УДК: 004.942, 004.421, 004.021, 004.048
ГРНТИ: 28.17.19, 76.01.11, 28.23.29

ПРЕДСКАЗАТЕЛЬНОЕ МОДЕЛИРОВАНИЕ МНОГОШАГОВЫХ КЛИНИЧЕСКИХ ПУТЕЙ НА ПРИМЕРЕ ЗАДАЧ РЕПРОДУКТИВНОЙ МЕДИЦИНЫ

Жданова Е.А. (Национальный исследовательский университет ИТМО)
Научный руководитель – с. н. с. НЦКР, к.т.н., Ковальчук С.В.
(Национальный исследовательский университет ИТМО)

Введение. В рамках многих существующих медицинских приложений существенное влияние на качество работы и применимость предсказательных моделей имеет учет контекста применения в рамках динамики состояния пациента и процессов оказания медицинской помощи. В такой ситуации зачастую традиционный подход машинного обучения с унифицированным учетом всех доступных на текущий момент факторов приводит к снижению качества и интерпретируемости предсказания[1]. Целью данной работы является построение прогностической модели на основании данных, полученных из Международного Центра Репродуктивной Медицины с анализом вопросов прикладной интерпретируемости и учета структуры процесса оказания медицинской помощи на примере лечения бесплодия. В качестве предикторов в работе будут использованы анамнезы пациентов, назначения врачей и показатели, полученные в результате обследований пациента. Планируется также выделить наиболее значимые предикторы и обеспечить высокий уровень интерпретируемости итоговой модели.

Основная часть. В рассматриваемом наборе данных из международного центра репродуктивной медицины целевая метка представлена столбцом результатов лечения. Для получения более полного представления о производительности модели была использована комбинацию таких метрик, как F1 score, precision, recall и ROC-AUC. В качестве метрики для оптимизации была использована ROC-AUC.

Так как в датасете присутствовали пропуски, то был использован метод интерполяции отсутствующих значений в наборе данных с помощью алгоритма K-ближайших соседей – KNN imputer. Он работает путем определения K ближайших наблюдений с не пропущенными значениями для данного наблюдения с пропущенными значениями. Затем недостающее значение вменяется путем взятия среднего (для числовых данных) или моды (для категориальных данных) значений K ближайших соседей. Метрика расстояния, используемая для определения ближайших соседей, может варьироваться, но для данного набора данных было использовано евклидово расстояние. Процесс интерполяции повторяется для каждого наблюдения с отсутствующими значениями до тех пор, пока все отсутствующие значения не будут заполнены. Так как, выбор числа ближайших соседей может повлиять на результаты интерполяции было решено попробовать различные значения K и оценить их влияние на результаты интерполяции. Значения K находились в диапазоне от 1 до 21. В качестве базовых оценщиков были рассмотрены алгоритмы SVC, Decision Tree, Random Forest. Лучшее значения целевой метрики в процессе обучения и оценки модели на валидационной выборке было получено для алгоритма Random Forest и числа ближайших соседей 21.

Далее было решено использовать библиотеку optuna с целью настройки гиперпараметров в алгоритме случайного леса для улучшения его производительность путем нахождения оптимальных значений для гиперпараметров. Известно, что алгоритм случайного леса склонен к переобучению из-за тенденции к слишком сильной подстройке под тренировочную выборку. В качестве гиперпараметров для оптимизации были выбраны количество деревьев в лесу, количество признаков, рассматриваемых для каждого разбиения, и максимальную глубину деревьев. Оптимизация значения для этих гиперпараметров позволила достичь значения 0.97 для метрики ROC-AUC.

Самыми значимыми предикторами для построенной модели являются

1) Количество плодов, визуализированных в полости матки.
2) Количество перенесённых эмбрионов в полость матки.

3) Количество фолликулов, которые были пропунктированы.

Для детектирования важности предикторов в обученной модели был использован метод объяснения результатов модели машинного обучения путем присвоения баллов важности признаков каждой входной переменной. В ходе исследования были выявлены следующие закономерности:

1. Чем больше было перенесено эмбрионов, тем выше вероятность наступления беременности.
2. Чем большее число фолликулов были пропунктированы, тем выше вероятность наступления беременности.
3. Чем больше количество зрелых ооцитов, которые были получены при пункции, тем выше вероятность наступления беременности.
4. При средняя оценке качества переносимого эмбриона равной 1, вероятность наступления беременности ниже.

Выводы. В ходе исследования были разработана модель, которая на основе числовых предикторов способна предсказать исход лечения бесплодия. Для модели были подобраны оптимальные значения гиперпараметров. Были выделены наиболее значимые признаки, участвующие в процессе принятия решений. Самыми значимыми предикторами для построенной модели являются:

1. Количество плодов, визуализированных в полости матки.
2. Количество перенесённых эмбрионов в полость матки.
3. Количество фолликулов, которые были пропунктированы.

Для детектирования важности предикторов в обученной модели был использован метод объяснения результатов модели машинного обучения путем присвоения баллов важности признаков каждой входной переменной. В ходе исследования были выявлены следующие закономерности:

1. Чем больше было перенесено эмбрионов, тем выше вероятность наступления беременности.
2. Чем большее число фолликулов были пропунктированы, тем выше вероятность наступления беременности.
3. Чем больше количество зрелых ооцитов было получено при пункции, тем выше вероятность наступления беременности.
4. При средней оценке качества переносимого эмбриона равной 1, вероятность наступления беременности ниже.

Список использованных источников:

1. Кузнецова А. В., Сенько О. В., Кузнецова Ю. О. Преодоление проблемы" черного ящика" при использовании методов машинного обучения в медицине //Врач и информационные технологии. – 2018. – №. S1. – С. 74-80.