Министерство науки и высшего образования Российской Федерации

Новосибирский Государственный технический университет

Кафедра автоматизированных систем управления



**Отчет по лабораторной работе 3**

**по дисциплине «Методы анализа данных»**

**«Классификация»**

Выполнили

студенты группы АВТ-812:

Бородина Алина

Березин Дмитрий

Глинин Евгений

Преподаватель:

Ганелина Наталья Давидовна,

к.т.н., доцент кафедры АСУ

г. Новосибирск

2021 г.

**Содержание**

[Цель работы и постановка задачи 3](#_Toc88252718)

[Описание исходных данных 4](#_Toc88252719)

[Результаты дескриптивного анализа 5](#_Toc88252720)

[Описание параметров метода (сети) 6](#_Toc88252721)

[1 Дерево решений 6](#_Toc88252722)

[2 Нейронные сети 10](#_Toc88252723)

[2.1 Стандартизация (нормировка) данных 10](#_Toc88252724)

[2.2 Построение нейронной сети 10](#_Toc88252725)

[Сравнительный анализ решений и интерпретация результатов 14](#_Toc88252726)

[Вывод 19](#_Toc88252727)

# Цель работы и постановка задачи

Цель работы: изучение методов решения задачи классификации.

Среда реализации: Deductor Academic.

Задание:

1. Выбрать массив данных, описать параметры.
2. Провести дескриптивный анализ, определить, согласуются ли выборки с нормальным распределением.
3. Стандартизировать переменные (при необходимости).
4. Изучить пример решения задачи классификации
5. Решить задачу классификации двумя методами (нейронная сеть, дерево решений, метод опорных векторов, дискриминантная функция и т.д.). Аргументировать выбор метода. Допустимо использовать две нейронные сети разных типов.
6. Оценить качество построенных моделей (в т.ч. точность, полноту классификатора). Провести сравнительный анализ решений.
7. Изменить параметры одного из методов (или параметры нейросети), оценить полученные результаты.
8. Интерпретировать результаты.

# Описание исходных данных

Был выбран массив данных «Caesarian Section Classification Dataset». Этот набор данных содержит информацию о результатах кесарева сечения у 80 беременных женщин с наиболее важными характеристиками проблем с родами в медицинской сфере.

Мы выбираем возраст, номер поступления, время поступления, артериальное давление и состояние сердца.

Время поступления классифицируется в датасете на преждевременное, своевременное и опоздание. Как и время поступления, артериальное давление рассматривается в трех состояниях: низкое, нормальное и высокое. Заболевания сердца разделяются на допустимые для кесарева сечения случаи и недопустимые.

Параметры:

1. Возраст (непрерывная);
2. Номер поступления (целое число: 1,2,3,4)

3.  Время поступления (целое число: 0 = своевременное, 1 = преждевременное, 2 = опоздание);

4. Кровяное давление (целое число: 0 = низкое, 1 = нормальное, 2 = высокое);

5. Заболевания сердца (целое число: 0 = есть, 1 = отсутствуют);

6. Кесарево сечение (целое число: 0 – нет, 1 – да).

# Результаты дескриптивного анализа

Проведем дескриптивный анализ выборки (рисунок 1).

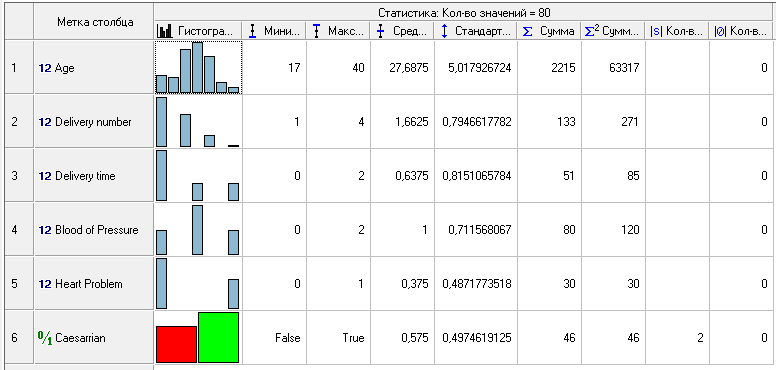


Рисунок 1 – Дескриптивный анализ выборки

На нормальное распределение похожа выборка по параметру «Возраст» и «Кровяное давление». Выборки по параметрам «Номер поступления» и «Время поступления» имеют правую асимметрию. Выборка по параметру «Заболевания сердца» также обладает правой асимметрией.

# Описание параметров метода (сети)

В выбранной среде реализации (Deductor Academic) для решения задач классификации можно воспользоваться двумя методами:

1. Дерево решений
2. Нейронная сеть

## 1 Дерево решений

Настроим значения столбцов. Все параметры кроме «Кесарево сечение» будут обладать входным назначением, а выходным будет «Кесарево сечение» (рисунок 2).

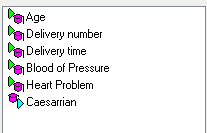


Рисунок 2 – Настройка назначений столбцов

Настроим разбиение исходного множества данных на обучающее и тестовое множества (рисунок 3).

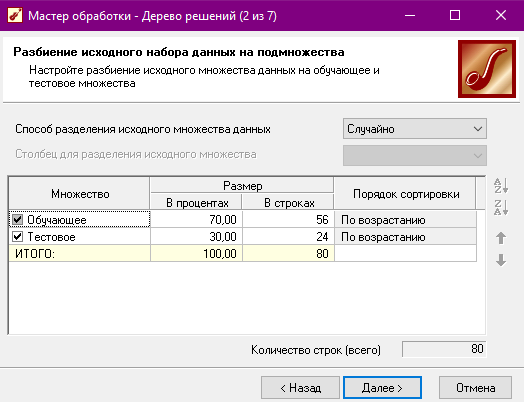


Рисунок 3 – Разбиение исходного набора данных на подмножества

Укажем значения параметров обучения дерева решений (рисунок 4).

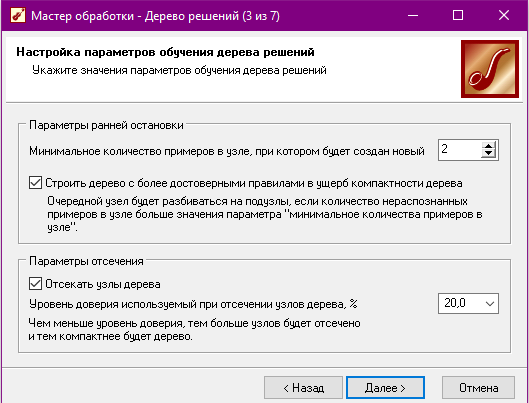


Рисунок 4 – Настройка параметров обучения дерева решений

Укажем желаемый способ построения дерева решений (см. рисунок 5).

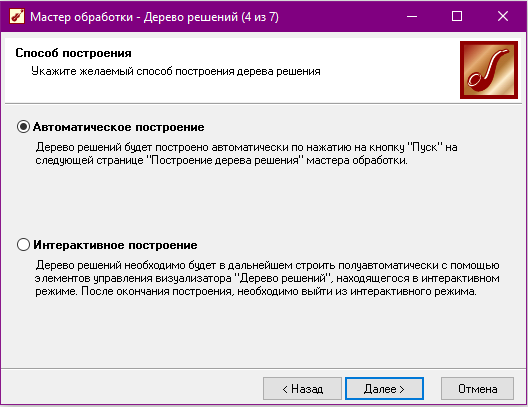


Рисунок 5 – Выбор способа построения дерева решений

Запустим процесс построения дерева решений (рисунок 6).

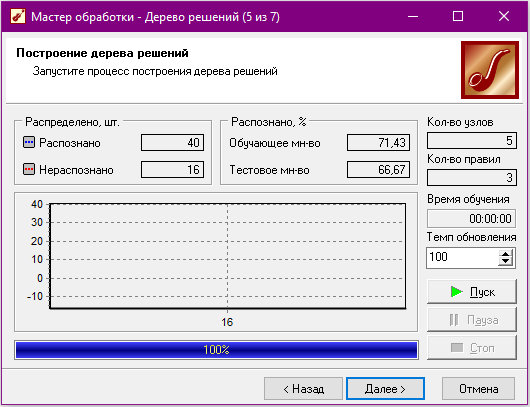


Рисунок 6 – Построение дерева решений

После окончания процесса построения дерева решений получаем модель оценки подлинности образцов банкнот (рисунок 7).

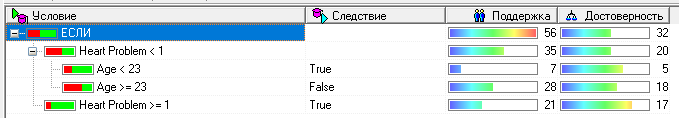


Рисунок 7 – Модель оценки подлинности образцов банкнот

Формализованные правила классификации, выраженные в форме «Если

<Условие>, тогда <Класс>», можно увидеть, выбрав визуализатор «Правила» (рисунок 8).

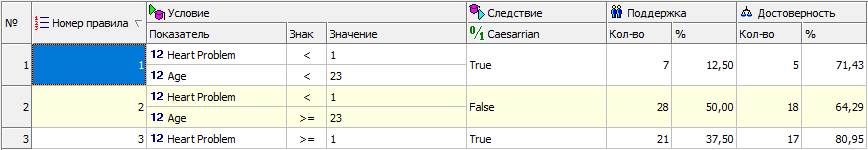


Рисунок 8 – Правила классификации

Важную информацию предоставляет визуализатор «Значимость атрибутов». С помощью него можно определить, насколько сильно выходное поле зависит от каждого из входных факторов. Чем больше значимость атрибута, тем больший вклад он вносит при классификации. Самым значимым атрибутом является параметр «Заболевания сердца» (см. рисунок 9).

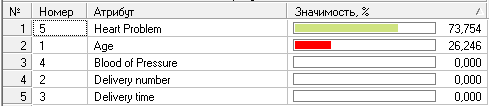


Рисунок 9 – Значимость атрибутов

Часто аналитику бывает полезно узнать, сколько примеров было распознано неверно, какие именно примеры были отнесены к какому классу ошибочно. На этот вопрос дает ответ визуализатор «Таблица сопряженности» (рисунок 10). Далее таблица сопряженности понадобится для оценки качества построенной модели (вычисление точности и полноты классификатора).

По диагонали таблицы расположены примеры, которые были правильно распознаны, в остальных ячейках – те, которые были отнесены к другому классу. В данном случае дерево правильно классифицировало большинство примеров.

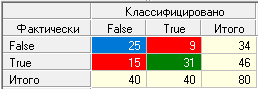


Рисунок 10 – Таблица сопряженности

## 2 Нейронные сети

### 2.1 Стандартизация (нормировка) данных

Все входные поля для нейронной сети должны быть представлены в числовом виде. Для этого все поля приводятся к диапазону значений [a, b] (в нашем случае мы приводили к диапазону [-1, 1]). Для полей с непрерывным видом данных в Deductor Academic можно использовать только Линейный нормализатор. Для нормализации используется формула:

– текущее значение, , – минимальные и максимальные значения поля соответственно, по умолчанию для входных полей , , для выходных , .

### 2.2 Построение нейронной сети

Далее настроим разбиение исходного множества данных на обучающее и тестовое множества (рисунок 11).

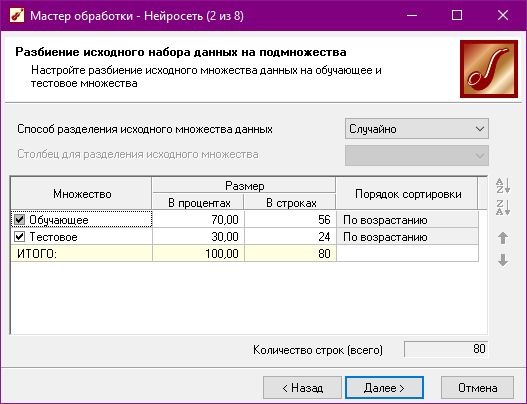


Рисунок 11 – Разбиение исходного набора данных на подмножества

Количество нейронов первого скрытого слоя экспериментально было выставлено в значение 1. Активационной функцией была выбрана сигмоида с величиной крутизны 1 (рисунок 12).

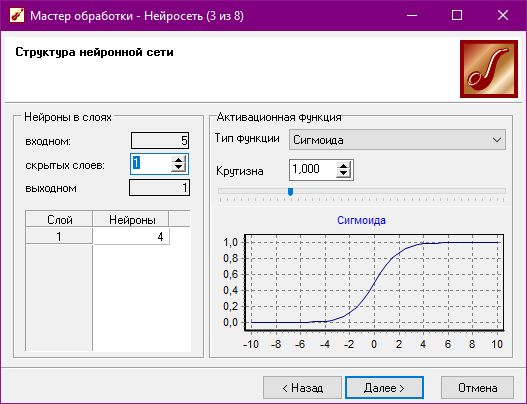


Рисунок 12 – Структура нейронной сети

Выберем алгоритм и зададим параметры обучения (рисунок 13).

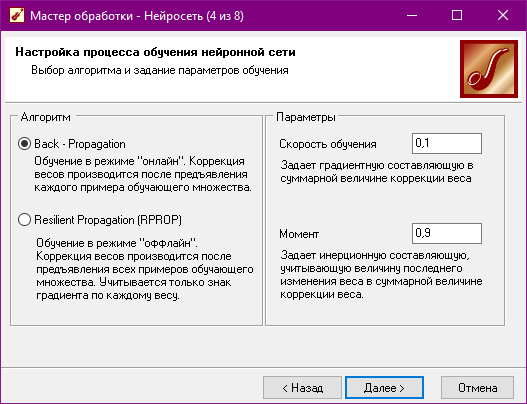


Рисунок 13 – Выбор алгоритма обучения

Укажем условия прекращения обучения (рисунок 14).

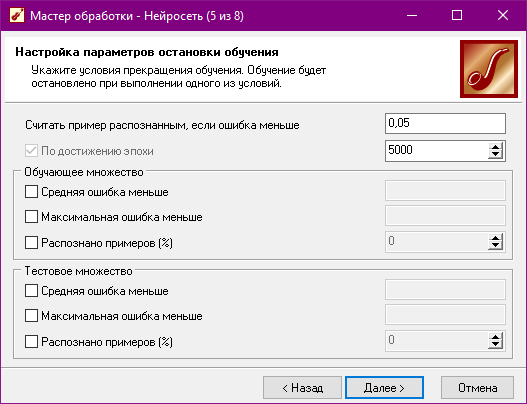


Рисунок 14 – Настройка параметров остановки обучения

Запустим процесс обучения нейронной сети (рисунок 15).

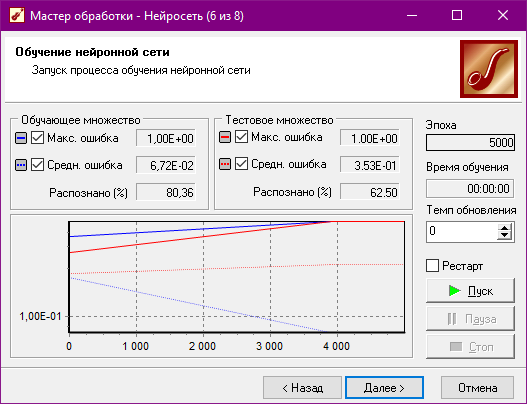


Рисунок 15 – Обучение нейронной сети

На графе нейросети видно, как выглядит обученная сеть (рисунок 16).

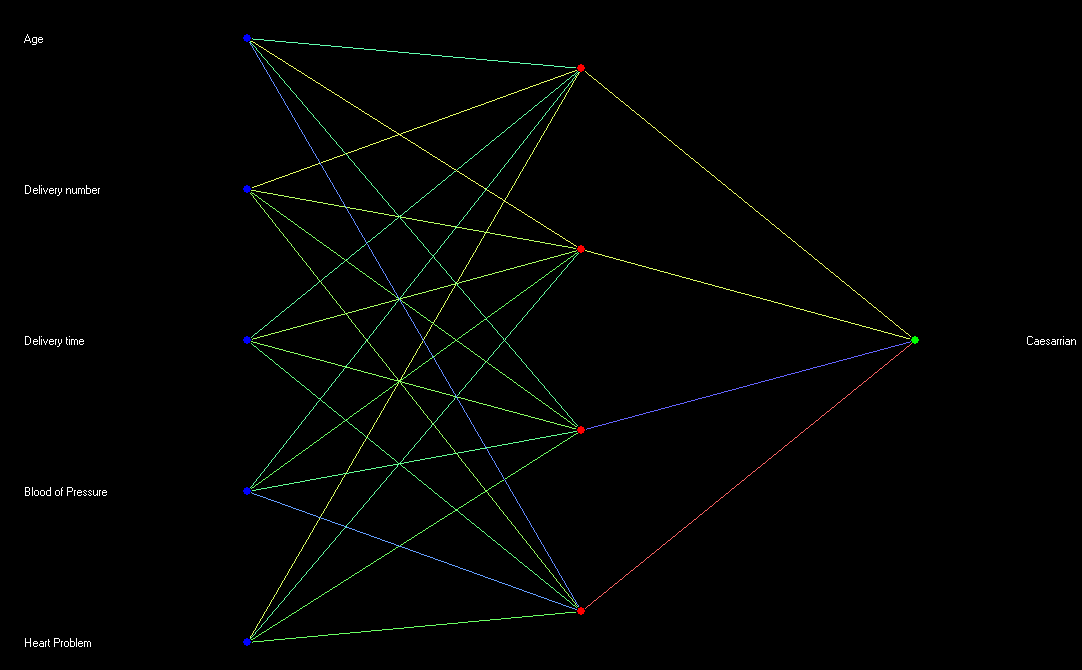


Рисунок 16 – Граф нейросети

На рисунке 17 приведена таблица сопряженности, на которой показано, что 14 примеров было распознано неверно.

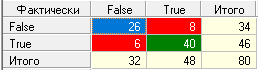


Рисунок 17 – Таблица сопряженности

# Сравнительный анализ решений и интерпретация результатов

Оценим качество построенных моделей (точность и полноту классификатора). Проведем сравнительный анализ решений.

Обратимся к рисунку 12 – таблица сопряженности для оценки качества построенного дерева решений:

Обратимся к рисунку 17 – таблица сопряженности для оценки качества обученной нейронной сети:

Отметим, что обученная нейронная сеть справилась с задачей классификации лучше (), чем построенное дерево решений .

Далее изменим параметры одного из методов (параметры нейросети) и

оценим полученные результаты.

Проведем эксперименты по изменению параметров обучения нейросети, и сделаем выводы по эффективности процесса обучения при разном количестве нейронов.

Изменим количество скрытых слоёв до 2 и количество нейронов скрытого слоя 1 будет 4, а 2 будет 2 (рисунок 18).

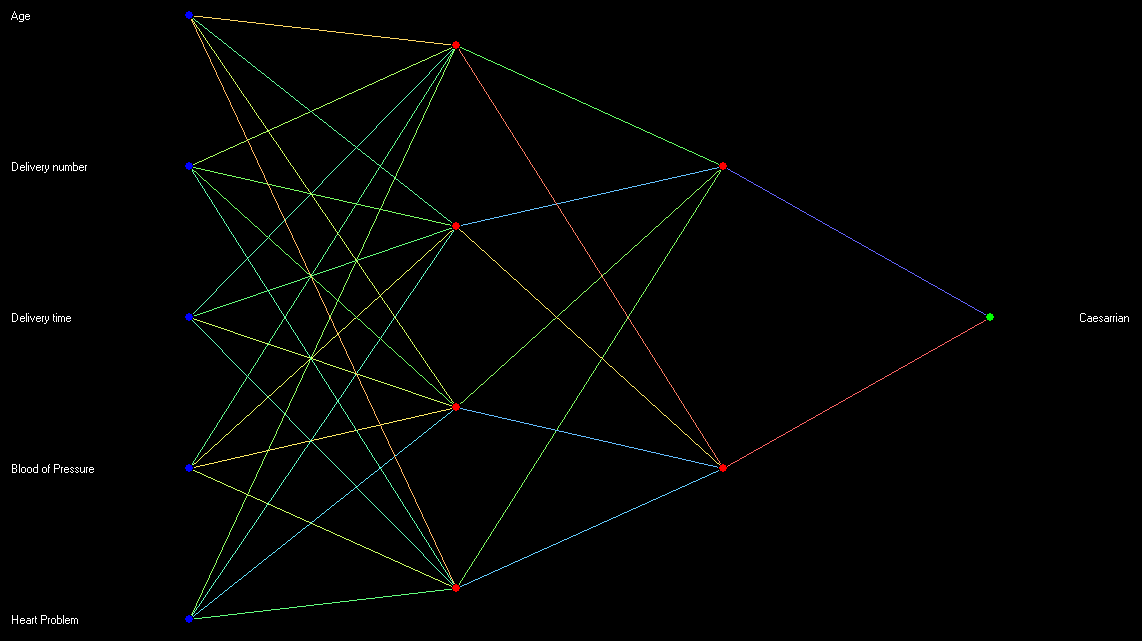


Рисунок 18 – Граф нейросети

На рисунке 19 приведена таблица сопряженности, на которой показано, что 7 примеров было распознано неверно.

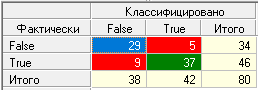


Рисунок 19 – Таблица сопряженности

Тогда получим следующие метрики качества классификации:

**С увеличением количества скрытых слоёв полнота классификатора повысилась (была: 0,78 стала: 0,805).**

В этот раз изменим крутизну с 1 до 0,7 (см. рисунок 20).

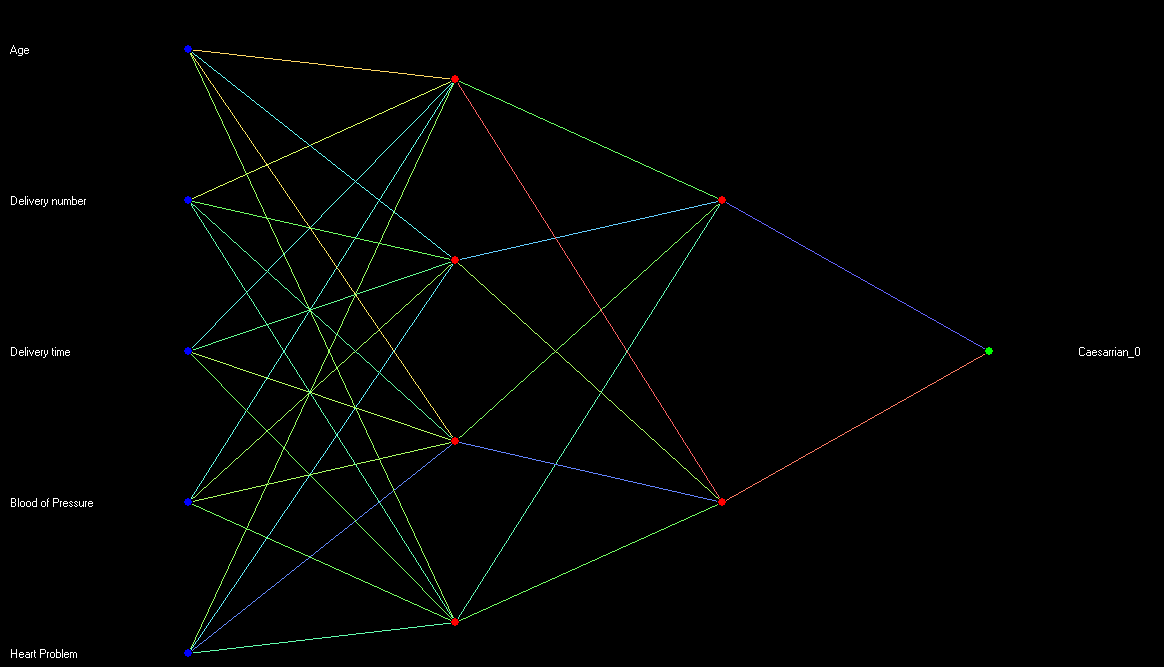


Рисунок 20 – Граф нейросети

На рисунке 21 приведена таблица сопряженности, на которой показано, что было ошибочно распознано 16 примеров.

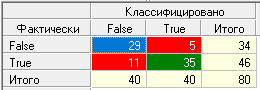


Рисунок 21 – Таблица сопряженности

Затем рассчитаем метрики качества классификации:

**Отсюда следует, что понижение крутизны привело к ухудшению полноты классификатора (была: 0,805 стала: 0,7837).**

В этот раз вернём значение крутизны (1) и изменим скорость обучения (с 0,01 до 0,5) с моментом (с 0,9 до 0,8) (рисунок 22-23).

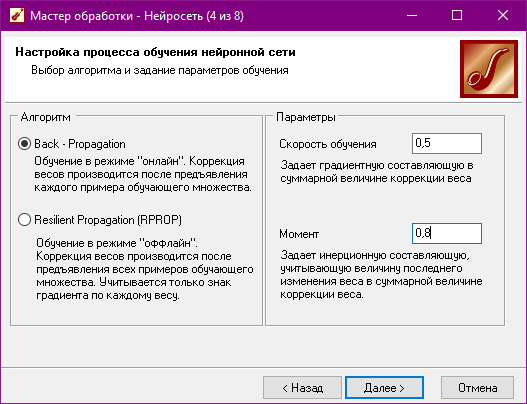


Рисунок 22 – Выбор алгоритма обучения

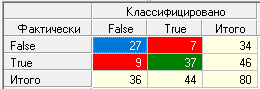


Рисунок 23 – Таблица сопряженности

Затем рассчитаем метрики качества классификации:

**Отсюда следует, что понижение скорости обучения и момента привело к ухудшению полноты классификатора (была: 0,805 стала: 0,7713).**

# Вывод

После выполнения лабораторной работы были изучены методы решения задач классификации. Работы была выполнена в Deductor Academic с использованием двух методов:

1) Дерево решений

2) Нейронная сеть

Оценка метода классификации нейронной сети был 0,805, а у

дерева решений 0,675.