МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ВИЗУАЛИЗАЦИЯ ДАННЫХ И ПРОГНОЗОВ В ЗАДАЧЕ ВЫЯВЛЕНИЯ КАРДИОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ У ЧЕЛОВЕКА НА ОСНОВЕ ФИЗИОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Никифорова Яна Николаевича

Научный руководитель
доцент, к. ф.-м. н. _____ О. В. Мещерякова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2023

ВВЕДЕНИЕ

Развитие информационных технологий обусловлено задачами и вовлеченностью людей в экономические и социальные процессы. Это позволяет собирать, обрабатывать и передавать большие объемы данных. Большие объемы данных требуют новых методов и инструментов. Интеллектуальный анализ данных (Data mining) позволяет извлекать ценные знания из больших объемов данных и применяется в различных областях, включая экономику, политику, социальные сети и здравоохранение.

В здравоохранении на данный момент уже применяются инструменты статистического анализа и Data Mining для выявления скрытых заболеваний и постановки верных диагнозов. Медицинские данные содержат ценные сведения, которые могут быть использованы для извлечения скрытых закономерностей. Системы Data Mining помогают в выборе средств лечения на основе правил, описывающих сочетания симптомов различных заболеваний. Прогностические модели и анализ данных могут быть полезны в борьбе с заболеваниями.

Data mining, становится ценным инструментом. Он позволяет анализировать данные, собранные от пациентов, включая клинические исследования, результаты обследований, медицинские карты и другие сведения, чтобы прогнозировать риски и разрабатывать индивидуализированные подходы к лечению и профилактике сердечно-сосудистых заболеваний.

Одной из задач интеллектуального анализа данных является построение модели дерева решений, которая широко используются для задач классификации и регрессии благодаря их способности представлять сложные процессы принятия решений иерархическим образом. Для удобства специалистов всех отраслей, модели, построенные с помощью алгоритмов, представлены в виде деревьев решений. Визуализация, как средство представления данных в визуальном формате, была признана мощным инструментом для облегчения понимания и интерпретации сложных структур данных, включая модели дерева решений.

Целью дипломной работы является разработка инструмента визуализации деревьев решений, которые используются в анализе данных и машинном обучении при решении задач построения прогнозов предрасположения и наличия кардиологических заболеваний. Для реализации данной цели были

решены следующие задачи:

— Сбор данных.
— Подготовка данных для анализа.
— Выбор и реализация модели Data Mining, ее обучение на полученных данных.
— Анализ производительности модели и ее поведения.

Источники исследования включают книги, статьи и электронные ресурсы, охватывающие различные аспекты интеллектуального анализа данных, баз данных, медицинских датасетов. Они предоставляют информацию о методах, техниках, инструментах и приложениях в области анализа и использования данных. Эти материалы являются ценным источником знаний для исследователей, профессионалов и студентов, которые интересуются разработкой и применением алгоритмов, моделей и инструментов в области больших данных и аналитики.

Структура работы охватывает введение, 3 основные части (теоретические основы интеллектуального анализа данных, применение модели для прогнозирования кардиологических заболеваний, практическую часть с подготовкой данных и анализом модели), а также заключение, список использованных источников и четырех приложений, которые содержат информацию о библиотеках и реализации дерева решений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе (Основные понятия, классификации, инструменты) рассматриваются актуальность задачи прогнозирования сердечно-сосудистых заболеваний, возможность использований современных аналитических средств для анализа данных, и применение популярных инструментов моделирования и анализа. Таким средством стала технология Data Mining. Важный аспект Data Mining, включающий задачу классификации, визуализацию данных и метод деревьев решений.

Data Mining – это процесс извлечения ценной информации из больших объемов данных. Он использует различные методы анализа и статистики для нахождения скрытых закономерностей и шаблонов в данных. Data Mining отличается от традиционных методов анализа, таких как статистика и OLAP, тем, что сам формулирует гипотезы о взаимосвязях в данных и может предска-зывать будущие события. Важным элементом успеха в Data Mining является качественная подготовка данных. Он представляет собой мощный инструмент для принятия обоснованных решений на основе скрытых закономерностей и шаблонов в данных.

Данные – это информация, представленная в различных формах, таких как факты, текст, графики, изображения, звуки, видео. Они могут быть получены измерениями, экспериментами и математическими операциями. Для удобства хранения, передачи и обработки данных необходимо их представить в удобной форме. Данные представляются в виде таблицы "объект-атрибут где объекты расположены в горизонтальной части таблицы, а их атрибуты – в вертикальной части. Атрибуты представляют свойства объектов, такие как цвет волос или высота здания. При анализе данных используется выборка – небольшой набор данных из генеральной совокупности. Для анализа данных также используются различные типы шкал измерений, такие как номинальная, порядковая, интервальная, относительная и дихотомическая. Данные могут быть хранены в базах данных, а классификация данных может основываться на их типе, постоянстве значений, функциях и периоде охвата.

Data Mining – процесс, состоящий из нескольких этапов, включающих анализ предметной области, постановку задачи, подготовку данных, построение и обучение модели, проверку и оценку модели, применение модели, а

также коррекцию и обновление модели.

Data Mining объединяет математический инструментарий и информационные технологии, включая различные методы анализа данных, такие как нейронные сети, деревья решений, кластерный анализ, поиск ассоциативных правил и другие.

Классификация является одной из самых простых и распространенных задач Data Mining. Она представляет собой системное распределение изучаемых объектов по родам, видам или типам на основе существенных признаков. Классификация может быть простой, когда деление основано на одном признаке, или сложной, когда используются разные признаки и синтезируются в единое целое.

Для выполнения классификации необходимо иметь формальное описание объекта, обучающее множество (содержащее входные и выходные значения примеров) и тестовое множество (для оценки качества модели).

Деревья решений – это структура данных и метод машинного обучения, используемый для классификации и прогнозирования объектов на основе их характеристик. Дерево решений имеет древовидную структуру, где каждый узел представляет собой тест на один из признаков, ветви соответствуют значениям признака, а листовые узлы содержат предсказания или классификации. Деревья решений могут быть использованы для классификации и регрессии. Они могут быть двоичными или многовариантными, а также могут быть построены индуктивно или с использованием обучения с подкреплением. Простые деревья имеют небольшую глубину и простую структуру, в то время как сложные деревья имеют большую глубину и сложную структуру. Деревья решений позволяют интерпретировать данные, объяснять причину принятия решений и обрабатывать пропущенные значения. Они могут быть построены на больших базах данных и обучаются быстрее, чем нейронные сети. Критерии расщепления, такие как мера энтропии, используются для выбора наилучшего атрибута расщепления и улучшения прогнозирования.

Во второй главе (Построение и анализ модели для прогнозирования наличия кардиологических заболеваний у человека) исследуется возможность построения модели Data Mining для предсказания наличия кардиологических заболеваний у пациентов на основе их физиологических и лабораторных показателей. Цель работы заключается в улучшении ранней диагности-

ки и принятии обоснованных решений в области кардиологии. Для анализа используется открытый набор данных "Сердечно-сосудистые заболевания"с информацией о пациентах и наличии или отсутствии заболеваний. Набор данных содержит 70000 образцов с 14 признаками, включая возраст, пол, артериальное давление, уровень холестерина, уровень глюкозы, курение, уровень алкоголя, физическую активность и наличие сердечно-сосудистого заболевания. Для построения и анализа модели используется язык программирования Python3 и соответствующие библиотеки для обработки данных, вычислений, машинного обучения и визуализации результатов. Исходный набор данных разделяется на тренировочную и тестовую выборки, где тренировочная выборка используется для обучения модели, а тестовая выборка – для проверки ее производительности на новых данных. Для улучшения модели применяется поиск оптимальных значений гиперпараметров, таких как глубина и ширина дерева, минимальное количество образцов в листе и количество листьев. После обучения модели на тренировочной выборке, ее производительность оценивается на тестовой выборке путем сравнения фактических и предсказанных значений показателей.

В третьей главе (Практическая часть) описывается практическая часть работы с набором данных о физиологических и лабораторных показателях пациентов. Изначально выполняется загрузка данных из файла с помощью библиотеки Pandas и сохранение в переменную df. Для ознакомления с данными и с их структурой на экран выводятся первые строки данных.

Далее проведена предварительная обработка данных, включающая проверку наличия пропущенных значений и дубликатов. Метод isnull применяется к набору данных для выявления пропущенных значений, а метод duplicated используется для поиска дубликатов строк.

Анализ пропущенных значений и дубликатов является важным этапом предварительной обработки данных, поскольку пропуски могут указывать на ошибки или проблемы с данными, а дубликаты могут привести к неоднозначности или проблемам с записью данных.

Во время предварительной обработки данных обнаружился факт наличия выбросов в некоторых строках данных, например, аномально высокое диастолическое давление. Для удаления выбросов устанавливаются пороговые значения на основе среднего значения и стандартного отклонения призна-

ков. После удаления выбросов из данных выводится количество удаленных значений.

Для получения сводной статистической информации о данных используется функция describe, которая позволяет оценить количество непропущенных значений, среднее, стандартное отклонение, минимальное и максимальное значения, а также процентили для каждого числового признака.

Также построятся гистограммы для каждого признака с помощью метода hist. Гистограммы позволяют визуально оценить форму распределения данных, центральную тенденцию и вариацию.

В результате предварительного анализа данных установлено, что набор данных подходит для обучения алгоритма дерева решений и последующей визуализации.

Следующим этапом является построение модели и ее обучение модели с использованием решающих деревьев для задачи классификации. Изначально данные разделяются на тренировочную и тестовую выборки. В работе выбрана модель DecisionTreeClassifier, которая обладает простотой интерпретации, универсальностью и способностью работать с различными типами данных.

Метод train_test_split используется для разделения данных, определение признаков и целевой переменной, а также сохранение разделенных данных в соответствующие переменные. Для обучения модели выбрана DecisionTreeClassifier, который обладает преимуществами простой интерпретации, универсальности и способности работать с разными типами данных. Однако, у этой модели также есть ограничения, такие как склонность к переобучению и неустойчивость к небольшим изменениям в данных.

Для улучшения точности модели и ее более тонкой настройки используются набор параметров, такие как критерий разделения признаков, максимальная глубина дерева, минимальное количество образцов наблюдений и максимальное количество узлов в дереве. Для нахождения оптимальных значений параметров используется метод случайного поиска (RandomizedSearchCV) в сочетании с кросс-валидацией. Этот подход позволяет автоматически настраивать параметры модели, улучшая ее производительность и гибкость.

После настройки параметров модель обучается на тренировочных данных и выполняется прогноз целевой переменной на тестовых данных. Точность модели вычисляется с помощью функции accuracy_score, которая срав-

нивает прогнозированные значения с истинными значениями целевой переменной. Это позволяет оценить, насколько хорошо модель справилась с предсказанием на новых данных. На выбранном наборе данных точность модели составляет 78.91%. Это означает, что модель правильно классифицирует достаточно большое количество тестовых примеров.

Заключительным этапом работы является визуализация. Для визуализации дерева решений использованы две библиотеки – dtreeviz и sklearn.tree.

Библиотеки dtreeviz и sklearn.tree используются для визуализации дерева решений. Dtreeviz позволяет создавать наглядные графические представления деревьев решений с признаками и значениями ветвей. Метод model из библиотеки dtreeviz используется для визуализации, принимая модель, значения признаков, целевую переменную и другие аргументы. Результатом является инфографика, показывающая процесс принятия решений и классификационные результаты. Инфографика помогает анализировать и интерпретировать результаты, сравнивая их с гипотезами. Визуализация показывает выбранные критерии решений и как они разделяют наблюдения. Ветви дерева указывают на значения, а листовые узлы показывают классы и прогнозы модели.

Второй способ визуализации дерева решений – использование метода tree.plot_tree из модуля sklearn. Аргументы, передаваемые этому методу, включают объект модели с оптимальными параметрами (best_model), список имен признаков (feature_names) и список имен классов (class_names). Метод создает графическую визуализацию дерева с узлами и ветвями, показывая условия принятия решений и значения признаков. Закрашивание узлов позволяет выделить преобладающий класс. Этот метод требует меньше вычислительных ресурсов, но может быть менее наглядным по сравнению с предыдущим способом.

Текстовое представление дерева решений, созданное с помощью функции tree.export_text, является третьим способом визуализации. Оно содержит информацию о структуре дерева, условиях разделения и классах, связанных с каждым узлом. Текстовое представление полезно для интерпретации модели, обмена информацией и проверки правильности модели. Оно может быть особенно полезным для больших и сложных деревьев.

ЗАКЛЮЧЕНИЕ

Сердечно-сосудистые заболевания лидируют среди причин смертности молодых людей (от 25 до 64 лет) в России. Главное коварство этой группы заболеваний в том, что нередко они протекают бессимптомно, и человек узнают о том, что у него есть проблемы с сердцем тогда, когда ему уже нельзя помочь.

Поэтому особую актуальность приобретают методы ранней диагностики заболеваний, обнаружения скрытых факторов, влияющих на предрасположенность к таким заболеваниям. Кроме специальных медицинских исследований существуют примеры использования новых технологий и приемов выявления заболевания или обнаружение предрасположенности к ним. К ним относятся и аналитические методы, а именно, анализ данных и машинное обучение.

Внедрение методов машинного обучения и анализа данных в кардиологические исследования может принести следующие практические преимущества: Диагностика и прогнозирование: модели машинного обучения могут помочь в определении риска возникновения сердечно-сосудистых заболеваний, выявлении и классификации различных типов заболеваний, а также прогнозировании их прогрессирования и результатов лечения.

Оптимизация лечения: анализ данных пациентов с помощью методов машинного обучения может помочь определить оптимальные методы лечения и терапии для индивидуальных пациентов. Это может включать выбор наиболее эффективных лекарственных препаратов, дозировок и методов лечения, основанных на предсказанных результатах и реакции пациента на терапию.

Разработка новых диагностических и прогностических моделей: применение методов машинного обучения позволяет выявлять новые паттерны и связи в больших объемах данных, что может привести к разработке новых диагностических и прогностических моделей.

Предупреждение и предотвращение: модели машинного обучения могут помочь в предупреждении развития сердечно-сосудистых заболеваний путем идентификации ранних признаков и факторов риска.

Все эти факторы делают исследование в области кардиологии с применением методов машинного обучения важным и перспективным направлени-

ем, способствующим более точной диагностике, более эффективному лечению и снижению заболеваемости и смертности от сердечно-сосудистых заболеваний.

В дипломной работе сделана попытка построить модель классификации пациентов, обучить ее, визуализировать модель и построить прогноз о наличии заболевания.

В ходе работы был сделан обзор известных моделей для классификации и выбрана модель в виде решений, проведено обучение модели, настройка оптимальных параметров модели, сделан выбор инструментов для анализа и выполнена визуализация дерева решений различными способами.

Поставленные задачи сбора, предварительной обработки данных, построения и реализации модели и ее визуализации выполнены полностью.

Инструменты для исследования в области кардиологии имеет высокую практическую ценность, поскольку кардиологические заболевания являются одной из основных причин смерти и проблем со здоровьем по всему миру.

Основные источники информации:

1 Рафалович В. Data mining, или интеллектуальный анализ данных для занятых / В. Рафалович // SmartBook. 2014. 110 с.

2 Луньков А.Д. Интеллектуальный анализ данных / А.Д. Луньков, А.В. Харламов // Московский технический университет связи и информатики. 2021. 96 с.

3 Michael J.H. Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design / J.H. Michael // Addison-Wesley Professional, 2nd edition. 2003. 611 с.

4 Сьоре Э. Проектирование и реализация систем управления базами данных / Э. Сьоре // ДМК-Пресс. 2021. 466 с.

5 Харрингтон Дж. Разработка баз данных / Дж. Харрингтон // ДМК-Пресс. 2005. 230 с.

6 Гарсиа-Молина Г. Системы баз данных / Г. Гарсиа-Молина, Дж. Ульман, Дж. Уидом // Вильямс. 2003. 1088 с.

7 Абруков В.С. Применение средств интеллектуального анализа данных (data mining) для исследования неполно определенных систем / В.С. Абруков, Я.Г. Николаева, Д.Н. Макаров, А.А. Сергеев, Е.В. Карлович // Вестник Чувашского университета. 2008. 9 с.

8 Павлов Н.А. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике / Н.А. Павлов, А.Е. Андрейченко, А.В. Владзимирский, А.А. Ревазян, Ю.С. Кирпичев, С.П. Морозов // Digital Diagnostics. 2021. 66 с.
9 Hand D. Principles of Data Mining / D. Hand, H. Mannila, P. Smyth // The MIT Press. 2001. 322 с.
10 Jiawei H. Data Mining Concepts and Techniques / H. Jiawei, P. Jian // Morgan Kaufmann Publishers is an imprint of Elsevier. 2012. 740 с.
11 Gutman A.J. Welcoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning / A.J. Gutman, J. Goldmeier // Wiley; 1st edition. 2021. 272 с.
12 Sequeda J. Designing and Building Enterprise Knowledge Graphs / J. Sequeda // Morgan and Claypool publishers. 2021. 142 с.
13 Дюк В.А. Data Mining: учебный курс / В.А. Дюк, А.П. Самойленко — СПб.: Питер, 2001. 185 с.
14 Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод // БХВ-Петербург. 2004. 337 с.
15 Zhiyong Z. Proceedings of the First International Forum on Financial Mathematics and Financial Technology / Z. Zhiyong // Springer; 1st edition. 2021. 248 с.
16 Neil C. MACHINE LEARNING FOR BEGINNERS: a comprehensive guide to understand machine learning. How it works and how is correlated to Artificial Intelligence and Deep Learning / C. Neil // Independently published. 2020. 187 с.
17 Preim B. Visualization, Visual Analytics and Virtual Reality in Medicine: State-of-the-art Techniques and Applications (The MICCAI Society book Series) / B. Preim, R. Raidou, N. Smit, K. Lawonn // Academic Press; 1st edition. 2023. 568 с.
18 Жаркова О.С. Построение систем поддержки принятия решений в медицине на основе деревьев решений / О.С. Жаркова, К.А. Шаропин, А.С. Сеидова, Е.В. Берестнева, И.А. Осадчая // Современные наукоемкие технологии. 2016. С. 33-37.

19 Маккэндлесс Д. Инфографика. Самые интересные данные в графическом представлени / Д. Маккэндлесс // МИФ. 2013. 267 с.
20 Tufte R.E. The Visual Display of Quantitative Information / R.E. Tufte // Graphics Press; 2nd edition. 2001. 191 с.
21 Мошков М.Ю. Деревья решений. Теория и приложения / М.Ю. Мошков // Издательство Нижегородского университета. 1994. 176 с.
22 Goetz T. The Decision Tree: How to make better choices and take control of your health / T. Goetz // Rodale Books. 2011. 336 с.
23 Takefuji Y. Open Source Machine Learning in Medicine / Y. Takefuji // Independently published. 2019. 59 с.
24 Kaggle [Электронный ресурс] URL: https://www.kaggle.com/docs (дата обращения: 01.04.2023)
25 Сердечно-сосудистые заболевания [Электронный ресурс] URL: https://www.kag
disease-dataset?select=cardio_train.csv (дата обращения: 01.04.2023)
26 Лутц М. Программирование на Python / М. Лутц // O'Reilly Media. 2011. 992 с.
27 Albon C. Machine Learning with Python Cookbook / C. Albon // O'Reilly Media. 2018. 366 с.
28 Geron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow / A. Geron // O'Reilly Media. 2017. 856 с.