

Федеральное государственное автономное образовательное учреждение
высшего образования «Уральский федеральный университет имени первого
Президента России Б.Н. Ельцина»
Институт фундаментального образования
Кафедра интеллектуальных информационных технологий

На правах рукописи



Синотова Светлана Леонидовна

РАЗРАБОТКА МАТЕМАТИЧЕСКОГО И АЛГОРИТМИЧЕСКОГО
ОБЕСПЕЧЕНИЯ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ
ВРАЧЕБНЫХ РЕШЕНИЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ЗДОРОВЬЯ
РЕБЕНКА, ЗАЧАТОГО ПРИ ПОМОЩИ ВСПОМОГАТЕЛЬНЫХ
РЕПРОДУКТИВНЫХ ТЕХНОЛОГИЙ

Специальность — 2.3.1. Системный анализ, управление и обработка
информации, статистика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата технических наук

Научный руководитель:
кандидат физико-математических наук, доцент
Солодушкин Святослав Игоревич

Екатеринбург – 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	4
1. АНАЛИЗ СОСТОЯНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ. ПОСТАНОВКА ЦЕЛИ И ЗАДАЧ ИССЛЕДОВАНИЯ.....	10
1.1 Вспомогательные репродуктивные технологии: основные определения и методы, рассмотренные для создания системы.....	11
1.2 Сравнительный анализ существующих работ, посвященных прогнозированию результатов протокола ВРТ	14
1.3 Описание наборов данных, переданных для разработки	20
1.4 Описание основных принципов работы проектируемой СППВР	27
1.5 Анализ алгоритмов отбора признаков для машинного обучения	31
1.6 Анализ методов, используемых для решения задачи машинного обучения с учителем.....	34
1.7 Методы и алгоритмы, используемые в работе	41
1.8 Постановка цели и задач исследования	46
2. РАЗРАБОТКА АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ ПРЕДОБРАБОТКИ ИСХОДНЫХ ДАННЫХ	48
2.1 Алгоритм получения итогового набора данных.....	51
2.2 Алгоритм обработки текстовых данных	53
2.3 Алгоритм обработки орфографических ошибок	62
Выводы	66
3. РАЗРАБОТКА МАТЕМАТИЧЕСКОГО И АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ СППВР.....	67
3.1 Входные и выходные данные для каждого блока системы	67
3.2 Структура разрабатываемой системы	70
3.3 Перечисление всех моделей машинного обучения в составе системы ...	73
3.4 Результаты сравнения алгоритмов для отбора признаков и машинного обучения	76
Выводы	78
4. АПРОБАЦИЯ СОЗДАННОГО МАТЕМАТИЧЕСКОГО И АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ.....	79
4.1 Описание наборов подготовленных данных.....	79
4.2 Значимые признаки и их распределение по моделям системы	88
4.3 Разработка прогностических моделей системы и оценка их качества....	92

4.4 Разработка пользовательского программного приложения для работы с системой	104
4.5 Апробация СППВР на новых данных	106
Выводы	118
ЗАКЛЮЧЕНИЕ	119
СПИСОК СОКРАЩЕНИЙ.....	120
СПИСОК ЛИТЕРАТУРЫ.....	121
Приложение А. Распределение моделей машинного обучения по блокам СППВР	130
Приложение Б. Распределение признаков по ML-моделям СППВР и значимость признаков.....	132
Приложение В. Пример одного из деревьев решений для предсказания наступления беременности.....	138
Приложение Г. Интерфейс пользовательского программного приложения	141
Приложение Д. Копии акта о внедрении результатов диссертационного исследования и свидетельства о регистрации программы для ЭВМ	142
Приложение Е. Код основного файла интерфейса созданного программного приложения	144

ВВЕДЕНИЕ

Актуальность темы исследования. В настоящее время в Российской Федерации проживает 34,3 млн. женщин репродуктивного возраста (15-49 лет) [1], при этом около 8 млн. из них страдают бесплодием [2].

Согласно плану основных мероприятий, проводимых в рамках Десятилетия детства поставлена задача увеличения количества циклов (протоколов) вспомогательных репродуктивных технологий (ВРТ) семьям, страдающим бесплодием, за счет ОМС с 70 тыс. циклов в 2019 г. до 80 тыс. циклов в 2024 г [3]. По данным отчета Регистра ВРТ Российской ассоциации репродукции человека, наблюдается ежегодный прирост числа проводимых протоколов и доля родов после применения ВРТ от общего числа родов в России [4]. Международный практический опыт также показывает, что ВРТ можно рассматривать как эффективную и востребованную меру повышения рождаемости [5,6], но до сих пор имеют место противоречивые мнения по поводу здоровья рожденных детей [7,8,9,10,11,12]. Поэтому сотрудники центров ВРТ и врачи, занимающиеся детьми, зачатыми в результате применения ВРТ, ставят перед специалистами по данным задачу создания системы поддержки принятия врачебных решений (СППРВ) прогнозирования осложнений течения беременности и родов, а также возникновения с течением времени заболеваний у ребенка, которые можно связать с наследственными факторами, информацией о беременности и родах матери.

Данные, собираемые регулярно, и включающие в себя информацию с момента обращения родителей в центр ВРТ о проведенных манипуляциях, течении беременностей, родов, а также информацию о полном долгосрочном наблюдении за рожденными детьми могут стать основой для построения достоверных выводов об успешности применения ВРТ. А также применяться для индивидуального прогнозирования исходов ВРТ в дальнейшем. В странах с высоким уровнем информатизации в медицине есть возможность собрать полную цепочку данных [13,14]. В Российской Федерации из-за разрозненности и

несогласованности информационных систем медицинских организаций взаимосвязь между здоровьем матери и здоровьем ребенка остается неисследованной, а значит и построение прогнозов является затруднительным.

Разработка СППВР является одним из перспективных направлений искусственного интеллекта, а сами системы становятся востребованным инструментом персонифицированной медицины [15,16,17,18]. Их внедрение в практику позволит получать основанное на данных и независимое от человеческого фактора решение, показывающее правильность выбранной стратегии ведения женщины, а также скорректировать ведение протокола и увеличить число благоприятных исходов. Кроме того, данная СППВР позволит наглядно продемонстрировать клиентам основные этапы предстоящего процесса. СППВР позволит построить корректную модель наблюдения за ребенком после рождения и оказания помощи ему и его семье в дальнейшем, если это будет необходимо.

Степень разработанности темы. Несмотря на многочисленные исследования в области прогнозирования результата применения ВРТ [19–30], нет единой системы, объединяющей множество моделей, которая позволяла бы охватывать широкий спектр входных факторов и осуществлять многоэтапное прогнозирование. Разработки СППВР для предсказания исходов ВРТ ведутся по всему миру [31,32], что подчеркивает актуальность данной темы. В Российской Федерации подобных систем не существует.

Цель исследования. На основе применения алгоритмов машинного обучения к медицинской документации осуществить разработку математического и алгоритмического обеспечения системы поддержки принятия врачебных решений для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи вспомогательных репродуктивных технологий, а также создать алгоритмы обработки, компоновки и анализа неструктурированных текстовых данных на примере данных медицинских организаций.

Для достижения поставленной цели были сформулированы и решены следующие **задачи**:

1) проанализировать состояние разработок в области создания математического и алгоритмического обеспечения для прогнозирования исходов ВРТ и формализовать постановку задачи принятия врачебных решений при проведении протокола ВРТ;

2) разработать алгоритмы обработки, компоновки и анализа разнородных текстовых данных на примере данных медицинских организаций для выделения признаков описания моделируемых объектов;

3) разработать структуру СППВР, которая позволяет осуществлять многоэтапное прогнозирование здоровья ребенка, зачатого при помощи ВРТ;

4) создать математическое и алгоритмическое обеспечение СППВР для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи ВРТ, реализовать СППВР в виде программного приложения и апробировать СППВР в клинической практике.

Объект исследования – данные о проведенных протоколах ВРТ и здоровье рожденных в результате этих протоколов детей в возрасте до трех лет.

Предмет исследования – математическое и алгоритмическое обеспечение системы поддержки принятия врачебных решений для прогнозирования здоровья ребенка, рожденного при помощи ВРТ.

Научная новизна результатов диссертационного исследования заключается в следующем:

1) предложена структура СППВР, позволяющая моделировать многоэтапные процессы, которая применена к имеющимся данным о здоровье ребенка, зачатого при помощи ВРТ;

2) предложено математическое и алгоритмическое обеспечение для объединения и обработки не обладающих функциональной совместимостью данных разных медицинских учреждений в единую базу на основе методов интеллектуального анализа данных;

3) разработано математическое и алгоритмическое обеспечение СППВР, состоящее из моделей машинного обучения, позволяющих на всех этапах проведения протокола ВРТ предсказать его исходы: от наступления беременности до здоровья ребенка в возрасте трех лет.

Методология и методы исследований. В работе использованы методы и алгоритмы интеллектуального анализа данных: обработки информации, математической статистики, алгоритмы отбора значимых признаков для моделей машинного обучения, алгоритмы машинного обучения с учителем, методы разработки программного приложения.

Теоретическая значимость заключается в разработке методов и алгоритмов интеллектуальной поддержки при принятии управленческих решений для прогнозирования здоровья ребенка, зачатого при помощи ВРТ. Методы и алгоритмы могут применяться для моделирования медицинских задач, которые формализуются в виде многоэтапных процессов.

Практическая значимость результатов работы заключается в разработке системы поддержки принятия врачебных решений и в создании программного приложения, готового к использованию в качестве вспомогательного инструмента принятия решений для сотрудников и пациентов центров ВРТ.

Основные положения, выносимые на защиту.

1. Предложенный метод обработки разнородных данных на основе алгоритмов токенизации коллекций текстовых медицинских документов, с использованием морфологического анализатора, исправления ошибок и кластеризации данных может использоваться для решения задач создания признакового описания объектов данных по текстовому описанию.
2. Разработанная структура системы поддержки принятия врачебных решений многоэтапного прогнозирования результатов использования ВРТ позволяет сделать прогноз группы здоровья ребенка с указанием диагнозов, по которым возможно возникновение осложнений. Струк-

тура предоставляет возможность уточнять прогноз, дополняя его новыми признаками, полученными с каждым следующим этапом протокола ВРТ.

3. Разработанное математическое и алгоритмическое обеспечение СППВР, состоящее из 63 моделей машинного обучения, позволяет осуществлять прогнозирование с уровнем точности, достаточным для клинического применения, а именно по результатам апробации средняя доля правильных ответов для прогнозируемых исходов составляет 84,5%, при этом 95% доверительный интервал – [78,7; 90,3].

Достоверность полученных в работе результатов обеспечивается корректным применением математических методов, подтверждается успешно проведенной апробацией разработанного математического и алгоритмического обеспечения системы поддержки принятия врачебных решений.

Внедрение результатов диссертационного исследования. Результаты диссертационной работы используются в АО «Центр семейной медицины», г. Екатеринбург, Россия (акт об использовании результатов от 11.04.2022).

Апробация работы. Основные положения и результаты работы докладывались и обсуждались на конференциях: VII Международная молодежная научная конференция «Физика. Технологии. Инновации ФТИ-2020» Екатеринбург, Россия, 18–22 мая 2020; VIII Международная молодежная научная конференция «Физика. Технологии. Инновации ФТИ-2021» Екатеринбург, Россия, 17–21 мая 2021.

Диссертационное исследование выполнено при финансовой поддержке гранта РФФИ № 19-37-90121 «Экспертная система прогнозирования здоровья детей, рожденных при помощи ВРТ».

Личный вклад. Автор провел анализ состояния предметной области, обосновал используемые методы и материалы исследования, выделил значимые признаки для машинного обучения из предоставленной медицинской документации, разработал алгоритм для составления признакового описания пациентов, создал структуру системы поддержки принятия врачебных решений,

построил модели машинного обучения, разработал интерфейс программного приложения и его функционал, создал пользовательское приложение для работы с созданной системой.

Публикации. По теме диссертации опубликовано 10 научных работ, в том числе 5 статей в рецензируемых научных изданиях, определенных ВАК РФ и Аттестационным советом УрФУ, из них 1 статья в изданиях, индексируемых в международных цитатно-аналитических базах WoS и Scopus. Получено 1 свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем работы. Диссертационная работа состоит из введения, трех глав, заключения, списка литературы, шести приложений. Работа изложена на 162 страницах, содержит 37 рисунков и 15 таблиц.

1. АНАЛИЗ СОСТОЯНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ.

ПОСТАНОВКА ЦЕЛИ И ЗАДАЧ ИССЛЕДОВАНИЯ

В терминах анализа данных запрос врачей можно формализовать как задачу создания набора моделей прогнозирования, созданных на основе алгоритмов машинного обучения (англ. machine learning, ML) с учителем, в которой объектами данных являются женщины и рожденные ими дети. Необходимо по признаковому описанию объектов данных и имеющимся исходам построить прогностические модели. Для каждого исхода необходимо предсказать вероятность его наступления (например, возникновения беременности) или оценить его числовую характеристику (например, срок родов в неделях, число прижившихся эмбрионов). То есть решить задачу классификации или регрессии.

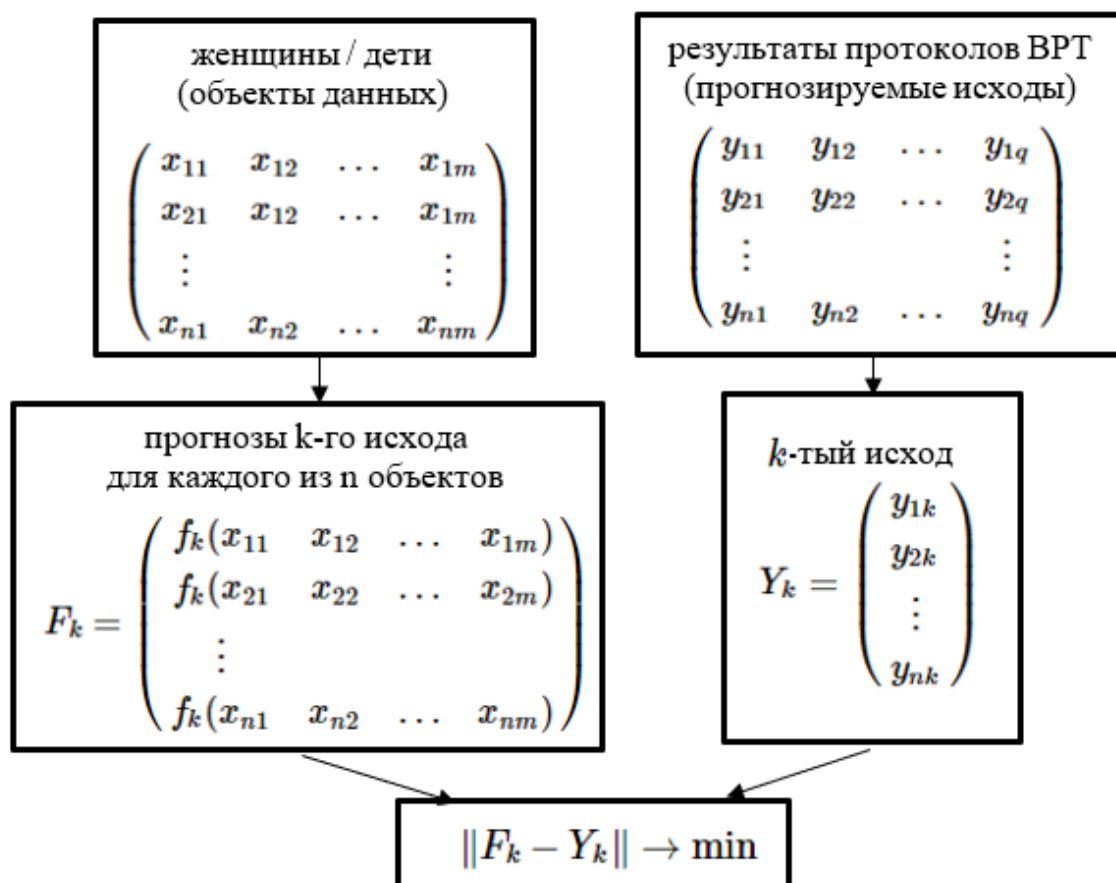


Рисунок 1.1 Описание задачи в терминах анализа данных

Описание задачи в терминах анализа данных представлено на рисунке 1.1. Пусть есть n пациентов. Каждый описывается m признаками. Для каждого пациента существует q исходов. Обозначим через $x_{ij}, i = 1, \dots, n, j = 1, \dots, m$, – значение j -го признака у i -го пациента, через $y_{ik}, i = 1, \dots, n, k = 1, \dots, q$, – значение k -го исхода у i -го пациента. Обозначим через F_k n -мерный вектор, координаты которого содержат прогнозы k -го исхода по всем n объектам, Y_k – значения k -го признака для всех n объектов. Требуется построить для всех исходов разрешающие правила f_k из некоторого класса функций или алгоритмов, минимизирующие тот или иной функционал невязки.

В случае, когда объекты данных – женщины, каждый признак представляет собой данные из их медицинской документации, существенно влияющие на возможность зачатия и успешность вынашивания, а также, характеризующие течение беременности и ее результат. В случае, когда объекты данных – рожденные этими женщинами дети, их признаковое описание помимо признакового описания матери включает данные, характеризующие здоровье ребенка. На каждом этапе нужно предсказать вероятность наступления исходов или оценить числовую характеристику. То есть решить задачу классификации или регрессии.

1.1 Вспомогательные репродуктивные технологии: основные определения и методы, рассмотренные для создания системы

В качестве программы ВРТ как медицинской услуги наибольшую популярность имеет экстракорпоральное оплодотворение (ЭКО), которое может быть осуществлено с помощью традиционной инсеминации (принято называть термином ЭКО) или с помощью интрацитоплазматической инъекции сперматозоида в цитоплазму яйцеклетки (ИКСИ). Обе программы зачастую называют термином ЭКО, но указывают способ оплодотворения: только ЭКО для всех ооцитов; частично ЭКО и ИКСИ; только ИКСИ.

Этапы проведения данных программ ВРТ схожи и состоят из [33]:

- 1) отбора пациентов;

- 2) овариальной стимуляции;
- 3) пункции фолликулов яичника;
- 4) переноса эмбрионов в полость матки;
- 5) поддержки лютеиновой фазы;
- 6) диагностики беременности.

Отбор пациентов включает в себя сбор информации о показаниях к проведению определенного протокола ВРТ, сбор анамнеза жизни и здоровья, проведение первичного осмотра и обследований, предусмотренных клиническими рекомендациями [33] в том числе для исключения наличия возможных противопоказаний к процедуре.

Стимуляция яичников – стимуляция с помощью медицинских препаратов одномоментного развития и созревания набора фолликулов с целью получения нескольких/ множества ооцитов при их пункции [33]. Она может проводиться по короткому, длинному протоколу или протоколу с естественным циклом (самостоятельная овуляция женщины). Для каждой пациентки предусмотрен индивидуальный подход, учитывающий данные анамнеза и здоровья, позволяющий осуществить выбор предварительного лечения (претритмент), выбор длительности протокола и триггера овуляции.

Пункция фолликулов яичников – это процесс получения ооцитов для дальнейшего оплодотворения. После пункции производится оценка качества ооцитов. Отдельным этапом можно выделить процесс оплодотворения ооцитов, который и отличает ЭКО от ИКСИ. ЭКО предполагает самостоятельное оплодотворение яйцеклетки наиболее сильным сперматозоидом без постороннего вмешательства, в то время как при ИКСИ отобранный сперматозоид вводится принудительно в яйцеклетку эмбриологом. Оба типа оплодотворения выполняются в чашке Петри с использованием различных сред, которые делят на среду – жидкость и среду – масло. Также возможно добавление донорских ооцитов или криоконсервированных ооцитов пациентки, полученных в результате предыдущих стимуляций.

После оценки качества эмбрионов, выполняется их перенос в полость матки пациентки на второй или третий день после пункции фолликулов на четырех–восьми клеточной стадии развития. Продление срока культивирования до пяти-шести дней приводит к селекции эмбрионов: часть останавливается в развитии, другая - достигает стадии бластоцисты. Перенос на стадии бластоцисты позволяют выбрать наиболее жизнеспособных, снизить общее число переносимых и замораживаемых эмбрионов и повысить вероятность имплантации [34]. Выбор дня переноса определяется в зависимости от числа, качества и характера дробления эмбрионов. Решение принимают лечащий врач и эмбриолог, обсуждается с пациенткой [35]. На практике допускается перенос не более двух эмбрионов, что обосновано стремлением к снижению вероятности наступления многоплодной беременности.

Далее продолжается гормональная поддержка лютеиновой фазы и выполняется последующая диагностика наступления беременности.

Таким образом, процесс получения беременности с помощью ЭКО / ИКСИ можно описать с помощью четырех основных этапов, имеющих вариативность течения, и двух заключительных, выполняющихся единообразно (рис. 1.2). Имеет смысл предсказывать здоровье планируемого ребенка на каждом основном этапе из-за большого числа факторов, способных повлиять на результат.

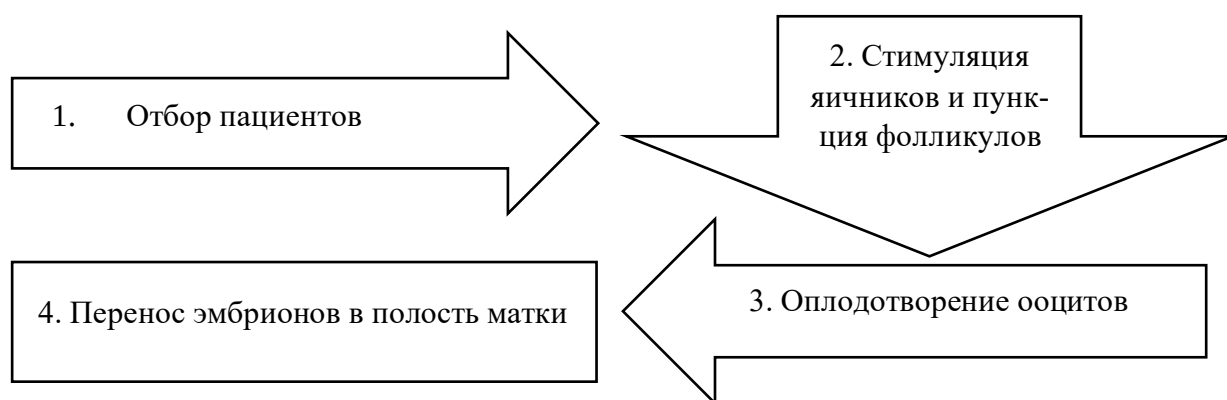


Рисунок 1.2 – этапы проведения протоколов ЭКО (ИКСИ), после которых возможна переоценка прогноза здоровья планируемого ребенка

В диссертационной работе рассматриваются также данные программы криопереноса эмбрионов. Программа исключает этапы стимуляции и оплодотворения, выполняется только перенос эмбрионов с предварительной разморозкой (рис. 1.3) и завершающие два этапа: поддержка лютеиновой фазы, диагностика беременности.

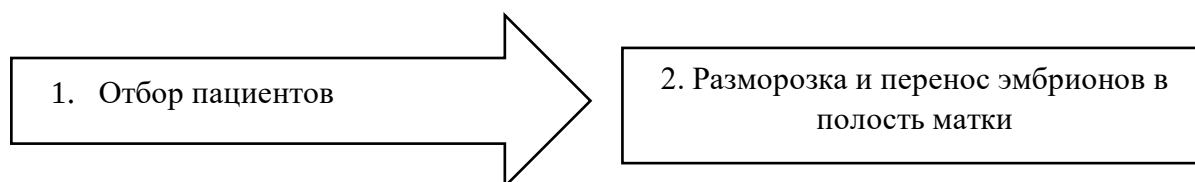


Рисунок 1.3 – этапы проведения протоколов криопереноса, после которых возможна переоценка прогноза здоровья планируемого ребенка

Далее в главе проведен анализ исследований по поиску и обоснованию факторов, влияющих на исход ВРТ и здоровье рожденного ребенка, и обзор имеющихся на сегодняшний день наработок в области прогнозирования результатов ВРТ.

1.2 Сравнительный анализ существующих работ, посвященных прогнозированию результатов протокола ВРТ

Большое число имеющихся факторов, наличие различных методик ВРТ и медикаментозного сопровождения, отсутствие систематических данных об исходах беременностей и рожденных младенцах создают сложности в выделении единого набора предикторов, оказывающих влияние на результаты протоколов и здоровье ребенка. Поиск и обоснование набора значимых признаков ведут ученые [20,21,25,29,36,37,38,39,40,41,42,43,44,45], используя при этом как методы статистического анализа медицинских данных, так и методы искусственного интеллекта.

Чаще всего работы по анализу клинических исходов посвящены поиску ответа на наличие влияния какого-либо одного фактора или группы смежных факторов на результат применения определенного метода ВРТ [36,37,38]. Однозначно сказать, какой способ отбора значимых признаков дает самые лучшие

результаты, нельзя, но работы, в которых сравниваются различные подходы к отбору значимых признаков, редки.

В исследовании по определению предикторов наступления беременности у бесплодных пар [39] для отбора наиболее значимых признаков был использован метод пошагового обратного отбора, при котором из набора факторов удаляли по одному наименее значимому фактору.

Для поиска значимых факторов при прогнозировании неблагоприятного акушерского исхода при беременности ЭКО в исследовании [40] был проведен статистический анализ с использованием критерия хи-квадрат и t-критерия Стьюдента. Многовариантный логистический регрессионный анализ использовался для проверки связи между акушерскими осложнениями и характеристиками матери, построения предсказательных номограмм.

В исследовании [21] по выявлению предикторов рождения живого ребенка при проведении процедуры ЭКО/ИКСИ для отбора факторов использовалась оценка их корреляции с значением выходной переменной (фактом рождения живого ребенка). Для оценки корреляции факторов с исходом использовался метод обобщенных оценочных уравнений. Для исключения переобучения модели на втором этапе исследования были удалены коллинеарные факторы, имеющие корреляцию между собой выше 0,8. В конечном итоге из 100 факторов 36 были признаны пригодными для предсказания, но лучшая модель содержала только 7 из них.

Есть ряд факторов, которые выделяются исследователями и применяются ими при построении прогностических моделей [20,21,25,41,42,43]. Самыми часто встречающимися факторами из анамнеза женщины являются: возраст [20,21,25,42,43], ИМТ [41,43], причина бесплодия [21,43], некоторые данные акушерского анамнеза [25,42,43]. Реже исследуют влияние следующих факторов: тип бесплодия [25,42], курение [41,46], данные первого протокола [21] и число неудачных попыток ВРТ [20,21,42]. В некоторых работах рассматривается влияние социальных факторов и образа жизни пациентки.

Безусловно, некоторые факторы, отражающие материнское здоровье и образ жизни имеют влияние на течение беременности без привязки к способу ее получения, а значит могут влиять как на ребенка, рожденного в результате применения ВРТ, так и на зачатого естественным путем. Например, в работе [45] говорится о том, что материнское ожирение связано с риском ожирения, ишемической болезни сердца, инсульта, диабета 2 типа и астмы у ее потомства, может привести к ухудшению когнитивных функций и повышению риска нарушений развития нервной системы, включая церебральный паралич. А авторы статьи [46] заключают, что материнское курение влияет на функцию легких у рожденных детей.

Результатом представленного обзора работ является вывод о том, что нет единого принятого алгоритма для поиска предикторов в задачах прогнозирования исходов ВРТ или здоровья ребенка. Также нет единого набора значимых предикторов, оказывающих влияние на исход беременности и здоровье ребенка, рожденного в результате применения ВРТ нет, но проводятся многочисленные медицинские и аналитические исследования в данной области.

В настоящий момент большинство исследований в области прогнозирования результатов использования ВРТ ориентированы на анализ и предсказание наступления зачатия или живорождения.

Например, в [19] построена ML-модель прогнозирования живорождения, основанная на алгоритме XGBoost в результате анализа протоколов ЭКО включая ИКСИ, с учетом всех циклов заморозки-разморозки, проведенных после одной стимуляции яичников. Значимые факторы были отобраны экспертами из более чем ста признаков, отбор производился на основе знаний предмета, патофизиологических механизмов или результатов предыдущих исследований и клинических руководств. Выборка составила 7188 женщин, из которых живорождение наблюдалось у 2797 (39%) человек. Модель может использоваться только для пар, которые никогда прежде не проходили через ЭКО, что ограничивает применимость модели. Не был учтен семейный генетический анамнез и факторы образа жизни из-за ограниченности набора данных.

Цель работы [20] – прогнозирование возможности живорождения в последующих циклах. Материалом послужили данные о 151 женщине, у которой было не менее пяти последовательных циклов ЭКО, из которых первые четыре оказались неудачными. Отбор предикторов выполнялся с использованием статистических методов и данных клинических исследований. Была построена логистическая модель с пошаговой стратегией, основанной на минимизации информационного критерия Акаике, пошаговый подход к отбору признаков был доработан экспертами: выбор предикторов обсуждался как с клинической, так и со статистической (р-значения и значения критерия Акаике) точки зрения. Прогнозирование построено на основе логистической регрессии.

Другие прогнозирующие живорождение ML-модели, например, [21,22,23] так же строятся на разных предикторах.

В статье [21] для построения модели использовалась информация о 5722 протоколах ЭКО/ИКСИ, проведенных с 1999 по 2014 года. Модель была разработана с использованием многомерной логистической регрессии. Исследование проведено только для технологий ЭКО / ИКСИ и переноса одного эмбриона.

Авторы работы [23] сравнили различные алгоритмы искусственного интеллекта, включая классическое машинное обучение (логистическая регрессия, метод k-ближайший соседей, многослойный персептрон, дерево решений), глубокое обучение (одномерная нейронная сеть с сигмовидным нейроном активации на выходном слое) и ансамблевые методы (алгоритм случайного леса, AdaBoost) для построения своей модели прогнозирования живорождения. Модель случайного леса имела лучшее значение критерия F-меры без проведения отбора признаков. Используемый набор данных содержал 141 160 запись о проведенных протоколах ЭКО и 25 признаков, после их приведения к бинарному или категориальному виду. При построении использовались как все признаки, так и отобранные самими алгоритмами. Работа показывает успешную применимость и разнообразие инструментов и методов искусственного интеллекта к построению прогностических моделей. Цель дальнейшей работы авторов – со-

брать данные из разных источников и использовать более обширный набор признаков для построения ML-модели, которую целесообразно будет применять на практике.

Истмико-цервикальная недостаточность является одной из частых причин поздних выкидышей или преждевременных родов. Прогнозированию возникновения данной патологии беременности посвящена модель, описанная в работе [24]. Исследование включало 4710 пациентов, прошедших ЭКО/ИКСИ в 2011 по 2018 года, случайным образом разделенных для построения и тестирования номограммы.

В статье [25] представлены алгоритмы прогнозирования исходов ЭКО и ЭКО/ИКСИ у супружеских пар при некоторых формах бесплодия. Все четыре алгоритма построены на основе бинарной логистической регрессии по данным 256 пациентов, в качестве значимых характеристик выбраны признаки из анамнеза и данные стимуляции. Успешность рассматривается только осуществлением переноса, но не проводится связь с наступлением беременности и ее результатом, со здоровьем ребенка.

В статье [25] представлена прогностическая ML-модель для оценки количества зрелых ооцитов, необходимых для получения хотя бы одного эуплоидного эмбриона для переноса при проведении ЭКО / ИКСИ. Использовалось отрицательное биномиальное распределение для моделирования количества эуплоидных бластоцист и метод LASSO для выбора переменных. Для построения предсказания применена логистическая регрессия. В исследование было включено 347 пациентов.

В работе [27] проведен обзор качества ML-моделей клинического прогнозирования ЭКО, который включает 35 различных моделей, построенных с 1989 по 2018 года, каждая из которых основана на различном числе проведенных протоколов: от 141 цикла ВРТ до выборки из 288161 женщины, каждая из которых прошла от одного до трех протоколов. Количество предикторов варьируется от 3 до 14 (медиана распределения – 7). В общей сложности более 50 раз

личных предикторов были включены в разработанные модели во всех исследованиях этого обзора. Женский возраст и продолжительность бесплодия являются двумя ключевыми факторами, используемыми для прогнозирования шансов на беременность и живорождение после ЭКО. Восемь моделей прогнозируют живорождение, другие сосредоточены на различных типах исходов беременности, например, наступление или продолжение беременности при определенных условиях. Большинство моделей предсказывают успех ЭКО при первом свежем переносе, а для разработки моделей чаще всего использовалась логистическая регрессия. В обзоре нет ни одной модели, прогнозирующей здоровье рожденного ребенка.

Работа [28] посвящена не только предсказанию рождения ребенка, но и возникновению преждевременных родов и низкой массы тела при рождении, что безусловно влияет на дальнейшее здоровье детей, но взаимосвязи с появлением возможных заболеваний и протоколом, течением беременности не проводится.

Рассмотрением здоровья ВРТ детей зачастую занимаются другие исследователи, чаще делающие выводы исходя из баз данных детских оздоровительных учреждений и больниц. Такие работы не акцентируют внимание на семейном анамнезе ребенка и примененной технологии ВРТ, больше являются аналитическими, ищущими взаимосвязи и значимые факторы, влияющие на здоровье, без построения прогнозирующих систем. Нередко выполняется сравнение детей ВРТ с детьми, рожденными естественным путем [47, 48, 49].

В статье [29] представлен разработанный алгоритм прогнозирования частых нарушений здоровья у детей первого года жизни, родившихся от одноплодной беременности после ЭКО, включающий выявление факторов риска с расчетом прогностических коэффициентов. Все выводы получены в результате проведения ретроспективного сплошного обсервационного исследования 210 детей раннего неонатального периода, родившихся в результате одноплодной беременности, полученной при помощи ЭКО в 2012 – 2013 годах. И с помощью групп сравнения детей ЭКО и зачатых естественным путем (по 121 ребенку в

каждой группе), за которыми было проведено наблюдение в течение первого года жизни.

Таким образом, большинство работ в области моделирования исходов ВРТ позволяют на основе статистических методов и алгоритмов машинного обучения спрогнозировать результат процедуры в виде наступления беременности или живорождения. Не определен общепринятый набор факторов, влияющих на здоровье ребенка и исход ВРТ. Систем, прогнозирующих долгосрочное здоровье ребенка, зачатого при помощи ВРТ нет, но есть работы, подтверждающие актуальность разработки таких систем.

1.3 Описание наборов данных, переданных для разработки

Набор данных АО «ЦСМ» г. Екатеринбург

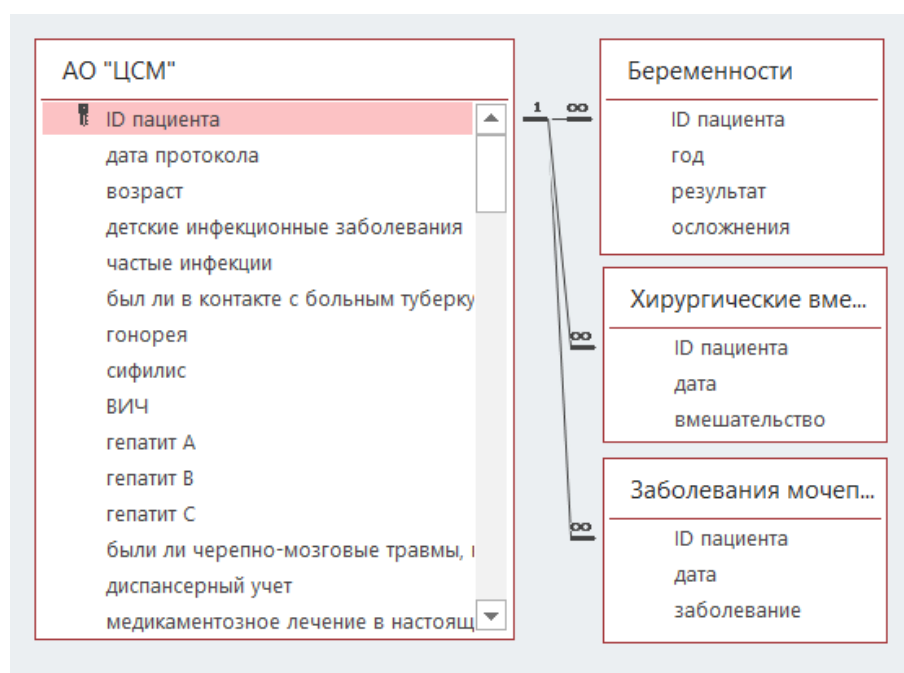


Рисунок 1.4 – Схема данных АО «ЦСМ»

Для исследования был получен набор из 8500 оригинальных записей о всех проведенных протоколах ВРТ с 2014 – 2018 гг. в клинике АО «Центр семейной медицины» (АО «ЦСМ») [50] г. Екатеринбурга с сохранением персональных данных. Данные находятся в электронных таблицах формата .xlsx, распределены на несколько таблиц, поскольку выгружены из реляционной базы данных

и имеют связь «один ко многим», где «Общая таблица» является родительской, остальные таблицы – дочерними, а ключом – ID пациента (рис. 1.4). Поля всех таблиц перечислены в таблице 1.

Таблица 1 – поля таблиц данных АО «ЦСМ»

общая таблица			
1	ID пациента	42	год вступления в брак
2	дата протокола	43	методы контрацепции
3	возраст	44	фолликулостимулирующий гормон
	анамнез женщины	45	пролактин
4	детские инфекционные заболевания		данные протокола ВРТ
5	частые инфекции	46	программа
6	был ли в контакте с больным туберкулезом	47	индекс массы тела
7	гонорея	48	номер попытки
8	сифилис	49	претритмент
9	вирус иммунодефицита человека	50	протокол
10	гепатит А	51	стартовая доза фолликулостимулирующего гормона
11	гепатит В	52	число дней стимуляции
12	гепатит С	53	эстрадиол перед пункцией
13	были ли черепно-мозговые травмы, в том числе с потерей сознания	54	прогестерон перед пункцией
14	диспансерный учет	55	триггер
15	медикаментозное лечение в настоящее время	56	толщина эндометрия
	наследственный анамнез	57	количество доминантных фолликулов
16	хронические заболевания	58	количество комплексов
17	гипертоническая болезнь	59	количество ооцитов всего
18	ишемическая болезнь сердца	60	количество зрелых ооцитов
19	инсульт	61	качество ооцитов
20	аллергические болезни	62	получение спермы
21	язвенная болезнь	63	качество спермы
22	желче-каменная болезнь	64	донор спермы
23	онкологические заболевания (с локализацией)	65	тип оплодотворения ИКСИ
24	ожирение	66	тип оплодотворения ЭКО
25	сахарный диабет	67	среды оплодотворения

26	тромбозы, варикозное расширение вен	68	количество зигот
27	психические расстройства	69	сутки 3 (ЭКО)
	аллергоанамнез	70	сутки 3 (ИКСИ)
28	на лекарства	71	сутки 5 УТРО (ЭКО)
29	прочее	72	сутки 5 УТРО (ИКСИ)
	профессиональная биография	73	сутки 6 (ЭКО)
30	особенности профессии	74	сутки 6 (ИКСИ)
31	имелся ли контакт с профессиональными вредными факторами: химическими, радиоактивными и т. д.	75	преимплантационная генетическая диагностика
32	другие профессиональные вредности	76	количество бластоцист хорошего качества на пятые сутки
	сведения о вредных привычках	77	количество бластоцист хорошего качества на шестые сутки
33	употребление алкоголя	78	общее количество бластоцист на шестые сутки
34	наличие эпизодов внутривенного употребления наркотиков	79	среда заморозки
35	табакокурение стаж	80	среда разморозки
	менструальная функция	81	номер соломины
36	возраст менархе	82	количество эмбрионов (перенос)
37	регулярность менструаций	83	качество эмбрионов
38	продолжительность цикла (дней)	84	среда для переноса
39	менструальная кровопотеря	85	эстрогены
40	нарушения цикла	86	гестагены
	репродуктивная функция	87	низкомолекулярный гепарин
41	возраст начала половой жизни	88	прегнил
хирургические вмешательства			
1	ID пациента	3	вмешательство
2	дата		
беременности			
1	ID пациента	3	результат
2	год	4	осложнения
заболевания мочеполовой системы			
1	ID пациента	3	заболевание
2	дата		

1. «общая таблица» содержит информацию о собранном анамнезе пациенток и данные проведенных протоколов (8500 записей). Все поля, кроме первых двух, содержат данные произвольного текстового формата. Поле «дата протокола» – тип дата в формате «ЧЧ.ММ.ГГГГ».

2. «хирургические вмешательства» – дополняет анамнез, содержит информацию о перенесенных хирургических вмешательствах пациентками в течение жизни (17608 записей). Поле «дата» – тип дата в формате «ЧЧ.ММ.ГГГГ», поле «вмешательство» – произвольный текст.

3. «беременности» – дополняет анамнез, содержит информацию о предыдущих беременностях пациенток (6466 записей). Поле «год» - тип дата в формате «ГГГГ», поля «результат» и «осложнение» - произвольный текст.

4. «заболевания мочеполовой системы» – дополняет анамнез, содержит информацию о перенесенных в течение жизни и хронических заболеваниях мочеполовой системы пациенток (1791 запись). Поле «дата» - тип дата в формате «ЧЧ.ММ.ГГГГ», поле «заболевание» - текст из выпадающего списка.

Набор данных «Программа мониторинга беременных» АС «РАМ. Версия 1.2.» ГАУЗ СО «ОДКБ»

Автоматизированная система «Региональный акушерский мониторинг» ГАУЗ СО «ОДКБ» Версия 1.2.» (АС «РАМ») содержит информацию обо всех беременных женщинах Свердловской области, в том числе и о родильницах, беременность которых была получена при помощи ВРТ. Первая информация о случае беременности вносится при постановке на учет, последняя – вносится после родов.

Из данной автоматизированной системы с помощью SQL-запросов был выгружен набор данных из 821 оригинальной записи о пациентках, беременность которых наступила в результате проведения ВРТ в АО «ЦСМ». Выгруженные данные находятся в электронной таблице формата .xlsx и содержат следующие поля:

1) ID пациента (в соответствии с данными АО «ЦСМ»);

- 2) ID первого ребенка (при наличии);
- 3) ID второго ребенка (при наличии);
- 4) ID третьего ребенка (при наличии);
- 5) возраст (числовой тип);
- 6) исход беременности (текстовые значения: «роды» или «прерывание беременности»);
- 7) дата исхода (дата в формате ЧЧ.ММ.ГГГГ);
- 8) информация об исходе (текстовые значения, характеризующие: место родов («роды в стационаре», «прерывание беременности», «исход неизвестен», срок родов: «срочные», «преждевременные», «до 12 недель», «до 22 недель», способ родоразрешения: «самостоятельные», «кесарево сечение плановое», «кесарево сечение экстренное»);
- 9) информация о пациенте (текстовые данные, включающие в себя произвольную информацию по четырем разделам: акушерско-гинекологический анамнез, экстрагенитальные заболевания, осложнения текущей беременности со стороны матери, осложнения текущей беременности со стороны плода);
- 10) диагноз первого ребенка (текстовые данные, выбранные из списка диагнозов по международной классификации болезней (МКБ-10[51]). Могут быть пустые ячейки);
- 11) диагноз второго ребенка (текстовые данные, выбранные из списка диагнозов по МКБ-10. Могут быть пустые ячейки);
- 12) диагноз третьего ребенка (текстовые данные, выбранные из списка диагнозов по МКБ-10. Могут быть пустые ячейки).

Набор данных «Регистр детей, нуждающихся в ранней помощи» ГАУЗ СО «ОДКБ»

После выписки из родильного дома дети, зачатые с использованием ВРТ, подлежат мониторингу в информационной системе «Регистр детей, нуждаю-

щихся в ранней помощи» Областного центра ранней помощи ГАУЗ СО «Областная детская клиническая больница» (ГАУЗ СО «ОДКБ») [52], с транслированием информации в Областной центр превентивной педиатрии ГАУЗ СО «Многопрофильный медицинский центр «Бонум» (ГАУЗ СО «МКМЦ «Бонум») [53].

Для исследования был получен набор из 14211 оригинальной записи из ИС «Регистр детей, нуждающихся в ранней помощи» ГАУЗ СО «ОДКБ» о детях, рожденных с 2014 – 2019 гг. и зачатых при помощи ВРТ с сохранением персональных данных. Данные находятся в электронной таблице формата .xlsx и содержат следующие поля:

- 1) ID ребенка (в соответствии с данными АС «РАМ»);
- 2) дата рождения ребенка (в формате «ЧЧ.ММ.ГГГГ»);
- 3) год рождения (тип дата в формате «ГГГГ»);
- 4) срок гестации (числовой тип);
- 5) масса тела при рождении (числовой тип);
- 6) диагноз основной при выписке (произвольный текст);
- 7) диагноз сопутствующий при выписке (произвольный текст);
- 8) диагноз основной при первой явке (произвольный текст);
- 9) диагноз сопутствующий при первой явке (произвольный текст);
- 10) диагноз основной в 6 месяцев (произвольный текст);
- 11) диагноз сопутствующий в 6 месяцев (произвольный текст);
- 12) диагноз основной в 1 год (произвольный текст);
- 13) диагноз сопутствующий в 1 год (произвольный текст);
- 14) диагноз основной в 1,5 года (произвольный текст);
- 15) диагноз сопутствующий в 1,5 года (произвольный текст);
- 16) диагноз основной в 2 года (произвольный текст);
- 17) диагноз сопутствующий в 2 года (произвольный текст);
- 18) диагноз основной в 2,5 года (произвольный текст);
- 19) диагноз сопутствующий в 2,5 года (произвольный текст);
- 20) диагноз основной в 3 года (произвольный текст);

21) диагноз сопутствующий в 3 года (произвольный текст).

Набор данных «Электронная карта пациента» ГАУЗ СО «МКМЦ «Бонум»

Для исследования был получен набор из 9800 оригинальной записи электронных карт пациентов ГАУЗ СО «МКМЦ «Бонум» о детях, рожденных с 2014 – 2019 гг. и зачатых при помощи ВРТ с сохранением персональных данных. Данные находятся в электронной таблице формата .xlsx и содержат следующие поля:

- 1) ID ребенка (в соответствии с данными АС «РАМ»);
- 2) срок гестации (числовой тип);
- 3) масса при рождении (числовой тип);
- 4) дата рождения (в формате «ЧЧ.ММ.ГГГГ»);
- 5) диагнозы (текст из выпадающего списка, все диагнозы, имеющиеся в процессе наблюдения за ребенком до даты выгрузки информации).

Для возможности решения задачи прогнозирования здоровья ребенка, зачатого при помощи ВРТ, на основе полученных данных, необходимо выполнить значительную их предобработку и преодолеть следующие проблемы:

- 1) отсутствует интероперабельность между базами данных различных учреждений;
- 2) полученные данные недостаточно формализованы и стандартизированы.

Другими словами, до построения и обучения моделей машинного обучения (ML-моделей) необходимо разработать

- 1) алгоритмы объединения данных, полученных из разных источников;
- 2) алгоритмы, которые позволят составить признаковое описание объектов по объединенным данным.

1.4 Описание основных принципов работы проектируемой СППВР

После получения данных и обсуждения со специалистами в области акушерства и гинекологии, педиатрии, эмбриологии, было принято решение о формализации задачи следующим образом: необходимо построить СППВР, которая позволяет

- 1) до вступления в протокол ВРТ предсказывать вероятность наступления беременности, осложнения ее течения, срок и способ родов, здоровье рожденного ребенка до трех лет по данным анамнеза пациентки;
- 2) после вступления в протокол ВРТ осуществлять корректировку прогнозов с учетом дополнительных данных, полученных на различных этапах проведения протокола.

Система должна представлять собой набор моделей машинного обучения с учителем, построенных на имеющихся данных (описанных в разделе 1.3).

Здоровье ребенка описывается принадлежностью к определенной группе здоровья [54]. Группа здоровья присваивается на основании самого тяжелого диагноза из имеющихся у ребенка. Построение прогноза осуществляется до рождения ребенка, поэтому прогнозированию подлежат диагнозы, возникновение которых можно связать с наследственными факторами, информацией о беременности и родах матери. То есть для прогнозирования здоровья ребенка прежде всего необходимо оценить вероятность наступления беременности, далее спрогнозировать осложнения течения беременности и родов матери, затем с учетом осложнений спрогнозировать диагнозы ребенка и в итоге с учетом диагнозов прогнозировать группу здоровья ребенка (рис. 1.5).

Возникающая задача прогнозирования многоэтапна, каждая прогностическая модель должна быть привязана к определенному временному промежутку, а результат прогнозирования текущей ML-модели должен являться признаком для следующей ML-модели. Группа здоровья – единственный исход, который

не является признаком, поскольку это финальный результат предсказания. Таким образом прогнозируется комплекс параметров, определяющих здоровье ребенка.



Рисунок 1.5 – многоэтапный процесс прогнозирования комплекса параметров, определяющих здоровье ребенка

Для прогнозирования комплекса параметров, определяющих здоровье ребенка, зачатого при помощи ВРТ, необходимо учитывать данные проведенного протокола ВРТ. В разделе 1.1 было отмечено, что протокол ЭКО включает в себя четыре вариативных этапа проведения, а протокол криопереноса – два. СППВР должна позволять для каждого этапа протокола ВРТ задавать набор необходимых входных параметров и выводить после его выполнения прогноз комплекса параметров, определяющих здоровье ребенка и результаты этапа, поскольку результаты влияют на набор входных параметров для следующего этапа ВРТ. В соответствии с этим алгоритм, по которому должна будет работать разрабатываемая СППВР формализован в виде блок-схемы (рис. 1.6). В соответствии клиническими рекомендациями [33] проведения протоколов ВРТ в проектируемой системе было выделено четыре блока. Блоки 1 и 4 предусмотрены и для программы ЭКО и для криопереносов, а блоки 2 и 3 отражают этапы только программы ЭКО. Далее будет описан каждый блок системы в отдельности.

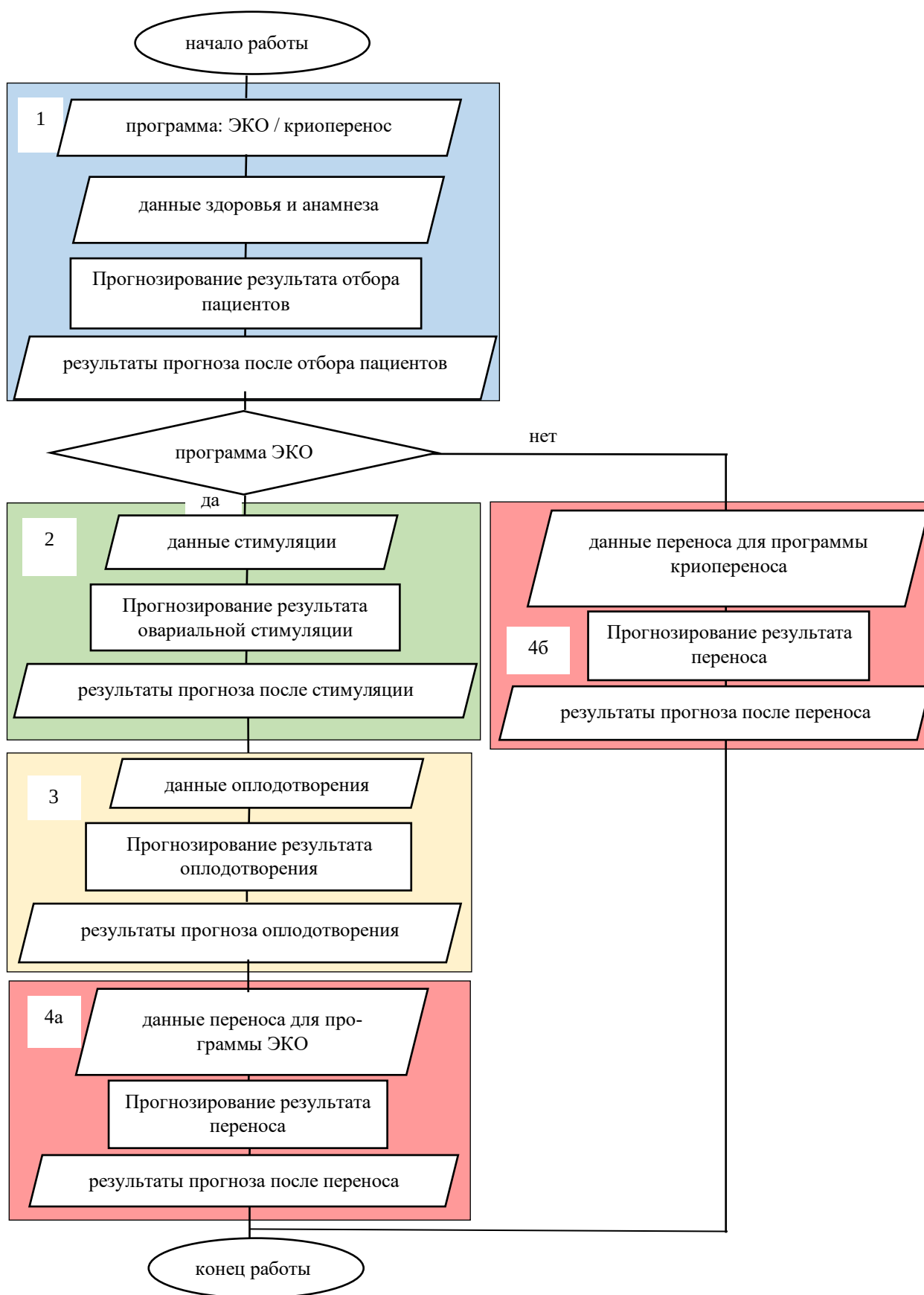


Рисунок 1.6 – Блок-схема логики работы СППВР

1. Отбор пациентов.

Входными данными для первого этапа являются данные жизни и здоровья пациентки, выбранная программа ВРТ (ЭКО или криоперенос). На данном этапе необходимо спрогнозировать комплекс параметров, определяющих здоровье ребенка, отдельно для многоплодной и одноплодной беременности.

2. Овариальная стимуляция.

На этапе стимуляции нужно, во-первых, скорректировать прогнозы, сделанные на предыдущем этапе (с учетом данных стимуляции). Во-вторых, спрогнозировать результаты стимуляции и пункции фолликулов: предполагаемое количество фолликулов, толщину эндометрия и качество ооцитов. Этап предусмотрен только для программы ЭКО.

3. Оплодотворение ооцитов.

Входными параметрами на третьем этапе должны быть данные, имеющиеся перед оплодотворением ооцитов, а результатом – прогнозирование качества яйцеклеток после оплодотворения. Этап предусмотрен только для программы ЭКО.

4. Перенос эмбрионов в полость матки.

На четвертом этапе нужно скорректировать прогноз комплекса параметров, определяющих здоровье ребенка, с учетом данных переноса эмбрионов в полость матки. Поскольку на этом этапе программы ЭКО и криопереноса отличаются набором параметров, необходимо создать возможность прогнозирования для обеих программ.

То есть ключевой особенностью разрабатываемой СППВР является то, что математическое и алгоритмическое обеспечение в ее основе – это не просто набор несвязанных классификаторов и регрессионных моделей, а иерархия моделей, позволяющих выполнять последовательное уточнение прогнозов на всех этапах ВРТ. Подобная структура будет позволять на основе прогнозов принимать управленческие решения, связанные с тактикой ведения пациентки, что способствует повышению эффективности проведения протоколов ВРТ и переходу к персонифицированной медицине.

После разработки системы предполагается создание пользовательского приложения для работы с ней.

1.5 Анализ алгоритмов отбора признаков для машинного обучения

Базы данных, полученные из информационных систем медицинских организаций, содержат большое количество признаков, многие из которых оказываются нерелевантными при прогнозировании конкретного исхода. В связи с этим, важной составляющей машинного обучения является отбор признаков (англ. *feature selection*), который либо является составной частью алгоритма построения моделей, либо предшествует ему, и в этом случае рассматривается как подготовительный этап.

Следует отметить, что задача отбора признаков существует в различных постановках. Во-первых, выделяют задачу поиска минимального множества признаков (англ. *minimal-optimal problem*), обеспечивающих наилучшее качество классификации или регрессии. Разумеется, качество классификации или регрессии должно быть формализовано в виде того или иного функционала качества, например, исправленного коэффициента детерминации (англ. *Adjusted R²*), информационного критерия Акаики или Байеса (англ. *AIC, BIC*), что в том числе приводит к разнообразию постановок и методов отбора. Во-вторых, рассматривается задача поиска всех релевантных признаков (англ. *all-relevant problem*), которая является более сложной [55].

Методы отбора признаков принято разделять на три класса в соответствии тем, как они комбинируются с алгоритмами построения моделей ML: фильтры, обертки и встроенные методы [56].

Методы-фильтры выбирают подмножество признаков до построения ML-модели, т. е. работают на этапе препроцессинга. Как следствие, фильтры, очевидно, выбирают признаки независимо от типа используемой ML-модели. Методы-фильтры вычисляют некоторую метрику для оценки силы и значимости влияния каждого отдельно взятого признака на прогнозируемую величину. В

качестве таких метрик наиболее часто используются взаимная информация, разница долей, коэффициент корреляции и др. [57, 58, 59].

Поскольку сложность методов-фильтров линейно зависит от числа признаков, они обычно не требовательны к вычислительным ресурсам. Фильтры хороши для задач высокой размерности, когда другие методы работают медленнее, они устойчивы к переобучению [60], их целесообразно применять как предобработку для вычислительно более затратных методов. Недостатком фильтров является то, что выбор подмножества признаков не учитывает специфику ML-модели. Кроме того, фильтры оценивают влияние признаков на прогнозируемую величину по отдельности. Между тем, известно, что линейная комбинация даже двух признаков, каждый из которых является слабым классификатором, может быть сильным классификатором; в многомерном случае эта проблема стоит еще более остро. Так же фильтры не учитывают взаимосвязи между самими признаками, что может приводить к проблеме мультиколлинеарности на этапе построения ML-модели.

Методы-обертки используют некоторую ML-модель как черный ящик, т. е. являются обертками над ней, что и обуславливает название данного класса методов. В самой общей формулировке обертки выбирают некоторое подмножество признаков, обучают на нем ML-модель и по качеству (например, предсказательной силе) полученной модели судят о важности использованных признаков. Итеративно перебирая различные подмножества признаков, обертки находят оптимальный набор.

Методы-обертки обычно позволяют находить наилучший набор признаков для конкретного типа ML-модели. Однако, обучение новой модели на каждом подмножестве признаков требует очень больших вычислительных ресурсов, поскольку задача поиска оптимального набора является NP-полной [61], и ее точное решение требует перебора 2^n подмножеств, где n – число признаков. В связи с этим применяются различные эвристические подходы к перебору подмножеств: метод ветвей и границ, имитации обжига, генетические алгоритмы [56]. Большинство методов-обертки построено на основе жадной стратегии

(например, пошаговая регрессия, RFE, ADD-DEL), что не гарантирует оптимальности решения, зато позволяет решать задачу за разумное время и избегать переобучения [62, 63, 64].

Встроенные методы производят отбор признаков в процессе обучения ML-модели. Классическим примером встроенного метода является Lasso регрессия, которая содержит в функционале качества штрафное слагаемое – L_1 норму вектора коэффициентов.

Хотя методы-обертки, использующие ML-алгоритмы как черный ящик, универсальны и просты в реализации, встроенные методы находят решения быстрее, избегая повторного обучения с нуля для каждого исследуемого подмножества признаков.

Рассмотрим отдельно Boruta – хорошо себя зарекомендовавший алгоритм-обертку над случайным лесом, предназначенный для поиска всех релевантных признаков [55].

Первая ключевая идея Boruta состоит в том, что решение о важности признака принимается не в результате сравнения признаков между собой (как делается в жадных реализациях оберток типа Forward LR или Stepwise regression), а в результате сравнения признаков с так называемыми «теневыми признаками». Именно, на очередной итерации Boruta создается копия исходного набора признаков, далее в каждом созданном столбце значения перемешиваются (англ. shuffle) для того, чтобы разрушить связи с целевой переменной, если таковые имелись; полученные таким образом признаки называются теневыми.

На полученном объединенном наборе, составленном из исходных и теневых признаков, обучается ML-модель и возвращает массив важности признаков (англ. feature importance). Те признаки, важность которых на данной итерации оказалась выше, чем максимум из важностей теневых признаков, получают отметку, так называемый хит (англ. hit).

Поскольку число хитов, рассчитанное после всех N итераций, является случайной величиной с биномиальным распределением с параметрами N и 0,5,

открывается возможность сделать вывод об итоговой важности признака, используя аппарат математической статистики. В этом состоит вторая ключевая особенность Boruta. Если число хитов больше, чем $B(N; 0,5; 1 - \alpha)$, где $B(N; 0,5; 1 - \alpha)$ – это $(1 - \alpha)$ -ая квантиль биномиального распределения с параметрами N и $0,5$, то считается, что признак оказывает важное влияние на значение целевой величины.

Отметим, что простая замена признака на случайный набор чисел, очевидно, разрушит связи с целевой переменной, если таковые имелись, однако приведет к существенным сдвигам и может внести систематические ошибки в обучающуюся модель и привести к неверным выводам. С другой стороны, перемешивание значений признака в пределах столбца связи разрушит, но к сдвигам не приведет, так как начальные и центральные моменты инвариантны относительно перемешиваний.

Отмеченные особенности стали причиной того, что Boruta в настоящее время стал одним из наиболее успешно используемых алгоритмов отбора признаков. В настоящей диссертации использована библиотека BorutaPy [65,66].

Отбор признаков, основанный на идее перемешивания признаков, реализован и в ряде других алгоритмов: PermutationImportance в библиотеке eli5 [67], PIMP [68].

Каждый алгоритм для отбора признаков имеет свои достоинства и недостатки и заслуживает внимания, поэтому, до принятия решения о выборе какого-то одного алгоритма, будет проведено опытное сравнение результатов применения всех алгоритмов к данным текущей задачи. Результаты этого исследования описаны в разделе 2.4.

1.6 Анализ методов, используемых для решения задачи машинного обучения с учителем

Формально задача прогнозирования исходов ВРТ является задачей машинного обучения с учителем, а именно задачей классификации или регрессии в зависимости от прогнозируемого исхода. Существуют различные алгоритмы

решения таких задач: построение регрессии, случайный лес, нейронные сети, дискриминантный анализ и т. д.

Априори сложно сказать, какой именно алгоритм машинного обучения с учителем позволит получить модель машинного обучения, обладающую наилучшей прогностической точностью. Имеет смысл использовать разные алгоритмы и по метрикам качества построенных моделей машинного обучения на тестовых выборках отобрать лучший алгоритм для решения конкретной задачи.

Рассмотрим некоторые наиболее популярные алгоритмы, использующиеся для решения задач классификации и регрессии при разнородных данных [69]. Результат сравнения их применения к решаемой задаче описан в разделе 2.4.

Множественная линейная регрессия

Рассматривается случайная величина Y линейно зависящая от m случайных величин $X_1, \dots, X_m$; при этом их распределения неизвестны, есть лишь набор из n наблюдений $(x_{11}, x_{12}, \dots, x_{1m}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{nm}, y_n)$. Предикторы $X_1, \dots, X_m$ и зависимая величина Y измеряются в интервальной шкале

Линейная модель, рассматриваемая здесь, имеет вид:

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \dots + w_m x_{im} + w_0 + e_i, i = 1, \dots, n,$$

где $w_0, w_1, \dots, w_m$ – весовые коэффициенты модели, e_i – случайная ошибка в i -ом наблюдении.

Требования к входным данным для достижения наилучших оценок параметров уравнения регрессии указаны в теореме Гаусса-Маркова [70]:

- 1) факторы и случайные ошибки – независимые случайные величины;
- 2) случайные ошибки модели гомоскедастичные, т. е. дисперсия ошибок постоянная, не зависит от значений предикторов;
- 3) отсутствует корреляция (автокорреляция) случайных ошибок разных наблюдений между собой: $cov(e_i, e_j) = 0, 1 \leq i < j \leq n$.

Итак, по набору из n наблюдений $(x_{i1}, x_{i2}, \dots, x_{im}, y_i), i = 1, \dots, n$, строим выборочное уравнение линейной регрессии $\hat{y} = w_1 x_1 + w_2 x_2 + \dots + w_m x_m +$

w_0 , где параметры $w_0, w_1, \dots, w_m$ будут соответственно выборочными оценками параметров регрессии.

Задача линейной регрессии заключается в нахождении таких $w_0, w_1, \dots, w_m$, что

$$\sum_{i=1}^n \|\hat{y}_i - y_i\| \rightarrow \min.$$

Коэффициенты w_i при признаках x_i легко интерпретируются: увеличение признака x_i на одну единицу своего измерения приводит к изменению зависимой величины на w_i своих единиц измерения.

На практике обычно встречаются модели множественной линейной регрессии, в которых число независимых переменных более двух. Данный тип моделей очень часто используется для создания прогностических моделей в медицине, например, в работах [71,72].

Многомерная логистическая регрессия

Использование линейной регрессии теряет смысл для решения задач классификации, поэтому при аналогичных требованиях к данным можно использовать алгоритм логистической регрессии. Значения предикторов измеряются в интервальной шкале.

Пусть p – это вероятность прогнозируемого события или принадлежности классу. Вероятность p оценивается как

$$p = \frac{e^z}{1 + e^z},$$

где $z = w_1x_1 + w_2x_2 + \dots + w_mx_m + w_0$ – «направленность» к событию.

Если для бинарной логистической регрессии метки классов $y_i, i = 0,1$, обозначить «1» и «-1», то задача оценки параметров логистической регрессии формализуется следующим образом:

$$\sum_{i=1}^n \log (1 + e^{-y_i z_i}) \rightarrow \min.$$

Если признаки мультиколлинеарны, то задача оценки параметров становится неустойчивой. Для того, чтобы этого избежать, в уравнение регрессии добавляют регуляризатор [72,73], ограничивающий величину коэффициентов при признаках. Это позволяет решить проблему переобучения модели. Чаще всего в качестве регуляризатора используют L_1 или L_2 нормы. Регуляризованная задача логистической регрессии с L_2 нормой имеет вид

$$\sum_{i=1}^n \log(1 + e^{-y_i z_i}) + \lambda \|w\|^2 \rightarrow \min, \quad \text{где } \|w\|^2 = \sum_{j=0}^m w_j^2,$$

где λ – параметр регуляризации.

Также как и линейная регрессия, логистическая регрессия часто используется для построения прогнозов в медицине, которые могут облегчить принятие решений, предоставляя индивидуальные оценки шансов, например, для прогнозирования вероятности развития силикоза у работников огнеупорного производства [74], отслойки плаценты [75]; хрупкости костей у детей с острым лимфобластным лейкозом [76]; тяжелой реакции на пероральный тест коровьего молока [77].

Если в данных имеются нелинейные зависимости, то целесообразно рассмотреть другие алгоритмы машинного обучения с учителем.

Деревья решений

Деревья решений – универсальные аппроксиматоры, позволяющие приближать сколь угодно сложные нелинейные зависимости в классе кусочно-постоянных функций. Деревья позволяют решать как задачи классификации, так и задачи регрессии [78].

Дерево состоит из узлов, содержащих предикаты – логические функции, истинность которых зависит от подставляемых в них объектов, – и листьев, дающих ответ. Например, в задачах классификации ответ – это решение о принадлежности объекта классу.

Алгоритм построения дерева есть, по сути, не что иное как поиск предикатов, последовательно разбивающих обучающее множество на подмножества,

каждое из которых содержит «более однородные» объекты, чем родительский узел. Так в корне содержатся все объекты обучающей выборки; те объекты, для которых корневой предикат истинный, относятся к левому потомку, а те, для которых ложный, – к правому. И так далее вплоть до листьев.

В задачах классификации понятие «однородности объектов» формализуется с использованием тех или иных функционалов или критериев, например, энтропийных. Так если все объекты, попавшие в лист, принадлежат одному классу, то энтропия равна нулю. И наоборот, если объекты узла равномерно распределены по классам, то энтропия равна единице. Большое распространение на практике получили критерии Джини

$$\sum_{k=1}^K p_k(1 - p_k),$$

где K – число классов в задаче, p_k — доля объектов класса k в текущей вершине X_m .

Увеличение глубины деревьев, означающее использование большего числа предикатов, очевидно позволяет более точно подогнать структуру дерева по обучающую выборку, что в свою очередь может позволить дать и более точные прогнозы. Однако, на практике увеличение глубины деревьев как правило приводит к переобучению – идеальному приближению обучающей выборки при низкой обобщающей способности модели. Для борьбы с переобучением используется регуляризация, которая заключается в задании одного или нескольких критериев, останавливающих процесс ветвления, тем самым делая дерево как можно более простым. К таким критериям относятся

- 1) ограничение по максимальной глубине дерева;
- 2) ограничение на минимальное количество объектов в листе;
- 3) ограничение на максимальное количество листьев в дереве;
- 4) требование, чтобы функционал качества при делении текущей подвыборки на две улучшался не менее чем на заданный процент.

Деревья использованы, например, для создания правил принятия решений при оценке риска детской астмы [79], где признаки риска астмы связаны с ее развитием нелинейно. Другие примеры использования деревьев решений в медицине: модель прогнозирования успешности грудного вскармливания [80], алгоритм принятия решений для отбора детей с тяжелой черепно-мозговой травмой для нейрохирургического вмешательства [81], модель прогнозирования развития сахарного диабета [82]. Авторы исследования [82] призывают медицинских работников использовать деревья решений для классификации и прогнозирования, чтобы улучшить клинические результаты.

С помощью деревьев решений можно создать понятные и простые для интерпретации модели, но из-за их склонности к переобучению лучшим вариантов решения задачи может стать использование композиции (ансамбля) деревьев решений [81].

Ошибка прогнозирования складывается из шума, смещения и разброса [83]. Величину шума изменить невозможно, смещение ансамбля моделей равно среднему смещению моделей в его составе. Разброс композиции из k некоррелированных моделей в k раз меньше разброса отдельной модели, что обосновывает преимущество использования ансамбля деревьев решений.

Случайный лес

Случайный лес – алгоритм, использующий ансамбль деревьев решений. Для решения задачи независимо друг от друга строится заданное количество решающих деревьев, при этом для построения каждого дерева формируется бутстреп выборка исходных данных, а для разбиения каждого узла отбирается случайное подмножество признаков. Благодаря такому подходу, все деревья в случайном лесе отличаются друг от друга. В задачах регрессии результат прогнозирования определяется как среднее ответов всех деревьев, входящих в лес; а задачах классификации – как наиболее частый класс [84,85,86].

Наборы предикторов для построения каждого дерева решений и число деревьев в лесу необходимо выбирать таким образом, чтобы решающие деревья оставались непохожими друг на друга. При этом глубину каждого решающего

дерева рекомендуется задавать максимальной для уменьшения смещения в ошибке модели.

Модели случайного леса могут стать лучшим вариантом решения задач при небольшом наборе данных, имеющих сложные зависимости, и при предсказании событий низкой вероятности возникновения, например [87]. Для табличных данных модели на основе случайного леса показывают лучший результат прогнозирования, чем логистическая регрессия, метод опорных векторов, наивный байесовский алгоритм, XGBoost, KNN-алгоритм, многослойный персептрон и нейронные сети [88,89,90].

Подход на основе композиции деревьев устраняет основной недостаток решающих деревьев – переобучение, но к недостаткам случайного леса можно отнести их требование к большому объему вычислительных ресурсов для построения глубоких деревьев и большого количества деревьев для решения сложных задач при неспособности неглубоких деревьев обнаруживать сложные закономерности в данных. Данные проблемы решает градиентный бустинг.

Бустинг

Алгоритмы на основе бустинга, например, градиентный бустинг [91] – это варианты ансамблевых алгоритмов на основе деревьев решений. Они более чувствительны к настройке параметров, требуют больших объемов данных, чем случайный лес, но могут дать более высокое качество, чем логистическая регрессия, случайный лес или многослойные нейронные сети, например в работах [92,94,95].

Отличие бустинга от случайного леса в том, что базовые модели строятся последовательно, при этом каждая следующая модель пытается уменьшить ошибку (смещение) всех предыдущих моделей. В связи с этим деревья решений должны обладать небольшим разбросом и могут обладать высоким смещением, а следовательно, глубина деревьев решений должна быть небольшой.

Количество деревьев в ансамбле и скорость обучения имеют обратную зависимость: чем меньше скорость обучения, тем больше деревьев требуется для

построения модели аналогичной сложности. Необоснованно большое число деревьев делает модель более сложной и может привести к переобучению.

В разделе 2.4 проведено сравнение результатов применения логистической регрессии, случайного леса и градиентного бустинга к данным настоящего диссертационного исследования.

1.7 Методы и алгоритмы, используемые в работе

Инструменты и методы для обработки данных

Для представления, обработки и анализа данных была использована библиотека pandas [96], входящая в состав Anaconda [97] – дистрибутива Python 3. Для анализа информации, выделенной из таблиц, были использованы библиотеки: numpy, rymorphy2, scikit-learn [98,99,100].

Алгоритмы подбора параметров для моделей машинного обучения, создания многоклассовых классификаторов, балансировки данных

Для построения всех моделей системы был использован алгоритм случайного леса в реализации RandomForestClassifier и RandomForestRegressor из библиотеки Scikit-learn. Обоснование выбранного алгоритма проведено в разделе 2.4.

В библиотеке sklearn реализовано два метода для подбора оптимальных гиперпараметров моделей:

- 1) GridSearchCV – исчерпывающий поиск по указанным значениям параметров;
- 2) RandomizedSearchCV – случайный поиск, проверяющий фиксированное количество параметров, либо случайно выбранных из заданного распределения, либо равновероятно выбранных из предложенного списка.

Эти методы используют кросс-валидацию для оптимизации параметров. Данные алгоритмы имеют существенный недостаток – ненаправленный поиск решения, алгоритм просто последовательно перебирает все предложенные наборы параметров, строит модели и сравнивает их качество. Иными словами,

новые модели строятся без учета опыта прошлых построений и оценок, хотя возможное решение могло бы получиться путем направленного изменения параметров в уже имеющихся моделях, что обеспечило бы большую скорость сходимости и меньшую вычислительную нагрузку.

Более удобный подход предоставляют библиотеки для автоматического отбора параметров, использующие эволюционные алгоритмы, градиентный спуск или другие базовые подходы к решению задач оптимизации значений моделей с большим числом параметров. Например, библиотека Hyperopt [101]. Пошаговое руководство по работе с библиотекой Hyperopt и подробный анализ того, почему байесовская настройка параметров работает лучше, чем случайный выбор, приведен в [102, 103].

Байесовская оптимизация имеет стратегию направленного поиска. Если при соответствующем изменении значений гиперпараметров, модель показывает более высокое качество, такое изменение считает перспективным, и поиск продолжается в данном направлении. Как следствие, требуется проверять и обучать меньшее число моделей, решение находится быстрее.

Для настройки параметров при работе с Hyperopt необходимо задать 1) целевую функцию, значение которой оптимизируется (кросс валидационный скор); 2) область возможных значений гиперпараметров, которая представлена вероятностью нахождения значения в некотором пространстве для числовых данных и списком возможных значений для категориальных; 3) алгоритм, определяющий правила перехода к очередному набору тестируемых параметров; и 4) априорную модель, которая в дальнейшем будет улучшаться.

Для данной работы была реализована функция настройки гиперпараметров для всех моделей при помощи библиотеки Hyperopt. Код функции представлен на рисунке 1.7.

```
1. import warnings
2. warnings.filterwarnings('ignore')
3.
4. NUM_EVALS = 1000
5. N_FOLDS = 5
```

```

6. RFC_MAX_DEPTH = 50
7. EVAL_METRIC_RFC_CLASS = 'roc_auc'
8. BEST_SCORE = 0
9.
10. def model_param_hyperopt(data, labels, package='rfc', num_evals=NUM_EVALS,
    diagnostic=False):
11.
12.     if package=='rfc':
13.
14.         print('Running {} rounds of RandomForestClassifier parameter opti-
    misation:'.format(num_evals))
15.         gc.collect()
16.
17.         integer_params = ['n_estimators', 'max_depth']
18.
19.         def objective(space_params):
20.
21.             for param in integer_params:
22.                 space_params[param] = int(space_params[param])
23.
24.             return{ 'loss': best_loss, 'status': STATUS_OK }
25.
26.             train_data, test_data, train_labels, test_labels =
    train_test_split(data, labels, test_size = 0.3)
27.             rfc = RandomForestClassifier(n_estimators=space_params['n_esti-
    mat'], criterion = space_params['criterion'], max_depth =
    space_params['max_depth'], max_features=space_params['max_feat'], bootstrap =
    space_params['boot-
    strap'], oob_score=space_params['oob_score']).fit(train_data, train_labels)
28.             rfc_predictions = rfc.predict(test_data)
29.             rfc_proba_predictions = rfc.predict_proba(test_data)
30.
31.             cv_strategy = StratifiedShuffleSplit(n_splits = 10, test_size =
    0.3, random_state = 2)
32.             scoring_cv = cross_val_score(rfc, test_data, test_labels, scor-
    ing = EVAL_METRIC_RFC_CLASS, cv = cv_strategy)
33.
34.             best_loss = 1 - scoring_cv['test-auc-mean'].iloc[-1]
35.
36.             return{ 'loss': best_loss, 'status': STATUS_OK }

```

```

37.
38.     criterion_list = ['gini', 'entropy' ]
39.     max_features_list = ['auto', 'sqrt', 'log2']
40.     bootstrap_list = ['True', 'False']
41.     oob_score_list = ['True', 'False']
42.
43.     space={'n_estimat': hp.quniform('n_estimat', 3, 100, 1),
44.           'max_depth': hp.quniform('max_depth', 2, 67, 1),
45.           'max_feat' : hp.choice('max_feat', max_features_list),
46.           'bootstrap' : hp.choice('bootstrap', bootstrap_list),
47.           'oob_score' : hp.choice('oob_score', oob_score_list),
48.           'criterion' : hp.choice('criterion', criterion_list)}
49.
50.     trials = Trials()
51.     best = fmin(fn=objective,
52.               space=space,
53.               algo=tpe.suggest,
54.               max_evals=num_evals,
55.               trials=trials)
56.
57.     best['criterion'] = criterion_list[best['criterion']]
58.     best['max_feat'] = max_features_list[best['max_feat']]
59.     best['bootstrap'] = bootstrap_list[best['bootstrap']]
60.
61.     print('{ ' + '\n'.join('{}: {}'.format(k, v) for k, v in
62.         best.items()) + ' }')
63.
64.     if diagnostic:
65.         return (best, trials)
66.     else:
67.         return (best)
68.
69.     else:
70.         print('not found')

```

Рисунок 1.7 – программная реализация функции настройки гиперпараметров ML-моделей

Случайный лес не требует длительной настройки параметров, важными являются параметры `n_estimators` – число деревьев, `max_features` – максимальное

число используемых признаков и опции предварительной обрезки деревьев, например, максимальная глубина (`max_depth`) [86]. Для `n_estimators` большее значение всегда дает лучший результат, но требует больше памяти и времени для обучения. Для параметра `max_features` лучше использовать значения по умолчанию: `max_features = sqrt(n_features)` для классификации и `max_features = n_features` для регрессии.

Имеющиеся данные не сбалансированы для большинства классов, что грозит построением классификаторов, неэффективных для предсказания малочисленных классов. Поэтому выполнена балансировка данных и, из-за небольшого размера выборки, применен алгоритм Adaptive Synthetic (ADASYN), реализованный в библиотеке `imbalanced-learn` [104,105]. Данный способ балансировки основан на подходе `over-sampling` [105], который заключается в произвольном дублировании и добавлении в выборку данных миноритарного класса.

Алгоритм дополняет выборку синтетическими объектами, созданными на основе имеющихся. Отличие ADASYN от других подобных алгоритмов заключается в генерации большего набора данных для тех объектов, которые сложнее прогнозировать, чем для тех, которые прогнозируются легче. ADASYN фокусируется на генерации объектов рядом с исходными, которые ошибочно классифицированы с использованием `kNN` классификатора. В результате такого подхода ADASYN улучшает данные за счет сокращения смещения, вносимого дисбалансом классов и дает классификатору больше наблюдений, помогающих настроиться на сложные случаи [105].

Используемые метрики качества и описание процесса получения их значений

Для оценки качества классификаторов использованы метрики AUC-ROC [105,106] и доля правильных ответов (англ. `assurasy`).

Доля правильных ответов или правильность – это метрика, дающая значение процента или доли верно предсказанных классов. В целом, если классы сбалансированы, то значение метрики можно применять для оценки бинарного

классификатора, однако, лучше не ориентироваться только на нее. При несбалансированных данных значение доли правильных ответов может вводить в заблуждение. Данную метрику удобно использовать для оценки качества многоклассового классификатора.

Для оценки качества регрессионных моделей использованы метрики среднеквадратичная ошибка (MSE) [107] и средняя абсолютная ошибка (MAE) [108], найден коэффициент детерминации R^2 .

Оценка значения метрик получена на тестовых данных с помощью использования перекрестной проверки [109,110]. В работе использована кросс-валидация на 5 блоков (фолдов) с помощью стратегии StratifiedKFold, реализованной в библиотеке Scikit-Learn [111].

Инструменты для разработки программного приложения

Для сохранения всех обученных моделей использована библиотека pickle [112]. Для разработки приложения использован PyQT5 (версия 5.9.2) – набор расширений для создания графического интерфейса на базе Qt5 для языка программирования Python [113] и Qt Designer – инструмент для быстрого создания графических пользовательских интерфейсов с виджетами из среды Qt GUI [114]. Сборка python-приложения в один файл осуществлялась при помощи программы PyInstaller.

1.8 Постановка цели и задач исследования

В первой главе проведен обзор современных исследований в области разработок математического и алгоритмического обеспечения для прогнозирования различных факторов, связанных с результатом применения ВРТ. Обзор показал, что задача актуальна и на сегодняшний день нет моделей предсказания успешности процедур ВРТ, выраженной в долгосрочном здоровье ребенка, также нет единого набора значимых признаков.

Также в первой главе описаны данные полученной медицинской документации, сделан вывод о том, что данные нуждаются в значительной предобработке. Далее представлены основные принципы работы создаваемой СППВР,

показана необходимость создания набора связанных прогностических моделей. Описаны использованные в работе методы, алгоритмы и инструменты.

Исходя из описания решаемой проблемы и проведенного анализа состояния предметной области сформулированы цель и задачи исследования.

Цель исследования. С использованием методов машинного обучения осуществить разработку математического и алгоритмического обеспечения системы поддержки принятия врачебных решений для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи вспомогательных репродуктивных технологий, а также создать алгоритмы обработки, компоновки и анализа неструктурированных текстовых данных медицинских организаций.

Для достижения поставленной цели были сформулированы и решены следующие **задачи**:

1. Разработать структуру СППВР, которая позволяет осуществлять многоэтапное прогнозирование здоровья ребенка, зачатого при помощи ВРТ;
2. Разработать алгоритмы обработки, компоновки и анализа разнородных текстовых данных на примере данных медицинских организаций для выделения признакового описания моделируемых объектов;
3. Создать математическое и алгоритмическое обеспечение СППВР для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи ВРТ, реализовать СППВР в виде программного приложения и апробировать СППВР в клинической практике.

2. РАЗРАБОТКА АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ ПРЕДОБРАБОТКИ ИСХОДНЫХ ДАННЫХ

Целью подготовки данных является извлечение из них признакового описания объектов данных и целевых значений. Понять, какие признаки существенно влияют, а какие нет, можно лишь после того, как все они будут извлечены и представлены в хорошем структурированном виде.

То есть задача заключается в получении таблицы размера (MxN), где N – число строк таблицы (или объектов данных), M – число столбцов таблицы (или признаков). На пересечении строки и столбца должна быть информация о наличии или отсутствии признака у объекта или количественная характеристика признака. Объекты данных – женщины или дети.

Задачу осложняют следующие недостатки в данных:

1. данные не обладают функциональной совместимостью (рис. 2.1);

Дата рождения	Год	срок гестации	масса тела при рождении	Диагноз основной при выписке	Диагноз сопутствующий при выписке	Диагноз основной при первой явке	Диагноз сопутствующий при первой явке	Диагноз основной в 6 месяцев	Диагноз сопутствующий в 6 месяцев	Диагноз основной в 1 год	Диагноз сопутствующий в 1 год	Диагноз основной в 1,5 года	Диагноз сопутствующий в 1,5 года	Диагноз основной в 2 года	Диагноз сопутствующий в 2 года	Диагноз основной в 3 года	Диагноз сопутствующий в 3 года
24.07.2017	2017	30	1470	P07.3 Нед-ть 30 недель. КРДН. ДН 2 ст.;	P59.9 Дакриоцистит справа.;			D89.9;	D50.8 Пеллозатизия слева. ФОО. РН задержка в анамнезе.;	D89.9;	D50.8 Пеллозатизия слева. ФОО. РН задержка в анамнезе.;	D89.9;	D50.8 Пеллозатизия слева. ФОО. РН задержка в анамнезе.;	D89.9;	D50.8 Пеллозатизия слева. ФОО. РН задержка в анамнезе.;		
18.05.2018	2018							G93.4;									
18.05.2018	2018	30 лет		Исход: Роды	24.02.2017	ИНФОРМАЦИЯ О ПАЦИЕНТЕ	ИНФОРМАЦИЯ ОБ ИСХОДЕ	ребенок 1	ребенок 2	ребенок 3	диагноз 1	диагноз 2					
15.11.2017	2017	31 год		Исход: Роды	24.07.2017	ИНФОРМАЦИЯ О ПАЦИЕНТЕ	ИНФОРМАЦИЯ ОБ ИСХОДЕ	ребенок 1	ребенок 2	ребенок 3	диагноз 1	диагноз 2					

Рисунок 2.1 – отсутствие функциональной совместимости данных разных медицинских организаций

В настоящий момент в нашей стране отсутствует единое цифровое пространство в здравоохранении, и это осложняет совместный анализ данных информационных систем различных организаций, поскольку все они обладают разной структурой, разработанной под свои цели, и не обладают интероперабельностью. Данные для работы получены из центра ВРТ, из женских консультаций и

1	ID	беременности
408	26078	мини- аборт
	26499	прерывание по медицинским показаниям в 24 недели
409		
410	27522	миниаборт
411	27278	Медицинский аборт
	27278	Внематочная беременность справа
412		
413	27019	тубная беременность справа.
414	25050	Медицинский аборт
	27444	регрессирующая беременность в сроке 8 недель
422		
	27141	регрессирующая беременность малого срока
426		
429	25215	регресс в 6 недель
	26544	неразвивающаяся беременность монозиготной двойни в 4-5 недель
431		
444	27402	тубная справа
	30455	Самопроизвольный выкидыш в

Рисунок 2.3 – замусоренность данных

4. данные плохо структурированы (рис. 2.4).

1	ID	беременности
2	24212	оперативные роды
3	25039	мини - аборт
4	23518	медицинский аборт в 12 недель. Выскабливание полости матки
	23518	Срочные роды. девочка, 3980/54. здорова.
5		
6	23518	Срочные роды, мальчик, 3800/54
	23518	Плановые оперативные роды. мальчик. 3700/51. здоров. Течение беременности без особенностей
7		
	28845	Преждевременные самопроизвольные роды в 32-33 нед. Дихориальная диамниотическая двойня
8		
	23826	неразвивающаяся беременность

Рисунок 2.4 – плохая структурированность данных

Например, какие-то признаки уже представлены отдельным столбцом («черепно-мозговые травмы»), а какие-то текстом в столбце («Хронические заболевания»). Или в данных одной и той же организации в одной таблице строка – пациентка, в другой – каждой пациентке соответствует несколько строк.

2.1 Алгоритм получения итогового набора данных

На основе данных за 2014-2018 гг., находящихся в АС «РАМ» ГАУЗ СО «ОДКБ» в количестве 248000 записей (передана 821 запись, созданная на основании запросов к АС), в базе данных АО «ЦСМ», 8500 записей, в базе данных – электронная карта пациента ГАУЗ СО «МКМЦ «Бонум», 9800 записей и в информационной системе «Регистр детей, нуждающихся в ранней помощи» ГАУЗ СО «ОДКБ», 14211 записей с сохранением персональных данных были оценены исходы и состояние здоровья детей, рожденных при помощи ВРТ.

После получения данных из АС «РАМ» был проведен поиск пациентов в данных из других баз по ID женщины или ID ребенка при помощи функции merge библиотеки pandas. В результате получилась одна таблица с ключом «ID ребенка» (рис. 2.5).

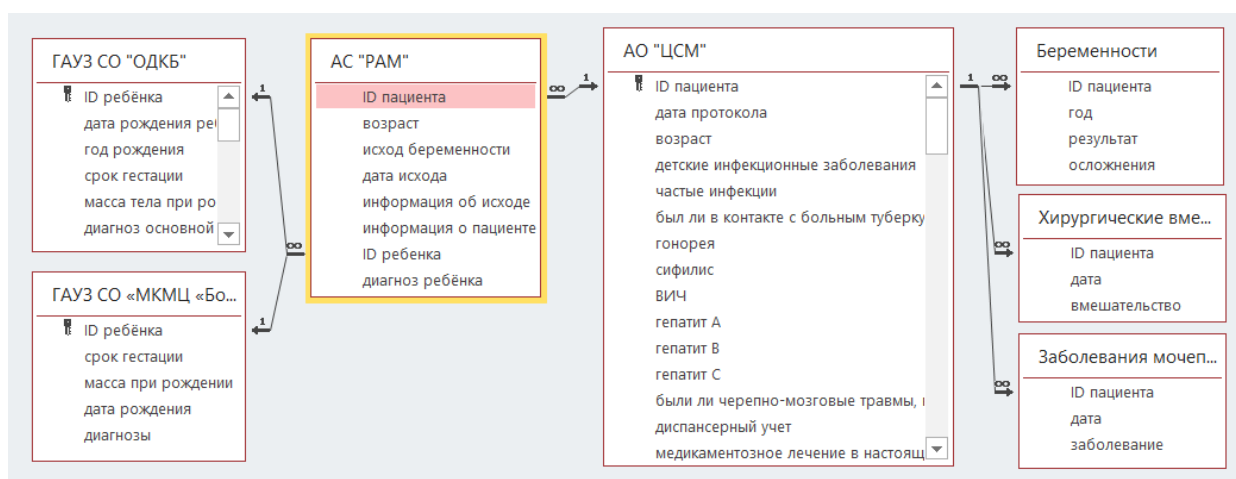


Рисунок 2.5 – Схема связи всех имеющихся данных

Предварительно в таблице данных АС «РАМ» был создан единый столбец «ID ребенка» вместо трех столбцов с идентификаторами детей и столбец «диагноз», вместо трех столбцов с диагнозами средствами библиотеки pandas (рис.2.6). «ID ребенка» - столбец, содержащий пропущенные значения при исходе «прерывание беременности», поэтому до предобработки были выбраны только строки с исходом беременности «роды».

```
1. df_csm = df_csm.merge(df_disease, how='outer')
```

```

2. df_csm = df_csm.merge(df_pregn, how='outer')
3. df_all = df_csm.merge(df_ram, how='inner')
4. df_bonum = df_bonum.merge(df_odkb, how='outer')
5. df_all = df_all.merge(df_bonum, how = 'left')

```

Рисунок 2.6 – Объединение всех таблиц с данными исследования

```

1. df_ram = pd.read_csv('AC PAM.csv', sep=';', encoding='ansi', decimal=',')
2. df_ram1 = df_ram.loc[df_ram['исход беременности'] == 'прерывание беременности']
3. df_ram = df_ram.loc[df_ram['исход беременности'] == 'роды']
4.
5. ID = []
6. diagnosis = []
7. for j in ['ID первого ребенка', 'ID второго ребенка', 'ID третьего ребенка']:
8.     for i in range(0, len(df_ram)):
9.         if not pd.isnull(df_ram[j][i]):
10.            ID.append(df_ram[j][i])
11.            if j == 'ID первого ребенка': diagnosis.append(df_ram['диагноз пер-
12.                вого ребенка'][i])
13.            if j == 'ID второго ребенка': diagnosis.append(df_ram['диагноз вто-
14.                рого ребенка'][i])
15.            if j == 'ID третьего ребенка': diagnosis.append(df_ram['диагноз
16.                третьего ребенка'][i])
17. df = pd.DataFrame()
18. df['ID ребенка'] = ID
19. df['diagnosis'] = diagnosis
20. df['ID пациента'] = ''
21. for i in range(0, len(df)):
22.     for j in range(0, len(df_ram)):
23.         if (df['ID ребенка'][i] == df_ram['ID первого ребенка'][j]) or (df['ID
24.             ребенка'][i] == df_ram['ID второго ребенка'][j]) or (df['ID ребенка'][i] ==
25.             df_ram['ID третьего ребенка'][j]):
26.             df['ID пациента'][i] = df_ram['ID пациента'][j]
27. df = df_ram.merge(df, how='inner')
28. df.drop(columns = ['ID первого ребенка', 'ID второго ребенка', 'ID третьего ре-
29.     бенка', 'диагноз первого ребенка ', 'диагноз второго ребенка ', 'диагноз треть-
30.     его ребенка '], axis = 0, inplace=True)
31. df
32. df_ram1.drop(columns = ['ID первого ребенка', 'ID второго ребенка', 'ID третьего
33.     ребенка', 'диагноз первого ребенка ', 'диагноз второго ребенка ', 'диагноз треть-
34.     его ребенка '], axis = 0, inplace=True)
35. df_ram1['ID ребенка'] = np.nan
36. df_ram1['diagnosis'] = np.nan
37. df_ram = pd.concat([df, df_ram1], ignore_index=True)

```

Рисунок 2.6 – Код предобработки таблицы АС «РАМ»

В результате объединения данных была получена 821 оригинальная запись женщин и рожденных ими 836 детей на этапе родильного дома, а также состояния здоровья 464 детей в возрасте до трех лет.

Из 821 женщин полная информация была найдена только для 572, которые использовались для построения моделей о результатах беременностей и родов.

Для построения предсказания вероятности наступления беременности, были использованы дополнительные данные, не включенные в 572 полученных образца. Для того, чтобы дополнить данные информацией о пациентках и протоколах, в результате которых беременность не наступила, были добавлены 282 протокола АО «ЦСМ», из которых 188 – протоколы ЭКО, 94 – протоколы криопереноса. Таким образом, общее число протоколов, включенных в анализ – 854.

2.2 Алгоритм обработки текстовых данных

Анализ структуры предоставленных данных для задачи позволил выделить этапы, прохождение которых последовательно упрощает задачу, делая данные более структурированными и чистыми.

Этап 1. Одна строка - один объект данных

Этап заключается в объединении в одну строку информации об одном объекте данных, разрозненной по нескольким ячейкам таблицы, но предоставляющей описание одной группы признаков (например, «акушерско-гинекологический» анамнез)

Этап 2. Создание атомарных ячеек

Суть действий этапа в создании гарантированного доступа к данным, то есть удовлетворении второго правила Кодда: каждое отдельное (атомарное) значение данных должно быть логически доступно с помощью комбинации имени таблицы, значения первичного ключа (объекта данных) и имени столбца.

Этап 3. Кластеризация на семантическом уровне

Кластеризация – объединение диагнозов в укрупненные группы с отсеиванием незначимых деталей. Если признаки будут не сгруппированы, то каждый пациент будет описан уникальным набором признаков и мы не сможем понять, влияет ли каждый на что-нибудь – большую выборку раздробим на маленькие кусочки. После группировки (кластеризации) пациенты будут объединены под укрупненными признаками (например, «внематочная беременность» включает в

себя внематочную слева/справа, трубную беременность. Или «гинекологическая операция», объединяющая все хирургические вмешательства в гинекологии. То есть кластеризуем диагнозы (столбцы), а не пациенток (строки) на основе словарей, которые создаем самостоятельно, вкладывая в них экспертные знания врача.

Блок-схема алгоритма представлена на рисунке 2.7.



Рисунок 2.7 – алгоритм обработки текстовых данных

Рассмотрим алгоритм на примере построения признакового описания объектов данных, когда объекты данных – женщины.

Этап 1. “одна строка - один объект данных”

Данные из клиники ВРТ представлены в виде нескольких таблиц. В основной находится описание пациенток, которое дополнено информацией из дополнительных таблиц, в которых каждая пациентка может быть представлена несколькими строками.

Поэтому мы объединили данные нескольких строк, относящихся к одинаковому пациенту, в одну строку (рис. 2.8).

	A	B
1	ID пациента	Беременности
163	22386	Срочные роды
6983 строки	33003	срочные роды
164	33003	медицинский аборт. Выскабливание полости матки
165	33003	Левосторонняя трубная беременность.
166	33003	Тубэктомия слева
167	33003	Правосторонняя трубная беременность.
168	33003	Левосторонняя тубэктомия
169	32796	Срочные оперативные роды
27692	мини-аборт	регрессирующая беременность в сроке 5-6
	A	B
	ID пациента	Беременности
59	32999	срочные роды медицинский аборт регрессирующая беременность малого срока. срочные оперативные
3404 строки	33000	срочные роды прерывание беременности в сроке 14 недель по медицинским показаниям (ВПР ЦНС у плода
50	33003	срочные роды медицинский аборт. Выскабливание полости матки Левосторонняя трубная беременность. Тубэктомия слева Правосторонняя трубная беременность. Левосторонняя тубэктомия
51	33005	Медицинский аборт, МВА
52		медицинский аборт самостоятельные роды - девочка 3125. 50 см регрессирующая беременность в сроке 8-9

Рисунок 2.8 – приведение к одной строке всех строк с информацией об одном объекте данных на примере таблицы «Беременности»

Этап 2. создание атомарных ячеек

После объединения, сделанного на прошлом этапе, в некоторых ячейках содержатся весьма пространственные описания пациентов, что затрудняет доступ к данным. Выделим из ячеек атомарные данные и распределим их по соответствующим столбцам.

Поскольку данные представляют собой произвольный текст, при токенизации число столбцов очень большое и требуется его сократить:

- 1) привести к нижнему регистру;
- 2) очистить “мусор” – в данном случае знаки препинания и буквы латинского алфавита;
- 3) выполнить лемматизацию - приведение к нормальной форме;
- 4) выполнить токенизацию;
- 5) выполнить первичное удаление стоп-слов.

беременность	биохимический	кесарев
беремеенность	билхимический	кесарев
беременнистить	биохимический	кесаревао
беременнность		кесарнво
беременнньсть		кесарово
беременность	выскабливание	кесерево
беременностьва	выскабливание	
беременностьмалый	выскабливание	медицинский
беременностьпосле	высблвание	медицинкий
беременностьть	выяскабливание	медицинский
беременнотситъ		медицинскиттъ
беременность	преждевременный	медициский
беременнсоть	преждевоеменный	медициский
беременнсть	преждевременный	медицицнский
беременность	преждевременный	междицинский
беременнность	преждевреме́ный	
береммность	прежденвременный	
беремнность	прждевременной	
берменность		
еременность		

Рисунок 2.10 – примеры группировки слов с ошибками написания к оригинальному слову

Создали словарь, в котором есть соотношение признака и укрупненной группы (рис. 2.11). Сгруппировали признаки по смыслу. Далее просматриваем все слова текста и слова семантических групп: если слово в тексте относится к какой-либо группе, то параметр `word.group` этого слова делаем равным `group.name`. Создаем список `words_groups` из названия группы для каждого слова. Список нужен для названия столбцов в итоговой таблице. Теперь каждое слово представлено именем группы, к которой он принадлежит (или 0).

Заменяем в матрице - векторайзере названия столбцов на `words_groups`: складываем столбцы с одинаковым названием и заменяем суммы, отличные от нуля на единицу, удаляем столбец с названием "0".

Таким образом, используя созданный алгоритм обработки и компоновки данных, удалось получить все желаемые признаки из каждой ячейки данных. На рисунке 2.12 столбец «Результат» содержит данные объекта, представленные в одной ячейке, а каждый из следующих столбцов содержит информацию о наличии у объектов данных одного из признаков.

```

semantic_groups = []
group = Group(0)
group.name = 'самопроизвольный выкидыш'
group.words = ['выкидыш', 'самопрои́за', 'самопроизвольный']
semantic_groups.append(group)
group = Group(1)
group.name = 'поздний самопроизвольный или индуцированный выкидыш'
group.words = ['13', '14', '15', '16', '17', '18', '19', '20', '21',
               '22', '23', '24', '25', '26', '27', '28', '29', '30', '31']
semantic_groups.append(group)
group = Group(2)
group.name = 'выкидыш или аборт в малом сроке'
group.words = ['01', '03', '04', '4неда', '06', '09', '10', '11', '12']
semantic_groups.append(group)
group = Group(3)
group.name = 'индуцированный выкидыш / аборт'
group.words = ['индукция', 'индуцировать', 'аборт', 'абот', 'абрт', 'мёд',
               'медаборт', 'медикаментозный', 'медицинский',
               'медицинсикть', 'миниаборт', 'прерваться', 'прерывание']
semantic_groups.append(group)
group = Group(4)

```

Рисунок 2.11 – пример созданных семантических групп

: res.head()

	внематочная беременность	выкидыш или аборт в малом сроке	индуцированный выкидыш / аборт	неразвивающаяся беременность	оперативные роды	поздний самопроизвольный или индуцированный выкидыш	преждевременные роды	самопроизвольный с выкидыш
Результат								
срочные оперативные роды (незрелая шейки матки). Мальчик 4110, 51 см.	0	0	0	0	1	0	0	0
кесарево сечение 37 недель - мальчик - 2810гр, 48 см	0	0	0	0	1	0	1	0
срочные оперативные роды в39-40 недель после ЭКО.девочка	0	0	0	0	1	0	0	0

Рисунок 2.12 – итоговая векторизация объектов

Завершив обработку таблиц и выделение признаков акушерского анамнеза, использовали тот же подход для обработки таблиц с хирургическим анамнезом.

Для хирургических вмешательств после объединения строк по ID пациента удалось сократить число строк с 17608 до 5039, то есть более, чем в 3 раза (рис. 2.13). В этой таблице информация сложнее всего структурировалась из-за большого числа возможных вариантов написаний и главное – большого числа возможных хирургических вмешательств.

(из данных АС «РАМ») и до достижения трехлетнего возраста (из данных ГАУЗ СО «ОДКБ» и ГАУЗ СО «МКМЦ «Бонум»).

В данных АС «РАМ» каждой строкой является пациентка, а дети описаны в столбцах (с учетом того, что родится у пациентки может до трех детей). Был создан единый столбец «ID ребенка» вместо трех столбцов с идентификаторами детей и столбец «диагноз», вместо трех столбцов с диагнозами, что подробно описано в разделе 2.1.

Данные по детям также представлены по-разному: где-то диагноз содержит код диагноза и его текстовое описание, где-то диагнозы представлены в виде списка кодов диагнозов для каждого ребенка, где-то представлены в виде произвольного текста, разбитого по месяцам ведения ребенка (рис. 2.17). В каждой ячейке могло быть по несколько диагнозов, могли быть пропущенные значения.

У	Н	И	Л	К	Л	М	Н	О	Р	
Диагноз основной при выписке	Диагноз сопутствующий при выписке	Диагноз основной при первой явке	Диагноз сопутствующий при первой явке	Диагноз основной в 6 месяцев	Диагноз сопутствующий в 6 месяцев	Диагноз основной в 1 год	Диагноз сопутствующий в 1 год	Диагноз основной в 1,5 года	Диагноз сопутствующий в 1,5 года	Диагноз основной в 2 года
P07.3 Нед-ть 30 недель. КРДСН. ДН 2 ст. ;	P59.9 Дакриоцистит справа. ;			D89.9 ; G93.4 Темповая задержка	D50.8 Пиелозктазия слева. ФОО. РН в анамнезе. ;	D89.9 ; G93.4 Темповая задержка	D50.8 Пиелозктазия слева. ФОО. РН в анамнезе. ;	D89.9 ; G93.4 Темповая задержка		D50.8 Пиелозктазия слева. ФОО. РН в анамнезе. ;
R17 ;	P91.6 ;	R17 ;	D50.8 ;	D89.9 ; G93.4 ;	E67.8 ;	G93.4 ;		L27.9 ;		M ;
					грамма	D	E	F	G	Va
						H04.3	K59.9			де
P2	ребен ок 2	ребен ок 3	диагноз 1		диагноз 2		диагноз 3		диагноз 4	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.1		H35.1	
головой,			Z38.0, Один ребенок, рожденный в стац,		нет дз					
головой,			P21.1, Средняя и умеренная асфикция ,		G93.4		Q65.6		M21.1	
головой,			Z37.0, Один живорожденный		G93.4		G93.8		H04.5	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H50.0		H52.0	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,			Z38.0, Один ребенок, рожденный в стац,		H35.1		H35.2		H31.	
головой,										

- 1) по принятой классификации диагнозов в укрупненные группы;
- 2) по степени “опасности” диагнозов (по группам здоровья).

На рисунке 2.18 представлены результаты группировки информации о диагнозах детей.

[illegible]

Рисунок 2.18 – примеры группировки диагнозов детей по классам в соответствии с МКБ-10 и по группам здоровья

2.3 Алгоритм обработки орфографических ошибок

Данные в формате произвольного текста содержат орфографические ошибки. Поэтому для корректировки правописания и устранения лишних столбцов в матрице частот слов (этап 2 алгоритма обработки текстовых данных), был использован алгоритм приведения всех слов с опечатками к правильному слову с использованием расстояния Левенштейна.

Блок-схема алгоритма представлена на рисунке 2.19.

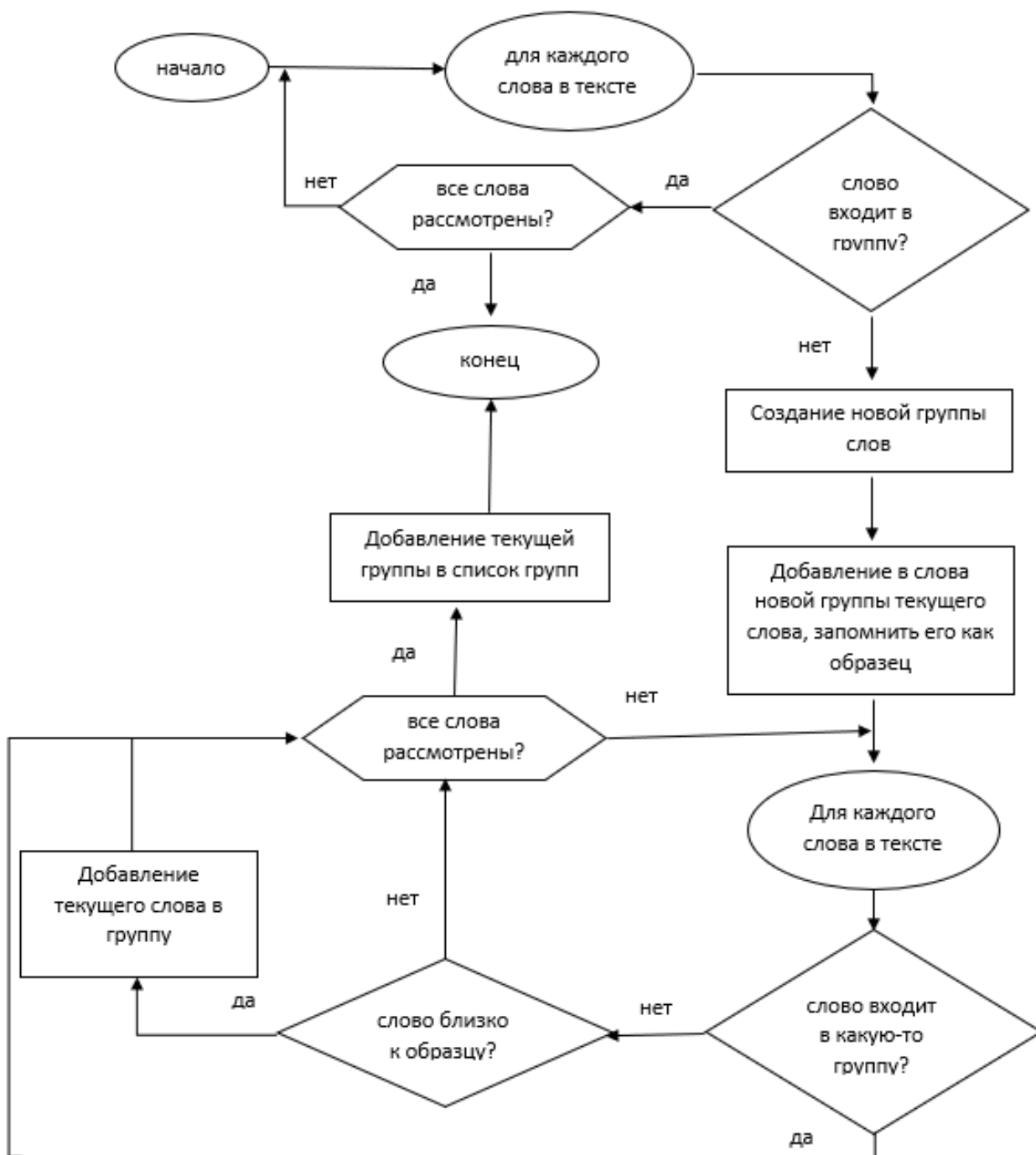


Рисунок 2.19 – блок-схема алгоритма обработки ошибок написания

До начала алгоритма каждое слово из матрицы счетчика слов представляем как объект класса «Word», который имеет четыре параметра: первый параметр – «word» – слово, задается при создании объекта, «in_group» – находится ли слово уже в какой-то группе (изначально «False»), «group» – номер группы (изначально «0»), «count» – число, сколько раз слово встречается в тексте (изначально «0»).

Для распределения слов на группы, был создан класс «Group», каждый объект которого имеет параметры: «number» – номер группы, задается при создании группы, «words» – список слов в группе (изначально пустой список), «name» – имя группы (изначально «none»).

Считаем, сколько раз каждое слово встречается в тексте. Далее создаем список групп по следующему алгоритму (код алгоритма представлен на рисунке 2.20):

- смотрим список слов, встречающихся в тексте:
- если слово еще не входит ни в одну группу, то:
 - создаем группу с новым номером;
 - добавляем в слова созданной группы это слово;
 - запоминаем это слово как образец;
 - просматриваем все имеющиеся слова в матрице счетчика слов:
 - если текущее слово еще не состоит ни в одной группе, то:
 - если расстояние Левенштейна от образца до текущего слова > 75 :
 - добавляем текущее слово в группу;
 - добавляем созданную группу в список групп;
- даем имя каждой группе по тому слову группы, которое наиболее часто встречается в тексте;
- создаем список из названия группы для каждого слова (список будет использован для названия столбцов в итоговой таблице);
- в матрице счетчика слов задаем новые названия столбцов: столбцы – название группы, которой принадлежит слово, заданное этим столбцом, строки – вместилища;
- складываем столбцы с одинаковым названием, заменяя все числа, кроме нуля на единицу.

```
1.      class Word:
2.          def __init__(self, word:str):
```



```

3.         self.in_group = False
4.         self.word = word
5.         self.group = 0
6.         self.count = 0
7.
8.     class Group:
9.         def __init__(self,n: int):
10.             self.words = []
11.             self.number = n
12.             self.name = 'none'
13.
14. words = []
15.
16. for word in cv2.get_feature_names():
17.     words.append(Word(word))
18. for x in lemm_phrases:
19.     for xx in x.split():
20.         for word in words:
21.             if xx == word.word:
22.                 word.count += 1
23.                 continue
24.
25. for group in groups:
26.     count = 0
27.     for word in group.words:
28.         if word.count > count:
29.             count = word.count
30.             k = word.word
31.     group.name = k
32.
33. words_groups = []
34. for group in groups:
35.     for word in group.words:
36.         words_groups.append(group.name)
37.
38. data2=pd.DataFrame(data=x2.toarray(), columns=words_groups,in-
    dex=df[df.columns[1]])
39.
40. def sjoin(x):
41.     if sum(x) > 1:

```

```

42.         return 1
43.     else:
44.         return sum(x)
45.
46. result = data2.groupby(level=0, axis=1).apply(lambda x: x.apply(sjoin,
axis=1))

```

Рисунок 2.20 – код алгоритма обработки орфографических ошибок

В результате получаем обновленную матрицу счетчика слов меньшего размера. Например, для столбца «хирургические вмешательства» число столбцов сократилось с 3782 до 1976.

Выводы

Во второй главе решена вторая задача диссертационного исследования, разработаны алгоритмы обработки, компоновки и анализа разнородных текстовых данных на примере данных медицинских организаций для выделения признакового описания моделируемых объектов, а именно:

- 1) представлен алгоритм слияния разнородных текстовых данных, полученных из баз данных различных медицинских организаций, в единую базу данных;
- 2) представлен алгоритм извлечения из данных признакового описания объектов данных и целевых значений;
- 3) представлен алгоритм обработки орфографических ошибок.

3. РАЗРАБОТКА МАТЕМАТИЧЕСКОГО И АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ СППВР

3.1 Входные и выходные данные для каждого блока системы

Согласно основным принципам работы СППВР, описанным в разделе 1.4, разрабатываемые прогностические модели были сгруппированы по четырем блокам. Входные и выходные данные для каждого блока представлены в таблице 2.

Первый блок «Прогнозирование результата отбора пациентов» включает в себя ML-модели, построенные на основе всех имеющихся пациенток. Как отдельный признак выделена программа, по которой проводится оплодотворение пациентки (ЭКО либо криоперенос). Остальные признаки представляют собой данные акушерского анамнеза женщины, информацию о проведенных хирургических вмешательствах, о хронических заболеваниях, особенностях менструального цикла, вредных производственных факторах, родственный анамнез, возраст, индекс массы тела (ИМТ) и номер попытки ВРТ.

Таблица 2 – входные и выходные данные по каждому блоку системы

Первый блок «Прогнозирование результата отбора пациентов»	
<u>Ввод:</u> - программа: ЭКО / криоперенос; - номер попытки ВРТ; - акушерский анамнез; - наследственный анамнез; - хирургические вмешательства; - хронические заболевания; - данные текущего здоровья.	<u>Вывод (для одноплодной и многоплодной беременности):</u> - вероятность наступления беременности; - возможные осложнения беременности; - срок родоразрешения; - способ родоразрешения; - группа здоровья ребенка; - классы диагнозов по МКБ-10.
Второй блок «Прогнозирование результата овариальной стимуляции»	
<u>Ввод:</u> - данные блока 1; - претритмент; - протокол; - триггер.	<u>Вывод (для одноплодной и многоплодной беременности):</u> - ожидаемое число фолликулов; - толщина эндометрия (норма или отклонение); - качество ооцитов (хорошее или плохое); - выводы блока 1 с поправкой на данные ввода блока 2

Третий блок «Прогнозирование результата оплодотворения»	
<u>Ввод:</u> - данные блоков 1,2; - количество зрелых ооцитов; - тип оплодотворения; - среды для проведения оплодотворения.	<u>Вывод:</u> - количество клеток хорошего качества на третьи сутки; - количество бластоцист хорошего качества на пятые сутки; - количество бластоцист хорошего качества на шестые сутки; - общее количество бластоцист на шестые сутки.
Четвертый-а блок «Прогнозирование результата переноса (для программы ЭКО)»	
<u>Ввод:</u> - данные блоков 1,2,3; - количество эмбрионов; - качество эмбрионов; - сутки переноса.	<u>Вывод (для одноплодной и многоплодной беременности):</u> - выводы блока 1 с поправкой на данные ввода блоков 2,3,4
Четвертый-б блок «Прогнозирование результата переноса (для программы криопереноса)»	
<u>Ввод:</u> - данные блока 1; - количество эмбрионов; - качество эмбрионов; - сутки переноса; - среда заморозки эмбрионов; - среда разморозки эмбрионов; - срок хранения эмбрионов.	<u>Вывод (для одноплодной и многоплодной беременности):</u> - выводы блока 1 с поправкой на данные ввода блока 4

Модели из первого блока выдают прогнозы следующих исходов: наступление беременности и (отдельно для многоплодной и одноплодной беременности) прогнозируемые осложнения беременности, срок и способ родов, группа здоровья ребенка с указанием классов диагнозов по МКБ-10, по которым возможно возникновение проблем со здоровьем.

Если в первом блоке выбрана программа ЭКО, то становятся доступными предсказания из второго и третьего блоков. Иначе, взаимодействие со вторым и третьим блоком недоступно для пользователя, и выводится сообщение о то, что данные этапы не предусмотрены программой криопереноса.

Второй блок «Прогнозирование результатов овариальной стимуляции» использует для построения выводов не только данные из первого блока, но и данные стимуляции: возможные варианты претритмента (предусмотрен выбор одного варианта, возможно, в сочетании с гистероскопией), протокол стимуляции

и используемый триггер. Результатом работы блока является вывод всех перечисленных в первом блоке результатов прогнозирования с учетом поправки на данные стимуляции, а также ожидаемых результатов стимуляции: ожидаемое число фолликулов, толщина эндометрия (норма или отклонение) и качество ооцитов (хорошее или плохое).

В третьем блоке «Прогнозирование результата оплодотворения» вводится количество зрелых ооцитов (оно может быть меньше ожидаемого числа фолликулов после стимуляции или, наоборот, больше, при добавлении донорских или ранее криоконсервированных ооцитов пациентки), тип оплодотворения (ЭКО, ИКСИ, ЭКО+ИКСИ), а также тип среды для проведения оплодотворения. Результаты прогнозирования блока: предсказание количества клеток хорошего качества на третьи сутки, количества бластоцист хорошего качества на пятые сутки, количества бластоцист на шестые сутки и общего количества бластоцист на шестые сутки.

Четвертый блок «Прогнозирование результата переноса» делится на две составные части в зависимости от выполняемого протокола: перенос после проведения ЭКО или перенос эмбрионов из крио. Для программы ЭКО признаки включают в себя количество эмбрионов (один или два), выбор суток переноса (третьи или позже) и выбор качества первого и второго эмбриона (или единственного эмбриона). Для программы криопереноса кроме количества и качества эмбрионов пользователь должен указать срок хранения криоконсервированных эмбрионов, среду их заморозки и разморозки и сутки переноса (пятые или шестые). Результаты прогнозирования аналогичны результатам работы моделей первого блока, но имеют скорректированный прогноз с учетом влияния всех используемых в системе признаков.

3.2 Структура разрабатываемой системы

Структура системы представлена на рисунке 3.1.

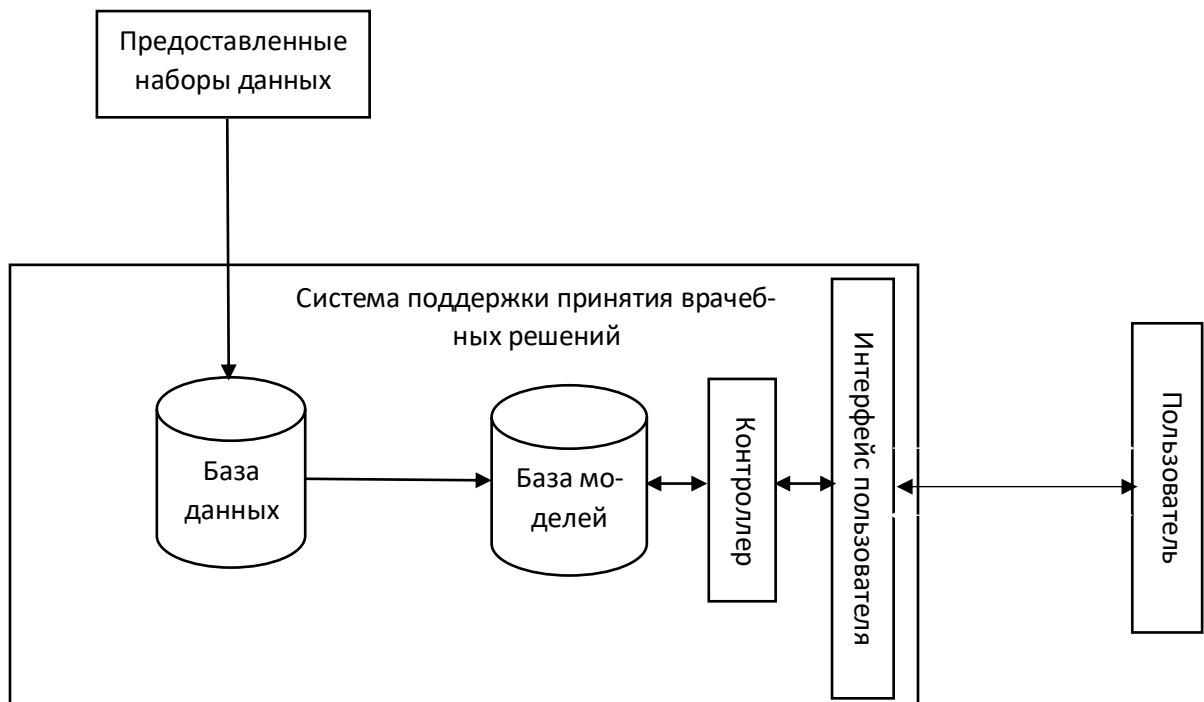


Рисунок 3.1 – Структура СППВР

Разработанная СППВР состоит из взаимодействующих между собой базы данных, базы моделей, контроллера и пользовательского интерфейса, которые описаны ниже.

Данные, представляющие собой описанные в разделе 1.3 выгрузки из информационных систем медицинских организаций, были обработаны с целью получения признакового описания объектов. Соответствующие признаковые описания сохранены в базу данных.

Заполнение базы данных сделано автоматизированным. Именно, если появляются дополнительные данные в выгрузках из ИС медицинских организаций, то базу данных можно обновить и дополнить автоматически без написания дополнительного программного кода.

После применения алгоритмов машинного обучения с учителем к данным, получена база моделей, являющаяся математическим обеспечением разрабатываемой СППВР.

Как отмечено в разделе 1.4 ключевая особенность разработанной СППВР в том, что это не просто набор несвязанных классификаторов и регрессионных моделей, а иерархия моделей, позволяющих выполнять последовательное уточнение прогнозов на всех этапах ВРТ. Логика взаимодействия ML-моделей между собой и порядок их работы составляют алгоритмическое обеспечение созданной СППВР и регулируются контроллером.

Пользовательский интерфейс позволяет пользователю вводить данные, описывающие конкретный новый объект, для которого нужно дать прогноз. То есть с помощью пользовательского интерфейса врач взаимодействует с СППВР, описывая женщину, вступающую в протокол ВРТ.

Входные данные передаются через контроллер соответствующим моделям. Модели проводят вычисления, т. е. осуществляют прогнозирование, и результаты отображаются пользователю посредством пользовательского интерфейса. На основе отображаемых результатов пользователь СППВР может принимать персонифицированные медицинские решения. Кроме того, пользователь может варьировать входные параметры и сравнивать соответствующие результаты.

Функциональная схема работы СППВР представлена на рисунке 3.2.

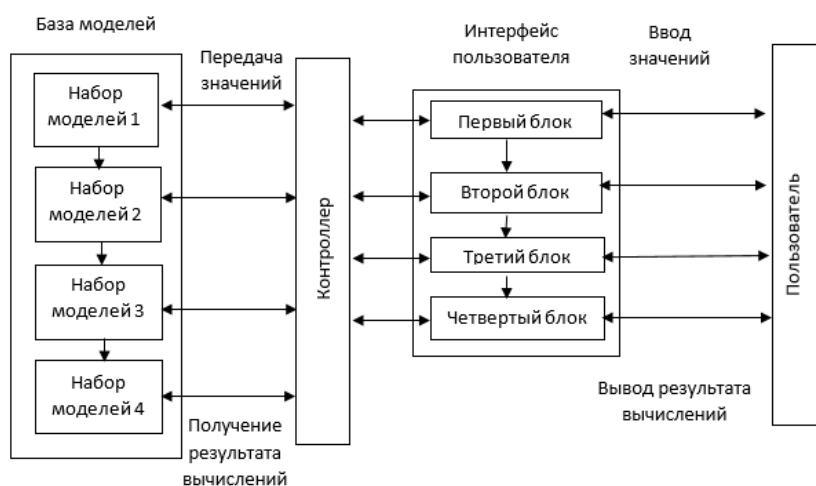


Рисунок 3.2 – функциональная схема работы системы

Согласно функциональной схеме работы системы, интерфейс пользователя предоставляет возможность вводить признаковое описание объектов для

четырёх различных блоков системы. Каждый блок отвечает за признаки, которые добавляются к описанию объекта на определенном этапе процедуры ВРТ. При этом данные, введенные пользователем в текущий блок, могут быть использованы в качестве значимых признаков для прогностических моделей последующих блоков.

Каждому блоку системы соответствует определенный набор моделей машинного обучения, в котором каждая модель принимает на вход набор значимых для нее признаков и возвращает контроллеру результаты вычислений – спрогнозированные значения.

Контроллер позволяет группировать входные данные в наборы признаков и отправлять в каждую из ML-моделей нужный набор. Также контроллер отвечает за последовательность прогнозирования в наборах моделей, то есть задает очередность построения прогнозов и передает спрогнозированные значения в следующие по логике модели, использующие эти значения в качестве предикторов. Пользователь получает вывод результатов вычислений для каждого блока.

Процесс создания базы моделей происходит вне СППВР. Он представлен блок-схемой на рисунке 3.3. Согласно блок-схеме, к предоставленным для исследования данным применяются алгоритмы их предобработки. В результате создается единая база данных, хранящая признаковые описания всех объектов и ответы (исходы). Далее для каждого блока системы рассматриваются все исходы, подлежащие прогнозированию в данном блоке, а для каждого исхода выбирается набор объектов и значимых признаков из базы данных. Затем для каждого созданного набора данных применяется алгоритм машинного обучения, с помощью которого создается модель машинного обучения – прогностическая модель, способная для любого объекта выдать достаточно точный ответ. Сохраненные модели подгружаются в СППВР при ее запуске.

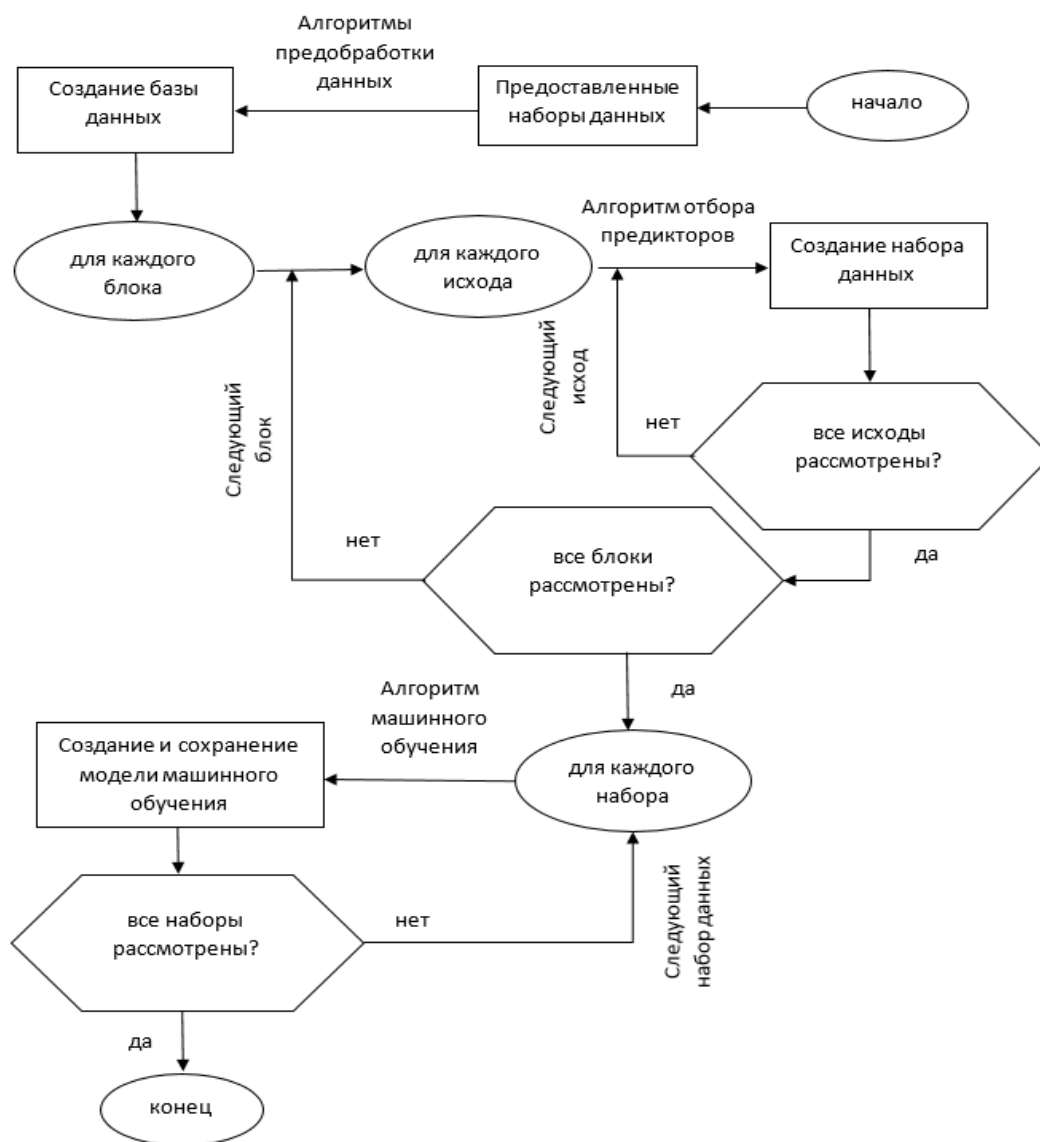


Рисунок 3.3 – алгоритм создания базы моделей

3.3 Перечисление всех моделей машинного обучения в составе системы

Распределение моделей по блокам системы представлено в Приложении А. Первый блок («Прогнозирование результата отбора пациентов») содержит 26 моделей:

- 1) бинарный классификатор для прогнозирования наступления беременности;
- 2) бинарный классификатор для прогнозирования прерывания беременности;

- 3) бинарные классификаторы для прогнозирования возможных осложнений беременности: истмико-цервикальная недостаточность, гипертензивные расстройства, предлежание плаценты, гестационный сахарный диабет, нарушения количества околоплодных вод, преждевременный разрыв плодных оболочек. Последние три модели используют в числе прочих признаков прогнозы, полученные от первых трех моделей (шесть моделей);
- 4) многоклассовые классификаторы для прогнозирования способа родов и группы здоровья ребенка (две модели);
- 5) регрессионная модель для оценки срока родов в неделях;
- 6) бинарные классификаторы для прогнозирования различных классов заболеваний ребенка по МКБ-10 (15 моделей).

Второй блок «Прогнозирование результата овариальной стимуляции» содержит 29 моделей:

- 1) модели из первого блока: классификаторы для прогнозирования прерывания беременности, гестационного сахарного диабета, нарушения количества околоплодных вод, преждевременного разрыва плодных оболочек, способа родов, регрессионная модель для оценки срока родов, а также девять классификаторов предсказания различных классов заболеваний ребенка (15 моделей);
- 2) модели, представляющие доработки моделей из первого блока, но с большим числом значимых признаков – с учетом данных стимуляции: классификаторы для прогнозирования наступления беременности, истмико-цервикальной недостаточности, гипертензивных расстройств, предлежания плаценты, группы здоровья ребенка, а также шесть классификаторов предсказания различных классов заболеваний ребенка (11 моделей);
- 3) модели для прогнозирования данных стимуляции: бинарные классификаторы для прогнозирования толщины эндометрия (норма или отклонение) и качества ооцитов (хорошее или плохое), регрессионная модель

для оценки количества фолликулов (три модели).

Третий блок «Прогнозирование результата оплодотворения» содержит четыре регрессионные модели для прогнозирования результатов оплодотворения: количество клеток хорошего качества на третьи сутки, количество бластоцист хорошего качества на пятые сутки, количество бластоцист хорошего качества на шесте сутки, общее количество бластоцист.

Четвертый-а блок «Прогнозирование результата переноса (для программы ЭКО)» содержит 27 моделей:

- 1) модели из первого блока и второго блоков: модели для прогнозирования возможных осложнений беременности, срока и способа родов, а также восемь классификаторов предсказания различных классов заболеваний ребенка (16 моделей);
- 2) модели, представляющие доработки моделей из первого и второго блоков, но с большим числом значимых признаков – с учетом данных оплодотворения и переноса: классификатор для прогнозирования наступления беременности, классификатор для прогнозирования прерывания беременности, группы здоровья ребенка, а также семь классификаторов предсказания различных классов заболеваний ребенка (10 моделей);
- 3) бинарный классификатор для прогнозирования количества плодов.

Четвертый-б блок «Прогнозирование результата переноса (для программы криопереноса)» содержит 27 моделей:

- 1) модели из первого блока: модели для прогнозирования возможных осложнений беременности, срока и способа родов, а также одиннадцать классификаторов предсказания различных классов заболеваний ребенка (19 моделей);
- 2) модели, представляющие доработки моделей из блока 1, но с большим числом значимых признаков – с учетом данных программы криопереноса эмбрионов и данных переноса: классификатор для прогнозирования наступления беременности, классификатор для прогнозирования прерывания беременности, группы здоровья ребенка, а также четыре

классификатора предсказания различных классов заболеваний ребенка (семь моделей);

3) бинарный классификатор для прогнозирования количества плодов.

Разработанная СППВР содержит 63 модели, из которых шесть являются регрессионными, пять – многоклассовыми классификаторами, остальные 52 – бинарными классификаторами.

3.4 Результаты сравнения алгоритмов для отбора признаков и машинного обучения

Для того чтобы строить модели машинного обучения, обладающие наилучшими прогностическими способностями / дающие наиболее точные прогнозы (в рамках настоящего исследования), необходимо было использовать соответствующий алгоритм отбора признаков (см. раздел 1.5), и алгоритм построения модели (см. раздел 1.6).

Покажем, как осуществлялся выбор алгоритма отбора признаков и алгоритма построения ML-модели на примере задачи прогнозирования наступления беременности. Результаты представлены в таблице 3. Полужирным выделены значения, являющиеся наиболее успешными. Для работы с алгоритмами была использована библиотека Scikit-learn.

Алгоритм логистической регрессии применен с добавлением регуляризации l_2 , параметр регуляризации равен единице. Для поиска коэффициентов модели был использован модификация алгоритма Бройдена-Флетчера-Гольдфарба-Шанно, позволяющая минимизировать функцию потерь.

Для построения модели случайного леса использовано 20 базовых деревьев максимальной глубины с числом признаков для построения каждого равным квадратному корню от общего числа признаков. Для оценки качества разбиения в узлах деревьев использован критерий Джини. Для построения каждого дерева использована бутстреп подвыборка и out-of-bag подход [91].

Для модели градиентного бустинга использованы сто решающих деревьев максимальной глубины три. Скорость обучения – 0,1. Максимальное число признаков для построения каждого дерева равно общему числу признаков, а количество листьев не ограничено. Критерий качества разбиения каждого узла в деревьях – MSE по Фридману [93].

Таблица 3 – число признаков, отобранных каждым методом отбора и средние значения AUC моделей

№	Название метода отбора	число признаков	AUC (Logistic Regression)	AUC (Random Forest Classifier)	AUC (Gradient Boosting Classifier)
1	Непараметрические критерии	5	0,971±0,022	0,983±0,024	0,976±0,014
2	Интервальная оценка долей	6	0,971±0,046	0,981±0,018	0,975±0,006
3	Z-критерий для разности двух долей	5	0,945±0,008	0,971±0,017	0,951±0,028
4	Взаимная информация	12	0,967±0,052	0,980±0,036	0,968±0,034
5	ReliefF	9	0,978±0,035	0,980±0,028	0,981±0,042
6	Lasso	5	0,976±0,025	0,978±0,024	0,979±0,032
7	Random Forest	4	0,954±0,016	0,977±0,009	0,963±0,018
8	PIMP	3	0,961±0,012	0,979±0,011	0,969±0,023
9	Boruta	3	0,967±0,016	0,983±0,007	0,976±0,017
10	PermutationImportance	16 / 13 / 10	0,968±0,040	0,985±0,022	0,980±0,025
11	RFECV	18 / 18 / 15	0,973±0,013	0,987±0,015	0,981±0,009
12	Add-Del	14 / 12 / 8	0,962±0,043	0,968±0,038	0,963±0,033

Используемый метод должен быть достаточно точным, это требование мы формализовали следующим образом: $AUC \geq 98\%$. В то же время метод должен использовать минимальное возможное число предикторов (что в свою очередь определяет сложность модели, измеряемую числом признаков, и «защищает» от переобучения).

По результатам анализа оптимальным набором предикторов выбран набор, полученный в сочетании модели случайного леса и алгоритма Boruta из-за меньшего числа предикторов при высоком качестве прогнозирующей модели ($AUC = 0,983 \pm 0,007$).

Аналогично, для построения всех моделей системы использовались именно эти алгоритмы: Boruta для отбора признаков и случайный лес для моделей классификационных и регрессионных моделей.

Отметим, что для построения регрессионных моделей была использована модель RandomForestRegressor, с максимальным числом признаков для построения каждого дерева решений равным числу всех признаков. Остальные параметры повторяют гиперпараметры модели-классификатора.

Выводы

В третьей главе решена третья задача диссертационного исследования, разработать структуру СППВР, которая позволяет осуществлять многоэтапное прогнозирование здоровья ребенка, зачатого при помощи ВРТ, а именно:

- 1) определены входные и выходные параметры для каждого блока СППВР;
- 2) разработана структура СППВР;
- 3) определены все модели машинного обучения, которые должны составлять математическое и алгоритмическое обеспечение СППВР;
- 4) показан способ выбора алгоритма отбора признаков и алгоритма построения ML-модели на примере задачи прогнозирования наступления беременности.

4. АПРОБАЦИЯ СОЗДАННОГО МАТЕМАТИЧЕСКОГО И АЛГОРИТМИЧЕСКОГО ОБЕСПЕЧЕНИЯ

4.1 Описание наборов подготовленных данных

Набор полученных признаков

В разделе перечислены признаки, которые были получены из данных исследования с помощью алгоритмов, представленных во второй главе.

После того, как все признаки были получены, данные были собраны в одну таблицу, строками которой являлись ID пациентов, а каждый столбец отвечал за какой-то один из признаков. В таблице 4 перечислены все полученные признаки, их тип данных, распределение значений и количество объектов выборки, имеющих данный признак. В колонке «значение» для номинальных и порядковых признаков указаны коды значений (например, нет аборта в анамнезе - 0, есть аборт - 1) и количество пациентов с соответствующим значением признака (0 - 686 и 1 - 168). Для признаков, представленных в интервальной шкале или шкале отношений, указаны среднее $\pm$ стандартное отклонение (для нормально распределенных признаков) или медиана, первый и третий квартили (для признаков, чье распределение статистически значимо отличалось нормально).

Таблица 4 – Все признаки, полученные из данных

	название признака	тип данных	значения	количество объектов
акушерский анамнез				
1	аборт на раннем сроке	номинальный	0 – 686 1 – 168	854
2	внематочная беременность	номинальный	0 – 706 1 – 148	854
3	индуцированный поздний выкидыш	номинальный	0 – 845 1 – 9	854
4	неразвивающаяся беременность	номинальный	0 – 719 1 – 135	854
5	оперативные роды	номинальный	0 – 778 1 – 76	854
6	поздний самопроизвольный выкидыш	номинальный	0 – 835 1 – 19	854
7	преждевременные роды	номинальный	0 – 837	854

			1 – 17	
8	самопроизвольный выкидыш	номинальный	0 – 770 1 – 84	854
9	самостоятельные роды	номинальный	0 – 698 1 – 156	854
10	срочные роды	номинальный	0 – 625 1 – 229	854
наследственный анамнез				
11	гипертония / ишемия / инсульт / тромбозы у родственников	номинальный	0 – 619 1 – 235	854
12	онкология у родственников	номинальный	0 – 741 1 – 113	854
13	сахарный диабет у родственников	номинальный	0 – 718 1 – 136	854
хирургические вмешательства				
14	вмешательства на молочных железах	номинальный	0 – 819 1 – 35	854
15	вмешательства на щитовидной железе	номинальный	0 – 851 1 – 3	854
16	выскабливание полости матки	номинальный	0 – 746 1 – 108	854
17	гистерорезектоскопия	номинальный	0 – 837 1 – 17	854
18	гистероскопия	номинальный	0 – 190 1 – 664	854
19	лапараскопия	номинальный	0 – 391 1 – 463	854
20	лапаротомия (в том числе кесарево сечение)	номинальный	0 – 683 1 – 171	854
21	резекция в мочеполовой системе	номинальный	0 – 751 1 – 103	854
22	эмбриоскопия	номинальный	0 – 832 1 – 22	854
данные текущего здоровья				
23	аллергия	номинальный	0 – 680 1 – 174	854
24	возраст	интервальный	34 ± 5	854
25	вредные профессиональные факторы	номинальный	0 – 779 1 – 75	854
26	гормональное лечение в настоящее время	номинальный	0 – 684 1 – 170	854
27	диспансерный учет	номинальный	0 – 844 1 – 10	854
28	индекс массы тела	интервальный	Me = 28	854

			Q1 = 26, Q3 = 30	
29	курение	номинальный	0 – 829 1 – 25	854
30	нарушения цикла	номинальный	0 – 732 1 – 122	854
31	нормальная продолжительность цикла	номинальный	0 – 121 1 – 733	854
32	применялась гормональная контрацепция	номинальный	0 – 686 1 – 168	854
33	регулярные менструации без препаратов	номинальный	0 – 76 1 – 778	854
34	черепно-мозговые травмы	номинальный	0 – 830 1 – 24	854
хронические заболевания				
35	аденомиоз	номинальный	0 – 709 1 – 145	854
36	аднексит	номинальный	0 – 818 1 – 36	854
37	анемия	номинальный	0 – 840 1 – 14	854
38	аутоиммунный тиреоидит	номинальный	0 – 815 1 – 39	854
39	вегетососудистая дистония	номинальный	0 – 822 1 – 32	854
40	гастрит	номинальный	0 – 794 1 – 60	854
41	гиперпролактинемия	номинальный	0 – 813 1 – 41	854
42	гипотиреоз	номинальный	0 – 750 1 – 104	854
43	инфекционные заболевания (ВИЧ и/или гепатит В)	номинальный	0 – 826 1 – 28	854
44	миома матки	номинальный	0 – 790 1 – 64	854
45	миопия	номинальный	0 – 816 1 – 38	854
46	синдром поликистоза яичников	номинальный	0 – 847 1 – 7	854
47	тромбофилия и/или варикозное расширение вен	номинальный	0 – 837 1 – 17	854
48	фиброзно-кистозная мастопатия	номинальный	0 – 680 1 – 174	854
49	холецистит и/или панкреатит	номинальный	0 – 821	854

			1 – 33	
50	цистит и/или пиелонефрит	номинальный	0 – 800 1 – 54	854
51	эндометриоз	номинальный	0 – 794 1 – 60	854
52	эутиреоз	номинальный	0 – 820 1 – 34	854
53	количество заболеваний из представленных (признаки 35 – 52)	интервальный	Me = 2 Q1 = 1 Q3 = 3	854
текущая беременность				
54	количество плодов (1 либо 2)	интервальный	1 – 486 2 – 86	572
55	номер попытки ВРТ	порядковый	0 – 397 1 – 457	854
56	программа (ЭКО / криоперенос)	номинальный	0 – 328 1 – 526	854
57	гестационный сахарный диабет	номинальный	0 – 422 1 – 150	572
58	гипертензивные расстройства	номинальный	0 – 487 1 – 49	536
59	истмико-цервикальная недостаточность	номинальный	0 – 492 1 – 80	572
60	нарушения количества околоплодных вод	номинальный	0 – 495 1 – 41	536
61	предлежание плаценты	номинальный	0 – 483 1 – 53	536
62	преждевременный разрыв плодных оболочек	номинальный	0 – 541 1 – 31	572
63	прерывание беременности	номинальный	0 – 536 1 – 36	572
64	способ родов (плановое кесарево сечение / экстренное кесарево сечение / самостоятельные роды)	номинальный	0 – 196 1 – 153 2 – 187	536
65	срок родов	интервальный	Me = 38 Q1 = 37 Q3 = 39	536
стимуляция (для ЭКО)				
66	гистероскопия	номинальный	0 – 283 1 – 243	526
67	претритмент (нет / пайпель биопсия / гестагены /	номинальный	0 – 101 1 – 22 2 – 39	526

	заместительная гормональная терапия / эстроген / комбинированные оральные контрацептивы до протокола)		3 – 46 4 – 50 5 – 32	
68	протокол (длинный / короткий / естественный цикл)	номинальный	0 – 158 1 – 342 2 – 26	526
69	триггер (диферелин / овитрель / прегнил / хронический гонадотропин)	номинальный	0 – 37 1 – 190 2 – 66 3 – 233	526
результат стимуляции (для ЭКО)				
70	хорошая толщина эндометрия	номинальный	0 – 70 1 – 456	526
71	качество ооцитов	номинальный	0 – 49 1 – 477	526
оплодотворение (для ЭКО)				
72	количество зрелых ооцитов	шкала отношений	Me = 7 Q1 = 4 Q2 = 10	526
73	количество клеток на 3 сутки хорошего качества	шкала отношений	Me = 4 Q1 = 2 Q3 = 6	526
74	количество клеток на 5 сутки хорошего качества	шкала отношений	Me = 2 Q1 = 1 Q3 = 3	489
75	количество клеток на 6 сутки хорошего качества	шкала отношений	Me = 3 Q1 = 1 Q3 = 3	489
76	общее количество клеток на 6 сутки	шкала отношений	Me = 4 Q1 = 2 Q3 = 5	489
77	среда жидкость (Cook / CSSM / Sage / Vitrolife / Irvine)	номинальный	0 – 150 1 – 140 2 – 53 3 – 27 4 – 156	526
78	среда масло (Cook Oil / Irvine Oil)	номинальный	0 – 167 1 – 359	526
79	тип оплодотворения (ИКСИ / ЭКО / ИКСИ+ЭКО)	номинальный	0 – 370 1 – 93 2 – 63	526
перенос				
80	качество первого эмбриона	номинальный	0 – 53	854

			1 – 801	
81	качество второго эмбриона	номинальный	0 – 618 1 – 236	854
82	количество эмбрионов при подсадке (1 либо 2)	интервальный	0 – 273 1 – 581	854
83	перенос на третьи сутки или позже (для ЭКО)	номинальный	0 – 453 1 – 73	526
перенос (только для криопереноса)				
84	среда заморозки (Irvine / Kitazato)	номинальный	0 – 109 1 – 219	328
85	среда разморозки (Irvine / Kitazato)	номинальный	0 – 173 1 – 155	328
86	срок хранения (до 2 мес / от 2 - 12 мес / от 12 мес - 1825 сут)	порядковый	0 – 92 1 – 176 2 – 60	328
87	сутки переноса (5 либо 6)	интервальный	0 – 276 1 – 52	328

Таким образом, полный набор выделенных признаков включает в себя 87 признаков, которые характеризуют:

- 1) возраст, ИМТ женщины, ее акушерско-гинекологический анамнез, наследственный анамнез, перенесенные или хронические заболевания, оперативные вмешательства (те, которые могут оказывать влияние на репродуктивную систему и вынашивание беременности), вредные привычки, вредные особенности профессии (53 признака);
- 2) текущую ВРТ беременность и ее результат: номер попытки и программа ВРТ, количество плодов, возникшие осложнения беременности, прерывание беременности, срок и способ родов (12 признаков);
- 3) данные протокола ВРТ: данные стимуляции яичников, оплодотворения ооцитов, информацию об эмбрионах из крио (среды, срок хранения), качество и количество перенесенных эмбрионов, сутки переноса (22 признака).

Некоторые признаки являются одновременно и исходами для прогнозирования. Например, осложнения текущей беременности, срок и способ родов. Осложнения беременности предсказываются до вступления в протокол, с учетом

полученных значений предсказываются срок и способ родов, которые являются одними из прогностических параметров для моделей, связанных со здоровьем ребенка.

Набор значений множеств ответов

В разделе перечислены ответы (исходы), подлежащие прогнозированию, которые были получены из данных исследования с помощью алгоритмов, представленных во второй главе. Исходы перечислены в таблице 5.

Таблица 5 – Все исходы, их тип данных, распределение значений

№	название исхода	тип данных	значения	количество объектов
1	наступление беременности	номинальный	0 – 282 1 – 572	854
2	количество плодов (1 / 2)	бинарный	1 – 486 2 – 86	572
3	истмико-цервикальная недостаточность	номинальный	0 – 492 1 – 80	572
4	гипертензивные расстройства	номинальный	0 – 487 1 – 49	536
5	предлежание плаценты	номинальный	0 – 483 1 – 53	536
6	гестационный сахарный диабет	номинальный	0 – 422 1 – 150	572
7	нарушения количества околоплодных вод	номинальный	0 – 495 1 – 41	536
8	преждевременный разрыв плодных оболочек	номинальный	0 – 541 1 – 31	572
9	прерывание беременности	номинальный	0 – 536 1 – 36	572
10	способ родов (плановое / экстренное кесарево сечение / самостоятельные роды)	номинальный	0 – 536 1 – 36	572
10	срок родов	интервальный	Me = 38 Q1 = 37 Q3 = 39	536

11	группа здоровья ребенка (1-2, 3-4, 5)	номинальный	0 – 60 1 – 354 2 – 50	464
12	некоторые инфекционные и паразитарные болезни (A00-B99)	номинальный	0 – 457 1 – 7	464
13	новообразования (C00-D48)	номинальный	0 – 428 1 – 36	464
14	болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм (D50-D89)	номинальный	0 – 413 1 – 51	464
15	болезни эндокринной системы, расстройства питания и нарушения обмена веществ (E00-E90)	номинальный	0 – 429 1 – 35	464
16	болезни нервной системы (G00-G99)	номинальный	0 – 155 1 – 309	464
17	болезни глаза и его придаточного аппарата (H00-H59)	номинальный	0 – 289 1 – 175	464
18	болезни уха и сосцевидного отростка (H60-H95)	номинальный	0 – 445 1 – 19	464
19	болезни системы кровообращения (I00-I99)	номинальный	0 – 453 1 – 11	464
20	болезни органов дыхания (J00-J99)	номинальный	0 – 419 1 – 45	464
21	болезни органов пищеварения (K00-K93)	номинальный	0 – 368 1 – 96	464
22	болезни кожи и подкожной клетчатки (L00-L99)	номинальный	0 – 384 1 – 80	464
23	болезни костно-мышечной системы и соединительной ткани (M00-M99)	номинальный	0 – 409 1 – 55	464
24	болезни мочеполовой системы (N00-N99)	номинальный	0 – 434 1 – 30	464
25	отдельные состояния, возникающие в перинатальном периоде (P00-P96)	номинальный	0 – 424 1 – 40	464
26	врожденные аномалии, деформации и хромосомные нарушения (Q00-Q99)	номинальный	0 – 365 1 – 99	464

27	толщина эндометрия	номинальный	0 – 96 1 – 466	562
28	качество ооцитов	номинальный	0 – 47 1 – 515	562
29	количество фолликулов	шкала отношений	Me = 10 Q1 = 6 Q3 = 13	562
30	количество клеток на 3 сутки хо- рошего качества	шкала отношений	Me = 7 Q1 = 4 Q3 = 10	562
31	количество клеток на 5 сутки хо- рошего качества	шкала отношений	Me = 4 Q1 = 2 Q3 = 6	489
32	количество клеток на 6 сутки хорошего качества	шкала отношений	Me = 2 Q1 = 1 Q3 = 3	489
33	общее количество бластоцист на 6 сутки	шкала отношений	Me = 3 Q1 = 1 Q3 = 3	489

Исходы «толщина эндометрия» и «качество ооцитов» нужны для построения вспомогательных моделей, помогающих прогнозировать результат беременности и здоровье ребенка после проведения стимуляции яичников.

Для более удобного и компактного представления диагнозов, они были сгруппированы в классы МКБ-10, а для понятного представления о том, насколько сильно болен ребенок, врачом-педиатром был просмотрен весь список поставленных диагнозов и проставлены метки группы здоровья: 1-2 группа, 3-4 группа или 5, которую получает носитель диагноза.

Созданный в результате проделанной работы набор исходов включает в себя 33 различных исхода, подлежащих прогнозированию, которые представляют:

- 1) возможность наступления беременности и осложнения течения беременности: количество плодов, истмико-цервикальную недостаточность, гипертензивные расстройства, предлежание плаценты, гестационный сахарный диабет, нарушения количества околоплодных вод, преждевременный разрыв плодных оболочек (восемь исходов);

- 2) возникшие болезни детей в соответствии с классом МКБ-10 (15 исходов);
- 3) прерывание беременности, срок и способ родов, группу здоровья рожденного ребенка (три исхода);
- 4) результаты стимуляции яичников и оплодотворения ооцитов (семь исходов).

4.2 Значимые признаки и их распределение по моделям системы

В первую очередь до использования алгоритма отбора признаков была исследована их взаимная корреляция. Наличие коррелирующих признаков может отрицательно сказаться на скорости обучения модели (особенно для итерационных методов), снизить интерпретируемость и производительность модели, а также способствует переобучению. При наличии коррелированных признаков алгоритмы, основанные на оценке важности перестановок признака могут ошибаться, поскольку при удалении одного из них модель не снизит качество. Следовательно, важность перестановки для них может быть низкой и все они будут отброшены вне зависимости от полезности. Из совокупности взаимно коррелирующих признаков можно удалить все кроме одного без значимой потери информации.

Корреляция признаков с друг другом была отражена с помощью матрицы корреляций. Диапазон корреляций составил: $[-0,61; 0,52]$. Корреляция у большинства признаков очень слабая согласно критерию V Крамера ($< 0,2$). Обратная связь средней степени ($-0,61$ и $-0,54$), максимальная для данной матрицы, установлена между признаками «аденомиоз» - «нет заболеваний мочеполовой системы», признаками «нет хронических заболеваний» и «одно хроническое заболевание», что объясняется распределением значений в признаках. Также есть две положительные связи средней степени, значения которых $0,52$ для корреляции признаков «гормональное лечение в настоящее время» «гипотиреоз» и $0,51$ для признаков «внематочная беременность» - «лапаротомия». В целом связи не настолько сильные, чтобы исключать какие-либо признаки из-за дублирования.

Итоговые признаки, используемые для моделей, перечислены в таблице 6. В Приложении Б представлена таблица распределения значимых признаков по моделям СППВР. Номера признаков, отраженные в приложении, совпадают с номерами признаков в таблице 6.

Таблица 6 – значимые признаки для всех моделей СППВР

№	название признака	№	название признака
1	аборт на раннем сроке	35	эндометриоз
2	внематочная беременность	36	эутиреоз
3	неразвивающаяся беременность	37	количество заболеваний из представленных
4	оперативные роды	38	количество плодов (1 / 2)
5	преждевременные роды	39	номер попытки ВРТ
6	самопроизвольный выкидыш	40	программа (ЭКО / криоперенос)
7	срочные роды	41	гестационный сахарный диабет
8	гипертония / ишемия / инсульт / тромбозы у родственников	42	гипертензивные расстройства
9	сахарный диабет у родственников	43	истмико-цервикальная недостаточность
10	гистероскопия	44	нарушения количества околоплодных вод
11	лапараскопия	45	преждевременный разрыв плодных оболочек
12	лапаротомия (в том числе кесарево сечение)	46	способ родов (плановое кесарево сечение / экстренное кесарево сечение / самостоятельные роды)
13	резекция в мочеполовой системе	47	срок родов
14	возраст	48	претритмент (нет / пайпель биопсия / гестагены / заместительная гормональная терапия / эстроген / комбинированные оральные контрацептивы до протокола)
15	гормональное лечение в настоящее время	49	протокол (длинный / короткий / естественный цикл)

16	индекс массы тела	50	триггер (диферелин / овитрель / прегнил / хронический гонадотропин)
17	нарушения цикла	51	хорошая толщина эндометрия
18	применялась гормональная контрацепция	52	качество ооцитов
19	аденомиоз	53	количество фолликулов
20	аднексит	54	количество зрелых ооцитов
21	аллергия	55	среда жидкость (Cook / CSSM / Sage / Vitrolife / Irvine)
22	анемия	56	среда масло (Cook Oil / Irvine Oil)
23	аутоиммунный тиреоидит	57	тип оплодотворения (ИКСИ / ЭКО / ИКСИ+ЭКО)
24	вегетососудистая дистония	58	количество клеток на 3 сутки хорошего качества
25	гастрит	59	количество клеток на 5 сутки хорошего качества
26	гиперпролактинемия	60	количество клеток на 6 сутки хорошего качества
27	гипотиреоз	61	качество первого эмбриона
28	инфекционные заболевания (ВИЧ / гепатит В)	62	качество второго эмбриона
29	миома матки	63	количество эмбрионов при подсадке (1 / 2)
30	миопия	64	сутки переноса (3 / 5 / 6)
31	тромбофилия/варикозное расширение вен	65	среда заморозки (Irvine / Kitazato)
32	фиброзно-кистозная мастопатия	66	среда разморозки (Irvine / Kitazato)
33	холецистит / панкреатит	67	срок хранения (до 2 мес / от 2 - 12 мес / от 12 мес - 1825 сут)
34	цистит / пиелонефрит		

После проведения отбора признаков, удалось выделить из 87 признаков 67 значимых признаков для всех ML-моделей системы. Из которых:

- 4) вводятся в первом блоке СППВР: номер попытки и программа ВРТ, возраст, ИМТ женщины, ее акушерско-гинекологический анамнез, наследственный анамнез, перенесенные или хронические заболевания, оперативные вмешательства (те, которые могут оказывать влияние на репродуктивную систему и вынашивание беременности), вредные привычки, вредные особенности профессии (39 признаков);
- 5) являются одновременно прогнозируемыми откликами: количество плодов, возникшие осложнения беременности, срок и способ родов (семь признаков);
- 6) вводятся во втором блоке СППВР: претритмент, протокол, триггер (три признака);
- 7) являются одновременно прогнозируемыми результатами овариальной стимуляции: хорошая толщина эндометрия, хорошее качество ооцитов, количество фолликулов (три признака);
- 8) вводятся в третьем блоке СППВР: хорошее качество ооцитов, количество зрелых ооцитов, тип оплодотворения, среда жидкость, среда масло (пять признаков);
- 9) являются одновременно прогнозируемыми результатами оплодотворения: количество клеток на 3 сутки хорошего качества, количество бластоцист на 5 сутки хорошего качества, количество бластоцист на 6 сутки хорошего качества (три признака);
- 10) вводятся в четвертом блоке СППВР: количество эмбрионов при переносе, сутки переноса, качество перенесенных эмбрионов, срок заморозки / разморозки эмбрионов и срок их хранения для программы криопереносов (семь признаков);
- 11) дополнительный прогнозируемый результат в четвертом блоке – количество плодов (один признак).

4.3 Разработка прогностических моделей системы и оценка их качества

Первый блок «Прогнозирование результата отбора пациентов»

В таблице 7 представлены значения количества объектов, имеющихся для построения каждой модели машинного обучения в составе системы до и после балансировки данных.

Первая модель – бинарный классификатор предсказания наступления беременности, содержит максимальное для разрабатываемой системы число протоколов (854). Включает в себя все данные о протоколах ВРТ, включая данные тех протоколов, в результате которых беременность не наступила.

572 протокола закончились наступлением беременности и использованы для создания моделей прогнозирования таких осложнений текущей беременности, как истмико-цервикальная недостаточность, гестационный сахарный диабет, преждевременный разрыв плодных оболочек. Также 572 протокола использованы для построения регрессионной модели предсказания срока родов и бинарного классификатора прерывания беременности.

536 протоколов закончились родами и использованы для построения классификаторов предсказания следующих осложнений беременности: гипертензия, предлежание плаценты, нарушения количества околоплодных вод. Эти протоколы использованы для многоклассового классификатора предсказания способа родов: плановое кесарево сечение, экстренное, самостоятельные роды.

Для предсказания группы здоровья рожденного ребенка и возникновения заболеваний определенного класса были использованы данные о всех 464 рожденных в результате ВРТ детях. Для каждого ребенка имелась полная информация о матери и проведенном протоколе, а также данные о его здоровье от рождения до достижения трех лет.

Значимые признаки для каждого классификатора представлены в Приложении Б. Они включают в себя только данные текущей беременности, акушерского анамнеза, текущего здоровья и хронических заболеваний, осложнений текущей беременности, срока и способа родов.

Таблица 7 – количество объектов для построения всех моделей первого блока СППВР

№ модели	название модели	количество объектов для построения	количество объектов после балансировки
1	наступление беременности	854	0 - 588 1 – 572
2	истмико-цервикальная недостаточность	572	0 - 492 1 – 501
3	гипертензивные расстройства	536	0 - 487 1 – 495
4	предлежание плаценты	536	0 - 483 1 - 503
5	гестационный сахарный диабет	572	0 - 422 1 - 438
6	нарушения количества околоплодных вод	536	0 - 495 1 - 494
7	преждевременный разрыв плодных оболочек	572	0 - 541 1 - 542
8	способ родов (плановое кесарево сечение / экстренное кесарево сечение / самостоятельные роды)	536	0 - 349 1 – 335 2 - 345
9	прерывание беременности	572	0 - 536 1 - 543
10	группы здоровья (1-2, 3-4, 5)	464	0 - 246 1 – 249 2 – 247
11	некоторые инфекционные и паразитарные болезни (A00-B99)	464	0 - 457 1 - 458
12	новообразования (C00-D48)	464	0 - 428 1 - 432
13	болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм (D50-D89)	464	0 - 413 1 - 410
14	болезни эндокринной системы, расстройства питания и нарушения обмена веществ (E00-E90)	464	0 - 429 1 - 430
15	болезни нервной системы (G00-G99)	464	0 - 298 1 - 309
16	болезни глаза и его придаточного аппарата (H00-H59)	464	0 - 289 1 - 348

17	болезни уха и сосцевидного отростка (H60-H95)	464	0 - 445 1 - 445
18	болезни системы кровообращения (I00-I99)	464	0 - 453 1 - 453
19	болезни органов дыхания (J00-J99)	464	0 - 419 1 - 414
20	болезни органов пищеварения (K00-K93)	464	0 - 368 1 - 351
21	болезни кожи и подкожной клетчатки (L00-L99)	464	0 - 384 1 - 392
22	болезни костно-мышечной системы и соединительной ткани (M00-M99)	464	0 - 409 1 - 408
23	болезни мочеполовой системы (N00-N99)	464	0 - 433 1 - 426
24	отдельные состояния, возникающие в перинатальном периоде (P00-P96)	464	0 - 424 1 - 414
25	врожденные аномалии, деформации и хромосомные нарушения (Q00-Q99)	464	0 - 365 1 - 361

Для подбора гиперпараметров классификаторов была использована функция, созданная на основе библиотеки `Hyperopt` и описанная в разделе 1.7. Для построения всех моделей случайного леса использовано от 20 до 40 базовых деревьев максимальной глубины с числом признаков для построения каждого равным квадратному корню от общего числа признаков для классификаторов и с числом признаков равным общему числу признаков для регрессионной модели. Для оценки качества разбиения в узлах деревьев использован критерий Джини. Значение метрик ROC-AUC и доли правильных ответов для моделей первого блока СППВР представлены в таблице 8.

Таблица 8 – значения метрик качества бинарных классификаторов первого блока СППВР

№ модели	название модели	ROC-AUC	доля правильных ответов
1	наступление беременности	0,884±0,007	0,792±0,029
2	истмико-цервикальная недостаточность	0,777±0,013	0,861±0,046
3	гипертензивные расстройства	0,841±0,022	0,912±0,013
4	предлежание плаценты	0,819±0,015	0,883±0,012
5	гестационный сахарный диабет	0,736±0,056	0,782±0,032

6	нарушения количества околоплодных вод	0,888±0,008	0,814±0,021
7	преждевременный разрыв плодных оболочек	0,927±0,034	0,861±0,017
8	способ родов (плановое кесарево сечение / экстренное кесарево сечение / самостоятельные роды)	-	0,716±0,065
9	прерывание беременности	0,912±0,054	0,840±0,068
10	группы здоровья (1-2, 3-4, 5)	-	0,684±0,072
11	некоторые инфекционные и паразитарные болезни (A00-B99)	0,988±0,004	0,961±0,013
12	новообразования (C00-D48)	0,925±0,013	0,880±0,043
13	болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм (D50-D89)	0,904±0,026	0,842±0,053
14	болезни эндокринной системы, расстройства питания и нарушения обмена веществ (E00-E90)	0,887±0,022	0,812±0,047
15	болезни нервной системы (G00-G99)	0,743±0,062	0,672±0,074
16	болезни глаза и его придаточного аппарата (H00-H59)	0,687±0,018	0,641±0,056
17	болезни уха и сосцевидного отростка (H60-H95)	0,893±0,007	0,872±0,012
18	болезни системы кровообращения (I00-I99)	0,988±0,005	0,960±0,052
19	болезни органов дыхания (J00-J99)	0,903±0,020	0,836±0,060
20	болезни органов пищеварения (K00-K93)	0,766±0,042	0,695±0,042
21	болезни кожи и подкожной клетчатки (L00-L99)	0,789±0,061	0,717±0,068
22	болезни костно-мышечной системы и соединительной ткани (M00-M99)	0,898±0,038	0,826±0,044
23	болезни мочеполовой системы (N00-N99)	0,914±0,027	0,861±0,042
24	отдельные состояния, возникающие в перинатальном периоде (P00-P96)	0,915±0,022	0,867±0,027
25	врожденные аномалии, деформации и хромосомные нарушения (Q00-Q99)	0,802±0,034	0,740±0,056

Показатели качества регрессионной модели срока родов: MSE = 1,540±0,341; MAE = 0,901±0,095; R² = 0,436±0,111.

В результате проделанной работы первый блок СППВР содержит 25 классификаторов и одну регрессионную модель. В блоке используется 47 значимых признаков. Их распределение по ML-моделям СППВР можно посмотреть в приложении Б. Первый блок позволяет получить прогнозирование вероятности наступления беременности в результате проведения протокола ВРТ, прогнозирование отдельно для одноплодной и многоплодной беременности возможных осложнений ее течения, срок и способ родов, а также прогноз группы здоровья рожденного ребенка с учетом вероятных заболеваний.

Второй блок «Прогнозирование результата овариальной стимуляции»

Второй блок включает в себя 29 моделей машинного обучения: 15 ML-моделей из первого блока (№ 5, 6, 7, 8, 9, 10, 14, 15, 17, 19, 20, 21, 22, 23, 25), 13 новых классификаторов (№ 27 – 37, 39, 40), одну новую регрессионную модель (№ 38).

В таблице 9 представлены значения количества объектов, имеющихся для построения каждой модели в составе второго блока системы до и после балансировки данных. На гестационный сахарный диабет были обследованы все женщины, беременность которых наступила. Нарушение количества околоплодных вод отслежено только у тех женщин, беременность которых закончилась родами. Для отслеживания здоровья детей использовались 328 объектов данных – дети, рожденные в результате 314 беременностей по протоколам ЭКО. Прогноз результатов стимуляции построен на данных всех женщин, вступивших в протокол ЭКО.

Новые классификаторы № 27 - 37 повторяют аналогичные классификаторы первого блока, но используют для построения предсказания не только данные анамнеза, но и данные стимуляции (претритмент, протокол, триггер, толщину эндометрия, качество ооцитов).

Бинарные классификаторы № 39 и № 40 предсказывают ожидаемую толщину эндометрия (хорошую: 8-15 мм или плохую: менее 8, более 15 мм) и ожидаемое качество ооцитов (хорошее / плохое). Для их построения используются

данные анамнеза и стимуляции. Регрессионная модель № 38 прогнозирует ожидаемое количество фолликулов, основываясь на данных анамнеза и стимуляции яичников пациентки. Значимые признаки для каждой модели представлены в Приложении Б.

Таблица 9 – количество объектов для построения всех моделей второго блока СППВР

№ модели	название модели	количество объектов для построения	количество объектов после балансировки
27	наступление беременности (2)	538	0 – 335 1 – 338
28	истмико-цервикальная недостаточность (2)	338	0 – 270 1 – 272
29	гипертензивные расстройства (2)	314	0 – 279 1 – 280
30	предлежание плаценты (2)	314	0 – 274 1 – 258
31	группы здоровья (1-2, 3-4, 5) (2)	328	0 – 202 1 – 204 2 – 201
32	некоторые инфекционные и паразитарные болезни (A00-B99) (2)	328	0 – 323 0 – 320
33	новообразования (C00-D48) (2)	328	0 – 283 1 – 285
34	болезни нервной системы (G00-G99) (2)	328	0 – 185 1 – 186
35	болезни глаза и его придаточного аппарата (H00-H59) (2)	328	0 – 207 1 – 204
36	болезни мочеполовой системы (N00-N99) (2)	328	0 – 276 1 – 277
37	врожденные аномалии, деформации и хромосомные нарушения (Q00-Q99) (2)	328	0 – 252 1 – 243
38	количество фолликулов	562	-
39	толщина эндометрия	562	0 – 494 1 – 466
40	качество ооцитов	562	0 – 512 1 – 515

Для подбора гиперпараметров классификаторов была использована функция, созданная на основе библиотеки `Hyperopt` и описанная в разделе 1.7. Для построения всех моделей случайного леса использовано от 20 до 40 базовых деревьев максимальной глубины с числом признаков для построения каждого равным квадратному корню от общего числа признаков для классификаторов и с числом признаков равным общему числу признаков для регрессионной модели. Для оценки качества разбиения в узлах деревьев использован критерий Джини. Значение метрик ROC-AUC и доли правильных ответов для моделей второго блока СППВР представлены в таблице 10.

Таблица 10 – значения метрик качества бинарных классификаторов второго блока СППВР

№ модели	название модели	ROC-AUC	доля правильных ответов
27	наступление беременности (2)	0,914±0,036	0,883±0,030
28	истмико-цервикальная недостаточность (2)	0,943±0,030	0,880±0,018
29	гипертензивные расстройства (2)	0,874±0,006	0,856±0,040
30	предлежание плаценты (2)	0,914±0,016	0,848±0,021
31	группы здоровья (1-2,3-4,5)	-	0,763±0,023
32	некоторые инфекционные и паразитарные болезни (A00-B99) (2)	0,970±0,024	0,965±0,028
33	новообразования (C00-D48) (2)	0,955±0,015	0,894±0,032
34	болезни нервной системы (G00-G99) (2)	0,922±0,002	0,825±0,015
35	болезни глаза и его придаточного аппарата (H00-H59) (2)	0,960±0,027	0,893±0,052
36	болезни мочеполовой системы (N00-N99) (2)	0,924±0,002	0,864±0,015
37	врожденные аномалии, деформации и хромосомные нарушения (Q00-Q99) (2)	0,832±0,027	0,772±0,052
39	толщина эндометрия	0,830±0,018	0,772±0,016
40	качество ооцитов	0,894±0,018	0,815±0,039

Показатели качества регрессионной модели: $MSE = 3,812 \pm 0,879$; $MAE = 2,234 \pm 0,212$; $R^2 = 0,467 \pm 0,210$.

В результате проделанной работы для второго блока СППВР было разработано 14 ML-моделей, использующих как признаки из первого блока СППВР, так и признаки, связанные со стимуляцией яичников (шесть признаков). Результат использования блока позволяет получить прогнозирование вероятности наступления беременности в результате проведения протокола ВРТ, прогнозирование отдельно для одноплодной и многоплодной беременности возможных осложнений ее течения, срок и способ родов, а также прогноз группы здоровья рожденного ребенка с учетом вероятных заболеваний на основе данных анамнеза и стимуляции яичников. Также позволяет вывести три параметра, получаемые после стимуляции: ожидаемое число фолликулов, качество эндометрия и качество ооцитов. Модели, дополняющие классификаторы первого блока новыми признаками, имеют более высокое качество, чем исходные.

Третий блок «Прогнозирование результата оплодотворения»

Третий блок включает в себя модели № 41 – 44: «количество клеток на 3 сутки хорошего качества», «количество бластоцист на 5 сутки хорошего качества», «количество бластоцист на 6 сутки хорошего качества», «общее количество бластоцист на 6 сутки».

Все модели являются регрессионными, используют для построения признаки, связанные с результатами стимуляции яичников и проведением оплодотворения ооцитов. Значимые признаки для каждой модели представлены в Приложении Б, количество объектов для построения моделей – в таблице 11. Для модели № 41 используется информация обо всех протоколах ЭКО, а для остальных моделей только то число объектов, для которых перенос бластоцист осуществлялся на 5 или 6 сутки.

Таблица 11 – количество объектов для построения всех моделей третьего блока СППВР

№ модели	Название модели	количество объектов для построения
41	количество клеток на 3 сутки хорошего качества	562

42	количество бластоцист на 5 сутки хорошего качества	489
43	количество бластоцист на 6 сутки хорошего качества	489
44	общее количество бластоцист	489

Для подбора гиперпараметров моделей была использована функция, созданная на основе библиотеки `Hyperopt` и описанная в разделе 1.7. Для построения всех моделей случайного леса использовано от 20 до 40 базовых деревьев максимальной глубины с числом признаков для построения каждого равным общему числу признаков. Для оценки качества разбиения в узлах деревьев использован критерий Джини. Значение метрик ROC-AUC и доли правильных ответов для моделей третьего блока СППВР представлены в таблице 12.

Таблица 12 – значение метрик качества регрессионных моделей третьего блока

№ модели	название модели	MSE	MAE	R ²
41	количество клеток на 3 сутки хорошего качества	4,685±1,453	1,585±0,210	0,472±0,085
42	количество бластоцист на 5 сутки хорошего качества	3,187±0,429	1,304±0,066	0,505±0,172
43	количество бластоцист на 6 сутки хорошего качества	3,091±0,421	1,299±0,085	0,550±0,115
44	общее количество бластоцист	3,241±0,446	1,328±0,093	0,580±0,207

В результате проделанной работы третий блок СППВР содержит четыре регрессионные модели. Результат использования блока позволяет получить прогнозирование общего количества бластоцист и количества бластоцист высокого качества.

Четвертый блок «Прогнозирование результата переноса»

Четвертый блок включает в себя 44 модели: 35 ML-моделей из предыдущих блоков и 19 новых ML-моделей (№ 45 – 63).

Новые ML-модели № 45 – 54, 56 – 62 повторяют аналогичные модели первого и второго блоков СППВР, но используют для построения предсказания не

только значимые признаки из первых трех блоков СППВР, но и признаки, связанные с четвертым вариативным этапом ВРТ – переносом эмбриона в полость матки.

Прогнозирование учитывает используемую программу ВРТ: модели № 45 – 55 созданы для программы ЭКО (учитывают данные стимуляции яйцеклеток, оплодотворения ооцитов и переноса эмбрионов), модели № 56 – 63 созданы для программы криопереноса (учитывают срок и среду хранения эмбрионов, данные переноса). Количество объектов данных для построения до и после балансировки представлены в таблице 13.

Таблица 13 – количество объектов для построения всех моделей четвертого блока СППВР

№ модели	название модели	количество объектов для построения	количество объектов после балансировки
Четвертый-а блок (ЭКО)			
45	наступление беременности (3)	538	0 – 335 1 – 338
46	прерывание беременности (3)	338	0 – 343 1 – 345
47	группа здоровья (1-2,3-4,5) (3)	328	0 – 202 1 – 204 2 – 201
48	новообразования (C00-D48) (3)	328	0 – 283 1 – 285
49	болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм (D50-D89) (2)	328	0 – 293 1 – 286
50	болезни глаза и его придаточного аппарата (H00-H59) (2)	328	0 – 225 1 – 232
51	болезни органов дыхания (J00-J99) (2)	328	0 – 297 1 – 300
52	болезни костно-мышечной системы и соединительной ткани (M00-M99) (2)	328	0 – 287 1 – 278
53	болезни мочеполовой системы (N00-N99) (2)	328	0 – 309 1 – 311
54	отдельные состояния, возникающие в перинатальном периоде (P00-P96) (2)	328	0 – 300 1 – 299

55	количество плодов (ЭКО)	338	1 – 280 2 – 278
Четвертый-б блок (криоперенос)			
56	наступление беременности (4)	234	0 – 176 1 – 175
57	прерывание беременности (3)	234	0 – 181 1 – 185
58	группа здоровья (1-2,3-4,5) (4)	136	0 – 125 1 – 119 2 – 116
59	болезни эндокринной системы, расстройства питания и нарушения обмена веществ (E00-E90) (2)	136	0 – 126 1 – 123
60	болезни органов дыхания (J00-J99) (3)	136	0 – 122 1 – 122
61	болезни костно-мышечной системы и соединительной ткани (M00-M99) (3)	136	0 – 122 1 – 127
62	отдельные состояния, возникающие в перинатальном периоде (P00-P96) (4)	136	0 – 124 1 – 127
63	количество плодов (криоперенос)	234	0 – 206 1 – 206

Для подбора гиперпараметров классификаторов была использована функция, созданная на основе библиотеки Nupurort и описанная в разделе 1.7. Для построения всех моделей случайного леса использовано от 20 до 40 базовых деревьев максимальной глубины с числом признаков для построения каждого равным квадратному корню от общего числа признаков для классификаторов и с числом признаков равным общему числу признаков для регрессионной модели. Для оценки качества разбиения в узлах деревьев использован критерий Джини. Значение метрик ROC-AUC и доли правильных ответов для моделей четвертого блока СППВР представлены в таблице 14.

Таблица 14 – значения метрик качества бинарных классификаторов четвертого блока СППВР

№ модели	название модели	ROC-AUC	доля правильных ответов
Четвертый-а блок (ЭКО)			
45	наступление беременности (3)	0,921±0,022	0,896±0,026

46	прерывание беременности (3)	0,981±0,019	0,912±0,074
47	группа здоровья (1-2,3-4,5) (3)	-	0,807±0,053
48	новообразования (C00-D48) (3)	0,981±0,013	0,904±0,048
49	болезни крови, кроветворных органов и отдельные нарушения, вовлекающие иммунный механизм (D50-D89) (2)	0,966±0,006	0,970±0,040
50	болезни глаза и его придаточного аппарата (H00-H59) (2)	0,960±0,034	0,903±0,047
51	болезни органов дыхания (J00-J99) (2)	0,971±0,021	0,946±0,054
52	болезни костно-мышечной системы и соединительной ткани (M00-M99) (2)	0,990±0,010	0,963±0,012
53	болезни мочеполовой системы (N00-N99) (2)	0,989±0,056	0,903±0,029
54	отдельные состояния, возникающие в перинатальном периоде (P00-P96) (2)	0,993±0,009	0,962±0,016
55	количество плодов (ЭКО)	0,779±0,012	0,707±0,082
Четвертый-б блок (криоперенос)			
56	наступление беременности (4)	0,981±0,019	0,912±0,074
57	прерывание беременности (3)	0,971±0,026	0,880±0,063
58	группа здоровья (1-2,3-4,5) (4)	-	0,795±0,042
59	болезни эндокринной системы, расстройства питания и нарушения обмена веществ (E00-E90) (2)	0,977±0,002	0,953±0,003
60	болезни органов дыхания (J00-J99) (3)	0,966±0,022	0,902±0,044
61	болезни костно-мышечной системы и соединительной ткани (M00-M99) (3)	0,920±0,013	0,897±0,062
62	отдельные состояния, возникающие в перинатальном периоде (P00-P96) (4)	0,937±0,021	0,827±0,099
63	количество плодов (криоперенос)	0,786±0,045	0,715±0,062

В результате проделанной работы для четвертого блока было разработано 19 новых ML-моделей. Результат использования блока позволяет получить прогнозирование вероятности наступления беременности в результате проведения протоколов ЭКО и криопереноса, прогнозирование количества плодов, с учетом которого осуществлено дальнейшее прогнозирование возможных осложнений течения беременности, срока и способа родов, а также группы здоровья рожден-

ного ребенка с учетом вероятных заболеваний. Модели, дополняющие классификаторы предыдущих блоков новыми признаками, имеют более высокое качество, чем исходные.

4.4 Разработка пользовательского программного приложения для работы с системой

Описание интерфейса приложения

Все признаки заполняются с помощью элемента интерфейса переключатель (radioButton), позволяющего выбрать один из допустимых значений признака («да»/ «нет» для большинства).

После заполнения признаков и нажатия на кнопку «Прогноз по анамнезу», в выделенном поле появятся результаты прогнозирования: вероятность наступления беременности в процентах и, отдельно для многоплодной и одноплодной беременности, прогнозируемые осложнения беременности, срок и способ родов, группа здоровья ребенка с указанием классов диагнозов по МКБ-10, по которым возможно возникновение проблем со здоровьем.

Если в первом блоке была выбрана программа ЭКО, то доступны для получения предсказания блоки 2 и 3. Иначе, при попытке получить прогноз, выводятся сообщения о непредусмотренности данных этапов для программы криопереноса.

Второй блок «Прогнозирование после стимуляции» использует для построения выводов не только данные из первого блока, но и данные стимуляции: возможные варианты претритмента (предусмотрен выбор одного варианта, возможно, в сочетании с гистероскопией), протокол стимуляции и используемый триггер. Все признаки заполняются с помощью элемента интерфейса переключатель. После заполнения признаков и нажатия на кнопку «Прогноз после стимуляции», в выделенном поле появятся все перечисленные в блоке 1 результаты прогнозирования с учетом поправки на данные стимуляции, а также ожидаемые результаты стимуляции: ожидаемое число фолликулов, толщина эндометрия (норма / отклонение) и качество ооцитов (хорошее / плохое).

В третьем блоке «Прогнозирование результатов оплодотворения» пользователю предлагается ввести количество зрелых ооцитов (оно может быть меньше ожидаемого числа фолликул после стимуляции или, наоборот, больше, при добавлении донорских или ранее криоконсервированных ооцитов пациентки), а также с помощью переключателей выбрать тип оплодотворения (ЭКО, ИКСИ, ЭКО+ИКСИ) и среды для проведения оплодотворения. После заполнения признаков и нажатия на кнопку «Прогноз результата оплодотворения», в выделенном поле появятся результаты предсказания количества клеток на 3 сутки (по качеству лучших, чем 6В), количество бластоцист на 5 сутки (лучших по качеству, чем 3ВВ), количество бластоцист на 5-6 сутки (лучших по качеству, чем 3ВВ) и общее количество бластоцист на 5-6 сутки.

Четвертый блок «Прогнозирование после переноса» можно разделить на две составные части в зависимости от выполняемого протокола: перенос после проведения ЭКО или перенос эмбрионов из крио. Пользователь видит признаки, используемые для обоих прогнозов, но для заполнения должен выбрать соответствующий блок. Для программы ЭКО признаки включают в себя количество эмбрионов (1/2), выбор суток переноса (третьи/ позже) и выбор качества первого и второго эмбриона (или единственного эмбриона). Для программы криопереноса кроме количества и качества эмбрионов нужно указать срок хранения криоконсервированных эмбрионов, среду их заморозки и разморозки и сутки переноса (5/6). Все признаки заполняются с помощью элемента интерфейса переключатель. После заполнения признаков и нажатия на кнопку «Прогноз после переноса», в выделенном поле появятся результаты работы моделей, аналогичные моделям первого блока, но имеющие скорректированный прогноз с учетом влияния всех используемых в системе признаков.

Пример работы приложения и его интерфейс представлены в Приложении Г. Приложение представляет собой единую форму, на которой расположены все четыре блока системы.

Разработка приложения

Приложение выполнено в виде исполняемого файла «IVF.exe». Для его создания было осуществлено следующее:

- 1) Сохранение всех моделей машинного обучения в отдельные файлы расширения .pkl;
- 2) Разработка интерфейса программы, сохраненного в виде файла «GUI2_all.ui»;
- 3) Создание файла программы на Python 3.7;
- 4) Сборка python-приложения в один файл при помощи программы PyInstaller. Для этого все файлы приложения помещаются в одну папку и запускается из консоли команда вида: «pyinstaller -F IVF.pyw».

Для работы с исполняемым файлом все файлы моделей машинного обучения и файл интерфейса должны лежать в той же папке, где и исполняемый файл.

4.5 Апробация СППВР на новых данных

Тестирование итоговой версии информационной системы, представленной в виде программного приложения, было проведено АО «ЦСМ» г. Екатеринбург.

Валидация (концептуальное тестирование) показало, что общая структура приложения соответствует ожидаемой, учтены все аспекты решаемой задачи, приложение понятно для пользователя и удобно в работе.

Для проверки на корректность работы системы (верификации) использовались данные о женщинах, прошедших 635 протоколов ВРТ с конца 2018 года по начало 2021 года, из которых беременность наступила в 435 случаях и завершилась родами в 335 случаях. Количество многоплодных беременностей – 72, из них 65 завершились родами. Информация об этих женщинах не использовалась при разработке системы, протоколы ВРТ были выбраны случайным образом. Всего было рождено 400 детей (355 детей 1-2 групп здоровья, 42 ребенка 3-4 групп здоровья, 3 детей 5 группы).

Группа здоровья ребенка оценена по диагнозам при рождении. Получены следующие результаты:

- 1) наступление беременности предсказано верно в 550 случаях (86,6 %);

- 2) способ родов предсказан верно в 218 случаях (65,1 %);
- 3) срок родов предсказан в неделях верно в 274 случаях (81,8 %) (ошибка предсказания ± 1 неделя);
- 4) осложнения текущей беременности предсказаны верно: истмико-цервикальная недостаточность – 391 случай (89,9 %), гипертензивные расстройства – 318 случаев (94,9 %), предлежание плаценты – 284 случаев (84,7 %), гестационный сахарный диабет – 346 случаев (79,5 %), нарушения количества околоплодных вод – 287 случаев (85,7 %), преждевременный разрыв плодных оболочек – 305 случаев (91,0 %);
- 5) группы здоровья рожденных детей предсказаны верно в 342 случаях (85,5 %).

Результаты прогнозирования представлены для каждого из перечисленных исходов в виде таблицы сопряженности (табл. 15) и в виде ROC-кривой (рис. 4.1).

Таблица 15 – таблицы сопряженности для прогнозируемых исходов

а) наступление беременности:

		реальные данные		всего (прогноз)
		беременность наступила	беременности нет	
прогноз	беременность наступила	402	52	454
	беременности нет	33	148	181
всего (реальные данные)		435	200	

чувствительность – 0,92%

специфичность – 0,74%

б) способ родоразрешения:

		реальные данные			всего (про- гноз)
		срочные роды	плановое кесарево сечение	экстренное кесарево сечение	
п р о	срочные роды	77	1	42	120

	плановое кесарево сечение	16	87	21	124
	экстренное кесарево сечение	30	7	54	91
всего (реальные данные)		123	95	117	

в) истмико-цервикальная недостаточность:

		реальные данные		всего (прогноз)
		есть истмико- цервикальная недостаточность	нет истмико-цер- викальной недостаточности	
прогноз	есть истмико-цер- викальная недостаточность	48	25	73
	нет истмико-цер- викальной недостаточности	19	343	362
всего (реальные данные)		67	368	

чувствительность – 0,72%

специфичность – 0,93%

г) гипертензивные расстройства:

		реальные данные		всего (прогноз)
		есть гипертензив- ные расстройства	нет гипертензив- ных расстройств	
прогноз	есть гипертензив- ные расстройства	32	9	41
	нет гипертензив- ных расстройств	8	286	294
всего (реальные данные)		40	295	

чувствительность – 0,80%

специфичность – 0,97%

д) предлежание плаценты:

		реальные данные		всего (прогноз)
		есть предлежани- е плаценты	нет предлежа- ния плаценты	
прогноз	есть предлежание плаценты	27	40	67
	нет предлежания плаценты	11	257	266
всего (реальные данные)		38	297	

чувствительность – 0,71%

специфичность – 0,87%

е) гестационный сахарный диабет:

		реальные данные		всего (прогноз)
		есть гестаци- онный сахар- ный диабет	нет гестацион- ного сахарного диабета	
прогноз	есть гестационный сахарный диабет	85	65	150
	нет гестационного сахарного диабета	24	261	285
всего (реальные данные)		109	326	

чувствительность – 0,78%

специфичность – 0,80%

ж) нарушения количества околоплодных вод:

		реальные данные		всего (прогноз)
		есть наруше- ния количества	нет нарушения количества око- лоплодных вод	

		ОКОЛОПЛОДНЫХ ВОД		
прогноз	есть нарушения количества около- плодных вод	18	43	61
	нет нарушения ко- личества около- плодных вод	5	269	274
всего (реальные данные)		23	312	

чувствительность – 0,78%

специфичность – 0,86%

з) преждевременный разрыв плодных оболочек:

		реальные данные		всего (прогноз)
		есть прежде- временный разрыв плод- ных оболочек	нет преждевре- менного раз- рыва плодных оболочек	
прогноз	есть преждевре- менный разрыв плодных оболочек	21	35	56
	нет преждевре- менного разрыва плодных оболочек	6	373	379
всего (реальные данные)		27	408	

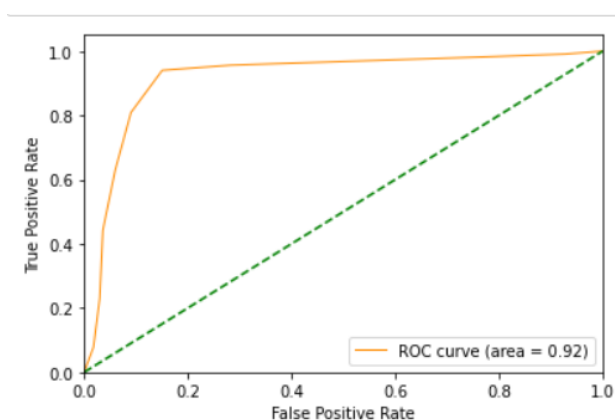
чувствительность – 0,78%

специфичность – 0,91%

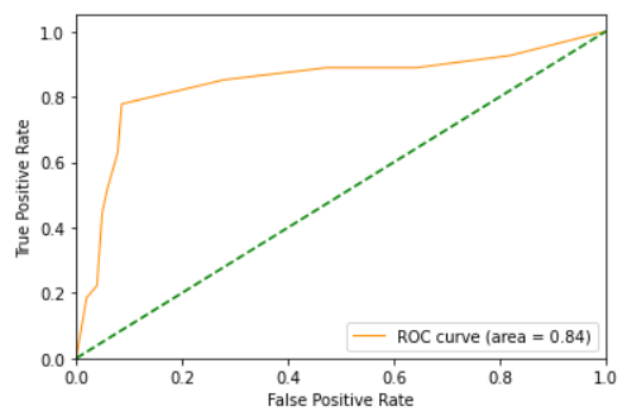
и) группы здоровья рожденных детей:

	реальные данные			всего (прогноз)
	1-2 группа	3-4 группа	5 группа	

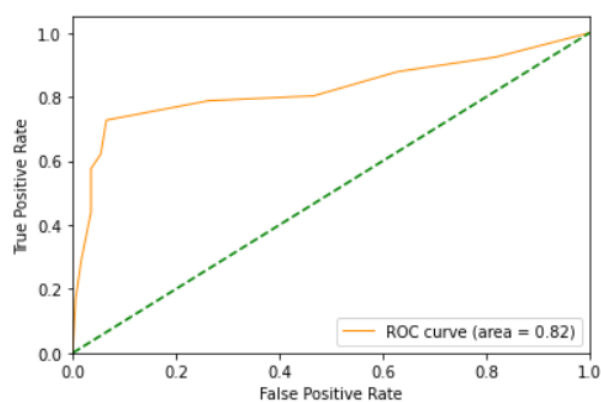
прогноз	1-2 группа	309	3	0	312
	3-4 группа	38	32	2	72
	5 группа	8	7	1	16
всего (реальные данные)		355	42	3	



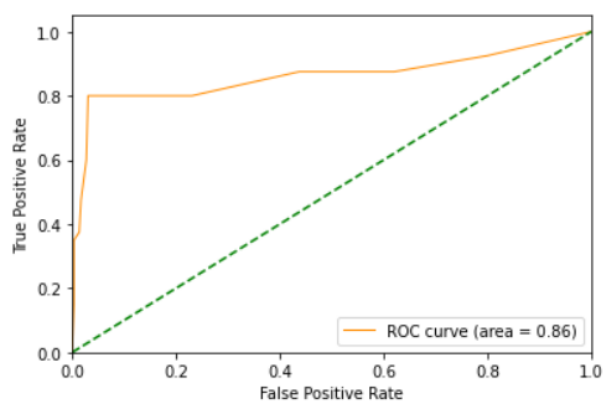
а) наступление беременности



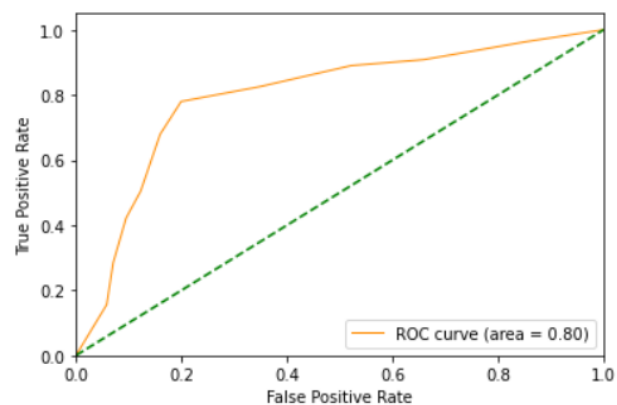
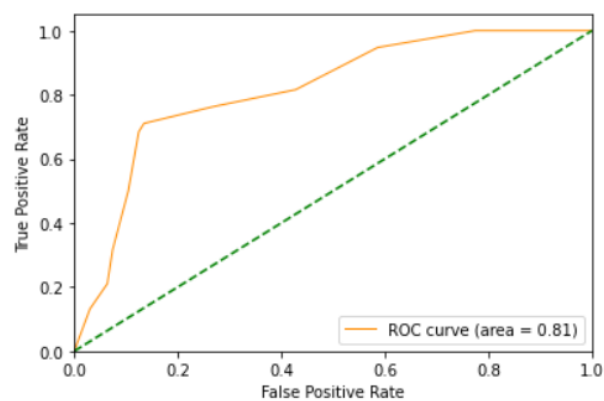
б) преждевременный разрыв плодных оболочек



в) истмико-цервикальная недостаточность

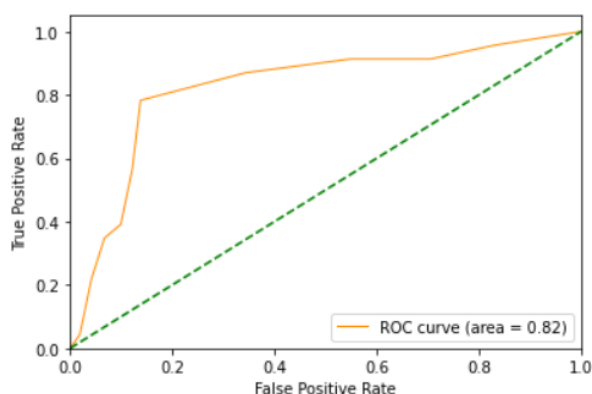


г) гипертензивные расстройства



д) предлежание плаценты

е) гестационный сахарный диабет



ж) нарушения количества околоплодных вод

Рисунок 4.1 – ROC-кривые для прогнозируемых исходов

Пример апробации на данных прошлых лет

Пациентка возрастом 38 лет, вторая попытка ЭКО, индекс массы тела равен 26. Из хронических заболеваний у пациентки имеется эндометриоз, аденомиоз, гипотериоз. Известно, что ближайшие родственники имеют сахарный диабет и гипертонию. Информация из акушерско-гинекологического анамнеза, хирургических вмешательств представлена на рисунке 4.2.

Регрессирующая беременность в 20 недель-индуцированный поздний выкидыш самопроизвольный выкидыш до 12 недель, ЛДВ	Кесарево сечение, послеоперационная гематома, переливание свежзамороженной плазмы
Срочные роды	Операция по поводу варизного расширения вен левой ноги (кроссэктомия)
Срочные оперативные роды	Операция по поводу варикозного расширения вен правой ноги (кроссэктомия).
мед. аборт в сроке до 12 недель	Гистероскопия. Аденомиоз. Рубец на матке
аборт в сроке до 12 недель	Гистероскопия, мва - ЦСМ

Рисунок 4.2 – а) пример акушерско-гинекологического анамнеза пациентки; б) пример хирургических вмешательств пациентки

После применения алгоритмов предобработки текстовых данных, описанных во второй главе диссертационного исследования, у данной пациентки были выделены следующие значимые признаки из данных акушерско-гинекологического анамнеза и хирургических вмешательств:

1) аборт на раннем сроке;

- 2) самопроизвольный выкидыш;
- 3) срочные роды;
- 4) кесарево сечение;
- 5) неразвивающаяся беременность;
- 6) гистероскопия.

Известно, что в результате проведенного длинного протокола, удалось получить четыре фолликула и три зрелых ооцита. После оплодотворения методом ИКСИ получился один эмбрион низкого качества на шестые сутки. В результате переноса этого эмбриона наступила беременность, осложнившаяся истмико-цервикальной недостаточностью и завершившаяся плановым кесаревым сечением в 40 недель беременности. На момент родов ребенок не имел заболеваний и был выписан из роддома вместе с матерью в плановом порядке.

После ввода всех имеющихся данных в приложение, получился прогноз, показанный на рисунке 4.3.

Программа
☒ ЭКО ☐ криопереносы
 Номер попытки ВРТ: 2
 ИМТ: 26
 Возраст: 38
 Акушерский анамнез:
☒ Аборт на раннем сроке
☐ Внезапная беременность
☒ Кесарево сечение
☒ Неразвивающаяся беременность
☐ Преждевременные роды
☒ Самопроизвольный выкидыш
☒ Срочные роды
 Хирургические вмешательства:
☒ Гистероскопия
☐ Лапароскопия
☐ Лапаротомия
☐ Резекция в мочеполовой системе
☐ Применялась гормональная контрацепция
☐ Есть нарушения цикла
☐ Гормональное лечение в наст. время
 Наличие у ближайших родственников:
☒ Сахарный диабет
☒ Гипертония/ ишемия/ инсульт/ тромбозы

Хронические заболевания
☒ Аденомиоз
☐ Аднексит
☐ Аллергия
☐ Анемия
☐ Аутоиммунный тиреоидит
☐ Вегето-сосудистая дистония
☐ ВИЧ / гепатит В
☐ Гастрит
☐ Гиперпролактинемия
☒ Гипотиреоз
☐ Миома матки
☐ Миопия
☐ Тромбофилия / варикоз
☐ Фиброзно-кистозная мастопатия
☐ Холецистит / панкреатит
☐ Цистит / пиелонефрит
☐ Эндометриоз
☐ Эутиреоз

СТимуляция (программа ЭКО)
 Претритмент:
☐ Пайпель биопсия ☐ Эстроген ☐ зГТ
☒ Гестагены ☐ КОК до протокола ☐ нет
 Протокол:
☒ длинный ☐ короткий ☐ естественный цикл
 Триггер:
☐ диферелин ☒ прегнил ☐ овитрель ☐ ХГЧ

Оплодотворение (программа ЭКО)
 Количество зрелых ооцитов: 3
 Тип оплодотворения:
☐ ЭКО ☒ ИКСИ ☐ ЭКО+ИКСИ
 Среда жидкость:
☒ Cook ☐ CSSM ☐ Irvine
☐ Sage ☐ Vitrolife
 Среда масло:
☒ Cook ☐ Vitrolife ☐ Irvine

Перенос
 Количество эмбрионов:
☒ 1 ☐ 2
 Качество эмбрионов:
☐ высокое ☐ низкое
☐ 1 высокое, 1 низкое ☒ 5 сутки ☐ 6 сутки
 Перенос:
☐ 3 сутки ☐ 5 сутки ☐ 6 сутки
 Дополнительно для программы Криопереносы:
 Среда заморозки:
☐ Irvine ☐ Kitazato ☐ Irvine ☐ Kitazato
 Срок хранения:
☒ до 2 мес. ☐ от 12 мес. до 5 лет
☐ от 2 мес. до 1 года ☐ более 5 лет

Прогноз по анамнезу
 вероятность наступления беременности: 4%
при одноплодной беременности:
 вероятность прерывания: 42%
возможные осложнения:
 истмико-цервикальная недостаточность: 100%
 гипертензивные расстройства: 8%
 предлежание плаценты: 12%
 гестационный сахарный диабет: 33%
 нарушения количества околоплодных вод: 36%
 преждевременный разрыв плодных оболочек: 0%
предполагаемый срок родов: 38
 плановое кесарево сечение
группа здоровья ребенка: 3-4
 (K00-K93,L00-L99,M00-M99,Q00-Q99)
при многоплодной беременности:
 вероятность прерывания: 0%
возможные осложнения:
 истмико-цервикальная недостаточность: 18%

Прогноз после стимуляции
 ожидаемое количество фолликул: 9
 толщина эндометрия 8-15 мм
 хорошее качество ооцитов
 вероятность наступления беременности: 40%
при одноплодной беременности:
 вероятность прерывания: 25%
возможные осложнения:

Прогноз результата оплодотворения
 количество клеток на 3 сутки (> 68): 1
 количество бластоцист на 5 сутки (> 38В): 0
 количество бластоцист на 6 сутки (> 38В): 0
 общее количество бластоцист на 6 сутки: 1

Прогноз после переноса
программа ЭКО:
 вероятность наступления беременности: 95%
одноплодная беременность
 вероятность прерывания: 15%
возможные осложнения:
 истмико-цервикальная недостаточность: 100%
 гипертензивные расстройства: 8%
 предлежание плаценты: 12%
 гестационный сахарный диабет: 33%
 нарушения количества околоплодных вод: 36%
 преждевременный разрыв плодных оболочек: 0%
предполагаемый срок родов: 38
 плановое кесарево сечение
группа здоровья ребенка: 3-4
 (K00-K93,L00-L99,Q00-Q99)

Рисунок 4.3 – апробация созданной СППВР на данных представленного примера

Видно, что СППВР правильно определила наступление беременности, истмико-цервикальную недостаточность, способ родов. Ожидаемое количество фол-

ликулов оказалось больше реального количества (девять вместо четырех), предполагаемый срок родов меньше на две недели чем реальный. Врожденных аномалий развития (Q00-Q99), предсказанных СППВР, у ребенка не обнаружено.

Пример работы с приложением во время ведения пациента

Пользовательское приложение позволяет врачу наглядно отобразить пациенту предстоящий процесс проведения протокола ВРТ и отображает независимый, основанный на данных прогноз.

До вступления в протокол ВРТ в первый блок вносится вся информация о пациентке (рисунок 4.4). Например, пациентка собирается вступить в первый протокол ЭКО, имеет неразвивающуюся беременность в анамнезе, индекс массы тела равный 28, возраст – 30 лет, нарушения цикла и хронический панкреатит.

The screenshot displays the SPPIVR application interface for patient data entry and prognosis. The interface is divided into several sections:

- Программа (Program):** Includes radio buttons for ☒ ЭКО (IVF) and ☐ криопереносы (Cryopreservation).
- Номер попытки ВРТ (IVF attempt number):** A text input field containing the value 1.
- ИМТ (BMI):** A text input field containing the value 28.
- Возраст (Age):** A text input field containing the value 30.
- Акушерский анамнез (Obstetric history):** A list of radio buttons for various conditions:
 - ☐ Аборт на раннем сроке (Early abortion)
 - ☐ Внематочная беременность (Ectopic pregnancy)
 - ☐ Кесарево сечение (Cesarean section)
 - ☒ Неразвивающаяся беременность (Miscarriage)
 - ☐ Преждевременные роды (Premature birth)
 - ☐ Самопроизвольный выкидыш (Spontaneous miscarriage)
 - ☐ Срочные роды (Term birth)
- Хирургические вмешательства (Surgical interventions):** A list of radio buttons:
 - ☒ Гистероскопия (Hysteroscopy)
 - ☐ Лапароскопия (Laparoscopy)
 - ☐ Лапаротомия (Laparotomy)
 - ☐ Резекция в мочеполовой системе (Resection in the genitourinary system)
- Применялась гормональная контрацепция (Hormonal contraception used):** ☒
- Есть нарушения цикла (Irregular cycle):** ☒
- Гормональное лечение в наст. время (Hormonal treatment at present):** ☐
- Наличие у ближайших родственников (Presence in relatives):**
 - ☐ Сахарный диабет (Diabetes)
 - ☐ Гипертония/ ишемия/ инсульт/ тромбозы (Hypertension/ischemia/stroke/thromboses)
- Хронические заболевания (Chronic diseases):** A grid of radio buttons for various conditions:
 - ☐ Аденомиоз (Adenomyosis)
 - ☐ Аднексит (Adnexitis)
 - ☐ Аллергия (Allergy)
 - ☐ Анемия (Anemia)
 - ☐ Аутоиммунный тиреоидит (Autoimmune thyroiditis)
 - ☐ Вегето-сосудистая дистония (Vegeto-vascular dystonia)
 - ☐ ВИЧ / гепатит В (HIV / Hepatitis B)
 - ☐ Гастрит (Gastritis)
 - ☐ Гиперпролактинемия (Hyperprolactinemia)
 - ☐ Гипотиреоз (Hypothyroidism)
 - ☐ Миома матки (Uterine fibroid)
 - ☐ Миопия (Myopia)
 - ☐ Тромбофилия / варикоз (Thrombophilia / varicose veins)
 - ☐ Фиброзно-кистозная мастопатия (Fibrocystic mastopathy)
 - ☒ Холецистит / панкреатит (Cholecystitis / pancreatitis)
 - ☐ Цистит / пиелонефрит (Cystitis / pyelonephritis)
 - ☐ Эндометриоз (Endometriosis)
 - ☐ Эутиреоз (Euthyroidism)
- Прогноз по анамнезу (Prognosis by history):** A summary box showing:
 - вероятность наступления беременности: 0%
 - при одноплодной беременности:**
 - вероятность прерывания: 0%
 - возможные осложнения:**
 - истмико-цервикальная недостаточность: 5%
 - гипертензивные расстройства: 12%
 - предлежание плаценты: 29%
 - гестационный сахарный диабет: 0%
 - нарушения количества околоплодных вод: 71%
 - преждевременный разрыв плодных оболочек: 0%
 - предполагаемый срок родов: 38**
 - экстренное кесарево сечение
 - группа здоровья ребенка: 5**
 - (C00-D48, H00-H59, L00-L99, M00-M99, Q00-Q99)
 - при многоплодной беременности:**
 - вероятность прерывания: 17%
 - возможные осложнения:**
 - истмико-цервикальная недостаточность: 0%

Рисунок 4.4 – пример работы с первым блоком СППВР

Результат работы моделей первого блока (рисунок 4.4), дает понять, что по данным пациентки сложно утверждать о возможном наступлении беременности (вероятность – 0%), причем, даже в случае наступления беременности, она осложнится нарушением количества околоплодных вод и, при многоплодной беременности, гипертензивным расстройством. Завершится экстренным или, в случае многоплодной беременности, плановым кесаревым сечением с высокой вероятностью рождения ребенка группы здоровья 5.

Далее можно посмотреть, при какой организации второго этапа протокола ЭКО, у пациентки будут лучшие шансы на деторождение. Например, при проведении длинного протокола с определенной подготовкой, СППВР дает прогноз вероятности наступления беременности – 25%. При этом прогноз результата наступившей беременности практически не меняется. А, к примеру, при проведении короткого протокола вероятность наступления беременности не возрастает (рисунок 4.5)

The image displays two side-by-side screenshots of the 'СППВР' (SPVPR) software interface, specifically the 'СТИМУЛЯЦИЯ (программа ЭКО)' (Stimulation (IVF program)) section. Both screens show the same set of radio button options for 'Претритмент' (Pre-treatment), 'Протокол' (Protocol), and 'Триггер' (Trigger). The left screen shows the 'длинный' (long) protocol selected, resulting in a 'вероятность наступления беременности: 25%' (probability of pregnancy: 25%). The right screen shows the 'короткий' (short) protocol selected, resulting in a 'вероятность наступления беременности: 0%' (probability of pregnancy: 0%). Both screens also display 'ожидаемое количество фолликул: 11' (expected number of follicles: 11) on the left and 'ожидаемое количество фолликул: 4' (expected number of follicles: 4) on the right. The 'возможные осложнения' (possible complications) section is visible at the bottom of both screens.

Рисунок 4.5 – пример работы со вторым блоком СППВР

При проведении оплодотворения в определенной среде можно посмотреть, какой тип оплодотворения будет лучшим вариантом. Например, в данном случае при вероятных семи зрелых ооцитах, часть из которых будет оплодотворена ЭКО, а

часть – ИКСИ, прогнозируется получение двух бластоцист хорошего качества (рисунок 4.6).

И далее, в случае получения эмбрионов высокого качества и переноса одного из них в полость матки, прогнозируемая вероятность наступления беременности равна 100%, а при эмбрионе низкого качества – 70% (рисунок 4.7). Причем в любом случае ожидается такое осложнение беременности, как нарушение количества околоплодных вод с высокой вероятностью проведения экстренного кесарева сечения на 38 неделе.

Количество зрелых ооцитов: 7

Тип оплодотворения:

- ☐ ЭКО
- ☐ ИКСИ
- ☒ ЭКО+ИКСИ

Среда жидкость:

- ☒ Cook
- ☐ CSSM
- ☐ Irvine
- ☐ Sage
- ☐ Vitrolife

Среда масло:

- ☒ Cook
- ☐ Vitrolife
- ☐ Irvine

Прогноз результата оплодотворения:

количество клеток на 3 сутки (> 6В):	4
количество бластоцист на 5 сутки (> 3ВВ):	2
количество бластоцист на 6 сутки (> 3ВВ):	2
общее количество бластоцист на 6 сутки:	3

Рисунок 4.6 – пример работы с третьим блоком СППВР

Таким образом, на основе всей представленной СППВР информации, врач может обсудить с пациенткой ее планы по вступлению в протокол, оценить возможные риски и принять управленческие, персонифицированные решения по тактике ведения пациентки.

Таким образом, клиника АО «ЦСМ» г. Екатеринбург признала тестирование успешным, а программу пригодной для использования в лечебном процессе, о чем получен акт внедрения от 11 апреля 2022 года.

Также данная программа для ЭВМ «Прогнозирование исходов и здоровья детей, зачатых при помощи вспомогательных репродуктивных технологий» получила высокую оценку и рекомендована к внедрению в других медицинских организациях для повышения качества оказываемой медицинской помощи Плаксиной А.Н. (главным специалистом Министерства здравоохранения Свердловской

области по медицинской реабилитации детей, заведующей детской поликлиникой ЕМЦ «УГМК-Здоровье»).

The image shows two identical panels of the SPBPV software interface, likely representing a before-and-after or a comparison of two scenarios. Each panel is titled 'ПЕРЕНОС' (Transfer).

Left Panel Settings:

- Количество эмбрионов:** 1 (selected), 2
- Качество эмбрионов:** высокое (selected), 1 высокое, 1 низкое, низкое
- Перенос:** 3 сутки, 5 сутки (selected), 6 сутки
- Дополнительно для программы Криопереносы:**
 - Среда заморозки:** Irvine (selected), Kitazato
 - Среда разморозки:** Irvine (selected), Kitazato
 - Срок хранения:** до 2 мес. (selected), от 2 мес. до 1 года, от 12 мес. до 5 лет, более 5 лет

Right Panel Settings:

- Количество эмбрионов:** 1 (selected), 2
- Качество эмбрионов:** высокое, 1 высокое, 1 низкое, низкое (selected)
- Перенос:** 3 сутки, 5 сутки (selected), 6 сутки
- Дополнительно для программы Криопереносы:**
 - Среда заморозки:** Irvine (selected), Kitazato
 - Среда разморозки:** Irvine (selected), Kitazato
 - Срок хранения:** до 2 мес. (selected), от 2 мес. до 1 года, от 12 мес. до 5 лет, более 5 лет

Prognosis (Прогноз после переноса):

программа ЭКО:

- вероятность наступления беременности: 100% (Left) / 70% (Right)
- одноплодная беременность**
- вероятность прерывания: 0%
- возможные осложнения:**
- истмико-цервикальная недостаточность: 5%
- гипертензивные расстройства: 12%
- предлежание плаценты: 29%
- гестационный сахарный диабет: 0%
- нарушения количества околоплодных вод: 71%
- преждевременный разрыв плодных оболочек: 0%
- предполагаемый срок родов: 38**
- экстренное кесарево сечение
- группа здоровья ребенка: 3-4**
- (N00-N59, L00-L99, Q00-Q99)

Рисунок 4.7 – пример работы с четвертым блоком СППВР

Получено свидетельство о государственной регистрации программы для ЭВМ «Прогнозирование исходов и здоровья детей, зачатых при помощи вспомогательных репродуктивных технологий» № 2021618399 от 26 мая 2021г.

Перспективой дальнейшей работы является:

- 1) сбор больших баз данных проведенных протоколов ВРТ, включающих информацию о долгосрочном здоровье рожденных в результате этих протоколов детей;
- 2) создание базы данных с помощью разработанных алгоритмов обработки данных;

3) обучение ML-моделей на больших объемах данных с возможным изменением входных и выходных параметров;

4) применение многоэтапного подхода к прогнозированию и алгоритма подготовки данных для работы с данными женщин, беременность которых наступила естественным путем и для создания систем поддержки принятия врачебных решений в других областях.

Выводы

В четвертой главе решена последняя задача диссертационного исследования, создать математическое и алгоритмическое обеспечение СППВР для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи ВРТ, реализовать СППВР в виде программного приложения и апробировать СППВР в клинической практике, а именно:

- 1) проведена обработка предоставленной медицинской документации при помощи описанного алгоритмического обеспечения (глава 2), в следствии которой получены наборы подготовленных данных;
- 2) выделены значимые признаки для каждой ML-модели в составе СППВР;
- 3) разработаны все ML-модели для СППВР;
- 4) разработано пользовательское программное приложение, позволяющее работать с созданным математическим и алгоритмическим обеспечением СППВР;
- 5) проведена успешная апробация созданного математического и алгоритмического обеспечения на новых данных: средняя доля правильных ответов для прогнозируемых исходов составляет 84,5%, при этом 95% доверительный интервал – [78,7; 90,3].

ЗАКЛЮЧЕНИЕ

1. Проведен анализ состояния предметной области, обоснована актуальность создания информационной системы поддержки принятия решений для прогнозирования здоровья детей, зачатых при помощи ВРТ, и отсутствие аналогов.

2. Разработаны алгоритмы обработки, компоновки и анализа разнородных текстовых данных на примере данных медицинских организаций для выделения признакового описания моделируемых объектов.

3. Разработана структура СППВР, которая позволяет осуществлять многоэтапное прогнозирование здоровья ребенка, зачатого при помощи ВРТ.

4. Создано математическое и алгоритмическое обеспечение СППВР для многоэтапного прогнозирования здоровья ребенка, зачатого при помощи ВРТ.

5. Созданное программное приложения, позволяющее работать с СППВР, получило высокую оценку по результатам тестирования, внедрено в АО «Центр семейной медицины» г. Екатеринбурга и рекомендовано к внедрению в других медицинских организациях.

Перспективы дальнейшей разработки темы исследования заключаются в применении созданного многоэтапного подхода и алгоритма подготовки данных к другим данным протоколов ВРТ, к данным женщин, беременность которых наступила естественным путем, а также в других прикладных областях.

СПИСОК СОКРАЩЕНИЙ

ВРТ	Вспомогательные репродуктивные технологии
СППВР	Система поддержки принятия врачебных решений (прогнозирования здоровья ребенка, зачатого при помощи ВРТ)
ML-модель	Модель машинного обучения
ЭКО	Экстракорпоральное оплодотворение
ИКСИ	Интрацитоплазматическая инъекция сперматозоида в цитоплазму яйцеклетки

СПИСОК ЛИТЕРАТУРЫ

1. Численность населения Российской Федерации по полу и возрасту на 1 января 2021 года (Статистический бюллетень). Официальный портал Федеральной службы государственной статистики. URL: https://rosstat.gov.ru/storage/mediabank/Bul_chislen_nasel-pv_01-01-2021.pdf (дата обращения: 16.05.2022).
2. Официальный сайт ФГБУ «НМИЦ АГП им. В.И. КУЛАКОВА» Минздрава России. URL: https://ncagp.ru/index.php?_t8=422 (дата обращения: 16.05.2022).
3. Об утверждении плана основных мероприятий, проводимых в рамках Десятилетия детства, на период до 2027 г. : распоряжение Правительства Российской Федерации от 23.01.2021 № 122-р.
4. Регистр ВРТ Российской ассоциации репродукции человека. URL: https://www.rahr.ru/d_registr_otchet/RegistrART2018.pdf (дата обращения: 02.09.2020).
5. International Committee for Monitoring Assisted Reproductive Technologies world report: assisted reproductive technology 2012 / J. de Mouzon [et al.] // Human Reproduction. – V. 35, I. 8. – P. 1900–1913.
6. Assisted Reproductive Technology Surveillance – United States, 2014 / S. Sunderam [et al.] // The Morbidity and Mortality Weekly Report (MMWR) Surveillance Summaries. – 2017. – V. 66, I. 6. – P. 1–24.
7. Вспомогательные репродуктивные технологии: перинатальные исходы и состояние детей / Г. М. Савельева, Г. В. Касьянова, М. А. Дронова, Е. М. Карачунская // Проблемы репродукции. – 2014. – Т. 20, № 24. – С. 35–39.
8. Кешишян, Е. С. Состояние здоровья и развитие детей, рожденных после экстракорпорального оплодотворения / Е. С. Кешишян, А. Д. Царегородцев, М. И. Зиборова // Российский вестник перинатологии и педиатрии. – 2014. – Т. 59, № 5. – С. 15–25.
9. Von Wolff, M. In vitro fertilization technology and child health risks, mechanisms and possible consequences / M. von Wolff, T. Haaf // Deutsches Ärzteblatt international. – 2020. – V. 117, № 3. – P. 23–30.
10. Wennerholm, U. B. Perinatal outcome in children born after assisted reproductive technologies / U. B. Wennerholm, C. Bergh // Upsala journal of medical sciences. – 2020. – V. 125, № 2. – P. 158–166.
11. McDonald, S. Systematic review and meta-analysis of perinatal outcomes of in vitro fertilization singletons / S. McDonald, K. Murphy, A. Óhlsson // Journal of obstetrics and gynaecology Canada. – 2005. – V. 25, № 5. – P. 449–459.

12. Intellectual Disability in Children Conceived Using Assisted Reproductive Technology / M. Hansen, K. R. Greenop, J. Bourke [et al.] // *Pediatrics*. – 2018. – Vol. 142, № 6. – e20181269.
13. Japan Environment & Children's Study Group. Complications and adverse outcomes in pregnancy and childbirth among women who conceived by assisted reproductive technologies: a nationwide birth cohort study of Japan environment and children's study / C. Nagata, L. Yang, K. Yamamoto-Hanada [et al.] // *Bio-Med Central pregnancy and childbirth*. – 2019. – Vol. 19, № 1. – P. 77.
14. The Massachusetts Outcomes Study of Assisted Reproductive Technologies (Perinatal Outcomes Associated With Assisted Reproductive Technology: The Massachusetts Outcomes Study of Assisted Reproductive Technologies (MOSART) / E. Declercq, B. Luke, C. Belanoff [et al.] // *Fertility and Sterility*. – 2015. – Vol. 103. – P. 888–895.
15. ISO/TS 22756:2020 Health Informatics — Requirements for a knowledge base for clinical decision support systems to be used in medication-related processes. URL: <https://www.iso.org/standard/73827.html> (дата обращения: 22.05.2021).
16. Платформа прогнозной аналитики Webiomed. Обзор Российских систем поддержки принятия врачебных решений (СППВР). URL: <https://webiomed.ru/blog/obzor-rossiiskikh-sistem-podderzhki-priniatiia-vrachebnykh-reshenii/> (дата обращения: 22.05.2021).
17. Decision time for clinical decision support systems / O'Sullivan D., Fraccaro P., Carson E., Weller P. // *Clinical Medicine*. 2014, 14 (4) 338-341;
18. Teufel A, Binder H: Clinical Decision Support Systems. *Visc Med* 2021;37:491-498.
19. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method / Qiu, J. [et al.] // *Journal of Translational Medicine*. – 2019. – V. 17.
20. Predicting live birth chances for women with multiple consecutive failing IVF cycles: a simple and accurate prediction for routine medical practice / Porcu G [et al.] // *Reproductive Biology and Endocrinology*. – 2013. – V. 11, I.1.
21. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers / K. K. Vaegter [et al.] // *Fertility and Sterility*. – 2017. – V. 107, I. 3. – P.641–648.
22. Predicting the likelihood of a live birth for women with endometriosis-related infertility / L. Benoit [et al.] // *European Journal of Obstetrics & Gynecology and Reproductive Biology*. – 2019. – V. 242. – P. 56–62.

23. Goyal A. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment / A. Goyal, M. Kuchana, K. P. R. Ayyagari // *Scientific Reports*. – 2020. – V. 10, № 20925
24. Development and validation of a model for individualized prediction of cervical insufficiency risks in patients undergoing IVF/ICSI treatment / Y. Wu [et al.] // *Reproductive Biology and Endocrinology*. – 2021. – V.19, № 6.
25. Колесниченко Т.В., Амирова А.А. Прогнозирование исходов ЭКО и ЭКО/ИКСИ у бесплодных супружеских пар при различных формах бесплодия. // *Материалы форума Информационные технологии и общество Турция, Кемер (Бельдиби)*. – 2008. – С. 166–167.
26. A Novel Predictive Model to Estimate the Number of Mature Oocytes Required for Obtaining at Least One Euploid Blastocyst for Transfer in Couples Undergoing in vitro Fertilization/Intracytoplasmic Sperm Injection: The ART Calculator / S. C. Esteves [et al.] // *Frontiers in Endocrinology*. – 2019. – V.10, № 99.
27. A systematic review of the quality of clinical prediction models in in vitro fertilization / M. B. Ratna [et al.] // *Human Reproduction*. – 2020. – Vol. 35, № 1. – P. 100–116.
28. Nelson S. M. Predicting Live Birth, Preterm Delivery, and Low Birth Weight in Infants Born from In Vitro Fertilisation: A Prospective Study of 144,018 Treatment Cycles / S. M. Nelson, D. A. Lawlor // *PLOS Medicine*. – 2011. – V. 8, I. 1.
29. Факторы риска и алгоритм прогнозирования нарушений здоровья к году жизни у детей, родившихся с очень низкой и экстремально низкой массой тела / О. М. Филькина [и др.] // *Анализ риска здоровью*. – 2016. – Т.1, № 13. – С. 68–75.
30. Плаксина А.Н. Прогнозирование здоровья и качества жизни детей, рожденных с помощью вспомогательных репродуктивных технологий: дис. ... канд. мед. наук. – Екатеринбург, 2011. – 174 с.
31. Deep learning for clinical decision support in in vitro fertilisation, IVF – проект Monash University (Мельбурн, Австралия). URL: <https://supervisorconnect.it.monash.edu/projects/honours/deep-learning-clinical-decision-support-vitro-fertilisation-ivf> (дата обращения: 20.04.2022).
32. Letterie G., Mac Donald A. Artificial intelligence in in vitro fertilization: a computer decision support system for day-to-day management of ovarian stimulation during in vitro fertilization // *Fertility and Sterility*. – 2020. – V. 114, I. 5. – P. 1026-1031.
33. Клинические рекомендации Вспомогательные репродуктивные технологии и искусственная инсеминация : Протокол лечения // Министерство здравоохранения Российской Федерации. – Письмо № 15-4/И/2-1908. – 5 марта 2019.

34. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology / D. Glujovsky [et al.] // *Cochrane Database of Systematic Reviews*. – 2016. – V. 6.
35. Consecutive transfer of day 3 embryos and of day 5-6 blastocysts increases overall pregnancy rates associated with blastocyst culture / S.J. Phillips [et al.] // *Journal of Assisted Reproduction and Genetics*. – 2003. – V. 20, I. 11. – P. 461-464.
36. Elective and Onco-fertility preservation: factors related to IVF outcomes / A Cobo [et al.] // *Human Reproduction*. – 2018. – V. 33, I. 12. – P. 2222–2231.
37. Endometrial thickness as a predictor of the reproductive outcomes in fresh and frozen embryo transfer cycles / T. Zhang [et al.] // *Medicine*. – 2018. – V. 97, I. 4.
38. Maternal serum concentration of anti-Mullerian hormone is a better predictor than basal follicle stimulating hormone of successful blastocysts development during IVF treatment / S. Sadruddin [et al.] // *PLOS One*. – 2020. – V. 15, I. 10.
39. Factors affecting clinical pregnancy rates after IUI for the treatment of unexplained infertility and mild male subfertility / M. Atasever [et al.] // *Journal Of The Turkish-German Gynecological Association*. – 2016. – V. 17. – P. 134–138.
40. Nomograms for predicting adverse obstetric outcome in IVF pregnancy: A preliminary study / A. Portal [et al.] // *Birth: Issues in Perinatal Care*. – 2021. – V. 48, I. 2. – P. 186–193.
41. Purewal S. A systematic review and meta-analysis of lifestyle and body mass index predictors of successful assisted reproductive technologies / S. Purewal, S. C. E. Chapman, O. B.A. van den Akker // *Journal of Psychosomatic Obstetrics & Gynecology*. – 2019. – V. 40, I. 1. – P. 2–18.
42. Individualized decision-making in IVF: calculating the chances of pregnancy / L.L. van Loendersloot [et al.] // *Human Reproduction*. – 2013. – V. 28, I. 11. – P. 2972–2980.
43. Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool / R. K. Dhillon [et al.] // *Human Reproduction*. – 2016. – V. 31, I. 1. – P. 84–92.
44. Факторы риска и прогнозирование нарушений здоровья у детей первого года жизни, родившихся от одноплодной беременности после экстракорпорального оплодотворения / Л.А. Пыхтина [и др.] // *Анализ риска здоровью*. – 2017. – №1. – С. 56–65.
45. Influence of maternal obesity on the long-term health of offspring / K. M. Godfrey [et al.] // *The Lancet Diabetes & Endocrinology*. – 2017. – V. 5, I. 1. – P. 53-64.
46. McEvoy C. T., Spindel E. R. Pulmonary Effects of Maternal Smoking on the Fetus and Child: Effects on Lung Development, Respiratory Morbidities, and Life Long Lung Health / C. T. McEvoy, E. R. Spindel // *Paediatric Respiratory Reviews*. – 2017. – V. 21. – P. 27-33.
47. Cerebral palsy after assisted reproductive technology: a cohort study / S. Goldsmith [et al.] // *Developmental Medicine and Child Neurology*. – 2018. – V. 60, I. 1. – P. 73–80.

- 48.Лалаян Т.Н. Состояние здоровья детей, рожденных с помощью новых репродуктивных технологий: дис. ... канд. мед. наук / Т. Н. Лалаян : С.-петерб. мед. акад. последипломного образования. – Санкт-Петербург : [б. и.], 2005. – 118 с.
- 49.Growth, health, and motor development of 5-year-old children born after preimplantation genetic diagnosis / M. Heijligers [et al.] // Fertility and Sterility – 2019. – V.111, I.6. – P.1151–1158.
- 50.АО «Центр семейной медицины». URL: <https://www.cfm.ru/> (дата обращения: 22.05.2021).
- 51.Международная классификация болезней 10-го пересмотра (МКБ-10). URL: <https://www.who.int/publications/i/item/International-statistical-classification-of-diseases-and-related-health-problems-icd-10-in-occupational-health> (дата обращения 22.05.2021).
- 52.ГАУЗ СО «Областная детская клиническая больница». URL: <http://odkb.ru/> (дата обращения: 22.05.2021).
- 53.ГАУЗ СО «МКМЦ «Бонум». URL: <http://bonum.info/> (дата обращения: 22.05.2021).
- 54.О комплексной оценке состояния здоровья детей : Приказ Министерства Здравоохранения Российской Федерации от 30.12.2003 № 621.
- 55.Kursa M. B. Feature Selection with the Boruta Package / M. B. Kursa, W. R. Rudnicki // Journal of Statistical Software. – 2010. - Vol. 36, I. 11. – P. 1–13.
- 56.Kohavi R, John G.H. Wrappers for feature selection / R. Kohavi, G.H. John // Artificial Intelligence. – 1997. – V.97. – P. 273–324.
- 57.Kenji K. The feature selection problem: traditional methods and a new algorithm / K. Kenji, A. L. Rendell // Association for the Advancement of Artificial Intelligence. – 1992. – P. 129–134.
- 58.Kononenko I. Estimating attributes: Analysis and extensions of RELIEF / I. Kononenko // Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). – 1994. – V. 784. – P. 171–182.
- 59.Robnik-Sikonja M. An adaptation of Relief for attribute estimation in regression / M. Robnik-Sikonja, I. Kononenko // ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning. – 1997. – P. 296–304.
- 60.Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale / J. Hamon // Applications [stat.AP]. Université des Sciences et Technologie de Lille. – 2013.
- 61.Amaldi E., Kann V. On the approximation of minimizing non zero variables or unsatisfied relations in linear systems / E. Amaldi, V. Kann // Theoretical Computer Science. – 1998. – V. 209. – P. 237–260.

62. Gene Selection for Cancer Classification using Support Vector Machines / I. Guyon [et al.] // Machine Learning. – 2002. – V. 46. – P. 389–422.
63. Реализация алгоритма RFECV в библиотеке Scikit-learn. URL: https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.RFECV (дата обращения: 22.05.2021).
64. Воронцов К. В. Лекции по методам оценивания и выбора моделей / К. В. Воронцов. URL: <http://www.ccas.ru/voron/download/Modeling.pdf> (дата обращения: 22.05.2021).
65. Homola D. BorutaPy. An all relevant feature selection method based on Random Forest estimators. URL: <https://danielhomola.com/feature%20selection/phd/borutapy-an-all-relevant-feature-selection-method/#python-implementation> (дата обращения: 22.05.2021).
66. Python-реализация метода Boruta. URL: https://github.com/scikit-learn-contrib/boruta_py (дата обращения 18.08.2020).
67. Библиотека ELI5. URL: <https://eli5.readthedocs.io/en/latest/index.html#> (дата обращения: 22.05.2021).
68. Permutation importance: a corrected feature importance measure / A. Altmann [et al.] // Bioinformatics. – 2010. – V. 26, I. 10. – P. 1340–1347.
69. Brink H., Richards J.W., Fetherolf M. Real-World Machine Learning. – 2017. – 264 p.
70. Кристофер Доугерти. Введение в эконометрику. — 2-е, пер. с англ. — М.: ИНФРА-М, 2004. — 419 с.
71. Estimation of Health-Related Physical Fitness Using Multiple Linear Regression in Korean Adults: National Fitness Award 2015–2019 / Sung-Woo K. [et al.] // Frontiers in Physiology. – 2021. – V. 12.
72. Renal length and volume prediction in healthy children / Bianchi M.E. [et al.] // Clinical Nephrology. – 2020. – V. 93. – P. 31–35.
73. Реализация алгоритма Logistic Regression в библиотеке Scikit-learn. URL: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (дата обращения: 22.05.2021).
74. Прогнозирование риска развития силикоза у работников огнеупорного производства / Обухова Т.Ю. [и др.] // Здоровье населения и среда обитания – ЗНисО. – 2020. – Т. 9. – P. 84–90.
75. A risk factor analysis and prediction model of placental abruption / Huang K.N. [et al.] // International journal of clinical and experimental medicine. – 2018. – V. 12, I. 11. – P. 13023–13031.
76. A validated risk prediction model for bone fragility in children with acute lymphoblastic leukemia / Verwaaijen E.J. [et al.] // Journal of bone and mineral research. – 2021. – V. 36(12). – P: 2290–2299.

77. Risk prediction of severe reaction to oral challenge test of cow's milk / Kawahara T. [et al.] // *European Journal of Pediatrics*. – 2019. – V. 178. – P. 181–188.
78. Реализация дерева решений в sklearn URL: <https://scikit-learn.org/stable/modules/tree.html#tree> (дата обращения: 23.05.2021).
79. Decision tree-based rules outperform risk scores for childhood asthma prognosis / Owora A.H [et al.] // *Pediatric Allergy and Immunology*. – 2021. – V. 32. – P. 1464–1473.
80. Early recognition of the preference for exclusive breastfeeding in current China: a prediction model based on Decision Trees / Wang Y.[et al.] // *Scientific Reports*. – 2020. – V. 10, № 6720.
81. Selection of children with ultra-severe traumatic brain injury for neurosurgical intervention / Greenan K.[et al.] // *Journal of Neurosurgery: Pediatrics PED*. – 2022. – V. 23, I. 6. – P. 670-679.
82. Decision tree modeling in R software to aid clinical decision making / Toth E.G.[et al.] // *Health and Technology*. – 2021. – V. 11. – P. 535–545.
83. Schapire R.E. The strength of weak learnability / R. E. Shapire // *Mach Learn*. – 1990. – V.5. – P.197–227.
84. Bias–variance decomposition // *Encyclopedia of Machine Learning*. — 2011.
85. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными (оригинал: *Introduction to Machine Learning with Python: A Guide for Data Scientists*). – 2016-2017. – 393 с.
86. Реализация алгоритма Random Forest Classifier в библиотеке Scikit-learn. URL: <https://scikit-learn.org/stable/modules/ensemble.html#forest> (дата обращения: 22.05.2021).
87. Development of a random forest model for hypotension prediction after anesthesia induction for cardiac surgery / X.F. Li [et al.] // *World Journal of Clinical Cases*. – 2021. – V. 9, I. 29. – P. 8729-8739.
88. Byeon. H. Comparing the accuracy and developed models for predicting the confrontation naming of the elderly in South Korea using weighted Random Forest, Random Forest, and Support Vector Regression / H. Byeon // *International Journal of Advanced Computer Science and Applications*. – 2021. – V. 12. I. 2. – P: 326-331.
89. Predicting children with ADHD using behavioral activity: A machine learning analysis // M. Maniruzzaman [et al.] / *Applied sciences-basel*. – 2022. – V. 12, I. 5. - № 2723.
90. Automated prediction of COVID-19 mortality outcome using clinical and laboratory data based on hierarchical feature selection and random forest classifier / N.

- Amini [et al.] // Computer Methods in Biomechanics and Biomedical Engineering. – 2022.
91. Реализация алгоритма Gradient Boosting Classifier в библиотеке Scikit-learn [Электронный ресурс] – Режим доступа: <https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting> (дата обращения: 22.05.2021).
 92. Early prediction of clinical response to etanercept treatment in juvenile idiopathic arthritis using machine learning / X. Mo [et al.] *Frontiers in Pharmacology*. – 2020. – V. 11, I. 1164.
 93. Friedman J.H. Greedy function approximation: A gradient boosting machine / J.H. Friedman // *Annals of Statistics*. – 2001. – V.29, I.5. – P. 1189-1232.
 94. A comparative study of patient and staff safety evaluation using tree-based machine learning algorithms / M. C. E. Simsekler [et al.] // *Reliability Engineering & System Safety*. – 2021. – V. 208. - № 107416.
 95. Stillbirth risk prediction using machine learning for a large cohort of births from Western Australia, 1980–2015 / E. Malacova [et al.] // *Scientific Reports*. – 2020. – V. 10, 5354.
 96. Библиотека для анализа и обработки данных pandas. URL: <https://pandas.pydata.org/> (дата обращения: 22.05.2021).
 97. Платформа для обработки данных и машинного обучения Anaconda. URL: <https://www.anaconda.com> (дата обращения: 22.05.2021).
 98. Пакет для научных вычислений. URL: <https://numpy.org/> (дата обращения: 22.05.2021).
 99. Морфологический анализатор pymorphy2. URL: <https://pymorphy2.readthedocs.io/en/latest/> (дата обращения 22.05.2021).
 100. Scikit-learn: Machine Learning in Python / Pedregosa [et al.] // *Journal of Machine Learning Research*. – 2011. – V. 12. – P. 2825–2830.
 101. Библиотека Hyperopt. URL: <http://hyperopt.github.io/hyperopt/> (дата обращения: 22.05.2021).
 102. KOEHRSEN W. Automated Model Tuning. URL: <https://www.kaggle.com/willkoehrsen/automated-model-tuning> (дата обращения: 22.05.2021).
 103. RNA. Parameter Tuning in One Function with Hyperopt. URL: <https://www.kaggle.com/bigironsphere/parameter-tuning-in-one-function-with-hyperopt> (дата обращения: 22.05.2021).
 104. ADASYN: Adaptive synthetic sampling approach for imbalanced learning / H. He [et. al] // 2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congr. on Computational Intelligence). – Hong Kong: 2008. – P. 1322–1328.

105. Lemaitre G. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning / G. Lemaitre, F. Nogueira, C. K. Aridas // Journal of Machine Learning Research. – 2017. – V. 18, № 17. – P. 1–5.
106. Hanley J. A. The meaning and use of the area under a receiver operating characteristic (ROC) curve / J. A. Hanley, B. J. McNeil // Radiology. – 1982. – V. 143, № 1. – P. 29–36.
107. Реализация расчета средней квадратической ошибки (MSE) в библиотеке Scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (дата обращения: 22.05.2021).
108. Реализация расчета средней абсолютной ошибки (MAE) в библиотеке Scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (дата обращения: 22.05.2021).
109. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // R. Kohavi // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. – San Mateo, CA: 1995. – Vol. 2, № 12. – P.1137—1143.
110. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions / M. Stone // Journal of the Royal Statistical Society, Series B (Methodological) – 1974. – V. 36, № 2. – P. 111–147.
111. Реализация стратегии кросс-валидации StratifiedKFold в библиотеке sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html (дата обращения: 22.05.2021).
112. Библиотека Pickle. URL: <https://docs.python.org/3.6/library/pickle.html>.
113. PyQt5 - набор расширений для создания графического интерфейса на базе Qt5 для языка программирования Python. URL: <https://pypi.org/project/PyQt5/>.
114. Qt Designer - инструмент для быстрого создания графических пользовательских интерфейсов. URL: <https://build-system.fman.io/qt designer-download>.

Приложение А. Распределение моделей машинного обучения по блокам СППВР

№	Первый блок	№	Второй блок	№	Третий блок	№	Четвертый блок ЭКО	№	Четвертый блок криоперенос
1	наступление беременности	27	наступление беременности (2)	41	количество клеток хорошего качества на 3 сутки	45	наступление беременности (3)	56	наступление беременности (4)
2	истмико-церви- кальная недостаточность	28	истмико-цервикальная недостаточность (2)	42	количество клеток хорошего качества на 5 сутки		из блока 1		из блока 1
3	гипертензивные расстройства	29	гипертензивные расстройства (2)	43	количество клеток хорошего качества на 6 сутки		из блока 1		из блока 1
4	предлежание плаценты	30	предлежание плаценты (2)	44	общее количество клеток на 6 сутки		из блока 1		из блока 1
5	гестационный сахарный диабет		из блока 1				из блока 1		из блока 1
6	нарушения количества околоплодных вод		из блока 1				из блока 1		из блока 1
7	преждевременный разрыв плодных оболочек		из блока 1				из блока 1		из блока 1
8	прерывание беременности		из блока 1			46	прерывание беременности (2)	57	прерывание беременности (3)
9	способ родов		из блока 1				из блока 1		из блока 1
10	срок родов		из блока 1				из блока 1		из блока 1
11	группа здоровья	31	группа здоровья (2)			47	группа здоровья (3)	58	группа здоровья (4)
12	A00-B99	32	A00-B99 (2)				из блока 2		из блока 1
13	C00-D48	33	C00-D48 (2)			48	C00-D48 (3)		из блока 1
14	D50-D89		из блока 1			49	D50-D89 (2)		из блока 1
15	E00-E90		из блока 1				из блока 1	59	E00-E90 (2)
16	G00-G99	34	G00-G99 (2)				из блока 2		из блока 1

17	H00-H59		H00-H59 (2)		50	H00-H59 (2)		из блока 1
18	H60-H95	35	из блока 1			из блока 2		из блока 1
19	I00-I99		из блока 1			из блока 1		из блока 1
20	J00-J99		из блока 1		51	J00-J99 (2)	60	J00-J99 (3)
21	K00-K93		из блока 1			из блока 1		из блока 1
22	L00-L99		из блока 1			из блока 1		из блока 1
23	M00-M99		из блока 1		52	M00-M99 (2)	61	M00-M99 (3)
24	N00-N99	36	N00-N99(2)		53	N00-N99 (3)		из блока 1
25	P00-P96		из блока 1		54	P00-P96 (2)	62	P00-P96 (4)
26	Q00-Q99	37	Q00-Q99 (2)			из блока 2		из блока 1
		38	количество фолликулов		55	количество плодов (ЭКО)	63	количество плодов (крио)
		39	толщина эндометрия					
		40	качество ооцитов					

**Приложение Б. Распределение признаков по ML-моделям
СППВР и значимость признаков**
ML-модели первого блока

номер признака (табл.6)	беременность	гестационный сахарный диабет	гипертензивные расстройства	истмико-цервикальная недостаточность	нарушения количества околоплодных вод	предлежание плаценты	преждевременный разрыв плодных оболочек	способ родов	прерывание беременности	срок родов	группы здоровья
1									0,029		
2					0,053						
3						0,064	0,068				
4								0,044			
5									0,004		
6									0,029		
7			0,078						0,058		
8							0,060	0,039	0,027		
9			0,044			0,064					
11					0,071						
12							0,058				
13		0,108				0,073					
14	0,423	0,367	0,253	0,498	0,388	0,323	0,233	0,316	0,183	0,241	0,262
15					0,082		0,080		0,048		
16	0,365	0,251	0,220	0,387	0,259	0,167	0,151	0,252	0,189	0,202	0,223
17	0,212	0,189	0,066								
18							0,040				
20			0,048			0,042					
21			0,084				0,062				
27											0,056
32						0,073					
37			0,142		0,147	0,116	0,090	0,170	0,074	0,168	0,150
38				0,115				0,062	0,258	0,173	
39		0,083	0,065			0,080	0,082	0,117		0,116	0,093
41							0,076		0,064		
43									0,021	0,050	
44									0,016		
45										0,050	
46											0,104
47											0,112

номер признака (табл.6)	A00-B99	C00-D48	D50-D89	E00-E90	G00-G99	H00-H59	H60-H95	I00-I99	J00-J99	K00-K93	L00-L99	M00-M99	N00-N99	P00-P96	Q00-Q99
1			0,036												
2			0,075			0,063									
3			0,036				0,064		0,058						
7	0,087							0,083					0,132		
8			0,051			0,058	0,041	0,085				0,042	0,107		
9				0,069					0,059						
10	0,062														
11			0,056		0,053			0,172			0,071			0,053	0,060
12		0,061	0,045				0,055	0,041	0,064				0,056	0,088	0,100
13															
14	0,086	0,198	0,151	0,229	0,228	0,227	0,140	0,133	0,185	0,249	0,256	0,178	0,229	0,160	0,234
15	0,082	0,102				0,039			0,053					0,056	
16	0,068	0,136	0,090	0,117	0,147	0,138	0,086	0,184	0,133	0,139	0,158	0,149	0,179	0,106	0,121
17							0,053								
18					0,042		0,058			0,076					
27		0,039	0,039			0,022	0,044		0,055						
32				0,056		0,044		0,038				0,053			
37	0,245	0,086	0,080	0,168	0,153	0,120	0,092		0,081	0,123	0,134	0,089		0,142	0,124
38	0,214			0,056	0,095			0,049				0,053			
39		0,098	0,042		0,091	0,099		0,081		0,097	0,105	0,074		0,066	0,088
41		0,051	0,063				0,107			0,108	0,047	0,051		0,047	
42									0,038			0,041		0,024	
43							0,067					0,053		0,038	0,043
46	0,096	0,155	0,060	0,084	0,075	0,073	0,192	0,056	0,079		0,098	0,113		0,058	0,095
47	0,059	0,075	0,115	0,123	0,117	0,117		0,078	0,107	0,139	0,131	0,103	0,112	0,117	0,135

ML-модели второго блока

номер признака (табл.6)	беременность	гипертензивные расстройства	истмико-цервикальная недостаточность	предлежание плаценты	группы здоровья	количество фолликулов	эндометрий 8-15 мм	качество ооцитов	A00-B99	C00-D48	G00-G99	H60-H95	N00-N99	Q00-Q99
1								0,064						
3				0,070		0,054		0,089				0,039		
7		0,088						0,053	0,040				0,101	
8												0,057	0,120	
9		0,043		0,050										
10									0,044					
11											0,036			0,079
12										0,056		0,039	0,069	0,092
13				0,072			0,037	0,051						
14	0,171	0,124	0,285	0,283	0,138	0,291	0,369	0,309	0,204	0,112	0,116	0,119	0,220	0,203
15							0,099		0,062	0,085				
21		0,028											0,078	
17	0,035	0,029						0,038				0,007		
16	0,621	0,100	0,221	0,149	0,134	0,141		0,269	0,095	0,121	0,100	0,124	0,135	0,116
18							0,045				0,108	0,032	0,040	
19		0,030		0,054										
20							0,038							
27					0,014					0,023		0,049		
32				0,074										
35							0,027							
37		0,104		0,059	0,052	0,137		0,126	0,071	0,083	0,026	0,089		0,086
39		0,057		0,048	0,039	0,105				0,070	0,070			0,056
40														
38			0,065						0,081		0,087			
41										0,059		0,062		
43												0,046		0,056
46					0,042				0,082	0,117	0,045	0,160		0,094
47					0,222				0,080	0,144	0,123		0,109	0,186
48	0,092		0,125			0,095	0,188		0,114				0,128	
49	0,051	0,061				0,046	0,197							
50	0,029	0,336				0,131			0,128					
51														
52					0,046						0,158	0,178		0,032
53			0,304	0,142	0,146					0,132	0,132			

ML-модели третьего блока

номер признака (табл. 6)	количество клеток на 3 сутки хорошего качества	количество бластоцист на 5 сутки хорошего качества	количество бластоцист на 6 сутки хорошего качества	общее количество бластоцист на 6 сутки
52	0,023	0,020		
54	0,824	0,232		
55	0,058	0,077		
56	0,030	0,038		
57	0,065	0,075		
58		0,557		
59			1	
60				1

ML-модели четвертого блока (ЭКО)

номер признака (табл.6)	беременность	прерывание беременности	группы здоровья	C00-D48	D50-D89	H00-H59	J00-J99	M00-M99	N00-N99	P00-P96	количество плодов
1		0,039			0,033						
2					0,049	0,040					
3					0,027		0,052				
5		0,008									
6		0,031									
7		0,060							0,111		
8		0,049			0,059	0,050		0,094	0,201		
9							0,078				
11					0,156					0,181	
12				0,051	0,033		0,067		0,119	0,078	
14	0,006	0,265	0,110	0,082	0,090	0,161	0,133	0,137	0,194	0,079	0,149
15		0,050		0,074		0,047	0,093			0,024	
16	0,032	0,098	0,070	0,057	0,044	0,124	0,104	0,095	0,099	0,085	
17	0,060										
21					0,029		0,072		0,055	0,056	
27			0,069	0,026	0,038	0,024	0,033				

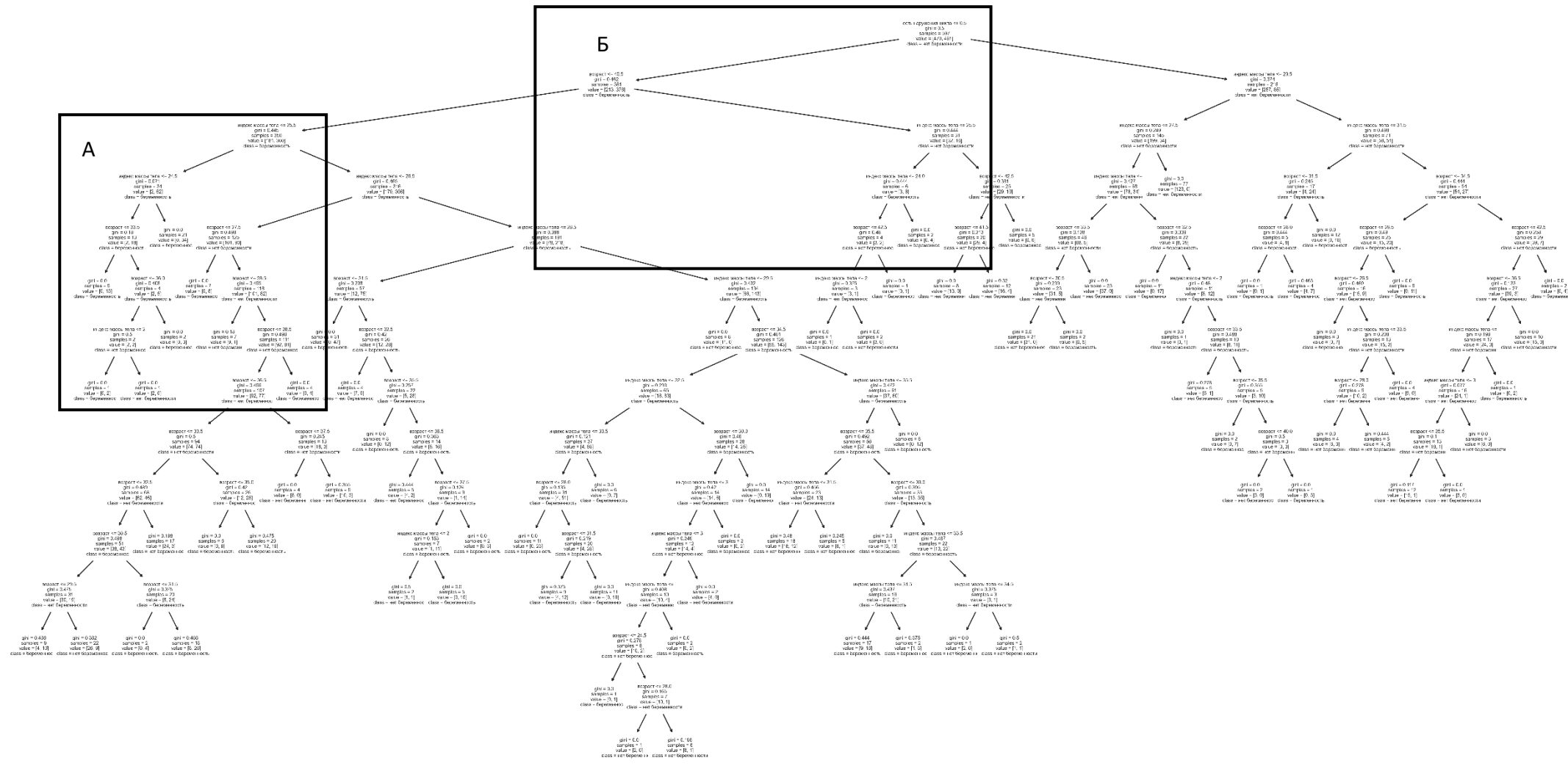
32						0,059		0,062			
33									0,009		
37		0,142	0,113	0,151	0,063	0,103	0,127	0,116		0,092	0,104
38		0,087						0,064			
39			0,071	0,086	0,072	0,086		0,088		0,092	0,115
41		0,059		0,078	0,094			0,039		0,072	
42							0,039	0,007		0,027	
43		0,063						0,033		0,023	
44		0,040									
46			0,091	0,116	0,091	0,087	0,071	0,164		0,086	
47			0,195	0,081	0,116	0,123	0,122	0,080	0,157	0,086	
48	0,002										0,096
49	0,026										0,074
50	0,124										0,059
52			0,026								
53			0,076								
54	0,232		0,067								0,179
61	0,481	0,003	0,008		0,008	0,015	0,008	0,023	0,054	0,008	
62	0,021					0,019				0,012	
63	0,015		0,056								0,193
64		0,008	0,049	0,199		0,062					0,031

ML-модели четвертого блока (Криоперенос)

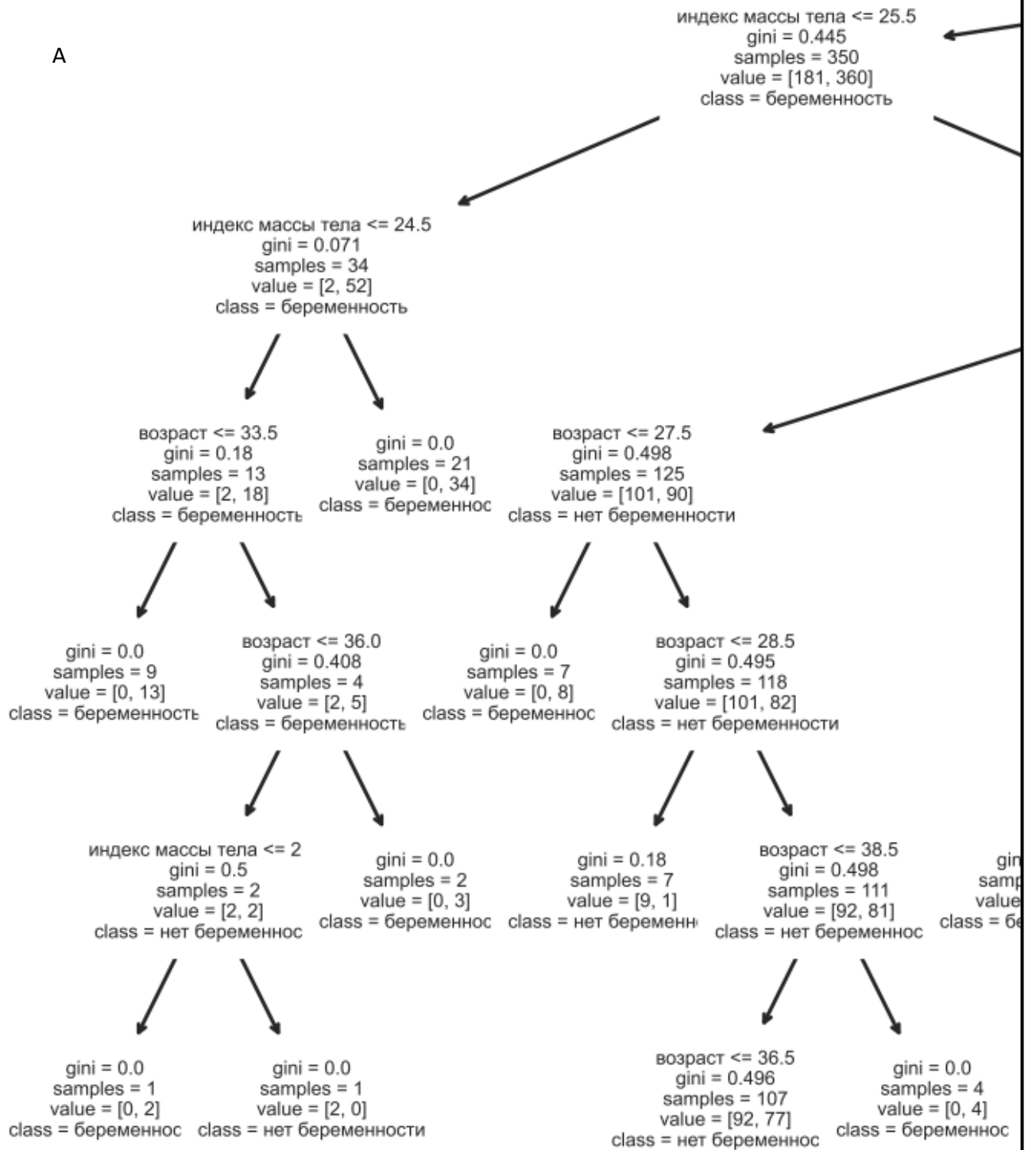
номер признака (табл. 6)	беременность	прерывание беременности	группы здоровья	E00-E90	J00-J99	M00-M99	P00-P96	количество плодов
1		0,040						
2								
6						0,069		
7		0,232			0,067			
8		0,043						
10		0,085						
11		0,134						
12				0,051				
15					0,035			
18					0,021	0,023		
19							0,056	
20							0,101	
21					0,038			
23	0,028	0,048	0,109	0,056			0,044	0,242
24					0,210			
25		0,003			0,116		0,076	
26								
27				0,033	0,044	0,068	0,203	
28	0,145							

29	0,079	0,013	0,075	0,010			0,019	
41					0,035			
42			0,099					
48						0,005		
49				0,027				
52					0,186	0,136		
53		0,053	0,065	0,156			0,090	0,182
54			0,087			0,076	0,014	
55								
56		0,070		0,113		0,064		
58		0,065				0,091	0,044	
59					0,095	0,055	0,030	
60		0,070				0,118	0,097	
61		0,023						
64			0,084	0,225	0,060	0,157	0,070	
65			0,212	0,092	0,058	0,020	0,076	
72			0,032					
73	0,231		0,088					
73	0,064		0,063					0,212
75	0,375	0,001			0,036			
76	0,079		0,035				0,060	0,148
77				0,076		0,059		
78						0,058		
79		0,074		0,162				0,128
80		0,047	0,051				0,020	

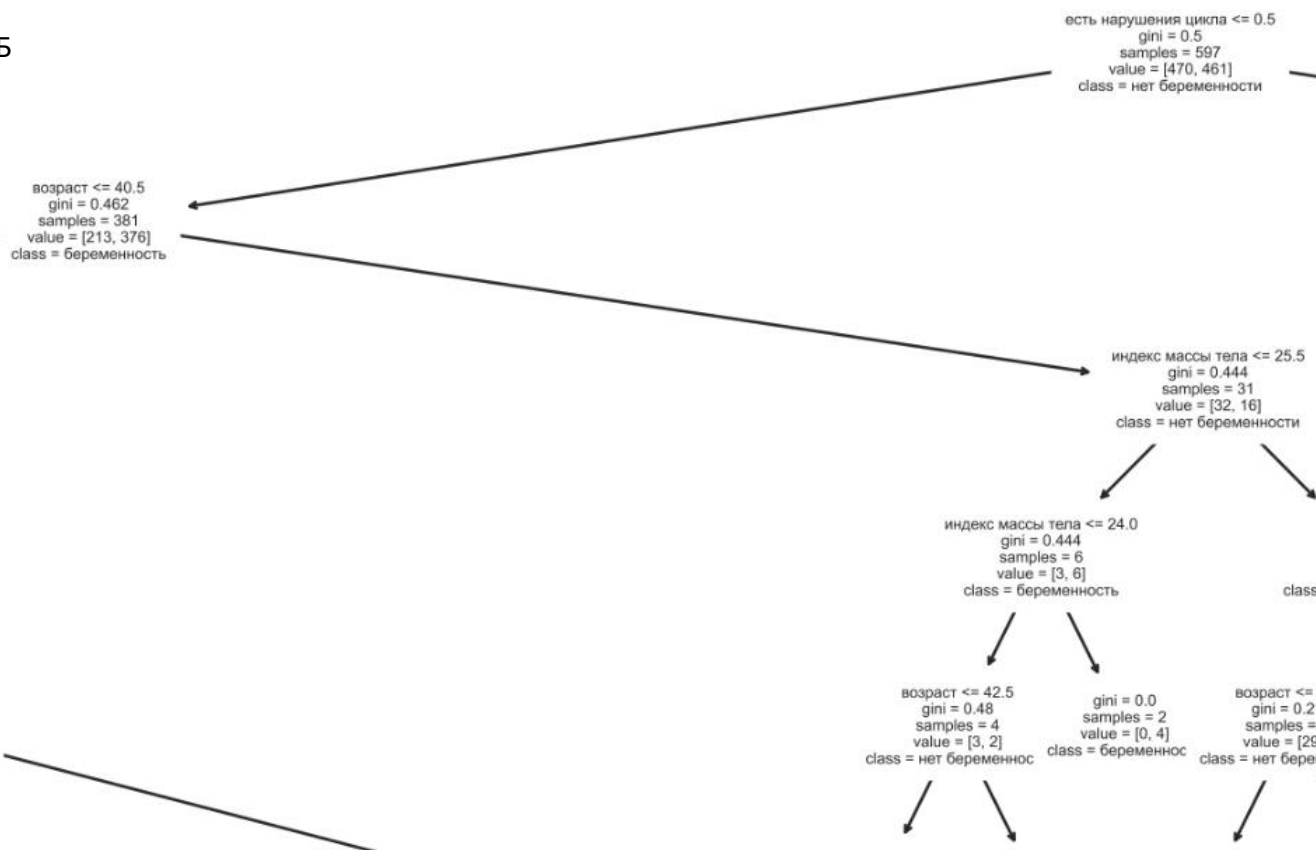
Приложение В. Пример одного из деревьев решений для предсказания наступления беременности



A



Б



Приложение Г. Интерфейс пользовательского программного приложения

Блок 1

Программа
☒ ЭКО ☐ криопереносы

Номер попытки ВРТ

ИМТ

Возраст

Акушерский анамнез

☐ Аборт на раннем сроке

☒ Внематочная беременность

☐ Кесарево сечение

☐ Неразвивающаяся беременность

☐ Преждевременные роды

☐ Самопроизвольный выкидыш

☐ Срочные роды

Хирургические вмешательства

☒ Гистероскопия

☒ Лапароскопия

☐ Лапаротомия

☐ Резекция в мочеполовой системе

☐ Применялась гормональная контрацепция

☒ Есть нарушения цикла

☐ Гормональное лечение в наст. время

Наличие у ближайших родственников:

☐ Сахарный диабет

☐ Гипертония ишемия инсульт тромбозы

Хронические заболевания

☐ Аденомиоз

☐ Аднексит

☐ Аллергия

☐ Анемия

☐ Аутоиммунный тиреоидит

☐ Вегето-сосудистая дистония

☐ ВИЧ / гепатит В

☐ Гастрит

☐ Гиперпролактинемия

☐ Гипотиреоз

☐ Миома матки

☐ Миопия

☐ Тромбофилия / варикоз

☐ Фиброзно-кистозная мастопатия

☐ Холецистит / панкреатит

☐ Цистит / пиелонефрит

☐ Эндометриоз

☐ Эутиреоз

Прогноз по анамнезу

вероятность наступления беременности: 60%

при одноплодной беременности:

вероятность прерывания: 8%

возможные осложнения:

истмико-цервикальная недостаточность: 2%

гипертензивные расстройства: 4%

предлежание плаценты: 55%

гестационный сахарный диабет: 25%

нарушения количества околоплодных вод: 0%

преждевременный разрыв плодных оболочек: 64%

предполагаемый срок родов: 38

экстренное кесарево сечение

группа здоровья ребенка: 3-4

(C00-D48,E00-E90,I00-I99,K00-K93,L00-L99,
M00-M99,P00-P96)

при многоплодной беременности:

вероятность прерывания: 8%

возможные осложнения:

Блок 2

СТИМУЛЯЦИЯ (программа ЭКО)

Претритмент

☐ Пайпель биопсия ☐ Эстроген ☐ ЗГТ

☐ Гестагены ☒ КОК до протокола ☐ нет

Протокол

☐ длинный ☒ короткий

☐ естественный цикл

Триггер

☒ диферелин ☐ прегнил

☐ овитрель ☐ ХГЧ

Прогноз после стимуляции

ожидаемое количество фолликул: 17

толщина эндометрия 8-15 мм

плохое качество ооцитов

вероятность наступления беременности: 75%

при одноплодной беременности:

вероятность прерывания: 8%

возможные осложнения:

ОПЛОДОТВОРЕНИЕ (программа ЭКО)

Количество зрелых ооцитов

Тип оплодотворения

☐ ЭКО

☐ ИКСИ

☒ ЭКО+ИКСИ

Среда жидкость

☒ Cook ☐ CSSM ☐ Irvine

☐ Sage ☐ Vitrolife

Среда масло

☒ Cook ☐ Vitrolife ☐ Irvine

Прогноз результата оплодотворения

количество клеток на 3 сутки (> 6B): 4

количество бластоцист на 5 сутки (> 3BV): 2

количество бластоцист на 6 сутки (> 3BV): 2

общее количество бластоцист на 6 сутки: 3

Блок 4

ПЕРЕНОС

Количество эмбрионов

☒ 1 ☐ 2

Качество эмбрионов

☒ высокое

☐ 1 высокое, 1 низкое

☐ низкое

Перенос

☐ 3 сутки

☒ 5 сутки

☐ 6 сутки

Дополнительно для программы Криопереносы:

Среда заморозки

☐ Irvine ☐ Kitazato

Среда разморозки

☐ Irvine ☐ Kitazato

Срок хранения

☒ до 2 мес.

☐ от 12 мес. до 5 лет

☐ от 2 мес. до 1 года

☐ более 5 лет

Прогноз после переноса

программа ЭКО:

вероятность наступления беременности: 100%

многоплодная беременность

вероятность прерывания: 0%

возможные осложнения:

истмико-цервикальная недостаточность: 37%

гипертензивные расстройства: 4%

предлежание плаценты: 55%

гестационный сахарный диабет: 25%

нарушения количества околоплодных вод: 0%

преждевременный разрыв плодных оболочек: 64%

предполагаемый срок родов: 36

плановое кесарево сечение

группа здоровья ребенка: 3-4

(C00-D48,G00-G99,H00-H59,I00-I99,K00-K93,
.....)

Приложение Д. Копии акта о внедрении результатов диссертационного исследования и свидетельства о регистрации программы для ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО
о государственной регистрации программы для ЭВМ
№ 2021618399

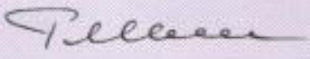
Прогнозирование исходов и здоровья детей, зачатых при помощи вспомогательных репродуктивных технологий

Правообладатель: *Федеральное государственное бюджетное образовательное учреждение высшего образования "Уральский государственный медицинский университет" Министерства здравоохранения Российской Федерации (ФГБОУ ВО УГМУ Минздрава России) (RU)*

Авторы: *Плаксына Анна Николаевна (RU), Синотова Светлана Леонидовна (RU), Лимановская Оксана Викторовна (RU), Макутина Валерия Андреевна (RU)*

Заявка № **2021617516**
Дата поступления **18 мая 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ **26 мая 2021 г.**



*Руководитель Федеральной службы
по интеллектуальной собственности*
 **Г.П. Ивлиев**



Акционерное общество
«Центр семейной медицины»

ИНН 6658112160 ОГРН1026602320502
620043, г. Екатеринбург,
ул. Начдива Васильева, дом 1, корпус 3.
Почтовый адрес: 620043, г. Екатеринбург, а/я 35
Телефон: (343) 237-1-999, 214-8-999
Сайт: www.cfm.ru; e-mail: cfm@cfm.ru

«УТВЕРЖДАЮ»

Генеральный директор АО «ЦСМ», к.м.н



Д.Р. Медведская

«11» апреля 2022 г.

№ _____ от _____

АКТ

об использовании результатов диссертационного исследования
Синотовой Светланы Леонидовны, а именно программы для ЭВМ
«Прогнозирование исходов и здоровья детей, зачатых при помощи вспомогательных
репродуктивных технологий» в АО «ЦСМ»

Комиссия в составе: председатель комиссии: главный врач, к.м.н. Рябко Е.В.

Члены комиссии:

Балезин С.Л. – заместитель генерального директора, к.м.н.

Медведева В.В. – заведующий консультативным отделением ВРТ, к.м.н.,

составила настоящий акт, подтверждающий, что программа для ЭВМ позволяет определить вероятность наступления беременности и ее осложнений, исходы беременности, группы здоровья детей и классы возможных заболеваний, что используется для корректной тактики ВРТ, разработки индивидуальных программ ранней помощи и последующего динамического наблюдения этой категории младенцев.

Председатель комиссии:

Главный врач АО «ЦСМ», к.м.н.

Е.В. Рябко

Члены комиссии:

Заместитель генерального
директора АО «ЦСМ», к.м.н.

С.Л. Балезин

Заведующий консультативным отделением ВРТ
АО «ЦСМ», к.м.н.

В.В. Медведева

Приложение Е. Код основного файла интерфейса созданного программного приложения

```
1. from PyQt5 import uic, QtWidgets
2. from PyQt5.QtWidgets import QApplication
3. import pandas as pd
4. import models_new as models
5. import result_in_listWidget_new as res
6. import features_from_radioButtons_new as ffrB
7.
8.
9. Form, _ = uic.loadUiType("GUI2 all-ред2.ui")
10.
11. class Ui (QtWidgets.QDialog, Form):
12.     def __init__(self):
13.         super(Ui,self).__init__()
14.         self.setupUi(self)
15.         self.pushButton.clicked.connect(self.count)
16.         self.pushButton_2.clicked.connect(self.count2)
17.         self.pushButton_5.clicked.connect(self.count3)
18.         self.pushButton_4.clicked.connect(self.count4)
19.
20.         delegate = ffrB.HTMLDelegate(self.listWidget)
21.         self.listWidget.setItemDelegate(delegate)
22.         self.listWidget_2.setItemDelegate(delegate)
23.         self.listWidget_5.setItemDelegate(delegate)
24.         self.listWidget_4.setItemDelegate(delegate)
25.
26.
27.     def count(self):
28.
29.         data_all = pd.DataFrame(index=[0],columns=range(0,68))
30.
31.         data = ffrB.features_b1(self,data_all)
32.
33.         # количество плодов
34.         data[data.columns[38]] = 1
35.
36.         if (int(window.lineEdit_3.text()) == 0) or (int(win-
            dow.lineEdit_2.text()) == 0) or (int(window.lineEdit_4.text()) == 0):
```



```

37.         self.listWidget.addItem('введите возраст, ИМТ, номер попытки ВРТ')
38.
39.     else:
40.
41.         predictpr = models.loaded_modelpr.predict_proba(data[data.columns[[14,17,16]]])
42.
43.         predict_icn = models.loaded_model_icn.predict_proba(data[data.columns[[14,16,38]]])
44.         predict_gt = models.loaded_model_gt.predict_proba(data[data.columns[[7,9,14,21,17,16,19,37,39]]])
45.         predict_pp = models.loaded_model_pp.predict_proba(data[data.columns[[3,9,13,14,16,19,32,37,39]]])
46.         predict_gsd = models.loaded_model_gsd.predict_proba(data[data.columns[[13,14,17,16,39]]])
47.         predict_nkov = models.loaded_model_nkov.predict_proba(data[data.columns[[2,11,14,15,16,37]]])
48.
49.         data[data.columns[41]] = models.loaded_model_gsd.predict(data[data.columns[[13,14,17,16,39]]])
50.         data[data.columns[42]] = models.loaded_model_gt.predict(data[data.columns[[7,9,14,21,17,16,19,37,39]]])
51.         data[data.columns[43]] = models.loaded_model_icn.predict(data[data.columns[[14,16,38]]])
52.         data[data.columns[44]] = models.loaded_model_nkov.predict(data[data.columns[[2,11,14,15,16,37]]])
53.
54.         predict_prpo = models.loaded_model_prpo.predict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
55.
56.         data[data.columns[45]] = models.loaded_model_prpo.predict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
57.
58.         predict_sposob = models.loaded_model_sposob.predict(data[data.columns[[4,8,14,16,37,39,38]]])
59.         predict_break = models.loaded_model_break.predict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44]]])
60.         predict_srok = models.loaded_model_srok.predict(data[data.columns[[14,16,37,39,38,43,45]]])
61.

```

```

62.         data[data.columns[46]] = predict_sposob
63.         data[data.columns[47]] = predict_srok
64.
65.         predict_groups = models.loaded_model_groups.pre-
        dict(data[data.columns[[14,16,27,37,39,46,47]]])
66.
67.
68.         self.listWidget.clear()
69.         pregnancy='вероятность наступления беременности:
        '+str(round(predictpr[0,1]*100))+ '%'
70.         self.listWidget.addItem(pregnancy)
71.         self.listWidget.addItem('<b>при одноплодной беременности:</b>')
72.
73.         res.result_in_listWidget(self.listWidget,predictpr,
74.                                   predict_icn,predict_gt,predict_pp,pre-
        dict_gsd,predict_nkov,predict_prpo,
75.                                   predict_break,predict_srok,pre-
        dict_sposob,predict_groups)
76.
77.         if ((predict_groups == 2) or (predict_groups == 3)):
78.             predict_a00_b99 = models.loaded_model_a00_b99.pre-
        dict(data[data.columns[[7,10,14,15,16,37,38,46,47]]])
79.             predict_c00_d48 = models.loaded_model_c00_d48.pre-
        dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47]]])
80.             predict_d50_d89 = models.loaded_model_d50_d89.pre-
        dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47]]])
81.             predict_e00_e90 = models.loaded_model_e00_e90.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47]]])
82.             predict_g00_g99 = models.loaded_model_g00_g99.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47]]])
83.             predict_h00_h59 = models.loaded_model_h00_h59.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47]]])
84.             predict_h60_h95 = models.loaded_model_h60_h95.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46]]])
85.             predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
86.             predict_j00_j99 = models.loaded_model_j00_j99.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47]]])
87.             predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])

```

```

88.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
89.         predict_m00_m99 = models.loaded_model_m00_m99.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47]]])
90.         predict_n00_n99 = models.loaded_model_n00_n99.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47]]])
91.         predict_p00_p96 = models.loaded_model_p00_p96.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47]]])
92.         predict_q00_q99 = models.loaded_model_q00_q99.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47]]])
93.
94.
95.         res.diagnosis_result(self.listWidget,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
96.                             predict_e00_e90,predict_g00_g99,predict_h00_h59,
97.                             predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
98.                             predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
99.
100.
101.         data[data.columns[38]] = 2
102.         self.listWidget.addItem('<b>при многоплодной беременности:</b>')
103.
104.         predictpr = models.loaded_modelpr.predict_proba(data[data.col-
        umns[[14,17,16]]])
105.
106.         predict_icn = models.loaded_model_icn.pre-
        dict_proba(data[data.columns[[14,16,38]]])
107.         predict_gt = models.loaded_model_gt.predict_proba(data[data.col-
        umns[[7,9,14,21,17,16,19,37,39]]])
108.         predict_pp = models.loaded_model_pp.predict_proba(data[data.col-
        umns[[3,9,13,14,16,19,32,37,39]]])
109.         predict_gsd = models.loaded_model_gsd.pre-
        dict_proba(data[data.columns[[13,14,17,16,39]]])
110.         predict_nkov = models.loaded_model_nkov.pre-
        dict_proba(data[data.columns[[2,11,14,15,16,37]]])
111.
112.         data[data.columns[41]] = models.loaded_model_gsd.pre-
        dict(data[data.columns[[13,14,17,16,39]]])

```

```

113.         data[data.columns[42]] = models.loaded_model_gt.pre-
            dict(data[data.columns[[7,9,14,21,17,16,19,37,39]]])
114.         data[data.columns[43]] = models.loaded_model_icn.pre-
            dict(data[data.columns[[14,16,38]]])
115.         data[data.columns[44]] = models.loaded_model_nkov.pre-
            dict(data[data.columns[[2,11,14,15,16,37]]])
116.
117.         predict_prpo = models.loaded_model_prpo.pre-
            dict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
118.
119.         data[data.columns[45]] = models.loaded_model_prpo.pre-
            dict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
120.
121.         predict_sposob = models.loaded_model_sposob.pre-
            dict(data[data.columns[[4,8,14,16,37,39,38]]])
122.         predict_break = models.loaded_model_break.pre-
            dict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44]]])
123.         predict_srok = models.loaded_model_srok.predict(data[data.col-
            umns[[14,16,37,39,38,43,45]]])
124.
125.         data[data.columns[46]]= predict_sposob
126.         data[data.columns[47]]= predict_srok
127.
128.         predict_groups = models.loaded_model_groups.pre-
            dict(data[data.columns[[14,16,27,37,39,46,47]]])
129.
130.         res.result_in_listWidget(self.listWidget,predictpr,
131.                                   predict_icn,predict_gt,predict_pp,pre-
            dict_gsd,predict_nkov,predict_prpo,
132.                                   predict_break,predict_srok,pre-
            dict_sposob,predict_groups)
133.
134.         if ((predict_groups == 2) or (predict_groups == 3)):
135.             predict_a00_b99 = models.loaded_model_a00_b99.pre-
                dict(data[data.columns[[7,10,14,15,16,37,38,46,47]]])
136.             predict_c00_d48 = models.loaded_model_c00_d48.pre-
                dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47]]])
137.             predict_d50_d89 = models.loaded_model_d50_d89.pre-
                dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47]]])

```

```

138.         predict_e00_e90 = models.loaded_model_e00_e90.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47]]])
139.         predict_g00_g99 = models.loaded_model_g00_g99.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47]]])
140.         predict_h00_h59 = models.loaded_model_h00_h59.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47]]])
141.         predict_h60_h95 = models.loaded_model_h60_h95.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46]]])
142.         predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
143.         predict_j00_j99 = models.loaded_model_j00_j99.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47]]])
144.         predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])
145.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
146.         predict_m00_m99 = models.loaded_model_m00_m99.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47]]])
147.         predict_n00_n99 = models.loaded_model_n00_n99.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47]]])
148.         predict_p00_p96 = models.loaded_model_p00_p96.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47]]])
149.         predict_q00_q99 = models.loaded_model_q00_q99.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47]]])
150.
151.
152.         res.diagnosis_result(self.listWidget,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
153.         predict_e00_e90,predict_g00_g99,predict_h00_h59,
154.         predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
155.         predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
156.
157.         return data
158. ##-----
        -----
159. ##-----этап 2-----
        -----
160.

```

```

161.     def count2(self):
162.
163.         data_all = Ui.count(self)
164.
165.         data = ffrB.features_b2(self,data_all)
166.
167.         predictfoln = models.loaded_model_follicle.predict(data[data.columns[
            3,14,16,37,39,48,49,50]]])
168.         predictend = models.loaded_model_end.predict(data[data.columns[
            13,14,15,18,20,35,48,49]]])
169.         predictoocq = models.loaded_model_oocq.predict(data[data.columns[
            1,3,7,13,14,17,16,37]]])
170.
171.         self.listWidget_2.clear()
172.         if (self.radioButton_53.isChecked()==False):
173.             self.listWidget_2.addItem('предусмотрена только для программы
            ЭКО')
174.
175.         else:
176.             if (self.radioButton_155.isChecked()==True):
177.                 foln='ожидаемое количество фолликул: 2'
178.             else:
179.                 foln='ожидаемое количество фолликул: '+str(int(predictfoln))
180.             self.listWidget_2.addItem(foln)
181.             if (predictend==1):
182.                 end='толщина эндометрия 8-15 мм'
183.             else:
184.                 end='толщина эндометрия менее 8
            мм'
185.             self.listWidget_2.addItem(end)
186.
187.             if (predictoocq==1):
188.                 ooc='хорошее качество ооцитов'
189.             else:
190.                 ooc='плохое качество ооцитов'
191.             self.listWidget_2.addItem(ooc)
192.
193.             data[data.columns[38]] = 1
194.
195.             data[data.columns[53]]=int(predictfoln)

```

```

196.         data[data.columns[51]]=predictend
197.         data[data.columns[52]]=predictoocq
198.
199.         predictpr2 = models.loaded_modelpr2.predict_proba(data[data.col-
        umns[[14,17,16,48,49,50]]])
200.
201.         predict_icn2 = models.loaded_model_icn2.pre-
        dict_proba(data[data.columns[[14,16,38,48,53]]])
202.         predict_gt2 = models.loaded_model_gt2.pre-
        dict_proba(data[data.columns[[7,9,14,21,17,16,19,37,39,49,50]]])
203.         predict_pp2 = models.loaded_model_pp2.pre-
        dict_proba(data[data.columns[[3,9,13,14,16,19,32,37,39,53]]])
204.         predict_gsd = models.loaded_model_gsd.pre-
        dict_proba(data[data.columns[[13,14,17,16,39]]])
205.         predict_nkov = models.loaded_model_nkov.pre-
        dict_proba(data[data.columns[[2,11,14,15,16,37]]])
206.
207.         data[data.columns[41]] = models.loaded_model_gsd.pre-
        dict(data[data.columns[[13,14,17,16,39]]])
208.         data[data.columns[42]] = models.loaded_model_gt2.pre-
        dict(data[data.columns[[7,9,14,21,17,16,19,37,39,49,50]]])
209.         data[data.columns[43]] = models.loaded_model_icn2.pre-
        dict(data[data.columns[[14,16,38,48,53]]])
210.         data[data.columns[44]] = models.loaded_model_nkov.pre-
        dict(data[data.columns[[2,11,14,15,16,37]]])
211.
212.         predict_prpo = models.loaded_model_prpo.pre-
        dict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
213.
214.         data[data.columns[45]] = models.loaded_model_prpo.pre-
        dict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
215.
216.         predict_sposob = models.loaded_model_sposob.pre-
        dict(data[data.columns[[4,8,14,16,37,39,38]]])
217.         predict_break = models.loaded_model_break.pre-
        dict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44]]])
218.         predict_srok = models.loaded_model_srok.predict(data[data.col-
        umns[[14,16,37,39,38,43,45]]])
219.
220.         data[data.columns[46]] = predict_sposob

```

```

221.         data[data.columns[47]] = predict_srok
222.
223.         predict_groups2 = models.loaded_model_groups2.pre-
        dict(data[data.columns[[14,16,27,37,39,46,47,52,53]]])
224.
225.
226.
227.         pregnancy='вероятность наступления беременности:
        '+str(round(predictpr2[0,1]*100))+'%'
228.         self.listWidget_2.addItem(pregnancy)
229.         self.listWidget_2.addItem('<b>при одноплодной беременности:</b>')
230.
231.         res.result_in_listWidget(self.listWidget_2,predictpr2,
232.                                   predict_icn2,predict_gt2,predict_pp2,predict_gsd,predict_nkov,predict_prpo,
233.                                   predict_break,predict_srok,predict_sposob,predict_groups2)
234.
235.         if ((predict_groups2 == 2) or (predict_groups2 == 3)):
236.             predict_a00_b99 = models.loaded_model_a00_b99_2.pre-
        dict(data[data.columns[[7,10,14,15,16,37,38,43,44,45,46,47,48,50]]])
237.             predict_c00_d48 = models.loaded_model_c00_d48_2.pre-
        dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47,53]]])
238.             predict_d50_d89 = models.loaded_model_d50_d89.pre-
        dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47]]])
239.             predict_e00_e90 = models.loaded_model_e00_e90.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47]]])
240.             predict_g00_g99 = models.loaded_model_g00_g99_2.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47,52,53]]])
241.             predict_h00_h59 = models.loaded_model_h00_h59.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47]]])
242.             predict_h60_h95 = models.loaded_model_h60_h95_2.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46,52]]])
243.             predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
244.             predict_j00_j99 = models.loaded_model_j00_j99.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47]]])
245.             predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])

```



```

246.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
247.         predict_m00_m99 = models.loaded_model_m00_m99.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47]]])
248.         predict_n00_n99 = models.loaded_model_n00_n99_2.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47,48]]])
249.         predict_p00_p96 = models.loaded_model_p00_p96.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47]]])
250.         predict_q00_q99 = models.loaded_model_q00_q99_2.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47,52]]])
251.
252.
253.         res.diagnosis_result(self.listWidget_2,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
254.         predict_e00_e90,predict_g00_g99,predict_h00_h59,
255.         predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
256.         predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
257.
258.
259.         data[data.columns[38]] = 2
260.         self.listWidget_2.addItem('<b>при многоплодной беременно-
        сти:</b>')
261.
262.         predictpr2 = models.loaded_modelpr2.predict_proba(data[data.col-
        umns[[14,17,16,48,49,50]]])
263.
264.         predict_icn2 = models.loaded_model_icn2.pre-
        dict_proba(data[data.columns[[14,16,38,48,53]]])
265.         predict_gt2 = models.loaded_model_gt2.pre-
        dict_proba(data[data.columns[[7,9,14,21,17,16,19,37,39,49,50]]])
266.         predict_pp2 = models.loaded_model_pp2.pre-
        dict_proba(data[data.columns[[3,9,13,14,16,19,32,37,39,53]]])
267.         predict_gsd = models.loaded_model_gsd.pre-
        dict_proba(data[data.columns[[13,14,17,16,39]]])
268.         predict_nkov = models.loaded_model_nkov.pre-
        dict_proba(data[data.columns[[2,11,14,15,16,37]]])
269.

```

```

270.         data[data.columns[41]] = models.loaded_model_gsd.pre-
        dict(data[data.columns[[13,14,17,16,39]]])
271.         data[data.columns[42]] = models.loaded_model_gt2.pre-
        dict(data[data.columns[[7,9,14,21,17,16,19,37,39,49,50]]])
272.         data[data.columns[43]] = models.loaded_model_icn2.pre-
        dict(data[data.columns[[14,16,38,48,53]]])
273.         data[data.columns[44]] = models.loaded_model_nkov.pre-
        dict(data[data.columns[[2,11,14,15,16,37]]])
274.
275.         predict_prpo = models.loaded_model_prpo.pre-
        dict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
276.
277.         data[data.columns[45]] = models.loaded_model_prpo.pre-
        dict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
278.
279.         predict_sposob = models.loaded_model_sposob.pre-
        dict(data[data.columns[[4,8,14,16,37,39,38]]])
280.         predict_break = models.loaded_model_break.pre-
        dict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44]]])
281.         predict_srok = models.loaded_model_srok.predict(data[data.col-
        umns[[14,16,37,39,38,43,45]]])
282.
283.         data[data.columns[46]] = predict_sposob
284.         data[data.columns[47]] = predict_srok
285.
286.         predict_groups2 = models.loaded_model_groups2.pre-
        dict(data[data.columns[[14,16,27,37,39,46,47,52,53]]])
287.
288.         res.result_in_listWidget(self.listWidget_2,predictpr2,
289.                                   predict_icn2,predict_gt2,pre-
        dict_pp2,predict_gsd,predict_nkov,predict_prpo,
290.                                   predict_break,predict_srok,pre-
        dict_sposob,predict_groups2)
291.
292.         if ((predict_groups2 == 2) or (predict_groups2 == 3)):
293.             predict_a00_b99 = models.loaded_model_a00_b99_2.pre-
        dict(data[data.columns[[7,10,14,15,16,37,38,43,44,45,46,47,48,50]]])
294.             predict_c00_d48 = models.loaded_model_c00_d48_2.pre-
        dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47,53]]])

```

```

295.         predict_d50_d89 = models.loaded_model_d50_d89.pre-
        dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47]]])
296.         predict_e00_e90 = models.loaded_model_e00_e90.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47]]])
297.         predict_g00_g99 = models.loaded_model_g00_g99_2.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47,52,53]]])
298.         predict_h00_h59 = models.loaded_model_h00_h59.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47]]])
299.         predict_h60_h95 = models.loaded_model_h60_h95_2.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46,52]]])
300.         predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
301.         predict_j00_j99 = models.loaded_model_j00_j99.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47]]])
302.         predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])
303.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
304.         predict_m00_m99 = models.loaded_model_m00_m99.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47]]])
305.         predict_n00_n99 = models.loaded_model_n00_n99_2.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47,48]]])
306.         predict_p00_p96 = models.loaded_model_p00_p96.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47]]])
307.         predict_q00_q99 = models.loaded_model_q00_q99_2.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47,52]]])
308.
309.
310.         res.diagnosis_result(self.listWidget_2,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
311.         predict_e00_e90,predict_g00_g99,predict_h00_h59,
312.         predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
313.         predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
314.
315.         return data
316. ##-----
        -----

```

```

317.##-----этап 3-----
    -----
318.
319.     def count3(self):
320.
321.         data_all = Ui.count2(self)
322.
323.         data = ffrB.features_b3(self,data_all)
324.
325.         self.listWidget_5.clear()
326.         if (self.radioButton_53.isChecked()==False):
327.             self.listWidget_5.addItem('предусмотрена только для программы
ЭКО')
328.
329.         else:
330.             predict3_6B = models.loaded_model_3dayamount.pre-
dict(data[data.columns[[52,54,57,55,56]]])
331.             p3_6B = 'количество клеток на 3 сутки (>
6B): '+str(int(predict3_6B))
332.             data[data.columns[58]]=int(predict3_6B)
333.             predict5_3BB = models.loaded_model_5dayamount.pre-
dict(data[data.columns[[52,54,57,55,56,58]]])
334.             data[data.columns[59]]=int(predict5_3BB)
335.             predict6_3BB = models.loaded_model_6dayamount.pre-
dict(data[data.columns[[59]]])
336.             data[data.columns[60]]=int(predict6_3BB)
337.             predict_BS = models.loaded_model_6dayamount_all.pre-
dict(data[data.columns[[60]]])
338.
339.             self.listWidget_5.addItem(p3_6B)
340.             p5_3BB = 'количество бластоцист на 5 сутки (>
3BB): '+str(int(predict5_3BB))
341.             p56_3BB = 'количество бластоцист на 6 сутки (>
3BB): '+str(int(predict6_3BB))
342.             p_BS = 'общее количество бластоцист на 6
сутки: '+str(int(predict_BS))
343.             self.listWidget_5.addItem(p5_3BB)
344.             self.listWidget_5.addItem(p56_3BB)
345.             self.listWidget_5.addItem(p_BS)
346.

```

```

347.         return data
348. ##-----
    -----
349. ##-----этап 4-----
    -----
350.
351.     def count4(self):
352.
353.         data_all = Ui.count3(self)
354.
355.         data = ffrB.features_b4(self,data_all)
356.
357.         if ((self.radioButton_53.isChecked()==True) and
            (int(self.lineEdit.text()) !=0)):
358.             self.listWidget_4.clear()
359.             self.listWidget_4.addItem('<b>программа ЭКО:</b>')
360.
361.             predictpr = models.loaded_modelpr3.predict_proba(data[data.col-
                umns[[14,17,16,48,49,50,54,63,61,62]]])
362.
363.             pregnancy='вероятность наступления беременности:
                '+str(round(predictpr[0,1]*100))+ '%'
364.             self.listWidget_4.addItem(pregnancy)
365.
366.             numbereco = models.loaded_model_numbereco.predict(data[data.col-
                umns[[14,37,39,48,49,50,54,63,64,65]]])
367.             data[data.columns[38]] = numbereco
368.
369.             if (numbereco != 1):
370.                 self.listWidget_4.addItem('<b>многоплодная беременность</b>')
371.             else:
372.                 self.listWidget_4.addItem('<b>одноплодная беременность</b>')
373.
374.             predict_icn = models.loaded_model_icn.pre-
                dict_proba(data[data.columns[[14,16,38]]])
375.             predict_gt = models.loaded_model_gt.predict_proba(data[data.col-
                umns[[7,9,14,21,17,16,19,37,39]]])
376.             predict_pp = models.loaded_model_pp.predict_proba(data[data.col-
                umns[[3,9,13,14,16,19,32,37,39]]])

```

```

377.         predict_gsd = models.loaded_model_gsd.pre-
        dict_proba(data[data.columns[[13,14,17,16,39]]])
378.         predict_nkov = models.loaded_model_nkov.pre-
        dict_proba(data[data.columns[[2,11,14,15,16,37]]])
379.
380.         data[data.columns[41]] = models.loaded_model_gsd.pre-
        dict(data[data.columns[[13,14,17,16,39]]])
381.         data[data.columns[42]] = models.loaded_model_gt.pre-
        dict(data[data.columns[[7,9,14,21,17,16,19,37,39]]])
382.         data[data.columns[43]] = models.loaded_model_icn.pre-
        dict(data[data.columns[[14,16,38]]])
383.         data[data.columns[44]] = models.loaded_model_nkov.pre-
        dict(data[data.columns[[2,11,14,15,16,37]]])
384.
385.         predict_prpo = models.loaded_model_prpo.pre-
        dict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
386.
387.         data[data.columns[45]] = models.loaded_model_prpo.pre-
        dict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
388.
389.         predict_sposob = models.loaded_model_sposob.pre-
        dict(data[data.columns[[4,8,14,16,37,39,38]]])
390.         predict_break = models.loaded_model_break_eco.pre-
        dict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44,64,61]]])
391.         predict_srok = models.loaded_model_srok.predict(data[data.col-
        umns[[14,16,37,39,38,43,45]]])
392.
393.         data[data.columns[46]]= predict_sposob
394.         data[data.columns[47]]= predict_srok
395.
396.         predict_groups = models.loaded_model_groups3.pre-
        dict(data[data.columns[[14,16,27,37,39,46,47,52,53,54,63,64,61]]])
397.
398.         res.result_in_listWidget(self.listWidget_4,predictpr,
399.                                predict_icn,predict_gt,predict_pp,pre-
        dict_gsd,predict_nkov,predict_prpo,
400.                                predict_break,predict_srok,pre-
        dict_sposob,predict_groups)
401.
402.         if ((predict_groups == 2) or (predict_groups == 3)):

```

```

403.         predict_a00_b99 = models.loaded_model_a00_b99.pre-
        dict(data[data.columns[[7,10,14,15,16,37,38,46,47]]])
404.         predict_c00_d48 = models.loaded_model_c00_d48_eco.pre-
        dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47,64]]])
405.         predict_d50_d89 = models.loaded_model_d50_d89_eco.pre-
        dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47,61]]])
406.         predict_e00_e90 = models.loaded_model_e00_e90.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47]]])
407.         predict_g00_g99 = models.loaded_model_g00_g99.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47]]])
408.         predict_h00_h59 = models.loaded_model_h00_h59_eco.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47,61,62,64]]])
409.         predict_h60_h95 = models.loaded_model_h60_h95.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46]]])
410.         predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
411.         predict_j00_j99 = models.loaded_model_j00_j99_eco.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47,61]]])
412.         predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])
413.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
414.         predict_m00_m99 = models.loaded_model_m00_m99_eco.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47,61]]])
415.         predict_n00_n99 = models.loaded_model_n00_n99_eco.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47,61]]])
416.         predict_p00_p96 = models.loaded_model_p00_p96_eco.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47,61,62]]])
417.         predict_q00_q99 = models.loaded_model_q00_q99.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47]]])
418.
419.
420.         res.diagnosis_result(self.listWidget_4,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
421.         predict_e00_e90,predict_g00_g99,predict_h00_h59,
422.         predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
423.         predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
424.

```

```

425.
426.
427.         elif ((self.radioButton_53.isChecked()==True) and
(int(self.lineEdit.text()) ==0)):
428.             self.listWidget_4.clear()
429.             self.listWidget_4.addItem('заполните ОПЛОДОТВОРЕНИЕ')
430.
431.
432.### крио
433.
434.         if (self.radioButton_54.isChecked()==True):
435.             self.listWidget_4.clear()
436.             self.listWidget_4.addItem('<b>программа Криопереносы:</b>')
437.
438.             predictpr = models.loaded_modelpr3.predict_proba(data[data.columns[
umns[[14,17,16,48,49,50,54,63,61,62]]]])
439.
440.             pregnancy='вероятность наступления беременности:
'+str(round(predictpr[0,1]*100))+ '%'
441.             self.listWidget_4.addItem(pregnancy)
442.
443.             numbercrio = models.loaded_model_numbercrio.predict(data[data.columns[
[14,37,39,48,49,50,54,63,64,65]]]])
444.             data[data.columns[38]] = numbercrio
445.
446.             if (numbereco != 1):
447.                 self.listWidget_4.addItem('<b>многоплодная беременность</b>')
448.             else:
449.                 self.listWidget_4.addItem('<b>одноплодная беременность</b>')
450.
451.             predict_icn = models.loaded_model_icn.predict_proba(data[data.columns[
[14,16,38]]]])
452.             predict_gt = models.loaded_model_gt.predict_proba(data[data.columns[
[7,9,14,21,17,16,19,37,39]]]])
453.             predict_pp = models.loaded_model_pp.predict_proba(data[data.columns[
[3,9,13,14,16,19,32,37,39]]]])
454.             predict_gsd = models.loaded_model_gsd.predict_proba(data[data.columns[
[13,14,17,16,39]]]])
455.             predict_nkov = models.loaded_model_nkov.predict_proba(data[data.columns[
[2,11,14,15,16,37]]]])

```



```

456.
457.         data[data.columns[41]] = models.loaded_model_gsd.pre-
            dict(data[data.columns[[13,14,17,16,39]]])
458.         data[data.columns[42]] = models.loaded_model_gt.pre-
            dict(data[data.columns[[7,9,14,21,17,16,19,37,39]]])
459.         data[data.columns[43]] = models.loaded_model_icn.pre-
            dict(data[data.columns[[14,16,38]]])
460.         data[data.columns[44]] = models.loaded_model_nkov.pre-
            dict(data[data.columns[[2,11,14,15,16,37]]])
461.
462.         predict_prpo = models.loaded_model_prpo.pre-
            dict_proba(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
463.
464.         data[data.columns[45]] = models.loaded_model_prpo.pre-
            dict(data[data.columns[[3,8,12,14,15,21,16,18,37,39,41]]])
465.
466.         predict_sposob = models.loaded_model_sposob.pre-
            dict(data[data.columns[[4,8,14,16,37,39,38]]])
467.         predict_break = models.loaded_model_break_crio.pre-
            dict_proba(data[data.columns[[1,5,6,7,8,14,15,16,37,38,41,43,44,64,61]]])
468.         predict_srok = models.loaded_model_srok.predict(data[data.col-
            umns[[14,16,37,39,38,43,45]]])
469.
470.         data[data.columns[46]]= predict_sposob
471.         data[data.columns[47]]= predict_srok
472.
473.         predict_groups = models.loaded_model_groups_crio.pre-
            dict(data[data.columns[[14,16,27,37,39,46,47,52,53,64,61]]])
474.
475.         res.result_in_listWidget(self.listWidget_4,predictpr,
476.                                predict_icn,predict_gt,predict_pp,pre-
            dict_gsd,predict_nkov,predict_prpo,
477.                                predict_break,predict_srok,pre-
            dict_sposob,predict_groups)
478.
479.         if ((predict_groups == 2) or (predict_groups == 3)):
480.             predict_a00_b99 = models.loaded_model_a00_b99.pre-
                dict(data[data.columns[[7,10,14,15,16,37,38,46,47]]])
481.             predict_c00_d48 = models.loaded_model_c00_d48.pre-
                dict(data[data.columns[[12,14,15,16,27,37,39,41,46,47]]])

```

```

482.         predict_d50_d89 = models.loaded_model_d50_d89.pre-
        dict(data[data.columns[[1,2,3,8,11,12,14,21,16,27,37,39,41,46,47]]])
483.         predict_e00_e90 = models.loaded_model_e00_e90_crio.pre-
        dict(data[data.columns[[9,14,21,16,32,37,38,46,47,65,67]]])
484.         predict_g00_g99 = models.loaded_model_g00_g99.pre-
        dict(data[data.columns[[11,14,16,18,37,39,38,46,47]]])
485.         predict_h00_h59 = models.loaded_model_h00_h59.pre-
        dict(data[data.columns[[2,8,14,15,16,27,32,37,39,46,47]]])
486.         predict_h60_h95 = models.loaded_model_h60_h95.pre-
        dict(data[data.columns[[3,8,12,14,17,16,18,27,37,41,43,46]]])
487.         predict_i00_i99 = models.loaded_model_i00_i99.pre-
        dict(data[data.columns[[7,8,11,12,14,16,32,39,38,46,47]]])
488.         predict_j00_j99 = models.loaded_model_j00_j99_crio.pre-
        dict(data[data.columns[[3,9,12,14,15,21,16,27,37,42,46,47,61,64]]])
489.         predict_k00_k93 = models.loaded_model_k00_k93.pre-
        dict(data[data.columns[[14,21,16,18,37,39,41,47]]])
490.         predict_l00_l99 = models.loaded_model_l00_l99.pre-
        dict(data[data.columns[[11,14,16,37,39,41,46,47]]])
491.         predict_m00_m99 = models.loaded_model_m00_m99_crio.pre-
        dict(data[data.columns[[8,14,16,32,37,39,38,41,42,43,46,47,65,66]]])
492.         predict_n00_n99 = models.loaded_model_n00_n99.pre-
        dict(data[data.columns[[7,8,12,14,21,16,18,47]]])
493.         predict_p00_p96 = models.loaded_model_p00_p96_crio.pre-
        dict(data[data.columns[[11,12,14,15,21,16,37,39,41,42,43,46,47,62,64]]])
494.         predict_q00_q99 = models.loaded_model_q00_q99.pre-
        dict(data[data.columns[[11,12,14,16,37,39,43,46,47]]])
495.
496.
497.         res.diagnosis_result(self.listWidget_4,predict_a00_b99,pre-
        dict_c00_d48,predict_d50_d89,
498.                             predict_e00_e90,predict_g00_g99,predict_h00_h59,
499.                             predict_h60_h95,predict_i00_i99,pre-
        dict_j00_j99,predict_k00_k93,
500.                             predict_l00_l99,predict_m00_m99,pre-
        dict_n00_n99,predict_p00_p96,predict_q00_q99)
501. if __name__ == '__main__':
502.     app = QApplication([])
503.     window = Ui()
504.     window.show()
505.     app.exec_()

```