УДК 004.62 , 616.379-008.64
doi:10.18720/SPBPU/2/id23-503

Сажнова Виктория Александровна [1],
студент магистратуры;
Нестеров Сергей Александрович [2],
доцент, канд. техн. наук, доцент

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ САХАРНОГО ДИАБЕТА В МЕДИЦИНСКИХ СИСТЕМАХ ПРИНЯТИЯ РЕШЕНИЙ

[1, 2] Россия, Санкт-Петербург, Санкт-Петербургский политехнический университет Петра Великого,
[1] sazhnova.va@edu.spbstu.ru, [2] nesterov@spbstu.ru

Аннотация. В статье рассматривается анализ медицинского опроса пациентов по наличию или отсутствию ряда характерных для сахарного диабета симптомов, на основе которого с помощью средств языка R была построена модель прогнозирования наличия сахарного диабета для случайного пациента. Результаты и методы, полученные в данном исследовании, могут быть использованы для разработки систем ранней диагностики и медицинских приложений самодиагностики сахарного диабета.

Ключевые слова: системы принятия решений, корреляция, интеллектуальный анализ данных, прогнозирование, сахарный диабет, ранняя диагностика заболеваний, язык программирования R.

Victoria A. Sazhnova [1],
Master's Student;
Sergey A. Nesterov [2],
Candidate of Technical Sciences, Associate Professor

APPLICATION OF DATA MINING FOR PREDICTING DIABETES IN CLINICAL DECISION SUPPORT SYSTEM

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia, [1] sazhnova.va@edu.spbstu.ru, [2] nesterov@spbstu.ru

Abstract. The article describes the analysis of the patients report about presence or absence of some diabetic symptoms. In the result, model was constructed for predicting the presence of diabetes for a random patient using the R language tools. The results and methods can further be used in the deployment of systems for the early diabetes diagnosis.

Keywords: decision support systems, correlation, data mining, predicting, diabetes, early diagnostic, R programming language.

Введение

На данный момент согласно отчетам Всемирной организации здравоохранения (ВОЗ) сахарный диабет находится на третьем месте по причинам смертности после сердечно-сосудистых заболеваний и онкологии. Поэтому важно своевременно диагностировать заболевание на ранних стадиях, и как можно раньше помочь человеку начать поддерживающее лечение для сохранения качества и увеличения продолжительности жизни.

Цель данной работы — определение вероятности наличия сахарного диабета по проявлению характерных ему симптомов у человека.

1. Описание набора данных и первичная обработка

Для проведения исследования был выбран язык программирования R, который предназначен для статистической обработки данных, а также имеет удобные инструменты для визуализации результатов [2].

В работе исследуется обезличенный набор данных с результатами анкетирования пациентов из диабетической больницы Силхета (Бангладеш) [1]. Помимо возраста и пола пациента в опросе уточнялось наличие или отсутствие следующих 14 симптомов: частое мочеиспускание, повышенная жажда, резкое снижение веса, общая слабость, повышенный аппетит, генитальный кандидоз, нечеткость зрения, кожный зуд, раздражительность, замедленное заживление ран, частичный парез, ригидность мышц, выпадение волос, ожирение.

Также в наборе данных каждому пациенту была сопоставлена информация о том, был ли у него диагностирован сахарный диабет. Сводная информация о количестве принявших участие в анкетировании женщин и мужчин с учётом наличия сахарного диабета приведена в таблице 1.

Таблица 1

Описание набора данных

Категория	С диабетом, чел.	Без диабета, чел.	Всего, чел.
Мужчины	173	19	192
Женщины	147	181	328
Общее	320	200	520

Для визуализации возраста принявших участие в анкетировании пациентов была построена гистограмма, которая изображена на рисунке 1, с выделением цветом по половой принадлежности: розовым и голубым для женщин и мужчин соответственно.

Как видно из гистограммы, возраст наибольшего количества опрошенных находится в диапазоне от 30 до 60 лет. Также для исследуемого набора были определены следующие характеристики: минимальный возраст опрошенного — 16 лет, максимальный — 90 лет, а средний возраст — 48 лет.

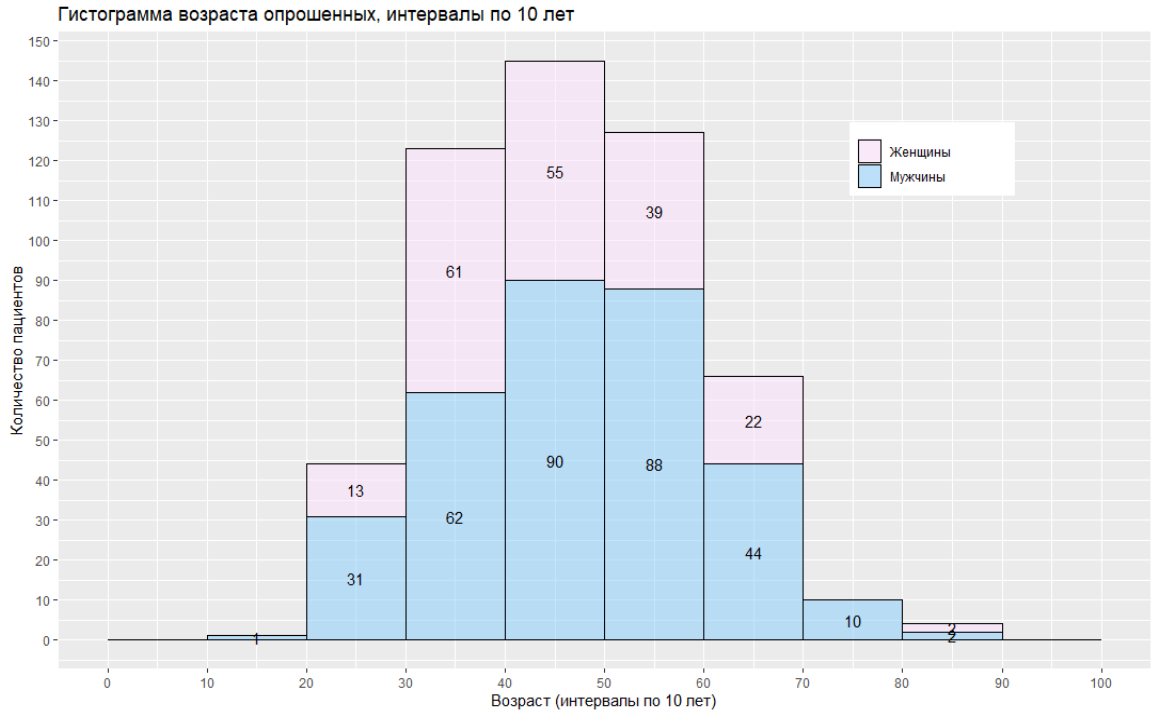

Рис. 1. Гистограмма возраста опрошенных (интервалы по 10 лет)

2. Корреляционный анализ атрибутов

Корреляция определяет меру зависимости между двумя и более величинами. Данная зависимость выражается через коэффициент корреляции. Коэффициент принимает значения в промежутке от –1 до 1. Интерпретировать значения коэффициента корреляции необходимо по знаку и модулю: если коэффициент положительный — связь между атрибутами прямая, ес-

ли коэффициент отрицательный — обратная; если модуль коэффициента близок к 1 или равен ему — связь между переменными сильная, если близка к 0 — слабая, а если равна 0 — связь отсутствует [4].

С целью определения зависимости между переменными исходного набора для них были попарно рассчитаны коэффициенты корреляции по Пирсону r по следующей формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad (1)$$

где x_i и y_i — значения двух переменных, $\bar{x}$ и $\bar{y}$ — их средние значения, а s_x и s_y — их стандартные отклонения; n — количество пар значений.

Полученные результаты были визуализированы с помощью тепловой карты, которая изображена на рисунке 2.

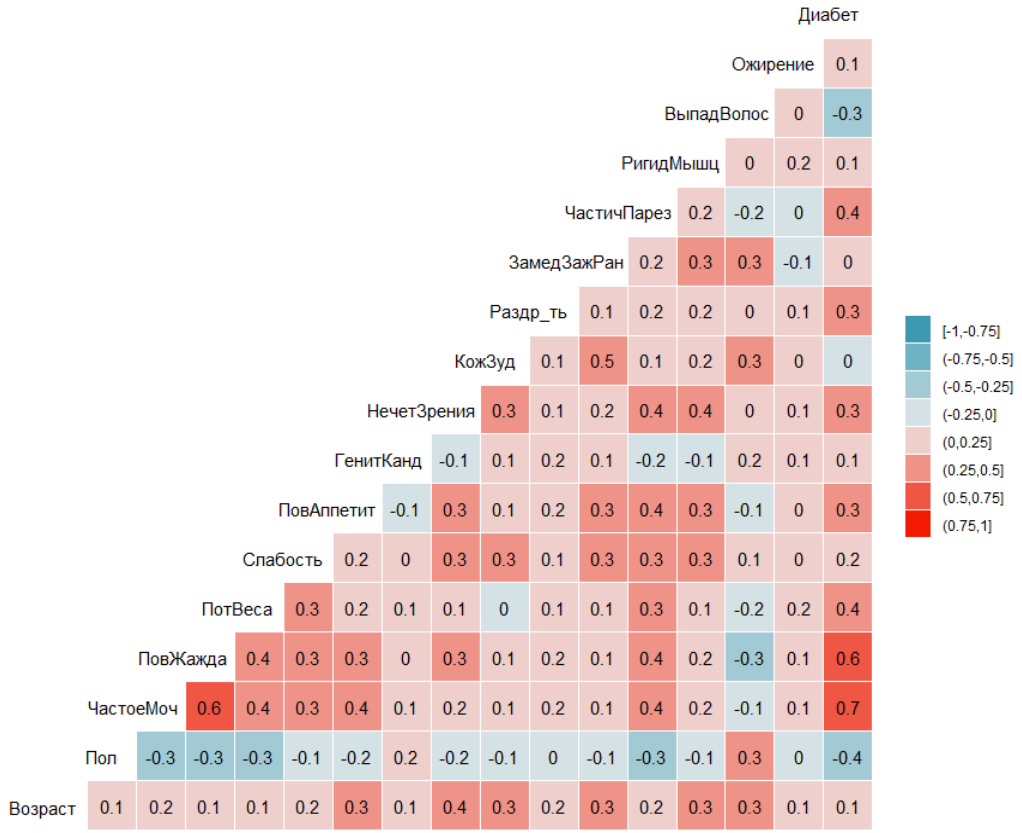

Рис. 2. Тепловая карта корреляции атрибутов

На данной тепловой карте палитра красных цветов характеризует положительную корреляцию, а палитра синих цветов — отрицательную. Согласно тепловой карте видна положительная корреляция между наличием сахарного диабета и таких атрибутов, как частое мочеиспускание и повышенная жажда.

На основе таблицы интерпретации силы корреляции можно сказать, что наличие таких симптомов, как частое мочеиспускание и повышенная жажда, имеют среднее влияние на наличие сахарного диабета, так как значения соответствующих коэффициентов принадлежат диапазону от 0,5 до 0,7 [4].

Таким образом, можно сказать, что данные симптомы являются главными признаками для ранней диагностики сахарного диабета. С медицинской точки зрения такую взаимосвязь можно обосновать тем, что почки интенсивно фильтруют и поглощают накопленный в крови избыток глюкозы.

Также стоит отметить, что коэффициенты корреляции между наличием сахарного диабета и таких симптомов, как кожный зуд и замедленное заживление ран, равны нулю, что показывает отсутствие зависимости между ними. Объяснить такую зависимость можно тем, что данные симптомы диабета чаще проявляются на поздних стадиях заболевания, поэтому не так распространены у опрошенных пациентов.

2. Построение и сравнение моделей предсказания

Цель исследования — определить риск наличия сахарного диабета по проявлению конкретных симптомов, на основе набора данных с заранее известными результатами. Эта задача является задачей построения предсказательной модели классификации [3].

В исследовании были построены четыре модели классификации на основе следующих алгоритмов: алгоритм k-ближайших соседей, алгоритм деревьев решений, алгоритм Байеса и алгоритм логистической регрессии.

Для обучения и проверки точности моделей изначальный набор данных был разделён на тренировочный и тестовый наборы в соотношении 70 на 30. В таблице 2 представлены результаты обучения построенных моделей на тестовых наборах.

Таблица 2

Сравнение результатов моделей

Алгоритм	Точность модели на основе матрицы ошибок	Точность модели на основе значения AUC
k-ближайших соседей	0,9352941	0,9428571
Деревья решений	0,8470588	0,8410287
Байес	0,8235294	0,8298992
Логистическая регрессия	0,8352941	0,8282448

Исходя из результатов таблицы наилучшую точность согласно матрице ошибок имеет модель, построенная на основе алгоритма k-ближайших соседей. Матрица ошибок данной модели показала, что из 100 диабетиков — верно классифицировано 90 диабетиков и 10 диабетиков неверно. А из 70 пациентов, не имеющих сахарный диабет, только

1 был отнесен к диабетикам. Таким образом, точность данной модели составляет 94 процента на тестовом наборе.

Также для оценки точности для каждой модели были вычислены значения площади (AUC) под ROC-кривой («кривая ошибок»). Полученные результаты, которые представлены в таблице 2, схожи с теми, что были вычислены при расчете доли правильно предсказанных значений по матрицам ошибок. Наилучший показатель точности также демонстрирует модель, полученная с помощью алгоритма k-ближайших соседей.

4. Пример использования модели предсказания для случайного пациента

Для построения модели прогнозирования вероятности наличия сахарного диабета у случайного пациента была взята модель на основе алгоритма k-ближайших соседей, так как она показала наилучший результат классификации пациентов на тестовом наборе данных.

Рассмотрим в качестве примера нового пациента с наличием восьми симптомов: повышенная жажда, резкое снижение веса, общая слабость, повышенный аппетит, нечеткость зрения, кожный зуд, раздражительность, выпадение волос.

Набор наличия симптомов из примера был передан в модель предсказания, результаты работы которой представлены на рисунке 3.

```
> example_patient$result <- predict(knn_model, example_patient)
> example_patient$result
[1] 0.6722467
```

Рис. 3. Результат прогнозирования модели для случайного пациента

Полученная вероятность наличия сахарного диабета для пациента с указанными выше симптомами равна 0,67 (67 %), то есть риск наличия сахарного диабета достаточно высокий.

Аналогичным образом можно определить вероятность наличия сахарного диабета для любого нового пациента с учётом проявления конкретных симптомов.

Заключение и выводы

В ходе работы был проведен корреляционный анализ атрибутов исходного набора данных, по результатам которого были выявлено, что основные взаимосвязи между атрибутами: наибольшее влияние на наличие сахарного диабета оказывает проявление у пациентов таких симптомов, как частое мочеиспускание и чрезмерная жажда.

В данной работе представлен результат сравнения моделей прогнозирования наличия сахарного диабета, построенных на четырёх алгоритмах: наилучший результат показала модель, основанная на алгоритме k-ближайших соседей, точность модели — 94 % на тестовых данных.

В результате работы была построена предсказательная модель вероятности наличия сахарного диабета для любого нового пациента с учётом проявления конкретных симптомов.

Результаты и методы, полученные в ходе проведения данного исследования, могут быть использованы для разработки систем ранней диагностики и медицинских приложений самодиагностики диабета.

Список литературы

1. UCI. Machine Learning Repository. – URL: https://archive.ics.uci.edu/ml/machine-learning-databases/00529/ (дата обращения: 20.09.2022).

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

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. Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей [Пер. с нем.] / Под ред. В.Е. Момота. – М. [и др.] : DiaSoft(DS), 2002. – 602 с.