|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ** | | | | | | | | | | | | | | | | | |
| **ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ** | | | | | | | | | | | | | | | | | |
| **УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ** | | | | | | | | | | | | | | | | | |
| **«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»** | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| Институт информационных технологий | | | | | | | | | | | | | | | | | | |
| (полное название института) | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| кафедра «Информационные системы» | | | | | | | | | | | | | | | | | | |
| (полное название кафедры) | | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| **Пояснительная записка** | | | | | | | | | | | | | | | | | |
| к расчетно-графической работе | | | | | | | | | | | | | | | | | |
| по дисциплине «Интеллектуальный анализ данных» | | | | | | | | | | | | | | | | | |
| на тему: | **Поиск ассоциативных правил и анализ временных рядов в среде** | | | | | | | | | | | | | | | | |
| **Deductor** | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
|  | | | |  | | |  | | |  | | | | | | | |
| Выполнили: студенты | | | | 3 | | | курса, группы: | | | ИС/б-20-2-о | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| Направления подготовки (специальности) | | | | | | | | | 09.03.02 | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| Информационные системы и технологии | | | | | | | | | | | | | | | | | |
| (код и наименование направления подготовки (специальности)) | | | | | | | | | | | | | | | | | |
| профиль (специализация) | | | | | | Информационные системы и технологии | | | | | | | | | | | |
| Белик Г. М. | | | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| (фамилия, имя, отчество студента) | | | | | | | | | | | | | | | | | |
| Руководитель | |  | | | | | | | | | | | | | | | |
|  | | (фамилия, инициалы, степень, звание, должность) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | | | |
| Защита « |  | | » | |  | | | 20 |  | | г. | |  | | Оценка |  | |
|  | | | | | | |  | | | | | | | | | | |
| Руководитель | | | | | | |  | | | | |  | |  | | | |
|  | | | | | | | (подпись) | | | | |  | | (инициалы, фамилия) | | | |
|  | | | | | | |  | | | | |  | |  | | | |
| Севастополь | | | | | | | | | | | | | | | | | |
| 20 | | | | | | | | 23 | г. | | | | | | | |

ВВЕДЕНИЕ

«Data Mining», также определенный как анализ больших данных или интеллектуальный анализ данных, является крайне актуальным инструментом, который используется для работы в различных современных областях науки и бизнеса, например: медицина, государственное управление, маркетинг, розничная торговля и многие другие.

Целью работы является изучение основ методов анализа экспериментальных данных и освоить технику их практического применения в Deductor Studio.

Для достижения поставленных целей необходимо выполнить следующие пункты:

1. Решение задачи поиска ассоциаций в Deductor
   1. Подготовить данные для поиска ассоциативных данных.
   2. Провести поиск ассоциативных правил, проанализировать полученные результаты Прогнозирование временного ряда.
2. Прогнозирование временного ряда.
   1. Подготовить данные для прогнозирования временного ряда
   2. Провести анализ временного ряда
   3. Провести прогнозирование временного ряда на 12 месяцев.

1 ОБЗОР МЕТОДОВ DATA MINING

Интеллектуальный анализ данных – это направление информационных технологий, охватывающее всю область проблем, связанных с извлечением знаний из массивов данных.

Часто интеллектуальный анализ данных рассматривают как трансформацию термина Data Mining (раскопка, разработка данных) и считают два этих понятия синонимами. В широком смысле это вполне допустимо. Однако считать их абсолютно идентичными было бы неверно, поскольку методы Data Mining сосредотачиваются главным образом на процессах анализа данных и интеллектуальном моделировании. В то же время считается, что интеллектуальный анализ данных рассматривает весь спектр проблем, связанных с процессом извлечения знаний из баз данных.

В основе интеллектуального анализа данных, как и в Data Mining, лежит идея активного применения математических методов, таких как оптимизация, генетические алгоритмы, распознавание образов, статистика, Data Mining и т.д., а также использующих визуальное представление информации.

Модели интеллектуального анализа данных могут применяться в конкретных бизнес-сценариях, а именно:

* Прогнозирование: оценка продаж, прогнозирование нагрузки сервера или времени простоя сервера;
* Риск и вероятность: выбор наиболее подходящих заказчиков для целевой рассылки, определение точки равновесия для рискованных сценариев, назначение вероятностей диагнозам или другим результатам;
* Рекомендации: определение продуктов, которые с высокой долей вероятности могут быть проданы вместе, создание кросс-пакетов;
* Поиск последовательностей: анализ выбора заказчиков во время совершения покупок, прогнозирование следующего возможного события;
* Группировка: разделение заказчиков или событий на кластеры связанных элементов, анализ и прогнозирование общих черт.

Внушительные объемы современных баз данных вызвали устойчивый спрос на новые масштабируемые алгоритмы анализа данных. Одним из популярных методов обнаружения знаний стали алгоритмы поиска ассоциативных правил.

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:

Полезные правила – содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение. Такие правила могут быть использованы для принятия решений, приносящих выгоду.

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.

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 ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING

* 1. Поиск ассоциативных правил

Подготовка данных.



Рисунок 1 – Данные для ассоциативного анализа

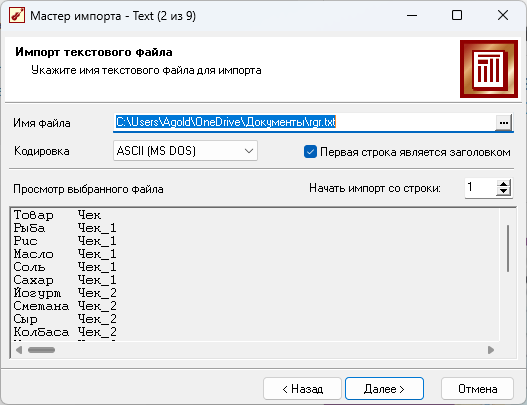


Рисунок 2 – Шаг 1. Указывание пути файла

Сформируем текущие данные в таблицу, состоящую из двух столбцов: Товар и Чек. Полученную таблицу импортируем в Deductor Studio.

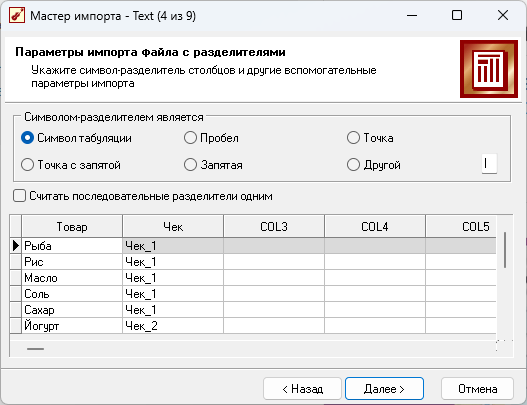


Рисунок 3 – Шаг 2. Импорт файла

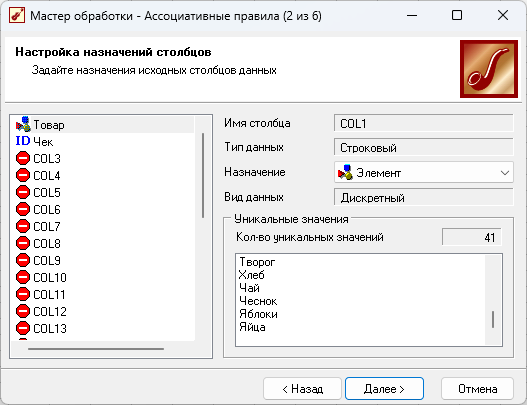


Рисунок 4 – Шаг 3. Указываем какой столбец идентификатор и элемент

Для поиска ассоциативных правил воспользуемся Мастером обработки, где выберем тип обработки “Ассоциативные правила”.

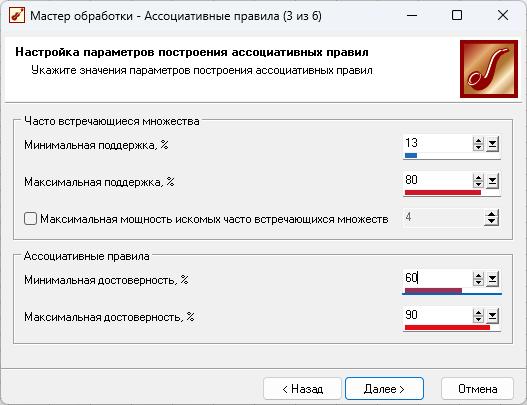


Рисунок 5 – Шаг 1. Настройка параметров построения правил

Следующий шаг позволяет запустить процесс поиска ассоциативных правил. На экране отображается информация о количестве множеств и найденных правил, а также числе часто встречающихся множеств.

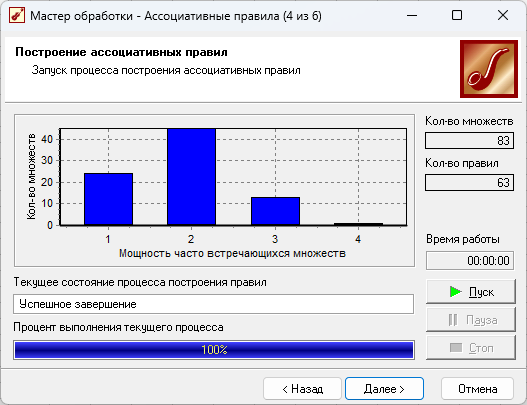


Рисунок 6 – Шаг 2. Результат поиска ассоциативных правил

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.

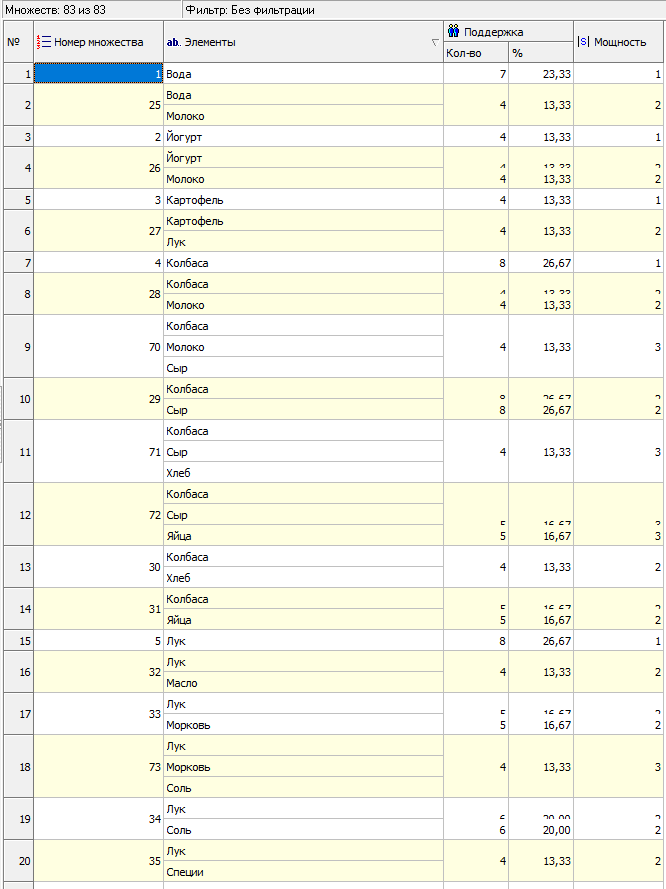


Рисунок 7 – Популярные наборы

Исследуя визуализатор поиска ассоциативных правил “Популярные наборы”, можно сделать вывод, что такие продукты, как: Вода, Йогурт, Картофель, Колбаса, Лук, Майонез, Масло, Молоко, Морковь являются часто используемыми в данной списке греческих рецептов.

Визуализатор "Правила" отображает ассоциативные правила в виде списка правил, которые описывают поведение покупателей.

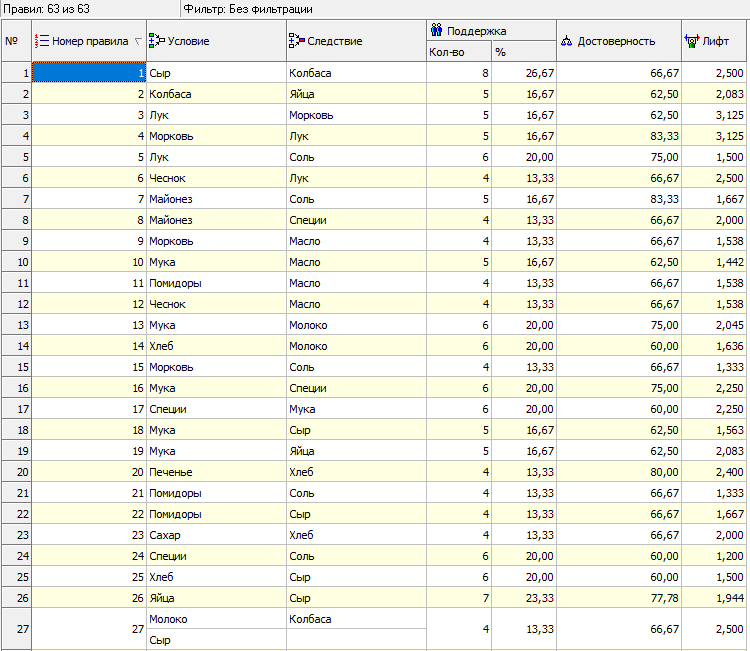


Рисунок 8 – Визуализатор Правила

Из полученных результатов на рисунке видно, что при использовании человеком Молока и Масла, с вероятностью 66,67% он использует Сыр и Яйца, а при покупке Моркови он приобретет Лук или Соль с вероятностью 66,67%.

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.

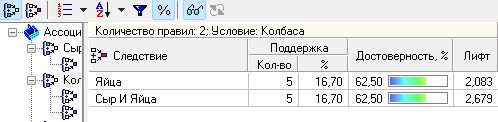


Рисунок 9 – Построение дерева правил по условию

Узлы Чеснок, Колбаса, Лук, Морковь находятся на верхнем уровне дерева и представляют собой условия. А Сыр, Яйца и т.д. – следствия. Это означает, что человек, использующий Колбасу, использует еще и Яйца достоверностью 62,5%, Сыр с достоверностью 62,5%.

Второй вариант дерева правил – дерево, построенное по следствию. Здесь на первом уровне располагаются узлы со следствием.

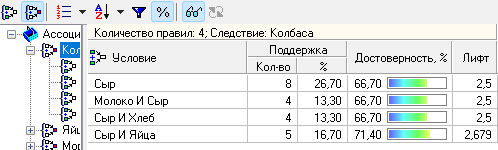


Рисунок 10 – Построение дерева правил по следствию

Например, для того чтобы человек приобрел Колбасу, он должен купить хотя бы один предмет из следующего списка: Сыр, Молоко, Хлеб, Яйца и т.д.

Анализ “Что-если” в ассоциативных правилах позволяет ответить на вопрос, что получим в качестве следствия, если выберем данные условия. Например, какие товары приобретаются совместно с выбранными товарами.

Пусть необходимо проанализировать, что, возможно, забыл человек купить, если он уже прибрел Колбасу и Лук. Для этого следует добавить в список условий эти ингредиенты и затем нажать на кнопку "Вычислить правила". При этом в списке следствий появятся ингредиенты, совместно приобретаемые с данными.

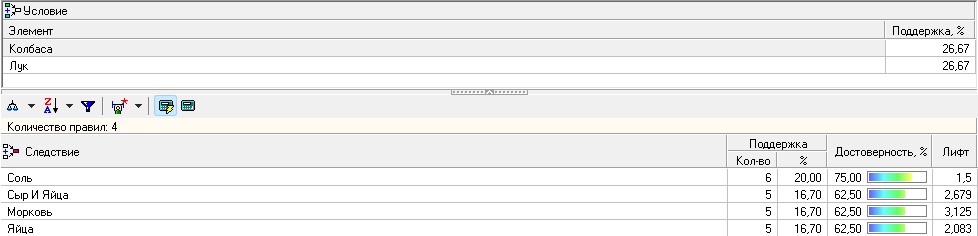


Рисунок 11 – Результат анализа Что-Если

В данном случае появится Соль, Сыр, Яйца, Морковь, т.е. именно эти ингредиенты человек возможно забыл купить с вероятностью 62-75%.

* 1. Прогнозирование временного ряда

Для анализа временного ряда была выбран эмиграция из России за 20 лет.

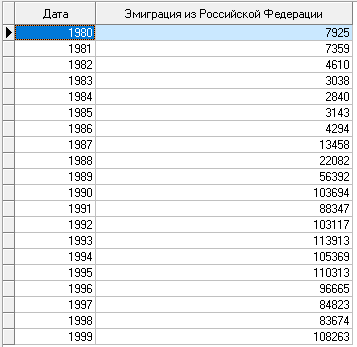


Рисунок 11 – Данные для анализа временного ряда

После импорта данных воспользуемся диаграммой для их просмотра и проведем удаление аномалий и сглаживание данных.

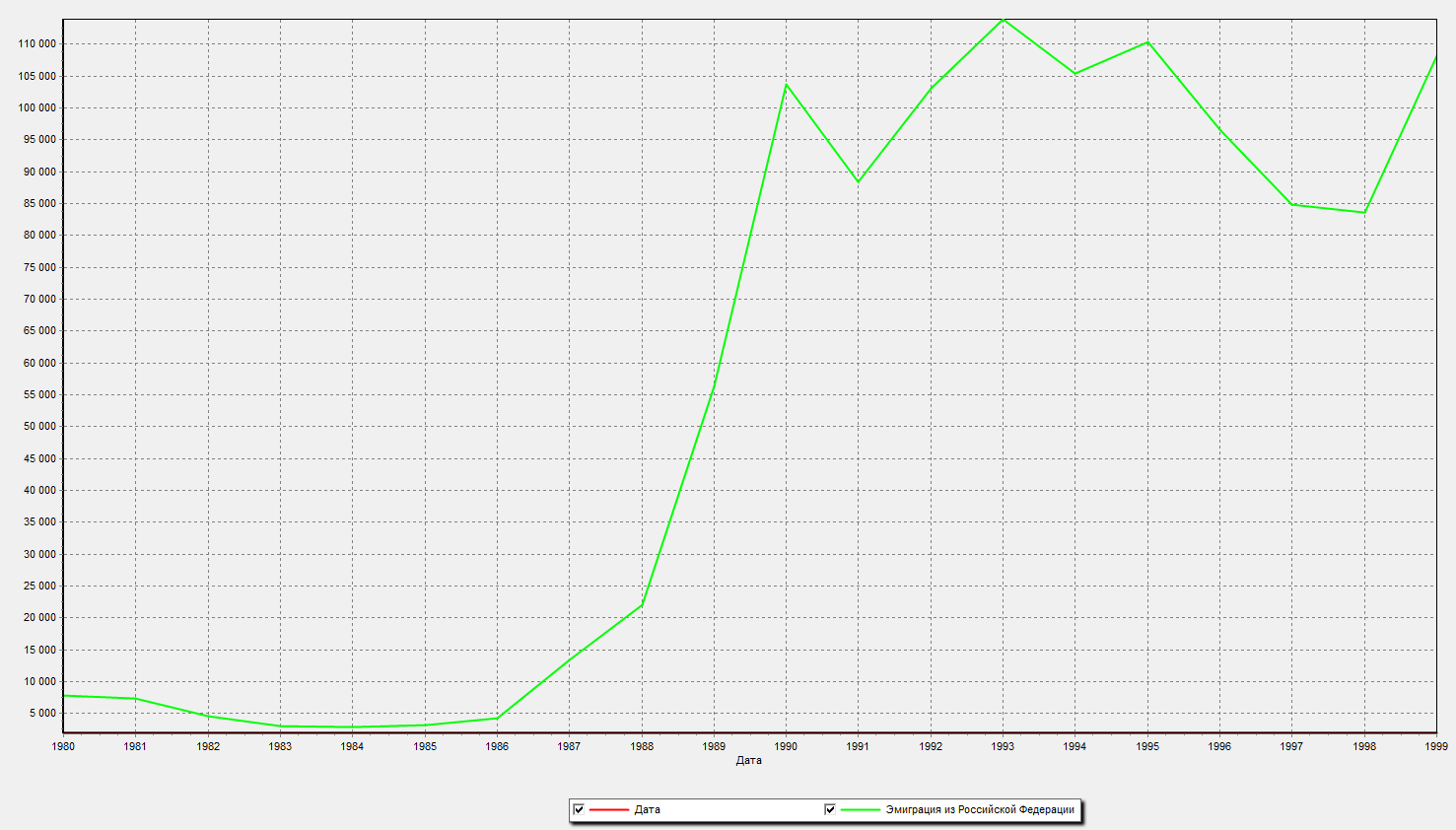


Рисунок 12 – Диаграмма до обработки

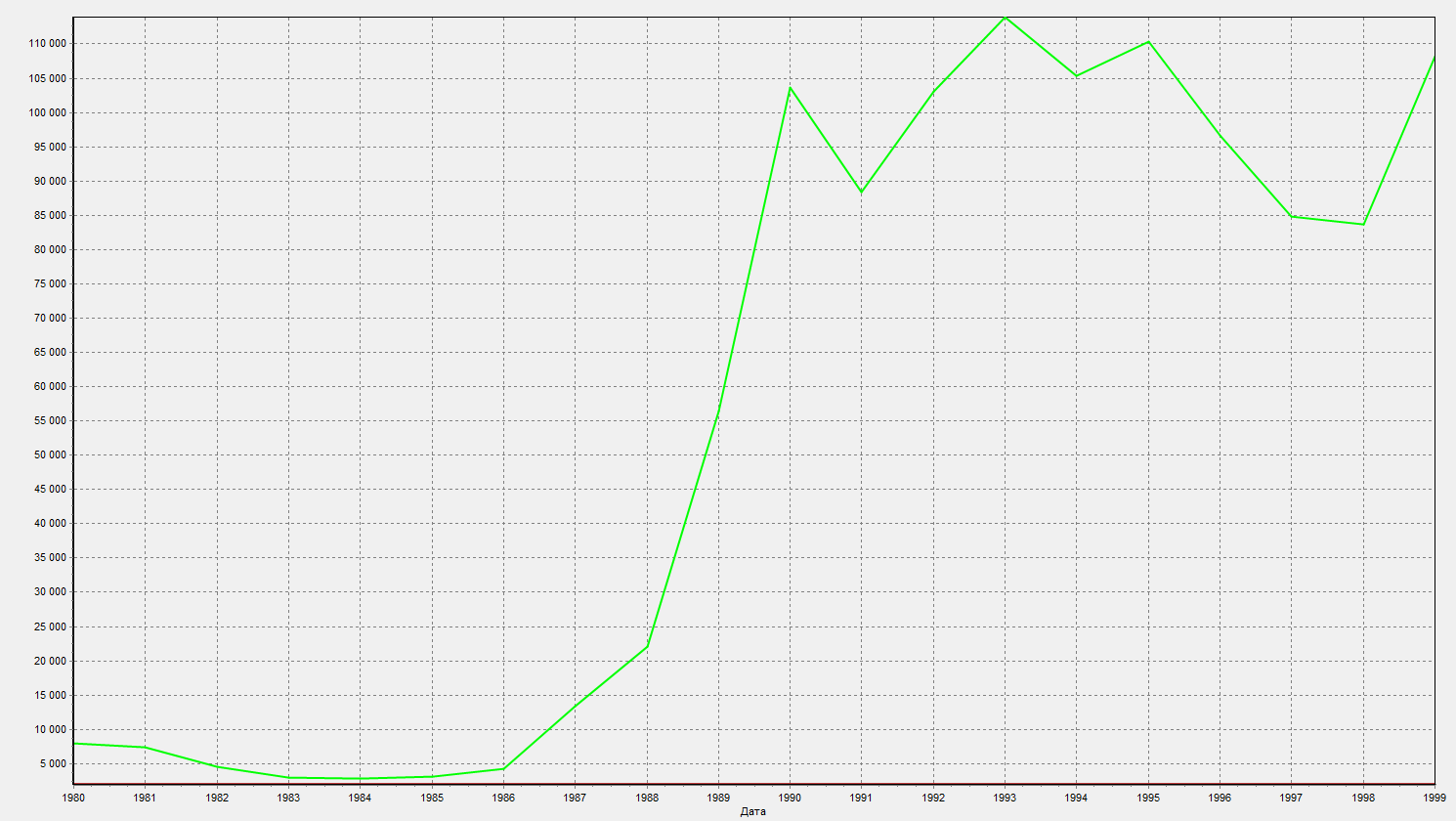


Рисунок 13 – Диаграмма после обработки

Отберем данные используя метод “Скользящее окно” с глубиной погружения 10 лет.

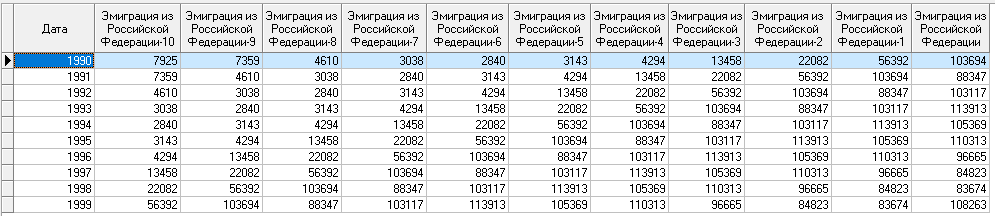


Рисунок 14 – Часть данных после использования метода “Скользящее окно”

Перейдем к обучению нейросети для дальнейшего построения прогноза.

В качестве входных полей будут поля с Эмиграцией-1 до Эмиграцией -10, а в качестве выходного Эмиграция.

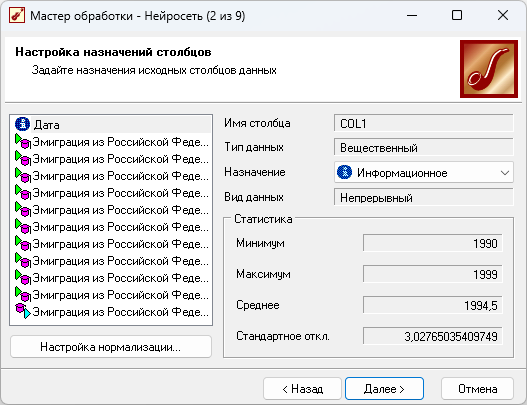


Рисунок 15 – Шаг 1. Настройка назначений столбцов

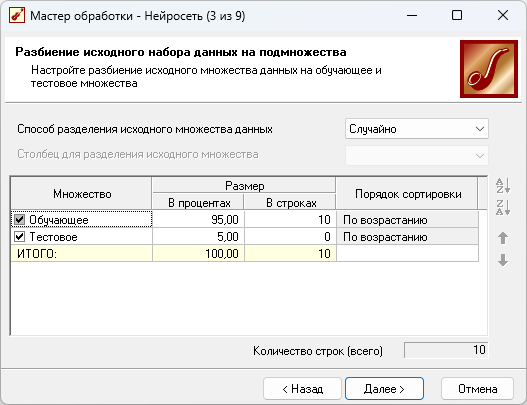


Рисунок 16 – Шаг 2. Разбиение исходного набора данных на подмножества

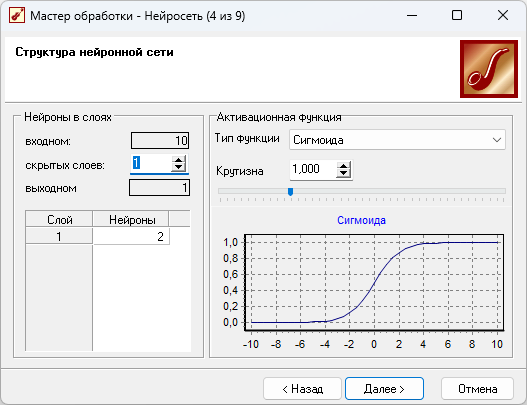


Рисунок 17 – Шаг 3. Настройка структуры нейронной сети

Результатом обучения нейросети будет две диаграммы:

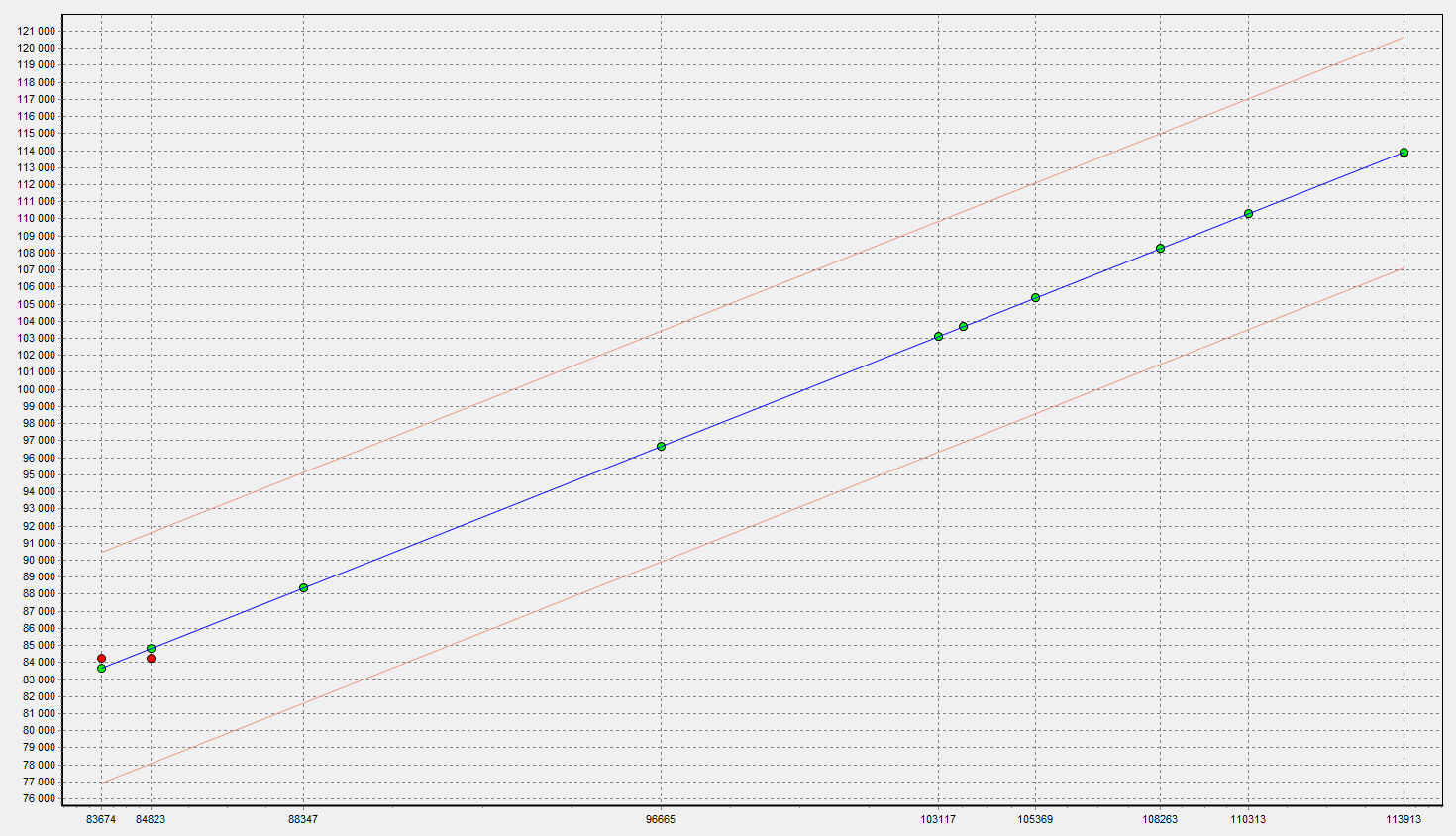


Рисунок 18 – Диаграмма рассеяния

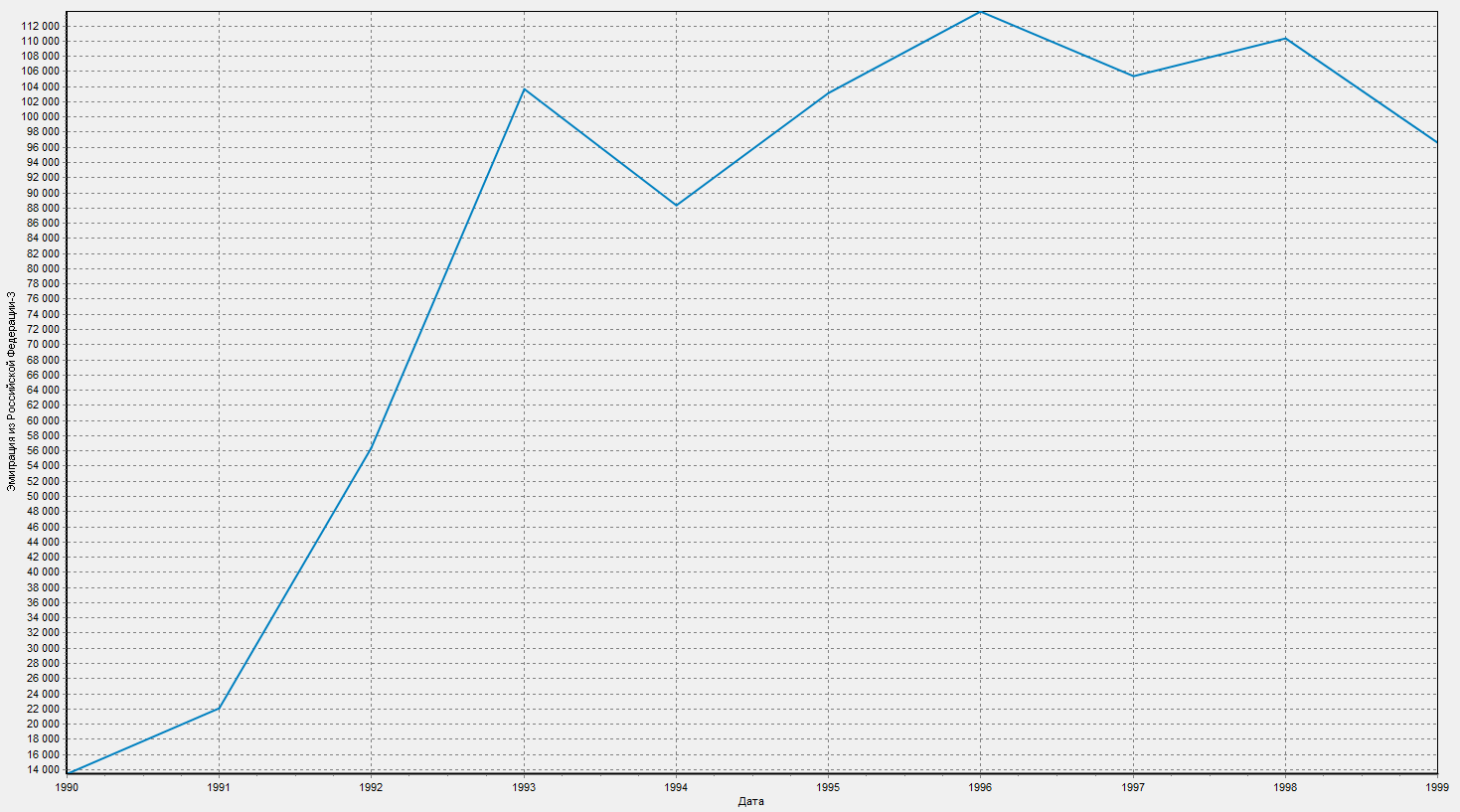


Рисунок 19 – Диаграмма по данным

После того, как нейросеть была обучена, построим прогноз с помощью обработчика “Прогнозирование”.

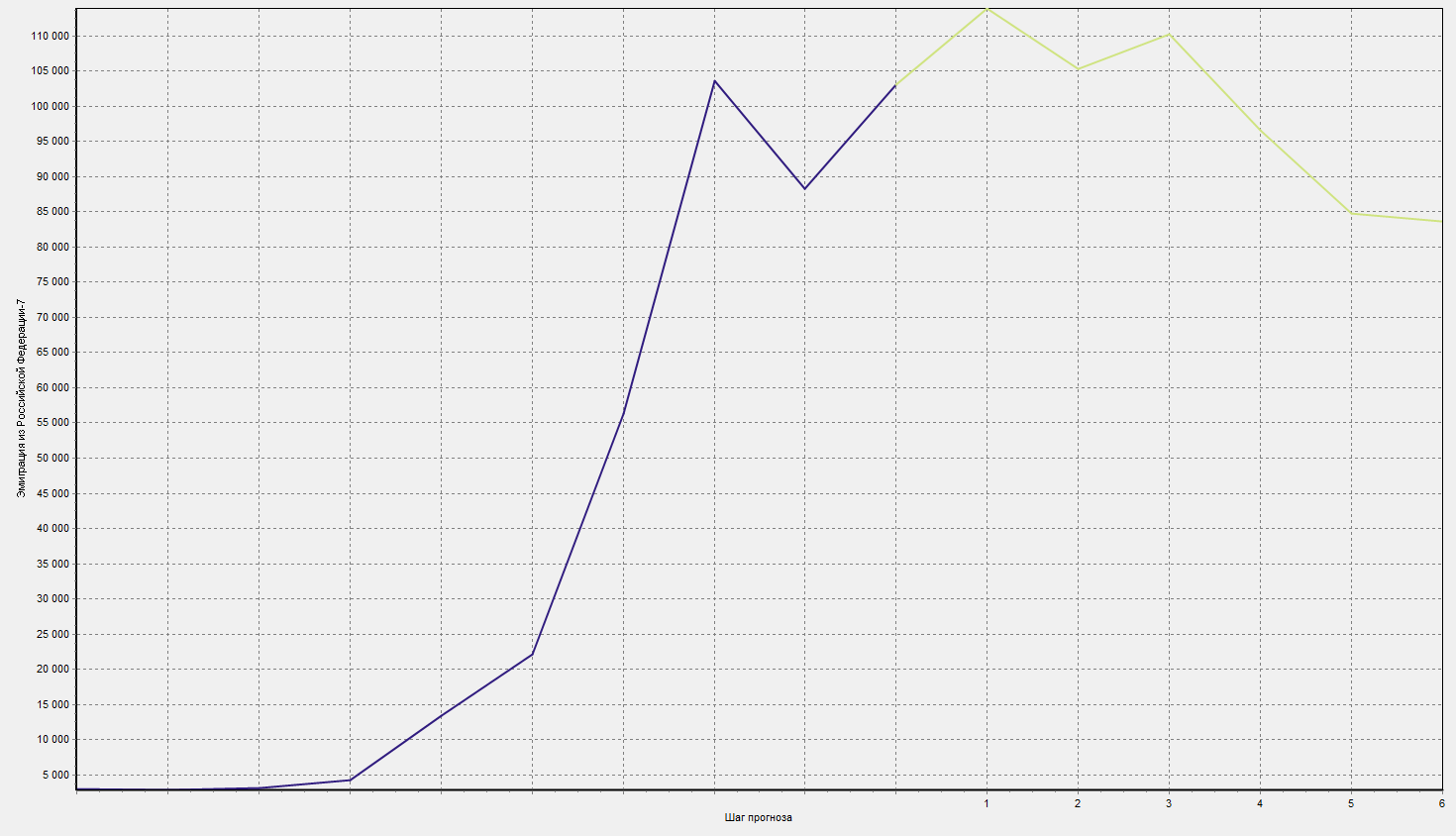


Рисунок 20 – Результат прогнозирования

На диаграмме прогнозирования желтой линией отображено изменение эмигрантов в следующие 5 лет. Система спрогнозировала небольшой спад эмигрантов со 100000 до 85000 ежегодно.

ВЫВОДЫ

В ходе выполнения расчетно-графического задания был проведен поиск ассоциативных правил для данных, представляющих собой чеки в продуктовом магазине. Были выявлены популярные наборы: Хлеб, Молоко, Колбаса, Сыр, Лук, Морковь. В визуализаторах “Правила”, “Дерево правил” и “Что-если” были определены условия и вероятности того, что еще человек купит, если он уже приобрел выбранный товар.

Также было проведено прогнозирование временного ряда эмиграции из России за 20 лет. 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 10 лет. Проведено обучение нейросети и построен прогноз изменения эмигрантов на следующие 10 лет. Система спрогнозировала среднее количество эмигрантов в 85000 людей.