UO‘K 681.4

MASHINAVIY O‘QITISHNING ANSAMBLLARI ASOSIDA QANDLI DIABETNI TASHXISLASH ALGORITMLARINI ISHLAB CHIQISH

Sariyev Sh.N.[1]

[1] Sharof rashidov nomidagi Samarqand davlat universiteti, Samarqand, O‘zbekiston

[+] sariyevshokhrukh@gmail.com

Annotatsiya. Ushbu tadqiqotda qandli diabet kasalligini tashxislashda mashinaviy o‘qitish algoritmlarining ansambl modellaridan LightGBM, AdaBoost va HGB qo‘llanilishi uchun algoritmlari ishlab chiqildi. Tadqiqotda ma’lumotlar to‘plamidagi yuqori korrelyatsiyaga ega atributlar tanlab olinib modellar o‘quv va test to‘plamlariga bo‘linib sinovdan o‘tkazildi. Har bir model uchun aniqlik accuracy, precision, recall, F1-o‘lchovlari hamda AUC-ROC ko‘rsatkichlari hisoblab chiqildi. Tajriba natijalaridan uchala modeldan HGB va LightGBM yuqori samaradorlikka ega ekanligi aniqlandi. Ushbu modellarning murakkab klinik ma’lumotlarni tahlil qilishda yuqori ustunligini ko‘rsatadi. Ushbu tadqiqot qandli diabetni erta tashxislashda mashinaviy o‘qitish algoritmlarining samarali vosita ekanligini ko‘rsatadi. Sog‘liqni saqlash sohasida avtomatlashtirilgan tashxislash tizimlarini rivojlantirish uchun muhim ilmiy asos bo‘ladi.

Kalit so‘zlar: AdaBoost, LightGBM, HGB, qandli diabet.

1 KIRISH

Hozirgi kunning dolzarb muammolaridan mashinaviy o‘qitish algoritmlarining tibbiyot sohasida keng qo‘llanilishi sezilarli darajada kun sayin oshib bormoqda. Mashinaviy o‘qitish va sun’iy intellekt algoritmlari tibbiyotda tashxis, davolash, kasalliklarni oldini olish va bemorlarning ma’lumotlarini monitoring qilish kabi jarayonlarni avtomatlashtirish va takomillashtirishda o‘zining katta xissasini qo‘shmoqda. Ayniqsa tibbiyot sohasida erishilayotgan yutuqlar bilan birga ushbu texnologiyalar sog‘liqni saqlash tizimining dolzarb muammolariga samarali yechimlar taklif qilmoqda. Dunyo bo‘ylab qandli diabetning tez tarqalishi va uning asoratlari sog‘liqni saqlash tizimlariga katta zararlar yetkazmoqda. Mashinaviy o‘qitish algoritmlari diabetni erta bosqichda aniqlashda o‘zining samarasini bermoqda. Bemorlarda kasalliklarning rivojlanishini kuzatish va individual davolash rejalarini ishlab, chiqishda muhim rol o‘ynamoqda. Ushbu tadqiqotda qandli diabet kasalligini tashxislash uchun mashinaviy o‘qitish algoritmlari bilan tashxishlash taklif etildi. Qandli diabetni erta tashxislash bemorlarning sog‘lig‘ini saqlash va asoratlarini kamaytirishda hal qiluvchi samarali usul hisoblanadi [1]. Shu sababli sog‘liqni saqlash sohasida zamonaviy texnologiyalarni qo‘llash xususan sun’iy intellekt va mashinaviy o‘qitish metodlaridan foydalanish orqali diabetni erta va yuqori aniqlikda aniqlash yo‘llari izlanmoqda.

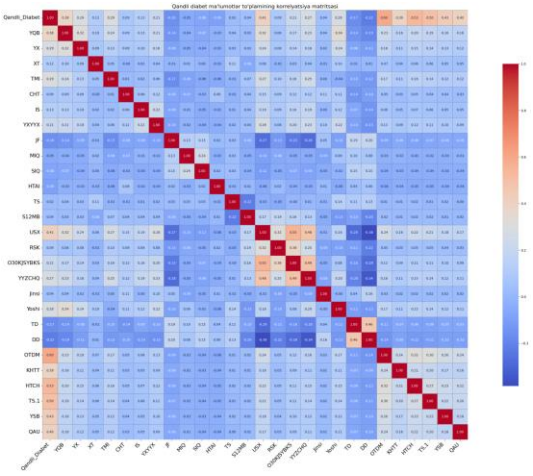
Mashinaviy o‘qitish algoritmlaridan ansambl usullari murakkab tibbiy ma’lumotlar to‘plamini tahlil qilish va klassifikatsiya qilishda yuqori samaradorligi bilan boshqa algoritmlardan ajralib turadi. Ushbu yondashuvlar mashinaviy o‘qitishning oddiy modellarga nisbatan ko‘proq aniqroq va barqarorlikni ta’minlaydi. Ushbu mashxur ansambl modellardan AdaBoost ansambli zaif klassifikatorlar ketma-ket o‘rgatilishi va ularning natijalarini vaznlab kuchli ansambl yaratish printsipiga asoslanib qandli diabet kasalligini tashxislashda noaniq va shovqinli ma’lumotlarga tez moslashadi. LightGBM ansambli gradient boosting tamoyillari asosida samaradorlik va tezlikni oshirgan holda katta o‘lchamdagi va yuqori o‘lchovli tibbiy ma’lumotlar to‘plamini qayta ishlashga mo‘ljallangan. HGB algoritmi kategorik o‘zgaruvchilar bilan samarali ishlash imkoniyatiga ega hisoblanadi. Tibbiy ma’lumotlar to‘plami ichida keng tarqalgan kategorik atributlarni optimallashtirishda yuqoir samaradorlikka ega. Ushbu tadqiqotda LightGBM, AdaBoost va HGB nsambllari asosida qandli diabet kasalligini tashxislash uchun mashinaviy o‘qitish modellarini ishlab chiqish va ularning tahlili amalga oshiriladi. Qandli diabetni tashxislashdagi samaradorlikni o‘rganish ansambl modellarning parametrlari optimallashtirilishi va amaliy sog‘liqni saqlash tizimlarida qo‘llanish imkoniyatlarini aniqlashdan iboratdir [2-3].

2 ASOSIY QISM

Mashinaviy o‘qitish algoritmlari uchun qandli diabet kasalligining ma’lumotlar to‘plamida mavjud bo‘lgan ko‘plab klinik belgilarning ko‘rsatkichlari orasida samarali va to‘g‘ri tashxislash uchun muhim

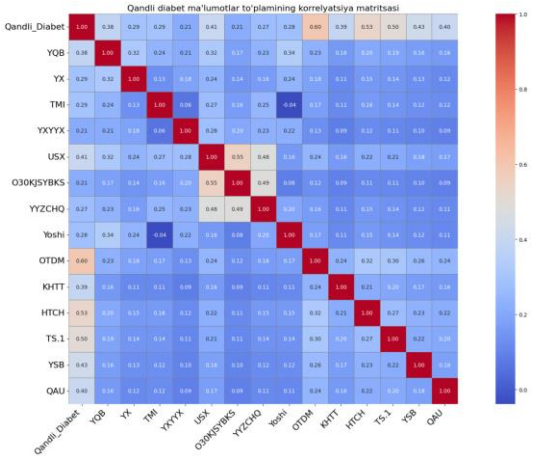
ahamiyatga ega bo‘lgan atributlarni aniqlash zarur bo‘ladi. Ushbu ma’lumotlar to‘plamidagi xususiyatlarni 1-rasmda Pearsin korrelyatsiya koeffitsiyentlarini hisoblash orqali o‘zaro kuchli bog‘langan va tashxisga eng katta ta’sir ko‘rsatadigan ustunlarni tanlab olish amaliyoti qo‘llaniladi.

Ma’lumotlar to‘plamini tahlil qilish va tozalash bosqichida ma’lumotlarni tayyorlash jarayonining muhim qismi hisoblanadi. Modellarni qurishda ortiqcha parametrlar bilan bog‘liq muammolarni kamaytiradi, va hisoblash vaqtlarini tejaydi. Bundan tashqari muhim atributlar orasidagi korrelyatsiyani chuqur tahlil qilish orqali muhim xususiyatlarini aniqlashda algoritmlarning o‘rganish tezligi va barqarorligini oshiradi.



1-rasm. Pearsin korrolyatsiya matritsasining natijasi dastlabki holatda

Qandli diabet tashxislashda 2-rasmda yuqori korrelyatsiyaga ega atributlarning tanlanishi natijasida, mashinaviy o‘qitish modellarining aniqroq va samaraliroq ishlashiga erishiladi, shuningdek, klinik ma’lumotlardan to‘g‘ri xulosa chiqarish imkoniyati kengayadi.



2-rasm. Korrolyatsiyasi 20 % dan past maydonlar tashlab yuborilgan holati

AdaBoost ansambl metod bo‘lib bir nechta zaif klassifikatorlarni ketma-ket o‘rgatadi va ularni ma’lum og‘irliklar asosida birlashtirib, kuchli klassifikator hosil qilishdan iboratdir. AdaBoost ansamblini qandli diabet kasalligini tashxislashda qo‘llash quyidagi matematik ifodalar orqali amalga oshiriladi. Berilgan ma’lumotlar to‘plamini quyidagicha kiritiladi:

$$\{(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)\}, \quad y_i \in \{-1; +1\},$$

bunda $x_i \in R^d$, i – bemorning d o‘lchovdagi sog‘liq ko‘rsatkichlari $y_i \in \{-1, +1\}$, i – namunani diabet kasalligini tashxisi bor bo‘lsa +1 yoki yo‘qligi -1 bildiruvchi yorliq. Maqsad: kuchli klassifikator $H(x)$ yaratish, ya’ni har qanday yangi bemorning kasallik holatini yuqori aniqlikda tashxislash.

Boshlang‘ich vaznlar taqsimoti. Dastlab har bir olinayotgan namunaga quyidagicha teng vazn beriladi:

$$w_t(i) = \frac{1}{N}, \quad i = 1,..., N,$$

bu yerda: $w_t(i)$ dagi $t-$iteratsiyadagi $i-$namunaga berilgan vazn, $t = 1.2...T$ iteratsiyalar soni. Bu vaznlar namunalar tanlanishining ehtimolini ifodalaydi. Dastlab barcha namunalarga teng vazn beriladi. Zaif klassifikatorni o‘rgatish. Vaznlangan trening to‘plami asosida zaif klassifikator iteratsiyaning har bir qadamida vaznlar taqsimoti asosida zaif klassifikator $h_t(x_i)$ o‘qitiladi.

$$h_t(x_i) : R^d \to \{-1, +1\}.$$

Klassifikatorning xatosini hisoblashda zaif klassifikatorning vaznlangan xatosi (1) kabi hisoblanadi:

$$\varepsilon_t = \sum_{i=1}^{N} w_t(i) \cdot I[h_t(x_i) \neq y_i], \tag{1}$$

$$\mathrm{I} = \begin{cases} h_t(x_i) \neq y_i, bo'lsa \;\; 1, \\ aks\, holda \;\; 0, \end{cases}$$

bu yerda: I – indikator funksiyasi, $h_t(x_i) \neq y_i$ bo‘lsa 1, aks holda 0.

Klassifikatorning ishonchliligini baholash. Klassifikatorning ishonchlilik koeffitsienti α_t quyidagicha hisoblanadi:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right).$$

Bu qiymat kichik xatoga ega bo‘lgan klassifikatorga katta vazn beradi. Vaznlarning yangilanishi (2) formula orqali vaznlari yangilanadi:

$$w_{t+1}(i) = \frac{w_t(i)}{Z_t} \exp(-\alpha_t y_i h_t(x_i)), \tag{2}$$

bu yerda: Z_t – normallashtirish koeffitsienti bo‘lib (3) orqali amalga oshiriladi. vaznlarning yig‘indisini $\sum_{i=1}^{N} w_{t+1}(i) = 1$ ga tenglashtiradi.

$$Z_t = \sum_{i=1}^{N} w_t(i) \exp(-\alpha_t y_t(x_i)). \tag{3}$$

Yangilanish prinsipi to‘g‘ri tasniflangan namunalarning vazni kamayadi. Noto‘g‘ri tasniflanganlarining vazni oshadi. Shu bilan AdaBoost model noto‘g‘ri tasniflangan namunalarni keyingi bosqichlarda yaxshiroq o‘rganishga harakat qiladi.

Yakuniy kuchli klassifikatorni qurishda T barcha iteratsiyalardan so‘ng kuchli klassifikator quyidagi formula orqali hisoblanadi:

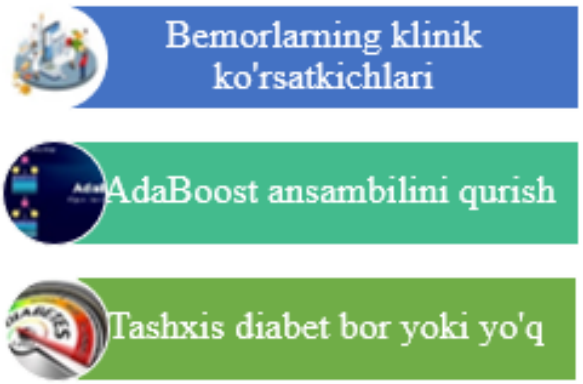
$$H(x) = sign\left(\sum_{t=1}^{T} a_t h_t(x)\right), \tag{4}$$

bu yerda: $x-$yangi namunani ifodalovchi xususiyatlar vektori(diabet tashxisini aniqlamoqchi bo‘lgan bemorning klinik belgilari); $h_t(x)$, $t-$zaif klassifikatorning x namunaga bo‘lgan bashorati; Zaif klassifikator $h_t(x) \in \{-1, +1\}$ da bo‘ladi. Ya’ni diabet yo‘q yoki bor degan qaror qabuladi; $H(x)-$barcha zaif klassifikatorlarning og‘irliklangan ovozlari yig‘indisining belgisi -1 diabet yo‘q yoki +1 diabet bor sifatida aniqlanadi.

AdaBoost ansamblini diabet tashxisiga moslashtirish quyidagicha amalga oshiriladi. AdaBoost yordamida yuqori xatolik darajasiga ega bo‘lgan parametrlar ko‘proq e’tibor oladinadi. Shuning uchun ham model qandli diabet tashxisida yuqori sezgirlikka ega bo‘ladi. 3-rasmda amalga oshirish bosqichi keltirilgan.

AdaBoost ansambl algoritmi qandli diabet kasalligini tashxislashda kuchli va samarali algoritm hisoblanadi. AdaBoost algoritmining o‘ziga xos jihati sifatida vaznlarning iterativ yangilanishi mavjud

bo‘lib bu jarayon zaif klassifikatorlarning xatolariga qarshi moslashuvchan o‘qitishni ta’minlaydi. Model tibbiy ma’lumotlardagi murakkabliklarni samarali o‘rganib tashxislash aniqligini sezilarli darajada oshiradi [4-5].



3-rasm. Qandli diabet kasalligini AdaBoost orqali tashxislash

LightGBM – bu gradient boosting algoritmlarining yengil, tez va samarali versiyasi bo‘lib, asosan yirik ma’lumotlar to‘plamlarida ishlash uchun optimallashtirilgan. U qaror daraxtlarini ketma-ket qurish orqali modellar ansamblini hosil qiladi.

Trening to‘plami quyidagicha belgilanadi:

$$D = \{(x_i, y_i)\}, i = 1.2...N,$$

$x_i \in R^d$, i – namunadagi xususiyatlar vektori, $y_i \in \{0,1\}$, i – namunadagi maqsad funksiyasi 0 diabet yo‘q 1 diabet bor holati. Maqsad funksiya $F(x)$ ni topish u ma’lumotga eng yaxshi mos tushadi, va yangi x uchun to‘g‘ri prognoz beradi. LightGBM gradient boosting algoritmiga asoslanadi. Gradient boosting modellarning ketma-ket qo‘shilishiga asoslangan ansambl usuli bo‘lib har bir yangi model oldingi modellarning xatolarini kamaytirishdan iborat. Maqsad funksiyasi quyidagicha hisoblanadi:

$$F(x) = \sum_{m=1}^{M} f_m(x), \tag{5}$$

bu yerda M – daraxtlar soni iteratsiyalar soni; $f_m(x)$, m – qaror daraxti modeli; $F(x)$ – yakuniy kuchli model. Sigmoid funksiyasini qo‘llash quyidagi formula orqali amalga oshiriladi:

$$\hat{y} = \sigma\left(F\left(x\right)\right) = \frac{1}{1+e^{-F(x)}}, \tag{6}$$

$\hat{y}$ – qandli diabet ehtimoli. Yo‘qotish funksiyasi. Modelni o‘rgatishda LGBM quyida keltirilgan formulada yo‘qotish funksiyasini minimallashtirishga xarakat qiladi

$$L = -\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{7}$$

bu yerda: $y_i \in \{0,1\}$, $\hat{y}_i = \sigma\left(F\left(x_i\right)\right)$ bashorat qilingan modelning ikkilik qiymati.

Gradient asosida qo‘shimcha daraxt qurish quyidagicha amalga oshiriladi. m – chi iteratsiyada yangi daraxt $f_m(x)$ oldingi modelni yangilaydi [6-7]

$$F_m(x) = F_{m-1}(x) + \eta f_m(x).$$

$F_{m-1}(x)$ – oldingi iteratsiyalardagi modellar yig‘indisi, η – o‘rganish tezligi kichik qiymat modelni sekinroq o‘rganishga va ortiqcha moslashishning oldini olishga yordam beradi. $f_m(x)$ – yangi qurilayotgan daraxt.

LightGBM modelni optimallashtirish Gradient va Hessian uchun yo‘qotish funksiyasini ikki tartibli Taylor qatori yordamida quyidagicha aniqlanadi:

$$L^{(m)} \approx \sum_{i=1}^{N}\left[l(y_i, F_{m-1}(x_i)) + g_i f_m(x_i) + \frac{1}{2} h_i f_m(x_i)^2 \right], \quad (8)$$

bu yerda $g_i = \frac{\partial L}{\partial F(x_i)} = \hat{y}_i - y_i$, yo‘qotish funksiyasining birinchi tartibli gradienti(birinchi tartibli hosila);

$h_i = \frac{\partial^2 L}{\partial F(x_i)^2} = \hat{y}_i\left(1-\hat{y}_i\right)$, yo‘qotish funksiyasining ikkinchi tartibli hosilasi.

Bu yondashuv modelni tez va aniq o‘rgatish imkonini beradi. Daraxt tuzilishini optimallashtirishda LightGBM daraxtlarini qurishda har bir tugun uchun quyidagi funksiyani minimal qilishga harakat qiladi:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} \right] - \gamma, \quad (9)$$

bu yerda I – joriy tugundagi namunalarning indeksi; I_L, I_R – chap va o‘ng bolalar tugunlaridagi indekslar; g_i, h_i – yuqoridagi gradientlar; λ – L2 regulyarizatsiya parametri; γ – daraxtning murakkabligini penalizatsiya qiluvchi parametri.

Bu funksiya daraxtda tugunlarni bo‘lish orqali maksimal yo‘qotish kamayishini ta’minlashga qaratilgan. $F(x)$ – yakuniy kuchli model, turli qaror daraxtlari yig‘indisi; $f_m(x)$ – har bir iteratsiyada o‘qitiladigan zaif model (daraxt); η – o‘rganish tezligi, modellar og‘irligini sozlaydi; L – yo‘qotish funksiyasi, modelning xatosini o‘lchaydi; g_i, h_i – birinchi va ikkinchi tartibli hosilalar, yo‘qotishni optimallashtirish uchun hisoblanadi; λ, γ – regulyarizatsiya parametrlari, ortiqcha moslashishni oldini olishga yordam beradi; L_{split} – daraxtning har bir tugunini bo‘lishda minimal qilishga harakat qilinadigan qiymat.

Hist Gradient Boosting (HGB) ansambli Gradient Boosting Decision Tree algoritmining optimallashtirilgan varianti bo‘lib har bir xususiyatlar bo‘yicha butun qiymatlar oralig‘ini oldindan ma’lum bo‘lgan kichik oraliqlarga ya’ni bin-histogram segmentlarga ajratadi va split bo‘linishlarni aynan ushbu histogram binlari bo‘yicha qidiradi. Model har bir xususiyatning aniq qiymatlari o‘rniga ularni diskret bo‘laklarga ajratilgan histogram binlar sifatida ko‘radi. Split uchun optimal nuqta topish jarayoni tezlashadi va xotira sarfini kamaytiradi. Bunday yondashuv millionlab namunalarda va ko‘plab xususiyatlarda ham samarali hisob-kitob qilish imkonini beradi. HGB klassik gradient boostingning tezlashtirilgan implementatsiyasi bo‘lib har bir ustunidagi qiymatlar sonini oldindan belgilangan binlar sonigacha diskretlashtiradi va gradient va hessian qiymatlarini aynan histogram segmentlari bo‘yicha yig‘ib splitlarni qidiradi. Bundan klassik “pre-sorted” split qidirish algoritmiga nisbatan hisoblash murakkabligini va resurs sarfini sezilarli kamaytiradigan [8-10]:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \dots x_{1p} \\ x_{21} & x_{22} & x_{23} \dots x_{2p} \\ \vdots & \vdots & \vdots \quad \vdots \\ x_{N1} & x_{N2} & x_{N3} \dots x_{Np} \end{pmatrix}, \quad (10)$$

bu yerda N – namunalar soni(qatorlar soni), p – xususiyatlar soni(ustunlar soni).

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix},$$

bu yerda y – target qiymatlari.

Multi klass uchun yo‘qotish funksiyasi quyidagi formula orqali ifodalanadi:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \ln P_{i,y_i}, \quad (11)$$

bu yerda L – umumiy yo‘qotish loss funksiyasi, i - namuna indeksi $i = 1,...,N$, y_i – namunadagi haqiqiy klass 0 va 1, P_{i,y_i} – i – namunaga model tomonidan berilgan haqiqiy klass uchun hisoblangan ehtimoli quyidagicha topiladi.

$$P_{i,y_i} = \frac{e^{F_{y_i}(x_i)}}{\sum_{k=0}^{K-1} e^{F_k(x_i)}},$$

bu yerda $F_k(x_i)$ – bu x_i namuna uchun k – klassning model chiqishi. Boshlang‘ich qiymat quyidagicha hisoblanadi:

$$F_k^{(0)}(x) = \ln\left(\frac{n_k}{N}\right),$$

bu yerda $F_k^0(x)$, k – sinf uchun boshlang‘ich qiymat(har bir klass uchun alohida), n_k – ma’lumotlar to‘plamida $y = k$ bo‘lgan namunalar soni.

Har bir iteratsiyadagi model yangilanishi quyidagi formula orqali yangilanadi:

$$F_k^{(m)}(x) = F_k^{(m-1)}(x) + \lambda \cdot h_k^{(m)}(x), \tag{12}$$

bu yerda $F_k^{(m)}(x)$, m – iteratsiyadagi k sinf uchun modelning chiqish qiymati, $m-$ iteratsiya daraxt raqami $m = 1,...,M$, $\lambda-$ learning rate(o‘rganish tezligi), $0 < \lambda \le 1$, $h_k^{(m)}(x)$, $k-$sinf uchun $m-$iteratsiyada o‘rgatilgan yangi daraxtning chiqishi.

Gradientni va hessiyani hisoblash, har bir namuna i va har bir klass k uchun quyidagicha hisoblanadi:

$$\begin{cases} g_{ik} = P_{ik} - \mathrm{I}(y_i = k), \\ h_{ik} = P_{ik}(1 - P_{ik}), \end{cases}$$

bu yerda $g_{ik}-$ model bu orqali keyingi daraxt uchun targetni aniqlaydi, $h_{ik}-$ hessian ikkinchi tartibli hosila Nyuton bo‘yicha optimizatsiyada ishlatiladi, P_{ik}, x_i – namunasi uchun model hisoblangan $k-$sinf ehtimoli. P_{ik} quyidagicha hisoblanadi:

$$P_{ik} = \frac{\exp(F_k(x_i))}{\sum_{s=1}^{K} \exp(F_s(x_i))}.$$

Yakuniy chiquvchi qiymat funksiyasi quyidagicha ifodalanadi:

$$\hat{y}_i = \arg\max_{k \in \{0,1\}} \arg\max F_k^{(M)}(x_i),$$

bu yerda $\hat{y}_i-$ model tomonidan bashorat qilingan klass, arg max – maksimal qiymatga mos indexni tanlaydigan operator klassning indeksini qaytaradi.

3 TAJRIBA NATIJALARI

Ushbu tadqiqot jarayonida qandli diabet kasalligini tashxislashda LightGBM, AdaBoost va HGB ansambl algoritmlari qo‘llanildi. Ushbu ansamblarning samaradorligi korrolyatsiya koeffitsiyentlari 20% dan yuqori bo‘lgan maydonlar yordamida baholandi. Ma’lumotlar to‘plami test va o‘quv to‘plamlariga ajratilib, har bir model test to‘plamidagi namunalar bo‘yicha sinovdan o‘tkazildi. Accuracy, precision, recall, F1-score va AUC-ROC ko‘rsatkichlari asosida olingan natijalar 1-jadval va 4-rasmda keltirilgan. Ushbu tadqiqotda tajriba natijalari AdaBoost, LightGBM va HGB modellarining qandli diabet kasalligini tashxislashda yuqori samaradorlikka ega ekanligini ko‘rish mumkin. Shu bilan birga klinik sharoitlarda

avtomatlashtirilgan diagnostika tizimlarini yaratishda mashinaviy o‘qitish algoritmlari ishonchli vositalar ekanligini tasdiqlaydi.

1-jadval. Mashinaviy o‘qitishning ansambl natijalari

	Model	Acc	Prec	Rec	F1-Score	AUC
1	HGB	0.9268	0.9267	0.9267	0.9267	0.981
2	LightGBM	0.9249	0.9232	0.9269	0.9251	0.978
3	AdaBoost	0.9244	0.9221	0.9273	0.9247	0.978

4 XULOSA

Ushbu olib borilgan tadqiqotda qandli diabet kasalligini tashxislash uchun LightGBM, AdaBoost va HGB ansambl algoritmlari sinovdan o‘tkazildi. Ushbu ansambl modellardan AdaBoost va LightGBM natijalari yuqori aniqlikka erishildi. Ma’lumotlar to‘plamida yuqori aniqlik va ishonchli bashorat natijalarini ko‘rsatadi. Accuracy, precision, recall, F1-score va AUC-ROC ko‘rsatkichlari asosida barcha modellarning samaradorligi o‘rganildi.Xulosa qilib aytganda HGB va LightGBM modellarining yuqori aniqlik va AUC-ROC ko‘rsatkichlari qandli diabetni tashxislash mumkin ekanligini ko‘rsatdi. Kelgusida tadqiqotlarda ushbu modellarni yanada rivojlantirish va giper parametrlarini optimallashtirishga qaratiladi.

ADABIYOTLAR

[1] *Nazarov F., Sariyev Sh., Yarmatov Sh.*, “Analyzing the Effectiveness of Ensemble Methods in Solving Multi-Class Classification Problems”, International Russian Smart Industry Conference (SmartIndustryCon), 2025, doi.org/10.1109/SmartIndustryCon65166.2025.10986248

[2] *Javed, A., Jalil, Z., Moqurrab, S., Abbas, S., Liu, X.*, 2020. Ensemble Adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. Trans. Emerg. Telecommun. Technol.https://doi.org/10.1002/ett.4088.

[3] *Nurmamatov M., Sariyev Sh., Izhar Uddin*, “Methods of Using Artificial Intelligence Algorithms in Human Resource Management”, International Russian Smart Industry Conference (SmartIndustryCon), 2025, doi.org/10.1109/SmartIndustryCon65166.2025.10986087

[4] *R¨atsch, G., Onoda, T., & Müller, K.-R.* (2001). Soft margins for AdaBoost. Machine Learning, 42(3), 287–320. https://doi.org/10.1023/A:1007618119488.

[5] *AllenZhu, Z., Li, Y.* (2020). Towards understanding ensemble, knowledge distillation and self-distillation in deep Learning. https://doi.org/10.48550/arXiv.2012.09816

[6] *Ke G, Meng Q, Finley T, Wang T, Chen W, Mal W, Yel Q, Liul T-Y.* LightGBM: A highly efficient gradient boosting decision tree. In: 31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA. 2017.

[7] *Nazarov, F., Nurmamatov, M., & Sariyev, S.* (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.

[8] *Ke, Guolin, et al.* “LightGBM a Highly Efficient Gradient Boosting Decision Tree” Advances in Neural Information Processing Systems (NeurIPS), 2017.

[9] *F. Taromideh, R. Fazloula, B. Choubin, M. Masoodi and A. Mosavi*, "Ensemble Machine Learning for Urban Flood Hazard Assessment," 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI), Stará Lesná, Slovakia, 2024, pp. 000525-000530, doi: 10.1109/SAMI60510.2024.10432902.

[10] *A. N. Nikhil, S. K. Thalanki, G. Rajaram and B. Renganathan*, "Enhancing the Accuracy in Food Image Recognition Using Recurrent Neural Network Model in Comparison With Graph Neural Network Model," 2024 9th International Conference on Applying New Technology in Green Buildings (ATiGB), Danang, Vietnam, 2024, pp. 507-511, doi: 10.1109/ATiGB63471.2024.10717820.

Поступила в редакцию 22.03.2025

Citation: *Sariyev Sh.N.* (2025). Mashinaviy o‘qitishning ansambllari asosida qandli diabetni tashxislash algoritmlarini ishlab chiqish. Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. 8(2). – B. 87-94. https://doi.org/10.62132/ijdt.v8i2.267.

DEVELOPMENT OF DIABETES DIAGNOSIS ALGORITHMS BASED ON MACHINE LEARNING ENSEMBLES

Sariyev Sh.N.[1]

[1] Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan

sariyevshokhrukh@gmail.com

Abstract. In this study, algorithms were developed for applying ensemble models of machine learning algorithms - LightGBM, AdaBoost, and HGB - in the diagnosis of diabetes mellitus. In the study, attributes with high correlation were selected from the dataset, and the models were tested by dividing them into training and test sets. For each model, accuracy, precision, recall, F1-scores, and AUC-ROC indicators were calculated. The results of the experiment revealed that HGB and LightGBM demonstrated higher efficiency among the three models. These models demonstrate a high advantage in the analysis of complex clinical data. This research demonstrates that machine learning algorithms are an effective tool for the early diagnosis of diabetes mellitus. It will serve as an important scientific foundation for the development of automated diagnostic systems in the healthcare sector.

Keywords: AdaBoost, LightGBM, HGB, diabetes.

РАЗРАБОТКА АЛГОРИТМОВ ДИАГНОСТИКИ ДИАБЕТА НА ОСНОВЕ АНСАМБЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Сариев Ш.Н.[1]

[1] Самаркандский государственный университет имени Шарофа Рашидова, Самарканд, Узбекистан

sariyevshokhrukh@gmail.com

Аннотация. В данном исследовании были разработаны алгоритмы для применения ансамблевых моделей алгоритмов машинного обучения LightGBM, AdaBoost и HGB - в диагностике сахарного диабета. В ходе исследования из набора данных были отобраны атрибуты с высокой корреляцией, а модели были протестированы путем разделения данных на обучающую и тестовую выборки. Для каждой модели были рассчитаны показатели точности, полноты, F1-меры и AUC-ROC. Результаты эксперимента показали, что HGB и LightGBM показали более высокую эффективность среди трех моделей. Эти модели демонстрируют значительное преимущество при анализе сложных клинических данных. Данное исследование показывает, что алгоритмы машинного обучения являются эффективным инструментом для ранней диагностики сахарного диабета. Это послужит важной научной основой для разработки автоматизированных диагностических систем в сфере здравоохранения.

Ключевые слова: AdaBoost, LightGBM, HGB, диабет.