Кравченко Елизавета Сергеевна
студентка
Полякова Анастасия Сергеевна
канд. техн. наук, доцент
ФГБОУ ВО «Сибирский государственный университет науки
и технологий им. академика М.Ф. Решетнева»
г. Красноярск, Красноярский край

СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА РЕШЕНИЯ ЗАДАЧ ДИАГНОСТИКИ В МЕДИЦИНСКИХ УЧРЕЖДЕНИЯХ

Аннотация*: в статье проводится анализ и сравнение различных методов машинного обучения для решения задачи диагностики сердечно-сосудистых заболеваний. Сердечно-сосудистые заболевания являются одной из основных причин смертности в мире, что делает задачу их своевременной диагностики и прогнозирования развития болезни крайне актуальной. Внедрение систем поддержки принятия решений в основные бизнес-процессы медицинской организации – является современным трендом в медицинских информационных системах. В исследовании рассматриваются алгоритмы, такие как случайный лес, метод k-ближайших соседей, логистическая регрессия и нейронные сети, с целью выявления их эффективности при работе с медицинскими данными.*

Ключевые слова*: качество медицинского обслуживания, поддержка принятия решений, обработка данных медицинской организации.*

Введение

Сердечно-сосудистые заболевания (ССЗ) остаются одной из ведущих причин смертности в мире, что подчеркивает важность ранней диагностики и эффективного лечения. В последние годы технологии машинного обучения (МО) приобрели значительное внимание в медицинских исследованиях, предоставляя новые возможности для анализа данных и улучшения профилактики заболева-

ний [1]. Данная статья направлена на анализ и сравнение различных методов машинного обучения, применяемых для диагностирования сердечно-сосудистых заболеваний.

Постановка задачи (цель исследования)

Современное медицинское учреждение применяет информационные технологии на всех этапах работы. При чем речь идет о всех возможных бизнес процессах: основных, управленческих и вспомогательных. К управленческим можно отнести системы управления персоналом или обеспечения документооборота [2]. Вспомогательные системы могут отвечать за управление хранением медикаментов или за закупки оборудования. Но, в настоящее время, наибольший интерес прикован к системам, обеспечивающим основные процессы, связанные с лечением пациентов.

Любая медицинская организация является генератором огромного объема данных. Данные медицинской диагностики обладают рядом существенных особенностей, заставляющих аналитиков искать особые подходы к решению задач медицины. Так, зачастую, медицинские данные не могут похвастать слишком большим объемом. Часто в результате проведенных исследований не удается собрать выборку значений больше чем несколько сотен объектов. Почти всегда медицинская информация является объектом врачебной тайны. Признаки описываемых объектов (пациентов) часто являются категориальными и ранговыми переменными [3]. В результате успешный опыт внедрения методов машинного обучения и искусственного интеллекта в других областях знаний требует доработки под задачи медицины.

Методы и материалы исследования

На сегодняшний день имеется множество методов для решения задачи классификации, использующих различные математические подходы и инструменты для своей реализации. Однако эффективность этих методов во многом зависит от конкретной задачи, которую необходимо решить. В дальнейшем будут рассмотрены несколько методов машинного обучения, а также их преимущества и недостатки.

Случайный лес (Random Forest) – это алгоритм машинного обучения, который использует ансамбль решающих деревьев для выполнения задач классификации и регрессии. Основная идея алгоритма заключается в том, что множество деревьев решений, обученных на случайных подмножествах данных и признаков, могут давать более надежные и обобщающие результаты, чем одно дерево.

Преимущества: высокая точность, устойчивость к переобучению, возможность обработки больших объемов данных и различных типов переменных.

Недостатки: сложность интерпретации результатов, необходимость больших вычислительных ресурсов.

Метод k ближайших соседей (k Nearest Neighbor, k-NN) основан на поиске k-ближайших соседей для каждого объекта из тестового набора данных и принятии решения на основе их классов или значений.

Преимущества k-NN: не требует предварительной обработки, может обрабатывать как линейные, так и нелинейные данные.

Недостатки k-NN: не учитывает взаимосвязь между признаками, чувствителен к шуму и выбросам в данных.

Логистическая регрессия – это статистический метод, используемый для классификации и прогнозирования вероятностей. Она применяется, когда зависимая переменная является категориальной (например, бинарной), то есть принимает два значения, такие как «да/нет», «успех/неудача» или «болен/здоров».

Преимущества: простота и интерпретируемость, эффективность при малом количестве данных, отсутствие предположений о распределении, гибкость

Недостатки: линейность, чувствительность к выбросам, неэффективность при сложных зависимостях, требования к количеству данных.

Нейронные сети (Artificial Neural Networks, ANN) – это класс алгоритмов машинного обучения, вдохновленный структурой и функцией человеческого мозга. Они используются для решения различных задач, включая классификацию, регрессию, обработку изображений, распознавание речи и многие другие. Нейронные сети состоят из взаимосвязанных узлов (нейронов), организованных в слои.

Преимущества ANN: может работать с большими объемами данных, может автоматически извлекать признаки из данных и создавать связи между ними.

Недостатки ANN: может быть склонны к переобучению на обучающих данных, требует большого количества обучающих данных для достижения высокой точности.

В таблице 1 расписаны основные параметры задачи классификации.

Таблица 1

Описание параметров

	Задача диагностики сердечно-сосудистых заболеваний
Кол-во признаков	8
Кол-во объектов выборки	450
Кол-во классов	2
Тип переменных	Integer

Полученные результаты

При сравнительном анализе методов машинного обучения для решения задачи классификации сердечно-сосудистых заболеваний необходимо учитывать точность классификации (ATrain – точность решения задачи классификации на обучающей выборки – accuracy).

Таблица 2

Результат исследования

	Accuracy	
	ATrain	ATest
Логистическая регрессия	0.60	0.68
Random Forest	0.98	0.78
k-NN	0.79	0.69
ANN	0.78	0.76

В таблице 2 представлен результат исследования сравнительного анализа методов машинного обучения. В ранее описанной задаче наибольшая эффективность была получена с помощью алгоритма случайный лес (Random Forest).

Заключение

На текущий момент продолжаются исследования предложенных методов на расширенном множестве задач [4]. В дальнейшем алгоритмы, прошедшие апробацию, будут переданы коллегам из медицинских организаций для проведения клинических испытаний.

Список литературы

1. Прогнозирование фентанил-ассоциированной нейротоксичности у больных с раком поджелудочной железы с помощью клинико-генетической модели / О.П. Боброва, Н.А. Шнайдер, М.М. Петрова [и др.] // Экспериментальная и клиническая гастроэнтерология. – 2021. – №3 (187). – С. 136–145. – DOI 10.31146/1682-8658-ecg-187-3-136-145. – EDN YEGGGN

2. Таракан Н.С. Автоматизация бизнес-процессов внутреннего контроля качества и безопасности медицинской деятельности / Н.С. Таракан, И.А. Панфилов // Право, экономика и управление: состояние, проблемы и перспективы: материалы Всероссийской научно-практической конференции с международным участием. – Чебоксары, 2024. – С. 156–160. – EDN VBQQQI

3. Моделирование процессов лабораторной информационной системы на производственном предприятии / И.А. Панфилов, А.В. Соинов, А.В. Безворотных, И.О. Степина // Перспективы науки. – 2023. – №1 (160). – С. 40–45. – EDN NKSSYL

4. Evolutionary algorithm for automated formation of decision-making models for predicting the safety of opioid therapy / L.V. Lipinskiy, O.D. Melnikova, A.S. Polyakova [et al.] // IOP Conference Series: Materials Science and Engineering. Krasnoyarsk Science and Technology City Hall. Krasnoyarsk, 2021. – С. 12126. – DOI 10.1088/1757-899X/1047/1/012126. – EDN AMARVK