МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА ДЕСКТОП ПРИЛОЖЕНИЯ ДЛЯ ОЦЕНКИ ВЕРОЯТНОСТИ НАСТУПЛЕНИЯ ПОВТОРНОГО ЖЕЛУДОЧНОГО КРОВОТЕЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование информационных систем
факультета КНиИТ
Алексеевой Марии Олеговны

Научный руководитель
к. ф.-м. н., доцент _____ Ю. Н. Кондратова

Заведующий кафедрой
к.ф.-м.н., доцент _____ С. В. Миронов

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы данной работы в современных условиях обусловлена тем, что традиционные методы диагностики зачастую основываются на субъективной оценке врача, что может приводить к ошибкам, особенно в случаях со сложной симптоматикой. В отличие от них, алгоритмы МО способны анализировать сотни параметров, выявляя даже незначительные корреляции, которые могут указывать на развитие патологии.

Таким образом, цель магистерской работы заключается в разработке Desktop приложения, основанного на алгоритмах машинного обучения и предназначенного для прогнозирования наступления повторного желудочного кровотечения.

Поставленная цель определила следующие задачи:

- изучить информацию в области использования средств машинного обучения в медицине;
- подобрать репрезентативную выборку пациентов;
- провести разведочных анализ и выполнить предобработку выборки;
- подобрать оптимальный алгоритм и провести его обучение;
- разработать Desktop приложение с интегрированным в него обученным алгоритмом;
- создать установщик разработанного приложения для операционной системы Windows.

Научная новизна работы заключается в разработке и внедрении нового подхода к прогнозированию повторного желудочно-кишечного кровотечения, основанного на алгоритмах машинного обучения. Этот подход включает в себя использование оптимального набора признаков, которые описывают клиническую картину предрецидивного симптома, что позволяет значительно повысить точность прогнозирования по сравнению с традиционными методами.

Практически значимой задачей является разработка эффективного инструмента, который поможет в прогнозировании и управлении рисками, связанными с желудочно-кишечными кровотечениями. Это откроет врачам возможность принимать обоснованные решения, минимизируя вероятность субъективных ошибок. В отличие от существующих решений, предлагаемый подход не является инвазивным и разработан на основе оптимального набора признаков, описывающих клиническую картину предрецидивного симптома.

Магистерская работа состоит из введения, 2 разделов, заключения, спис-

ка использованных источников и 5 приложений. Общий объем работы — 73 страницы, из них 62 страницы — основное содержание, включая 28 рисунков и 8 таблиц, цифровой носитель в качестве приложения, список использованных источников информации — 40 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические аспекты применения моделей машинного обучения» посвящен анализу существующей литературы и обзору теоретической составляющей задействованных алгоритмов машинного обучения.

Область здравоохранения является перспективным направлением для внедрения и развития технологий машинного обучения. Применение искусственного интеллекта может принести заметную пользу в задачах систематизации больных и постановке диагноза, улучшить сканирование и сегментацию снимков [1], поддержать принятие решений [2], дать вероятностную оценку риска заболевания [3, 4]. Также алгоритмы МО могут быть интегрированы для решения задач в области нейровизуализации [5].

Одной из первых работ, посвященных использованию алгоритмов машинного обучения в медицине, является статья M. M. Churpek, Yuen T. C., Winslow C. и д.р., опубликованная в 2016 году [6]. В работе описываются варианты прогнозирования остановки сердца, перевода больного в отделение интенсивной терапии и смерти. За основу была взята выборка, состоящая из 269 999 историй болезней пациентов. В данной задаче наилучшую точность классификации показал метод случайного леса: метрика AUC составила 0,8.

Помимо сердечных заболеваний алгоритмы МО, в частности, также исследовались для прогнозирования госпитализации и амбулаторного применения кортикостероидов у пациентов с воспалительными заболеваниями кишечника. Этой теме была посвящена публикация авторов Waljee A. K., Lipson R., Wiitala W. L. и д.р [7]. В работе была использована выборка по 20 368 пациентам с воспалительными заболеваниями кишечника и 351 112 визитам к лечащему врачу, данные датируются периодом 2002-2009 гг. В ходе эксперимента лучший показатель качества показала модель, использующая случайны лес — метрика AUC составила 0,85.

В августе 2020-го года была опубликована работа авторов Weegar R., Sundström K., в которой прогнозирование рака шейки матки осуществлялось при помощи алгоритмов машинного обучения [8]. Выборка для обучения содержала данные ENR о 1723 больных. В качестве признаков использовались записи о лабораторных данных и проведенном лечении. Помимо этого при помощи алгоритмов обработки текста были извлечены признаки из записей EHR. В ходе эксперимента наилучший показатель точности показала модель, исполь-

зующая случайный лес: AUC=0,97.

В результате, алгоритмы машинного обучения демонстрируют высокую прогностическую ценность в области медицины. Среди используемых алгоритмов наилучшую точность показывали модели, основанные на алгоритме случайного леса.

В данной работе исследуется способ оценки вероятности наступления повторного кишечно-желудочного кровотечения, являющегося задачей классификации — одной из основных задач машинного обучения, цель которой заключается в определении принадлежности объектов к одному из заранее определенных классов (категорий).

Для данной задачи были выбраны следующие алгоритмы:

— логистическая регрессия (Logit model).
— метод K-ближайших соседей (KNN);
— метод опорных векторов (SVM);
— наивный байесовский классификатор (Naive Bayes classifier);
— дерево решений (Decision Tree);
— метод случайного леса (Random Forest);
— градиентный бустинг (XGBoost).

Метод K-ближайших соседей принадлежит к семейству алгоритмов, основанных на вычислении оценок сходства между объектами. Для повышения надежности классификации объект относится к тому классу, которому принадлежит большинство его соседей — k-ближайших к нему объектов из обучающей выборки. Расстояние между объектами рассчитывается на основе метрик: Евклидова метрика, Манхэттенское расстояние, расстояние Чебышева.

Метод опорных векторов является одним из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора.

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

Решающее дерево состоит из следующих компонентов: корневой узел, ветви (левая и правая), решающие и листовые (терминальные) узлы. Корневой и решающие узлы представляют из себя вопросы с пороговым значением для разделения тренировочного набора на части (левая и правая), а листья являются конечными прогнозами: среднее значений в листе для регрессии и статистическая мода для классификации.

Случайный лес применяется для задач классификации и регрессии. Он представляет собой улучшение алгоритма решающих деревьев, используя ансамбль деревьев для повышения точности классификации или регрессии.

Наивный байесовский классификатор используется для задач классификации и основанный на теореме Байеса и предполагает, что все признаки являются независимыми друг от друга.

По сути, байесовский классификатор представляет собой вероятностную модель. Пусть задано множество наблюдений, каждое из которых представлено вектором признаков $x = (x_1, x_2, \cdots, x_n)$. Модель присваивает каждому наблюдению условную вероятность $p(C_k|x_1, x_2, \cdots, x_n)$, C_k — класс. Вероятностный классификатор предсказывает класс с самой большой условной вероятностью для заданного вектора признаков x.

Логистическая регрессия использует линейную комбинацию входных признаков и соответствующих весов, которая описывает линейную гиперплоскость в пространстве признаков. Затем этот результат проходит через логистическую функцию, которая переводит линейную комбинацию в вероятность принадлежности объекта к одному из классов.

Градиентный бустинг последовательно объединяет деревья решений в ансамбль. В отличие от бэггинга, каждое новое дерево корректирует ошибки

предыдущих, минимизируя отклонения предсказаний. Процесс продолжается до достижения минимальной ошибки или срабатывания критериев остановки.

В итоге, машинное обучения в медицине имеет широкие перспективы в области улучшения диагностики, прогнозирования заболеваний и принятия клинических решений, что подтверждается успешными примерами применения различных алгоритмов. При этом теоретическая часть рассматривает ключевые алгоритмы с их основными принципами.

Второй раздел «Разработка десктопного приложения» посвящен описанию разработанного приложения и отражению его особенностей.

В качестве среды разработки для моделей машинного обучения был выбран Jupyter Notebook, предоставляющий возможность запуска отдельных ячеек кода, что ускоряет процесс разработки.

Для работы с данными использовались библиотеки Python: pandas, numpy и scikit-learn.

Для реализации Desktop приложения был выбран модуль PySide6, обеспечивающий доступ к платформе Qt 6.0+ и позволяющий свободное использование в открытых и коммерческих проектах.

Для управления версиями проекта использовалась система контроля версий Git, что способствовало организованному процессу разработки. Для разрешения конфликтов между зависимостями применялось виртуальное окружение.

Для построения модели медиками Саратовской городской клинической больницы №2 им. В. И. Разумовского был предоставлен набор данных, содержащий реальные данные о пациентах. Каждый из 569 объектов в выборке описывается 30 -ю признаками. Большая часть признаков (27 из 30) формируется на основе анкеты больного с кровоточащей гастродуоденальной язвой, заполняемой дежурным врачом.

Целевым признаком является признак, описывающий состояние пациента. Для прогнозирования вероятности наступления повторного желудочного кровотечения нужно определить вероятность наступления предрецидивного синдрома, являющегося одним из значений целевого признака.

В ходе разведочного анализа стало понятно, что предрецидивный синдром наиболее часто встречается при большом объеме кровопотери (в среднем 750 мл. — 1400 мл.).

Также при наличии предрецидивного симптома размер обнаруженной яз-

вы у пациента в большинстве случаев больше чем у пациентов с консервативным лечением. В обоих случаях размер язвы с увеличением возраста увеличивается.

Состояние пациента во время осмотра при условии наличия у него предрецидивного симптома и язвы большого размера чаще удовлетворительное, реже — тяжелое.

В ходе обработки пропущенных значений был удален 1 объект, имеющий наибольшее количество пропусков. Пропуски в признаке «ДЦК» были заменены на значение -1; в признаке «Размер_язвы» — на значение 0. Пропуски в возрасте были заполнены медианным значением. В признаках «Пульс» и «Гемоглобин» присутствовали аномальное значение — 0. Объекты, обладающие аномальными значениями в признаках «Пульс» и «Гемоглобин», были удалены из выборки. В итоге, в наборе данных осталось 565 объектов (из 569 начальных).

Также была проведена нормализация количественных признаков («Пульс», «АД», «ДЦК», «Эритроциты», «Гемоглобин», «Размер_язвы», «Частота_дыхания»). Для проведения нормализации была использована часть библиотеки Scikit-learn — MinMaxScaler.

В процессе обучения для всех 7 алгоритмов подбирался оптимальный набор гиперпараметров. В итоге, лучшей моделью оказался обученный алгоритм случайного леса (см. рис. 1), со значением метрики F1-мера равной 0.783.

	Модель	Precision	Recall	F1-score
0	Random Forest (N=100, max_depth=14, entropy)	0.861000	0.740000	0.783000
1	SVM (C=2.31, kernel=linear)	0.957000	0.719000	0.782000
2	XGBoost (N=26, max_depth=2, l_r=0.1)	0.957000	0.719000	0.782000
3	Logistic Regression (solver=liblinear)	0.768000	0.724000	0.743000
4	Decision Tree (max_depth=5, gini)	0.724000	0.662000	0.684000
5	KNN (N=15, eculidian)	0.945000	0.625000	0.671000
6	Naive Bayes (GaussianNB)	0.648000	0.704000	0.666000

Рисунок 1 – Результаты обучения моделей классификации

Для улучшения пользовательского опыта при работе с моделью было написано Desktop приложение, работающее на базе ОС Windows. Функциональные возможности приложения позволяют пользователю заполнить анкету на больного с кровоточащей гастродуоденальной язвой; автоматически сформировать отчет, сформированный на данных анкеты (см. рис. 2), и получить предсказание модели машинного обучения.

Дата составления отчета: 2025-05-08 20:52:27.765840+04:00

Данные анкеты

Вопрос анкеты	Ответ
ФИО пациента	Иванов Иван Иванович
Дата поступления	01.01.01
Длительность кровотечения	До 6 часов
Анамнез	Гастритический
Операции в анамнезе	Ушивание прободной язвы
Способствующие факторы	Употребление лекарств накануне (гормоны, салицилаты и др.)

Данные о пациенте

Вывод модели:

Класс	Вероятность принадлежности к классу
Консервативное лечение	0.22
Предрецидивный синдром	**0.78**

Вывод модели

Рисунок 2 – Структура приложенного PDF-файла

Для удобства распространения и установки приложения был разработан инсталлятор на базе Inno Setup, включающий исполняемый файл, созданный с помощью PyInstaller, и все необходимые ресурсы. В процессе установки инсталлятор распаковывает файлы в целевую директорию, создает записи в реестре Windows для хранения пользовательских настроек и устанавливает требуемые зависимости.

В итоге, разработанное приложение предоставляет пользователям полный функционал для получения выводов модели машинного обучения и передачи полученных результатов. Данное решение имеет потенциальную возможность для значительного улучшения взаимодействия с медицинскими данными и повышения эффективность работы врачей.

ЗАКЛЮЧЕНИЕ

В ходе выполнения данной работы были выполнены все вспомогательные задачи:

- изучение существующих методов: в процессе работы был проведен анализ литературы по применению машинного обучения в медицине, что позволило выявить актуальные подходы и алгоритмы, используемые для решения задач классификации в области здравоохранения;
- сбор и предобработка данных: для обучения модели была получена репрезентативная выборка данных о пациентах, содержащая как демографические, так и клинические признаки; проведенный разведочный анализ данных позволил выявить важные зависимости и подготовить данные для обучения;
- обучение моделей: в ходе работы были протестированы различные алгоритмы машинного обучения, такие как логистическая регрессия, метод K-ближайших соседей, метод опорных векторов, наивный байесовский классификатор, дерево решений и случайный лес; наилучшие результаты показал алгоритм случайного леса, который продемонстрировал высокую точность и устойчивость к переобучению;
- разработка Desktop приложения: приложение было реализовано с использованием библиотеки PySide6, что обеспечило удобный интерфейс для пользователей, в приложении реализованы функции ввода данных, валидации, а также автоматической отправки результатов на электронную почту;
- создание инсталлятора: для удобства распространения приложения был разработан инсталлятор с использованием Inno Setup, что позволило упростить процесс установки и настройки приложения для конечных пользователей.

В итоге, была достигнута поставленная цель работы — разработано Desktop приложение с интегрированной моделью машинного обучения, предназначенное для прогнозирования ЖКК.

Результат проведенной работы выразился в создании эффективного решения, которое может быть использовано в клинической практике для прогнозирования ЖКК. Приложение позволяет медицинским работникам быстро оценивать риски и принимать обоснованные решения на основе данных о паци-

ентах.

Несмотря на достигнутые результаты, работа имеет ряд направлений для дальнейшего развития. Во-первых, можно рассмотреть возможность интеграции дополнительных источников данных, таких как результаты лабораторных исследований и медицинская история пациентов, что может повысить точность прогнозирования. Во-вторых, стоит обратить внимание на улучшение алгоритмов обработки данных, включая методы отбора признаков и оптимизации гиперпараметров, что может привести к улучшению качества модели.

Отдельные части магистерской работы были представлены на XIV Всероссийской неделе науки с международным участием. По итогам работы была отдана на печать научная статья «Выбор оптимальной модели прогнозирования рецидива язвенного гастродуоденального кровотечения» [9] по направлению клинической медицины.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Li, W.* Automatic segmentation of liver tumor in ct images with deep convolutional neural networks / W. Li, F. Jia, Q. Hu // *Journal of Computer and Communications*. — 2015. — Vol. 3, no. 11. — Pp. 146–151.

2 *Rao, S. R.* Electronic health records in small physician practices: availability, use, and perceived benefits / S. R. Rao, C. M. DesRoches, K. Donelan, E. G. Campbell, P. D. Miralles, A. K. Jha // *Journal of the American Medical Informatics Association*. — 2011. — Vol. 18, no. 3. — Pp. 271–275.

3 Predicting post-stroke pneumonia using deep neural network approaches / Y. Ge, Q. Wang, L. Wang, H. Wu, C. Peng, J. Wang, Y. Xu et al. // *International journal of medical informatics*. — 2019. — Vol. 132. — Pp. 103986–103986.

4 Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records / B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, C. T. Tran, C. R. Simpson // *Computer methods and programs in biomedicine*. — 2019. — Vol. 182. — Pp. 105055–105055.

5 *Faturrahman, M.* Structural mri classification for alzheimer's disease detection using deep belief network / M. Faturrahman, I. Wasito, N. Hanifah, R. Mufidah. — 2017. — Pp. 37–42.

6 *Churpek, M. M.* Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards / M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, D. P. Edelson // *Critical care medicine*. — 2016. — Vol. 44, no. 2. — Pp. 74–368.

7 Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning / A. K. Waljee, R. Lipson, W. L. Wiitala, Y. Zhang, B. Liu, J. Zhu, B. Wallace et al. // *Inflammatory bowel diseases*. — 2017. — Vol. 24, no. 1. — Pp. 45–53.

8 *Weegar, R.* Using machine learning for predicting cervical cancer from swedish electronic health records by mining hierarchical representations / R. Weegar, K. Sundström // *PloS one*. — 2020. — Vol. 15, no. 8. — P. e0237911.

9 ОМУС СГМУ [Электронный ресурс]. — URL: https://omus-sgmu.ru/applications/check? (Дата обращения 17.05.2025). Загл. с экр. Яз. рус.