**МИНИСТЕРСТВО** НАУКИ И ВЫСШЕГО **ОБРАЗОВАНИЯ РФ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

кафедра «Информационные системы»

**Пояснительная записка**

к расчётно-графической работе

по дисциплине «Интеллектуальный анализ данных»

«Компьютерные методы анализа данных и прогнозирования»

Выполнил: студент 3 курса, группы: ИС/б-21-2-о

Направления подготовки (специальности): 09.03.02 Информационные системы и технологии

профиль (специализация): Информационные системы и технологии

Мовенко Константин Михайлович

Старший преподаватель кафедры ИС Сырых О.А.

Защита « » 2024 г. Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись) (инициалы, фамилия)

20 24 г.

# СОДЕРЖАНИЕ

[СОДЕРЖАНИЕ 2](#_Toc168615279)

[ВВЕДЕНИЕ 3](#_Toc168615280)

[1 АНАЛИТИЧЕСКАЯ ЧАСТЬ 4](#_Toc168615281)

[1.1. Интеллектуальный анализ данных 4](#_Toc168615282)

[1.2. Поиск ассоциативных правил 5](#_Toc168615283)

[1.3. Прогнозирование временного ряда 6](#_Toc168615284)

[Вывод раздела 1 7](#_Toc168615285)

[2 ПРАКТИЧЕСКАЯ ЧАСТЬ 8](#_Toc168615286)

[2.1 Поиск ассоциативных правил 8](#_Toc168615287)

[2.2 Анализ временного ряда 13](#_Toc168615288)

[Вывод раздела 2 18](#_Toc168615289)

[ЗАКЛЮЧЕНИЕ 19](#_Toc168615290)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 20](#_Toc168615291)

# ВВЕДЕНИЕ

Цели и задачи расчётно-графической работы: ознакомление с двумя методиками Data Mining – поиском ассоциативных правил и прогнозированием временных рядов. Стоит задача изучить теоретическую основу методов и область их применения, а также попрактиковаться на реальных данных в их применении.

В качестве прикладного инструмента для работы с данными предложена программа Deductor Studio − аналитическая платформа, основа для создания законченных прикладных решений. Технологии Deductor позволяют реализовать все этапы построения аналитической системы: от создания хранилища данных до автоматического подбора моделей и визуализации полученных результатов.

В ходе работы необходимо подготовить соответствующие наборы данных, экспортировать их в Deductor, провести соответствующие манипуляции и проанализировать результаты. При анализе результатов сформулировать конкретные выводы, оценить качество прогнозов.

По окончании выполнения работы будут получены навыки сбора, преобразования и анализа экспериментальных данных, а также прогнозирования временных рядов с помощью имеющихся компьютерных средств.

# 1 АНАЛИТИЧЕСКАЯ ЧАСТЬ

## Интеллектуальный анализ данных

Интеллектуальный анализ данных (Data Mining) – процесс поиска и обнаружения в больших наборах данных информации, неизвестной явно, но имеющей важное практическое применение. Благодаря интеллектуальному анализу возможно из наборов данных, плохо поддающихся ручному анализу, выводить закономерности и строить прогнозы на будущее.

Data Mining, как мощный инструмент аналитики, используется при решении различного ряда задач. Вот основные из них:

* классификация – отнесение объектов к классам на основе имеющихся примеров, предсказание классов ранее неизвестных объектов;
* кластеризация – распределение группы объектов на кластеры, определение признаков кластеров, выделение структур данных;
* ассоциативный анализ – обнаружение правил и паттернов, анализ продуктовых чеков и частых комбинаций товаров;
* прогнозирование – предсказание развития будущего на основе тенденций изменения данных в прошлом и статистики;
* выявление аномалий – поиск в паттернах данных отклонений и ошибок, требующих рассмотрения.

Эта дисциплина используется множеством предприятий для принятия организационных и стратегических решений, позволяет компаниям понимать клиентов и рынок, оптимизировать бизнес-процессы. Благодаря всеобщему распространению информационных технологий аналитики данных являются ценными специалистами в практически любой области труда. Это делает Data Mining крайне перспективной и актуальной областью знаний.

## Поиск ассоциативных правил

Ассоциативные правила позволяют находить закономерности между связанными объектами. Правило заключается в том, что из некоторого события X с некоторой вероятностью следует событие Y. Установление подобных зависимостей даёт возможность находить простые и интуитивно понятные правила.

Основной пример применения методики − анализ рыночной корзины. Алгоритм выявляет типичные шаблоны покупок и совместно приобретаемых товаров. Полученные результаты позволяют оптимизировать ассортимент и его размещение в торговых залах, улучшить управление запасами, увеличить объёмы продаж за счёт предложения клиентам сопутствующих товаров.

Также метод эффективен при составлении адресной рассылки рекламных предложений потенциальным и существующим покупателям – это высокоэффективный, простой и дешевый маркетинговый инструмент. Для увеличения количества откликов на письма необходимо производить тщательный отбор объектов для рассылки посланий, чему способствует рассматриваемый алгоритм.

Помимо вышерассмотренных примеров ассоциативные правила используются во многих других областях, включая Web Mining, обнаружение вторжений, непрерывное производство и биоинформатику.

Алгоритм поиска предполагает рассмотрение всех возможных пар элементов и выборку тех, что удовлетворяют заданным значениям достоверности и поддержки (ассоциативных).

Выявление часто встречающихся наборов элементов – операция, требующая больших вычислительных и временных ресурсов. В прикладных системах для её реализации используется алгоритм Apriori.

Взяв в качестве примера анализ чеков покупателей, поддержка – это отношение числа чеков с рассматриваемой комбинацией товаров к общему числу чеков, а достоверность – к числу чеков с первым товаром комбинации.

## Прогнозирование временного ряда

Прогнозирование временных рядов − это методика, которая позволяет предсказывать будущие значения на основе данных о прошлом. Метод предполагает поиск в данных зацикленности, а также трендов и тенденций, позволяющих судить о дальнейшем их развитии.

Основная цель прогнозирования временных рядов заключается в выявлении трендов и паттернов, которые помогут принимать обоснованные решения в бизнесе и науке.

Метод прогнозирования учитывает различные факторы: сезонные изменения, выбросы и аномалии, неопределённость будущего. Тем не менее, он может быть уязвим к непредсказуемым внешним обстоятельствам.

В качестве примеров применения прогнозирования можно привести: предсказание спроса на товары или услуги, выявление сезонных колебаний и трендов и др. Особо полезен метод при принятии решений в условиях рисков и вероятностей.

Прогнозирование предполагает создание модели нейросети, полученной через глубокого обучение на исторических данных. Модель использует временные срезы уже известных данных чтобы изучить тенденции, сезонность, закономерности и взаимосвязи между прошлыми данными и ожидаемыми значениями переменных.

В программе Deductor Studio прогнозирование реализуется за счёт мастеров обработки: редактирование выбросов, спектральная обработка, скользящее окно, нейросеть, диаграмма прогноза. В конечном итоге полученная нейросеть дорисует имеющуюся диаграмму значений на будущие временные периоды.

## Вывод раздела 1

В этом разделе было рассмотрено, что такое интеллектуальный анализ в общем, в чём цели этой дисциплины и какими методами она располагает. Проанализирована актуальность Data Mining в современных реалиях.

В рамках работы наиболее подробно были рассмотрены два метода: поиск ассоциативных правил и прогнозирование временных рядов.

Первый метод даёт возможность в ряду сложных комбинаций данных находить правила, по которым некоторые сочетания объектов более вероятны чем другие.

Второй метод позволяет на основе регулярных данных о прошедшем временном периоде формировать статистику, выявлять общие тенденции и строить прогноз на то, как будет развиваться рассматриваемая область в ближайшем будущем.

# 2 ПРАКТИЧЕСКАЯ ЧАСТЬ

## 2.1 Поиск ассоциативных правил

Перед проведением анализа был подготовлен набор данных, содержащий ингредиенты блюд русской кухни (рисунок 1). Данные были оформлены в виде таблицы формата .csv.

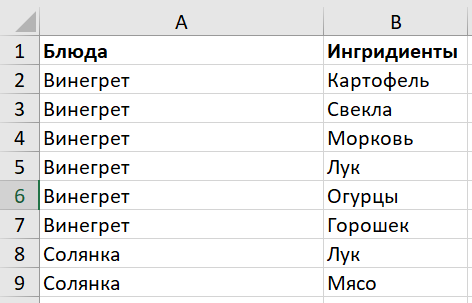


Рисунок 1 – Данные в MS Excel

Затем была открыта программа Deductor Studio и создан новый проект. В него с помощью мастера импорта были импортированы данные из csv-файла (рисунок 2).

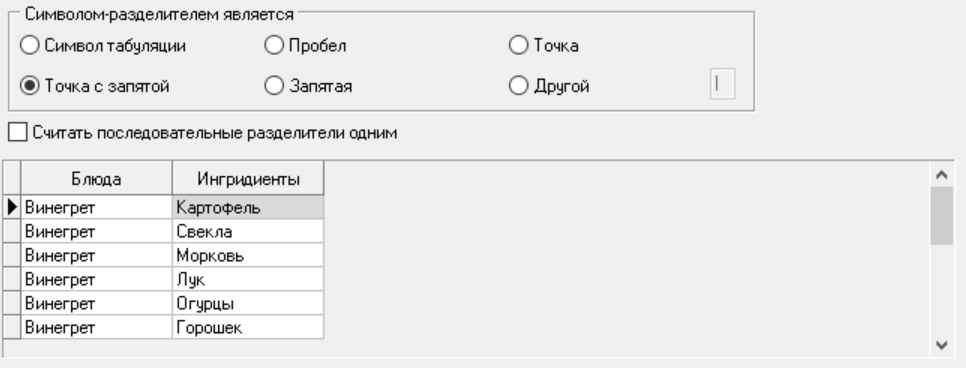


Рисунок 2 – Мастер импорта данных из файла

Данные были отображены в программе в виде таблицы (рисунок 3).

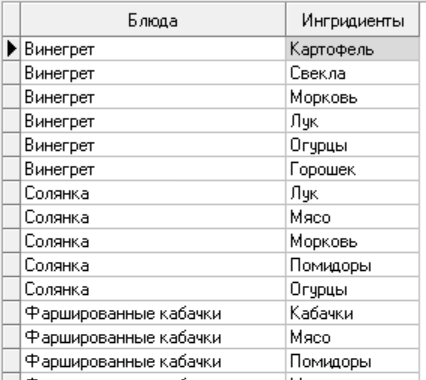


Рисунок 3 – Таблица данных в Deductor

Далее с помощью мастера обработки в Deductor в указанной таблице был проведён поиск ассоциативных правил. В качестве идентификатора транзакции был установлен столбец «Блюда», а в качестве её элемента – «Ингредиенты» (рисунок 4).

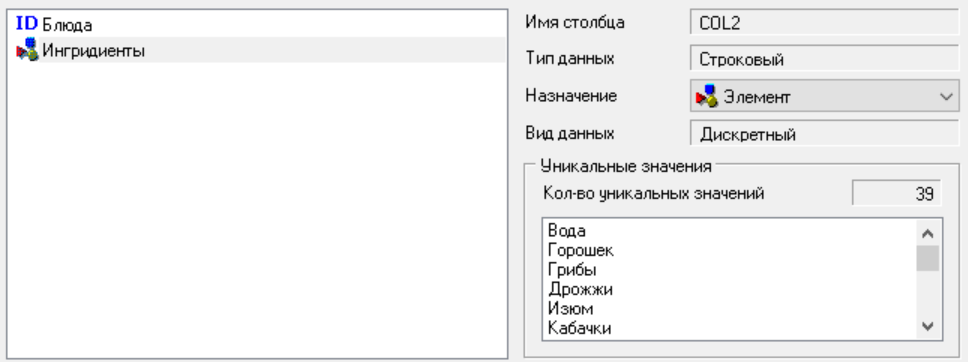


Рисунок 4 – Настройка столбцов для поиска ассоциаций

Был задан поиск правил, имеющих поддержку в диапазоне 10-80% и достоверность 60-90% (рисунок 5).

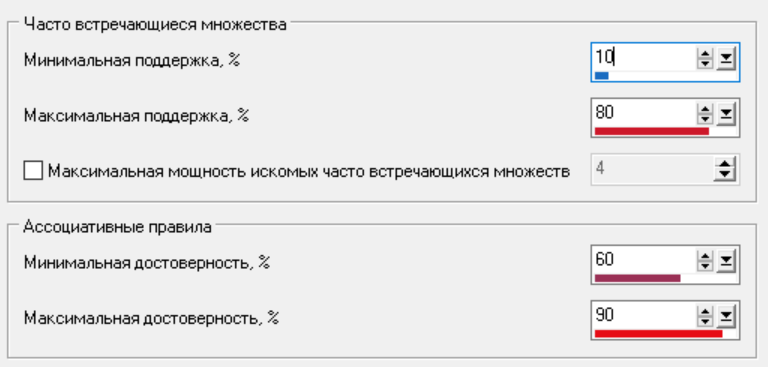


Рисунок 5 – Значения параметров построения ассоциативных правил

Процесс поиска ассоциативных правил был запущен. Предварительные данные показали, что было найдено 6 ассоциативных правил, а также 14 часто встречающихся множеств от 1 до 3 элементов (рисунок 6).



Рисунок 6 – Результаты поиска ассоциативных правил

В качестве способа отображения полученных данных были выбраны следующие визуализаторы: «Правила», «Популярные наборы», «Дерево правил» и «Что-если» (рисунок 7).

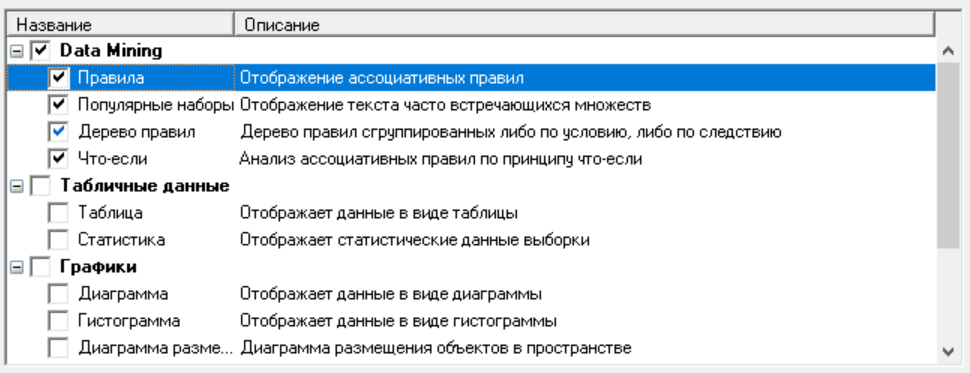


Рисунок 7 – Выбор способа визуализации

Первым был рассмотрен визуализатор Популярные наборы (рисунок 8). Из него можно сделать вывод, что наиболее часто в русской кухне используются лук (65%), яйца (50%), мука (53%).

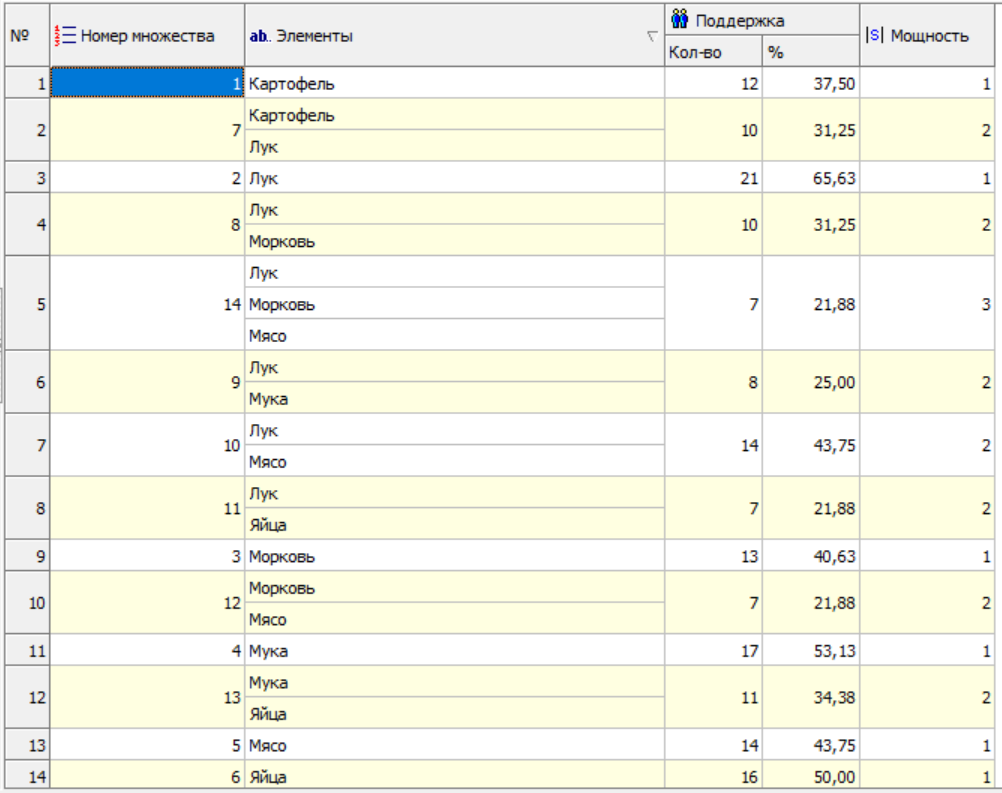


Рисунок 8 – Популярные наборы

Затем был рассмотрен визуализатор Правила (рисунок 9). По нему можно сделать ряд выводов о совместном использовании ингредиентов. Так лук встречается в блюдах с вероятностью 83% при появлении картофеля, с 76% при появлении моркови. Также при сочетании лука и моркови с вероятностью 70% в блюде имеется мясо.

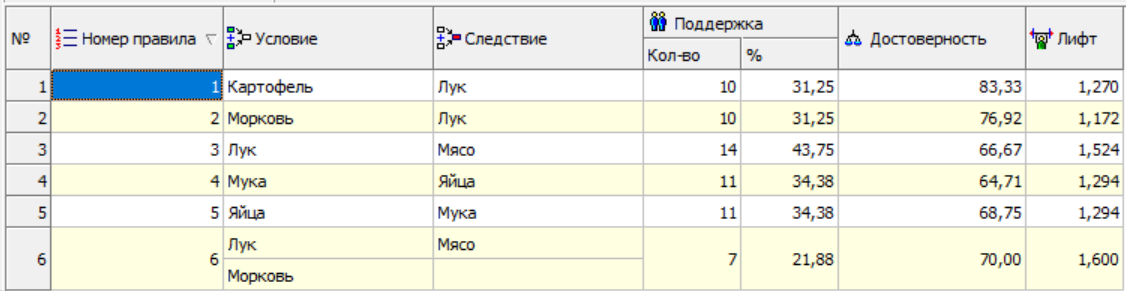


Рисунок 9 – Правила

Был рассмотрен визуализатор Дерево правил. Сначала было выведено дерево правил по условию (рисунок 10). На нём можно последить те же зависимости, что и в визуализаторе Правила. Каждый узел условия соединён с соответствующим следствием.

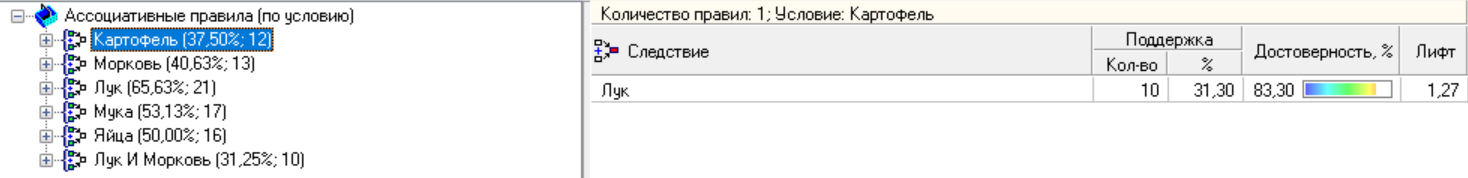


Рисунок 10 – Дерево правил (по условию)

Затем была рассмотрена вторая форма дерева правил – по следствию (рисунок 11). Информация в нём аналогична, однако на первом уровне располагаются узлы следствия. Например, можно сделать вывод, что лук с наибольшей вероятностью имеется в блюдах, где есть картофель (83%) или морковь (76%).

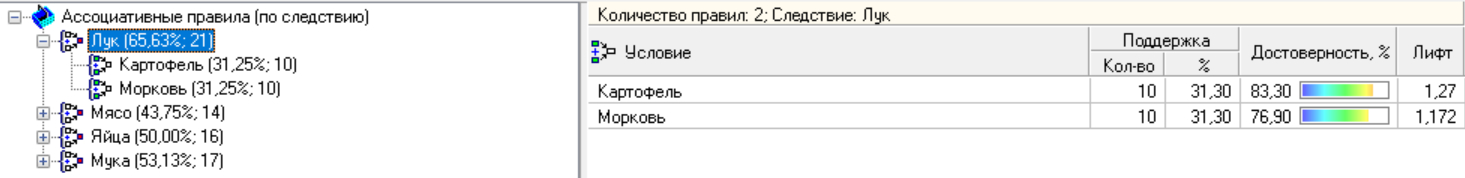


Рисунок 11 – Дерево правил (по следствию)

Наконец, был открыт визуализатор Что-если (рисунок 12). По нему можно определить, какие следствия могут возникнуть при определённых наборах правил. Например, если в одном блюде используются лук и яйца, то, вероятно, вместе с ними идут мука и мясо (вероятности 68% и 67%).

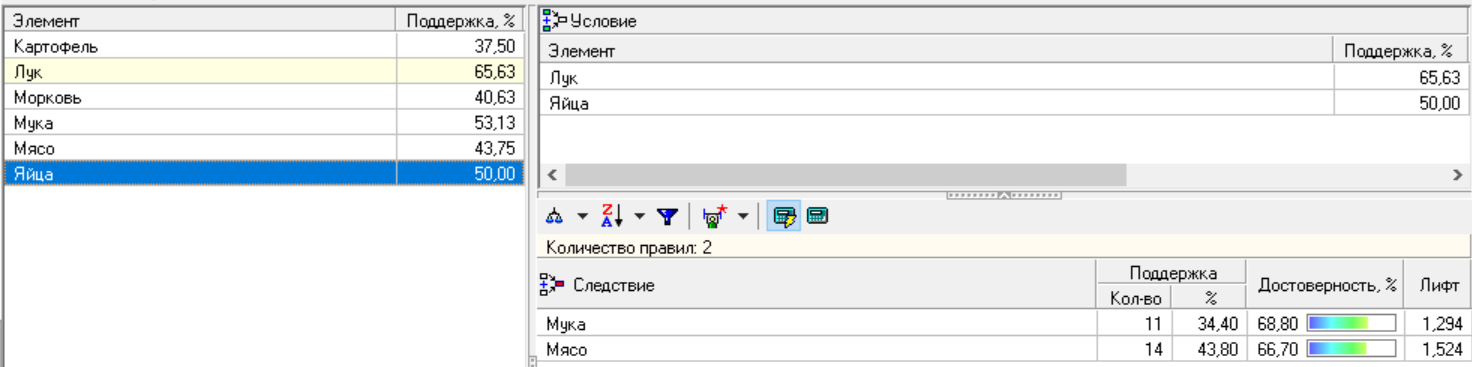


Рисунок 12 – Что-если

## 2.2 Анализ временного ряда

Для прогнозирования временного ряда был выбран набор данных о ежесуточной температуре в городе Москва на период 2021-2023 годов. Набор представлен в виде csv-файла (рисунок 13). Затем данные были экспортированы в Deductor Studio и представлены в виде таблицы (рисунок 14).

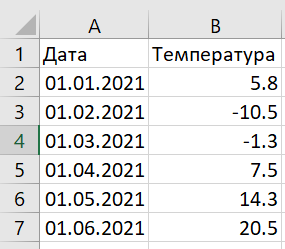


Рисунок 13 – Набор данных в Excel

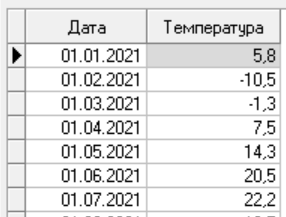


Рисунок 14 – Экспорт данных в Deductor (таблица)

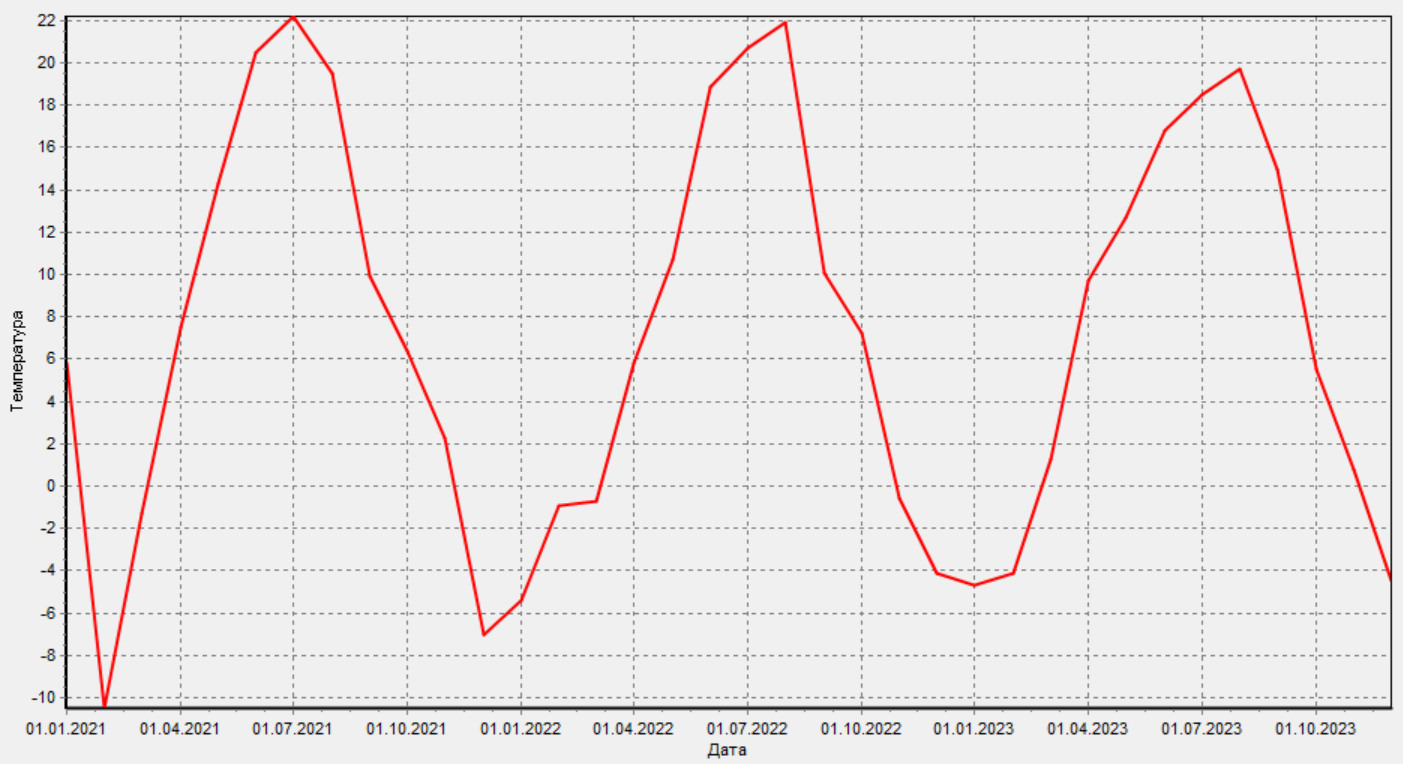


Рисунок 15 – Экспорт данных в Deductor (диаграмма)

Затем с помощью мастера обработки данные были прогнаны через редактирование выбросов и спектральную обработку, данные были сглажены, аномалии и шумы практически исчезли (рисунок 16).

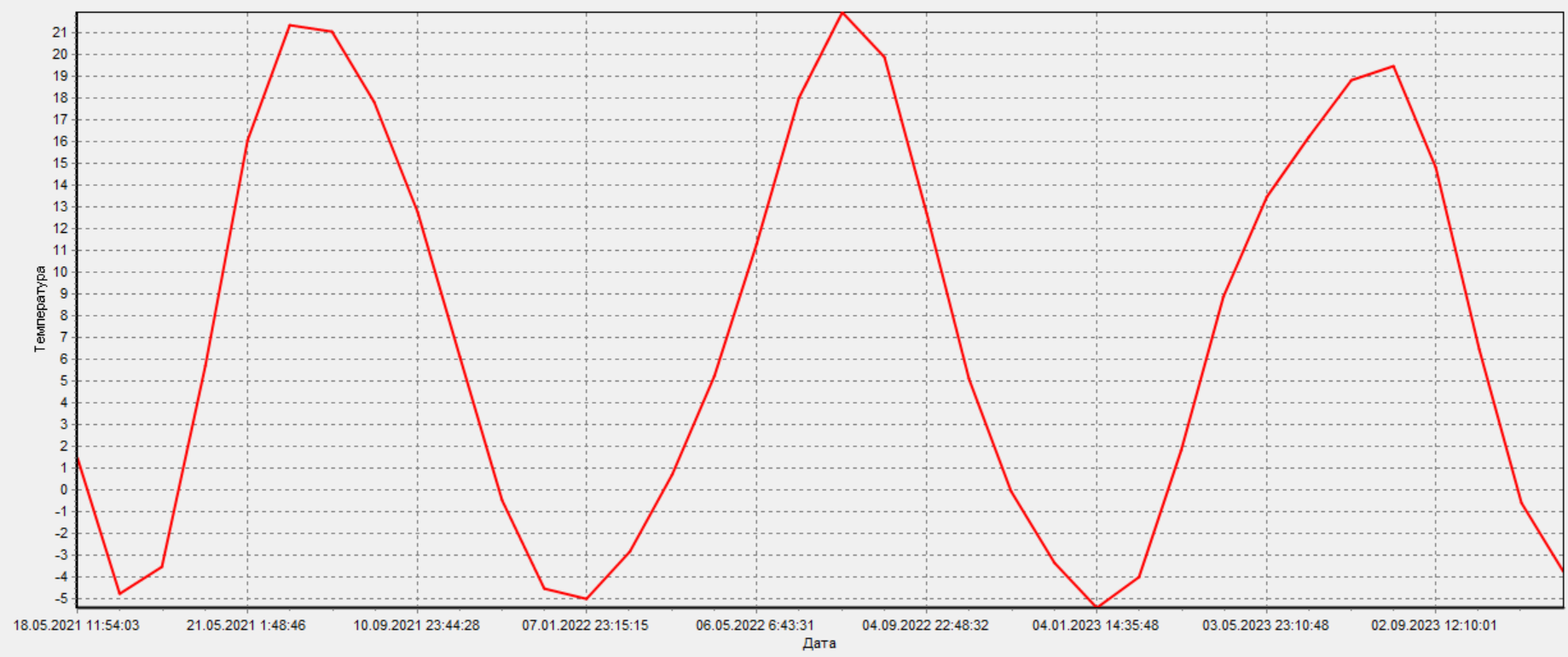


Рисунок 16 – Диаграмма после сглаживания

После этого данные были отобраны с помощью метода «Скользящее окно с глубиной погружения» 12 месяцев (рисунок 17).

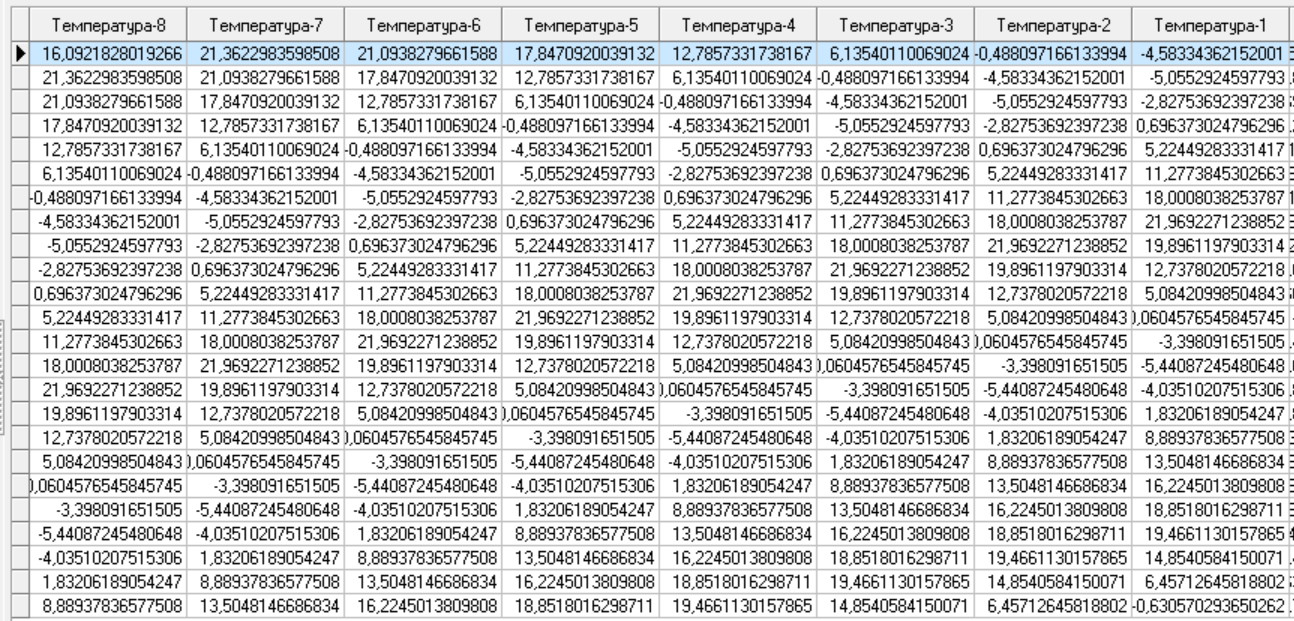


Рисунок 17 − Данные после использования метода «Скользящее окно»

Был открыт мастер обработки «Нейросеть». В качестве входных полей были установлены Температура-1, 2, 3 и 4, в качестве выходного – Температура, остальные в качестве информационных (рисунок 18).

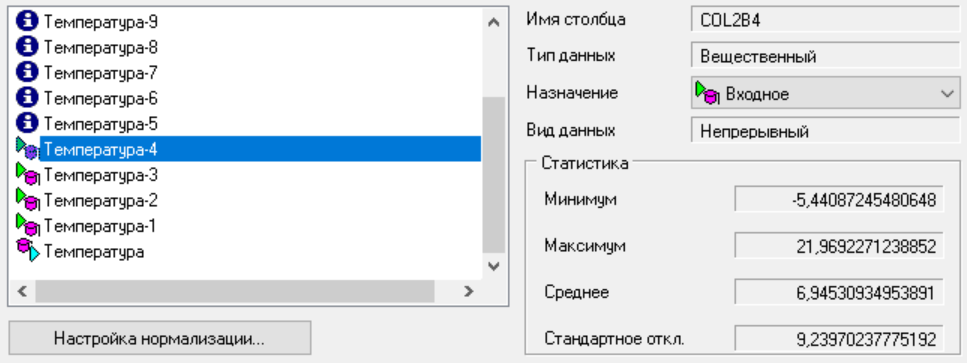


Рисунок 18 – Настройка назначений столбцов

На следующем шаге было указано разбиение тестового и обучающего множеств (рисунок 19).

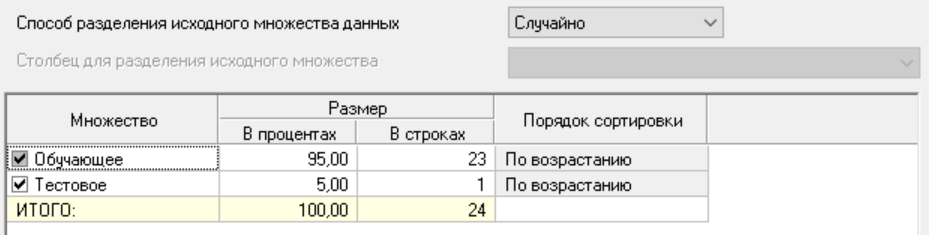


Рисунок 19 – Разбиение данных на подмножества

На следующем этапе было отмечено необходимое количество слоёв и нейронов в нейросети (рисунок 20). В результате была обучена нейросеть (рисунок 21).

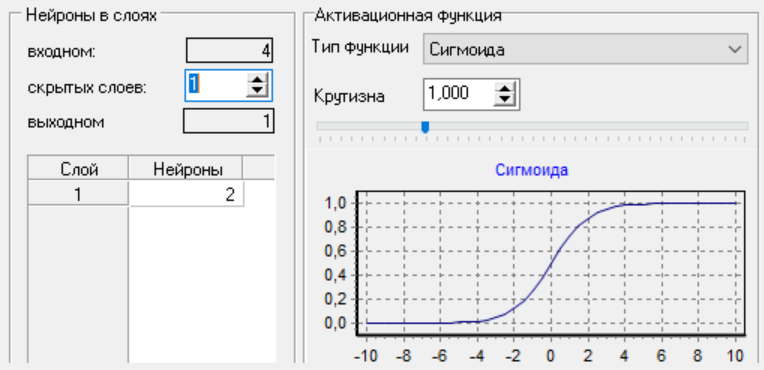


Рисунок 20 – Структура нейронной сети

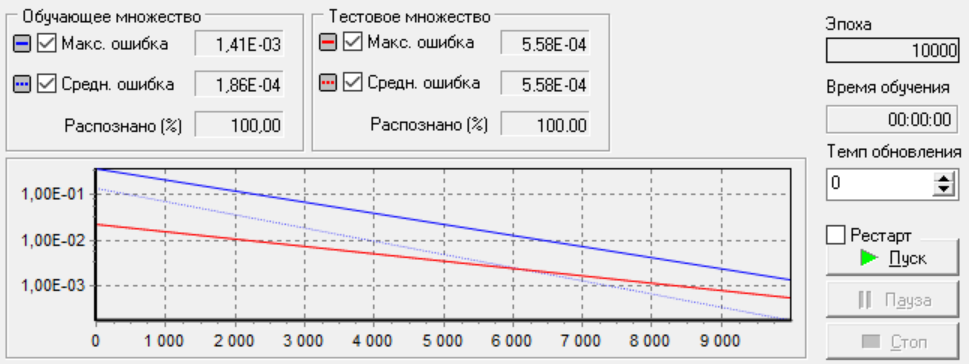


Рисунок 21 – Обучение нейронной сети

В результате обучения нейросети были выведены её граф (рисунок 22) и диаграмма рассеяния (рисунок 23).

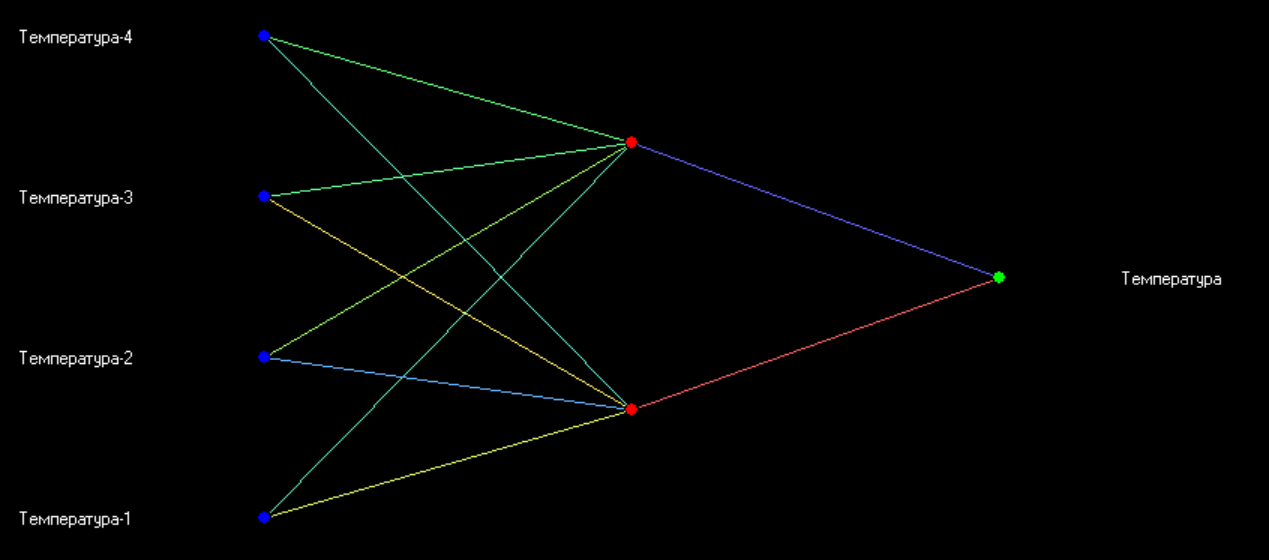


Рисунок 22 – Граф нейросети

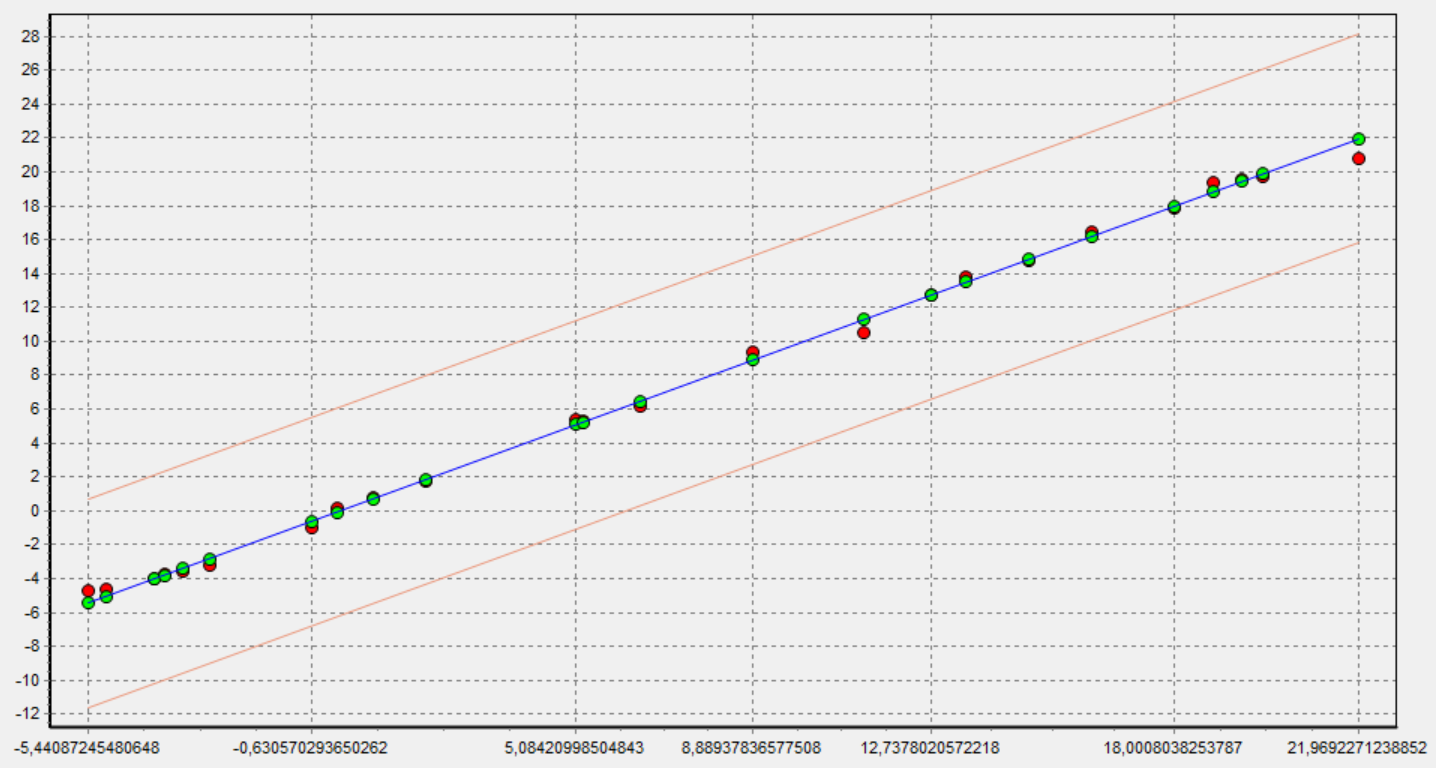


Рисунок 23 – Диаграмма рассеяния нейросети

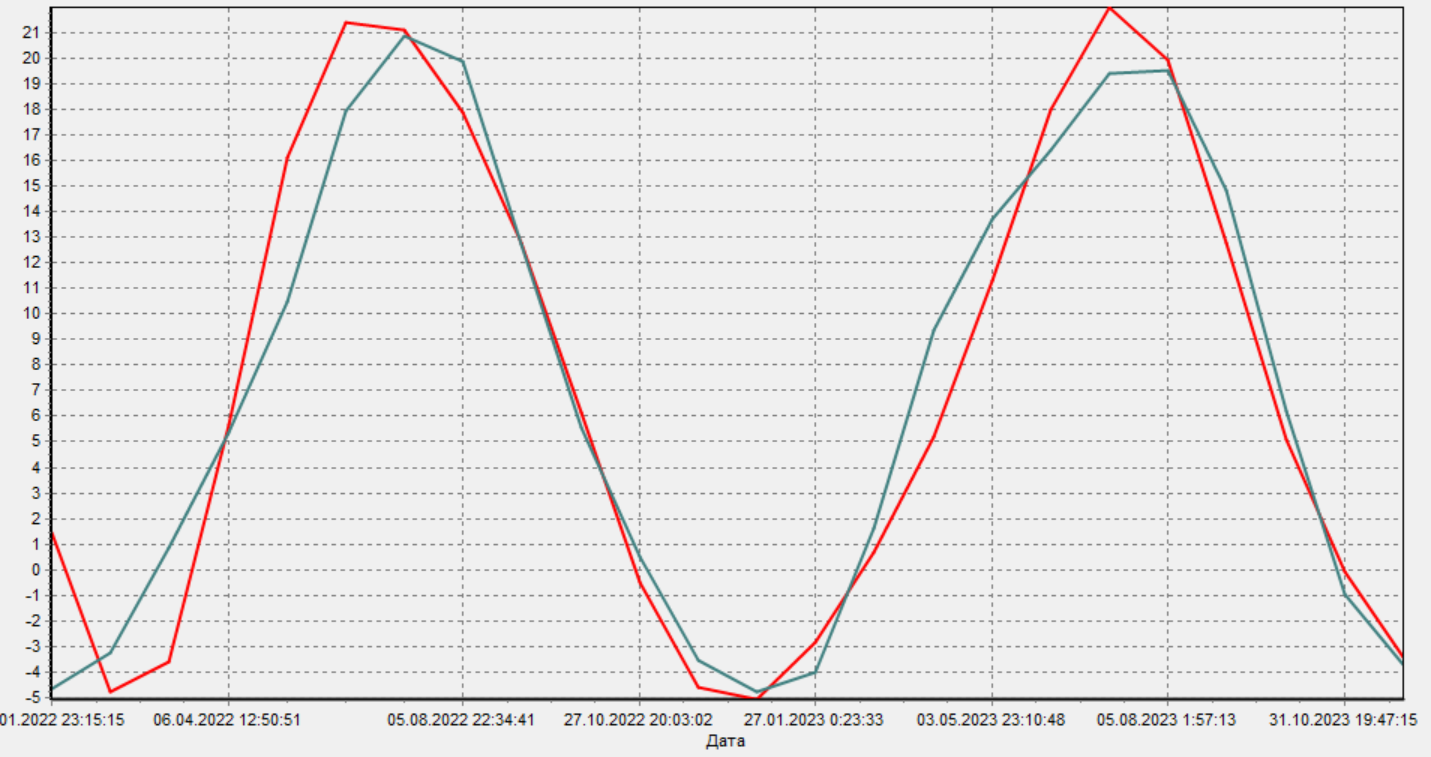


Рисунок 24 – Диаграмма данных нейросети

С помощью мастера обработки Прогнозирование с помощью имеющейся нейросети был построен прогноз временного ряда на следующие 5 месяцев (рисунок 25).

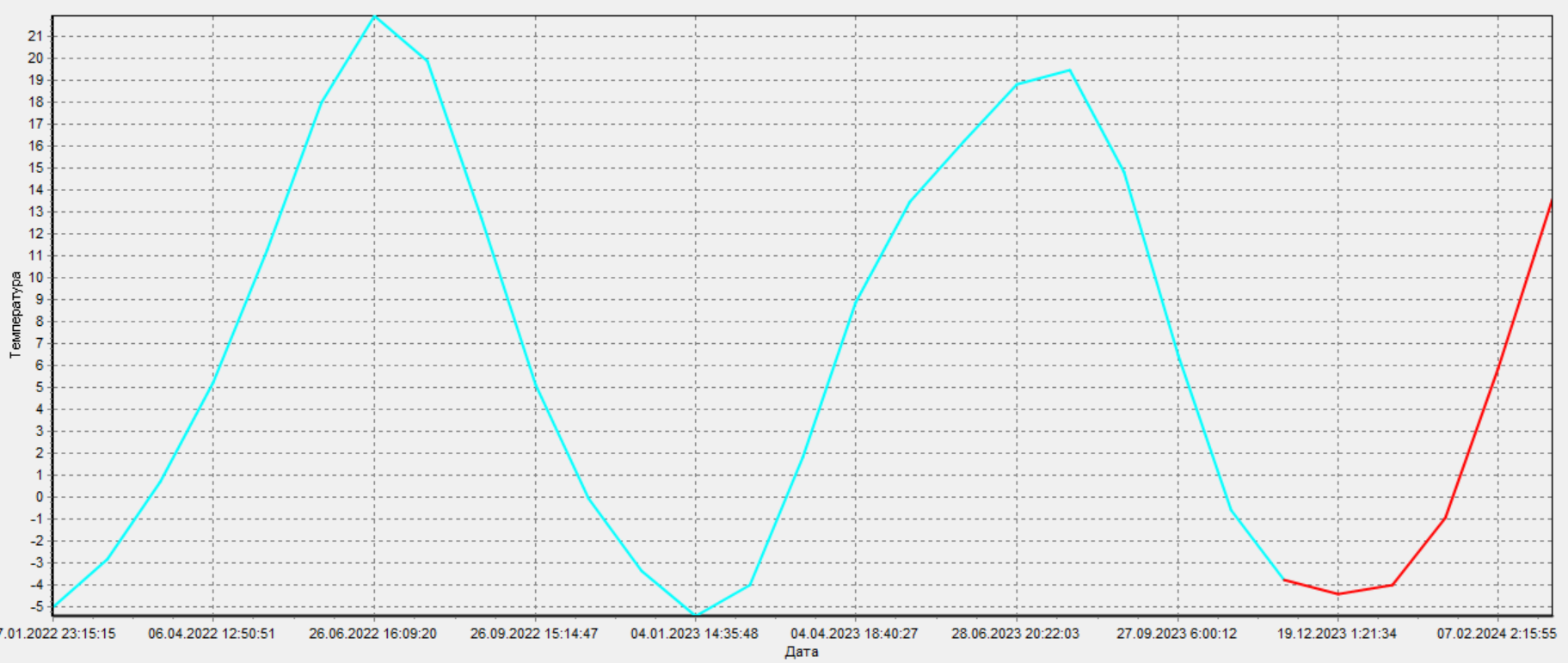


Рисунок 25 − Диаграмма прогноза

Голубой линией обозначены известные данные, красной – прогноз. Система спрогнозировала непродолжительный микроспад и последующий рост температуры.

## Вывод раздела 2

В данном разделе представлено практическое применение двух методик анализа данных – поиска ассоциативных правил и прогнозирования временных рядов.

На примере анализа набора данных о рецептах русской кухни была продемонстрирована эффективность поиска ассоциативных правил. С помощью него были выявлены закономерности в использовании определённых комбинаций ингредиентов, что может служить основой при подборе продуктов или вынесении суждений об особенностях русской кухни.

В качестве примера прогнозирования временного ряда были взяты ежемесячные данные о температуре воздуха в Москве за 3 года. На имеющихся данных была построена модель прогнозирования температуры на 5 месяцев 2024 года. Сравнение прогноза с реальными данными за этот период не выявило серьёзных отклонений (не считая резкого похолодания в январе).

# ЗАКЛЮЧЕНИЕ

В ходе выполнения расчётно-графической работы были продемонстрированы навыки практического применения методов анализа данных. В качестве методик анализа были рассмотрены поиск ассоциативных правил и прогнозирование временных рядов.

Поиск ассоциативных правил был проведён на наборе данных о наиболее популярных блюдах русской кухни. Были проанализированы закономерности в сочетании ингредиентов в рамках одного блюда. В результате были выявлены наиболее часто встречающиеся из них, вычислены вероятности появления одних ингредиентов в случае наличия других, определены случаи, где эти вероятности наиболее высоки. Это позволило сформировать рекомендации по сочетанию продуктов в блюдах и формированию общего видения русской кухни.

Прогнозирование временного ряда было проведено для данных о среднемесячной температуре воздуха в городе Москва за три года. Диаграмма данных была сглажена с помощью спектральной обработки, был использован метод Скользящее окно с глубиной 12 месяцев. На этих данных была обучена нейросеть прогнозирования погоды. Сравнение прогноза нейросети на первое полугодие 2024 года с имеющими данными на момент выполнения работы не выявило критических различий.

Таким образом, в ходе работы была продемонстрирована эффективность применения методик анализа данных для извлечения полезной информации из наборов экспериментальных данных. Информация, полученная методами Data Mining, может как давать основу для суждений о предметных областях, так и предсказывать будущие тенденции их развития.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Кацко И.А., Паклин Н.Б. Практикум по анализу данных на компьютере: Учеб. пособие для вузов. — М.: Издательство "КолосС", 2009. — 278 с. — 978-5-9532-0624-2.
2. Deductor. Руководство аналитика. Версия 5.2.
3. Годфред, Дэвид. "Мастеринг Data Mining: методы и инструменты. Вильямс", 2012. – 672 с.
4. Дюк Б., Самойленко А. Data Mining: учебный курс. — М.: Издательство: Питер, 2001. — 368 с. — 5-318-00227-7.
5. Вагнер, Майкл. "Применение Data Mining и машинного обучения в бизнесе: Практическое руководство", 2023. – 126 с. /