УДК 61:615.1(062)(476-25)
ББК 52я73
А 43
ISBN 978-985-21-1008-2

Рустянова Д. Р.

ПРОГНОЗИРОВАНИЕ РИСКА РАЗВИТИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ

Научные руководители: д-р мед. наук, доц. Казакова А. В,
канд. мед. наук, доц. Yen-Kuang Lin
Кафедра акушерства и гинекологии
Самарский государственный медицинский университет, г. Самара
National Taiwan Sport University, Taoyuan

Актуальность. Рак молочной железы (РМЖ) является распространенным видом онкологического заболевания среди женского населения. По данным Всемирной Организации Здравоохранения в 2020 году было зарегистрировано свыше 2300000 случаев данного заболевания, скончалось – 685000 человек. Существует большое количество факторов риска, среди которых мутации генов BRCA 1, BRCA 2, дисгормональные заболевания молочных желез, раннее менархе, поздняя менопауза, эндокринопатии, психо-эмоциональные нагрузки, приводящие женщину к развитию онкологии груди. Надежный способ борьбы против РМЖ — предиктивная диагностика, которая возможна с внедрением в медицинскую практику возможностей машинного обучения (Machine Learning, ML).

Цель: создание алгоритма ML, способного прогнозировать предрасположенность к развитию рака груди на основании анамнестических особенностей пациенток.

Материалы и методы. В основу работы лег принцип работы модели Гейла. В нашем распоряжении был файл с закодированными медицинскими данными 14055 пациенток. Мы строили модель используя XGBoost-Classifier – ML-алгоритм, который учится самостоятельно генерировать правила для выполнения предсказаний на основании размеченных данных. Модель XGBoost является индустриальным стандартом для работы с большими данными. Программирование производилось на языке Python.

Результаты и их обсуждение. По извлеченной из исходного файла информации мы узнали, что 14 055 пациенток были из разных стран, разных рас и возрастов. Имелись закодированные данные о возрасте менархе и первых родов, наличии мутации в генах BRCA 1, BRCA 2, лобулярной или протоковой карциномы in situ, гиперпластических процессов в молочной железе и родственников первой степени родства с РМЖ. Средний возраст ($М_{возраст}$) исследуемых женщин составил 54 года, первых родов - 27 лет, менархе – 15,2 лет. Доля пациенток с мутациями в генах BRCA 1, BRCA 2 составила всего 5%. По тепловой карте корреляции, построенной для графического представления данных, отражающих насколько один признак линейно зависим от другого, мы увидели наибольшую взаимосвязь между РМЖ и наличием мутаций в генах (0,25), возрастом (0,17), наличием атипичной гиперплазии в анамнезе (0,2). По построенным Boxplots мы выявили, что медианный возраст ($Ме_{возраст}$) женщин с геном BRCA 1 – 50 лет, BRCA 2 - 55 лет. Лобулярный рак груди развивался у женщин, первые роды которых произошли в позднем репродуктивном периоде - от 40 лет и старше ($p<0.001$), $Ме_{возраст}$ - 60 лет. Была определена зависимость менархе от наличия генов BRCA 1, BRCA 2: у женщин с BRCA2 менархе наступает раньше - с 11 лет ($p<0.001$). Но $М_{возраст}$ начала менструации - 13 лет - одинаков для всех обследованных женщин. Достоверных взаимосвязей между принадлежностью к определенной расе и развитием РМЖ обнаружено не было. Построенный на основании XGBoost-Classifier ML-алгоритм был проверен при помощи трёх метрик качества: Roc-Auc (кривая ошибок) - 0,9; Accuracy (точность) – 0,93; Specificity (специфичность) – 0,98.

Выводы. На основании модели Гейла мы разработали собственный алгоритм машинного обучения, который по минимальному количеству необходимой информации о пациентах позволяет предсказывать риск развития РМЖ. При этом в нем будет минимальное количество ложноположительных срабатываний (2%).