K-MEANS ASOSIDA KLASTERLASH ORQALI IMBALANCED MA'LUMOTLAR TO'PLAMINI BALANSLASH VA TIBBIY KLASSIFIKATSIYA UCHUN TAYYORLASH

Ass. ***Sariyev Shohruh Norqo'zi o'g'li***
Sh.Rashidov nomidagi Samarqand davlat universiteti, Samarqand, O'zbekiston
sariyevshokhrukh@gmail.com

Annotatsiya. *Tibbiyotda amaliy mashinaviy o'qitish modellarining samaradorligi ko'pincha ma'lumotlar to'plamining balansiga bog'liq. Qandli diabetni erta aniqlash vazifasida ko'plab ma'lumotlar to'plamlari imbalanced ya'ni sinflararo nomutanosiblikka ega bo'lib kam uchraydigan sinf prediabet to'g'ri klassifikatsiya qilinmaydi. Ushbu maqolada 0: sog'lom 20 000, 1: prediabet 4604, 2: diabet 10 000 kabi sinflardan iborat imbalanced tibbiy ma'lumotlar to'plami K-Means algoritmi yordamida klasterlash asosida balanslanib prediabet sinfi asos sifatida tanlanib har bir sinfdan 4604 ta namunani klasterlar bo'yicha tanlash orqali yangi balansli ma'lumotlar to'plami shakllantirildi. Yondashuv statistik vizual va amaliy jihatdan baholanib samaradorligi ko'rsatildi.*

Kalit so'zlar*: imbalanced data, K-Means, ma'lumotlarni balanslash, HistGB, LightGBM, AutoGluon.*

1. Kirish

Mashinaviy o'qitish tibbiyotda keng qo'llanilayotgan yo'nalishlardan biridir. Ayniqsa diabet va prediabetni erta bosqichda aniqlash xavf guruhiga kiruvchi bemorlarni skrining qilish orqali kasallikni oldini olish mumkin[1-2]. Biroq amaliy tibbiy ma'lumotlar to'plamlari ko'pincha imbalanced holatda bo'ladi. Sog'lom shaxslar soni ko'p prediabet esa kam uchraydi. Bu nomutanosiblik mashinaviy o'rganish algoritmlarida major sinfga yon bosish bias holatiga olib keladi va kam sonli ammo klinik ahamiyatli sinflar masalan prediabet noto'g'ri klassifikatsiya qilinadi. Bu esa F1-score va aniqlik kabi baholash mezonlarining

pasayishiga olib keladi. Muammoni hal qilish uchun an’anaviy oversampling va undersampling usullari mavjud biroq ular ma’lumotlarning tuzilmasini buzadi yoki overfittingga olib keladi. Shu sabab ushbu tadqiqotda K-Means klasterlashga asoslangan balanslash yondashuvi taklif etiladi. Bu yondashuv har bir sinfni alohida klasterlarga ajratib har bir klasterdan teng namunalar tanlash orqali tuzilmani saqlagan holda balansli to‘plam hosil qiladi[3-4]. Eksperimental qismda sinflar soni 0: sog‘lom 20000, 1: prediabet 4604, 2: diabet 10000 bo‘lgan ma’lumotlar to‘plami ishlatilgan va har bir sinfdan 4604 ta namunaga tenglashtirilgan yangi ma’lumotlar to‘plami tayyorlangan.

2. Asosiy qism

Ma’lumotlar to‘plamining umumiy tavsifi. Tahlil uchun ishlatilgan ma’lumotlar to‘plami quyidagi sinflarni o‘z ichiga oladi.

Klass	Turi	Namuna soni
0	Sog‘lom	20,000
1	Prediabet	4,604
2	Diabet	10,000

Ko‘rinib turibdiki 1-sinf prediabet namunalari eng kam sonli bo‘lib ushbu sinfga nisbatan ML modellar sezgir emas. Shuning uchun prediabet sinfi asos sifatida tanlandi.

K-Means asosidagi balanslash algoritmi. Yangi balanslangan to‘plam yaratish quyidagi bosqichlarda amalga oshirildi

1. Klasterlash

- Har bir sinf alohida holda **K-Means** yordamida **K=10** ga ajratildi.
- Shunday qilib, 0, 1, va 2 sinflar har biri 10 ta klasterdan iborat guruhlarga ajratildi.

2. Klaster ichidan tanlash

- Har bir klaster ichidan teng miqdorda namunalar tanlab olindi.
- Prediabet sinfida har bir klasterdan taxminan 460 ta namunani tashkil qiladi ($10 \times 460 \approx 4600$).

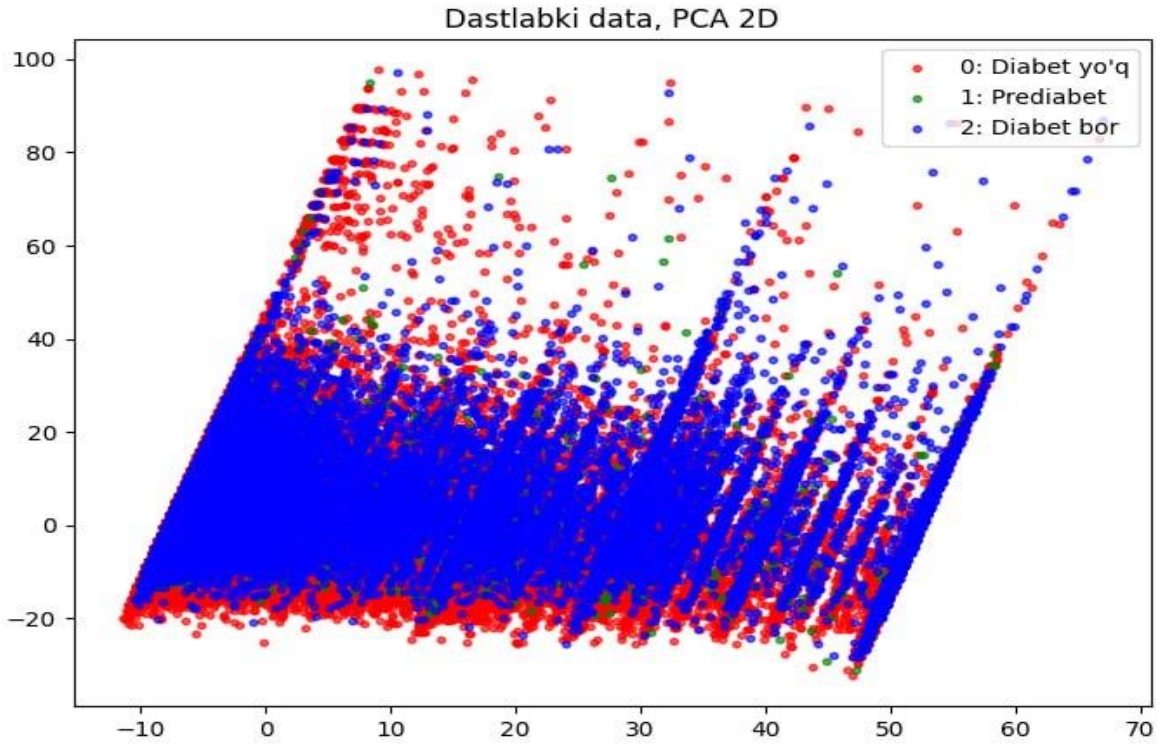
3. Boshqa sinflarni balanslash

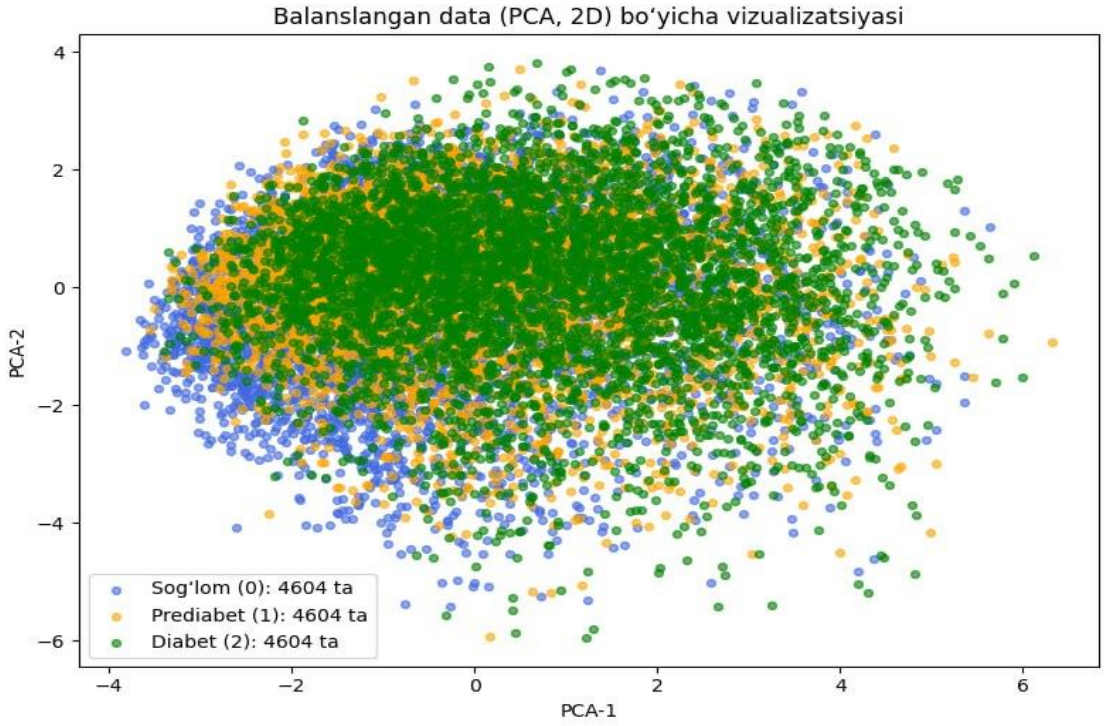
- Sog‘lom (0) va diabet (2) sinflaridan ham aynan har bir klasterdan 460 ta namunani tanlab olindi (10×460 = 4600).

4. Yakuniy balanslangan to‘plam. Quyidagi jadvalda balanslangan ma’lumotlar to‘plami hosil qilindi.

Klass	Namuna soni
0	4604
1	4604
2	4604

Bu yondashuv tuzilmani buzmasdan sinflar o‘rtasida tenglik yaratildi va sinflararo vakillikni ta’minladi[5-6]. Quyidagi rasmlardan ham ko‘rish mumkin.





Exspriment natijalar

Balanslashdan oldin ML modellardan **LighGBM**, **HistGB** va **AutoGluon** modellarida quyidagi natijalarga erishildi.

Balanslashdan keyin esa quyidagi natijalar qayd etildi.

Model	Auccarcy	Precision	Recall	F1-score
LighGBM	0.9055	0.9154	0.9055	0.9042
HistGB	0.9055	0.9151	0.9055	0.9042
AutoGluon	0.9066	0.9160	0.9066	0.9053

Natijalar shuni ko‘rsatadiki K-Means asosida balanslangan to‘plam sezilarli darajada prediabet sinfining aniqligini oshirgan.

Munozara

K-Means asosidagi balanslash usuli quyidagi ustunliklarni ko‘rsatdi. Klasterlar o‘rtasida strukturani saqladi: bu modelning turli qiyofadagi namunalarni o‘rganishiga yordam berdi. Oversampling emas — ya’ni ma’lumotlar

nusxalanmadi. Haqiqiy klasterlar ichidan tanlandi. Kam sinfning vakilligi oshdi, prediabet holatlarini aniqlash darajasi sezilarli yaxshilandi. Shu bilan birga bu yondashuv quyidagi cheklovlarga ega[7-8]. K-Means klaster soni K tanlash subyektivdir. Klasterlar notekis bo‘lsa kichik klasterdan tanlashda muammolar bo‘lishi mumkin. Kompyuter resurslari nisbatan ko‘proq talab qilinadi 10×3 klasterlarni ajratish. Shunga qaramay bu yondashuv tibbiy klassifikatsiyada amaliy jihatdan juda foydali bo‘lib, ayniqsa imbalanced tibbiy holatlarda erta bosqichdagi kasalliklar qo‘llash mumkin.

Xulosa

Imbalanced ma’lumotlar bilan ishlashda K-Means asosida klasterlash orqali balanslash — bu amaliy strukturaviy va klinik nuqtai nazardan maqbul yondashuv hisoblanadi. Ushbu usul tibbiy klassifikatsiyada kam sonli sinflarning prediabet kabi aniqligini sezilarli darajada oshiradi.

Model F1-score Recall kabi muhim ko‘rsatkichlarda yaxshilanishga erishgan bo‘lib, bu tibbiy qaror qabul qilishda ayniqsa muhim hisoblanadi[9-11]. Yondashuvni boshqa klasterlash algoritmlari bilan solishtirish, shuningdek, real vaqt monitoringda sinab ko‘rish bo‘yicha keyingi izlanishlar olib borilishi mumkin.

FOYDALANILGAN ADABIYOTLAR

1. Akmal Akhatov, Fayzullo Nazarov, Mekhriddin Nurmamatov, Shokhrukh Sariyev. (2024). Genetic algorithm application technology in multi-parameter optimization problems AIP Conf. Proc. 3244, 030025. https://doi.org/10.1063/5.0242074.

2. Alkalifah, B., Shaheen, M. T., Alotibi, J., Alsubait, T., & Alhakami, H. (2025). Evaluation of machine learning-based regression techniques for prediction of diabetes levels fluctuations. Heliyon, 11(1).

3. Nazarov, F., Nurmamatov, M., & Sariyev, Sh. (2024). ma’lumotlarni intellektual tahlil qilish uchun genetik algoritmlar va ularni qo‘llanilishi. digital transformation and artificial intelligence, 2(6), 162–168. retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i630

4. Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P., & Pilakoutas, K. (2021). Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. Cement and Concrete Research, 145, Article 106449.

5. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE TKDE*, 21(9), 1263–1284.

6. Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *JAIR*, 16, 321–357.

7. Liu, X. Y., et al. (2009). Exploratory undersampling for class-imbalance learning. *TKDE*, 21(9), 129–143.

8. Han, H., et al. (2005). Borderline-SMOTE. *PAKDD*, 878–887.

9. M. Nurmamatov, S. Sariyev and I. Uddin, "Methods of Using Artificial Intelligence Algorithms in Human Resource Management," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 566-571, doi: 10.1109/SmartIndustryCon65166.2025.10986087

10. N. Fayzullo, S. Sariyev and Y. Sherzodjon, "Analyzing the Effectiveness of Ensemble Methods in Solving Multi-Class Classification Problems," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 788-793, doi: 10.1109/SmartIndustryCon65166.2025.10986248

11. Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in AI*, 5(4), 221–232.