1-дәріс. Пәнге кіріспе. Курстың мақсаты мен міндеттері. Мәліметтер түсінігі және типтері

КУРСТЫҢ МАЗМҰНЫ

Курстың зерттеу нысандары. «Вконтакте», «YouTube» және «Твиттер» әлеуметтік желілердегі мәтіндер.

Тақырыптың өзектілігі. Бұл курс мәтіндегі экстремистік бағытты анықтау үшін семантикалық талдау модельдерін құруды зерттеудің бөлігі болып табылады. Машиналық оқыту әдістерін қолдана отырып, «экстремистік» және «бейтарап» категорияларды жіктеу үшін қолданылатын экстремистер жиі қолданатын кілт сөздерді анықтау. Қазақ тілі үшін мұндай мәліметтер базасы жоқ. Осы зерттеу нәтижесінде эксперименттік корпус және қазақ тіліндегі түйінді сөздер тізімі жасалады. Әр түрлі морфологиялық белгілері бар мәліметтер базасына кілт сөздер қосылады. Экстремистік кілт сөздердің бар-жоғындағы кіріс мәтінін тексеретін және табылған сөздерді қайтаратын бағдарлама жасалады.

Курстың негізгі сипаты. Экстремистік мазмұнды анықтау үшін веб-ресурстарды қазақ тіліндегі деректерді семантикалық талдау алгоритмдерін, экстремистік мазмұндағы мәтіндерді анықтау үшін машиналарды оқыту әдістерінің оңтайлы жиынтығын қалыптастыру әдістерін, қатысушыларды идентификациялау әдістері мен кибер-тергеу алгоритмдерін және жалпыға ашық үлкен көлемді ақпарат ішінен көлеңкелі нысандарды жасанды нейрондық желілерді қолдану арқылы танитын транзакция мәліметтерін интеллектуалды талдау әдісін құру.

КУРСТЫҢ МАЗМҰНЫ

Курстың мақсаты. Бұл курс интернет желісіндегі ақпаратты пайдалана отырып, қауіпсіздікті қамтамасыз ету, терроризм мен экстремизмге қарсы тұру мәселелерін шешу үшін машиналық оқыту әдістерін қолдануға арналған. Бұл міндеттерге әлеуетті террористік және экстремистік ақпаратты қамтитын электрондық хабарламаларды, құжаттарды және web ресурстарды іздеу, осындай ақпаратты тарататын пайдаланушылар топтарының және интернет қоғамдастықтардың құрылымын анықтау, осындай қоғамдастықтарда айналатын ақпарат ағындарының мониторингін және тақырыптық модельдеуді жүзеге асыру, алынған мониторинг нәтижелері негізінде қатерлерді бағалау және тәуекелдерді болжау.

Зерттеу әдістері. ЭБ (экстремистік бағыт) анықтау үшін веб-мазмұнды жинауға және талдауға арналған бағдарламалық модуль, ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту және тестілеуге арналған мәтіндер корпусы, ЭБ мәтіндерін семантикалық талдау моделі, морфологиялық анализатор, кілт сөздер базасы, ЭБ мәтіндерін анықтауға арналған машиналық әдістерді оқыту мен тестілеуге арналған белгілер жиынтығы.

КУРСТЫҢ МАЗМҰНЫ

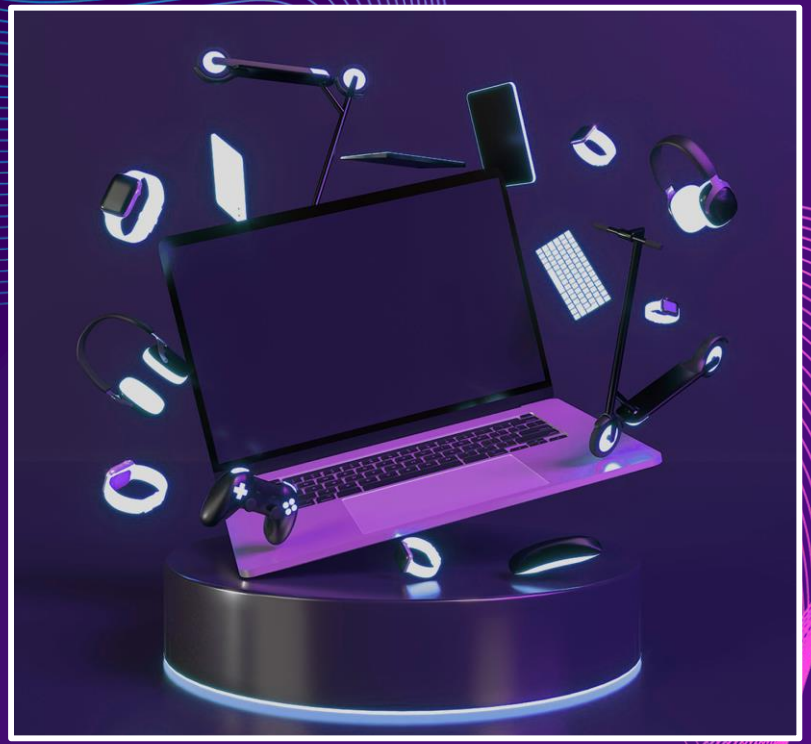
Курстың міндеттері:

- Террористік немесе экстремистік мазмұндағы ақпарат таратылатын интернет желісіндегі пайдаланушылар топтарын,қоғамдастықтар мен ресурстарды анықтау.
- Осындай топтарда таратылатын хабарламалар мен құжаттар ағындарының сипаттамаларын бақылау, алу және болжау.
- Осындай қауымдастық мүшелерінің қауіп-қатерін бағалау және болжау.
- Экстремистік бағытты анықтауға арналған бағдарламалық жасақтама жасау.

Курс нәтижелерінің қолданылу аясы. Алынған нәтижелердің мақсатты тұтынушылары – іргелі нәтижелерді әлемдік ғылыми қауымдастық пайдалана алады, әдіснама, алгоритмдер, патенттер және прототип түріндегі қолданбалы нәтижелерді ақпараттық қауіпсіздікті, сыни инфрақұрылымды қамтамасыз ету, интернет-экстремизммен күрес жөніндегі уәкілетті органдар пайдалануы мүмкін.

Мәліметтер (данные; data) - автоматты құралдардың көмегімен, кей жағдайда адамның қатысуымен, өңдеуге l ыңғайлы түрде берілген мағлұмат.

Мәліметтер типі (ағылш. data type) — программалауда мәліметтердің мүмкін мәндерінің ауқымын, оларға рұқсат етілген операциялар жиынын және бұл мәліметтерді аяқтаудың тәсілін анықтау арқылы оларды бір типке біріктіру жолы. Мынадай типтер: сандық, символдық, логикалық, мерзімдік (дата) және т.б. болады.



МӘЛІМЕТ ТИПТЕРІ

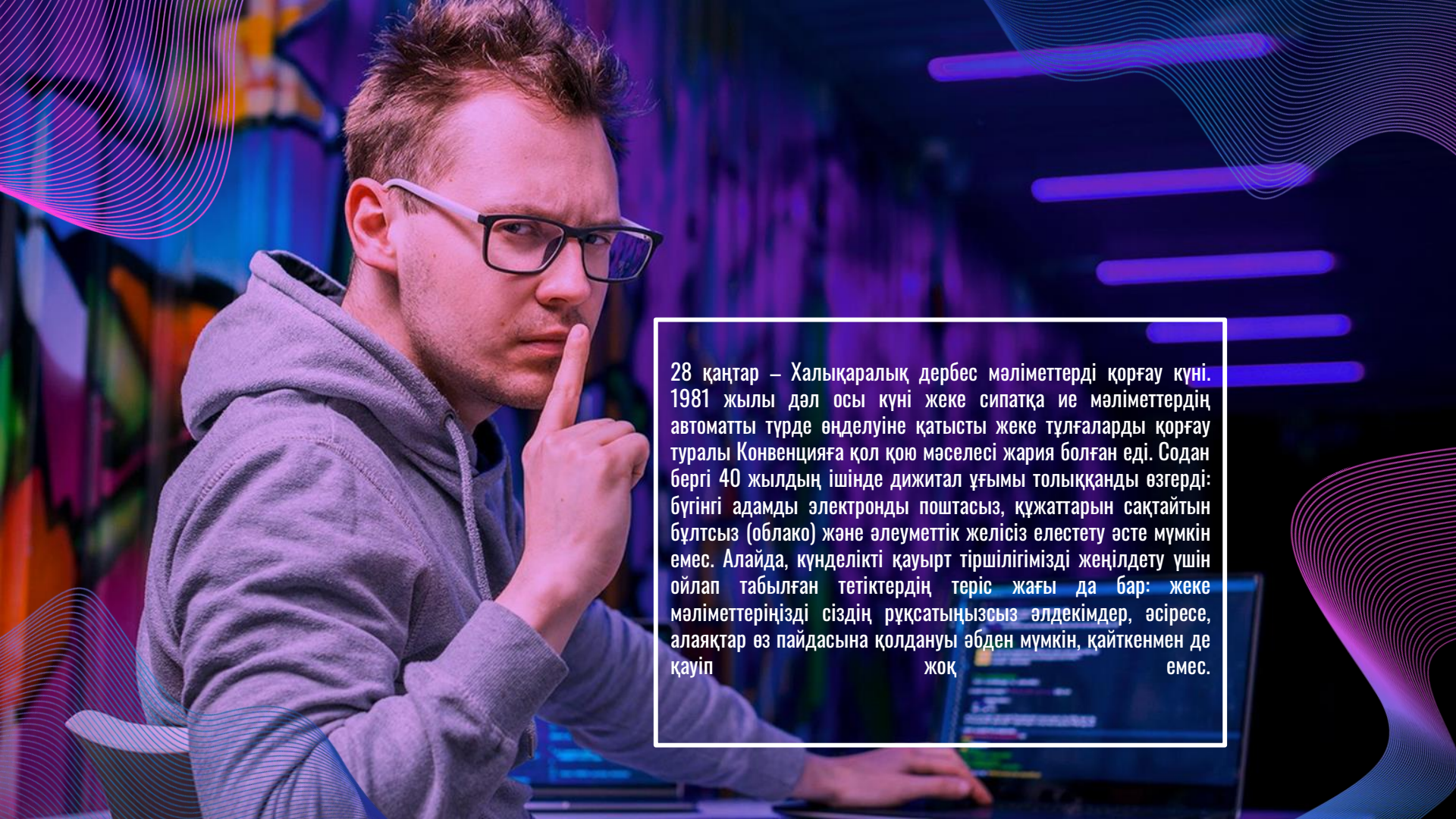
- Мәліметтердің кірістік, шығыстық, басқару, проблемалық, сандық, мәтіндік, графикалық және т.б. түрлерін атап өтуге болады.
- Арифметикалық (сандық) мәліметтер (арифметические (числовые) данные; arithmetic data) — арифметикалық амалдар орындауға болатын мәліметтер, яғни сандар.
- Әріптік мәліметтер (Буквенные данные; alphadetic data) — алфавит әріптері мен бос орындардан тұратын мәліметтер.
- Басқару мәліметтері (Управляющие данные; (управляющая информация) control information) — құрылғының, программаның, жүйенің қандай да бір басқару функцияларын орындауға қажетті мәліметтер.

МӘЛІМЕТ ТИПТЕРІ

- Графикалық мәліметтер (Графические данные; graphic data) — 1) графикалық
- бейне түріндегі (сурет, схема) мәліметтер; 2) графикалық объектілерді бейнелеуге жеткілікті болатын олардың машиналық өрнегі.
- Дискрет (цифрлық) мәліметтер (Дискретные (цифровые) данные; digital data) — цифрлық кодпен өрнектелген мәліметтер.
- Екілік мәліметтер (Двоичные данные; binary data) — екілік кодпен өрнектелген мәліметтер; мәндері екілік санау жүйесіндегі сандар болып келетін мәліметтер.
- Жабық (қорғалған) мәліметтер (Закрытые (защищенные) данные; restricted data) — кейбір өзгертулердің ғана пайдалану мүмкіндігі бар мәліметтер. Әдетте, қорғау пароль жүйесі арқылы іске асырылады.

МӘЛІМЕТ ТИПТЕРІ

- Кірістік мәліметтер (Входные данные; input data) — өңдеу немесе сақтау үшін жүйеге енгізілетін мәліметтер.
- Ондық мәліметтер (Десятичные данные; decimal data) — ондық санау жүйесінде өрнектелген мәліметтер.
- Реттелген мәліметтер (Упорядоченные (отсортированные, ранжированные) данные; ranked data) — берілген реттік қатынаспен орналастырылған мәліметтер (<- өспелі, >-кемімелі, <=кемімейтін, >= - өспейтін).
- Шифрланған мәліметтер (Зашифрованные данные; cipher data) — компьютердің жадында шифрланған түрде сақталатын мәліметтер (яғни криптографиялық қорғау тәсілі қолданылған мәліметтер).
- Шығыстық мәліметтер (Выходные данные; output data) — мәліметтерді шығару құрылғыларына компьютерден түсетін мәліметтер; программаның орындалу нәтижелері.



28 қаңтар – Халықаралық дербес мәліметтерді қорғау күні. 1981 жылы дәл осы күні жеке сипатқа ие мәліметтердің автоматты түрде өңделуіне қатысты жеке тұлғаларды қорғау туралы Конвенцияға қол қою мәселесі жария болған еді. Содан бергі 40 жылдың ішінде дижитал ұғымы толыққанды өзгерді: бүгінгі адамды электронды поштасыз, құжаттарын сақтайтын бұлтсыз (облако) және әлеуметтік желісіз елестету әсте мүмкін емес. Алайда, күнделікті қауырт тіршілігімізді жеңілдету үшін ойлап табылған тетіктердің теріс жағы да бар: жеке мәліметтеріңізді сіздің рұқсатыңызсыз әлдекімдер, әсіресе, алаяқтар өз пайдасына қолдануы әбден мүмкін, қайткенмен де қауіп жоқ емес.

ЖЕКЕ МӘЛІМЕТТЕР

Жеке мәліметтер дегенде ресми түрде адамның аты-жөні, тегі; туған күні мен туған жері; жеке куәліктегі дерегін және биометрлік мәліметтері; отбасылық мәртебесі және жанұя мүшелері, байланыс құралдары туралы қандай да бір ақпаратты түсінеміз. Десек те, бүгінде жеке мәліметтер дегеннің ауқымы әлдеқайда кең. Дүкеннен сатып алған дүниенің ақысын картамен төлегенде, әлеуметтік желілерге сурет салып, қай жерде, кіммен жүргеніңді белгілегенде, өзің туралы ашық ақпарат жаза бастағанда әр жерде цифрлық ізімізді қалдырып кете барамыз. Сондықтан дербес мәліметтердің қауіпсіздігі үкімет қабылдайтын заңнамалық актілерден бөлек, адамның бұл мәселеден қаншалықты хабардар болуымен де тікелей байланысты.

Біз достарымызбен ақпарат алмасып қана қоймай, кибер-алаяқтар үшін оңтайлы орта қалыптастырып жатырмыз. Инстаграм арқылы адам өміріне қатысты көп ақпарат алуға болатын күнге жеттік. Мұндай мүмкіндікті алаяқтар мүлт жіберіп жатқан жоқ.

Мұнда да бірнеше бағыт бойынша бір уақытта әрі кешенді жұмыс атқарылуы қажет. Ең алдымен, әрбір адам дұрыс мағынасында өзіне деген бақылауды (цензура) күшейтуі тиіс. Ғаламтор тұтынушысы желіге жүктелген кез келген ақпараттың «жеке» болудан қалатынын түсінуі тиіс. Бұл орайда, жабық топтар мен жабық парақшалардың тигізетін пайдасы шамалы. Сондықтан, желіде жария болуын қаламайтын дүниеңіз бар болса, оны мүлдем шығармаңыз.

«Хакерлер «мекендейтін» бірқатар ресустардан кез келген ақпаратты сатып алуға болады. Мәселен, 2019 жылдың жазында 11 миллиондай Қазақстан азаматтарының дербес мәліметтері ғаламторда ашық дереккөздерде жария болды. Сол уақытта бұл мәліметтерді жүктеп үлгергендер бүгінде онысын хакерлік ресурстарда сатумен әлек.

Бұзумен айналысатын хакерлерді өзара жіктеуге болады. White hat (ағылшын – «ақ қалпақ») және Black hat (ағылшын тілінен – «қара қалпақ») дейтін екі негізгі түрін атап өтуге болады. Киберқылмыскерлерге «қара қалпақ» деген атау берілген де, заң аясында ақпараттық қауіпсіздікпен айналысатын (әсіресе, ірі IT-компанияларда қызмет ететін) мамандар мен IT-жүйелерді зерттеушілер «ақ қалпақ» деп аталып кеткен. Бұдан бөлек, кішігірім құқықбұзушылыққа жол беретін, яки мүлдем заң бұзбайтын, бірақ қандай да бір интернет-сервистің ішкі ережелерін бұзғандарға Gray hat (ағылшын тілінен – «сұр қалпақ») деген термин тән».

Кәсіби «қоңырау шалғыштар» бар, олар қоңырау шалып, банктің қауіпсіздік қызметінен хабарласып тұрғандай сыңай танытады. Банктің жария болған базасынан алған деректерді айту арқылы болашақ құрбанының сеніміне кіреді. Бұл орайда, азаматтар құпиясөзін, әңгіме барысында хабарлама ретінде келген код-белгілерді өзі-ақ айтып береді. Құпиясөз бен қажет белгілерді біліп алған алаяқтар үшін жасанды адамның картасына әп-сәтте ақша аудару түкке тұрмайды. Сондай-ақ, алаяқтар банктердің клон (ойдан құрастырылған) сайттарын жасап, әлеуметтік желілерде акциялар мен конкурстар туралы жарнама тартатып, айласын асырып жатады. Нәтижесінде, адамдар жалған сайттарға тіркелу арқылы жеке мәліметтерін иеленуге жол ашады.

Қарапайым кеңес: тек тексерілген интернет-дүкендерді пайдаланыңыз, бөгде бір тұлғаға, тіпті банк қызметкері есебінде хабарласып тұрған кеңесші болса да, төлем карталары, хабарлама арқылы келетін код, құпиясөз туралы ақпарат бермеңіз, Qiwi-әмиян немесе басқа да электронды әмиян арқылы алдын ала төлем жасауға ешуақытта келіспеңіз. Интернет арқылы тауар сатып алғыңыз келсе, сатушыны әбден зерттеңіз. Саудагер басқа қаладан болса, сондағы туған-туыс немесе танысыңызға өтініш жасаңыз, сатушымен бетпе-бет кездесіп, тауарды тыңғылықты тексеріп алғаны абзал. Сонымен қатар, құпиясөзді жиі ауыстырыңыз, ғаламтордағы «оңай олжа» дегеннен өте сақ болыңыз.

Дербес мәліметтерге жекеменшік мүлік ретінде қарайтындар аз, ал дұрысында бұл тек жеке тұлғаға тән мүлік. Мәселен, АҚШ, дәлірек айтсақ, ондағы бірқатар құзыретті ұйымдар мемлекет пен бизнес өкілдерінің азаматтардың жеке деректерін қандай да бір мақсатта пайдалану мәселесіне қатаң қарайды. Калифорнияда түр-әлпетті тану құралдарын енгізу туралы мәселе көтерілді. Саясаткерлер мұндай технологияларды енгізуді және оның қорында сақталатын суреттерді жинақтауды болдырмау туралы шешімге келді. Өйткені, жүйе деген қателесуі мүмкін: дәл сол Калифорнияда адамның жүзін танитын жүйе полицияның мәліметтер базасынан 21 шенеуікті қылмыскер деп таныған.

Атауын мүлдем естіп көрмеген дүкеннен хабарласып, тауар яки қызмет ұсынған жағдаймен бетпе-бет келіп көрген боларсыз.Заманауи технологиялар мен цифрланудың бүгінгі даму қарқыны жеке цифрлық деректерімізбен бөлісуге «мәжбүрлеп» отыр. Алайда, біз бұл мәліметтерге кім және қалай иелік ететінін нақты білуіміз керек. Дербес мәліметтер дегеніміз біздің түр-әлпетіміз сияқты жеке өзімізге ғана тән бірегейлігіміз екенін әркім түсінуі шарт. Өзіңе ғана тиесілі автономияның айналасын бақылап, қорғауға тиіспіз. Мемлекет өз кезегінде дербес мәліметтерді жинау қажеттігін негіздеп отыруға міндетті.

Өлшемдер мен шкалалар

- Өлшем (measurement) зерттелетін нысандар мен құбылыстардың сипаттамаларына белгілі бір ережеге сай сандарды тағайындауды білдіреді.
- Шкала (scale) зерттелетін нысандар мен құбылыстарға сандар тағайындауда қолданылатын ереже немесе алгоритм.

Мәліметтер (data)

- **әрі қарай талдау мен зерттеу мақсатында жинақталатын бақылаулар мен зерттеулер нәтижесі.**

Респондент	Жасы	Жынысы	Білімі	Отбасылық жағдайы
1	29	0	12	2
2	23	1	14	1
3	37	1	16	2
4	46	0	10	4
5	34	1	14	1

Белгі

Белгі – заттың немесе құбылыстың оны басқа заттар мен құбылыстардан ерекшелейтін сипаттамасы.

Белгі түрлері:

- Сапалық, категориялық:
  - номиналды
  - дихотомиялық
  - реттік, ординалды, ранжирленетін
- Сандық, интервалды
  - дискретті
  - үзіліссіз

Номиналды шкала
(nominal scale)

Нысандар мен құбылыстарды қандай да бір белгі бойынша сұрыптауға немесе жіктеуге арналған атаулардан немесе категориялардан тұрады.

Номиналды шкала көмегімен алынған өлшеу нәтижелерін ретпен орналастыру мүмкін емес және олармен арифметикалық операциялар орындалмайды.

Мысал келтіріңіздер.

Сапалық, категориялық ***номиналды***

- ✓ Тікелей өлшеу мүмкін емес
- ✓ Ретпен орналастыру мүмкін емес
- ✓ Олармен арифметикалық операциялар орындалмайды.

диагноз, мамандық, отбасылық жағдай

Сапалық, категориялық *дихотомиялық (binary)*

Екі мәннің бірін ғана қабылдай алатын қарама-қарсы екі категорияға ғана қатысты болады.

Мысал келтіріңіз.

Дихотомиялық шкала (dichotomous scale)

- ✓ Екі категориядан тұратын номиналды шкала.

иә/жоқ, жұмысшы/жұмыссыз

Сапалық, категориялық реттік *(ordinal)*

Табиғи ретпен орналастыруға болады, бірақ шамалар арасында сандық өлшем болмайды.

Мысал келтіріңіз.

Реттік шкала (ordinal scale)

- ✓ Салыстырмалы позицияларды көрсету үшін нысандарға сандар тағайындалады, бірақ ол сандар нысандар арасындағы айырмашылықты көрсете алмайды.

Ауырлық деңгейі
Ауру кезеңі
Денсаулық жағдайына өзіндік баға беру

Интервалды шкала (interval scale)

Екі шама арасындағы айырмашылықты табуға мүмкіндік береді. Номиналды және реттік шкалалардың сипаттарына ие, алайда ол өлшенетін белгінің сандық мәнін көрсетуге мүмкіндік береді. Кемшілігі – есептеу нүктесі ретінде абсолютті нөл шамасының болмауы.

Мысал келтіріңіз.

Сандық немесе интервалдық

- ✓ Сандық өлшемі нақты анықталған белгілер.

Т, АҚ,, бой, салмақ, холестерин деңгейі, жұмысқа жарамсыздық күндері

Аталғандардың қайсысы үзіліссіз, қайсысы дискретті?

Сандық, *үзіліссіз*

- ✓ Үзіліссіз шкаладағы кез келген мәнді қабылдайды.

Дене массасы, температура, қанның биохимиялық көрсеткіштері

Сандық, *дискретті*

- ✓ Өлшеу диапазонындағы белгілі бір мәндерді ғана қабылдайды, әдетте бүтін сандар болып келеді.

Отбасыдағы бала саны, бір адамның бойындағы аурулар саны

Салыстырмалы шкала (ratio scale)

- ✓ Есеп нүктесі ретінде абсолют нөлді қабылдай алады, бұл оған интервалды шкаланың барлық қасиеттерін алуға мүмкіндік береді. Бұл шкаладағы мәліметтер үшін азайту мен бөлшектерді қоса алғанда барлық операцияларды орындауға болады.

Математика бойынша тестті орындауға берілетін уақыт

Шкалалар – қорытындылайық:

- номиналды
- Дихотомиялық
- Реттік
- Интервалды
- Салыстырмалы

- Тек категориялар болады, мәліметтерді реттестіру мүмкін емес.
- Номиналды шкала түрі, тек екі категориядан тұрады.
- Реттеуге болатын категориялардан тұрады, бірақ айырмашылық маңызды емес.
- Мәндер арасындағы айырмашылықты есептеуге болады, бірақ есептеу нүктесі жоқ.
- Есептеу нүктесі бар, мәндер арасында қатынас құруға болады.

Бәйге жарысының нәтижелері:

- Дихотомиялық белгі. Бұл жылқы бірінші келді ме?
0 – нет, 1- да

- Реттік. Бұл жылқы финишке нешінші болып келді?
1 – бірінші, 2 – екінші, 3 – үшінші және т.с.с.

- Сандық белгі. Бұл жылқының нәтижесі қандай?
60 сек., және т.с.с.

Туынды (екінші ретті) мәліметтер

- Проценттер. Науқастың жағдайы ем шарасынан кейін *24 % -ға жақсарды, яғни абсолютті мәліметтер емес, жақсару деңгейі маңызды.*

- Пропорциялар немесе қатынастар. *Дене массасы индексі*

- Қарқындылық. Аурудың салыстырмалы жиілігі, бұл шама ауру санын науқастарды бақылау жүргізілетін жылдардың жалпы санына бөледі.
- Белгілер, бағалар. Санды есептеу мүмкін болмаған жағдайда қолданылады. Мысалы, өмір сүру сапасына сауалнама жүргізу

Мәліметтерді редукциялау

- Талдауды оңайлату үшін мәліметтер жинағындағы категориялар санын азайту.

- Мәліметтерді жіктеу схемалар мен арифметикалық амалдар арқылы қосу.

- Индекс түріндегі мәліметтер мен мәліметтер жинағын қосу, мысалы, күтілетін өмір сүру ұзақтығы немесе жиынтық ішкі өнім.

Мәліметтерді редукциялау

Жас:
Жылдар (16 жас) – сандық белгі

Онжылдықтар (10-20 жас) – интервалды

Периодтар (жастық) – ординалды

Жас, жасөспірім - номиналды

Назарларыңызға рақмет!