

АДАПТАЦИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОСТПРАНДИАЛЬНОГО ГЛИКЕМИЧЕСКОГО ОТВЕТА У БЕРЕМЕННЫХ С ГЕСТАЦИОННЫМ ДИАБЕТОМ

**Исаков А.О. (ИТМО, ФГБУ НМИЦ им. В.А. Алмазова), Сашин В.А. (ИТМО)
Научный руководитель – кандидат технических наук, доцент, старший научный сотрудник Ватьян А.С. (ИТМО)**

Введение. Множество передовых исследований посвящено теме прогнозирования эпизодов гипергликемии, а также расчету необходимой дозы инсулина для больных диабетом первого и второго типов. Однако, в силу особенностей причин манифестации и протекания, относительно небольшая часть накопленных знаний может быть свободно перенесена в область терапии гестационного сахарного диабета (ГСД). ГСД характеризуется острым, но контролируемым нарушением углеводного обмена во время беременности. Поэтому, если это возможно, курирующие специалисты стремятся избежать преждевременного назначения инсулина, обращаясь к диетотерапии. Предложенное решение, основанное на алгоритмах машинного обучения, позволяет прогнозировать постпрандиальный гликемический ответ (ППГО) для пациенток с ГСД с целью информирования пациента о возможном патологическом изменении уровня глюкозы в интерстициальной жидкости (ИСЖ).

Основная часть. Поставленный вопрос сводится к решению задачи регрессии. Для формирования набора данных на базе медицинского центра ФГБУ НМИЦ им. В.А. Алмазова “Центр Персонализированной Медицины” были собраны записи о 105 женщинах с ГСД на 24-36 неделях беременности [1]. Целевыми переменными являются параметры вариабельности ППГО, выраженные следующими значениями: пиковый уровень глюкозы спустя час после приема пищи (BGMax), площадь под кривой гликемии через 1 час (AUC60) и инкрементальная площадь через 2 часа (iAUC120). Объем набора данных после применения фильтрации, предложенной в [2], составил 2800 записей о приемах пищи.

Основными категориями прогностических признаков были:

- параметры приема пищи: макро и микронутриенты;
- параметры непрерывного мониторинга глюкозы: набор дискретизированных значений кривой уровня глюкозы в ИСЖ за установленные интервалы до приема пищи (до 1 часа с шагом в 5 минут);
- параметры анкетирования: относительная частота потребления различных категорий продуктов питания в неделю;
- генетические параметры;
- параметры бактериального состава кишечника (экспериментальные данные).

Для решения задачи прогнозирования каждой из 3 целевых переменных были проведены эксперименты по обучению моделей машинного обучения 2 популярных архитектур:

1. экстремального градиентного бустинга деревьев решений (фреймворк XGBoost);
2. нейронной сети (многослойного перцептрона) с 2 скрытыми слоями (фреймворки PyTorch Lightning и ClearML).
3. Алгоритм экспериментов включал в себя следующие этапы:
 1. Разделение предобработанного набора данных на обучающую, валидационную и тестовую выборки в процентном соотношении 75/15/10 таким образом, чтобы записи об одном пациенте не присутствовали в разных выборках.
 2. Для нейронной сети: итеративное заполнение немногочисленных пропусков в

данных при помощи байесовской гребневой регрессии (IterativeImputer с BayesianRidge предиктором из библиотеки scikit-learn), а также стандартизация данных.

3. Подбор оптимальных значений гиперпараметров моделей при помощи алгоритма Tree-structured Parzen Estimator фреймворка Optuna для обеих архитектур. Дополнительно, для оркестрации и трекинга экспериментов при подборе гиперпараметров нейронной сети был применен фреймворк ClearML.
4. Выбор лучшего варианта модели по метрике RMSE на валидационной выборке.
5. Подсчет метрик качества лучших моделей на тестовом сете: MAE, MSE, RMSE.

Выводы. В результате экспериментов метрики моделей XGBoost на тестовой выборке составили:

1. AUC60: MAE=0.35, MSE=0.20, RMSE=0.44.
2. BGMax: MAE=0.54, MSE=0.52, RMSE=0.72.
3. iAUC120: MAE=0.38, MSE=0.26, RMSE=0.51.

Метрики нейронных сетей оказались значительно хуже:

1. AUC60: MAE=0.45, MSE=0.36, RMSE=0.6.
2. BGMax: MAE=0.9, MSE=2.58, RMSE=1.61.
3. iAUC120: MAE=0.84, MSE=2.68, RMSE=1.64.

Таким образом, модели XGBoost показали наилучшие метрики качества предсказания всех 3 целевых переменных.

Более того, в ходе работы были выделены направления будущих исследований, такие как: исследование влияния данных микробиома на качество предсказания параметров variability ППГО, прогнозирование большего количества таких параметров, а также возможное усовершенствование использованной архитектуры нейронной сети для улучшения качества прогнозирования.

Список использованных источников:

1. Popova P, Anopova A, Vasukova E, Isakov A, Eriskovskaya A, Degilevich A, Pustozero E, Tkachuk A, Pashkova K, Krasnova N, Kokina M, Nemykina I, Pervunina T, Li O, Grineva E and Shlyakhto E (2023) Trial protocol for the study of recommendation system DiaCompanion with personalized dietary recommendations for women with gestational diabetes mellitus (DiaCompanion I). *Front. Endocrinol.* 14:1168688. doi: 10.3389/fendo.2023.1168688 – URL: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1168688/full>
2. E. A. Pustozero et al., "Machine Learning Approach for Postprandial Blood Glucose Prediction in Gestational Diabetes Mellitus," in *IEEE Access*, vol. 8, pp. 219308-219321, 2020, doi: 10.1109/ACCESS.2020.3042483 – URL: <https://ieeexplore.ieee.org/document/9281297>

Автор _____ Исаков А.О.
Научный руководитель _____ Ватьян А.С.