ML ALGORITMLARINI TAQQOSLASH: O'QUVCHI O'ZLASHTIRISH DARAJASINI BASHORATLASHDA XGBOOST VA LOGISTIC REGRESSION SAMARADORLIGI

Yunusxo'jayev Lutfullo Zafarxo'ja o'g'li
Magistratura talabasi
yunusxojayevlutfullo@gmail.com
Ilmiy rahbar: **Jalolov Tursunbek Sadriddinovich**
p.f.f.d. (PhD), dotsent Osiyo xalqaro universiteti

ANNOTATSIYA

Ushbu maqolada o'quvchilarning akademik o'zlashtirish darajasini bashoratlashda XGBoost va Logistic Regression algoritmlarining qiyosiy tahlili amalga oshirilgan. Tadqiqot davomida 750 nafar o'quvchining 18 ta belgidan iborat ma'lumotlar to'plami ustida turli mashinaviy o'qitish algoritmlarining ishlash ko'rsatkichlari o'rganildi. Natijalar shuni ko'rsatadiki, XGBoost algoritmi 91.2% aniqlik ko'rsatkichi bilan eng yuqori natijaga erishdi, Logistic Regression esa 78.4% bilan sodda talqin etiluvchi alternativ sifatida tavsiya etildi. Ushbu tadqiqot ta'lim muassasalarida algoritmni tanlash bo'yicha amaliy ko'rsatmalar beradi.

KALIT SO'ZLAR: XGBoost, Logistic Regression, qiyosiy tahlil, o'quvchi o'zlashtirishi, mashinaviy o'qitish, Learning Analytics, ta'lim texnologiyalari.

KIRISH

Ta'lim sohasida sun'iy intellekt va mashinaviy o'qitish usullarini qo'llash so'nggi yillarda jadal sur'atda rivojlanmoqda. Dunyo bo'ylab minglab ta'lim muassasalari o'quvchilarning akademik ko'rsatkichlarini monitoring qilish va erta ogohlantirish tizimlari yaratish maqsadida turli xil algoritmlardan foydalanmoqda [1].

Biroq, turli algoritmlarning o'ziga xos kuchli va zaif tomonlari mavjud bo'lib, ularni to'g'ri tanlash tadqiqot natijalarining sifatini bevosita belgilaydi. Murakkab gradient boosting algoritmlari (XGBoost, LightGBM) yuqori aniqlik ko'rsatsa-da, ularning talqin qilinishi qiyinroq. Oddiy modellar (Logistic Regression, Decision Tree) esa past aniqlik evaziga tushunarliroq natijalar beradi [2].

Ushbu tadqiqot XGBoost va Logistic Regression algoritmlarini to'rt asosiy mezon bo'yicha — aniqlik, tezlik, talqin qilinishi va miqyoslanish imkoniyati — qiyosiy tahlil qilishni maqsad qilgan. Bundan tashqari, ikki algoritm o'rtasida gibrid yondashuv (stacking ensemble) qo'llash imkoniyati ham o'rganildi.

Tadqiqotning dolzarbligi shundaki, ta'lim muassasalari aksariyat hollarda resurs cheklovi va texnik salohiyat yetishmasligi sababli eng murakkab algoritmni emas, balki maqsadga

muvofiq algoritmni tanlashga majbur bo'ladi. Shu sababli qiyosiy tahlil natijalarimiz amaliy qarorlar qabul qilishda muhim ko'rsatma bo'lib xizmat qiladi.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

2.1. Adabiyotlar sharhi

Chen va Guestrin (2016) XGBoost algoritmini taqdim etib, u gradient boosting framework'ining parallellashtirish va regularizatsiya imkoniyatlarini takomillashtirishi orqali raqobatbardosh natijalar berishini ko'rsatdi [3]. Hosmer va Lemeshow (2018) esa Logistic Regression modelining ta'lim sohasidagi o'ziga xos afzalliklarini ta'kidlab o'tdi [4].

Mduma va boshq. (2019) o'quvchi tashlab ketishini bashoratlashda 12 ta algoritm taqqoslagan va XGBoost 88.9% bilan birinchi o'ringa chiqganini aniqlagan [5]. Hämäläinen va Vinni (2010) esa Logistic Regression ning interpretatsiya imkoniyati ta'lim amaliyotchilariga natijalani tushuntirishda muhim ekanligini ko'rsatgan [2].

2.2. Tadqiqot dizayni va ma'lumotlar

Tadqiqotda Toshkent va Samarqand viloyatlaridan 750 nafar o'rta maktab o'quvchisining 2 yillik ma'lumotlari ishlatildi. Ma'lumotlar to'plami 18 ta belgini o'z ichiga oldi. Ikkala algoritm ham bir xil sharoitda: Python 3.11, scikit-learn 1.3, XGBoost 2.0 da sinovdan o'tkazildi. Baholash protokoli: stratified k-fold cross-validation (k=10), har bir fold uchun aniqlik, F1-score, AUC-ROC hisoblandi.

1-jadval. Ma'lumotlar to'plami tarkibi

Kategoriya	Belgilar	Soni
Akademik	Baholar, test natijalari, fanlar kesimida o'rtacha	6 ta
Xulq-atvor	Davomati, uy vazifasi, sinfdagi faollik, kechikish	5 ta
Ijtimoiy	Oila tarkibi, ota-ona ta'limi, daromad, yashash joyi	5 ta
Demografik	Yosh, jinsi	2 ta
Jami	—	18 ta

Ma'lumotlarni oldindan qayta ishlash (preprocessing) bosqichlari: yo'qolgan qiymatlarni KNN imputation bilan to'ldirish, kategorik o'zgaruvchilarni one-hot encoding bilan kodlash, raqamli o'zgaruvchilarni StandardScaler bilan normalizatsiya qilish. Sinf muvozanati tekshirildi: past (22%), o'rta (51%), yuqori (27%).

NATIJALAR

3.1. Asosiy samaradorlik ko'rsatkichlari

Tajriba natijalari to'rt algoritm ichida XGBoost eng yuqori natija ko'rsatganini tasdiqladi: 91.2% aniqlik, 0.91 F1-score va 0.94 AUC-ROC. Cross-validation natijasida XGBoost 90.8% ± 0.9% barqaror natija ko'rsatdi. Logistic Regression esa 78.4% aniqlik bilan to'rtinchi o'rinda bo'lsa ham, eng qisqa o'quv vaqti (0.8 soniya) va eng yuqori talqin qilinish qobiliyati bilan ajralib turdi.

2-jadval. Algoritmlarni qiyosiy tahlil natijalari

Mezon	XGBoost	Log. Reg.	Rand. Forest	Naive Bayes	Decision Tree
Aniqlik (%)	91.2	78.4	89.3	72.1	81.6
F1-Score	0.91	0.78	0.89	0.71	0.81
AUC-ROC	0.94	0.84	0.93	0.79	0.85
O'quv vaqti (s)	12.4	0.8	18.6	0.3	1.2
Talqin qilish	O'rta	Yuqori	Past	Yuqori	Yuqori
Miqyoslanish	A'lo	Yaxshi	Yaxshi	A'lo	O'rta

3.2. XGBoost giperaparametrlarini sozlash

Bayesian optimization yordamida XGBoost ning optimal giperaparametrlari aniqlandi: learning_rate=0.05, n_estimators=300, max_depth=6, subsample=0.8, colsample_bytree=0.7, reg_alpha=0.1, reg_lambda=1.0. Sozlashdan oldingi aniqlik 88.7% bo'lsa, sozlashdan keyin 91.2%ga ko'tarildi (2.5 foizlik oshish).

3.3. Stacking Ensemble natijasi

XGBoost va Logistic Regression ni meta-learner sifatida birlashtirilgan stacking ensemble modeli yaratildi. Meta-learner sifatida Ridge Regression ishlatildi. Stacking modeli 92.1% aniqlikka erishib, individual modellardan ham yaxshiroq natija ko'rsatdi. Biroq bu yondashuv o'quv vaqtini ikki hissa oshirdi (26.4 soniya).

3-jadval. Stacking Ensemble natijasi taqqosi

Model	Aniqlik (%)	F1-Score	O'quv vaqti (s)
XGBoost (yagona)	91.2	0.91	12.4
Logistic Regression (yagona)	78.4	0.78	0.8
Stacking (XGB + LR)	92.1	0.92	26.4

Random Forest (taqqos uchun)	89.3	0.89	18.6

MUHOKAMA

4.1. Algoritmni tanlash mezonlari

Olingan natijalar shuni ko'rsatadiki, algoritmni tanlash ta'lim muassasasining maqsad va imkoniyatlariga qarab amalga oshirilishi lozim. Yuqori aniqlik talab etiladigan, katta hajmdagi ma'lumotlar bilan ishlaydigan va texnik resurslar yetarli bo'lgan holatlarda XGBoost algoritmi optimal tanlov hisoblanadi. Biroq natijalari o'qituvchi va ota-onalarga tushuntirilishi zarur bo'lgan hollarda Logistic Regression afzalroq bo'lishi mumkin [5].

Stacking ensemble yondashuvi eng yuqori aniqlikni ta'minlasa-da, hisoblash resurslariga qo'shimcha talab qo'yadi. Shu sababli cheklangan IT infratuzilmasiga ega maktablar uchun XGBoost yagona model sifatida tavsiya etiladi, katta shahar maktablari uchun esa stacking yondashuvi maqbulroq.

4.2. Logistic Regression ning noyob afzalliklari

Garchi Logistic Regression aniqlik jihatidan XGBoost dan 12.8 foiz past bo'lsa-da, uning bir qator noyob afzalliklari mavjud. Birinchidan, koeffitsientlarni to'g'ridan-to'g'ri talqin qilish imkoniyati: masalan, 'dars davomati 1 foizga oshishi o'quvchining yuqori guruhga kirish ehtimolini 3.2 foizga oshiradi' degan aniq bayonotlar berish mumkin. Bu o'qituvchilar va ota-onalar bilan muloqotda juda muhim [2].

Ikkinchidan, Logistic Regression ning o'quv vaqti (0.8 soniya) XGBoost nikiga nisbatan 15 barobar qisqa. Bu katta hajmdagi ma'lumotlarda yoki resurs cheklangan tizimlarda muhim afzallik. Uchinchidan, model statistik ahamiyatlilik testlarini (p-value) qo'llash imkonini beradi, bu esa akademik tadqiqotlarda muhim.

4.3. Cheklovlar va kelajakdagi yo'nalishlar

Tadqiqotning asosiy cheklovi: ma'lumotlar to'plami O'zbekistonning faqat ikkita viloyatini qamrab olgan. Turli ijtimoiy-iqtisodiy sharoitlardagi maktablar o'rtasidagi farqlar modelning umumlashuvchanligiga ta'sir qilishi mumkin. Bundan tashqari, onlayn ta'lim platformalari (Moodle, Zoom) dan olingan ma'lumotlar hali qo'shilmagan.

Kelajakdagi tadqiqotlarda quyidagi yo'nalishlar tavsiya etiladi: (1) LightGBM va CatBoost algoritmlarini qo'shib, taqqoslash doirasini kengaytirish; (2) SHAP (SHapley Additive exPlanations) yordamida XGBoost ning talqin qilinishini yaxshilash; (3) real vaqt rejimida ishlash uchun stream processing arxitekturasini ishlab chiqish; (4) ko'p tilli ma'lumotlar bilan ishlash (o'zbek, rus, ingliz tilidagi matnlar).

XULOSA

Ushbu qiyosiy tadqiqot XGBoost algoritmining o'quvchilar o'zlashtirishini bashoratlashda eng yuqori aniqlik (91.2%, AUC-ROC 0.94) ni ta'minlashini isbotladi. XGBoost va Logistic

Regression dan iborat stacking ensemble modeli 92.1% aniqlik bilan individual modellardan ustun chiqdi.

Biroq, amaliy ta'lim muhitida algoritmni tanlash faqat aniqlik ko'rsatkichi bilan chegaralanmasligi, balki talqin qilish imkoniyati, hisoblash resurslari va amalga oshirish murakkabligi ham hisobga olinishi kerakligi ta'kidlandi. Resurs cheklangan muhitlar uchun Logistic Regression, yuqori aniqlik zarur bo'lgan hollarda XGBoost, ikkalasining afzalliklarini birlashtirish uchun stacking ensemble tavsiya etiladi.

Ushbu tadqiqot natijalari ta'lim muassasalari va ta'lim texnologiyalari sohasidagi amaliyotchilar uchun algoritmni tanlashda muhim amaliy ko'rsatma bo'lib xizmat qiladi va O'zbekistonda ta'lim sifatini oshirish yo'lida raqamlashtirish strategiyasiga hissa qo'shadi.

ADABIYOTLAR RO’YXATI

1. Dutt A., Ismail M.A., Herawan T. A Systematic Review on Educational Data Mining // IEEE Access. 2017. №5. Pp. 15991–16005.

2. Hämäläinen W., Vinni M. Classifiers for Educational Data Mining // Handbook of Educational Data Mining. 2010. Pp. 57–71.

3. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of KDD. 2016. Pp. 785–794.

4. Hosmer D.W., Lemeshow S. Applied Logistic Regression. 3rd ed. New York: Wiley, 2018. 528 p.

5. Mduma N., Kalegele K., Machuve D. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction // The Electronic Journal of Information Systems in Developing Countries. 2019. №85(3). Pp. 1–11.

6. Lundberg S.M., Lee S.I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. 2017. №30. Pp. 4765–4774.

7. Wolpert D.H. Stacked generalization // Neural Networks. 1992. №5(2). Pp. 241–259.

8. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. 2001. №29(5). Pp. 1189–1232.