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-20-1-о\_\_\_

Скороходова В.А.\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Проверил:** ст. пр. Сырых О.А.\_ \_

**Севастополь**

2023 г.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc136379764)

[1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 5](#_Toc136379765)

[2 ПРАКТИЧЕСКАЯ ЧАСТЬ 9](#_Toc136379766)

[2.1. Решение задачи поиска ассоциаций в Deductor. 9](#_Toc136379767)

[2.1.1. Популярные наборы 12](#_Toc136379768)

[2.1.2. Правила 13](#_Toc136379769)

[2.1.3. Дерево правил 13](#_Toc136379770)

[2.1.4. Что-если 14](#_Toc136379771)

[2.2. Прогнозирование временного ряда 15](#_Toc136379772)

[ЗАКЛЮЧЕНИЕ 25](#_Toc136379773)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 27](#_Toc136379774)

# ВВЕДЕНИЕ

Задачи прогнозирования решаются в самых разнообразных областях человеческой деятельности, таких как наука, экономика, производство и множество других сфер. Прогнозирование является важным элементом организации управления как отдельными хозяйствующими субъектами, так и экономики в целом. Развитие методов прогнозирования непосредственно связано с развитием информационных технологий, в частности, с ростом объемов хранимых данных и усложнением методов и алгоритмов прогнозирования, реализованных в инструментах Data Mining [1].

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. Подготовка данных для поиска ассоциативных правил.

2. Проведение поиска ассоциативных правил, анализ полученных результатов.

3. Подготовка данных для прогнозирования временного ряда.

4. Проведение анализа временного ряда, проведение прогнозирования временного ряда на 12 месяцев.

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 итальянской кухни 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 Нью-Йорке 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.

# 1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Глубокие исследования данных проводятся с помощью специальных методов – методов интеллектуального анализа (ИА) или извлечения данных (Data Mining), которые включают в себя поиск закономерностей и зависимостей между данными с целью поиска шаблонов. Особенность Data Mining – необычность разыскиваемых шаблонов, скрытые знания. По определению одного из основателей этого направления Г. Пиатецкого-Шапиро: «Data Mining – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1].

В интеллектуальном анализе данных при помощи математического анализа выявляются закономерности и тенденции, которые существуют в данных и которые нельзя обнаружить при традиционном просмотре данных из-за их чрезмерного объема или по причине сложности связей.

Существует несколько методов Data Mining [1], которые могут использоваться в конкретных задачах, а именно:

1. Ассоциативные правила (Association Rules) − метод поиска скрытых связей между различными элементами данных. Ассоциативные правила могут использоваться для анализа покупательского поведения, выявления связей между симптомами и заболеваниями, определения связей между словами в тексте и т.д.
2. Анализ временных рядов (Time Series Analysis)  метод анализа данных, которые являются функцией времени. Анализ временных рядов может использоваться для прогнозирования временных рядов, таких как продажи, температура, финансовые показатели, и т.д.
3. Кластерный анализ (Cluster Analysis) − метод разделения данных на группы (кластеры) на основе сходства между наблюдениями. Кластерный анализ может использоваться для сегментации клиентов, выделения групп товаров, выявления аномалий в данных и т.д.
4. Классификация (Classification) − метод построения модели, которая может предсказывать принадлежность объекта к определенному классу на основе известных признаков. Классификация может использоваться для решения задач, таких как распознавание образов, выявление мошеннических операций на банковских счетах, определение токсичности сообщений в социальных сетях и т.д.
5. Регрессионный анализ (Regression Analysis) − метод, который позволяет оценить влияние одной или нескольких независимых переменных на зависимую переменную. Регрессионный анализ может использоваться для прогнозирования продаж, оценки влияния рекламных кампаний на продажи, определения влияния факторов на здоровье и т.д.

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

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий "Хлеб", приобретет и "Молоко". Впервые эта задача была предложена для поиска ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины.

Все множество ассоциативных правил можно разделить на три вида [4]:

1. Полезные правила – содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение. Такие правила могут быть использованы для принятия решений, приносящих выгоду.
2. Тривиальные правила – содержат действительную и легко объяснимую информацию, которая уже известна. Такие правила, хотя и объяснимы, но не могут принести какой-либо пользы, т.к. отражают или известные законы в исследуемой области, или результаты прошлой деятельности. При анализе рыночных корзин в правилах с самой высокой поддержкой и достоверностью окажутся товары-лидеры продаж. Практическая ценность таких правил крайне низка.
3. Непонятные правила – содержат информацию, которая не может быть объяснена. Такие правила могут быть получены или на основе аномальных значений, или глубоко скрытых знаний. Напрямую такие правила нельзя использовать для принятия решений, т.к. их необъяснимость может привести к непредсказуемым результатам. Для лучшего понимания требуется дополнительный анализ.

Для поиска обычных ассоциативных правил в программе Deductor Studio служит обработчик «Ассоциативные правила».

Временным рядом (рядом динамики, динамическим рядом) называется последовательность значений показателя или признака, упорядоченная в хронологическом порядке, т.е. в порядке возрастания временного параметра. Отдельные наблюдения временного ряда называются уровнями этого ряда.

Каждый временной ряд содержит два элемента:

1) значения времени;

2) соответствующие им значения уровней ряда.

Прогнозирование результата на определенное время вперед, основываясь на данных за прошедшее время, – задача, встречающаяся довольно часто. К примеру, перед большинством торговых фирм стоит задача оптимизации складских запасов, для решения которой требуется знать, что и сколько должно быть продано через неделю и т.п., задача предсказания стоимости акций какого-нибудь предприятия через день и т.д. и другие подобные вопросы. Deductor Studio предлагает для этого инструмент «Прогнозирование».

# 2 ПРАКТИЧЕСКАЯ ЧАСТЬ

## 2.1. Решение задачи поиска ассоциаций в Deductor.

Был подготовлен набор данных для метода поиска ассоциаций – список рецептов итальянской кухни [3]. Как выглядят данные, иллюстрирует рисунок 1.

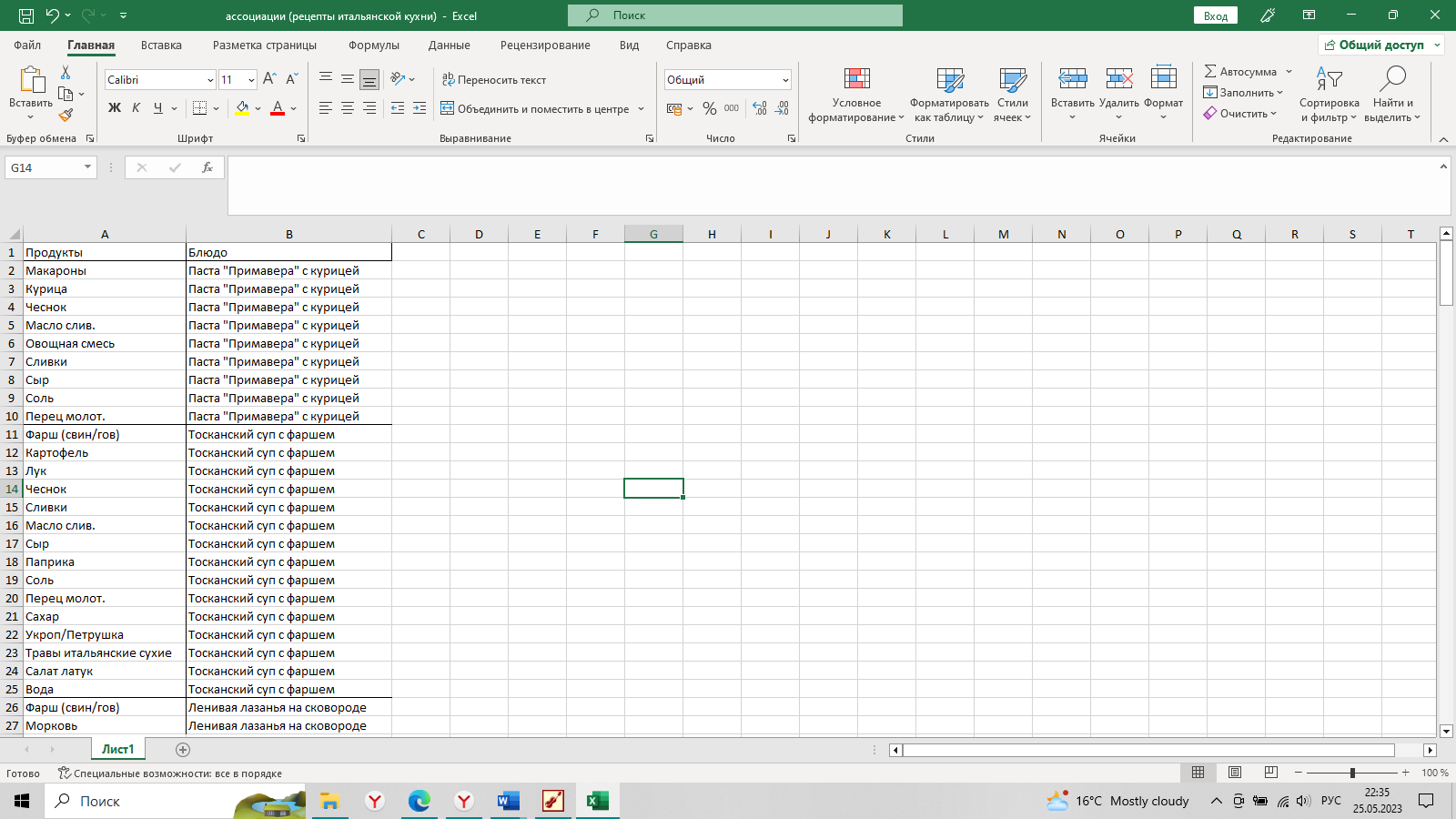


Рисунок 1 – Набор данных для поиска ассоциаций

Данные из Excel были преобразованы в файл с расширением .txt и импортированы в Deductor Studio. Импорт данных изображен на рисунке 2.

В результате импорта данные в Deductor были представлены в виде таблицы как на рисунке 3. Столбцу «Продукт» присвоено назначение элемента транзакции, а столбцу «Блюдо» назначение индикатора транзакции.

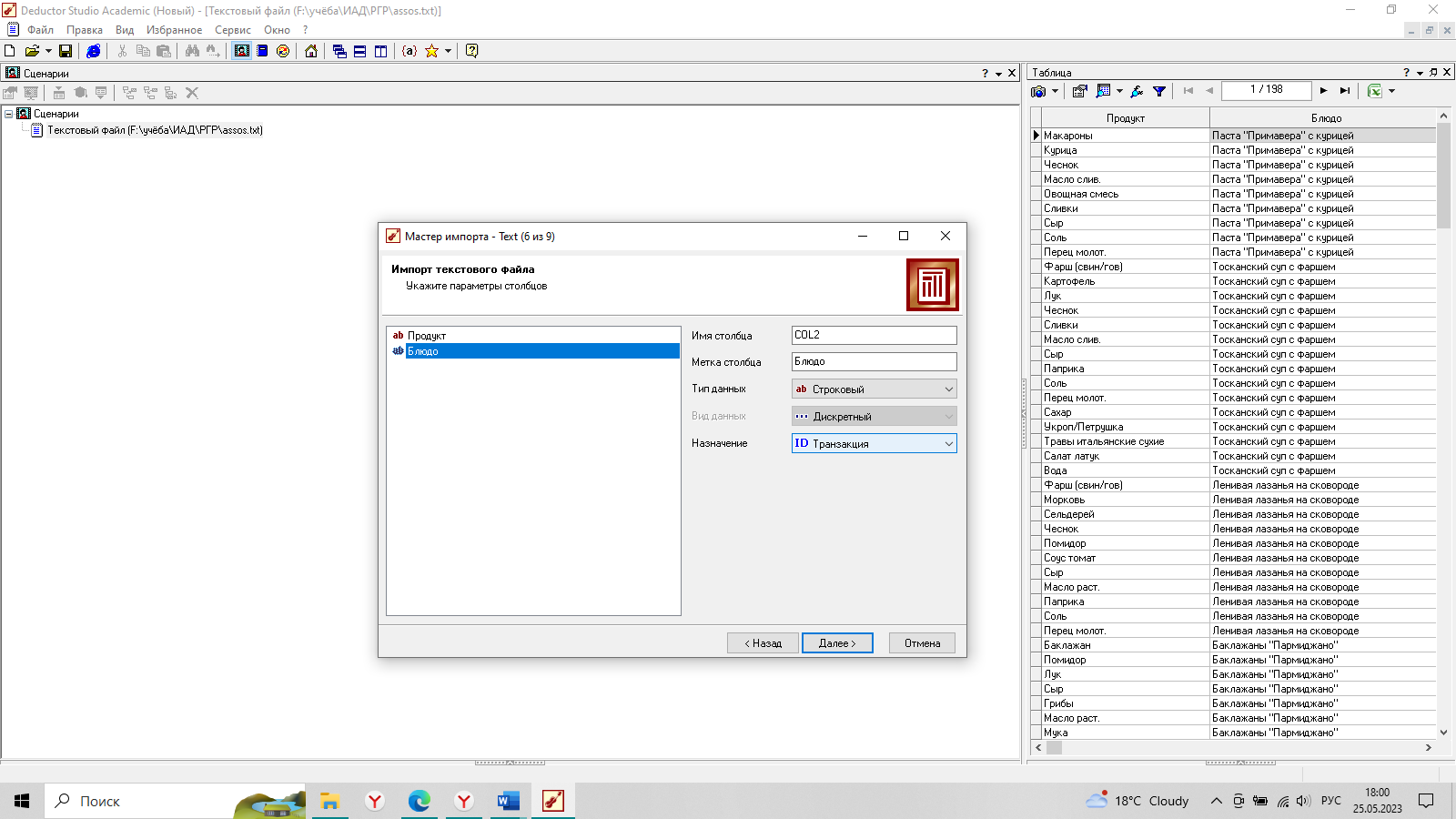
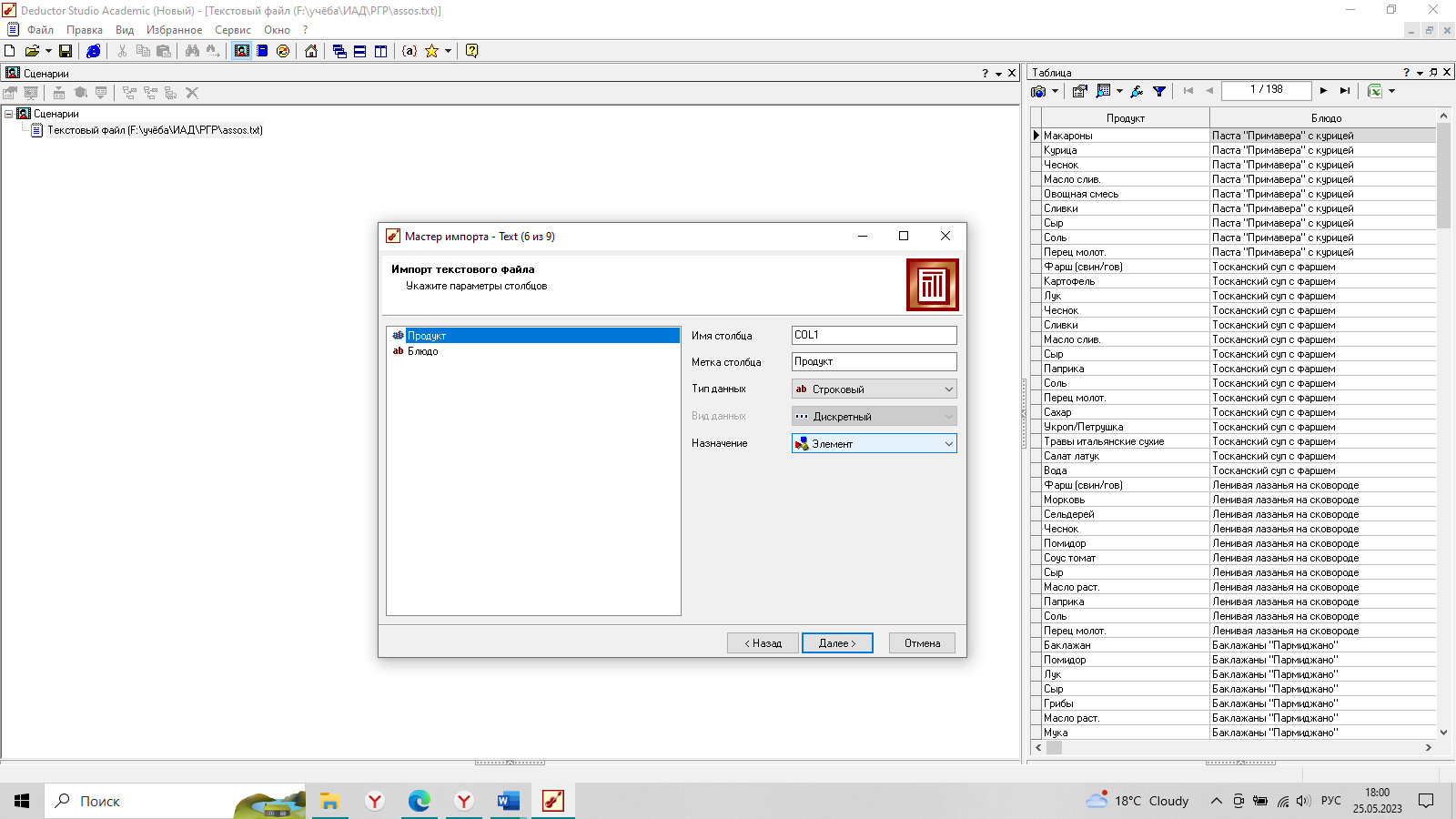
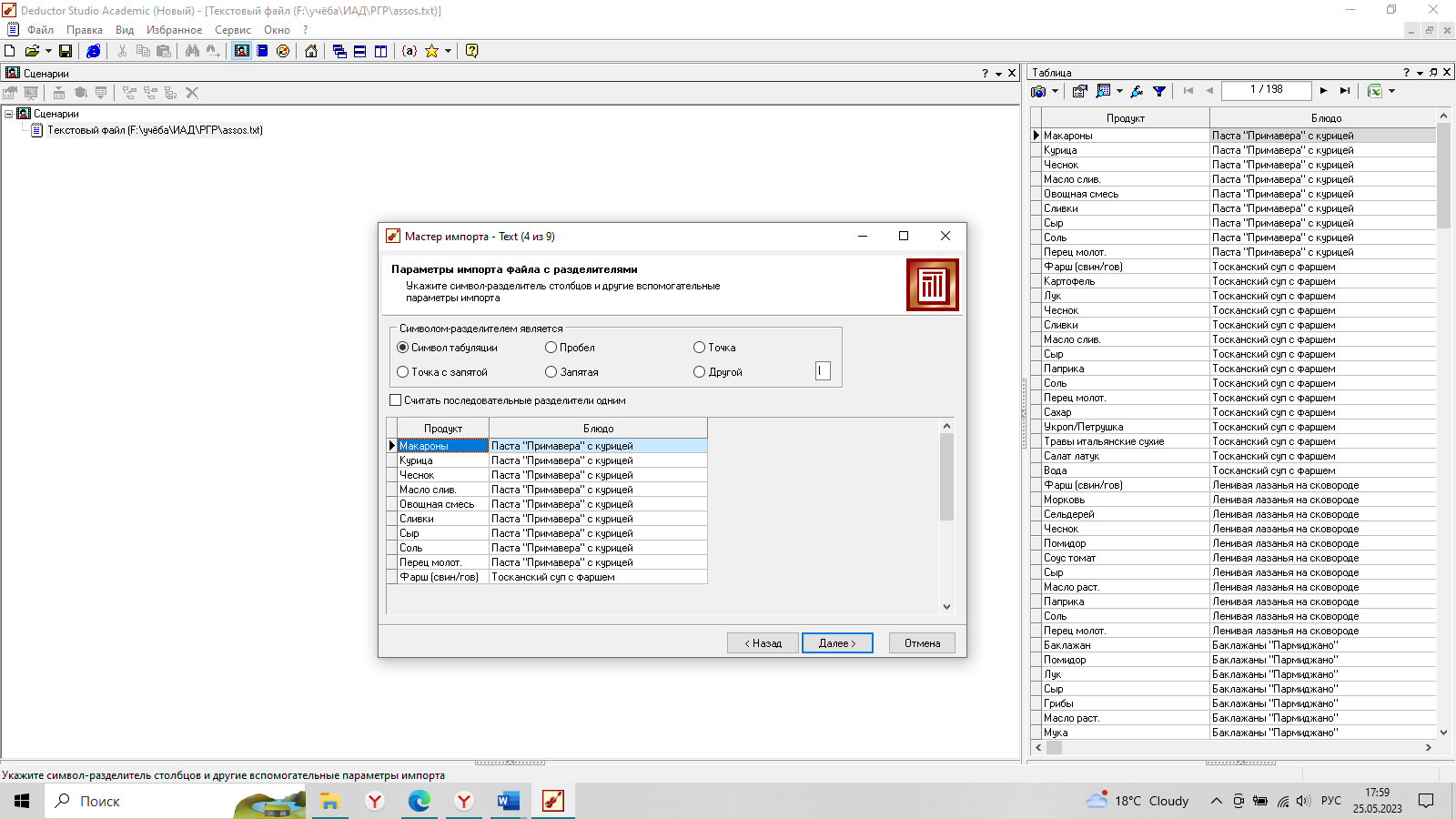
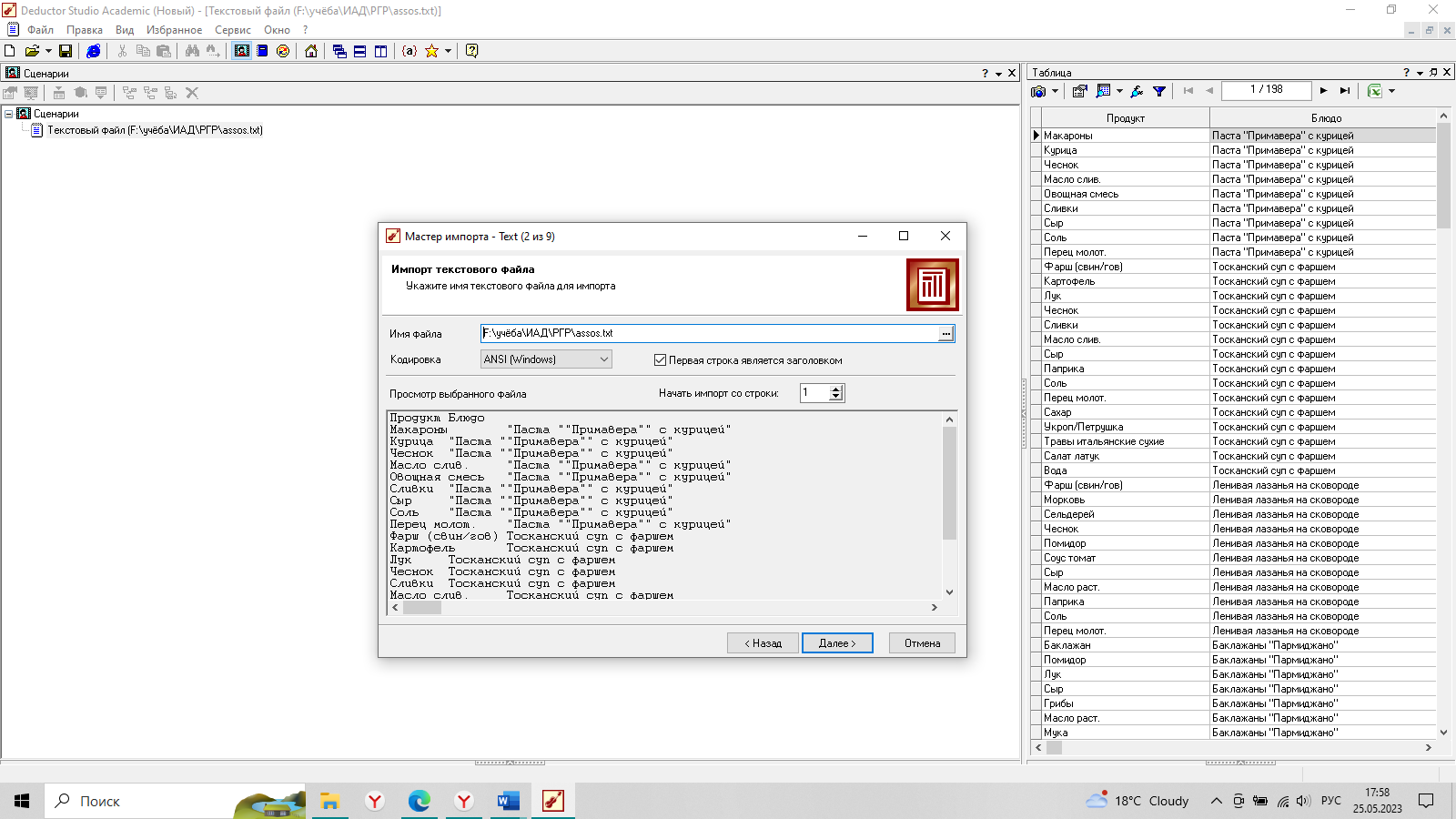


Рисунок 2 – Процесс импорта данных

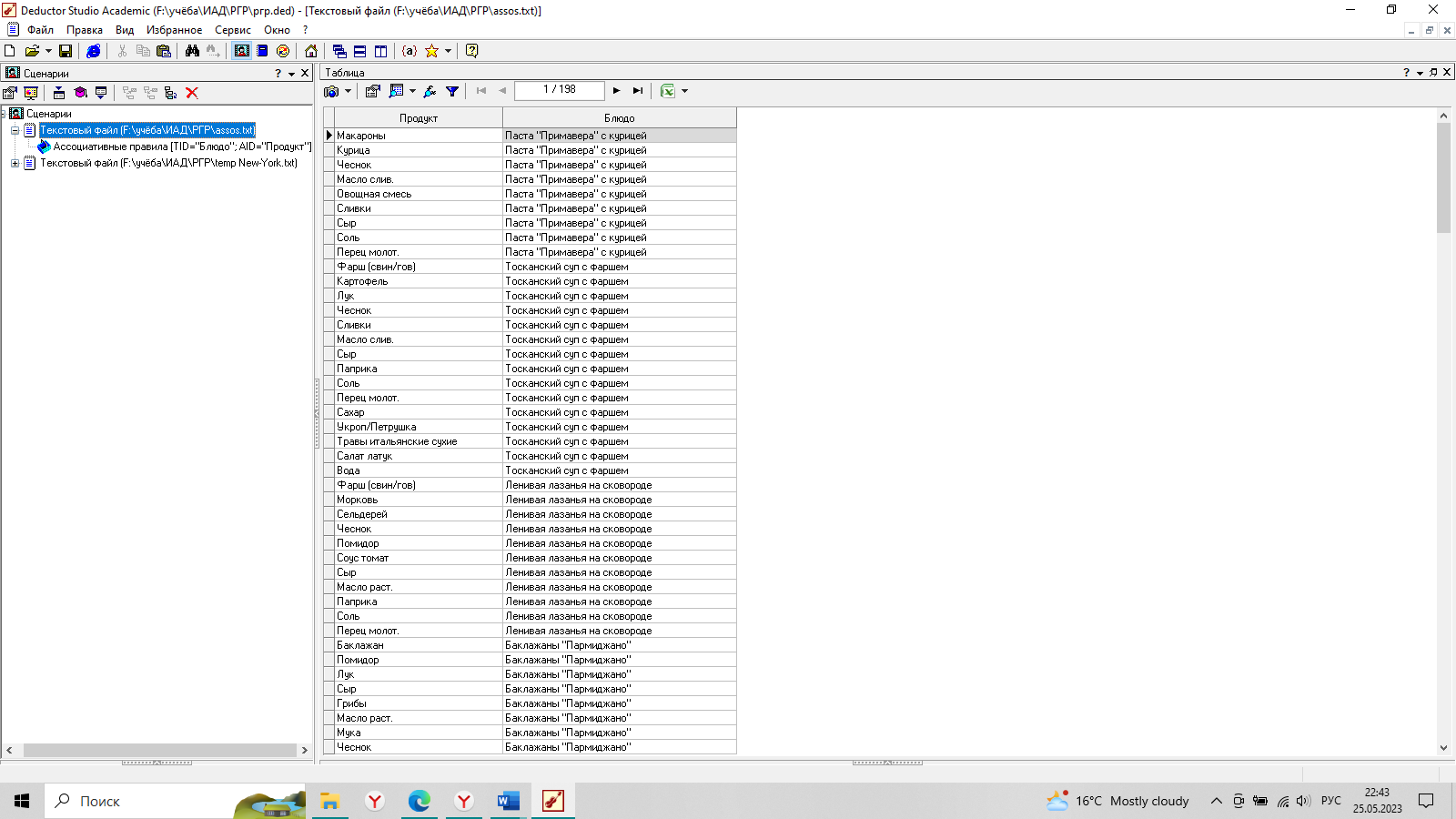


Рисунок 3 – Представление данных в Deductor

Далее с помощью «Мастера обработки» в разделе Data Mining был выбран пункт «Ассоциативные правила».

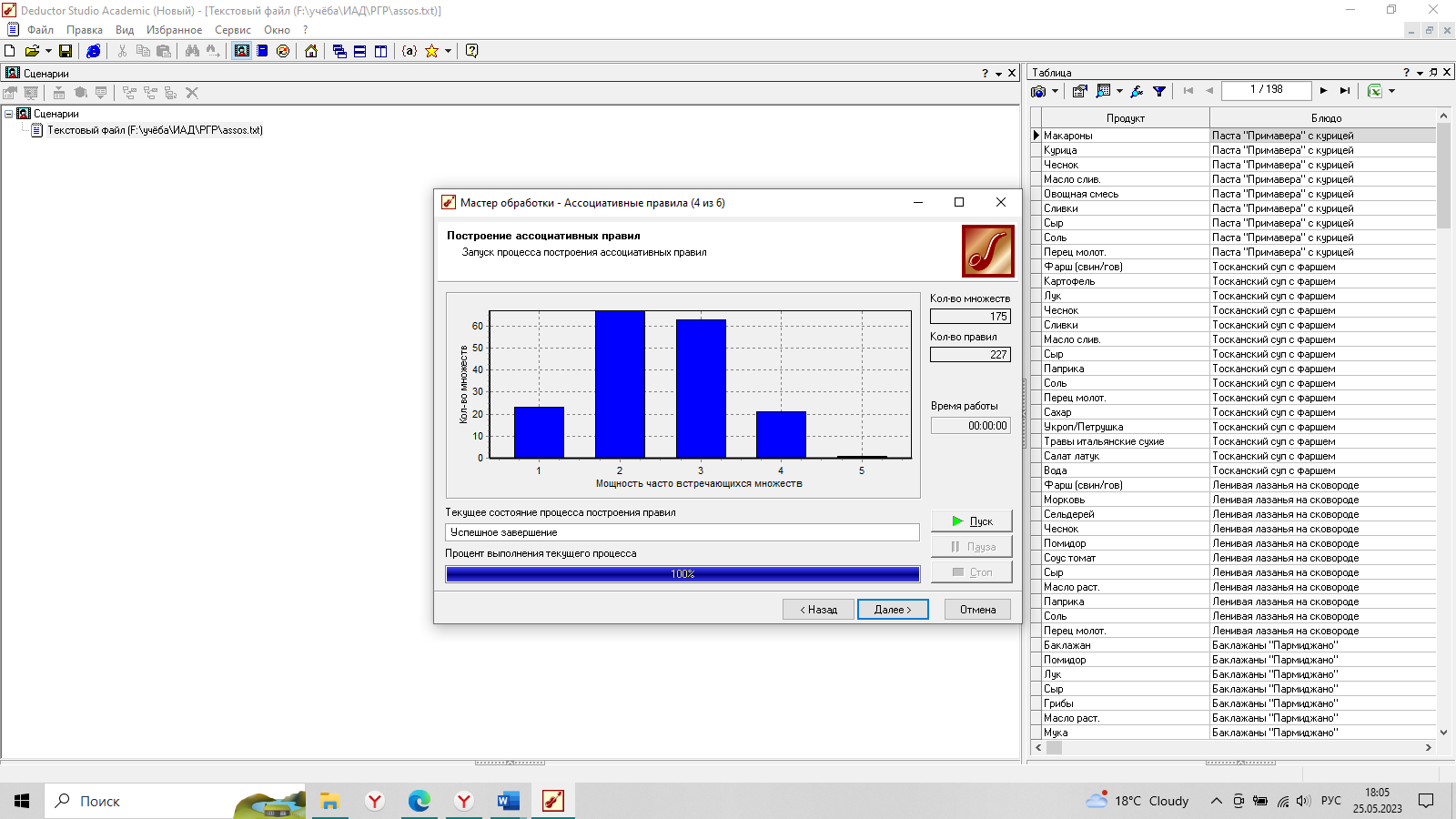
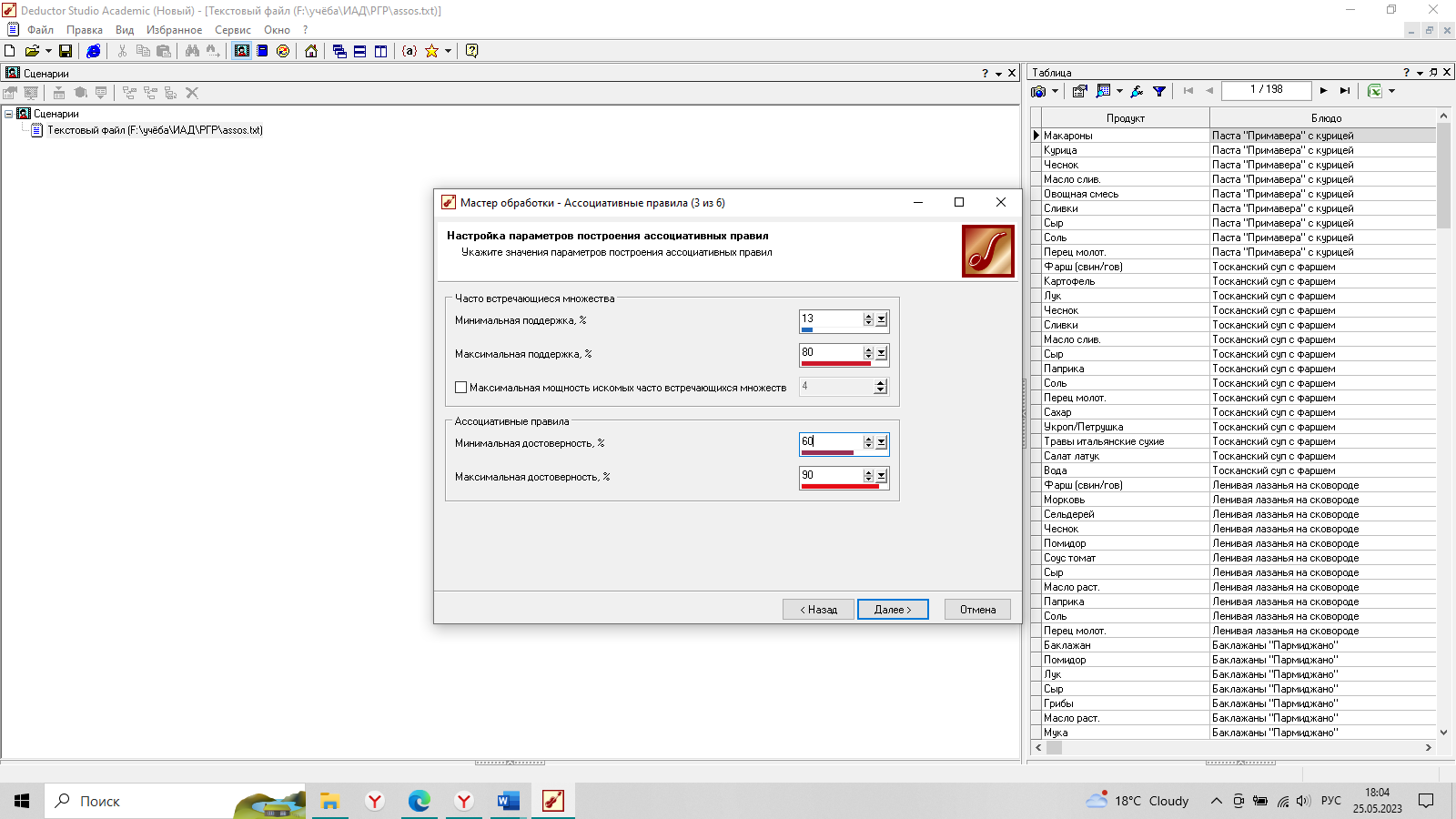
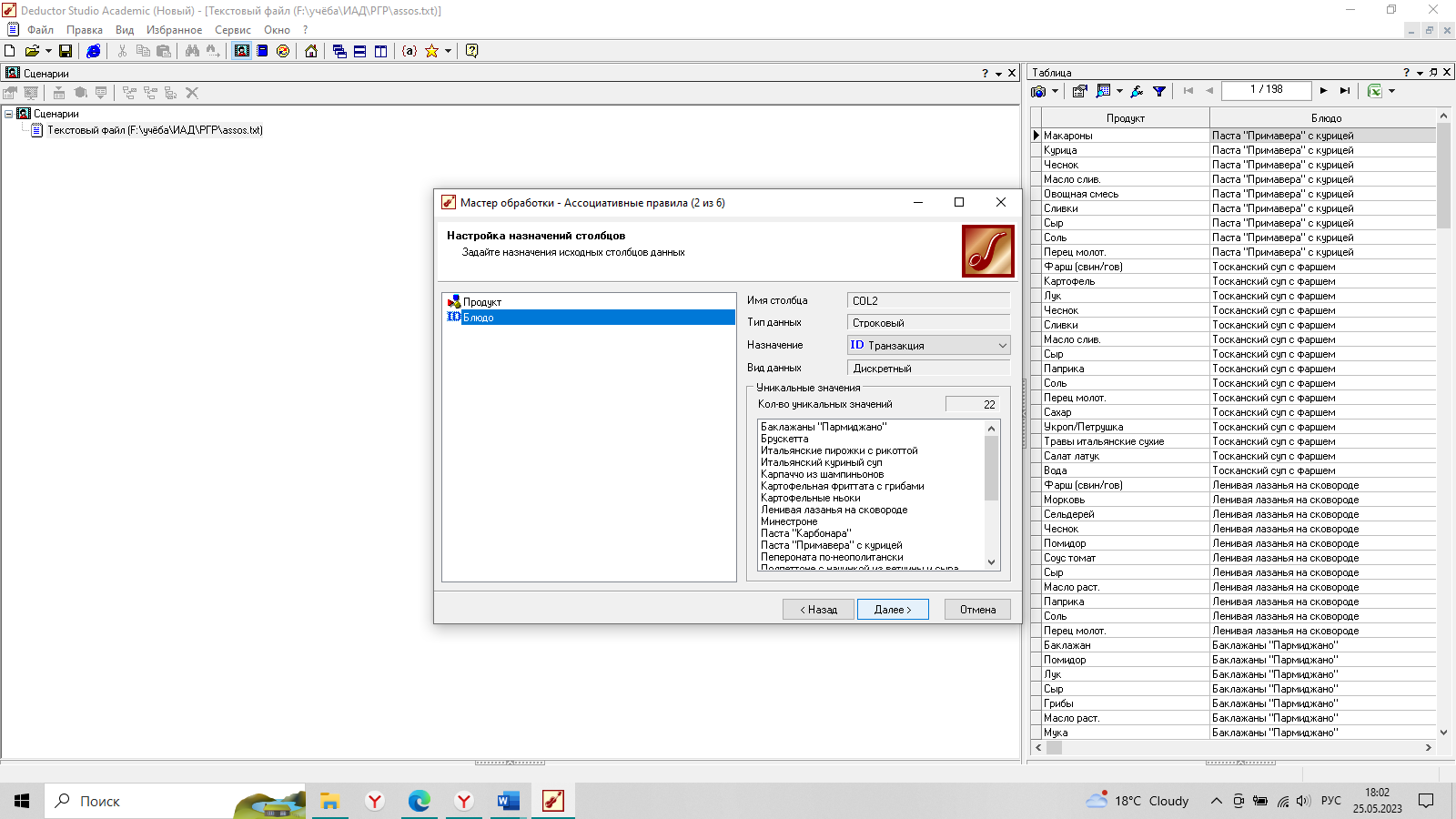


Рисунок 4 – Проведение поиска ассоциативных правил

Был проведен поиск ассоциаций, представленный на рисунке 4. Были настроены параметры построения ассоциативных правил: минимальная и максимальная поддержка, минимальная и максимальная достоверность, а также максимальная мощность множества. Эти параметры необходимо выставлять исходя из характера имеющихся данных. Например, границы поддержки следует указать – 13% и 80% и достоверности 60% и 90%. Количество множеств равно 175, количество правил – 227.

Способы отображения были выбраны все из секции Data Mining, а именно: «правила», «популярные наборы», «дерево правил», «что-если».

### 2.1.1. Популярные наборы

Популярные наборы – это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно. Насколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Данный визуализатор отображает множества в виде списка (рисунок 5).

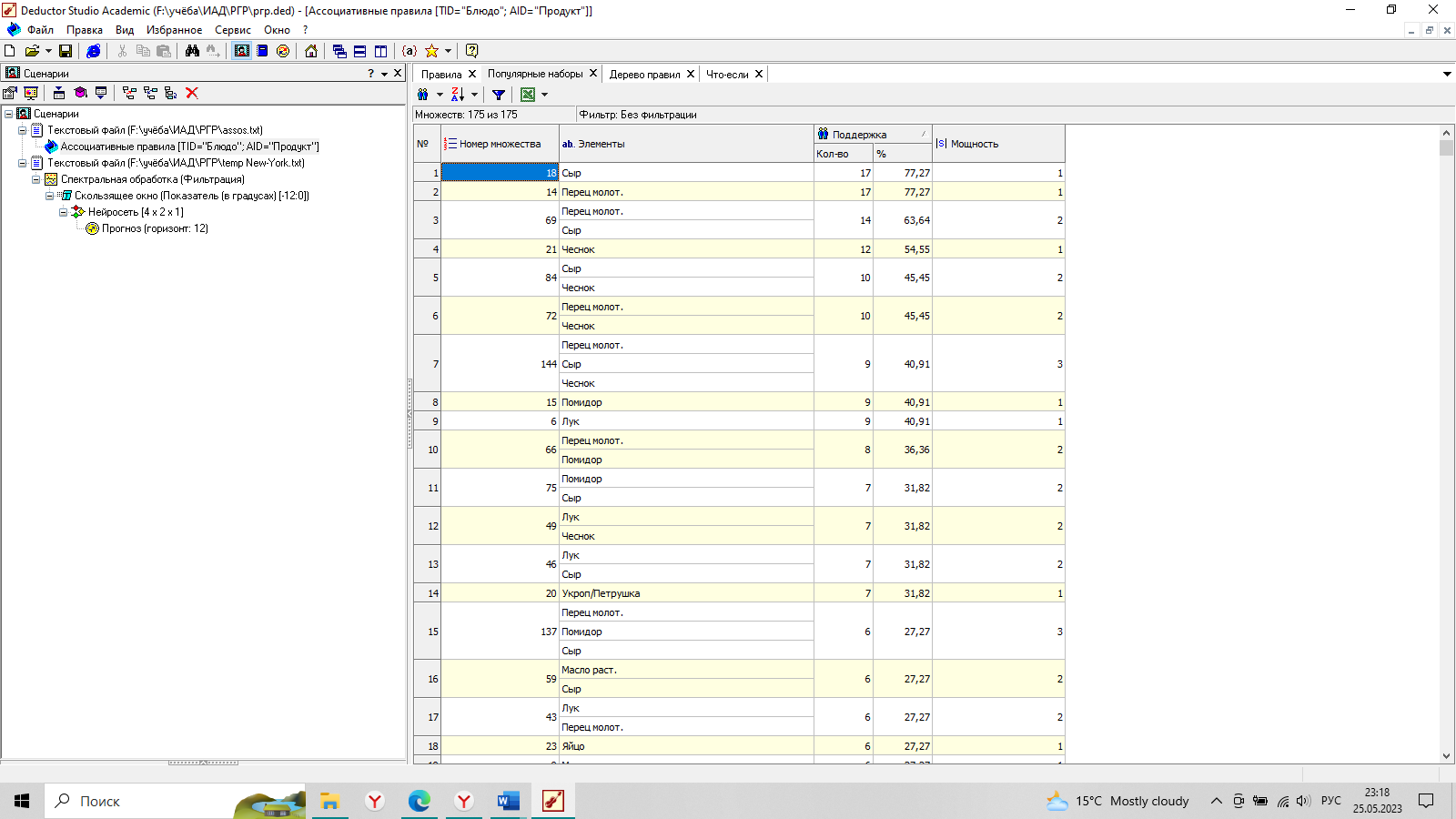


Рисунок 5 – Популярные наборы

Исходя из рисунка 5, можно заметить, что такие продукты, как сыр, перец молот., чеснок, помидор, лук, наиболее часто используются в приготовлении блюд.

### 2.1.2. Правила

Визуализатор «Правила» отображает ассоциативные правила в виде списка правил, описывающих поведение чего-либо.

По результатам, изображенным на рисунке 6, можно сделать следующий вывод: если блюдо содержит в составе сыр и чеснок, то с вероятность 90% в него будет добавлен и перец, а также если в составе будет яйцо, то с вероятностью 83,33% будет добавлен сыр.

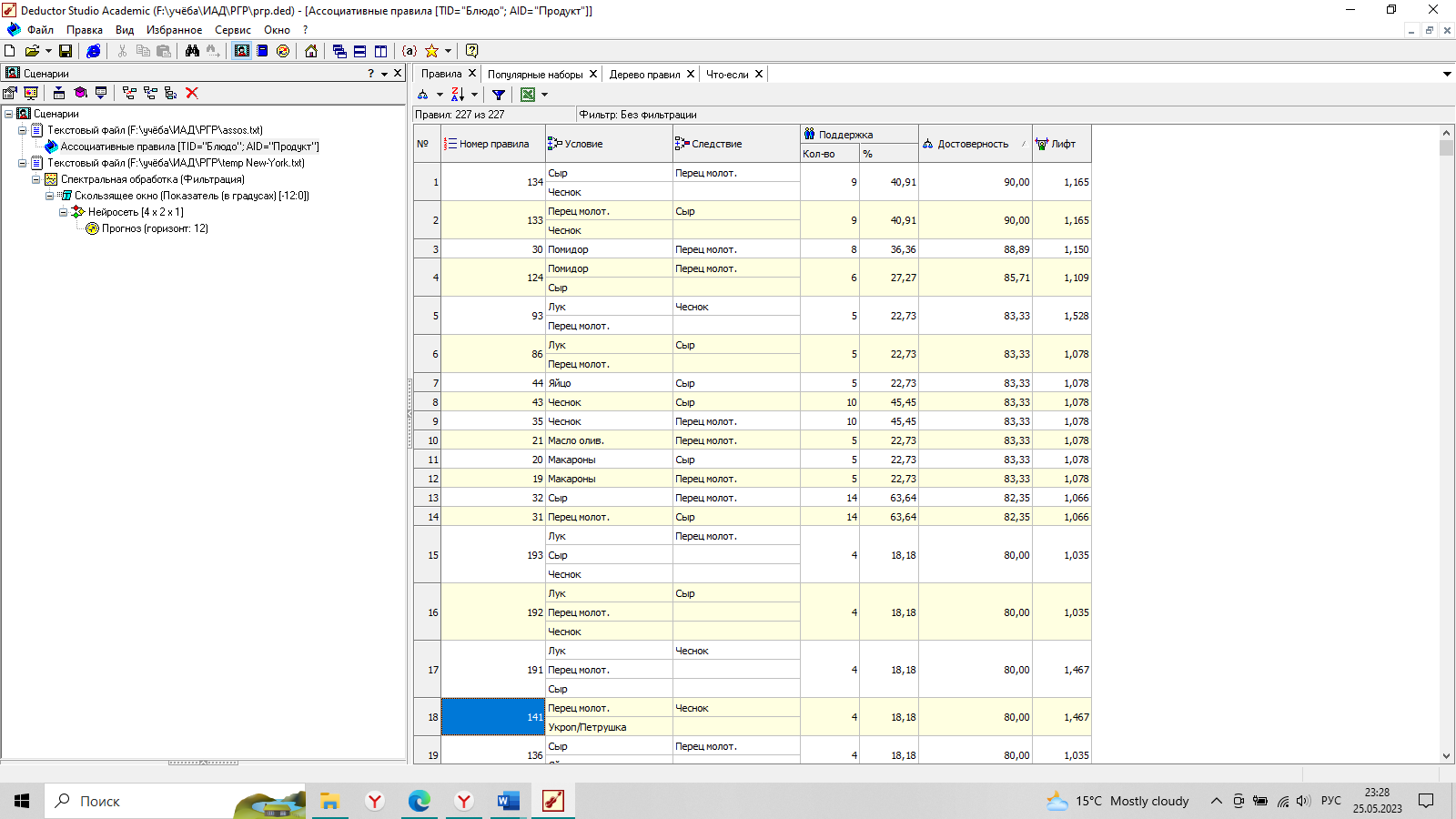


Рисунок 6 – Правила

### 2.1.3. Дерево правил

Визуализатор «Дерево правил» – это двухуровневое дерево, которое 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. На рисунках 7 и 8 изображены деревья, построенные по условию и следствию соответственно.

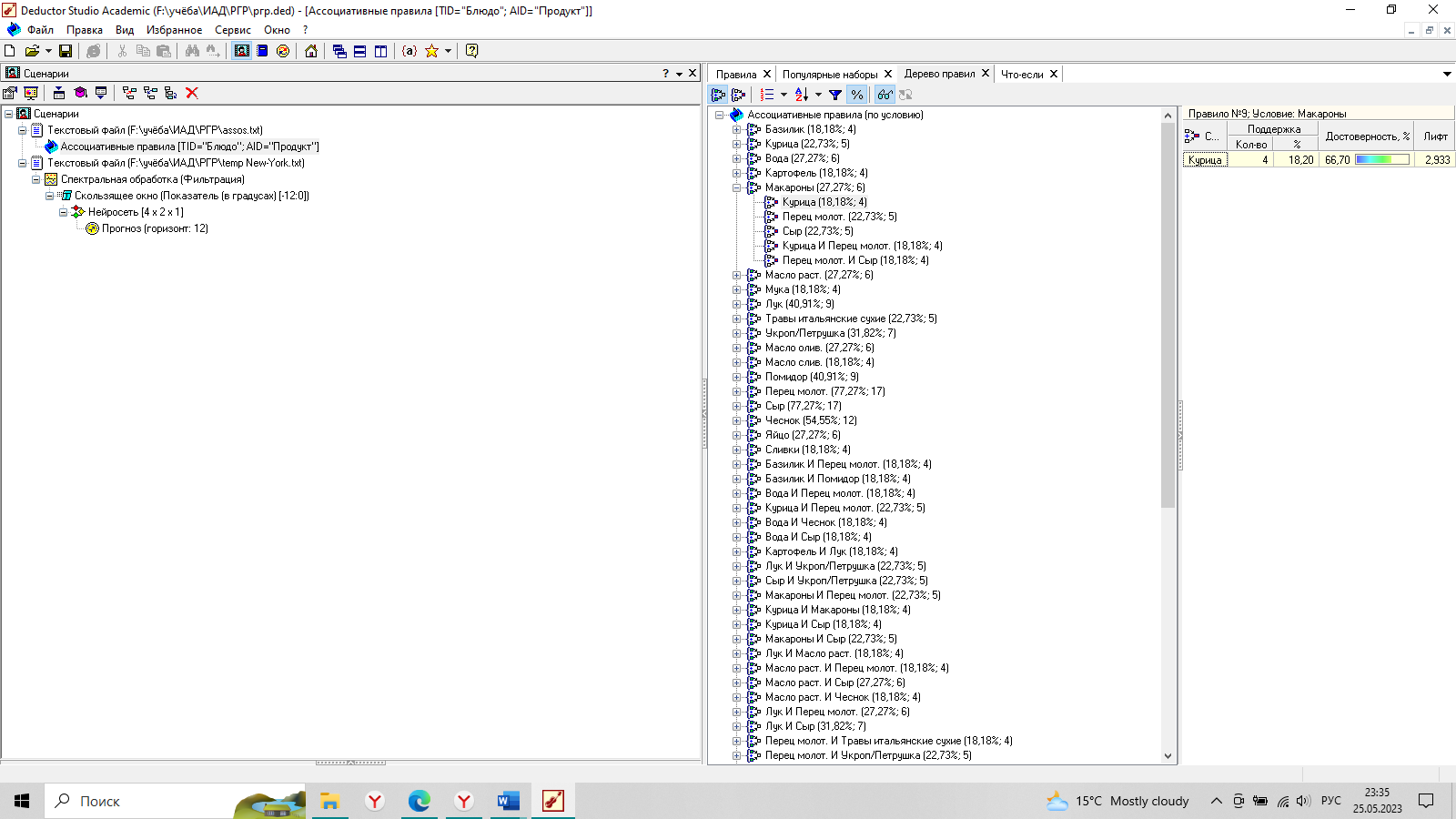


Рисунок 7 – Дерево правил по условию

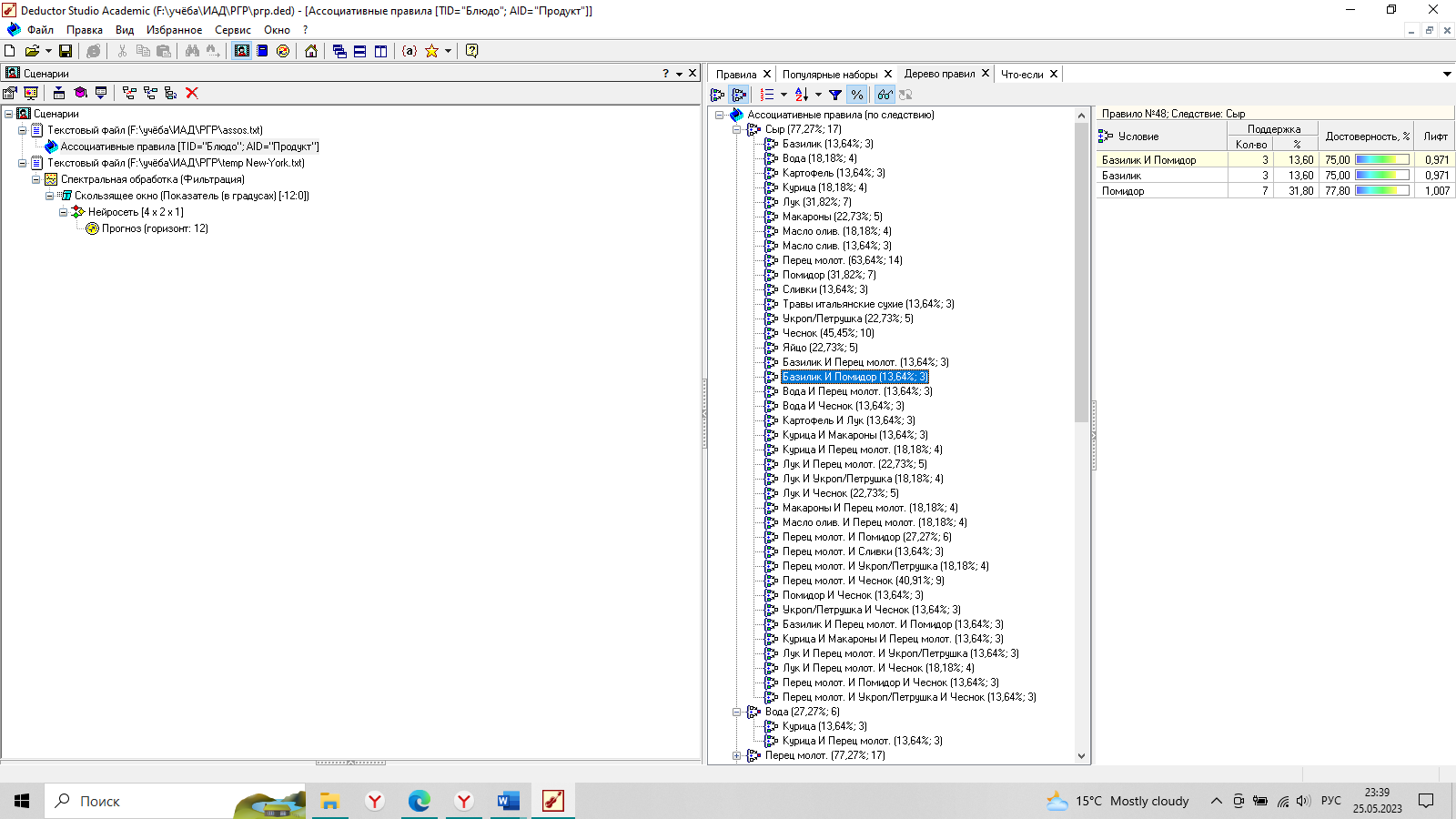


Рисунок 8 – Дерево правил по следствию

Как видно из рисунка 7, если блюдо содержит макароны, то с вероятностью 66,7% в блюдо будет добавлена курица.

Рисунок 8 иллюстрирует, что для того, чтобы добавить в блюдо сыр, в его составе с вероятностью 75% должны быть бализик и помидор или только помидор и с вероятностью 77,8% должен быть только помидор.

### 2.1.4. Что-если

Анализ «Что-если» в ассоциативных правилах позволяет ответить на вопрос, что получим в качестве следствия, если выберем данные условия. Например, какие ингредиенты используются в приготовлении блюд совместно с другими.

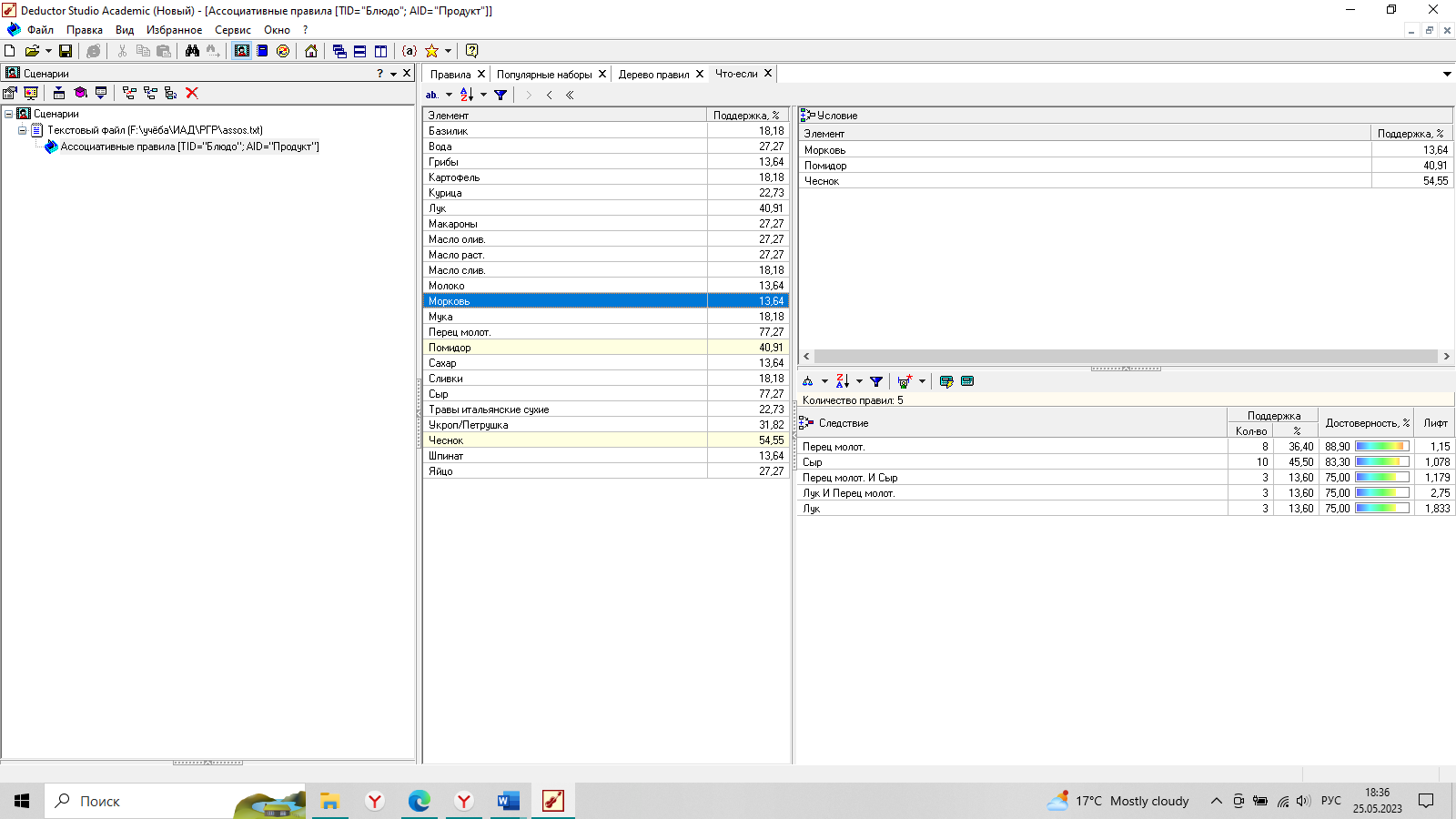
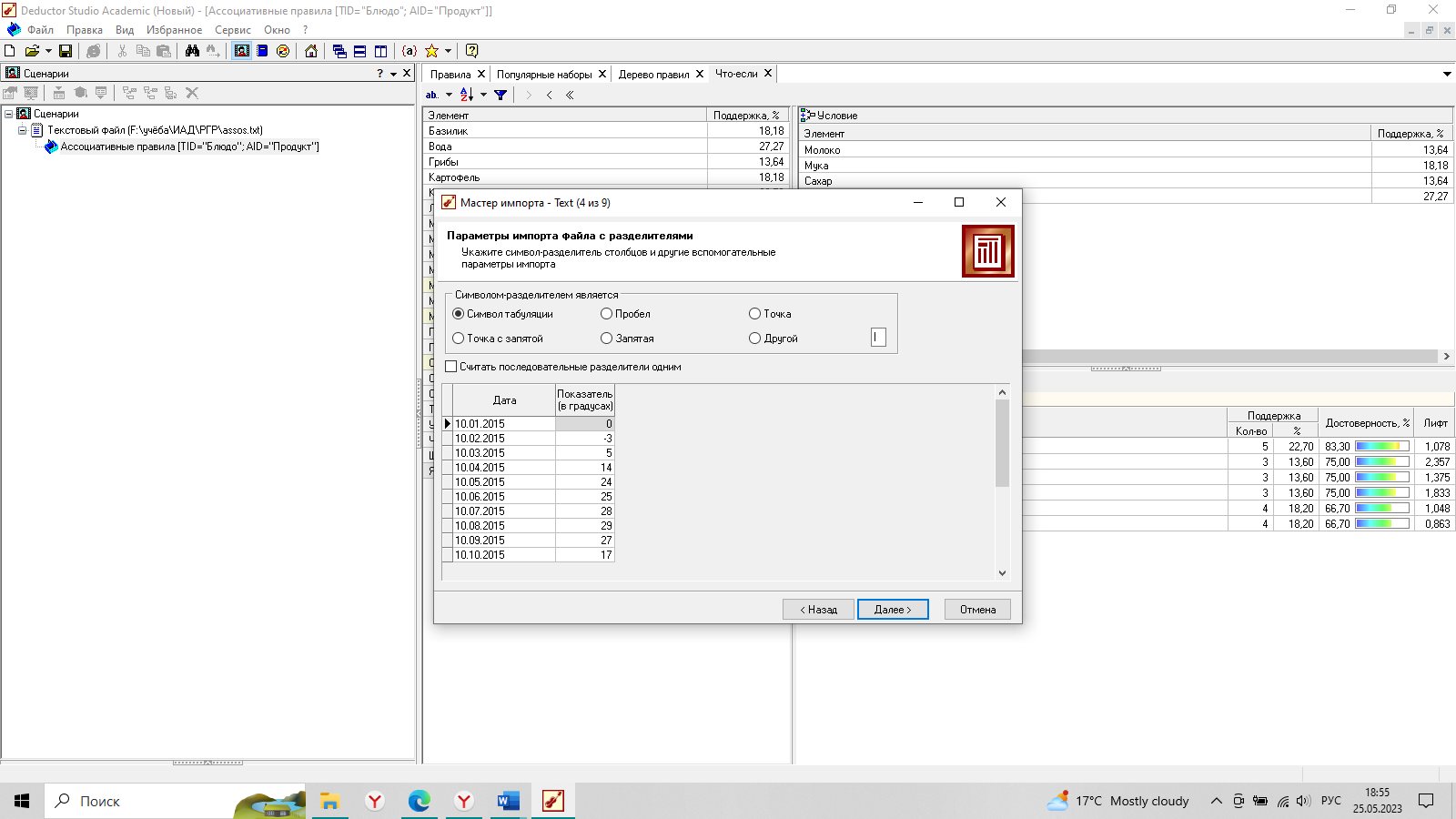
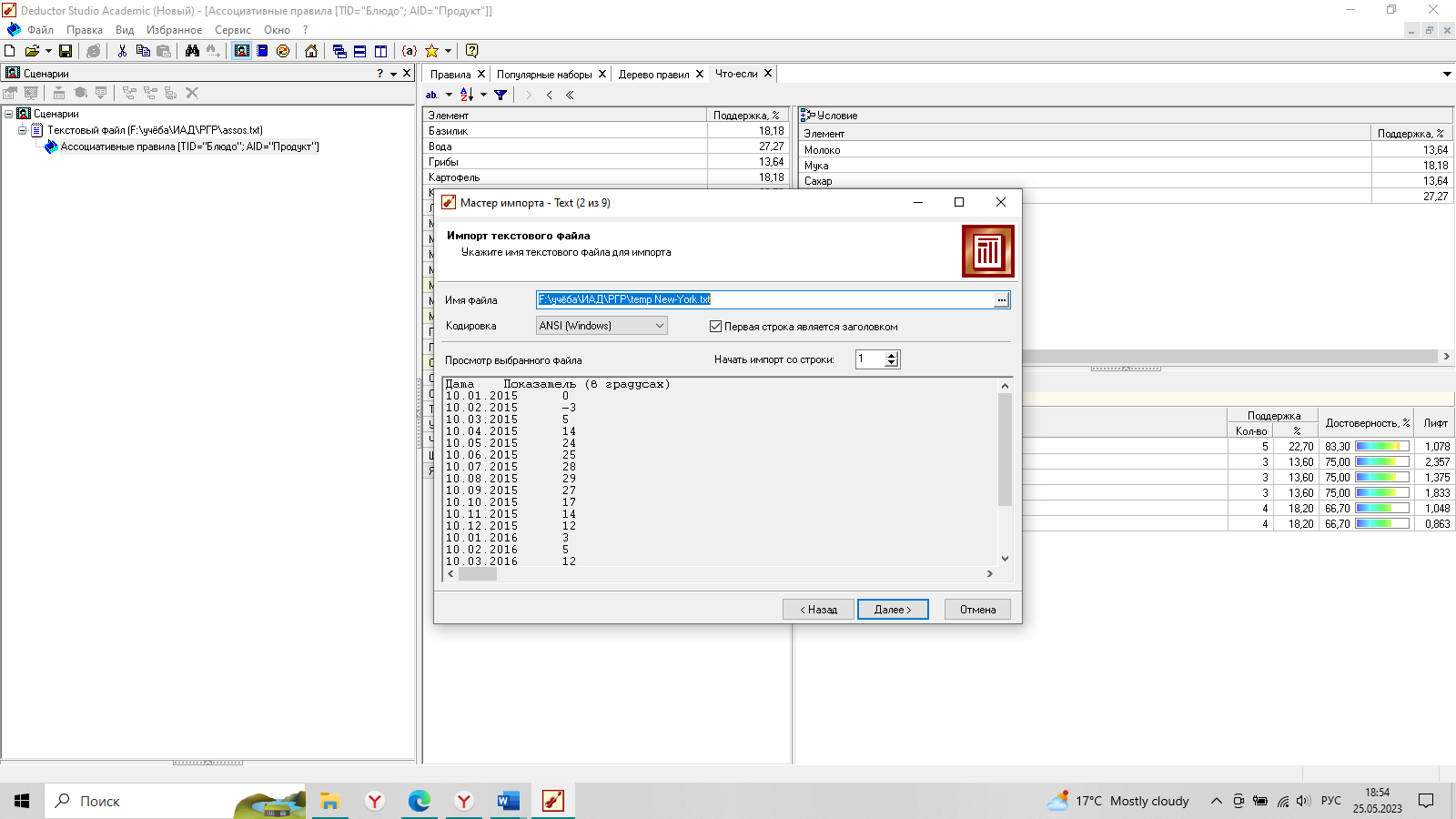


Рисунок 9 – Анализ «Что-если»

На рисунке 9 изображена следующая ситуация: необходимо было узнать, какие ингредиенты с какой вероятностью понадобятся, чтобы приготовить блюдо, в чьем составе есть морковь, помидор и чеснок. В «Следствие» вывелись ингредиенты: перец молот., сыр, лук. С вероятностью 88,9% в блюдо забыли положить перец, с 83,3% – сыр, с вероятностью 75% – только лук, либо лук и перец, либо перец и сыр.

## 2.2. Прогнозирование временного ряда

Аналогично с предыдущим набором данных сформированные данные о среднемесячной температуре в Нью-Йорке [2] были сохранены в формате .txt и импортированы. Процесс импорта изображен на рисунке 10.



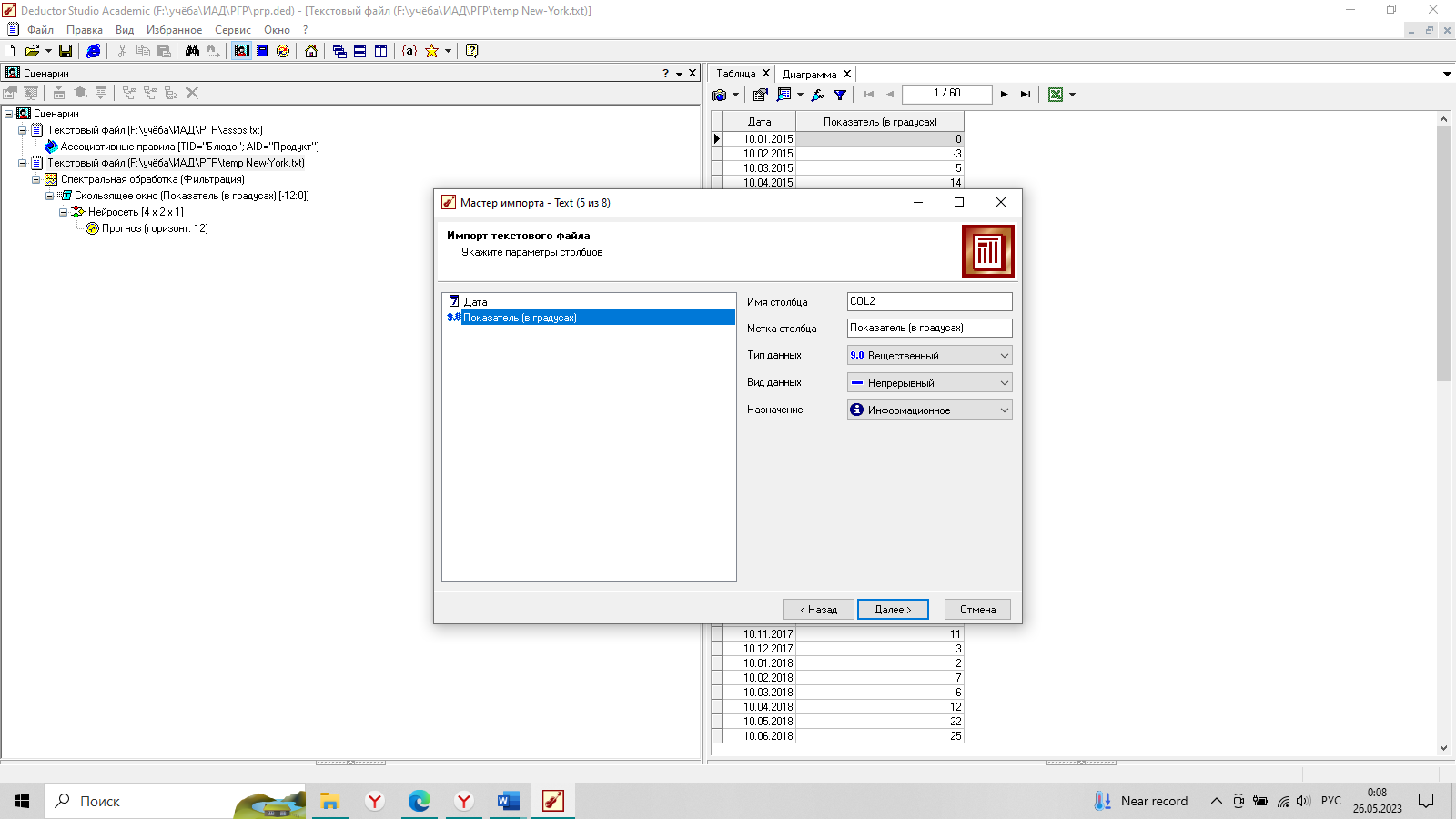
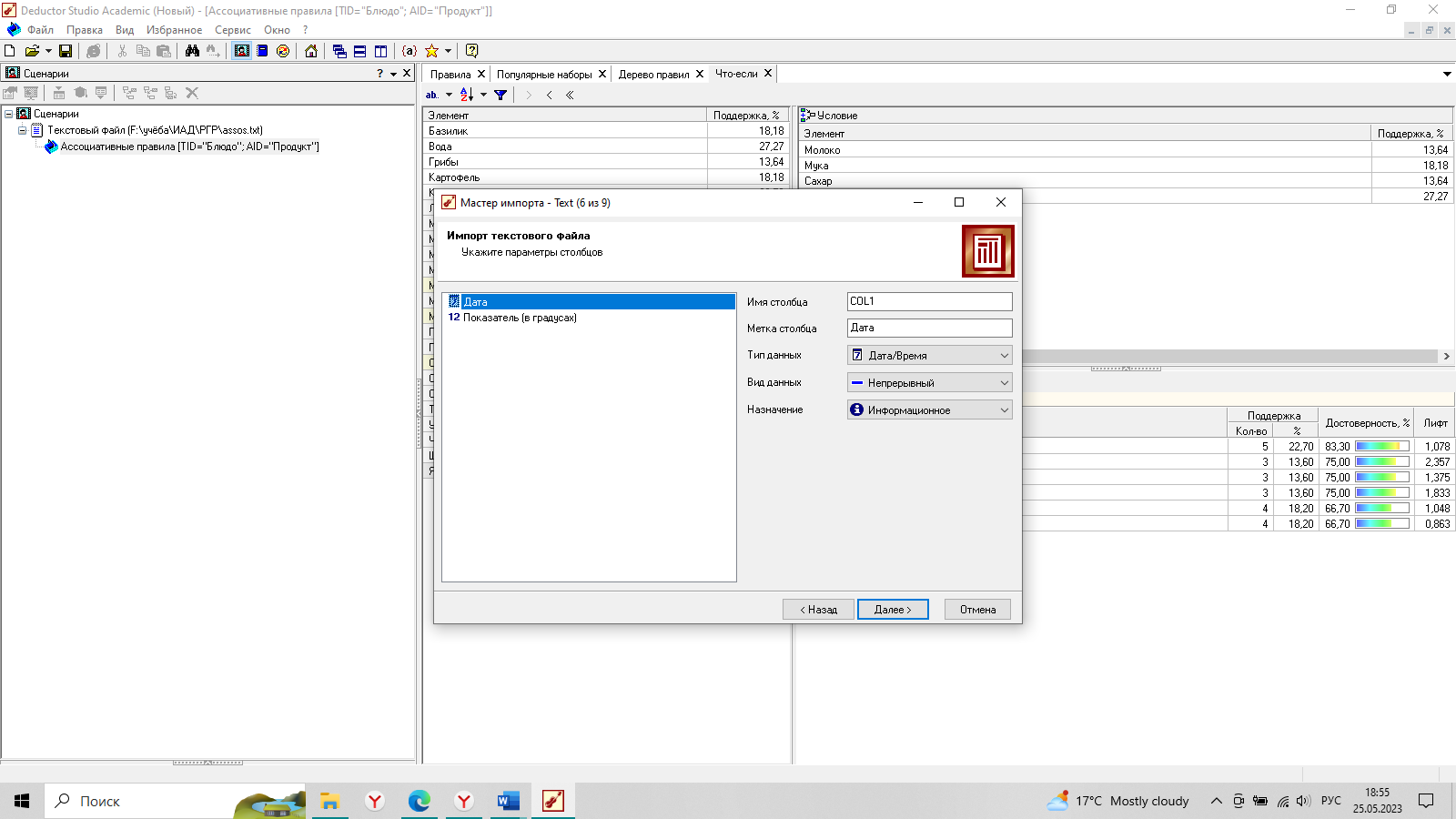


Рисунок 10 – Импорт данных

Вид данных обоих столбцов – непрерывный. Данные представлены в виде диаграммы на рисунке 11.

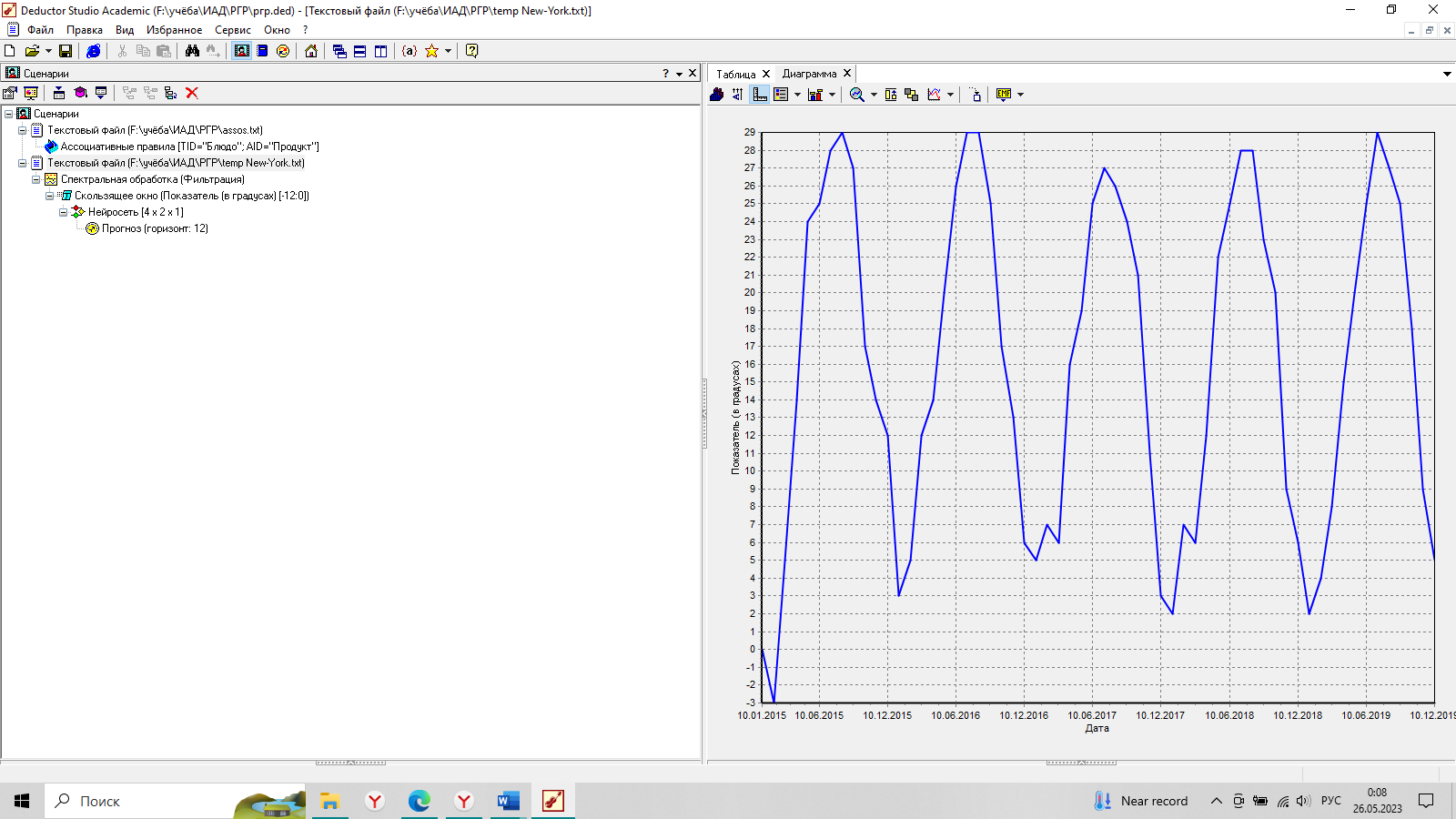


Рисунок 11 – Диаграмма данных

С помощью инструментов «Мастер обработки» 🡪 «Спектральная обработка» было проведено сглаживание данных с помощью вейлет-преобразования (рисунок 12). Результат проиллюстрирован на рисунке 13.

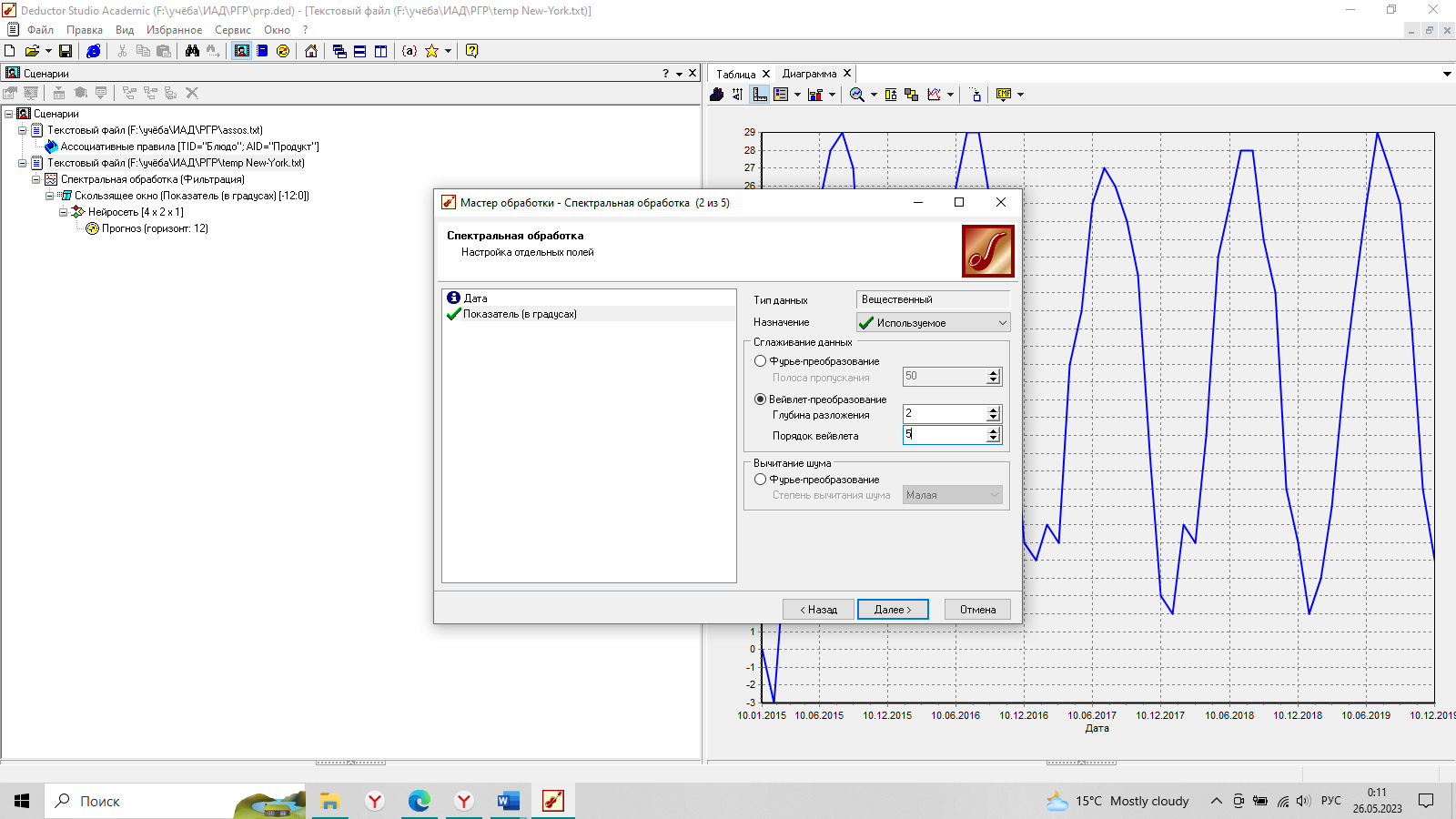


Рисунок 12 – Сглаживание данных

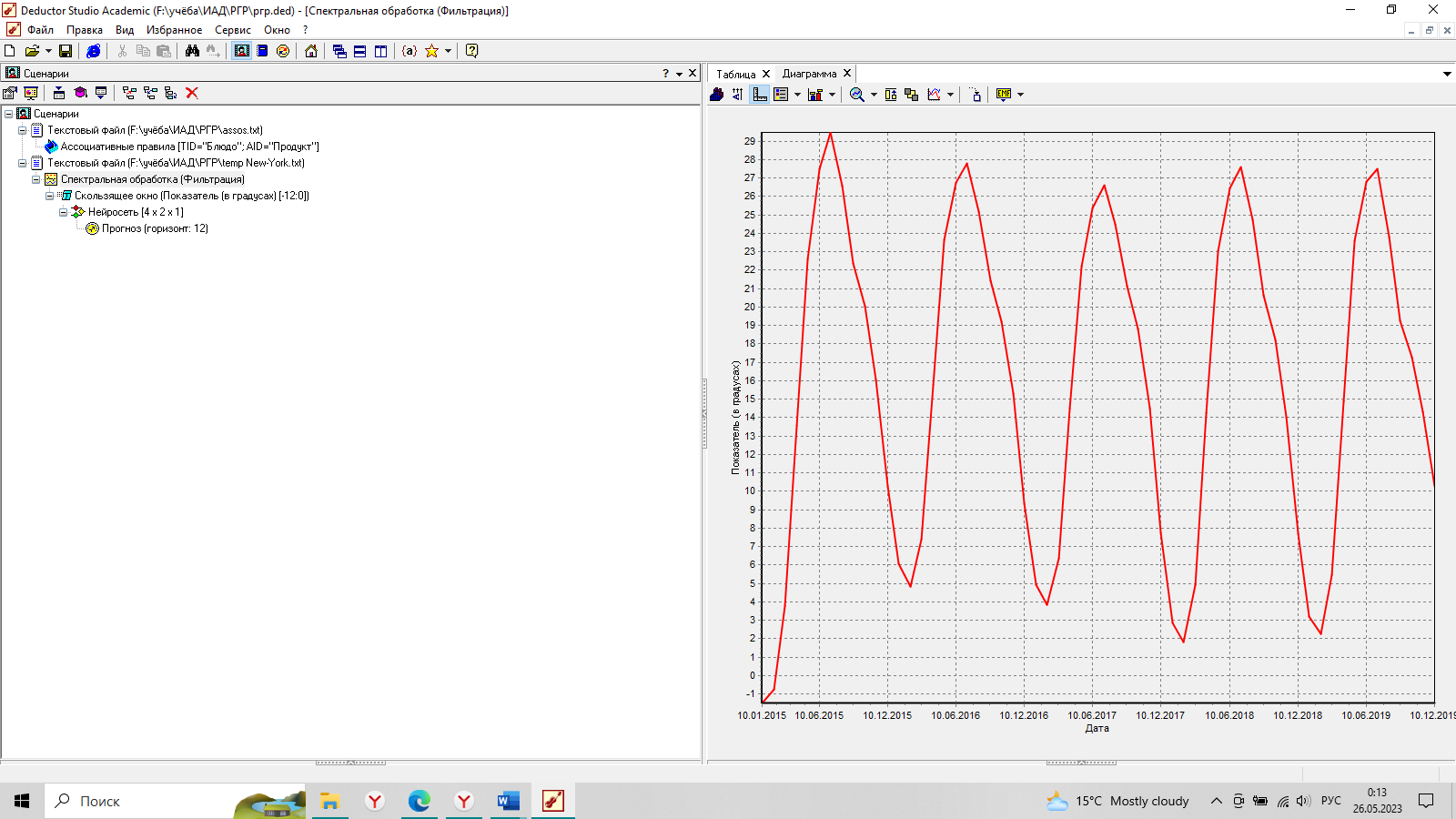


Рисунок 13 – Диаграмма сглаженных данных

Далее был снова запущен «Мастер обработки», выбрано в качестве обработчика «Скользящее окно». Требовалось выбрать глубину погружения 12, назначив поле «Показатель (в градусах)» используемым. Результат представлен на рисунке 14. Способ отображения был задан в виде таблицы, которая представлена на рисунке 15.

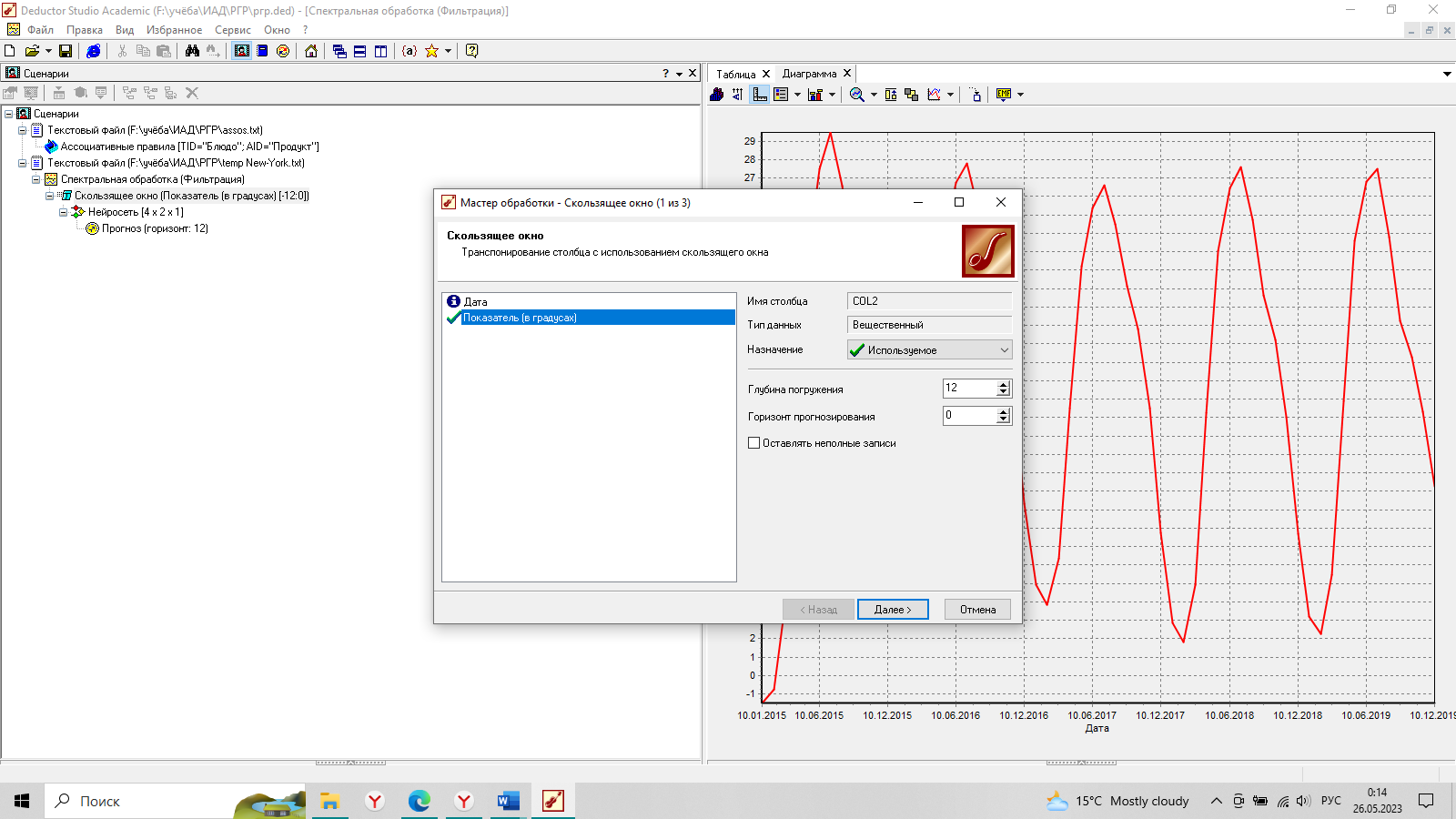


Рисунок 14 – Обработчик «Скользящее окно»

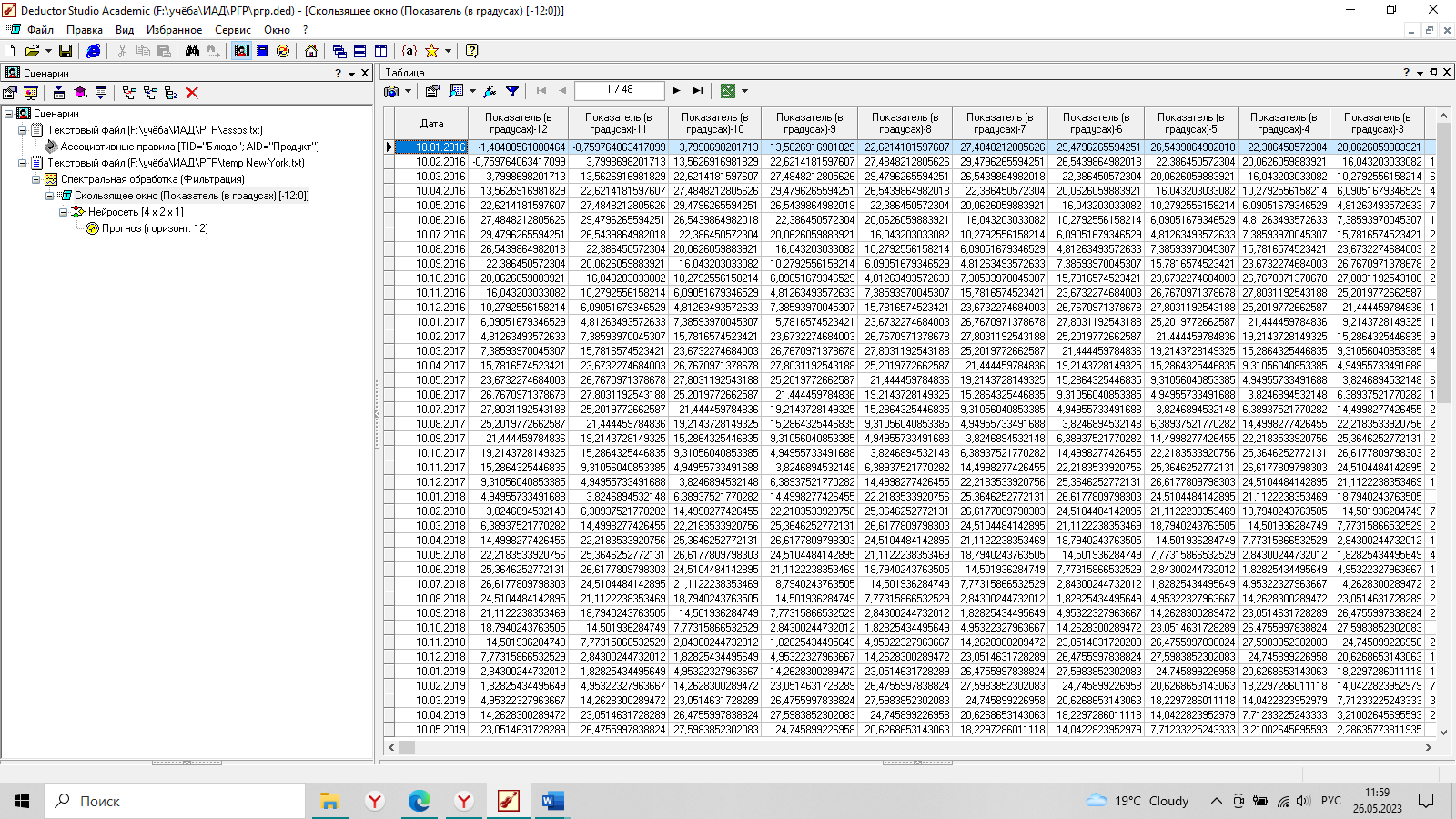


Рисунок 15 – Результат

Затем было произведено обучение нейросети с помощью «Мастер обработки» 🡪 «Нейросеть». Некоторые настройки представлены на рисунках 16-20.

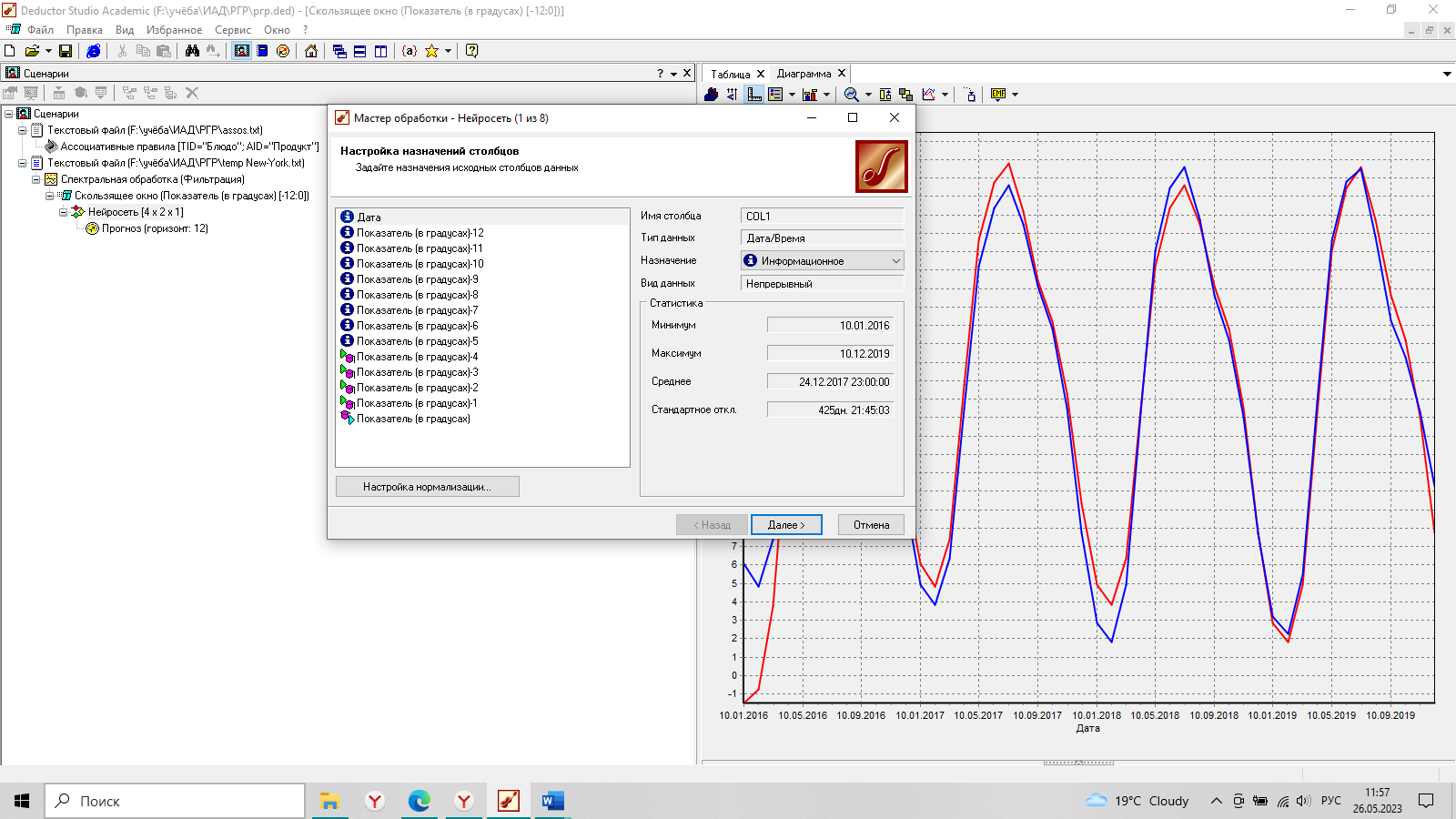


Рисунок 16 – Настройка назначений столбцов

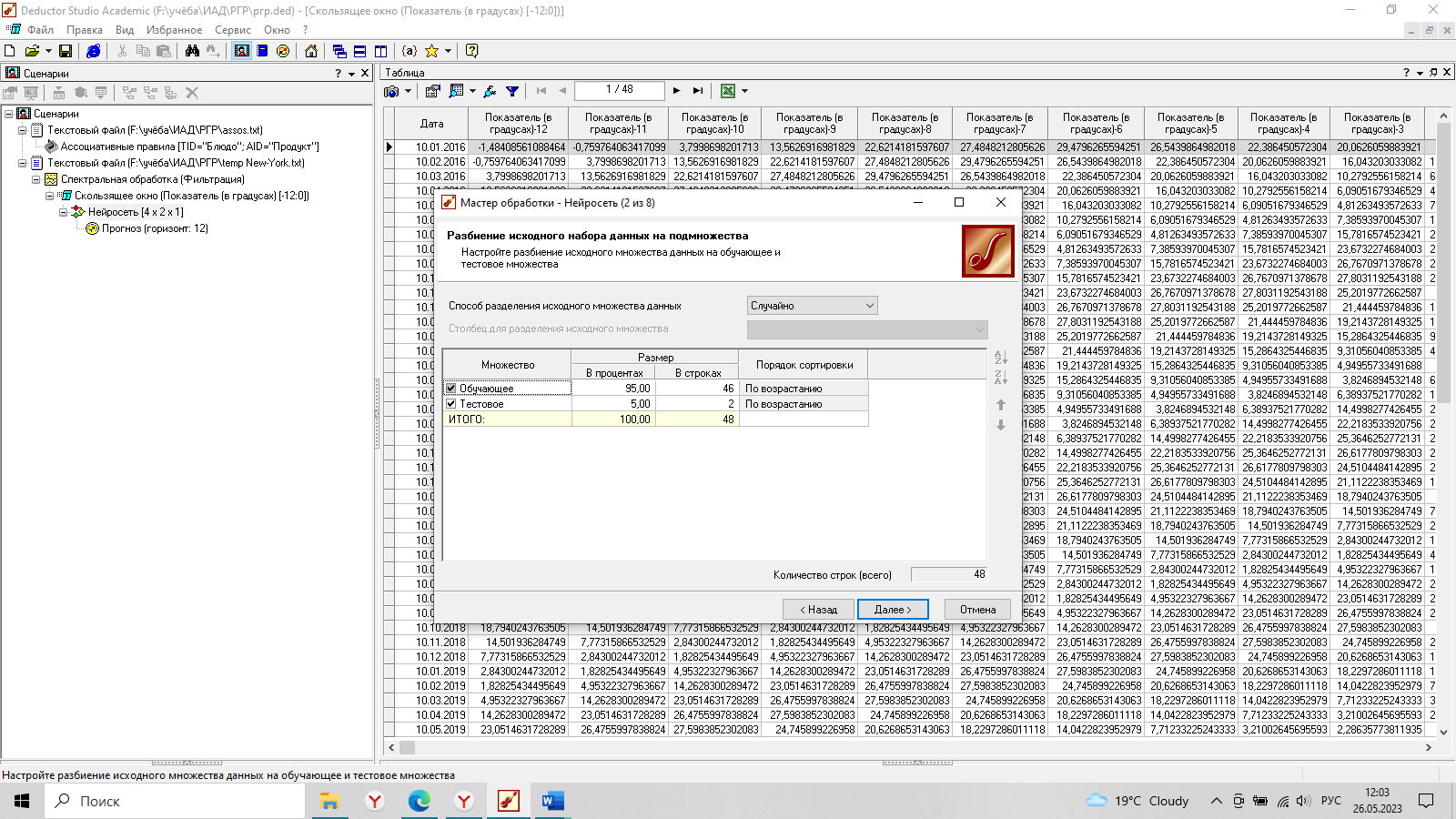


Рисунок 17 – Разбиение данных на обучающее и тестовое подмножества

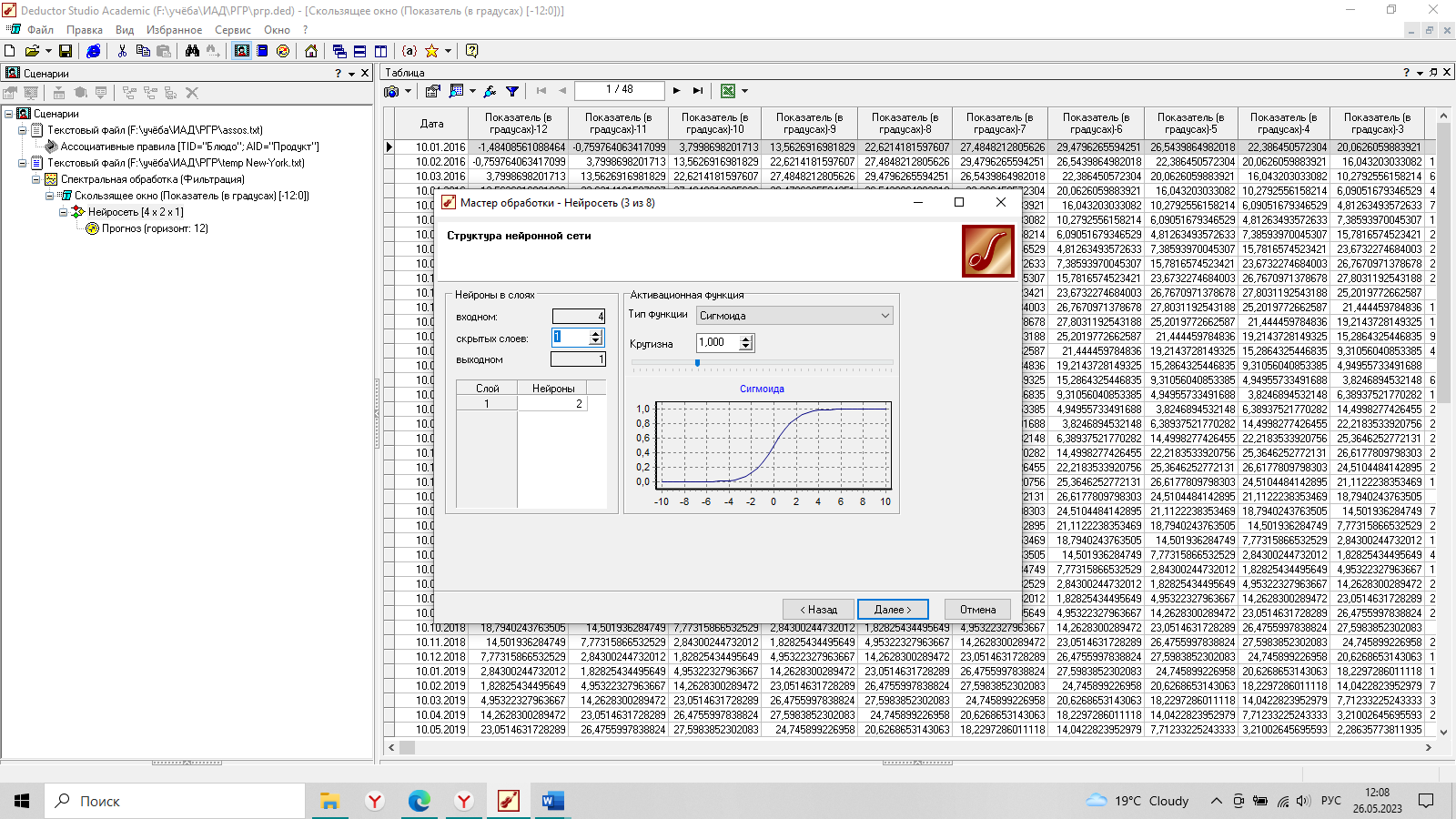


Рисунок 18 – Структура нейронной сети

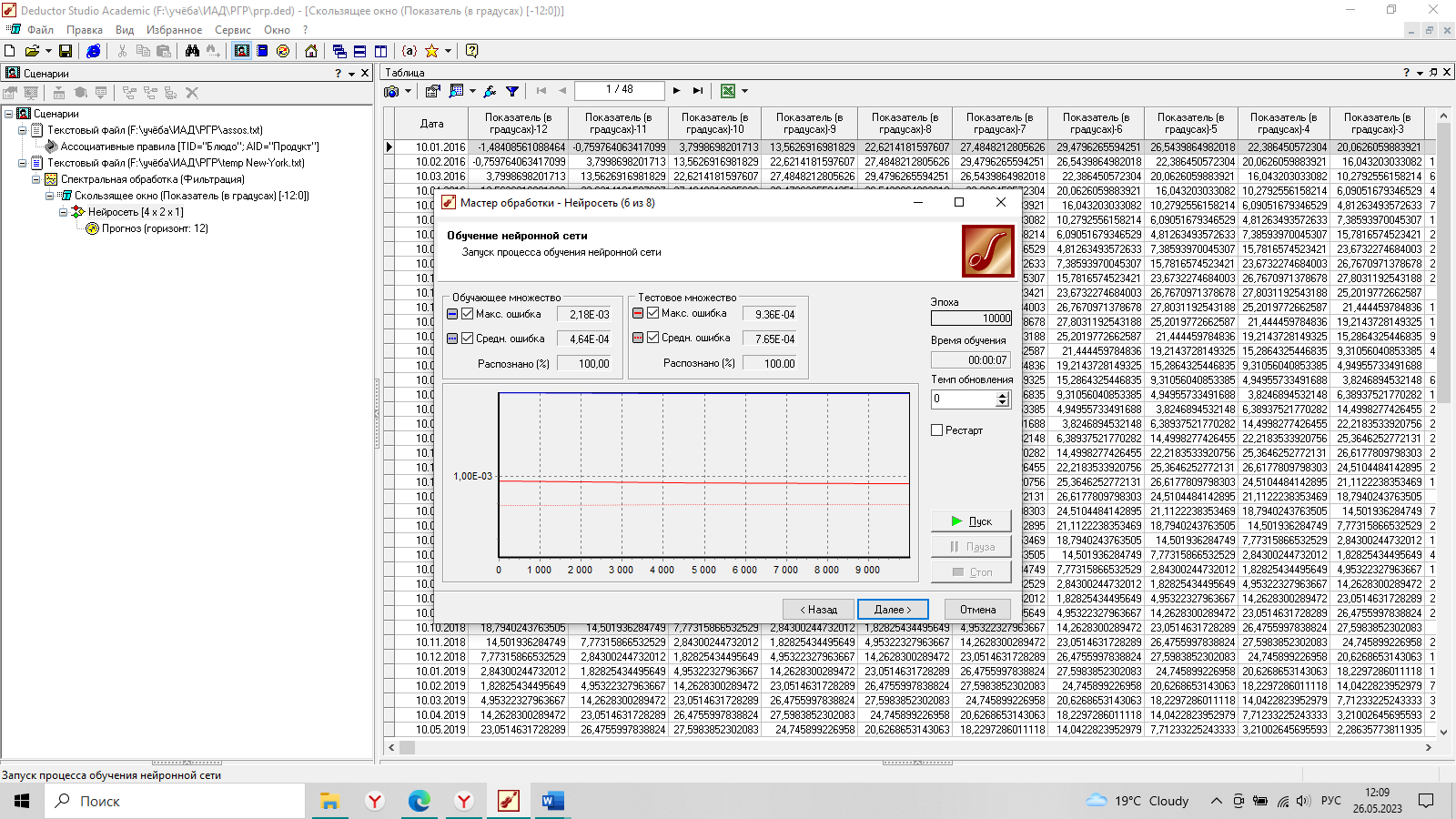


Рисунок 19 – Запуск обучения нейронной сети

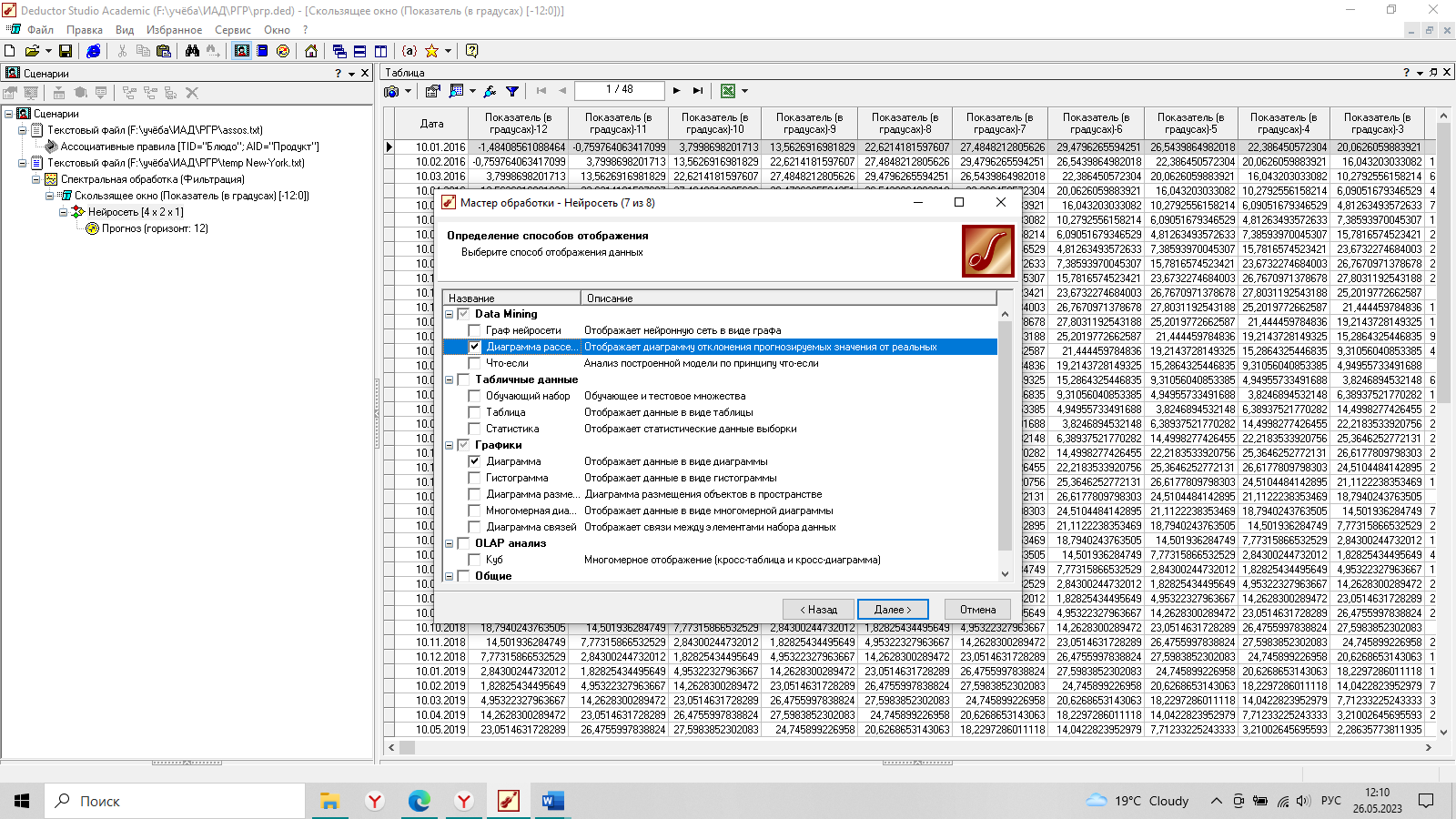


Рисунок 20 – Выбор отображения данных

На рисунке 21 представлена диаграмма рассеяния, на рисунке 22 – диаграмма, где розовый график – реальное значение, а фиолетовый – спрогнозированное.

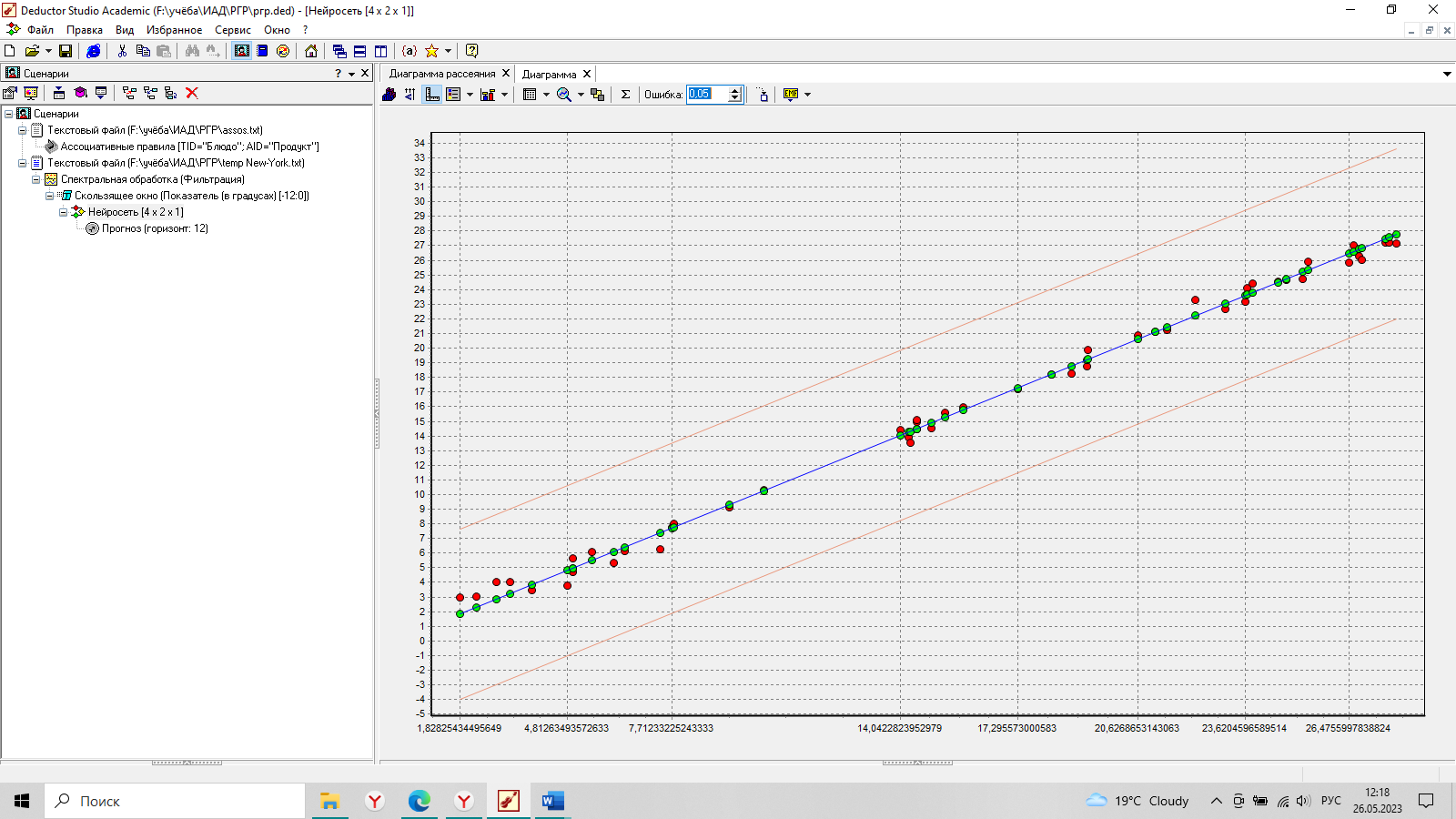


Рисунок 21 – Диаграмма рассеяния

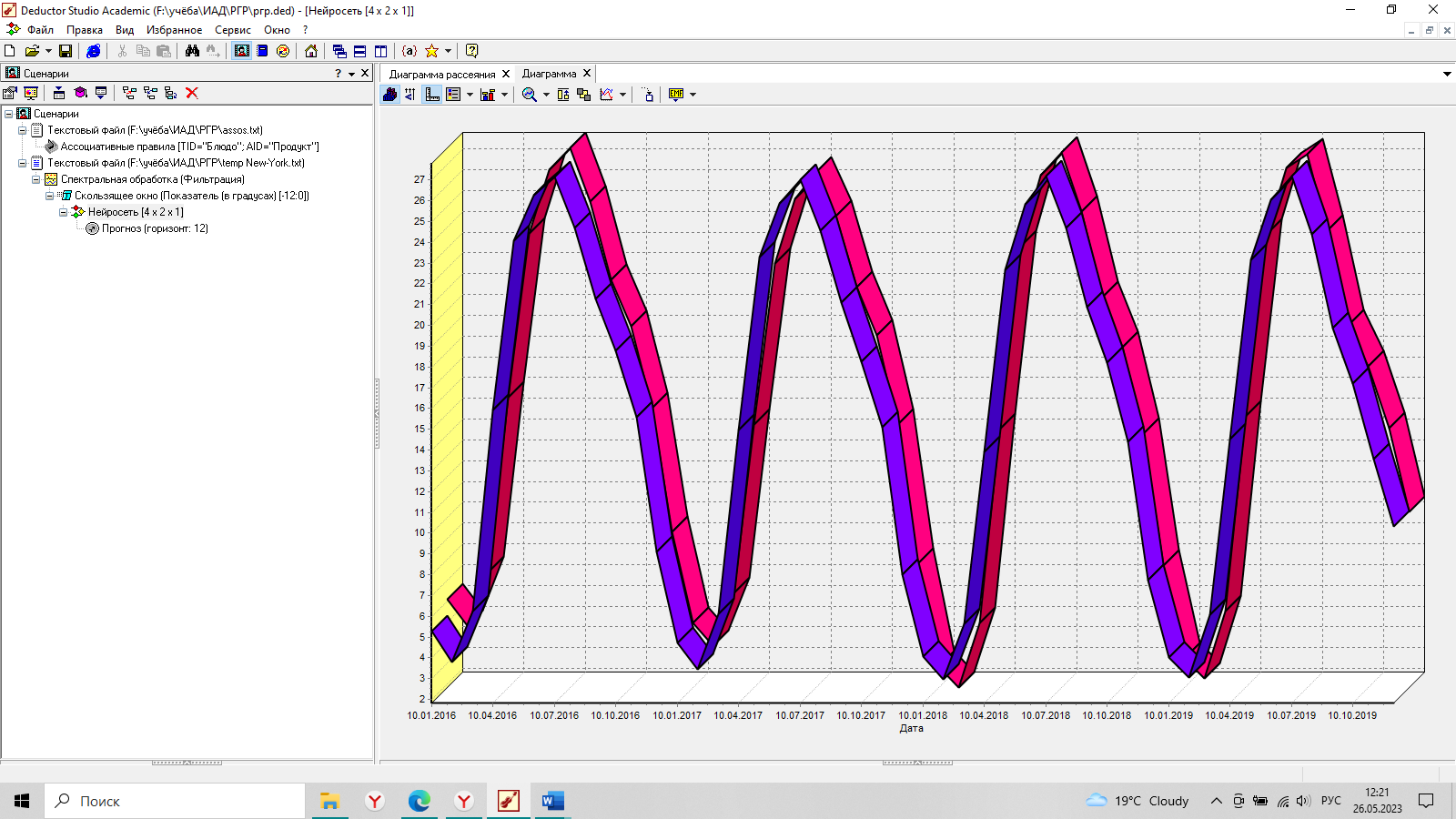


Рисунок 22 – Диаграмма реального и спрогнозированного значений

Далее был построен прогноз на следующие 12 месяцев (горизонт прогноза равен 12). Для этого были использованы «Мастер обработки» 🡪 «Прогнозирование». Результат отображался в виде диаграммы прогноза (рисунок 24).

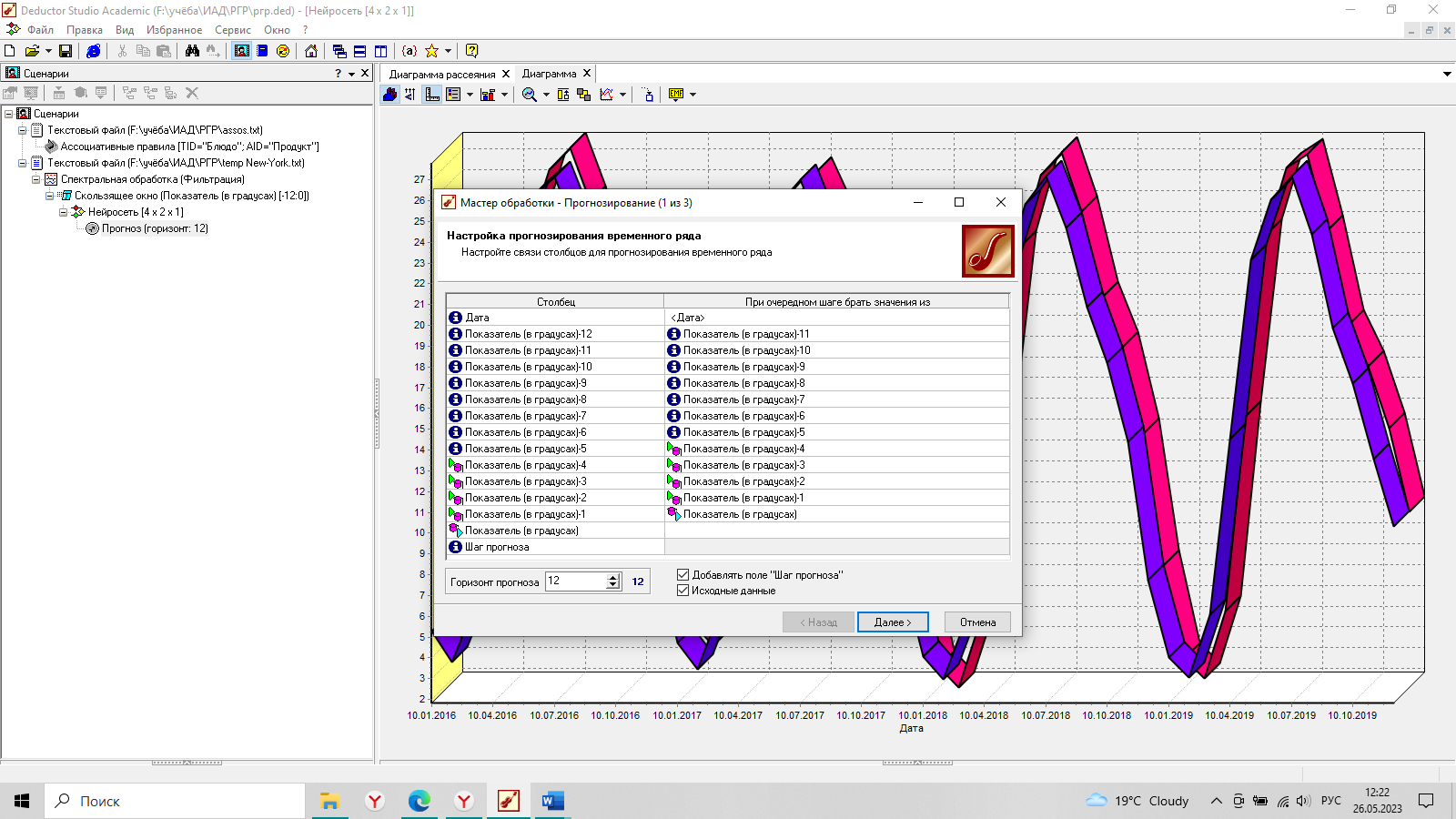


Рисунок 23 – Настройка прогнозирования временного ряда

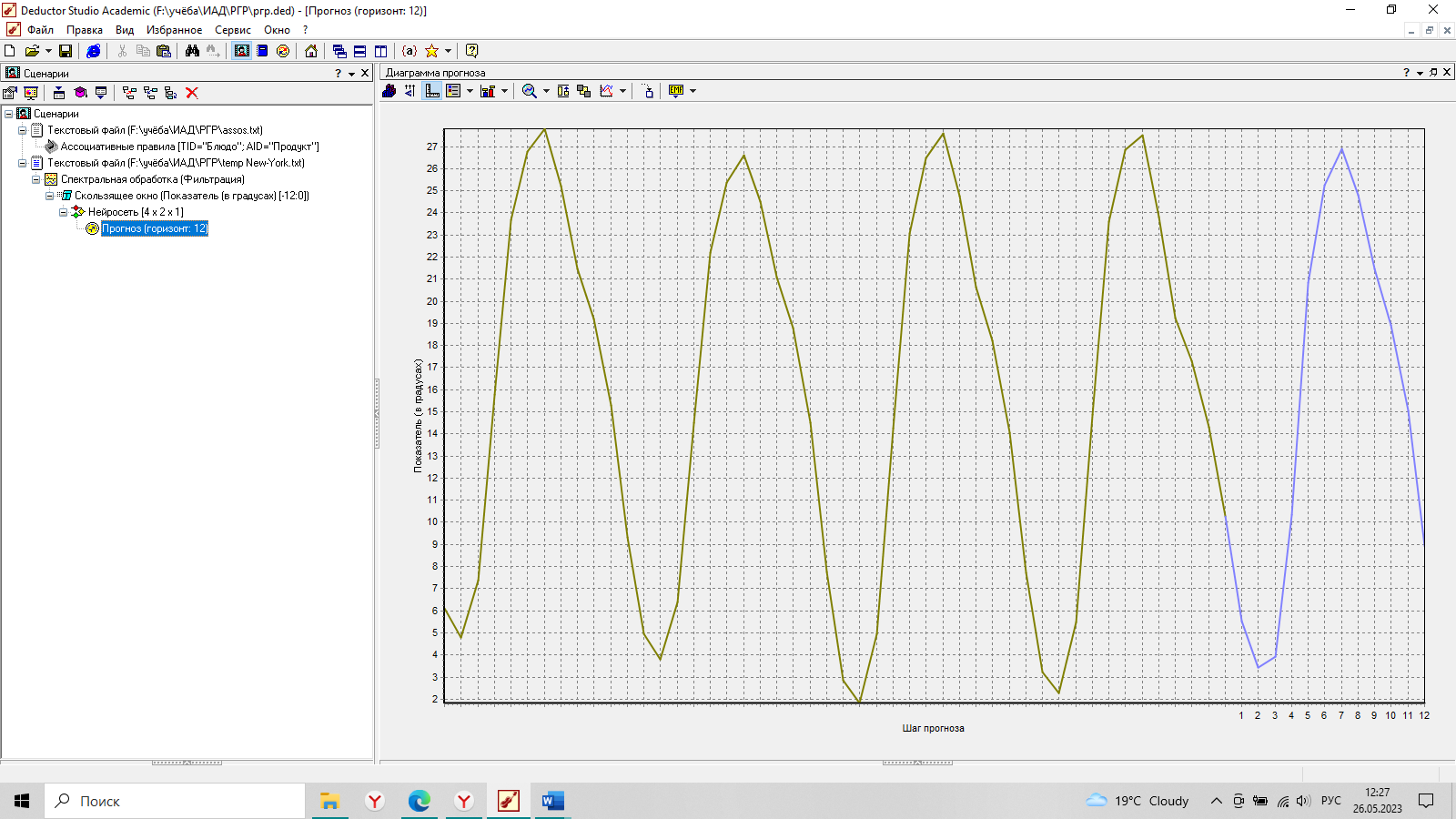


Рисунок 24 – Диаграмма прогноза

Полученные значения были сравнены с фактическими и оба были занесены для удобства в таблицу, представленную на рисунке 25.

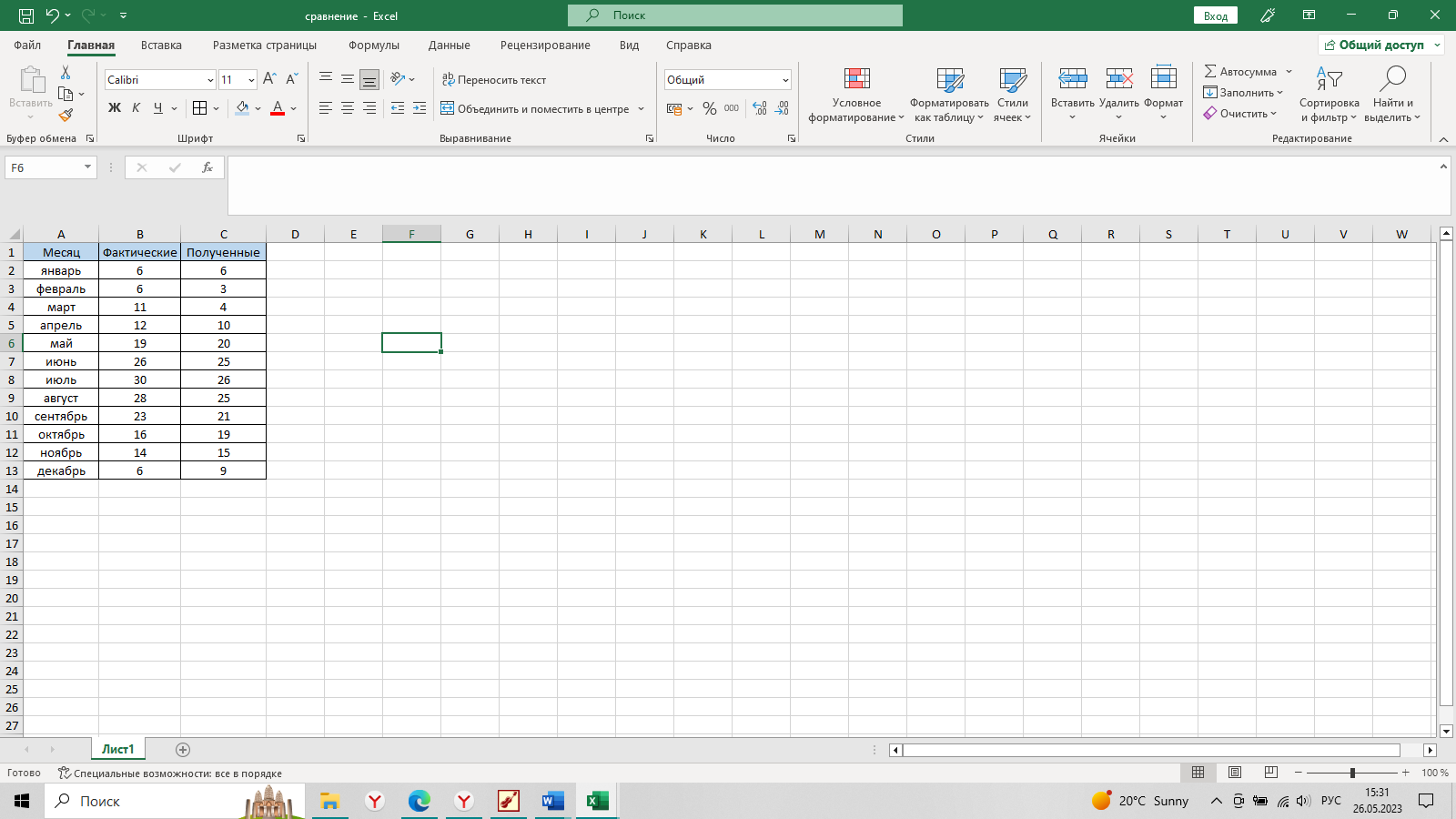


Рисунок 25 – Таблица сравнения экспериментальных и фактических данных

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, что логично. 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:

1. в большинстве итальянских блюд для приготовления используется молотый перец и сыр;
2. если блюдо содержит в составе сыр и чеснок, то с вероятность 90% в него будет добавлен и перец, а также если в составе будет яйцо, то с вероятностью 83,33% будет добавлен сыр;
3. если блюдо содержит макароны, то с вероятностью 66,7% в блюдо будет добавлена курица.

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, хоть и не совсем точным.

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.

На рисунке 26 в виде дерева сценариев представлены результаты проведенных операций над данными.

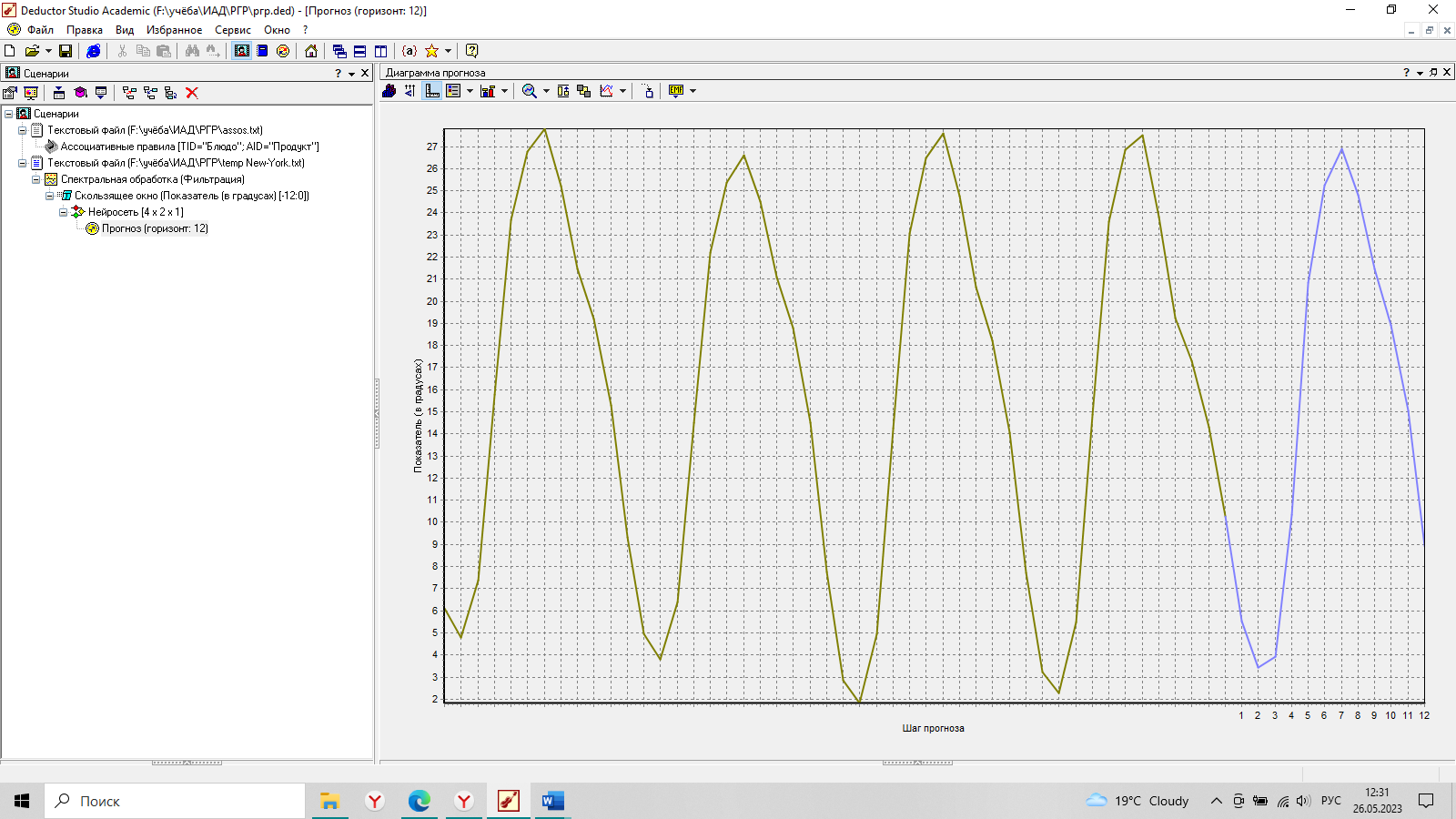


Рисунок 26 – Дерево сценариев всего РГР

# СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Замятин, А. В. Интеллектуальный анализ данных : учебное пособие / А. В. Замятин. - Томск : Издательский Дом Томского государственного университета, 2020. - 196 с. - ISBN 978-5-94621-898-6. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785946218986.html (дата обращения: 31.05.2023). - Режим доступа : по подписке.

2. Погода в Нью-Йорке по месяцам: [сайт]. URL: <https://world-weather.ru/pogoda/usa/new_york/2020/>

3. Итальянская кухня, фото с рецептами на RussianFood.com: [сайт]. URL: <https://www.russianfood.com/recipes/bytype/?fid=110>

4. Методические указания к РГР\_ИАД