

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

кандидата технических наук, доцента Хальясмаа Александры Ильмаровны на диссертационную работу Синотовой Светланы Леонидовны на тему «Разработка математического и алгоритмического обеспечения системы поддержки принятия врачебных решений для прогнозирования здоровья ребенка, зачатого при помощи вспомогательных репродуктивных технологий», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика.

1. Актуальность темы исследования

Тема диссертационной работы актуальна, поскольку вспомогательные репродуктивные технологии (ВРТ) являются востребованной мерой повышения рождаемости, но до сих пор имеет место большое число неуспешных протоколов ВРТ и противоречивые мнения по поводу здоровья рожденных детей. Внедрение в клиническую практику системы поддержки принятия врачебных решений (СППВР) для прогнозирования здоровья ребенка, зачатого при помощи ВРТ, позволит скорректировать ведение протокола ВРТ, а также спрогнозировать осложнения течения беременности у пациентки и возможные заболевания у будущего ребенка, таким образом увеличив число успешных протоколов ВРТ. Несмотря на многочисленные исследования в области прогнозирования результата применения ВРТ, нет единой СППВР, которая позволяла бы охватывать широкий спектр входных факторов и осуществлять прогнозирование долгосрочного здоровья ребенка на каждом из этапов проведения ВРТ.

2. Научная новизна полученных результатов

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 до трех лет.

3. Обоснованность и достоверность научных положений, сформулированных в диссертации

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

Основные положения и результаты работы докладывались и обсуждались на научных семинарах кафедры интеллектуальных информационных технологий ИнФО УрФУ, кафедры информационных технологий и систем управления ИРИТ-РТФ УрФУ, а также на международных научных 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. По теме диссертации опубликовано 11 научных работ, в том числе 5 статей в рецензируемых научных изданиях, определенных ВАК РФ и Аттестационным советом УрФУ, из них 1 статья в издании, индексируемом в международных библиографических базах WoS и Scopus, а также 1 свидетельство о государственной регистрации программы для ЭВМ.

4. Практическая значимость результатов диссертации

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. Результаты диссертационной работы используются в АО «Центр семейной медицины», г. Екатеринбург, Россия.

5. Оценка содержания диссертации и её оформления

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

Для достижения поставленной цели работы были сформулированы и решены четыре 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 признакового 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.

Каждой задаче соответствует своя глава в тексте диссертационной работы.

Первая глава посвящена анализу состояния предметной области, обоснованию необходимости создания СППВР для прогнозирования здоровья ребенка, зачатого при помощи ВРТ. Показано, что требуемое математическое и алгоритмическое обеспечение СППВР должно представлять собой набор иерархических моделей машинного обучения, а предоставленные для разработки СППВР данные нуждаются в обработке. Проведен анализ алгоритмов отбора признаков для машинного обучения и анализ методов, используемых для решения задачи машинного обучения с учителем. Описаны методы и алгоритмы, использованные в работе.

Во второй главе представлена созданная база данных, объединяющая информацию от всех медицинских организаций, и описана разработка алгоритмического обеспечения для предобработки исходных данных. Представлен алгоритм получения итогового набора данных, алгоритм обработки текстовых данных, включающий в себя алгоритм исправления орфографических ошибок и кластеризации слов в семантические группы. Описаны особенности обработки каждой предоставленной таблицы данных.

В третьей главе описана разработка математического и алгоритмического обеспечения СППВР. Представлены выходные и входные данные для каждого блока СППВР. Далее в работе описаны структура и функциональная схема работы СППВР. Представлен процесс создания базы моделей, перечислены все модели машинного обучения в составе СППВР и показано их распределение по блокам системы.

Четвертая глава содержит результаты применения алгоритмов предобработки данных исследования, описание процесса создания моделей машинного обучения и полученные результаты, описание процесса создания программного приложения, позволяющего работать с СППВР, и данные апробации разработанного приложения.

В заключении содержатся выводы о достижении поставленной цели диссертационной работы. После заключения написаны перспективы дальнейшей разработки темы исследования.

6. Замечания и вопросы по диссертации

Следует отметить некоторые замечания и вопросы по содержанию диссертационной работы:

1. Недостаточно подробно описаны используемые модели данных.

1.1. Представленные на Рисунках 1.4 и 2.5 схемы данных содержат некорректные названия сущностей предметной области. АО «ЦСМ», ГАУЗ СО «ОДКБ», ГАУЗ СО «МКЦ ««Бонум» являются названиями медицинских организаций, а не сущностей, следовало бы использовать «Протокол ВРТ», «Электронная карта ребенка» и т.д.

1.2. Описывается процесс предварительной обработки данных и объединения нескольких таблиц, но не приведены модели данных до и после него. Из схемы на Рисунке 2.5 следует, что поле ID ребенка не может быть пустым, но в тексте на стр. 51 указано, что может. В представленных на Рисунках 1.4 и 2.5 схемах поля, представляющие диагнозы, являются текстовыми, а после предобработки формируется множество бинарных признаков, как показано на Рисунке 2.12. Итоговая схема данных описана в тексте работы, но не визуализирована.

2. Выбор моделей машинного обучения ограничен логистической регрессией и двумя ансамблевыми методами.

2.1. На стр. 40 указано: «Для табличных данных модели на основе случайного леса показывают лучший результат прогнозирования, чем ... XGBoost» и там же: «Алгоритмы на основе бустинга... могут дать более высокое качество, чем логистическая регрессия, случайный лес». Одновременно утверждается превосходство случайного леса над бустингом и превосходство бустинга над случайным лесом.

2.2. Утверждение на стр. 40 «Количество деревьев в ансамбле и скорость обучения имеют обратную зависимость: чем меньше скорость обучения, тем больше деревьев требуется для построения модели аналогичной сложности» является спорным, оно не подтверждается ссылками на источники или результатами собственных исследований.

2.3. Неясно, почему не проведен сравнительный анализ большего числа моделей, таких как экстремальный градиентный бустинг, адаптивный бустинг, метод опорных векторов, многослойный перцептрон.

3. Результаты работы, посвященные отбору признаков, представляют большой интерес, но описаны недостаточно полно.

3.1. В Таблице 3 на стр. 77 приведены результаты различных алгоритмов отбора признаков. Насколько совпадают между собой полученные различными алгоритмами наборы признаков?

3.2. В Таблице 3 на стр. 77 показано, что метод отбора признаков «Voruta» выбрал лишь три признака. Затем при использовании отобранных признаков был сделан выбор в пользу случайного леса для всех решаемых далее задач. Корректно ли сделан этот выбор, если, во-первых, в ряде задач число отобранных признаков намного больше трех, как показано в Приложении Б, во-вторых, анализ проведен для задачи классификации, а часть решаемых далее задач относятся к задачам регрессии?

3.3. Отсутствует интерпретация результатов отбора признаков.

3.4. Выполнена ли верификация, показывающая, что отбор признаков повышает точность прогнозирования для всех решаемых далее задач машинного обучения?

4. Ряд вопросов вызывают используемые метрики и их значения.

4.1. В диссертационной работе указано: «Оценка значения метрик получена на тестовых данных с помощью использования перекрестной проверки ... использована кросс-валидация на 5 блоков». Здесь перекрестная проверка и кросс-валидация – это одно и то же? Как именно было выполнено разделение выборки на часть, которая использовалась для обучения моделей и выбора их гипер-параметров, и тестовую часть?

4.2. В работе не приводится сравнение значений метрик на обучающей, валидационной и тестовой выборках.

4.3. Необходима интерпретация относительно низких значений коэффициента детерминации 0,436 (стр. 95), 0,467 (стр. 98), значений в Таблице 12 на стр. 100. Является ли точность моделей достаточной для практического использования и намного ли она выше по сравнению с наиболее простыми моделями, такими как линейная регрессия?

4.4. В задачах медицинской диагностики большое внимание следует уделять объяснимости моделей машинного обучения. В работе этот аспект не рассматривается. Судя по результатам сравнения логистической регрессии и ансамблей деревьев решений на стр. 77, логистическая регрессия уступает в точности, но не критично. В то же время это одна из наиболее объясняемых моделей машинного обучения. С учетом вышесказанного не будет ли выбор логистической регрессии более обоснованным по совокупности критериев?

5. Содержит ли разработанное программное приложение проверку корректности введенных данных и защиту от подмены моделей машинного обучения?

Указанные замечания направлены на совершенствование работы и не снижают общего положительного впечатления от диссертации.

7. Общая характеристика работы

Диссертационная работа выполнена на актуальную тему. В работе проведен анализ предметной области, формализована поставленная пользователями задача, в полной мере описаны материалы и методы исследования, представлены алгоритмы предобработки данных исследования, описаны процессы разработки СППВР в целом и ее составляющих, представлены результаты апробации.

Диссертационная работа Синотовой Светланы Леонидовны хорошо и логично структурирована, соответствует требованиям, предъявляемым к научным работам. Автореферат полностью отражает содержание текста диссертации.

8. Заключение по работе

Исследовательская работа изложена грамотным научно-техническим языком, в полной мере отвечает требованиям по актуальности, научной новизне, практической значимости, личному вкладу автора, отражению результатов в публикациях, а также полностью соответствует п. 9 Положения о присуждении ученых степеней в УрФУ и специальности 2.3.1. Системный анализ, управление и обработка информации, статистика. Автор диссертации Синотова Светлана Леонидовна заслуживает присуждения учёной степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика.

Официальный оппонент:

Кандидат технических наук, доцент

Заведующий лабораторией цифровых двойников в электроэнергетике Уральского энергетического института,

Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»,

Тел.: +7 (343) 375-47-51

e-mail: a.i.khaliasmaa@urfu.ru

Адрес: 620002, Свердловская область, г. Екатеринбург, ул. Мира, 19

Хальясмаа Александра Ильмаровна

(подпись)

Подпись
заверяю



10.02.2023
НАЛЬНИК
МИНИСТРАТИВНОГО ОТДЕЛА
Е. Л. ЗИНОВЬЕВА