3. Альтшулер А.И., Альтшулер Ю.В. Особенности системного подхода в теории организации // Казанская наука. 2010. № 8(2). С. 262–268.

4. Ehaidib R., Dahanayake A. Cultural behavior features for adapting hospital information systems // J. Pokorný, M. Ivanović, B. Thalheim, P. Šaloun (eds.) Proceedings of the 20th East European Conference on Advances in Databases and Information Systems (ADBIS 2016), August 2016, Prague, Czech Republic. Lecture Notes in Computer Science. Vol. 637. Cham, Switzerland: Springer International Publishing, 2016. P. 180–192. DOI: 10.1007/978-3-319-44066-8_19.

5. The 7Rs of Process Innovation. URL: http://www.stephenshapiro.com/pdfs/7rs.pdf (дата обращения: 24.05.2020).

УДК 004.62 : 004.85
doi:10.18720/SPBPU/2/id20-223

Аверина Анастасия Александровна,
студент

ПРИМЕНЕНИЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В РАЗРАБОТКЕ МЕДИЦИНСКИХ СИСТЕМ ПРИНЯТИЯ РЕШЕНИЙ

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия,
averina.a@edu.spbstu.ru

Аннотация. Основная цель данной работы — выявление закономерности развития у человека болезни сердца по возрасту, полу и максимальной частоте сердечного ритма в момент выполнения физических упражнений для дальнейшего построения медицинской системы принятия решений. В исследовании подробно рассматривается возможность применения моделей логистической регрессии для прогнозирования вероятности выявления болезни сердца у конкретного пациента, построенных с использованием реальных данных об учете заболеваний сердца нескольких медицинских центров. Результаты экспериментов, описанные в работе, далее могут быть использованы для внедрения аналогичных систем принятия решения в медицинские организации.

Ключевые слова: системы принятия решений, язык программирования R, сердечно-сосудистые заболевания, модели логистической регрессии, *t*-тесты, хи-квадрат тесты.

Anastasiia A. Averina,
Master Student, BSc

APPLICATION OF DATA MINING METHODS IN THE DEVELOPMENT OF CLINICAL DECISION SUPPORT SYSTEMS

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia,
averina.a@edu.spbstu.ru

Abstract. In this paper, the relations between human sex, age, maximum heart rate at exercise and the development of cardiovascular diseases are analyzed. Logistic regression models are used to predict the possibility of the development of cardiovascular diseases for a particular patient. For the experiments I have used multiple cardiovascular disease datasets [1], which contain actual data from four medical centers. The results of the experiments can further be used in the deployment of the decision support systems in health organizations.

Keywords: decision support systems, R programming language, cardiovascular diseases, logistic regression models, t-tests, chi-squared tests.

Введение

Сердечно-сосудистые заболевания (ССЗ) – это группа болезней сердца и кровеносных сосудов, включающая гипертонию, инфаркт, инсульт, сердечную недостаточность, ревматические заболевания сердца, врожденные пороки сердца и т.д., которые могут протекать в скрытой форме долгое время, клинически никак себя не проявляя. По данным ВОЗ именно эти заболевания являются основной причиной преждевременной смерти людей в развитых странах мира. Поэтому организации здравоохранения все чаще внедряют в свою структуру системы принятия решений, использующие методы интеллектуального анализа данных и помогающие клиническим специалистам в постановке диагнозов, прогнозировании болезней, разработке направлений лечения. Таким системам приходится обрабатывать большие объемы данных.

1. Описание набора данных

Основная цель данной работы – выявление закономерности развития у человека болезни сердца по возрасту, полу и максимальной частоте сердечного ритма в момент выполнения физических упражнений.

В работе используется обезличенный набор данных, включающий в себя информацию об учете заболеваний сердца медицинских центров на базе клиник Лонг-Бич и Кливленда, Венгерского университета кардиологии, клиник при университетах в Базеле и Цюрихе [1]. В наборе содержатся характеристики 916-ти пациентов, база данных учитывает 76 атрибутов (возраст, пол, характер болей в грудной клетке, кровяное давление и т. д.) и 1 искусственный атрибут – идентификатор кортежа, представленные в виде единой структуры данных expdataset. Для обработки данных используется язык программирования R, предназначенный для статистической обработки данных, и свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R – R-Studio [2].

Исходный набор данных был приведен к удобному для анализа виду – дискретный атрибут, описывающий наличие у пациента болезни сердца и его тяжесть, был представлен как «событие», т. к. характеризовался слишком большим диапазоном значений. В рамках данной работы тяжесть заболевания сердца не важна, поэтому диапазон значений атрибута

был сокращён до наличия у пациента болезни или ее отсутствия (события 1/0). Исходный атрибут «пол» был преобразован в специальный класс векторов, предназначенный для хранения кодов соответствующих уровней номинальных признаков – фактор с метками «Женщина» и «Мужчина».

2. Определение значимости влияния атрибутов набора на наличие болезни сердца

Для первичного определения атрибутов набора, оказывающих наибольшее влияние на наличие или отсутствие у пациента болезни сердца, были выбраны статистические *t*-тесты и хи-квадрат тесты, на выходе дающие показательные значения – тест-статистику и достигнутый уровень значимости (*p*-значение). В качестве нулевой гипотезы положена статистическая значимость равенства среднего значения каждого атрибута среди пациентов с болезнью сердца и без. В терминах статистики нулевая гипотеза заключается в том, что обе эти выборки происходят из нормально распределенных генеральных совокупностей с одинаковыми средними значениями: H_0:μ_1=μ_1. В общем виде проверка (тест) гипотезы осуществляется с помощью *t*-критерия, который рассчитывается как отношение разницы между выборочным средним и известным значением к стандартной ошибке выборочного среднего. Так как эти генеральные средние оцениваются в работе при помощи выборочных средних значений, формула *t*-критерия имеет вид

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

где в знаменателе находится стандартная ошибка разницы между выборочными средними, а s_1^2 и s_2^2 – выборочные оценки дисперсии.

Например, для атрибутов «возраст» и «пульс в режиме нагрузки», по результатам *t*-тестов достигнутый уровень значимости равен $2{,}2 \cdot 10^{-16}$. Таким образом, средние значения возраста и максимального пульса пациентов из рассматриваемых весовых групп (с болезнью и без) статистически значимо различаются и отвергая нулевую гипотезу о равенстве этих средних значений, можно ошибиться с вероятностью лишь $2{,}2 \cdot 10^{-14}$ %. Критическое значение хи-квадрат для таблицы сопряжённости, описывающей частоту возникновения болезней сердца среди мужчин и женщин, равно 85,605 с 1 степенью свободы. Т. к. критическое значение критерия хи-квадрат Пирсона при уровне значимости ниже 0,001 составляет менее 0,0000016, зависимость наличия болезни сердца от пола – статистически значима. Для иллюстрации была построена гистограмма

(см. рис. 1), из которой видно, что болезням сердца чаще подвержены мужчины – в 90 % случаев и лишь в 10 % случаев – женщины.

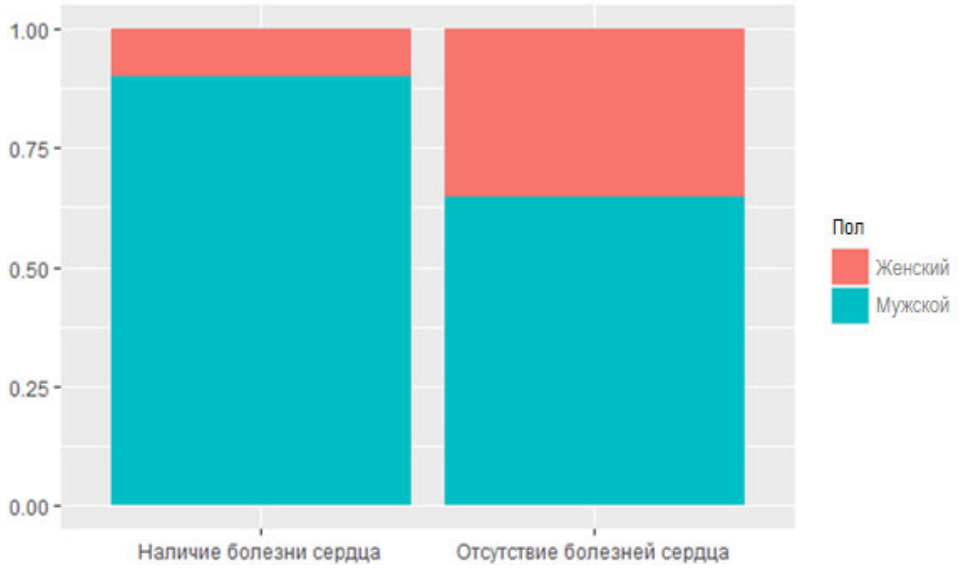

Рис. 1. Гистограмма зависимости наличия болезни сердца от пола

Статистические тесты и диаграммы, построенные также для других 72-х атрибутов, показали, что значения атрибутов: пол, возраст и максимальная частота пульса в нагрузке, – оказывают наибольшее влияние на возникновение у пациента болезни сердца.

3. Построение модели прогнозирования вероятности возникновения болезни сердца у пациента

Неформальной целью данной работы является построение модели прогнозирования вероятности приобретения человеком болезни сердца по возрасту, полу и частоте пульса в режиме нагрузки. В работе используются методы логистической регрессии, т. к. выходной атрибут один (наличие болезни сердца) и принимает бинарные значения, а количество входных атрибутов (пол, возраст, пульс в нагрузке) более двух. Бинарный атрибут зависим, поэтому исследованию подвергается влияние независимого атрибута на возможность получения определенного значения зависимого атрибута [3]. В среде R для применения обобщенных линеаризованных моделей (регрессий) к бинарным выходным атрибутам выбрана команда glm().

```
typemodel ← glm(data = expdataset, cl_reduced ~ age + sex + maxheartrate,
                family ="binomial")
```

Чтобы доказать значимость влияния атрибутов «пол», «возраст» и «пульс в нагрузке» на прогнозируемый атрибут, построенная модель регрессии позволяет определить количественное описание тесноты связи

признака А с признаком Б или отношение шансов. То есть, как сильно наличие или отсутствие признака влияет на наличие или отсутствие значения выходного атрибута. В языке R при использовании функции glm() отношение шансов вычисляется для каждого входного атрибута по умолчанию. В полученной модели показателем отношения шансов для трех входных атрибутов является экспоненцированное свойство 'OR' (см. рис. 2).

	term	estimate	std.error	statistic	p.value	OR	lower_CI	upper_CI
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	0.431	0.805	0.535	5.93e- 1	1.54	0.317	7.46
2	age	0.0471	0.00920	5.11	3.16e- 7	1.05	1.03	1.07
3	sexMale	1.49	0.198	7.54	4.87e-14	4.44	3.01	6.53
4	maxheartrate	-0.0281	0.00353	-7.96	1.78e-15	0.972	0.966	0.979

Рис. 2. Таблица коэффициентов модели прогнозирования

Отношение шансов для входного атрибута «пол» намного больше 1, а для атрибутов «возраст» и «максимальный пульс» примерно равно 1, что говорит о положительном влияние наличия этих признаков на прогнозируемый атрибут. Таким образом, значимость влияния выбранных входных атрибутов доказана статистически, визуально и аналитически.

Чтобы построенная выше модель предсказывала вероятность проявления болезни сердца у каждого нового пациента, была создана структура данных new_expdataset, из которой задаются определенные значения входных атрибутов, определяющих вероятность – к примеру, 23 года, женщина, 173 уд./мин. Модель настраивается на прогнозирование выходного атрибута при каждом добавлении нового кортежа в созданную структуру. В исходный набор записывается свойство, указывающее, что вероятность возникновения болезни более 50 %. Для отображения вероятности возникновения сердечной болезни для каждого нового пациента, значения, рассчитанные функцией, сохраняются в отдельную переменную.

```
pred_prob ← predict(typemodel, expdataset, type = "response")
expdataset$pred_cl ← ifelse(pred_prob >= 0.5, 1, 0)
new_expdataset ← data.frame(age = 23, sex = "Female", maxheartrate = 173)
pred_new ← predict(typemodel, new_expdataset, type = "response")
```

Полученная вероятность возникновения сердечной болезни для 23-летней женщины с максимальной частотой сердечного ритма 173 ударов в минуту в режиме физической активности равна 0,034 (3,4 %), то есть возникновение болезни сердца маловероятно.

4. Оценка точности модели прогнозирования

Для оценки точности построенной прогнозирующей модели были использованы наиболее распространенные метрики оценивания моделей логистической регрессии – доля правильно предсказанных значений, ошибка классификации, площадь под ROC-кривой (AUC) и матрица ошибок. Было решено использовать 4 разные метрики, т.к. каждая оценка имеет свои преимущества и недостатки – посчитанная ошибка классификации может ввести в заблуждение, если выходное значение не сбалансировано или у одного из классов довольно высокая априорная вероятность, а преимуществом AUC является инвариантность относительно отношения цены ошибки первого или второго рода.

В R для построения всех трех видов метрик используется библиотека Metrics, а именно функции auc(), accuracy() и ce(), которые рассчитывают площадь под ROC-кривой, долю правильно предсказанных значений и ошибку классификации соответственно. Метрики дают схожие значения вплоть до третьего порядка после запятой – в соответствие с ними, построенная модель точна приблизительно на 69 % (0,69), так как в некоторых случаях данные были отнесены к неверному классу, что видно в матрице ошибок. Из 389 человек, не имеющих болезни сердца, 304 были правильно классифицированы, а 85 – нет. А из 472 человек, имеющих болезнь сердца, правильно классифицированы были 366. Такие показания точности были расценены, как удовлетворительные для решения поставленной задачи.

Заключение и выводы

По результатам можно сделать вывод о том, что, для мужчины в преклонном возрасте и с более низкой максимальной частотой сердечного ритма в режиме физической активности, наличие болезни сердца вероятнее всего, т.к. эти три показателя являются самыми серьезными факторами риска возникновения болезни сердца. Построенная модель может быть применена для расчета вероятности возникновения болезни у случайного пациента и использована для диагностики болезней сердца.

Модель является достаточно точной для решения практических задач – 69 % (0,69). Один из способов повышения точности модели - включение в модель других подходящих входных атрибутов из исходного набора и расширения набора за счет новых экспериментальных данных.

На данный момент в качестве задачи для дальнейшего исследования была выбрана оценка эффективности использования алгоритмов кластеризации в определении направлений лечения пациентов с болезнями сердца. Пациенты с одинаковыми или похожими значениями атрибутов должны откликаться на один и тот же метод лечения, а изучение данного вопроса позволит предсказывать результат такого лечения.

Список литературы

1. UCI. Machine Learning Repository. URL: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ (дата обращения: 01.05.2020).

2. Lantz B. Machine learning with R. 2nd. ed. Birmingham, UK: Packt Publishing, 2015. 452 p.

3. Барсегян А. А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

4. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science: 50 важнейших понятий / Пер. с англ. СПб.: БХВ-Петербург, 2018. 304 с.

УДК 004.422
doi:10.18720/SPBPU/2/id20-224

Сердюкова Мария Александровна [1],
студент магистратуры;
Курбесов Александр Валерьянович [2],
канд. экон. наук, доцент кафедры

ВЕБ-ПРИЛОЖЕНИЕ «ЭЛЕКТРОННЫЙ РЕЕСТР ВЕТЕРАНОВ ВОЙН» (ЭР ВВ)

[1, 2] Ростовский государственный экономический университет (РИНХ),
Ростов-на-Дону, Россия,
[1] maria_sun777@mail.ru, [2] akurbesov@yandex.ru

Аннотация. В статье рассмотрен комплекс современных решений в сфере безопасности и хранения данных в медицинских учреждениях. Разработанный программный продукт позволяет осуществить процесс накопления необходимой информации на основе современных интернет-технологий. Основной задачей разработанного веб-приложения, является централизация и хранение данных о ветеранах войн Ростовской области. «Электронный реестр ветеранов войн» (ЭР ВВ) представляет собой легко структурированное веб-приложение. В статье описаны функции и механизмы веб-приложения, его структура, сформулированы особенности его построения. Задача веб-приложения «Электронный реестр ветеранов войн» (ЭР ВВ) – это ведение актуального учета качественных и количественных отчетов, в рамках заданной предметной области. На основе полученных данных была проведена исследовательская работа реальных статистических данных о состоянии здоровья ветеранов войн. Вносимые данные способствовали оперативному принятию адекватных лечебно-диагностических решений.

***Ключевые слова*:** веб-приложение, «Электронный реестр ветеранов войн», качественные отчеты, количественные отчеты, хранение данных, MVC, C#, HTML, фреймворк, веб-фреймворк, jQuery, JavaScript.