УДК 004

Нурматова Елена Вячеславовна,
доцент кафедры аппаратного, программного и математического обеспечения вычислительных систем, кандидат технических наук, Института кибербезопасности и цифровых технологий РТУ МИРЭА
nurmatova@mirea.ru

Пылаев Константин Витальевич,
магистрант 2 курса специальности «Информационные системы и технологии» Института кибербезопасности и цифровых технологий РТУ МИРЭА

ПОДХОД К РЕШЕНИЮ ЗАДАЧИ ДИАГНОСТИКИ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Аннотация. *В статье описан выбор наилучшей модели для прогнозирования сердечно-сосудистых заболеваний и этапы работы, которую следует провести с данными перед их обучением.*

Ключевые слова: *машинное обучение, статистические данные, прогнозирование, анализ данных.*

При рассмотрении темы, связанной с диагностикой сердечно-сосудистых заболеваний (ССЗ), возникает вопрос, как спрогнозировать заболевание? Ответом будет являться машинное обучение[1]. Алгоритм самостоятельного нахождения решений путем комплексного использования статистических данных, из которых выводятся закономерности и на основе которых будут производиться прогнозы.

Модели машинного обучения часто используются для прогнозирования многих распространённых заболеваний, включая диагностику ССЗ[2]. Часто существует множество факторов, которые способствуют выявлению пациентов, подверженных риску из этой группы заболеваний. Методы машинного обучения помогают выявлять скрытые закономерности в этих факторах, которые в противном случае могут быть попросту пропущены человеком[3].

[1] Хабр // URL: https://habr.com/ru/post/448892/ (дата обращения: 15.11.2022).

[2] Medium // URL: https://medium.com/nuances-of-programming/все-модели-машинного-обучения-за-5-минут-9270566197e7 (дата обращения: 15.11.2022).

[3] BMC // URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5 (дата обращения: 15.11.2022).

Существует достаточно моделей, которое могли помочь с решением поставленной задачи. Их характеристики приведены в табл. 1.

Получается, что каждая из моделей имеет, как и существенные достоинства в сравнении с другими, так и значительные недостатки. При этом нужно понимать, что у моделей как положительных, так и отрицательных сторон намного больше нежели, чем представлено в таблице. Для более детального выяснения и выбора наилучшего метода необходимо провести ряд тестов и ознакомиться с каждым поглубже[4].

Таблица 1. Сравнение моделей машинного обучения

Название модели	***Плюсы***	***Минусы***
1. Logistic Regression (Логистическая регрессия)	1. Проще реализовать, интерпретировать и очень эффективно обучать. 2. Очень быстро классифицирует неизвестные записи. 3. Может интерпретировать коэффициенты модели как показатели важности функции.	1. Если количество наблюдений меньше количества признаков, логистическую регрессию не следует использовать, может привести к переобучению. 2. Основным ограничением является предположение о линейности между зависимой переменной и независимыми переменными.
2. K Nearest Neighbours (k-ближайшие соседи)	1. Высокая точность. 2. Нечувствительность к выбросам. 3. Отсутствие допущений о вводе данных.	1. Высокая временная сложность 2. Большая пространственная сложность.
3. Kernel SVM (Метод опорных векторов)	1. Эффективен в пространствах больших размеров. 2. Универсальность.	1. В случае превышения количество выборок, следует избегать чрезмерной подгонки при выборе функций ядра.
4. Naive Bayes (Наивный байесовский классификатор)	1. Простота реализации 2. Низкие вычислительные затраты при обучении и классификации.	1. Низкое качество классификации в большинстве реальных задач.
5. Decision Tree (Деревья решений)	1. Простота. 2. Не нужна нормализация. 3. Встроенный отбор признаков работает одновременно с дискретными и непрерывными признаками.	1. Для новых наблюдений требуется полная перестройка всего дерева.

[4] BMC // URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5 (дата обращения: 15.11.2022).

Окончание табл. 1

Название модели	***Плюсы***	***Минусы***
6. Random Forest Classifier (Классификация случайного леса)	1. Имеет высокую точность предсказания. 2. Не чувствителен к масштабированию. 3. Хорошо работает с пропущенными данными	1. Нет формальных выводов, доступных для оценки важности переменных. 2. Алгоритм склонен к переобучению на некоторых задачах, особенно на зашумленных данных

Основываясь на исследованиях, проведенных по диагностировании риска сердечной недостаточности лучший результат показал: Random Forest Classifier[5].

Таблица 2. Точность и стандартное отклонение моделей

Название модели	***Точность***	***Стандартное отклонение***
1. Logistic Regression (Логистическая регрессия)	78.20 %	8.53 %
2. K Nearest Neighbours (k-ближайшие соседи)	75.02 %	6.22 %
3. Kernel SVM (Метод опорных векторов)	75.93 %	7.84 %
4. Naive Bayes (Наивный байесовский классификатор)	79.55 %	9.67 %
5. Decision Tree (Деревья решений)	74.13 %	5.70 %
6. Random Forest Classifier (Классификация случайного леса)	83.08 %	7.02 %

[5] Хабр // URL: https://habr.com/ru/post/448892/ (дата обращения: 15.11.2022).

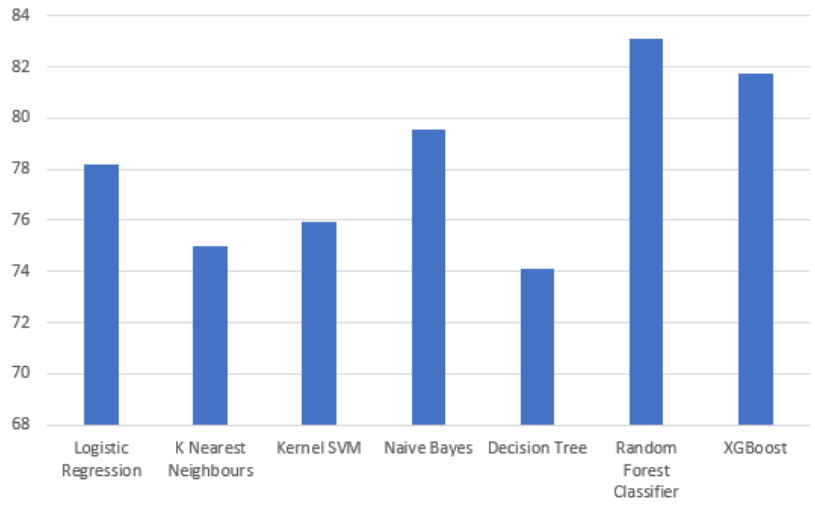

Рисунок 1. Отображение точности моделей

По данным результатам (рис. 1) предпочтительнее взять за основу модель Классификация случайного леса, так как точность по сравнению с другими ненамного, но все же выше и достигает 83%, при этом стандартное отклонение в пределах 7%, что также является хорошим результатом в сравнении с моделью Дерево решений, у которого 5%, но это нивелирует высокая точность первого.

А уже по результатам другого исследования, ориентированного на выявление ССЗ лучшим себя, показал метод Логистической регрессии, но стоит заметить, что тестирование проводилось с использованием лишь трех моделей машинного обучения (рис. 2).

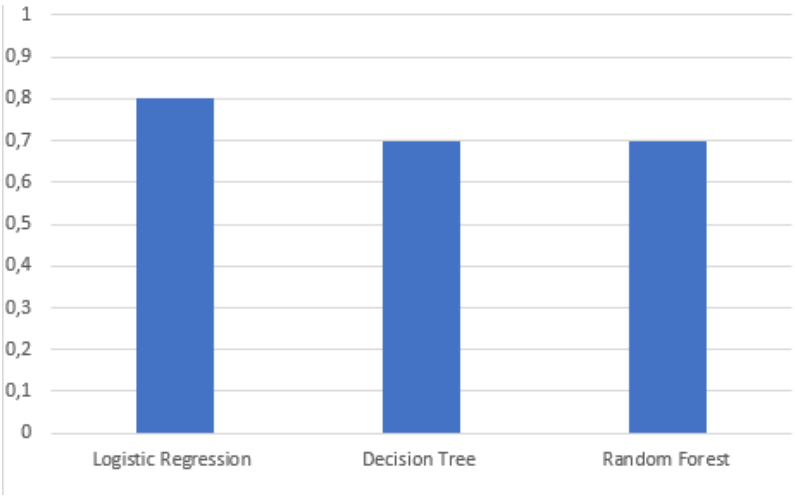

Рисунок 2. Точность 3-х моделей

Лучший результат среди всех показала, как ранее было сказано, логическая регрессия с точностью 80%. В связи с такими результатами стоит рассмотреть в

научной работе два вида моделей: Логистическую регрессию, классификатор случайного леса.

Также на точность результата можно будет повлиять, используя k-кратную перекрестную проверку, а после чего настроить гиперпараметры.

Стоит отметить, что перед тем, как отправлять данные на обучение с ними необходимо будет «поработать», используя выборку данных, очистку данных, генерацию признаков, интеграцию, форматирование и прочее.

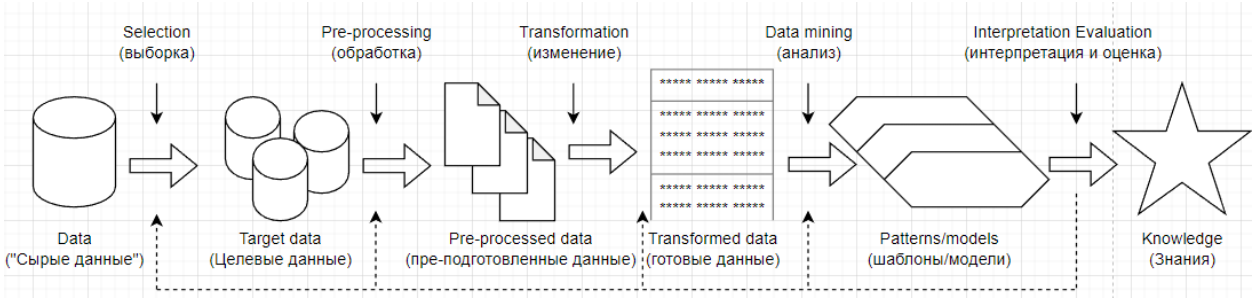

Рисунок 3. Работа с «сырыми» данными

По итогу можно сказать, что, решая задачу диагностики стоит много подзадач, исходя из которых сложится результат решения проблемы, но проведя уже такой небольшой анализ методов и описав подцели работы с “сырыми” данными:

—во-первых — это сильно упростит реализацию поставленной задачи;

—во-вторых — данный анализ, описанный выше, дал представление, как должна выглядеть блок-схема, визуализирующая процесс обработки данных (рис. 3).

Список источников и литературы

1. Хабр // URL: https://habr.com/ru/post/448892/ (дата обращения: 15.11.2022).

2. Medium // URL: https://medium.com/nuances-of-programming/все-модели-машинного-обучения-за-5-минут-9270566197e7 (дата обращения: 15.11.2022).

3. BMC // URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5 (дата обращения: 15.11.2022).