



**RĪGAS TEHNISKĀ
UNIVERSITĀTE**

RĪGAS TEHNISKĀ UNIVERSITĀTE

Datorzinātnes un informācijas tehnoloģijas fakultāte

2.Praktiskais darbs

mācību priekšmetā

“Mākslīgā intelekta pamati”

Autors: Daniels Kisels

St.ap.nr.: 211RDB368

Grupa: 9

[Links uz github](#)

2022/2023 m.g.

Saturs

Darba uzdevums	3
I daļas apraksts - Datu pirmapstrāde/izpēte.....	5
I daļas izpilde - Datu pirmapstrāde/izpēte.....	6
1. darbība	6
2., 3. un 4. Darbība	10
5. Darbība	11
Secinājumi par I daļas izpildi	16
II daļas apraksts – Nepārraudzītā mašīnmācīšanās	17
II daļas izpilde - Nepārraudzītā mašīnmācīšanās	18
1. Darbība	18
<i>Hierarhiska klasterizācija hiperparametri 14.attēlā</i>	18
<i>K-vidējo hiperparametri 15. attēls</i>	19
2. Darbība	20
3. Darbība	21
Secinājumi par II daļas izpildi	23
III daļas apraksts - Pārraudzītā mašīnmācīšanās	24
III daļas izpilde - Pārraudzītā mašīnmācīšanās	25
1. Darbība	25
2., 3., 4., un 5. Darbība	26
6. Darbība	30
Secinājumi	31

Darba 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.

Lai atrastu datu kopu darba izpildei, studenti var izmantot šādas plaši zināmās krātuves:

- UC Irvine Machine Learning Repository <https://archive.ics.uci.edu/ml/index.php>
- R Datasets on Github <https://vincentarelbundock.github.io/Rdatasets/>
- Kaggle Datasets <https://www.kaggle.com/datasets>
- Awesome Lists: Public Datasets <https://github.com/caesar0301/awesome-public-datasets>
- Yahoo! Webscope Datasets <https://webscope.sandbox.yahoo.com/?guccounter=1>
- Reddit: <https://www.reddit.com/r/datasets>

Izvēloties datu kopu, studentiem ir jāņem vērā šādi aspekti:

- ir jāizvēlas datu kopa, kas ir piemērota klasifikācijas uzdevumam. Students nedrīkst izvēlēties Iris ziedu (Iris dataset) vai Pingvīnu (Palmer Archipelago (Antarctica) penguin data) datu kopas. Turklāt 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 jāsaturs 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 daļas apraksts - 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;
 - b) ir jāizveido vismaz 2 histogrammas, kas parāda klašu atdalīšanu, pamatojoties uz interesējošām pazīmēm (atribūtiem);
 - c) ir jāatspoguļo 2 interesējošo pazīmju (atribūtu) sadalījums;
 - d) ir jāaprēķina statistiskie rādītāji (vismaz vidējās vērtības un dispersiju).

I daļas izpilde - Datu pirmapstrāde/izpēte

1. darbība

Es izvēlējos datu kopu par Ķermeņa veiktspējas datiem, jo viņa atbilst visiem iepriekš minētajiem nosacījumiem darba uzdevumu aprakstā. Datu bazē ir 15 kategorijas

Datu kopu "Body performance Data"; autors: KUKUROO3; tiek paņemta no Kaggle Datasets krātuves: <https://www.kaggle.com/datasets/kukuroo3/body-performance-data>

Autora apraksts angļu valodā:

dataset

This is data that confirmed the grade of performance with age and some exercise performance data.

columns

data shape : (13393, 12)

- age : 20 ~64
- gender : F,M
- height_cm : (If you want to convert to feet, divide by 30.48)
- weight_kg
- body fat_%
- diastolic : diastolic blood pressure (min)
- systolic : systolic blood pressure (min)
- gripForce
- sit and bend forward_cm
- sit-ups counts
- broad jump_cm
- class : A,B,C,D (A: best) / stratified

Source

[link \(Korea Sports Promotion Foundation\)](#)

Some post-processing and filtering has done from the raw data.

Tulkots autora apraksts latviešu valodā:

datu kopa

Tie ir dati, kas apstiprināja snieguma pakāpi atkarībā no vecuma un dažī dati par fizisko aktivitāšu veikumu.

koloni

datu forma : (13393, 12)

- vecums : 20 ~ 64 gadi
- dzimums : S,V

- garums_cm
- svars_kg
- ķermeņa tauku_%
- diastoliskais : diastoliskais asinsspiediens (min)
- sistoliskais : sistoliskais asinsspiediens (min)
- satvērienaSpēks
- sēdēt un noliekties uz priekšu_cm
- sēdēšanas pietupienu skaits
- platais lēciens_cm
- klase : A,B,C,D (A: labākais) / stratificēta

Avots

saite (Korejas Sporta veicināšanas fonds)

No neapstrādātiem datiem ir veikta pēcapstrāde un filtrēšana.

Uz 1. attelā var redzēt kā izskatas dati Kaggle platformā:

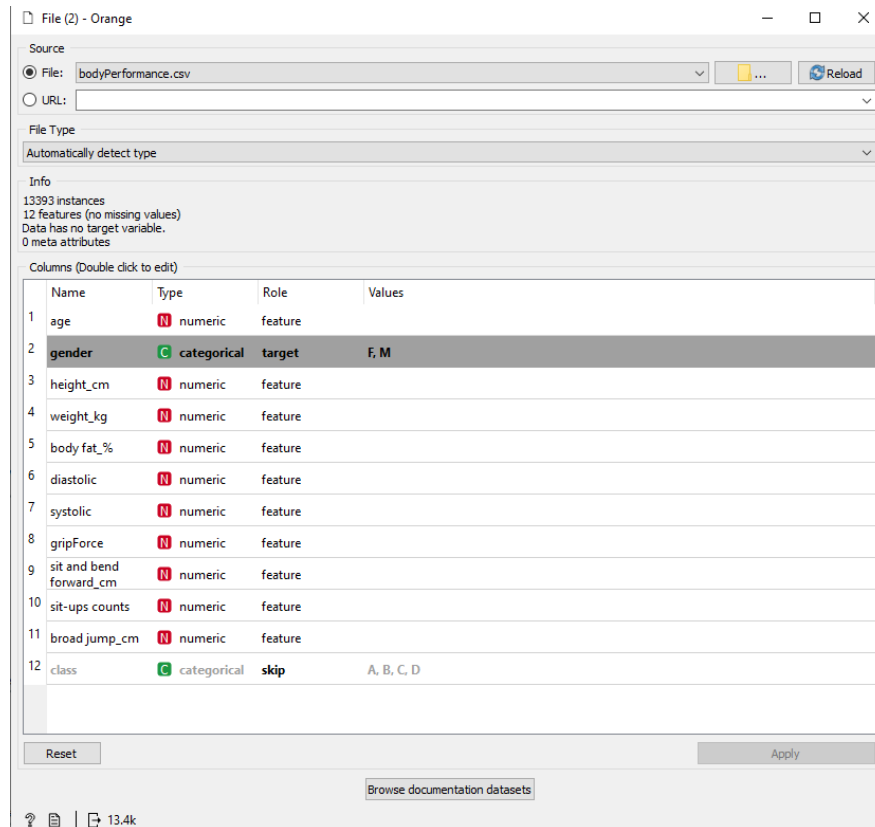
bodyPerformance.csv (761.84 kB) ↓ ↗ >

Detail Compact Column 12 of 12 columns

# age	gender	# height_cm	# weight_kg	# body fat_%	# diastolic	# systolic	# gripForce	# sit and bend forw
	M 63% F 37%							
21		125	26.3	3	0	0	0	-25
27.0	M	172.3	75.24	21.3	80.0	130.0	54.9	18.4
25.0	M	165.0	55.8	15.7	77.0	126.0	36.4	16.3
31.0	M	179.6	78.0	20.1	92.0	152.0	44.8	12.0
32.0	M	174.5	71.1	18.4	76.0	147.0	41.4	15.2
28.0	M	173.8	67.7	17.1	70.0	127.0	43.5	27.1
36.0	F	165.4	55.4	22.0	64.0	119.0	23.8	21.0
42.0	F	164.5	63.7	32.2	72.0	135.0	22.7	0.0
33.0	M	174.9	77.2	36.9	84.0	137.0	45.9	12.3
54.0	M	166.8	67.5	27.6	85.0	165.0	40.4	18.6
28.0	M	185.0	84.6	14.4	81.0	156.0	57.9	12.1
42.0	M	169.2	65.4	19.3	63.0	110.0	43.5	16.0
57.0	F	153.0	49.0	20.9	69.0	106.0	21.5	30.0
27.0	F	156.0	53.9	35.5	69.0	116.0	23.1	13.1
22.0	M	175.7	67.9	11.3	71.0	103.0	52.5	19.2
24.0	M	181.0	84.4	20.4	80.0	120.0	48.9	7.2
45.0	F	159.0	63.1	30.9	93.0	144.0	34.1	19.0
25.0	F	164.2	66.6	30.2	82.0	120.0	25.7	22.9
26.0	M	179.9	71.5	9.7	64.0	135.0	59.6	17.8
26.0	M	169.2	70.6	21.0	63.0	129.0	41.3	15.1
21.0	F	162.7	47.2	18.9	78.0	133.0	25.4	20.5
25.0	F	161.7	63.36	31.3	89.0	128.0	25.0	10.7

1. attēls

Uz 2. un 3. attēlā ir redzams kā izskatās dati Orange platformā:



2. attēls Atribūti Orange rīkā

Data Table - Orange

Info

13393 instances (no missing data)
10 features
Target with 2 values
No meta attributes.

Variables

☒ Show variable labels (if present)

☐ Visualize numeric values

☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

☒ Send Automatically

	gender	age	height_cm	weight_kg	body_fat_%	diastolic	systolic	gripForce	and bend forward_cm	sit-ups count
1	M	27	172.3	75.24	21.3	80.0	130.0	54.90	18.40	
2	M	25	165.0	55.80	15.7	77.0	126.0	36.40	16.30	
3	M	31	179.6	78.00	20.1	92.0	152.0	44.80	12.00	
4	M	32	174.5	71.10	18.4	76.0	147.0	41.40	15.20	
5	M	28	173.8	67.70	17.1	70.0	127.0	43.50	27.10	
6	F	36	165.4	55.40	22	64.0	119.0	23.80	21.00	
7	F	42	164.5	63.70	32.2	72.0	135.0	22.70	0.80	
8	M	33	174.9	77.20	36.9	84.0	137.0	45.90	12.30	
9	M	54	166.8	67.50	27.6	85.0	165.0	40.40	18.60	
10	M	28	185.0	84.60	14.4	81.0	156.0	57.90	12.10	
11	M	42	169.2	65.40	19.3	63.0	110.0	43.50	16.00	
12	F	57	153.0	49.00	20.9	69.0	106.0	21.50	30.00	
13	F	27	156.0	53.90	35.5	69.0	116.0	23.10	13.10	
14	M	22	175.7	67.90	11.3	71.0	103.0	52.50	19.20	
15	M	24	181.0	84.40	20.4	80.0	120.0	48.90	7.20	
16	F	45	159.0	63.10	30.9	93.0	144.0	34.10	19.00	
17	F	25	164.2	66.60	30.2	82.0	120.0	25.70	22.90	
18	M	26	179.9	71.50	9.7	64.0	135.0	59.60	17.80	
19	M	26	169.2	70.60	21	63.0	129.0	41.30	15.10	
20	F	21	162.7	47.20	18.9	78.0	133.0	25.40	20.50	
21	F	25	161.7	63.36	31.3	89.0	128.0	25.00	10.70	
22	F	59	155.9	62.70	30.2	76.0	143.0	36.80	29.10	
23	M	38	166.7	67.30	23.2	70.0	111.0	26.10	19.70	
24	M	44	170.0	63.30	12.9	65.0	115.0	44.50	11.60	
25	F	23	164.1	59.40	29.6	91.0	126.0	24.60	27.50	
26	M	62	169.0	70.70	30.5	96.0	146.0	39.30	4.00	
27	F	47	158.3	53.50	29.2	70.0	117.0	25.90	8.10	
28	M	48	175.8	84.50	31.4	83.0	125.0	33.80	3.70	
29	M	36	176.0	81.30	24.5	81.0	139.0	46.20	8.10	
30	F	50	159.8	57.10	24.4	63.0	103.0	30.80	24.40	
31	M	25	170.9	70.70	15.7	80.0	127.0	36.40	26.40	
32	M	26	176.7	77.20	16.3	66.0	129.0	50.00	9.10	
33	F	28	159.5	51.54	24.5	82.0	123.0	37.20	23.00	
34	M	30	172.1	79.50	26.7	91.0	148.0	34.70	-2.00	

13.4k | 13.4k | 13.4k

3. attēls Dati Orange rīkā

Datu bazē ietver sevī 12 pazīmju vērtības, kuri attiecas katram atsevišķam cilvēkam:

1. Atribūts: Vecums(līdz 64 gadu)
2. Atribūts: Dzimums(vīrietis vai sieviete)
3. Atribūts: Garums, centimetros(līdz 195 cm)
4. Atribūts: Svars, kilogramos(līdz 138 kg)
5. Atribūts: Tauki ķermeņā, procentos (līdz 79%)
6. Atribūts: Diastoliskais asinsspiediens minūtes (līdz 156)
7. Atribūts: Sistaliskais asinsspiediens minūtes (līdz 201)
8. Atribūts: Satveriena spēks(līdz 70 kg)
9. Atribūts: Attālums uz kuru cilvēks var noliekties uz priekšu sežot(līdz 213 cm)
10. Atribūts: Pietupienu skaits (līdz 80 reizēm)
11. Atribūts: Plašs lēciens centimetros(līdz 300 cm)
12. Atribūts: Klase, pie kura cilvēks attiecās(no labāka – A līdz sliktāka - D)

2., 3. un 4. Darbība

Es samazināju 36393 ierakstus līdz 803 ierakstiem lai pārraudzītais un nepārraudzītais mašīnmācīšanās strādātu atrāk. Dati tiek attēloti pēc nejaušības principa tāpēc tās neitekmēs mašīnmācīšanas kvalitātei. Arī dzimumā atribūtam bija tekstveida vērtības, tos es samainīju uz skaitliskiem: 0 – Sieviete, 1 – Vīrietis(4. attēls). Šajā datu kopā nebija trūkstošu vērtības.

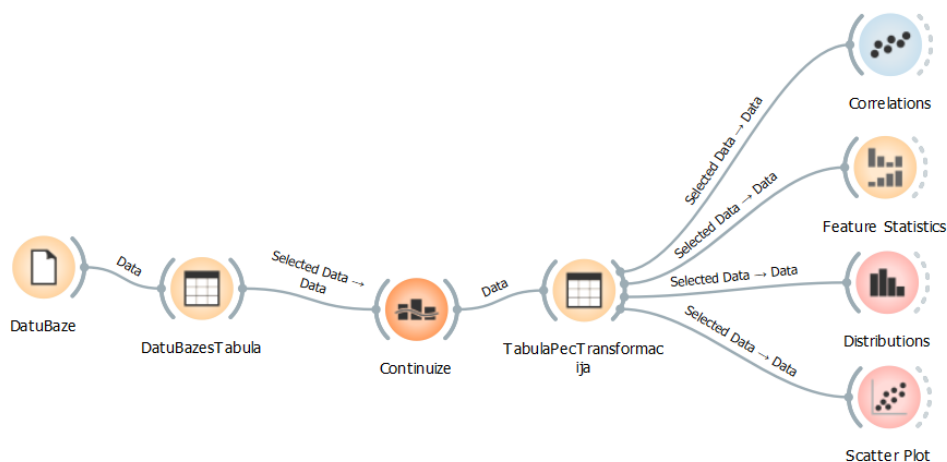
Name	Type	Role	Values
1 age	N numeric	feature	
2 gender	C categorical	target	F, M
3 height_cm	N numeric	feature	
4 weight_kg	N numeric	feature	
5 body fat_%	N numeric	feature	
6 diastolic	N numeric	feature	
7 systolic	N numeric	feature	
8 gripForce	N numeric	feature	
9 sit and bend	N numeric	feature	

Name	Type	Role	Values
1 age	N numeric	skip	
2 gender	C categorical	target	0, 1
3 height_cm	N numeric	feature	
4 weight_kg	N numeric	feature	
5 body fat_%	N numeric	skip	
6 diastolic	N numeric	skip	
7 systolic	N numeric	skip	

4. attēls Izmaiņās datu kopā(Kreisajā puse pirms, Labajā pēc)

5. Darbība

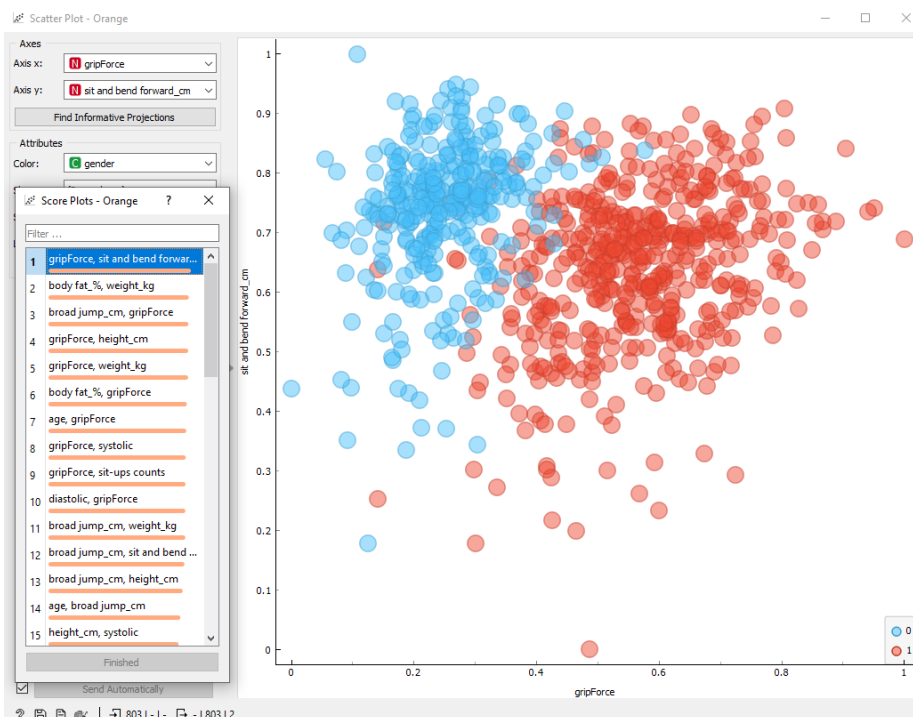
Lai atspoguļotu datu kopu vizuāli Orange rikā es izpildīju sekojošu shēmu(5. attēls), kur vispirms dati tiek paņemti no faila, tiek normalizēti skaitliskas vērtības un tika izpētīti.



5. attēls Datu kopu vizualizācijas shēma

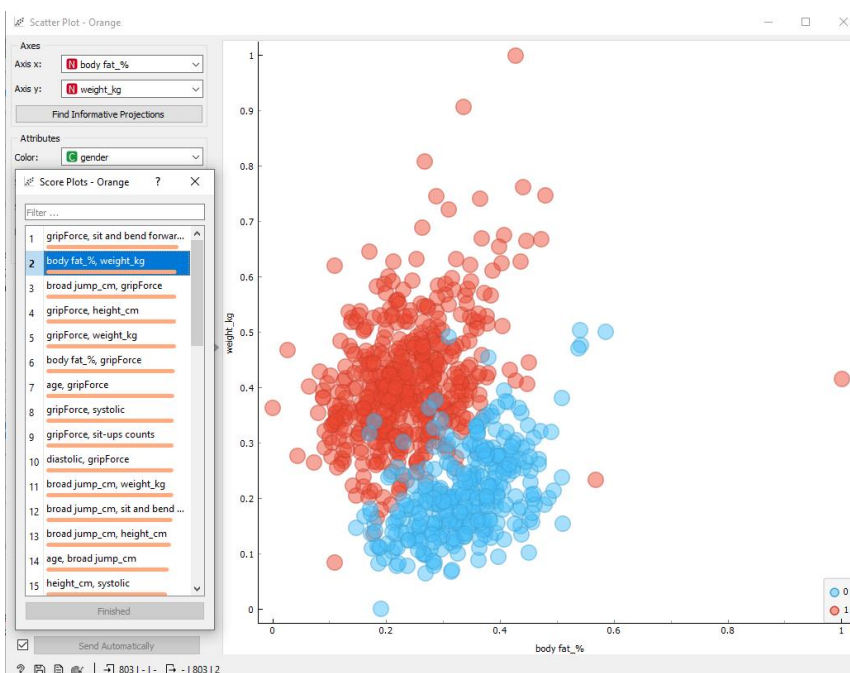
a) Izklīdes diagramma

Pirmajā izklīdes diagrammā(6. attēls) par iezīmes mainīgu es paņēmu Dzimuma atribūtu. Izmantojot Atrodiet informatīvas prognozes funkciju ir redzams kā vislabāk tiek atlasīti Satverina spēks un Attālums uz kuru cilvēks var noliekties uz priekšu Atribūti(gripForce un sit and bend forward_cm). Ir redzams kā vīriešiem ir labāka satveriena spēks un noliekšana ir labāka nekā sievietēm, tāpēc tie ir labi atlasīti jo ir likumsakarība.



6. attēls Pirmā izklīdes diagramma

Otrajā diagrammā(7.attēls) arī par iezīmes mainīgais tiek pieņemts Dzimums. Atrodiet informatīvas prognozes funkcija piedāvā paņemt body_fat un weight_kg atribūtus, tur arī ir likumsakarība un vīrieši starp sievietēm tiek labi atdalīti.

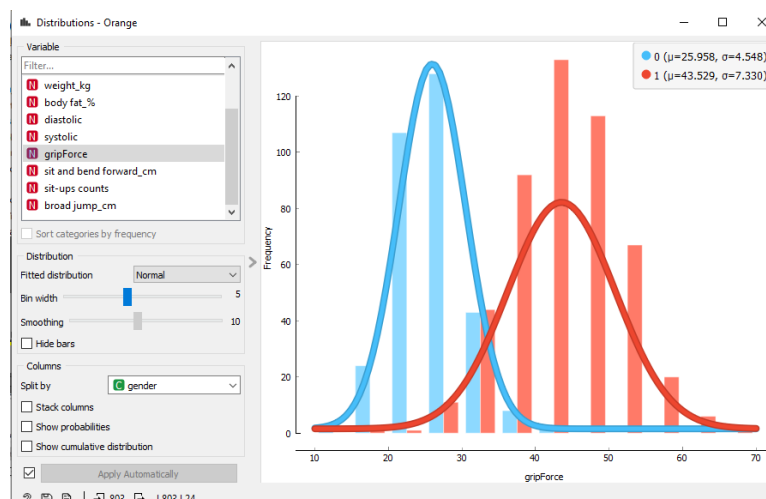


7. attēls Otra izklaides diagramma

b) Histogrammas

Pirmajā histogrammā (8. attēlā) tiek apskatīti dzimuma un satveriena spēka atribūtu saistība. Tabulā ir atspoguļoti kopēji un atsevišķie procenti starp vīriešiem un sievietēm un viņa saistība ar satveriena spēku:

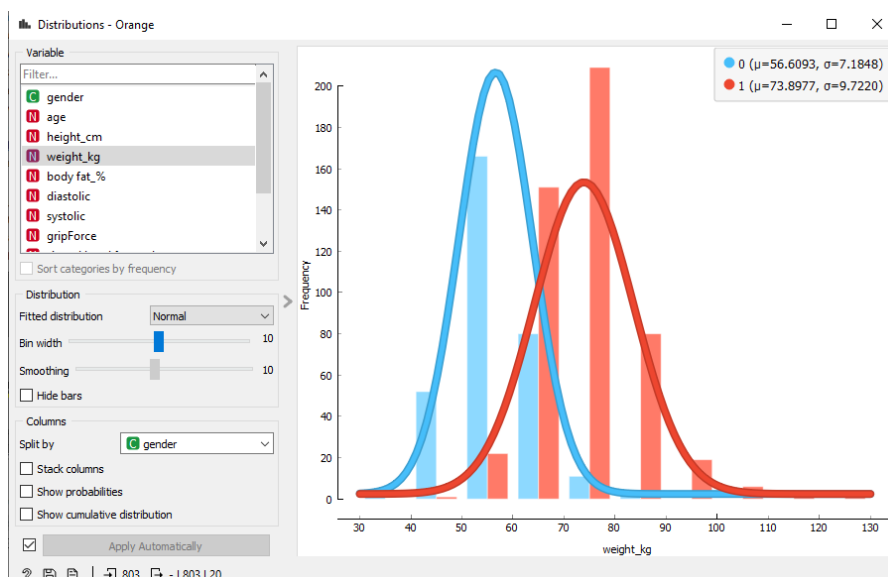
Diapazons	Sieviešu procents	Vīriešu procents	Kopējais svars
< 15 kg	100%	0%	0.12%
no 15 līdz 20 kg	92.3%	7.7%	3.24%
no 20 līdz 25 kg	99.1%	0.9%	13.45%
no 25 līdz 30 kg	92.1%	7.9%	17.3%
no 30 līdz 35 kg	49.4%	50.6%	10.8%
no 35 līdz 40 kg	8%	92%	12.45%
no 40 līdz 45 kg	1.5%	98.5%	16.81%
no 45 līdz 50 kg	0%	100%	14.07%
no 50 līdz 55 kg	0%	100%	8.34%
no 55 līdz 60 kg	0%	100%	2.49%
no 60 līdz 65 kg	0%	100%	0.75%
65 kg >	0%	100%	0.12%



8. attēls Satveriena spēka un dzimuma klašu atdalīšana histogrammā

Otrajā histogrammā (9. attēls) es tajā pašā veidā ka pirmajā salīdzināšu vīriešu un sievietes svāru klašu atdalīšanas ietvaros:

<i>Diapazons</i>	<i>Sieviešu procents</i>	<i>Vīriešu procents</i>	<i>Kopējais svars</i>
< 40 kg	100%	0%	0.12%
no 40 līdz 50 kg	98.1%	1.9%	6.6%
no 50 līdz 60 kg	88.3%	11.7%	23.41%
no 60 līdz 70 kg	34.6%	65.4%	28.77%
no 70 līdz 80 kg	5%	95%	27.4%
no 80 līdz 90 kg	3.6%	96.4%	10.34%
no 90 līdz 100 kg	0%	100%	2.38%
no 100 līdz 110 kg	0%	100%	0.75%
no 110 līdz 120 kg	0%	100%	0.12%
120kg >	0%	100%	0.12%

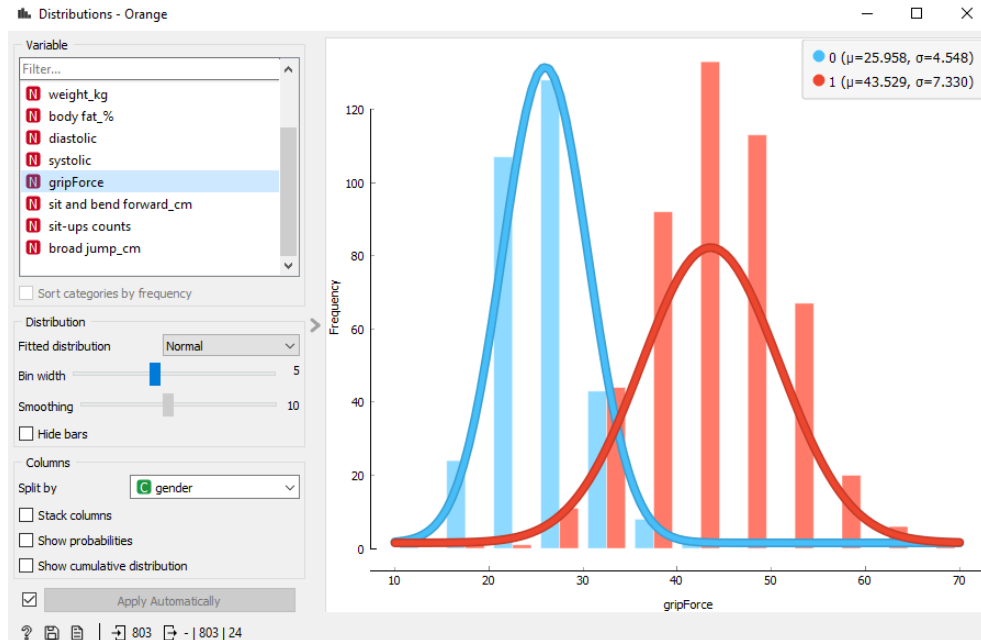


9. attēls Svāra un dzimuma klašu atdalījums histogrammā

c) Interesējošo pazīmju sadalījumu

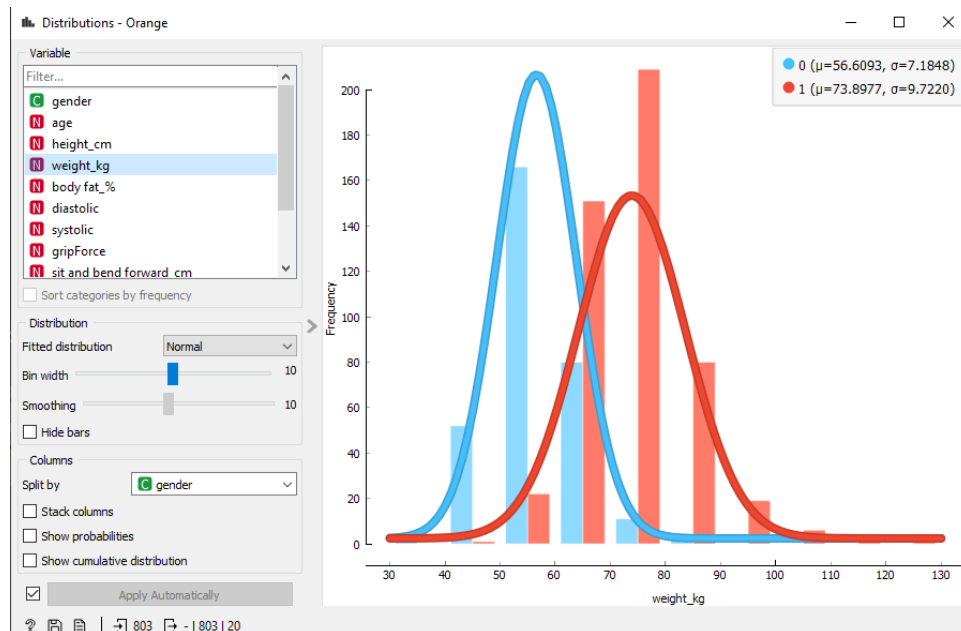
Balstoties uz iepriekšējo diagramma datiem man interese 2 atribūti sadalījumiem: satveriena spēks un svars jo tiem ir vislielākā sadalījuma attālums starp diviem.

10. attēlā ir redzams, ka sievietēm (0) ir mazāka satverina spēks, aptuveni 25, savukārt vīriešiem(1) ir lielākā:



10.attēls Satveriena un dzimuma sadalījums

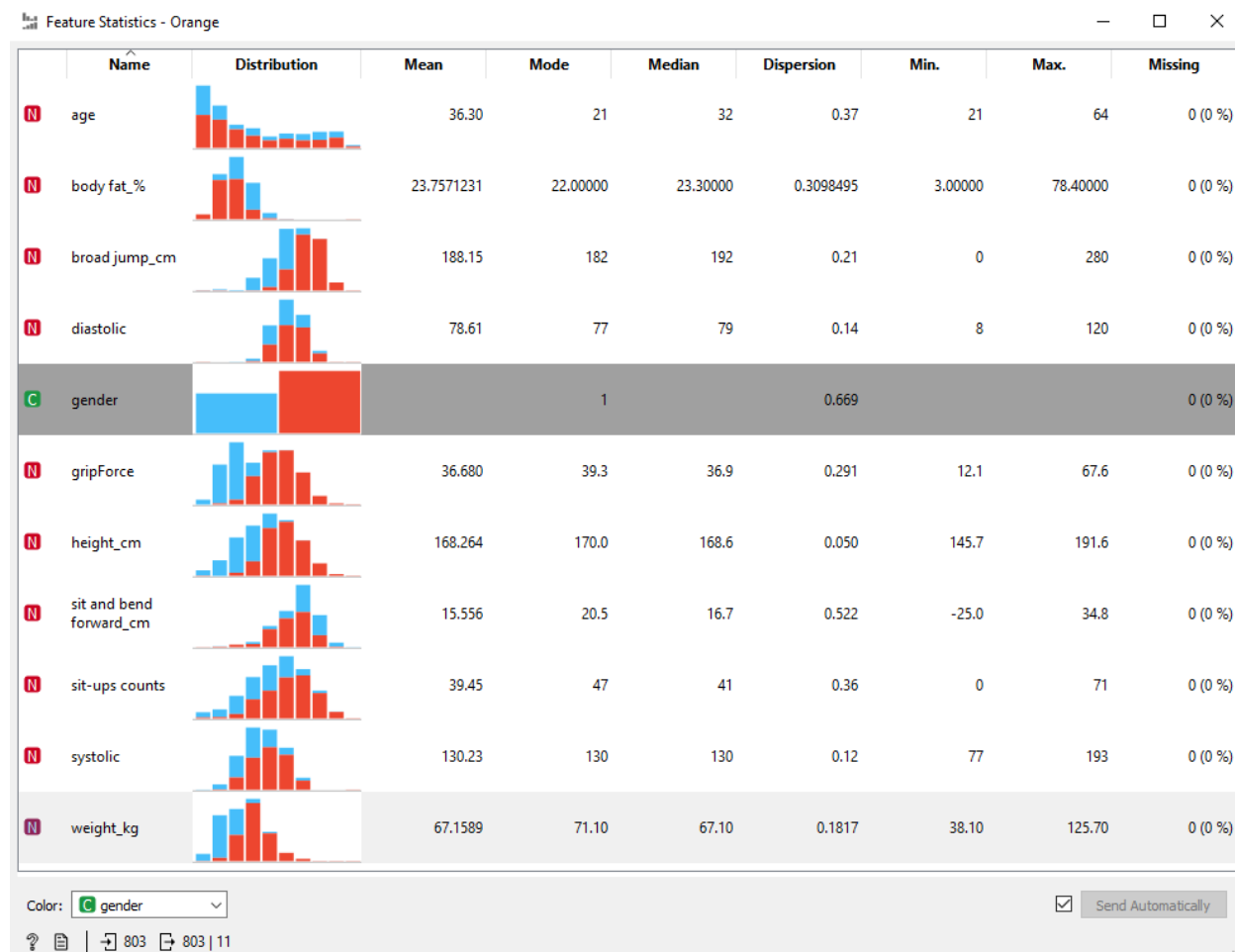
11. attēlā ir redzams, ka sievietēm(0) svārs ir mazāks par vīriešiem par ko arī liecina sadalījuma attālums un amplitūda:



11. attēls Svārs un dzimuma sadalījums

d) Aprēķināti statistiskie rādītāji

Ar statistika funkciju var viegli aprēķināt dažādus statistisku rādītājus un viņa atlasīt. Es atlasīju pa dažiem rādītājiem: Moda, Mediāna, Dispersija. Galveno parametru es atstāju dzimumu un atlasīju datus pēc vārda. Piemēram visizplatītākais vecums šajā datu kopā ir 21 gads; Vidēja vērtība vecumam ir 32; Dispersija vecumam ir 0.37. Visi citi dati ir redzami 12.attēlā.



12. attēls Statistiskie rādītāji.

Secinājumi par I daļas izpildi.

Izpildot I daļu un izpētot Ķermeņa veikspējas datu kopu es varu secināt, datu kopa līdzsvarotas. Ir atribūti, kur dominē vīrieši –satvērienā spēks, svars, bet ir atribūti kur dominē sievietes – attālums uz kuru cilvēks var noliekties uz priekšu sēžot, tauku ķermeņa. Vizuālie datu atspoguļojums ļauj redzēt datu struktūru. Piemēram izklaidēs diagrammā 6. attēlā var viegli redzēt ka skaidri atdalāmi 2 atribūti.

Vecums, piemēram, neietekmē datu kopai, jo tur ir savākti daži cilvēki ar dažu vecumu un ar to salīdzināt nav vērts.

Tālākai izpētei man interese šie atribūta: svars, tauki ķermeņa, satvērienā spēks, noliekšana uz priekšu sēžot.

II daļas apraksts – Nepārraudzītā mašīnmācīšanās

Šajā darba daļā studenti veiks iepriekš izvēlētās datu kopas klasterizāciju. Darba I daļa sniedza studentiem izpratni par to, kādas pazīmes (atribūti) un klases ir datu kopā un cik labi datu objekti sadalās klasēs. Šīs darba daļas mērķis ir, izmantojot klasterizācijas metodes, vēl vairāk izpētīt datu kopu, lai noskaidrotu, vai iepriekš izdarītie secinājumi par datu kopas struktūru ir spēkā.

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.

2. Hierarhiskās klasterizācijas algoritmam ir jāveic vismaz 3 eksperimenti, brīvi pārvietojot atdalošo līniju un analizējot,

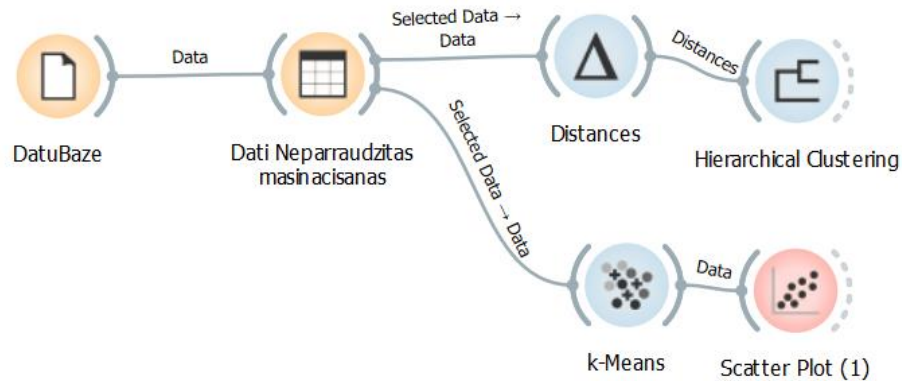
kā mainās klasteru skaits un saturs;

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

II daļas izpilde - Nepārraudzītā mašīnmācīšanās

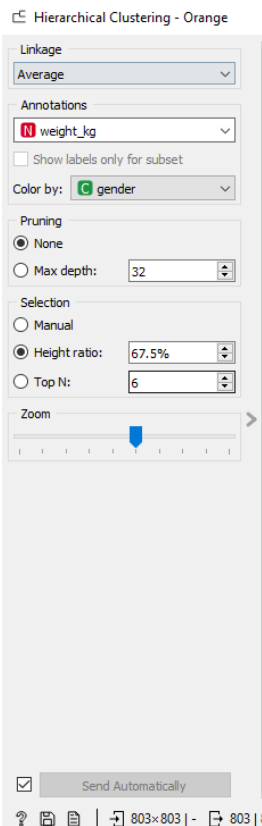
1. Darbība

Lai izpildītu nepārraudzītās mašīnmācīšanās algoritmus (hierarhiska klasterizācija un K-vidējo algoritmu) es izveidoju Orange rīkā shēmu katram algoritmam nolūkam. 13. attēlā augša ir hierarhiska klasterizācijas algoritms, apakša ir k-vidējais algoritms.



13. attēls Algoritma shēma Orange rīkā

Hierarhiska klasterizācija hiperparametri 14.attēlā



- Saiknes:

Viena saikne aprēķina attālumu starp abu kopu tuvākajiem elementiem

Vidējā saikne aprēķina vidējo attālumu starp abu kopu elementiem

Svērtā saikne izmanto WPGMA metodi

Pilnīga saikne aprēķina attālumu starp klasteru visattālākajiem elementiem

Ward linkage aprēķina kvadrātu kļūdas summas pieaugumu.

14.attēls Hierarhiska klasterizācija hiperparametri

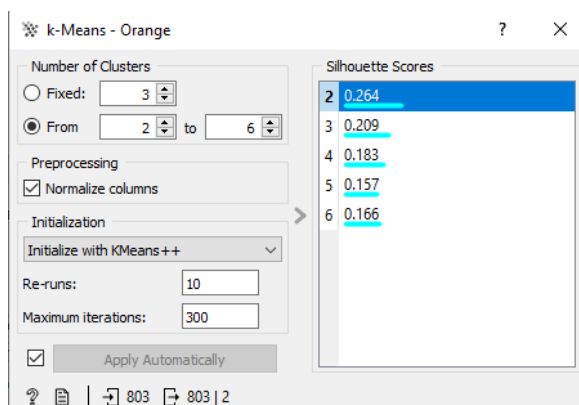
- Dendrogrammas mezglu etiķetes var izvēlēties Anotācijas lodziņā.
- Atzarošanas lodziņā var apgriezt milzīgas dendrogrammas, izvēloties maksimālo dendrogrammas dziļumu. Tas ietekmē tikai displeju, nevis faktisko kopu veidošanu.
- Logrīks piedāvā trīs dažādas atlasēšanas metodes:

Rokasgrāmata (noklikšķinot dendrogrammas iekšpusē, tiks atlasīts klasteris. Vairākas kopas var izvēlēties, turot Ctrl / Cmd. Katrs atlasītais klasteris tiek parādīts citā krāsā un tiek uzskatīts par atsevišķu kopu izvadē.)

Augstuma attiecība (noklikšķinot uz dendrogrammas apakšējā vai augšējā lineāla, grafikā tiek ievietota griezumuma līnija. Tiek atlasīti vienumi pa labi no līnijas.)

Top N (izvēlas augšējo mezglu skaitu.)

K-vidējo hiperparametri 15. attēls



15. attēls K-vidējo hiperparametri

- Izvēlieties kopu skaitu.

Novērsts: algoritms klasteru datus uz noteiktu skaitu klasteru.

No X līdz Y: logrīks parāda klasterizācijas rādītājus izvēlētajam klasteru diapazonam, izmantojot Silhouette rezultātu (kontrastē vidējo attālumu līdz elementiem tajā pašā klasterī ar vidējo attālumu līdz elementiem citās kopās).

- *Priekšapstrāde*: ja opcija ir atlasīta, kolonnas tiek normalizētas (Vidējais centrēts uz 0 un standartnovirze mērogots līdz 1).
- Inicializācijas metode (veids, kā algoritms sāk klasterizāciju):

k-Nozīmē++ (pirmais Centrs tiek izvēlēts nejauši, pēc tam tiek izvēlēti no atlikušajiem punktiem ar varbūtību proporcionāli kvadrātā attālumam no tuvākā centra)

Nejauša inicializācija (kopas vispirms tiek piešķirtas nejauši un pēc tam atjauninātas ar turpmākām iterācijām)

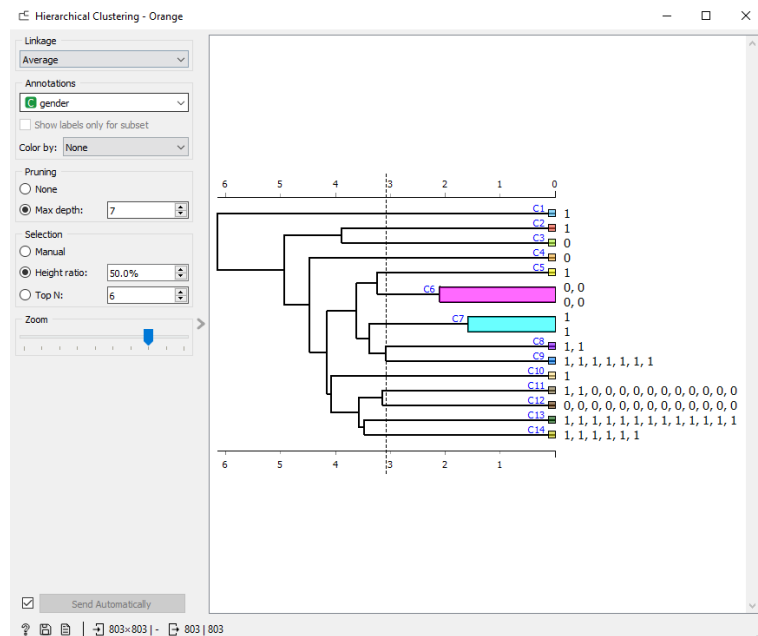
- *Atkārtoti palaiž* (cik reizes algoritms tiek palaists no nejaušām sākotnējām pozīcijām; tiks izmantots rezultāts ar zemāko kvadrātu kopu summu) un *maksimālās iterācijas* (maksimālo atkārtojumu skaitu katrā algoritma izpildījumā) var iestatīt manuāli.

2. Darbība

Lai paveikt eksperimentu, man jāpārvieto atdalošo līniju un paskatīties ka mainā klasteru skaits un saturs.

1. eksperiments

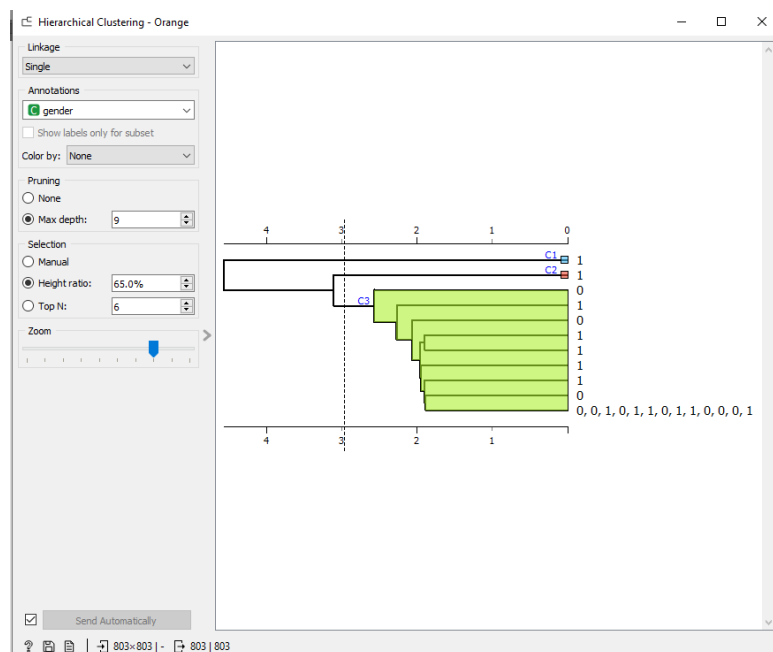
1. eksperimentā es uztaisīju (16. attēlā) Average linkage, max depth 7, height ratio 50%



16. attēls 1. eksperiments

2. eksperiments

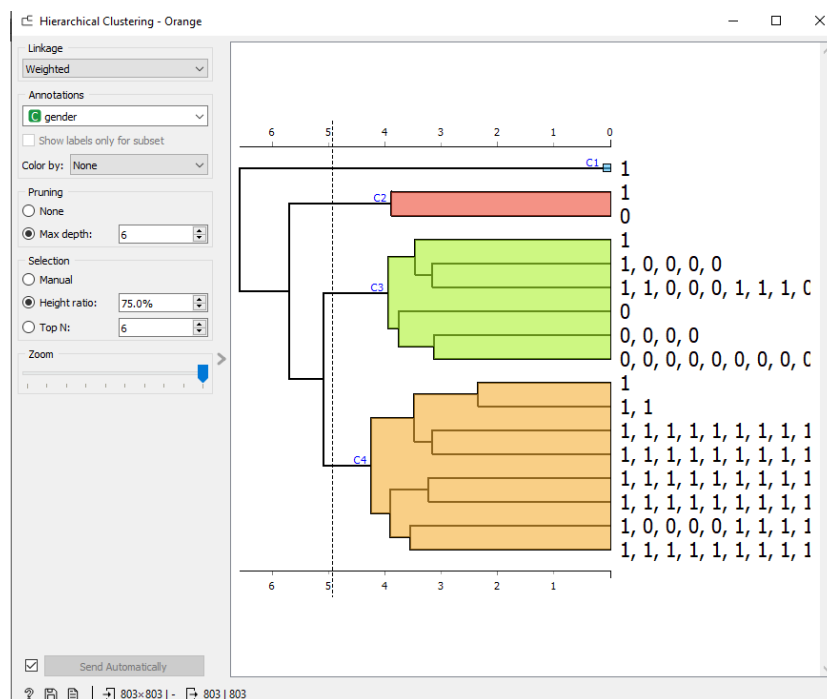
2. eksperimentā es uztaisīju (17. attēlā) Single linkage, max depth 9, height ratio 65%



17. attēls 2. eksperiments

3. eksperiments

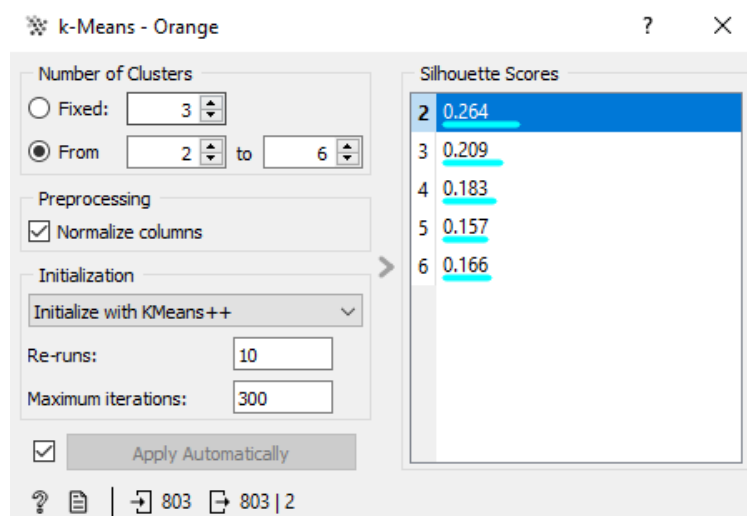
3. eksperimentā es uztaisīju (18. attēls) Weighted linkage, max depth 6, height ratio 75 %



18. attēls 3. eksperiments

3. Darbība

K-vidēja funkcija (19. attēls) ļauj ļoti izvēlēties klasteru diapazonu. Nav jēgas taisīt lielāko skaitu, jo būs liels attālums un būs mazāks Silhouette scores, es izvēlējos no 2 līdz 6 klasteriem. Lai dati būtu vienādi jāizveido normalize check box. Es inicializēju to ar KMeans++ un uztaisīju 100 maksimuma iterācijās.



19. attēls K-vidējais algoritms

Kā redzams Silhoutte Score ir lielāks jo ir mazāka klasteru skaits. 20. attēlā ir salīdzinājums izkļiedes diagrammās, kur ir redzams klasteru skaits ietekmi uz datu vērtībām.



20. attēls Izkļiedes diagramma klasteriem

Secinājumi par II daļas izpildi

Balstoties uz algoritmu darbības izpildi es varu izveidot secinājumu, ka datu kopā klases ir nav labi atdalāmas. Tās nav visiem atribūtiem, bet tiem kuru es minēju iepriekšēja daļa atdalās labāk un ar viņiem var veikt neparaudzītu mašīnmācīšanu. Šie 2 algoritmi savukārt nepalīdzēs pareizi noteikt jaunu klasi, bet lai izpētītu tālāk viņi var palīdzēt.

III daļas apraksts - Pārraudzītā mašīnmācīšanās

Šajā darba daļā studentiem ir jāpielieto vismaz 3 klasifikācijas algoritmi iepriekš izvēlētajai datu kopai. Viens no algoritmiem, kura izmantošana ir obligāta, ir mākslīgie neironu tīkli. Divus citus algoritmus studenti var brīvi izvēlēties.

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.
2. Ir jāsadala datu kopa apmācību un testa datu kopās.
3. Katram algoritmam, lietojot apmācību datu kopu, ir jāveic vismaz 3 eksperimenti, mainot algoritma hiperparametru vērtības un analizējot algoritmu veikspējas metrikas;
4. Katram algoritmam ir jāizvēlas tas apmācītais modelis, kas nodrošina labāko algoritma veikspēju;
5. Katra algoritma apmācītais modelis ir jāpielieto testa datu kopai.
6. Ir jānovērtē un jāsalīdzina apmācīto modeļu veikspēja.

Darba atskaitē ir jāiekļauj šāda informācija par šo darba daļu:

- Īss apraksts (1/3 no A4 lapas) diviem brīvi izvēlētajiem algoritmiem un to izvēles motivācijai (izņemot mākslīgo neironu tīklu), norādot arī atsaucis uz izmantotajiem informācijas avotiem.
- Katram algoritmam ir jāapraksta Orange rīkā pieejamie hiperparametri un to nozīme.
- Informācija par testu un apmācību datu kopām:
 - o kopējais testa un apmācību datu kopām pievienoto datu objektu skaits (skaits un %);
 - o informācija par to, cik datu objektu no katras klases ir iekļauts apmācību un testa datu kopās (skaits un %);
- Katram algoritmam tabulas veidā ir jāatspoguļo eksperimentos izmantotās hiperparametru vērtības;
- Secinājumi par modeļu veikspēju veiktajos eksperimentos, skaidri identificējot modelī, kas tiks izmantots testēšanā;
- Apmācīto modeļu testēšanas rezultāti un to veikspējas salīdzinājums un interpretācija, tos skaidri nodalot no apmācības eksperimentiem.

III daļas izpilde - Pārraudzītā mašīnmācīšanās

1. Darbība

Lai izdarītu pārraudzīto mašīnmācīšanu es izvelējos kNN, Logistic Regression un obligāto Neural Network algoritmu.

Apakšā ir īss apraksts par kNN un loģistikas regresijas algoritmu.

kNN algoritms:

KNN (K-tuvāko kaimiņu) algoritms ir metrisks klasifikācijas algoritms, kas izmanto objektu tuvumu, lai izlemtu klasificēt vai paredzēt atsevišķa datu punkta grupēšanu. Tas attiecas uz neparametriskiem mašīnmācīšanās algoritmiem un ir viens no vienkāršākajiem klasifikācijas algoritmiem.

Tas ir balstīts uz pieņēmumu, ka vienas klases objekti atrodas tuvu viens otram pazīmju telpā. KNN algoritms darbojas šādi: katram testa datu kopas objektam ir k tuvākie mācību datu kopas objekti. Pēc tam objektam tiek piešķirta klase, kas ir visizplatītākā starp k tuvākajiem kaimiņiem.

Hipeparametri:

- Nosaukums, ar kuru tas parādīsies citos logrīkos. Noklusējuma nosaukums ir "kNN".
- Iestatiet tuvāko kaimiņu skaitu, attāluma parametru (metriku) un svarus kā modeļa kritērijus.
- Metrika var būt:

Eiklida ("taisna līnija", attālums starp diviem punktiem)

Manhetena (visu atribūtu absolūto atšķirību summa)

Maksimālais (lielākais no absolūtajām atšķirībām starp atribūtiem)

Mahalanobis (attālums starp punktu un sadalījumu).

- Svari, kurus varat izmantot, ir:

Vienveidīgs: visi punkti katrā apkārtnē tiek svērti vienādi.

Attālums: tuvākiem vaicājuma punkta kaimiņiem ir lielāka ietekme nekā kaimiņiem tālāk.

- Sagatavot ziņojumu.

Logistic Regression algoritms:

Loģistika regresija algoritms ir metrisks klasifikācijas algoritms, kas izmanto loģistikas funkciju, lai izlemtu klasificēt vai paredzēt atsevišķa datu punkta grupēšanu. Tas pieder pie parametru mašīnmācīšanās algoritmiem un ir viens no populārākajiem klasifikācijas algoritmiem.

Loģistika regresija tiek izmantota binārai klasifikācijai, tas ir, kad dati jāsadala divās klasēs. Tas darbojas šādi: katram testa datu kopas objektam tiek aprēķināta tā piederības varbūtība vienai no divām klasēm. Pēc tam objektam tiek piešķirta klase, kuras varbūtība ir lielāka.

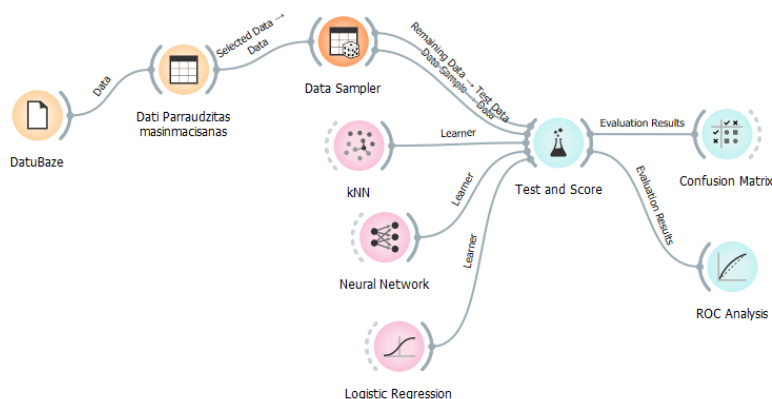
Hiperparametri:

- Nosaukums, ar kuru skolēns parādās citos logrīkos. Noklusējuma nosaukums ir "Logistikas regresija".
- *Regularizācijas* veids (vai nu L1, vai L2). Iestatiet izmaksu stiprumu (noklusējums ir $C=1$).
- Nospiediet Lietot, lai veiktu izmaiņas. Ja tiek atzīmēts pieteikums Automātiski, izmaiņas tiks paziņotas automātiski.

Izvēles motivācija:

Logistika regresija un kNN algoritma izvēles iemesls klasifikācijas problēmai ir tas, ka tos var izmantot, lai atrisinātu plašu klasifikācijas un regresijas problēmu loku. Ir pieejama dokumentācija Orange vietnē un viņa darbības var viegli saprast.

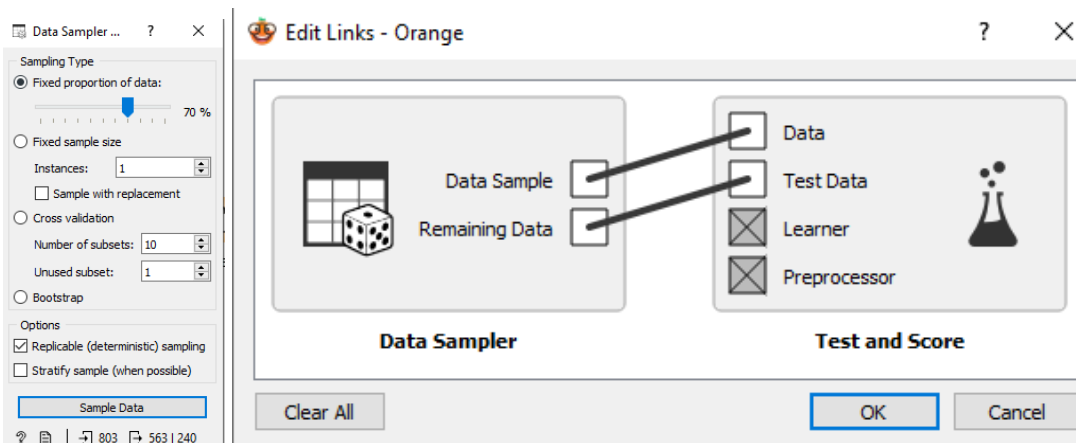
21.attēlā ir izveidota Orange shēma kNN, Logistic Regression un Neural Network algoritma darbībai



21. attēls Pārraudzītā mašīnmācīšanas Orange shēma

2., 3., 4., un 5. Darbība

Ar Data Sampler funkciju es sadalīju datu kopu apmācību un testa datu kopu 70% uz 20% (22. attēls).



22. attēls Testa un apmācības datu kopa

Talāk es izpildīju 3 testus ar visiem 3 algoritmiem mainot hiperparametru vērtības:

1. Tests

23.attēls kNN1

24. attēls NN1

25.attēls LR1

Model	AUC	CA	F1	Precision	Recall
kNN	0.999	0.988	0.988	0.988	0.988
Neural Network	0.725	0.625	0.481	0.391	0.625
Logistic Regression	0.993	0.983	0.983	0.984	0.983

26. attēls 1. testa rezultāti

		Predicted		Σ
		0	1	
Actual	0	89	1	90
	1	2	148	150
Σ		91	149	240

27.attēls kNN1 Rez

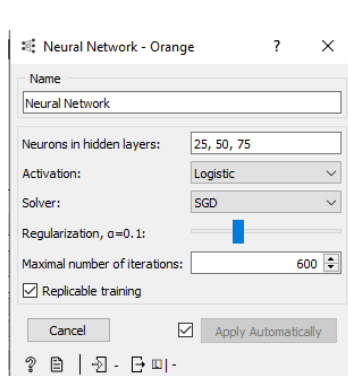
		Predicted		Σ
		0	1	
Actual	0	89	1	90
	1	3	147	150
Σ		92	148	240

28. attēls LR1 Rez

		Predicted		Σ
		0	1	
Actual	0	0	90	90
	1	0	150	150
Σ		0	240	240

29. attēls NN1 Rez

2. Tests



Neural Network - Orange

Name: Neural Network

Neurons in hidden layers: 25, 50, 75

Activation: Logistic

Solver: SGD

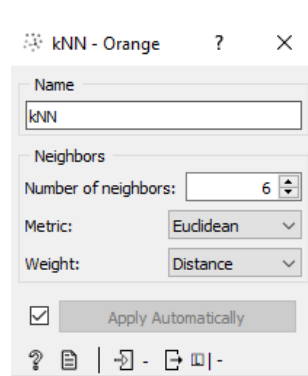
Regularization, $\alpha=0.1$: [Slider]

Maximal number of iterations: 600

☒ Replicable training

Buttons: Cancel, Apply Automatically

30. attēls NN2



kNN - Orange

Name: kNN

Neighbors

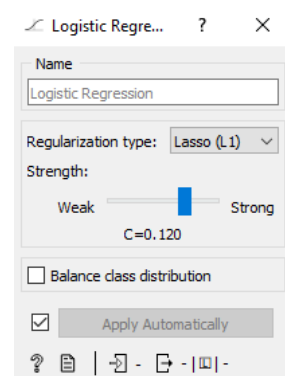
Number of neighbors: 6

Metric: Euclidean

Weight: Distance

☒ Apply Automatically

31. attēls kNN2



Logistic Regression - Orange

Name: Logistic Regression

Regularization type: Lasso (L1)

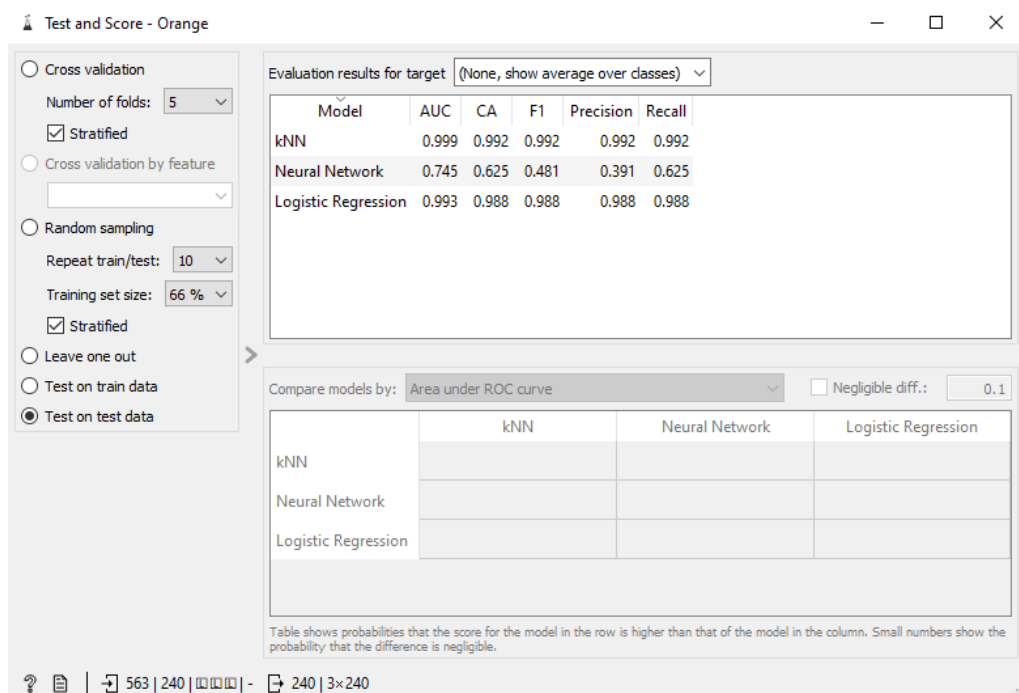
Strength: [Slider] Weak Strong

C=0.120

☐ Balance class distribution

☒ Apply Automatically

32. attēls LR1



Test and Score - Orange

☐ Cross validation

Number of folds: 5

☒ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 66 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☒ Test on test data

Evaluation results for target: (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.999	0.992	0.992	0.992	0.992
Neural Network	0.745	0.625	0.481	0.391	0.625
Logistic Regression	0.993	0.988	0.988	0.988	0.988

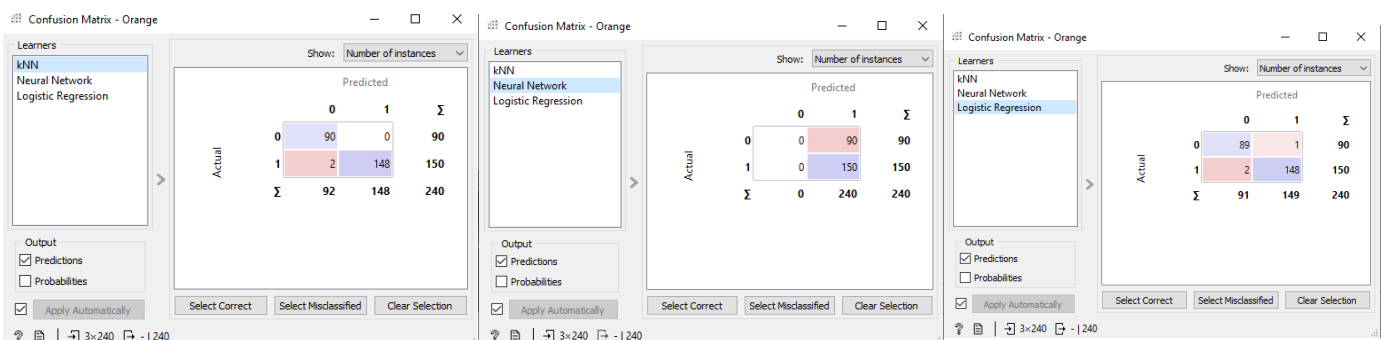
Compare models by: Area under ROC curve

☐ Negligible diff.: 0.1

	kNN	Neural Network	Logistic Regression
kNN			
Neural Network			
Logistic Regression			

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.

33. attēls 2. testa rezultāti



Confusion Matrix - Orange

Learners: kNN, Neural Network, Logistic Regression

Show: Number of instances

	0	1	Σ
0	90	0	90
1	2	148	150
Σ	92	148	240

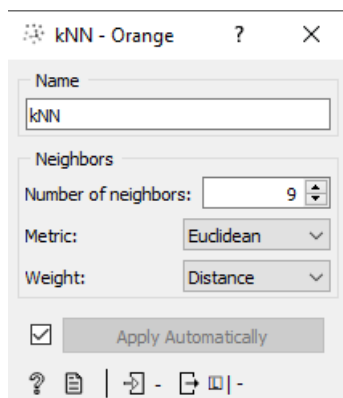
Buttons: Select Correct, Select Misclassified, Clear Selection

34. attēls kNN2 Rez

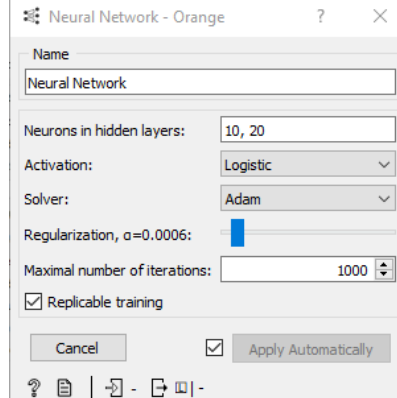
35. attēls NN2 Rez

36. attēls LR2 Rez

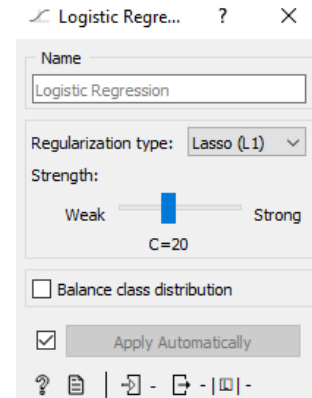
3. tests



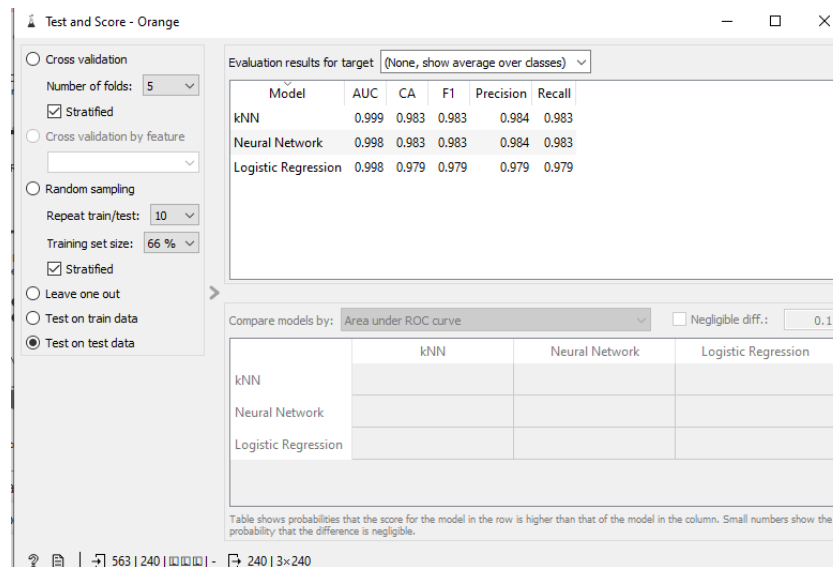
37. attēls kNN3



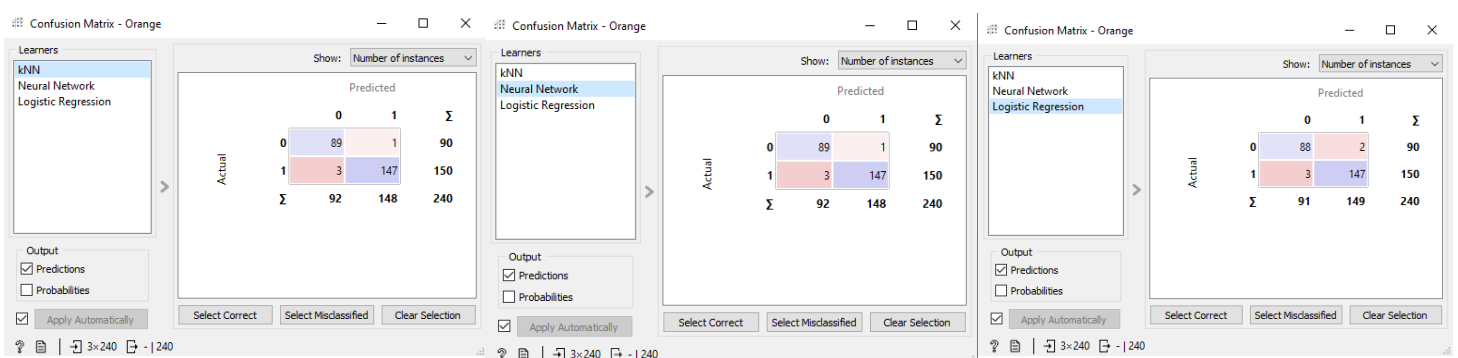
38. attēls NN2



39. attēls LR3



40. attēls 3. testa rezultāti



41. attēls kNN3 Rez

42. attēls NN3 Rez

43. attēls LR3 Rez

6. Darbība

Izpildot 3 eksperimentus sanāca šadi rezultāti:

kNN un Logistiskās regresijas algoritms vislabāk nostrādāja ar 2 testā iesaistījumiem, bet Neirona Tiklu algoritms ar 3 testa hiperparametriem.

kNN vislabāk strādā ar neparāk lielu un neparaki mazu kaimiņa skaitu (mazāk par 9, bet lielāk par 6). Ir redzams, ka jā ietvert parāk lielu kaimiņa skaitu, samazinās kopēja precizitāte.

Logistiskā regresija algoritms ar Lasso regularizāciju stiprību ietver sevī sekojošo sakarību: jo lielāka stiprība, jo lielāka precizitāte.

Neirona tīkla algoritms vislabāk nostrādāja ar stohastisko gradienta optimizētāju, viņš parādīja 0.983 precizitāti. Ar stohastisku gradienta nolaišanu man nesanāca palielināt precizitāti par 0.5.

Kopumā, atsevišķi skatoties uz 3 eksperimentu man sanāca iegūt ļoti labi precizitāti ar visiem algoritmiem, tātad apmācīt modeli.

Secinājumi

Darbs ar Orange rīku bija ļoti interesants un ērts, veicot dažādus uzdevumus, es iemācījos analizēt, izvēlēties un izmantot nekontrolētas un kontrolētas mašīnmācīšanās algoritmus. Strādājot ar savu datu kopu, es izvēlējos vispiemērotākos atribūtus un izmantoju tikai tos, kas man palīdzēja tos palaist algoritmos. Manuprāt, visgrūtākais uzdevums bija izvēlēties piemērotu datu kopu, jo dažiem no tiem ir slikta klasifikācija, un vēlāk mašīnmācīšanās var radīt problēmas. Orange rīkā ir ērti skatīties un novērst šādas problēmas.

Darba rezultātu var izmantot, lai turpinātu mācīties un attīstītu mašīnmācīšanos.

Izmantoti avoti

- 1) <https://orangedatamining.com/widget-catalog/unsupervised/kmeans/>
- 2) <https://orangedatamining.com/widget-catalog/unsupervised/hierarchicalclustering/>
- 3) <https://orangedatamining.com/widget-catalog/model/logisticregression/>
- 4) <https://orangedatamining.com/widget-catalog/model/neuralnetwork/>
- 5) <https://orangedatamining.com/widget-catalog/model/knn/>
- 6) <https://www.kaggle.com/datasets/kukuroo3/body-performance-data>