ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

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

DOI 10.51885/1561-4212_2023_1_109
МҒТАА 50.41.29

А.Д. Кубегенова[1], К.Т. Искаков[2], Е.С. Кубегенов[3], О.И. Криворотько[4]
[1]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал қ., Қазақстан
*E-mail: aigul-03@mail.ru**
[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан
E-mail: kazizat@mail.ru
[3]Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті, Орал қ., Қазақстан
E-mail: erlando78@mail.ru
[4]Новосібір мемлекеттік университеті, Новосібір қ., Ресей
E-mail: krivorotko.olya@mail.ru

ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ АРҚЫЛЫ ӘЛЕУМЕТТІК МАҢЫЗЫ БАР АУРУЛАР МОДЕЛІН ЖӘНЕ КЛАСТЕРЛЕУ АЛГОРИТМІН ҚҰРУ

ПОСТРОЕНИЕ МОДЕЛИ СОЦИАЛЬНО-ЗНАЧИМЫХ ЗАБОЛЕВАНИИ И АЛГОРИТМА КЛАСТЕРИЗАЦИИ В ИНТЕЛЛЕКТУАЛЬНОМ АНАЛИЗЕ ДАННЫХ

BUILDING A MODEL OF SOCIALLY SIGNIFICANT DISEASES AND A CLUSTERING ALGORITHM IN DATA MINING

Аңдатпа. *Қазіргі кезде ақпаратты технология қоғамдағы өзекті және маңызды, сұранысқа ие міндеттердің бірі, үлкен деректер көлемінің ішінен пайдалы және мағыналы ақпаратты анықтау болып келеді. Ең маңызды аспектілердің бірі интеллектуалды деректерді талдау. Интеллектуалды деректерді талдау – бұл көптеген мәліметтері бар, кейбір үлгілерді анықтау мен және шығару процесі. Деректерді талдаудың ондаған, тіпті жүздеген әртүрлі әдістері бар, маркетингтік зерттеулермен статистиканың ақпараттық құралдар жинағынан құралған. Жалпы деректерді өндіру, денсаулық сақтау саласында деректерді талдау, шешім қабылдау немесе оқиғаларды ерте анықтау үшін математикалық модельдерді қолдану кең таралған. Осыған байланысты зерттеу жұмысында деректерді интеллектуалды талдау технологиясы, машиналық оқыту әдістері және кластерлік талдау негізінде есептерді сәйкестендіруді реттеу жүргізіледі, әдістерді, алгоритмдерді және нәтижелерді қолдану тұрғысынан осы салаға қатысты әртүрлі дереккөздерден мәлімет алынып қарастырылды. Қазақстанда әлеуметтік маңызы бар АИТВ-инфекциясының таралуының математикалық моделі үшін сандық шешім алгоритмдерін сипаттайды. АИТВ бойынша жағдайды модельдеуде Data mining технологиясы ерекше өзекті болып табылады, өйткені оның негізінде Қазақстанда және ел өңірлерінде сырқаттанушылықтың қысқа мерзімді болжамының карталары жасалады. Соңғы 10 жылда Қазақстанда (2010-2020 жж.) АИТВ-инфекциясының таралуы бойынша статистикалық деректер алынды. АИТВ жұқтырған адамдар туралы мәлімет алынып, анықтап және олардың жағдайын талдау үшін Data Mining технологиясының көмегімен жіктеу әдістері қолданылды. Халықтың аурушаңдық көрсеткіштерін зерттеу статистикалық деректер негізінде алынып Statistica және* SPSS *қолданбалы бағдарламалар пакетін пайдалана отырып жүргізілді.*

Түйін сөздер: *Кластерлік талдау, дендрограмма, Data Mining, медиан, мода.*

Аннотация. *В настоящее время информационная технология является одной из актуальных и важных, востребованных задач в обществе, выявление полезной и значимой информации из большого объема данных. Одним из наиболее важных аспектов является анализ интеллектуальных данных. Интеллектуальный анализ данных – это процесс выявления и извлечения некоторых образцов с большим количеством данных. Существуют десятки или даже сотни различных методов анализа данных, составленных из набора информационных инструментов статистики с маркетинговыми исследованиями. В области общего интеллектуального анализа данных, здравоохранения широко распространено использование математических моделей для анализа данных, принятия решений или раннего выявления событий. В связи с этим в исследовательской работе проводится регулирование идентификации задач на основе технологии интеллектуального анализа данных, методов машинного обучения и кластерного анализа, получены данные из различных источников, относящихся к данной области, с точки зрения применения методов, алгоритмов и результатов. Характеризует алгоритмы численного решения математической модели распространения социально значимой ВИЧ-инфекции в Казахстане. В моделировании ситуации по ВИЧ особенно актуальна технология Data mining, так как на ее основе составляются карты краткосрочного прогноза заболеваемости в Казахстане и регионах страны. Последние 10 лет в Казахстане (2010-2020 гг.). Получены статистические данные по распространенности ВИЧ-инфекции. Для получения и выявления данных о ВИЧ-инфицированных и анализа их состояния использовались методы классификации с помощью технологии Data Mining. Исследование показателей заболеваемости населения проводилось на основе статистических данных с использованием пакета прикладных программ Statistica и SPSS.*

Ключевые слова: *Кластерный анализ, дендрограмма, Data Mining, медиана, мода.*

Abstract: *Currently, information technology is one of the urgent and important, in-demand tasks in society, identifying useful and meaningful information from a large volume of data. One of the most important aspects is the analysis of intellectual data. Data mining is the process of identifying and extracting some samples with a large amount of data. There are dozens or even hundreds of different methods of data analysis compiled from a set of statistical information tools with marketing research. In the field of general data mining, healthcare, the use of mathematical models for data analysis, decision-making or early detection of events is widespread. In this regard, the research work regulates the identification of tasks based on data mining technology, machine learning methods and cluster analysis, data from various sources related to this field are obtained from the point of view of the application of methods, algorithms and results. Characterizes the algorithms of numerical solution of the mathematical model of the spread of socially significant HIV infection in Kazakhstan. Data mining technology is particularly relevant in modeling the HIV situation, as it is used to make maps of the short-term prognosis of morbidity in Kazakhstan and the regions of the country. For the last 10 years in Kazakhstan (2010-2020), statistical data on the prevalence of HIV infection have been obtained. To obtain and identify data on HIV-infected people and analyze their condition, classification methods using Data Mining technology were used. The study of morbidity indicators of the population was carried out on the basis of statistical data using the Statistica application software package.*

Keywords: *Cluster analysis, dendrogram, Data Mining, median, mode.*

Kipicne. Қазіргі уақытта денсаулық сақтаудың жаһандық және маңызды әлеуметтік мәселелерінің бірі АИТВ-инфекциясы пандемия сипатына ие болып отыр. Бүкіл әлемдегі жұқпалы аурулар адам денсаулығы мен халықаралық қауіпсіздікке үлкен қауіп төндіреді. Эпидемиялық таралуға бейім инфекциялар барлық елдерде үлкен қоғамдық алаңдаушылық тудырады. Қазақстанда да АИТВ жұқтырғандардың өсуіне байланысты қоғамдық денсаулық сақтауда жедел шешуді және көлемді деректерді талдауды талап ететін мәселелер жиі туындауда. Бұл аурудың қауіптілік сипаты, ең алдымен, еліміздің жас ұрпақтын және еңбекке қабілетті азаматтарды зақымдануынан тұрады.

ЮНЭЙДС (2021) сарапшыларының деректері бойынша АИТВ-мен өмір сүретін адамдардың жалпы әлемдік саны 38,4 млн астам адамды құрады [33,9 млн – 43,8 млн].

ЖҚТБ-мен байланысты аурулардан қайтыс болған адамдардың саны 650 000 құрады [510,000-860,000] адам [1].

Біріккен БҰҰ бағдарламасының мәліметтері АИТВ пандемиясының салыстырмалы тұрақтылығын көрсетсе де, ауру деңгейі әлі де жоғары. 2019-2021 жылдардағы коронавирустық инфекция пандемиясы бақылаусыз сипатқа ие инфекциялар қаупінің айқын мысалы болып табылады және барлық мемлекеттердің бұрын-соңды болмаған шараларымен еңсерілген міндеттерді бүкіл адамзатқа көрсетті.

АИТВ-инфекциясының проблемасы таралудың басқа сипатына ие болса да, өте қауіпті болып қала береді, өйткені емдеудің жетістіктері ремиссия кезеңіндегі науқастардың түпкілікті қалпына келуіне әкелген жоқ.

ЖИТС-тың алдын алу және оған қарсы күрес жөніндегі республикалық орталықтың деректері бойынша 30.09.2020 ж. өсу қорытындысымен АИТВ-инфекциясының 27 100 жағдайы тіркелді, оның ішінде ерлер – 16344, әйелдер – 10756, балалар – 146. Бұдан басқа, елімізде АИТВ оң статусы бар әйелдерден туылған 4464 бала тіркелген [1].

БҰҰ-ның ВИЧ/СПИД жөніндегі біріккен бағдарламасының (ЮНЭЙДС) жаңа стратегиясы 2030 жылға қарай әлемде ЖҚТБ індетін тоқтатуға міндеттеме алды.

Бұл Қазақстан Республикасының Денсаулық сақтау жүйесін дамытудың 2016-2030 жылдарға арналған «Денсаулық» мемлекеттік бағдарламасында көрініс тапты.

Бұл мәселенің өзектілігі Қазақстанда, әсіресе жұқтыру қаупі жоғары халық топтарында АИТВ-инфекциясы эпидемиясының сипатын зерттеу қажеттілігінен туындады. Эпидемиологиялық аурудың дамуын болдырмау үшін популяциядағы ауруды алдын-ала анықтауға мүмкіндік беретін терең талдау әдістері қолданылады. Қазақстан медицинасы оның қызметінің барлық салаларына статистикалық өңдеуді енгізу қажеттілігін түсінді. Алайда, статистикалық өңдеу құралдарының кеңінен енгізілуімен сапалы талдау ғана емес, сонымен қатар деректерді визуализациялау процестерін егжей-тегжейлі және терең зерттеу қажеттілігі туралы түсінік пайда болды.

Статистикалық талдауға арналған бағдарламалар пакетін білу ғана емес, сонымен қатар оларды әр нақты жағдайға нақтылау қажет.

Медициналық зерттеулер жүргізудегі зерттеушінің маңызды міндеті – деректерді статистикалық талдаудың нақты әдісін таңдау.

Медициналық ақпарат жүйесінің ауқымы мен күрделілігі күрт өсті және оны дамыту мен басқару жөніндегі қызметті бақылау қиын. Математикалық статистиканың дәстүрлі әдістері мен қарапайым әдістері саласында мәліметтер мен ақпараттың қарқынды өсуінен туындаған мәселелерді шешу қиын, бұл медициналық ақпараттық қызмет жүйесін басқаруға теріс әсер етеді. Сондықтан бағдарламалық жасақтаманы әзірлеу мен техникалық қызмет көрсетуді басқару үшін бағдарламалық жасақтама деректерін жинау өте маңызды болып келеді.

Компьютерлік және ақпараттық технологиялардың, сондай-ақ сақтау технологиясының қарқынды дамуымен көптеген деректерді сақтауға болады [2]

Әдеби шолу. Деректерді іздеу технологиясы көптеген мәліметтерден потенциалды керек білімді іздеп және шығара алады. Деректер қоры технологиясы – бұл мәліметтер базасын басқаратын бағдарламалық жасақтама туралы ғылым. Деректер базасындағы деректерді құрылымдау, жобалау және қолдану әдістерін зерттеу арқылы талданады [3]. Ақпараттық технологияның қарқынды дамуымен мәліметтер базасының ауқымы, көлемі және тереңдігі кеңеюде, «бай деректер мен жаман ақпарат» деген ұғымға сәйкес келеді [4].

Деректерді іздеу дегеніміз – мәліметтер шаблонын іздеу процесі, яғни көптеген толық емес, анық емес, кездейсоқ мәліметтерден алынған мәліметтермен жұмыс жасау [5].

Деректерді іздеу – бұл мәліметтер базасы мен жасанды интеллект саласындағы өте

белсенді зерттеу саласы [6-8].

Компьютерлік деректерді іздеу технологиясын әзірлеуге және қолдануға көп көңіл бөлу керек, өйткені деректерді іздеу технологияларын қолдана отырып, біз тұрақты дамуға ықпал ететін тиімді стратегияларды біріктіреміз [9].

Деректерді іздеу технологиясы тану параметрлерін және коэффициенттерді таңдауды егжей-тегжейлі талдайды, содан кейін деректерді іздеу моделі шығарылады [10].

Пациенттердің үлкен анонимді деректерін талдау үшін авторлар жағдайды өңдеу және құрылымдау технологиясына негізделген әдісті қолдануды ұсынады. Осы әдісті, арнайы модельді қолдана отырып әр жағдайда негізгі ақпаратты дәл және тиімді алуға болады [11].

Сандық векторлар жиынтығы ретінде, сипатталған әдістерге сәйкес кластерлерге топтастырылып Хопкинс статистикасының мәнімен есептеуге, кластерлеуді sklearn кітапханасының құралдарын қолдану арқылы жүзеге асыруға болады. Кластерлеудің екі түрлі K-Medium, АВТО-конфигурациясы бар тығыздыққа негізделген әдісін қолданылады. Салыстыру жағдайында кластерлік құрылым бір алгоритмнің әртүрлі параметрлерін өзгерту арқылы бағаланып (мысалы, k топтарының саны); алынған және дайындалған объектілерде модель (немесе бірнеше) құрылады және оның параметрлері түзетіліп, тестілеу және нәтижелерді талдау жүргізіледі [12].

Эпидемиологияның математикалық моделінің мысалы (АИТВ және туберкулездің коинфекциясы) математикалық модельдердің сәйкестендірілуіне арналған зерттеулерді көрсетеді [13]. Модель параметрлерін анықтау міндеті квадраттық мақсатты функцияны азайтуға дейін азаяды.

Сызықтық емес жүйелер қарастырылғандықтан, эпидемиологияның кері есептерін шешу екі түрлі болуы мүмкін, сондықтан кері есептердің сәйкестендірілуін талдау тәсілдері сипатталған. Бұл тәсілдер белгісіз параметрлердің қайсысын (немесе олардың комбинациясын) қолда бар Қосымша ақпарат бойынша бір мәнді және тұрақты қалпына келтіруге болатындығын анықтауға мүмкіндік береді [14].

Эпидемиологиялық модельдің коэффициенттері халықтың ерекшеліктері мен аурудың дамуын сипаттайды. Эксперементталдық деректерден статистикалық деректерге квадраттың ауытқуы болып сипатталатыны, математикалық модельдегі параметрлерді сәйкестендірудің кері есебінің функциясының төмендеу болып келеді. Статистикалық және оңтайландыру алгоритмдерінің жиынтығы 30 % салыстырмалы дәлдікпен параметрлерді сәйкестендіруді көрсетеді. Нәтижелерді Денсаулық сақтау ұйымдары модельдеу деректерін тарихи деректермен салыстыру арқылы белгілі бір аймақтағы жұқпалы аурулардың эпидемиясын болжау үшін қолдана алады [15].

Деректерді талдау құралы ретінде Statistica қолданбалы бағдарламалар пакеті қолданылды: StatisticaBase, StatisticaAdvanced, Data Mining. Графикалық формаларды қолдану арқылы деректерді кластерлеу талдау уақытын қысқартуға, сондай-ақ сырқаттанушылықты болжау алгоритмін әзірлеуге мүмкіндік берді [16].

Медициналық ақпаратты талдау үшін статистика әдістерін қолдану қазіргі уақытта Қазақстанда жеткілікті кең таралмаған, сондықтан біздің зерттеулеріміздің мақсаты Data Mining технологиясы бойынша эпидемиологиялық жағдайды талдау, болжау және алдын ала анықтау болып табылады.

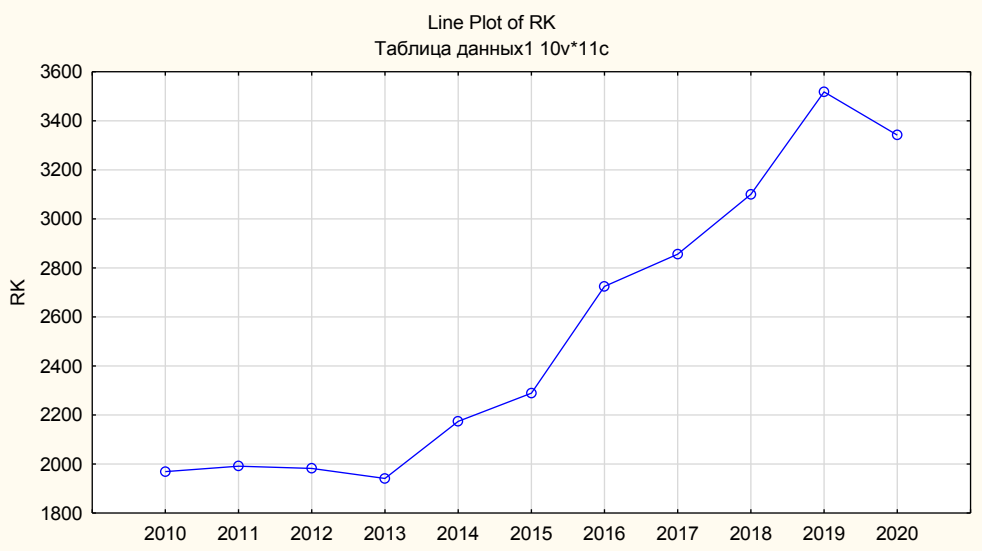
Материалдар және зерттеу әдістері. Зерттеу жұмысында Қазақстан Республикасында 2010-2020 жылдар аралығындағы АИТВ инфекциясын жұқтырған туралы деректер алынды. Алынған мәліметтерді жіктеу Data Mining технологиясының көмегімен талдау жүргізілді. Деректерді талдау құралы ретінде Statistica және SPSS. қолданбалы бағдарламалар пакеті қолданылды: Statistica Base, StatisticaAdvanced, деректерді өндіруге

арналған Data Mining технологиясының құралдары.

Кластерлеудің жаңа әдістері бір байланыс әдісіне негізделген графикалық формалардың көмегімен талдау жасауға мүмкіндік берді. Графикалық формаларды қолдану арқылы деректерді кластерлеу талдау уақытын қысқартуға, сондай-ақ сырқаттанушылықты болжау алгоритмін әзірлеуге мүмкіндік берді.

Кластерлік талдауды деректерге қолданудың практикалық маңыздылығы мен өзектілігі күмән тудырмайды, өйткені қазіргі ақпараттық қоғамда деректер мен оларды талдау нәтижелері үлкен рөл атқарады, ал кластерлеу бұл деректерді жақсы түсінуге мүмкіндік береді.

Нәтижелері және оларды талқылау. Эксперименттік мәліметтерді өңдеу компьютерде статистикалық пакеттерде жүргізілді.



1-сурет. Қазақстан Республикасының 2010-2020 жылдарға арналған сызықтық графигі

10 жылдық кезеңнің жиынтық деректерін ескере отырып, АИТВ-инфекциясының сырқаттанушылардың сызықтық графигі құрылды 2010-2020 жыл аралығында (1-сурет).

АИТВ ауруын жұқтырған абсцисса осі бойынша, зерттеу жылдары берілген, координаталар осі бойынша – АИТВ жұқтырғандардың абсолюттік саны келтірілген. (100 000 адам шаққанда). Бұл графикте аурудың 2010-2013 жылдар аралығында тұрақты тенденциясын көрсетеді. 2014 жылдан бастап аурудың нәтижесі екі есе көбейіп нашарлағаны байқалады. 2019 жылы зерттеудің алғашқы жылдарымен салыстырғанда халықтың ауруы бірнеше есеге артып күшейгенің көреміз. 2020 жылдары аурудың аздап төмендегенін байқаймыз, бұл көрсеткішті жаңа короновирустық инфекция пандемиясы кезеңінде ақпарат жинау жүйесінің нашарлануымен, оның өлім түріндегі салдарымен түсінуге болады. Осылайша, Қазақстан Республикасында АИТВ-жұқпасымен сырқаттанушылықтың көпжылдық серпінін бағалау кезінде 2013-2019 жылдары жылдам көтерілу және 2019-2020 жылдар аралығында төмендеу анықталады. Тігінен берілген жолында 1,3 %-ға дейін төмендеуі бұл тенденция абсолютті дегенді білдірмейді, өйткені онда ауытқу бар – нәтиже жақсарғанымен біртіндеп нашарлағаның байқаймыз. АИТВ-жұқтырғандардың сырқаттанушылық деңгейі бойынша сызықтық кестені талдау негізінде үш тапқа бөліп көрсетуге

болады:

1) орташа көтерілу жылдары (2010-2013);
2) жоғары көтерілу жылдары (2013-2019);
3) құлдырау жылдары (2019-2020);
4) аралық жылдар (2014, 2016, 2018).

	Descriptive Statistics (Исходные данные 2010-2020)							
Variable	Valid N	% Valid obs.	Mean	Confidence -95,000%	Confidence 95,000%	Median	Mode	Frequency of Mode
Total revealed	11	100,0000	46,45455	35,93322	56,97587	43,00000	43,00000	2

2-сурет. АИТВ жұқтырғандар нәтижелерінің сипаттама статистикасы бар электрондық кесте. (100 000 адамға шаққанда)

Бақыланатын айнымалының іріктемелі орташа мәні төмендегі формула бойынша анықталады (1):

$$\bar{x} = \frac{i\sum_{i=1}^{n} x_i}{n} \quad (1)$$

мұндағы n-үлгінің көлемі (айнымалы х бақылауларының нақты саны болып келеді).

Медиан(Median) жоғарғы және төменгі жағынан екіге тең біркелкі бөлінген, реттелген мәндерден тұрады. Мода(Mode) – бұл мәліметтер жиынтығында жиі кездесетін мән.

Іріктемелі дисперсия айнымалының өзгергіштігін сипаттайды және формула бойынша есептеледі (2):

$$\bar{S}_x^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1} \quad (2)$$

Мұндағы $\bar{x}$ – үлгінің орташа мәні. Дисперсия 0-ден шексіздікке дейін өзгереді. 0-дің төтенше мәні өзгергіштіктің болмауы – тұрақты айнымалылар.

Бастапқы деректер файлында Қазақстан Республикасының 16 өңірінде және 2 қаласында АИТВ жұқтырғандар туралы ақпарат бар. Осы кластерлік талдаудың мақсаты кластерлерге бөлу және тәуекел топтарын анықтау үшін тиісті кластерді анықтау болып табылады. Бұл мәселені шешу үшін кластерлік талдауды қолдану негізгі тиімді және жаппай қолданылатын әдістердің бірі болып саналады.

Жақындық шарасы ретінде Евклид арақашықтығын (Euclideandistances), ал кластерлерді біріктіру үшін – жалғыз байланыс әдісін(SingleLinkage) немесе (жақын көршілерді) әдісті пайдалана отырып, кластерлік талдаудың иерархиялық рәсімінің көмегімен 16 өңірді жіктеуді жүргіземіз. Осы әдістердің көмегімен екі кластерді бір-бірімен байланыстыруға болады. Кез келген екі кластер бір болғанда, олар бір-біріне жақындап және байланыс қашықтығымен ерекшеленеді.

Сондықтан біріктірілген кластерлер кездейсоқ қалған бөліктерден бөлек элементтерге айналады. Бұл құбылыс нысандарды бір-бірімен байланыстырады және кластерлерді құрайды. Алынған кластерлер ұзын тізбектермен ұсынылған. Кластерлердің табиғи санын анықтау аймақтарды кластерлерге біріктіру арқылы жүргізілді. Өңірлерді кластерлерге біріктіру тәртібі иерархиялық ағашта берілген 3-сурет.

$$\pi = \frac{a_i + a_j}{2b_{iJ}} \tag{3}$$

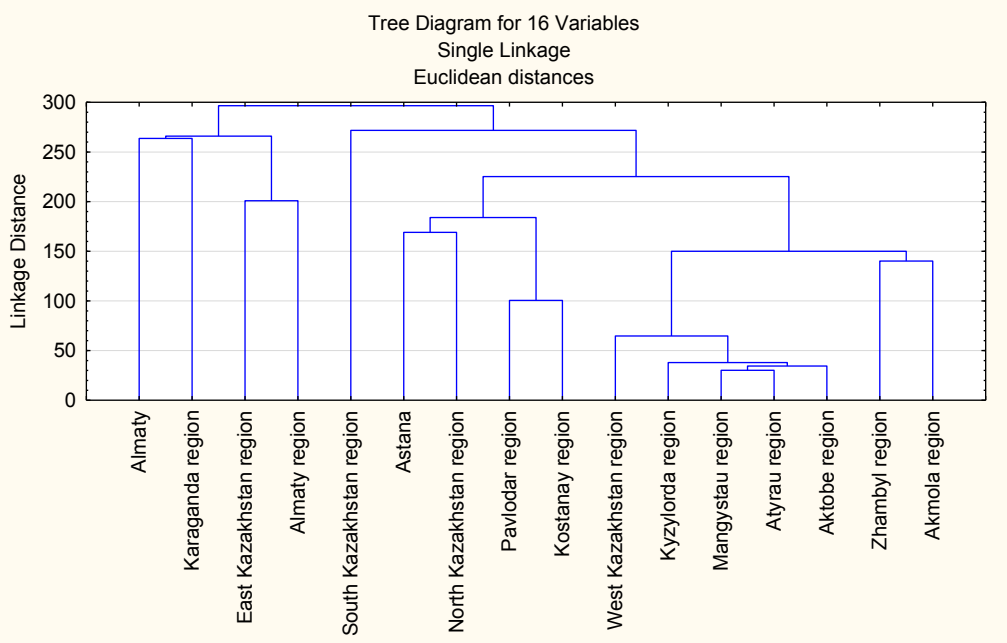
мұндағы a_i, a_j – сыныптардың орта кластық арақашықтығы;

$i : j; b_{ij}$ – осы сыныптар арасындағы орташа кластық арақашықтықтар.

Бағалауды дәстүрлі бөлу мынадай формула бойынша жүргізіледі:

$$S = \frac{1}{k}\sum_{i=1}^{\bar{e}} max\pi ij \tag{4}$$

Жоғарыда сипатталған алгоритмнің көмегімен алынған бөліктер біреуіне тең немесе 1-ден аспайды. Сондықтан, барлық нысандар бір кластерге біріктіріліп, соңында біреуіне тең болады деп қорытынды жасауға болады.



3-сурет. 2010-2020 жыл арлығындағы халықтың аурушаңдығы бойынша Қазақстан Республикасы өңірлерінің сыныптама дендограммасы

Жалғыз байланыс (Одиночная связь,SingleLinkade) – әдісі ең ұғымды әдіс болып келеді жәнеде кең таралған атауы «Жақын көрші» (Ближайший сосед, NearestNeighbor). Алгоритмнің жұмысы ең жақын екі нысанды іздеумен ұсынылған, олардың комбинациясы бастапқы кластерді құрумен байланысты. Әрбір келесі объект осы жақын орналасқан кластерге қосылады. Объектілер жиынтығы бөлінген кластерлердің табиғи санын анықтау үшін иерархиялық кластерлеудің әр деңгейінде жиынтықты осы кластар санына бөлу жүргізілді. Кластерлердің әр жұбымен олардың бір-бірімен ішкі байланысы бағаланып, әр кластер үшін орташа кластерлік қашықтықты есептеу болып келеді. Байланысты бағалау ретінде орташа кластерішілік қашықтықтың кластераралық қашықтыққа қатынасы алынады. Дендрограммада объектілер кластерлерге біріктірілетін қашықтықтар (шартты

бірліктерде) көлденеңінен белгіленеді. Көлденең келген ось бақылауды, ал тігінен – біріктіру қашықтығын білдіреді (3-сурет).

Алғашқы қадамдарда Қазақстан өңірлерінің кластерлері құрылуда: (Atyrauregion, Mangystauregion,Aktoberegion). Әрі қарай (WestKazakhstanregion, Kyzylordaregion) кластерлер пайда болды, бұл аймақтар арасындағы кластерлерді біріктіргенде үлкен болып келеді, алдыңғы қадамдарда қарағанда. (Pavlodarregion, Kostanayregion) келесі кластерлерге біріктіріледі (NorthKazakhstanregion, Astana). Бұдан әрі бір кластерге (Karaganda region, Almaty) және (Akmola region,East Kazakhstan region) және т.б. кластерлер біріктіріледі. Процесс барлық нысандарды бір кластерге біріктірумен аяқталады. Сонымен, дендограмма бойынша, бұл жағдайда үш кластерді бөлуге болады (1, 2-кестелер). Бастапқы бөлуді анықтаудың әртүрлі тәсілдері бар k-орташа сынақ үлгісінің үш кластеріне бөлудің жалпы нәтижелері (3-кесте).

1-кесте. Кластерлер арасындағы Евклид қашықтықтарының матрицасы

	Кластер 1	Кластер 2	Кластер 3
Кластер 1	0,0000	30464,68	101881,3
Кластер 2	174,5413	0,00	21707,3
Кластер 3	319,1884	147,33	0,0

2-кесте. Кластерлерге, қашықтыққа және аймақ бойынша айнымалыға бөлінген мәліметтер кестесі

	Spreadsheet6		
	1 VARIABLE	2 CLUSTER	3 DISTANCE
Akmola region	1	1	77,31
Aktobe region	2	1	177,89
Almaty region	3	2	211,43
Atyrau region	4	1	195,55
East Kazakhstan regi	5	2	131,75
Zhambyl region	6	1	75,14
West Kazakhstan regi	7	1	142,01
Karaganda region	8	2	79,87
Kostanay region	9	1	253,19
Kyzylorda region	10	1	208,81
Mangystau region	11	1	191,59
North Kazakhstan reg	12	1	120,97
South Kazakhstan reg	13	1	192,63
Almaty	14	2	273,38
Astana	15	1	192,72
Pavlodar region	16	1	276,28

3-кесте. Үш кластерге аймақ бойынша бөлінген жалпы нәтижесі

1-кластер	2-кластер	3-кластер
Алматы облысы, Қарағанды облысы, Шығыс Қазақстан облысы	Шығыс Қазақстан облысы, Астана қаласы, Солтүстік Қазақстан облысы,	Батыс Қазақстан облысы, Қызылорда облысы, Манғыстау облысы,

	Павлодар облысы, Қостанай облысы	Атырау облысы, Ақтөбе облысы, Жамбыл облысы, Ақмола облысы

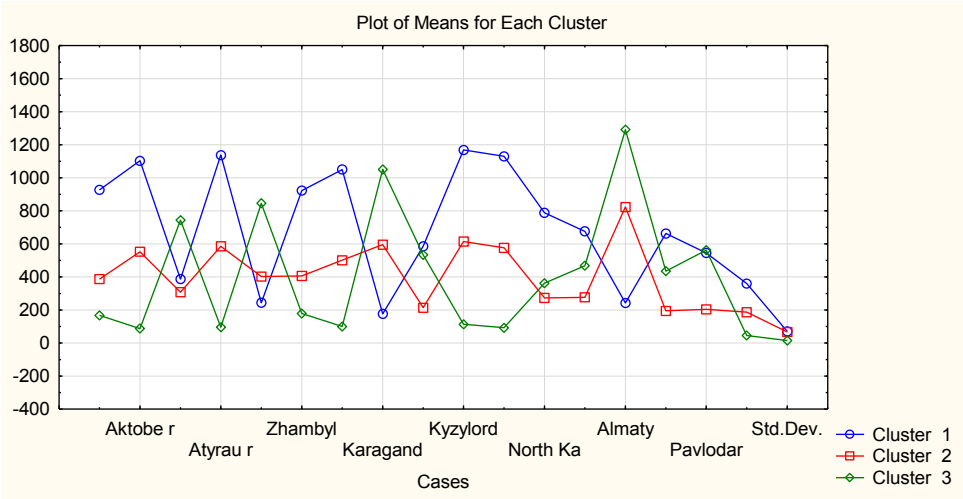
3-суретте жақындық өлшемін 250 деңгейінде кесу кезінде 3 кластер ерекшеленетіні көрсетілген. Алынған кластерлердің құрамы 2-кестеде анықталған. Алынған кластерлердің ерекшеліктерін талдап, аймақтардағы сыныптар бойынша АИТВ жұқтырғандар көрсеткіштерінің орташа мәнін салыстыра отырып, біз келесі нәтижелерге қол жеткіздік:

Бірінші кластер жалпы орта деңгейде кең таралған – АИТВ-инфекциясының жаңалықтары ересек тұрғындар мен халықтың осал топтары – есірткі тұтынушылар, сотталғандар арасында және АИТВ-инфекциясының жыныстық трансмиссиясының үлес салмағын алады.

Екінші кластер үшін бірінші кластермен салыстырғанда АИТВ – жұқпасының төмен көрсеткіші тән. Екінші кластерге тәуекелі жоғары, алайда алкоголизмнен, нашақорлықтан және басқа да әлеуметтік аурулардан зардап шегетін адамдар тобы кіреді;

Үшінші кластерде инъекциялық есірткіні тұтынушылар арасында АИТВ-инфекциясының таралуы мен берілу жолы, сондай-ақ ауру анадан құрсаққа жұқтыру жолдары көрсетілген.

3-суретте анықталғандай аймақтардың үш тобына қатысты айырмашылықтарды атап өтіп, 4-суретте әр кластер үшін орташа мәндерін көруге болады.



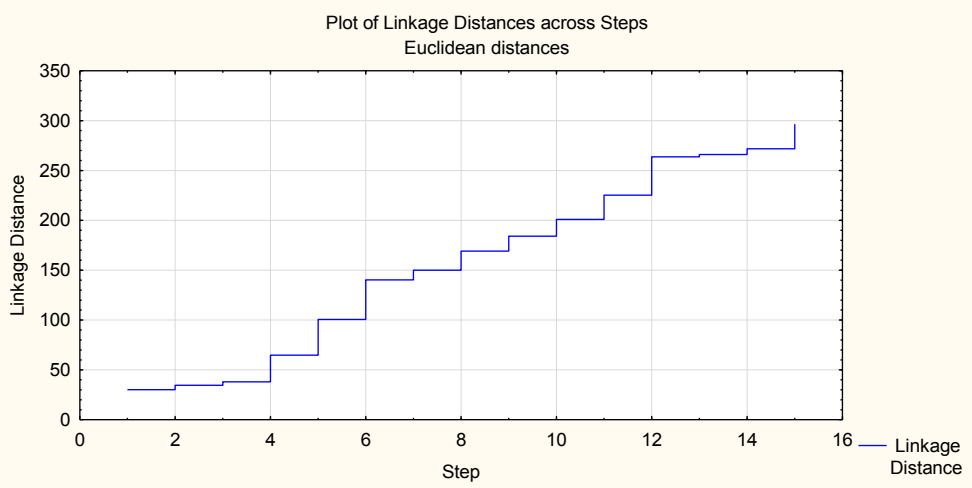
4-сурет. Әр кластер үшін орташа мәндер графигі

5-суретте кластар арасындағы қашықтық келтірілген. Кластер нөмірлері тігінен және көлденеңінен көрсетілген. Осылайша, жолдар мен бағандар қиылысқан кезде сәйкес кластар арасындағы қашықтық көрсетіледі. Сонымен қатар, диагональдан жоғары (нөлдер) квадраттар көрсетілген, ал төменде – евклидтік қашықтығы анықталған.

	1 Akmola region	2 Aktobe region	3 Almaty region	4 Atyrau region	5 East Kazakhstan region	6 Zhambyl region	7 West Kazakhstan region	8 Karaganda region	9 Kostanay region	10 Kyzylorda region	11 Mangystau region	12 North Kazakhstan region	13 South Kazakhstan region	14 Almaty	15 Astana	16 Pavlodar region
Akmola region	0,00000	191,92186	607,96464	226,11059	710,01408	140,23552	149,99000	915,89847	394,87720	253,93306	212,26870	225,23321	377,19358	153,31999	291,61619	421,70132
Aktobe region	191,92186	0,00000	784,53107	42,07137	886,44458	208,65282	64,73793	1092,47151	570,37532	68,43975	34,52535	387,05814	504,94851	332,58358	468,52748	599,95000
Almaty region	607,96464	784,53107	0,00000	816,00797	200,97512	608,80457	735,32986	376,07313	340,11028	846,46500	809,28796	477,18026	376,66431	582,51009	355,00000	296,53162
Atyrau region	226,11059	42,07137	816,00797	0,00000	920,75621	236,32181	101,90682	1126,62682	607,01236	38,00000	30,19934	418,30133	531,69822	366,56613	504,00595	634,45725
East Kazakhstan	710,01408	886,44458	200,97512	920,75621	0,00000	704,76237	831,97897	265,98120	377,49172	950,24418	912,25983	583,11234	467,40026	464,70959	454,24883	334,31123
Zhambyl region	140,23552	208,65282	608,80457	236,32181	704,76237	0,00000	166,86821	911,56130	424,16270	265,97744	234,74667	306,13722	314,47893	151,14074	335,64565	447,18229
West Kazakhstar	149,99000	64,73793	735,32986	101,90682	831,97897	166,86821	0,00000	1037,95327	512,76798	127,03149	92,78470	348,28580	455,66435	279,24040	415,74151	543,42341
Karaganda region	915,89847	1092,47151	376,07313	126,62682	265,98120	911,56130	1037,95327	0,00000	574,37618	1157,41782	119,02189	777,70303	662,41301	263,71386	650,65505	534,83268
Kostanay region	394,87720	570,37532	340,11028	607,01236	377,49172	424,16270	512,76798	574,37618	0,00000	635,14408	594,40895	323,49343	337,12757	805,75368	184,02717	100,59821
Kyzylorda region	253,93306	68,43975	846,46500	38,00000	950,24418	265,97744	127,03149	1157,41782	635,14408	0,00000	43,77214	445,75778	560,73791	397,73352	532,41337	663,62489
Mangystau region	212,26870	34,52535	809,28796	30,19934	912,25983	234,74667	92,78470	1119,02189	594,40895	43,77214	0,00000	408,05882	531,56843	358,95143	493,45314	623,72430
North Kazakhstar	225,23321	387,05814	477,18026	418,30133	583,11234	306,13722	348,28580	777,70303	323,49343	445,75778	408,05882	0,00000	348,84237	004,36597	169,10943	317,72944
South Kazakhstar	377,19358	504,94851	376,66431	531,69822	467,40026	314,47893	455,66435	662,41301	337,12757	560,73791	531,56843	348,84237	0,00000	897,70597	271,71124	321,57270
Almaty	1153,31999	1332,58358	582,51009	366,56613	464,70959	1151,14074	1279,24040	263,71386	805,75368	1397,73352	358,95143	1004,36597	897,70597	0,00000	885,65738	765,47175
Astana	291,61619	468,52748	355,00000	504,00595	454,24883	335,64565	415,74151	650,65505	184,02717	532,41337	493,45314	169,10943	271,71124	885,65738	0,00000	189,02381
Pavlodar region	421,70132	599,95000	296,53162	634,45725	334,31123	447,18229	543,42341	534,83268	100,59821	663,62489	623,72430	317,72944	321,57270	765,47175	189,02381	0,00000

5-сурет. Кластар арасындағы қашықтық кестесі

Ағаш тәрізді (древовидная кластеризация) кластерлеу нәтижелері. Диаграммада ось бойынша көлденең қадамдар, тігінен – қашықтықты бейнелейді. Барлық нысандарды бір кластерге біріктіру үшін 6-суретте көрсетілгендей алгоритмге 16 қадам қажет болды.



6-сурет. Қадамдармен біріктіру процесі

Алынған жіктеме бірінші кластерге біріктірілген өңірлерде АИТВ жұқтырғандардың жоғары өсуі бар кластерлерді анықтады. Өңірлерді біртекті топтарға біріктіру және кері міндеттерді шешу жолымен алынған статистикалық болжамдау нәтижелері инъекциялық есірткіні тұтынушылар, сырқаттанушылықтың болжамдаушылары болып табылатынын көрсетті. Халықтың осы тобы АИТВ-инфекциясы індетінің өсуін ынталандыруды, жалғастырып отырғанын Data mining технологиясының көмегімен өңделгенін көрсетті. Жыныстық жолмен берілетін инфекциялар құрылымында маңызды рөл атқаратын коинфекциялар үлесінің артуы үлкен алаңдаушылық тудырады.

Қорытынды. Мақалада Қазақстан Республикасында АИТВ-инфекциясымен сырқаттанушылықтың 10 жылдық кезеңінің деректер алынып, Data Mining технологиясының көмегімен халықтың сырқаттанушылығы бойынша талдау жүргізілді.

Нақты статистикалық деректер бойынша зерттеу жұмысы Statistica және SPSS қолданбалы бағдарламалар пакетінің көмегімен талданды. «Жақын көрші» (NearestNeighbor) әдісі бойынша иерархиялық кластерлік талдауды қолдану Қазақстан аумағында АИТВ инфекциясын жұқтыру және халық өлімі деңгейінің жалпы үрдістерімен ерекшеленді. Кластерлік талдау нәтижелері бойынша 3 кластерге бөлуге мүмкіндік берді, медициналық көмектің жиынтық коэффициентінің мәні бойынша: рейтингі жоғары, орташадан жоғары, орташадан төмен болып анықталды. Зерттеу кезінде талдау әдісі, қашықтық формуласының түрі жалғыз байланыс (SingleLinkade) және анықтамалық алгоритмдегі кластерлер саны анықталды.

Жүргізілген зерттеу нәтижесінде Қазақстан Республикасында АИТВ инфекциясымен сырқаттанған 2010-2020 жылдар аралығының мәліметі бойынша жүргізілген талдау АИТВ инфекциясын жұқтырған адамның артуы және сырқаттанушылықтың тұрақты үрдісін анықтауға мүмкіндік берді. Кластерлік жіктеу алгоритмі бойынша объектілер арасындағы ішкі байланыс құрылып, АИТВ эпидемиологиясының математикалық моделінің дұрыстығын көрсетті. Статистикалық талдау нәтижелері АИТВ инфекциясын жұқтырудың жоғары қаупіне ұшырайтын халық топтары анықталды.

Қазақстанда АИТВ-инфекциясына қарсы күрестегі терапевтік сұлбаларды, нақты антиретровирустық препараттарды модельдеуге, таңдалған базаны оңтайландыруға, алынған кластерлердің кластерленуі мен сипатын қалыптастыруға мүмкіндік берді.

Әдебиеттер тізімі

1. ЮНЭЙДС Информационный бюллетень 2022 https://www.unaids.org/ru/resources/fact-sheet
2. Статистический сборник. Здоровье населения РеспубликиКазахстан и деятельность организаций здравоохранения в 2020 году. https://amanbol.kz/news/vich-v-kazahstane-dannye/https://masa.media/ru/site/
3. Xinyi Wang. The Role of Data Mining Technology in Advertising Marketing. J. Phys.: Conf. Ser. 1744 042202, 2021. DOI: 10.1088/1742-6596/1744/4/042202
4. Yang, J, Li, Y, Liu, Q, et al. Brief introduction of medical database and data mining technology in big data era. J Evid Based Med. 2020; 13: 57-69. https://doi.org/10.1111/jebm.12373
5. Jianguo Liu & Sheng Zhou (2021) Application Research of Data Mining Technology in Personal Privacy Protection and Material Data Analysis, Integrated Ferroelectrics, 216:1, 29-42, DOI: 10.1080/10584587.2021.1911255
6. Bijalwan V, Kumar V, Kumari P et al (2014) KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application.Vol.7, No.1 (2014), pp.61-70.http://dx.doi.org/10.14257/ijdta.2014.7.1.06
7. Yukselturk E., Ozekes S., Turel Y. K. Predicting dropout student: An application of data mining methods in an online education program //European Journal of Open, Distance and e-learning. – 2014. – Vol. 17. – No. 1. – P. 118-133.
8. He, Wu, Gongjun Yan, and Li Da Xu. "Developing vehicular data cloud services in the IoT environment." IEEE transactionsonindustrialinformatics 10.2 (2014): 1587-1595.
9. Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." Expertsystemswith applications 41.4 (2014): 1432-1462.
10. Liu, L. (2020). Development and Application of Computer Data Mining Technology. In: Abawajy, J., Choo, KK., Islam, R., Xu, Z., Atiquzzaman, M. (eds) International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019. ATCI 2019.Advances in Intelligent Systems and Computing, vol 1017.Springer, Cham. https://doi.org/10.1007/978-3-030-25128-4_121
11. Liu, M., Qu, M. & Zhao, B. Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm. Mobile NetwAppl 22, 418-426 (2017).https://doi.org/10.1007/s11036-016-0797-2

12. Kubegenova, A.D., Zhakhiena, A.G., Baigubenova, S.K., Utyasheva, G.S., Omarov, A.N.Clustering and data mining on the example of hiv-infected people data (2022*)* Journal of Theoretical and Applied Information Technology*,* 100 (13). – P. 5010-5018.
13. Romero, C., Ventura, S. (2020). Educational Data Mining and Learning Analytics: An Updated Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Vol. 10, no. 3, pp. 1-21, doi: https://doi.org/10.1002/widm.1355
14. Zhenhua HUANG, Zhenyu WANG, Li JIANG, Rui ZHANG, Chang LEI, Xingwei LIU, Xiaohui XIE. Analysis of COVID-19 spread characteristics and infection numbers based on large-scale structured case data. SCIENTIA SINICA Informationis, Volume 50, Issue 12: 1882(2020) https://doi.org/ 10.1360/SSI-2020-0029
15. KABANIKHIN, S. I. et al. Determination of the coefficients of nonlinear ordinary differential equations systems using additional statistical information. International Journal of Mathematics and Physics, [S.l.]. – V. 10. – N. 1. – P. 36-42. june 2019. ISSN 2409-5508. Available at: <https://ijmph. kaznu.kz/index.php/kaznu/article/view/276>. Dateaccessed: 11 apr. 2022. doi: https://doi.org/ 10.26577/ijmph-2019-i1-5.
16. Кубегенова А.Д., Такуадина А.И., Криворотько О.И., Нурушева Ж.Т. Батыс Қазақстан облысындағы АИТВ-инфекциясының эпидемиологиялық жағдайын болжаудағы интеллектуалды талдау технологиясы. ҚР Ұлттық инженерлік академиясының хабаршысы. 2022 ж. – № 3(85). – С. 28-42. https://doi.org/10.47533/2020.1606-146X.174

References

1. YUNEJDS Informacionnyj byulleten' 2022https://www.unaids.org/ru/resources/fact-sheet
2. Statisticheskij sbornik. Zdorov'e naseleniya RespublikiKazahstan i deyatel'nost' organizacij zdravoohraneniya v 2020 godu.https://amanbol.kz/news/vich-v-kazahstane-dannye/https://masa. media/ru/site/
3. Xinyi Wang. The Role of Data Mining Technology in Advertising Marketing.J. Phys.: Conf. Ser. 1744 042202, 2021. DOI: 10.1088/1742-6596/1744/4/042202
4. Yang, J, Li, Y, Liu, Q, et al. Brief introduction of medical database and data mining technology in big data era. J Evid Based Med. 2020; 13: 57– 69. https://doi.org/10.1111/jebm.12373
5. Jianguo Liu & Sheng Zhou (2021) Application Research of Data Mining Technology in Personal Privacy Protection and Material Data Analysis, Integrated Ferroelectrics, 216:1, 29-42, DOI: 10.1080/10584587.2021.1911255
6. Bijalwan V, Kumar V, Kumari P et al (2014) KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application.Vol.7, No.1 (2014), pp.61-70.http://dx.doi.org/10.14257/ijdta.2014.7.1.06
7. Yukselturk E., Ozekes S., Turel Y. K. Predicting dropout student: An application of data mining methods in an online education program //European Journal of Open, Distance and e-learning. – 2014. – Vol.17. – No. 1. – P. 118-133.
8. He, Wu, Gongjun Yan, and Li Da Xu. "Developing vehicular data cloud services in the IoT environment." IEEE transactionsonindustrialinformatics 10.2 (2014): 1587-1595.
9. Peña-Ayala, Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." Expertsystemswith applications 41.4 (2014): 1432-1462.
10. Liu, L. (2020). Development and Application of Computer Data Mining Technology. In: Abawajy, J., Choo, KK., Islam, R., Xu, Z., Atiquzzaman, M. (eds) International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019. ATCI 2019.Advances in Intelligent Systems and Computing, vol 1017.Springer, Cham. https://doi.org/10.1007/978-3-030-25128-4_121
11. Liu, M., Qu, M. & Zhao, B. Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm. Mobile NetwAppl 22, 418–426 (2017).https://doi.org/10.1007/s11036-016-0797-2
12. Kubegenova, A.D., Zhakhiena, A.G., Baigubenova, S.K., Utyasheva, G.S., Omarov, A.N.Clustering and data mining on the example of hiv-infected people data (2022*)* Journal of Theoretical and Applied Information Technology, 100 (13), pp. 5010-5018.
13. Romero, C., Ventura, S. (2020). Educational Data Mining and Learning Analytics: An Updated Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Vol. 10, no. 3, pp. 1-21, doi: https://doi.org/10.1002/widm.1355
14. Zhenhua HUANG, Zhenyu WANG, Li JIANG, Rui ZHANG, Chang LEI, Xingwei LIU, Xiaohui XIE. Analysis of COVID-19 spread characteristics and infection numbers based on large-scale structured case data. SCIENTIA SINICA Informationis, Volume 50, Issue 12: 1882(2020) https://

doi.org/10.1360/SSI-2020-0029
15. KABANIKHIN, S. I. et al. Determination of the coefficients of nonlinear ordinary differential equations systems using additional statistical information. International Journal of Mathematics and Physics, [S.l.], v. 10, n. 1, p. 36-42, june 2019. ISSN 2409-5508. Available at: <https://ijmph. kaznu.kz/index.php/kaznu/article/view/276>. Dateaccessed: 11 apr. 2022. doi: https://doi.org/ 10.26577/ijmph-2019-i1-5
16. Kubegenova, A.I. Takuadina, O.I. Krivorot'ko, ZH.T. Nurusheva. Batys Қazaқstan oblysyndaғy AITV-infekciyasynyң epidemiologiyalyқ zhaғdajyn bolzhaudaғy intellektualdy taldau tekhnologiyasy. ҚR Ұlttyқ inzhenerlik akademiyasynyң habarshysy. 2022 zh. – № 3(85). – S. 28-42 https://doi.org/ 10.47533/2020.1606-146X.174