Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра автоматизированных систем управления

“УТВЕРЖДАЮ”
ДИРЕКТОР ИСТР
д.соц.н. Осьмук Л. А.
“___”_____ _____ г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Методы анализа данных

Образовательная программа: 09.03.01 Информатика и вычислительная техника, профиль: Автоматизированные системы обработки информации и управления в социальной сфере

Курс: 4, семестр : 7

Институт социальных технологий и реабилитации,

		Семестр
№	**Вид деятельности**	**7**
1	Всего зачетных единиц (кредитов)	3
2	Всего часов	108
3	Всего занятий в контактной форме, час.	45
4	Лекции, час.	18
5	Практические занятия, час.	0
6	Лабораторные занятия, час.	18
7	из них в активной и интерактивной форме, час.	18
8	Аттестация, час.	2
9	Консультации, час.	7
10	Самостоятельная работа, час.	63
11	Виды самостоятельной работы (курсовой проект, курсовая работа, РГЗ, подготовка к контрольной работе)	РГЗ
12	Вид аттестации	ДЗ

Новосибирск 2017

Рабочая программа составлена на основании федерального государственного образовательного стандарта высшего образования по направлению (специальности): 09.03.01 Информатика и вычислительная техника

ФГОС введен в действие приказом №5 от 12.01.2016 г. , дата утверждения: 09.02.2016 г.

Место дисциплины в структуре учебного плана: Блок 1, вариативная

Рабочая программа разработана на основе компетентностной модели выпускника по направлению (специальности): 09.03.01 Информатика и вычислительная техника

Рабочая программа обсуждена на заседании кафедры

АСУ, протокол заседания кафедры №7 от 20.06.2017

Утверждена на совете института социальных технологий и реабилитации, протокол № 6 от 21.06.2017

Программу разработал:

доцент, к.т.н. Ганелина Н. Д.

Заведующий кафедрой:

профессор, д.т.н. Гриф М. Г.

Ответственный за образовательную программу:

заведующий кафедрой Гриф М. Г.

1. Внешние требования

Таблица 1.1

***Компетенция ФГОС*: ОПК.2 способность осваивать методики использования программных средств для решения практических задач; *в части следующих результатов обучения*:**
з3. знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности
у5. уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности
***Компетенция ФГОС*: ОПК.3 способность разрабатывать бизнес-планы и технические задания на оснащение отделов, лабораторий, офисов компьютерным и сетевым оборудованием; *в части следующих результатов обучения*:**
у1. уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ
***Компетенция ФГОС*: ПК.3 способность обосновывать принимаемые проектные решения, осуществлять постановку и выполнять эксперименты по проверке их корректности и эффективности; *в части следующих результатов обучения*:**
у1. уметь ставить и решать задачи статистической обработки экспериментальных данных
у8. уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности

2. Требования НГТУ к результатам освоения дисциплины

Таблица 2.1

Результаты изучения дисциплины по уровням освоения (иметь представление, знать, уметь, владеть)	Формы организации занятий
Методы анализа данных	
ОПК.2.з3 знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности	
1.знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности	Лекции; Лабораторные работы; Самостоятельная работа
ОПК.2.у5 уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности	
2.уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности	Лекции; Лабораторные работы; Самостоятельная работа
ОПК.3.у1 уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ	
3.уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ	Лабораторные работы; Самостоятельная работа
ПК.3.у1 уметь ставить и решать задачи статистической обработки экспериментальных данных	
4.уметь ставить и решать задачи статистической обработки экспериментальных данных	Лекции; Лабораторные работы; Самостоятельная работа
ПК.3.у8 уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности	
5.уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности	Лекции; Лабораторные работы; Самостоятельная работа

3. Содержание и структура учебной дисциплины

Таблица 3.1

Темы лекций	Активные формы, час.	Часы	Ссылки на результаты обучения	Учебная деятельность
Семестр: 7				
Дидактическая единица: Основы Data Mining				
1. Введение в Data Mining. Определение понятия Data Mining. Основные этапы Data Mining. Основные методы Data Mining, их сравнительный анализ. Проблемы, возникающие в процессе анализа данных и пути их решения. Наиболее перспективные и развиваемых направления Data Mining.	0	1	1, 4	
Дидактическая единица: Технологии компьютерного анализа данных				
6. Обзор компьютерных систем и технологий анализа данных	0	1	1, 4, 5	
Дидактическая единица: Предварительный анализ данных.				
9. Предварительтный анализ данных. Анализ выбросов, пропущенных значений. Визуализация. Описательная статистика	0	1	1, 2, 4, 5	
Дидактическая единица: Дисперсионный анализ				
8. Однофакторный дисперсионный анализ	0	1	1, 2, 4, 5	
Дидактическая единица: Статистические закономерности. Корреляционный анализ. Регрессионный анализ.				
10. Корреляционный анализ. Коэффициент Пирсона. Ранговый корреляционный анализ. Линейная регрессия. Множественная линейная регрессия. Коэффициент детерминации. График остатков.	0	1	1, 2, 4, 5	
Дидактическая единица: Ассоциативные правила				
11. Ассоциативные правила. Численные ассоциативные правила. Обобщенные ассоциативные правила. Алгоритм Apriori.	0	0,5	1, 2, 4, 5	
Дидактическая единица: Классификация. Дискриминатный анализ. Деревья решений				
1. Задача классификации. Дискриминантная функция. Метод опорных векторов. Деревья решений. Оценка качества построенной модели.	0	1,5	1, 2, 4	
Дидактическая единица: Основы кластерного анализа				
4. Методы кластерного анализа. Меры сходства. Иерархические методы кластеризации. Метод k-средних.	0	2	1, 2, 4, 5	Учатся применять на практике методы кластерного анализа и интерпретировать полученные результаты.
Дидактическая единица: Факторный анализ				

12. Факторный анализ. Метод главных компонент. Методы вращения. Метод поиска сгущений.	0	2	1, 2, 4, 5	
Дидактическая единица: Анализ временных рядов				
13. Модели временных рядов. Методология ARIMA	0	3	1, 4, 5	
Дидактическая единица: Основы нейроинфоматики				
4. Введение в нейроинфоматику. История развития нейронных сетей. Схема искусственного нейрона, виды функций активации. Виды архитектур нейронных сетей. Принцип работы однослойного персептрона. Обобщение на случай многослойного персептрона. Алгоритм обратного распространения ошибки. Основные принципы работы и области применения сетей Хопфилда. Основные принципы работы и области применения сетей Кохонена.	0	4	1, 2, 4, 5	

Таблица 3.2

Темы лабораторных работ	Активные формы, час.	Часы	Ссылки на результаты обучения	Учебная деятельность
Семестр: 7				
Дидактическая единица: Предварительный анализ данных.				
7. Предварительный анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР.	4	4	1, 2, 3, 4, 5	Студент проводит предварительный анализ данных, применяет на практике знания, полученные в ходе изучения темы в ходе лекции и самостоятельной работы.
Дидактическая единица: Дисперсионный анализ				
5. Однофакторный дисперсионный анализ	2	2	1, 2, 3, 4, 5	Студент выбирает массив данных для проведения дисперсионного анализа. Проводит предварительный анализ данных. Выдвигает гипотезу, проводит расчеты, анализирует результаты. Оформляет отчет в соответствии с требованиями. Защищает выполненную работу.
Дидактическая единица: Статистические закономерности. Корреляционный анализ. Регрессионный анализ.				

8. Корреляционный анализ. Регрессионный анализ (линейный)	2	2	1, 2, 3, 4, 5	Студент выбирает массив данных для проведения регрессионного анализа. Проводит предварительный анализ данных. Выдвигает гипотезу, проводит расчеты, анализирует результаты. Оформляет отчет в соответствии с требованиями. Защищает выполненную работу.
Дидактическая единица: Ассоциативные правила				
9. Поиск ассоциативных правил	2	2	1, 2, 3, 4, 5	Студент выбирает массив данных для поиска ассоциативных правил. Проводит предварительный анализ данных. Выдвигает гипотезу, проводить расчеты, анализирует результаты. Оформляет отчет в соответствии с требованиями. Защищает выполненную работу. Учится применять на практике знания, полученные на лекциях и в ходе самостоятельной подготовки.
Дидактическая единица: Основы кластерного анализа				
10. Кластеризация	4	4	1, 2, 3, 4, 5	Студент выбирает массив данных для проведения кластеризации. Проводит предварительный анализ данных. Выдвигает гипотезу, проводит расчеты, анализирует результаты. Оформляет отчет в соответствии с требованиями. Защищает выполненную работу. Учится применять на практике знания, полученные на лекциях и в ходе самостоятельной подготовки.
Дидактическая единица: Факторный анализ				
11. Факторный анализ	4	4	1, 2, 3, 4, 5	Студент выбирает массив данных для проведения факторного анализа. Проводит предварительный анализ данных. Выдвигает гипотезу, проводит расчеты, анализирует результаты. Оформляет отчет в соответствии с требованиями. Защищает выполненную работу. Учится применять на практике знания, полученные на лекциях и в ходе самостоятельной подготовки.

4. Самостоятельная работа обучающегося

№	Виды самостоятельной работы	Ссылки на результаты обучения	Часы на выполнение	Часы на консультации
Семестр: 7				
1	РГЗ	1, 2, 3, 4, 5	36	4
Постановка и решение задачи классификации: Авдеенко Т. В. Компьютерные методы анализа данных и прогнозирования [Электронный ресурс] : учебно-методическое пособие / Т. В. Авдеенко ; Новосиб. гос. техн. ун-т. - Новосибирск, [2011]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000162664. - Загл. с экрана. Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана.				
2	Подготовка к занятиям	1, 2, 4, 5	7	0
: Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана.				
3	Подготовка к аттестации	1, 4, 5	20	3
: Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана.				

5. Технология обучения

Для организации и контроля самостоятельной работы обучающихся, а также проведения консультаций применяются информационно-коммуникационные технологии (табл. 5.1).

Таблица 5.1

Деятельность	Информационно-коммуникационные технологии
Информирование	e-mail; Среда электронного обучения НГТУ
Консультирование	e-mail; Среда электронного обучения НГТУ
Контроль	e-mail; Среда электронного обучения НГТУ
Размещение учебных материалов	Среда электронного обучения НГТУ; ЭБС

Таблица 5.2

Активные и интерактивные формы проведения занятий

№	Наименование активных форм	Коды формируемых компетенций
1	Дискуссия	ОПК.2;
Формируемые умения: з3. знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности; у5. уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности		
Краткое описание применения: Студенты изучают массив данных, описывающий некоторую предметную область. Формулируют гипотезу о наличии изучаемого вида зависимостей.		
Подробная информация об использовании технологии приводится в "Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана."		
2	Кейс-стади	ОПК.2; ОПК.3; ПК.3;
Формируемые умения: з3. знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности; у1. уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ; у1. уметь ставить и решать задачи статистической обработки экспериментальных данных; у5. уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности; у8. уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности		
Краткое описание применения: Студенты изучают массив данных, описывающий некоторую предметную область. Формулиуют гипотезы о наличии рассматрвиаемых видов зависимостей, проводят расчеты, интерпретирют полученные результаты.		
Подробная информация об использовании технологии приводится в "Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана."		

6. Правила аттестации обучающихся по учебной дисциплине

Для аттестации обучающихся по дисциплине используется балльно-рейтинговая система (БРС), позволяющая выставлять оценки по традиционной шкале и 15-уровневой ECTS. Краткая информация о БРС приведена в табл. 6.1.

Таблица 6.1

Оцениваемые виды деятельности обучающихся	Мин. балл	Максимальный балл
Семестр: 7		
Самостоятельное изучение теоретического материала:	0	
Лабораторная:	20	40
РГЗ:	20	40
Зачет:	10	20

В таблице 6.2 представлено соответствие форм контроля заявляемым требованиям к результатам освоения дисциплины.

Таблица 6.2

Коды компетенций ФГОС	Результаты обучения	Формы контроля	
		Защита РГЗ	Зачет
ОПК.2	з3. знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности	+	+
	у5. уметь применять методы и специализированные инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности	+	+
ОПК.3	у1. уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ		+
ПК.3	у1. уметь ставить и решать задачи статистической обработки экспериментальных данных	+	+
	у8. уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности	+	+

Фонд оценочных средств по дисциплине представлен в приложении № 1 к рабочей программе.

7. Литература

Основная литература

1. Цильковский И. А. Методы анализа знаний и данных : конспект лекций / И. А. Цильковский, В. М. Волкова ; Новосиб. гос. техн. ун-т. - Новосибирск, 2010. - 66, [2] с. : ил. - Режим доступа:http://elibrary.nstu.ru/source?bib_id=vtls000134804

2. Чубукова И. А. Data Mining : учебное пособие / И. А. Чубукова. - М., 2006. - 382 с. : ил.

3. Авдеенко Т. В. Компьютерные методы анализа временных рядов и прогнозирования : учебное пособие / Т. В. Авдеенко ; Новосиб. гос. техн. ун-т. - Новосибирск, 2008. - 270, [1] с. : ил., табл.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000088320. - Инновационная образовательная программа НГТУ «Высокие технологии».

4. Авдеенко Т. В. Компьютерные методы анализа данных и прогнозирования. Лекция 1 [Электронный ресурс] : конспект лекций / Т. В. Авдеенко ; Новосиб. гос. техн. ун-т. - Новосибирск, [2013]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000180086. - Загл. с экрана.

5. Боровиков В. П. Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере : учебное пособие для вузов по специальности "Прикладная математика" / В. П. Боровиков, Г. И. Ивченко. - М., 2006. - 367, [1] с. : ил.

Дополнительная литература

1. Бериков В. Б. Анализ статистических данных с использованием деревьев решений : учебное пособие [для 4-5 курсов ИДО (направление 0617-Статистика)] / В. Б. Бериков ; Новосиб. гос. техн. ун-т. - Новосибирск, 2002. - 60 с. : ил.

2. Халафян А. А. STATISTICA 6. Статистический анализ данных : [учебное пособие для вузов по специальности "Статистика" и другим экономическим специальностям] / А. А. Халафян. - М., 2007. - 503, [5] с. : ил.

3. Халафян А. А. STATISTICA 6. Статистический анализ данных : учебное пособие для вузов / А. А. Халафян. - М., 2008. - 503, [5] с. : ил.

4. Боровиков В. П. STATISTICA: искусство анализа данных на компьютере / Владимир Боровиков. - СПб., 2003. - 688 с. : ил. + 1 электрон. опт. диск (CD-ROM).

5. Боровиков В. П. STATISTICA: искусство анализа данных на компьютере : Для профессионалов. - СПб., 2001. - 650 с. : ил.. - В прилож.: CD-ROM.

6. Боровиков В. П. STATISTICA: Статистический анализ и обработка данных в среде Windows / В. П. Боровиков, И. П. Боровиков. - М., 1997. - 608 с.

7. Боровиков В. П. Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере : учебное пособие для вузов по специальности "Прикладная математика" / В. П. Боровиков, Г. И. Ивченко. - М., 2000. - 384 с. : ил.

8. Боровиков В. П. Программа STATISTICA для студентов и инженеров. - М., 2001. - 300 с. : ил.

9. Нейронные сети. Statistica neural networks : методология и технологии современного анализа данных / под ред. В. П. Боровикова. – 2-е изд., перераб. и доп. – М. : Горячая линия-Телеком, 2008. – 392с. : ил.

Интернет-ресурсы

1. ЭБС НГТУ : http://elibrary.nstu.ru/

2. StatSoft [Электронный ресурс] : компания : сайт. – Режим доступа: http://www.statsoft.ru/. – Загл. с экрана.

3. ЭБС «Издательство Лань» : https://e.lanbook.com/

4. ЭБС IPRbooks : http://www.iprbookshop.ru/

5. ЭБС "Znanium.com" : http://znanium.com/

6. :

8. Методическое и программное обеспечение

8.1 Методическое обеспечение

1. Обнаружение и анализ закономерностей в эмпирических данных : методическое пособие для 5 курса ФПМИ / Новосиб. гос. техн. ун-т ; [сост.: Г. С. Лбов, В. М. Неделько]. - Новосибирск, 2008. - 19, [1] с. : табл.. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000087336

2. Ганелина Н. Д. Методы анализа данных [Электронный ресурс] : электронный учебно-методический комплекс [для студентов 4 курса кафедры АСУ и ИСР, направления «Информатика и вычислительная техника»] / Н. Д. Ганелина ; Новосиб. гос. техн. ун-т. - Новосибирск, [2014]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000208325. - Загл. с экрана.

3. Авдеенко Т. В. Компьютерные методы анализа данных и прогнозирования [Электронный ресурс] : учебно-методическое пособие / Т. В. Авдеенко ; Новосиб. гос. техн. ун-т. - Новосибирск, [2011]. - Режим доступа: http://elibrary.nstu.ru/source?bib_id=vtls000162664. - Загл. с экрана.

8.2 Специализированное программное обеспечение

1 Statistica

2 Denwer

3 Microsoft Office

4 Deductor Academic

9. Материально-техническое обеспечение

Компьютерный класс

№	Наименование	Назначение
1	Компьютерный класс (Компьютеры объединены в локальную сеть с выходом в Internet)	Компьютеры объединены в локальную сеть с выходом в Internet

Специальное оборудование

№	Наименование	Назначение
1	Мультимедийный проектор	Для проведения занятий с о студентами ИСР
2	Акустическая система	Для проведения занятий с о студентами ИСР

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра автоматизированных систем управления

“УТВЕРЖДАЮ”
ДИРЕКТОР ИСТР
д.соц.н., профессор Л.А. Осьмук
“___”_____ _____ г.

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

УЧЕБНОЙ ДИСЦИПЛИНЫ

Методы анализа данных

Образовательная программа: 09.03.01 Информатика и вычислительная техника, профиль: Автоматизированные системы обработки информации и управления в социальной сфере

Новосибирск 2017

1. **Обобщенная структура фонда оценочных средств учебной дисциплины**

Обобщенная структура фонда оценочных средств по **дисциплине** «Методы анализа данных» приведена в Таблице 1.

Таблица 1.

Формируемые компетенции	**Показатели сформированности компетенций (знания, умения, навыки)**	**Темы**	**Этапы оценки компетенций**	
			Мероприятия текущего контроля (РГЗ)	**Промежуточная аттестация (экзамен, зачет)**
ОПК.2 способность осваивать методики использования программных средств для решения практических задач	з2. знать методы и инструментальные средства анализа и статистической обработки данных о функционировании объектов профессиональной деятельности	Ассоциативные правила. Численные ассоциативные правила. Обобщенные ассоциативные правила. Алгоритм Apriori. Введение в Data Mining. Определение понятия Data Mining. Основные этапы Data Mining. Основные методы Data Mining, их сравнительный анализ. Проблемы, возникающие в процессе анализа данных и пути их решения. Наиболее перспективные и развиваемых направления Data Mining. Введение в нейроинфоматику. История развития нейронных сетей. Схема искусственного нейрона, виды функций активации. Виды архитектур нейронных сетей. Принцип работы однослойного персептрона. Обобщение на случай многослойного персептрона. Алгоритм обратного распространения ошибки. Основные принципы работы и области применения сетей Хопфилда. Основные принципы работы и области применения сетей Кохонена. Задача классификации. Дискриминантная функция. Метод опорных векторов. Деревья решений. Оценка качества построенной модели. Кластеризация Корреляционный анализ. Коэффициент Пирсона. Ранговый корреляционный анализ. Линейная регрессия. Множественная линейная регрессия. Коэффициент детерминации. График остатков. Корреляционный анализ. Регрессионный анализ (линейный) Методы кластерного анализа. Меры сходства. Иерархические методы кластеризации. Метод k-средних. Модели временных рядов. Методология ARIMA Обзор компьютерных систем и технологий анализа данных Однофакторный дисперсионный анализ Поиск ассоциативных правил Предварительный анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР. Предварительтный анализ данных. Анализ выбросов, пропущенных значений. Визуализация. Описательная статистика Факторный анализ Факторный анализ. Метод главных компонент. Методы вращения. Метод поиска сгущений.	РГЗ. Темы и разделы РГР: предварительный анализ данных (корреляционный анализ; регрессионный анализ); задача классификации.	Зачет. Вопросы 1-9, 33-34, 15-22.
ОПК.2	у4. уметь применять методы и специализированны	Задача классификации. Дискриминантная функция. Метод опорных векторов. Деревья решений. Оценка качества построенной модели. Кластеризация	РГЗ. Основной раздел – задача	Зачет. Вопросы 1-35.

	е инструментальные средства анализа и обработки данных, компьютерные технологии анализа данных для исследования объектов профессиональной деятельности	Корреляционный анализ. Регрессионный анализ (линейный) Однофакторный дисперсионный анализ Поиск ассоциативных правил Предварительный анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР. Факторный анализ	классификации. Предварительный анализ – корреляционный, регрессионный анализ, описательная статистика. Постановка задачи и выбор инструментальных средств.	
ОПК.3 способность разрабатывать бизнес-планы и технические задания на оснащение отделов, лабораторий, офисов компьютерным и сетевым оборудованием	у1. уметь оформлять отчеты по научно-исследовательской работе в соответствии с требованиями ГОСТ	Кластеризация Корреляционный анализ. Регрессионный анализ (линейный) Однофакторный дисперсионный анализ Поиск ассоциативных правил Предварительный анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР. Факторный анализ	РГЗ. Оформление отчета.	Зачет. Вопросы 34-35.
ПК.9.В/НИ готовность обосновывать принимаемые проектные решения, осуществлять постановку и выполнять эксперименты по проверке их корректности и эффективности	у1. уметь ставить и решать задачи статистической обработки экспериментальных данных	Ассоциативные правила. Численные ассоциативные правила. Обобщенные ассоциативные правила. Алгоритм Apriori. Введение в Data Mining. Определение понятия Data Mining. Основные этапы Data Mining. Основные методы Data Mining, их сравнительный анализ. Проблемы, возникающие в процессе анализа данных и пути их решения. Наиболее перспективные и развиваемых направления Data Mining. Введение в нейроинфоматику. История развития нейронных сетей. Схема искусственного нейрона, виды функций активации. Виды архитектур нейронных сетей. Принцип работы однослойного персептрона. Обобщение на случай многослойного персептрона. Алгоритм обратного распространения ошибки. Основные принципы работы и области применения сетей Хопфилда. Основные принципы работы и области применения сетей Кохонена. Задача классификации. Дискриминантная функция. Метод опорных векторов. Деревья решений. Оценка качества построенной модели. Кластеризация Корреляционный анализ. Коэффициент Пирсона. Ранговый корреляционный анализ. Линейная регрессия. Множественная линейная регрессия. Коэффициент детерминации. График остатков. Корреляционный анализ. Регрессионный анализ (линейный) Методы кластерного анализа. Меры сходства. Иерархические методы кластеризации. Метод k-средних. Модели временных рядов. Методология ARIMA Обзор компьютерных систем и технологий анализа данных Однофакторный дисперсионный анализ Поиск ассоциативных правил Предварительный	РГЗ. Постановка задачи и выбор инструментальных средств.	Зачет. Вопросы 1-2, 34-36.

		анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР. Предварительтный анализ данных. Анализ выбросов, пропущенных значений. Визуализация. Описательная статистика Факторный анализ Факторный анализ. Метод главных компонент. Методы вращения. Метод поиска сгущений.		
ПК.9.В/НИ	у8. уметь обосновывать выбор математических методов (моделей), компьютерных технологий и средств для решения задач исследования объектов профессиональной деятельности	Ассоциативные правила. Численные ассоциативные правила. Обобщенные ассоциативные правила. Алгоритм Apriori. Введение в нейроинфоматику. История развития нейронных сетей. Схема искусственного нейрона, виды функций активации. Виды архитектур нейронных сетей. Принцип работы однослойного персептрона. Обобщение на случай многослойного персептрона. Алгоритм обратного распространения ошибки. Основные принципы работы и области применения сетей Хопфилда. Основные принципы работы и области применения сетей Кохонена. Кластеризация Корреляционный анализ. Коэффициент Пирсона. Ранговый корреляционный анализ. Линейная регрессия. Множественная линейная регрессия. Коэффициент детерминации. График остатков. Корреляционный анализ. Регрессионный анализ (линейный) Методы кластерного анализа. Меры сходства. Иерархические методы кластеризации. Метод k-средних. Модели временных рядов. Методология ARIMA Обзор компьютерных систем и технологий анализа данных Однофакторный дисперсионный анализ Поиск ассоциативных правил Предварительный анализ данных для проведения дальнейшего анализа в соответствии с заданиями лабораторной работы и РГР. Предварительтный анализ данных. Анализ выбросов, пропущенных значений. Визуализация. Описательная статистика Факторный анализ Факторный анализ. Метод главных компонент. Методы вращения. Метод поиска сгущений.	РГЗ. Постановка задачи и выбор инструментальных средств.	Зачет. Вопросы 1-2, 34-36.

2. Методика оценки этапов формирования компетенций в рамках дисциплины.

Промежуточная аттестация по **дисциплине** проводится в 7 семестре - в форме дифференцированного зачета, который направлен на оценку сформированности компетенций ОПК.2, ОПК.3, ПК.9.В/НИ.

Зачет проводится в письменной форме, по билетам. Варианты билетов составляются из вопросов, приведенных в паспорте зачета, позволяющих оценить показатели сформированности соответствующих компетенций. Билет включает четыре теоретических вопроса из разных разделов или (по выбору студента) два теоретических вопроса и две задачи.

Кроме того, сформированность компетенций проверяется при проведении мероприятий текущего контроля, указанных в таблице раздела 1.

В 7 семестре обязательным этапом текущей аттестации является расчетно-графическое задание

(РГЗ). Требования к выполнению РГЗ, состав и правила оценки сформулированы в паспорте РГЗ.

Общие правила выставления оценки по дисциплине определяются балльно-рейтинговой системой, приведенной в рабочей программе учебной дисциплины. Баллы за РГЗ и лабораторные работы суммируются без коэффициентов (общий максимальный балл - 80).

На основании приведенных далее критериев можно сделать общий вывод о сформированности компетенций ОПК.2, ОПК.3, ПК.9.В/НИ, за которые отвечает дисциплина, на разных уровнях.

Общая характеристика уровней освоения компетенций.

Ниже порогового. Уровень выполнения работ не отвечает большинству основных требований, теоретическое содержание курса освоено частично, пробелы могут носить существенный характер, необходимые практические навыки работы с освоенным материалом сформированы не достаточно, большинство предусмотренных программой обучения учебных заданий не выполнены или выполнены с существенными ошибками.

Пороговый. Уровень выполнения работ отвечает большинству основных требований, теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые виды заданий выполнены с ошибками.

Базовый. Уровень выполнения работ отвечает всем основным требованиям, теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые из выполненных заданий, возможно, содержат ошибки.

Продвинутый. Уровень выполнения работ отвечает всем требованиям, теоретическое содержание курса освоено полностью, без пробелов, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному.

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»
Кафедра автоматизированных систем управления

Паспорт зачета

по дисциплине «Методы анализа данных», 7 семестр

1. Методика оценки

Зачет проводится в письменной форме, по билетам. Билет формируется по следующему правилу: четыре теоретических вопроса или два теоретических вопроса и две задачи (по решению студента) выбираются случайным образом из списка вопросов (представлен ниже) и задач из разных разделов. Преподаватель вправе задать уточняющие вопросы по ответам и дополнительные вопросы, если ответы неоднозначные. Время на выполнение заданий составляет 2 ак. часа: все студенты одной группы выполняют задания одновременно в течение одной пары; уточняющие вопросы преподаватель задает после проверки работ.

Форма билета для зачета

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Факультет АВТФ

Билет № 12
к зачету по дисциплине «Методы анализа данных»

---

1. Вопрос 1. Метод главных компонент.
2. Вопрос 2. Деревья классификации (деревья решений).
3. Вопрос 3. Основные методы поиска ассоциативных правил. Характеристики качества правил ассоциации.
4. Вопрос 4. Корреляционный анализ количественных переменных. Основные характеристики связи.

Утверждаю: зав. кафедрой_____должность, ФИО
(подпись)
(дата)

2. Критерии оценки

- Ответ на билет для зачета считается **неудовлетворительным**, если студент при ответе на вопросы не дает определений основных понятий, не способен показать причинно-следственные связи явлений, при решении задачи допускает принципиальные ошибки, оценка составляет менее 11 баллов.
- Ответ на билет для зачета засчитывается на **пороговом** уровне, если студент при ответе на вопросы дает определение основных понятий, может показать причинно-следственные связи явлений, при решении задачи допускает непринципиальные ошибки, например, вычислительные, при описании алгоритма метода приводит основные шаги, но не в состоянии подробно описать процедуру применения метода, оценка составляет 11 баллов.
- Ответ на билет для зачета билет засчитывается на **базовом** уровне, если студент при ответе на

вопросы формулирует основные понятия, законы, дает характеристику процессов, явлений, проводит анализ причин, условий, может представить качественные характеристики процессов, не допускает ошибок при решении задачи, оценка составляет 12-18 баллов.

- Ответ на билет для зачета билет засчитывается на **продвинутом** уровне, если студент при ответе на вопросы проводит сравнительный анализ подходов, проводит комплексный анализ, выявляет проблемы, предлагает механизмы решения, способен представить количественные характеристики определенных процессов, приводит конкретные примеры из практики, не допускает ошибок и способен обосновать выбор метода решения задачи,оценка составляет 19-20 баллов.

3. Шкала оценки

Зачет считается сданным, если сумма баллов по всем заданиям билета оставляет не менее 11 баллов (из 20 возможных). В общей оценке по дисциплине баллы за зачет учитываются в соответствии с правилами балльно-рейтинговой системы, приведенными в рабочей программе дисциплины. Студент допускается к зачету, если сумма набранных баллов в течение семестра не менее 41 (максимум – 80). Баллы за зачет и баллы, набранные в течение семестра, суммируются без коэффициентов.

4. **Вопросы к** зачету **по дисциплине «**Методы анализа данных**»**

1. Основные понятия АД, типы задач АД, примеры задач. Математический аппарат АД.
2. Основные этапы компьютерной технологии анализа данных, особенности.
3. Графический разведочный анализ данных. Методы визуализации данных на примере прикладного пакета Statistica. Достоинства и недостатки визуальных методов анализа данных.
4. Первичный разведочный анализ данных. Основные числовые характеристики данных.
5. Первичный разведочный анализ данных. Исследование закона распределения данных. Критерии согласия.
6. Корреляционный анализ количественных переменных. Основные характеристики связи. Примеры задач АД.
7. Ранговый корреляционный анализ. Основные характеристики связи. Примеры задач АД.
8. Анализ таблиц сопряженности. Основные характеристики связи, статистические критерии. Примеры задач АД.
9. Регрессионный анализ. Коэффициент детерминации. Многомерная регрессия.
10. Основные методы поиска ассоциативных правил. Характеристики качества правил ассоциации.
11. Дисперсионный анализ. Однофакторный дисперсионный анализ. Двухфакторный дисперсионный анализ.
12. Кластерный анализ. Постановка задачи. Примеры задач АД. Методологические этапы проведения кластерного анализа. Используемые меры расстояния между объектами и выбор меры. Достоинства и недостатки различных мер расстояний.
13. Кластерный анализ. Классификация методов кластерного анализа данных. Иерархические агломеративные методы (метод одиночной связи, метод полной связи, метод Варда), метод К-средних. Достоинства и недостатки методов кластеризации.
14. Кластерный анализ. Оценка адекватности полученных решений: кофенетическая корреляция; тесты значимости признаков на основе дисперсионного анализа; метод повторных выборок; процедура Монте-Карло.
15. Дискриминантный анализ. Постановка задачи интерпретации и классификации. Примеры задач.

16. Дискриминантный анализ. Задача классификации. Понятие классификационной функции. Прогноз принадлежности объекта к классу на основе классификационной функции. Примеры задач АД. Оценка адекватности функции классификации выборочным данным.
17. Деревья классификации (деревья решений). Примеры задач АД. Основные понятия, связанные с деревьями классификации, показатели сложности и точности деревьев решений. Методологические этапы построения дерева решений.
18. Деревья классификации (деревья решений). Выбор критерия точности прогноза. Метод CART построения дерева решений.
19. Деревья классификации (деревья решений). Определение момента прекращения ветвления, правила останова. Процедуры проверки прогнозной способности построенного дерева решений.
20. Деревья классификации (деревья решений). Выбор размера дерева, кросс-проверочное отсечение по минимальной цене сложности.
21. Математическая постановка задачи снижения размерности и выбор наиболее информативных показателей. Основные методы.
22. Метод главных компонент как метод снижения размерности признакового пространства.
23. Методы факторного анализа при решении задачи снижения размерности признакового пространства и выбор наиболее информативных показателей. Условия применения методов.
24. Анализ и прогнозирование временных рядов. Примеры задач АД. Определение временного ряда. Классификация факторов, влияющих на формирование значений временного ряда, структурная модель временного ряда.
25. Анализ и прогнозирование временных рядов. Этапы решения задачи прогнозирования временного ряда на примере.
26. Анализ и прогнозирование временных рядов. Методы определения вида функции тренда.
27. Анализ и прогнозирование временных рядов. Гармонический анализ для определения вида сезонной составляющей временного ряда.
28. Анализ и прогнозирование временных рядов. Анализ случайных остатков. Определение адекватности построенной модели (верификация модели) данным наблюдения.
29. Генетические алгоритмы. Определение, задачи АД, решаемые на основе использования генетических алгоритмов. Основные понятия и термины, принятые в области генетических алгоритмов. Характеристики генетических алгоритмов.
30. Генетические алгоритмы. Этапы работы генетического алгоритма на примере конкретной задачи.
31. Генетические алгоритмы. Показатели эффективности работы генетического алгоритма. Основные отличия генетических алгоритмов от традиционных методов оптимизации, их достоинства и недостатки. Перспективы развития генетически алгоритмов.
32. Эволюционное моделирование. Основные определения. Применение алгоритмов эволюционного моделирования для решения задач структурной идентификации модели.
33. Нейронные сети.
34. Компьютерные технологии решения разных типов задач АД в современном статистическом программном обеспечении на примере реальных пакетов.
35. Технологии Data Mining, основные этапы, особенности.
36. Направления развития методов, технологий и средств решения задач АД.

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»
Кафедра автоматизированных систем управления

Паспорт
расчетно-графического задания

по дисциплине «Методы анализа данных», 7 семестр

1. Методика оценки

В рамках расчетно-графического задания по дисциплине студенты должны решить задачу классификации с помощью нейронных сетей. Допускается изменение постановки задачи при использовании студентом новых массивов данных.

При выполнении расчетно-графического задания студенты должны провести предварительный анализ массива данных, сформулировать постановку задачи, выбрать метод и среду решения задачи, оценить и интерпретировать полученные результаты. При использовании нейронных сетей необходимо также аргументировать выбор параметров сети и алгоритма обучения.

Обязательные структурные части РГЗ.

1. Описание массива данных.
2. Предварительный анализ массива данных. При необходимости – снижение размерности признакового пространства. Корреляционный анализ.
3. Постановка задачи.
4. Выбор и настройка параметров сети.
5. Анализ полученных результатов.

Оцениваемые позиции: все обязательные разделы.

2. Критерии оценки

- Работа считается **не выполненной,** если выполнены не все части РГЗ, отсутствует анализ массива данных, описание процесса постановки и решения задачи, студент не может показать процесс решения задачи, оценка составляет менее 21 балла.
- Работа считается выполненной **на пороговом** уровне, если части РГЗ выполнены формально: анализ массива данных присутствует, но без пояснений, полученные результаты достаточны для понимания, но не объяснены, студент может показать процесс решения задачи, оценка составляет 21 балл.
- Работа считается выполненной **на базовом** уровне, если оформление отчета в целом соответствует требованиям, но некоторые разделы не содержат пояснений или пропущены, студент может уверенно ответить на вопросы о процедуре решения задачи, оценка составляет 22 – 34 балла.
- Работа считается выполненной **на продвинутом** уровне, если анализ массива данных проведен полностью, предложенные варианты решения задачи, полученные результаты корректны и интерпретированы, студент уверенно отвечает на все вопросы по постановке задачи и процессу ее решения, оценка составляет 35-40 баллов.

3. Шкала оценки

В общей оценке по дисциплине баллы за РГЗ учитываются в соответствии с правилами балльно-рейтинговой системы, приведенными в рабочей программе дисциплины. Баллы за РГЗ суммируются с баллами за лабораторные работы, выполненные в течение семестра, без коэффициентов. Выполнение и защита РГЗ являются основным требованием для допуска студента

к зачету.

4. **Типовое РГЗ**

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
КАФЕДРА АВТОМАТИЗИРОВАННЫХ СИСТЕМ УПРАВЛЕНИЯ

Отчет по расчетно-графической работе
по дисциплине «Методы анализа данных»
Задача классификации с помощью нейронных сетей в среде STATISTICA,
DEDUCTOR и SPSS

Выполнили студенты 4 курса АВТФ
Группы АВТ-312:
Прокофьев А.С
Григорьев И.Е.

Проверил к.т.н., доцент кафедры АСУ:
Ганелина Н. Д.

Новосибирск, 2016

Оглавление

1 Цель работы

Познакомиться с теорией и практикой применения анализа с использованием нейронной сети.

2 Постановка задачи

Для выбранного массива данных решить задачу классификации с помощью нейронной сети в пакете *Statistica* и в пакете *Deductor*. Исследовать влияние параметров сети на качество решения в каждом из пакетов. Сравнить полученные результаты.

3 Ход работы

3.1 Описание исходных данных и их предварительный анализ

В качестве данных для корреляционного, регрессионного и дисперсионного анализа были выбраны данные «**Abalone**» (Морское ушко).

Описание: набор данных «Abalone» содержит информацию об исследовании морского ушка, проводимых отделом морских ресурсов Морской Научно-исследовательской лаборатории – «Taroona» в 1995 году.

Количество экземпляров: 1000

Количество атрибутов: 8

Аттрибут	**Описание**
Длина	Длина тела ушка (мм)
Диаметр	Диаметр тела ушка, перпендикулярно длине(мм)
Высота	Высота тела с мясом в оболочке (мм)
Вес ушка	Весь вес ушка (гр)
Вес мяса	Вес только мяса (гр)
Вес кишечника	Вес внутренностей ушка (гр)
Вес панцеря	Вес только панцеря после сушки (гр)
Возраст	Возраст, определенный с помощью анализа колец на теле ушка (года)

Длина	Диаметр	Высота	Вес_ушка	Вес_мяса	Вес_кишечника	Вес_панцеря	Возраст
0,455	0,365	0,095	0,51	0,2245	0,1010	0,150	15
0,350	0,265	0,090	0,23	0,0995	0,0485	0,070	7
0,530	0,420	0,135	0,68	0,2565	0,1415	0,210	9
0,440	0,365	0,125	0,52	0,2155	0,1140	0,155	10
0,330	0,255	0,080	0,21	0,0895	0,0395	0,055	7
0,425	0,300	0,095	0,35	0,1410	0,0775	0,120	8
0,530	0,415	0,150	0,78	0,2370	0,1415	0,330	20
0,545	0,425	0,125	0,77	0,2940	0,1495	0,260	16
0,475	0,370	0,125	0,51	0,2165	0,1125	0,165	9
0,550	0,440	0,150	0,89	0,3145	0,1510	0,320	19
0,525	0,380	0,140	0,61	0,1940	0,1475	0,210	14
0,430	0,350	0,110	0,41	0,1675	0,0810	0,135	10
0,490	0,380	0,135	0,54	0,2175	0,0950	0,190	11
0,535	0,405	0,145	0,68	0,2725	0,1710	0,205	10
0,470	0,355	0,100	0,48	0,1675	0,0805	0,185	10
0,500	0,400	0,130	0,66	0,2580	0,1330	0,240	12
0,355	0,280	0,085	0,29	0,0950	0,0395	0,115	7
0,440	0,340	0,100	0,45	0,1880	0,0870	0,130	10
0,365	0,295	0,080	0,26	0,0970	0,0430	0,100	7
0,450	0,320	0,100	0,38	0,1705	0,0750	0,115	9
0,355	0,280	0,095	0,25	0,0955	0,0620	0,075	11
0,380	0,275	0,100	0,23	0,0800	0,0490	0,085	10
0,565	0,440	0,155	0,94	0,4275	0,2140	0,270	12
0,550	0,415	0,135	0,76	0,3180	0,2100	0,200	9

Рисунок 1 – Фрагмент таблицы с исходными данными

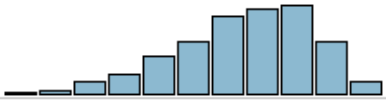
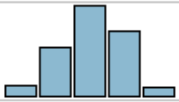
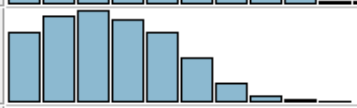
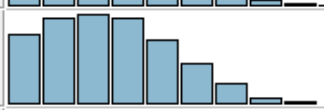
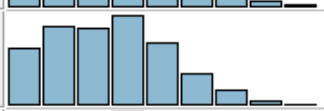
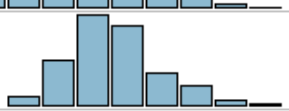
	Метка столбца	Статистика: Кол-во значений = 4177	
		Гистограмма	Кол-во пустых значений
1	9.0 Длина		0
2	9.0 Диаметр		0
3	9.0 Высота		0
4	9.0 Вес ушка		0
5	9.0 Вес мяса		0
6	9.0 Вес кишечника		0
7	9.0 Вес панцеря		0
8	9.0 Возраст		0

Рисунок 2 – Гистограммы распределения величин

На гистограммах видно, что больших выбросов не обнаружено, а выпадающие из общей картины значения имеют случайный характер и их часть незначительна.

3.2 Описание применяемого алгоритма классификации при помощи

нейронных сетей

В задаче классификации сеть должна отнести каждое наблюдение к одному из нескольких классов. Для классификации используется номинальная выходная переменная – различные её значения соответствуют различным классам. Решение такой задачи – одна из наиболее важных областей применения нейронных сетей. В таких задачах входные данные представляют собой результаты измерений некоторых характеристик объекта. Цель состоит в том, чтобы определить, к какому из нескольких заданных классов принадлежит этот объект.

Классификацию можно осуществлять с помощью сетей следующих типов:

- многослойного персептрона;
- радиальной базисной функции;
- сети Кохонена;
- вероятностной нейронной сети и линейной сети.

Чтобы нейронная сеть могла классифицировать объекты, её нужно обучить. Сеть обучается, чтобы для некоторого множества входов давать желаемое множество выходов. Каждое такое входное (или выходное) множество рассматривается как вектор. Обучение осуществляется последовательным предъявлением входных векторов с одновременной подстройкой весов в соответствии с определённой процедурой. В процессе обучения веса сети становятся такими, чтобы каждый входной вектор вырабатывал правильный выходной вектор.

Различают алгоритмы обучения с учителем и без учителя. Обучение с учителем предполагает, что для каждого входного вектора существует целевой вектор, представляющий собой требуемый выход. Вместе они называются обучающей парой. Обычно сеть обучается на некотором числе таких обучающих пар. Предъявляется входной вектор, вычисляется выход сети и сравнивается с соответствующим целевым вектором, разность (ошибка) с помощью обратной связи подаётся в сеть, и веса изменяются в соответствии с алгоритмом, стремящимся минимизировать ошибку. Векторы обучающего множества предъявляются последовательно, вычисляются ошибки и веса подстраиваются для каждого вектора до тех пор, пока ошибка по всему обучающему массиву не достигнет приемлемо низкого уровня.

3.3 Решение задачи в среде *Statistica*

3.3.1 Настройка и обучение сети

Запустим автоматизированные нейронные сети. Для этого выберем задачу классификации и установим автоматический режим (*ANS*) подбора сетей (рис.3).

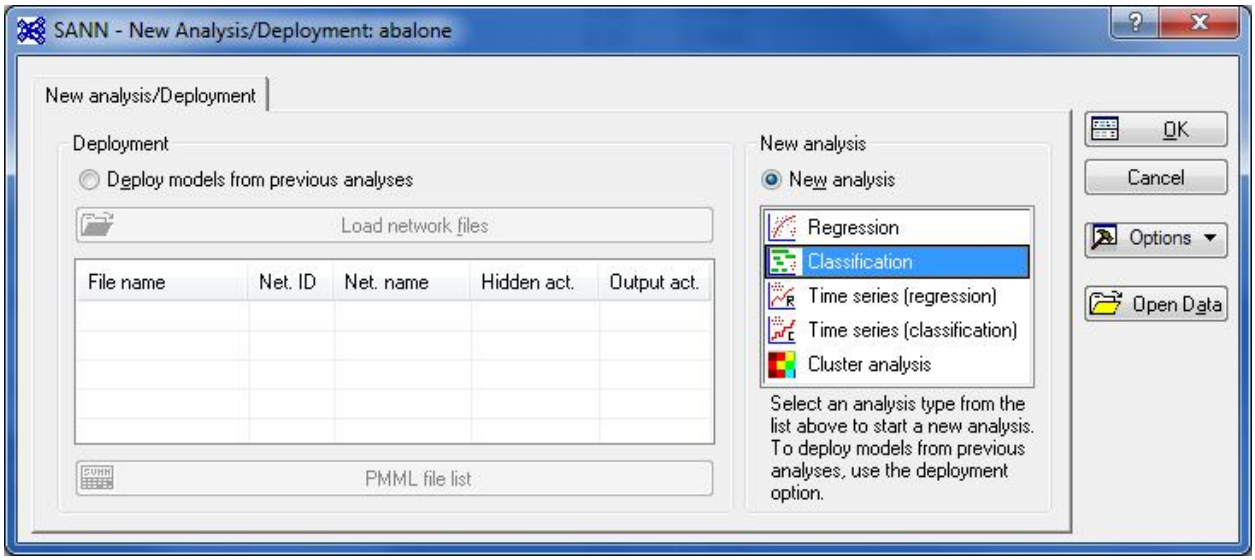


Рисунок 3 – Выбор задачи классификации

Далее необходимо выбрать переменные для анализа (*Variables*). Категориальная целевая переменная – это *Vozrast (Возраст)*. Непрерывные входные характеристики – это параметры с 1 по 7.

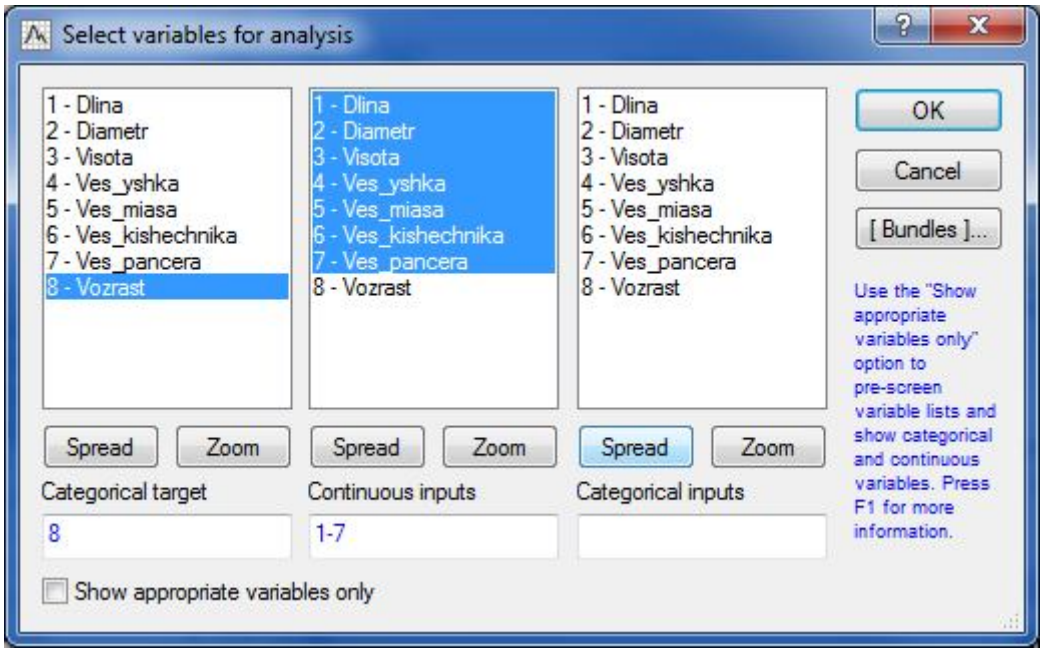


Рисунок 4 – Выбор параметров для анализа.

Теперь необходимо разбить данные на три под-выборки.

1. Обучающая – обучение сети, подбор весовых коэффициентов.
2. Тестовая – если тестовая ошибка перестала убывать или даже стала расти, это указывает на то, что сеть начала слишком близко аппроксимировать данные и обучение следует остановить.
3. Контрольная – итоговая модель тестируется на данных из этого контрольного множества, чтобы убедиться, что результаты, достигнутые на обучающем и тестовом множествах реальны, а не являются артефактами процесса обучения.

Разделение данных оставим стандартным: 70% 15% 15% (рис. 3).

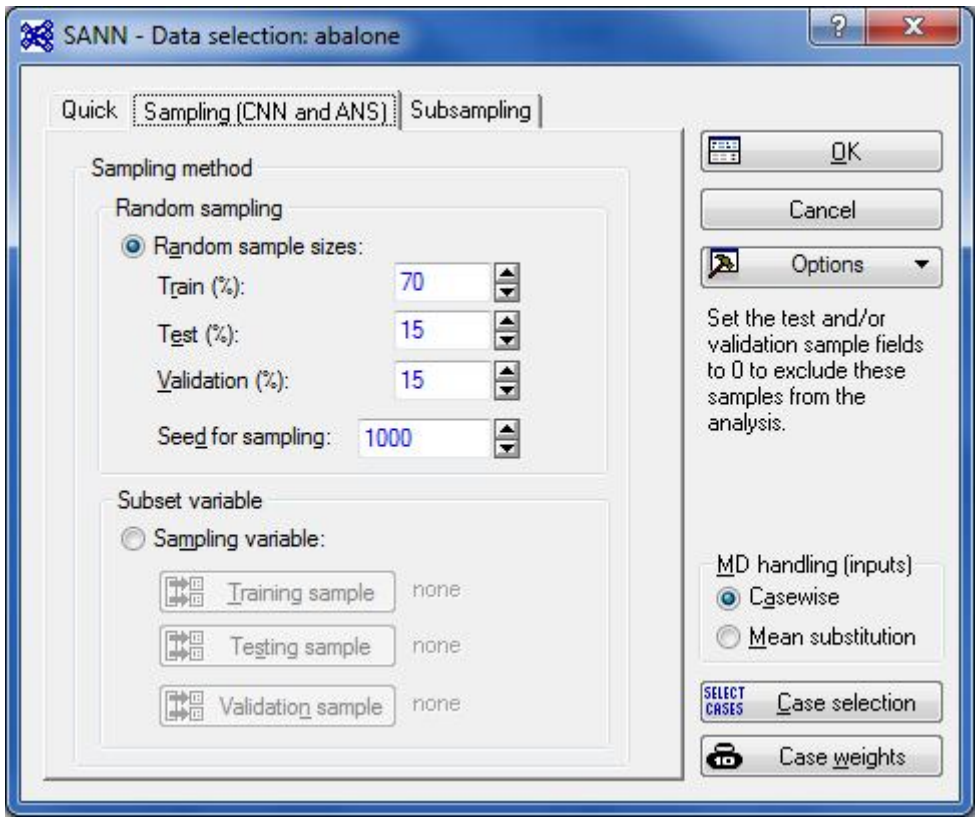


Рисунок 5 – Разделение данных на три подвыборки.

Далее необходимо задать тип нейронной сети, минимальное и максимальное количество скрытых нейронов. Так же нужно задать число генерируемых сетей и количество лучших сетей, которые нужно отобрать из числа генерируемых (рис.6 и рис.7).

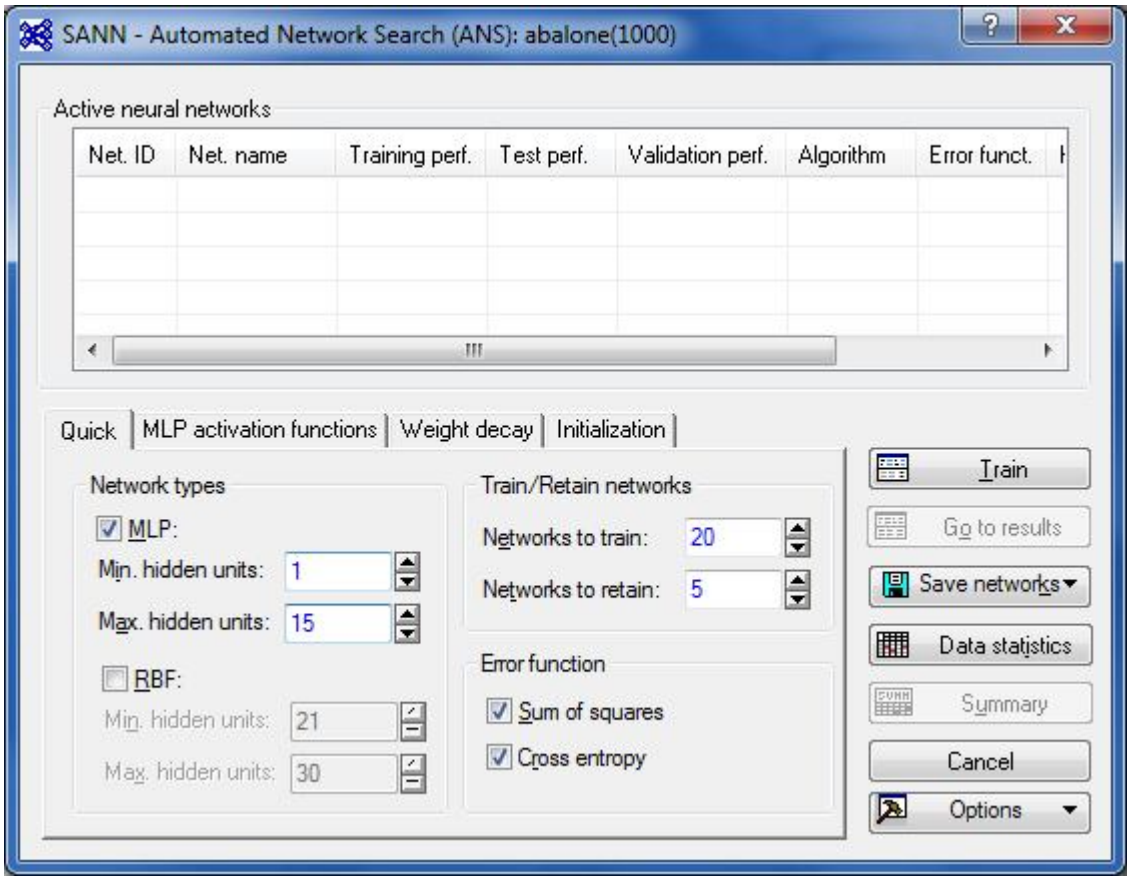


Рисунок 4 – Параметры для генерации сетей

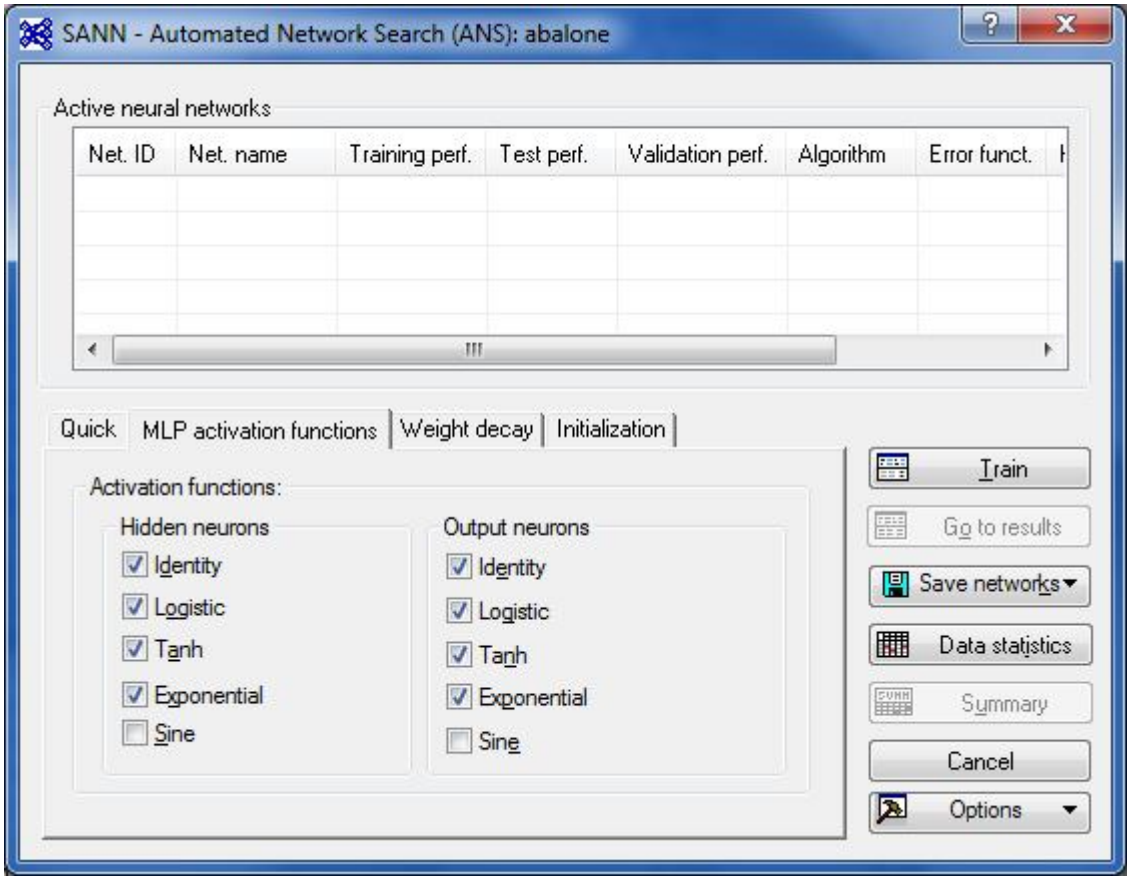


Рисунок 5 – Выбор функций активации для внешних и скрытых слоев

Теперь, когда все параметры сети установлены, можно приступать к генерации и последующему оцениванию сети. (рис.6)

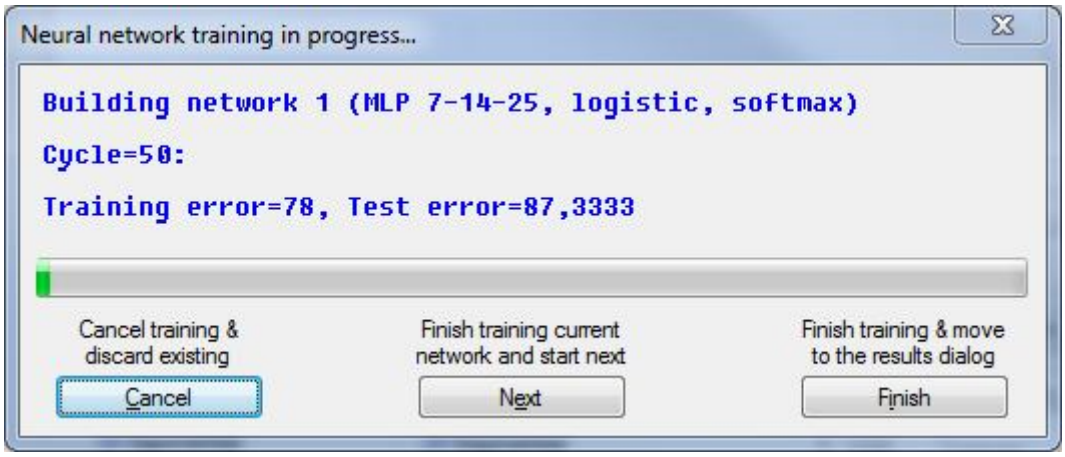


Рисунок 6 – Построение нейронных сетей

Index	Net. name	Training perf.	Test perf.	Validation perf.	Training algorithm	Error function	Hidden activation	Output activation
1	MLP 7-4-25	19,25000	27,00000	27,00000	BFGS 24	Entropy	Identity	Softmax
2	MLP 7-10-25	17,62500	24,00000	27,00000	BFGS 14	Entropy	Tanh	Softmax
3	MLP 7-4-25	18,50000	24,00000	26,00000	BFGS 18	Entropy	Tanh	Softmax
4	MLP 7-5-25	16,87500	23,00000	27,00000	BFGS 20	Entropy	Exponential	Softmax
5	MLP 7-4-25	21,25000	25,00000	28,00000	BFGS 167	SOS	Logistic	Tanh

Рисунок 7 – Результаты генерации и обучения сетей

Видно, что все представленные сети имеют плохие показатели качества по всем трем подгруппам данных. Лучшая сеть (5) имеет показатели эффективности равный всего 21,25%.

Сгенерируем набор сетей еще раз, но в этом случае ограничим параметры теми, которые присутствуют у лучшей сети:

- *Error function = SOS;*
- *Hidden activation = Logistic;*
- *Output activation = Tanh.*

Index	Net. name	Training perf.	Test perf.	Validation perf.	Training algorithm	Error function	Hidden activation	Output activation
1	MLP 7-4-25	21,85714	20,66667	23,33333	BFGS 42	Entropy	Logistic	Softmax
2	MLP 7-12-25	20,57143	21,33333	25,33333	BFGS 51	SOS	Logistic	Tanh
3	MLP 7-4-25	22,42857	22,00000	26,00000	BFGS 81	SOS	Logistic	Tanh
4	MLP 7-13-25	23,71429	22,00000	24,00000	BFGS 79	SOS	Logistic	Tanh
5	MLP 7-4-25	18,71429	20,66667	23,33333	BFGS 32	Entropy	Logistic	Softmax

Рисунок 8 – Результаты повторной генерации

Для дальнейшего анализа из серии проделанных генераций зафиксированы

характеристики лучшей сети (4):

1. *Training perf.* – производительность обучения = 23,71%;
2. *Test. perf* – тестовая производительность = 22,00%;
3. *Validation perf.* – контрольная производительность = 24,00%;
4. *Training algorithm* – обучающий алгоритм *BFGS*;
5. *Error function* – функция ошибки *SOS*;
6. *Hidden activation* – функция скрытых нейронов *Logistic*;
7. *Output activation* – функция выходных нейронов *Tanh*.

3.3.2 Анализ полученных результатов

Рассмотрим более подробно выделенную ранее сеть.

В данной матрице (рис.9) указано число правильно и неправильно классифицированных наблюдений по каждому классу. По таблице видно, что в обоих случаях система ошибается достаточно часто при отнесении объекта к классам.

Сеть *MPL* 7-13-25 содержит 13 скрытых нейронов. Даже этого количества не хватает, чтобы обеспечить высокую степень обучаемости и предотвратить возможность переобучения.

Vozrast (Classification summary) (abalone(1000))
Samples: Train

		Vozrast-1	Vozrast-10	Vozrast-11	Vozrast-12	Vozrast-13	Vozrast-14
4.MLP 7-13-25	Total	1,0000	75,00000	72,00000	57,00000	51,00000	38,00000
	Correct	0,0000	17,00000	25,00000	20,00000	13,00000	1,00000
	Incorrect	1,0000	58,00000	47,00000	37,00000	38,00000	37,00000
	Correct (%)	0,0000	22,66667	34,72222	35,08772	25,49020	2,63158
	Incorrect (%)	100,0000	77,33333	65,27778	64,91228	74,50980	97,36842

Рисунок 9 – Матрица ошибок полученной сети

Исходя из матрицы чувствительности (рис.10), можно заметить, что наиболее значащим параметром является *Вес панцеря* Однако нетрудно заметить, что и остальные параметры являются практически равнозначащими, так как их значения колеблются в одном интервале.

Networks	**Ves_pancera**	Ves_miasa	Dlina	Diametr	Ves_kishechnika	Ves_yshka	Visota
4.MLP 7-13-25	1,088692	1,071477	1,044169	1,038438	1,018091	1,008829	1,006066

Рисунок 10 – Матрица чувствительности

Для того чтобы посмотреть, какие конкретно элементы были ошибочно классифицированы, воспользуемся матрицей доверительных уровней (рис.11).

Case name	**Vozrast Target**	Vozrast - Output 2. MLP 7-9-25
21	11	7
23	12	11
24	9	10
30	11	8
32	15	17
33	18	11
35	13	17
36	8	7
37	16	13
38	8	9
39	11	11
40	9	9
41	9	7
42	14	11
43	5	5
44	5	5
46	7	7
47	9	9
48	7	10
49	6	7
50	9	11
53	10	9
54	10	9
55	7	9
56	8	11
58	8	10

Рисунок 11 – Фрагмент матрицы доверительных уровней

Это можно объяснить тем, что возраст морского ушка может достигать 25 лет определить с вероятностью 100% точный возраст не возможно. Но Statistica учитывает только точные результаты, поэтому такой высокий процент ошибки.

3.4 Решение задачи в среде *Deductor*

Запустим мастер обработки и выберем пункт меню *Нейросеть*. В открывшемся окне зададим назначения параметров (рис.12). Атрибут «Возраст» являются выходным параметром, все остальные входные. Все входные параметры являются вещественными и дискретными.

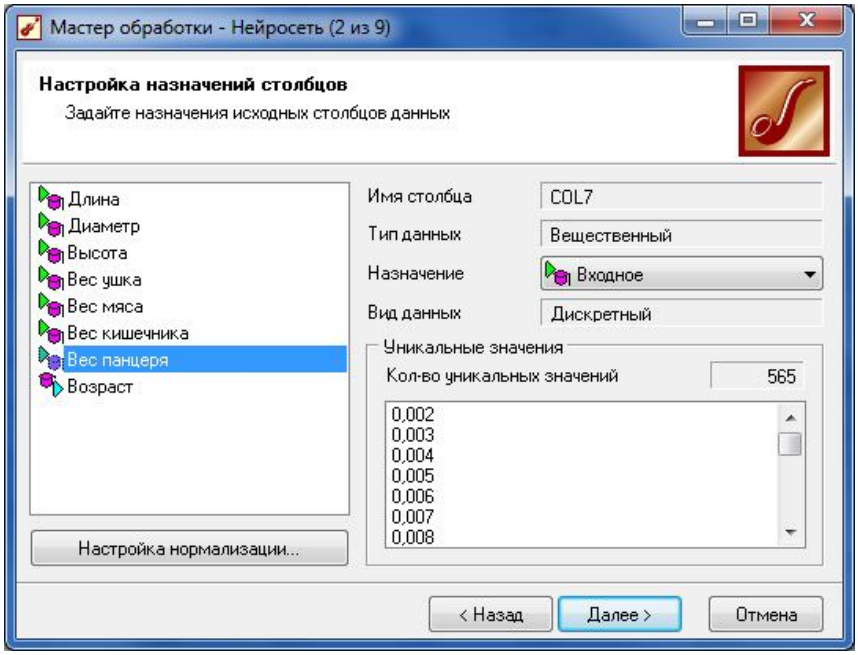


Рисунок 12– Настройка параметров столбцов

Сформируем нейронную сеть со следующими параметрами (рис. 13):

- Скрытых слоев – 6;
- Активационная функция – сигмоида;
- Алгоритм обучения – *Resilient Propagation* (шаг спуска = 0,5; шаг подъема = 1,2).

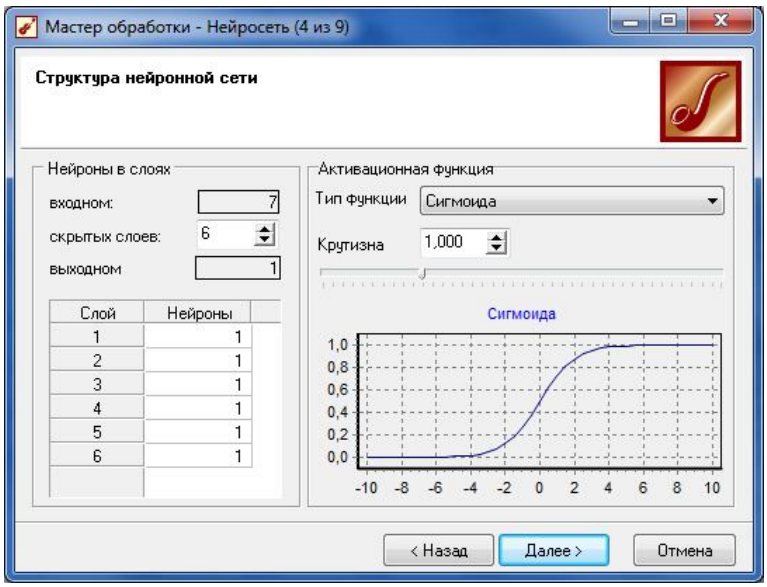


Рисунок 13 – Структура нейронной сети

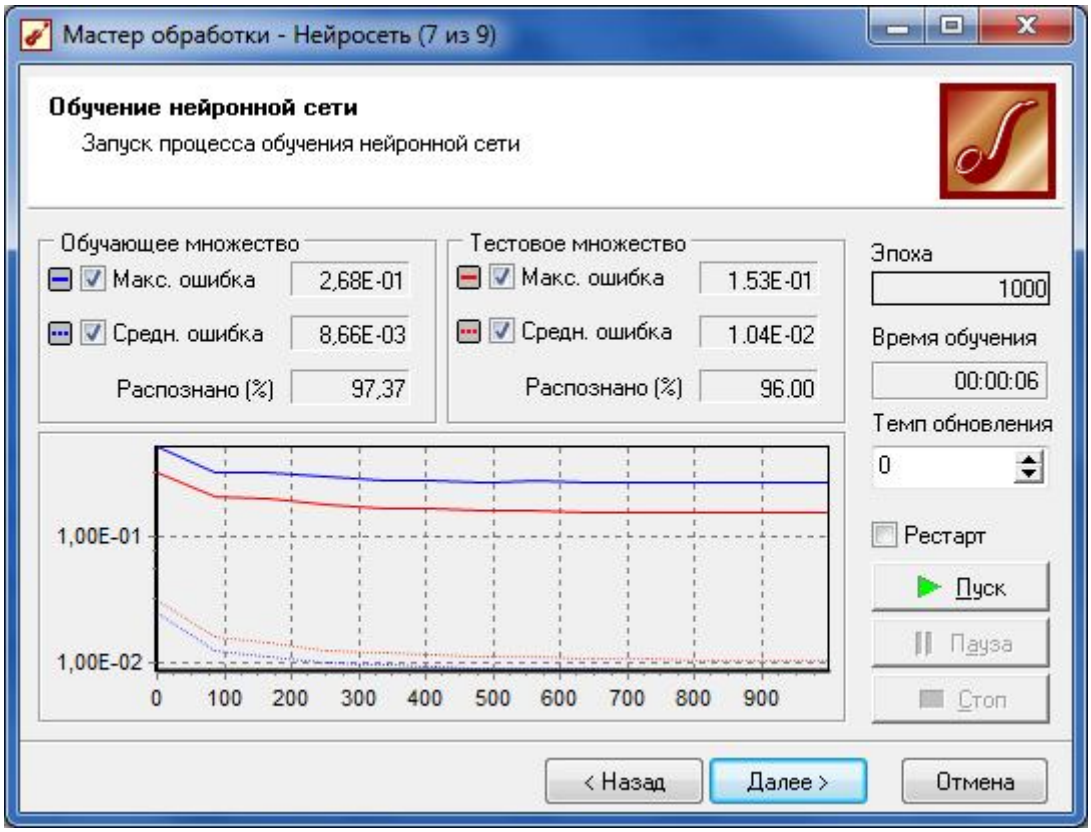


Рисунок 14 – Обучение сети в среде Deductor

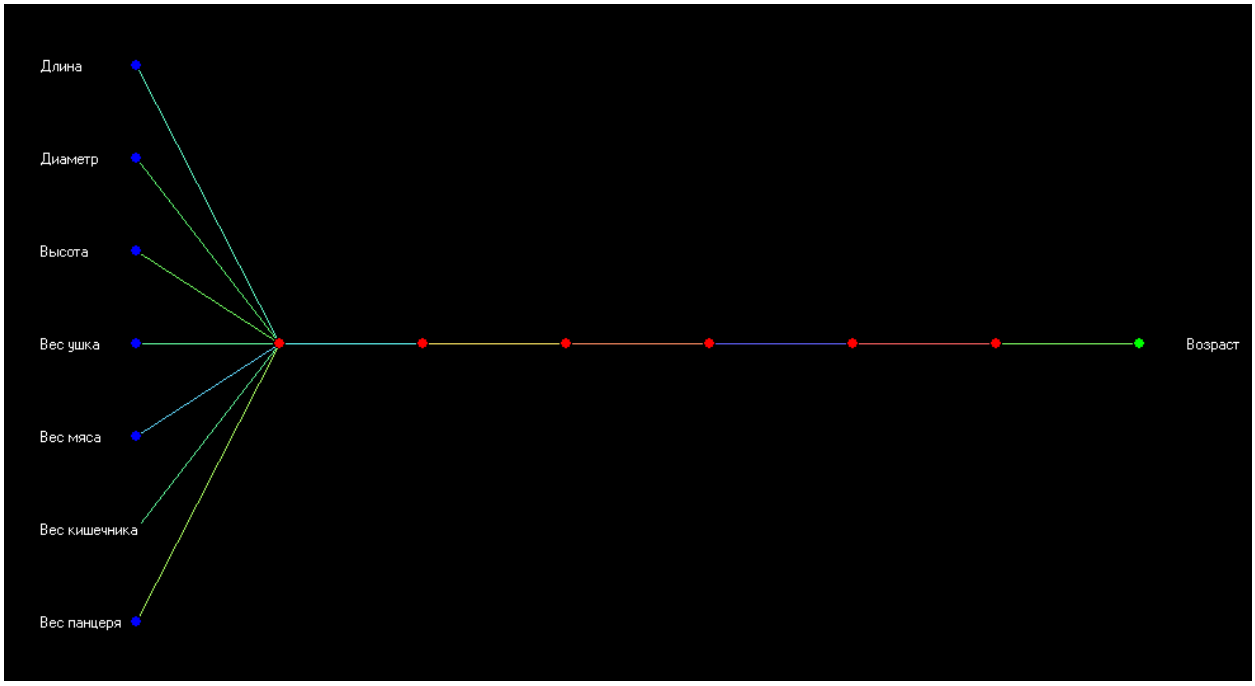


Рисунок 15 – Граф нейросети

Длина	Диаметр	Высота	Вес ушка	Вес мяса	Вес кишечника	Возраст	Возраст_OUT	Возраст_ERR
0,615	0,48	0,18	1,16	0,4845	0,2165	13	12,1566977563028	0,000907090145694858
0,61	0,485	0,17	1,02	0,419	0,2405	12	13,2686211165283	0,00205280553227229
0,58	0,45	0,15	0,93	0,276	0,1815	14	13,7786688198102	6,2484045056397E-5
0,355	0,275	0,085	0,22	0,092	0,06	8	9,12314047132007	0,00160898535499627
0,51	0,4	0,14	0,81	0,459	0,1965	10	7,58016948493683	0,00746885168575369
0,5	0,405	0,155	0,77	0,346	0,1535	12	11,1118972563432	0,00100602867766664
0,505	0,41	0,15	0,64	0,285	0,145	11	10,9421759708159	4,26481932536529E-6
0,64	0,5	0,185	0,3	0,4445	0,2635	16	14,2109746903685	0,00408241270217117
0,56	0,45	0,16	0,92	0,432	0,178	15	10,6279710082633	0,0243809151844215
0,585	0,46	0,185	0,92	0,3635	0,213	10	12,7349199894775	0,00954054508781073
0,45	0,345	0,12	0,42	0,1655	0,095	9	9,09313810787384	1,10646774723456E-5
0,5	0,4	0,165	0,83	0,254	0,205	13	13,2156330970275	5,93082047623383E-5
0,5	0,4	0,145	0,63	0,234	0,1465	12	11,89271145748	1,46821828521246E-5
0,53	0,435	0,17	0,82	0,2985	0,155	13	12,875622873316	1,97317214823555E-5
0,42	0,335	0,115	0,37	0,171	0,071	8	8,50408344598003	0,00032410729657027 8
0,44	0,34	0,14	0,48	0,186	0,1085	9	10,1652493068412	0,00173189534068087
0,4	0,3	0,11	0,32	0,109	0,067	9	8,89544968872187	1,39423055973902E-5
0,435	0,34	0,11	0,38	0,1495	0,085	8	8,64877662920797	0,00053687642169190 9
0,525	0,415	0,17	0,83	0,2755	0,1685	13	13,4350867652469	0,00024145471083294 9
0,37	0,28	0,095	0,27	0,122	0,052	7	7,30190940979867	0,00011626185169002 8
0,49	0,365	0,145	0,63	0,1995	0,1625	10	11,7491266228084	0,00390235196762401
0,335	0,25	0,09	0,18	0,0755	0,0415	7	7,17288317242149	3,81232031970894E-5
0,415	0,325	0,105	0,38	0,1595	0,0785	12	8,28445551772976	0,0176087637751643
0,5	0,405	0,14	0,62	0,241	0,1355	9	11,111549851924	0,00568704435862258
0,485	0,395	0,16	0,66	0,2475	0,128	14	12,2352777669561	0,00397225071402974
0,55	0,405	0,14	0,8	0,244	0,1635	10	11,9640552100391	0,00492029702561462
0,45	0,35	0,13	0,46	0,174	0,111	8	9,40775955851051	0,00252778950838994
0,405	0,3	0,12	0,32	0,1265	0,07	7	8,63613727023104	0,00341447087632537
0,47	0,36	0,135	0,5	0,1665	0,115	10	10,5220573346313	0,00034763247530915 9
0,415	0,305	0,13	0,32	0,1305	0,0755	8	8,75304199663938	0,00072330643967171 5
0,445	0,325	0,125	0,46	0,1785	0,1125	9	8,95314286179412	2,80049923577158E-6
0,47	0,35	0,145	0,52	0,187	0,1235	11	10,7823791733531	6,04066635082844E-5
0,49	0,375	0,15	0,58	0,22	0,144	9	10,9687565469992	0,00494388053743924

Рисунок 16 – Фрагмент таблицы обучающего набора

Результат, полученный в данной среде, незначительно отличается от данных, полученных в среде *Statistica*, и это при меньшем количестве нейронов на внутреннем слое.

Результат можно назвать успешным. Точность классификации 97,37% на обучающем и 96,00% на тестовом подмножестве можно назвать эффективным, не смотря на меньшую размерность внутреннего слоя.

По данным «*таблицы обучающего набора*» (рис. 15) видно, что прогнозируемый возраст (столбец «Возраст_OUT) не всегда совпадает с исходными данными. Но при этом ошибка незначительна (столбец «Возраст_ERR). Это можно объяснить тем, что возраст морского ушка может достигать 25 лет определить с вероятностью 100% точный возраст не возможно. Ошибка в 1-2 года не является существенной. Deductor не воспринимает эту разницу, как ошибка, поэтому такой высокий процент обучаемости сети (в отличии от Statictica).

Теперь изменим количество элементов внутреннего слоя на 6 (рис. 16) и сравним полученные результаты с предыдущими.

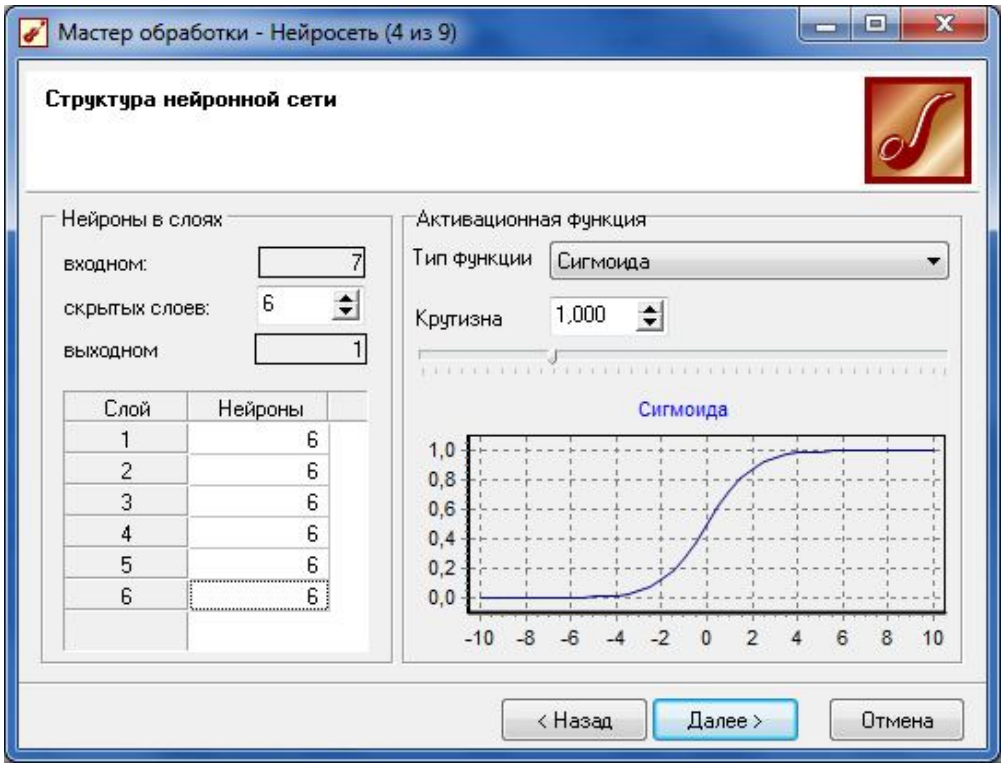


Рисунок 17 – Структура нейронной сети

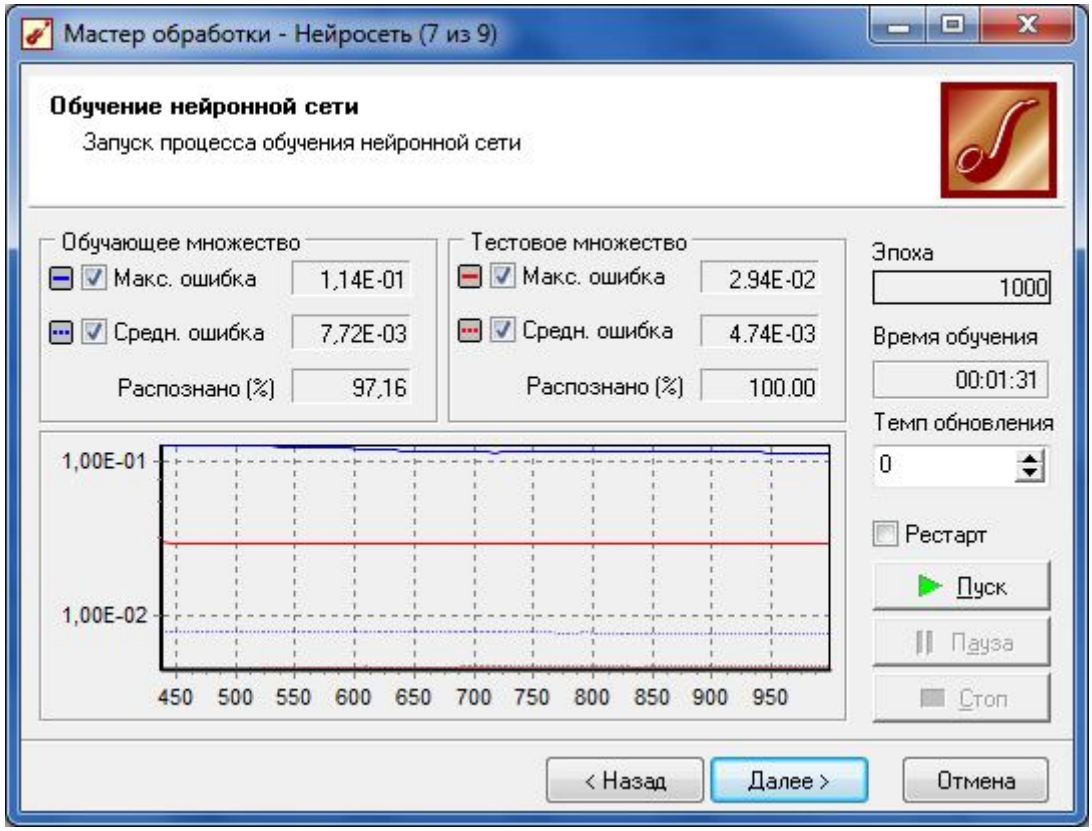


Рисунок 18 – Обучение сети в среде Deductor (для 6 нейронов внутреннего слоя)

Длина	Диаметр	Высота	Вес ушка	Вес мяса	Вес кишечника	Вес панцеря	Возраст	Возраст_OUT	Возраст_ERR
0,425	0,335	0,095	0,32	0,1205	0,061	0,125	10	9,26689726328219	0,000685509722682585
0,38	0,305	0,095	0,28	0,1255	0,0525	0,09	8	7,68417075579994	0,000127229734045897
0,53	0,415	0,145	0,94	0,3845	0,185	0,265	21	11,0210094564997	0,127015627891925
0,34	0,265	0,085	0,18	0,077	0,046	0,065	10	7,42881523795487	0,00843238658236359
0,475	0,365	0,115	0,49	0,223	0,1235	0,134	9	8,12728554633181	0,000971467496991543
0,43	0,34	0,12	0,39	0,1555	0,095	0,141	7	10,0154509295578	0,0115981432507286
0,46	0,365	0,125	0,47	0,1895	0,0945	0,158	10	9,94674366571721	3,61764941484718E-6
0,47	0,36	0,13	0,52	0,198	0,1065	0,165	9	10,4250972164977	0,00259043632202759
0,36	0,295	0,1	0,21	0,066	0,0525	0,075	9	8,3680648961439	0,000509364764650018
0,355	0,265	0,09	0,17	0,05	0,041	0,063	8	7,86411055908293	2,35534950927979E-5
0,38	0,235	0,1	0,26	0,1055	0,054	0,08	7	7,40179751323066	0,000205919951069311
0,355	0,26	0,085	0,19	0,081	0,0485	0,055	6	6,86212883926672	0,00094804354017269
0,44	0,345	0,12	0,49	0,1965	0,108	0,16	14	10,2588723493166	0,0178520868605967
0,51	0,4	0,13	0,57	0,219	0,1365	0,195	13	10,9019299458686	0,00561466575515655
0,325	0,24	0,085	0,17	0,0795	0,038	0,05	7	6,5835643896519	0,000221197216283153
0,62	0,485	0,18	1,18	0,4675	0,2655	0,39	13	13,3634634923766	0,000168502181493143
0,59	0,45	0,16	0,9	0,358	0,156	0,315	19	12,9019954959091	0,0473187493712428
0,33	0,255	0,095	0,19	0,0735	0,045	0,06	7	7,49735922015268	0,000315518104427139
0,45	0,34	0,13	0,37	0,1605	0,0795	0,105	9	8,23047759525078	0,000755312157412013
0,445	0,33	0,12	0,35	0,12	0,084	0,105	11	9,09848646440883	0,00461193077300565
0,33	0,215	0,075	0,11	0,045	0,0265	0,035	6	5,94287392378493	4,16248543843098E-6
0,48	0,375	0,145	0,78	0,216	0,13	0,17	9	11,9007212455654	0,0107323772250947
0,46	0,35	0,12	0,49	0,193	0,105	0,155	11	9,90384657604107	0,00153259225619501
0,475	0,36	0,125	0,45	0,1695	0,081	0,14	9	9,44804296951503	0,000256049110372258
0,255	0,18	0,065	0,08	0,034	0,014	0,025	5	4,97644220890608	7,07869287276739E-7
0,335	0,245	0,09	0,17	0,0595	0,04	0,06	6	7,48167232750571	0,00280019500777574
0,47	0,35	0,13	0,47	0,1845	0,099	0,145	11	9,66903794643556	0,00225951529085263
0,31	0,225	0,08	0,13	0,054	0,024	0,05	7	6,72681467758784	9,51916076293829E-5
0,37	0,28	0,11	0,23	0,0945	0,0465	0,075	10	7,90381252771939	0,00560459428437014
0,295	0,215	0,075	0,13	0,05	0,0295	0,04	7	6,19878652462242	0,000818804889192115
0,555	0,435	0,165	0,97	0,336	0,2315	0,295	17	13,2478833015995	0,0179571169877754
0,615	0,515	0,17	0,2	0,4305	0,2245	0,42	16	14,7073730156148	0,00213123025607249

Рисунок 19 – Фрагмент таблицы обучающего набора (для 6 нейронов внутреннего слоя)

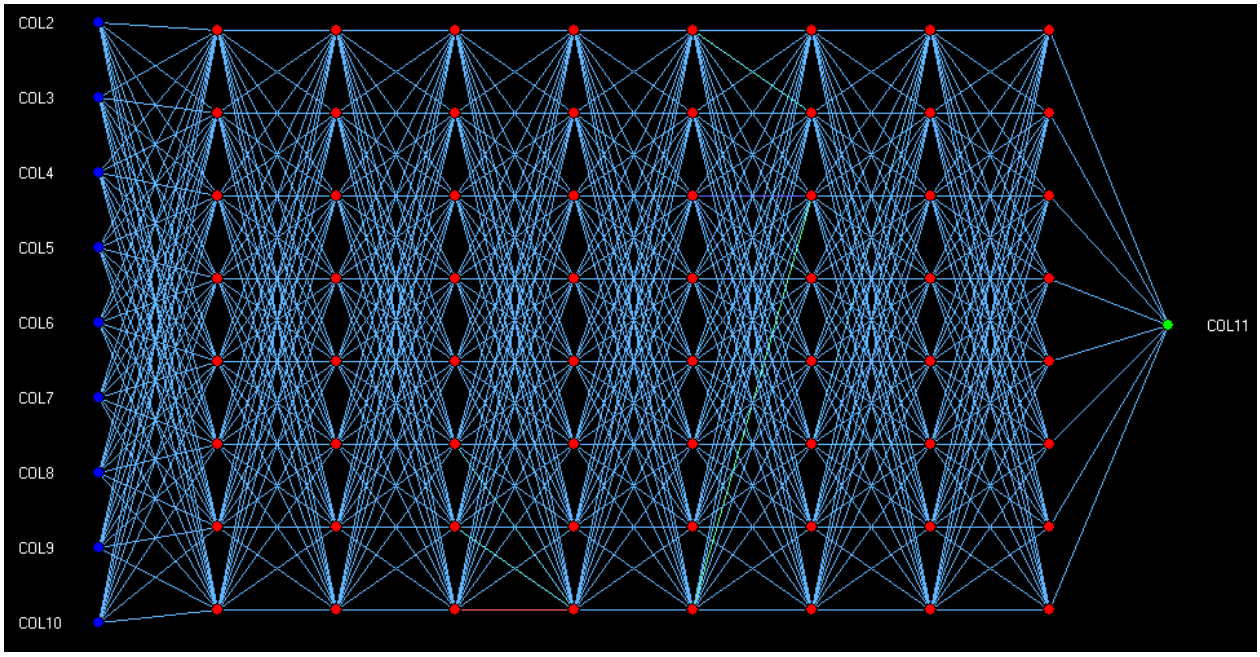


Рисунок 20 – Граф нейросети (для 8 нейронов внутреннего слоя)

Полученный результат улучшился. Точность классификации 97,16% на обучающем и 100,00% тестовом подмножестве можно назвать эффективным. Точность классификации удалось повысить, но для этого потребовалось больше времени.

3.5 Решение задачи в среде *SPSS*

Запустим мастер обработки и выберем пункт меню Анализ-> *Нейронные сети-> Многослойный Перспптрон*. В открывшемся окне зададим назначения параметров (рис.21). Атрибут «Возраст» являются зависимой переменной, все остальные факторы.

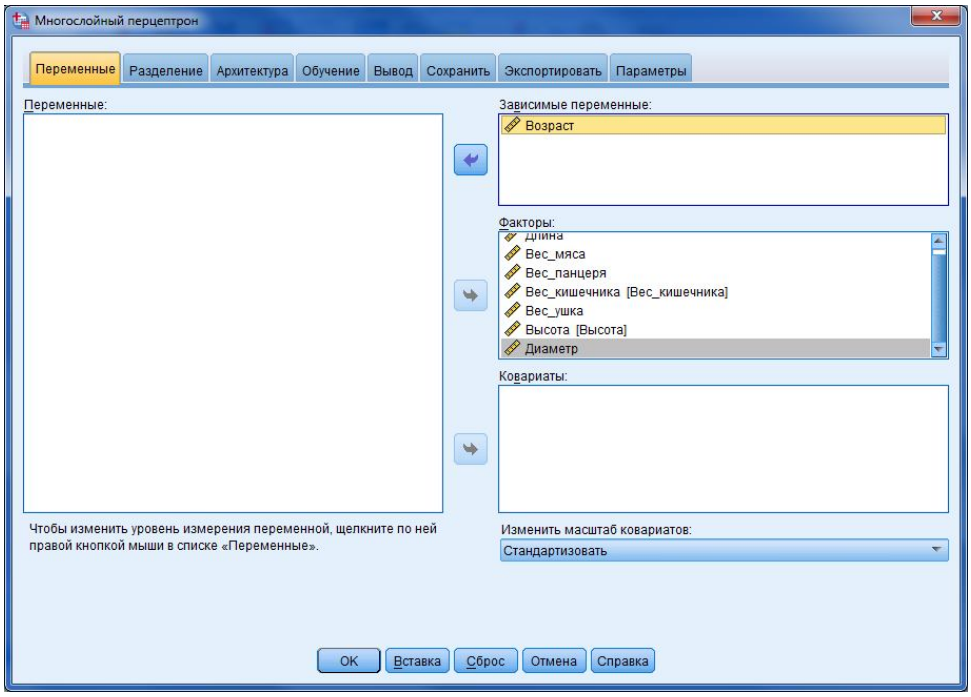


Рисунок 21 – Настройка параметров

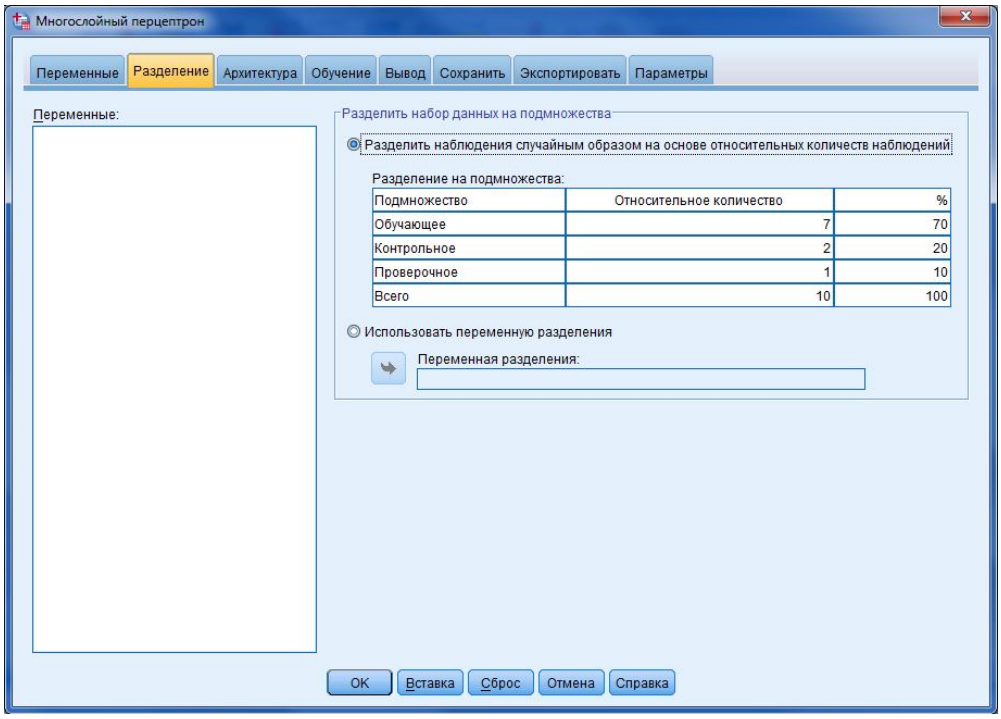


Рисунок 22 – Разделение данных

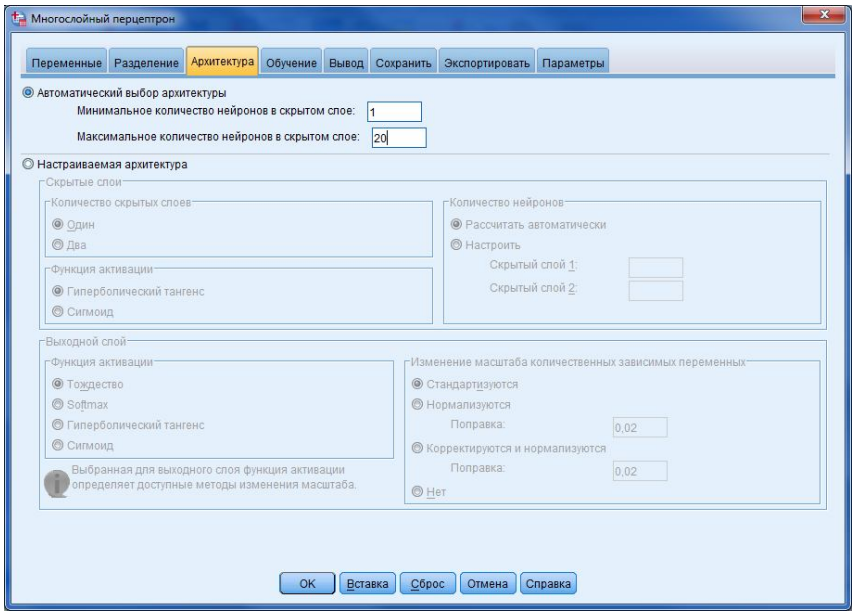


Рисунок 23 – Архетектура нейронной сети

Разделим параметры следующим образом: 70%- обучение, 20%- тестирование, 10%-проверка.

Минимальное количество нейроннов в скрытом слое: 1;

Максимальное количество нейроннов в скрытом слое: 20. (рис23)

Обучение: пакетное(рис. 24)

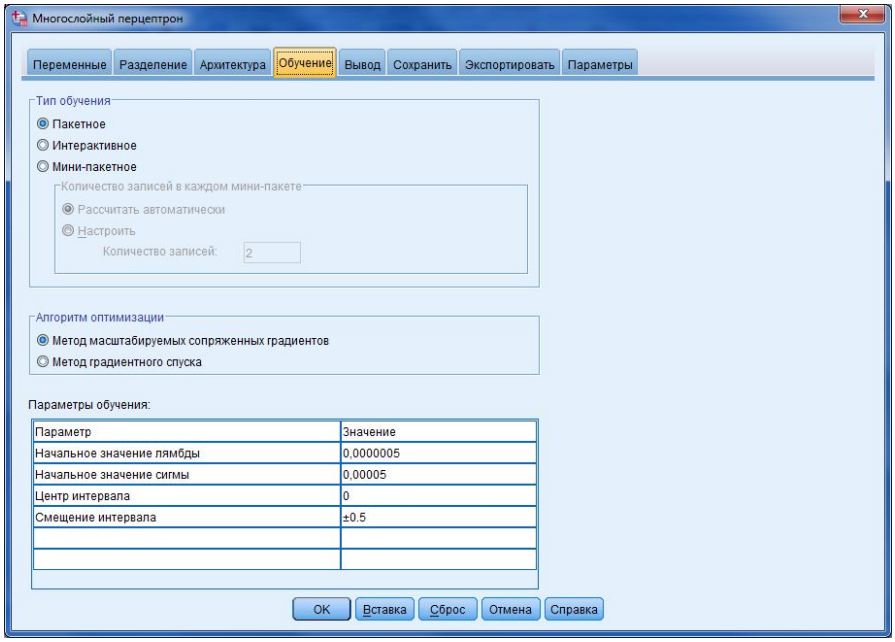


Рисунок 24 – Выбор способа обучения

3.5.1 Результаты

		N	Процент
Выборка	Обучающая	694	91,2%
	Контрольная	44	5,8%
	Проверочная	23	3,0%
Валидные		761	100,0%
Исключенные		239	
Всего		1000	

Рисунок 25 – Сводка обработки наблюдений

Входной слой	Факторы	1	Длина
		2	Вес_кишечника
		3	Вес_панцеря
		4	Вес_мяса
		5	Вес_ушка
		6	Высота
		7	Диаметр
	Количество нейронов[a]		1626
Скрытые слои	Количество скрытых слоев		1
	Количество нейронов в скрытом слое 1[a]		13
	Функция активации		Гиперболический тангенс
Выходной слой	Зависимые переменные	1	Возраст
	Количество нейронов		1
	Метод изменения масштаба для количественных зависимых переменных		Стандартизировано
	Функция активации		Единичная матрица
	Функция ошибки		Сумма квадратов

Рисунок 26 – Информация о сети

Результат, полученный в среде SPSS, по числу скрытых нейронов (13) совпал с результатом Statictic, при этом точность обучения (91,2%) ближе к результатам среды Deductor. Результат можно назвать успешным. Точность классификации 91,2% на обучающем подмножестве можно назвать эффективным.

4 Сравнительный анализ

	Statistica	**Deductor**	**SPSS**
Тип обучения	*Алгоритм BFGS*	*Алгоритм RPROP*	*Пакетное*
Тип нейронной сети	*Многослойный персептрон*	*Многослойный персептрон*	*Многослойный персептрон*
Количество нейронов во входном/выходном слоях	*7/1*	*7/1*	*7/1*
Количество скрытых слоев	*13*	*6*	*13*
Активационная функция	*Identity*	*Сигмоида*	*Гиперболический тангенс*
Точность на обучающем множестве	23,71%;	*96,37%*	*91,2%*
Точность на тестовом множестве	22,00%;	*96,00%*	-