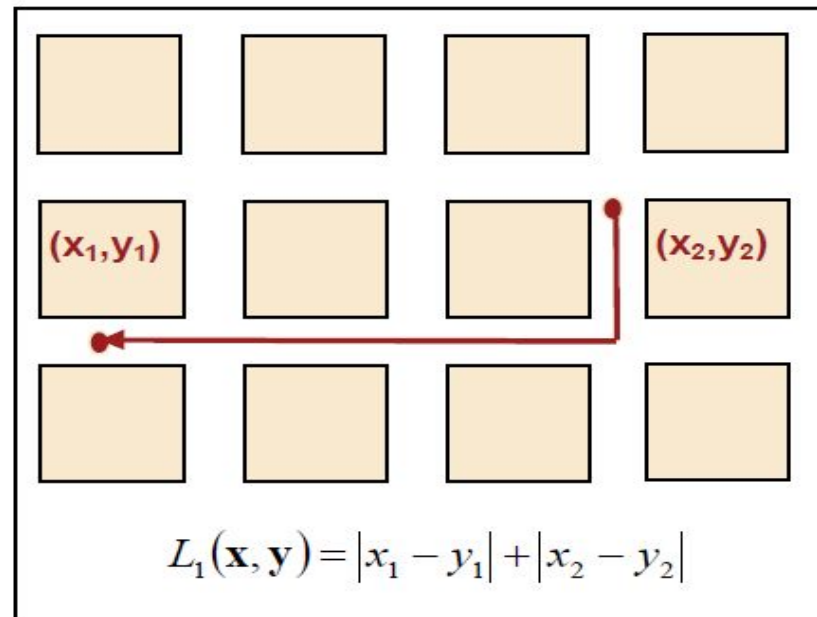
$$d_{Maha}(X, Y) = \sqrt{(X - Y)^T S^{-1} (X - Y)}$$

где S – это матрица ковариаций для i -х координат.

Расстояние Хэмминга (Hamming distance) — число позиций, в которых различаются соответствующие символы двух строк одинаковой длины.

МЕТРИКИ В КЛАСТЕРНОМ АНАЛИЗЕ

- Множество точек, равноудаленных от некоторого центра при использовании евклидовой метрики будет образовывать сферу (или круг в двумерном случае), и кластеры, полученные с использованием **евклидова расстояния**, также будут иметь форму, близкую к **сферической**.
- Преимущество метрики *Расстояния Манхэттена* заключается в том, что ее использование позволяет снизить влияние аномальных значений на работу алгоритмов. Кластеры, построенные на основе расстояния Манхэттена, стремятся к **кубической форме**.

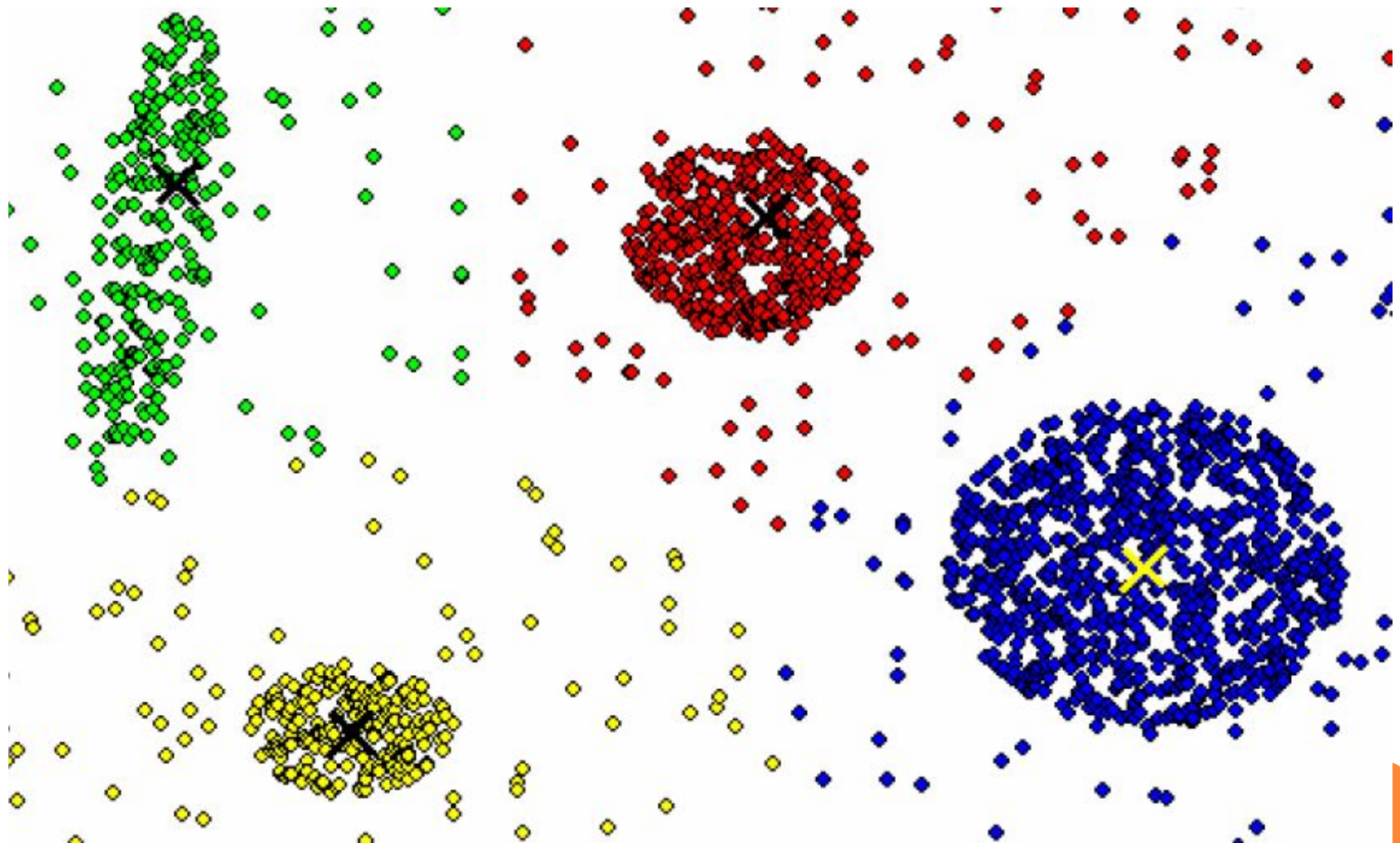


АЛГОРИТМЫ В КЛАСТЕРНОМ АНАЛИЗЕ

- Наибольшее распространение в популярных статистических пакетах получили две группы алгоритмов кластерного анализа: **иерархические методы и итеративные методы группировки** (алгоритм *k-means* и сети *Кохонена*).
- Кластеризация не приносит каких-либо результатов без содержательной интерпретации каждого кластера.
- Интерпретация предполагает присвоение каждому кластеру емкого названия, отражающего его суть, например «Пожилые женщины с факторами риска», «Пациенты, ведущие неактивный образ жизни» и т. д.
- Для интерпретации аналитик детально исследует каждый кластер: **его статистические характеристики, распределение значений признаков объекта в кластере, оценивает мощность кластера — число объектов, попавших в него.**
- Интерпретация значительно облегчается, если имеются способы представления результатов кластеризации в специализированном виде: **дендограммы, кластерограммы, карты.**



Алгоритмы в кластерном анализе

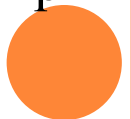


АЛГОРИТМЫ В КЛАСТЕРНОМ АНАЛИЗЕ

Одним из наиболее простых и эффективных алгоритмов кластеризации является алгоритм *K-means*

Он состоит из четырех шагов.

1. Задается число кластеров k , которое должно быть сформировано из объектов исходной выборки.
2. Случайным образом выбирается k записей исходной выборки, которые будут служить начальными центрами кластеров. Такие начальные точки, из которых потом «вырастает» кластер, часто называют «семенами» (от англ. *seeds* – семена, посевы). Каждая такая запись будет представлять собой своего рода «эмбрион» кластера, состоящий только из одного элемента.
3. Для каждой записи исходной выборки определяется ближайший к ней центр кластера.
4. Производится вычисление *центроидов* – центров тяжести кластеров. Это делается путем определения среднего для значений каждого признака для всех записей в кластере. Затем старый центр кластера смещается в его центроид.



K-MEANS

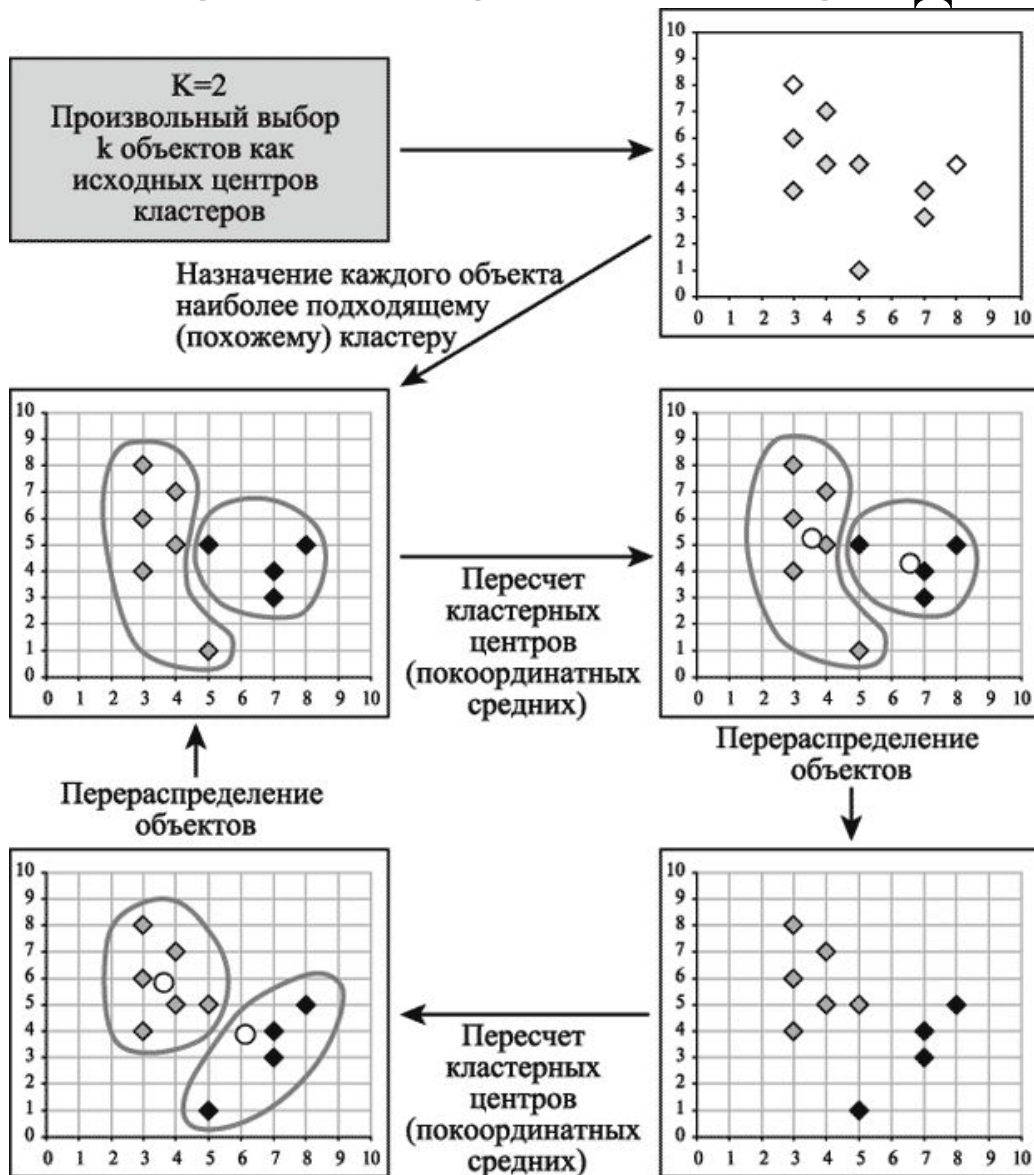
Центроиды становятся новыми центрами кластеров для следующей итерации алгоритма.

- Шаги 3 и 4 повторяются до тех пор, пока выполнение алгоритма не будет прервано либо не будет выполнено условие в соответствии с некоторым критерием сходимости.
- Остановка алгоритма производится тогда, когда границы кластеров и расположения центроидов не перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор записей. На практике алгоритм *k-means* обычно находит набор стабильных кластеров за несколько десятков итераций.
- В качестве критерия сходимости чаще всего используется критерий суммы квадратов ошибок между центроидом кластера и всеми вошедшими в него записями.

$$E = \sum_{i=1}^k \sum_{p \in C_i} (p - m_i)^2,$$

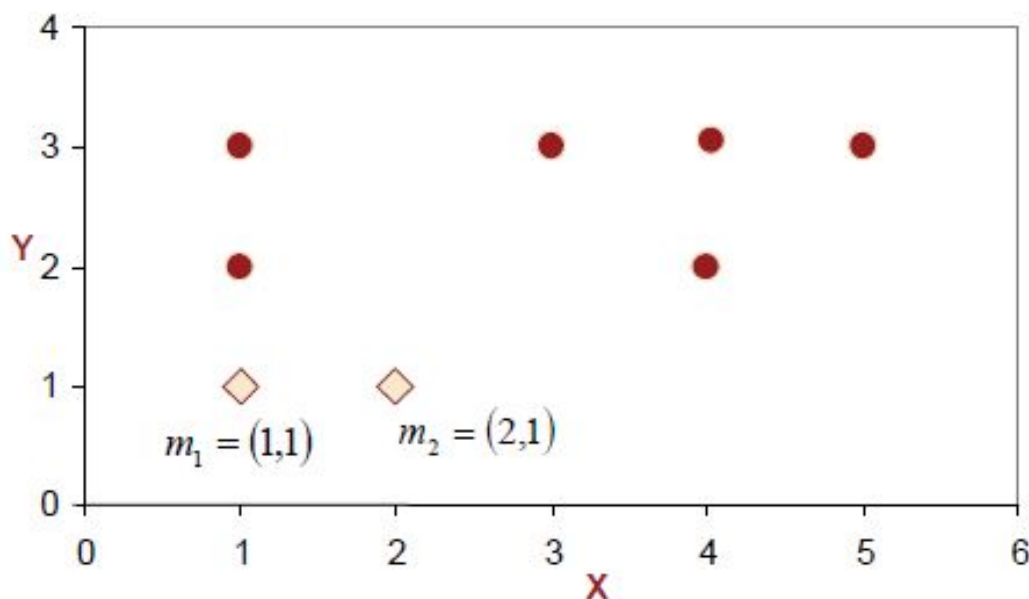
где $p \in C_i$ – произвольная точка данных, принадлежащая кластеру C_i , m_i – центроид данного кластера. Иными словами, алгоритм остановится тогда, когда ошибка E достигнет достаточно малого значения.

ПРИМЕР РАБОТЫ АЛГОРИТМА К-СРЕДНИХ (K=2)



ПРИМЕР

A	B	C	D	E	F	G	H
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)



Шаг 1. Определим число кластеров, на которое требуется разбить исходное множество $k=2$.

Шаг 2. Случайным образом выберем две точки, которые будут начальными центрами кластеров. Пусть это будут точки $m_1=(1;1)$ и $m_2=(2;1)$.

Шаг 3, проход 1. Для каждой точки определим к ней ближайший центр кластера с помощью расстояния Евклида.

ПРИМЕР

Точка	Расстояние от m_1	Расстояние от m_2	Принадлежит кластеру
A	2,00	2,24	1
B	2,83	2,24	2
C	3,61	2,83	2
D	4,47	3,61	2
E	1,00	1,41	1
F	3,16	2,24	2
G	0,00	1,00	1
H	1,00	0,00	2

Таким образом, кластер 1 содержит точки A, E, G, а кластер 2 – точки B, C, D, F, H. Как только определятся члены кластеров, может быть рассчитана сумма квадратичных ошибок:

$$E = \sum_{i=1}^k \sum_{p \in C_i} (p - m_i)^2 = 2^2 + 2,24^2 + 2,83^2 + 3,61^2 + 1^2 + 2,24^2 + 0^2 + 0^2 = 36.$$

Шаг 4, проход 1. Для каждого кластера вычисляется его центроид, и центр кластера перемещается в него. Центроид для первого кластера вычисляется как

$$[(1+1+1)/3, (3+2+1)/3] = (1; 2).$$

Центроид для кластера 2 будет равен

$$[(3 + 4 + 5 + 4 + 2)/5, (3 + 3 + 3 + 2 + 1)/5] = (3,6; 2,4).$$

Вывод.

При помощи **метода к-среднего** реализуется процедура построения **усредненных профилей** каждого класса, что дает возможность проводить качественный **анализ выраженности признаков** у представителей каждого класса.



ПРИМЕР

Симптомы поздней депрессии, сгруппированные методом кластерного анализа

Кластер №	Симптомы (в порядке убывания межгрупповых расстояний)
1	Астения психическая и физическая, пониженная самооценка; моторная заторможенность; снижение массы тела; сенестопатии, ипохондрия, снижение аппетита, адинамия, пессимизм; симпатикотония, тревога постоянная, гипотимия, идеи самоуничтожения; печаль; соматические жалобы, бессонница; снижение продуктивности деятельности
2	Фобии; снижение или отсутствие критики; раздражительность; ажитация; плаксивость; эмоциональная лабильность; чувство печали; тревога пароксизмальная; вегетативные кризы («панические атаки»); ускорение мышления; конверсионная симптоматика; импульсивность
3	Апатия; замедление мышления; суточная циркадность; витальная тоска; суицидальные тенденции; идеи виновности; тоска; ангедония; соматическая деперсонализация; гипобулия; снижение либидо; потеря способности к самообслуживанию в быту
4	Психическая деперсонализация; обсессии; вязкость мышления; идеи ущерба, обнищания; психопатоподобное поведение; ухудшение внимания; амнезия; субъективное и объективное снижение интеллекта; иллюзии; галлюцинации; дисфория; моторное возбуждение; нецеленаправленность действий; монотонность невербальная и стереотипии речи; непоследовательность и обеднение содержания мышления; обманы памяти

ПРИМЕР

Карта оценки психического статуса при депрессии пожилого возраста: список оцениваемых симптомов и их определение

Симптом	Определение симптома
Гипотимия	Снижение настроения
Тоска	Беспричинная тяжелая подавленность, уныние, безнадежность
«Витальная» тоска	Ощущение психического страдания (тоска) в форме физического гнета в загрудинной области
Ажитация	Тревога с психомоторным возбуждением
Апатия	Вялость, безразличие, отсутствие эмоционального отклика, апатический ступор
Печаль	Грусть, подавленность, мрачная окраска переживаний
Ангедония	Потеря чувства удовольствия, «радости»
Тревога постоянная	Беспричинное волнение, предчувствие опасности, внутреннее напряжение, боязливое ожидание, беспредметное беспокойство
Тревога пароксизмальная	То же в форме пароксизмов
Эмоциональная лабильность	Повышение силы эмоционального отклика и снижение порога для его возникновения
Дисфория	Угрюмость, ожесточенность, сварливость, претензии к окружающим
Плаксивость	Неспособность сдерживать слезы
Раздражительность	Негативная словесная или поведенческая реакция в ответ на незначительное событие
Замедление мышления	От замедления речи до мутизма



K-MEANS

Номера объектов первого класса

	1	3	2
Distance	1.484488	1.097134	0.693889

Номера объектов второго класса

	8	4	9	7	6	5
Distance	1.357421	1.566430	0.535758	0.855267	1.272938	0.822147

	Межкл. расст.	Внутрикл. расст.	Степ.свободы	F	p
D	9.38	17.50	7	3.75	0.093821
RESP	3.55	19.33	7	1.28	0.293890
LABACT	37.55	4.00	7	65.72	0.000084
WTG	18.00	6.00	7	21.00	0.002536
GOAL	22.22	12.00	7	12.96	0.008735
MOTIV	0.055	10.16	7	0.038	0.850495

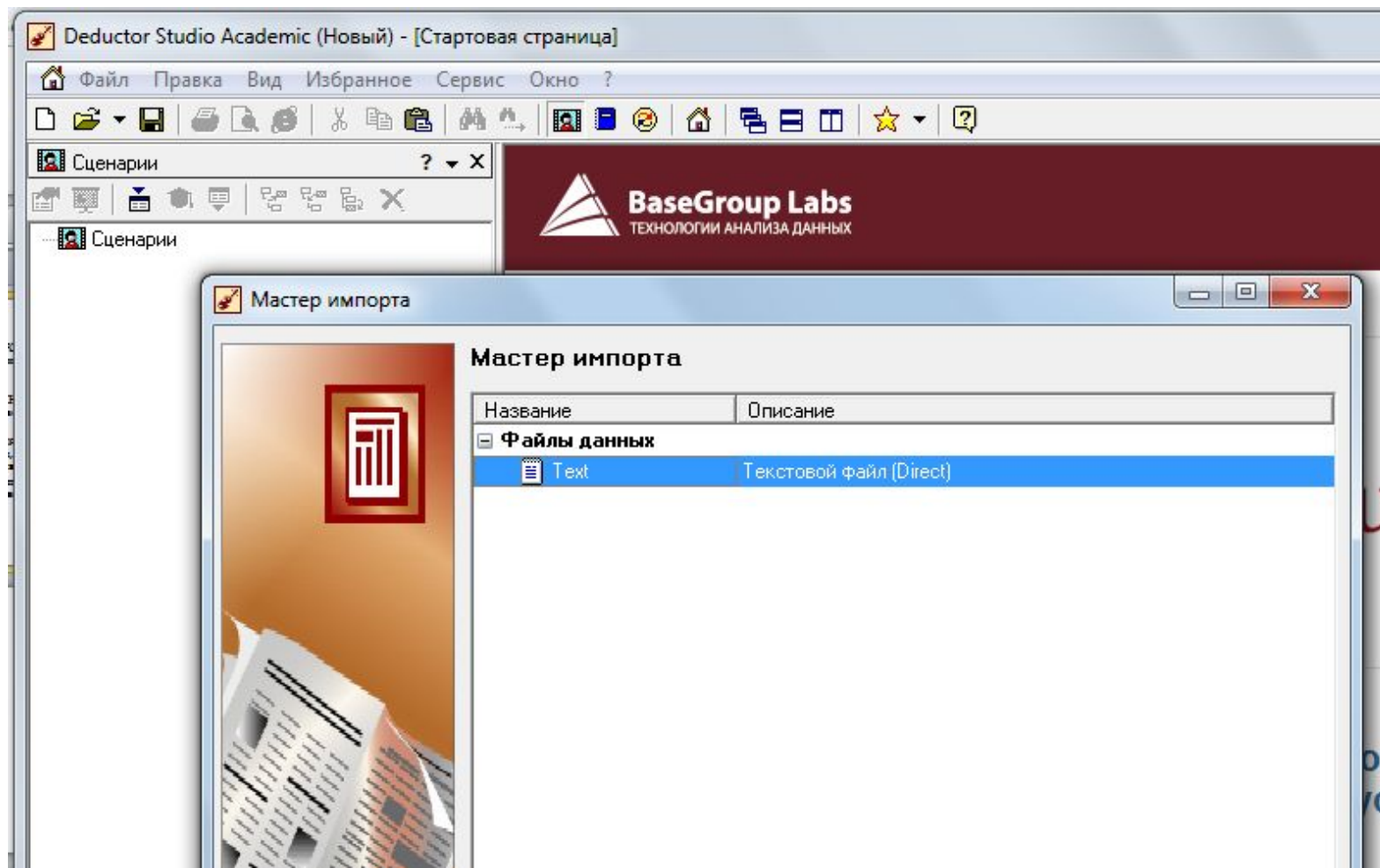


ПЛАТФОРМА DEDUCTOR

Модуль	Academic (руб.)	Professional (руб.)	Enterprise (руб.)
Studio	бесплатно	29 000 	<u>по запросу</u>
Viewer	-	8 700 	<u>по запросу</u>
Warehouse	бесплатно	бесплатно	бесплатно
Server	-	-	<u>по запросу</u>
Client	-	-	бесплатно



ПЛАТФОРМА DEDUCTOR



ПЛАТФОРМА DEDUCTOR

Deductor Studio Academic (Новый) - [Стартовая страница]

Файл Правка Вид Избранное Сервис Окно ?

Сценарии

BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

Мастер импорта - Text (2 из 9)

Импорт текстового файла
Укажите имя текстового файла для импорта

Имя файла: C:\Users\Andrey\Desktop\111.txt

Кодировка: ANSI (Windows) Первая строка является заголовком

Просмотр выбранного файла Начать импорт со строки: 1

N	Пол	Возраст	Перебои	Сердцебиение	Агв	Агн	ЧС	ПЖ	
47	1	70	0	0	130	85	68	2,2	3,2
48	0	46	0	0	130	90	78	2,4	3,2
49	1	61	1	0	130	80	57	2,4	3,2
50	0	27	0	1	115	70	59	2,5	3,2
51	0	38	0	1	120	70	65	2,0	2,8
52	1	31	0	1	100	80	70	2,2	3,6
53	1	46	0	0	120	70	78	2,4	3,2
54	0	44	1	1	110	80	90	2,0	3,1
55	1	42	0	1	110	70	92	1,9	2,8
56	0	41	0	0	130	70	79	2,7	2,8
57	0	21	0	0	120	80	90	1,7	3,0
58	0	22	0	0	120	80	90	1,8	3,4



ПЛАТФОРМА DEDUCTOR


The screenshot displays the 'Deductor Studio Academic' application window. The title bar reads 'Deductor Studio Academic (Новый) - [Стартовая страница]'. The menu bar includes 'Файл', 'Правка', 'Вид', 'Избранное', 'Сервис', and 'Окно'. The active window is 'Мастер импорта - Text (6 из 9)'. The main area is titled 'Импорт текстового файла' with the instruction 'Укажите параметры столбцов'. A list of columns is shown on the left, including '9.0 N', '0/1 Пол', '9.0 Возраст', '0/1 Перебои', '0/1 Сердцебиение', and various alphanumeric codes like '9.0 Adv', '9.0 Адн', '9.0 ЧС', '9.0 ПЖ', '9.0 ЛП', '9.0 А01', '9.0 А02', '9.0 КДО', '9.0 КСО', and '9.0 ФИ'. The right side of the dialog has fields for 'Имя столбца', 'Метка столбца', and 'Назначение'. The 'Тип данных' dropdown is open, showing options: '0/1 Логический' (selected), '7 Дата/Время', '9.0 Вещественный', '12 Целый', and 'ab Строковый'. A secondary dropdown at the bottom right shows categories like 'Информационное', 'Измерение', 'Атрибут', 'Факт', and 'Транзакция'.



ПЛАТФОРМА DEDUCTOR


Мастер импорта - Text (8 из 9)

Определение способов отображения
Выберите способ отображения данных



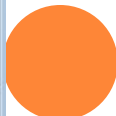
Название	Описание
<input checked="" type="checkbox"/> Табличные данные	
<input checked="" type="checkbox"/> Таблица	Отображает данные в виде таблицы
<input checked="" type="checkbox"/> Статистика	Отображает статистические данные выборки
<input checked="" type="checkbox"/> Графики	
<input checked="" type="checkbox"/> Диаграмма	Отображает данные в виде диаграммы
<input type="checkbox"/> Гистограмма	Отображает данные в виде гистограммы
<input type="checkbox"/> Многомерная диаграмма	Отображает данные в виде многомерной диаграммы
<input type="checkbox"/> Диаграмма размещения	Диаграмма размещения объектов в пространстве
<input type="checkbox"/> OLAP анализ	
<input type="checkbox"/> Куб	Многомерное отображение (кросс-таблица и кросс-диаграмма)
<input type="checkbox"/> Общие	
<input type="checkbox"/> Сведения	Сведения о параметрах

Название процесса
Успешное завершено

Процент выполнения


Время выполнения

< Назад Далее > Отмена




ПЛАТФОРМА DEDUCTOR

Мастер импорта - Text (9 из 10)

Настройка столбцов диаграммы

Настройка различных параметров столбцов диаграммы

Метка столбца	Тип данных	Цвет
<input checked="" type="checkbox"/> N	9.0 Вещественный	■
<input checked="" type="checkbox"/> Возраст	9.0 Вещественный	■
<input checked="" type="checkbox"/> Адв	9.0 Вещественный	■
<input type="checkbox"/> Адн	9.0 Вещественный	■
<input type="checkbox"/> ЧС	9.0 Вещественный	■
<input type="checkbox"/> ПЖ	9.0 Вещественный	■
<input checked="" type="checkbox"/> ЛП	9.0 Вещественный	■
<input type="checkbox"/> АО1	9.0 Вещественный	■
<input type="checkbox"/> АО2	9.0 Вещественный	■
<input type="checkbox"/> КДО	9.0 Вещественный	■
<input type="checkbox"/> КСО	9.0 Вещественный	■
<input type="checkbox"/> ФИ	9.0 Вещественный	■

Тип:  Точки

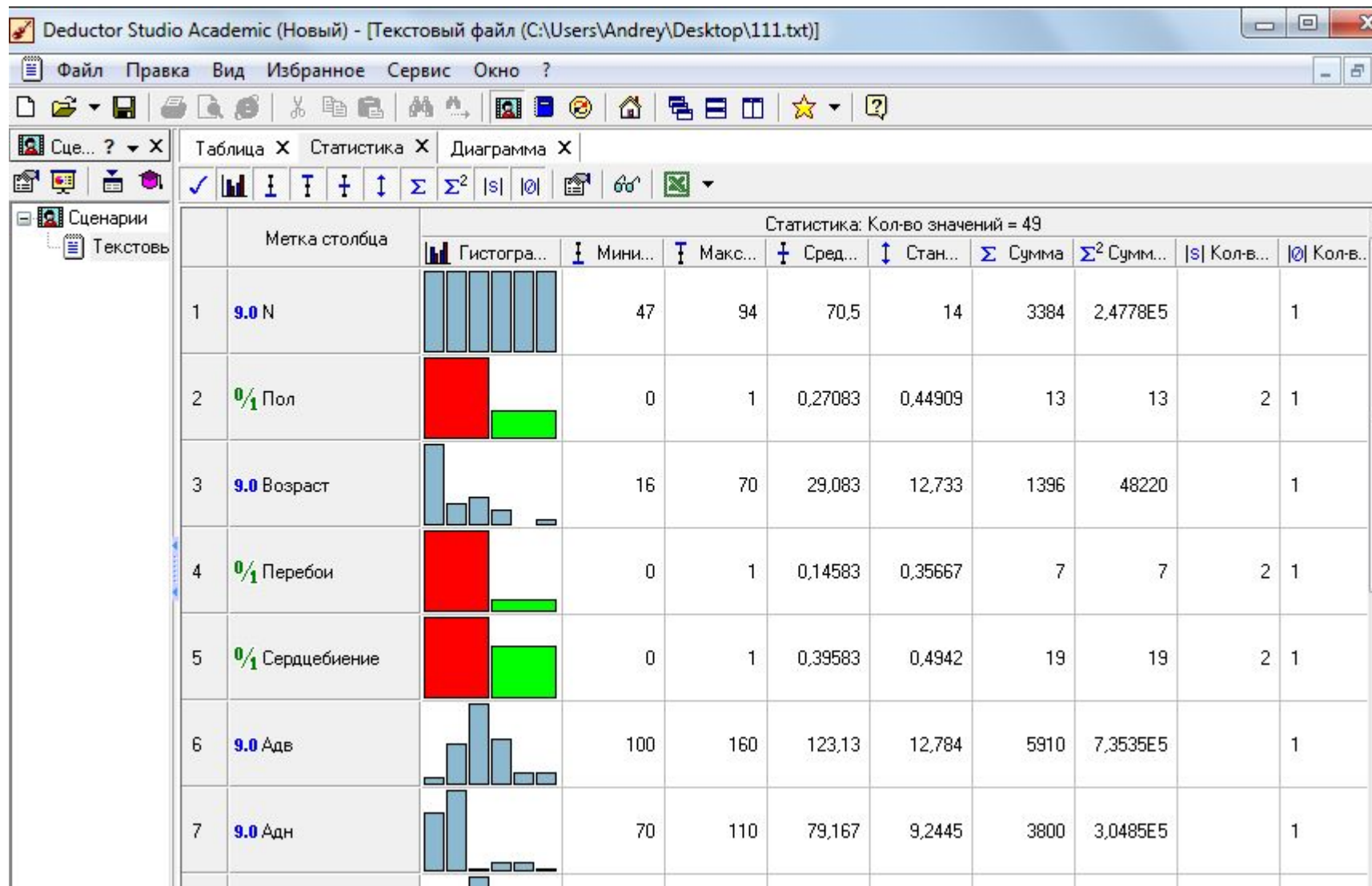
Подписи по X: КДО

Значения по X: Возраст

< Назад Далее > Отмена



ПЛАТФОРМА DEDUCTOR



ПЛАТФОРМА DEDUCTOR

Deductor Studio Academic (Новый) - [Текстовый файл (C:\Users\Andrey\Desktop\111.txt)]

Файл Правка Вид Избранное Сервис Окно ?

Сценарии ? X

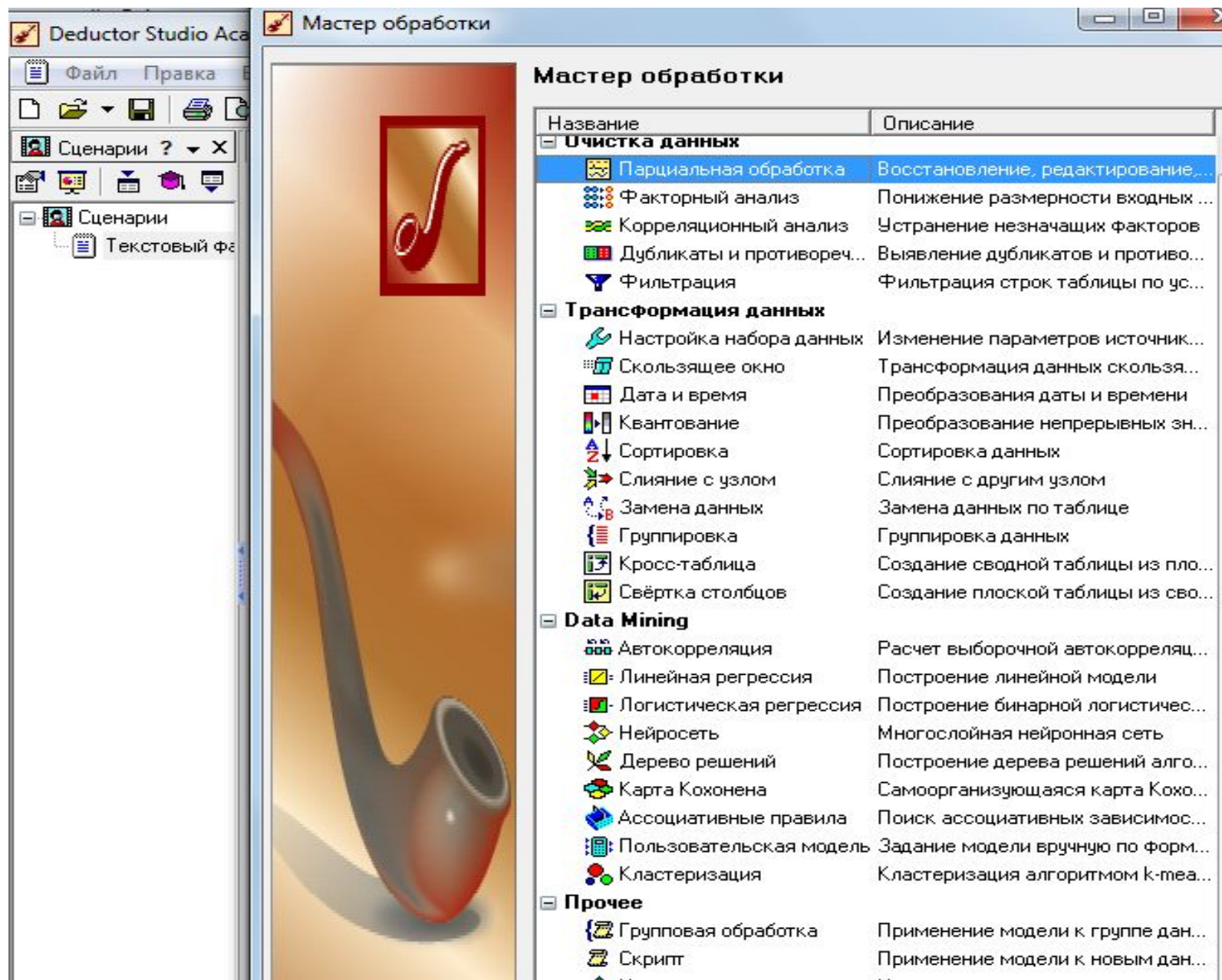
Таблица X Статистика X Диаграмма X

1 / 49

N	Пол	Возраст	Перебои	Сердцебиение	Адв	ЧС	ПЖ	A01	A02	КДО	КСО	ФИ
47	<input checked="" type="checkbox"/>	70	<input type="checkbox"/>	<input type="checkbox"/>	130	68	2.2	3.8	2.1	124	40	60
48	<input type="checkbox"/>	46	<input type="checkbox"/>	<input type="checkbox"/>	130	78	2.4	3	2.1	124	40	69
49	<input checked="" type="checkbox"/>	61	<input checked="" type="checkbox"/>	<input type="checkbox"/>	130	57	2.4	3	2.1	124	40	69
50	<input type="checkbox"/>	27	<input type="checkbox"/>	<input checked="" type="checkbox"/>	115	59	2.5	3.2	2.1	111	36	68
51	<input type="checkbox"/>	38	<input type="checkbox"/>	<input checked="" type="checkbox"/>	120	65	2	2.6	1.9	142	51	64
52	<input checked="" type="checkbox"/>	31	<input type="checkbox"/>	<input checked="" type="checkbox"/>	100	70	2.2	2.5	1.9	130	50	62
53	<input checked="" type="checkbox"/>	46	<input type="checkbox"/>	<input type="checkbox"/>	120	78	2.4	3	2.1	124	40	69
54	<input type="checkbox"/>	44	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	110	90	2	3	1.9	100	27	72
55	<input checked="" type="checkbox"/>	42	<input type="checkbox"/>	<input checked="" type="checkbox"/>	110	92	1.9	2.8	1.5	108	31	69
56	<input type="checkbox"/>	41	<input type="checkbox"/>	<input type="checkbox"/>	130	79	2.7	2.8	1.8	109	34	69
57	<input type="checkbox"/>	21	<input type="checkbox"/>	<input type="checkbox"/>	120	90	1.7	2	2.1	111	32	73
58	<input type="checkbox"/>	22	<input type="checkbox"/>	<input type="checkbox"/>	120	90	1.8	3.5	2	113	44	66
59	<input checked="" type="checkbox"/>	17	<input type="checkbox"/>	<input type="checkbox"/>	100	70	2.4	3	2.1	124	40	69
60	<input type="checkbox"/>	40	<input type="checkbox"/>	<input type="checkbox"/>	120	70	2.4	3	2.1	124	40	69
61	<input type="checkbox"/>	19	<input type="checkbox"/>	<input type="checkbox"/>	120	74	2.4	2.7	1.9	138	51	63
62	<input type="checkbox"/>	27	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	120	78	2.4	3.5	2.2	107	34	67
63	<input type="checkbox"/>	20	<input type="checkbox"/>	<input checked="" type="checkbox"/>	120	49	2.2	2.9	2.1	114	28	75



ПЛАТФОРМА DEDUCTOR



The screenshot displays the 'Deductor Studio' application window with the 'Master of Processing' (Мастер обработки) dialog box open. The dialog box features a large, stylized pipe icon in the background. The main content is a table listing various data processing tasks, organized into categories. The table has two columns: 'Name' (Название) and 'Description' (Описание).

Название	Описание
Учистка данных	
Парциальная обработка	Восстановление, редактирование...
Факторный анализ	Понижение размерности входных ...
Корреляционный анализ	Устранение незначущих факторов
Дубликаты и противореч...	Выявление дубликатов и противо...
Фильтрация	Фильтрация строк таблицы по ус...
Трансформация данных	
Настройка набора данных	Изменение параметров источник...
Скользящее окно	Трансформация данных скользя...
Дата и время	Преобразования даты и времени
Квантование	Преобразование непрерывных зн...
Сортировка	Сортировка данных
Слияние с узлом	Слияние с другим узлом
Замена данных	Замена данных по таблице
Группировка	Группировка данных
Кросс-таблица	Создание сводной таблицы из пло...
Свёртка столбцов	Создание плоской таблицы из сво...
Data Mining	
Автокорреляция	Расчет выборочной автокорреляц...
Линейная регрессия	Построение линейной модели
Логистическая регрессия	Построение бинарной логистичес...
Нейросеть	Многослойная нейронная сеть
Дерево решений	Построение дерева решений алго...
Карта Кохонена	Самоорганизующаяся карта Кохо...
Ассоциативные правила	Поиск ассоциативных зависимос...
Пользовательская модель	Задание модели вручную по форм...
Кластеризация	Кластеризация алгоритмом k-mea...
Прочее	
Групповая обработка	Применение модели к группе дан...
Скрипт	Применение модели к новым дан...



СЕТИ КОХОНЕНА

- Термин «сети Кохонена» (англ.: *Kohonen network*, *KCN*) был введен в 1982 финским ученым Тойв Кохоненом.
- Изначально они применялись для обработки изображений и звука, но сети Кохонена также являются эффективным средством кластерного анализа.
- Кохонен внес большой вклад в развитие теории искусственных нейронных сетей, ассоциативной памяти и распознавании образов.
- Основной целью KCN является преобразование сложных многомерных данных в более простую структуру малой размерности.
- Самоорганизующаяся карта Кохонена является разновидностью нейронной сети.
- Применяется, когда необходимо решить задачу кластеризации, т.е. распределить данные по нескольким кластерам.
- Механизм же построения карты Кохонена позволяет отобразить многомерное пространство в двумерном, которое более удобно и для визуализации и для интерпретации результатов аналитиком.



СПАСИБО ЗА ВНИМАНИЕ.

