

Министерство науки и высшего образования Российской Федерации
 федеральное государственное автономное
 образовательное учреждение высшего образования
 «Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа информационных технологий и робототехники
 Направление подготовки 09.04.03. Прикладная информатика
 Отделение школы (НОЦ) Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Система поддержки научных исследований медицинской диагностики
УДК 004.422.63:519.816:616-08-035

Студент

Группа	ФИО	Подпись	Дата
8KM71	Сидоров Владислав Викторович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина О.В.	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор отделения СГН	Сосковец Л.И.	д.и.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ООД ШБИП	Атепаева Н.А.			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина О.В.	к.т.н.		

Запланированные результаты обучения по программе

Код результата	Результат обучения (выпускник должен быть готов)
Профессиональные компетенции	
Р1	Применять базовые и специальные знания в области современных информационно-коммуникационных технологий для решения междисциплинарных инженерных задач.
Р2	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научнотехнической информации, математическое моделирование, проведение эксперимента, анализ и интерпретацию полученных данных в области информатизации и автоматизации прикладных процессов и создания, внедрения, эксплуатации и управления информационными системами в прикладных областях.
Р3	Внедрять, сопровождать и эксплуатировать современные информационные системы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья и безопасности труда, выполнять требования по защите окружающей среды.
Р4	Активно владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты инновационной инженерной деятельности.
Р5	Владеть и применять методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе глобальных компьютерных сетей.
Р6	Эффективно работать индивидуально, в качестве члена и руководителя группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.
Р7	Самостоятельно учиться и непрерывно повышать квалификацию в течение всего периода профессиональной деятельности.
Профиль «Системы корпоративного управления»	
Р8	Применять глубокие профессиональные знания основ построения информационных технологий и систем, достаточные для решения научных и профессиональных задач производства. Знать современные проблемы и методы прикладной информатики и научно-технического развития

	информационных технологий.
P9	Ставить и решать задачи комплексного анализа, связанные с информатизацией и автоматизацией прикладных процессов; созданием, внедрением, эксплуатацией и управлением информационными системами в прикладных областях, с использованием базовых и специальных знаний, современных аналитических методов и моделей.
P10	Организовывать работы по моделированию прикладных информационных систем и реинжинирингу прикладных и информационных процессов предприятия и организации. Управлять проектами по информатизации прикладных задач и созданию информационных систем предприятий и организаций.

		Графическое изображение кластеризации Полученное дерево решений
Консультанты по разделам выпускной квалификационной работы		
Раздел	Консультант	
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Сосковец Любовь Ивановна, д. и. н., профессор СГН	
Социальная ответственность	Атепаева Наталья Александровна, старший преподаватель ШБИП	
Английский язык	Диденко Анастасия Владимировна, к.ф.н., доцент ОИЯ	
Названия разделов, которые должны быть написаны на русском и иностранном языках:		
Введение		
Глава 1 Анализ предметной области		
Глава 3 Реализация проекта		

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	05.02.2019
---	------------

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина О. В.	к. т. н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8KM71	Сидоров В.В.		

Министерство образования и науки Российской Федерации
 Федеральное государственное автономное образовательное учреждение
 высшего профессионального образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
 ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Инженерная школа информационных технологий и робототехники
 Специальность 09.04.03. Прикладная информатика
 Уровень образования Магистратура
 Отделение Информационных технологий
 Период выполнения Весенний семестр 2019 учебного года
 Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	
--	--

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
01.02.2019	Получение задания на ВКР	
15.03.2019	Получение задания по финансовому менеджменту	
18.03.2019	Получение задания по социальной ответственности	
31.03.2019	Глава 1. Анализ предметной области	
08.04.2019	Глава 2. Методы и инструменты для решения поставленной задачи	
20.04.2019	Глава 3. Реализация проекта	
25.04.2019	Глава 4. Финансовый менеджмент	
30.04.2019	Глава 5. Социальная ответственность	
23.05.2019	Проверка оформления. Заключение.	
27.05.2019	Сдача выполненной работы.	

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина Ольга Владимировна	к. т. н.		

СОГЛАСОВАНО:

Руководитель ООП 09.04.03 Прикладная информатика	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина Ольга Владимировна	к. т. н.		

**ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И
РЕСУРСОСБЕРЕЖЕНИЕ»**

Студенту:

Группа	ФИО
8KM71	Сидорову Владиславу Викторовичу

Школа	Инженерная школа информационных технологий и робототехники	Отделение школы (НОЦ)	Информационных технологий
Уровень образования	Магистратура	Направление/специальность	09.04.03 Прикладная информатика

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	1. Стоимость расходных материалов 3240 руб. 2. Оплата руководителю 23618,49 руб., оплата исполнителя 10766 руб.
2. Нормы и нормативы расходования ресурсов	Норматив потребления электроэнергии 4 руб/кВтч
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	1. Отчисления во внебюджетные фонды 27,1% 2. Районный коэффициент 30% 3. Коэффициент дополнительной заработной платы 12%

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. Оценка коммерческого и инновационного потенциала НТИ	Оценка потенциальных потребителей исследования, анализ конкурентных решений, SWOT анализ, QuaD технология.
2. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок	Планирование этапов работ, определение трудоемкости и построение календарного графика, формирование бюджета, определение рисков НТИ.
3. Определение ресурсной, финансовой, экономической эффективности	Оценка показателей эффективности исследования.

Перечень графического материала (с точным указанием обязательных чертежей):

1. Оценка конкурентоспособности технических решений
2. QuaD -технология
3. SWOT- анализ
4. График проведения НТИ
5. Бюджет НТИ
6. Оценка ресурсной, финансовой и экономической эффективности НТИ

Дата выдачи задания для раздела по линейному графику

01.03.2019

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор отделения СГН	Сосковец Любовь Ивановна	Д. И. Н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8KM71	Сидоров Владислав Викторович		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8KM71	Сидорову Владиславу Викторовичу

Школа	Инженерная школа Информационных технологий и робототехники	Отделение (НОЦ)	Информационных технологий
Уровень образования	Магистратура	Направление/специальность	09.04.03 Прикладная информатика

Тема ВКР:

Система поддержки научных исследований медицинской диагностики	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Технологический процесс представляет собой работу с языком программирования R в программной среде RStudio. Основным оборудованием, на котором производится работа, является персональный компьютер. Областью применения данной разработки являются медицинские исследования и медицинские учреждения.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– СанПин 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы. – ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования. – ГОСТ 12.2.061-81 ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам. – Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019). – Закон Томской области от 9 июля 2003 года №83-ОЗ «Об охране труда в Томской области» (с изменениями на 4 июля 2014 года).
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	Вредные факторы: – отклонение показателей микроклимата; – недостаточная освещенность рабочей зоны; – превышение уровня шума; – опасные и вредные факторы, связанные с электромагнитными полями; – психофизиологические факторы; Опасные факторы: – повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека;
3. Экологическая безопасность:	– Воздействие объекта на окружающую природную среду: утилизация отходов, связанных с выходом из строя персональных компьютеров, люминесцентных ламп и другого офисного оборудования.

4. Безопасность в чрезвычайных ситуациях:	Возможные чрезвычайные ситуации на объекте: – Землетрясение; – Пожар (наиболее вероятен).
--	---

Дата выдачи задания для раздела по линейному графику	01.03.2019
---	-------------------

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ООД ШБИП	Атепаева Наталья Александровна			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8KM71	Сидоров Владислав Викторович		

РЕФЕРАТ

Выпускная квалификационная работа содержит 144 с., 62 рис., 25 табл., 74 источника, 3 приложения.

Ключевые слова: бронхиальная астма, набор данных, восстановление пропущенных значений, сокращение размерности, корреляция, кластеризация, логические правила, дерево решений.

Объект исследования – модель классификации условных типов бронхиальной астмы, разработанная учеными СибГМУ.

Предмет исследования – массив данных, содержащий значения клинико-лабораторных, а также психофизиологических показателей пациентов с различными формами бронхиальной астмы.

Целью исследования – создание системы поддержки научных исследований медицинской диагностики, изучение и применение методов анализа данных для подтверждения или опровержения возможности применения новой классификации, предложенной учеными СибГМУ.

Метод исследования – теоретический, практический, получение информации о предметной области, её исследование и анализ.

Во введении приведено обоснование актуальности исследования, выявлена проблема.

В первой главе представлен анализ предметной области, определен объект и цели исследования, представлена характеристика экспериментальных данных.

Вторая глава содержит описание методов и инструментов для решения поставленной задачи. Обзор методов сокращения размерности, кластеризации, построения деревьев решений. А также произведен выбор инструментария.

Третья глава содержит описание программной реализации проекта, приведены и проанализированы результаты исследования.

Четвертая глава содержит расчеты экономической эффективности разработки.

В пятой главе приводится описание социальной ответственности: условия работы и возможные риски для здоровья человека участвующего в исследовании.

Выпускная квалификационная работа выполнена в текстовом редакторе Microsoft Word 2010 и представлена на диске (в конверте на обороте обложки).

Оглавление

Введение.....	15
Обзор литературы.....	17
Глава 1 Анализ предметной области.....	21
1.1 Объект и цели исследования.....	21
1.2 Исходные данные и классификации бронхиальной астмы	21
1.3 Формальная постановка задачи	24
Глава 2 Методы и инструменты для решения поставленной задачи.....	26
2.1 Выбор методов решения.....	26
2.1.1 Методы заполнения пропущенных значений	26
2.1.2 Методы сокращения размерности.....	34
2.1.3 Кластеризация, анализ качества кластеризации	36
2.1.4 Дерево решений	37
2.2 Выбор инструментов	41
2.2.1 Скриптовый язык R	45
2.2.2 Типы данных языка R.....	46
2.2.3 Импортирование данных в R.....	47
Глава 3 Реализация проекта	48
3.1 Последовательность действий выполнения поставленной задачи	48
3.2 Первичная обработка данных	49
3.2.1 Анализ выбросов.....	49
3.2.2 Восстановление пропущенных значений.....	50
3.3 Сокращение размерности.....	52
3.4 Выявление однородных групп пациентов.....	54
3.5 Построение СППР в виде дерева решений	57
3.6 Анализ полученных результатов.....	59
Глава 4 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение.....	60
Введение.....	60

4.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения	61
4.1.1 Потенциальные потребители результатов исследования	61
4.1.2 Анализ конкурентных технических решений.....	62
4.1.3 Технология QuaD	63
4.1.4 SWOT-анализ	65
4.2 Планирование научно-исследовательских работ	66
4.2.1 Структура работ в рамках научного исследования.....	66
4.2.2 Определение трудоемкости выполнения работ.....	67
4.2.3 Разработка графика проведения научного исследования.....	69
4.3 Бюджет научно-технического исследования (НТИ)	72
4.3.1 Расчет материальных затрат НТИ.....	72
4.3.2 Основная заработная плата исполнителей темы	73
4.3.3 Дополнительная заработная плата	74
4.3.4 Отчисления во внебюджетные фонды (страховые отчисления).....	75
4.3.5 Накладные расходы	75
4.3.6 Формирование бюджета затрат научно-исследовательского проекта	76
4.4 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования ..	76
Глава 5 Социальная ответственность.....	80
Введение.....	80
5.1 Правовые и организационные вопросы обеспечения безопасности	80
5.2 Производственная безопасность	82
5.2.1 Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения.....	83
5.2.1.1 Отклонение показателей микроклимата.....	83
5.2.1.2 Недостаточная освещенность рабочей зоны.....	84
5.2.1.3 Превышение уровня шума	87
5.2.1.4 Опасные и вредные производственные факторы, связанные с электромагнитными полями	88

5.2.1.5 Психофизиологические факторы	89
5.2.2 Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения.....	90
5.2.2.1 Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	90
5.3 Экологическая безопасность.....	92
5.4 Безопасность в чрезвычайных ситуациях	92
5.5 Выводы по разделу	94
Заключение	96
Список использованных источников	97
Приложение А	106
Приложение Б.....	128
Приложение В.....	137

Введение

Бронхиальная астма – это заболевание, в основе которого лежит повышенная чувствительность бронхов к различным раздражителям (в частности, аллергенам). В ответ на действие этих раздражителей развивается сужение (обструкция) бронхов. Этот процесс обусловлен несколькими причинами: повышением тонуса бронхов, избыточного отделяемого в просвет бронхов и их воспаления. При астме приступы чаще всего возникают эпизодически, например, после контакта с раздражителем. При тяжелой астме обструкция бронхов часто сохраняется и между приступами. Данное заболевание является хроническим, поэтому ему требуется постоянный контроль. [26]

В настоящее время, по данным ВОЗ, с каждым годом увеличивается количество человек с таким заболеванием как «Бронхиальная астма» (БА). За последние двадцать лет количество больных людей с данным заболеванием стало составлять около трехсот миллионов человек в мире. Данное заболевание является одним из самых распространенных хронических заболеваний в мире, которому подвержены и мужчины, и женщины вне зависимости от возраста. При этом в большинстве случаев при появлении данного заболевания в детстве оно сохраняется и во взрослой жизни. [43]

Благодаря работе ученых, которая заключалась в анализе 34 исследований заболеваемости бронхиальной астмой в Европе, мы узнали, что в Австрии за 10 лет в период с 1992 по 2002 год заболеваемость детей выросла в 4 раза. В Италии в период с 1974 по 1998 увеличение составило 6% с 7% до 13%. Во многих странах Европы заболеваемость росла до середины 90-х годов XX века, но на текущий момент есть незначительное снижение [27].

Хоть в настоящее время идет эффективная терапия данного заболевания, но распространение астмы и увеличение количества больных ею, привело к тому, что исследователей вновь заинтересовал вопрос типирования этого заболевания. При этом методы типирования различались.

Одни исследовали клинические показатели больных, на основе которых принимали решение. Другие же рассматривали психологическое направление, при этом больные расценивались как однородная популяция людей в плане соматического статуса, но с разными психоэмоциональными состояниями. Но многообразие психологических и соматических состояний необходимо изучать комплексно.

В связи с этим актуальность работы не вызывает сомнений.

В основе исследования лежит предложенная томскими медиками (Языков К.Г., Немеров Е.В.) психосинергетическая модель психогенной бронхиальной астмы. В соответствии с данной моделью выделены эмпирические группы условно названные: БАПИ – бронхиальная астма психогенно - индуцированная; БАСП – бронхиальная астма соматопсихогенная (предварительный термин); БАНП – бронхиальная астма непсихогенная; ПО – психогенная одышка. [42]

Классификация является одним из самых важных и фундаментальных процессов в научных исследованиях и дальнейших применениях на практике. Внедрение же данной классификации бронхиальной астмы может привести к более конкретным методам лечения больных с различным типом данного заболевания.

Обзор литературы

Проблеме обработки и диагностики многомерных медицинских данных и использование для этого различных математических методов в настоящее время посвящено множество публикаций. В данной области исследований существует множество различных подходов – в зависимости от характера данных и поставленных целей. По данной теме, как в России, так и за границей, опубликовываются статьи, в которых не только предлагаются новые методики анализа, но также подчёркивается важность данной тематики в современном мире. В таблице 1 представлен обзор статей, основной темой которых является использование математических методов в области медицинской диагностики.

Таблица 1 – Обзор литературы

№	Авторы	Название статьи	Краткий обзор
1	Макарова Л.С., Берестнева О.Г., Семерякова Е.Г.	Математические методы в задачах медицинской диагностики. [1]	В работе рассматриваются задачи для получения правил решения для СППР в области дифференциальной диагностики бронхиальной астмы. Показаны и обработаны данные, полученные двумя различными методами: нейронной сетью и дискриминантным анализом.
2	Орлов А.И.	Математические методы теории классификации. [2]	В данной статье получен обзор методов для построения и использования классификаций. Были раскрыты наиболее часто используемые подходы для решения задач кластерного анализа и группировки.
3	Н. Бейли	Математика в биологии и медицине. [3]	Поверхностно знакомит с возможностями математических и вычислительных методов в медицинской диагностике.
4	Новосельцев В.Н.	Математические методы в медицине. [4]	Рассказывает о математических методах, используемых в медицине.
5	Черкашина Ю. А.	Применение математических методов в задаче диагностики состояния здоровья детей первого года жизни. [5]	Рассказывается о применении неоднородной последовательной процедуры распознавания (НППР) для распознавания заболевания.

6	Г.Г.Устинов, М.Л.Жмудяк, А.В. Стребуков, А.В.Гайнер, А.Н. Повалихин, А.Л.Жмудяк.	Диагностика заболеваний методами теорий вероятностей. [6]	В данном изыскании представлены задачи по совмещению математического подхода с медицинских подходом в области медицинской диагностики. А именно поиск необходимого диагностического признака для дальнейшего уточнения диагноза.
7	М.Д. Степанова, С.А. Самодумкин, Т.Л. Лемешева	Математические методы диагностики в медицинских интеллектуальных системах. [7]	В книге формулируются проблемы классификации, дается постановка задачи диагностики как классификационной задачи.
8	В.А. Биллиг, О.В. Иванова, Н.А. Царегородцев	Построение ассоциативных правил в задаче медицинской диагностики. [8]	В данной статье приведен алгоритм построения ассоциативных правил AprioriScale. Данный алгоритм необходим для решения диагностической задачи в области медицины. Данный алгоритм реализован на языке C#.
9	Компанченко А.С.	Сравнение эффективности ансамблей интеллектуальных технологий в задачах медицинской диагностики. [9]	Проведен анализ эффективности различных алгоритмов классификации в системе RapidMiner, а также анализ эффективности ансамблей этих алгоритмов, на примере решения задач медицинской диагностики.
10	Локтионова Н.Н. Фильчакова К.А.	Применение математических методов исследования в медицине. [10]	Рассказывается о применении математических методов в изучении процессов, происходящих на уровне целостного организма.
11	Горбунова Т.И.	Интеллектуальная система поддержки принятия решений в медицинской диагностике. [11]	В данном научном исследовании рассматривается создание лингвистической модели для помощи в принятии решений в медицинской диагностике.
12	Черкашина Ю.А., Вадутова Ф.А.	Оценка состояния здоровья детей с экстремально низкой массой тела на основе технологии искусственных нейронных сетей. [12]	Статья рассказывает о применении нейронных сетей для оценки здоровья детей.
13	Черкашина Ю.А.	Статистический анализ медико-биологических показателей у детей, больных ожирением. [13]	В данной статье автор рассказывает о применении статистического анализа для оценки состояния здоровья детей больных ожирением.
14	Заброда Н.Н., Артеменко	Математические методы анализа	Проверялось существование различий корреляции между значениями

	М.В.	заболеваемости населения региона Курской магнитной аномалии. [14]	заболеваний, которые были получены в городах с разной напряженностью магнитных полей, с помощью математических методов анализа.
15	Кувина В. Н., Самойленко Н. Э., Кувин С. С.	Комплексный анализ медицинских данных. [15]	Данная статья посвящена изложению методов комплексного анализа слабо формализованных данных в задачах, где требуется выявлять скрытые закономерности. Тематика данной работы имеет широкий спектр актуальных задач для применения, одна из них – это системы медицинской диагностики.
16	Стегалин И.С., Ананьев К.И., Демьянова В.В., Демьянов В.Ф., Кокорина А.В., Свистун С.Я., Демьянов В.Ф.	Оптимизационные методы в задачах диагностики. [16]	Рассматриваются результаты, которые получены в результате применения негладкого дискриминантного анализа. На основании изложенного подхода проводится анализ базы данных больных шизофренией.
17	Харитонов Т.С.	Задача классификации по уровню стрессоустойчивости с помощью нейронных сетей. [17]	В данной статье рассмотрена область применения нейронных сетей для классификации людей по уровню влияния на них стресса.
18	Мазуров В.Д., Шестаков А.А., Мазуров А.Д.	Математические модели диагностики и прогнозирования в медицине и биологии. [18]	В данной научной статье рассматриваются варианты обучения по прецедентам на основе нейронных сетей. Дается развернутое обоснование применения предложенных методов.
19	Шайдуров А.А.	Построение многокритериальной нейросетевой модели оптимизации в задачах диагностики заболеваний со слабо выраженной симптоматикой. [19]	В исследовании предлагается решение диагностики перинатального поражения ЦНС (центральной нервной системы). Диагностика строится на применении слоистых нейронных сетей.
20	Антонова Н.Е.	Сравнительный анализ применения статистических методов и нейронных сетей для диагностики заболеваний опорно-двигательного аппарата. [20]	В данной статье автор рассказывает об оценке результативности в использовании статистических методов и нейронных сетей для диагностики заболеваний опорно-двигательного аппарата. Была проведена оценка точности предсказания и работоспособности системы.
21	Головин П.А., Нечаев В.А., Нечаев Д.А.	Экспертные системы для классификации болезней в медицинской	Произведен анализ методов разработки экспертных систем для классификации болезней в медицинской диагностике. Изложены достоинства и недостатки

		диагностике. [21]	каждого из методов. Приведен обзор существующих медицинских экспертных систем.
22	Мишин А.С., Арзамасцев А.А., Зенкова Н.А.	Экспертная система для прогнозирования результатов хирургического лечения больных колоректальным раком, осложненных острой кишечной непроходимостью. [22]	В исследовании рассматривается создание экспертной системы, которая будет прогнозировать результаты хирургического лечения колоректального рака, осложненного острой непроходимостью. Данная система будет прогнозировать результаты при помощи нейронных сетей.
23	Алпеева А.В.	Результаты прогнозирования и распознавание характера заболевания системы внешнего дыхания у детей дошкольного возраста с помощью математических методов. [23]	Рассказывается о результатах прогнозирования с помощью математических методов.
24	Оразбаев Б.Б.	Экспертные системы для медицинской диагностики с применением методов теории нечетких множеств. [24]	На основе методов теории нечетких множеств предлагается подход к созданию экспертной системы для медицинской диагностики с учетом нечеткости исходной информации и оценок.
25	Ильсова Н.Ю., Куприянов А.В., Парингер Р.А.	Формирование признаков для повышения качества медицинской диагностики на основе методов дискриминантного анализа. [25]	В данной статье рассматривается вариант дополнения алгоритмом формирования информативных признаков системы диагностики глазных заболеваний.

На основе проведенного обзора литературы сделан вывод об актуальности разрабатываемой системы поддержки научных исследований медицинской диагностики. Практическая значимость выполненного исследования заключается в возможности соотнесения пациента к конкретной группе больных и применения соответствующего метода лечения.

Глава 1 Анализ предметной области

1.1 Объект и цели исследования

Объект исследования – модель классификации условных типов бронхиальной астмы, разработанная учеными СибГМУ.

Целью исследования – создание системы поддержки научных исследований медицинской диагностики, изучение и применение методов анализа данных для подтверждения или опровержения возможности применения новой классификации, предложенной учеными СибГМУ.

1.2 Исходные данные и классификации бронхиальной астмы

Городской клинической больницей №3 по городу Томску предоставлен массив данных, в котором содержатся данные о взрослых пациентах страдающих бронхиальной астмой, а именно их показатели до и после лечения. Пациентами являются взрослые люди с заболеванием «Бронхиальная астма». Далее на основе этих показателей будет выполнена дальнейшая работа. Ниже приведены показатели, содержащиеся в исходных данных:

Пол – пол больного;

Возраст – возраст больного, лет;

Рост – рост больного, см;

Вес – вес больного, кг;

Степ.ВН — степени вентиляционной недостаточности;

Ст.тяж.БА — степень тяжести бронхиальной астмы;

Возр.1эп. б-ни — возраст пациента во время первого эпизода болезни, лет;

Прод.б-ни — продолжительность болезни (лет);

Кл.пр.атопии – клинические проявления (признаки) атопии;

Показатели вентиляции легких: 1-до, 2- после АВС

МОД – Минутный объем дыхания, л/мин;

ЖЕЛ – Жизненная Емкость Легких, % (к должным величинам);

ФЖЕЛ – Форсированная ЖЕЛ, %;

ОФВ1 – Объем Форсированного Выдоха за 1 секунду, %;
ОФВ1/ЖЕЛ – отношение ОФВ1/ЖЕЛ, %;
МВЛ – Максимальная Вентиляция Легких за 1 мин, %;
ПОС выд. – Пиковая Объемная Скорость выдоха, %;
МОС25 – Максимальная Объемная Скорость выдоха на уровне 25% ФЖЕЛ;
МОС50 – Максимальная Объемная Скорость выдоха на уровне 50% ФЖЕЛ;
МОС75 – Максимальной Объемная Скорость выдоха на уровне 75% ФЖЕЛ;

Все показатели вентиляции легких (кроме МОД) выражены в % к должным величинам

Показатели механики дыхания: 1-до, 2- после АВС

C_{dyn} – динамическая растяжимость (C) легких, л/см.вод.ст.;

C_{stat} – статическая растяжимость C) легких, л/см.вод.ст.;

$R_{вд}$ – бронхиальное сопротивления на вдохе, см.вод.ст./л/сек;

$R_{выд}$ – бронхиальное сопротивление на выдохе, см.вод.ст./л/сек;

$W_{общ}$ – общая работа дыхания при спонтанном дыхании, кгм/мин;

$W_{уд}$ – удельной работы дыхания (на литр вентиляции), кгм/л;

$W_{н.эл}$ – неэластическая фракция общей работы дыхания, кгм/мин;

$W_{эл}$ – эластическая фракция общей работы дыхания, кгм/мин;

$W_{МОД10}$ – общая работа дыхания при МОД10, кгм/мин;

$W_{МОД15}$ – общая работа дыхания при МОД15, кгм/мин;

$W_{МВЛобщ}$ - общая работа дыхания при МВЛ, кгм/мин;

$W_{МВЛуд}$ - работы дыхания при МВЛ, кгм/л;

Показатели механики дыхания в дальнейшем приводятся к системе СИ

Существует несколько различных классификаций астмы такие как [51]:

1. По уровню контроля:

- Хорошо контролируемая;
- Частично контролируемая;

- Плохо контролируемая;

2.В соответствии со степенью тяжести. Данная классификация предложили академики РАМН: А. Д. Адо и Б. К. Булатов. Она основана на частоте и силе приступов и том, как сильно они влияют на жизнь пациента.

Существуют следующие виды:

- Интермиттирующая (эпизодическая) астма. Самая легкая форма по проявлениям, которая характеризуется длительной ремиссией;
- Персистирующая астма легкой степени тяжести. Симптомы проявляются уже чаще – за неделю приступы могут возникать несколько раз, причем как днем, так и ночью;
- Персистирующая астма средней степени тяжести. Для нее характерны дневные приступы, беспокоящие ежедневно. Ночные же ухудшения состояния случаются около двух раз в неделю. Эти приступы в значительной мере ухудшают жизнь пациента;
- Персистирующая бронхиальная астма тяжелой степени. Симптомы заболевания постоянно беспокоят больного, значительно затрудняется работоспособность.

3.Особые формы бронхиальной астмы:

- Профессиональная;
- Рефлюкс-индуцированная форма развивается при сниженном тонусе сфинктера пищевода;
- Аспириновая астма представляет собой псевдоаллергическую воспалительную реакцию организма, вызванную возникновением гиперчувствительности к нестероидным противовоспалительным препаратам, содержащим ацетилсалициловую кислоту;
- Физического напряжения – приступ наблюдается во время физических усилий, а также уже после нагрузки;
- Ночная астма. Приступы преимущественно возникают в ночное время.

4. По причинам, вызывающим приступ. Данная классификация получила распространение за границей, хотя она в последнее время подвергается критике. Она основана на причинах вызывающих приступы. Есть следующие виды астмы по данной классификации:

- Экзогенная форма бронхиальной астмы;
- Эндогенная форма бронхиальной астмы;
- Смешанного генеза.

В основе же данной работы лежит классификация, как было сказано ранее, предложенная томскими медиками: Языков К.Г., Немеров Е.В.

В основе исследования лежит психосинергетическая модель психогенной бронхиальной астмы. В соответствии с данной моделью выделены эмпирические группы, которые имеют следующие рабочие названия:

Первую группу – БАПИ, составляют люди, у которых первый приступ удушья развился после перенесенного эмоционального потрясения. Её ухудшение также связано с психологическими потрясениями.

Вторую группу – БАСП, составляют люди, у которых «обычное» течение болезни было нарушено жизненным стрессом, после которого психоэмоциональные триггеры вызывали тяжелые приступы удушья, обострения болезни.

Третью группу – БАНП, составляют люди с классической экзогенной формой бронхиальной астмы.

Четвертую группу – ПО, составляют люди с невротическим нарушением дыхания при таких явлениях как: панические расстройства, истерические состояния и различных психосоматических заболеваниях.

1.3 Формальная постановка задачи

Исходный массив данных нашей задачи – матрица X_{nt} ,

где n – количество пациентов;

m – число медицинских показателей.

Для решения поставленной задачи применяется многомерный анализ данных. Целью данного анализа становятся сложные системы, в которых элементы имеют множество зависимостей, таких как объекты и признаки. Все это представлено в виде матрицы, где строки – это объекты, а столбцы признаки.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

где n – число объектов (строк);

m – число признаков (столбцов);

x_{nm} – конкретное значение признака m у n -го элемента.

Для того чтобы выполнить поставленную задачу нужно провести анализ существующих методов и подходов, выбрать один из методов и посмотреть возможность его компьютерной реализации, который позволит обрабатывать предоставленный массив данных и достигнуть запланированной цели.

В результате решения задачи будет создана система поддержки научных исследований медицинской диагностики, основанная на методах интеллектуального анализа данных.

Глава 2 Методы и инструменты для решения поставленной задачи

2.1 Выбор методов решения

2.1.1 Методы заполнения пропущенных значений

Достаточно часто так происходит, что данные, с которыми необходимо работать, содержат пропущенные значения. Из-за этого человек, работающей с такими данными, сталкивается с проблемой выбора, что с такими данными делать: игнорировать, отбросить или же заполнить пропущенные значения. Заполнение пропущенных значений чаще всего оказывается более предпочтительным решением, но это не всегда так. Если при заполнении будет выбран не подходящий метод, то это может не только не улучшить, но и сильно ухудшить результат анализа. В некоторых случаях лучше исключать строки с пропущенными значениями, например при сильной разрежённости исходных данных.

На заре развития технологий заполнения и обработки пропущенных значений широко использовались «ad-hoc методы». Под данными методами подразумевается заполнение пропусков нулями, медианой, средним арифметическим значением. Данные методы достаточно просты, что и служит причиной их выбора. При этом сегодня уже существуют исследования, которые объясняют, что применение этих методов может приводить к искажению статистических свойств исходного набора данных, вследствие чего ухудшаются полученные результаты после такой обработки [Horton, 2007], но их по-прежнему часто используют. [32]

Применение методов «ad-hoc» или простое удаление строк с пропущенными значениями могут таить в себе множество скрытых камней, о которых нам необходимо знать. Поэтому мы коротко посмотрим данные методы и поймем их основные проблемы.

Для того чтобы понять, как правильно работать с пропусками, надо узнать как они могут появиться. Существует три основных механизма формирования пропусков:

MCAR (Missing Completely At Random) – механизм формирования пропусков, при котором вероятность пропуска для каждой записи набора одинакова. Например, при каких-либо опросах или тестированиях заведомо один из десяти человек не получал один случайно выбранный вопрос, при этом на все остальные вопросы были предоставлены и на них были получены ответы, то имеет место механизм MCAR. В таком случае игнорирование/исключение записей содержащих пропущенные данные не ведет к искажению результатов.

MAR (Missing At Random) – на практике данные обычно пропущены не случайно, а ввиду некоторых закономерностей. Пропуски относят к MAR, если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации, не содержащей пропуски. В таком случае удаление или замена пропусков на значение «Пропуск», как и в случае MCAR, не приведет к существенному искажению результатов.

MNAR (Missing Not At Random) – механизм формирования пропусков, при котором данные могут отсутствовать по неизвестным причинам. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

Далее уже рассмотрим сами методы борьбы с пропущенными значениями.

Complete-case Analysis (он же Listwise Deletion Method) – метод обработки пропусков, применяемый во множестве прикладных пакетов как метод по умолчанию. Заключается в исключении из набора данных записей/строк или атрибутов/колонок, содержащих пропуски.

В случае первого механизма пропусков (MCAR) применение данного метода не приведет к существенному искажению параметров модели. Однако удаление строк приводит к тому, что при дальнейших вычислениях используется не вся доступная информация, стандартные отклонения

возрастают, полученные результаты становятся не столь репрезентативными. В случаях, когда пропусков в данных много, это становится ощутимой проблемой.

Кроме того, в случае второго (MAR) и, особенно, третьего механизма пропусков (MNAR) смещение статистических свойств выборки, значений параметров построенных моделей и увеличение стандартных отклонений становятся еще сильнее.

Таким образом, несмотря на широкое распространение, применение данного метода для решения практических задач ограничено.

Available-case analysis (он же Pairwise Deletion) – методы обработки, основанные на игнорировании пропусков в расчетах. Эти методы, как и Complete-case Analysis, тоже часто применяются по умолчанию.

Статистические характеристики, такие как средние значения, стандартные отклонения, можно рассчитать, используя все непропущенные значения для каждого из атрибутов/столбцов. Как и в случае Complete-case Analysis, при условии выполнения гипотезы MCAR, применение данного метода не приведет к существенному искажению параметров модели.

Преимущество данного подхода в том, что при построении модели используется вся доступная информация. Главный же недостаток данных методов заключается в том, что они применимы для расчета далеко не всех показателей и, как правило, сопряжены с алгоритмическими и вычислительными сложностями, приводящие к некорректным результатам.

Когда гипотеза MCAR не выполняется, методы Available-case analysis так же, как и методы Complete-case Analysis приводят к существенным искажениям статистических свойств выборки (среднего значения, медианы, вариации, корреляции...).

К недостаткам первых двух методов обработки пропусков (Complete-case Analysis и Available-case analysis) относится и то, что, далеко не всегда исключение строк в принципе приемлемо. Нередко процедуры последующей обработки данных предполагают, что все строки и колонки участвуют в

расчетах (например, когда пропусков в каждой колонке не очень много, но при этом строк, в которых нет ни одного пропущенного поля мало).

Далее рассмотрим методы, которые предполагают заполнение пропусков на основе имеющейся информации. Часто эти методы объединяют в одну группу, называемую Single-imputation methods.

Заполнение пропуска средним значением (Mean Substitution) (другие варианты: заполнение нулем, медианой и тому подобные.) – подразумевает заполнение пропущенных значений средним значением.

Всем вариантам данного метода свойственны одни и те же недостатки. Данный метод приводит к существенному искажению распределения характеристики даже в случае MCAR.

«**knnImpute**» – метод основан на простейшем, но чрезвычайно эффективном алгоритме kNN.

Метод kNN предполагает, что необходимый и тестируемый объект d имеет примерно те же показатели и признаки, что и объекты в его соседнем окружении.

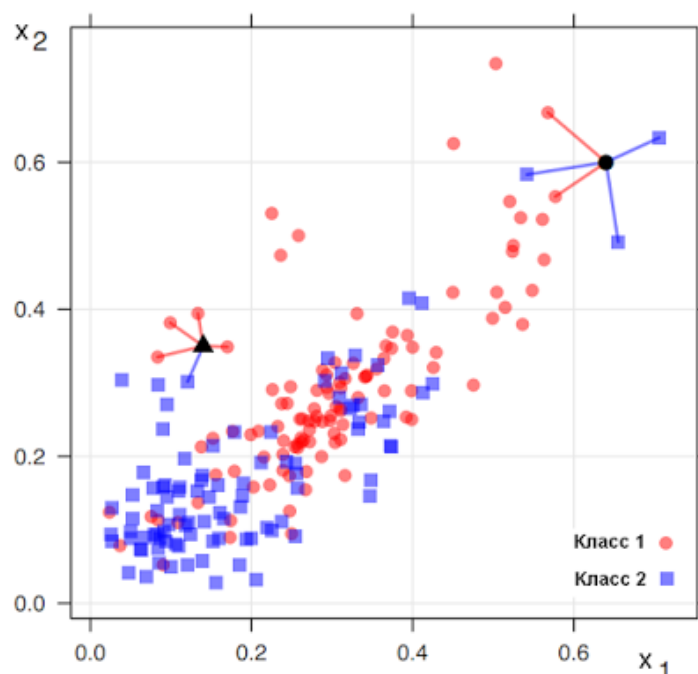


Рисунок 1 – Метод knn

Вначале данного алгоритма нам необходимо задать число k , т.е. количество соседей, на основании которых и будет работать алгоритм

дальше. Если принять $k = 1$, то алгоритм теряет обобщающую способность, ведь в данном случае он просто будет копировать ближайший объект, при этом если установить слишком большое значение, то многие локальные особенности будут утеряны.

Если речь идет о классификации, то неизвестный класс объекта определяется голосованием k его ближайших соседей (на рис. 1 $k = 5$). kNN-регрессия оценивает неопределенное значение неизвестной координаты Y , усредняя известные ее величины для тех же k соседних точек.

Одна из важных проблем kNN – выбор метрики, на основе которой оценивается близость объектов. Наиболее общей формулой для подсчета расстояния в m -мерном пространстве между объектами X_1 и X_2 является мера Минковского [56]:

$$D_s(X_1, X_2) = \left[\sum_i |X_{1i} - X_{2i}|^p \right]^{1/r},$$

где i изменяется от 1 до m , а r и p – задаваемые исследователем параметры, с помощью которых можно осуществить нелинейное масштабирование расстояний между объектами. Мера расстояния по Евклиду получается, если принять в метрике Минковского $r=p=2$, и является, по-видимому, наиболее общей мерой расстояния, знакомой всем по школьной теореме Пифагора. При $r=p=1$ имеем "манхеттенское расстояние", не столь контрастно оценивающее большие разности координат x . При этом необходимо помнить, что все признаки должны быть измерены «в одном масштабе», иначе признаки с наибольшими числовыми значениями будут преобладать, остальные же будут попросту утеряны. При кластерном анализе есть два основных способа нормализации данных:

Мини-макс нормализация [55]:

$$x' = (x - MIN[X]) / (MAX[X] - MIN[X])$$

В этом случае все значения будут лежать в диапазоне от 0 до 1. Дискретные бинарные значения определяются как 0 и 1.

Z-нормализация:

$$x' = (x - M[X])/\sigma[X],$$

где σ – среднеквадратическое отклонение. В этом случае большинство значений попадет в диапазон $(-3\sigma; 3\sigma)$.

Вторая проблема с kNN заключается в решении вопроса о том, на мнение скольких конкретно соседей k нам целесообразно положиться? В практических задачах значение k задают с помощью критерия скользящего контроля, так же возможен вариант, в котором используется исключение объектов по одному.

Наиболее часто в реальных задачах данный метод имеет сомнительную эффективность. Так же проблемой данного классификатора является скорость классификации.

Метод «bagInput».

Данный метод реализует для каждой из имеющихся переменных построение множественной бутстреп-агрегированной модели или бэггинг-модели (англ. bagging) на основе деревьев регрессии, принимая все остальные переменные в качестве предикторов. Этот метод точен, но требует значительных затрат времени на вычисление, особенно при работе с данными большого объема.

Bagging (от Bootstrap aggregation) – это один из первых и самых простых видов ансамблей. Он был придуман Лео Брэйманом в 1994 году. Бэггинг основан на статистическом методе бутстрэпа, который позволяет оценивать многие статистики сложных распределений. [50]

Метод бутстрэпа заключается в следующем. Пусть имеется выборка X размера N . Равномерно возьмем из выборки N объектов с возвращением. Это означает, что мы будем N раз выбирать произвольный объект выборки (считаем, что каждый объект «достаётся» с одинаковой вероятностью $1/N$), причем каждый раз мы выбираем из всех исходных N объектов. Можно

представить себе мешок, из которого достают шарики: выбранный на каком-то шаге шарик возвращается обратно в мешок, и следующий выбор опять делается равновероятно из того же числа шариков. Отметим, что из-за возвращения среди них окажутся повторы. Обозначим новую выборку через X_1 . Повторяя процедуру M раз, сгенерируем M подвыборок $X_1, \dots, X_m$. Теперь мы имеем достаточно большое число выборок и можем оценивать различные статистики исходного распределения. На рисунке 2 вы можете видеть схему бутстрэпа.

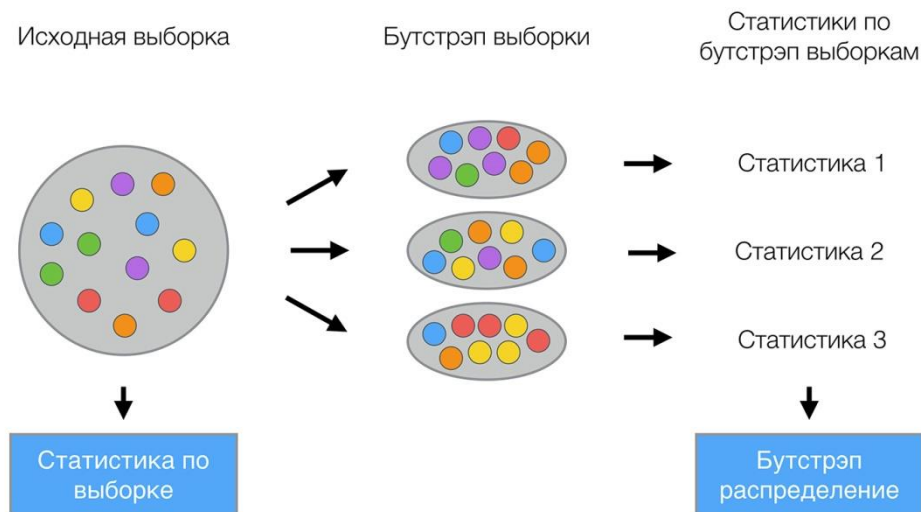


Рисунок 2 – Схема бутстрэпа

Теперь имея представление о бутстрэпе, мы можем перейти непосредственно к бэггингу. Пусть имеется обучающая выборка X . С помощью бутстрэпа сгенерируем из неё выборки $X_1, \dots, X_m$. Теперь на каждой выборке обучим свой классификатор $a_i(x)$. Итоговый классификатор будет усреднять ответы всех этих алгоритмов (в случае классификации это соответствует голосованию) [50]:

$$a(x) = \frac{1}{M} \sum_{i=1}^M a_i(x)$$

Эту схему можно представить картинкой ниже – рисунок 3.

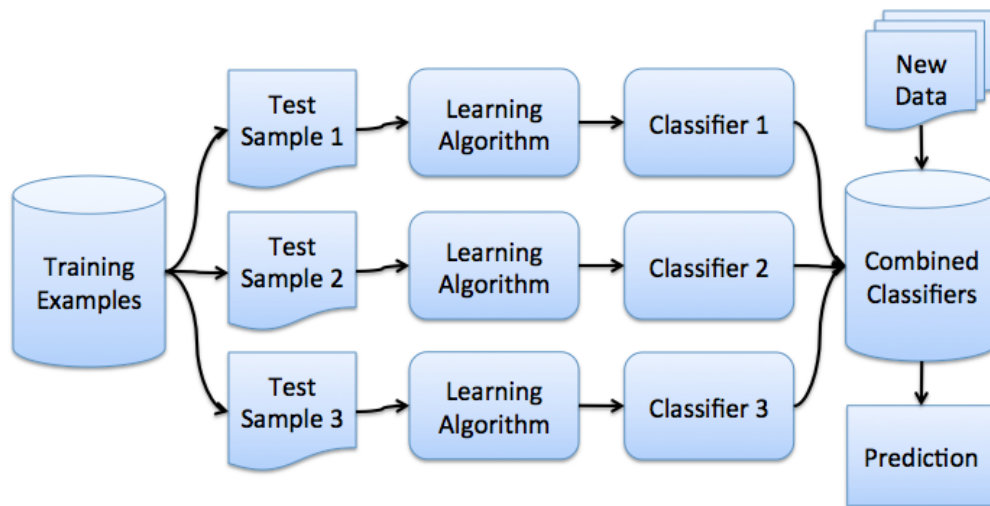


Рисунок 3 – Схема бэггинга

Рассмотрим задачу регрессии с базовыми алгоритмами $b_1(x), \dots, b_n(x)$. Предположим, что существует истинная функция ответа для всех объектов $y(x)$, а также задано распределение на объектах $p(x)$. В этом случае мы можем записать ошибку каждой функции регрессии:

$$\varepsilon_i(x) = b_i(x) - y(x), i = 1, \dots, n$$

и записать математическое ожидание среднеквадратичной ошибки

$$E_x(b_i(x) - y(x))^2 = E_x \varepsilon_i^2(x)$$

Средняя ошибка построенных функций регрессии имеет вид

$$E_1 = \frac{1}{n} E_x \sum_{i=1}^n \varepsilon_i^2(x)$$

Предположим, что ошибки несмещены и некоррелированы:

$$E_x \varepsilon_i(x) = 0,$$

$$E_x \varepsilon_i(x) \varepsilon_j(x) = 0, i \neq j$$

Построим теперь новую функцию регрессии, которая будет усреднять ответы построенных нами функций:

$$a(x) = \frac{1}{n} \sum_{i=1}^n b_i(x)$$

Найдем ее среднеквадратичную ошибку:

$$\begin{aligned}
E_n &= E_x \left(\frac{1}{n} \sum_{i=1}^n b_i(x) - y(x) \right)^2 \\
&= E_x \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \right)^2 \\
&= \frac{1}{n^2} E_x \left(\sum_{i=1}^n \varepsilon_i^2(x) + \sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x) \right) \\
&= \frac{1}{n} E_1
\end{aligned}$$

Таким образом, усреднение ответов позволило уменьшить средний квадрат ошибки в n раз.

Бэггинг дает нам возможность уменьшить дисперсию (variance) классификатора, который мы обучаем. Уменьшая величину, насколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность данного метода обуславливается тем, что базовые алгоритмы, обученные по различным подвыборкам, имеют достаточно различий. Вследствие чего при голосовании на выбранный вариант возможные ошибки, допущенные в каждом отдельном классификаторе, компенсируют друг друга. Еще за счёт того, что объекты-выбросы, появляющиеся в некоторых подвыборках, могут не попадать в другие подвыборки.

2.1.2 Методы сокращения размерности

При просмотре исходного набора данных мы увидели, что количество строк пациентов меньше чем количество столбцов с параметрами из-за этого дальнейшей его анализ осложняется возможной недостоверностью и неудобством работы. Для того чтобы исключить данные факторы нам необходимо уменьшить количество переменных, т.е. сократить размерность. Существует несколько методов сокращения размерности:

- Метод главных компонент;
- Метод экстремальной группировки параметров;
- Метод корреляционных плеед;
- Многомерное шкалирование;

Метод главных компонент является одним из наиболее часто используемых методов снижения размерности. Основная его идея состоит в последовательном выявлении направлений, в которых данные имеют наибольший разброс.

Для визуального анализа данных часто используют проекции исходных векторов на плоскость первых двух главных компонент. Обычно хорошо видна структура данных, выделяются компактные кластеры объектов и отдельно выделяющиеся вектора. [33]

Метод экстремальной группировки параметров используется, когда изменение какого-либо общего фактора неодинаково на измеряемых частных признаках. В частности, возможна ситуация, когда исходная совокупность из p признаков обнаруживает расщепление на сравнительно небольшое число групп, при котором изменение признаков, относящихся к какой-либо одной группе, обуславливается в основном каким-то одним общим фактором, своим для каждой такой группы. [34]

Метод корреляционных плеяд также как и метод экстремальной группировки, и кластерный анализ, предназначен для нахождения таких групп признаков – ‘плеяд’, когда корреляционная связь, которая определяется как сумма модулей коэффициентов корреляции между параметрами одной группы (внутриплеядная связь), достаточно велика, а связь между параметрами из разных групп (межплеядная) – мала. На основе корреляционной матрицы строится граф связей, который разбивается по тем или иным признакам на подграфы. Вершины подграфа и образуют плеяду. [34]

Многомерное шкалирование. Основная идея этого класса методов состоит в представлении каждого объекта точкой геометрического пространства (обычно размерности 1, 2 или 3), координатами которой служат значения скрытых (латентных) факторов, в совокупности достаточно адекватно описывающих объект. При этом отношения между объектами заменяются отношениями между точками – их представителями. Так, данные

о сходстве объектов – расстояниями между точками, данные о превосходстве – взаимным расположением точек. [33]

2.1.3 Кластеризация, анализ качества кластеризации

Кластеризация – это группировка объектов на основе данных, описывающих сущность этих объектов. Объекты внутри кластера должны быть «похожими» друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем больше объекты похожи внутри кластера и чем больше различий между кластерами, тем точнее кластеризация. [35]

Существуют различные методы кластерного анализа, их можно разделить на две группы:

1. Иерархические:

- Иерархические агломеративные методы (Agglomerative Nesting, AGNES);
- Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA).

2. Неиерархические (итеративные) методы:

- Алгоритм k -средних (k -means);
- Алгоритм PAM (partitioning around Medoids).

Используя различные методы кластерного анализа, можем получать различные решения для одних и тех же данных. Алгоритмы неиерархического разделения осуществляют декомпозицию набора данных, состоящего из n наблюдений, на k групп (кластеров) с заранее неизвестными параметрами [22].

В работе применён наиболее распространенный алгоритм k -средних. Алгоритм разбивает объекты на k групп, где каждая группа представляет один кластер. Принципиальное отличие метода k -средних от иерархического кластер-анализа заключается в том, что необходимо изначально определить число кластеров, на которое требуется разбить изучаемую совокупность. Соответственно, желательно еще до начала анализа иметь гипотезу о структуре исследуемой совокупности. С помощью данного метода была

исследована структура медицинских данных, проводились случаи кластеризации для $k = 3, 4, 5$.

Для оценки качества кластеризации используется индекс Данна [37]. Индекс Данна сравнивает межкластерное расстояние с диаметром кластера. Считается, что если диаметр кластера мал по сравнению с межкластерным расстоянием, то кластеры полученной структуры достаточно компактны и отделимы. Следовательно, чем больше индекс, тем лучше кластеризация.

$$D = \frac{\min_{1 \leq k < k^* \leq K} d(i, j)}{\max_{1 \leq k \leq K} d(k)}$$

$d(i, j)$ – расстояние между кластерами i и j .

$d(k)$ – внутрикластерное расстояние для k -того кластера.

Индекс Данна чувствителен к выбросам, поэтому мы предварительно исследовали данные на выбросы.

2.1.4 Дерево решений

Деревья решений – один из методов автоматического анализа данных. Первые идеи создания деревьев решений восходят к работам Ховленда (Hoveland) и Ханта (Hunt) конца 50-х годов XX века. Однако, основополагающей работой, давшей импульс для развития этого направления, явилась книга Ханта (Hunt, E.B.), Мэрина (Marin J.) и Стоуна (Stone, P.J) "Experiments in Induction", увидевшая свет в 1966г. [49]

Дерево принятия решений (также может называться деревом классификации или регрессионным деревом) – средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко

используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе. Пример дерева решений представлен на рисунке 4.



Рисунок 4 – Дерево решений

Каждый лист представляет собой значение целевой переменной, изменённой в ходе движения от корня по листу. Каждый внутренний узел соответствует одной из входных переменных. Дерево может быть также «изучено» разделением исходных наборов переменных на подмножества, основанные на тестировании значений атрибутов. Это процесс, который повторяется на каждом из полученных подмножеств. Рекурсия завершается тогда, когда подмножество в узле имеет те же значения целевой переменной, таким образом, оно не добавляет ценности для предсказаний. Процесс, идущий «сверху вниз», индукция деревьев решений (TDIDT), является примером поглощающего «жадного» алгоритма, и на сегодняшний день является наиболее распространённой стратегией деревьев решений для данных, но это не единственная возможная стратегия. В интеллектуальном анализе данных, деревья решений могут быть использованы в качестве математических и вычислительных методов, чтобы помочь описать,

классифицировать и обобщить набор данных, которые могут быть записаны следующим образом:

$$(x, Y) = (x_1, x_2, x_3 \dots x_k, Y)$$

Зависимая переменная Y является целевой переменной, которую необходимо проанализировать, классифицировать и обобщить. Вектор x состоит из входных переменных x_1, x_2, x_3 и т. д., которые используются для выполнения этой задачи.

Деревья решений, используемые в Data Mining, бывают двух основных типов:

- Дерево для классификации, когда предсказываемый результат является классом, к которому принадлежат данные;
- Дерево для регрессии, когда предсказываемый результат можно рассматривать как вещественное число (например, цена на дом, или продолжительность пребывания пациента в больнице).

Упомянутые выше термины впервые были введены Брейманом и др. Перечисленные типы имеют некоторые сходства (рекурсивный алгоритмы построения), а также некоторые различия, такие, как критерии выбора разбиения в каждом узле.

Есть различные способы выбирать очередной атрибут [52]:

- Алгоритм ID3, где выбор атрибута происходит на основании прироста информации (англ. Gain), либо на основании критерия Джини.
- Алгоритм C4.5 (улучшенная версия ID3), где выбор атрибута происходит на основании нормализованного прироста информации (англ. Gain Ratio).
- Алгоритм CART и его модификации – IndCART, DB-CART.
- Автоматический детектор взаимодействия Хи-квадрат (CHAID). Выполняет многоуровневое разделение при расчёте классификации деревьев;
- MARS: расширяет деревья решений для улучшения обработки цифровых данных.

На практике в результате работы этих алгоритмов часто получаются слишком детализированные деревья, которые при их дальнейшем применении дают много ошибок. Это связано с явлением переобучения. Для сокращения деревьев используется отсечение ветвей (англ. pruning).

Среди прочих методов Data Mining, метод дерева принятия решений имеет несколько достоинств:

1.Простой в понимании и интерпретации. Люди способны интерпретировать результаты модели дерева принятия решений после краткого объяснения

2.Не требует подготовки данных. Прочие техники требуют нормализации данных, добавления фиктивных переменных, а также удаления пропущенных данных.

3.Способен работать как с категориальными, так и с интервальными переменными. Прочие методы работают лишь с теми данными, где присутствует лишь один тип переменных. Например, метод отношений может быть применён только на номинальных переменных, а метод нейронных сетей только на переменных, измеренных по интервальной шкале.

4.Использует модель «белого ящика». Если определённая ситуация наблюдается в модели, то её можно объяснить при помощи булевой логики. Примером «черного ящика» может быть искусственная нейронная сеть, так как результаты данной модели поддаются объяснению с трудом.

5.Позволяет оценить модель при помощи статистических тестов. Это даёт возможность оценить надёжность модели.

6.Является надёжным методом. Метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включённые в модель.

7.Позволяет работать с большим объёмом информации без специальных подготовительных процедур. Данный метод не требует специального оборудования для работы с большими базами данных.

При этом данный метод имеет и ряд недостатков:

1. Проблема получения оптимального дерева решений является *NP*-полной с точки зрения некоторых аспектов оптимальности даже для простых задач. Таким образом, практическое применение алгоритма деревьев решений основано на эвристических алгоритмах, таких как алгоритм «жадности», где единственно оптимальное решение выбирается локально в каждом узле. Такие алгоритмы не могут обеспечить оптимальность всего дерева в целом.

2. В процессе построения дерева решений могут создаваться слишком сложные конструкции, которые недостаточно полно представляют данные. Данная проблема называется переобучением. Для того, чтобы её избежать, необходимо использовать метод «регулирования глубины дерева».

3. Существуют концепты, которые сложно понять из модели, так как модель описывает их сложным путём. Данное явление может быть вызвано проблемами XOR, чётности или мультиплексарности. В этом случае мы имеем дело с непомерно большими деревьями. Существует несколько подходов решения данной проблемы, например, попытка изменить репрезентацию концепта в модели (составление новых суждений), или использование алгоритмов, которые более полно описывают и репрезентируют концепт (например, метод статистических отношений, индуктивная логика программирования).

4. Для данных, которые включают категориальные переменные с большим набором уровней (закрытий), большой информационный вес присваивается тем атрибутам, которые имеют большее количество уровней.

2.2 Выбор инструментов

Качество и скорость выполняемой работы связаны с инструментами, которые будут использоваться в ходе работы. Поэтому хоть мы и можем провести обработку исходного набора данных только лишь с помощью ручки и бумаги, но использование специализированных инструментов, таких как

программное обеспечение, для статистической обработки данных, позволит нам сделать это эффективнее и быстрее.

Все статистическое программное обеспечение (ПО) можно поделить на два основных вида: ПО с графическим интерфейсом и ПО без графического интерфейса. Наличие данной возможности не характеризует качество программы с точки зрения статистической обработки данных, но является одним из самых важных, т.к. графический интерфейс повышает удобство работы с программой и не все люди могут комфортно работать с помощью командной строки. Такое деление размывается в серьезном ПО для статистики, потому что в таких приложениях графический интерфейс объединяют с возможностью работы с помощью командной строки со встроенным языком программирования.

На данный момент насчитываются десятки качественных статистических пакетов, рассмотрим наиболее популярные из них. [38]

1. Система SAS. Появилась данная система в 60-х годах прошлого века в университете США как приложение для анализа результатов сельскохозяйственных исследований. На текущий момент времени данная программа развивается в компании SAS Institute и имеет девять версий. Область применения данного программного обеспечения широка: от обычной бизнес аналитики до разнообразных научных исследований.

Данная система состоит из модулей, где каждый модуль отвечает за определенный круг задач и может быть не использован, если круг задач за который он отвечает, не задействована. SAS работает на собственном языке программирования. Она позволяет загружать данные непосредственно в терминал или же можно использовать для этого внешние файлы. SAS дает возможность проводить статистический анализ данных значений различного уровня сложности, в зависимости от поставленной задачи. Работа с программным продуктом возможна в двух формах: в консольном режиме и через графический интерфейс пользователя.

2. Язык программирования R. Данный язык программирования произошел от языка S, с которым он имеет много общих черт, этим объясняется его надежность, т.к. язык S был очень мощным языком программирования в свое время. Стандартный язык программирования R не имеет графического интерфейса, а работает с командной строкой. Графический интерфейс легко устанавливается как отдельное приложение и работает с уже установленной версией R на данном компьютере. Пример такого интерфейса это среда разработки RStudio, которая распространяется по бесплатной модели и имеет минималистический интерфейс. Язык R имеет обширные возможности для статистической обработки данных, в том числе и для работы с графиками. Но надо понимать, что графические интерфейсы для R, такие как RStudio, заметно слабее, чем в других аналогичных программах. Одной из причин данного факта является и его преимущество перед другими программами – это его бесплатное распространение. Синтаксис данного языка программирования достаточно прост и легок в изучении. Установить же его можно на компьютеры с такими системами как: Windows, Linux, MacOS. Так же одним из плюсов данного языка является свободная возможность написания пакетов с реализациями различных методов. Это привело к тому, что на текущий момент в R реализовано огромное количество функций и библиотек руками заинтересованных пользователей. Но сейчас очень мало книг по R на русском языке.

3. Stat – программа для анализа данных использующая преимущественно возможности командной строки. Данная программа разрабатывается компанией StataCorp. Установить же его можно на компьютеры с такими системами как: Windows, Linux, MacOS. Ввод же данных осуществляется с помощью встроенного редактора таблиц, который хоть довольно и прост, но позволяет все необходимые операции над таблицами. Также можно осуществлять ввод с помощью внешних файлов. Методы работы с данным программным обеспечением не отличаются от методов при использовании описанных выше программ.

4. Кроме уже описанных выше программ для анализа данных, которые имеют возможность работать с данными из различных отраслей и разными типами, существуют также программы конкретно для ученых, работающих в области медицинских исследований. Например, программа MedCalc, разрабатываемая компанией MedCalc Software. Разработчики позиционируют его как полноценное статистическое приложение для исследователей в области медицины. Главным пунктом, на котором заостряют свое внимание разработчики это удобство использования данного ПО для анализа ROC-кривых. Загрузка данных в MedCalc производится во встроенном табличном редакторе или же путем импорта файлов различных форматов, таких как *.csv, excel и др. Работать с данным приложением могут только пользователи Windows.

5. Statistica – это программный продукт разрабатываемый компанией StatSoft. Программа имеет собственный графический интерфейс, а также встроенный язык программирования. Имеет очень широкие возможности в статистической обработке, но это приводит к высоким требованиям к самому компьютеру. Данная программа работает только под управлением системы Windows.

6. SPSS названием данной программы является её аббревиатура (Statistical Package for the Social Sciences), в 2009 году данную программу выкупили IBM и сменили название на PASW (Predictive Analytics SoftWare). Данная система имеет широкие графические возможности. Она имеет в своем арсенале большое количество различных категорий и типов графиков таких как: научные, деловые, трехмерные, двухмерные, выполненные в различных системах координат, специализированные статистические графики и другие. К сожалению требования данного программного продукта к производительности системы, по сравнению с остальными, являются высокими, необходимо не менее 1 Гб оперативной памяти для того чтобы начать работать.

Проведем сравнительный анализ данных систем на наибольшую эффективность в задачах статистического анализа (таблица 2).

Таблица 2 - Сравнительный анализ инструментов решений

Признак/Инструмент	SAS	R	Stata	MedCalc	Statistica	SPSS
Наличие графического интерфейса	+	+	+	+	+	+
Наличие консольного ввода	+	+	+	+	+	+
Свободное распространение	-	+	-	-	-	-
Работа с графикой	+	+	+	+	+	+
Поддержка ОС Linux и Windows	+	+	+	-	-	+
Простота в понимании и изучении	+	+	+	+	-	-
Ресурсоэффективная	+	+	+	+	-	-

На основе проведенного сравнительного анализа для решения поставленной задачи был выбран язык R. В качестве среды разработки выбрано – Rstudio.

2.2.1 Скриптовый язык R

R – это среда вычислений, была разработана Россом Айхэкой (англ. Ross Ihaka) и Робертом Джентлменом (англ. Robert Gentleman) для обработки данных, математического моделирования и работы с графикой. Его применение широко. R можно применять как калькулятор, как редактор таблиц. С его помощью можно проводить простой статистический анализ или сложный. [44]

R – это язык программирования, что позволяет нам писать свои собственные программы и создавать для него пакеты или брать готовые, внутри которых будут реализованы необходимые нам функции. Язык и среда поддерживаются и развиваются организацией R Foundation. [39]

R – это бесплатный продукт с открытым кодом. Из-за этого его разработкой, тестированием и отладкой занимается не отдельная компания с нанятым персоналом, а сами пользователи. Благодаря этому библиотека пакетов языка непрерывно расширяется, а сам R занимает ведущие позиции в статистике, в анализе данных и в добыче данных. [40]

2.2.2 Типы данных языка R

Все объекты данных (а, следовательно, и переменные) в R можно разделить на следующие классы (т.е. типы объектов) [41]:

- `numeric` – объекты, к которым относятся целочисленные (`integer`) и действительные числа (`double`);
- `logical` – логические объекты, которые принимают только два значения: `FALSE` (сокращенно `F`) и `TRUE` (`T`);
- `character` – символьные объекты (значения переменных задаются в двойных, либо одинарных кавычках).

В отличие от вектора или матрицы, которые могут содержать данные только одного типа, в список (`list`) или таблицу (`data frame`) можно включать сочетания любых типов данных. Это позволяет эффективно, т.е. в одном объекте, хранить разнородную информацию. Каждый компонент списка может являться переменной, вектором, матрицей, фактором или другим списком. Кроме того, эти элементы могут принадлежать к различным типам: числа, строки символов, булевы переменные. Списки являются наиболее общим средством хранения внутрисистемной информации: в частности, результаты большинства статистических анализов в программе R хранятся в объектах - списках.

Таблица данных (`data frame`) представляет собой объект R, по структуре напоминающий лист электронной таблицы Microsoft Excel. Каждый столбец таблицы является вектором, содержащим данные определенного типа. При этом действует правило, согласно которому все столбцы должны иметь одинаковую длину (собственно, с «точки зрения» R таблица данных является частным случаем списка, в котором все компоненты - векторы имеют одинаковый размер). Таблицы данных – это основной класс объектов R, используемых для хранения данных. Обычно такие таблицы подготавливаются при помощи внешних приложений (особенно популярна и удобна программа Microsoft Excel) и затем загружаются в среду R.

2.2.3 Импортирование данных в R

Правила подготовки загружаемых файлов:

- В импортируемой таблице с данными не должно быть пустых ячеек. Если некоторые значения по тем или иным причинам отсутствуют, вместо них следует ввести NA.
- Импортируемую таблицу с данными рекомендуется преобразовать в простой текстовый файл с одним из допустимых расширений. На практике обычно используются файлы с расширением .txt, в которых значения переменных разделены знаками 38 табуляции (tab-delimited files), а также файлы с расширением .csv (comma separated values), в которых значения переменных разделены запятыми или другим разделяющим символом.
- В качестве первой строки в импортируемой таблице рекомендуется ввести заголовки столбцов-переменных. Такая строка – удобный, но не обязательный элемент загружаемого файла. Если она отсутствует, то об этом необходимо сообщить в описании команды, которая будет управлять загрузкой файла (например, `read.table()` – см. ниже). Все последующие строки файла в качестве первого элемента могут содержать заголовки строк (если таковые предусмотрены), после которых следуют значения каждой из имеющихся в таблице переменных.

Имена столбцов таблицы лучше присвоить с соблюдением правил идентификации переменных R, т.е. исключить пробелы и другие специальные символы, кроме точки и подчеркивания. Во избежание проблем, связанных с кодировкой, все текстовые величины в импортируемых файлах стоит создавать с использованием букв латинского алфавита.

Глава 3 Реализация проекта

3.1 Последовательность действий выполнения поставленной задачи

Для обработки и анализа полученных медицинских показателей, была проведена подготовка данных и их предобработка, корреляционный анализ, в ходе которого были выявлены существующие зависимости и произведено сокращение размерности матрицы исходных значений. Выполнен кластерный анализ для выявления однородных групп пациентов, произведена оценка качества кластеризации. И была построена система поддержки принятия решения в виде дерева решений. Все это было выполнено с применением возможностей языка R в среде разработки RStudio. В ходе работы была разработана последовательность действий, по которой и осуществлялся анализ исходных значений. Она изображена в форме схемы на рисунке 5.

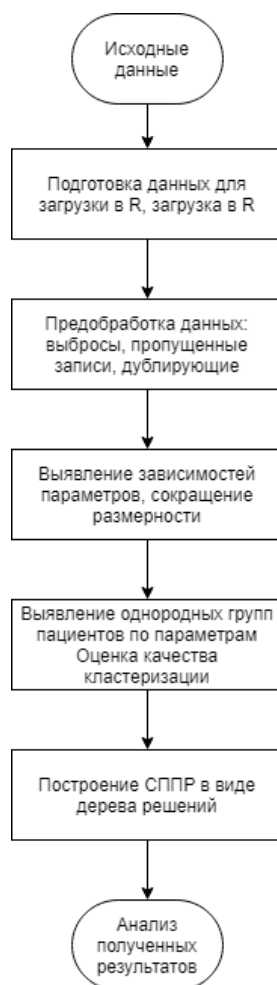


Рисунок 5 – Схема действий выполнения работы

3.2 Первичная обработка данных

Особо важным моментом при анализе данных является подготовка исходных значений к предстоящей работе, так как без данного шага дальнейшая работа может не иметь смысла из-за искаженности полученных результатов. Исходные данные для просмотра находятся в таблице Excel предоставленной томскими врачами.

Исходные данные были проверены на форматирование, так чтобы R мог их хорошо прочитать, загружены в среду разработки R и отредактированы, а именно: все столбцы были приведены к одному типу переменных, были заполнены пропущенные значения методом «*bagImput*», проверены на выбросы.

3.2.1 Анализ выбросов

После загрузки данных в среду R, первым делом необходимо проверить данные на выбросы, т.к. иначе они могут сильно повлиять на конечный результат.

Для наглядности были построены диаграммы размаха для каждого показателя отдельно с помощью функции *boxplot()* реализованной в языке R.

Для примера на рисунках 6 и 7 приведены изображения только двух диаграмм размаха. В Приложении Б находятся диаграммы размаха для данных с выбросами.

```
boxplot(MedDix_1$Rвд1, main='Rвд1')
```

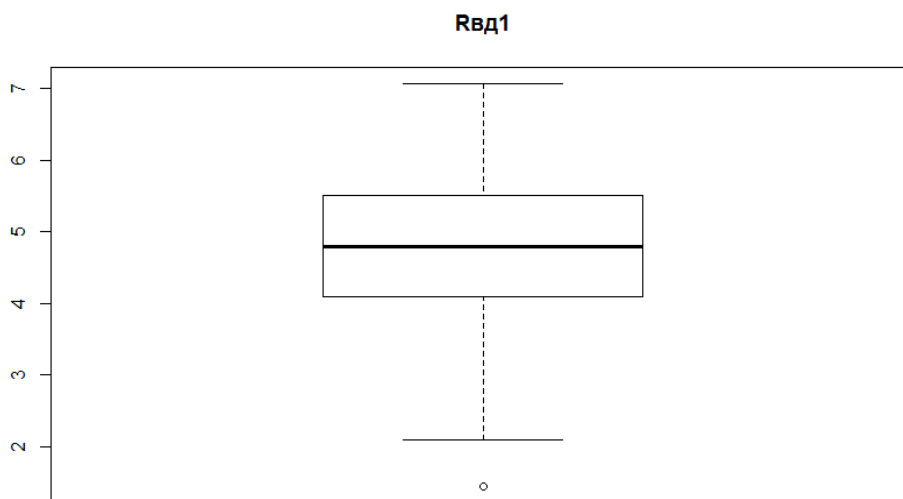


Рисунок 6 – Диаграмма размаха для Rвд1

```
boxplot(MedDix_1$Rвыд1, main='Rвыд1')
```

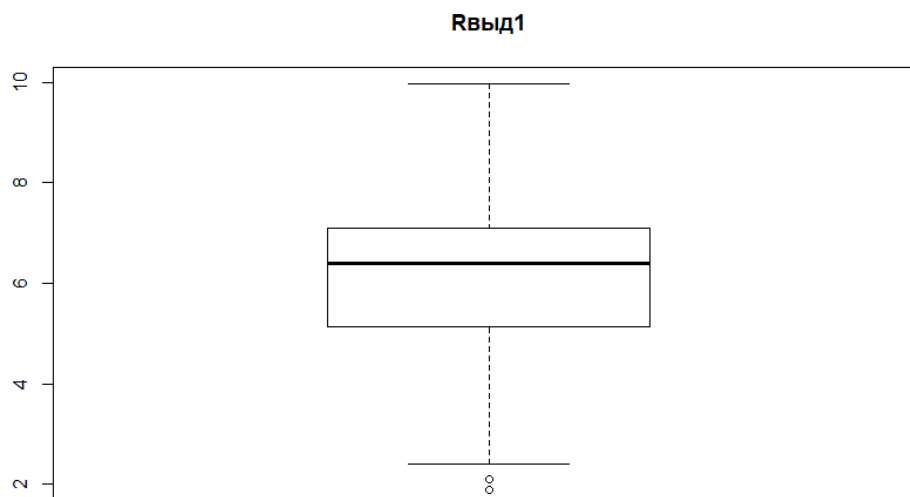


Рисунок 7 – Диаграмма размаха для Rвыд1

Проанализировав полученные диаграммы размаха, мы видим, что выбросы действительно имеются, но их не большое количество и при этом они находятся в допустимых пределах, поэтому данные «аномальные» точки не могут быть удалены.

3.2.2 Восстановление пропущенных значений

Скриптовый язык R командой `summary()` позволяет нам посмотреть общую информацию обо всех числовых столбцах, а именно нас интересует наличие пропущенных значений. После просмотра мы видим, что из всех столбцов 2 столбца имеют пропущенные значения – рисунок 8: «Вес» и «МОС50_2». На рисунке 9 мы можем видеть часть таблицы исходных данных с пропущенными значениями.

	Вес		МОС50_2
Min.	: 34.00	Min.	:13.50
1st Qu.	: 56.50	1st Qu.	:40.30
Median	: 65.00	Median	:54.15
Mean	: 66.03	Mean	:53.04
3rd Qu.	: 75.50	3rd Qu.	:66.42
Max.	:105.00	Max.	:96.50
NA's	:8	NA's	:1

Рисунок 8 – Наличие пропущенных значений

	№ группы	Пол	Возраст	Рост	Вес	МОД1	МОД2
1	1	2	18	168	55	11.50	9.00
2	1	2	48	154	NA	10.30	7.70
3	1	2	31	170	58	8.20	8.50
4	1	2	36	160	62	7.90	7.70
5	1	2	40	167	85	11.40	12.70
6	1	2	48	158	58	10.00	9.60
7	1	1	32	181	66	8.80	8.00
8	1	2	41	164	NA	9.00	6.50
9	1	2	49	156	NA	7.50	7.70
10	1	2	43	156	52	9.32	8.92

Рисунок 9 – Фрагмент с наличием пропущенных значений

Теперь перейдем к восстановлению пропущенных значений в наборе данных. Для того чтобы заполнить пропущенные значения необходимо выполнить следующую последовательность скриптов:

```
> pPbI <- preProcess(MedDix_1[, 5:6], method = 'bagImpute')
> MedDix_1[, 5:6] <- predict(pPbI, MedDix_1[, 5:6])
> pPbI <- preProcess(MedDix_1[, 22:23], method = 'bagImpute')
> MedDix_1[, 22:23] <- predict(pPbI, MedDix_1[, 22:23])
```

Данные скрипты позволяют заполнить пропущенные значения методом «bagImput». Рисунок 10: «Вес» и «МОС50_2» после заполнения. На рисунке 11 мы можем видеть часть таблицы исходных данных уже после заполнения пропущенных значений.

	Вес		МОС50_2
Min.	: 34.00	Min.	:13.50
1st Qu.	: 57.00	1st Qu.	:40.30
Median	: 62.14	Median	:54.10
Mean	: 65.66	Mean	:53.05
3rd Qu.	: 74.00	3rd Qu.	:66.35
Max.	:105.00	Max.	:96.50

Рисунок 10 – Отсутствие пропущенных значений после заполнения

	№ группы	Пол	Возраст	Рост	Вес	МОД1	МОД2
1	1	2	18	168	55.00000	11.50	9.00
2	1	2	48	154	62.13899	10.30	7.70
3	1	2	31	170	58.00000	8.20	8.50
4	1	2	36	160	62.00000	7.90	7.70
5	1	2	40	167	85.00000	11.40	12.70
6	1	2	48	158	58.00000	10.00	9.60
7	1	1	32	181	66.00000	8.80	8.00
8	1	2	41	164	61.25138	9.00	6.50
9	1	2	49	156	59.48977	7.50	7.70
10	1	2	43	156	52.00000	9.32	8.92

Рисунок 11 – Фрагмент с отсутствием пропущенных значений после заполнения

Как мы видим по данным изображениям, пропущенные значения до заполнения отображались символами «NA». После же заполнения появились численные значения, которые соответствуют интервалам допустимых значений во всех столбцах. Из чего мы можем сделать вывод, что заполнение пропущенных значений методом «bagImput» было произведено успешно.

3.3 Сокращение размерности

При выполнении работы было обнаружено, что количество столбцов с различными медицинскими показателями превышает количество строк, где каждая строка является запись об одном пациенте. В данных условиях мы не можем продолжить нашу работу, т.к. методы кластеризации плохо работают с такими типами таблиц. Чтобы решить данную проблему был произведен корреляционный анализ с использованием коэффициента корреляции «Спирмена», для того чтобы выявить зависимости параметров и сократить размерность [46]. Данным способом нам удалось сократить количество столбцов.

Для построения участка матрицы корреляции необходимо выполнить следующую последовательность действий:

```
CR <- cor(MedDix_1_1_Short, method = "spearman")
CR_0_0 <- as.data.frame.matrix(CR)
CR_1_0 <- CR_0_0[1:12, 1:12]
CR_1_0_0 <- as.matrix(CR_1_0)
```

`corrplot(CR_1_0_0, method = "number")`

Далее на рисунках 12, 13, 14 вы можете видеть фрагменты построенной корреляционной матрицы [47]. Вся же корреляционная матрица прикреплена в Приложение В.

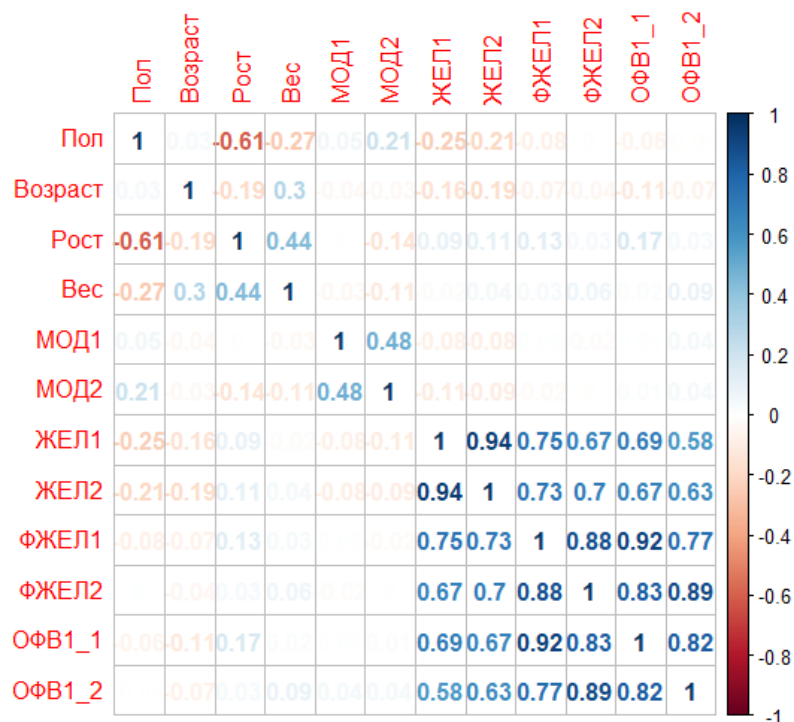


Рисунок 12 – Фрагмент 1 корреляционной матрицы

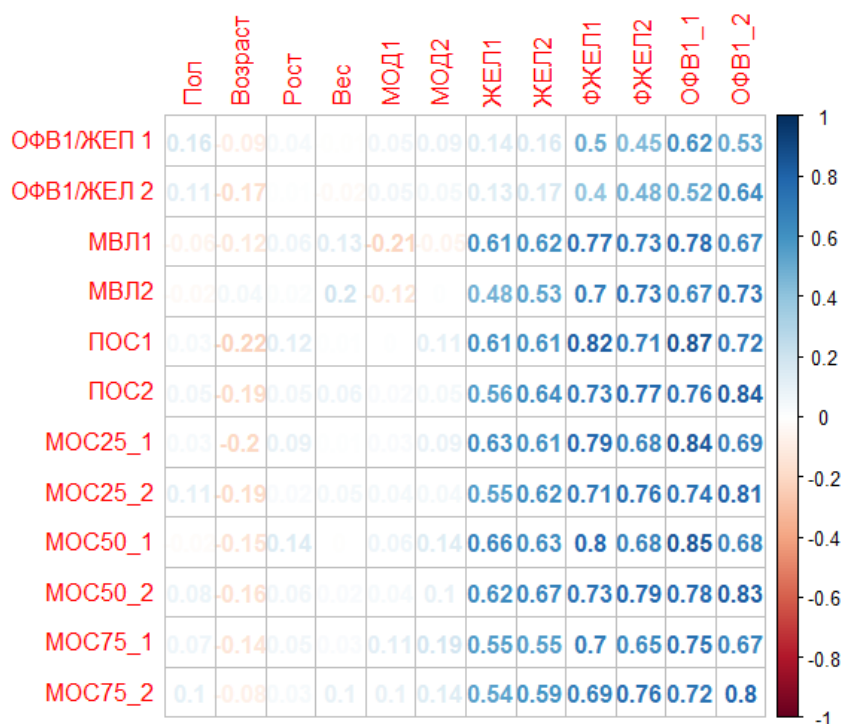


Рисунок 13 – Фрагмент 2 корреляционной матрицы

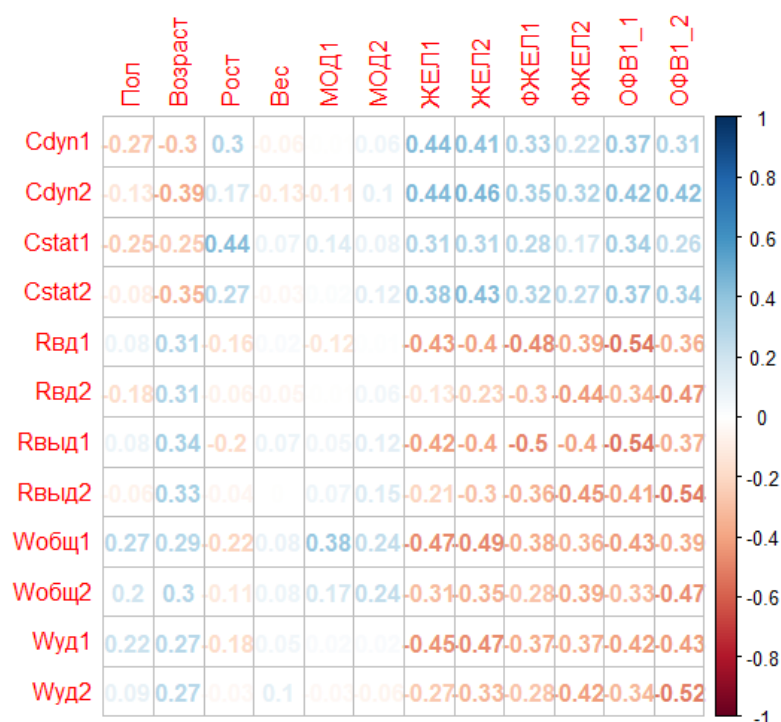


Рисунок 14 – Фрагмент 3 корреляционной матрицы

3.4 Выявление однородных групп пациентов

Т.к. предложенная томскими врачами классификация больных астмой на 4 группы является экспериментальной, то вначале нам необходимо показать, можно ли делить больных на 4 подгруппы. Для этого и был выбран метод кластеризации. [45]

После проведенных до этого манипуляций можно приступить к выполнению кластерного анализа. Выполним кластерный анализ методом *k*-средних для 4 кластеров с помощью функции *kmeans*:

```
> clus <- kmeans(MedDix_1_1, 4)
```

Результат кластеризации представлен на рисунке 15.

```

> clus
K-means clustering with 4 clusters of sizes 23, 23, 24, 13

Cluster means:
# группы  Пол  Возраст  Рост  Вес  МОД1  МОД2  ЖЕЛ1  ЖЕЛ2  ФЖЕЛ1  ФЖЕЛ2  ОФВ1_1  ОФВ1_2  ОФВ1/ЖЕП 1
1 2.130435 1.608696 37.78261 166.3478 63.60556 8.400000 7.197826 67.25261 67.97087 60.17348 65.01826 54.42739 59.67391 69.84739
2 2.434783 1.695652 35.34783 168.4783 64.90250 8.580000 7.957391 86.96739 89.74783 84.50870 88.88696 79.34783 85.93478 87.74174
3 2.125000 1.541667 38.25000 163.4583 66.86490 8.207500 7.589167 83.41042 83.95417 73.67500 78.85833 67.60833 76.36250 75.13125
4 3.000000 1.461538 31.38462 169.8462 69.61564 8.139231 7.326923 96.23077 103.52308 106.28923 114.16923 100.59385 110.87692 95.95385
ОФВ1/ЖЕЛ 2  МВЛ1  МВЛ2  ПОС1  ПОС2  МОС25_1  МОС25_2  МОС50_1  МОС50_2  МОС75_1  МОС75_2  Сdyn1  Сdyn2  Сstat1
1 75.06043 44.25174 45.48087 37.81913 41.15957 35.81478 38.48826 30.19826 31.45174 25.95957 26.56609 0.09047826 0.08730435 0.1325652
2 92.48652 74.55043 80.34087 69.09130 75.00870 66.74348 74.64348 59.79130 65.83478 51.87826 56.98261 0.13334783 0.14234783 0.1635652
3 82.97875 69.73333 73.67500 53.34458 58.46333 50.85417 56.14583 45.12083 48.85427 42.43750 44.54167 0.10954167 0.11162500 0.1358333
4 100.20615 95.31923 105.45385 77.62308 86.43077 74.85615 82.60769 65.37692 76.36308 58.76923 68.54615 0.12046154 0.12738462 0.1520769
Сstat2  Rvд1  Rvд2  Rvвд1  Rvвд2  Wобш1  Wобш2  Wуд1  Wуд2  Wн.эл1  Wн.эл2  Wэл1  Wэл2
1 0.1178261 5.582609 4.443478 7.395652 5.715652 0.6857174 0.4971304 0.08182609 0.06786957 0.6210000 0.4587826 0.3151304 0.2685217
2 0.1696087 3.980000 3.293478 5.344348 4.125217 0.3925217 0.3211739 0.04656522 0.04130435 0.3281304 0.2819130 0.2193913 0.1859130
3 0.1343333 4.880000 4.084167 6.540417 4.947083 0.5783333 0.4665000 0.07037500 0.06283333 0.5105000 0.4190000 0.2725000 0.2513750
4 0.1546923 3.564615 2.837692 4.277692 3.413846 0.4357692 0.3058462 0.05230769 0.04061538 0.3740769 0.2652308 0.2171538 0.1583077
W МОД10_1 W МОД10_2 W МОД15_1 W МОД15_2 W МВлобш1 WМВлобш2 WМВлуд1 WМВлуд2
1 0.8122174 0.6238696 1.3847826 1.1151739 13.52570 14.37026 0.3094783 0.3167826
2 0.4716522 0.3976957 0.7549565 0.6391304 21.45604 24.55304 0.2713478 0.3148261
3 0.6710833 0.5879167 1.1159583 0.9225000 23.11500 26.22325 0.3380000 0.3597917
4 0.5033846 0.3700000 0.9423077 0.6584615 22.44423 25.13462 0.2256154 0.2366923

Clustering vector:
[1] 2 3 2 2 2 1 3 1 4 2 2 1 3 2 1 1 3 4 1 3 3 2 3 2 1 1 3 4 1 3 1 3 1 3 2 4 3 3 1 3 1 1 1 3 3 3 1 3 2 4 1 4 3 3 1 1 3 4 3 3 2 3 1 2 2
[67] 1 2 1 2 2 4 2 2 2 2 2 4 4 4 4 2 4 4 2

within cluster sum of squares by cluster:
[1] 66134.45 48840.25 55609.73 44635.44
(between_SS / total_SS = 60.1 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"

```

Рисунок 15 – Результат работы скрипта

Для более удобного визуального восприятия был построен типичный график для кластерного анализа, он изображен на рисунке 16. [29]

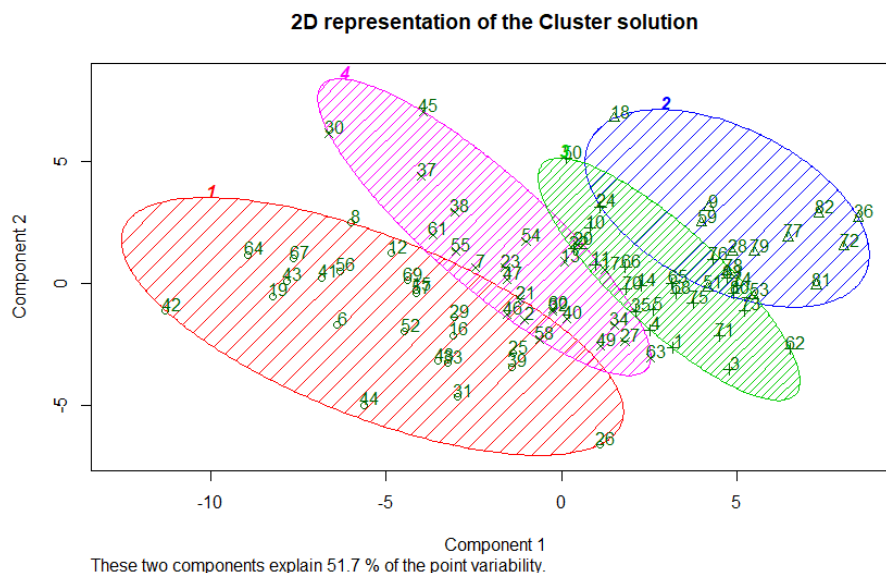


Рисунок 16 – График кластерного анализа

Далее произведем анализ достоверности полученных результатов в таблице 3.

Таблица 3 - Анализ достоверности полученных результатов

Номер кластера	Процент достоверности данных, %
Кластер 1 - БАПИ	88,46
Кластер 2 - БАНП	79,20
Кластер 3 - БАСП	72,22
Кластер 4 - ПО	85,71

Для оценки качества кластеризации использовался индекс «Данна».

[30]

```
> dunn_clust<-kmeans(MedDix_1_1,4)
> Dist<-dist(MedDix_1_1, method='euclidean')
> cluster<-dunn_clust$cluster
> dunn(Dist, cluster)
```

Значение индекса «Данна» составило 0,268213. Это показывает, что данный набор данных достаточно хорошо кластеризуется. [31]

3.5 Построение СППР в виде дерева решений

После того как был произведен кластерный анализ, и показано что можно делить пациентов на 4 подгруппы можно приступить к следующему этапу выполнения задания. Необходимо выявить условия соотнесения пациентов к различным группам и построить систему поддержки принятия решения в виде дерева решений.

Есть несколько основных алгоритмов, которые используются для построения деревьев. В данной работе для создания дерева решений используется функция «`rpart`» из пакета «`rpart`» для языка R [48]. Данная функция делает рекурсивный выбор для каждого следующего узла. Она выбирает такие разделяющие значения, которые приводят к минимальной сумме квадратов внутригрупповых отклонений D_t для всех t узлов дерева.

```
dtm<- rpart(Groups~., MedDix_1_1, method="class")

> dtm
n= 83

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 83 54 3 (0.28915663 0.21686747 0.34939759 0.14457831)
2) Rвд1>=3.15 70 41 3 (0.34285714 0.24285714 0.41428571 0.00000000)
4) Rвд2< 4.935 55 31 1 (0.43636364 0.30909091 0.25454545 0.00000000)
8) МОС50_2>=45.095 35 16 1 (0.54285714 0.11428571 0.34285714 0.00000000)
16) МОС25_1< 60.4 18 4 1 (0.77777778 0.11111111 0.11111111 0.00000000) *
17) МОС25_1>=60.4 17 7 3 (0.29411765 0.11764706 0.58823529 0.00000000) *
9) МОС50_2< 45.095 20 7 2 (0.25000000 0.65000000 0.10000000 0.00000000) *
5) Rвд2>=4.935 15 0 3 (0.00000000 0.00000000 1.00000000 0.00000000) *
3) Rвд1< 3.15 13 1 4 (0.00000000 0.07692308 0.00000000 0.92307692) *
```

Рисунок 17 – Условия соотнесения пациентов

Данной командой мы построили полный набор условий для соотнесения пациентов к одной из четырех групп, который изображен на рисунке 17. Далее построим СППР с использованием полученных условий в виде дерева решений. С помощью функции «`rpr()`» из пакета «`rpart.plot`» [28]. Она изображена на рисунке 18.

prp(dtm)

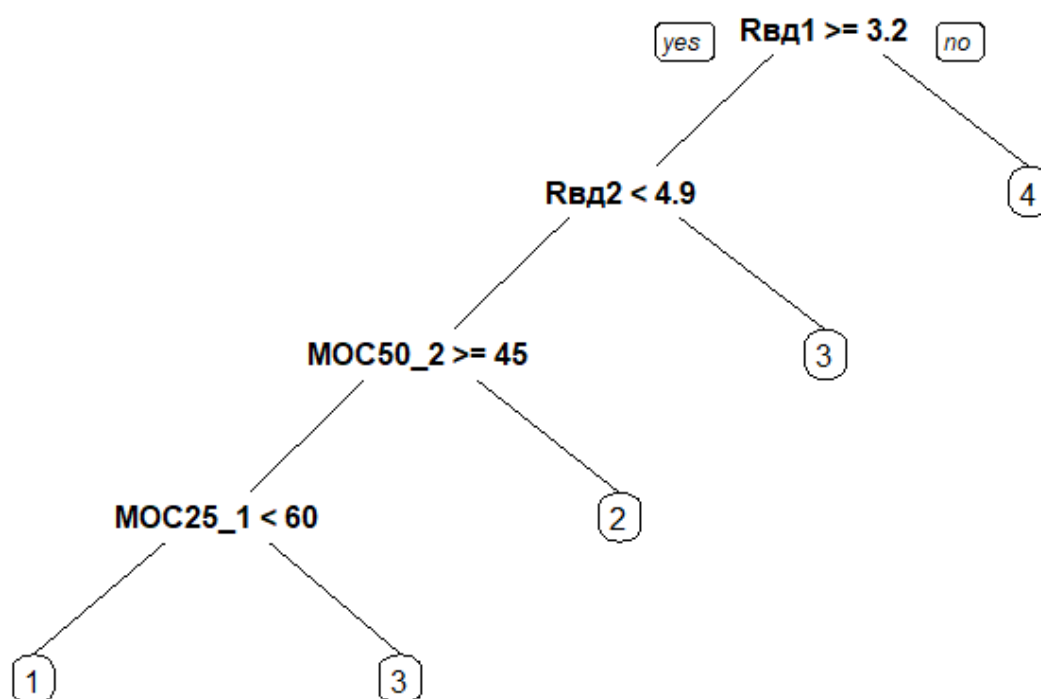


Рисунок 18 – Полученное дерево решений

Полученное дерево дает нам следующие результаты:

1. Если $Rвд1 < 3.2$, то группа 4, т.е $БА = PD$.
2. Если $Rвд1 \geq 3.2$ и $Rвд2 \geq 4.9$, то группа 3, $БА = BANP$.
3. Если $Rвд1 \geq 3.2$ и $Rвд2 < 4.9$ и $МОС50_2 < 45$, то группа 2, $БА = BASP$.
4. Если $Rвд1 \geq 3.2$ и $Rвд2 < 4.9$ и $МОС50_2 \geq 45$ и $МОС25_1 \geq 60$, то группа 3, $БА = BANP$.
5. Если $Rвд1 \geq 3.2$ и $Rвд2 < 4.9$ и $МОС50_2 \geq 45$ и $МОС25_1 < 60$, то группа 1, $БА = BAPI$.

Далее произведем анализ достоверности полученных результатов в таблице 4.

Таблица 4 - Анализ достоверности полученных результатов

Логическое правило	Процент достоверности и данных, %
1. Если $R_{вд1} < 3.2$, то группа 4, т.е. БА=PD.	92,31
2. Если $R_{вд1} \geq 3.2$ и $R_{вд2} \geq 4.9$, то группа 3, БА=BANP.	100
3. Если $R_{вд1} \geq 3.2$ и $R_{вд2} < 4.9$ и $МОС50_2 < 45$, то группа 2, БА = BASP.	74,07
4. Если $R_{вд1} \geq 3.2$ и $R_{вд2} < 4.9$ и $МОС50_2 \geq 45$ и $МОС25_1 \geq 60$, то группа 3, БА = BANP.	73,91
5. Если $R_{вд1} \geq 3.2$ и $R_{вд2} < 4.9$ и $МОС50_2 \geq 45$ и $МОС25_1 < 60$, то группа 1, БА = BAPI.	81,82

Как мы можем видеть из таблицы 4, построенное дерево решений имеет достаточно высокий показатель достоверности данных.

3.6 Анализ полученных результатов

В ходе выполнения задания в рамках разработанного алгоритма действий были совершены следующие действия:

1. подготовка данных;
2. предобработка данных;
3. выявлены зависимости параметров и сокращена размерность путем корреляционного анализа;
4. проведен кластерный анализ;
5. произведено построение дерева решений.

При кластерном анализе данные были поделены на 4 кластера, что соответствует 4-м заявленным группам бронхиальной астмы. И было построено дерево решений для соотнесения каждого пациента к одной из четырех групп. Так как достоверность результатов в среднем выше 80%, то можно сделать вывод, что предложенная томскими медиками классификации бронхиальной астмы имеет право на существование.

Глава 4 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Введение

В нижеприведенном разделе дипломного проекта дается оценка экономического и инвестиционного эффекта от применения потребителями разработанного в магистерском дипломном проекте программного продукта.

Целью представленной магистерской диссертации является создание системы поддержки научных исследований медицинской диагностики. Для достижения данной цели необходимо было выполнить ряд задач, в том числе дать оценку коммерческих возможностей проведенного исследования, выявить его ресурсосберегающий потенциал, определить финансовую эффективность исследования и его инвестиционный потенциал. Составить график времени проведения данного исследования и финансовые затраты на его реализацию.

Объектом нашего исследования выступает набор данных, а именно, медицинские показатели пациентов с бронхиальной астмой. В работе рассматривается создание системы поддержки научных исследований медицинской диагностики на основе полученных медицинских данных. Данное исследование призвано создать программный продукт, который будет помогать врачам и исследователям, связанным с областью исследования и лечения бронхиальной астмы.

В данном разделе: выявлены потенциальные потребители нашего продукта исследования, дан анализ конкурентных технических решений, измерены характеристики качества этой разработки и ее перспективность использования в работе, произведен SWOT-анализ, описаны и проанализированы финансово-экономические аспекты выполненной работы. Произведена оценка денежных затрат на исследование. В заключение дана комплексная оценка научно-техническому уровню работы.

4.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

4.1.1 Потенциальные потребители результатов исследования

Для анализа потребителей результатов исследования необходимо рассмотреть целевой рынок и провести его сегментирование.

Целевой рынок – сегменты рынка, на котором будет продаваться в будущем разработка. В свою очередь, сегмент рынка – это особым образом выделенная часть рынка, группы потребителей, обладающих определенными общими признаками. Целевой рынок проекта – медицинские учреждения.

Главными потребителями нашего продукта могут быть врачи и исследователи, которые занимаются вопросами лечения и изучения бронхиальной астмы, а также различные исследовательские группы и центры, занимающиеся данным вопросом.

Сегментирование – это разделение покупателей на однородные группы, для каждой из которых может потребоваться определенный товар (услуга).

Выделим наиболее значимые для разрабатываемой информационной системы критерии сегментации рынка потребителей: область применения и масштабность применения предлагаемого продукта. На основе выявленных данных построим сегментирование рынка, оно представлено в таблице 5.

Таблица 5 – Сегментирование рынка

		Область применения	
		Психология	Медицина
Масштабность применения	В пределах страны	Долгосрочная перспектива	
	В пределах региона	Краткосрочная перспектива	Долгосрочная перспектива
	В пределах города	Краткосрочная перспектива	

4.1.2 Анализ конкурентных технических решений

Проведем сравнительный анализ конкурентных технических решений, существующих на рынке информационных систем.

Конкурент номер один – информационная система «Здоровье». ИС предназначена для оценки уровня функциональных резервов и формирования мотивации здорового образа жизни. Оценка состояния здоровья проводится в шкале «норма - донозологическое состояние - преморбидное состояние - клинически значимые отклонения». Визуализация результатов производится цветами соответственно состоянию: «зеленый - голубой - желтый - красный».

Конкурент номер два – информационная система «Анализ данных». Данная информационная система предназначена для анализа вводимой информации и составления вывода по полученным результатам.

Анализируемые данные представлены в таблице 6.

Таблица 6 – Оценочная карта для сравнения конкурентных технических решений
(разработок)

Критерии оценки	Вес критерия	Баллы			Конкурентоспособность		
		Б _ф	Б _{к1}	Б _{к2}	К _ф	К _{к1}	К _{к2}
1	2	3	4	5	6	7	8
Технические критерии оценки ресурсоэффективности							
1. Удобство в заполнении методик	0,1	5	4	5	0,5	0,4	0,5
2. Продолжительность обработки массива данных	0,2	5	3	4	1	0,6	0,8
3. Качество полученных результатов	0,1	3	5	4	0,3	0,5	0,4
4. Возможности технологии	0,05	4	4	5	0,2	0,2	0,25
5. Простота алгоритма	0,1	5	5	5	0,5	0,5	0,5
6. Пути внедрения	0,05	4	4	5	0,2	0,2	0,25
Экономические критерии оценки эффективности							
1. Конкурентоспособность технологии	0,1	4	5	3	0,4	0,5	0,3
2. Цена технологии	0,15	5	5	4	0,75	0,75	0,6
3. Срок внедрения	0,1	3	3	2	0,3	0,3	0,2
4. Наличие сертификации разработки	0,05	4	5	5	0,2	0,25	0,25
Итого	1				4,35	4,2	4,05

Анализ конкурентных технических решений определяется по формуле:

$$K = \sum B_i * B_i,$$

где K – конкурентоспособность научной разработки или конкурента;

B_i – вес показателя (в долях единицы);

B_i – балл i -го показателя.

Для оценки ресурсоэффективности были выбраны следующие критерии: удобство в заполнении методик, продолжительность обработки массива данных, качество полученных результатов, возможности технологии, простота алгоритма, пути внедрения технологии.

Наиболее значимым критерием является продолжительность обработки массива данных. Для оценки эффективности были выбраны следующие экономические критерии: конкурентоспособность технологии, цена, срок внедрения и наличие сертификации разработки.

Результаты анализа выявили, что разработанная нами технология выгодно отличается от конкурентов благодаря высокой скорости обработки данных и удобством заполнения соответствующих таблиц.

4.1.3 Технология QuaD

Данная технология использована для анализа проекта, чтобы измерить характеристики качества этой разработки и ее перспективность использования в работе.

В основе технологии QuaD лежит нахождение средневзвешенной величины следующих групп показателей: показателей оценки коммерческого потенциала разработки и показателей оценки качества разработки.

Анализ проводится в виде оценочной карты (таблица 7).

Таблица 7 – Оценочная карта QuaD

Критерии оценки	Вес критерия	Баллы	Максимальный балл	Относительное значение	Средневзвешенное значение
1	2	3	4	5	6
Показатели оценки качества разработки					
1. Удобство в заполнении методик	0,1	70	100	0,7	0,07
2. Продолжительность обработки массива данных	0,2	90	100	0,9	0,18
3. Качество полученных результатов	0,1	50	100	0,5	0,05
4. Возможности технологии	0,05	80	100	0,8	0,04
5. Простота алгоритма	0,1	100	100	1	0,1
6. Пути внедрения	0,05	80	100	0,8	0,04
Показатели оценки коммерческого потенциала разработки					
1. Конкурентоспособность технологии	0,1	90	100	0,9	0,09
2. Цена технологии	0,15	95	100	0,95	0,1425
3. Срок внедрения	0,1	90	100	0,9	0,09
4. Наличие сертификации разработки	0,05	100	100	1	0,05
Итого:	1				0,8525

Оценка качества и перспективности по технологии QuaD определяется по формуле:

$$P_{cp} = \sum B_i \cdot B_i ,$$

где P_{cp} – средневзвешенное значение показателя качества и перспективности научной разработки;

B_i – вес показателя (в долях единицы);

B_i – средневзвешенное значение i -го показателя.

Получаем, что $P_{cp} = 0,8525$.

Значение P_{cp} позволяет говорить о перспективах разработки и что в нее стоит инвестировать.

4.1.4 SWOT-анализ

SWOT-анализ является одним из самых распространенных методов анализа и стратегического планирования в менеджменте и маркетинге. SWOT-анализ применяют для исследования внешней и внутренней среды проекта и осуществляется в несколько этапов.

Первый этап заключается в описании сильных и слабых сторон проекта, определении его возможностей и угроз.

Второй этап состоит в выявлении соответствия сильных и слабых сторон научно-исследовательского проекта внешним условиям окружающей среды. Это соответствие или несоответствие должны помочь выявить степень необходимости проведения стратегических изменений.

В рамках третьего этапа должна быть составлена итоговая матрица SWOT-анализа. Итоговые результаты представлены в таблице 8.

Таблица 8 – SWOT-анализ

	Сильные стороны: С1. Простота алгоритма, заложенного в основе технологии. С2. Исключение ошибок связанных с человеческим фактором. С3. Низкие цены на технологию. С4. Упрощение работы с медицинскими данными.	Слабые стороны: Сл.1 Неквалифицированные пользователи. Сл.2 Высокая скорость развития конкурентных технологий.
Возможности: В1. Расширение географии внедрения технологии. В2. Увеличение скорости проведения анализа медицинских данных. В3. Повышение квалификации сотрудников.	1. Внедрение технологии на территории всей страны за счет низкой стоимости. 2. Хранить медицинские данные в электронном виде в необходимом для системы формате. 3. Быстрый и точный сбор статистики приведет к исключению работы людей с многочисленными данными и позволит получить информативные данные.	1. Разработать систему обучения пользователей. 2. Своевременное развитие технологий в соответствии с возникающими новшествами.

Угрозы: У1. Ужесточение конкуренции. У2. Выход на рынок новых конкурентов. У3. Сбои в работе системы.	1. Поддерживать существующую ценовую политику для повышения конкурентоспособности. 2. Разработать систему скидок для многоцелевого использования технологии. 3. Сбои в работе можно исключить за счет грамотного выбора разработчика и предоставления лучших инструментов для работы.	1. Снижать издержки и оптимизировать ресурсы. 2. Стараться развивать технологию быстрее конкурентов.
--	--	--

Таким образом, в результате *SWOT*-анализа были рассмотрены сильные и слабые стороны разработки информационной системы, выявлены возможные перспективы ее создания и рассмотрены варианты минимизации влияния угроз, которые могут этому помешать.

Наиболее сильными сторонами являются: простота алгоритма, заложенного в основе технологии и низкие цены на технологию. Слабости проекта обнаруживаются в том, что для работы с технологией необходимо быть квалифицированных в области работы с использованной технологией, т.е. есть вероятность неквалифицированных пользователей, а также высокая скорость развития конкурентных технологий. Для того чтобы уменьшить влияние слабых сторон на наш продукт необходимо разработать систему обучения пользователей и поддерживать активное своевременное развитие технологии в соответствии с возникающими новшествами.

4.2 Планирование научно-исследовательских работ

4.2.1 Структура работ в рамках научного исследования

Планирование комплекса предполагаемых работ осуществляется в следующем порядке:

- определение структуры работ в рамках научного исследования;
- определение участников каждой работы;
- установление продолжительности работ;
- построение графика проведения проектной работы.

Для выполнения научного исследования сформирована рабочая группа, в состав которой входят студент и преподаватель – руководитель проекта. По каждому виду запланированных работ установлена соответствующая должность исполнителей, эти данные представлены в таблице 9.

Таблица 9 - Перечень этапов, работ и распределение исполнителей

Основные этапы	№ раб.	Содержание работ	Должность исполнителя
Подготовительный этап	1	Продумать идею и записать основные данные	Руководитель, студент
	2	Получение технического задания	Руководитель, студент
	3	Подбор материала, его анализ и обобщение	Студент
	4	Формирование возможных решений поставленной задачи, выбор оптимального решения	Студент
Основной этап	5	Программная реализация придуманной идеи	Студент
	6	Тестирование реализованной программы	Студент
	7	Исправление ошибок, доработка программы	Студент
	8	Проведение исследований, получение результатов, их анализ	Студент
Заключительный этап	9	Составление отчета о проделанной работе	Студент
	10	Оценка эффективности полученных результатов	Студент
	11	Защита ВКР	Студент

4.2.2 Определение трудоемкости выполнения работ

Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от многих факторов. Для определения ожидаемого (среднего) значения трудоемкости $t_{ожi}$ используется формула:

$$t_{\text{ож}i} = \frac{3t_{\text{мин}i} + 2t_{\text{макс}i}}{5},$$

где $t_{\text{ож}i}$ – ожидаемая трудоемкость выполнения i -ой работы чел.-дн.;

$t_{\text{мин}i}$ – минимально возможная трудоемкость выполнения заданной i -ой работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), чел.-дн.;

$t_{\text{макс}i}$ – максимально возможная трудоемкость выполнения заданной i -ой работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), чел.-дн.

Исходя из ожидаемой трудоемкости работ, определяется продолжительность каждой работы в рабочих днях T_{pi} , учитывающая параллельность выполнения работ несколькими исполнителями. Такое вычисление необходимо для обоснованного расчета заработной платы, так как удельный вес зарплаты в общей сметной стоимости научных исследований составляет около 65 %. Продолжительно одной работы определяется по формуле:

$$T_{pi} = \frac{t_{\text{ож}i}}{Ч_i},$$

где T_{pi} – продолжительность одной работы, раб. дн.;

$t_{\text{ож}i}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.

$Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

Составим таблицу 10 и подсчитаем значения $t_{\text{ож}i}$ и T_{pi} для каждой из работ.

Таблица 10 – Перечень работ и подсчет показателей

Перечень	t min	t max	t ожид	Ч	Tpi
Продумать идею и записать основные данные	2	5	3,2	2	1,6
Получение технического задания	1	2	1,4	2	0,7
Подбор материала, его анализ и обобщение	12	15	13,2	1	13,2
Формирование возможных решений поставленной задачи,	12	15	13,2	1	13,2

выбор оптимального решения					
Программная реализация придуманной идеи	15	17	15,8	1	15,8
Тестирование реализованной программы	2	5	3,2	1	3,2
Исправление ошибок, доработка программы	5	7	5,8	1	5,8
Проведение исследований, получение результатов, их анализ	1	2	1,4	1	1,4
Составление отчета о проделанной работе	10	15	12	1	12
Оценка эффективности полученных результатов	3	5	3,8	1	3,8
Защита ВКР	5	8	6,2	1	6,2

4.2.3 Разработка графика проведения научного исследования

Составим график проведения научных работ в форме диаграммы Ганта.

Диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Для удобства построения графика, длительность каждого из этапов работ из рабочих дней переведем в календарные дни. Для этого воспользуемся формулой:

$$T_{ki} = T_{pi} \cdot k_{\text{кал}},$$

где T_{ki} – продолжительность выполнения i -й работы в календарных днях;

T_{pi} – продолжительность выполнения i -й работы в рабочих днях;

$k_{\text{кал}}$ – коэффициент календарности.

Коэффициент календарности определяется по формуле:

$$k_{\text{кал}} = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{вых}} - T_{\text{пр}}},$$

где $T_{\text{кал}}$ – количество календарных дней в году;

$T_{\text{вых}}$ – количество выходных дней в году;

$T_{\text{пр}}$ – количество праздничных дней в году.

$$k_{\text{кал}} = 365 / (365 - 125) = 1,5.$$

Рассчитанные значения в календарных днях по каждой работе T_{ki} округлим до целого числа.

Все рассчитанные значения сведены в таблицу 11.

Таблица 11 – Расчет продолжительности работ в календарных днях

Перечень	T_{pi}	T_{ki}
Продумать идею и записать основные данные	1,6	3
Получение технического задания	0,7	1
Подбор материала, его анализ и обобщение	13,2	20
Формирование возможных решений поставленной задачи, выбор оптимального решения	13,2	20
Программная реализация придуманной идеи	15,8	24
Тестирование реализованной программы	3,2	5
Исправление ошибок, доработка программы	5,8	9
Проведение исследований, получение результатов, их анализ	1,4	3
Составление отчета о проделанной работе	12	18
Оценка эффективности полученных результатов	3,8	6
Защита ВКР	6,2	10

По данным расчетам, программа будет разработана за 119 дней:

На основе таблицы 11 построен календарный план-график, представленный в таблице 12. Работы на графике выделены различной штриховкой в зависимости от исполнителей, ответственных за ту или иную работу.

Таблица 12 – Диаграмма Ганта

№	Название работы	Исполнители	T_{ki} , кал. дн.	Продолжительность выполнения работ																	
				январь			февраль			март			апрель			май			июнь		
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	Продумать идею и записать основные данные	Студент Руководитель	3																		
2	Получение технического задания	Студент Руководитель	1																		
3	Подбор материала, его анализ и обобщение	Студент	20																		
4	Формирование возможных решений поставленной задачи, выбор оптимального решения	Студент	20																		
5	Программная реализация придуманной идеи	Студент	24																		
6	Тестирование реализованной программы	Студент	5																		
7	Исправление ошибок, доработка программы	Студент	9																		
8	Проведение исследований, получение результатов, их анализ	Студент	3																		
9	Составление отчета о проделанной работе	Студент	18																		
10	Оценка эффективности полученных результатов	Студент Руководитель	6																		
11	Защита ВКР	Студент	5																		
	-студент	-руководитель																			

4.3 Бюджет научно-технического исследования (НТИ)

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

- материальные затраты НТИ;
- расчет амортизации;
- основная заработная плата исполнителей темы;
- дополнительная заработная плата исполнителей темы;
- отчисления во внебюджетные фонды (страховые отчисления);
- накладные расходы.

4.3.1 Расчет материальных затрат НТИ

Расчет материальных затрат осуществляется по формуле:

$$З_m = (1 + k_T) \cdot \sum_{i=1}^m C_i \cdot N_{расхi},$$

где m – количество видов материальных ресурсов, потребляемых при выполнении научного исследования;

$N_{расхi}$ – количество материальных ресурсов i -го вида, планируемых к использованию при выполнении научного исследования (шт., кг, м, м² и т.д.);

C_i – цена приобретения единицы i -го вида потребляемых материальных ресурсов (руб./шт., руб./кг, руб./м, руб./м² и т.д.);

k_T – коэффициент, учитывающий транспортно-заготовительные расходы.

Для разработки данного продукта необходимы следующие материальные ресурсы:

- Системный блок;
- Принтер;
- Монитор;
- Бумага для принтера;
- Различные канцелярские принадлежности.

При этом орг. технику, оборудование и электроэнергию предоставляет университет.

Расчет материальных затрат представлен в таблице 13.

Таблица 13 - Материальные затраты

Наименование	Ед. изм.	Количество	Цена за ед., руб.	Затраты на материалы, (З _м), руб.
Бумага	шт.	3	580	1740
Канцелярские предметы				1500
Итого:				3240

4.3.2 Основная заработная плата исполнителей темы

Основная заработная плата рассчитывается по формуле:

$$C_{осн/зн} = \sum_{i=1}^n t_i \cdot C_{зн_i},$$

где n – количество видов работ;

t_i – затраты труда на выполнение i -го вида работ, в днях;

$C_{зн_i}$ – среднедневная заработная плата работника, выполняющего i -ый вид работы, руб./день.

Среднедневная заработная плата рассчитывается по формуле:

$$C_{зн_i} = \frac{D \cdot K \cdot M_p}{F_0},$$

где D – месячный должностной оклад работника;

K – коэффициент, учитывающий коэффициент по премиям и районный коэффициент ($K=1,3$);

M_p – количество месяцев работы без отпуска в течение года;

F_0 – действительный годовой фонд рабочего времени работника, в днях.

При отпуске 28 дня $M_p=11,08$.

Результаты расчета действительного годового фонда проведены в таблице 14.

Таблица 14 - Годовой фонд рабочего времени

Показатели рабочего времени, дни	Исполнение
Календарное число дней в году	365
Количество нерабочих дней	104
Выходные	13
Праздники (фактически по каждому году)	
Планируемые потери отпуска	28
Действительный годовой фонд	220

Расчет затрат на основную заработную плату приведен в таблице 15. При этом затраты на оплату труда студента-дипломника определяются как оклад студента ($D = 2477,8$), а оклад руководителя проекта составляет 33664. Коэффициент K , учитывающий районный коэффициент равный 1,3.

Таблица 15 – Затраты на основную заработную плату

Исполнители	Среднедневная заработная плата $C_{зн}$ (руб.)	Трудоемкость (t_i), чел-дни	Затраты на основную зарплату (руб.)
Руководитель	1989,24	10,6	21087,94
Студент	125	76,9	9612,5
Итого			30700,44

4.3.3 Дополнительная заработная плата

Дополнительная заработная плата включает заработную плату за не отработанное рабочее время, но гарантированную действующим законодательством.

Расчет дополнительной заработной платы ведется по формуле:

$$Z_{доп} = k_{доп} \cdot Z_{осн},$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (на стадии проектирования принимается равным 0,12 – 0,15).

$k_{доп}$ равен 0,12. Результаты по расчетам дополнительной заработной платы сведены в таблицу 16.

Таблица 16 - Затраты на дополнительную заработную плату

Исполнители	Основная зарплата(руб.)	Коэффициент дополнительной заработной платы ($k_{\text{доп}}$)	Дополнительная зарплата(руб.)
Руководитель	21087,94	0,12	2530,55
Студент	9612,5	0,12	1153,5
Итого			3684,05

4.3.4 Отчисления во внебюджетные фонды (страховые отчисления)

Величина отчислений во внебюджетные фонды определяется исходя из формулы:

$$З_{\text{внеб}} = k_{\text{внеб}} \cdot (З_{\text{осн}} + З_{\text{доп}}),$$

где $k_{\text{внеб}}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.).

На основании пункта 1 ст.58 закона №212-ФЗ для учреждений осуществляющих образовательную и научную деятельность в 2014 году вводится пониженная ставка страховых взносов – 27,1%.

Отчисления во внебюджетные фонды представлены в таблице 17.

Таблица 17 - Отчисления во внебюджетные фонды

Исполнитель	Основная заработная плата, руб.	Дополнительная заработная плата, руб.
Руководитель проекта	21087,94	2530,55
Студент	9612,5	1153,5
Коэффициент отчислений во внебюджетные фонды	27,1%	
Итого		
Исполнение	9318,2	

4.3.5 Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые и телеграфные расходы, размножение материалов и т.д. Их величина определяется по формуле:

$$З_{\text{накл}} = (\text{сумма статей } 1 \div 4) \cdot k_{\text{нр}},$$

Где $k_{нр}$ – коэффициент, учитывающий накладные расходы.

Величина коэффициента накладных расходов равна 16%.

Исполнение: $(3240+30700,44+3684,05+9318,2)*0,16=7510,83$

4.3.6 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанная величина затрат научно-исследовательской работы (темы) является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

Определение бюджета затрат на научно-исследовательский проект приведен в таблице 18.

Таблица 18 - Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.
1. Материальные затраты НТИ	3240
2. Затраты по основной заработной плате исполнителей темы	30700,44
3. Затраты по дополнительной заработной плате исполнителей темы	3684,05
4. Отчисления во внебюджетные фонды	9318,2
5. Накладные расходы	7510,83
6. Бюджет затрат НТИ	54453,52

Вывод: Основываясь на данных, полученных в пунктах 4.3.1 – 4.3.5, был рассчитан бюджет затрат научно-исследовательской работы.

4.4 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный финансовый показатель разработки определяется по формуле:

$$I_{\text{финр}} = \frac{\Phi_p}{\Phi_{\text{max}}},$$

где $I_{\text{финр}}$ – интегральный финансовый показатель разработки;

Φ_p – стоимость исполнения;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Расчет:

$$I_{\text{финр}} = \frac{54453,52}{80000} = 0,68$$

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в размах (значение больше единицы), либо соответствующее численное удешевление стоимости разработки в размах (значение меньше единицы, но больше нуля).

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное удешевление стоимости разработки в размах.

Интегральный показатель ресурсоэффективности исполнения объекта исследования можно определить по формуле:

$$I_p = \sum a \cdot b,$$

где I_p – интегральный показатель ресурсоэффективности;

a – весовой коэффициент;

b – бальная оценка, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

Расчет интегрального показателя ресурсоэффективности приведен в таблице 19.

Таблица 19 - Сравнительная оценка характеристик варианта исполнения проекта

Объект исследования Критерии	Весовой коэффициент параметра	Оценка выполнения
1. Способствует росту производительности труда пользователя	0,15	4
2. Удобство в эксплуатации (соответствует требованиям потребителей)	0,25	3
3. Экономия времени	0,2	5
4. Простота исполнения	0,3	5
5. Надежность	0,1	4
ИТОГО:	1	

$$I_p = 4 \cdot 0,15 + 3 \cdot 0,25 + 5 \cdot 0,2 + 5 \cdot 0,3 + 4 \cdot 0,1 = 4,25;$$

Исходя из вычислений, показатель ресурсоэффективности выполнения имеет достаточно высокое значение (по 5-бальной шкале), что говорит об эффективности использования технического проекта.

Интегральный показатель эффективности вариантов исполнения разработки ($I_{\text{исп.}}$)

Определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{\text{исп}} = \frac{I_p}{I_{\text{финр}}} = \frac{4,25}{0,68} = 6,25,$$

Полученное значение интегрального показателя эффективности исполнения разработки превысил максимальный балл в системе оценивания. Это говорит о том, что результат нашей работы можно считать положительным, так как оценка интегрального показателя ресурсоэффективности близка к максимальной.

Выводы по разделу:

В ходе работы был определен целевой рынок результатов проекта, построена карта его сегментирования. Дана оценка конкурентных систем.

Произведена оценка качества и перспективности предлагаемого продукта по технологии *QuaD*, где было выявлено, что обсуждаемая

разработка является перспективной.

Произведен *SWOT*-анализ, в ходе, которого были рассмотрены сильные и слабые стороны разработки медицинской системы, выявлены возможные перспективы ее создания и рассмотрены варианты минимизации влияния слабых сторон, которые могут этому помешать.

В рамках анализа определена трудоемкость данного проекта, она равна 119 дням. На основании трудоемкости была построена диаграмма Ганта.

Так же были рассчитаны величины затрат научно-исследовательских работ. В результате проведенных расчетов, бюджет затрат НИИ составил 54453,52.

После проведенных сравнений и расчетов мы можем сделать вывод, что разработка данной медицинской системы является перспективной и привлекательной для инвесторов.

Глава 5 Социальная ответственность

Введение

Направление данной магистерской диссертации связано с разработкой программного обеспечения на языке программирования R для анализа набора данных. Суть проекта по ВКР состоит в том, что нам были предоставлены медицинские данные по больным бронхиальной астмой в форме таблиц, с целью их обработки, анализа и выявления необходимой информации. Областью применения данной работы это дальнейшие более глубокие медицинские исследования, а также практическое применения полученных результатов в медицине. Потенциальными пользователями являются научные сотрудники в области медицинских исследований, а также врачи.

Работа научного сотрудника связана во взаимодействии с персональным компьютером. Рабочий зоной для разработки технологии, описанной в данной работе, является персональное место за компьютером в 204 аудитории Кибернетического центра Национального исследовательского Томского политехнического университета расположенного в городе Томск.

При этом данная работа является актуальной и значимой, т.к. является исследованием в области бронхиальной астмы, а данное заболевание является одним из самых распространённых хронических заболеваний на текущий момент.

5.1 Правовые и организационные вопросы обеспечения безопасности

Для обеспечения безопасности при работе определяют следующие требования к организации рабочих мест пользователей:

- рабочее место должно быть организовано с учетом эргономических требований согласно ГОСТ 12.2.032-78 [61] и ГОСТ 12.2.061-81 [62];
- конструкция рабочей мебели (компьютерный стол, офисное кресло, подставка для ног) должна предусматривать возможность регулировки в соответствии с индивидуальными особенностями

пользователя для создания комфортных условий для выполнения работы. Вокруг ПК должно быть обеспечено свободное пространство в радиусе как минимум 60-120см;

- оригинал-держатель должен быть установлен на уровне экрана.

В соответствии с государственными стандартами и правовыми нормами обеспечения безопасности предусмотрена рациональная организация труда в течение смены, которая предусматривает:

- продолжительность рабочей смены, не превышающей 8 часов;
- длительность обеденного перерыва не меньше 40 минут;
- установление двух регламентируемых перерывов (не меньше 20 минут после 1-2 часов работы и не меньше 30 минут после 2 часов работы).

Обязательно должен быть предусмотрен предварительный медосмотр, который осуществляется при приеме на работу, и периодические медосмотры.

Также перед приемом на работу каждый сотрудник должен пройти инструктаж по технике безопасности, а в дальнейшем с работником должен быть проведен инструктаж по электробезопасности и охране труда.

Рабочей зоной является помещение, ширина которого – 6 м, длина – 6 м, высота – 2,8 м, оборудованное 10 рабочими местами, но одновременно в работе задействованы 3-4 человека. Каждое рабочее место включает в себя персональный компьютер, расположенный на персональном столе и офисный стул. Площадь аудитории составляет 36 м^2 , объём помещения - $100,8 \text{ м}^3$. В аудитории имеется 2 окна оборудованных жалюзи, поэтому в дневное время можно использовать естественное освещение, в остальное время используется равномерное искусственное освещение. Имеется кондиционер, а также естественная вентиляция: двери, окна. Стены и потолок имеют светлые оттенки, пол застелен линолеумом светло-серого цвета.

Следовательно, в среднем на одного сотрудника приходится не менее 25 м^3 объема помещения и не менее 9 м^2 площади, что удовлетворяет

требованиям санитарных норм, согласно которым для одного работника должны быть предусмотрены площадь величиной не менее 6 м² и объем не менее 24 м³, с учетом максимального числа одновременно работающих в смену.

5.2 Производственная безопасность

Так как на состояние офисных работников (программистов) химические и биологические факторы не оказывают существенного влияния, то основное внимание будет уделено физическим и психофизиологическим факторам. Для представления всех вредных и опасных факторов необходимо классифицировать их в соответствии с нормативными документами. Классификация представлена в таблице 20.

Таблица 20 – Классификация вредных и опасных факторов

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разрабо тка	Изготов ление	Эксплу тация	
1.Отклонение показателей микроклимата	+	+	+	1.ГОСТ 12.1.005-88 ССБТ. Общие санитарно-гигиенические требования к воздуху рабочей зоны. [65] 2.СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату производственных помещений. [69]
2.Недостаточная освещенность рабочей зоны		+	+	3.СанПиН 2.2.1/2.1.1.1278–03. Гигиенические требования к естественному, искусственному и совмещённому освещению жилых и общественных зданий. [67] 4.СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*[70]
3.Превышение уровня шума	+	+	+	5.ГОСТ 12.1.003-2014 ССБТ. Шум. Общие требования безопасности. [64]
4.Опасные и вредные производственные факторы, связанные с электромагнитными полями	+	+	+	6.ГОСТ 12.1.006-84 ССБТ. Электромагнитные поля радиочастот. Общие требования безопасности. [66] 7.СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах. [72] 8.СанПиН 2.2.2/2.4.1340–03. Санитарно-

				эпидемиологические правила и нормативы «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы». [68]
5.Психофизиологические факторы	+	+		9.ТОИ Р-45-084-01. Типовая инструкция по охране труда при работе на персональном компьютере" (утв. Приказом Минсвязи РФ от 02.07.2001 N 162) [74]
6.Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	+	+	+	10.ГОСТ 12.1.038–82 Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов. [73]

5.2.1 Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения

5.2.1.1 Отклонение показателей микроклимата

Гигиенические нормативы на параметры микроклимата в рабочей зоне даны в ГОСТ 12.1.005 – 88. Микроклимат в рабочей зоне определяется действующими на организм человека сочетаниями влажности, температуры воздуха и окружающих поверхностей и скорости движения воздуха. Мероприятия по доведению микроклиматических показателей до нормативных значений включаются в комплексные планы предприятий по охране труда. Санитарные нормы устанавливают оптимальные и допустимые метеорологические условия в рабочей зоне помещения. Согласно СанПиН 2.2.4.548-96 выполняемая работа относится к категории легкая (16) – интенсивность энергозатрат в пределах 121-150 ккал/час (140-174 Вт), это работы: сидя, стоя или связанные с ходьбой с некоторым физическим напряжением. Оптимальные и допустимые показатели микроклимата приведены в таблицах 21 и 22.

Таблица 21 – Оптимальные величины показателей микроклимата на рабочих местах производственных помещений

Период года	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	Температура воздуха, °С	Температура поверхностей, °С
Холодный	60-40	0,1	21-23	20-24
Тёплый	60-40	0,1	23-25	22-26

Таблица 22 – Допустимые величины показателей микроклимата

Период года	Относительная влажность воздуха, %	Скорость движения воздуха, м/с		Температура воздуха, °С		Температура поверхностей, °С
		Для диапазона температур воздуха ниже оптимальных величин, не более	Для диапазона температур воздуха выше оптимальных величин, не более	Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин	
Холодный	15-75	0,1	0,2	19,0-20,9	23,1-24,0	18,0-25,0
Тёплый	15-75	0,1	0,3	20,0-21,9	24,1-28,0	19,0-29,0

В рассматриваемом помещении в холодный период температура поверхностей и температура воздуха составляет 22⁰С и 23⁰С соответственно, а влажность воздуха 45%, а в теплый период температура поверхностей и температура воздуха – 24⁰С и 25⁰С соответственно. Сравнивая со значениями из таблицы, отклонений от норм не выявлено. В данном случае температура воздуха и температура поверхностей составляют 22⁰С и 21⁰С при относительной влажности 45% в холодный период года, 24⁰С и 23⁰С при относительной влажности воздуха 50% в теплый период года, что соответствует нормам.

5.2.1.2 Недостаточная освещенность рабочей зоны

Освещение оказывает влияние на общее самочувствие и настроение, определяет эффективность труда. Нерационально организованное освещение может явиться причиной травматизма: недостаточно освещенные опасные зоны, слепящие источники света и блики от них, резкие тени и пульсации освещенности ухудшают видимость и могут вызвать неадекватное восприятие наблюдаемого объекта. В компьютерных комнатах должно быть как естественное, так и искусственное освещение. Естественное освещение обеспечивается за счет оконных проемов, коэффициент искусственного

освещения (КОЕ) которых должен быть не менее 1,2% в местах, где имеется снежный покров и не менее 1,5% на остальной территории. Свет из окна должен падать с левой стороны от пользователя. Естественное освещение в аудитории осуществляется через два оконных проема размером 1 на 1.35 метра в наружной стене. Нормируемые показатели естественного, искусственного и совмещенного освещения в соответствии с СанПиН 2.2.1/2.1.1.1278-03 указаны в таблице 23.

Таблица 23 – Нормируемые показатели естественного, искусственного и совмещенного освещения

Помещения	Рабочая поверхность и плоскость нормирования КЕО и освещенности и высота плоскости над полом, м	Естественное освещение		Совмещенное освещение		Искусственное совмещение				
		КЕО е_н, %		КЕО е_н, %		Освещенность, лк				
		При верхнем или комбинированном освещении	При боковом освещении	При верхнем или комбинированном освещении	При боковом освещении	При комбинированном освещении		При общем освещении	Показатель диска форта, М, не более	Коэффициент пульсации освещенности, К п, %, не более
1	2	3	4	5	6	7	8	9	10	11
Кабинеты, рабочие комнаты	Г – 0,8	3,0	1,0	1,8	0,6	400	200	300	40	15
Помещения для работы с дисплеями и видеотерминалами, залы ЭВМ	Г – 0,8 Экран монитора: В – 1,2	3,5 –	1,2 –	2,1 –	0,7 –	500 –	300 –	400 200	15 –	10

Для организации искусственного освещения в помещениях, в которых работают за персональными компьютерами, рекомендуется применять светильники типа ЛПО36. Также допустимо применять светильники прямого

света, преимущественно отраженного света типа ЛПО5, ЛПО13, ЛСО4, ЛПО34, ЛПО31 с люминесцентными лампами типа ЛБ. Ещё допускается применение светильников местного освещения с лампами накаливания. Светильники должны располагаться линиями (прямыми или прерывающимися) так, чтобы при различном расположении компьютеров они были параллельны линии зрения пользователя. Защитный угол светильников должен быть не менее 40 градусов.

Для того чтобы производственное освещение в помещении соответствовало всем нормам, нужно не менее двух раз в год мыть стекла и светильники, а также следить за работой светильников и при необходимости менять вышедшие из строя лампы.

Когда естественного освещения недостаточно, необходимо использовать общее искусственное освещение. В качестве основных источников искусственного освещения используются лампы белого и дневного света ЛБ-20 и ЛД-20.

В помещении, в котором проводилась работа, используются рядные расположенные потолочные светильники с люминесцентными лампами. В результате анализа освещенности рабочего места отклонений от норм выявлено не было. Уровень освещенности соответствует нормам в разные периоды светового дня.

Рассматриваемое помещение имеет светлый цвет потолков и стен, серое покрытие пола. Длина помещения (a) – 6 м., ширина (b) – 6 м., высота (h) – 2,8 м. В качестве источника света используются светильники, каждый из которых содержит по $n=4$ люминесцентные лампы мощностью 18 Вт.; общая яркость светового потока (Φ) 1150 Лм. Помещение предназначено для работы за персональным компьютером, поэтому нормой освещенности (E) для него согласно СНиП 23-05-95 станет 200-300 Лк, рабочая плоскость стола находится на расстоянии (h_1) 0,8 м. над уровнем пола, коэффициент запаса (k_3) равняется 1,4, а коэффициенты отражения: для потолка – 0,7; для стен – 0,5; для пола – 0,3.

Сначала находим площадь помещения (S): $6 \cdot 6 = 36 \text{ м}^2$. Далее находим индекс помещения по формуле:

$$\frac{S}{(h - h_1) * (a + b)} = \frac{36}{(2,8 - 0,8) * (6 + 6)} = 1,5.$$

Теперь на основании показателей отражения поверхностей и высчитанного индекса можно из таблицы определить коэффициент использования ($k_{исп}$). В данном случае он равняется 64.

И, наконец, определим необходимое количество светильников:

$$N = \frac{E * S * 100 * k_3}{U * n * \Phi} = \frac{300 * 36 * 100 * 1,4}{64 * 4 * 1150} = 5,14 \approx 6.$$

В помещении, в котором проводилась работа, используются рядно расположенные потолочные светильники с люминесцентными лампами. Проведенные расчеты показали, что минимальное число светильников должно быть равно 6. В результате анализа освещенности рабочего места отклонений от норм выявлено не было. Уровень освещенности соответствует нормам в разные периоды светового дня.

5.2.1.3 Превышение уровня шума

В случае постоянного нахождения при шуме более 85 децибел могут наблюдаться нарушения слуха. Также шум может мешать сконцентрироваться и являться фактором стресса, тем самым повышая систолическое кровяное давление; может привести к несчастным случаям, препятствуя получению предупредительных сигналов.

Для аудитории, в которой осуществлялась работа магистранта, основными источниками шума являются расположенные в помещении компьютеры и кондиционер. Уровни шума для различных категорий рабочих мест служебных помещений регламентирует ГОСТ 12.1.003-2014 [7].

Помещения, в которых для работы используются компьютеры, не должны соседствовать с помещениями, в которых уровни шума превышают нормируемые значения. В помещениях, которые оборудованы компьютерами, которые являются основным источником шума, уровень шума на рабочем месте должен быть не более 50 дБ.

Рассматриваемая аудитория по уровню производственных шумов не выходит за рамки допустимых значений. Уровень шума менее 50 дБ.

5.2.1.4 Опасные и вредные производственные факторы, связанные с электромагнитными полями

Во время работы за компьютером человек подвергается воздействию электромагнитных и электростатических полей. Создаваемое персональным компьютером электромагнитное излучение имеет электрическую (Е) и магнитную (Н) составляющие, а также сложный спектральный состав с диапазоном частот от 0 до 1000 МГц. Основным источником электромагнитных излучений является монитор, в состав которого входит трансформатор высокой частоты строчной развертки. На сегодняшний день ЭЛТ-мониторы потеряли свою популярность. Их вытеснили ЖК-мониторы, уровень электромагнитного излучения, которых гораздо меньше. СанПиН 2.2.4.1191-03 определяет нормы допустимых уровней напряженности электрических полей. Они зависят от времени пребывания человека в контролируемой зоне. Время допустимого пребывания в рабочей зоне в часах рассчитывается по формуле $T=50/E-2$. Если напряженность электрического поля лежит в диапазоне 20–25 кВ/м, то работа не может продолжаться более 10 минут. При напряженности, не превышающей 5 кВ/м, деятельность людей в рабочей зоне может осуществляться в течение 8 часов. Еще одним нормативным документом в данной сфере является СанПиН 2.2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы», регламентирующие безопасные уровни излучений. В таблицах 24-25 представлены предельно-допустимые уровни напряженности на рабочих местах и допустимые уровни электромагнитных полей.

Таблица 24 – Предельно-допустимые уровни напряженности на рабочих местах

Время воздействия за рабочий день, мин	Условия воздействия			
	Общее		Локальное	
	ПДУ напряженности кА/м	ПДУ магнитной индукции мТл	ПДУ напряженности кА/м	ПДУ магнитной индукции мТл
0-10	24	30	40	50
11-60	16	20	24	30
61-480	8	10	12	15

Таблица 25 – Допустимые уровни электромагнитных полей

Наименование параметра	
Напряженность электромагнитного поля на расстоянии 50 см вокруг дисплея до электрической составляющей, В/м, не более:	
в диапазоне частот 5 Гц – 2 кГц	25
в диапазоне частот 2 – 400 кГц	2,5
Плотность магнитного потока на расстоянии 50 см вокруг дисплея, нТл, не более:	
в диапазоне частот 5 Гц – 2 кГц	250
в диапазоне частот 2 – 400 кГц	25
Поверхностный электростатический потенциал, В, не более	500

Уровень напряженности электромагнитного поля в рассматриваемой аудитории не превышает предельно-допустимого значения. Все рабочие машины прошли сертификацию, а рабочие места аттестованы, индивидуальная защита пользователей не требуется.

5.2.1.5 Психофизиологические факторы

Во время длительной работы за компьютером человек также может подвергаться воздействию психофизиологических факторов, таких как эмоциональные перегрузки, умственное перенапряжение, монотонность труда и другие. Эмоциональные перегрузки вызывают изменения функционального состояния центральной нервной системы, что может негативно отразиться на состоянии организма в целом. Они могут быть вызваны необходимостью выполнения большого объема работы, конфликтными или стрессовыми ситуациями. Умственное перенапряжение может наступать вследствие отсутствия необходимого времени на отдых после продолжительной работы, нарушения режима сна или режима питания.

Оно может накапливаться и приводить к возникновению заболеваний. Отличительными признаками монотонной работы служат однообразие рабочих действий, их многократное повторение и небольшая длительность. Таковой является работа за компьютером. В результате работающий теряет интерес к работе, и у него возникает состояние «производственной скуки». Монотонная работа отрицательно сказывается на эффективности производства: ухудшаются экономические показатели, повышается аварийность, травматизм, растет текучесть кадров. Также вредным фактором производства может служить фиксированная рабочая поза. Она вызывает нарушение кровообращения в нижних конечностях и органах тазовой области, которое может приводить к профессиональным заболеваниям, например, варикозное расширение вен. Для снижения эмоциональных перегрузок и умственных перенапряжений предусмотрены перерывы в работе, возможность выбора удобного времени для выполнения работы. Для уменьшения рисков возникновения последствий от фиксированной рабочей позы, установленное в аудитории оборудование имеет регулировки: стул регулируется по высоте и наклону спинки, монитор позволяет подобрать наклон под индивидуальные особенности человека.

5.2.2 Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения

5.2.2.1 Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека

Опасное и вредное воздействия на людей электрического тока и электрической дуги проявляются в виде профессиональных заболеваний и электротравм.

Помещение, в котором расположены персональные компьютеры, относится к помещениям без повышенной опасности, потому что отсутствуют следующие факторы: высокая температура, токопроводящая пыль, токопроводящие полы, сырость, возможность одновременного прикосновения человека к имеющим соединение с землёй

металлоконструкциям зданий, технологическим аппаратам и механизмам и металлическим корпусам электрооборудования.

Мероприятия, направленные на предотвращение возможности поражения электрическим током, включают в себя следующее:

- при выполнении монтажных работ необходимо использовать только исправно работающий инструмент, аттестованный службой КИПиА;
- заземление корпусов приборов и инструментов, которое поможет защитить от поражения электрическим током, который может возникнуть между корпусом приборов и инструментом при пробое сетевого напряжения на корпус;
- запрет на выполнение работ на задней панели при включенном сетевом напряжении;
- выполнение работ по устранению неисправностей должно производиться компетентными людьми;
- нужно постоянно наблюдать за исправностью электропроводки и в случае обнаружения неисправностей незамедлительно принимать действия по их устранению.

Перед началом работы необходимо проверить, чтобы не было свешивающихся со стола или висящих под столом проводов электропитания, убедиться в целостности вилки и провода электропитания, в отсутствии видимых повреждений аппаратуры и рабочей мебели, в отсутствии повреждений и наличии заземления приэкранного фильтра. Токи статического электричества, которые могут возникнуть в процессе работы компьютера на корпусах системного блока, монитора и клавиатуры, могут провоцировать разряды при прикосновении к этим элементам, которые не представляют опасности для человека, но могут привести к поломке компьютера. Для уменьшения величин токов статического электричества применяются нейтрализаторы, увлажнение воздуха (местное и общее), используются покрытия полов с антистатической пропиткой.

5.3 Экологическая безопасность

Охрана окружающей среды заключается в устранении отходов жизнедеятельности человека и бытового мусора. Если персональные компьютеры теряют свою работоспособность, их списывают и отправляют на специализированный склад, на котором уже принимаются меры по утилизации техники и комплектующих.

По статистике вышедшие из строя люминесцентные лампы являются одним из самых распространенных источников ртутного загрязнения. Помимо стекла и алюминия каждая лампа содержит приблизительно 60 мг ртути, поэтому отработавшие люминесцентные лампы являются опасным источником токсичных веществ.

Утилизация таких ламп заключается в их передаче перерабатывающим предприятиям, которые имеют специальное оборудование для переработки вредных ламп в безвредное сырье – сорбент, которое может являться материалом для других производств.

Согласно ГОСТ Р 57740-2017 и ГОСТ Р 51768-2001, отработанные люминесцентные лампы относятся к отходам, которые собираются и сортируются отдельно, поэтому их утилизация и хранение должны отвечать определенным требованиям.

5.4 Безопасность в чрезвычайных ситуациях

Для аудитории, в которой проходит написание ВКР, наиболее вероятно возникновение такой ЧС как пожар, который может возникнуть при замыкании электропроводки оборудования, обрыве проводов или же при несоблюдении мер пожарной безопасности.

Помещение, в котором выполняется работа по написанию выпускной работы, относится к категории «В» по пожарной и взрывной опасности.

К противопожарным мероприятиям в помещении относятся следующие действия:

- 1) помещение должно быть оборудовано средствами тушения пожара, такими как огнетушители, стенд с противопожарным инвентарем, ящик с

песком; средствами связи; электрическая проводка осветительных приборов и электрооборудования должна быть в исправном состоянии.

2) каждый сотрудник должен знать месторасположение средств тушения пожара и средств связи; знать номера телефонов экстренных служб для оповещения о пожаре; уметь использовать средства пожаротушения.

Рассматриваемое помещение оснащено средствами пожаротушения в соответствии с нормами: огнетушитель пенный ОП-10 – 1 шт., огнетушитель углекислотный ОУ-5 – 1 шт.

В помещении и на этаже присутствуют следующие средства оповещения: световая индикация направления движения к выходу в коридорах этажа, звуковая индикация, которая представляет собой систему оповещения о пожаре через громкоговоритель, пассивные датчики задымленности.

Чтобы минимизировать вероятность возникновения пожара нужно своевременно проводить профилактические работы, направленные на устранение возможных источников возникновения пожара, такие как: систематическое наблюдение за состоянием электропроводки, выключение питания оборудования при завершении работы и покидании рабочего места, периодическое проведение инструктажа по пожаробезопасности для персонала.

Для того чтобы увеличить устойчивость рабочего помещения к возможному возникновению ЧС, необходимо устанавливать системы противопожарной сигнализации, которые будут реагировать на задымленность и другие продукты горения, размещать необходимые огнетушители, обеспечивать помещение и инструктировать рабочих о плане эвакуации, а также назначить лиц, ответственных за данные мероприятия. Два раза в год (в разные периоды – зимой и летом) проводить учебные тревоги для отработки действий при пожаре.

Если же пожар все-таки возник, необходимо эвакуировать персонал из рабочего помещения в соответствии с планом эвакуации, который изображен

на рисунке 19. При отсутствии прямых угроз жизни и здоровью необходимо предпринять возможные меры по тушению источника возгорания огнетушителем. В случае потери контроля над пожаром, необходимо немедленно эвакуироваться по плану эвакуации и ждать приезда специалистов соответствующих служб. При возникновении пожара должна сработать система пожаротушения, издав предупредительные сигналы, и передав на пункт пожарной станции сигнал о ЧС, в случае если система не сработала, по каким-либо причинам, необходимо нажать тревожную кнопку или самостоятельно произвести вызов пожарной службы по телефону 101, сообщить место возникновения ЧС и ожидать приезда специалистов.



Рисунок 19 – План эвакуации

5.5 Выводы по разделу

В данном разделе были рассмотрены основные требования к производственному помещению, в соответствии с текущим законодательством и нормами РФ, основные аспекты производственной, экологической и техногенной безопасности. В рамках этого раздела были изучены показатели микроклимата рабочего помещения, а именно: освещенность, шум, электромагнитные поля, влажность. Также были рассмотрены возможные психофизиологические факторы воздействия на организм сотрудника. Выполнены необходимые расчеты показателей

освещенности. В результате анализа всех факторов, рассматриваемое помещение является полностью безопасным для работы и соответствует нормативам.

Экологическая же безопасность сводится к правильной утилизации вредных отходов, таких как сломанное оборудование и люминесцентные лампы.

Техногенная безопасность заключается в проведении мер по предотвращению возникновения чрезвычайных ситуаций и обучению персонала правильному поведению при ЧС.

Заключение

В результате выполнения магистерской диссертации, была получена система поддержки научных исследований медицинской диагностики в виде дерева решений, на основе полученных клинико-лабораторных показателей.

В процессе разработки, были рассмотрены объекты и цели исследования, были описаны исходные данные. Был выбран метод для заполнения пропущенных значений, метод для сокращения размерности, описаны виды кластеризации и метод построения дерева решений, а также были выбраны и описаны инструменты для реализации поставленной задачи. Был разработан алгоритм выполнения поставленной задачи. Была полностью реализована поставленная задача, в соответствии с шагами, описанными в алгоритме.

Список использованных источников

1. Макарова Л.С., Берестнева О.Г., Семерякова Е.Г. Математические методы в задачах медицинской диагностики [Электронный ресурс]. – URL: <https://science-education.ru/pdf/2012/6/180.pdf> (дата обращения: 06.02.2019).
2. Орлов А.И. Математические методы теории классификации [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/v/matematicheskie-metody-teorii-klassifikatsii> (дата обращения: 06.02.2019).
3. Н. Бейли. Математика в биологии и медицине [Электронный ресурс]. – URL: http://stu.sernam.ru/book_m_bio.php?id=71 (дата обращения: 06.02.2019).
4. Новосельцев В.Н. Математические методы в медицине [Электронный ресурс]. – URL: http://бмэ.орг/index.php/МАТЕМАТИЧЕСКИЕ_МЕТОДЫ (дата обращения: 06.02.2019).
5. Черкашина Ю.А. Применение математических методов в задаче диагностики состояния здоровья детей первого года жизни [Электронный ресурс]. – URL: <https://applied-research.ru/pdf/2015/5-1/6752.pdf> (дата обращения: 06.02.2019).
6. М.Л.Жмудяк, А.Н. Повалихин, А.В. Стребуков, А.В.Гайнер, А.Л.Жмудяк, Г.Г.Устинов. Диагностика заболеваний методами теорий вероятностей [Электронный ресурс]. – URL: http://medbooks.org/index.files/book/Matematika/012708/Diagnostika_zabolevanij_metodami_teorii_verojatnostej.pdf (дата обращения: 06.02.2019).
7. М.Д. Степанова, С.А. Самодумкин, Т.Л. Лемешева. Математические методы диагностики в медицинских интеллектуальных системах [Электронный ресурс]. – URL: http://900501.ucoz.ru/_ld/0/31_--_.pdf (дата обращения: 06.02.2019).
8. В.А. Биллиг, О.В. Иванова, Н.А. Царегородцев. Построение ассоциативных правил в задаче медицинской диагностики [Электронный

ресурс]. – URL: <https://cyberleninka.ru/article/n/postroenie-assotsiativnyh-pravil-v-zadache-meditsinskoy-dagnostiki> (дата обращения: 06.02.2019).

9. Компанченко А.С. Сравнение эффективности ансамблей интеллектуальных технологий в задачах медицинской диагностики [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/sravnenie-effektivnosti-ansambley-intellektualnyh-tehnologiy-v-zadachah-meditsinskoy-dagnostiki> (дата обращения: 06.02.2019).

10. Локтионова Н.Н. Фильчакова К.А. Применение математических методов исследования в медицине [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/primenenie-matematicheskikh-metodov-issledovaniya-v-meditsine> (дата обращения: 07.02.2019).

11. Горбунова Т.И. Интеллектуальная система поддержки принятия решений в медицинской диагностике [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/intellektualnaya-sistema-podderzhki-prinyatiya-resheniy-v-meditsinskoy-dagnostike> (дата обращения: 07.02.2019).

12. Черкашина Ю.А., Вадутова Ф.А. Оценка состояния здоровья детей с экстремально низкой массой тела на основе технологии искусственных нейронных сетей [Электронный ресурс]. – URL: <https://elibrary.ru/item.asp?id=29810157&> (дата обращения: 07.02.2019).

13. Черкашина Ю.А. Статистический анализ медико-биологических показателей у детей, больных ожирением [Электронный ресурс]. – URL: <https://elibrary.ru/item.asp?id=23172055&> (дата обращения: 07.02.2019).

14. Заброда Н.Н., Артеменко М.В. Математические методы анализа заболеваемости населения региона Курской магнитной аномалии [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/matematicheskie-metody-analiza-zabolevaemosti-naseleniya-regiona-kurskoy-magnitnoy-anomalii> (дата обращения: 07.02.2019).

15. Самойленко Н. Э., Кувина В. Н., Кувин С. С. Комплексный анализ медицинских данных [Электронный ресурс]. – URL:

<https://cyberleninka.ru/article/n/kompleksnyy-analiz-meditsinskih-dannyh> (дата обращения: 07.02.2019).

16. Ананьев К.И., Демьянова В.В., Демьянов В.Ф., Кокорина А.В., Свистун С.Я., Стегалин И.С. Оптимизационные методы в задачах диагностики [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/optimizatsionnye-metody-v-zadachah-diagnostiki> (дата обращения: 07.02.2019).

17. Харитонов Т.С. Задача классификации по уровню стрессоустойчивости с помощью нейронных сетей [Электронный ресурс]. – URL: <https://elibrary.ru/item.asp?id=32330369> (дата обращения: 08.02.2019).

18. Мазуров В.Д., Мазуров А.Д., Шестаков А.А. Математические модели диагностики и прогнозирования в медицине и биологии [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/matematicheskie-modeli-diagnostiki-i-prognozirovaniya-v-meditsine-i-biologii> (дата обращения: 08.02.2019).

19. Шайдуров А.А. Построение многокритериальной нейросетевой модели оптимизации в задачах диагностики заболеваний со слабо выраженной симптоматикой [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/postroenie-mnogokriterialnoy-neyrosetevoy-modeli-optimizatsii-v-zadachah-diagnostiki-zabolevaniy-so-slabo-vyrazhennoy-simptomatiko> (дата обращения: 08.02.2019).

20. Антонова Н.Е. Сравнительный анализ применения статистических методов и нейронных сетей для диагностики заболеваний опорно-двигательного аппарата [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/v/cravnitelnyy-analiz-primeneniya-statisticheskikh-metodov-i-neyronnyh-setey-dlya-diagnostiki-zabolevaniy-oporno-dvigatel'nogo-apparata> (дата обращения: 08.02.2019).

21. Головин П.А., Нечаев В.А., Нечаев Д.А. Экспертные системы для классификации болезней в медицинской диагностике [Электронный ресурс].

– URL: <https://cyberleninka.ru/article/n/ekspertnye-sistemy-dlya-klassifikatsii-bolezney-v-medsinskoy-diagnosticske> (дата обращения: 08.02.2019).

22. Мишин А.С., Арзамасцев А.А., Зенкова Н.А. Экспертная система для прогнозирования результатов хирургического лечения больных колоректальным раком, осложненных острой кишечной непроходимостью [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/ekspertnaya-sistema-dlya-prognozirovaniya-rezultatov-hirurgicheskogo-lecheniya-bolnyh-kolorektalnym-rakom-oslozhnennyh-ostroy> (дата обращения: 08.02.2019).

23. Алпеева А.В. Результаты прогнозирования и распознавание характера заболевания системы внешнего дыхания у детей дошкольного возраста с помощью математических методов [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/rezultaty-prognozirovaniya-i-raspoznavanie-haraktera-zabolevaniya-sistemy-vneshnego-dyhaniya-u-detey-doshkolnogo-vozrasta-s-pomoschyu> (дата обращения: 08.02.2019).

24. Оразбаев Б.Б. Экспертные системы для медицинской диагностики с применением методов теории нечетких множеств [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/ekspertnye-sistemy-dlya-medsinskoy-diagnosticski-s-primeneniem-metodov-teorii-nechetkih-mnozhestv> (дата обращения: 08.02.2019).

25. Ильясова Н.Ю., Парингер Р.А., Куприянов А.В. Формирование признаков для повышения качества медицинской диагностики на основе методов дискриминантного анализа [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/formirovanie-priznakov-dlya-povysheniya-kachestva-medsinskoy-diagnosticski-na-osnove-metodov-diskriminantnogo-analiza> (дата обращения: 08.02.2019).

26. Бронхиальная астма - воспаление бронхов // Альфа центр здоровья [Электронный ресурс]. – URL: <https://www.alfazdrav.ru/zabolevania/bronkhialnaya-astma/> (дата обращения 10.02.2019).

27. Бронхиальная астма – проблемы и достижения. По материалам 15-го ежегодного конгресса европейского респираторного общества // «Клиническая иммунология. Аллергология. Инфектология». Издательский дом «Здоров'я України». – 2005. – № 1 [Электронный ресурс]. – URL: <https://web.archive.org/web/20080605045429/http://immuno.health-ua.com/article/3.html> (дата обращения: 10.02.2019).
28. PRP // RDocumentation [Электронный ресурс]. – URL: <https://www.rdocumentation.org/packages/rpart.plot/versions/3.0.7/topics/prp> (дата обращения: 12.02.2019).
29. Clusplot.default // RDocumentation [Электронный ресурс]. – URL: <https://www.rdocumentation.org/packages/cluster/versions/2.0.8/topics/clusplot.default> (дата обращения: 12.02.2019).
30. Оценка качества в задаче кластеризации // Университет ИТМО [Электронный ресурс]. – URL: <https://clck.ru/GLTUr> (дата обращения: 18.02.2019).
31. МЕТОДЫ ОЦЕНКИ КАЧЕСТВА ЧЁТКОЙ КЛАСТЕРИЗАЦИИ // CYBERLENINKA [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/metody-otsenki-kachestva-chyotkoy-klasterizatsii> (дата обращения: 18.02.2019).
32. Обработка пропусков в данных – часть 1 // BaseGroup Labs Технологии анализа данных [Электронный ресурс]. – URL: <https://basegroup.ru/community/articles/missing> (дата обращения: 22.02.2019).
33. Прикладная статистика // Орлов А.И. – М.: Издательство «Экзамен», 2004. [Электронный ресурс]. – URL: <http://www.aup.ru/books/m163/> (дата обращения: 22.02.2019).
34. АЛГОРИТМЫ И МЕТОДЫ СНИЖЕНИЯ ПРОСТРАНСТВА ДИАГНОСТИЧЕСКИХ ПРИЗНАКОВ [Электронный ресурс]. – URL: <https://cyberleninka.ru/article/n/lgoritmy-i-metody-snizheniya-prostranstva-diaagnosticheskikh-priznakov> (дата обращения: 01.03.2019).

35. Кластеризация // Университет ИТМО [Электронный ресурс]. – URL: <http://neerc.ifmo.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F> (дата обращения: 01.03.2019).
36. Классификация, регрессия и другие алгоритмы Data Mining с использованием R // Шитиков В.К., Мاستицкий С.Э. (2017). [Электронный ресурс]. – URL: <https://github.com/ranalytics/data-mining> (дата обращения: 09.03.2019).
37. Оценка качества обучения. Часть 2 // Coursera [Электронный ресурс]. – URL: <https://www.coursera.org/lecture/vvedeniye-v-nauku-o-dannykh/otsienka-kachiestva-obucheniia-chast-2-Db3Xm> (дата обращения: 09.03.2019).
38. Обзор статистических программ // ScienceFiles Medical Writing&Biostatistics [Электронный ресурс]. – URL: <http://www.sciencefiles.ru/section/46/> (дата обращения: 10.03.2019).
39. Язык программирования R и его место среди статистических программ // Samoedd [Электронный ресурс]. – URL: <http://samoedd.com/soft/r-introduction> (дата обращения: 10.03.2019).
40. R (язык программирования) // Национальная библиотека им. Н. Э. Баумана Bauman National Library [Электронный ресурс]. – URL: [https://ru.bmstu.wiki/R_\(язык_программирования\)](https://ru.bmstu.wiki/R_(язык_программирования)) (дата обращения: 10.03.2019).
41. Статистический анализ и визуализация данных с помощью R // Шитиков В.К., Мастыцкий С.Э. (2014). [Электронный ресурс]. – URL: http://www.socd.univ.kiev.ua/sites/default/files/library/elopen/mastitsky_and_shitikov_2014_r_tutorials.pdf#2 (дата обращения: 12.03.2019).
42. Языков К.Г., Немеров Е.В. К вопросу изучения личностных свойств в психофизиологической реактивности больных бронхиальной

астмой на аудиовизуальную стимуляцию // Вестник ТГПУ. 2011. № 6(108). – С. 134-138.

43. Global strategy for asthma management and prevention // Pharmther [Электронный ресурс]. – URL: http://pharmther.nuph.edu.ua/wp-content/uploads/2015/10/GINA_Report_2015_Aug11.pdf (дата обращения: 18.03.2019).

44. Необходимо ли вам изучать язык R? // Olap [Электронный ресурс]. – URL: <http://www.olap.ru/home.asp?artId=2895> (дата обращения: 15.03.2019).

45. Кластерный анализ в R: отделяем мух от котлет // Мои волшебные инструменты URL: <http://mymagictools.blogspot.ru/2015/07/r.html> (дата обращения: 25.03.2019).

46. Коэффициент корреляции // Портал знаний [Электронный ресурс]. – URL: <http://statistica.ru/glossary/general/koeffitsient-korrelyatsii/> (дата обращения: 26.03.2019).

47. An introduction to corrplot Package [Электронный ресурс]. – URL: <https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html> (дата обращения: 26.03.2019).

48. Деревья принятия решений в R // Re9ulus Blog [Электронный ресурс]. – URL: http://re9ulus.github.io/2015/12/07/trees_in_r/ (дата обращения: 02.04.2019).

49. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей // Habr [Электронный ресурс]. – URL: <https://habr.com/ru/company/ods/blog/322534/> (дата обращения: 02.04.2019).

50. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес // Habr [Электронный ресурс]. – URL: <https://habr.com/ru/company/ods/blog/324402/> (дата обращения: 05.04.2019).

51. Формы и виды бронхиальной астмы: основные принципы и критерии классификации [Электронный ресурс]. – URL: <https://bronhialnaya->

astma.com/vidy-astmy/formy-i-vidy-bronhialnoj-astmy#i-6 (дата обращения: 06.04.2019).

52. Деревья решений – общие принципы работы // BaseGroup Labs Технологии анализа данных [Электронный ресурс]. – URL: <https://basegroup.ru/community/articles/description> (дата обращения 08.04.2019).

53. Бэггинг и случайный лес // Coursera [Электронный ресурс]. – URL: <https://www.coursera.org/lecture/vvedenie-mashinnoe-obuchenie/beghghingh-i-sluchainyi-lies-n6JbA> (дата обращения 12.04.2019).

54. Бэггинг // MachineLearning [Электронный ресурс]. – URL: <http://www.machinelearning.ru/wiki/index.php?title=Бэггинг> (дата обращения 15.04.2019).

55. Алгоритм ближайшего соседа // BaseGroup Labs Технологии анализа данных [Электронный ресурс]. – URL: <https://basegroup.ru/community/articles/knn> (дата обращения 15.04.2019).

56. Метод ближайших соседей // MachineLearning [Электронный ресурс]. – URL: <http://www.machinelearning.ru/wiki/index.php?title=KNN> (дата обращения 15.04.2019).

57. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета. – 70 с.

58. Социальная ответственность: Методические указания по разработке раздела «Социальная ответственность» выпускной квалификационной работы магистра, специалиста и бакалавра всех направлений (специальностей) и форм обучения ТПУ [Электронный ресурс] Сост. Е.Н. Пашков, И.Л. Мезенцева. – Томск: Изд-во Томского политехнического университета, 2019. – 24 с.

59. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019).

60. Закон Томской области от 9 июля 2003 года №83-ОЗ «Об охране труда в Томской области» (с изменениями на 4 июля 2014 года).

61. ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования.
62. ГОСТ 12.2.061-81 ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам.
63. ГОСТ 12.0.003-2015 ССБТ. Опасные и вредные производственные факторы. Классификация.
64. ГОСТ 12.1.003-2014 ССБТ. Шум. Общие требования безопасности.
65. ГОСТ 12.1.005-88 ССБТ. Общие санитарно-гигиенические требования к воздуху рабочей зоны.
66. ГОСТ 12.1.006-84 ССБТ. Электромагнитные поля радиочастот. Общие требования безопасности.
67. СанПиН 2.2.1/2.1.1.1278–03. Гигиенические требования к естественному, искусственному и совмещённому освещению жилых и общественных зданий.
68. СанПиН 2.2.2/2.4.1340–03. Санитарно-эпидемиологические правила и нормативы «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы».
69. СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату производственных помещений.
70. СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95.
71. ГОСТ Р 22.0.01-2016. Безопасность в ЧС. Основные положения.
72. СанПиН 2.2.4.3359-16 Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.
73. ГОСТ 12.1.038–82 Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов.
74. ТОИ Р-45-084-01. Типовая инструкция по охране труда при работе на персональном компьютере" (утв. Приказом Минсвязи РФ от 02.07.2001 N 162).

Приложение А (справочное)

Разделы:
Introduction
Chapter 1 Subject domain analysis
Chapter 3 Project implementation

Студент:

Группа	ФИО	Подпись	Дата
8KM71	Сидоров Владислав Викторович		

Консультант отделения ИТ:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Марухина О.В.	к.т.н.		

Консультант – лингвист отделения ИЯ:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ, ШБИП	Диденко А.В.	к.ф.н		

Introduction

Bronchial asthma is a disease based on the increased sensitivity of the bronchi to various irritators (in particular, allergens). In response to the action of these irritators, the narrowing (obstruction) of the bronchi develops. This process is conditioned by several reasons: the increased bronchial tone, excessive discharge into the bronchial lumen and their inflammation. In asthma, the attacks are typically episodic, for example, after contact with an irritant. In severe asthma, bronchial obstruction often persists between attacks as well. This disease is chronic, so it requires constant monitoring.

Currently, according to the WHO, the number of people suffering from such a disease as the Bronchial Asthma (BA) is increasing every year. Over the past twenty years, the number of people with this disease has grown to about three hundred million globally. This disease is one of the most common chronic diseases in the world, which affects both men and women, regardless of age. Moreover, in most cases, if this disease onsets in childhood, it persists when the person becomes adult.

Owing to the work of scientists, which consisted in the analysis of 34 studies of the incidence of bronchial asthma in Europe, we learned that in Austria the incidence in children increased 4 times in the 10-year period from 1992 to 2002. In Italy the increase was 6% (from 7% to 13%) in the period between 1974 and 1998. In many European countries, the incidence continued to increase until the mid-90s of the twentieth century, but at the moment there is a slight decrease. [27]

Despite the fact that currently there is an effective therapy for this disease, the spread of asthma and the increase in the number of asthma patients has resulted in the renewed researchers' interest in the problem of typing this disease. Furthermore, the typing methods differ. Some researchers investigate the clinical indicators of patients and make their decisions based on them. Others consider the psychological vector, while patients are regarded as a homogeneous population of people in terms of somatic status, but with different psycho-emotional states.

However, the diversity of psychological and somatic must be studied comprehensively. In this context, the relevance of the work is beyond doubt.

The study is based on the psycho-synergetic model of psychogenic bronchial asthma proposed by Tomsk medical researchers (K.G. Yazykov, E.V. Nemerov). In line with this model, empirical groups are selected, which are named as follows: PIBA – psychogenically induced bronchial asthma; SPBA – somatopsychogenic bronchial asthma (a preliminary term); NPBA – non-psychogenic bronchial asthma; PsD – psychogenic dyspnea.

Classification is one of the most important and fundamental processes in scientific research and further practical applications. The introduction of this classification of bronchial asthma may produce more specific treatment methods for patients with different types of this disease.

Chapter 1 Subject domain analysis

1.1 The research object and purposes

The study object is the data array, namely, the medical records of patients with bronchial asthma.

The study purpose is developing the support system for research studies of medical diagnostics based on the obtained medical data.

1.2 Initial data and classifications of bronchial asthma

The City Clinical Hospital No. 3 for the city of Tomsk has provided an array of data containing the data on adult patients with bronchial asthma, namely, their values before and after treatment. The patients are adults with the bronchial asthma disease. Now, based on these indicators further work will be performed. The initial data contain the following:

Sex - patient's sex;

Age - patient's age, years;

Height - patient's height, cm;

Weight - patient's weight, kg;

Deg. RF - degree of respiratory failure;

Deg. BA sev. - degree of bronchial asthma severity;

Age 1 att. - patient's age at the moment of the first attack of the disease, years;

Dur.dis. - duration of disease (years);

Cl. manif.atopy - clinical manifestations (signs) of atopy;

Lung ventilation values: 1-before, 2- after ABC

PMV– pulmonary minute volume, l/min;

PC - pulmonary capacity, % (to the due values);

FPC– Forced PC, %;

FEV1– forced expiratory volume per sec, %;

FEV1/PC – FEV1/PC ratio, %;

MPV – maximal pulmonary ventilation per minute, %;

PFR ex. – peak expiratory flow rate %;

MFR25 – maximal expiratory flow rate at the level of 25% FPC;

MFR50 – maximal expiratory flow rate at the level of 50% FPC;

MFR75 – maximal expiratory flow rate at the level of 75% FPC;

All the lung ventilation values (except for PMV) are expressed in % to due values.

Pulmonary mechanics values: 1-before, 2- after ABC

Cdyn – dynamic pulmonary compliance (C), l/cm w.c.;

Cstat – static pulmonary compliance (C), l/cm w.c.;

Rin – inspiratory airways resistance, cm w.c./l/s;

Rex – expiratory airways resistance, cm w.c./l/s;

Wtot – total breathing work in case of spontaneous breathing, kgm/min;

Wsp – specific breathing work (per litre of ventilation), kgm/l;

Wn.el – non-elastic fraction of the total breathing work, kgm/min;

Wel – elastic fraction of the total breathing work, kgm/min;

W PMV10 – total breathing work at PMV10, kgm/min;

W PMV15 – total breathing work at PMV15, kgm/min;

W MPVtot – total breathing work at MPV, kgm/min;

W MPVsp – breathing work at MPV, kgm/min;

Pulmonary mechanics values are further brought to the SI system.

Below are given the psychological test values:

DeprB – Boeck depression;

DeprZ – Zung depression;

StAnx – state anxiety;

StAnx_12 – grades of state anxiety;

TrAnx – trait anxiety;

TrAnx _12 – grades of trait anxiety;

ShAnx – Sheehan anxiety;

ShAnx _12 – grades of Sheehan anxiety;

Deg. RF - degree of respiratory failure;

Deg. BA sev. – degree of bronchial asthma severity;

Age 1 att. – patient's age at the moment of the first attack of the disease;
Dur.dis. – duration of disease;
Cattel test values (16 scales designated with letters from A to Q4) before (1)
and after (2) ABC course;

MMPI test scale values before (1) and after (2) ABC course;

LSI (lifestyle index) test scale values before (1) and after (2) ABC course;

D:Boeck1 – Boeck depression before the ABC course;

D:Boeck2 – Boeck depression after the ABC course;

D:Zung1 – Zung depression before the ABC course;

D:Zung2 – Zung depression after the ABC course;

RA1 – reactive anxiety before the ABC course;

RA2 – reactive anxiety after the ABC course;

TA1 – trait anxiety before the ABC course;

TA2 – trait anxiety after the ABC course;

T:Sheehan1 – Sheehan anxiety before the ABC course;

T:Sheehan2 – Sheehan anxiety after the ABC course;

TAS1 – TAS alexithymia scale before the ABC course;

TAS2 – TAS alexithymia scale after the ABC course;

There are also several asthma classifications such as:

1. By the control level:

- Well-controlled;
- Partially controlled;
- Poorly controlled;

2. In accordance with the severity level. This classification is offered by the RAMS academicians: A.D. Ado and B.K. Bulatov. It is based on the frequency and intensity of the attacks and on the level of their impact on patient's life. The following types are distinguished:

- Intermittent (episodic) asthma. The mildest form of manifestations, which is characterized by long-term remission;

- Mild persistent asthma. The symptoms are more frequent - several attacks can occur in a week, both during the day and night;
- Moderate persistent asthma. It is characterized by daily attacks occurring during the day time. Night time deteriorations occur about two times a week. These attacks significantly worsen the patient's life;
- Persistent severe bronchial asthma. Symptoms of the disease constantly disturb the patient; labour performance is significantly hampered.

3. Specific forms of bronchial asthma:

- Professional;
- Reflux-induced form develops in case of reduced oesophageal sphincter tone;
- Aspirin asthma represents a pseudo-allergic inflammatory reaction of the organism, which is caused by the occurrence of hypersensitivity to nonsteroidal anti-inflammatory drugs containing acetylsalicylic acid;
- Physical stress - an attack occurs in case of physical exertion, as well as after the loading;
- Night asthma. Attacks mainly occur at night.

4. By causes of attacks. This classification has become widespread abroad, although it has recently been criticized. It is based on the causes of attacks. There are the following types of asthma in this classification:

- Exogenous form of bronchial asthma;
- Endogenous form of bronchial asthma;
- Bronchial asthma of mixed genesis.

This work is based on the classification, as mentioned earlier, proposed by Tomsk medical professionals: K.G. Yazaykov, Ye.V. Nemerov.

The study is based on the psycho-synergetic model of psychogenic bronchial asthma. In line with this model, empirical groups are selected, which are named as follows:

The first group - PIBA, is composed of people whose first asthma attack had developed after suffering an emotional distress. Its deterioration is also associated with psychological shocks.

The second group - SPBA is composed of people whose “normal” course of the disease was disturbed by life stress, after which psycho-emotional triggers caused severe asthma attacks, exacerbations of the disease.

The third group - NPBA, is composed of people with the classic exogenous form of bronchial asthma.

The fourth group - PsD is composed of people with neurotic respiratory disorders in such phenomena as panic disorders, hysterical conditions and various psychosomatic diseases.

1.3 Formal problem statement

The initial data array of our problem is the matrix X_{nm} ,

where n is the number of patients;

m – number of medical values.

In solving the problem, the multidimensional data analysis is used. The multidimensional analysis is applied to complex systems, which elements are characterized by a multitude of interdependent objects and attributes represented in the form of a matrix, which lines correspond to the observed objects, and the columns correspond to the sign characterizing them.

$$\hat{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix},$$

Where n is the number of objects (lines);

m – the number of signs (columns);

x_{nm} – the specific value of the attribute m at the k -th x element of the n -th object.

To solve the problem, the existing methods and approaches should be analyzed with selecting one of the methods and checking the possibility of its computer implementation, which will allow processing the provided data array and

achieving the planned goal. As a result of solving the problem, a support system for research studies of medical diagnostics will be created based on data mining methods.

Chapter 3 Project implementation

3.1 The algorithm for solving the problem

To process and analyse the obtained medical values, data are prepared and preprocessed, correlation analysis is performed during which the existing correlations are revealed and the dimension of the matrix of initial values is reduced. A cluster analysis is performed to identify homogeneous groups of patients, and the clustering quality is assessed. The decision support system is developed in the form of a decision tree. All this is accomplished through the use of the R language capabilities in the RStudio development environment. In the course of the work, a sequence of actions is developed serving as the basis for analysing the initial values. It is shown in the form of a diagram in Figure 3.1.

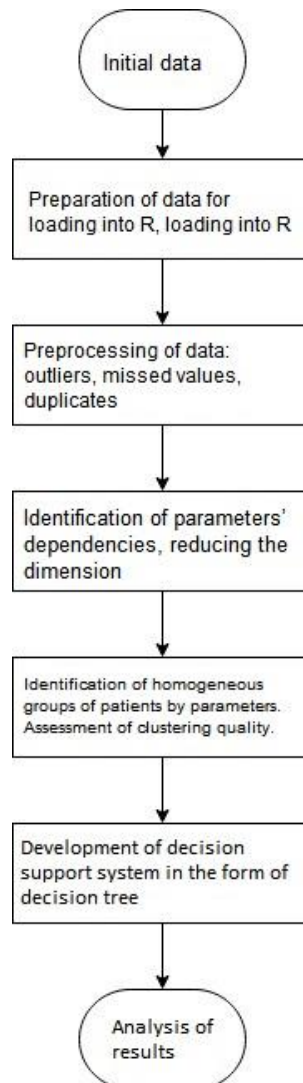


Figure 5 - Work flow diagram

3.2 Primary data processing

A particularly important aspect in analysing the data is the preparation of the initial values for the upcoming work, because the omission of this step will make further work senseless due to the distorted results. The initial data for viewing are in the Excel spreadsheet provided by Tomsk doctors (Annex A).

The initial data are checked for formatting, so that R could read them well, loaded into the R development environment and edited, namely: all columns are reduced to the same type of variables, missing values are filled in by the “bagImput” method, and checked for outliers.

3.2.1 Outlier analysis

After downloading the data to the R environment, the first goal is to check the data for outliers, since otherwise they can greatly affect the final result.

For clarity, box plots are developed for each indicator separately using the `boxplot()` function implemented in the R language.

As an example, the images of only two box plots are given here, the rest are shown in Annex B.

```
boxplot(MedDix_1$Rin1, main='Rin1')
```

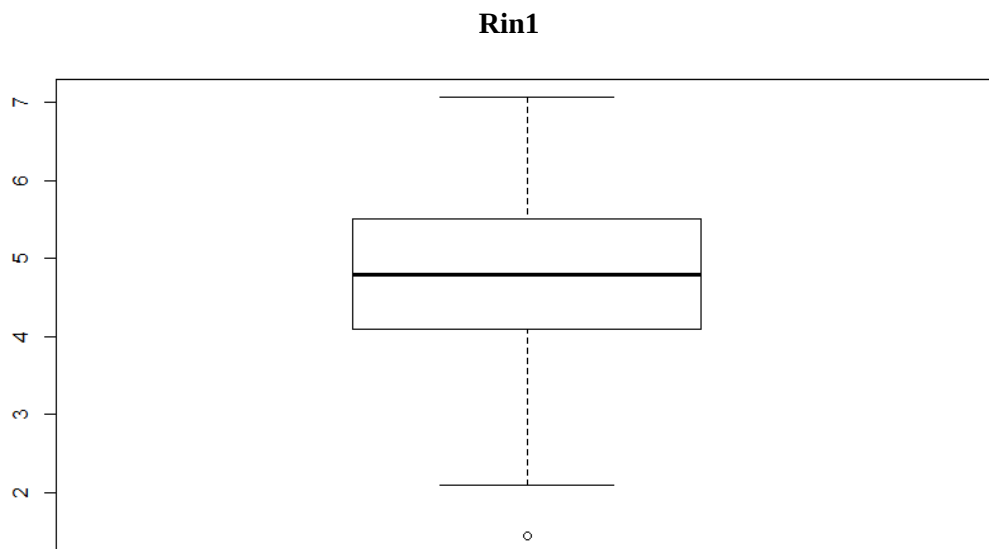


Figure 6 – Box plot for Rin1

```
boxplot(MedDix_1$Rex1, main='Rex1')
```

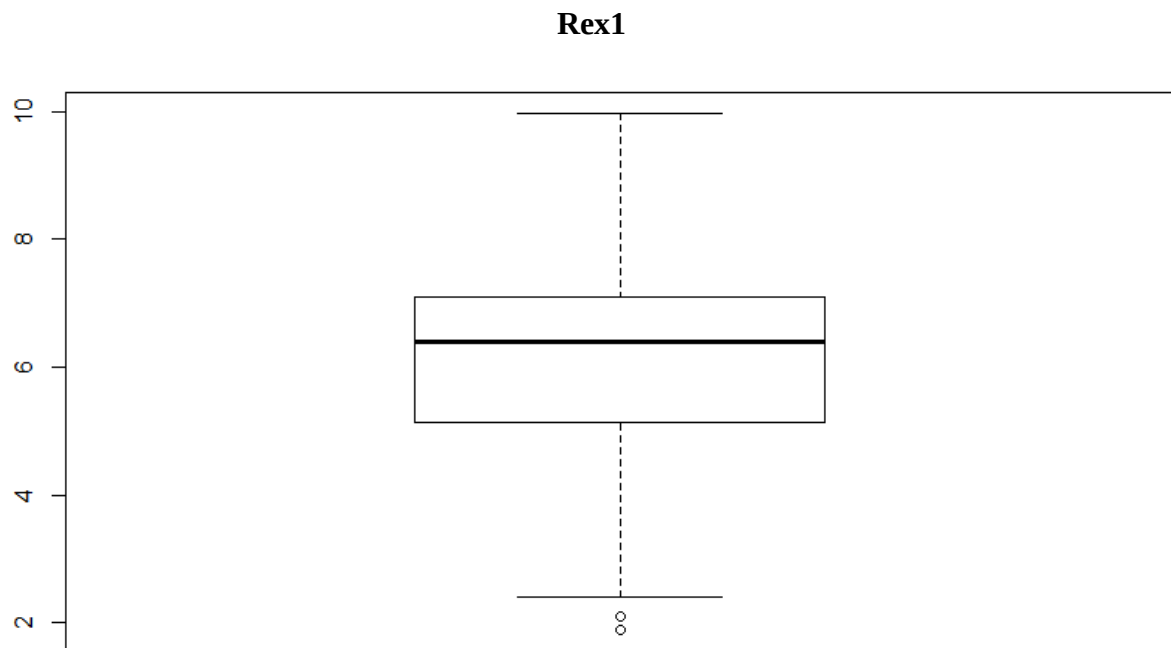


Figure 7 – Box plot for Rex1

Having analysed the obtained box plots, we see that outliers are really present but in small quantity, and furthermore they are within acceptable limits, therefore, these “abnormal” points cannot be deleted.

3.2.2 Recovering missing values

The script language R with the command `summary()` allows viewing general information about all numeric columns, namely, the presence of missing values being of interest for us. After viewing we can see that 2 columns out of 48 have missing values: Weight and MFR50_2.

Weight	MFR50_2
Min. : 34.00	Min. :13.50
1st Qu.: 56.50	1st Qu.:40.30
Median : 65.00	Median :54.15
Mean : 66.03	Mean :53.04
3rd Qu.: 75.50	3rd Qu.:66.42
Max. :105.00	Max. :96.50
NA's :8	NA's :1

Figure 8 – Presence of missing values

	Group №	Sex	Age	Height	Weight	PMV1	PMV2
1	1	2	18	168	55	11.50	9.00
2	1	2	48	154	NA	10.30	7.70
3	1	2	31	170	58	8.20	8.50
4	1	2	36	160	62	7.90	7.70
5	1	2	40	167	85	11.40	12.70
6	1	2	48	158	58	10.00	9.60
7	1	1	32	181	66	8.80	8.00
8	1	2	41	164	NA	9.00	6.50
9	1	2	49	156	NA	7.50	7.70
10	1	2	43	156	52	9.32	8.92

Figure 9 – Fragment with missing values

Next, we proceed to recover the missing values in the data array. To fill in the missing values, the following sequence of scripts should be executed:

```
> pPbI <- preProcess(MedDix_1[, 5:6], method = 'bagImpute')
> MedDix_1[, 5:6] <- predict(pPbI, MedDix_1[, 5:6])
> pPbI <- preProcess(MedDix_1[, 22:23], method = 'bagImpute')
> MedDix_1[, 22:23] <- predict(pPbI, MedDix_1[, 22:23])
```

These scripts allow filling in the missing values using the “bagImpute” method.

```
Weight      MFR50_2
Min.       : 34.00 Min.       :13.50
1st Qu.    : 57.00 1st Qu.    :40.30
Median     : 62.14 Median     :54.10
Mean       : 65.66 Mean       :53.05
3rd Qu.    : 74.00 3rd Qu.    :66.35
Max.       :105.00 Max.       :96.50
```

Figure 10 - The absence of missing values after filling in

	Group №	Sex	Age	Height	Weight	PMV1	PMV2
1	1	2	18	168	55.00000	11.50	9.00
2	1	2	48	154	62.13899	10.30	7.70
3	1	2	31	170	58.00000	8.20	8.50
4	1	2	36	160	62.00000	7.90	7.70
5	1	2	40	167	85.00000	11.40	12.70
6	1	2	48	158	58.00000	10.00	9.60
7	1	1	32	181	66.00000	8.80	8.00
8	1	2	41	164	61.25138	9.00	6.50
9	1	2	49	156	59.48977	7.50	7.70
10	1	2	43	156	52.00000	9.32	8.92

Figure 11 - The fragment with the absence of missing values after filling-in

As we can see in these images, the missing values before filling-in are displayed with the symbols “NA”. After filling-in, numerical values appeared that

correspond to the intervals of admissible values in all columns. Based on it a conclusion may be drawn that the filling-in of missing values by the “bagInput” method is successful.

3.3 Reducing dimension

When performing the work, it is found that the number of columns with different medical values exceeds the number of lines, where each line is a record of one patient. In these conditions, the work cannot be continued since the clustering methods poorly perform in case of such types of tables. To solve this problem, a correlation analysis is performed using the Spearman correlation factor in order to identify the dependencies of the parameters and reduce the dimension. Using this method, we manage to reduce the number of columns.

To build a section of the correlation matrix, the following sequence of actions should be performed:

```
CR <- cor(MedDix_1_1_Short, method = "spearman")
CR_0_0 <- as.data.frame.matrix(CR)
CR_1_0 <- CR_0_0[1:12, 1:12]
CR_1_0_0 <- as.matrix(CR_1_0)
corrplot(CR_1_0_0, method = "number")
```

Further in Figures 3.8-3.10 fragments of the developed correlation matrix may be built. The entire correlation matrix is attached in Annex B.

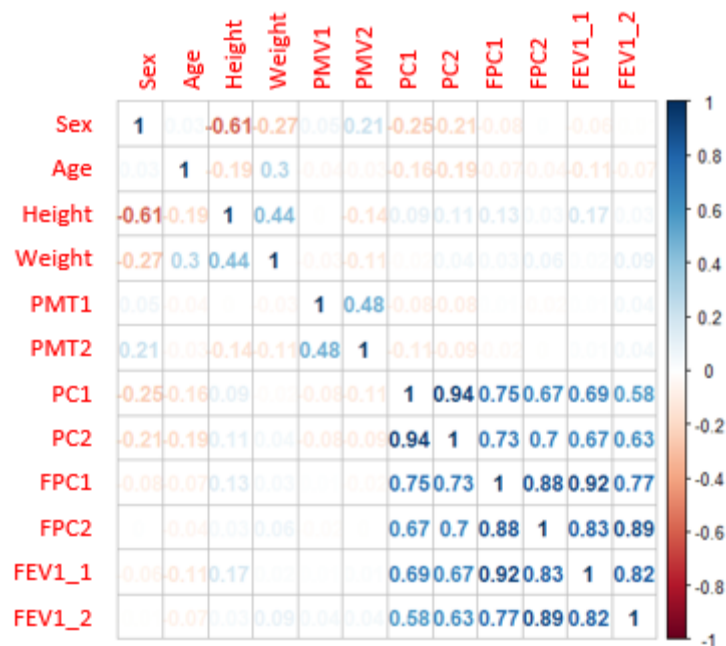


Figure 12 - Fragment 1 of the correlation matrix

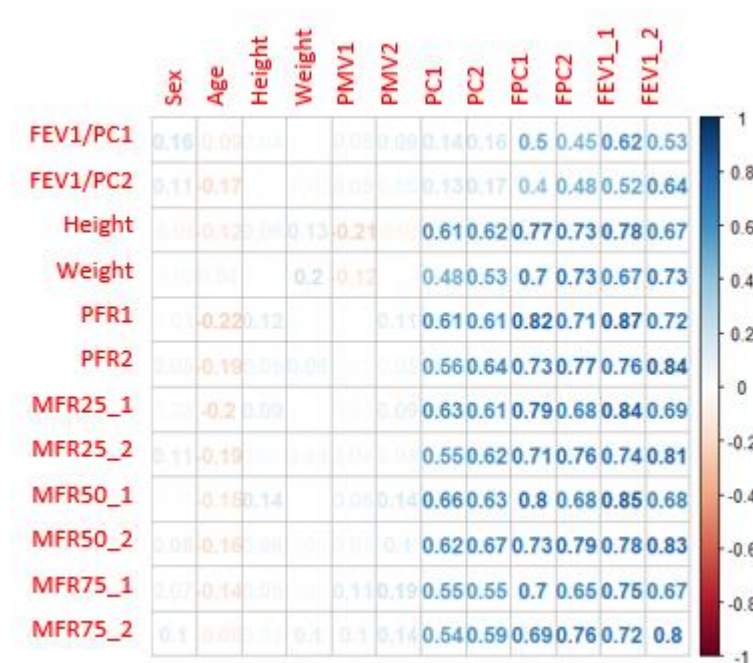


Figure 13 - Fragment 2 of the correlation matrix

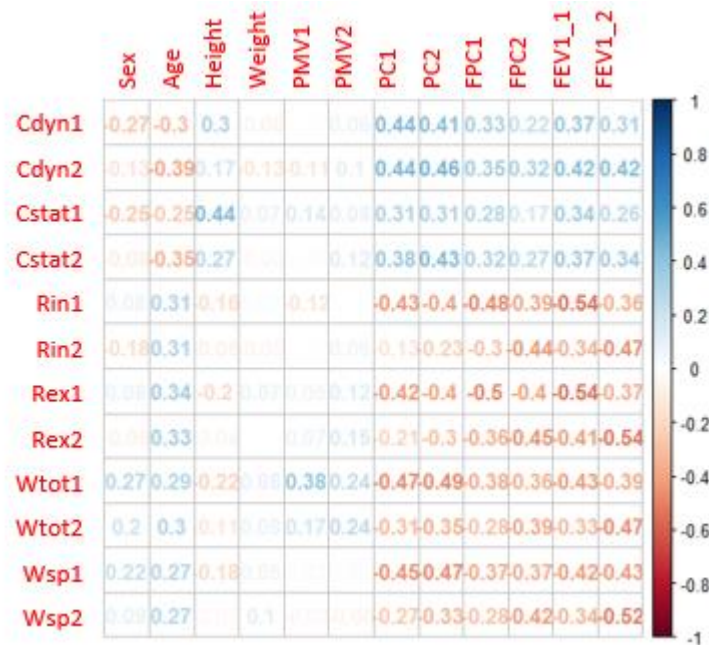


Figure 14 - Fragment 3 of the correlation matrix

3.4 Detecting homogeneous groups of patients

Since the classification of asthma patients into 4 groups proposed by Tomsk medical professionals is experimental, first of all, it should be demonstrated whether patients can be divided into 4 subgroups. To this end, the clustering method is selected.

After the previous manipulations, we can proceed to cluster analysis. Next, we perform the cluster analysis by the *k* means method for 4 clusters using the *kmeans* function:

```
> clus <- kmeans(MedDix_1_1, 4)
```

The clustering result is given in fig. 3.11.

```

> clus
K-means clustering with 4 clusters of sizes 23, 23, 24, 13

Cluster means:
№ группы  Пол  Возраст  Рост  Вес  МОД1  МОД2  ЖЕЛ1  ЖЕЛ2  ФЖЕЛ1  ФЖЕЛ2  ОФБ1_1  ОФБ1_2  ОФБ1/ЖЕЛ1
1 2.130435 1.608696 37.78261 166.3478 63.60556 8.400000 7.197826 67.25261 67.97087 60.17348 65.01826 54.42739 59.67391 69.84739
2 2.434783 1.695652 35.34783 168.4783 64.90250 8.580000 7.957391 86.96739 89.74783 84.50870 88.88696 79.34783 85.93478 87.74174
3 2.125000 1.541667 38.25000 163.4583 66.86490 8.207500 7.589167 83.41042 83.95417 73.67500 78.85833 67.60833 76.36250 75.13125
4 3.000000 1.461538 31.38462 169.8462 69.61564 8.139231 7.326923 96.23077 103.52308 106.28923 114.16923 100.59385 110.87692 95.95385
ОФБ1/ЖЕЛ2  МВЛ1  МВЛ2  ПОС1  ПОС2  МОС25_1  МОС25_2  МОС50_1  МОС50_2  МОС75_1  МОС75_2  Сdyn1  Сdyn2  Сstat1
1 75.06043 44.25174 45.48087 37.81913 41.15957 35.81478 38.48826 30.19826 31.45174 25.95957 26.56609 0.09047826 0.08730435 0.1325652
2 92.48652 74.55043 80.34087 69.09130 75.00870 66.74348 74.64348 59.79130 65.83478 51.87826 56.98261 0.13334783 0.14234783 0.1635652
3 82.97875 69.73333 73.67500 53.34458 58.46333 50.85417 56.14583 45.12083 48.85427 42.43750 44.54167 0.10954167 0.11162500 0.1358333
4 100.20615 95.31923 105.45385 77.62308 86.43077 74.85615 82.60769 65.37692 76.36308 58.76923 68.54615 0.12046154 0.12738462 0.1520769
Сstat2  РВд1  РВд2  РВЫд1  РВЫд2  ВОБш1  ВОБш2  УУд1  УУд2  УН.эл1  УН.эл2  УЭл1  УЭл2
1 0.1178261 5.582609 4.443478 7.395652 5.715652 0.6857174 0.4971304 0.08182609 0.06786957 0.6210000 0.4587826 0.3151304 0.2685217
2 0.1696087 3.980000 3.293478 5.344348 4.125217 0.3925217 0.3211739 0.04656522 0.04130435 0.3281304 0.2819130 0.2193913 0.1859130
3 0.1343333 4.880000 4.084167 6.540417 4.947083 0.5783333 0.4665000 0.07037500 0.06283333 0.5105000 0.4190000 0.2725000 0.2513750
4 0.1546923 3.564615 2.837692 4.277692 3.413846 0.4357692 0.3058462 0.05230769 0.04061538 0.3740769 0.2652308 0.2171538 0.1583077
W МОД10_1 W МОД10_2 W МОД15_1 W МОД15_2 W МВЛобш1 МВЛобш2 МВЛуд1 МВЛуд2
1 0.8122174 0.6238696 1.3847826 1.1151739 13.52570 14.37026 0.3094783 0.3167826
2 0.4716522 0.3976957 0.7549565 0.6391304 21.45604 24.55304 0.2713478 0.3148261
3 0.6710833 0.5879167 1.1159583 0.9225000 23.11500 26.22325 0.3380000 0.3597917
4 0.5033846 0.3700000 0.9423077 0.6584615 22.44423 25.13462 0.2256154 0.2366923

Clustering vector:
[1] 2 3 2 2 2 1 3 1 4 2 2 1 3 2 1 1 3 4 1 3 3 2 3 2 1 1 3 4 1 3 1 3 1 3 2 4 3 3 1 3 1 1 1 1 3 3 3 1 3 2 4 1 4 3 3 1 1 3 4 3 3 2 3 1 2 2
[67] 1 2 1 2 2 2 4 2 2 2 2 4 4 4 2 4 4 2

within cluster sum of squares by cluster:
[1] 66134.45 48840.25 55609.73 44635.44
(between_SS / total_SS = 60.1 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"

```

Figure 15 - Script operation result

For a more convenient visual perception, a typical diagram for the cluster analysis is developed.

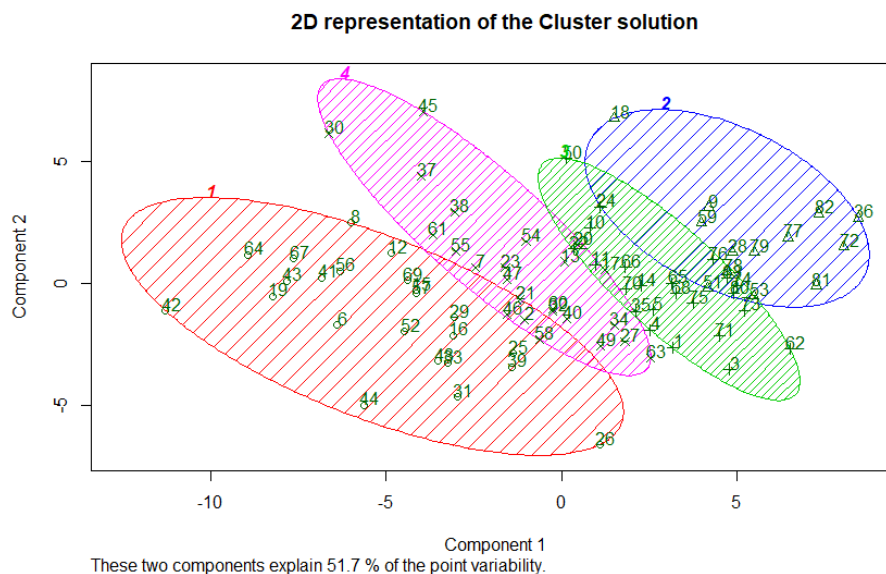


Figure 16 - Cluster analysis diagram

Next, analyse the reliability of the obtained results in table 3.

Table 3 - Result reliability analysis

Cluster number	Data reliability percentage, %
Cluster 1 - PIBA	88.46
Cluster 2 - NPBA	79.20
Cluster 3 - SPBA	72.22
Cluster 4 - PsD	85.71

To assess the clustering quality, the Dunn index is used.

```
> dunn_clust<-kmeans(MedDix_1_1,4)
> Dist<-dist(MedDix_1_1, method='euclidean')
> cluster<-dunn_clust$cluster
> dunn(Dist, cluster)
```

The value of the Dunn index is 0.268213. This indicates that this data array clustering is quite good.

3.5 Building a DSS in the form of a decision tree

After the cluster analysis has demonstrated that patients may be divided into 4 subgroups, the next stage of the task may be performed. The conditions for assigning the patients to different groups should be detected with building a decision support system in the form of a decision tree.

There are several basic algorithms that are used to build trees. In this work, the function “rpart” from the package “rpart” for the language R is used to create a decision tree. This function performs a recursive selection for each next node. It selects such dividing values which lead to the minimum sum of squares of intragroup deviations D_t for all t nodes of the tree.

```
dtm<- rpart(Groups~., MedDix_1_1, method="class")

> dtm
n= 83

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 83 54 3 (0.28915663 0.21686747 0.34939759 0.14457831)
 2) Rbд1>=3.15 70 41 3 (0.34285714 0.24285714 0.41428571 0.00000000)
   4) Rbд2< 4.935 55 31 1 (0.43636364 0.30909091 0.25454545 0.00000000)
     8) MOC50_2>=45.095 35 16 1 (0.54285714 0.11428571 0.34285714 0.00000000)
       16) MOC25_1< 60.4 18 4 1 (0.77777778 0.11111111 0.11111111 0.00000000) *
       17) MOC25_1>=60.4 17 7 3 (0.29411765 0.11764706 0.58823529 0.00000000) *
     9) MOC50_2< 45.095 20 7 2 (0.25000000 0.65000000 0.10000000 0.00000000) *
   5) Rbд2>=4.935 15 0 3 (0.00000000 0.00000000 1.00000000 0.00000000) *
 3) Rbд1< 3.15 13 1 4 (0.00000000 0.07692308 0.00000000 0.92307692) *
```

Figure 17 - Patient assigning conditions

Using this command, we have built a complete set of conditions for assigning patients to one of four groups. Next, we can build a DSS using the obtained conditions in the form of a decision tree. Using the function "prp()" from the package "rpart.plot".

prp(dtm)

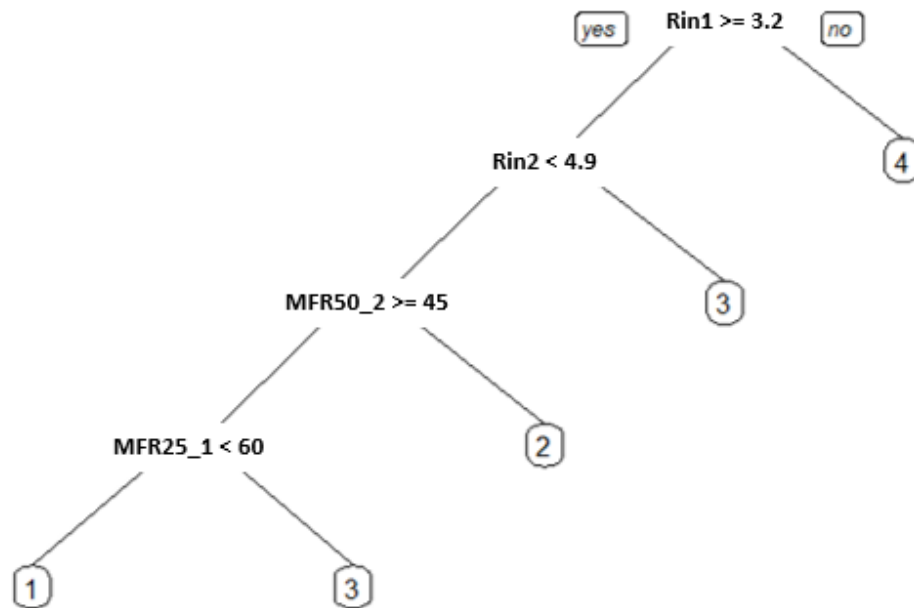


Figure 18 - The resulting decision tree

The resulting tree gives the following results:

1. If $Rin1 < 3.2$, then the group is 4, i.e. BA=PD.
2. If $Rin1 \geq 3.2$ and $Rin2 \geq 4.9$, then the group is 3, BA=BANP.
3. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 < 45$, then the group is 2, BA = BASP.
4. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 \geq 45$ and $MFR25_1 \geq 60$, than the group is 3, BA = BANP.
5. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 \geq 45$ and $MFR25_1 < 60$, than the group 1, BA = BAPI.

Next, analyse the reliability of the obtained results in table 4.

Table 4 - Result reliability analysis

Logical rule	Data reliability percentage, %
1. If $Rin1 < 3.2$, then the group is 4, i.e. BA=PD.	92.31
2. If $Rin1 \geq 3.2$ and $Rin2 \geq 4.9$, then the group is 3, BA=BANP.	100
3. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 < 45$, then the group is 2, BA = BASP.	74.07
4. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 \geq 45$ and $MFR25_1 \geq 60$, than the group is 3, BA = BANP.	73.91
5. If $Rin1 \geq 3.2$ and $Rin2 < 4.9$ and $MFR50_2 \geq 45$ and $MFR25_1 < 60$, than the group 1, BA = BAPI.	81.82

As can be seen from table 4, the developed decision tree has a rather high data reliability indicator.

3.6 Analysis of the results

In the course of the assignment, the following actions are performed within the framework of the developed action algorithm:

1. Data preparation;
2. Data preprocessing;
3. The dependences of parameters are revealed and dimensionality is reduced by correlation analysis;
4. The cluster analysis is performed;
5. The decision tree is built.

In the cluster analysis, the data are divided into 4 clusters, which corresponds to the 4th stated bronchial asthma groups. And a decision tree is built to relate each patient to one of the four groups. Since the accuracy of the results is above 80% at the average, a conclusion can be drawn that the classification of bronchial asthma proposed by Tomsk medical professionals is viable.

Conclusion

As a result of the master's thesis, a support system for research studies of medical diagnostics in the form of a decision tree has obtained, based on the obtained clinical and laboratory parameters. In the development process, the study objects and objectives have considered, the initial data have described. A method for filling-in the missing values, a method for reducing the dimension have selected, types of clustering and a method for building a decision tree have described, and tools for implementing the set problem have selected and described. An algorithm for solving the set problem has developed. The set up problem has fully solved in accordance with the steps described in the algorithm

Приложение Б (справочное)

Диаграммы размаха для столбцов с выбросами

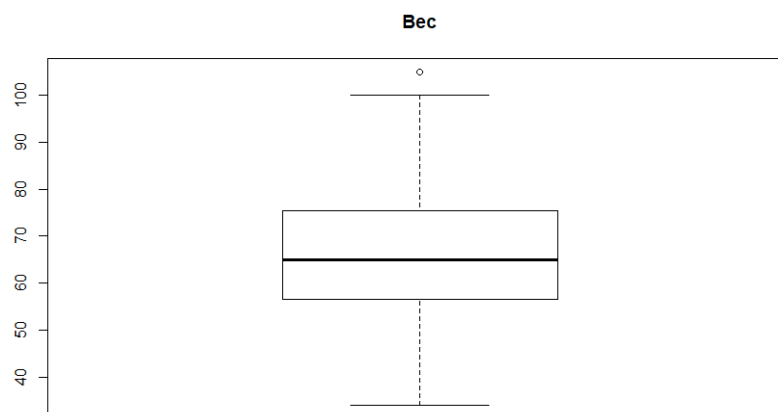


Рисунок 20 – Диаграмма размаха для Вес

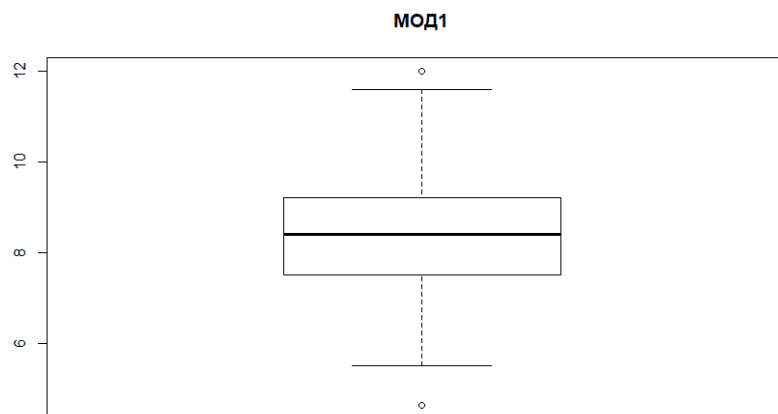


Рисунок 21 – Диаграмма размаха для МОД1

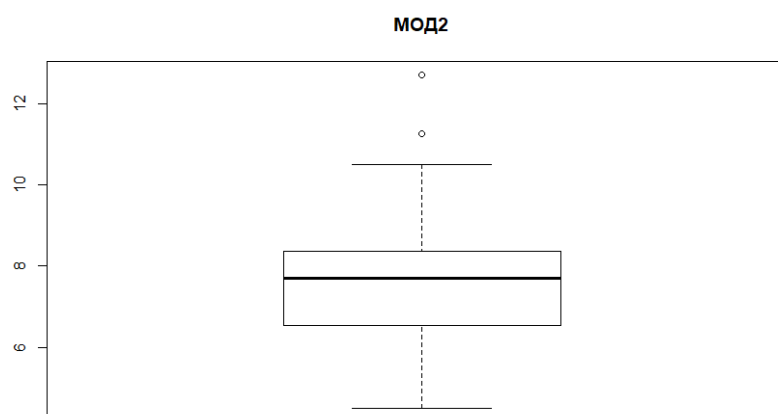


Рисунок 22 – Диаграмма размаха для МОД2

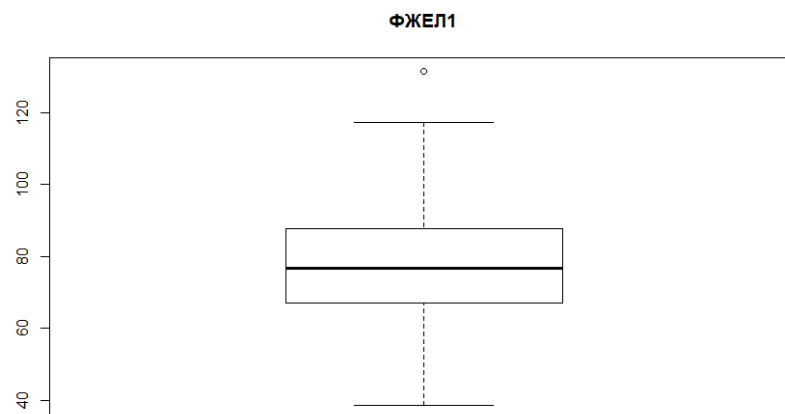


Рисунок 23 – Диаграмма размаха для ФЖЕЛ1

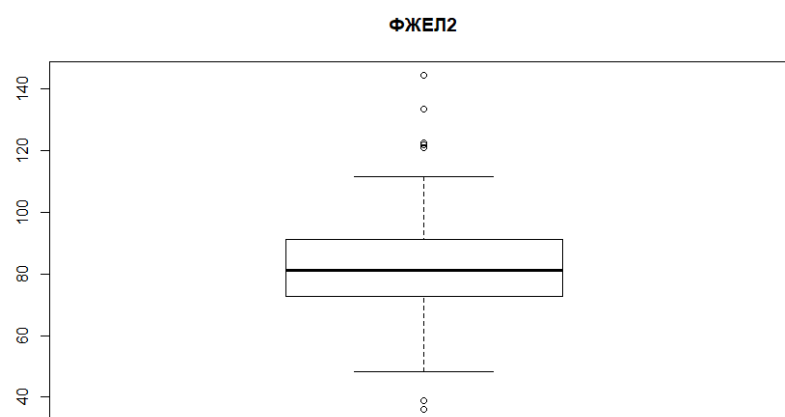


Рисунок 24 – Диаграмма размаха для ФЖЕЛ2

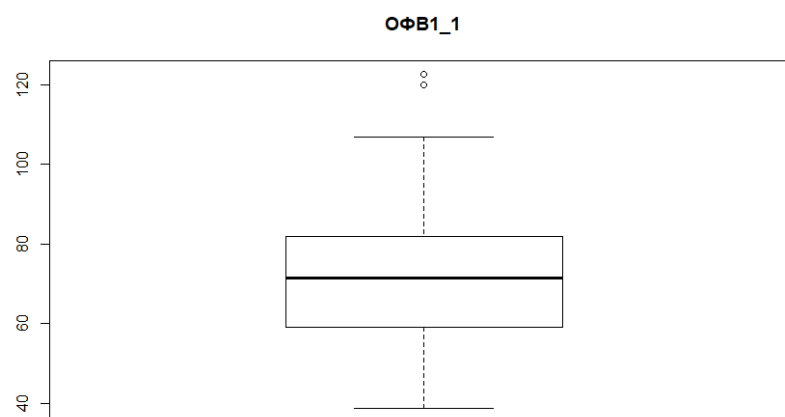


Рисунок 25 – Диаграмма размаха для ОФВ1_1

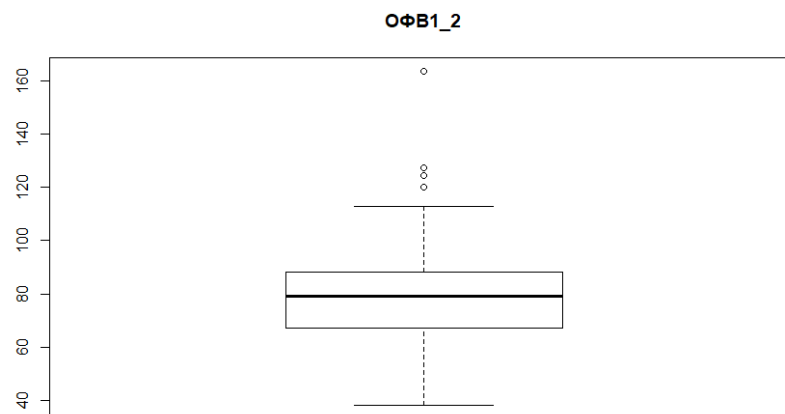


Рисунок 26 – Диаграмма размаха для ОФВ1_2

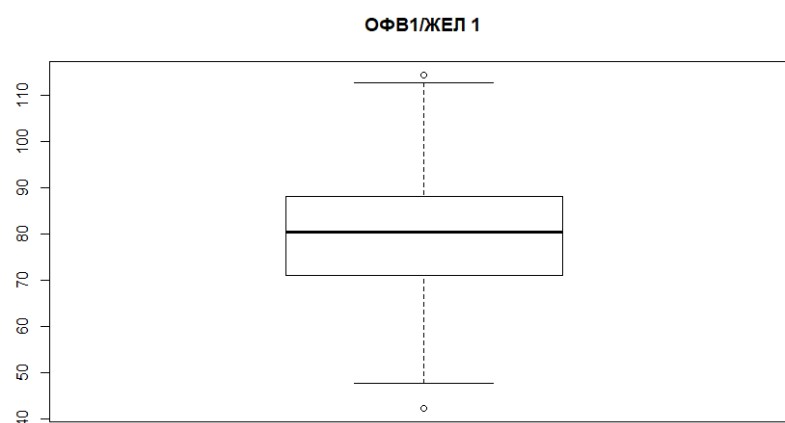


Рисунок 27 – Диаграмма размаха для ОФВ1/ЖЕЛ1

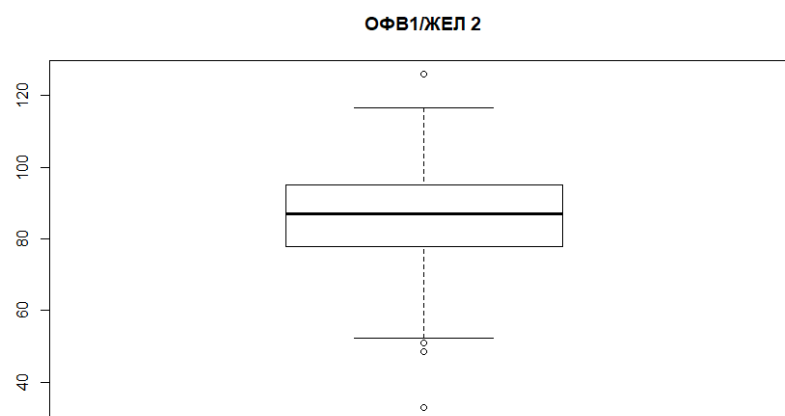


Рисунок 28 – Диаграмма размаха для ОФВ1/ЖЕЛ2

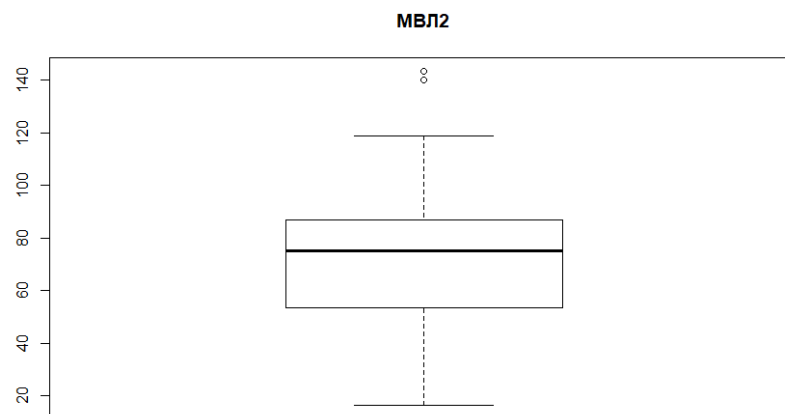


Рисунок 29 – Диаграмма размаха для МВЛ2

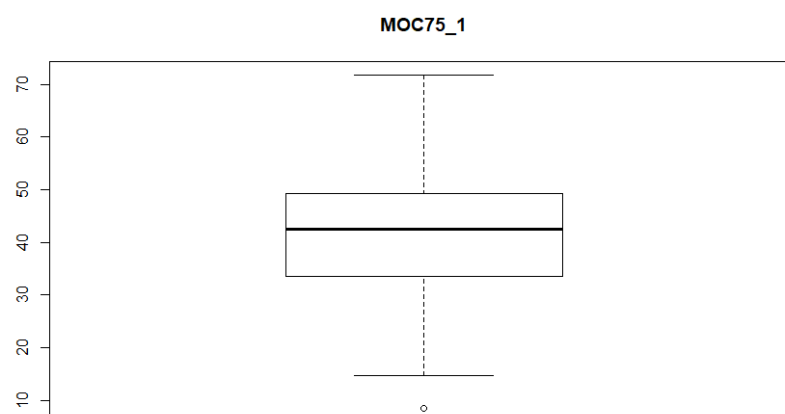


Рисунок 30 – Диаграмма размаха для МОС75_1

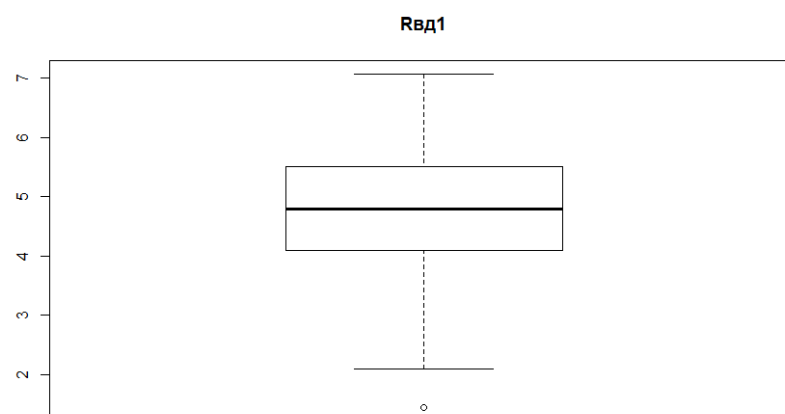


Рисунок 31 – Диаграмма размаха для Рвд1

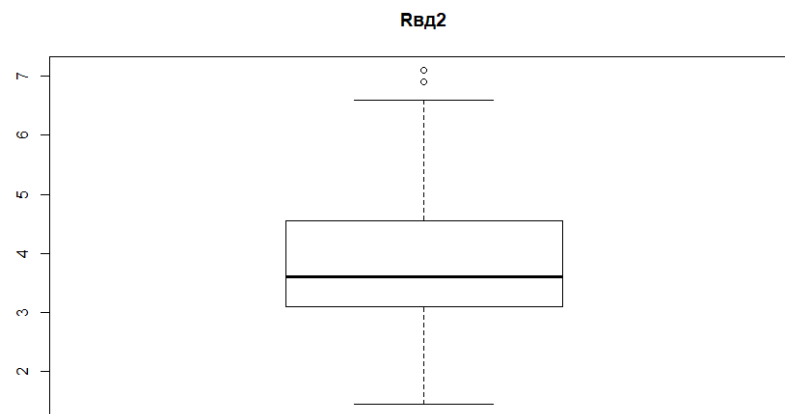


Рисунок 32 – Диаграмма размаха для Rвд2

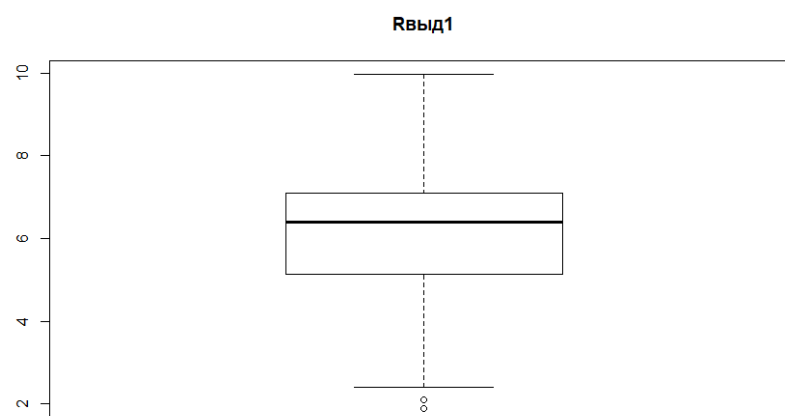


Рисунок 33 – Диаграмма размаха для Rвд1

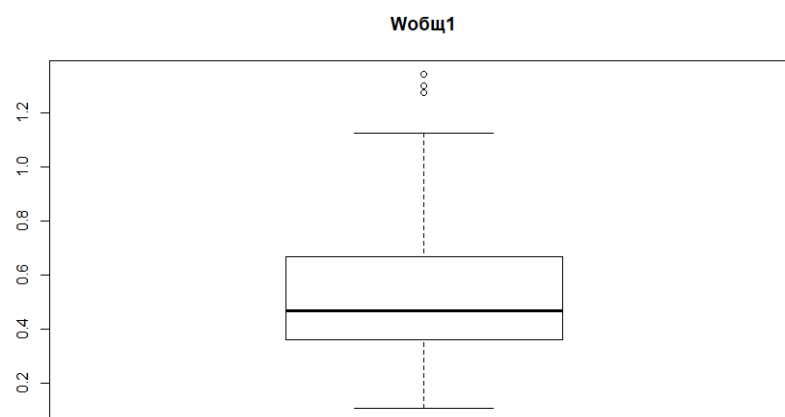


Рисунок 34 – Диаграмма размаха для Rвд2

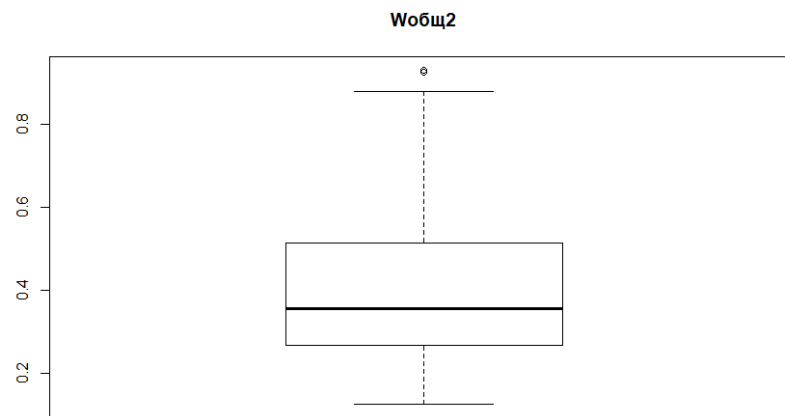


Рисунок 35 – Диаграмма размаха для Wобщ2

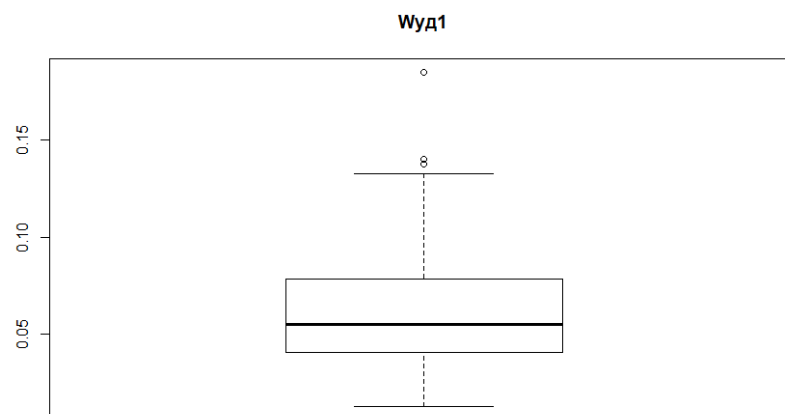


Рисунок 36 – Диаграмма размаха для Wуд1

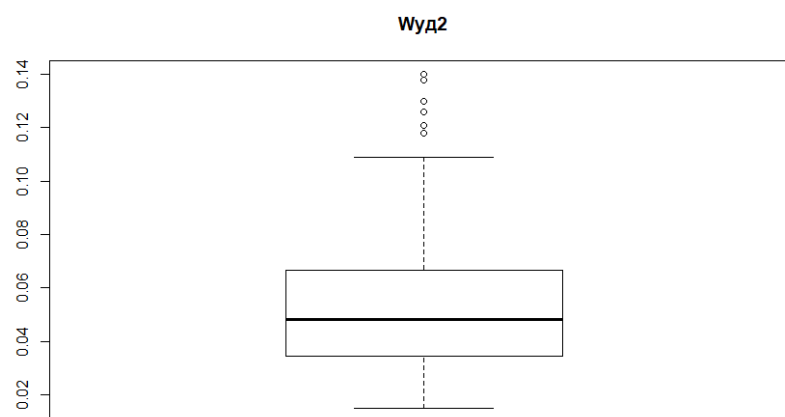


Рисунок 37 – Диаграмма размаха для Wуд2

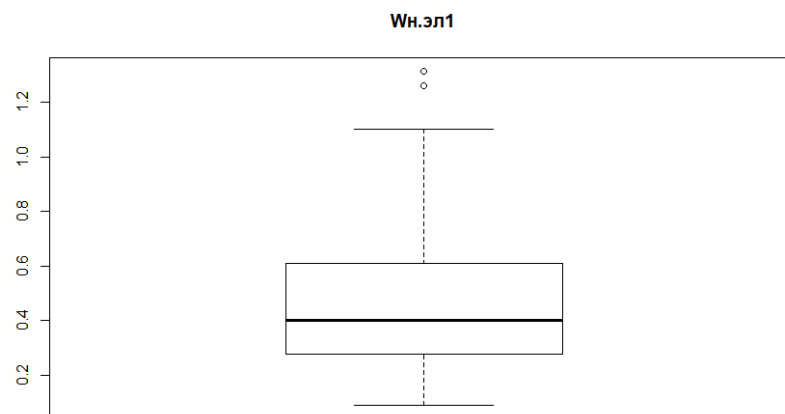


Рисунок 38 – Диаграмма размаха для Wн.эл1

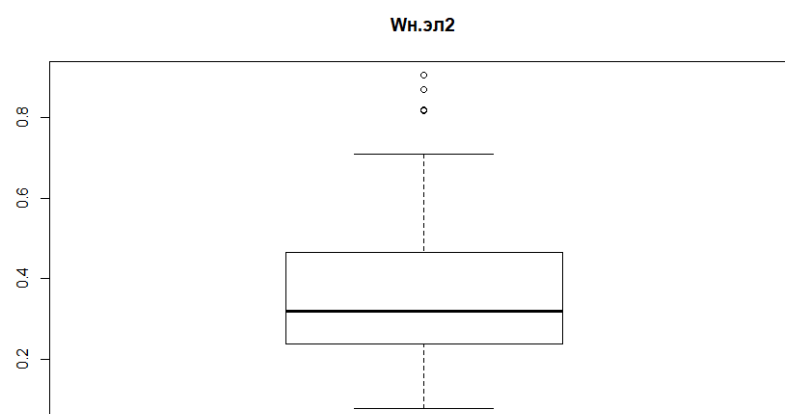


Рисунок 39 – Диаграмма размаха для Wн.эл2

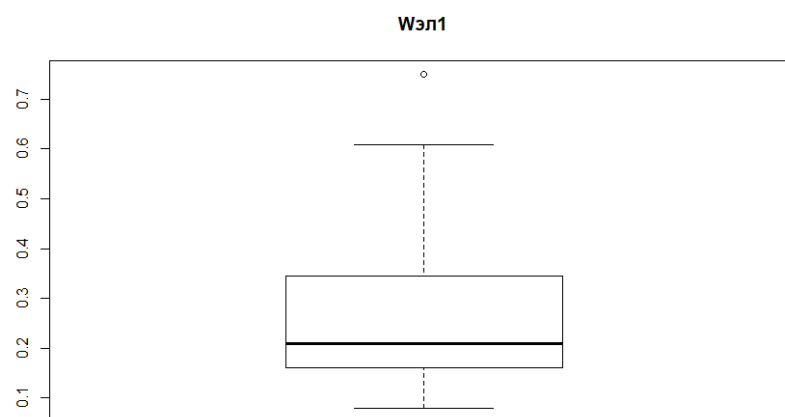


Рисунок 40 – Диаграмма размаха для Wэл1

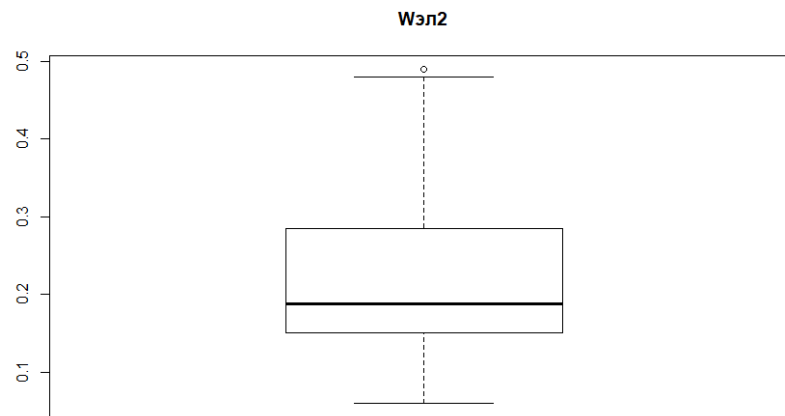


Рисунок 41 – Диаграмма размаха для Wэл2

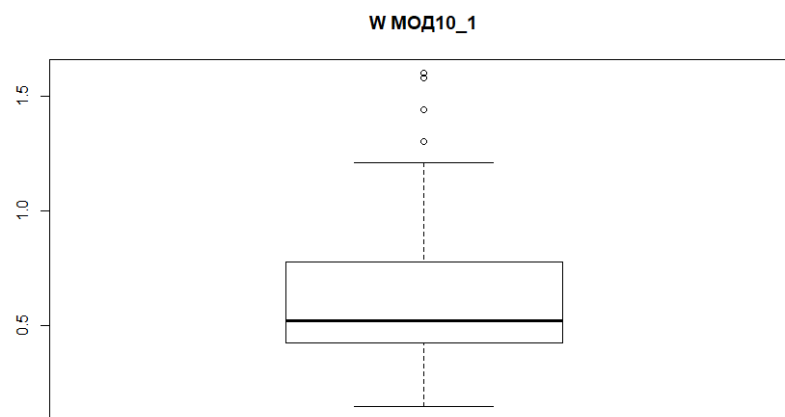


Рисунок 42 – Диаграмма размаха для W МОД10_1

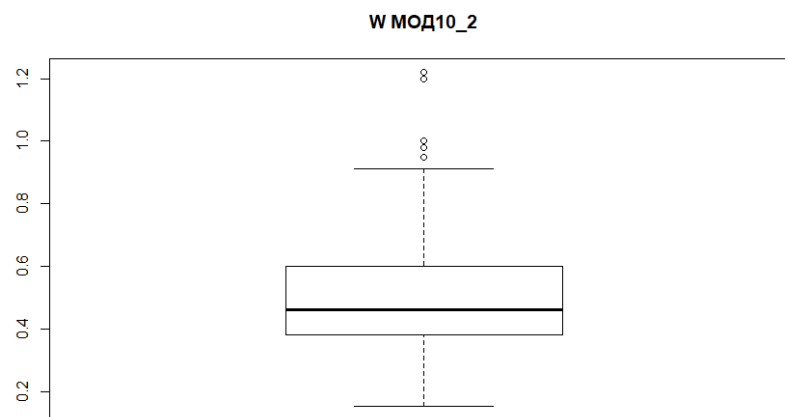


Рисунок 43 – Диаграмма размаха для W МОД10_2

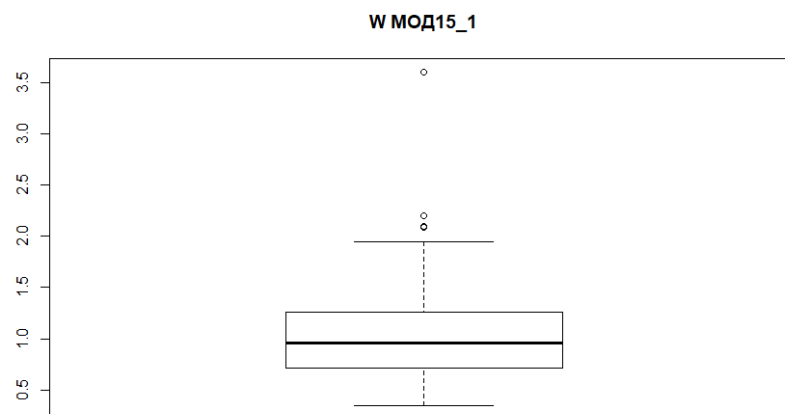


Рисунок 44 – Диаграмма размаха для W МОД15_1

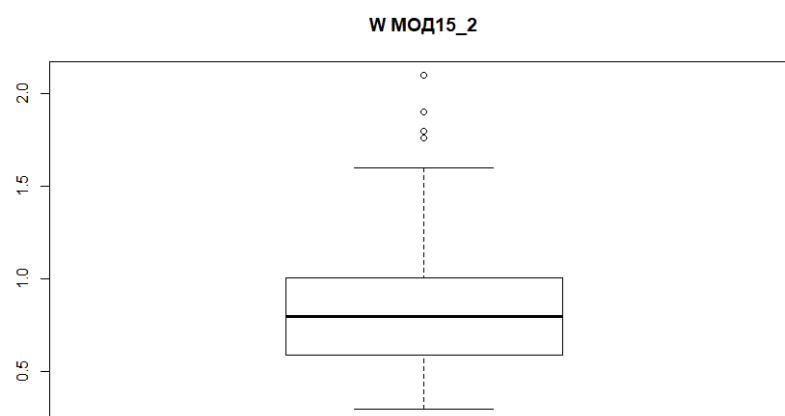


Рисунок 45 – Диаграмма размаха для W МОД15_2

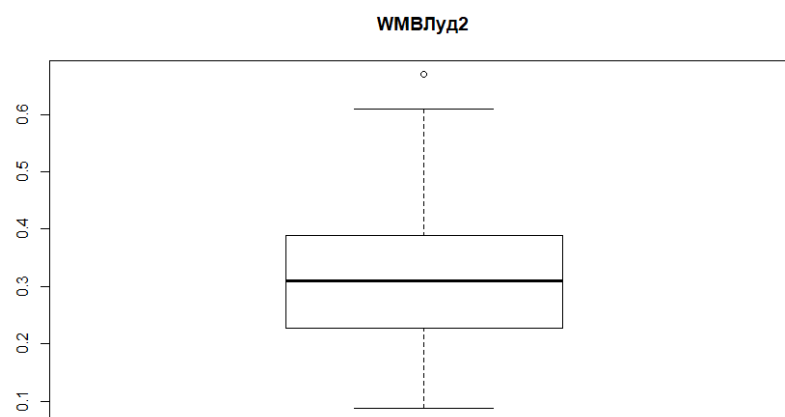


Рисунок 46 – Диаграмма размаха для WMBЛуд2

Приложение В (справочное) Матрица корреляции

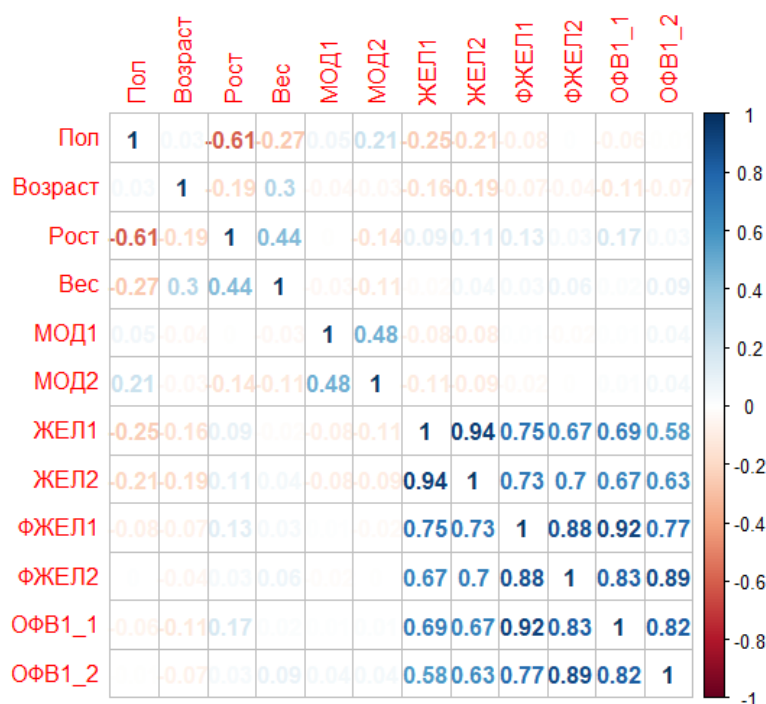


Рисунок 47 – Фрагмент 1 корреляционной матрицы

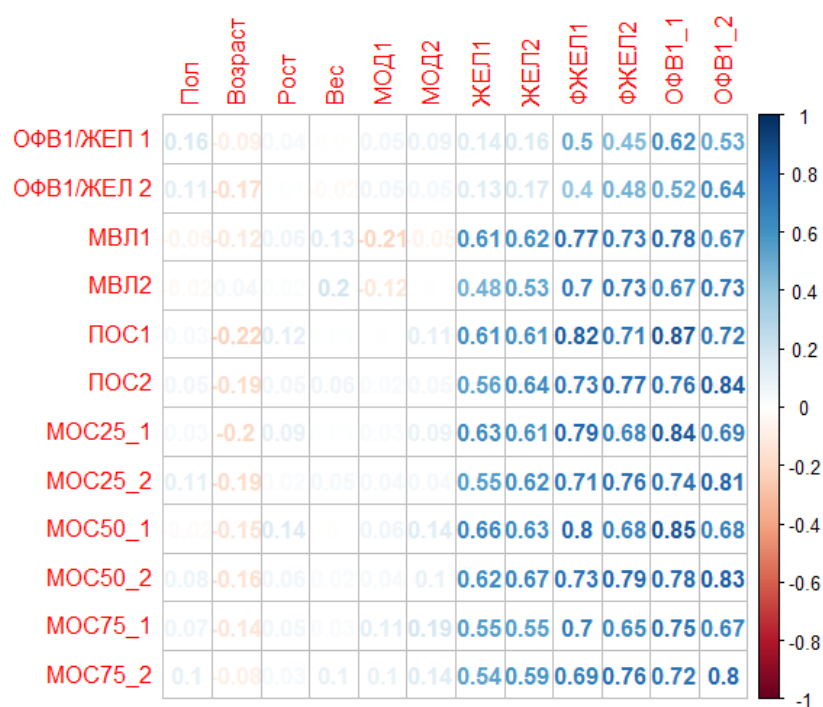


Рисунок 48 – Фрагмент 2 корреляционной матрицы

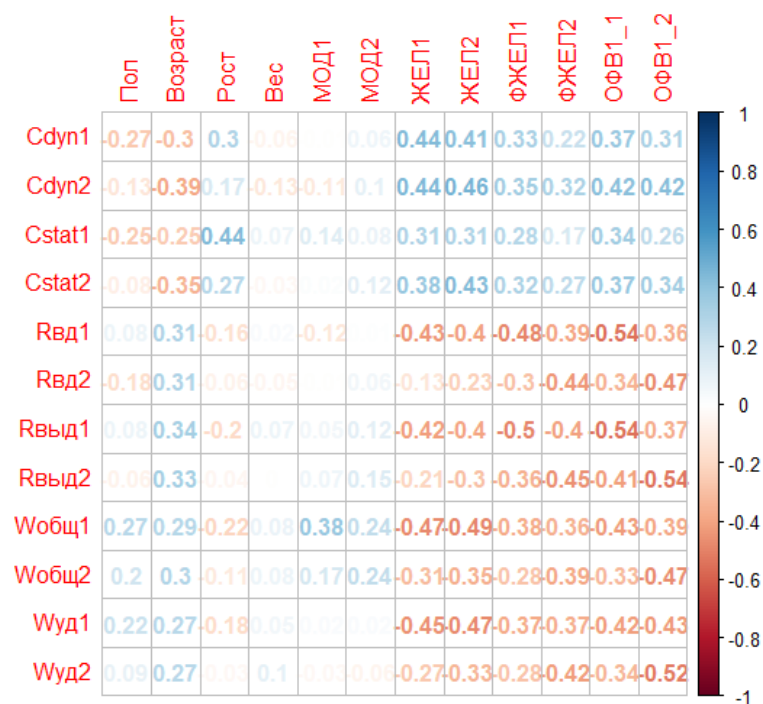


Рисунок 49 – Фрагмент 3 корреляционной матрицы

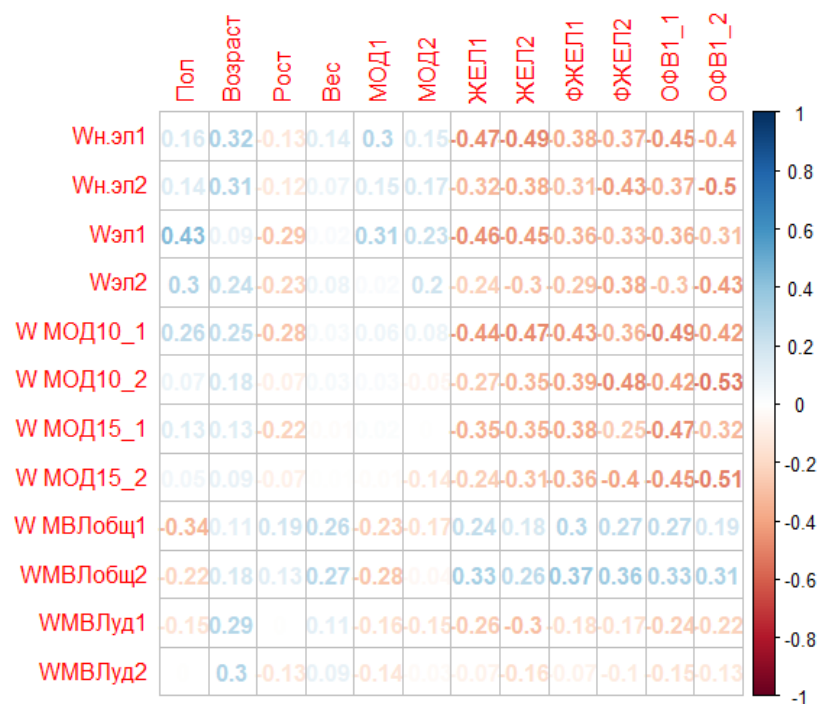


Рисунок 50 – Фрагмент 4 корреляционной матрицы

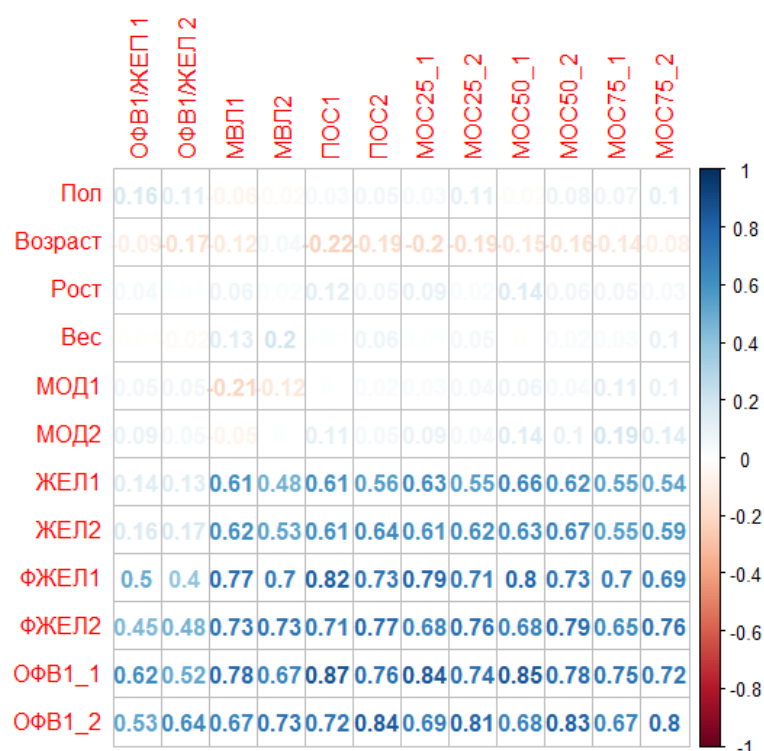


Рисунок 51 – Фрагмент 5 корреляционной матрицы

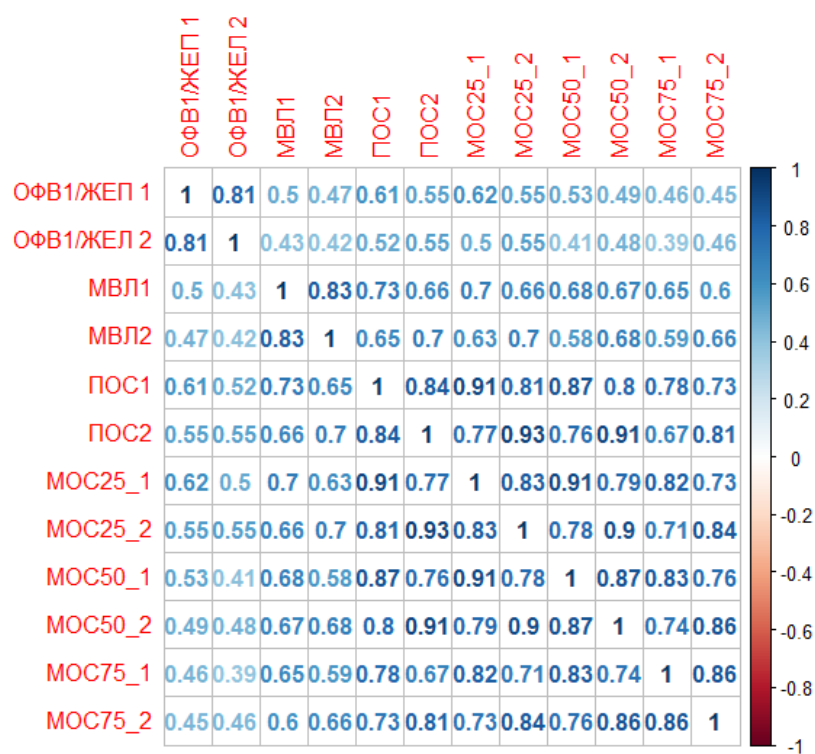


Рисунок 52 – Фрагмент 6 корреляционной матрицы

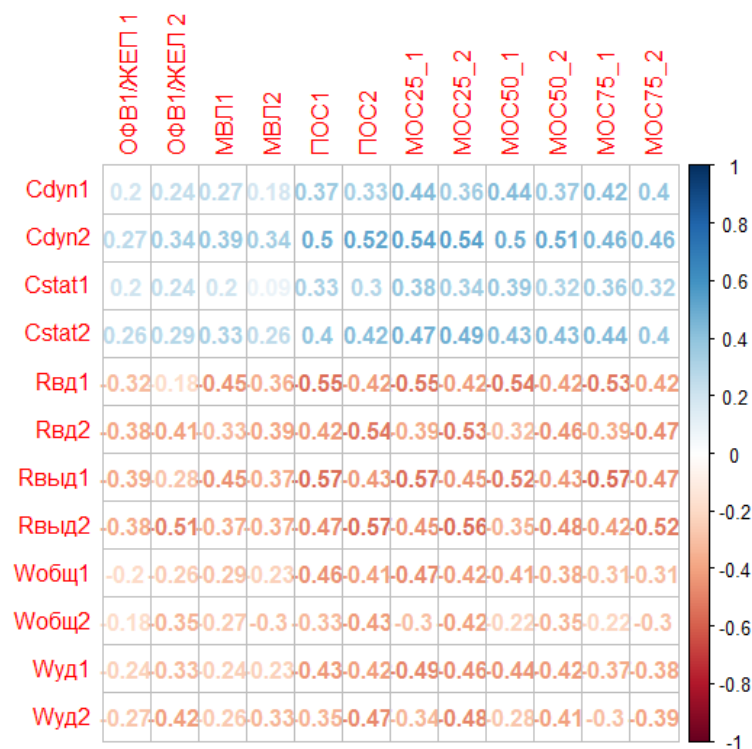


Рисунок 53 – Фрагмент 7 корреляционной матрицы

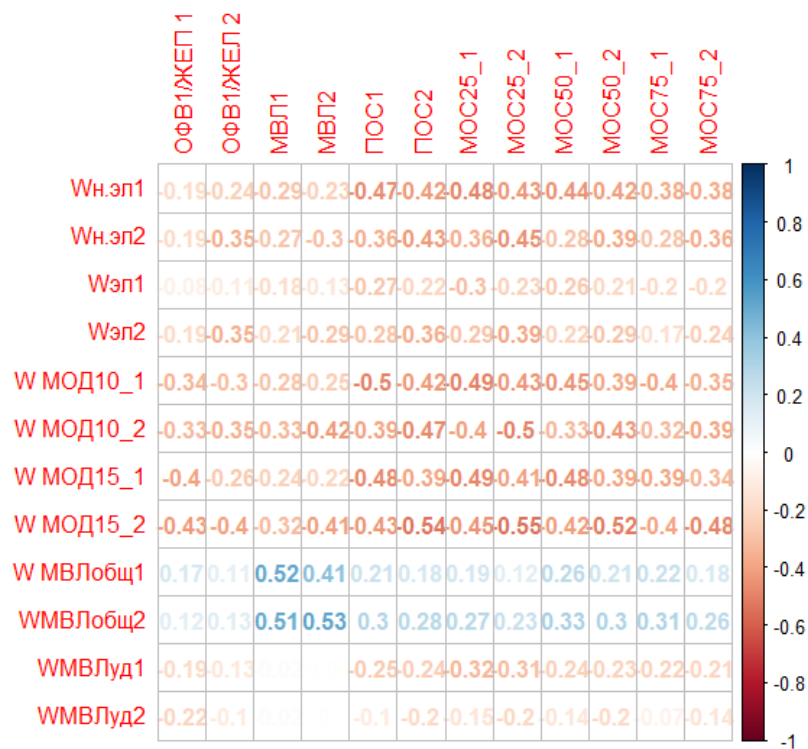


Рисунок 54 – Фрагмент 8 корреляционной матрицы

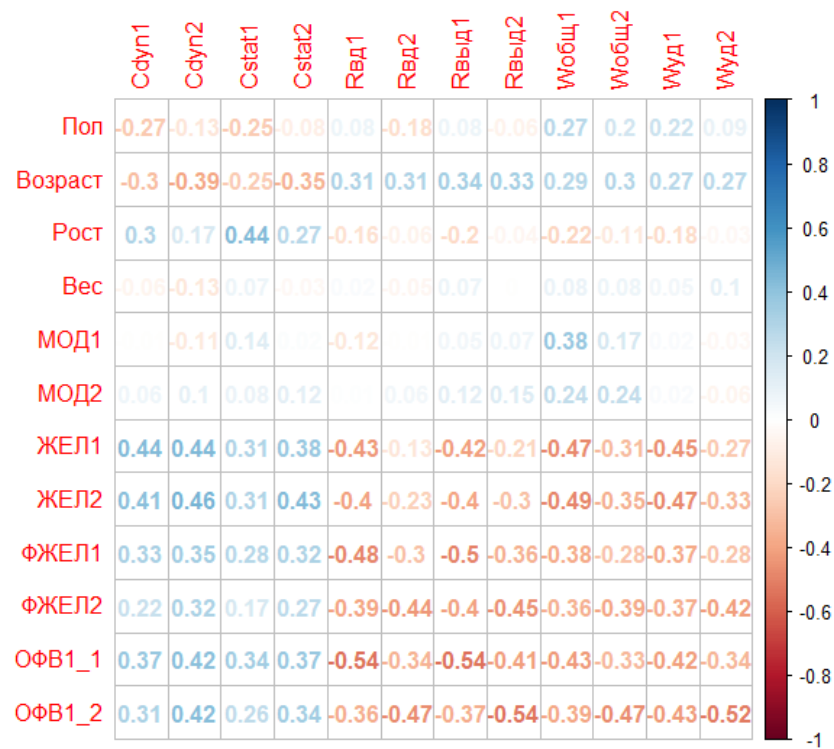


Рисунок 55 – Фрагмент 9 корреляционной матрицы

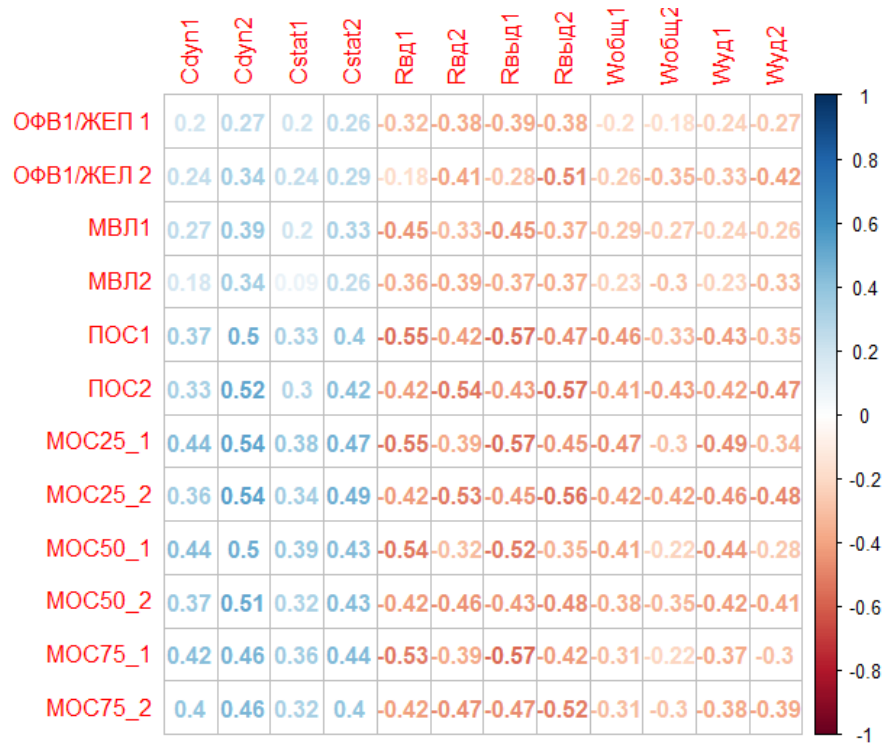


Рисунок 56 – Фрагмент 10 корреляционной матрицы

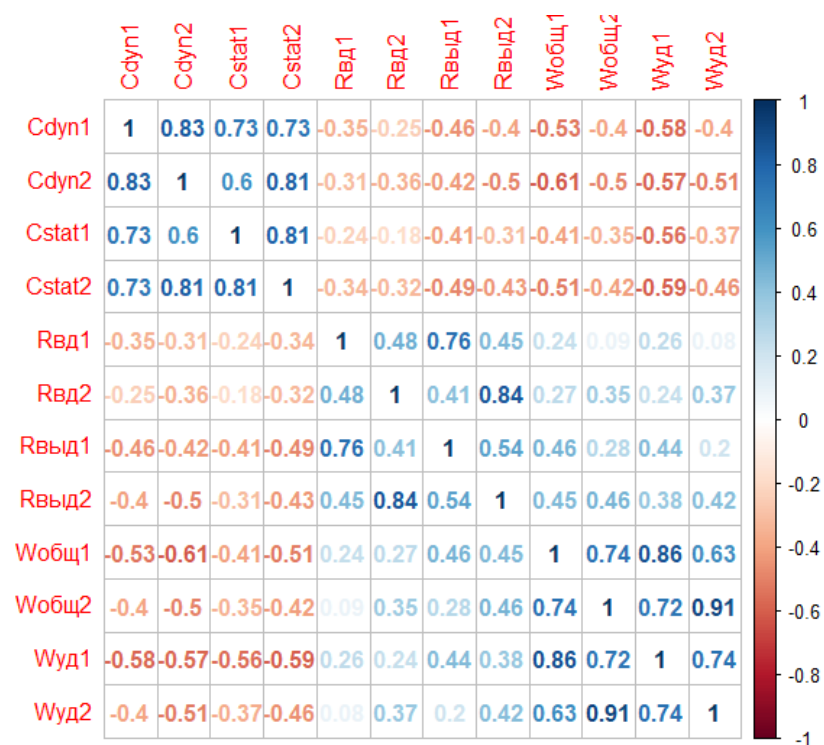


Рисунок 57 – Фрагмент 11 корреляционной матрицы

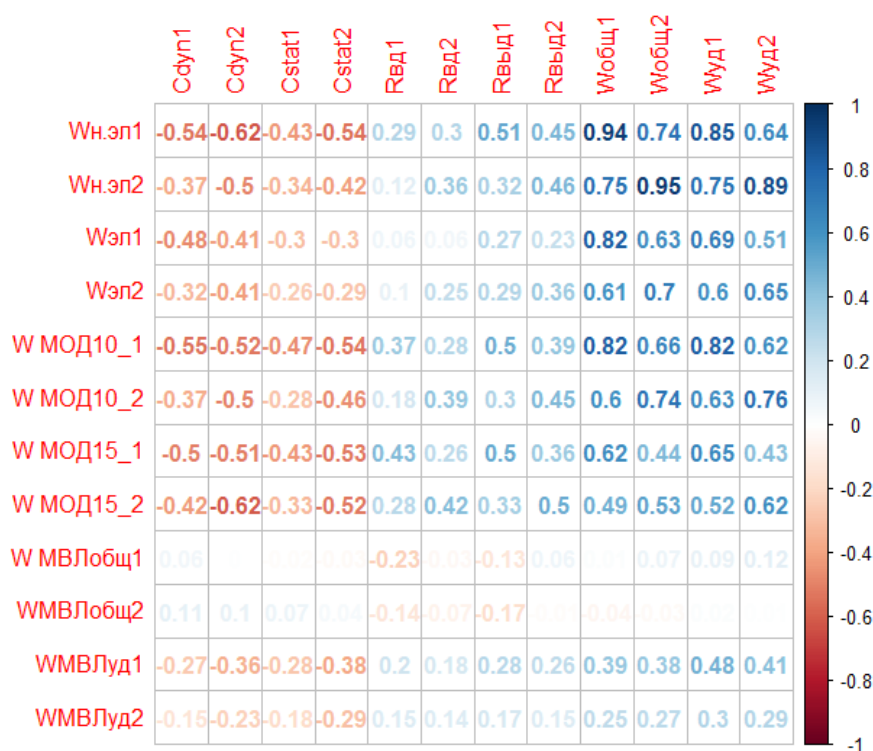


Рисунок 58 – Фрагмент 12 корреляционной матрицы

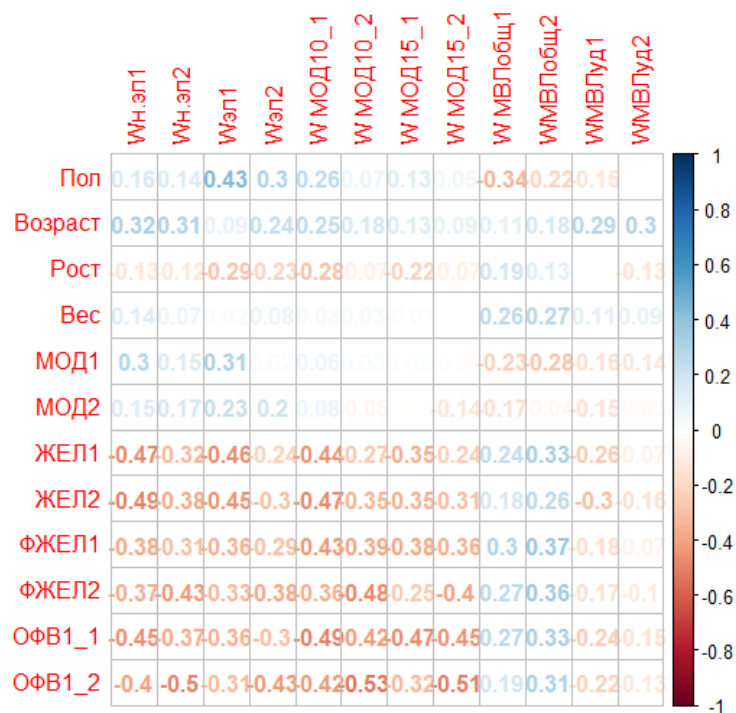


Рисунок 59 – Фрагмент 14 корреляционной матрицы

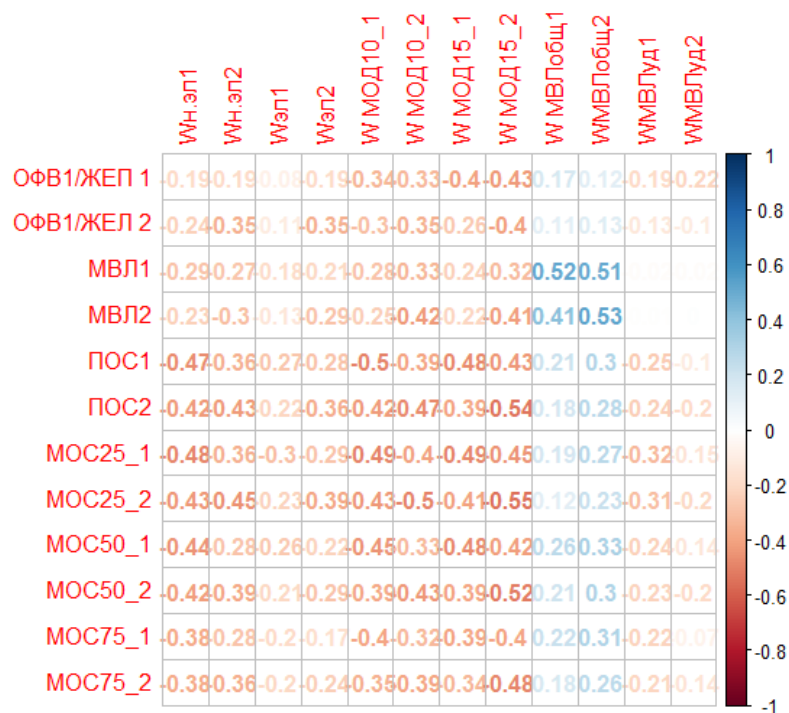


Рисунок 60 – Фрагмент 13 корреляционной матрицы

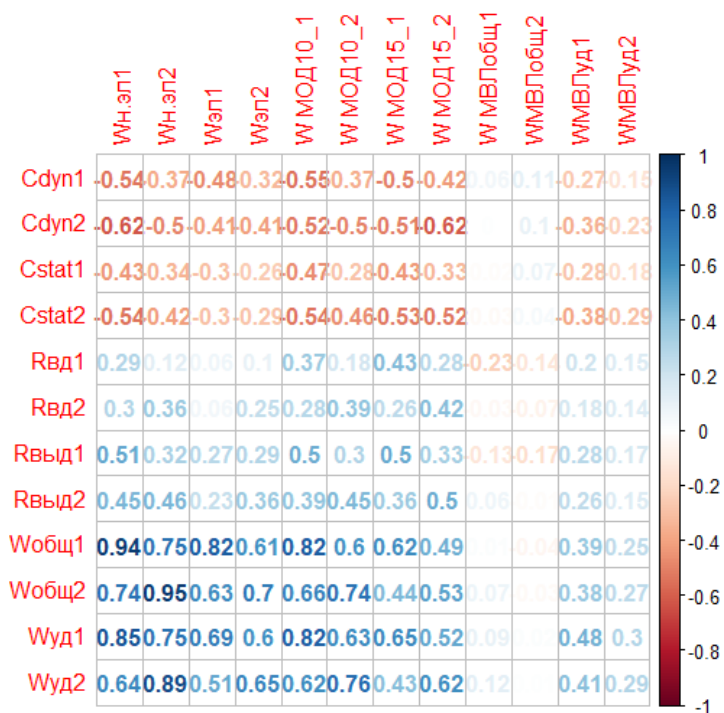


Рисунок 61 – Фрагмент 15 корреляционной матрицы

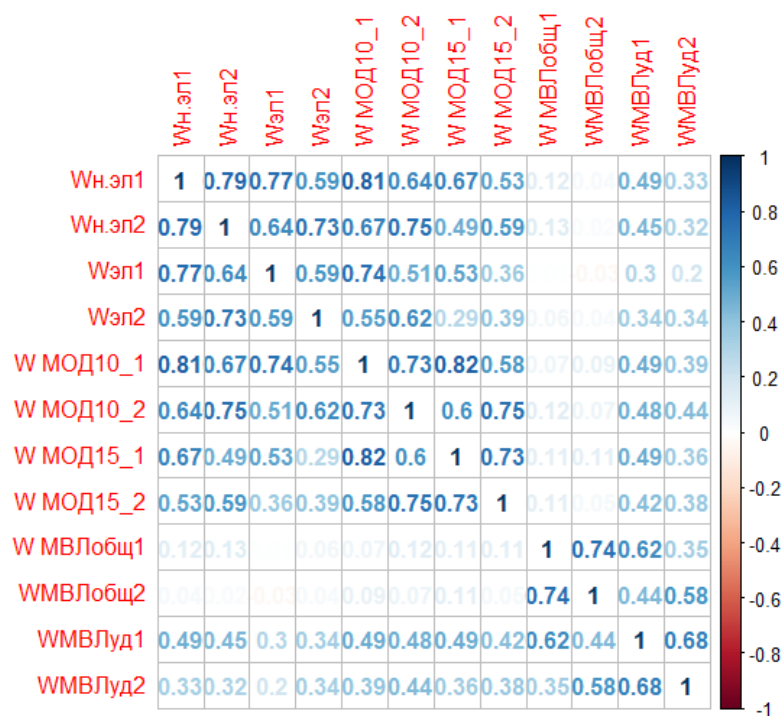


Рисунок 62 – Фрагмент 16 корреляционной матрицы