ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ»
МИНИСТЕРСТВА ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Медико-профилактический факультет с отделением биологии
Кафедра медицинской физики с курсом информатики

Серегин Владимир Сергеевич

ОПТИМИЗАЦИОННЫЕ МОДЕЛИ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Научный руководитель:
доктор физико-математических наук, доцент
заведующий кафедрой медицинской
физики с курсом информатики

А. А. Кудрейко

Уфа - 2023

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ»
МИНИСТЕРСТВА ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Медико-профилактический факультет с отделением биологии
Кафедра медицинской физики с курсом информатики

Серегин Владимир Сергеевич

ОПТИМИЗАЦИОННЫЕ МОДЕЛИ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Научный руководитель:
доктор физико-математических наук, доцент
заведующий кафедрой медицинской
физики с курсом информатики А. А. Кудрейко

Уфа - 2023

Оглавление

Введение

Актуальность проблемы. Современные объемы накопленных данных настолько внушительны, что человеку не по силам самостоятельно их анализировать. Необходимость выполнения такого анализа вполне очевидна, поскольку в необработанных данных заключены знания, которые могут быть востребованы при принятии решений, в том числе и врачебных. Для выполнения автоматического анализа данных, используется технология Data Mining. Структурированные и неструктурированные большие объемы данных называют большими данными (Big Data). Как известно, термин «большие данные» впервые ввёл редактор американского журнала Nature Клиффорд Линч в 2008 году в специальном выпуске на тему взрывного роста мировых объемов информации [10]. В действительности, большие данные – это не только размер данных с расширенными возможностями их обработки, это ещё и технологии анализа, визуализации прогнозирования результатов. В России термин «большие данные» вошел в употребление несколько позже. Для анализа больших данных используют разные способы обработки, обобщенные термином «анализ данных». Развитие методов анализа данных не обошло стороной и систему здравоохранения России.

В федеральном проекте «Создание единого цифрового контура в здравоохранении на основе единой государственной информационной системы здравоохранения (ЕГИСЗ)» утверждается необходимость комплексного внедрения медицинских информационных систем во всех медицинских организациях. Развитие данного направления включает в себя разработку систем поддержки принятия врачебных решений.

Сегодня в системе здравоохранения генерируются цифровые данные, которые поступают от разных источников, например, результаты лабораторных анализов, информационное взаимодействие между подразделениями организации и так далее. Большое количество цифровой медицинской информации позволяет совершенствовать здравоохранение. Широкое применение интеллектуальных систем и методов машинного

обучения можно увидеть в примерах от предварительного анализа медицинских данных до постановки диагноза и этапов лечения [31].

Методы машинного обучения используются в здравоохранении [1,3,9]. Прогнозы алгоритмов машинного обучения для здравоохранения, проверяются врачом или поставщиком медицинских услуг и применяются при поддержке принятия врачебных решений [38]. Однако процесс тестирования модели и сравнения разных алгоритмов обучения часто остается понятным лишь для специалистов по анализу данных. Необходимо внедрение понятных медицинскому персоналу инструментов анализа данных.

Исходя из обозначенного круга вопросов анализа данных системы здравоохранения, возникла следующая **цель**: оценка возможности применения алгоритмов машинного обучения для повышения эффективности принятия врачебных решений.

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

1. Поиск способов визуализации работы алгоритмов диагностики.

2. Создание диагностической модели ансамбля алгоритмов машинного обучения на основе исходных данных.

3. Тестирование комплексной диагностической модели.

4. Определение основных факторов, влияющих на работу алгоритмов при выявлении заболеваний по симптомам.

Область исследования. Выпускная квалификационная работа выполнена в рамках освоения компетенций ОПК-2 (способен творчески использовать в профессиональной деятельности знания фундаментальных и прикладных разделов дисциплин (модулей), определяющих направленность магистратуры), и ОПК-6 (способен творчески применять и модифицировать современные компьютерные технологии, работать с профессиональными базами данных, профессионально оформлять и представлять результаты новых разработок) ФГОС ВО – Магистратура по направлению подготовки

06.04.01 Биология. В процессе написания выпускной квалификационной работы использовались труды отечественных и зарубежных исследователей в области моделирования медицинских процессов на основе методов математической статистики, искусственного интеллекта и машинного обучения.

Решение поставленных задач выполнено в программе Orange Data Mining. Она является программой визуального отображения данных, машинного обучения и интеллектуального анализа данных. В программе Orange Data Mining применяется визуальное программирование, представленное в виде предопределенных или разработанных пользователем блоков (виджетов). В результате математического анализа закономерностей в большом объеме данных, решаются задачи анализа данных.

В выпускной квалификационной работе выполнен анализ данных на примере симптоматических показателей по диагностике заболеваний для оценки возможности применения алгоритмов машинного обучения для повышения эффективности принятия врачебных решений.

Объектом исследования является открытая база данных, загруженная с платформы Kaggle. Исследуемая база данных состоит из двух CSV-файлов для обучения и тестирования модели.

Файл для обучения алгоритмов содержит 4920 случаев заболеваний, а файл для тестирования содержит 41 случай заболевания.

Глава 1. Современное состояние применения интеллектуального анализа данных в медицине

1.1 Роль больших данных в медицинской деятельности

Правильную постановку диагноза осложняют динамика заболевания, интерпретация результатов исследований, организация процесса диагностики. На помощь врачу в решении указанных проблем приходят технологии машинного обучения. Одной из областей применения машинного обучения в медицине является постановка диагноза [16,39,51]. При помощи машинного обучения можно решить большое количество задач, например, задачи статистического анализа результатов исследований ЭКГ, УЗИ или МРТ. В большинстве случаев задача любой модели сводится к предсказанию заболевания, болен ли человек и чем именно. Выявление некоторых трудно диагностируемых заболеваний часто зависит от квалификации врача и стадии заболевания. Обучив модель на большом массиве данных, полученных из результатов исследований и множества форм патологий (сопутствующих тем или иным заболеваниям), можно повысить качество постановки диагнозов и количество выявленных заболеваний на ранних стадиях. Таким образом технологии машинного обучения способны повысить качество работы медучреждений, автоматизировав трудоёмкую и ответственную часть работы врачей.

Системы поддержки принятия врачебных решений можно определить, как программное обеспечение, в котором характеристики отдельного пациента используются для представления клиницисту конкретных оценок или рекомендаций по принятию решения [11].

Первая медицинская экспертная система разработана в 1976 году и предназначалась для определения режимов приема антибиотиков при тяжелых бактериальных инфекциях [48]. К сожалению, она фактически не использовалась на практике из-за отсутствия системной интеграции в клиническую работу.

Существенным ограничением экспертных систем является большой объем правил, необходимых при принятии сложных клинических решений. Машинное обучение помогает преодолеть это ограничение. В машинном обучении инженеры программируют алгоритмы, способные определять свои собственные правила на основе исходных данных. Правила, закодированные человеком вручную, заменяются искусственным поиском правил из данных. Это позволяет системам, построенным на машинном обучении извлекать значимую информацию из данных и интерпретировать неизвестные ситуации. Производительность и способность машины к обучению зависят от объема и качества предоставляемых данных. Доступность медицинских данных резко возросла благодаря электронным системам ведения медицинской документации и появлению подключенных устройств [7].

Оптимальный анализ и интерпретация больших данных, нуждается в эффективных алгоритмах и вычислительной мощности современных машин, поэтому инструменты машинного обучения становятся популярными.

1.2 Системы поддержки врачебных решений

Система поддержки принятия врачебных решений относится к компьютеризированным системам, которые оказывают помощь медицинскому персоналу или самим пациентам для улучшения здоровья, медицинских услуг и внедряются с помощью медицинских технологий, например, искусственного интеллекта, машинного обучения и интеллектуального анализа данных [6,34,49].

При применении для профилактики и ведения заболеваний поддержка принятия врачебных решений с использованием машинного обучения включает в себя интеллектуальный анализ данных и вывод для оказания помощи врачам путем облегчения индивидуального изучения профиля риска, индивидуального лечения, прогностического моделирования состояния здоровья пациента и так далее.

Общая архитектура системы поддержки принятия врачебных решений изображена на Рисунок 1. Рассмотрим данную схему более подробно:

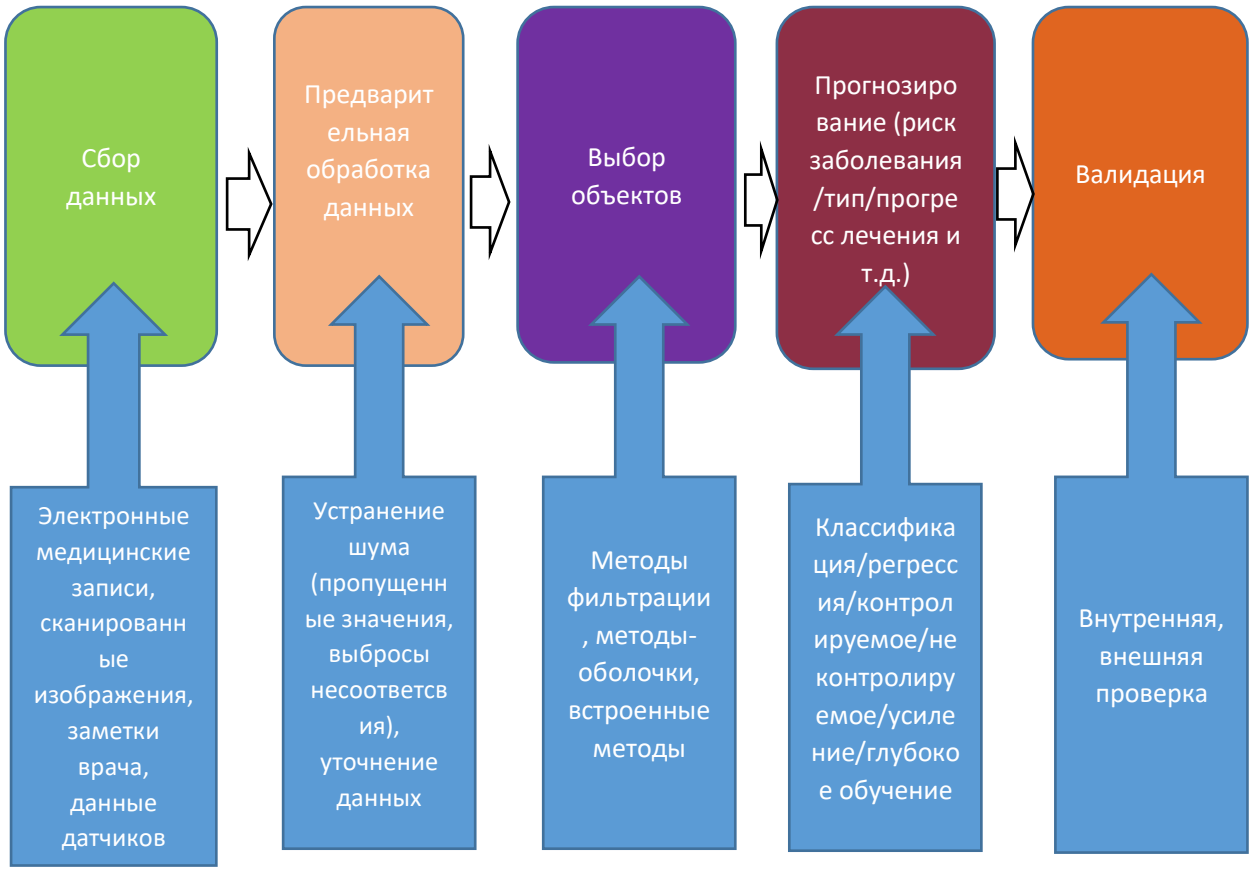


Рисунок 1. Архитектура систем поддержки принятия врачебных решений для профилактики и ведения заболеваний.

1) Сбор данных. Сбор медицинских данных из доступных источников, (базы данных больниц, общедоступные онлайн-хранилища, предоставляемые исследовательскими организациями, правительствами и так далее). Характер данных зависит от проблемы и цели исследования. Различные медицинские изображения, рецепты, данные, собранные медицинскими датчиками, широко используются в научных исследованиях.

2) Предварительная обработка данных. Для медицинских данных характерно количество недостающих значений, наличие выбросов и классовый дисбаланс. С целью устранения этих недостатков, данные необходимо предварительно обработать.

3) Выбор объектов. Объекты являются строительными блоками любого алгоритма машинного обучения. Выбор объектов – это процесс извлечения подмножества объектов, которые наилучшим образом представляют различные классы в наборе данных, а также максимизируют точность алгоритма обучения. Существует три класса методов отбора признаков [28].

а) Методы фильтрации. Извлекают релевантные признаки на основе присущих им статистических свойств данных для получения подмножества функций, имеющих большее отношение к цели [37]. Корреляция, взаимная информация, дисперсия – это показатели релевантности в методах фильтрации.

б) Методы-оболочки. Включают поиск в пространстве признаков и алгоритм классификации для выбора подмножества признаков, обеспечивающего максимальную точность классификации. В отличие от методов фильтрации, считается, что оболочки сильно зависят от алгоритма классификации.

в) Встроенные методы. Это комбинации методов фильтрации и методов-оболочек, где поиск оптимального подмножества признаков является неотъемлемой частью обучения классификатора, а не независимым шагом.

4) Методы прогнозирования. Применение этих методов осуществляется поэтапно: обучение и тестирование [19,45]. Исходный набор данных разделен на обучающий и тестовый наборы данных. При обучении, набор данных представляется в качестве входных данных алгоритма обучения, который учится обобщать или прогнозировать выходные данные с помощью распознавания образов. Тестирование включает в себя определение производительности прогнозирования алгоритма обучения по ранее неиспользуемым примерам из тестируемого набора данных. В системах поддержки врачебных решений прогностические алгоритмы машинного обучения включают в себя прогнозирование диагноза, прогнозирование риска заболевания, типа, прогресса, лечения, осложнений и так далее.

5) Валидация или проверка производительности классификатора на предмет точности и надежности его прогнозов. Часто это выполняется на внутреннем тестовом наборе данных, полученном из исходного набора данных, известного как внутренняя проверка. Когда для проверки результатов используется внешний набор данных, который не является частью входного набора данных, это называется внешней проверкой. Валидация является важным шагом, особенно в диагностических приложениях поддержки принятия врачебных решений, если они используются в клинических условиях [17].

Приложения для поддержки принятия врачебных решений на основе машинного обучения можно разделить на четыре класса: диагностика, оценка риска, прогнозирование и прогностика (см. Рисунок 2). Приложения для поддержки принятия врачебных решений не ограничиваются профилактикой и ведением заболеваний, но и направлены на улучшение медицинского обслуживания.

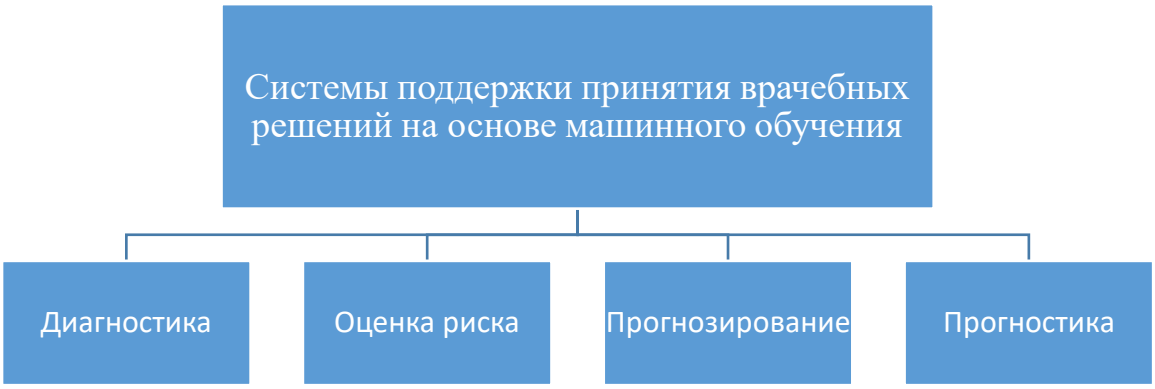


Рисунок 2. Классификация приложений машинного обучения для систем поддержки принятия врачебных решений

1) Диагностика относится к выявлению медицинского состояния или болезни на основе признаков или симптомов. Диагностика заболеваний на основе машинного обучения может быть использована для разработки инструментов массового скрининга, портативных диагностических программных пакетов на месте оказания медицинской помощи для

отдаленного сельского района, неинвазивных или минимально инвазивных систем диагностики для пожилых людей.

2) Оценка риска. Объективная оценка факторов риска, приводящих к заболеванию, и их соответствующего индивидуального вклада в прогрессирование заболевания. Прогнозирование заболеваемости, прогресса, побочных эффектов лекарств, госпитализации и так далее включает оценку риска в качестве решающего шага.

3) Прогнозирование. Прогнозирование типа заболевания, связанных с ним медицинских затрат, потенциального лечения, потребности в госпитализации являются важными проблемами моделей прогнозирования на основе машинного обучения в здравоохранении.

4) Прогностика относится к прогрессированию заболевания, отмечает возрастающую тяжесть и связанное с этим ухудшение здоровья с течением времени. Прогностическое моделирование заболевания и связанных с ним осложнений представляет собой поддержку принятия прогностических врачебных решений на основе машинного обучения.

1.3 Примеры применения машинного обучения в медицине

В последнее время методы машинного обучения всё чаще применяются для оценки риска сахарного диабета 2 типа. Авторы работы [8] воспользовались искусственной нейронной сетью, логистической регрессией (см. раздел 2.2.2), деревом решений (см. раздел 2.2.1), методом опорных векторов, случайным лесом и наивным Байесом (см. раздел 2.2.3) для прогнозирования сахарного диабета. Анализ риска был проведен с помощью одномерной и многомерной взвешенной логистической регрессии для определения связи отдельных факторов риска с исходом сахарного диабета 2 типа. Авторы определили время сна и частоту медицинских осмотров как новые факторы риска сахарного диабета 2 типа.

Авторы исследования [46] представили стратификацию риска диабета и гипертонии на основе машинного обучения для раннего выявления

субъектов высокого риска с краткосрочным интервалом в 2 месяца. Они использовали деревья решений, регуляризованную логистическую регрессию, k-ближайших соседей (kNN), случайный лес и классификаторы AdaBoost, а также сравнили результаты их классификации с пятью различными наборами функций. Полученные результаты показали, что алгоритм случайного леса оказался оптимальным классификатором, он показал улучшение площади под ROC-кривой на 35,5% для диабета в сравнении с показателями риска в США и Великобритании.

В исследовании [33] представлена работа по прогнозированию риска развития диабета 2 типа на основе машинного обучения и продемонстрирован метод автоматического объяснения риска развития диабета 2 типа. Концепция исследования основана на построении отдельных моделей прогнозирования и объяснения риска.

Коллектив авторов работы [34] воспользовался критерием Хи-квадрат и бинарной логистической регрессией для статистического анализа и прогнозирования факторов риска. Они также разработали веб-приложение для прогнозирования риска диабета второго типа для конкретного пациента в режиме реального времени с использованием дерева решений.

Возможность и эффективность оценки риска развития диабета, с использованием регулярно собираемых данных (таких как записи в электронную медицинскую карту) исследована в работе [29] путем тестирования шести моделей оценки риска. Авторы сообщают, что оценки риска могут обеспечить надежные результаты в отношении записей в электронную медицинскую карту только в том случае, если недостающие данные и несоответствия обрабатываются надлежащим образом.

Оценка генетического риска в сочетании с существующими моделями прогнозирования машинного обучения используется для оценки улучшения дискриминационной способности для прогнозирования сахарного диабета 2 типа [20] и сравнивается с традиционными моделями пропорциональных рисков Кокса. Авторы аналогичной работы [35] предложили две оценки

неинвазивного риска на основе набора данных для исследования реакции и в сочетании с моделями машинного обучения являются прогнозом сахарного диабета 2 типа.

Авторы работы [32] предложили алгоритмы наивного Байеса (см. раздел 2.2.3) и логистической регрессии (см. раздел 2.2.2) для идентификации факторов риска сахарного диабета 2 типа с использованием фенотипов и антропометрических показателей. Связь различных фенотипов, состоящих из антропометрических показателей наряду с показателями триглецеридов, была изучена с использованием бинарной логистической регрессии.

В исследовании [18] выполнена оценка роли различных метаболитов в определении будущего риска развития диабета с помощью метода масс-спектрометрии и логистического регрессионного анализа. Классификация достигнута с использованием регуляризованного моделирования методом наименьших квадратов.

Использование машинного обучения в диагностике сахарного диабета имеет свои особенности. Клинический диагноз сахарный диабет 2 типа устанавливается с помощью стандартных диагностических анализов крови FPG (уровень глюкозы в плазме крови натощак), OGTT (пероральный тест на толерантность к глюкозе) или HbA1c (гликированный гемоглобин). Все эти тесты требуют укола пациента чтобы собрать образцы крови и проверить соответствие уровня глюкозы в крови нормальным показателям. Если пациент страдает диабетом, ему требуется регулярный контроль уровня глюкозы в крови. Это означает регулярное прохождение инвазивных анализов крови. Чтобы найти альтернативу, многие исследователи сосредотачиваются на разработке неинвазивных методов диагностики диабета.

Неинвазивная диагностика диабета на основе машинного обучения была целью многих исследований, о которых недавно сообщалось в более ранних исследованиях. Некоторые исследователи сосредоточились на создании быстрых и точных диагностических инструментов для массового

скрининга без зависимости от лабораторных процедур, другие сосредоточились, на устранении тяжелого положения пациентов, испытывающих физический дискомфорт от инвазивных тестов, путем поиска альтернативных методов. Наиболее распространенные неинвазивные особенности относились к демографии пациента, физикальному обследованию, анамнезу истории болезни пациента, данным, собранным с помощью анкетирования, и так далее. В последние годы для неинвазивной диагностики сахарного диабета 2 типа были использованы ногти на ногах [12,14], особенности языка [15], и изображения радужной оболочки глаза [47]. В исследовании [12] авторы рассмотрели анализ химического состава различных элементов, обнаруженных в ногтях на ногах пациентов с диабетом, для диагностики с использованием прогностического моделирования машинного обучения.

Значительное количество исследований оценки риска развития диабета с помощью машинного обучения основано на клинических характеристиках пациентов, собранных с помощью инвазивных лабораторных процедур. Основная причина включения таких признаков заключается в том, что они предоставляют убедительные доказательства и медицинский контекст для характеристики диабета, который установлен повсеместно. Авторы исследования [4] предложили полуавтоматическую платформу машинного обучения с разработкой функций обнаружения пограничных диабетиков, пропущенных обычными экспертными алгоритмами, путем включения самооценки диабетических симптомов и осложнений наряду с клиническими особенностями. Подход к прогнозированию сердечно-сосудистых заболеваний, преддиабета и диабета с использованием исчерпывающего набора характеристик, состоящего из демографических, диетических, физикальных обследований, лабораторных показателей и анкет, представлен в работе [2]. Для классификации использовалась модель взвешенного ансамбля машинного обучения, которая включала 131 признак для сердечно-сосудистых заболеваний и 123 признака преддиабета и диабета.

В прогностическом моделировании сахарного диабета используется машинное обучение. Прогностическое моделирование заболевания включает в себя моделирование прогрессирования заболевания на различных стадиях, где последовательные стадии соответствуют более серьезным осложнениям и ухудшению здоровья. Это помогает прогнозировать риск будущих осложнений и планировать лечение таким образом, чтобы уменьшить тяжесть.

Прогностическое моделирование сахарного диабета 2 типа с использованием машинного обучения может быть классифицировано по двум категориям:

1) Вероятность развития сахарного диабета 2 типа у здоровых, с нормальной гликемией людей.

2) Прогноз распространенного сахарного диабета 2 типа, приводящего к сопутствующим осложнениям.

Моделирование прогноза развития сахарного диабета включает нормогликемию предшествующую заболеваемости в качестве предварительного условия. Нормогликемия относится к нормальному уровню глюкозы в крови, обнаруживаемому у здорового человека. Поскольку прогрессирование происходит постепенно, прогноз нормогликемии до развития диабета в будущем имеет важное значение для оценки неизбежного риска, связанного с возрастом, а также с изменением привычек образа жизни. Некоторые предшествующие состояния здоровья или патологические связи тесно связаны с возникновением диабета, например, преддиабет, метаболический синдром и гипертония. Считается, что эти заболевания сами по себе повышают риск развития диабета.

Авторы [44] разработали оценки риска для прогнозирования перехода преддиабета в диабет с использованием многомерного логистического регрессионного анализа. Они исследовали взаимосвязь разницы в ИМТ (конечный ИМТ – начальный ИМТ) в течение периода наблюдения 4,7 года с начала диабета и пришли к выводу, что пациенты с высокой разницей в ИМТ

были более подвержены риску, и потеря веса может существенно снизить риск.

Авторы [43] применили модель деревьев с градиентным усилением для моделирования прогрессирования преддиабета до диабета с использованием лассо на логистической регрессии, используемой для извлечения признаков. Была проведена как внутренняя, так и внешняя валидация результатов.

Осложнения, возникающие при диабете, имеют широкий спектр последствий для основных органов и приводят к серьезному ухудшению здоровья. Наиболее распространенными являются сердечно-сосудистые осложнения, ретинопатия, нефропатия и невропатия. Авторы исследования [13] предложили ансамбли искусственных нейронных сетей для оценки 5-летнего риска сахарного диабета 2 типа и связанных с ним сердечно-сосудистых осложнений в результате. Авторы [36] предсказали диабетические ретинопатию, невропатию и нефропатию у пациентов с сахарным диабетом 2 типа в разные временные интервалы с использованием логистической регрессии и пошагового отбора признаков. Были представлены клинически значимые графические номограммы.

1.4 Заключение к главе 1

Большинство из вышеперечисленных работ были сосредоточены на разработке систем раннего, быстрого или минимально инвазивного прогнозирования сахарного диабета 2 типа. Множество исследовательских работ оценили прогностическую способность различных классификаторов для прогнозирования сахарного диабета 2 типа с помощью сравнительного анализа, а некоторые другие проверили эффективность предсказаний классификатора с различными комбинациями предварительной обработки.

Логистическая регрессия - традиционный статистический метод решения бинарных задач и многомерный анализ рассматривался во многих работах благодаря его простоте и способности моделировать взаимосвязи между зависимыми и независимыми переменными. Прогнозирование диабета

как правило основано на следующих алгоритмах: наивный Байес, k – ближайших соседей (k-NN), машины опорных векторов, кластеризация k-средних, деревья решений, нейронные сети, модели ансамблей, такие как случайный лес, повышение градиента и так далее.

Авторы процитированных работ претендуют на понимание моделей поддержки принятия врачебных решений, но никто из них не предложил понятной медицинским работникам комплексной диагностической модели, которая бы работала на основе ансамбля диагностических методов. Такой подход позволил бы сравнивать между собой результативность методов без написания кода программы.

Дальнейшее решение поставленных задач выполнено на основе открытой базы данных, которая содержит 4920 записей заболеваний, 132 симптома и 1 прогноз. В следующей главе рассмотрены алгоритмы машинного обучения, которые были применены для обучения и тестирования модели.

Глава 2. Интеллектуальный анализ данных и машинное обучение

2.1 Методы интеллектуального анализа данных

Интеллектуальный анализ данных – это область науки о данных, направленная на извлечение информации из набора данных с использованием статистических методологий. Извлеченные данные можно использовать для выполнения исследовательских, прогностических функций. Добытые данные преобразуются и могут быть отфильтрованы так, что будут содержать только полезную информацию. Интеллектуальный анализ данных опирается на различные методы и инструменты. Рассмотрим некоторые из них.

Выявление закономерностей (отслеживание патернов) – это один из основных методов интеллектуального анализа данных, он включает в себя распознавание закономерностей внутри наборов данных. Алгоритмы распознают аномалии и аберрации данных в зависимости от времени и других переменных.

Классификация – метод интеллектуального анализа данных, который обучает алгоритмы машинного обучения сортировать данные по отдельным категориям. В классификации для определения категории используются деревья решений, метод ближайшего соседа и другие статистические методы.

Ассоциация – это принципиальное соединение переменных и элементов друг с другом посредством выводов, ориентированных на данные. Ассоциация использует различные события и атрибуты, которые пропорциональны или связаны по своей природе, а затем делает вывод на основе этой информации.

Обнаружение выбросов. Метод интеллектуального анализа данных включает в себя распознавание аберраций и аномалий. Это помогает прогнозировать будущие события и способствует эффективному и действенному решению проблем.

Кластеризация – объединение нескольких точек данных в группы на основе их сходства. Кластеризация отличается от классификации тем, что не

может различать данные по определенным категориям, но может находить закономерности в их сходстве. Формирование кластера зависит от различных параметров, таких как кратчайшее расстояние, графики и плотность точек данных. Результатом работы алгоритма является группировка в кластеры, она осуществляется путем нахождения меры сходства между объектами на основе некоторой метрики.

Регрессия - метод интеллектуального анализа данных, который используется для моделирования и планирования прогнозируемых событий. Метод регрессии используется для выявления связи и взаимосвязи между более чем одной переменной, принадлежащих определенному набору данных.

Прогнозирование. Один из основных методов интеллектуального анализа данных, позволяющий прогнозировать будущие данные и события. Примером моделей прогнозирования временных рядов является: модель на цепях Маркова, регрессионные модели прогнозирования, модели экспотенциального сглаживания и так далее.

Машинное обучение – это исследование алгоритмов, которые делают машину способной к обучению без явных инструкций. Основная цель машинного обучения – изучить обучающие данные и оценить модель с помощью тестовых данных.

Способы машинного обучения:

– Обучение с учителем является наиболее распространенным способом машинного обучения. Данный способ объединяет алгоритмы и методы построения моделей на основе множества примеров, содержащих пары «известный вход – известный выход». Алгоритм обучения с учителем стремится создать модель, обнаружив взаимосвязи между функциями и выходными данными, затем делает прогнозы значений для нового набора данных. Тестовый набор данных, который не использовался для обучения, обычно используется для прогнозирования производительности алгоритма и его проверки. К числу алгоритмов

обучения с учителем относятся логистическая регрессия (см. раздел 2.2.2), дерево решений (см. раздел 2.2.1), наивный Байес (см. раздел 2.2.3.).

- Обучение без учителя – неконтролируемые алгоритмы машинного обучения. Они выводят закономерности из набора данных без ссылки на известные или помеченные результаты. В отличии от машинного обучения с учителем, неконтролируемые алгоритмы нельзя напрямую применять к регрессии или задаче классификации, потому что неизвестно, какими могут быть значения выходных данных, что делает невозможным обучение алгоритма обычным способом. Неконтролируемое обучение можно использовать для обнаружения базовой структуры данных. Например, метод k – средних.
- Обучение с частичным привлечением учителя. В этом способе обучающие данные представляют собой комбинацию как размеченных, так и не размеченных данных. Размеченных данных используется намного меньше, чем неразмеченных. Обучение с частичным привлечением учителя представляет собой промежуточный этап между алгоритмами обучения с учителем и без учителя.
- Обучение с подкреплением – это способ машинного обучения на основе обратной связи, в котором агент учится вести себя в среде, выполняя действия и наблюдая результаты действий. За каждое хорошее действие агент получает положительную обратную связь, а за каждое плохое действие агент получает отрицательную обратную связь или штраф. В обучении с подкреплением агент учится автоматически, используя обратную связь без каких-либо помеченных данных. Без размеченных данных агент должен учиться только на своем опыте.
- Глубинное обучение — это способ машинного обучения, представляющий собой нейронную сеть с тремя и более слоями. Данные нейронные сети пытаются имитировать поведение

человеческого мозга, позволяя ему обучаться на больших объемах данных.

2.2 Алгоритмы машинного обучения

2.2.1 Дерево решений

Дерево решений – это классификатор, построенный на основе совокупности правил для принятия решений. Применение метода деревьев решений в задачах классификации заключается в том, чтобы осуществлять процесс деления исходных данных на группы, пока не будут получены однородные их множества. Деревья решений являются основой для многих классических алгоритмов машинного обучения, таких как случайный лес, Bagging и Boosted Decision Trees.

Результатом работы классификатора не всегда может быть четкий ответ или решение. Алгоритм извлекает правила из набора данных и предлагает варианты специалисту для принятия самостоятельного обоснованного решения. Деревья решений имитируют человеческое мышление, поэтому специалистам обычно легко понять и интерпретировать результаты. Как следует из названия, критерий основан на понятиях теории информации, а именно – информационной энтропии данного множества

$$H = \sum_{i=1}^{n} \frac{N_i}{N} \log\left(\frac{N_i}{N}\right),$$

Формула 1. Информационная энтропия множества

где n – число классов в исходном подмножестве, N_i – число примеров i-го класса, N – общее число примеров в подмножестве. Следовательно, энтропия рассматривается как мера неоднородности подмножества по представленным в нем классам. Если классы представлены в равных долях и неопределенность классификации наибольшая, энтропия максимальна. Если все примеры в узле относятся к одному классу *N=Ni, log(1)=0*, то *H=0*.

Структурно классификатор дерево решений состоит из следующих объектов:

- **Корневой узел:** начальный узел дерева решения.
- **Разделение:** процесс разделения узла на несколько подузлов.
- **Узел принятия решения:** подузел далее разбивается на последующие подузлы.
- **Конечный узел:** подузел не делится на последующие подузлы; представляет возможные результаты.
- **Обрезка -** процесс удаления подузлов дерева решений.
- **Ветвь -** подраздел дерева решений, состоящий из нескольких узлов.

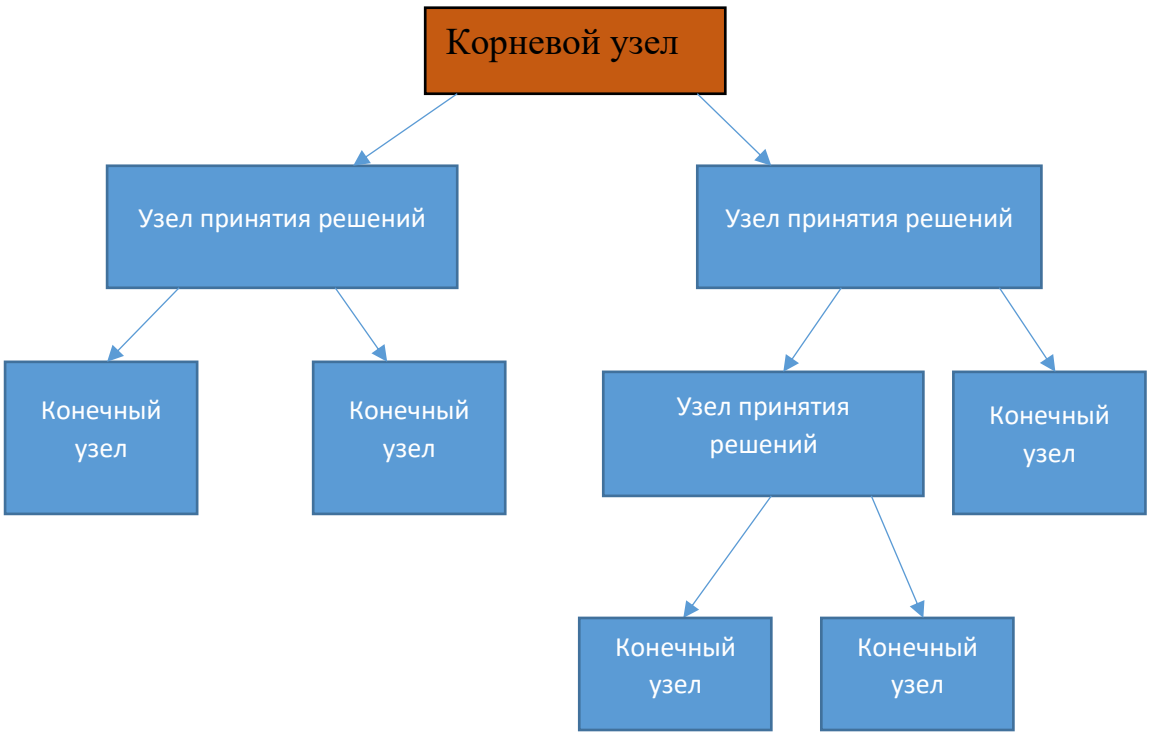


Рисунок 3. Структура дерева решений.

Дерево решений имеет сходное строение с деревом и позволяет пользователю увидеть лежащую в его основе логику (Рисунок 3). Основой дерева является корневой узел. Из корневого узла вытекает ряд узлов решений, которые изображают решения, которые необходимо принять. Из узлов решений получаются конечные узлы, представляющие последствия этих решений. Каждый узел решения представляет собой вопрос или точку разделения, а конечные узлы, вытекающие из узла решения, представляют

возможные ответы. Каждый подраздел дерева решений называется «ветвью».

Деревья решений делятся на два основных типа: категориальный и непрерывный. Подразделение основано на типе используемых целевых переменных.

В дереве решений категориальных переменных ответ четко вписывается в ту или иную категорию. В этом типе дерева решений данные помещаются в одну категорию на основе решений в узлах по всему дереву.

Дерево решений с непрерывными переменными — это дерево, в котором нет простого ответа «да» или «нет». Дерево решений с непрерывными переменными также известно, как дерево регрессии, неизвестная переменная зависит от других решений, расположенных выше по дереву, или от типа выбора, связанного с решением.

Преимущество дерева решений с непрерывными переменными заключается в том, что результат можно предсказать на основе нескольких переменных, а не на основе одной переменной, как в дереве решений с категориальными переменными. Деревья решений с непрерывными переменными используются для создания прогнозов. Систему можно использовать как для линейных, так и для нелинейных зависимостей, если выбран правильный алгоритм.

Деревья решений не предполагают независимости между атрибутами, в отличии от наивных байесовских классификаторов, что делает их применимыми во многих сценариях. Предыдущие исследования показали, что классификатор хорошо работал при решении проблем, связанных с управлением трафика, маркетингом, индустрией медицинского страхования, идентификацией генов и медицинскими диагнозами [5,23,25,26,27,40,41,52].

При принятии медицинских решений (классификация, диагностика и так далее) существует множество ситуаций, когда решение должно быть принято эффективно и надежно. Концептуальные простые модели принятия

решений с возможностью автоматического обучения являются наиболее подходящими для выполнения таких задач.

2.2.2 Логистическая регрессия

Логистическая регрессия — это алгоритм контролируемой классификации, в основе которого логистические функции для прогнозирования вероятности бинарного результата. Логическая регрессия анализирует взаимосвязь между одной или несколькими независимыми переменными и классифицирует данные по дискретным классам. Метод широко используется в прогнозном моделировании, где модель оценивает математическую вероятность того, принадлежит ли экземпляр к определенной категории или нет.

В алгоритме логистической регрессии применяется регрессионное уравнение (логит-преобразование):

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

Формула 2. Регрессионное уравнение (логит-преобразование)

где p – вероятность того, что произойдет интересующее событие; e – основание натурального логарифма; y – стандартное уравнение регрессии ($y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n$). Зависимость, связывающая вероятность события p и величину y, показана на графике Рисунок 4:

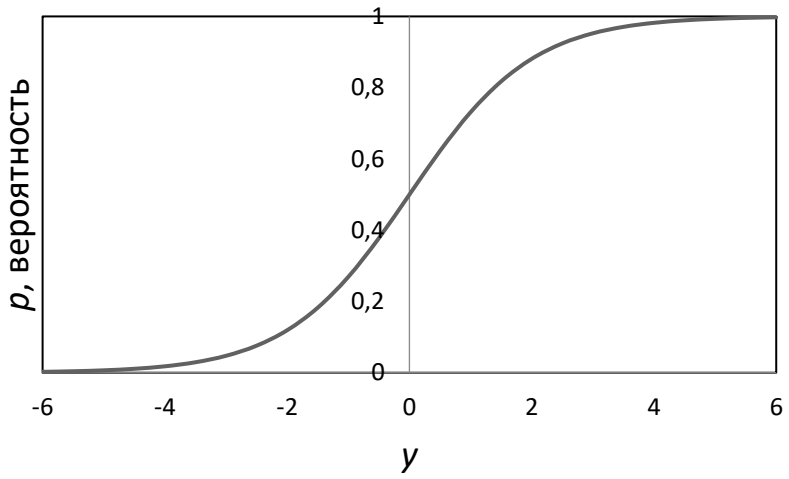


Рисунок 4. Зависимость вероятности события от величины y.

Типы логистической регрессии:

1. Бинарная. Категорический ответ имеет только два 2 возможных исхода. Пример: спам или не спам.

2. Полиномиальная. Категориальная зависимая переменная имеет два или более дискретных результата в типе полиномиальной регрессии. Это означает, что этот тип регрессии имеет более двух возможных результатов. Пример: используя переменные годовой доход и возраст предсказать вероятность того, что человек проголосует за одного из четырех кандидатов в президенты.

3. Порядковая. Порядковая логистическая регрессия применяется, когда зависимая переменная находится в упорядоченном состоянии (то есть порядковом). Зависимая переменная (y) определяет заказ с двумя или более категориями, или уровнями. Пример: используя переменные общее время показа и жанр предсказать вероятность того, что данный фильм получит рейтинг от 1 до 10.

Преимущества логистической регрессии:

1. Простота реализации методов машинного обучения. Модель машинного обучения можно настроить с помощью обучения и тестирования. Обучение выявляет закономерности во входных данных (изображение) и связывает их с некоторой формой вывода (меткой). Обучение логистической модели алгоритмом регрессии не требует больших вычислительных мощностей. Таким образом, логистическую регрессию легче реализовать, интерпретировать и обучить, чем другие методы машинного обучения.

2. Подходит для линейно разделимых наборов данных. Линейно разделимый набор данных относится к графику, на котором прямая линия разделяет два класса данных. В логистической регрессии переменная y принимает только два значения. Следовательно, можно эффективно классифицировать данные на два отдельных класса, если используются линейно разделимые данные.

3. Предоставляет ценную информацию: логистическая регрессия измеряет, насколько актуальна независимая переменная/предиктор (размер коэффициента), а также показывает направление их взаимосвязи или ассоциации (положительное или отрицательное).

Сильной стороной логистической регрессии являются хорошо зарекомендовавшие себя математические алгоритмы. Это означает, что алгоритм широко поддерживается большинством пакетов программного обеспечения статистического анализа. Более того, взаимосвязь между каждым предиктором и результатом четко определяется константами в формуле регрессии, что позволяет легко определить относительную важность каждого предиктора. А поскольку результирующая модель кратко представлена простой математической формулой, ее можно легко развернуть в конечной среде приложения.

Логистический регрессионный анализ широко используется в медицинских исследованиях. Чаще всего он используется в ситуациях, в которых возникновение бинарного результата должно быть предсказано на основе одной или нескольких прогнозирующих переменных [22,30,50].

2.2.3 Наивный Байес

Наивный Байес – это вероятностная модель машинного обучения, которая используется для задач классификации. Суть классификатора основана на теореме Байеса:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Формула 3. Теорема Байеса

где, *P(A/B*) – апостериорная вероятность (вероятность гипотезы *A* при наступлении события *B*), *P(B/A)* – вероятность вероятности (вероятность наступления события B при истинности гипотезы A), *P(A*) – априорная вероятность (вероятность гипотезы до наблюдения доказательств), *P(B)* – предельная вероятность (вероятность свидетельства)

Теорема Байеса позволяет определить вероятность события при условии, что произошло другое статистически взаимосвязанное с ним событие. Предполагается, что предикторы независимы, наличие одного конкретного признака не влияет на другой, поэтому его называют наивным.

Наивные байесовские классификаторы могут быть чрезвычайно быстрыми в сравнении с более сложными методами. Разделение распределений условных признаков класса означает, что каждое распределение может быть независимо оценено как одномерное распределение.

Виды наивных байесовских классификаторов:

1) Полиномиальный наивный байесовский классификатор – используется, когда данные распределены полиномиально.

2) Наивный байесовский классификатор Бернулли – работает аналогично полиномиальному классификатору, но переменные-предикторы являются независимыми булевыми переменными.

3) Гауссовский наивный байесовский классификатор – предполагает, что функции соответствуют нормальному распределению. Это означает, что если предикторы принимают непрерывные значения вместо дискретных, то модель предполагает, что эти значения взяты из распределения Гаусса.

Алгоритмы (классификаторы), которые описаны в предыдущих параграфах реализованы в программе Orange Data Mining. В следующем параграфе рассмотрим функционал этой программы.

2.3 Функционал программы Orange Data Mining

Программа Orange Data Mining - предназначена для визуализации данных, машинного обучения и интеллектуального анализа данных. Данная программа имеет открытый исходный код. Разработчиком данной программы является Университет Любляны. Программа предоставляет собой платформу обработки данных, систем рекомендаций и прогнозного моделирования и используется в биомедицине, биоинформатике, геномном исследовании и

обучении. В научных исследованиях программа используется в качестве платформы для тестирования новых алгоритмов машинного обучения и внедрения новых методов в медицинских исследованиях, а также в области генетики и биоинформатики [24,42].

Рабочие процессы Orange Data Mining состоят из компонентов ввода информации, обработки и визуализации. Такие компоненты называются виджетами. И они размещаются на рабочем поле. Обмен данными между виджетами выполняется по каналам связи. Выход из одного виджета используется как вход для другого. Создание рабочих процессов происходит посредством помещения виджетов на рабочее поле и соединения их от передающего виджета к принимающему виджету. Выходы виджетов находятся справа, а входы слева. Программа содержит несколько алгоритмов классификации и регрессии. Например, дерево решений, случайный лес, линейная регрессия, логистическая регрессия, наивный Байес. Orange Data Mining может читать файлы в своем ows расширении и ряда других формата данных. Классификация использует два типа объектов: обучающиеся и классификаторы. Обучающиеся рассматривают данные, помеченные классом, и возвращают классификатор. Методы регрессии в Orange очень похожи на классификацию. Оба предназначены для контролируемого интеллектуального анализа данных, им требуются данные с маркировкой класса.

В программе Orange Data Mining процесс анализа данных выполняется с помощью визуального программирования. С помощью комбинаций виджетов, можно создать структуру анализа данных. В пакет входит более 100 стандартных виджетов, с помощью которых можно решить большинство задач анализа данных в том числе и по биоинформатике. Программа позволяет разрабатывать собственные виджеты, а интерфейс сценариев можно расширять для создания автономных надстроек, интегрирующихся с остальной частью Orange Data Mining, позволяющих повторно использовать компоненты и программный код.

Рассмотрим используемые в данной работе виджеты более подробно:

I. Виджет данных

«Таблица данных» – виджет принимает один или несколько наборов данных из его входных данных (например, файлов других приложений) и представляет их в виде таблицы. Экземпляры данных могут быть отсортированы по значениям атрибутов. Виджет также поддерживает ручной выбор экземпляров данных.

II. Виджеты моделей

1. «Дерево решений» – это алгоритм, разбивающий данные на узлы по чистоте класса (энтропии). Может обрабатывать как категориальные, так и числовые наборы данных. Виджет можно использовать как для задач классификации, так и регрессии.

2. «Логистическая регрессия» - алгоритм классификации логистической регрессии с регуляризацией гребнем (L2). Работает только для задач классификации.

3. «Наивный Байес» - быстрый и простой вероятностный классификатор, основанный на теореме Байеса с предположением о независимости признаков. Работает только для задач классификации.

III. Виджеты визуализации

1. «Просмотрщик дерева» – это универсальный виджет с двухмерной визуализацией деревьев классификации и регрессии. Пользователь может выбрать узел, предписывая виджету выводить данные, связанные с этим узлом, что позволяет проводить исследовательский анализ данных.

2. «Номограмма» - номограмма обеспечивает визуальное представление классификаторов (наивного байесовского классификатора и классификатора логистической регрессии). Она дает представление о структуре обучающих данных и влиянии атрибутов на вероятности класса. Помимо визуализации классификатора, виджет предлагает интерактивную поддержку прогнозирования вероятностей классов.

IV. Виджеты оценки

1. Виджет «предсказания» получает набор данных и один или несколько предикторов (предиктивные модели, а не алгоритмы обучения), выводит данные и прогнозы. Виджет показывает вероятности и окончательные решения прогностических моделей. Результатом виджета является другой набор данных, к которому добавляются прогнозы в виде новых метаатрибутов. Результат можно наблюдать в таблице данных. Если прогнозируемые данные включают истинные значения класса, результат прогнозирования можно наблюдать в матрице путаницы.

2. Виджет «тест и оценка» - во-первых он демонстрирует таблицу с различными показателями производительности классификатора, такими как точность классификации и площадь под кривой. Во-вторых, он выводит результаты оценки, которые могут использоваться другими виджетами для анализа производительности классификаторов, таких как ROC-анализ или матрица путаницы.

Известно, что виджет «тест и оценка» поддерживает различные методы выборки:

а) Перекрестная проверка, которая разбивает данные на заданное количество кратностей. Алгоритм проверяется путем показа примеров из одной складки. Модель индуцируется из других складок тем самым классифицируются примеры из протянутой складки.

б) Перекрестная проверка по признаку, выполняет перекрестную проверку, но складки определяются выбранным категориальным признаком из метапризнаков.

в) Случайная выборка, позволяет случайным образом разбивать данные на обучающую и тестовую выборку, в заданной пропорции, процедура повторяется определенное количество раз.

г) Метод исключения одного, аналогичен случайной выборке, но он удерживает по одному экземпляру за раз, создавая модель из всех остальных, а затем классифицирует оставшиеся экземпляры. Данный метод очень

стабилен и надежен, но в тоже время очень медленный по скорости получения результатов.

д) Тест на обучающих данных, использует весь набор данных для обучения, а затем и для тестирования.

е) Тест на тестовых данных, использует только тестовые данные.

Виджет «тест и оценка» позволяет вычислять ряд статистических данных о производительности, такие как:

а) Площадь под ROC кривой.

б) Точность классификации, равная доли правильно классифицированных примеров.

в) F1 – это средневзвешенное гармоническое значение точности и полноты

г) Точность – доля истинно положительных результатов среди случаев, классифицированных как положительные.

д) Полнота как доля истинно положительных результатов среди всех положительных случаев.

е) Специфичность – это доля истинно отрицательных результатов среди всех отрицательных случаев.

ж) Кросс-энтропийная потеря учитывает неопределенность прогноза в зависимости от того, насколько он отличается от фактической метки.

з) Время обучения – время в секундах, используемое для обучения модели.

и) Время тестирования – время в секундах, используемое для тестирования модели.

Попарное сравнение моделей доступно только для перекрестной проверки. Число в таблице показывает вероятность того, что модель, соответствующая строке, лучше, чем модель, соответствующая столбцу.

3. «Матрица путаницы» показывает долю экземпляров между предсказанным и фактическим классом. Выбор элементов в матрице передает соответствующие данные в выходной сигнал. Таким образом, можно

наблюдать, какие конкретные экземпляры были неправильно классифицированы и как.

2.4 Заключение к главе 2

Рассмотренные в данной главе алгоритмы являются лишь небольшой частью доступных алгоритмов, используемых в инструментах поддержки принятия врачебных решений, но они очень распространены и служат моделями, из которых многие другие являются производными.

Имеющийся в программе Orange Data Mining инструментарий позволил визуализировать работу алгоритмов диагностики, создать диагностическую модель ансамбля алгоритмов машинного обучения на основе исходных данных, протестировать комплексную диагностическую модель и определить основные факторы, влияющие на работу алгоритмов при выявлении заболеваний по симптомам.

Глава 3. Исследование оптимизационных моделей поддержки принятия врачебных решений

3.1 Постановка задачи

Исследуемая база данных состоит из двух файлов CSV-файлов. Первый файл предназначен для обучения, а второй - для тестирования модели. Каждый файл CSV-файл содержит 133 столбца, из них 132 столбца – это симптомы, 133-й столбец – это диагноз заболевания. Все симптомы сопоставлены с 41 заболеванием.

Файл тренировки алгоритма содержит 4920 случаев заболеваний, а файл для его тестирования - 41 случай заболевания. Все представленные атрибуты являются категориальными, то есть представляют собой характеристики.

Симптомы заболеваний включают Таблица 1:

	Англоязычный термин	**Перевод**
1	itching	зуд
2	skin rash	кожная сыпь
3	nodal skin eruptions	узловые высыпания на коже
4	continuous sneezing	непрерывное чихание
5	shivering	дрожь
6	chills	озноб
7	Joint pain	боль в суставах
8	stomach pain	боль в животе
9	acidity	кислотность желудочного сока
10	ulcers on tongue	язвы на языке
11	muscle wasting	атрофия мышц
12	vomiting	рвота
13	burning micturition	жжение при мочеиспускании
14	spotting urination	кровянистые выделения при мочеиспускании
15	fatigue	усталость
16	weight gain	увеличение веса
17	anxiety	беспокойство
18	cold hands and feets	холодные руки и ноги
19	mood swings	перепады настроения
20	weight loss	потеря веса
21	restlessness	беспокойство
22	lethargy	вялость

23	patches in throat	пятна в горле
24	irregular sugar level	нерегулярный уровень сахара
25	cough	кашель
26	high fever	высокая температура
27	sunken eyes	запавшие глаза
28	breathlessness	одышка
29	sweating	потливость
30	dehydration	обезвоживание
31	indigestion	расстройство желудка
32	headache	головная боль
33	yellowish skin	желтоватая кожа
34	dark urine	темная моча
35	nausea	тошнота
36	loss of appetite	потеря аппетита
37	pain behind the eyes	боль за глазами
38	back pain	боль в спине
39	constipation	запор
40	abdominal pain	боль в животе
41	diarrhoea	диарея
42	mild fever	слабая лихорадка
43	yellow urine	желтая моча
44	yellowing of eyes	пожелтение глаз
45	acute liver failure	острая печеночная недостаточность
46	fluid overload	перегрузка жидкостью
47	swelling of stomach	вздутие живота
48	swelled lymph nodes	опухшие лимфатические узлы
49	malaise	недомогание
50	blurred and distorted vision	размытое и искаженное зрение
51	phlegm	мокрота
52	throat irritation	раздражение горла
53	redness of eyes	покраснение глаз
54	sinus pressure	давление в пазухах
55	runny nose	насморк
56	congestion	скопление, закупорка
57	chest pain	боль в груди
58	weakness in limbs	слабость в конечностях
59	fast heart rate	тахикардия
60	pain during bowel movements	боль во время дефекации
61	pain in anal region	боль в анальной области
62	bloody stool	кровавый стул
63	irritation in anus	раздражение в анусе
64	neck pain	боль в шее
65	dizziness	головокружение

66	cramps	судороги
67	bruising	кровоподтеки
68	obesity	ожирение
69	swollen legs	опухшие ноги
70	swollen blood vessels	опухшие кровеносные сосуды
71	puffy face and eyes	опухшее лицо и глаза
72	enlarged thyroid	увеличенная щитовидная железа
73	brittle nails	ломкие ногти
74	swollen extremeties	опухшие конечности
75	excessive hunger	чрезмерный голод
76	extra marital contacts	внебрачные связи
77	drying and tingling lips	сухость и покалывание губ
78	slurred speech	невнятная речь
79	knee pain	боль в колене
80	hip joint pain	боль в тазобедренном суставе
81	muscle weakness	мышечная слабость
82	stiff neck	скованность мышц шеи
83	swelling joints	опухание суставов
84	movement stiffness	скованность движений
85	spinning movements	вращательные движения
86	loss of balance	потеря равновесия
87	unsteadiness	неустойчивость
88	weakness of one body side	слабость одной стороны тела
89	loss of smell	потеря обоняния
90	bladder discomfort	дискомфорт мочевого пузыря
91	foul smell of urine	неприятный запах мочи
92	continuous feel of urine	постоянное выделение мочи
93	passage of gases	газообразование, газовыделение
94	internal itching	внутренний зуд
95	toxic look (typhos)	тифозный вид
96	depression	депрессия
97	irritability	раздражительность
98	muscle pain	боли в мышцах
99	altered sensorium	нарушение сенсорики (восприятия)
100	red spots over body	красные пятна на теле
101	belly pain	боль в животе
102	abnormal menstruation	аномальные менструации
103	dischromic patches	дисхромические пятна
104	watering from eyes	слезотечение из глаз
105	increased appetite	повышенный аппетит
106	polyuria	полиурия
107	family history	история семьи
108	mucoid sputum	мокрота, слизь

109	rusty sputum	ржавая мокрота
110	lack of concentration	нехватка концентрации
111	visual disturbances	нарушения зрения
112	receiving blood transfusion	получение переливания крови
113	receiving unsterile injections	получение нестерильных инъекций
114	coma	кома
115	stomach bleeding	желудочное кровотечение
116	distention of abdomen	вздутие живота
117	history of alcohol consumption	история употребления алкоголя
118	fluid overload	скопление жидкости
119	blood in sputum	кровь в мокроте
120	prominent veins on calf	выступающие вены на икрах
121	palpitations	учащенное сердцебиение
122	painful walking	болезненная ходьба
123	pus filled pimples	гнойные прыщи
124	blackheads	угри
125	scurring	суета
126	skin peeling	шелушение кожи
127	silver like dusting	серебристая пыль
128	small dents in nails	небольшие вмятины на ногтях
129	inflammatory nails	воспаленые ногти
130	blister	волдырь
131	red sore around nose	красная язва вокруг носа
132	yellow crust ooze	желтая корочка (гнойная)

Таблица 1. Симптомы заболеваний

На основе симптомов заболеваний ставится диагноз. В Таблица 2 перечислены диагнозы, которые поставлены на основе симптомов.

1	(vertigo) Paroymsal Positional Vertigo	Пароксизмальное позиционное головокружение
2	AIDS	СПИД
3	Acne	Акне
4	Alcoholic hepatitis	Алкогольный гепатит
5	Allergy	Аллергия
6	Arthritis	Артрит
7	Bronchial Asthma	Бронхиальная астма

8	Cervical spondylosis	Шейный спондилез
9	Chicken pox	Ветреная оспа
10	Chronic cholestasis	Хронический холестаз
11	Common Cold	Простуда
12	Dengue	Лихорадка Денге
13	Diabetes	Диабет
14	Dimorphic hemmorhoids(piles)	Геморрой
15	Drug Reaction	Лекарственная реакция
16	Fungal infection	Грибковая инфекция
17	GERD	Гастроэзофагеальная рефлюксная болезнь (ГЭРБ)
18	Gastroenteritis	Гастроэнтерит
19	Heart attack	Инфаркт миокарда
20	Hepatitis B	Гепатит В
21	Hepatitis C	Гепатит С
22	Hepatitis D	Гепатит D
23	Hepatitis E	Гепатит Е
24	Hypertension	Гипертония
25	Hyperthyroidism	Гипертиреоз
26	Hypoglycemia	Гипогликемия
27	Hypothyroidism	Гипотиреоз
28	Impetigo	Импетиго
29	Jaundice	Желтуха
30	Malaria	Малярия
31	Migraine	Мигрень
32	Osteoarthristis	Остеоартрит
33	Paralysis (brain hemorrhage)	Паралич (кровоизлияние в мозг)

34	Peptic ulcer diseae	Язвенная болезнь
35	Pneumonia	Пневмония
36	Psoriasis	Псориаз
37	Tuberculosis	Туберкулез
38	Typhoid	Брюшной тиф
39	Urinary tract infection	Инфекции мочевыводящих путей
40	Varicose veins	Варикозное расширение вен
41	hepatitis A	Гепатит А

Таблица 2. Заболевания

В следующих разделах рассмотрим работу алгоритмов дерево решений, логистическая регрессия и наивный Байес с исследуемой базой данных (см. разделы 2.2.1, 2.2.2, 2.2.3)

3.2 Визуализация алгоритма дерево решений

Пусть задано обучающее множество S, содержащее 4920 примеров, для каждого атрибута из которых задана метка класса и 132 атрибута, которые определяют принадлежность объекта к тому или иному классу.

Рабочий процесс алгоритма дерево решений изображен на Рисунок 5:

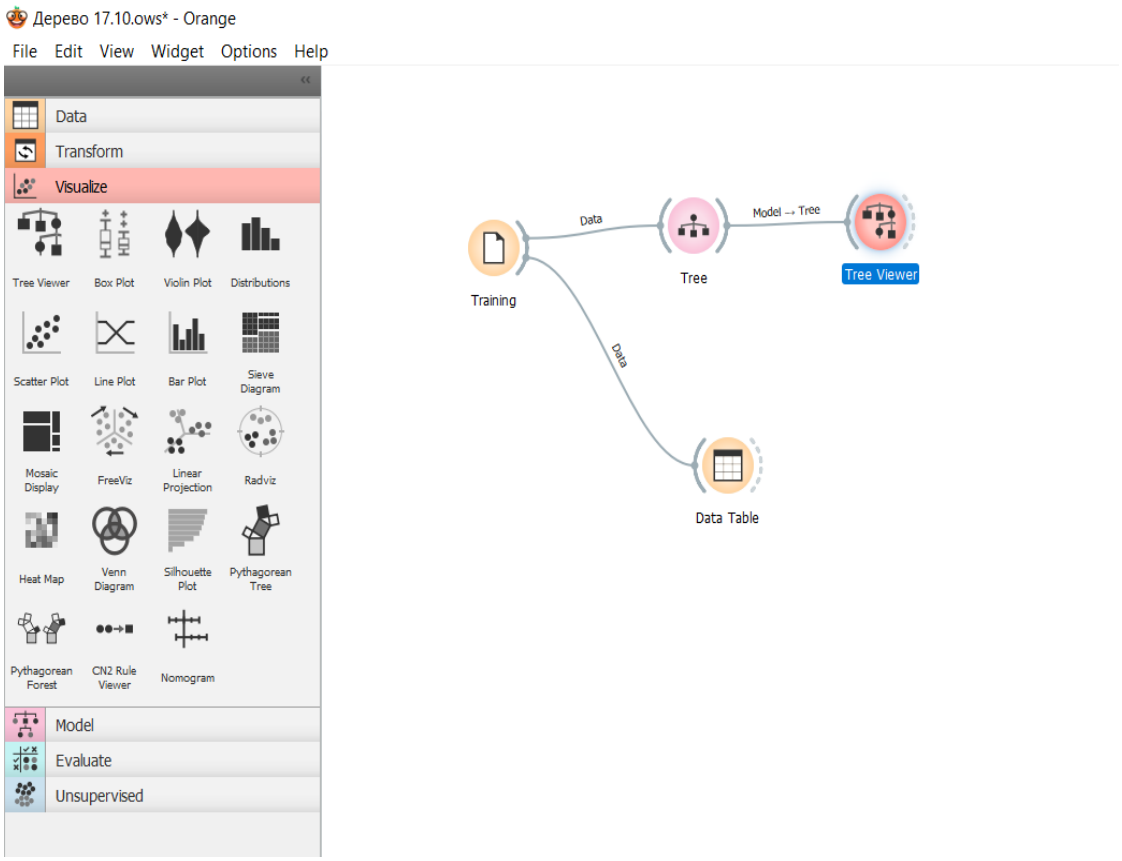


Рисунок 5. Рабочий процесс алгоритма дерево решений

Из базы данных был выбран файл Training, который был связан с несколькими виджетами.

Для представления данных в виде таблицы и возможности сортировки данных по значениям атрибутов, отправляем данные в таблицу данных (Рисунок 6):

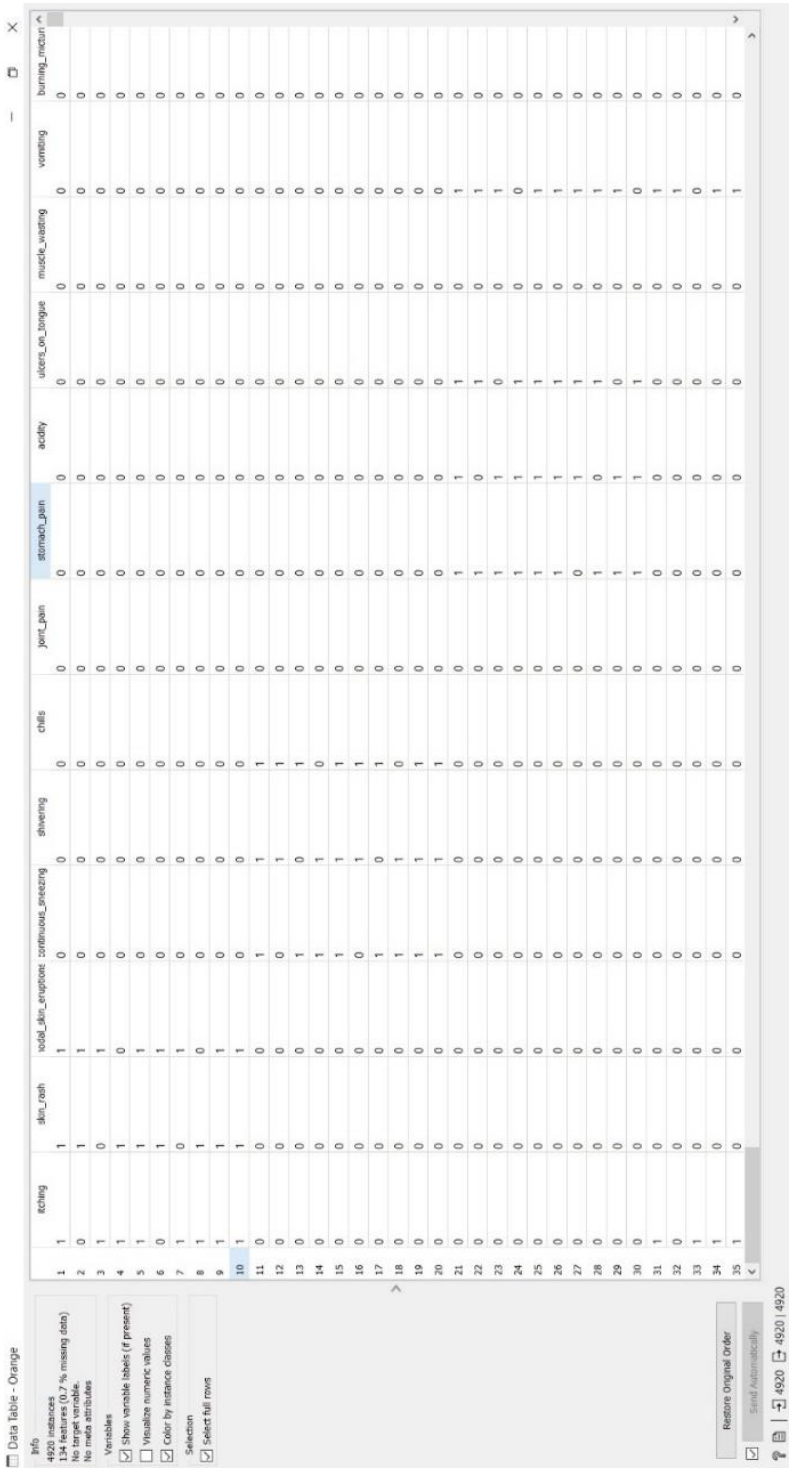


Рисунок 6. Таблица данных 4920 случаев заболеваний с 133 атрибутами, включающими 132 симптома и 1 прогноз

Отправляем данные в виджет «дерево решений», затем для визулизации результата работы алгоритма отправляем их в виджет «просмоторщик дерева».

На определенном этапе выбранный атрибут разбивает множество наблюдений в узле так, чтобы результирующие подмножества содержали

примеры с одинаковыми метками класса или были максимально приближены к этому. Если неопределенность классификации наибольшая, энтропия H (Формула 1) максимальна. Лучшим атрибутом разбиения (симптомом заболевания) будет тот, который обеспечит максимальное снижение энтропии результирующего подмножества относительно родительского. Когда атрибут выбран, например, yellowing of eyes =1, то множество значений атрибута разбивается на 2 подмножества так, что должен быть максимальный прирост информации – величине обратно пропорциональной энтропии H.

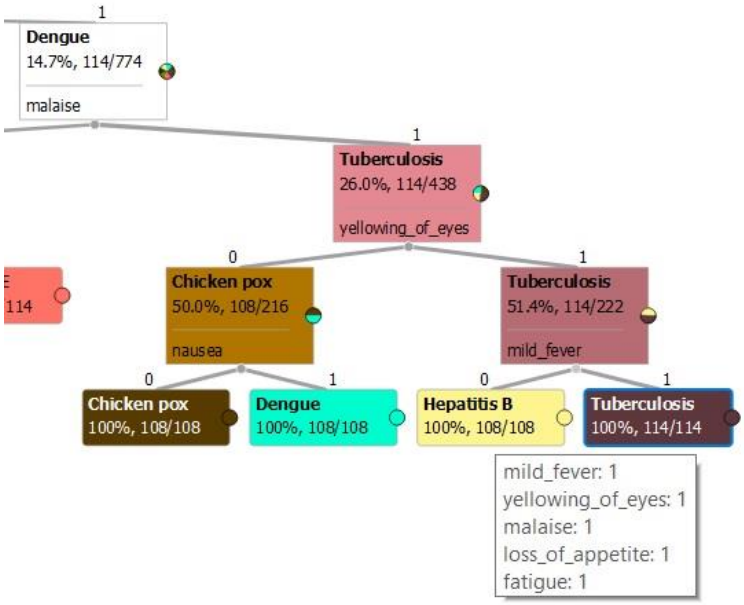


Рисунок 7. Процесс классификации по симптомам заболевания туберкулез.

На Рисунок 7 представлен один из путей работы алгоритма дерево решений при классификации заболевания туберкулез. В 114 случаях при умеренной лихорадке, пожелтении глаз, недомогании, отсутствии аппетита, усталости, алгоритм определяет предполагаемый диагноз туберкулез.

В 6 случаях алгоритм дерево решений классифицирует заболевание туберкулез по наличию: озноба, кашля, боли в груди и рвоте и отсутствию: боли в животе, тошноты и усталости.

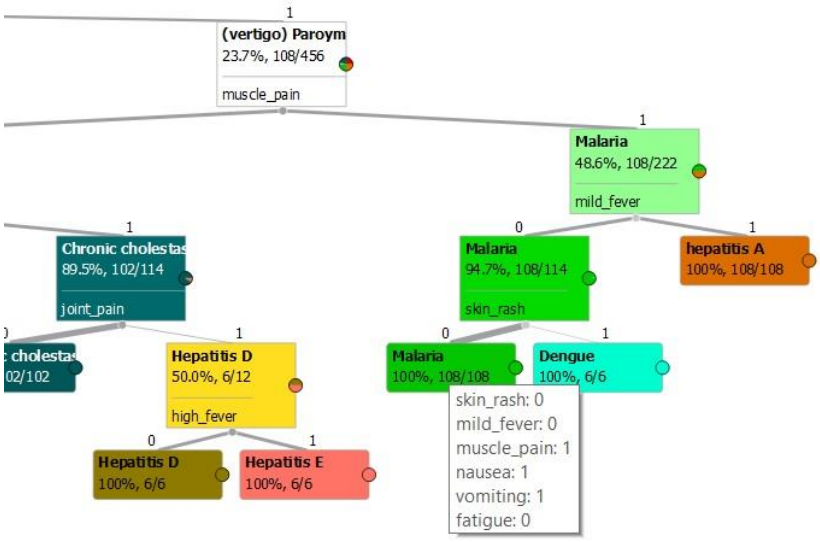


Рисунок 8. Процесс классификации по симптомам заболевания малярия

На Рисунок 8 представлен один из путей процесса классификации заболевания малярия. В 108 случаях при отсутствии кожной сыпи, умеренной лихорадки, усталости и при наличии боли в мышцах, тошноты, рвоты алгоритм показывает, что предполагаемый диагноз малярия.

В 6 случаях алгоритм дерево решений классифицирует заболевание малярия по наличию озноба, диареи и рвоты и отсутствию боли в груди, боли в животе, тошноты и усталости.

В третьем случае алгоритм классифицирует заболевание по наличию озноба и головной боли, а также отсутствию непрерывного чихания, повышенной кислотности желудочного сока, потери равновесия, кожной сыпи, рвоты и усталости.

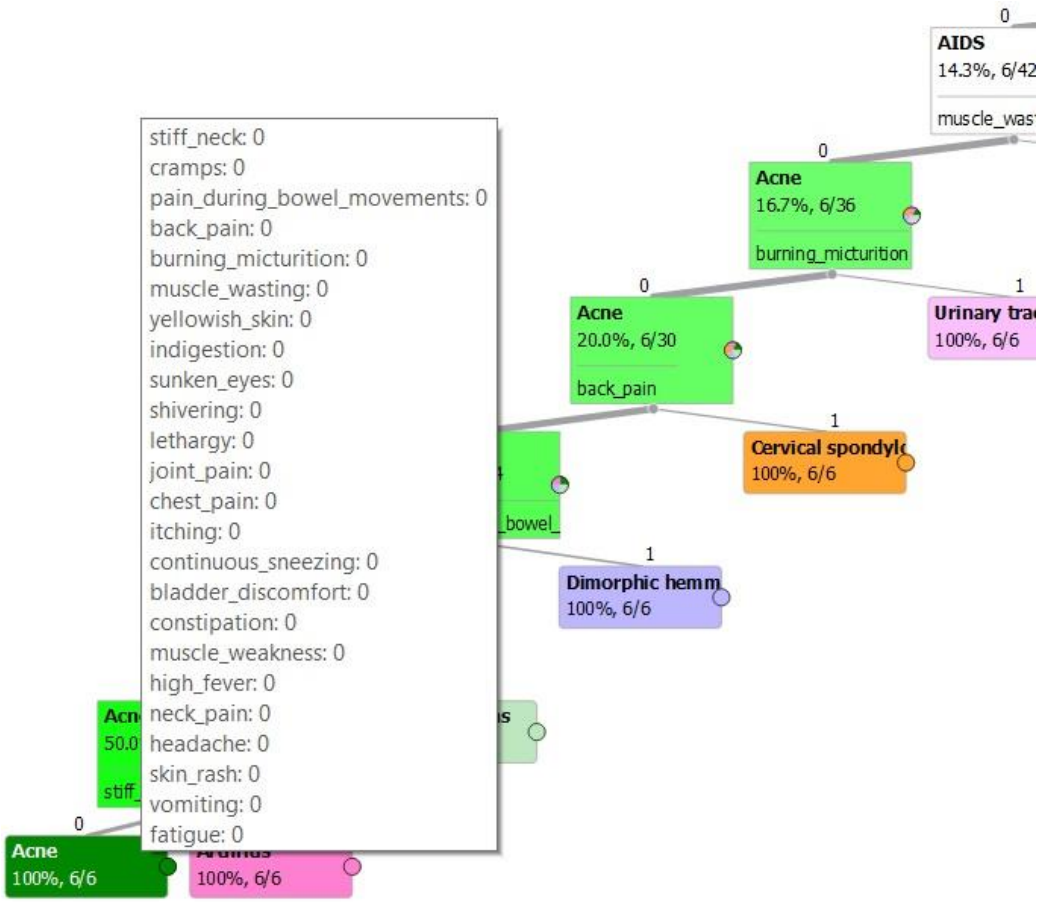


Рисунок 9. Процесс классификации по симптомам заболевания акне

На Рисунок 9 представлен один из путей процесса классификации симптомов при прогнозировании заболевания акне. В 6 случаях классификация при постановке диагноза происходит не по наличию симптомов, а по их отсутствию.

В 12 случаях заболевание акне классифицируется по наличию угрей и кожной сыпи, а также отсутствию узловых высыпаний на коже, гнойных прыщей, волдырей, боли в суставах, зуда, рвоты и усталости.

Ещё в 102 случаях заболевание акне классифицируется по наличию гнойных прыщей и кожной сыпи, и отсутствию волдырей, боли в суставах, зуда, рвоты и усталости.

Основной особенностью алгоритма дерево решений является возможность классификации заболеваний по разным симптомам.

3.3. Визуализация алгоритма логистической регрессии

Рабочий процесс алгоритма логистической регрессии показан на Рисунок 10:

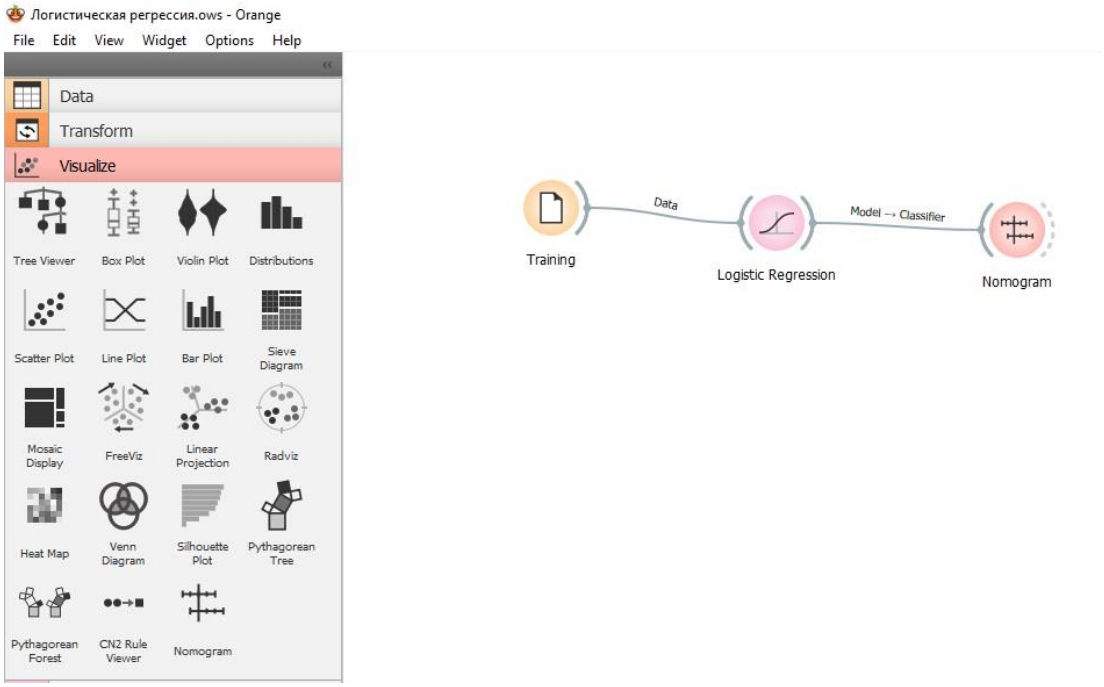


Рисунок 10. Рабочий поток исследований логистическая регрессия.

Файл Training из базы данных был связан с виджетами «логистическая регрессия» и «номограмма».

Заболевания для сравнения были взяты, что и при рассмотрении алгоритма Дерево решений. Для наглядности представлены 10 наиболее значимых симптомов:

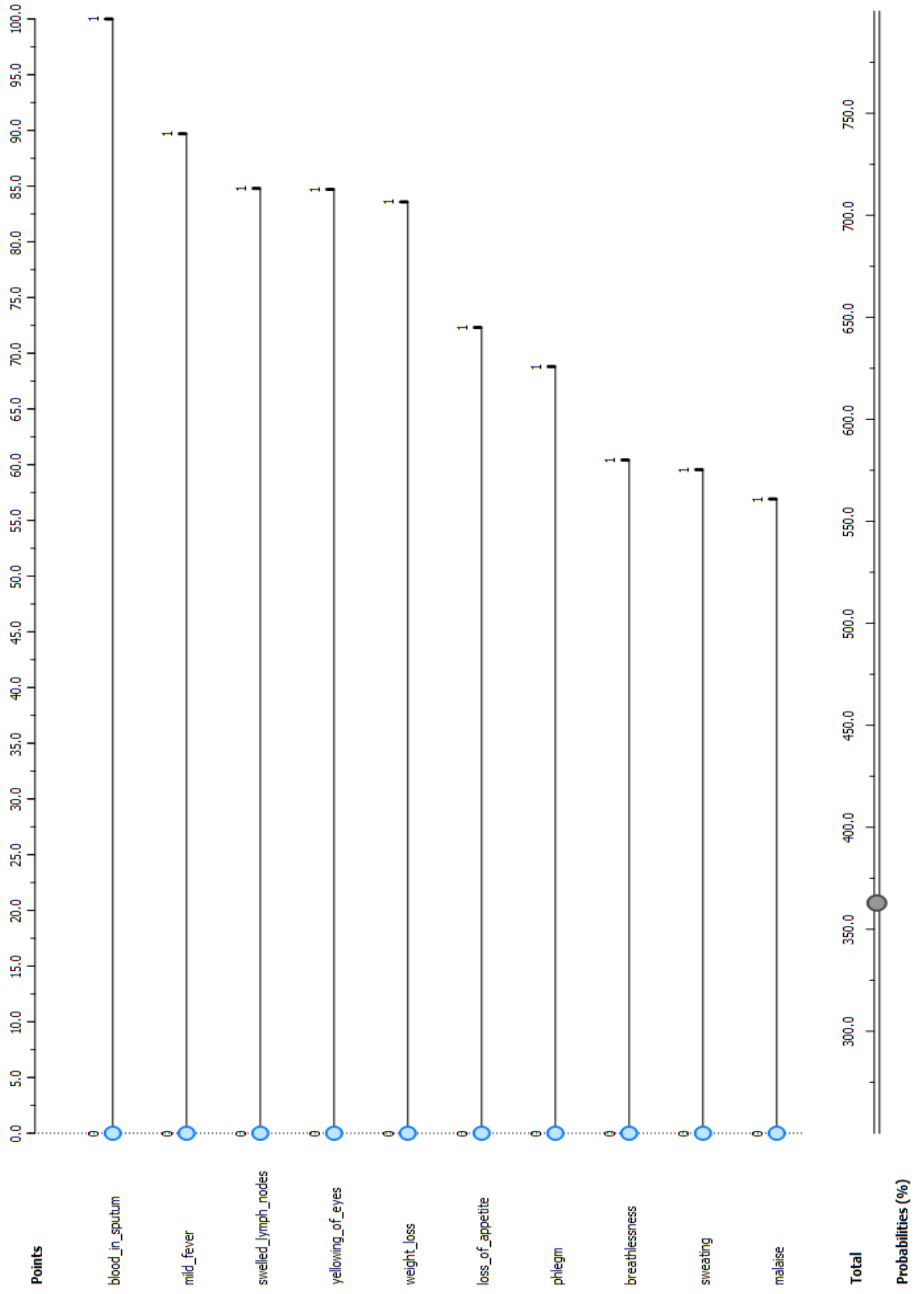


Рисунок 11. Номограмма структура обучающих данных и влияние атрибутов заболевания туберкулез

На Рисунок 11 изображена визуализация алгоритма логистической регрессии при классификации туберкулеза. Основные симптомы (атрибуты), влияющие на классификацию туберкулеза - это кровь в мокроте, умеренная лихорадка, увеличенные лимфатические узлы, пожелтение глаз, потеря веса, отсутствие аппетита, мокрота, отдышка, потливость, недомогание. По номограмме можно сделать вывод о степени влияния каждого симптома на классификацию данного заболевания.

Для облегчения понимания и интерпретации используется Шкала баллов. Единица получается путем перемасштабирования логарифмических шансов таким образом, чтобы максимальное абсолютное логарифмическое отношение шансов в номограмме составляло 100 баллов.

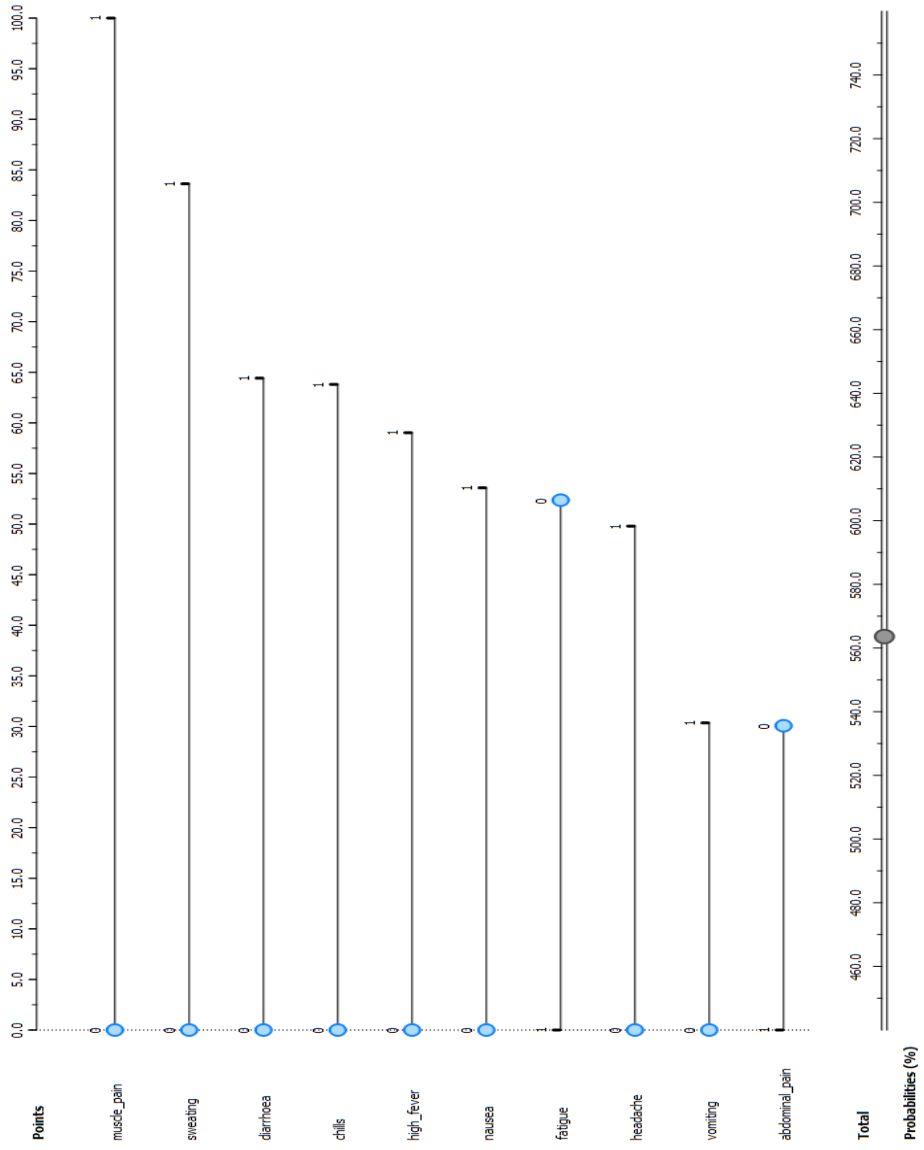


Рисунок 12. Номограмма структура обучающих данных и влияние атрибутов заболевания малярия.

На Рисунок 12 изображена визуализация алгоритма логистической регрессии при заболевании малярия. Основные симптомы влияющие на классификацию малярии – это боли в мышцах, потливость, диарея, озноб, высокая температура, тошнота, усталость, головная боль, рвота и боль в

животе. Алгоритм учитывает не наличие усталости и боли в животе, а их отсутствие.

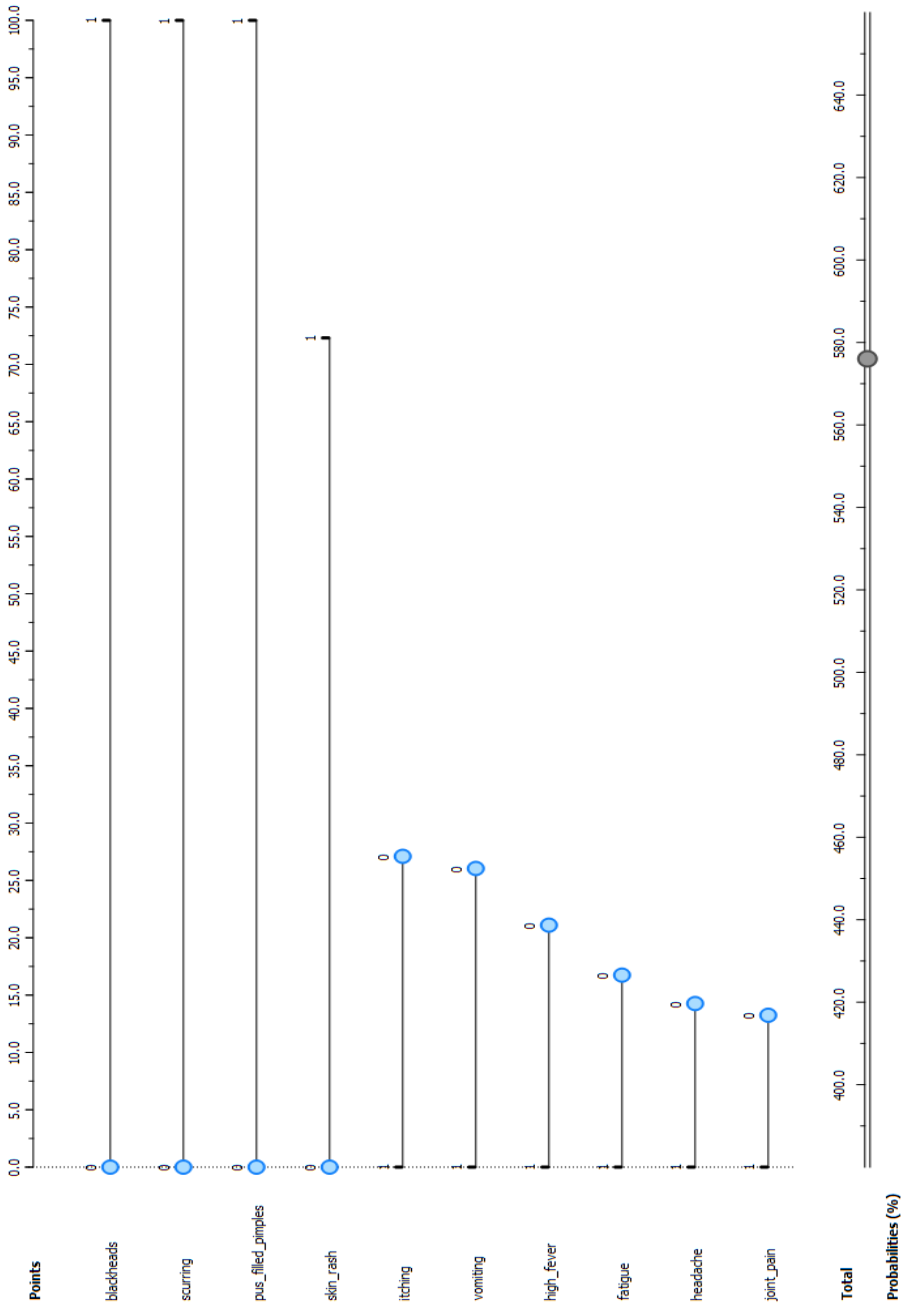


Рисунок 13. Номограмма структура обучающих данных и влияние атрибутов заболевания акне.

На Рисунок 13 изображена визуализация алгоритма логистической регрессии при классификации акне. На классификацию акне в первую очередь влияет наличие следующих симптомов: угри, возбужденность, гнойные прыщи и кожной сыпи, и отсутствие: зуда, рвоты, высокой температуры, усталости, головной боли и боли в суставах.

Рассмотрим работу алгоритма логистическая регрессия на конкретном примере данных. Для этого возьмем из обучающей базы данных клинический случай диабета и проанализируем его номограмму (см. Рисунок 14).

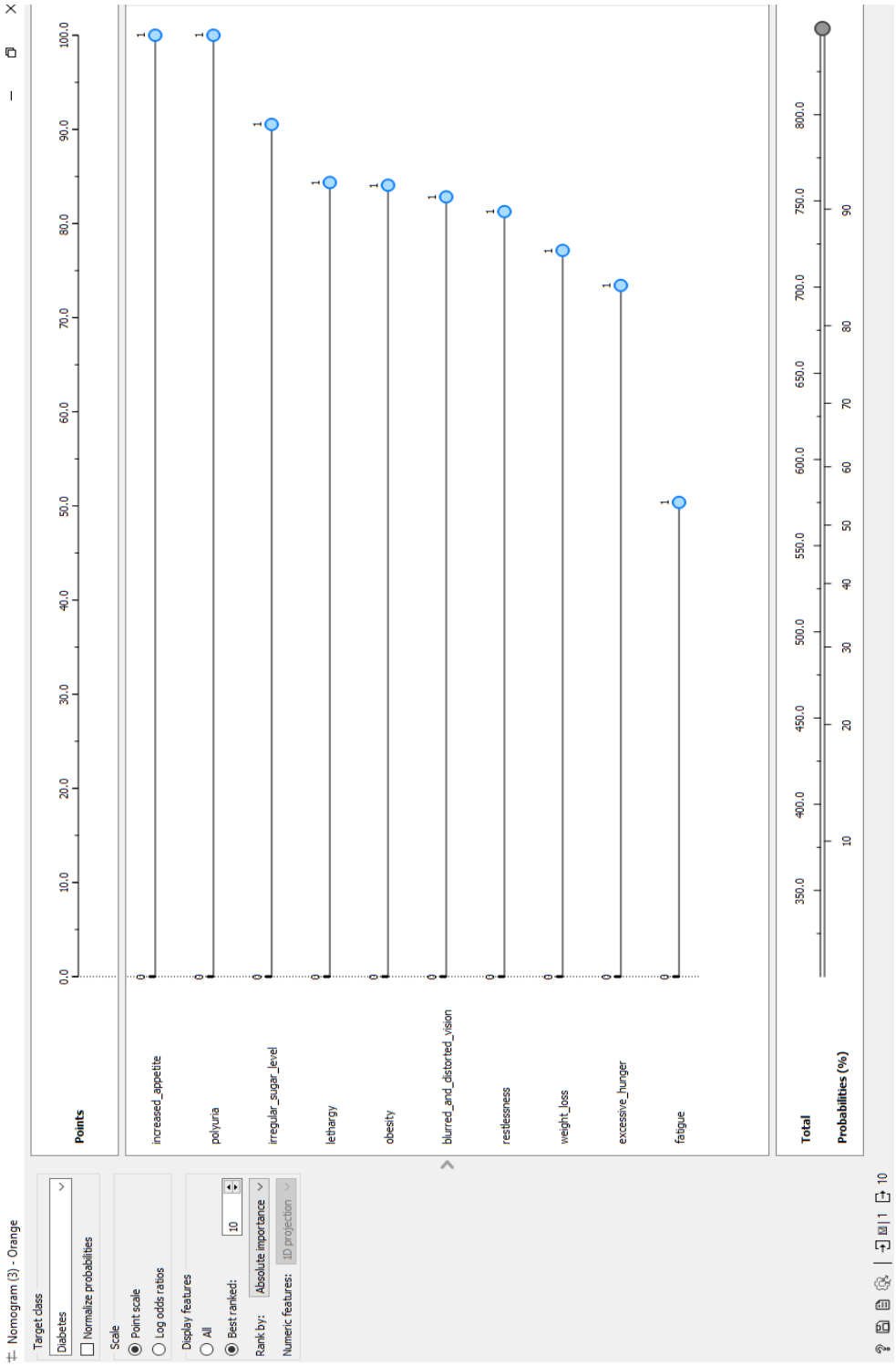


Рисунок 14. Номограмма клинического случая диабета.

Номограмма изображенная на Рисунок 14 позволяет выполнить оценку баллами вклада каждого симптома, отдельные баллы суммируются для

определения вероятности заболевания. Наибольший вклад в диагностику, в порядке убывания вносят следующие симптомы: повышенный аппетит, полиурия, не стабильный уровень сахара, вялость, ожирение, размытое и искаженное зрение, беспокойство, потеря веса, чрезмерный голод и усталость.

Для полноты картины рассмотрим этот же клинический случай (диабет), в сравнении с бронхиальной астмой (Рисунок 15):

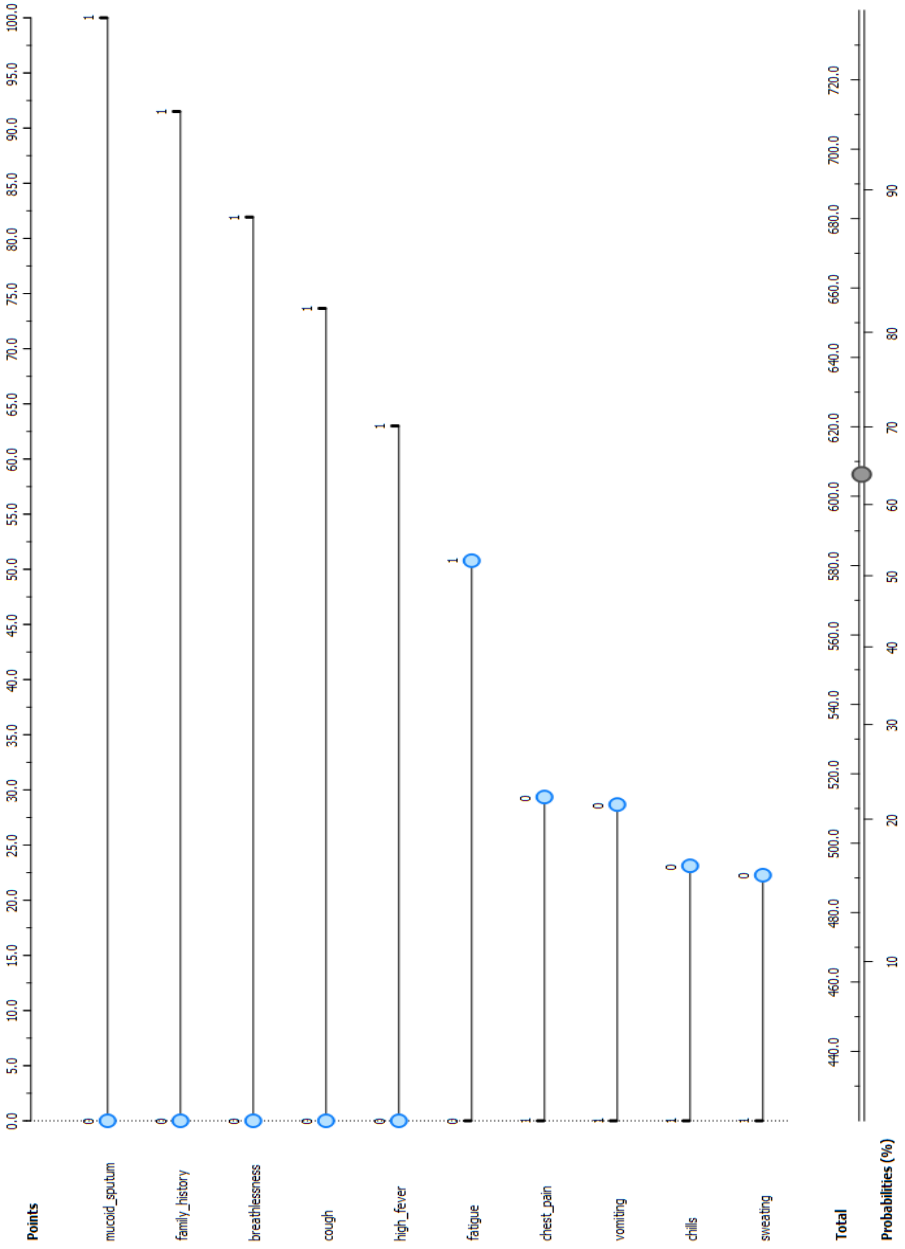


Рисунок 15. Номограмма клинического случая диабета в сравнении с бронхиальной астмой

Можно сделать вывод, что алгоритм определяет вероятность диабета, намного выше, чем вероятность бронхиальной астмы. Но вероятность того что в данном случае это бронхиальная астма алгоритм оценил в 64%. Это вызвано тем что алгоритм логистической регрессии оценивает только положительный вклад каждого симптома в классификацию заболевания. И даже если нет одинаковых симптомов, то есть совпадения в отсутствующих симптомах, которые внесут вклад в вероятность того или иного заболевания, в данном случае бронхиальной астмы.

Таким образом, можно судить о том, что алгоритм логистической регрессии дает более полное представление о влиянии симптомов заболевания на его классификацию в сравнении с алгоритмом дерево решений, но имеет свои недостатки.

3.4. Визуализация алгоритма наивный Байес

Рабочий процесс алгоритма наивный Байес отображен на Рисунок 16:

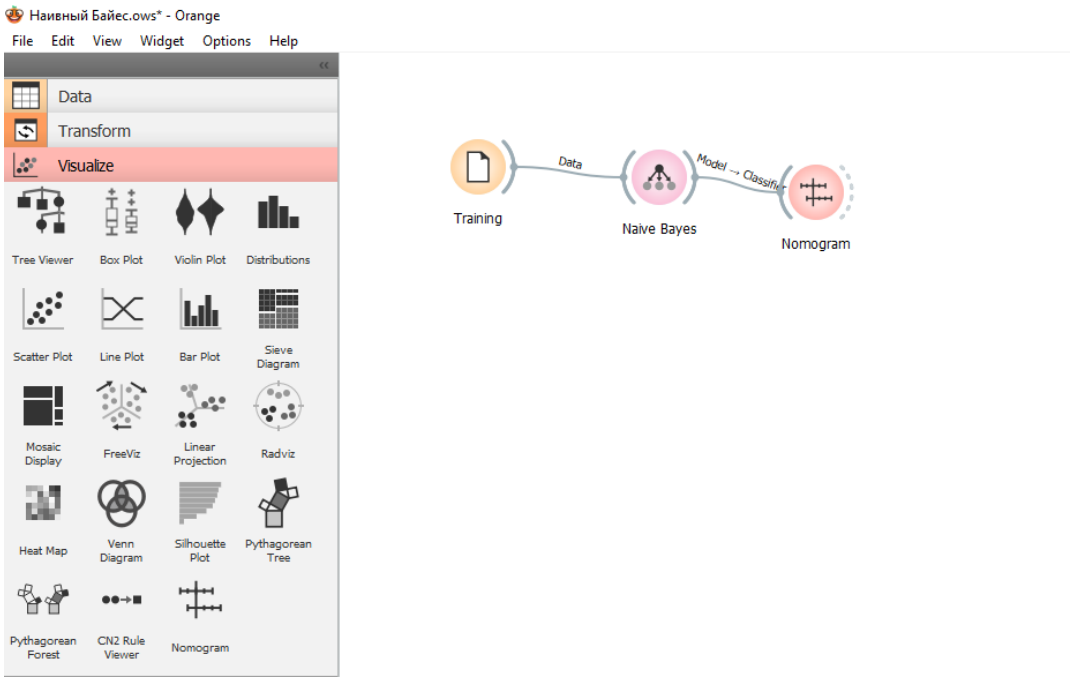


Рисунок 16. Рабочий поток исследований наивный Байес

Файл Training соединен с виджетами «наивный байес» и «номограмма».

Рассмотрим 10 наиболее значимых симптомов для тех же заболеваний, что и в предыдущих алгоритмах:

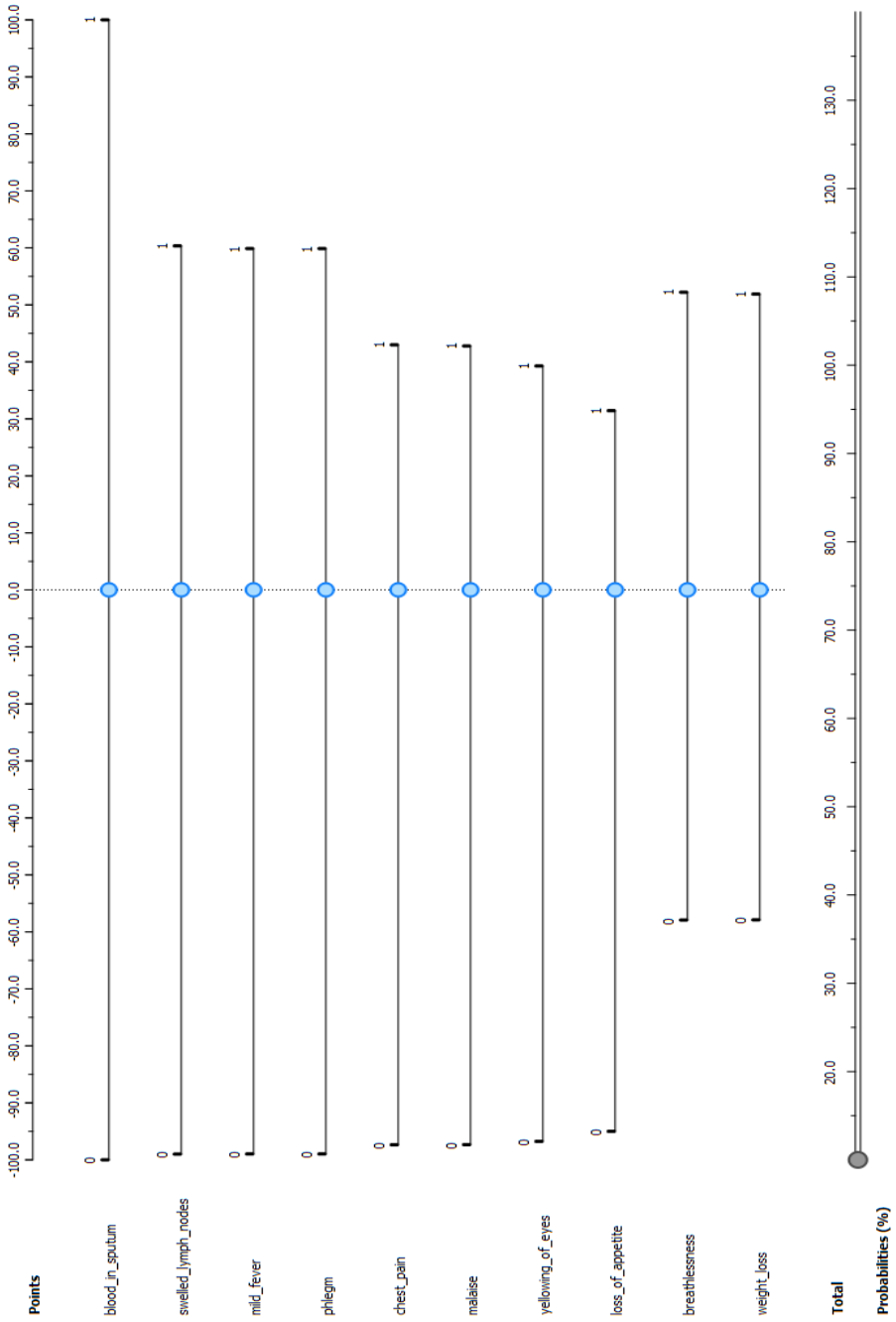


Рисунок 17. Номограмма структура обучающих данных и влияние атрибутов заболевания туберкулез.

На Рисунок 17 видно что основные симптомы, влияющие на классификацию признаков в диагностике туберкулеза, при использовании алгоритма наивный Байес - это кровь в мокроте, увеличенные лимфатические

узлы, умеренная лихорадка, мокрота, боль в груди, недомогание, пожелтение глаз, отсутствие аппетита, отдышка, потеря веса.

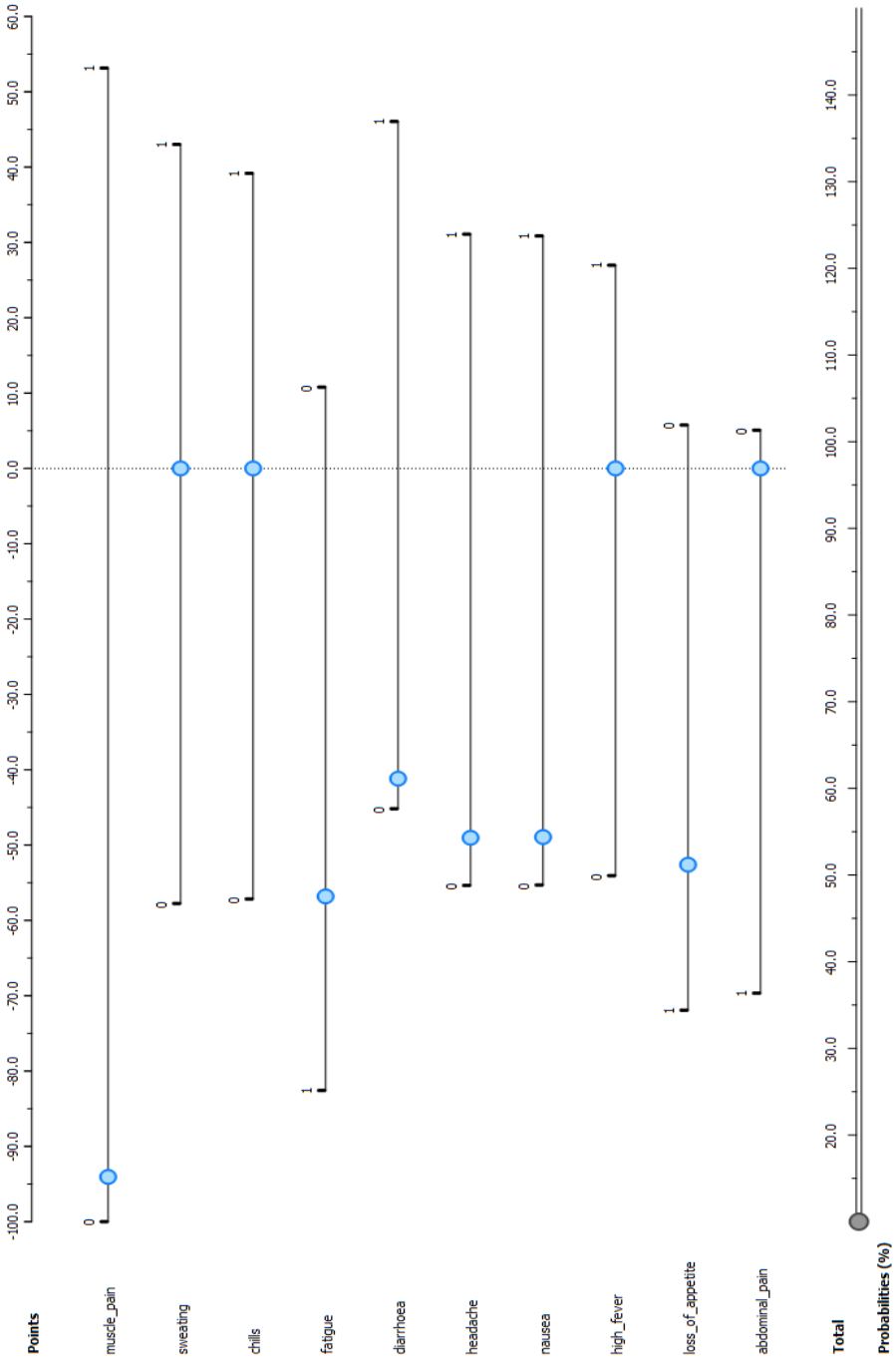


Рисунок 18. Номограмма структура обучающих данных и влияние атрибутов заболевания малярия.

Как следует из Рисунок 18 основными симптомами при классификации малярии будут боли в мышцах, потливость, озноб, усталость, диарея, головная боль, тошнота, высокая температура, отсутствие аппетита и боли в животе. Также как при использовании алгоритма логистической регрессии, в

данном случае учитывается не наличие, а отсутствие таких симптомов как усталость, отсутствие аппетита, боль в животе.

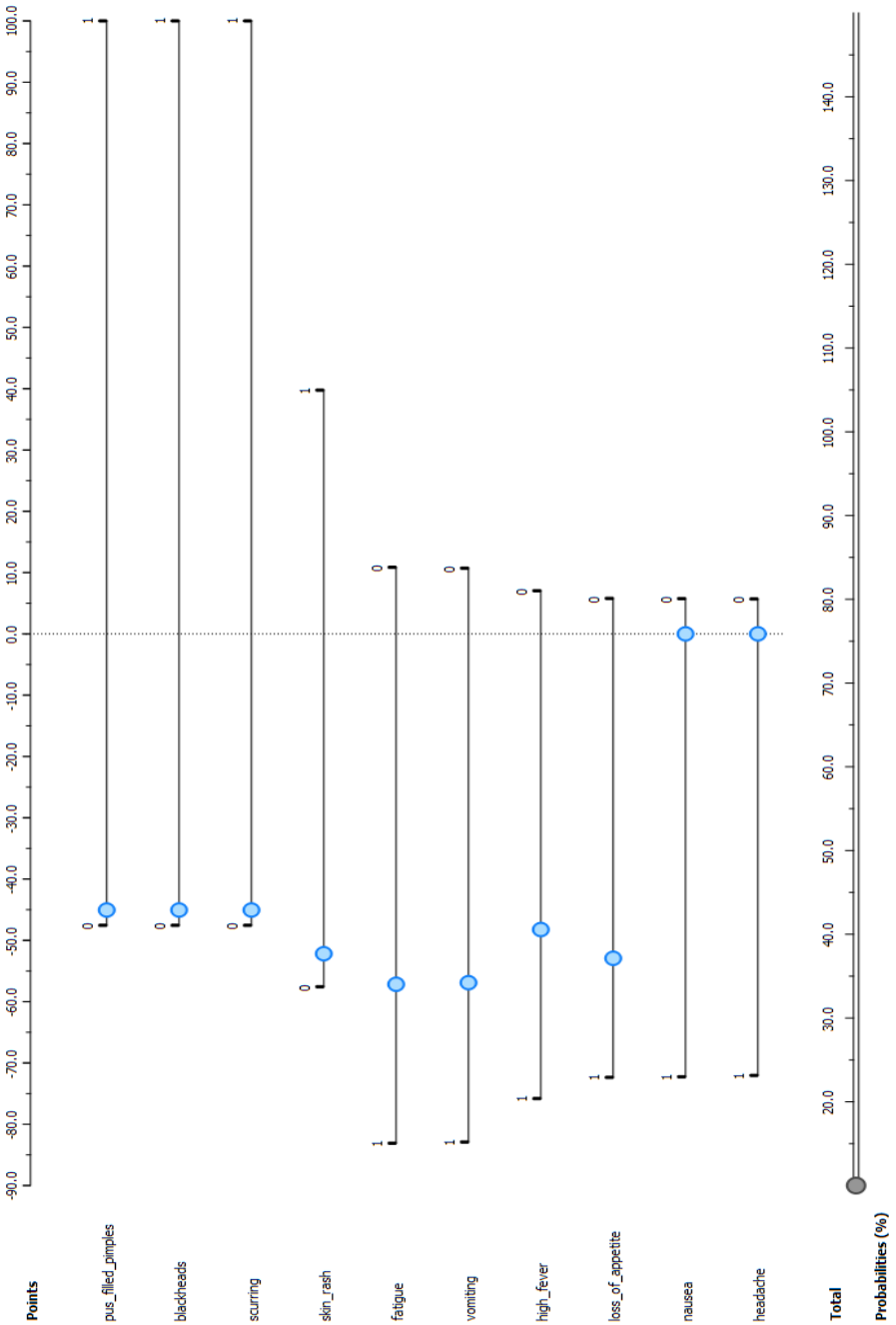


Рисунок 19. Номограмма структура обучающих данных и влияние атрибутов заболевания акне.

Основные симптомы при классификации акне (Рисунок 19) - это гнойные прыщи, угри, возбуждённость, кожная сыпь, а также отсутствие

усталости, рвоты, высокой температуры, плохого аппетита, тошноты и головной боли.

Рассмотрим работу алгоритма наивный Байес на конкретном примере данных. Для этого возьмем из обучающей базы данных конкретный клинический случай диабета и проанализируем его номограмму (Рисунок 20):

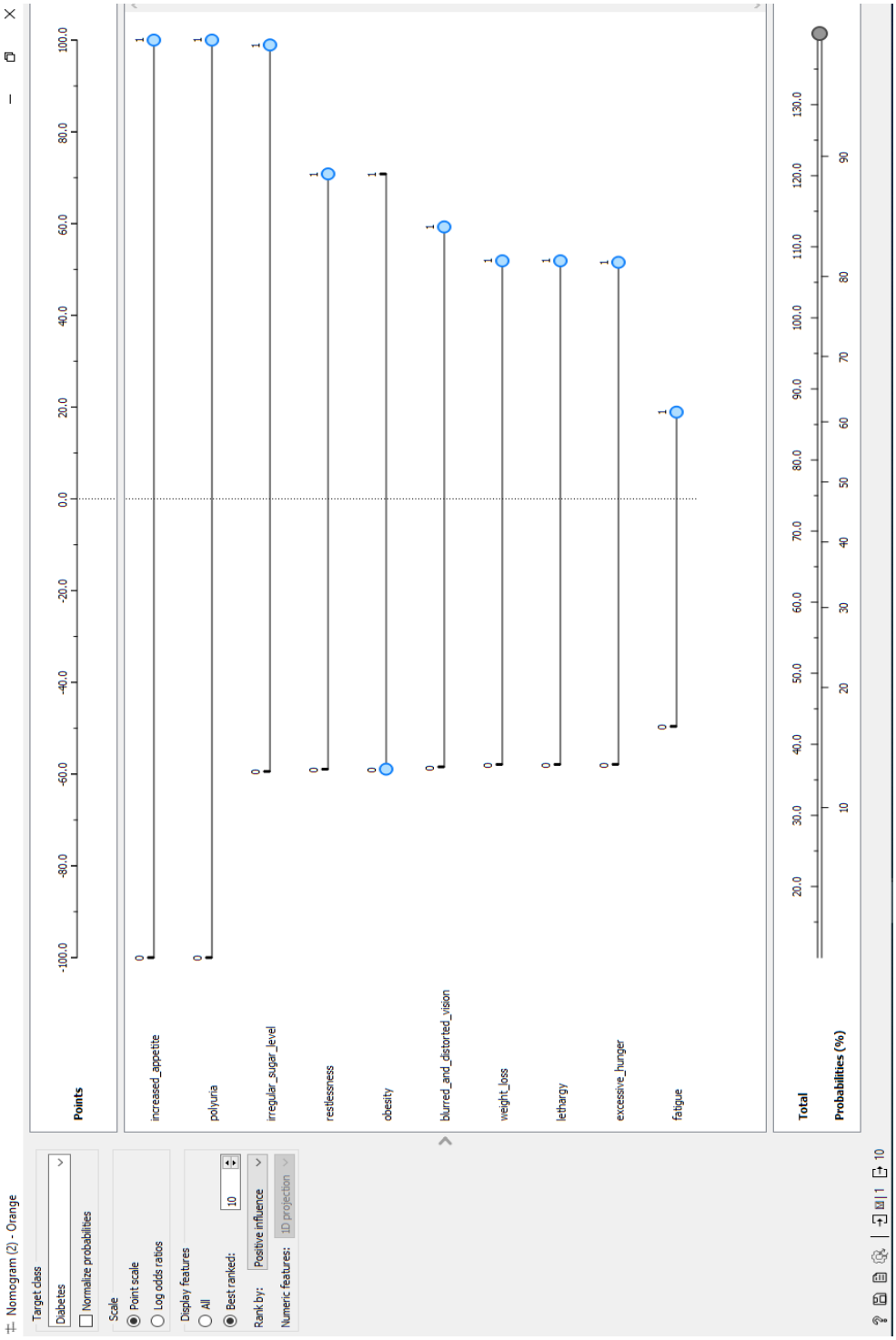


Рисунок 20. Номограмма клинического случая диабета (наивный Байес)

На Рисунок 20 можно оценить вклад каждого симптома в виде баллов. Отдельные баллы суммируются для определения вероятности. В отличии от алгоритма логистической регрессии в данном примере можно оценить не только положительный, но и отрицательный вклад каждого симптома в классификацию заболевания. Например, у больного не было ожирения и отсутствие данного симптома алгоритм оценил отрицательными баллами.

Рассмотрим этот же клинический случай в сравнении с бронхиальной астмой (Рисунок 21):

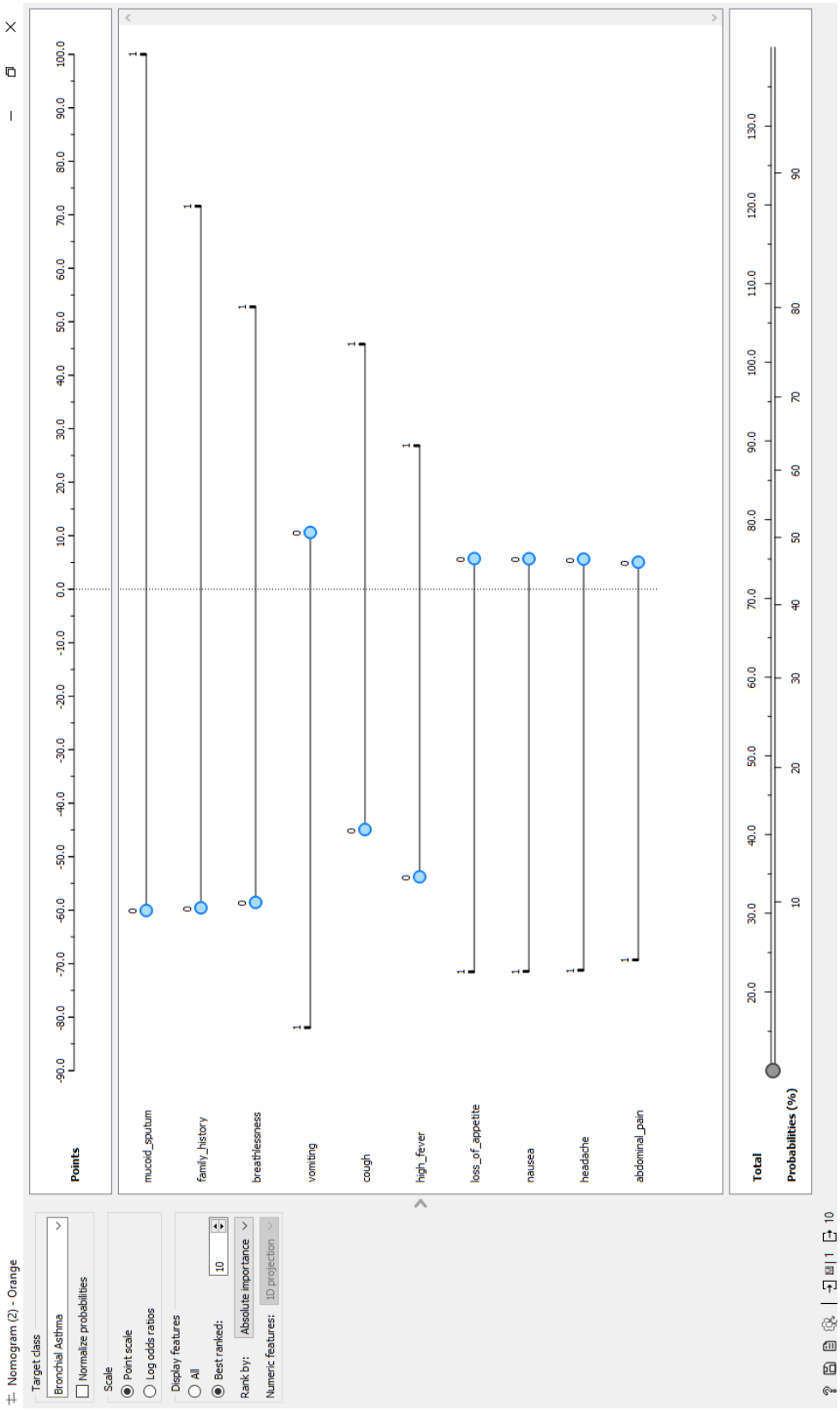


Рисунок 21. Номограмма клинического случая диабета в сравнении с бронхиальной астмой.

На Рисунок 21 видно что алгоритм правильно классифицировал заболевание как диабет, а не бронхиальная астма. По полученным

результатам можно судить о том, что алгоритм наивного Байеса дает наиболее полное представление о структуре обучающих данных и влиянии атрибутов на вероятности класса.

Таким образом номограмма правильно выделяет уникальные для данного заболевания наиболее значимые атрибуты. Такой способ визуализации способствует принятию врачебных решений на основе статистических опросов.

3.5. Прогнозы диагнозов моделей на основе данных о заболеваниях

Рабочий процесс прогноза диагнозов на основе данных о заболеваниях изображен на Рисунок 22:

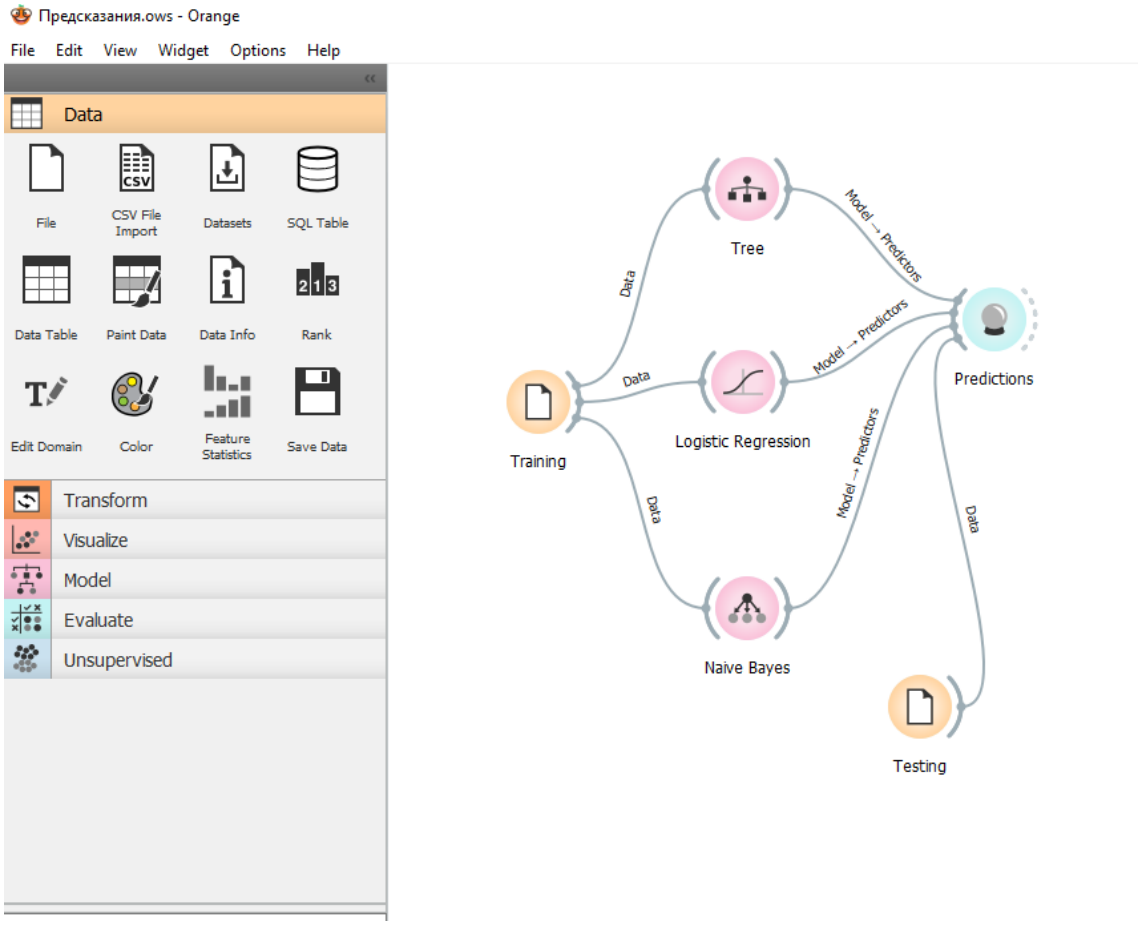
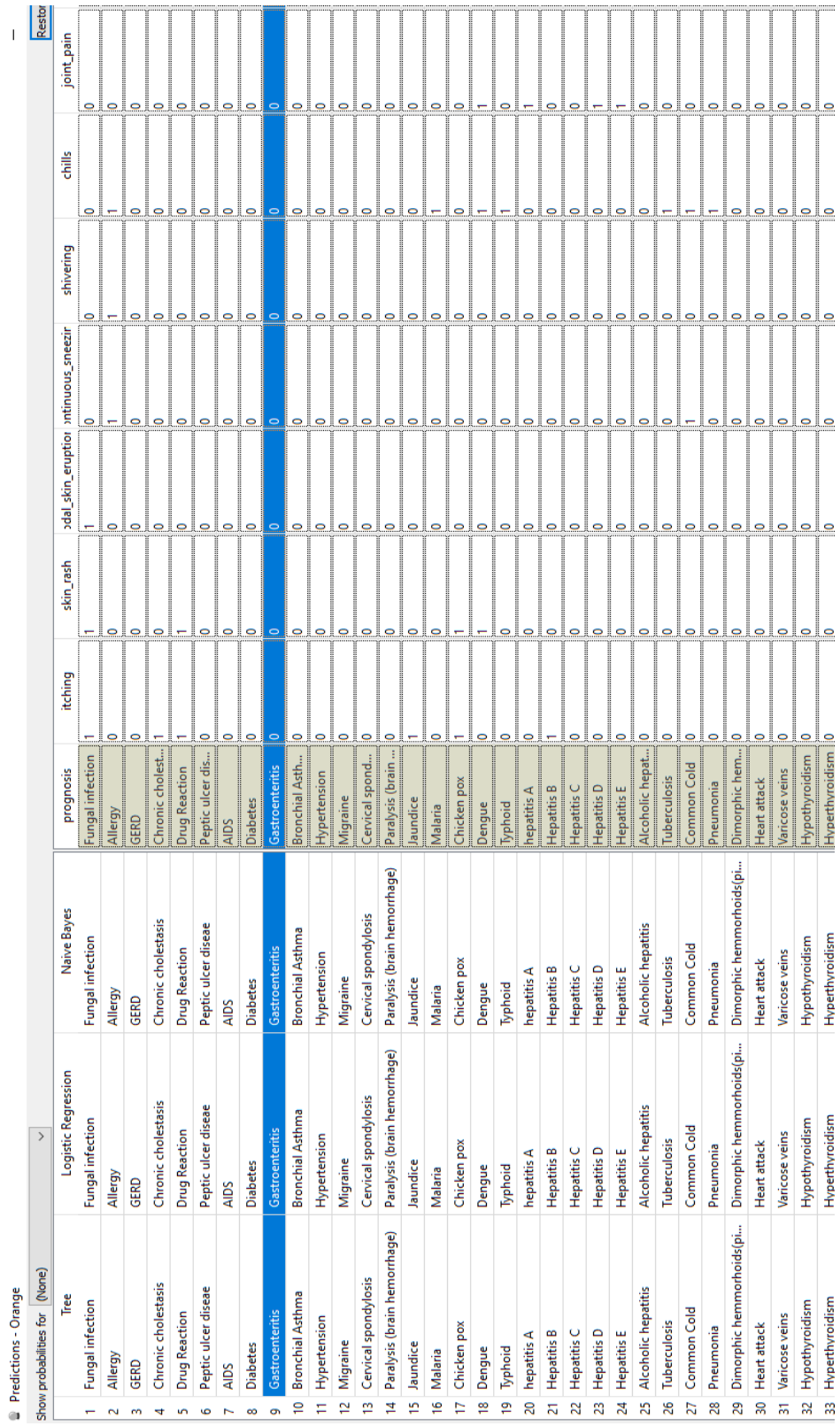


Рисунок 22. Рабочий процесс прогноза диагнозов моделей на основе данных о заболеваниях.

Для прогнозов необходимы обучающие данные (файл Training) и тестовые данные (файл Testing). Файл для обучения в нашей работе содержит

4920 случаев заболеваний, а файл для тестирования 41 случай заболеваний. В результате получаем набор данных, к которому можем добавить прогнозы диагнозов в виде новых метаатрибутов:


Predictions - Orange

Show probabilities for (None)

	Tree	Logistic Regression	Naive Bayes	prognosis	itching	skin_rash	dal_skin_eruptio	ntinuous_sneezir	shivering	chills	joint_pain
1	Fungal infection	Fungal infection	Fungal infection	Fungal infection	1	1	1	0	0	0	0
2	Allergy	Allergy	Allergy	Allergy	0	0	0	1	1	1	0
3	GERD	GERD	GERD	GERD	0	0	0	0	0	0	0
4	Chronic cholestasis	Chronic cholestasis	Chronic cholestasis	Chronic cholest...	1	0	0	0	0	0	0
5	Drug Reaction	Drug Reaction	Drug Reaction	Drug Reaction	1	1	0	0	0	0	0
6	Peptic ulcer diseae	Peptic ulcer diseae	Peptic ulcer diseae	Peptic ulcer dis...	0	0	0	0	0	0	0
7	AIDS	AIDS	AIDS	AIDS	0	0	0	0	0	0	0
8	Diabetes	Diabetes	Diabetes	Diabetes	0	0	0	0	0	0	0
9	Gastroenteritis	Gastroenteritis	Gastroenteritis	Gastroenteritis	0	0	0	0	0	0	0
10	Bronchial Asthma	Bronchial Asthma	Bronchial Asthma	Bronchial Asth...	0	0	0	0	0	0	0
11	Hypertension	Hypertension	Hypertension	Hypertension	0	0	0	0	0	0	0
12	Migraine	Migraine	Migraine	Migraine	0	0	0	0	0	0	0
13	Cervical spondylosis	Cervical spondylosis	Cervical spondylosis	Cervical spond...	0	0	0	0	0	0	0
14	Paralysis (brain hemorrhage)	Paralysis (brain hemorrhage)	Paralysis (brain hemorrhage)	Paralysis (brain ...	0	0	0	0	0	0	0
15	Jaundice	Jaundice	Jaundice	Jaundice	1	0	0	0	0	0	0
16	Malaria	Malaria	Malaria	Malaria	0	0	0	0	0	1	0
17	Chicken pox	Chicken pox	Chicken pox	Chicken pox	1	1	0	0	0	0	0
18	Dengue	Dengue	Dengue	Dengue	0	1	0	0	0	1	1
19	Typhoid	Typhoid	Typhoid	Typhoid	0	0	0	0	0	1	0
20	hepatitis A	hepatitis A	hepatitis A	hepatitis A	0	0	0	0	0	0	1
21	Hepatitis B	Hepatitis B	Hepatitis B	Hepatitis B	1	0	0	0	0	0	0
22	Hepatitis C	Hepatitis C	Hepatitis C	Hepatitis C	0	0	0	0	0	0	0
23	Hepatitis D	Hepatitis D	Hepatitis D	Hepatitis D	0	0	0	0	0	0	1
24	Hepatitis E	Hepatitis E	Hepatitis E	Hepatitis E	0	0	0	0	0	0	1
25	Alcoholic hepatitis	Alcoholic hepatitis	Alcoholic hepatitis	Alcoholic hepat...	0	0	0	0	0	0	0
26	Tuberculosis	Tuberculosis	Tuberculosis	Tuberculosis	0	0	0	0	0	1	0
27	Common Cold	Common Cold	Common Cold	Common Cold	0	0	0	1	0	1	0
28	Pneumonia	Pneumonia	Pneumonia	Pneumonia	0	0	0	0	0	1	0
29	Dimorphic hemmorhoids(pi...	Dimorphic hemmorhoids(pi...	Dimorphic hemmorhoids(pi...	Dimorphic hem...	0	0	0	0	0	0	0
30	Heart attack	Heart attack	Heart attack	Heart attack	0	0	0	0	0	0	0
31	Varicose veins	Varicose veins	Varicose veins	Varicose veins	0	0	0	0	0	0	0
32	Hypothyroidism	Hypothyroidism	Hypothyroidism	Hypothyroidism	0	0	0	0	0	0	0
33	Hyperthyroidism	Hyperthyroidism	Hyperthyroidism	Hyperthyroidism	0	0	0	0	0	0	0

Рисунок 23. Набор данных с добавленными прогнозами диагнозов заболеваний

Из Рисунок 23 видно, что все исследуемые алгоритмы показали правильный прогноз заболеваний.

Для того чтобы убедиться в правильности работы алгоритмов, уберем в тестовом файле столбец «прогноз». В результате набор данных имеет следующий вид (Рисунок 24):

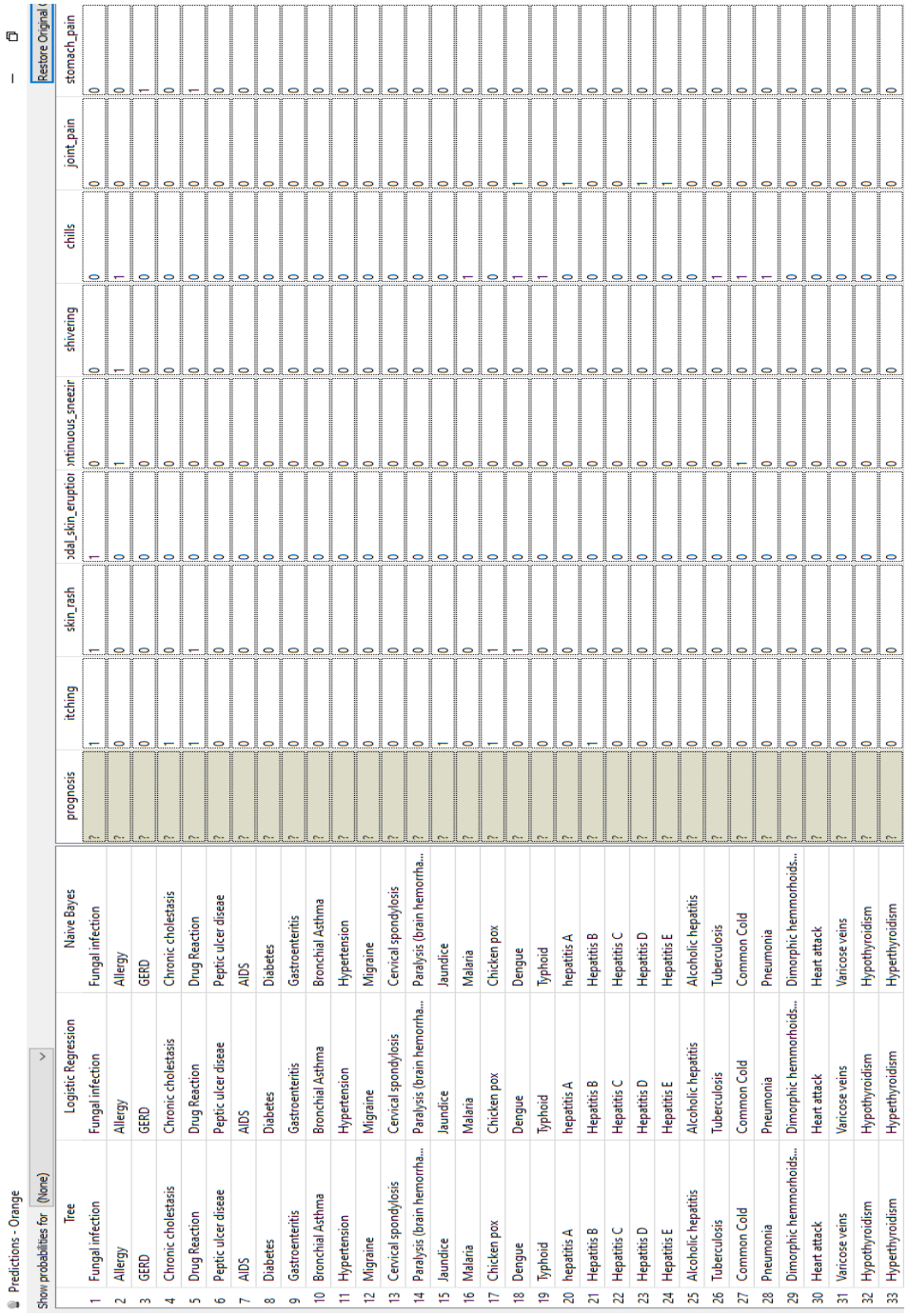


Рисунок 24. Набор данных с добавленными прогнозами диагнозов заболеваний (без изначального прогноза)

Прогнозы на Рисунок 24 не отличались от прогнозов диагнозов на Рисунок 23. Все три модели правильно диагностировали заболевания.

3.6. Тестирование алгоритмов классификации

Для проверки алгоритмов классификации на данных создан рабочий процесс следующего вида:

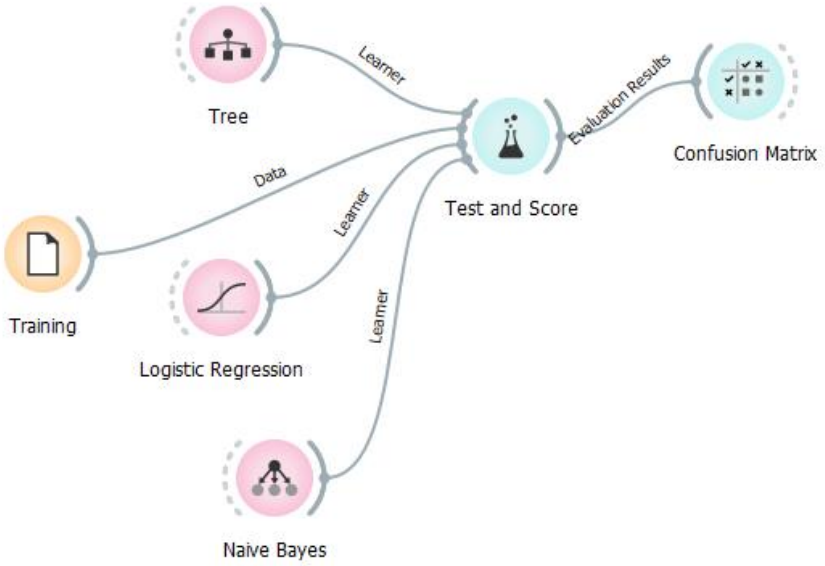


Рисунок 25. Рабочий процесс тестирования алгоритмов классификации.

Тестирование алгоритмов проведено на обучающей базе данных (файл Training). Сигнал Ученик (Learner) можно подключить более чем к одному виджету для тестирования нескольких алгоритмов с помощью одних и тех же процедур.

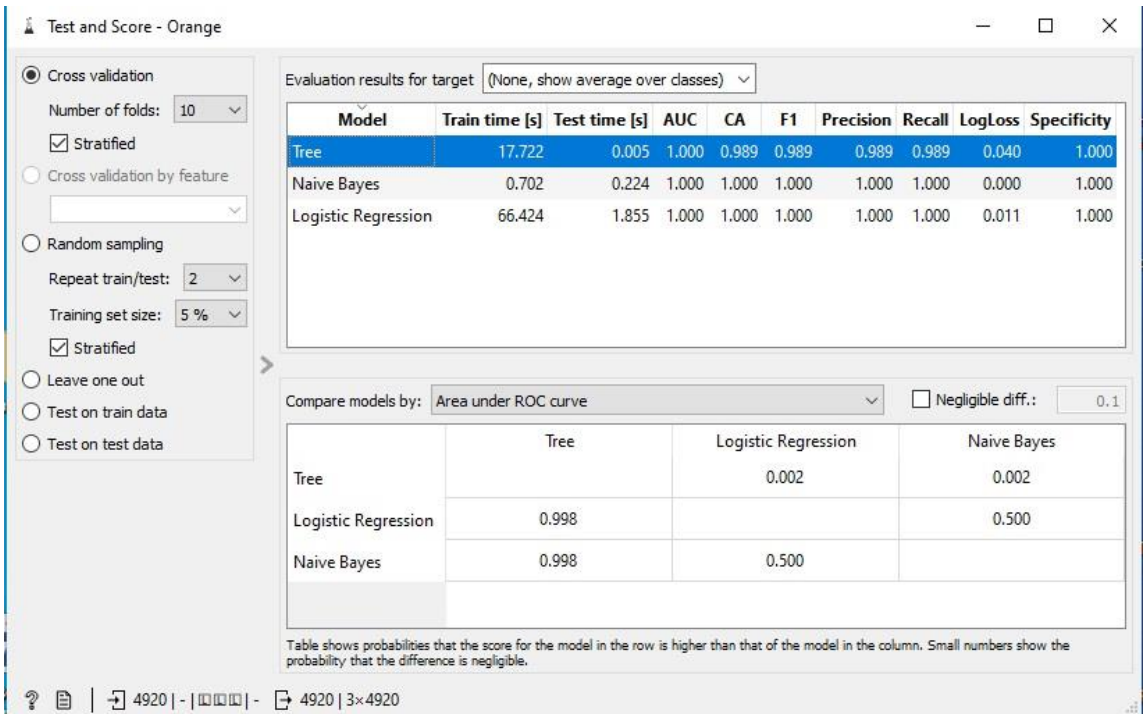


Рисунок 26. Результаты тестирования алгоритмов классификации

На Рисунок 26 представлены результаты тестирования исследуемых алгоритмов классификации. Использовалась перекрестная проверка, разбивающая данные на 10 кратностей.

Точность классификации алгоритма дерево решений составил 0,989. Однако алгоритмы логистической регрессии и наивного Байеса определили принадлежность атрибутов определенному классу с вероятностью 1.000. Следовательно, алгоритмы логистической регрессии и наивного Байеса правильно классифицируют заболевания по симптомам в 100% случаев.

При попарном сравнении моделей вероятность того, что алгоритм дерева решений лучше, алгоритма логистической регрессии и алгоритма наивного Байеса составляет 0,002. Вероятность того, что алгоритмы логистической регрессии и наивного Байеса лучше, чем алгоритм дерево решений составляет 0,098. Алгоритмы логистической регрессии и наивного Байеса продемонстрировали одинаковый результат.

Неправильно классифицированные алгоритмом дерево решений экземпляры можно найти в виджете «матрица путаницы».

Confusion Matrix

Confusion matrix for Tree (showing number of instances)

Actual \ Predicted	(vertigo) Paroymsal Positional Vertigo	AIDS	Acne	Alcoholic hepatitis	Allergy	Arthritis	Bronchial Asthma	Cervical spondylosis	Chicken pox	Chronic cholestasis	Common Cold	Dengue	Diabetes
(vertigo) Paroymsal Positional Vertigo	118	0	0	0	0	0	0	0	0	2	0	0	0
AIDS	0	120	0	0	0	0	0	0	0	0	0	0	0
Acne	0	0	120	0	0	0	0	0	0	0	0	0	0
Alcoholic hepatitis	0	0	0	116	0	0	0	0	0	0	0	0	0
Allergy	0	0	0	0	120	0	0	0	0	0	0	0	0
Arthritis	0	0	0	0	0	115	0	0	0	0	0	0	0
Bronchial Asthma	0	0	0	0	0	0	118	0	0	0	0	0	0
Cervical spondylosis	0	0	0	0	0	0	0	119	0	0	0	0	0
Chicken pox	0	0	0	0	0	0	0	0	120	0	0	0	0
Chronic cholestasis	0	0	0	2	0	0	0	0	0	118	0	0	0
Common Cold	0	0	0	0	0	0	0	0	0	0	120	0	0
Dengue	0	0	0	0	0	0	0	0	0	0	0	113	0
Diabetes	0	0	0	0	0	0	0	0	0	0	0	0	120
Dimorphic hemmorhoids(piles)	0	0	0	0	0	0	0	0	0	0	0	0	0
Drug Reaction	0	0	0	0	0	0	0	0	0	0	0	0	0
Fungal infection	0	0	0	0	0	0	0	0	0	0	0	0	0
GERD	0	0	0	0	0	0	0	0	0	0	0	0	0
Gastroenteritis	0	0	0	0	0	0	0	0	0	0	0	0	0
Heart attack	0	0	0	0	0	0	0	0	0	0	0	0	0
Hepatitis B	0	0	0	0	0	0	0	0	0	0	0	0	0
Hepatitis C	0	0	0	0	0	0	0	0	0	0	0	0	0
Hepatitis D	0	0	0	0	0	0	0	0	0	0	0	0	0
Hepatitis E	0	0	0	0	0	0	0	0	0	0	0	0	0
Hypertension	0	0	0	0	0	0	0	0	0	0	0	0	0
Hyperthyroidism	0	0	0	0	0	0	0	0	0	0	0	0	0
Hypoglycemia	1	0	0	0	0	0	0	0	0	0	0	0	0
Hypothyroidism	0	0	0	0	0	0	0	0	0	0	0	0	0
Impetigo	0	6	0	0	0	0	0	0	0	0	0	0	0
Jaundice	0	0	0	0	0	0	2	0	0	0	0	0	0
Malaria	0	0	0	0	0	0	0	0	0	0	0	0	0
Migraine	0	0	0	0	0	1	0	0	0	0	0	0	0
Osteoarthristis	0	0	0	0	0	8	0	0	0	0	0	0	0
Paralysis (brain hemorrhage)	0	0	0	0	0	0	0	0	0	0	0	0	0
Peptic ulcer diseae	0	0	0	2	0	0	0	0	0	0	0	0	0
Pneumonia	0	0	0	0	0	0	0	0	0	0	0	0	0
Psoriasis	0	0	0	0	0	0	0	0	0	0	0	0	0
Tuberculosis	0	0	0	0	0	0	0	0	0	0	0	0	0
Typhoid	0	0	0	0	0	0	0	0	0	0	0	0	0
Urinary tract infection	0	0	0	0	0	0	0	0	0	0	0	0	0
Varicose veins	0	0	0	0	0	0	0	0	0	0	0	0	0
hepatitis A	0	0	0	0	0	0	0	0	0	0	0	0	0
Σ	119	126	120	120	120	124	120	119	120	120	120	113	120

Рисунок 27. Матрица путаницы

Как видно из Рисунок 27. Матрица путаницы, алгоритм дерево решений неправильно классифицировал некоторые заболевания, например в 6 случаев импетиго классифицировал как СПИД, а 8 случаев остеоартрита классифицировал как артрит.

Неправильно классифицированные случаи отправим в виджет «таблица данных» (Рисунок 28*)*. Таким образом из 4920 экземпляров, алгоритм дерево решений неправильно классифицировал 56 экземпляров.

Data Table

Data instances: 56
Features: 133
Meta attributes: 1
Target: Class 'prognosis'

	prognosis	prognosis(Tree)
1	Hepatitis D	Hepatitis C
2	Alcoholic hepatitis	Peptic ulcer diseae
3	Hepatitis D	Hepatitis C
4	Alcoholic hepatitis	Peptic ulcer diseae
5	Osteoarthristis	Arthritis
6	Osteoarthristis	Arthritis
7	Jaundice	Bronchial Asthma
8	Dengue	Malaria
9	Hepatitis D	Hepatitis C
10	Osteoarthristis	Arthritis
11	Hepatitis D	Hepatitis C
12	Jaundice	Bronchial Asthma
13	Impetigo	AIDS
14	Chronic cholestasis	Alcoholic hepatitis
15	Osteoarthristis	Arthritis
16	Hepatitis D	Hepatitis C
17	(vertigo) Paroymsal Positional Vertigo	Chronic cholestasis
18	Chronic cholestasis	Alcoholic hepatitis
19	(vertigo) Paroymsal Positional Vertigo	Chronic cholestasis
20	Cervical spondylosis	Osteoarthristis
21	Bronchial Asthma	Jaundice
22	Bronchial Asthma	Jaundice
23	Alcoholic hepatitis	Peptic ulcer diseae
24	Arthritis	Osteoarthristis
25	Alcoholic hepatitis	Peptic ulcer diseae
26	Arthritis	Osteoarthristis
27	Malaria	Gastroenteritis
28	Osteoarthristis	Arthritis
29	Impetigo	AIDS
30	Dengue	Hepatitis D
31	Hepatitis D	Hepatitis C
32	Tuberculosis	GERD
33	Dengue	Hepatitis D
34	Dengue	Malaria
35	Tuberculosis	GERD
36	Osteoarthristis	Arthritis
37	Impetigo	AIDS
38	Malaria	Gastroenteritis
39	Hepatitis D	Hepatitis C
40	Arthritis	Osteoarthristis
41	Arthritis	Osteoarthristis
42	GERD	Heart attack
43	Dengue	Malaria
44	Impetigo	AIDS
45	GERD	Heart attack
46	Arthritis	Osteoarthristis
47	Peptic ulcer diseae	Alcoholic hepatitis
48	Peptic ulcer diseae	Alcoholic hepatitis
49	Migraine	Arthritis
50	Impetigo	AIDS
51	Hypoglycemia	(vertigo) Paroymsal Positional Vertigo
52	Dengue	Malaria
53	Osteoarthristis	Arthritis
54	Osteoarthristis	Arthritis
55	Dengue	Malaria
56	Impetigo	AIDS

Рисунок 28. Неправильно классифицированные экземпляры

Основные результаты и выводы

В данной работе автором проанализирован ряд алгоритмов, позволяющих поддерживать принятие врачебных решений с большим количеством атрибутов, которые определяют принадлежность к определенной метке класса. Исследование принадлежности атрибутов к определенному заболеванию позволяет сделать следующие выводы:

а) Алгоритм дерева решений может проводить классификацию заболеваний по разным симптомам, но учитывает при классификации не все имеющиеся в конкретном случае симптомы. В некоторых случаях алгоритм дерева решений классифицирует заболевание не по наличию симптомов, а их отсутствию, поскольку это обеспечивает минимальное значение энтропии.

б) Алгоритм логистической регрессии при визуализации дает представление о влиянии каждого симптома заболевания на его классификацию.

в) Результаты применения алгоритма наивного Байеса являются содержательно интерпретируемыми и обоснованными.

г) Совокупность алгоритмов поддержки принятия решений способна повысить эффективность решения инновационных задач при обработке медицинской информации.

Созданная диагностическая модель ансамбля алгоритмов машинного обучения на тестовых данных показала, что все исследуемые алгоритмы дали правильное прогнозирование диагнозов заболевания.

Тестирование комплексной диагностической модели показало, что точность классификации алгоритма дерево решений составил 0,989, но точность классификации алгоритмов логистической регрессии и наивного Байеса составила 1,000.

По времени обучения алгоритм наивный Байес показывает лучший результат – 0,70 секунды. Время обучения алгоритма Дерево решений около 18 секунд, а алгоритма логистической регрессии около 66 секунд.

По времени тестирования лучший результат у алгоритма Дерево решений – 0,005 секунды. Время тестирования алгоритма наивный Байес 0,224 секунды, а алгоритма логистическая регрессия – 1,855 секунды.

В этом исследовании были сделаны прогнозы диагнозов заболеваний по симптомам.

В ходе исследования было проведено сравнение между различными алгоритмами, которые могли бы быть использованы в системах поддержки принятия врачебных решений.

Из полученных результатов видно, что алгоритм наивный Байес генерирует наилучший результат среди исследуемых алгоритмов.

Список литературы

1. A correlation-based feature analysis of physical examination indicators can help predict the overall underlying health status using machine learning / H. Wang, P. Shuai, Y. Deng [et al.] // Sci. Rep. - 2022. – Vol. 12, №1. - DOI: 10.1038/s41598-022-20474-3.
2. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. / A. Dinh, S. Miertschin, A. Young, S. D. Mohanty // BMC Med. Inform. Decis. Mak. - 2019. – Vol.19, №211. -DOI: 10.1186/s12911-019-0918-5.
3. A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology / L. Coombs, A. Orlando, X. Wang [et al.] // npj Digital Medicine. – 2022. – Vol.117, №2022. - DOI:10.1038/s41746-022-00660-3.
4. A machine learning-based framework to identify type 2 diabetes through electronic health records. / T. Zheng, W. Xie, L. Xu [et al.] // Int. J. Med. Inform. – 2017. – Vol.97. – P.120-127. - DOI: 10.1016/j.ijmedinf.2016.09.014.
5. Araújo, F. H. D. Using machine learning to support healthcare professionals in making preauthorisation decisions. / F. H. D. Araújo, A. M. Santana, P. de A. Santos Neto // Int. J. Med. Inform. – 2016. – Vol.94. – P.1-7. - DOI: 10.1016/j.ijmedinf.2016.06.007.
6. Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods / S. Abhari, S. R. N. Kalhori, M. Ebrahimi, [et al.] // Healthc. Inform. Res. – 2019. – Vol.25, №4. – P.248-261. - DOI 10.4258/hir.2019.25.4.248.
7. Beam, A. L. Big Data and machine learning in health care / A. L. Beam, I. S. Kohane // JAMA. – 2018. – Vol.319, №13. – P.1317-1318. - DOI:10.1001/jama.2017.18391.
8. Building risk prediction models for type 2 diabetes using machine learning techniques / Z. Xie, O. Nikolayeva, J. Luo, D. Li // Preventing Chronic Disease. – 2019. – Vol.16. - DOI: 10.5888/pcd16.190109.
9. Chen, A. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data / A. Chen, D. O. Chen // Sci. Rep. – 2022. – Vol.12, №1. - DOI: 10.1038/s41598-022-23011-4.
10. Clifford Lynch. How do your data grow? / C. Lynch // Nature. - 2008. – Vol.455, №7209. – P.28-29.
11. Clinical decision support systems for the practice of evidence-based medicine / I Sim, P Gorman, R A Greenes [et al.] // J. Am. Med. Inform. Assoc. – 2001. – Vol.8, №6. – P.527–534. - DOI:10.1136/jamia.2001.0080527.
12. Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes / J. A.

Carter, C. S. Long, B. P. Smith [et al.] //Expert Systems with Applications. – 2019. – Vol.115. – P.245-255 – DOI: 10.1016/j.eswa.2018.08.002.
13. Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications. / K. Dalakleidi, K. Zarkogianni, A. Thanopoulou, K. Nikita // Expert Systems. – 2017. - Vol.34, №6. - DOI:10.1111/exsy.12214.
14. Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine/ N. Nirala, R. Periyasamy, B. K. Singh, A. Kumar // Biocybernetics and Biomedical Engineering. – 2019. – Vol.39, №1 – P.38-51 - DOI:10.1016/j.bbe.2018.09.007.
15. Diagnostic method of diabetes based on support vector machine and tongue images / J. Zhang, J. Xu, X. Hu [et al.] // Biomed. Res. Int. – 2017. – Vol.2017, №7961494 - DOI: 10.1155/2017/7961494.
16. Doguc, O. Recent applications of data mining in medical diagnosis and prediction / O. Doguc, Z. N. Canbolat, G. Silahtaroglu // Big Data Analytics for Healthcare. – 2022. – P.95-109. - DOI:10.1016/B978-0-323-91907-4.00006-6.
17. Driving type 2 diabetes risk scores into clinical practice: performance analysis in hospital settings / A. Martinez-Millana, M. Argente-Pla, B. V. Martinez [et al.] // J. Clin. Med. – 2019. - Vol.8, №1. -DOI:10.3390/jcm8010107.
18. Early metabolic markers identify potential targets for the prevention of type 2 diabetes / G. Peddinti, J. Cobb, L. Yengo [et al.] // Diabetologia. – 2017. – Vol.60, №9. – P.1740–1750. - DOI: 10.1007/s00125-017-4325-0.
19. Fractional lévy stable motion: finite difference iterative forecasting model / H. Liu, W. Song, M. Li [et al.] // Chaos, Solitons & Fractals. – 2020. – Vol.133. – DOI: 10.1016/j.chaos.2020.109632.
20. Genetic risk score increased discriminant efficiency of predictive models for type 2 diabetes mellitus using machine learning: Cohort Study / Y. Wang, L. Zhang, M. Niu [et al.] // Front. Public. Health. – 2021. – Vol.9. - DOI: 10.3389/fpubh.2021.606711.
21. Hassan, A. Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: a retrospective cross-sectional study / A.Hassan, T.Khan // IEEE Access 8. – 2020. - DOI:10.1109/ACCESS.2020.3035026.
22. Hosmer, D. W. Applied logistic regression / D. W. Hosmer, S. Lemeshow. - John Wiley & Sons, Inc, 2000. - DOI:10.1002/0471722146.
23. How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach / D. Ichikawa, T. Saito, W. Ujita, H. Oyama // J. Biomed. Inform. – 2016. – Vol.64. – P.20-24. - DOI: 10.1016/j.jbi.2016.09.012.

24. Husam, E. Prediction of chronic kidney disease using data mining techniques / E. Husam, Y. A. M. A. Al-Abiary, M. F. A. Kadir // International Journal of Special Education. – 2022. – Vol.37, №3. – P. 9534-9544 – DOI: /10.2139/ssrn.4022160.
25. In silico prediction of gamma-aminobutyric acid type-a receptors using novel machine-learning-based SVM and GBDT approaches / Z. Liao, Y. Huang, X. Yue [et al.] // Biomed. Res. Int. – 2016. – Vol.2016, № 2375268. - DOI: 10.1155/2016/2375268.
26. Jowkar, G.-H. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification / G.-H. Jowkar, E. G. Mansoori // Comput. Biol. Chem. – 2016.- Vol.64. – P.263-270. - DOI: 10.1016/j.compbiolchem.2016.07.004.
27. Karim, M. Decision Tree and Naïve Bayes algorithm for classification and generation of actionable knowledge for direct marketing / M. Karim, R. M. Rahman // Journal of Software Engineering and Applications. – 2013. – Vol.6, №4. – P.196-206. - DOI: 10.4236/jsea.2013.64025.
28. Kaur, A. Feature selection in machine learning: methods and comparison // A. Kaur, K. Guleria, N. K. Trivedi // 2021 International Conference on Advance Computing andInnovative Technologies inEngineering (ICACITE 2021). – 2021. – P.789-795. - DOI:10.1109/ICACITE51222.2021.9404623.
29. Kaur, H. Predictive modelling and analytics for diabetes using a machine learning approach / H. Kaur, V. Kumari // Applied Computing and Informatics. – 2022. - DOI:10.1016/j.aci.2018.12.004.
30. King, G. Logistic Regression in rare events data / G. King, L. Zeng // Political Analysis. – 2001. – Vol.9, №2. – P.137-163. - DOI: 10.1093/oxfordjournals.pan.a004868.
31. Kose, U. Deep learning for medical decision support systems / U. Kose, O. Deperlioglu, J. Alzubi, B. Patrut. – Springer,2021. – P.171
32. Lee, B.J. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning / B. J. Lee, J. Y. Kim // IEEE J. Biomed. Health. Inform. – 2016. – Vol.20, №1. – P. 39-46 - DOI: 10.1109/JBHI.2015.2396520.
33. Luo, G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction / G. Luo // Health. Inf. Sci. Syst. - 2016. – Vol. 4 - DOI: 10.1186/s13755-016-0015-4.
34. Machine learning and data mining methods in diabetes research / I. Kavakiotis, O. Tsave, A. Salifoglou [et al.] // Comput. Struct. Biotechnol J. – 2017. – Vol.15. – P.104-116. - DOI: 10.1016/j.csbj.2016.12.005.

35. Machine learning for tuning, selection, and ensemble of multiple risk scores for predicting type 2 diabetes / Y. Liu, S. Ye, X. Xiao, [et al.] // Risk Manag. Healthc. Policy. – Vol.12. – P.189—198. – DOI: 10.2147/RMHP.S225762.
36. Machine learning methods to predict diabetes complications / A. Dagliati , S. Marini , L. Sacchi [et al.] //J. Diabetes Sci. Techno l – 2018. – Vol.12, №2. – P.295-302. - DOI: 10.1177/1932296817706375.
37. Meng, J. Lightweight relevance filtering for machine-generated resolution problems / J. Meng, L. C. Paulson // Journal of Applied Logic. – 2009. – Vol.7, №1. – P.41-57. – DOI:10.1016/j.jal.2007.07.004.
38. Mišić, V.V. A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission / V. Mišić, K. Rajaram, E. Gabel // npj Digital Medicine. – 2021. – Vol.98, №2021. - DOI:10.1038/s41746-021-00468-7.
39. Nezhad, S. N. Detecting diseases in medical prescriptions using data mining methods / S. N. Nezhad, M. H. Zahedi, E. Farahani // BioData Min. – 2022. – Vol.15, №1. - DOI: 10.1186/s13040-022-00314-w.
40. Ontology driven decision support for the diagnosis of mild cognitive impairment / X. Zhang, B. Hu, X. Ma [et al.] //Comput. Methods Programs Biomed. – 2014. – Vol.113, №3. – P.781-791. - DOI: 10.1016/j.cmpb.2013.12.023.
41. Palaniappan, S. Intelligent heart disease prediction system using data mining techniques / S. Palaniappan, R. Awang // International Journal of Computer Science and Network Security. – 2008. - №8. – P.343-350. - DOI: 10.1109/AICCSA.2008.4493524.
42. Peker, M. Use of Orange Data Mining toolbox for data analysis in clinical decision making: the diagnosis of diabetes disease / M. Peker, O. Özkaraca, A. Şaşar // Expert System Techniques in Biomedical Science Practice. – 2018. – P.25. - DOI: 10.4018/978-1-5225-5149-2.ch007.
43. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model / A. Cahn, A. Shoshan, T. Sagiv [et al.] //Diabetes Metab. Res. Rev. – 2020. – Vol.36, №2. - DOI: 10.1002/dmrr.3252.
44. Predictive models for conversion of prediabetes to diabetes / N. Yokota, T. Miyakoshi, Y. Sato [et al.] //J. Diabetes Complications. – 2017. – Vol.31, №8. – P.1266-1271. - DOI: 10.1016/j.jdiacomp.2017.01.005.
45. Remaining useful life prediction for lithium-ion batteries using fractional brownian motion and fruit-fly optimization algorithm / H. Wang, W. Song, E. Zio [et al.] // Measurement. – 2020. – Vol.161. – DOI: 10.1016/j.measurement.2020.107904.
46. Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis / J. J. Boutilier, T. C. Y. Chan, M.

Ranjan, S. Deo // J. Med. Internet Res. – 2021. – Vol.23, №1. - DOI: 10.2196/20123.
47. Samant, P. Machine learning techniques for medical diagnosis of diabetes using iris images / P. Samant, R. Agarwal // Comput. Methods. Programs Biomed. – 2018. – Vol.157. – P.121-128 - DOI: 10.1016/j.cmpb.2018.01.004.
48. Shortliffe, E. Computer-Based Medical Consultations: MYCIN / E. Shortliffe // Journal of Clinical Engineering. - 1976. – DOI:10.1097/00004669-197610000-00011.
49. Sumathi, A. Semi supervised data mining model for the prognosis of pre-diabetic conditions in type 2 diabetes mellitus / A. Sumathi, S. Meganathan // Bioinformation. – 2019. – Vol.15, №12. – P.875-881. - DOI:10.6026/97320630015875.
50. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes / J.V.Tu // J. Clin. Epidemiol. – 1996. – Vol.49, №11. – P.1225-1231. - DOI: 10.1016/s0895-4356(96)00002-9.
51. Wasan, S. K. The impact of data mining techniques on medical diagnostics / S. K. Wasan,V. Bhatnagar, H. Kaur // Data Science Journal. – 2006. – Vol.5. – P.119-126. - DOI: 10.2481/dsj.5.119.
52. Williams, N. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification / N. Williams, S. Zander, G. Armitage // Computer Communication Review. – 2006. – Vol.36, №5. – P.5-16. – DOI: 10.1145/1163593.1163596.

АНТИПЛАГИАТ
ОБНАРУЖЕНИЕ ЗАИМСТВОВАНИЙ

СПРАВКА

о результатах проверки текстового документа
на наличие заимствований

Башкирский государственный медицинский университет

ПРОВЕРКА ВЫПОЛНЕНА В СИСТЕМЕ АНТИПЛАГИАТ.ВУЗ

Автор работы: Серегин Владимир Сергеевич
Самоцитирование рассчитано для: Серегин Владимир Сергеевич
Название работы: ОПТИМИЗАЦИОННЫЕ МОДЕЛИ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ
Тип работы: Выпускная квалификационная работа
Подразделение: ФГБОУ ВО БГМУ Минздрава России

РЕЗУЛЬТАТЫ

СОВПАДЕНИЯ	5.28%
ОРИГИНАЛЬНОСТЬ	82.81%
ЦИТИРОВАНИЯ	11.92%
САМОЦИТИРОВАНИЯ	0%

ДАТА ПОСЛЕДНЕЙ ПРОВЕРКИ: 13.06.2023

Структура документа: Проверенные разделы: основная часть с.1, 3-65, содержание с.2, библиография с.66-68

Модули поиска: ИПС Адилет; Модуль поиска "БГМУ"; Библиография; Сводная коллекция ЭБС; Интернет Плюс*; Сводная коллекция РГБ; Цитирование; Переводные заимствования (RuEn); Переводные заимствования по eLIBRARY.RU (EnRu); Переводные заимствования по Интернету (EnRu); Переводные заимствования издательства Wiley ; eLIBRARY.RU; СПС ГАРАНТ: аналитика; СПС ГАРАНТ: нормативно-правовая документация; Медицина; Диссертации НББ; Коллекция НБУ; Перефразирования по eLIBRARY.RU; Перефразирования по СПС ГАРАНТ: аналитика; Перефразирования по Интернету; Перефразирования по Интернету (EN); Перефразирования по коллекции издательства Wiley ; Патенты СССР, РФ, СНГ; СМИ России и СНГ; Шаблонные фразы; Кольцо вузов; Издательство Wiley; Переводные заимствования

Работу проверил: Кобзева Наталья Рудольфовна
ФИО проверяющего

Дата подписи: 13.06.2023

ФГБОУ ВО БГМУ
Минздрава России
НАУЧНАЯ БИБЛИОТЕКА
Подпись проверяющего

Чтобы убедиться в подлинности справки, используйте QR-код, который содержит ссылку на отчет.

Ответ на вопрос, является ли обнаруженное заимствование корректным, система оставляет на усмотрение проверяющего. Предоставленная информация не подлежит использованию в коммерческих целях.