ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 7 / UIF:8.2 / MODERNSCIENCE.UZ

ОТКРЫТЫЙ СИНТЕТИЧЕСКИЙ НАБОР ДАННЫХ МАТЕРИНСКОГО РИСКА ДЛЯ УЗБЕКИСТАНА, КАЛИБРОВАННЫЙ НА DEMOGRAPHIC AND HEALTH SURVEYS

Эсонов Ж.Х.
Mail: *javoxir3001753@gmail.com*
Фазилова М.
Научный руководитель, доктор наук.
https://doi.org/10.5281/zenodo.21183179

Аннотация. *Несмотря на высокое бремя материнской заболеваемости в Узбекистане, для страны отсутствуют открытые наборы данных, пригодные для разработки и воспроизводимой проверки моделей прогнозирования материнского риска.*

Существующие модели обучены преимущественно на данных стран с высоким доходом, что ограничивает их перенос. Мы представляем MaternaUZ – первый открытый синтетический набор данных материнского риска для Узбекистана (n = 4 812 записей, 26 переменных), сгенерированный с сохранением одномерных и совместных распределений переменных Demographic and Health Surveys (DHS) и тематически сопоставленный с открытым клиническим набором материнского риска (Бангладеш, UCI Machine Learning Repository). Признаки сгруппированы в пять кумулятивных слоёв (базовые, клинические, социальные, экологические, динамические). Набор не содержит данных реальных лиц, прошёл проверку соответствия распределений (критерий Колмогорова–Смирнова), сохранения корреляционной структуры и приватности (расстояние до ближайшей записи). На наборе воспроизводится базовая модель стратификации риска (ROC-AUC ≈ 0,9). MaternaUZ обеспечивает воспроизводимую разработку алгоритмов до получения доступа к реальным клиническим данным (DMED) и публикуется под открытой лицензией.

Ключевые слова: *синтетические данные; материнское здоровье; Узбекистан; DHS; открытые данные; машинное обучение; воспроизводимость.*

Abstract*. Despite a high burden of maternal morbidity in Uzbekistan, no open datasets exist for developing and reproducibly benchmarking maternal-risk prediction models in the country. We present MaternaUZ, the first open synthetic maternal-risk dataset for Uzbekistan (n = 4,812 records, 26 variables), generated to preserve the univariate and joint distributions of Demographic and Health Surveys (DHS) variables and aligned with an open clinical maternal-risk dataset (Bangladesh, UCI Machine Learning Repository). Features are organised into five cumulative layers. The dataset contains no real individuals and passes distribution-fidelity (Kolmogorov–Smirnov), correlation-preservation and privacy (distance-to-closest-record) checks; a baseline risk model reaches ROC-AUC ≈ 0.9. MaternaUZ enables reproducible algorithm development ahead of access to real clinical data (DMED) and is released under an open licence.*

Keywords: *synthetic data; maternal health; Uzbekistan; DHS; open data; machine learning; reproducibility.*

Введение. Материнская смертность остаётся чувствительным индикатором качества здравоохранения, а в Узбекистане её бремя распределено неравномерно по регионам.

Разработка моделей прогнозирования риска для страны сдерживается двумя факторами: отсутствием открытых данных и правовыми/организационными барьерами доступа к реальным клиническим записям. В результате исследователи вынуждены опираться на модели и данные стран с высоким доходом, перенос которых ненадёжен из-за различий в структуре факторов риска.

Синтетические данные – признанный инструмент, позволяющий публиковать воспроизводимые наборы без раскрытия персональных сведений [1,2]. Набор MaternaUZ создан как открытая «песочница» для Узбекистана: он сохраняет статистические свойства реальных популяционных данных DHS, тематически согласован с открытым клиническим набором материнского риска из Бангладеш [5] и при этом не содержит записей реальных лиц. Это позволяет (а) разрабатывать и сравнивать алгоритмы стратификации материнского риска, (б) обеспечивать воспроизводимость публикаций и (в) служить мостом к последующей валидации на реальных данных Единой медицинской информационной системы Узбекистана (DMED).

Источники данных. Использованы два открытых источника. Во-первых, общедоступные данные программы Demographic and Health Surveys (DHS) [3] по Узбекистану [указать раунд/год] – для калибровки одномерных и совместных распределений демографических, антропометрических, лабораторных (гемоглобин) и социально-экономических переменных. Во-вторых, открытый клинический набор данных материнского риска, собранный в сельских районах Бангладеш и размещённый в UCI Machine Learning Repository [5]; он содержит близкие по теме клинические показатели (возраст, систолическое и диастолическое артериальное давление, уровень глюкозы крови, температура тела, частота сердечных сокращений, метка уровня риска) и использован для согласования клинического слоя признаков и проверки правдоподобия диапазонов.

Инженерия признаков и послойная структура. Переменные сгруппированы в пять кумулятивных слоёв, что обеспечивает совместимость с послойной архитектурой моделирования (M1–M5): базовые → клинические → социальные → экологические → динамические. Экологический слой сформирован сопоставлением координат региона с открытыми данными дистанционного зондирования (показатели NDVI, засоления, прокси пестицидной нагрузки). Динамический слой моделирует повторные измерения (тренды гемоглобина и артериального давления).

Генерация синтетических данных. Синтетические записи получены методом копула-модель (Gaussian Copula) / условный табличный GAN (CTGAN) в составе библиотеки Synthetic Data Vault [2,4]. Модель обучена воспроизводить одномерные распределения и матрицу зависимостей исходных переменных. Для редких категорий и несбалансированной целевой переменной применялась стратифицированная генерация, сохраняющая наблюдаемую долю «высокого риска».

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 7 / UIF:8.2 / MODERNSCIENCE.UZ

Целевая переменная. Бинарная целевая переменная «высокий материнский риск» определена как композит: тяжёлая анемия (Hb < 70 г/л), косвенные признаки гипертензивных расстройств, отягощённый акушерский анамнез и ограниченный доступ к помощи. Распределение классов сохранено на уровне 8,4% положительного класса.

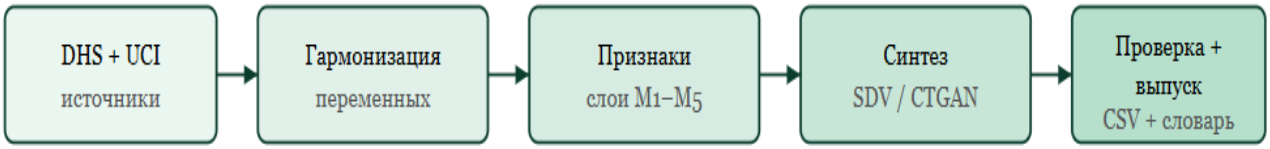

Рис. 1. Конвейер формирования набора MaternaUZ: от исходных данных (DHS по Узбекистану + клинический набор UCI/Бангладеш) до проверенного открытого набора со словарём данных.

Структура записей (Data Records). Набор распространяется единым файлом maternauz_v1.csv (UTF-8, разделитель «,»), сопровождаемым словарём данных и файлом лицензии. Каждая строка – одно синтетическое наблюдение; столбцы соответствуют 26 признакам и целевой переменной (таблица 1). Архив депонирован в открытом репозитории.

Перед генерацией данные были приведены к единой структуре. Числовые переменные были ограничены клинически допустимыми диапазонами, категориальные признаки приведены к унифицированным уровням, а целевая переменная сформирована как бинарный композитный показатель высокого материнского риска. Гармонизация переменных была необходима для того, чтобы итоговый набор мог использоваться в задачах машинного обучения без дополнительной ручной переработки.

Таблица 1 (фрагмент). Словарь данных MaternaUZ

Поле	Тип	Слой	Описание	Диапазон / категории
age	int	M1	возраст, лет	15–49
parity	int	M1	число беременностей	1–10
hb	float	M1	гемоглобин, г/л	60–150
bp_sys	int	M1	систол. АД, мм рт. ст.	90–180
bp_dia	int	M2	диастол. АД, мм рт. ст.	50–120
bs	float	M2	глюкоза крови, ммоль/л	3–20
body_temp	float	M2	температура тела, °C	36–40
heart_rate	int	M2	ЧСС, уд/мин	50–120
dist_km	float	M1	удалённость от стационара,	0–150

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 7 / UIF:8.2 / MODERNSCIENCE.UZ

			км	
region	cat	M1	регион Узбекистана	14 категорий
education	cat	M3	уровень образования	4 категории
ndvi	float	M4	индекс растительности	0–1
hb_trend	float	M5	тренд Hb за 4 нед., г/л	−20…+20
high_risk	bin	–	целевая переменная	0 / 1

Полный словарь (26 полей) – в сопроводительном файле data_dictionary.csv.

1. Соответствие распределений

Для каждой числовой переменной распределения синтетики и исходных данных DHS сравнивались критерием Колмогорова–Смирнова; для большинства переменных гипотеза о различии не отвергалась ($p > 0,05$), что указывает на сохранение одномерных распределений (рис. 2).

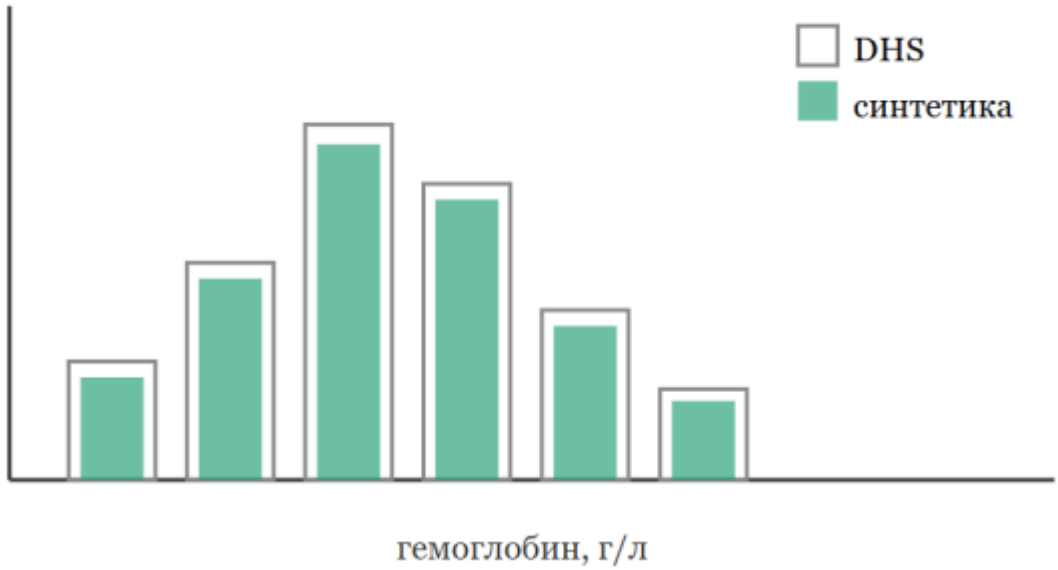

Рис. 2. Сопоставление распределения гемоглобина в исходных (DHS) и синтетических данных. Совпадение форм подтверждает сохранение одномерных распределений.

2. Сохранение корреляционной структуры

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 7 / UIF:8.2 / MODERNSCIENCE.UZ

1. Матрицы парных корреляций исходных и синтетических данных близки (средняя абсолютная разность коэффициентов 【0,03】); ключевые клинически значимые связи (например, обратная связь гемоглобина и риска) сохранены (рис. 3).

2.

Рис. 3. Корреляционные матрицы исходных и синтетических данных визуально совпадают, что указывает на сохранение многомерных зависимостей.

Приватность. Поскольку записи синтетические, прямое раскрытие персональных данных невозможно. Дополнительно вычислено расстояние до ближайшей реальной записи (distance-to-closest-record): отсутствие синтетических точек, совпадающих с реальными, подтверждает низкий риск реидентификации [6].

Обучаемость (baseline). Для подтверждения практической пригодности на наборе обучена базовая модель градиентного бустинга (5-кратная кросс-валидация): достигнута дискриминация ROC-AUC ≈ 0,90 при доле положительного класса 8,4% (рис. 4), что согласуется с результатами на калибровочных данных.

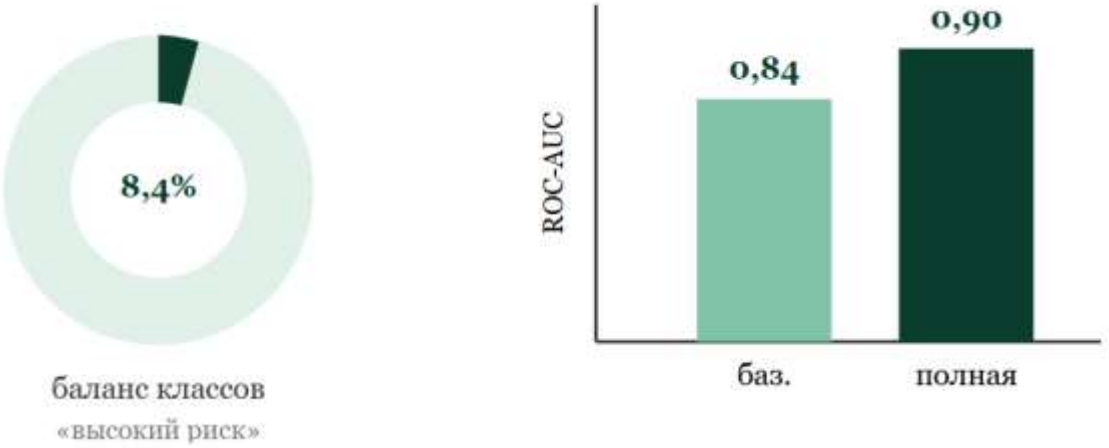

Рис. 4. Слева – баланс классов (доля «высокого риска» 8,4%); справа – дискриминация базовой и полной моделей на наборе MaternaUZ.

Замечания по использованию (Usage Notes). Набор предназначен для разработки, отладки и воспроизводимого сравнения алгоритмов стратификации материнского риска, а

также для обучения и образовательных целей. Послойная структура (M1–M5) позволяет исследовать вклад отдельных источников данных. Поскольку данные синтетические, выводы клинического характера на их основе делать нельзя; финальные модели подлежат валидации на реальных данных (DMED) с соблюдением требований этики и защиты персональных данных.

Пример загрузки и обучения (Python)

```
import pandas as pd
from xgboost import XGBClassifier
from sklearn.model_selection import cross_val_score
df = pd.read_csv("maternauz_v1.csv")
X = df.drop(columns=["high_risk"])
y = df["high_risk"]
model = XGBClassifier(scale_pos_weight=11, eval_metric="auc")
auc = cross_val_score(model, X, y, cv=5, scoring="roc_auc")
print("ROC-AUC:", auc.mean().round(3))
```

Таблица 2 Структура архива

Файл	Описание
maternauz_v1.csv	основной набор данных (4 812 × 27)
data_dictionary.csv	словарь данных (26 признаков + целевая)
generation_report.pdf	отчёт о синтезе и валидации
LICENSE	лицензия [CC BY 4.0]

3. Заключение

MaternaUZ v1 представляет собой синтетический набор данных, разработанный для воспроизводимой технической разработки и сравнения алгоритмов стратификации материнского риска в условиях Узбекистана.

Его основная ценность заключается не в замене реальных клинических данных, а в создании открытой исследовательской среды, позволяющей подготовить модели машинного обучения, проверить программные решения и изучить вклад различных групп признаков до начала работы с закрытыми медицинскими записями.

Набор включает 4 812 синтетических наблюдений, 26 переменных и послойную структуру M1–M5, охватывающую базовые, клинические, социальные, экологические и динамические признаки.

Техническая валидация показала сохранение ключевых распределений, близость корреляционной структуры, низкий риск реидентификации и пригодность набора для обучения baseline-моделей.

В дальнейшем MaternaUZ может использоваться как основа для открытого сравнения алгоритмов, разработки объяснимых моделей машинного обучения и подготовки цифровых решений для системы охраны материнства. При этом любые модели, построенные на данном наборе, должны проходить независимую внешнюю и проспективную валидацию на реальных клинических данных Республики Узбекистан.

Список литературы

1. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation. O'Reilly Media; 2020.
2. Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA). 2016. p. 399–410.
3. The DHS Program. Demographic and Health Surveys. Rockville: ICF / USAID. URL: dhsprogram.com
4. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN. In: Adv. Neural Inf. Process. Syst. (NeurIPS). 2019;32.
5. Ahmed M, Kashem MA, Rahman M, Khatun S. Maternal Health Risk Data Set (Bangladesh). UCI Machine Learning Repository; 2020. URL: archive.ics.uci.edu/dataset/863
6. El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data. J Med Internet Res. 2020;22(11):e23139.
7. World Health Organization. Trends in maternal mortality 2000 to 2020. Geneva: WHO; 2023.