Министерство науки и высшего образования Российской Федерации
Севастопольский государственный университет
Кафедра ИС

Расчетно-графическая работа
«Компьютерные методы анализа данных и прогнозирования»
по дисциплине
«Интеллектуальный анализ данных»

Выполнил студент группы ИС/б-17-2-о
Горбенко К. Н.
Проверила
Сырых О.А.

Севастополь
2020

1 ЦЕЛЬ РАБОТЫ

- Приобрести базовые навыки работы в Deductor Studio;
- Изучить основы методов анализа экспериментальных данных и освоить технику их практического применения в Deductor Studio.

2 ПОСТАНОВКА ЗАДАЧИ

1. Скачать и установить Deductor Studio Academic.
2. Создать проект, заполнить его свойства, просмотреть файл проекта через любой текстовый редактор.
3. Создать текстовый файл с данными, импортировать его в Deductor, настроить метки к столбцам, экспортировать файл, присоединить предыдущую ветвь к новому узлу импорта.
4. Настроить следующие визуализаторы: «Таблица», «Статистика».
5. Ответить на контрольные вопросы.
6. Подобрать данные и произвести поиск ассоциативных правил и прогнозирование временного ряда.

3 ИЗУЧЕНИЕ СИСТЕМЫ DEDUCTOR

3.1 Создание проекта

Создадим проект в Deductor. Изменим его свойства и откроем в любом текстовом редакторе. Сделаем видимой вкладку «Подключения». Результат представлен на рисунке 1.

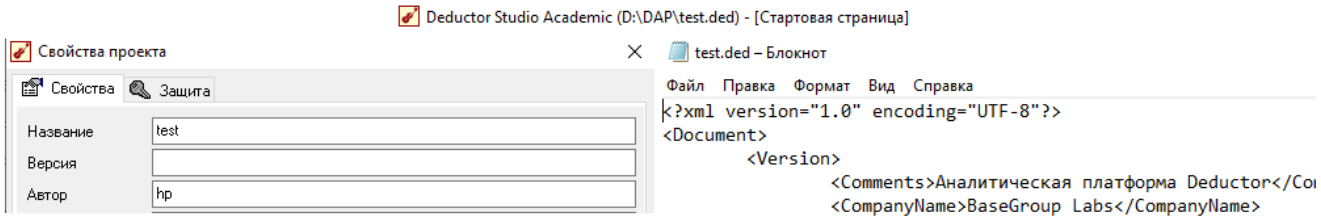


Рисунок 1 – Создание проекта в Deductor

3.2 Импорт и экспорт текстового файла

Создадим текстовый файл и импортируем его в Deductor. После исправления в программе экспортируем результат в текстовый файл и отобразим. Результат представлен на рисунке 2.

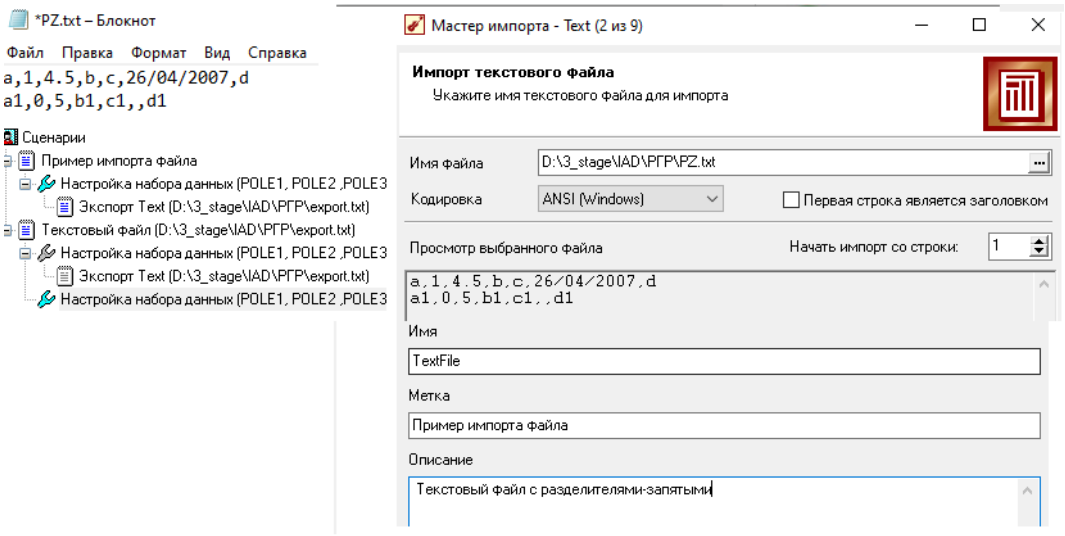


Рисунок 2 – Импорт и экспорт в Deductor

3.3 Настройка визуализаторов

В проекте из предыдущего задания настроим визуализаторы. В визуализаторе таблицы настроим, чтобы при отображении поля 3 добавлялась единица измерения. Результат представлен на рисунке 3.

Рисунок 3 – Визуализаторы в Deductor

3.4 Ответы на контрольные вопросы

1. Дедуктор состоит из Warehouse, Studio, Viewer, Server, Client.

2. Чтобы скрыть столбец из набора данных, нужно задать ему назначение «Неиспользуемое».

3. Категории пользователей: аналитик, пользователь, администратор, программист.

4. Функции аналитика: создание в Deductor Studio сценариев — последовательности шагов, которую необходимо провести для получения нужного результата, построение, оценка и интерпретация моделей, настройка панели отчетов для пользователей Deductor Viewer, настройка сценария на поточную обработку новых данных.

5. В deductor studio ключевым понятием является проект. Это файл с расширением *.ded, по структуре соответствующий стандартному xml-файлу. Он хранит в себе последовательности обработки данных (сценарии), настроенные визуализаторы, переменные проекта и служебную информацию.

3.5 Выводы

В ходе практической работы изучены основные возможности и компоненты программы Deductor Studio. Изучены основные функции аналитиика, пользователя, администратора, программиста в программе Deductor Studio.

4 ПОИСК АССОЦИАТИВНЫХ ПРАВИЛ

4.1 Подготовка данных

В качестве исходных данных для анализа возьмем 30 чеков покупателей:

```
moloko  chek 1
smetana chek 1
kolbasa chek 1
xleb    chek 1
sir     chek 1
```

Рисунок 4 – Исходные данные

4.2 Составление таблицы

Сформируем текущие данные в таблицу, состоящую из двух столбцов: «Продукт» и «Номер чека» и укажем идентификатор и элемент транзакции (рисунок 5).

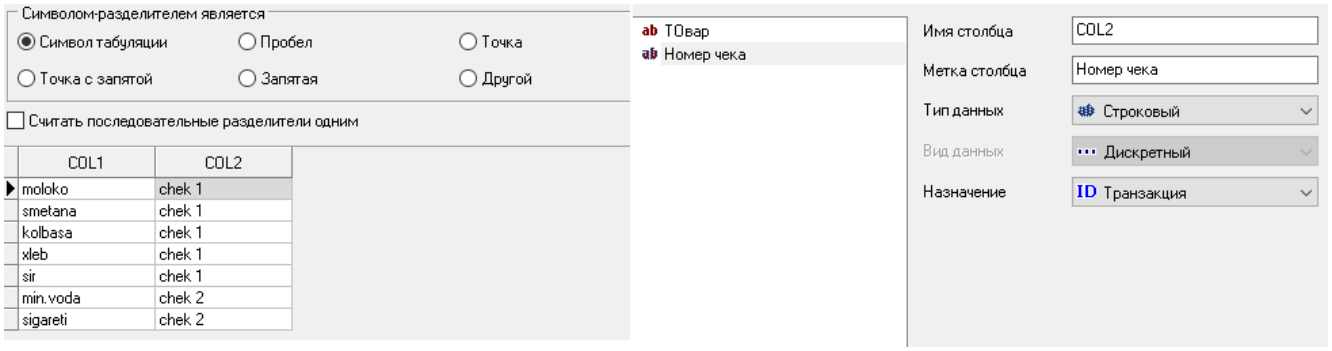


Рисунок 5 – Определение столбца идентификатора транзакции и ее элемента

4.3 Поиск ассоциативных правил

Для поиска ассоциативных правил воспользуемся Мастером обработки, где выберем тип обработки «Ассоциативные правила».

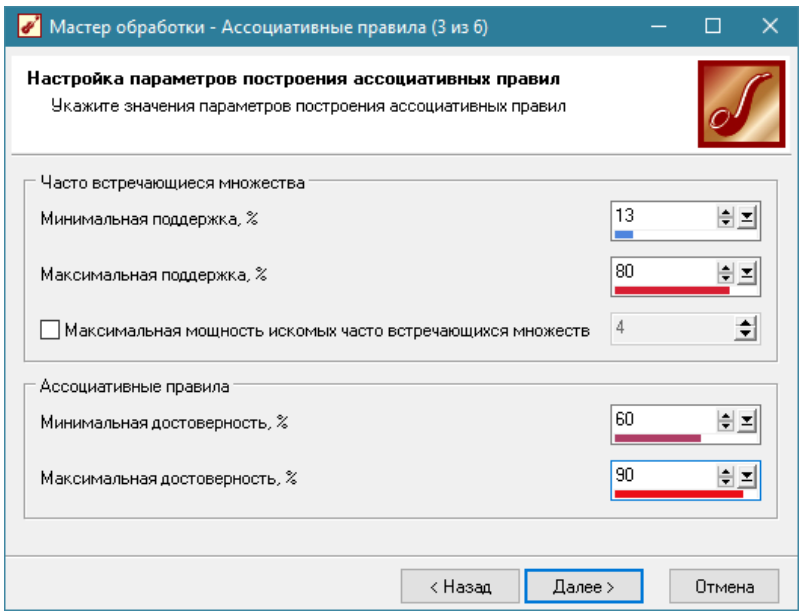


Рисунок 6 – Настройка параметров построения правил

Последующие действия позволяет запустить процесс поиска ассоциативных правил. Результат изображен на рисунке 7.

Рисунок 7 – Результат поиска ассоциативных правил

4.4 Определение популярных наборов

Популярные наборы (рисунок 8):

№	Номер множества	ab. Элементы	Поддержка		Мощность
			Кол-во	%	
95	54	sodovaya vipechka	4	13,33	2
96	55	sodovaya wino	4	13,33	2
97	56	sodovaya xleb	9	30,00	2

Рисунок 8 – Популярные наборы

Исследуя визуализатор поиска ассоциативных правил «Популярные наборы», можно сделать вывод, что такие продукты, как колбаса, хлеб, минеральная вода являются приоритетными к покупке в торговой точке.

4.5 Визуализатор «Правила»

Визуализатор «Правила» (рисунок 9).

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	kolbasa	sigareti	9	30,00	60,00	1,286
2	2	sigareti	kolbasa	9	30,00	64,29	1,286
3	3	kolbasa	xleb	9	30,00	60,00	0,857
4	4	kolbasa	yayza	9	30,00	60,00	1,000

Рисунок 9 – Визуализатор Правила

Из полученных результатов на рисунке 7 видно, что при покупке конфет, покупатель с вероятностью 80% купит и яйца и содовую, при покупке колбасы он купит сигареты с вероятностью 60%.

4.6 Визуализатор «Дерево правил»

При построении дерева правил по следствию на первом (верхнем) уровне находятся узлы со следствиями, а на втором уровне – узлы с условиями.

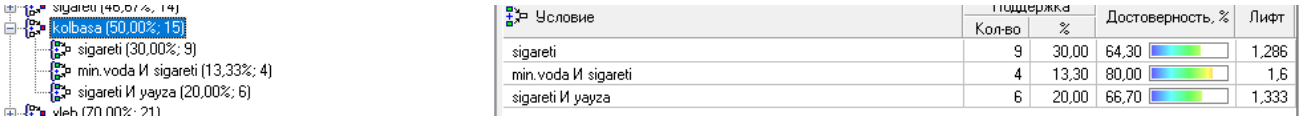


Рисунок 10 – Построение дерева правил по следствию

Например, для того чтобы человек приобрел колбасу, он должен купить хотя бы один предмет или несколько предметов из списка: сигареты, минеральную воду и сигареты, сигареты и яйца.

Второй вариант дерева правил – дерево, построенное по условию. Здесь на первом уровне располагаются узлы с условием (рисунок 11).

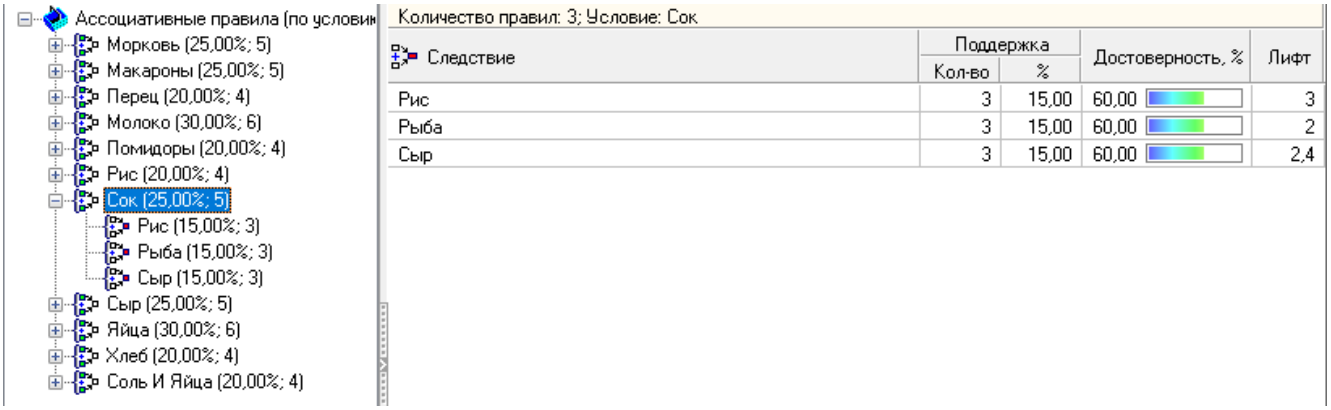
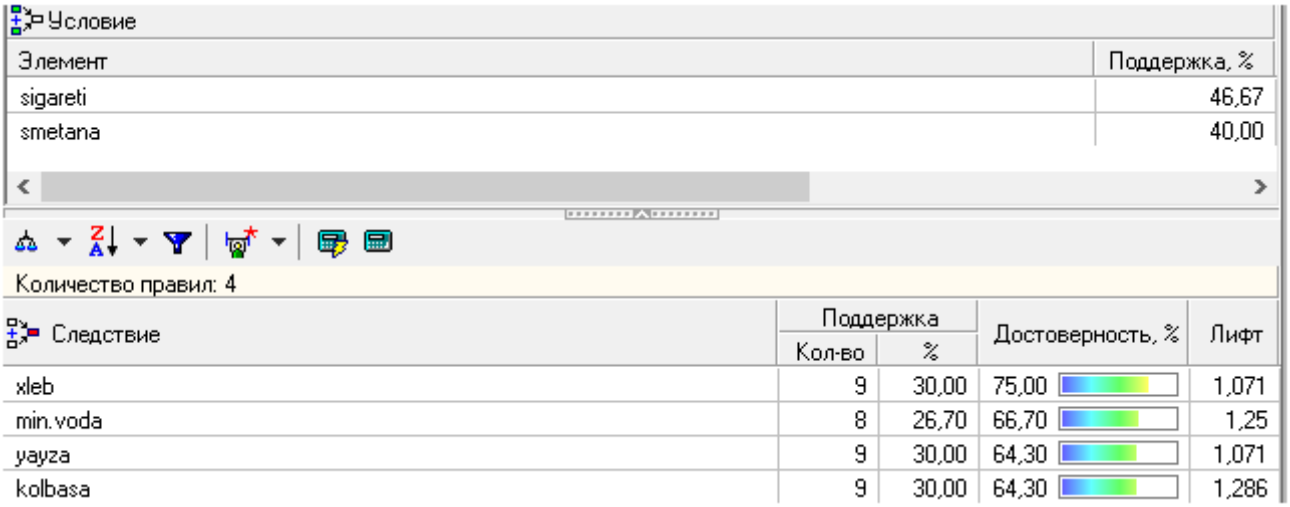


Рисунок 11 – Построение дерева правил по условию

Узлы - верхний уровень дерева и условие. А ветви – следствия. Это означает, что покупатель, купивший колбасу, так же купит сигареты, хлеб и яйца с достоверностью 60%.

4.7 Анализ «Что-если»

Анализ “Что-если” позволяет определить, что получим в качестве следствия, если выберем определенные условия. Например, какие товары приобретаются совместно с выбранными товарами. Пусть необходимо проанализировать, что, возможно, забыл покупатель приобрести, если он уже взял сигареты и сметану.



Условие

Элемент	Поддержка, %
sigareti	46,67
smetana	40,00

Количество правил: 4

Следствие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
xleb	9	30,00	75,00	1,071
min.voda	8	26,70	66,70	1,25
yayza	9	30,00	64,30	1,071
kolbasa	9	30,00	64,30	1,286

Рисунок 12 – Результат анализа Что-Если

В данном случае появится минеральная вода, хлеб, яйца и колбаса. Именно эти продукты покупатель возможно забыл приобрести с вероятностью 65-75%.

5 ПРОГНОЗИРОВАНИЕ ВРЕМЕННОГО РЯДА

5.1 Импорт данных

Для прогнозирования временного ряда были выбраны данные по продаже шампуня за три года по месяцам:

```
"Month","Sales"
"2018-01",266.0
"2018-02",145.9
"2018-03",183.1
"2018-04",119.3
"2018-05",180.3
"2018-06",168.5
"2018-07",231.8
"2018-08",224.5
"2018-09",192.8
"2018-10",122.9
...
```

Построим по исходным данным диаграмму:

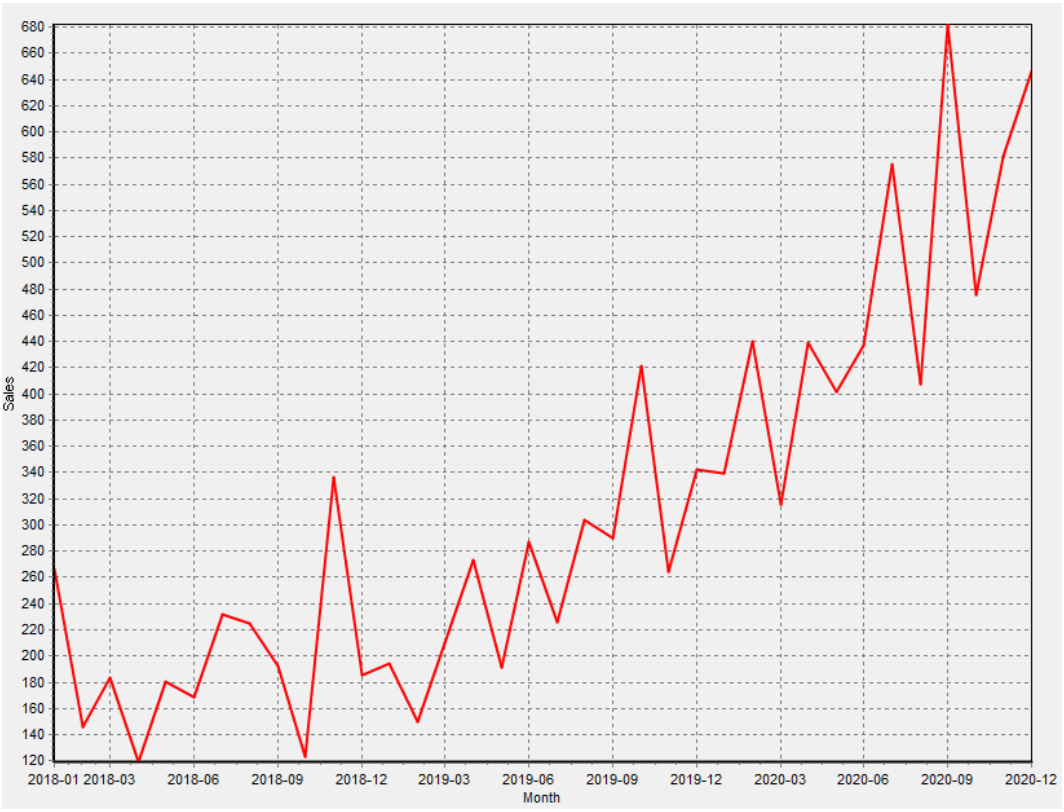


Рисунок 13 – Диаграмма, построенная по исходным данным

5.2 Обработка исходных данных

По диаграмме видно, что в данных содержатся выбросы и шумы. Произведем «Редактирование выбросов и экстремальных значений» и «Спектральную обработку». Результат представлен на рисунке 14.

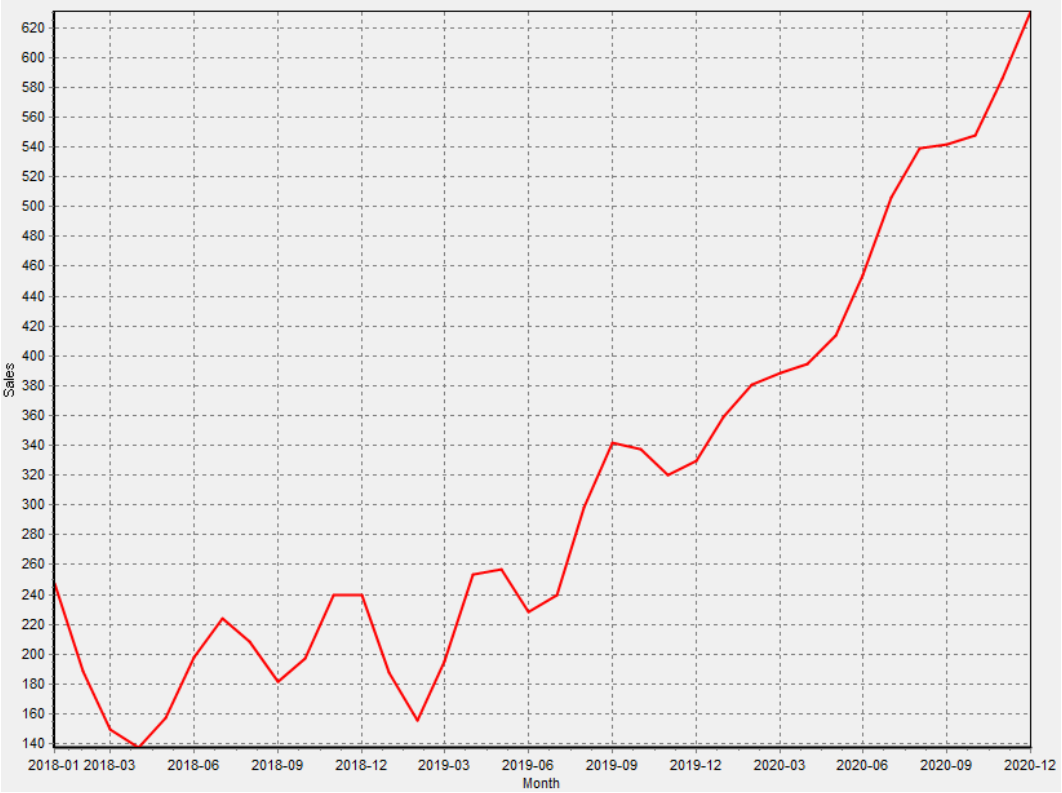


Рисунок 14 – Диаграмма после обработки

Отберем данные используя метод «Скользящее окно» с глубиной погружения 12 месяцев. Результат представлен на рисунке 15.

Month	Sales-12	Sales-11	Sales-10	Sales-9	Sales-8	Sales-7	Sales-6	Sales-5	Sales-4	Sales-3	Sales-2	Sales-1	Sales
2019-01	266	145.9	183.1	119.3	180.3	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3
2019-02	145.9	183.1	119.3	180.3	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5
2019-03	183.1	119.3	180.3	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1
2019-04	119.3	180.3	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3
2019-05	180.3	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4
2019-06	168.5	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287
2019-07	231.8	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287	226
2019-08	224.5	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287	226	303.6
2019-09	192.8	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287	226	303.6	289.9
2019-10	122.9	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287	226	303.6	289.9	421.6
2019-11	336.5	185.9	194.3	149.5	210.1	273.3	191.4	287	226	303.6	289.9	421.6	264.5
2019-12	185.9	194.3	149.5	210.1	273.3	191.4	287	226	303.6	289.9	421.6	264.5	342.3
2020-01	194.3	149.5	210.1	273.3	191.4	287	226	303.6	289.9	421.6	264.5	342.3	339.7
2020-02	149.5	210.1	273.3	191.4	287	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4
2020-03	210.1	273.3	191.4	287	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9
2020-04	273.3	191.4	287	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3
2020-05	191.4	287	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3
2020-06	287	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4
2020-07	226	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5
2020-08	303.6	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5	407.6
2020-09	289.9	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5	407.6	682
2020-10	421.6	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5	407.6	682	475.3
2020-11	264.5	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5	407.6	682	475.3	581.3
2020-12	342.3	339.7	440.4	315.9	439.3	401.3	437.4	575.5	407.6	682	475.3	581.3	646.9

Рисунок 15 – Результат работы метода «Скользящее окно»

5.3 Обучение нейросети

Воспользуемся мастером обработки, выберем в нем нейросеть. В качестве входных полей используем первые 6 столбцов. Разобьем данные на тестовые и обучающие Процесс изображен на рисунках 17 и 16.

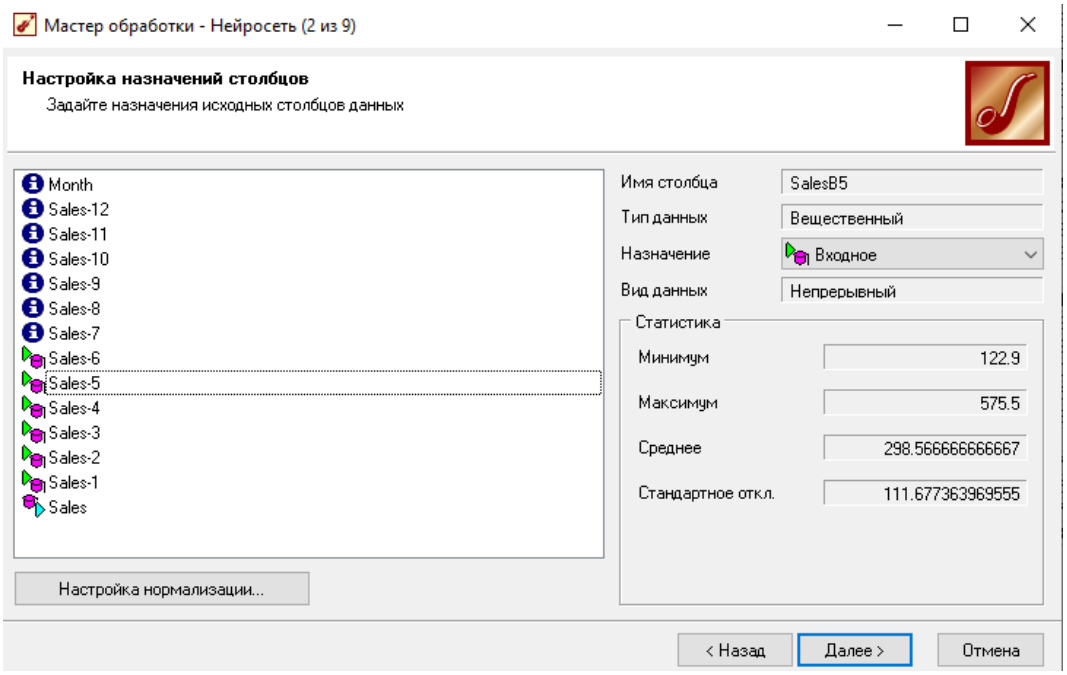


Рисунок 16 – Выбор столбцов входных данных для нейронной сети

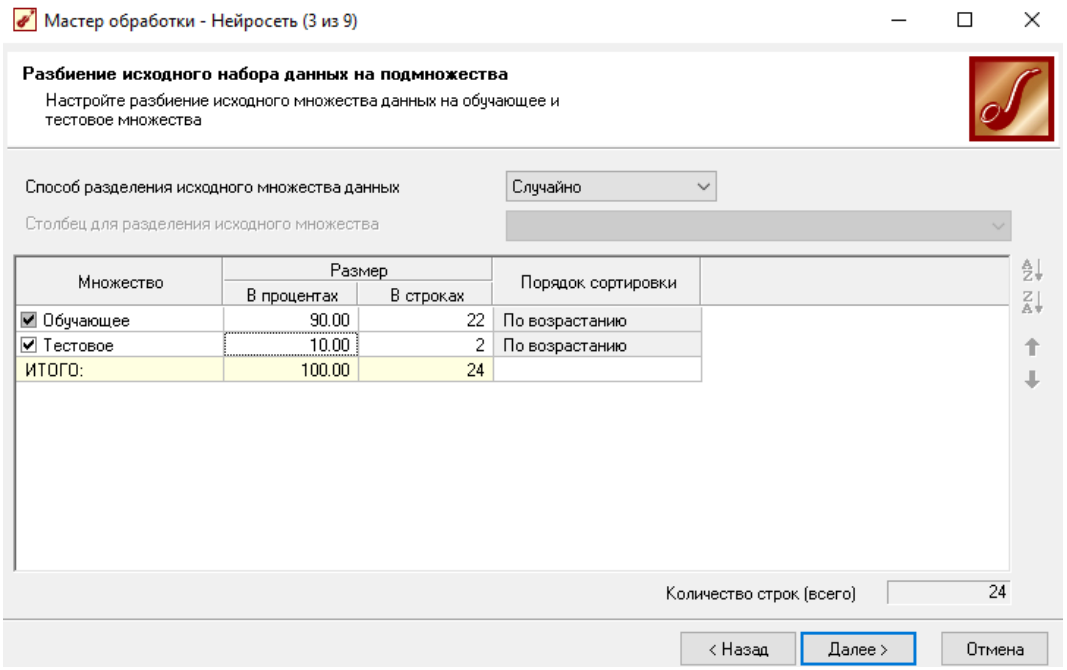


Рисунок 17 – Распределение данных на обучающее и тестовое подмножества

Результат обучения нейросети представлен диаграммой рассеяния (рисунок

18) и диаграммой, изображающей исходные данные вместе с моделью, полученной с помощью нейросети (рисунок fig:neuron-vs-data):

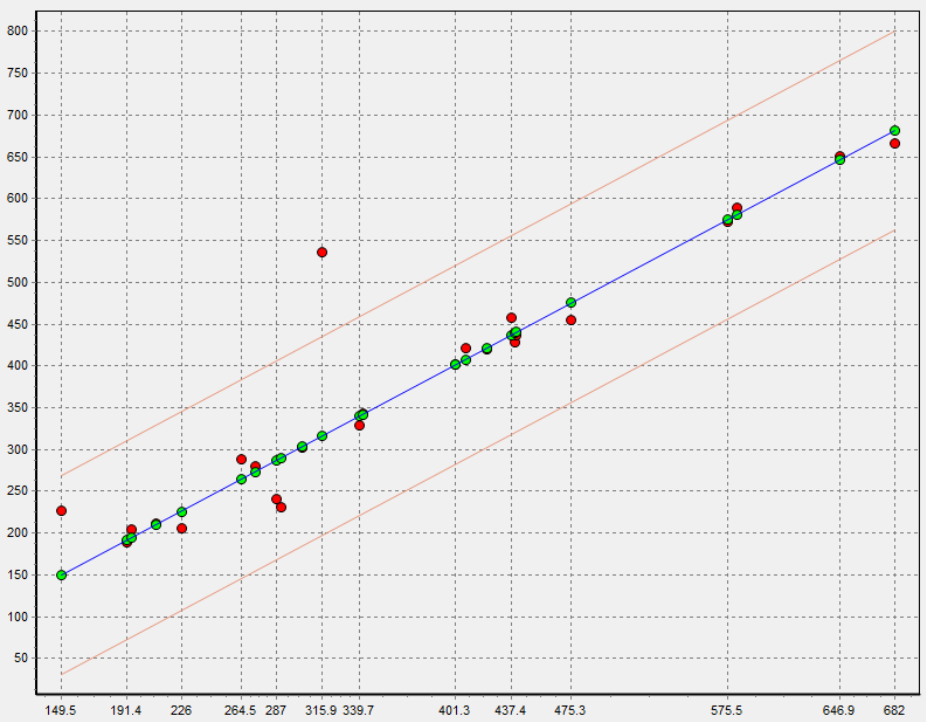


Рисунок 18 – Диаграмма рассеяния

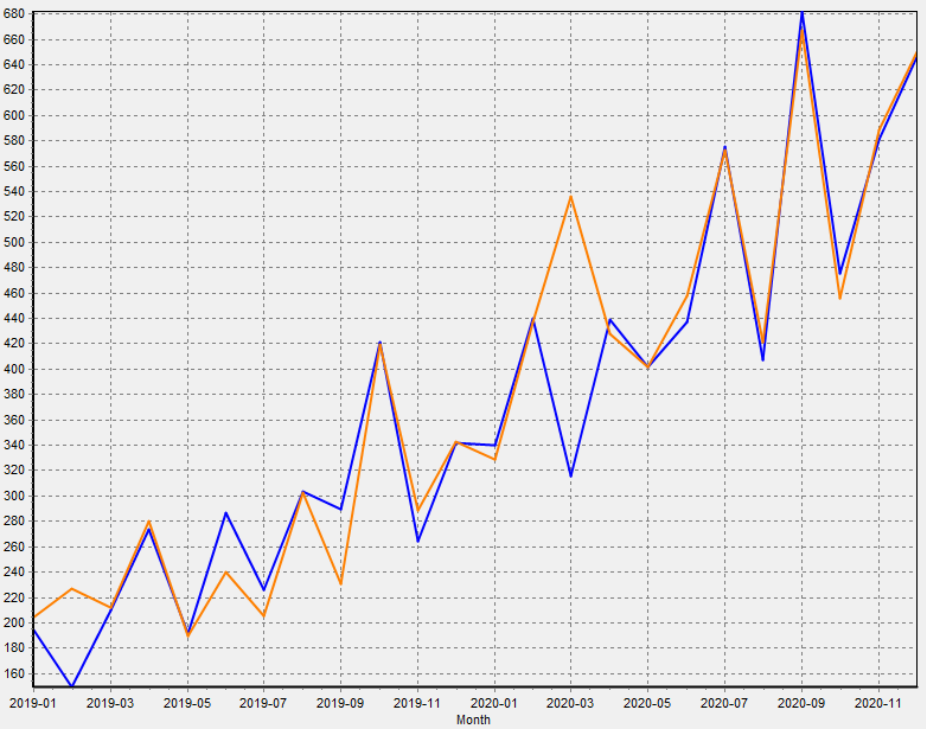


Рисунок 19 – Диаграмма по данным

На диаграмме видно, что во втором месяц е 2018 г. и во втором месяце

2020 года модель, построенная нейросетью не показывает реального падения в продажах.

5.4 Прогнозирование продаж

После того, как нейросеть была обучена, построим прогноз с помощью обработчика «Прогнозирование». Результат представлен на рисунке 20.

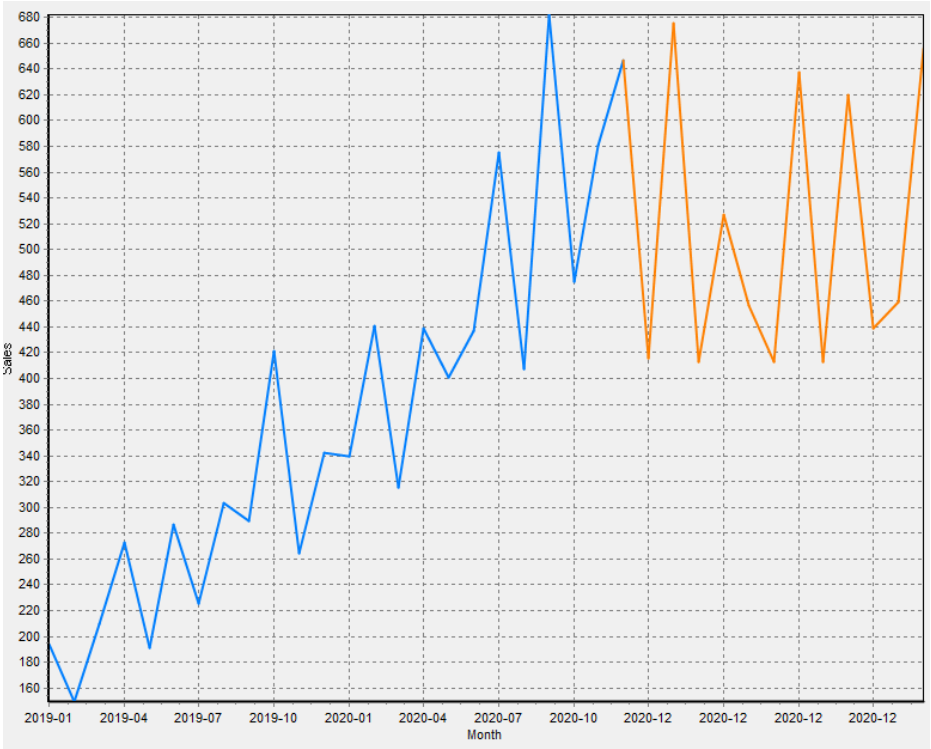


Рисунок 20 – Прогноз продаж на 1 год

На диаграмме синим цветом (первые два года) изображены текущие продажи, оранжевым цветом изображен прогноз продаж на год вперед.

ВЫВОДЫ

В ходе выполнения расчетно-графического задания был проведен поиск ассоциативных правил для данных, представляющих собой чеки покупателей продуктового магазина. Были выявлены популярные наборы: вода, картофель, колбаса, соль, яйца, лук, молоко. В визуализаторах «Правила», «Дерево правил» и «Что-если» были определены условия и вероятности того, что приобретет посетитель магазина, если он уже купил определенную товарную позицию в магазине.

Также было проведено прогнозирование временного ряда количества продаж шампуня за 3 года. При помощи «Редактирование выбросов и экстремальных значений» и «Спектральная обработка» была проведена обработка данных от аномалий и шумов, мешающих построению дальнейшей тенденции. Для прогнозирования временного ряда при помощи нейросети было проведена обработка данных методом «Скользящее окно» с глубиной погружения 12 месяцев. Проведено обучение нейросети и построен прогноз продаж шампуня на год вперед.