Пәні: Деректерді интеллектуалды талдау және визуализациялау

Кластерлеу, болжау және визуализация тапсырмалары

Дәріс №7

Шаяхметова А.С., PhD, қауымдастырылған профессор
Жасанды интеллект және Big Data кафедрасы
2026-2027 оқу жылы

Мақсаты және күтілетін оқу нәтижелері

Мақсаты

Деректерді топтастыру, болжау және бейнелеудің өзара байланысын түсіндіру. Әдісті таңдау нәтижені қолдану мақсатына сәйкес болуы керек.

Теориялық нәтиже

Кластер, қашықтық, центроид, тренд, маусымдылық және болжам көкжиегі ұғымдарын ажырату. Формуладағы шамаларды түсіндіру.

Практикалық нәтиже

Деректерді дайындау ретін құру, бірнеше алгоритмді салыстыру және графиктің негізгі қорытындысын айту.

Бағалау нәтижесі

Кластер сапасын silhouette арқылы, болжам қатесін MAE және RMSE арқылы бағалау. Оқу дерегіндегі жақсы нәтиженің жеткіліксіз екенін түсіну.

Жоспары

1. Кластерлеу негіздері және алгоритмдері
2. Кластерлеу сапасы және қолданылуы
3. Болжау және уақыт қатарлары
4. Визуализация және нәтижені түсіндіру

1. Кластерлеу негіздері және алгоритмдері

Кластерлеудің анықтамасы

Белгілері ұқсас объектілерді топтарға біріктіру. Кластар алдын ала берілмейді. Алгоритм топтарды деректердің орналасуына қарай анықтайды.

Кластердің екі қасиеті

Ішкі біртектілік: бір топтағы объектілер жақын. Сыртқы оқшаулану: әртүрлі топтардың объектілері бір-бірінен ерекшеленеді.

Математикалық көрінісі

Әр объект $x_i = (x_{i1}, \ldots, x_{ip})$ белгі векторы ретінде беріледі. Ұқсастық қашықтықпен немесе арнайы ұқсастық өлшемімен бағаланады.

Қолданылатын жағдайлар

Белгісіз клиент топтарын, ұқсас құжаттарды немесе датчик режимдерін іздеу. Табылған топтардың мағынасын маман түсіндіреді.

Жіктеу, кластерлеу және регрессияны салыстыру

Сипаттама	Жіктеу	Кластерлеу	Регрессия
Мақсаты	Белгілі класс таңдау	Жасырын топ табу	Сандық мән болжау
Мақсатты белгі	Класс берілген	Класс берілмейді	Сандық мән берілген
Оқыту	Мұғаліммен	Мұғалімсіз	Мұғаліммен
Нәтиже	Спам / қалыпты	1, 2, 3 топ	Баға немесе көлем
Мысал	Ақау түрін анықтау	Клиент сегменті	Жеткізу уақытын болжау

Белгілерді дайындау және қашықтық өлшемі

Дерек сапасын тексеру

Қайталанатын жазбалар, жетіспейтін мәндер және ерекше бақылаулар тексеріледі. Әр түзету талдау мақсатына негізделуі керек.

Масштабты сәйкестендіру

$z = (x - \mu) / \sigma$. Табыс мыңдаған бірлікпен, ал жиілік ондықпен өлшенсе, үлкен масштабтағы белгі қашықтыққа басым әсер етеді.

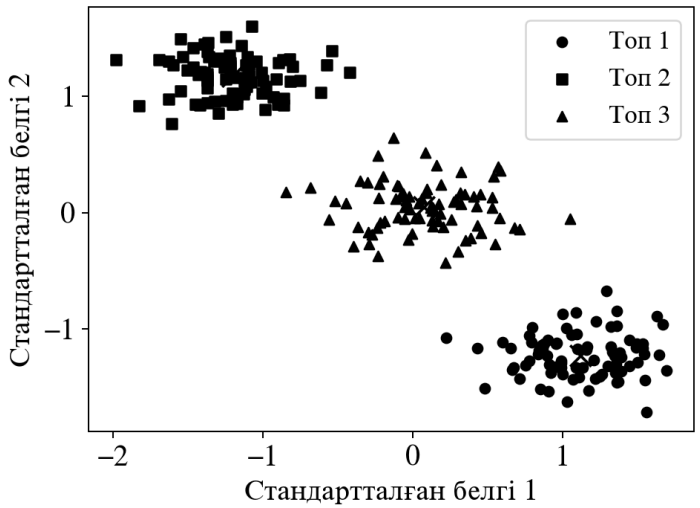
Қашықтық формуласы

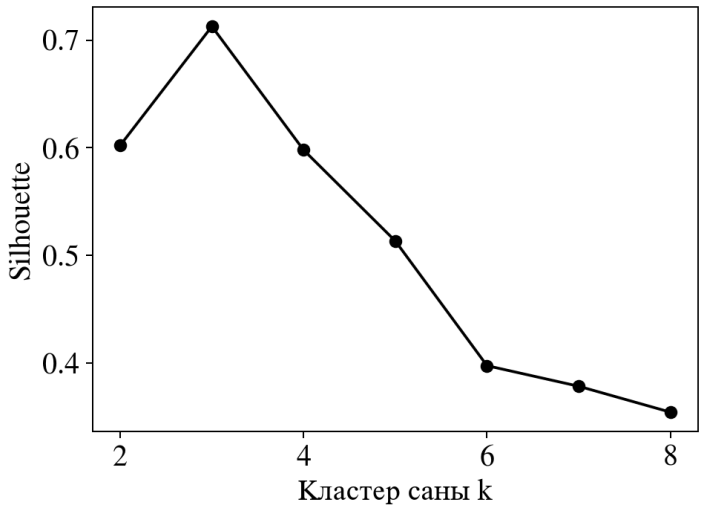
Евклидтік қашықтық: $d(x,y) = \sqrt{\Sigma(x_j - y_j)^2}$. Мәтін үшін бағыттық ұқсастық маңызды болса, косинустық өлшем қолданылады.

Белгілерді таңдау

ID сияқты техникалық идентификаторды ұқсастық белгісі ретінде қолдану орынсыз. Бір мағынаны қайталайтын белгілер оның салмағын арттыруы мүмкін.

K-means: центроидтар және объектілерді бөлу

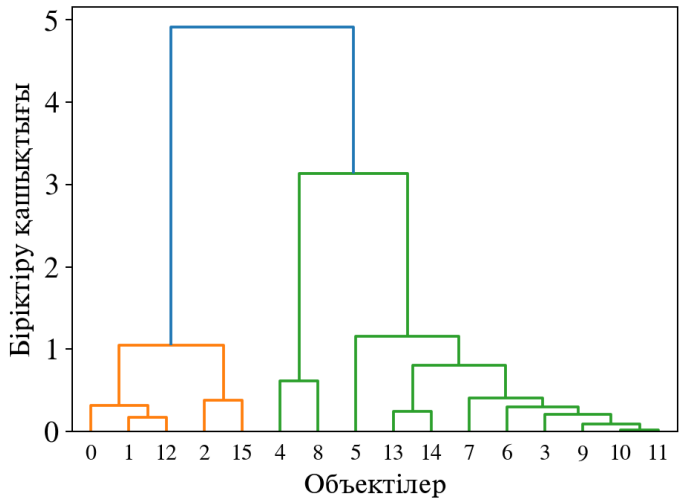


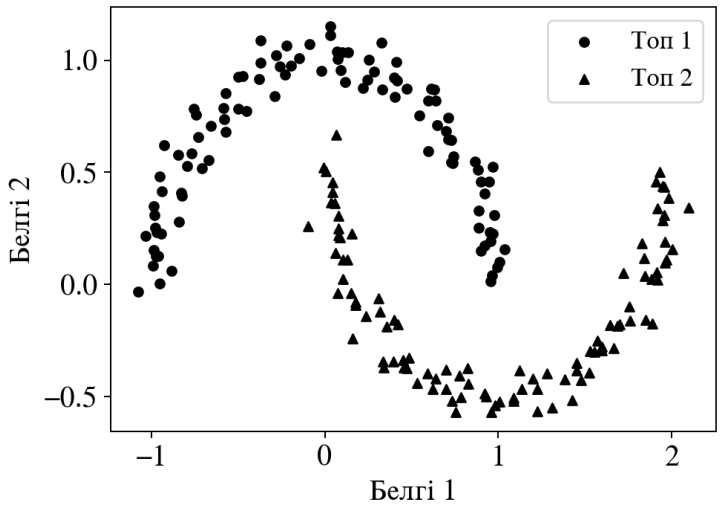


Түсіндірме

240 синтетикалық объект, екі стандартталған белгі. k = 3 кезінде топтар бөлек көрінеді. «×» таңбалары центроидтарды көрсетеді. Silhouette = 0.713.

Иерархиялық кластерлеу және DBSCAN





Түсіндірме

Сол жақта: 16 объектінің Ward дендрограммасы. Біріктіру биіктігі топтар айырмашылығын көрсетеді. Оң жақта: екі иілген топты DBSCAN анықтады.

Кластерлеу алгоритмдерін таңдау

Әдіс	Негізгі параметр	Қолайлы жағдай	Шектеуі
K-means	k, инициализация	Ықшам дөңгелек топтар	Масштаб пен шеткі мәнге сезімтал
Иерархиялық	Linkage, қашықтық	Сатылы құрылым	Ірі деректе баяу
DBSCAN	eps, min_samples	Еркін пішін және шу	Тығыздық айырмасы қиындық туғызады
Gaussian mixture	Компонент саны	Топқа тиістілік ықтималдығы	Таралу болжамдарына тәуелді
MiniBatch K-means	k, batch_size	Көп объект, шағын пакеттер	Жылдамдық үшін жуықтау жасайды

2. Кластерлеу сапасы және қолданылуы

Негізгі сұрақ

Топтардың бөлінуі сандық жағынан жақсы ма және практикалық мағынасы бар ма? Екі сұраққа да жауап қажет.

Ішкі бағалау

Класс белгілері жоқ кезде топ ішіндегі жақындық пен топтар арасындағы алшақтық тексеріледі. Мысалы, silhouette қолданылады.

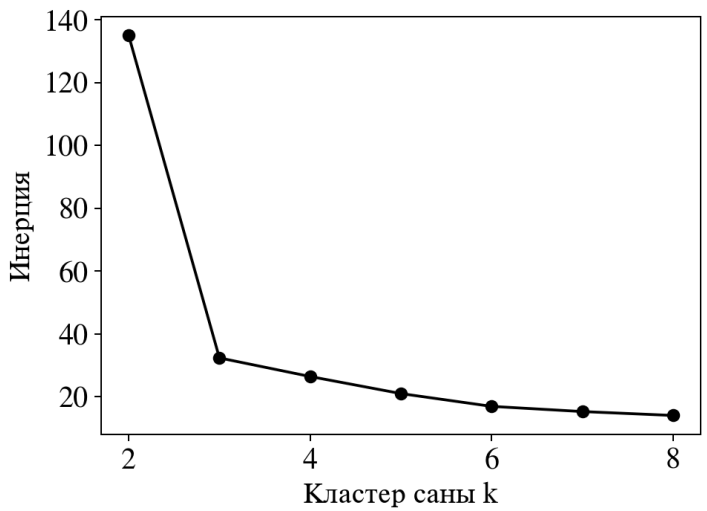
Сыртқы бағалау

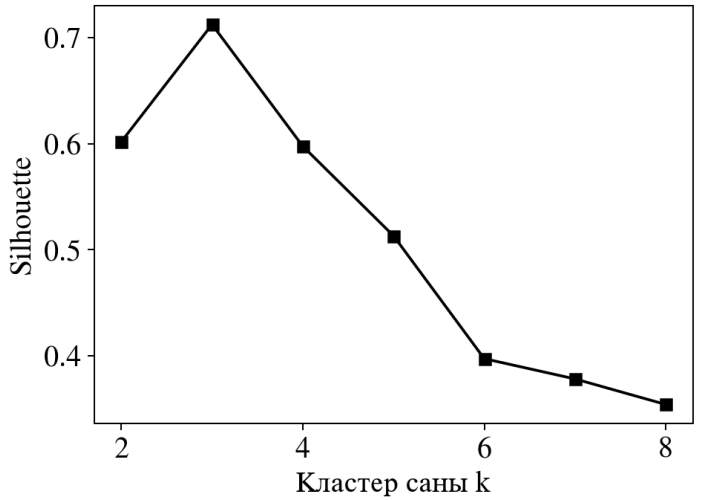
Егер салыстыруға арналған сенімді белгілер бар болса, алынған топтар сол белгілермен салыстырылады. Кластер нөмірлері ауысуы мүмкін.

Тұрақтылық және түсіндіру

Деректің бір бөлігі немесе параметр өзгергенде топтар сақтала ма? Топ профилі нақты әрекет ұсынуға мүмкіндік бере ме?

Кластер санын таңдау: elbow және silhouette





Түсіндірме

Инерция k артқан сайын төмендейді, сондықтан ең кіші мәнді ғана іздеу жеткіліксіз. «Иілу нүктесі» артық топтың пайдасы азайған жерді көрсетеді.

Кластер сапасының өлшемдері

Өлшем	Қалай түсіндіріледі	Практикалық ескерту
Silhouette	$s = (b - a) / \max(a,b)$	a: өз тобына орташа қашықтық, b: ең жақын өзге топқа
$s \approx 1$	Топтан жақсы бөлінген объект	Орташа мәнмен бірге жеке топтарды қарау
$s \approx 0$	Шекара маңындағы объект	Топтар қабаттасуы мүмкін
$s < 0$	Басқа топқа жақынырақ объект	Белгі, масштаб және әдісті қайта тексеру
Тұрақтылық	Қайта үлгілеуде ұқсас құрылым	Топ нөмірін емес, құрамын салыстыру

Практикалық мысал: клиенттерді сегменттеу

RFM белгілері → **Масштабтау** → **K-means** → **Топ профилі**

R: соңғы сатып алудан өткен күн. F: сатып алу саны. M: жалпы шығын.

Төмен R және жоғары F тұрақты белсенді клиенттерді сипаттайды.

Әр топтың профилі мен көлемін талдап, нақты ұсыныс әзірлеу керек.

3. Болжау және уақыт қатарлары

Уақыт қатары

Уақыт ретімен жиналған бақылаулар: күндік жүктеме, айлық сатылым немесе сағаттық энергия тұтынуы. Уақыт ретін сақтау маңызды.

Мақсатты көрсеткіш

Нені болжау керек: сұраныс саны, ақша көлемі немесе жеткізу ұзақтығы? Бірлік пен жинақтау жиілігі алдын ала анықталады.

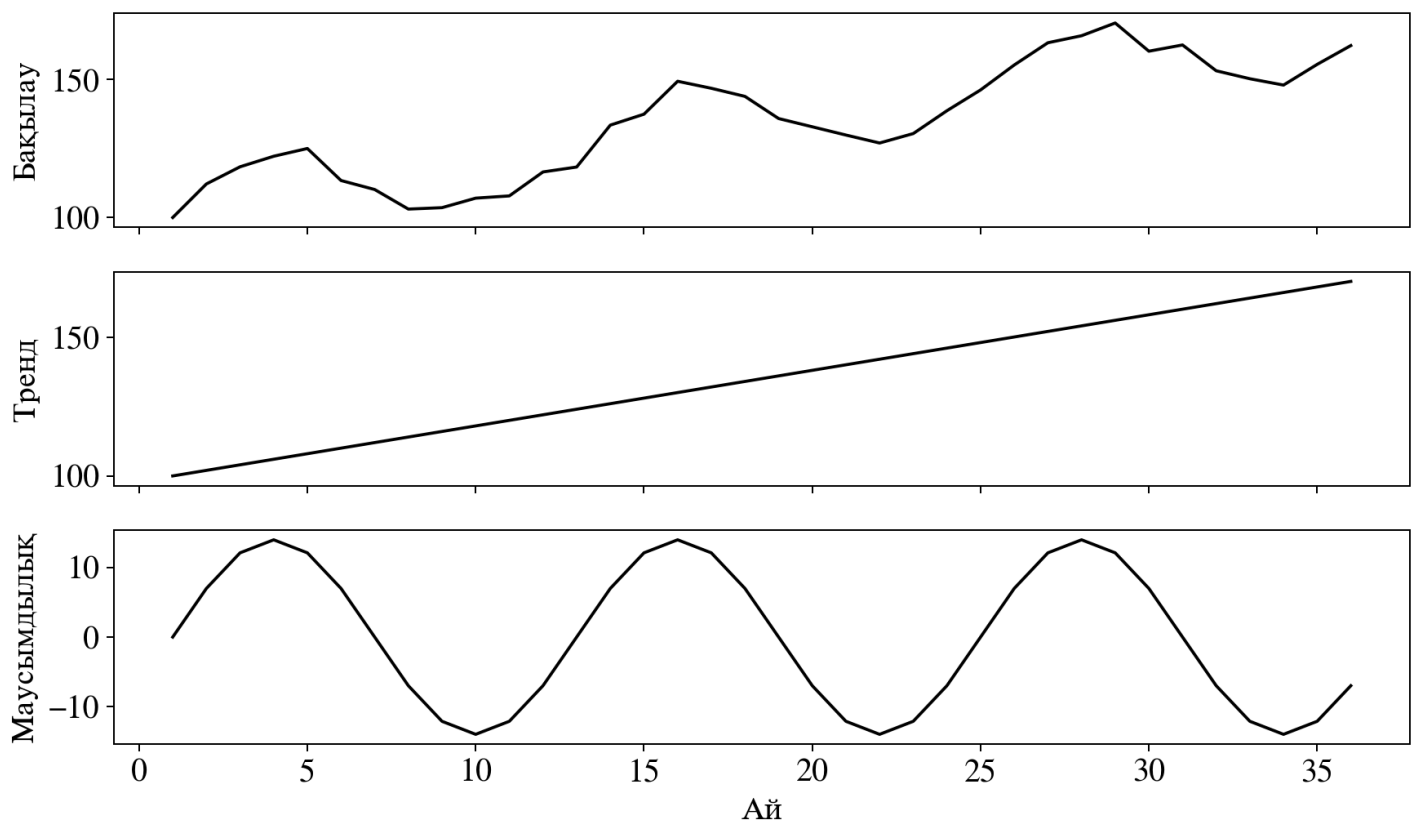
Негізгі заңдылықтар

Тренд ұзақ мерзімді бағытты, маусымдылық тұрақты қайталануды сипаттайды. Қалдық бөлік осы компоненттерден кейін қалады.

Нәтижені қолдану

Болжам қорды жоспарлау, қызметкерлер санын анықтау немесе қуатты бөлу үшін қолданылады. Қате құны шешімнің мақсатына байланысты.

Уақыт қатары: тренд және маусымдылық



Түсіндірме

36 айлық синтетикалық қатар. $y_t = 100 + 2t + 14\sin(2\pi t/12) + \varepsilon_t$. Тренд: ай сайын 2 бірлік өсу. Маусымдық период: 12 ай. Қалдыққа кездейсоқ шу қосылған.

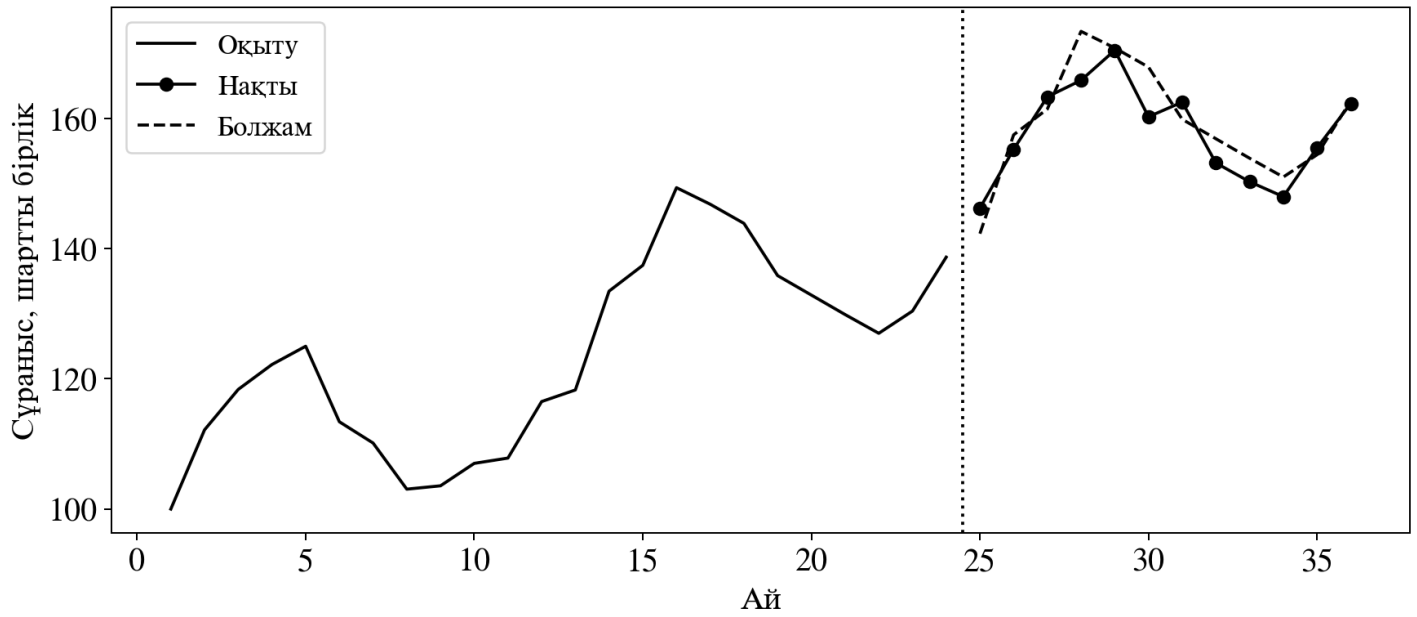
Болжаудың уақыт параметрлері

Параметр	Анықтамасы	Айлық сұраныс мысалы
Кезең	Бір бақылаудың уақыт бірлігі	1 ай
Көкжиек	Болашақта қамтылатын кезең саны	Алдағы 6 ай
Жаңарту аралығы	Болжамды қайта есептеу жиілігі	Ай сайын
Маусымдық период	Қайталанудың ұзақтығы	12 ай
Оқыту терезесі	Модельге берілетін тарихи аралық	Соңғы 24 немесе 36 ай

Болжау әдістерін салыстыру

Әдіс	Негізгі идея	Қолдану шарты	Ескерту
Naive	Соңғы мәнді қайталау	Қарапайым бастапқы салыстыру	Трендті арнайы модельдемейді
Seasonal naive	Өткен маусым мәнін алу	Тұрақты маусымдылық	Маусым өзгерсе қателік өседі
ETS	Деңгей, тренд, маусымды тегістеу	Тұрақты уақыт құрылымы	Компоненттерді дұрыс таңдау қажет
ARIMA	Автокорреляция және айырмалау	Уақыттық тәуелділік	Диагностика және параметр таңдау
ML регрессия	Lag және сыртқы белгілер	Көп факторлы болжам	Болашақ белгілер қолжетімді болуы керек

Оқу тәжірибесі: сұранысты 12 айға болжау



Түсіндірме

Алғашқы 24 ай: оқыту бөлігі. Соңғы 12 ай: тест бөлігі. Болжам: өткен жылдың сәйкес айы + 24 бірлік. MAE = 3.16; RMSE = 3.90.

Болжам дәлдігін бағалау

Метрика / тәсіл	Формула немесе мағына	Қолданудағы ескерту
Қате	$e_t = y_t - \hat{y}_t$	Нақты мән минус болжам
MAE	$\Sigma\|e_t\| / n$	Бастапқы өлшем бірлігінде
RMSE	$\sqrt{(\Sigma e_t^2 / n)}$	Ірі қателіктерге көбірек салмақ береді
MAPE	$100 \times \Sigma\|e_t / y_t\| / n$	Нақты мән нөл болса анықталмайды
Rolling validation	Оқыту терезесін уақытпен алға жылжыту	Әр тест кезеңінде тек өткен дерек қолданылады

4. Визуализация және нәтижені түсіндіру

Зерттеу міндеті

Деректегі таралуды, шеткі мәндерді, топтарды және тәуелділіктерді байқау. Бір график барлық сұраққа жауап бермейді.

Нәтижені ұсыну

Көрермен графиктен қандай қорытынды алуы керек екенін анықтау. Ось атауын, бірлікті, уақыт аралығын және дерек көзін көрсету.

Салыстыру міндеті

Әдістердің нәтижелерін бірдей масштабта көрсету. Кестеде нақты сан, графикте құрылым және тенденция жақсы көрінеді.

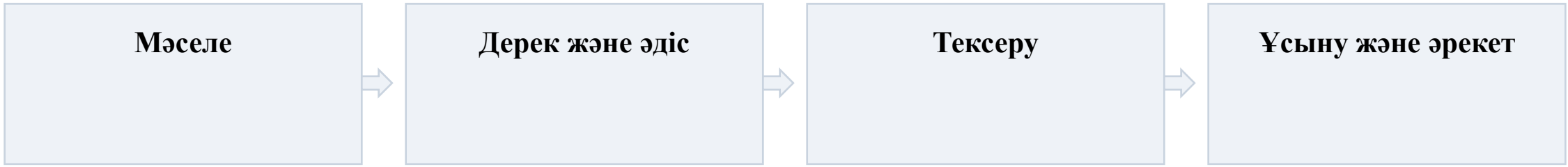
Шешімге өту

Көрсеткіш өзгерісі ұсынылған әрекетпен байланыстырылуы керек. Белгісіздік пен мысал деректің мәртебесі жасырылмайды.

Талдау сұрағына сәйкес визуализация

Сұрақ	Қолайлы көрініс	Нені көрсету керек
Көрсеткіш уақытпен қалай өзгереді?	Сызықтық график	Уақыт, бірлік, болжам аймағы
Топтар қаншалықты ерекшеленеді?	Бағандық диаграмма	Бірдей шкала, топ көлемі
Мәндер қалай таралған?	Гистограмма / boxplot	Интервал, медиана, шеткі мән
Екі белгі байланысқан ба?	Нүктелік диаграмма	Белгі атауы, топ таңбасы
Қадамдар қалай орындалады?	Процесс схемасы	Кезек, кіріс, шығыс
Нақты нәтижелер қандай?	Кесте	Метрика, әдіс, тексеру дерегі

Толық аналитикалық процестің схемасы



1. Мақсат пен бағалау өлшемін анықтау.
2. Деректерді тазалау және белгілерді дайындау.
3. Әдістерді бірдей шарттармен салыстыру.
4. Нәтижені график, кесте және схема арқылы ұсыну.
5. Жаңа дерек түскенде сапаны қайта тексеру.

Нәтижені қате түсіндірудің алдын алу

Жиі кездесетін қателер

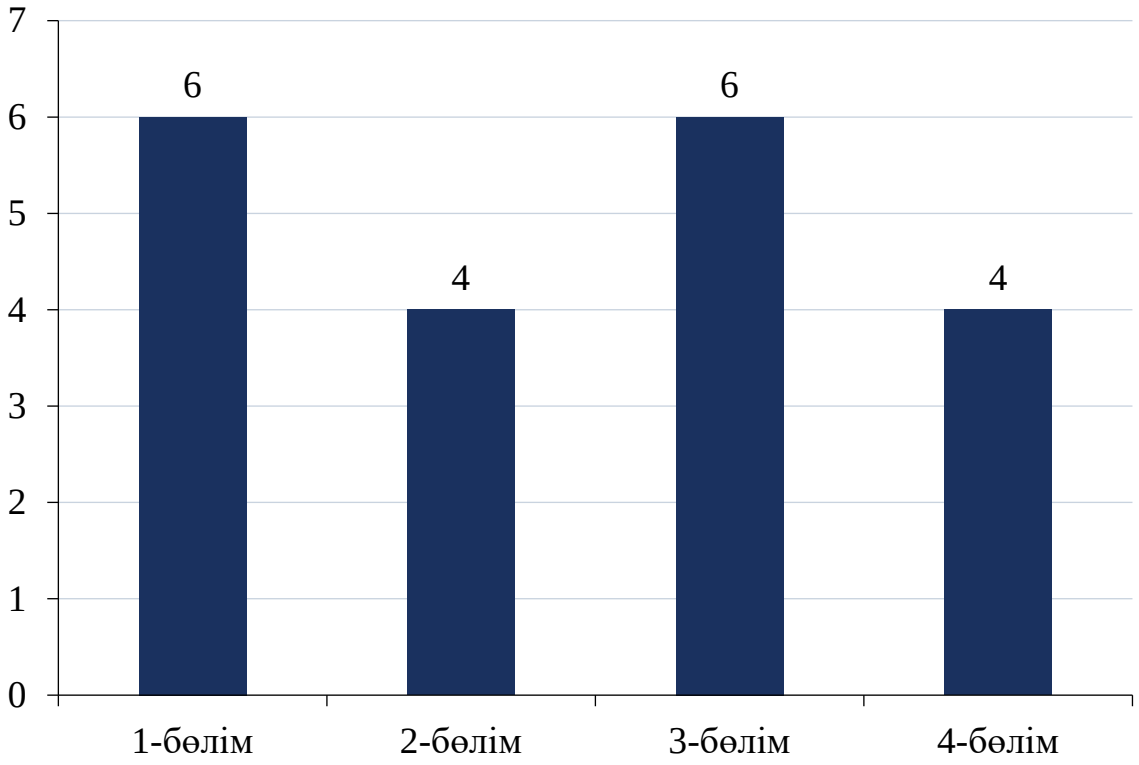
Масштабталмаған белгілерден алынған топты нақты құрылым деп қабылдау. Кластер нөмірін рейтингке теңестіру. Болжамды тек оқыту қатесімен бағалау.

Дұрыс түсіндіру

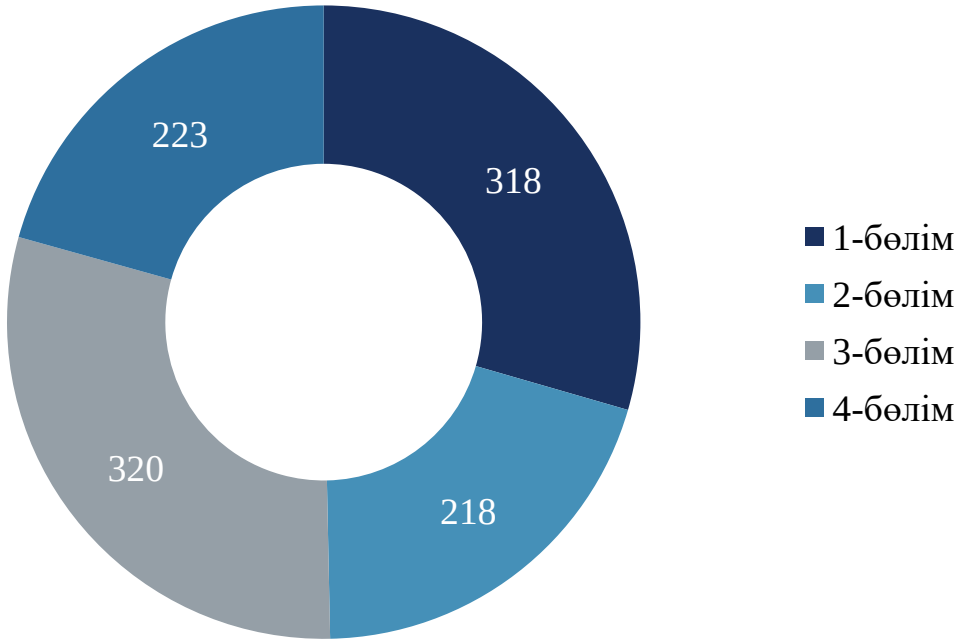
Топ профилін белгілермен және объект саны арқылы сипаттау. Таңдаудың бірнеше параметрге тұрақтылығын тексеру. Болжамды кейінгі кезеңмен салыстыру.

Диаграммалар

Бөлімдер бойынша слайд саны



Бөлімдер бойынша мәтін көлемі (сөз саны)



Қорытынды

1. Кластерлеу ұқсас объектілерді топтап, жасырын құрылымды анықтайды.
2. Әдісті таңдау дерек пішініне, масштабына және талдау мақсатына байланысты.
3. Болжауды жаңа уақыт кезеңінде MAE және RMSE арқылы бағалау қажет.
4. График, кесте және схема нәтижені түсіндіруге және шешім қабылдауға көмектеседі.

Дереккөздер: №7 лекция; scikit-learn (Clustering, KMeans, DBSCAN, Silhouette); Hyndman және Athanasopoulos, Forecasting: Principles and Practice.

Кластерлеу, болжау және визуализация тапсырмалары

Дәріс №7