

Įvadas į dirbtinio intelekto metodus IV gimnazijos klasėje

Renata Burbaitė,

Panevėžio Juozo Balčikonio gimnazijos informatikos mokytoja

Kauno technologijos universiteto dėstytoja

DI tematika III-IV gimnazijos klasėje: pasiekimų raida

(<https://emokykla.lt/bendrosios-programos/visos-bendrosios-programos/3?ach-1=6&ach-2=6&ach-3=6&ach-4=6&ach-5=6&ach-6=6&clases=3677,3658&educations=&st=3&types=5,6&ct=6>)

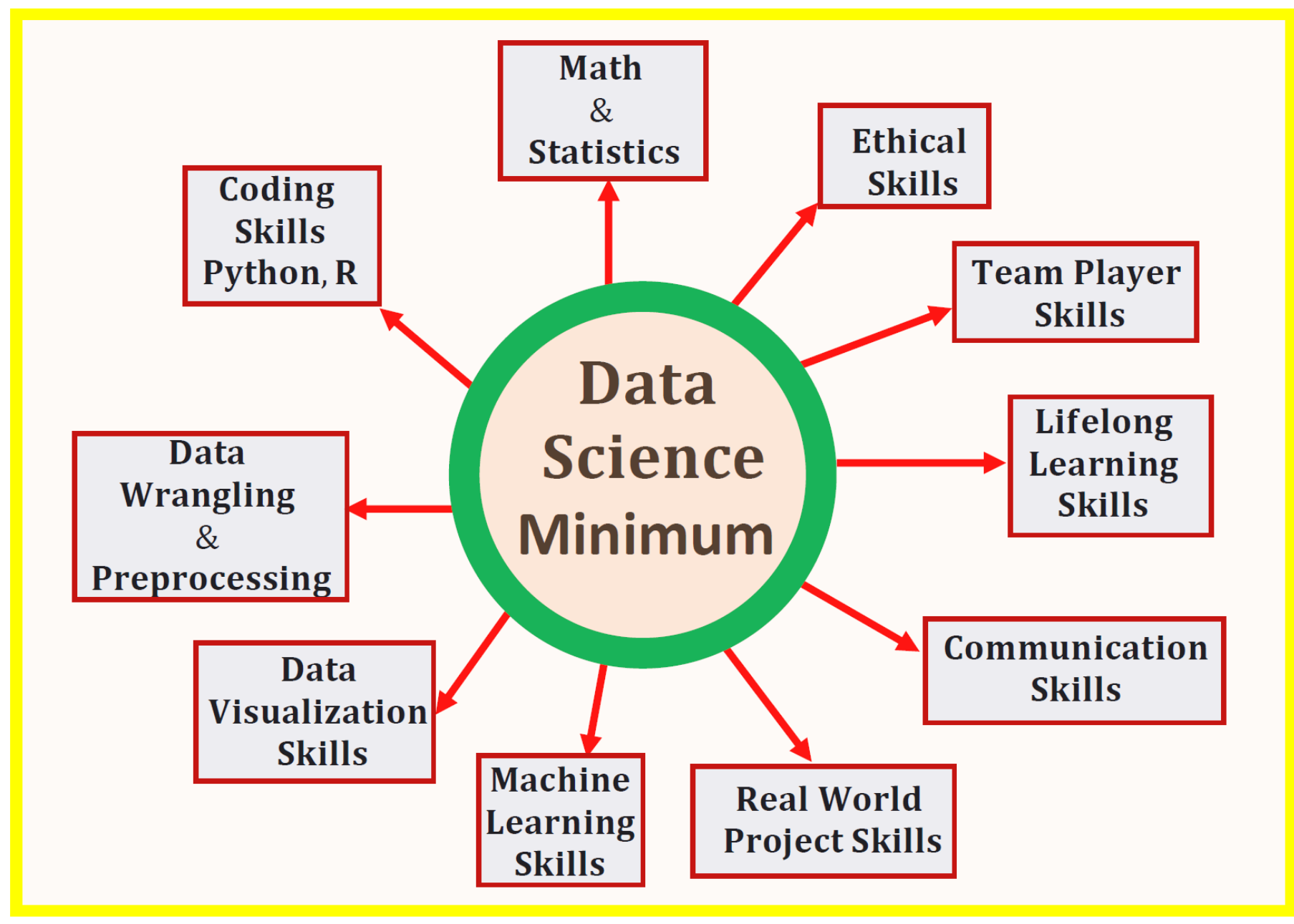
Duomenų tyryba ir informacija (C)

- **C2. Tyrinėja duomenis ir atlieka veiksmus su jais.**
 - Tyrinėja ir apibendrina viešai prieinamus ar automatizuotai renkamus duomenis ir išgauna reikalingą informaciją (C2.3).
- **C3. Vertina duomenų ir informacijos patikimumą, privatumą.**
 - Atpažįsta dirbtinio intelekto, neuroninių tinklų taikymo sritis, vertina privalumus ir grėsmes (C3.3).

DI tematika IV gimnazijos klasėje: mokymo(si)

turinys <https://emokykla.lt/bendrosios-programos/visos-bendrosios-programos/3?ach-1=6&ach-2=6&ach-3=6&ach-4=6&ach-5=6&ach-6=6&clases=3677,3658&educations=&st=3&types=5,7&ct=6>

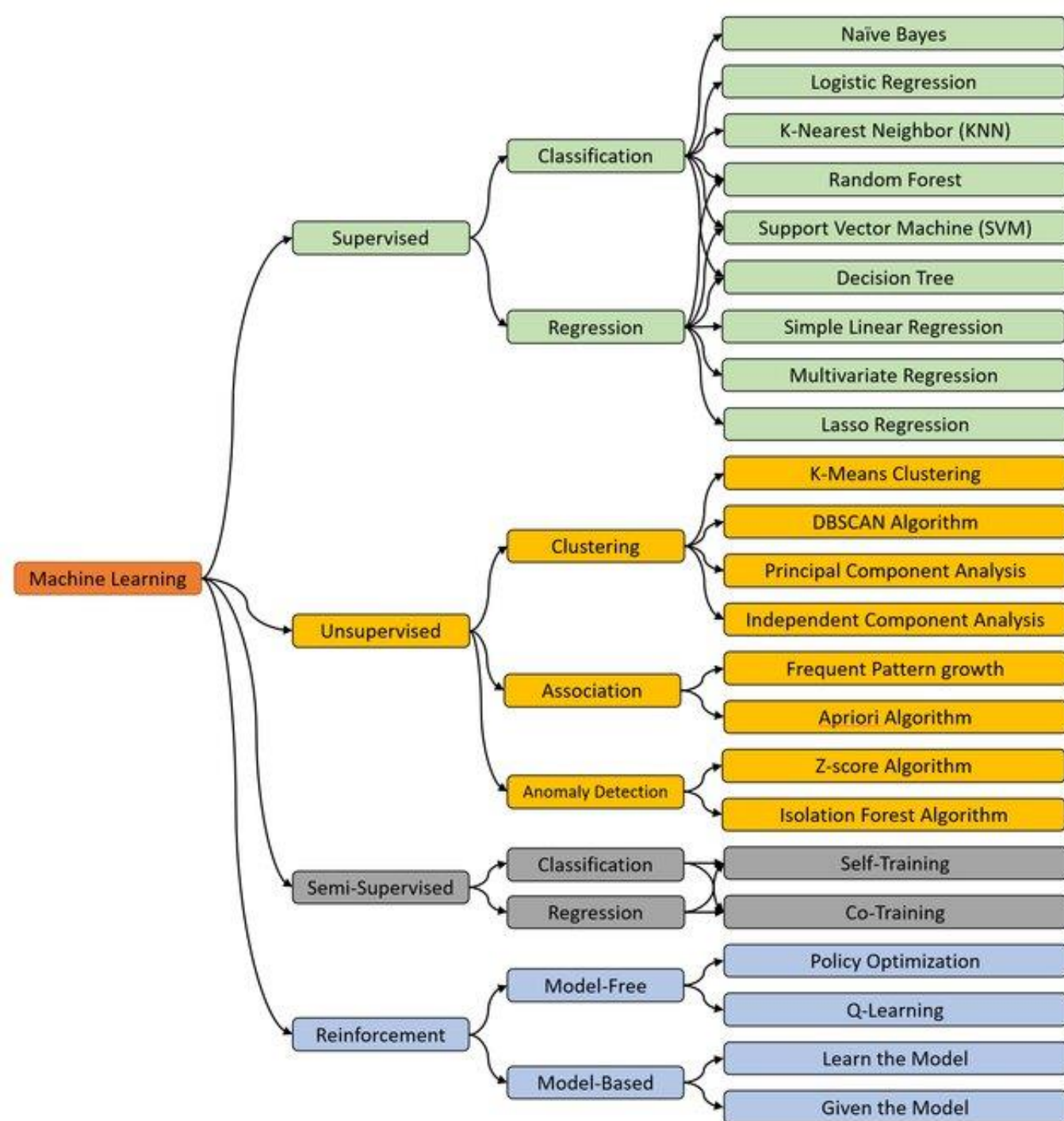
- Nagrinėjami dirbtinio intelekto taikymai duomenims klasifikuoti (pavyzdžiui, gėlių klasifikavimas naudojant „Iris“ duomenų rinkinį), atpažinti (pavyzdžiui, tekstų analizei, kalbos, veido atpažinimui), prognozuoti (pavyzdžiui, oro temperatūrai prognozuoti).
- Aptariamos sprendimų medžio, dirbtinio neuroninio tinklo ir kitos naujausios dirbtinio intelekto technologijos.
- Prisimenama ir gilinamasi, kaip vyksta dirbtinio neuroninio tinklo apmokymas. Galima atlikti įvairius eksperimentus, pavyzdžiui, su *Orange*, *Python* ar kt., naudojant giliajam mokymuisi skirtas *TensorFlow* ir *Keras API* sąsajas.

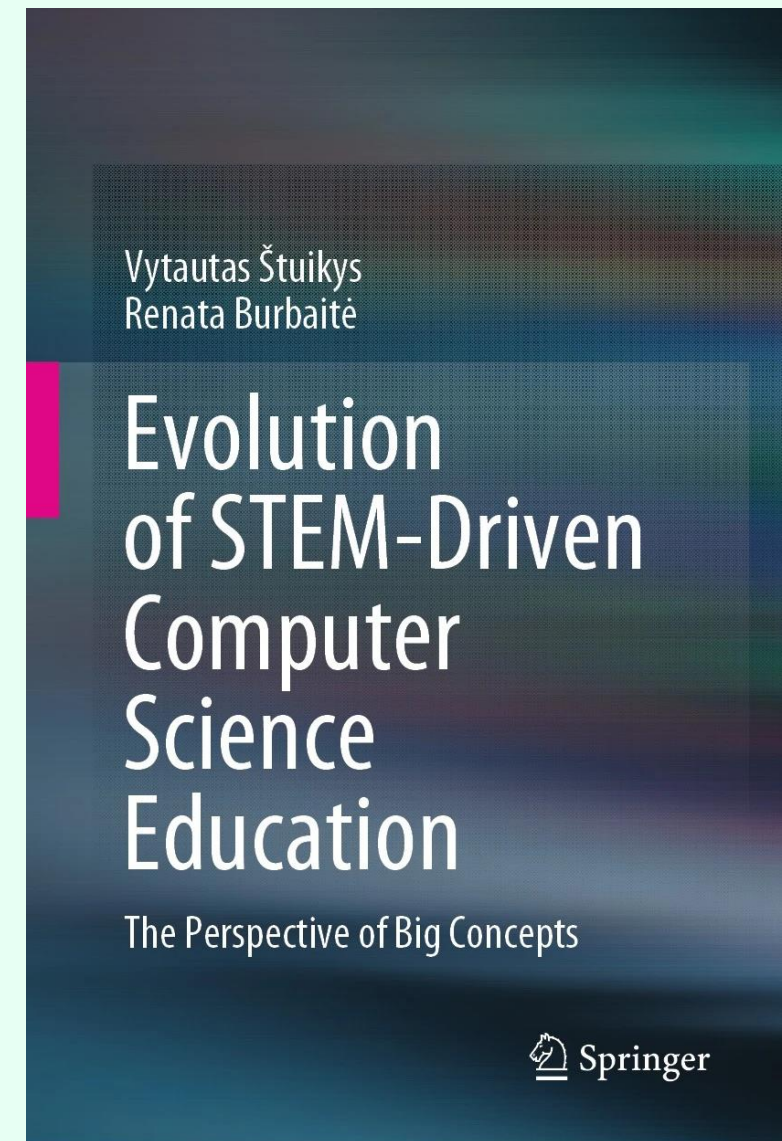
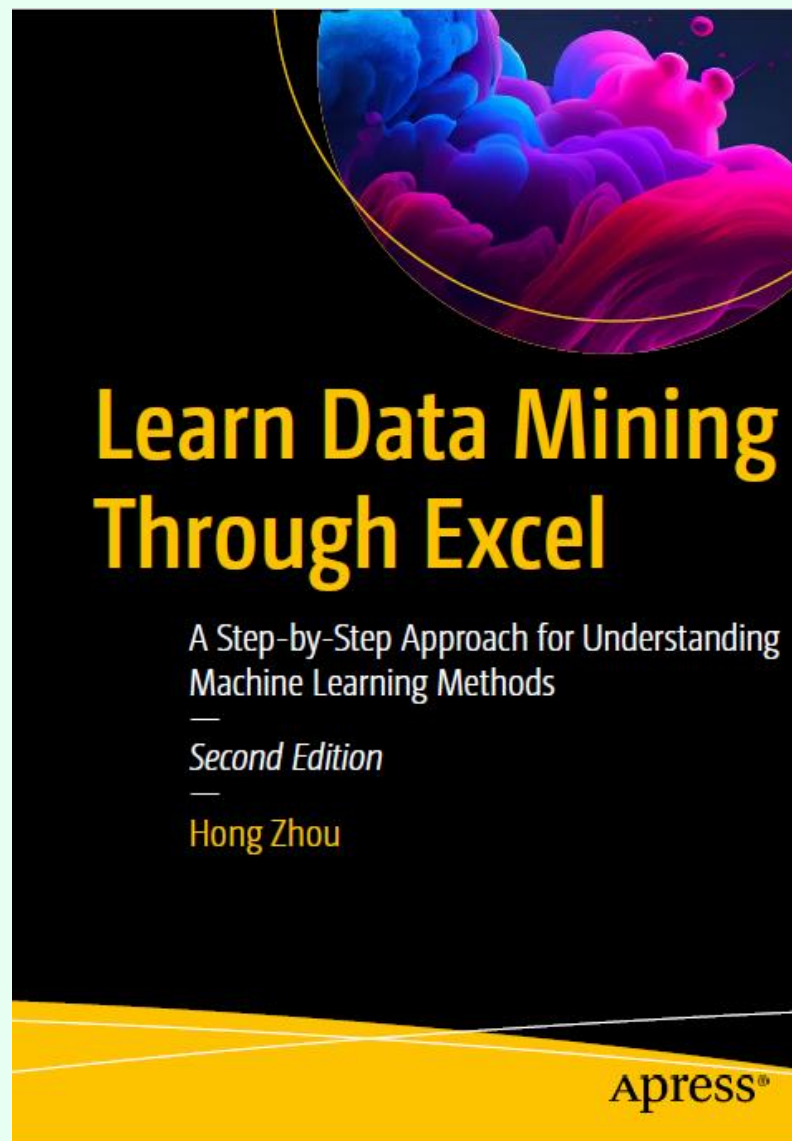
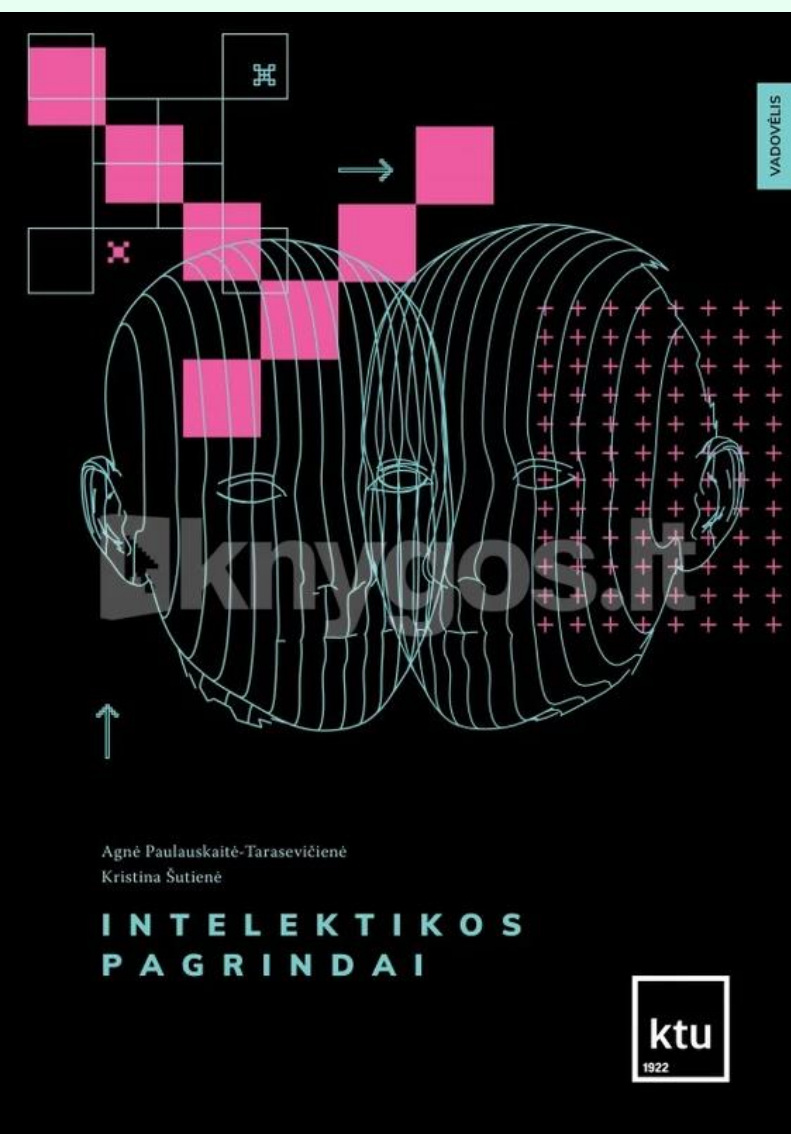


Mašininis mokymasis

(https://www.researchgate.net/publication/373838363_A_Review_of_Machine_Learning-based_Security_in_Cloud_Computing/figures)

- Mašininis mokymasis – dirbtinio intelekto algoritmų klasė, kuriai būdingas ne tiesioginis problemos sprendimas, o mokymasis kaip pritaikyti daugelio panašių problemų sprendimus (šaltinis: A. Paulauskaitė-Tarasevičienė, K. Šutienė „Intelektikos pagrindai“, Kaunas, 2022).





Praktikos darbų struktūra

- Metodo esmė, algoritmas
- Metodo panaudojimo pavyzdžiai Excel
- Metodo panaudojimo pavyzdžiai Orange
- Metodo taikymo sritys, privalumai ir trūkumai
- Savarankiško darbo užduotys

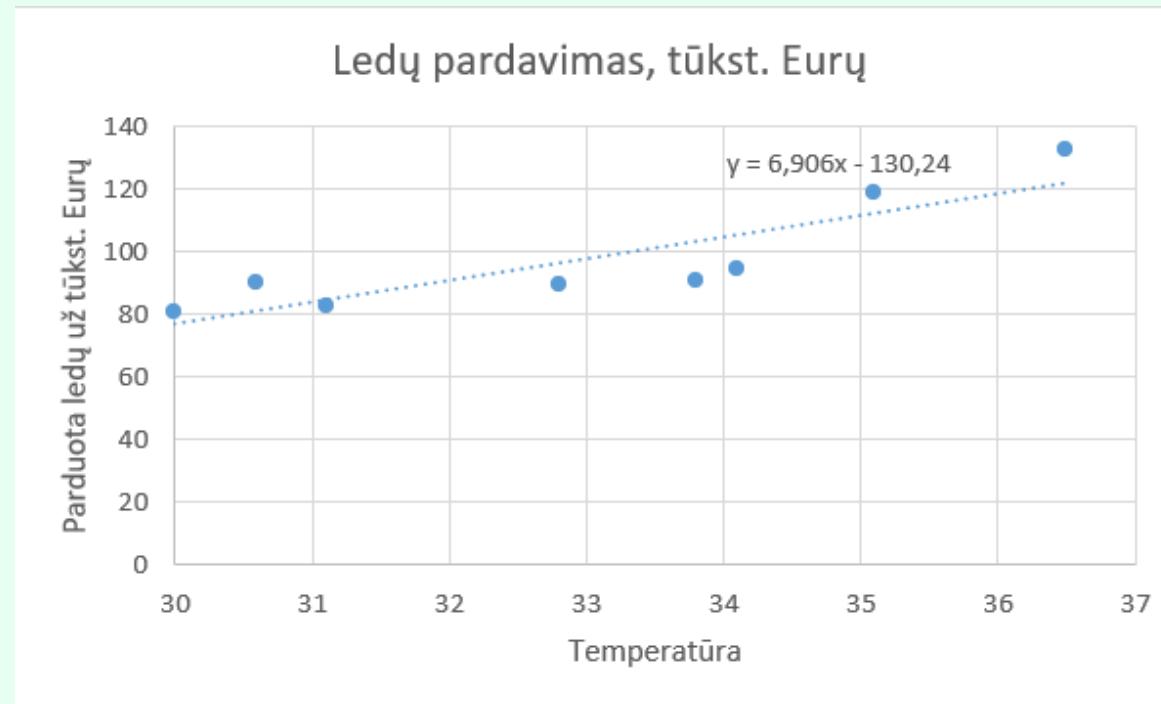
Praktikos darbai, parengti remiantis „Learn Data Mining Through Excel“, Orange dokumentacija

- [Tiesinė regresija](#)
- [Duomenų parengimas analizei](#)
- [Asociacijų analizė](#)
- [Sprendimų medžiai](#)
- [K-vidurkių klasterizacijos metodas](#)
- [Hierarchinė klasterizacija](#)
- [Tiesinė diskriminantinė analizė](#)
- [Logistinė regresija](#)
- [K-artimiausių kaimynų metodas](#)
- [Naivusis Bajeso klasifikatorius](#)
- [Dirbtinis neuroninis tinklas](#)
- [Teksto tyryba](#)

Tiesinė regresija

Tiesinės regresijos modelis

- Regresijos modelis – statistinis modelis, leidžiantis vieno kintamojo reikšmės prognozuoti pagal kito kintamojo reikšmės. Regresija – statistinė vieno atsitiktinio dydžio reikšmių priklausomybė nuo kito – neatsitiktinio – dydžio (arba kelių kitų dydžių), turinti griežtą funkcinį ryšį (t.y., aprašoma lygtimi).



Tiesinės regresijos taikymai

Ekonomika ir finansai:

- **Akcijų rinkos prognozės.** Tiesinė regresija naudojama siekiant prognozuoti akcijų kainas remiantis įvairiais ekonominiais rodikliais, tokiais kaip infliacija, palūkanų normos ar ekonomikos augimas.
- **Kainų nustatymas.** Modeliai gali padėti prognozuoti prekių ar paslaugų kainas, atsižvelgiant į gamybos sąnaudas, paklausą ir pasiūlą.

Sveikatos apsauga:

- **Ligų prognozavimas.** Naudojama analizuoti sveikatos rodiklius, kaip kraujospūdį, cukraus kiekį kraujyje, ir numatyti ligų riziką, pavyzdžiui, širdies ligas ar diabetą.
- **Medikamentų efektyvumas.** Gali būti naudojama vertinant, kokią įtaką tam tikri veiksniai (pvz., dozė, paciento amžius) turi gydymo rezultatams.

Rinkodara ir pardavimai:

- **Reklamos efektyvumas.** Naudojama analizuoti reklamos išlaidų poveikį pardavimams ir optimizuoti marketingo kampanijas.
- **Klientų elgsenos analizė.** Gali būti taikoma siekiant suprasti, kokią įtaką skirtingi veiksniai (pvz., pajamos, amžius) turi klientų pasirinkimams.

Socialiniai mokslai:

- **Demografiniai tyrimai.** Tiesinė regresija naudojama siekiant nustatyti, kaip įvairūs demografiniai veiksniai (pvz., amžius, pajamos, išsilavinimas) daro įtaką tokiems rodikliams kaip nedarbo lygis ar šeimos dydis.
- **Psichologiniai tyrimai.** Naudojama analizuoti sąsajas tarp psichologinių kintamųjų, pvz., kaip stresas veikia darbo našumą.

Inžinerija ir gamyba:

- **Kokybės kontrolė.** Gali būti naudojama prognozuoti gaminių kokybę ar defektus atsižvelgiant į gamybos parametrus.
- **Išteklių optimizavimas.** Inžinieriai naudoja regresijos modelius optimizuojant gamybos procesus ir sumažinant sąnaudas.

Aplinkosauga ir klimato kaita:

- **Klimato pokyčiai.** Tiesinė regresija naudojama modeliuojant klimato duomenis, pvz., analizuojant temperatūros pokyčių įtaką ledynų tirpimui ar jūros lygio kilimui.
- **Teršalų analizė.** Gali būti naudojama analizuoti, kokia tam tikrų veiksmų įtaka (pvz., gamyklų skaičius, transporto priemonių kiekis) oro taršai.

Sportas:

- **Sporto rezultatų prognozavimas.** Tiesinė regresija gali būti naudojama numatyti komandos ar individualaus žaidėjo pasirodymus, remiantis tokiais duomenimis kaip ankstesni rezultatai, sužaidžiamų rungtynių skaičius, komandos sudėtis.
- **Žaidėjų vertinimas.** Naudojama siekiant nustatyti, kaip tam tikri veiksniai, pvz., treniruočių intensyvumas ar žaidėjo amžius, gali paveikti žaidėjų meistriškumą.

Nekilnojamojo turto vertinimas:

- **Nekilnojamojo turto kainos.** Tiesinė regresija dažnai naudojama siekiant nustatyti, kaip veiksniai, pvz., vieta, būsto dydis, metų skaičius nuo statybos, veikia nekilnojamojo turto kainas.

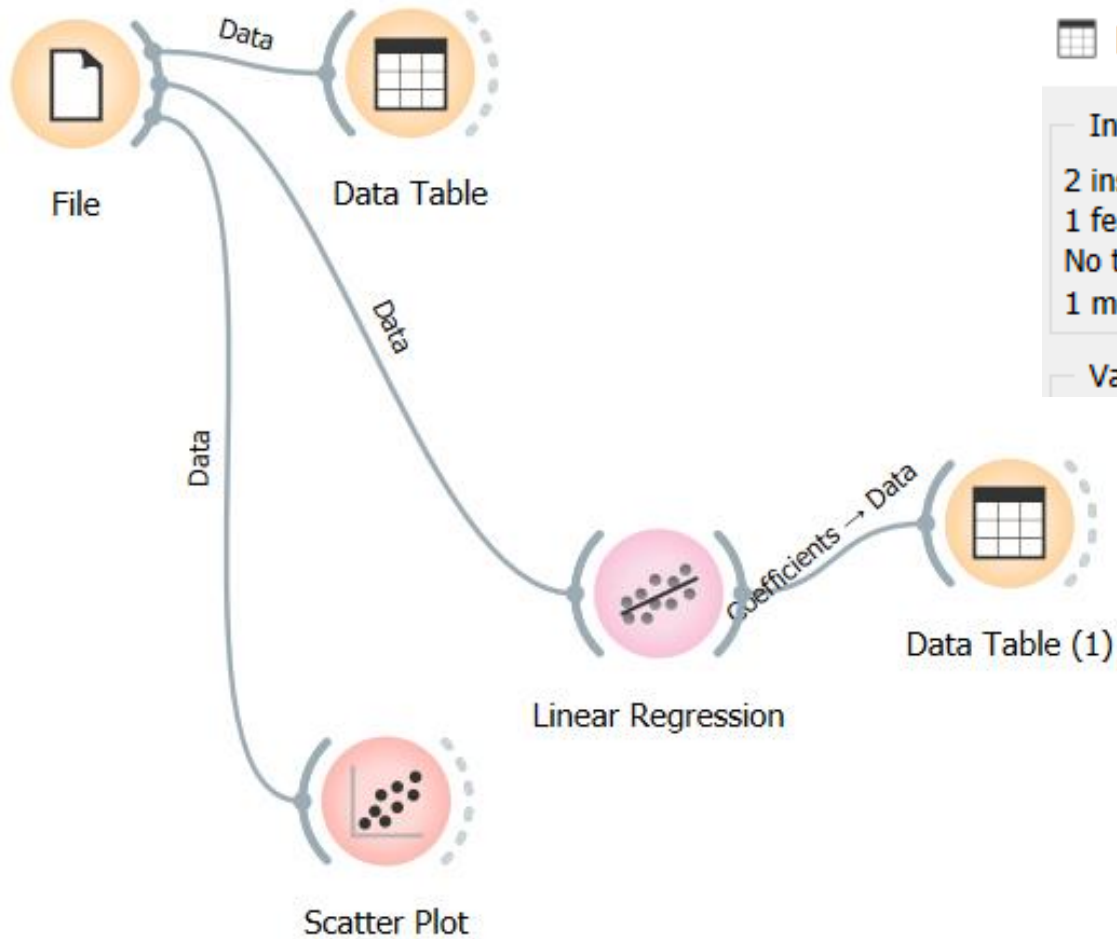
Transportas ir logistika:

- **Kuro suvartojimo prognozavimas.** Analizuoja, kaip įvairūs veiksniai, pvz., kelio būklė, transporto priemonės amžius, paveikia degalų suvartojimą.
- **Eismo prognozės.** Naudojama numatant eismo intensyvumą ir optimizuojant transporto sistemas.

Daugialypė tiesinė regresija

	A	B	C	D	E	F
1	Temperatūra, °C	Turistų skaičius	Saulėtų dienų per savaitę skaičius	Ledų pardavimas, tūkst. Eurų	Prognozė	Paklaidos
2	32,8	998	4	89,8	88,773508	1,0536855
3	30,6	1256	7	90,2	90,412367	0,0450996
4	30	791	6	81,1	81,092432	5,728E-05
5	31,1	705	5	83	83,187129	0,0350173
6	33,8	1089	3	90,9	89,760346	1,298812
7	35,1	1135	6	119	117,07283	3,7139981
8	34,1	1076	4	94,9	97,966192	9,401534
9	36,5	1198	7	132,4	133,0352	0,4034789
10						
11	Temperatūra	1	6,016574579	Paklaidų suma	15,951683	
12	Turistų skaičius	2	-0,002189783			
13	Saulėtų dienų per savaitę skaičius	3	7,202682329			
14	y-intercept	4	-169,3553671			
15						

Tiesinās regresijas realizavimas Orange



Data Table (1) - Orange

Info

2 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

	name	coef
1	intercept	-130.236
2	Temperatūra, 0C	6.90601

Scatter Plot - Orange

Attributes

Color: N Ledu pardavimas, tūkst. Ei

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

Symbol size:

Opacity:

Jittering:

☐ Jitter numeric values

☐ Show color regions

☒ Show legend

☒ Show gridlines

☒ Show all data on mouse hover

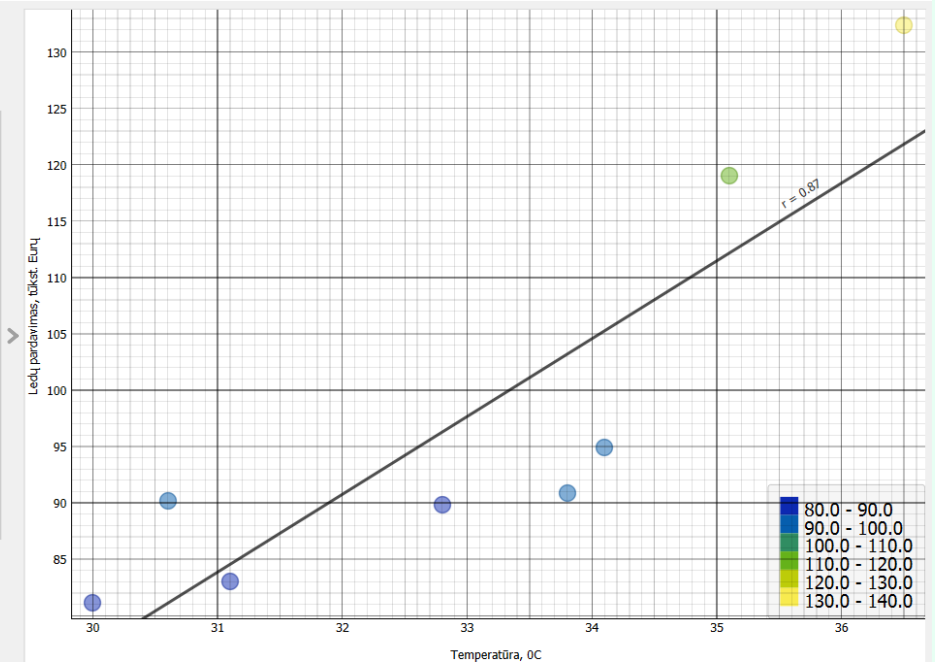
☒ Show regression line

☐ Treat variables as independent

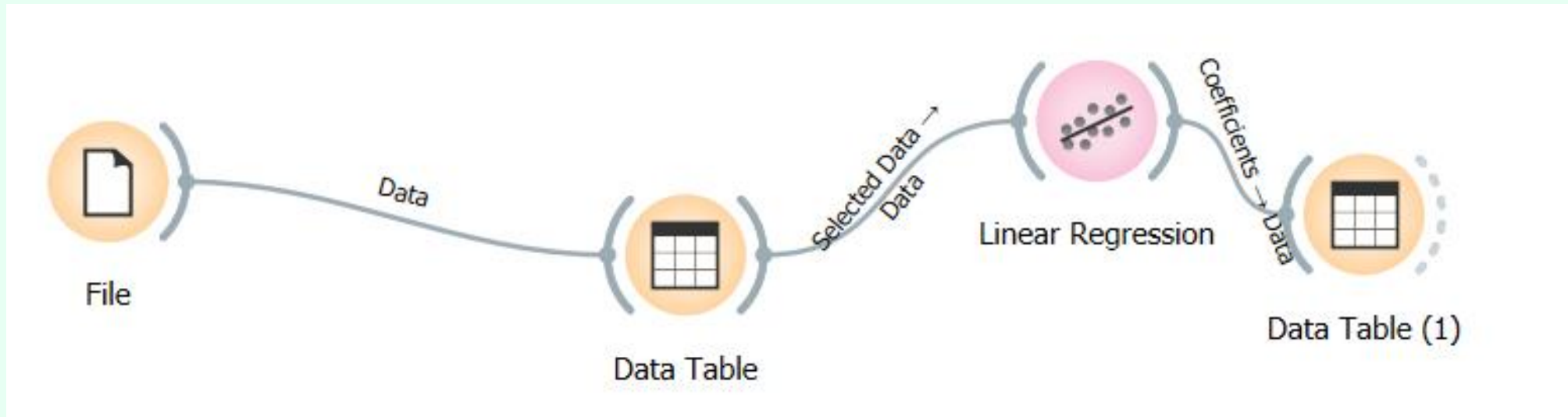
☐ Show confidence ellipse

Zoom/Select

☐ Send Selection



Daugialypės tiesinės regresijos realizavimas Orange



Data Table (1) - Orange

Info

4 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

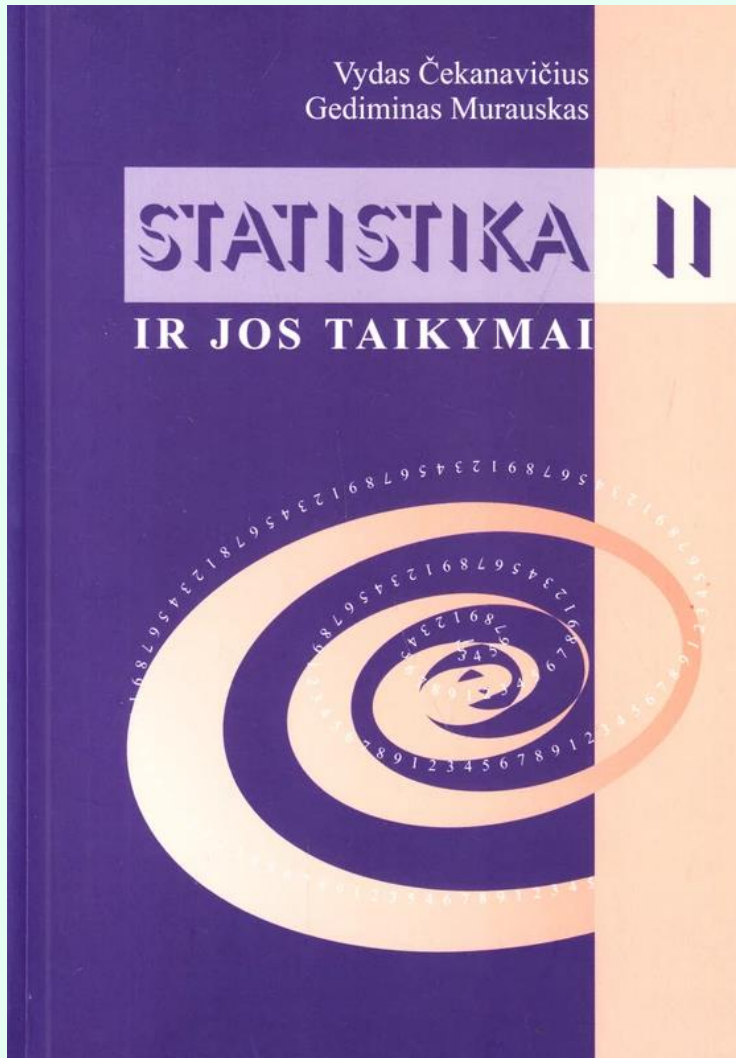
☒ Show variable labels (if present)

	name	coef
1	intercept	-169.355
2	Temperatūra, 0C	7.20268
3	Turistų skaičius	-0.00218978
4	Saulėtų dienų p...	6.01657

Paklaidų priežastys atliekant tiesinę regresiją

- Kartais regresijos modelis neapima visų reikšmingų kintamųjų, kurie gali turėti įtakos priklausomam kintamajam. Pvz., jei prognozuojame būsto kainą, bet neįtraukiame kintamųjų, tokių kaip kriminogeninė padėtis ar mokyklų kokybė, modelio tikslumas gali sumažėti, o prognozavimo paklaidos padidėti. Jei reikšmingi veiksniai neįtraukti, modelis negali tiksliai prognozuoti rezultatų.
- Jei realus duomenų ryšys nėra tiesinis, bet pasirenkamas tiesinis modelis, prognozės bus netikslios. Tokiais atvejais reikalingi sudėtingesni modeliai (pvz., polinominė regresija), kurie geriau atspindėtų realų duomenų ryšį.
- Kai kurie kintamieji gali būti neteisingai išmatuoti ar užfiksuoti. Pvz., klaidos gali kilti dėl prietaisų netikslumų, subjektyvių vertinimų ar duomenų įvedimo klaidų. Šios klaidos tiesiogiai padidina paklaidas ir sumažina modelio tikslumą.
- Paklaidos gali atsirasti dėl veiksnių, kurie negali būti tiksliai numatyti arba išmatuoti, tačiau daro įtaką priklausomam kintamajam. Pvz., ekonominiai sukrėtimai, gamtos nelaimės ar netikėti visuomeniniai pokyčiai gali sukelti nenumatytas pokyčių tendencijas, kurios neatitinka modelio.
- Duomenys natūraliai kinta, todėl neišvengiama, kad atsiranda tam tikra atsitiktinė paklaida, net jei modelis gerai atspindi pagrindinę tendenciją.
- Tiesinės regresijos modelis daro prielaidą, kad paklaidų dispersija yra pastovi visoms nepriklausomų kintamųjų reikšmėms. Jei paklaidos didėja arba mažėja priklausomai nuo kintamųjų reikšmių, modelio tikslumas gali sumažėti.
- Jei nepriklausomi kintamieji yra stipriai susiję tarpusavyje, regresijos koeficientai tampa nepatikimi. Tai gali sukelti didesnes paklaidas prognozuojant rezultatus ir sunkinti modelio interpretaciją.
- Jei duomenų rinkinyje yra ekstremalių arba išskirtinių taškų (anomalijų), jie gali neproporcingai paveikti tiesinę regresiją, todėl bendros tendencijos tampa klaidingos. Šiuos taškus būtina identifikuoti ir įvertinti, ar tai klaidingi duomenys, ar anomalija.

Savarankiško darbo užduotys



- 1-7 užduotys – tiesinė regresija
- 8-10 užduotys – daugialypė tiesinė regresija
- [Savarankiško darbo užduočių rinkinys](#)

Duomenų parengimas analizei

Duomenų parengimas (1)

- Įrašų skaičius
- Atributų skaičius
- Kiekvieno atributo duomenų tipas
- Kiekvieno skaitinio požymio suma, vidurkis, mediana ir dispersija
- Kiekvieno skaitinio požymio verčių diapazonas
- Kategorinio tikslo kintamojo ir atributų unikalios reikšmės
- Tikslo kintamojo verčių pasiskirstymas klasėje
- Trūkstanti duomenys
- Duomenų anomalijos ir kt.

Duomenų parengimas (2)

- Duomenų valymas
 - panaudojant skaičiuoklės funkcijas,
 - uždedant filtrus.
- Atributų pasirinkimas
 - atributų pasirinkimo principai:
 - atributai, kurių reikšmės yra toje pačioje skalėje, veikia geriau,
 - atributai su didesne dispersija turi didesnę įtaką,
 - susiję atributai pateikia tą pačią informaciją apie duomenis.
 - duomenų transformavimas į vieną skalę (normalizavimas ir standartizavimas),
 - koreliacijos matrica.

Savarankiško darbo užduotys

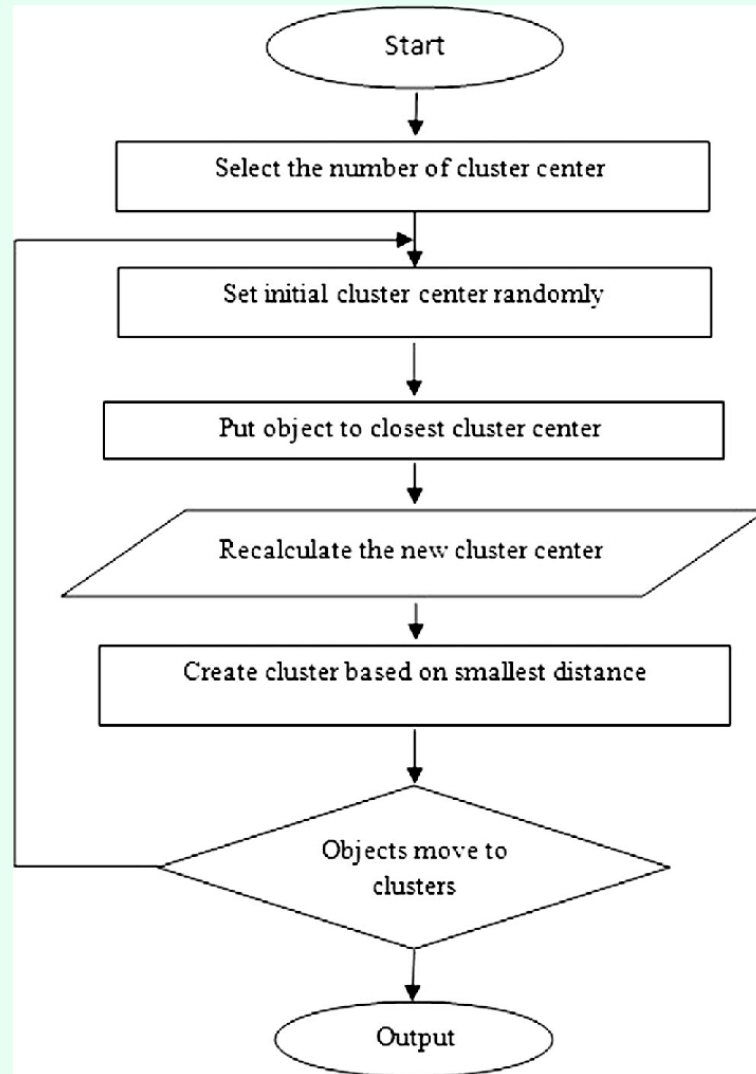
- Įrašų, atributų skaičiaus, atributų duomenų tipų, skaitinių atributų pagrindinių statistinių charakteristikų skaičiavimas, kategorinių duomenų unikalių verčių nustatymas, tikslo kintamojo verčių pasiskirstymas.
- Netinkamų duomenų pašalinimas.
- Kategorinių verčių kodavimas.
- Koreliacijos matricos formavimas ir atributų pasirinkimas pirminiam analizės etapui.
- [Savarankiško darbo užduočių failas.](#)

K-vidurkių klasterizacija

Metodo esmė

- Klasterizacija yra klasifikavimo metodas, kurį taikant duomenys yra suskirstomi į skirtingas grupes (klasterius), kurių kiekvienas turi daugybę būdingų savybių.
- Klasterizacija yra neprižiūrimasis duomenų tyrybos metodas. Klasterizuojant nereikia mokymo duomenų rinkinio. Du populiariausi klasterizacijos metodų tipai yra skaidymo ir hierarchinė klasterizacija. K-vidurkių klasterizacija, kur k reiškia norimą klasterių skaičių, yra skaidymo klasterizacijos tipas. Kiekvienas klasteris apibrėžiamas klasterio duomenų taškų centru (arba vidurkiu).
- K-vidurkių klasterizacijos metodas reikalauja, kad visi duomenys būtų skaitiniai.

Algoritmas https://www.researchgate.net/figure/Flowchart-of-k-means-clustering-algorithm_fig2_271915066



Pagrindinės taikymo sritys

- **Rinkodara ir klientų segmentavimas.** Naudojamas segmentuoti klientus pagal jų elgesį, pirkimo įpročius ar kitus rodiklius, siekiant sukurti tikslines rinkodaros kampanijas.
- **Vaizdų apdorojimas.** Naudojamas spalvų kvantavimui arba segmentavimui, kad būtų galima suskirstyti vaizdus į skirtingas spalvų zonas ar objektus.
- **Biologija ir medicina.** Naudojamas genetinių duomenų, biologinių stebėjimų grupavimui ir medicininių vaizdų segmentavimui (pvz., aptinkant anomalijas arba skirtingus audinius).
- **Finansų analizė.** Naudojamas nustatyti investicijų grupes, finansines anomalijas arba segmentuoti akcijų rinkos tendencijas.
- **Rekomendacijų sistemos.** Naudojamas vartotojų elgsenos segmentavimui, siekiant personalizuoti rekomendacijas (pvz., filmuose, muzikos platformose).
- **Tiekimo grandinių valdymas.** Naudojamas optimizuoti sandėlių išdėstymą, klientų aptarnavimo zonų formavimą arba tiekimo tinklų planavimą.
- **Socialinių tinklų analizė.** Naudojamas analizuoti vartotojų grupes pagal jų sąveiką socialiniuose tinkluose arba aptikti bendruomenes.
- **Geografinė analizė ir žemėlapių kūrimas.** Naudojamas suskirstyti geografines zonas pagal tam tikrus rodiklius, pvz., gyventojų tankį ar vietovių charakteristikas.

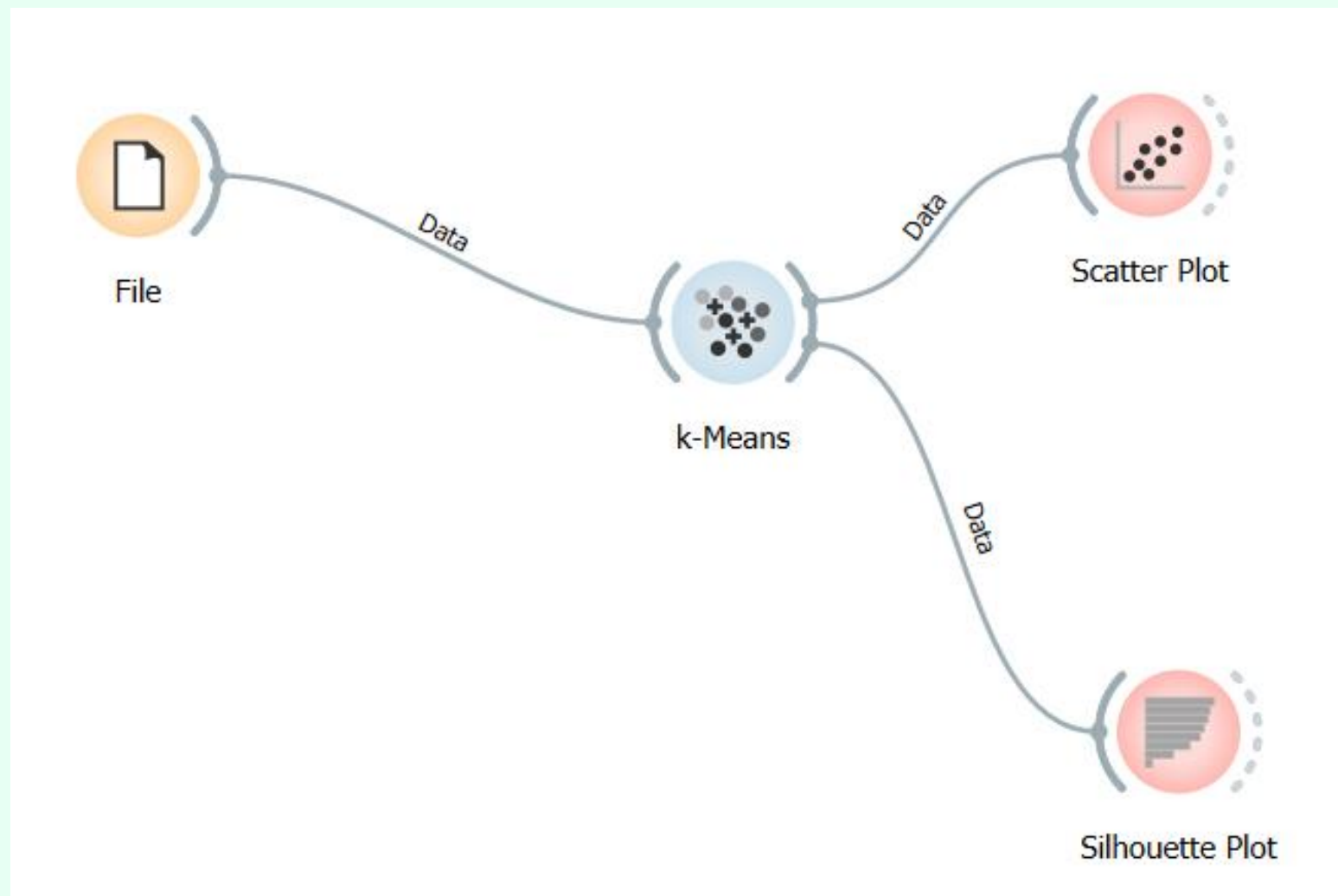
Mokomasis pavyzdys (1)

	A	B	C	D	E	
1						
2	Šalis	Nusikaltimai_1	Nusikaltimai_2	Gyventojų skaičius	Nusikaltimai_3	
3	L1	12	193	51	21	
4	L2	9	222	39	42	
5	L3	6	226	72	30	
6	L4	8	177	50	17	
7	L5	7	235	86	39	
8	L6	5	170	78	37	
9	L7	0	35	75	7	
10	L8	4	198	70	15	
11	L9	13	259	74	30	
12	L10	15	136	52	24	
13	L11	5	22	22	17	

Mokomasis pavyzdys (2)

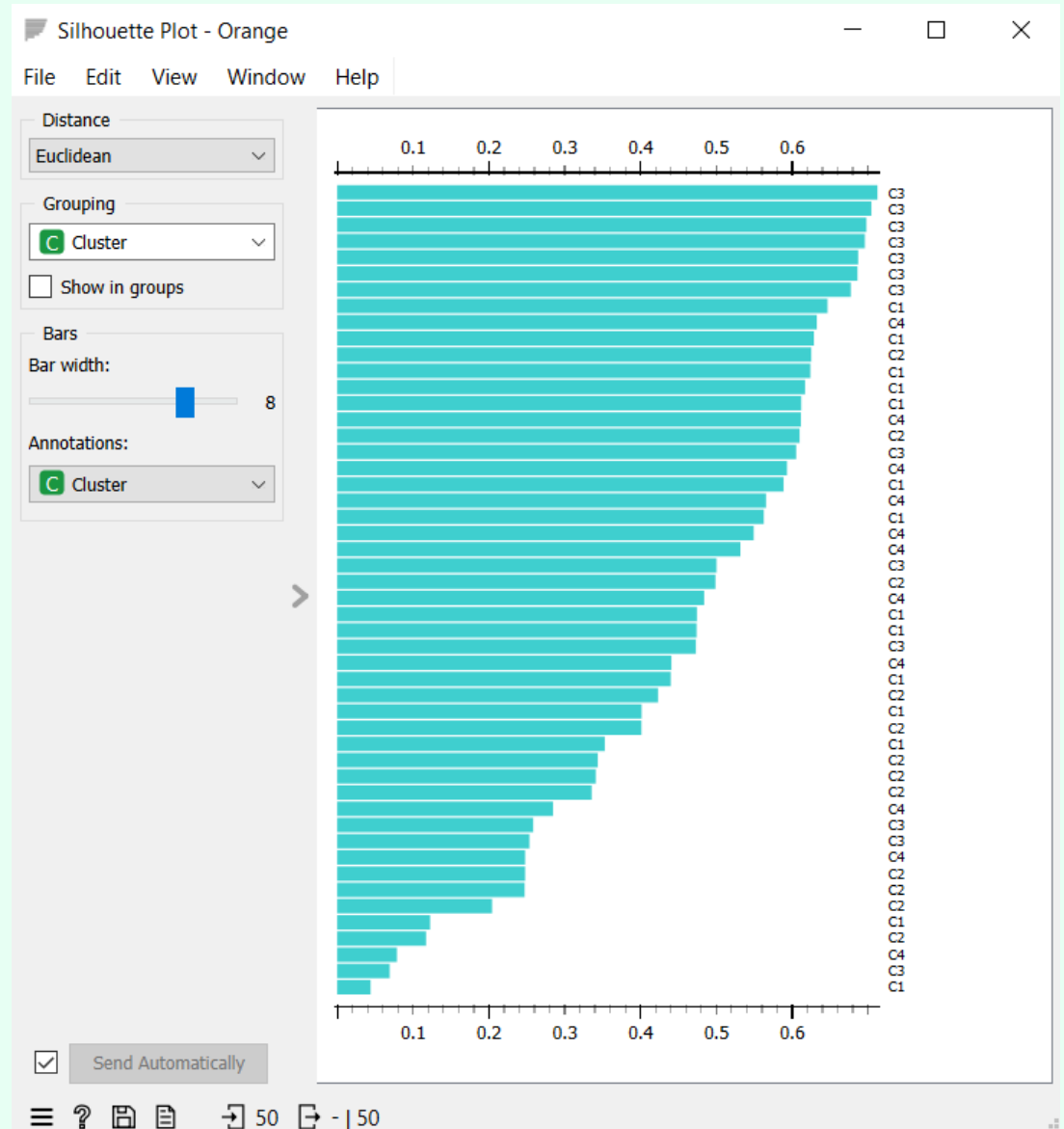
F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
5	Vidurkis						Atstumai					1		
	Klasteris	K1	K2	K3	K4		K1	K2	K3	K4	Klasteris	2	Senas klasteris	Skirtumas
2	Nusikaltimai_1	2,31	3,009091	7,8	10,28571		184,88133	136,48940	61,63181	42,68943	K4	3	K4	0
3	Nusikaltimai_2	8,64	57,09091	133,8	233,5714		216,10378	168,19538	95,25692	30,12762	K4	4	K4	0
4	Gyventojų skaičius	52,6	57,81818	67,6	62,57143		219,06343	170,15264	92,84115	13,11332	K4	5	K4	0
5	Nusikaltimai_3	11,2	15,45455	20,2	27,28571		168,57599	120,27723	46,75767	58,90151	K3	6	K3	0
							230,54120	181,70354	104,56615	26,43784	K4	7	K4	0
							165,39373	116,72163	41,33618	66,34511	K3	8	K3	0
							34,92257	29,38974	100,25607	200,25581	K2	9	K2	0
	Konvergencija		0				190,20322	141,43888	64,56686	38,87106	K4	10	K4	0
							252,20191	203,32368	125,85341	28,14177	K4	11	K4	0
							128,63050	80,48169	17,73358	98,31052	K3	12	K3	0
							38,09102	34,93803	102,90131	202,83551	K2	13	K2	0
							45,76970	12,14806	83,18942	181,18448	K2	14	K2	0
							211,34130	162,51135	84,99694	23,78089	K4	15	K4	0

K-vidurkių klasterizacijos realizavimas Orange

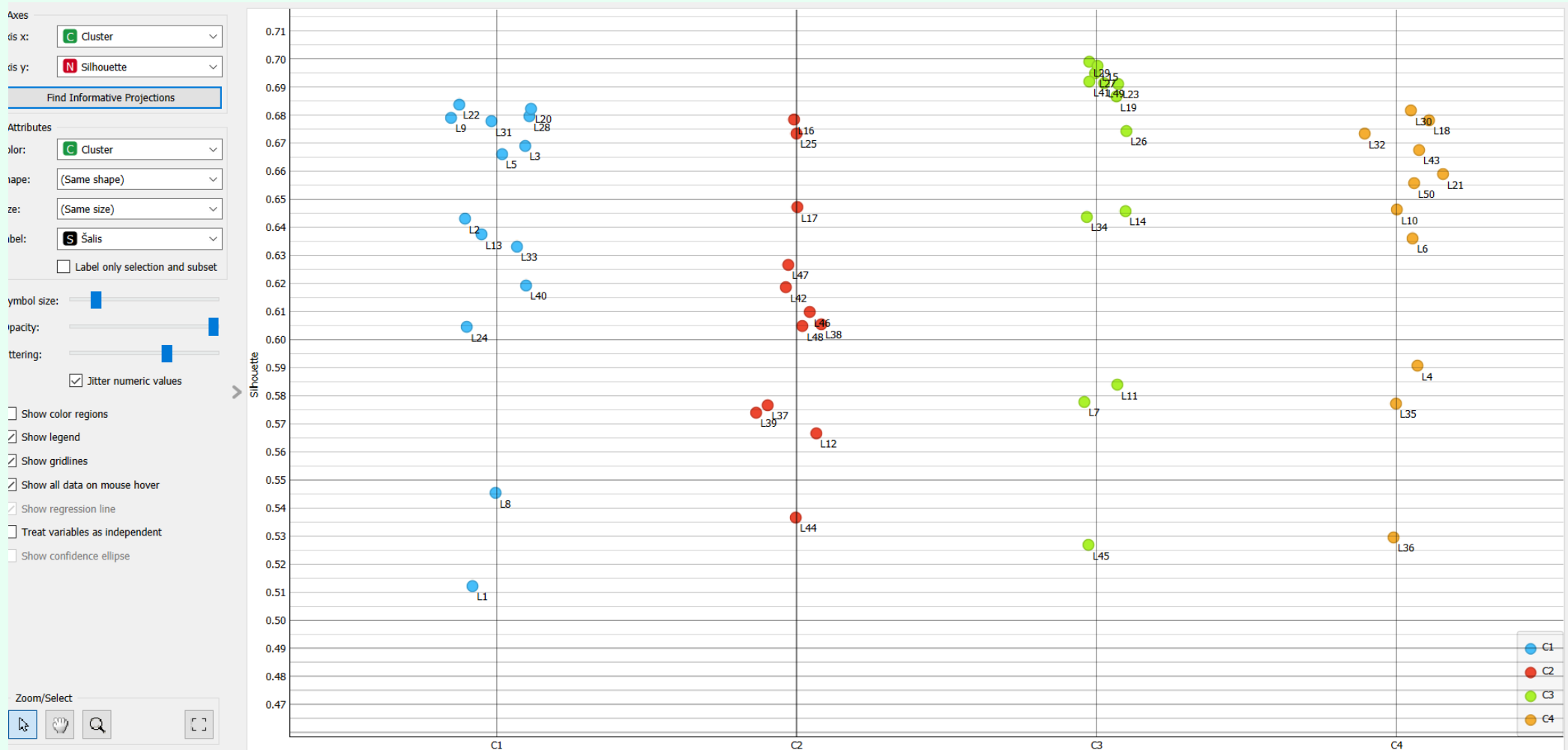


Rezultatų vizualizacija Silhouette Plot

- Silhouette Plot leidžia vizualiai įvertinti klasterio kokybę. Šioje vizualizacijoje rodomas vidutinis atstumas tarp egzempliorių klasteryje ir egzempliorių artimiausiame klasteryje. Tam tikro duomenų egzemplioriaus siluetas, artimas 1, rodo, kad duomenų egzempliorius yra arti klasterio centro. Atvejai, kurių silueto balai artimi 0, yra dviejų grupių riboje.



Rezultatų vizualizacija Scatter Plot



Savarankiško darbo užduotys

1. Genomas gali būti naudojamas vėžiu sergančių pacientų prognozei. Paprastai genominiis parašais turi daugiau nei 20 genų ar net šimtus genų. Remiantis genų ekspresijos lygiais, galima apytiksliai numatyti pacientų išgyvenamumą, todėl galima daryti prielaidą, kad galima suskirstyti pacientus į skirtingas grupes (skirtingų išgyvenamumo metų ir tikimybių). Tarkime, kad minėti 7 genai gali būti naudojami gaubtinės žarnos vėžiu sergantiems pacientams sugrupuoti. Atlikite k-vidurkių klasterizaciją. Pradiniai duomenys pateikti faile [3Skyrius-1SD.xlsx](#).
2. Faile [3Skyrius-2SD.xlsx](#) pateikti duomenys apie tuberkuliozės atvejus Lietuvoje (duomenys adaptuoti iš <https://data.gov.lt/datasets/1894/>).

Hierarchinė klasterizacija

Metodo esmė

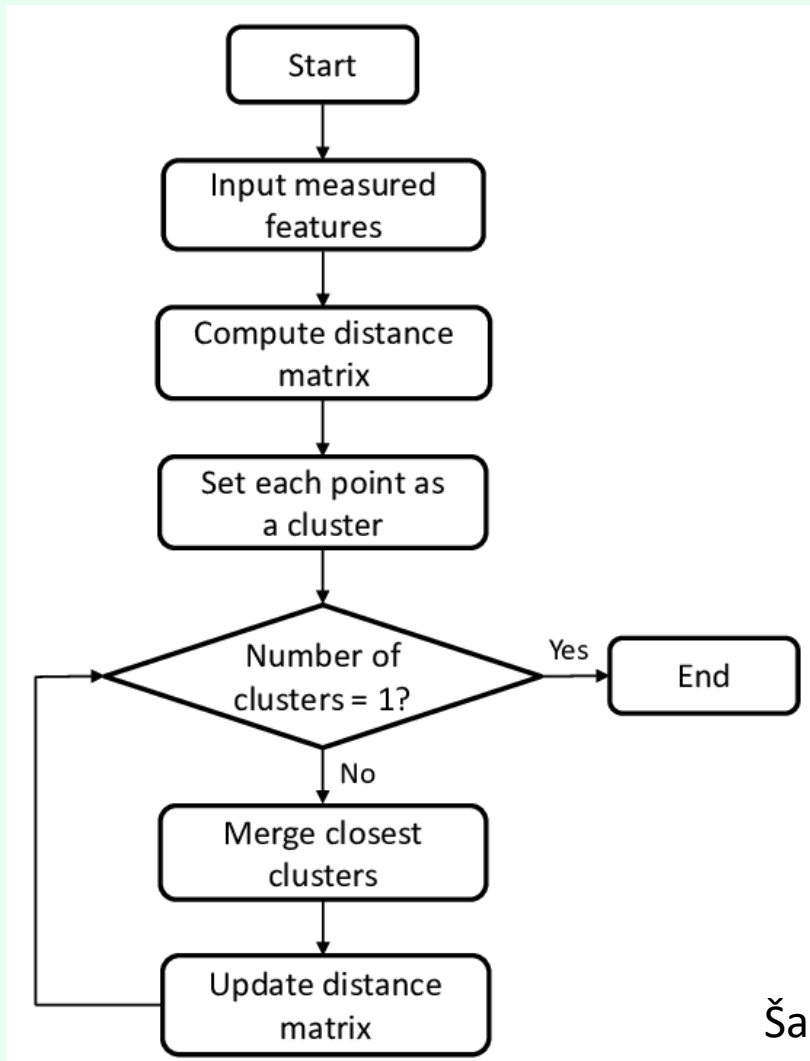
- Neprižiūrimas klasifikavimo metodas, suskirstantis subjektus (duomenų taškus) į skirtingas grupes (klasterius), remiantis tam tikrų savybių panašumu.
- Hierarchinė klasterizacija sukuria klasterių medį. Galutinis hierarchinės klasterizacijos produktas paprastai pateikiamas dendrograma, kuri atvaizduoja santykinius atstumus nuo pradinių duomenų taškų arba sugrupuotų duomenų taškų.
- Hierarchinis grupavimas prasideda nuo artumo matricos ir niekada nereikia iš anksto nustatyti grupių skaičiaus.

Hierarchinės klasterizacijos algoritmas (1)

Hierarchinės klasterizacijos algoritmą galima paaiškinti pavyzdžiu, kuriame yra šeši duomenų taškai A, B, C, D, E ir F.

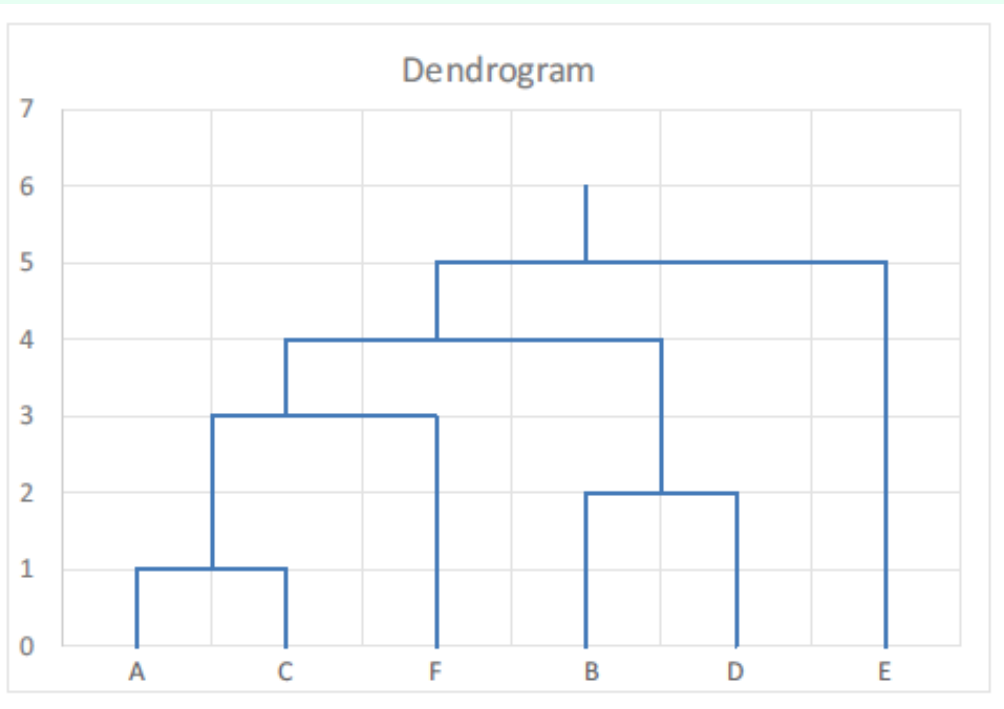
1. Kiekvienas duomenų taškas iš pradžių traktuojamas kaip klasteris.
2. Sukuriama artumo matrica, rodanti atstumus tarp šešių grupių.
3. Tarkime, kad mažiausias atstumas yra tarp A ir C, tada A ir C pirmajame rate dedami į tą patį klasterį ir vaizduojami kaip AC.
4. Dabar yra tik penkios grupės: AC, B, D, E ir F.
5. Pakartokite 2 veiksmą, kad sukurtumėte penkių grupių artumo matricą. Tarkime, kad tarp B ir D yra mažiausias atstumas, tada yra tik keturi klasteriai: AC, BD, E ir F.
6. Pakartokite 2 veiksmą, kad sukurtumėte keturių grupių artumo matricą. Tarkime, kad tarp AC ir F yra mažiausias atstumas, tada yra tik trys klasteriai: ACF, BD ir E.
7. Pakartokite 2 veiksmą, kad sukurtumėte trijų grupių artumo matricą. Tarkime, kad tarp ACF ir BD yra mažiausias atstumas, tada yra tik du klasteriai: ACFBD ir E.
8. Hierarchinis klasterizavimas baigiamas, kai susiformuoja vienas klasteris: ACFBDE.

Hierarchinės klasterizacijos algoritmas (2)



Šaltinis: https://www.researchgate.net/figure/Flow-chart-of-agglomerative-hierarchical-clustering_fig1_313238175

Dendrograma

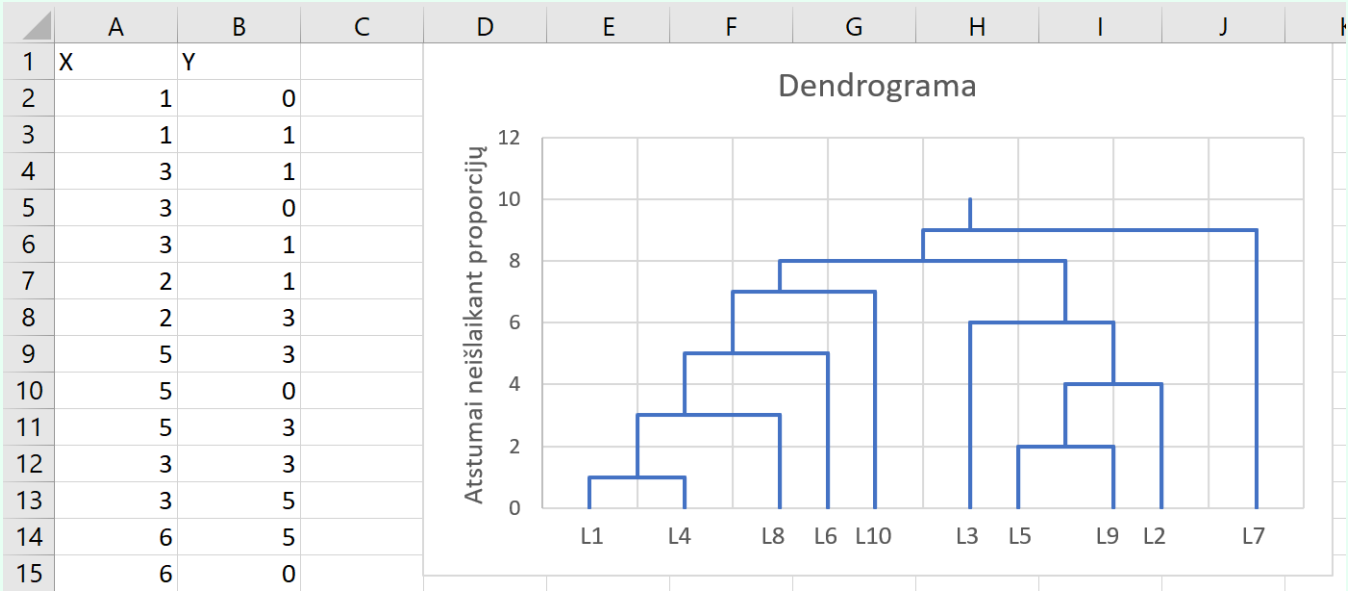


1. **Vertikalios linijos** atspindi klasterių sujungimo lygį. Kuo aukščiau linija, tuo didesnis atstumas tarp sujungiamų klasterių ar duomenų taškų.
2. **Horizontali linija** nurodo klasterių susijungimo tašką. Kiekvienas šis taškas rodo, kurioje vietoje du ar daugiau klasterių susijungia, o dendrograma leidžia stebėti šiuos susijungimus hierarchiškai.
3. **Duomenų taškai** arba **objektai**, pradedant nuo apačios, vaizduoja individualius duomenų taškus arba objektus. Kylant aukštyn, jie susijungia, formuodami didesnius klasterius.
4. **Pjūvio (angl. cut-off) linijos**: Norint nustatyti klasterių skaičių, dendrogramoje galima brėžti horizontalias linijas tam tikru aukščiu. Darant dendrogramos pjūvius tam tikrame aukštyje, galima atskirti, kiek klasterių bus suformuota.

Dendrogramos interpretacijoje galima matyti, kaip individualūs taškai (objektai) apačioje susijungia į grupes (klasterius). Šios grupės toliau jungiasi, kol susiformuoja vienas didelis klasteris arba keli stambūs klasteriai, priklausomai nuo pasirinkto suskirstymo lygio.

Mokomasis pavyzdys

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Fiksuota	Šalis	Nužudymai	Apiplėšimai	Gyventojų skaičius	Išprievartavimai	1 veiksmas	2 veiksmas	3 veiksmas	4 veiksmas	5 veiksmas	6 veiksmas	7 veiksmas	8 veiksmas	9 veiksmas	10 veiksmas
2	1	L1	12	193	51	21	11									
3	2	L2	9	222	39	42	12									
4	3	L3	6	226	72	30	13									
5	4	L4	8	177	50	17	14									
6	5	L5	7	235	86	39	15									
7	6	L6	5	170	78	37	16									
8	7	L7	0	35	75	7	17									
9	8	L8	4	198	70	15	18									
10	9	L9	13	259	74	30	19									
11	10	L10	15	136	52	24	20									



Hierarchinė klasterizacija Orange



Distances - Orange

File View Window Help

Compare

☒ Rows ☐ Columns

Distance Metric

☐ Euclidean (normalized) ☐ Cosine

☒ Euclidean ☐ Pearson

☐ Manhattan (normalized) ☐ Pearson (absolute)

☐ Manhattan ☐ Spearman

☐ Mahalanobis ☐ Spearman (absolute)

☐ Hamming ☐ Jaccard

☒ Apply Automatically

Linkage

Complete

Annotations

Salis

☐ Show labels only for subset

Color by: None

Pruning

☐ None

☒ Max depth: 20

Selection

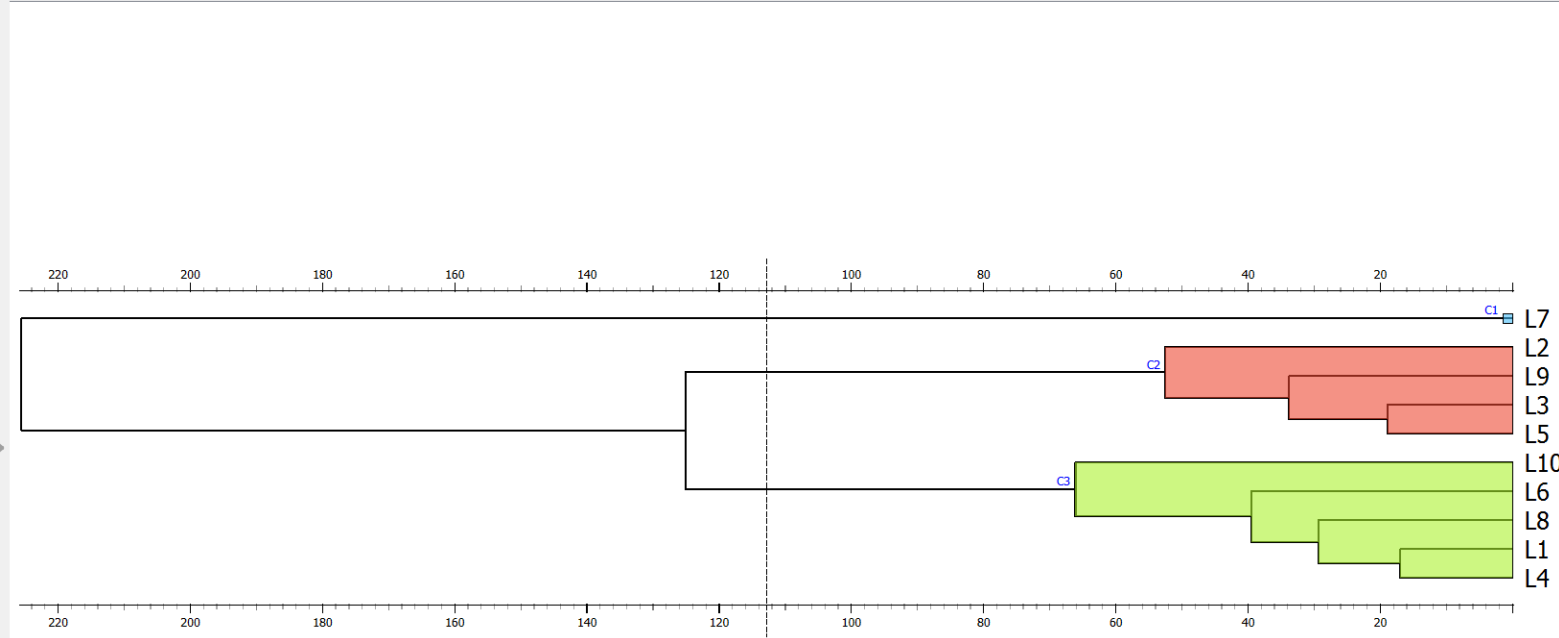
☐ Manual

☒ Height ratio: 50,0 %

☐ Top N: 3

Zoom

10 10x10



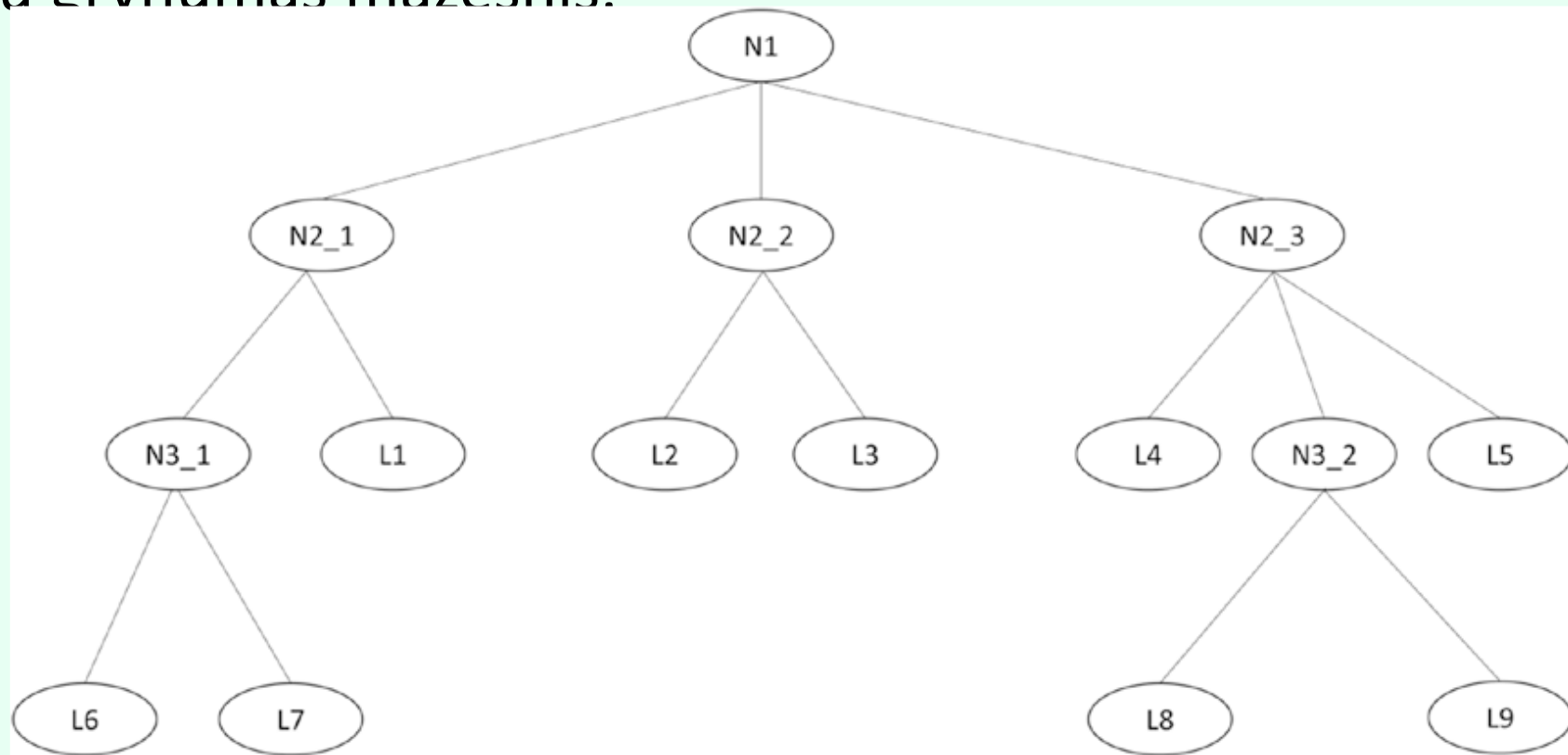
Savarankiško darbo užduotys

1. Išnagrinėkite ir praktiškai atlikite hierarchinės klasterizacijos pavyzdžius, pateiktus Orange dokumentacijoje:
 - <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html>
 - video įrašuose:
 - <https://www.youtube.com/watch?v=ZiOv5YAiE4I>,
 - https://www.youtube.com/watch?v=UYz5vIxB_UI,
 - https://www.youtube.com/watch?v=j_iqLi-NHF8.
1. Duomenų rinkinyje, pateiktame faile [Mall_Customers.csv](#) (šaltinis: <https://www.kaggle.com/code/devahuja2808/hierarchical-clustering/input>) pateikti duomenys apie klientus. Pasirinktu įrankiu (pvz., Excel, Orange, Python ir kt.) atlikite hierarchinę klasterizaciją ir paaiškinkite gautus rezultatus.

Sprendimų medžiai

Metodo esmė

- Sprendimų medis yra taisyklėmis pagrįstas metodas.
- Mazgo padalijimas medyje pagrįstas entropijos skaičiavimu: entropija parodo, kiek medžio mazgo duomenys yra „gryni“. Kuo didesnė entropija, tuo duomenų grynumas mažesnis.

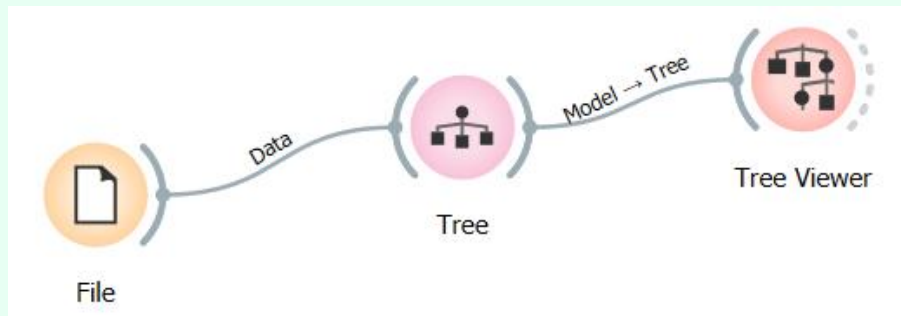


Mokomasis pavyzdys: sprendimų medžio taikymas prognozei

	A	B	C	D	E	F	G
1	Temperatūra	Drėgmė	Vėjas	Oras	Žaidimas	Tikimybė	
2	Žema	Didelė	Vėjuota	Apsiniaukę	Taip	1	
3	Žema	Normali	Vėjuota	Saulėta	Taip	1	
4	Tinkama	Didelė	Vėjuota	Lietinga	Ne	1	
5	Tinkama	Didelė	Vėjuota	Lietinga	Ne	1	
6	Aukšta	Didelė	Nevėjuota	Saulėta	Ne	1	
7	Aukšta	Normali	Vėjuota	Apsiniaukę	Taip	1	
8	Žema	Didelė	Vėjuota	Saulėta	Ne	1	
9	Tinkama	Didelė	Vėjuota	Lietinga	Ne	1	
10	Tinkama	Normali	Vėjuota	Lietinga	Ne	1	
11	Žema	Didelė	Vėjuota	Lietinga	Ne	1	
12	Tinkama	Didelė	Nevėjuota	Saulėta	Ne	1	
13	Tinkama	Normali	Nevėjuota	Saulėta	Taip	1	
14	Aukšta	Didelė	Nevėjuota	Apsiniaukę	Taip	1	
15	Žema	Didelė	Nevėjuota	Apsiniaukę	Taip	1	
16							
17				Oras			
18							
19		Lietinga(3,2)		Apsiniaukę(0,0)		Saulėta(2,3)	
20							
21	Vėjuota(0,2)	Nevėjuota(3,0)				Didelė drėgmė(0,3)	Normali drėgmė(2,0)
22							

Sprendimų medžio realizacija Orange (1)

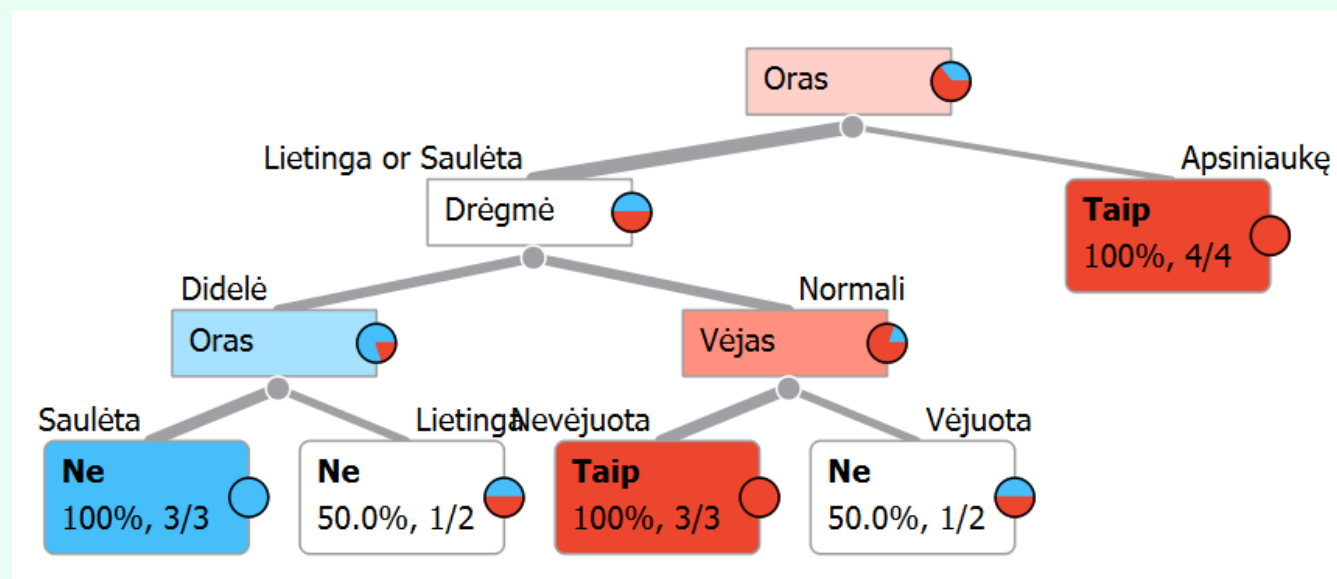
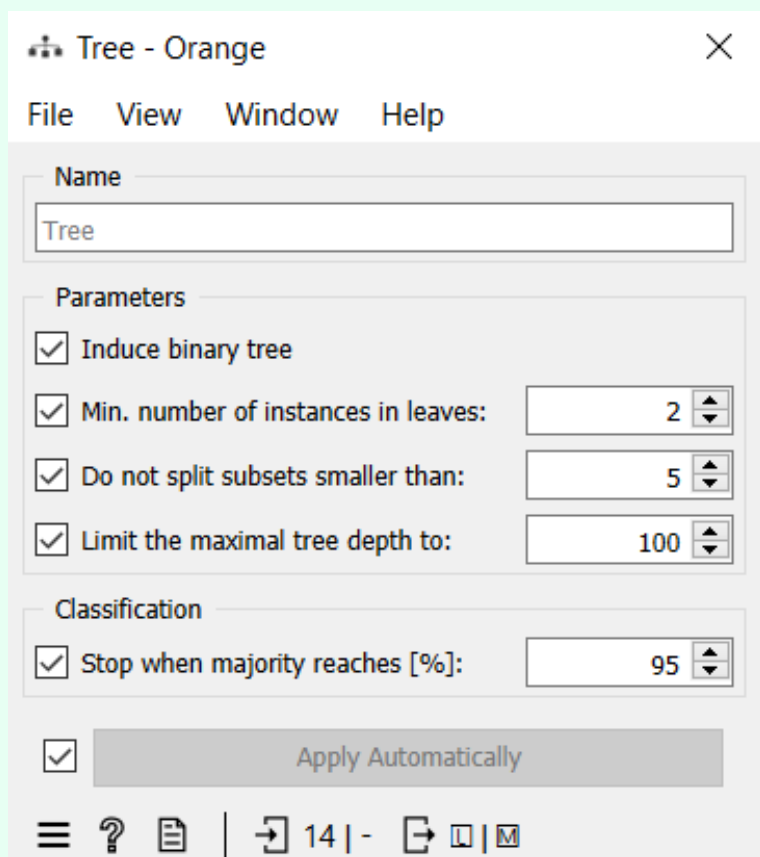
- Sprendimų medžio modelis
- Modelio duomenų atributų reikšmės



Info				
14 instances				
5 features (no missing values)				
Data has no target variable.				
0 meta attributes				
Columns (Double click to edit)				
	Name	Type	Role	Values
1	Temperatūra	C categorical	feature	Aukšta, Tinkama, Žema
2	Drėgmė	C categorical	feature	Didelė, Normali
3	Vėjas	C categorical	feature	Nevejuta, Vėjuta
4	Oras	C categorical	feature	Apsiniaukę, Lietinga, Saulėta
5	Žaidimas	C categorical	target	Ne, Taip

Sprendimų medžio realizacija Orange (2)

- Valdiklio Tree nustatymai
- Sprendimų medis



Sprendimų medžių privalumai ir trūkumai

Privalumai

- Intuityvūs ir lengvai interpretuojami.
- Tinkami apdoroti ir skaitiniams, ir kategoriniams duomenims.
- Nereikia papildomai apdoroti duomenų (normalizuoti, standartizuoti).
- Gebėjimas apdoroti sudėtingų struktūrų duomenis.
- Tinka svarbiausių atributų atrankai.

Trūkumai

- Priklauso nuo duomenų kokybės.
- Persimokymo rizika.
- Trūkstamų duomenų valdymo sudėtingumas.
- Kintamojo pasirinkimo šališkumas (didesnį skaičių reikšmių turintys kintamieji gali būti dažniau pasirenkami kaip pagrindiniai atributai, nes jie turi daugiau potencialių padalijimų).

Sprendimų medžių taikymo sritys

Finansai:

- **Kredito rizikos vertinimas:** Naudojant sprendimų medžius, galima įvertinti klientų kredito riziką pagal jų finansinę istoriją, pajamų lygį ir kitus rodiklius.
- **Sukčiavimo aptikimas:** Sprendimų medžiai gali padėti aptikti netipines finansines operacijas, kurios gali būti sukčiavimo atvejai, remiantis operacijų istorija ir kliento elgsenos modeliu.

Medicinos diagnostika:

- **Ligos prognozavimas:** Sprendimų medžiai gali būti naudojami prognozuoti paciento riziką susirgti tam tikra liga, remiantis simptomais, amžiumi ir kitais rodikliais.
- **Vėžio tipų klasifikacija:** Jie naudojami nustatyti, ar navikas yra piktybinis ar gerybinis, remiantis pacientų duomenimis ir naviko savybėmis.

Rinkodara ir klientų analizė:

- **Klientų segmentavimas:** Sprendimų medžiai gali segmentuoti klientus pagal elgseną, pirkimo įpročius ar demografinius duomenis.
- **Produktų rekomendacijos:** Naudojami prognozuoti, kokie produktai galėtų būti aktualūs klientui pagal jo ankstesnius pirkimus ir panašių klientų pasirinkimus.

Aplinkos apsauga ir tyrimai:

- **Aplinkos duomenų analizė:** Sprendimų medžiai naudojami prognozuoti oro kokybės pokyčius, vandens taršos lygį ar net ekologinius pavojus, remiantis meteorologiniais duomenimis ir kitais veiksniais.
- **Gyvūnų elgsenos analizė:** Jie gali būti naudojami analizuoti ir klasifikuoti gyvūnų rūšis pagal jų savybes arba prognozuoti gyvūnų populiacijų pokyčius pagal aplinkos veiksnus.

Žmogiškieji ištekliai:

- **Darbuotojų atranka:** Sprendimų medžiai gali padėti prognozuoti, kurie kandidatai turi didžiausią sėkmės potencialą pagal jų patirtį, išsilavinimą ir testų rezultatus.
- **Išėjimo iš darbo prognozė:** Jie gali numatyti darbuotojų išėjimo iš darbo riziką pagal jų darbo laiką, našumą, pasitenkinimą darbu ir kitus atributus.

Savarankiško darbo užduotys

Pritaikykite sprendimų medžius klasifikacijos uždaviniams spręsti. Remdamiesi gautais rezultatais suformuluokite išvadas. Siūlomi duomenų rinkiniai:

1. Nacionalinių mokinių pasiekimų patikrinimų (NMPP) 2022-2023 m. m. duomenys (šaltinis: <https://data.gov.lt/datasets/1930/>, failas [Pasiekimai_P1.xlsx](#)).
2. Duomenų rinkinyje yra pateikiami mobilios interneto prieigos duomenų perdavimo spartos kontrolinių matavimų rezultatai. Matavimai atliekami operatorių UAB „Bitė Lietuva“, AB „Telia Lietuva“ ir UAB „TELE2“ mobiliojo ryšio tinkluose visoje Lietuvos teritorijoje važiuojant keleivinio geležinkelio maršrutais (šaltinis: <https://data.gov.lt/datasets/1420/resource/9715>, failas [Internetas_P2.xlsx](#)).

Kas pavyko

- Sudominti mokinius ir išlaikyti motyvaciją.
- Išsamūs pamokų scenarijai palengvino darbą.
- Mokiniams suprantamiausi vizualiai atvaizduoti rezultatai.
- Remdamiesi gautais rezultatais mokiniai noriai formulavo įžvalgas.
- Mokiniai lygino skaičiuokle ir Orange gautus rezultatus, bandė įvardinti, kodėl rezultatai taikant kai kuriuos metodus skiriasi.
- Keli mokiniai rimtai nusprendė susieti ateitį su duomenų mokslu (data science).

Kas buvo sunkiausia

- Surasti tinkamas sąvokas lietuvių kalba (ypač teksto tyrybai).
- Parinkti tinkamus duomenų rinkinius analizei.
- Pritrūko laiko projektiniams darbams.

Kas toliau?

- Papildyti praktikos darbus mašininio mokymosi metodų realizacijomis Python programavimo kalba.
- Papildyti praktikos darbus video pamokomis.
- Papildyti užduočių bazę naujais duomenų rinkiniais.
- Domėtis naujovėmis ir jas pritaikyti sukuriant praktikos darbus.

Visa medžiaga

- <https://drive.google.com/drive/folders/1f7iBc-HlI0nqthrFZvnbub8fxrMTerv-?usp=sharing>