KAUNO TECHNOLOGIJOS UNIVERSITETAS

INFORMATIKOS FAKULTETAS

Programavimo kalbų teorija (P175B124)

Laboratorinių darbų ataskaita

Atliko:

IFF-7/14 gr. studentas

Eigijus Kiudys

2020 m. kovo 3 d.

Priėmė:

lekt. Evaldas Guogis

lekt. Fyleris Tautvydas

KAUNAS 2020

# Turinys

[Turinys 2](#_Toc34089641)

[Įvadas 3](#_Toc34089642)

[Laboratorinio darbo užduotis 3](#_Toc34089643)

# Įvadas

Laboratorinis darbas Nr.1. Duomenų apdorojimas rinkinio analizė

## Laboratorinio darbo užduotis

Pasirinkti (susikurti) duomenų rinkinį, su kuriuo atliksite šį ei sekančius laboratorinius darbus. Jūsų pasirinkimą turi patvirtinti vienas iš laboratorinių darbų dėstytojų. 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., ∞ > m ≥ 500 ir atributų n nemažiau nei 8 (stulpeliai) ∞ > n ≥ 8. Jeigu atributų n pasirinktame duomenų rinkinyje yra mažiau, privalote pridėti išvestinius (sukurtus) atributus

Užduotys:

1. 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.

1. 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.

1. Nupaišyti atributų histogramas. Pateikti aprašymus, koks tai pasiskirstymas ir kokias išvadas pagal tai galima formuluoti.
2. Identifikuoti duomenų kokybės problemas:

* trūkstamas reikšmes
* kardinalumo problemas
* triukšmus– ekstremalias reikšmes (angl. outliers).
* Pateikti šių problemų sprendimo planą, kuris bus realizuotas programiškai.

1. 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 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

1. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą. Rezultatus pakomentuoti.
2. Atlikti duomenų normalizaciją ( režiai [0;1] arba [-1;1]).
3. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.

## Pasirinkti darbo įrankiai:

Pasirinkta programavimo kalba: Python.

Pasirinkta integruota kūrimo aplinka: PyCharm.

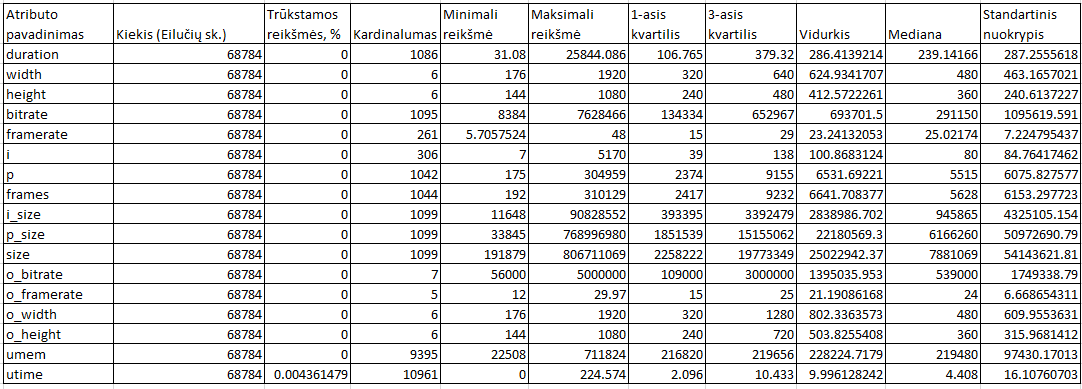
Duomenų rinkinys: Internetinių vaizdo įrašų charakteristikos ir perkodavimo laikas (angl. Online Video Characteristics and Transcoding Time)

|  |  |
| --- | --- |
| Duomenų rinkinys | |
| Įrašų kiekis | 68785 |
| Atributų kiekis | 20 |
| Naudojamų atributų kiekis | 19 |
| Tolydinio tipo atributų kiekis | 17 |
| Kategorinio tipo atributų kiekis | 2 |
| Tolydus atributai | |
| Trukmė (angl. Duration) | |
| Plotis (angl. Width) | |
| Aukštis (angl. Height) | |
| Pralaidumas (angl. Bitrate) | |
| Kadrų dažnis (angl. Framerate) | |
| I (angl. I) | |
| P (angl. P) | |
| Kadrai (angl. Frames) | |
| I\_dydis (angl. I\_size) | |
| P\_dydis (angl. P\_size) | |
| Dydis (angl. Size) | |
| Išėigos pralaidimas (angl. O\_bitrate) | |
| Išeigos kadrų dažnis (angl. O\_framerate) | |
| Išeigos plotis (angl. O\_width) | |
| Išeigos aukštis (angl. O\_height) | |
| Sunaudotas RAM kiekis (angl. Umem) | |
| Užtruktas laikas (angl. Utime) | |
| Kategoriniai atributai | |
| Kodekas (angl. Codec) | |
| Konvertuotas Kodekas (angl. O\_codec) | |

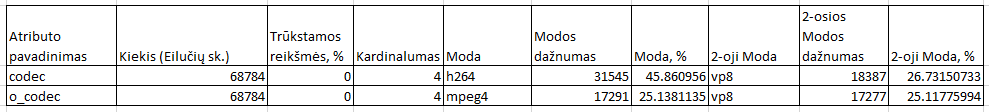
# Užduočių sprendimai

Atmečiau šį atributą: ID.

## Atlikti duomenų rinkinio kokybės analizę. Kiekvienam tolydinio tipo atributui.



## Atlikti duomenų rinkinio kokybės analizę. Kiekvienam kategorinio tipo atributui.



## Nupaišyti atributų histogramas.