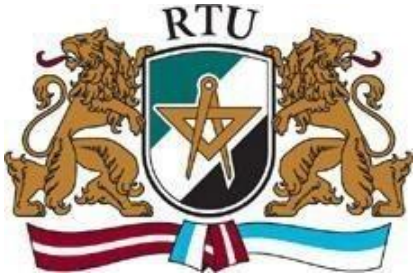
Datorzinātnes un informācijas tehnoloģijas fakultāte

Datorvadības, automātikas un datortehnikas institūts



Mākslīgā intelekta pamati

Referāts

2.praktiskā darba atskaite

Jekaterina Leitarte

Stud.apl.Nr. 181RDB219

Saite uz GitHub projektu:

https://github.com/jekmiz/MI_Orange_2023

Rīga, 2023

SATURS

1. DATU KOPAS APRAKSTS

Datu kopa ir pieejama: https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman

Saite uz GitHub projektu: https://github.com/jekmiz/MI_Orange_2023

Datu kopas autors: Utkarsh Singh

Šīs datu kopas veidotāja veidoja šo datu kopu ar mērķi analizēt, kā arī trenēties datu vizualizācijā. Šī datu kopa satur pacientu ierakstus no 1984.gada līdz 1989.gada pētījuma, ko veica Vācijas Krūts vēža pētījumu grupa. Tajā piedalījās 720 pacienti ar krūts vēzi. Datu kopā ir saglabāti 686 pacienti ar pilnīgiem datiem un prognostiskajiem mainīgajiem.

Pēc datu sākotnējās sagatavošanas .csv fails tika ielādēts Orange rīkā. Pēc datu ielādēšanas var redzēt, kā izskatās ielādēti dati tam speciāli paredzētajā rīkā. Uzreiz tiks parādīta informācija par to, cik daudz ir datu objektu, kā arī , ka ir 12 pazīmes un nav izvēlēta mērķa pazīme, pēc kuras tiks veikta datu analīze.

Datu kopas satura apraksts

Darba autorei ir jānorāda, ka dati netika papildināti no darba autores puses.

?	pid	age	meno	size	grade	nodes	pgr	er	hormon	rfstime	status
1	132	49	0	18	2	2	0	0	0	1838	0
2	1575	55	1	20	3	16	0	0	0	403	1
3	1140	56	1	40	3	3	0	0	0	1603	0
4	769	45	0	25	3	1	0	4	0	177	0
5	130	65	1	30	2	5	0	36	1	1855	0
6	1642	48	0	52	2	11	0	0	0	842	1
7	475	48	0	21	3	8	0	0	0	293	1
8	973	37	0	20	2	9	0	0	1	42	0
9	569	67	1	20	2	1	0	0	1	564	1
10	1180	45	0	30	2	1	0	0	0	1093	1
11	97	62	1	12	2	7	0	0	0	436	1
12	131	59	1	30	2	8	0	0	1	238	1
13	1574	62	1	21	2	2	0	0	1	723	0
14	762	51	1	25	2	2	0	80	0	503	1
15	820	32	0	57	3	24	0	13	0	448	1

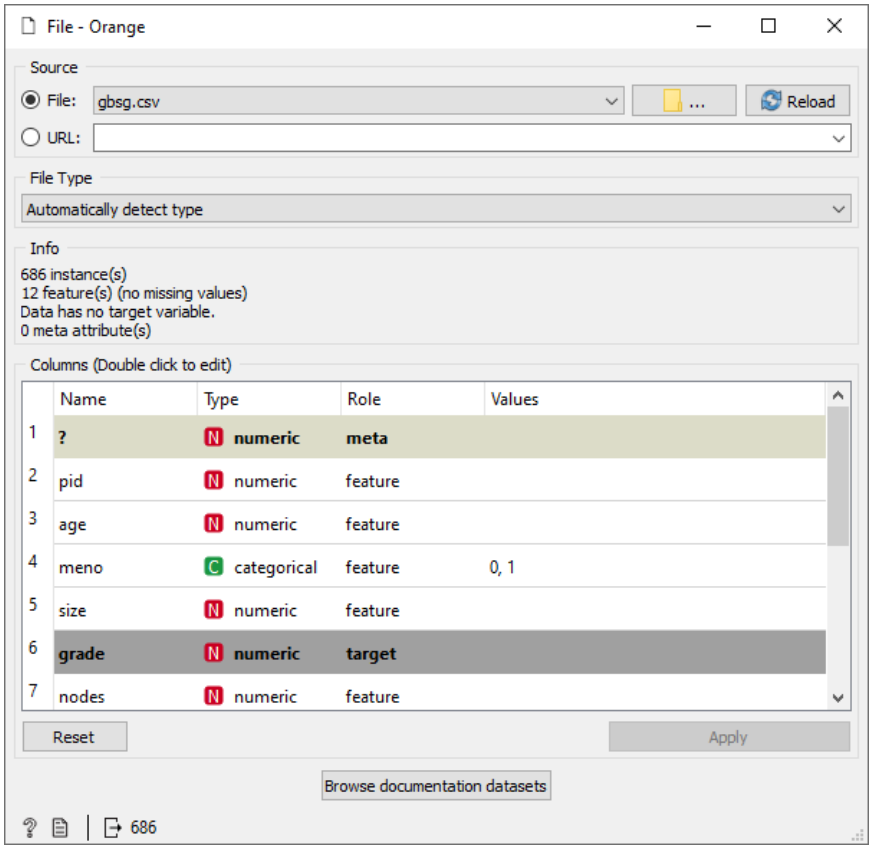
1.1.attēls. Datu kopa

Dati .csv formātā tika ielādēti Orange rīkā.

Kā mērķa pazīmi darba autore ir norādījusi grade (pakāpe).

Datu objektu skaits 686 objekti.

Kopai ir 12 pazīmes, nav tukšo vērtību, ko var redzēt 1.1.attēlā.



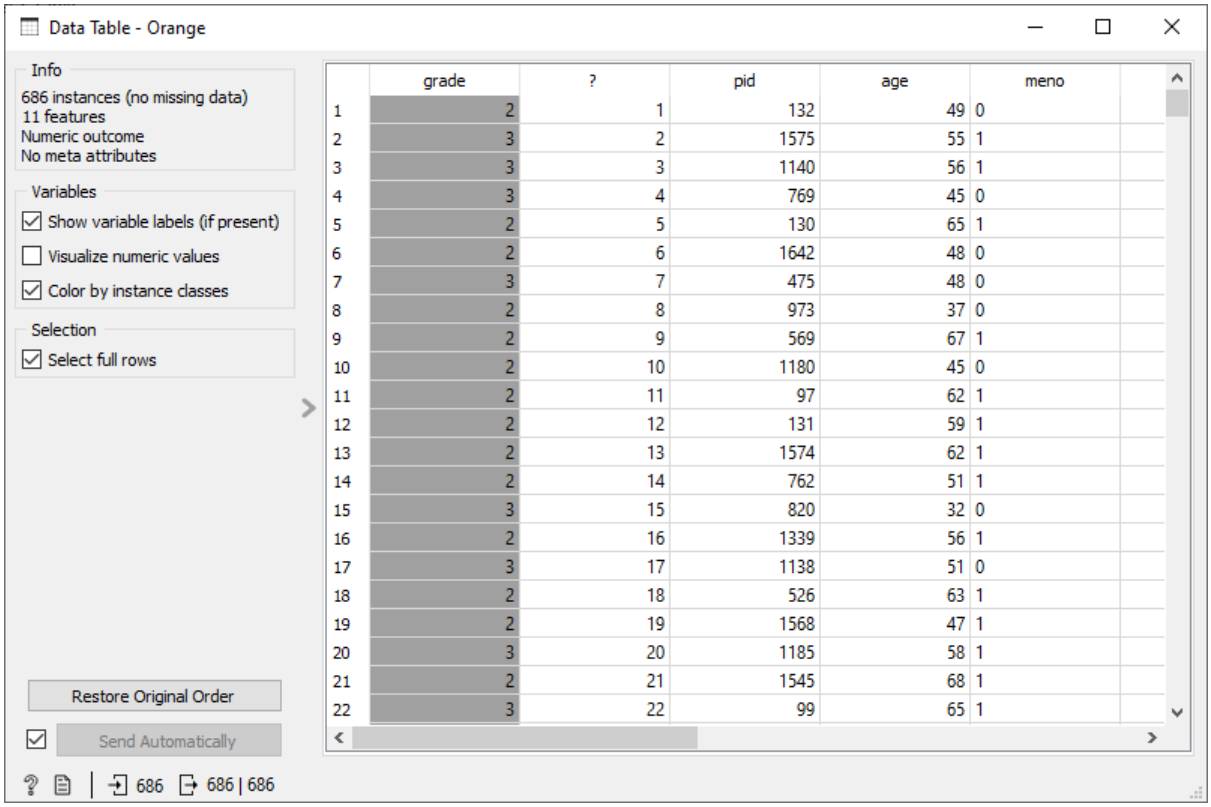
1.2.attēls. Faila ielādēšana Orange rīkā

Pazīmju (atribūtu) skaits un to atspoguļojums:

- Pid – personas ID;
- Age – vecums;
- Meno – vai ir menopauze (menopauzes status, 0 = pre-menopauze, 1 = pēcmenopauze)
- Size – audzēja lielums, mm;
- Grade – audzēja pakāpe;
- Nodes – cik limfmezgli ir skarti;
- Pgr – progesterona receptori (fmol/l);
- Er – estrogena receptori (fmol/l);
- Hormon – hormonālā terapija (0 = nav, 1 = ir);
- Rfstime – izdzīvošanas rādītājs pēc remisijas;

- Status – 0 = dzīve bez atkārtota gadījuma, 1 = atkārtotais gadījums vai nāve.

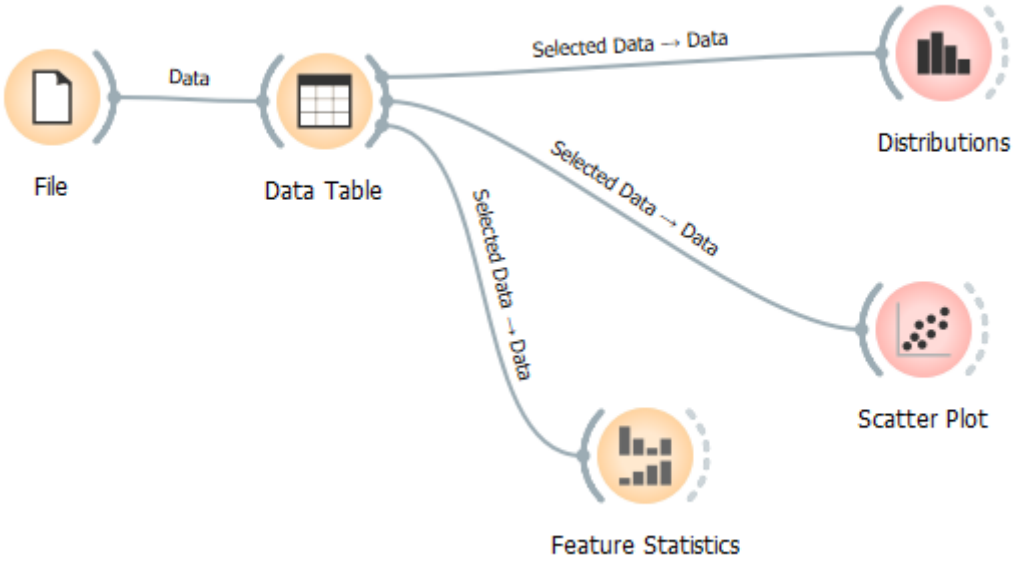
Kā mērķis tika izvēlēts audzēja pakāpe (Grade), saskaņā ar ko tika veikta analīze. Tam ir vairākas vērtības. Ielādētos datus var redzēt datu tabulā.



	grade	?	pid	age	meno
1	2	1	132	49	0
2	3	2	1575	55	1
3	3	3	1140	56	1
4	3	4	769	45	0
5	2	5	130	65	1
6	2	6	1642	48	0
7	3	7	475	48	0
8	2	8	973	37	0
9	2	9	569	67	1
10	2	10	1180	45	0
11	2	11	97	62	1
12	2	12	131	59	1
13	2	13	1574	62	1
14	2	14	762	51	1
15	3	15	820	32	0
16	2	16	1339	56	1
17	3	17	1138	51	0
18	2	18	526	63	1
19	2	19	1568	47	1
20	3	20	1185	58	1
21	2	21	1545	68	1
22	3	22	99	65	1

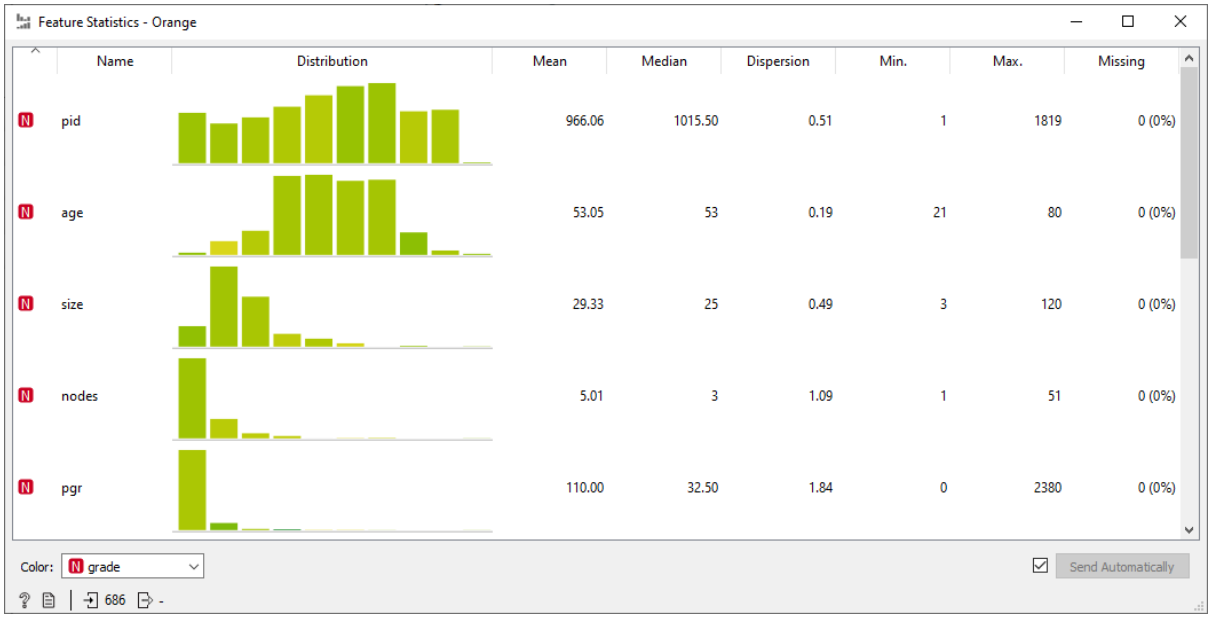
1.3.attēls. Datu tabula

Pēc datu ielādes un sagatavošanas darba autore izvēlējās datu analīzei Feature Statistics, Distributions, Scatter Plot, ko piedāvā Orange rīks, lai varētu izdarīt secinājumus par izliedes diagrammām, histogrāmmām un sadalījumu par datu kopas klašu atdalāmību.



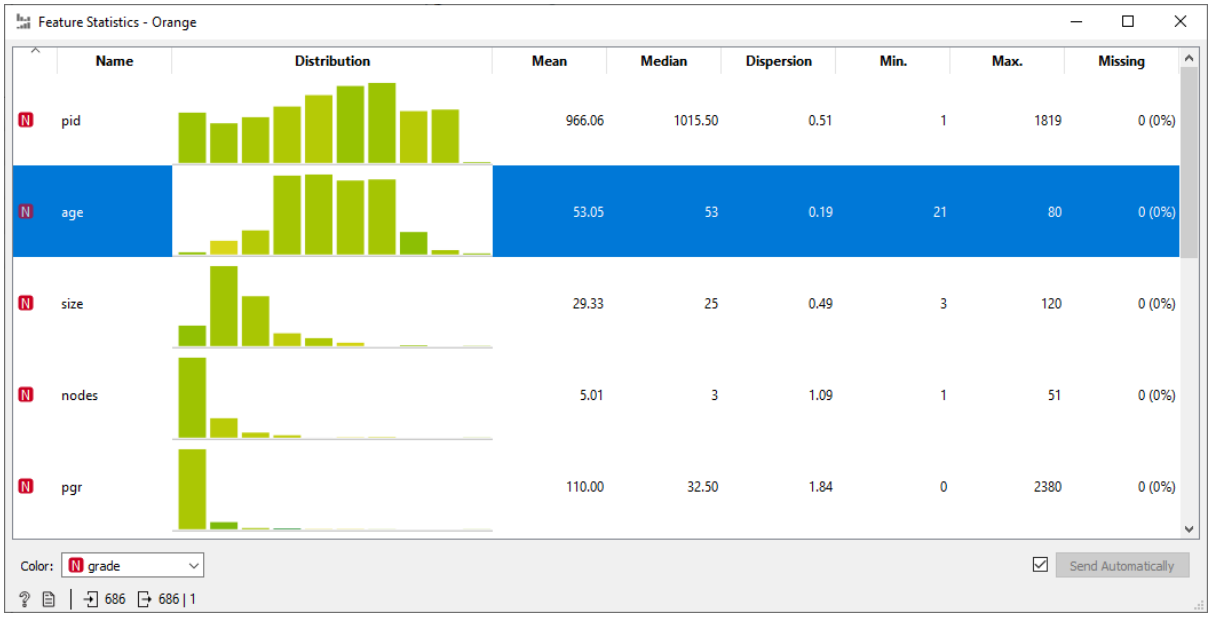
1.4.attēls. Pievienotās programmas daļas

Pirmais, ko darba autore izmantoja, ir Feature Statistics, kura ir atspoguļota 1.4.attēlā.

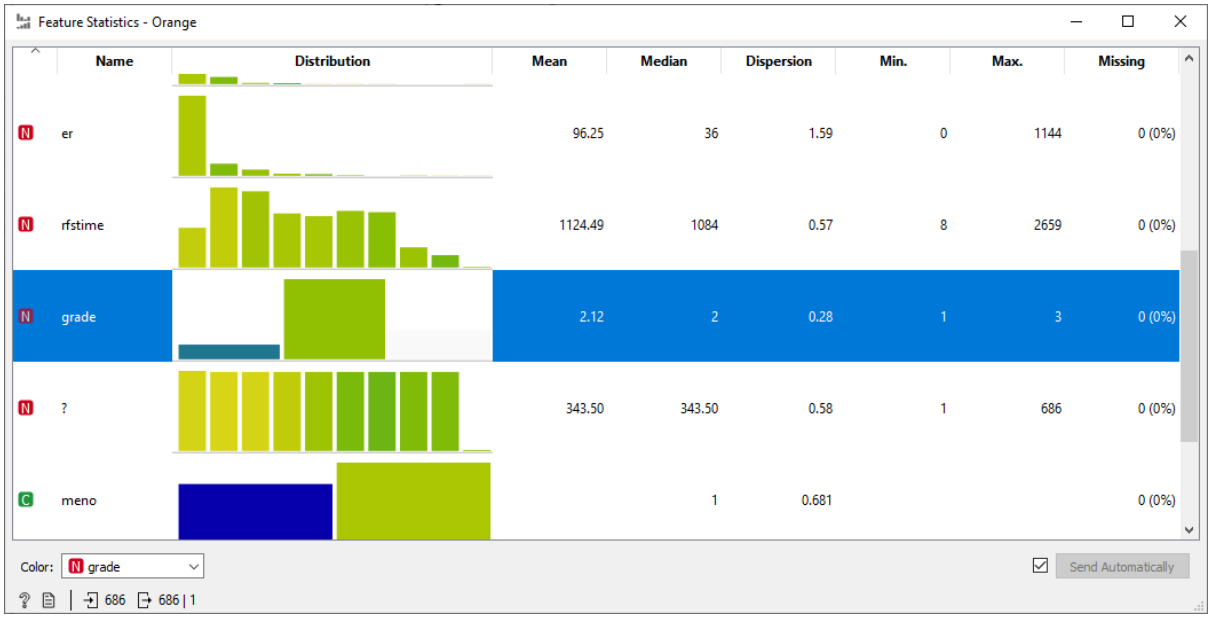


1.5.attēls. Statistikas dati

Zemāk var redzēt vecuma sadalījumu (sk.1.6.att.).

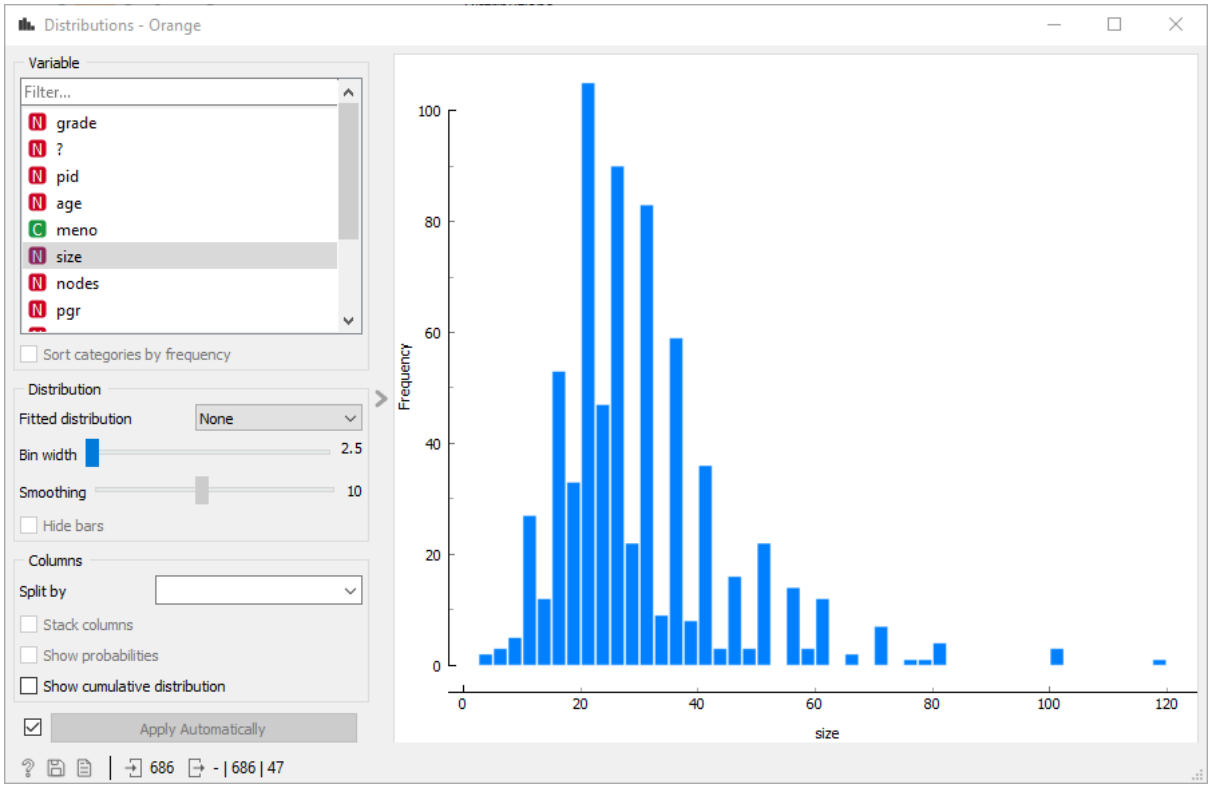


1.6.attēls. Vecuma sadalījums



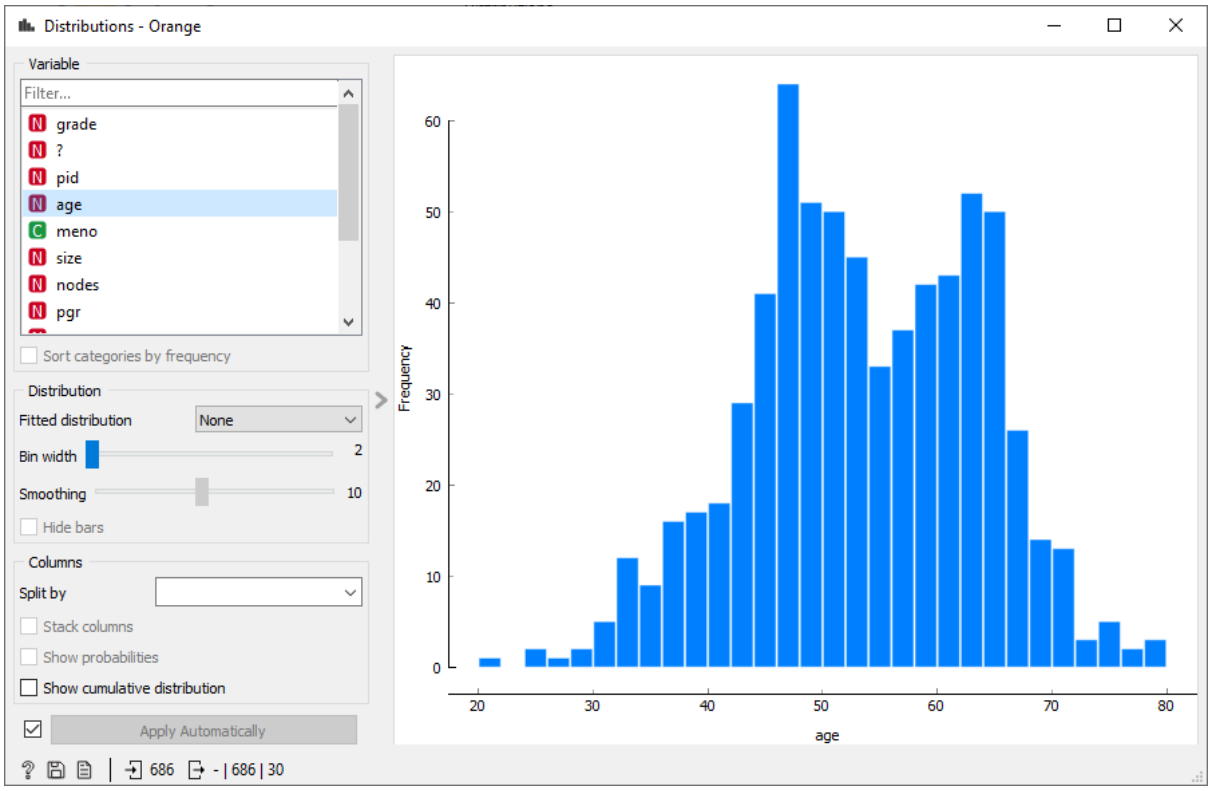
1.7.attēls. Pakāpes (stadijas) sadalījums

Distributions iespēja. 1.8.attēlā ir parādīta informācija par to, kā dalās dati. Distributions funkcija sniedz iespēju aplūkot tās pašas iespējas, tikai detalizētāk. Piemēram, 1.8.attēlā ir parādīta informācija, kā sadalās audzēju izmēri.



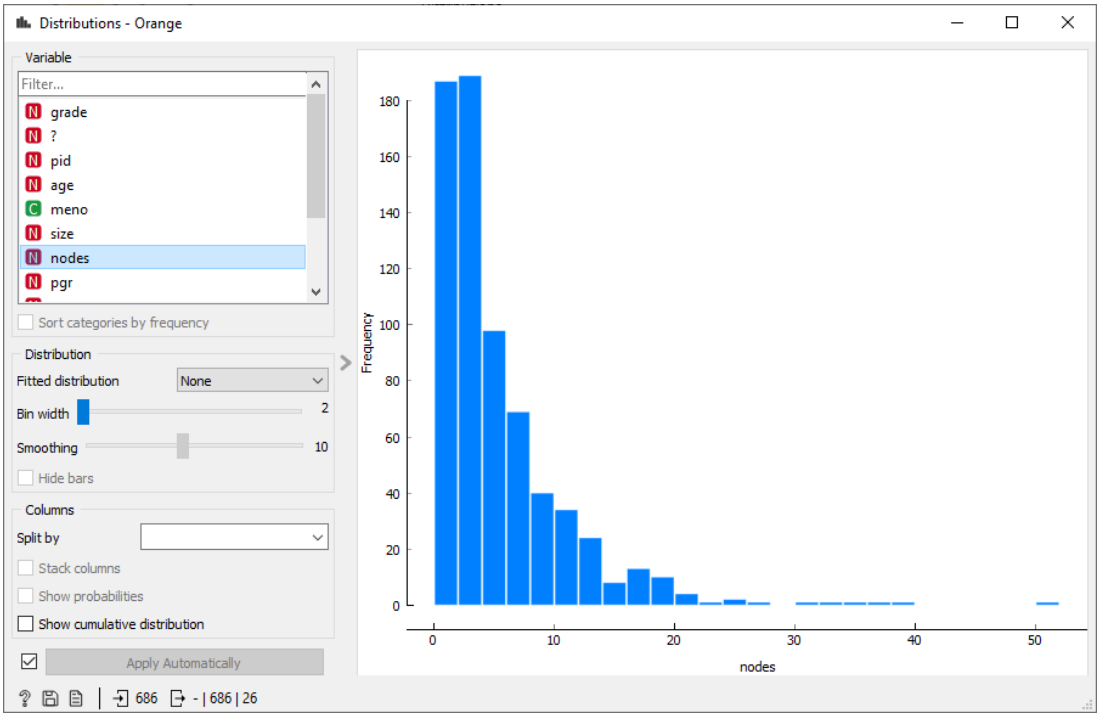
1.8.attēls. Audzēju izmēru sadalījums

1.9.attēlā var aplūkot arī vecuma sadalījumu.



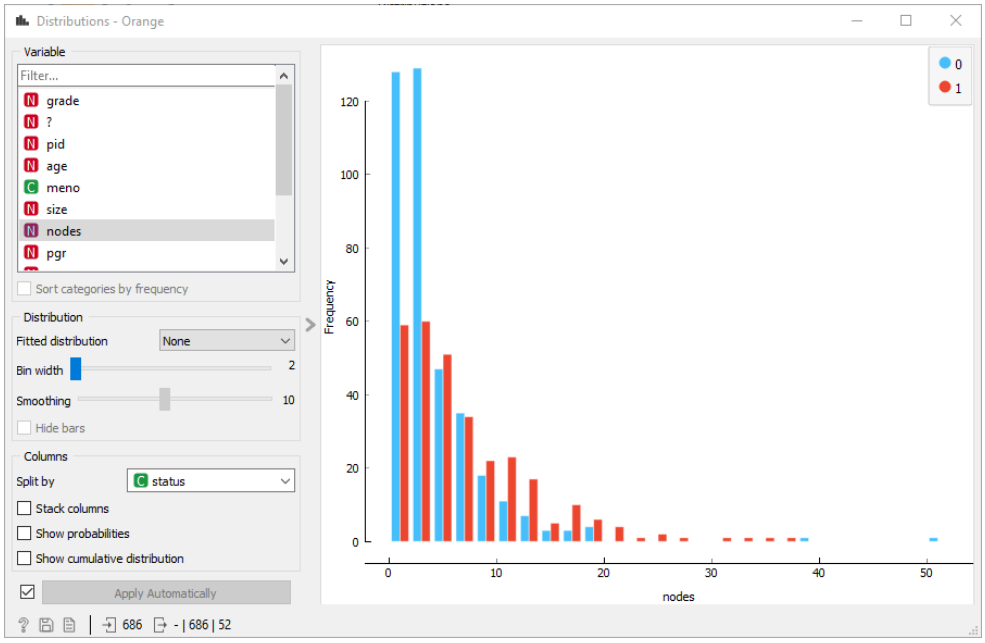
1.9.attēls. Vecuma sadalījums

Tāpat var aplūkot skarto limfmezglu skaitu (sk.1.10.att.).



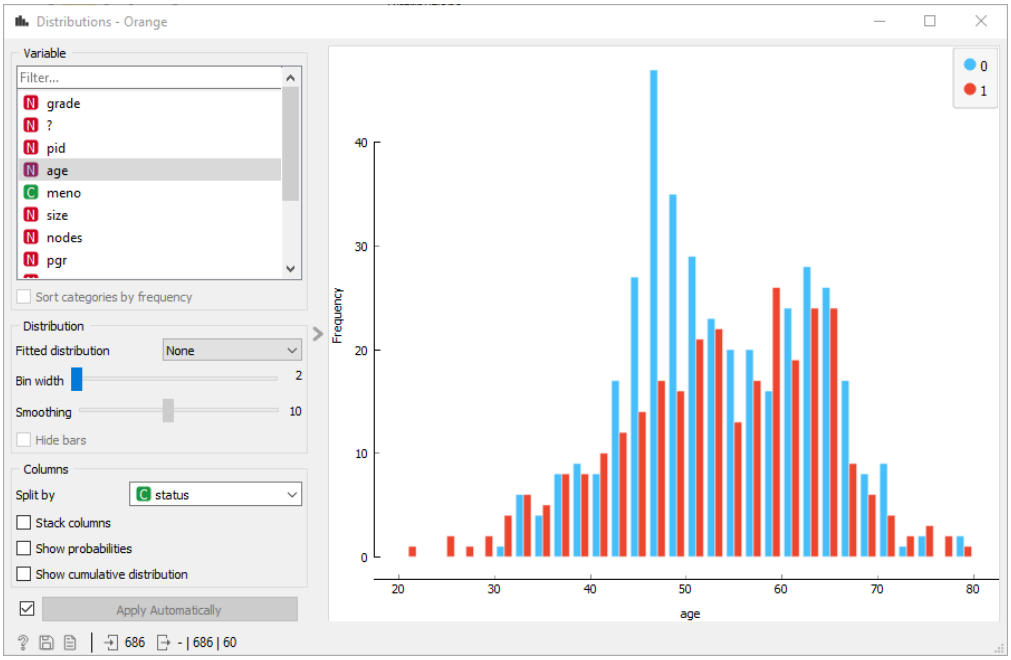
1.10.attēls. Skarto limfmezglu skaits

1.11.attēlā var redzēt sadalījumu, kur 0 ir krūts audzēja pacienti remisijā, bet 1 – pacienti bez remisijas, ar atkārtotiem audzēja gadījumiem vai letāls gadījums. Tāpat var aplūkot, cik daudz šiem pacientiem bija skarto limfmezglu.



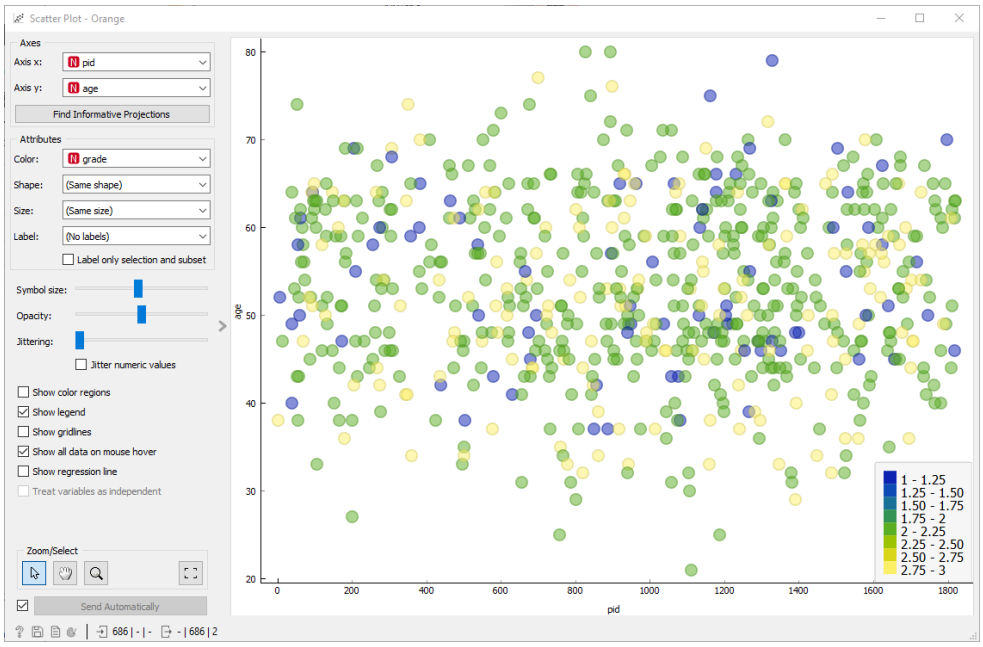
1.11.attēls. Sadalījums atkarībā no skarto limfmezglu un izdzīvošanas

1.11.attēlā ir parādīts vecuma un pacientu statusa, kuriem 0 ir krūts audzēja pacienti remisijā, bet 1 – pacienti bez remisijas, ar atkārtotiem audzēja gadījumiem vai letāls gadījums, sadalījums.



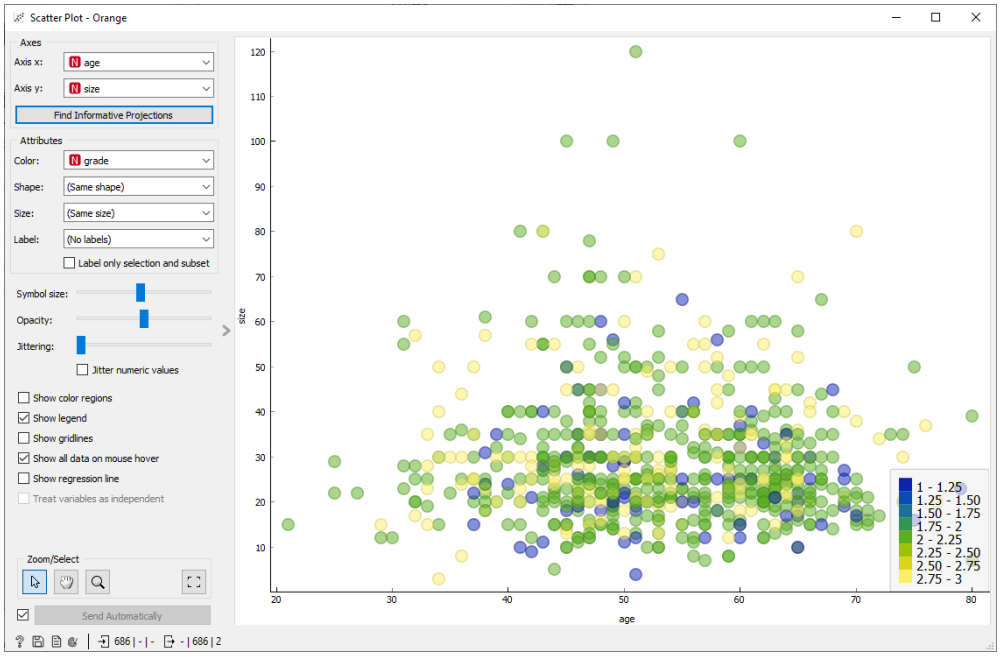
1.12.attēls. Vecuma un statusa sadalījums

Tāpat darba autore ir analizējusi rezultātus, izmantojot Scatter Plot funkcionalitāti (sk.1.13.att.).



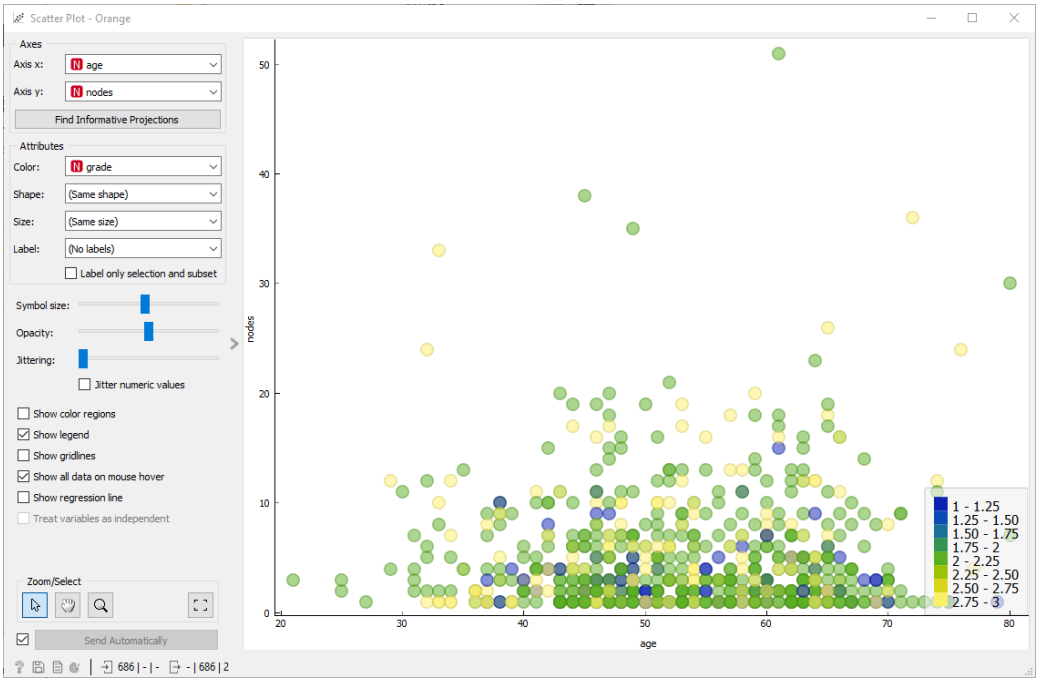
1.13.attēls. Scatter plot

1.14.attēlā ir atspoguļots x asī – vecums, y asī – audzēja izmērs.



1.14.attēls. Vecuma un audzēja izmēra sadalījums

1.15.attēlā ir atspoguļots x asī – vecums, bet y asī – skarto limfmezglu skaits.



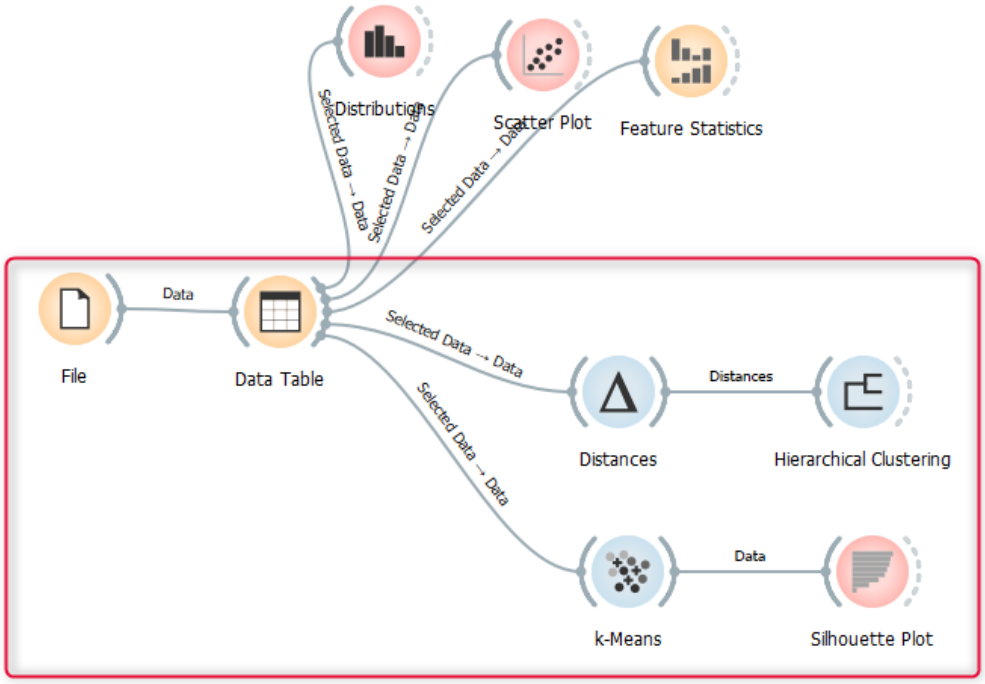
1.15.attēls. Vecuma un skarto limfmezglu sadalījums

Secinājumi:

1. Klašu datu kopas nav līdzsvarotas.
2. Datu vizuālais atspoguļojums, pēc darba autores domām, neļauj līdz galam redzēt datu struktūru, jo dati pārklājas. Datu objekti, kuri pieder dažādām klasēm, nav skaidri atdalāmi. Datu grupējumi atrodas tuvu viens otram, kas apgrūtina datu analīzi.

2. NEPĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS

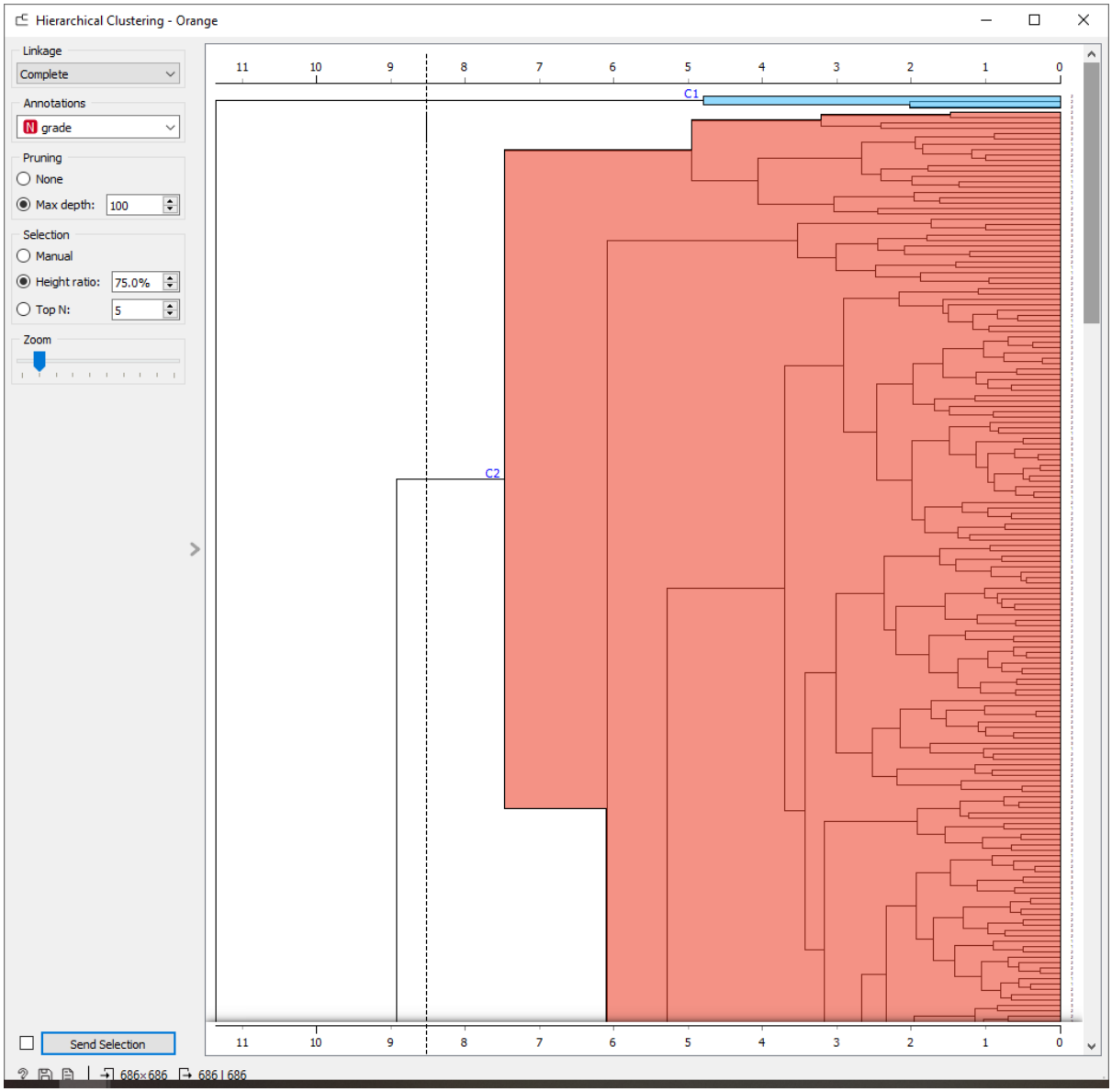
Tika sagatavota darba virsma Orange rīkā.



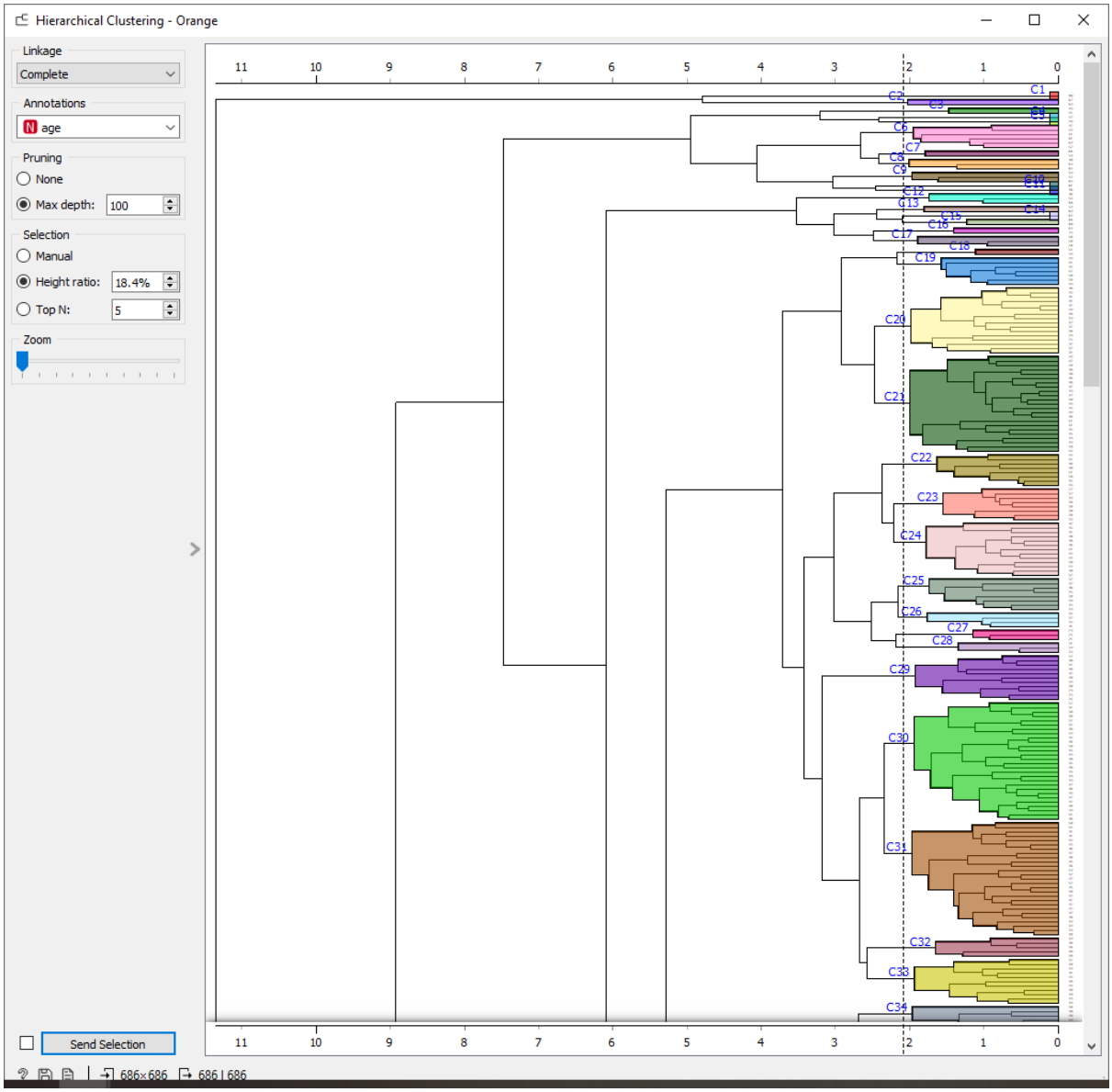
2.1.attēls. Sagatavota darba virsma

2.1. Hierarhiska klasterizācija

Klasterizācijas rezultātā var grupēt datus pēc noteiktām pazīmēm. Klasterus katrā līmenī veido, apvienojot klasterus nākamajā – zemākajā līmenī. Viszemākajā hierarhijas līmenī katram datu objektam ir savs klasteris.



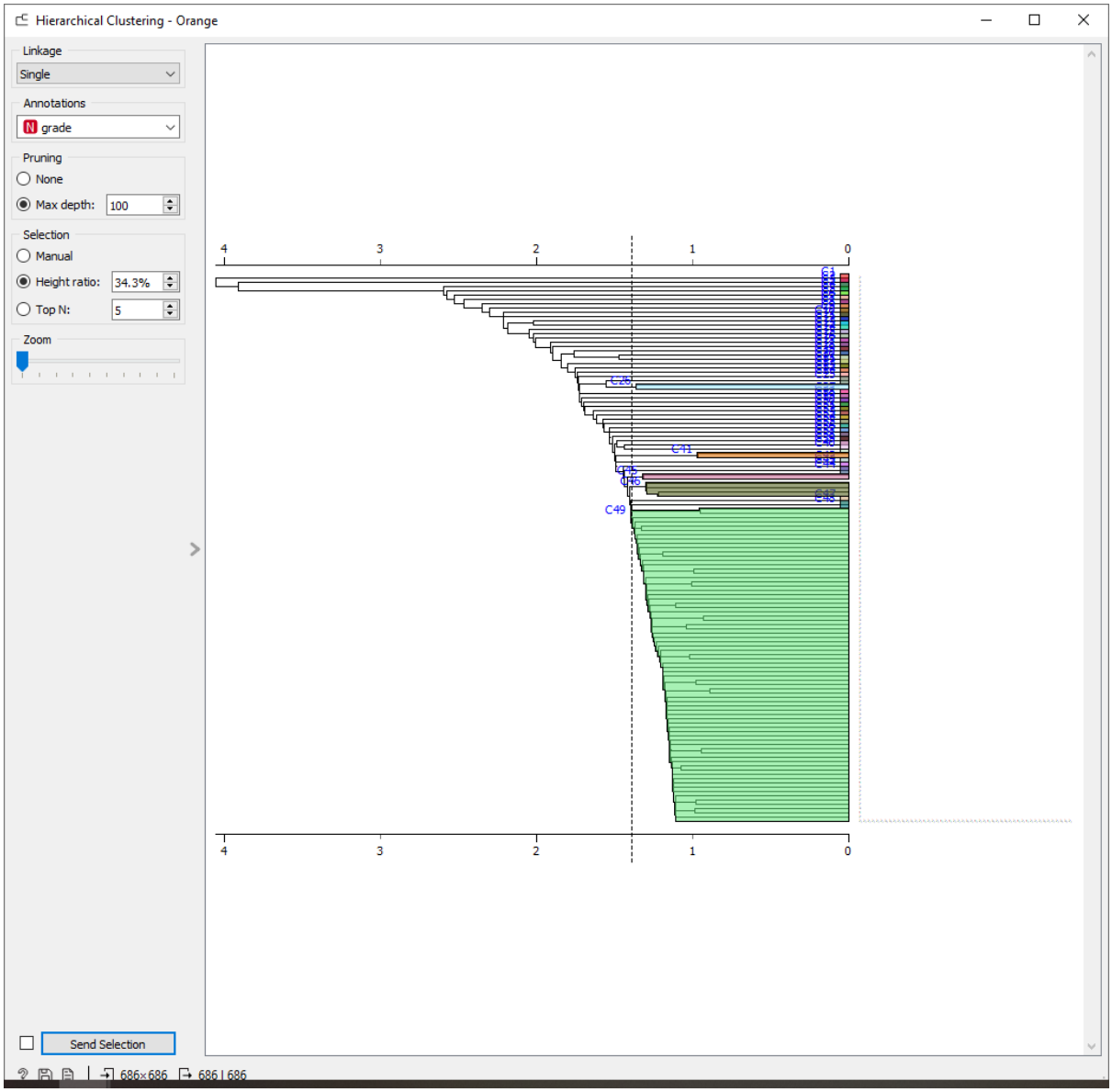
2.2.attēls. Hierarhiskās klasterizācijas piemērs, izmantojot Grade



2.3.attēls. Klasterizācijas piemērs

Secinājumi:

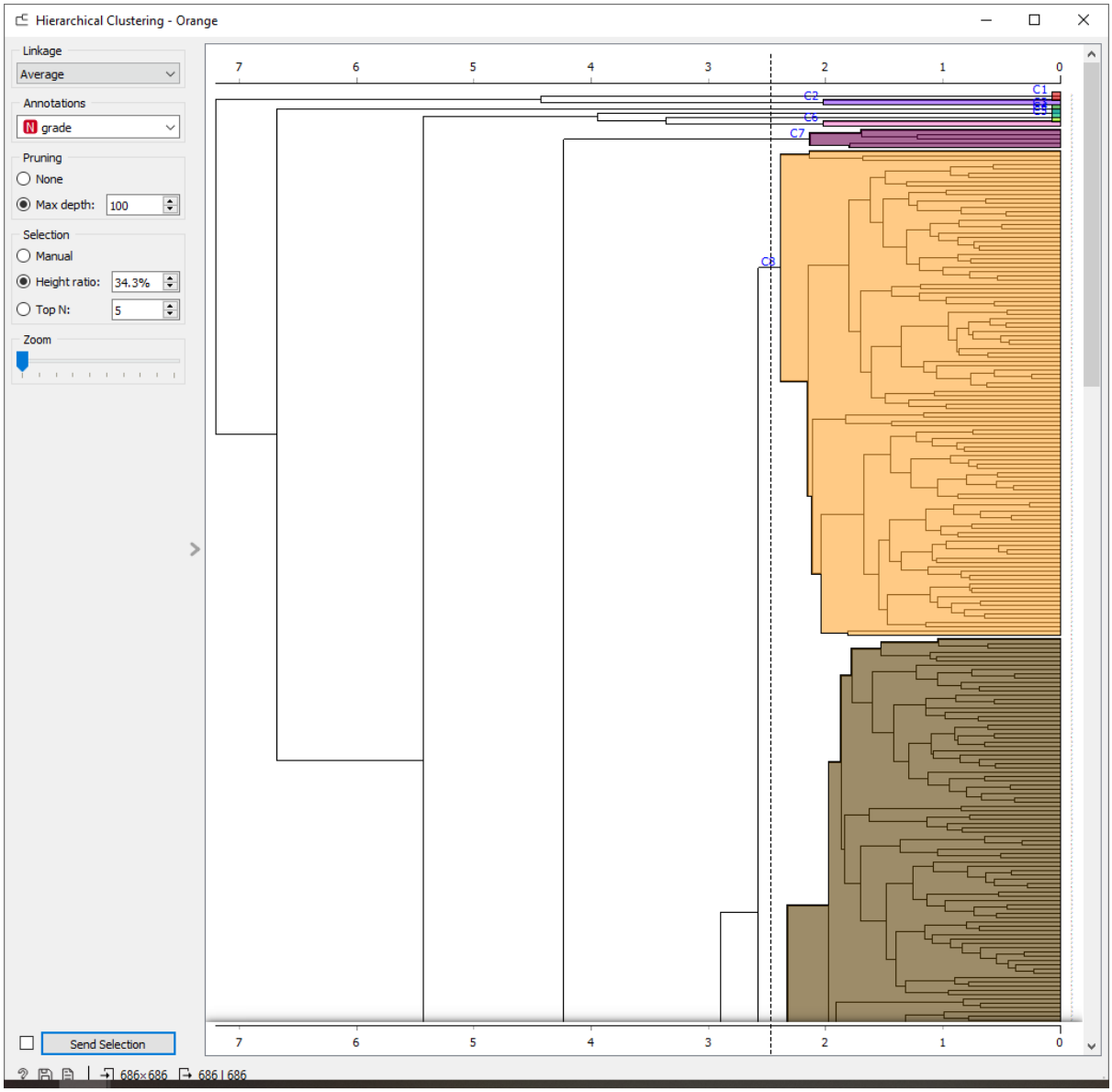
1. Distances netika normalizētas, kā rezultātā rezultāts sanāca nepārskatāms.
2. Eksperimenta rezultātā sanāca ļoti liels klasteru daudzums.



2.4.attēls. Klasterizācijas piemērs (single)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas single vērtības, kā rezultātā arī klasteru daudzums sanāca lielāks.
2. Lai izveidotu šo klasteru, distances tika normalizētas.
3. Mazāki klasteri ir grūtāk pārskatāmi.



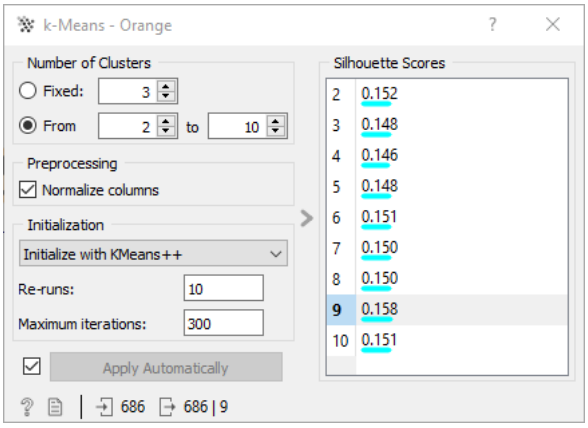
2.5.attēls. Klasterizācijas piemērs (average)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas vidējās vērtības.
2. Distances tika normalizētas, tomēr arī šajā gadījumā rezultāts sanāca ļoti liels.
3. Mazāki klasteri ir grūtāk pārskatāmi.
4. Grupējot datus pēc klasēm, klasteri ir labāk pārskatāmi.

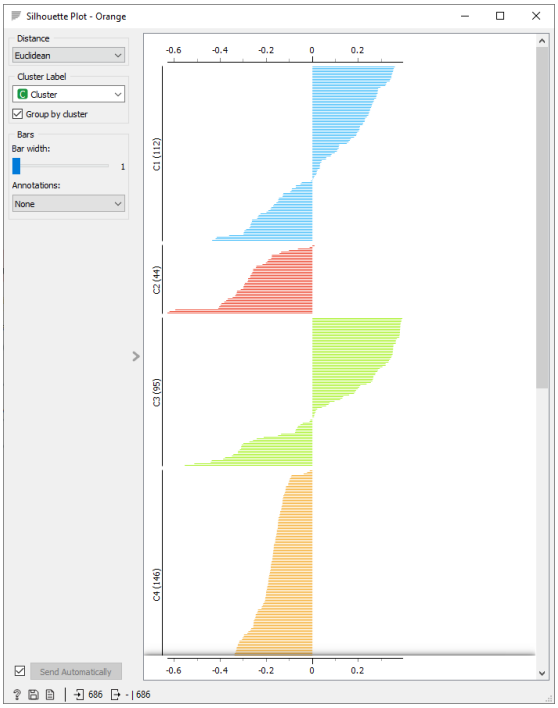
2.2. K-vidējo algoritms

Šis algoritms ir salīdzinoši vienkāršs. To ir viegli lietot. Pēc k-means var redzēt, cik daudz klasteru algoritmam būtu jāizveido. Tas nozīmē, ka K ir šī algoritma hiperparametrs. Var redzēt, ka vislabākais rezultāts ir 9 klasteri.



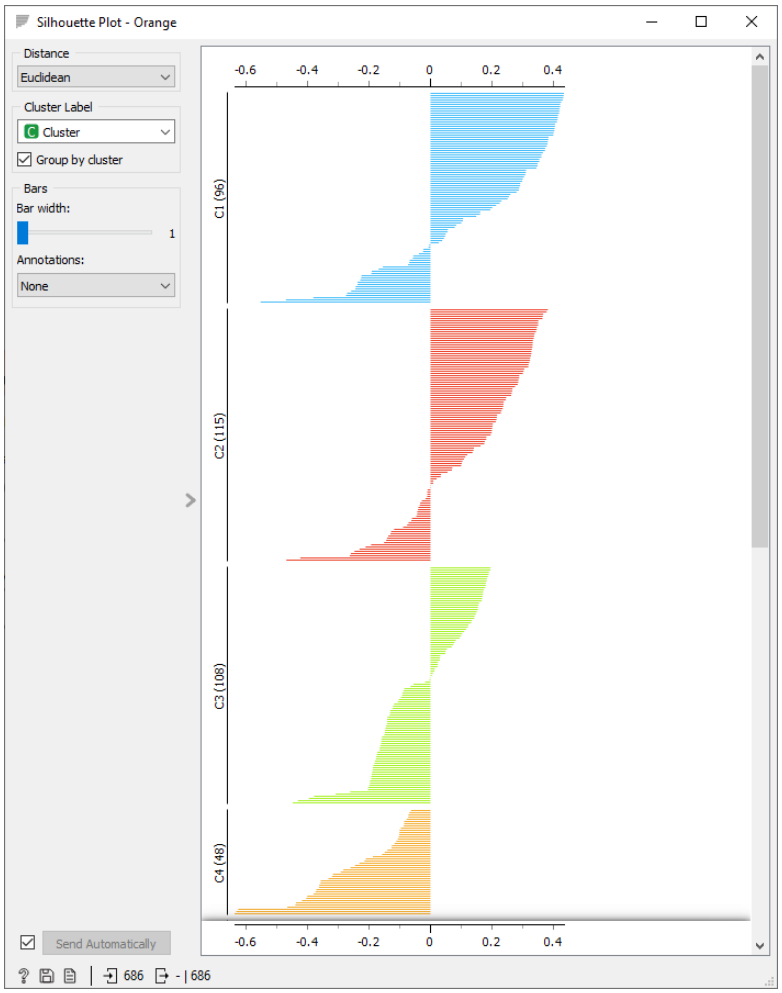
2.6.attēls. Silueta koeficienti

Darba autorei ir jānorāda, ka silueta koeficienti ir tuvāki nullei nekā tiecas uz 1. Jo tuvāk 1, jo labāk klasteri ir atdalāmi. Ja vērtības ir tuvāk 0, tad tas norāda, ka sadalījums starp klasteriem nav visai nozīmīgs. Vispirms eksperiments tika veikts ar 9 klasteriem.



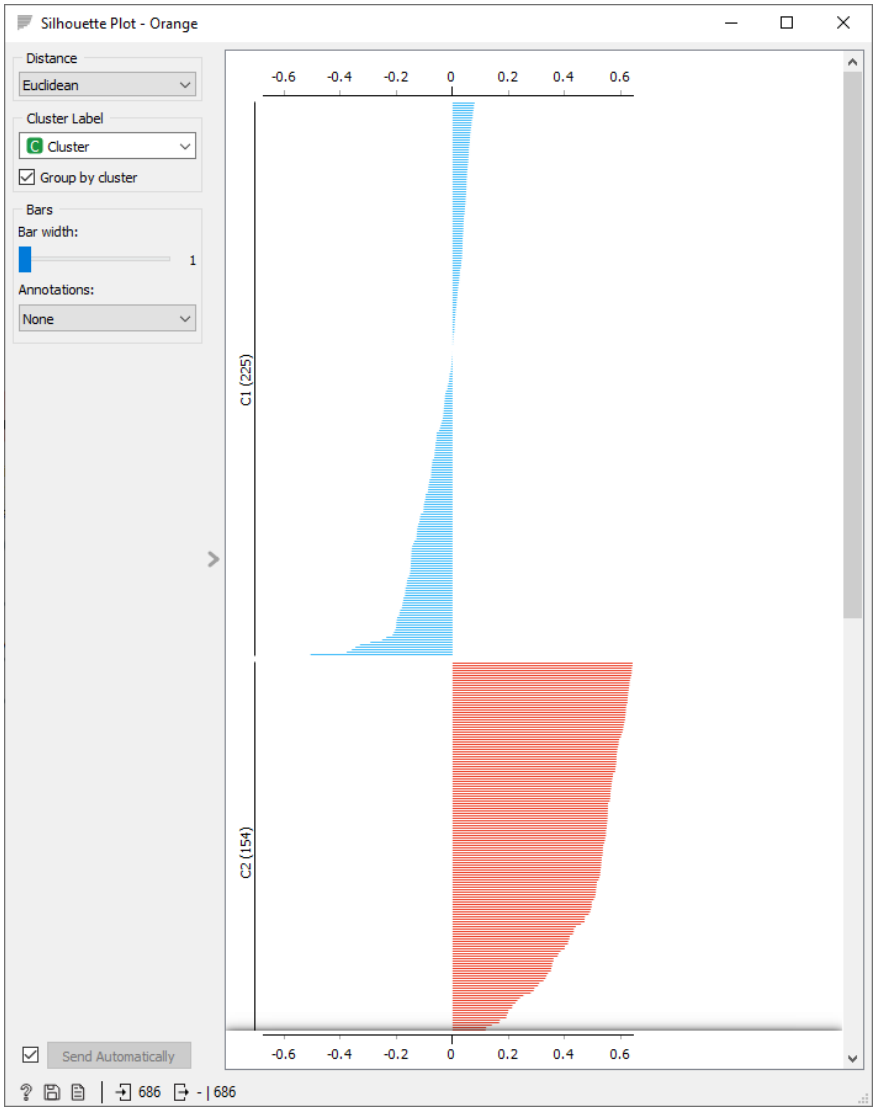
2.7.attēls. 9 klasteri

Darba autore vēlējās uzzināt, kā izskatīsies algoritma darbības rezultāts, ja tiks izmantoti 8 klasteri.



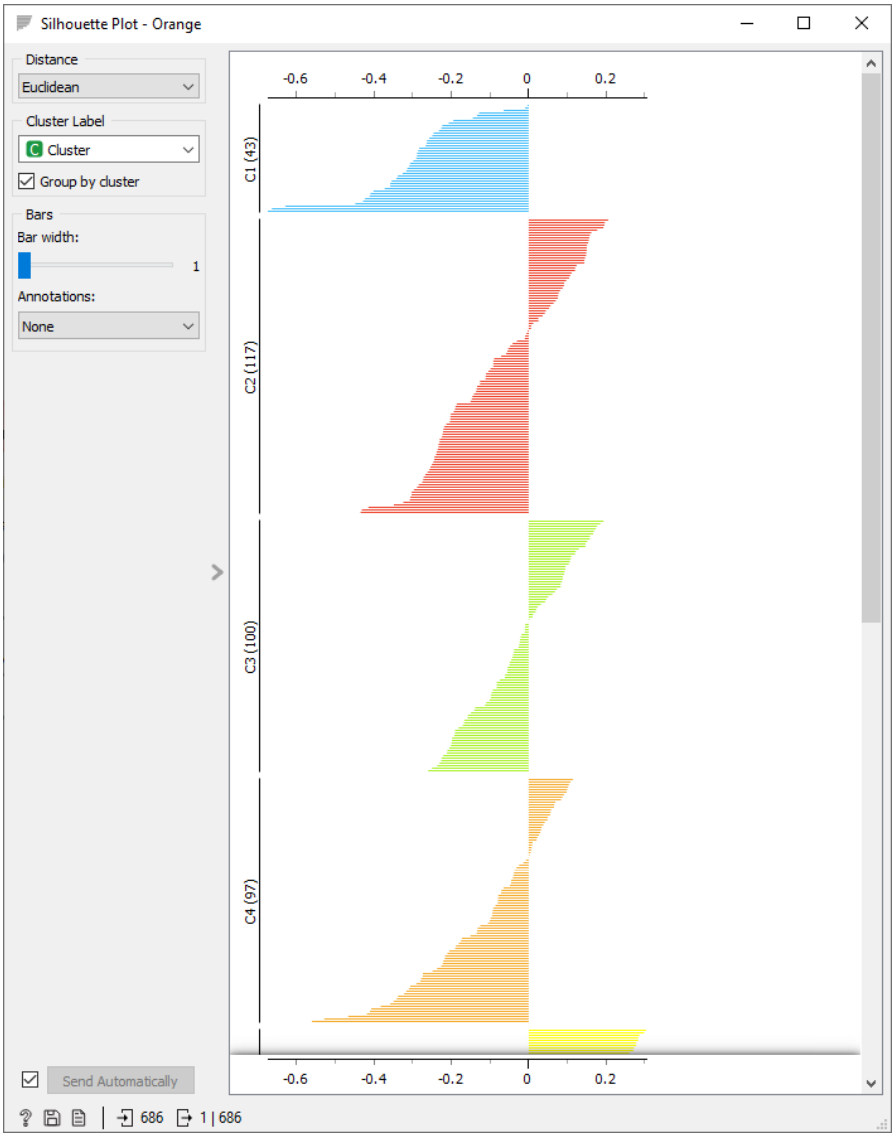
2.8.attēls. 8 klasteri

Izvērtējot iegūto informāciju, ir jānorāda, ka neliels klasteru skaits padara datus labāk pārskatāmus. Tomēr pēc darba autores domām, ir nepieciešams turpināt eksperimentu, izmantot 4 vai 5 klasterus, kā ir parādīts zemāk.



2.9.attēls. 5 klasteri

Pārskatot rezultātus, ir jānorāda, ka, palielinot klasteru skaitu, dati var palikt mazāk pārskatāmi, ko var redzēt 2.10.attēlā. Tāpat ir jānorāda, ka, palielinoties klasteru skaitam, parādās arī vērtējumi ar – zīmi, kas norāda uz to, ka objekti netika pareizi attiecināti uz noteikto klasteri.



2.10.attēls. 10 klasteri

SECINĀJUMI

1. Darba ietvaros tika analizēta un izvērtēta datu kopa, kas tiek darīts jau ne pirmo reizi. Tika izvēlēta laba datu kopa, uz kā var trenēties ne tikai vizualizēt datus, bet arī analizēt to.
2. Pirms darba uzsākšanas bija nepieciešams aplūkot datus, analizēt tos, kā arī sagatavot darbam ar tiem. Pirms algoritmu izmantošanas dati tika aplūkoti un ielādēti programmā, kā arī aprakstīti.
3. Pēc pirmās darba daļas pabeigšanas darba autore var izdarīt sekojošus secinājumus:
    - Klašu datu kopas nav līdzsvarotas;
    - Datu vizualizācija līdz galam neļauj redzēt datu struktūru, jo dati pārklājas;
    - Datu pārklāšanās traucē atdalīt vienu informāciju no otras.
    - Datu objekti nav skaidri atdalāmi.
4. Darbā tika izmantoti nepārraudzītas mašīnmācīšanās algoritmi – hierarhiskā klasterizācija un K-vidējo algoritms. Dati tika sadalīti klasteros. Tika veikti arī eksperimenti ar klasteriem, kā rezultātā parādījās vērtējumi ar – zīmi. Tas nozīmē, ka objekti netika pareizi attiecināti pret klasteriem.
5. Darba autorei ir jānorāda, ka Orange rīku viņa izmantoja jau otro reizi, kā rezultātā pārliecinājās, ka tas ir labs un ērts rīks datu analīzei.