ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2022. Т. 4, № 2 (9). С. 30–36 СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ http://siit.ugatu.su

УДК 004.048 doi 10.54708/26585014_2022_42930

Анализ эндокринологических данных на основе моделей классификации

Г. Р. Шахмаметова [1], А. Д. Христодуло [2], С. П. Береговая [3]

[1] shakhgouzel@mail.ru, [2] annhrist@mail.ru, [3] sofyaklimets@gmail.com

[1, 2] ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)
[2] Университет г. Тренто, Италия
[3] МОГБУЗ «Городская поликлиника», Магадан

Поступила в редакцию 24.06.2022 г.

Аннотация. Приведены результаты исследования данных эндокринологического мониторинга по Республике Башкортостан с помощью различных моделей классификации: моделей на основе логистической регрессии, деревьев решений и случайного леса. В качестве исходных данных использовались данные эндокринологического мониторинга по заболеваемости сахарным диабетом первого типа по Республике Башкортостан. Используемым инструментом является облачный сервис «Microsoft Azure: Machine Learning Studio». Целью исследования является обоснованный, экспериментально подтвержденный выбор модели классификации для анализа и интерпретации данных эндокринологического мониторинга для определения закономерностей по заболеваемости сахарным диабетом I типа. В исследовании проводится классификация набора данных, включающего более 150 тыс. записей, тремя алгоритмами классификации (логистичекой регрессией, деревьями решений и случайным лесом), оценка каждой модели различными метриками, интерпретация полученных результатов и выбор наиболее подходящей модели классификации для указанного набора данных для проведения дальнейших исследований данных эндокринологического мониторинга.

Ключевые слова: анализ данных; машинное обучение; эндокринологические данные; классификация; деревья решений; логистическая регрессия; случайный лес.

ВВЕДЕНИЕ

Искусственный интеллект (ИИ) и машинное обучение сегодня успешно находят свое применение в сфере медицины, где помогают решать широкий спектр задач [1]. Одним из перспективных направлений применения методов машинного обучения и ИИ является анализ биомедицинских данных для выявления скрытых закономерной и знаний [2]. Рассматриваемая тема является актуальной на сегодняшний день, так как ввиду постоянного и интенсивного совершенствования информационных технологий, с большой скоростью накапливаются большие массивы данных, в том числе биомедицинских [3]. В связи с этим появляется потребность в обработке и анализе накапливаемых данных с помощью методов машинного обучения [4].

Повышение эффективности лечения сахарного диабета является на сегодняшний день одним из актуальных направлений в современной медицине и медицинской информатике, так как, согласно данным экспертов, данное заболевание представляет собой реальную угрозу за счет ранней инвалидизации и высокой смертности от сосудистых заболеваний [5]. Эндокринологический мониторинг важен для контроля сахарного диабета, так как данное заболевание без должного лечения ведет к тяжелым последствиям [6]. В глобальном докладе ВОЗ 2020 г. отмечается, что с 1980 года число людей, страдающих диабетом, увели-

чилось в 4 раза, а к 2030 году диабет может стать 7-й по счету причиной смерти во всем мире [7]. Таким образом, сфера анализа эндокринологических данных, а именно сахарного диабета, заслуживает отдельного внимания из-за остроты и динамики роста заболевания.

Во втором разделе данной статьи рассматриваются особенности инструмента «Microsoft Azure» для анализа биомедицинских данных; третий раздел посвящен предобработке исходного набора эндокринологических данных; в четвертом разделе рассматривается анализ данных на основе моделей классификации, таких как логистическая регрессия, деревья решений и случайный лес; пятый раздел посвящен анализу полученных результатов.

MICROSOFT AZURE В АНАЛИЗЕ БИОМЕДИЦИНСКИХ ДАННЫХ

Microsoft Azure – сервис облачных вычислений для управления приложениями через дата центры Microsoft, предоставляет пользователям такие модели как: SaaS – программное обеспечение как услуга, PaaS – платформа как услуга и IaaS – инфраструктура как услуга, а также поддерживает множество языков программирования, инструментов и программных сред [8].

Одним из основных компонентов цифровой медицины сегодня является медицинский клиентский портал, который позволяет собирать данные о пациентах и организовывать взаимодействие пациентов с медицинскими специалистами [9]. Одним из порталов с такой архитектурой является инфраструктура Azure Well Architected Framework, предоставляемая Microsoft Azure. Потенциальные варианты использования данной архитектуры:

– отслеживание статистики с носимого устройства;

– взаимодействие с медицинским специалистом на расстоянии и предоставление пользователю доступа к медицинским данным;

– возможность составления личного графика приема лекарств, с помощью которого можно автоматически пополнять их запасы и отслеживать их прием;

– контроль лишнего веса и уровня глюкозы через взаимодействие с куратором по здоровому питанию.

Среди вариантов применения сервиса Microsoft Azure в здравоохранении можно также отметить:

– непрерывный онлайн мониторинг состояния пациентов;

– клинический анализ медицинских изображений с помощью искусственного интеллекта Azure;

– получение сведения о геноме человека с помощью технологии Genomics;

– служба Health Bot, разрабатывающая виртуальных медицинских помощников.

Инструменты Microsoft Azure для анализа биомедицинских данных:

– Azure Synapse Analytics – хранение, обработка, анализ и визуализация имеющихся клинических данных;

– API Azure для здравоохранения – безопасное и ускоренное управление защищенными данными;

– высокопроизводительные вычисления – использование высокопроизводительной биоинформационной инфраструктуры для обработки данных в области геномики;

– технология blueprint – прогнозирование количества требуемого персонала, свободных койко-мест в госпитале и др. [10].

Анализ достоинств и недостатков использования Microsoft Azure в анализе биомедицинских данных приведен в табл. 1.

Таблица 1

Анализ достоинств и недостатков Microsoft Azure для обработки биомедицинских данных

	Достоинства	Недостатки
1	Безопасность. Microsoft Azure –облачная служба, что означает надежную защиту всех данных клиентов и организаций	Проблема пользования новыми сервисами, их нестабильность и ненадежность сразу после выпуска
2	Мощные инструменты для анализа данных и реализации алгоритмов машинного обучения	Требуется ИТ-специалист для эффективного управления данными
3	Децентрализация клинических испытаний с удаленным мониторингом, что очень востребовано, если пациент страдает хроническим заболеванием	Необходимо наличие опыта работы с платформой при переходе в работе с локальных сервисов на облачные
4	Удобная визуализация больших объемов данных, унификация больших объемов данных в облачных хранилищах	Стратегия «единого поставщика», что увеличивает риск потери данных

Учитывая достоинства и недостатки Microsoft Azure, было принято решение об использовании сервиса для анализа данных эндокринологического мониторинга.

ИСХОДНЫЕ ДАННЫЕ И ПРЕДОБРАБОТКА

Общее количество записей в исходных данных составляет 168340 строк. Каждая запись имеет следующие характеристики: Код пациента, Дата рождения, Пол, Дата постановки на учет, Тип СД, Инсулин, HbA1c и др., всего 27.

Сценарий предобработки включает несколько этапов: сортировка, преобразование формата, перевод, очистка.

Исходный набор данных включает записи о пациентах со следующими типами сахарного диабета – I тип СД, II тип СД, гестационный СД. Для дальнейшего анализа необходимо отсортировать данные по типу СД. В результате получаем 3 выборки, соответствующие первому, второму и гестационному типам диабета.

Далее файл с данными необходимо преобразовать в формат csv с разделительными запятыми. Для этого каждая ячейка должна соответствовать общему формату данных. В исходном наборе присутствуют даты, которые являются числовыми форматами данных, необходимо перевести их в общий формат.

После конвертирования файла в формат csv, данные следует перевести с русского на английский язык, так как Microsoft Azure не поддерживает русский язык.

По окончании предобработки данных исходный набор данных состоит из трех выборок формата csv с первым, вторым и гестационным типами диабета, готовый к анализу в Microsoft Azure.

Заключительным этапом предобработки является очистка пропущенных значений в данных (рис. 1).

Рис. 1. Статистика очищенного набора данных на примере столбца «Medicine»

АНАЛИЗ ЭНДОКРИНОЛОГИЧЕСКИХ ДАННЫХ

Для анализа предобработанных данных был использован сервис Microsoft Azure Machine Learning [11]. До начала обучения моделей и анализа данных необходимо создать облачный вычислительный кластер и загрузить набор данных. Визуализация этапа представлена на рис. 2.

Рис. 2. Визуализация выборки данных СД I типа

В процессе визуализации были отображены все данные пациентов с СД первого типа: 27 столбцов с признаками и первые 50 из 10503 строк, соответствующие записям о пациентах.

Алгоритм классификации на основе логистической регрессии.

Для построения и обучения классификационной модели необходимо наличие конвейера с вычислительным модулем двухклассовой логистической регрессии. Процесс создания конвейера и обучения модели по параметру «Пол» приведен на рис. 3.

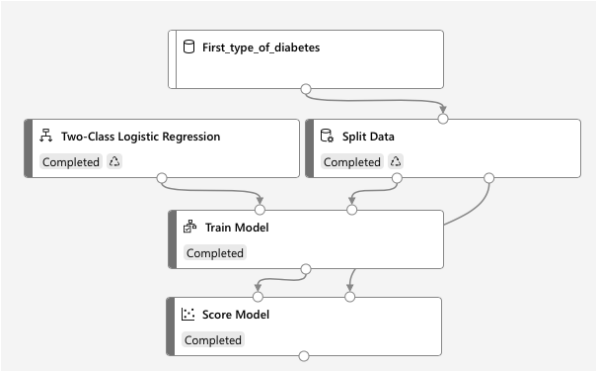

Рис. 3. Конвейер создания и обучения модели логистической регрессии

Визуализация результатов в блоке оценки модели приведена на рис. 4.

Рис. 4. Визуализация результатов классификации

В колонке справа приведены результаты классификации. В колонке пола лишь 2 уникальных значения (мужской и женский пол), отсутствуют пропущенные значения, а тип признака – строка (String Label). Добавился столбец меток с оценкой, где содержатся спрогнозированные значения меток, а также столбец с оценкой, которое указывает вероятность положительного прогноза. Столбец оценок указывает на количество пациентов мужского и женского пола среди больных сахарным диабетом первого типа. Количество пациентов мужского пола значительно превалирует (3473 значения, тогда как женщин – 2244).

Добавим модуль анализа модели (Evaluate Model) к уже существующему конвейеру, оценивающую качество модели, запустим эксперимент и визуализируем результаты (рис. 5).

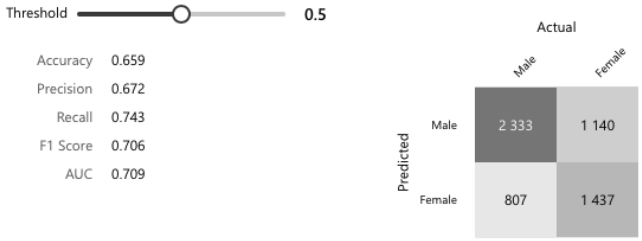

Рис. 5. Оценка модели на основе логистической регрессии

Слева представлены метрики оценки модели классификации:

– правильность (accuracy) – оценивает показатель качества модели (0,659);

– точность (precision) – пропорция фактических результатов ко всем положительным результатам (0,672);

– полнота (recall) – доля объема соответствующих экземпляров, которые были получены (0,743);

– оценка F1 (F1 score) – взвешенное среднее значение точности и полноты, где наилучшим значением считается «1» (0,706);

– AUC – определяет качество прогнозов модели независимо от порога классификации.

Справа представлена матрица ошибок для модели, которая представляет собой сетку 2×2, отображающую прогнозируемое (Predicted) и фактическое (Actual) значение для классов мужского и женского пола.

Верхняя левая ячейка матрицы, содержащая значение 2333, указывает количество истинных положительных результатов для значения показателей мужского пола. Это означает, что модель спрогнозировала верное значение для пациентов мужчин в 2333 случаях.

Нижняя левая ячейка с показателем 807 высчитывает количество ложных положительных результатов, другими словами, количество раз, когда предполагалось, что пациент – женщина, но в результате пациентом оказывался мужчина.

Верхняя правая ячейка матрицы со значением 1140 указывает количество ложных положительных результатов для значения показателей женского пола. Это означает, что в 1140 случаях модель спрогнозировала,

что пациентом является мужчина, когда пациентом оказывалась женщина.

Наконец, нижняя правая ячейка со значением 1437 указывает количество истинных положительных результатов для целевого значения показателей женского пола. Таким образом, в 1437 случаях модель правильно предсказала, что пациент – женщина.

Сложив показатели истинных положительных результатов по диагонали, и ложных положительных результатов по другой диагонали, можно получить общее число точных прогнозов и ложных прогнозов соответственно. Итак, количество точных прогнозов – 3770, а количество ошибочных – 1947, что является достаточно большой ошибкой.

Алгоритм классификации на основе деревьев решений.

Для реализации модели на основе деревьев решений необходимо создать вычислительный конвейер, добавив модуль алгоритма деревьев решений (Two-Class Boosted Decision Tree) и обучив модель по признаку наличия инсулинотерапии у пациента (рис. 6), используется двухклассовый алгоритм дерева принятия решений.

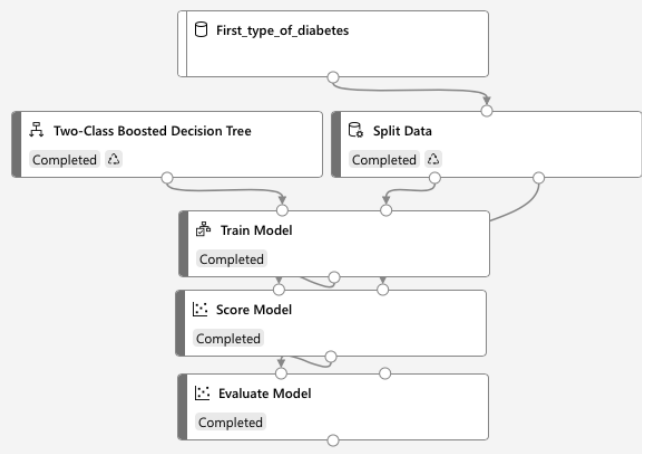

Рис. 6. Конвейер создания и обучения модели на основе деревьев решений

Оценка полученной модели представлена на рис. 7.

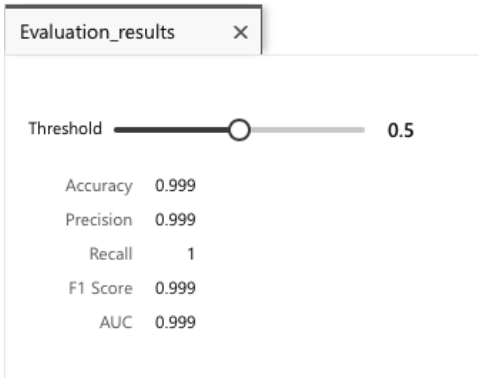

Рис. 7. Оценка модели на основе деревьев решений

Входными данными являлись все данные пациентов, страдающих первым типом сахарного диабета, на выходе – два класса: пациенты, которым была назначена инсулинотерапия и пациенты, которым инсулинотерапия не была назначена.

Показатели правильности, точности, оценка F1 и AUC модели равны 0,999, что является наилучшим результатом, так как результат максимально приближен к значению 1.

Алгоритм классификации на основе случайного леса.

Следующий алгоритм классификации для анализа эндокринологических данных – алгоритм случайного леса (Multiclass Decision Forest) (рис. 8). Модель обучена также по признаку назначения пациенту инсулинотерапии для сравнения метрик.

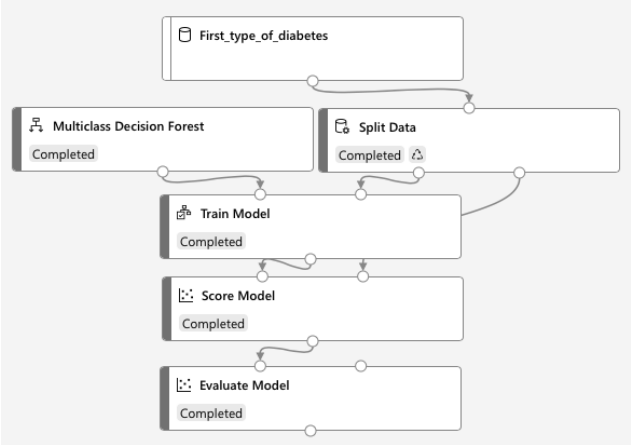

Рис. 8. Конвейер создания и обучения модели случайного леса

Оценка модели приведена на рис. 9.

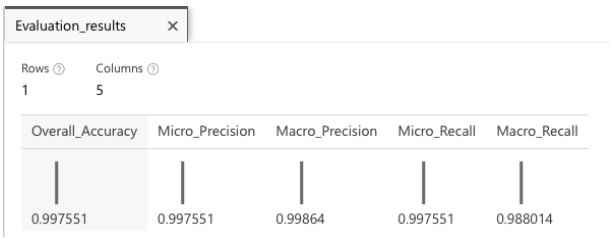

Рис. 9. Оценка модели на основе случайного леса

Показатели метрик достаточно качественны, что говорит о точности модели.

ОЦЕНКА РЕЗУЛЬТАТОВ

Сравнительные результаты трех исследованных моделей классификации для анализа эндокринологических данных по сахарному диабету I типа приведены в табл. 2.

Таблица 2

Сравнение метрик моделей классификации

Модель классификации	Правильность (Accuracy)	Точность (Precision)
Логистическая регрессия	0,659	0,672
Деревья решений	0,999	0,999
Случайный лес	0,997	0,998

Модель классификации на основе деревьев решений является наиболее точным алгоритмом для анализа рассматриваемых данных, показатели правильности, точности, оценки F1 и AUC модели равны 0,999, что является наилучшим результатом, максимально приближенным к значению. Модель классификации на основе случайного леса находится на втором месте с незначительной потерей качества. Алгоритм логистической регрессии среди используемых алгоритмов оказался наименее точным.

ЗАКЛЮЧЕНИЕ

В ходе проведенного исследования были проанализированы и интерпретированы данные эндокринологического мониторинга по Республике Башкортостан, в частности, набор данных пациентов, страдающих сахарным диабетом I типа. В результате анализа с помощью инструмента «Microsoft Azure: Machine Learning Studio» была выявлена наиболее точная модель для классификации рассматриваемого набора данных – модель на основе деревьев решений, которая и была отобрана для проведения дальнейших исследований данных эндокринологического мониторинга

СПИСОК ЛИТЕРАТУРЫ

1. **Бурсов А. И.** Применение искусственного интеллекта для анализа медицинских данных // Альманах клинической медицины. 2019. Т. 47, № 7. С. 630–633. **[** A. I. Bursov, “Application of artificial intelligence in medical data analysis”, (in Russian), in Al`manah klinicheskoj mediciny, vol 47, no. 7, pp. 630-633, 2019. **]**

2. **Анализ** методов машинного обучения для решения задач медицинского профиля. [Электронный ресурс]. URL: http://elibrary.asu.ru/xmlui/bitstream/handle/asu/6414/vkr.pdf?sequence=1&isAllowed=y (дата обращения 24.04.2022). **[** Analysis of machine learning methods for solving medical problems (2022, Apr. 24). [Online]. Available: http://elibrary.asu.ru/xmlui/bitstream/handle/asu/6414/vkr.pdf?sequence=1&isAllowed=y **]**

3. **Медицинские** информационные системы: обзор возможностей и примеры использования. [Электронный ресурс]. URL: https://evergreens.com.ua/ru/articles/medical-information-systems.html (дата обращения 24.04.2022). **[** Medical information systems: an overview of opportunities and examples of use (2022, Apr. 24). [Online]. Available: https://evergreens.com.ua/ru/articles/medical-information-systems.html **]**

4. **Информационная** система распознавания сахарного диабета. [Электронный ресурс]. URL: https://elib.pnzgu.ru/files/eb/doc/AB2wpOCX862A.pdf (дата обращения 24.04.2022). **[** Diabetes recognition information system (2022, Apr. 24). [Online]. Available: https://elib.pnzgu.ru/files/eb/doc/AB2wpOCX862A.pdf **]**

5. **Сахарный** диабет 1 и 2 тип. [Электронный ресурс]. URL: https://www.immun.ru/main/endokrinologiya/diabetes/ (дата обращения 24.04.2022). **[** Diabetes mellitus type 1 and 2 (2022, Apr. 24). [Online]. Available: https://www.immun.ru/main/endokrinologiya/diabetes/ **]**

6. **Типы** диабета. [Электронный ресурс]. URL: https://www.medtronic-diabetes.ru/cto-takoe-diabet/tipy-diabeta (дата обращения 24.04.2022). **[** Types of diabetes (2022, Apr. 24). [Online]. Available: https://www.medtronic-diabetes.ru/cto-takoe-diabet/tipy-diabeta **]**

7. Эпидемиологические характеристики сахарного диабета в Российской Федерации: клинико-статистический анализ по данным регистра сахарного диабета на 01.01.2021. [Электронный ресурс]. URL: https://www.dia-endojournals.ru/jour/article/view/12759/0 (дата обращения 24.04.2022). **[** Epidemiological characteristics of diabetes mellitus in the Russian Federation: clinical and statistical analysis according to the data of the diabetes registry as of 01.01.2021 (2022, Apr. 24). [Online]. Available: https://www.dia-endojournals.ru/jour/article/view/12759/0 **]**

8. **Аналитика** медицинских данных с использованием Microsoft Cloud для здравоохранения. [Электронный ресурс]. URL: https://docs.microsoft.com/ru-ru/azure/architecture/example-scenario/mch-health/medical-data-insights

(дата обращения 24.04.2022). **[** Health Analytics with Microsoft Cloud for Health (2022, Apr. 24). [Online]. Available: https://docs.microsoft.com/ru-ru/azure/architecture/example-scenario/mch-health/medical-data-insights **]**

9. **Стандарт** организации «Информационные системы в здравоохранении. Общие требования». [Электронный ресурс]. URL: http://www.miacso.ru/Documents/images/Site/Standart.pdf (дата обращения 24.04.2022). **[** Standard of the organization "Information systems in healthcare. General requirements" (2022, Apr. 24). [Online]. Available: http://www.miacso.ru/Documents/images/Site/Standart.pdf **]**

10. **Implement** the Azure healthcare blueprint for AI. [Electronic resource]. URL: https://docs.microsoft.com/ru-ru/azure/architecture/industries/healthcare/healthcare-ai-blueprint (accessed 24.04.2022).

11. **Документация** по Azure. [Электронный ресурс]. URL: https://learn.microsoft.com/ru-ru/azure/?product=compute (дата обращения 24.04.2022). **[** Documentation for Azure (2022, Apr. 24). [Online]. Available: https://learn.microsoft.com/ru-ru/azure/?product=compute **]**

ОБ АВТОРАХ

ШАХМАМЕТОВА Гюзель Радиковна, проф. каф. вычислительной математики и кибернетики. Дипл. инж. по инф. системам (УАИ, 1992). Д-р техн. наук по сист. анализу, управлению и обработке информации (УГАТУ, 2013). Иссл. в обл. интеллект. поддержки принятия решений, распознавания образов, обр. биомед. данных методами машинного обучения и иск. интеллекта.

ХРИСТОДУЛО Анна Дмитриевна, магистрант направления 09.04.01 Информатика и вычислительная техника (профиль «Компьютерный анализ и интерпретация данных»), магистрант Университета г. Тренто, Италия. Иссл. в обл. обр. эндокринолог. данных методами машинного обучения.

БЕРЕГОВАЯ Софья Павловна, врач-эндокринолог МОГБУЗ «Городская поликлиника». Дипл. о высш. мед. обр. по спец. врач-лечебник (БГМУ, 2019), дипл. об оконч. ординатуры по спец. врач-эндокринолог (БГМУ, 2021). Иссл. в обл. осложн. сах. диабета I и II типов.

METADATA

Title: Analysis of endocrinological data based on classification models.

Authors: G. R. Shakhmametova [1], A. D. Khristodulo [2], S. P. Beregovaya [3]

Affiliation:

[1, 2] Ufa State Aviation Technical University (UGATU), Russia.

[2] University of Trento, Italy.

[3] City Polyclinic of Magadan, Russia.

Email: [1] shakhgouzel@mail.ru, [2] annhrist@mail.ru, [3] sofyaklimets@gmail.com

Language: Russian.

Source: SIIT (scientific journal of Ufa State Aviation Technical University), vol. 4, no. 2 (9), pp. 30-36, 2022. ISSN 2686-7044 (Online), ISSN 2658-5014 (Print).

Abstract: The results of the study of endocrinological monitoring data in the Republic of Bashkortostan using various classification models are presented: models based on logistic regression, decision trees and random forest. As initial data, endocrinological monitoring data on the incidence of type 1 diabetes mellitus in the Republic of Bashkortostan were used. The tool used is the cloud service "Microsoft Azure: Machine Learning Studio." The aim of the study is an informed, experimentally validated selection of a classification model for the analysis and interpretation of endocrinological monitoring data to determine patterns in the incidence of type I diabetes mellitus. The study classifies a dataset of more than 150 thousand records, three classification algorithms (logistic regression, decision trees and random forest), evaluates each model with different metrics, interprets the results obtained and selects the most suitable classification model for the specified dataset for further studies of endocrinological monitoring data.

Key words: data analysis; machine learning; endocrinological data; classification; decision trees; logistic regression; random forest.

About authors:

SHAKHMAMETOVA, Gyuzel Radikovna, Prof., Dept. of Computational Mathematics and Cybernetics. Dipl. Eng. in Information Systems (UAI, 1992). Dr. of Tech. Sci. in System Analysis, Management and Information Processing (USATU, 2013). Research in the field of intelligent decision support, pattern recognition, biomedical data processing by machine learning and artificial intelligence.

KHRISTODULO, Anna Dmitrievna, master's student in the direction of 09.04.01 Informatics and Computer Engineering (profile "Computer Analysis and Interpretation of Data"), master's student at the University of Trento, Italy. Research in the field of processing endocrinological data by machine learning methods.

BEREGOVAYA, Sofia Pavlovna, endocrinologist of "City Polyclinic." Diploma of Higher Medical Education in the specialty of medical doctor (BSMU, 2019), dipl. of residency in the specialty of endocrinologist (BSMU, 2021). Research in the field of Type I and Type II diabetes complication.