



DATORZINĀTNES UN INFORMĀCIJAS TEHNOLOGIJAS FAKULTĀTE
Priekšmeta “Mākslīgā intelekta pamati(1), 22/23-P”

Otrais praktiskais darbs

Mašīnmācīšanās algoritmu lietojums

<https://github.com/nthehemk/practicalworkML2>

Karīna Dubkova

2.kurss 7.grupa

Studenta apl. nr. 211RDB362

Rīga, 2023

SATURS

UZDEVUMS	3
PRASĪBAS	4
I DAĻA – DATU PIRMAPSTRĀDE/IZPĒTE.....	5
DATU KOPAS APRAKSTS	5
DATU KOPAS SATURA APRAKSTS	5
DATU KOPAS ANALĪZE.....	11
SECINĀJUMS.....	19
II DAĻA - NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS.....	20
HIERARHISKĀ KLASTERIZĀCIJA	20
K-VIDĒJO ALGORITMS	22
SIHOUETTE PLOT	26
SECINĀJUMS.....	27
III DAĻA – PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS.....	27
INFORMĀCIJA PAR ALGORITMIEM	27
TESTI	28
TESTU REZULTĀTU SALĪDZINĀJUMS UN ANALĪZE	37
SECINĀJUMS.....	39
IZMANTOTIE INFORMĀCIJAS AVOTI	40

UZDEVUMS

Šī darba izpildei studentiem ir nepieciešams izvēlēties datu kopu un izmantot tās apstrādei pārraudzītās un nepārraudzītās mašīnmācīšanās algoritmus. Darba mērķis ir attīstīt studentu prasmes izmantot mašīnmācīšanās algoritmus un analizēt iegūtos rezultātus. Šī darba galarezultāts ir studenta sagatavotā atskaite par darba izpildi. Darba izstrādei studentiem ir ieteicams izmantot Orange rīks. Tā lietotāja pamācība ir pieejama e-studiju kursa sadala “Praktiskie darbi”. Darba izpildes kontekstā īpaši vērtīgi ir šādi Orange logrīki: File, Data table, Data Sampler, Bar Plot, Scatter plot, Feature Statistics, Distributions, Test and Score, Predictions, Confusion matrix, Silhouette plot, Roc analysis, kā arī dažādu mašīnmācīšanās algoritmu logrīki. Tajā pašā laikā students var izvēlēties izpildīt darbu Python valodā. Tomēr tālākais uzdevuma apraksts pamatā attiecas uz rīku Orange, bet tās pašas prasības tiek piemērotas, ja students izmanto Python valodu. Ir jāņem vērā, ka darba izpildes nolūkam studentiem, iespējams, būs nepieciešams patstāvīgi meklēt un pētīt papildu informācijas avotus, lai atbildētu uz šī darba jautājumiem vai sniegtu iegūto rezultātu analīzi un interpretāciju.

PRASĪBAS

Izvēloties datu kopu, studentiem ir jāņem vērā šādi aspekti:

- ir jāizvēlas datu kopa, kas ir piemērota klasifikācijas uzdevumam. Students nedrīkst izvēlēties Iris ziedu (Iris data set) vai Pingvīnu (Palmer Archipelago (Antarctica) penguin data) datu kopas. Turklat ir jāpiedomā pie klasifikācijas jēgpilnuma, piemēram, klasificēt kontinentus pēc Covid-19 gadījumiem ir bezjēdzīgi, jo, pirmkārt, ir tikai 6 kontinenti un jaunie drīz vai tuvākajā laikā parādīsies un, otrkārt, Covid-19 gadījumu skaits nav kontinentu raksturojošā īpašība;
- ir vēlams izvēlēties datu kopu, kas jau ir dota .csv datu faila formātā; • datu kopai ir jābūt labi dokumentētai (ir jābūt pieejamai informācija par datu kopas izveidotāju, laiku, kad tā tika izveidota, un datu avotu);
- datu kopai ir jābūt saprātīga izmēra (vismaz 200 datu objekti);
- datu kopai ir jābūt detalizētam aprakstam par datu kopā esošajām datu pazīmēm (atribūtiem) un to nozīmi;
- datu pazīmju (atribūtu) skaitam ir jābūt diapazonā no 5 līdz 15;
- datu kopai ir jāsatur klašu iezīmes;
- studentiem ir jāizvairās no datu kopām, kurās ir daudz Būla tipa (patiess/nepatiess, 1/0 utt.) vai kategoriskā tipa pazīmju (atribūtu) vērtību. Ir vēlams izmantot datu kopas, kurās lielākā daļa no pazīmēm ir atspoguļota ar nepārtrauktām pazīmju vērtībām;
- studentiem ir jāizvairās no datu kopām, kurās klašu iezīmes nav dotas (piemēram, teksta korpusiem un neapstrādātiem attēliem).

I DALĀ – DATU PIRMAPSTRĀDE/IZPĒTE

DATU KOPAS APRAKSTS

Nosaukums: Diabetes Dataset

Autors: Pima Indians

Avots: <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>

Diabēts ir hroniska slimība, ko raksturo augsts cukura līmenis asinīs. Tā rodas tāpēc, ka organisms nespēj efektīvi izmantot vai ražot insulīnu.

Tādi atribūti kā grūtniecība, glikozes līmenis asinīs, asinsspiediens, ādas biezums, insulīna līmenis asinīs, ķermeņa masas indekss, diabēta procentuālā daļa, sievietes vecums un rezultāti.

Šie dati tika iegūti no Nacionālā diabēta un gremošanas un nieru slimību institūta. Tā tika apkopota, lai prognozētu diabēta klātbūtni pacientiem, pamatojoties uz datu kopā iekļautajiem diagnostiskajiem mērījumiem. Dati sastāv no 9 atribūtiem, un objektu apjoms ir 768. Tādi atribūti kā grūtniecība, glikozes līmenis asinīs, asinsspiediens, ādas biezums, insulīna līmenis asinīs, ķermeņa masas indekss, diabēta procentuālā daļa, sievietes vecums un rezultāti.

Licence: CC0: Public Domain

DATU KOPAS SATURA APRAKSTS

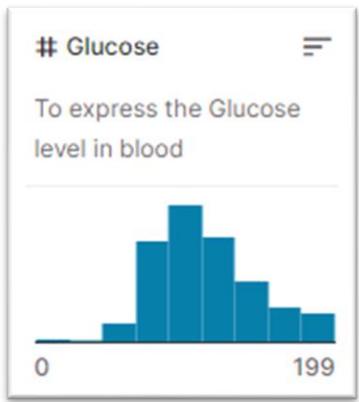
Atribūti to tips, veids un apraksts.

1. att.



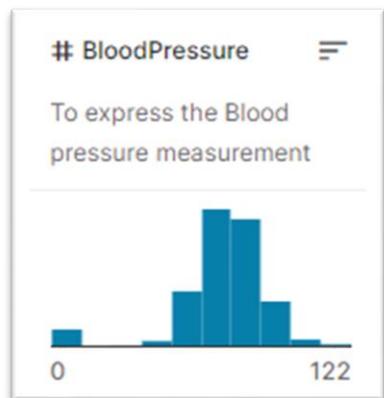
Grūtniecības atribūts parāda, cik daudzām sievietēm ir bijusi grūtniecība. (no 0 līdz 17 sievietēm) (1. att.)

2. att.



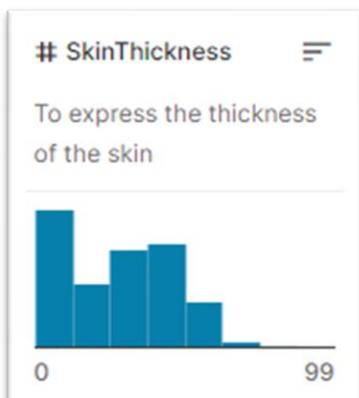
Atribūts glikozes līmenis asinīs parāda, kāds bija sievietes glikozes līmenis (no 0 līdz 199 mmol/l). (2. att.)

3. att.



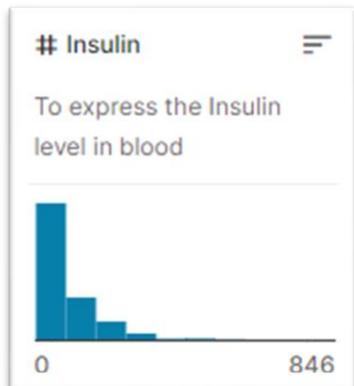
Asinsspiediena atribūts parāda, kāds bija asinsspiediens sievietēm (0 līdz 122 pulss). (3. att.)

4. att.



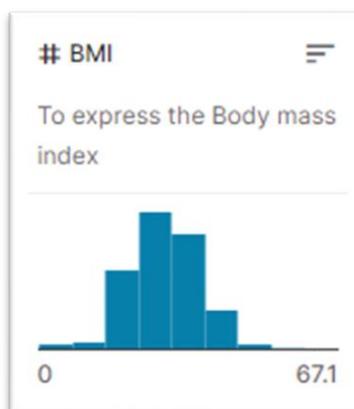
Atribūts "ādas biezums" parāda, cik bija sieviešu āda (no 0 līdz 99 mm). (4. att.)

5. att.



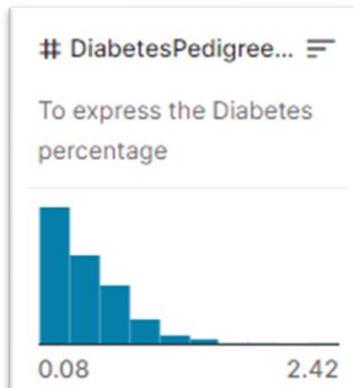
Insulīna atribūts parāda, kāds bija insulīna līmenis sievietēm (0 līdz 846 uiu/ml). (5. att.)

6. att.



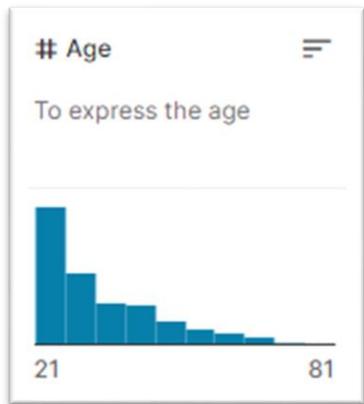
Ķermeņa masas indeksa atribūts parāda, kāds bija ķermeņa masas indeksa līmenis sievietēm (no 0 līdz 67,1 kg²/m). (6. att.)

7. att.



Diabēta atribūts procentos parāda, kāds bija diabēta līmenis sievietēm (no 0,08 līdz 2,42 mmol/l). (7.att.)

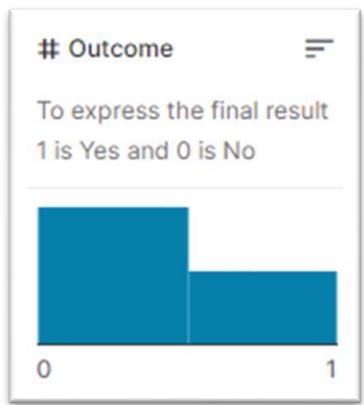
8. att.



Atribūts "vecums" norāda, cik vecas bija sievietes (no 21 līdz 82 gadiem).

Rezultāta atribūts parāda, kāds ir galīgais rezultāts (0 ir nē, 1 ir jā). (8. att.)

9. att.



Rezultāta atribūts parāda, kāds ir galīgais rezultāts (0 ir nē, 1 ir jā). (9. att.)

1.tab

Klases nosaukums	Objektu skaits
Sievietēm nav diabēta (0)	500
Sievietēm ir diabēts (1)	268

2. tab

Atribūta nosaukums	Vērtības	Tips
grūtniecība	no 0 līdz 17	skaitliskais
likozes līmenis asinīs	no 0 līdz 199	skaitliskais
asinsspiediens	no 0 līdz 122	skaitliskais
ādas biezums	no 0 līdz 99	skaitliskais
insulīna līmenis asinīs	no 0 līdz 846	skaitliskais

ķermenē masas indekss	no 0 līdz 67	skaitliskais
diabēta procentuālā daļa	no 0.08 līdz 2.42	skaitliskais
sievietes vecums	no 21 līdz 81	skaitliskais
rezultāti	no 0 līdz 1	skaitliskais

10. att. datu kopas pazīmju (atribūtu) atspoguļojums kopā ar to lomām Orange rīkā

Info

768 instances
 9 features (no missing values)
 Data has no target variable.
 0 meta attributes

Name	Type	Role	Values
1 Pregnancies	N numeric	feature	
2 Glucose	N numeric	feature	
3 BloodPressure	N numeric	feature	
4 SkinThickness	N numeric	feature	
5 Insulin	N numeric	feature	
6 BMI	N numeric	feature	
7 DiabetesPedigr...	N numeric	feature	
8 Age	N numeric	feature	
9 Outcome	C categorical	target	0, 1

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

11. att. datu bāze scv formātā

Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1 0=Blood D	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69
2 0=Blood D	32	m	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5
3 0=Blood D	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3
4 0=Blood D	32	m	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7
5 0=Blood D	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76	29.9	68.7
6 0=Blood D	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111	91	74
7 0=Blood D	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70	16.9	74.5
8 0=Blood D	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.6	109	21.5	67.1
9 0=Blood D	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.1	83	13.7	71.3
10 0=Blood D	32	m	42.4	86.3	20.3	20	35.2	5.46	4.45	81	15.9	69.9
11 0=Blood D	32	m	44.3	52.3	21.7	22.4	17.2	4.15	3.57	78	24.1	75.4
12 0=Blood D	33	m	46.4	68.2	10.3	20	5.7	7.36	4.3	79	18.7	68.6
13 0=Blood D	33	m	36.3	78.6	23.6	22	7	8.56	5.38	78	19.4	68.7
14 0=Blood D	33	m	39	51.7	15.9	24	6.8	6.46	3.38	65	7	70.4
15 0=Blood D	33	m	38.7	39.8	22.5	23	4.1	4.63	4.97	63	15.2	71.9
16 0=Blood D	33	m	41.8	65	33.1	38	6.6	8.83	4.43	71	24	72.7
17 0=Blood D	33	m	40.9	73	17.2	22.9	10	6.98	5.22	90	14.7	72.4
18 0=Blood D	33	m	45.2	88.3	32.4	31.2	10.1	9.78	5.51	102	48.5	76.5
19 0=Blood D	33	m	36.6	57.1	38.9	40.3	24.9	9.62	5.5	112	27.6	69.3
20 0=Blood D	33	m	42	63.1	32.6	34.9	11.2	7.01	4.05	105	19.1	68.1
21 0=Blood D	33	m	44.3	49.8	32.1	21.6	13.1	7.44	5.59	103	30.2	74
22 0=Blood D	33	m	46.7	88.3	23.4	23.9	7.8	9.42	4.62	78	29.5	74.3
23 0=Blood D	34	m	42.7	65.3	46.7	30.3	23.4	10.95	5.06	75	99.6	69.1
24 0=Blood D	34	m	43.4	46.1	97.8	46.2	11.3	7.99	3.62	71	35.3	69.6
25 0=Blood D	34	m	40.5	32.4	29.6	27.1	5.8	10.5	4.56	91	26.6	72
26 0=Blood D	34	m	44.8	77.7	36.9	31	19.5	10.51	5.59	80	23.7	78.9
27 0=Blood D	34	m	42.6	27	21.4	21.7	7.2	8.15	6.79	85	13.9	67.7
28 0=Blood D	34	m	29	41.6	29.1	16.1	4.8	6.82	4.03	62	14.5	53.2
29 0=Blood D	34	m	44.6	84.1	19.6	29.8	5.8	7.6	5.07	95	9.9	71.9
30 0=Blood D	34	m	46.8	61.7	24.5	24.2	23.1	10.99	4.6	83	23.8	73.1
31 0=Blood D	34	m	41.8	75.8	30.9	35.5	6.1	9.97	5.94	89	48.5	71.3
32 0=Blood D	34	m	46.1	70.6	35.8	30	7.6	7.7	4.2	93	14.3	78.7
33 0=Blood D	34	m	43.6	58.9	47.1	31.1	18.5	9.14	4.99	95	22.2	69.3
34 0=Blood D	35	m	37.5	69.8	37.1	25	7.8	11.66	5.73	84	27.3	71
35 0=Blood D	35	m	42.1	68.3	37.2	56.2	11.1	9.3	4.63	99	16.8	73.6

Pregnancy	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1
1	97	66	15	140	23.2	0.487	22	0
13	145	82	19	110	22.2	0.245	57	0
5	117	92	0	0	34.1	0.337	38	0
5	109	75	26	0	36	0.546	60	0
3	158	76	36	245	31.6	0.851	28	1
3	88	58	11	54	24.8	0.267	22	0
6	92	92	0	0	19.9	0.188	28	0
10	122	78	31	0	27.6	0.512	45	0
4	103	60	33	192	24	0.966	33	0

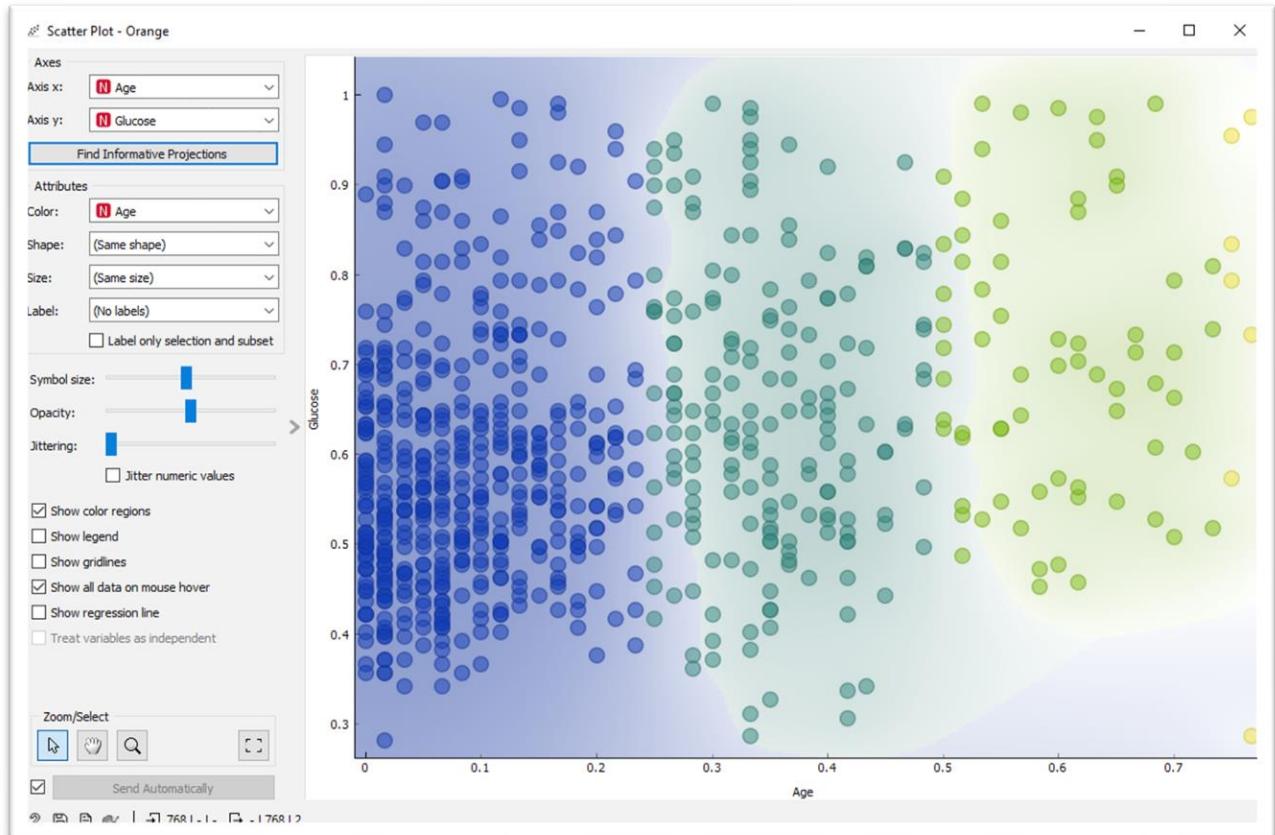
Nebija tukšu vērtību vai teksta vērtību, tāpēc nekas netika mainīts.

DATU KOPAS ANALĪZE

- a) ir jāizveido vismaz divas 2- vai 3-dimensiju izkliedes diagrammas (scatter plot), kas ilustrē klasses atdalāmību, balstoties uz dažādām pazīmēm

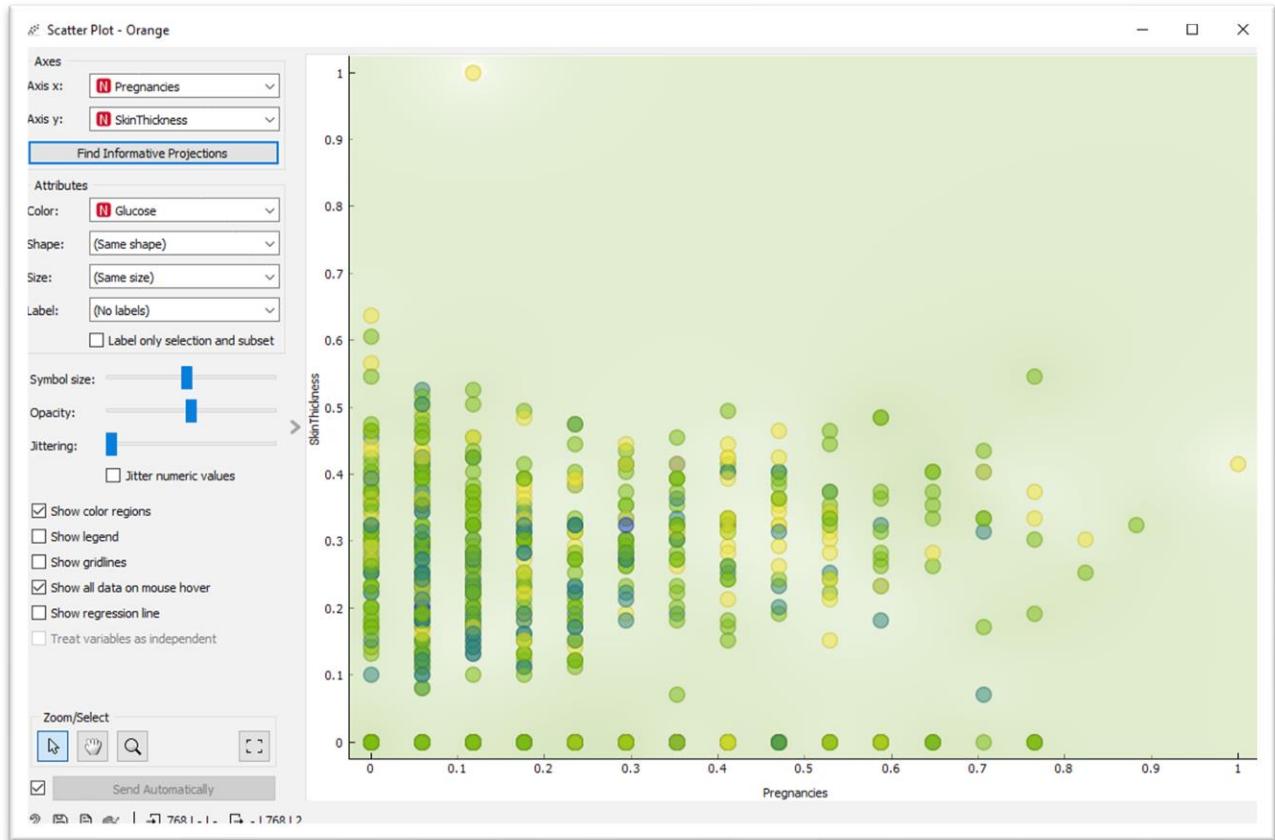
(atribūtiem); studentam ir jāizvairās izmantot datu objekta ID vai klases iezīmi kā mainīgo izkliedes diagrammā;

13. att.



Pirmajā diagrammā par x-vektoru tika izvēlēts vecums, bet par y-vektoru - glikozes līmenis asinīs; vecums tika izvēlēts kā atsevišķs atribūts. (13. att.)

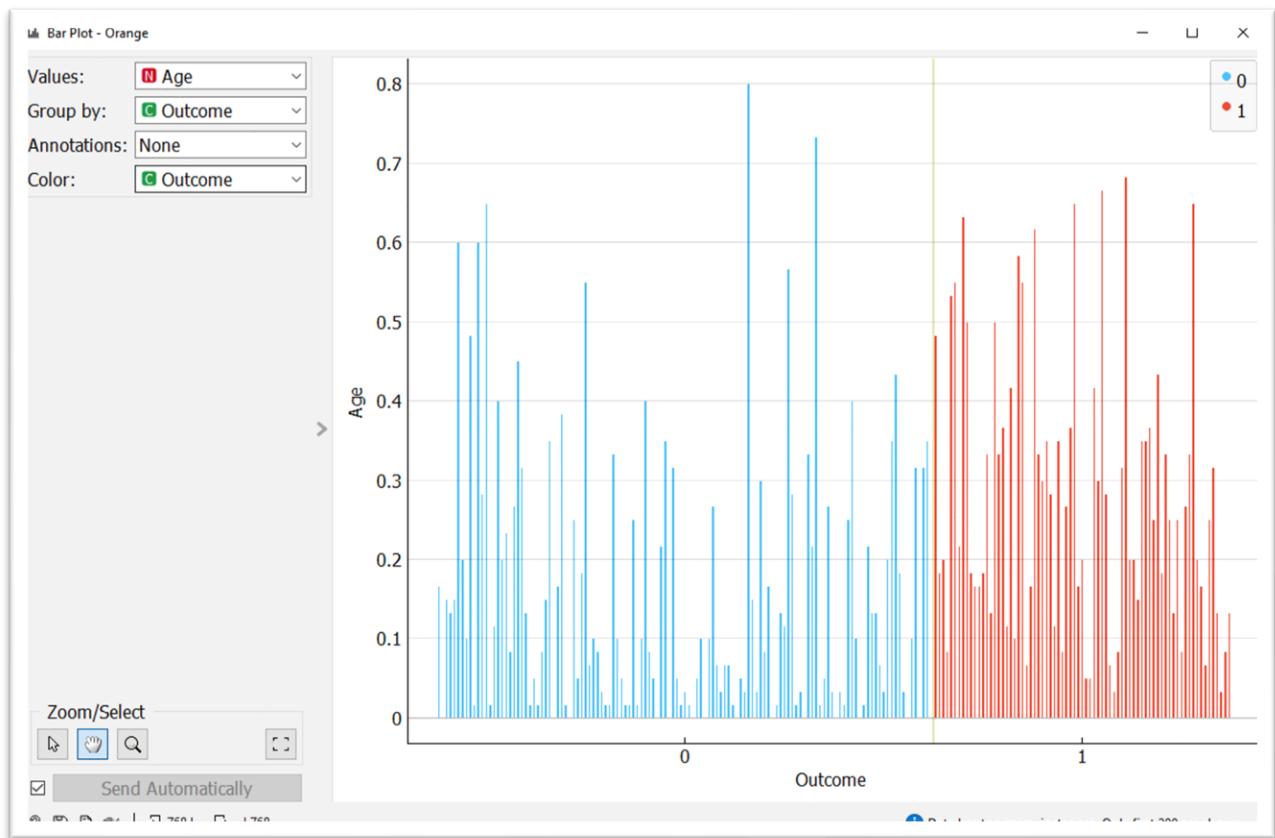
14. att.



Otrajā diagrammā par vektoru x tika izvēlēta grūtniecība, par vektoru y - ādas biezums, bet par atsevišķu atribūtu - glikoze. (14. att)

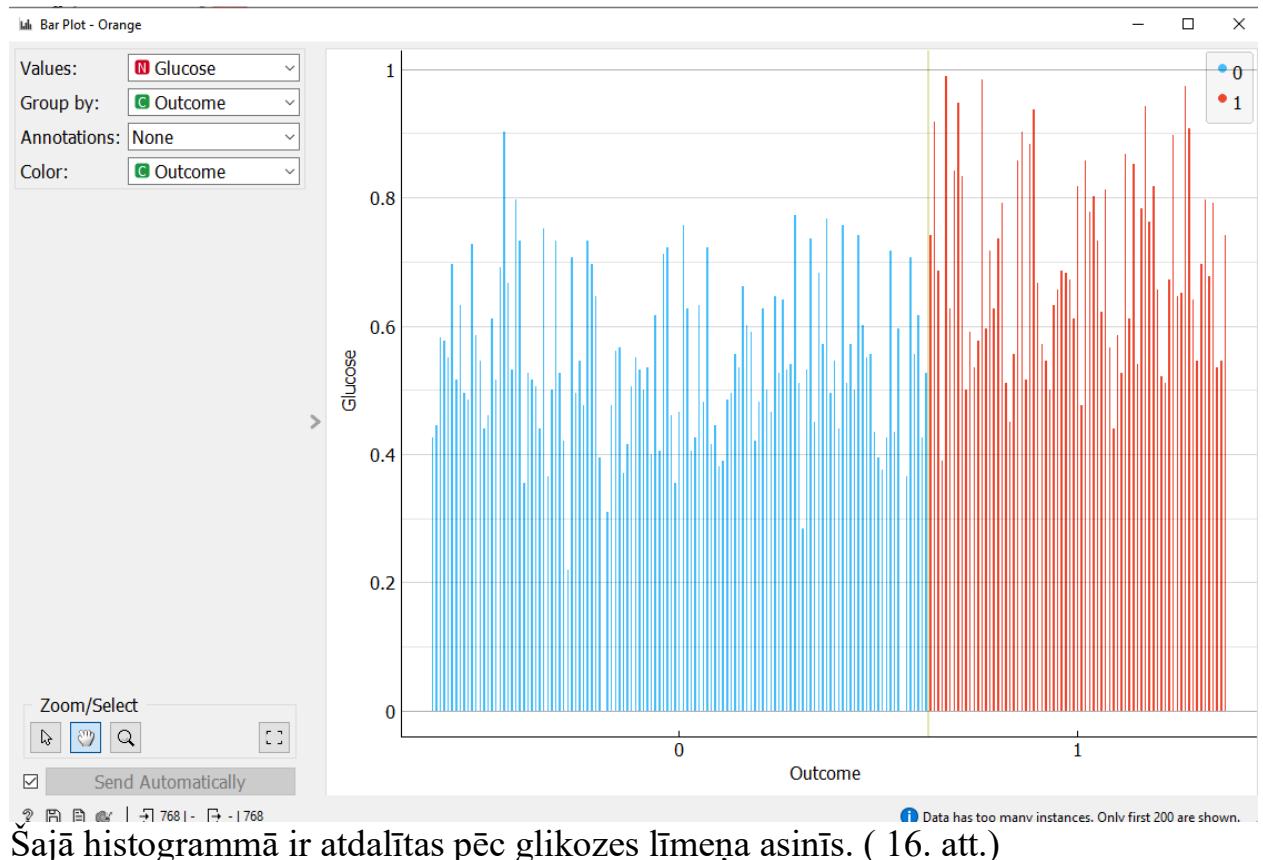
- b) ir jāizveido vismaz 2 histogrammas, kas parāda klašu atdalīšanu, pamatojoties uz interesējošām pazīmēm (atribūtiem);

15. att.



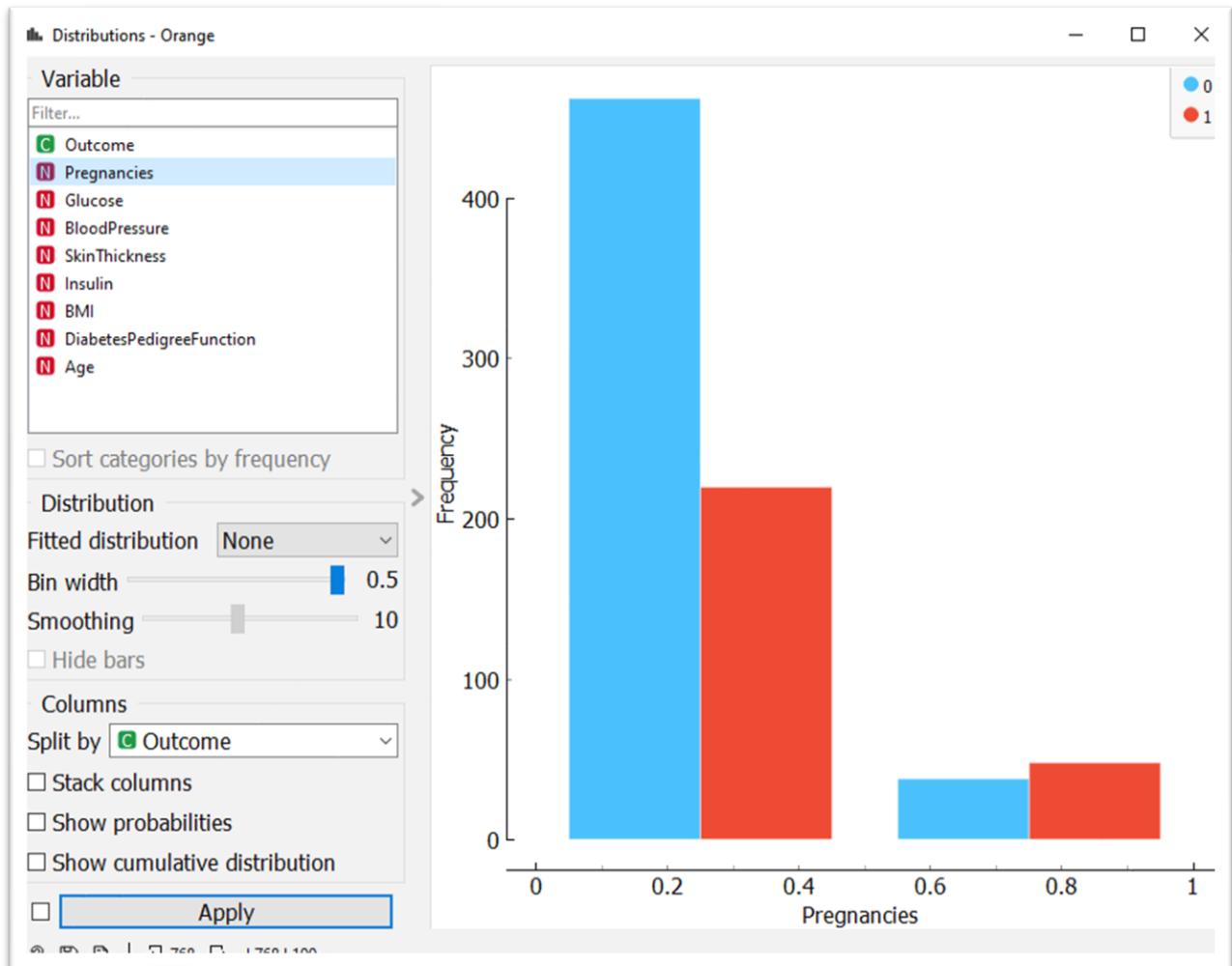
Šajā histogrammā ir atdalītas pēc vecuma. (15. att)

16. att.



Šajā histogrammā ir atdalītas pēc glikozes līmeņa asinīs. (16. att.)

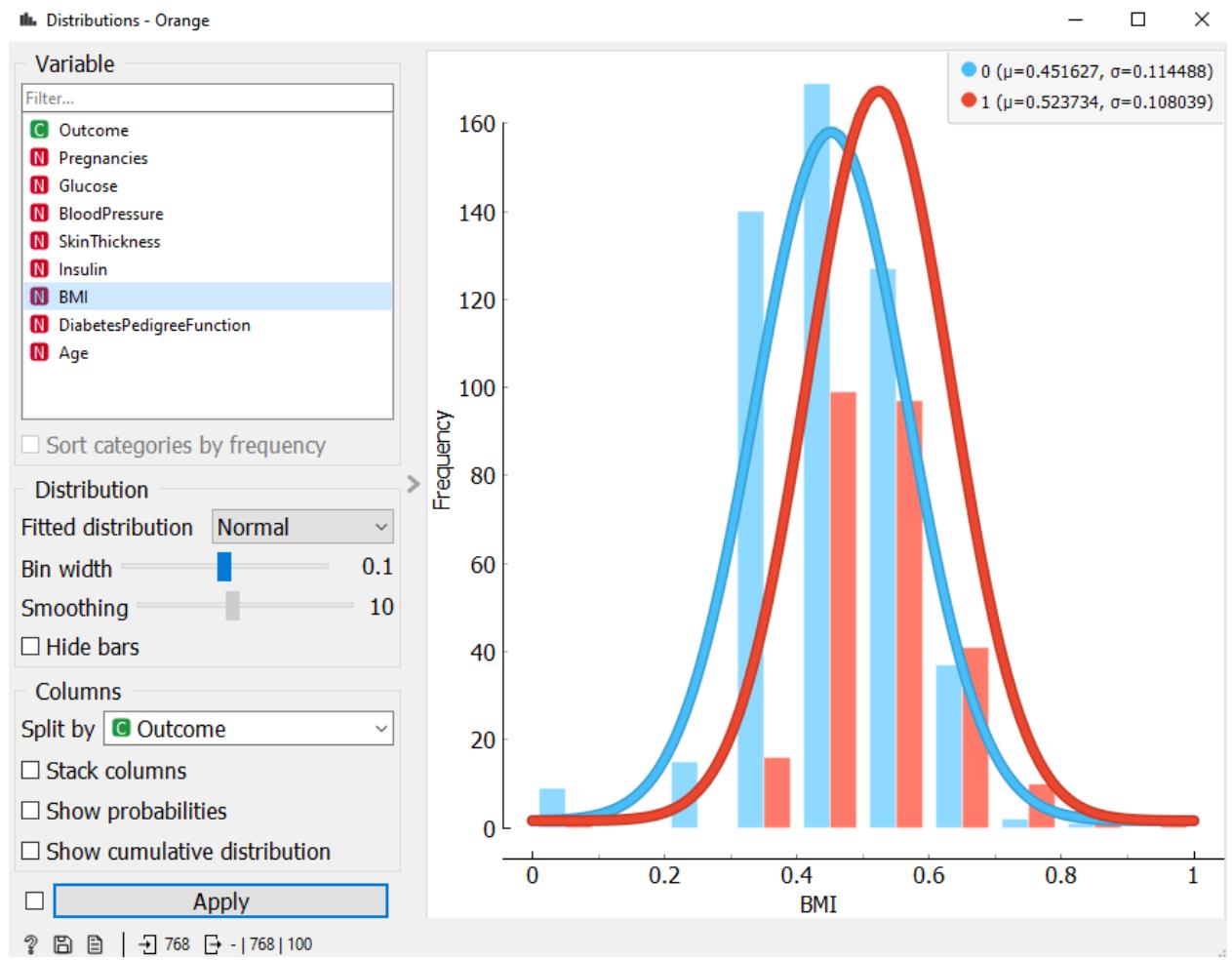
17. att.



Tiek parādīts atribūta "grūtniecība" sadalījums. (17. att.)

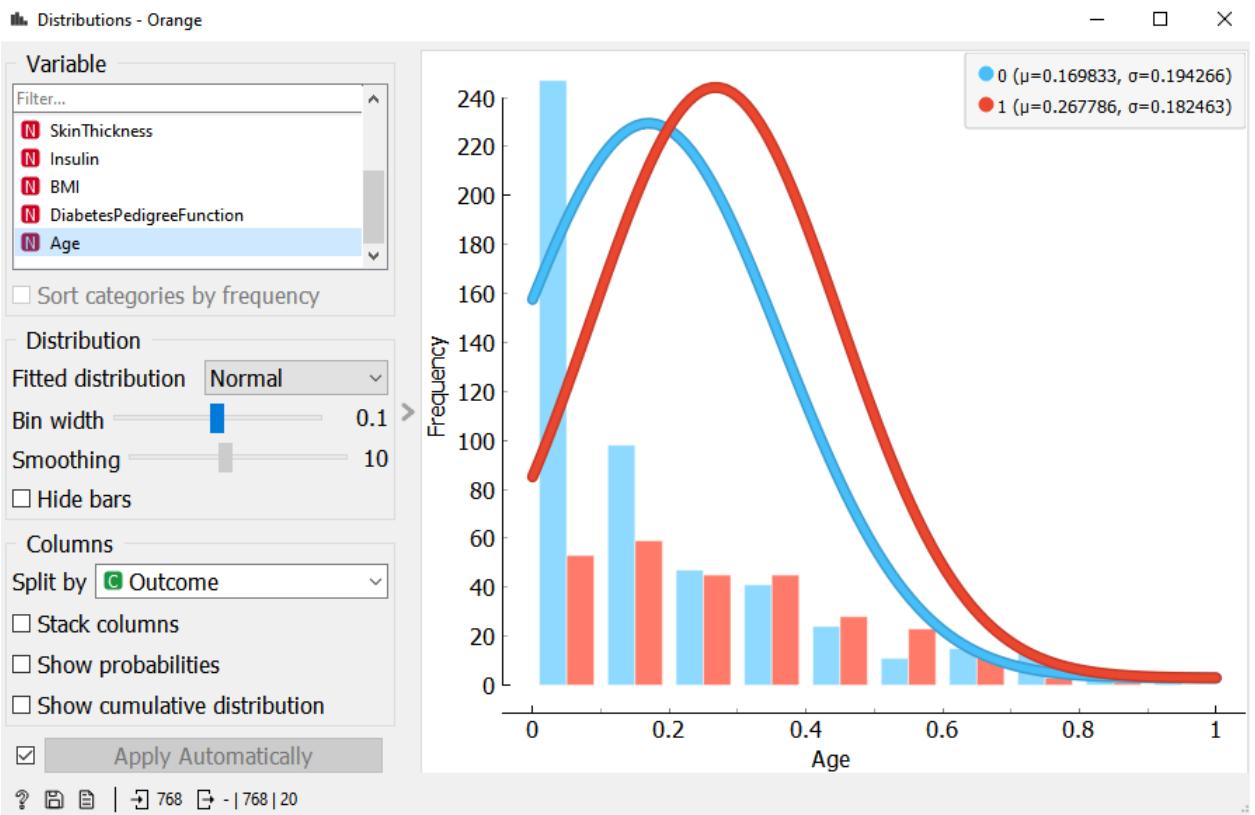
- c) ir jāatspoguļo 2 interesējošo pazīmju (atribūtu) sadalījums;

18. att.



Tiek parādīts atribūta "ķermeņa masas indekss" sadalījums. (18. att.)

19. att.



Tiek parādīts atribūta "vecums" sadalījums. (19. att.)

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

20. att.



Vidējā vērtība un dispersija tiek aprēķināta automātiski, izmantojot "Feature Statistics". (20. att.)

SECINĀJUMS

Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)?
Datu kopā nav dominējošu klašu, tās visas ir līdzsvarotas un tām ir vienāds objektu skaits. (768)

Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?

Izkliedes diagrammā dati ir pietiekami tālu viens no otra, bet tomēr starp to "robežām" dati ir tuvu viens otram. Datu struktūras var viegli redzēt. (13. att.) Tas attiecas arī uz histogrammām un sadalījumiem, kur dati arī ir skaidri redzami un nav problēmu. (18. att.) (15. att.)

Cik datu grupējums ir iespējams identificēt, pētot datu vizuālo atspoguļojumu? Aplūkojot 13.att. sniegto informāciju, var secināt, ka var izdalīt 3 datu grupas.

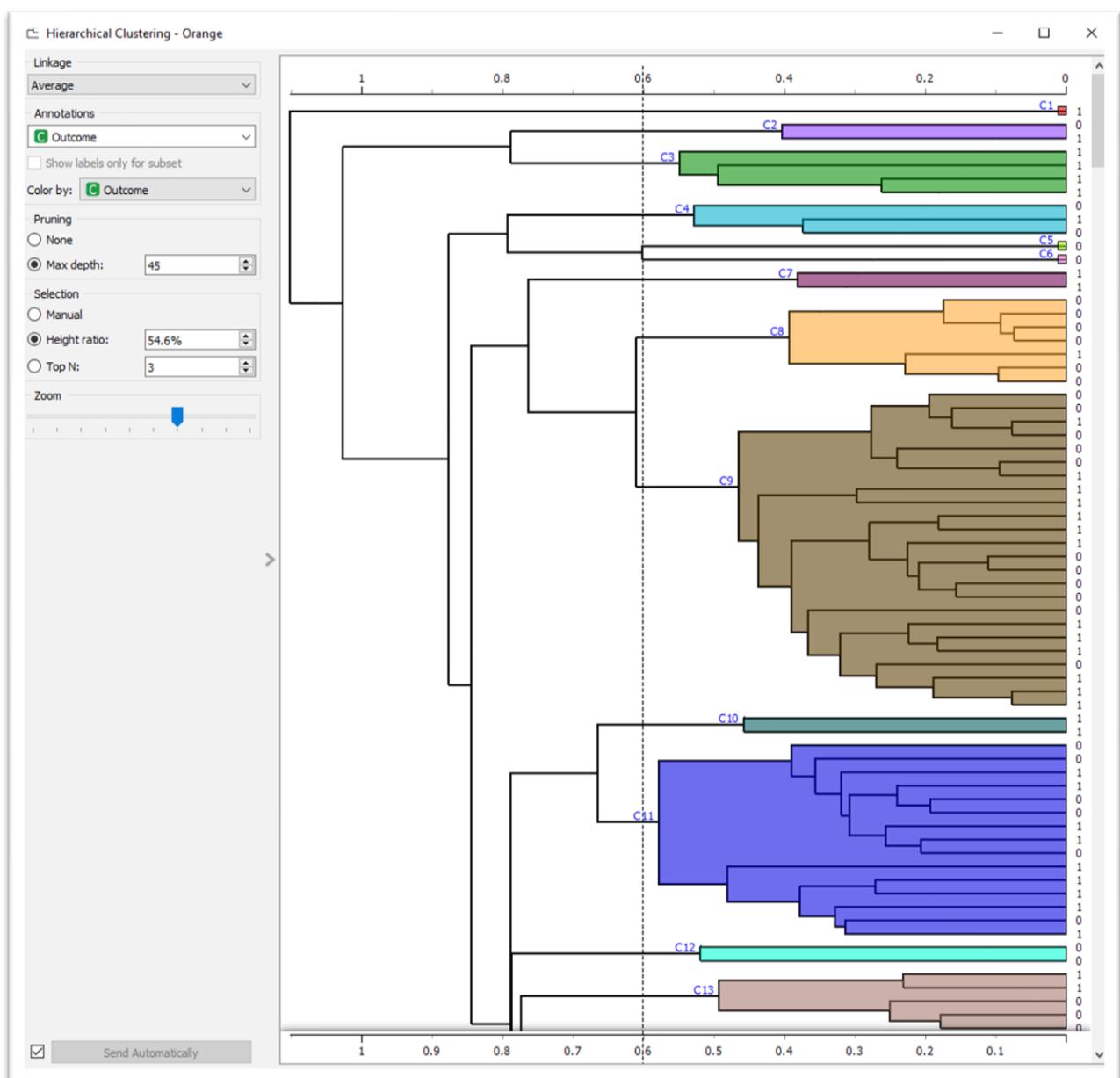
Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?

Kā atbildēts pirmajā jautājumā, tie ir ļoti tuvi viens otram. (13. att.)

II DAĀA - NEPĀRRAUDZĪTĀ MAŠINMĀCĪSANĀS

HIERARHISKĀ KLASTERIZĀCIJA

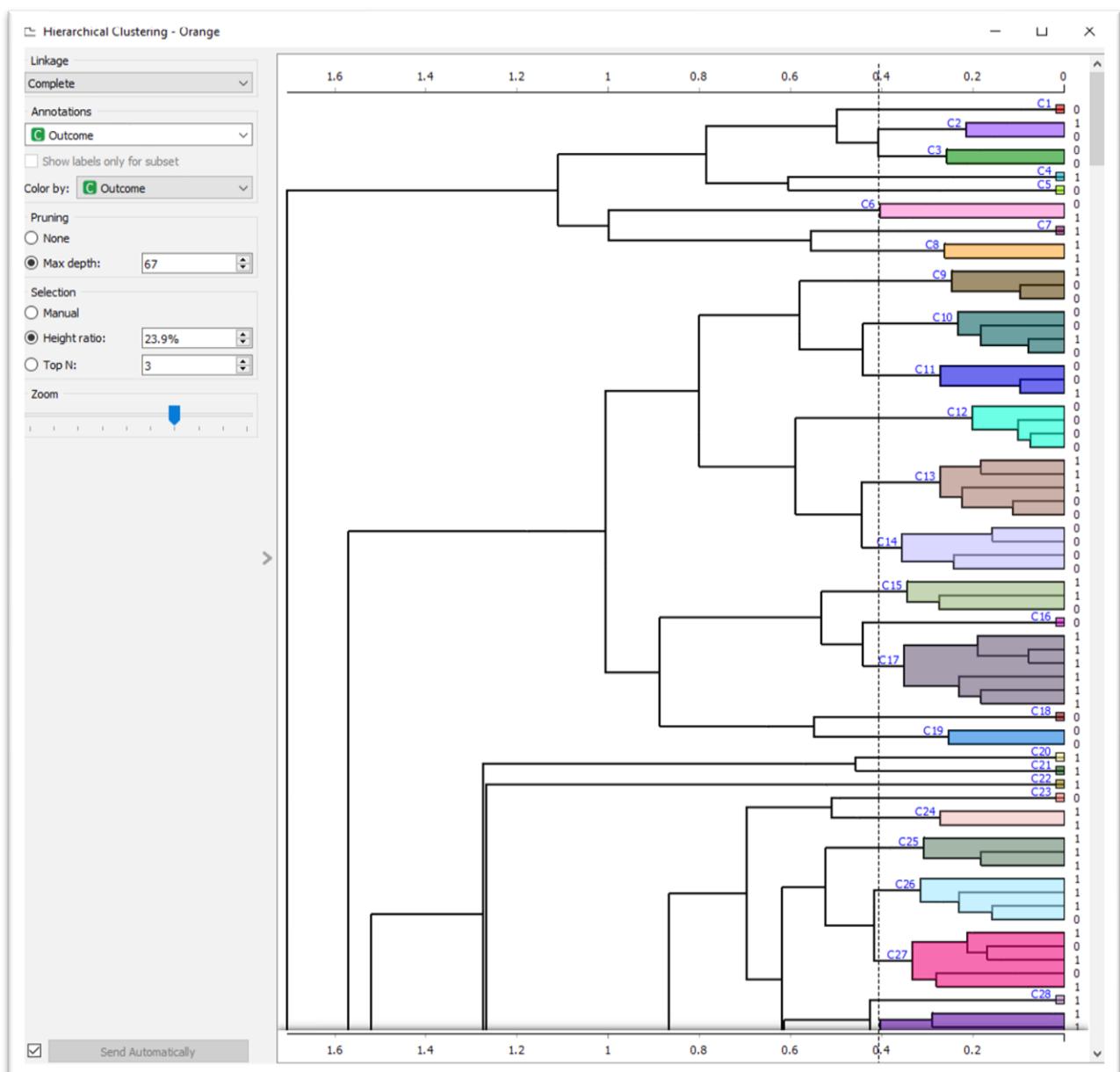
Hierarhiskās klasterizācijas algoritmam ir jāveic vismaz 3 eksperimenti, brīvi pārvietojot atdalošo līniju un analizējot, kā mainās klasteru skaits un saturs
21. att



Pirmajā eksperimentā dendogrammai ir šādas vērtības: linkare - average, max depth - 45, height ratio - 54,6 %, dalījuma līnija - 0,6.

Kopumā ir 18 klasteri. 1. klasterī ir tikai 1 objekts. 2. klasterī ir 2 objekti. 3. klasterī ir 4 objekti. 4. klasterī ir 3 objekti. 5. klasterī ir 1 objekts. 6. klasterī ir 1 objekts. 7. klasterī ir 2 objekti. 8. klasterī ir 7 objekti. 9. klasterī ir 24 objekti. 10. klasterī ir 2 objekts. (21. att.)

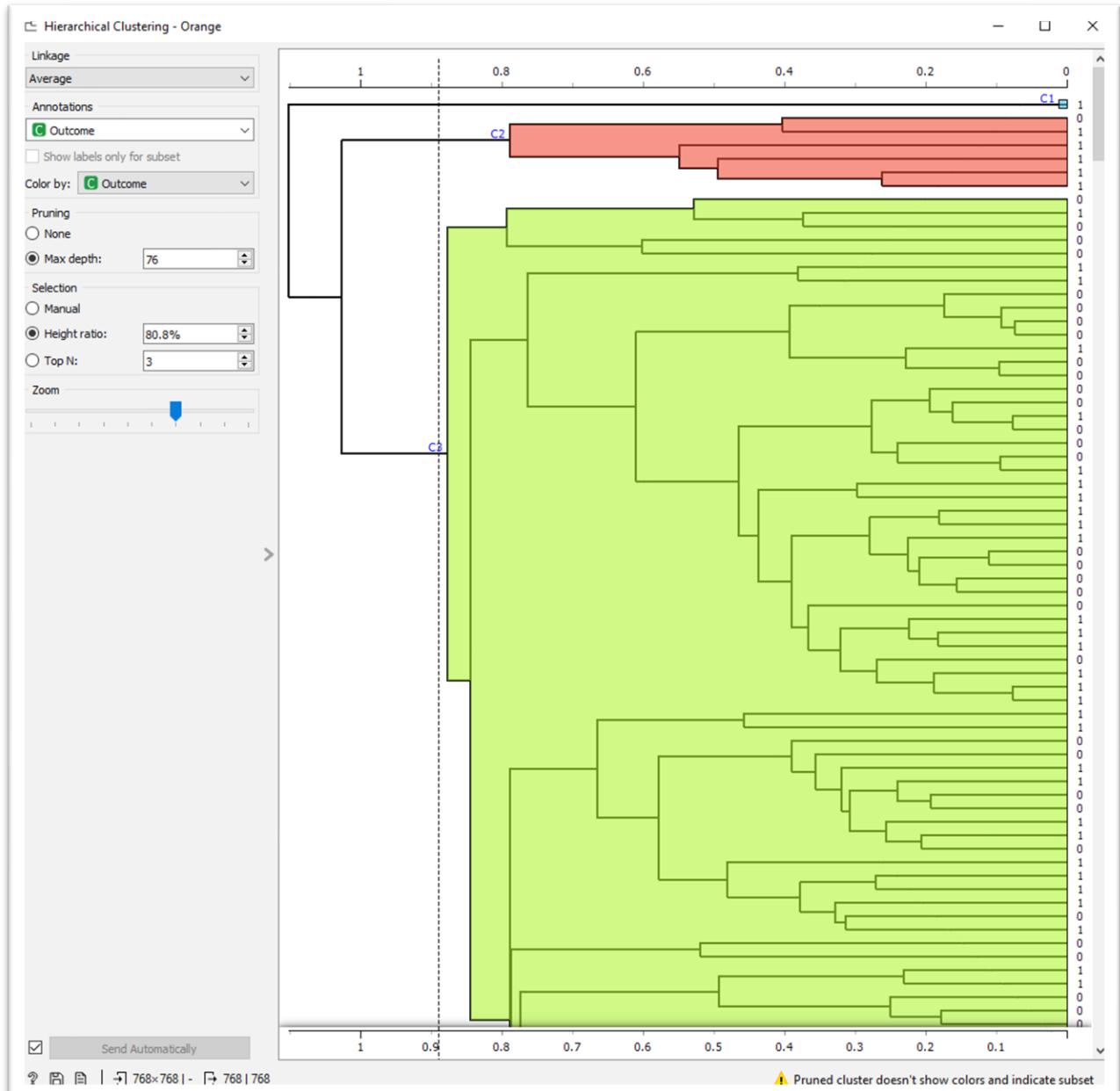
22. att.



Otrajā eksperimentā dendogrammai ir šādas vērtības: linkare - complete, max depth - 67, height ratio – 23.5 %, dalījuma līnija - 0,4.

Kopumā ir 189 klasteri. 1. klasterī ir 1 objekts. 2. klasterī ir 2 objekti. 3. klasterī ir 2 objekti. 4. klasterī ir 1 objekti. 5. klasterī ir 1 objekts. 6. klasterī ir 1 objekts. 7. klasterī ir 1 objekti. 8. klasterī ir 1 objekti. 9. klasterī ir 2 objekti. 10. klasterī ir 3 objekts. (22.att.)

23. att.

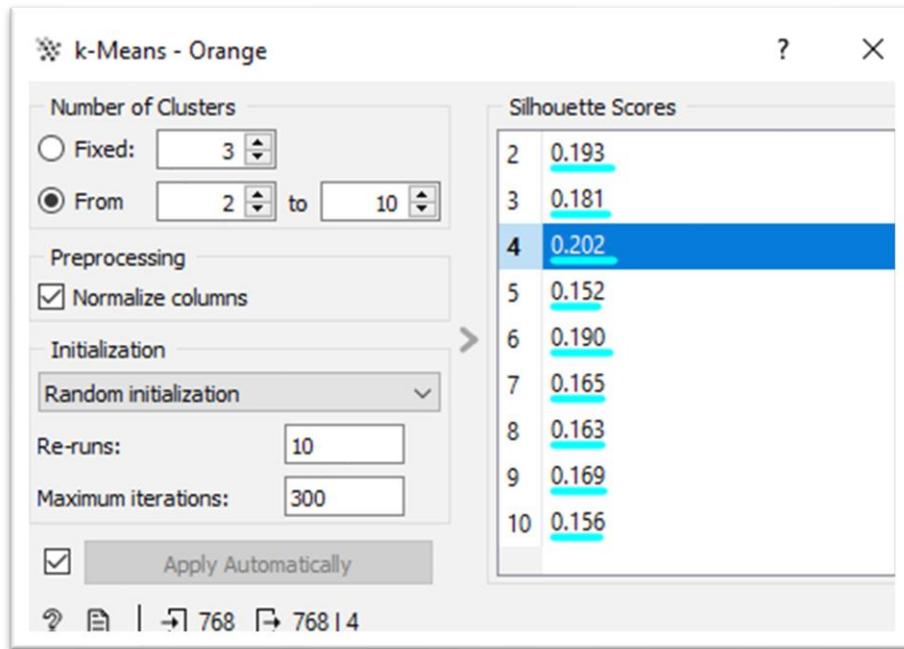


Trešajā eksperimentā dendogrammai ir šādas vērtības: linkare - average, max depth - 76, height ratio – 81.7 %, dalījuma līnija - 0.9.

Kopumā ir 2 klasteri. 1. klasterī ir 6 objekti. 2. klasterī ir 761 objekti. (23.att.)

K-VIDĒJO ALGORITMS

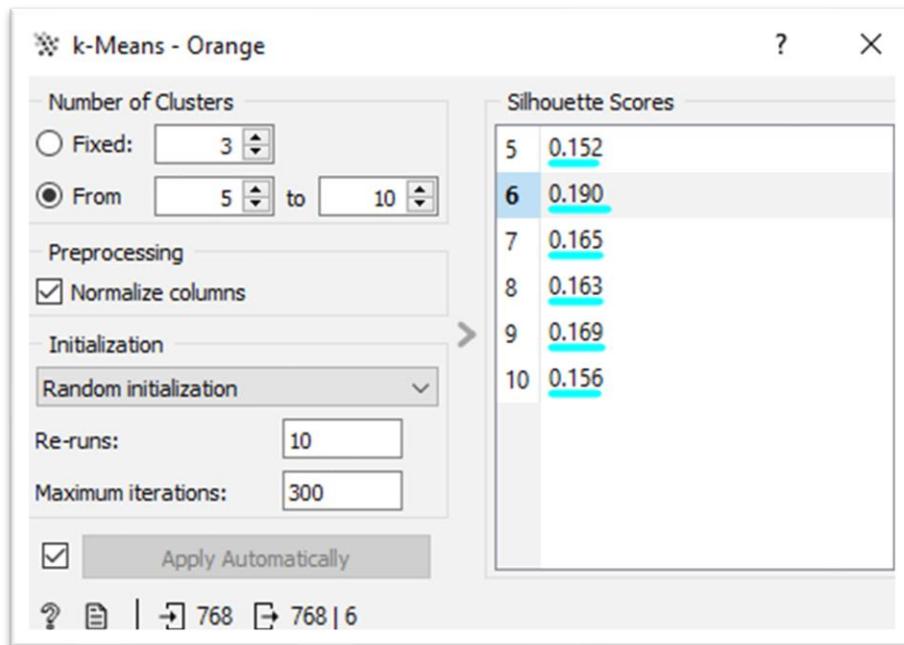
24.att.



Fixed – klasteru skaits; From ... To (klasteru diapazons); Initialization (random initialization/initialize with Kmeans++); Re-runs: (gadījuma kārtībā izvēlēto pozīciju skaits, no kurām sākas algoritms.); Maximum iterations (algoritmam atļauto iterāciju skaits)

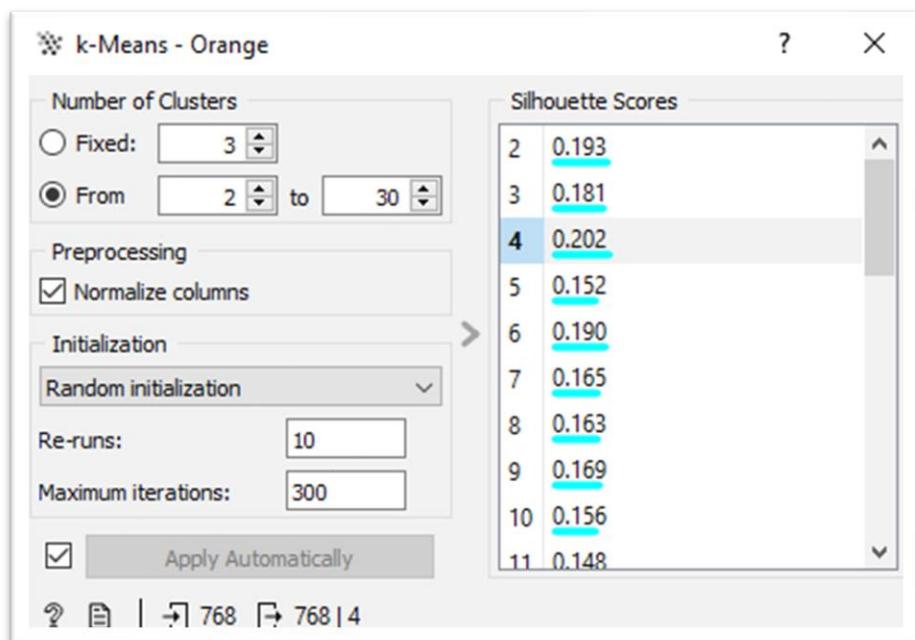
Vislabākais klasteru skaits ir 4. (24.att.)

25.att.



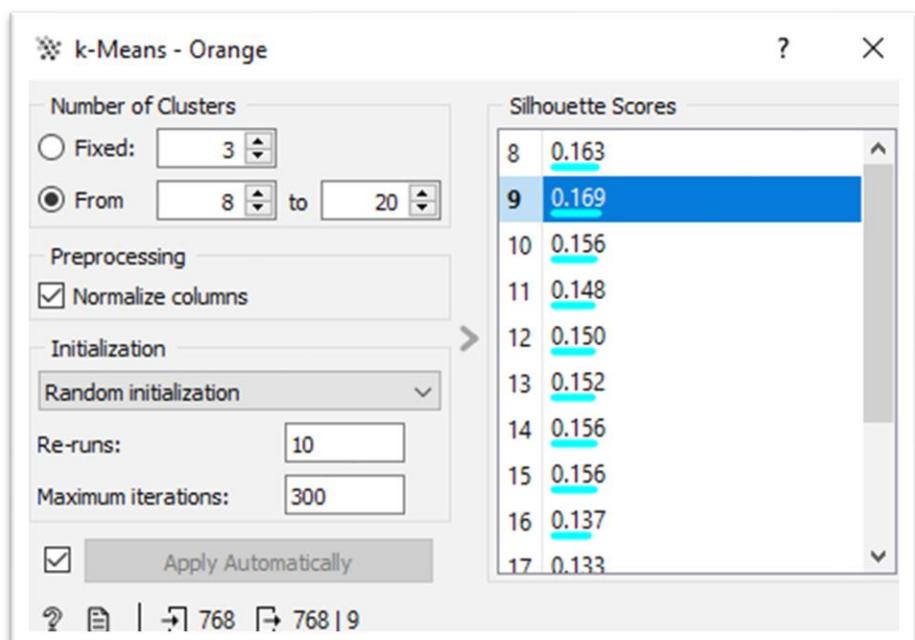
Vislabākais klasteru skaits ir 6. (25.att.)

26. att.



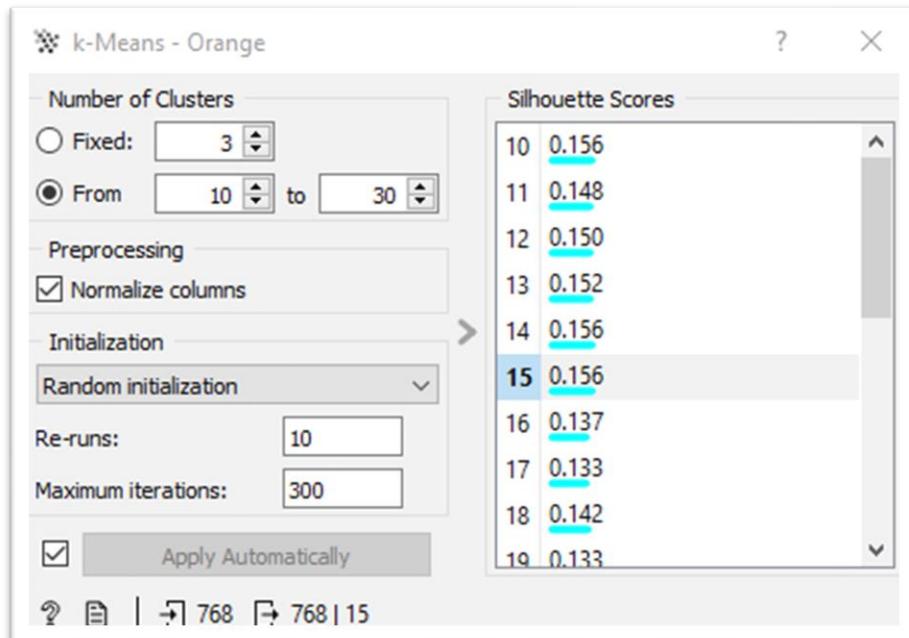
Vislabākais klasteru skaits ir 4. (26.att.)

27. att



Vislabākais klasteru skaits ir 9. (27.att.)

28.att.

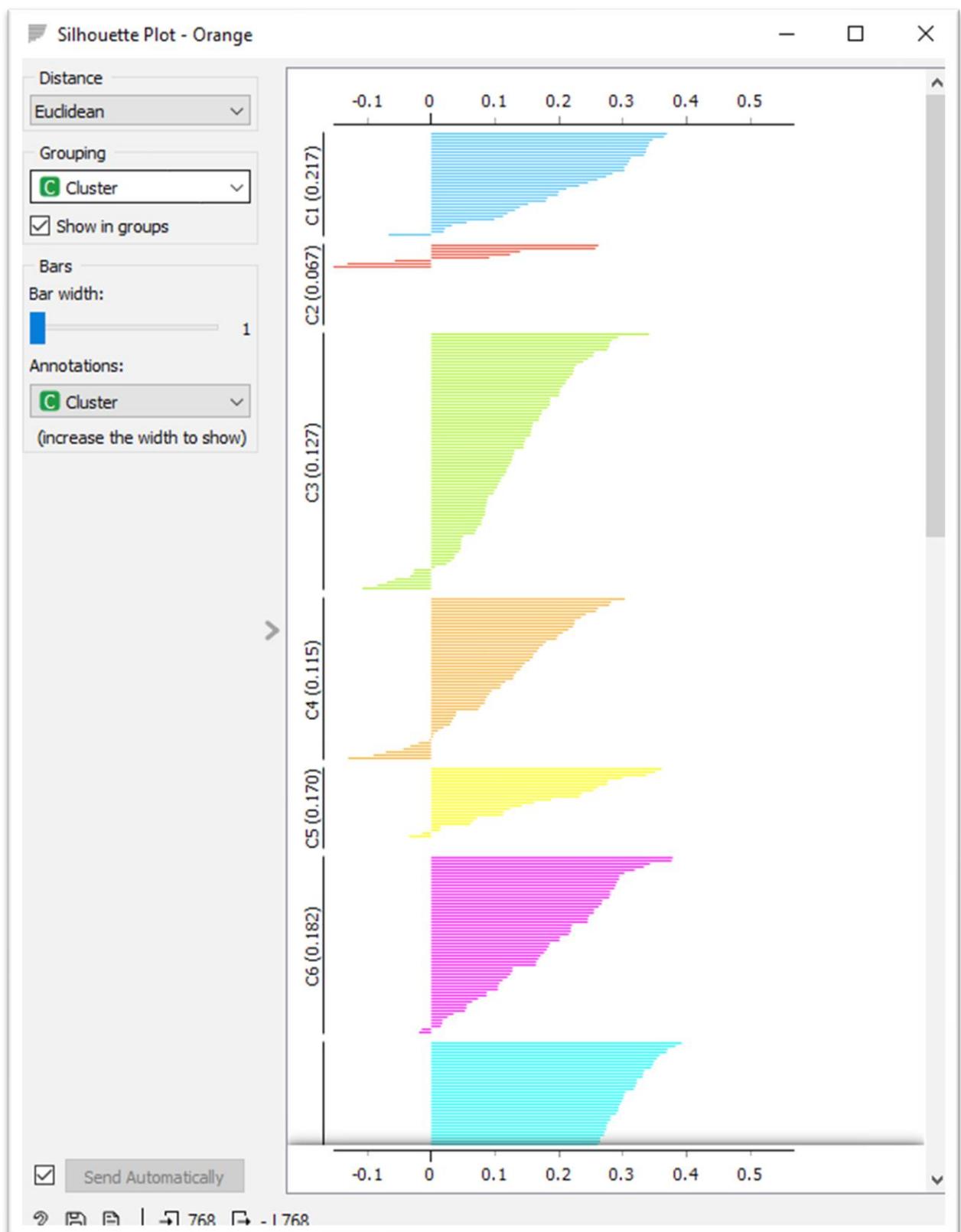


Vislabākais klasteru skaits ir 15. (28.att.)

Pamatojoties uz 5 dažādām vērtībām, varam secināt, ka labākais klasteris ir 4 ar vērtību 0,202, pārējie nav piemēroti, jo citi skaitļi ir mazāki par 0,202.

SIHOUETTE PLOT

29. att.



redzams, ka klasteri ir sadalīti diezgan pozitīvi un labi, jo klasteris lielākoties ir pozitīvās vērtībās. Ir pāris klasteri, kur tas pārsniedz negatīvās vērtības, bet tas nav ļoti būtiski. (29. att)

SECINĀJUMS

Pamatojoties uz visiem rezultātiem, varam secināt, ka labākais klasteris ir 4, jo tā vērtība ir 0,202. Šādu rezultātu varam pieņemt, jo sākotnēji nav daudz datu, kuros varētu redzēt lielas atšķirības starp vērtībām. Protams, galu galā klasteri nav perfekti nodalīti, un to var redzēt tālāk darbā, jo 0,202 ir mazs skaitlis, bet negatīvās vērtības nav tik lielas, kā iepriekš pieņemts.

III DAĻA – PĀRRAUDZĪTĀ MAŠĪNMACIŠANĀS

3.tab Ir jāsadala datu kopa apmācību un testa datu kopās.

Klases nosaukums	Datu kopa apmācība	Testa datu kopas
Sievietēm nav diabēta (0)	349	151
Sievietēm ir diabēts (1)	189	79

INFORMĀCIJA PAR ALGORITMIEM

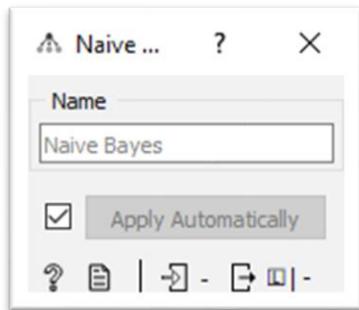
Tika izvēlēti trīs algoritmi: Naive Bayes, kNN, neironu tīkls.

Naive Bayes algoritms. (30. att.)

Ātrs un vienkāršs varbūtiskais klasifikators, kura pamatā ir Bejas teorēma ar pieņēmumu par pazīmju neatkarību. Iepriekšēja apstrāde

Naive Bayes pēc noklusējuma izmanto pirmapstrādi, ja nav norādīti citi pirmapstrādes procesori. Tā veic tās šādā secībā: 1.dzēš tukšās kolonnas 2.diskretizē skaitliskās vērtības 4 bintos ar vienādu frekvenci 3.Lai noņemtu noklusējuma pirmapstrādi, pievienojiet izglītojamajam tukšu Preprocess widžetu. Algoritmam nav hiperparametru un to vērtības. Šis algoritms tika izvēlēts tā paša iemesla dēļ. Izmantotais informācijas avots - <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/naivebayes.html>

30. att.



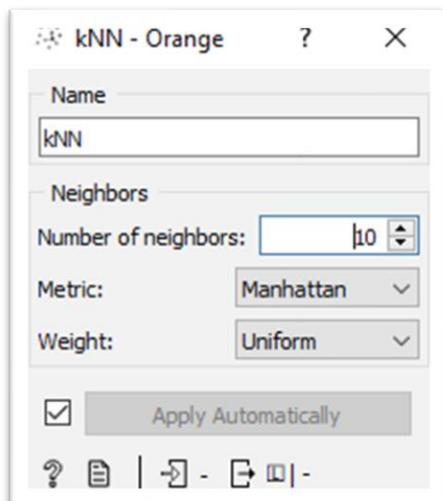
kNN algoritms. (31. att.)

kNN algoritms ir neparametrisks uzraudzītas mācīšanās klasifikators, kas izmanto tuvumu, lai klasificētu vai prognozētu atsevišķu datu punktu grupēšanu. Lai gan to var izmantot gan regresijas, gan klasifikācijas problēmām, parasti to izmanto kā klasifikācijas algoritmu, pamatojoties uz pienēmumu, ka līdzīgus punktus var atrast tuvu vienu otram. Ir tādi hiperparametri kā Number of neighbors. Metric (Euclidean, Manhattan, Maximal, Mahalanobis). Weight (Uniform, Distance). Šis algoritms tika izvēlēts, jo tas bija pieejams kursā un ir labāk pazīstams un saprotams. Izmantotie informācijas avoti -

<https://www.ibm.com/topics/knn#:~:text=The%20k-nearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,%20an%20individual%20data%20point>.

<https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html>

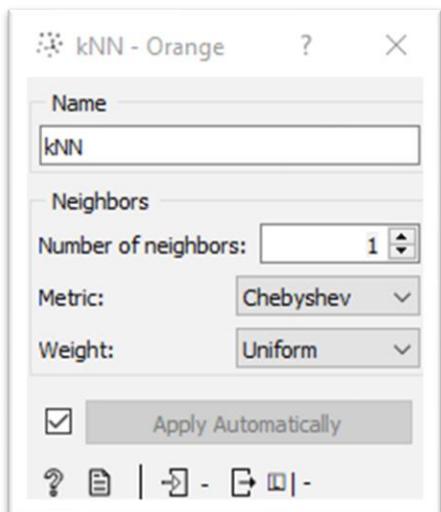
31. att.



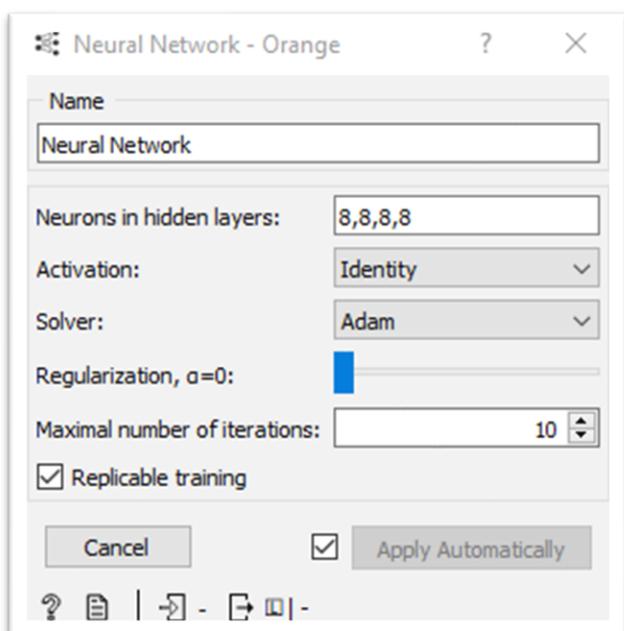
TESTI

1. tests

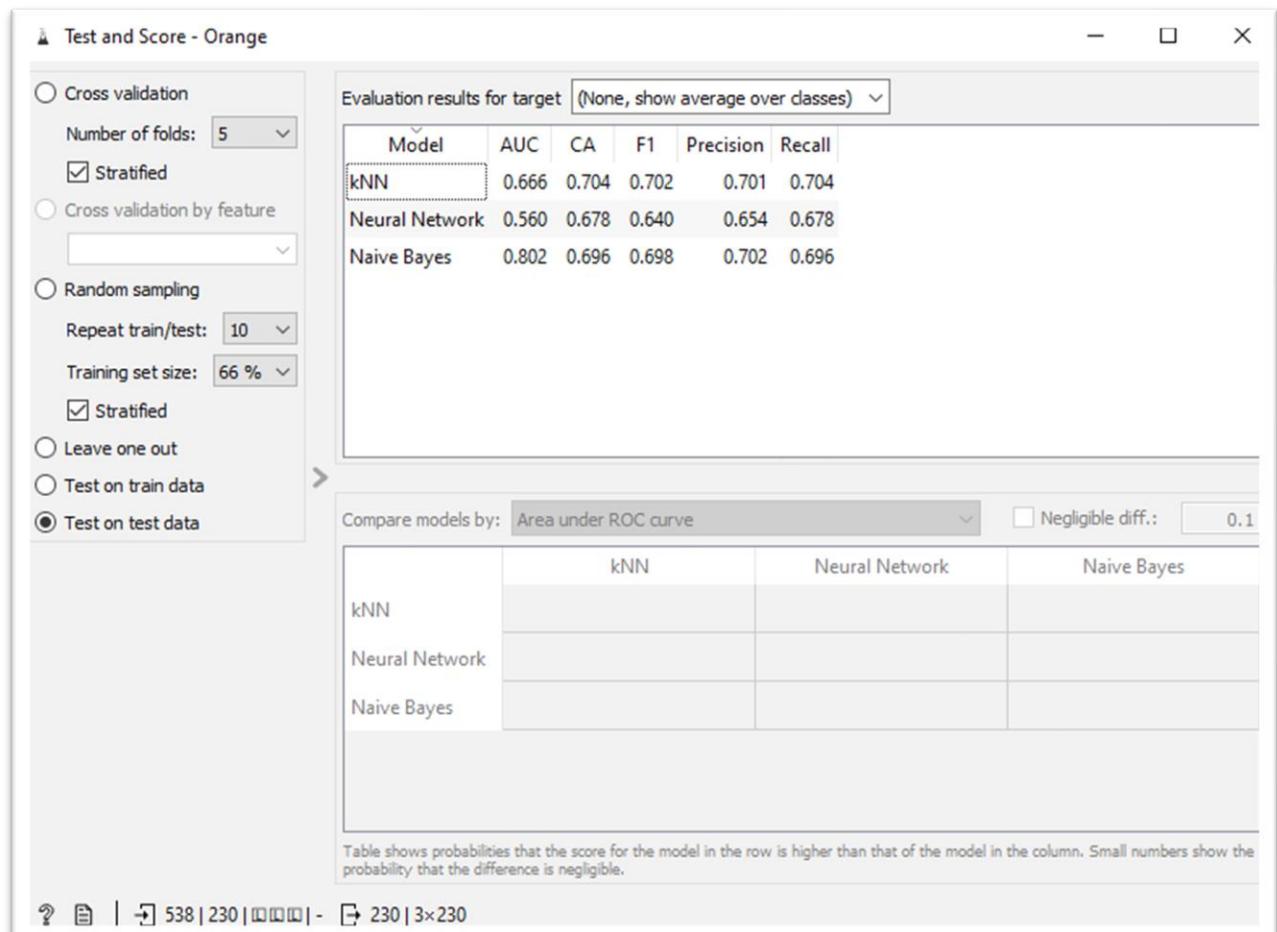
32.att. kNN algoritms



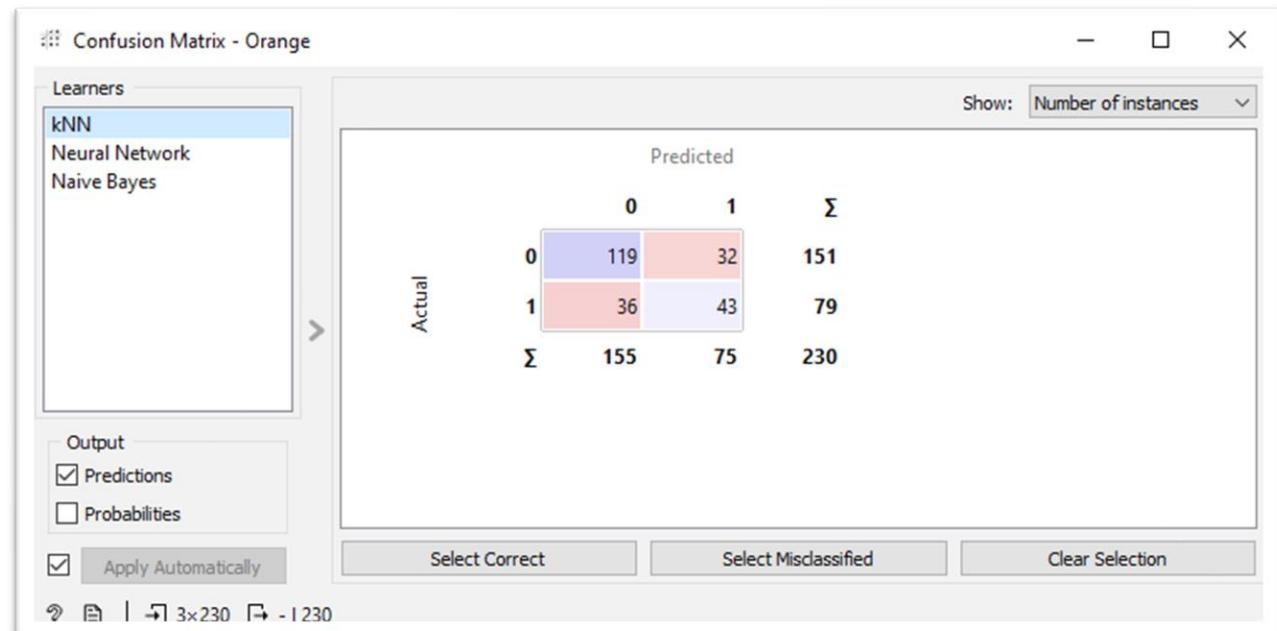
33. att. neironu tīkls



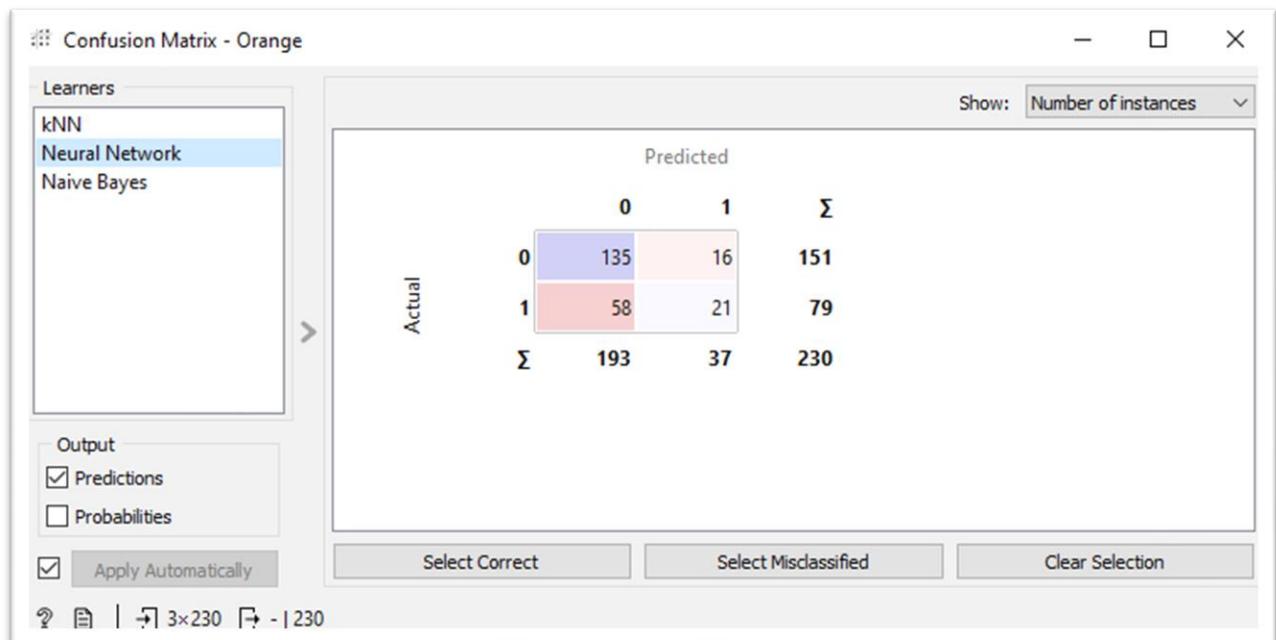
34. att. rezultāti



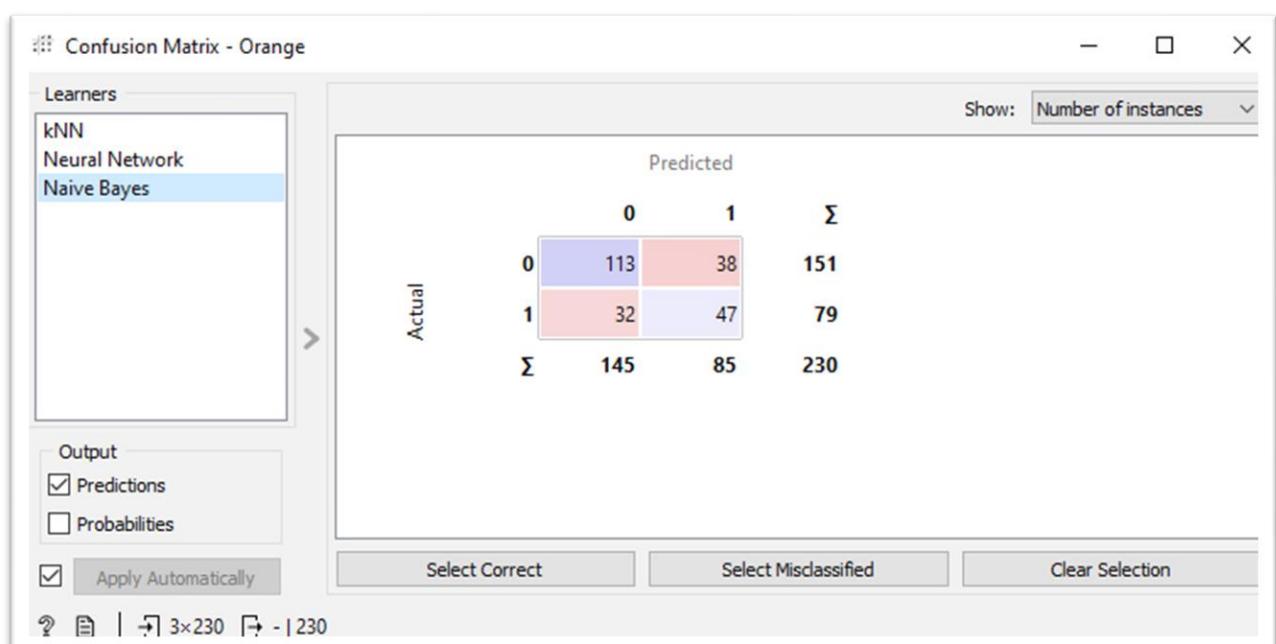
35.att. kNN matrix



36.att. neural network matrix

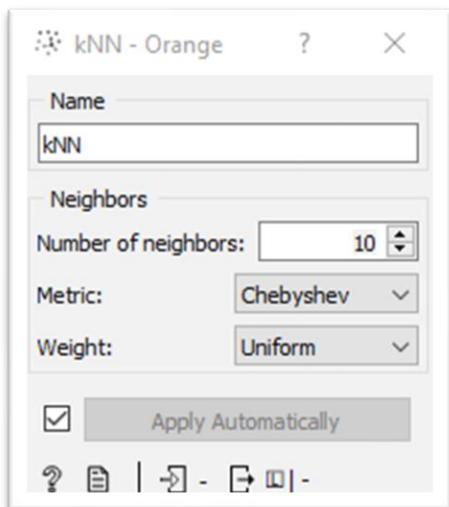


37.att. naive bayes matrix

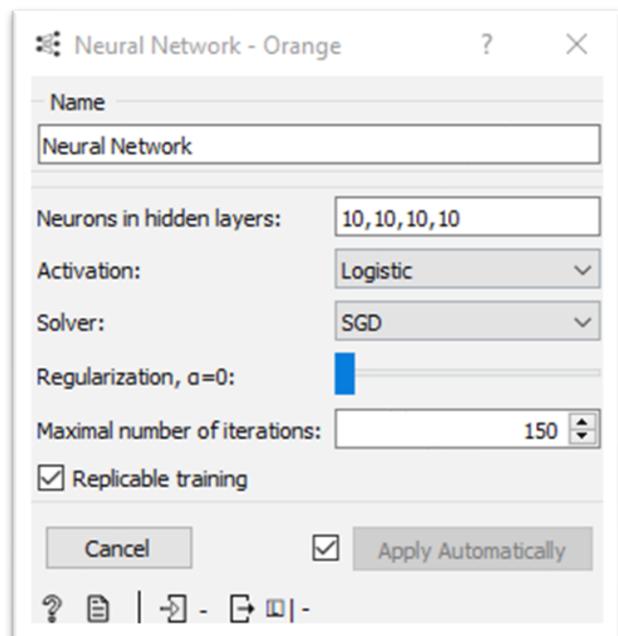


2. tests

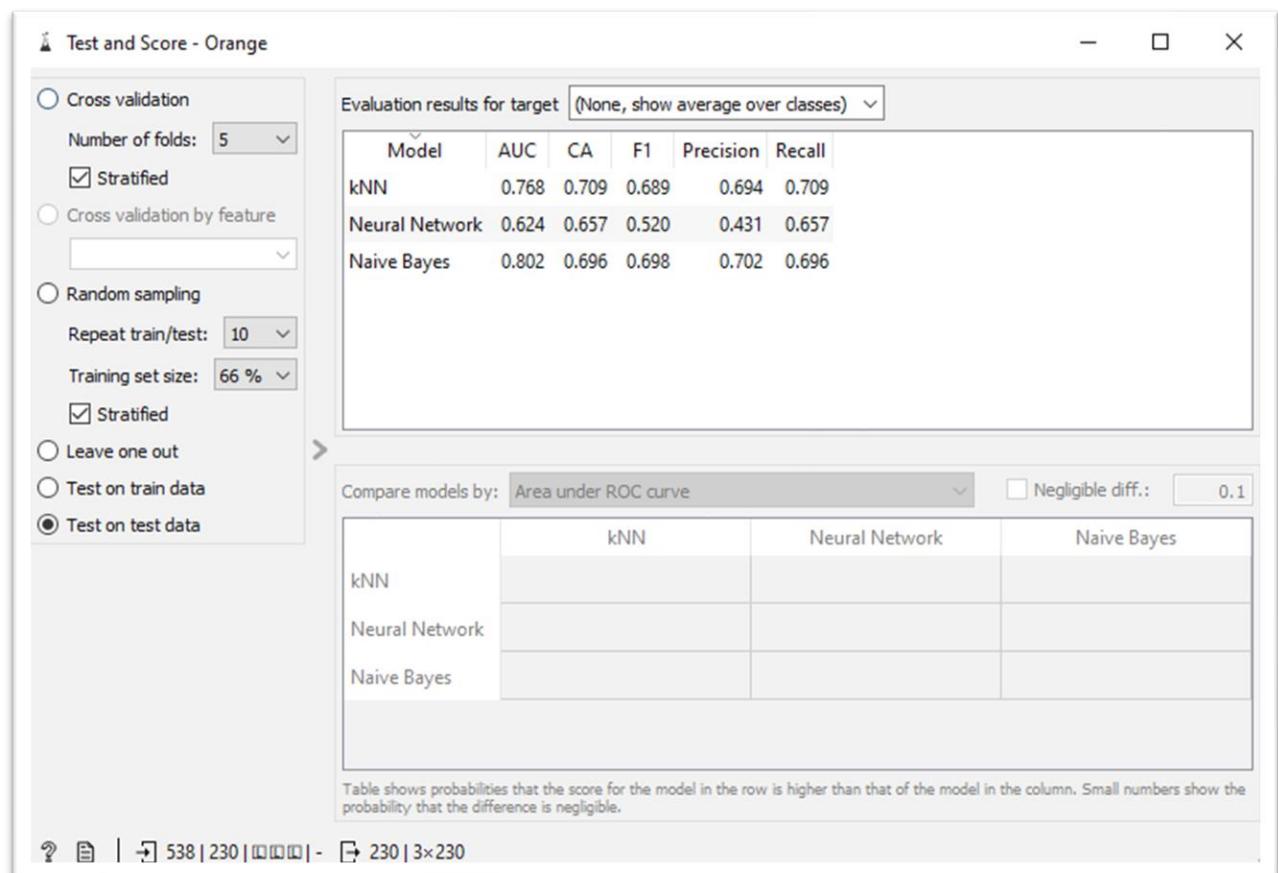
38.att. kNN algoritms



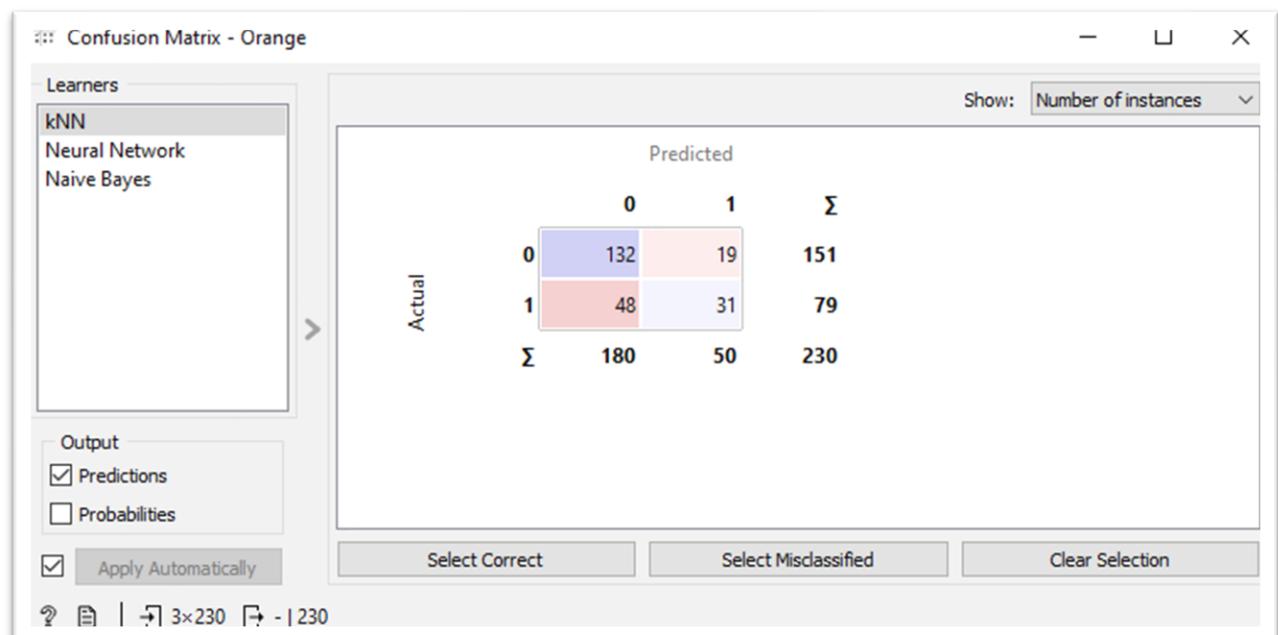
39.att neironu tīkls



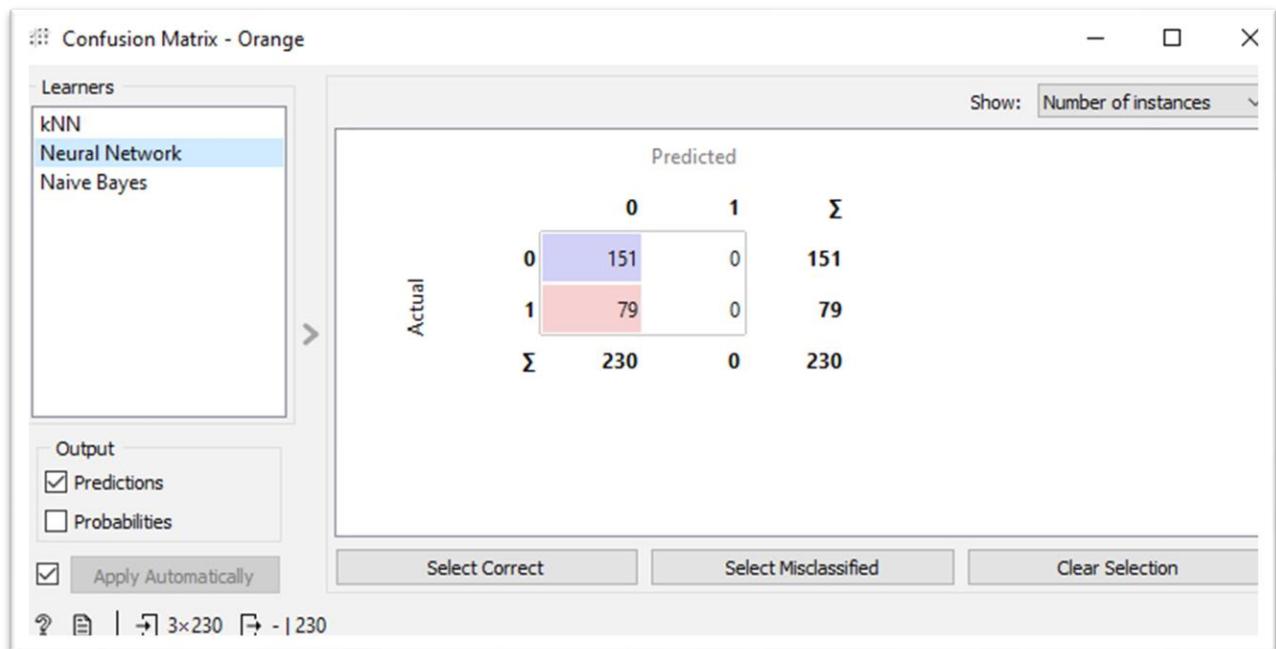
40.att rezultāti



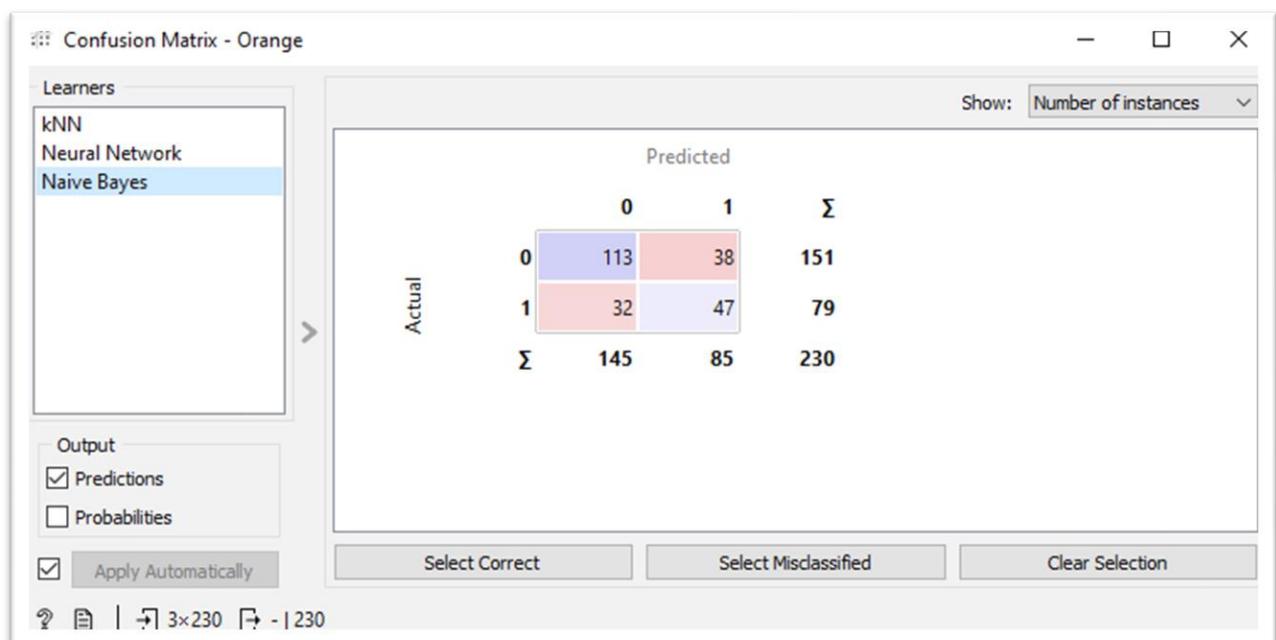
41.att. kNN matrix



42. att. neural network

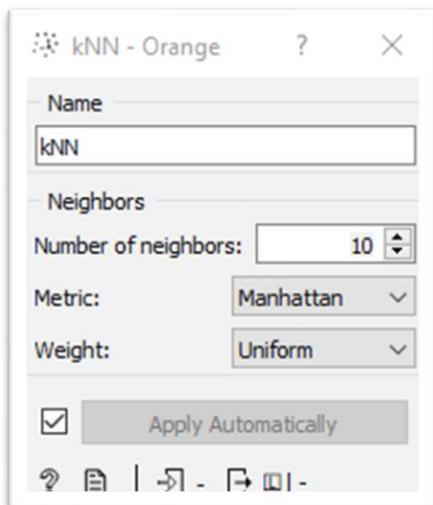


43.att. naive bayes matrix

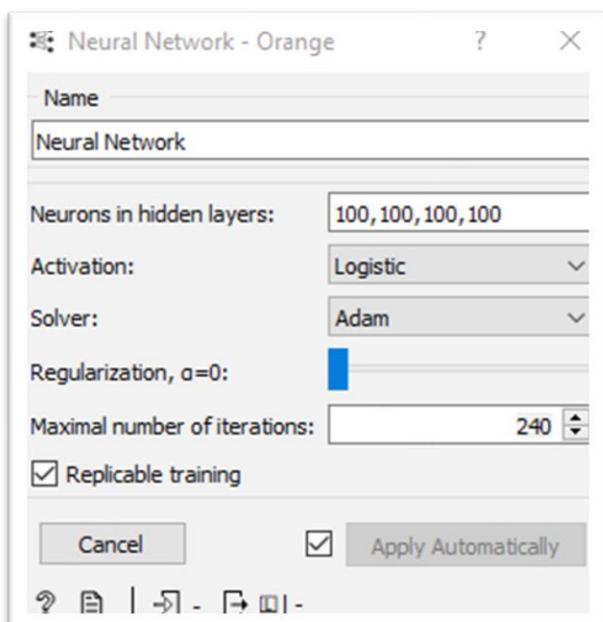


3. tests

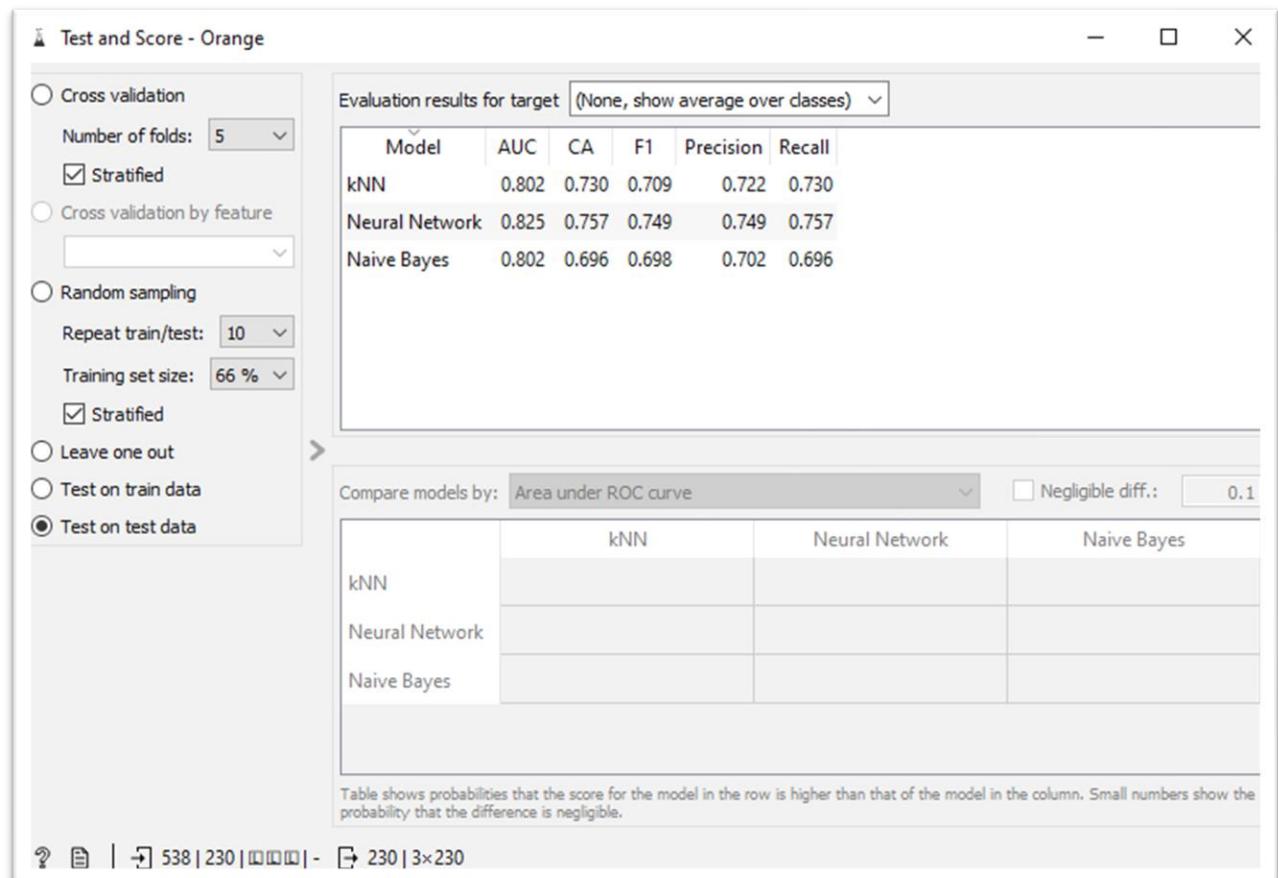
44. att. kNN algoritms



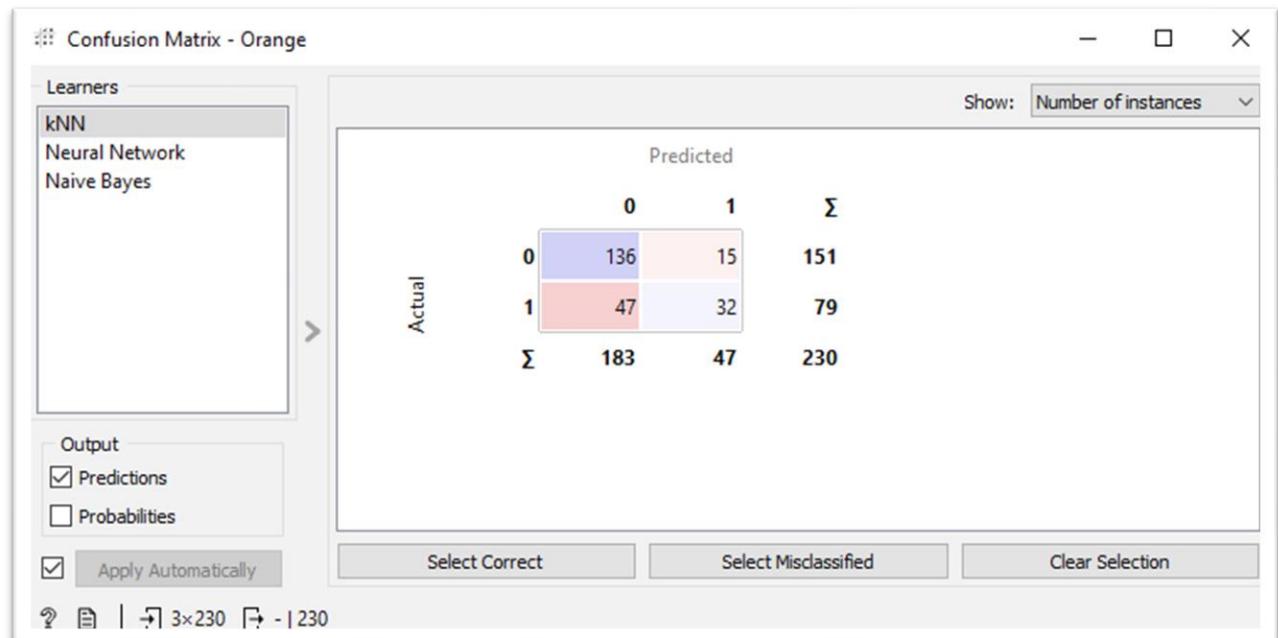
45. att. neironu tīkls



46.att. rezultāti



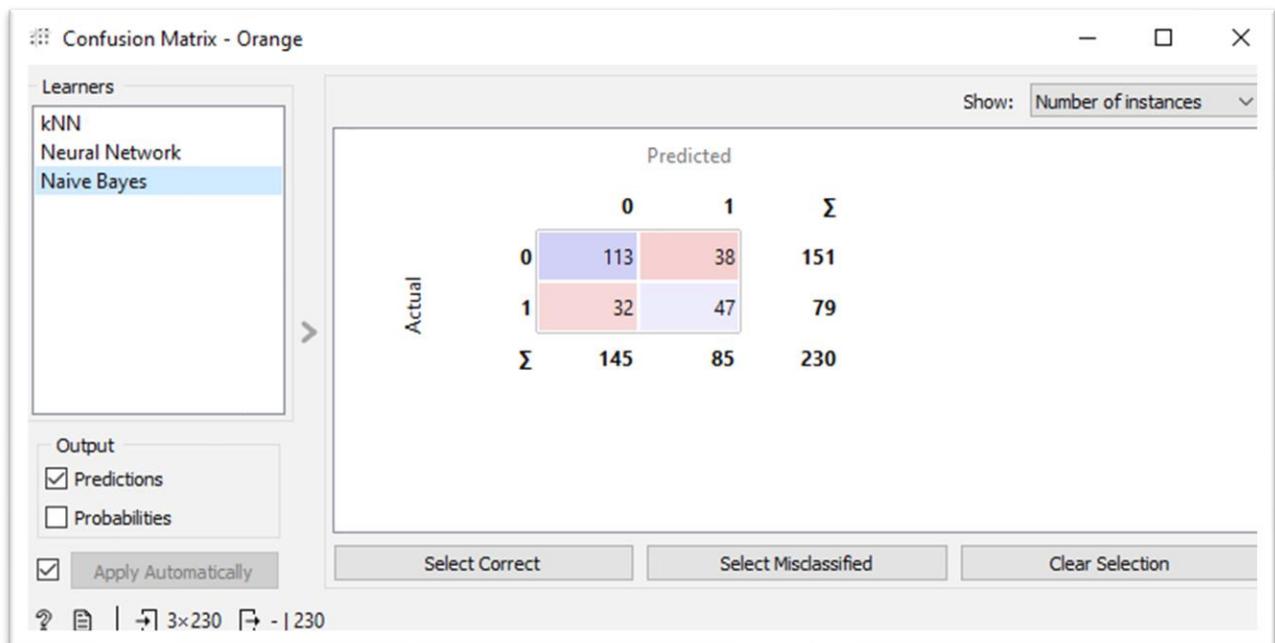
47.att. kNN matrix



48.att. neural network matrix



49. att. naive bayes matrix



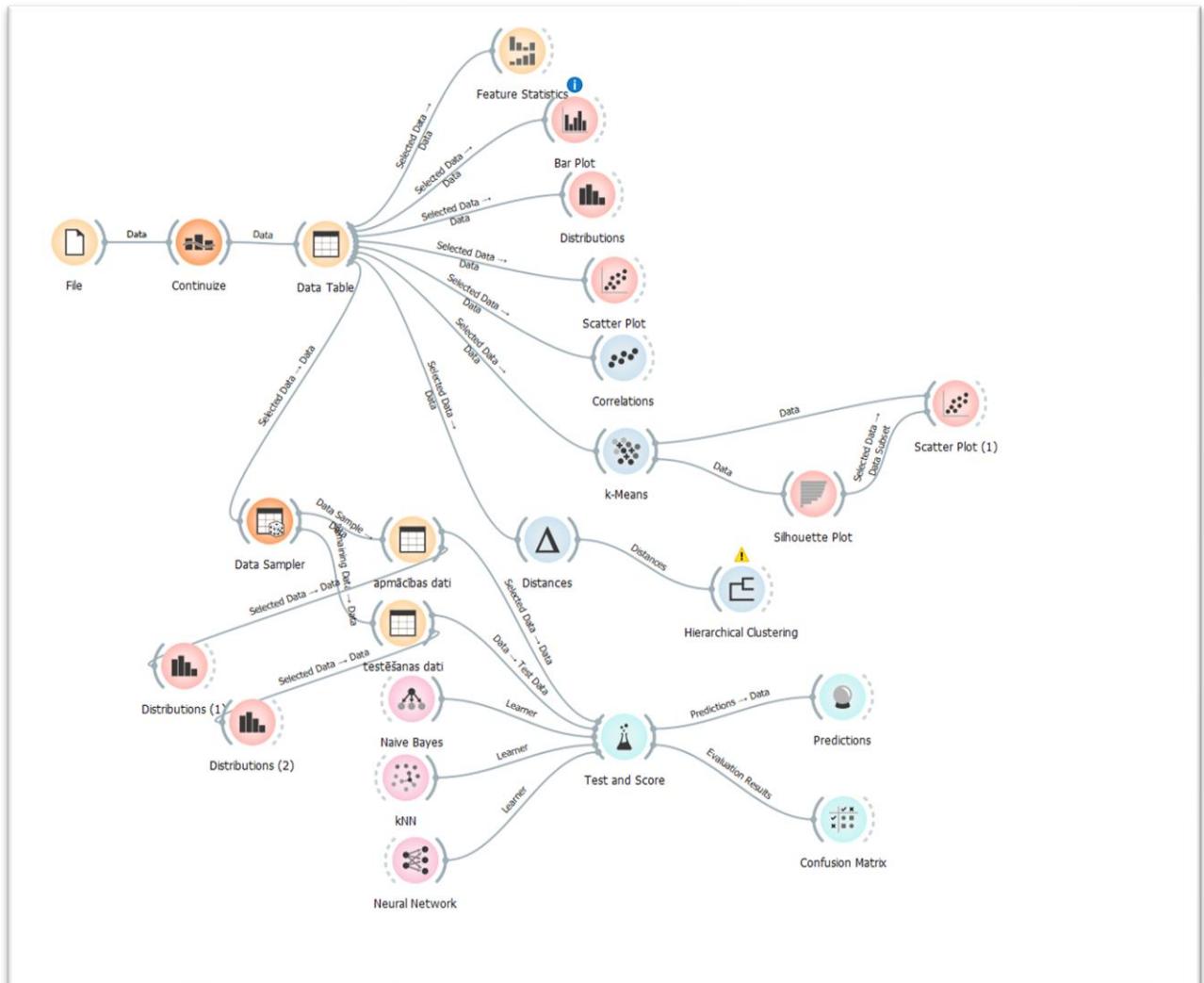
TESTU REZULTĀTU SALĪDZINĀJUMS UN ANALĪZE

Šajā testā mērķis bija tuvoties 1. Tas tika sasniegts tikai trešajā testā. Sliktākais rezultāts tika sasniegts 1. testā. Lai iegūtu galīgos rezultātus, kNN un neuronu tīkla algoritmi mainīja parametrus. Aplūkojot 1. testu (32, 33, 34, 35, 36, 37 att.), tika mainīti šādi hiperparametri: kNN algoritms: number of neighbors - 1, metric - chebyshev, weight- uniform. Neuronu tīkls: neurons in hidden layers - 8,8,8,8 ,

activation - identity, solver - adam, maximal number of iterations - 10. Aplūkojot 2 .testu (38, 39, 40, 41,42, 43 att.), tika mainīti šādi hiperparametri: kNN algoritms: number of neighbors - 10, metric - chebyshev, weight- uniform. Neironu tīkls: neurons in hidden layers - 10,10,10,10 , activation - logistic, solver - sgv, maximal number of iterations - 150. Aplūkojot 3. testu (44, 45, 46,47, 48, 49 att.), tika mainīti šādi hiperparametri: kNN algoritms: number of neighbors - 10, metric - manhattan, weight- uniform. Neironu tīkls: neurons in hidden layers - 100,100,100,100 , activation - logistic, solver - adam, maximal number of iterations - 240. Analizējot matricas (3. testā) (47, 48, 49 att.) , var izdarīt šādas darbības un secinājumus. Datu pareizas klasificēšanas iespēja ar kNN algoritmu ir $(74,3 \% + 68,1 \%) / 2 = 0.62\%$. Datu pareizas klasificēšanas iespēja ar neironu tīklu ir $(78,0 \% + 67,7 \%) / 2 = 0.65\%$ Datu pareizas klasificēšanas iespēja, izmantojot Naive Bayes algoritmu, ir $(77,9\% + 55,3\%) / 2 = 0.60\%$ Datu nepareizas klasificēšanas iespēja ar kNN algoritmu ir $(31,9 \% + 25,7 \%) / 2 = 0.21\%$ Datu nepareizas klasificēšanas iespēja ar neironu tīklu ir $(32,3 \% + 22,0 \%) / 2 = 0.19\%$ Datu nepareizas klasificēšanas iespēja, izmantojot Naivās Beijes algoritmu, ir $(44,7 \% + 22,1 \%) / 2 = 0.27\%$

SECINĀJUMS

50.att



Šajā praktiskajā darbā man bija pirmā pieredze orange programmā. Visi uzdevumi ir izpildīti. Tika izveidota Orange rīka darbplūsmas atspoguļojums (50.att.). Pirmajā daļā veiksmīgi tika analizēta datu bāze. Otrajā daļā veiksmīgi tika analizēti klasteri un to sadalījums. Trešajā daļā veiksmīgi tika veikti eksperimenti ar algoritmiem. Darbs bija grūts, bet galvenokārt neuzmanības dēļ.

IZMANTOTIE INFORMĀCIJAS AVOTI

<https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>
<https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/naivebayes.html>
<https://www.ibm.com/topics/knn#:~:text=The%20k-nearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20r,of%20an%20individual%20data%20point>
<https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html>
<https://www.youtube.com/watch?v=bmwH3EcTBEM>
<https://www.youtube.com/watch?v=ojaxlQSylLr0>
<https://www.youtube.com/watch?v=UiGH4v3VKPc>