МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической теории
упругости и биомеханики

Проектирование и частичная реализация системы выявления параметров для прогнозирования результатов лечения для СППВР

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 442 группы

направления 09.03.03 – Прикладная информатика

механико-математического факультета

Задворной Ольги Александровны

Научный руководитель
к.ф.-м.н., доцент _____ Л.В. Бессонов
подпись, дата

Зав. кафедрой
д.ф.-м.н., профессор _____ Л. Ю. Коссович
подпись, дата

Саратов 2021

Введение. Работа раскрывает тему выявления наиболее значимых параметров для прогнозирования результатов лечения. На данный момент мир живет в режиме постоянного накопления данных, в том числе и в медицинской сфере. Различные показатели и данные опросников пациентов представляют собой большой объем информации, которая со временем только увеличивается. *Big Data* или *большие данные* представляют из себя внушительный объем информации, увеличивающийся быстрее, чем его можно обработать с помощью традиционного прикладного программного обеспечения для обработки данных.

Всё более возрастающее значение приобретает информационное обеспечение самых различных медицинских задач. Однако не все из возможных направлений разработки на данный момент получили развитие. Медицинские информационные системы и системы предоперационного планирования составляют реализованную важную основу технологий. Полноценные системы прогнозирования и оценки в рамках СППВР в таком случае предстают актуальным и логичным следующим шагом разработок, способным привести к дальнейшему качественному скачку в качестве оказания медицинских услуг.

В бакалаврской работе раскрываются вопросы по темам методов анализа данных и их возможного применения при выявлении параметров для дальнейшего прогнозирования. Определяется, что они собой представляют. Рассматривается на практике комбинация изученных алгоритмов на примере медицинской выборки, состоящей из 30 случаев.

Целью данной работы является реализация системы выявления параметров для прогнозирования результатов лечения для СППВР и анализ существующих алгоритмов.

В ходе работы будут рассмотрены следующие задачи, способствующие осуществления обозначенной цели:

- Проанализировать предметную область и существующие в ней решения;
- Оценить роль систем поддержки принятия врачебных решений (СППВР) в медицине;
- Исследовать методы анализа данных;

- Выявить возможности применения и пути использования данных методов;
- Изучить исходные данные для анализа;
- Реализовать систему алгоритмов для выявления параметров;
- Визуализировать полученные модели и данные;
- Проверить полученную модель на реальных данных из выборки;
- Спроектировать интерфейс разрабатываемой системы.

Работа состоит из введения, трёх глав, заключения, списка использованных источников и двух приложений. В первой главе «Анализ предметной области» даются общие сведения об области применения и роли систем поддержки принятия врачебных решений, проводится анализ существующих решений и оценивается применение разрабатываемой системы. Вторая глава «Описание методов и данных» содержит описания основных алгоритмов, возможных для выявления параметров для прогнозирования, а также в ней начинается описание реализации с представления используемой медицинской выборки. Третья глава «Проведение анализа данных и выявление параметров для прогнозирования результатов лечения для СППВР» посвящена непосредственно применению выбранной системы алгоритмов и визуализации полученных данных. В ней приводится применение полученной модели для одного из случаев тестовой выборки и реализуется интерфейс создаваемой системы.

Основное содержание работы.

В первой главе «Анализ предметной области» приводятся сведения о медицинской области и текущей роли систем поддержки принятия врачебных решений, анализируются существующие СППВР и рассматриваются возможности применения создаваемой системы.

С началом XXI века тесно связан рост количества заболеваний опорно-двигательной системы (ОДС). В первую очередь это связано с распространением сидячего образа жизни, повышением неестественной нагрузки на позвоночник и полномасштабным входом в нашу жизнь автотранспорта. В хирургии заболевания позвоночника и таза относят к тяжёлым типам заболеваний, которые связаны с определенным риском. Травмы позвоночника нередко чреваты повреждением центральной нервной системы, а таз, в свою очередь, является одним из центральных звеньев ОДС и опорой для внутренних органов человека. Операция, хоть и является крайней мерой лечения, нередко бывает необходима. Она помогает вернуть прежнюю функциональность основной части скелета, избавляет от болей и различных патологических новообразований.

Серьезность и распространенность данных заболеваний и травм также подтверждаются цифрами из статистики. В связи с этим улучшение качества операций в области позвоночника и таза может значительным образом повлиять на общую картину возвращения качества жизни немалой доли трудоспособного населения.

Существует большое количество важных медицинских задач, которые могут быть улучшены или оптимизированы с помощью современных информационных технологий. Одной из центральных задач является уменьшение человеческого фактора за счет снижения числа врачебных ошибок.

Среди направлений подобного информационного обеспечения здравоохранения особенно стоит отметить разработку *систем поддержки принятия врачебных решений* (СППВР). Их определение и описание было рассмотрено в данном разделе.

СППВР необходима, для того чтобы помогать врачу в диагностике заболевания, в процессе проведения лечения. Подобная система должна производить поиск подобных случаев травм или заболеваний в медицинской практи-

ке. Должен производиться контроль оказываемой пациенту помощи и должна быть дана возможность планирования дальнейшей терапии.

Предусматривается применение таких систем на всех этапах работы с пациентами в медицине.

Наиболее эффективная СППВР включает в себя все четыре отдельных компонента, способных работать как цельный комплекс:

1. Систему предоперационного планирования.
2. Систему биомеханического моделирования.
3. Медицинскую информационную систему (медицинскую базу данных).
4. Систему прогнозирования.

Стоит отметить, что в настоящее время полноценную реализацию получили только *медицинские информационные системы* и системы, относящиеся к предоперационному *геометрическому планированию*, однако они составляют лишь одну из сторон полноценной подготовки к операции.

Потенциально интересной, но мало развитой являются *системы биомеханического моделирования*. Не меньший интерес вызывают *системы прогнозирования*, способные оценить успешность исхода операции на основе статистического анализа данных.

Одной из трудностей создания подобной системы поддержки принятия врачебных решений является необходимость иметь дело с изменяющимися большими данными. Поэтому перед разработкой системы стоит в том числе и задача отбора тех особо значимых параметров.

Важно предложить подход к выбору модели в отношении определения количества значимых факторов, использовать несколько разных алгоритмов на разных стадиях анализа данных, чтобы сравнивать эффективность нескольких критериев выбора оптимального количества факторов и использовать результаты предыдущей модели в последующей для увеличения достоверности полученного результата, а также всестороннего изучения данных.

Наконец, важна демонстрация результатов в удобной визуальной форме. Для этого важно разработать дополнительно интерфейс для представления всех необходимых данных в максимально компактном формате без лишней информации.

Во второй главе «Описание методов и данных» содержится описание некоторых алгоритмов и рассмотрение возможности их применения, а также изучаются исходные данные.

В первом разделе описываются такие виды анализа, как корреляционный, кластерный, дискриминантный, а также алгоритм «Деревья решений». Перед анализом данных важно провести их дополнительную первоначальную обработку для того, чтобы убрать наименее значимые показатели, исходя из их корреляции. А после важно также провести сам всесторонний анализ данных для поиска наиболее значимых параметров.

Корреляционный анализ — это метод, который используется в первую очередь, чтобы проверить гипотезу о статистической значимости всех или некоторых параметров выборки. С его помощью можно изучить зависимости в первоначальном наборе данных.

В соответствующем подразделе были описаны решаемые с помощью корреляционного анализа задачи. Такой метод анализа может применяться в тех случаях, когда есть большая выборка данных, в которой параметры являются количественными и измеримыми в понятных величинах. В зависимости от исходных параметров применяются различные коэффициенты корреляции. Когда создаётся многофакторная корреляционная модель выбирают показатели, которые имеют коэффициент парной корреляции не более 0.85. Результат модели может быть представлен в разном виде.

Таким образом, данный метод нужен в тех случаях, когда задан большой объём схожих параметров, чтобы первоначально очистить их от коррелирующих показателей, чтобы не создавать лишнего шума в данных, а также чтобы значительно сократить количество значимых для анализа и выявления результирующего показателя параметров.

Кластерный анализ — набор методологий для разделения первоначального набора данных в несколько групп таким образом, чтобы выборки внутри одной группы были аналогичны, а элементы, принадлежащие разным, непохожи.

В данном подразделе рассмотрены задачи кластерного анализа, типы входных данных для его применения и один из распространенных методов подсчета — Евклидова расстояния.

Существует множество различных подходов и алгоритмов для выполнения задач кластеризации. Одна из проблем, связанных с кластеризацией, заключается в том, что в большинстве случаев неизвестно количество кластеров в наборе данных до начала анализа. Для того чтобы решить эту проблему, в работе использовалась иерархическая кластеризация. Были изучены понятия иерархической кластеризации и агломеративного метода кластеризации, а также приведен пример дендродраммы, строящейся по его итогу.

Методология кластеризации подходит для исследования взаимосвязей для предварительной оценки структуры выборки и особенно важно использовать там, где нет четкого финализирующего показателя, по которому можно было бы оценить достоверность результата. Поэтому она будет использоваться на втором этапе анализа данных для выделения схожих случаев с пациентами и общего анализа имеющегося набора данных.

Дискриминантный анализ в первую очередь нужен для того, чтобы выяснить, по каким элементам расходятся (дискриминируют) отдельные совокупности (группы). При дискриминантном анализе группы известны заранее, что является его отличительной особенностью.

Его основная идея состоит в определении различий между совокупностями и проверке, отличаются ли совокупности по среднему какой-либо переменной (или линейной комбинации переменных).

В соответствующем подразделе приведено описание исходных данных для дискриминантного анализа, его задач, построение основной дискриминантной функции и алгоритм проведения анализа. Дискриминантный анализ был сравнен с кластерным анализом.

Линейный дискриминантный анализ может быть использован для уменьшения размерности входных данных путем проецирования их в наиболее различительных направлениях с использованием метода преобразования.

Таким образом, данный метод нужен в первую очередь для того, чтобы проверить и проанализировать уже имеющуюся структуру разделения, которая может быть получена, например, при проведении кластерного анализа. И этот метод, помимо прочего, является хорошим вариантом визуализации моделей.

Дерево решений — непараметрический контролируемый метод обучения, используемый для классификации и регрессии. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных.

В подразделе рассматривается набор данных для построения дерева решений, преимущества данного метода и основные типи деревьев решений, включая методы, использующие несколько таких моделей. Среди алгоритмов отдельно был рассмотрен «случайный лес».

Подробнее был рассмотрен один из основных параметров — средневзвешенное загрязнение `Gini` (Джини).

Последний алгоритм позволяет максимально точно отсортировать данные благодаря возможностям случайного леса к минимизации ошибки. Этот метод подходит для оптимального выбора параметров, потому что в каждом узле можно отследить параметр разбиения выборки. В том числе можно использовать данные других моделей для улучшения работы алгоритма.

Исходные данные составляю записи о 30 пациентах в общей сложности более, чем по ста параметрам. Возраст пациентов при этом варьировался от 18 до 57 лет и в среднем составил 36,5 лет. Операции проходили в период с 2016 по 2019 гг. Срок до последующего заполнения опросников составил от 1,5 месяцев до 2,8 года.

Во втором разделе подробно описываются имеющиеся в выборке общие и предоперационные параметры, такие как возраст, под, механизм травмы, пульс и т. д.

Помимо прочего в исходных данных представлены результаты заполнения опросников SF-36 и S.A. Majeed и количество дней от получения травмы до их непосредственного заполнения.

В третьей главе «Проведение анализа данных и выявление параметров для прогнозирования результатов лечения для СППВР» показываются полученные результаты работы с исходными данными и реализованная система.

В качестве эксперимента для достижения поставленной цели было рассмотрено последовательное применение алгоритмов, описанных ранее.

Были представлены используемые в работе средства и библиотеки. Вся работа была выполнена на языке `Python`.

Перед построением основных моделей необходимо произвести очистку исходных данных. С этой целью была построена специальная *корреляционная матрица*. Этот вид анализа позволит выявить скрытую корреляцию данных и в дальнейшем отказаться от схожих параметров.

В соответствующем подразделе был приведен и рассмотрен кусок кода с использованием данного алгоритма. Для большего удобства на второй схеме параметры были сгруппированы. Подобная группировка позволяет более наглядно увидеть, какие параметры образуют кластеры по «похожести» и определиться с их включением в дальнейший анализ.

Было выяснено, что наибольшие коррелирующие группы составляют параметры, относящиеся к постоперационным опросникам. Это замечено вполне справедливо, так как отдельные вопросы взаимосвязаны и складываются затем в единую результирующую оценку по рассматриваемым опросникам.

Первоначально в выборке хранится информация о 30 примерах с 88 параметрами. Сократив коррелирующие группы параметров, анализ был проведен повторно, на этот раз рассматривая только 22 параметра. Получившиеся матрицы также были приведены в работе. Таким образом, корреляционный анализ позволил сократить первоначальное количество параметров в 4 раза.

Очистив первоначальные данные, мы обратились к *кластерному анализу* с целью выделить отдельные группы пациентов по схожести их предоперационного состояния.

Для объединения случаев была проведена иерархическая кластеризация исходных данных. Это позволило визуально увидеть объединение отдельных случаев в кластеры по схожим предоперационным параметрам.

Кусок кода с данным анализом и полученная дендрограмма приведены в соответствующем разделе. Из полученного графика видно, что данные делятся на три кластера при евклидовом расстоянии более 100. Получившиеся кластеры были сравнены со средней оценкой результата лечения пациентов, имеющейся в исходных данных, и пронумерованы в порядке увеличения такой оценки.

Таким образом, кластерный анализ позволил разделить данные на кластеры по предоперационным данным и получилось оценить значение кластеров.

Следующим был применен *дискриминантный анализ* для визуализации получившихся в прошлом подразделе кластеров. Это позволило удостовериться в качестве полученного разделения.

Код проведения и визуализации результата дискриминантного анализа приведен в соответствующем подразделе. Результат работы дискриминантного анализа был представлен на графике. На нем можно увидеть четкое разделение трех кластеров, что подтвердило сделанные в прошлом подразделе выводы.

Полученное разделение на кластеры было использовано в качестве меток для построения *модели «случайного леса»*. Это позволило оценить параметры по их вкладу в разделение данных.

В соответствующем подразделе был приведен код обучение модели «случайного леса». Несколько из построенных в ходе анализа деревьев были приведены на рисунках. Таким образом, получилось перейти от первоначальной таблицы к дереву.

В рамках выявления параметров было проанализировано в общей сложности 50 деревьев решений. Результатом анализа стало определение следующих предоперационных параметров как значимых: пульс, возраст, АД систолическое, койко-день, BE(act), тяжесть повреждения. Данные параметры встречались на корневом уровне деревьев и близких к нему узлам чаще других и вносили больший вклад в итоговое предсказание.

Был приведен *пример* использования одной из полученных моделей на неиспользованном при обучении случае.

В ходе всестороннего анализа исходных данных были выявлены наиболее значимые для рассматриваемой выборки параметры, которые можно будет использовать для дальнейшего полноценного прогнозирования результатов лечения в рамках СППВР.

Одной из важных составляющих любой системы является удобный для пользователя *интерфейс*. В работе был представлен общий вид интерфейса

реализованной программы, некоторые ее отдельные окна и описаны ее текущие возможности.

Далее были рассмотрены варианты исходных параметров и некоторые реализованные в программе функции.

У системы прогнозирования потенциально можно выделить три направления оценки успешности лечения: геометрическое соответствие норме, уменьшение уровня боли, восстановление качества жизни. Каждое из данных направлений было рассмотрено подробнее.

Геометрическое соответствие норме заключается в оценке приближенности послеоперационных показателей к нормальным показателям до получения травмы.

Уменьшение уровня боли. Успешность операции в первую очередь оценивается по ослаблению болей или полному от них освобождению как по основному параметру, на который жалуются пациенты, нуждающиеся в операции на позвоночнике.

Была рассмотрена одна из распространенных шкал интенсивности боли — визуальную аналоговую шкалу боли (ВАШ, VAS). Эта шкала также была реализована в программе.

Восстановление качества жизни оценивается по специальным анкетам качества жизни. Была рассмотрена самая широко применяемая из них — анкета качества жизни Освестри (ODI), которая также присутствует в программе.

Дополнительно была отмечена *оценка исхода лечения*, являющаяся важным критерием исхода лечения и измеряющая «удовлетворенность» пациента. Одна из наиболее часто упоминаемых таких шкал — *субъективная оценочная шкала Макнаб* (Macnab).

Отдельно были подробно рассмотрены еще две шкалы, которые были реализованы в программе: опросник SF-36 и шкала S.A. Majeed.

В приложениях представлены полные перечени вопросов и их оценки для опросников SF-36 и S.A. Majeed.

Заключение. По итогу выполнения данной работы была проведена реализация системы выявления параметров для прогнозирования результатов лечения для СППВР.

В ходе написания бакалаврской работы было совершено следующее: изучены основные виды систем поддержки принятия врачебных решений, используемых сейчас в медицине; рассмотрена важность СППВР при оптимизации врачебной деятельности; изучены некоторые методы анализа данных; оценены возможности применения рассматриваемой системы в рамках СППВР для прогнозирования; изучены исходные данные и произведен анализ полученного результата на примере реального случая. По итогу была реализована система алгоритмов для выявления параметров для дальнейшего прогнозирования и разработан интерфейс подобной системы.

Таким образом, была дана оценка текущему состоянию развития технологий СППВР и была рассмотрена система алгоритмов, с помощью которой можно выявить данные для прогнозирования. Данная работа представляет собой один из возможных путей развития систем прогнозирования. Была доказана важность СППВР в медицине, чьё внедрение в клиническую практику позволит повысить качество медицинской помощи и оптимизировать расходы здравоохранения.