RĪGAS TEHNISKĀ UNIVERSITĀTE

2. praktiskais darbs
Mācību priekšmetā
“Mākslīgā intelekta pamati(1)”
Mašīnmācīšanās algoritmu lietojums
https://github.com/LRSSmeiksts/Maksliga_intelekta_pamati_PD2

Izstrādāja: Lauris Šmeiksts
12. grupa, 211RDB064
Saite uz projektu un datu kopu: Saite
Pārbaudīja:
Asoc.prof. A.Anohina-Naumeca

2022./2023. Māc. gads

Saturs

1 Datu pirmapstrāde/izpēte

1.1 Datu kopas apraksts

Datu kopa "User knowledge modeling data set" tika ievākta 2009. gada semestrī Gazi universitātē, Ankārā, Turcijā. Datu kopa sevī iekļauj mācību priekšmeta "Elektriskās līdzstrāvas iekārtas" studentu zināšanu līmeņu novērtējumus. Datu kopas izveidotāji: Hamdi Tolga Kahraman, Ilhami Colak, Seref Sagiroglu.

Datus apstrādāja datu kopas izveidotāji, Datu kopa tika publicēta 2013. gada 26. jūnijā ar citācijas pieprasījumu: H. T. Kahraman, Sagiroglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, pp. 283-295, 2013.[1]

1.2 Veiktās datu kopas transformācijas

Lai ar datu kopu varētu strādāt Orange rīkā, tā tika transformēta. Pieejamajā xls failā tika izdzēsta "information" lapa, kā arī "Test_data" lapa. Iekš "Training_data" lapas tika izdzēsts teksta lauks ar "Attribute Information", jo tas neļāva Orange rīkā pareizi nolasīt atribūtus. Kā arī tika rediģēti mērķa atribūta vērtības uz skaitliskām vērtībām manuāli."very_low" uz 0; "Low" uz 1; "Middle" uz 2 un "High" uz 3. Pēc šo izmaiņu viekšanas datu kopas kvalitāte netika kompromizēta.

1.3 Datu kopas satura apraksts

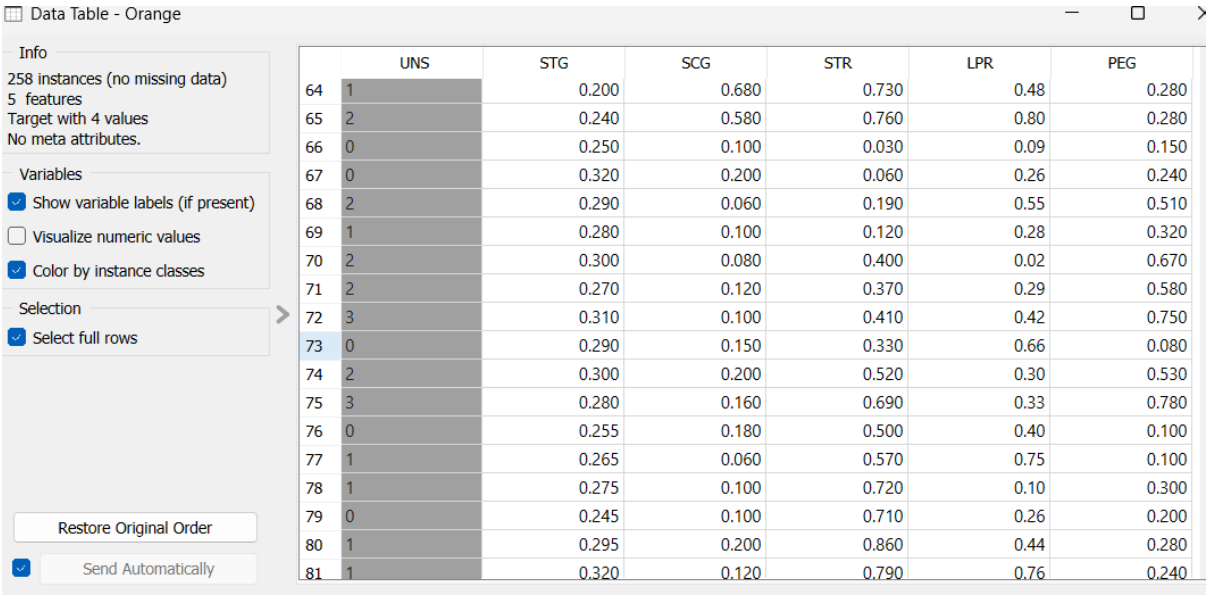
Datu kopa sevī iekļauj 258 datu objektus, un 6 atribūtus, kur dati tiek klasificēti pēc lietotāja zināšanu līmeņa. (skat. 1.tab)[2],(skat. 1.att) .

Atribūts	Nozīme	Diapazons	Loma Orange rīkā
STG	Mācību laika pakāpe priekš mērķa objektu materiāliem	[0.00 – 0.990]	feature
SCG	Lietotāja atkārtojuma pakāpe priekš mērķa objektu materiāliem	[0.00 – 0.900]	feature
STR	Lietotāja izpētes laika pakāpe saistītiem objektiem ar mērķa objektu	[0.00 – 0.950]	feature
LPR	Lietotāja veiktspēja eksāmenā priekš ar mērķi saistītiem objektiem	[0.00 – 0.99]	feature
PEG	Lietotāja veiktspēja eksāmenā priekš mērķa objekta	[0.00 – 0.930]	feature
UNS	Lietotaja zināšanu līmenis	[0 – 3]	target

1. Tab. Atribūti, to raksturojumi

[1] User Knowledge Modeling Data Set. Retreived 2023 May 11. From: https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling

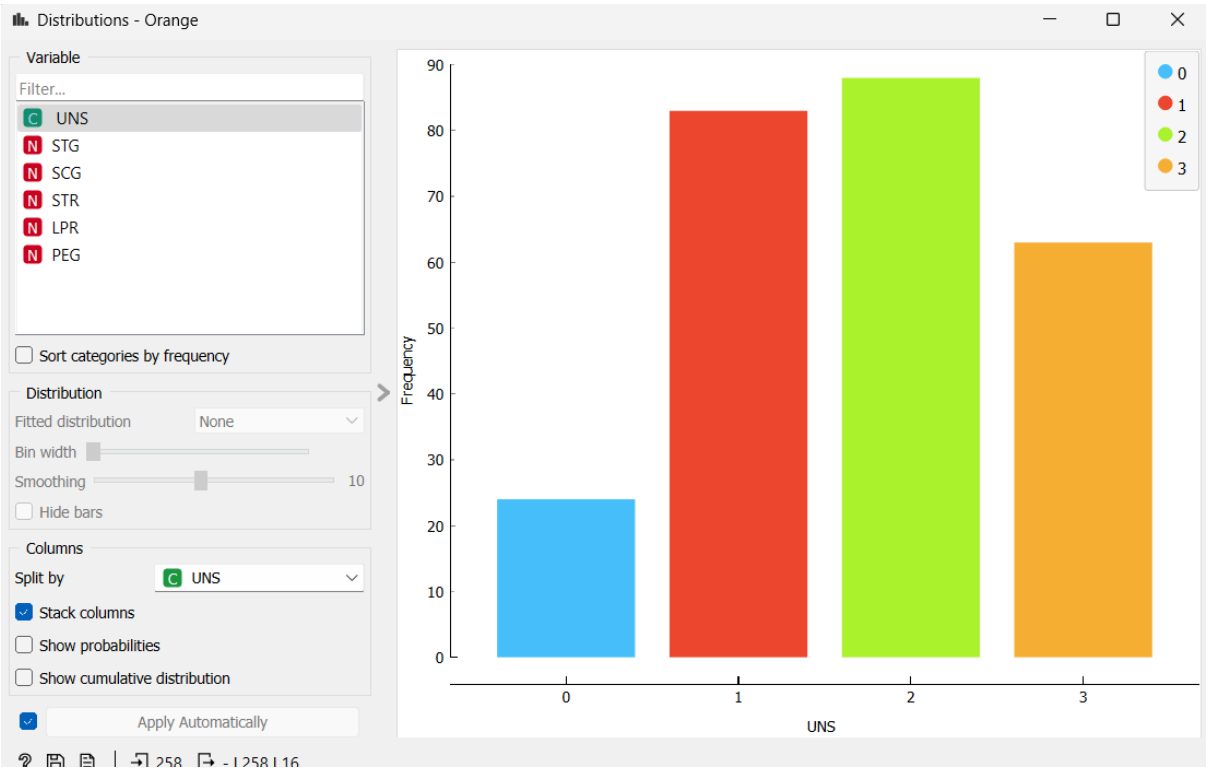
[2] Turpat



	UNS	STG	SCG	STR	LPR	PEG
64	1	0.200	0.680	0.730	0.48	0.280
65	2	0.240	0.580	0.760	0.80	0.280
66	0	0.250	0.100	0.030	0.09	0.150
67	0	0.320	0.200	0.060	0.26	0.240
68	2	0.290	0.060	0.190	0.55	0.510
69	1	0.280	0.100	0.120	0.28	0.320
70	2	0.300	0.080	0.400	0.02	0.670
71	2	0.270	0.120	0.370	0.29	0.580
72	3	0.310	0.100	0.410	0.42	0.750
73	0	0.290	0.150	0.330	0.66	0.080
74	2	0.300	0.200	0.520	0.30	0.530
75	3	0.280	0.160	0.690	0.33	0.780
76	0	0.255	0.180	0.500	0.40	0.100
77	1	0.265	0.060	0.570	0.75	0.100
78	1	0.275	0.100	0.720	0.10	0.300
79	0	0.245	0.100	0.710	0.26	0.200
80	1	0.295	0.200	0.860	0.44	0.280
81	1	0.320	0.120	0.790	0.76	0.240

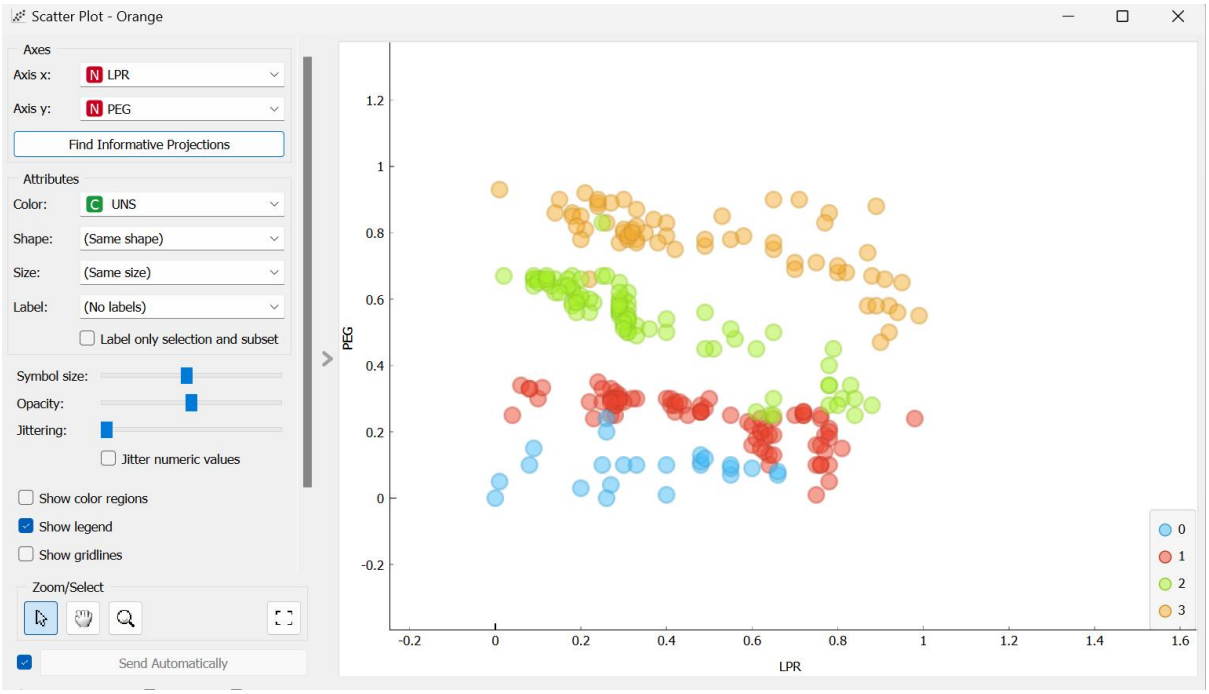
1. Att. Datu faila struktūras fragments

Datu kopā esošie datu objekti pieder pie kādas no mērķa atribūta klasēm: 0 – “very low” ; 1 – “Low” ; 2 – “Middle” un 3 – “High”.(skat. 2.att) Var secināt, ka dati līdzinās normālajam sadalījumam, kur tomēr, “High” ir lielāks pārsvars nekā “very low”, kas arī ir gaidāms, šādā situācijā. Respektīvi pie klases “0” pieder 24 datu objekti; “1” – 83; “2” – 88; “3” – 63 datu objekti.

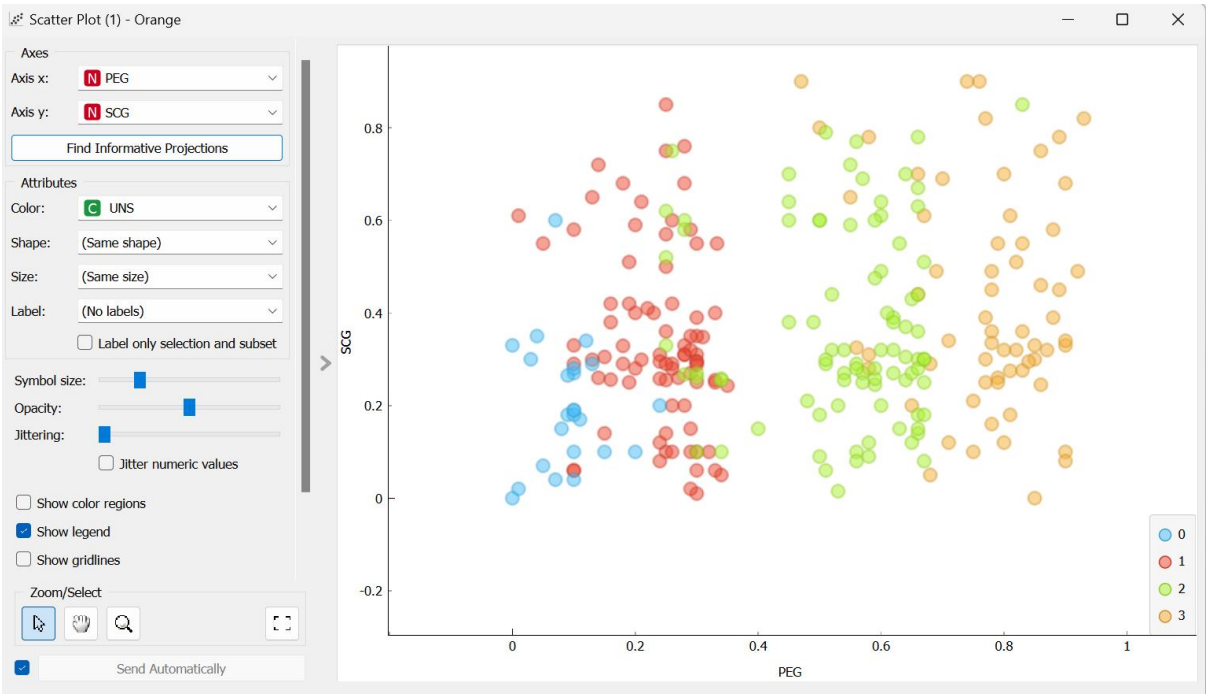


2. Att. Datu objektu sadalījums pa klasēm

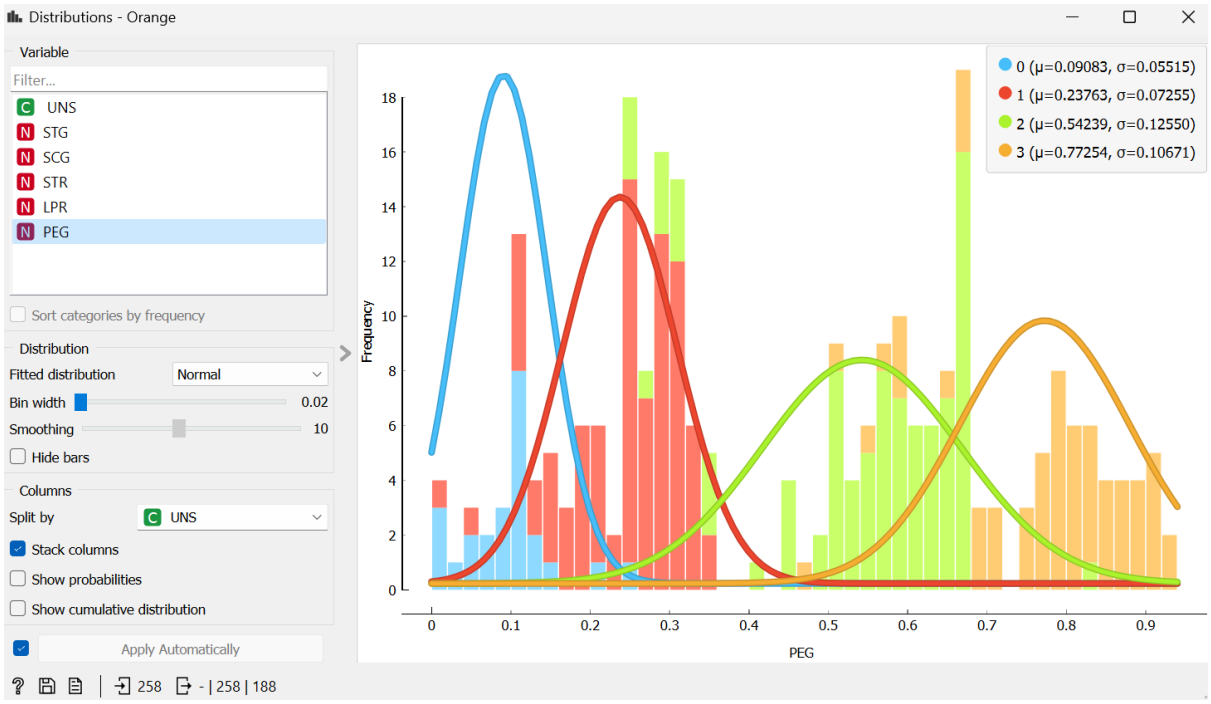
1.4 Datu vizualizācija



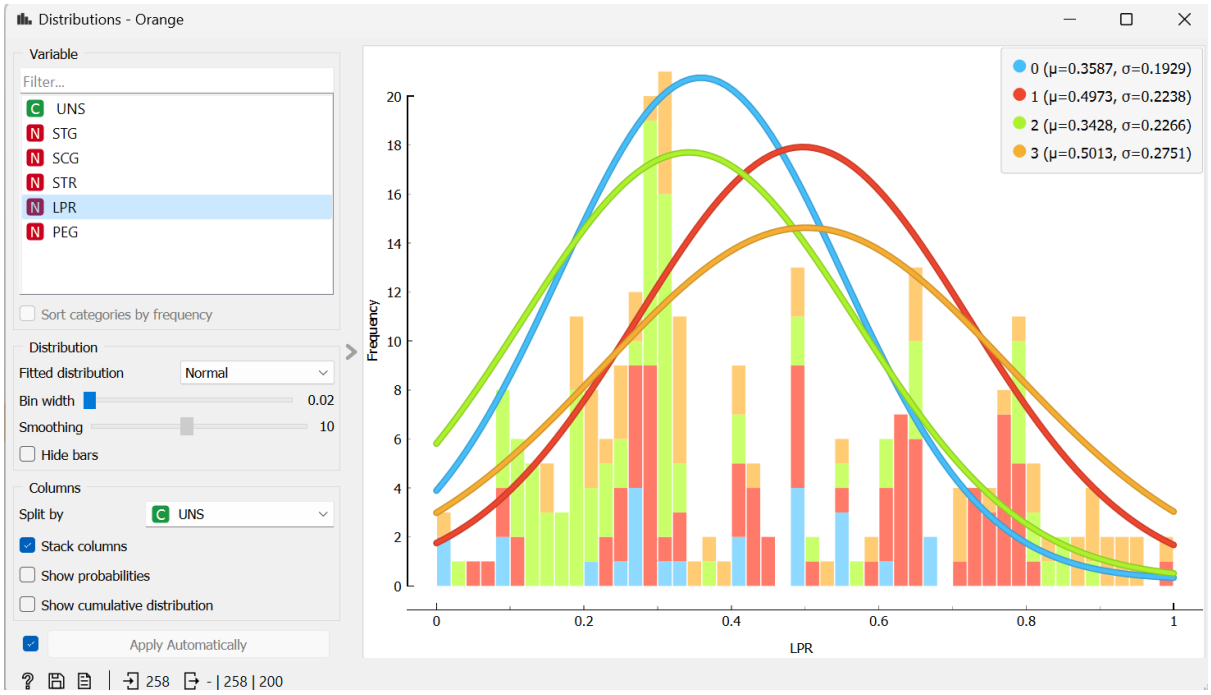
3. Att. Izkliedes dendogramma: (X – LPR; Y – PEG)



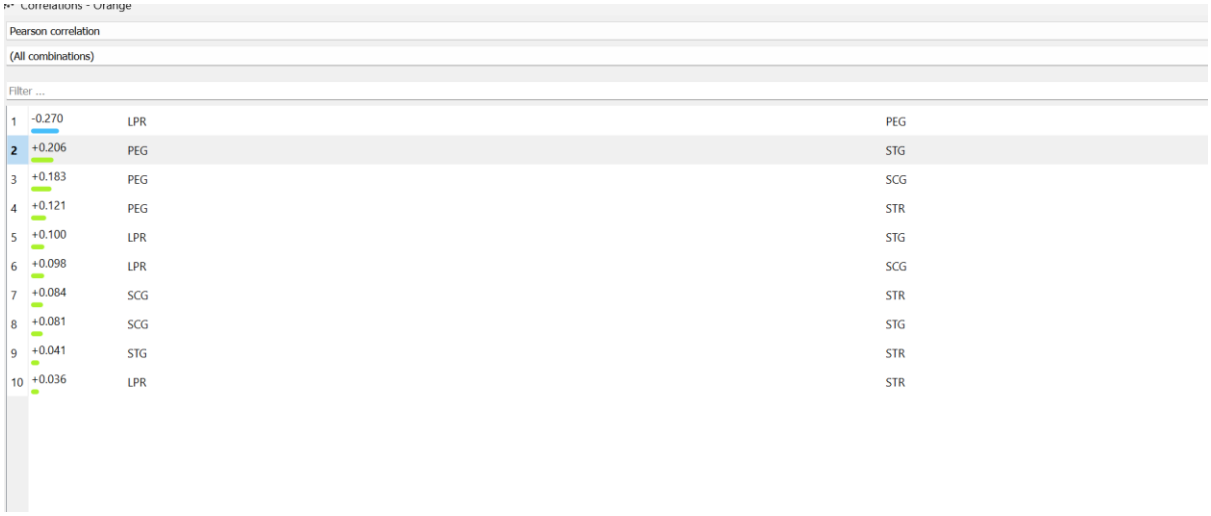
4 Att. Izkliedes dendogramma: (X – PEG; Y – SCG)



5. Att. Histogramma. Klašu atdalīšana pēc “PEG”



6. Att. Histogramma. Klašu atdalīšana pēc “LPR”



7. Att. Korelācijas matrica

8. Att. Klašu statistiskie rādītāji

1.5 Datu izpētes secinājumi

Pēc datu vizualizācijas var secināt, ka klases “1” un “2” ir dominējošās, jo šīm klasēm pieder vairāk datu objektu neka pārējām klasēm

Veicot dendogrammu izpēti, vislabākā atribūtu kombinācijas izrādījās, kur (X – LPR; Y – PEG) (skat. Att 3.) , jo dato objekti ir skaidri atdalāmi, kā arī tie atrodas relatīvi tālu viens no otra, taču ir arī tādi objekti, kuri saplūst ar citām klasēm. Dendogrammā, kur (X – PEG; Y – SCG) arī ir novērojama samērā laba klašu atdalāmība.(skat. Att. 4.)

Analizējot histogrammu, kur klases tiek atdalītas pēc “PEG” (skat. Att. 5.)var novērot skaidru klašu atdalāmību, taču arī ir redzams, ka ir datu objekti, kas “ieplūst” arī citās klasēs, kas iespējams var sagātāt sarežģījumus turpmākā mašīnmācīšanās modelī. Savukārt, atdalot klases pēc atribūta “LPR” (skat. Att. 6.) klašu atdalāmība vairs nav tik skaidri redzama.

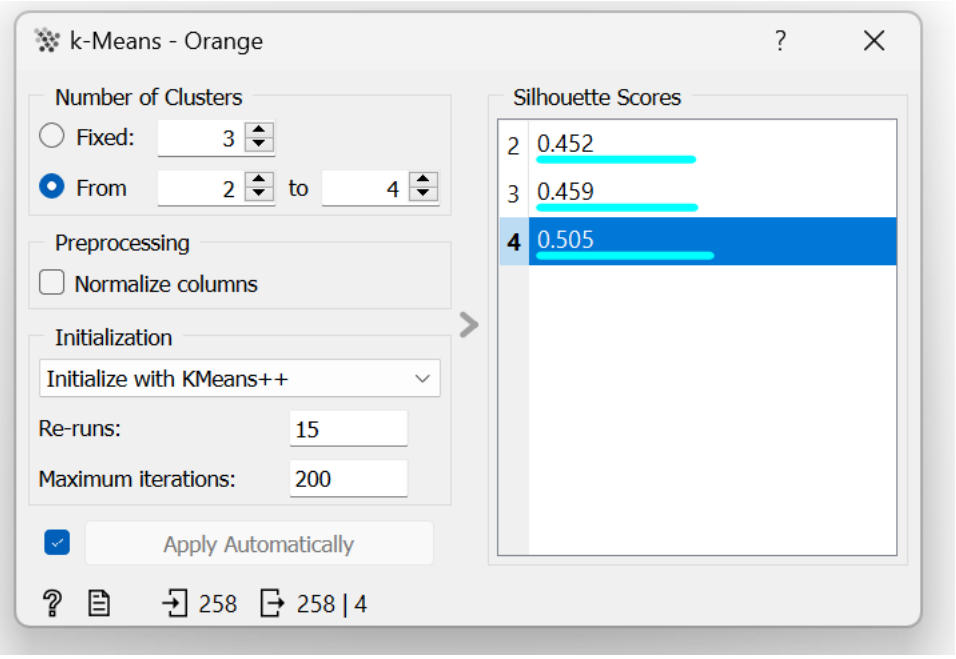
Attēlā nr. 7 var novērot, ka “PEG” un “STG” savā starpā vislabāk korelē, kas nozīmē, ka šīs pazīmes varēs turpmāk izmantot klasifikācijas uzdevumiem.

Visbeidzot, pētot klašu statistiskos rādītājus, var secināt, ka atribūtam “STR” ir vismazākā dispersija, kas liecina par to, ka datu objekti ir vairāk sakoncentrēti pie vidējās vērtības, respektīvi, tos ir grūtāk atdalīt vienu no otra.

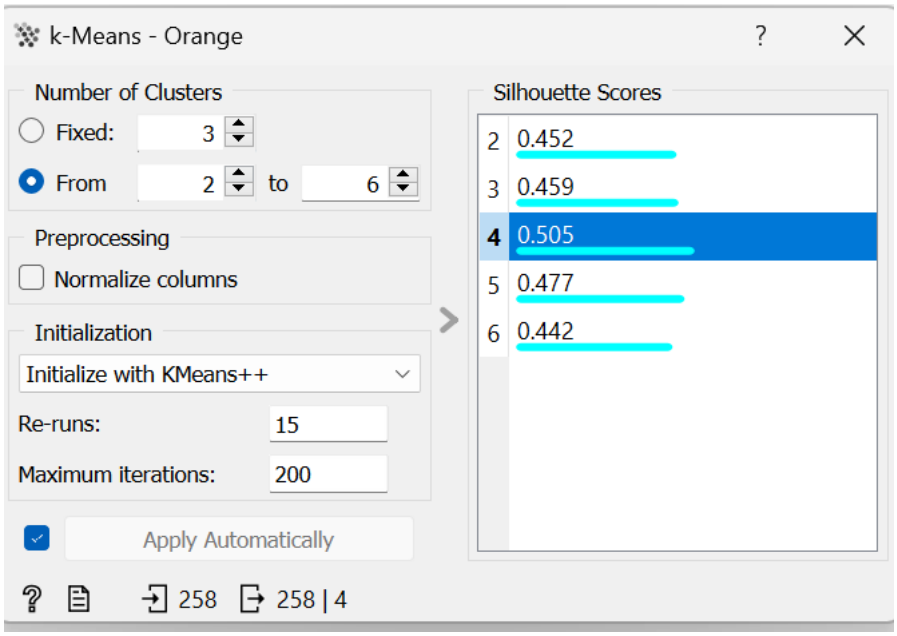
2 Nepārraudzīta mašīnmācīšnās

Ņemot vērā iepriekšējo datu analīzi, tika izmantoti tikai “PEG”, “LPR” un “UNS” atribūti, savukārt pārējie tika atmesti

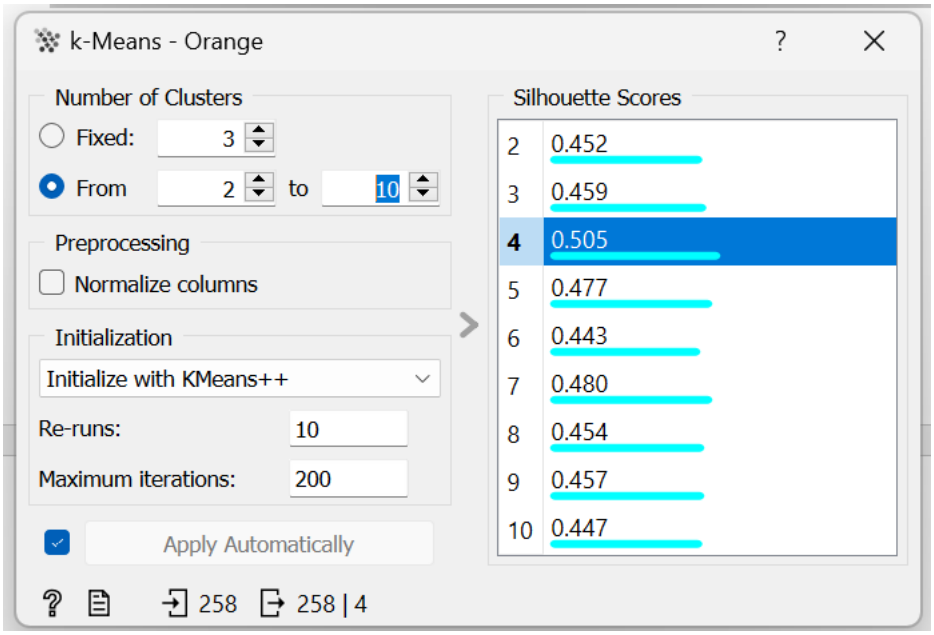
2.1 K-vidējo algoritms



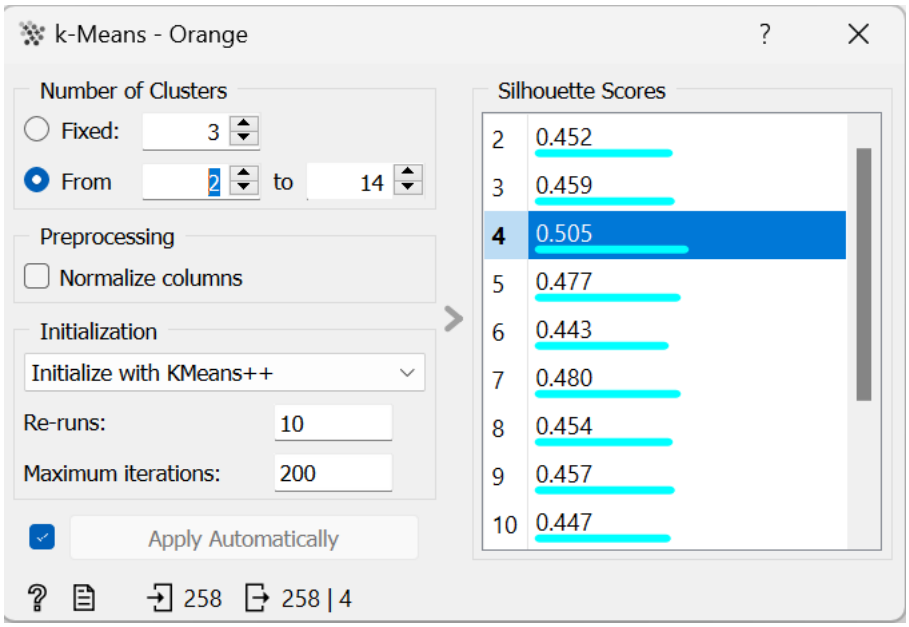
9. Att. “K-means”, klasteru skaits: 4



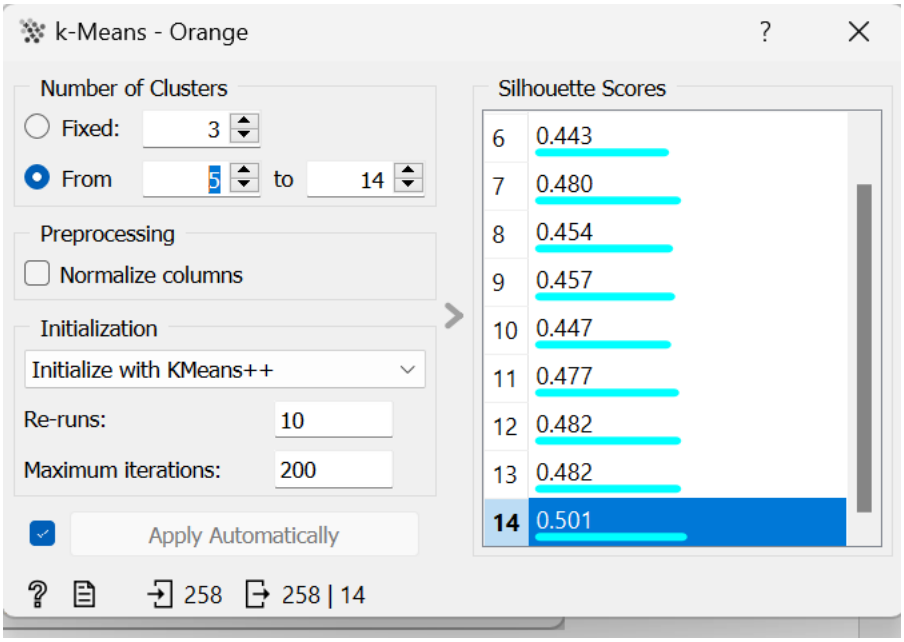
10. Att. “K-means”, klasteru skaits: 6



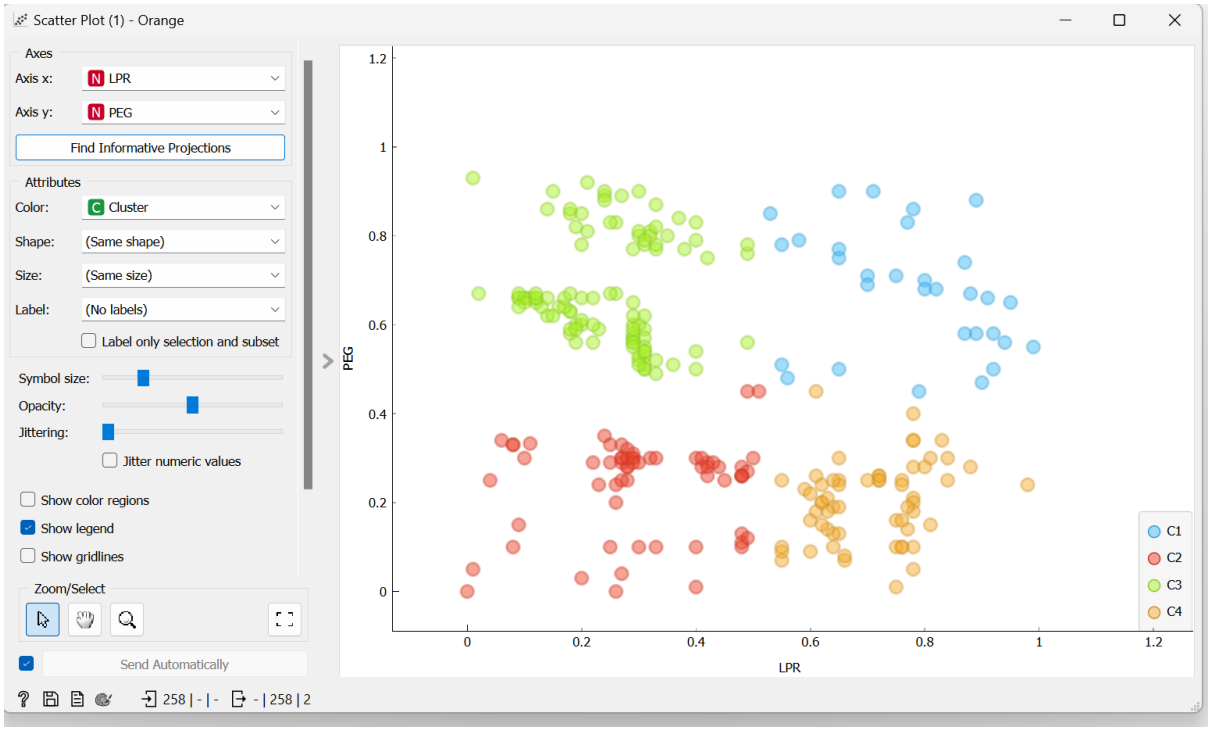
11. Att. “K-means”, klasteru skaits 10



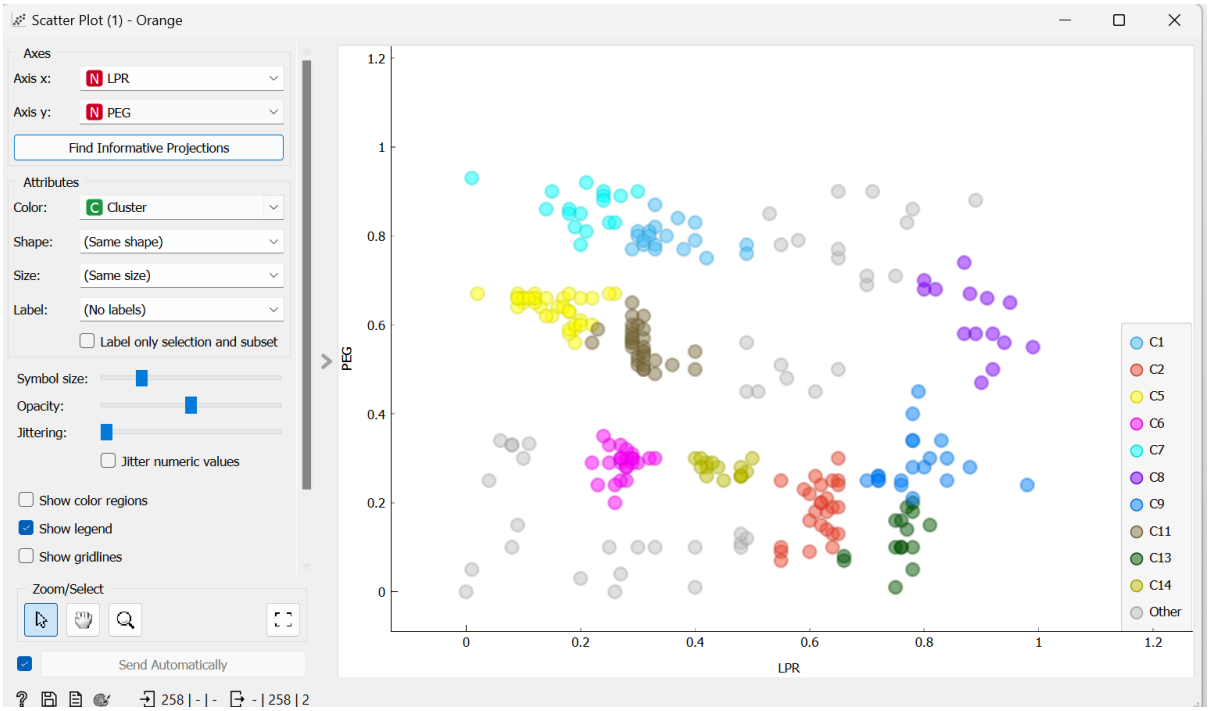
12. Att. “K-means”, klasteru skaits 14



13. Att. “K-means” min. klasteru skaits- 5; max. klasteru skaits – 14



14. Att. Izkliedes diagramma, (X- LPR; Y- PEG) klasteru sadalījums



15.Att. Izkliedes diagramma, (X- LPR; Y- PEG) 14 klasteru sadalījums

Eksperimentējot ar K-means algoritmu var novērot, ka maksimālo klasteru skaita palielināšana īpaši neietekmēja rezultātu, jo tika atrasts optimālais klasteru skaits, kuram jau bija lielākais silhouette koeficents. Skat.(Att. 9.,10.,11.,12). Tāpēc, eksperimenta nolūkos tika palielināts minimālais klasteru skaits līdz 5. Līdz ar to varēja novērot, ka daudzi datu objekti nemaz netika iekļauti klasteros. Skat.(Att. 15.)

Izmainot Re-runs vērtību netika manīta būtiska atšķirība, ja tas jau bija lielāks par 10.

14.Attēlā ir redzams K-means rezultāts pie k = 14, var secināt, lai gan datu objekti ir sadalīti pa klasteriem, tie ir samērā tālu viens no otra, līdz ar to silhuette score ir tik zems.

K- means tika arī testēts ar citiem atributiem, taču tur silhuette score nepārsniedza 0.4 un nedeva vēlamo rezultātu

Vērtības netika normalizētas, jo tas jau bija izdarīts iepriekš.

Pēc testēšanas var secināt, ka maksimālais klasteru skaits ir vissvarīgākais hiperparametrs, jo tas rezultātus ietekmēja visvairāk.

Silhouette koeficents – mērījums, kas parāda, cik līdzīgs datu punkts ir klasterī salīdzinājumā ar citām kopām[3]

Orange rīkā ir iespējams izvēlēties vai klasteru skaits būs fiksēts, vai diapazonā

Fiksēts – algoritms sagrupē datus noteiktam klasteru skaitam;

Diapazonā- logrīks parāda klasterizācijas rezultātus atlasītajam klasteru diapazonam, izmantojot Silhouette koeficentu.

Ir arī iespējami divi inicializācijas veidi:

- KMeans++ - pirmais centroids tiek izvietots nejauši, taču nākamie tiek izvēlēti no atlikušajiem punktiem ar varbūtību, kas proporcionāla kvadrātiskajam attālumam no tuvākā centra
- Nejauša incializācija – Klasteri tiek piešķirti nejauši;

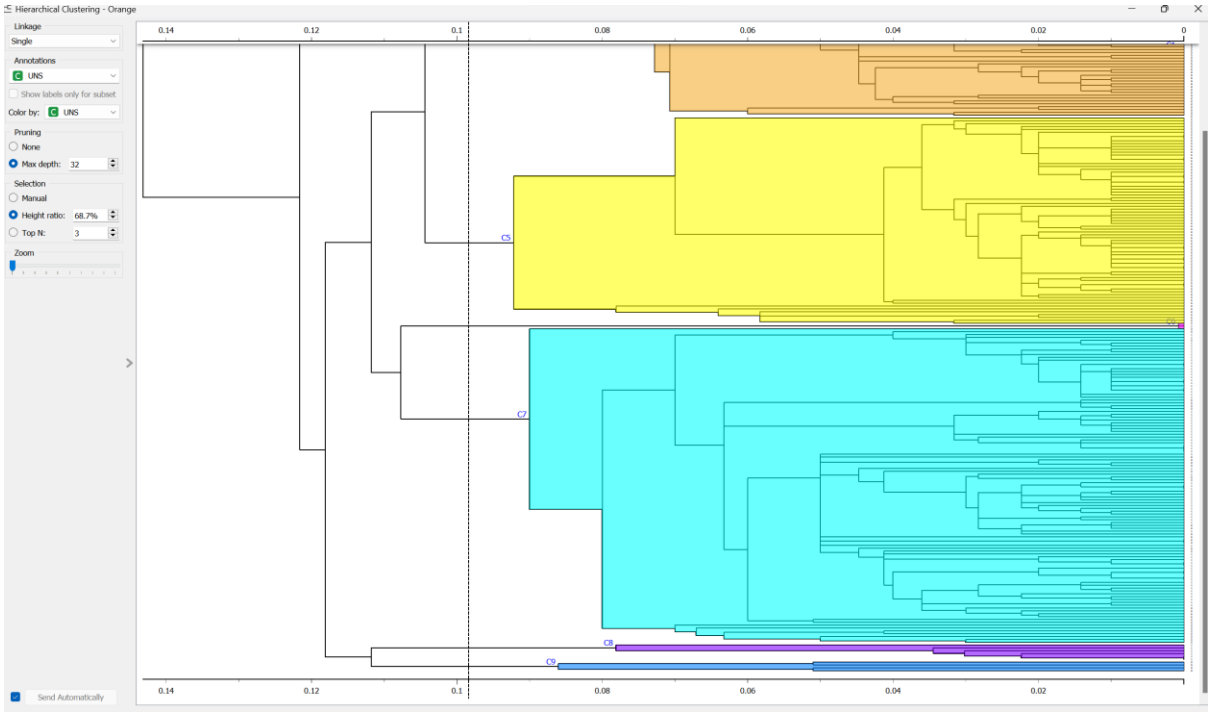
Atkārtotās izpildes – šajā parametrā tiek norādīts, cik reižu algoritms tiks palaists no nejaušām sākotnējām pozīcijām

Maksimālās izpildes – šajā parametrā tiek norādīts maksimālais iterāciju skaits.[4]

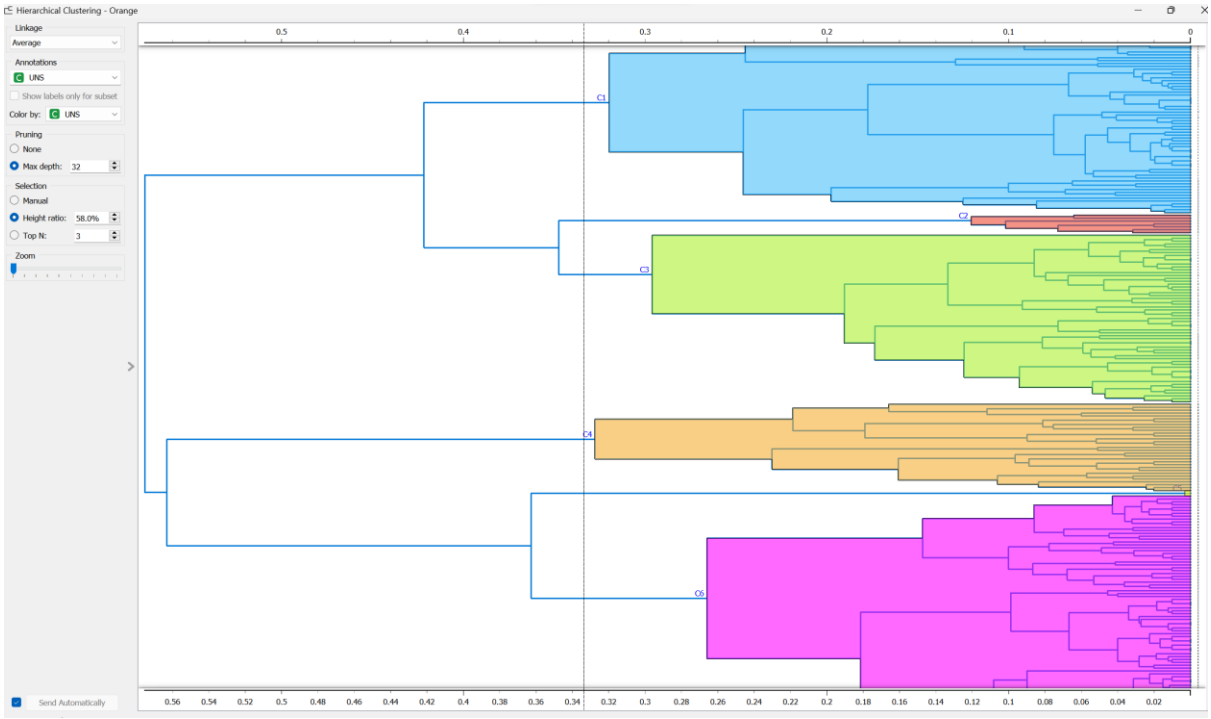
[3] K-Mean: Getting the Optimal Number Of Clusters. Retreived 2023 May 12. From: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=The%20silhouette%20coefficient%20or%20silhouette,scikit%2Dlearn%2Fsklearn%20library.

[4] K-Mean, Retreived 2023 May 12. From: https://orangedatamining.com/widget-catalog/unsupervised/kmeans/

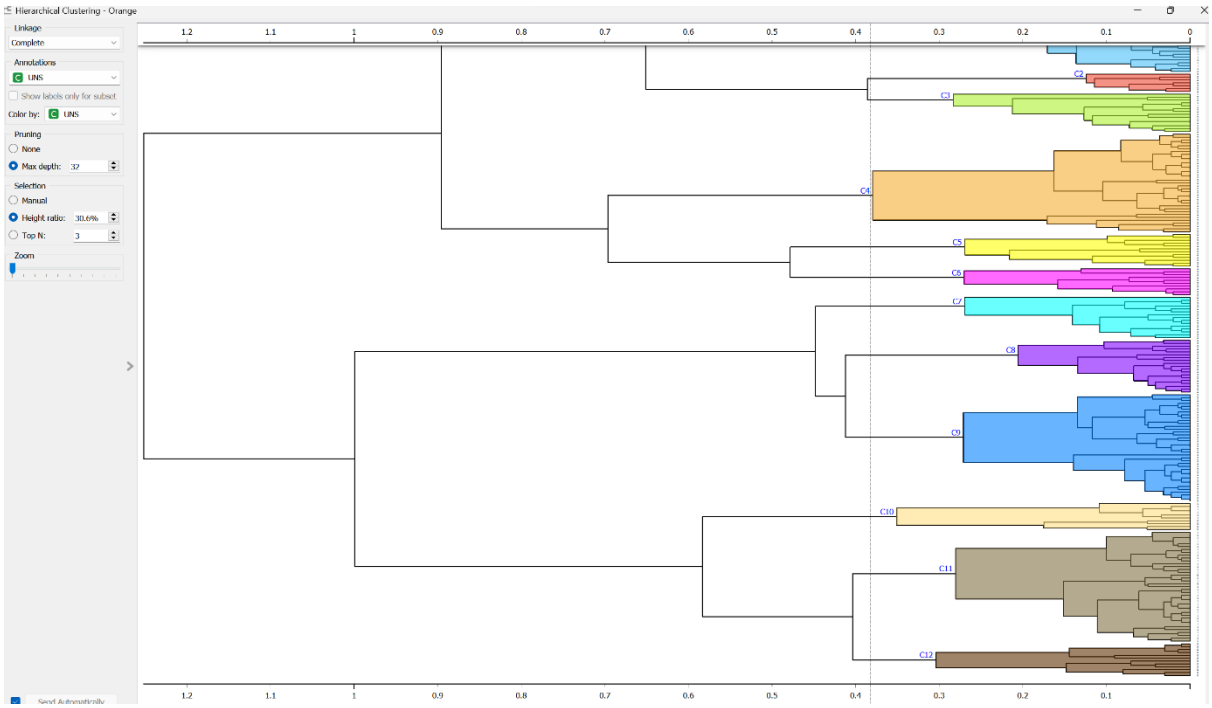
2.2 Hierarhiskās klasterizācijas algoritms



15. Att. Dendogramma “Single” saistīšanas metode



16. Att. Dendogramma “Average” saistīšanas metode



17. Att. Dendogramma "Complete" saistīšanas metode

Pirms tika īstenota heiristiskā klasterizācija, bija nepieciešams izveidot attālumu matricu, kas aprēķina attālumu starp datu objektu aprēķiniem.[5] Iekš kuras bija iespējams izvēlēties distances metriku. Tika izvēlēts Eiklīda attālums. Bija arī iespējams normalizēt datus, taču ši opcija netika izmantota, jo dati jau tika normalizēti datu pirmastrādē.

Pēc šī soļa izpildes tika izmantota heiristiskās klasterēšanas opcija.

Pētījuma ietvaros tika konfigurēts Linkage hiperparamets.

"Single" saistīšanas metodē "Height ratio" tika uzstādīts uz 68,7%., lai gan rezultātā iznāca 9 klasteri, datu objekti bija diezgan precīzi sadalīti. Taču vietām bija arī nepareizs datu objekts(skat. 15. Att.)

"Average" saistīšanas metode bija salīdzinoši neprecīza, palielināt "Height ratio" vērtību nebīja izdevīgi, jo palielinoties klasteru skaitam precizitāte nepaaugstinājās proporcionāli.(Skat. 16. Att)

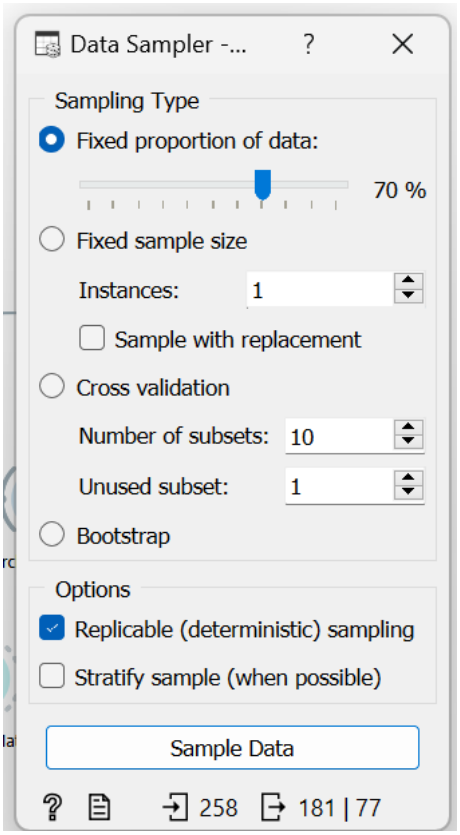
No visām metodēm "Complete" saistīšans metode bija visneprecīzākā, pat uzstādot "Height ratio" vērtību uz 30,14%, bija daudz datu objektu, kas atradās nepareizos klasteros(Skat. 17. Att)

Var secināt, ka Heiristiskās klasifikācijas algoritms šim uzdevumam bija nepiemerots, tas ir iespējams, jo jau datu izpētes etapā bija novērojams, ka daudziem atribūtiem bija diezgan līdzīgas un zemas dispersijas, kas nozīmē datu koncentrēšanos pie vidējās vērtības.

[5] Example distance Retreived 2023 May 12 From: http://docs.biolab.si/orange/2/widgets/rst/unsupervized/exampledistance.html

3 Pārraudzītā mašīnmācīšanās

3.1 Testa un apmācību kopa



18. Att apmācības un testa kopas izveide

Turpmākajai algoritmu izpētei tika izveidota datu apmācības kopa, kas ir 70% un testa datu kopa, kas ir 30%

3.2 Mākslīgo neironu tīkls

Turpmākajos eksperimentos tiks izmēģinātas dažādas hiperparametru variācijas;
Tiks konfigurēts "Activation" hiperparametrs, tas iekļauj sevī dažādas aktivizācijas funkcijas priekš slēptā slāņa:

- Identitāte: becoperācijas aktivizēšana, noderīga, lai īstenotu lineāro sašaurinājumu
- Loģisktika: loģiskā sigmoida funkcija;
- Tanh: hiperboliskā iedeguma funkcija;
- ReLu: rektificētas lineārās vienības funkcija.

Kā arī "Solver"hiperparametrs, kas regulē svaru optimizāciju:

- L-BFGS-B: optimizētājs kvaziņūtona metožu saimē;
- SGD: stohastiskā gradienta nolaišanās;
- Adam: uz stohastisku gradientu balstīts optimizētājs[6]

3.2.1 Eksperiments Nr. 1

Pirmā eksperimenta ietvaros, parametrs "Activation" tika konfigurēts uz "Logistic", savukārt "Solver" uz "SGD"

		Predicted			
		0	**1**	**2**	**3**	**∑**
Actual	**0**	NA	NA	5.2 %	NA	**4**
	1	NA	NA	36.4 %	NA	**28**
	2	NA	NA	33.8 %	NA	**26**
	3	NA	NA	24.7 %	NA	**19**
	∑	**0**	**0**	**77**	**0**	**77**

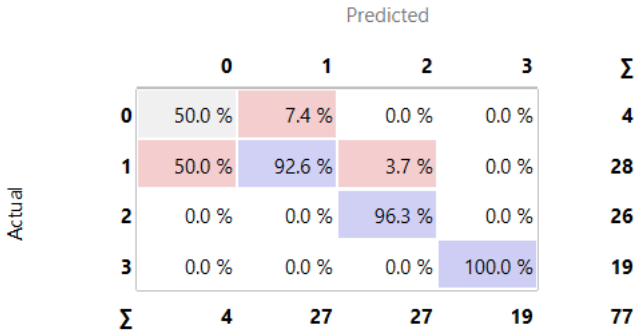
19.Att. 1. Eksperimenta rezultāti

Precizitāte: 0.114;
F1: 0.170

[6] Neural Network Retreived 2023 May 12 From: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/neuralnetwork.html

3.2.2 Eksperiments Nr. 2

Otrā eksperimenta ietvaros, parametrs “Activation” tika konfigurēts uz “tanh”, savukārt “Solver” uz “L-BFGS-B”

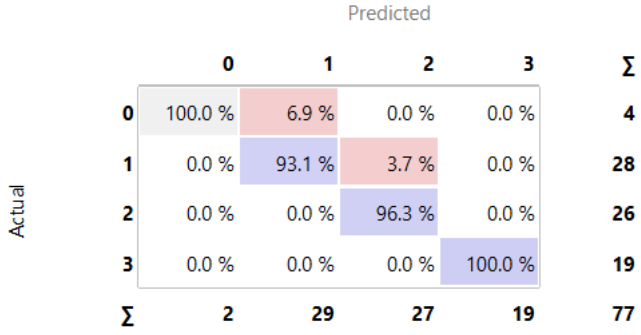


20.Att. 2. Eksperimenta rezultāti

Precizitāte: 0.935
F1: 0.935

3.2.3 Eksperiments Nr.3

Otrā eksperimenta ietvaros, parametrs “Activation” tika konfigurēts uz “ReLu”, savukārt “Solver” uz “Adam”



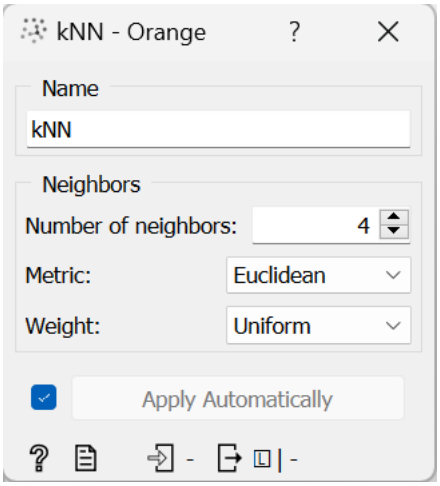
21.Att. 3. Eksperimenta rezultāti

Precizitāte: 0.962
F1: 0.957

3.2.4 Eksperimentu apkopojums

No visām testētajām variācijām, labākos rezultātus uzrādija 3. Eksperimenta konfigurācija. Vissliktākos rezultātus uzrādija “Logistic” un “SGD” savienojums.

3.3 kNN algoritms



22.Att. kNN algoritma parametri

kNN algoritmam piemīt 3 hiperparametri:

- Kaimiņu skaits – tuvāko kaimiņu skaits
- Metrika:
  - Eiklīds – attālums starp diviem punktiem;
  - Manhetena – visu atribūtu absolūto atšķirību summa;
  - Maksimāls – lielākā absolūtā atšķirība starp atribūtiem;
  - Mahalonobis – attālums starp punktu un sadalījumu.

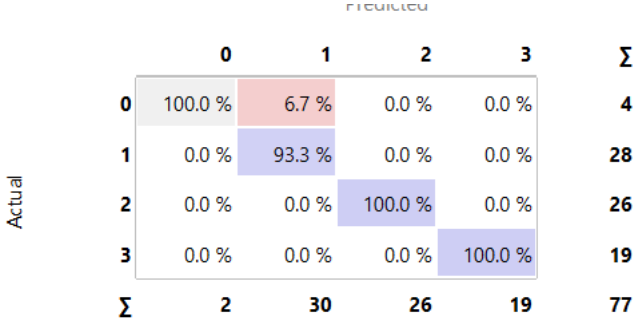
- Svari:
  - Vienots – visi punkti katrā apkaimē tiek svērti vienādi;
  - Attālums – tuvākiem vaicājuma punkta kaimiņiem ir lielāka ietekme nekā tālākiem kaimiņiem.[7]

Turpmākajos eksperimentos tiks izmēģinātas dažādas hiperparametru variācijas; Tiks konfigurēts “Metric” un “Weight” hiperparametri

[7] kNN. Retreived 2023 May 12 From: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html

3.3.1 Eksperiments Nr. 1

Pirmā eksperimenta ietvaros, parametrs “Metric” tika konfigurēts uz “Euclidean”, savukārt “Weight” uz “Uniform”

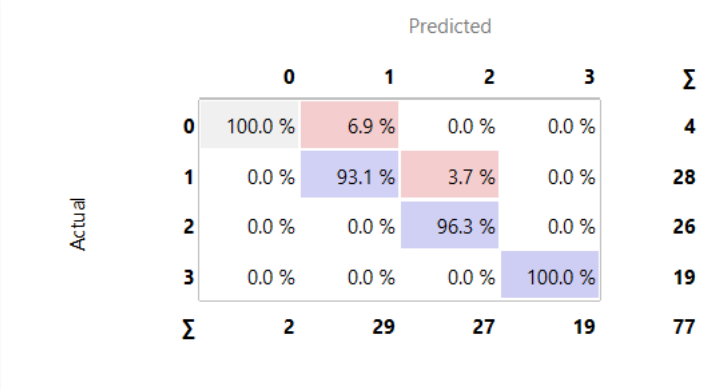


23.Att. kNN 1. Eksperimenta rezultāti

Precizitāte: 0.976
F1: 0.970

3.3.2 Eksperiments Nr. 2

Otrā eksperimenta ietvaros, parametrs “Metric” tika konfigurēts uz “Manhattan”, savukārt “Weight” uz “Distance”

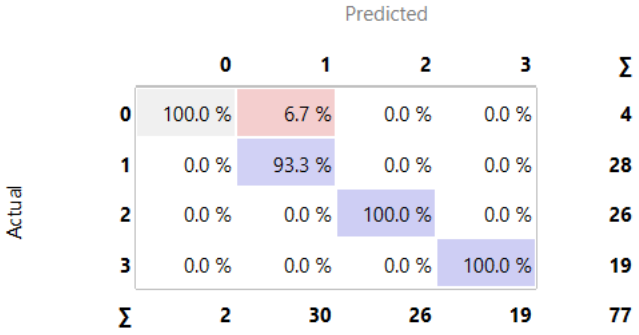


24.Att. kNN 2. Eksperimenta rezultāti

Precizitāte: 0.962
F1: 0.957

3.3.3 Eksperiments Nr. 3

Trešā eksperimenta ietvaros, parametrs “Metric” tika konfigurēts uz “Mahalanobis”, savukārt “Weight” uz “Uniform”



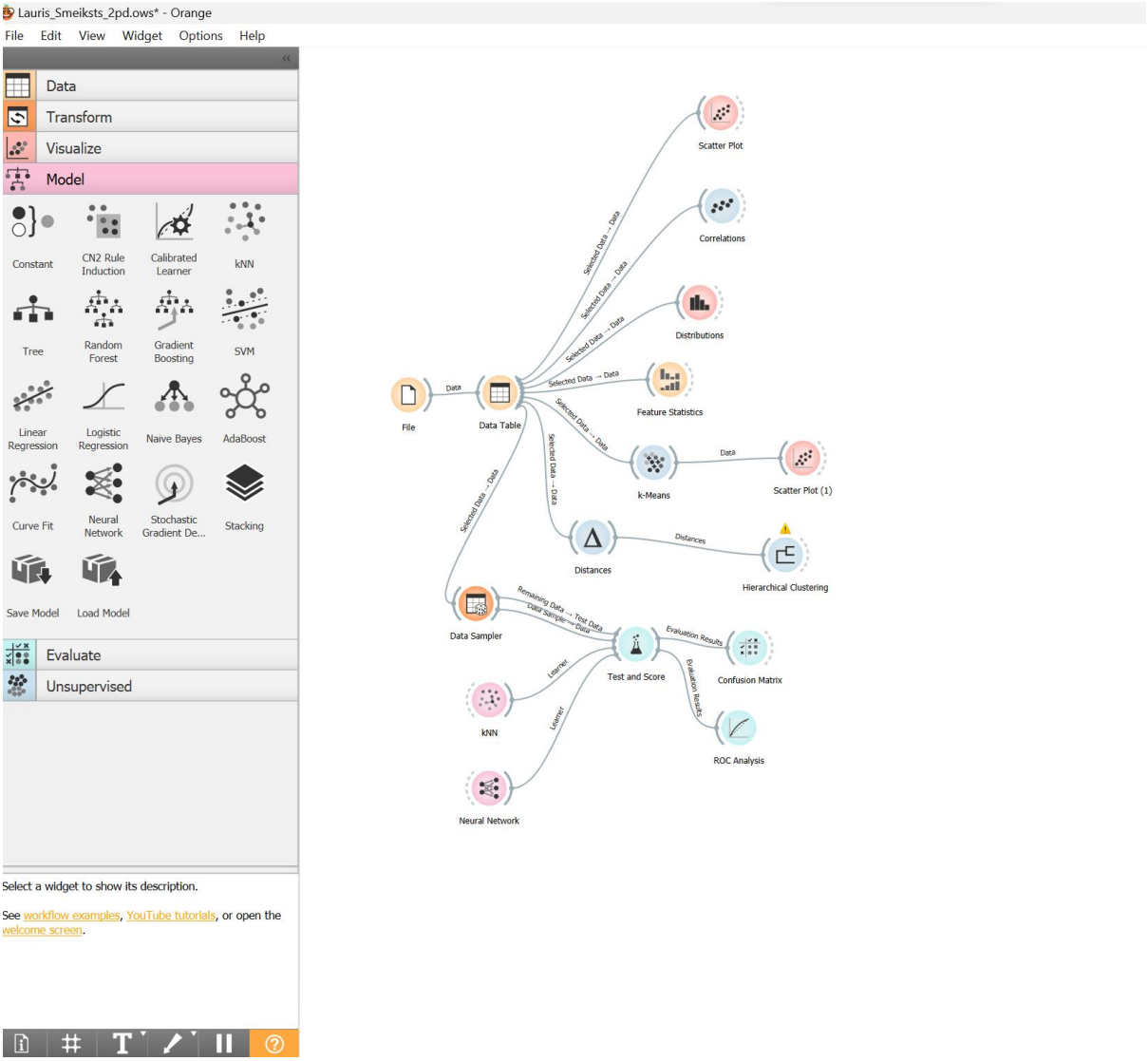
25. Att. kNN 3. Eksperimenta rezultāti

Precizitāte: 0.976
F1: 0.970

3.3.4 kNN Eksperimentu apkopojums

Savā starpā rezultāti bija ļoti līdzīgi. Precizitāte un F1 gandrīz nemainījās, 1. un 3. eksperimentam rezultāti bija identiski

4 Orange darbaplūsmas atspoguļojums



26.Att. Orange darbaplūsmas atspoguļojums

5 Atsauces un izmantotā literatūra

1. User Knowledge Modeling Data Set. Retreived 2023 May 11. From: https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling
2. K-Mean: Getting the Optimal Number Of Clusters. Retreived 2023 May 12. From: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=The%20silhouette%20coefficient%20or%20silhouette,scikit%2Dlearn%2Fsklearn%20library.
3. K-Mean. Retreived 2023 May 12. From: https://orangedatamining.com/widget-catalog/unsupervised/kmeans/
4. Example distance Retreived 2023 May 12 From: http://docs.biolab.si/orange/2/widgets/rst/unsupervized/exampledistance.html
5. Neural Network Retreived 2023 May 12 From: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/neuralnetwork.html
6. kNN. Retreived 2023 May 12 From: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html