

---

## Laboratorinis darbas Nr.1. Duomenų apdorojimas rinkinio analizė

---



1. Pasirinkti (susikurti) duomenų rinkinį<sup>1,2</sup>, su kuriuo atliksite šį ei sekančius laboratorinius darbus. Jūsų pasirinkimą turi patvirtinti vienas iš laboratorinių darbų dėstytojų<sup>3</sup>.

Duomenų rinkinio reikalavimai:

- Turi egzistuoti skaitinės (*integer* ir *real* tipo) ir /arba kategorinės reikšmės. Duomenų rinkinys kuriame yra tik kategorinio tipo atributai **yra netinkamas**.
- Duomenų rinkinyje įrašų (eilučių)  $m$  turi būti ne mažiau nei 500, t.y.,  $\infty > m \geq 500$  ir atributų  $n$  nemažiau nei 8 (stulpeliai)  $\infty > n \geq 8$ . Jeigu atributų  $n$  pasirinktame duomenų rinkinyje yra mažiau, privalote pridėti išvestinius (sukurtus) atributus (žr. pav. 1.)

**Svarbu.** Sekančios užduotys turi būti realizuotos programiškai naudojant *Matlab* arba *Python*.

Atributai					
	1	2	3	4	n
1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	...
2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	...
3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	...
...					
m	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	...

pav. 1. Duomenų aibės grafinis atvaizdavimas

2. Atlikti duomenų rinkinio kokybės analizę (žr. 2 pav.). Kiekvienam **tolydinio** tipo atributui paskaičiuoti:
- bendrą reikšmių skaičių,
  - trūkstamų reikšmių procentą,
  - kardinalumą,
  - minimalią (*min*) ir maksimalią (*max*) reikšmes,
  - 1-ąją ir 3-ją kvartilius,
  - vidurkį,
  - medianą,
  - standartinį nuokrypį.
3. Kiekvienam **kategorinio** tipo atributui paskaičiuoti:
- bendrą reikšmių skaičių,
  - trūkstamų reikšmių procentą,
  - kardinalumą,
  - modą,
  - modos dažnumo reikšmę
  - modos procentinę reikšmę
  - 2-ąją modą,
  - 2-osios modos dažnumo reikšmę,
  - 2-osios modos procentinę reikšmę.

---

<sup>1</sup> <https://archive.ics.uci.edu/ml/datasets.php>

<sup>2</sup> <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

<sup>3</sup> A.Tarasevičienė, G.Budnikas, A.Nečiūnas

Tolydinio tipo reikšmės										
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-iasis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
Kategorinio tipo reikšmės										
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %	

pav. 2. Tolydinio ir kategorinio tipo duomenų analizės kokybės parametrų lentelės

- Nupaišyti atributų histogramas. Ataskaitoje pateikti aprašymus, koks tai pasiskirstymas (pvz., *normalusis*, *vien(a)modalis*, *eksponentinis* ir t.t.) ir kokias išvadas pagal tai galima formuluoti.
- Identifikuoti duomenų kokybės problemas: trūkstamos reikšmės, kardinalumo problemas, išskirtis – ekstremalias reikšmes (angl. *outliers*). Pateikti šių problemų sprendimo planą, kuris bus realizuotas programiškai (pvz., bus įtraukiamos trūkstamos kategorinio atributo reikšmės remiantis atributo moda įverčiu, ekstremalios reikšmės yra šalinamos ar koreguojamos).
- Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus:
  - Tolydinio tipo atributams:** naudojant „scatter plot“ tipo diagramą pateikti kelis (2-3) pavyzdžius su stipria tiesine atributų priklausomybe (tiesioginė arba atvirkštinė koreliacija) bei kelis pavyzdžius su tarpusavyje nekoreliuojančiais (silpnai koreliuojančiais) atributais. Pakomentuoti rezultatus.
  - Pateikti SPLOM diagramą (Scatter Plot Matrix).
  - Kategorinio tipo atributams:** naudojant „bar plot“ tipo diagramą pateikti keletą (2-3) atributų priklausomybės pavyzdžių ir pakomentuoti rezultatus.
  - Pateikti keletą (2-3) histogramų ir „box plot“ diagramų pavyzdžių, vaizduojančių sąryšius tarp **kategorinio** ir **tolydinio** tipo kintamųjų<sup>4</sup>.
- Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą. Rezultatus pakomentuoti.
- Atlikti duomenų normalizaciją.
- Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius<sup>2</sup>.

<sup>4</sup> Tik tokiu atveju jeigu duomenų rinkinyje yra ir kategorinio ir tolydinio tipo kintamųjų.