Пәні: Деректерді интеллектуалды талдау және визуализациялау

Деректерді интеллектуалды талдау әдістері (Data Mining)

Дәріс №3

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

Дәрістің мақсаты

Data Mining ұғымын және оның деректерді талдаудағы орнын түсіндіру. KDD процесінің кезеңдерімен және деректерді алдын ала өңдеумен таныстыру. Классификация, кластерлеу және ассоциативті ережелер әдістерін салыстыру. Data Mining-тің қолдану салалары мен бағдарламалық құралдарын қарастыру.

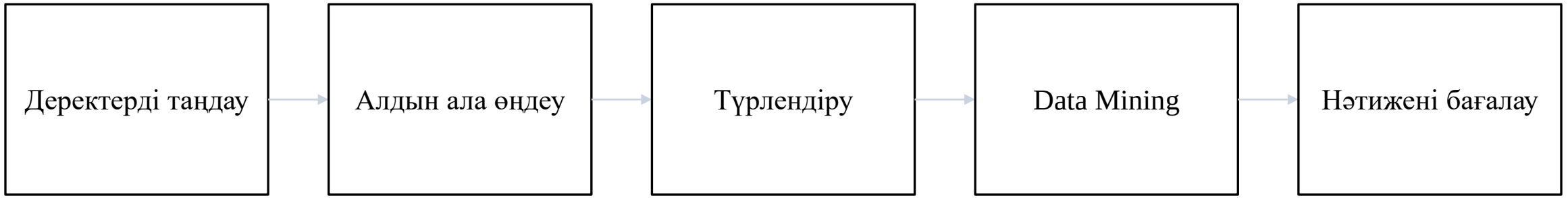
Дәріс жоспары

1. Data Mining ұғымы және KDD процесі
2. Деректерді алдын ала өңдеу
3. Негізгі әдістер: классификация, кластерлеу, ассоциация
4. Қолдану салалары және құралдары

Data Mining дегеніміз не?

- Data Mining - үлкен көлемді деректерден бұрын белгісіз, пайдалы заңдылықтарды табу процесі
- Деректер қоймасы, статистика және машиналық оқыту әдістеріне сүйенеді
- Нәтиже: модельдер, ережелер, топтар және болжамдар
- Қолданушыға шешім қабылдауға көмектесетін білім алуға бағытталған

KDD процесінің кезеңдері



KDD (Knowledge Discovery in Databases) - деректерден білім алудың толық циклі; Data Mining осы циклдің ең маңызды кезеңі.

Data Mining және классикалық статистика

Белгі	Классикалық статистика	Data Mining
Деректер көлемі	Шағын таңдама	Өте үлкен көлем
Тәсіл	Алдын ала гипотеза тексеріледі	Заңдылықтар автоматты табылады
Нәтиже	Параметрлер бағасы	Модельдер мен ережелер

Data Mining есептерінің түрлері

Сипаттау

Кластерлеу, ассоциативті ережелер, аномалияларды іздеу

Болжау

Классификация, регрессия, уақыттық қатарларды болжау

Алдын ала өңдеу кезеңдері

Тазалау
Қателер мен шуды жою

Біріктіру
Әртүрлі көздерді бір жүйеге жинау

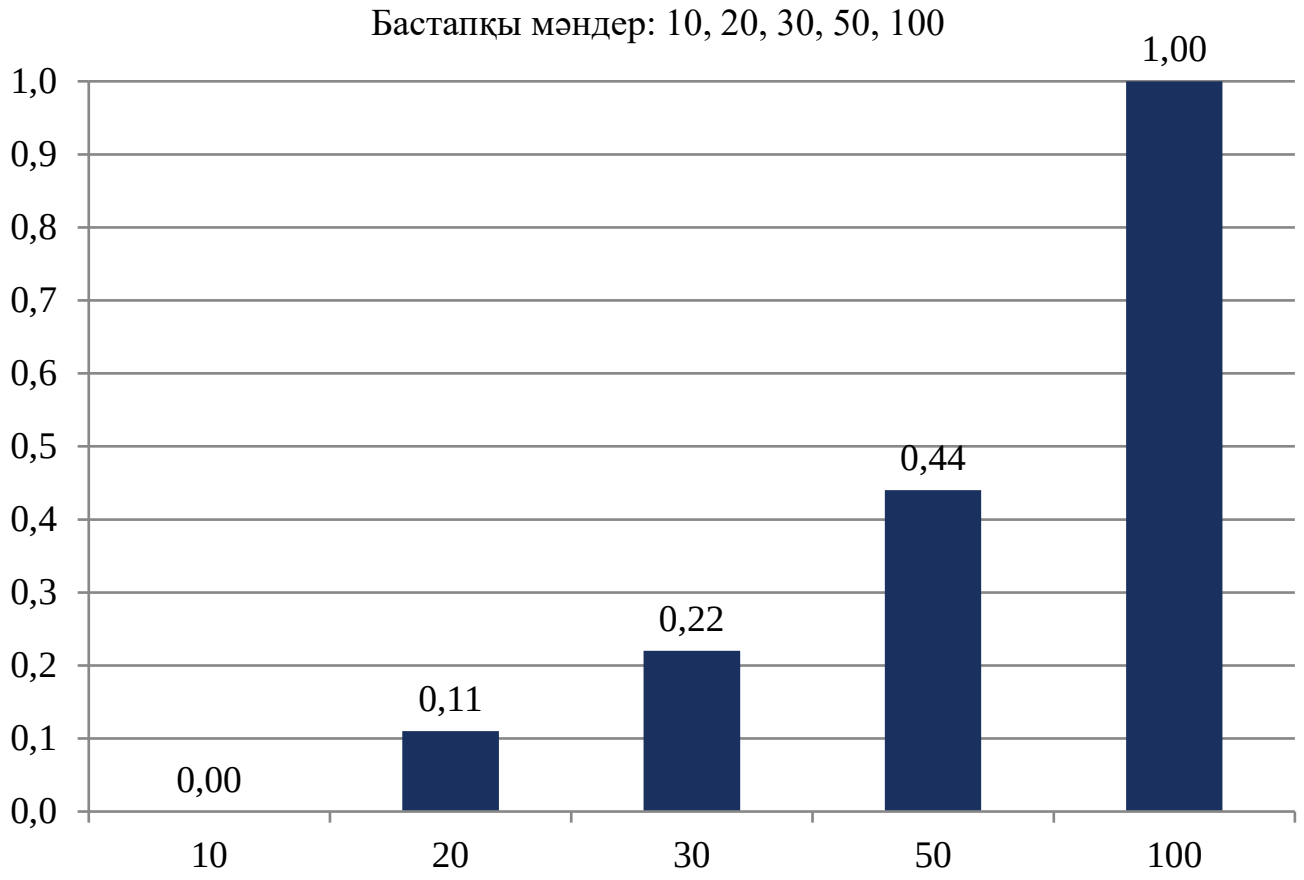
Түрлендіру
Нормалау, кодтау

Қысқарту
Белгілерді таңдау, таңдама алу

Жетіспейтін мәндермен жұмыс

Әдіс	Мәні
Жолды жою	Қалдығы аз болғанда қолданылады
Орташамен толтыру	Сандық белгілер үшін
Жиі кездесетін мән	Категориялық белгілер үшін
Модельмен болжау	Басқа белгілер арқылы толтыру

Нормалау мысалы: min-max әдісі



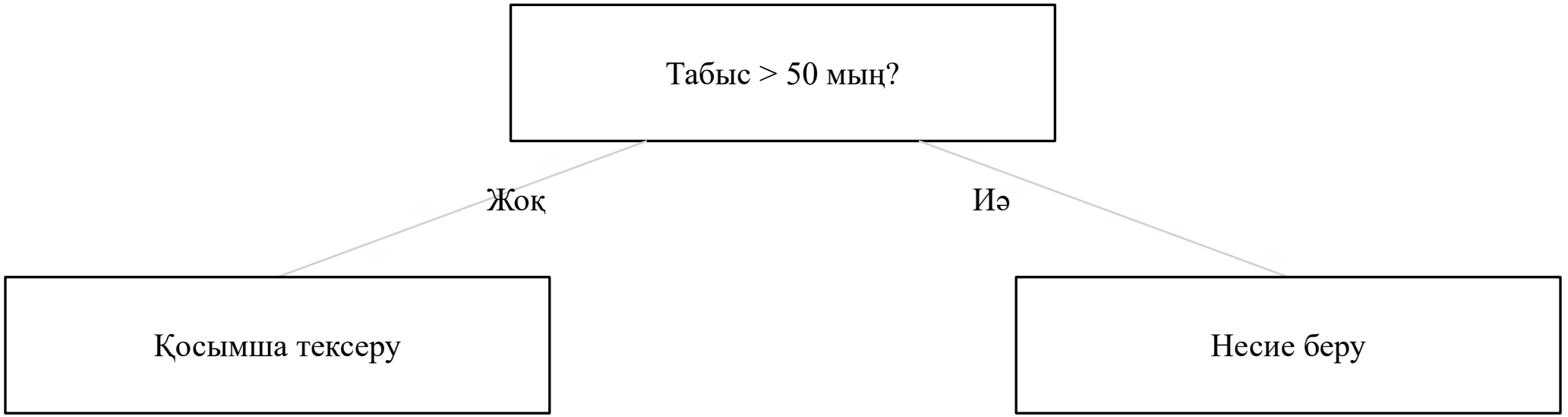
$x' = (x - min) / (max - min)$

Мәндер [0; 1] аралығына келтіріледі, сондықтан әртүрлі шкалалы белгілерді салыстыруға болады.

Белгілерді таңдау және таңдама алу

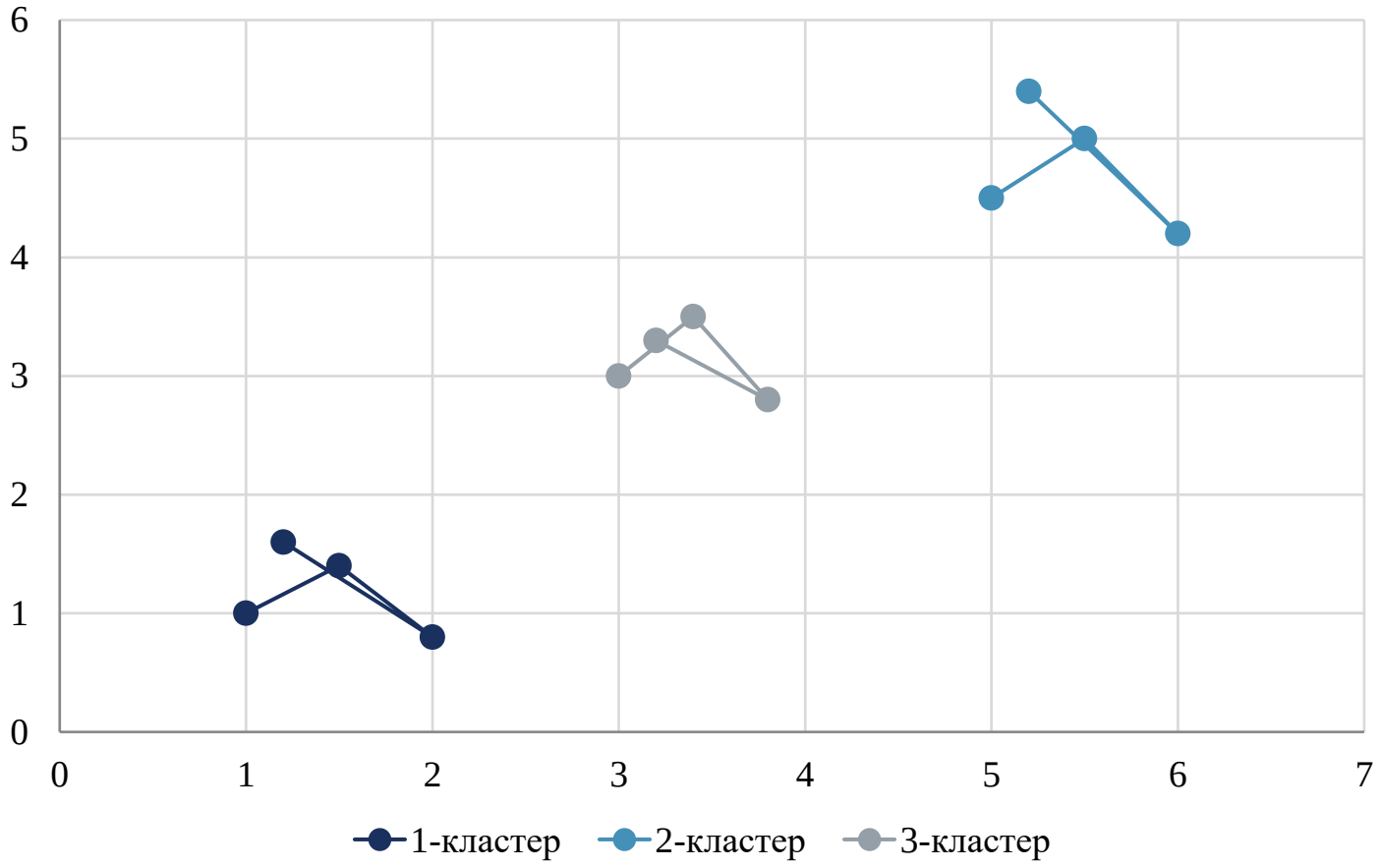
- Маңызды емес және қайталанатын белгілер алынып тасталады
- Корреляциясы жоғары белгілердің біреуі қалдырылады
- Үлкен деректерден кездейсоқ немесе стратификацияланған таңдама алынады
- Қысқарту есептеу уақытын азайтады, модель сапасын төмендетпейді

Классификация: шешім ағашы



Классификация объектіні алдын ала белгілі сыныпқа жатқызады (мысалы, несие тәуекелі).

Кластерлеу: объектілерді топтау



Алдын ала сыныптар жоқ: объектілер ұқсастығы бойынша топталады. Мысал алгоритмдер - k-means, DBSCAN.

Ассоциативті ережелер

Ереже	Support	Confidence
нан → сүт	40%	67%
сүт → нан	40%	80%

Мысал: нан 60%, сүт 50%, екеуі бірге 40% чектерде кездеседі. Support — ереженің жиілігі, confidence = support(A,B) / support(A).

Әдістерді салыстыру

Әдіс	Тапсырма	Алгоритм мысалы
Классификация	Сыныпты болжау	C4.5, kNN, SVM
Регрессия	Санды шаманы болжау	Сызықтық регрессия
Кластерлеу	Топтарды табу	k-means, DBSCAN
Ассоциация	Ережелерді табу	Apriori

Қолдану салалары

Банк
Несиелік скоринг, алаяқтықты анықтау

Медицина
Диагностикаға көмек, емдеу нәтижесін болжау

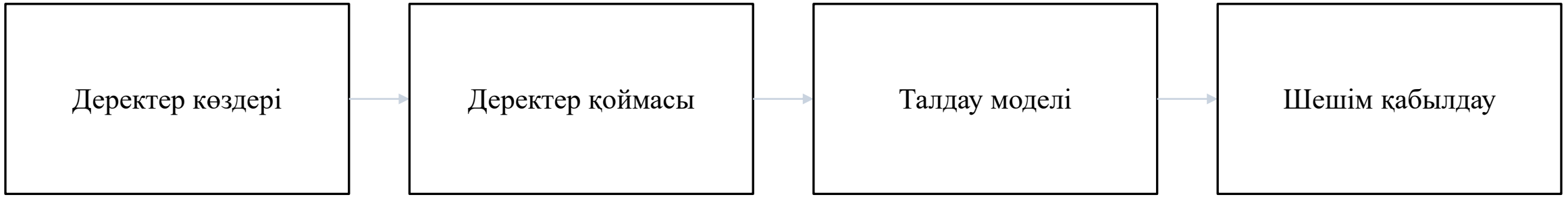
Сауда
Себет талдауы, сатылымды болжау

Телеком
Абоненттердің кетуін болжау

Бағдарламалық құралдар

Құрал	Ерекшелігі
Weka	Тегін, оқу мақсатына ыңғайлы
RapidMiner	Визуалды жұмыс процестері
KNIME	Модульдік, ашық платформа
Python (scikit-learn)	Икемді, кітапханалары көп

Data Mining жүйесінің схемасы



Қиындықтар мен тәуекелдер

- Деректер сапасының төмендігі нәтижені бұрмалайды
- Шамадан тыс үйрену (overfitting) модельдің жалпылау қабілетін төмендетеді
- Жеке деректердің құпиялылығын қамтамасыз ету қажет
- Нәтижені салада білетін маманның түсіндіруі керек

Қорытынды

1. Data Mining - деректерден жасырын заңдылықтарды табатын KDD процесінің негізгі кезеңі
2. Алдын ала өңдеу модельдің сапасына тікелей әсер етеді
3. Классификация, кластерлеу және ассоциация әртүрлі есептерге қолданылады
4. Әдісті таңдау есептің түріне және деректердің сипатына байланысты



Деректерді интеллектуалды талдау әдістері (Data Mining)

Дәріс №3