12-дәріс.

Интеллектуалды деректер талдауы (Data Mining)

Пәні: Интеллектуалды ақпараттық жүйелер Дәрістің тақырыбы: Интеллектуалды деректер талдауы (Data Mining)

Бұл дәріс алдыңғы дәрістердегі машиналық оқыту, нейрондық желілер және интеллектуалды жүйелер туралы білімді жалғастырады. Материал деректерді жинаудан бастап, оларды өңдеу, жасырын заңдылықтарды анықтау, нәтижелерді бағалау және интеллектуалды жүйеде қолдануға дейінгі толық үдерісті түсіндіреді.

Дәрістің мақсаты

Студенттерге интеллектуалды деректер талдауының теориялық негіздерін, Data Mining технологиясының пайда болу себептерін, деректерден білім алу кезеңдерін, негізгі алгоритмдерін және оларды интеллектуалды ақпараттық жүйелерде қолдану ерекшеліктерін түсіндіру.

Күтілетін оқу нәтижелері

Дәріс соңында студенттер:

1. Data Mining ұғымын және оның интеллектуалды ақпараттық жүйелердегі рөлін түсіндіреді.
2. Data Mining, Machine Learning, статистика және KDD ұғымдарының айырмашылықтарын анықтайды.
3. Деректерді дайындау және тазалау әдістерін сипаттайды.
4. Классификация, кластеризация, ассоциациялық ережелер және аномалияларды анықтау міндеттерін ажыратады.
5. K-means алгоритмі бойынша қарапайым есептеулер жүргізеді.
6. Support, Confidence және Lift көрсеткіштерін есептейді.
7. Деректерді талдау нәтижелерін графиктер мен кестелер арқылы түсіндіреді.
8. Data Mining әдістерін нақты интеллектуалды жүйелерде қолданудың артықшылықтары мен шектеулерін бағалайды.

1. Интеллектуалды деректер талдауына кіріспе

Қазіргі ақпараттық жүйелер күн сайын үлкен көлемде деректер жинайды.

Мысалы, университеттің ақпараттық жүйесінде студенттердің бағалары, сабаққа қатысуы, пәндерге тіркелуі, тапсырмаларды орындау уақыты және электрондық оқу платформасындағы белсенділігі сақталады.

Бұл мәліметтер ақпараттық жүйенің күнделікті жұмысын қамтамасыз етеді. Бірақ деректердің жиналуы олардың автоматты түрде пайдалы білімге айналғанын білдірмейді.

Мысалы, университеттің деректер базасында 10 000 студент туралы ақпарат бар делік.

Біз келесі сұрақтарға жауап іздеуіміз мүмкін:

- Қандай факторлар студенттің академиялық үлгеріміне байланысты?
- Қандай студенттерге қосымша оқу қолдауы қажет болуы мүмкін?
- Студенттердің қандай топтары ұқсас оқу белсенділігін көрсетеді?
- Қандай пәндер жиі бірге таңдалады?
- Қандай көрсеткіштер әдеттегі жағдайдан ерекше ауытқиды?

Бұл сұрақтарға жауап беру үшін деректерді жай сақтау жеткіліксіз.

Оларды өңдеу, салыстыру, заңдылықтарды анықтау және алынған нәтижелерді түсіндіру қажет.

Осы міндеттерді шешуге **Data Mining** — интеллектуалды деректер талдауы технологиясы көмектеседі.

2. Data Mining дегеніміз не?

Data Mining — үлкен немесе күрделі деректер жиындарынан бұрын анық байқалмаған, пайдалы және түсіндіруге болатын заңдылықтарды, байланыстарды, құрылымдарды немесе ерекшеліктерді анықтауға бағытталған әдістер жиынтығы.

Ағылшын тіліндегі Mining сөзі «**қазып алу**» деген мағынаны білдіреді.

Бұл атау кездейсоқ таңдалмаған.

Пайдалы қазбаны жер қойнауынан іздегендей, Data Mining технологиясы да көптеген деректер арасынан пайдалы ақпаратты анықтайды.

Бірақ Data Mining тек ақпарат іздеу емес. Ол статистикалық, алгоритмдік және машиналық оқыту әдістері арқылы деректердегі заңдылықтарды анықтауды қамтиды.

Қарапайым мысал

Дүкеннің ақпараттық жүйесінде сатып алушылардың чектері сақталған делік.

Деректерді қарапайым сұрыптау арқылы бүгін қанша нан сатылғанын анықтаймыз.

Ал Data Mining көмегімен нан сатып алған клиенттердің қандай бөлігі сүтті де бірге сатып алатынын зерттей аламыз.

Мұндай заңдылық тауарларды орналастыру немесе ұсыныстар қалыптастыру кезінде пайдалы болуы мүмкін.

Негізгі айырмашылық: қарапайым сұраныс белгілі фактіні табады, ал Data Mining деректердегі заңдылықтарды анықтауға бағытталады.

3. Data Mining технологиясының пайда болуы

Data Mining бағытының дамуы бірнеше ғылыми және технологиялық өзгерістермен байланысты.

Біріншіден, деректер базалары кең таралып, ұйымдар үлкен көлемде ақпарат жинай бастады.

Екіншіден, компьютерлердің есептеу мүмкіндіктері артты.

Үшіншіден, статистика, машиналық оқыту және деректер базалары салаларындағы әдістер бір-бірімен тығыз байланыса бастады.

1990-жылдары деректерден білім алу (Knowledge Discovery in Databases, KDD) бағыты белсенді дамыды.

Data Mining осы үдерістің маңызды кезеңдерінің бірі ретінде қарастырылды.

Қазіргі уақытта Data Mining қаржы, медицина, білім беру, сауда, өндіріс, ауыл шаруашылығы және киберқауіпсіздік салаларында қолданылады.

4. Data Mining, Machine Learning және статистика

Бұл ұғымдар өзара байланысты болғанымен, олардың негізгі міндеттері бірдей емес.

Ұғым	**Негізгі мақсаты**	**Қарапайым мысал**
Статистика	Деректерді сипаттау, бағалау және гипотезаларды тексеру	Студенттердің орташа GPA көрсеткішін анықтау
Machine Learning	Деректер негізінде болжау немесе шешім қабылдау моделін үйрету	Студенттің келесі семестрдегі үлгерімін болжау
Data Mining	Деректерден пайдалы заңдылықтар мен құрылымдарды анықтау	Үлгерімі ұқсас студенттер топтарын табу
KDD	Деректерден пайдалы білім алудың толық үдерісін ұйымдастыру	Деректерді таңдау, тазалау, талдау және нәтижені түсіндіру

Data Mining машиналық оқыту алгоритмдерін қолдануы мүмкін.

Мысалы, K-means алгоритмі — машиналық оқыту әдісі. Оны студенттердің оқу белсенділігі бойынша топтарын анықтау үшін қолдансақ, ол Data Mining міндетін орындайды.

Сондықтан бұл салалар бір-біріне қарсы қойылмайды. Олар бір жобаның әртүрлі бөліктерінде бірге қолданылуы мүмкін.

5. Деректерден білім алу үдерісі — KDD

Knowledge Discovery in Databases (KDD) — бастапқы деректерден пайдалы, түсіндіруге болатын білім алуға бағытталған көпкезеңді үдеріс.

Data Mining — осы үдерістің алгоритмдік талдау кезеңі.

1. Деректерді таңдау (Selection)
2. Деректерді тазалау (Preprocessing)
3. Түрлендіру (Transformation)
4. Data Mining — заңдылықтарды анықтау
5. Нәтижелерді бағалау және түсіндіру

Бірінші кезең. Деректерді таңдау

Алдымен қандай мәселені зерттейтінімізді анықтаймыз.

Мысалы, студенттердің академиялық белсенділігін талдау үшін GPA, сабаққа қатысу және тапсырмаларды орындау мәліметтері таңдалады.

Екінші кезең. Деректерді тазалау

Қате, қайталанатын және жетіспейтін мәндер анықталады.

Үшінші кезең. Деректерді түрлендіру

Әртүрлі масштабтағы белгілер бір форматқа келтіріледі. Қажет болса, категориялық мәндер сандық түрде кодталады.

Төртінші кезең. Data Mining

Таңдалған алгоритм деректердегі заңдылықтарды анықтайды.

Мысалы, студенттерді ұқсас оқу белсенділігі бойынша топтарға бөледі.

Бесінші кезең. Нәтижелерді түсіндіру

Алынған нәтижелердің қаншалықты пайдалы, сенімді және қолдануға жарамды екені бағаланады.

Бұл кезең өте маңызды, себебі алгоритм анықтаған кез келген заңдылық міндетті түрде пайдалы білім бола бермейді.

6. Деректер жиынының құрылымы

Data Mining алгоритмдері деректер жиынымен (Dataset) жұмыс істейді.

Кестелік деректерде әр жол — бір объект, ал әр баған — белгілі бір белгі.

Мысалы:

Студент	GPA	Қатысу (%)	Кеш тапсырмалар	Тәуекел белгісі
A	3.8	95	0	Төмен
B	2.4	60	5	Жоғары
C	3.2	85	1	Төмен
D	2.1	55	6	Жоғары
E	3.5	90	0	Төмен
F	2.8	72	3	Орташа

Бұл — оқыту мақсатында құрастырылған шартты деректер.

Мұнда GPA, қатысу және кеш тапсырмалар саны — белгілер (Features).

Ал тәуекел белгісі — егер классификация моделі құрылса, мақсатты айнымалы (Target).

Деректердің қандай белгілері таңдалатыны зерттеу мақсатына байланысты.

7. Деректер сапасы және алдын ала өңдеу

Data Mining нәтижесінің сапасы бастапқы деректердің сапасына тікелей байланысты.

Егер деректер қате немесе толық емес болса, алгоритм де қате заңдылықтарды анықтауы мүмкін.

Деректер сапасын бағалауда келесі қасиеттер маңызды:

- толықтық (Completeness);
- дұрыстық (Accuracy);
- бірізділік (Consistency);
- өзектілік (Timeliness);
- бірегейлік (Uniqueness).

Мысалы, студенттің сабаққа қатысуы 150% деп көрсетілсе, бұл қате мән.

Егер бір студенттің жазбасы бірнеше рет қайталанса, кейбір статистикалық нәтижелер бұрмалануы мүмкін.

Сондықтан алгоритмді іске қоспас бұрын деректерді тексеру қажет.

8. Жетіспейтін мәндерді өңдеу

Деректер жиынында кейбір ұяшықтар бос болуы мүмкін.

Мысалы:

Студент	GPA	Қатысу (%)
A	3.2	90
B	2.8	—
C	3.6	80
D	3.0	85

B студентінің қатысу көрсеткіші жоқ.

Жетіспейтін мәндерді өңдеудің бірнеше тәсілі бар:

1. Жазбаны алып тастау.
2. Орташа мәнмен алмастыру.
3. Медианамен алмастыру.
4. Басқа белгілер негізінде бағалау.
5. Жетіспеудің өзін жеке белгі ретінде пайдалану.

Мысалы, белгілі қатысу мәндерінің орташа шамасы:

$$\bar{x} = \frac{90 + 80 + 85}{3} = 85$$

Сонда бос мәнді 85 санымен алмастыруға болады.

Бірақ бұл әдіс әрдайым дұрыс емес.

Егер деректің болмауы кездейсоқ емес, белгілі бір жағдайға байланысты болса, орташа мәнмен алмастыру нәтижені бұрмалауы мүмкін.

9. Нормализация және стандарттау

Деректердегі белгілер әртүрлі өлшемдермен берілуі мүмкін.

Мысалы, GPA 0–4 аралығында, қатысу 0–100 аралығында, ал кеш тапсырмалар саны 0–20 аралығында болуы мүмкін.

Мұндай айырмашылық қашықтыққа негізделген алгоритмдердің жұмысына әсер етеді.

Min-Max нормализациясы

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Мысалы, қатысу көрсеткіші 80%, ал минималды және максималды мәндер 50% және 100% болсын.

$$x' = \frac{80 - 50}{100 - 50} = \frac{30}{50} = 0.6$$

Нәтиже 0 мен 1 аралығына келтірілді.

Стандарттау

$$z = \frac{x - \mu}{\sigma}$$

Мұндағы μ — орташа мән, σ — стандартты ауытқу.

Стандарттау белгі мәнінің орташа шамадан қаншалықты ауытқығанын көрсетеді.

Маңызды ереже: болжау моделін бағалау кезінде нормализация немесе стандарттау параметрлері тек оқыту деректерінен есептелуі тиіс. Тест деректерін алдын ала пайдалану деректердің ағып кетуіне **(Data Leakage)** әкелуі мүмкін.

10. Data Mining негізгі міндеттері

Data Mining бірнеше негізгі міндетті шешеді.

Міндет	**Мақсаты**	**Мысал**
Классификация	Объектіні белгілі класқа жатқызу	Академиялық тәуекел деңгейін анықтау
Регрессия	Сандық мәнді болжау	Келесі емтихан балын болжау
Кластеризация	Ұқсас объектілерді топтау	Студенттерді белсенділік бойынша бөлу
Ассоциациялық ережелер	Бірге кездесетін белгілерді табу	Бірге таңдалатын пәндерді анықтау
Аномалияларды анықтау	Ерекше бақылауларды табу	Әдеттен тыс жүйелік әрекетті анықтау

Бұл міндеттер бір жобада қатар қолданылуы мүмкін.

Мысалы, университет жүйесі алдымен студенттерді кластерлерге бөліп, кейін белгілі топтағы академиялық тәуекелді болжауы мүмкін.

11. Классификация

Классификация (Classification) — объектінің белгілеріне сүйене отырып, оны алдын ала анықталған кластардың біріне жатқызу міндеті.

Мысалы, студенттің академиялық тәуекелі:

- төмен;
- орташа;
- жоғары

деп жіктелуі мүмкін.

Классификация моделін үйрету үшін әдетте белгілері мен дұрыс жауаптары белгілі деректер пайдаланылады.

Мысалы:

$$X = (GPA,\ Attendance,\ LateSubmissions)$$

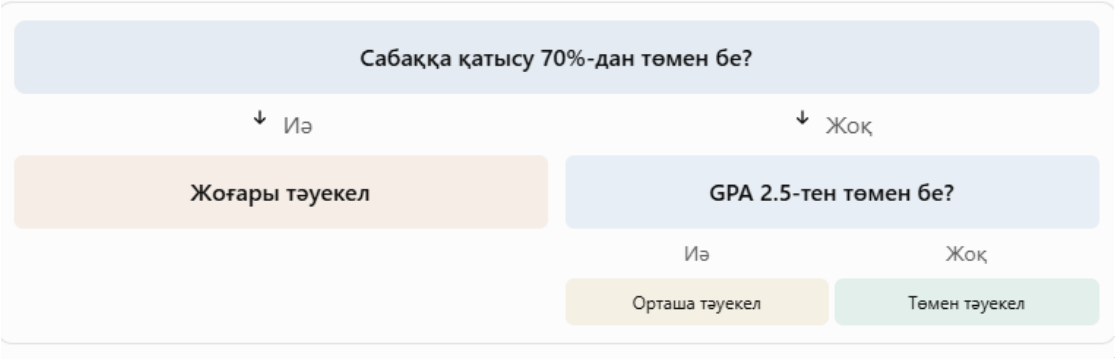
$$Y = \{\text{Төмен},\ \text{Орташа},\ \text{Жоғары}\}$$

Классификация үшін шешім ағаштары, логистикалық регрессия, кездейсоқ орман және нейрондық желілер қолданылуы мүмкін.

12. Шешім ағашы — Decision Tree

Шешім ағашы — белгілерге қатысты шарттарды ретімен тексеріп, соңғы шешімге келетін модель.

Мысалы, оқу мақсатындағы қарапайым ережелерді қарастырайық.



Мысалы, студенттің қатысуы 62% болса, осы шартты ағаш оны жоғары тәуекел класына жатқызады. Бірақ бұл шектер нақты университеттің ресми ережелері емес. Олар алгоритмді түсіндіру үшін ғана берілген.

Нақты Data Mining жобасында шешім ағашының бөліну шарттары оқыту деректерінен анықталады.

13. Кластеризация

Кластеризация (Clustering) — объектілерді олардың ұқсастығы бойынша топтарға бөлу әдісі.

Классификациядан негізгі айырмашылығы — кластеризацияда объектілердің дұрыс класс белгілері алдын ала берілмеуі мүмкін.

Мысалы, университетте студенттерді электрондық платформадағы белсенділігі бойынша топтастыруға болады.

Бірінші топта жүйеге жиі кіретін студенттер, екінші топта орташа белсенді студенттер, үшінші топта сирек кіретін студенттер шоғырлануы мүмкін.

Алгоритм топтарды белгілердегі ұқсастық бойынша анықтайды.

Топтардың педагогикалық мағынасы кейін талданады.

14. K-means алгоритмі

K-means — объектілерді алдын ала берілген *K* кластерге бөлуге арналған кең таралған алгоритм.

Алгоритмнің негізгі кезеңдері:

1. Кластерлер саны *K* таңдалады.
2. Бастапқы кластер орталықтары анықталады.
3. Әр объект ең жақын орталыққа бекітіледі.
4. Әр кластердің жаңа орталығы есептеледі.
5. Объектілерді бөлу мен орталықтарды жаңарту қадамдары тоқтау шарты орындалғанша қайталанады.

Қарапайым мысал

Алты студенттің аптасына электрондық оқу жүйесіне кіру саны:

Студент	Кіру саны
A	2
B	3
C	4
D	10
E	11
F	12

Екі кластер құрайық:

K =2

Алғашқы орталықтар 3 және 11 деп алынсын.

Бірінші кластерге 2, 3, 4 мәндері, екінші кластерге 10, 11, 12 мәндері жақын орналасады.

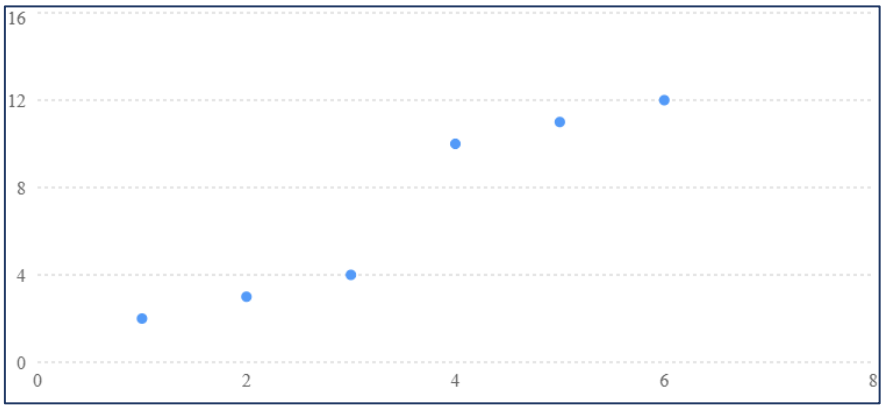
Жаңа орталықтар:

$$C_1 = \frac{2+3+4}{3} = 3$$

$$C_2 = \frac{10+11+12}{3} = 11$$

Орталықтар өзгермегендіктен, алгоритм осы мысалда тұрақты топтарға келді.

Студенттердің оқу платформасына кіру белсенділігі



Бұл деректерден екі айқын топ көрінеді.

Бірақ кластерлердің «жақсы» немесе «жаман» студенттерді білдіретіні туралы қорытынды жасауға болмайды. Кіру саны — оқу нәтижесінің тек бір жанама көрсеткіші.

15. Евклид қашықтығы

K-means алгоритмінде объектілер арасындағы жақындықты анықтау үшін Евклид қашықтығы жиі қолданылады.

Екі өлшемді кеңістікте:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Мысалы:

$$A = (2,3), \qquad B = (5,7)$$

Онда:

$$d(A,B) = \sqrt{(5-2)^2 + (7-3)^2}$$

$$d(A,B) = \sqrt{9+16} = \sqrt{25} = 5$$

Яғни екі нүктенің арасындағы қашықтық 5-ке тең.
Көпөлшемді деректерде формула:

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

түрінде жазылады.
Бұл формула GPA, қатысу және басқа белгілер бойынша объектілерді салыстыруға қолданылуы мүмкін.
Бірақ белгілердің масштабтары әртүрлі болса, қашықтық нәтижесі бұрмалануы ықтимал. Сондықтан K-means алдында деректерді стандарттау немесе нормализациялау жиі қажет.

16. Ассоциациялық ережелер

Ассоциациялық ережелер (Association Rules) — деректер жиынында қандай объектілердің, оқиғалардың немесе белгілердің бірге кездесетінін анықтайтын әдіс.
Ең танымал мысал — сатып алушылардың себетін талдау (Market Basket Analysis).
Мысалы, дүкен деректерінде нан сатып алатын адамдардың сүтті де жиі сатып алатыны байқалуы мүмкін.
Ассоциациялық ереже:

$$A \Rightarrow B$$

түрінде жазылады.
Мұндағы *A* — шарт бөлігі, *B* — нәтиже бөлігі.
Мысалы:

$$\{\text{Нан}\} \Rightarrow \{\text{Сүт}\}$$

Бұл ереже нан сатып алу сүт сатып алуға себеп болады дегенді білдірмейді.
Ол тек екі тауардың деректерде бірге кездесуі туралы ақпарат береді.

17. Support, Confidence және Lift

Ассоциациялық ережелердің маңыздылығын бағалау үшін үш негізгі көрсеткіш қолданылады.

Support — қолдау көрсеткіші

$$Support(A \Rightarrow B) = \frac{N(A \cap B)}{N}$$

Мұндағы *N(A∩B)* — A және B бірге кездескен транзакциялар саны, *N* — барлық транзакциялар саны.

Confidence — сенімділік көрсеткіші

$$Confidence(A \Rightarrow B) = \frac{N(A \cap B)}{N(A)}$$

Бұл көрсеткіш A кездескен жағдайлардың қандай үлесінде B де кездесетінін анықтайды.

Lift — байланыс күші

$$Lift(A \Rightarrow B) = \frac{Confidence(A \Rightarrow B)}{Support(B)}$$

Lift мәнін түсіндіру:

Lift	**Түсіндірме**
1-ден жоғары	A және B бірге күтілгеннен жиірек кездеседі
1-ге тең	Бақыланған бірлескен жиілік тәуелсіздік жағдайына сәйкес келеді
1-ден төмен	A және B бірге күтілгеннен сирегірек кездеседі

Бұл көрсеткіштер байланыстарды талдауға көмектеседі, бірақ өздігінен себеп-салдарлық қатынасты дәлелдемейді.

18. Ассоциациялық ережені есептеу мысалы

Дүкендегі 10 транзакцияны қарастырайық.

Транзакция	**Сатып алынған тауарлар**
T1	Нан, сүт
T2	Нан, сүт, май
T3	Нан, сүт
T4	Нан, сүт, жұмыртқа
T5	Нан, сүт
T6	Нан, май
T7	Нан
T8	Сүт
T9	Сүт, жұмыртқа
T10	Жұмыртқа

Нан 7 транзакцияда, сүт 7 транзакцияда, ал нан мен сүт бірге 5 транзакцияда кездеседі.

Ереже:

$$\{\text{Нан}\} \Rightarrow \{\text{Сүт}\}$$

Support:

$$Support = \frac{5}{10} = 0.5 = 50\%$$

Confidence:

$$Confidence = \frac{5}{7} \approx 0.714 = 71.4\%$$

Lift:

$$Lift = \frac{0.714}{0.7} \approx 1.02$$

Нәтижені түсіндірейік.

Транзакциялардың 50%-ында нан мен сүт бірге сатып алынған.

Нан сатып алынған транзакциялардың шамамен 71.4%-ында сүт те бар.

Бірақ Lift шамамен 1.02 ғана. Демек, осы шағын деректерде нан мен сүттің бірге кездесуі тәуелсіздік жағдайындағы күтілетін жиіліктен сәл ғана жоғары.

Бұл мысал Confidence жоғары болғанымен, байланыс міндетті түрде күшті болмайтынын көрсетеді.

19. Аномалияларды анықтау

Аномалия (Anomaly) — деректер жиынындағы әдеттегі үлгілерден елеулі ерекшеленетін бақылау.

Аномалияларды анықтау қаржылық алаяқтықты, желідегі күмәнді әрекеттерді және өндірістік жабдықтардың қалыптан тыс көрсеткіштерін зерттеуде қолданылады.

Мысалы, сервердегі минутына келетін сұраныстар саны:

45, 48, 50, 47, 52, 49, 310

болсын.

310 мәні басқа көрсеткіштерден елеулі ерекшеленеді.

Бірақ бұл мән міндетті түрде шабуыл болғанын білдірмейді. Ол заңды пайдаланушылардың белсенділігімен немесе жоспарланған жүктememен байланысты болуы мүмкін.

Z-score тәсілі

$$z = \frac{x - \mu}{\sigma}$$

Z-score мәнінің абсолют шамасы жоғары болса, бақылау орташа мәннен едәуір алыста орналасқанын көрсетеді.

Алайда аномалияны анықтау үшін бір ғана шек барлық деректерге жарамайды. Деректердің таралуы мен зерттеу мақсаты ескерілуі тиіс.

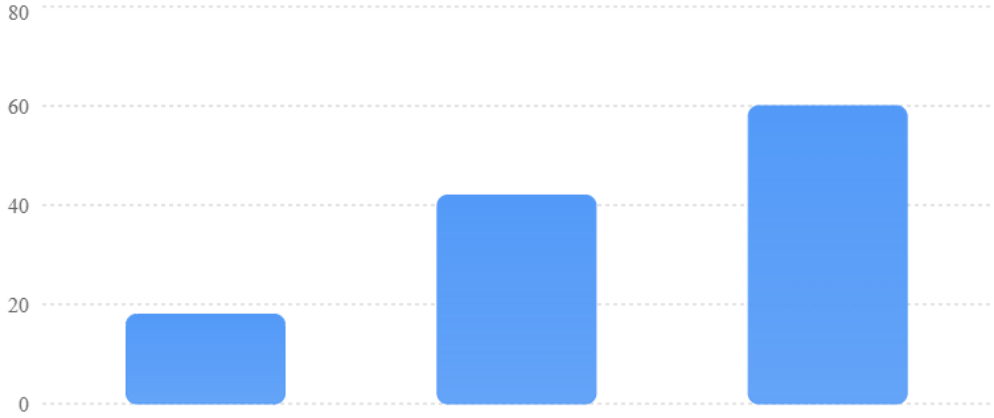
20. Деректерді визуализациялау

Data Mining нәтижелерін түсіндіру үшін графиктер мен диаграммалар қолданылады.

Визуализация күрделі нәтижелерді түсінікті түрде көрсетуге көмектеседі.

Мысалы, студенттердің сабаққа қатысу деңгейлерін қарастырайық.

Студенттердің сабаққа қатысу деңгейлері (шартты деректер)



Графикте 120 студенттің 60-ы жоғары, 42-сі орташа және 18-і төмен қатысу деңгейіне жататыны көрсетілген.

Мұндай диаграмма деректердің таралуын түсінуге мүмкіндік береді.

Бірақ диаграмманың өзі төмен қатысудың себебін немесе оның оқу нәтижесіне әсерін дәлелдемейді.

21. Data Mining нәтижелерін бағалау

Алгоритм нәтижесін бағалау әдісі шешілетін міндетке байланысты.

Классификацияда Accuracy, Precision, Recall және F1-score қолданылуы мүмкін.

Кластеризацияда кластерлердің ішкі тығыздығы мен өзара бөлінуі зерттеледі.

Ассоциациялық ережелерде Support, Confidence және Lift пайдаланылады.

Міндет	**Негізгі көрсеткіштер**
Классификация	Accuracy, Precision, Recall, F1
Регрессия	MAE, MSE, RMSE
Кластеризация	Silhouette Score, Inertia
Ассоциациялық ережелер	Support, Confidence, Lift
Аномалияларды анықтау	Precision, Recall және сарапшылық тексеру

Мысалы, кластеризация нәтижесін бағалауда Silhouette Score қолдануға болады:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Мұндағы *a(i)* — объектінің өз кластеріндегі басқа объектілерге орташа қашықтығы, *b(i)*— ең жақын басқа кластерге орташа қашықтығы.

Бұл көрсеткіш кластерлердің қаншалықты анық бөлінгенін бағалауға көмектеседі.

Алайда жоғары метрика практикалық маңыздылықтың толық кепілі емес.

22. Data Mining технологиясының қолданылуы

Data Mining әртүрлі салаларда пайдаланылады.

Білім беру: студенттердің оқу белсенділігін талдау, академиялық қолдауды жоспарлау, білім беру ресурстарын пайдалануды зерттеу.

Медицина: клиникалық деректердегі заңдылықтарды анықтау, аурулар арасындағы байланыстарды зерттеу, диагностикалық модельдерді қолдау.

Қаржы: күмәнді транзакцияларды анықтау, клиенттердің мінез-құлқын талдау, тәуекелдерді бағалау.

Сауда: тауарларды бірге сатып алу заңдылықтарын зерттеу, клиенттерді сегменттеу, ұсыныстар жүйесін дамыту.

Өндіріс: жабдықтардың сенсор деректерін талдау, ақауларды ерте анықтау, өндірістік процестерді оңтайландыру.

Киберқауіпсіздік: желілік трафиктегі әдеттен тыс әрекеттерді анықтау.

Әр салада деректердің мазмұны мен шешімнің салдары әртүрлі болғандықтан, алгоритмді таңдау және нәтижені тексеру талаптары да өзгереді.

23. Университеттің интеллектуалды жүйесіндегі кешенді мысал

Университет студенттерге академиялық қолдау көрсету процесін жетілдіргісі келеді делік.

Жүйе келесі деректерді жинайды:

- GPA;
- сабаққа қатысу;
- тапсырмаларды тапсыру мерзімдері;
- оқу платформасына кіру белсенділігі;
- аралық бақылау нәтижелері.

Алдымен деректердің толықтығы мен дұрыстығы тексеріледі.

Содан кейін кластеризация көмегімен ұқсас белсенділік көрсететін студенттер топтары анықталады.

Классификация моделі бұрынғы деректер негізінде академиялық тәуекелге қатысты болжам жасауы мүмкін.

Ассоциациялық талдау белгілі оқу әрекеттерінің қандай нәтижелермен бірге жиі кездесетінін анықтауға көмектеседі.

Аномалияларды анықтау әдістері ерекше өзгерістерді, мысалы, белсенділіктің кенет төмендеуін көрсетуі мүмкін.

Соңында нәтижелер эдвайзерге немесе жауапты оқытушыға түсінікті түрде беріледі.

Маңыздысы — жүйе студентке қатысты соңғы шешімді автоматты түрде қабылдамауы керек. Модель қорытындысы адам тарапынан тексерілетін қосымша ақпарат ретінде қолданылуы тиіс.

24. Data Mining артықшылықтары мен шектеулері

Артықшылықтары

Data Mining үлкен көлемдегі деректерден пайдалы заңдылықтарды анықтауға мүмкіндік береді.

Ол ақпараттық жүйелерде шешім қабылдауды қолдауға, процестерді жетілдіруге және тәуекелдерді ертерек байқауға көмектесуі мүмкін.

Сонымен қатар бұрын назар аударылмаған байланыстарды анықтауға жағдай жасайды.

Шектеулері

Data Mining нәтижесі деректердің сапасына тәуелді.

Алгоритм кездейсоқ немесе жалған байланыстарды да анықтауы мүмкін.

Кейбір модельдердің шешімін түсіндіру қиын.

Деректерде тарихи бұрмаланулар болса, талдау нәтижесінде әділетсіз қорытындылар пайда болуы мүмкін.

Сондықтан деректердің құпиялығы, пайдалануға рұқсат, модельдің сенімділігі және нәтижелердің түсіндірілетіндігі міндетті түрде ескерілуі қажет.

25. Дәрісті қорытындылау

Бүгін біз интеллектуалды деректер талдауының негізгі ұғымдарын және оның интеллектуалды ақпараттық жүйелердегі рөлін қарастырдық.

Data Mining технологиясы бастапқы деректерден пайдалы заңдылықтарды анықтауға бағытталатынын түсіндік.

KDD үдерісінің деректерді таңдау, тазалау, түрлендіру, талдау және нәтижелерді бағалау кезеңдерінен тұратынын зерттедік.

Классификация, кластеризация, ассоциациялық ережелер және аномалияларды анықтау әдістерінің айырмашылықтарын талдадық.

K-means алгоритмі, Евклид қашықтығы, Support, Confidence және Lift формулалары бойынша қарапайым есептер шығардық.

Дәрістің негізгі қорытындысы: деректердің үлкен көлемде жиналуы өздігінен білім қалыптастырмайды. Пайдалы білім алу үшін деректерді дұрыс дайындау, сәйкес алгоритмді таңдау, нәтижелерді тексеру және оларды зерттеу мақсатына сай түсіндіру қажет.

Практикалық тапсырма

Университеттің ақпараттық жүйесінде келесі мәліметтер сақталған:

Студент	GPA	Қатысу (%)	Кеш тапсырмалар
A	3.7	95	0
B	2.3	58	6
C	3.5	90	1
D	2.1	55	7
E	3.0	78	3
F	2.9	75	2

Тапсырмалар:

1. Осы деректердегі объектілер мен белгілерді анықтаңыз.
2. Қандай Data Mining міндеттерін қолдануға болатынын түсіндіріңіз.
3. Студенттерді екі кластерге бөлудің ықтимал жолын ұсыныңыз.
4. Қатысу көрсеткіштерінің орташа мәнін есептеңіз.
5. Қатысуы 60%-дан төмен студенттерді анықтаңыз.
6. Қандай қосымша деректер талдау сапасын арттыруы мүмкін екенін түсіндіріңіз.
7. Неліктен осы алты жазба бойынша сенімді қорытынды жасауға болмайтынын негіздеңіз.

Бақылау сұрақтары

1. Data Mining дегеніміз не?
2. Data Mining мен Machine Learning арасындағы айырмашылық қандай?
3. KDD қандай кезеңдерден тұрады?
4. Деректерді тазалау не үшін қажет?
5. Нормализация мен стандарттау арасындағы айырмашылық қандай?
6. Классификация қандай міндетті шешеді?
7. Кластеризацияның классификациядан айырмашылығы қандай?
8. K-means алгоритмі қалай жұмыс істейді?
9. Евклид қашықтығы қалай есептеледі?
10. Ассоциациялық ережелер дегеніміз не?
11. Support көрсеткіші нені білдіреді?
12. Confidence көрсеткіші қалай есептеледі?
13. Lift көрсеткішінің мағынасы қандай?
14. Аномалия дегеніміз не?
15. Data Mining нәтижелерін бағалау не үшін қажет?
16. Data Mining технологиясын білім беру саласында қалай қолдануға болады?
17. Деректерді талдауда қандай этикалық мәселелер туындайды?
18. Неліктен статистикалық байланыс себеп-салдарлық қатынасты дәлелдемейді?

Пайдаланылған әдебиеттер

1. Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann, 2011.
2. Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V. Introduction to Data Mining. 2nd ed. Pearson, 2019.
3. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. Morgan Kaufmann, 2016.
4. Aggarwal, C. C. Data Mining: The Textbook. Springer, 2015.
5. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996.