ИИ в промышленности

ИТМО

Прогнозирование рецидива диабетического кетоацидоза

Студент:
Садыхов Олег Бахадурович, гр. J4151
Научный руководитель:
Леоненко Василий Николаевич, к. ф.-м. н., доцент

Постановка проблемы

Диабетический кетоацидоз (ДКА) – острое, угрожающее жизни метаболическое осложнение сахарного диабета. Рецидивы ДКА – ключевой индикатор, приводящий к тяжелым нарушениям функциональной способности организма. Существующие клинические рекомендации основаны на небольшом числе известных факторов риска (например, содержание HbA1c), но их способность прогнозировать рецидив часто недостаточна. Недавние исследования [1-7] в области прогнозирования DKA с применением машинного обучения указывают на распространенную проблему "черного ящика", то есть отсутствие понимания глубинных причин их прогнозов, что подрывает доверие врачей.

Цель и задачи

Цель исследования: Разработка интерпретируемой модели, объясняющей ключевые факторы риска, для задачи прогнозирования рецидива ДКА для российской клинической практики.

Задачи:

- **Систематический обзор современных научных публикаций по применению методов машинного обучения для прогнозирования ДКА и его факторов риска.**
- **Предобработка и анализ реальных клинических данных пациентов.**
- **Удаление аномалий, масштабирование данных.**
- **Сравнение эффективности различных алгоритмов машинного обучения на подготовленных данных.**
- **Отбор ключевых признаков, в частности проведение SHAP-анализа.**
- **Сравнение итоговой интерпретируемой модели по комплексу метрик (ROC-AUC, F1-score, Precision, Recall).**

Обзор аналогичных решений/исследований

Источник	Решения	Недостатки
Ihalapathirana A. et al., 2023	Применены бустинговые алгоритмы, отбор признаков методом ReliefF, стратификация по полу.	Ограничение набора данных - рассматриваются только взрослые пациенты с 1 типом диабета.
Williams D. D. et al., 2023	Разработана и протестирована модель глубокого обучения типа LSTM для прогнозирования риска госпитализации по поводу ДКА у молодежи.	Отсутствие SHAP/XAI анализа.
Eid W. M. et al., 2023	Подтверждение эффективности классических алгоритмов ML на структурированных данных.	Ограниченный набор признаков.

Вывод: Существующие решения подтверждают эффективность ML для прогнозирования ДКА, но имеют существенные ограничения: недостаток интерпретируемости моделей, ограниченные наборы признаков и отсутствие исследования на российской популяции. Наша работа направлена на создание интерпретируемой модели, учитывающей эти особенности.

Методы и инструменты исследования

Методы: машинное обучение, анализ данных, интерпретация моделей.

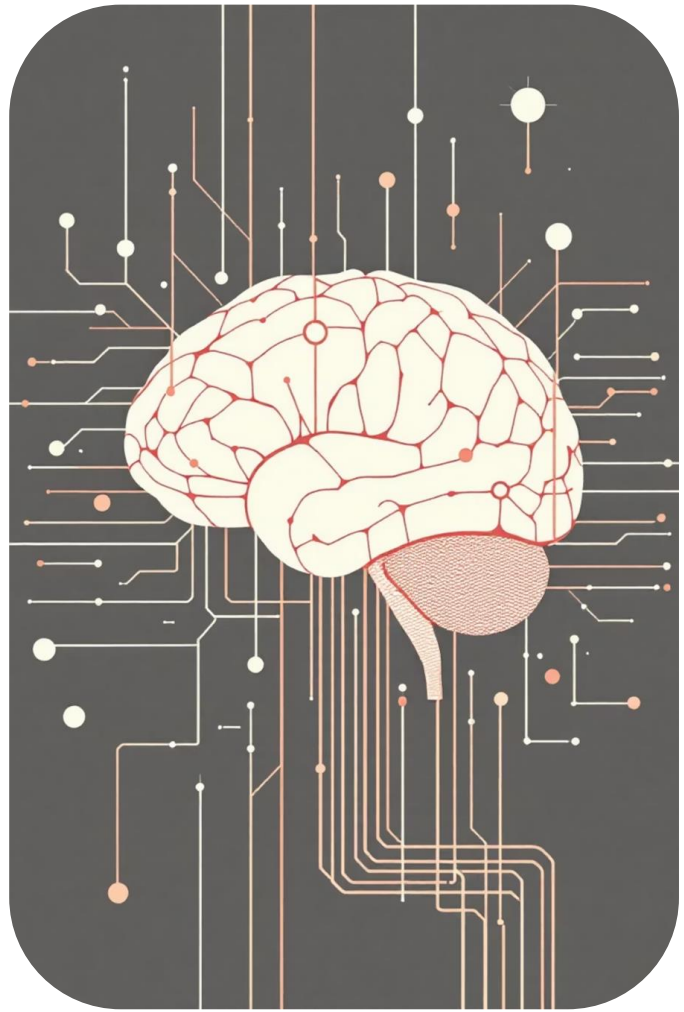
Инструменты: Python, Google Colab, Pandas, NumPy, Scikit-learn, CatBoost, XGBoost, SHAP, Matplotlib, Seaborn.

Данные: Анонимизированный клинический датасет пациентов с ДКА, предоставленный одной из больниц Санкт-Петербурга.

Научная новизна

Впервые предложен комплексный метод анализа клинических данных для прогнозирования рецидива ДКА в России, позволяющий учитывать и ранжировать вклад широкого спектра признаков с помощью методов XAI (SHAP).

Результаты исследования выявляют основные факторы риска рецидива ДКА в российской популяции. Это может помочь в создании моделей, учитывающих индивидуальные особенности пациентов с целью преждевременного обнаружения рецидива ДКА.



Ход решения

Сбор и подготовка данных

- Использован реальный клинический датасет
- Собрана первичная база данных пациентов с эпизодами ДКА (195 человек)
- Проведен разведочный анализ данных для оценки качества, распределения признаков.
- Удалены квазиконстантные и неинформативные признаки
- Удалены аномальные данные

Отбор признаков

- Выполнена предварительная фильтрация признаков: удалены колонки с применением жадного алгоритма
- Проведен экспертный отбор признаков совместно с врачами-эндокринологами для формирования финального набора переменных, имеющих клиническую интерпретацию и релевантность для прогноза рецидива ДКА

Ход решения

Предварительное обучение модели

- Использованы архитектуры ML (CatBoost, XGBoost, Случайный лес, Логистическая регрессия)
- Данные разделены на train и test 80/20
- Проведена кросс валидация на train для 4 фолдов
- Для оценки стабильности моделей проведена технология бутстрепа с последующим расчетом доверительных интервалов ROC-AUC.
- На данном этапе модели обучены без углубленного подбора гиперпараметров для получения baseline-оценки их прогнозной способности.

Исследование разделено на два подхода для анализа влияния анамнеза

- С учетом количества ДКА в анамнезе.
- Без учета количества ДКА в анамнезе.

Оценка качества

- AUC
- Classification report (precision, recall, f1-score, accuracy)

Ход решения

Предварительная точность моделей на кросс-валидации:

1. Вариант с ДКА в анамнезе

CV	AUC	F1-score для 1 класса	Accuracy
CatBoost	0.931	0.89	0.92
XGBoost	0.930	0.85	0.89
Random Forest	0.920	0.82	0.88
Logistic regression	0.921	0.81	0.87

2. Вариант без ДКА в анамнезе

CV	AUC	F1-score для 1 класса	Accuracy
CatBoost	0.737	0.49	0.71
XGBoost	0.597	0.34	0.60
Random Forest	0.731	0.29	0.68
Logistic regression	0.688	0.45	0.64

Ход решения

Предварительная точность моделей на тесте:

1. Вариант с ДКА в анамнезе

TEST	AUC	F1-score для 1 класса	Accuracy
CatBoost	0.993	0.92	0.94
XGBoost	0.970	0.92	0.94
Random Forest	0.997	0.92	0.94
Logistic regression	0.980	0.78	0.86

2. Вариант без ДКА в анамнезе

TEST	AUC	F1-score для 1 класса	Accuracy
CatBoost	0.836	0.53	0.75
XGBoost	0.729	0.43	0.64
Random Forest	0.826	0.38	0.72
Logistic regression	0.742	0.55	0.72

Ход решения

Для рассмотрения второго варианта работы проведен подбор гиперпараметров методом GridSearchCV для четырёх моделей и отбор признаков с использованием жадного алгоритма с последующим согласованием с экспертами-эндокринологами.

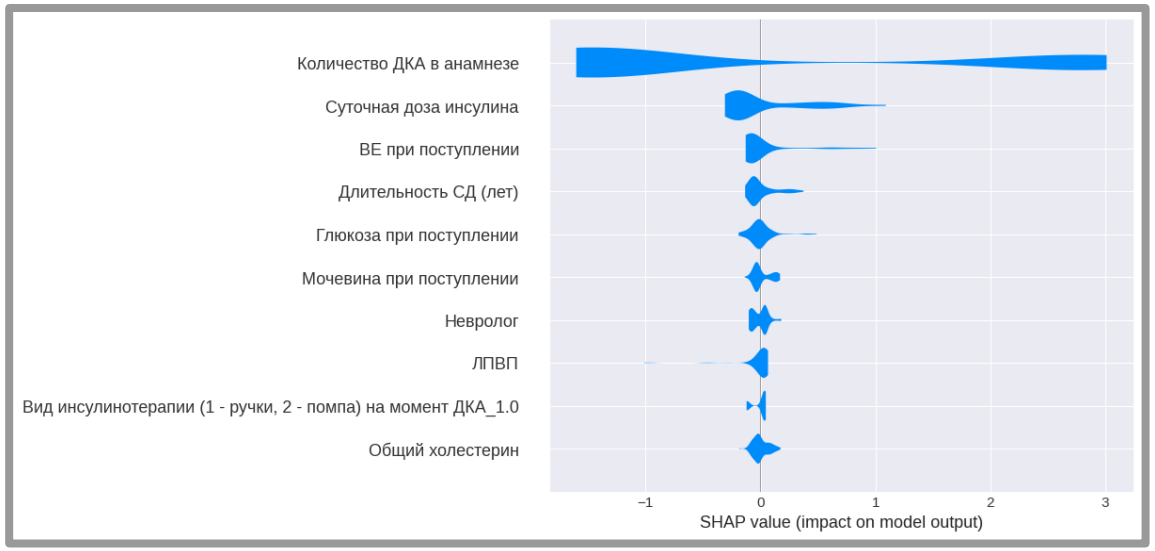
Наилучший результат показала модель случайного леса.

TEST	AUC	F1-score для 1 класса	Accuracy
Random Forest	0.829	0.64	0.78

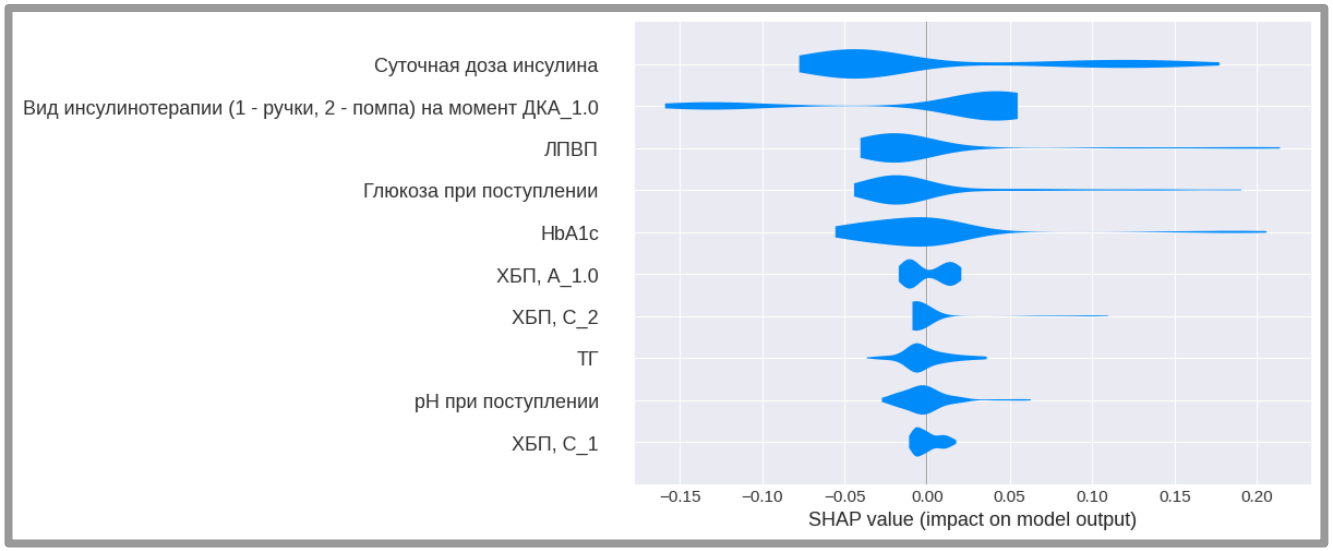
Ход решения

В каждом из вариантов проведен SHAP-анализ для лучших моделей для определения ключевых признаков, на которые ссылается модель при прогнозировании первого класса таргета:

1 Вариант



2 Вариант



Основные результаты:

GIT REPOSITORY: https://github.com/Lempopone/dka-relapse-prediction.git

1 вариант	2 вариант
Accuracy на тестовой выборке: 94%	Accuracy на тестовой выборке: 78%
AUC: 0.997	AUC: 0.829

Топ 5 признаков по важности SHAP		Топ 5 признаков по важности SHAP	
Количество ДКА в анамнезе	1.917134	**Суточная доза инсулина**	0.063852
Суточная доза инсулина	0.290616	**Вид инсулинотерапии**	0.034719
BE при поступлении	0.144938	**ЛПВП**	0.023290
Длительность СД (лет)	0.089690	**Глюкоза при поступлении**	0.030826
Глюкоза при поступлении	0.063292	HbA1c	0.029524

Основные результаты:

Выводы:

- **Получен и полностью обработан клинический датасет.**
- **Разработаны и обучены ML-модели по двум стратегиям: с учетом анамнеза ДКА и без.**
- **Методом SHAP выявлены и ранжированы ключевые факторы риска рецидива.**

Ожидаемые результаты

Получение и анализ расширенной базы данных для выявления ключевых для российской популяции предикторов ДКА.

Разработка интерпретируемой ML-модели, превосходящей baseline и валидированной на новых данных.

Создание рабочего прототипа системы (API + интерфейс) для поддержки врачебных решений в реальном времени.

Заключение

В рамках первого семестра была сформулирована цель исследования, поставлены задачи и выполнен ключевой подготовительный этап: произведён сбор и обработка клинического датасета, а также обучены и протестированы базовые модели машинного обучения для прогнозирования рецидива ДКА.

Наследующем этапе работы планируется расширение и обогащение датасета с целью повышения точности прогноза, а также разработка функционального прототипа системы поддержки врачебных решений в реальном времени.

Степень готовности результатов к публикации

Полученные результаты демонстрируют прочную основу и потенциал для разработки интерпретируемой модели прогнозирования рецидивов ДКА. С учетом высокой ответственности и специфики медицинских положений для последующей публикации необходимы расширение клинического датасета и получение разрешения от медицинского учреждения.

Литература:

1. Ihalapathirana A. et al. Explainable Artificial Intelligence to predict clinical outcomes in type 1 diabetes and relapsing-remitting multiple sclerosis adult patients //Informatics in Medicine Unlocked. – 2023. – Т. 42. – С. 101349.
2. Williams D. D. et al. An “All-Data-on-Hand” deep learning model to predict hospitalization for diabetic ketoacidosis in youth with type 1 diabetes: Development and validation study //JMIR diabetes. – 2023. – Т. 8. – С. e47592.
3. Eid W. M. et al. Predicting diabetic ketoacidosis in pediatric patients using machine learning //F1000Research. – 2023. – Т. 12. – С. 611.
4. Alsamhori J. F. et al. The implication of artificial intelligence in diabetic ketoacidosis //Avicenna. – 2025. – Т. 2025. – №. 2. – С. 16.
5. Earnest A. et al. Machine learning techniques to predict diabetic ketoacidosis and HbA1c above 7% among individuals with type 1 diabetes—A large multi-centre study in Australia and New Zealand //Nutrition, Metabolism and Cardiovascular Diseases. – 2025. – С. 103861.
6. Subramanian D., Sonabend R., Singh I. A Machine Learning Model for Risk Stratification of Postdiagnosis Diabetic Ketoacidosis Hospitalization in Pediatric Type 1 Diabetes: Retrospective Study //JMIR diabetes. – 2024. – Т. 9. – №. 1. – С. e53338.
7. Fan T. et al. Predicting the risk factors of diabetic ketoacidosis-associated acute kidney injury: A machine learning approach using XGBoost //Frontiers in public health. – 2023. – Т. 11. – С. 1087297.

Спасибо
за внимание!

ITMO more than a UNIVERSITY