УДК 621.3.049.77–048.24:537.2

СРАВНЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ ПРЕДСКАЗАНИЯ ПОРОКА СЕРДЦА

Дановский В.Д.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Хорошко В. В. – к.т.н., доцент, заведующий кафедры ПИКС

Аннотация. Исследованы характеристики, влияющие на диагностирование пороков сердца. Разработан набор предсказательных моделей на основе методов и моделей машинного обучения. Проведено сравнение результатов и определение наиболее подходящей для поставленной задачи модели. Предложены мероприятия по улучшению предсказательной точности модели.

Ключевые слова: порок сердца, электрокардиограмма, нейронная сеть, машинное обучение

Введение. Современная медицина активно использует методы машинного обучения для анализа больших медицинских данных и поддержки принятия клинических решений. Одной из важных задач является прогнозирование развития заболеваний, в частности пороков сердца, на основе различных факторов риска и биомаркеров.

Пороки сердца относятся к врожденным аномалиям развития сердца и крупных сосудов, которые могут быть выявлены уже во внутриутробном периоде с помощью ультразвукового скрининга или в период новорожденности, но чаще их обнаруживают уже в зрелом возрасте пациента с помощью общего обследования и электрокардиограммы. Правильная диагностика на ранних этапах развития имеет решающее значение для назначения эффективного лечения и профилактики осложнений.

Такие задачи, как классификация и прогнозирование состояния здоровья пациентов, традиционно решаются с помощью статистических и машинных методов обучения. Однако до сих пор отсутствуют систематические сравнения различных подходов машинного обучения, применяемых к конкретной задаче предсказания пороков сердца.

Данное исследование направлено на устранение этого пробела и сравнение эффективности популярных алгоритмов обучения: нейронных сетей, деревьев решений, логистической регрессии и других методов на реальных медицинских данных. Полученные результаты могут быть использованы для построения систем поддержки принятия клинических решений педиатрами и кардиологами.

Основная часть. Пример данных, используемых для этой работы представлен в таблице 1. Данные состоят из 1190 наблюдений по 11-ти признакам, а проведением наблюдений занимались Венгерские, Швейцарские и Американские университеты и больницы. 12-й признак является результатом, следствием произошедшим с пациентом.

Таблица 1 – Пример исследуемых данных о пациентах

Age	Sex	ChestPainType	RestingBP	Cholestero	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	ST	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
48	M	ATA	130	283	0	LVH	98	N	0.0	Up	0
37	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	F	NAP	150	195	0	Normal	122	N	0.0	Up	0

Каждый признак имеет следующее значение:

– Age: возраст пациента;

– Sex: пол пациента (M: Мужчина, F: Женщина);

– ChestPainType: тип боли в груди (TA: типичная стенокардия, ATA: атипичная стенокардия, NAP: неангинальная боль, ASY: бессимптомная);

– RestingBP: артериальное давление в состоянии покоя (мм рт. ст.);

– Cholestero: концентрация холестерина (мм/дл);

– FastingBS: уровень сахара в крови натощак (1: если FastingBS > 120 мг/дл, 0: в противном случае);

– RestingECG: результаты электрокардиограммы покоя (Normal: нормальный, ST: наличие аномалий ST-T (инверсия зубца T и/или подъем или депрессия ST > 0,05 мВ), LVH: вероятная или определенная гипертрофия левого желудочка по критериям Эстеса);

– MaxHR: достигнутая максимальная частота пульса (числовое значение от 60 до 202);

– ExerciseAngina: стенокардия, вызванная физической нагрузкой (Y: Да, N: Нет);

– Oldpeak: oldpeak = ST (Числовое значение, измеренное при депрессии);

– ST_Slope: наклон пикового сегмента ST при нагрузке (Up: вверх, Flat: плоский, Down: вниз);

– HeartDisease: выходной класс (1: заболевание сердца, 0: нормальное).

Так как исследуемые признаки описаны двумя форматами: числовым и текстовым, для работы с ними далее построенных моделей, данные будут векторизованы, тем самым придя в общий числовой формат. Однако до этого, возможно исследовать признаки в общем виде.

На рисунке 1 изображена тепловая карта корреляции числовых значений исследуемых признаков.

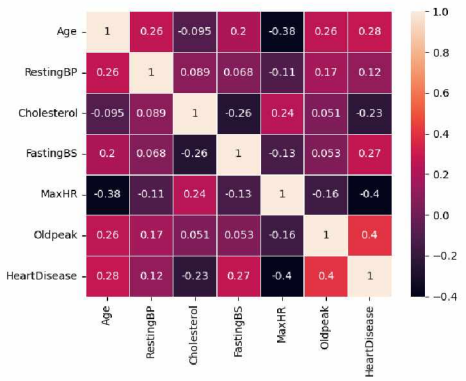

Рисунок 1 – Тепловая карта корреляции исследуемых признаков.

К сожалению, сильных корреляций между признаками не наблюдается, а даже наоборот видна отрицательная корреляция, что может быть использовано в будущем для исключения тех или иных конечных результатов.

Для признаков, представляющих собой общую характеристику и значения которых обозначено текстом, была оценена прямая зависимость влияния значений исследуемого признака к следствию. На рисунке 2 изображены гистограммы, визуализирующие частоту наличия или отсутствия порока сердца для характеристики всех признаков.

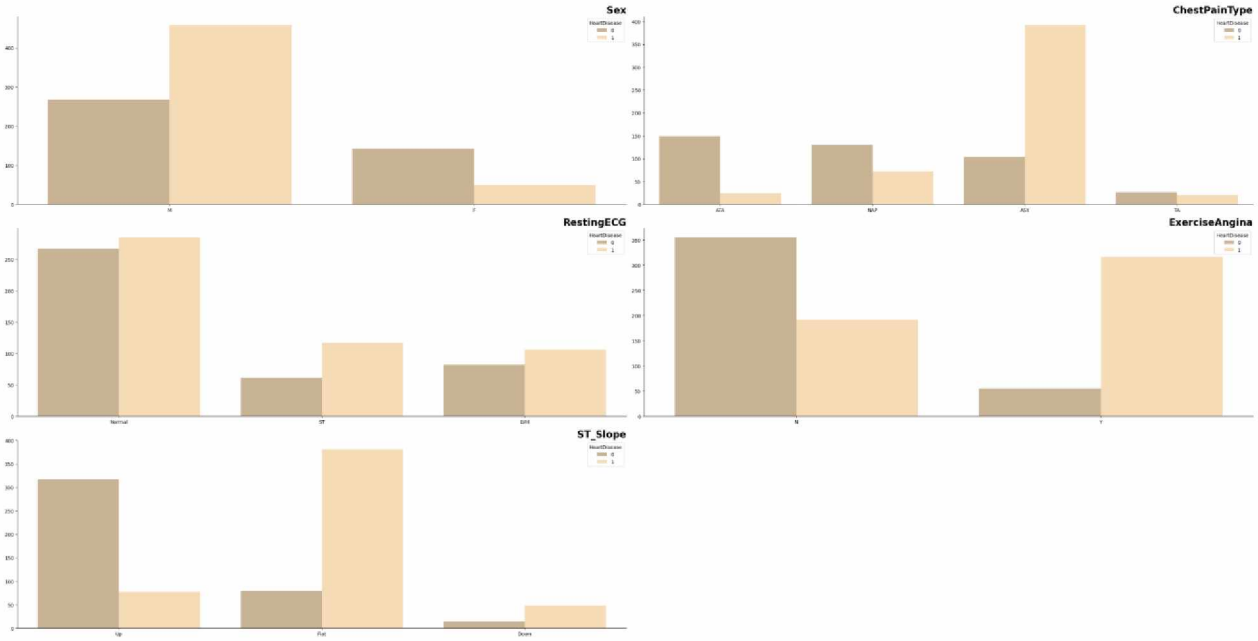

Рисунок 2 – Гистограммы отношения каждой из характеристик рассматриваемых признаков к наличию или отсутствию пороков сердца, где темным оттенком отображено наличие порока сердца, а светлым – его отсутствие

Из полученных гистограмм можно отметить следующие особенности наблюдений:

– ChestPainType: пациенты, у которых нет болей в груди или болезнь протекает бессимптомно, чаще других имеют порок сердца;

– ExerciseAngina: пациенты, страдающие стенокардией при физических нагрузках чаще имеют порок сердца;

– ST_Slope: пациенты с плоским или отсутствующим наклоном пикового сегмента ST при нагрузке чаще имеют порок сердца;

– Sex: у мужчин в 10 раз чаще встречается порок сердца, чем у женщин;

– RestingECG: вердикт о нормальных значениях ЭКГ не всегда может отобразить отсутствие порока сердца, в то время как при идентификации проблем на ЭКГ, вероятность встретить у пациента порок сердца выше.

Данные наблюдения подтверждают, что для диагностирования наличия у пациента порока сердца, необходимо проводить обследование по многим признакам, чем и удобны в использовании статистические и модели машинного обучения для комплексной оценки полученных результатов обследований.

Диаграмма рассеяния, демонстрирующая отношение максимальной частоты пульса (MaxHR) к возрасту представлены отдельно на рисунке 3. Из этой диаграммы видно, с увеличением возраста пациента, снижается частота сердечных сокращений и повышается вероятность развития сердечно-сосудистых заболеваний.

Несмотря на кажущуюся разнородность данных, особенно в вопросе значения каждого из признаков, с помощью статистических и моделей машинного обучения

возможна комплексная оценка признаков и предсказание наличия у пациента патологий, связанных с сердечно-сосудистой системой.

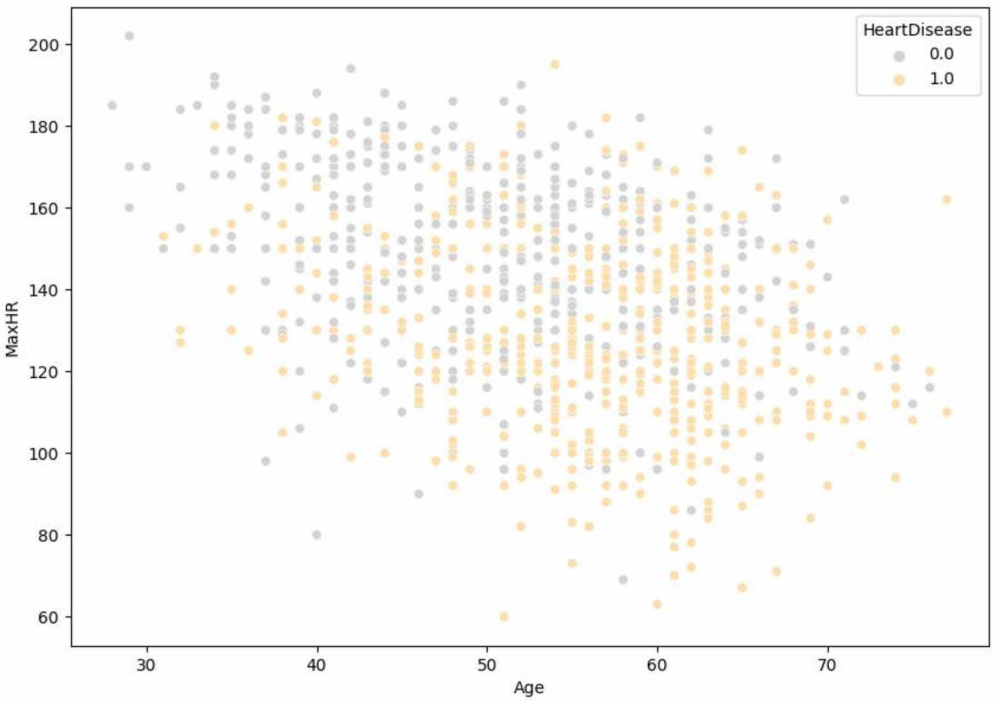

Рисунок 3 – Диаграмма рассеяния максимальной частоты пульса от возраста

В таблице 2 представлены данные о качестве работы шести методов и моделей машинного обучения: К-ближайших соседей, дерево решений, наивный Байес, метод опорных векторов, случайный лес и градиентный бустинг.

Таблица 2 – Оценка методов и моделей машинного обучения для задачи предсказания наличия порока сердца

Модель/Метод	**Точность**	**Полнота**	**F1-Score**
К ближайших соседей	73%	72%	73%
Дерево решений	80%	80%	80%
Наивный Байес	88%	87%	87%
Метод опорных векторов	69%	76%	72%
Случайный лес	88%	87%	85%
Градиентный бустинг	89%	89%	87%

Точность показывает долю правильно предсказанных случаев наличия у пациента порока сердца. Полнота показывает на сколько хорошо модель находит правильные случаи среди всех фактических. F1-Score – это среднее гармоническое точности и полноты. Этот критерий учитывает как точность, так и полноту и дает единую оценку производительности модели.

Исходя из таблицы 2 можно утверждать, что наибольшую точность в предсказании наличия у пациента порока сердца имеет метод градиентного бустинга. Однако стоит отметить, что модели были оценены с их рекомендованными параметрами по-умолчанию. В связи с этим, в теории, точность предсказания каждой из них можно улучшить, проведя индивидуальный подбор гиперпараметров для модели или метода. В связи с этими же причинами не была рассмотрена нейронная сеть, так как даже архитектура многослойного перцептрона отдельно нуждается в сравнении результатов её работы от множества вариаций гиперпараметров.

Заключение. Опираясь на всё вышесказанное, лучшим для предсказания наличия порока сердца у пациента возможно назвать метод градиентного бустинга. Метод способен работать с большим комплексом разнородных признаков и предоставлять верные результаты с точностью до 89%, что очень важно во время принятия решения о постановке диагноза. Однако стоит помнить, что система построенная с использованием этого метода будет носить рекомендательный характер, а конечное решение принимает лечащий врач.

Также стоит отметить, что возможно улучшить результаты как лидирующей, так и остальных рассмотренных моделей, проведя мероприятия по оптимизации как самих моделей, так и расширением набора исследуемых данных, а также преобразованием их для концентрации в них максимальной смысловой нагрузки и коррелирующих между собой или прямо влияющих на результат признаков.

Список литературы

1. Кушаковский, М. С., Журавлева Н.Б. Некоторые вопросы электрофизиологии сердца, механизмы сердечных аритмий и блокад / Кушаковский, М. С., Журавлева Н.Б // Аритмии и блокады сердца: Атлас электрокардиограмм– Питер, 2018. – Гл. 1. – С. 5–16.

2. Шляхто, Е. В. Клинические и лабораторные методы диагностики в кардиалогии / Е. В. Шляхто // Кардиология. Национальное руководство. Краткое издание. – 2023. – № 1 – С. 25 – 37.

3. Николенко, С. Быстрее, глубже, сильнее или Об оврагах, долинах и трамплинах / С. Николенко, А. Кадурин, Е. Архангельская. // Глубокое обучение. Погружение в мир нейронных сетей. – Питер, 2018. – Гл. 4. – С. 137–176.

4. Николенко, С. Теорема Байеса / С. Николенко, А. Кадурин, Е. Архангельская. // Глубокое обучение. Погружение в мир нейронных сетей. – Питер, 2018. – Гл. 2. – С. 39–53.

UDC 621.3.049.77–048.24:537.2

COMPARISON OF MACHINE LEARNING MODELS FOR THE TASK OF HEART DEFECT PREDICTION

Danovski V.D.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Khoroshko V.V. – Cand. of Sci, assistant professor, head of ICSD department

Annotation. Characteristics affecting the occurrence of heart defects and ECG plaques were investigated. A set of predictive models based on machine learning methods and models was developed. The results were compared and the most appropriate model for the task was determined. Measures to improve the predictive accuracy of the model are proposed.

Keywords. heart defect, electrocardiogram, neural network, machine learning