Пәні: Деректерді интеллектуалды талдау және визуализациялау

Data Mining технологиялық және статистикалық әдістері

Дәріс №4

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

Дәрістің мақсаты

Data Mining-тің технологиялық әдістерін (нейрондық желілер, шешім ағаштары, генетикалық алгоритмдер) қарастыру. Статистикалық әдістердің негіздерін: сипаттамалық статистика, корреляция, регрессия түсіндіру. Гипотезаларды тексеру және уақыттық қатарларды талдау тәсілдерімен таныстыру. Модель сапасын бағалау және нәтижені визуализациялау әдістерін көрсету

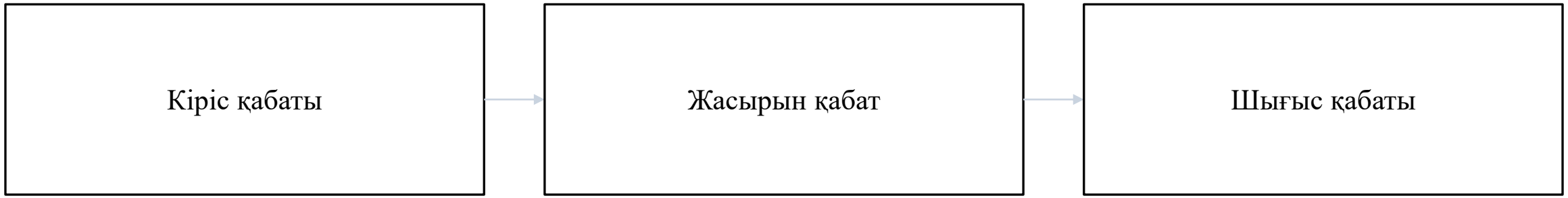
Дәріс жоспары

1. Data Mining технологиялық әдістері
2. Статистикалық әдістердің негіздері
3. Гипотезаларды тексеру және болжау
4. Нәтижелерді бағалау және визуализациялау

Технологиялық әдістерге шолу

Әдіс	Негізі	Қолдану мысалы
Нейрондық желі	Өзара байланысқан нейрондар	Кескінді тану
Шешім ағашы	Белгілер бойынша тармақтау	Несие скоринг
k ең жақын көрші	Объектілер арасындағы қашықтық	Ұсыныс жүйесі
Генетикалық алгоритм	Іріктеу және мутация	Параметрлерді оңтайландыру

Жасанды нейрондық желі



Кіріс қабаты деректерді қабылдайды, жасырын қабаттар салмақтар арқылы белгілерді түрлендіреді, шығыс қабаты нәтиже (сынып немесе сан) береді. Желі қатені азайту арқылы үйренеді.

Шешім ағашы және k ең жақын көрші

Шешім ағашы

Ереже түрінде түсіндіріледі, сандық және категориялық деректермен жұмыс істейді

k ең жақын көрші

Жаңа объект өзіне жақын k көршінің көпшілік сыныбына жатқызылады

Генетикалық алгоритм

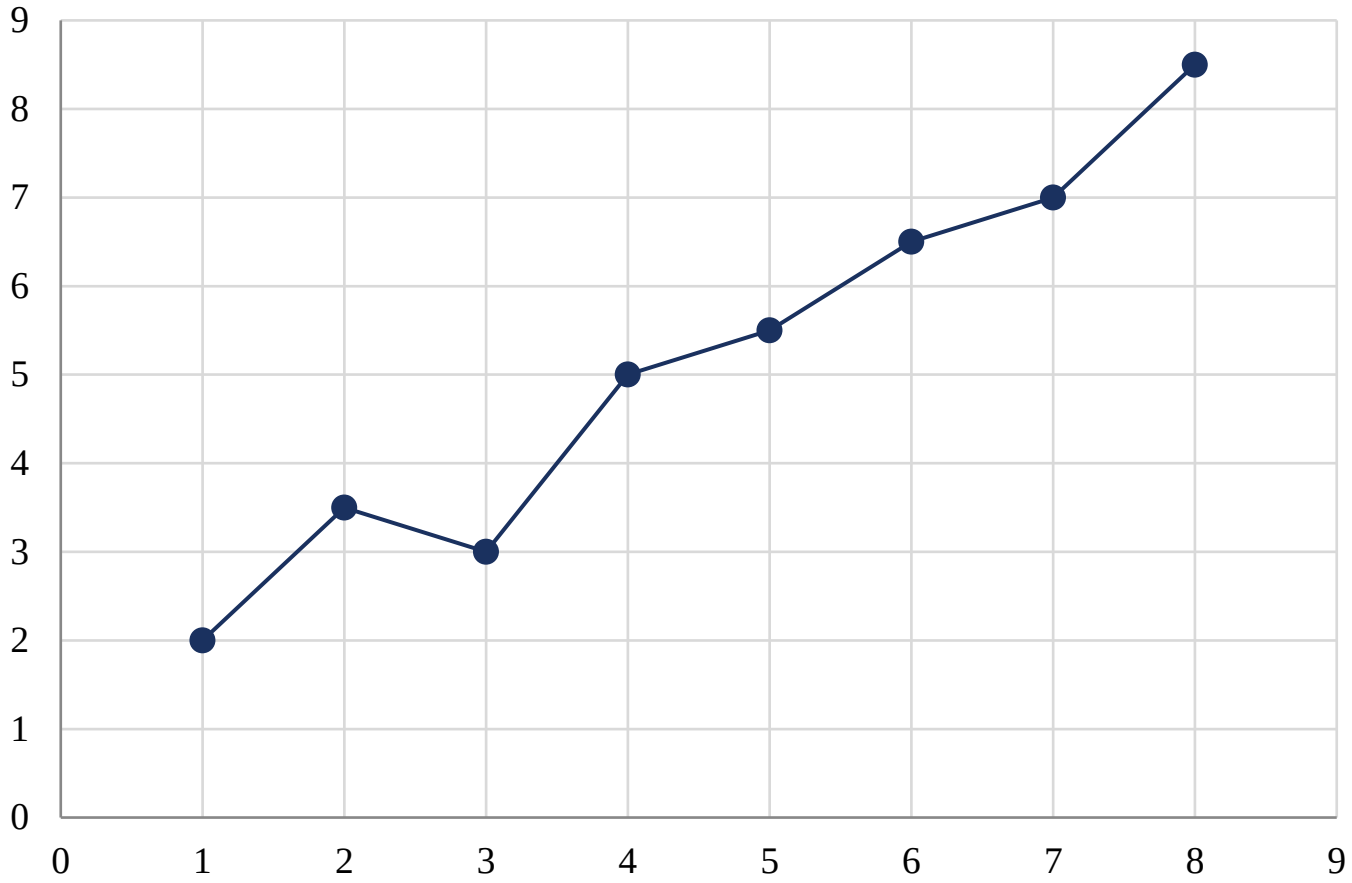


Процесс жақсы шешім табылғанша немесе ұрпақ саны шектен асқанша қайталанады.

Сипаттамалық статистика

Көрсеткіш	Мағынасы
Орташа мән	Мәндердің арифметикалық ортасы
Медиана	Реттелген қатардың ортасындағы мән
Мода	Ең жиі кездесетін мән
Стандартты ауытқу	Мәндердің орташадан шашырауы

Корреляция



Корреляция коэффициенті r екі шаманың сызықтық байланысын өлшейді және -1 мен +1 аралығында болады.

r ≈ +1 - тура, r ≈ −1 - кері, r ≈ 0 - байланыс жоқ.

Сызықтық регрессия

$y = a + b \cdot x$, $b = \Sigma(x - \bar{x})(y - \bar{y}) / \Sigma(x - \bar{x})^2$, $a = \bar{y} - b \cdot \bar{x}$

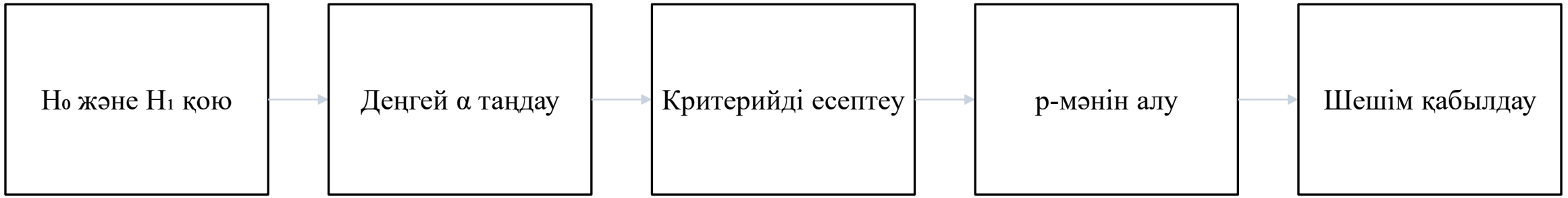
x	1	2	3	4	5
y	2	4	5	4	5

Мысал: $\bar{x} = 3$, $\bar{y} = 4$, $b = 0{,}6$, $a = 2{,}2$. Модель: $y = 2{,}2 + 0{,}6x$; $x = 6$ болғанда болжам $y = 5{,}8$.

Деректер түріне қарай әдіс таңдау

Деректер	Мақсат	Әдіс
Сандық	Байланысты табу	Корреляция
Сандық	Мәнді болжау	Регрессия
Категориялық	Байланысты тексеру	χ^2 критерийі
Уақыттық қатар	Болашақты болжау	Жылжымалы орташа

Гипотезаларды тексеру схемасы

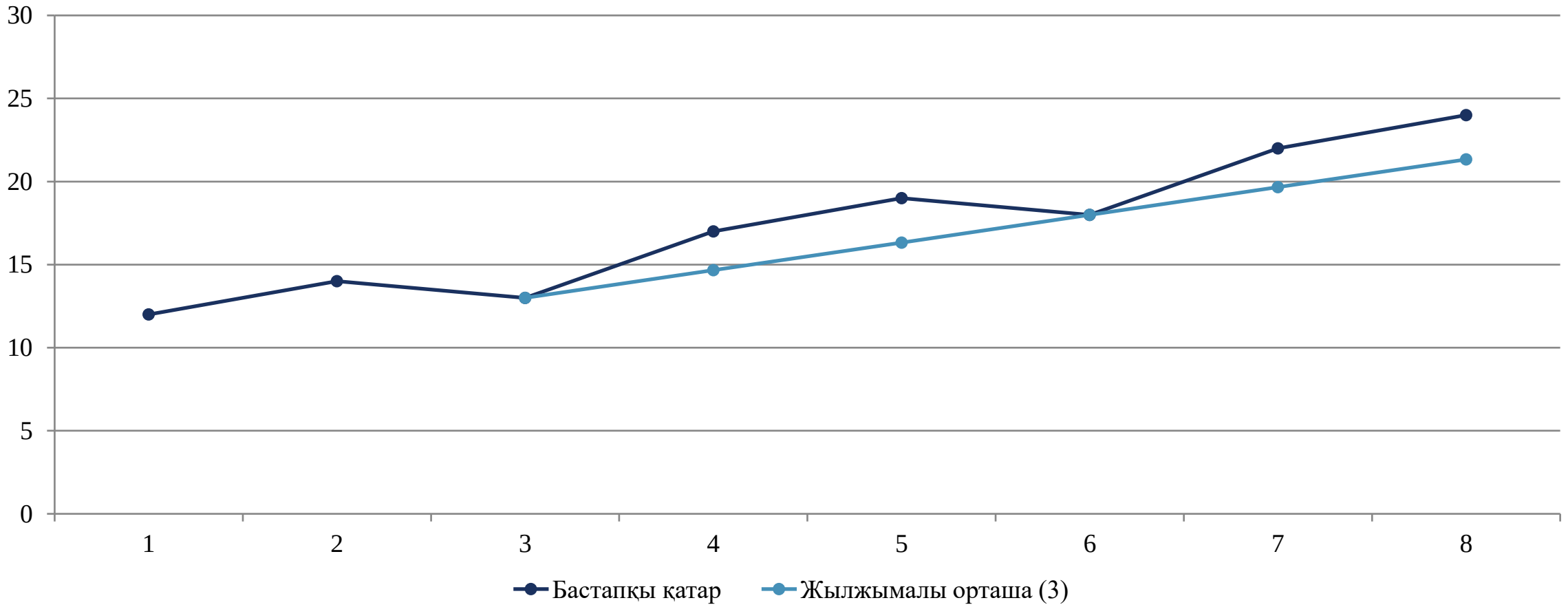


Егер p-мәні α деңгейінен аз болса, нөлдік гипотеза H_0 қабылданбайды.

Статистикалық критерийлер

Критерий	Қашан қолданылады
t-критерий	Екі топтың орташасын салыстыру
ANOVA	Үш және одан көп топты салыстыру
χ^2 критерийі	Категориялық белгілердің тәуелділігі
Манн–Уитни	Қалыпты емес таралған деректер

Уақыттық қатар: жылжымалы орташа



Гипотезаны тексеру қателері

Шешім	H_0 шын	H_0 жалған
H_0 қабылдау	Дұрыс шешім	II текті қате
H_0 қабылдамау	I текті қате (α)	Дұрыс шешім

Қателіктер матрицасы

	Болжам: иә	Болжам: жоқ
Шын мәні: иә	TP = 40	FN = 5
Шын мәні: жоқ	FP = 10	TN = 45

Матрица классификация моделінің қайсы жағдайларда дұрыс, қайсысында қате болжағанын көрсетеді (мысал 100 бақылау бойынша).

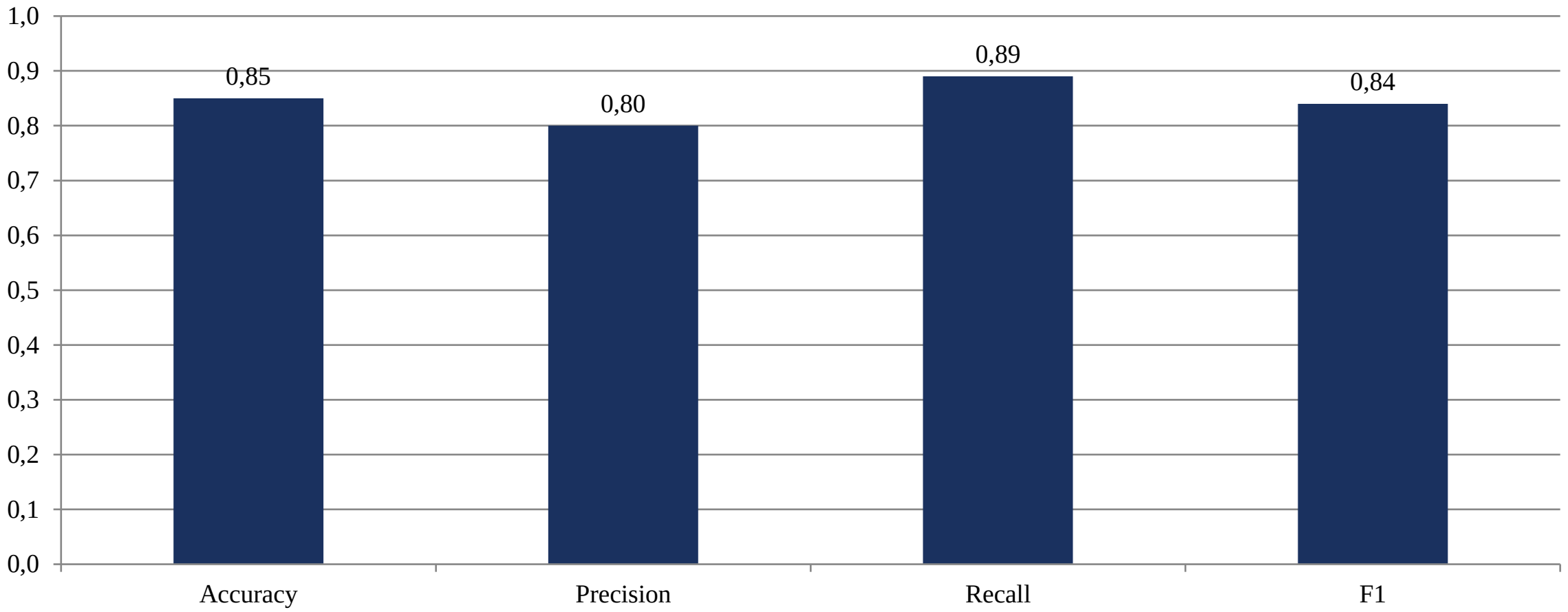
Модель сапасының көрсеткіштері

Көрсеткіш	Формула	Мысал
Accuracy	(TP+TN) / барлығы	0,85
Precision	TP / (TP+FP)	0,80
Recall	TP / (TP+FN)	0,89
F1	2PR / (P+R)	0,84

Визуализация түрлері

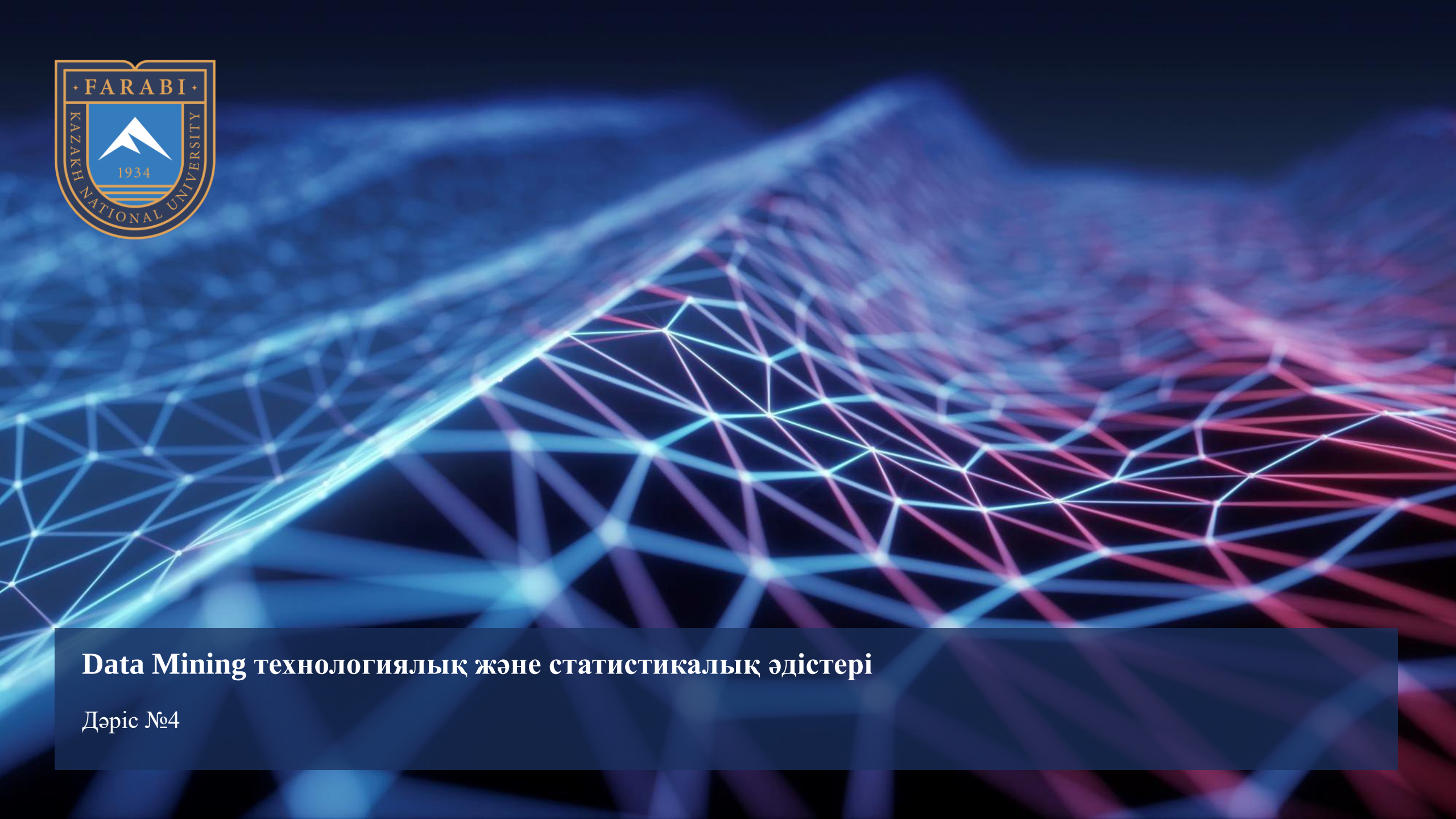
График	Қашан қолданылады
Гистограмма	Бір белгінің таралуы
Шашырау диаграммасы	Екі шаманың байланысы
Сызықтық график	Уақыт бойынша өзгеріс
Жылу картасы	Корреляция матрицасы

Модель көрсеткіштерін салыстыру



Қорытынды

1. Технологиялық әдістер (нейрондық желі, ағаш, kNN, генетикалық алгоритм) күрделі заңдылықтарды табуға мүмкіндік береді
2. Статистикалық әдістер байланысты өлшеп, гипотезаларды дәлелді тексереді
3. Модель сапасы қателіктер матрицасы мен Accuracy, Precision, Recall, F1 арқылы бағаланады
4. Визуализация нәтижені түсінікті етіп жеткізудің негізгі құралы



Data Mining технологиялық және статистикалық әдістері

Дәріс №4