УДК 004.89

ВЫЯВЛЕНИЕ ЭТНИЧЕСКОЙ И ГЕНДЕРНОЙ ПРЕДВЗЯТОСТИ В МЕДИЦИНСКИХ ML-СИСТЕМАХ

Айсер Б.Х., студент гр. ONT-2, II курс
Билькентский университет, г. Анкара

Выявление этнической и гендерной предвзятости в медицинских системах машинного обучения представляет собой одну из самых актуальных задач, стоящих перед современной медициной и наукой о данных. Системы машинного обучения становятся важными инструментами в области диагностики, принятия решений и прогнозирования, однако их широкое внедрение в медицинскую практику вызывает опасения по поводу того, как алгоритмические предсказания могут отражать или усиливать социальные и культурные предвзятости. Этническая и гендерная предвзятость в медицинских ML-системах может привести к неравенству в предоставлении медицинских услуг, неправильной интерпретации состояния здоровья и даже ухудшению исходов лечения для отдельных групп пациентов. Поэтому выявление и минимизация этих предвзятостей становится необходимым шагом в развитии справедливых и этичных алгоритмов [1].

Этническая и гендерная предвзятость может проявляться на разных стадиях разработки и использования медицинских систем машинного обучения, начиная от сбора данных и заканчивая предсказаниями и решениями, принимаемыми на основе этих данных. Эти предвзятости могут быть результатом как исторических, так и современных факторов, таких как неравенство в доступе к медицинским услугам, недостаточное представление определённых групп в обучающих наборах данных, а также системные социальные и экономические различия.

Например, если в тренировочном наборе данных преобладают пациенты определённой этнической группы или пол, модель, обученная на таких данных, может несправедливо относиться к другим группам, что приведёт к ошибочным диагнозам или предсказаниям, основывающимся на недооценке риска для этих пациентов. В медицинских приложениях, где ошибки могут быть крайне дорогостоящими, такой дисбаланс может иметь серьёзные последствия для здоровья пациентов [2].

Один из первых шагов в выявлении этнической и гендерной предвзятости – это анализ составляющих тренировочных данных, поскольку данные являются основой для обучения моделей. Проблемы могут возникать, если в

данных недостаточно представлены определённые этнические или гендерные группы, что приводит к их неправильному учету при принятии медицинских решений. Например, если в наборе данных для классификации заболеваний, таких как диабет или сердечно-сосудистые болезни, больше всего представлены мужчины, модель может недооценивать риск этих заболеваний для женщин. В таких случаях важно провести анализ дисбаланса в данных и применить методы, такие как перераспределение выборки или взвешивание примеров, чтобы минимизировать влияние этого дисбаланса. Это позволяет моделям лучше учитывать разнообразие пациентов и их здоровье, что снижает риск принятия неправильных решений.

Кроме того, необходимо применять методы аудита моделей на этапе тестирования для выявления скрытых предвзятостей, связанных с этнической или гендерной принадлежностью. В этом контексте важным инструментом является использование метрик справедливости, которые помогают измерить, как модель работает для разных групп пациентов. Например, можно использовать метрики ошибочных классификаций, такие как ошибка первого рода (ложноположительные результаты) и ошибка второго рода(ложноотрицательные результаты), чтобы увидеть, есть ли значительные различия в ошибках, которые модель делает для различных этнических или гендерных групп. Если модель оказывает более высокую частоту ложных диагнозов для одной группы, это может указывать на наличие предвзятости.

Методы контрфактической справедливости также могут быть полезны для выявления этнической и гендерной предвзятости. Этот метод позволяет анализировать, как изменения в чувствительных признаках, таких как пол или раса, влияют на результат предсказания модели. Например, можно проверить, как диагноз для пациента изменится, если изменить только его пол или этническую принадлежность, оставив остальные характеристики постоянными. Если модель изменяет диагноз при изменении только этих признаков, это может свидетельствовать о наличии этнической или гендерной предвзятости, что необходимо устранить [3].

Для более точного анализа можно использовать методы объясняемого ИИ, такие как SHAP или LIME, чтобы понять, какие признаки были наиболее важными для принятия решения моделью. Если чувствительные признаки, такие как пол или этническая принадлежность, оказывают значительное влияние на предсказания модели, это может быть поводом для пересмотра алгоритма. Применение таких методов позволяет не только выявить предвзятость, но и прояснить причины, по которым модель приняла определённое решение.

Кроме того, важно разрабатывать и внедрять методы регуляризации, чтобы минимизировать влияние чувствительных признаков на предсказания. Регуляризация позволяет корректировать веса признаков, что уменьшает

влияние тех, которые могут быть связаны с предвзятостью, и помогает моделью ориентироваться на более объективные и значимые характеристики, такие как медицинская история пациента, его состояние здоровья, результаты анализов и другие факторы.

Ещё одной важной частью работы с этнической и гендерной предвзятостью является мониторинг и обновление моделей в процессе их эксплуатации. В реальной практике данные могут изменяться со временем, и если модель не будет корректировать свои предсказания в ответ на эти изменения, она может начать работать менее эффективно или несправедливо. Например, если данные о пациентах изменяются, потому что они стали более разнообразными в этническом и гендерном плане, модель, которая была обучена на менее разнообразных данных, может начать ошибаться. Поэтому важно регулярно проводить тестирование и аудит моделей, чтобы убедиться в том, что они продолжают работать справедливо и точно для всех групп пациентов [4].

В заключение, выявление этнической и гендерной предвзятости в медицинских системах машинного обучения является неотъемлемой частью работы по созданию этичных и справедливых алгоритмов. Использование методов аудита, анализа данных, контрфактической справедливости, а также методов объясняемого ИИ помогает выявить и устранить скрытые формы предвзятости, что способствует улучшению качества медицинских услуг и предотвращает усиление существующих социальных неравенств.

Это требует от разработчиков моделей внимательного отношения к составу данных, корректной регуляризации и постоянного мониторинга, чтобы гарантировать, что технологии искусственного интеллекта служат на благо всех пациентов, независимо от их этнической принадлежности или гендерной идентичности.

Список литературы:

1. Obermeyer, Z. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations / Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan // Science. – 2019. – Vol. 366, № 6464. – P. 447–453. – DOI 10.1126/science.aax2342.

2. Gianfrancesco, M. A. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data / M. A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk // JAMA Internal Medicine. – 2018. – Vol. 178, № 11. – P. 1544–1547. – DOI 10.1001/jamainternmed.2018.3763.

3. Vyas, D. A. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms / D. A. Vyas, L. G. Eisenstein, D. S. Jones // New England Journal of Medicine. – 2020. – Vol. 383, № 9. – P. 874–882. – DOI 10.1056/NEJMms2004740.

4. Chen, I. Y. Ethical Machine Learning in Healthcare / I. Y. Chen, P. Szolovits, M. Ghassemi // Annual Review of Biomedical Data Science. – 2021. – Vol. 4. – P. 123–144. – DOI 10.1146/annurev-biodatasci-092820-114757.