TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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
Разработка системы поддержки принятия решения о выборе траектории лечения детей с избыточным весом

УДК 004.822:519.81:613.25-053.4

Студент

Группа	ФИО	Подпись	Дата
8КМ71	Сподина Мария Константиновна		

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	Атепаева Н.А.			

ДОПУСТИТЬ К ЗАЩИТЕ:

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

Томск – 2019 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

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 научно-технической 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.
Р9	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.
Р10	Организовывать работы по моделированию прикладных ИС и реинжинирингу прикладных и информационных процессов предприятия и организации. Управлять проектами по информатизации прикладных задач и созданию ИС предприятий и организаций.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Марухина О.В.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8КМ71	Сподиной Марии Константиновне

Тема работы:

Разработка системы поддержки принятия решения о выборе траектории лечения детей с избыточным весом	
Утверждена приказом директора (дата, номер)	07.03.2019, №1787

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Объектом данного исследования являются клинико-лабораторные показатели детей с избыточным весом, собранные Томским НИИ Курортологии и физиотерапии.
Перечень подлежащих исследованию, проектированию и разработке вопросов	Анализ предметной области Выбор методов решения задачи Реализация проекта Описание полученных результатов
Перечень графического материала	Схема алгоритма обработки данных Диаграммы размахов Диаграмма визуализации пропущенных значений Графическое построение дерева решений
Консультанты по разделам выпускной квалификационной работы *(с указанием разделов)*	

Раздел	Консультант
Раздел 1-3	Марухина Ольга Владимировна
Раздел 4	Сосковец Любовь Ивановна
Раздел 5	Атепаева Наталья Александровна
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Глава 1. Анализ предметной области	
Глава 2. Выбор методов решения задачи	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	01.02.2019

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
8КМ71	Сподина М.К.		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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
Уровень образования Магистратура
Отделение школы (НОЦ) Отделение информационных технологий
Период выполнения Весенний семестр 2018 /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	
26.02.2019	Получение задания по финансовому менеджменту	
11.03.2019	Получение задания по социальной ответственности	
29.03.2019	Глава 1. Анализ предметной области	
17.04.2019	Глава 2. Выбор методов решения задачи	
15.05.2019	Глава 3. Реализация проекта и описание полученных результатов	
23.05.2019	Глава 4. Финансовый менеджмент	
31.05.2019	Глава 5. Социальная ответственность	
27.05.2019	Раздел на иностранном языке	

Составил руководитель ВКР:

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

СОГЛАСОВАНО:
Руководитель ООП

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

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8КМ71	Сподиной Марии Константиновне

Школа	Инженерная школа информационных технологий и робототехники	**Отделение школы (НОЦ)**	Отделение 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 1493 руб. 2. Оклад руководителя 26300 руб., исполнителя 1906 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
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. Матрица SWOT 3. График проведения НТИ 4. Диаграмма Ганта 5. Бюджет НТИ 6. Оценка ресурсной, финансовой и экономической эффективности НТИ	

Дата выдачи задания для раздела по линейному графику	26.02.2019

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
8КМ71	Сподина Мария Константиновна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8КМ71	Сподиной Марии Константиновне

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

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Система поддержки принятия решений будет использоваться на персональном компьютере. Рабочее место представляет собой офисное помещение с персональным компьютером.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019). – СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	Анализ выявленных вредных и опасных факторов: – отклонение показателей микроклимата; – превышение уровня шума; – недостаточная освещенность рабочей зоны; – повышенный уровень электромагнитных излучений – повышенное значение напряжения в электрической цепи; – нервно-психические перегрузки
3. Экологическая безопасность:	Анализ негативного воздействия переработки бумажных отходов, утилизации люминесцентных ламп, компьютеров и другой оргтехники на окружающую природную среду.
4. Безопасность в чрезвычайных ситуациях:	Возможные чрезвычайные ситуации: пожар в здании (наиболее типичная ситуация); авария на коммунальных системах жизнеобеспечения; землетрясение.

Дата выдачи задания для раздела по линейному графику	11.03.2019

Задание выдал консультант:

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

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8КМ71	Сподина Мария Константиновна		

РЕФЕРАТ

Выпускная квалификационная работа содержит 132 страницы, 48 рисунков, 24 таблицы, 56 источников, 5 приложений.

Ключевые слова: многомерные данные, анализ данных, система поддержки принятия врачебных решений, медицинские исследования, восстановление данных, снижение размерности, логические правила.

Объектом исследования является проблема избыточного веса у детей и подростков.

Целью исследования является разработка алгоритмической базы системы поддержки принятия врачебных решений при выборе траектории лечения детей с избыточным весом.

В процессе исследования проведено исследование структуры пропущенных данных клинико-лабораторных показателей, осуществлено восстановление пропущенных значений методом множественного восстановления и проведено сокращение размерности с использованием пакетов языка R. На основе полученных результатов построена модель принятия решений, и формирован набор правил для выбора траектории лечения.

В результате, сформирован алгоритм обработки данных для принятия решения о выборе группы лечения детей с избыточным весом.

Данное исследование носит научный характер, результаты данной работы могут быть использованы в дальнейших исследованиях по данной тематике.

Определения

Система поддержки принятия решений – компьютерная автоматизированная система, целью которой является помощь людям, принимающим решение в сложных условиях для полного и объективного анализа предметной деятельности.

Система поддержки принятия врачебных решений – медицинская информационная система, предназначенная для помощи врачам и иным медицинским специалистам в работе с задачами, связанными с принятием клинических решений.

Избыточный вес – результат формирования аномальных или чрезмерных жировых отложений, которые могут наносить вред здоровью.

Ожирение – заболевание с аномальным или избыточным накоплением жира, которое может вызвать нарушение здоровья.

Данные – совокупность сведений, зафиксированных на определенном носителе в форме, пригодной для постоянного хранения, передачи и обработки.

Выброс – результат измерения, выделяющийся из общей выборки.

Снижение размерности – преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных.

Дерево принятия решений – средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике для прогнозирования результатов.

Сокращения и обозначения

СППР – система поддержки принятия решений;

СППВР – система поддержки принятия врачебных решений ;

НИИ – научно-исследовательский институт;

bt – before treatment (до лечения);

WS – объем талии;

TS – объем бедер;

BMI – индекс массы тела;

SAD – систолическое артериальное давление;

DAD – диастолическое артериальное давление;

HOMA – индекс инсулинорезистентности Хома;

TFN – толерантность к физической нагрузке;

DP – двойное произведение (насыщение миокарда кислородом);

TAG – содержание триацилглицеридов;

aXC – содержание альфа-холестерина;

LDL – липопротеиды низкой плотности;

VLDL – липопротеиды очень низкой плотности;

MDA – малоновый диальдегид в сыворотке;

T3 – содержание трийодтиронина;

T4 – содержание тетрайодтиронина;

TTG – содержание тиреотропного гормона;

АТТРО – антитела к тиреопероксидазе;

IL_6 – содержание интерлейкина-6;

FNO – содержание фактора некроза опухоли;

IgA – концентрация иммуноглобулина А;

IgG – концентрация иммуноглобулина G;

IgM – концентрация иммуноглобулина М;

CIC – циркулирующие иммунные комплексы;

T_lym – циркулирующие иммунные комплексы;

KK – уровень каллекриина;

PI – протромбиновый индекс;

MG – макроглобулин в сыворотке крови;

APF – ангиотензинпревращающий фермент;

NO – содержание оксида азота;

OKLDL – окисление липопротеидов низкой плотности.

Оглавление

Введение

На данный момент одной из проблем области здравоохранения становится рост численности детей и подростков, страдающих избыточным весом. Данная проблема охватывает весть мир, оказывается негативное влияние на общее состояние здоровья подрастающего поколения и влечет серьезные последствия в будущем.

Одним из подходов медицинской диагностики заболеваний является обработка накопленных массивов данных пациентов. Сбора и хранения данных предлагают большие возможности их использования для помощи специалистам в постановке диагноза, выборе подходящего метода лечения, оценке эффективности лечения и других процессах оказания медицинской помощи. Исходя из этого, задача разработки технологий для оперативной и качественной обработки медицинских данных является очень актуальной в настоящее время.

Объектом исследования является проблема избыточного веса у детей и подростков. Предметом исследования является многомерный набор данных клинического исследования, содержащий клинико-лабораторные показатели пациентов с разной степенью ожирения.

Целью исследования является разработка алгоритмической базы системы поддержки принятия врачебных решений при выборе траектории лечения детей с избыточным весом.

Для достижения поставленной цели были определены следующие задачи выпускной квалификационной работы:

- обзор литературных источников и практик решения подобных задач для поиска метолов и инструментов проведения исследования;
- выбор методов и инструментов исследования;
- оценка исходных данных, предварительная обработка, исследование выбросов и пропущенных значений;
- восстановление однородности исходных данных;

- снижение размерности пространства признаков и формирования набора данных для анализа;
- построение модели и формирование набора правил для выбора траектории лечения;
- формирование алгоритма обработки данных для принятия решения о выборе траектории лечения детей с избыточным весом;
- анализ ресурсоэффективности и ресурсосбережения исследования;
- анализ основных аспектов социальной ответственности при проведении исследования.

Актуальность разработки систем поддержки принятия решений обусловлена возникновением врачебных ошибок, которые могут возникать по разным причинам. Среди них ограниченность временных ресурсов, недостаточность знаний, отсутствие возможности привлечения компетентных экспертов, неполнота информации о состоянии больного. Данные факторы могут привести серьезным проблемам со здоровьем в будущем.

Новизна исследования заключается в создании нового алгоритма определения индивидуальной траектории лечения пациентов на основе анализа клинико-лабораторных показателей.

Обзор литературы

В настоящее время многие учреждения здравоохранения разного масштаба и профилей во всем мире широко применяют автоматизацию медицинских технологий и соответствующих бизнес-процессов. Для этого используются медицинские информационные системы. Одним из важных направлений является разработка систем поддержки принятия решений (СППР). СППР используют для диагностики и прогнозирования заболеваний. Подобные системы не могут нести ответственность за принятые с ее помощью решения, но способны существенно упростить и ускорить работу врача. Цель СППР заключается в осуществлении кооперации системы и человека в процессе принятия решений.

Анализа научных публикаций по данной тематике показал, что исследования и разработки в области СППР ведутся по всему миру и в различных направлениях. Так, например, была представлена система поддержки принятия врачебных решений (СППВР), разработанная авторами (Кудряшов Ю.Ю., Атьков О.Ю., Касимов О.В., Прохоров А.А.) для использования в области сердечно-сосудистых заболеваний [1]. Система предназначена для выбора клинических решений по восьми основным заболеваниям сердечно-сосудистой системы (артериальная гипертония, острый коронарный синдром, др.). Позднее на основе использования системы был проведен анализ проблем, возникающих при внедрении технологий персональной телемедицины [2]. Предложены пути преодоления существующих препятствий и повышения эффективности телемедицинских технологий для управления здоровьем, профилактики и реабилитации кардиоваскулярных заболеваний.

В исследовании на базе Башкирского государственного медицинского университета г. Уфы[3, 4] были проанализированы возможности современных СППР в практике врача-хирурга. Работа авторов РНИМУ им. Н.И. Пирогова совместно с НМИЦ кардиологии [5] направлена на разработку архитектуры

базы знаний СППР для инструментальной диагностики стенокардии, в основе которой лежит онтологический подход с использованием графовой базы данных. Также ведутся разработки врачебных систем для хронических заболеваний, в области дерматовенерологии, анализа больных с ишемической болезнью сердца [6-9] и др. Обзор публикаций по теме систем поддержки принятия решений приведен в Приложении А.

Многие системы находятся не только на стадии разработки, но уже успешно внедрены. Первыми разработками (60-70 гг. XX века) были такие системы, как AAHelp (поиск причин возникновения резких болей, определение необходимости оперативного вмешательства); MICIN (постановка диагноза при инфекционных заболеваниях); INTERNIST (постановка диагноза по наблюдаемым симптомам). Известными классическими система являются: Germwatcher (больничная эпидемиология); PEIRS (отчеты по химическим патологиям); Puff (анализ нарушений деханий); SETH (анализ токсичности лекарственных средств); ATTENDING (поиск ошибок в решении); MicroScan, BD Phoenix (клиническая микробиология) и др.

С 2000-го года и успешно развиваются системы: IndiGO (формирование индивидуальных протоколов диагностики и лечения); Auminence (формирование диагностического плана); DiagnosisOne (поиск ошибок и формирование планов лечения), Isabel Healthcare (анализ симптомов); VisualDx (дифференциальная диагностика); Nuance (СППР для радиологии); IBM Watson (суперкомпьютер IBM с вопросно-ответной системой искусственного интеллекта). В направлении СППВР развиваются и более простые программы для конечного пользователя (WebMD Symptom Checker, DrNow, iParmacy, EasyDiagnosis). Большинство ведущих СППВР систем имеют в настоящее время мобильные и онлайн-версии.

В России в 80-90 гг. XX века были разработаны экспертные медицинские системы: ДИН (диагностика неотложных состояний); «Айболит» (диагностика острых расстройств кровообращения); ДИАГЕН (диагностика наследственных заболеваний); SYNGEN (хромосомная патология) и ряд других

[10]. Но эти системы существенно устарели и не получили дальнейшего развития, на смену им пришли другие, с более совершенными IT-инструментами, алгоритмами, методами, средствами, а также используемыми медицинскими знаниями.

В настоящее время в России освоены и успешно реализованы следующие системы: OncoFinder (анализ внутриклеточных процессов и подбор терапии для различных типов рака); ONDOC (содержит сервисы для безопасного хранения и совокупного анализа персонифицированной медицинской информации). Некоторые отечественные компании ведут разработку симптом-чекеров, среди них компания Medme, Mail.ru и др.

Отдельные исследования ведутся в вузах и НИИ. На базе Института трансляционной медицины (совместно с Национальным медицинским исследовательским центром им. В.А. Алмазова) в Международной лаборатории «Системы поддержки принятия решений в медицине» исследуют технологии модели и методы создания СППР для персонифицированной медицинской помощи [11].

Аннотации нескольких существующих зарегистрированных систем приведены в Приложении Б. Все запатентованные разработки содержат модули: хранения результатов обследования, поддержки принятия решений, управления знаниями, систему управления базами данных. Все модули являются необходимыми для полноценного функционирования СППР.

В настоящее время в медицине накоплен достаточно большой опыт эффективного использования СППР, особенно хорошо системы поддержки принятия решений зарекомендовали себя при дифференциальной диагностике и выборе подходящего лечения. Компьютерные медицинские системы могут очень облегчить работу специалиста, предоставляют возможность врачу проверить его диагностические предположения и использовать систему при работе со сложными клиническими случаями [10].

Глава 1 Анализ предметной области

1.1 Анализ медицинских данных

Анализ данных – это совокупность методов и средств извлечения из организованных данных информации для принятия решений. Анализ данных имеет много подходов и методов в различных областях деятельности человека. Марио Фариа, вице-президент исследовательской компании Gartner, дал такое объяснение: «Анализ – преобразование данных в выводы, на основе которых будут приниматься решения и строиться действия с помощью людей, процессов и технологий».

Данные представляют собой совокупность сведений, зафиксированных на определенном носителе в форме, пригодной для постоянного хранения, передачи и обработки. Преобразование и обработка данных позволяет получить информацию. Знания составляют зафиксированную и проверенную практикой обработанную информацию, которая может использоваться для принятия решений. Процесс принятия решений осуществляются на основе полученной имеющихся знаний и информации. При решении поставленной задачи данные обрабатываются и анализируются при помощи имеющихся знаний. На основании проведенного анализа представляются все возможные решения, из которых принимается одно наилучшее решение по некоторому критерию отбора. База знаний пополняется результатами решений (рисунок 1) [18].

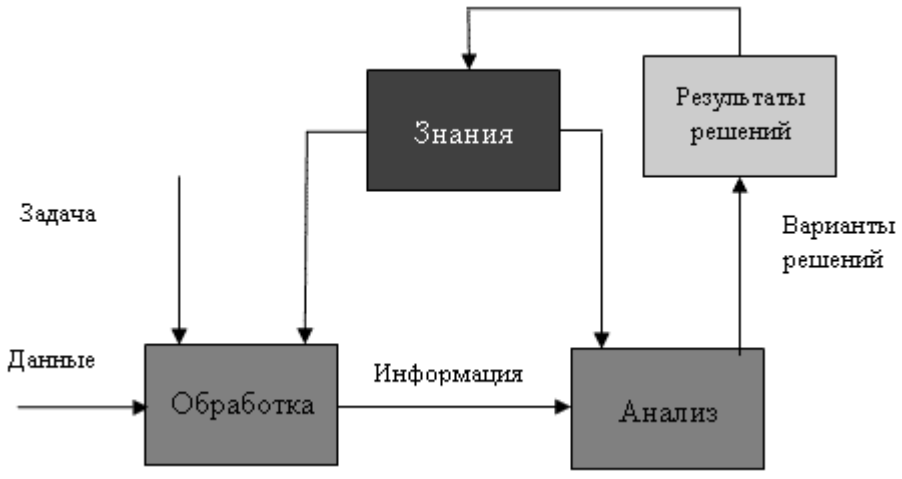

Рисунок 1 – Процесс принятия решений

В зависимости от количества признаков данные могут быть одномерными, двухмерными и многомерными. Многомерные данные содержат информацию о трех или более признаках для каждого объекта. Они используются для получения информации о зависимостях между признаками, для предсказания значения одной переменной на основании значений других. Анализ многомерных данных (АМД) – это подход, основанный на применении математических методов, в котором исследуются явления, процессы и системы, характеризуемые совокупностью переменных.

Медицинские данные – это любые данные, относящиеся к медицине, в более узком (персонифицированном) смысле – данные состояния здоровья конкретного человека. При работе с медицинскими данными следует учитывать определенные особенности:

- необходимо не допускать потерю информации, использовать данные исходной точности;
- не приводить избыточной информации при представлении результатов статистической обработки данных, достаточно исходной точности представления информации;
- часто при обработке данных необходим перевод качественных данные в количественные;
- при пропущенных значениях нельзя использовать «обнуление» (приписывать кодовое число нуль), это может совпадать с кодированием нормы по значениям признака;
- использовать среднее по классу значение для восстановления пропущенных данных неверно, особенно при малых выборках, потому как классы не всегда являются однородными. В этом случае лучше исключить подобные наблюдения или кодировать пропуски специальным знаком или числом, при этом обработка данных должна проводиться только по известным значениям [19].

Анализ медицинских данных, обнаружение закономерностей между ними и извлечение данных, как правило, сопровождается проблемой большой

размерности данных. Размерность хранимых данных определяется числом признаков, описывающих состояние здоровья пациента. Порой размерность медицинских данных весьма велика и может достигать нескольких десятков и сотен показателей. Поэтому очень важным аспектом при работе с медицинскими данными является снижение размерности пространства признаков и выделение наиболее информативных из них.

Под системой поддержки принятия решений обычно понимается информационная система, которая собирает данные, обрабатывает их и способна влиять на процессы принятия решений в различных областях человеческой деятельности. В здравоохранении используется термин «системы поддержки принятия врачебных решений» (СППВР).

В медицине СППР предназначены для решения таких задач как: помощь диагностирования, подача тревожных сигналов и напоминаний, поиск прецедентов (подходящих случаев), распознавание и интерпретация образов, планирование и контроль терапии. Важной функцией СППВР является распространение практик со всего мира. Чаще всего СППВР используются именно как помощь постановки диагноза, для назначения и корректировки выбранного лечения [11].

СППР обладают следующими преимуществами:

- помощь в определении ожидаемых диагностических решений основывается на собранных данных;
- диагностические гипотезы, принимаемые к рассмотрению, становятся менее зависимыми от личного опыта и знаний врача, уменьшается вероятность пропустить важные признаки;
- база знаний содержит знания экспертов и современные представления по проблемной области, что позволяет системе использовать необходимые инструменты и алгоритмы принятия решений.

1.2 Описание предметной области

Избыточный вес и ожирение у детей и подростков является одной из самых серьезных проблем общественного здравоохранения в XXI веке. В качестве основного показателя для оценивания избыточного веса широко применяется индекс массы тела (ИМТ), который рассчитывается как отношение массы тела в килограммах к квадрату роста в метрах (кг/м2). ИМТ не зависит от гендерного признака и возрастных категорий взрослых и поэтому является наиболее удобной мерой оценки уровня ожирения и избыточного веса. Однако у разных людей он может соответствовать разной степени полноты, поэтому ИМТ следует считать приблизительным критерием. При определении избыточного веса и ожирения у детей следует учитывать возраст.

Для детей в возрасте до 5 лет избыточный вес и ожирение определяются следующим образом: если соотношение «масса тела/рост» более чем на два стандартных отклонения превышает медианное значение (согласно Стандартным показателям физического развития детей, ВОЗ), то имеет место избыточный вес; если превышает более чем на три стандартных отклонения – ожирение.

Для детей возраста 5-19 лет: если соотношение «ИМТ/возраст» превышает медианное значение (согласно Стандартным показателям физического развития детей, ВОЗ), более чем на одно стандартное отклонение, то это избыточный вес; более чем на два стандартных отклонения – ожирение [20].

Неправильное питание и малоподвижный образ жизни являются основными факторами роста количества детей с избыточным весом. Реже ожирением страдают вследствие заболеваний эндокринной системы, опухолей головного мозга и других серьезных болезней. Исходя из причины развития ожирения у ребенка, различают алиментарное и обменное (эндокринное) ожирение. Алиментарное ожирение вызвано малоподвижным образом жизни и

неправильным питанием, а эндокринное ожирение, исходя из названия, развивается при различных заболеваниях эндокринной системы.

Ожирение в детском возрасте способно привести к развитию серьезных осложнений. Среди осложнений может быть раннее развитие таких заболеваний, нехарактерных для детского возраста, как гипертоническая болезнь, сахарный диабет, ишемическая болезнь сердца, цирроз печени и др. Это, в свою очередь, заметно ухудшает качество жизни и, к сожалению, уменьшает ее продолжительность [21].

Согласно статистике в 2016 году около 41 миллиона детей в возрасте до 5 лет имели избыточный вес или ожирение. В Африке с 2000 г. число детей в возрасте до 5 лет, страдающих ожирением, выросло почти на 50%. В 2016 году почти половина детей в возрасте до 5 лет с избыточным весом или ожирением проживала в Азии. В 2016 г. 340 миллионов детей и подростков в возрасте от 5 до 19 лет страдали избыточным весом или ожирением. С 1975 по 2016 гг. доля избыточного веса и ожирения среди детей и подростков в возрасте от 5 до 19 лет возросла с 4% до 18% и составило 124 миллионов человек. От последствий избыточного веса и ожирения умирает больше людей, чем от последствий аномально низкой массы тела. Ежегодно в результате избыточного веса или ожирения умирает в среднем 2,6 миллиона человек [20].

1.3 Описание исходных данных

В рамках сотрудничества с научно-исследовательским институтом курортологии и физиотерапии г. Томска (ТНИИКиФ) был получен набор данных клинического исследования, проведенного на базе детского отделения. Данные представлены в виде многомерной таблицы и содержат значения групп клинико-лабораторных показателей пациентов. Пациентами являются дети и подростки в возрасте от 7 до 18 лет, страдающие проблемами с весом. Набор данных содержит 345 объектов (пациентов), 160 показатель (значения до и после лечения) и 5 групп (методик лечения).

Фрагмент набора данных представлен на рисунке 2.

N	Name	Groupe 1	Groupe 2	Se	Receipt da	Discharge date	Age	Degree of obesit	The degree of obesity by body mass index (BMI)	Degree of goite	Related diagnos	Complicati ons	Complair
1	Patient 1	3	1	ж	9 мар 05	2 апр 05	11	I	1	I			
3	Patient 2	6	2	м	11 мар 05	25 июн 05	13	I	изб	I	7		2.3
4	Patient 3	8	3	ж	12 мар 05	10 авг 05	9	I	1	0	1,8,9		5
5	Patient 4	10	3	м	13 мар 05	4 ноя 05	10	I	изб	I	1,2,3,4		1.3
6	Patient 5	11	3	ж	14 мар 05	13 ноя 05	14	I	изб	0			
7	Patient 6	11	3	ж	15 мар 05	2 дек 05	9	II	изб	I	1.12		
8	Patient 7	11	3	ж	16 мар 05	2 дек 05	10	II	2	I	1.11		1,2,3
9	Patient 8	11	3	ж	17 мар 05	2 дек 05	11	I	изб	I	7.4		2.3
10	Patient 9	11	3	ж	18 мар 05	8 дек 05	10	I	изб	I	1.4		
11	Patient 10	11	3	м	19 мар 05	2 дек 05	15	II	изб	0			
12	Patient 11	12	1	ж	20 мар 05	29 дек 05	14	I	изб	I	7		2.3
18	Patient 12	3	1	м	24 фев 06	20 мар 06	10	II	2	0			
24	Patient 13	3	1	ж			12						
27	Patient 14	4	1	ж	17 апр 06	11 май 06	10	I	1	I			
28	Patient 15	4	1	ж			8	III	2	I			
29	Patient 16	4	1	ж	18 апр 06	11 май 06	13	III	1	0			
31	Patient 17	5	2	м	12 май 06	5 июн 06	13	III	2				
33	Patient 18	5	2	ж	19 май 09		14						

Рисунок 2 – Фрагмент исходного набора данных

Все показатели относятся к разным физиологическим группам. Перечислим в каждой группе основные показатели, по которым проводится исследование (таблица).

Таблица 1 – Основные клинико-лабораторные показатели исследования

Физиологическая группа	Клинико-лабораторный показатель, ед. изм.	Значение показателя
Клиника	ОТ	объем талии
	ОБ	объем бедер
	Масса, кг	масса тела в кг
	Избыток, %	избыток массы тела в %
	ИМТ	индекс массы тела
Сердечнососудистая система	САД, мм.рт.ст.	систолическое артериальное давление
	ДАД мм рт. ст.	диастолическое артериальное давление
Физическая работоспособность	HOMA	индекс инсулинорезистентности Хома
	ТФН	толерантность к физической нагрузке
	Дв.пр.	двойное произведение (насыщение миокарда кислородом)
	Общая р-ть	общая работоспособность
Липидный обмен	Общие липиды (4-8 г/л)	общие липиды крови
	ТАГ (менее 1,7ммоль/л)	содержание триацилглицеридов
	аХС (более 1,03 ммоль/л)	содержание альфа-холестерина
	ЛПНП (ммоль/л)	липопротеиды низкой плотности
	ЛПОНП (ммоль/л)	липопротеиды очень низкой плотности
Биохимия крови	Щелочная фосфатаза (ед.)	щелочная фосфатаза в сыворотке крови
	Каталаза (мкатал/л)	каталаза в сыворотке крови
	МДА	малоновый диальдегид в сыворотке
	Церулоплазмин	содержание церулоплазмина
Углеводный обмен	Глюкоза (< 5,6ммоль/л)	уровень глюкозы крови натощак
Гормональный статус	Т3 (1,0-2,8)	содержание трийодтиронина
	Т4 (53-158)	содержание тетрайодтиронина
	ТТГ (0,23-3,4)	содержание тиреотропного гормона
	АТ ТПО (до 30 у.е.)	антитела к тиреопероксидазе
	Инсулин	содержание инсулина в сыворотке

		крови
	Лептин	содержание лептина в сыворотке крови
Цитокиновый статус	ИЛ-6	содержание интерлейкина 6
	ФНО (не >2,5 пг/мл)	содержание фактора некроза опухоли
Иммунологический статус	IgA (0,7-3,0 г/л)	концентрация иммуноглобулина А
	IgG (8,0-16,0 г/л)	концентрация иммуноглобулина G
	IgM (0,8-2,5 г/л)	концентрация иммуноглобулина М
	ЦИК (40-100 у.е.)	циркулирующие иммунные комплексы
	Лизоцим (40-60 у.е.)	активность лизоцима
	Т-лимфоциты (45-70) %	циркулирующие иммунные комплексы
Состояние калликреин-кининовой системы	КК	уровень каллекриина
	ПИ	протромбиновый индекс
	МГ	макроглобулин в сыворотке крови
	АПФ	ангиотензинпревращающий фермент
Окислительная способность плазмы крови	NO	содержание оксида азота
	ОК ЛПНП	окисление липопротеидов низкой плотности

Помимо перечисленных показателей набор данных содержит информацию о принадлежности пациента к группе лечения (1-5). Для решения поставленной задачи необходимо построить систему, которая могла бы для нового пациента определить необходимую ему траекторию (группу) лечения, основываясь на его первичных клинико-лабораторных показателях.

Глава 2 Выбор методов решения задачи

2.1 Методы поддержки принятия решений

Существуют разные методы информационных технологий для осуществления поддержки принятия решений в СППР. Кратко рассмотрим некоторые из них:

1. *Информационный поиск* – представляет поиск неструктурированной документальной информации согласно определенной информационной потребности. Процесс поиска состоит из операций сбора, обработки и предоставления нужной информации. Процесс поиска состоит из следующих этапов:

1) определение потребностей в информации и формулировка информационного запроса;

2) определение числа возможных информационных источников (массивов);

3) извлечение информации из выявленных информационных массивов;

4) изучение полученной информации и оценка результатов.

Поиск бывает полнотекстовый (по всему содержимому документа), поиск по метаданным (по атрибутам документа, поддерживаемым системой) и по изображению.

Методы, с помощью которых осуществляется информационный поиск:

- адресный (поиск документов по формальным признакам, указанным в запросе);
- фактографический (поиск фактов, которые соответствуют информационному запросу);
- семантический (поиск документов по их содержанию);
- документальный (поиск соответствующих запросу пользователя документов в хранилище или в базе данных).

2. *Поиск знаний в базах данных* – это процесс обнаружения полезных знаний в базах данных. Знания могут быть представлены в виде правил,

закономерностей, прогнозов, взаимосвязей между элементами данных, др. Главным инструментом поиска знаний в базах данных являются аналитические технологии data mining. Процесс извлечения знаний состоит из выполнения последовательности операций для поддержки аналитического процесса. К ним относятся операции: консолидация данных, подготовка выборок данных, очистка данных от некорректных факторов, оптимизация данных для решения определенных задач, анализ, визуализация и интерпретация результатов анализа, применение результатов на практике.

3. *Интеллектуальный анализ данных* (*data mining*) представляет собой выявление скрытых закономерностей или взаимосвязей между переменными в больших массивах необработанных данных. Такой анализ предусматривает решение задач классификации, моделирования и прогнозирования, кластеризации, сокращения описания, ассоциации, анализа отклонений, визуализации и др.

4. *Рассуждение на основе прецедентов* – решение новой задачи на основе использования или адаптации опыта решения подобных задач. Прецедент представляет собой случай, произошедший ранее и являющийся примером для решения подобных случаев. Методы рассуждений предусматривают, как правило, четыре основных этапа, образующие цикл рассуждения на основе прецедентов (CBR-цикл): поиск прецедента, повторное использование, адаптация и сохранение. Для извлечения прецедентов и их модификаций существует ряд методов: метод ближайшего соседа; метод извлечения прецедентов на основе знаний; метод извлечения с учетом применимости прецедентов; метод извлечения прецедентов на основе деревьев решений. Также для извлечения прецедентов могут успешно применяться и другие методы, например, искусственные нейронные сети.

5. *Генетический алгоритм* представляет алгоритм поиска для решения задач оптимизации и моделирования путем последовательного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию. Генетический

алгоритм является разновидностью эволюционных вычислений. Отличительной особенностью является акцент на использовании оператора "скрещивания", который производит рекомбинацию решений-кандидатов, роль которой аналогична роли скрещивания в живой природе. Генетические алгоритмы служат для поиска решений в очень больших, сложных пространствах поиска.

6. *Имитационное моделирование* (ИМ) – метод исследования с использованием моделей, описывающих процессы реального мира. Подобные модели можно использовать как для множества испытаний. Полученные данные являются достаточно устойчивыми и правдивыми. Популярными системами ИМ являютяс: Arena, AnyLogic, eM-Plant, GPSS, Powersim.

Имитационное моделирование можно разделить на виды:

- агентное моделирование, которое используется для исследования децентрализованных агентов и систем в целом;
- дискретно-событийное моделирование, основанное на абстрагировании от непрерывной природы событий и рассмотрении только основных событий моделируемой системы (ожидание, обработка заказа, движение с грузом, разгрузка и др.);
- системная динамика, при которой для исследуемой системы строятся графические диаграммы причинных связей и глобального влияния одних параметров на другие во времени, далее на основе этих диаграмм строится модель, которая имитируется на компьютере.

7. *Искусственные нейронные сети* – это математические модели и их аппаратные или программные реализации, построенные по принципу организации и функционирования биологических нейронных сетей. Понятие возникло при изучении и моделировании процессов мозговой деятельности при мышлении. Впоследствии такие модели стали использовать в практических целях, как правило, в задачах прогнозирования. В машинном обучении нейронная сеть является частным случаем методов распознавания образов, методов кластеризации, дискриминантного анализа, и т.п. В области математики обучение нейронных сетей представляет собой

многопараметрическую задачу нелинейной оптимизации. В кибернетике нейронная сеть используется в робототехнике и в задачах адаптивного управления.

8. *Искусственный интеллект* – это разработка интеллектуальных машин и систем, направленных на понимание человеческого интеллекта. Используемые при этом методы не должны быть биологически правдоподобными. Учеными исследованы только некоторые механизмы интеллекта, поэтому в пределах этой науки под интеллектом понимают только вычислительную часть способности достигнуть поставленных целей. Когда основу работы СППР составляют методы искусственного интеллекта, то система называется интеллектуальной системой поддержки принятия решений [22, 23].

2.2 Технология Data Mining

Понятие интеллектуального анализа данных соответствует широко распространенному термину *Data Mining*, который переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний. Data Mining – это процесс, основная цель которого заключается в извлечении информации из набора данных и преобразовании ее в понятную структуру для дальнейшего использования.

Технология Data Mining включает в себя достаточно большое количество методов, основу которых составляют различные методы моделирования, классификации и прогнозирования, основанные на применении искусственных нейронных сетей, деревьев решений, эволюционного программирования, ассоциативной памяти, нечёткой логики, в том числе алгоритмы *k*-средних и *k*-медианы; методы поиска ассоциативных правил, метод ограниченного перебора, генетические алгоритмы и разнообразные методы визуализации данных.

К методам Data Mining часто относят статистические методы, например: корреляционный и регрессионный анализ, факторный анализ, дисперсионный

анализ, компонентный анализ, дескриптивный анализ, анализ временных рядов, дискриминантный анализ, и др.

Методы Data Mining предназначены для решения задач, которые можно условно разделить на классы:

1. Классификация – нахождение у объектов исследования специфических признаков, определяющих их отношение к одному из заранее определенных классов.

2. Кластеризация – разбиение множества объектов и признаков исследования на однородные в определенном смысле кластеры или группы. В этой задаче классы заранее не сформированы, их необходимо определить. В остальном идеи классификации сохраняются.

3. Последовательность – выявление закономерностей событий, взаимосвязанных по времени.

4. Ассоциация – выявление закономерностей среди взаимосвязанных одновременно произошедших событий и зависимости между произошедшими явлениями.

5. Регрессия и прогнозирование – поиск зависимости выходных данных от входных переменных и предсказание на основе выявленных зависимостей новых результатов.

6. Визуализация – графическое отображение анализируемых данных в виде наглядных таблиц, диаграмм, графов, схем, и т.д.

2.3 Методы анализа исходных данных

2.3.1 Анализ выбросов

Процесс исправления выбросов и другой неправильной и нежелательной информации называется очисткой данных при их предварительной обработке. Существует несколько типов процесса очистки данных, которые зависят от типа данных, подлежащих очистке. Для методов количественных данных обнаружение выбросов может быть использовано для устранения аномалии в данных.

Выбросами называют показатели, которые значительно отличаются от остальных в пределах рассматриваемой выборки. Выбросы могут образовываться в результате ошибок измерений, ввода или интерпретации данных, в виду особенности природы данных.

Для получения более точных результатов анализа, выбросы необходимо удалять. Для обнаружения выбросов используются простейшие способы, основанные на межквартильном расстоянии: выбросами считаются значения, не попавшие в диапазон:

$$\left[\left(x_{25} - 1{,}5 * (x_{75} - x_{25})\right), \left(x_{25} + 1{,}5 * (x_{75} - x_{25})\right)\right] \quad (2.1)$$

Более тонкими критериями являются критерий Шовене, критерий Граббса, критерий Пирса и критерий Диксона [24].

Обычно для обнаружения выбросов используют диаграммы размахов, которые также называются «ящиками с усами». Диаграмма размаха упрощает данные, показывая следующие параметры: медиану, верхнюю и нижние квартили, наименьшее и наибольшее значения (кроме выбросов), сами выбросы, существенно меньшие или большие остальных чисел в ряду. Как правило, этот способ применяется при необходимости визуального сравнения нескольких категорий данных.

2.3.2 Восстановление пропущенных значений

На практике наборы данных не всегда имеют полный перечень значений, часто приходится сталкиваться с пропусками в данных. В этом случае необходимо оценить причину их появления и влияние на дальнейший анализ. Если оставить пропущенные значения должной обработки, то полученные результаты в итоге могут не соответствовать действительности.

Чтобы понять, как правильно обработать пропуски, необходимо определить механизмы их формирования. Существуют три типа пропусков, различающиеся способами возникновения:

1. Полностью случайные пропуски (*missing completely at random, MCAR*). Такие пропуски независимы от наблюдаемых параметров и переменных и

происходят совершенно случайно, вероятность пропуска для каждого элемента в массиве одинакова. На этапе восстановления пропуска элемента в массиве данных, допускается игнорировать и изменять значение пропущенного элемента, так как это не ведёт к искажению результата.

2. Случайные пропуски (*missing at random, MAR*) возникают не случайно, а из-за некоторых закономерностей. Пропуски элементов в массивах данных распределены внутри определенных подгрупп переменных. Вероятность пропуска элемента может быть определена на основе имеющейся в наборе информации, которая не содержит пропуски. Исключение или замена пропущенного элемента на некоторое значение не ведёт к существенному искажению результатов.

3. Отсутствуют не случайно (*missing not at random, MNAR*). Данные отсутствуют в зависимости от неизвестных факторов. Вероятность пропуска могла бы быть описана на основе других атрибутов, но информация в наборе данных отсутствует. Такие пропуски являются не игнорируемыми [25].

Необходимо знать, какая доля данных пропущена. Слишком большое количество пропущенных значений является проблемой для дальнейшей обработки и анализа данных. Обычно максимальный порог пропущенных данных составляет 15% от общего количества выборки. Если пропущенные данные превышают порог, то необходимо отказаться от этого показателя.

Методы обработки пропущенных значений делятся на базовые (простейшие и простые) и продвинутые (рисунок 3) [26].

Наилучшим на данный момент из методов восстановления пропущенных данных является *метод множественного восстановления (multiple imputation, MI)*. Этот метод основан на повторяющемся моделировании данных при использовании методов Монте-Карло. Метод множественного восстановления данных состоит в одновременном генерировании нескольких значений пропущенной величины вместо замены пропущенной информации одним значением. Для практического использования

всех восстановленных таким образом данных генерируются наборы баз данных с различными вариантами подстановок пропущенных значений [27].

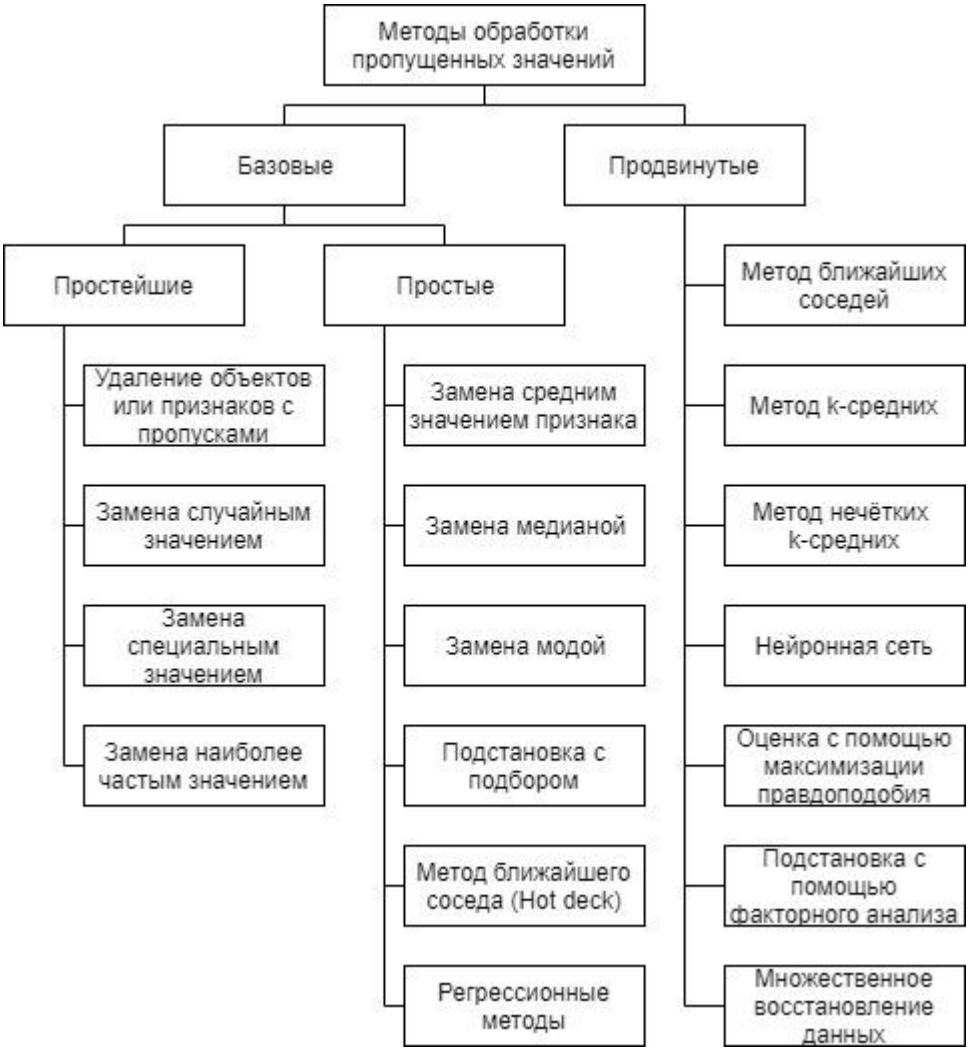

Рисунок 3 – Классификация методов восстановления пропущенных значений

2.3.3 Сокращение размерности многомерных данных

Снижение размерности – это преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных (информативных признаков). Сокращение размерности может проводиться разными методами, рассмотрим подробнее некоторые из них.

1. *Метод главных компонент (principal component analysis, PCA).* Является одним из наиболее часто используемых методов снижения размерности пространства признаков при минимальной потере полезной информации. Основу метода составляет переход от исходной системы координат (x_1,x_2,…,x_n) многомерного пространства признаков к новой системе

координат (y_1,y_2,…,y_n), которая является системой ортонормированных комбинаций:

$$f(\mathrm{n}) = \left\{ y_{\mathrm{j}}(x) = w_{\mathrm{ij}}(x_1 - m_1) + \cdots + w_{\mathrm{nij}}(x_{\mathrm{n}} - m_{\mathrm{n}}),\ \sum_{\mathrm{i}=1}^{\mathrm{n}} w_{\mathrm{ij}}^2 \right\}, \qquad (2.2)$$

где j = 1,…,n, $\sum_{\mathrm{i}=1}^{\mathrm{n}} w_{\mathrm{ij}} w_{\mathrm{ik}} = 0$, j, k = 1,…,n ($j \neq k$)

где m_{i} – математическое ожидание значений признака x_{i}. Линейные комбинации выбираются так, чтобы первая главная компонента $y_1(x)$ имела бы максимальную дисперсию. Таким образом, новая ось будет ориентирована вдоль направления наибольшей вытянутости эллипсоида рассеяния точек объектов данных в пространстве признаков.

Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеяния, перпендикулярное первой главной компоненте. Следующие главные компоненты определяются по аналогичной схеме [28].

2. *Факторный анализ (factor analysis, FA)* имеет отличие в том, что не опирается на заранее заданный перечень факторов, которые влияют на исследуемые объекты или процессы, а, наоборот, способствует обнаружению наиболее важных из них, причем скрытых.

Задача состоит в том, чтобы выявить скрытые обобщенные факторы, которые объясняют изменения изучаемого показателя. Эти факторы позволяют строить аналитические модели с относительно небольшим числом переменных, что упрощает их реализацию и интерпретацию пользователем, снижает вычислительные затраты и время, требуемое на получение решений, что в свою очередь повышает оперативность принятия решений на основе полученных результатов [29].

3. *Сингулярное разложение (Singular Value Decomposition, SVD)*

В интеллектуальном анализе данных этот алгоритм можно использовать для лучшего понимания базы данных, показывая количество важных измерений, а также для упрощения, уменьшая количество атрибутов, которые

используются в процессе интеллектуального анализа данных. Сингулярное разложение удаляет ненужные данные, которые линейно зависят с точки зрения линейной алгебры.

С помощью этого метода можно удалить шумы и линейно зависимые элементы, используя только самые важные сингулярные значения. Это очень полезно при извлечении данных, поскольку трудно определить, понятна ли база данных, и, если нет, какие элементы полезны для анализа.

4. *Выбор признаков*.

Методы выбора признаков можно разделить на два подхода: ранжирование признаков и выбор подмножества признаков. В первом подходе объекты ранжируются по некоторым критериям, а затем выбираются объекты, превышающие определенный порог. Во втором подходе происходит поиск подмножеств пространства признаков для создания оптимального подмножества. Второй подход можно разделить на три части:

1) Фильтрация: выбираются признаки, после чего выбранное подмножество используется для выполнения алгоритма классификации;

2) Включение: выбор характеристик происходят как часть алгоритма классификации;

3) Формирование: алгоритму классификации применяется к набору данных для определения наилучших характеристик [30].

После изучения методов и алгоритмов, для решения поставленной задачи восстановления пропущенных данных будет использован метод множественного восстановления пропущенных данных. В качестве методов сокращения размерности будут рассмотрены методы выбора признаков.

2.4 Выбор инструментария

На рынке программных продуктов имеется достаточно большой выбор средств для анализа данных. Рассмотрим наиболее популярные их них.

1. R – язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом в рамках проекта GNU.

R поддерживает большое количество статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов – библиотек для работы специфических функций или специальных областей применения. В базовую поставку R включен основной набор пакетов, по состоянию на 2017 год всего доступно более 11778 пакетов[31]. В системе R имеются возможности, в том числе, и для работы с графикой, оконный интерфейс можно установить как дополнительное приложение. Но простота использования и масштабируемость системы очень сильно повысили популярность R в последние годы.

RStudio – это среда для общения с системой R, которая позволяет более удобно создавать и передавать запросы на обработку системе R. RStudio является свободной средой разработки программного обеспечения с открытым исходным кодом для языка программирования R. RStudio доступна в двух версиях: RStudio Desktop как обычное приложение с локальным доступом; и RStudio Server с доступом через браузер к RStudio, установленной на удаленном Linux-сервере [32].

Достоинством продукта является большой выбор статистических методов и более 7000 тысяч проверенных пакетов; быстрый переход к функциям; интегрированная документация R. К недостаткам можно отнести небольшой объем документации на русском языке.

2. SPSS (Statistical Package for Social Science) – компьютерная программа для статистической обработки данных, является одним из лидеров рынка в области коммерческих статистических продуктов для проведения прикладных исследований в области социальных наук. Отличается гибкостью и мощностью применения для всех видов статистических расчетов.

Возможности программы: ввод и хранение данных; возможность использования переменных разных типов; первичная описательная статистика;

маркетинговые исследования; анализ данных маркетинговых исследований. Недостатками продукта можно назвать дороговизну лицензий и отсутствие гибкости в расчетах [33].

3. Statstica – программный пакет для статистического анализа, реализующий функции анализа, управления, добычи и визуализации данных с привлечением статистических методов. Пакет имеет широкие графические возможности, позволяет выводить информацию в виде различных типов графиков (включая деловые, научные, трёхмерные и двухмерные графики в разных системах координат, специализированные статистические графики: гистограммы, категорированные, матричные графики и др.), все компоненты графиков могут настраиваться под нужды исследования [34]. Statistica содержит более 250 статистических функций, генерирует понятные, настраиваемые отчеты, но разработана только под Windows, что является одним из ее недостатков.

4. Stata – программный пакет для анализа данных в сферах экономики, социологии, политики, биомедицины и др. Включает в себя: базовые статистические методы, линейные модели, непараметрические и многомерные методы, кластерный анализ, графику, операции над данными. Относительно дешевый аналог SPSS. Недостатком продукта является довольно узкая специализация [35].

5. SAS – это программный продукт, разработанный с целью создания точных предсказательных и описательных моделей на основании больших объемов информации. Областью применения являются разнообразные научные исследования, бизнес аналитика и т. д. Программный пакет может обрабатывать, изменять, управлять и извлекать данные из различных источников и выполнять статистический анализ. В системе SAS реализован собственный язык программирования. Взаимодействие с программой возможно через консоль и через графический интерфейс, представляющий собой графическую оболочку для упрощенного ввода команд языка программирования SAS.

Достоинствами системы являются гибкий интерфейс обмена данными (интеграции), быстрота расчетов на больших массивах данных и наличие инструментария для работы с кластерами. К недостаткам можно отнести сложность поддержки написанных скриптов, сложность освоения и дороговизну лицензий [36, 37].

6. Rapid Miner – бесплатная среда среда для прогнозной аналитики. Эта программа, написанная на Java, предлагает продвинутые возможности анализа данных, реализованные в виде основанного на шаблонах фреймвока. При этом пользователю вообще не требуется писать код. RapidMiner предлагается скорее в виде сервиса, а не отдельной программы. Кроме того, RapidMiner обеспечивает предварительную обработку и визуализацию данных, предиктивный анализ, статистическое моделирование. Он поддерживает учебные схемы, модели и алгоритмы из WEKA, а также скрипты на R. Возможности RapidMiner могут быть расширены с помощью дополнений, отдельные из которых также доступны бесплатно. Система поддерживает все этапы глубинного анализа данных, включая результирующую визуализацию, проверку и оптимизацию [38]. К недостаткам можно отнести то, что не все задачи могут выполняться на сервере, это характерно, когда модель требует всей базы, чтобы выполнить алгоритм частями.

7. MATLAB – пакет прикладных программ для решения задач технических вычислений. Предоставляет пользователю большое количество функций для анализа данных, которые покрывают практически все области математики. Работа на большинстве современных операционных систем, удобный графический интерфейс и простота в работе являются достоинствами программы. Недостатком является дороговизна лицензии и неполная поддержка статистических функций [35].

Проведем сравнительный анализ описанных выше программных продуктов (таблица 2) [39].

В ходе анализа для выполнения задач исследования были выбраны следующие инструменты: система R (среда разработки RStudio) и частично

возможности Rapid Miner. Язык R является одним из ведущих статистических инструментов в мире. Он активно применяется в различных областях науки и все больше используется в обработке медицинских данных.

Таблица 2 – Сравнительный анализ инструментов решений

	R	SPSS	Statstica	Stata	SAS	Rapid Miner	MATLAB
Свободная лицензия	+	-	-	-	-	+	-
Наличие графического интерфейса	+	+	+	+	+	+	+
Наличие консольного ввода	+	+	-	+	+	-	+
Простота понимания и изучения	+	-	-	+	-	+	-
Поддержка операционных систем	Windows, Mac OS, Linux, Unix	Windows, Mac OS, Linux, Unix	Windows	Windows, Mac OS, Linux, Unix	Windows, Linux, Unix	Windows, Mac OS, Linux	Windows, Mac OS, Linux

Глава 3 Реализация проекта и описание полученных результатов

3.1 Первичная обработка данных

Первичная обработка данных направлена на упорядочивание информации об исследуемом объекте. На данной стадии исходные сведения группируется по тем или иным признакам, заносятся по таблицам. Исходные данные необходимо анализировать, выявлять информацию с некорректными значениями, предоставлять в удобной для работы форме. Удобство представления сведений дает исследователю понять общую картину всей совокупности данных. Обработка может быть направлена на упорядочивание исходных данных; обнаружение и устранение ошибок, некорректно введенных данных, выбросов, пробелов в сведениях; выявление скрытых закономерностей и связей.

Обработка данных при проведении исследования осуществлялась по следующему алгоритму (рисунок 4):

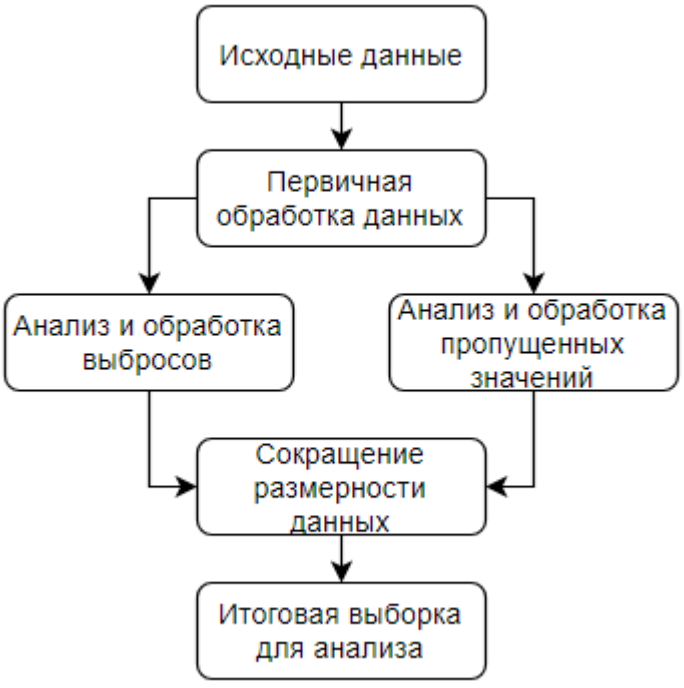

Рисунок 4 – Схема алгоритма обработки данных

3.1.1 Предварительная подготовка данных

Исходные данные представляют собой набор данных клинического исследования, который содержит 345 объектов, 160 признаков. Признаками являются различные клинико-лабораторные показали обследования пациентов до и после лечения. Для дальнейшего анализа необходимо привести данные к подходящему виду и определить рабочую выборку.

Данные исходного набора очень разрежены, в частности, некоторые признаки являются полностью пустыми на всём промежутке записей. Подобные признаки были исключены из рабочей выборки. Также были исключены все признаки до лечения и разности между значений до и после лечения. Задачей исследования является определение траектории лечения новых пациентов, поэтому в рабочей выборке содержатся данные о пациентах по клинико-лабораторным показателям до лечения и назначенные им методики лечения (группы). Подготовленный к анализу набор данных содержит записи о 276 пациентах по 66 признакам (рисунок 5).

WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	SAD_bt	DAD_bt	HOMA_bt	TFN_bt	DP
NA	NA	46.0	35.3	24.2	NA	100	50	NA	NA	
48	95	72.0	10.7	25.0	44.4	100	60	0.7098214	78.0	
NA	NA	40.0	29.0	22.2	29.8	100	50	NA	NA	
43	83	39.0	22.0	20.7	28.4	100	55	NA	NA	
NA	NA	70.0	18.0	24.0	45.6	110	55	0.4285714	NA	
48	76	34.0	30.7	20.1	27.7	100	50	3.0598214	48.3	
53	83	77.0	40.0	29.4	47.0	110	55	0.4303125	90.0	

Showing 1 to 9 of 276 entries, 66 total columns

Рисунок 5 – Фрагмент исходного набора данных

3.1.2 Анализ выбросов

Важным этапом подготовки данных к восстановлению является анализ выбросов, который позволяет избежать отклонения восстановленных данных от разумных значений. Для определения выбросов воспользуемся графическим анализом наблюдений посредством диаграмм размахов, с помощью которых легко распознать значения, лежавшие за пределами наблюдений.

В R этот график реализован с помощью функции boxplot. Построим диаграмму размаха для показателя WS_bt (объем талии до лечения) в зависимости от группы лечения Groupe (рисунок 6)

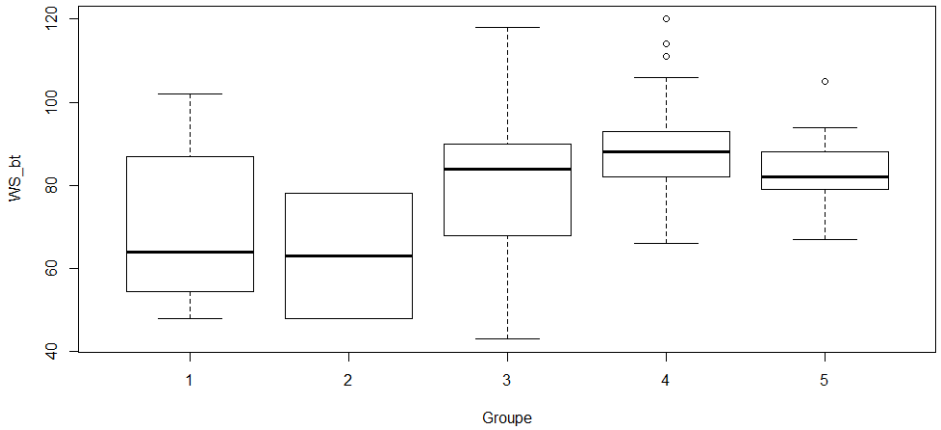

Рисунок 6 – Диаграмма размаха показателя WS_bt

Проверим наличие выбросов показателя:

```
> boxplot(WS_bt~Groupe, data = DataBT)$out
[1] 111 120 114 105
```

Выбросы наблюдаются в четвертой и пятой группах, но действительным выбросом будем считать только максимальное значение 120, остальные значения выбросов лежат в пределах значений других групп, поэтому могут быть допустимыми. Диаграмма размаха после удаления не допустимых значений представлена на рисунке 7.

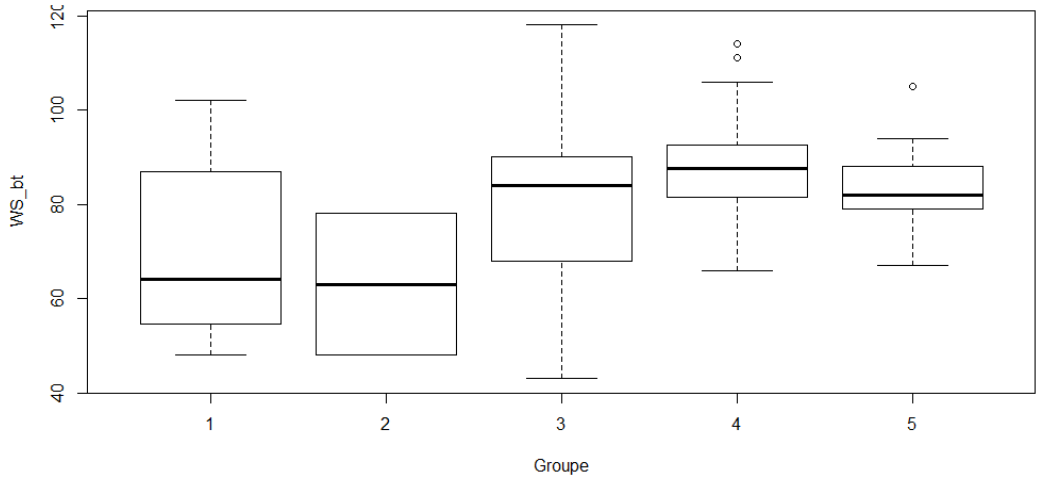

Рисунок 7 – Диаграмма размаха показателя WS_bt после удаления выбросов

Диаграмма размаха показателя TS_bt (рисунок 8) показывает, что значения выбросов находятся в пределах допустимых значений групп, поэтому значения данной группы остаются без изменений.

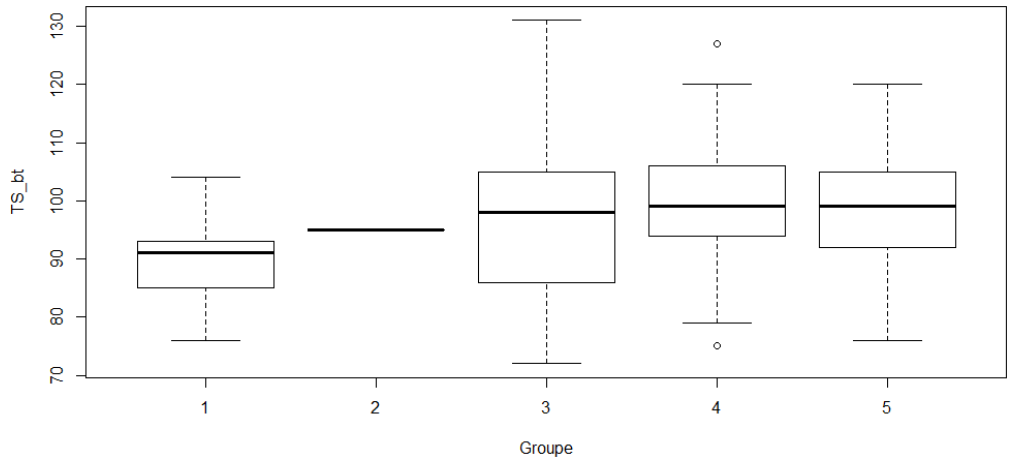

Рисунок 8 – Диаграмма размаха показателя TS_bt

При первичной обработке данных были проанализированы различные случаи наличия и обработки выбросов. Попадались показатели без выбросов, с множеством выбросов, с одним или двумя выбросами. Примером диаграммы размах с одним выбросом является показатель T_sup_bt (рисунок 9).

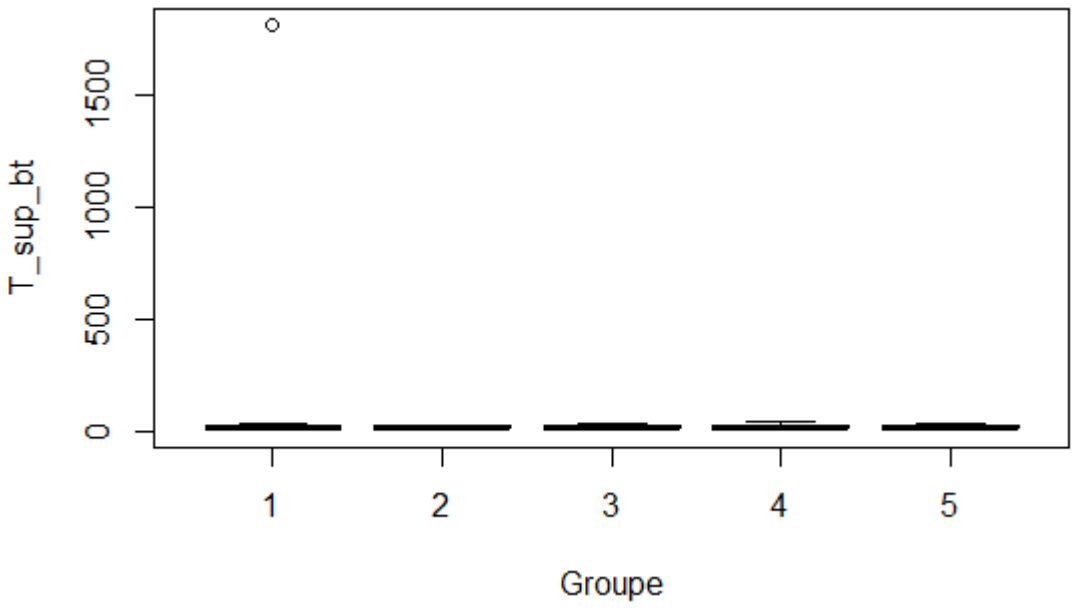

Рисунок 9 – Диаграмма размаха показателя T_sup_bt

Значение выброса значительно отличается от диапазона значений во всех группах и является экстремальным. Можно предположить, что этот выброс произошел в результаты ошибки ввода данных или опечатки

специалиста. Диаграмма размаха показателя T_sup_bt после удаления выброса представлена на рисунке 10.

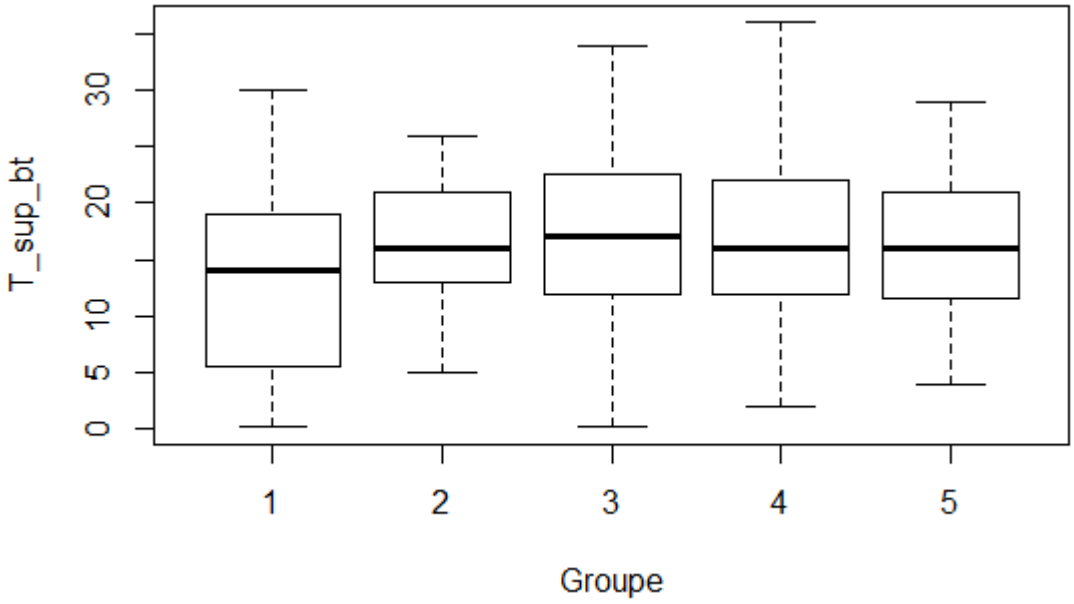

Рисунок 10 – Диаграмма размаха показателя T_sup_bt после удаления выбросов

Среди показателей набора данных были и такие, выбросы в которых были удалены полностью, т.к. не попадали в предел допустимых значений каждой из рассматриваемых групп лечения. Показатель HOMA_bt (индекс инсулин резистентности Хома) имеет значения в пределах от 0 до 7,5, выбросы находят выше границы пределов и подлежат удалению (рисунок 11).

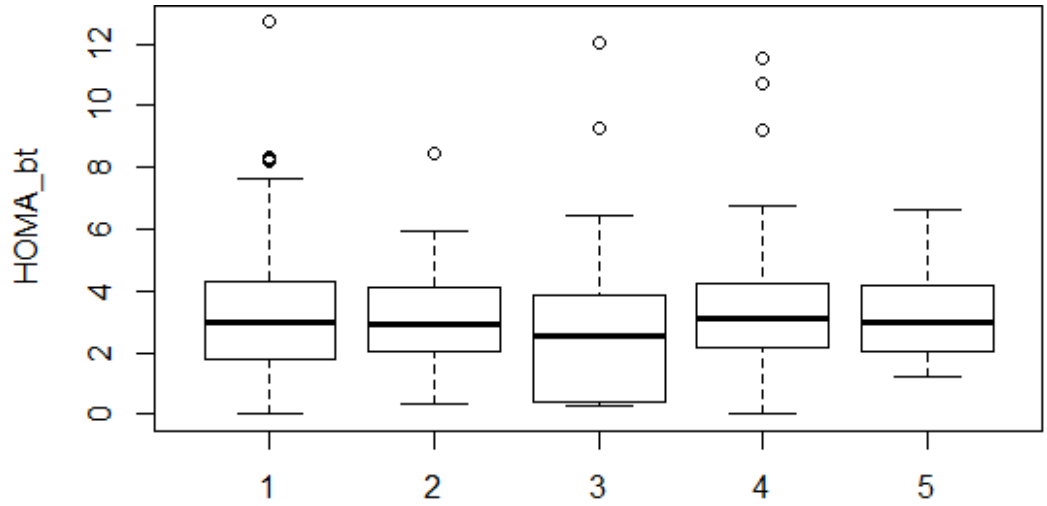

Рисунок 11 – Диаграмма размаха показателя HOMA_bt

После того как выбросы были удалены получена полностью очищенная от выбросов база, которую можно использовать для дальнейшего анализа и восстановления данных.

3.2 Восстановление пропущенных значений

3.1.2 Исследование пропущенных данных

Идентификация пропущенных значений обычно производится при помощи функций is.na() и complete.cases(). Проверка на наличие пропущенных значений по каждому признаку представлена ниже (рисунок12):

```
> summary(is.na(DataBT))
   Name          Groupe          Sex             Age          Degree_of_obesity Degree_of_obesity_by_BMI
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical     Mode :logical
 FALSE:276       FALSE:276       FALSE:270       FALSE:255       FALSE:243         FALSE:244
                                 TRUE :6         TRUE :21        TRUE :33          TRUE :32
 Degree_of_goiter  How_long        WS_bt           TS_bt          Mass_bt        Overweight_bt     BMI_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:159       FALSE:136       FALSE:114       FALSE:114       FALSE:244       FALSE:240       FALSE:240
 TRUE :117       TRUE :140       TRUE :162       TRUE :162       TRUE :32        TRUE :36        TRUE :36
   TMT_bt          SAD_bt          DAD_bt          HOMA_bt         TFN_bt          DP_bt          WorkC_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:136       FALSE:247       FALSE:244       FALSE:174       FALSE:124       FALSE:124       FALSE:54
 TRUE :140       TRUE :29        TRUE :32        TRUE :102       TRUE :152       TRUE :152       TRUE :222
   TBL_bt          TAG_bt          OXS_bt          aXC_bt          VLDL_bt          LDL_bt        Atherogenic_index_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:223       FALSE:221       FALSE:235       FALSE:222       FALSE:217       FALSE:216       FALSE:223
 TRUE :53        TRUE :55        TRUE :41        TRUE :54        TRUE :59        TRUE :60        TRUE :53
  Timol_bt       Glucose_bt        AP_bt         Catalase_bt       MDA_bt        sialic acids_bt Ceruloplasmin_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:28        FALSE:231       FALSE:27        FALSE:99        FALSE:97        FALSE:70        FALSE:95
 TRUE :248       TRUE :45        TRUE :249       TRUE :177       TRUE :179       TRUE :206       TRUE :181
 STTG_fasting    STTG_120m         IgA_bt          IgG_bt          IgM_bt          CIC_bt        Lysozyme_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:74        FALSE:73        FALSE:225       FALSE:223       FALSE:219       FALSE:219       FALSE:217
 TRUE :202       TRUE :203       TRUE :51        TRUE :53        TRUE :57        TRUE :57        TRUE :59
  T_lym_bt        T_hel_bt        T_sup_bt        B-lym_bt         T3_bt         T3_free_bt        T4_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:222       FALSE:221       FALSE:221       FALSE:221       FALSE:154       FALSE:87        FALSE:157
 TRUE :54        TRUE :55        TRUE :55        TRUE :55        TRUE :122       TRUE :189       TRUE :119
 T4_free_bt        TTG_bt         ATTPO_bt        Insulin_bt      Cortisol_bt     Leptin_bt        IL_4_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:97        FALSE:252       FALSE:114       FALSE:214       FALSE:210       FALSE:106       FALSE:66
 TRUE :179       TRUE :24        TRUE :162       TRUE :62        TRUE :66        TRUE :170       TRUE :210
  IL_6_bt          FNO_bt          IL_1B_bt        Garkavi_bt      Serotonin+bt      KK_bt           PI_bt
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:165       FALSE:157       FALSE:67        FALSE:93        FALSE:37        FALSE:211       FALSE:214
 TRUE :111       TRUE :119       TRUE :209       TRUE :183       TRUE :239       TRUE :65        TRUE :62
```

Рисунок 12 – Проверка на наличие пропущенных значений

Проверка на наличие строк, в которых нет пропущенных значений, показала, что в наборе данных таких строк нет. Список строк, в которых хотя бы одно пропущенное значение (рисунок 13), можно получить при помощи скрипта:

```
> DataBT[!complete.cases(DataBT),]
```

```
# A tibble: 276 x 66
   Name  Groupe   Sex   Age Degree_of_obesi~ Degree_of_obesi~ Degree_of_goiter How_long WS_bt TS_bt Mass_bt
   <chr>  <dbl> <dbl> <dbl>            <dbl>            <dbl>            <dbl>     <dbl> <dbl> <dbl>   <dbl>
 1 Pati~      1     1    11                1                1                1        NA    NA    NA      46
 2 Pati~      2     2    13                1                0                1        NA    48    95      72
 3 Pati~      3     1     9                1                1                0         3    NA    NA      40
 4 Pati~      3     2    10                1                0                1         4    43    83      39
 5 Pati~      3     1    14                1                0                0        NA    NA    NA      70
 6 Pati~      3     1     9                2                0                1         1    48    76      34
 7 Pati~      3     1    10                2                2                1         0    53    83      77
 8 Pati~      3     1    11                1                0                1         1    48    76      58
 9 Pati~      3     1    10                1                0                1         1    48    76      40
10 Pati~      3     2    15                2                0                0        NA    NA    NA      59
# ... with 266 more rows, and 55 more variables: Overweight_bt <dbl>, BMI_bt <dbl>, TMT_bt <dbl>, SAD_bt <dbl>,
```

Рисунок 13 – Список строк, в которых хотя бы одно пропущенное значение

Таким образом, набор данных не имеет полностью укомплектованных строк и имеет 276 строк хотя бы с одним пропущенным значением.

Дополнительную статистическую информацию о пропусках предоставляет функция md.pattern(), а функции aggr(), matrixplot() и scattMiss() позволяют отобразить на графиках закономерности во встречаемости отсутствующих наблюдений. Например, функция aggr() графически отображает число наблюдений для каждой отдельной переменной и для каждой комбинации переменных (рисунок 14).

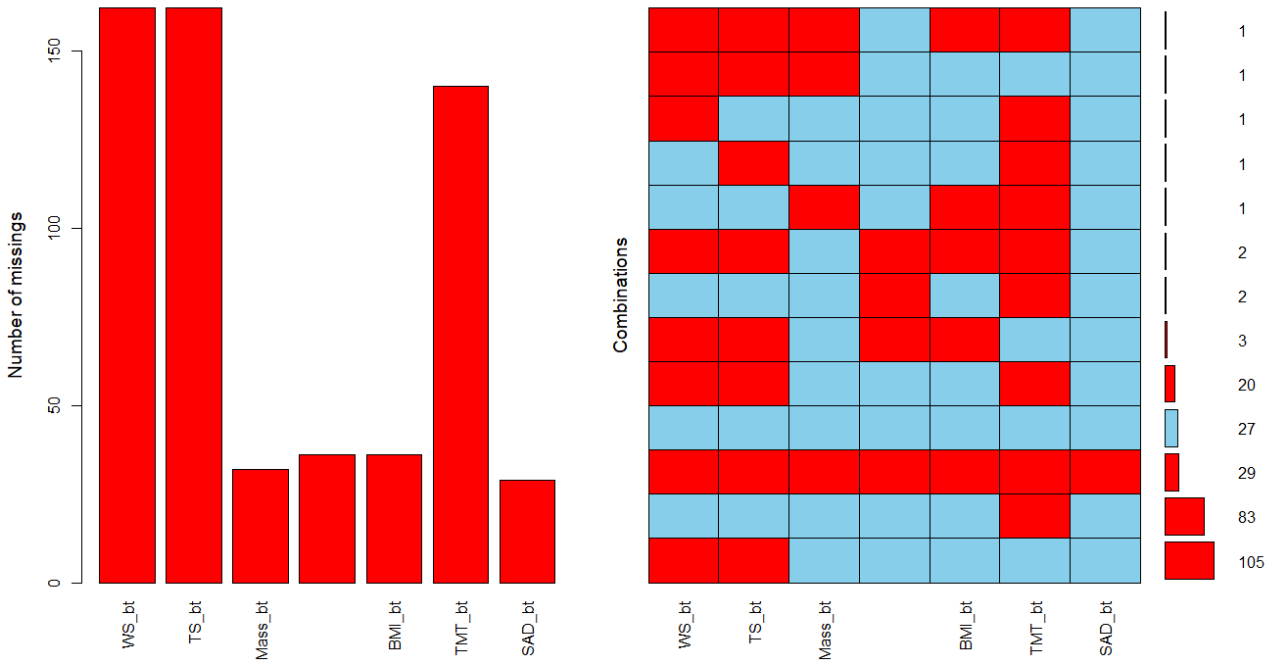

Рисунок 14 – Визуализация пропущенных значений функцией aggr()

На диаграмме слева представлено количество пропущенных значений для каждого показателя в отдельности, а справа – для комбинации показателей.

Далее исследуем закономерности появления отсутствующих значений в наборе данных. Для этого вычислим корреляцию между переменными. Заменим данные условными значениями: 1 – пропущенное значение, 0 – имеющееся значение. Полученную таблицу также называют матрицей теней (shadow matrix).

Результаты замены значений исходной таблицы представлены на рисунке 15. Для наглядности представлена часть исходной таблицы.

```
> x <- as.data.frame(abs(is.na(DataBT[,9:20])))
> head(DataBT[,9:20], n=10)
# A tibble: 10 x 12
   WS_bt TS_bt Mass_bt Overweight_bt BMI_bt TMT_bt SAD_bt DAD_bt HOMA_bt TFN_bt DP_bt WorkC_bt
   <dbl> <dbl>   <dbl>         <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl> <dbl>    <dbl>
 1    NA    NA      46          35.3   24.2   NA      100     50  NA        NA    NA         NA
 2    48    95      72          10.7   25     44.4    100     60   0.710    78   225         NA
 3    NA    NA      40          29     22.2   29.8    100     50  NA        NA    NA         NA
 4    43    83      39          22     20.7   28.4    100     55  NA        NA    NA         NA
 5    NA    NA      70          18     24     45.6    110     55   0.429    NA    NA         NA
 6    48    76      34          30.7   20.1   27.7    100     50   3.06     48.3  272         NA
 7    53    83      77          40     29.4   47      110     55   0.430    90    256.        NA
 8    48    76      58           9     23.5   39.7    105     55   0.331    65    195         NA
 9    48    76      40          21.2   21     31      105     55   0.266    56.6  245         NA
10    NA    NA      59          37.2   24.8   39.2    100     80   0.388    NA    NA          NA
> head(x, n=10)
   WS_bt TS_bt Mass_bt Overweight_bt BMI_bt TMT_bt SAD_bt DAD_bt HOMA_bt TFN_bt DP_bt WorkC_bt
1      1     1       0             0      0      1      0      0       1      1     1        1
2      0     0       0             0      0      0      0      0       0      0     0        1
3      1     1       0             0      0      0      0      0       1      1     1        1
4      0     0       0             0      0      0      0      0       1      1     1        1
5      1     1       0             0      0      0      0      0       0      1     1        1
6      0     0       0             0      0      0      0      0       0      0     0        1
7      0     0       0             0      0      0      0      0       0      0     0        1
8      0     0       0             0      0      0      0      0       0      0     0        1
9      0     0       0             0      0      0      0      0       0      0     0        1
10     1     1       0             0      0      0      0      0       0      1     1        1
```

Рисунок 15 – Замена исходных значений

Далее необходимо извлечь переменные, в которых пропущено несколько значений:

```
> y <- x[, which(colSums(x)>0)]
```

Вычисление корреляции между преобразованными переменными происходит следующим образом:

```
> print(cor(y),4)
```

В результате получаем корреляционную матрицу между частотами пропущенных значений (рисунок 16).

	WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	SAD_bt	DAD_bt	HOMA_bt	TFN_bt	DP_bt	WorkC_bt
WS_bt	1.0000	0.9851	0.2808	0.2812	0.3030	-0.4294	0.2874	0.2578	0.15444	0.5737	0.5737	0.55083
TS_bt	0.9851	1.0000	0.2808	0.2812	0.3030	-0.4294	0.2874	0.2578	0.15444	0.5885	0.5885	0.56938
Mass_bt	0.2808	0.2808	1.0000	0.8342	0.9014	0.3343	0.9462	0.9293	0.44955	0.2816	0.2816	0.17861
Overweight_bt	0.2812	0.2812	0.8342	1.0000	0.9361	0.3172	0.8847	0.8342	0.43898	0.3282	0.3282	0.19101
BMI_bt	0.3030	0.3030	0.9014	0.9361	1.0000	0.3172	0.8847	0.8678	0.46127	0.3282	0.3282	0.19101
TMT_bt	-0.4294	-0.4294	0.3343	0.3172	0.3172	1.0000	0.3377	0.3117	0.31921	-0.2637	-0.2637	-0.41303
SAD_bt	0.2874	0.2874	0.9462	0.8847	0.8847	0.3377	1.0000	0.9462	0.44753	0.2857	0.2857	0.16899
DAD_bt	0.2578	0.2578	0.9293	0.8342	0.8678	0.3117	0.9462	1.0000	0.42610	0.2816	0.2816	0.17861
HOMA_bt	0.1544	0.1544	0.4495	0.4390	0.4613	0.3192	0.4475	0.4261	1.00000	0.1332	0.1332	-0.03867
TFN_bt	0.5737	0.5885	0.2816	0.3282	0.3282	-0.2637	0.2857	0.2816	0.13319	1.0000	1.0000	0.54605
DP_bt	0.5737	0.5885	0.2816	0.3282	0.3282	-0.2637	0.2857	0.2816	0.13319	1.0000	1.0000	0.54605
WorkC_bt	0.5508	0.5694	0.1786	0.1910	0.1910	-0.4130	0.1690	0.1786	-0.03867	0.5460	0.5460	1.00000

Рисунок 16 – Корреляционная матрица

Рассмотрим связь между наличием пропущенных значений одной переменной и наблюдаемыми значениями других переменных:

```
> print(cor(DataBT[,9:20], y, use="pairwise.complete.obs"),4)
```

Результат представлен на рисунке 17.

	WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	SAD_bt	DAD_bt	HOMA_bt	TFN_bt	DP_bt	WorkC_bt
WS_bt	NA	-0.02381	0.23134	0.086238	0.2313388	0.604069	NA	0.12928	0.203268	-0.008869	-0.008869	-0.34382
TS_bt	0.077805	NA	0.27212	0.096147	0.2721239	0.345468	NA	0.10477	0.207519	-0.031000	-0.031000	-0.25780
Mass_bt	-0.009513	-0.02462	NA	0.006374	-0.0003187	0.025434	NA	-0.08998	0.064520	-0.144279	-0.144279	-0.10583
Overweight_bt	0.223605	0.21683	0.25488	NA	0.2916847	-0.087376	NA	0.07170	0.092857	0.133994	0.133994	0.12408
BMI_bt	0.067994	0.05728	0.15648	0.061812	NA	0.008128	NA	-0.08611	0.137503	-0.073870	-0.073870	-0.03165
TMT_bt	0.502010	0.50201	0.11315	0.139653	0.1396525	NA	NA	-0.04733	0.102565	0.097210	0.097210	0.19342
SAD_bt	-0.063560	-0.07315	0.10841	0.023522	0.0273644	0.126351	NA	0.01527	-0.007214	-0.102741	-0.102741	-0.15076
DAD_bt	-0.277530	-0.28627	0.09867	-0.101891	-0.0668051	0.286334	NA	NA	0.029662	-0.333317	-0.333317	-0.33301
HOMA_bt	-0.031844	-0.01077	0.09368	-0.123769	-0.0194540	0.245502	NA	-0.02136	NA	-0.136190	-0.136190	-0.03431
TFN_bt	-0.037250	-0.03861	-0.02474	0.050464	0.0504640	0.233809	0.05046	0.09576	0.033047	NA	NA	-0.13847
DP_bt	0.066554	0.06214	-0.06630	0.031851	0.0318514	0.154838	0.03185	0.12029	0.166540	NA	NA	-0.05581
WorkC_bt	-0.067082	-0.08016	NA	NA	NA	0.072578	NA	NA	-0.096349	NA	NA	NA

Рисунок 17 – Связь между наличием пропущенных значений

В полученной матрице строки являются наблюдаемыми переменные, а столбцы – преобразованными переменными, несущими информацию о наличии пропущенных значений. В первом столбце этой корреляционной матрицы можно заметить, что частота пропусков показателя WS_bt с большей вероятностью связана с показателем TMT_bt (0.502), показателем DAD_bt (0.277) и с показателем Overmeight_bt (0.223).

Другие столбцы интерпретируются аналогичным образом. Коэффициенты корреляции в матрице не слишком велики, исходя из чего, можно предположить, что пропуски не сильно отклоняются от полностью случайных величин и могут считаться случайными.

Согласно проведенному анализу на наличие пропущенных значений (рисунок 12) существует ряд параметров, в которых число пропусков очень велико. Создадим набор данных, в которых число пропусков в параметрах не превышает 176:

```
> f1 <- DataBT[,(colSums(is.na(DataBT)) < 176)]
```

После данного действия число параметров уменьшилось на 17. Проверка на наличие пропущенных значений в каждой строке таблицы также показывает, что число пропусков после удаления столбцов значительно уменьшилось (рисунок 18, 19).

```
> rowSums(is.na(DataBT))
  [1] 21 22 20 19 19 15 15 16 14 23 17 20 50 20 43 24 41 48 54 19 48 49 25 24 26 32 30 23 57 22 25 23 50 23 55
 [36] 26 22 18 19 19 55 24 16 24 26 17 22 26 49 22 20 22 23 24 24 20 22 21 22 49 21 22 26 27 25 51 25 52 24 22
 [71] 23 22 25 48 16 22 22 21 25 53 51 47 23 24 25 16 23 22 20 22 22 10 12 20 31 20 20 19 23 24 21 23 21 19 22
[106] 21 17 17 28 20 50 21 20 23 22 23 56 23 52 25 49 22 21 25 19 23 21 16 17 18 32 24 21 23 15 20 23 21 20 20
[141] 22 21 20 21 21 19 19 16 17 16 20 21 15 24 24 26 21 18 32 25 15 23 17 17 43 15 21 21 21 20 19 20 21 17 21
[176] 21 16 17 23 22 22 21 26 53 25 24 28 28 18 24 16 18 25 16 16 19 24 21 19 40 31 19 50 17 16 45 18 16 13 14
[211] 20 16 17 21 15 13 15 13 21 17 18 24 13 20 15 15 18 14 13 44 15 19 20 14 17 16 18 57 19 29 25 18 14 20 26
[246] 20 20 16 22 19 20 26 18 23 22 61 20 56 28 26 27 20 48 20 24 25 54 21 20 54 54 53 54 54 55 50
```

Рисунок 18 – Количество пропущенных значений по строкам до удаления параметров

```
> rowSums(is.na(f1))
  [1] 10 11 10  8  8  4  4  5  3 12  7  8 33  7 26 11 24 35 39  8 33 34 14 13 12 19 17 11 40 11 13 12 33 12 40
 [36] 15 11  7  8  7 38 11  5 10 13  5  9 12 32  9  7  9 10 11 10  4  5  6  6 33  4  6 11 10  8 34  9 37  9  6
 [71]  8  7  9 33  4  7  7  6  9 36 34 34  7  7  8  4 10  8  8 11 11  1  2  7 17  9  8  7  6 15  6  8  9  7  8
[106] 11  9  7 20 11 35 10  9 10  9 10 39 10 35 11 34  6 10 14  8  6  6  7  6  6 15  8  6 12  4  9 10 10  8  7
[141]  9  8  7  7  7  8  6  7  8  7  9  7  3  9  7  9  4  8 17  8  4 10  6  5 28  8  7  9  9  8  7  8  8  5  7
[176]  9  6  6 10 10  9 12 12 36 12 10 14 14  4 10  7  9 16  7  7  8 15  9  8 25 18  8 34  6  7 31  9  7  4  4
[211]  9  7  6  9  5  5  6  5 10  6  8 10  5  9  6  5  7  5  5 29  6  8  8  5  8  6  7 44 10 15 13  9  6 10 11
[246] 10 10  6  9  8 10 12  7 12 11 46 10 42 17 15 16 10 37 10 12 14 42 11 10 41 41 40 41 41 39 37
```

Рисунок 19 – Количество пропущенных значений по строкам после удаления параметров

3.1.3 Восстановление пропущенных данных

Для восстановления пропущенных значений воспользуемся методом множественного восстановления, при котором заполнение пропусков происходит при помощи повторного моделирования. Из существующего набора данных с пропущенными значениями создается несколько полных наборов данных, к каждому из которых применяются стандартные статистические методы для формирования окончательных результатов. Идея множественного восстановления пропущенных данных хорошо реализована в таком пакете R, как mice.

В mice многомерное восстановление данных реализуется при помощи связанных уравнений. Пропущенные значения замещаются при помощи выборок Гиббса. По умолчанию значения каждой переменной, содержащей пропущенные значения, предсказываются по значениям остальных переменных. Полученные уравнения используются для замещения пропущенных данных подходящими значениями. Этот процесс повторяется, пока значения для пропущенных данных не сойдутся.

Для восстановления данных выполним следующую последовательность скриптов:

```
> f2_mice <- mice(f2)
> f2_mice_output <- complete(f2_mice)
```

Результатом работы функции mice() стала база с восстановленными значениями. Фрагменты выборки до и после восстановления данных представлены на рисунке 20.

	WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	SAD_bt	DAD_bt	HOMA_bt	TFN_bt	DP_bt
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	NA	NA	46	35.3	24.2	NA	100	50	NA	NA	NA
2	48	95	72	10.7	25	44.4	100	60	0.710	78	225
3	NA	NA	40	29	22.2	29.8	100	50	NA	NA	NA
4	43	83	39	22	20.7	28.4	100	55	NA	NA	NA
5	NA	NA	70	18	24	45.6	110	55	0.429	NA	NA
6	48	76	34	30.7	20.1	27.7	100	50	3.06	48.3	272
7	53	83	77	40	29.4	47	110	55	0.430	90	256.
8	48	76	58	9	23.5	39.7	105	55	0.331	65	195
9	48	76	40	21.2	21	31	105	55	0.266	56.6	245
10	NA	NA	59	37.2	24.8	39.2	100	80	0.388	NA	NA

	WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	SAD_bt	DAD_bt	HOMA_bt	TFN_bt	DP_bt
1	84	76	46	35.3	24.2	50.9	100	50	2.2540179	85.0	268.6
2	48	95	72	10.7	25.0	44.4	100	60	0.7098214	78.0	225.0
3	78	76	40	29.0	22.2	29.8	100	50	2.0718750	71.7	226.8
4	43	83	39	22.0	20.7	28.4	100	55	2.9142857	65.0	267.2
5	75	99	70	18.0	24.0	45.6	110	55	0.4285714	85.0	278.4
6	48	76	34	30.7	20.1	27.7	100	50	3.0598214	48.3	272.0
7	53	83	77	40.0	29.4	47.0	110	55	0.4303125	90.0	256.5
8	48	76	58	9.0	23.5	39.7	105	55	0.3312946	65.0	195.0
9	48	76	40	21.2	21.0	31.0	105	55	0.2656250	56.6	245.0
10	54	98	59	37.2	24.8	39.2	100	80	0.3881250	85.0	216.0

Рисунок 20 – Восстановление пропущенных значений

3.3 Сокращение размерности пространства признаков

Для сокращения размерности существует два типа методов (рисунок21):

1. Выбор параметров (при этом сохраняются только самые релевантные переменные из исходного набора данных).

2. Сокращение размерности (сопровождается преобразованием параметров; при этом находится меньший набор новых переменных, каждая из которых является комбинацией входных переменных, содержащих в основном ту же информацию, что и входные переменные).

Рисунок 21 – Классификация методов уменьшения размерности

Методы сокращения размерности с преобразованием переменных, как правило, сопровождаются проблемой интерпретации новых переменных и определения вклада исходных переменных. Поэтому при решении поставленных задач данные методы использоваться не будут. Рассмотрим методы выбора параметров для уменьшения размерности нашего набора данных.

Высокая корреляция между двумя переменными означает, что они имеют сходные тенденции и, вероятно, несут схожую информацию. Это может резко снизить производительность некоторых моделей. Возможно рассчитать корреляцию между независимыми числовыми переменными, если коэффициент корреляции пересекает определенное пороговое значение, можно отбросить одну из переменных. Отбрасывание переменной очень субъективно и всегда должно выполняться с учетом предметной области. Как правило, предлагается сохранять те переменные, которые показывают среднюю или высокую корреляцию с целевой переменной.

Найдем атрибуты, которые сильно коррелированы. Функция findCorrelation() находит предикторы, чей уровень корреляции с другими предикторами в среднем превышает некоторый заданный порог (cutoff):

```
> highCor <- findCorrelation(cor(f2mo_red), cutoff = 0.7)
> names(f2mo_red[highCor])
```

```
[1] "BMI_bt" "Overweight_bt" "Atherogenic_index_bt" "T_lym_bt" "HO
MA_bt" "LDL_bt"
```

Переменные, связанные линейными зависимостями, можно считать малоинформативными, потому как они дублируют информацию. Для нахождения и исключения таких переменных можно использовать функцию findLinearCombos():

```
> findLinearCombos(f2mo_red)
$linearCombos
list()
$remove
NULL
```

Таким образом, переменные с линейными комбинациями отсутствуют.

Переменные с низкой дисперсией не влияют на целевую переменную, поэтому их можно отбросить для дальнейшего рассмотрения. Выведем информацию со значениями дисперсии для каждой переменной:

```
> summary(aov(Groupe~., data=f2mo_red))
                         Df Sum Sq Mean Sq F value   Pr(>F)
Sex                       1   0.29    0.29   0.424 0.515371
Age                       1   0.31    0.31   0.444 0.506044
Degree_of_obesity         1  34.16   34.16  49.430 2.32e-11 ***
Degree_of_obesity_by_BMI  1   0.20    0.20   0.294 0.587922
Degree_of_goiter          1   1.51    1.51   2.188 0.140428
How_long                  1   4.80    4.80   6.946 0.008971 **
WS_bt                     1  82.97   82.97 120.056  < 2e-16 ***
TS_bt                     1  29.06   29.06  42.047 5.38e-10 ***
Mass_bt                   1   1.21    1.21   1.754 0.186686
Overweight_bt             1   9.31    9.31  13.467 0.000302 ***
BMI_bt                    1   3.29    3.29   4.754 0.030247 *
TMT_bt                    1  21.42   21.42  30.993 7.19e-08 ***
SAD_bt                    1   6.08    6.08   8.799 0.003332 **
DAD_bt                    1  22.01   22.01  31.854 4.87e-08 ***
HOMA_bt                   1   0.44    0.44   0.633 0.427219
TFN_bt                    1   0.20    0.20   0.295 0.587608
DP_bt                     1  16.49   16.49  23.864 1.94e-06 ***
TBL_bt                    1   0.00    0.00   0.005 0.946493
TAG_bt                    1   9.77    9.77  14.143 0.000215 ***
OXS_bt                    1   4.00    4.00   5.781 0.016994 *
aXC_bt                    1   9.16    9.16  13.253 0.000336 ***
LDL_bt                    1   6.35    6.35   9.184 0.002720 **
Atherogenic_index_bt      1   0.26    0.26   0.382 0.537120
Glucose_bt                1   0.57    0.57   0.825 0.364629
IgA_bt                    1   0.38    0.38   0.548 0.460017
IgG_bt                    1   0.57    0.57   0.831 0.362980
IgM_bt                    1   5.18    5.18   7.502 0.006648 **
CIC_bt                    1   0.98    0.98   1.418 0.235030
Lysozyme_bt               1   0.58    0.58   0.841 0.359950
T_lym_bt                  1  11.02   11.02  15.951 8.75e-05 ***
T_hel_bt                  1   0.02    0.02   0.034 0.853340
T_sup_bt                  1   0.40    0.40   0.583 0.445783
T3_bt                     1   3.66    3.66   5.299 0.022232 *
T4_bt                     1   3.16    3.16   4.578 0.033447 *
TTG_bt                    1   0.89    0.89   1.285 0.258224
ATTPO_bt                  1  33.78   33.78  48.887 2.92e-11 ***
Insulin_bt                1   3.66    3.66   5.296 0.022268 *
Cortisol_bt               1   1.39    1.39   2.018 0.156783
Leptin_bt                 1   0.82    0.82   1.184 0.277659
```

```
IL_6_bt                   1   0.02    0.02   0.024 0.876238
FNO_bt                    1   0.01    0.01   0.011 0.917037
KK_bt                     1   0.08    0.08   0.110 0.740962
PI_bt                     1   4.07    4.07   5.884 0.016047 *
MG_bt                     1   0.01    0.01   0.008 0.927473
APF_bt                    1   9.49    9.49  13.735 0.000264 ***
Residuals               230 158.95    0.69
```

Элементы в последнем столбце обычно называются p-значениями. Если значение p меньше 0,05 (одна звезда, *), то влияние фактора, связанного с этой строкой, можно с уверенностью объявить не равным нулю. Если оно больше, чем 0,05, то из эксперимента нет убедительных доказательств того, что конкретный фактор или взаимодействие оказывает влияние. Метод сравнения средних значений ANOVA очень полезен для определения того, какие переменные наиболее важны для их воздействия на переменную ответа.

Важность переменных может быть оценена из данных путем построения модели. Некоторые методы, такие как деревья решений, имеют встроенный механизм для сообщения о важности переменной. Для других алгоритмов важность может быть оценена с использованием анализа кривой ROC, проведенного для каждого атрибута. Проведем классификацию показателей по их важности с использованием пакета Caret:

```
> control <- trainControl(method="repeatedcv",number=10,repeats=3)
> model <- train(Groupe~.,data=f2mo_red,method="treebag",trControl
=control)
> importance <- varImp(model, scale=FALSE)
> print(importance)
treebag variable importance
  only 20 most important variables shown (out of 45)
                Overall
DAD_bt           0.7821
Leptin_bt        0.7770
TS_bt            0.7549
ATTPO_bt         0.7377
T3_bt            0.6511
Overweight_bt    0.6506
APF_bt           0.6439
WS_bt            0.6333
Cortisol_bt      0.5906
FNO_bt           0.5897
TMT_bt           0.5663
KK_bt            0.5598
T_lym_bt         0.5563
MG_bt            0.5550
HOMA_bt          0.4897
Insulin_bt       0.4828
T_hel_bt         0.4781
T4_bt            0.4637
aXC_bt           0.4555
```

```
PI_bt          0.4528
```

График важности отдельных показателей пациентов представлен на рисунке 22.

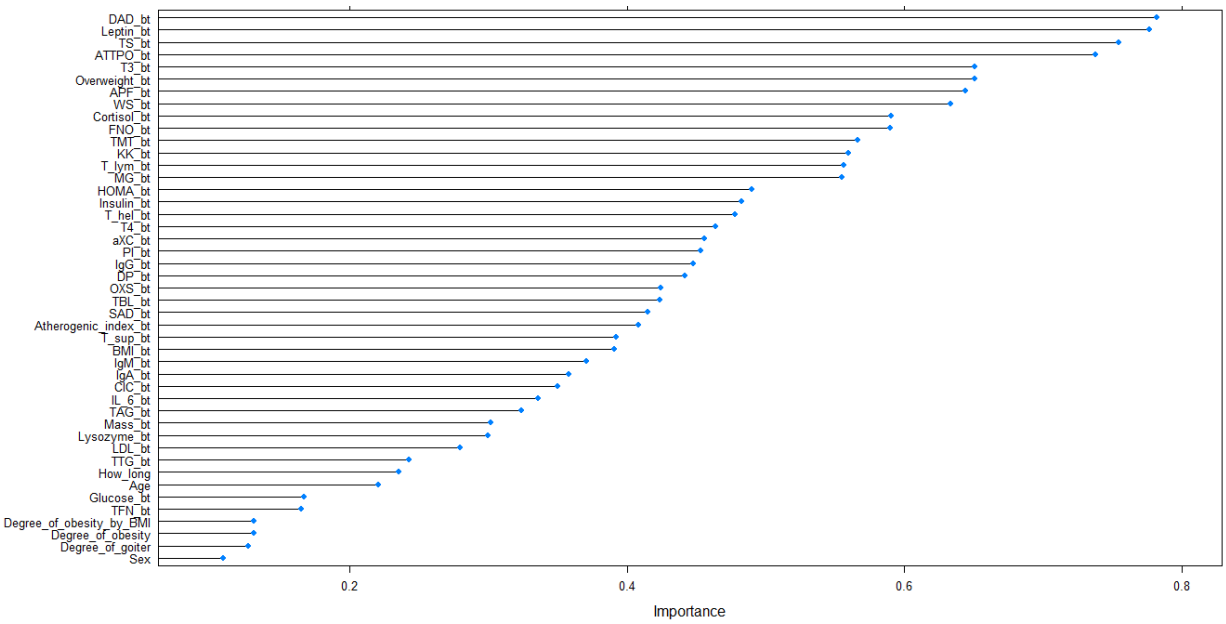

Рисунок 22 – Классификация важности показателей

Случайный лес (Random Forest) является одним из наиболее широко используемых алгоритмов выбора объектов. В него уже встроена функция важности, поэтому нет необходимости рассчитывать ее отдельно. Случайный лес может помочь выбрать меньшее подмножество функций. Выведем значения важности для переменных:

```
> rf <- randomForest(Groupe ~ ., data=f2mo_rf, importance=TRUE)
> importance(rf, type = 2)
                           MeanDecreaseGini
Sex                               0.7209427
Age                               3.2137342
Degree_of_obesity                 1.7246332
Degree_of_obesity_by_BMI          1.8039548
Degree_of_goiter                  0.8701017
How_long                          3.7189668
WS_bt                             7.4034923
TS_bt                             6.1718883
Mass_bt                           4.5994988
Overweight_bt                     5.4348166
BMI_bt                            4.4892699
TMT_bt                            7.6281254
SAD_bt                            3.9879660
DAD_bt                            7.6960611
HOMA_bt                           4.7683790
TFN_bt                            2.4449685
DP_bt                             3.7902276
TBL_bt                            5.1391466
TAG_bt                            4.1032930
OXS_bt                            4.7019532
aXC_bt                            5.2306074
LDL_bt                            3.7658734
```

Atherogenic_index_bt	4.2811051
Glucose_bt	3.1768048
IgA_bt	3.6480348
IgG_bt	4.7648218
IgM_bt	4.5826829
CIC_bt	7.3325545
Lysozyme_bt	4.5361678
T_lym_bt	6.0374172
T_hel_bt	4.8838829
T_sup_bt	4.3475970
T3_bt	4.7805505
T4_bt	4.4564026
TTG_bt	4.4214703
ATTPO_bt	7.6851737
Insulin_bt	5.1110219
Cortisol_bt	4.5862843
Leptin_bt	6.9143389
IL_6_bt	5.8171877
FNO_bt	4.2549143
KK_bt	5.7680306
PI_bt	5.2744550
MG_bt	4.7098210
APF_bt	5.0398580

Переменные с высокой важностью являются движущими факторами результата, и их значения оказывают значительное влияние на итоговые значения. Напротив, переменные с низкой важностью могут быть опущены в модели, что упрощает и ускоряет процесс подбора и прогнозирования.

Наглядно результаты градации признаков по степени важности представлены на рисунке 23.

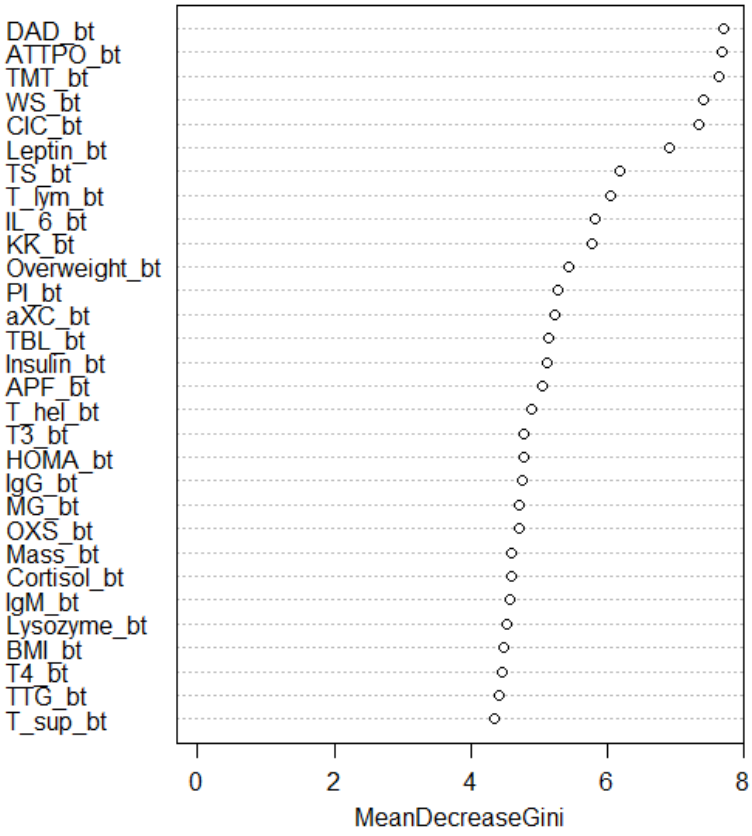

Рисунок 23 – График градации признаков по степени важности (Random Forest)

Сводная таблица признаков для сокращения размерности (таблица 3) содержит значения переменных, рекомендованных к удалению, со знаком «-», и выявленных как значимые переменные – со знаком «+».

Таблица 3 – Результат анализа на сокращение размерности признаков

Correlation (-)	**Variance (+)**	**Caret Importance (+)**	**Random Forest (+)**
BMI_bt	Degree_of_obesity	DAD_bt	DAD_bt
Overweight_bt	WS_bt	Leptin_bt	ATTPO_bt
Atherogenic_index_bt	TS_bt	TS_bt	TMT_bt
T_lym_bt	DAD_bt	ATTPO_bt	WS_bt
HOMA_bt	Overweight_bt	T3_bt	CIC_bt
LDL_bt	DP_bt	Overweight_bt	Leptin_bt
	TMT_bt	APF_bt	TS_bt
	aXC_bt	WS_bt	T_lym_bt
	TAG_bt	Cortisol_bt	IL_6_bt
	T_lym_bt	FNO_bt	KK_bt
	ATTPO_bt	TMT_bt	Overweight_bt
	APF_bt	KK_bt	PI_bt
		T_lym_bt	aXC_bt
		MG_bt	TBL_bt
		HOMA_bt	Insulin_bt
		Insulin_bt	APF_bt
		T_hel_bt	T_hel_bt
		T4_bt	T3_bt
		aXC_bt	HOMA_bt
		PI_bt	IgG_bt
			MG_bt
			OXS_bt
			Mass_bt
			Cortisol_bt
			IgM_bt
			Lysozyme_bt
			BMI_bt
			T4_bt
			TTG_bt
			T_sup_bt

Исходя из проведенного анализа, оставим в качестве информативных признаков только признаки из колонок 2-4. Таким образом, полученный набор данных содержит 35 атрибутов (рисунок 24).

	Groupe	Degree_of_obesity	WS_bt	TS_bt	Mass_bt	Overweight_bt	BMI_bt	TMT_bt	DAD_bt	HOMA_bt
1	1	1	84.0	76.0	46.0	35.3	24.20	50.9	50	2.2540179
2	2	1	48.0	95.0	72.0	10.7	25.00	44.4	60	0.7098214
3	3	1	78.0	76.0	40.0	29.0	22.20	29.8	50	2.0718750
4	3	1	43.0	83.0	39.0	22.0	20.70	28.4	55	2.9142857
5	3	1	75.0	99.0	70.0	18.0	24.00	45.6	55	0.4285714
6	3	2	48.0	76.0	34.0	30.7	20.10	27.7	50	3.0598214
7	3	2	53.0	83.0	77.0	40.0	29.40	47.0	55	0.4303125
8	3	1	48.0	76.0	58.0	9.0	23.50	39.7	55	0.3312946

Showing 1 to 10 of 276 entries, 35 total columns

Рисунок 24 – Фрагмент набора данных после сокращения размерности

3.4 Построение дерева решений

Исходными данными для построения дерева решений является набор данным, полученный в результате сокращения размерности. Данные были поделены на обучающую и тестовую выборки в соотношении 80/20. Реализация построения и графическое представление (рисунок 25) дерева решения в R выглядит следующим образом:

```
> model <- rpart(Groupe~., train, method = "class")
> prp(model, faclen = 0, cex = 0.8, extra = 0, fallen.leaves=TRUE)
```

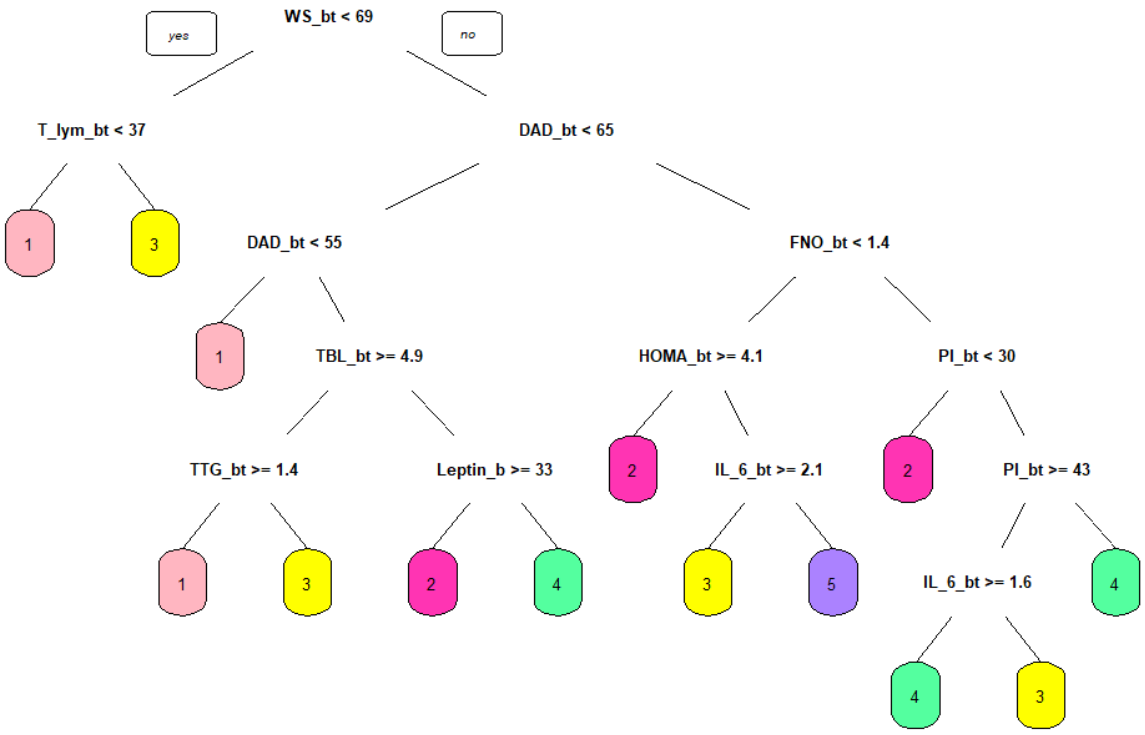

Рисунок 25 – Графическое построение дерева решений в R

Логические правила полученной модели дерева решений:

- Если WS_bt< 68.5 и
  - если T_lym_bt< 36.5, то 1 группа
  - если T_lym_bt>=36.5, то 3 группа
- Если WS_bt>=68.5 и
  - если DAD_bt< 64.5 и
    - если DAD_bt< 54.5, то 1 группа
    - если DAD_bt>=54.5 и
      - если TBL_bt>=4.85 и
        - если TTG_bt>=1.35, то 1 группа
        - если TTG_bt< 1.35, то 3 группа
      - если TBL_bt< 4.85 и
        - если Leptin_bt>=33.05, то 2 группа
        - если Leptin_bt<33.05, то 4 группа
    - если DAD_bt>=64.5 и
      - если FNO_bt< 1.35 и
        - если HOMA_bt>=4.079464, то 2 группа
        - если HOMA_bt< 4.079464 и
          - если IL_6_bt>=2.1, то 3 группа
          - если IL_6_bt< 2.1, то 5 группа
      - если FNO_bt>=1.35 и
        - если PI_bt< 29.61, то 2 группа
        - если PI_bt>=29.61 и
          - если PI_bt>=42.535 и
            - если IL_6_bt>=1.55, то 4 группа
            - если IL_6_bt< 1.55, то 3 группа
        - если PI_bt< 42.535, то 4 группа

Сделаем прогноз на тестовых данных и рассчитаем точность модели по данным испытаний:

```
> predicted <- predict(model, test, type ="class")
> mean(predicted == test$Groupe)
[1] 0.8284906
```

Построим дерево решений в RapidMiner. Исходным набором данных является полученный в результате анализа и сокращения размерности в R набор, который был поделен на тренировочную и тестовую выборку. Процесс построения дерева решений в RapidMiner с использованием оператора Decision Tree представлен на рисунке 26. Графическое построение дерева решений представлено на рисунке 27.

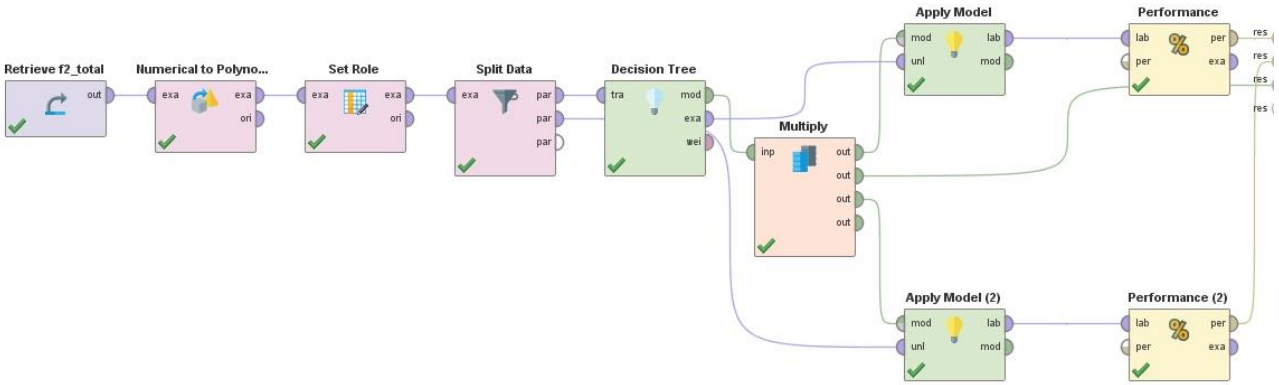

Рисунок 26 – Процесс построения дерева решений (Decision Tree)

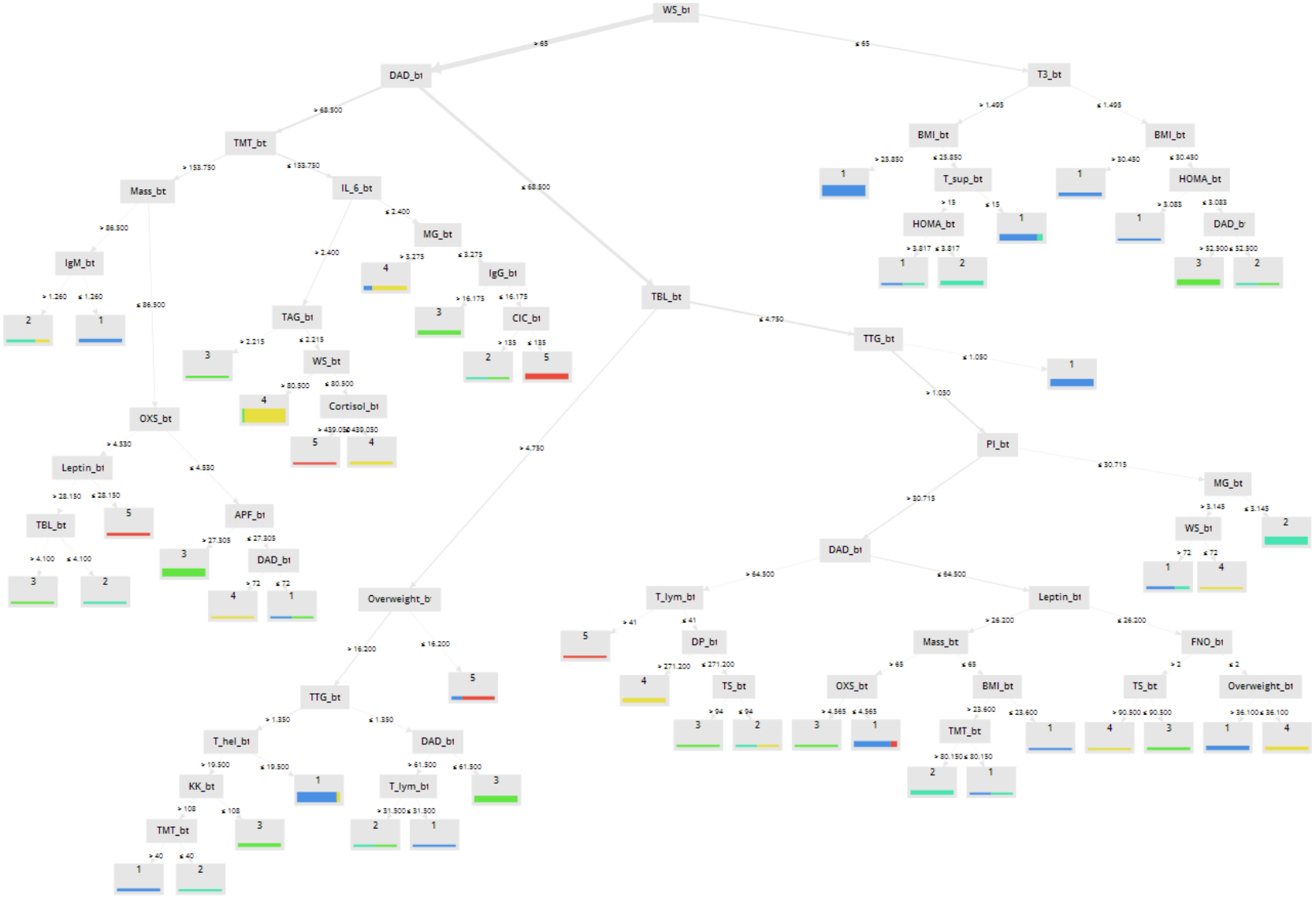

Рисунок 27 – Графическое построение дерева решений в RapidMiner (Decision Tree)

Полученный в результате построения дерева решений набор правил приведен в Приложении Г. Фрагмент набора правил представлен на рисунке 28.

Tree

```
WS_bt > 65
|   DAD_bt > 68.500
|   |   TMT_bt > 153.750
|   |   |   Mass_bt > 86.500
|   |   |   |   IgM_bt > 1.260: 2 {1=0, 2=2, 3=0, 4=1, 5=0}
|   |   |   |   IgM_bt ≤ 1.260: 1 {1=4, 2=0, 3=0, 4=0, 5=0}
|   |   |   Mass_bt ≤ 86.500
|   |   |   |   OXS_bt > 4.530
|   |   |   |   |   Leptin_bt > 28.150
|   |   |   |   |   |   TBL_bt > 4.100: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   |   |   TBL_bt ≤ 4.100: 2 {1=0, 2=2, 3=0, 4=0, 5=0}
|   |   |   |   |   Leptin_bt ≤ 28.150: 5 {1=0, 2=0, 3=0, 4=0, 5=3}
|   |   |   |   OXS_bt ≤ 4.530
|   |   |   |   |   APF_bt > 27.305: 3 {1=0, 2=0, 3=10, 4=0, 5=0}
|   |   |   |   |   APF_bt ≤ 27.305
|   |   |   |   |   |   DAD_bt > 72: 4 {1=0, 2=0, 3=0, 4=2, 5=0}
|   |   |   |   |   |   DAD_bt ≤ 72: 1 {1=1, 2=0, 3=1, 4=0, 5=0}
|   |   TMT_bt ≤ 153.750
|   |   |   IL_6_bt > 2.400
|   |   |   |   TAG_bt > 2.215: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   TAG_bt ≤ 2.215
|   |   |   |   |   WS_bt > 80.500: 4 {1=0, 2=0, 3=1, 4=17, 5=0}
|   |   |   |   |   WS_bt ≤ 80.500
|   |   |   |   |   |   Cortisol_bt > 439.050: 5 {1=0, 2=0, 3=0, 4=0, 5=2}
|   |   |   |   |   |   Cortisol_bt ≤ 439.050: 4 {1=0, 2=0, 3=0, 4=3, 5=0}
|   |   |   IL_6_bt ≤ 2.400
|   |   |   |   MG_bt > 3.275: 4 {1=1, 2=0, 3=0, 4=4, 5=0}
|   |   |   |   MG_bt ≤ 3.275
|   |   |   |   |   IgG_bt > 16.175: 3 {1=0, 2=0, 3=5, 4=0, 5=0}
|   |   |   |   |   IgG_bt ≤ 16.175
|   |   |   |   |   |   CIC_bt > 135: 2 {1=0, 2=1, 3=1, 4=0, 5=0}
|   |   |   |   |   |   CIC_bt ≤ 135: 5 {1=0, 2=0, 3=0, 4=0, 5=7}
WS_bt ≤ 65
```

Рисунок 28 – Фрагмент набора правил (Decision Tree)

Точность построения модели составила 93% (рисунок 29). Точность тестирования составила 67%.

accuracy: 93.21%

	true 1	true 2	true 3	true 4	true 5	class precision
pred. 1	75	4	1	1	1	91.46%
pred. 2	0	30	3	2	0	85.71%
pred. 3	0	0	45	0	0	100.00%
pred. 4	1	0	1	39	0	95.12%
pred. 5	1	0	0	0	17	94.44%
class recall	97.40%	88.24%	90.00%	92.86%	94.44%	

Рисунок 29 – Точность построения модели (Decision Tree)

Процесс построения дерева решений в RapidMiner с использованием оператора Random Forest представлен на рисунке 30.

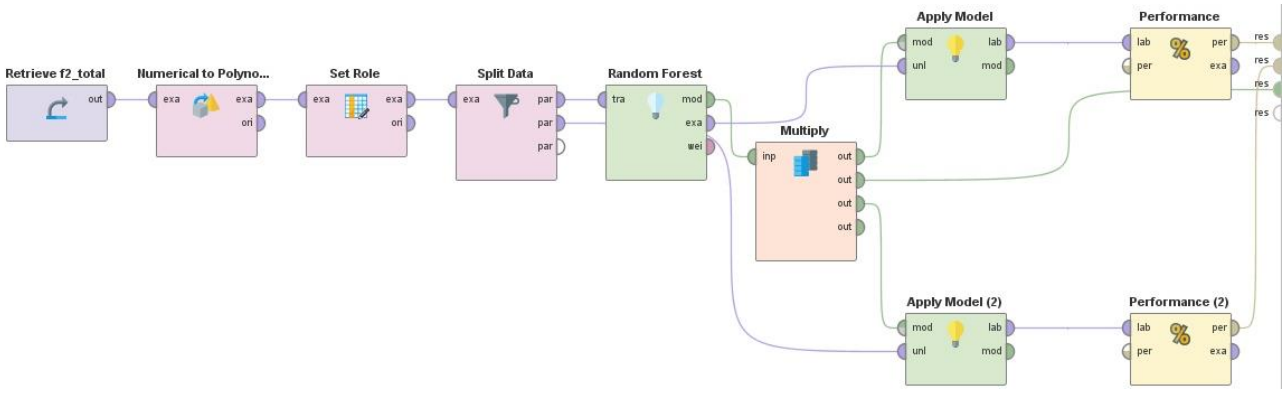

Рисунок 30 – Процесс построения дерева решений (Random Forest)

Точность построения модели составила 100% (рисунок 31). Точность тестирования составила 71%.

accuracy: 100.00%

	true 1	true 2	true 3	true 4	true 5	class precis
pred. 1	77	0	0	0	0	100.00%
pred. 2	0	34	0	0	0	100.00%
pred. 3	0	0	50	0	0	100.00%
pred. 4	0	0	0	42	0	100.00%
pred. 5	0	0	0	0	18	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	

Рисунок 31 - Точность построения модели (Random Forest)

Полученный в результате построения дерева решений набор правил приведен в Приложении Г. Фрагмент набора правил представлен на рисунке 32.

Tree

```
CIC_bt > 97.500
|   IgM_bt > 1.160
|   |   TTG_bt > 2.250
|   |   |   Lysozyme_bt > 38.500
|   |   |   |   T3_bt > 1.850
|   |   |   |   |   TMT_bt > 38.600: 2 {1=0, 2=9, 3=0, 4=0, 5=0}
|   |   |   |   |   TMT_bt ≤ 38.600: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   T3_bt ≤ 1.850
|   |   |   |   |   BMI_bt > 28.950: 3 {1=0, 2=0, 3=1, 4=0, 5=0}
|   |   |   |   |   BMI_bt ≤ 28.950: 5 {1=0, 2=0, 3=0, 4=0, 5=3}
|   |   |   Lysozyme_bt ≤ 38.500
|   |   |   |   PI_bt > 48.735
|   |   |   |   |   DAD_bt > 63.500: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   |   DAD_bt ≤ 63.500: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   PI_bt ≤ 48.735
|   |   |   |   |   DAD_bt > 77.500: 5 {1=0, 2=0, 3=0, 4=0, 5=1}
|   |   |   |   |   DAD_bt ≤ 77.500: 1 {1=8, 2=0, 3=0, 4=0, 5=0}
|   |   TTG_bt ≤ 2.250
|   |   |   Cortisol_bt > 358.550
|   |   |   |   DAD_bt > 72.500: 5 {1=0, 2=0, 3=0, 4=0, 5=2}
|   |   |   |   DAD_bt ≤ 72.500
|   |   |   |   |   DP_bt > 310.750: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   DP_bt ≤ 310.750: 4 {1=0, 2=0, 3=0, 4=5, 5=0}
|   |   |   Cortisol_bt ≤ 358.550
```

Рисунок 32 – Фрагмент набора правил (Random Forest)

Графическое построение дерева решений представлено на рисунке 33.

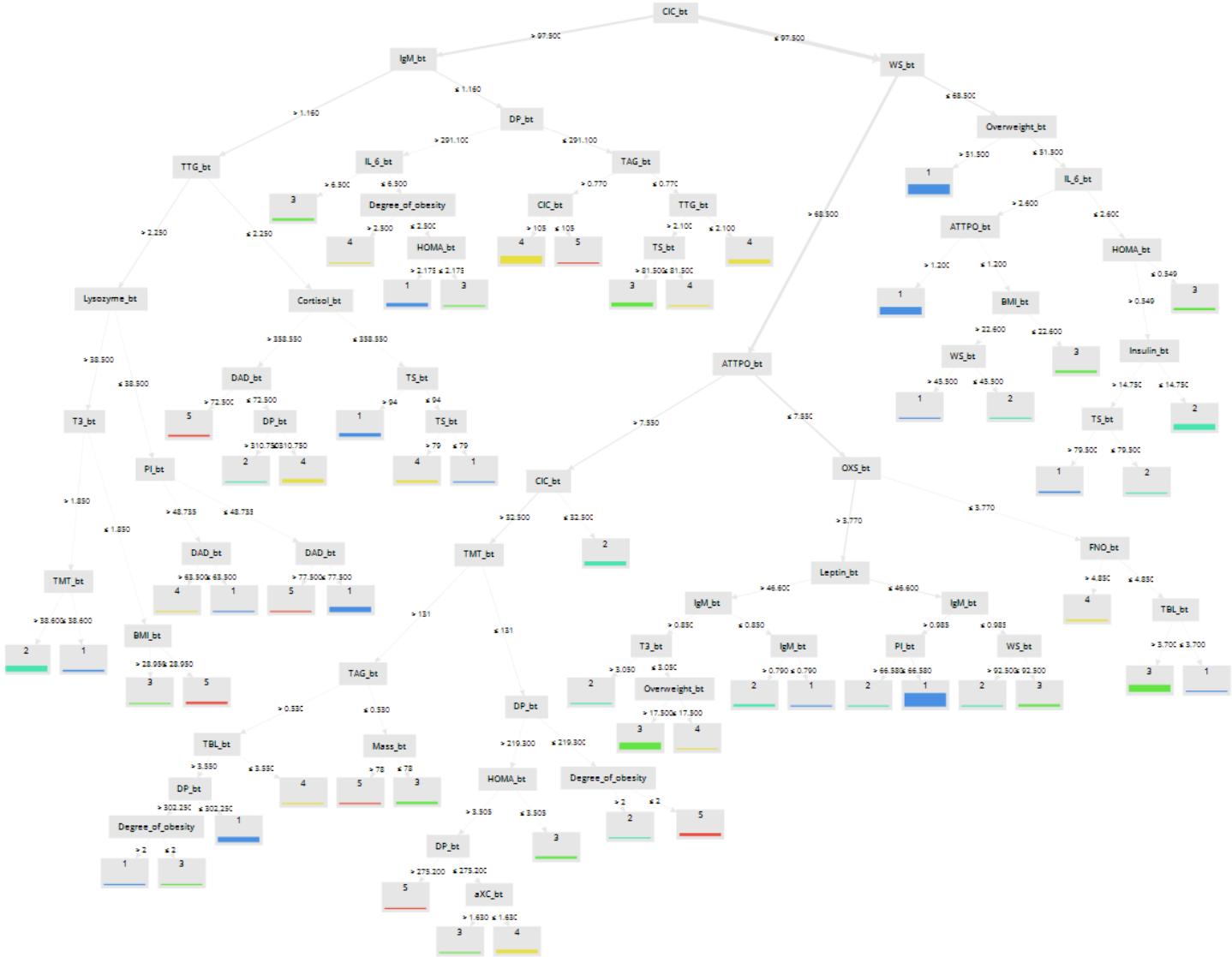

Рисунок 33 – Графическое построение дерева решений в RapidMiner (Random Forest)

3.5 Выбор решения и проверка результатов

В результате построения деревьев решений и анализа полученных результатов было решено выбрать модель дерева, построенного в R, точность которой оказалась выше. Для проверки качества построенной модели приведем пример самостоятельного визуального тестирования модели с данными пациентов. Выведем данные первого пациента из итогового набора данных (рисунок 34):

```
> f2_total[1,c("Groupe", "WS_bt", "T_lym_bt", "DAD_bt", "TBL_bt", "TTG_bt", "Leptin_bt",
 "FNO_bt", "HOMA_bt", "IL_6_bt", "PI_bt")]
  Groupe WS_bt T_lym_bt DAD_bt TBL_bt TTG_bt Leptin_bt FNO_bt  HOMA_bt IL_6_bt PI_bt
1      1    84       39     50    6.4    2.6      46.5    4.7 2.254018       0 21.38
```

Рисунок 34 – Показатели пациента №1

Результат проверки принадлежности пациента к первой группе лечения по модели дерева решений R представлен на рисунке 35.

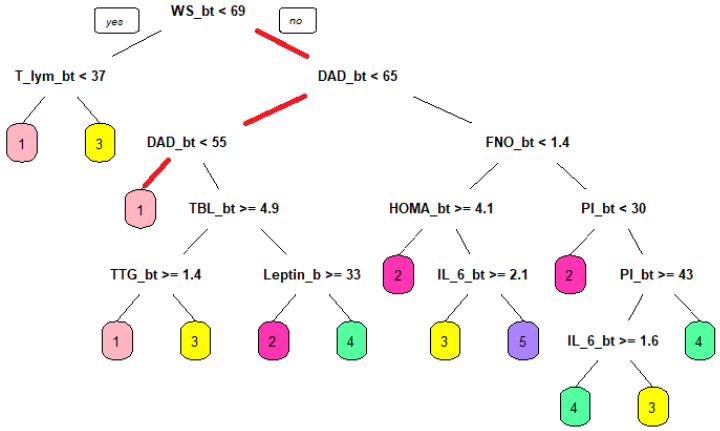

Рисунок 35 – Определение группы лечения пациента №1

Выведем данные семьдесят второго пациента из итогового набора данных (рисунок 36):

```
> f2_total[72,c("Groupe", "WS_bt", "T_lym_bt", "DAD_bt", "TBL_bt", "TTG_bt", "Leptin_bt",
 "FNO_bt", "HOMA_bt", "IL_6_bt", "PI_bt")]
   Groupe WS_bt T_lym_bt DAD_bt TBL_bt TTG_bt Leptin_bt FNO_bt  HOMA_bt IL_6_bt PI_bt
72      2    95       32     64      4    3.8        72      0 1.499107       3 51.05
```

Рисунок 36 – Показатели пациента №72

Результат проверки принадлежности пациента ко второй группе лечения по модели дерева решений представлен на рисунке 37.

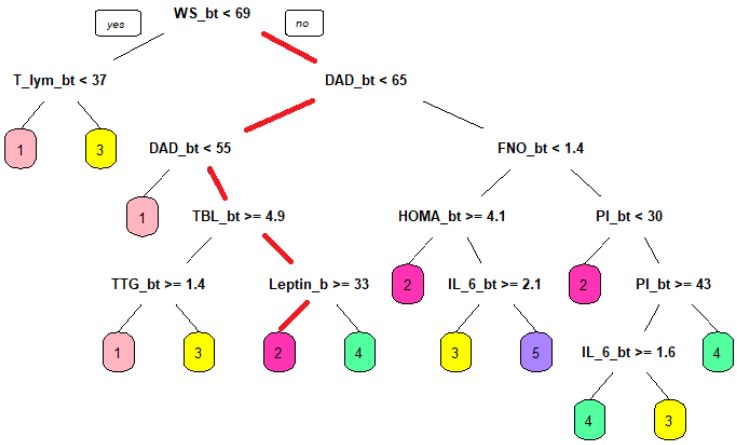

Рисунок 37 – Определение группы лечения пациента №72

Выведем данные семьдесят пациента №191 из итогового набора данных (рисунок 38):

```
> f2_total[191,c("Groupe", "WS_bt", "T_lym_bt", "DAD_bt", "TBL_bt", "TTG_bt", "Leptin_bt",
 "FNO_bt", "HOMA_bt", "IL_6_bt", "PI_bt")]
    Groupe WS_bt T_lym_bt DAD_bt TBL_bt TTG_bt Leptin_bt FNO_bt  HOMA_bt IL_6_bt PI_bt
191      5    94       27     80    4.1    1.1      23.8    1.1 1.499107       2 32.78
```

Рисунок 38 – Показатели пациента №191

Результат проверки принадлежности пациента к пятой группе лечения по модели дерева решений представлен на рисунке 39.

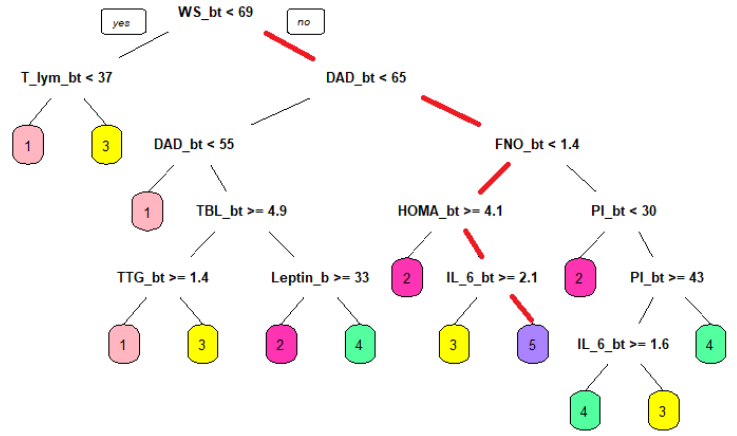

Рисунок 39 – Определение группы лечения пациента №191

Проверим качеством модели на данных пациентах исходного набора, которые не имели пропущенных значений в необходимых показателях. Выведем данные исходной выборки для девятого пациента с показателями полученной модели дерева решений (рисунок 40):

```
> DataBT[9,c("Name", "Groupe", "WS_bt", "T_lym_bt", "DAD_bt", "TBL_bt", "TTG_bt", "Leptin_bt",
 "FNO_bt", "HOMA_bt", "IL_6_bt", "PI_bt")]
# A tibble: 1 x 12
  Name      Groupe WS_bt T_lym_bt DAD_bt TBL_bt TTG_bt Leptin_bt FNO_bt HOMA_bt IL_6_bt PI_bt
  <chr>      <dbl> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>     <dbl>  <dbl>   <dbl>   <dbl> <dbl>
1 Patient 9      3    48       43     55    4.9    2.9      49.4    2.9   0.266     3.2  38.2
```

Рисунок 40 – Показатели пациента №9

Результат проверки принадлежности пациента к третьей группе лечения по модели дерева решений представлен на рисунке 41:

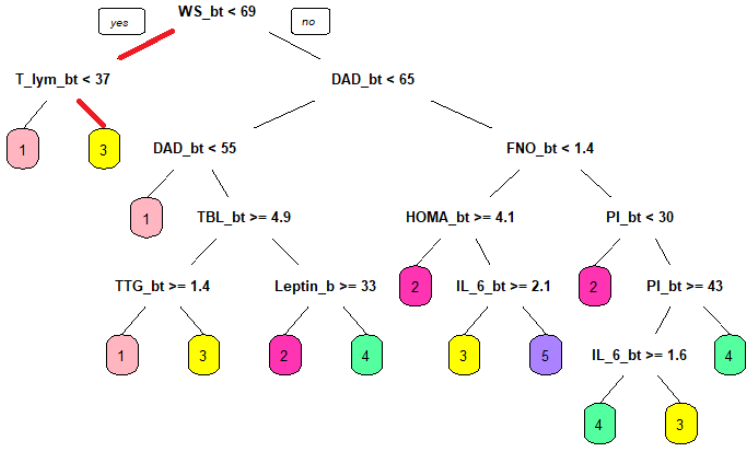

Рисунок 41 – Определение группы лечения пациента №9

Выведем данные исходной выборки для пациента №209 (рисунок 42):

```
> DataBT[209,c("Name", "Groupe", "WS_bt", "T_lym_bt", "DAD_bt", "TBL_bt", "TTG_bt", "Leptin_bt",
 "FNO_bt", "HOMA_bt", "IL_6_bt", "PI_bt")]
# A tibble: 1 x 12
  Name    Groupe WS_bt T_lym_bt DAD_bt TBL_bt TTG_bt Leptin_bt FNO_bt HOMA_bt IL_6_bt PI_bt
  <chr>    <dbl> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>     <dbl>  <dbl>   <dbl>   <dbl> <dbl>
1 Patien~      4    92       29     90    4.2    3.4      16.2    2.5    2.43     4.3  45.2
```

Рисунок 42 – Показатели пациента №209

Результат проверки принадлежности пациента к четвертой группе лечения по модели дерева решений представлен на рисунке 41:

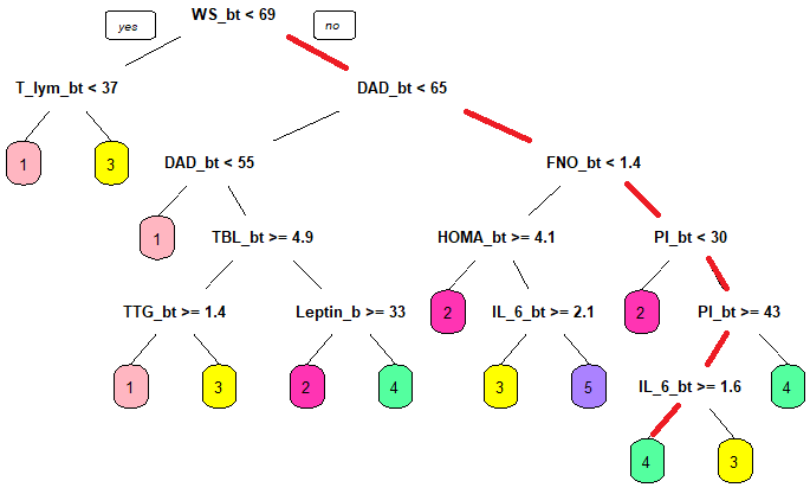

Рисунок 43 – Определение группы лечения пациента №209

Таким образом, в результате проведенного исследования был разработан алгоритм поддержки принятия решения о выборе траектории лечений детей с избыточным весом на основе построения дерева решения. Для разработки алгоритма были проведены следующие этапы:

- Получены результаты клинического исследования, проведенного на базе детского отделения Томского НИИ курортологии и физиотерапии.
- Проведена первичная обработка данных с помощью методов языка R. Был проведен анализ выбросов и восстановление пропущенных значений методом множественного восстановления.
- Сокращена размерность признакового пространства средствами анализа уровня корреляции, линейной зависимости, дисперсии и степени важности признаков.
- Сформирована итоговая выборка для построения и поиска модели дерева решения, которая была разделена на обучающую и тестовую.
- Проведено построение разных моделей деревьев решений и выбрана модель с наилучшими значениями точности.
- Проведена самостоятельная проверка качества работы модели принятия решения на основе полученного дерева решений.

Дерево решений позволило получить логические правила, по которым для нового пациента с диагностированным избыточным весом можно определить группу лечения по его клинико-лабораторным показателям.

Глава 4 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

В нижеприведенной главе дается оценка коммерческого потенциала и перспективности применения потребителями системы поддержки принятия решений, разработанной в дипломном проекте.

Магистерская диссертация направлена на разработку системы поддержки принятия решений о выборе траектории лечения для детей с избыточным весом. Система предназначена для усовершенствования процесса определения диагноза и назначения подходящего лечения. Для достижения поставленных целей необходимо было выполнить ряд задач, в том числе дать оценку коммерческих возможностей проведенного исследования, выявить его ресурсосберегающий потенциал, определить финансовую эффективности исследования.

Целью данного раздела является определение перспективности научно-исследовательского проекта. Задачами раздела являются:

- оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения;
- планирование работ по научно-исследовательскому проекту с использованием линейного графика;
- расчет бюджета научного-технического исследования;
- определение экономической эффективности исследования.

4.1 Оценка коммерческого потенциала и перспективности проведения научных исследований

4.1.1 Потенциальные потребители результатов исследования

Для анализа потенциальных потребителей необходимо провести анализ целевого рынка и его сегментирование.

Целевым рынком данного исследования являются медицинские учреждения. Можно выделить следующие критерии сегментирования рынка по разработке медицинских информационных систем: тип медицинского учреждения и тип медицинской информационной системы. На основе выбранных данных построим карту сегментирования (таблица 4).

Таблица 4 – Карта сегментирования рынка по разработке медицинских информационных систем

Критерии		Тип информационной системы		
		Система комплексной компьютеризации медицинских организаций и подразделений	Автоматизированная учетная медицинская информационная система	Экспертная система (система поддержки принятия решений)
Тип медицинского учреждения	Учреждения стационарного типа (диагностические центры, больницы, клиники)	ЕГИСЗ, Medesk	ПК «ЭПИКО» (разработка «АБ Систем»), ЕГИСЗ, Барс Здравоохранение	
	Амбулаторно-поликлинические учреждения	ЕГИСЗ, Medesk	ПК «ЭПИКО» (разработка «АБ Систем»), ЕГИСЗ, Барс Здравоохранение	
	Медицинские научно-исследовательские учреждения, НИИ	ЕГИСЗ	МИС qMS, Galenos («ТехЛАБ»)	Galenos («ТехЛАБ»)

В результате построения карты сегментирования выявлено, какие ниши на рынке услуг не заняты конкурентами или где уровень конкуренции низок. Такими нишами являются экспертные системы для учреждений стационарного типа и для амбулаторно-поликлинических учреждений.

Разнообразие и функциональные возможности существующих систем поддержки принятия решений очень малы для области решения проблем избыточного веса. Поэтому было решено разрабатывать экспертную систему, потребителями которой могут быть как научно-исследовательские институты, так и учреждения стационарного и поликлинического типа.

4.1.2 Анализ конкурентных технических решений

Необходимо систематически проводить анализ конкурирующих разработок, существующих на рынке. Это способствует своевременному усовершенствованию научного исследования и успешному противостоянию соперникам. Анализ конкурентных технических решений с позиции ресурсоэффективности и ресурсосбережения позволяет провести оценку эффективности разработки и определить направления для дальнейшего усовершенствования.

Для сравнительного анализа были выбраны существующие медицинские системы: «Galenos» и «МИС qMS». В последних реализациях системы «МИС qMS» были представлены возможности поддержки принятия врачебных решений, поэтому будет рассматривать ее как потенциальное конкурентное решение. Оценочная карта представлена в таблице 5.

Таблица 5 – Оценочная карта для сравнения конкурентных технических решений

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. Повышение производительности труда пользователя	0,15	5	5	3	0,75	0,75	0,45
2. Удобство в эксплуатации	0,1	4	4	4	0,40	0,40	0,40
3. Надежность	0,1	5	5	5	0,50	0,50	0,50
4. Функциональные возможности	0,15	5	5	4	0,75	0,75	0,60
5. Качество интеллектуального интерфейса	0,05	4	4	5	0,20	0,20	0,25
Экономические критерии оценки эффективности							
1. Конкурентоспособность продукта	0,1	5	5	3	0,50	0,50	0,30
2. Уровень востребованности среди потребителей	0,1	4	4	4	0,40	0,40	0,40
3. Цена	0,05	4	3	3	0,20	0,15	0,15
4. Послепродажное обслуживание	0,05	5	4	5	0,25	0,20	0,25
5. Финансирование научной разработки	0,05	4	3	3	0,20	0,15	0,15
Итого	**1**				**4,15**	**4,00**	**3,45**

Анализ конкурентных технических решений определяется по следующей формуле:

$$К = \sum В_i \cdot Б_i, \quad (4.1)$$

где К – конкурентоспособность научной разработки или конкурента;

Bi – вес показателя (в долях единицы);

Бi – балл i-го показателя.

Исходя из расчётов, можно сделать вывод, что наша разработка имеет достаточно высокий уровень конкурентоспособности. Разрабатываемая система имеет следующие преимущества: повышение производительности труда (за счёт чего будут уменьшаться врачебные ошибки), функциональные возможности и послепродажное обслуживание. Позиции конкурентов наиболее уязвимы в ценовом диапазоне и технических возможностях, что определяет конкурентное преимущество нашей разработки.

4.1.3 SWOT-анализ

В целях исследования внешней и внутренней среды объекта был проведен SWOT-анализ, который отражает сильные и слабые стороны, возможности и угрозы разрабатываемого продукта (таблица 6, таблица 7).

Таблица 6 – Матрица SWOT

	Сильные стороны: С1. Сокращение ошибок в назначениях врачей. С2. Отсутствие аналогичных систем для выбора траектории лечения детей с избыточным весом. С3. Отсутствие зависимости диагнозов от субъективного мнения, опыта и знаний врача. С4. Увеличение скорости принятия решений о тактике лечения.	**Слабые стороны:** Сл1. Ошибки в формировании базы знаний. Сл2. Неквалифицированные пользователи. Сл3. Значительные временные и интеллектуальные затраты на реализацию. Сл4. Срок выхода на рынок.

Возможности: В1. Расширение географии внедрения технологии. В2. Рост спроса, появление новых клиентов В3. Повышение квалификации сотрудников. В4. Осваивание новых отраслей.	За счет отсутствия аналогов системы внедрение технологии может выйти на большой уровень. Увеличение скорости и точности выбора лечения и отсутствие аналогов повышают спрос на разрабатываемую систему и внедрение ее в других отраслях.	Широкое внедрение технологии позволит увеличить квалификацию пользователей. Затраты на реализацию нивелируются увеличением числа клиентов системы и расширением числа освоенных отраслей. Квалификация сотрудников повышается по мере освоения системы.
Угрозы: У1. Сбои в работе системы. У2. Утечка персональных данных. У3. Увеличение конкуренции. У4. Отсутствие спроса на технологию разработки.	Сбои в работе системы могут повлиять на ошибки и скорость выбора лечения. Отсутствие аналогов снижает риск большой конкуренции на рынке, но может увеличить угрозу утечки персональных данных с целью разработки аналогичной системы. Увеличение скорости и точности выбора лечения и отсутствие аналогов повышают спрос на разрабатываемую систему.	Ошибки в формировании базы знаний могут произойти и по вине сбоев системы. Из-за отсутствия должной квалификации пользователей системы могут произойти сбои и утечка данных. Ошибки сформированной базы знаний могут оказать влияние на спрос и привести к увеличению конкуренции с подобными системами, базы знаний которых будут составлены правильно.

Таблица 7 – Интерактивная матрица проекта

		Сильные стороны проекта				Слабые стороны проекта			
		С1	С2	С3	С4	Сл1	Сл2	Сл3	Сл4
Возможности проекта	В1	-	+	-	-	-	+	0	0
	В2	-	+	+	+	-	-	+	0
	В3	0	-	0	0	-	+	-	-
	В4	+	+	+	+	-	0	+	-
Угрозы проекта	У1	+	-	-	+	+	+	-	-
	У2	-	+	-	-	-	+	-	-
	У3	0	+	0	0	+	-	0	0
	У4	+	+	+	+	+	-	-	-

В ходе проведенного анализа внутренней и внешней среды были выявлены проблемы, которые могут возникнуть при реализации проекта, согласно чему можно обозначить ряд рекомендаций:

- выбрать и провести тестирование инструментария для формирования базы знаний;
- проверить достоверность данных и обеспечить их надежную защиту;
- разработать систему обучения и поддержки пользователей;

– развивать подобную технологию на другие отрасли применения;
– расширять функционал и базу знаний, совершенствовать данные;
– производить постоянную проверку системы.

Таким образом, в результате SWOT-анализа были выявлены слабые и сильные стороны, а также возможные варианты повышения эффективности и минимизации угроз.

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 научных исследований.

Для выполнения научного исследования формируется рабочая группа, по каждому виду запланированных работ устанавливается соответствующая должность исполнителей. Перечень этапов и работ, распределение исполнителей по видам работ приведен в таблице 8.

Таблица 8 – Перечень этапов, работ и распределение исполнителей

Основные этапы	**№ работ**	**Содержание работ**	**Должность исполнителя**
Разработка технического задания	1	Выбор темы научного исследования	Студент, руководитель
	2	Составление и утверждения технического задания	Студент, руководитель
	3	Календарное планирование работ по теме исследования	Руководитель
Анализ предметной области	4	Подбор и изучение материалов по теме	Студент
	5	Описание предметной области	Студент
	6	Анализ исходных данных	Студент
Выбор методов реализации проекта	7	Поиск методик работы с данными	Студент, руководитель
	8	Выбор инструментария	Студент, руководитель
	9	Формирование рабочей выборки данных	Студент
	10	Обработка данных	Студент

Реализация проекта	11	Анализ данных с использованием статистических пакетов	Студент
	12	Сравнение результатов применения разных методов	Студент, руководитель
	13	Разработка алгоритма принятия решения	Студент, руководитель
Обобщение и оценка результатов	14	Оценка эффективности полученных результатов исследования	Студент, руководитель
Оформление отчета по исследованию	15	Составление пояснительной записки	Студент
	16	Подготовка презентационного материала	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 множества трудно учитываемых факторов.

Ожидаемое (среднее) значение трудоемкости $t_{ожi}$ вычисляется по формуле:

$$t_{\text{ож i}} = \frac{3t_{\min i}+2t_{\max i}}{5} \quad (4.2)$$

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

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

$t_{max\ i}$ – 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 $T_р$, учитывающая параллельность выполнения работ несколькими исполнителями.

$$T_{pi} = \frac{t_{\text{ож}\,i}}{\text{Ч}_i} \quad (4.3)$$

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

$t_{ожi}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.

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

Расчеты ожидаемой трудоемкости и продолжительности работ представлены в таблице 6.

4.2.3 Разработка графика проведения научного исследования

Для удобства построения графика, длительность каждого из этапов работ следует перевести из рабочих в календарные дни. Для этого необходимо воспользоваться следующей формулой:

$$T_{ki} = T_{pi} \times k_{кал} \quad (4.4)$$

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

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

$k_{кал}$– коэффициент календарности.

Коэффициент календарности определяется по следующей формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}} = \frac{365}{248} = 1{,}4718, \quad (4.5)$$

где $T_{кал}$ – календарные дни;

$T_{вых}$ – выходные дни;

$T_{пр}$ – праздничные дни.

Тогда длительность каждого из этапов работ в календарных днях будет равна: $T_{ki} = T_{pi} * 1{,}4718$.

Временные показатели проведения научного исследования представлены в таблице 9.

Таблица 9 – Временные показатели проведения научного исследования

№ работ	Трудоёмкость работ			Исполнители	Длительность работ в рабочих днях $T_{рi}$	Длительность работ в календарных днях $T_{кi}$
	t_{min}, чел-дни	t_{max}, чел-дни	$t_{ож\ i}$, чел-дни			
1	2	5	3,2	Студент, руководитель	1,6	2
2	5	7	5,8	Студент, руководитель	2,9	4
3	1	3	1,8	Руководитель	1,8	3
4	14	21	16,8	Студент	16,8	25
5	4	7	5,2	Студент	5,2	8
6	10	14	11,6	Студент	11,6	17
7	7	10	8,2	Студент, руководитель	4,1	6
8	3	5	3,8	Студент, руководитель	1,9	3
9	2	5	3,2	Студент	3,2	5
10	4	7	5,2	Студент	5,2	8
11	8	11	9,2	Студент	9,2	13
12	3	5	3,8	Студент, руководитель	1,9	3
13	7	10	8,2	Студент, руководитель	4,1	6
14	7	10	8,2	Студент, руководитель	4,1	6
15	20	32	24,8	Студент	24,8	36
16	2	4	2,8	Студент	2,8	4
Итого	Всего				101,2	149
	Руководитель				22,4	33
	Студент				99,4	146

На основании таблицы 9 строится календарный план-график. Работы выделены цветом в зависимости от исполнителей. Некоторые работы могут выполняться одновременно. План-график проведения научного исследования представлен на рисунке 44 в виде диаграммы Ганта.

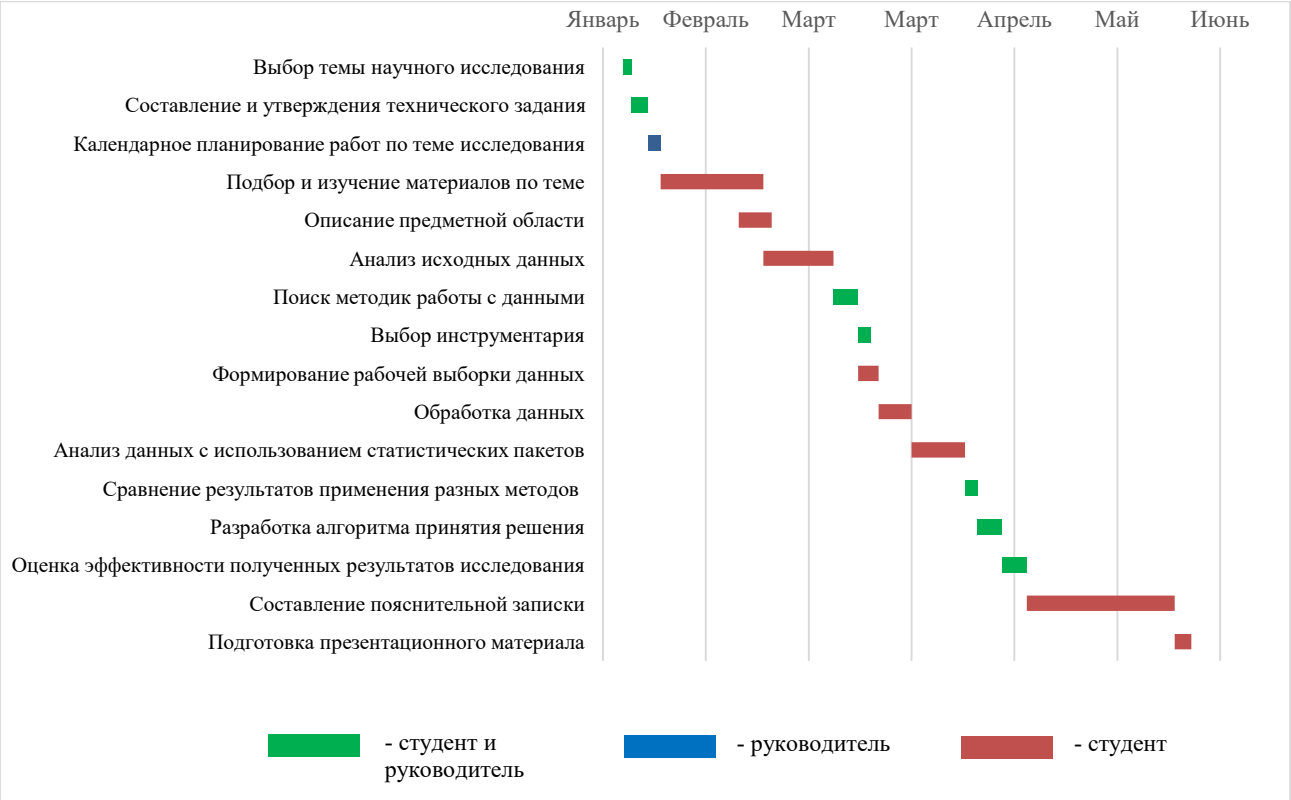

Рисунок 44 – Диаграмма Ганта

4.2.4 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 используются следующие статьи затрат:

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

4.2.4.1 Расчет материальных затрат

В эту статью включаются затраты на приобретение всех видов материалов, комплектующих изделий и полуфабрикатов, необходимых для выполнения работ по данной теме. Количество потребных материальных ценностей определяется по нормам расхода.

Расчет стоимости материальных затрат производится по действующим прейскурантам или договорным ценам. В стоимость материальных затрат включают транспортно-заготовительные расходы. В эту же статью включаются затраты на оформление документации (канцелярские принадлежности, тиражирование материалов).

Расчет материальных затрат осуществляется по следующей формуле:

$$C_{\text{м}} = (1 + k_T) \cdot \sum_{i=1}^{m} Ц_i \cdot N_{\text{расх}i} , \quad (4.6)$$

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

$N_{\text{расх}i}$ – количество материальных ресурсов i-го вида, планируемых к использованию при выполнении научного исследования (шт., кг, м, м^2 и т.д.);

$Ц_i$ – цена приобретения единицы i-го вида потребляемых материальных ресурсов (руб./шт., руб./кг, руб./м, руб./м^2 и т.д.);

k_T – коэффициент, учитывающий транспортно-заготовительные расходы.

Материальные затраты, необходимые для данной разработки, заносятся в таблицу 10.

Таблица 10 – Материальные затраты

Наименование	**Единица измерения**	**Кол-во**	**Цена за единицу, руб.**	**Сумма, руб.**
Бумага для принтера А4	уп	1	310	310
Картридж для принтера	шт	1	750	750
Папка для брошюровки	шт	1	140	140
Ручка шариковая	шт	2	25	50
Блокнот	шт	2	100	200
			Всего за материалы	1450
			Транспортно-заготовительные расходы (3-5%)	43,5
			Итого затраты на материалы ($C_{\text{м}}$)	**1493,5**

4.2.4.2 Основная заработная плата исполнителей

В настоящую статью включается основная заработная плата работников, непосредственно участвующих в выполнении работ по данной теме (включая премии, доплаты) и дополнительная заработная плата:

$$C_{\text{зп}} = З_{\text{осн}} + З_{\text{доп}} , \quad (4.7)$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата.

Основная заработная плата ($З_{осн}$) руководителя (лаборанта, инженера) от предприятия (при наличии руководителя от предприятия) рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot T_p, \quad (4.8)$$

где $З_{осн}$ – основная заработная плата одного работника;

$Т_р$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн. (из таблицы 6);

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м \cdot М}{F_д}, \quad (4.9)$$

где $З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года:

- при отпуске в 24 раб. дня М =11,2 месяца, 5-дневная неделя;
- при отпуске в 48 раб. дней М=10,4 месяца, 6-дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн. (таблица 11).

Таблица 11 – Баланс рабочего времени

Показатели рабочего времени	**Руководитель**	**Студент**
Календарное число дней	365	365
Количество нерабочих дней - выходные дни - праздничные дни	117	117
Потери рабочего времени - отпуск - невыходы по болезни	24	24
Действительный годовой фонд рабочего времени	224	224

Месячный должностной оклад работника:

$$З_м = З_б \cdot (1 + k_{пр} + k_д) \cdot k_р, \quad (4.10)$$

где $З_б$ – базовый оклад, руб.;

$k_{пр}$ – премиальный коэффициент (30% от $З_{тс}$);

$k_д$ – коэффициент доплат и надбавок (в НИИ и на промышленных предприятиях – за расширение сфер обслуживания, за профессиональное мастерство, за вредные условия: 15-20% от $З_{тс}$);

$k_р$ – районный коэффициент (1,3 для Томска).

Оклад студента равен 1906 руб., а оклад руководителя проекта (доцент, к.т.н.) составляет 26300 руб.

Расчёт основной заработной платы приведён в таблице 12.

Таблица 12 – Расчет основной заработной платы

Исполнители	**$З_б$, руб.**	**$k_{пр}$**	**$k_д$**	**$k_р$**	**$З_м$, руб**	**$З_{дн}$, руб.**	**$Т_р$, раб. дн.**	**$З_{осн}$, руб.**
Руководитель	26300	0,3	0,15	1,3	49575,5	2478,8	22,4	55525,12
Студент	1906			1,3	2477,8	123,89	99,4	12314,67
							Итого $З_{осн}$	67839,79

4.2.4.3 Дополнительная заработная плата исполнителей

Затраты по дополнительной заработной плате исполнителей темы учитывают величину предусмотренных Трудовым кодексом РФ доплат за отклонение от нормальных условий труда, а также выплат, связанных с обеспечением гарантий и компенсаций.

Расчет дополнительной заработной платы ведется по следующей формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}, \quad (4.11)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (10-15% от основной заработной платы). Примем коэффициент равный 0,12.

Расчёт дополнительной заработной платы приведён в таблице 13.

Таблица 13 – Расчет дополнительной заработной платы

Исполнитель	Основная заработная плата, руб.	$k_{доп}$	Дополнительная заработная плата, руб.
Руководитель	55525,12	0,12	6663
Студент	12314,67	-	-
		Итого $З_{доп}$	**6663**

4.2.4.4 Отчисления во внебюджетные фонды (страховые отчисления)

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}), \quad (4.12)$$

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

С 2016 г., в соответствии с Федеральным законом от 24.07.2009 №212-ФЗ, установлен размер страховых взносов равный 30%. На основании пункта 1 ст.58 закона №212-ФЗ для учреждений, осуществляющих образовательную и научную деятельность, действует пониженная ставка – 27,1%.

Отчисления во внебюджетные фонды представлены в таблице 14.

Таблица 14 – Отчисления во внебюджетные фонды

Исполнитель	Основная заработная плата, руб.	Дополнительная заработная плата, руб.	$k_{внеб}$	$З_{внеб}$
Руководитель	55525,12	6663	0,271	16852,98
Бакалавр	12314,67	-	-	-
			Итого $З_{внеб}$	**16852,98**

4.2.4.5 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 следующей формуле:

$$З_{накл} = (\text{сумма статей}) \cdot k_{нр}, \quad (4.13)$$

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

Величину коэффициента накладных расходов можно взять в размере 16%. Таким образом, накладные расходы по проекту составили:

$З_{накл}$ = (1493,5+67839,79+6663+16852,98) * 0,16 = 92849,27 * 0,16 = 14855,88 руб.

4.2.4.6 Формирование бюджета затрат научно-исследовательского проекта

Определение бюджета затрат на научно-исследовательский проект приведен в таблице 15.

Таблица 15 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.	Примечание
1. Материальные затраты НТИ	1493,5	Пункт 4.3.4.1
2. Затраты по основной заработной плате исполнителей	67839,79	Пункт 4.3.4.2
3. Затраты по дополнительной заработной плате исполнителей	6663	Пункт 4.3.4.3
4. Отчисления во внебюджетные фонды	16852,98	Пункт 4.3.4.4
5. Накладные расходы	14855,88	16 % от суммы ст. 1-4
6. Бюджет затрат НТИ	**107705,15**	Сумма ст. 1- 5

4.3 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 как:

$$I_{ф}^{p} = \frac{Ф_{p}}{Ф_{\max}}, \qquad (4.14)$$

где $Ф_р$ – стоимость исполнения;

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

$$I_{ф}^{p} = \frac{107705,15}{150000} = 0,72$$

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

Интегральный показатель ресурсоэффективности исполнения объекта исследования можно определить следующим образом:

$$I_p = \sum a \cdot b, \quad (4.15)$$

где a – весовой коэффициент параметра;

b – бальная оценка параметра для аналога и разработки, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

Расчет интегрального показателя ресурсоэффективности приведен в таблице 16.

Таблица 16 – Расчет интегрального показателя ресурсоэффективности

Критерии / **Объект исследования**	**Весовой коэффициент параметра**	**Оценка выполнения**
1. Способствует росту производительности труда пользователя	0,25	4
2. Удобство в эксплуатации (соответствует требованиям потребителей)	0,2	5
3. Функциональные возможности	0,2	4
4. Надежность	0,2	4
5. Экономия времени	0,15	5
Итого	1	

$I_р = 0,25*4 + 0,2*5 + 0,2*4 + 0,2*4 + 0,15*5 = 4,35$

Интегральный показатель эффективности исполнения разработки ($I_{исп.}$) определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{исп} = \frac{I_p}{I_ф} = \frac{4,35}{0,72} = 6,04 \quad (4.16)$$

Полученное значение интегрального показателя эффективности исполнения разработки превышает максимальный балл системы оценивания. Результат работы можно считать положительным, так как оценка интегрального показателя ресурсоэффективности близка к максимальной, при этом стоимость разработки ниже, чем у ряда аналогов, рассмотренных при анализе конкурентных решений.

Таким образом, полученные при анализе конкурентных решений данные позволяют сделать вывод, что разработка является привлекательной для

инвесторов. Продукт имеет много преимуществ перед рассмотренными конкурентами, в особенности по таким критериям, как повышение производительности, функциональные возможности и цена. SWOT-анализ позволил выявить слабые и сильные стороны, возможные перспективы и угрозы, а также предложены рекомендации по минимизации их влияния.

Также была построена структура работ проекта и определены ответственные должности для их выполнения. В соответствии с назначенными работами была рассчитана их трудоемкость и составлен план-график работ в виде диаграммы Ганта. Общая длительность проектирования и разработки программного продукта составила 149 дней. Общий бюджет НТИ составил 107 705,15рублей. Бюджет включает в себя затраты на основную и дополнительную заработную плату работников, материальные затраты, отчисления во внебюджетные фонды и накладные расходы.

Глава 5 Социальная ответственность

Выпускная квалификационная работа направлена на разработку системы поддержки принятия решений о выборе траектории лечения для детей с избыточным весом. Система предназначена для усовершенствования процесса определения диагноза и назначения подходящего лечения. Работа с системой заключается во взаимодействии потенциальных пользователей (медицинских работников) с компьютером.

Разработка технологии предполагает работу с информацией, проводимую за персональным компьютером в учебной аудитории Кибернетического центра Национального исследовательского Томского политехнического университета. В помещении для выполнения исследования было использовано одно оборудованное компьютерном место. Общая площадь аудитории составляет 25м^2, объем – 87 м^3, ширина – 4,5 м, длина – 5,5 м, высота – 3,5 м.

Данный раздел предназначен для разработки комплекса мер технического, организационного и правового характера, уменьшающих негативные последствия разработки информационной системы.

5.1 Правовые и организационные вопросы обеспечения безопасности

5.1.1 Специальные правовые нормы трудового законодательства

Регулирование отношений в сфере охраны труда в Томской области осуществляется Трудовым кодексом Российской Федерации, федеральными законами и иными нормативными правовыми актами Российской Федерации, содержащими нормы трудового права.

При работе с персональным компьютером очень важную роль играет соблюдение правильного режима труда и отдыха. В таблице 17 представлены сведения о регламентированных перерывах, которые необходимо делать при работе на компьютере, в зависимости от продолжительности рабочей смены, видов и категорий трудовой деятельности с ПЭВМ [41]. При работе над проектом уровень нагрузки соответствовал III категории, группа В.

Таблица 17 – Регламентированные перерывы при работе на компьютере

Категория работы с ПЭВМ	Уровень нагрузки за рабочую смену при видах работ с ПЭВМ			Суммарное время регламентированных перерывов, мин	
	Группа А, кол-во знаков	Группа Б, кол-во знаков	Группа В, ч	при 8-часовой смене	при 12-часовой смене
I	до 20000	до 15000	до 2	50	80
II	до 40000	до 30000	до 4	70	110
III	до 60000	до 40000	до 6	90	140

При несоответствии фактических условий труда требованиям Санитарных правил и норм время регламентированных перерывов следует увеличить на 30%. Эффективность перерывов повышается при сочетании с производственной гимнастикой или организации специального помещения для отдыха персонала с удобной мягкой мебелью, аквариумом, зеленой зоной и т.п.

Каждый студент, руководитель или сотрудник должен быть ознакомлен с техникой безопасности, инструкцией по охране труда. После ознакомления он должен расписаться в специальном журнале.

Работа за компьютером и разработка программного обеспечения относятся к категории не тяжелых работ, не предусматривает ненормированный рабочий график. Поэтому в данной работе не предусмотрено применение режима сокращённого рабочего дня, запрещение использования труда женщин и подростков, компенсаций за вредный труд, ночные смены [40].

5.1.2 Организационные мероприятия при компоновке рабочей зоны

При организации рабочего места необходимо учитывать требования безопасности, промышленной санитарии, эргономики, технической эстетики. Невыполнение этих требований может привести к получению работником производственной травмы или развитию у него профессионального заболевания. Согласно требованиям [41] при организации работы на ПЭВМ должны выполняться следующие условия:

- персональный компьютер (ПК) и рабочее место должно располагаться так, чтобы свет падал сбоку, лучше слева;
- расстояние от ПК до стен должно быть не менее 1 м, следует

избегать расположения рабочего места в углах либо лицом к стене;

- ПК лучше установить так, чтобы можно было увидеть удаленный предмет в помещении или на улице;
- при наличии нескольких компьютеров расстояние между экраном одного монитора и задней стенкой другого должно быть не менее 2 м, а расстояние между боковыми стенками соседних мониторов – не менее 1,2 м;
- окна в помещениях с ПЭВМ должны быть оборудованы регулируемыми устройствами (жалюзи, занавески, внешние козырьки и т.д.);
- монитор, клавиатура и корпус компьютера должны находиться прямо перед оператором; высота рабочего стола с клавиатурой должна составлять 680 – 800 мм над уровнем стола; а высота экрана (над полом) –900–1280 см;
- монитор должен находиться от оператора на расстоянии 60-70 см на 20 градусов ниже уровня глаз;
- пространство для ног должно быть: высотой не менее 600 мм, шириной не менее 500 мм, глубиной не менее 450 мм. Должна быть предусмотрена подставка для ног работающего шириной не менее 300 мм с регулировкой угла наклона до 20 градусов;
- рабочее кресло должно иметь мягкое сиденье и спинку, с регулировкой сиденья по высоте, с удобной опорой для поясницы;
- положение тела пользователя относительно монитора должно соответствовать направлению просмотра под углом 90 или 75 градусов.

Правильная поза и положение рук оператора являются очень важными для исключения нарушений в опорно-двигательном аппарате и возникновения синдрома постоянных нагрузок.

5.2 Профессиональная социальная безопасность.

Специфика труда и режим работы аналитика и разработчика характеризируются большими зрительными нагрузками одновременно с малой двигательной активностью, монотонностью выполняемых операций,

вынужденной рабочей позой. Эти факторы отрицательно сказываются на самочувствии работающего человека. В связи с этим к концу рабочего дня пользователи ПК ощущают головную боль, боли в глазах, мышцах шеи, рук, спины, зуд кожи лица и т.д. Постоянные недомогания приводят к мигреням, частичной потере зрения, сколиозу, кожным воспалениям и другим нежелательным явлениям.

Существует два вида производственных факторов: опасные, которые вызывают травмы, и вредные, вызывающие заболевания. По оказываемому влиянию на человека эти факторы делятся на 4 группы: физические; химические; биологические и психофизиологические [42].

Поскольку на состояние здоровья программиста химические и биологические факторы существенного влияния не оказывают, то рассматриваться будут две группы факторов: физические и психофизические.

К физически опасным факторам можно отнести опасность поражения электрическим током. К группе вредных факторов при работе с компьютером следует отнести: показатели микроклимата (влажность воздуха, температура); освещенность; уровень шума; уровень электромагнитных излучений.

К психофизиологическим факторам, воздействующим на разработчика, относятся: умственное перенапряжение; монотонность труда; эмоциональные нагрузки.

Для представления всех выявленных вредных и опасных факторов классифицируем их в соответствии с нормативными документами (таблица 18).

Таблица 18 – Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Изготовление	Эксплуатация	
1. Повышенное значение напряжения в электрической цепи	+	+	+	ПУЭ - Правила устройства электроустановок. Изд. 7. [43] СанПиН 2.2.2/2.4.1340–03 [44]
2. Отклонение показателей микроклимата	+	+	+	СанПиН 2.2.4.548-96 [45]
3. Недостаточная освещенность рабочей зоны	+	+		СанПиН 2.2.1/2.1.1.1278-03 [46]

4. Превышение уровня шума	+	+	+	ГОСТ 12.1.003-2014 [47] СанПиН 2.2.4.3359-16 [48]
5. Повышенный уровень электромагнитных излучений		+	+	ГОСТ 12.1.006–84 [49] СанПиН 2.2.4.3359-16 [48]
6. Нервно-психические перегрузки	+	+		Р 2.2.2006-05 [50]

Рассмотрим более подробно опасные и вредные факторы и мероприятия по снижению их воздействия.

5.2.1 Повышенное значение напряжения в электрической цепи

Поражение человека электрическим током возможно при замыкании электрической цепи через тело человека, т. е. при прикосновении человека к сети не менее чем в двух точках. При этом повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека, является опасным фактором.

Работа велась в помещении без повышенной опасности, т.е. отсутствовали такие условия как: повышенная влажность (относительная влажность воздуха длительно превышает 75%), высокая температура (более 35°С), токопроводящая пыль, токопроводящие полы, возможность одновременного соприкосновения к имеющим соединения с землей металлическим элементам и металлическим корпусам электрооборудования [43, 44].

Электрические установки, к которым относится ПК, представляют для человека большую потенциальную опасность, т.к. в процессе эксплуатации или проведения профилактических работ человек может коснуться комплектующих компьютера, находящихся под напряжением. Корпуса оборудования, оказавшегося под напряжением в результате повреждения или пробоя изоляции, не подают каких-либо сигналов об опасности.

Причинами электропоражений являются: провода с поврежденной изоляцией, розетки сети без предохранительных кожухов (при использовании приборов с европейскими вилками).

Основные технические средства защиты от поражения электрическим током:

– изоляция токопроводящих частей (проводов) и ее непрерывный контроль;

– установка оградительных устройств;

– предупредительная сигнализация и блокировки;

– использование знаков безопасности и предупреждающих плакатов; применение малых напряжений;

– защитное заземление;

– зануление;

– защитное отключение.

Согласно нормативам [43], питание устройства в помещении, в котором выполнялась работа, осуществляется от силового щита через автоматический предохранитель, который срабатывает при коротком замыкании нагрузки. Для снижения величин возникающих разрядов применяются покрытия из антистатического материала.

К организационно-техническим мероприятиям относится первичный инструктаж по технике безопасности, который является обязательным условием для допуска к работе в данном помещении.

В кабинете Кибернетического центра используются приборы, потребляющие напряжение 220В переменного тока с частотой 50Гц. Это напряжение опасно для жизни, поэтому обязательны следующие меры предосторожности:

1) перед началом работы необходимо убедиться, что выключатели и розетка закреплены и не имеют оголённых токоведущих частей;

2) при обнаружении неисправности оборудования и приборов, необходимо не делая никаких самостоятельных исправлений сообщить ответственному за оборудование;

3) запрещается загромождать рабочее место лишними предметами.

При возникновении несчастного случая следует немедленно освободить пострадавшего от действия электрического тока и, вызвав врача, оказать ему необходимую помощь.

5.2.2 Отклонение показателей микроклимата

Микроклимат производственных помещений характеризуется следующими параметрами: температурой, относительной влажностью, скоростью движения воздуха. Все эти параметры влияют на организм человека как сами по себе, так и в комплексе.

Энергозатраты организма измеряются в ккал/ч (Вт). По затраченной энергии работы разделяются на категории. Работа программиста относится к категории Iа – интенсивность энергозатрат до 120 ккал/ч (до 139 Вт) [45]. Работы производятся в основном сидя и сопровождаются незначительным физическим напряжением. Параметры микроклимата на рабочем месте для категории Iа приведены в таблице 19:

Таблица 19 – Оптимальные и допустимые величины показателей воздушной среды на рабочих местах производственных помещений

Сезон года	Температура воздуха, °С			Температура поверхностей, °С		Относительная влажность, %		Скорость движения воздуха, м/с		
	Оптим. значение	Допуст. значение		Оптим. значение	Допуст. значение	Оптим. значение	Допуст. значение	Оптим. значение	Допуст. значение	
		Ниже оптим.	Выше оптим.						Ниже оптим.	Выше оптим.
Холодный	22-24	20-21,9	24,1-25	21-25	19-26	60-40	15-75	0,1	0,1	0,1
Теплый	23-25	21-22,9	25,1-28	22-26	20-29	60-40	15-75	0,1	0,1	0,2

Параметры микроклимата помещения, регулирующийся системой центрального отопления, а также приточно-вытяжной вентиляцией, имеют следующие средние значения: влажность 40%, скорость движения воздуха 0,1 м/с, температура летом 23-25°С, зимой 20-23°С, что соответствует нормам.

К мероприятиям по оздоровлению воздушной среды в производственном помещении относится: правильная организация вентиляции и кондиционирования воздуха, отопление помещений. Вентиляция должна осуществляться как естественным, так и механическим путём. В рабочем

помещении необходима подача следующего объёма наружного воздуха: при объёме помещения до 20м^3 на человека – не менее 30м^3 в час на человека; при объёме помещения более 40м^3 на человека и отсутствии выделения вредных веществ допускается естественная вентиляция. В кабинете принудительная вентиляция отсутствует, однако имеется естественная посредством поступления и удаления воздуха через окна, двери, щели. Объём воздуха на одного человека в аудитории КЦ составляет 22 м^3, следовательно, необходимо наличие принудительной вентиляции.

В зимнее время система отопления обеспечивает достаточное, постоянное и равномерное нагревание воздуха. В помещениях с повышенными требованиями к чистоте воздуха должно использоваться водяное отопление. В аудиториях используется водяное отопление со встроенными нагревательными элементами и стояками.

5.2.3 Недостаточная освещенность рабочей зоны

Для обеспечения нормативных условий работы в помещениях необходимо провести оценку освещенности рабочей зоны в соответствии с СанПиН 2.2.1/2.1.1.1278-03[46].

Недостаточная освещенность снижает зрительную работоспособность, влияет на психику человека, эмоциональное состояние, может вызывать усталость центральной нервной системы, которая возникает в результате приложения дополнительных усилий для опознания четких или сомнительных сигналов. Для оптимизации условий труда большую роль играет освещение рабочих мест. Организация освещённости рабочих мест должна выполнять два требования: обеспечивать различимость рассматриваемых предметов и уменьшать напряжение и утомляемость органов зрения. Производственное освещение должно быть устойчивым и равномерным, иметь правильное направление, исключать слепящее действие и образование резких теней.

Основным качественным показателям световой среды является коэффициент пульсации освещенности (Кп). Для рабочих мест с ПЭВМ этот показатель не должен превышать 5%. Оптимальная яркость экрана дисплея

составляет 75–100 кд/м2. При такой яркости экрана, а также яркости поверхности стола в пределах от 100 до 150 кд/м2 обеспечивается работоспособность зрительного аппарата на уровне 80–90 % и сохраняется постоянный размер зрачка на допустимом уровне 3–4 мм. Местное освещение не должно создавать блики на поверхности экрана и не должно увеличивать освещенность экрана ПЭВМ более чем 300 лк.

Работа проводится в кабинете, где используется смешанное освещение, помещение освещается 6 светильниками, в каждом из которых установлено 4 люминесцентных лампы типа ЛБ-18 (Osram L18W/765). Светильники расположены равномерно по всей площади потолка в ряд, создавая при этом равномерное освещение рабочих мест.

В помещении два оконных проема. Коэффициент естественной освещенности при совмещенном освещении и боковом естественном освещении для данного типа помещений составляет 0,7. Уровень искусственного освещения должен быть не менее 300 лк. Нормируемые параметры систем естественного и искусственного освещения на рабочих местах приведены таблице 20.

Таблица 20 – Параметры систем естественного и искусственного освещения

Рабочее место	Рабочая повер-ть, м (Г-горизонт., В –вертик.)	Коэффициент естественной освещенности, КЕО, %		Коэффициент совмещ. освещения, КЕО, %		Искусственное освещение				
		при верхнем или комб. освещении	при боковом освещении	при верхнем или комб. освещении	при боковом освещении	Освещенность, лк			Показатель дискомфорта, М, не более	Коэфф. пульсации освещ-ти, %, не более
						при комб. освещ.		при общем освещ.		
						всего	от общ.			
1	2	3	4	5	6	7	8	9	10	11
Кабинет информатики и ВТ	Г – 0, 8 Экран дисплея: В – 1	3,5 -	1,2 -	2,1 -	0,7 -	500 -	300 -	400 200	15 -	10 -

Для того чтобы 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.

5.2.4 Превышение уровня шума

Источниками шума могут быть: работающее оборудование, установки кондиционирования воздуха, преобразователи напряжения, работающие осветительные приборы, внешние факторы. Шум ухудшает условия труда, оказывает вредное воздействие на организм человека. Он может затруднять разборчивость речи, вызывать снижение работоспособности, повышать утомляемость, вызывать необратимые изменения в органах слуха человека, ослаблять внимание, ухудшать память, реакцию, увеличивать число ошибок.

Источником шума на предприятиях, эксплуатирующих вычислительную технику, являются сами вычислительные машины (встроенные в стойки ПЭВМ вентиляторы, принтеры и т.д.), центральная система вентиляции и кондиционирования воздуха и другое оборудование [47]. Производственные помещения, в которых для работы используются ПЭВМ, не должны находиться по соседству с помещениями, в которых уровень шума и вибрации превышают нормируемые значения.

Во время выполнения ВКР основными источниками шума являлись компьютер, клавиатура, принтер. При работе за компьютером нормированный эквивалентный уровень звука не должен превышать 80 дБА [48]. При регулярных проверках оборудования можно избежать превышения допустимого уровня шума. По субъективным ощущениям шумовая обстановка на рабочем месте соответствовала норме, за исключением времени проведения каких-либо ремонтных работ.

5.2.5 Повышенный уровень электромагнитных излучений

Электрические приборы производят электромагнитное излучение, воздействие которого на организм человека зависит от напряжённостей

электрического, магнитного поля, потока энергии, частоты колебаний, а также от размера облучаемого тела.

При воздействии электромагнитных полей низкой напряжённости нарушения в организме человека носят обратимый характер. При превышении предельно допустимого уровня страдает нервная система, сердечнососудистая системы, органы пищеварения, ухудшаются некоторые биологические показатели крови.

Большая часть электромагнитных излучений происходит не от экрана монитора, а от видеокабеля и системного блока. В портативных компьютерах практически всё электромагнитное излучение идет от системного блока.

На расстоянии 50см вокруг видеотерминала напряженность электромагнитного поля должна быть не более 25 В/м, если частота находится в диапазоне 5 Гц ÷ 2 кГц, и не более 2,5 В/м, если частота находится в диапазоне 2 кГц ÷ 400кГц. Плотность магнитного потока не должна превышать 250 нТл, если частота находится в диапазоне 5 Гц ÷ 2 кГц и 25 нТл, если частота находится в диапазоне 2 кГц ÷ 400кГц [49].

Основные методы по уменьшению воздействия электромагнитного излучения: увеличение расстояния от источника (экран видеомонитора не должен находится ближе 50 см от пользователя); использование приэкранного фильтра, специального экрана, а также других средств индивидуальной защиты, которые прошли испытание в аккредитованных лабораториях имеют соответствующий гигиенический сертификат.

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

5.2.6 Нервно-психические перегрузки

Во время длительной работы за компьютером человек может подвергаться воздействию таких факторов как эмоциональные перегрузки, умственное перенапряжение, монотонность труда и другие.

Эмоциональные перегрузки вызывают изменения функционального состояния центральной нервной системы. Они могут быть вызваны необходимостью выполнения большого объема работы, конфликтными или стрессовыми ситуациями.

Умственное перенапряжение может наступать при отсутствии времени на отдых после длительной работы, нарушения режима сна или питания. Оно может накапливаться и приводить к возникновению заболеваний [51].

Монотонная работа характеризуется однообразием рабочих действий и их многократным повторением. В результате теряется интерес к работе, возникает состояние производственной скуки, ухудшаются экономические показатели, повышается аварийность, травматизм, растет текучесть кадров [52].

Вредным фактором также может служить фиксированная рабочая поза. Она вызывает нарушение кровообращения в нижних конечностях и органах тазовой области, что может приводить к профессиональным заболеваниям (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 оборудование должно иметь регулировки: высота и наклон спинки стула, наклон монитора под индивидуальные особенности человека.

5.3 Экологическая безопасность

Охрана окружающей среды при разработке и предполагаемом использовании системы заключается в устранении отходов жизнедеятельности человека и бытового мусора. Основные виды загрязнения литосферы – твердые бытовые и промышленные отходы. В ходе выполнения ВКР образовывались такие твердые отходы как: бумага, батарейки, лампочки, использованные картриджи, отходы от продуктов питания и личной гигиены, отходы от канцелярских принадлежностей и т.д. Защита почвенного покрова и недр от

твердых отходов реализуется за счет сбора, сортирования и утилизации отходов и их организованного захоронения.

Компьютеры и оргтехнику в случае выхода из строя или морального износа подлежат технической экспертизе. Если оборудование не подлежит ремонту, то оно признается неработоспособным и рекомендуется к списанию (замене); в случае естественного отказа оборудования и нецелесообразности его ремонта и модернизации даются рекомендации о необходимости его списания и утилизации [53]. В том случае, если оргтехнику просто выбрасывают в мусорные баки, на старое оборудование могут воздействовать атмосферные осадки, что может привести к возникновению различных химических реакций с выделением вредных веществ.

На рабочем месте разработчика используются люминесцентные лампы, которые помимо стекла и алюминия содержат ртуть. Поэтому вышедшие из строя 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. Все ртутьсодержащие отходы и приборы, содержащие ртуть, подлежат сбору и возврату для последующей регенерации ртути в специализированных организациях [54].

Пластмассовые, железные детали, использованная и ненужная бумага подвергаются переработке. Рядом со зданием предусмотрены контейнеры для отходов, вывоз мусора осуществляется ежедневно сторонней организацией.

5.4 Безопасность в чрезвычайных ситуациях

К наиболее вероятным чрезвычайным ситуациям (ЧС) можно отнести следующие: пожар (взрыв) в здании, авария на коммунальных системах жизнеобеспечения, землетрясение. 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.

Мероприятия по обеспечению пожарной безопасности:

- проведение противопожарного инструктажа;
- обучение правилам техники безопасности;
- проверка проводки;
- наличие в помещении планов эвакуации;
- соблюдение эксплуатационных норм оборудования;
- наличие средств пожаротушения в помещении [55].

Помещение, в котором выполняется работа по написанию выпускной работы, относится к категории В по взрывопожарной и пожарной опасности [56]. В помещении и на этаже присутствуют средства оповещения: световая индикация направления движения к выходу в коридорах этажа; звуковая индикация; пассивные датчики задымленности. В аудитории имеется углекислотный огнетушитель типа ОУ-2.

В случае угрозы возникновения ЧС необходимо отключить электропитание, вызвать по телефону пожарную команду, эвакуировать людей из помещения согласно плану эвакуации (рисунок 45). При наличии небольшого очага пламени можно воспользоваться подручными средствами с целью прекращения доступа воздуха к объекту возгорания.

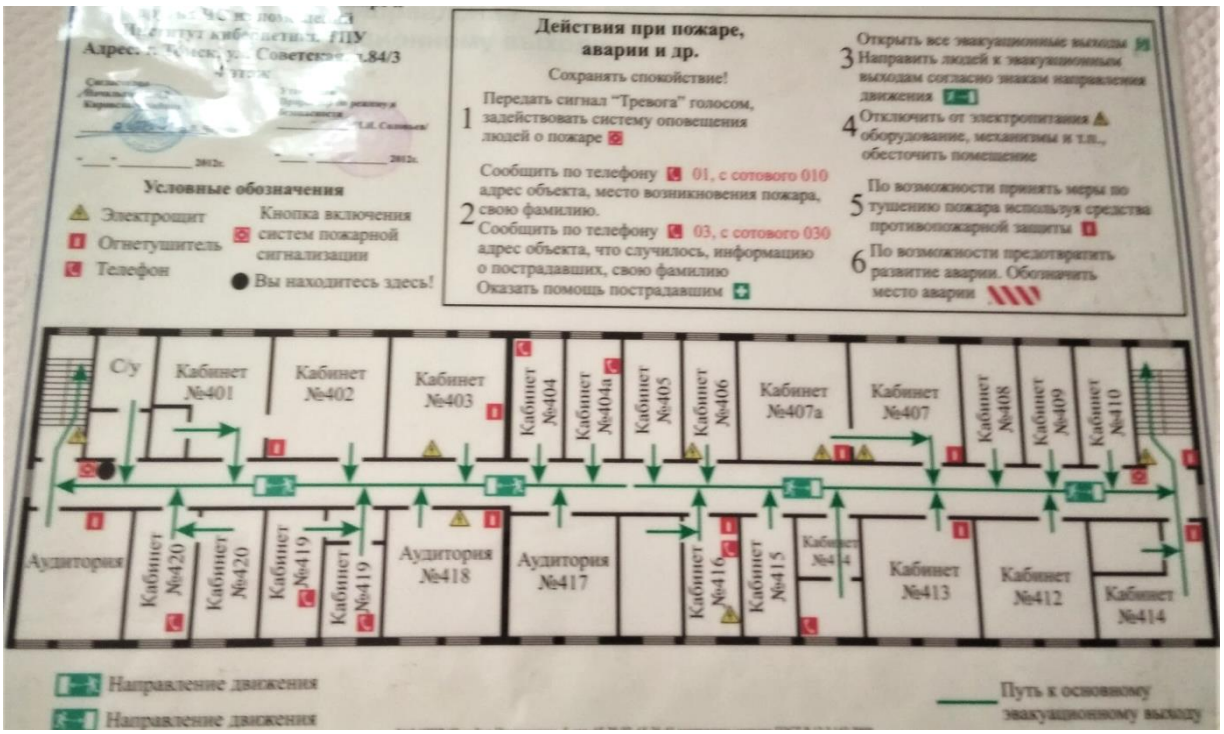

Рисунок 45 – План эвакуации при пожаре и других ЧС из помещений Института кибернетики ТПУ

Выводы по разделу

В данном разделе были рассмотрены и описаны правовые и организационные вопросы обеспечения безопасности, аспекты производственной и экологической безопасности, меры безопасности в чрезвычайных ситуациях, изложены требования к рабочему месту сотрудника. На основании изученной литературы, были указаны оптимальные параметры микроклимата для комфортной работы. Соблюдение условий, определяющих оптимальную организацию рабочего места научного сотрудника, позволит сохранить хорошую работоспособность в течение всего рабочего дня и будет способствовать продуктивной разработке системы.

Заключение

В результате выполнения выпускной квалификационной работы была разработана алгоритмическая база системы поддержки принятия решений при выборе траектории лечения детей с избыточным весом на основе анализа клинико-лабораторных показателей.

В рамках выполнения ВКР были решены следующие задачи:

- проведен обзор литературы по тематике исследования и изучены практики решения задач анализа многомерных данных;
- рассмотрены методы решения задач анализа многомерных данных и выбран инструментарий решения поставленных в ВКР задач;
- проведена первичная обработка и сокращение размерности данных;
- построена модель принятия решений на основе деревьев решений и проведена оценка качества модели;
- сформирован алгоритм обработки данных для принятия решения о выборе группы лечения детей с избыточным весом.

На данном этапе работы точность модели принятия решения составляет 82%. Это говорит о том, что исследования в данном направлении могут и должны быть продолжены для достижения наилучшего результата и разработки предельно точной системы поддержки принятия врачебных решений. Разнообразие методов и инструментов работы с многомерными данными позволяют проводить исследования с разных точек зрения, что позволит улучшить качество прогнозируемых решений.

В ходе написания выпускной работы в разделе «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» была проанализирована экономическая эффективность исследования и рассчитаны затраты в ходе его выполнения. В разделе «Социальная ответственность» рассмотрены параметры организации рабочего места, проанализированы вредные и опасные факторы труда и разработаны меры необходимые меры профилактики.

Список используемых источников

1. Атьков О.Ю., Кудряшов Ю.Ю., Прохоров А.А., Касимов О.В. Система поддержки принятия врачебных решений // Врач и информационные технологии. – 2013. – №6. – С. 67-75

2. Кудряшов Ю.Ю., Атьков О.Ю., Касимов О.В. Телемедицинская профилактика, реабилитация и управление здоровьем: проблемы и решения // Врач и информационные технологии. – 2016. – №2. – С.73-80.

3. Зарипова Г.Р., Богданова Ю.А., др. Современные модели систем поддержки принятия врачебных решений в хирургической практике. Состояние проблемы // Медицинский вестник Башкортостана. – 2016. – С. 96-101.

4. Богданова Ю.А., Зарипова Г.Р., др. Современные модели экспертных медицинских систем в прогнозировании операционного риска при наиболее распространенных интраабдоминальных вмешательствах // Медицинский альманах. – 2017. – С. 9-12.

5. Киселев К.В., Ноева Е.А., др. Разработка архитектуры базы знаний системы поддержки принятия врачебных решений, основанной на графовой базе данных // Медицинские технологии. Оценка и выбор. – 2018. – С.42-48.

6. Спирячин А.А., Бурковский В.Л. Интеллектуальная система принятия решений в условиях выбора тактики лечения хронических заболеваний на основе облачных технологий // Вестник Воронежского государственного технического университета. – 2014. –№5-1.

7. Раводин Р.А. Интеллектуальная система поддержки принятия врачебных решений в дерматовенерологии // Проблемы медицинской микологии. – 2014. – №3. – С.59-65.

8. Купеева И.А., Разнатовский К.И., др. Оценка эффективности интеллектуальной системы поддержки принятия врачебных решений // Вестник СПбГУ. – 2016. – сер.10, вып.2. – С. 62-68.

9. Кнышов Г.В., Руденко А.В., др. Особенности проектирования медицинской информационной системы поддержки принятия решений,

основанной на интеллектуальном анализе данных // Кибернетика и вычислительная техника. – 2014. – Вып. 177. – С. 79-87.

10. Кобринский Б.А. Системы поддержки принятия решений в здравоохранении и обучении // Врач и информационные технологии. – 2010. – №2. – С. 39-44.

11. Фролов С.В., Куликов А.Ю. , Остапенко О.А., Стрыгина Е.В. Системы поддержки врачебных решений в медицине // Научный журнал. – 2018. – С. 9-15.

12. Ватазин А.В., Карпов Л.Е., Юдин В.Н. Виртуальная интеграция и консолидация знаний в распределенной системе поддержки врачебных решений // Альманах клинической медицины. – 2009. – №20. – С.83-86.

13. Юдин В.Н., Карпов Л.Е., Ватазин А.В. Методы интеллектуального анализа данных и вывода по прецедентам в программной системе поддержки врачебных решений // Альманах клинической медицины. – 2008. – С. 266-269.

14. Юдин В.Н. Система информационной поддержки врачебных решений, основанная на модифицированном методе динамического кластерного анализа // Труды института системного программирования РАН. – 2002. – С. 103-119.

15. Лукашевич И.П., Дмитрова Е.Д., др. Системы поддержки принятия врачебных решений // Врач и информационные технологии. – 2007. – №4. – С. 99-101.

16. Поворознюк А.И. Система поддержки принятия решения в медицине на основе синтеза структурированных моделей объектов диагностики // Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. – 2009. – С. 170-176.

17. Литвин А.А., Литвин В.А. Системы поддержки принятия решений в хирургии // Новости хирургии. – 2014. – №1. – С. 96-100.

18. Экономическая информатика и информация [Электронный ресурс] // Обучение в интернет. URL: https://www.lessons-tva.info/edu/e-inf1/inf1-1-2.html

19. Кобринский Б.А., Зарубина Т.В. Медицинская информатика. – М.: Академия, 2009. – 192 с.

20. Ожирение и избыточный вес [Электронный ресурс] // Всемирная организация здравоохранения. URL: https://www.who.int/ru/news-room/fact-sheets/detail/obesity-and-overweight

21. Все об ожирении у детей и подростков [Электронный ресурс]. URL: https://polismed.ru/nwobesity-post002.html

22. Система поддержки принятия решений [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Система_поддержки_принятия_решений

23. Методы принятия управленческих решений [Электронный ресурс]. URL: https://studme.org/31911/menedzhment/metody_podderzhki_prinyatiya_upravlencheskih_resheniy_osnove_informatsionnyh_tehnologiy

24. Выброс [Электронный ресурс]. URL:https://ru.wikipedia.org/wiki/Выброс

25. Обработка пропусков в данных [Электронный ресурс]. URL: https://basegroup.ru/community/articles/missing

26. Методы восстановления пропусков в данных [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/images/4/48/Methods_for_missing_value.pdf

27. Метод множественного восстановления данных [Электронный ресурс]. URL: https://publications.hse.ru/mirror/pubs/share/folder/21tn35z9vl/direct/92272011

28. Метод главных компонент [Электронный ресурс]. URL: https://wiki.loginom.ru/articles/principal-component-analysis.html

29. Факторный анализ [Электронный ресурс]. URL: https://wiki.loginom.ru/articles/factorial-analysis.html

30. Data Mining Algorithms In R/Dimensionality Reduction [Электронный ресурс]. URL:

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Dimensionality_Reduction

31. R (язык программирования) [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/R_(язык_программирования)

32. RStudio [Электронный ресурс]. URL:https://ru.wikipedia.org/wiki/RStudio

33. SPSS [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/SPSS

34. Statistica [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Statistica

35. Сравнение программных продуктов для анализа данных [Электронный ресурс]. URL: http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-r-matlab-scipy-ms-excel-sas-spss-stata

36. Обзор статистических программ [Электронный ресурс]. URL: http://www.sciencefiles.ru/section/46/

37. Сравнение программных продуктов для анализа данных [Электронный ресурс]. URL: http://www.mbureau.ru/blog/sravnenie-programmnyh-produktov-dlya-analiza-dannyh-r-matlab-scipy-ms-excel-sas-spss-stata

38. Наиболее полный список инструментов для анализа данных и машинного обучения [Электронный ресурс]. URL: http://ru.datasides.com/big-data-analytic-tools/#RapidMiner

39. Comparison of statistical packages [Электронный ресурс]. URL: https://en.wikipedia.org/wiki/Comparison_of_statistical_packages

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

41. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.

42. ГОСТ 12.0.003-2015. Опасные и вредные производственные факторы. Классификация.

43. ПУЭ - Правила устройства электроустановок. Изд. 7.

44. СанПиН 2.2.2/2.4.1340–03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы

45. СанПиН 2.2.4.548-96. Гигиенические требования к микроклимату производственных помещений.

46. СанПиН 2.2.1/2.1.1.1278-03. Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий.

47. ГОСТ 12.1.003-2014 ССБТ. Шум. Общие требования безопасности

48. СанПиН 2.2.4.3359-16. Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.

49. ГОСТ 12.1.006-84 ССБТ. Электромагнитные поля радиочастот. Общие требования безопасности.

50. Р 2.2.2006-05 Гигиена труда. Руководство по гигиенической оценке факторов рабочей среды и трудового процесса. Критерии и классификация условий труда.

51. Умственный труд, его физиологические особенности. Меры профилактики умственного утомления [Электронный ресурс]. URL: http://fbuz19.ru/about/news/detail.php?ID=736

52. Монотонность труда [Электронный ресурс]. URL: http://kursak.net/monotonnost-truda/

53. ГОСТ Р 56397-2015. Техническая экспертиза работоспособности радиоэлектронной аппаратуры, оборудования информационных технологий, электрических машин и приборов. Общие требования.

54. ГОСТ 12.3.031-83 Система стандартов безопасности труда (ССБТ). Работы с ртутью. Требования безопасности.

55. СНиП 21-01-97* Пожарная безопасность зданий и сооружений.

56. СП 12.13130.2009 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.

Приложение А

(справочное)

Обзор публикаций по теме исследования

Таблица А.1 – Публикации по теме исследования

№	Название работы, авторы	Аннотация
1	Система поддержки принятия врачебных решений. (Врач и информационные технологии, 2013, №6, 67-75 с.) Атьков О.Ю., Кудряшов Ю.Ю., Прохоров А.А., Касимов О.В. [1]	В статье представлена система поддержки принятия врачебных решений для диагностики сердечнососудистых заболеваний. Описаны возможности персонализации СППВР, работающей в составе информационной системы клиники.
2	Телемедицинская профилактика, реабилитация и управление здоровьем: проблемы и решения. (Врач и информационные технологии, 2016, №2, 73-80 с.) Кудряшов Ю.Ю., Атьков О.Ю., Касимов О.В. [2]	В работе приводится анализ проблем внедрения технологий персонализированной телемедицины в практическое здравоохранение и пути усовершенствования технологий для применения в области кардиоваскулярных заболеваний.
3	Современные модели систем поддержки принятия врачебных решений в хирургической практике. Состояние проблемы. (Медицинский вестник Башкортостана, 2016, 96-101 с.) Зарипова Г.Р., Богданова Ю.А. Галимов О.В., Катаев В.А., Биккинина Г.М. [3]	В статье приведен анализ возможностей современных интеллектуальных систем поддержки принятия врачебных решений в практике врача-хирурга. Описывается общее состояние проблемы с учетом существующих систем, анализируются структуры и механизмы, лежащие в основе конструирования СППВР. Приводится описание принципов работы существующих СППР при распространенных видах хирургических вмешательств.
4	Современные модели экспертных медицинских систем в прогнозировании операционного риска при наиболее распространенных интраабдоминальных вмешательствах (Медицинский альманах, 2017, 9-12 с.) Богданова Ю.А., Зарипова Г.Р., Катаев В.А., Галимов О.В. [4]	В статье приводится обзор моделей систем поддержки принятия врачебных решений для распространенных патологий. Описываются принципы построения современных экспертных систем, проводится их анализ и выявление существующих преимуществ. Описывается система поддержки принятия решений с использованием искусственных нейронных сетей и клинико-диагностические параметры электронного алгоритма.
5	Разработка архитектуры базы знаний системы поддержки принятия врачебных решений, основанной на графовой базе данных. (Медицинские технологии. Оценка и	Среди форм представления знаний выделяется семантическая сеть, по своей структуре повторяющая модель графовых баз данных, обладающих необходимыми преимуществами для работы со знаниями.

	выбор, 2018, 42-48 с.) Киселев К.В., Ноева Е.А., Выборов О.Н., Зорин А.В., Потехина А. В., Осяева М.К., Швырев С.Л., Мартынюк Т.В., Чазова И.Е., Зарубина Т.В. [5]	Цель исследования. Разработка архитектуры базы знаний системы поддержки принятия врачебных решений для инструментальной диагностики стенокардии, основанную на онтологическом подходе, с использованием графовой базы данных
6	Интеллектуальная система принятия решений в условиях выбора тактики лечения хронических заболеваний на основе облачных технологий (Вестник Воронежского государственного технического университета, 2014) Спирячин А.А., Бурковский В.Л. [6]	В статье описывается обобщённая модель интеллектуальной системы принятия решений в условиях выбора тактики лечения хронических заболеваний на основе облачных технологий.
7	Интеллектуальная система поддержки принятия врачебных решений в дерматовенерологии. (Проблемы медицинской микологии, 2014, №3, 59-65 с.) Раводин Р.А. [7]	В статье рассмотрены интеллектуальные системы в медицине; особое внимание уделено особенностям разработки интеллектуальной системы поддержки врачебных решений в дерматовенерологии.
8	Оценка эффективности интеллектуальной системы поддержки принятия врачебных решений. (Вестник СПбГУ, 2016, сер.10, вып.2, 62-68 с.) Купеева И.А., Разнатовский К.И., Раводин В.А., Карелин В.В., Буре В.М., Гусаров М.В. [8]	В статье рассматривается ИСППВР по дерматовенерологии “Logoderm”. Разработанная авторами система осуществляет автоматизированный анализ вносимых симптомов и способна в сложных случаях проводить телемедицинские консультации с экспертами.
9	Особенности проектирования медицинской информационной системы поддержки принятия решений, основанной на интеллектуальном анализе данных. (Кибернетика и вычисл. техника. 2014. Вып. 177, 79-87 с.) Кнышов Г.В., Руденко А.В., Настенко Е.А., Яковленко А.В., Сиромаха С.О., Галич С.С. [9]	Описаны структура и этапы проектирования информационной системы поддержки принятия врачебных решений при ведении больных с ишемической болезнью сердца. В основе системы лежит интеллектуальный анализ данных, представленный математическими моделями прогноза развития острой сердечной недостаточности, что позволяет на дооперационном этапе выявить риск развития осложнения в раннем послеоперационном периоде. Разработанная технология позволяет выявлять риск развития осложнения и принимать решения с целью коррекции лечебного процесса.
10	Системы поддержки принятия решений в здравоохранении и обучении. (Врач и информационные технологии, 2010, №2, 39-45 с.)	В статье рассматриваются особенности, вопросы применения и перспективы развития СППР в клинической практике и в образовательном процессе. Обращено

	Кобринский Б.А. [10]	внимание на эффект самообучения, присущий системам на основе знаний.
11	Системы поддержки врачебных решений в медицине. (Научный журнал, 2018, 9-15 с.) Фролов С.В., Куликов А.Ю. , Остапенко О.А., Стрыгина Е.В. [11]	В статье анализируются основные направления в области разработки и применения интеллектуальных систем поддержки принятия врачебных решений в медицине, существующие на сегодняшний день.
12	Виртуальная интеграция и консолидация знаний в распределенной системе поддержки врачебных решений. (Альманах клинической медицины, 2009, №20, 83-86 с.) Ватазин А.В., Карпов Л.Е., Юдин В.Н. [12]	В работе описывается переход к функционированию системы поддержки принятия врачебных решений с интеграцией баз прецедентов автономных систем, описываются проблемные моменты и варианты подобной интеграции.
13	Методы интеллектуального анализа данных и вывода по прецедентам в программной системе поддержки врачебных решений. (Альманах клинической медицины, 2008, 266-269с.) Юдин В.Н., Карпов Л.Е., Ватазин А.В. [13]	В работе приводится описание подхода построения СППРВ на основе интеллектуального анализа данных алгоритмов вывода по прецедентам. Метод построен на разбиении базы прецедентов на классы. Приводится решение проблемы определения прецедентов для не полностью описанных объектов.
14	Система информационной поддержки врачебных решений, основанная на модифицированном методе динамического кластерного анализа. (Труды института системного программирования РАН, 2002, 103-119 с.) Юдин В.Н. [14]	В статье приводится описание СППВР, основанной на использовании метода аналогий и дифференциального ряда. Математическим аппаратом системы является модифицированный метод кластерного анализа.
15	Системы поддержки принятия врачебных решений. (Врач и информационные технологии, 2007, №4, 99-101 с.) Лукашевич И.П., Дмитрова Е.Д., Киселева О.А., Мачинская Р.И., Ткачёва Т.В., Фишман М.Н., Шкловский В.М. [15]	В работе представлена методика структурирования медицинской информации, при которой выявляются ключевые характеристики в минимально возможном объеме, достаточные для принятия решения.
16	Система поддержки принятия решения в медицине на основе синтеза структурированных моделей объектов диагностики. (Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика, 2009, 170-176 с.)	В статье представлен разработанный метод синтеза структурированных моделей объектов диагностики. Метод применим для всех этапах преобразования информации при построении компьютерных СППР в медицине. В методе учитываются экспертные оценки и неопределенность параметров модели.

	Поворознюк А.И. [16]	
17	Системы поддержки принятия решений в хирургии. (Новости хирургии, 2014, №1, 96-100с.) Литвин А.А., Литвин В.А. [17]	В статье приводится анализ использования СППР в области хирургии. Существующие медицинские системы позволяют проверять предположения и использовать технологии в сложных клинических случаях.

Приложение Б

(справочное)

Обзор зарегистрированных патентных разработок

Таблица Б.1 – Зарегистрированные патентные разработки

№	Название изобретения	№ и дата охранного документа	Правообладатель
1	Система и способ для поддержки принятия клинических решений для планирования терапии с помощью логического рассуждения на основе прецедентов	№ 0002616985 от 19.04.2017	КОНИНКЛЕЙКЕ ФИЛИПС ЭЛЕКТРОНИКС Н.В. (NL)
	Предложен постоянный машиночитаемый носитель данных, на котором хранится совокупность команд, исполняемых процессором, при этом совокупность команд приводится в действие для того, чтобы: принимать данные рассматриваемого пациента; сопоставлять данные рассматриваемого пациента с множеством данных предшествующих пациентов; выбирать множество данных предшествующих пациентов на основе уровня сходства между выбранным множеством данных предшествующих и рассматриваемого пациентов; предоставлять множество выбранных данных предшествующих пациентов пользователю; генерировать план лечения на основе соответствующих планов лечения из множества выбранных данных пациентов; оснащать весовыми коэффициентами каждый из планов лечения на основе сходства пациентов; и представлять пользователю посредством устройства отображения графическое сопоставление между данными пациентов. Предложена система для поддержки принятия клинических решений для планирования терапии, содержащая: интерфейс пользователя, принимающий совокупность данных рассматриваемого пациента; базу данных, хранящую множество совокупностей данных предшествующих пациентов; механизм поиска сходства; систему генерирования планов. Группа изобретений позволяет генерировать план лечения высокого качества за счет использования априорной информации.		
2	Автоматизированная система распределенной когнитивной поддержки принятия диагностических решений в медицине	№ 0002609737 от 02.02.2017	ООО "СИАМС"
	Для поддержки принятия диагностических решений используют автоматизированную систему, которая содержит модуль хранения результатов обследования, модуль поддержки принятия решений, модуль анализа изображений, модуль распределенного хранения результатов обследования, модуль управления знаниями, а также систему управления базами данных, при этом модуль хранения результатов обследования и модуль распределенного хранения результатов обследования содержат подсистему управления данными, подсистему анализа данных и подсистему удаленного доступа, модуль поддержки принятия решений содержит подсистему управления расчетами, подсистему визуализации, подсистему классификации, модуль анализа изображений содержит подсистему анализа изображений, модуль управления знаниями содержит подсистему машинного обучения, система управления базами данных содержит базу анкет, базу изображений и базу дескрипторов, а также блок управления анкетами и блок управления изображениями. Система оптимизирует процесс диагностики в режиме реального времени за счет автоматического анализа данных.		
3	Поддержка принятия клинических решений	№ 0002573218 от 20.01.2016	КОНИНКЛЕЙКЕ ФИЛИПС

			ЭЛЕКТРОНИКС Н.В. (NL)
	Технический результат - повышение точности диагностирования пациента. Система для поддержки принятия клинических решений содержит подсистемы, предназначенные для идентификации множества типов информации о пациенте, используемых в правиле принятия решений, входящем в систему поддержки принятия клинических решений; обращения по меньшей мере к одному банку данных с целью проверки наличия элементов информации о пациенте, относящихся к конкретному пациенту; определения того, какие типы информации о пациенте, используемые в правиле принятия решений, имеют соответствующие элементы информации о пациенте, относящиеся к конкретному пациенту; и предоставления указания полноты имеющейся информации, относящейся к конкретному пациенту, с точки зрения типов информации о пациенте, в которых имеются соответствующие элементы информации о пациенте, или типов информации о пациенте, в которых недостает соответствующих элементов информации о пациенте. Указанное можно выполнять с использованием рабочей станции, предназначенной для создания медицинских изображений.		
4	Система поддержки принятия клинических решений на основе принятия решений по сортировке пациентов	№ 0002679572 от 11.02.2019	КОНИНКЛЕЙКЕ ФИЛИПС Н.В. (NL)
	Система поддержки принятия клинических решений содержит машиночитаемый носитель данных для поддержки принятия клинических решений, закодированный машиночитаемыми командами. Вычислительная система включает в себя: вычислительный процессор; средства ввода/вывода и машиночитаемый носитель данных, закодированный модулем сортировки пациентов, при этом средства ввода/вывода выполнены с возможностью приема электрического сигнала, который включает в себя набор измеренных физиологических параметров пациента. Команды модуля сортировки пациентов включают: сравнение указанных физиологических параметров с заданным диапазоном параметров на основании выходных нормативов датчика в электронном формате; идентификацию данных, необходимых для определения вероятности и исследуемой степени тяжести пациента исходя из нормативных данных; получение указанных идентифицированных данных в электронном формате; определение вероятности и степени тяжести исследуемого состояния пациента исходя из принятых идентифицируемых данных; определение рекомендуемого порядка действий для пациента, исходя из полученных вероятности, степени тяжести ресурсов медицинского учреждения и нормативных событий; и вывод на экран дисплея визуального представления вероятности, степени тяжести и рекомендуемого порядка действий. Группа изобретений обеспечивает облегчение выполнения раннего обнаружения, сортировки пациентов и вмешательства при определенных исследуемых состояниях пациента		

Приложение В

(справочное)

Набор правил для Decision Tree

Tree

```
WS_bt > 65
|   DAD_bt > 68.500
|   |   TMT_bt > 153.750
|   |   |   Mass_bt > 86.500
|   |   |   |   IgM_bt > 1.260: 2 {1=0, 2=2, 3=0, 4=1, 5=0}
|   |   |   |   IgM_bt ≤ 1.260: 1 {1=4, 2=0, 3=0, 4=0, 5=0}
|   |   |   Mass_bt ≤ 86.500
|   |   |   |   OXS_bt > 4.530
|   |   |   |   |   Leptin_bt > 28.150
|   |   |   |   |   |   TBL_bt > 4.100: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   |   |   TBL_bt ≤ 4.100: 2 {1=0, 2=2, 3=0, 4=0, 5=0}
|   |   |   |   |   Leptin_bt ≤ 28.150: 5 {1=0, 2=0, 3=0, 4=0, 5=3}
|   |   |   |   OXS_bt ≤ 4.530
|   |   |   |   |   APF_bt > 27.305: 3 {1=0, 2=0, 3=10, 4=0, 5=0}
|   |   |   |   |   APF_bt ≤ 27.305
|   |   |   |   |   |   DAD_bt > 72: 4 {1=0, 2=0, 3=0, 4=2, 5=0}
|   |   |   |   |   |   DAD_bt ≤ 72: 1 {1=1, 2=0, 3=1, 4=0, 5=0}
|   |   TMT_bt ≤ 153.750
|   |   |   IL_6_bt > 2.400
|   |   |   |   TAG_bt > 2.215: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   TAG_bt ≤ 2.215
|   |   |   |   |   WS_bt > 80.500: 4 {1=0, 2=0, 3=1, 4=17, 5=0}
|   |   |   |   |   WS_bt ≤ 80.500
|   |   |   |   |   |   Cortisol_bt > 439.050: 5 {1=0, 2=0, 3=0, 4=0, 5=2}
|   |   |   |   |   |   Cortisol_bt ≤ 439.050: 4 {1=0, 2=0, 3=0, 4=3, 5=0}
|   |   |   IL_6_bt ≤ 2.400
|   |   |   |   MG_bt > 3.275: 4 {1=1, 2=0, 3=0, 4=4, 5=0}
|   |   |   |   MG_bt ≤ 3.275
|   |   |   |   |   IgG_bt > 16.175: 3 {1=0, 2=0, 3=5, 4=0, 5=0}
|   |   |   |   |   IgG_bt ≤ 16.175
|   |   |   |   |   |   CIC_bt > 135: 2 {1=0, 2=1, 3=1, 4=0, 5=0}
|   |   |   |   |   |   CIC_bt ≤ 135: 5 {1=0, 2=0, 3=0, 4=0, 5=7}
|   DAD_bt ≤ 68.500
|   |   TBL_bt > 4.750
|   |   |   Overweight_bt > 16.200
|   |   |   |   TTG_bt > 1.350
|   |   |   |   |   T_hel_bt > 19.500
|   |   |   |   |   |   KK_bt > 108
|   |   |   |   |   |   |   TMT_bt > 40: 1 {1=3, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   |   |   TMT_bt ≤ 40: 2 {1=0, 2=2, 3=0, 4=0, 5=0}
|   |   |   |   |   |   KK_bt ≤ 108: 3 {1=0, 2=0, 3=4, 4=0, 5=0}
|   |   |   |   |   T_hel_bt ≤ 19.500: 1 {1=12, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   TTG_bt ≤ 1.350
|   |   |   |   |   DAD_bt > 61.500
|   |   |   |   |   |   T_lym_bt > 31.500: 2 {1=0, 2=1, 3=1, 4=0, 5=0}
|   |   |   |   |   |   T_lym_bt ≤ 31.500: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   DAD_bt ≤ 61.500: 3 {1=0, 2=0, 3=8, 4=0, 5=0}
|   |   |   Overweight_bt ≤ 16.200: 5 {1=1, 2=0, 3=0, 4=0, 5=3}
|   |   TBL_bt ≤ 4.750
|   |   |   TTG_bt > 1.050
|   |   |   |   PI_bt > 30.715
|   |   |   |   |   DAD_bt > 64.500
|   |   |   |   |   |   T_lym_bt > 41: 5 {1=0, 2=0, 3=0, 4=0, 5=2}
|   |   |   |   |   |   T_lym_bt ≤ 41
|   |   |   |   |   |   |   DP_bt > 271.200: 4 {1=0, 2=0, 3=0, 4=5, 5=0}
```

```
|   |   |   |   |   |   |   DP_bt ≤ 271.200
|   |   |   |   |   |   |   |   TS_bt > 94: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   |   |   |   |   TS_bt ≤ 94: 2 {1=0, 2=1, 3=0, 4=1, 5=0}
|   |   |   |   |   DAD_bt ≤ 64.500
|   |   |   |   |   |   Leptin_bt > 26.200
|   |   |   |   |   |   |   Mass_bt > 65
|   |   |   |   |   |   |   |   OXS_bt > 4.565: 3 {1=0, 2=0, 3=2, 4=0, 5=0}
|   |   |   |   |   |   |   |   OXS_bt ≤ 4.565: 1 {1=6, 2=0, 3=0, 4=0, 5=1}
|   |   |   |   |   |   |   Mass_bt ≤ 65
|   |   |   |   |   |   |   |   BMI_bt > 23.600
|   |   |   |   |   |   |   |   |   TMT_bt > 80.150: 2 {1=0, 2=5, 3=0, 4=0, 5=0}
|   |   |   |   |   |   |   |   |   TMT_bt ≤ 80.150: 1 {1=1, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   |   |   BMI_bt ≤ 23.600: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   |   Leptin_bt ≤ 26.200
|   |   |   |   |   |   |   FNO_bt > 2
|   |   |   |   |   |   |   |   TS_bt > 90.500: 4 {1=0, 2=0, 3=0, 4=2, 5=0}
|   |   |   |   |   |   |   |   TS_bt ≤ 90.500: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   |   |   |   |   FNO_bt ≤ 2
|   |   |   |   |   |   |   |   Overweight_bt > 36.100: 1 {1=5, 2=0, 3=0, 4=0,
5=0}
|   |   |   |   |   |   |   |   Overweight_bt ≤ 36.100: 4 {1=0, 2=0, 3=0, 4=4,
5=0}
|   |   |   |   PI_bt ≤ 30.715
|   |   |   |   |   MG_bt > 3.145
|   |   |   |   |   |   WS_bt > 72: 1 {1=2, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   WS_bt ≤ 72: 4 {1=0, 2=0, 3=0, 4=2, 5=0}
|   |   |   |   |   MG_bt ≤ 3.145: 2 {1=0, 2=10, 3=0, 4=0, 5=0}
|   |   |   TTG_bt ≤ 1.050: 1 {1=9, 2=0, 3=0, 4=0, 5=0}
WS_bt ≤ 65
|   T3_bt > 1.495
|   |   BMI_bt > 25.850: 1 {1=14, 2=0, 3=0, 4=0, 5=0}
|   |   BMI_bt ≤ 25.850
|   |   |   T_sup_bt > 15
|   |   |   |   HOMA_bt > 3.817: 1 {1=1, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   HOMA_bt ≤ 3.817: 2 {1=0, 2=5, 3=0, 4=0, 5=0}
|   |   |   T_sup_bt ≤ 15: 1 {1=7, 2=1, 3=0, 4=0, 5=0}
|   T3_bt ≤ 1.495
|   |   BMI_bt > 30.450: 1 {1=4, 2=0, 3=0, 4=0, 5=0}
|   |   BMI_bt ≤ 30.450
|   |   |   HOMA_bt > 3.083: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   HOMA_bt ≤ 3.083
|   |   |   |   DAD_bt > 52.500: 3 {1=0, 2=0, 3=7, 4=0, 5=0}
|   |   |   |   DAD_bt ≤ 52.500: 2 {1=0, 2=1, 3=1, 4=0, 5=0}
```

Приложение Г

(справочное)

Набор правил для Random Forest

Tree

```
CIC_bt > 97.500
|   IgM_bt > 1.160
|   |   TTG_bt > 2.250
|   |   |   Lysozyme_bt > 38.500
|   |   |   |   T3_bt > 1.850
|   |   |   |   |   TMT_bt > 38.600: 2 {1=0, 2=9, 3=0, 4=0, 5=0}
|   |   |   |   |   TMT_bt ≤ 38.600: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   T3_bt ≤ 1.850
|   |   |   |   |   BMI_bt > 28.950: 3 {1=0, 2=0, 3=1, 4=0, 5=0}
|   |   |   |   |   BMI_bt ≤ 28.950: 5 {1=0, 2=0, 3=0, 4=0, 5=3}
|   |   |   Lysozyme_bt ≤ 38.500
|   |   |   |   PI_bt > 48.735
|   |   |   |   |   DAD_bt > 63.500: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   |   DAD_bt ≤ 63.500: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   PI_bt ≤ 48.735
|   |   |   |   |   DAD_bt > 77.500: 5 {1=0, 2=0, 3=0, 4=0, 5=1}
|   |   |   |   |   DAD_bt ≤ 77.500: 1 {1=8, 2=0, 3=0, 4=0, 5=0}
|   |   TTG_bt ≤ 2.250
|   |   |   Cortisol_bt > 358.550
|   |   |   |   DAD_bt > 72.500: 5 {1=0, 2=0, 3=0, 4=0, 5=2}
|   |   |   |   DAD_bt ≤ 72.500
|   |   |   |   |   DP_bt > 310.750: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   DP_bt ≤ 310.750: 4 {1=0, 2=0, 3=0, 4=5, 5=0}
|   |   |   Cortisol_bt ≤ 358.550
|   |   |   |   TS_bt > 94: 1 {1=5, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   TS_bt ≤ 94
|   |   |   |   |   TS_bt > 79: 4 {1=0, 2=0, 3=0, 4=3, 5=0}
|   |   |   |   |   TS_bt ≤ 79: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   IgM_bt ≤ 1.160
|   |   DP_bt > 291.100
|   |   |   IL_6_bt > 6.500: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   IL_6_bt ≤ 6.500
|   |   |   |   Degree_of_obesity > 2.500: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   Degree_of_obesity ≤ 2.500
|   |   |   |   |   HOMA_bt > 2.175: 1 {1=5, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   HOMA_bt ≤ 2.175: 3 {1=0, 2=0, 3=1, 4=0, 5=0}
|   |   DP_bt ≤ 291.100
|   |   |   TAG_bt > 0.770
|   |   |   |   CIC_bt > 105: 4 {1=0, 2=0, 3=0, 4=12, 5=0}
|   |   |   |   CIC_bt ≤ 105: 5 {1=0, 2=0, 3=0, 4=0, 5=1}
|   |   |   TAG_bt ≤ 0.770
|   |   |   |   TTG_bt > 2.100
|   |   |   |   |   TS_bt > 81.500: 3 {1=0, 2=0, 3=6, 4=0, 5=0}
|   |   |   |   |   TS_bt ≤ 81.500: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   TTG_bt ≤ 2.100: 4 {1=0, 2=0, 3=0, 4=6, 5=0}
CIC_bt ≤ 97.500
|   WS_bt > 68.500
|   |   ATTPO_bt > 7.550
|   |   |   CIC_bt > 32.500
|   |   |   |   TMT_bt > 131
|   |   |   |   |   TAG_bt > 0.530
|   |   |   |   |   |   TBL_bt > 3.550
|   |   |   |   |   |   |   DP_bt > 302.250
|   |   |   |   |   |   |   |   Degree_of_obesity > 2: 1 {1=1, 2=0, 3=0, 4=0,
5=0}
```

```
|   |   |   |   |   |   |   |   Degree_of_obesity ≤ 2: 3 {1=0, 2=0, 3=1, 4=0,
5=0}
|   |   |   |   |   |   |   DP_bt ≤ 302.250: 1 {1=8, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   |   TBL_bt ≤ 3.550: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   |   TAG_bt ≤ 0.530
|   |   |   |   |   |   Mass_bt > 78: 5 {1=0, 2=0, 3=0, 4=0, 5=1}
|   |   |   |   |   |   Mass_bt ≤ 78: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   |   TMT_bt ≤ 131
|   |   |   |   |   DP_bt > 219.300
|   |   |   |   |   |   HOMA_bt > 3.505
|   |   |   |   |   |   |   DP_bt > 275.200: 5 {1=0, 2=0, 3=0, 4=0, 5=1}
|   |   |   |   |   |   |   DP_bt ≤ 275.200
|   |   |   |   |   |   |   |   aXC_bt > 1.630: 3 {1=0, 2=0, 3=1, 4=0, 5=0}
|   |   |   |   |   |   |   |   aXC_bt ≤ 1.630: 4 {1=0, 2=0, 3=0, 4=5, 5=0}
|   |   |   |   |   |   HOMA_bt ≤ 3.505: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   |   |   DP_bt ≤ 219.300
|   |   |   |   |   |   Degree_of_obesity > 2: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   Degree_of_obesity ≤ 2: 5 {1=0, 2=0, 3=0, 4=0, 5=4}
|   |   |   CIC_bt ≤ 32.500: 2 {1=0, 2=6, 3=0, 4=0, 5=0}
|   |   ATTPO_bt ≤ 7.550
|   |   |   OXS_bt > 3.770
|   |   |   |   Leptin_bt > 46.600
|   |   |   |   |   IgM_bt > 0.850
|   |   |   |   |   |   T3_bt > 3.050: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   T3_bt ≤ 3.050
|   |   |   |   |   |   |   Overweight_bt > 17.500: 3 {1=0, 2=0, 3=11, 4=0, 5=0}
|   |   |   |   |   |   |   Overweight_bt ≤ 17.500: 4 {1=0, 2=0, 3=0, 4=1, 5=0}
|   |   |   |   |   IgM_bt ≤ 0.850
|   |   |   |   |   |   IgM_bt > 0.790: 2 {1=0, 2=3, 3=0, 4=0, 5=0}
|   |   |   |   |   |   IgM_bt ≤ 0.790: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   Leptin_bt ≤ 46.600
|   |   |   |   |   IgM_bt > 0.985
|   |   |   |   |   |   PI_bt > 66.580: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   PI_bt ≤ 66.580: 1 {1=22, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   IgM_bt ≤ 0.985
|   |   |   |   |   |   WS_bt > 92.500: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   |   WS_bt ≤ 92.500: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   OXS_bt ≤ 3.770
|   |   |   |   FNO_bt > 4.850: 4 {1=0, 2=0, 3=0, 4=2, 5=0}
|   |   |   |   FNO_bt ≤ 4.850
|   |   |   |   |   TBL_bt > 3.700: 3 {1=0, 2=0, 3=11, 4=0, 5=0}
|   |   |   |   |   TBL_bt ≤ 3.700: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   WS_bt ≤ 68.500
|   |   Overweight_bt > 51.500: 1 {1=16, 2=0, 3=0, 4=0, 5=0}
|   |   Overweight_bt ≤ 51.500
|   |   |   IL_6_bt > 2.600
|   |   |   |   ATTPO_bt > 1.200: 1 {1=12, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   ATTPO_bt ≤ 1.200
|   |   |   |   |   BMI_bt > 22.600
|   |   |   |   |   |   WS_bt > 45.500: 1 {1=1, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   |   WS_bt ≤ 45.500: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   BMI_bt ≤ 22.600: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
|   |   |   IL_6_bt ≤ 2.600
|   |   |   |   HOMA_bt > 0.549
|   |   |   |   |   Insulin_bt > 14.750
|   |   |   |   |   |   TS_bt > 79.500: 1 {1=2, 2=0, 3=0, 4=0, 5=0}
|   |   |   |   |   |   TS_bt ≤ 79.500: 2 {1=0, 2=1, 3=0, 4=0, 5=0}
|   |   |   |   |   Insulin_bt ≤ 14.750: 2 {1=0, 2=9, 3=0, 4=0, 5=0}
|   |   |   |   HOMA_bt ≤ 0.549: 3 {1=0, 2=0, 3=3, 4=0, 5=0}
```

Приложение Д

(обязательное)

Разделы:

Chapter 1 Analysis of the subject area

Chapter 2 Choice of methods for solving the problem

Студент:

Группа	ФИО	Подпись	Дата
8КМ71	Сподина Мария Константиновна		

Руководитель ВКР:

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

Консультант – лингвист ОИЯ ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ	Диденко Анастасия Владимировна	к.ф.н.		

Chapter 1 Analysis of the subject area

1.1 Medical Data Analysis

Data analysis is a set of methods and tools for extracting information from the organized data for making decisions. Data analysis has many approaches and methods in different areas of human activity.

Data is a collection of information recorded on a particular medium for permanent storage, transmission and processing. Conversion and processing of data allow getting information. Knowledge contains recorded and processed information that can be used for making decisions. The process of solving the problem is as follows: the data are processed and analyzed using available knowledge; there are offered all feasible solutions and one of them becomes the best one. The results of this decision enrich the knowledge base (Figure 1) [18].

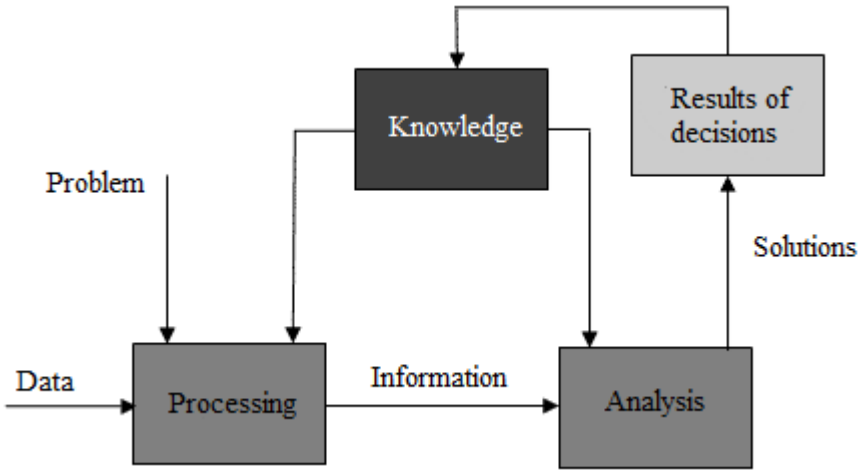

Figure 1 – Decision making process

Depending on the number of features the data can be one-dimensional, two-dimensional and multi-dimensional. Multi-dimensional data contains information about three or more features for each object. They are used to obtain information about dependencies between features, to predict the value of one variable based on the values of others. Multidimensional Data Analysis (MDA) is an approach based on the use of mathematical methods to investigate processes and systems characterized by a set of variables.

Medical data is any data related to medicine, in the narrow sense, it is data about the health status of a particular person. Working with medical data we should consider certain features:

- it is necessary to prevent loss of information;
- it is important not to provide redundant information after statistical data processing;
- it is often necessary to convert qualitative data into quantitative one;
- it is not recommended to use “zeroing” for missing values, zero may coincide with the coding of the norm of the characteristic;
- it is wrong to use the class average, especially for small samples, because the classes are not always homogeneous. In this case, it is better to exclude such observations or encode the missing values with a special sign or number [19].

The dimension of the stored data is determined by the number of different features of the object (patient). Sometimes the dimension can reach several tens or hundreds of features. That is why the reduction of dimension and the selection of the most informative features is a very important aspect in work with medical data.

A decision support system (DSS) is an information system that collects and processes data. DSS is able to influence decision-making processes in various areas of human activity. Similar system in health care is called as a clinical decision support system (CDSS).

CDSS are intended for solving the following tasks: giving alarms and reminders, assistance in the process of diagnostics, searching for suitable cases (precedents), monitoring and planning therapy, recognizing and interpreting of patterns. CDSSs are used more often to help in making a diagnosis, prescribing and adjusting the prescribed treatment. [11].

DSSs have the following advantages:

- the collected data is used as a basis to assist in determining of expected diagnostic decisions;

- diagnostic hypotheses become less dependent on the personal experience and knowledge of the doctor, the probability of missing important features decreases;
- the knowledge base contains the knowledge of experts and modern views on the problem area, which allow the system to use the necessary tools and decision-making algorithms.

1.2 Description of subject domain

Overweight and obesity among children and teenagers is one of the most serious public health problems in the 21st century. The body mass index (BMI) is widely used as the main indicator for estimating overweight. The BMI is defined as the body mass divided by the square of the body height (kg/m2). BMI is the most convenient measure of obesity and overweight, it is the same for all gender and age groups of adults. But BMI should be defined as an approximate criterion, because for different people it may match different degrees of overweight. BMI is used differently for children, it is calculated in the same way as for adults, but then compared to typical values for other children of the same age.

For children up to 5 years of age, overweight and obesity are defined as follows: if the ratio “body weight / height” exceeds the median value by more than two standard deviations, then it is overweight; if it exceeds more than three standard deviations, then it is obesity. For children at the ages from 5 to 19 overweight and obesity are defined as follows: if the ratio “BMI / age” exceeds the median value by more than one standard deviation, then it is overweight; if it exceeds more than two standard deviations, then it is obesity. [20].

The increase in the number of children with overweight is mainly due to unhealthy diet and a sedentary lifestyle. Obesity can also be as the result of diseases of the endocrine system, brain tumors and other serious diseases. Overweight in children can lead to early development of diseases such as hypertension, diabetes, coronary artery disease, cirrhosis, etc. [21]

According to statistics, about 41 million children up to 5 years old had overweight or obesity in 2016. In Africa this number has almost doubled since 2000.

In 2016, 340 million children and teenagers at the ages from 5 to 19 suffered from the problem of overweight. From 1975 to 2016 the proportion of overweight and obesity among children and teenagers at the ages from 5 to 19 increased from 4% to 18% and amounted to 124 million people. More people die from the effects of overweight and obesity than from the effects of abnormally low body weight. About 2.6 million people die from overweight or obesity every year [20].

1.3 Description of the source data

We have obtained a set of data from the clinical research conducted on the basis of the children's department of Tomsk Scientific Research Institute of Balneology and Physiotherapy. The data are presented in the form of a multidimensional table with the values of groups of clinical and laboratory parameters of patients. Patients are children and teenagers at the ages from 7 to 18 with overweight problems. The data set contains 345 objects (patients), 161 indicators (values before and after treatment) and 5 groups of treatment methods. A fragment of the data set is presented in Figure 2.

N	Name	Groupe 1	Groupe 2	Se	Receipt da	Discharge date	Age	Degree of obesit	The degree of obesity by body mass index (BMI)	Degree of goite	Related diagnos	Complicati ons	Complair
1	Patient 1	3	1	ж	9 мар 05	2 апр 05	11	I	1	I			
3	Patient 2	6	2	м	11 мар 05	25 июн 05	13	I	изб	I	7		2.3
4	Patient 3	8	3	ж	12 мар 05	10 авг 05	9	I	1	0	1,8,9		5
5	Patient 4	10	3	м	13 мар 05	4 ноя 05	10	I	изб	I	1,2,3,4		1.3
6	Patient 5	11	3	ж	14 мар 05	13 ноя 05	14	I	изб	0			
7	Patient 6	11	3	ж	15 мар 05	2 дек 05	9	II	изб	I	1.12		
8	Patient 7	11	3	ж	16 мар 05	2 дек 05	10	II	2	I	1.11		1,2,3
9	Patient 8	11	3	ж	17 мар 05	2 дек 05	11	I	изб	I	7.4		2.3
10	Patient 9	11	3	ж	18 мар 05	8 дек 05	10	I	изб	I	1.4		
11	Patient 10	11	3	м	19 мар 05	2 дек 05	15	II	изб	0			
12	Patient 11	12	1	ж	20 мар 05	29 дек 05	14	I	изб	I	7		2.3
18	Patient 12	3	1	м	24 фев 06	20 мар 06	10	II	2	0			
24	Patient 13	3	1	ж			12						
27	Patient 14	4	1	ж	17 апр 06	11 май 06	10	I	1	I			
28	Patient 15	4	1	ж			8	III	2	I			
29	Patient 16	4	1	ж	18 апр 06	11 май 06	13	III	1	0			
31	Patient 17	5	2	м	12 май 06	5 июн 06	13	III	2				
33	Patient 18	5	2	ж	19 май 09		14						

Figure 2 – Fragment of the source data set

All indicators belong to different physiological groups. The study was conducted on the main clinical and laboratory indicators of each group (Table 1).

Table 1 – The main clinical and laboratory indicators of the research

Physiological group	Clinical laboratory indicator	Indicator meaning
Clinical information	WS	Waist size
	TS	Thigh size
	Weight, kg	Body mass, kg
	Overweight, %	Overweight, %
	BMI	Body mass index
The cardiovascular system	SBP, mm Hg	Systolic blood pressure
	DBP, mm Hg	Diastolic blood pressure
Physical Working capacity	HOMA	Homa Insulin Resistance Index
	EST	Exercise and sport tolerance
	WorkC	General working capacity
Lipid metabolism	TBL (4-8 g/l)	Total blood lipids
	TAG (less than 1.7 mmol\L)	Triacylglycerides level
	aXC (more than 1.03 mmol/l)	Acholesterol level
	LDL (mmol/L)	Low density lipoproteins
	VLDL bt (mmol/l)	Very low density lipoproteins
Blood biochemistry	Alkaline phosphatase	Alkaline phosphatase in the blood serum
	Catalase (mcatal/l)	Catalase of the blood serum
	MDA	Malonic dialdehyde in the blood serum
	Ceruloplasmin	Content of ceruloplasmin in the serum
Carbohydrate metabolism	Glucose (less than 5.6 mmol/l)	Blood glucose on an empty stomach
Hormonal status	T3 (1,0-2,8)	Triiodothyronine level
	TTG (0,23-3,4)	Thyroid-stimulating hormone level
	AT TPO (до 30 у.е.)	Antibodies to thyroperoxidase
	Insulin	Insulin level in blood serum
	Leptin	Leptin level in blood serum
Cytokine status	IL-6	Interleukine-6 level in blood serum
	TNF (not > 2.5 pg / ml)	Tumor necrosis factor level
Immunological status	IgA (0,7-3,0 g/l)	Concentration of immunoglobulin A
	IgG (8,0-16,0 g/l)	Concentration of immunoglobulin G
	IgM (0,8-2,5 g/l)	Concentration of immunoglobulin M
	CIC (40-100)	Circulating immune complexes
	Lysozyme (40-60)	Activity of lysozyme
	T-lymphocytes (45-70), %	Circulating immune complexes
State of the kallikrein-kinin system	KK	Kalleriin level
	PI	Prothrombin index
	MG	Macroglobulin
	ACE	Angiotensin-converting enzyme
Blood plasma oxidative ability	NO	Nitric oxide level
	OX LDL	oxidation of low-density lipoproteins

In addition to the listed indicators, the data set contains information on the patient's belonging to the treatment group (1-5). To solve this problem, it is necessary to build a system to determine the required trajectory (group) of treatment for a new patient, based on its primary clinical and laboratory indicators.

Chapter 2 Choice of methods for solving the problem

2.1 Decision Making Techniques

Decision making in the DSS can be implemented by different methods:

- information retrieval,
- data mining,
- knowledge discovery in databases,
- case-based reasoning,
- simulation modeling,
- genetic algorithm,
- artificial neural network,
- cognitive modelling, etc. [22]

These methods are considered in more detail:

1. *Information retrieval* is the process of finding unstructured documentary information. The search process includes the sequence of operations of collecting, processing and providing the necessary information to interested parties. In general, information retrieval consists of four stages:

1) definition of information needs and formulation of information request;

2) establishment of a set of possible holders of information arrays (sources);

3) extracting information from identified information arrays;

4) familiarization with the received information and evaluation of search results.

There are several types of search:

- full text search (by full document content);
- search by metadata (according to the document attributes supported by the system, for example, title, creation date, author, etc.);
- image search (by image content).

2. *Data mining* is the identification of hidden patterns or ratios of variables in large arrays of raw data. The term data mining appeared around 1990 in the database community. Such an analysis involves solving problems of classification, modeling

and prediction, clustering, reducing descriptions, associations, variance analysis, visualization, etc.

3. *Knowledge discovery in databases (KDD)* is the process of discovering useful knowledge in databases. This knowledge can be presented in the form of patterns, rules, predictions, ratios between data elements, etc. The main tools for searching knowledge in the process of extraction in databases are analytical data mining technologies that implement the tasks of classification, clustering, regression, prediction, etc. The process of extracting knowledge in databases includes the execution of the operations sequence necessary to support the analytical process. These operations are data consolidation, preparation of analyzed data samples (including training ones), clearing data from the factors that prevent their correct analysis, transformation (data optimization for solving a specific task), data analysis, interpretation and visualization of analysis results, their application in business applications.

4. *Case-based reasoning (CBR)* is an approach to solve a new problem, using or adapting the solution of similar past problems. Case-based reasoning has been formalized for purposes of computer reasoning as a four-step process (*CBR-cycle)*: Retrieve, Reuse, Revise, Retain. There are a number of methods for extracting cases and their modifications: the nearest neighbor method, decision tree method, knowledge-based use case extraction method, extraction method for the precedents applicability. In addition to the considered methods, others can be also successfully used to extract cases, for example, the approach of artificial neural networks.

5. *Simulation modeling* is a method that allows building models for describing the processes as they would be in reality. Such a model can be played for one test or a given set of tests. The results will be determined by the random nature of the processes. It is easy to get a fairly stable statistics from these data. Popular simulation systems include: AnyLogic, Arena, eM-Plant, Powersim, GPSS.

6. *Genetic algorithm (GA)* – is a heuristic search algorithm for solving optimization and modeling problems by selecting, combining, and varying the required parameters using mechanisms resembling biological evolution. The genetic

algorithm is a type of evolutionary computation. Its distinguishing feature is the emphasis on the use of the "crossing" operator, which performs the operation of recombination of candidate solutions, whose role is similar to the role of crossing in living nature. Genetic algorithms are mainly used to search for solutions in very large, complex search spaces.

7. *Artificial neural networks (ANN)* are mathematical models, their software or hardware implementations, built on the principle of organization and functioning of biological neural networks (networks of nerve cells of a living organism). The original goal of the ANN approach was to solve problems in the same way that a human brain would. But, over time, attention moved to performing specific tasks, leading to deviations from biology. ANN have been used on a variety of tasks, including computer vision, machine translation, speech recognition, social network filtering, medical diagnosis. From the point of view of machine learning, a neural network is a special case of pattern recognition methods, discriminant analysis, clustering methods, etc.

8. *Artificial intelligence (AI)* is the science and development of intelligent machines and systems designed to understand human intelligence. Scientists have studied only some of the mechanisms of intelligence, so within this science intelligence is the computational part of the ability to achieve some goals. If the DSS work is based on artificial intelligence methods, then such a system is called an intelligent decision-making support system. [23].

2.2 Data Mining Technology

The term “Data Mining” translates as the data search, in-depth data analysis, knowledge extraction and knowledge discovery in databases. Data mining is the process of extracting information from a dataset and transforming it into a clear structure for future use.

Data Mining technology includes a large number of modeling, classification and prediction methods based on the use of artificial neural networks, decision trees, evolutionary programming, associative memory, fuzzy logic, including k-means and

k-median algorithms, methods of searching for associative rules, limited search method, genetic algorithms and various data visualization methods.

Data mining methods often include statistical methods such as: descriptive analysis, correlation and regression analysis, factor analysis, analysis of variance, component analysis, discriminant analysis, time series analysis, and many others. Data mining methods are designed to solve problems that can be divided into classes:

1. *Classification.* This is the finding specific features of the objects that determine their relation to one of the predefined classes.

2. *Clustering.* This is the partition of the set of objects and features into homogeneous groups or clusters. For this task, the classes are previously unknown, they need to be formed. The rest of the classification ideas are preserved.

3. *Association.* This is the identification of patterns among interrelated events, which is based on the consideration of simultaneously occurring events.

4. *Progression.* This is the finding of patterns among time-related events.

5. *Regression and prediction.* This is a search for the dependence of the output data on the input variables and the prediction of new results based on the identified dependencies.

6. *Visualization.* This is a graphical representation of the analyzed data. Large amounts of raw data are displayed as visual tables, charts, diagrams, graphs, etc.

2.3 Methods of source data analysis

2.3.1 Analysis of outliers

The correcting process of outliers and other incorrect information is called data cleansing when they are pre-processed. There are several types of data cleansing process that depend on the type of data to be cleared.

Outliers are data points that differ significantly from other observations within the considered sample. Outliers can occur as a result of measurement errors, data entry or interpretation, due to the nature of the data. To obtain more accurate analysis results, outliers must be removed. The simplest methods based on the interquartile range are used to detect outliers: outliers are values that are not in the range:

$$\left[\left(Q_1 - 1{,}5 \times (Q_3 - Q_1)\right), \left(Q_1 + 1{,}5 \times (Q_3 - Q_1)\right)\right],$$

where Q_1 and Q_2 are the lower and upper quartiles respectively [24].

Typically, box plots are used to identify outliers, which also called “box-and-whisker diagram”. The box plot simplifies the data by showing the following parameters: median, upper and lower quartiles, smallest and largest values, outliers that are significantly smaller or larger than the rest of the numbers in the series. As a rule, this method is applied when it is necessary to visually compare several categories of data.

2.3.2 Dealing with missing values

In practice, data sets do not always have a complete list of values, missing values are common. If we leave the missing values due to processing, then the final results may not correspond to reality. In this case, it is necessary to assess the cause of their appearance and the impact on further analysis.

To understand how to properly handle missing values, it is necessary to determine the mechanisms of their formation. There are three types of missing data:

1. *Missing completely at random (MCAR)*

Such missing values are independent of the observed parameters and variables and occur completely randomly. At the stage of restoring the missing element in the data array, it is allowed to ignore or change the value of the missing element; this does not lead to a deformation of the result.

2. *Missing at random (MAR)*

They occur due to some regularity. Missing elements in data arrays are distributed within certain subgroups of variables. The exclusion or replacement of the missing element for a certain value does not lead to a significant distortion of the results.

3. *Missing not at random (MNAR)*

They occur due to unknown factors. Such missing values are not ignored [25].

It is necessary to know what proportion data is missing. Too many missing values are the problem for further data processing and analysis. Usually, the maximum value of the missing data is 15% of the total sample. If the share of missing data exceeds 15%, then it is necessary to abandon this indicator.

Methods for handling missing values are divided into basic (elementary and simple) and advanced (Figure 3) [26].

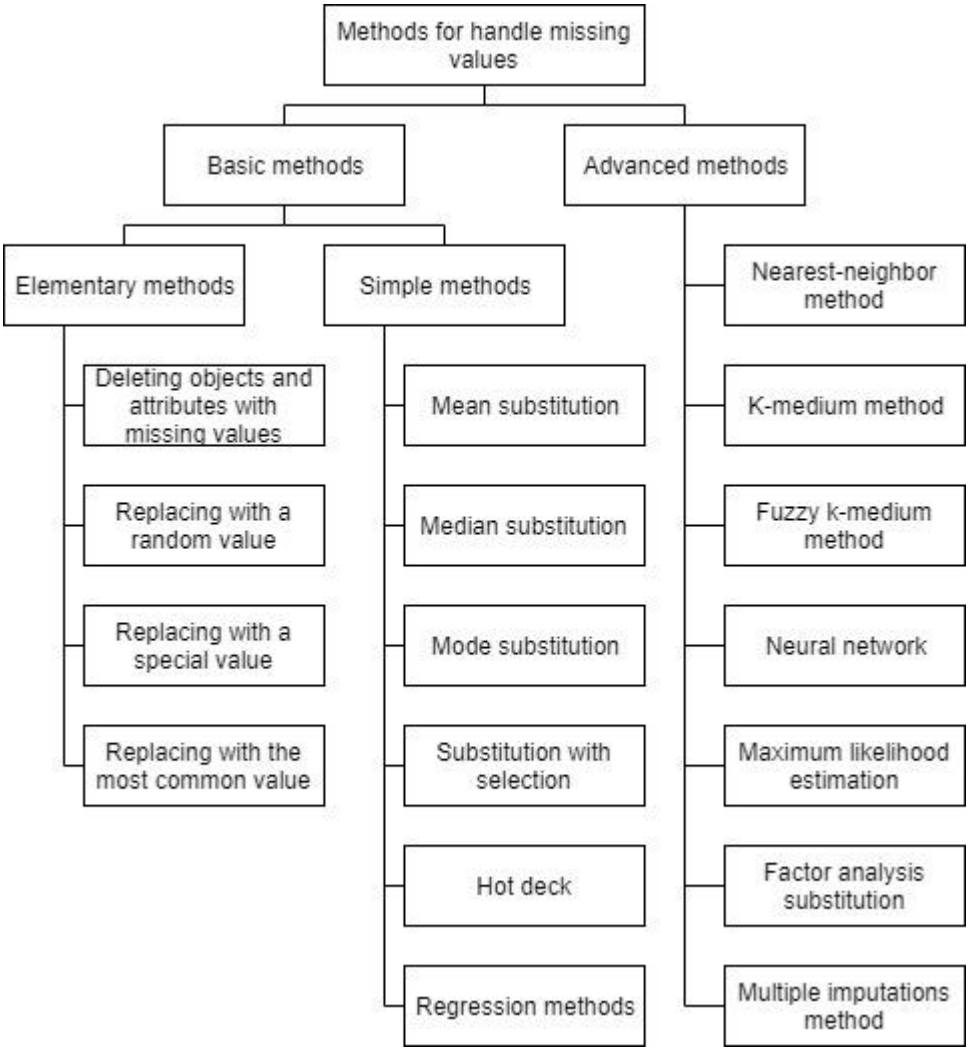

Figure 3 – Classification of Methods for handle missing values

The multiple imputation method is one of the best methods for handling missing values at the moment. This method is based on repeated data modeling using Monte-Carlo methods. The idea of the multiple imputation method is to simultaneously generate several values of the missing value instead of replacing the missing information with a single value. For practical use of all the data recovered in this way, database sets are generated with various options for missing value substitutions [27].

2.3.3 Dimension reduction of multidimensional data

Dimension reduction is a process of data transformation, consisting in reducing the number of variables by obtaining the principal variables (informative

features). Reduction of dimension can be carried out by different methods, some of them are considered in more detail.

1. *Principal component analysis (PCA)*. The method is based on the transition to a new coordinate system (y1, y2, ..., yn) from the original coordinate system (x1, x2, ..., xn) of the multidimensional feature space, which is a system of orthonormal combinations:

$$f(n) = \{y_j(x) = w_{ij}(x1 - m1) + \cdots + w_{nij}(x_n - m_n), \sum_{i=1}^{n} w_{ij}^2\}, \qquad j=1,\ldots,n$$

$$\sum_{i=1}^{n} w_{ij}w_{ik} = 0,\ j,\ k = 1,\ldots,n\ (j \neq k)$$

where m_i – expected value of the characteristic x_i.

Linear combinations are chosen so that the first principal component $y_1(x)$ has the maximum variance. In other words, the new axis will be oriented along the direction of the greatest elongation of the ellipsoid scattering of points in the feature space. The second principal component has the largest variance among all the remaining linear transformations uncorrelated with the first principal component. It is interpreted as the direction of the greatest elongation perpendicular to the first principal component. Next principal components are determined by a similar scheme [28]. Using the method of principal components leads to the problem of interpreting new variables. This method is not suitable for solving our problem.

2. *Factor analysis (FA).* This method does not use a predetermined list of factors, but helps to detect the most important of them (also, the hidden ones). The task is to identify hidden generalized factors that explain the changes in the studied indicator to a sufficient degree. The identified factors allow the construction of analytical models with a relatively small number of variables. This simplifies their implementation and interpretation by the user, reduces computational costs and time required for obtaining decisions, increases the speed of decision-making based on the analysis results [29].

3. *Informative value assessment*.

Kullback method offers a measure of divergence between the two classes (which is called divergence) as a measure of informativeness. Kullback method has been widely used in medicine for the consideration of individual factors affecting the

diagnosis. According to this method, the information content (or Kullback divergence) is calculated by the formula:

$$I(x_j) = \sum_{i=1}^{G} [P_{i1} - P_{i2}] \cdot \log_2 \frac{P_{i1}}{P_{i2}},$$

where G – number of characteristic gradations;

P_{i1} – probability of i-th gradation in the first grade:

$$P_{i1} = \frac{m_{i1}}{\sum_{i=1}^{G} m_{i1}},$$

where m_{i1} – frequency of i-th gradation in the first grade;

P_{i2} – probability of i-th gradation in the second grade:

$$P_{i2} = \frac{m_{i2}}{\sum_{i=1}^{G} m_{i2}}$$

where mi,2 – frequency of i-th gradation in the second grade.

Shannon method proposes to evaluate the information content as a weighted average of information per different gradations of the attribute. The information content of the attribute x is calculated by the formula [30]:

$$I(x_i) = 1 + \sum_{i=1}^{G} \left(P_i \cdot \sum_{k=1}^{K} P_{i,k} \cdot log_K P_{i,k} \right),$$

where G – number of characteristic gradations; K – number of classes;

P_i – probability of i-th gradation of characteristic:

$$P_i = \frac{\sum_{k=1}^{K} m_{i,k}}{N},$$

where $m_{i,k}$ – frequency of i-th gradation in the K-th class, N – total number of observations;

$P_{i,k}$ – probability of i-th grades in the K-th class:

$$P_{i,k} = \frac{m_{i,k}}{\sum_{k=1}^{K} m_{i,k}}$$

Kullback method is chosen to solve the problem of the data dimension reduction. After studying the methods and algorithms to solve the problem of

recovering the missing data we will consider the method of multiple recovery of missing data and the method of data recovery based on regression trees. Factor analysis and methods for exclusion non-informative features will be considered as methods for reducing the dimension.

2.4 Choice of the decision tool

In the software market there is a fairly large selection of tools for data analysis. We will consider the most popular tools.

1. *R* is a programming language for statistical data processing and work with graphics, a free open-source computing software environment. R supports a wide range of statistical and numerical methods and has good extensibility by using packages. Packages are libraries for specific functions or special applications. R has the ability to create high-quality graphics, which can include mathematical symbols [31].

RStudio is the environment for communicating with the R system, which makes it more convenient to create and transmit processing requests to the R system. RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics [32]. The advantage of the product is a large selection of statistical methods and more than seven thousand checked packages, fast transition to functions, integrated R documentation. The disadvantage is a small amount of documentation in Russian.

2. *SPSS (Statistical Package for Social Science)* is a computer program for statistical data processing, one of the market leaders in the field of commercial statistical products intended for conducting applied research in the social sciences. SPSS is notable for its application power for all types of statistical calculations. Program features are data entry and storage, the possibility of using variables of different types, primary descriptive statistics, marketing research, analysis of marketing research data. Disadvantages of the product include the high cost of licenses and lack of flexibility in the calculations. [33].

3. *Statstica* is a software package for statistical analysis that implements data analysis, data management, data mining, and data visualization functions using

statistical methods. The package has wide graphic capabilities, allows displaying information in the form of various types of graphs (including scientific, business, three-dimensional and two-dimensional graphs in different coordinate systems, specialized statistical graphs - histograms, matrix, categorized graphs, etc.), all components of graphs are customized [34]. Statistica contains more than 250 statistical functions, generates clear, customizable reports, but it is designed only for Windows, which is one of its drawbacks.

4. *Stata* is a software package for data analysis in the fields of economics, sociology, politics, biomedicine, etc. It includes basic statistical methods, linear models, non-parametric and multidimensional methods, cluster analysis, graphics, data operations. The disadvantage of the product is quite narrow specialization. [35].

5. *SAS* is a software product for creating accurate predictive and descriptive models based on large amounts of information. SAS is applied in a variety of scientific studies, business intelligence, etc. This package can process, modify, manage and extract data from various sources and perform statistical analysis. The advantages of the system include a flexible data exchange interface (integration), availability of tools for working with clusters, speed calculations on large data arrays. The disadvantages are the difficulty of supporting written scripts, expensive licenses and the complexity of mastering the program [36, 37].

6. *Rapid Miner* is a free environment for predictive analytics. The user does not need to write code at all. RapidMiner is offered as a service rather than as a separate program. RapidMiner provides pre-processing and visualization of data, predictive analysis and statistical modeling, supports training schemes, models and algorithms from WEKA and R scripts. RapidMiner features can be enhanced with additions, some of which are also available for free [38]. The disadvantages include the fact that not all tasks can be performed on the server, this is typical when the model requires the entire database to perform the algorithm in parts.

7. MATLAB is an application package for solving technical computing problems. MATLAB provides the user with a large number (several hundred) of functions for data analysis, covering almost all areas of mathematics. Work on most

modern operating systems, user-friendly graphical interface and ease of operation are the advantages of the program. The disadvantage is the high cost of the license and incomplete support of statistical functions [35].

Comparative analysis of the software products described above is shown in Table 2 [39].

Table 2 – Comparative analysis of decision tools

	R	SPSS	Statstica	Stata	SAS	Rapid Miner	MATLAB
Free license	+	-	-	-	-	+	-
Graphical interface	+	+	+	+	+	+	+
Console input	+	+	-	+	+	-	+
Ease of understanding and learning	+	-	-	+	-	+	-
Operating system support	Windows, Mac OS, Linux, Unix	Windows, Mac OS, Linux, Unix	Windows	Windows, Mac OS, Linux, Unix	Windows, Linux, Unix	Windows, Mac OS, Linux	Windows, Mac OS, Linux

After analyzing of decision tools we chose the R system with the RStudio development environment and partly the capabilities of Rapid Miner to perform research tasks. The R language is one of the leading statistical tools in the world. It is actively used in various fields of science and is increasingly used in the processing of medical data.