Vilniaus universitetas
Matematikos ir informatikos fakultetas

Duomenų analizė

prof. dr. Olga Kurasova
Olga.Kurasova@mif.vu.lt

Duomenys ir jų analizė

- Tarkime turime **objektus**, kuriuos apibūdina tam tikri **požymiai**.
- **Objektais gali būti** pacientai, įrenginiai, gamybos procesai, gamtos reiškiniai ir kt.
- Objektus žymėkime $X_1, X_2, \ldots, X_m$, o požymius $x_1, x_2, \ldots, x_n$. Čia m – objektų skaičius analizuojamoje aibėje, n – juos apibūdinančių požymių skaičius.
- Tam tikras visų požymių reikšmių rinkinys, nusako vieną konkretų analizuojamos aibės objektą $X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), i \in \{1, \ldots, m\}$.

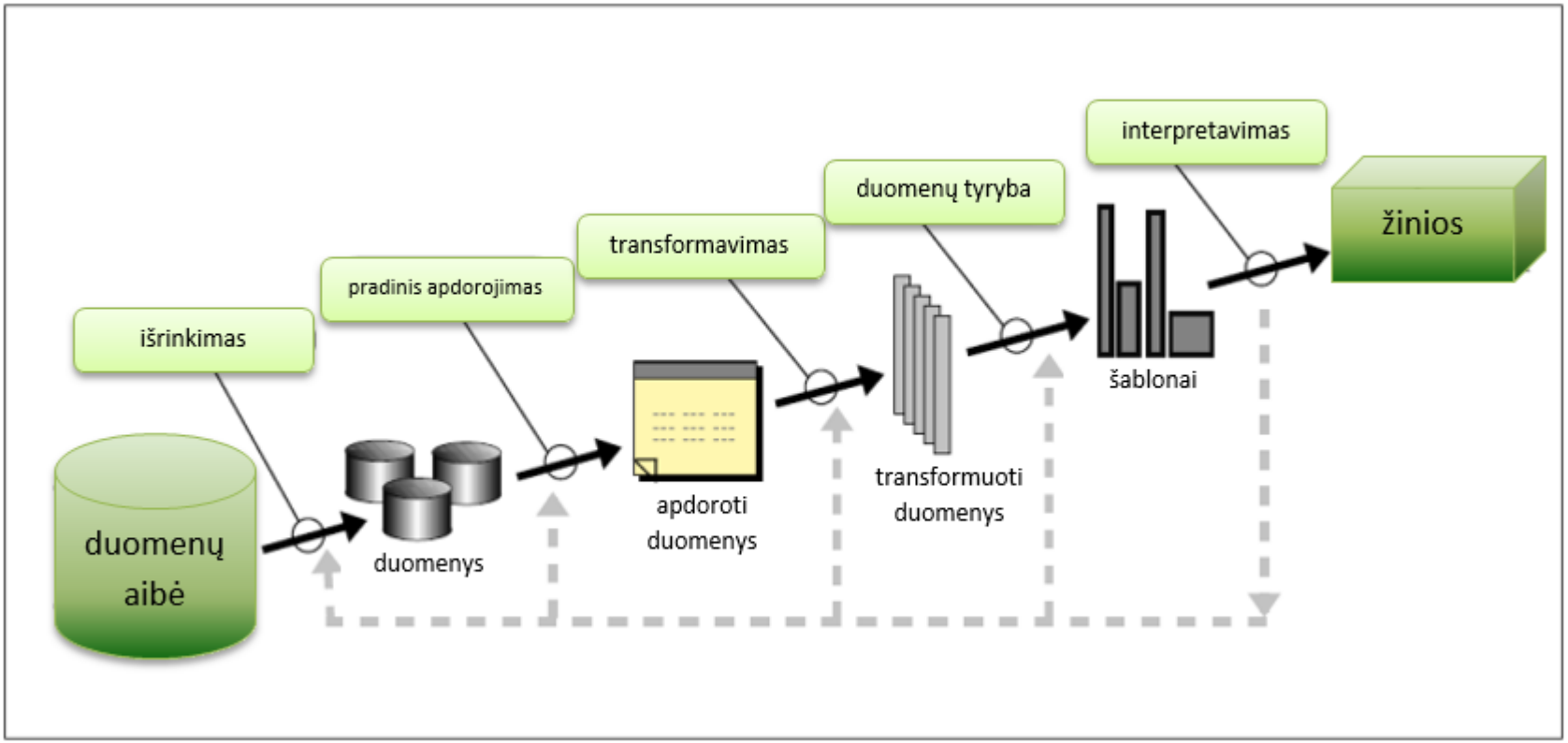
Duomenų analizė

- Vienas iš dažniausiai sprendžiamų uždavinių, pasitelkus skaitmeninio intelekto metodus, yra **duomenų analizė**.
- **Duomenų analizė** – tai procesas, kurio metu iš pradinių duomenų, juos apdorojant įvairiais metodais, gaunama naudinga informacija ir žinios.
- **Duomenų analizė** – tai duomenų nagrinėjimo procesas, kuriuo siekiama išryškinti naudingą informaciją, padaryti išvadas ir padėti sprendimų priėmėjui įgyti naujų žinių.
- Čia svarbios trys sąvokos – **duomenys**, **informacija**, **žinios**.

Duomenys, informacija, žinios

- **Duomenys** – tai objektyviai egzistuojantys faktai, vaizdai, garsai, kurie gali būti naudingi tam tikram uždaviniui spręsti.
- **Informacija** – tai duomenys, kurių forma ir turinys yra pateikti būdu, tinkamu naudoti sprendimų priėmimo procese. Duomenys virsta informacija, kai jiems suteikiamas kontekstas ir jie susiejami su tam tikra problema ar sprendimu.
- **Žinios** – tai gebėjimas spręsti problemas, atnaujinti arba sukurti naujas vertes remiantis ankstesne patirtimi, įgūdžiais ar išmokimu. Tai žmogaus proto abstrakcija apie duomenis, jų prasmę, naudą ir sąryšius. Turimos žinios gali virsti informacija, kuri gali būti panaudota naujoms žinioms įgyti.

Duomenų analizė (tyryba) žinių radimo procese



Data Mining in Knowledge Discovery in Databases

Žinių radimo procesą sudarantys žingsniai:

- Iš visos duomenų aibės **išrenkami** analizuojami duomenys;
- Atliekamas **pradinis duomenų apdorojimas** (valymas, filtravimas, transponavimas, požymių atrinkimas, normavimas);
- Atliekamas **duomenų transformavimas**, kurio metu duomenys paruošiami duomenų analizės metodui ir programinei įrangai tinkama forma;
- **Duomenys analizuojami** įvairias duomenų analizės (tyrybos) metodais;
- Interpretuojami ir vertinami gauti rezultatai, ko pasėkoje įgyjamos **naujos žinios**.

Duomenys ir jų analizė

- Apsiribosime duomenų analize, kai požymiai įgyja tam tikras **skaitines reikšmes**. Tuomet duomenų aibė yra matrica (lentelė)

$$X = \{X_1, X_2, \ldots, X_m\} = \{x_{ij}, i = 1, \ldots, m, j = 1, \ldots, n\}.$$

kurios i-oji eilutė yra vektorius (taškas) $X_i \in R^n$, čia $X_i = (x_{i1}, x_{i2}, \ldots, x_{in}), i \in \{1, \ldots, m\}$.

		Požymiai			
		x_1	x_2	...	x_n
Objektai	X_1	x_{11}	x_{12}	...	x_{1n}
	X_2	x_{21}	x_{22}	...	x_{2n}
	...	...	...	...	...
	X_m	x_{m1}	x_{m2}	...	x_{mn}

Duomenys ir jų analizė

Duomenų bazės terminais:

- Duomenų lentelės **eilutės** (objektai) atitinka **įrašus**,
- **Stulpeliai** (požymiai) atitinka **atributus** (laukus).

Duomenų pradinis apdorojimas

Duomenų valymas –

tai veikla, kurios metu patikrinama, ar duomenyse nėra trūkstamų reikšmių (*missing value*),

nustačius, kad tokių reikšmių yra, joms arba priskiriamos tam tikros reikšmės, pavyzdžiui, to požymio reikšmių vidurkis,

arba objektai su trūkstamomis reikšmėmis iš analizuojamos duomenų aibės yra pašalinami.

Duomenų pradinis apdorojimas

- **Duomenų filtravimas** – tai tam tikromis savybėmis pasižyminčių objektų atmetimas iš nagrinėjamos duomenų aibės.
- Duomenys gali būti filtruojami pagal tam tikras **taisykles**, pavyzdžiui, paliekamos tik tos reikšmės, kurios yra didesnės (ar mažesnės) už nustatytą dydį.
- Tokiu būdu iš duomenų **atmetamos išskirtys** (outliers) – nuo pagrindinės duomenų masės labai nutolusius ir prieštaraujančius jos tendencijoms objektus.
- **Pavyzdžiui**, skaičiuojant gyventojų vidutinį atlyginimą tikslinga atmesti labai didelį atlyginimą gaunančių asmenų įrašus, kadangi tai gali stipriai iškreipti vidurkį.

Duomenų pradinis apdorojimas

- **Duomenų požymių atrinkimas** – tai naujos duomenų aibės, sudarytos tik iš pasirinktų analizuojamų duomenų požymių reikšmių, suformavimas.
- **Duomenų normavimas** – tai procesas, kurio metu suvienodinamos duomenų požymių skalės; įprastai reikšmės suvedamos į intervalą [0; 1].
- **Duomenų standartizavimas** – tai procesas, kurio metu duomenų požymiai pakeičiami taip, kad jų vidurkis būtų lygus 0, o dispersija 1.

Duomenų normavimas

- Tegu $X_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ i = 1, \dots, m.$ Duomenys **normuojami** kiekvieno požymio reikšmę pakeičiant:

$$x_{ij} \leftarrow \frac{x_{ij} - \min(x_{1j}, x_{2j}, \dots, x_{mj})}{\max(x_{1j}, x_{2j}, \dots, x_{mj}) - \min(x_{1j}, x_{2j}, \dots, x_{mj})}.$$

- Čia $\min(x_{1j}, x_{2j}, \dots, x_{mj})$ yra j-tojo požymio reikšmių **minimali reikšmė**; $\max(x_{1j}, x_{2j}, \dots, x_{mj})$ yra j-tojo požymio reikšmių **maksimali reikšmė**.

- Po normavimo kiekvieno požymio minimalios reikšmės tampa lygios 0, o maksimalios 1.

Duomenų standartizavimas

- Tegu $X_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ i = 1, \dots, m.$ Duomenys **standartizuojami** kiekvieno požymio reikšmę pakeičiant:

$$x_{ij} \leftarrow \frac{x_{ij} - \mu(x_{1j}, x_{2j}, .., x_{mj})}{\sigma(x_{1j}, x_{2j}, .., x_{mj})}$$

- Čia $\mu(x_{1j}, x_{2j}, .., x_{mj})$ yra j-tojo požymio reikšmių **vidurkis**; $\sigma(x_{1j}, x_{2j}, .., x_{mj})$ yra j-tojo požymio reikšmių **standartinis nuokrypis**.
- Po standartizavimo kiekvieno požymio reikšmių vidurkiai tampa lygūs 0, o dispersijos 1.

Pagrindiniai duomenų analizės uždaviniai

- **Klasifikavimas** – duomenų priskyrimas klasėms;
- **Atpažinimas** – pagal turimą informaciją ir žinias, atpažįstami žinomi objektai.
- **Prognozavimas** – duomenų reikšmių numatymas;
- **Klasterizavimas** – duomenų suskirstymas į grupes (klasterius) pagal jų panašumą.

Duomenų klasifikavimas (1)

- **Klasifikavimas** – vienas dažniausių duomenų analizėje sprendžiamų uždavinių.
- Klasifikavimo tikslas – **priskirti** duomenis tam tikrai **klasei**.
- Įprastai daliai duomenų klasės yra žinomos. Pritaikius klasifikavimo metodą, klasės yra **nustatomos duomenims**, kurių klasės nebuvo žinomos.

Duomenų klasifikavimas (2)

- Klasifikavimo uždaviniai dažnai sprendžiami **medicinoje**, siekiant nustatyti **preliminarią diagnozę**.
- Tarkime, turime pacientų kraujo tyrimų duomenis ir žinome, kad dalis pacientų serga tam tikra liga, kiti pacientai yra sveiki. Vadinasi, turime dviejų klasių duomenis: **sergantys**, **sveiki**.
- Taip pat turime pacientus, kurių kraujo tyrimo rezultatai yra žinomi, tačiau **jie nėra priskirti** nei vienai klasei.
- **Klasifikavimo tikslas** – priskirti šiuos pacientus vienai iš dviejų klasių.

Duomenų klasifikavimas (3)

	x_1	x_2	x_3	x_4	**Klasė**
X_1	85	0,001	24,1	1025	sveikas
X_2	77	0,002	21,3	2036	sveikas
X_3	68	0,015	35,8	1059	sveikas
…	…	…	…	…	…
X_{101}	101	0,001	22,4	3011	serga
X_{102}	95	0,001	28,0	2645	serga
…	…	…	…	…	…
X_{201}	86	0,002	30,1	2987	**???**
X_{202}	72	0,010	19,5	1259	**???**

Atpažinimo uždavinys

- Vienas iš klasifikavimo uždavinių – yra **atpažinimo uždavinys**.
- Čia duomenys dažniausiai yra **vaizdai** arba **garsai** (signalai).
- Atpažinimo uždavinio tikslas – pagal apmokytus duomenis kitus duomenis priskirti tam tikroms klasėms (**juos atpažinti**).
- **Charakteringi pavyzdžiai** – ranka rašyto teksto atpažinimas, veido atpažinimas, automobilio numerio atpažinimas.

Prognozavimo uždavinys

- Dar vienas populiarus duomenų analizės uždavinys yra **prognozavimas**, kurio metu žinant duomenų dalies požymių reikšmes, nustatomos reikšmės požymiui, kurio reikšmė nežinoma.
- **Prognozavimo tikslas** – iš „istorinių“ duomenų nustatyti reikšmes „ateities“ duomenims.
- Prognozavimo uždaviniai sprendžiami įvairiose srityse. **Pavyzdžiui**, meteorologijoje remiantis ankstesnių metų to paties laikotarpio duomenimis bei pastarojo laikotarpio pokyčių prognozuojama oro temperatūra ateinančiai savaitei.
- Taip pat prognozavimas atliekamas vertybinių popierių biržoje, įvairiose finansinėse rinkose ir kitur.

Duomenų klasterizavimas

- **Klasterizavimo** tikslas – suskirstyti objektus taip, kad panašūs objektai patektų į tą patį klasterį, o skirtingi – į skirtingus.
- **Klasteris** – tai panašių objektų grupė.
- Čia svarbu nustatyti objektų **panašumo matą**. Vienas paprasčiausių panašumo matų – gerai žinomas Euklido atstumas.