Секция «Математика, информационные технологии и их приложения»

# Применение генеративных моделей машинного обучения для увеличения выборки в задачах медицинского прогнозирования.

***Сайфетдинов Салават Фаильевич***
*Аспирант*
Мордовский государственный университет им. Н.П. Огарёва, Факультет математики и информационных технологий, Саранск, Россия
*E-mail: ssayfetdinov@gmail.com*

Одной из фундаментальных проблем анализа и обработки информации в медицине является малый объем выборок в исследованиях. Построение прогностических моделей на таких данных классическими методами машинного обучения часто приводит к переобучению. В данной работе исследуется эффективность применения генеративных нейросетей: табличных вариационных автоэнкодеров (TVAE) и условных генеративно-состязательных сетей (CTGAN) для искусственного увеличения выборки данных пациентов, которые прошли терапию ожирения.

Исходная выборка составила 28 пациентов. Целевой переменной выступило клинически значимое снижение веса, которого достигли 21 человек. Алгоритмом Sequential Feature Selector были отобраны 5 ключевых предикторов: возраст, уровень инсулина, АЛТ, АСТ и мочевины. Логистическая регрессия на исходных данных продемонстрировала низкую прогностическую способность: ROC AUC на кросс-валидации (LOOCV) составил 0,639.

Результаты выявили ограничения методов SMOTE и CTGAN при работе с микровыборками. Алгоритм SMOTE показал нереалистично высокое качество (среднее значение ROC AUC $0.975 \pm 0.003$), поскольку этот метод формирует новые данные с помощью интерполяции между близкими объектами, что неизбежно ведет к переобучению. Нейросеть CTGAN, в свою очередь, не смогла стабильно обучиться на малом объеме данных (ROC AUC варьировался от 0.095 до 0.959). Результаты этой модели имели слишком большой разброс, и она оказалась нестабильной для увеличения нашей малой выборки.

Оптимальные результаты показал метод TVAE (среднее значение ROC AUC = $0.943 \pm 0.035$). Оценка качества сгенерированных данных показала, что он удовлетворительно воспроизводит индивидуальные характеристики реальных пациентов (совпадение показателей — 83.85%) и корректно сохраняет сложные взаимосвязи между медицинскими анализами (совпадение трендов — 72.37 %). Дополнительный анализ подтвердил, что алгоритм генерирует новые, уникальные профили, а не просто копирует исходные данные.

Полученные результаты приводят нас к выводу, что синтетическое увеличение выборки с помощью табличных вариационных автоэнкодеров (TVAE) можно рассматривать как перспективный подход для повышения прогностической ценности в условиях экстремального дефицита клинических данных.

## Источники и литература

1) www.who.int (Всемирная организация здравоохранения).
2) Gholampour A. Impact of medical data nature on ML classifiers for minority classes: SMOTE validity questionable // Machine Learning and Knowledge Extraction. 2024.
3) Hernández M. et al. A comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy // Front Digit Health. 2025.
4) Riley et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare // Lancet Digital Health. 2025.