Пәні: Деректерді интеллектуалды талдау және визуализациялау

Шешім ағашы

Дәріс №8

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

Мақсаты және күтілетін оқу нәтижелері

Мақсаты

Шешім ағашының объектіні қалай бөлетінін түсіндіру және оның сапасын тәуелсіз деректе бағалау тәсілін меңгеру.

Құрылымдық нәтиже

Тамыр, ішкі түйін, тармақ және жапырақты ажырату. Тамырдан жапыраққа дейінгі жолды логикалық ереже ретінде жазу.

Математикалық нәтиже

Gini мен энтропияны қарапайым мысалда есептеу. Түйінді бөлу сапасын салмақталған араласу арқылы салыстыру.

Практикалық нәтиже

Гиперпараметрлерді таңдауды түсіндіру, артық үйренуді графиктен көру және confusion matrix бойынша қателерді талдау.

Жоспары

1. Шешім ағашының мәні және құрылымы

2. Ағашты құру және бөліну критерийлері

3. Күрделілікті реттеу және сапаны бағалау

4. Практикалық қолдану және әдістерді салыстыру

1. Шешім ағашының мәні және құрылымы

Анықтамасы

Шешім ағашы деректер кеңістігін шарттар арқылы бөліп, әр аймақта болжам береді. Шарттар тізбегі иерархиялық құрылымға бірігеді.

Жіктеу міндеті

Мақсатты айнымалы категориялық: мысалы, «ақау бар» және «ақау жоқ». Жапырақ класс немесе оның ықтималдығын береді.

Регрессия міндеті

Мақсатты айнымалы сандық: мысалы, жеткізу уақыты. Қарапайым регрессиялық ағашта жапырақтың болжамы сол топтың орташа мәні болуы мүмкін.

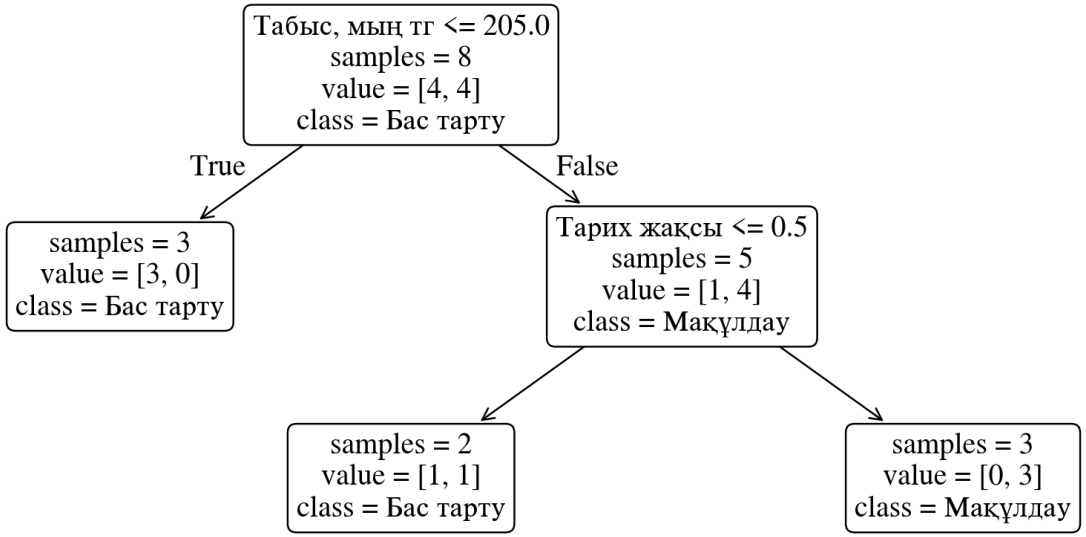
Негізгі артықшылық

Объектінің неге белгілі шешім алғанын жолдағы шарттармен түсіндіруге болады. Өте үлкен ағаштың түсініктілігі төмендейді.

Ағаш элементтері және олардың қызметі

Элемент	Қызметі	Мысал
Тамыр	Алғашқы бөлу, барлық оқу объектілері	Табыс ≤ 200 мың ₸?
Ішкі түйін	Қосымша шартты тексеру	Қайтару тарихы жақсы ма?
Тармақ	Шарт нәтижесі бойынша өту	Иә / жоқ
Жапырақ	Класс немесе сандық болжам	Мақұлдау / бас тарту
Жол	Бір объектіге қолданылатын шарттар	Табыс > шек және тарих жақсы

Оқу мысалы: несие туралы шешім ағашы



Түсіндірме

8 жасанды жазба, 2 белгі. Тамыр табыс деңгейін бөледі. Кейінгі шарт қайтару тарихын тексереді. value: кластар саны; samples: түйіндегі жазбалар саны.

Ағаш жолын логикалық ережеге айналдыру

Ереже қалай құрылады?

Әр ішкі түйіндегі шарт жолға қосылады. Бір жапыраққа апаратын шарттар «ЖӘНЕ» арқылы бірігеді. Егер $x_1 \leq a$ және $x_2 > b$ болса, онда класс = C.

Неліктен пайдалы?

Шешімнің қандай белгілерге сүйенгенін көруге болады. Шекті мәнге жақын объектілерді бөлек тексеруге мүмкіндік береді.

Жіктеу ағашы және регрессиялық ағаш

Өлшем	Жіктеу	Регрессия
Мақсат	Категорияны анықтау	Сандық көрсеткішті бағалау
Жапырақ	Класс үлестері және класс	Мысалы, орташа мән
Бөліну өлшемі	Gini немесе энтропия	Мысалы, квадраттық қателік
Метрика	Precision, Recall, F1	MAE, RMSE
Нәтиже мысалы	Тапсырыс кешігеді / кешікпейді	Жеткізу 3,5 күн болады

2. Ағашты құру және бөліну критерийлері

Бөлу мақсаты

Балалардағы объектілер мақсатты мәні жағынан ата-ана түйінге қарағанда біртекті болуы керек. Сапа критериймен өлшенеді.

Жергілікті таңдау

Алгоритм ағымдағы түйінде белгілер мен шектерді салыстырып, ең пайдалы бөлуді таңдайды. Бұл бүкіл ағаштың жаһандық оңтайлы екенін кепілдемейді.

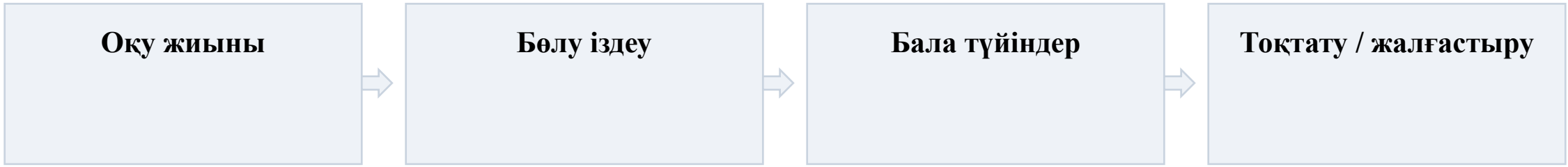
Қайталанатын процесс

Таңдалған шарт деректі екі топқа бөледі. Әр топта процесс тоқтату шарты орындалғанша қайталанады.

CART туралы

CART жіктеу және регрессия міндеттеріне арналған екілік ағаштар құрады. scikit-learn осы тәсілдің оңтайландырылған нұсқасын қолданады.

CART оқыту процесінің схемасы

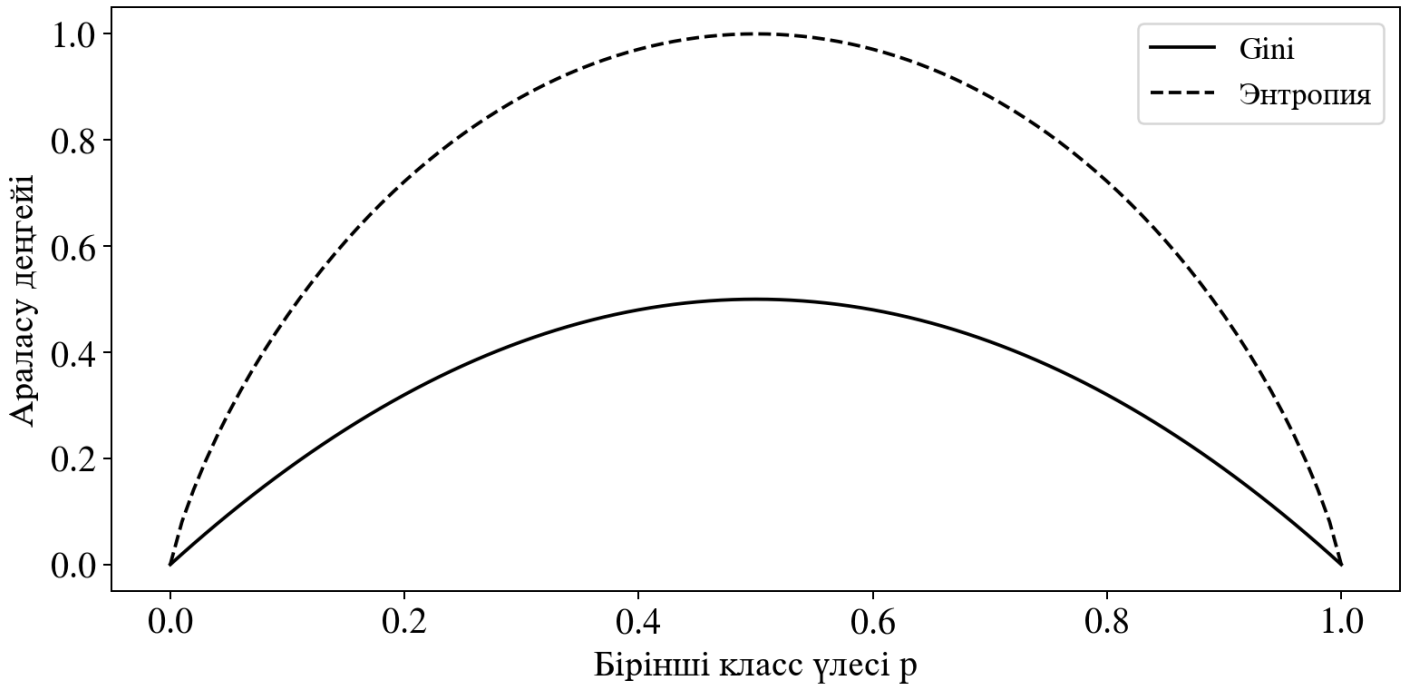


1. Түйіндегі класс үлестерін есептеу.
2. Белгілер мен шектерді салыстыру.
3. Ең пайдалы бөлінуді таңдау.
4. Объектілерді бала түйіндерге бөлу.
5. Тоқтату шартына дейін қадамдарды қайталау.

Gini, энтропия және ақпараттық өсім

Өлшем	Формула	Мағынасы
Gini	$G = 1 - \Sigma p_j^2$	Кластар араласуын бағалайды
Энтропия	$H = -\Sigma p_j \log_2 p_j$	Класс туралы белгісіздік
Салмақталған араласу	$I_{split} = (nL/n)IL + (nR/n)IR$	Балалардың көлемі ескеріледі
Ақпараттық өсім	$IG = H_{parent} - H_{split}$	Энтропия азаюы
Таза түйін	$G = 0, H = 0$	Барлық объект бір класта

Екі класс үшін араласу өлшемдерінің графигі



Түсіндірме

p = 0 немесе p = 1: түйін таза. p = 0,5: екі класс тең араласқан. Gini максимум 0,5; энтропия максимум 1 бит.

Сандық мысал: екі бөлуді Gini арқылы салыстыру

Түйін	A / B саны	Gini	Салмақ
Ата-ана	6 / 4	$1 - 0{,}6^2 - 0{,}4^2 = 0{,}48$	10 объект
A бөлу: сол	4 / 0	0	4/10
A бөлу: оң	2 / 4	$1 - (2/6)^2 - (4/6)^2 = 0{,}444$	6/10
B бөлу: сол	3 / 2	0,48	5/10
B бөлу: оң	3 / 2	0,48	5/10

Регрессиялық ағашта сандық болжам

Жапырақтағы болжам

Оқу мысалы: бір жапыраққа жеткізу уақыты 2, 3 және 4 күн болатын үш тапсырыс түсті. Орташа болжам: $\hat{y} = (2 + 3 + 4) / 3 = 3$ күн.

Бөліну логикасы

Алгоритм мақсатты мәндері жақын объектілерді бір топқа жинауға тырысады. Сапа балалардағы қателікпен бағаланады. Кәдімгі ағашта болжам әр аймақта тұрақты.

3. Күрделілікті реттеу және сапаны бағалау

Негізгі мәселе

Өте терең ағаш нақты заңдылықпен бірге кездейсоқ ерекшеліктерді жаттап алуы мүмкін. Оқыту дәлдігі жоғары болса да, жаңа деректе қате өседі.

Ерте шектеу

max_depth, min_samples_split және min_samples_leaf параметрлері түйіндердің құрылуын реттейді.

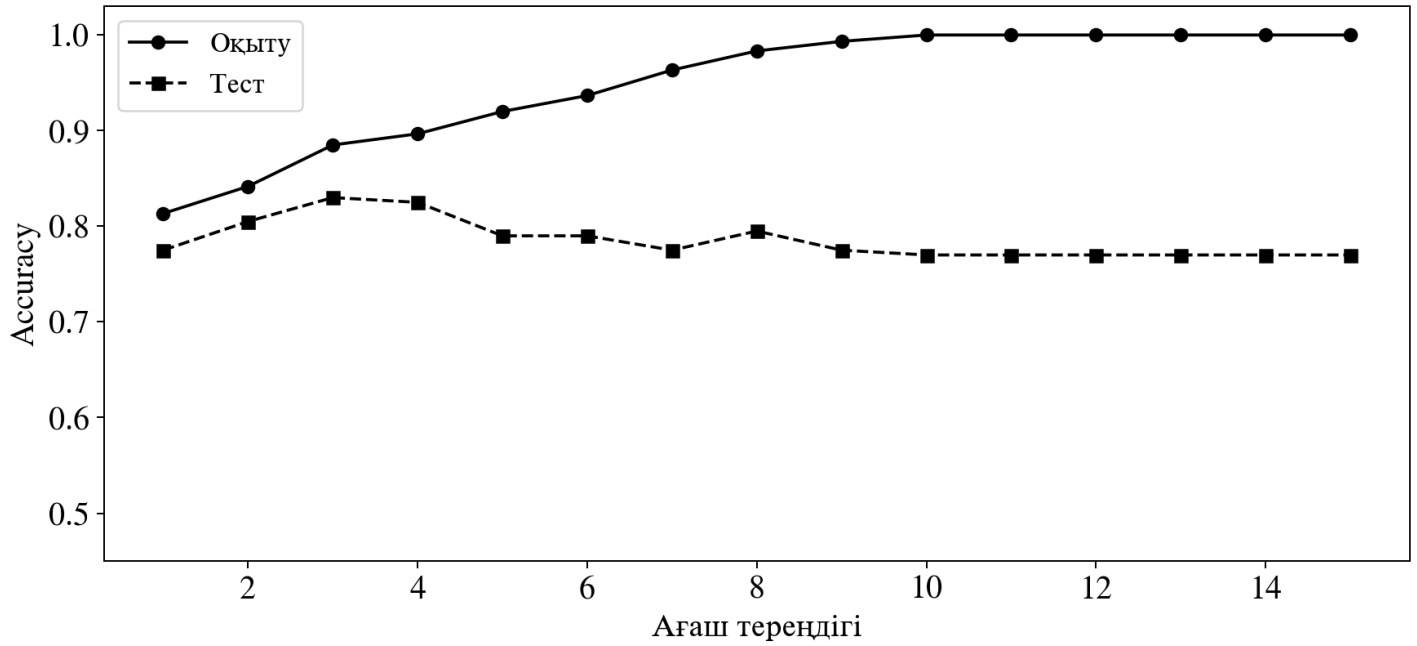
Кейін қысқарту

Pruning дайын ағаштың әлсіз тармақтарын алып тастайды. Қысқарту күші валидацияда таңдалады.

Тәуелсіз бағалау

Оқыту модельді құру үшін, валидация параметр таңдау үшін, тест соңғы сапаны бағалау үшін қолданылады.

Артық үйрену: тереңдік пен дәлдік байланысы



Түсіндірме

800 синтетикалық объект. Оқыту: 600, тест: 200. Тереңдік өскенде оқу дәлдігі артады. Тест дәлдігі бірдей бағытта тұрақты өспейді. Бұл график түсіндіру үшін.

Ағаштың негізгі гиперпараметрлері

Параметр	Қызметі	Практикалық әсері
max_depth	Максималды тереңдік	Кішірек мән қарапайым ағаш береді
min_samples_split	Бөлуге қажет ең аз объект	Өте шағын түйіндерді бөлуді тежейді
min_samples_leaf	Жапырақтағы ең аз объект	Сирек ережелерді азайтады
max_leaf_nodes	Жапырақ санының шегі	Ережелер санын бақылайды
ccp_alpha	Қысқарту күші	Үлкен мән көбірек қысқартады
class_weight	Кластардың салмағы	Теңгерімсіз кластарды ескереді

Pruning және модельді дұрыс тексеру

Cost-complexity pruning

Мақсат: қате мен ағаш көлемінің теңгерімі. Rα(T) = R(T) + α|T|. |T|: жапырақ саны. α: күрделілік айыбы. Әлсіз тармақтарды алып тастау ықшам ағаш береді.

Тексеру тәртібі

1. Оқу және тест жиындарын алдын ала бөлу. 2. Оқу жиыны ішінде валидация немесе cross-validation жүргізу. 3. Параметрлерді тек осы нәтижемен таңдау. 4.

Confusion matrix және классификация метрикалары

Көрсеткіш	Формула	Мағынасы
TP / TN	Дұрыс оң / дұрыс теріс	Қай шешім дұрыс болды?
FP / FN	Жалған оң / жалған теріс	Қате қай бағытта жасалды?
Accuracy	(TP + TN) / N	Барлық дұрыс болжам үлесі
Precision	TP / (TP + FP)	Оң болжамның дұрыстығы
Recall	TP / (TP + FN)	Оң класты табу үлесі
F1	2PR / (P + R)	Екі метриканың теңгерімі

4. Практикалық қолдану және әдістерді салыстыру

Деректерді дайындау

Мақсатты белгіні анықтау, қайталануды тексеру және категорияларды кодтау. scikit-learn ағаштарына сандық енгізу дайындау қажет.

Белгілерді тексеру

Болжам кезінде қолжетімсіз ақпаратты енгізуге болмайды. Мысалы, тапсырысты алғаннан кейінгі шағымды алдын ала тәуекел болжамына қосу дерек ағуы болады.

Интерпретация

Тармақтарды ережеге айналдыру және шешім шектерін түсіндіру. Белгі маңыздылығы себеп-салдарлық әсерге тең емес.

Әдісті дамыту

Бір ағашты Random Forest немесе boosting әдістерімен салыстыру. Күрделілік, сапа және түсіндіру мүмкіндігі бірге бағаланады.

Синтетикалық тәжірибенің нақты есептелген нәтижесі

Көрсеткіш	Мән	Түсіндірме
Дерек көлемі	800 объект, 6 белгі	Жасанды екі класты дерек
Бөлу	600 оқу / 200 тест	Стратификация, random_state = 12
Модель	max_depth = 4; min_samples_leaf = 10	Алдын ала белгіленген параметрлер
Accuracy	0.800	Тесттегі дұрыс болжам үлесі
Confusion matrix	TN=78, FP=22, FN=18, TP=82	Оң класс = 1
Тәжірибе мәртебесі	Оқу демонстрациясы	Нақты банкке немесе өндіріске қорытынды емес

Бір ағаш және ансамбльдік әдістер

Әдіс	Құрылымы	Артықшылығы	Негізгі шектеуі
Decision Tree	Бір ережелер ағашы	Ықшам болса түсіндіру оңай	Дерек өзгерісіне сезімтал
Random Forest	Көп ағаш, кездейсоқ үлгілер мен белгілер	Бір ағаштың тұрақсыздығын азайтады	Толық модельді түсіндіру күрделі
Gradient Boosting	Кезекпен қосылатын ағаштар	Алдыңғы модель қателерін түзетеді	Параметрлерді мұқият таңдау қажет
Таңдау критерийі	Бірдей бөлу мен метрика	Әділ салыстыру	Сапа жалғыз талап емес

Артықшылықтар, шектеулер және қолдану шарттары

Артықшылықтар

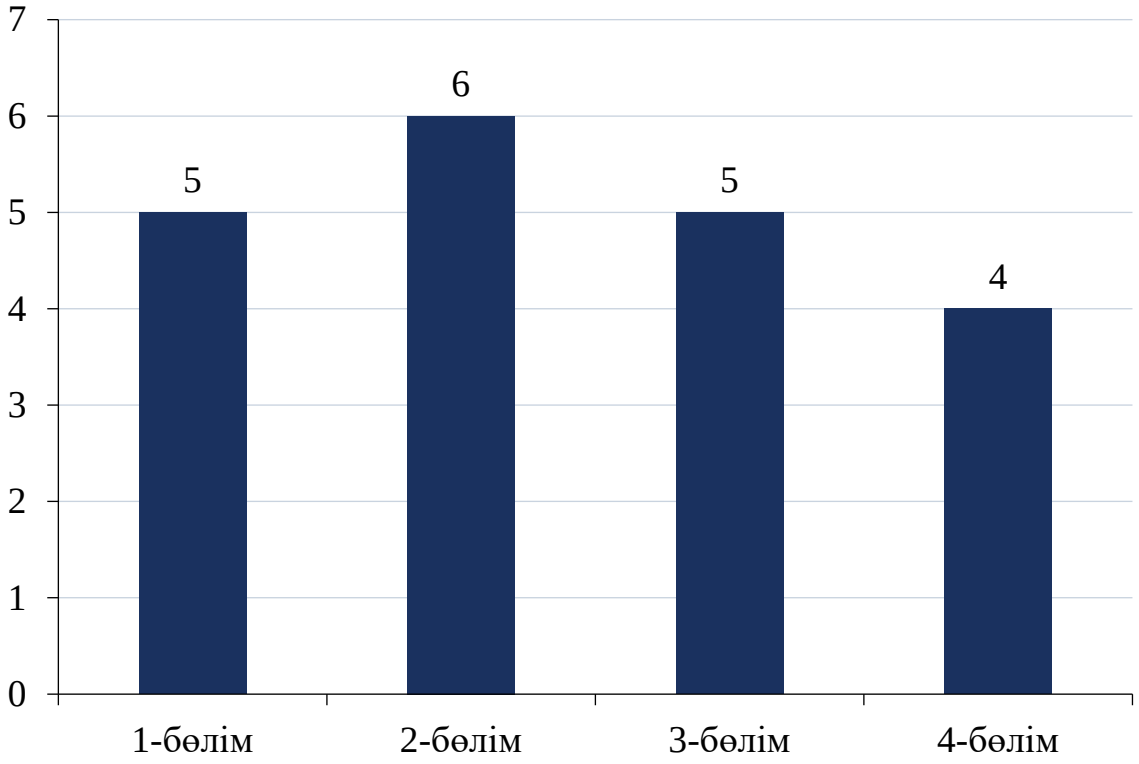
Шарттар арқылы түсіндірілетін шешім. Сызықтық емес байланыстар мен белгілердің өзара әсерін көрсете алады. Сандық белгілерді стандарттау әдетте міндетті емес.

Шектеулер

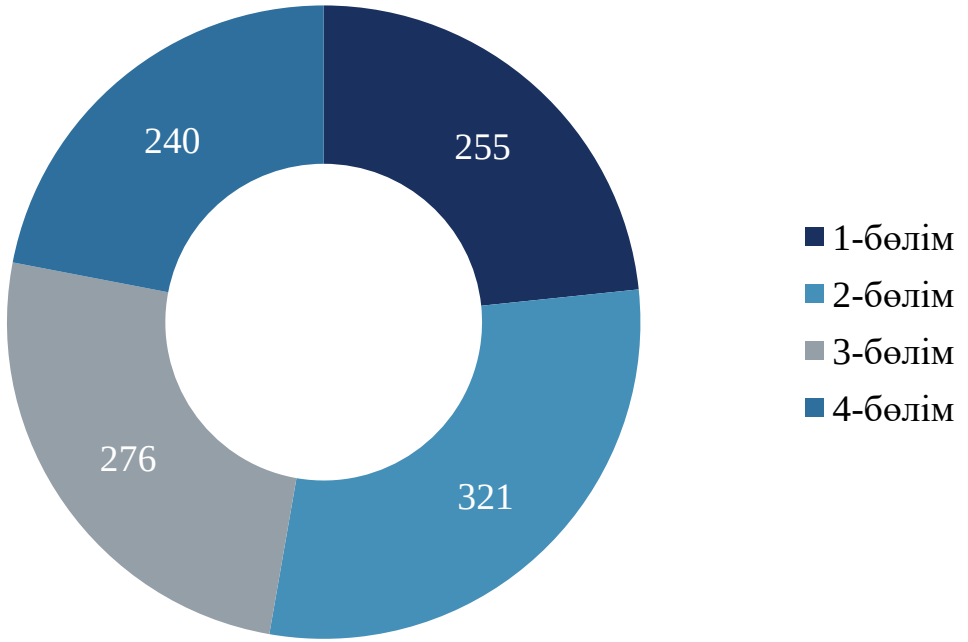
Өте терең ағаш артық үйренеді. Деректегі шағын өзгеріс басқа құрылым беруі мүмкін.

Диаграммалар

Бөлімдер бойынша слайд саны



Бөлімдер бойынша мәтін көлемі (сөз саны)



Қорытынды

1. Шешім ағашы белгілерді шарттармен бөліп, класс немесе сандық болжам береді.
2. Gini мен энтропия түйіндердің біртектілігін бағалайды.
3. Тереңдік пен pruning ағаштың күрделілігін және артық үйренуін реттейді.
4. Модельдің сапасын тәуелсіз деректе және қателердің маңызын ескере отырып бағалау қажет.

Дереккөздер: №8 лекция; scikit-learn (Decision Trees, DecisionTreeClassifier, Metrics, Pruning, Ensembles).

Шешім ағашы

Дәріс №8