PIMA өгөгдлийн багцад суурилан Random Forest алгоритмаар чихрийн шижингийн төлөвийг урьдчилан таамаглах нь

Ч.Баасандорж[1], Б.Одгэрэл[2], Л.Ажнай[1], Ц.Жавзмаа[1], У.Нямдаваа[1], Л.Анхбаяр[1], С.Оюунцэцэг[3], Э.Энх-Үрэл[1], Л.Пүрэвдолгор[1]*

[1]АШУҮИС, Био-АС, Физик мэдээлэл зүйн эрдмийн тэнхим
[2]АШУҮИС, СС, Ерөнхий эрдмийн тэнхим
[3]АШУҮИС, Био-АС, Харилцааны ур чадварын тэнхим
Цахим шуудан: purevdolgor@mnums.edu.mn, Утас: 95091107

Түлхүүр үг:

PIMA өгөгдлийн багц
Машин сургалт
Ангиллын загвар
ROC-AUC
Эрсдэлийн үнэлгээ

Товч утга:

Үндэслэл: Чихрийн шижин нь дэлхий даяар тархалт нэмэгдэж буй архаг бодисын солилцооны эмгэг бөгөөд эрт илрүүлэлт, эрсдэлийн үнэлгээ нь өвчний хүндрэлээс урьдчилан сэргийлэхэд чухал ач холбогдолтой. Сүүлийн жилүүдэд машин сургалтын аргуудыг эмнэлгийн өгөгдөлд ашиглан өвчний төлөвийг урьдчилан таамаглах, эмнэлзүйн шийдвэр гаргалтыг дэмжих боломжийг судлах болсон хэдий ч манай улсад энэ чиглэлээр хийгдсэн судалгааны ажил хомс байгаа нь энэхүү судалгааны үндэслэл боллоо. **Зорилго:** PIMA Indians Diabetes Database өгөгдлийн багцад Random Forest ангиллын алгоритм ашиглан чихрийн шижингийн төлөвийг урьдчилан таамаглах загвар боловсруулж, түүний ангиллын гүйцэтгэлийг үнэлэх. **Арга, аргачлал:** Судалгаанд 768 хүний ажиглалт, найман тайлбарлагч хувьсагч болон нэг зорилтот хувьсагч бүхий PIMA өгөгдлийн багцыг ашиглав. Физиологийн хувьд боломжгүй тэг утгатай зарим хэмжилтийг дутуу утга гэж үзэн NA болгон хувиргаж, медиан утгаар нөхсөн. Өгөгдлийг 80:20 харьцаагаар сургалтын болон тестийн хэсэгт хуваж, сургалтын өгөгдөлд 10-fold cross-validation ашиглан Random Forest загварын гиперпараметрийг тохируулав. Загварын гүйцэтгэлийг Accuracy, Sensitivity, Specificity, Balanced Accuracy, Cohen's Kappa болон ROC-AUC үзүүлэлтээр үнэлж, хувьсагчдын ач холбогдлыг тодорхойлов. Статистик болон машин сургалтын боловсруулалтыг R программчлалын орчинд гүйцэтгэв. **Үр дүн:** Хөндлөн баталгаажуулалтын үр дүнд mtry=4 тохиргоо хамгийн сайн гүйцэтгэл үзүүлж, Accuracy 78.05%, Cohen's Kappa 0.5062 байв. Тест өгөгдөл дээр загварын Accuracy 72.55%, Sensitivity 58.49%, Specificity 80.00%, Balanced Accuracy 69.25%, Cohen's Kappa 0.3883 гарсан. ROC муруйн доорх талбай (AUC) 0.7802 байсан нь загвар чихрийн шижинтэй болон чихрийн шижингүй тохиолдлыг санамсаргүй ангиллаас илүү сайн ялгах чадвартайг харууллаа. Хувьсагчдын ач холбогдлын шинжилгээгээр цусан дахь глюкозын хэмжээ хамгийн өндөр ач холбогдолтой байсан бөгөөд биеийн жингийн индекс (BMI), нас, жирэмслэлтийн тоо дараагийн чухал хувьсагчид байв. **Дүгнэлт:** Random Forest алгоритм нь PIMA өгөгдлийн багцад чихрийн шижингийн төлөвийг урьдчилан таамаглах боломжтой боловч тест өгөгдөл дээрх мэдрэг чанар харьцангуй бага байв. Иймээс уг загварыг эмнэлзүйн оношилгоог орлох хэрэгсэл бус, эрсдэлийн урьдчилсан үнэлгээ, эрт илрүүлэлтийг дэмжих боломжит арга хэлбэрээр авч үзэх нь зүйтэй. Цаашид мэдрэг чанарыг сайжруулах, бусад машин сургалтын алгоритмтай харьцуулах, илүү өргөн хүрээний өгөгдөл дээр баталгаажуулах шаардлагатай байна.

Үндэслэл: Чихрийн шижин нь дэлхий даяар өргөн тархсан, хүн амын эрүүл мэндэд ихээхэн нөлөө үзүүлдэг архаг бодисын солилцооны эмгэг юм. Дэлхийн Эрүүл Мэндийн Байгууллагын (WHO) мэдээлснээр чихрийн шижин нь цусан дахь глюкозын хэмжээ удаан хугацаанд ихсэх байдлаар илэрдэг бөгөөд 2021 онд уг өвчин нь дэлхий даяар 1.6 сая гаруй хүний нас баралтын шууд шалтгаан болсон байна. Чихрийн шижингийн тархалт нэмэгдэж байгаа нь эрт илрүүлэлт, эрсдэлийн үнэлгээ, урьдчилан сэргийлэх үр дүнтэй арга хэрэгслийн хэрэгцээг нэмэгдүүлж байна. [1]

Америкийн Чихрийн Шижингийн Холбоо (American Diabetes Association, ADA)-ны оношилгооны шалгууруар өлөн үеийн плазмын глюкозын хэмжээ ≥126 mg/dL, эсвэл бусад баталгаат шинжилгээний үзүүлэлтүүд тогтоосон босго хэмжээнд хүрсэн тохиолдолд чихрийн шижин гэж оношилдог. [2] Харин өлөн үеийн глюкозын хэмжээ 100–125 mg/dL байх нь чихрийн шижингийн өмнөх үе гэж тодорхойлогддог бөгөөд энэ нь цаашид II хэлбэрийн чихрийн шижин үүсэх эрсдэл нэмэгдсэнийг илэрхийлдэг. [2, 3]

Чихрийн шижингийн архаг хүндрэл нь том болон жижиг судасны гэмтлээр илэрч, зүрх судас, тархи, бөөр, нүд, мэдрэлийн тогтолцоог гэмтээдэг. Иймээс цусан дахь глюкозын түвшнийг эрт үеэс үр дүнтэй хянах нь хүндрэл, нас баралтыг бууруулахад чухал ач холбогдолтой. [4, 5]

Сүүлийн жилүүдэд хиймэл оюун ухаан (Artificial Intelligence, AI) болон машин сургалтын аргууд нь чихрийн шижингийн эрт илрүүлэлт, эрсдэлийн үнэлгээ, цусан дахь глюкозын хяналт, эмчилгээний шийдвэр гаргалтыг дэмжихэд өргөн хэрэглэгдэх болсон бөгөөд эмнэлзүйн практикт үр дүнтэй хэрэглэгдэх боломжтойг олон судалгаагаар харуулж байна. [6]

Олон судлаачид Pima Indian Diabetes Dataset (PIDD) өгөгдлийн багцыг ашиглан чихрийн шижинг урьдчилан таамаглах зорилгоор машин сургалтын (Machine Learning, ML) аргуудыг хэрэглэж байна. Энэхүү өгөгдлийн багц нь эмэгтэй өвчтөнүүдийн 768 ажиглалт болон 9 үндсэн үзүүлэлтийг агуулдаг бөгөөд чихрийн шижингийн эрсдэлийг үнэлэх, ангилах машин сургалтын загваруудыг боловсруулахад өргөн ашиглагддаг. [7, 8] PIDD өгөгдөлд суурилсан судалгаанд Random Forest, SVM, Naïve Bayes, Decision Tree зэрэг алгоритмыг ашиглан чихрийн шижинг таамаглаж байгаа ч өгөгдлийн тэнцвэргүй байдал, хувьсагчдын хамаарал, загварын тайлбарлах чадвар нь гол сорилт хэвээр байна. [9]

Зорилго: PIMA Indians Diabetes Database өгөгдөлд Random Forest ангиллын загвар байгуулж, чихрийн шижингийн төлөвийг таамаглах гүйцэтгэлийг Accuracy, Sensitivity, Specificity болон ROC-AUC үзүүлэлтээр үнэлэхэд оршино.

Судалгааны өгөгдөл: Энэхүү судалгаанд UCI Machine Learning Repository-д байршуулсан PIMA Indians Diabetes Database өгөгдлийн багцыг ашигласан. [10] Тус өгөгдлийн багц нь чихрийн шижингийн эрсдэлийг машин сургалтын аргаар ангилах зорилгоор өргөн хэрэглэгддэг бөгөөд нийт 768 хүний мэдээлэл багтсан бөгөөд бүх оролцогчид нь эмэгтэй, 21 ба түүнээс дээш насны хүмүүс байна. Энэ багц нь 8 тайлбарлагч хувьсагч, нэг зорилтот хувьсагч бүхий нийт 9 хувьсагчаас (attribute) бүрдэнэ. Үүнд:

1. Жирэмслэлтийн тоо (Pregnancy)
2. Биеийн жингийн индекс (BMI)
3. Инсулины түвшин (Insulin level)
4. Нас (Age)
5. Цусны даралт (Blood pressure)
6. Арьсны зузаан (Skin thickness)
7. Цусан дахь глюкозын хэмжээ (Glucose)
8. Чихрийн шижингийн удамшлын индекс (Diabetes pedigree function)
9. Чихрийн шижингийн төлөв (Outcome)

Зорилтот хувьсагч болох Outcome нь чихрийн шижингийн төлөвийг хоёртын ангиллаар илэрхийлсэн байна. Энд DiabetesPedigreeFunction нь физик хэмжих нэгжгүй, өгөгдлийн багцад удамшлын эрсдэлийг илэрхийлэх индекс утга хэлбэрээр өгөгдсөн байна.

Өгөгдлийг урьдчилан боловсруулах (Data preprocessing): PIMA Diabetes өгөгдлийн багцын Glucose, BloodPressure, SkinThickness, Insulin, BMI хувьсагчдын бодит бус 0 утгыг дутуу утга (NA) болгон хувиргаж, медианаар нөхөв (Хүснэгт 2). Медиан нь хэт ялгаатай утгын нөлөөнд бага өртдөг тул эмнэлзүйн өгөгдөлд тохиромжтой. Урьдчилан боловсруулсан өгөгдлийг машин сургалтын загвар сургах, туршихад ашиглав.

Хүснэгт 2. Diabetes өгөгдлийн багц дахь дутуу утгын тоо

№	Хувьсагч	Дутуу утгын тоо
1	Glucose (Цусан дахь глюкозын хэмжээ)	5
2	Blood Pressure (Цусны даралт)	35
3	Skin Thickness (Арьсны зузаан)	227
4	Insulin (Инсулины хэмжээ)	374
5	BMI (Биеийн жингийн индекс)	11

Машин сургалтын Random Forest ангиллын загвар байгуулах ба үнэлэх: PIDD өгөгдлийн багцын Outcome хувьсагчийг таамаглах Random Forest ангиллын загвар байгуулж, өгөгдлийг сургалтын 80%, тестийн 20% харьцаагаар санамсаргүй хуваав. Загварын ерөнхийлөх чадвар, тогтвортой байдлыг үнэлэхийн тулд сургалтын өгөгдөлд 10-fold cross-validation ашиглаж, 10 давталтын үнэлгээний үзүүлэлтүүдийн дунджаар гүйцэтгэлийг тодорхойлов. Тестийн өгөгдлийг загварын урьдчилан таамаглах чадварыг үнэлэхэд ашигласан. [11]

Random Forest нь Breiman-ийн боловсруулсан ансамбль сургалтын алгоритм бөгөөд bootstrap түүвэрлэлээр олон шийдвэрийн мод байгуулж, хувьсагчдын санамсаргүй дэд олонлогийг ашиглан салаалалтыг сонгодог. Ангиллын эцсийн үр дүнг мод бүрийн таамаглалыг нэгтгэн, олонхийн саналаар тодорхойлдог. [12]

Random Forest нь олон шийдвэрийн модны үр дүнг нэгтгэснээр хэт тохируулгыг бууруулж,

өгөгдлийн хэлбэлзэлд тэсвэртэй ажиллан, шугаман бус хамаарлыг илрүүлэх давуу талтай.[12]

Судалгаанд 500 мод (ntree=500) ашиглаж, загварын гүйцэтгэлийг сайжруулахын тулд модны салаалалт бүрд сонгогдох хувьсагчдын тоог илэрхийлдэг mtry параметрийг тохируулан үнэлэв.[13]

Хөндлөн баталгаажуулалт (Cross-validation): Загварын гиперпараметрийг сонгох, гүйцэтгэлийг найдвартай үнэлэхийн тулд сургалтын өгөгдлийг 10 хэсэгт хуваах, давталт бүрд 9 хэсгээр сурган 1 хэсгээр баталгаажуулах 10-fold cross-validation аргыг ашиглав. Үнэлгээний үзүүлэлтүүдийн дунджаар загварын гүйцэтгэлийг тодорхойлсон.[14,15,11]

Загварын үнэлгээ: Загварын гүйцэтгэлийг ангиллын алдааны матрицад үндэслэн Accuracy, Cohen's Kappa, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Balanced Accuracy болон ROC-AUC үзүүлэлтээр үнэлэв.[14,16,17]

Програм хангамж: Өгөгдөл боловсруулах, загвар сургах, 10-fold cross-validation хийхэд R программчлалын орчин, caret болон randomForest багц; ROC-AUC тооцоолоход pROC багцыг ашиглав.[18,14,19]

Үр дүн: Өгөгдөл урьдчилан боловсруулсны дараа 768 ажиглалт, 8 тайлбарлагч хувьсагчийг загварт ашиглав: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age. Outcome зорилтот хувьсагч нь 0 (чихрийн шижингүй), 1 (чихрийн шижинтэй) гэсэн хоёр ангилалтай.

Өгөгдлийн багцыг санамсаргүй хуваалтын аргаар 80% сургалтын өгөгдөл (615 ажиглалт), 20% тестийн өгөгдөл (153 ажиглалт) болгон хуваасан.

Random Forest загварын гиперпараметрийг тохируулах зорилгоор зөвхөн сургалтын өгөгдөл дээр 10-fold cross-validation үнэлгээг хийсэн. Шийдвэрийн модны салаалалт бүрт санамсаргүй сонгогдох хувьсагчдын тоог илэрхийлдэг mtry параметрийн 2, 3, 4, 5, 6 гэсэн утгуудыг туршин харьцуулсан. (Хүснэгт 3)

Хүснэгт 3. Random Forest загварын mtry параметрийн утгуудын гүйцэтгэлийн харьцуулалт

mtry	Нарийвчлал (Accuracy)	Kappa коэффициент
2	0.7789001	0.4961948
3	0.7772607	0.4945433
4	0.7805394	0.5061553
5	0.7805130	0.5036002
6	0.7740085	0.4896629

10-fold cross-validation үнэлгээний үр дүнгээс харахад mtry=4 тохиргоо нь хамгийн өндөр Accuracy (0.7805; 78.05%) болон Cohen's Kappa коэффициент (0.5062) үзүүлсэн байна. Иймээс Accuracy үзүүлэлтийн хамгийн өндөр утгыг үндэслэн mtry=4 параметрийг Random Forest загварын оновчтой гиперпараметр болгон сонгож, тусдаа тест өгөгдөл дээрх гүйцэтгэлийн үнэлгээнд ашигласан.

Сонгосон Random Forest загвар нь PIMA Diabetes өгөгдлийн багц дээр чихрийн шижингийн төлөвийг ангилахад 78.05%-ийн Accuracy үзүүлэлттэй байсан. Cohen's Kappa коэффициент 0.5062 гарсан нь загварын таамагласан болон бодит ангиллын хоорондын тохирол санамсаргүй тохирлоос давсан, дунд түвшний (moderate agreement) байгааг харуулсан. 10-fold cross-validation аргыг ашигласнаар сургалтын өгөгдлийг олон хэсэгт хуваан давтан сургах, баталгаажуулах замаар загварын үнэлгээний хэлбэлзлийг бууруулж, Random Forest загварын ерөнхийлөх чадварыг илүү найдвартай үнэлэх боломж бүрдсэн.

Загварын тест өгөгдөл дээрх гүйцэтгэл: 10-fold cross-validation үнэлгээгээр сонгосон оновчтой тохиргоотой Random Forest загварын ерөнхийлөх чадварыг үнэлэх зорилгоор тусдаа 20%-ийн тест өгөгдөл дээр шалгав. Загварын ангиллын гүйцэтгэлийг ангиллын алдааны матриц (confusion matrix)-д үндэслэн Accuracy, Cohen's Kappa, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Balanced Accuracy үзүүлэлтүүдээр үнэлэв. (Хүснэгт 4)

Хүснэгт 4. Загварын тест өгөгдөл дээрх ангиллын гүйцэтгэл

Үзүүлэлт	Үр дүн
Accuracy	72.55%
Sensitivity	58.49%
Specificity	80.00%
PPV	60.78%
NPV	78.43%
Balanced Accuracy	69.25%
Kappa	0.3883

Тестийн 153 ажиглалтаас 111-ийг зөв ангилж, Accuracy=72.55%, Cohen's Kappa=0.3883 үзүүлэв. Sensitivity=58.49%, Specificity=80.00%, PPV=60.78%, NPV=78.43%, Balanced Accuracy=69.25% байв.

Загвар чихрийн шижинтэй 53 тохиолдлын 31-ийг зөв илрүүлсэн бол чихрийн шижингүй 100 тохиолдлын 80-ыг зөв ангилжээ. Иймээс чихрийн шижингүй тохиолдлыг ялгах чадвар харьцангуй өндөр боловч өвчтэй тохиолдлыг илрүүлэх мэдрэг чанарыг сайжруулах шаардлагатай.

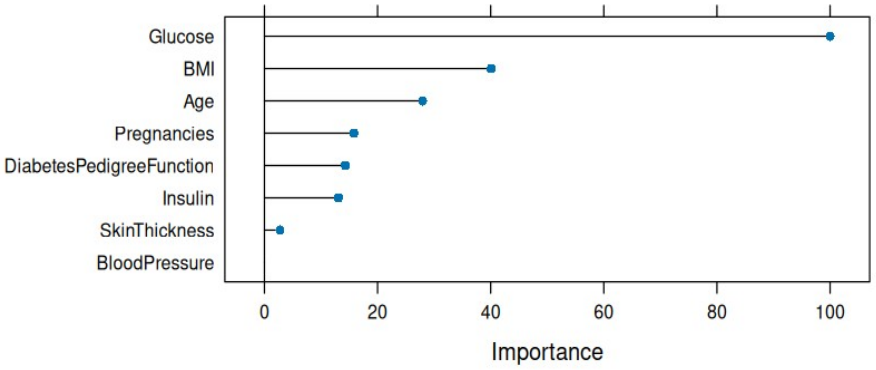
Ангиллын алдааны матриц: Random Forest загварын тест өгөгдөл дээрх ангиллын үр дүнг алдааны матриц (confusion matrix)-д үндэслэн үзүүлэв. (Хүснэгт 5) Нийт 153 тестийн ажиглалтаас:

1. 80 тохиолдлыг чихрийн шижингүй (Outcome=0) гэж зөв ангилсан (True Negative, TN),
2. 31 тохиолдлыг чихрийн шижинтэй (Outcome=1) гэж зөв илрүүлсэн (True Positive, TP),
3. 20 тохиолдлыг бодитоор чихрийн шижингүй боловч чихрийн шижинтэй гэж буруу ангилсан (False Positive, FP),
4. 22 тохиолдлыг бодитоор чихрийн шижинтэй боловч чихрийн шижингүй гэж буруу ангилсан (False Negative, FN) байна.

Хүснэгт 5. Random Forest загварын тест өгөгдөл дээрх алдааны матриц

	Бодит Outcome=0	Бодит Outcome=1
Таамагласан Outcome=0	TN=80	FN=22
Таамагласан Outcome=1	FP=20	TP=31

Random Forest загварын хувьсагчийн ач холбогдол: Random Forest загварын хувьсагчийн ач холбогдлын (Variable Importance) шинжилгээний үр дүн. (Зураг 1)



Зураг 1. Random Forest загварын хувьсагчийн ач холбогдол

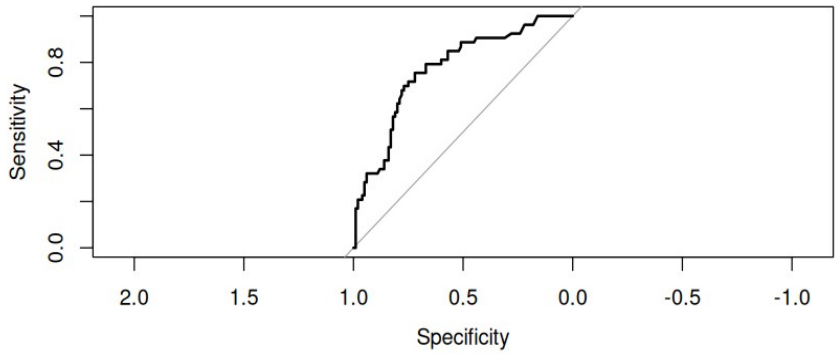
Хувьсагчдын ач холбогдлыг Mean Decrease Gini үзүүлэлтээр үнэлсэн бөгөөд энэ нь тухайн хувьсагч модны салаалалтын үед ангиллын ялгаралд хэр их хувь нэмэр оруулж байгааг илэрхийлдэг.

Үр дүнгээс харахад Glucose хувьсагч хамгийн өндөр ач холбогдолтой (Importance=100) байсан бөгөөд Random Forest загварын шийдвэр гаргалтад хамгийн их нөлөө үзүүлсэн байна. Дараагийн өндөр ач холбогдолтой хувьсагчид нь BMI, Age болон Pregnancies байсан бөгөөд эдгээр нь чихрийн шижинг ангилахад чухал мэдээлэл өгч байгааг харууллаа. Харин DiabetesPedigreeFunction, Insulin, SkinThickness болон BloodPressure хувьсагчдын ач холбогдол харьцангуй бага байсан боловч загварын ангиллын шийдвэрт тодорхой хэмжээгээр оролцсон байна.

Эдгээр үр дүн нь Random Forest загвар нь чихрийн шижингийн эрсдэлийг таамаглахдаа ялангуяа цусан дахь глюкозын түвшин, биеийн жингийн индекс (BMI), нас зэрэг эмнэлзүйн үндсэн үзүүлэлтүүдэд илүү их тулгуурлаж байгааг харуулж байна.

Тест өгөгдөл дээрх Cohen's Kappa=0.3883 байсан нь загварын таамагласан болон бодит ангиллын тохирол санамсаргүй таамгаас давсан боловч төгс бус байгааг харуулна. Хөндлөн баталгаажуулалтын Kappa=0.5062-оос бага байгаа нь загварын гүйцэтгэлийг шинэ өгөгдөл дээр цаашид баталгаажуулах шаардлагатайг илтгэнэ.

ROC муруйн шинжилгээ: Random Forest ангиллын загварын ялгах чадварыг үнэлэх зорилгоор Receiver Operating Characteristic (ROC) муруй болон түүний доорх талбай (Area Under the Curve, AUC)-г тооцоолсон. ROC муруй нь ангиллын босго утгыг өөрчлөх үед загварын эерэг тохиолдлыг зөв илрүүлэх чадвар (True Positive Rate, TPR буюу Sensitivity) болон хуурамч эерэг ангиллын түвшин (False Positive Rate, FPR) хоорондын хамаарлыг харуулдаг.



Зураг 2. ROC муруй.

Энэхүү судалгаанд Random Forest загварын ROC муруйн доорх талбай AUC=0.7802 гарсан. Энэ нь загвар нь чихрийн шижинтэй болон чихрийн шижингүй тохиолдлуудыг ялгах санамсаргүй ангиллаас мэдэгдэхүйц өндөр чадвартай болохыг харуулж байна. Өөрөөр хэлбэл, санамсаргүй сонгосон нэг чихрийн шижинтэй болон нэг чихрийн шижингүй тохиолдлын хувьд загвар нь эерэг тохиолдлыг илүү өндөр магадлалтайгаар зөв эрэмбэлэх боломж ойролцоогоор 78% байна.

Хэлцэмж: Судалгаагаар Random Forest загвар нь PIMA өгөгдлийн багцад чихрийн шижингийн төлөвийг урьдчилан таамаглах боломжтойг харууллаа. 10-fold cross-validation үнэлгээгээр mtry=4 тохиргоо хамгийн сайн үр дүн үзүүлсэн (Accuracy=78.05%,

Kappa=0.5062).

Тусгаарласан тест өгөгдөл дээр Accuracy=72.55%, Kappa=0.3883, AUC=0.7802 байв. Загварын Specificity (80.00%) нь Sensitivity (58.49%)-ээс өндөр байсан нь чихрийн шижинтэй зарим тохиолдлыг илрүүлж чадаагүйг харуулна. Скринингийн хэрэглээнд мэдрэг чанарыг нэмэгдүүлэхийн тулд ангиллын босго тохируулах, class weighting эсвэл SMOTE зэрэг аргыг цаашид судалж болно. [20]

ROC-AUC=0.7802 нь загвар хоёр ангиллыг санамсаргүй таамгаас илүү сайн ялгах чадвартайг илтгэнэ. Хувьсагчийн ач холбогдлын шинжилгээгээр Glucose хамгийн өндөр нөлөөтэй, дараа нь BMI, Age, Pregnancies орсон нь эмнэлзүйн үндсэн үзүүлэлтүүдийн ач холбогдлыг харуулж байна.

Өмнөх судалгаанд PIMA өгөгдөлд Decision Tree, SVM, Naïve Bayes зэрэг алгоритм хэрэглэж, машин сургалтын аргууд чихрийн шижинг таамаглах боломжтойг мэдээлсэн. [21] Гэсэн хэдий ч эмнэлзүйн хэрэглээнд нэвтрүүлэхийн өмнө илүү өргөн хүрээний өгөгдөл дээр баталгаажуулах шаардлагатай. [22]

Энэхүү судалгаанд зөвхөн Random Forest загварыг үнэлсэн тул цаашид Logistic Regression, Support Vector Machine, XGBoost болон Neural Network зэрэг алгоритмтай харьцуулж, гаднын өгөгдөл дээр баталгаажуулах нь зүйтэй. [23]

Дүгнэлт: PIMA Indians Diabetes Database өгөгдөлд Random Forest загвар байгуулж, чихрийн шижингийн төлөвийг таамаглах гүйцэтгэлийг үнэлэв. 10-fold cross-validation-ийн үр дүнд mtry=4 оновчтой тохиргоо болж, Accuracy=78.05%, Kappa=0.5062 үзүүлсэн.

Тест өгөгдөл дээр загварын Accuracy=72.55%, Sensitivity=58.49%, Specificity=80.00%, Balanced Accuracy=69.25%, AUC=0.7802 байв. Glucose хамгийн чухал хувьсагч байсан бөгөөд BMI, Age, Pregnancies дараагийн байранд орсон. Загвар нь эрсдэлийн урьдчилсан үнэлгээ, эрт илрүүлэлтийг дэмжих боломжтой боловч оношилгоог орлохгүй. Мэдрэг чанарыг нэмэгдүүлж, бусад алгоритм болон өргөн хүрээний өгөгдөл дээр баталгаажуулах шаардлагатай.

Ном зүй:

1. *Organization WH. Diabetes. 2026. Available at:* https://www.who.int/health-topics/diabetes.
2. *Association AD. Diagnosis. 2026. Available at:* https://diabetes.org/about-diabetes/diagnosis.
3. *Clinic. M. Prediabetes: Diagnosis and treatment. July 23, 2026. Available at:* https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284.
4. *American Diabetes Association Professional Practice Committee. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2025. Diabetes Car. 2025;no. Suppl 1:S27-S49.*
5. *M. Brownlee. The pathobiology of diabetic complications: a unifying mechanism. Diabetes. 2005;21(no. 6,):1615–1625.*
6. *I. Contreras and J. Vehi. Artificial intelligence for diabetes management and decision support: Literature review. Journal of Medical Internet Research. 2018;20(no. 5, doi: 10.2196/10775.):e10775.*
7. *e. a. T.M. Alam. Informatics in medicine unlocked a model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019;16:100204.*
8. *Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang. Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in Genetics. 2018;9(doi: 10.3389/fgene.2018.00515):515.*
9. *S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science. 2016;82:115-121.*
10. *J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. in Annual Symposium on Computer Applications in Medical Care. 1988:261-265.*
11. *R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection,» in Proceedings of the 14th. International Joint Conference on Artificial Intelligence (IJCAI). 1995;2:1137-1143.*
12. *L. Breiman. Random forests. Machine Learning. 2001;45(doi: 10.1023/A:1010933404324):5-32.*
13. *P. Probst, M. N. Wright, and A.-L. Boulesteix. Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery. 2019;9(no. 3, doi: 10.1002/widm.1301):e1301.*
14. *M. Kuhn and K. Johnson. Applied Predictive Modeling. Springer. 2013.*
15. *T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 2009.*
16. *T. Fawcett. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(no. 8, doi: 10.1016/j.patrec.2005.10.010.):861-874.*
17. *J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(no. 1, doi: 10.1177/001316446002000104):37-46.*
18. *Available. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2024. Available at:* https://www.R-project.org/.
19. *A. Liaw and M. Wiener. Classification and regression by randomForest. R News. 2002;2(3):18-22.*
20. *N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16(doi: 10.1613/jair.953):321-357.*
21. *D. Sisodia and D. S. Sisodia. Prediction of Diabetes using Classification Algorithms. Procedia Computer Science. 2018;132(doi: 10.1016/j.procs.2018.05.122):1578-1585.*
22. *J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan,. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University – Computer and Information Sciences. 2022;34(no. 6):3204-3225.*
23. *A. D. Association. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2024. Diabetes Care. 2024;47(no. Supplement_1, doi: 10.2337/dc24-S002):S20-S42.*

Predicting Diabetes Status Using the Random Forest Algorithm Based on the PIMA Dataset

*Baasandorj Ch[1], Odgerel B[2], Ajnai L[1], Javzmaa Ts[1], Nyamdavaa U[1], Oyuntsetseg B[3], Enkh-Urel E, Purevdolgor L*1*

[1]Department of Physics and Informatics, School of Biomedicine, MNUMS
[2]Department of Physics and Informatics, School of Biomedicine, MNUMS
[3]Department of General Education, School of Nursing, MNUMS
[3]Department of Communication Skills, School of Biomedicine, MNUMS
E-mail: purevdolgor@mnums.edu.mn, Tel: +976-95091107

Abstract:
Background: Diabetes mellitus is a chronic metabolic disorder with an increasing global prevalence. Early detection and risk assessment are essential for preventing disease-related complications. In recent years, machine learning methods have been increasingly investigated for predicting disease status using medical data and supporting clinical decision-making.
Aim: This study aimed to develop a predictive model for diabetes status using the Random Forest classification algorithm based on the PIMA Indians Diabetes Database and to evaluate its classification performance.
Materials and Methods: The study used the PIMA dataset, comprising 768 observations, eight predictor variables, and one target variable. Zero values in selected measurements that were physiologically implausible were treated as missing values and imputed using the median. The dataset was split into training and test sets at an 80:20 ratio. Hyperparameters of the Random Forest model were optimized using 10-fold cross-validation on the training data. Model performance was evaluated using Accuracy, Sensitivity, Specificity, Balanced Accuracy, Cohen's Kappa, and the area under the receiver operating characteristic curve (ROC-AUC). Variable importance was also assessed. All statistical analyses and machine learning procedures were performed in the R programming environment.
Results: Cross-validation identified mtry=4 as the optimal configuration, achieving an Accuracy of 78.05% and a Cohen's Kappa of 0.5062. On the test set, the model achieved an Accuracy of 72.55%, Sensitivity of 58.49%, Specificity of 80.00%, Balanced Accuracy of 69.25%, and Cohen's Kappa of 0.3883. The area under the ROC curve (AUC) was 0.7802, indicating that the model could distinguish between individuals with and without diabetes better than random classification. Variable importance analysis identified blood glucose level as the most influential predictor, followed by body mass index (BMI), age, and number of pregnancies.
Conclusion: The Random Forest algorithm demonstrated the potential to predict diabetes status using the PIMA dataset; however, its sensitivity on the test set was relatively low. Therefore, the model should be considered a potential tool for preliminary risk assessment and supporting early detection rather than a replacement for clinical diagnosis. Further research is needed to improve sensitivity, compare the model with other machine learning algorithms, and validate its performance using larger and more diverse datasets.

Keywords: PIMA Indians Diabetes Database, machine learning, classification model, ROC-AUC, risk assessment.