JADVALLI MA’LUMOTLARDAGI BO‘SH VA ANOMAL QIYMATLARNI ANIQLASH HAMDA QAYTA TIKLASH ASOSIDA DASTLABKI MA’LUMOTLARNI QAYTA ISHLASH ALGORITMI

Nishanov Axram Xasanovich[1], Mengturayev Farxod Ziyatovich[2]

[1]Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti tizimli va amaliy dasturlash kafedrasi professori
[2]Denov tadbirkorlik va pedagogika instituti katta o‘qituvchisi
E-mal: nishanov_akram@mail.ru, f.mengtoraev@dtpi.uz
ORCID: 0000-0003-0562-8377

K E Y W O R D S

Jadvalli ma’lumotlar, dastlabki ishlov berish, bo‘sh qiymat, anomal qiymat, ma’lumotlarni qayta tiklash, mashinali o‘qitish, klassifikatsiya, mediana, moda, qandli diabet, Accuracy, F1-score

A B S T R A C T

Ushbu maqolada mashinali o‘qitish masalalarida foydalaniladigan jadvalli ma’lumotlarning sifatini oshirishga qaratilgan adaptiv dastlabki ishlov berish algoritmi tadqiq etilgan. Jadvalli ma’lumotlar tarkibida uchraydigan bo‘sh qiymatlar, noto‘g‘ri kiritilgan hamda predmet sohasi uchun mantiqan yaroqsiz bo‘lgan anomal qiymatlar klassifikatsiya modellarining samaradorligiga salbiy ta’sir ko‘rsatishi mumkin. Mazkur muammoni bartaraf etish maqsadida belgining turi, ruxsat etilgan qiymatlar sohasi va obyektning sinf tegini hisobga oluvchi algoritmik yondashuv taklif etilgan. Algoritmda sonli va tartibli belgilar uchun mediana, nominal va binar belgilar uchun esa moda qiymatidan foydalaniladi. Bunda statistik qiymatlar butun dataset asosida emas, faqat o‘qitish to‘plamidagi to‘g‘ri qiymatlar asosida shakllantiriladi. Bu esa test ma’lumotlarining o‘qitish jarayoniga bilvosita ta’sir ko‘rsatishi, ya’ni ma’lumotlarning sizib chiqishi xavfini kamaytiradi.

I. KIRISH

Bugungi kunda raqamli texnologiyalarning jadal rivojlanishi natijasida turli sohalarda katta hajmdagi ma’lumotlarni yig‘ish, saqlash va avtomatik qayta ishlash imkoniyatlari kengayib bormoqda. Ayniqsa, sog‘liqni saqlash, moliya, ishlab chiqarish, qishloq xo‘jaligi va boshqa amaliy sohalarda qaror qabul qilish jarayonlarini avtomatlashtirishda mashinali o‘qitish algoritmlaridan keng foydalanilmoqda. Bunday algoritmlarning yakuniy natijasi tanlangan model bilan bir qatorda modelga uzatilayotgan ma’lumotlarning sifatiga ham bevosita bog‘liq[1-5].

Jadvalli ma’lumotlar mashinali o‘qitishning eng ko‘p uchraydigan ma’lumot shakllaridan biridir. Amaliy ma’lumotlarni shakllantirish jarayonida ayrim yacheykalarning bo‘sh qolishi, noto‘g‘ri qiymat kiritilishi yoki mavjud qiymatning predmet sohasi uchun mantiqan yaroqsiz bo‘lishi mumkin [6-7]. Shu sababli modelni qurishdan oldin ma’lumotlarni dastlabki qayta ishlash muhim bosqich hisoblanadi.

An’anaviy usullarda bo‘sh qiymatlarni o‘rtacha arifmetik qiymat yoki mediana bilan to‘ldirish, kategorik qiymatlar uchun esa moda qiymatidan foydalanish keng tarqalgan [8-11]. Biroq bunday yondashuvlar bo‘sh bo‘lmagan, lekin predmet sohasi nuqtai nazaridan noto‘g‘ri qiymatlarni har doim ham aniqlay olmaydi. Masalan, fizik o‘lchovni ifodalovchi belgida manfiy qiymat statistik jihatdan mavjud bo‘lishi mumkin, biroq predmet sohasi bo‘yicha anomal hisoblanadi [12-16].

Shu nuqtai nazardan, dastlabki qayta ishlash jarayonida nafaqat statistik xususiyatlar, balki har bir belgining predmet sohasi bo‘yicha ruxsat etilgan qiymatlar sohasi ham hisobga olinishi lozim.

Ushbu tadqiqotning maqsadi jadvalli ma’lumotlarda mavjud bo‘sh va anomal

(cc) BY *dtai.uz*

qiymatlarni aniqlash, ularni belgi turi va obyekt sinfini hisobga olgan holda qayta tiklash hamda taklif etilgan yondashuvning klassifikatsiya natijalariga ta’sirini eksperimental baholashdan iborat.

II. ASOSIY QISM

Mazkur tadqiqot ishida uch xil dataset ustida amaliyot olib borilgan. Ular:

- 1-tur qandli diabet kasalligi bo‘yicha,
- 2-tur qandli diabet kasalligi bo‘yicha,
- Dry Bean ochiq dataseti.

Birinchi va ikkinchi datasetlar bular muallif tomonidan shakllantirilgan. Qandli diabetning 1-turi 140 ta obyektdan kasallik varaqalari asosida eng asosiy bo‘lgan 10 ta belgi bo‘lib bu belgilarning qabul qilishi mumkin bo‘lgan qiymatlar ajratib olindi. Qandli diabetning 2-turi 340 ta obyektdan kasallik varaqalari asosida eng asosiy bo‘lgan 19 ta belgi bo‘lib bu belgilarning qabul qilishi mumkin bo‘lgan qiymatlar ajratib olindi.

Uchinchi dataset sifatida obyektlar soni ko‘pligi bilan ajralib turuvchi Dry Bean ochiq datasetidan foydalanildi. Mazkur dataset asosida taklif etilgan algoritmning samaradorligi baholandi. Buning uchun ma’lumotlarga ikki turdagi sun’iy buzilish bo‘sh qiymatlar va anomal qiymatlar kiritilib, ularni aniqlash va qayta tiklash amalga oshirildi. Qayta ishlangan ma’lumotlarning sifati bir nechta klassifikatsiya algoritmlari yordamida baholandi.

Ushbu ishda taklif etilayotgan algoritm shundan iborat bo‘lib, jadvalli ma’lumotlarga dastlabki ishlov berish jarayoni faqat bo‘sh qiymatlarni to‘ldirish bilan cheklanib qolmasdan balki har bir belgining predmet sohasiga xos mantiqiy cheklov shartlari asosida anomal qiymatlarni aniqlaydi va tuzatadi.

Odatda dastlabki ishlov berish jarayonida bo‘sh qiymatlar o‘rta arifmetik va mediana orqali to‘ldiriladi. Biroq bunday usullarda bo‘sh bo‘lmagan, lekin mantiqan noto‘g‘ri bo‘lgan qiymatlar har doim ham aniqlanmaydi. Masalan, Yuza = -10, Perimetr = 0 kabi qiymatlar bo‘sh emas, lekin real predmet sohasi nuqtai nazaridan qabul qilib bo‘lmaydi. Taklif etilgan dastlabki ishlov berish algoritmi aynan shu turdagi qiymatlarni ham aniqlash va tuzatishni nazarda tutadi.

Taklif etilayotgan ilmiy yangilik quyidagi jihatlar bilan asoslanadi:

1. **Bo‘sh va anomal qiymatlarni yagona algoritmik jarayonda qayta ishlash.** Algoritm bo‘sh qiymatlar bilan bir qatorda belgi cheklovlariga zid qiymatlarni ham tuzatilishi kerak bo‘lgan yacheykalar sifatida belgilaydi.
2. **Belgi mazmuniga asoslangan cheklovlarni hisobga oluvchi yondashuv.** Geometrik belgilar uchun $x > 0$, nisbiy belgilar uchun $0 < x \leq 1$, nominal belgilar uchun $x_i^j \in A_j$ shartlari kiritiladi. Bu dastlabki ishlov berish jarayonini oddiy statistik to‘ldirishdan ma’lumot sohasida xos cheklovlar asosidagi intellektual dastlabki ishlov berish bosqichiga aylantiradi.
3. **O‘qitishga asoslangan statistik to‘ldirish.** To‘ldirish uchun ishlatiladigan mediana qiymatlari butun dataset bo‘yicha emas, balki faqat o‘qitish qismi asosida hisoblanadi. Bu test ma’lumotlari dastlabki ishlov berish statistikasiga ta’sir qilishining oldini oladi.
4. **Bir nechta klassifikatsiya algoritmlari orqali tasdiqlash.** Algoritm samaradorligi Logistic Regression, KNN, SVM-RBF, Decision Tree, Random Forest, Extra Trees, Gradient Boosting va Gaussian Naive Bayes modellari orqali baholandi. Barcha 8 ta modelda to‘g‘rilangan holat buzilgan holatga nisbatan yaxshiroq natija berdi.

Masalaning qo‘yilishi

Quyidagicha ma’lumotlar to‘plami berilgan bo‘lsin:

$$D = \{X, y\},$$

bu yerda $X \in \mathbb{R}^{m \times N}$ - belgilar matritsasi, m - obyektlar soni, N - belgilar soni, y - sinf belgilari vektori.

Berilgan D to‘plamdagi bo‘sh va anomal qiymatlarni aniqlash va ularning qiymatlarini qayta tiklashga asoslangan dastlabki ishlov berish algoritmi orqali yangi $D^{rep} = \{X^{rep}, y^{rep}\}$ to‘plami shakllantirilsin.

(cc) BY

Dastlabki ishlov berish algoritmini shakllantirish va testdan o‘tkazish bosqichlari

1-bosqich. Datasetni tayyorlash. Dataset Excel fayldan o‘qiladi. Qaytalangan qatorlar olib tashlanadi, maqsad ustuni sifatida SINF belgisi ajratiladi.

$$D = \{X, y\}$$

bu yerda $X \in \mathbb{R}^{m \times N}$ - belgilar matritsasi, m - obyektlar soni, N - belgilar soni, y - sinf belgilari vektori.

2-bosqich. Belgilar uchun ruxsat etilgan qiymatlar sohasini belgilash. Ma’lumotlar to‘plamidagi belgilar ularning mazmuni va qabul qilishi mumkin bo‘lgan qiymatlariga ko‘ra geometrik o‘lchov, shakl koefitsienti va nominal belgilar guruhlariga ajratiladi:

$$F = F_g \cup F_r \cup F_n \qquad (1)$$

bu yerda F_g – musbat sonli qiymatlarni qabul qiluvchi geometrik o‘lchov belgilar, F_r – $(0,1]$ oralig‘idagi qiymatlarni qabul qiluvchi shakl koefisienti belgilari, F_n - belgi uchun ruxsat etilgan nominal belgilar to‘plami.

3-bosqich. Har bir belgi uchun ruxsat etilgan qiymatlar sohasini belgilash. j - belgi uchun ruxsat etilgan qiymatlar sohasi quyidagicha belgilanadi:

$$R_j = \begin{cases} (0, +t), & j \in F_g \\ \{0,1], & j \in F_r \\ A_j, & j \in F_n \end{cases}$$

bu yerda R_j – j-belgi uchun ruxsat etilgan qiymatlar to‘plami, A_j – aynan j-nominal belgi uchun ruxsat etilgan kategoriyalar to‘plami, t – belgining qabul qilishi mumkin bo‘lgan maksimum musbat qiymati.

4-bosqich. Butun datasetga sun’iy buzilish kiritish. Har bir geometrik belgi bo‘yicha ma’lum ulushdagi qiymatlar bo‘sh ya’ni NaN qilinadi va yana bir qismi anomal qiymatlar bilan almashtiriladi. Shunday ko‘rinishda buzilgan dataset shakllantiriladi:

$$D^{corr} = \{X^{corr}, y^{corr}\}$$

5-bosqich. Bo‘sh va anomal qiymatlarni aniqlash. Buzilgan datasetdagi bo‘sh yoki anomal qiymatlarni aniqlash uchun quyidagi indikator funksiya kiritiladi:

$$M_i^j = \begin{cases} 1, x_i^j = NaN \; yoki \; C_j(x_i^j) = 0 \\ 0, aks \; holda \end{cases}$$

$M_i^j = 1$ bo‘lgan qiymatlar bo‘sh yoki anomal hisoblanadi. $C_j(x_i^j)$ quyidagicha hisoblanadi:

$$C_j(x_i^j) = \begin{cases} 1, \; x_i^j \neq NaN \; \wedge x_i^j \in R_j \\ 0, \; x_i^j = NaN \; \vee x_i^j \notin R_j \end{cases}$$

Bunda $C_j(x_i^j) = 1$ qiymat belgining to‘g‘ri va ruxsat etilgan qiymatga ega ekanini, $C_j(x_i^j) = 0$ esa qiymat bo‘sh yoki belgilangan soha cheklovlariga mos emasligini anglatadi.

6-bosqich. Train/test indekslarini bir xil belgilash. *Original*, *Buzilgan* va *To‘g‘rilangan* datasetlar bir xil train/test indekslari asosida baholanadi. Bu natijalarni adolatli taqqoslash imkonini beradi.

Indekslar to‘plami:

$$I = \{1,2, \dots, m\},$$

bu yerda I – indekslar to‘plami, m – obyektlar soni.

Ushbu indekslar train va test qismlarga ajratiladi:

$$I = I_{train} \cup I_{test}$$

7-bosqich. Bo‘sh va anomal qiymatlarni to‘ldirish. Buzilgan datasetning train va test qismlardagi $M_i^j = 1$ bo‘lgan qiymatlar median yoki mode qiymati bilan to‘ldiriladi. To‘ldirish uchun har bir j-belgi bo‘yicha faqat o‘qitish qismidagi to‘g‘ri qiymatlardan mediana hisoblanadi:

$$\bar{x}_i^j = median\{x_i^j \mid i \in I_{train}, y_i = c, j \in F_g \cup F_r, C_j(x_i^j) = 1\},$$

(cc) BY

$$\bar{x}_i^j = mode\{x_i^j \mid i \in I_{train}, y_i = c, j \in F_n,\ C_j(x_i^j) = 1\},$$

$$\hat{x}_i^j = \begin{cases} \bar{x}_i^j, & M_i^j = 1, \\ x_i^j, & M_i^j = 0. \end{cases}$$

bu yerda $\hat{x}_i^j$ – to‘ldirilgan yoki tuzatilgan qiymat, $\bar{x}_i^j$ – j-belgi bo‘yicha o‘qitish qismidagi to‘g‘ri qiymatlar medianasi yoki mod qiymati, y_i – y sinflar vektoridagi i-obyektning tegishli bo’lgan sinf raqami, c – qaralayotgan obyekt tegishli bo‘lgan sinf raqami.

8-bosqich. To‘g‘rilangan datasetni shakllantirish. To‘ldirilgan qiymatlardan tashkil topgan dataset shakllantiriladi va uni quyidagicha belgilanadi:

$$D^{rep} = \{X^{rep}, y^{rep}\}$$

Shundan kelib chiqib quyidagi shart bajarilishi talab qilinadi:

$$I_{train}^{orig} = I_{train}^{corr} = I_{train}^{rep}$$

$$I_{test}^{orig} = I_{test}^{orig} = I_{test}^{orig}$$

9-bosqich. Uchta datasetni klassifikatsiya jarayonini testdan o‘tkazish.

Algoritm samaradorligini baholash uchun asl, buzilgan va qayta tiklangan datasetlar bir nechta klassifikatsiya algoritmlarida testdan o‘tkaziladi.

Tepada keltirilgan jarayondan kelib chiqib dastlabki ishlov berish algoritmini quyidagicha keltiriladi.

III. TADQIQOT NATIJALARI

Tadqiqotning maqsadi har xil turdagi belgilardan tashkil topgan uchta datasetlar tarkibidagi bo‘sh va anomal qiymatlarni klasslarni hisobga oluvchi moda yoki mediana asosida qayta tiklash va ushbu jarayonning klassifikatsiya sifatiga ta’sirini baholashdan iborat. Tajribada har bir dataset asl, buzilgan va qayta tiklangan holatlarda sakkizta klassifikatsiya modeli orqali baholandi.

Oddiy moda yoki mediana qiymati bilan to‘ldirish usulida qiymat barcha o‘qitish obyektlari bo‘yicha hisoblanadi. Taklif etilgan algoritmda esa qaralayotgan obyektning klass belgisi hisobga olinib, faqat shu klassga tegishli o‘qitish obyektlari orasidan tanlanadi.

1-jadval.

Ma’lumotlar to‘plamlarining umumiy tavsifi

Dataset	1-dataset	2-dataset	3-dataset
Obyektlar soni	140	340	13611
Belgilar soni	10	19	16
Klasslar soni	7	16	7
Sun’iy shakllantirilgan bo‘sh qiymatlar	70	323	10832
Sun’iy shakllantirilgan anomal qiymatlar	42	196	6496

1-dataset 140 ta obyekt, 10 ta nominal belgi va 7 ta klassdan iborat, 2-dataset esa 340 ta obyekt, 19 ta nominal belgi va 16 ta klassni o‘z ichiga oladi. 3-datasetda esa 13611 obyekt, 16 ta belgi bo‘lib shundan 12 ta geometrik o‘lcham belgisi va 4 shakl koefitsientini ifodalovchi belgidan iborat.

2-jadval.

Datasetlar bo‘yicha klassifikatsiya modellarining natijalarini taqqoslash

Model	Asl dataset						Buzilgan dataset						Qayta tiklangan dataset					
	Dry Bean dataset		1-dataset		2-dataset		Dry Bean dataset		1-dataset		2-dataset		Dry Bean dataset		1-dataset		2-dataset	
	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1	Aniqlik	F1

(cc) BY

Logistic Regression	91,9	93,0	91,0	90,4	89,4	88,7	59,9	58,1	88,1	87,2	85,7	85,0	90,1	91,4	90,0	89,3	87,8	87
KNN	91,5	92,7	91,8	91,0	90,3	89,4	76,3	76,9	68,3	67,7	66,3	65,7	90,1	91,4	90,5	89,6	88,7	87,7
SVM-RBF	92,3	93,4	92,7	91,7	91,6	90,5	84,1	85,2	82,2	81,6	80,9	79,9	91,5	92,6	91,6	90,5	90,1	88,9
Decision Tree	89,5	90,9	90,2	89,3	88,7	88,1	86,7	87,6	57,8	57,4	55,1	54,7	88,0	89,3	88,7	87,7	86,9	86,2
Random Forest	91,9	93,0	93,4	92,7	92,2	91,4	91,3	92,4	81,6	81,0	80,6	79,9	91,6	92,7	92,3	91,5	90,8	89,9
Extra Trees	91,4	92,6	92,9	92,0	91,8	90,8	90,6	91,7	85,3	84,7	84,1	83,2	91,1	92,3	91,9	90,9	90,4	89,3
Gradient Boosting	91,9	93,0	92,1	91,3	90,9	90	91,3	92,3	76,5	75,7	76,6	75,9	91,2	92,5	91,2	90,3	89,3	88,3
Gaussian NB	89,7	90,7	89,5	88,5	88,1	87,4	52,3	51,2	39,6	39,1	38,9	38,6	88,9	89,9	87,9	86,8	85,8	85
O‘rtacha	91,3	92,4	91,7	90,9	90,4	89,5	79,1	79,4	72,4	71,8	71,0	70,4	90,3	91,5	90,5	89,6	88,7	87,8

Jadvalda taklif etilgan ma’lumotlarni qayta tiklash algoritmining samaradorligi Logistic Regression, KNN, SVM-RBF, Decision Tree, Random Forest, Extra Trees, Gradient Boosting va Gaussian Naive Bayes klassifikatsiya modellari asosida baholangan. Tajribada asl, sun’iy buzilgan va qayta tiklangan datasetlar natijalari Accuracy (Aniqlik) hamda F1-score ko‘rsatkichlari bo‘yicha taqqoslangan.

Natijalarga ko‘ra, sun’iy buzilishlar kiritilganda barcha datasetlarda klassifikatsiya samaradorligining pasayishi kuzatilgan. Ayniqsa, Gaussian Naive Bayes va Decision Tree modellarida pasayish sezilarli bo‘lgan. Masalan, Dry Bean datasetida Gaussian Naive Bayes modelining F1-score qiymati 90,7% dan 51,2% gacha, Logistic Regression modelida esa 93,0% dan 58,1% gacha kamaygan.

Taklif etilgan algoritm yordamida ma’lumotlar qayta tiklangandan so‘ng barcha modellarda natijalar sezilarli darajada yaxshilangan. Dry Bean datasetida o‘rtacha Accuracy 79,1% dan 90,3% gacha, F1-score esa 79,4% dan 91,5% gacha oshgan. 1-datasetda ushbu ko‘rsatkichlar mos ravishda 72,4% dan 90,5% gacha va 71,8% dan 89,6% gacha, 2-datasetda esa 71,0% dan 88,7% gacha va 70,4% dan 87,8% gacha tiklangan.

Umuman olganda, qayta tiklangan dataset natijalari asl dataset natijalariga yaqinlashgani taklif etilgan algoritm bo‘sh va anomal qiymatlarning klassifikatsiya jarayoniga salbiy ta’sirini kamaytirishini ko‘rsatadi. Eng yuqori barqarorlik Random Forest, Extra Trees va Gradient Boosting modellarida kuzatilgan. Mazkur natijalar ishlab chiqilgan algoritmning

(cc) BY

turli hajmdagi jadvalli ma’lumotlarda samarali qo‘llanishi mumkinligini tasdiqlaydi.

IV. XULOSA

Ushbu tadqiqotda jadvalli ma’lumotlarda uchraydigan bo‘sh va anomal qiymatlarni aniqlash hamda qayta tiklashga mo‘ljallangan adaptiv dastlabki ishlov berish algoritmi ishlab chiqildi va uning samaradorligi eksperimental baholandi.

Taklif etilgan algoritmning asosiy xususiyati ma’lumotlarni oddiy statistik to‘ldirish bilan cheklanmasdan, har bir belgining turi, ruxsat etilgan qiymatlar sohasi hamda obyektning sinfga mansubligini hisobga olishidir. Sonli va tartibli belgilar uchun mediana, nominal va binar belgilar uchun moda qiymatidan foydalanildi.

Tajriba 1-tur qandli diabet, 2-tur qandli diabet va Dry Bean datasetlari asosida olib borildi. Datasetlarning original, sun’iy buzilgan va qayta tiklangan holatlari sakkizta mashinali o‘qitish modeli yordamida taqqoslandi.

Barcha datasetlarda qayta tiklangan ma’lumotlar asosidagi natijalar buzilgan dataset natijalaridan yuqori bo‘ldi. Bu taklif etilgan algoritmning jadvalli ma’lumotlar sifatini yaxshilash va mashinali o‘qitish modellarining klassifikatsiya samaradorligini tiklashdagi amaliy samaradorligini tasdiqlaydi.

Kelgusidagi tadqiqotlarda qayta tiklash jarayoniga qo‘shni obyektlar, ehtimollik modellari va ansambl yondashuvlarini integratsiya qilish, shuningdek real klinik ma’lumotlarda algoritmni prospektiv sinovdan o‘tkazish maqsadga muvofiq.

FOYDALANILGAN ADABIYOTLAR

1. Barrabes M., Perera M., Novelle Moriano V., Giro-I-Nieto X., Mas Montserrat D., Ioannidis A. G. Advances in biomedical missing data imputation: A survey // IEEE Access. – 2025. – Vol. 13. – P. 16918–16932. DOI: 10.1109/ACCESS.2024.3516506.
2. Afkanpour M., Hosseinzadeh E., Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review // BMC Medical Research Methodology. – 2024. – Vol. 24, No. 1. – P. 1–13. DOI: 10.1186/s12874-024-02310-6.
3. Hu Y.-H., Wu R.-Y., Lin Y.-C., Lin T.-Y. A novel MissForest-based missing values imputation approach with recursive feature elimination in medical applications // BMC Medical Research Methodology. – 2024. – Vol. 24, No. 1. – P. 1–12. DOI: 10.1186/s12874-024-02392-2.
4. Agbo B., Al-Aqrabi H., Alsboui T., Hussain M., Hill R. Imputation of missing clinical covariates for downstream classification problems // IEEE Access. – 2023. – Vol. 11. – P. 102935–102943. DOI: 10.1109/ACCESS.2023.3317775.
5. Ribeiro C. E., Freitas A. A data-driven missing value imputation approach for longitudinal datasets // Artificial Intelligence Review. – 2021. – Vol. 54, No. 8. – P. 1–31. DOI: 10.1007/s10462-021-09963-5.
6. Shin K., Han J., Kang S. MI-MOTE: Multiple imputation-based minority oversampling technique for imbalanced and incomplete data classification // Information Sciences. – 2021. – Vol. 575. – P. 80–89. DOI: 10.1016/j.ins.2021.06.043.
7. Altukhova O. Choice of method imputation missing values for obstetrics clinical data // Procedia Computer Science. – 2020. – Vol. 176. – P. 976–984. DOI: 10.1016/j.procs.2020.09.093.
8. Alghawli A. S. Complex methods detect anomalies in real time based on time series analysis // Alexandria Engineering Journal. – 2022. – Vol. 61, No. 1. – P. 549–561. DOI: 10.1016/j.aej.2021.06.033.
9. Hubbard R. A., Lett E., Ho G. Y. F., Chubak J. Characterizing bias due to differential exposure ascertainment in electronic health record data // Health Services and Outcomes Research Methodology. – 2021. – Vol. 21, No. 3. – P. 309–323. DOI: 10.1007/s10742-020-00235-3.
10. Jarrett D., Cebere B. C., Liu T., Curth A., van der Schaar M. HyperImpute: Generalized iterative imputation with automatic model selection // Proceedings of

(cc) BY *dtai.uz*

the 39th International Conference on Machine Learning (ICML). – PMLR, 2022. – P. 9916–9937. DOI: 10.48550/arXiv.2206.07769

11. Linkon A. A. Evaluation of feature transformation and machine learning models on early detection of diabetes mellitus // IEEE Access. – 2024. – Vol. 12. – P. 165425–165440. DOI: 10.1109/ACCESS.2024.3488743
12. Nishanov A. Kh., Djurayev G. P. Solving stages of classification in the preliminary treating of medical examples // Tekhnicheskie nauki: problemy i resheniya: sbornik statey XXXIII mezhdunarodnoy nauchno-prakticheskoy konferentsii. No. 2 (30). – Moskva, 2020. – P. 69–76.
13. Nishanov A. Kh., Jo‘rayev G‘. P., Azatov B. A. Ma’lumotlarga dastlabki ishlov berishning ayrim masalalarini yechish bosqichlari // V Mezhdunarodnaya nauchno-prakticheskaya konferentsiya “Nauka i obrazovanie v sovremennom mire: vyzovy XXI veka”. Tom I. – Nur-Sultan, 10–12 dekabrya 2019 g. – P. 23–27.
14. Nishanov A. Kh., Jurayev G. P., Narziyev N. B. Adaptable genetic algorithm for primary processing of medical data based on random search // Shestaya mezhdunarodnaya nauchno-prakticheskaya konferentsiya “BIG DATA and Advanced Analytics”. Ch. 1. – Minsk, 20–21 maya 2020 g. – P. 147–155.
15. Nishanov A. Kh., Ollamberganov F. F., Khasanova M. A., Mengturayev F. Z., Allayarov U. B. Algorithms for clustering, classification and selection of informative features for early detection of brain cancer // Acta IMEKO. – 2025. – Vol. 14, No. 3. – P. 1–8. DOI: 10.21014/actaimeko.v14i3.2053
16. Nishonov A.X., Mengturayev F. Z., Ollamberganov F.F., Allayarov U.B., Khasanova M.A., Doniyorova G.T. An algorithm for creating a semi-synthetic dataset for diabetes KazNU Bulletin. Mathematics, Mechanics, Computer Science Series ISSN 1563–0277 Vol. 129 No. 1 (2026): -P. 114-128

(cc) BY *dtai.uz*