

ჯანდაცვის სფეროს მონაცემთა მართვა - ტექნიკური, ეთიკური და იურიდიული ასპექტები NoSQL ბაზების მაგალითზე

ზებურ სურმანიძე, ბესიკ ბერიძე

ბათუმის შოთა რუსთაველის სახელმწიფო უნივერსიტეტი

zebur.surmanidze@bsu.edu.ge, b.beridze@bsu.edu.ge

რეზიუმე

ნაშრომში ხაზგასმულია მონაცემთა ბაზების მართვის გადამწყვეტი როლი ჯანდაცვის სფეროში მანქანური სწავლის ალგორითმების ინტეგრაციის პროცესში, განსაკუთრებით სიმპტომებზე დაფუძნებული სავარაუდო დიაგნოზის სისტემის მაგალითზე [1]. როგორც ჯანდაცვის ციფრული ტრანსფორმაციის საფუძველი, მონაცემთა ბაზები უზრუნველყოფს მონაცემთა მოქნილ შეგროვებას, საიმედო შენახვას, ეფექტურ დამუშავებასა და უსაფრთხო ანონიმიზაციას, რაც საშუალებას აძლევს მანქანური სწავლის მოდელებს გენერირებული იქნას ზუსტი და პერსონალიზებული სამედიცინო რეკომენდაციები. MongoDB-ის მაგალითზე ნაჩვენებია, თუ როგორ უწყობს ხელს NoSQL ბაზის მოქნილობა კლინიკური მონაცემების რთული სტრუქტურის მართვას, ხოლო უსაფრთხოების მექანიზმები, როგორიცაა დამიფვრა და წვდომის კონტროლი, იცავს პაციენტთა კონფიდენციალურობას.

ნაშრომში წარმოდგენილია მონაცემთა მოდელის დეტალური აღწერა, უსაფრთხოების მექანიზმების იმპლემენტაცია და პრაქტიკული მაგალითები, რომლებიც ნათლად აჩვენებს, თუ როგორ უწყობს ხელს მონაცემთა ბაზის ეფექტური მართვა ჯანდაცვის სექტორის ციფრულ ტრანსფორმაციას. კვლევა ხაზს უსვამს, რომ მონაცემთა ბაზების ეფექტური მართვა არა მხოლოდ ტექნიკური საჭიროებაა, არამედ ეთიკური და იურიდიული ვალდებულებაც ჯანდაცვის სფეროში სენსიტიური მონაცემების დაცვის კუთხით [8].

საკვანძო სიტყვები: მონაცემთა ბაზები, ჯანდაცვა, მანქანური დასწავლა, ანონიმიზაცია, MongoDB, უსაფრთხოება.

1. შესავალი.

თანამედროვე ჯანდაცვის სისტემები სულ უფრო მეტად ეყრდნობა მანქანურ სწავლას, რომელიც საშუალებას იძლევა გაანალიზდეს დიდი მოცულობის მონაცემები და გენერირდეს პროგნოზები, რომლებიც დაეხმარება ექიმებს და პაციენტებს უფრო ინფორმირებული გადაწყვეტილებების მიღებაში. მაგალითად, სიმპტომებზე დაფუძნებული სისტემები შეუძლიათ შეაფასონ პოტენციური რისკები და შესთავაზონ შესაბამისი მოქმედებები, რაც განსაკუთრებით მნიშვნელოვანია რეგიონებში, სადაც სამედიცინო რესურსები შეზღუდულია. თუმცა, ამგვარი სისტემების ეფექტურობა პირდაპირ დამოკიდებულია მონაცემთა ხარისხზე და მათ მართვაზე, რაც მონაცემთა ბაზების სპეციალისტების ძირითადი პასუხისმგებლობაა. ამ სტატიაში, როგორც მონაცემთა ბაზების სპეციალისტი, ყურადღება გამახვილებულია იმაზე, თუ როგორ უზრუნველყოფს ბაზა მონაცემთა საიმედო ნაკადს მანქანური სწავლის პროცესებისთვის [1,2].

ქვეყანაში ჯანდაცვის სისტემაში მონაცემთა მართვა ხშირად რჩება ფრაგმენტულ დონეზე, რაც ართულებს მანქანური სწავლის ტექნოლოგიების სრულფასოვან დანერგვას და იწვევს ისეთ პრობლემებს, როგორიცაა მონაცემთა დუბლირება, არასრული ინფორმაცია ან უსაფრთხოების დარღვევები. კვლევის მიზანია წარმოდგენილი იქნეს მონაცემთა ბაზის

არქიტექტურა, რომელიც საშუალებას იძლევა ეფექტურად განხორციელდეს მანქანური სწავლის ალგორითმების ინტეგრაცია სიმპტომებზე დაფუძნებულ სისტემაში. კონკრეტულად, განხილული იქნება ისეთი ალგორითმები, როგორიცაა გადაწყვეტილების ხე, შემთხვევითი ტყე, საყრდენი ვექტორის მანქანა და ხელოვნური ნეირონული ქსელები, რომლებიც იყენებენ ბაზაში შენახულ მონაცემებს პროგნოზების გენერირებისთვის [7]. ამოცანები მოიცავს მონაცემთა მოდელის შექმნას, უსაფრთხოების მექანიზმების იმპლემენტაციას და პრაქტიკულ მაგალითებს, რომლებიც აჩვენებენ ბაზის როლს სისტემის მთლიან ეფექტურობაში. ეს მიდგომა არა მხოლოდ ტექნიკურად ამართლებს თავს, არამედ ხელს უწყობს ჯანდაცვის სერვისების ხელმისაწვდომობის გაუმჯობესებას საქართველოს კონტექსტში, სადაც რეგიონული უთანაბრობა და მონაცემთა ციფრული ლიტერატურის დაბალი დონე კვლავ გამოწვევად რჩება [9].

2. ძირითადი ნაწილი

ჯანდაცვის სისტემაში არსებული მონაცემების მართვაში NoSQL ტიპის ბაზები, როგორცაა MongoDB, სულ უფრო პოპულარული ხდება მათი მოქნილობისა და სკალირებადობის გამო, რაც საშუალებას იძლევა დაამუშავდეს არასტრუქტურირებული მონაცემები, როგორცაა პაციენტთა სიმპტომური აღწერები ან კლინიკური ჩანაწერები [10]. კვლევები აჩვენებენ, რომ ასეთი ბაზები ეფექტურად უმკლავდებიან დიდ მოცულობის მონაცემებს, რაც კრიტიკულია მანქანური სწავლის მოდელების გაწვრთნისთვის. შესაბამისი ალგორითმების ეფექტურობა, სიზუსტე და სანდოობა პირდაპირ დამოკიდებულია მონაცემთა ხარისხზე, ხელმისაწვდომობაზე და მართვის სისტემის საიმედოობაზე, რასაც მონაცემთა ბაზა უზრუნველყოფს. მაგალითად, მონაცემთა ანონიმიზაციის ტექნიკები, როგორცაა k-anonymity ან differential privacy, აუცილებელია იმისთვის, რომ დაცული იყოს პაციენტთა კონფიდენციალურობა, რაც შეესაბამება GDPR-ის მოთხოვნებს და ხელს უწყობს ეთიკურ მიდგომას ჯანმრთელობის მონაცემების გამოყენებაში.

მანქანური სწავლის კონტექსტში, მონაცემთა წინასწარი დამუშავება ხდება ბაზის დონეზე, რაც უზრუნველყოფს მოდელების სიზუსტეს და ამცირებს გამოთვლით რესურსებს [2]. საქართველოში მსგავსი სისტემები ჯერ კიდევ ნაკლებად განვითარებულია, რაც ხაზს უსვამს ამ კვლევის აქტუალობას, განსაკუთრებით იმ ფონზე, როდესაც ქვეყანაში მიმდინარეობს ციფრული ტრანსფორმაციის პროცესები, მაგრამ მონაცემთა მართვის სტანდარტები კვლავ საჭიროებენ გაუმჯობესებას. უნდა აღვნიშნოთ, რომ მონაცემთა ბაზების ინტეგრაცია მანქანურ სწავლასთან არა მხოლოდ აუმჯობესებს პროგნოზების ხარისხს, არამედ ხელს უწყობს რესურსების ოპტიმიზაციას, რაც განსაკუთრებით მნიშვნელოვანია განვითარებადი ქვეყნებისთვის.

მეთოდოლოგია: ჩვენს მიერ შემუშავებული სისტემა იყენებს კლინიკურ მონაცემებს, რომლებიც "clinical_data.csv" ფაილიდან და იმპორტირდება MongoDB-ში, რაც უზრუნველყოფს მონაცემთა სწრაფ წვდომას და მოქნილ შენახვას. მონაცემთა სქემა შექმნილია ისე, რომ იგი მოერგოს ცვალებად კლინიკურ მოთხოვნებს, მაგალითად სიმპტომების მასივი, რომელიც შეიცავს დინამიურ ელემენტებს, როგორცაა სიმპტომების სახელწოდებები, და რომლებიც შეიძლება განახლდეს რეალურ დროში. შეგროვება ხდება ანონიმურად API-ის მეშვეობით, რომელიც შექმნილია Python პლათფორმაზე Flask-ის გამოყენებით, სადაც პაციენტი შეიყვანს სიმპტომებს და სისტემა ავტომატურად აგენერირებს ანონიმიზებულ იდენტიფიკატორს hashing ტექნიკის მეშვეობით, როგორცაა SHA-256, რაც უზრუნველყოფს მონაცემთა კონფიდენციალურობას და შეესაბამება საერთაშორისო სტანდარტებს [4,8].

ბაზაში მონაცემები გადიან წინასწარ დამუშავებას, რაც მოიცავს გამოტოვებული მნიშვნელობების მართვას გაფილტვრის მეთოდებით ან ამოშლით, რათა უზრუნველყოფილი იყოს მონაცემთა სრულყოფილება მანქანური სწავლის მოდელების გაწვრთნისთვის. სიმპტომები გარდაიქმნება ბინარულ ვექტორებად One-Hot Encoding-ის მეშვეობით, რაც ხდება Python-ის Pandas ბიბლიოთეკის გამოყენებით, და შემდეგ ინახება ბაზაში, რათა მოდელებმა სწრაფად წვდომა მიიღონ მზა ფორმატზე. რაოდენობრივი მახასიათებლები, როგორიცაა ტემპერატურა ან გულისცემა, სტანდარტიზებულია StandardScaler-ის მეშვეობით, რაც უზრუნველყოფს მოდელების სტაბილურობას. MongoDB-ის Aggregation Pipeline გამოიყენება მონაცემთა აგრეგაციისთვის, რაც საშუალებას იძლევა კომპლექსური მოთხოვნები შესრულდეს ეფექტურად, მაგალითად, სიმპტომების ფილტრაცია და დიაგნოზების დათვლა [6,11].

MongoDB-ში იმპლემენტირებულია უსაფრთხოების მექანიზმები, რომლებიც მოიცავს ველების დონეზე დაშიფვრას სენსიტიური ინფორმაციისთვის, როგორიცაა სიმპტომები, რათა თავიდან იქნას აცილებული არავტორიზებული წვდომა. წვდომის კონტროლი ეფუძნება როლებზე დაფუძნებულ სისტემას, სადაც ადმინისტრატორებს აქვთ წვდომა ნედლ მონაცემებზე, ხოლო სხვა მომხმარებლები მხოლოდ ანონიმიზებულ ვერსიას ხედავენ. აუდიტი უზრუნველყოფს ყველა წვდომის ლოგირებას, რაც საშუალებას იძლევა მონიტორინგი და რეაგირება პოტენციურ საფრთხეებზე. სარეზერვო კოპიები შენახულია შიფრირებულ ფორმატში, რაც ხელს უწყობს მონაცემთა აღდგენას კატასტროფების შემთხვევაში. მანქანური სწავლის მოდელების გაწვრთნა ხდება მხოლოდ ანონიმიზებულ მონაცემებზე, რაც მნიშვნელოვნად ამცირებს რისკებს და უზრუნველყოფს ეთიკურ სტანდარტებს [8].

მოდელები, როგორიცაა გადაწყვეტილების ხე ან შემთხვევითი ტყე, გაწვრთნილია Scikit-learn ბიბლიოთეკის გამოყენებით და ინახება pickle ფორმატში, მაგრამ მათი მუშაობა დამოკიდებულია ბაზიდან ამოღებულ მონაცემებზე. მოთხოვნები, როგორიცაა სიმპტომების ფილტრაცია, შესრულდება ბაზის დონეზე, რაც უზრუნველყოფს სწრაფ პროგნოზებს და ინტეგრაციას API-სთან.

შედეგი. დასწავლა ჩატარდა ანონიმიზებულ ჩანაწერებზე, „ანონიმიზებული“ მონაცემები გულისხმობს ისეთ მონაცემებს, რომლებიც დამუშავებულია ისე, რომ მათი საშუალებით შეუძლებელია კონკრეტული პიროვნების იდენტიფიცირება. ამ შემთხვევაში, სტატიაში აღწერილი სისტემა იყენებს ანონიმიზაციის მეთოდებს, როგორიცაა SHA-256 hashing, რათა პაციენტის პერსონალური იდენტიფიკატორები (მაგ., სახელი, პირადი ნომერი) გარდაიქმნას უნიკალურ, მაგრამ არა-პერსონალიზებულ კოდებად. ეს პროცესი უზრუნველყოფს, რომ მონაცემები (როგორიცაა სიმპტომები ან დიაგნოზები) გამოყენებულ იქნას მანქანური სწავლის მოდელების გაწვრთნისთვის, მაგრამ ისე, რომ პაციენტის კონფიდენციალურობა არ დაირღვეს. ანონიმიზაციის მიზანია შეესაბამებოდეს მონაცემთა დაცვის საერთაშორისო სტანდარტებს, მაგალითად, GDPR-ს, რაც განსაკუთრებით მნიშვნელოვანია ჯანდაცვის სფეროში, სადაც მონაცემები სენსიტიურია. სტატიის კონტექსტში, ანონიმიზაციამ „სრული კონფიდენციალურობის“ უზრუნველყოფა ნიშნავს, რომ ტესტირებისას არც ერთი ჩანაწერი არ იქნეს დაკავშირებული კონკრეტულ პიროვნებასთან, რაც ასევე გამორიცხავს მონაცემთა გაჟონვის რისკს.

ტესტირება ჩატარდა 500 ანონიმიზებულ ჩანაწერზე, მაგრამ MongoDB-ის მოთხოვნის შესრულების სიჩქარის (<50ms) შეფასება განხორციელდა სიმულირებულ მონაცემთა ბაზაზე, რომელიც შეიცავდა მილიონობით ჩანაწერს. მანქანური დასწავლის მოდელების სიზუსტე მერყეობდა 85-დან 92 პროცენტამდე, სადაც შემთხვევითი ტყის ალგორითმი ყველაზე ეფექტური აღმოჩნდა.

ცხრილი 1: მოდელების შედარება ბაზის მონაცემებზე

ალგორითმი	სიზუსტე (%)	მოთხოვნის დრო (მს)	მონაცემთა მოცულობა (მბ)
გადაწყვეტილების ხე (Decision Tree)	85	30	10
შემთხვევითი ტყე (Random Forest)	92	45	15
საყრდენი ვექტორის მანქანა (Support Vector Machine - SVM)	88	40	12
ხელოვნური ნეირონული ქსელები (Artificial Neural Networks - ANN)	90	50	14

შედეგების მიხედვით შეიძლება ვიმსჯელოთ, რომ მონაცემთა ბაზები ჯანდაცვაში არ წარმოადგენს მხოლოდ შენახვის ინსტრუმენტს, არამედ ხდება მანქანური სწავლის საფუძველი, რადგან უზრუნველყოფს მონაცემთა საიმედო ნაკადს და უსაფრთხოებას [3]. MongoDB-ის მოქნილობამ საშუალება მოგვცა სწრაფად ინტეგრირებულიყო სისტემაში, თუმცა გამოწვევები მოიცავს მონაცემთა მოცულობის ზრდას და კიბერუსაფრთხოების საკითხებს, რაც საჭიროებს მუდმივ მონიტორინგს. შედარებით ტრადიციულ რელაციურ ბაზებთან, NoSQL უკეთესად უმკლავდება არასტრუქტურირებულ მონაცემებს, რაც განსაკუთრებით მნიშვნელოვანია ჯანდაცვის სფეროში არსებული მონაცემთა ბაზებისათვის, რადგან ეს სფერო ხასიათდება მონაცემთა მრავალფეროვანი და რთული ბუნებით, სადაც მონაცემები სენსიტიურია (მაგ., პაციენტთა სამედიცინო ისტორია, სიმპტომები, დიაგნოზები). მომავალში რეკომენდებულია blockchain ტექნოლოგიის ინტეგრაცია, რათა კიდევ უფრო გაიზარდოს უსაფრთხოება, მონაცემთა დაცვა არავტორიზებული წვდომისგან, გაუზიარდოს ან მანიპულა ასევე მნიშვნელოვანია მონაცემთა წვდომის, მართვისა და გამოყენების პროცესების „გამჭვირვალობა“ და „თვალსაჩინოება“, ისე, რომ ყველა ჩართული მხარისთვის (პაციენტები, ექიმები, ადმინისტრატორები) ნათელი იყოს, თუ როგორ, სად და რატომ გამოიყენება მონაცემები.

3. დასკვნა.

მონაცემთა ბაზები კრიტიკული ელემენტია ჯანდაცვის მანქანური სწავლის სისტემებში, როგორც ეს აჩვენა სიმპტომებზე დაფუძნებულმა პროტოტიპმა, რომელიც უზრუნველყოფს მონაცემთა ეფექტურ მართვას და უსაფრთხოებას [5,10]. კვლევა ხაზს უსვამს მონაცემთა ბაზების სპეციალისტების როლს ჯანდაცვის ციფრულ სტანდარტებზე გადასვლის შემთხვევაში და მიუთითებს სამომავლო კვლევების აუცილებლობაზე, როგორცაა ჰიბრიდული მოდელების განვითარება და ხელოვნური ინტელექტის უფრო ღრმა ინტეგრაცია.

გამოყენებული ლიტერატურა:

1. J. D. Kelleher et al., Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, 2nd ed. MIT Press, 2020.
2. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. O'Reilly Media, 2019.
3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.

4. H. M. Abbas, "Ethical and Legal Considerations in Artificial Intelligence in Healthcare: A Review," *Health Information Science and Systems*, vol. 9, no. 1, pp. 1–7, 2021.
5. World Health Organization (WHO), "Digital Health Strategy 2020–2025", [Online]. Available: <https://www.who.int/publications/i/item/9789240020924>
6. D. Ghosh, S. Dey, and S. Ghosh, "Artificial Neural Networks in Healthcare: A Review," *Advances in Intelligent Systems and Computing*, vol. 1216, pp. 23–37, Springer, 2021.
7. S. M. Raza and A. S. Haque, "Comparative Analysis of Decision Tree, Random Forest, SVM and ANN in Healthcare Diagnosis," *International Journal of Computer Applications*, vol. 182, no. 19, pp. 7–12, Mar. 2019.
8. General Data Protection Regulation (GDPR), European Union, 2016. [Online]. Available: <https://gdpr.eu>
9. Ministry of Internally Displaced Persons from the Occupied Territories, Labour, Health and Social Affairs of Georgia, "Health System Performance Assessment", Tbilisi, 2021.
10. N. Elmqvist, "Data Fusion and Feature Engineering in Healthcare Machine Learning," *IEEE Computer Graphics and Applications*, vol. 40, no. 6, pp. 92–97, 2020.
11. L. Deng and D. Yu, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

Healthcare Data Management - Technical, Ethical, and Legal Aspects Using NoSQL Databases as an Example

Zebur Surmanidze, Besik Beridze

Batumi Shota Rustaveli State University

zebur.surmanidze@bsu.edu.ge b.beridze@bsu.edu.ge

Abstract

The paper highlights the crucial role of databases in the process of integrating machine learning algorithms in the healthcare sector, especially on the example of a symptom-based presumptive diagnosis system. As the foundation of the digital transformation of healthcare, databases provide flexible data collection, reliable storage, efficient processing, and secure anonymization, which allows machine learning models to generate accurate and personalized medical recommendations. The example of MongoDB shows how the flexibility of a NoSQL database helps manage complex clinical data structures, while security mechanisms such as encryption and access control protect patient confidentiality.

The paper presents a detailed description of the data model, implementation of security mechanisms, and practical examples that clearly demonstrate how effective database management contributes to the digital transformation of the healthcare sector. The study emphasizes that effective database management is not only a technical necessity, but also an ethical and legal obligation in terms of protecting sensitive data in the healthcare sector.

Keywords: databases, healthcare, machine learning, anonymization, MongoDB, security.