МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ

ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

кафедра «Информационные системы»

Расчетно-графическая работа

«Компьютерные методы анализа данных и прогнозирования»

по дисциплине «Интеллектуальный анализ данных»

**Выполнил**: ст. гр. ИС/б-20-2-о\_\_\_

Юрьева П.В.\_ \_\_ \_\_\_\_\_\_\_\_\_\_\_

**Проверил:** ст. пр. Сырых О.А.

**Севастополь**

2023 г.

Содержание

Введение

В современном информационном обществе огромное количество данных генерируется и накапливается ежедневно. Понимание и использование этих данных становятся ключевыми факторами для принятия обоснованных решений в различных областях. В этом контексте прогнозирование и поиск ассоциативных правил с помощью data mining (анализа данных) становятся мощными инструментами интеллектуального анализа данных (ИАД).

Прогнозирование – это процесс предсказания будущих событий или значений на основе имеющихся данных и трендов. Будь то прогнозирование продаж, финансовых показателей или климатических условий, точные прогнозы помогают организациям принимать стратегические решения и планировать свою деятельность.

Поиск ассоциативных правил, с другой стороны, направлен на выявление связей и зависимостей между различными элементами данных. Ассоциативные правила позволяют нам определить, какие элементы обычно встречаются вместе или какие события происходят параллельно. Это может быть полезно для принятия решений, планирования маркетинговых кампаний или выявления скрытых паттернов в данных.

В данной расчетно-графической работе рассматриваются два метода ИАД: прогнозирование временного ряда и поиск ассоциаций. Прогнозирование позволяет предугадывать будущие значения переменных на основе их исторических значений и других факторов. Поиск ассоциаций, в свою очередь, позволяет выявлять связи между различными переменными и определять, какие факторы влияют на другие.

Целью настоящей расчетно-графической работы являются изучение основ вышеупомянутых методов анализа экспериментальных данных, а также освоение техники их практического применения в Deductor Studio и закрепление знаний и умений по дисциплине «Интеллектуальный анализ данных».

Для достижения цели были решены следующие задачи:

1. Изучить возможности и инструменты Data minig

2.Использовать методы Data mining для поиска ассоциативных правил и прогнозирования временного ряда.

Пояснительная записка содержит полную информацию о расчетно-графической работе.

В разделе «Теоретическая часть» содержится методах Data Mining, а также более подробно рассмотрены методы ассоциативных правил и временного ряда.

В разделе «Практическая часть» представлены два подраздела.

Первый подраздел – «Поиск ассоциативных правил» – содержит описание последовательных шагов анализа данных о рецептах десертов с целью получения ассоциативных правил по ним; здесь же в отдельные подразделы вынесены результаты, отображенные в каждом визуализаторе.

Второй раздел – «Прогнозирование временных рядов» – содержит описание последовательных шагов анализа среднемесячной температуры в Париже на основе данных по месяцам за пять лет и прогнозирование данных на следующий год, а также сравнение полученного прогноза с фактическими данными.

В разделе «Заключение» содержатся выводы по результатам работы обоих методов, подведены итоги о достижении цели расчетно-графической работы.

# ОБЗОР МЕТОДОВ АНАЛИЗА ДАННЫХ И ПРОГНОЗИРОВАНИЯ

* 1. Интеллектуальный анализ данных:

Интеллектуальный анализ данных (ИАД) – это процесс извлечения полезной информации из больших объемов данных с использованием методов и техник машинного обучения, статистики, искусственного интеллекта и других подходов. Он помогает обнаруживать скрытые закономерности, выявлять взаимосвязи и делать прогнозы на основе имеющихся данных.

* 1. Data Mining

Data Mining – это процесс извлечения полезной информации, шаблонов и закономерностей из больших объемов данных. Он включает в себя применение различных методов, алгоритмов и статистических моделей для анализа данных с целью выявления скрытых знаний и принятия обоснованных решений.

В методах Data Mining используются различные техники и подходы для обработки данных и извлечения информации. Некоторые из основных методов включают:

1. Кластерный анализ: Этот метод используется для группировки схожих объектов или данных в кластеры. Кластерный анализ помогает выявить скрытые структуры и позволяет классифицировать данные на основе их сходства.
2. Классификация и регрессия: Классификация и регрессия являются методами, которые используются для прогнозирования и классификации объектов на основе их характеристик. Классификация относит объекты к заранее определенным классам, в то время как регрессия предсказывает непрерывные значения.
3. Ассоциативные правила: Как уже упоминалось, ассоциативные правила используются для обнаружения взаимосвязей между различными элементами набора данных. Они позволяют выявить часто встречающиеся комбинации и закономерности, которые могут быть полезны для прогнозирования и принятия решений.
4. Анализ временных рядов: анализ временных рядов, как было упомянуто ранее, позволяет предсказывать значения переменных в будущем на основе их предыдущих значений и трендов. Этот метод включает в себя использование статистических моделей и алгоритмов для прогнозирования будущих значений.

1.3. Поиск ассоциативных правил:

Поиск ассоциативных правил – это один из методов анализа данных, который позволяет выявить взаимосвязи между различными элементами набора данных. Ассоциативные правила определяют статистические связи между различными товарами, событиями или характеристиками, на основе которых можно делать выводы о поведении или предсказывать будущие события. Например, анализ продаж в супермаркете может показать, что клиенты, покупающие памперсы, часто также покупают детское питание, что позволяет магазину разрабатывать стратегии маркетинга или размещать товары в определенных зонах.

Все эти методы Data Mining имеют широкое применение в различных областях, включая бизнес, маркетинг, финансы, медицину, науку о данных и другие. Они помогают обнаруживать скрытые паттерны, выявлять взаимосвязи и делать прогнозы, что в свою очередь способствует принятию обоснованных решений и улучшению результатов.

В ходе выполнения расчетно-графического задания были подробно рассмотрены и применены методы поиска ассоциативных правил и анализа временных рядов.

# 2. ПРАКТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В АНАЛИТИЧЕСКОЙ СИСТЕМЕ DEDUCTOR

## 2.1. Решение задачи поиска ассоциаций в Deductor.

Для выполнения расчетно-графической работы были получены базовые навыки работы с Deductor Studio.

Deductor Studio – это инструмент для анализа данных и машинного обучения, который позволяет решать задачи классификации, регрессии, кластеризации и другие.

Был подготовлен набор данных для метода поиска ассоциаций – список рецептов десертов [2]. Как выглядят данные, иллюстрирует рисунок 2.1.1.



Рисунок 2.1.1 – Данные для исследования

Далее, переформатировав данные в файл с расширением .txt данные были загружены в Deductor с помощью функции «Мастер импорта».

Рисунок 2.1.2 демонстрируют работу с «Мастером импорта» а также метки и назначение столбцов. Для колонки «Продукт» установлено значение «Элемент», для колонки «Десерт» – транзакция. Оба столбца являются строковым и дискретным типом и видом данных соответственно.

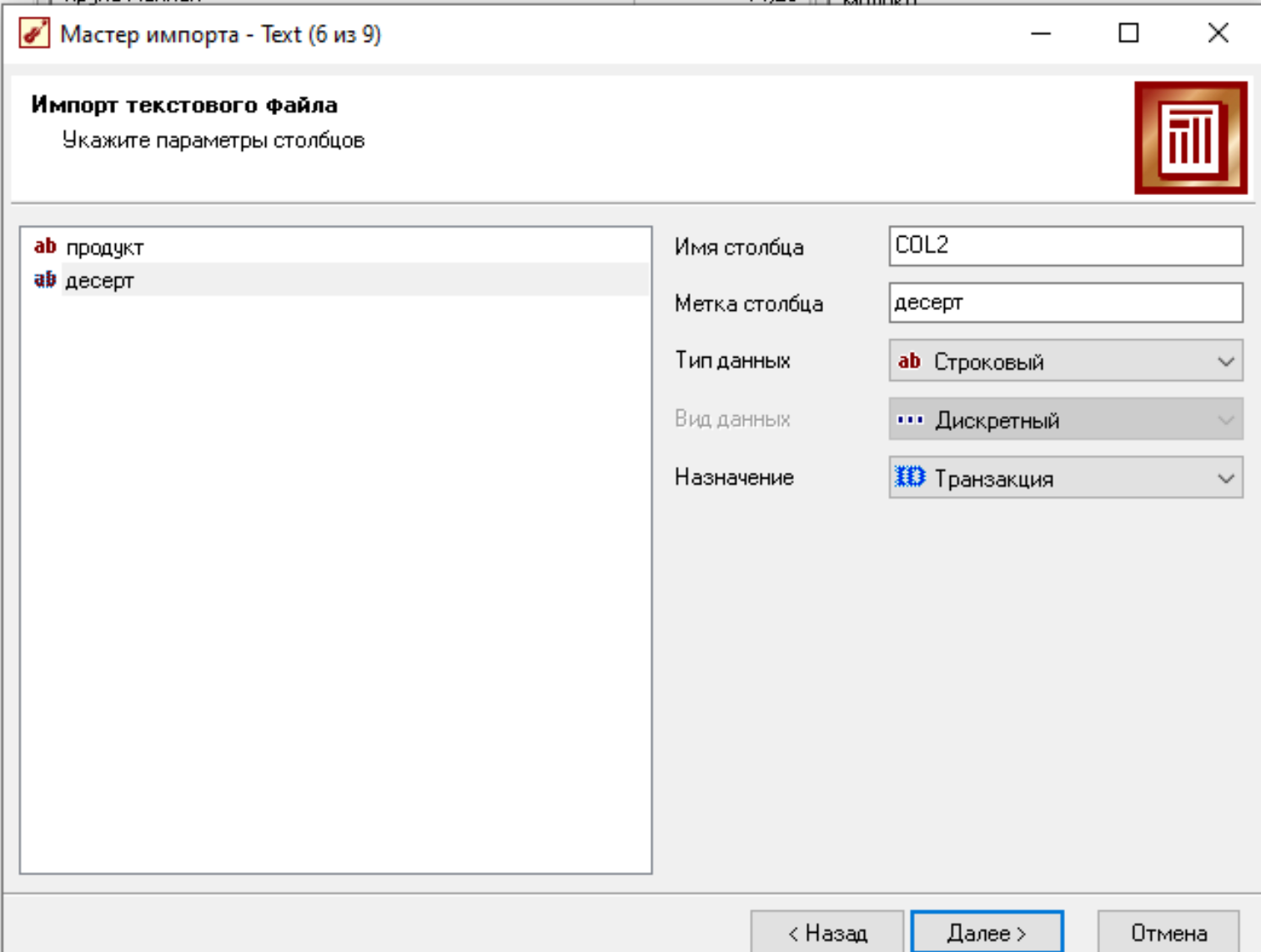
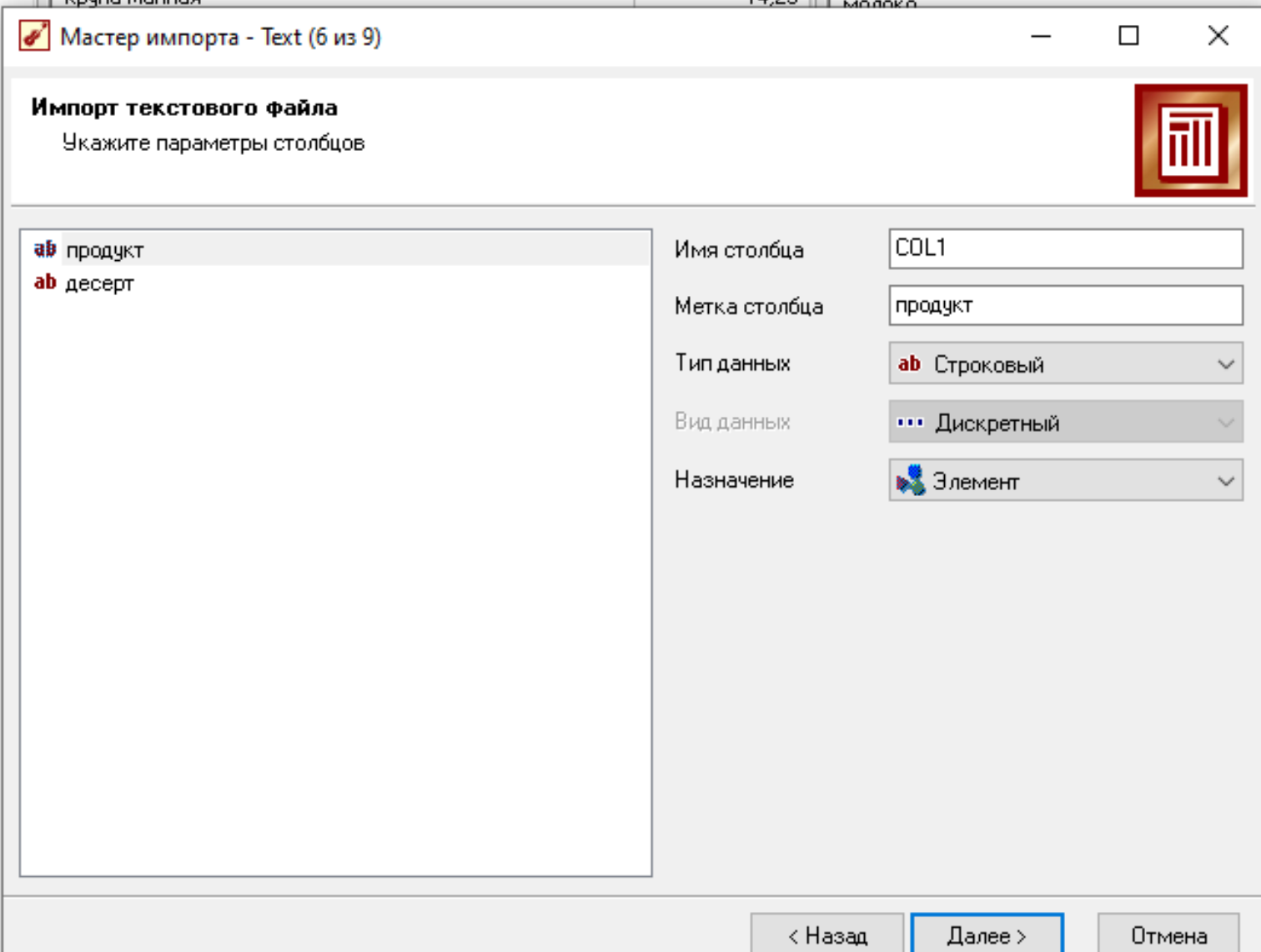


Рисунок 2.1.2 – Данные для исследования

Далее, на рисунке 2.1.3 продемонстрирована настройка отображения данных в формате таблицы.

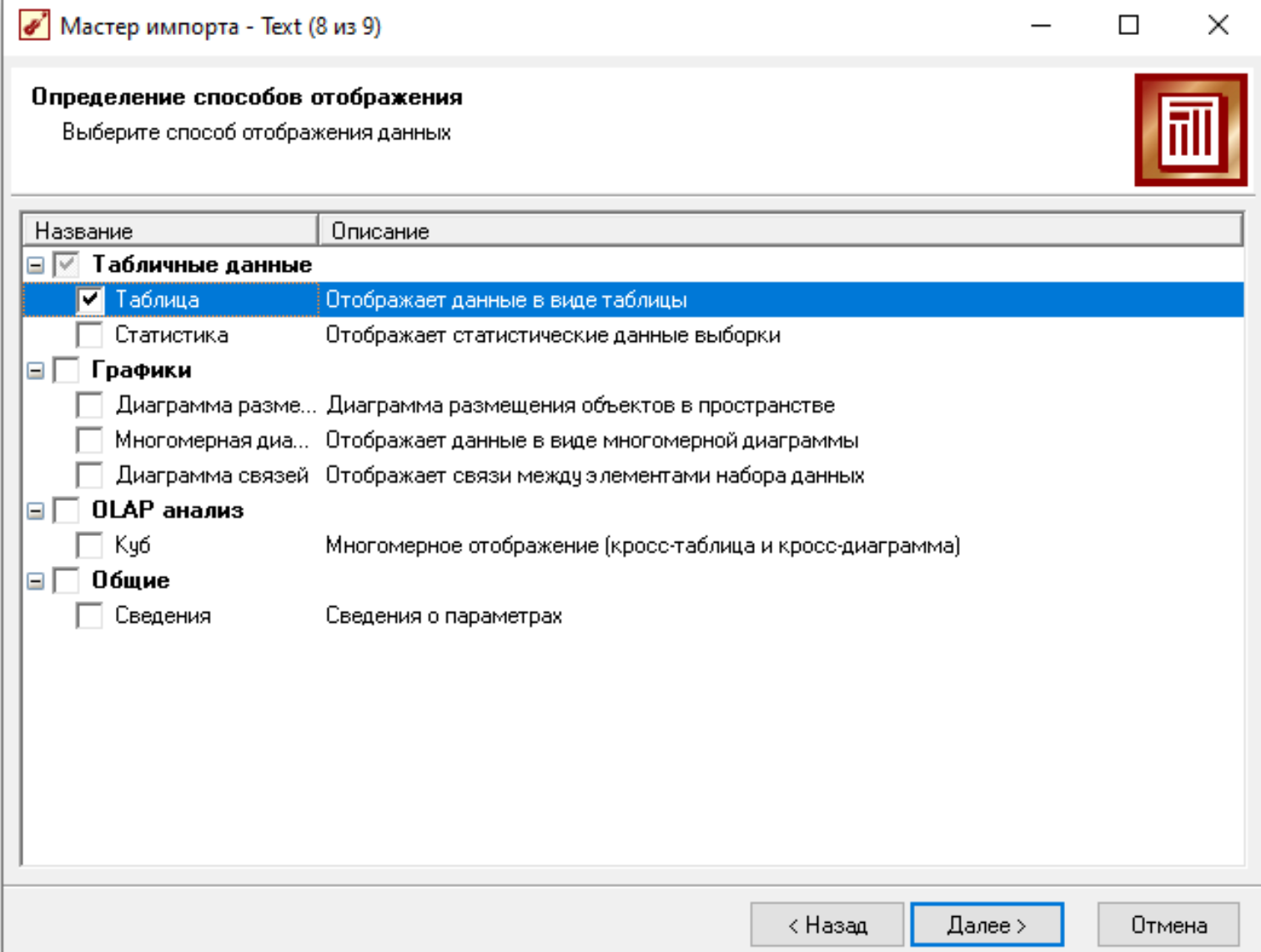


Рисунок 2.1.3 – Выбор способа отображения данных

2.2. Ассоциативные правила

Рассмотрим инструмент «Ассоциативные правила» в Deductor Studio. На рисунке 2.2.1 продемонстрировано окно «Мастер обработки», где можно выбрать один из методов Data Mining – ассоциативные правила.

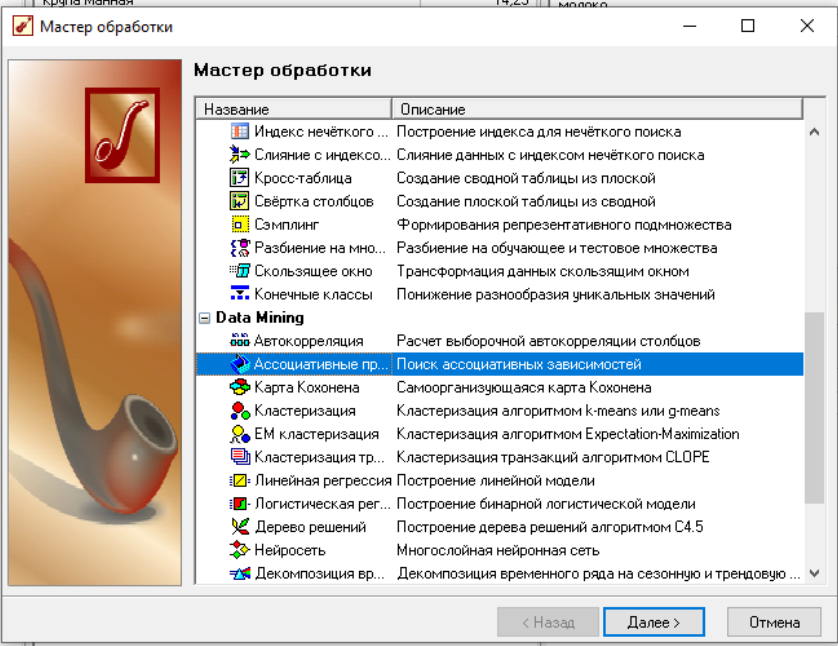


Рисунок 2.2.1 – Мастер обработки

Рисунки 2.2.2 и 2.2.3 отображают настройку параметров ассоциативных правил. Рисунок 2.2.4 демонстрирует успешное выполнение поиска ассоциативных правил.

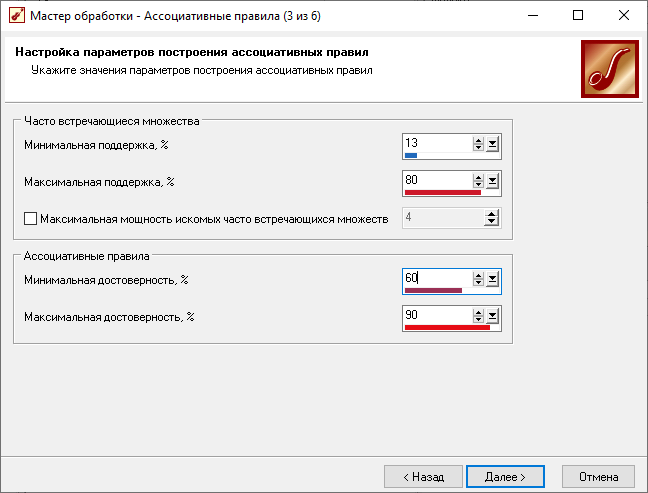
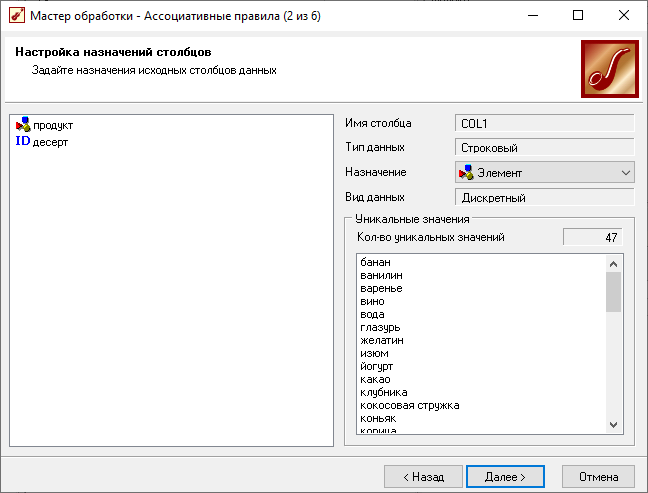


Рисунок 2.2.2 – Настройка параметров

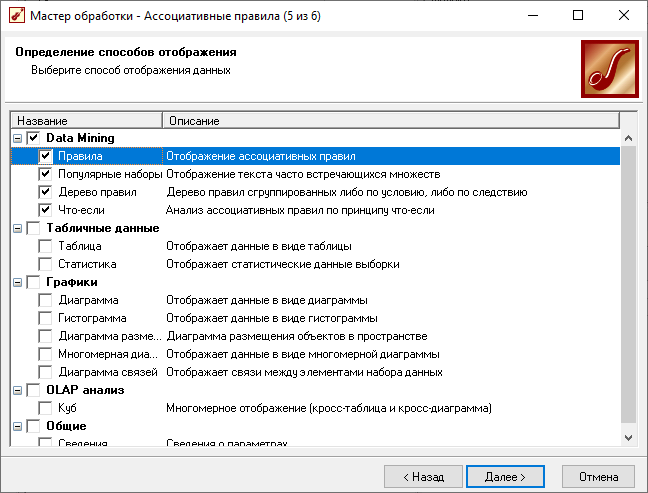
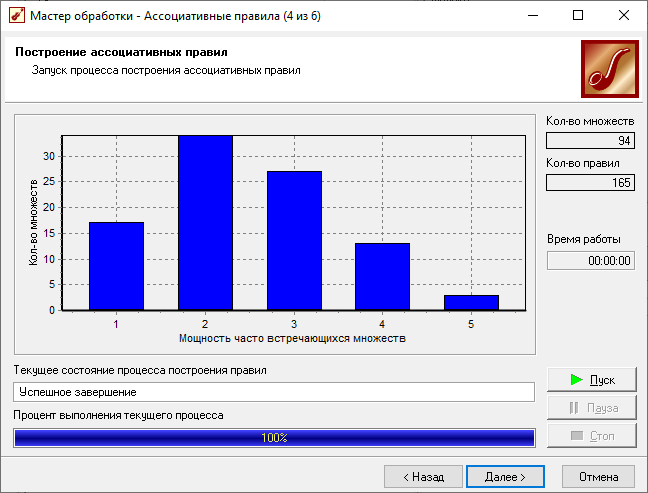


Рисунок 2.2.3 – Статус построения ассоциативных правил и варианты отображения результатов

Всего есть 4 визуализатора – «Правила», «Популярные наборы», «Дерево правил» и «Что-если».

Рассмотрим результаты в визуализаторе «Правила». Данный визуализатор отображает ассоциативные правила в виде списка правил, которые описывают поведение. Например, исходя их рисунка 2.2.4 можно сделать вывод, если есть какао и масло сливочное, то с вероятностью 87,5% будет добавлен сахар.

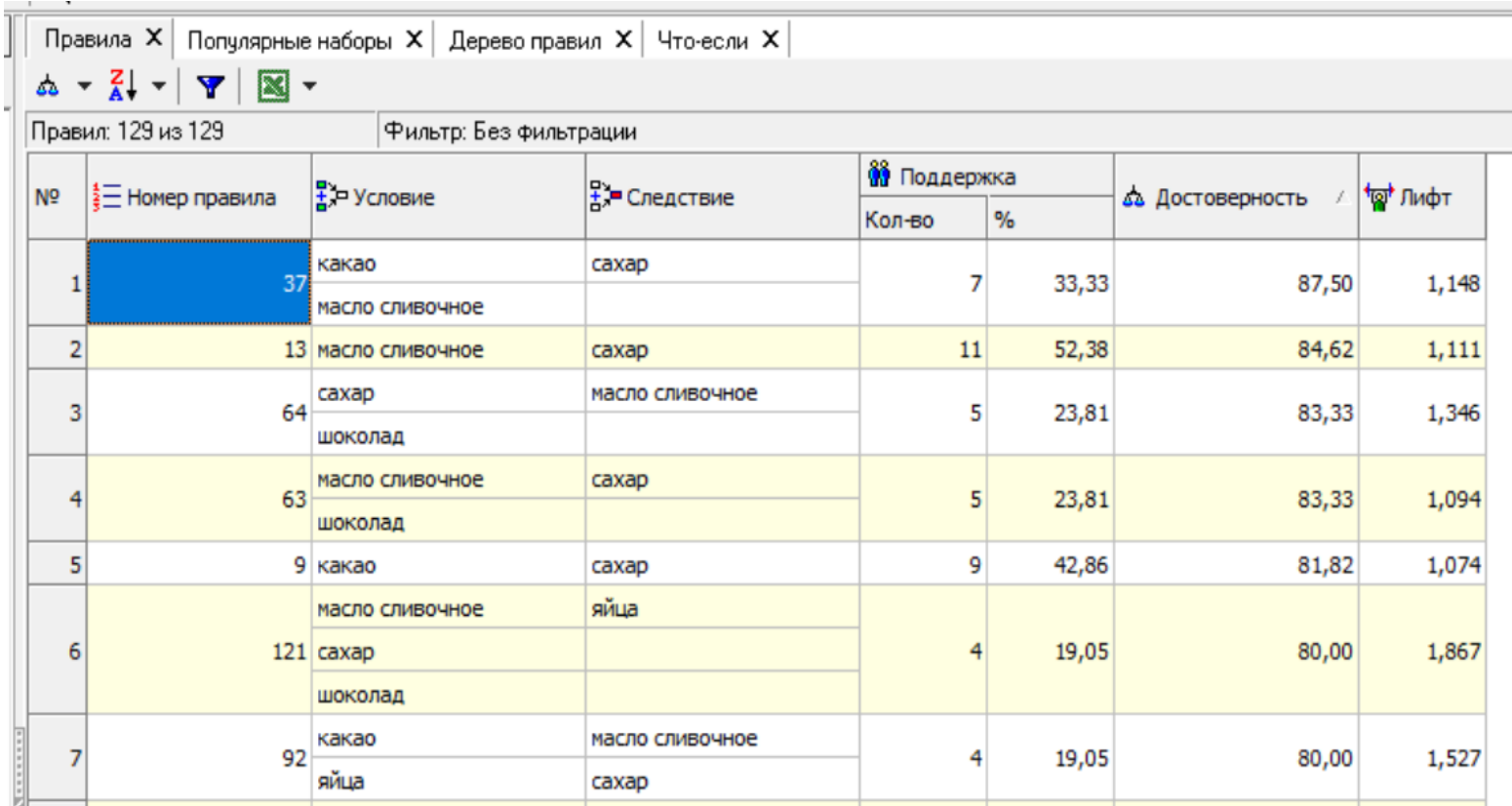


Рисунок 2.2.4 – Результаты построения правил

Популярные наборы – это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно. На сколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Из рисунка 2.2.5 видно, что самыми популярными продуктами являются «сахар», «масло сливочное» и «какао».

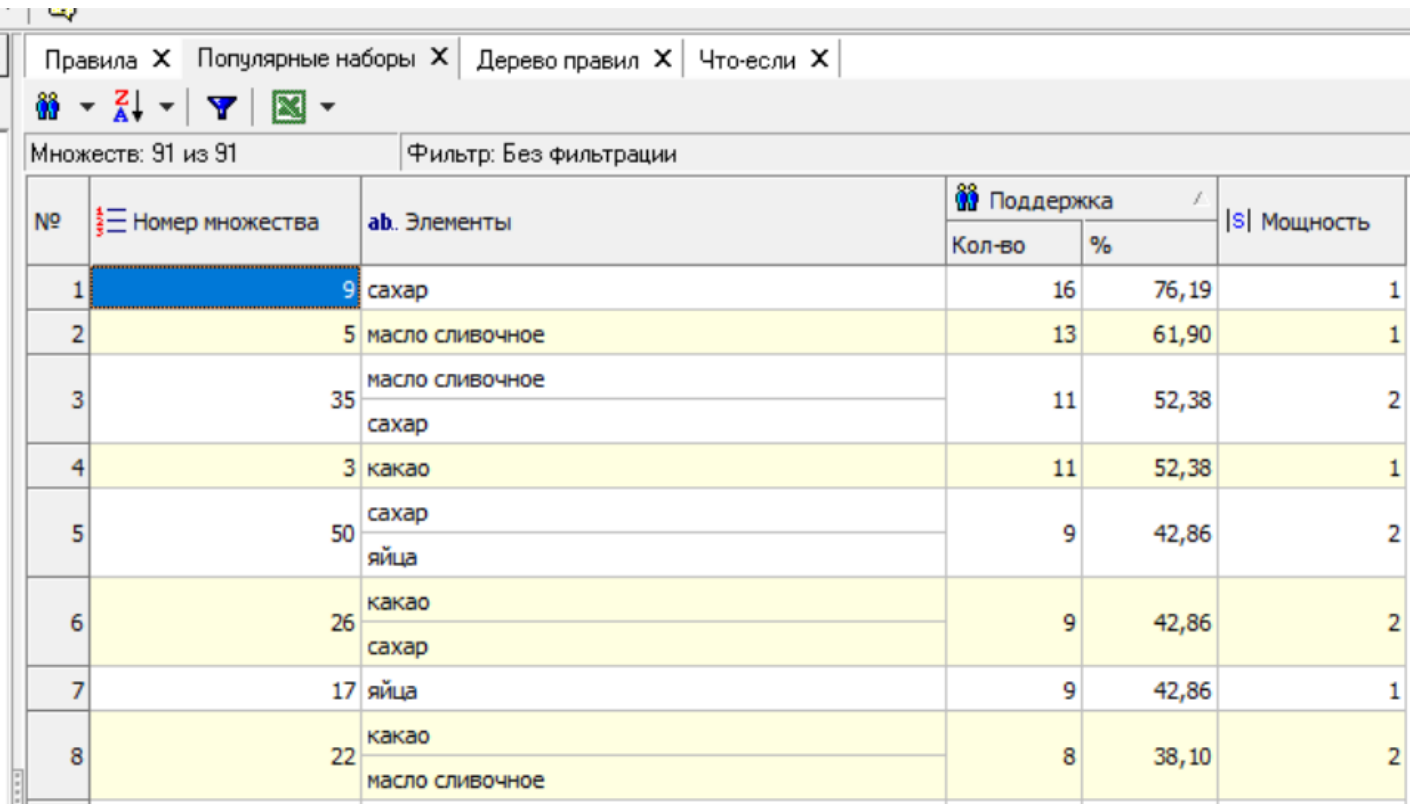


Рисунок 2.2.5 – Результаты построения популярных наборов

Визуализатор «Дерево правил» – это всегда двухуровневое дерево. Оно может быть построено либо по условию, либо по следствию. При построении дерева правил по условию на первом (верхнем) уровне находятся узлы с условиями, а на втором уровне – узлы со следствием.

Исходя из рисунка 2.2.6, на котором представлено дерево правил по условию, можно сделать вывод, что если добавлен продукт «какао», то с вероятностью 72,2% будет добавлен продукт «масло сливочное», с вероятностью 81,8% будет добавлен «сахар», и с вероятностью в 63,6% – оба продукта.

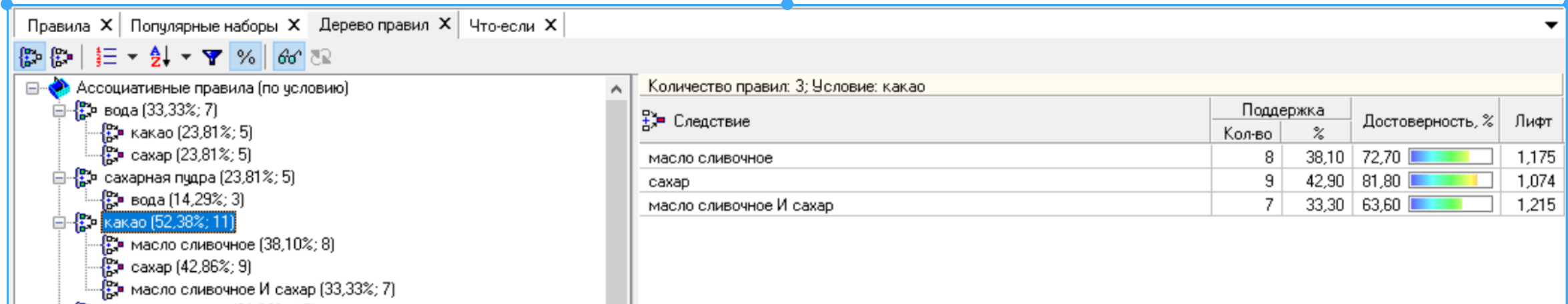


Рисунок 2.2.6 – Дерево правил (по условию)

Рассмотрим набор правил по следствию в дереве правил. По представленному дереву на рисунке 2.2.7 например, можно сделать такое заключение: чтобы добавить какао, с вероятностью 71,4% в рецепте должна также быть вода.

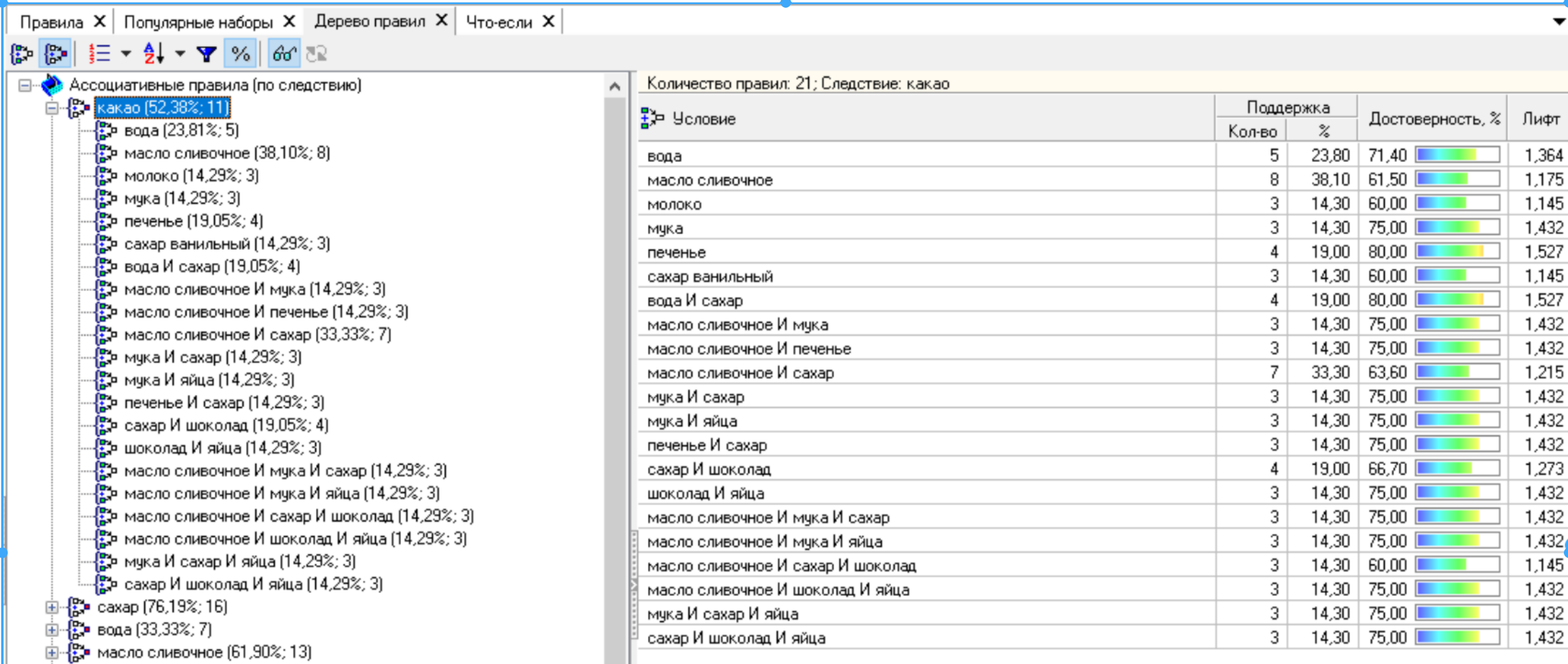


Рисунок 2.2.7 – Дерево правил (по следствию)

Анализ «Что-если» в ассоциативных правилах позволяет ответить на вопрос, что получим в качестве следствия, если выберем данные условия. Например, какие еще продукты и с какой вероятностью мы забыли положить в блюдо.

Пусть необходимо приготовить десерт с ингредиентами «вода», «какао», «молоко», «сахар» и необходимо понять, что еще необходимо добавить в блюдо. Из рисунка 2.2.8 можно понять, что с вероятностью в 80% необходимо добавить продукт «масло сливочное» и с вероятностью в 60% – «сахар ванильный».

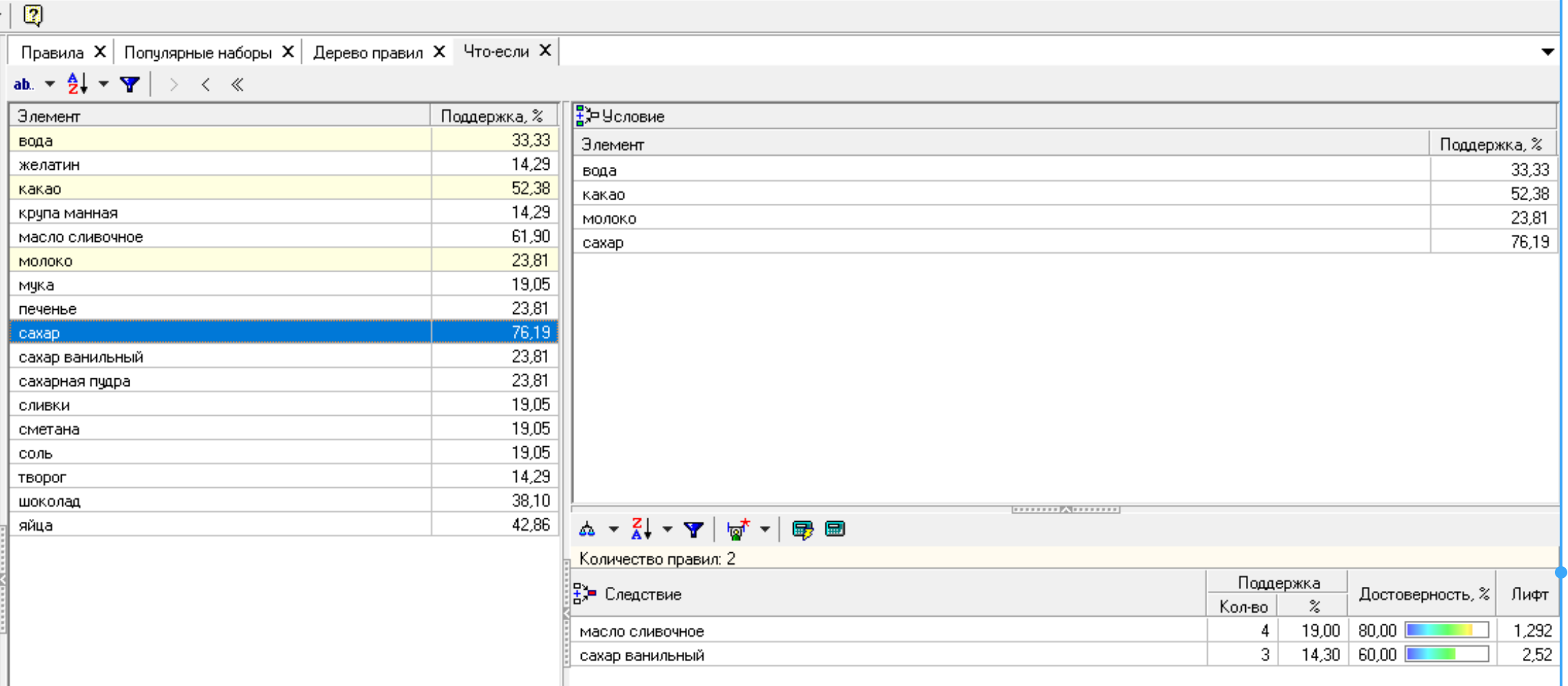


Рисунок 2.2.8 – Визуализатор «Что–если»

2.3. Временной ряд

Для анализа временного ряда были выбраны данные изменения температуры воздуха в Париже в течение трех лет. Вид данных представлен на рисунке 2.3.1. Рисунок 2.3.2. демонстрирует импорт данных.

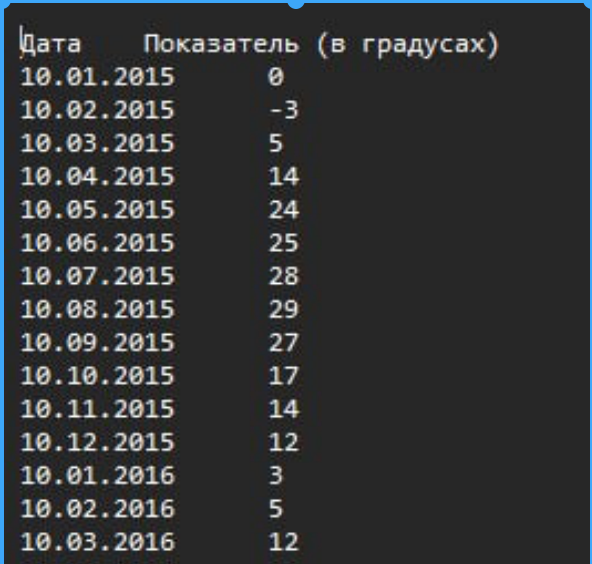
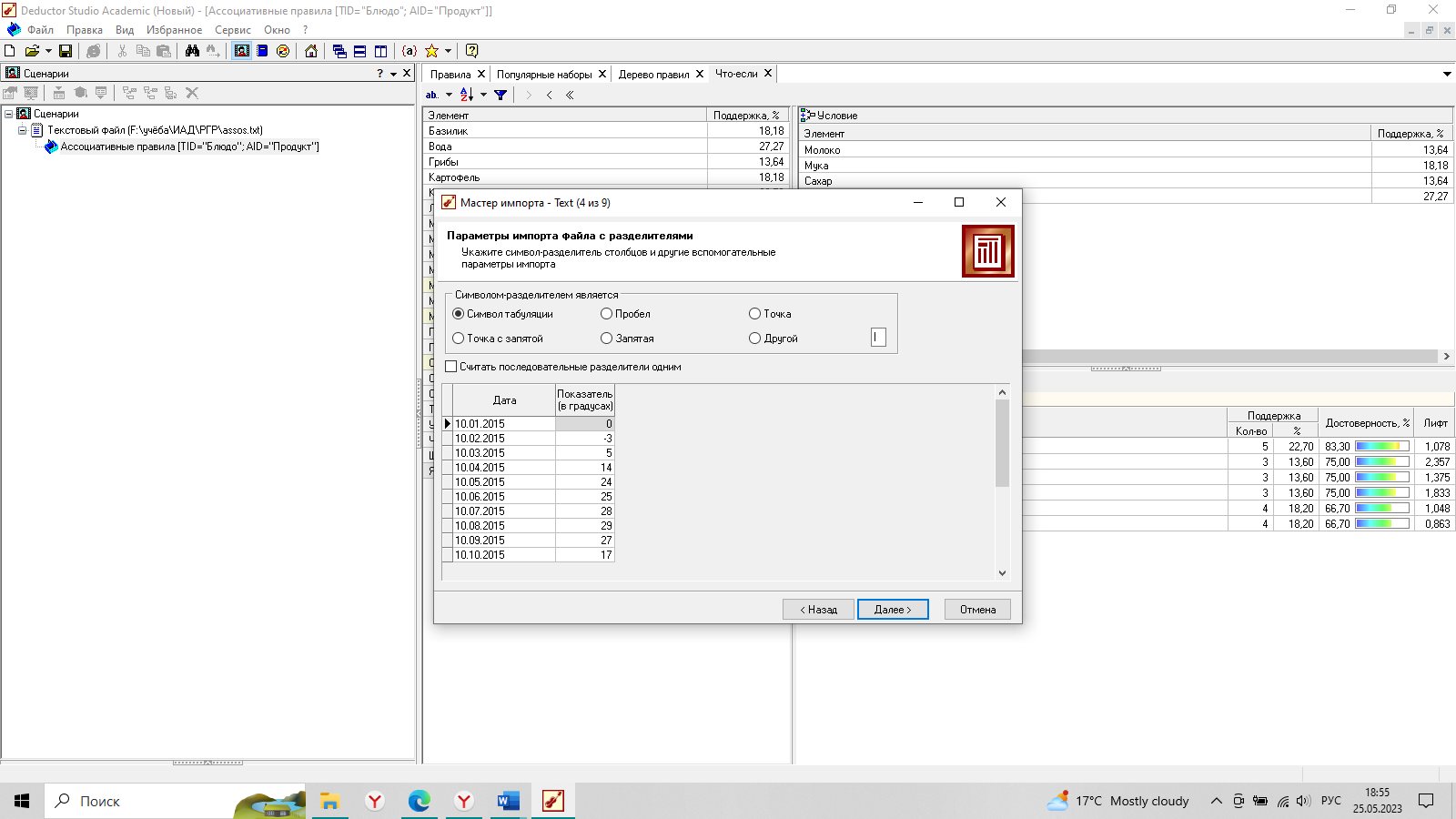
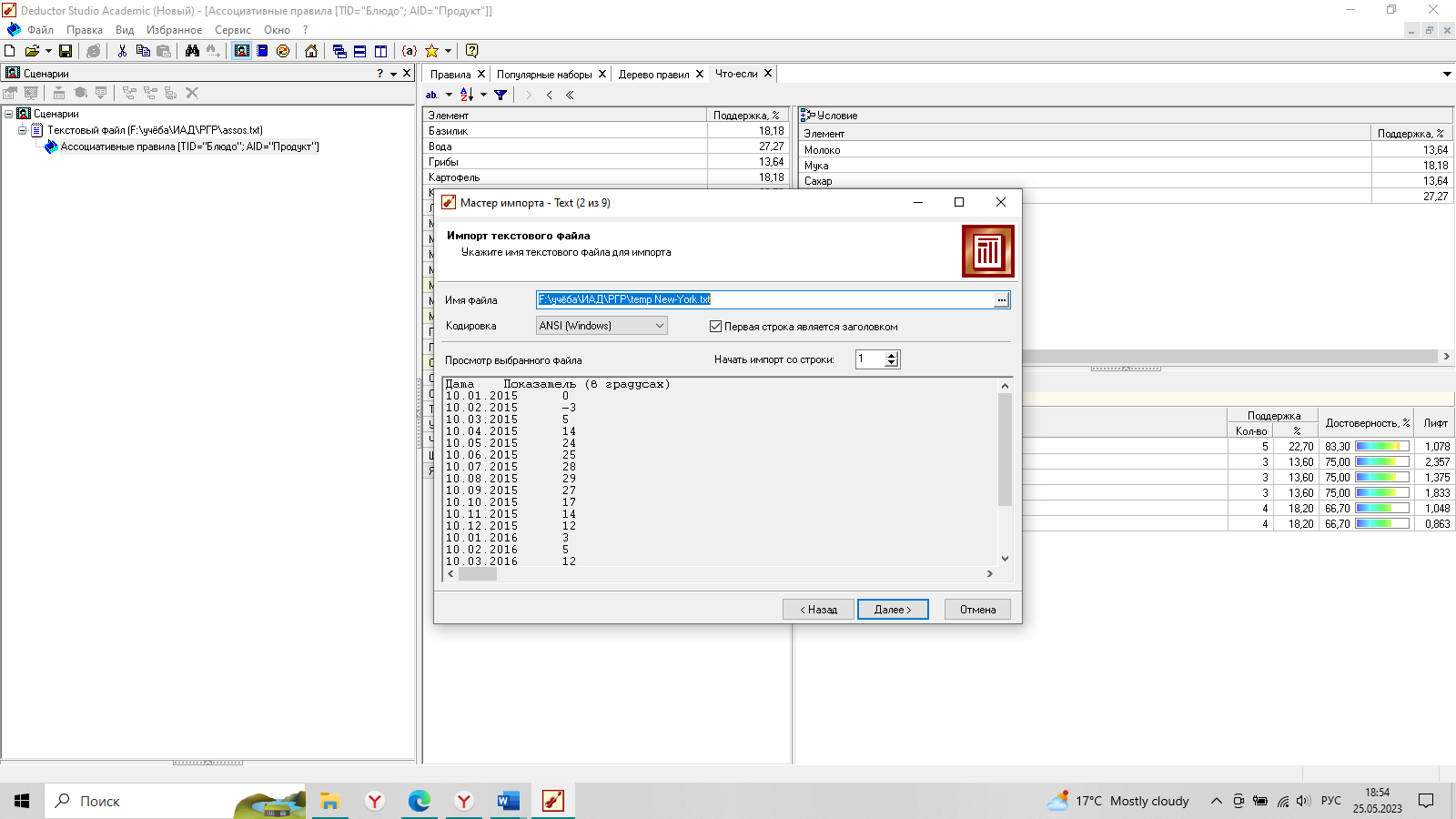


Рисунок 2.3.1 – Данные температуры в Париже



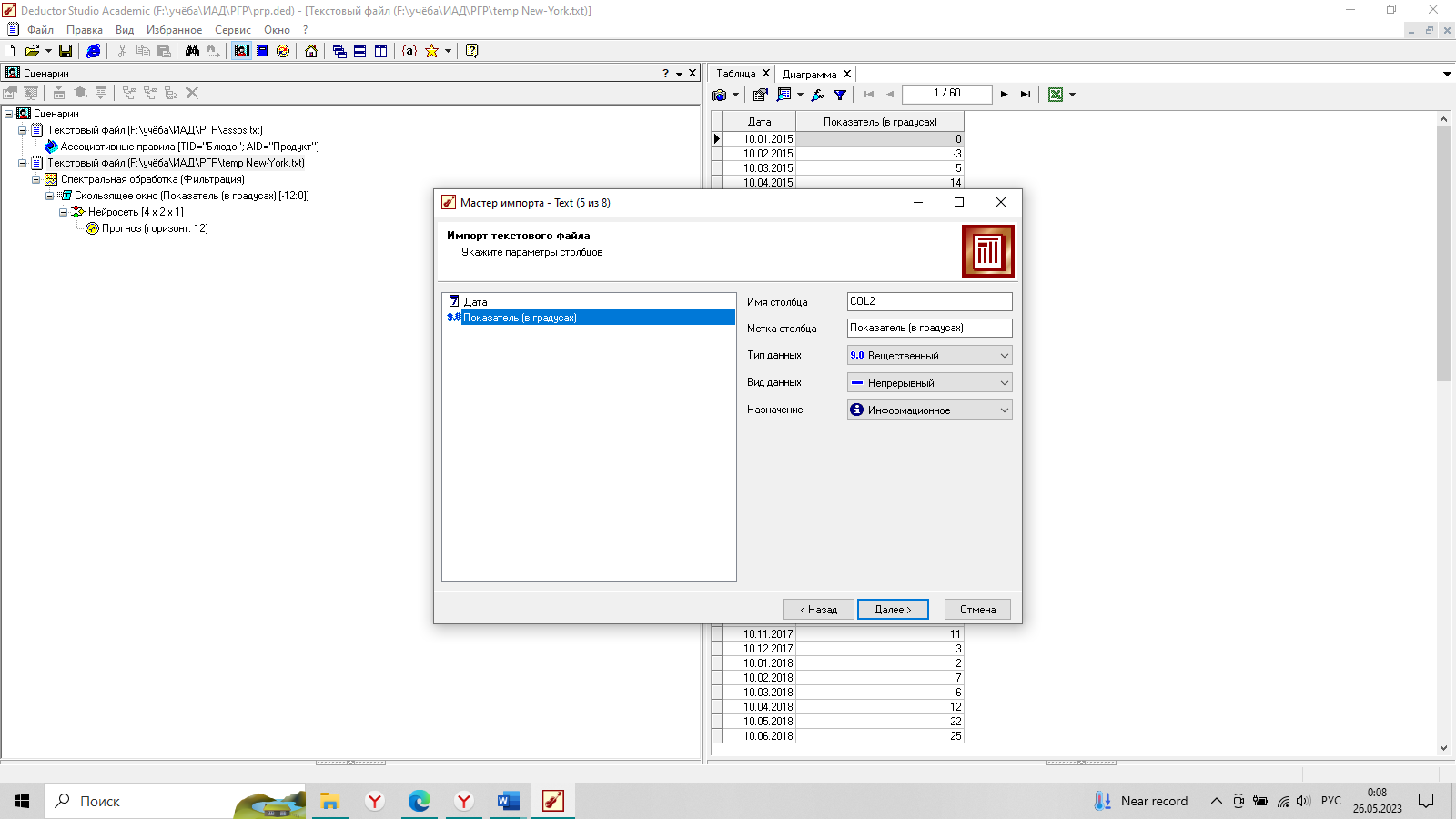
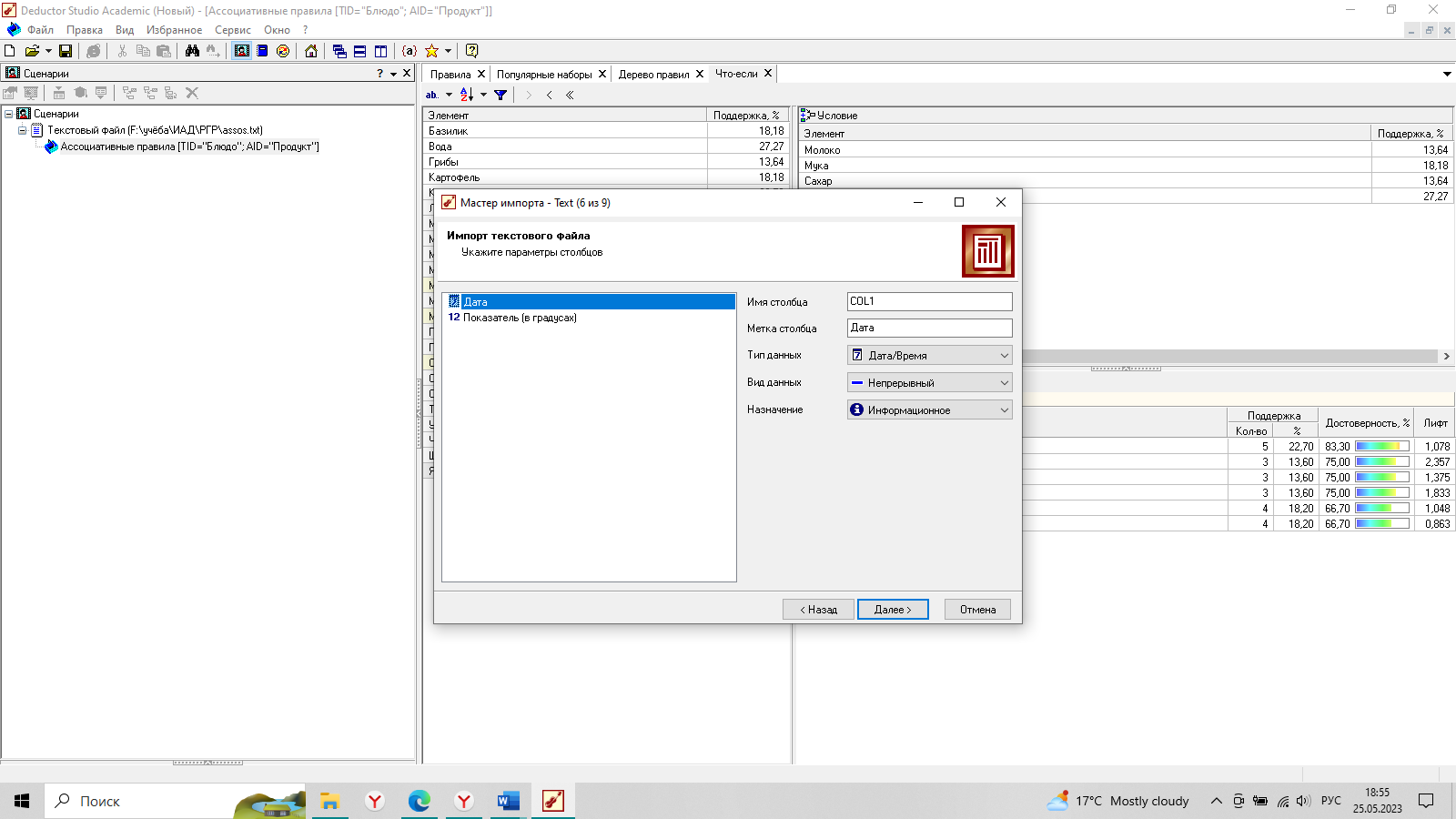


Рисунок 2.3.2. – Импорт данных

После импорта данных было проведено удаление аномалий и сглаживание данных. Рисунок 2.2.3 представляет график без сглаживания.

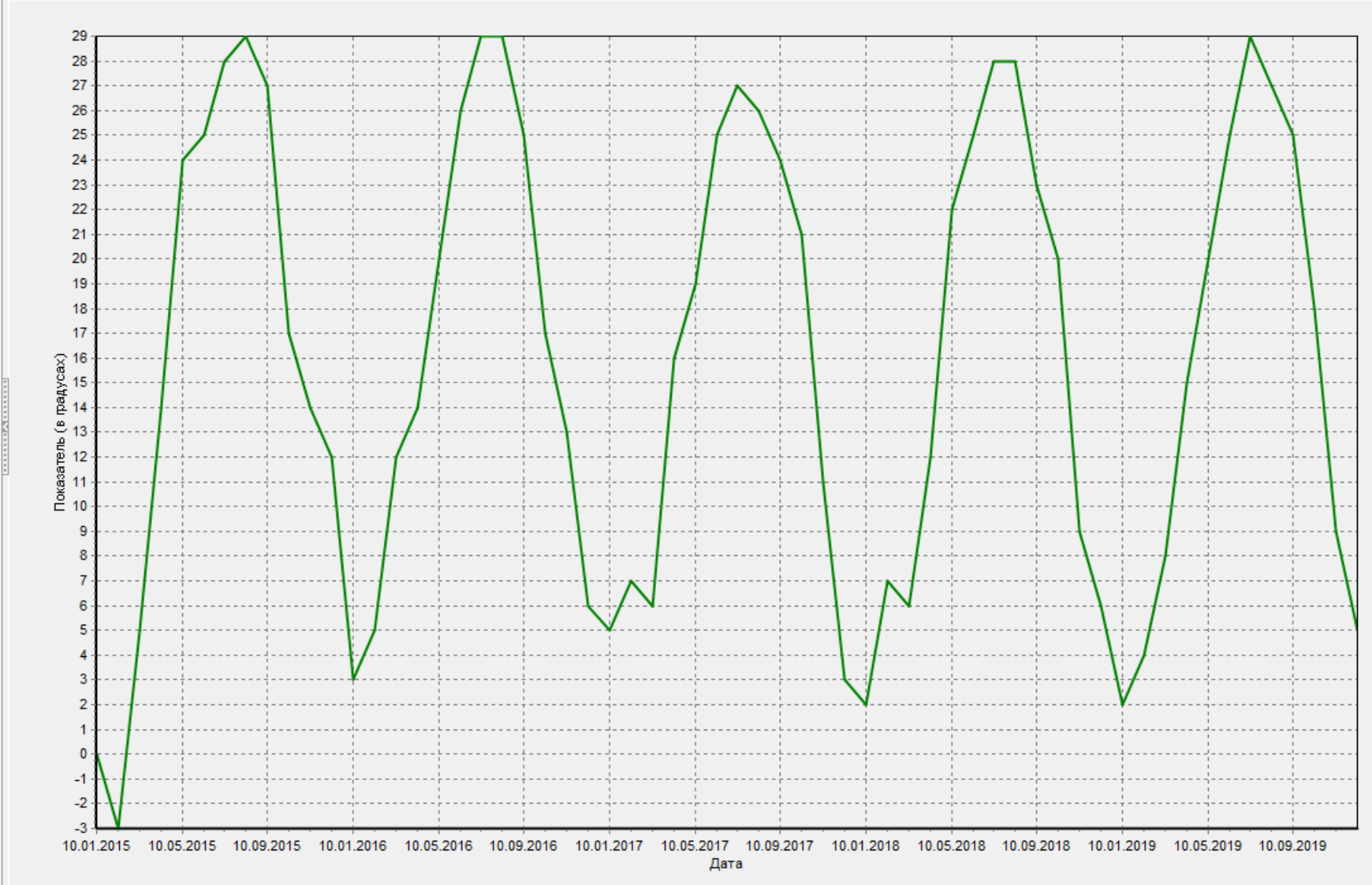


Рисунок 2.3.3 – Исходные данные без сглаживания

Сглаживание выполнялось с помощью Вейвлет-преобразования с параметрами 1 и 4 (чем меньше эти показатели, тем оптимальнее – данные не сильно искажены от сглаживания, и аномалии исчезнут). Настройка параметров сглаживания представлена на рисунке 2.3.4.

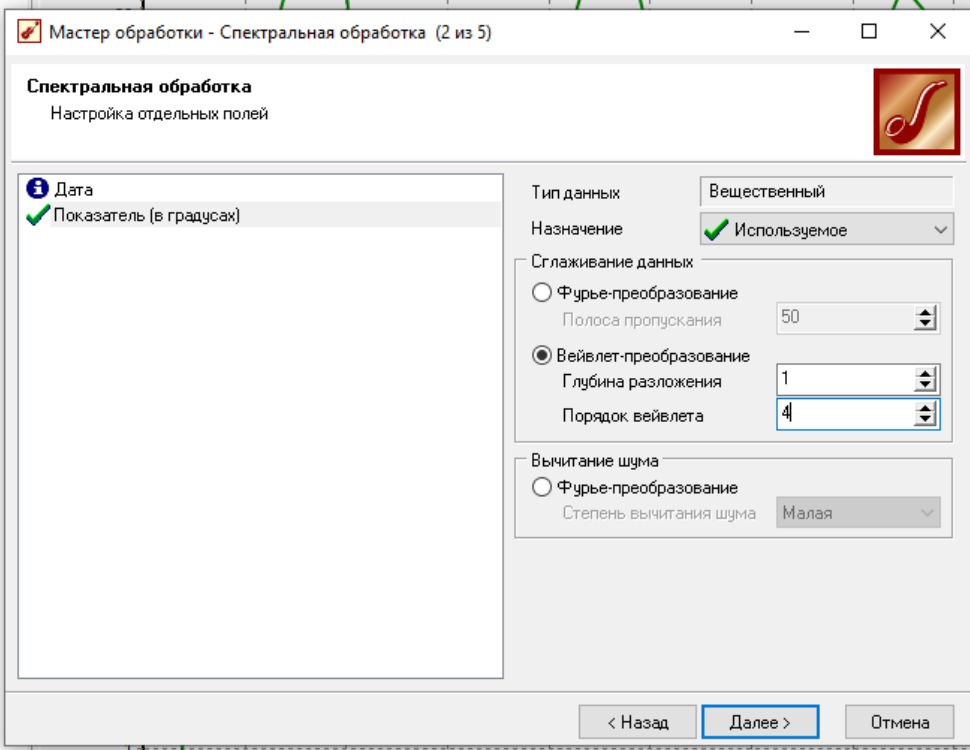


Рисунок 2.3.4 – Мастер обработки для инструмента «Спектральная обработка»

Рисунок 2.3.5. представляет диаграмму после преобразований. Из рисунка видно, что данные сгладились,а аномалии и шумы исчезли.

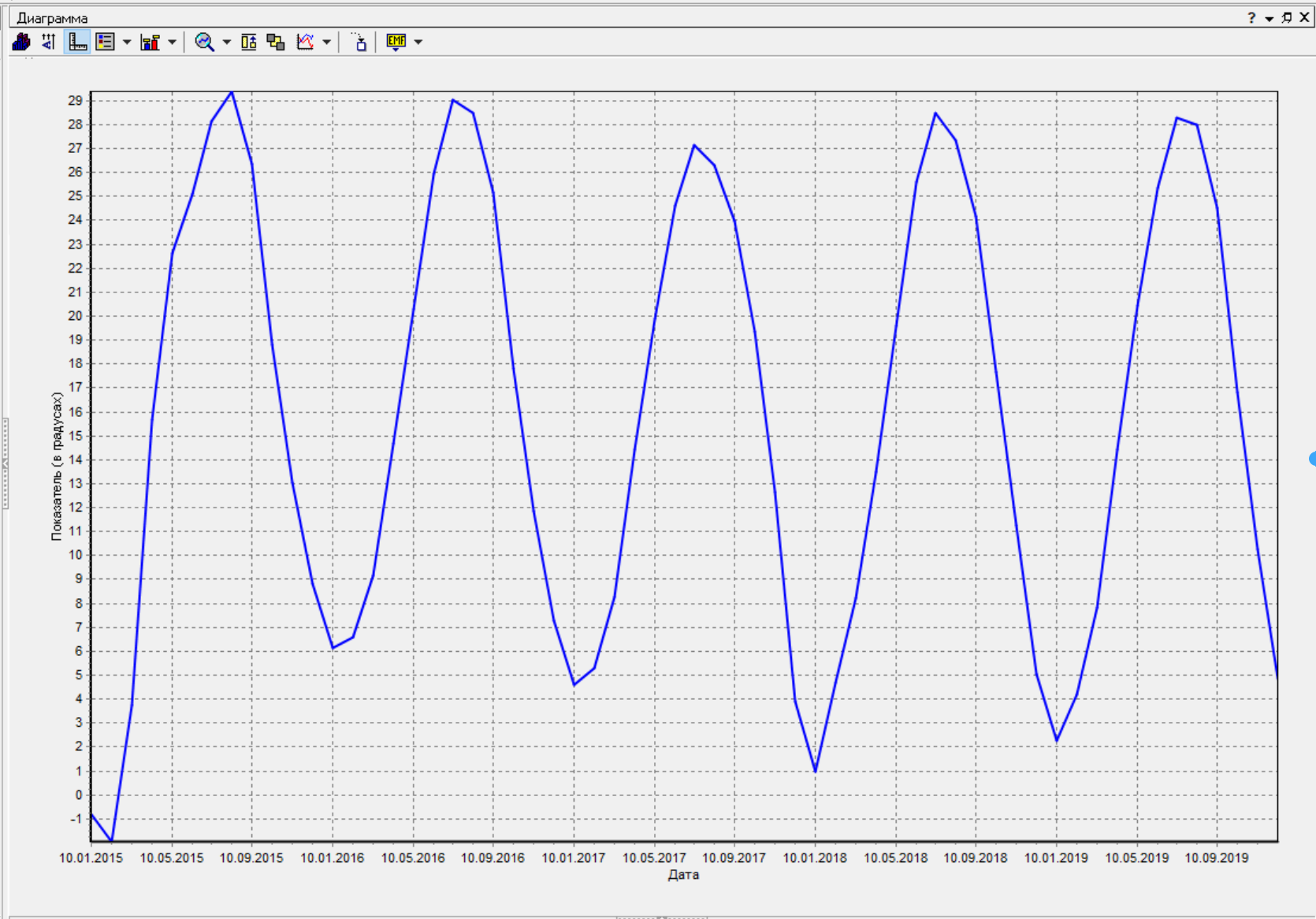


Рисунок 2.3.5 – Диаграмма после преобразований.

Далее были отобраны данные методом «Скользящее окно» с глубиной погружения 12 месяцев. Рисунок 2.3.6 демонстрирует параметры транспонирования столбца для использования метода «Скользящее окно», на рисунке 2.3.7 продемонстрированы результаты выполнения.

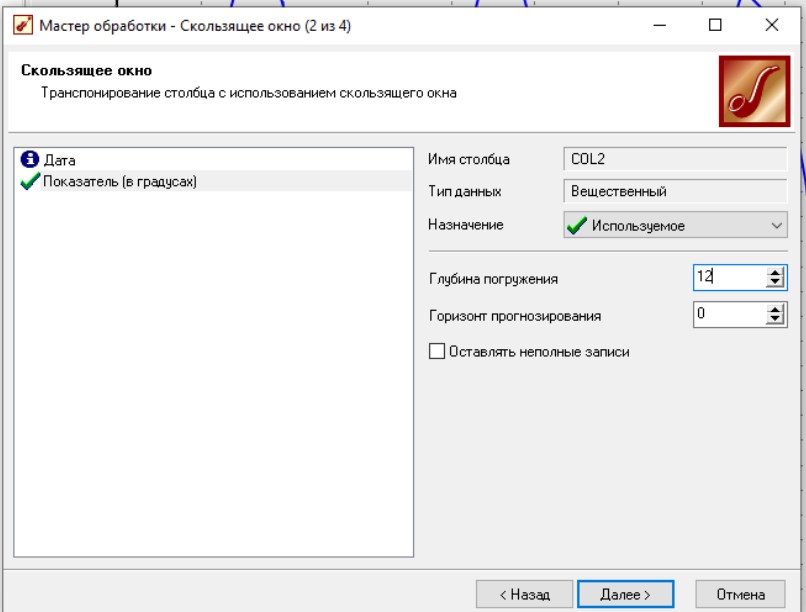


Рисунок 2.3.6 – Диаграмма после преобразований.

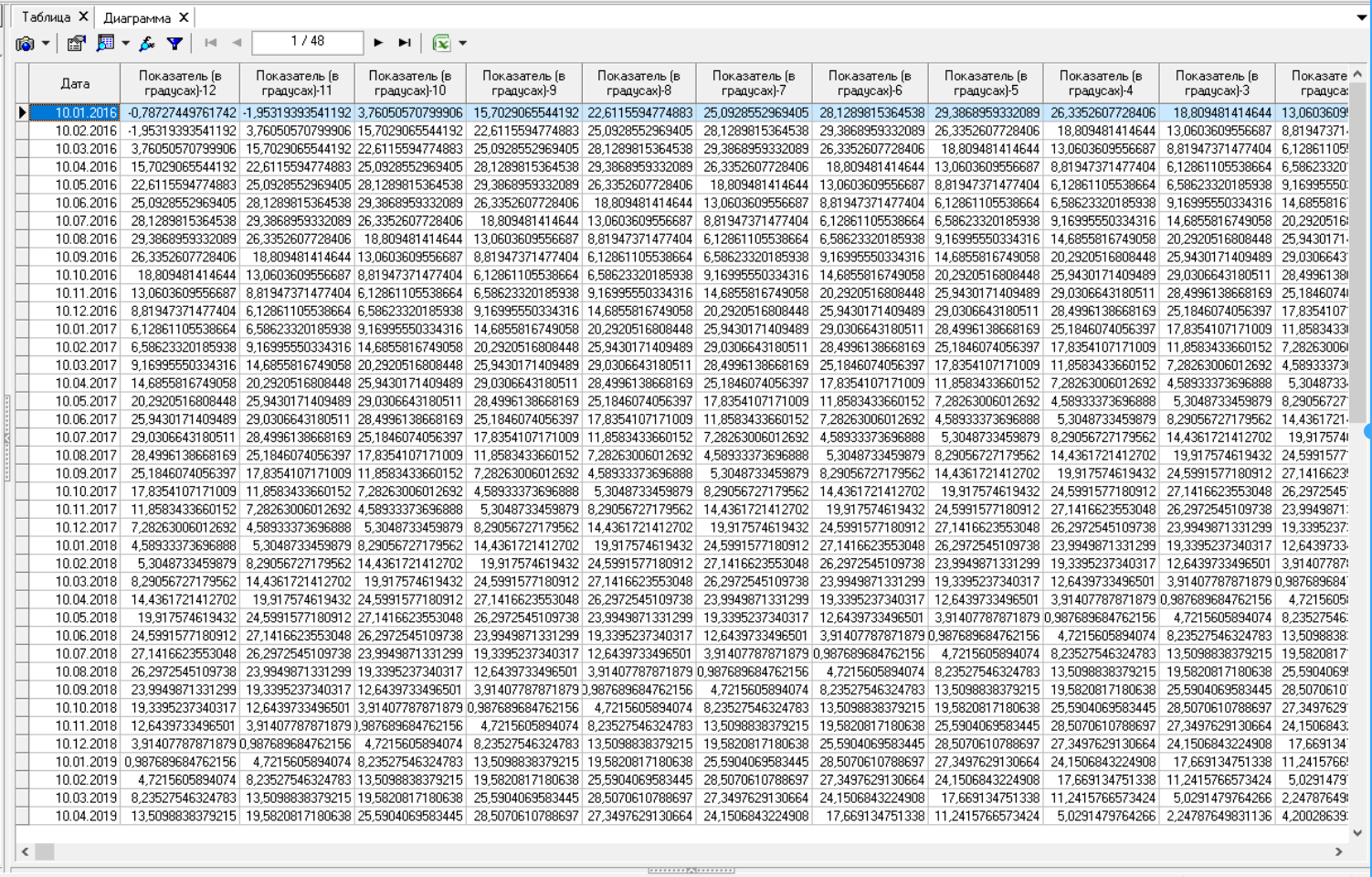


Рисунок 2.3.7 – Полученные результаты после использования метода «Скользящее окно»

Затем было произведено обучение нейросети с помощью инструмента «Нейросеть». Некоторые настройки представлены на рисунках 2.3.8 – 2.3.11

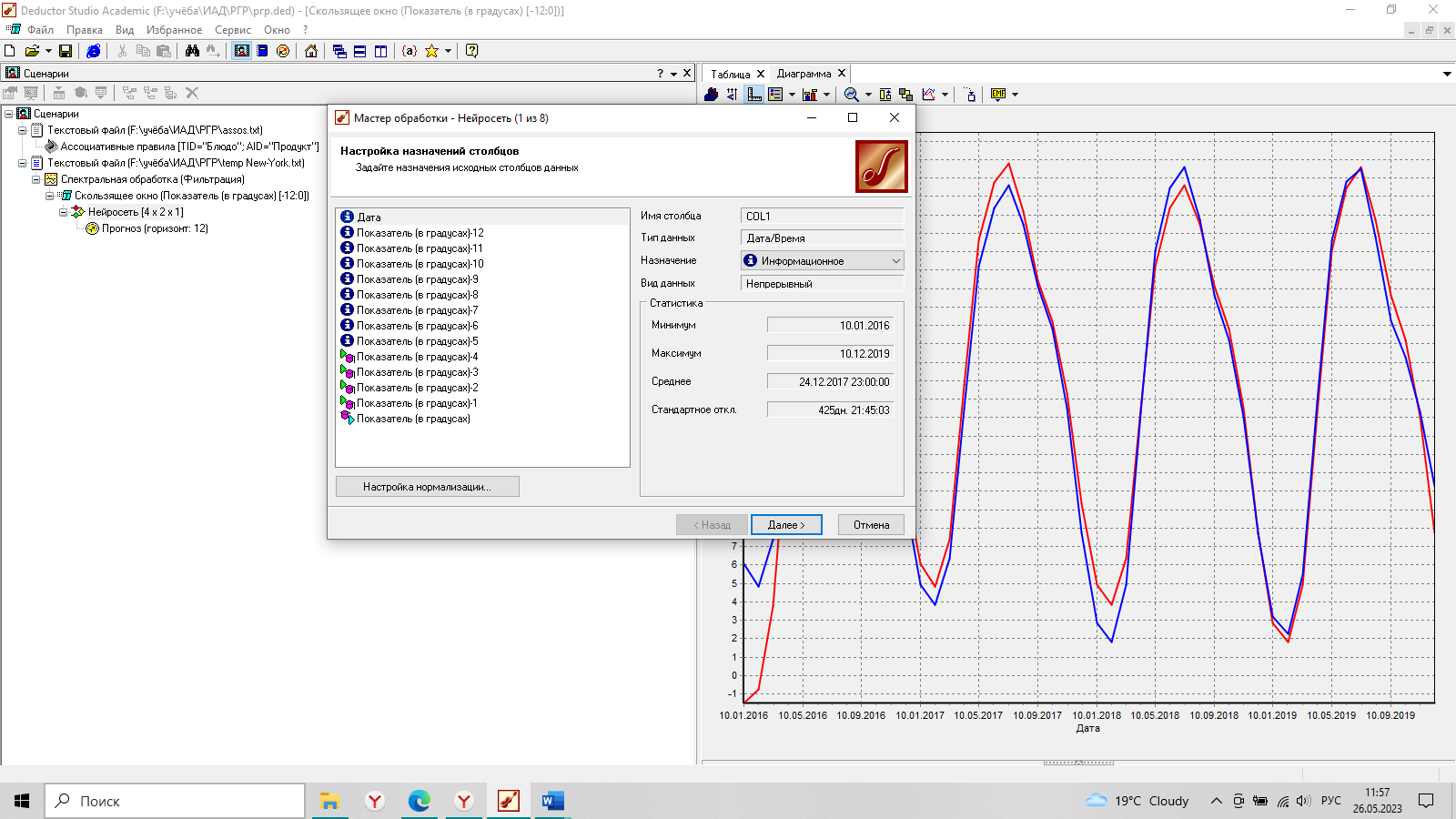


Рисунок 2.3.8 – Настройка назначений столбцов

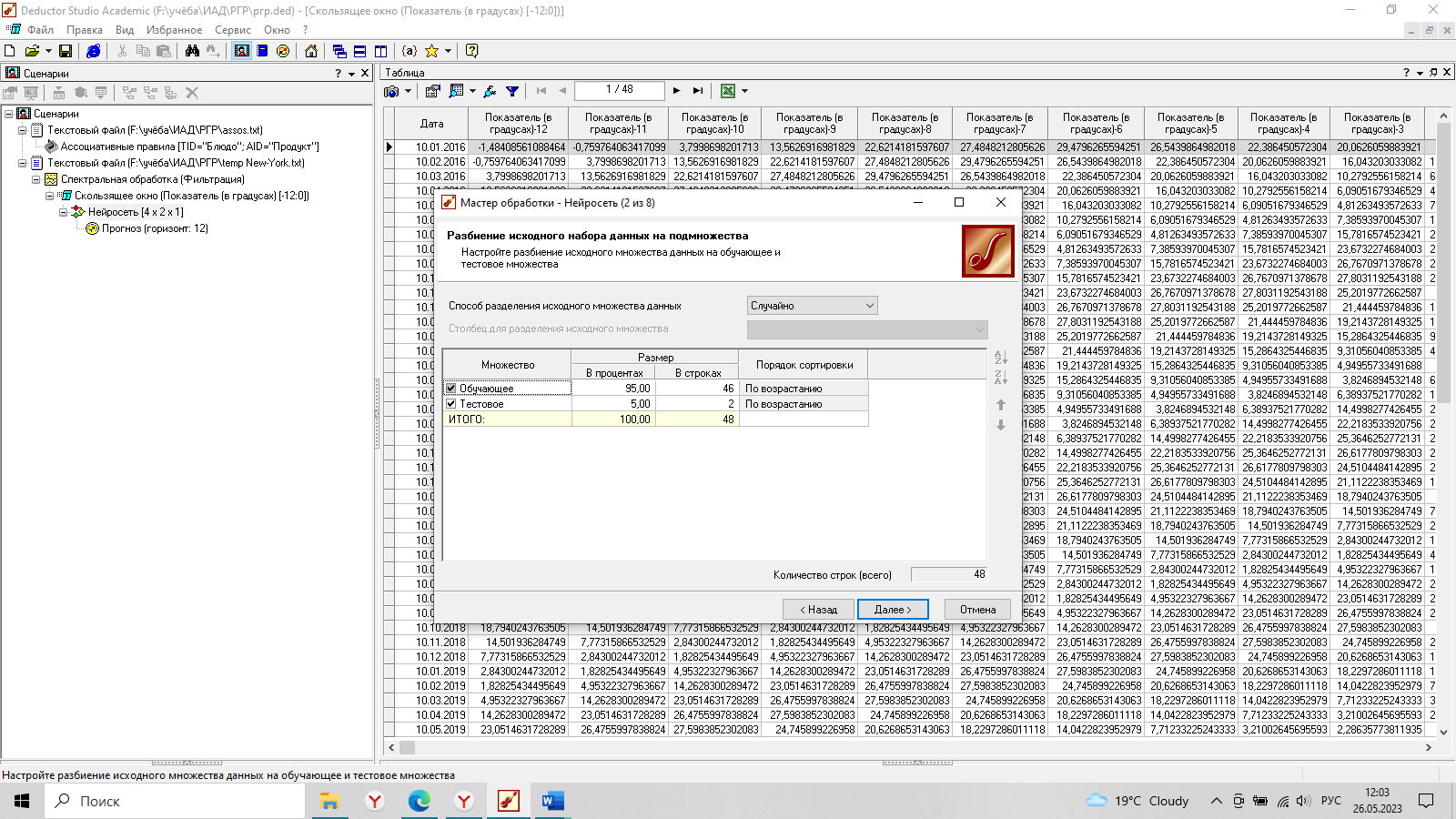


Рисунок 2.3.9 – Разбиение данных на обучающее и тестовое подмножества

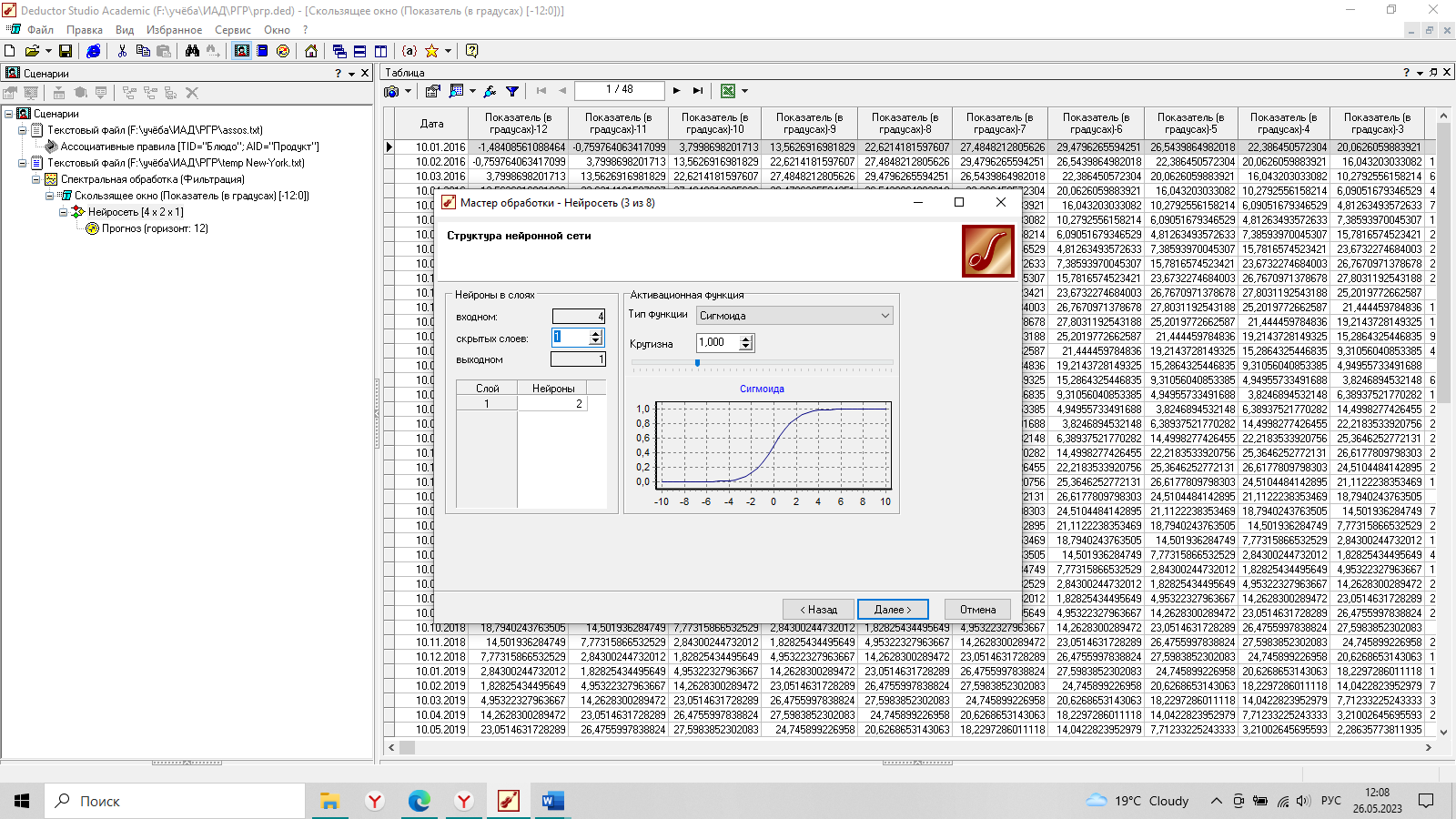


Рисунок 2.3.10 – Структура нейронной сети

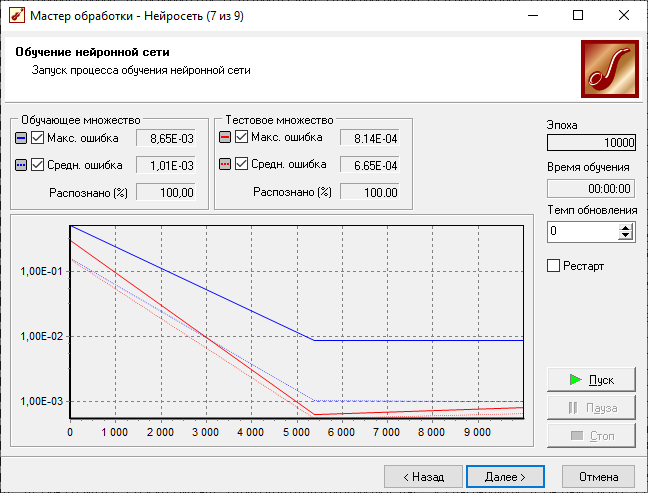


Рисунок 2.3.11 – Запуск обучения нейронной сети

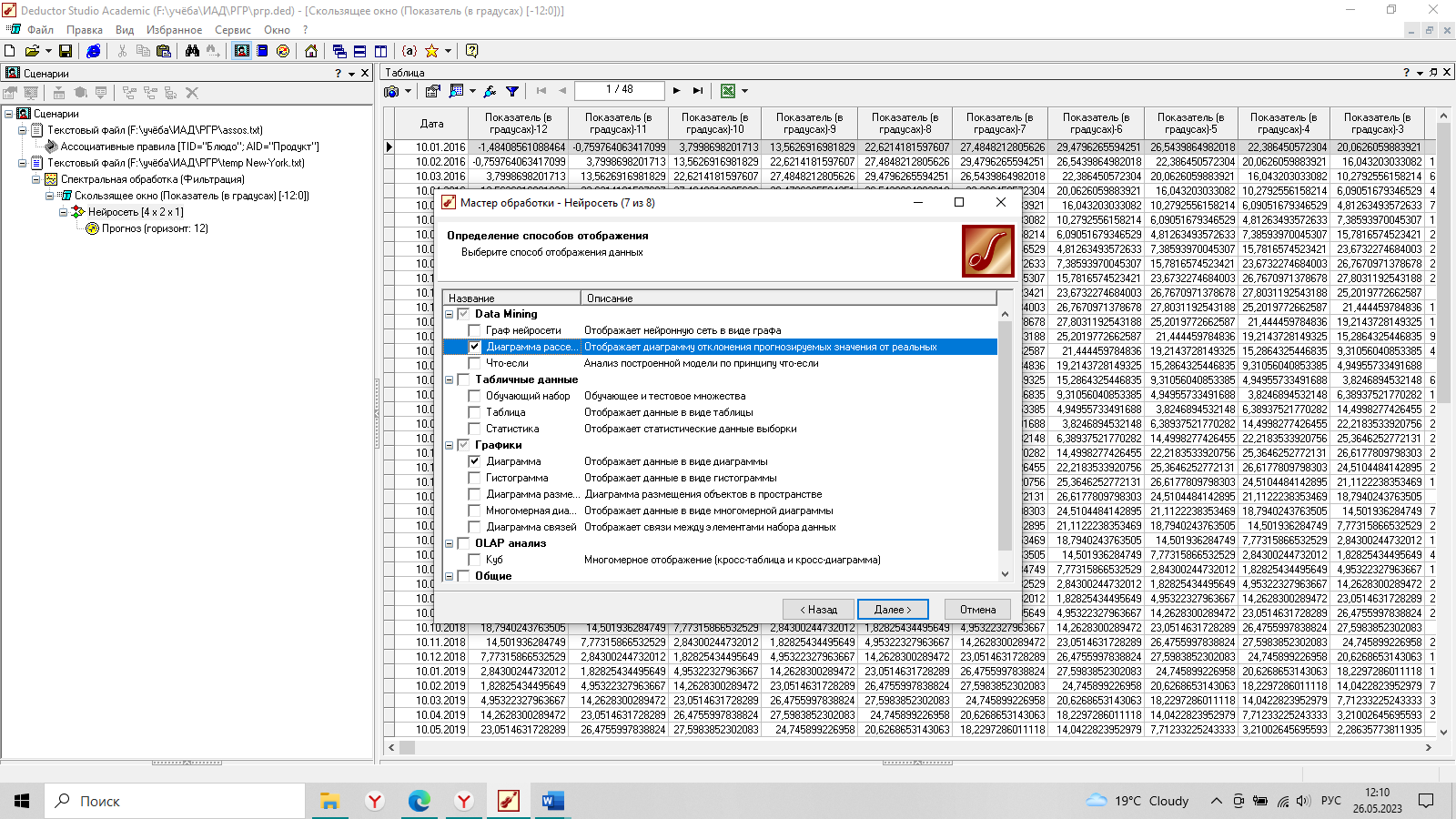


Рисунок 2.3.11 – Выбор отображения данных

На рисунке 2.3.12 представлен граф результатов обучения нейросети, рисунок 2.3.13 – диаграмма рассеяния, на рисунке 2.3.14 – диаграмма, где розовый график – реальное значение, а синий – спрогнозированное.

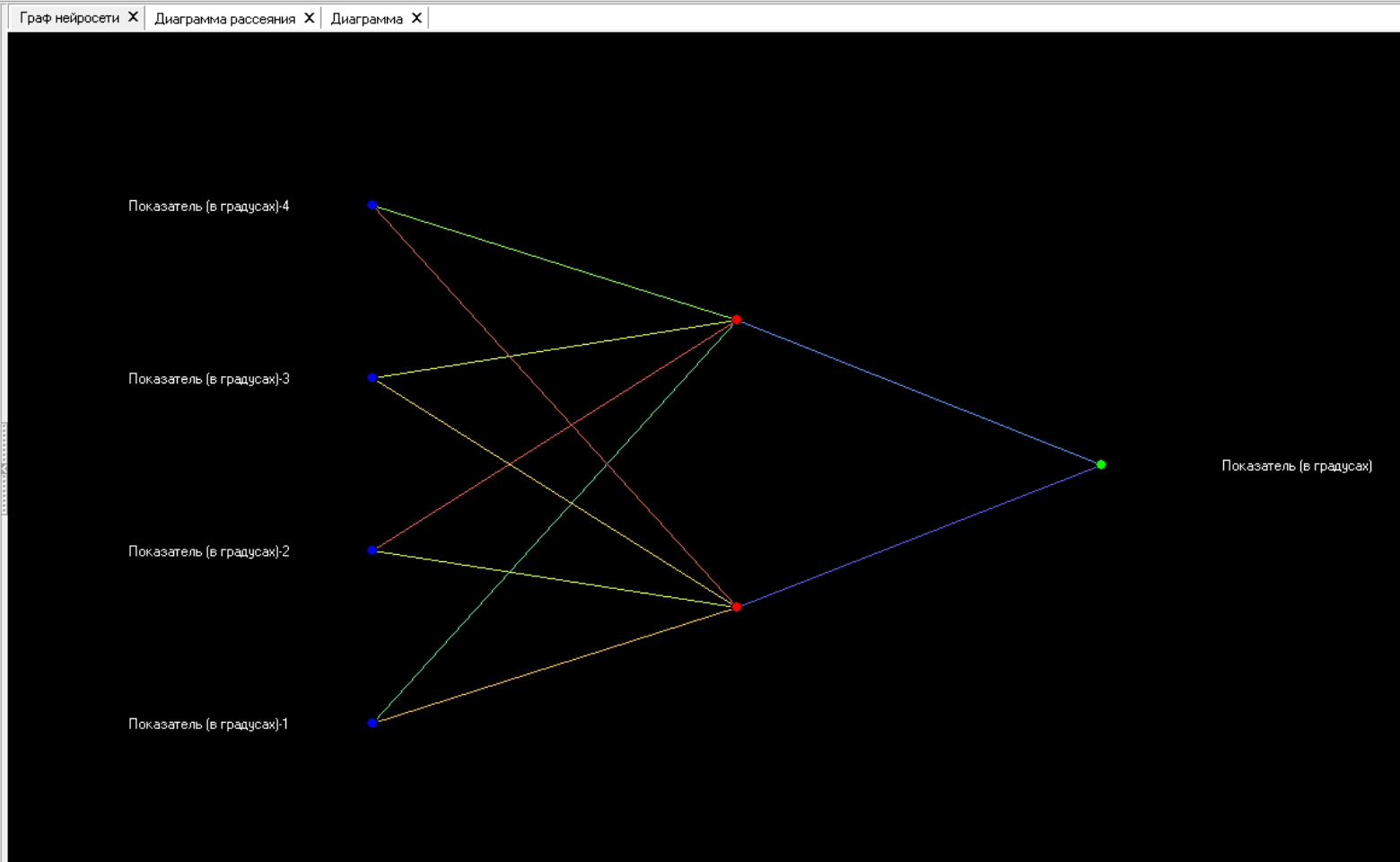


Рисунок 2.3.12 – Выбор отображения данных

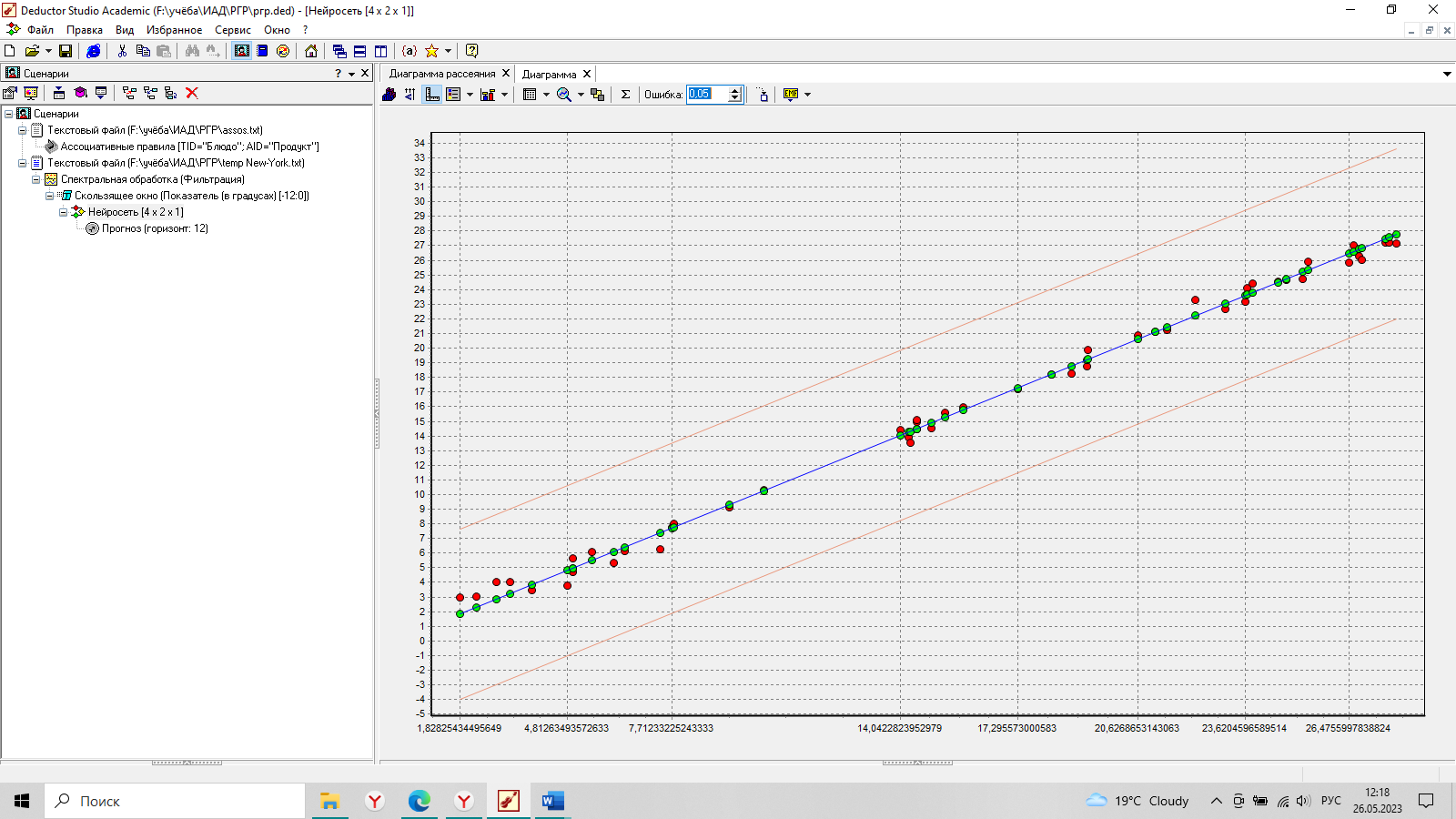


Рисунок 2.3.13 – Диаграмма рассеяния

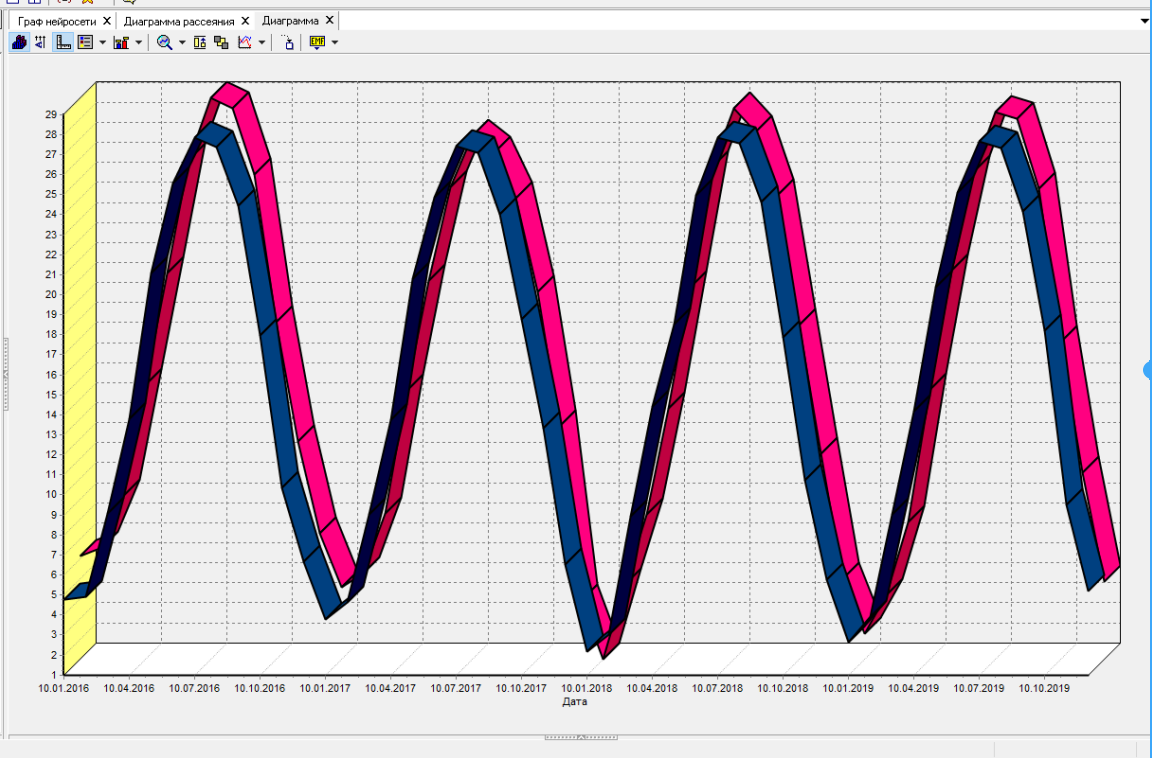


Рисунок 2.3.13 – Диаграмма прогноза

Далее был построен прогноз на следующие 12 месяцев (горизонт прогноза равен 12). Для этого был использован инструмент «Прогнозирование». Настройки представлены на рисунке 2.3.14. Результат отображался в виде диаграммы прогноза (рисунок 2.3.15).

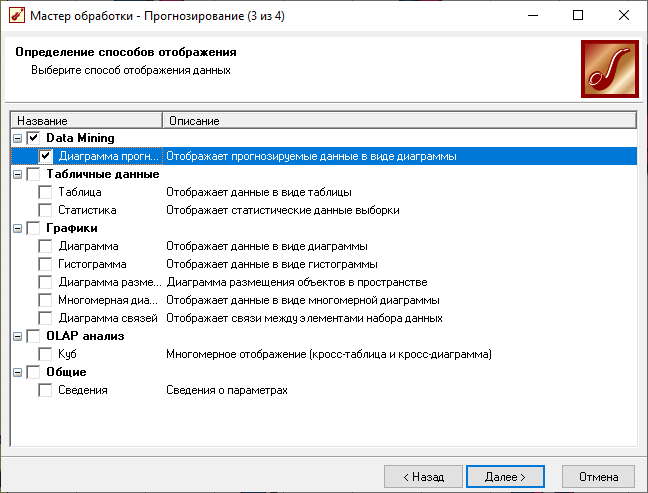
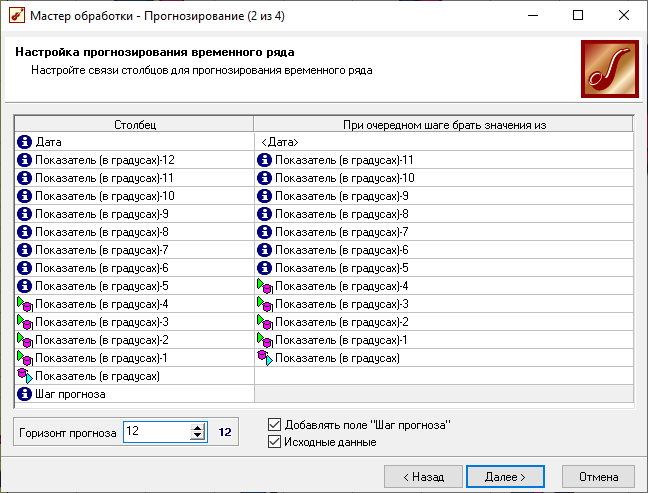


Рисунок 2.3.13 – Настройка параметров инструмента «Прогнозирование»



На диаграмме прогнозирования оранжевой линией отображено изменение температуры в городе Париж в ближайшие 12 месяцев. Согласно прогнозу температура будет изменятся от 5 до 27 градусов по Цельсию в ближайшие 6 месяцев, а в последующие 6 от 27 до 7 градусов.

Полученные значения были сравнены с фактическими и оба были занесены для удобства в таблицу, представленную в таблице 1.

Таблица 1 – Сравнение результатов

|  |  |  |
| --- | --- | --- |
| Месяц | Реальные | Экспериментальные |
| январь | 6 | 3 |
| февраль | 6 | 4,6 |
| март | 11 | 10 |
| апрель | 12 | 16,7 |
| май | 19 | 22 |
| июнь | 26 | 26 |
| июль | 30 | 28,2 |
| август | 28 | 26,5 |
| сентябрь | 23 | 22 |
| октябрь | 16 | 16 |
| ноябрь | 14 | 15 |
| декабрь | 6 | 5 |

Как можно видеть, в некоторых местах экспериментальные данные сильно расходятся с фактическими. Связано это со сглаживанием, произведенным в начале исследованием, а также с самим характером данных, поскольку погоду в целом трудно предсказать из-за воздействия на этот показатель других факторов.

Что касается самих данных, в них четко прослеживается присутствие сезонной компоненты. Также можно проследить возрастающий тренд – температура зимой за последние годы повышается.

Заключение

В ходе выполнения расчетно-графической работы средствами аналитической платформы Deductor Studio был произведен поиск ассоциаций в данных о рецептах итальянской кухни и анализ временного ряда о среднемесячной температуре в Нью-Йорке. В итоге исследований были получены следующие результаты.

После выполнения метода поиска ассоциаций были сделаны следующие выводы:

* самыми популярными продуктами в десертах являются «сахар», «масло сливочное» и «какао».
* если добавлен продукт «какао», то с вероятностью 72,2% будет добавлен продукт «масло сливочное», с вероятностью 81,8% будет добавлен «сахар», и с вероятностью в 63,6% – оба продукта.
* чтобы добавить какао, с вероятностью 71,4% в рецепте должна также быть вода.

Далее, после выполнения прогнозирования временного ряда на 12 месяцев вперед были сделаны выводы о присутствии сезонной компоненты и возрастающего тренда. Прогноз был приближен к реальным данным.

По завершении выполнения данной расчетно-графической работы были достигнуты все поставленные цели, а также изучены основы вышеупомянутых методов анализа экспериментальных данных и освоены техники их практического применения в Deductor Studio

.СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. G. Piatetsky-Shapiro. Статья: Что такое data mining? [Электронный ресурс] / G. Piatetsky-Shapiro. – Электрон. текстовые дан. – Режим доступа: http://bizoffice.ru/crm/arhiv/chto-takoe-data-mining.html, свободный.

2. Десерты, фото с рецептами на RussianFood.com: [сайт]. URL: https://www.russianfood.com/recipes/bytype/?fid=45

3. Погода в Париже по месяцам: [сайт]. URL: https://world-weather.ru/pogoda/france/paris/2020/