Кіріспе

Деректерді өндіру және білімді ашу

Ұсынған
Ардан Жанегизов

Жоспары

- Анықтама
- Деректерді өндіру тапсырмалары
- Деректерді өндіру әдістері – жақсы және жағымсыз жақтары
- Деректерді түсіну
  - Атрибуттардың классификациясы
  - Деректер сапасы
- Деректерді дайындау және интеграциялау
- Деректерді өндіру
  - анықтамалар және мысалдар
- Қиындықтар

Кіріспе

- Деректерді қалай пайдалануға болады?
- **Деректерді өндіру** бұл машиналық оқыту, статистика және дерекқор жүйелерінің қиылысындағы әдістерді қамтитын үлкен деректер жиынындағы үлгілерді табу процесі.
- **Деректерді өндіру мақсаты** деректерді үлгіге сәйкестендіру.
- **Күтілетін нәтижелер** бастапқы деректерді қарағанкезде анық болмауы мүмкін жоғары деңгейлі мета ақпарат.
- Білімді ашу мен деректерді өндірудің айырмашылығы неде?
  - **Деректерді өндіру** білімді ашудың ажырамас бөлігі.

Білімді ашу дегеніміз не?

- **Білімді ашу** бастапқы деректерді пайдалы ақпаратқа түрлендіру процесі

Білімді ашу келесі кезеңдерден тұрады

- (1) Шуды және сәйкес келмейтін деректерді жою үшін деректерді тазалау
- (2) Бірнеше деректер көздерін біріктіру үшін деректер интеграциясы
- (3) Дерекқорлардан деректерді таңдау
- (4) Деректерді өңдеуге сәйкес пішіндерге келтіру үшін деректерді түрлендіру
- (5) Деректер үлгілерін шығару үшін **деректерді өндіру**
- (6) Қызықты үлгілерді анықтау үшін үлгіні бағалау
- (7) Пайдаланушыларға алынған білімді презентациялау

Деректерді өндіру дегеніміз не?

- Деректерді өндіру бұл деректерден пайдалы ақпаратты алу немесе ашу үшін машина көмегімен қолайлы әдістерді әзірлеу, анықтау және пайдалану туралы ғылым.

- Деректерді өндіру бұл үлкен деректер репозитарийлеріндегі пайдалы ақпаратты автоматты түрде табу процесі.

Деректерді өндіру дегеніміз не?

- Алдын ала бар үлкен деректер репозитарийлеріндегі деректер
  - Деректер жиналып қойған және сақталған
  - Деректер жинағы бекітілген
- Деректердің үлкен көлемі
  - Қолмен өңдеуге тым көп
- Сәйкес әдістерді әзірлеу және анықтау
  - Берілген тапсырма үшін дұрыс әдісті таңдау
  - Егер қолайлы әдіс табылмаса, онда қолайлы әдісті әзірлеу
  - Деректерді, тапсырманы, қолжетімді әдістерді білу мен түсінуді талап етеді
- Жаңа, пайдалы, түсінікті нәтижелер
  - Нәтиже бойынша әрекет ету мүмкіндігі болуы керек
  - Адамдар үлгілер мен нәтижелерді түсіндіре білуі керек

Деректер ғалымы

- Деректерді өндірудегі әдістерді білетін сарапшы
- Деректерді іздеу тапсырмасының күрделілігін түсінеді
- Берілген тапсырма үшін ең қолайлы әдісді анықтайды
- Берілген тапсырманың жаңа әдістерін әзірлеу қабілетіне ие
- Нәтижелерді оқу, түсіндіру және түсіну қабілетіне ие

Деректерді өндіру тапсырма түрлері

Деректерді өндіру міндеттерін келесідей жіктеуге болады

1. Болжамдық тапсырмалар
   - тапсырмалардың мақсаты басқа атрибуттардың мәніне негізделген белгілі бір атрибуттың мәнін болжау
2. Сипаттама тапсырмалар
   - тапсырмалардың мақсаты деректердегі негізгі қатынастарды қорытындылайтын корреляциялар, трендтер, кластерлер, траекториялар және аномалиялар сияқты үлгілерді алу
3. Немесе екеуі де

Болжамдық және Сипаттама тапсырмалар

Болжамдық

1. классификация
2. регрессия
3. уақыт қатарын талдау
4. болжау

Сипаттама

1. кластерлеу
2. қорытындылау
3. ассоциация ережелері
4. ретін ашу

Деректерді өндіру тапсырмалары

- **Классификация.** Деректерді алдын ала анықталған топтарға немесе сыныптарға салу.
  - Бақыланатын оқу
  - Үлгіні тану
  - Болжау

- **Регрессия.** деректер элементін нақты бағаланған болжам айнымалысына салыстыру.
  - Функцияның жуықтауы

- **Кластерлеу.** Ұқсас деректерді бірге кластерлерге топтау.
  - Бақылаусыз оқыту
  - Сегменттеу
  - Бөлу

Деректерді өндіру тапсырмалары

- **Қорытындылау.** Деректерді байланысты қарапайым сипаттамалары бар ішкі жиындарға салыстыру.
  - Сипаттама
  - Жалпылау

- **Сілтемені талдау.** Деректер арасында қарым-қатынастарды ашу.
  - Сәйкестік талдауы
  - Қауымдастық ережелері
  - Тізбекті талдау ретті үлгілерді анықтайды

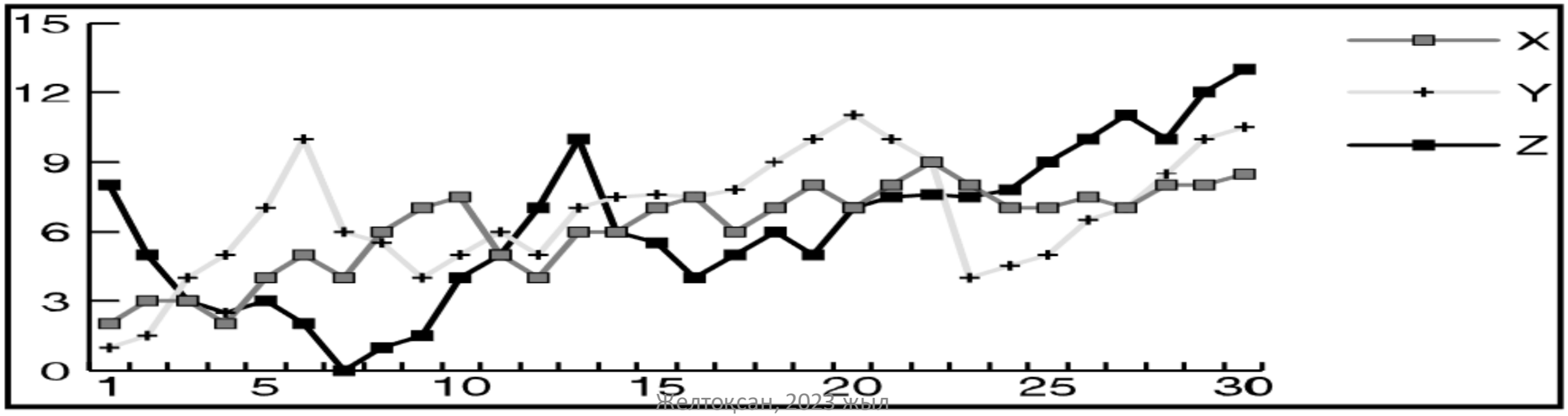
Деректерді өндіру тапсырма мысалдары

Кейбір мысалдар

- Болжамдық тапсырмалар
- Кластерлік талдау
- Ассоциация талдау
- Аномалияны анықтау

Болжамдық модельдеу

- Болжамдық модельдеу өткен бақылаулардың түсіндірмелі айнымалыларының функциясы ретінде мақсатты айнымалы үшін үлгі құру міндетін білдіреді.
- Мысалдар: Қор нарығындағы болашақ құндылықтарды болжау, электр энергиясын тұтыну, ...



Кластерлік талдау

- Кластерлік талдау бір кластерге жататын бақылаулар басқа кластерге жататын бір-біріне көбірек ұқсас болатындай тығыз байланысты бақылаулар топтарын табады.

- Мысалы, газеттегі мақалаларды сәйкес тақырыптар бойынша топтастыру.

- Сөз жиілік жұптарын талдау негізінде (w,c) w - сөз және c - сөздің мақалада қанша рет кездесетіндігі. экономикаға сәйкес келетін баптар кластерін және денсаулық сақтау саласына сәйкес келетін мақалалар кластерін анықтауға болады.

Ассоциация талдауы

- Ассоциация талдауы деректердегі қатты байланыстырылған мүмкіндіктерді сипаттайтын үлгілерді табу үшін пайдаланылады.
- Мысалы, тұтынушылар жиі бірге сатып алатын заттарды табу.

Тұтынушылар себетін талдау негізінде келесі ережелердің шығарылуы

- {нан} → {сары май}, бұл нан сатып алатын тұтынушылардың да май сатып алуға бейім екенін көрсетеді
- {доллар} → {индустрия}, бұл доллар сөзін пайдаланатын газет мақалаларында да сөз индустриясын қолдануды ұсынады
- {салбырап тұрған анықтама} → {жадтың ағып кетуі} сілтеме қатесі бар бағдарламалық құралда да жадтың ағып кетуін көрсетеді

Деректерді өндіру әдістері

Кейбір кең таралған деректерді өндіру әдістері

- K-means
- DBscan
- Ең жақын көрші классификаторы
- Нейрондық желілер
- Векторлық машиналар
- Шешім ағашының классификаторлары
- Bayes классификаторы
- Ассоциация ережесін өндіру
- …басқа да көп

Ең жақын көрші классификаторы

Артықшылықтары

- Жылдам жаттығу
- Қарапайым және түсінуге оңай

Кемшіліктері

- Қолдану кезінде баяу
- Функция таңдау жоқ
- Шуға және шектен тыс әсерлерге сезімтал
- Теңгерімсіз деректерге сезімтал

Шешім ағашы

Артықшылықтары

- Ақылға қонымды жаттығу уақыты
- Түсіндіру оңай
- Іске асыру оңай
- Көптеген мүмкіндіктерді өңдей алады

Кемшіліктері

- Мүмкіндіктер арасындағы күрделі қатынасты өңдеу мүмкін емес
- Қарапайым шешім шекаралары
- Жетіспейтін деректерге сезімтал
- Шуға және шектен тыс әсерлерге сезімтал

Нейрондық желілер (NN)

Артықшылықтары

- Күрделі сынып шекараларын біле алады
- Көптеген мүмкіндіктерді өңдей алады
- Функция таңдауы жоқ
- Шулар мен шектен тыс әсерлерге сезімтал емес
- Параллельді компьютерлерде масштабтауға болады

Кемшіліктері

- Жаттығу уақыты баяу
- Түсіндіру қиын
- Іске асыру қиын
- Параметрлерді сынақ және қате арқылы орнатуды талап етеді
- Кейбір NN теңгерімсіз деректерге сезімтал
- Сирек деректерге сезімтал болуы мүмкін

Векторлық машиналар

Артықшылықтары

- Күрделі сынып шекараларын біле алады
- Көптеген мүмкіндіктерді өңдей алады
- Нейрондық желілерге қарағанда қарапайым
- Теңгерімсіз деректерге сезімтал емес

Кемшіліктері

- Ақылға қонымды жаттығу уақыты, бірақ өте үлкен деректерге масштабталмайды
- Түсіндіру қиын
- Сирек деректерге сезімтал болуы мүмкін

Деректерді өндіру әдістері

Осы мысалдардан біз мыналарды түйеміз

- Деректерді өндіруге барлығына сәйкес келетін тәсіл жоқ
- Әрбір әдістің өз артықшылықтары мен кемшіліктері бар
  - әрбір әдістің күшті және әлсіз жақтарын түсіну маңызды
  - деректердің қасиетін түсіну және берілген тапсырманы түсіну маңызды

- Проблемаларға ұшырау арқылы қолайлы әдістерді анықтау
- Тәжірибе ең жақсы мұғалім болады
  - істей отырып үйрен

Деректер

Деректерді өндірудегі алғашқы қадам

- Деректер қасиетін түсін
- Деректердің сапасын біл
- Доменмен таныс
    - Деректер қайдан келеді, деректер қалай жиналды, деректер қалай сақталады және оған қалай қол жеткізуге болады, деректер толық па?

Енгізу деректері

Енгізілген деректерді келесідей сипаттауға болады

Деректер жинағы

- Деректер жинағы – деректер объектілерінің (жазбалар, нүктелер, векторлар, графиктер, бақылаулар, т.б.) жиынтығы

Атрибут

- Атрибут - бір объектіден екіншісіне немесе бір уақыттан екіншісіне өзгеруі мүмкін объектінің қасиеті, сипаттамалары

Атрибут түрі

- Атрибут түрі төлсипаттың негізгі қасиеттеріне сәйкес келетін оның мәндерінің қасиеттерімен анықталады

Атрибут түрлері

Номиналды

- номиналды атрибуттың мәндері әртүрлі атаулар (ақ, нан, Адам)

Реттік

- объектілерді кезектестіруге мүмкіндік береді (1 ақпан, 3 жүлде)

Интервалды

- арасында айырмашылық бар мағыналы интервалдар (10-30 жас)

Қатынас

- айырмашылықтар да, арақатынастар да маңызды (1/3)

Деректер сапасы

Өлшеу және деректерді жинау қателері

- Өлшеу қатесі жазылған мән шынайы мәннен өзгеше болғанда орын алады.
- Деректерді жинау қатесі деректер нысандарын немесе атрибуттарын өткізіп жіберуді немесе деректер нысанын сәйкессіз қосуды білдіреді.

Шу және артефактілер

- Шу – өлшеу қателігінің кездейсоқ құрамдас бөлігі, ол мәнді бұрмалайды немесе жалған нысандарды қосады.
- Артефакт – деректердің детерминирленген бұрмалануы.

Деректер сапасы

Нақтылық, қиғаштық, және дәлдік.

- Нақтылық деп қайталанған өлшемдердің бір-біріне жақындығын білдіреді.
- Қиғаштық өлшемдердің өлшенетін шамадан жүйелі түрде өзгеруін білдіреді.
- Дәлдік деп өлшенетін шаманың шын мәнінің өлшемдерінің жақындығын білдіреді.

Шеткі мәндер

- Шеткі мәндер - бұл деректер жиынындағы басқа деректер нысандарының көпшілігінен өзгеше сипаттамалары бар деректер нысандары немесе ...
- ... немесе сол төлсипаттың типтік мәндеріне қатысты әдеттен тыс атрибуттардың мәндері.

Деректер сапасы

Мәндердің болмауы

- Мәндердің болмауы бір немесе бірнеше төлсипат мәндерінің деректер нысандарында қолжетімді еместігін білдіреді.
- Мәндер болмауы себебі ақпарат жиналмаған, кейбір атрибуттар қолданылмайды, оның болуы басқа мәндердің болуына байланысты және т.б.

Сәйкес емес мәндер

- Сәйкес емес мәндер - берілген сәйкестік шектеулерін бұзатын мәндер.

Қайталанатын деректер

- Қайталанатын деректер бір-бірінің қайталанатын немесе дерлік қайталанатын деректер нысандары болып табылады.

Деректерді алдын ала өңдеу

Деректерді өндірудегі екінші қадам

- деректерді алдын ала өңдеу және
- деректерді біріктіру және түрлендіру

Деректерді алдын ала өңдеу

Агрегация

- Агрегация екі немесе одан да көп нысандарды бір нысанға біріктіред

Іріктеу

- Іріктеу талданатын деректер нысандарының жиынын таңдайды

Өлшемді азайту

- Өлшемді азайту ескілердің тіркесімі болып табылатын жаңасын жасайтын нысанды сипаттайтын атрибуттардың жалпы санын азайтады

Мүмкіндіктер жиынын таңдау

- Мүмкіндіктің ішкі жиынын таңдау маңызды емес атрибуттарды жою арқылы төлсипаттардың жалпы санын азайтады

Деректерді алдын ала өңдеу

Функцияны жасау

- Функцияны жасау бастапқыдан жаңа атрибуттар жинағын жасауды білдіреді.

Дискретизация және бинаризация

- Дискретизация – үздіксіз атрибуттың категориялық атрибутқа айналуы.
- Бинаризация – үздіксіз және дискретті атрибуттардың екілік атрибутқа айналуы.

Айнымалы түрлендіру

- Айнымалы түрлендіру деп айнымалының (атрибуттың) барлық мәндеріне қолданылатын түрлендіруді айтады.

Деректерді біріктіру

- Деректерді біріктіру бірнеше деректер көздерінен алынған деректерді когерентті деректер қоймасына біріктіруді білдіреді.

Схеманы біріктіру және нысанды сәйкестендіру

- Схеманы біріктіру нақты әлем нысандарын ортақ схемаға сәйкестендіруді білдіреді.
- Нысанды сәйкестендіру дегеніміз, сипаттамалары сәл басқаша болатын бірдей нақты дүние нысандарын сәйкестендіру.

Артық жұмыстарды жою

- Артықшылықтарды жою мәндері басқа атрибуттардан алынуы мүмкін атрибуттарды табуды білдіреді, мысалы. корреляциялық талдау арқылы.

Деректерді біріктіру

Деректер мәнінің қайшылықтарын анықтау және шешу

- Деректер мәнінің қайшылықтарын анықтау және шешу бір нақты әлем субъектісі үшін әртүрлі көздерден бірдей атрибуттардың мәндері әртүрлі болуы мүмкін барлық жағдайларды анықтау және жоюды білдіреді.

Деректерді өндіру

Деректерді өндірудегі үшінші қадам

- Өндіру

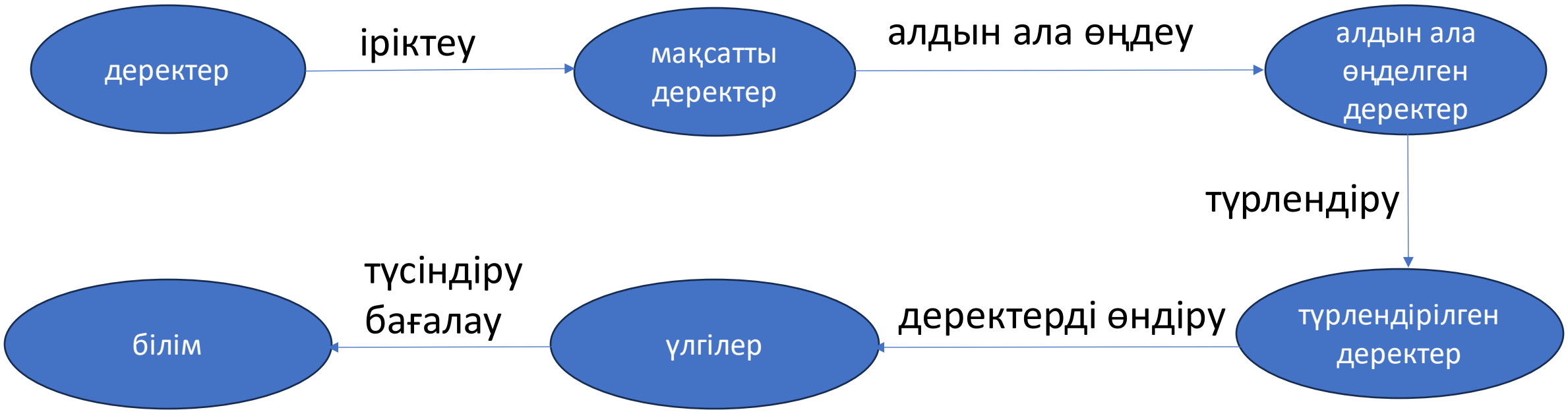
Бірақ қандай әдістерді қолдану керек?

- Деректерге байланысты (алғашқы және екінші қадамда қамтылған)
- Тапсырмаға, доменге және мәселеге байланысты

Деректерді өндіру

Деректерді өндірудегі төртінші қадам

- Түсіндіру және бағалау



Деректерді өндіру тапсырмалары

- Классификация
- Кластерлеу
- Тізбекті үлгіні табу
- Регрессия
- Ауытқуды анықтау

Классификация анықтамасы

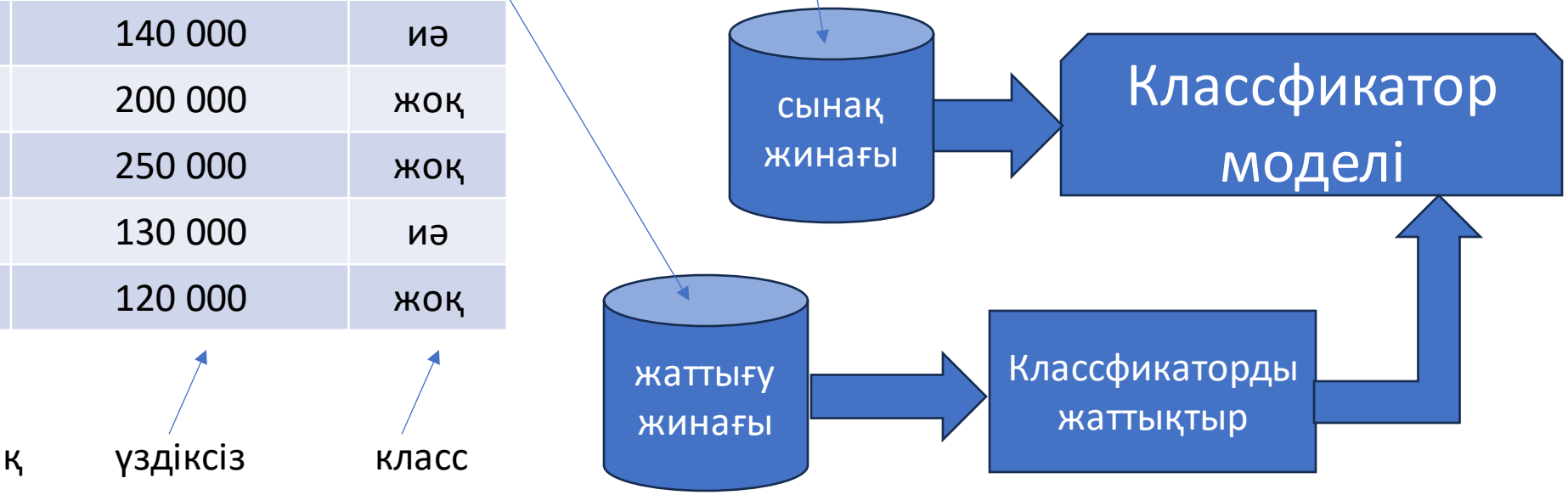
- Жазбалар жинағы берілген (жаттығу жинағы)
  - әрбір жазбада атрибуттар жиыны бар
  - атрибуттардың бірі класс болып табылады
- **Мақсат**
  - басқа атрибуттар мәнінің функциясы ретінде класс атрибут моделін тап
  - **бұрын көрмеген** жазбаларға классты мүмкіндігінше дәл беру керек
- Модельдің дәлдігін анықтау үшін сынақ жинағы пайдаланылады
  - әдетте, берілген деректер жаттығу және сынақ жинағына деп бөлінеді
  - үлгіні жаттықтыру үшін жаттығу жинағы қолданылады
  - оны тексеру үшін сынақ жинағы қолданылады

Классификация мысалы

id	Қайтару	Отбасылық жағдайы	Салық салынатын табыс	Алдау
1	иә	бойдақ	150 000	жоқ
2	жоқ	үйленген	100 000	жоқ
3	иә	үйленген	170 000	жоқ
4	жоқ	ажырасқан	200 000	жоқ
5	жоқ	бойдақ	140 000	иә
6	иә	үйленген	200 000	жоқ
7	иә	үйленген	250 000	жоқ
8	жоқ	ажырасқан	130 000	иә
9	иә	бойдақ	120 000	жоқ

категориялық категориялық үздіксіз класс

id	Қайтару	Отбасылық жағдайы	Салық салынатын табыс	Алдау
1	иә	бойдақ	100 000	?
2	жоқ	үйленген	180 000	?
3	иә	ажырасқан	120 000	?



Классификацияны қолдан, тікелей жарнама

Мақсат

- пошта жөнелтілімінің шығының азайту
- жаңа смартфон алатын тұтынушылар тобына тікелей жіберу арқылы

Тәсіл

- бұрын ұсынылған ұқсас өнімге арналған деректерді пайдалан
- қай тұтынушы сатып алды, қайсысы сатып алмағаны белгілі
- {сатып алды, сатып алмады} шешімі класс атрибутын құрайды
- тұтынушы туралы демографиялық, өмір салты тұралы ақпаратты жина
  - жұмыс істейтің жері, қайда тұрады, қанша алады және т.б.
- модельді жаттықтыруға ақпаратты кіріс атрибуттары ретінде пайдалан

Классификацияны қолдан, алаяқты анықта

Мақсат

- карта транзакцияларындағы алаяқтық жағдайларды болжау

Тәсіл

- транзакция, шот иесі ақпаратын атрибуттар ретінде пайдалан
  - тұтынушы қашан сатып алады, не сатып алады, қанша төледі, т.б
- өткен транзакцияларда {алаяқтық, әділ} класс атрибутың белгіле
- транзакциялар класының моделін жаттықтыр
- алаяқтықты анықтау үшін есептік жазбадағы карта транзакцияларына модельді пайдалан

Кластер анықтамасы

Деректер нүктелерінің жиынтығы берілген

- әрқайсысында атрибуттар жиынтығы бар
- және олардың арасындағы ұқсастық өлшемі бар

Кластерлерді тап

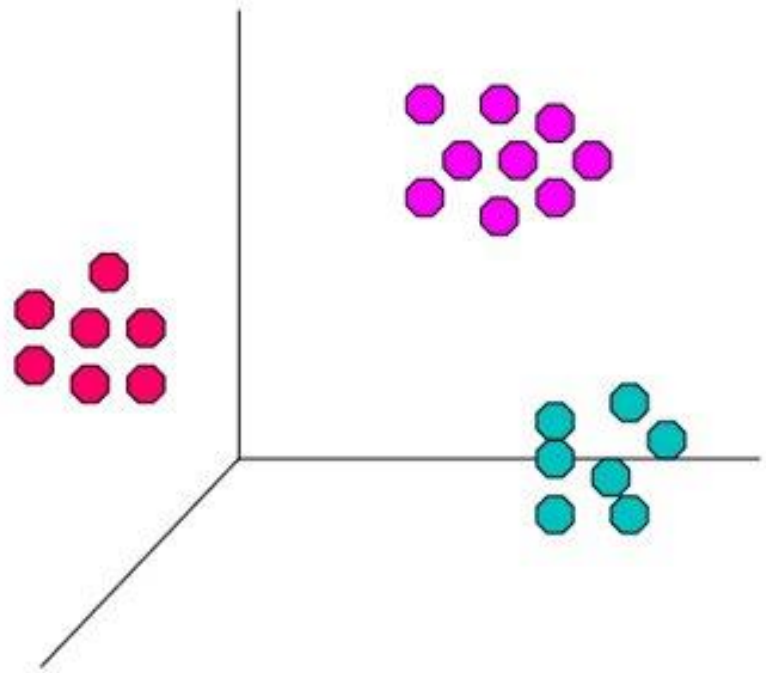
- бір кластердегі деректер нүктелері бір-біріне көбірек ұқсас
- жеке кластерлердегі деректер нүктелері бір-біріне ұқсамайды

Ұқсастық шаралары

- атрибуттар үздіксіз болса, Евклидтік қашықтық
- мәселеге қатысты басқа шаралар

Кластер иллюстрациясы

- Кластер ішілік қашықтық минималды
- Кластер аралық қашықтық максималды



Кластерді қолдан, нарық сегменті

Мақсат. нарықты тұтынушылардың әр түрлі ішкі жиындарына бөлу

- мұнда кез келген ішкі жиын нақты маркетинг кешенімен қол жеткізу үшін нарықтық мақсат ретінде таңдалуы мүмкін

Тәсіл

- географиялық және өмір салтына қатысты ақпарат негізінде тұтынушылар туралы әртүрлі атрибуттар жина
- ұқсас тұтынушылардың кластерлерін тап
- бір кластердегі тұтынушылардың әртүрлі кластерлердегі сатып алу үлгілерімен салыстыру арқылы кластер сапасын өлше

Кластерді қолдан, құжаттар кластері

Мақсат

- құжатта кездесетін терминдер негізінде бір-біріне ұқсас құжат топтары

Тәсіл

- әр құжатта жиі кездесетін терминдерді анықта
- әртүрлі терминдердің жиіліктеріне негізделген ұқсастық өлшемін жаса
- Оны кластерлеу үшін пайдалан

Пайда

- жаңа құжатты кластер тобына байланыстыруға пайдалана алады
- кластерленген құжаттарда терминді іздеуге пайдалана алады

Құжаттар кластері иллюстрациясы

Кластер нүктелері

- Қазақстан газетінің 3204 мақаласы

Ұқсастық өлшемі

- бұл құжаттарда қанша сөз ортақ (кейбір сөздерді сүзгілеуден кейін)

Санат	Жалпы мақалалар	Дұрыс орналастырылған
Қаржылық	555	364
Шетелдік	341	260
Ұлттық	1216	1023
Спорт	738	573
Ойын-сауық	354	278

Ассоциация ережесін ашу анықтамасы

- Жазбалар жиынтығы берілген, олардың әрқайсысында берілген жинақтағы элементтердің кейбір саны бар
  - басқа элементтердің пайда болуына негізделген элементтің пайда болуын болжайтын тәуелділік ережелерін жаса

- Табылған ережелер
  - {Сүт} --> {Май}
  - {Қағаз, Сүт} --> {Жұмыртқа}

Id	Элементтер
1	Нан, май, сүт
2	Жұмыртқа, нан
3	Жұмыртқа, май, қағаз, сүт
4	Жұмыртқа, нан, қағаз, сүт
5	Май, қағаз, сүт

Ассоциация ережесін қолдан, сату ынтасы

Табылған ереже

- {жаңғақ, ...} --> {сүт}
- жаңғақ себебі => дүкен жаңғақтар сатуды тоқтатса, қандай өнімдерге әсер ететінін көру үшін пайдалануға болады
- сүт салдары => сүт сатылымын арттыру үшін не істеу керектігін анықтау үшін пайдаланылуы мүмкін
- жаңғақ себебі және сүт салдары => сүтті сатуды ынталандыру үшін жаңғақтармен қандай өнімдерді сату керектігін көру үшін пайдалануға болады

Ассоциация ережесін қолдан, дүкен сөресі

Мақсаты

- көптеген тұтынушылар бірге сатып алатын заттарды анықтау

Тәсіл

- элементтер арасындағы тәуелділіктерді табу үшін штрих-код сканерлерімен жиналған сату орнындағы деректерді өңде
- классикалық ереже
  - {Қағаз, Сүт} --> {Жұмыртқа}
  - егер тұтынушы қағаз бен сүт алса, онда жұмыртқа алу ықтималы жоғары
  - ендеше, сүт қасында жұмыртқа пакетті тапсаң, таң қалма

Тізбекті үлгіні табу анықтамасы

- Объектілер жиынтығы берілген. Әрбір объект оқиғалардың өз уақыт сызбасымен байланысты. Әртүрлі оқиғалар арасындағы күшті дәйекті тәуелділікті болжайтын ережелерді тап.

(А Б) (С) -> (Д Е)

- Ережелер үлгілерді табу арқылы қалыптасады. Үлгілердегі оқиғалардың пайда болуы уақыт шектеулерімен реттеледі.

Регрессия анықтамасы

Тәуелділіктің сызықтық немесе сызықты емес моделін қабылдай отырып, басқа айнымалылардың мәндеріне негізделген берілген үздіксіз мәнді айнымалының мәнін болжаңыз.

Статистикада, нейрондық желілерде жақсы зерттелген.

Мысалдар

- Жарнамалық шығындар негізінде өнімді сату көлемін болжау.
- Температура, ылғалдылық, ауа қысымы және т.б. функциясы ретінде желдің жылдамдығын болжау.
- Қор нарығының индекстерінің уақыттық қатарын болжау.

Ауытқу/аномалияны анықтау анықтамасы

Қалыпты мінез-құлықтан елеулі ауытқуларды анықтайды

Қолданбалар

- Карта бойынша алаяқтықты анықтау

- Желіге шабуылды анықтау

Университет деңгейіндегі әдеттегі желілік трафик тәулігіне 100 миллионнан астам қосылымға жетуі мүмкін

Ақырында, деректерді өндіру деген не?

Алгоритм

- деректерді өңдеудің стратегиялары бар

Деректердің үлкен көлемі

- деректерді өндіру шағын деректер жиынына назар аудармайды

Білімді ашу

- деректерді өндіру арқылы біз гипотезаны растайтын немесе жоққа шығаратын жаңа білім табамыз деп күтеміз
- деректерден "жасырын" байланыстарды табамыз деп күтеміз

Деректерді өндіруге не жатпайды, жатады?

Деректерді өндіруге жатпайды?

- телефон каталогынан телефон нөмірін іздеген
- "зат" туралы ақпарат алу үшін веб-іздеу жүйесінен сұраған (ақпаратты іздеу)

Деректерді өндіруге не жатады?

- туыстары мен достарының қалауы бойынша әлеуетті пайдаланушыға кітап ұсынған (бірлескен сүзгілеу)
- іздеу жүйесі қайтарған ұқсас құжаттарды контекстіне сәйкес топтаған (құжаттарды кластерлеу)

Деректерді өндірудің қиындықтары

- Масштабтылық
- Өлшемділік
- Күрделі және гетерогенді деректер
- Домен туралы білім
- Деректер сапасы
- Деректерді иелену және тарату
- Құпиялылықтың сақталуы
- Деректер ағыны

Деректерді өндірудің шығу тегі

Машиналық оқытудан, үлгіні танудан, статистикадан және дерекқор жүйелерінен идеяларды тартады.

Дәстүрлі әдістер мыналарға байланысты жарамсыз болуы мүмкін.

- деректердің үлкендігі
- мәліметтердің жоғары өлшемділігі
- мәліметтердің гетерогенді, таралған сипаты

Әдебиеттер

1. Панг-Нинг Тан және Майкл Штайнбах және Випин Кумар, Деректерді өндіруге кіріспе, Пирсон, Аддисон Уэсли, 2-ші басылым, 2018, 1 және 2 тараулар
2. Цзявэй Хан және Мишелин Камбер, Деректерді өндіру тұжырымдамалары мен әдістері, 3-ші басылым, 2012, 1 және 2 тараулар