МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный университет аэрокосмического приборостроения»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

на тему: Программные пакеты и статистические методы в постановке медицинского диагноза

выполнена: Алдохиной Юлией Александровной
фамилия, имя, отчество студента в творительном падеже

по специальности: 01.03.02 Прикладная математика и информатика
код | наименование специальности

наименование специальности

по направлению подготовки/ специальности: 03 Прикладная математика и информатика
код | наименование направленности

в наукоемком производстве
наименование направленности

Студент группы: М611 _____ Ю.А. Алдохина
номер | подпись, дата | инициалы, фамилия

Санкт-Петербург 2020

Оглавление

ВВЕДЕНИЕ

Принятие правильного решения становится ключевым фактором для успешного достижения наших целей во всех областях практической деятельности. Способов найти правильное решение столько же, сколько и людей, которые должны его принять. Можно ожидать, что вновь принятые решения станут лучше и надежнее, но для отдельных лиц и групп, которые должны принимать решения, это на самом деле становится все более и более сложным, поскольку они просто не могут обрабатывать огромные объемы данных. И там возникает необходимость в хорошей технике поддержки принятия решений. Она должна иметь возможность обрабатывать эти огромные объемы данных и помогать экспертам принимать решения легче и надежнее. Таким образом, эксперт может решить, является ли предложенное решение подходящим или нет.

Потребность в системе анализа совпала с появлением интеллектуального анализа данных - процесса обнаружения знаний, который представлял собой смесь машинного обучения, экспертных систем, статистики и т. д. Такая система показала лучшее понимание процесса и предсказания будущего. Основной целью интеллектуального анализа данных является извлечение скрытых знаний из очень больших наборов данных, которые невозможно наблюдать с помощью простого статистического анализа.

Интеллектуальный анализ данных, также называемый обнаружением знаний в базах данных, в области

компьютерных наук, представляет собой процесс обнаружения интересных и полезных моделей и взаимосвязей в больших объемах данных, а также предоставляет лучше понять зависимость между атрибутами выборки в большом наборе данных и интерпретировать процессы подсистемы, создавать законы и предсказания поведения соответствующей подсистемы. Эта область объединяет инструменты из статистики и искусственного интеллекта (такие как нейронные сети и машинное обучение) с управлением баз данных для анализа больших цифровых коллекций, известных как наборы данных. Процесс извлекает информацию высокого качества, которую можно использовать для выработки выводов на основе отношений или структуры данных.

Интеллектуальный анализ данных является результатом использования реализованных алгоритмов в программном обеспечении для удовлетворения потребностей медицинской науки в каждом разделе с построением аналитических моделей, категоризации, информационного прогноза (прогнозирования) и представления.

В настоящее время все интенсивнее развиваются статистические методы и программные пакеты в биологии и экологии [10], географии [12, 15], психологии [8], социологии [3], бизнесе (страхование, банковское дело, розничная торговля), государственной безопасности (обнаружение преступников и террористов) и т.д.

На **актуальность темы** указывают такие факторы, как сокращение времени, необходимого для постановки диагноза, который позволяет медицинским работникам лучше

расставить приоритеты в случае пациента. Интеллектуальный анализ данных и глубокое обучение могут анализировать гораздо больше факторов и случаев, чем работник. Чтобы быть более точным, мы можем использовать его для исследования генома, разработки лекарств, медицинской визуализации. Устройства на основе интеллектуального анализа данных могут изучать, анализировать большие объемы информации и принимать решения гораздо быстрее, чем люди.

Важной целью исследований в области диагностической медицины является оценка и сравнение точности диагностических тестов, которые служат двум целям:

1) предоставление достоверной информации о состоянии пациента и

2) формированию плана лечение пациента на основе установленного диагноза.

Как следует из вышеизложенного, здесь существенные продвижения могут быть получены именно за счет применения интеллектуального анализа данных, осуществляемого с помощью современных программных средств.

Целью данного исследования является оценка возможности эффективного использования алгоритмов для прогнозирования пациентов при поступлении в больницу, что, в свою очередь, открывает и возможности прогнозировать необходимое лечение для пациентов, а также обеспечить необходимые меры для пациентов с травмами, которые находятся до входа в критическую ситуацию. Для реализации цели должны быть решены следующие задачи:

- исследование возможности применения статистического метода***кластерного анализа***
- исследование возможности применения алгоритма построения ***дерева решений***

Объектом исследования являются медицинские данные пациентов одной из поликлиник г. Котлас.

Предметом исследования являются статистические программные пакеты.

1. СОВРЕМЕННЫЕ ПРОГРАММНЫЕ ПАКЕТЫ, ИСПОЛЬЗУЕМЫЕ ДЛЯ РЕШЕНИЯ ЗАДАЧ В МЕДИЦИНЫ

В настоящее время существует достаточно большой выбор программ и программных пакетов, которые использует современная медицина при решении как оперативных, так и относительно долговременных задач. Диапазон этих задач достаточно широк. Так, например, программные пакеты используются для решения различных задач, как: построение баз данных;интерпретации

медицинских данных; контроль качества (оценить эффективность лечения);анализ тенденций заболеваемости, статистических данных о пациентах и информации об использовании;информация о состоянии здоровья населения и постановки диагноза.

Рассмотрим основные программные средства, распространенные в медицинской практике.

1.1. MSEXCEL

В данное время существует большое множество отраслей и предприятий, нуждающихся в аналитики, но стоимость и чрезмерная сложность данного программного обеспечения часто вынуждает отказаться от идеи построения собственной аналитической системы и отдается предпочтение MS Excel.

Для изучения MS Excel, как статистический пакет, в медицине были рассмотрены учебные пособия Шеламова М. А. «Использование программы Excel в работе с базой медико-биологических данных» [17] иКорсунова Е.С., Тишакова К.Д. «Применение пакета STATISTICA и MS EXCEL для обработки биомедицинской информации» [2].

MicrosoftExcel (MS Excel) - широко используемая программа для сбора данных и статистического анализа.Excel наиболее часто используемая электронная таблица для ПК. Программа легко доступна без дополнительных затрат для всех, кто пользуется настольным компьютером. Многие компьютеры часто поставляются с уже загруженным Excel. Так же он может быть полезной платформой для ввода и ведения данных исследований. Excel довольно прост в освоении и использовании. Исследователи

могут использовать простые статистические и графические функции Excel, чтобы помочь лучше понять их данные.

Функции, которые имеет MSExcel: импорт и работу с данными; выявление, классификация и представление данных; установление границ и проверка гипотез; проверка среднего значения; проверка шаблонов; визуализация данных и пространственный анализ; интерпретация дисперсии; анализ текста и многое другое.

Excelвесьма удобен для ввода данных и быстрой обработки строк и столбцов перед статистическим анализом.

Однако эта программа имеет много ограничений, включая меньшее количество функций, которые можно использовать для анализа, и ограниченное количество ячеек по сравнению со специализированными статистическими программами.

Одним из наиболее известных задокументированных недостатков Excelявляется его набор вычислительных алгоритмов. В основном это касается вычислительных алгоритмов для базовой статистики. Перечислим основные недостатки:

- Excel использует плохие алгоритмы, чтобы найти стандартное отклонение;
- Excel не обрабатывает связанные наблюдения правильно при ранжировании;
- расчеты регрессии часто ошибочны из-за плохих алгоритмов.

Кроме того, Excel обычно отображает гораздо больше цифр, чем нужно (естественно для восприятия).

Регрессия в Excelимеет следующий ряд трудностей с регрессионными процедурами:

- не относится к моделям с нулевым перехватом;
- иногда получает отрицательные суммы квадратов;
- не справляется с мультиколлинеарностью правильно;
- вычисляет стандартизированные остатки неправильно;
- отображает нормальные вероятностные графики, которые полностью неверны;
- делает выбор переменных очень сложным.

1.2. STATISTICA

Линейка продуктов StatSoft представляет собой набор аналитических программных продуктов STATISTICA. STATISTICA предоставляет наиболее полный набор процедур анализа данных, управления данными, визуализация данных, интеллектуальный анализ данных, машинное обучение, процедуры анализа текста. Его методы включают в себя широкий выбор методов прогностического моделирования, кластеризации, классификации и исследования в одной программной платформе. Аналитические возможности STATISTICA дополняются множеством уникальных функций, в том числе:

- Запросы к базам данных;
- Визуализация данных.

Графические средства в STATISTICA сочетают в себе чрезвычайно широкий выбор научных и технических диаграмм (со встроенными аналитическими средствами) с

возможностями настройки, рисования и мультиграфического управления, которые обычно присутствуют только в специально предназначенных графических программах и программах для рисования. STATISTICA предлагает сотни типов 2- и 3-мерных графических дисплеев, в том числе 2- и 3-мерные троичные графы, специальные 4-мерные графы, многомерные графы, матрицы графиков, спектральные 2- и 3 трехмерные графы, составные графы и многие другие специализированные процедуры. Кроме того, гибкие и очень простые в использовании средства позволяют настраивать совершенно новые типы графиков и постоянно добавлять их в меню или плавающие панели инструментов.

Исследователь может использовать различные виды классификации и методы регрессионного анализа, реализованные в пакете STATISTICA, в том числе:

- Классификация и регрессия;
- Автоматизированная нейронная сеть;
- Общая классификация и регрессия с помощью деревьев;
- Общая модель CHAID;
- Случайный лес (Randomforest) др.

Комплексные реализации специализированных методов для анализа данных используются в различных сферах(например, интеллектуальный анализ данных; бизнес, социальные науки и биомедицинские исследования [13]).

Одним из главных преимуществ программы STATISTICA является широкий спектр алгоритмов (автокорреляции, множественная регрессия, аппроксимация, графический анализ таблиц, вычисление экстремумов, подгонка

распределений, байесовский анализ и т.д.), вторым по важности преимуществом это высокая точность расчетов.

Также эта программа имеет и ряд недостатков, таких как:

- использование пакета STATISTICA требует больших знаний теории «Теория вероятности и математическая статистика»;
- относительно сложный интерфейс.

1.3. SPSS STATISTICS

SPSS - сокращение от «Статистический пакет для социальных наук».SPSS Statisticsиспользуется исследователями рынка, исследователями здравоохранения, исследовательскими компаниями, государственными структурами, исследователями в области образования, маркетинговыми организациями и многими другими для обработки и анализа данных обследований, например, в области медицины использовалась для нахождения ролинейропротекции в терапии гипертонической энцефалопатии.

Это программное обеспечение является одним из самых популярных статистических пакетов, которые могут выполнять очень сложные манипуляции и анализ данных при выполнении относительно простых инструкций. SPSS может брать данные практически из любого типа файлов и использовать их для создания табличных отчетов, диаграмм и графиков распределений и тенденций, описательной статистики и проведения комплексного статистического анализа.

Это программное обеспечение широко использовалось исследователями для проведения количественного анализа с момента его разработки в 1960-х годах Норманом Х. и сотрудничестве с К. ХадлайХаллом и Дейлом Бентом.

Программное обеспечение SPSS может считывать и записывать данные из других статистических пакетов, баз данных и электронных таблиц.

Есть много статистических методов, которые можно использовать в SPSS, а именно:

- прогнозирование разнообразных данных для определения групп, включая такие методологии, как кластерный анализ, факторный анализ, дисперсионный анализ и т. д.;
- описательные статистические данные, в том числе методологии SPSS, представляют собой статистические данные о частотах, перекрестных таблицах и описательных соотношениях, которые очень полезны;
- кроме того, двумерная статистика, включая методологии, такие как дисперсионный анализ (ANOVA), средние значения, корреляционные и непараметрические тесты и т. д.;
- прогноз числового результата, такой как линейная регрессия.

Преимущества данного пакета:

- более легкий доступ, управление и анализ практически любого типа данных;
- надежные результаты с широким спектром испытаний и процедур;

- отчет о результатах в простых для понимания форматах.

Основныминедостатками SPSS являются: нельзя использовать для анализа очень большой набор данных и высокая цена.

1.4. SAS VISUAL ANALYTICS

SAS («Система статистического анализа») - это набор статистического программного обеспечения, разработанный институтом SAS для управления данными, расширенной аналитики, многомерного анализа, бизнес-аналитики, уголовного расследования и прогнозной аналитики.

SAS программный комплекс, который может добывать, изменять, управлять и извлекать данные из различных источников, а также выполнять статистический анализ. Он не только предоставляет организациям все необходимые инструменты для мониторинга, но также предоставляет мощную аналитику и отчеты для лиц, принимающих решения, для принятия обоснованных решений.

SAS VisualAnalytics помогает анализировать большие данные предприятия и генерировать из них мощную информацию настолько простым способом, что бизнес-пользователи сами могут сделать вывод из всего этого процесса, таким образом снимая с ИТ-команды эту нагрузку. Этот инструмент позволяет компаниям выявлять тенденции, выявлять корреляцию между данными, выявлять выбросы, осознавать исключения, выявлять причину таких изменений и предлагать новые идеи и идеи, о которых они не знали.

Преимущества:

- пользователи могут применять возможности аналитики SAS к огромным объемам данных;
- позволяет создавать визуальные отчеты и информационные панели на основе обычных таблиц и графиков;
- простое создание моделей и исследование данных.

Недостатки:

- недостаток эффективности в подготовке данных и управлении данными;
- требуется ручная работа в прогнозировании;
- пользовательский интерфейс не является достаточно удобным.

1.5. STATA

Stata - это пакет статистического программного обеспечения общего назначения, созданный в 1985 году компанией StataCorp. Большинство его пользователей работают в области исследований, особенно в области экономики, социологии, политологии, биомедицины и эпидемиологии.

Возможности Stata включают управление данными, статистический анализ, графику, моделирование, регрессию и пользовательское программирование.

Эта программа включает в себя импорт и управление наборами данных, очистку и подготовку данных, создание и управление переменными, создание описательной статистики и значимых графиков, отличную встроенную поддержку моделирования структурных уравнений, а также центральные количественные методы, такие как линейные и

бинарные логистические регрессии, и сопоставления. Дополнительная информация о диагностических тестах гарантирует, что эти методы дают действительные и правильные результаты, соответствующие академическим стандартам.

Преимущества:

- имеет различные пакеты дополнений, такие как скрытый кластерный анализ, пространственные модели AR, нелинейные многоуровневые модели, модели конечных смесей, пороговая регрессия и т. д.;
- обеспечивают расширенное моделирование выбора;
- предлагает широкий спектр статистических анализов;
- хорошая система поддержки;
- надежные оценки и тесты, методы продольных данных, многомерный временной ряд.

Недостатки:

- очень мало литературы на русском языке по работе в программном пакете Stata;
- несколько ограниченная графика;
- не такой гибкий, как программы статистического анализа.

1.6. RAPIDMINER

RapidMiner [21] - это среда для машинного обучения, интеллектуального анализа данных, анализа текста, прогнозной аналитики и бизнес-аналитики. Проект RapidMiner был начат в 2001 году Ральфом Клинкенбергом, ИнгоМиерсвой и Саймоном Фишером из отдела

искусственного интеллекта Технического университета Дортмунда.

С 2007 года RapidMiner был значительно расширен и стал одним из наиболее важных инструментов для анализа и анализа данных. Он интенсивно используется на вводных курсах и в академических целях в университетах по всему миру. RapidMiner также используется в промышленных целях многими компаниями и консультантами для решения различных задач таких как: обнаружение спама [1], определение типа дефекта поверхности в нержавеющих стальных пластинах [5], определение и создание пользовательских групп путем обработки данных использования сайтов [11], поиск управляющих компаний [14] и т.д..

RapidMiner - это централизованное решение с очень мощным и надежным графическим пользовательским интерфейсом, позволяющим пользователям создавать, предоставлять и поддерживать прогностическую аналитику. Помимо предоставления истинной прогностической аналитики, пакет приложений RapidMiner также включает интеграцию данных, преобразование, машинное обучение и интеграцию приложений. Благодаря такому унифицированному подходу RapidMiner ускоряет процесс обучения, улучшает стандартизацию и упрощает обслуживание и расширяемость, что значительно повышает производительность и эффективность.

Программное обеспечение написано на языке программированияJava и запускает так называемые процессы. Процесс в основном представляет собой XML-файл,

сгенерированный пользователем и содержащий последовательность задач, которые представлены операторами. Более 500 операторов уже включены в программное обеспечение. Их функциональные возможности охватывают основные аспекты анализа данных, такие как загрузка и преобразование данных, предварительная обработка и визуализация данных, моделирование и оценка моделей, прогнозную аналитику и статистическое моделирование.

Комбинируя этих операторов, можно выполнять основные задачи машинного обучения, такие как интеллектуальный анализ данных, анализ текста, анализ временных рядов и прогнозирование, веб-анализ, а также анализ настроений и анализ мнений.

Обзор преимуществ RapidMiner:

- RapidMiner предлагает надежный и очень мощный интегрированный набор инструментов и функций, каждый компонент которого представляет собой удобный интерфейс, который помогает пользователям добиться значительного повышения производительности с самого начала. Его инструмент визуального конструктора рабочих процессов предлагает пользователям простую в использовании визуальную среду, которая позволяет им проектировать, создавать и развертывать аналитические процессы, визуальные презентации и модели без проблем;
- система упрощает доступ к данным и управление ими, позволяя получать, загружать и оценивать все виды данных, включая тексты, изображения и аудиодорожки.

RapidMiner позволяет вам структурировать их так, чтобы вам и вашей команде было легко их понять;

- позволяет создавать модели и планы, чтобы вы могли извлекать критическую статистику и информацию, на которой вы будете основывать свои решения и стратегии.

1.7. ВЫВОД К ПЕРВОЙ ГЛАВЕ

В настоящей главе был проведен обзор наиболее известных, и получивших широкое распространение, статистических программных пакетов.

Рассмотрены пакеты программного обеспечения для статистического анализа данных, такие как: MSExcel, STATISTICA, SPSS Statistics, SAS VisualAnalytics, Stata и RapidMinerStudio.

В результате проведенного анализа имеющихся на настоящий момент программных средств, для решения задач прикладной математической статистики для целей медицины можно сделать следующие выводы:

1. Предлагаемая линейка программных пакетов решает практически весь спектр задач прикладной статистики, возникающих в практической медицинской деятельности.

2. Программный пакет **RAPIDMINER,** несомненно, является наиболее эффективным средством решения задачоперативной медицинской практики – диагностики и прогнозирования.

2. ТЕХНОЛОГИЯ РЕШЕНИЯ ЗАДАЧ ДИАГНОСКИ С ИСПОЛЬЗОВАНИЕМ RAPIDMINER

В настоящей главе мы рассмотрим ряд специфических задач прикладной математической статистики, решение которых, с одной стороны, естественным образом определяется задачами и потребностями медицинской оперативной практики, а с другой стороны, наиболее удобно осуществимо в рамках программного пакета RAPIDMINER.

2.1. ЗАДАЧА КЛАССИФИКАЦИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Классификация - это процесс прогнозирования классазаданных точек данных. Нередко задача классификации на языке математической статистики формулируется как задача разделения смеси распределений. В последнем случае она называется кластерным анализом или кластеризацией (см. раздел 2.2).Классификационное прогнозирующее моделирование - это задача приближения функции отображения (f) от входных переменных (x) к дискретным выходным переменным (y).

Задачи интеллектуального анализа данных делятся на две большие группы: прогнозирующие и описательные.

Задачи прогнозирования связаны с построением модели, которая может использоваться для прогнозирования поведения анализируемой системы в ситуации, которая ранее не наблюдалась.

Целью решения описательных задач является поиск скрытых закономерностей в данных, их описание и вывод правил, которые могут быть использованы в будущем для

повышения эффективности работы. Поэтому эту группу задач также называют задачами структурированного интеллектуального анализа данных.

В настоящее время в задачах интеллектуального анализа данных обычно используются различные варианты классификации, показанные на рис. 1.

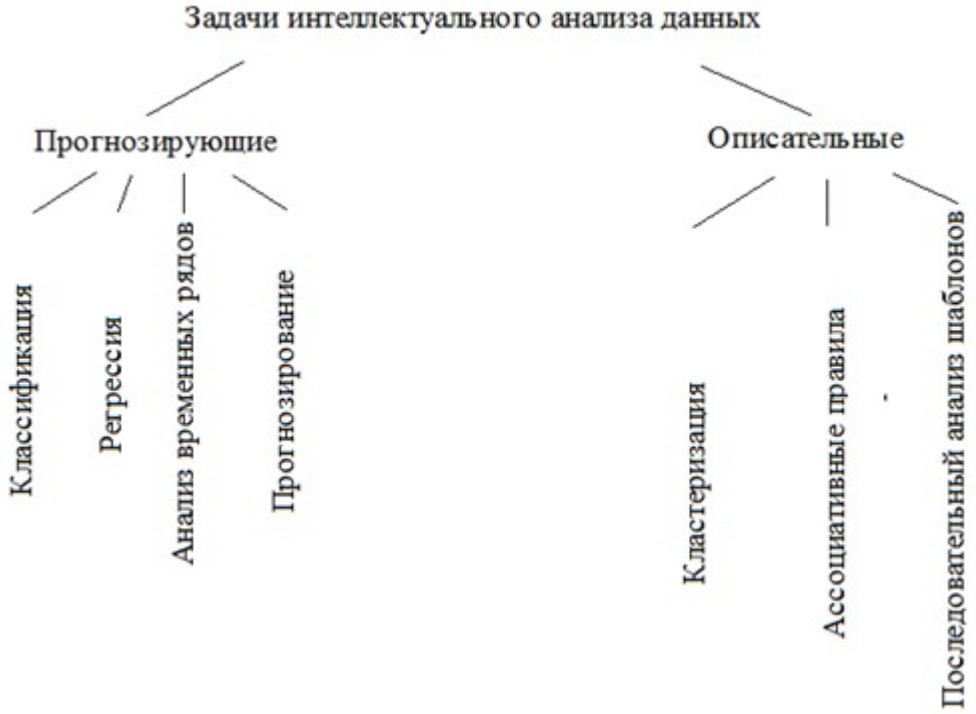

Рис.1. Классификация задач анализа данных

В рамках задач *классификации* решается проблема разделения определенного набора данных на заранее определенный набор классов. Решение задачи классификации позволяет не только изучить существующие данные, но и позволяет прогнозировать будущее поведение системы.

По количеству классов, на которые делится входная выборка, необходимо различать проблемы бинарной и полинарной классификации. С двоичной классификацией во входной выборке выделяются только два класса (точнее, один

класс и все остальное). С помощью полинарной классификации входная выборка делится на три или более классов.

Практическим применением методов классификации в медицине является определение вида опухоли груди (доброкачественная или злокачественная), клиническое исследование о сердечной недостаточности.

Решение проблемы построения*регрессии* включает в себя выявление взаимосвязи между независимыми (входными) переменными и зависимыми (выходными). Суть решения сводится к выводу математической формулы эвристическим или аналитическим способом, выражающим взаимосвязь между входными и выходными данными.

М. Какимото [25] предложил использовать анализ логической регрессии для извлечения правил отношений и изучения взаимосвязи функций мозга с движением пальцев и человеческой речью.

Целью решения задачи *прогнозирования* является аппроксимация (определение) значений некоторых показателей в будущем на основе заданных значений в прошлом и настоящем.

Один из классических примеров в медицине, «Маска Гиппократа», описывает процедуру прогнозирования надвигающейся смерти, основанную на наблюдении отличительных признаков и симптомов, которые он выявил [22].

Целью решения задачи *анализа временных рядов* является прогнозирование будущих значений определенного

набора данных, где значение выходной переменной зависит не только от прошлых значений переменной, но и от времени. Характерной особенностью временных рядов является равномерное распределение входных данных во времени. Анализ временных рядов является своего рода проблемой регрессии, но поскольку он использует конкретные входные данные и методы принятия решений, он выделяется в отдельный класс задач.

Модели анализа временных рядов были использованы для характеристики механизмов почечной ауторегуляции и для выявления взаимодействия между различными ритмами регуляции потока нефронного давления. Они также использовались при изучении тенденций в оказании медицинской помощи. Временные ряды повсюду в нефрологии, и их анализ может привести к открытию ценных знаний.

Давайте перейдем к задачам описательной группы.

При решении проблемы *кластеризации*(см. раздел 2.2) необходимо найти закономерности в массиве входных данных, чтобы выделить несколько зон (кластеров) в нем и распределить данные по этим кластерам. Задача кластеризации напоминает проблему классификации, с существенным отличием в том, что сами классы не определены заранее. Для решения проблемы кластеризации используются алгоритмы обучения без учителя.

Методы кластеризации часто используются для анализа массивов генетической информации. В работе [27] была выполнена кластеризация массива, содержащего ДНК 86

видов опухолей груди. Было получено два кластера. В первый кластер вошли опухоли, дающие рецидив в 34 % случаев, во второй — в 70 %. Первый кластер условно можно назвать «плохо прогнозируемыми опухолями», а второй — «хорошо прогнозируемыми опухолями». Далее эта информация использовалась для повышения точности прогнозирования развития опухолей.

Поиск *правил ассоциации* позволяет устанавливать связи и связи между переменными в больших базах данных. Ассоциативные правила позволяют нам находить закономерности среди связанных событий, то есть они дают возможность ответить на вопрос: «С какой вероятностью связаны события A и B?» Последовательность возникновения событий не имеет значения.

Методы поиска ассоциативных правил находят применение в медицинской диагностике, например, правила базы знаний из экспертной системы, используемой для диагностики заболеваний сердца, проверяются с использованием правила ассоциации.

Последовательный анализ шаблонов. В отличие от поиска ассоциативных правил, последовательный анализ паттернов подразумевает идентификацию причинно-следственных правил, то есть учитывает фактор времени и позволяет ответить на вопрос: «С какой вероятностью возникновение события A влечет за собой событие B?» Обычно предполагается, что события описываются дискретными значениями, что отличает эту задачу от задачи анализа временных рядов.

Например, в медицине были проанализированы последовательность белка и классификация белка по шаблонам, то есть они извлекали последовательные образцы белков, которые затем использовались для классификации неизвестных белков [23].

2.2. КЛАСТЕРНЫЙ АНАЛИЗ

В области интеллектуального анализа данных изучение кластеров является популярным и хорошо изученным способом обнаружения интересных результатов среди огромной базы данных.

К примеру, многомерная классификация данных находит широкое применение в медицинских исследованиях и психологии. Так, В.А. Альбахели в исследовании проводит кластерный анализ работы медицинской техники с целью повышения качества диагностики заболеваний с помощью МРТ. В работе В.П. Пономарева и И.Ю. Белоглазовой исследование показателей крови больных проведено на основе кластерного и факторного видов анализа. Применение кластерного анализа для обработки данных психологических исследования показано в работе. Автор рассматривает теоретико-методические, а также прикладные вопросы применения этого вида анализа, предлагает варианты развития методики классификации и пути совершенствования алгоритмов анализа данных, реализуемых в современных пакетах прикладных программ [18].

Кластерный анализ - это многомерный метод, целью которого является классификация выборки субъектов (или

объектов) на основе набора измеряемых переменных в ряде различных групп, так что похожие предметы помещаются в одну группу. Кластерный анализ также называется классификационным анализом или числовой таксономией. В кластерном анализе нет никакой предварительной информации о членстве в группе или кластере ни для одного из объектов.

Кластерный анализ направлен на обнаружение естественного разделения объектов. Другими словами, он группирует наблюдения, которые похожи на однородные подмножества. Эти подклассы могут выявить закономерности, связанные с изучаемым явлением. Функция расстояния используется для оценки доступности сходства между объектами и широким разнообразием алгоритмов кластеризации, основанных на различных концепциях. Меры подобия сначала вычисляются между наблюдениями и между кластерами, когда наблюдения начинают группироваться в кластеры. Несколько метрик, таких как евклидово и Махаланобиса расстояние [9], могут использоваться для вычисления сходства. Евклидово расстояние рассматривает каждую переменную как одинаково важную при расчете расстояния. Альтернативный подход заключается в масштабировании вклада отдельных переменных в значение расстояния в соответствии с изменчивостью каждой переменной. Этот подход иллюстрируется расстоянием Махаланобиса, которое является мерой расстояния между каждым наблюдением в многомерном облаке точек и центром тяжести облака.Кроме того, возможны несколько стратегий слияния, которые приводят к различным шаблонам

кластеризации. Поэтому результаты кластеризации являются несколько субъективными, поскольку они в значительной степени зависят от выбора пользователей.

Существует множество алгоритмов кластеризации, такие как: иерархическая кластеризация, кластеризация на основе центроидов (например, k-means) и даже нечеткие методы, в которых один объект может в различной степени принадлежать более чем одному кластеру. Какой бы алгоритм вы ни использовали для кластеризации, центральным в подходе является выбор функций (или параметров), на которых будет основана кластеризация.Рассмотрим несколько из них: k-means, k-mediods и x-means.

Кластеризация k-means является наиболее популярным методом. Алгоритм k-means определяет набор из k кластеров и присваивает каждому примеру точное количество кластеров. Кластеры состоят из похожихпримеров. Сходство между примерами основано на измерении расстояния между ними.

Кластер в алгоритме k-means определяется положением центра в n-мерном пространстве из n атрибутов. Эта позиция называется центроид.

Алгоритм k-means начинается с k точек, которые рассматриваются как центроид k потенциальных кластеров. Эти начальные точки являются либо положением k случайно выбранных примеров входных данных, либо определяются эвристикой k-means ++, если для определения хороших начальных значений задано значение true.

Все примеры присваиваются к ближайшему кластеру (ближайший определяется типом меры). Затем центроиды кластеров пересчитываются путем усреднения по всем примерам одного кластера. Предыдущие шаги повторяются для новых центроидов до тех пор, пока центроиды не перестанут двигаться или не будет достигнут допустимый максимум количества шагов оптимизации.

Процедура повторяется максимальное время прогонов с каждым разным набором начальных точек. Поставляется набор кластеров с минимальной суммой квадратов расстояний всех примеров до соответствующих центроидов.

Алгоритм кластеризации k-means чувствителен к выбросам, поскольку среднее значение легко зависит от экстремальных значений.

Работа алгоритма*кластеризации k-medoids*. Предполагая, что мы используем евклидово расстояние или нечто подобное в качестве меры, мы можем определить центр тяжести кластера как точку, для которой каждое значение атрибута является средним значением значений соответствующего атрибута для всех точек в кластере. Центроид кластера всегда будет одной из точек в кластере. В этом главное отличие алгоритма k-means и k-medoids. В алгоритме k-meansцентроид кластера часто будет воображаемой точкой, а не частью самого кластера, которую мы можем взять, чтобы отметить его центр.

Основным недостатком алгоритмов k-medoid является то, что он не подходит для кластеризации несферических (произвольных форм) групп объектов. Это потому, что он основан на минимизации расстояний между немедоидными

объектами и медоидом (центром кластера) - вкратце, он использует компактность в качестве критерия кластеризации, а не связности.

Он может получить разные результаты для разных прогонов одного и того же набора данных, поскольку первые k медоиды выбираются случайным образом.

В статистике и интеллектуальном анализе данных *кластеризация x-means* представляет собой разновидность кластеризации k-means, которая уточняет назначения кластеров путем многократной попытки разделения и сохранения наилучших результирующих разбиений, пока не будет достигнут какой-либо критерий, такого как байесовский информационный критерий.

Основное преимущество кластеризации перед классификацией состоит в том, что она адаптируется к изменениям и помогает выделить полезные функции, которые отличают разные группы.

Так же кластерный анализ широко используется в других сферах помимо медицины таких как:

- анализ текстовых документов [16];
- анализ внутренних затрат на научные исследования и разработки по субъектам Российской Федерации [7];
- классификация регионов по уровню инновационного развития [4];
- изучение экономической деятельности судостроительных и судоремонтных предприятий [6].

2.3. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ

Решения играют важную роль и в медицине, особенно в медицинских диагностических процессах. Системы поддержки принятия решений, помогающие врачам, становятся очень важной частью принятия медицинских решений, особенно в тех ситуациях, когда решение должно приниматься эффективно и надежно. Поскольку для выполнения таких задач следует рассматривать простые концептуальные модели принятия решений с возможностью автоматического обучения, деревья решений являются очень подходящим кандидатом. В 1997 году органы здравоохранения штата Сан-Паулу Бразилии разработала кампанию вакцинации против кори на основе модели принятия решений, которая использует нечеткую логику [28]. Выбранная стратегия массовой вакцинации осуществила и изменила естественный ход эпидемии в этом состоянии. Авторы построилимодель с использованием дерева решений и сравнил его с моделью нечеткой логики.В 2001 году Куо и Чанг рассмотрели и классифицировали результаты ультразвуковых исследований у пациентов с раком молочной железы на основе дерева решений [26]. Также деревья решений использовались в выявление сигналов о возможных побочных реакциях на лекарства были показаны Джонсом [24].

Дерево решений – это инструмент поддержки принятия решений, который использует древовидную диаграмму или модель решений и их возможных последствий, включая случайные исходы событий, затраты ресурсов и полезность. Это один из способов отображения алгоритма, который содержит только условные операторы управления [20].

Дерево решений – это непараметрический метод обучения под наблюдением, используемый для классификации и регрессии. Это древовидный граф, в которой каждый внутренний узел представляет «тест» для атрибута (например, подбрасывание монеты вверх или вниз), каждая ветвь представляет результат теста, а каждый конечный узел представляет метку класса (решение принимается после вычисления всех атрибутов) [20].

Цель состоит в том, чтобы создать модель классификации, которая прогнозирует значение целевого атрибута (часто называемого классом или меткой) на основе нескольких входных атрибутов. В RapidMiner атрибут с ролью метки прогнозируется оператором дерева решений. Каждый внутренний узел дерева соответствует одному из входных атрибутов. Количество ребер номинального внутреннего узла равно количеству возможных значений соответствующего входного атрибута. Исходящие ребра числовых атрибутов помечены непересекающимися диапазонами. Каждый листовой узел представляет значение атрибута label, учитывая значения входных атрибутов, представленных путем от корня до листа. Это описание можно легко понять, изучив прилагаемый пример процесса на рис.1.

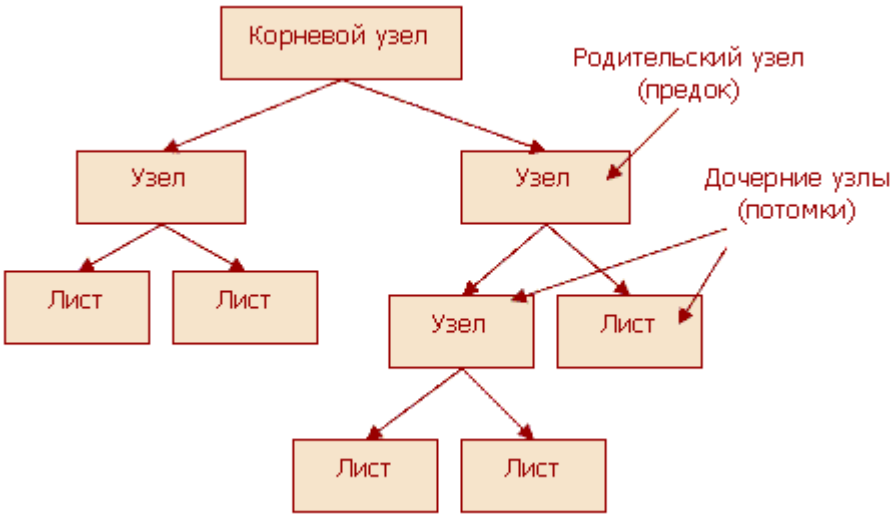

Рис.1. Пример дерева решений

Деревья решений создаются путем рекурсивного разбиения. Рекурсивное разбиение означает многократное разбиение по значениям атрибутов. В каждой рекурсии алгоритм выполняет следующие шаги:

• Атрибут A выбран для разделения. Правильный выбор атрибутов для разделения на каждом этапе имеет решающее значение для создания полезного дерева. Атрибут выбирается в зависимости от критерия выбора, который может быть выбран параметром критерия.

• Примеры сортируются в подмножества, по одному для каждого значения атрибута A в случае номинального атрибута. В случае числовых атрибутов подмножества формируются для непересекающихся диапазонов значений атрибутов.

• Дерево возвращается с одним ребром или ветвью для каждого подмножества. Каждая ветвь имеет дочернее поддерево или значение метки, полученное путем рекурсивного применения одного и того же алгоритма.

Обычно рекурсия останавливается, когда все примеры или экземпляры имеют одинаковое значение метки, то есть подмножество является чистым. Или рекурсия может прекратиться, если большинство примеров имеют одно и то же значение метки.

Сила дерева решений заключается в том, что оно используется для определения атрибутов данных, которые наиболее откровенно отражают классификацию записей для соответствующих возможных результатов. На первом разделении находится атрибут, который является наиболее показательным индикатором классификации членства, а на следующем разделении отражаются другие данные после сегментирования данных из предыдущего разделения. Таким образом, этот процесс является последовательным и повторяющимся, причем каждое разбиение влияет на последующее. В зависимости от настроек предварительного сокращения, применяемых к деревьям решений, деревья могут разрешаться в пределах одного или нескольких уровней. В зависимости от данных, может быть, а может и нет дерево решений, которое выводится из него.

Дерево решений один из самых простых типов визуализации данных, который можно интерпретировать из процесса машинного обучения, потому что его процессы достаточно прозрачны, а кульминация визуализации данных показывает взаимосвязь данных довольно легко читаемым и интерпретируемым человеком способом. Дерево решений информирует исследователя, какие атрибуты наиболее связаны с определенными классификациями. Исходя из этой

информации, исследователь может выдвинуть гипотезу из наблюдений для более глубокого понимания.

В результате дерево принятия решений является одним из наиболее популярных алгоритмов классификации, используемых в интеллектуальном анализе данных и машинном обучении.

Из-за своей простоты древовидные диаграммы используются в широком спектре отраслей и дисциплин решая такие задачи, как:

- оценка возможностей расширения бренда для бизнеса с использованием исторических данных о продажах;
- определение вероятных покупателей продукта с использованием демографических данных для обеспечения ограниченного рекламного бюджета;
- оптимизации ремонтных программ предприятий электроэнергетики РФ [19].

2.4. ВЫВОД К ВТОРОЙ ГЛАВЕ

По результатам второй главы можно сделать следующие выводы:

1. Рассмотрены задачи классификации интеллектуального анализа, применяемые в медицине. Эти задачи позволяют построить модели, которые могут использоваться для прогнозирования поведения анализируемой системы в ситуации, которая ранее не наблюдалась; использовать поиск скрытых закономерностей в данных, их описание и вывод правил, которые могут быть использованы в будущем для повышения эффективности работы.

2. В результате проведенного анализа типов задач математической статистики, решение которых осуществимо в рамках использования программного пакета RapidMiner, наиболее интересной с точки зрения оперативной медицинской практики, несомненно, является, задача построения дерева решений, которая в свою очередь опирается на решение задачи кластеризации. Более подробно рассмотрены и описаны кластерный анализ и деревья решений.

3. Уникальность на настоящий момент пакета RapidMiner в том и состоит, что он пока является единственным широко распространенным программным продуктом, решающим эту последовательность задач. При этом он обладает удобным интерфейсом и предлагает надежный и очень мощный интегрированный набор инструментов и функций, который помогает пользователям добиться значительного повышения производительности с самого начала. Его инструмент визуального конструктора рабочих процессов предлагает пользователям простую в использовании визуальную среду, которая позволяет им проектировать, создавать и развертывать аналитические процессы и модели без проблем.

3. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ПАКЕТА RAPIDMINER

Задачей настоящей главы является проработка и демонстрация решения задач кластеризации и построения дерева решений на реальных медицинских данных, с целью исследования возможностей программного пакета RapidMiner в решении актуальных задач, непосредственно возникающих и имеющих важное значение в оперативной медицинской практике. Начнем с описания исходных данных.

3.1. ИСХОДНЫЕ ДАННЫЕ И ИХ ЗАГРУЗКА

Рассмотрим в качестве исходных данных 49 пациентов различных возрастных категорий, 24 мужчин и 25 женщин в таблице А.1. Пациенты распределились на 9 групп:

1 группа – заболевание АГ 1 степени *(3 пациента),*

2 группа – заболевание АГ 1 степени на фоне почечной недостаточности *(10 пациентов)*,

3 группа – заболевание АГ 1 степени на фоне сердечной недостаточности *(6 пациентов)*,

4 группа – заболевание АГ 2 степени *(6 пациентов)*,

5 группа – заболевание АГ 2 степени на фоне почечной недостаточности *(8 пациентов)*,

6 группа – заболевание АГ 2 степени на фоне сердечной недостаточности *(4 пациентов)*,

7 группа – заболевание АГ 3 степени *(4 пациента)*,

8 группа – заболевание АГ 3 степени на фоне почечной недостаточности *(7 пациентов)*,

9 группа – заболевание АГ 3 степени на фоне сердечной недостаточности *(1 пациент)*.

При обследовании пациентов:

- измерялись температура, систолическое(верхнее) давление, диастолическое (нижнее) давление, пульс;

- проводились анализы на креатинин, СКФ, холестерин, СОЭ, глюкозу.

Количественные показатели гипертонии и их статистический анализ сведены в табл. 1.

Таблица 1 – Показатели гипертонии

Имя	Тип данных	Минимум	Максимум	Среднее значение
А	polynominal			
Диагноз	polynominal			
Пол	integer	0	1	
Возраст	integer	20	80	50
Температура	real	35,5	37,5	36,592
Систолическое давление	integer	117	250	166,449
Диастолическое давление	integer	78	138	102,714
Пульс	integer	78	129	107,653
Креатинин	integer	34	148	98,347
СКФ	integer	40	125	87,898
Холестерин	real	3,6	6,5	4,757
СОЭ	integer	2	52	17,633
Глюкоза	real	4,1	6,7	5,282

Вышеперечисленные входные данные загружаются в репозиторий. В результате загрузки задается тип численных (real, integer) и качественных (polinominal) данных, представленных на рис.2:

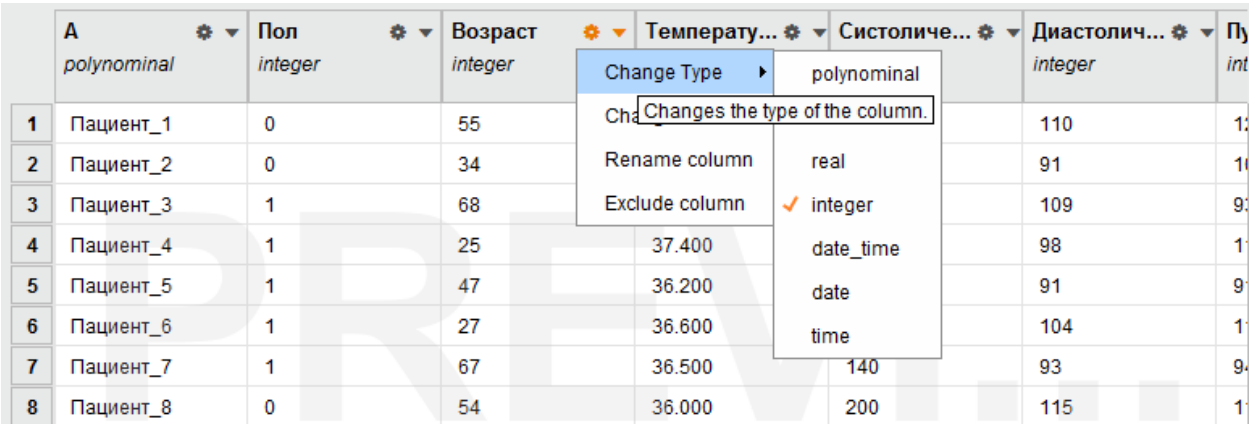

Рис.2. Пример ввода типа данных

В качестве атрибутов (label) выбираются: заболевание АГ 1 степени, заболевание АГ 1 степени на фоне почечной недостаточности, заболевание АГ 1 степени на фоне сердечной недостаточности, заболевание АГ 2 степени, заболевание АГ 2 степени на фоне почечной недостаточности, заболевание АГ 2 степени на фоне сердечной недостаточности, заболевание АГ 3 степени, заболевание АГ 3 степени на фоне почечной недостаточности, заболевание АГ 3 степени на фоне сердечной недостаточности.

3.2. ТЕХНОЛОГИЯ РЕШЕНИЯ ЗАДАЧИКЛАСТЕРНОГО АНАЛИЗА

Цель кластеризации: во-первых, кластеризация стремится разделить элементы данных на ряд групп, так что элементы в одной группе больше похожи на другие элементы в той же группе; во-вторых, он направлен на то, чтобы предметы в одной группе отличались от предметов в другой группе.

a) K-medoids

Компьютерная модель кластерного анализа с использованием k-medoids представлена на рис.3:

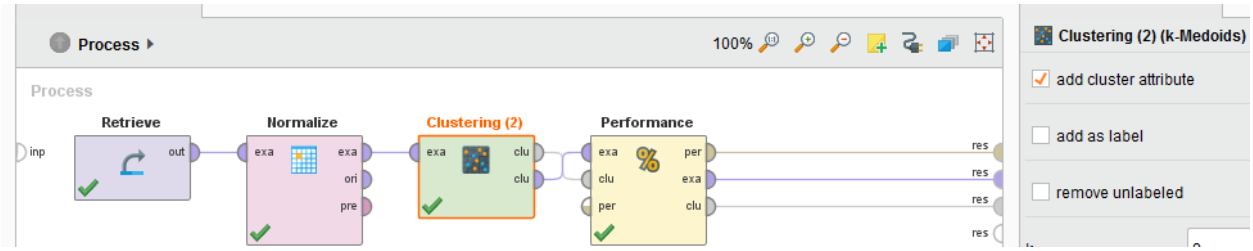

Рис.3. Компьютерная модель кластерного анализа с использованием k-medoids

Схемасостоитиз 4 операторов: Retrive, Normalize, Clustering и Performance.

Оператор Retrieve может получить доступ к хранимой информации в хранилище и загрузить ее в процесс. В Retrive загружаются в репозиотрий исходные данные из программы MSExcel.

ОператорNormalize нормализует значения выбранных атрибутов. Нормализация используется для масштабирования значений, чтобы они соответствовали определенному диапазону. Регулировка диапазона значений очень важна при работе с атрибутами разных единиц и шкал. Нормализация полезна для сравнения атрибутов, которые различаются по размеру.

Оператор Clustering выполняет кластеризацию, используя алгоритм k-medoids. Кластеризация связана с группированием объектов, которые похожи друг на друга и не похожи на объекты, принадлежащие другим кластерам. Кластеризация k-medoids является эксклюзивным алгоритмом кластеризации, то есть каждый объект назначается точно одному из набора кластеров.

Оператор Performance используется для оценки производительности. Предоставляет список значений критериев эффективности. Эти критерии эффективности определяются автоматически, чтобы соответствовать типу задачи обучения.

Результат

На рис.4 представлена модель кластера:

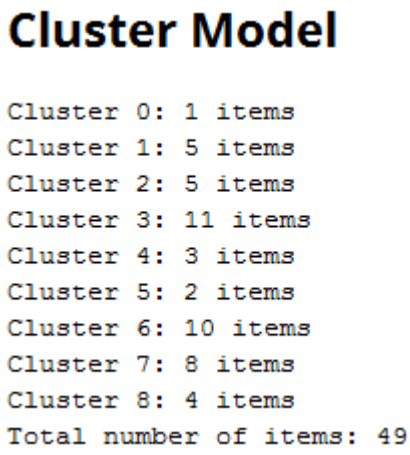

Рис.4. Модель кластера с использованием алгоритма k-medoids

Диагнозы, разделенные на кластеры алгоритмом k-medoids, показаны в таблице 2:

Таблица 2 – Состав кластеров с использованием алгоритма k-medoids

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
АГ 3 степени на фоне сердечной недостаточности	АГ 1 степени	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени на фоне почечной недостаточности	АГ 1 степени
	АГ 2 степени на фоне почечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне почечной недостаточности	АГ 2 степени
	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени	АГ 2 степени на фоне почечной недостаточности	АГ 2 степени
	АГ 2 степени	АГ 3 степени	АГ 3 степени на фоне почечной недостаточности	
	АГ 3 степени	АГ 3 степени	АГ 2 степени на фоне почечной недостаточности	
			АГ 1 степени на фоне почечной недостаточнос	

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
			ти	
			АГ 1 степени на фоне почечной недостаточности	
			АГ 2 степени на фоне почечной недостаточности	
			АГ 2 степени на фоне почечной недостаточности	
			АГ 1 степени на фоне почечной недостаточности	
			АГ 1 степени на фоне почечной недостаточности	

Продолжение таблицы 2

Cluster 5	Cluster 6	Cluster 7	Cluster 8
АГ 2 степени на фоне сердечной недостаточности	АГ 2 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 1 степени
АГ 1 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 2 степени
	АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 2 степени
	АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 2 степени на фоне почечной недостаточности
	АГ 1 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	
	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	
	АГ 1 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	
	АГ 1 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	
	АГ 1 степени на фоне почечной недостаточности		
	АГ 3 степени на		

Cluster 5	Cluster 6	Cluster 7	Cluster 8
	фоне почечной недостаточности		

Программой RapidMiner в сформированный cluster_0 (1 пациент) отобраны 1 пациент с диагнозом АГ 3 степени на фоне сердечной недостаточности. Cluster_1 (5 пациентов) состоит из 1 пациента с диагнозом АГ 1 степени, 1 пациента с АГ 2 степени на фоне почечной недостаточности, 1 пациента с АГ 3 степени на фоне почечной недостаточности, 1 пациента с АГ 2 степени и 1 пациента с АГ 3 степени. Cluster_2 (5 пациентов) состоит из 2 пациентов с диагнозом АГ 3 степени, 1 пациента с АГ 2 степени и 2 пациента с АГ 3 степени на фоне почечной недостаточности. Cluster_3 (11 пациентов) состоит из 5 пациентов с диагнозом АГ 2 степени на фоне почечной недостаточности, 1 пациента с АГ 3 степени на фоне почечной недостаточности и 5 пациентов с АГ 1 степени на фоне почечной недостаточности. Cluster_4 (3 пациентов) состоит из 2 пациентов с диагнозом АГ 2 степени и 1 пациента с АГ 1 степени. Cluster_5 (2 пациент) состоит из 1 пациента с диагнозом АГ 2 степени на фоне сердечной недостаточности и 1 пациента с диагнозом АГ 1 степени на фоне сердечной недостаточности. Cluster_6 (10 пациентов) состоит из 1 пациента АГ 2 степени на фоне почечной недостаточности, 4 пациентов с АГ 2 степени на фоне почечной недостаточности и 5 пациентов АГ 1 степени на фоне почечной недостаточности. Cluster_7 (8 пациентов) состоит из 3 пациентов с диагнозом АГ 2 степени на фоне сердечной недостаточности и 5 пациентов с АГ 1 степени на фоне сердечной недостаточности. И cluster_8 (4 пациента) состоит из 2 пациентов с диагнозом АГ 2 степени, 1 пациента

с АГ 2 степени на фоне почечной недостаточности и 1 пациента с АГ 1 степени.

График отличий кластеров, построенный в программе RapidMiner представлен на рис.5.

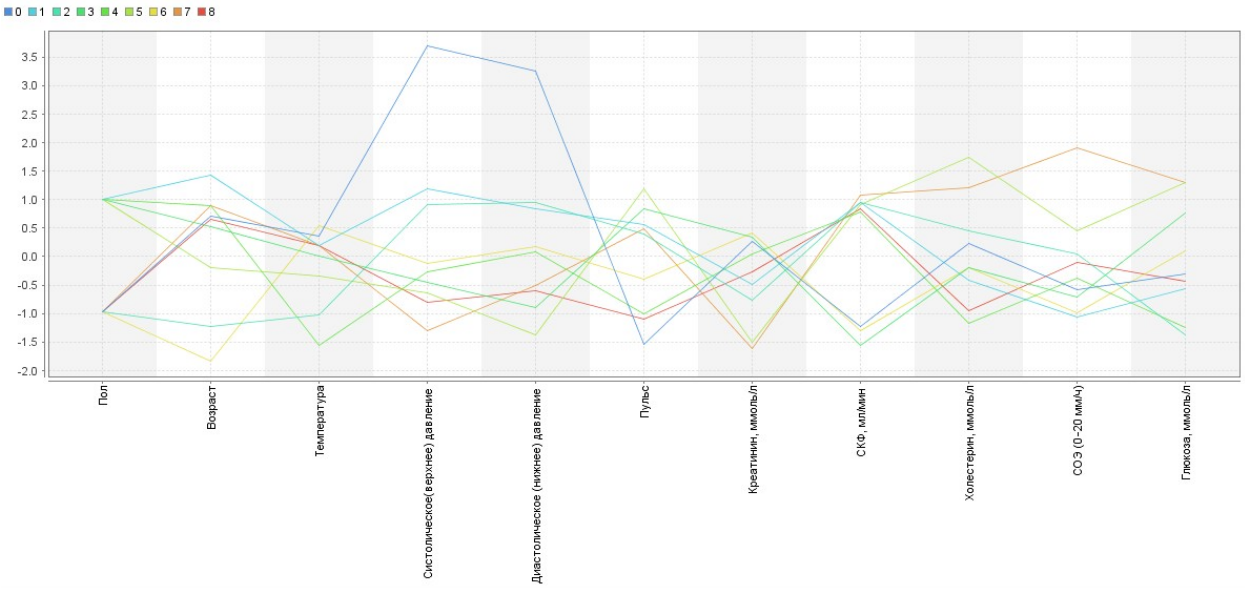

Рис.5. График отличий кластеров с использованием алгоритма x-means

Этот график показывает, что для пациентов, которые входят в cluster_0, характерно повышены систолическое и диастолическое давления; в cluster_1 пониженное значение СОЭ и зрелый возраст; пациент cluster_2 отличается пониженнойглюкозой; cluster_3 отличается пониженным СКФ; cluster_4 отличается пониженнымитемпературой и холестерином; в cluster_5 отличительным значением будет пониженное диастолическое давление, повышенные пульс и холестерин; cluster_6 отличается повышеннымкреатинином и ранним возрастом; в cluster_7 повышенныеСКФ и СОЭ, пониженные систолическое давление и креатинин.

После результатов, которые мы получили, нужно определить количество совпадений объектов кластерного анализа с использованием алгоритмak-medoids. Cluster_0

содержит 100% совпаденийобъектов; cluster_1 – 14%; cluster _2 – 75%; cluster _3 – 62%; cluster_4 – 33%; cluster _5 – 25%; cluster _6 – 50%; cluster _7 – 83%; cluster _8 – 33% совпадений. Существует эмпирическое правило – устойчивая группировка должна сохраняться при изменении методов кластеризации: к примеру, в случае если итоги кластерного анализа имеют долю совпадений больше 70% с группировкой по методу k-medoids, то предположение об устойчивости принимается. Количество совпадений объектов кластерного анализа в программной среде RapidMiner в общем случае составляет 41,66%, что считается признаком плохой кластеризации.

b) K-means

Компьютерная модель кластерного анализа с использованием k-meansпредставлена на рис.6:

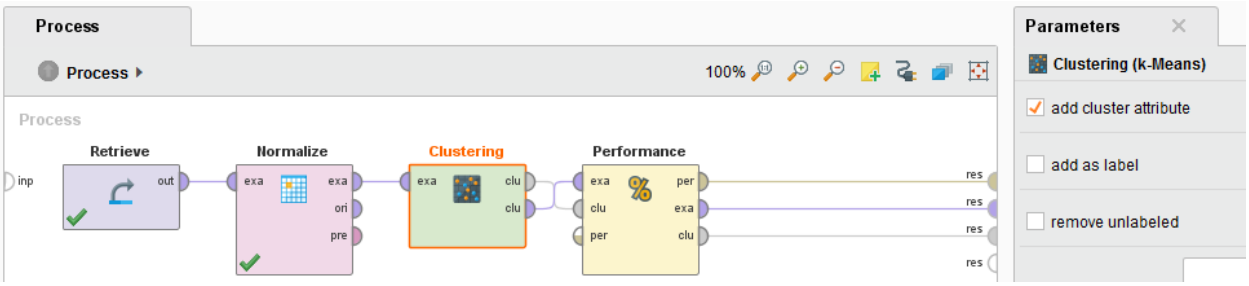

Рис.6. Компьютерная модель кластерного анализа с использованием k-means

Оператор Clustering выполняет кластеризацию, используя алгоритм k-means.

Результат:

На рис.7 представлена модель кластера:

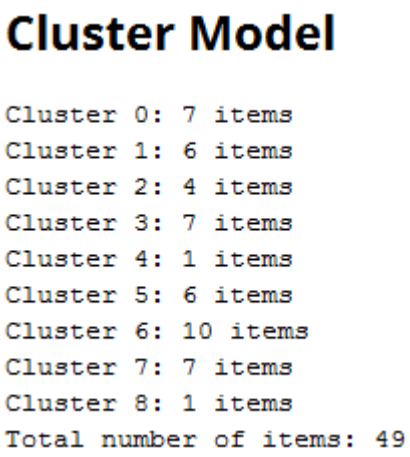

Рис.7. Модель кластера с использованием алгоритма k-means

Диагнозы, разделенные на кластеры алгоритмом k-means, показаны в таблице 3:

Таблица 3 – Состав кластеров с использованиемалгоритмаk-means

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
АГ 1 степени	АГ 2 степени на фоне почечной недостаточности	АГ 2 степени	АГ 1 степени	АГ 3 степени на фоне сердечной недостаточности
АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне почечной недостаточности	АГ 3 степени	АГ 2 степени	
АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне почечной недостаточности	АГ 3 степени	АГ 2 степени	
АГ 2 степени на фоне почечной недостаточности	АГ 2 степени на фоне почечной недостаточности	АГ 3 степени	АГ 2 степени	
АГ 2 степени	АГ 1 степени на фоне почечной недостаточности		АГ 1 степени	
АГ 2 степени	АГ 1 степени на фоне почечной недостаточности		АГ 1 степени на фоне почечной недостаточности	

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
АГ 3 степени			АГ 2 степени на фоне почечной недостаточности	

Продолжение таблицы3

Cluster 5	Cluster 6	Cluster 7	Cluster 8
АГ 3 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	АГ 3 степени на фоне почечной недостаточности
АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	
АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 2 степени на фоне почечной недостаточности	
АГ 3 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	
АГ 3 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	
АГ 3 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 2 степени на фоне почечной недостаточности	
	АГ 2 степени на фоне сердечной недостаточности	АГ 1 степени на фоне почечной недостаточности	
	АГ 1 степени на фоне сердечной недостаточности		
	АГ 1 степени на фоне сердечной недостаточности		
	АГ 1 степени на фоне сердечной недостаточности		

Программой RapidMiner в сформированный cluster_0 (7 пациентов) отобраны 3 пациента с диагнозом АГ 2 степени на фоне почечной недостаточности, 2 пациента с АГ 2 степени, 1 пациента с АГ 3 степени и 1 пациент АГ 1 степени. Cluster_1 (6 пациентов) состоит из 2 пациентов с диагнозом АГ 2 степени на фоне почечной недостаточности и 4 пациентов с АГ 1 степени на фоне почечной недостаточности.

Cluster_2 (4 пациента) состоит из 3 пациентовс диагнозом АГ 3 степени и 1 пациента с АГ 2 степени. Cluster_3 (7 пациентов) состоит из 3 пациентов с диагнозом АГ 2 степени, 2 пациентов с АГ 1 степени на фоне почечной недостаточности, 1 пациента с АГ 1 степени на фоне почечной недостаточности и 1 пациента с АГ 2 степени на фоне почечной недостаточности. Cluster_4 (1 пациента) состоит из 1 пациента с диагнозом АГ 3 степени на фоне сердечной недостаточности. Cluster_5 (6 пациентов) состоит из 6 пациентов с диагнозом АГ 3 степени на фоне почечной недостаточности. Cluster_6 (10 пациентов) состоит из 6 пациентов АГ 1 степени на фоне сердечной недостаточности и 4 пациента АГ 2 степени на фоне сердечной недостаточности. Cluster_7 (7 пациентов) состоит из 5 пациентов с диагнозом АГ 1 степени на фоне почечной недостаточности и 2 пациентов с АГ2 степени на фоне почечной недостаточности. И cluster_8 (1 пациент) состоит из 1 пациента с диагнозом АГ 3 степени на фоне почечной недостаточности.

График отличий кластеров, построенный в программе RapidMiner представлен на рис.8.

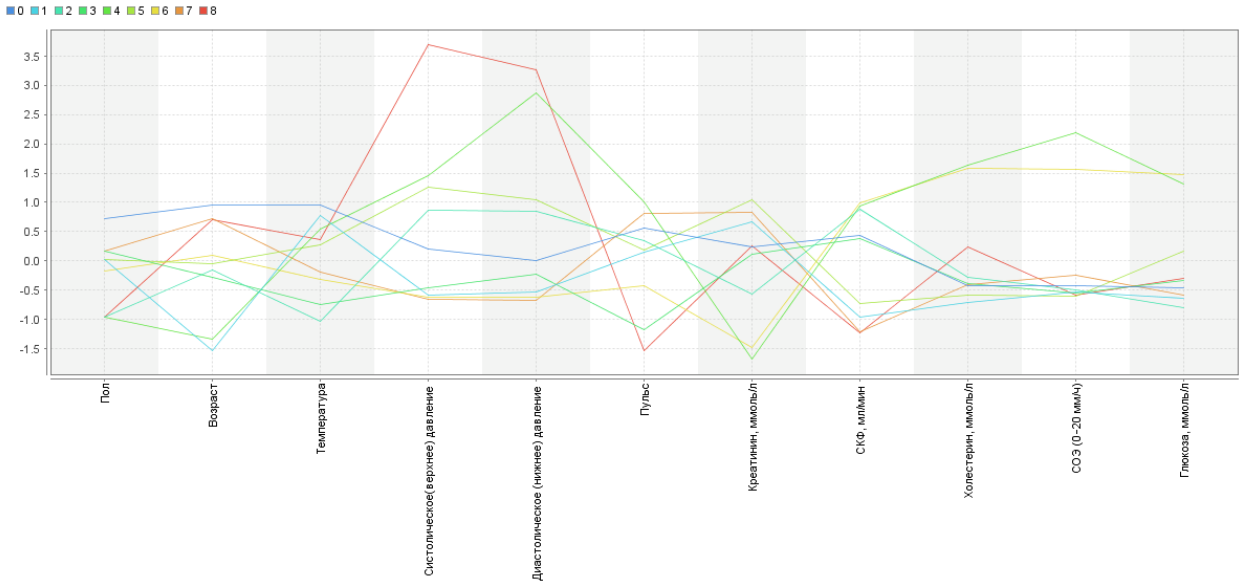

Рис.8. График отличий кластеров с использованием алгоритма x-means

Этот график показывает, что для пациентов, которые входят в cluster_0, характерно повышенная температура и зрелый возраст; в cluster_1 ранним возрастом; пациент cluster_2 отличается повышеннойтемпературой и пониженной глюкозой; cluster_3 отличается пониженным СОЭ; cluster_4 отличается повышеннымипульсом, СОЭ и низким креатинином; в cluster_5 отличительным значением будет повышенный креатинин; cluster_6 отличается повышенной глюкозой; в cluster_7 пониженнымидиастолическим давлением и систолическим давлением; cluster_8 отличается повышенными систолическим давлением и диастолическим давлением.

После результатов, которые мы получили, нужно определить количество совпадений объектов кластерного анализа с использованием алгоритмak-means. Cluster_0 содержит37,5% совпаденийобъектов; cluster_1 – 0%; cluster _2 – 75%; cluster _3 – 50%; cluster_4 – 100%; cluster _5 – 85%; cluster _6 – 100%; cluster _7 – 50%; cluster _8 – 0% совпадений.

Существует эмпирическое правило – устойчивая группировка должна сохраняться при изменении методов кластеризации: к примеру, в случае если итоги кластерного анализа имеют долю совпадений больше 70% с группировкой по методу k-means, то предположение об устойчивости принимается. Количество совпадений объектов кластерного анализа в программной среде RapidMiner в общем случае составляет 55,28%, что считается признаком плохой кластеризации.

c) X-means

Компьютерная модель кластерного анализа с использованием x-means представлена на рис.9:

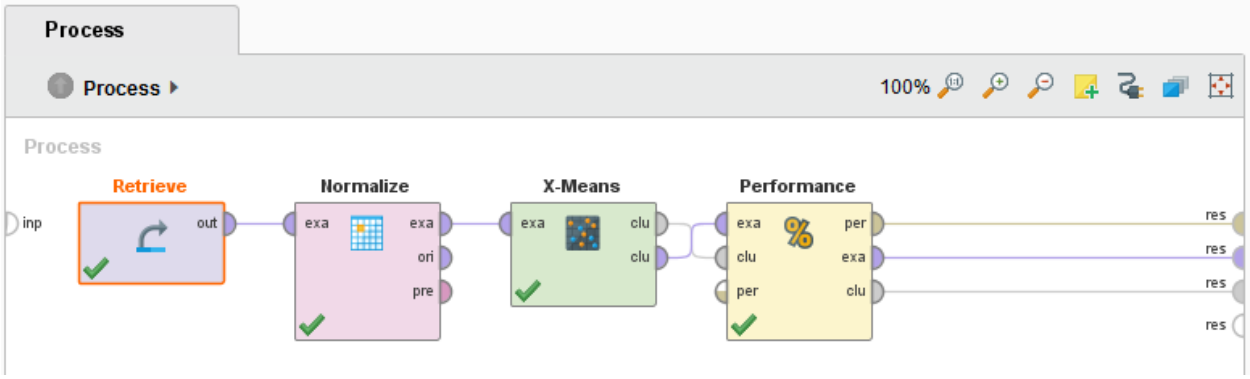

Рис.9.Компьютерная модель кластерного анализа с использованием x-means

ОператорX-Means реализует алгоритм кластеризации с использованием x-means, опубликованный Дэном Пеллегом и Эндрю Муром [29].

X-Means - это алгоритм кластеризации, который определяет правильное количество центроидов на основе эвристики. Он начинается с минимального набора центроидов, а затем итеративно эксплуатируется, если использование большего количества центроидов имеет смысл

в соответствии с данными. Если кластер разделен на два подкластера, определяется байесовским информационным критерием, который компенсирует компромисс между точностью и сложностью модели.

Результат:

На рис.10 представлена модель кластера:

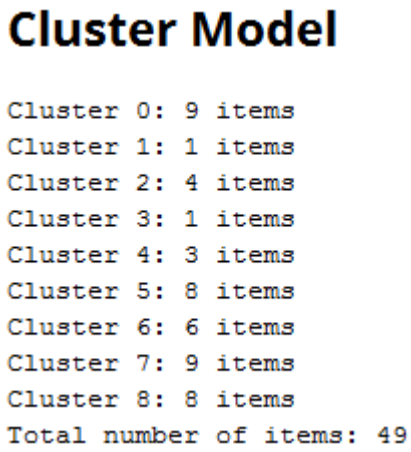

Рис.10. Модель кластера с использованием алгоритма x-means

Диагнозы, разделенные на кластеры алгоритмом x-means, показаны в таблице 4:

Таблица 4 – *Состав кластеров с использованиемалгоритмах-means*

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
АГ 1 степени на фоне почечной недостаточности	АГ 3 степени на фоне сердечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 1 степени	АГ 3 степени
АГ 1 степени на фоне почечной недостаточности		АГ 2 степени на фоне сердечной недостаточности		АГ 3 степени
АГ 1 степени на фоне почечной недостаточности		АГ 2 степени на фоне сердечной недостаточности		АГ 3 степени
АГ 1 степени на фоне		АГ 2 степени на фоне		

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
почечной недостаточности		сердечной недостаточности		
АГ 1 степени на фоне почечной недостаточности				
АГ 1 степени на фоне почечной недостаточности				
АГ 1 степени на фоне почечной недостаточности				
АГ 1 степени на фоне почечной недостаточности				
АГ2 степени на фоне почечной недостаточности				

Продолжение таблицы 4

Cluster 5	Cluster 6	Cluster 7	Cluster 8
АГ 1 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне почечной недостаточности
АГ 3 степени	АГ 1 степени на фоне сердечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени
АГ 2 степени на фоне почечной недостаточности	АГ 2 степени на фоне сердечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 1 степени
АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени
АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени
АГ 2 степени на фоне почечной недостаточности	АГ 1 степени на фоне сердечной недостаточности	АГ 3 степени на фоне почечной недостаточности	АГ 2 степени
АГ 2 степени на фоне почечной недостаточности		АГ 3 степени на фоне почечной недостаточности	АГ 2 степени
АГ 1 степени		АГ 2 степени на фоне почечной недостаточности	АГ 2 степени
		АГ 3 степени на	

Cluster 5	Cluster 6	Cluster 7	Cluster 8
		фоне почечной недостаточности	

Программой RapidMiner в сформированный cluster_0 (9 пациентов) отобраны 8 пациентов с диагнозом АГ 1 степени на фоне почечной недостаточности и 1 пациент АГ 2 степени на фоне почечной недостаточности. Cluster_1 (1 пациент) состоит из 1 пациента с диагнозом АГ 3 степени на фоне сердечной недостаточности. Cluster_2 (4 пациента) состоит из 4 пациентов с диагнозом АГ 2 степени на фоне сердечной недостаточности. Cluster_3 (1 пациент) состоит из 1 пациента с диагнозом АГ 1 степени. Cluster_4 (3 пациента) состоит из 3 пациентов с диагнозом АГ 3 степени. Cluster_5 (8 пациент) состоит из 1 пациента с диагнозом АГ 1 степени на фоне почечной недостаточности, 1 пациента с диагнозом АГ 3 степени, 1 пациента с диагнозом АГ 1 степени и 5 пациентов с диагнозом АГ 2 степени на фоне почечной недостаточности. Cluster_6 (6 пациентов) состоит из 5 пациентов АГ 1 степени на фоне сердечной недостаточности и 1 пациента АГ 2 степени на фоне сердечной недостаточности. Cluster_7 (9пациентов) состоит из 7 пациентов с диагнозом АГ 3 степени на фоне почечной недостаточности и 2 пациентов с АГ2 степени на фоне почечной недостаточности. И cluster_8 (8 пациентов) состоит из 6 пациентов с диагнозом АГ 2 степени, 1 пациента с АГ 1 степени на фоне почечной недостаточности и 1 пациента с АГ 1 степени.

График отличий кластеров, построенный в программе RapidMiner представлен на рис.11.

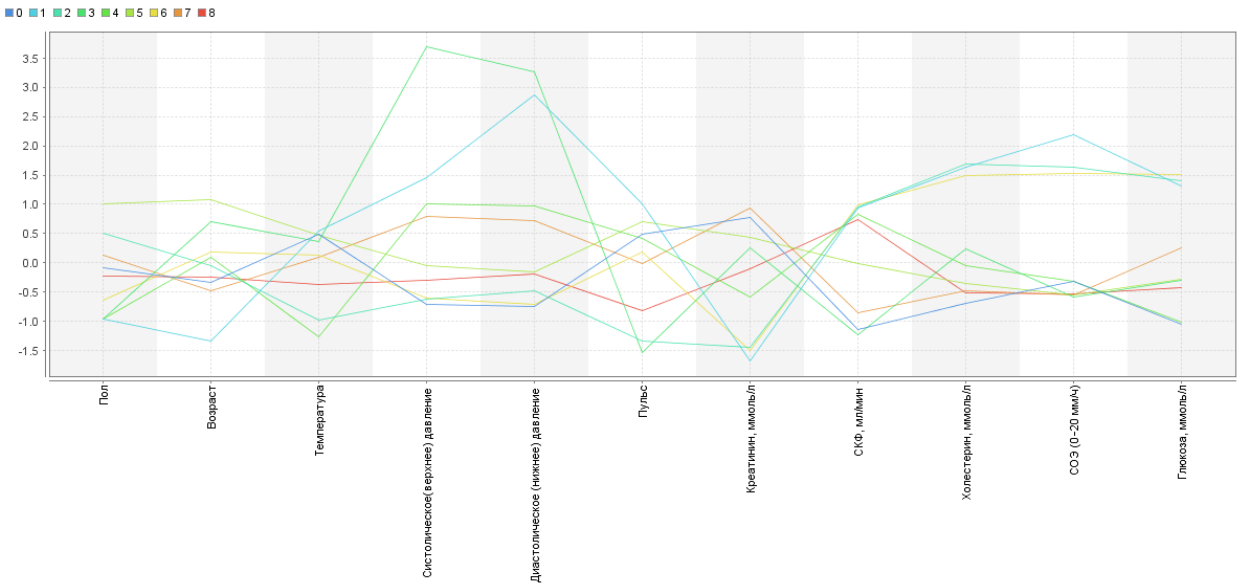

Рис.11. График отличий кластеров с использованием алгоритма x-means

Этот график показывает, что для пациентов, которые входят в cluster_0, характерно низкие значениясистолического давления, диастолическое давление и холестерина; в cluster_1 высокие значения СОЭ, пульс, температура, низкий креатинин и юный возраст; пациент cluster_2 отличается повышенным холестерином; cluster_3 отличается от всего остального низкой температурой; cluster_4 отличается повышенными систолическим давлением и диастолическим давлением; в cluster_5 отличительным значением будет зрелый возраст; cluster_6 отличается повышенной глюкозой; в cluster_7 повышенный креатинин; cluster_8 отличается пониженной СОЭ.

После результатов,которые мы получили, нужно определить количество совпадений объектов кластерного анализа с использованием алгоритма x-means. Cluster_0 содержит 80% совпаденийобъектов; cluster_1 – 100%; cluster _2 – 100%; cluster _3 – 33%; cluster_4 – 75%; cluster _5 – 62%; cluster _6 – 83%; cluster _7 – 100%; cluster _8 – 100%

совпадений. Существует эмпирическое правило – устойчивая группировка должна сохраняться при изменении методов кластеризации: к примеру, в случае если итоги кластерного анализа имеют долю совпадений больше 70% с группировкой по методу k средних, то предположение об устойчивости принимается. Количество совпадений объектов кластерного анализа в программной среде RapidMiner в общем случае составляет 72,55%, что считается признаком качественной кластеризации.

d) Сравнение алгоритмов

В таблицах 5 и 6 сравнение трех кластеров с использованием алгоритмов k-means, k-medoids и x-means. Сравнение сделано с точки зрения среднего расстояние между кластерами, генерируемыми каждым алгоритмом.

Таблица 5 – Сравнение алгоритмов

Кластер	k-means	k-medoids	x-means
cluster_0	-3.321	-0.000	-5.278
cluster_1	-3.632	-2.502	-0.000
cluster_2	-2.280	-4.845	-3.173
cluster_3	-4.723	-5.194	-0.000
cluster_4	-0.000	-0.833	-1.907
cluster_5	-4.064	-2.093	-4.147
cluster_6	-4.249	-5.141	-2.935
cluster_7	-4.633	-3.387	-5.074
cluster_8	-0.000	-1.551	-4.395

Таблица 6 – Сравнение алгоритмов с точки зрения среднего расстояния в кластере

	k-means	k-medoids	x-means
Avg.	-3.807	-3.720	-4.031

withincentroiddistance			

Расстояние между каждым кластером можно наблюдать из результатов в таблице 5 и 6, которые показывают результаты от каждого алгоритма кластеризации k-means, k-medoids и x-means в терминах среднего расстояния между кластерами, генерируемыми ими.Также вспомним совпадение объектов у алгоритмов k-means – 55,28%, k-medoids – 41,66% и x-means – 72,55%, которые мы получили. У k-means и k-medoids очень маленькая разница между кластерами в соответствии с средним расстоянием.

Таким образом, можно сделать вывод, что алгоритмx-means показал хорошие результаты по сравнению с другими двумя алгоритмами, он кластеризовал данные с -4.031 средним расстоянием в каждом кластере и совпадение объектов кластера больше 70%.

3.3. ТЕХНОЛОГИЯ ПОСТРОЕНИЯДЕРЕВА РЕШЕНИЙ

Цель состоит в том, чтобы создать модель классификации, которая прогнозирует значение метки на основе нескольких входных атрибутов. Каждый внутренний узел дерева соответствует одному из входных атрибутов. Количество ребер внутреннего узла равно количеству возможных значений соответствующего входного атрибута. Каждый листовой узел представляет значение метки с учетом значений входных атрибутов, представленных путем от корня до листа.

Основная компьютерная модель классификации дерева решений и анализа данных представлена на рис. 12.

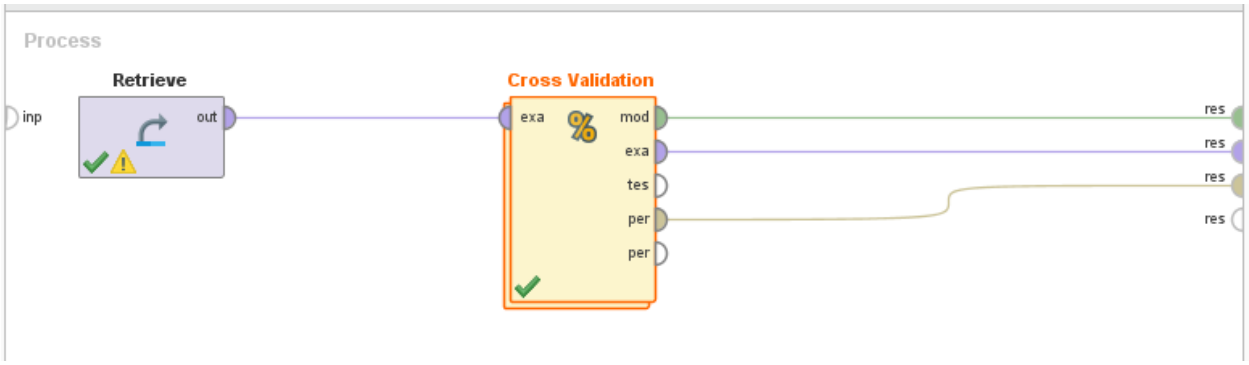

Рис.12. Основная компьютерная модель дерево решений

Схема состоит из двух операторов: Retrieve и CrossValidation. В операторе CrossValidation создана другая схема, разделенная на два подпроцесса: в обучающем подпроцессе используется оператор DecisionTree, а в подпроцессе тестирования операторы ApplyModel и Performance.

Оператор Retrieve может получить доступ к хранимой информации в хранилище и загрузить ее в процесс. В Retrive загружаются в репозиотрий исходные данные из программы MSExcel.

Оператор CrossValidation выполняет простую перекрестную проверку, то есть случайным образом разбивает ExampleSet (пример набора) на два подпроцесса: обучающий подпроцесс и подпроцесс тестирования. Этот оператор выполняет разделенную проверку для оценки производительности оператора обучения (обычно для невидимых наборов данных). Он в основном используется для оценки того, насколько точно модель (усвоенная конкретным оператором обучения) будет работать на практике.

Вложенные подпроцессы оператора CrossValidation представлены на рис.13:

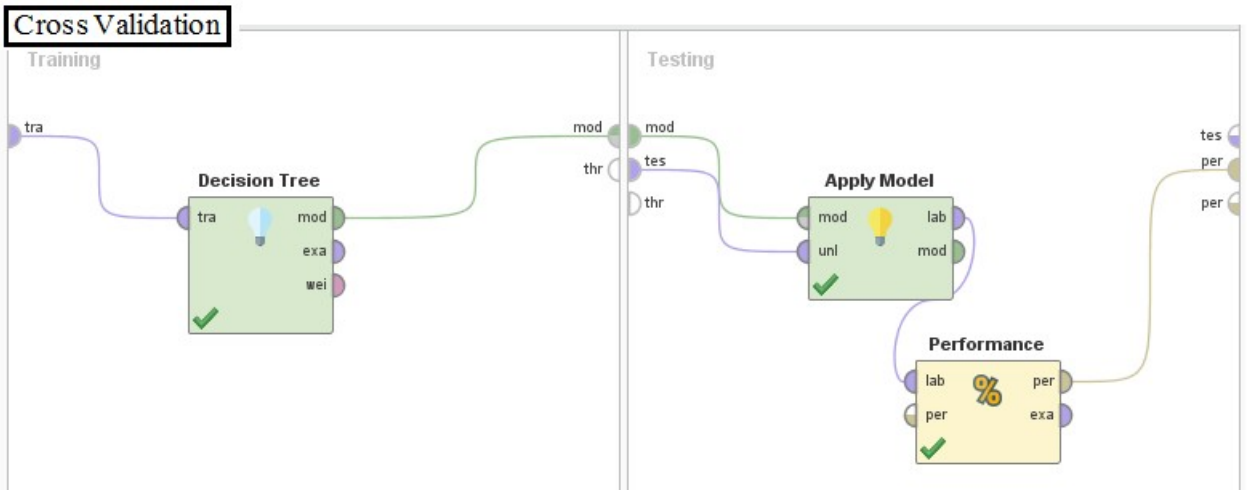

Рис.13. Компьютерная модель процесса оператора CrossValidation

Оператор DecisionTree создает модель дерева решений, которую можно использовать для классификации и регрессии. Дерево решений – это древовидная совокупность узлов, предназначенная для принятия решения о принадлежности значений к классу или оценке числового целевого значения. Каждый узел представляет правило разделения для одного конкретного атрибута. Для классификации это правило разделяет значения, принадлежащие разным классам, для регрессии оно разделяет их, чтобы уменьшить ошибку оптимальным способом для выбранного критерия параметра.

Критерии может иметь одно из следующих параметров:

- information_gain: вычисляются энтропии всех атрибутов, а для разделения используется один с наименьшей энтропией. Этот метод имеет тенденцию к выбору атрибутов с большим количеством значений.
- gain_ratio: вариант получения информации, который регулирует усиление информации для каждого атрибута, чтобы обеспечить широту и однородность значений атрибута.

- gini_index: мера неравенства между распределениями характеристик метки. Индекс Джини, задает при необходимости меру добавления, создает разветвления дерева по бинарному разделению.
- accuracy: атрибут выбран для разделения, что максимизирует точность всего дерева.
- less_square: для разделения выбран атрибут, который минимизирует квадратное расстояние между средними значениями в узле по отношению к истинному значению.

Оператор дерева решений принимает только полиномиальные, числовые и биномиальные атрибуты, а также биномиальные и полиномиальные метки (целевые атрибуты).

Оператор ApplyModel делает построение модели дерева решений.

Оператор Performance используется для оценки производительности. Предоставляет список значений критериев эффективности. Эти критерии эффективности определяются автоматически, чтобы соответствовать типу задачи обучения.

Результат.

В результате построения дерева решений с использование пакета RapidMiner была получена модель, представленная на рис.14.

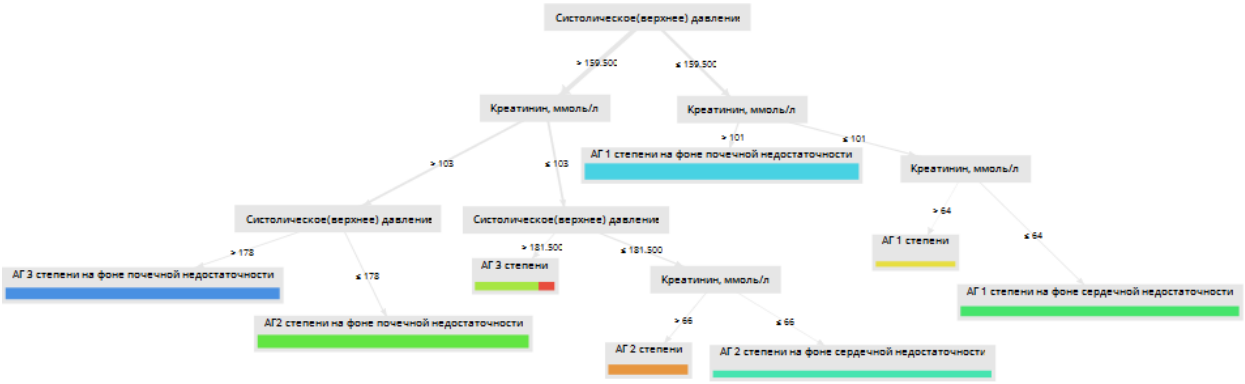

Рис. 14. Дерево решений в виде графа

В текстовом представлении можно увидеть сводку по дереву, а также конфиденциальность ветвей:

Tree

```
Систолическое(верхнее) давление  > 159.500
|   Креатинин, оль/л > 103
|   |   Систолическое(верхнее) давление  > 178: АГ 3 степени на фоне почечной
недостаточности {АГ 3 степени на фоне почечной недостаточности=7, АГ 1
степени на фоне почечной недостаточности=0, АГ 2 степени на фоне сердечной
недостаточности=0, АГ 1 степени на фоне сердечной недостаточности=0, АГ2
степени на фоне почечной недостаточности=0, АГ 3 степени=0, АГ 1 степени=0,
АГ 2 степени=0, АГ 3 степени на фоне сердечной недостаточности=0}
|   |   Систолическое(верхнее) давление  ≤ 178: АГ2 степени на фоне почечной
недостаточности {АГ 3 степени на фоне почечной недостаточности=0, АГ 1
степени на фоне почечной недостаточности=0, АГ 2 степени на фоне сердечной
недостаточности=0, АГ 1 степени на фоне сердечной недостаточности=0, АГ2
степени на фоне почечной недостаточности=8, АГ 3 степени=0, АГ 1 степени=0,
АГ 2 степени=0, АГ 3 степени на фоне сердечной недостаточности=0}
|   Креатинин, оль/л ≤ 103
|   |   Систолическое(верхнее) давление  > 181.500: АГ 3 степени {АГ 3
степени на фоне почечной недостаточности=0, АГ 1 степени на фоне почечной
недостаточности=0, АГ 2 степени на фоне сердечной недостаточности=0, АГ 1
степени на фоне сердечной недостаточности=0, АГ2 степени на фоне почечной
недостаточности=0, АГ 3 степени=4, АГ 1 степени=0, АГ 2 степени=0, АГ 3
степени на фоне сердечной недостаточности=1}
|   |   Систолическое(верхнее) давление  ≤ 181.500
|   |   |   Креатинин, оль/л > 66: АГ 2 степени {АГ 3 степени на фоне
почечной недостаточности=0, АГ 1 степени на фоне почечной недостаточности=0,
АГ 2 степени на фоне сердечной недостаточности=0, АГ 1 степени на фоне
сердечной недостаточности=0, АГ2 степени на фоне почечной недостаточности=0,
АГ 3 степени=0, АГ 1 степени=0, АГ 2 степени=6, АГ 3 степени на фоне
сердечной недостаточности=0}
|   |   |   Креатинин, оль/л ≤ 66: АГ 2 степени на фоне сердечной
недостаточности {АГ 3 степени на фоне почечной недостаточности=0, АГ 1
степени на фоне почечной недостаточности=0, АГ 2 степени на фоне сердечной
недостаточности=4, АГ 1 степени на фоне сердечной недостаточности=0, АГ2
степени на фоне почечной недостаточности=0, АГ 3 степени=0, АГ 1 степени=0,
АГ 2 степени=0, АГ 3 степени на фоне сердечной недостаточности=0}
Систолическое(верхнее) давление  ≤ 159.500
|   Креатинин, оль/л > 101: АГ 1 степени на фоне почечной недостаточности
{АГ 3 степени на фоне почечной недостаточности=0, АГ 1 степени на фоне
```

```
почечной недостаточности=10, АГ 2 степени на фоне сердечной
недостаточности=0, АГ 1 степени на фоне сердечной недостаточности=0, АГ2
степени на фоне почечной недостаточности=0, АГ 3 степени=0, АГ 1 степени=0,
АГ 2 степени=0, АГ 3 степени на фоне сердечной недостаточности=0}
|   Креатинин, оль/л ≤ 101
|   |   Креатинин, оль/л > 64: АГ 1 степени {АГ 3 степени на фоне почечной
недостаточности=0, АГ 1 степени на фоне почечной недостаточности=0, АГ 2
степени на фоне сердечной недостаточности=0, АГ 1 степени на фоне сердечной
недостаточности=0, АГ2 степени на фоне почечной недостаточности=0, АГ 3
степени=0, АГ 1 степени=3, АГ 2 степени=0, АГ 3 степени на фоне сердечной
недостаточности=0}
|   |   Креатинин, оль/л ≤ 64: АГ 1 степени на фоне сердечной
недостаточности {АГ 3 степени на фоне почечной недостаточности=0, АГ 1
степени на фоне почечной недостаточности=0, АГ 2 степени на фоне сердечной
недостаточности=0, АГ 1 степени на фоне сердечной недостаточности=6, АГ2
степени на фоне почечной недостаточности=0, АГ 3 степени=0, АГ 1 степени=0,
АГ 2 степени=0, АГ 3 степени на фоне сердечной недостаточности=0}
```

Примеры построенных деревьев решений с различными критериями приведены в табл. 7.

Таблица 7– точность деревев решений

Критерий построения	Точность построения	Количество листьев	Количество ветвей	Количество узлов	Корень дерева
gain_ratio	90.00%	8	16	8	Креатинин
information_gain	88.00%	8	16	8	Креатинин
gini_index	90.00%	8	14	7	Систолическое давление
accuracy	88.00%	8	16	7	Систолическое давление

В соответствии с построенной таблицей точности можно предположить, что оптимальным является дерево решений с критерием построения gain_index. Второй столбец точности в таблице 2 показывает измерения. Для этого атрибута все полученные результаты имеют высокую точность, и обученная модель может быть использована для прогнозов.

На рис. 15 показана точность модели для дерева решений и точность маркировки данных составляет 90% с критерием gain_index:

accuracy: 90.00% +/- 14.14% (micro average: 89.80%)

	true АГ 3 ст...	true АГ 1 ст...	true АГ 2 ст...	true АГ 1 ст...	true АГ2 сте...	true АГ 3 ст...	true АГ 1 ст...	true АГ 2 ст...	true АГ 3 ст...	class precis...
pred. АГ 3 с...	6	0	0	0	1	0	0	0	0	85.71%
pred. АГ 1 с...	0	10	0	0	0	0	0	0	0	100.00%
pred. АГ 2 с...	0	0	3	0	0	0	0	1	0	75.00%
pred. АГ 1 с...	0	0	1	6	0	0	0	0	0	85.71%
pred. АГ2 ст...	1	0	0	0	7	0	0	0	0	87.50%
pred. АГ 3 с...	0	0	0	0	0	4	0	0	1	80.00%
pred. АГ 1 с...	0	0	0	0	0	0	3	0	0	100.00%
pred. АГ 2 с...	0	0	0	0	0	0	0	5	0	100.00%
pred. АГ 3 с...	0	0	0	0	0	0	0	0	0	0.00%
class recall	85.71%	100.00%	75.00%	100.00%	87.50%	100.00%	100.00%	83.33%	0.00%	

Рис.15. Оценка точности моделис критериемgain_index

Точность – это проверка на соответствие исходных данных, а также проверка на внутреннюю достоверность на основе исходных данных.

Таким образом, с точностью 90% разделены входные данные на 9 выборок: АГ 3 степени на фоне почечной недостаточности (6 пациентов), АГ 1 степени на фоне почечной недостаточности (10 пациентов), АГ 2 степени на фоне сердечной недостаточности (3 пациента), АГ 1 степени на фоне сердечной недостаточности (6 пациентов), АГ 2 степени на фоне почечной недостаточности (7 пациентов), АГ 3 степени (4 пациента), АГ 1 степени (3 пациента), АГ 2 степени (5 пациентов), АГ 3 степени на фоне сердечной недостаточности (0 пациентов).

Верно распознает АГ 3 степени на фоне почечной недостаточности в 85,71%, АГ 1 степени на фоне почечной недостаточности в 100%, АГ 2 степени на фоне сердечной недостаточности в 75%, АГ 1 степени на фоне сердечной недостаточности в 100%, АГ 2 степени на фоне почечной недостаточности в 87,5%, АГ 3 степени в 100%, АГ 1 степени в 100%, АГ 2 степени в 83,33%, АГ 3 степени на фоне сердечной недостаточности в 0%. Верно предсказывает АГ 3

степени на фоне почечной недостаточности в 85,71%, АГ 1 степени на фоне почечной недостаточности в 100%, АГ 2 степени на фоне сердечной недостаточности в 75%, АГ 1 степени на фоне сердечной недостаточности в 85,71%, АГ 2 степени на фоне почечной недостаточности в 87,5%, АГ 3 степени в 80%, АГ 1 степени в 100%, АГ 2 степени в 100%, АГ 3 степени на фоне сердечной недостаточности в 0%.

В основе рис. 15 создадим таблицу 8 ошибок.

После проведенного анализа (рис. 15), видно, что диагноз гипертонии прогнозируются у6+10+3+6+7+4+3+5+0=44 пациентов, не точно поставлена степень и выраженность диагноза у 1+1+1+1+1=5пациентов.Прогнозируемый АГ 2 степени на фоне почечной недостаточности есть у 1 пациента, а дерево решений распознает его как АГ 3 степени на фоне почечной недостаточности, но и распознанный АГ 2 на фоне почечной недостаточности есть у 1 пациента с прогнозируемым АГ 3 степени на фоне почечной недостаточности. Прогнозируемый АГ 1 степени на фоне сердечной недостаточности есть у 1 пациента с распознанным АГ 2 степени на фоне сердечной недостаточности. Распознанный АГ 2 степени есть у 1 пациента с прогнозируемым АГ 2 степени на фоне сердечной недостаточности. Также распознанный АГ 3 степени на фоне сердечной недостаточности есть у 1 пациента с прогнозируемым АГ 3 степени. Из результатов можно сделать вывод, что мы получаем 1+1+1=3 пациента имеют предполагаемый диагноз, но не распознаются методом классификации дерево решений.

Таблица 8 – Таблица диагностики сопряженности дерева решений с параметром gain_index

Параметр	Предлагаемый диапазон		Итог
	есть	нет	
Дерево решений – есть	44	0	44
Дерево решений – нет	2	3	5
Итог	46	3	49

3.4. ВЫВОД К ТРЕТЬЕЙ ГЛАВЕ

По результатам третьей главы можно сделать следующие выводы:

1. В исследованиях кластерного анализа были использованы k-mean, k-medoids и X-meansалгоритмы кластеризации с использованием инструмента RapidMiner. Эти три алгоритмы были применены для оценки производительности каждого алгоритма с точки зрения кластеризации медицинских данных, результаты сравнивались с каждым алгоритмом, сравнение показала, что лучше использовать с точки зрения кластеризации алгоритм x-means.

Совпадение объектов у алгоритмов k-means – 55,28%, k-medoids – 41,66% и x-means – 72,55%, которые мы получили. У k-means и k-medoids очень маленькая разница между кластерами в соответствии с средним расстоянием.

Таким образом, можно сделать вывод, что алгоритм x-means показал хорошие результаты по сравнению с другими двумя алгоритмами, он кластеризовал данные с -4.031

средним расстоянием в каждом кластере и совпадение объектов кластера больше 70%.

2. В исследовании построения дерева решений мы получаем оптимальное дерево решений с критерием gain_index, где точность построения дерева 90% . Проведя анализ, видно, что диагноз гипертонии прогнозируется у 44 пациентов, не точно поставлена степень у 5 пациентов из них 3 пациента имеют предполагаемый диагноз, но не распознаются методом классификации дерево решений.

3. Благодаря этим исследованиям, мы выяснили, что:

Полученные результаты кластерного анализа позволяют сформировать решение, обобщить собранные данные и разработать соответствующие рекомендации. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы.

Метод дерева решений является мощным статистическим инструментом для классификации, прогнозирования, интерпретации и обработки данных, который имеет несколько потенциальных применений в медицинских исследованиях. Использование моделей дерева решений для описания результатов исследований имеет следующие преимущества:

• Упрощает сложные отношения между входными переменными и целевыми переменными, разделяя исходные входные переменные на значимые подгруппы.

• Легко понять и интерпретировать.
• Устойчив к выбросам.

ЗАКЛЮЧЕНИЕ

С быстрым увеличением численности населения, существует значительное количество роста заболеваний, связанных со здоровьем. Некоторые заболевания тесно связаны с симптомами, которые создают врачам сложность прогнозировать точные заболевания сразу. Вот где появляется техника интеллектуального анализа данных, которая помогает в прогнозировании заболевания. Это исследование сосредоточено на исследованиинесколькихметодах и их алгоритмов.

В результате проделанной работы проведено исследованиенаиболее известных, и получивших широкое распространение, статистических программных пакетов.Рассмотрены пакеты программного обеспечения для статистического анализа данных, такие как: MSExcel, STATISTICA, SPSS Statistics, SAS VisualAnalytics, Stata и RapidMinerStudio.

В результате проведенного анализа имеющихся на настоящий момент программных средств, для решения задач прикладной математической статистики для целей медицины можно сделать следующие выводы:

1. Предлагаемая линейка программных пакетов решает практически весь спектр задач прикладной статистики, возникающих в практической медицинской деятельности.

2. Программный пакет RAPIDMINER,несомненно, является наиболее эффективным средством решения задачоперативной медицинской практики – диагностики и прогнозирования.

Рассмотрены задачи классификации интеллектуального анализа. Эти задачи позволяют построить модели, которые могут использоваться для прогнозирования поведения анализируемой системы в ситуации, которая ранее не наблюдалась; использовать поиск скрытых закономерностей в данных, их описание и вывод правил, которые могут быть использованы в будущем для повышения эффективности работы медицинских учреждений.

Более подробно рассмотрены и описаны кластерный анализ и построение дерева решений.

В ходе работы выполнены поставленные задачи, а именно:

- исследование возможности применения статистического метода ***кластерного анализа***.В исследованиях кластерного анализа были использованы k-means, k-medoids и x-means алгоритмы с использованием инструмента RapidMiner. Эти три алгоритма были применены

для оценки производительности каждого алгоритма с точки зрения кластеризации медицинских данных. Результаты сравнивались с каждым алгоритмом с точки зрения среднего значения и совпадения объектов. В результате исследования мы получили, что алгоритм x-means показал лучший результат по сравнению с другими двумя алгоритмами, он кластеризовал данные с -4.031 средним расстоянием в каждом кластере и совпадение объектов кластера больше 70%, что считается признаком хорошей кластеризации.

- исследование возможности применения алгоритма построения ***дерева*** решений. Это исследование является попыткой использовать функции инструмента RapidMiner для анализа данных и представить некоторые из возможностей, которые предлагаются для анализа данных и было создано для того, чтобы представить деревья решений как один из инструментов современных возможностей машинного обучения (интеллектуального анализа данных). Следуя из таблицы точности, мы получаем оптимальное дерево решений с критерием gain_index, где точность построения дерева 90%. Проведя анализ, видно, что диагноз гипертонии прогнозируется у 44 пациентов, не точно поставлена степень у 5 пациентов из них 3 пациента имеют предполагаемый диагноз, но не распознаются методом классификации дерево решений.

На основании проведенного исследования можно заключить, что на настоящий момент RapidMinerявляется наиболее эффективным продуктом в решении задач оперативной медицинской практики.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

Литература

1. Дресвянский Д.В. О применении методов интеллектуального анализа данных в задаче обнаружения спама / Д.В. Дресвянский, Е.С. Семёнкин // Электронный сборник материалов международной конференции студентов, аспирантов и молодых ученых «Проспект свободный – 2015» посвященной 70-летию великой победы. -2015.-С.22-24.

2. Корсунова Е.С. Применение пакета STATISTICA и MS EXCEL для обработки биомедицинской информации/ Е.С. Корсунова, К.Д. Тишакова // Технические науки: тенденции, перспективы и технологии развития. / Сборник научных трудов по итогам международной научно-практической конференции. № 4. г. Волгоград, 2017. -64 с.

3. Крыштановский А.О.Анализ социологических данных с помощью пакета SPSS [Текст]: учеб, пособие для вузов / А. О. Крыштановский- Москва: ВШЭ, 2006.-с. 225—281.

4. Мальцева А.А. Классификация регионов по уровню инновационного развития на основе кластеризации / А.А. Мальцева, А.Л. Баскакова // Вестник Тверского государственного университета. Серия: Экономика и управление. 2015.- № 4. -С. 167-176.

5. Мамонтов Д.Ю. Применение методов интеллектуального анализа данных для задачи классификации повреждений стальных пластин / Д.Ю. Мамонтов, Е.С. Семёнкин //Электронный сборник материалов

международной конференции студентов, аспирантов и молодых ученых «Проспект свободный – 2015» посвященной 70-летию великой победы. 2015.-С.36-38.

6. Неслухов Д.С. Использование кластерного и регрессионного анализа в изучении экономической деятельности судостроительных и судоремонтных предприятий / Д.С. Неслухов // Интернет-журнал «НАУКОВЕДЕНИЕ», 2016-Том 8, №4.

7. Овсянников А.О. Анализ внутренних затрат на научные исследования и разработки по субъектам Российской Федерации при помощи кластерного анализа RapidMiner / А.О. Овсянников // Научно-практический электронный журнал Аллея Науки. 2018.- №6(22).

8. Орестова В.Р., Применение статистического пакета анализа данных SPSS Statistics в психологических исследованиях на примере факторного анализа / В. Р. Орестова, А. А. Бастрон // История и архивы. – 2017. – №2.– С. 38-51.

9. Пичугин Ю.А. О классификации летних режимов погоды в Санкт-Петербурге / Ю.А. Пичугин // Метеорология и гидрология. 2000.- № 5. -С. 31–39.

10. Сиделев, С. И. Математические методы в биологии и экологии: введение в элементарную биометрию: учебное пособие / С. И. Сиделев; Яросл. гос. ун-т им. П. Г. Демидова. – Ярославль: ЯрГУ, 2012. – 140 с.

11. Сылова С.Д. Создание групп для маркетинговых целей из данных использования веб-сайта / С.Д. Сылова // Вестник Удмуртского университета. Математика. Механика. Компьютерные науки- 2017.- Т. 27, вып. 3.- С. 470-478.

12. Третьяков А.С. Статистические методы в прикладных географических исследованиях: Учебно-методическое пособие / А.С. Третьяков; науч. ред. проф. И.Г. Черванев – Х.: Шрифт, 2004. – 96 с.

13. Трухачева Н. В. Математическая статистика в медико-биологических исследованиях с применением пакета Statistica. / Н.В. Трухачева. - М.: ГЭОТАР-Медиа, 2012.-384 с.

14. Ульянов Е.А. Кластеризация паевых инвестиционных фондов по прибыльности/ Е.А. Ульянов, Д.Ш. Бесаев // Научно-практический электронный журнал Аллея Науки. - 2018.-№6.

15. Филандышева Л. Б. Статистические методы в географии: учебно-методическое пособие / Л. Б. Филандышева, Е. С. Сапьян; отв. ред. А.В. Пучкин; Том. гос. ун-т. - Томск : Издательский Дом Томского государственного университета, 2015. – 164 с.

16. Чернышова Г.Ю. Применение методов интеллектуального анализа данных для кластеризации текстовых документов. / Г.Ю. Чернышова, А.Н. Овчинников // Информационная безопасность регионов: научно-практический журнал. 2015. -№4 (21). – С.5-12.

17. Шеламова, М. А. Использование программы Excel в работе с базой медико-биологических данных: учеб.-метод. пособие / М. А. Шеламова // Минск : БГМУ, 2011. –С.56.

18. Шубат О. М. Кластерный анализ в исследовании социально-экономических процессов: опыт критического анализа / О. М. Шубат, А. П. Караева // Проблемы моделирования социальных процессов: Россия и страны АТР : материалы Второй всероссийской научно-практической

конференции с международным участием — Владивосток : Дальневост. федерал. ун-т, 2016. — С. 325-328.

19. Эльрих Ю. Применение метода «дерево решений» в целях оптимизации ремонтных программ предприятий электроэнергетики РФ/ Ю. Эльрих, Э. Петровский // РИСК: Ресурсы, информация, снабжение, конкуренция, 2012. – № 1. – С. 385-388.

Интернет-ресурсы

20. Википедия – свободная энциклопедия [Электронный ресурс]. - https://ru.wikipedia.org/wiki/Дерево_решений . - (дата обращения: 21.03.2020).

21. RapidMiner [Электронный ресурс]. URL: https://rapidminer.com/ (дата обращения: 17.03.2020).

Иностранные источники

22. Lloyd GER. Hippocratic Writings. / J. Chadwick, N.W. Mann, Trans // London: Penguin Books, 1983.-pp. 223.

23. Exarchos TP. Mining sequential patterns for protein fold recognition. / T.P. Exarchos, C. Papaloukas, C. Lampros, D.I. Fotiadis // Biomed Inform. 2008.-pp.165–179.

24. Jones J.K. The role of data mining technology in the identification of signals of possible adverse drug reactions: value and limitations, current therapeutic research-clinical and experimental. / J.K. Jones // 2001-vol. 62, num. 9.-pp. 664-672.

25. Kakimoto M. Data Mining from Functional Brain Image / M. Kakimoto, C. Morita, H. Tsukimoto // Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD Conference. Boston, 2000.-pp. 91-97.

26. Kuo W.J. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. / W.J. Kuo, R.F. Chang, D.R. Chen, C.C. Lee //Breast Cancer Res Treat, 2001. -66(1) - pp.51–57.

27. Laura J. van’t Veer. Gene expression profiling predicts clinical outcome of breast cancer / Laura J. van’t Veer, Hongyue Dai, Marc J. Van De Vijver et al. // Nature. 2002.-V. 415. № 6871. -pp. 530–536.

28. Ohno-Machado L. Decision trees and fuzzy logic: A comparison of models for the selection of measles vaccination strategies in Brazil. / L. Ohno-Machado, R. Lacson, E. Massad // Journal of the American medical informatics association: Suppl., 2000.- pp. 625-629.

29. Pelleg Dan. Accelerating Exact k-means Algorithms with Geometric Reasoning / Dan Pelleg, Andrew Moore // Carnegie Mellon University, Pittsburgh, 1999 - pp. 277-281.

ПРИЛОЖЕНИЕ А

Таблица А.1

	Пол	Возраст	Температура	Систолическое (верхнее) давление	Диастолическое (нижнее) давление	Пульс	Креатинин, ммоль/л	СКФ, мл/мин	Холестерин, ммоль/л	СОЭ, мм/ч	Глюкоза, ммоль/л	Диагноз
Пациент_1	0	55	37	180	110	120	140	50	4	9	5,2	АГ 3 степени на фоне почечной недостаточности
Пациент_2	0	34	37,5	158	91	103	144	43	5,1	5	5	АГ 1 степени на фоне почечной недостаточности
Пациент_3	1	68	36,6	161	109	93	49	119	5,9	44	6,3	АГ 2 степени на фоне сердечной недостаточности
Пациент_4	1	25	37,4	150	98	119	115	67	3,8	20	4,5	АГ 1 степени на фоне почечной недостаточности
Пациент_5	1	47	36,2	141	91	91	36	121	6,5	52	6,7	АГ 1 степени на фоне сердечной недостаточности
Пациент_6	1	27	36,6	164	104	117	124	70	3,9	20	5,7	АГ2 степени на фоне почечной недостаточности
Пациент_7	1	67	36,5	140	93	94	108	46	4,1	19	4,9	АГ 1 степени на фоне почечной недостаточност

												и
Пациент_8	0	54	36	200	115	115	90	111	4,6	3	5	АГ 3 степени
Пациент_9	1	63	36,8	157	95	128	94	124	4,9	12	5,1	АГ 1 степени
Пациент_10	1	38	35,8	169	106	91	133	47	5,2	11	5,9	АГ2 степени на фоне почечной недостаточности
Пациент_11	0	47	37,1	158	91	100	41	113	6,4	46	6,3	АГ 1 степени на фоне сердечной недостаточности
Пациент_12	1	64	37,4	176	102	98	134	89	4	13	4,6	АГ2 степени на фоне почечной недостаточности
Пациент_13	1	51	36	198	113	111	143	66	3,9	7	5,9	АГ 3 степени на фоне почечной недостаточности
Пациент_14	1	60	36,8	164	107	122	146	41	3,7	17	4,1	АГ2 степени на фоне почечной недостаточности
Пациент_15	0	74	36,1	155	96	116	34	113	5,9	29	6,7	АГ 1 степени на фоне сердечной недостаточности
Пациент_16	1	51	35,5	167	104	99	52	114	6,5	42	6,4	АГ 2 степени на фоне сердечной недостаточности
Пациент_17	0	62	36,8	250	138	91	107	50	5	10	5,1	АГ 3 степени на фоне почечной

												недостаточности
Пациент_18	0	60	37,4	152	92	124	117	44	3,8	12	4,2	АГ 1 степени на фоне почечной недостаточности
Пациент_19	0	31	35,9	140	99	96	136	78	5	9	5,3	АГ 1 степени на фоне почечной недостаточности
Пациент_20	0	71	35,6	184	114	112	67	113	4,4	19	4,4	АГ 3 степени
Пациент_21	0	35	36,4	179	109	110	79	119	3,9	4	5,2	АГ 2 степени
Пациент_22	1	32	35,6	147	96	97	89	118	4,3	15	4,9	АГ 1 степени
Пациент_23	1	68	37,4	168	106	116	139	84	4	3	5,1	АГ2 степени на фоне почечной недостаточности
Пациент_24	0	40	36,8	209	116	122	143	78	3,6	11	5,3	АГ 3 степени на фоне почечной недостаточности
Пациент_25	0	53	37,3	177	102	108	94	117	4,8	20	5,3	АГ 2 степени
Пациент_26	0	20	36,8	141	98	118	122	52	3,7	13	4,2	АГ 1 степени на фоне почечной недостаточности
Пациент_27	0	47	37,2	164	105	103	54	124	6,1	46	6,5	АГ 2 степени на фоне сердечной недостаточност

												и
Пациент_28	0	57	35,9	147	94	129	142	66	4,3	20	5	АГ 1 степени на фоне почечной недостаточности
Пациент_29	0	61	36,4	196	112	104	147	76	5,2	10	5,7	АГ 3 степени на фоне почечной недостаточности
Пациент_30	1	25	36,8	191	121	95	117	43	4	15	5,5	АГ 3 степени на фоне почечной недостаточности
Пациент_31	1	80	37,1	169	103	120	111	60	4,6	16	4,8	АГ2 степени на фоне почечной недостаточности
Пациент_32	1	23	37	156	95	100	114	70	3,6	2	4,3	АГ 1 степени на фоне почечной недостаточности
Пациент_33	0	57	36,6	171	101	118	125	46	5,2	17	4,7	АГ2 степени на фоне почечной недостаточности
Пациент_34	1	64	37,5	205	118	112	128	78	4,7	6	5,1	АГ 3 степени на фоне почечной недостаточности
Пациент_35	0	40	36,5	160	100	109	41	125	6,3	51	6,6	АГ 2 степени на фоне сердечной недостаточности
Пациент_36	0	32	35,8	152	93	90	36	116	6,5	30	6,1	АГ 1 степени на фоне сердечной недостаточности

Пациент_37	1	61	37,3	170	109	119	91	117	4	19	5,1	АГ 2 степени
Пациент_38	0	28	36,9	201	134	120	35	117	6,3	50	6,3	АГ 3 степени на фоне сердечной недостаточности
Пациент_39	0	65	36,7	140	99	114	37	121	5,9	46	6,3	АГ 1 степени на фоне сердечной недостаточности
Пациент_40	0	30	36	189	114	113	69	117	5,2	19	4,3	АГ 3 степени
Пациент_41	0	20	36,9	166	106	104	113	48	4,6	4	5,4	АГ2 степени на фоне почечной недостаточности
Пациент_42	0	52	37	169	105	93	87	117	4,4	6	5,7	АГ 2 степени
Пациент_43	1	75	35,6	147	98	119	148	69	5,1	10	5,4	АГ 1 степени на фоне почечной недостаточности
Пациент_44	1	41	36,4	171	104	96	78	111	4,4	2	4,3	АГ 2 степени
Пациент_45	1	47	36,4	155	90	122	41	116	6,4	25	6,3	АГ 1 степени на фоне сердечной недостаточности
Пациент_46	1	74	36,7	195	113	115	79	117	4,4	3	4,9	АГ 3 степени

Пациент_47	1	65	35,7	163	105	97	99	112	3,7	13	4,4	АГ 2 степени
Пациент_48	1	59	36,6	159	95	118	110	40	4,6	8	5,9	АГ 1 степени на фоне почечной недостаточности
Пациент_49	0	61	36,7	151	98	96	87	114	3,9	17	5	АГ 1 степени