УДК 618.19

ДИАГНОСТИКА И ПРОГНОЗ ОНКОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ ГРУДИ С ПОМОЩЬЮ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ЯЗЫКЕ PYTHON И МЕТОДА ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

ПАСЕДЬКО[1] В.В., ЯКИМОВ[2] Д.А., ВЫГОВСКАЯ[3] Н.В.

[1]ООО «Техартгруп» (г. Могилев, Республика Беларусь)
[2]УЗ «Могилевская областная больница» (г. Могилев, Республика Беларусь)
[3]Белорусско-Российский университет, (г. Могилев, Республика Беларусь)

Аннотация. Статья посвящена описанию методики создания автоматизированной программной системы для диагностики и прогнозирования онкологических заболеваний груди у женщин. В материале рассматривается выполнение экспериментальной части, построение модели для машинного обучения и предистория создания такой системы. На основании полученных данных было выявлено, что точность модели на тестовой выборке составляет 99,4 %.

Ключевые слова: метод логистической регрессии, биопсия новообразований груди, язык программирования Python, приложение, прогноз.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

DIAGNOSTICS AND PREDICTIONS OF BREAST CANCER USING THE PYTHON SOFTWARE AND THE LOGISTIC REGRESSION METHOD

VYACHESLAV V. PASEDKO[1], DMITRIY A. YAKIMOV[2], NATALIA V. VYGOVSKAYA[3]
[1]iTechArt Group (Mogilev, Republic of Belarus)
[2]Health care institution «Mogilev Regional Clinical Hospital» (Mogilev, Republic of Belarus)
[3]Belarusian-Russian university, (Mogilev, Republic of Belarus)

Abstract. The article is devoted to the description of the methodology for creating an automated software system for diagnosing and predicting breast cancer in women. The material discusses the implementation of the experimental part, building a model for machine learning and the prehistory of creating such a system. Based on the data obtained, it was revealed that the accuracy of the model on the test sample is 99,4 %.

Keywords: logistic regression method, breast biopsy, Python programming language, application, prognosis.
Conflict of interests. The authors declare no conflict of interests.

Введение

Распространение методов обучения искусственного интеллекта упирается в проблему четкого разграничения визуальных образов. Это требует исходного задания множества параметров, что, в свою очередь, приводит к необходимости поиска оптимального программного аппарата их учета. Наш опыт работы с медицинскими данными базировался на анализе исследований пациенток с опухолями груди, которые были выполнены в штате Висконсин, США. Данные были взяты из Kaggle – системы организации конкурсов по исследованию данных, а также социальной сети специалистов по обработке данных и машинному обучению и получены из оцифрованного изображения биопсии новообразований груди, собранных доктором Уильямом Х. Вольбергом [1] в университете Висконсин, больница Мэдисон, США. Данные являются характеристиками ядер клеток, пример изображения которых представлен на рис. 1.

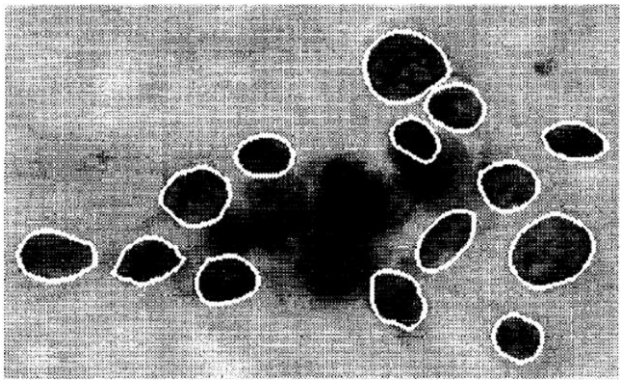

Рис. 1. Увеличенное изображение клеток при злокачественной опухоли груди
Fig. 1. Enlarged image of cells in breast cancer

Теоретический анализ

Для анализа данных использовался язык программирования Python [2] и его библиотеки для визуализации. В качестве метода для построения предиктивной математической модели был выбран метод логистической регрессии [3, 4], который также был имплементирован на языке Python. Целью

разработки программного обеспечения (ПО) было построение модели, способной из представленного материала давать заключение о наличии или отсутствии злокачественного роста. Для прогнозирования требуется взять пункционную биопсию и оцифровать её изображение. По этим новым данным можно будет сделать заключение о вероятности рака груди у пациентки.

Методика

В процессе работы был проведён пилотный анализ данных, а именно:

- рассчитаны статистические показатели для каждой переменной-предиктора;
- рассчитано распределение целевой переменной;
- построены графики распределений переменных-предикторов;
- рассчитана ядерная оценка плотности для переменных-предикторов;
- построены графики корреляций между переменными-предикторами, а также между предикторами и целевой переменной (рис.2);

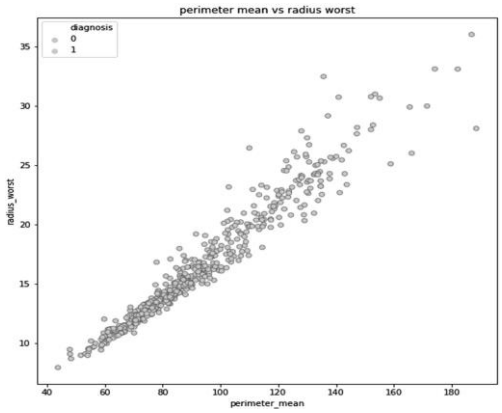

Рис. 2. Корреляция переменных *perimeter mean* и *radius worst*
Fig. 2. Correlation of *perimeter mean* and *radius worst* variables

- построен график тепловой карты, который отображает корреляции между переменными;
- найдены позитивно и негативно коррелирующие между собой предикторы.

Далее был выполнен этап подготовки к построению модели, а именно:

- был определен метод логистической регрессии для решения задачи классификации;
- был определен метод оценки результата выполнения модели.

Экспериментальная часть

Следующим шагом явилась подготовка датасета:

- была определена матрица переменных-предикторов, а также вектор целевой переменной;
- данные были стандартизированы (Feature Scaling);
- датасет был разделён на тестовую и обучающую выборки;
- были подобраны оптимальные гипер-параметры для построения модели;
- было произведено обучение модели на обучающей выборке, в процессе обучения были подобраны коэффициенты (веса) для каждой переменной-предиктора.

Результаты и их обсуждение

После применения модели на тестовой выборке были получены следующие результаты: точность модели на тестовой выборке – 99,4%;

Предложенная модель явилась удовлетворительным тестом принятия решения на основании математической обработки медицинских данных. В 1994 году Уильямом Вольбергом на кафедре компьютерных наук в Университете Висконсина впервые были описаны два медицинских приложения линейного программирования. Методы машинного обучения на основе линейного программирования были использованы для повышения точности рака груди и прогнозирования рецидива болезни. Первое приложение для диагностики рака груди использовало характеристики отдельных ядер кле-

ток, полученных методом аспирации тонкой иглой, чтобы отличить доброкачественные образования от злокачественных опухолей молочной железы. В настоящее время, с ростом возможностей фотографической техники и ее широким внедрением в медицинскую практику, появляется возможность поиска значимых параметров для описания микропрепарата. Причем такой подход может быть востребован на любом этапе, где необходимо выполнять гистологическое исследование.

Предложенная точность модели не является ограничителем по внедрению в клиническую практику, так как даже предварительный результат, полученный в краткие сроки клинически востребован. Появляется возможность сократить сроки госпитализации и временной нетрудоспособности.

Прогнозирование на основании только лишь картины микропрепарата выглядит менее востребованным. В то же время, всеобщая компьютеризация и облегчение систематизации данных позволят вносить в удачную программную модель новые параметры описания пациента: сроки госпитализации, методики лечения и тому подобное. Появляется возможность сравнивать прогностические результаты разных клиник, выявлять наиболее удачные лечебные процедуры.

Разработанное нами программное обеспечение не прошло апробацию в реальных клиниках, однако по схожести методик диагностики предполагается получить эффект от его использования.

Заключение

Разработана диагностическая система на современном языке программирования Python с удобным для пользователя интерфейсом, которая имеет практическую значимость в медицине. Была усовершенствована методика использования машинного обучения для разработки подобных систем путем использования метода логистической регрессии и экспериментальной обработки данных.

Список литературы

1. Mangasarian O.L., Street W.N., WolbergBreast W.H. Cancer Diagnosis and Prognosis Via Linear Programming / [Электрон. ресурс] Operations research. 1995. Т. 43. Режим доступа : https://doi.org/10.1287/opre.43.4.578. Дата доступа: 12.03.2020.

2. Коэльо Л.П., Ричарт В.В. Построение систем машинного обучения на языке Python. 2-е изд. / пер. с англ. Слимкина А.А. М. : ДМК Пресс, 2016. 302 с.

3. Леонов В. Логистическая регрессия в медицине и биологии [Электрон. ресурс] Биометрика. 2020. Режим доступа : http://www.biometrica.tomsk.ru/logit_0.htm. Дата доступа: 10.03.2020.

4. Паседько В.В., Выговская Н.В. Использование логистической регрессии при анализе медицинских данных // Материалы, оборудование и ресурсосберегающие технологии: материалы междунар. науч.-техн. конф. : М. Е. Лустенков (гл. ред.) [и др.]; Могилев, 23–24 апреля 2020 г. Могилев : Белорус.-Рос. ун-т, 2020. С. 499–500.

References

1. Mangasarian O.L., Street W.N., WolbergBreast W.H. Cancer Diagnosis and Prognosis Via Linear Programming / [Electronic resource] Operations research. 1995. Vol. 43. Access mode: https://doi.org/10.1287/opre.43.4.578. Access date: 12.03.2020.

2. Koehlo L.P., Richart V.V. Building machine learning systems in Python. 2nd ed. M.: DMK Press, 2016. 302 p.

3. Leonov V. Logistic regression in medicine and biology / [Electronic resource] Biometrika. 2020. Access mode : http://www.biometrica.tomsk.ru/logit_0.htm. Access date: 10.03.2020.

4. Pasedko V.V., Vygovskaya N.V. Using logistic regression in medical data analysis // Materials, equipment and resource-saving technologies: materialy mezhdunar. nauch.-tekhn. konf. : M.E. Lustenkov (gl. red) [I dr.]; Mogilev, 23–24 aprelya 2020 g. Mogilev: Belorus.-Ros. un-t, 2020. S.499–500.

Сведения об авторах

Паседько В.В., специалист, ООО «Техартгруп».
Якимов Д.А., к.м.н., врач, УЗ «Могилевская областная больница».
Выговская Н.В., старший преподаватель, Белорусско-Российский университет

Information about the authors

Pasedko V.V., specialist, iTechArt Group.
Yakimov D.A., PhD, doctor, Health care institution «Mogilev Regional Clinical Hospital»
Vygovskaya N.V., senior lecturer, Belarusian-Russian university.

Адрес для корреспонденции

212026, Республика Беларусь, г. Могилев, ул.

Address for correspondence

212026 Belarus, Mogilev, str Byalynitsky-Birulya,

Бялыницкого-Бирули, 12,
УЗ «Могилевская областная больница»

тел. +375 29 312 92 07;
Якимов Дмитрий Анатольевич

12,
Health care institution «Mogilev Regional Clinical Hospital»
tel. +375 29 312 92 07;
Yakimov Dmitry Anatolevich