Rīgas tehniskā universitāte

Mākslīgie intelekta pamati

2.praktiskais darbs

https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download

https://github.com/DZUDMENS/PD2.git

Izstrādāja: Gvido Urniežus

211RDB436 6.grupa

Saturs

1 daļa – Datu pirmapstrāde/izpēte

Lai izpildītu šī darba daļu, studentiem ir jāveic šādas darbības:
1. Ir jāizvēlas un jāapraksta datu kopa, pamatojoties uz informāciju, kas sniegta krātuvē, kurā datu kopa ir pieejama.

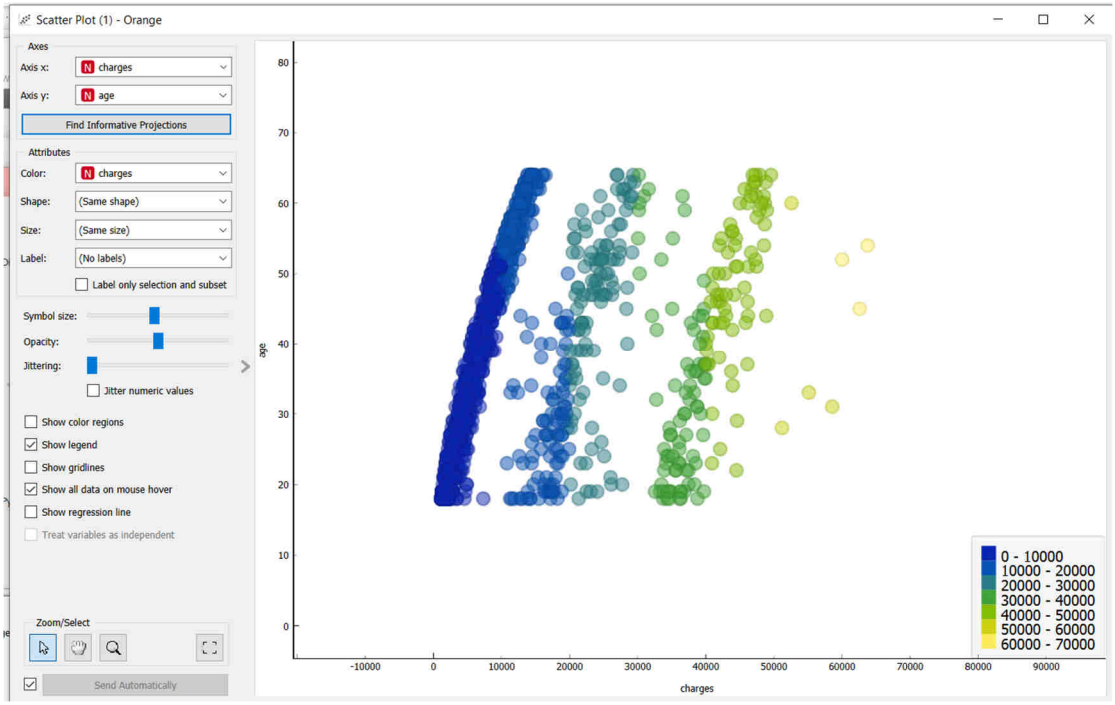
2. Ja no krātuves iegūtā datu kopa nav formātā, ar kuru ir viegli strādāt (piemēram, komatatdalītās vērtības vai .csv fails), ir jāveic tās transformācija vajadzīgajā formātā.

3. Ja kādu pazīmju (atribūtu) vērtības ir tekstveida vērtības (piemēram, yes/no, positive/neutral/negative, u.c.), tās ir jātransformē skaitliskās vērtībās.

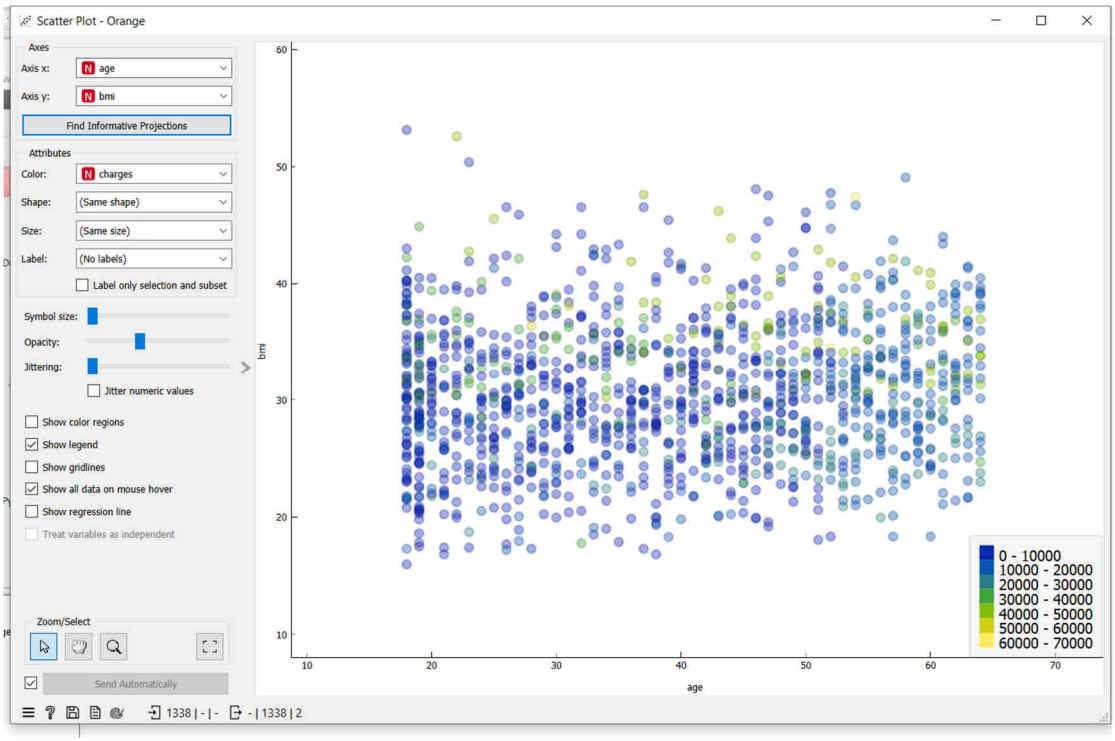
4. Ja kādiem datu objektiem trūkst atsevišķu pazīmju (atribūtu) vērtības, ir jāatrod veids, kā tās iegūt, studējot papildu informācijas avotus.

5. Ir jāatspoguļo datu kopa vizuāli un jāaprēķina statistiskie rādītāji:

a) ir jāizveido vismaz divas 2- vai 3-dimensiju izkliedes diagrammas (scatter plot), kas ilustrē klases atdalāmību, balstoties uz dažādām pazīmēm (atribūtiem); studentam ir jāizvairās izmantot datu objekta ID vai klases iezīmi kā mainīgo izkliedes diagrammā;

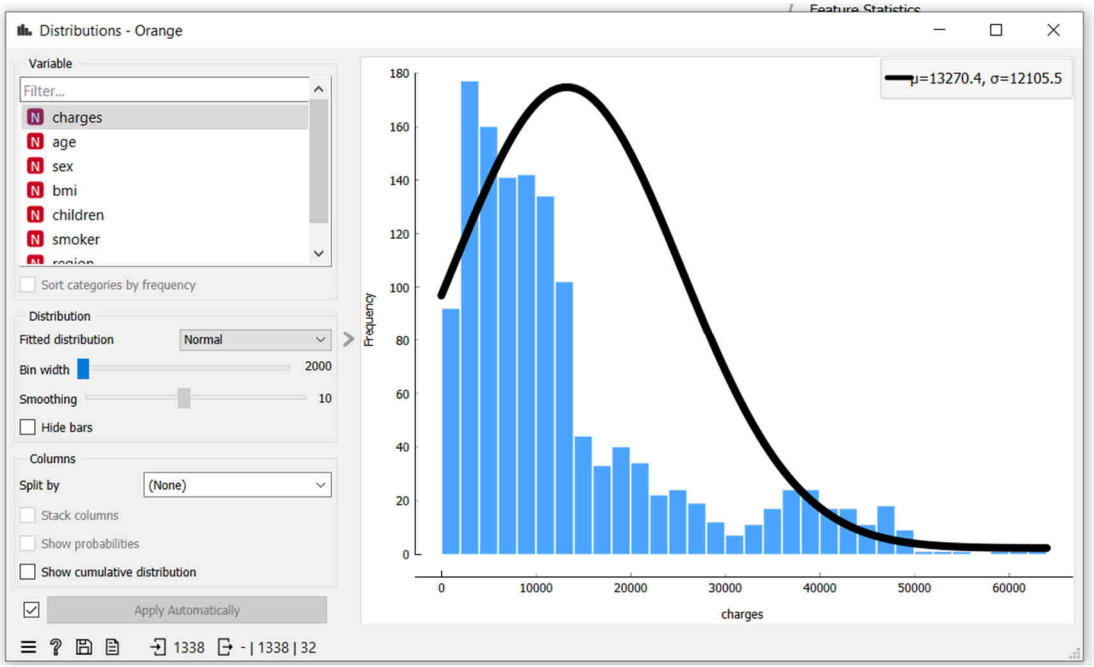


(1.att)

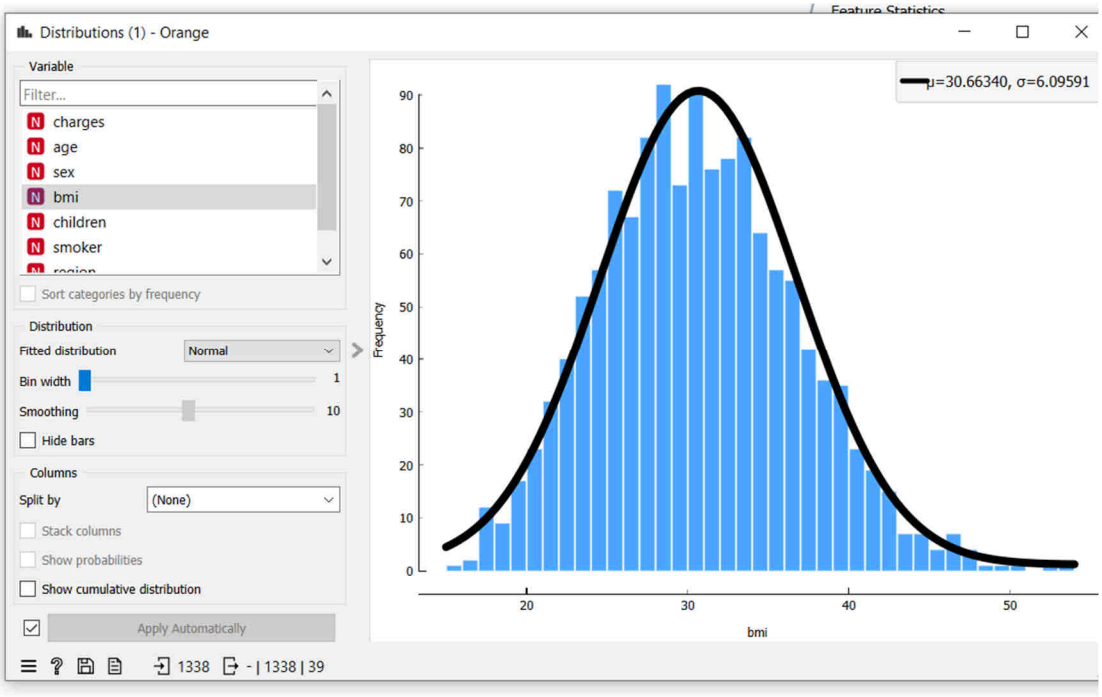


(2.att)

b) ir jāizveido vismaz 2 histogrammas, kas parāda klašu atdalīšanu, pamatojoties uz interesējošām pazīmēm (atribūtiem);

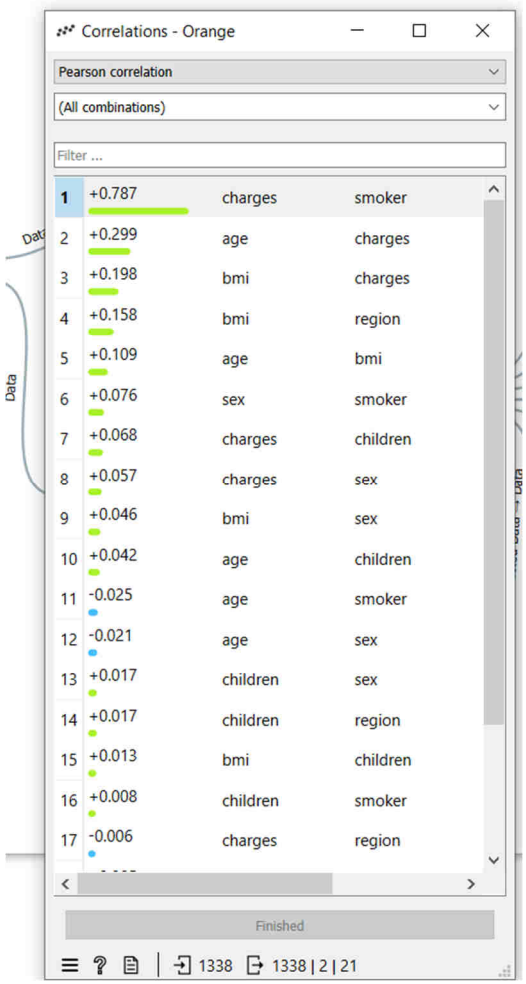


(3.att)



(4.att)

c) ir jāatspoguļo 2 interesējošo pazīmju (atribūtu) sadalījums;



(5.att)

d) ir jāaprēķina statistiskie rādītāji (vismaz vidējās vērtības un dispersiju).

Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
age		39.21	18	39	0.36	18	64	0 (0 %)
sex		0.505232	1	1	0.989591	0.00	1	0 (0 %)
bmi		30.66340	32.300	30.400	0.19880	15.960	53.130	0 (0 %)
children		1.09	0	1	1.10	0	5	0 (0 %)
smoker		0.204783	0.00	0.00	1.97059	0.00	1	0 (0 %)
region		1.5157	2	2	0.72869	0.00	3	0 (0 %)
charges		13270.4	1639.56	9382.03	0.912216	1121.87	63770.4	0 (0 %)

(6.att)paskaidrojumu skatīties pie 9.att.

1.Daļas atskaite

Datu kopas apraksts (sniedzot arī atsauces izmantotajiem informācijas avotiem)

- Datu kopas nosaukums, avots, izveidotājs un/vai īpašnieks;

  Nosaukums-US Health Insurance Dataset
  Avots - https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download
  Izveidotājs-Anirban Datta

- Datu kopas problēmsfēras apraksts;

  Šī datu kopa var būt noderīga vienkāršā, bet izgaismojošā pētījumā, lai izprastu riska parakstīšanu veselības apdrošināšanā, dažādu apdrošinātā atribūtu mijiedarbību un noskaidrotu, kā tie ietekmē apdrošināšanas prēmiju.

- Datu kopas licencēšanas nosaucījumi (jā tādi ir);

  https://creativecommons.org/publicdomain/zero/1.0/

- Veids, kā datu kopa tika savākta;

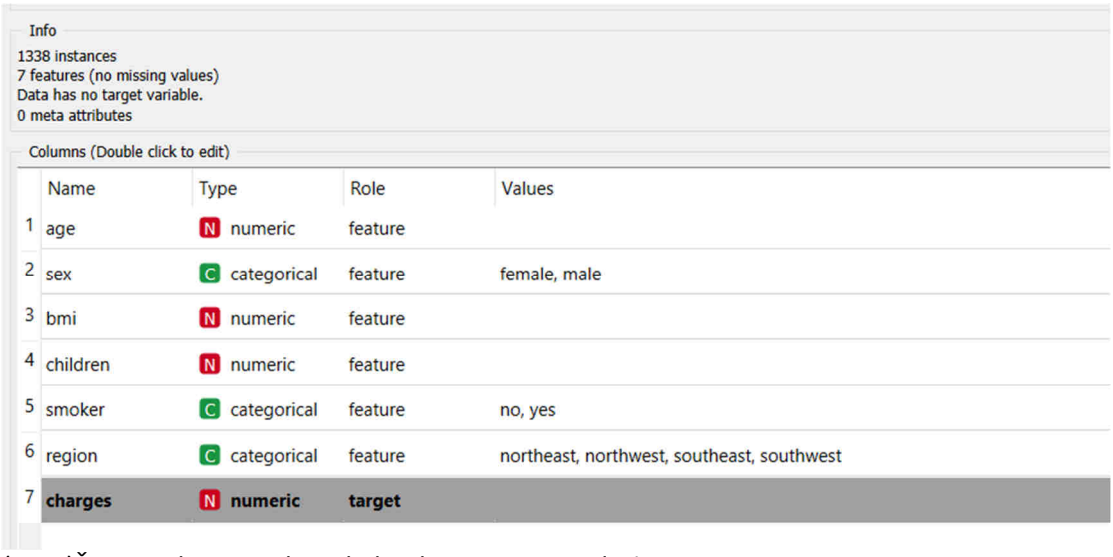
  Nav norādīts.

Datu kopas satura apraksts (sniedzot arī atsauces uz izmantotajiem informācijas avotiem)

- Datu objektu skaits datu kopā;

  1338 kopējie ieraksti.

- Datu kopas pazīmju (atribūtu) atspoguļojums kopā ar to lomām Orange rīkā;



Info
1338 instances
7 features (no missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	age	N numeric	feature	
2	sex	C categorical	feature	female, male
3	bmi	N numeric	feature	
4	children	N numeric	feature	
5	smoker	C categorical	feature	no, yes
6	region	C categorical	feature	northeast, northwest, southeast, southwest
7	**charges**	**N numeric**	**target**	

(7.att)Šajā attēli ir noradīts cik daudz parametri ir doti Dataset.

- Klašu skaits datu kopā, katras klases nozīme un klašu atspoguļošanas veids (klasēm atbilstošo iezīmju skaidrojums); ja datu kopa nodrošina vairākas iespējamas datu klasifikācijas, tad atskaitē skaidri ir jāidentificē, kāda tieši klasifikācija tiek apskatīta darbā;

  Sex-šī kategorija norāda kāds dzimums ir šai personai.
  Smoker-šī kategorija norāda vai persona pīpe tabakas izstrādājumus vai nē
  Region-šī kategorija norāda kādā reģionā US dzīvo.

- Datu objektu skaits, kas pieder katrai klasei;

  Sex-male(vīrietis),female(sieviete)
  Smoker-yes(jā),no(nē)
  Region-
  northeast(ziemeļaustrumi),nortwest(ziemeļrietumi),southeast(dienvidaustrumos),southwest(dienvidrietumos)

- Pazīmju (atribūtu) skaits un nozīme datu kopā, kā arī to vērtību tipi un diapazoni (šī informācija būtu jāatspoguļo tabulā, norādot pazīmes (atribūta) apzīmējumu, skaidrojumu, vērtību tipu un datu kopā pieejamo vērtību diapazonu);

Nr.	Name(nosaukums)	Type(tips)	Skaidrojums	Role(loma)	Values(vērtības)
1	Age	Skaitlisks	Vecums	feature	Skaitlis(18-64)
2	Sex	Kategorija	Dzimums	feature	(Female,male)(0,1)
3	Bmi(body mass index)	Skaitlisks	Ķermeņa masas indeks	feature	Skaitlis(16-53,1)
4	Children	Skaitlisks	Cik bērni ir cietušajam	feature	Skaitlis(0-5)
5	Smoker	Kategorija	Tabakas lietošana	feature	(Yes,no)(1,0)
6	Region	Kategorija	Reģions	feature	(northeast, nortwest Southeast, southwest) (0,1,2,3)
7	Charges	Skaitlisks	Izmaksas	target	Skaitlis(1120-63800)

- Datu faila struktūras fragments, kurā ir redzamas visas datu faila kolonnas un to vērtības vismaz dažiem datu objektiem;

	charges	age	sex	bmi	children	smoker	region
1	16884.9	19	0	27.900	0	1	3
2	1725.55	18	1	33.770	1	0	2
3	4449.46	28	1	33.000	3	0	2
4	21984.5	33	1	22.705	0	0	1
5	3866.86	32	1	28.880	0	0	1
6	3756.62	31	0	25.740	0	0	2
7	8240.59	46	0	33.440	1	0	2
8	7281.51	37	0	27.740	3	0	1
9	6406.41	37	1	29.830	2	0	0
10	28923.1	60	0	25.840	0	0	1
11	2721.32	25	1	26.220	0	0	0
12	27808.7	62	0	26.290	0	1	2
13	1826.84	23	1	34.400	0	0	3
14	11090.7	56	0	39.820	0	0	2
15	39611.8	27	1	42.130	0	1	2
16	1837.24	19	1	24.600	1	0	3
17	10797.3	52	0	30.780	1	0	0
18	2395.17	23	1	23.845	0	0	0
19	10602.4	56	1	40.300	0	0	3
20	36837.5	30	1	35.300	0	1	3
21	13228.8	60	0	36.005	0	0	0
22	4149.74	30	0	32.400	1	0	3

(8.att)Šajā attēlā ir norādīts data table kurā ir atspoguļoti dati jau pārveidoti uz numerācija kā piem. (yes ir 1, un no ir 0).

Secinājumi, kas izriet no izkliedes diagrammu, histogrammu un sadalījumu analīzes (sk. I daļas 5. solis) par datu kopas klašu atdalāmību. Studentiem ir jāatbild uz šādiem jautājumiem:

- Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)? Tas tiek noteikts, spriežot pēc tā, cik daudz datu objektu pieder katrai kopai.

  Nevarētu teikt ka visas klases ir līdzsvarotas jo tomēr ir būla tipa klases, šajā dataset dominē viena klase kas ir (charges) kas norāda cik daudz patērētu līdzekļu tika izmatoti uz personu veselības apdrošināšanā.

- Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru? Runa ir par to, vai datu objekti, kuri pieder dažādām klasēm, ir skaidri atdalāmi.

  Šajā tabulā ir atspoguļojama datu struktūra, jo katru klasi var skaidri atdalīt.

- Cik datu grupējums ir iespējams identificēt, pētot datu vizuālo atspoguļojumu? Runa ir par to, vai ir kaut cik atdalāmi datu grupējumi, ja gadījumā dažādu klašu datu objekti saplūst kopā.

  Ir iespēja atdalīt datu grupējumus un labāk parskatīt datus.

- Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?

  Identificētie datu grupējumi atrodas viens otram tuvu.

Secinājumi, kas izriet no statistisko rādītāju (vidējo vērtību un dispersijas vērtību) analīzes.

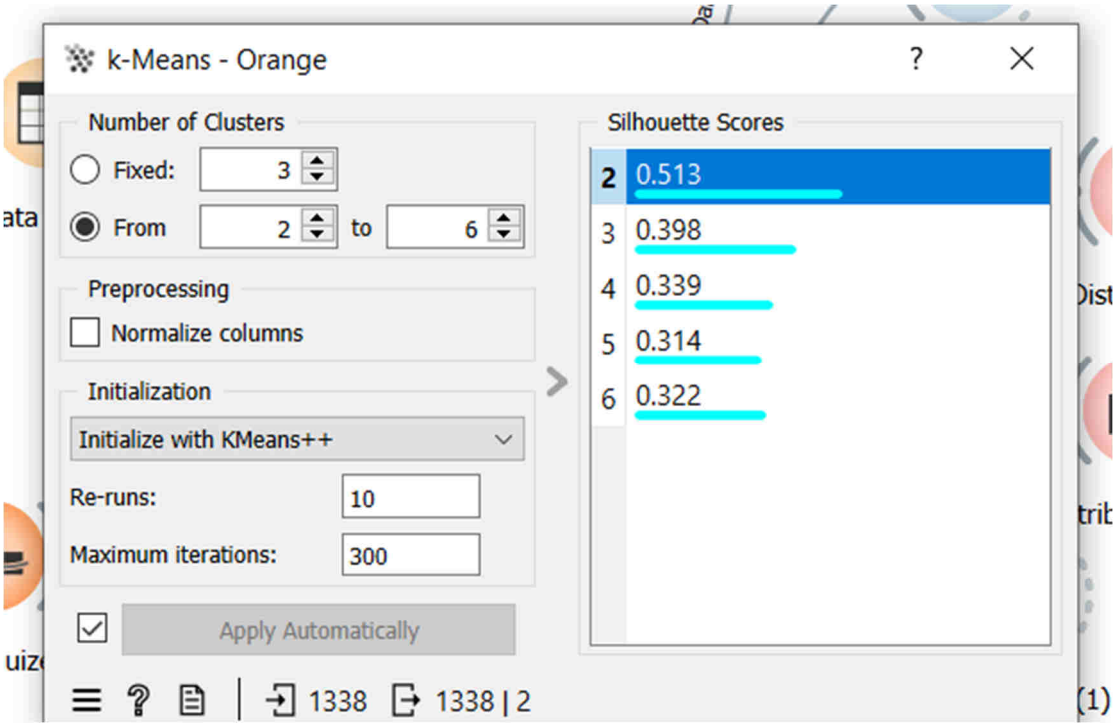
Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
age		39.21	18	39	0.36	18	64	0 (0 %)
sex		0.505232	1	1	0.989591	0.00	1	0 (0 %)
bmi		30.66340	32.300	30.400	0.19880	15.960	53.130	0 (0 %)
children		1.09	0	1	1.10	0	5	0 (0 %)
smoker		0.204783	0.00	0.00	1.97059	0.00	1	0 (0 %)
region		1.5157	2	2	0.72869	0.00	3	0 (0 %)
charges		13270.4	1639.56	9382.03	0.912216	1121.87	63770.4	0 (0 %)

(9.att)

Šajā statiskajā radijumā ir aprēķināta videja vērtiba un dispersija. Vidējais vecums ir 39 gadi un vidējie patērētie līdzekļi ir 13270.4$ kas ir uz pusi sievietēm un uz pusi vīriešiem 1 vienu bērnu ģimenē. Dispersija patērētajiem līdzekļiem ir 0.912216.

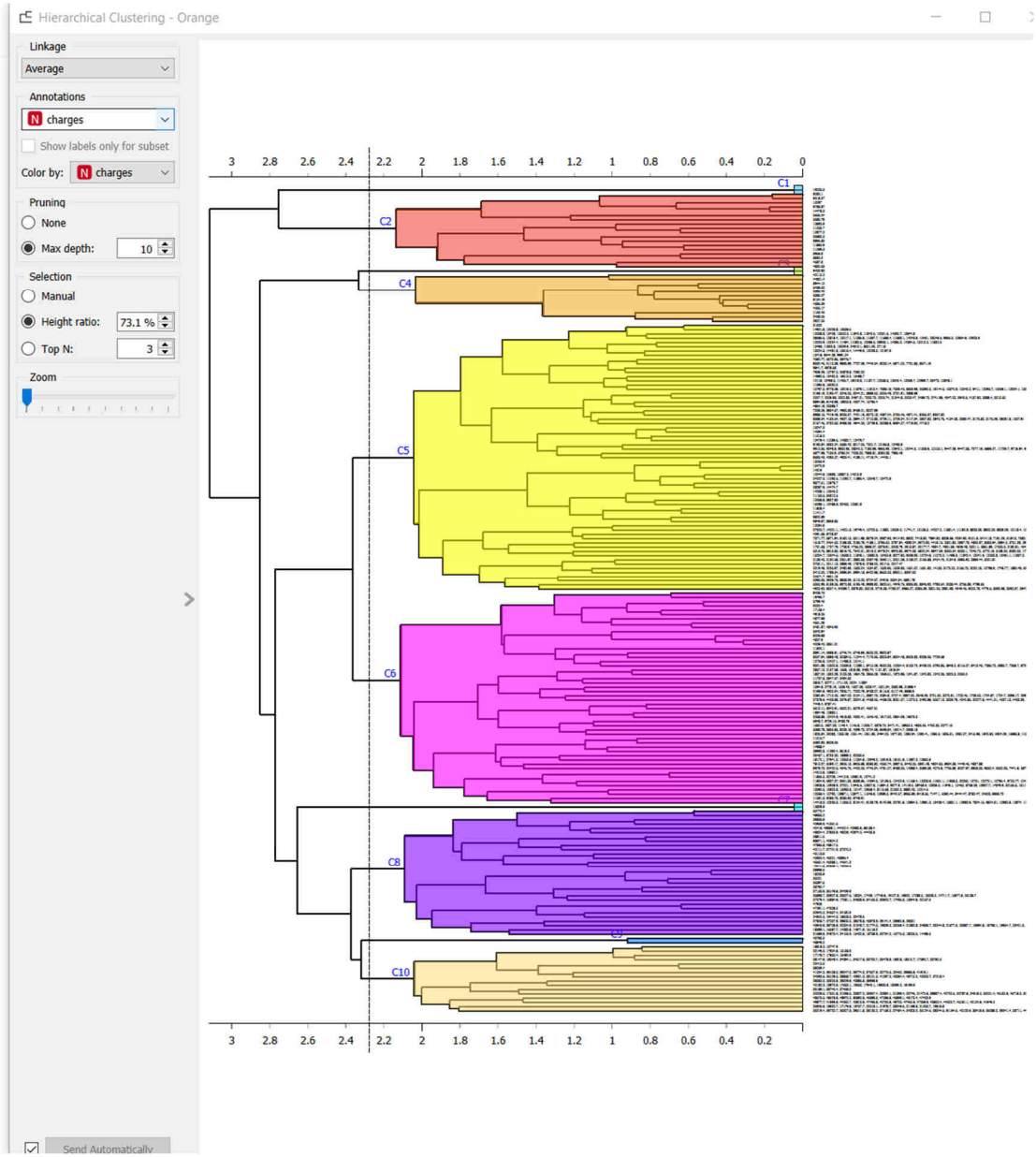
2.daļa-Nepārraudzītā mašīnmācīšanās

Lai izpildītu šo darba daļu, studentiem ir jāveic šādas darbības:
1. Jāpielieto divi studiju kursā apskatītie nepārraudzītās mašīnmācīšanās algoritmi: (1) hierarhiskā klasterizācija un (2) K-vidējo algoritms.

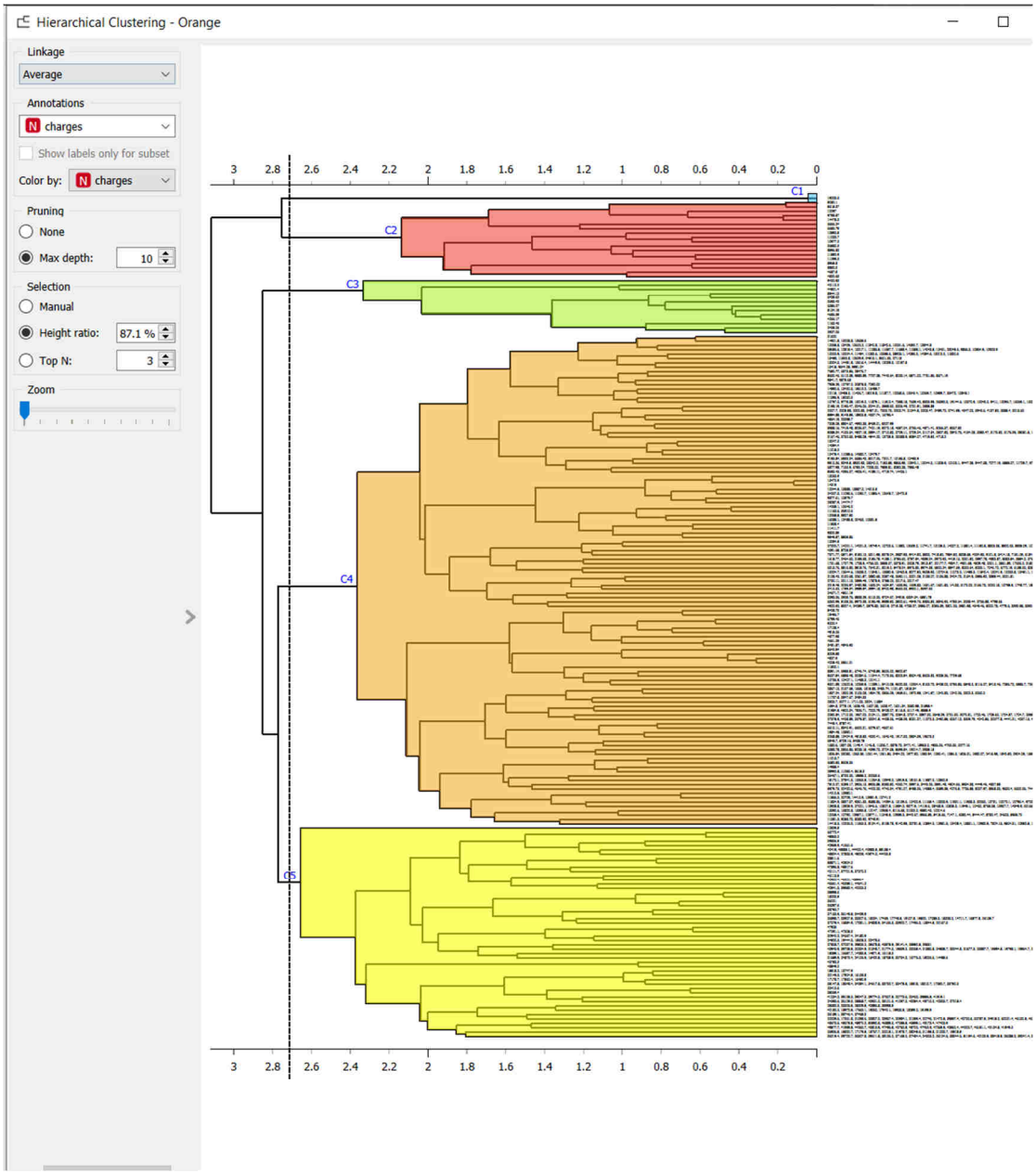


(10.att)

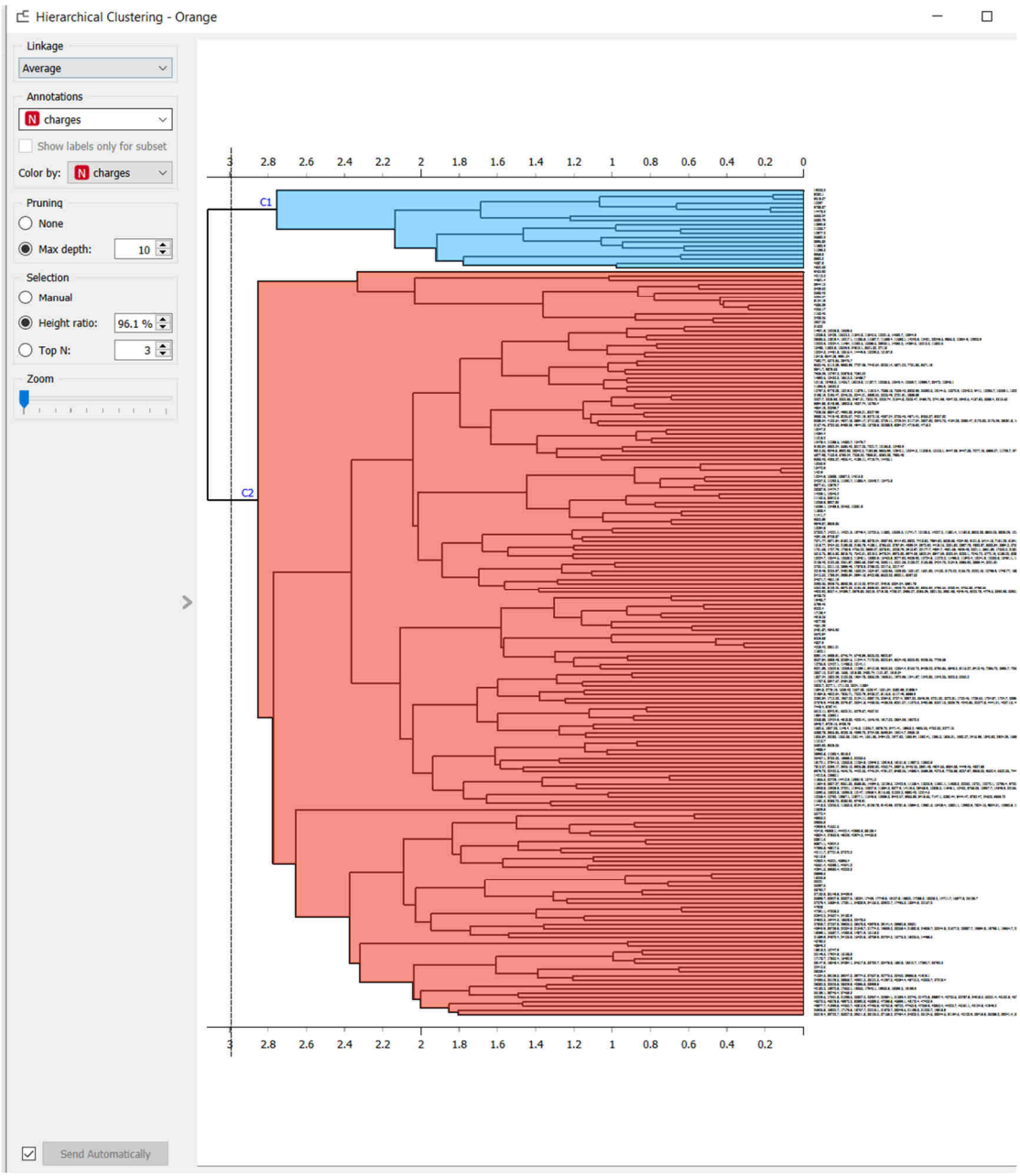
2. Hierarhiskās klasterizācijas algoritmam ir jāveic vismaz 3 eksperimenti, brīvi pārvietojot atdalošo līniju un analizējot, kā mainās klasteru skaits un saturs;



(11.att)


Hierarchical Clustering - Orange
Linkage
Average
Annotations
charges
Show labels only for subset
Color by:
charges
Pruning
None
Max depth:
10
Selection
Manual
Height ratio:
87.1 %
Top N:
3
Zoom
3 2.8 2.6 2.4 2.2 2 1.8 1.6 1.4 1.2 1 0.8 0.6 0.4 0.2 0
C1
C2
C3
C4
C5
3 2.8 2.6 2.4 2.2 2 1.8 1.6 1.4 1.2 1 0.8 0.6 0.4 0.2

(12.att)

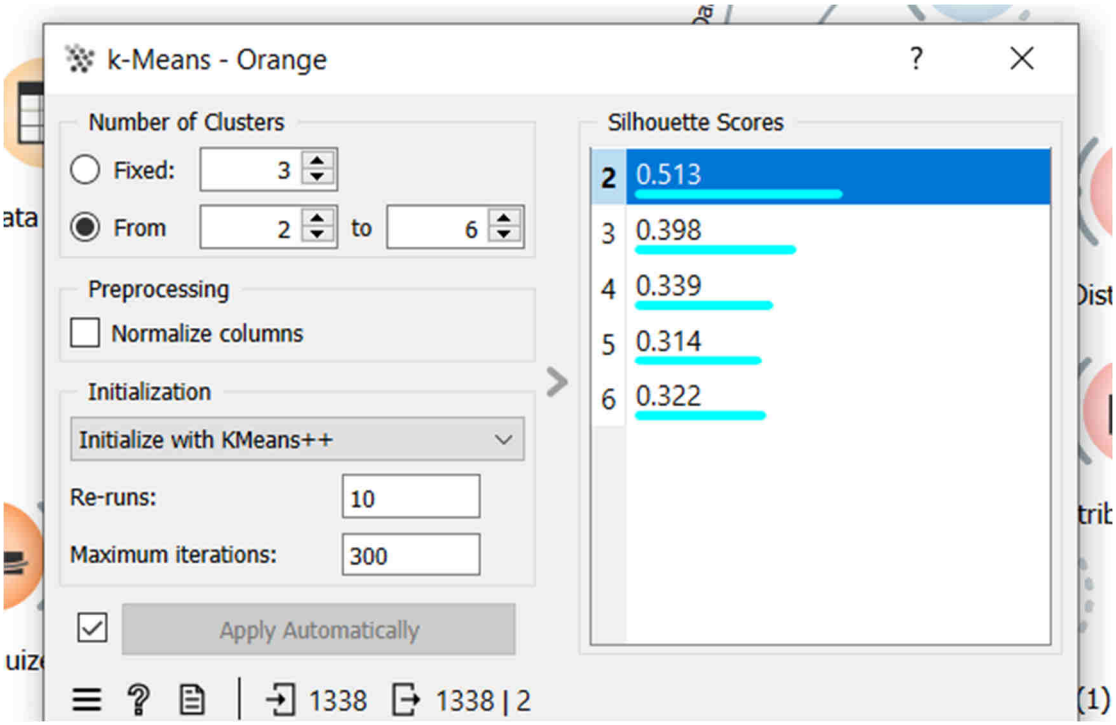

Hierarchical Clustering - Orange
Linkage
Average
Annotations
charges
Show labels only for subset
Color by:
charges
Pruning
None
Max depth:
10
Selection
Manual
Height ratio:
96.1 %
Top N:
3
Zoom
Send Automatically
C1
C2
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0

(13.att)

3. K-vidējo algoritmam ir jāaprēķina Silhouette Score vismaz 5 dažādām *k* vērtībām, un jāanalizē algoritma darbība.

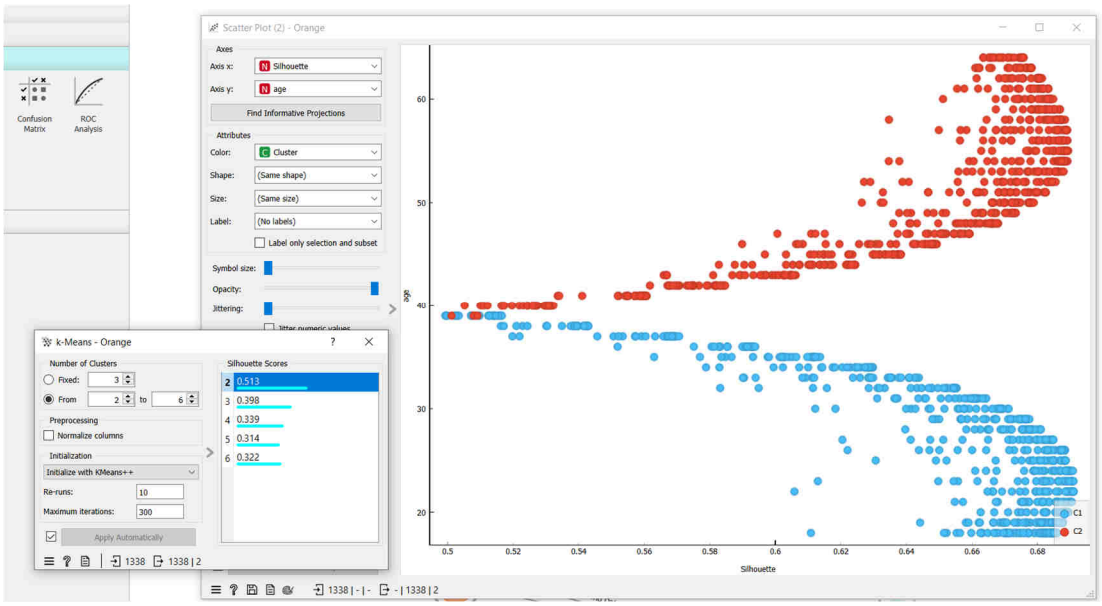
Darba atskaitē ir jāiekļauj šāda informācija par šo darba daļu:

• Katram algoritmam ir jāapraksta Orange rīkā pieejamie hiperparametri un to nozīme.

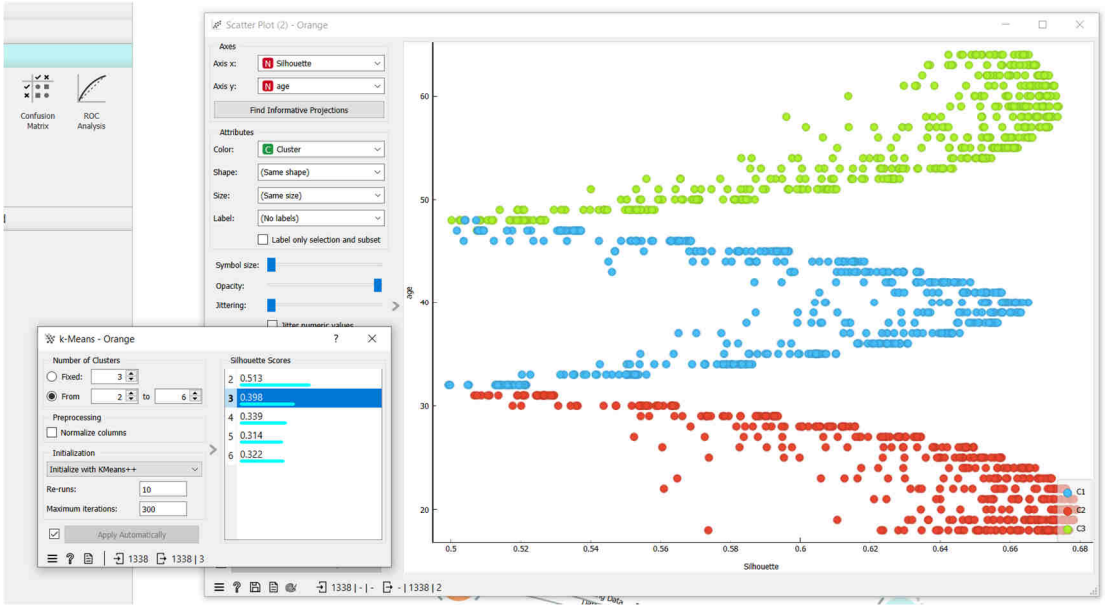


(14.att)

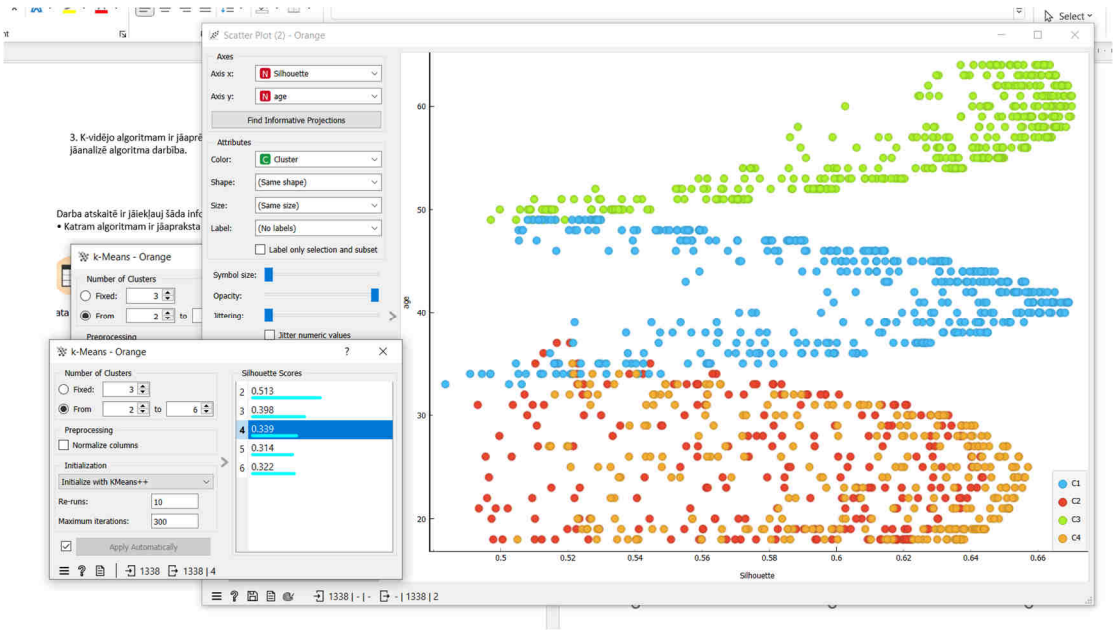
• Katram algoritmam ir jāapraksta veiktie eksperimenti, skaidri norādot izmantotās hiperparametru vērtības, un sniedzot secinājumus par algoritma darbību no tā viedokļa, cik iegūtie rezultāti atbilst zināmajam klašu skaitam datu kopā.



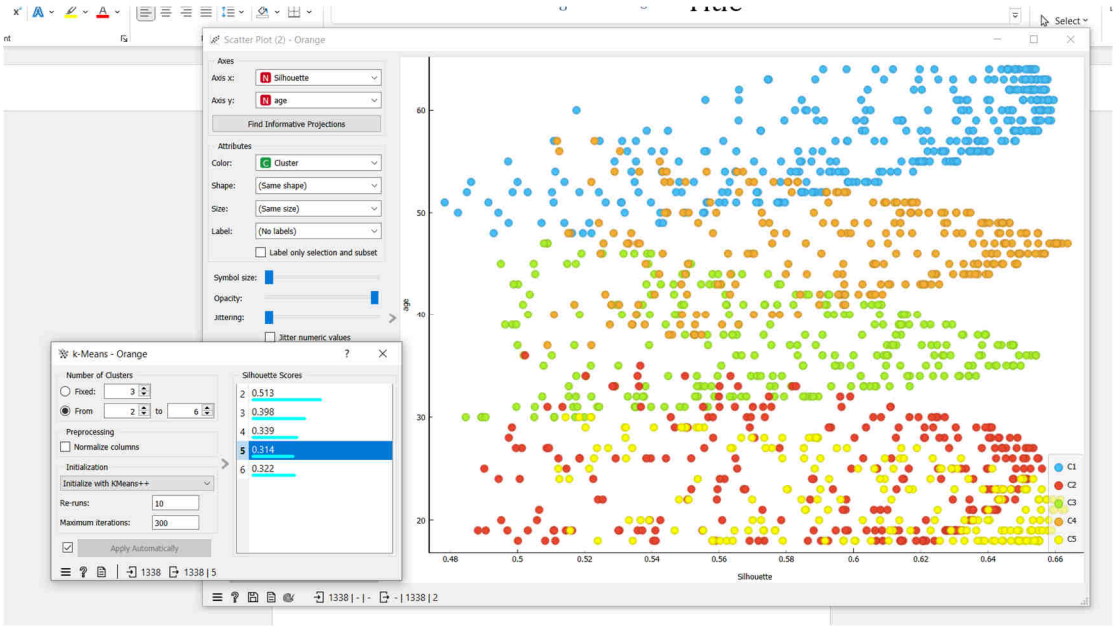
(15.att)



(16.att)



(17.att)



(18.att)skatoties uz šo eksperimetnu un 19.att eksperimentu var saprast cik būtiski ir tomēr K-Means score, jo atšķiroties tikai par 0.008 grafiks izmainās.

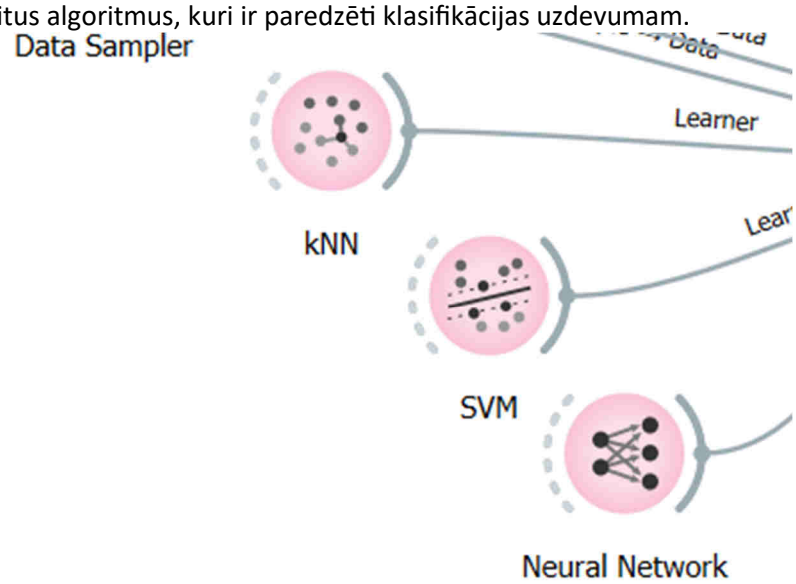
(19.att)

• Balstoties uz abu algoritmu darbības analīzi, ir jādod studenta secinājumi par to, vai datu kopā esošās klases ir labi vai slikti atdalāmas.

Balstoties uz abu algoritmiem ir slikti atdalāmas , jo parak daudz cluster.

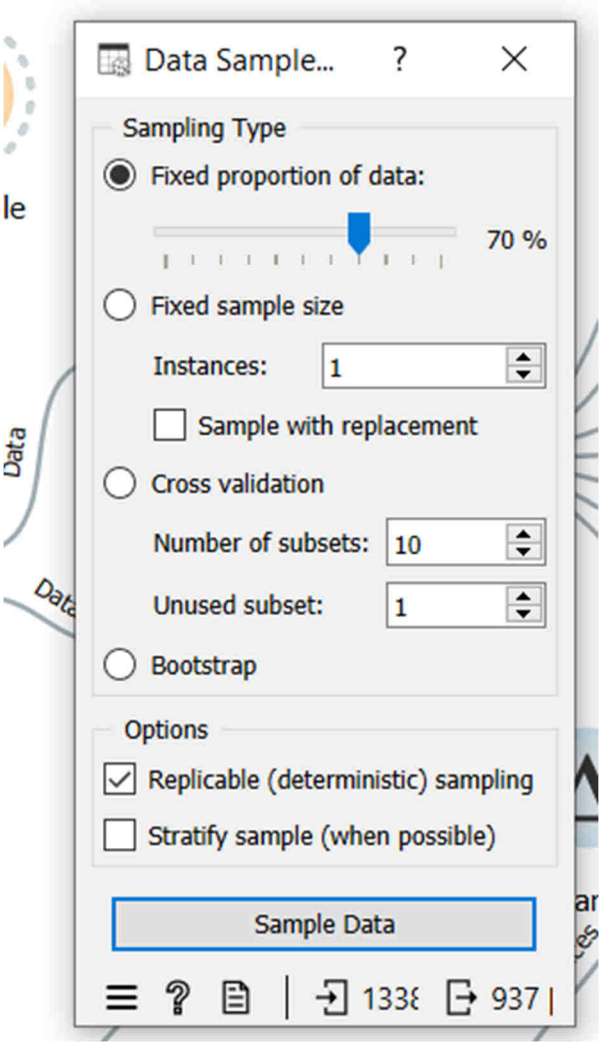
3.daļa-Pārraudzītā mašīnmācīšanās

Lai izpildītu šo darba daļu, studentiem ir jāveic šādas darbības:
1. Ir jāizvēlas vismaz divi pārraudzītās mašīnmācīšanās algoritmi, kas ir paredzēti klasifikācijas uzdevumam. Studenti drīkst izmantot studiju kursā aplūkotos algoritmus vai arī jebkurus citus algoritmus, kuri ir paredzēti klasifikācijas uzdevumam.



(20.att)

2. Ir jāsadala datu kopa apmācību un testa datu kopās.

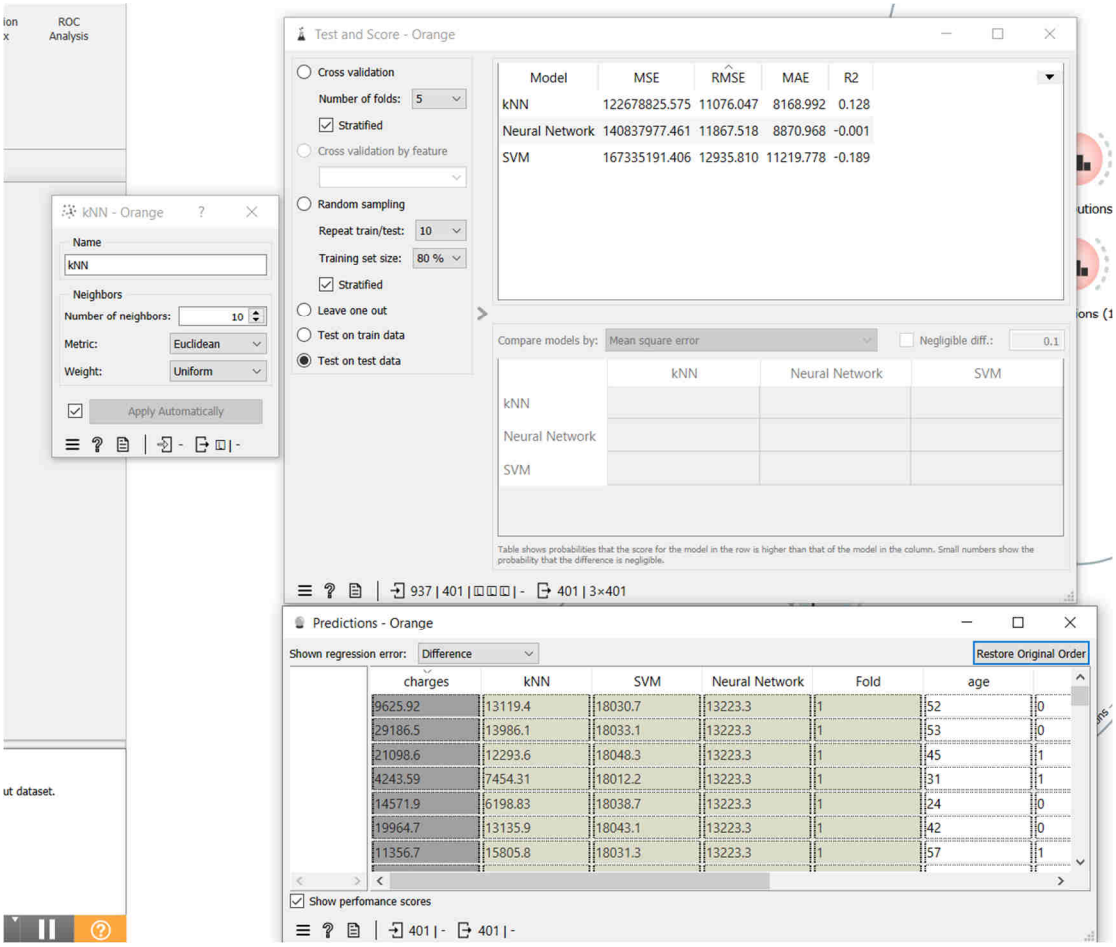


(21.att)

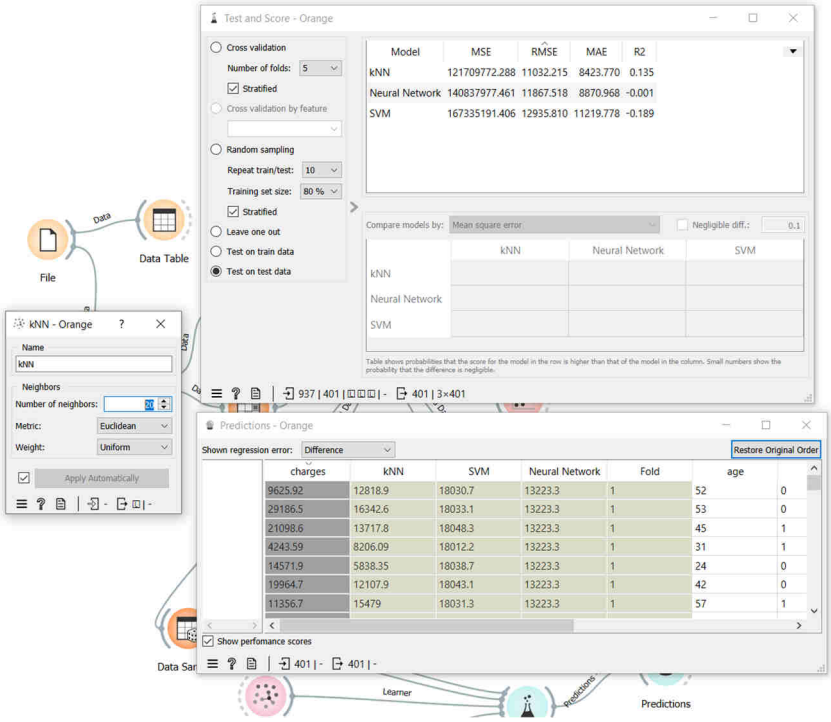
3. Katram algoritmam, lietojot apmācību datu kopu, ir jāveic vismaz 3 eksperimenti, mainot algoritma hiperparametru vērtības un analizējot algoritmu veiktspējas metrikas;

4. Katram algoritmam ir jāizvēlas tas apmācītais modelis, kas nodrošina labāko algoritma veiktspēju;

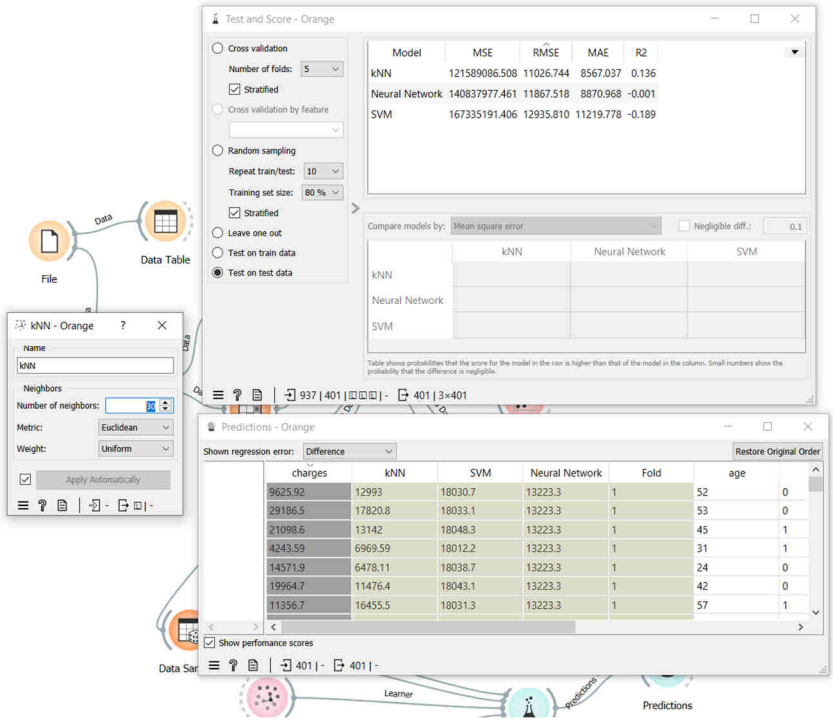
5. Katra algoritma apmācītais modelis ir jāpielieto testa datu kopai.



(22.att)


Test and Score - Orange
Cross validation
Number of folds: 5
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 80 %
Stratified
Leave one out
Test on train data
Test on test data
Model MSE RMSE MAE R2
kNN 121709772.288 11032.215 8423.770 0.135
Neural Network 140837977.461 11867.518 8870.968 -0.001
SVM 167335191.406 12935.810 11219.778 -0.189
Compare models by: Mean square error
Negligible diff.: 0.1
kNN Neural Network SVM
kNN
Neural Network
SVM
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.
937 | 401 | 401 | 3×401
Data
Data Table
File
kNN - Orange
Name
kNN
Neighbors
Number of neighbors: 5
Metric: Euclidean
Weight: Uniform
Apply Automatically
Predictions - Orange
Shown regression error: Difference
Restore Original Order
charges kNN SVM Neural Network Fold age
9625.92 12818.9 18030.7 13223.3 1 52 0
29186.5 16342.6 18033.1 13223.3 1 53 0
21098.6 13717.8 18048.3 13223.3 1 45 1
4243.59 8206.09 18012.2 13223.3 1 31 1
14571.9 5838.35 18038.7 13223.3 1 24 0
19964.7 12107.9 18043.1 13223.3 1 42 0
11356.7 15479 18031.3 13223.3 1 57 1
Show perfomance scores
401 | - 401 | -
Data Sar
Learner
Predictions

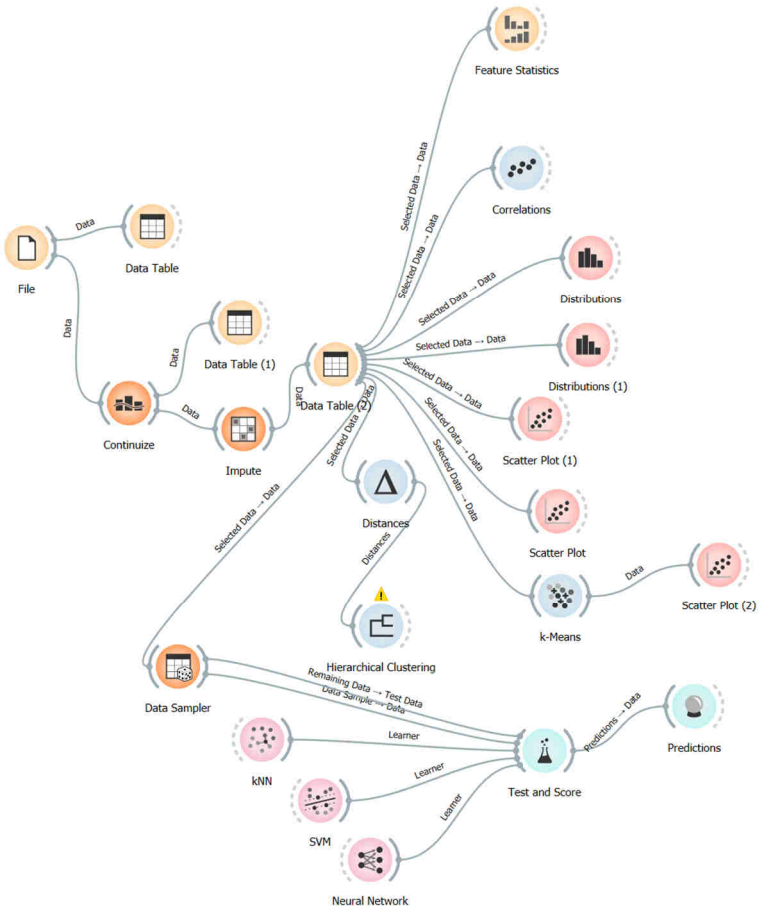
(23.att)


Test and Score - Orange
Cross validation
Number of folds: 5
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 80 %
Stratified
Leave one out
Test on train data
Test on test data
Model MSE RMSE MAE R2
kNN 121589086.508 11026.744 8567.037 0.136
Neural Network 140837977.461 11867.518 8870.968 -0.001
SVM 167335191.406 12935.810 11219.778 -0.189
Compare models by: Mean square error
Negligible diff.: 0.1
kNN Neural Network SVM
kNN
Neural Network
SVM
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.
937 | 401 | 401 | 3×401
Data
Data Table
File
kNN - Orange
Name
kNN
Neighbors
Number of neighbors: 5
Metric: Euclidean
Weight: Uniform
Apply Automatically
Predictions - Orange
Shown regression error: Difference
Restore Original Order
charges kNN SVM Neural Network Fold age
9625.92 12993 18030.7 13223.3 1 52 0
29186.5 17820.8 18033.1 13223.3 1 53 0
21098.6 13142 18048.3 13223.3 1 45 1
4243.59 6969.59 18012.2 13223.3 1 31 1
14571.9 6478.11 18038.7 13223.3 1 24 0
19964.7 11476.4 18043.1 13223.3 1 42 0
11356.7 16455.5 18031.3 13223.3 1 57 1
Show perfomance scores
401 | - 401 | -
Data Sar
Learner
Predictions

(24.att)

6. Ir jānovērtē un jāsalīdzina apmācīto modeļu veiktspēja.
Skatoties uz 22 23 un 24 attēlu un mainot knn datus mainijās beigu prediction dati un testa dati.

Orange datu shemas attēlojums



Izmantotā avoti

https://estudijas.rtu.lv/course/view.php?id=252548
https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset?resource=download
https://orangedatamining.com/