Министерство образования и науки РФ

ФГБОУ ВО «Омский государственный технический университет»

Кафедра «Автоматизированные системы обработки информации и управления»

**ОТЧЕТ**

**О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ**

по теме «Применение нечеткой логики при ранней диагностике заболеваний

больных неалкогольной жировой болезнью печени»

студента Серобабова Александра Сергеевича группы ИВТм-172

**Пояснительная записка**

Направление 09.04.01

Зав кафедрой, д. т. н., проф.

\_\_\_\_\_\_\_\_\_\_\_\_ Никонов А.В.

« » \_\_\_\_\_ 20\_\_ г.

Научный руководитель, д. т. н., проф.

\_\_\_\_\_\_\_\_\_\_\_\_ Денисова Л.А.

« » \_\_\_\_\_ 20\_\_ г.

Разработал студ.

\_\_\_\_\_\_\_\_\_\_\_ Серобабов А.С.

« » \_\_\_\_\_ 20\_\_ г.

Омск 2019

Реферат

Пояснительная записка 66 с., 17 рис., 16 табл., 3 прил, 78 ист.

НЕЧЕТКИЕ ПОДМНОЖЕСТВА, МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА, НЕЧЕТКИЙ КЛАССИФИКАТОР, ЭКСПЕРТНАЯ СИСТЕМА, НЕАЛКОГОЛЬНАЯ БОЛЕЗНЬ ПЕЧЕНИ,

Объектом анализа настоящей работы является сочетание методов статистического анализа данных с применение нечеткой логики для ранней диагностики неалкогольной жировой болезни печени (НАЖБП).

Цель работы – с помощью применений нечеткой логики повысить процент правильности постановки диагноза заболевания печени на ранней стадии развития.

В процессе работы установлена актуальность и значимость работы. Выделены основные проблемы в выявлении болезни на ранней стадии. Проанализированы аналогичные решения проблемы и выделены ключевые особенности предлагаемого подхода. Произведен первичный анализ данных, где выявлены значащие признаки для построения экспертной системы.

**Содержание**

[Список используемых определений и сокращений 4](#_Toc1640141)

[Введение 5](#_Toc1640142)

[1. Анализ состояния проблемы 6](#_Toc1640143)

[1.1 Актуальность исследования 6](#_Toc1640144)

[1.2 Обзор предметной области и методов диагностирования болезни печени. 7](#_Toc1640145)

[1.3 Научная новизна работы 11](#_Toc1640146)

[1.4 Практическое использование полученных результатов 11](#_Toc1640147)

[1.5 Методология и методы исследования 12](#_Toc1640148)

[1.6 Методы выявления достоверности результатов 12](#_Toc1640149)

[2 Обоснование проектирования и постановка задач 13](#_Toc1640150)

[2.1 Постановка цели и задач 13](#_Toc1640151)

[3 Исследование ранней диагностики заболевания печени 14](#_Toc1640152)

[3.1 Первичный анализ данных 16](#_Toc1640153)

[3.1.1 Предобработка данных 18](#_Toc1640154)

[3.1.2 Описательная статистика 23](#_Toc1640155)

[3.1.3 Построение диаграммы размаха 27](#_Toc1640156)

[3.1.4 Регрессионный анализ 29](#_Toc1640157)

[3.2 Модуль «Выявление значащих параметров для экспертной системы» 30](#_Toc1640158)

[3.3 Проверка физических и лабораторных параметров на соответствие нормальному распределению 34](#_Toc1640159)

[3.4 Обобщение результатов и построение графа зависимости 37](#_Toc1640160)

[3.5 Экспертная система 38](#_Toc1640161)

[3.5.1 Проектирование экспертной системы ранней диагностики заболевания с использованием нечеткого классификатора 39](#_Toc1640162)

[3.5.2 Создание базы правил экспертной системы 41](#_Toc1640163)

[3.5.3 Алгоритм работы модуля нечеткого классификатора 45](#_Toc1640164)

[Заключение 48](#_Toc1640165)

[Список использованных источников 49](#_Toc1640166)

[Приложение А 53](#_Toc1640167)

[Приложение Б 54](#_Toc1640168)

[Приложение В 55](#_Toc1640169)

[Приложение Г 56](#_Toc1640170)

# Список используемых определений и сокращений

НАЖБП – неалкогольная жировая болезнь печени.

ЭД – Эндотелиальная дисфункция.

ММП9 – матриксная металлопротеиназа-9.

Фиброз – разрастание соединительной ткани с появлением рубцовых изменений в различных органах, возникающие, как правило, в результате хронического воспаления.

ALT – Аланинаминотрансфероза.

AST – Аспартатаминотрансфераза.

ВОЗ – Всемирная организация здравохранения.

Нечеткое подмножество – множество A состоящее из совокупности упорядоченных пар, составленных из элементов x универсального множества X и соответствующих степеней принадлежности.

АЖБП – алкогольная жировая болезнь печени.

# Введение

Применение нечеткой логики, как инструмент классификации обеспечивает принципиально новый подход к проектированию экспертных систем, в которых данные носят слабо прослеживающуюся взаимосвязь, а искажения, вызываемые внешними факторами, являются недостаточно определенными или не поддаются точному математическому описанию.

Ввиду того, что на современном этапе развития медицина все чаще использует теоретическую базу, опирающуюся на математический аппарат [24, 31-33], применение принципов нечеткой логики является перспективным и научно-доказуемым способом в рамках прикладной задачи ранней диагностики заболевания.

На фоне глобального нарастания распространѐнности избыточной массы и ожирения на сегодняшний день неалкогольной жировая болезнь печени занимает ключевое место в списке причин заболеваний печени. Распространенность НАЖБП в мире составляет 20 – 40% по данным литературных источников [64, 65]. По последним данным в России, на основании исследования DIREG2, составляет 37%[3]. Особую тревогу вызывает то, что всего за 7 лет процент больных увеличился на 10% по сравнению с данными исследованием DIREG1 [66, 67].

В настоящее время все чаще врачи прибегают к ранней диагностике заболеваний. Несмотря на бурный рывок в научно-техническом прогрессе, внедрении современных технологий в медицине и широкий выбор, имеющихся методов диагностирования заболевания, высокой точности диагностирования на ранней стадии заболевания печени так и не удалось. Основными сложностями, с которыми сталкиваются методы диагностики, являются: недостаточность данных, искаженность и пропуски данных, скрытое развитие болезни и сложность выявления различий между схожими болезнями.

Данный спектр проблем нивелирует те преимущества, которые дают методы, нацеленные на диагностику заболевания в стадии полного проявления. По этим причинам для данного исследования выбраны методы: статистических анализа данных и нечеткой логики, зарекомендовавшие себя в работе с неполными и неточными данными.

# Анализ состояния проблемы

# 1.1 Актуальность исследования

Ранняя диагностика заболеваний является актуальной проблемой современной медицины. Поскольку многие заболевания на ранней стадии поддаются полному излечению, и в результате сокращается время выздоровление пациента, что благоприятно сказывается на здоровье пациента. Конкретно в этой работе производится ранняя диагностика заболевания пациента с НАЖБП. По данным всемирной организации здравохранения в мире наблюдается отчетливая тенденция к росту заболевания печени. Еще в 19 веке немецкий клиницист Фридрих Теодор фон Фридрихс впервые описал изменения, происходящие в печени у больных сахарным диабетом.[1] В 1952 г. S. Zelman описал изменения печени у пациентов с ожирением [2]. И была подтверждена статистическая зависимость между снижением веса пациента, отклонения печеночных тестов и морфологических изменений.

На основании данных статьи исследования DIREG 1 проведенного в 2007 г. доля лиц с выявленным диагнозом НАЖБП, первично или повторно обратившихся в амбулаторные лечебно-профилактических учреждений. Вне зависимости от причины обращения, включая пациентов с подозрением на диагноз НАЖБП, в России составила 27%. В исследовании же 2014 г. DIREG 2 результаты уже 37.3%[3]. Всего за 7 лет доля больных увеличилась на 10%, такое увеличение имеет эпидемический характер.

Важность проблемы так же определяет, что на ранней стадии заболевания печени отсутствует специфическая клиническая симптоматики. Пациент с болезнью или не проявляет жалоб, или они являются специфичными – в виде слабости, быстрой утомляемости, тошнота, нарушения стула, тяжесть в правом подреберье, тупые боли в правом подреберье. Отсутствие клинических проявлений приводит к тому, что пациенты отменяют или откладывают посещения врача. Само заболевание чаще всего выявляется случайно при исследовании другой патологии больного.

Еще одной сложностью, в классификации заболевания неалкогольной составляющей болезни печени является большое сходство с заболеванием алкогольной болезнью печени, так как отсутствуют объективные

критерии дифференцирования. На данное время лучшим решением является проведение анкетирования или опрос родственников пациента. С этой целью разработаны опросники AUDIT[35] и CAGE(скриннинговая методика оценки хронической алкогольной интоксикации)[36]. Оба теста специфицируются на выявлении алкоголизма у испытуемого. Различия их только в том, что один направлен на выявление группы лиц подверженных риску алогизма, другой же предназначен для выявления систематического употребление алкоголя.

Следует так же отметить, что лица страдающие ожирением или сахарным диабетом 2-го типа, распространенность заболевания печени достигает 70-90%[61]. При отсутствии лечения заболевание может перетечь в следующую стадию болезнь, а в конечном итоге и к циррозу печени.

В настоящее время не разработано строгих алгоритмов ведения пациентов с НАЖБП. Выявлено, что постепенное снижение массы тела и физическая активность является важнейшим составляющим лечения.

Несмотря на рост больных с заболевание печени патогенетические механизмы, лежащие в ее основе, недостаточно изучены. Отсутствуют общепринятые подходы в диагностике и лечении. Взаимосвязь между показателями сахарного диабета 2 степени и болезни печени являются предметом дискуссий. А малая осведомленность людей ведет к процентному увеличению числа больных.

# 1.2 Обзор предметной области и методов диагностирования болезни печени.

Для диагностирования заболеваний используется огромный спектр различных методов:

- Пункционная биопсия печени;

- Ультразвуковое исследование (УЗИ);

- Статические методы прогноза;

Пункционная биопсия печени является точным методом и является своего рода «золотым стандартом» в области гепатологии, но высокий потенциальный риск процедуры и большое количество противопоказаний ограничивает данный метод к его повсеместному использованию. Хирургическая процедура сопровождается 30% ошибкой от всех проведенных операций[37]. Статистика по смертности, которая свидетельствует, что самые высокие показатели после проведения биопсии зарегистрированы в течение 3 месяцев и составляет от 0,2 до 0,4%[10]. Основной причиной послужившей летальных исходов является кровотечение брюшной полости.

Использование ультразвукового исследования для диагностирования заболевания дает хорошие результаты, в этой ветви исследования выявляются новые взаимосвязи между другими заболеваниями организма[6], но данный метод не дает права поставить окончательный диагноз НАЖБП. Он может быть установлен лишь при совокупности анамнестических данных и результатов лабораторных исследований. Взаимосвязь между компонентами метаболического синдрома и НАЖБП[7]. В данной области так же широкое применение приобретает использование ультразвуковых изображений, как входных данных для нейронных сетей, которые выполняют классификацию по болезни печени[63].

Перспективным направлением в области ранней диагностики, является использование систем поддержки принятия врачебных решений (CDDSS) [39-47]. В качестве теоретического базиса выступает теория нечетких множеств, сформированная Лотфи Заде[58]. MYCIN[69] – одна из первых и известных экспертных систем показавшие высокие результаты выступающая в роли помощника терапевта в диагностике инфекционных заболеваний, в частности менингита и бактериальной септицемии. Продолжателем MYCIN является экспертная система CADUCEUS[72,73], которая может диагностировать до 1000 различных заболеваний связанных с кровью.

Современные исследования и создания экспертных систем диагностики базирующиеся на болезнях печени: выявление вируса гепатита[8, 52-53], стадий фиброза печени [38,48-51].

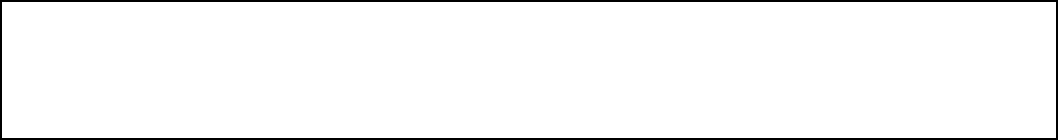
Растущий интерес к использованию нейронных сетей не мог ни отразиться на прикладном использовании в медицине. Созданы системы диагностики заболевания: гепатита на основе SVM (метод опорных векторов)[54-56]; фиброза на основе SVM[57], CCNN(кросс-контрастная нейронная сеть)[59] эволюционно-обученная нейронная сеть[60], ANN(искусственная нейронная сеть)[62].

Анализ литературных данных и результаты собственных исследований позволяют сделать вывод о том, что значительное число задач прогнозирования, ранней медицинской диагностики характеризуется высокой сложностью формализации описания классов состояния здоровья организма, а в задачах прогнозирования и ранней диагностики различные классы сильно пересекаются в пространстве информативных признаков[25].

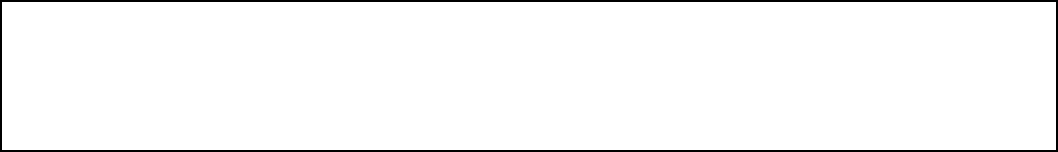
В диагностике медики интересуются не только в выводе, но и в том, как система получает необходимые заключения. Экспертные системы, построенные на основе нечеткого классификатора с использованием нечеткой логики над нейронными сетями – это ее «прозрачность». Каждое правило определено и понятно даже не специалисту информационных технологий. Исходя из явного преимущества нечеткой логики в выявлении связей между совокупностью данных и проверенных временем статистических методов. Дынный метод станет основой в построение нечеткого классификатора.

В основу исследования легли результаты лабораторных анализов 149 больных НАЖБП, предоставленные Омским государственным медицинским университетом. Схема алгоритмы исследования (рисунок 1) поэтапно отражает весь процесс исследования.

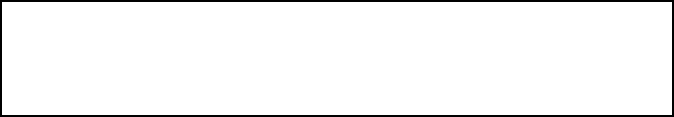
**Формирование исследовательской когорты на основании критериев включения/исключения**



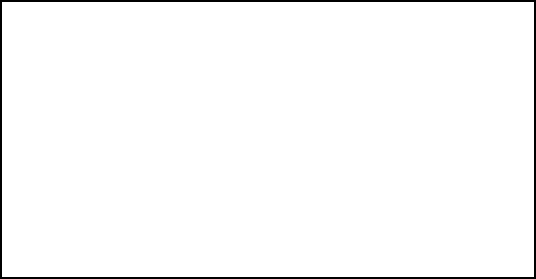
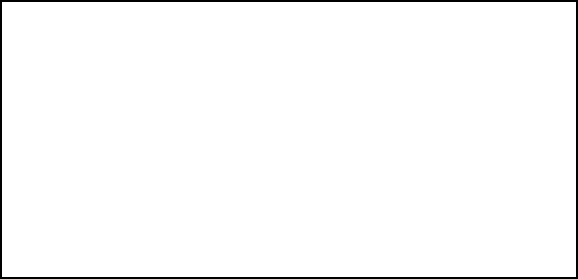
**Общеклиническое, лабораторное исследование, очистка**



**данных от зашумлений, выбросов и дубликатов.**



**Распределение на группы сравнения**

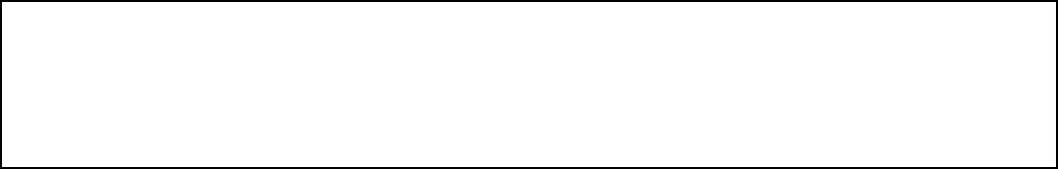


Основная группа

**Лица, страдающие НАЖБП с различной стадией фиброза печени**

Группа сравнения

**Лица, страдающие НАЖБП с отсутствием фибротических изменений в печени**



**Передача данных в экспертную систему**

Рисунок 1 – Алгоритм сбора данных

Данные анализов представляют собой таблицу (приложение А), состоящей из более 100 параметров включающих в себя лабораторные анализы и результаты опросов пациента. Большое количество данных образует проблему их правильной статистической обработки и использование

методов, которые адекватно подходят для решения задач связанных с ранней диагностикой

Данные представлены различными типами, а именно: логическим, целочисленным, числами с плавающей запятой и строковыми данными. В свою очередь сами данные лабораторных исследований имеют пропущенные значения, шумы и выбросы. Такая специфика данных приводит к необходимости подготовки первичных данных и применению большого диапазона статистических методов исследования.

В свою очередь данные первичного анализа станут основой для построения нечеткой системы и статистического анализа по выявления взаимосвязей. Основными методами статистики, используемые в медицине, являются:

- Регрессионный анализ;

- Дескриптный анализ;

- Дисперсионный анализ.

Регрессионный анализ используется для изучения связей между зависимой переменной и одной или нескольких независимых переменных. Используется в случаях: выявления взаимозависимости переменных, предсказания зависимой переменной и для определения значимости переменной. Из недостатков его нельзя использовать для определения наличия связи между переменными, так наличие связи и есть предпосылка для применения анализа.

Дескриптный анализ его главной задачей сформировать представления о исследуемых значениях, но при этом не выдвигать причины или другого рода гипотезы.

Дисперсионный анализ – метод в статистике применяемый для исследования виляние одной или нескольких качественных характеристик.

Таким образом, все перечисленные статистические методы подходят, для использования в работе. Дескриптный анализ охарактеризует результаты лабораторных исследований. Регрессионный анализ будет применен для данных, взаимосвязь которых очевидна и доказана предыдущими работами. Дисперсионный анализ выявит новые взаимосвязи внутри выбранных групп.

Анализ данных основанных на лингвистических переменных, полученных в результате опросов, и кластеризация лабораторных данных будет производиться с помощью нечеткой логикой. Алгоритм выполнения классификации состоит из пяти шагов[11]:

- Представление образа;

- Определение близости образа;

- Кластеризация;

- Абстракция данных;

- Оценка вывода.

Заключительным шагом будет построение нечеткой экспертной системы, основанной на данных полученных в результате этапа статистического анализа. Данные о связях будут учитываться при построении базы лингвистических правил[12].

На основании вышеизложенного представляется актуальным создание экспертной системы, которая позволит решить

# 1.3 Научная новизна работы

Впервые применена нечеткая логика для диагностирования ранней стадии НАЖБП. Выявление взаимосвязи совокупности значений лептина и obr фактору риска развития фиброза.

# 1.4 Практическое использование полученных результатов

Результаты исследования в дальнейшем будут использоваться в практической работе поликлинических учреждений города Омска. Так же в рамках клинической практики, медицинских учреждений любой формы собственности и направленности.

# 1.5 Методология и методы исследования

Научная методология исследования основывается на системном подходе и комплексом воздействующих факторов на раннее выявление заболевания неалкогольной жировой болезни печени. Методологической базой послужили труды отечественных и зарубежных ученых в области диагностики неалкогольной жировой болезни печени.

Работа представляет собой прикладное научное исследование, решающие проблему диагностирования болезни печени на ранней стадии проявления. Путем использования нечетких подмножеств.

Объектом для исследования является когорта пациентов с НАЖБП на основании критериев включения и исключения. Когорта формировалась на амбулаторном этапе из лиц, проходивших диспансеризацию взрослого населения в различных поликлинических учреждениях города Омска. В исследование включались пациенты в возрасте от 18 до 65 лет с диагностированной НАЖБП, подписавшие информированное согласие.

Предмет исследования – изменение показателей взаимосвязи групп лабораторных анализов и физиологических данных пациента. В качестве методов исследования применялись математические (статистические) методы. В качестве приемов, позволяющих найти новые решения

# 1.6 Методы выявления достоверности результатов

Достоверность исследования подтверждается на основе совокупности отношений данных полученных с помощью нечеткой логики и результатов генной типизации.

**1.7 Резюме к разделу 1**

При анализе пациентов ожидается множество проблем связанные с формированием экспериментальной выборки. Поэтому с самого начала исследования необходимо определить размер выборки и критерия ограничения, чтобы уменьшить количество накапливаемых ошибок и затрат ресурсов.

**1.8 Выводы к разделу 1**

Выполнен анализ работ посвященных диагностике в медицине. Сделан вывод, что современные методы диагностирования строятся на подход медицинского толка и с использование статистического, машинного обучения и нечеткой логики. В результате отмечается следующее:

1. Существующий процент заболеваемости носит эпидемный характер и количество подверженных болезней печени с каждым только возрастает.
2. Текущее положение в диагностике заболеваний печени на данный момент имеет методики с недостаточной точностью результатов. Использование биопсии, хоть и является «золотым стандартом», но совершенно не применим в ранней диагностике, и имеет высокий процент ошибки хирургического вмешательства, что приводит к причинению вреда или летальному исходу пациента.
3. Не выявлены параметры пациентов, отвечающие за возникновение болезни и ее дальнейшего хода развития.
4. Выявлена необходимость формирования нового подхода с прикладным использованием нечеткой логики для построения классифицирующего модуля стадий НАЖБП в экспертной системе на основании входных данных о пациенте, позволяющей систематизировать ранее накопленные знания о болезни благодаря созданию экспертной базы правил.
5. Создание экспертной системы на основе нечеткой логики является перспективнее и надежнее, так как вся система представляет собой «белый ящик».

# 

# 2 Обоснование проектирования и постановка задач

# 2.1 Постановка цели

При постановке диагноза пациентам много времени занимает проведение подготовительных работ, взятие анализов и выполнение лабораторных исследований. При отсутствии четкого понимания течения заболевания НАЖБП и ее стадийных дефиниций. Врачу сложно, а иногда невозможно поставить правильный диагноз, чтобы уточнить свой диагноз доктор отправляет пациента на биопсию печени, что является опасной и дорогостоящей процедурой. Чаще всего пациенты с заболеванием печени выявляются при диспансеризации. Поэтому важно иметь инструмент диагностики, который имеет высокую точность постановки диагноза, при этом способен справляться с большим потоком пациентов.

Цель моей работы: с помощью использования нечеткой логики увеличить вероятность правильного диагностирования болезни на ранней стадии проявления.

Целью исследования является выявление новых взаимосвязей влияющих на развитие НАЖБП, которые могут быть получены в результате обработки данных. В качестве основных алгоритмов выявления зависимостей используются: статистические инструменты и нечеткая классификация.

# 2.2 Постановка задачи

Основные задачи исследования:

- Провести изучение и анализ основных технологий диагностирования болезней печени;

- Провести «очистку данных» на наличие дубликатов, выбросов и зашумлений.

- Провести первоначальный анализ решаемой проблемы;

- Выбрать критерии включения и исключения пациентов в когорту больных с НАЖБП;

- Провести статистический анализ данных лабораторных анализов;

- Построить математическую модель системы;

- Выполнить вычислительные эксперименты с нечеткой моделью;

- Группировка входных параметров;

- Проанализировать полученные результаты и сформировать соответствующие выводы;

- Определение категории заболевания печени;

- Коррекция и доработка нечеткой системы.

# 3 Исследование ранней диагностики заболевания печени

В ходе анализа состояния проблемы и обзора предметной области, сделан вывод о необходимости создания автоматизированной системы для ранней диагностики заболевания печени. Система должна повысить процент правильности постановки диагноза заболевания «Неалкогольная жировая болезнь печени» на ранней стадии развития. Для этого необходимо провести исследование данных на предмет «полезной» информации. Полученные данные используются как входные параметры для экспертной системы. Общий алгоритм исследования изображен на рисунке 1.

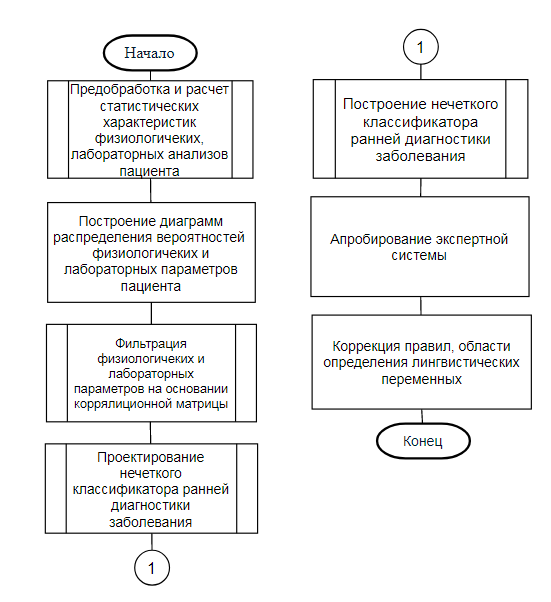


Рисунок 1 – Общий алгоритм исследования выявления взаимосвязи лабораторных, физических параметров, сопутствующих заболеваний и принимаемых препаратов на изменение стадии болезни.

Алгоритм состоит из следующих основных шагов. Вначале вводятся данные подготовленные и предоставленные Омским государственным медицинским университетом.

1. Проверка и фильтрация информации на наличие ошибок и дубликатов. Для этого выборка подвергается разбиению на группы по признаку представляемого типа данных.

Описательная статистика подразумевает под собой обобщение первичных результатов наблюдений и включает в себя четыре этапа:

* Сбор данных;
* Категоризацию данных;
* Обобщение данных;
* Представление данных.

1. Построение диаграмм распределения вероятности, как один из инструментов визуального исследования выборки для физиологических и лабораторных параметров пациента. Главной задачей является выявление зависимостей на основе визуальной оценки. Основой для построения диаграмм выступает «ящик с усами» или по-другому диаграмма размаха. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартиль.
2. Фильтрация параметров выборки *X* не несущих полезной информации. Для этой цели выполняется построение корреляционной матрицы выборки *X*. С помощью анализа корреляций эксперт может установить существует ли зависимость между двумя величинами.
3. Проектирование системы, которая включает в себя формирование базы нечетких правил и выбор функций принадлежности нечетких множеств.
4. Построение системы, для этого используется пакет прикладных программ «MATLAB» и пакет расширения «FuzzyLogic».
5. Апробирование экспертной системы в медицинском учреждение. Для этого проводятся контрольные испытания на новой выборке больных и сверяются процентные соотношение правильно поставленных диагнозов.
6. Коррекция правил, области определения лингвистических переменных по результатам восьмого шага и замечаний экспертов прикладной области.

# 3.1 Первичный анализ данных

Первичный анализ данных – раздел математической статистики, изучающий способы обработки массивов статистических данных с целью нахождения обобщающих характеристик элементов массива[16]. В ходе первичной статистической обработки данных необходимо решить следующие задачи: преобразование параметров в номинальную или порядковую шкалу, статистическое описание исходных совокупностей с определением пределов варьирования переменных, анализ выбросов данных, восстановление пропущенных наблюдений. Для этого необходимо провести анализ основных статистических показателей.

Имеется выборка *X* из 149 пациентов с выявленной неалкогольной жировой болезнью печени (с различными стадиями фиброза), характеризующихся 129 переменными (параметры, взятые из таблицы ). Каждый пациент был отобран в результате диспансеризации населения из различных поликлинических учреждений города Омска. В исследование включались пациенты в возрасте от 18 до 65 лет с диагностированной НАЖБП, подписавшие информированное согласие.

Критерии исключения в исследуемую выборку являются:

1. Возраст старше 65 лет.

2. Вирусное поражение печени.

3. Алкогольная болезнь печени.

4. Наличие или подозрение на наркотическую зависимость пациента.

5. Аутоиммунные заболевания печени.

6. Болезни накопления (гемохроматоз, болезнь Вильсона-Коновалова).

7. Лекарственное поражение печени.

8. Цирроз печени.

9. Тяжелые сопутствующие заболевания (артериальная гипертензия III степени, некорригированная, СД 2 тип, декомпенсированный, хроническая сердечная недостаточность III-IV ФК, наличие в анамнезе инфарктов, инсультов).

10. Беременность, период лактации.

11. Онкологические заболевания.

12. Операции на органах ЖКТ (еюноилеальное шунтирование, резекция желудка, резекция тощей кишки, наложение билиарно-панкреатической стомы и др.).

Начальным этапом работы с данными, является этап первичной обработки данных, заключающийся в задачах: ознакомления с исследуемой выборкой, устранении дубликатов и аномалий, выявлении взаимосвязи и разбиение на подгруппы. На рисунке 2 продемонстрирована структура программного комплекса для первичной обработки данных анализов пациента. Она состоит из трех главных модулей: предобработки данных, анализа и визуализации данных. На стадии предобработки данных неподготовленные сведения группируются по категориальным критериям, проводится их очистка, результаты передаются на вход модуль анализа данных и определения оптимальных параметров. Далее производятся расчеты статистических параметров, включающие в себя статистическое описание данных физиологических параметров, лабораторных анализов и наличие сопутствующих заболеваний пациентов. Полученные данные подвергаются преобразованию и подаются на вход модуля визуализации.



Рисунок - 2 Структура программного комплекса для первичного анализа данных пациента.

Для обработки данных в исследовании будет использоваться язык программирования Python. Данный язык на основе статьи [13] в 19-м ежегодном опросе занял лидирующее место среди инструментов анализа данных и машинного обучения. Благодаря веб-приложению «Jupyter Notebook», исследовательские данные можно визуализировать с помощью подключения дополнительных библиотек.

# 3.1.1 Предобработка данных

Данный структурный модуль отвечает за выбор формата представления исследуемых текстовых данных, а так же устранение ошибок и дубликатов допущенных вследствие сбора сведений о пациенте. Алгоритм модуля представлен на рисунке 3.



Рисунок 3 – Алгоритм модуля предобработки данных и расчета статистических характеристик входной выборки пациентов

Алгоритм модуля предобработки данных и расчета описательной статистики входной выборки пациентов состоит из шести шагов перечисленных ниже:

1. Указывается физический путь к данным на ЭВМ для их непосредственной загрузки в программный комплекс первичной обработки данных. Учитывая тот факт, что исходный файл с хранимой информацией о пациенте создан в электронной таблице «Excel», то в качестве входного формата целесообразно выбрать «CSV» [14]. Формат, предназначенный для представления табличных данных, где каждая строка файла – это одна строка таблицы. Разделитель значений колонок является символ запятой.
2. Присвоение уникального идентификатора (ID) пациенту. Для этого фамилия и имя заменятся порядковым индексом, что требует соблюдения по договору о персональных данных больных.
3. Для повышения точности исследования производится обработка пропусков данных. Для этого выборка подвергается проверке на их целость преобразований типов, чтобы не отбросить данные из за их неправильной обработки программой. Так как данные являются медицинской направленности и должны удовлетворять высоким критериям достоверности, то такие методы восстановления данных как: заполнение нулями, значением медианы, средним арифметическим значением, является необоснованным и рискованным способом, который может сильно исказить результаты исследования. Поэтому данные, у которых невозможно восстановить пропуски подвергаются исключению из исследования.
4. Расчет статистических характеристик. Вычисление значений дисперсии, математического ожидания, размаха и квартилей.
5. Импортирование данных в «Jupyter» и группировка данных по их статистическим видам. В результате было получено, что данные по пациентам состоят из 4 групп: лабораторные исследования, сопутствующие заболевания, данные по пациенту, принимаемые препараты и представлены следующими типами статистических данных: количественные, биноминальные, качественно номинальные, качественно порядковые. Группировка параметров представлена в таблицах 1-4, где параметры разбиты по типам статистического параметра: бинарные, количественные, категориальные.
6. Табличное представление сгруппированных данных средствами «Jupyter».

Таблица 1 – Данные принадлежащие к группе сопутствующие заболевания

|  |  |  |  |
| --- | --- | --- | --- |
| Заболевания | Тип статистического параметра | Заболевания | Тип статистического параметра |
| (наличие заболевания неалкогольного стеатогепатита (стеатоз (1),  гепатит (2))) | Категориальный | (степень заболевания НАЖБП) | Категориальный |
| **(**сахарный диабет 2 типа) | Бинарный | **(**нарушенная толерантность к сахару) | Бинарный |
| **(**ишемическая болезнь сердца) | Бинарный | **(**артериальная гипертензия) | Бинарный |
| **(**ожирение) | Бинарный | болезнь бронхов | Бинарный |
| **(**остеоартроз) | Бинарный | **(**желчнокаменная болезнь) | Бинарный |
| **(**рубиновые пятна) | Бинарный | (болезнь Жильбера) | Бинарный |

Данные таблицы 1 представляют собой сопутствующие заболевания пациентов. В данную группу включены параметры, которые могут быть одной из причин появления заболевания, а также вносить свои значения в лабораторные анализы.

Таблица 2 – Данные принадлежащие к группе лабораторные исследования

|  |  |  |  |
| --- | --- | --- | --- |
| Лабораторное исследование | Тип статистического параметра | Лабораторное исследование | Тип статистического параметра |
| (содержание гемоглобин в крови) | Количественный | (содержание тромбоцитов в крови) | Количественный |
| **(**скорость оседания эритроцитов) | Количественный | **(**содержание глюкозы в крови) | Количественный |
| (тиреотропный гормон натощак) | Количественный | (тиреотропный гормон натощак) через 2 часа | Количественный |
| **(**аланинаминотрансфе-раз) | Количественный | (аспартатаминотронс-фераза) | Количественный |
| / **(**отношение аспартатаминотронсфе-  разы к аспартатамино  тронсферазе **)** | Количественный | **(**обхват бедер) | Количественный |
| **(**щелочная фосфатаз) | Количественный | (триглецириды) | Количественный |
| **(**гамма-глютамилтранспентидаза) | Количественный | уровень глюкозы в крови | Количественный |
| липопротеиды высокой плотности | Количественный | **(**липопротеиды низкой плотности) | Количественный |
| (количество в общего билирубин) мкмоль/л | Количественный | (содержание железа в организме) | Количественный |
| (содержание железа в крови) мкг/л | Количественный | (содержание церула в крови) | Количественный |
| **(**протомбиновый  индекс) | Количественный | **(**международное нормализованное соотношение) | Количественный |
| (содержание инсулин в организме) | Количественный | (содержание лептина в организме) | Количественный |
|  | Количественный | **(**количестворецептор лептина) | Количественный |
| - отношение содержания лептина в организме к количеству рецептор воспринимающих лептин | Количественный | **(**индекс инсулинорезистен-  тности) | Количественный |
| **(**матриксная металлопротеиназа 9) | Количественный | **(**тканевой ингибитор матриксных  протеиназ 1**)** | Количественный |
| **(**тканевой ингибитор матриксных протеиназ 2**)** | Количественный |  |  |

Данные таблицы 2 представляют собой параметры лабораторных исследований пациентов. В данную группу включены лабораторные исследования, являющиеся количественными мерами состояния организма.

Таблица 3 – Данные принадлежащие к группе принимаемые препараты

|  |  |  |  |
| --- | --- | --- | --- |
| Принимаемые препараты | Тип статистического параметра | Принимаемые препараты | Тип статистического параметра |
| **(**метформин) | Бинарный | **(**галвус) | Бинарный |
| глибенкламид) | Бинарный | (инсулин) | Бинарный |
| **(**эналаприл) | Бинарный | **(**периндоприл) | Бинарный |
| **(**лизиноприл) | Бинарный | (лозартан) | Бинарный |
| **(**диуретики) | Бинарный | **(**бета-адреноблокаторы) | Бинарный |
| **(**амлодипин) | Бинарный | **(**ксантелазмы) | Бинарный |

Данные таблицы 3 представляют собой список принимаемых препаратов. В данную группу включены препараты, непосредственно воздействующие на состояние печени.

Таблица 4 – Данные принадлежащие к группе параметры пациента

|  |  |  |  |
| --- | --- | --- | --- |
| Параметр пациента | Тип статистического параметра | Параметр пациента | Тип статистического параметра |
| (пол пациента) | Категориальный | (боль в правом подлеберье) | Категориальный |
| (дискомфорт) | Категориальный | (cтепень отрыжки) | Категориальный |
|  | Категориальный | (общая слабость) | Категориальный |
| (степень утомляемости) | Категориальный | **(**возраст пациента) | Количественный |
| наличие боли в эпигастрии) | Бинарный | (бета-адреноблокаторы) | Бинарный |
| (увеличение печени) | Бинарный | (гепатомегалия УЗИ) | Бинарный |
| (xолецитсэктомия) | Бинарный | (эхогенность образования) | Бинарный |
| (наличие конрементов) | Бинарный | (эхогенность поджелудочной железы) | Бинарный |
| (наличие неодинарность и зернистость ПЖЖ) | Бинарный | (кальцинаты поджелудочной железы) | Бинарный |
|  | Количественный | (давность заболевания НАЖБП) | Количественный |
| (систолическое артериальное давление) | Количественный | (диастолическое давление) | Количественный |
| (диастолическое давление) | Количественный | (частота сердечных сокращений) | Количественный |
| рост пациента) | Количественный | **(**вес пациента) | Количественный |
| (индекс массы тела) | Количественный | (обхват талии  ) | Количественный |
| (обхват бедер) | Количественный | (отношение обхвата бедер к обхватуталии) | Количественный |
| (длина желудочного пузыря) | Количественный | (ширина желудочного пузыря) | Количественный |
| (толщина стенки желудочного пузыря) | Количественный | (площадь селезенки пациента) | Количественный |
| (время динамики) | Количественный |  |  |

# 3.1.2 Описательная статистика

**Описательная статистика** — один из разделов статистической науки, в рамках которого изучаются методы описания и представления основных свойств данных. Позволяет обобщать первичные результаты, полученные при наблюдении или в эксперименте [18]. Применение описательной статистики включает следующие этапы:

1. Сбор данных
2. Категоризация данных
3. Обобщение данных
4. Представление данных

В таблицах 5-10 рассчитаны следующие описательные статистические показатели:

* минимум – минимальное значение исследуемого параметра;
* математическое ожидание – среднее значение исследуемого параметра

|  |  |
| --- | --- |
| , | (1) |

где *j* – индекс параметра;

*i* – индекс элемента параметра;

– элемент *j* параметра;

– количество элементов *j* параметра.

* дисперсия – характеризует меру изменчивости исследуемой величины.

|  |  |
| --- | --- |
|  | (2) |

где *j* – индекс параметра;

– математическое ожидание.

* среднеквадратическое отклонение – характеризует величину отклонений значений от среднего.

|  |  |
| --- | --- |
|  | (3) |

где – дисперсия j параметра.

* квартиль – характеризует часть распределения вероятностей случайной величины.

|  |  |
| --- | --- |
|  | (4) |

где *j* – индекс параметра;

*;*

– количество значений выборки;

*i* – номер квартиля.

Таблица 5 – Статистические характеристики физиологические параметров пациентов

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | муж(1)  жен(2) | , мм | мм | , мм | , см |  |  | *см* | |
| Количе-ство элементов | 149 | 149 | 110 | 81 | 110 | 138 | 138 | 115 | | 113 |
| Среднее | 48,49 | 1,23 | 75,18 | 2,15 | 28,19 | 173,4 | 98,62 | 35,86 | | 108 |
|  | 10,34 | 0,43 | 11,43 | 0,42 | 6,38 | 8,5 | 14,55 | 9,97 | | 9,9 |
|  | 23 | 1 | 46 | 1 | 18 | 152 | 64 | 17 | | 87 |
|  | 41 | 1 | 66 | 2 | 23,25 | 168,3 | 89 | 29 | | 101 |
|  | 57 | 1 | 83 | 2 | 31.75 | 179 | 106 | 40 | | 115 |
|  | 73 | 2 | 111 | 3 | 55 | 190 | 147 | 74 | | 140 |
| Медиана | 46 | 1.24 | 75.5 | 2 | 28 | 174 | 98 | 35 | | 107 |
| – обхват талии пациента, – рост пациента, – вес пациента, , – площадь селезенки пациента, – ширина желчного пузыря, , – длина стенки желчного пузыря, – толщина стенки желчного пузыря | | | | | | | | | | |

Таблица 6 – Статистические характеристики физиологические параметров пациентов

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| Количе-ство элементов | 65 | 149 | 149 | 149 | 149 | 72 | 149 |
| Среднее | 20,71 | 0,34 | 0,34 | 0,44 | 0,5 | 0,11 | 82,51 |
|  | 33,34 | 0,75 | 0,77 | 0,77 | 0,8 | 0,31 | 6,81 |
|  | 0 | 0 | 0 | 0 | 0 | 0 | 70 |
|  | 1 | 0 | 0 | 0 | 0 | 0 | 80 |
|  | 24 | 0 | 3 | 1 | 1 | 0 | 90 |
|  | 144 | 3 | 0 | 3 | 3 | 1 | 100 |
| Медиана | 5 | 0 | 0 | 0 | 0 | 0 | 89 |
| T – давность заболевания НАЖБП, – дисдолическое давление, – общая слабость, – cтепень отрыжки, – , – xолецитсэктомия, – диастолическое давление | | | | | | | |

Таблица 7 – Статистические характеристики сопутствующих заболеваний пациентов

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  | |
| Количе-ство элементов | 149 | 149 | 149 | 149 | 149 | 149 | 149 | 149 | |
| Среднее | 1,52 | 0,13 | 0,23 | 0,64 | 0,14 | 0,97 | 0,05 | 0,13 | |
|  | 0,5 | 0,34 | 0,42 | 0,48 | 0,35 | 0,16 | 0,23 | 0,34 | |
|  | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
|  | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | |
|  | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | |
|  | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| Медиана | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | |
| – наличие заболевания артериальная гипертензия у пациента,  – наличие заболевания ишемическая болезнь сердца у пациента,  – наличие заболевания нарушенной толерантности к глюкозе, болезнь бронхов, – наличие остеоартроза у пациента, – наличие заболевание сахарного диабета 2 типа у пациента,  – наличие заболевания неалкогольного стеатогепатита (стеатоз (1), гепатит (2)). | | | | | | | | |

Таблица 8 - Статистические характеристики лабораторных анализов пациентов.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | , нг/мл | , нг/мл | , нг/мл | , мкг/мл | нг/мл | мкг/мл | нг/мл | % | ед/л |
| Количе-ство | 87 | 87 | 87 | 87 | 35 | 111 | 108 | 72 | 63 |
| Среднее | 1464 | 127.36 | 391.18 | 6.76 | 9.42 | 18.89 | 21.55 | 97.49 | 81.63 |
|  | 579,71 | 45,18 | 219,92 | 7,37 | 10,35 | 13,06 | 18,45 | 11,94 | 165,8 |
|  | 570 | 70,5 | 61 | 0,12 | 2,46 | 0,07 | 1,35 | 18 | 5 |
|  | 1105 | 93,5 | 250,5 | 1,44 | 4,52 | 7,04 | 9,69 | 93,75 | 31,9 |
|  | 1582 | 153.75 | 486 | 10,15 | 9,92 | 27,45 | 26,43 | 104 | 75 |
|  | 4105 | 286 | 1636 | 43,64 | 64,32 | 61,2 | 108,8 | 116 | 200 |
| Медиа-на | 1345 | 113 | 342 | 4,54 | 7,03 | 3,27 | 16,31 | 99 | 53,3 |
| – тканевой ингибитор матриксных протеиназ 1, – тканевой ингибитор матриксных протеиназ 2, – матриксная металлопротеиназа 9, (Homeostasis Model, Assessment of Insulin Resistance) – индекс инсулинорезистентости, – протромбированный индекс,  **-** гамма-глютамилтрансептидаза. – адипонектин, – содержание рецепторов воспринимающих лептин в крови . – содержание лептина , | | | | | | | | | |

Таблица 9 - Статистические характеристики лабораторных анализов пациентов.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **,** мкМЕ/мл | ,  ммоль/л | ,  ед/л | , ед/л | ,  ед/л | ,  г/л | ммоль/л | ,  ммоль/л |
| Количе-ство | 19 | 19 | 137 | 52 | 137 | 137 | 129 | 143 | 23 |
| Среднее | 6,52 | 7,41 | 5,72 | 156,22 | 26,33 | 17,53 | 76 | 5,8 | 23,89 |
|  | 1,59 | 2,49 | 1,2 | 158,66 | 34,61 | 20,9 | 5,99 | 1,3 | 12,71 |
|  | 4 | 3,9 | 3,7 | 45,0 | 0,13 | 0,1 | 61 | 3,64 | 9,6 |
|  | 5,45 | 5,45 | 4,9 | 75,5 | 0,51 | 0,3 | 72 | 4,78 | 16,9 |
|  | 7,5 | 8,65 | 6,3 | 198 | 41 | 28 | 80 | 6,5 | 26,7 |
|  | 11 | 14.3 | 10.5 | 1115 | 181 | 110 | 91 | 10,61 | 61 |
| Медиа-на | 6,3 | 7 | 5,4 | 104 | 15 | 16 | 76 | 5,61 | 20,5 |
| – содержание глюкозы в крови,  **-** териатропный гормон натощак,  **-** териатропный гормон после 2 часов приема пищи**, –** щелочная фосфата,  **–** аланинаминотрансфераза,  **–** аспартатаминотрансфераза,  **-** уровень глюкозы в крови, – содержание железа в крови | | | | | | | | | |

Таблица 10 - Статистические характеристики лабораторных анализов пациентов.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ед/л | ммоль/л | ммоль/л | МкЕд/мл |  |
| Количе-ство | 71 | 54 | 52 | 92 | 35 |
| Среднее | 1,05 | 1,21 | 3,72 | 25,37 | 5,33 |
|  | 0,29 | 0,6 | 1,26 | 27,53 | 6,86 |
|  | 0,78 | 0,45 | 1,5 | 0,55 | 0,08 |
|  | 0,97 | 0,95 | 2,92 | 6,24 | 1,44 |
|  | 1,07 | 1,29 | 4,58 | 37,51 | 5,46 |
|  | 3,38 | 5,06 | 7 | 163,66 | 28,24 |
| Медиа-на | 1,01 | 1,09 | 3,79 | 16,3 | 15 |
| международное нормализованное соотношение, липопротеиды высокой плотности, липопротеиды низкой плотности, – инсулин, – отношение лептина к obr, | | | | | |

# 3.1.3 Построение диаграммы размаха

Построение диаграммы размаха используется для визуализации областей наиболее достоверных значений, выделенных на основании анализа диаграмм распределения. Строится данная диаграмма на основании формулы:

|  |  |
| --- | --- |
|  | (5) |

где

первый квартиль;

третий квартиль;

*k* – коэффициент, наиболее часто употребляемое значение, которого равно 1,5. [15].

Анализ гистограммы позволяет определить: область наиболее достоверных значений и выбросов, отличных от всей совокупности выборки.

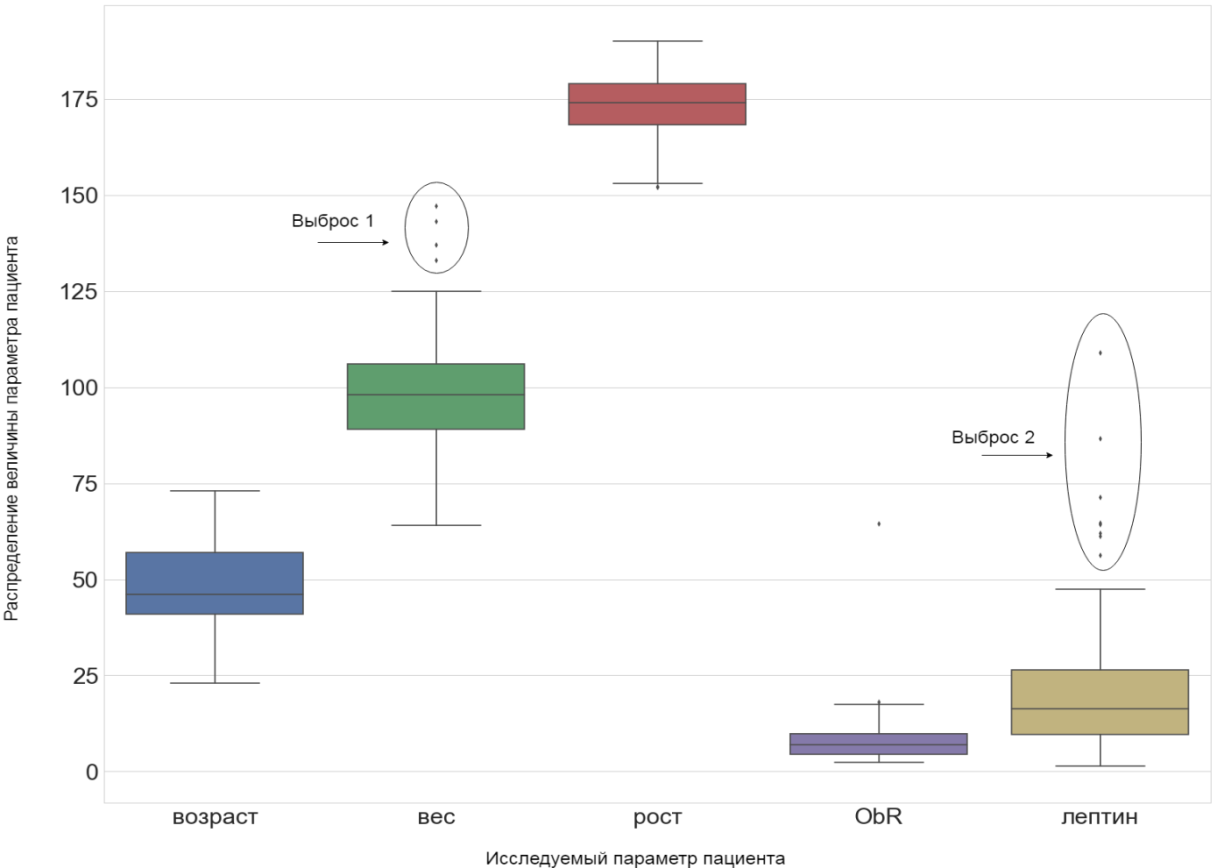


Рисунок 4 – Диаграмма размаха

На рисунке 4 изображена диаграмма размаха для следующих параметров: , , , , . Оцениваемые параметры возраст, , не имеют четко выраженных выбросов. Однако, имеются области выброса: у и веса .

Учитывая специфику заболевания проявляющуюся у людей с ожирением, результаты «выброса 1» не стоит исключать из общей выборки, а лишь необходимо принять во внимание и при постановке диагноза учитывать принадежность пациента к выбросу. Второй выброс объясняется биологическим фактором, разным средним количеством лептина у мужчин и женщин, рисунок 5.

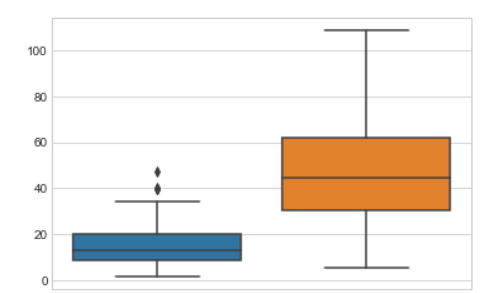


Рисунок 5 – Распределение параметра . Слева – муж, справа – жен.

Так же прослеживается связь принадлежности некоторых пациентов к двум выборкам сразу, что позволяет сделать предположение о том, что существует зависимость между параметрами лабораторных значений и физиологичекого показателя вес.

# 3.2 Модуль «Выявление значащих параметров для экспертной системы»

На вход модуля «Выявления значащих параметров для экспертной системы» передаются данные полученные из модуля «Предобработки данных».

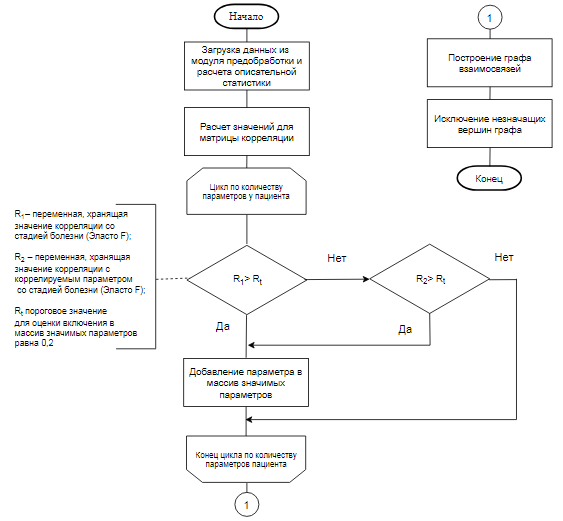


Рисунок 6 – Алгоритм модуля «выявления значащих параметров для экспертной системы»

Алгоритм модуля «выявления значащих параметров для экспертной системы» на первом шаге производит загрузку данных, полученных на выходе из модуля «предобработки данных и расчета описательной статистики».

На втором шаге производятся вычисление корреляции по принципу каждый параметр с каждым» по формуле (6).

|  |  |
| --- | --- |
| , | (6) |

где– среднее значение параметров пациента.

Корреляционная матрица представляет собой симметричную квадратную матрицу размером *M\*M*, где *M* – число параметров пациентов, главная диагональ, которой заполнена единицами, а недиагональные элементы представляют собой коэффициент корреляции. Для лучшей визуализации накладывается цветовой градиент соответствующий цветовому спектру от тёмно-синего до тёмно-красного, где красный цвет отвечает за положительное значение, а синий отрицательное значение корреляции.

На основании полученных результатов, параметры, значения которых удовлетворяет условиям (7) относятся к значимым для исследования и добавляются в граф взаимосвязей параметров с параметром .

(категориальный параметр, отвечающий за стадию болезни)

|  |  |
| --- | --- |
| || && , | (7) |

где – переменная, хранящая значение корреляции с ;

– переменная, хранящая значение корреляции с коррелируемым параметром с

В результате данной фильтрации будет получен рисунок 8 с наглядной визуализацией коррелированности.

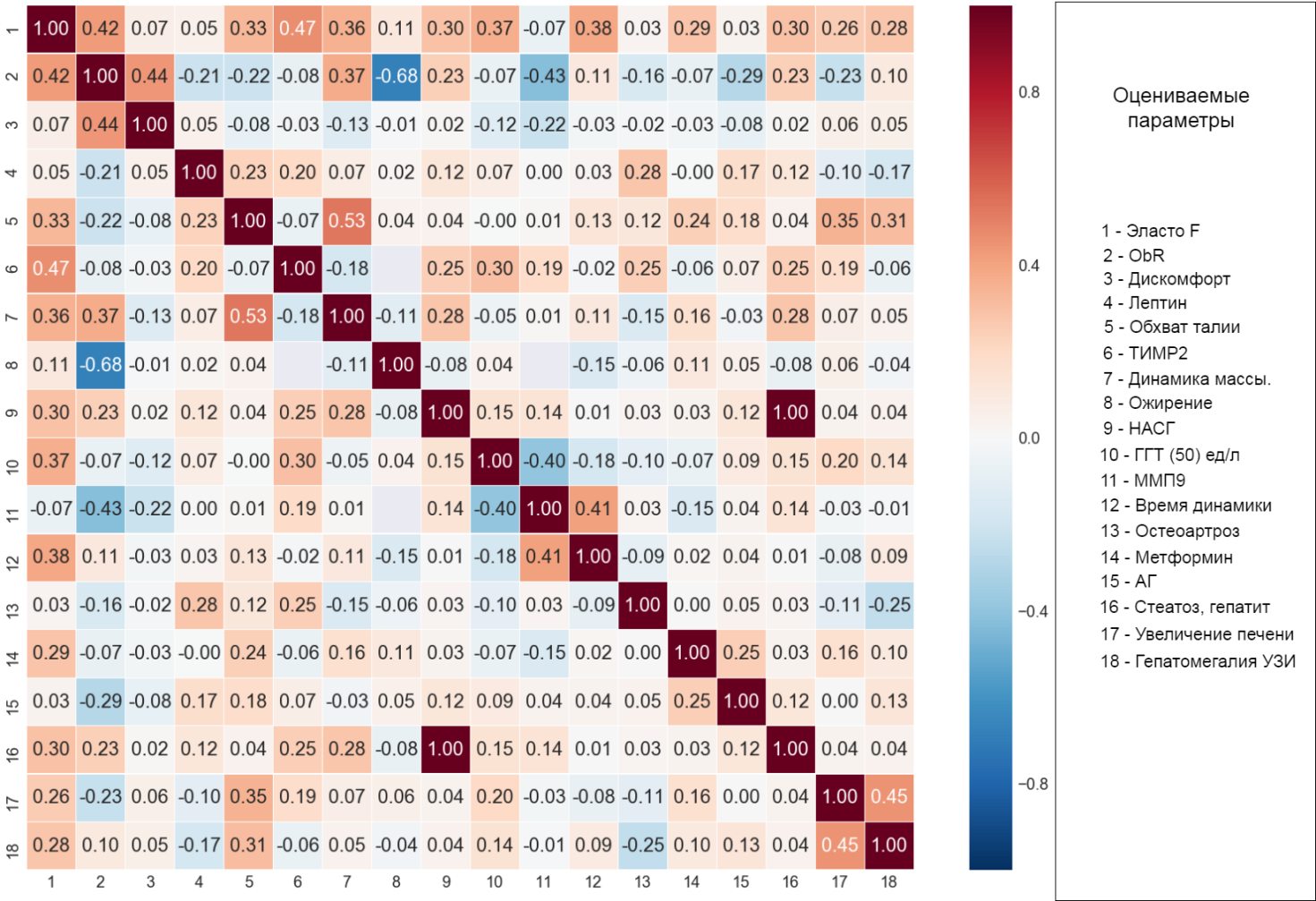


Рисунок 7 – Карта взаимосвязи результатов исследования

Анализ рисунка 7 позволяет выделить следующие взаимосвязи: имеет слабую корреляцию с параметрами , , , , , , , , .

# 3.3 Проверка физических и лабораторных параметров на соответствие нормальному распределению

Тестирование данных на нормальность часто является первым этапом их анализа, так как большое количество статистических методов исходит из предположения нормальности распределения изучения данных.

Для исследования физических и лабораторных параметров задается значение значимости

Эксцесс – мера остроты пика распределения случайной величины. Характеризует остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение (пик заострен). Отрицательный эксцесс обозначает относительное сглаженное распределение (пик закруглен). Таким образом, можно установить отклонения заданного закона от нормального по формуле (8).

|  |  |
| --- | --- |
| , | (8) |

где –среднее математическое значение;

– стандартное отклонение.

Необходимым условием нормальности распределения является отклонение значение эксцесса не более чем на 1 [25].

Достоверность принимаемой гипотезы устанавливается в результате использования критерия согласия Пирсона. Критерий хи-квадрат Пирсона – это непараметрический метод, который позволяет оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Применяется для проверки гипотезы о принадлежности эмпирической выборки объемом некоторому теоретическому закону распределения и подтверждении или отвержении нулевой гипотезы.

Так как статистика Пирсона измеряет разницу между эмпирическим и теоретическим распределениями, то, чем больше ее наблюдаемое значение , тем сильнее довод против основной гипотезы. Поэтому критическая область для этой статистики всегда правосторонняя: [Kkp;+∞). Используя таблицу распределения критерия Пирсона [29], составляется таблица 11. Число степеней свободы рассчитывается по формуле (9).

|  |  |
| --- | --- |
| , | (9) |

где – количество интервалов разбиения исследуемого параметра;

– параметры распределения.

Число интервалов равно 5, 2 параметра нормального распределения (математическое ожидание и среднеквадратическое отклонение) и минус 1. Следовательно, значение *k* = 2. Зная значение уровня значимости α по таблице 11, устанавливается значение критерия для значимости 0.05 .

Таблица 11 – Распределение критерия Пирсона при количестве степеней свободы *k* = 2

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *k*/*α* | 0,99 | 0,98 | 0,95 | 0,90 | 0,80 | 0,20 | 0,10 | 0,05 | 0,02 | 0,01 | 0,001 |
| 2 | 0,0201 | 0,0404 | 0,103 | 0,211 | 0,446 | 3,219 | 4,605 | 5,991 | 7,824 | 9,210 | 13,815 |

Для автоматизации расчетов используется библиотека для языка программирования Python с открытым кодом, предназначенная для выполнения научных и инженерных расчетов [26]. Код программы в приложение В.

Используя функции «scipy.stats.kurtosis()»[27] и «scipy.stats.normaltest()» [28] вычислены значения хи-квадрат Пирсона и эксцесса. Фрагмент результатов вычислений представлен на рисунке 8.

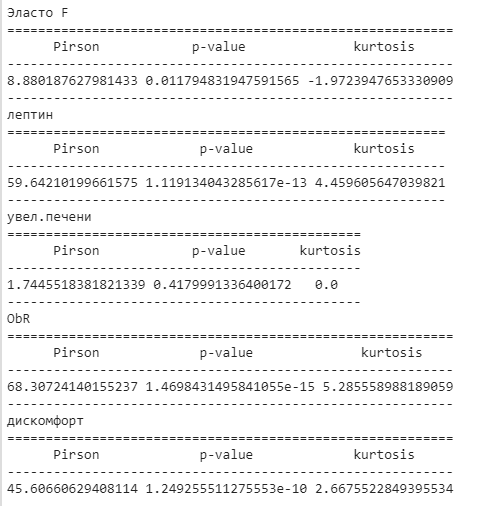


Рисунок 8 – Фрагмент вывода значений критерия Пирсона и p-value вычисленного с помощью библиотеки статистики «SciPy»

По полученным результатам построена таблица 12. Преобладающее большинство параметров имеет значение критерия Пирсона больше . Следующие параметры: , , имеют значения Пирсона больше чем Ограничению значения расхождение эксцесса с нормальным распределением удовлетворяют параметры: , , , , , .Приняв результаты значений Пирсона, эксцесса и p-value можно выделить следующие параметры, подходящие под нормальное распределение: и .

Таблица 12 –Значения критерия Пирсона , p-value, эксцесса исследуемых физических и лабораторных параметров

|  |  |  |  |
| --- | --- | --- | --- |
| Оцениваемый физический или лабораторный параметр | Значение критерия Пирсона | Значение p-value | Значение расхождение эксцесса исследуемого параметра со значением эксцесса у нормального распределения |
|  | 8,88 | 0,012 | -1,972 |
|  | 59,642 | 0 | 4,46 |
|  | 4,504 | 0,105 | 1,06 |
|  | 68,307 | 0 | 5,29 |
|  | 22,321 | 0 | 2,256 |
|  | 68,62 | 0 | 5,322 |
| ( | 22,91 | 0 | 3,057 |
|  | 76,31 | 0 | -7,752 |
|  | 8,609 | 0,014 | 0 |
|  | 129,389 | 0 | 6,815 |
|  | 0,122 | 0,941 | 0 |
|  | 45,6 | 0 | 2,668 |
|  | 8,73 | 0,013 | 0 |
|  | 205,962 | 0 | 8,071 |
|  | 68,89 | 0 | 3,689 |
|  | 1,744 | 0,418 | 0 |
|  | 14,44 | 0 | 2,216 |

# 3.4 Регрессионный анализ

Сталкиваясь с задачей диагностики и прогнозирования, часто требуется установить взаимосвязи между несколькими параметрами. Наиболее простым и эффективным способом выявления функциональной зависимости.

Регрессионный анализ – метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной) [19].

Задачи регрессионного анализа:

* определение степени детерминированности вариации критериальной переменой предикторами (независимыми переменными);
* предсказание значения зависимой переменной с помощью независимой(-ых);
* определение вклада отдельных независимых переменных в вариацию зависимой.

На основе результатов полученных при построении корреляционной матрицы, выбрав признаки с наибольшей корреляции, и предполагаемой связи c строятся графики линейной регрессии. Для этого используются функция Python paitplot()[70], которая формируют график, по главной диагонали которого изображается распределение данных, а по побочной линейная регрессия.

Построение прямой осуществляется подбором коэффициентов свободного члена и углового коэффициента прямой. Для этого используется метод наименьших квадратов[71] формула (10), чтобы минимизировать сумму квадратов остатков.

|  |  |
| --- | --- |
| , | (10) |

где – значение наблюдаемого параметра;

– аппроксимирующее значение наблюдаемого параметра.

Результаты приведены на рисунках 9 – 10.

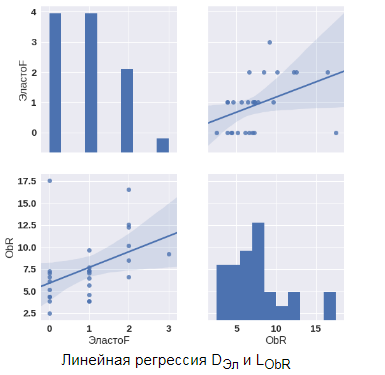
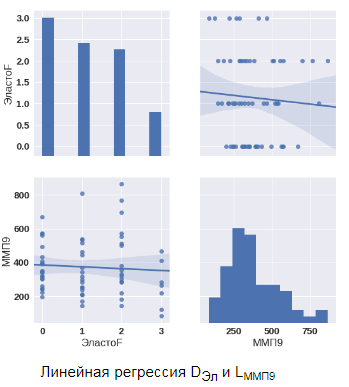


Рисунок 9 – Графики линейной регрессии , *,*

На рисунке 9 линейные регрессии и недостаточно описывают взаимосвязь данных. Многие данные остаются за пределом прямой, что позволяет сделать вывод о нелинейной зависимости.

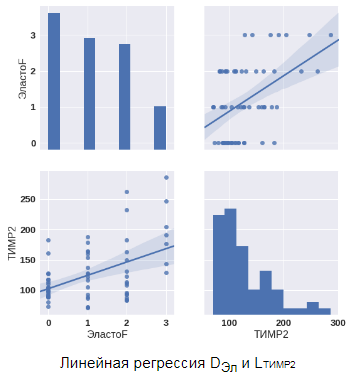
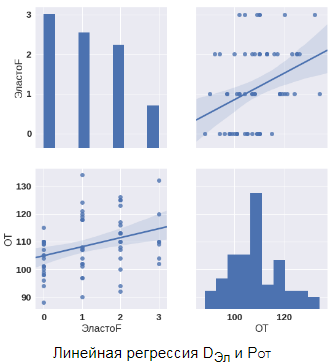
 

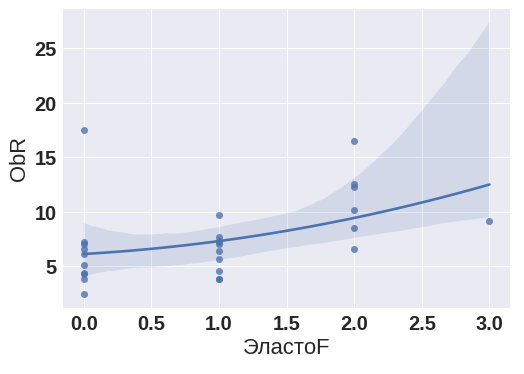
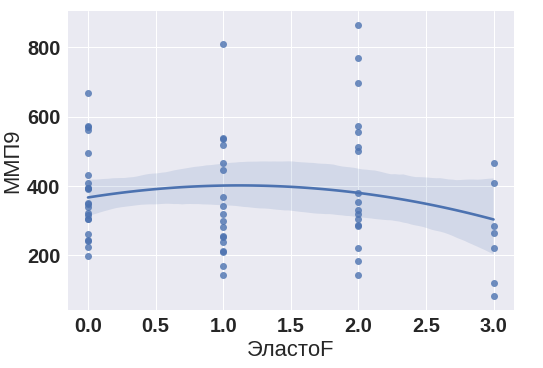
Рисунок 10 – Графики линейной регрессии , ,

На рисунке 10 линейные регрессии и . Данные расположены ближе к аппроксимирующей прямой, но основная часть все же лежит за пределами аппроксимации, что характеризует об отсутствие линейной зависимости.

Таблица 13 – Значение p-value для линейной регрессии

|  |  |
| --- | --- |
| Наименование параметров | Значение p-value |
|  | 0,637 |
|  | 0,637 |
|  | 0,602 |
|  | 0,063 |

Исходя из графиков 9,10 и таблицы 13, основные коррелирующие признаки не имеют линейной зависимости с . Поэтому параметры необходимо рассматривать группами, что соответствует раннее предположенной гипотезе о совокупности влияющих параметров.



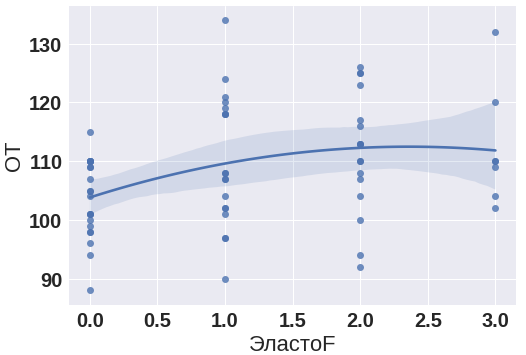
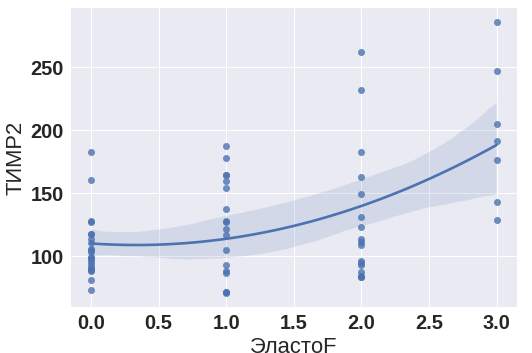
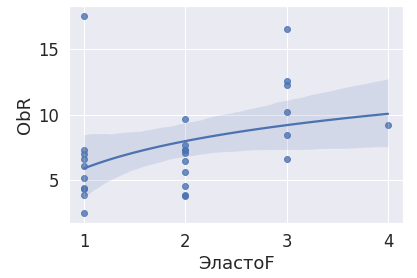
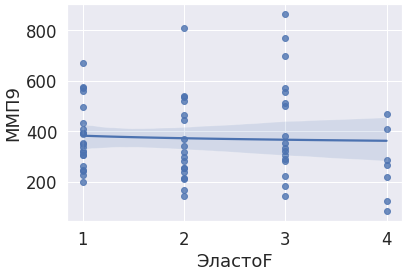


Рисунок 11 – Графики квадратичной регрессии , , ,

Таблица 14 – Значение p-value для квадратичной регрессии

|  |  |
| --- | --- |
| Наименование параметров | Значение p-value |
|  | 0,08 |
|  | 0,07 |
|  | 0,02 |
|  | 0,315 |

Исходя из графиков 11 и таблицы 14, один из основных коррелирующих признаков имеют квадратичную зависимости с .



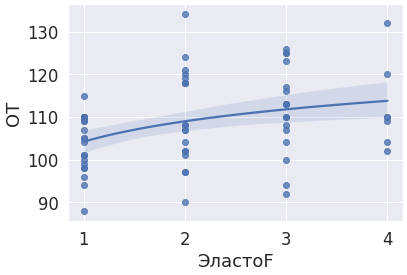
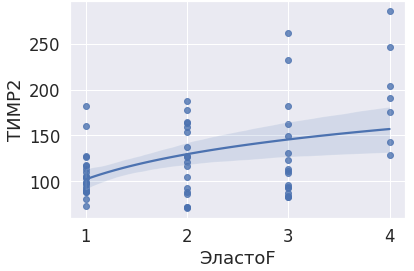


Рисунок 12 – Графики логарифмической регрессии , , ,

Таблица 15 – Значение p-value для логарифмической регрессии

|  |  |
| --- | --- |
| Наименование параметров | Значение p-value |
|  | 0,06 |
|  | 0,205 |
|  | 0,03 |
|  | 0,013 |

Исходя из графиков 12 и таблицы 15, два из основных коррелирующих параметров иимеют логарифмическую зависимость.

# 3.5 Обобщение результатов и построение графа зависимости

По результатам, полученным на рисунке 9, строится граф зависимостей. В центре графика располагается параметр . Относительно его откладываются переменные, которые в результате построения корреляционной матрицы дали значения выше 0,2 по модулю и формируют собой первый уровень зависимости. Далее предполагая, что существует множество параметров, которые зависимые между собой, находятся такие, что имеют зависимость с первым уровнем и значение корреляции больше 0,2. Результаты представлены на рисунке 12.

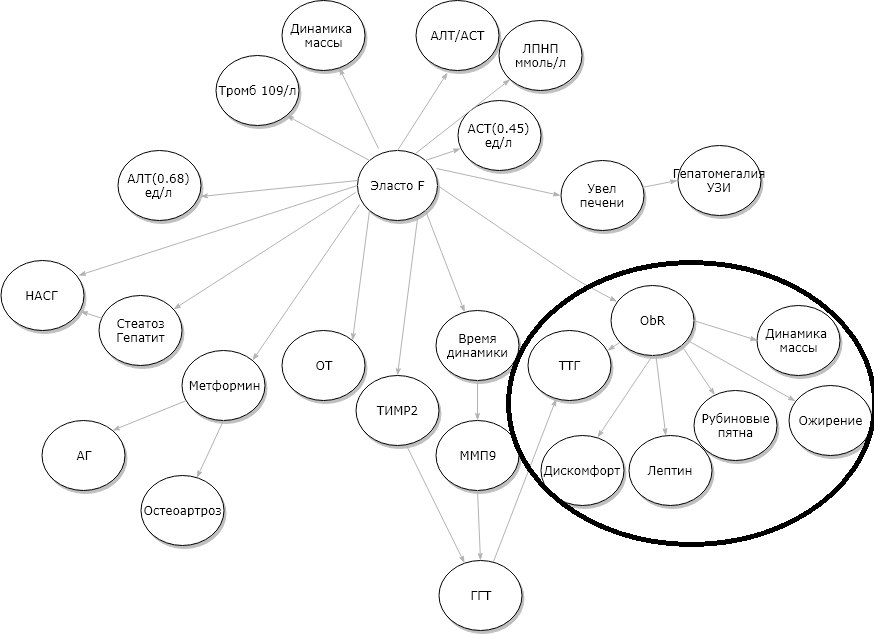


Рисунок 12 – Граф взаимосвязей параметров

Исходя из рисунка 12 можно выделить 2 подгруппы с наличием четко прослеживающийся зависимости это: и .

# 3.6 Экспертная система

Экспертная система – компьютерная система, способная частично заменить специалиста-эксперта в разрешении проблемной ситуации [19]. При ранней диагностике используется система на основе нечеткой логике.

Причиной данного выбора является наработанный ранее материал экспертами и простота построение нечетких систем по сравнению системами, базирующимися на нейронных сетях.

- Повышенная доступность. Для обеспечения доступа к экспертным знаниям могут применяться любые подходящие компьютерные аппаратные средства. В определенном смысле вполне оправдано утверждение, что экспертная система – это средство массового производства экспертных знаний.

- Уменьшенные издержки. Стоимость предоставления экспертных знаний в расчете на отдельного пользователя существенно снижается.

- Уменьшенная опасность. Экспертные системы могут использоваться в таких вариантах среды, которые могут оказаться опасными для человека. Постоянство. Экспертные знания никуда не исчезают. В отличие от экспертов-людей, которые могут уйти на пенсию, уволиться с работы или умереть, знания экспертной системы сохраняются в течение неопределенно долгого времени

- Возможность получения экспертных знаний из многих источников. С помощью экспертных систем могут быть собраны знания многих экспертов и привлечены к работе над задачей, выполняемой одновременно и непрерывно, в любое время дня и ночи. Уровень экспертных знаний, скомбинированных путем объединения знаний нескольких экспертов, может превышать уровень знаний отдельно взятого эксперта.

- Повышенная надежность. Применение экспертных систем позволяет повысить степень доверия к тому, что принято правильное решение, путем предоставления еще одного обоснованного мнения эксперту-человеку или посреднику при разрешении несогласованных мнений между несколькими экспертами-людьми. (Разумеется, такой метод разрешения несогласованных мнений не может использоваться, если экспертная система запрограммирована одним из экспертов, участвующих в столкновении мнений.) Решение экспертной системы должно всегда совпадать с решением эксперта; несовпадение может быть вызвано только ошибкой, допущенной экспертом, что может произойти, только если эксперт устал или находится в состоянии стресса.

- Объяснение. Экспертная система способна подробно объяснить свои рассуждения, которые привели к определенному заключению. А человек может оказаться слишком усталым, не склонным к объяснениям или неспособным делать это постоянно. Возможность получить объяснение способствует повышению доверия к тому, что было принято правильное решение.

- Быстрый отклик. Для некоторых приложений может потребоваться быстрый отклик или отклик в реальном времени. В некоторых экстремальных ситуациях может потребоваться более быстрая реакция, чем у человека; в таком случае приемлемым вариантом становится применение экспертной системы, действующей в реальном времени.

- Неизменно правильный, лишенный эмоций и полный ответ при любых обстоятельствах. Такое свойство может оказаться очень важным в реальном времени и в экстремальных ситуациях, когда эксперт-человек может оказаться неспособным действовать с максимальной эффективностью из-за воздействия стресса или усталости.

- Возможность применения в качестве интеллектуальной обучающей программы. Экспертная система может действовать в качестве интеллектуальной обучающей программы, передавая учащемуся на выполнение примеры программ и объясняя, на чем основаны рассуждения системы.

- Возможность применения в качестве интеллектуальной базы данных. Экспертные системы могут использоваться для доступа к базам данных с помощью интеллектуального способа доступа. В качестве примера можно привести анализ скрытых закономерностей в данных [19].

# 3.6.1 Проектирование экспертной системы ранней диагностики заболевания с использованием нечеткого классификатора

Экспертные системы обеспечивают поддержку принятия решений в ситуациях, для которых алгоритм заранее не известен и формируется в виде цепочки рассуждений из базы правил. Основой проектирования системы является определение стадия заболевания при ранней диагностики болезни. Применение аппарата нечеткой логики позволяет работать с существующей неопределенностью, неполнотой и нечеткостью информации. Выбор решения по числовой информации в условиях неопределенности возникает в том случае, как с каждым принимаемым решением связано некоторое множество возможных результатов с известными условными вероятностями.

Поэтому использование аппарата нечеткой логики при разработке базы знаний экспертной системы позволит формализовать процедуру оценки стадии болезни печени на базе формализованного экспертного опыта.

Таким образом, основной целью использования нечеткой логики является создание аппарата способного моделировать рассуждения эксперта по данным полученным в результате первичного анализа данных и экспертного опыта в области гастрологии [22 – 26].

В разработке экспертных систем участвуют представители разных специальностей. Наиболее часто в разработке участвуют три специалиста. Это, как правило:

* эксперт прикладной области
* инженер по знанию
* программист по разработке инструментальных средств[68,74]

Формирование знаний для экспертной системы является одним из узких мест, которые оборачиваются в проблему «новичок и эксперт» для инженера знаний и эксперта. Инженер знаний, выступая в роли посредника, помогает специалисту строить систему, но так как инженер имеет гораздо меньше знаний о предметной области, чем эксперт, то возникают проблемы взаимодействия, что вызывает сложности переноса компетентности в программу. Словарь, используемый первоначально экспертом в разговоре с новичком о своей предметной области, часто оказывается недееспобным для решения задач; поэтому инженер и эксперт должны проработать совместно, чтобы расширить и уточнить его, рисунок 13.

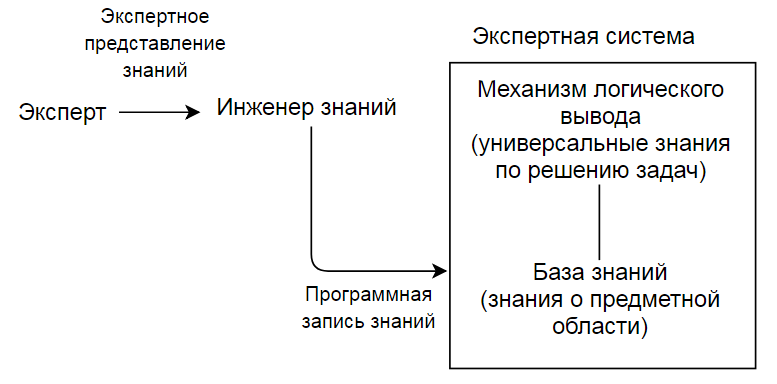


Рисунок 13 – Процесс инженерий знаний (передача знаний от эксперта в базу знаний)

Перед непосредственным проектированием экспертной системы, чтобы ограничить систему, сделаны основные предположения:

* система выполняет только постановку диагноза и не предполагает дальнейших инструкций по уточнению диагноза или назначения лечения;
* вводимые данные оператором(врачом) имеют высокую достоверность и не имеют резко выраженных шумов;
* входные данные не имеют пропусков и человеческих ошибок;
* система не предоставляет возможности добавления и изменения правил;
* не предполагается мутационная изменчивость;
* не учитывается влияние других болезней, которым подвержен пациент
* значение у женщин нормализовано к значению у мужчин.

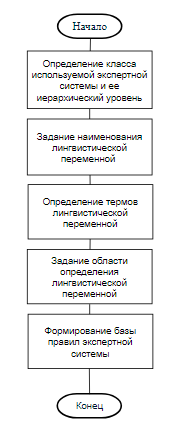


Рисунок 14 – Алгоритм модуля «проектирования экспертной системы ранней диагностики заболевания»

Алгоритм модуля «проектирования экспертной системы ранней диагностики заболевания» состоит из 5 шагов перечисленных ниже.

1. Определение класса используемой экспертной системы и ее иерархический уровень.

2. Формирование наименования лингвистических переменных. В результате работы модуля «первичный анализ данных» найдены два лабораторных параметра, которые имеют корреляцию с и образуют возле себя зависимые группы коррелирующих параметров со стадией болезни печени. Поэтому на вход системы нечеткого вывода будут поданы значения трех переменных: значение , и   
, которые и будут являться наименованиями лингвистических переменных. Использование , как дополнительной входной переменной, обусловлено работами рассматривающие совокупную составляющую болезни печени со стадией [21].

3. Задаются термы лингвистической переменной. Исходя из предметной области, термы {«малое», «среднее», «большое»}; термы {« малое», «среднее», «большое»};

4. Определяется область определения лингвистических переменных.

5. Создание и описание базы правил экспертной системы.

Перед тем как приступить к разработке экспертной системы, инженер по знания должен выполнить оценку области применение и ее необходимости, основываясь на следующих требованиях:

* эксперты сходятся в оценке предлагаемого решения, иначе нельзя будет оценить качество разработанной экспертной системы;
* эксперты могут выразить на естественном языке и объяснить используемые ими предположения и методы.
* должны быть выделены основные отношения, понятия и способы получения решения задачи[68].

Применение экспертной системы оправдано в том случае, если решение задачи принесет значительный эффекта, например, экономический, замена человека эксперта в связи его территориальной недоступности, невозможности выполнение экспертизы одновременно в различных местах.

# 3.6.2 Создание базы правил экспертной системы

Целью базовых правил нечеткой логики является формализация и применение человеческого умозаключения. Таким образом, нечеткая логика является частью искусственного интеллекта. Базы правил нечеткой логики являются наиболее часто используемым инструментов в приложениях с нечеткой логикой. Базы правил нечеткой логики представляют собой набор правил, которые обычно используются параллельно, но в некоторых приложениях могут быть объединены.

Требования, обеспечивающие построение базы нечетких продукционных правил для экспертной системы ранней диагностик заболевания печени:

* данные и знания должны быть надежными;
* данные и знания должны быть неизменными;
* используемые правила должны быть конкретизирована, а сама база оптимизирована (отсутствие дубликатов, неиспользуемых правил).

Для задания области определения лексической переменной необходимо определить принадлежность стадии болезни к границам интервала. По этой причине строится диаграмма распределения в зависимости от стадии болезни рисунок 15.

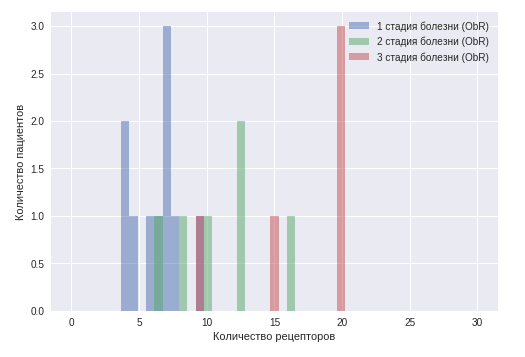


Рисунок 15 – график распределения количества в зависимости от стадии болезни

Исходя из графика 15, стадии имеют четко выраженное распределение по интервалам. Первой стадии болезни печени соответствуют значения где [3; 8], второй стадии соответствуют значения [6; 16], третьей стадии соответствуют значения [10; 20].

Для задания области определения лексической переменной необходимо определить диапазон возможных принимаемых значений. Исходя из рисунка 16 и таблицы 8, значения принимают значения в диапазоне от 1,35 до 108,8 нг/мл. Так как четкой взаимосвязи на рисунке 16 не просматривается, то основополагаясь на знания и опыт экспертов области медицины были выделены следующие диапазоны: Первой стадии болезни печени соответствуют значения , где [0; 50], второй стадии соответствуют значения [5;40], третьей стадии соответствуют значения [10; 70]. Поведение не имеет четко определяемой структуры принадлежности к стадии болезни печени и носит лишь нечеткую распределённость по интервалам.

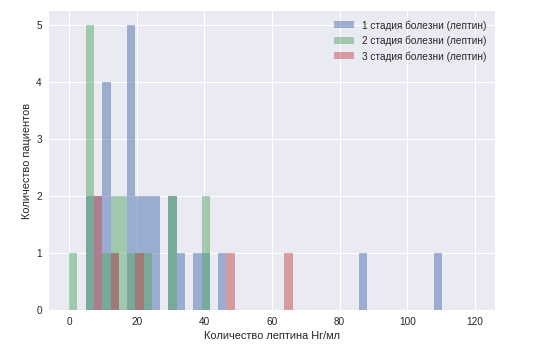


Рисунок 16 – график распределения количества в зависимости от стадии болезни.

Поведение исследуемой системы описывается в терминах лингвистических переменных [79, 80]. Опираясь на полученные диапазоны данных по и строиться нечеткая модель базы данных**,** где каждое правило состоит из двух частей: часть ЕСЛИ (IF), называемая посылкой или условием, и часть ТО (THEN), называемая заключением. IF < посылка> THEN [30].

Правила могут содержать одну или несколько посылок. Конъюнктивные правила (conjunction) - правила «И» или «AND» требуют одновременного выполнения всех посылок. IF < посылка 1> AND < посылка 2> …AND < посылка n> THEN < заключение > Дизъюнктивные правила (disjunction) - правила «ИЛИ» или «OR» требуют выполнения хотя бы одной из посылок. IF < посылка 1> OR < посылка 2>…OR < посылка n> THEN < заключение > Часть «THEN» может содержать несколько заключений: IF < посылка > THEN < заключение 1> < заключение 2>…< заключение m> Посылка содержит две части: лингвистический объект и его значение (характеристика). Заключение содержит оператор, объединяющий объект и значение его характеристики.

В результате получено 11 правил нечетких продукций следующего вида:

RULE0: ***ЕСЛИ*** *ЕСТЬ S* ***И*** *ЕСТЬ S* ***ТО*** *ЕСТЬ*

RULE1: ***ЕСЛИ*** *ЕСТЬ S* ***И*** *ЕСТЬ M* ***ТО*** *ЕСТЬ*

RULE2: ***ЕСЛИ*** *ЕСТЬ M* ***И*** *ЕСТЬ S* ***ТО*** *ЕСТЬ*

RULE3: ***ЕСЛИ*** *ЕСТЬ M* ***И*** *ЕСТЬ M* ***ТО*** *ЕСТЬ*

RULE4: ***ЕСЛИ*** *ЕСТЬ M* ***И*** *ЕСТЬ L* ***И*** *ЕСТЬ (2)* ***ТО*** *ЕСТЬ*

RULE5: ***ЕСЛИ*** *ЕСТЬ M* ***И*** *ЕСТЬ L* ***И*** *ЕСТЬ (1)* ***ТО*** *ЕСТЬ*

RULE6: ***ЕСЛИ*** *ЕСТЬ L* ***И*** *ЕСТЬ M* ***И*** *ЕСТЬ (2)* ***ТО*** *ЕСТЬ*

RULE7: ***ЕСЛИ*** *ЕСТЬ L* ***И*** *ЕСТЬ M* ***И*** *ЕСТЬ (1)* ***ТО*** *ЕСТЬ*

RULE8*:* ***ЕСЛИ*** *ЕСТЬ L* ***И*** *ЕСТЬ S* ***ТО*** *ЕСТЬ*

RULE9*:* ***ЕСЛИ*** *ЕСТЬ S* ***И*** *ЕСТЬ L* ***ТО*** *ЕСТЬ*

RULE10: ***ЕСЛИ*** *ЕСТЬ L* ***И*** *ЕСТЬ L* ***ТО*** *ЕСТЬ*

*ЕСТЬ S* [0 ; 40]

*ЕСТЬ M* [20 ; 40]

*ЕСТЬ L* [30 ; 60]

*ЕСТЬ S* [3 ; 8]

*ЕСТЬ M*  [6 ; 16]

*ЕСТЬ L* [10 ; 20]

– , характеризует воспалительный процесс в организме.

–стадия заболевания печени,

Диапазоны изменения входных переменных разбиваются на термы значений: – малое, – среднее, – большое.

Четкое число задающее заключение каждого правила, принимает значения стадий ; , .

# 3.6.3 Алгоритм работы модуля нечеткого классификатора

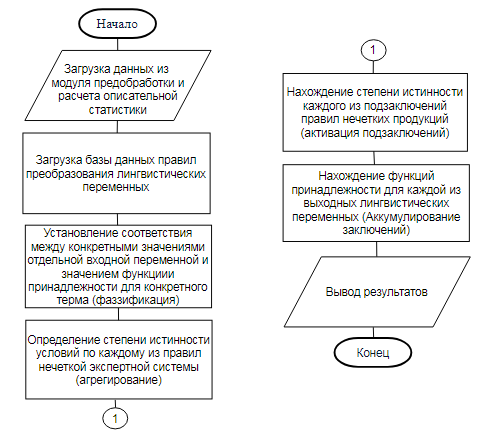


Рисунок 17 – Алгоритм модуля нечеткого классификатора системы ранней диагностики заболевания

Нечеткий классификатор [6] – алгоритм классификации, основанный на извлечение нечетких правил из массивов данных. Основная идея нечеткого классификатора состоит в описании предполагаемого кластера нечетким прототипом, размерность которого определена размерностью пространства исследуемых данных [7]. Таким образом, каждый кластер определяется нечетким правилом на основе базы нечетких процедурных правил. В процессе нечеткой классификации объект относится к каждому классу с определенной степенью принадлежности.

Алгоритм модуля нечеткого классификатора экспертной системы ранней диагностики заболевания состоит из семи шагов, перечисленных ниже.

* 1. Загрузка параметров пациентов, отобранных на модуле выявления значащих переменных.
  2. Загрузка базы данных правил, предназначенная для формального представления эмпирических знаний или знаний экспертов. База правил представляет собой конечное множество правил нечетких продукций, согласованных относительно используемых в них лингвистических переменных. В данном случае формируется база данных правил использующих входные лингвистические переменные значений параметров полученных на модуле «выявления значащих переменных». Выходной лингвистической переменной является принадлежность к стадии фиброза или
  3. Фаззификация множества m – количество входных лингвистических переменных. При этом значения используются в качестве аргумента . Тем самым находится количественное значение по формуле (11).

|  |  |
| --- | --- |
| *,* | (11) |

где – функция принадлежности.

Получившиеся значения, являются результатами фаззификации и данный шаг, является выполненным, когда найдены все значения для каждого из подусловий всех правил, входящих в рассматриваемую базу правил преобразования лингвистических переменных.

Целью этапа является установление соответствия между конкретным значением отдельной входной переменной системы нечеткого вывода и значением функции принадлежности соответствующей ей терма входной переменной.[20]

* 1. Определения степени истинности условий по каждому из правил созданных и загруженных на предыдущем этапе. Этап агрегирования считается законченным, когда будут найдены все значения принадлежности для каждого из правил входящих в базу правил.

1. Нахождение степени истинности каждого из подзаключений правил нечетки продукций.
2. Нахождение функции принадлежности для каждой из выходных лингвистических переменных. Цель аккумуляции заключается в том, чтобы объединить все степени истинности заключений (подзаключений) для получения функции принадлежности каждой из выходной переменной.
3. Вывод информации на ЭВМ пользователя. Данные включают в себя предполагаемую стадию заболевания, значение и при которых было поставлен диагноз, а также вероятность принадлежности к другим стадиям болезни печени.

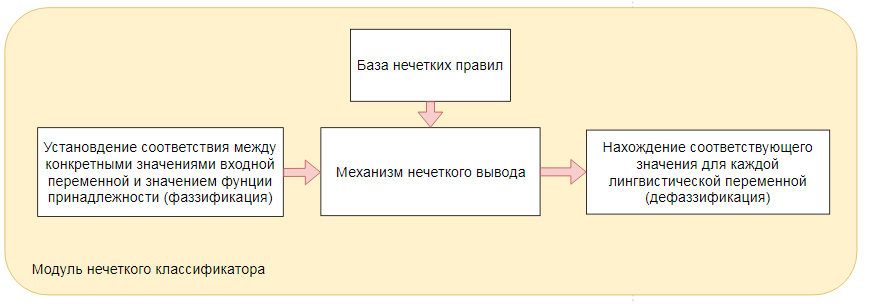


Рисунок 18 – Структурная схема нечеткого классификатора системы ранней диагностики заболевания

# Выбор функций принадлежности для фаззификации

В качестве функций принадлежности для задания лингвистических термов входных переменных принята – образная функция (Z – form), а для принята – образная функция, формируемые в соответствие с выражением [5]

|  |  |
| --- | --- |
|  | (12) |

где и – числовые параметры, упорядоченные отношением < . Параметр c задает координату перегиба функции, а коэффициент определяет наклон функции в точке перегиба (с увеличением значения наклон растет).

При этом случай >0 соответствует S – образной функции принадлежности, а случай <0 соответствует Z – образной функции принадлежности.

В качестве функций принадлежности для задания лингвистических термов входных переменных прията симметричная гауссова функция (gaussmf), формируемая в соответствии с выражением

|  |  |
| --- | --- |
|  | (13) |

где параметр задает модальное значение функции,

ширину. Для лингвистических термов Z (малая, низкая) модальное значение функции принадлежности = 0.

Выше описанные функции принадлежности для термов входных переменных выбраны в связи с тем, что данные функции принадлежности на всей области определения являются гладкими и принимают ненулевые значения, что в рамках прикладной задачи является существенным фактором, так как значения могут выходить за установленные ранее пределы.

Экспертная система на рисунке ? содержит 2 подсистемы. Первая подсистема выполняет операцию нормализации значения В зависимости от пола пациента, в случае женщины, значение делится на коэффициент нормализации 1.5[76, 77], что получено в ходе работы и изображено на рисунке 5.

Синтезированная поверхность «входы-выход» при постановку диагноза болезни печени.

# Заключение

Во вводной части рассмотрены проблемы ранней диагностики НАЖБП. Основная проблема – это отсутствие уже изученных признаков присущих стадии полного проявления.

Использование нечеткой логики дает возможность пересмотреть сам подход к проблеме, что в свою очередь позволит выявить новые зависимости между результатами анализов.

Сформулирована научная новизна работы и определены методы для выявления достоверности результатов работы.

В результате работы сформулированы первоначальные требования к формированию данных, определена цель и назначение работы, составлен алгоритм исследования, а также поставлены задачи, которые будут решаться в ходе дальнейшей работы.

В разделе первичный анализ данных рассмотрены основные этапы первичного анализа данных. Параметры структурированы по 4 группам: лабораторные исследования, сопутствующие заболевания, данные по пациенту, принимаемые препараты. Следующим шагом было построение описательной статистики по всем имеющимся параметрам, для ознакомления с представленными данными и нахождение простых зависимостей. На основе корреляционной матрицы было произведена фильтрация данных, в ходе которой исключены параметры с малым значением корреляции с параметром . В результате был построен граф взаимосвязи параметров рисунок 7, где выделены две группы: и , которые вносят наибольший «вклад» в развитие болезни.

В разделе проектирование экспертной системы рассмотрены зависимости и со стадией болезни печени. Построены графики распределения значений и выявлены соответствующие интервалы, принадлежащие к стадии болезни. Полученные данные были формализованы и использованы в формировании нечеткой базы правил экспертной системы.

# Список использованных источников

1. Von Frerichs F. T. Uber den Diabetes. Berlin. 1884 316с.
2. Zelman S. The liver in obesity. AMA archives of internal medicine. 1952; 90 (2): 141-156.
3. «Распространенность неалкогольной жировой болезни печени у пациентов амбулаторно-поликлинической практики в Российской Федерации: результаты исследования DIREG 2» В.Т. Ивашкин, О.М. Драпкина, И.В. Маев, А.С. Трухманов, Д.В. Блинов, Л.К. Пальгова, В.В. Цуканов, Т.И. Ушакова // Российский журнал гастроэнтерологии, гепатологии, колопроктологии. — 2015. — № 6. — С.31-41.
4. «Значение эндотелиальной дисфункции у больных неалкогольной жировой болезнью печени» Гейвандова Наталья Иогановка, Ягода Александр Валентинович, Нигиян Заринэ Викторовна, Бабашева Гаянэ Гамлетовка// Журнал ДОКТОР.РУ , гепатологии, колопроктологии. — 2015. — № 6. — С.31-41.
5. Нигиян Заринэ Викторовна «Эндотелиальная дисфункция при неалкогольной жировой болезни печени»: Диссертация канд. мед. наук

14.01.04 // Нигиян Заринэ Викторовна 2016. 149с.

1. Ханарин Николай Владимирович «Возможности ультразвукового исследования в диагностике неалкогольной жировой болезни печени у больных с метаболическим синдромом» Диссертация канд. мед. наук

14.01.04 // Ханарин Николай Владимирович 2015. 108с.

1. Mehta S., Insulin resistance, lipotoxicity, type 2 diabetes and atherosclerosis: the missing links. The Claude Bernard Lecture 2009 Diabetologia (2010) 53:1270– 1287.
2. «Prognosis of liver disorders in DNA postitive HBV patiens based on fuzzy soft sets» Ramadhan A. Mohammed a,\*, Ahmed Mohammed Salihb, Muayad Aghali Mirzac, Tahir H. Ismaild, Ahmed A. Allame
3. Gilmor I.T., Burroughs A., Murray-Lyon I.M. et al. Indications, methods, and outcomes of percutaneous liver biopsy in England and Wales: an audit by the British Society of Gastroenterology and the Royal College of Physicians of London // Gut. – 1995. – Vol. 36. – P. 437–441.
4. Петров Д.В. Диагностика, лечение и профилактика расстройств,вызванных употреблением алкоголя. – Ярославль, 2003, С. 90-92 .
5. Жетимекова Г.Ж. Нечеткая классификация с использованием нечеткого анализа кластеризации //ВЕСТНИК КГУСТА — 2014. — № 4. — С.117-120.
6. А. Пегат «Нечеткое моделирование и управление» / А. Пегат – Москва:

«БИНОМ. Лаборатория знаний» 2013. — с 180.

1. KDnuggets [Электронный ресурс]. – Режим доступа : https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html, свободный. – Загл. с экрана.
2. Текстовый формат данных CSV [Электронный ресурс]. – Режим доступа : https://ru.wikipedia.org/wiki/CSV, свободный. – Загл. с экрана.
3. Диаграмма распределения [Электронный ресурс]. – Режим доступа : https//ru.wikipedia.org/wiki/Ящик\_с\_усами, свободный. – Загл. с экрана.
4. Анализ данных [Электронный ресурс]. – https://www.intuit.ru/studies/courses/6/6/lecture/172, , свободный. – Загл. с экрана.
5. Описательная статистика [Электронный ресурс]. – <http://www.machinelearning.ru/wiki/index.php?title=Описательная> статистик, ,свободный. – Загл. с экрана.
6. Регрессионный анализ [Электронный ресурс]. – http://www.machinelearning.ru/wiki/index.php?title= Регрессионный анализ, свободный. – Загл. с экрана.
7. Джарратано, Джозеф, Райли, Гари. Экспертные системы: принципы разработки и программирование, 4-е издание. : Пер. с англ. — М. : ООО “И.Д. Вильямс”, 2007. — 1152 с. : ил. — Парал. тит. англ.
8. Леоненков А.В. Нечеткое моделирование в среде MATLAB и FuzzyTECH СПб: БХВ-Перебург, 2005. – 736 с.
9. Неалкогольная жировая болезнь печение и неалкогольный стеатогепатит [Электронный ресурс]. –http://www.worldgastroenterology.org/UserFiles/file/guidelines/nafld-nash-russian-2012.pdf – Загл. с экрана.
10. Zadeh L.A. Is there a need for fuzzy logic? // Information Sciences, Elsevier– 2008. – № 178. – P. 2751-2779.
11. Интеллектуальные системы: коллективная монография. Вып. 3 / Редкол.: В.М. Курейчик и др. – М.: Физматлит, 2009. –196 с
12. Полковникова Наталья Анатольевна, Курейчик Виктор Михайлович Разработка модели экспертной системы на основе нечёткой логики // Известия ЮФУ. Технические науки. 2014. №1 (150). [Электронный ресурс]. – https://cyberleninka.ru/article/n/razrabotka-modeli-ekspertnoy-sistemy-na-osnove-nechyotkoy-logiki – Загл. с экрана.
13. Безмельцын, Н.Г. Подготовка выпускных квалификационных работ в высших учебных заведениях физической культуры: учебное пособие. – 2-е изд., перераб. и доп / Н.Г. Безмельцын, Н.А. Астафьев. – Омск, 2007. – 124с.
14. SciPy [Электронный ресурс]. – <https://www.scipy.org/> – Заглавие с экрана
15. Scipy.stats.kurtosis [Электронный ресурс] – <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kurtosistest.html> – Заглавие с экрана. Дата обращения:09.01.2019
16. Scipy.stats.normaltest [Электронный ресурс] – <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html> – Заглавие с экрана. Дата обращения:09.01.2019
17. Критерии точки распределения [Электронный ресурс]. – Режим доступа : https://math.semestr.ru/group/xixi.php, свободный. – Загл. с экрана. Дата обращения: 10.01.2019
18. Спицын В.Г. C 72 Разработка экспертных систем на основе нечетких правил вывода: методические указания к лабораторным работам. – Томск: Изд-во ТПУ, 2011. – 33 с.
19. Ливзан М.А., Лаптева И.В., Кролевец Т.С., Черкащенко Н.А. Медицинский совет // Лептинорезистентность у пациентов с неалкогольной жировой болезнью печени, ассоциированной с ожирением и избыточной массой тела. – 2015. – №13. – C.58-63.
20. Б.А. Кобринский Система поддержки принятия врачебных решений // Необходимость в медицине и необходимость ее отражения в экспертных системах. – 2016. – №5. – C.6-14.
21. Б.Б. Оразбаев // Экспертные системы для медицинской диагностики с применением метод теории нечетких множеств. [Электронный ресурс] – Режим доступа:http://itportal.ru/science/tech/ekspertnye-sistemy-dlya-meditsinsko/, свободный. – Загл. с экрана. Дата обращения: 21.02.2019
22. А.И. Поворознюк, Харченко Е.С Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование // Применение нечеткой логики в компьютерных системах медицинской диагностики. – 2015. – №33. – C.125-133.
23. Saunders, J.B., Aasland, O.G., Babor, T.F., de la Fuente, J.R. and Grant, M. // Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption. II. Addiction, 88, 791-804, 1993.
24. Nerys Williams // The CAGE questionnaire, Occupational Medicine, Volume 64, Issue 6, September 2014, Pages 473–474, <https://doi.org/10.1093/occmed/kqu058>
25. Sanai, F.; Keefe, E. // Liver biopsy for histological assessment-the case against. Saudi J. Gastroloenterol. 16(2), 124-132(2010)
26. S. S. Sweidan, S. El-Sappagh, and H. Elbakry // «Liver ﬁbrosis diagnosis with Mamdani FIS» J. Adv. Res. Des., vol. 1, no. 1, pp. 17–24, 2018
27. Saleh, E.; Valls, A.; Moreno, A.; Romero, P.: Diabetes retinopathy risk estimation using fuzzy rules on electronic health record data. In: Modeling Decision for Artificial Intelligence MDAI Lecture Notes in Computer Science vol. 2016, no. (9880), pp. 263–274 (2016)
28. Tsipouras, M.; Fotiadis, D.; Naka, K.; Michalis, L. // Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans. Inf. Technol. Biomed. 12(2), 447–458 (2008)
29. Satarkar, S.; Ali, M.: Fuzzy expert system for the diagnosis of common liver disease. Int. Eng. J. Res. Dev. 1(1), 2–7 (2014)
30. Bohacik, J.; Zabovsky, M. // Dissimilarity measure for comparison of fuzzified instance and its application in a fuzzy rule-based system for heart failure domain. In: 14th International Symposium on Applied Machine Learning and Informatics, pp. 339–344. IEEE (2016)
31. Orczyk, T.; Porwik, P. // Liver fibrosis diagnosis support system using machine learning methods. Adv. Comput. Syst. Secur. 395, 111–121 (2015)
32. Malmir, B.; Amini, M.; Chang, S. // A medical decision support system for disease diagnosis under uncertainty. Expert Syst. Appl. 88, 95–108 (2017)
33. Farokhzad, M.; Ebrahimi, L. // A novel adaptive neuro-fuzzy inference system for the diagnosis of liver disease. Int. J. Acad. Res. Comput. Eng. IJARCE 1(1), 61–66 (2016)
34. Gorunescu, F.; Belciug, S.; Gorunescu, M.; Badea, R. // Intelligent decision making for liver fibrosis stadialization based on tandem feature selection and evolutionary driven neural network. Expert Syst. Appl. 39, 12824–12832 (2012)
35. Gadaras, I.; Mikhailov, L. // An interpretable fuzzy rule-based classification methodology for medical diagnosis. Artif. Intell. Med. 47(1), 25–41 (2009)
36. Badria, F.; Gabr, S. // Prediction of liver fibrosis and cirrhosis among egyptians using noninvasive index. J. Pure Appl. Microbiol. 1(1),45–50 (2007)
37. El-Sappagh, S.; Ali, F.; Ali, A.; Hendawi, A.; Badria, F.; Suh, D. // Clinical decision support system for liver fibrosis prediction in hepatitis patients: A case comparison of two soft computing techniques. IEEE Access 6, 52911–52929 (2018)
38. Kumar, Y.; Sahoo, G. // Prediction of different types of liver diseases using rule based classification model. Technol. Health Care 21, 417–432 (2013)
39. Raoufy, M.; Vahdani, P.; Alavian, S.; Fekri, S.; Eftekhari, P.; Gharibzadeh, S. // A novel method for diagnosing cirrhosis in patients with chronic hepatitis B: artificial neural network approach. J. Med. Syst. 35, 121–126 (2011)
40. Sartakhti, J.; Zangooei, M.; Mozafari, K. // hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Comput. Method Program Biomed. I08, 570–579 (2012)
41. Sweidan, S.; Elbakry, H.; Elsappagh, S.; Sabah, S.; Mastorakis, N. // viral hepatitis diagnosis: a survey of artificial intelligent techniques. Int. J. Biol. Biomed. 1, 106–116 (2016)
42. L. Ozyilmaz, T. Yildirim // Artificial neural networks for diagnosis of hepatitis disease Proceedings of the International Joint Conference on Neural Networks, 2003, vol. 1 (2003), pp. 586-589
43. Z. Jiang // Support vector machine-based feature selection for classification of liver fibrosis grade in chronic hepatitis C J. Med. Syst., 30 (2006), pp. 389-394
44. Javad Salim iSartakhti, Mohammad Hossein Zangooei, Kourosh Mozafari // Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA) Computer Methods and Programs in BiomedicinVolume 108, Issue 2, November 2012, Pages 570-579
45. Kumar, Y.; Sahoo, G. // Prediction of different types of liver diseases using rule based classification model. Technol. Health Care 21, 417–432 (2013)
46. Zadeh, L. // Fuzzy sets. Inf. Control **8**, 338–353 (1965)
47. Yudan Huanga, Ying Chena, Haochuan Zhu, Weifeng Li, Yun Ge, Xiaolin Huang, Jian He // A liver fibrosis staging method using cross-contrast network Expert Systems with Applications (2019)
48. Florin Gorunescu Smaranda Belciug Marina Gorunescu Radu Badea // Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network Expert Systems with Applications Volume 39, Issue 17, 1 December 2012, Pages 12824-12832
49. Бирюкова Елена Валерьевна, Родионова Светлана Владимировна Сахарный диабет 2-го типа и неалкогольная жировая болезнь печени - болезни современности // Медицинский альманах. 2017. №6 (51). URL: https://cyberleninka.ru/article/n/saharnyy-diabet-2-go-tipa-i-nealkogolnaya-zhirovaya-bolezn-pecheni-bolezni-sovremennosti (дата обращения: 15.04.2019).
50. Salvador Resino, Jose Antonio Seoane, Jose Maria Bellon, Julian Dorado, Fernando Martin-Sanchez, Emilio Alvarez Jaime, Cosinf Juan, Carlo Lopez,Guilllermo Lopéz, Pilar Miralles, Juan Berenguerf // An artificial neural network improves the non-invasive diagnosis of significant fibrosis in HIV/HCV coinfected patients
51. Deepti Mittal, Vinod Kumar, Suresh Chandra Saxena, Niranjan Khandelwal, Naveen Kalrab // Neural network based focal liver lesion diagnosis using ultrasound images // Computerized Medical Imaging and Graphics Volume 35, Issue 4, June 2011, Pages 315-323
52. Макаров, И.О. Распространенность неалкогольной жировой болезни печени у беременных с ожирением / И.О. Макаров, Е.И. Боровкова, Р.Д. Казаков //Акушерство, гинекология и репродукция. – 2012. – №4. – С.18 – 21.
53. Pacana, T. The cardiovascular link to nonalcoholic disease / T. Pacana // Clin. Liver Dis. – 2012. – №16. – Р. 599 – 613.
54. Ивашкин, В.Т. Распространенность неалкогольной жировой болезни печени у пациентов амбулаторно-поликлинической практики в Российской Федерации: результаты исследования Direg 2 // В.Т.Ивашкин, О.М.Драпкина, И.В.Маев и др. / Российский журнал гастроэнтерологии, гепатологии и колопроктологии. – 2015. – Т.XXV. – №6. – С. 31 – 41.
55. Drapkina, O.M. Prevalence of nonalcoholic fatty liver disease in the Russion Federation: the open, multicenter, prospective study, DIREG I // Am. J. Clin. Med. Res. – 2015. – №3(2). – P. 31 – 36.
56. Построение экспертных систем: Пер. с англ/Под.ред. П 63 Ф, Хейеса-Рота, Д.Уотермана, Д.Лената. – М.: Мир, 1987. – 441с., ил.
57. Shortliffe, Edward. (1976). Computer-based medical consultations: MYCIN. Artificial Intelligence - AI. 388.
58. Seaborn.pairlot [Электронный ресурс] – <http://seaborn.pydata.org/generated/seaborn.pairplot.html?highlight=plot> – Загл. с экрана. Дата обращения: 25.05.2018
59. Метод наименьших квадратов (статика, динамика, модели с уточняемой структурой) // Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий). 2017. №2. URL: https://cyberleninka.ru/article/n/metod-naimenshih-kvadratov-statika-dinamika-modeli-s-utochnyaemoy-strukturoy (дата обращения: 17.04.2019).
60. Pople H.E., Jr., J.D. Myers, and R.A. Miller. 1975. DIALOG: A model of diagnostic logic for interval medicine. In IJCAI 4, pp. 848-855
61. Pople H.E., E., Jr. 1977. The formation on composite hypotheses in fiagnostic problem solving: An exercise in synthetic reasoning. In IJCAI 5, pp. 1030-1037.
62. Дошина А. Д. Экспертная система. Классификация. Обзор существующих экспертных систем // Молодой ученый. – 2016. –№21. – С. 756-758. – URL https://moluch.ru/archive/125/34485/ (дата обращения: 18.04.2019).
63. Штовба С.Д. Проектирование нечетких систем средствами MATLAB. – М.: Горячая линия – Телеком 2007. – 288с., ил.
64. Considine R.V., Considine E.L., Williams C.J., et al. The hypothalamic leptin receptor in humans: identification of incidental sequence polymorphisms and absence of the db/db mouse and fa/fa rat mutations. Diabetes 1996; 45: 992–4.
65. Чубаненко Е. А., Беляева О.Д., Беркович О.А., Баранова Е.И. Значение лептина в формировании метаболического синдрома // Проблемы женского здоровья. – Медиком – 2010 – №1 – с. 45-60.
66. H. Ishibuchi, T. Nakashima, and T. Murata. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems, IEEE Trans. on Systems, Man, and Cybernetics, vol.29, pp. 601-618, May 1999
67. [Mir90] D.P. Miranker. TREAT: A better match algorithm for AI production systems. PhD thesis, 1990.
68. [BW09] A. Burns and A. Wllings. Real-Time Systems and Programming Languages. Pearson Education Limited, 2009.

# Приложение А

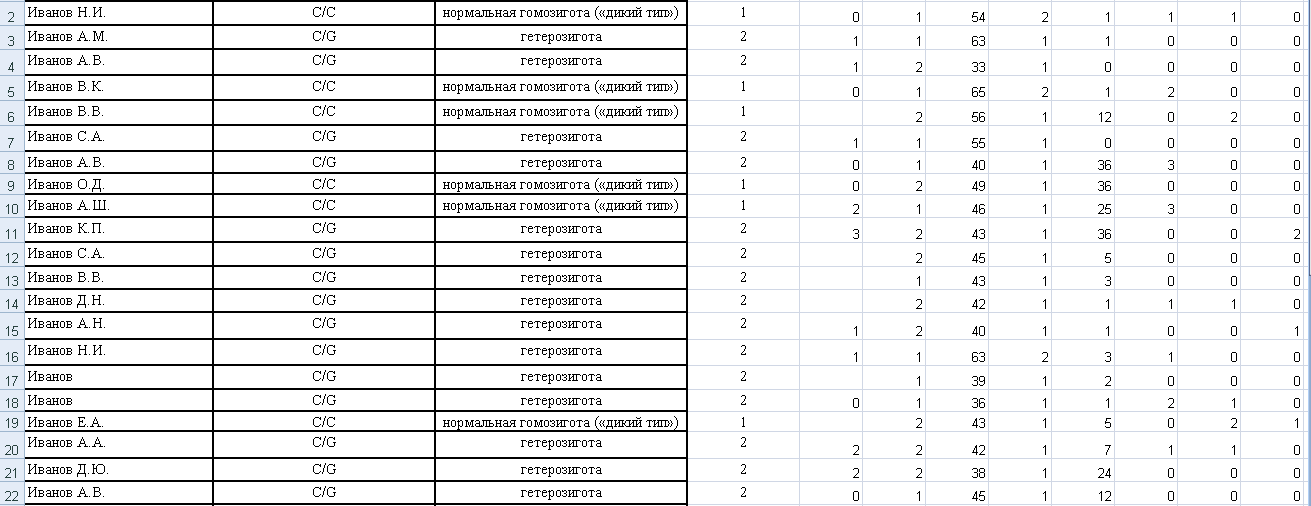
Фрагмент данных параметров пациентов

Рисунок А.1 – Фрагмент данных параметров пациентов полученных от ОмГМУ

# Приложение Б

Листинг программы «проверка физических и лабораторных параметров на соответствие нормальному распределению»

import seaborn as sns

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import warnings

from scipy.stats import kurtosistest

!pip install scipy

import scipy.stats as stats

from scipy.stats import skew

from statsmodels.iolib.table import SimpleTable

warnings.simplefilter('ignore')

pd.set\_option('display.max\_columns', 100)

pd.set\_option('display.max\_rows', 100)

from statsmodels.iolib.table import SimpleTable

row = ['Эласто F', 'лептин', 'увел.печени', 'ObR', 'дискомфорт', 'ТИМР2', 'ожирение','ММП9', 'время динамики', 'остеоартроз', 'метформин', 'стеатоз-1, гепатит -2', 'ОТ', 'АГ', 'Динамика массы', 'ГГТ (50) ед/л', 'гепатомегалия УЗИ']

row2 = [u'Pirson', u'p-value', u'kurtosis']

for f in row:

pirson\_test = stats.normaltest(data[f].dropna())

a = np.array([pirson\_test])

kurt = kurtosistest(data[f].dropna())

kurt = np.array([kurt])

s = np.hstack((a, kurt))

itog = SimpleTable(s, row2)

print(f)

print(itog)

# Приложение В

**Свидетельство о разработке автоматизированной системы прогноза фиброза при неалкогольной жировой болезни печени**

