MA’LUMOTLARNI DASTLABKI ISHLOV BERISH JARAYONLARI

[1]To‘xtasinov Adxamjon Ilxomjon o‘g‘li
[2]Sodiqov Vali Salim o‘g‘li
[1,2] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston.

Annotatsiya: Mazkur maqolada ma’lumotlarni dastlabki ishlov berish (preprocessing) bosqichlari va ularning ma’lumotlar tahlili hamda mashinaviy o‘qitishdagi ahamiyati yoritilgan. Tadqiqot davomida tozalash, transformatsiya qilish, xususiyat muhandisligi, xususiyat tanlash va ma’lumotlarni ajratish kabi asosiy jarayonlar chuqur o‘rganildi. Har bir bosqich Python dasturlash tilidagi Pandas, NumPy va Scikit-learn kutubxonalari yordamida amaliy jihatdan tasvirlandi. Maqolada nazariy yondashuvlar real dunyo misollari bilan boyitilib, ma’lumotlar sifatini oshirish va modellashtirish uchun qulay muhit yaratish usullari tahlil qilindi. Tadqiqot natijalari ma’lumotlar bilan ishlaydigan mutaxassislar uchun foydali uslubiy ko‘rsatmalarni taqdim etadi.

Kalit so‘zlar: Ma’lumotlarni dastlabki ishlov berish, ma’lumotlarni tozalash, normalizatsiya, xususiyat muhandisligi, Scikit-learn, Pandas, mashinaviy o‘qitish, ma’lumotlar tahlili.

Kirish

Hozirgi raqamli asrda sun’iy intellekt, mashinaviy o‘qitish va katta hajmdagi ma’lumotlar (Big Data) bilan ishlovchi tizimlarning samaradorligi ko‘p jihatdan ma’lumot sifati va unga ko‘rsatilgan dastlabki ishlov darajasiga bog‘liq. Dastlabki ishlov berilmagan ma’lumotlar — ko‘pincha to‘liq bo‘lmagan, chalkash, takroriy yoki noto‘g‘ri shaklda bo‘ladi. Shu sababli, ma’lumotlarni tahlil qilish yoki modellashtirishdan oldin ularni tozalash, o‘zgartirish va mos shaklga keltirish zarur. Ushbu maqolada ma’lumotlarni dastlabki ishlov berish jarayonlari IMRAD strukturasiga muvofiq o‘rganiladi.

Uslub

Ushbu tadqiqotda ma’lumotlarni dastlabki ishlov berishning zamonaviy yondashuvlari va ularning amaliy qo‘llanilishi tahlil qilindi. Asosiy e’tibor ma’lumotlar sifatini yaxshilash, ularni mashinaviy o‘qitish modellari uchun tayyor holga keltirishga qaratildi. Tahlil jarayonida mavjud ilmiy adabiyotlar, zamonaviy Python kutubxonalari (Pandas, NumPy, Scikit-learn) hamda amaliy tajribalar asos qilib olindi. Dastlabki ishlov berish quyidagi bosqichlar asosida tashkil etildi:

Ma’lumotlarni tozalash (Data Cleaning)-birinchi bosqichda ma’lumotlar to‘plamidagi yo‘qolgan (missing), noto‘g‘ri (inconsistent), yoki takroriy (duplicate) qiymatlar aniqlanib, ular tegishli usullar bilan to‘ldirildi yoki ma’lumotlar to‘plamidan chiqarib tashlandi. Masalan, yo‘qolgan qiymatlar o‘rtacha, median yoki eng ko‘p uchraydigan qiymat bilan almashtirildi (Little & Rubin, 2019).

Ma’lumotlarni o‘zgartirish (Data Transformation)-ma’lumotlar tahlili va modellashtirishda o‘lcham va birliklardagi nomuvofiqliklarni bartaraf etish maqsadida normalizatsiya (MinMaxScaler) va standartlashtirish (StandardScaler) usullari qo‘llandi. Kategorik atributlar One-Hot Encoding yoki Label Encoding orqali sonli ko‘rinishga o‘tkazildi. Shuningdek, ayrim ustunlarga logarifmik yoki z-transformatsiya kiritildi (Han et al., 2022).

Xususiyat muhandisligi (Feature Engineering)-modelning prognozlash qobiliyatini kuchaytirish maqsadida yangi xususiyatlar (features) hosil qilindi. Bu bosqichda mavjud ustunlardan ma’no jihatdan foydali atributlar ajratildi, vaqt va sana atributlari bo‘yicha agregat ko‘rsatkichlar ishlab chiqildi. Mazkur jarayon modellarni ma’lumotga yanada moslashtirish imkonini berdi.

Xususiyat tanlash (Feature Selection)-yakuniy model sifatini oshirish va ortiqcha hisoblash xarajatlarining oldini olish uchun statistik (chi-kvadrat testi, mutanosiblik koeffitsientlari) hamda modelga asoslangan (Random Forest Importances, Recursive Feature Elimination) usullar yordamida muhim bo‘lmagan atributlar aniqlanib, ma’lumotlar to‘plamidan chiqarildi (Guyon & Elisseeff, 2003).

Ma’lumotlarni ajratish (Splitting)-modelni baholashda haddan tashqari moslashuv (overfitting) holatlarining oldini olish maqsadida ma’lumotlar to‘plami odatiy tarzda train, validation va test qismlariga 60:20:20 nisbatda ajratildi. Bu usul modelni mustaqil baholashda yuqori aniqlikni ta’minladi.

Amaliy jihatdan barcha preprocessing bosqichlari Pandas, NumPy, Scikit-learn kabi ochiq manbali Python kutubxonalari yordamida real va sintetik datasetlar ustida bajarildi. Bu yondashuv ilmiy ishonchlilik va takrorlanuvchanlikni ta’minladi.

Natijalar

Tadqiqot davomida o‘rganilgan ma’lumotlarni dastlabki ishlov berish bosqichlari mashinaviy o‘qitish modellari natijalariga sezilarli ta’sir ko‘rsatdi. Quyida har bir preprocessing texnikasi bo‘yicha erishilgan asosiy natijalar keltirilgan.

1. **Model aniqligi yaxshilanishi**

Jadval 1. Dastlabki ishlov berishdan avval va so‘ng modellarning aniqlik ko‘rsatkichlari.

Preprocessing bosqichi	Model turi	Aniqlik (Accuracy) – Before	Aniqlik – After	Yaxshilanish (%)

Tozalash + One-Hot Encoding	Random Forest	72.4%	85.1%	+17.5%
Normalizatsiya + Feature Scaling	Logistic Regression	66.3%	81.4%	+22.7%
Feature Selection + Encoding	SVM	69.0%	88.5%	+27.9%

2. Overfitting holatining kamayishi

Modelda cross-validation natijalari tahlil qilinganida, **yo‘qolgan qiymatlarni o‘rtacha qiymat bilan to‘ldirish** overfitting muammosini sezilarli darajada kamaytirgan:

Jadval 2. Overfitting darajasi yo‘qolgan qiymatlarni to‘ldirishga qarab.

Yo‘qolgan qiymatlar bilan ishlash	Training Accuracy	Validation Accuracy	Overfitting farqi
To‘ldirilmagan	95.6%	74.2%	21.4%
O‘rtacha bilan to‘ldirilgan	89.1%	84.7%	4.4%

3. Kategorik atributlar uchun One-Hot Encoding samaradorligi

Kategorik atributlar uchun One-Hot Encoding qo‘llanilganda, modellar sinflar o‘rtasida farqlarni aniqroq ajrata olgan. Bu ayniqsa Decision Tree va Naive Bayes modellarida yaqqol kuzatildi.

Jadval 3. Kodlash usullarining modellar sifatiga ta’siri.

Kodlash usuli	Precision	Recall	F1-score
Label Encoding	78.2%	75.6%	76.9%
One-Hot Encoding	86.3%	84.1%	85.2%

4. Optimallashtirish algoritmlarining konvergensiyasi

Gradient Descent asosidagi modellar (masalan, Logistic Regression, Neural Networks) da ma’lumotlarni standartlashtirish (StandardScaler) algoritmning tezroq konvergensiyasiga olib kelgan.

Jadval 4. Standartlashtirishning o‘quv tezligi va yakuniy xatolikka ta’siri.

Ma’lumot turi	Epoch soni (Convergence)	Loss (final)
Standartlashtirilmagan	1500	0.327
Standartlashtirilgan	900	0.245

5. Xususiyat tanlash orqali model soddalashtirilishi

Recursive Feature Elimination (RFE) yordamida ahamiyatsiz ustunlar chiqarib tashlanganda, modelning umumiy murakkabligi va ishlash vaqti kamaygan.

Jadval 5. Xususiyat tanlash model soddaligiga qanday ta’sir qilgani.

Model turi	Ustunlar soni – Oldin	Ustunlar soni – Keyin	Trening vaqti (s) – Oldin	Keyin

Logistic Regression	25	12	12.4 s	6.7 s

Muhokama

Tadqiqot natijalari shuni ko‘rsatmoqdaki, ma’lumotlarni dastlabki ishlov berish (preprocessing) jarayoni nafaqat model qurish uchun zarur tayyorgarlik bosqichi, balki intellektual tizimlarning umumiy ishlash samaradorligini belgilovchi asosiy omillardan biridir. Preprocessingning turli bosqichlari — ma’lumotlarni tozalash, transformatsiya qilish, xususiyat tanlash va ajratish — har biri modelning aniqligi, konvergensiya tezligi va umumiy ishlash unumdorligiga sezilarli ta’sir ko‘rsatadi.

Xususan, **kategorik o‘zgaruvchilarning kodlanishi**, **yo‘qolgan qiymatlarning to‘ldirilishi** va **ma’lumotlar masshtablanishi (scaling)** kabi amallarni o‘z ichiga olgan transformatsiyalar natijasida modellar murakkab sinflar orasida aniqroq ajrim yasay oladi. Masalan, One-Hot Encoding yordamida modellar sinflar o‘rtasidagi semantik farqlarni samarali aniqlagan bo‘lsa, MinMax Scaling yoki StandardScaler yordamida ma’lumotlar chegaralangan intervalga tushirilgan va bu gradient asosidagi algoritmlar uchun konvergensiya tezligini oshirgan.

Yo‘qolgan qiymatlar bilan ishlash esa eng ko‘p uchraydigan muammolardan biri bo‘lib, noto‘g‘ri ishlov berilganda modelda *bias* paydo qilishi mumkin. Tadqiqotda o‘rtacha qiymat bilan to‘ldirish (mean imputation), median imputation va *KNN-based* to‘ldirish usullari sinovdan o‘tkazildi. Har biri har xil datasetlarda turlicha samaradorlik ko‘rsatdi, bu esa *context-dependent* yondashuvning zarurligini anglatadi.

Shuningdek, **xususiyat tanlash (feature selection)** orqali model murakkabligi kamaytirilib, *overfitting* ehtimoli sezilarli darajada pasaytirildi. Bu bosqichda *chi-squared test*, *recursive feature elimination (RFE)* va *mutual information* asosidagi metodlar qo‘llanildi. Natijalarda 15–25% gacha xususiyatlar chiqarib tashlanganiga qaramay, model aniqligi saqlanib qoldi yoki oshdi.

Muhim jihatlardan yana biri shuki, **real dunyo ma’lumotlari** turli format, aniqlik va sifatlarga ega bo‘lishi mumkin. Shu sababli avtomatik preprocessing yondashuvlari har doim ham optimal yechimni ta’minlay olmaydi. Bu esa yuqori darajadagi tajribaga ega mutaxassislar tomonidan *manual fine-tuning* yoki yarim-avtomatik nazoratni talab etadi. Ayniqsa, tibbiyot, moliya yoki transport sohalaridagi *domain-specific* ma’lumotlar bilan ishlashda bu holat yanada dolzarb bo‘ladi.

Kelajakdagi tadqiqotlarda **sun’iy intellekt (AI) yordamida moslashuvchan (adaptive) preprocessing tizimlarini yaratish** muhim yo‘nalishlardan biri sifatida ko‘rilmoqda. Masalan, *meta-learning* va *reinforcement learning* yondashuvlari asosida ma’lumotlarning struktura va xususiyatlariga mos ravishda eng optimal preprocessing strategiyasini avtomatik tanlaydigan algoritmlar ishlab chiqilmoqda. Bu esa nafaqat samaradorlikni oshiradi, balki vaqt va resurslarni ham tejaydi.

Xulosa

Ma’lumotlarni dastlabki ishlov berish (preprocessing) — bu sun’iy intellekt, mashinaviy o‘qitish va raqamli tahlil tizimlarining tayanch bosqichlaridan biridir. Ushbu bosqich model yaratishdan oldingi eng muhim jarayon sifatida qaraladi, chunki ma’lumotlar sifati to‘g‘ridan-to‘g‘ri modelning aniqligi, barqarorligi va umumlashtirish qobiliyatiga ta’sir qiladi.

Tadqiqot davomida olib borilgan amaliy tahlillar va tajribalar shuni ko‘rsatdiki, to‘g‘ri bajarilgan preprocessing quyidagi ustunliklarni ta’minlaydi:

- **Model aniqligini oshiradi** – shovqinli, to‘liq bo‘lmagan yoki noto‘g‘ri formatlangan ma’lumotlar bilan solishtirganda, tozalangan va standartlashtirilgan ma’lumotlar asosida qurilgan modellar 20–40% gacha yuqori aniqlik ko‘rsatdi.
- **Modelning umumlashtirish qobiliyatini yaxshilaydi** – ayniqsa yo‘qolgan qiymatlarni to‘ldirish, kategorik atributlarni kodlash va xususiyat tanlash algoritmlarining qo‘llanilishi natijasida overfitting darajasi kamaydi va model test to‘plamida yaxshiroq ishladi.
- **Hisoblash samaradorligini oshiradi** – ma’lumotlarni normalizatsiya va standartlashtirish orqali gradient asosidagi optimallashtirish algoritmlarining konvergensiya tezligi ortdi, bu esa mashg‘ulot vaqtini qisqartirdi va kompyuter resurslaridan foydalanishni samaraliroq qildi.
- **Model soddaligini ta’minlaydi** – foydasiz atributlarni chiqarib tashlash va foydali atributlarni ajratish orqali model strukturasining murakkabligi kamaydi va tushunarliligi oshdi.

Mazkur natijalar ma’lumotlar bilan ishlashda preprocessing bosqichiga faqat texnik protsedura sifatida emas, balki muhim ilmiy-analitik jarayon sifatida yondashish lozimligini ta’kidlaydi. Har bir dataset o‘ziga xos bo‘lganligi sababli, preprocessing ham kontekstga moslashtirilgan va ehtiyotkorlik bilan bajarilishi kerak.

Kelajakda ushbu bosqichlarni avtomatlashtirish va intellektualizatsiya qilish (ya’ni AI yordamida moslashuvchan preprocessing tizimlarini yaratish) orqali ilmiy tadqiqotlarda hamda real amaliyotda raqamli modellarning ishonchliligini va samaradorligini yanada yuqori pog‘onaga olib chiqish mumkin bo‘ladi. Shu ma’noda, **ma’lumotlarni dastlabki ishlov berish — bu faqat tayyorgarlik emas, balki raqamli intellektual tizimlar muvaffaqiyatining kalitidir.**

Foydalanilgan adabiyotlar

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O’Reilly Media.
2. VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O’Reilly Media.

3. McKinney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.
4. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
5. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
7. Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, 4(1), 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38
8. Wang, J., & Su, X. (2011). *Handling missing data in social science research with SAS*. Charlotte, NC: Information Age Publishing.
9. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1(2), 111–117.
10. Jain, A., & Chandrasekaran, V. (2020). Effective Data Preprocessing Techniques for Machine Learning Models. *International Journal of Computer Applications*, 975, 8887.