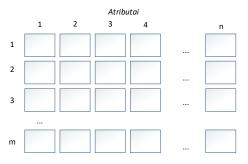
Laboratorinis darbas Nr.1. Duomenų apdorojimas rinkinio analizė

- Pasirinkti (susikurti) duomenų rinkinį^{1,2}, su kuriuo atliksite šį ei sekančius laboratorinius darbus. Jūsų pasirinkimą turi patvirtinti vienas iš laboratorinių darbų dėstytojų³. 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., $\infty > m \ge 500$ ir atributų n nemažiau nei 8 (stulpeliai) $\infty > n \ge 8$. 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*.



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,
 - kardinaluma,
 - minimalią (min) ir maksimalią (max) reikšmes,
 - 1-aja ir 3- ja kvartilius,
 - vidurki,
 - mediana,
 - standartinį nuokrypį.
- 3. Kiekvienam **kategorinio** tipo atributui paskaičiuoti:
 - bendrą reikšmių skaičių,
 - trūkstamų reikšmių procentą,
 - kardinaluma,
 - modą,
 - modos dažnumo reikšmę
 - modos procentinę reikšmę
 - 2-aja moda,
 - 2-osios modos dažnumo reikšmę,
 - 2-osios modos procentinę reikšmę.

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

² https://vincentarelbundock.github.io/Rdatasets/datasets.html

³ A.Tarasevičienė, G.Budnikas, A.Nečiūnas

Tolydinio tipo rei	kšmėms									
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-iasis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
Kategorinio tipo	reikšmėms									
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

- 4. 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.
- 5. Identifikuoti duomenų kokybės problemas: trūkstamas reikšmes, kardinalumo problemas, išskirstis 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).
- 6. 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 diagrama (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 kintamųjų⁴.
- 7. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą. Rezultatus pakomentuoti.
- 8. Atlikti duomenų normalizaciją.
- 9. Kategorinio tipo kintamuosius paversti i tolydinio tipo kintamuosius².

_

⁴ Tik tokiu atveju jeigu duomenų rinkinyje yra ir kategorinio ir tolydinio tipo kintamųjų.