

Duomenų tyrybos įrankių taikymo galimybės

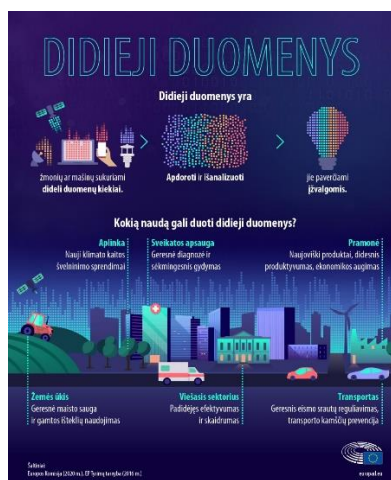
1. Duomenų tyryba

- a. **Duomenų tyryba** – tai procesas, kurio metu, naudojant įvairius duomenų analizės įrankius, bandoma nustatyti ir atrasti „užslėptas“ duomenų struktūras ir ryšius.
- b. **Duomenų tyrybos tikslas** – iš įprastai didelių duomenų aibių išgauti svarbią informaciją ir atmesti nereikšmingą. Duomenų tyrybos metodai padeda „nepasiklysti“ duomenų ir informacijos gausoje.
- c. **Duomenų tyryba** atliekama žinių radimo duomenų aibėse procese (angl. knowledge discovery in databases), kurio metu didelių apimčių duomenų aibėse ieškoma naujos informacijos, padėsiančios įgyti žinias apie analizuojamus duomenis ir priimti tinkamiausius sprendimus.

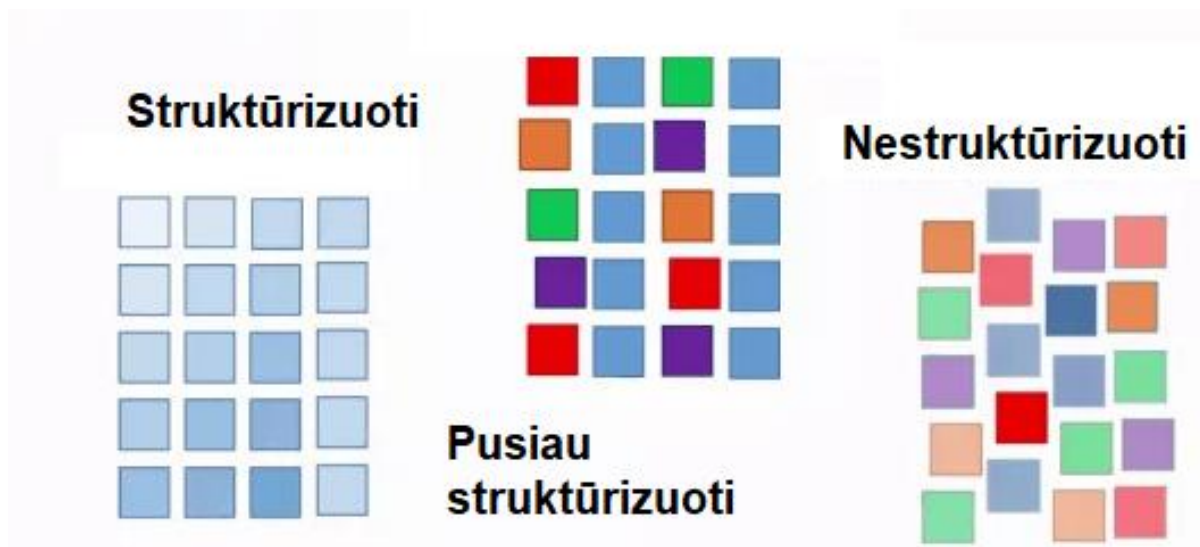


(Šaltinis: https://www.linkedin.com/pulse/data-mining-software-market-swot-analysis-size-status-kdrnf?trk=public_post_main-feed-card_feed-article-content)

2. **Didieji duomenys** (angl. *big data*) – tai žmonių ar mašinų sukuriama dideli duomenų kiekiai (pirkimo sandorių duomenys, naudojimosi socialiniais tinklais įpročiai, interesai ir pomėgiai, padėties nustatymo sistemų, pvz., GPS, signalai, jutiklių registruojama klimato informacija ir kt.).



3. Didžiųjų duomenų tipai ir šaltiniai



(Šaltinis: <https://www.astera.com/type/blog/structured-semi-structured-and-unstructured-data/>)

- a. Struktūrizuotų duomenų šaltiniai
 - i. Daiktų internetas (IoT), įskaitant įvairių jutiklių, pavyzdžiui, temperatūros ar slėgio ir kt., duomenis, suteikia daug naudingų duomenų. Pavyzdžiui, jei namuose ar mokykloje įrengti daiktų interneto jutikliai, galime automatiškai valdyti šildymą ir apšvietimą, kad sumažintume išlaidas.
 - ii. Apklausa, kurias prašome naudotojų užpildyti po pirkimo arba apsilankius interneto svetainėje.
 - iii. Elgesio analizė gali, pavyzdžiui, padėti suprasti, kaip giliai naudotojas įeina į svetainę ir kokia yra tipinė svetainės palikimo priežastis.
- b. Nestrukūrizuotų duomenų šaltiniai
 - i. Tekstai gali būti turtingas įžvalgų šaltinis, pavyzdžiui, bendras nuotaikų įvertinimas arba raktinių žodžių ir semantinės reikšmės išskyrimas.
 - ii. Vaizdai arba vaizdo įrašai. Stebėjimo kameros vaizdo įrašas gali būti naudojamas siekiant įvertinti eismo intensyvumą kelyje ir informuoti žmones apie galimas spūstis.
 - iii. Interneto serverio žurnalai gali būti naudojami siekiant suprasti, kurie mūsų svetainės puslapiai lankomi dažniausiai ir kiek laiko.
- c. Pusiau struktūrizuotų duomenų šaltiniai
 - i. Socialinių tinklų grafikai gali būti puikus duomenų apie naudotojų asmenybes ir galimą informacijos sklaidos veiksmingumą šaltinis.
 - ii. Kai turime krūvą nuotraukų iš vakarėlio, galime pabandyti išgauti grupės dinamikos duomenis sudarydami žmonių, fotografuojančių vienas kitą, grafiką.

Iš kur paimti didelius duomenų kiekius?

1. Lietuvos atvirų duomenų portalas - <https://data.gov.lt/>

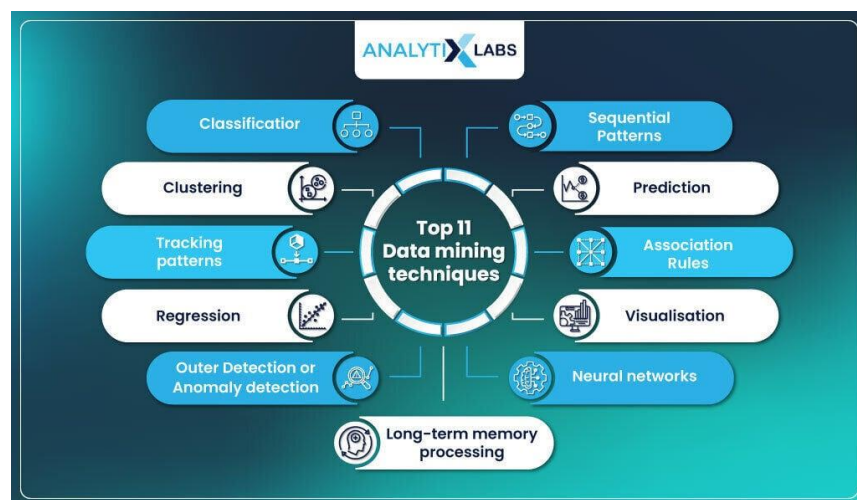
Duomenų rinkinių kategorijos



2. Google DataSet Search – <https://datasetsearch.research.google.com/>
3. Kaggle - <https://www.kaggle.com/datasets>
4. Data Hub – <https://datahub.io/collections>
5. Earth Data – <https://www.earthdata.nasa.gov/>
6. CERN Open Data Portal – <https://opendata.cern.ch/>
7. Global Health Observatory data repository – <https://apps.who.int/gho/data/node.home>
8. Daugiau galima rasti:
 - a. <https://towardsdatascience.com/data-repositories-for-almost-every-type-of-data-science-project-7aa2f98128b>
 - b. <https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/?sh=54b27882b54d>
 - c. <https://www.interviewquery.com/p/free-datasets>
 - d. <https://www.linkedin.com/pulse/30-amazing-free-big-data-ai-public-sources-2018-bernard-marr>

4. Didžiųjų duomenų apdorojimo technikos ir įrankiai

4.1. Duomenų tyrybos technikos



(Šaltinis: <https://www.analytixlabs.co.in/blog/data-mining-techniques/>)

4.2. Duomenų tyrybos įrankių klasifikacija

Duomenų tyrybos įrankiai gali būti klasifikuojami į kelias pagrindines kategorijas, atsižvelgiant į jų paskirtį, funkcionalumą ir taikymo sritis.

a. Duomenų surinkimo ir tvarkymo įrankiai

Šie įrankiai skirti duomenų rinkimui iš įvairių šaltinių, duomenų valdymui, duomenų apdorojimui, valymui ir transformacijai. Pavyzdžiui, programavimo kalbos (Python, R), duomenų bazės valdymo sistemos (MySQL, PostgreSQL), ETL (Extract, Transform, Load) įrankiai ir duomenų valdymo platformos.

b. Duomenų analizės ir modeliavimo įrankiai

Šie įrankiai skirti duomenų analizei, statistikai, mašiniam mokymuisi, prognozavimui ir modeliavimui. Tai apima statistinius paketus (SPSS, SAS), programavimo kalbas (Python, R), mašininio mokymosi bibliotekas (scikit-learn, TensorFlow, Keras), duomenų analizės ir vizualizavimo platformas (Tableau, Power BI) ir kt.

c. Duomenų vizualizacijos įrankiai

Šie įrankiai skirti duomenų vizualizacijai, kad būtų lengviau suprasti duomenų struktūrą, tendencijas ir santykius. Tai apima specializuotas vizualizacijos bibliotekas (matplotlib, ggplot2), vizualizacijos platformas (Tableau, Plotly), diagramų kūrimo įrankius (Microsoft Excel, Google Sheets) ir kt.

d. Duomenų saugos ir privatumo įrankiai

Šie įrankiai skirti užtikrinti duomenų saugumą, privatumą ir atitikimą teisės aktams. Tai apima duomenų šifravimą, duomenų anonimizavimą, prieigos kontrolę, duomenų atkūrimo sistemų kūrimą ir kt.

e. Internetinių duomenų surinkimo ir analizės įrankiai

Šie įrankiai specializuojasi surinkti duomenis iš interneto šaltinių, tokie kaip socialiniai tinklai, interneto svetainės ir kt., ir atlikti analizę. Tai apima interneto skraperius, API klientus, interneto monitoringo įrankius ir kt.

f. Skaitmeninės reklamos ir rinkodaros įrankiai

Šie įrankiai skirti surinkti, analizuoti ir interpretuoti duomenis apie skaitmeninės reklamos ir rinkodaros kampanijų veiksmingumą. Tai apima analitikos platformas (Google Analytics, Facebook Insights), reklamos valdymo įrankius (Google Ads, Facebook Ads Manager), A/B testavimo platformas ir kt.

Daugelis įrankių gali turėti kelių kategorijų savybių, nes jie gali būti naudojami įvairiais tikslais ir įvairiems veiksams atlikti.

5. Konkrečių įrankių panaudojimas

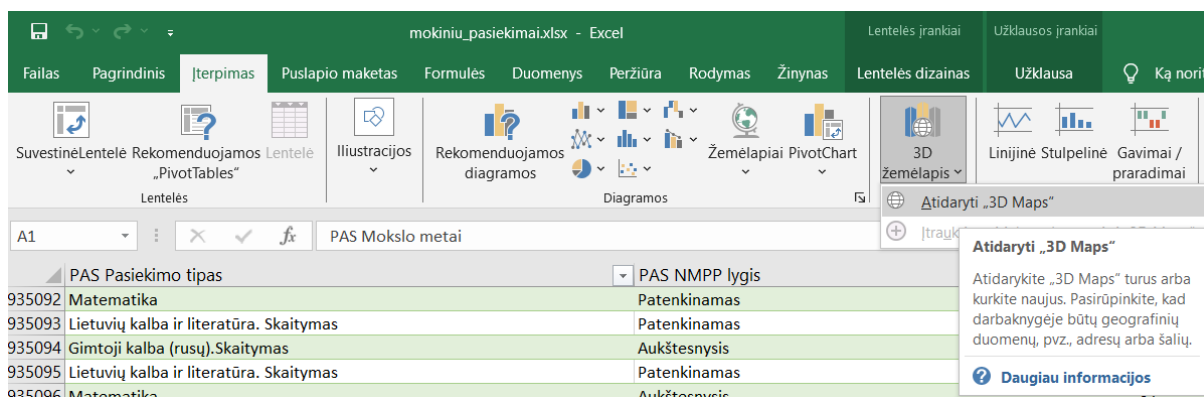
5.1. Skaičiuoklės

Privalumai	Trūkumai
Prieinamumas. Skaičiuoklės yra lengvai prieinamos ir naudojamos, nes jos dažnai įtrauktos į įprastinę biuro programinės įrangos paketą arba prieinamos kaip nemokamos internetinės paslaugos. Paprastumas. Skaičiuoklėse galima lengvai kurti ir redaguoti duomenų lentelės formatą, atlikti skaičiavimus, taikyti formules ir pan. Duomenų vizualizavimas. Jos turi įvairių funkcijų duomenų vizualizacijai, leidžiančių kurti diagramas, grafikus ir kitus vizualinius elementus. Paieška, filtravimas, rūšiavimas ir kt., kad būtų lengviau tvarkyti ir analizuoti duomenis.	Duomenų dydžių apribojimai. Skaičiuoklės dažnai turi ribotą galimybę tvarkyti didelius duomenų kiekius. Kai reikia tvarkyti labai didelius duomenų rinkinius, skaičiuoklės gali tapti nepakankamai efektyviomis arba netinkamomis dėl duomenų apdorojimo greičio ir našumo apribojimų. Duomenų saugumas. Skaičiuoklės gali turėti ribotą duomenų saugumo funkcionalumą, o duomenų redagavimas ir dalijimasis gali sukelti problemų dėl duomenų nuoseklumo ir autentiškumo. Duomenų analizės funkcionalumas. Skaičiuoklės neturi tokių išplėstinės duomenų analizės funkcijų kaip specializuotos duomenų analizės programos.

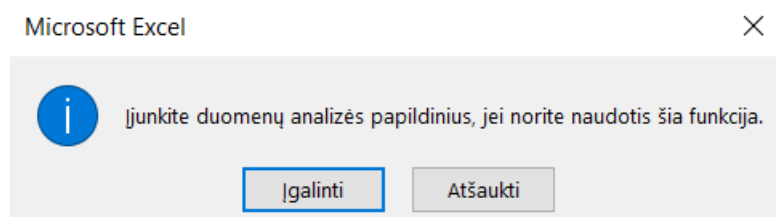
Nepaisant šių trūkumų, skaičiuoklės dažnai naudojamos kaip pradinis duomenų analizės įrankis dėl savo paprastumo ir prieinamumo, tačiau norint atlikti išsamias analizes, gali būti naudinga pasitelkti ir kitus, specializuotus duomenų analizės įrankius.

Skaičiuoklės panaudojimas duomenų vizualizacijai panaudojant 3D žemėlapių įrankį

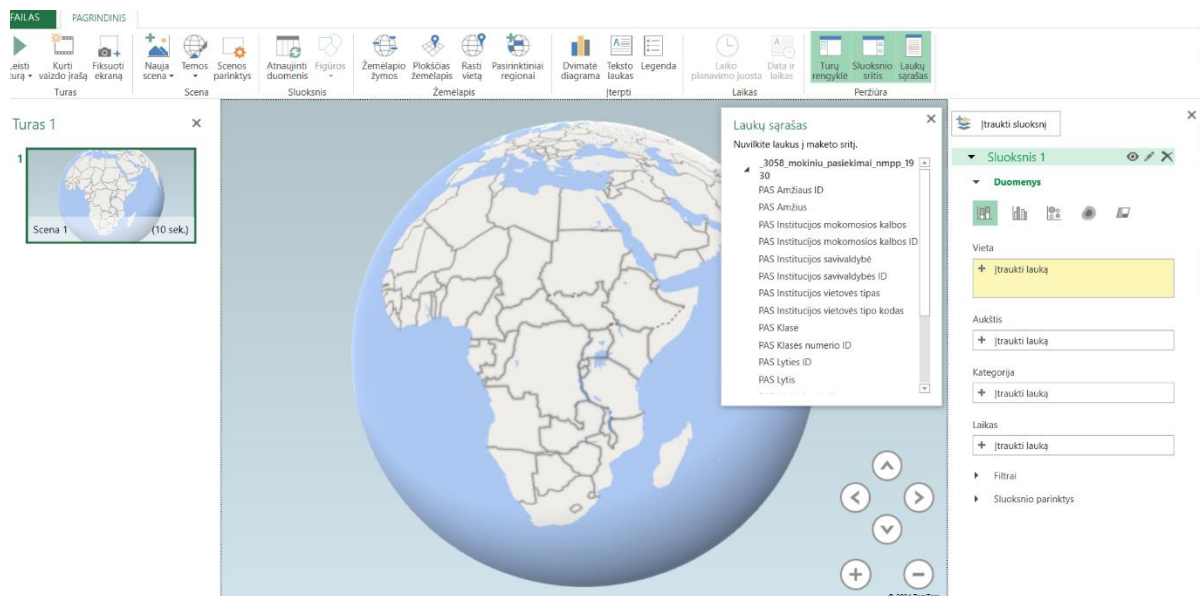
1. Skaičiuoklėje atverkite failą 3058_mokiniu_pasiekimai_nmpp-1930.txt (failą rasite kataloge Duomenų tyrybos įrankiai). Įkeltą failą išsaugokite mokiniu_pasiekimai.xlsx. Šį failą taip pat galite rasti kataloge. Jei nepavyko įkelti *.txt failo, atverkite *.xlsx failą.
2. Pasirinkite:



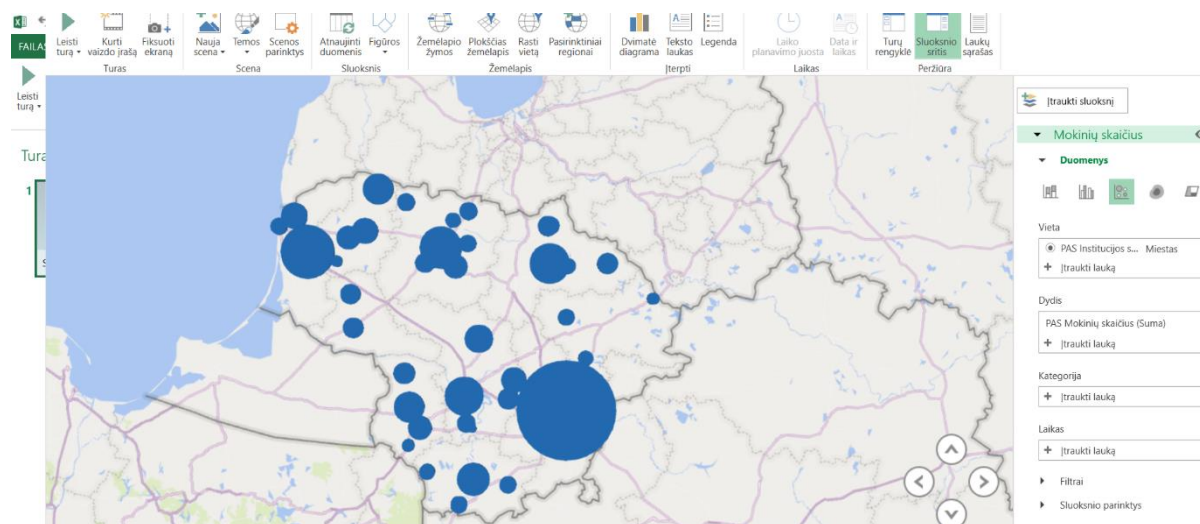
Gali būti prašoma įjungti duomenų analizės papildinius. Įgalinimas gali šiek tiek užtrukti.



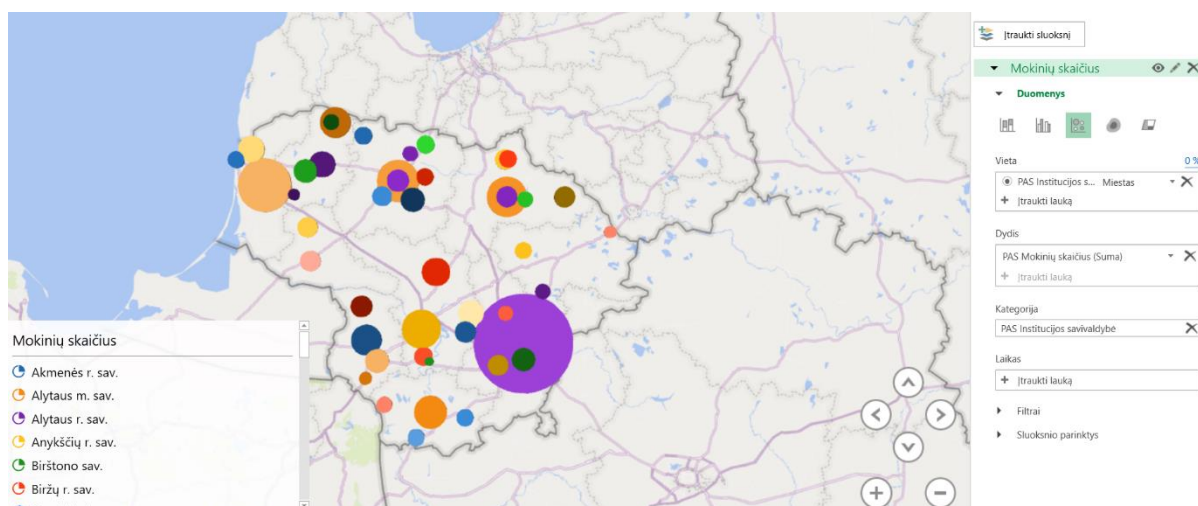
3. Įgalinę duomenų analizės papildinius turėtumėte matyti panašų vaizdą:



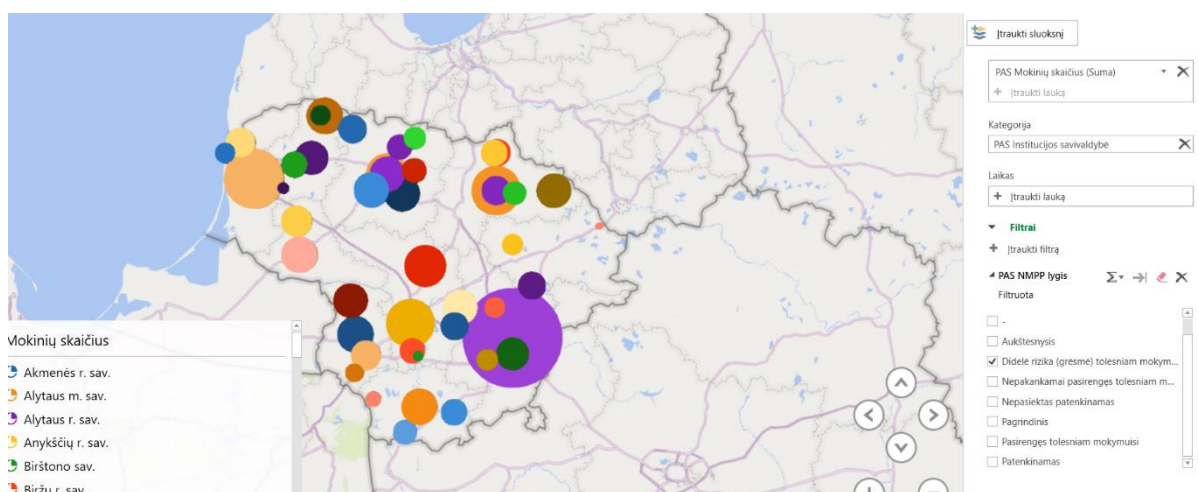
4. Pasirinkite parametrus ir priartinę žemėlapi matysite, kaip pasiskirsto mokinių skaičius įvairiose savivaldybėse.



5. Papildykite žemėlapiį įtraukdami savivaldybę į kategorijų lauką.



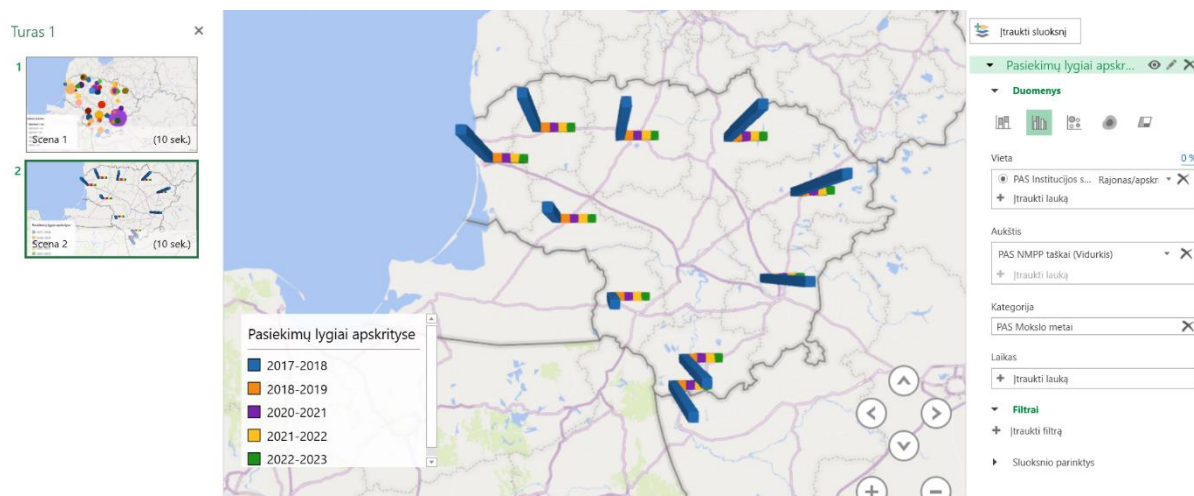
6. Įtraukite filtrą pagal pasiekimus ir stebėkite, kaip keičiasi vaizdas žemėlapyje.



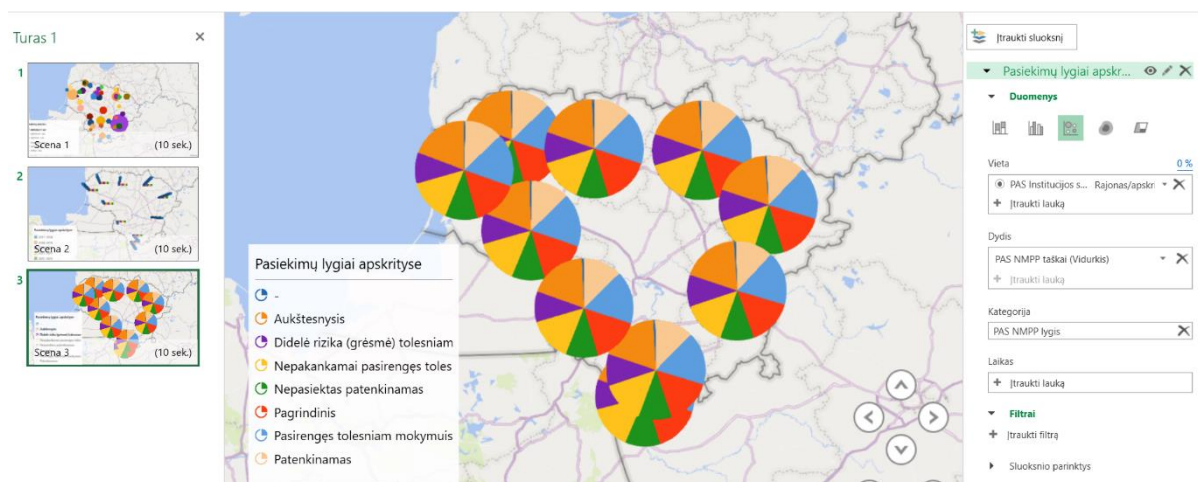
7. Įtraukite naują sceną su pasaulio žemėlapiu:

Nauja scena	Temos	Scenos parinktys	Atnaujinti duomenis	Figūros	Žemėlapių žymos	Plokštumos žemėlapis
Kopijuoti Scena 2						
Sukurkite aktyvios scenos kopiją.						
Pasaulio žemėlapis						
Sukurti naują sceną su pasaulio fonu.						
Naujas pasirinktinis žemėlapis						
Sukurkite naują sceną su naujo paveikslėlio fonu.						

8. Pasirinkite tinkamus laukus norėdami tinkamai pateikti mokinius pagal pasiekimų lygius apskrityse skirtingais mokslo metais:



9. Sukurkite 3 sceną ir joje atvaizduokite Lietuvos apskričių mokinių pasiskirstymą pagal pasiekimų lygius skritulinėmis diagramomis.



10. Išbandykite parametrų pasirinkimo ir keitimo galimybes. Sukurkite ir išsaugokite vaizdo įrašą (pateikiamas pavyzdys mokiniu_pasiekimai.mp4)

5.2. Nemokami duomenų gavybos įrankiai, skirti 11-12 klasių mokiniams (<https://www.predictiveanalyticstoday.com/top-free-data-mining-software/>)

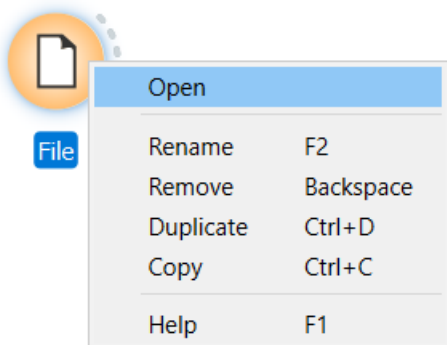
5.3. Orange – dažniausiai rekomenduojamas įrankis duomenų gavybos galimybėms pademonstruoti.

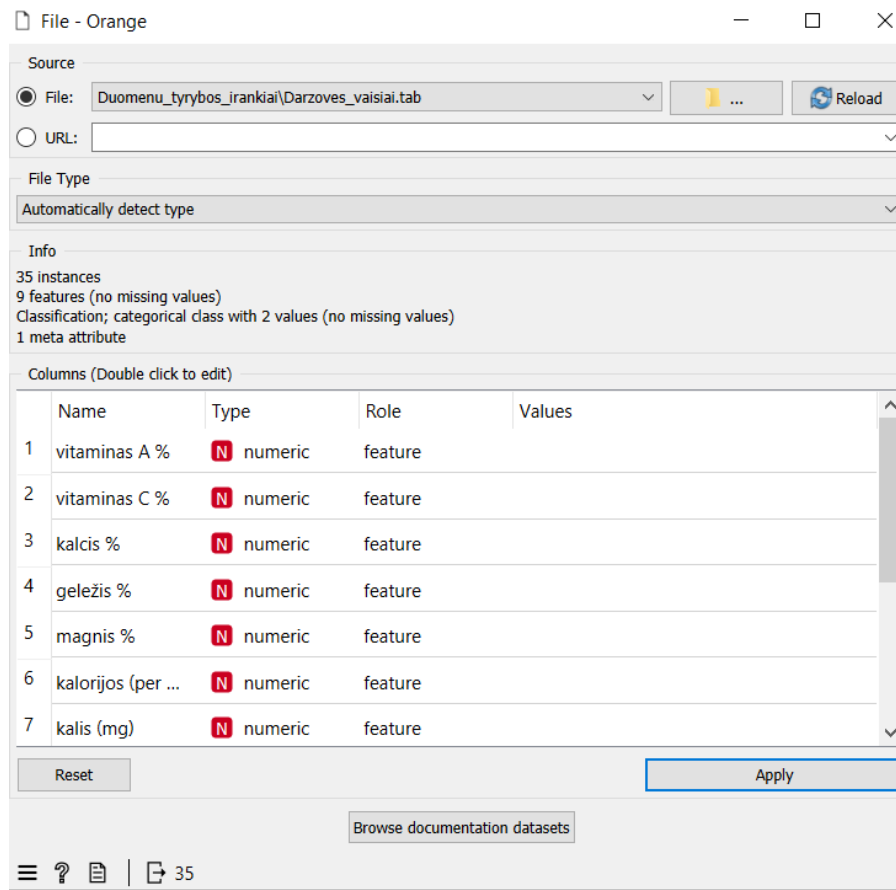
Orange valdikliai



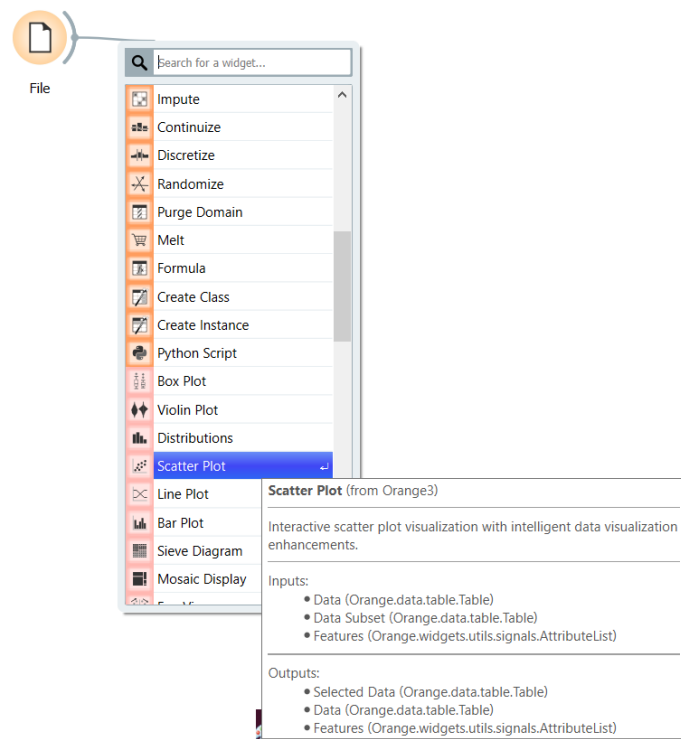
1. Duomenų įkėlimas:

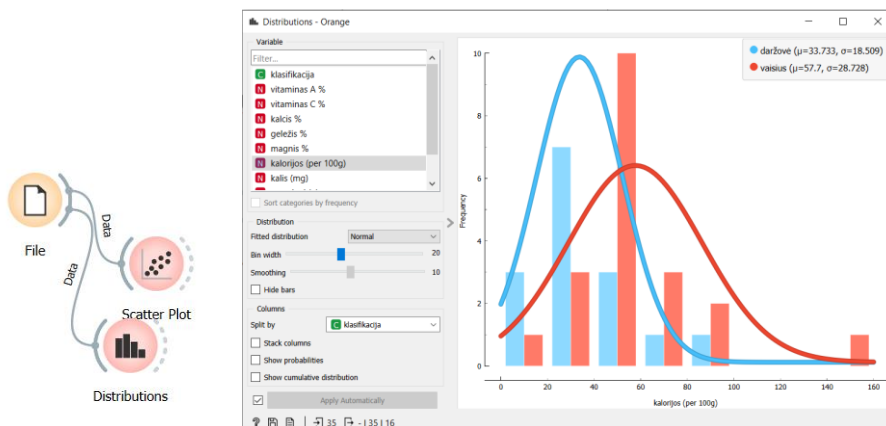
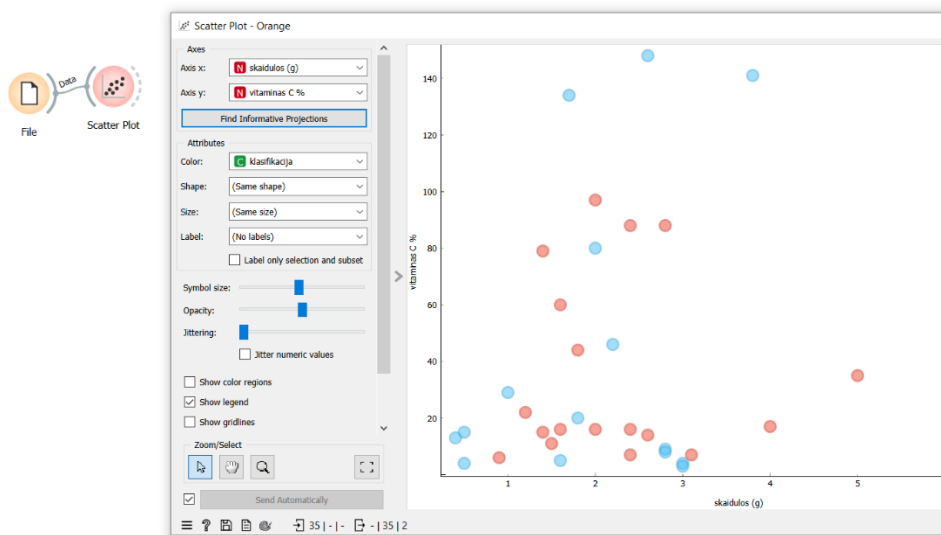
- Skaitiniai duomenys.
- Kategoriniai duomenys – tai duomenys, kuriuos galima suskirstyti į grupes arba kategorijas. Kategorinių duomenų pavyzdžiai: lytis, akių spalva ir plaukų spalva. Kategoriniai duomenys yra diskretieji. Kategoriniai duomenys yra kokybiniai.
- Datos duomenų tipas.
- Eilutės (string) duomenų tipas.
- Įkelsime failą Darzoves_vaisiai.tab.





2. Įkėlus duomenis, galima juos vizualizuoti.



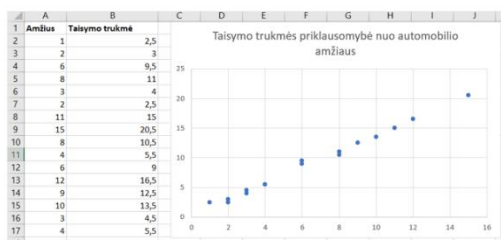


3. Tiesinės regresijos realizacija.

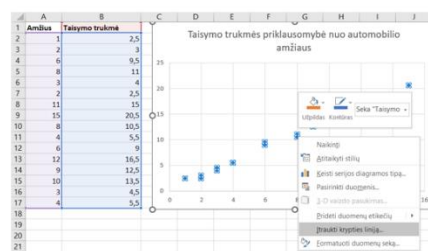
- Tiriama kaip kintamojo Y reikšmės priklauso nuo X reikšmių.
- Sudaromas modelis: parenkama kintamuosius siejanti funkcija: $y = ax + b$.
- Užduotis: patikrinti, ar tinka tiesinės regresijos modelis, jei žinomas automobilio amžius metais ir mechaninių gedimų taisymo laikas. Duomenys pateikti faile automobiliai.xlsx.
- Užduotį atliksime panaudodami skaičiuoklę Excel ir ORANGE.

Tiesinė regresija naudojant skaičiuoklę (1)

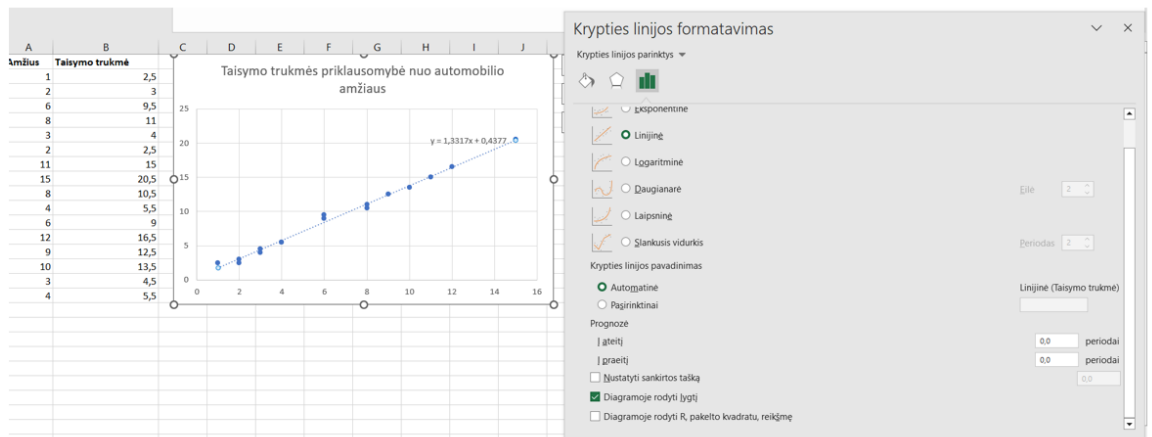
- 1. Sukuriame sklaidos diagramą.



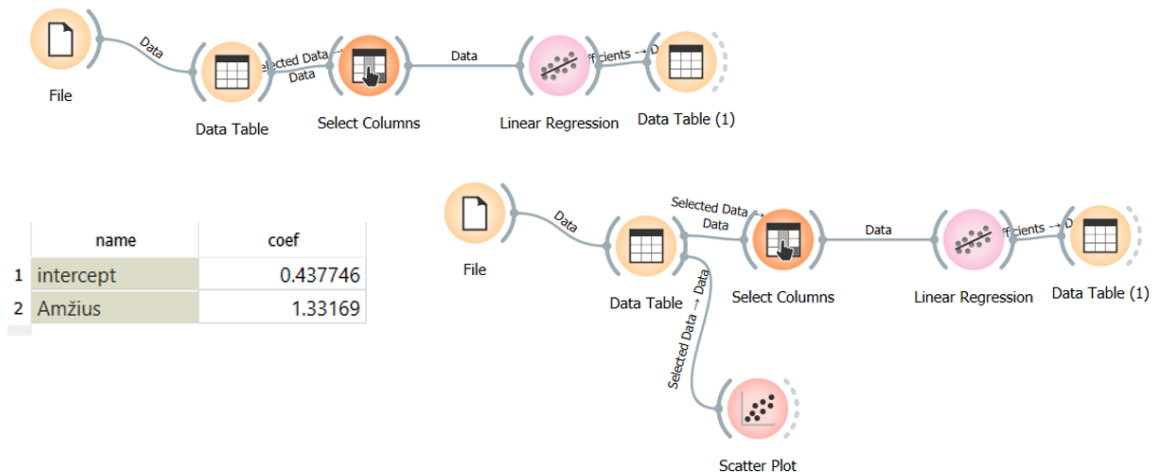
- 2. Pažymime taškus diagramoje ir kontekstiniame meniu pasirenkame Įtraukti krypties liniją..



- Pasirenkame, kad bus tiesinė priklausomybė (linijinė) ir kad diagramoje būtų rodoma lygtis.



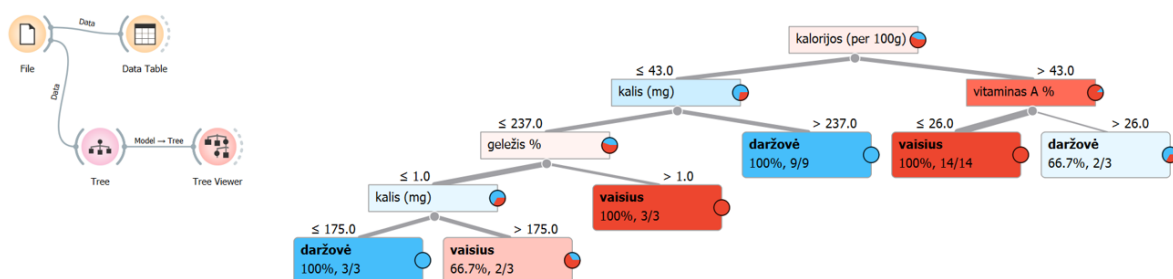
Tiesinė regresija naudojant ORANGE



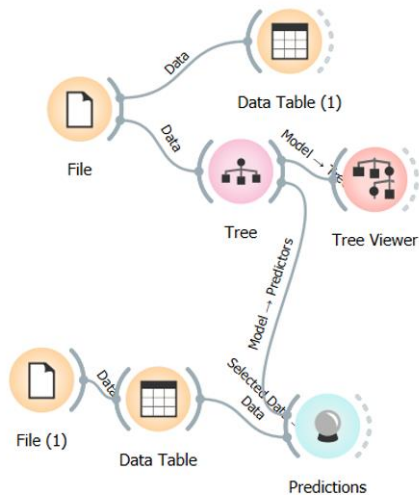
4. Logistinė regresija.

- Kai priklausomas kintamasis yra dvireikšmis, yra naudojama **logistinė regresija**. Kategorinio kintamojo reikšmių prognozavimas tam tikra prasme yra klasifikavimo uždavinys.
- Įkelsime failą Darzoves_vaisiai.tab.
- Prognozavimui įkelsime failą Darzoves_vaisiai_prognosei.tab.

Medžio kūrimas:



Prognozavimas

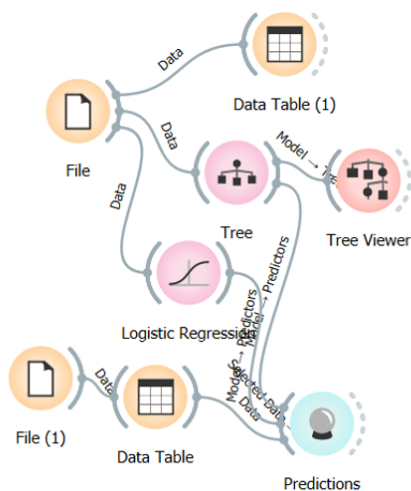


Predictions - Orange

Show probabilities for: Classes known to the model Restore Original Order

	Tree	pavadinimas	vitaminas A %	vitaminas C %	k ₂
1	0.00 : 1.00 → vais...	?	1.0	154.0	3.0
2	0.00 : 1.00 → vais...	?	15.0	300.0	2.0
3	0.00 : 1.00 → vais...	?	0.0	43.0	2.0

Logistinės regresijos ir klasifikavimo medžio prognozių palyginimas



Predictions - Orange

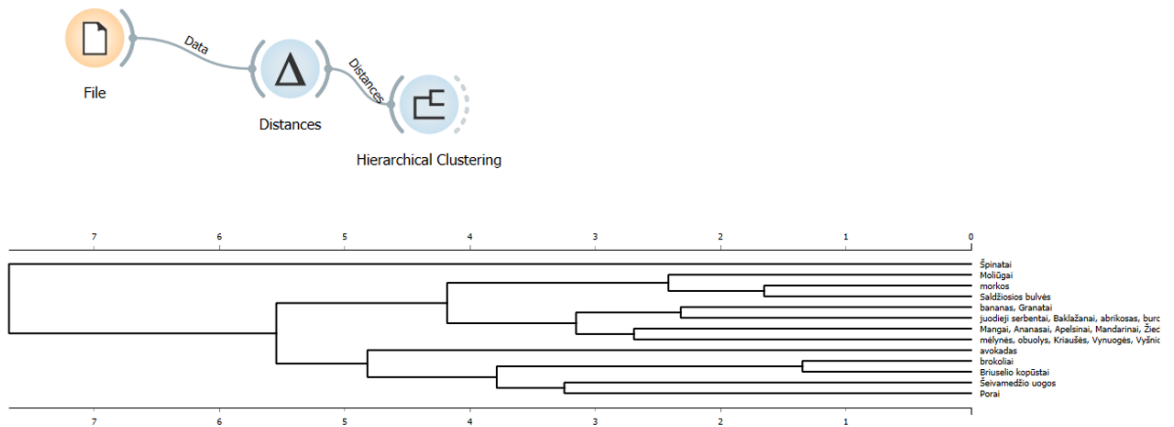
Show probabilities for: Classes known to the model Restore Original Order

	Tree	Logistic Regression	pavadinimas	vitaminas A %	vitami
1	0.00 : 1.00 → vais...	0.01 : 0.99 → vaisius	?	1.0	154.0
2	0.00 : 1.00 → vais...	0.98 : 0.02 → darž...	?	15.0	300.0
3	0.00 : 1.00 → vais...	0.01 : 0.99 → vaisius	?	0.0	43.0

5. Klasterinė analizė

- Klasteris – panašių objektų grupė.
- Klasterinės analizės tikslas – suskirstyti objektus taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.
- Klasterinė analizė – egzistuojančių struktūrų paieška.
- Rezultatus interpretuoja tyrėjas.
- Hierarchinis klasterizavimas – visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos – dar mažesni ir t. t.
- Taikant hierarchinius metodus nustatoma klasterių tarpusavio priklausomybių struktūra ir po to sprendžiama, koks klasterių skaičius optimalus.

- g. Hierarchiniai metodai skirstomi į jungimo ir skaidymo metodus. Jungimo metodai smulkius klasterius jungia į stambesnius. Skaidymo metodai yra loginė jungimo metodų priešingybė.
- h. Klasterizavimo rezultatai vaizduojami dendogramomis.
- i. Plačiau apie klasterizavimą: V. Čekanavičius, G. Murauskas. Statistika ir jos taikymai II. Vilnius, TEV, 2002.



Daugiau:

- [Model evaluation and scoring](#)
- [Add-ons](#)
- [Principal component analysis](#)
- [Feature scoring and ranking](#)
- [k-Means](#)
- [k-Means explained](#)
- [Silhouette](#)
- [Image analytics-clustering](#)
- [Image analytics-classification](#)
- [Text preprocessing](#)
- [Text clustering](#)
- [Text classification](#)
- [How to import text documents](#)
- [Multivariate projection](#)