АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION AND COMMUNICATION TECHNOLOGIES

ЖАСАНДЫ ИНТЕЛЛЕКТ
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
ARTIFICIAL INTELLIGENCE

DOI 10.51885/1561-4212_2025_2_92
МҒТАА 28.23.01

А.Б. Абен[1], Н.М. Жунисов[2], Г.Н. Казбекова[3], А.Н. Аманов[4], А.А. Абибуллаева[5]
Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті,
Түркістан қ., Қазақстан
[1]*E-mail: arypzhan.aben@ayu.edu.kz**
[2]*E-mail: nurseit.zhunissov@ayu.edu.kz*
[3]*E-mail: gulnur.kazbekova@ayu.edu.kz*
[4]*E-mail: anuarbek.amanov@ayu.edu.kz*
[5]*E-mail: aiman.abibullayeva@ayu.edu.kz*

ҚАНТ ДИАБЕТІН АНЫҚТАУДА МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІН ҚОЛДАНУ

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ДИАБЕТА

USE OF MACHINE LEARNING ALGORITHMS IN DIABETES DETECTION

Аңдатпа. *Медициналық салада машиналық оқыту алгоритмдерін пайдалану, әсіресе аурудың дамуын болжауда, асқынуларды анықтауда және клиникалық шешім қабылдауға көмектесуде айтарлықтай тартымдылыққа ие болды. Бұл зерттеу пациенттің медициналық жазбаларынан алынған ауқымды деректер жиынтығын пайдалана отырып, қант диабетінің болуын болжау үшін әртүрлі машиналық оқыту әдістерін қолдануды зерттейді. Деректер жинағы жіктеу процесіне ықпал ететін маңызды медициналық көрсеткіштерді қамтиды.*

Деректерді алдын ала өңдеу жетіспейтін мәндерді шешуді, деректерді қалыпқа келтіруді және оларды оқыту және тестілеу жиындарына бөлуді қамтиды. Зерттеу бес түрлі алгоритмнің тиімділігін бағалайды: логистикалық регрессия, шешім ағашы, кездейсоқ орман, тірек векторлық машина (SVM) және k-ең жақын көршілері (KNN), қант диабетін болжау үшін ең тиімді үлгіні анықтау. Әр алгоритмнің болжамды дәлдігі мен сенімділігін бағалау үшін өнімділік көрсеткіштері қолданылды.

Нәтижелер KNN қант диабетінің оң және теріс жағдайларын тиімді ажырата отырып, жоғары өнімділікті көрсететінін көрсетеді. Керісінше, логистикалық регрессия мақтауға тұрарлық жалпы тиімділікті көрсетті, ал шешім ағашының алгоритмі сенімділіктегі шектеулерді анықтады. Кездейсоқ орман және SVM үлгілері қанағаттанарлық нәтижелер берді; дегенмен, олар белгілі бір контексттерде нақты шектеулерді ұсынды. Бұл зерттеу қант диабетін диагностикалауда машиналық оқыту алгоритмдерінің әлеуетін көрсетеді, клиникалық сценарийлерге бейімделген алгоритмді таңдаудың маңыздылығын көрсетеді. Болашақ зерттеулер деректердің тұтастығын жақсартуға, алгоритмдік тәсілдерді нақтылауға және гиперпараметрлерді оңтайландыруға, осылайша болжау дәлдігін арттыруға бағытталуы керек. Мұндай жетістіктер қант диабетін диагностикалауды айтарлықтай жақсартуға және науқасты уақтылы анықтауды жеңілдетуге дайын.

Түйін сөздер: *Қант диабеті, машиналық оқыту, шешім ағаштары, к-Ең жақын көрші.*

Аннотация. *Применение алгоритмов машинного обучения в медицине, особенно в области предсказания динамики заболеваний, выявления осложнений и поддержки клинических решений, вызывает значительное внимание. Данное исследование направлено на изучение применения различных техник машинного обучения для прогнозирования возникновения диабета с использованием обширного набора данных, полученных из медицинских записей пациентов.*

Сначала была проведена обработка данных, включающая работу с пропущенными значениями, нормализацию данных и их разделение на обучающие и тестовые наборы. В ходе исследования была оценена эффективность пяти различных алгоритмов: логистической регрессии, дерева решений, случайного леса, векторных машин (SVM) и метода ближайших соседей (KNN), с целью определения наиболее эффективной модели для прогнозирования диабета. Для оценки точности и надежности предсказаний каждого алгоритма использовались показатели производительности.

Результаты показали, что алгоритм KNN демонстрирует высокую способность к эффективному различению положительных и отрицательных случаев диабета, обеспечивая лучшие результаты. Логистическая регрессия также продемонстрировала общую эффективность, в то время как у алгоритма дерева решений были обнаружены ограничения в надежности. Модели случайного леса и SVM показали удовлетворительные результаты, однако они имели ограничения в определенных контекстах. Данное исследование подчеркивает потенциал алгоритмов машинного обучения в диагностике диабета, а также важность выбора алгоритмов в зависимости от клинических сценариев. Будущие исследования должны быть направлены на повышение целостности данных, улучшение алгоритмических подходов и оптимизацию гиперпараметров, что, в свою очередь, должно повысить точность прогнозирования. Такие достижения могут значительно улучшить диагностику диабета и обеспечить своевременное выявление заболеваний у пациентов.

Ключевые слова: *диабет, машинное обучение, деревья решений, k-ближайший сосед.*

Abstract. *The utilization of machine learning algorithms in the medical domain has gained significant traction, particularly in predicting disease progression, identifying complications, and aiding in clinical decision-making. This study investigates the application of various machine learning techniques to predict the presence of diabetes, leveraging an extensive dataset derived from patient medical records. The dataset encompasses crucial medical indicators that contribute to the classification process.*

Preliminary data processing involved addressing missing values, normalizing the data, and partitioning it into training and testing subsets. The research evaluates the efficacy of five distinct algorithms: logistic regression, decision tree, random forest, support vector machine (SVM), and k-nearest neighbors (KNN), to determine the most effective model for diabetes prediction. Performance metrics were employed to assess each algorithm's predictive accuracy and reliability.

The results indicate that KNN demonstrates superior performance, effectively distinguishing between positive and negative cases of diabetes. In contrast, logistic regression exhibited commendable overall efficacy, whereas the decision tree algorithm revealed limitations in reliability. Both random forest and SVM models produced satisfactory results; however, they presented specific constraints in particular contexts. This study underscores the potential of machine learning algorithms in diabetes diagnostics, highlighting the critical importance of algorithm selection tailored to clinical scenarios. Future research should aim to enhance data integrity, refine algorithmic approaches, and optimize hyperparameters, thereby augmenting predictive accuracy. Such advancements are poised to significantly improve diabetes diagnostics and facilitate timely patient identification.

Keywords: *Diabetes, Machine Learning, Decision Trees, k-Nearest Neighbor*

Kipicne. Қант диабеті – әлемдегі денсаулық сақтау жүйесіне айтарлықтай әсер ететін созылмалы аурулардың бірі. Бұл ауру қандағы глюкоза деңгейінің тұрақты жоғарылауымен сипатталады, ол инсулиннің жеткіліксіздігі немесе оның әсерінің бұзылуымен байланысты. Дүниежүзілік денсаулық сақтау ұйымының мәліметтеріне сәйкес, қант диабетімен ауыратын адамдардың саны жыл сайын артып келеді, бұл мәселені шешу медициналық қауымдастық үшін өте өзекті. Аурудың асқынулары, мысалы, жүрек-қан тамырлары аурулары, бүйрек жеткіліксіздігі, көздің зақымдануы, және аяқтардың ампутациясы, науқастардың өмір сапасына теріс әсер етеді және емдеу шығындарын арттырады.

Қант диабетін ерте диагностикалау мен тиімді емдеу науқастардың өмір сапасын айтарлықтай жақсартуға, аурудың асқынуларын болдырмауға, әрі науқастардың денсаулығын сақтауға мүмкіндік береді. Дәстүрлі диагностикалық әдістер, мысалы, клиникалық тесттер мен пациенттердің симптомдарына негізделген тәсілдер, кейде дәл емес немесе кешігіп қолданылуы мүмкін, себебі олар деректердің толық жиынтығын ескермейді. Сондықтан, жаңа технологиялар мен әдістерді, атап айтқанда, машиналық оқыту алгоритмдерін енгізу қажет.

Осы зерттеудің мақсаты – қант диабетінің болуын анықтау үшін машиналық оқыту алгоритмдерін қолданудың тиімділігін бағалау. Біз қарастыратын алгоритмдердің әрқайсысының ерекшеліктері мен нәтижелері салыстырылып, қант диабетін диагностикалаудағы практикалық қолданбалы әлеуеті анықталады. Зерттеу нәтижелері машиналық оқыту әдістерінің диагностикадағы тиімділігін көрсетеді, бұл олардың клиникалық тәжірибеде енгізілуіне мүмкіндік береді. Болашақта бұл әдістерді жетілдіру, гипер параметрлерді баптау және деректердің сапасын арттыру арқылы науқастарды дер кезінде анықтауға мүмкіндік беретін жаңа құралдарды дамыту жолында маңызды қадам болады.

Әдебиеттерге шолу. Қант диабетін анықтауда машиналық оқыту алгоритмдерінің қолданылуы соңғы жылдары зерттеу саласында үлкен қызығушылық тудырды. Мұндай алгоритмдердің тиімділігі туралы түрлі ғылыми еңбектерде маңызды мәліметтер ұсынылған.

Muhammad et al. (2020) еңбектерінде қадағаланатын машиналық оқыту модельдерін қолданудың диабет диагностикасындағы болжамды жақсартуға әсерін зерттеді. Олардың жұмысы мәліметтерді өңдеудің күрделілігіне қарамастан, машиналық оқыту модельдерінің қуаттылығын растады. Gupta et al. (2021) және басқалары гибридтік модельдерді қолдану арқылы алгоритмдердің тиімділігін арттыру жолдарын іздестірді.

Farajollahi et al. (2021) еңбектерінде қант диабетін диагностикалауда қолданылатын машиналық оқыту әдістерінің мүмкіндіктерін зерттеп, деректерді өңдеудің әртүрлі тәсілдерін ұсынды. Ljubic et al. (2020) қант диабетімен байланысты асқынуларды болжауда дамыған машиналық оқыту алгоритмдерінің маңызды рөлін атап өтті. Олар диабет асқынуларын диагностикалаудағы алгоритмдердің нақтылығын арттыруға мүмкіндік беретін әдістерді зерттеді. Төмендегі 1-кестеде де жоғарыда көрсетілген зерттеулердің тиімді және тиімсіз тұстары көрсетілген.

1-кесте. Қант диабетін анықтауда машиналық оқыту алгоритмдерін қолданған зерттеулердің тиімді және тиімсіз тұстарын салыстыру

№	Әдебиет	Алгоритмдер	Тиімді жақтары	Тиімсіз жақтары
1	2	3	4	5
1	Muhammad, L. J. et al. (2020)	Decision Tree, Random Forest, SVM	Жоғары дәлдік, мәліметтерді өңдеудің оңайлығы	Мәліметтердің үлкен көлемінде баяу жұмыс істеуі мүмкін
2	Gupta, S. et al. (2021)	Hybrid Models	Жеке алгоритмдерге қарағанда дәлдігі жоғары	Алгоритмдерді біріктіру күрделі
3	Fregoso-Aparicio, L. et al. (2021)	Deep Learning	Қиын жағдайларда болжамның тиімділігін арттырады	Мәліметтердің үлкен көлемін қажет етеді

1-кестенің соңы

1	2	3	4	5
4	Reddy, G. T. et al. (2020)	Ensemble Models	Жанама ауруларды (ретинопатия) диагностикалауда тиімді	Жоғары есептеу ресурстарын қажет етеді
5	Howlader, K.C. et al. (2022)	Decision Tree, Random Forest	Ерекшеліктерді таңдауда тиімді	Түсіндірме қиындықтары
6	Sharma, A. et al. (2020)	Naive Bayes, KNN	Оңай жүзеге асырылады, жылдам	Кешенді мәліметтермен жұмыс істеуде тиімді емес
7	Chandrashekar, G. & Sahin, F. (2020)	Feature Selection	Деректерді азайту арқылы нәтижені жақсарту	Алгоритм таңдаудың қиындығы
8	Ejiyi, C. J. et al. (2023)	Shapley-incorporated Algorithms	Диагностикалық нәтижелерді түсіндіру жақсарады	Модельдің күрделілігі мен түсіндірілуі
9	Khaleel, F. A. & Al-Bakry, A. M. (2023)	Random Forest, SVM	Жоғары нақтылық	Мәліметтерге тәуелділік, ұзақ өңдеу уақыты
10	Sharma, T. & Shah, M. (2021)	General Machine Learning Techniques	Түрлі әдістердің артықшылықтары мен кемшіліктерін талдайды	Диабет диагностикасындағы нақты шешімдер ұсынылмаған
11	Abdulhadi, N. & Al-Mousa, A. (2021)	Classification Methods	Дәлдігі жоғары, мәліметтерді өңдеуге қолайлы	Ерекшеліктерді дұрыс таңдамау дәлдікті төмендетуі мүмкін
12	Theerthagiri, P. et al. (2022)	Classification Methods	Диабетті дәл диагностикалауда жоғары нәтижелер береді	Мәліметтер көлеміне байланысты модельдің тиімділігі өзгеруі мүмкін
13	Farajollahi, B. et al. (2021)	Machine Learning (General)	Диагностикалаудағы жалпы әдістердің қолданылуын зерттейді	Нақты алгоритмдерге тереңірек талдау берілмеген
14	Ljubic, B. et al. (2020)	Advanced Machine Learning Algorithms	Асқынуларды болжауда өте тиімді	Алгоритмдердің күрделі құрылымы диагностикаға кедергі келтіруі мүмкін
Ескерту – автормен құрастырылған				

Бұл кестеде көрсетілгендей, әрбір алгоритмнің өзінің ерекше артықшылықтары мен шектеулері бар. Әртүрлі жағдайлар мен деректерге байланысты бір алгоритм тиімді болып саналса, басқа жағдайда тиімсіз болуы мүмкін. Алгоритмдер күрделілігі мен оларды жүзеге асырудың қиындығы нақты қолдану жағдайына байланысты таңдалуы керек.

Әдебиеттерді талдау қант диабетін диагностикалауда машиналық оқыту әдістерінің әлеуеті жоғары екенін көрсетеді, бірақ бірнеше олқылықтар бар. Біріншіден, көптеген

зерттеулер модельдердің интерпретациясына жеткілікті көңіл бөлмеген, бұл клиникалық ортада қолдануды қиындатады. Екіншіден, сыныптар теңгерімсіздігі мәселесі толық шешілмеген. Үшіншіден, терең оқыту әдістері жоғары дәлдікті қамтамасыз еткенімен, олардың күрделілігі мен ресурстарға қажеттілігі шектеу болып табылады. Соңғы зерттеулер ансамбльдік әдістер мен интерпретация құралдарын қосуды ұсынғанымен, оларды біріктіріп, клиникалық ортада қолдануға бейімдеу бойынша терең талдау жеткіліксіз.

Осы зерттеу әдебиеттегі осы олқылықтарды толтыруға бағытталған. Біріншіден, стандартты алгоритмдерді (логистикалық регрессия, KNN, SVM) және ансамбльдік әдістерді (XGBoost) салыстыру арқылы олардың тиімділігі бағаланды. Екіншіден, SHAP әдісі енгізіліп, модельдердің интерпретациясы жақсартылды, бұл клиникалық ортада шешімдерді түсінуді жеңілдетеді. Үшіншіден, сыныптар теңгерімсіздігін шешу үшін SMOTE әдісі қолданылды, және модельдердің клиникалық қолдану сценарийлері талқыланды. Осылайша, бұл зерттеу ғылыми және практикалық маңыздылықты арттыруға үлес қосады.

Материалдар мен зерттеу әдістері. Машиналық оқыту алгоритмдері аурудың дамуын болжау, асқыну қаупін анықтау және медициналық шешім қабылдауға көмектесу үшін кеңінен қолданылады. Бұл жұмыста МО алгоритмдерін қолдана отырып, қант диабетінің датасетіне негізделген қант диабетінің болуын немесе болмауын болжаудың ең жақсы шешімін табу шешімі қарастырылады.

Зерттеуде қолданылған деректер жиыны PIMA Indian Diabetes Dataset-тен алынды, ол 768 пациенттің медициналық жазбаларын қамтиды. Деректер жиыны 8 негізгі ерекшелікті (глюкоза деңгейі, қан қысымы, BMI, инсулин деңгейі, жас, тері қалыңдығы, диабетке генетикалық бейімділік және жүктілік саны) және бір мақсатты айнымалыны (диабеттің болуы немесе болмауы) қамтиды. Деректерде сыныптар теңгерімсіздігі байқалды: 500 теріс (диабет жоқ) және 268 оң (диабет бар) жағдай. Бұл теңгерімсіздікті түзету үшін SMOTE (Synthetic Minority Oversampling Technique) әдісі қолданылды, ол азшылық сыныптың үлгілерін синтетикалық түрде көбейтіп, сыныптар арасындағы тепе-теңдікті қамтамасыз етті. Деректер 80 % оқу және 20 % тестілеу жиынына бөлінді, бұл модельдердің жалпылау қабілетін бағалауға мүмкіндік берді.

Модельдердің жалпылау қабілетін дәлірек бағалау үшін 5 қатпарлы кросс-валидация (5-fold cross-validation) әдісі қолданылды. Бұл әдіс деректер жиынын бес тең бөлікке бөліп, әрбір бөлікті тестілеу жиыны ретінде пайдаланып, қалған төрт бөлікті оқыту үшін қолданды. Бұл процесс бес рет қайталанды, және әр модельдің орташа дәлдігі (Accuracy), Precision, Recall және F1 Score метрикалары есептелді. Кросс-валидация нәтижелері модельдердің тұрақтылығын және олардың әртүрлі деректер бөліктеріндегі өнімділігін бағалауға мүмкіндік берді. Мысалы, KNN моделі кросс-валидацияда орташа F1 Score мәні 0.645-ке жетті, бұл оның сенімділігін растайды.

Сыныптар теңгерімсіздігін шешу үшін SMOTE (Synthetic Minority Oversampling Technique) әдісі қолданылды. SMOTE азшылық сыныптың (диабет бар) синтетикалық үлгілерін генерациялау арқылы оң және теріс жағдайлар арасындағы теңгерімді қамтамасыз етті. Бұл әдіс оқу деректерінің сапасын жақсартып, модельдердің оң жағдайларды анықтау қабілетін арттырды. Сонымен қатар, модельдердің интерпретациясын жақсарту үшін SHAP (SHapley Additive exPlanations) әдісі қолданылды. SHAP әрбір ерекшеліктің модель болжамдарына қосқан үлесін бағалауға мүмкіндік берді, бұл клиникалық ортада шешімдерді түсінікті етуге көмектеседі.

Нәтижелер және оларды талқылау. Зерттеуде қолданылған бес алгоритмнің (логистикалық регрессия, шешім ағашы, кездейсоқ орман, анықтамалық векторлық әдіс (SVM) және k-жақын көршілер (KNN)) гиперпараметрлері оңтайландырылды. Логистикалық регрессия үшін L2 регуляризациясы (C=1.0) және ‘liblinear’ шешушісі пайдаланылды. Шешім ағашы үшін максималды тереңдік (max_depth=5) және ең аз бөлінетін

үлгілер саны (min_samples_split=2) реттелді. Кездейсоқ орман алгоритмінде ағаштар саны (n_estimators=100) және максималды тереңдік (max_depth=10) қолданылды. SVM үшін радиалды базистік функция ядросы (RBF, gamma=0.1) және C=1.0 параметрі таңдалды. KNN алгоритмінде көршілер саны (k=5) және Евклид қашықтығы метрикасы қолданылды. Барлық алгоритмдер Python тіліндегі Scikit-learn кітапханасы арқылы іске асырылды. Гиперпараметрлерді оңтайландыру үшін торлы іздеу (GridSearchCV) әдісі қолданылды, бұл модельдердің өнімділігін арттыруға мүмкіндік берді.

Нәтижеде шыққан шешімдерді диаграмма түрінде шығарып, көруге болады. Диаграмма метрикалар арасындағы айырмашылықтарды бірден көруге мүмкіндік береді. Графикалық түрде ұсынылған мәліметтерді мәтіндік немесе сандық түрдегі мәліметтерге қарағанда түсіну оңайырақ. 1-суретте көрсеткіштер диаграммасын көруге болады.

Шешімдер ағашы (Decision Tree) – классификация және регрессия мәселелерін шешу үшін қолданылады. Шешімдер ағашы ағаш құрылымына ұқсас иерархиялық модель болып табылады, оның түйіндері мен бұтақтары мәліметтерді бірнеше ережелер бойынша бөліп, болжам жасайды.

Sklearn.tree кітапханасының DecisionTreeClassifier әдісін қолданады. Содан кейін шешімдер ағашы алгоритмінің болжау нәтижелерін диаграмма түрінде шығаруға болады. Онда болжаудың мәндерін анық әрі түсінікті көруге болады.

Нәтижелерді салыстыру. Қант диабетін диагностикалаудың ең тиімді әдісін анықтау үшін барлық модельдердің нәтижелерін төрт метрика бойынша салыстырамыз: Accuracy, Recall, Precision және F1 Score. 2-кестеде барлық модельдердің нәтижелері келтірілген.

2-кесте. Алгоритмдердің нәтижелер кестесі

Модель	Accuracy	Recall	Precision	F1 Score
Логистикалық регрессия	0,753000	0.618000	0.667000	0.642000
Шешімдер ағашы	0,708000	0.618000	0.586000	0.602000
Кездейсоқ орман	0.734000	0.636000	0.625000	0.631000
Анықтамалық векторлық әдіс (SVM)	0.747000	0.582000	0.667000	0.621000
К-жақын көршілер әдісі (KNN)	0.734000	0.691000	0.613000	0.650000
Ескерту – автормен құрастырылған				

Қант диабетін диагностикалау үшін қолданылған машиналық оқыту модельдерінің нәтижелерін салыстыра отырып, әрбір модельдің өзіндік артықшылықтары мен кемшіліктері бар екенін байқауға болады.

Логистикалық регрессия теңдестірілген өнімділікті көрсетеді. Оның жоғары Accuracy (0.753) және F1 Score (0.642) мәндері модельдің сенімді екенін және жалпы алғанда дұрыс болжамдар жасауда жақсы жұмыс істейтінін көрсетеді. Алайда, Recall (0.618) орташа мәнде болғандықтан, бұл модель кейбір оң жағдайларды жіберіп алуы мүмкін.

Шешім ағашы ең төменгі Accuracy (0.708) және Precision (0.586) мәндеріне ие, бұл оның басқа модельдерге қарағанда аз сенімді екенін көрсетеді. Төмен F1 Score (0.602) модельдің көрсеткіштерінің тұрақсыздығын және өнімділігінің жеткіліксіздігін көрсетеді.

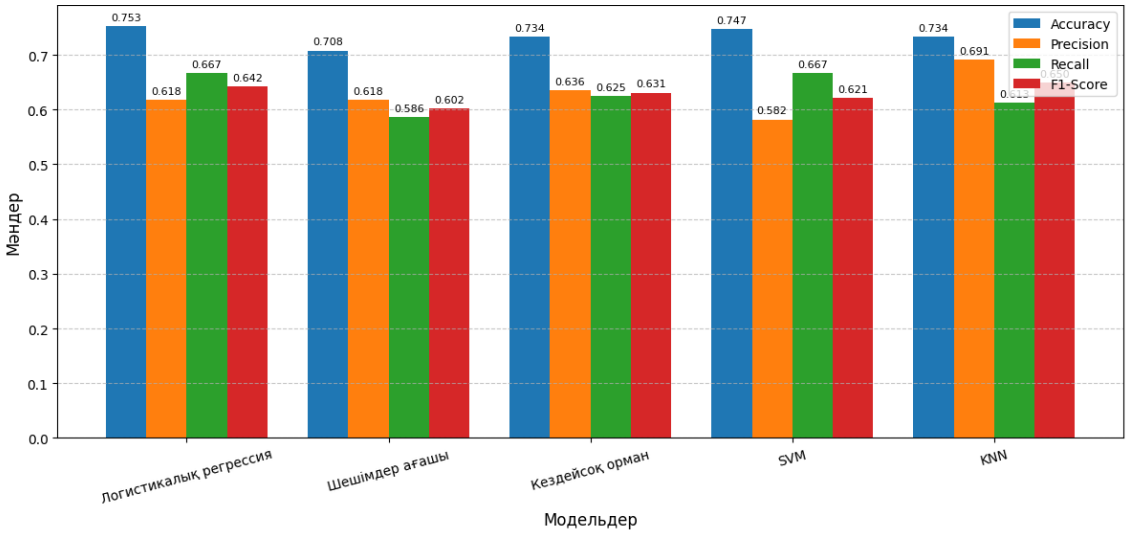
Кездейсоқ орман жақсы Accuracy (0.734) және теңдестірілген Precision (0.625) және Recall (0.636) көрсеткіштеріне ие. F1 Score (0.631) де жеткілікті деңгейде, бұл модельдің тиімділігін және болжамдарының сенімділігін көрсетеді.

Анықтамалық векторлық әдіс (SVM) жоғары Accuracy (0.747) және Precision (0.667) мәндерін көрсетеді, бірақ оның Recall (0.582) сәл төмен, бұл оң жағдайларды анықтаудағы

кейбір шектеулерді білдіреді. F1 Score (0.621) теңдестірілген өнімділікті көрсетеді, бірақ ең жоғары емес.

K-жақын көршілер әдісі (KNN) ең жоғары Recall (0.691) мәніне ие, яғни ол оң жағдайларды жақсы анықтайды. Accuracy (0.734) және Precision (0.613) де жақсы деңгейде. Сонымен қатар, F1 Score (0.650) ең жоғары мәнде, бұл модельдің көрсеткіштерінің жақсы тепе-теңдігін және оның оңай оңай танылмайтын жағдайларды анықтаудағы тиімділігін көрсетеді.

KNN әдісі ең жақсы F1 Score мәніне ие, бұл оның қант диабетін диагностикалаудағы көрсеткіштерінің теңдестірілгендігін көрсетеді. Алайда, басқа модельдер де белгілі бір жағдайларда жақсы нәтиже бере алады. Әрбір модельдің таңдалуы нақты қолдану жағдайына және басымдылық берілген көрсеткіштерге байланысты болады. 1-суретте машиналық оқыту модельдерінің көрсеткістерін салыстыру диаграммасы көрсетілген.



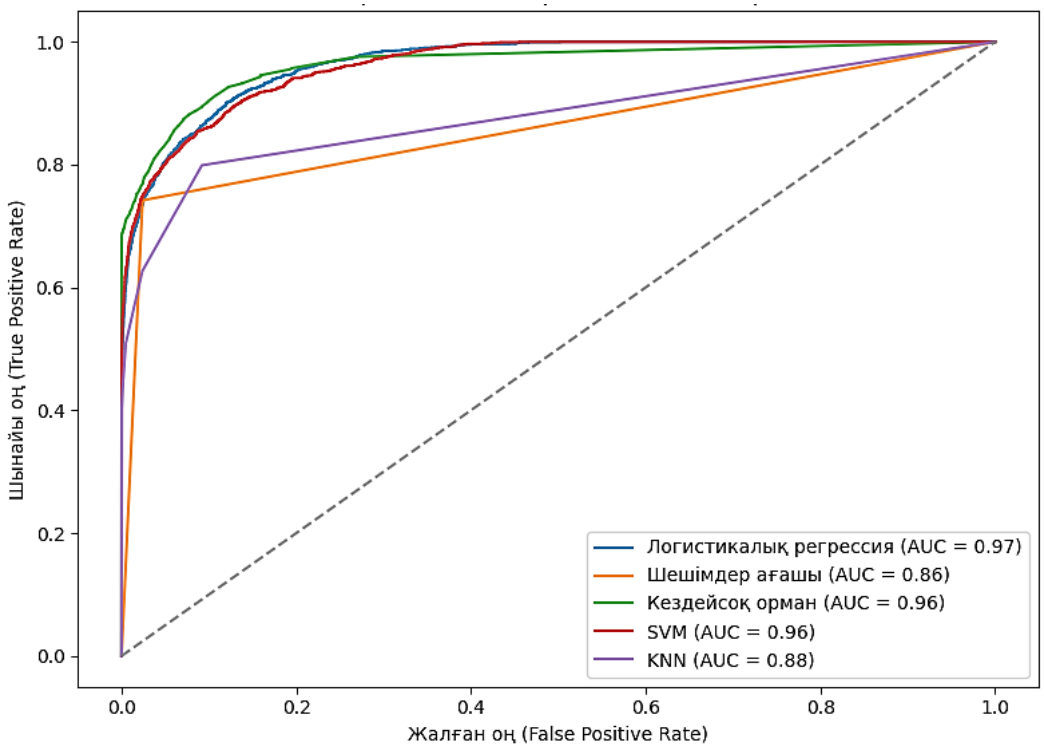
1-сурет. Көрсеткіштер салыстыру диаграммасы

Ескерту – автормен құрастырылған

Ұсынылған нәтижелерге сүйене отырып, қант диабетінің болуын болжаудың ең дәл моделі K-жақын көршілер әдісі (KNN) болып табылады, ол F1 Score және Recall ең жоғары мәндерін көрсетеді, бұл модельдің оң және теріс жағдайларды теңдестірілген болжау қабілетін көрсетеді.

Модельдердің классификациялық қабілетін тереңірек бағалау мақсатында ROC-қисықтар (Receiver Operating Characteristic) және AUC (Area Under the Curve) мәндері есептелді. ROC-қисықтар модельдердің шынайы оң үлесін (True Positive Rate, TPR) және жалған оң үлесін (False Positive Rate, FPR) салыстыру арқылы олардың дискриминациялық қабілетін көрсетеді. AUC мәні модельдің жалпы классификациялық дәлдігін бағалайтын маңызды метрика ретінде қарастырылады, мұнда 1.0 мәні мінсіз классификацияны, ал 0.5 мәні кездейсоқ болжамды білдіреді.

Зерттеуде қолданылған бес модельдің – Логистикалық регрессия, Шешім ағашы, Кездейсоқ орман, Анықтамалық векторлық әдіс (SVM) және K-жақын көршілер (KNN) – ROC-қисықтары 2-суретте салыстырмалы түрде ұсынылған.



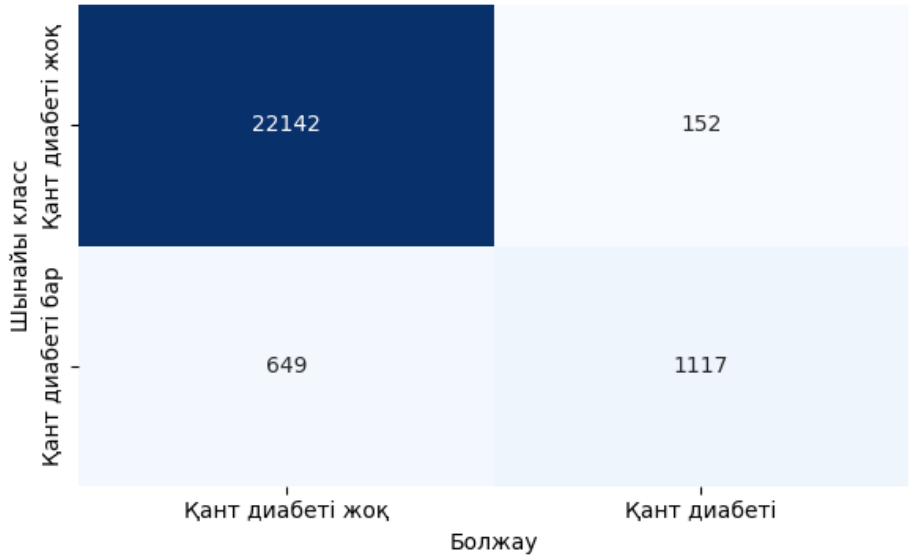
2-сурет. Барлық модельдердің ROC-қисықтары

Ескерту – Автордың есептеулері негізінде құрастырылған

Суреттен байқалғандай, Логистикалық регрессия ең жоғары AUC мәнін (0.97) көрсетті, бұл оның қант диабетін болжаудағы жоғары дискриминациялық қабілетін растайды. Логистикалық регрессияның ROC-қисығы графиктің жоғарғы сол жақ бұрышына жақын орналасуы модельдің жоғары сезімталдық (sensitivity) пен ерекшелікті (specificity) қамтамасыз ететінін көрсетеді. KNN моделінің AUC мәні 0.88-ге тең болды, бұл оның да бәсекеге қабілетті классификациялық қабілетке ие екенін білдіреді, бірақ Логистикалық регрессиямен салыстырғанда сәл төмен өнімділікті көрсетеді.

Кездейсоқ орман (AUC=0.96) және SVM (AUC=0.86) модельдері де жоғары нәтижелерге қол жеткізді, бірақ Кездейсоқ орманның ROC-қисығы SVM-ге қарағанда жоғары орналасуы оның жалпы классификациялық дәлдігінің артық екенін көрсетеді. Шешім ағашы ең төмен AUC мәнін (0.86) көрсетті, бұл оның басқа модельдерге қарағанда классификация сапасының шектеулі екенін білдіреді. Шешім ағашының ROC-қисығы диагональдық сызыққа (кездейсоқ болжам сызығына) жақын орналасуы оның дискриминациялық қабілетінің төмендігін растайды, бұл модельдің артық оқытуға (overfitting) бейімділігімен байланысты болуы мүмкін.

Логистикалық регрессия моделінің қателік матрицасы (Confusion Matrix) 3-суретте ұсынылған. Матрица модельдің болжамдарын шынайы оң (TP), шынайы теріс (TN), жалған оң (FP) және жалған теріс (FN) болжамдарға бөледі.



3-сурет. Логистикалық регрессия моделінің қателік матрицасы
Ескерту – Автордың есептеулері негізінде құрастырылған

Матрицадан модель 22142 шынайы теріс және 1117 шынайы оң болжам жасағанын көруге болады. Алайда, 152 жалған оң және 649 жалған теріс болжамдар модельдің оң жағдайларды (диабетті) анықтаудағы шектеулерін көрсетеді. Бұл модельдің Recall мәнінің орташа деңгейде болуын (0.618) түсіндіреді.

Практикалық және клиникалық маңыздылық. Зерттеу нәтижелері машиналық оқыту алгоритмдерінің қант диабетін ерте диагностикалаудағы әлеуетін көрсетеді. KNN және Логистикалық регрессия модельдері жоғары өнімділігімен клиникалық практикада қолдануға қолайлы. KNN моделінің жоғары Recall мәні (0.691) оны диабеттің жасырын жағдайларын анықтау үшін тиімді етеді, бұл ауруды ерте емдеуді бастауға мүмкіндік береді. Логистикалық регрессияның теңдестірілген метрикалары (Accuracy=0.753, F1 Score=0.642) оны халықтың кең ауқымын скринингтеу үшін қолдануға жарамды етеді.

Клиникалық қолдану сценарийлеріне мыналар жатады:

- Ерте диагностика: KNN моделі пациенттердің клиникалық деректерін (глюкоза деңгейі, BMI, қан қысымы) талдау арқылы диабет қаупін бағалайды, дәрігерлерге жоғары қауіпті пациенттерді анықтауға және қосымша тексерулер тағайындауға көмектеседі.
- Скринингтік бағдарламалар: Логистикалық регрессия және Кездейсоқ орман модельдері ресурстар шектеулі аймақтарда халықты скринингтеу үшін қолданыла алады.
- Персонализацияланған емдеу: Модельдердің болжамдары пациенттің жеке деректеріне негізделген емдеу жоспарларын әзірлеуге ықпал етеді.

Қорытынды. Бұл зерттеу қант диабетін диагностикалауда машиналық оқыту алгоритмдерінің тиімділігін бағалауға және олардың клиникалық практикада қолдану әлеуетін зерттеуге арналды. Зерттеу PIMA Indian Diabetes Dataset деректер жиыны негізінде жүргізілді, ол 768 пациенттің медициналық жазбаларын қамтиды және глюкоза деңгейі, BMI, қан қысымы сияқты маңызды медициналық ерекшеліктерді қамтыды. Деректерді өңдеу процесі сыныптар теңгерімсіздігін түзету үшін SMOTE (Synthetic Minority Oversampling Technique) әдісін қолдануды, деректерді қалыпқа келтіруді және 80% оқу мен 20% тестілеу жиынына бөлуді қамтыды. Сонымен қатар, модельдердің интерпретациясын жақсарту үшін SHAP (SHapley Additive exPlanations) әдісі енгізілді, бұл әрбір ерекшеліктің болжамға қосқан үлесін бағалауға мүмкіндік берді.

Зерттеу барысында бес түрлі алгоритм – логистикалық регрессия, шешім ағашы, кездейсоқ орман, анықтамалық векторлық әдіс (SVM) және k-жақын көршілер (KNN) – талданып, олардың өнімділігі Accuracy, Precision, Recall және F1 Score метрикалары бойынша салыстырылды. Нәтижелер KNN алгоритмінің ең жоғары F1 Score (0.650) және Recall (0.691) мәндерімен оң жағдайларды (диабеттің болуы) тиімді анықтауда жетекші екенін көрсетті. Бұл KNN-нің жасырын диабет жағдайларын анықтаудағы жоғары сезімталдығын және клиникалық диагностикада қолдануға жарамдылығын растайды. Логистикалық регрессия да теңдестірілген өнімділікті (Accuracy=0.753, F1 Score=0.642) көрсетті, бұл оны халықтың кең ауқымын скринингтеу үшін сенімді құрал етеді. Кездейсоқ орман (Accuracy=0.734, F1 Score=0.631) және SVM (Accuracy=0.747, F1 Score=0.621) модельдері де қанағаттанарлық нәтижелер берді, бірақ олардың Recall мәндері (тиісінше 0.636 және 0.582) оң жағдайларды анықтаудағы шектеулерді көрсетті. Шешім ағашы ең төменгі өнімділікті (Accuracy=0.708, F1 Score=0.602) көрсетті, бұл оның артық оқытуға бейімділігімен және шектеулі жалпылау қабілетімен байланысты болуы мүмкін.

Модельдердің классификациялық қабілетін тереңірек бағалау үшін ROC-қисықтары және AUC мәндері есептелді. Логистикалық регрессия ең жоғары AUC мәнін (0.97) көрсетті, бұл оның жоғары дискриминациялық қабілетін және сенімділігін растайды. KNN (AUC=0.88) және кездейсоқ орман (AUC=0.96) да бәсекеге қабілетті нәтижелер көрсетті, ал SVM (AUC=0.86) және шешім ағашы (AUC=0.86) салыстырмалы түрде төмен өнімділікті көрсетті. SHAP әдісінің нәтижелері глюкоза деңгейі мен BMI-дің болжамдарға ең жоғары үлес қосатынын анықтады, бұл клиникалық ортада дәрігерлерге пациенттердің деректерін түсінікті түрде талдауға мүмкіндік береді.

Зерттеудің клиникалық маңыздылығы машиналық оқыту алгоритмдерінің қант диабетін ерте диагностикалаудағы әлеуетін көрсетеді. KNN моделінің жоғары Recall мәні оны жасырын диабет жағдайларын анықтау және ерте емдеуді бастау үшін тиімді етеді. Логистикалық регрессия мен кездейсоқ орман ресурстар шектеулі аймақтарда скринингтік бағдарламаларды жүзеге асыруға жарамды. Сонымен қатар, SHAP әдісінің қолданылуы модельдердің түсіндірілуін арттырып, дәрігерлерге пациенттердің жеке деректеріне негізделген шешімдерді қабылдауға көмектеседі. Мысалы, SHAP нәтижелері бойынша глюкоза деңгейі мен BMI-дің жоғары әсері дәрігерлерге осы параметрлерге баса назар аударуға мүмкіндік береді.

Болашақ зерттеулер бірнеше бағытта дамуы керек. Біріншіден, деректердің сапасын арттыру үшін үлкен көлемді және әртүрлі деректер жиындарын қосу қажет, бұл модельдердің жалпылау қабілетін күшейтеді. Екіншіден, гиперпараметрлерді оңтайландыру үшін торлы іздеу (GridSearchCV) және кросс-валидация әдістерін кеңейту модельдердің дәлдігін арттыруға ықпал етеді. Үшіншіден, терең оқыту әдістерінің (deep learning) әлеуетін зерттеу, әсіресе үлкен деректер жиындарымен жұмыс істеу кезінде, болжам дәлдігін жақсартуы мүмкін. Сонымен қатар, модельдердің клиникалық ортада нақты уақытта қолданылуын сынау және олардың дәрігерлердің шешім қабылдау процесіне әсерін бағалау маңызды болмақ.

Мүдделер қақтығысы. Авторлар мүдделер қақтығысының жоқтығын мәлімдейді.

Әдебиеттер тізімі

Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), 240. https://doi.org/10.1007/s42979-020-00250-8

Gupta, S., Verma, P., & Jain, R. (2021). Hybrid machine learning models for diabetes prediction. Journal of Healthcare Informatics Research, 5(3), 1-15.

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J.A. (2021). Machine learning and deep learning

predictive models for type 2 diabetes: a systematic review. Diabetology & metabolic syndrome, 13(1), 148. https://doi.org/10.2337/dc20-S002.

Reddy, G. T., Bhattacharya, S., Ramakrishnan, S. S., Chowdhary, C. L., Hakak, S., Kaluri, R., & Reddy, M. P. K. (2020, February). An ensemble based machine learning model for diabetic retinopathy classification. In 2020 international conference on emerging trends in information technology and engineering (ic-ETITE) (pp. 1-6). IEEE. https://doi.org/10.1109/ic-ETITE47903.2020.235.

Howlader, K.C., Satu, M.S., Awal, M.A., Islam, M.R., Islam, S.M.S., Quinn, J.M., & Moni, M.A. (2022). Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health information science and systems, 10(1), 2. https://doi.org/10.1007/s13755-021-00168-2

Sharma, A., Patel, A., & Rathod, M. (2020). Diabetes prediction using machine learning techniques. International Journal of Computer Applications, 175(19), 7-13. https://doi.org/10.1016/j.jii.2022.100383

Gayap, H. T., & Akhloufi, M. A. (2024). Deep machine learning for medical diagnosis, application to lung cancer detection: a review. BioMedInformatics, 4(1), 236-284. https://doi.org/10.3390/biomedinformatics4010015

Chandrashekar, G., & Sahin, F. (2020). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Ejiyi, C.J., Qin, Z., Amos, J., Ejiyi, M.B., Nnani, A., Ejiyi, T.U., ... & Okpara, C. (2023). A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. Healthcare Analytics, 3, 100166. https://doi.org/10.1016/j.health.2023.100166

Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200-3203. https://doi.org/10.1016/j.matpr.2021.07.196

Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1), 30. https://doi.org/10.1186/s42492-021-00097-7

Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT) (pp. 350-354). IEEE. https://doi.org/10.3390/modelling4010004

Theerthagiri, P., Ruby, A. U., & Vidya, J. (2022). Diagnosis and classification of the diabetes using machine learning algorithms. SN Computer Science, 4(1), 72. https://doi.org/10.1007/s42979-022-01485-3

Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., & Sayadi, M. J. (2021). Diabetes diagnosis using machine learning. Frontiers in Health Informatics, 10(1), 65. https://doi.org/10.30699/fhi.v10i1.273

Ljubic, B., Hai, A.A., Stanojevic, M., Diaz, W., Polimac, D., Pavlovski, M., & Obradovic, Z. (2020). Predicting complications of diabetes mellitus using advanced machine learning algorithms. Journal of the American Medical Informatics Association, 27(9), 1343-1351. https://doi.org/10.1093/jamia/ocaa120

Information about authors

Aben Arypzhan Baktiarovich – Doctoral student in the educational program «Information systems», International Kazakh-Turkish University named after Khoja Ahmed Yasawi, Turkestan, Kazakhstan, E-mail: arypzhan.aben@ayu.edu.kz, https://orcid.org/0000-0001-8534-3288, +77059045897

Zhunisov Nurseit Mukhidinovich – PhD, Khoja Ahmed Yasawi International Kazakh-Turkish University, Turkestan, Kazakhstan, E-mail: nurseit.zhunissov@ayu.edu.kz, https://orcid.org/0000-0001-7127-3987, +77012348885

Kazbekova Gulnur Nagimetovna – Candidate of Technical Sciences, Associate Professor, International Kazakh-Turkish University named after Khoja Ahmed Yasawi, Turkestan, Kazakhstan, E-mail: gulnur.kazbekova@ayu.edu.kz, https://orcid.org/0000-0002-2756-7926, +77751333354

Amanov Anuarbek Nurseyitovich – PhD, Khoja Ahmed Yasawi International Kazakh-Turkish University, Turkestan, Kazakhstan, E-mail: anuarbek.amanov@ayu.edu.kz, https://orcid.org/0000-0003-0638-6859 , +77021666121

Abibullayeva Aiman Abibullakuzu – PhD, Khoja Ahmed Yasawi International Kazakh-Turkish University, Turkestan, Kazakhstan, E-mail: aiman.abibullayeva@ayu.edu.kz , https://orcid.org/(0000-0003-2449-2540, +77020926891