RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Informācijas tehnoloģijas institūts

2.praktiskais darbs
mācību priekšmetā
“Mākslīgā intelekta pamati”

Mašīnmācīšanās algoritmu lietojums

Izstrādāja: Sabīne Ošiņa
3.grupa, 201RDB121

Pārbaudīja: Alla Anohina - Naumeca

2021./2022. mācību gads

ANOTĀCIJA

Šis darbs tiek izveidots ar *(Orange: Data Mining Toolbox in Python)* ***Orange3-3.32.0-Miniconda-x86_64.exe*** (Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B., 2013) aplikāciju. Darbā tiek aprakstīts pasniedzēja izveidotais darba uzdevums, autores izvēlētā datu kopa, datu pirmapstrāde, nepārraudzītā mašīnmācīšanās, kā arī pārraudzītā mašīnmācīšanās. Darbā tiek aprakstīti arī datu iegūtie secinājumi un izmantotā literatūra.

Šis darbs tiek veidots, lai autorei ir padziļinātākas zināšanas mākslīgajā intelektā, apgūstot aplikācijas *Orange* iespējas. Šajā darbā var aplūkot vairāku veidu algoritmu izpildes. Visiem algoritmiem ir to paskaidrojumi un interpretācijas. Daži algoritmi ir arī salīdzināti viens ar otru, lai spētu noteikt labāko izpildes rezultātu un izveidot to analīzi.

Šajā darbā tiek pielietotas zināšanas no pasniedzējas parādītā video mācību kursā, kā arī papildus zināšanas un algoritmu apraksti tiek iegūti no Orange oficiālās mājaslapas.

Darbam ir arī pieejama pilna dokumentācija *GitHub* mājaslapā ar visiem failiem un dokumentiem, kas tika pielietoti šī darba izpildē.

SATURS

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 jāizmanto *Orange* rīks. Tā lietotāja pamācība ir pieejama e-studiju kursa sekcijā "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. 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.

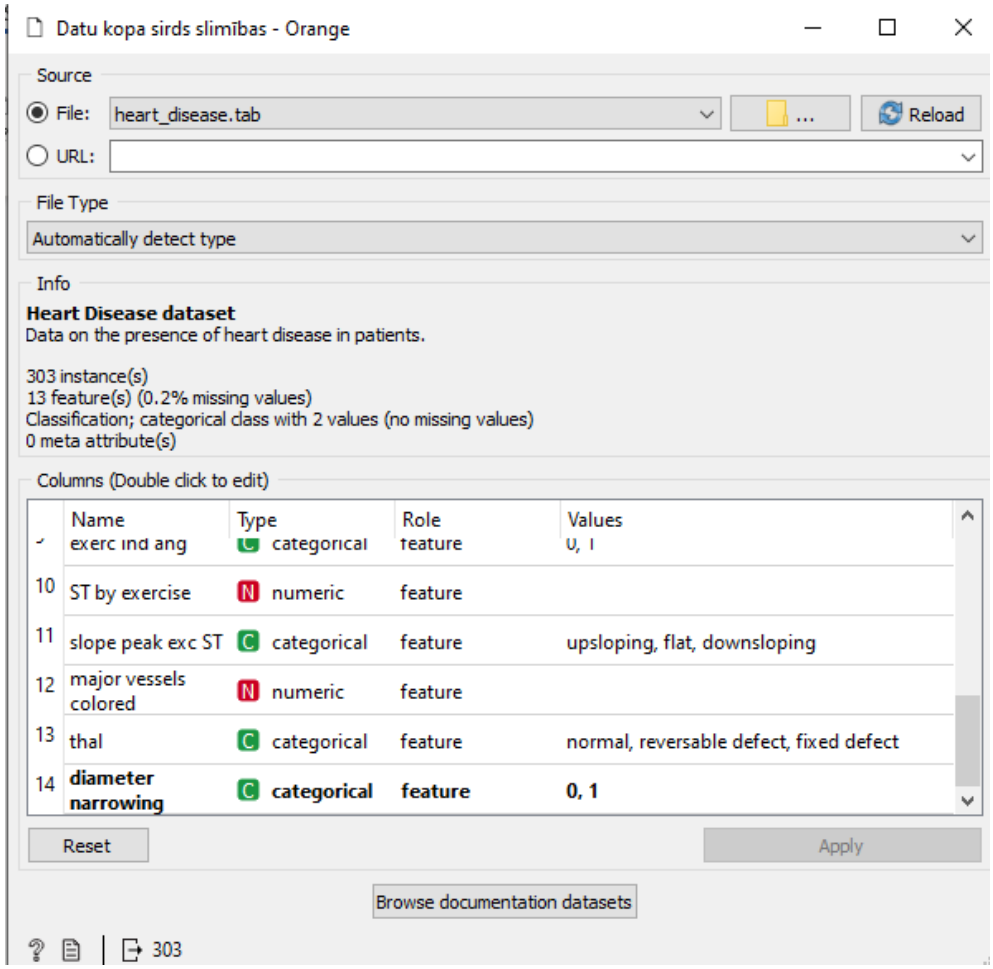
DATU KOPAS IZVĒLE

Lejupielādējot *Orange* aplikāciju, autore ieraudzīja to, ka ir pieejamas datu kopas, kuras jau ir gatavas, kā piemēri. Autore šobrīd pielieto *Orange* aplikācijas "*heart_disease.tab*" piemēru. Datu kopa nav dota kā *.csv* datu fails, bet tas neaizliedz to pilnvērtīgi pielietot. Autore saprot, ka datu kopas izveidotājs ir aplikācija *Orange*. Meklējot internetā, autore neatrada paplašinātu informāciju par laiku, kad datu kopa ir izveidota. Licence *(License)* pievienota beigās (Orange University of Ljubljana, 2022).

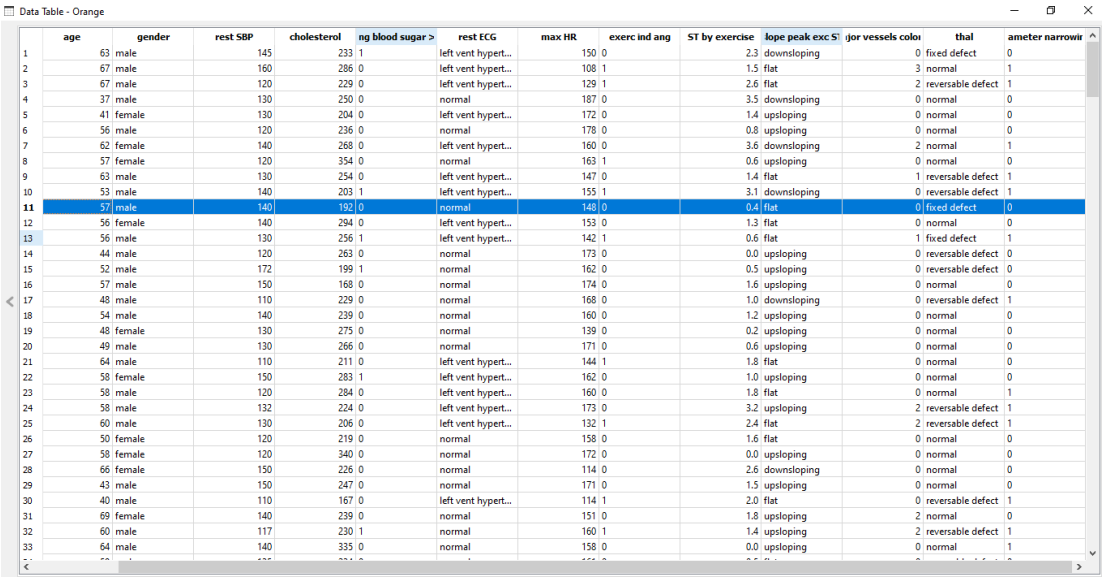
Datu kopai ir 303 datu objekti- cilvēki, kuriem ir problēmas ar sāpēm krūtīs. Datu kopa satur kategoriskās klasifikācijas klašu iezīmes, kur ir tikai 2 vērtības. Datu kopai ir 14 atribūti, iekļaujot mērķa atribūtu:

1) *age – numeric – feature,*
2) *gender – categorical – feature – female/male,*
3) *chest pain – categorical – **target**- asymptomatic, atypical ang, non-anginal, typical ang,*
4) *rest SBP – numeric – feature,*
5) *cholesterol – numeric – feature,*
6) *fasting blood sugar >120 – categorical – feature – 0/1,*
7) *rest ECG – categorical – feature – normal, left vent hypertrophy, ST-T abnormal,*
8) *max HR – numeric – feature.,*
9) *exerc ind ang – categorical – feature – 0/1,*
10) *ST by exercise – numeric – feature,*
11) *slope peak exc ST – categorical – feature – upsloping, flat, downsloping,*
12) *major vessels colored – numeric, feature,*
13) *Thal – categorical – feature – normal, reversable defect, fixed defect,*
14) *diameter narrowing – categorical – feature – 0/1,*

Autore no sākuma gribēja dažus datus izņemt jeb "*skip*", bet vēlāk saprata, ka tam īsti nebūs vajadzība, jo, ja ir vairāk datu, tad ir lielāka precizitāte. Datu kopā ir daži dati, kas iekļauj būla tipa 1/0 vērtības, bet to nav tik daudz, lai datu kopa būtu neizmantojama. Datu kopā nav iekļauti teksta korpusi un neapstrādāti attēli.



1.att. "*heart_disease.tab*" datu kopas attēlojums.

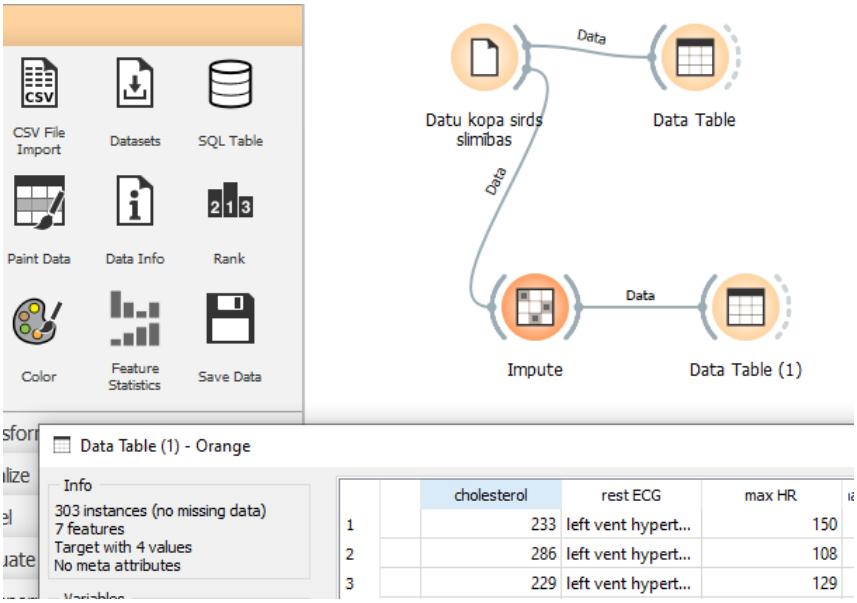


	age	gender	rest SBP	cholesterol	ng blood sugar >	rest ECG	max HR	exerc ind ang	ST by exercise	lope peak exc S	jor vessels colo	thal	ameter narrowi
1	63	male	145	233	1	left vent hypert...	150	0	2.3	downsloping	0	fixed defect	0
2	67	male	160	286	0	left vent hypert...	108	1	1.5	flat	3	normal	1
3	67	male	120	229	0	left vent hypert...	129	1	2.6	flat	2	reversable defect	1
4	37	male	130	250	0	normal	187	0	3.5	downsloping	0	normal	0
5	41	female	130	204	0	left vent hypert...	172	0	1.4	upsloping	0	normal	0
6	56	male	120	236	0	normal	178	0	0.8	upsloping	0	normal	0
7	62	female	140	268	0	left vent hypert...	160	0	3.6	downsloping	2	normal	1
8	57	female	120	354	0	normal	163	1	0.6	upsloping	0	normal	0
9	63	male	130	254	0	left vent hypert...	147	0	1.4	flat	1	reversable defect	1
10	53	male	140	203	1	left vent hypert...	155	1	3.1	downsloping	0	reversable defect	1
11	57	male	140	192	0	normal	148	0	0.4	flat	0	fixed defect	0
12	56	female	140	294	0	left vent hypert...	153	0	1.3	flat	0	normal	0
13	56	male	130	256	1	left vent hypert...	142	1	0.6	flat	1	fixed defect	1
14	44	male	120	263	0	normal	173	0	0.0	upsloping	0	reversable defect	0
15	52	male	172	199	1	normal	162	0	0.5	upsloping	0	reversable defect	0
16	57	male	150	168	0	normal	174	0	1.6	upsloping	0	normal	0
17	48	male	110	229	0	normal	168	0	1.0	downsloping	0	reversable defect	1
18	54	male	140	239	0	normal	160	0	1.2	upsloping	0	normal	0
19	48	female	130	275	0	normal	139	0	0.2	upsloping	0	normal	0
20	49	male	130	266	0	normal	171	0	0.6	upsloping	0	normal	0
21	64	male	110	211	0	left vent hypert...	144	1	1.8	flat	0	normal	0
22	58	female	150	283	1	left vent hypert...	162	0	1.0	upsloping	0	normal	0
23	58	male	120	284	0	left vent hypert...	160	0	1.8	flat	0	normal	1
24	58	male	132	224	0	left vent hypert...	173	0	3.2	upsloping	2	reversable defect	1
25	60	male	130	206	0	left vent hypert...	132	1	2.4	flat	2	reversable defect	1
26	50	female	120	219	0	normal	158	0	1.6	flat	0	normal	0
27	58	female	120	340	0	normal	172	0	0.0	upsloping	0	normal	0
28	66	female	150	226	0	normal	114	0	2.6	downsloping	0	normal	0
29	43	male	150	247	0	normal	171	0	1.5	upsloping	0	normal	0
30	40	male	110	167	0	left vent hypert...	114	1	2.0	flat	0	reversable defect	1
31	69	female	140	239	0	normal	151	0	1.8	upsloping	2	normal	0
32	60	male	117	230	1	normal	160	1	1.4	upsloping	2	reversable defect	1
33	64	male	140	335	0	normal	158	0	0.0	upsloping	0	normal	1

2.att. "*heart_disease.tab*" datu failu kolonnas.

I DAĻA – DATU PIRMAPSTRĀDE/IZPĒTE

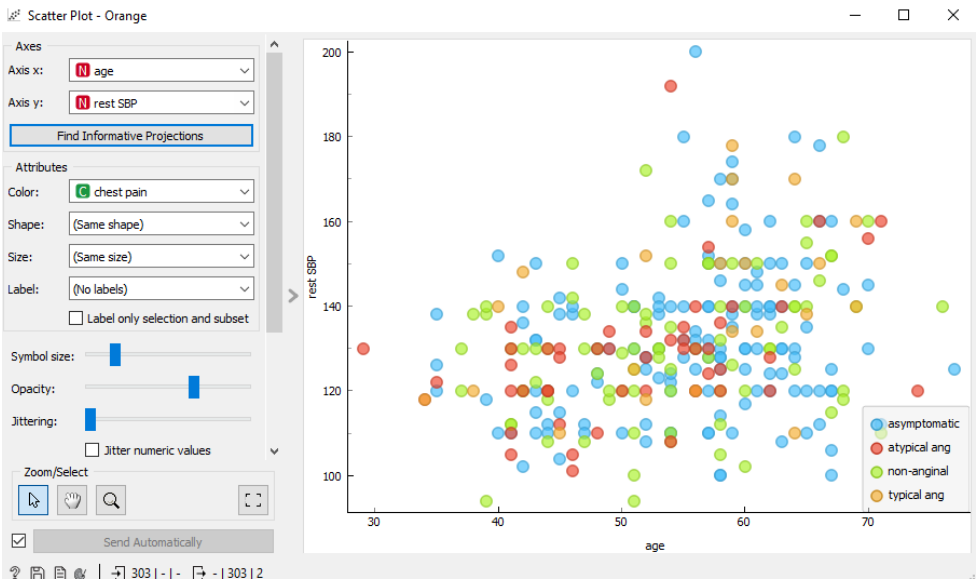
1) Tiek izvēlēta datu kopa "*heart_diseases.tab*" no *Orange* aplikācijas pieejamajiem piemēriem.
2) Datu kopa ir ideālā formātā *.tab*, lai ar to uzreiz varētu strādāt. Datu objekti, atverot tabulu, tiek uzrādīti pēc kārtas no 1 līdz 303, nākošajā kolonnā, uzrādot mērķa klases iezīmi, tālāk pārējās kolonnās, uzrādot pārējās atribūtu vērtības.
3) Datu kopā nav nevienas teksta veida vērtības, kuras būtu jātransformē kā skaitliskās vērtības.
4) Dažiem datu objektiem trūka dažas vērtības, tāpēc autore pielietoja "*Impute*", lai nebūtu vairāk tukšās vērtības.



3.att. Tukšo vērtību noņemšana ar "*Impute*" palīdzību.

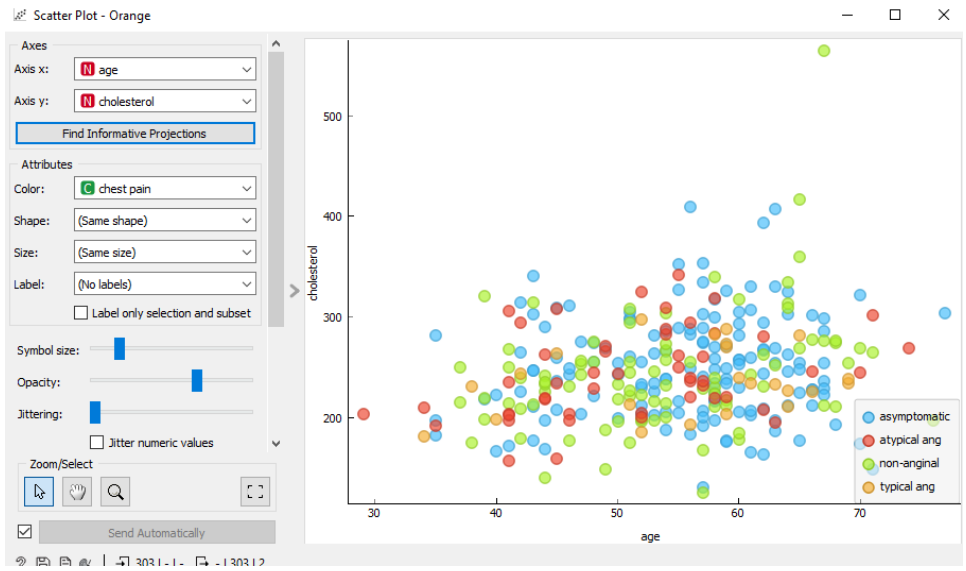
Bija skaidri redzams, ka 0.2% vērtību nav, tāpēc papildus izveidojam "*Impute*" un vēl vienu "*Data Table*", lai vairāk nebūtu tukšuma vērtību.

5) Tiek atspoguļota datu kopa vizuāli, arī aprēķinot statistiskos rādītājus:
   a) Izvairoties izmantot datu objekta *ID* kā mainīgo ir jāizveido divas 2 vai 3 dimensiju izkliedes diagrammas (*scatter plot*):



4.att. "*Scatter plot*" diagrammas izveide 1.

Šajā diagrammā tiek aplūkoti dati uz x ass salīdzinot vecumu, uz y ass salīdzinot miera stājas asinsspiedienu. Krāsa tiek iekrāsota, uzrādot kāda veida krūšu sāpes ir pacientam. Autores prāt, datu kopā klases nav līdzsvarotas, pārsvarā dominē zilās krāsas klase. Datu vizuālais atspoguļojums ļauj redzēt datu struktūru. Ir iespējams identificēt četrus datu grupējumus kaut vai tie saplūst kopā. Identificētie datu grupējumi lielākoties atrodas tuvu viens otram, bet ir daži, kuri ir arī tālāk.

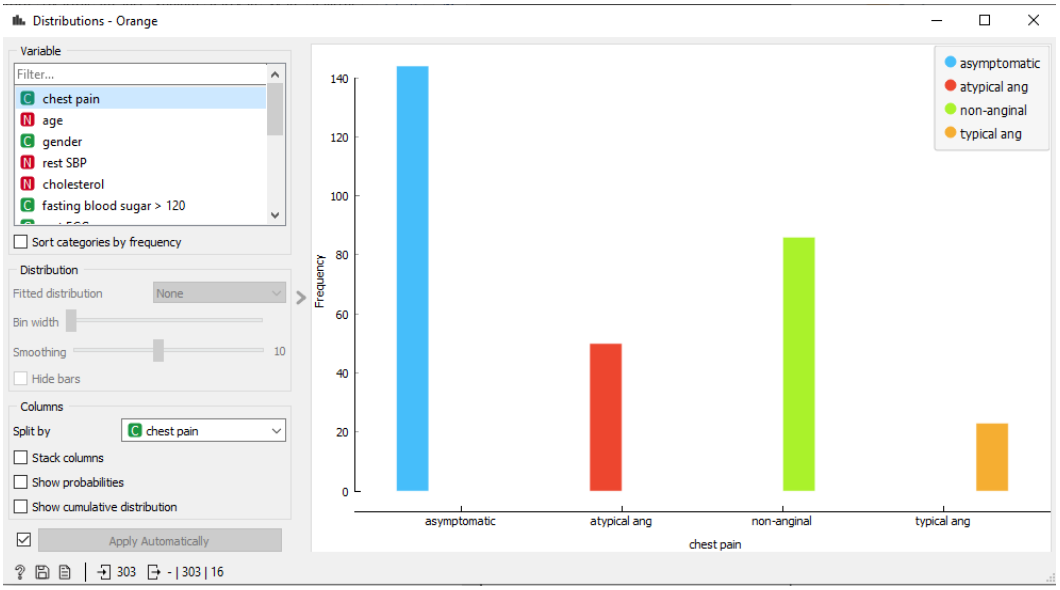


5.att. "*Scatter plot*" diagrammas izveide 2.

Šajā diagrammā tiek aplūkoti dati uz x ass salīdzinot vecumu, uz y ass salīdzinot holesterīna līmeni. Krāsa tiek iekrāsota, uzrādot kāda veida krūšu sāpes ir pacientam.

Autores prāt, datu kopā klases nav līdzsvarotas, pārsvarā dominē zilās un zaļās krāsas klases. Datu vizuālais atspoguļojums ļauj redzēt datu struktūru. Ir iespējams identificēt četrus datu grupējumus kaut vai tie saplūst kopā. Identificētie datu grupējumi lielākoties atrodas tuvu viens otram, bet ir daži, kuri ir arī tālāk.

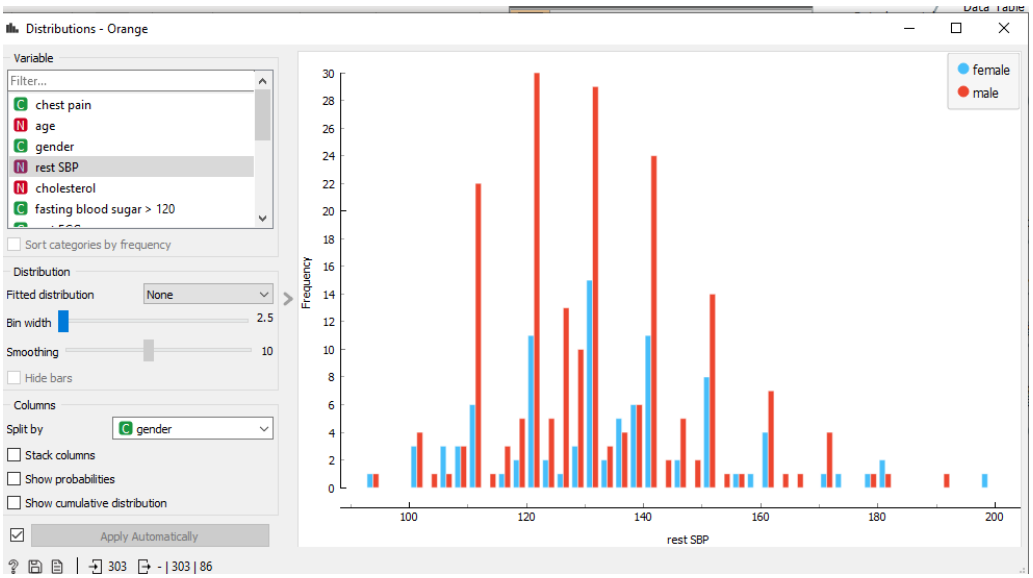
b) ir jāizveido vismaz 2 histogrammas, kas parāda klašu atdalīšanu, pamatojoties uz interesējošām pazīmēm:



6.att. "*Distributions*" histogrammas izveide 1.

Šajā histogrammā tiek aplūkoti dati uz x ass salīdzinot krūšu sāpes, uz y ass salīdzinot daudzumu. Dati tiek atdalīti, sadalot tos pa krūšu sāpju kategorijām. Šajā attēlā ir redzams tas, kuras tad ir visbiežākās krūšu sāpes.

Autores prāt, datu kopā klases nav līdzsvarotas, pārsvarā dominē zilās krāsas klase. Datu vizuālais atspoguļojums ļauj redzēt datu struktūru. Ir iespējams identificēt četrus datu grupējumus. Dati ir atdalīti viens no otra.

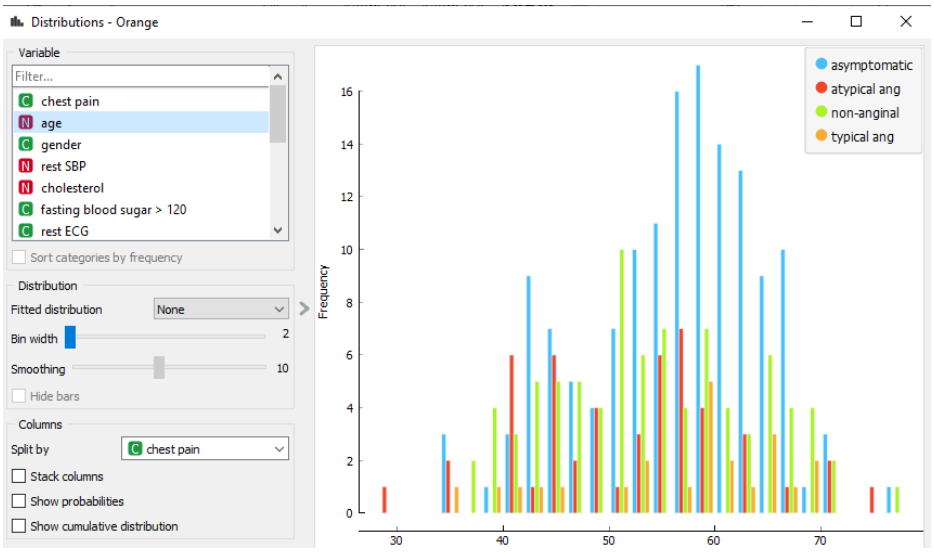


7.att. "*Distributions*" histogrammas izveide 2.

Šajā histogrammā tiek aplūkoti dati uz x ass salīdzinot asinsspiedienu miera stājā, uz y ass salīdzinot daudzumu. Dati tiek atdalīti, sadalot tos pa dzimumiem. Šajā attēlā ir redzams tas, kuram dzimumam lielākoties ir vislielākais asinsspiediens.

Autores prāt, datu kopā klases nav līdzsvarotas, pārsvarā dominē sarkanās krāsas klase. Datu vizuālais atspoguļojums ļauj redzēt datu struktūru. Ir iespējams identificēt divus datu grupējumus. Identificētie datu grupējumi no sākuma un beigās atrodas tuvu viens otram, bet pa vidu ir milzīga atšķirība starp dzimumiem. Vīriešiem lielākoties asinsspiediens ir 120 sitieni minūtē, bet sievietēm 130 sitieni minūtē.

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

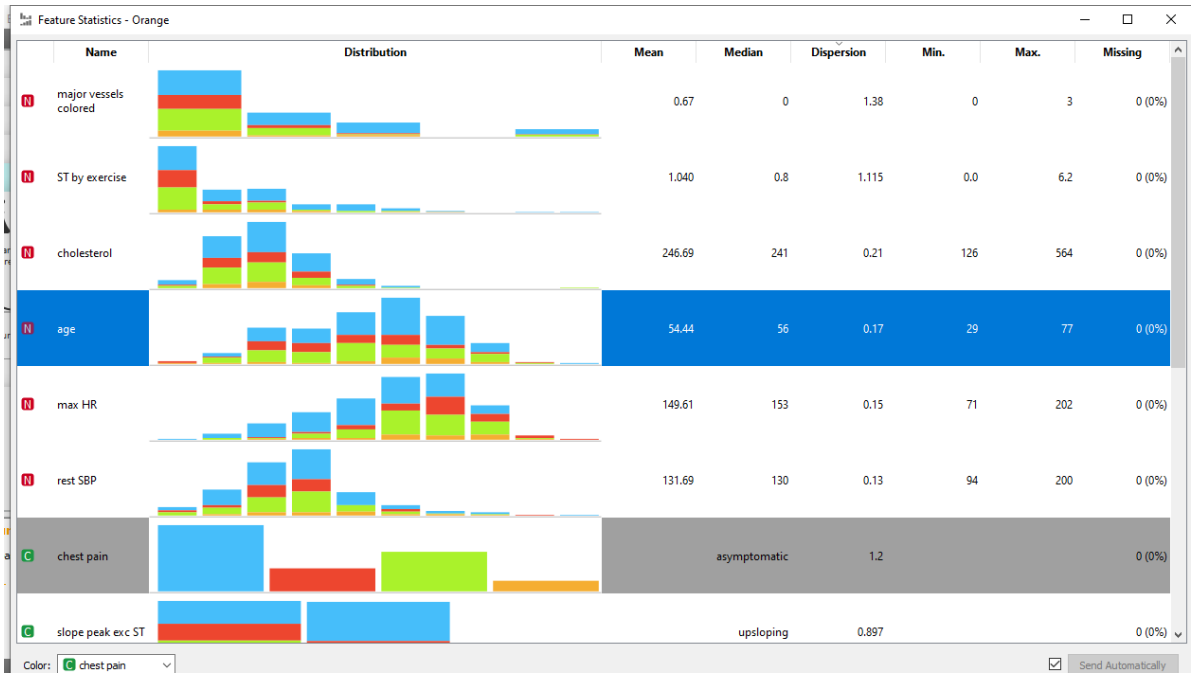


8.att. "*Distributions*" histogrammas izveide 3.

Šajā histogrammā tiek atspoguļoti dati par krūšu sāpēm un cilvēku vecumu.

Autores prāt, datu kopā klases nav līdzsvarotas, pārsvarā dominē zilās krāsas klase. Datu vizuālais atspoguļojums ļauj redzēt datu struktūru. Ir iespējams identificēt četrus datu grupējumus. Identificētie datu grupējumi lielākoties atrodas tuvu viens otram, bet ir zilais datu grupējums, kas ir tālāk. Lielāties asimptomātiskas sāpes ir pacientiem ap 60 gadu vecumu.

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



9.att. Statistisko rādītāju aprēķināšana ar "*Feature Statistics*".

Pie otrās datu tabulas tiek pievienots "*Feature Statistics*" un tiek apskatīti dati, kas iekļauj min, max, mediānu, dispersiju, kā arī datus, kuri trūkst. Ir redzami daži dati, kur vidējā vērtība pārsniedz mediānas vērtību, kas nosaka to, ka lielākoties ir cilvēki, kuriem atribūtu vērtības lielākas par vidējā cilvēka atribūtu vērtību, tas ir, paaugstinātas vērtības.

II DAĻA – NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS

Jāpielieto divi studiju kursā apskatītie nepārraudzītās mašīnmācīšanās algoritmi: (1) hierarhiskā klasterizācija un (2) K-vidējo algoritms.

1) Hierarhiskās klasterizācijas (*Hierarchical Clustering*) algoritmam ir jāveic vismaz 3 eksperimenti, brīvi mainot hiperparametru vērtības, un analizējot algoritma darbību:

   Logrīks atbalsta tālāk norādītos veidus, kā mērīt attālumus starp kopām:

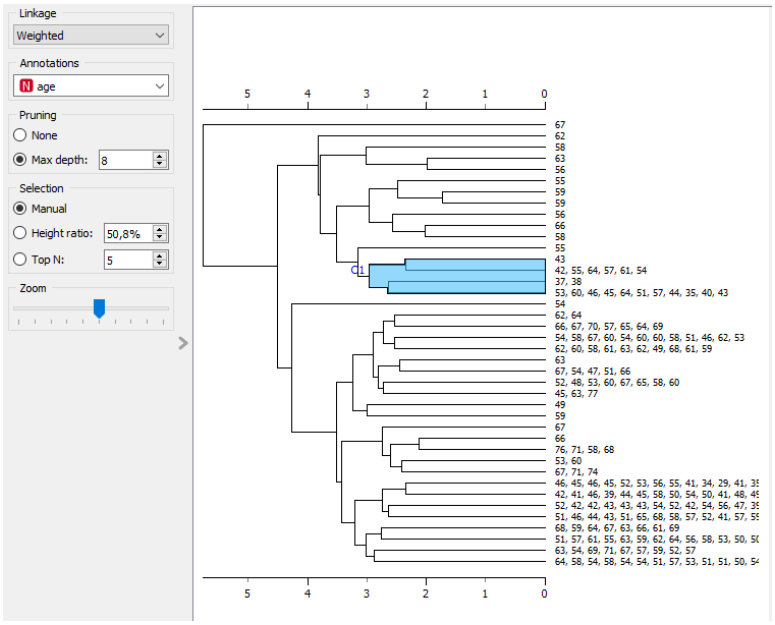
- Viena saite aprēķina attālumu starp tuvākajiem divu klasteru elementiem.
- Vidējā saite aprēķina vidējo attālumu starp divu klasteru elementiem.
- Svērtajā savienojumā tiek izmantota *WPGMA* metode.
- Pilnīga saite aprēķina attālumu starp klasteru visattālākajiem elementiem.
- *Ward* saite aprēķina kvadrātu kļūdu summas pieaugumu. Citiem vārdiem sakot, nodaļas minimālās dispersijas kritērijs samazina kopējo klastera dispersiju.

   Dendrogrammas mezglu etiķetes var izvēlēties lodziņā Anotācija.

Milzīgas dendrogrammas var apgriezt lodziņā Atzarošana, izvēloties maksimālo dendrogrammas dziļumu. Tas ietekmē tikai displeju, nevis faktisko klasterizāciju.

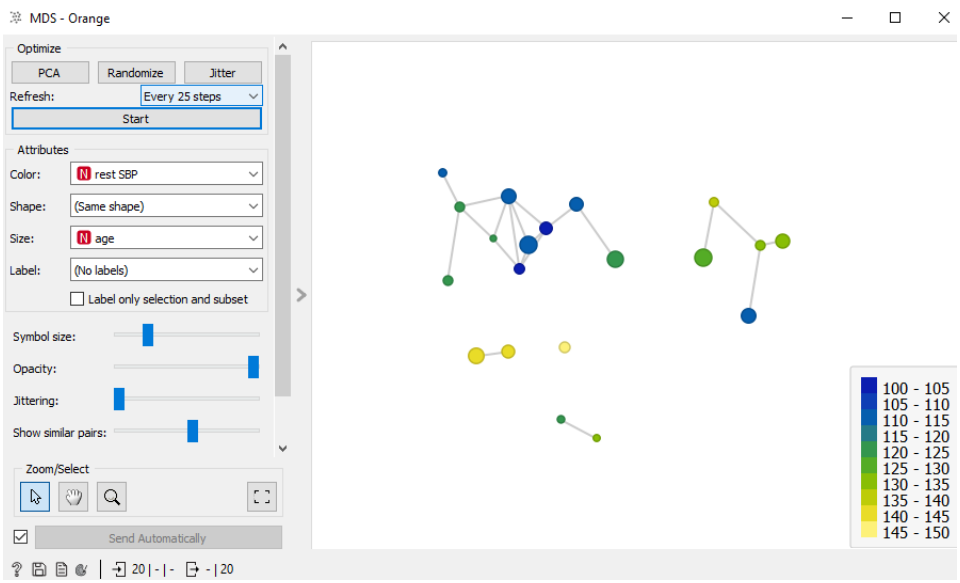
Logrīks piedāvā trīs dažādas atlases metodes:

- Manuāli (Noklikšķinot dendrogrammas iekšpusē, tiks atlasīts klasteris. Vairākus klasteri var atlasīt, turot nospiestu taustiņu *Ctrl/Cmd*. Katrs atlasītais klasteris tiek parādīts citā krāsā un izvadā tiek uzskatīts par atsevišķu klasteru.).
- Augstuma attiecība (Noklikšķinot uz dendrogrammas apakšējā vai augšējā lineāla, grafikā tiek parādīta robežlīnija. Tiek atlasīti vienumi pa labi no līnijas.).
- Top N (atlasa augšējo mezglu skaitu.) (Orange Data Mining, 2015).



10.att. "*Hierarchical Clustering*" datu aplūkošana 1.

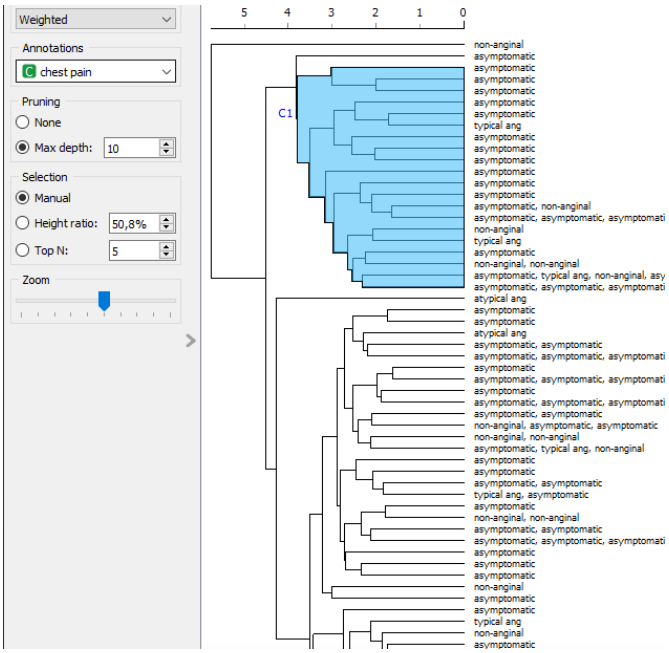
Pie otrās datu tabulas tiek pievienots "*Hierarchical Clustering*" un tiek apskatīti dati svērtajā savienojumā, par vecumu, kur maksimālais dziļums ir 8. Tā, kā autore vēlas apskatīt datus padziļinātāk, tad dati tiks apskatīti ar funkcijas "*MDS*" pievienošanu pie "*Hierarchical Clustering*".



11.att. "*MDS*" datu aplūkošana 1.

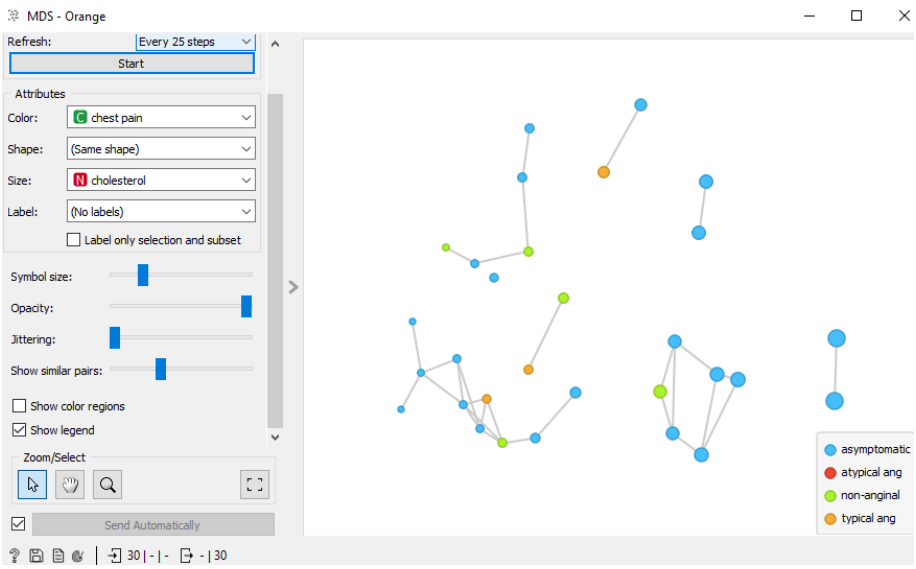
Autore "*MDS*" datos aplūko miera stājas asinsspiedienu salīdzinājumā ar vecumu. Autore ir atzīmējusi, lai parādās līdzīgo pāru salīdzinājumi. Tur, kur ir lielāku krāsu aplīši, tur ir lielāks vecums. Ir skaidri redzams tas, ka dati tiek sadalīti pa vecuma

grupām, tādējādi vecuma grupās uzrādot asinsspiediena augstumus. Dažām grupām pievienojas arī citu grupu dati, bet tas autores prāt ir datu līdzības dēļ, jo ir pieprasīts, lai tiek uzrādīti līdzīgi pāri.



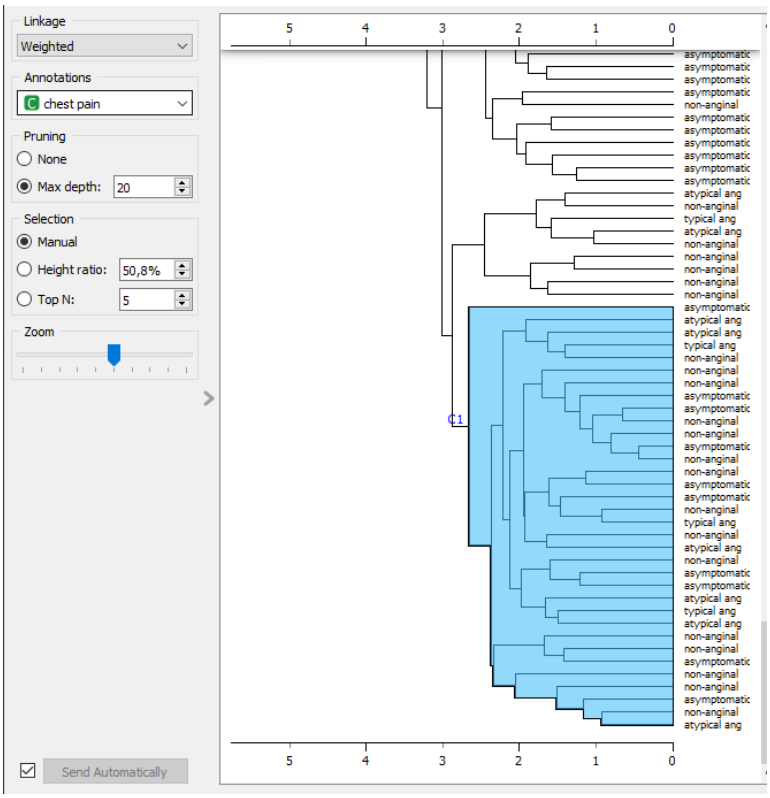
12.att. "*Hierarchical Clustering*" datu aplūkošana 2.

Pie otrās datu tabulas tiek pievienots "*Hierarchical Clustering*" un tiek apskatīti dati svērtajā savienojumā, par krūšu sāpēm, kur maksimālais dziļums ir 10.



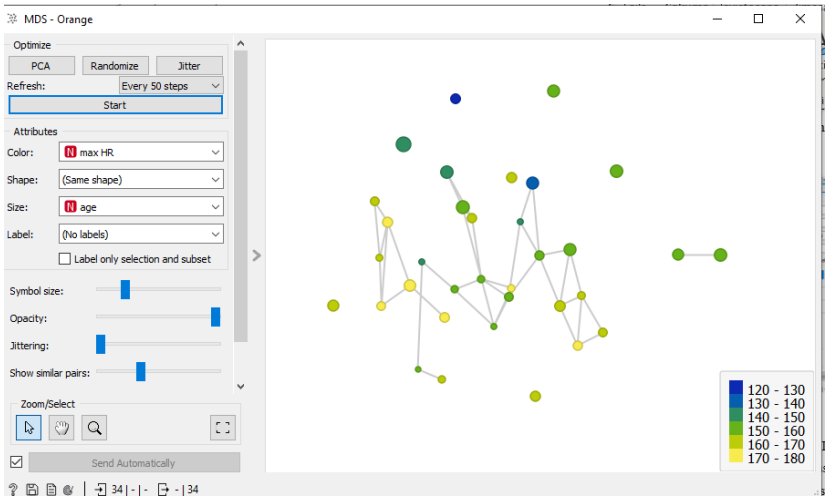
13.att. "*MDS*" datu aplūkošana 2.

Autore “*MDS*” datos aplūko krūšu sāpes salīdzinājumā ar holesterīna līmeni. Autore ir atzīmējusi, lai parādās līdzīgo pāru salīdzinājumi. Tur, kur ir lielāku krāsu aplīši, tur ir lielāks holesterīna lielums. Ir skaidri redzams tas, ka dati tiek sadalīti pa holesterīna lieluma grupām. Šeit ir iespējams secināt to, ka lielākoties visās holesterīna grupās ir asimptomātiskas krūšu sāpes. Vismazāk ir tipiskas krūšu sāpes.



14.att. “*Hierarchical Clustering*” datu aplūkošana 3.

Pie otrās datu tabulas tiek pievienots “*Hierarchical Clustering*” un tiek apskatīti dati svērtajā savienojumā, par krūšu sāpēm, kur maksimālais dziļums ir 10.



15.att. "*MDS*" datu aplūkošana 3.

Autore "*MDS*" datos aplūko maksimālo sirdspukstu skaitu minūtē salīdzinājumā ar vecumu. Autore ir atzīmējusi, lai parādās līdzīgo pāru salīdzinājumi. Tur, kur ir lielāku krāsu aplīši, tur ir lielāks vecums. Ir skaidri redzams tas, jo lielāks cilvēka vecums, jo mazāks ir sirdspukstu skaits minūtē un otrādāk.

2) K-vidējo algoritmam ir jāveic eksperimenti ar vismaz 5 k vērtībām, jāaprēķina Silhouette Score, un jāanalizē algoritma darbība:

Atlasiet klasteru (*k-Means*) skaitu:

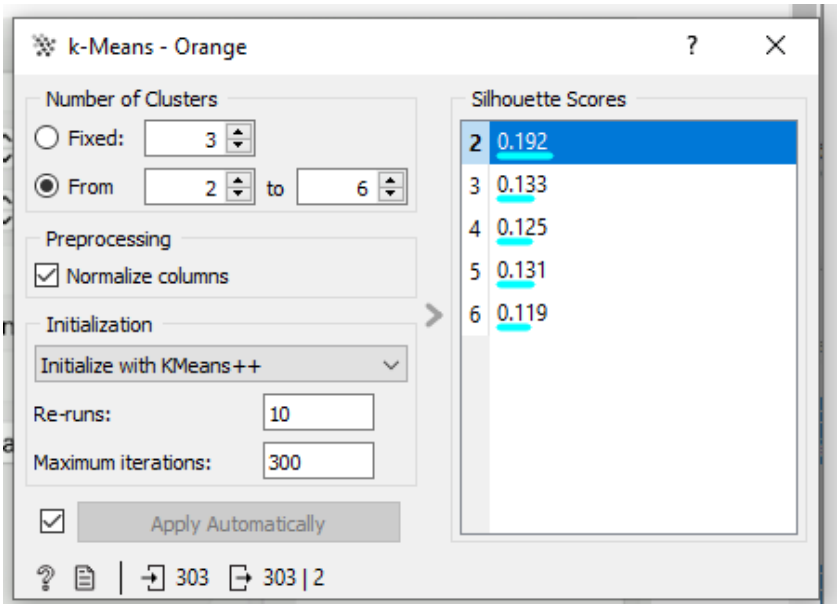
- Izlabots: algoritms sagrupē datus noteiktam klasteru skaitam.
- No X līdz Y: logrīks parāda klasterizācijas rādītājus atlasītajam klasteru diapazonam, izmantojot Silueta punktu (vidējo attālumu līdz elementiem vienā klasterī kontrastē ar vidējo attālumu līdz elementiem citās kopās).

Iepriekšēja apstrāde: ja ir atlasīta opcija, kolonnas tiek normalizētas (vidējais centrēts līdz 0 un standarta novirze mērogota līdz 1).

Inicializācijas metode (veids, kā algoritms sāk klasterizāciju):

- *k-Means++* (pirmais centrs tiek izvēlēts nejauši, nākamie tiek izvēlēti no atlikušajiem punktiem ar varbūtību, kas proporcionāla attālumam kvadrātā no tuvākā centra).
- Nejaušas inicializācijas (klasteri sākumā tiek piešķirti nejauši un pēc tam atjaunināti ar turpmākām iterācijām).

Atkārtotas palaišanas (cik reižu algoritms tiek palaists no nejaušām sākotnējām pozīcijām; tiks izmantots rezultāts ar zemāko kvadrātu summu klasterī) un maksimālās iterācijas (maksimālais iterāciju skaits katrā algoritma izpildē) var iestatīt manuāli (Orange University of Ljubljana, 2022).



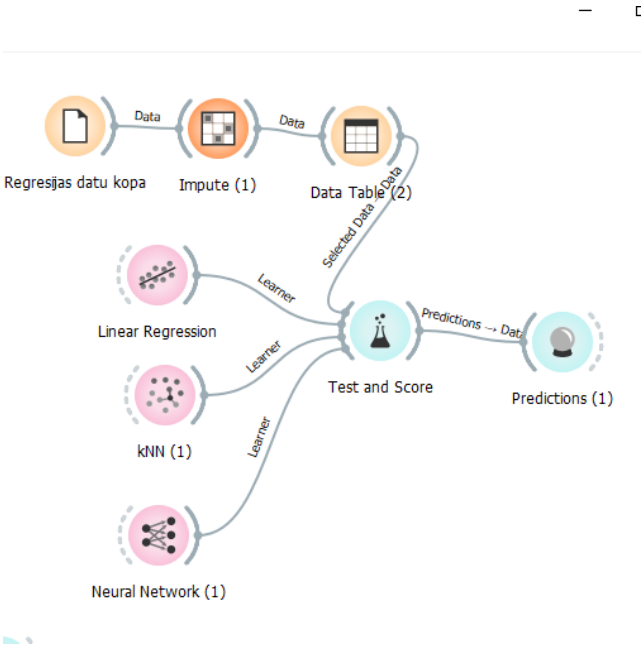
16.att. "*k-Means*" datu aplūkošana.

Autore "*k-Means*" datos tiek veikti aprēķini ar piecām k vērtībām, sākot tās skaitīt no 2 līdz 6 ieskaitot. Autore izvēlējās izmantot inicializāciju ar *KMeans++*. Tiek aprēķināti "*Silhouette Scores*". Maksimālās iterācijas ir 300, atkārtotās palaišanas reizes ir 10. Vislabākais datu sadalījums ir 2 klāsteros. Jo tuvāk vieniniekam, jo labāk klāsteri ir atdalāmi. Tomēr ir redzams, ka vērtējums ir diezgan tuvs nullei, kas uzrāda to, ka šie dati nav īpaši nozīmīgi. Tā, kā ir 4 veidu krūšu sāpes, ir interesanti tas, ka tiek parādīts, ka vislabāk būtu sadalīt 2 daļās, jo dati ir diezgan līdzīgi.

III DAĻA – PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS

Autore izvēlējās pielietot lineāro regresiju un *kNN* datu klasifikācijas metodes. Šie datu izveides veidi tika izvēlēti, jo *Orange* aplikācijā šiem algoritmiem ir pieejams apraksts, kur ir aprakstīts, ka šīs ir klasifikācijas metodes. Pārējiem algoritmiem nav aprakstīts, ka tās būtu klasifikācijas metodes.

1) Tiek izveidots datu tīkls.



17.att. Otrā datu tīklu aplūkošana.

Autore no sākuma izvēlējās "*File*", pēc tam pievienoja "*Impute*", lai nebūtu pazudušie dati, pievienoja datu tabulu, kā arī testa datus, pie testa datiem autore pievienoja 3 algoritmus:

- "*Linear Regression"*
- "*kNN*"
- "*Neural Network*".

Beigās pie testa datiem tika pievienots arī "*Predictions*".

Testa un apmācību datu kopai tiek pievienoti 66% no datiem. 20 reizes ir testu atkārtošana, lai iegūtu pareizākos datus. 66% no 303 pacientiem ir 199,98 pacientu dati. Visi trīs algoritmi tad arī darbojas ar 199,98 pacientu datiem jeb 66% no kopējiem datiem.

Testa datu (*Test and Score)* detalizēts apraksts:

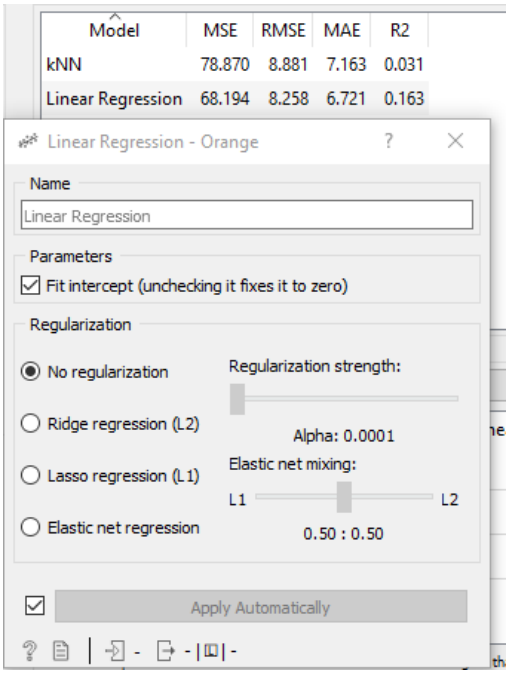
- *MSE* mēra kļūdu vai noviržu kvadrātu vidējo lielumu (starpība starp novērtētāju un aprēķināto).
- *RMSE* ir kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības (novērtētāja atbilstības datiem nepilnības mērs).
- *MAE* izmanto, lai noteiktu, cik tuvu prognozes vai prognozes ir iespējamajiem rezultātiem.
- *R2* tiek interpretēts kā atkarīgā mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma.
- *CVRMSE* ir *RMSE*, kas normalizēta ar faktisko vērtību vidējo vērtību (Orange University of Ljubljana, 2022).

2) "*Linear Regression*" algoritma izpilde.

Ir iespējami četri regulāciju veidi:

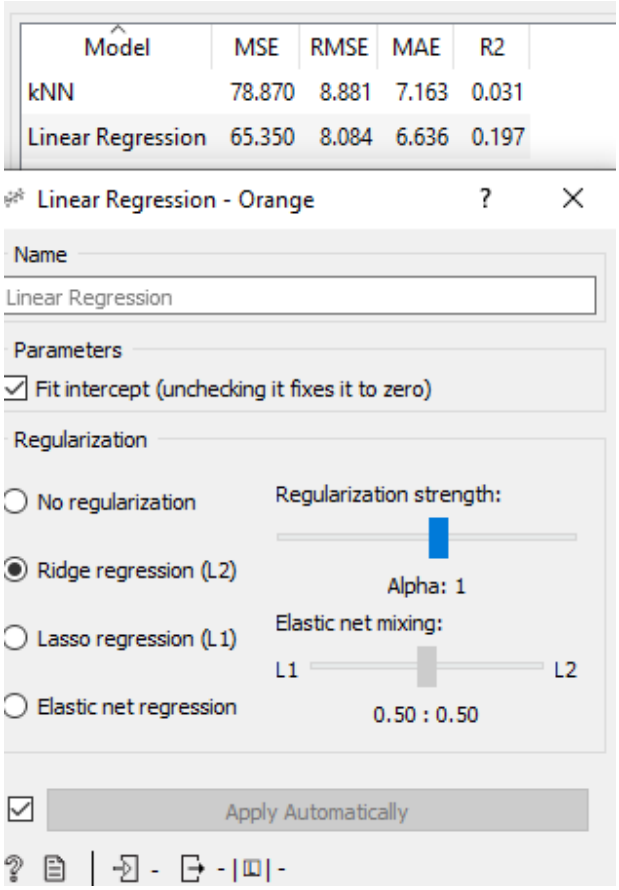
- Bez regulācijas
- *Ridge* regresija
- *Lasso* regresija
- Elastīgā tīkla regresija

Kā arī ir iespējams nomainīt regulācijas stiprumu.



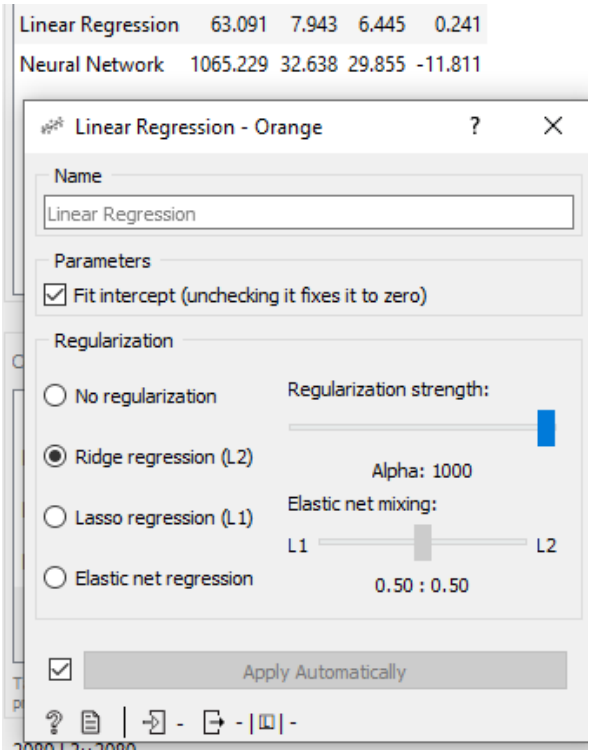
18.att. Lineārās regresijas datu aplūkošana 1.

Autore izvēlējās izveidot datus bez regulācijas, tās stiprums ir alfa 0.0001. Kļūdu vai noviržu kvadrātu vidējais lielums ir 68.194, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 8.258, prognozes ir 6.721 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.163.



19.att. Lineārās regresijas datu aplūkošana 2.

Autore izvēlējās izveidot datus ar *Ridge* regresiju, tās stiprums ir alfa 1. Kļūdu vai noviržu kvadrātu vidējais lielums ir 65.350, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 8.084, prognozes ir 6.636 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.197.



20.att. Lineārās regresijas datu aplūkošana 3.

Autore izvēlējās izveidot datus ar *Ridge* regresiju, tās stiprums ir alfa 1000. Kļūdu vai noviržu kvadrātu vidējais lielums ir 63.091, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 7.943, prognozes ir 6.445 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.241. Autores prāt, vislabākais veids ir pielietot "*Ridge regression*" un alfa 1000, jo šajā gadījumā ir vismazākais kļūdu vai noviržu kvadrātu vidējais lielums, kas nosaka to, ka datos ir vismazāk kļūdu.

3) kNN - (*kNN*) algoritma izpilde:

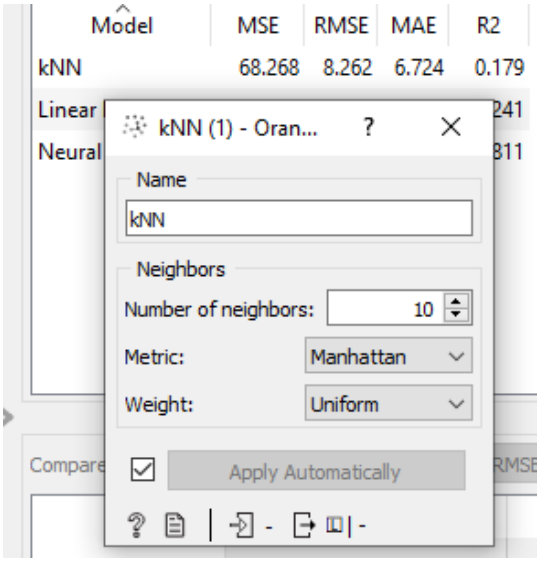
Metrika var būt:

- Eiklīds ("taisna līnija", attālums starp diviem punktiem).
- Manhetena (visu atribūtu absolūto atšķirību summa).
- Maksimālais (lielākā absolūtā atšķirība starp atribūtiem).
- Mahalanobis (attālums starp punktu un sadalījumu).

Svari, kurus varat izmantot, ir:

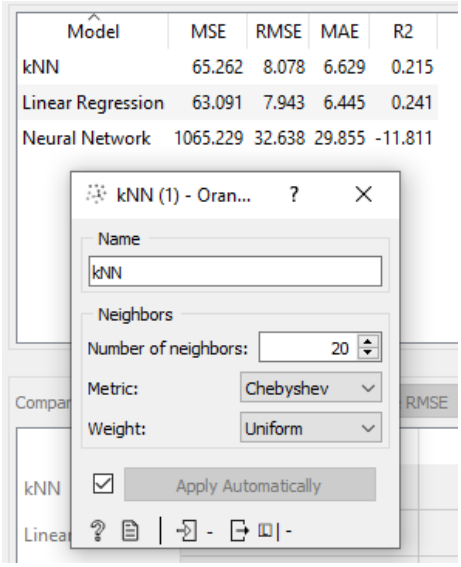
- Vienots: visi punkti katrā apkaimē tiek svērti vienādi.

- Attālums: tuvākiem vaicājuma punkta kaimiņiem ir lielāka ietekme nekā tālākiem kaimiņiem (Orange Data Mining, 2015).



21.att. *kNN* datu aplūkošana 1.

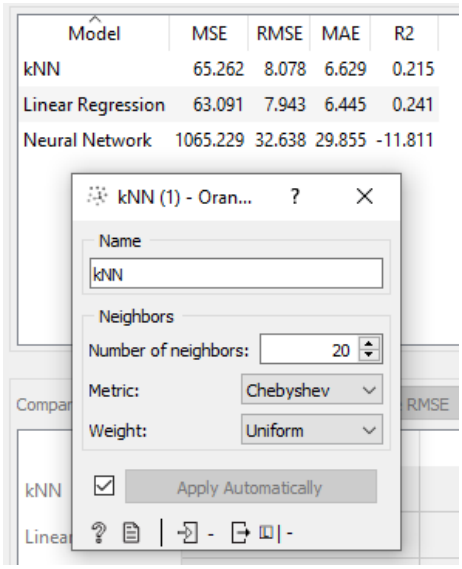
Autore izvēlējās izveidot datus ar *Manhattan*, tās kaimiņi ir 10. Kļūdu vai noviržu kvadrātu vidējais lielums ir 68.268, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 8.262, prognozes ir 6.724 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.179.



22.att. *kNN* datu aplūkošana 2.

Autore izvēlējās izveidot datus ar *Chebyshev*, tās kaimiņi ir 20. Kļūdu vai noviržu kvadrātu vidējais lielums ir 65.262, kvadrātsakne no skaitļu kopas kvadrātu

aritmētiskās vidējās vērtības ir 8.078, prognozes ir 6.629 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.179.



23.att. *kNN* datu aplūkošana 3.

Autore izvēlējās izveidot datus ar *Manhattan*, tās kaimiņi ir 100. Kļūdu vai noviržu kvadrātu vidējais lielums ir 72.299, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 8.503, prognozes ir 6.445 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.241.

Autores prāt, vislabākais bija otrais variants ar "*kNN*" algoritma izmantošanu, jo tur bija vismazākais kļūdu vai noviržu kvadrātu vidējais lielums.

4) Neirona tīkla - (*Neural Network)* algoritma izpilde:

Neironi slēptā slānī: definēts kā i-tais elements, kas apzīmē neironu skaitu i-tajā slēptajā slānī. Piem. neironu tīklu ar 3 slāņiem var definēt kā 2, 3, 2.

Slēptā slāņa aktivizēšanas funkcija:

- Identitāte: bezoperācijas aktivizēšana, noderīga, lai īstenotu lineāro sašaurinājumu.
- Loģistika: loģistikas sigmoīda funkcija.
- *tanh*: hiperboliskā iedeguma funkcija.
- *ReLu*: rektificētas lineārās vienības funkcija.

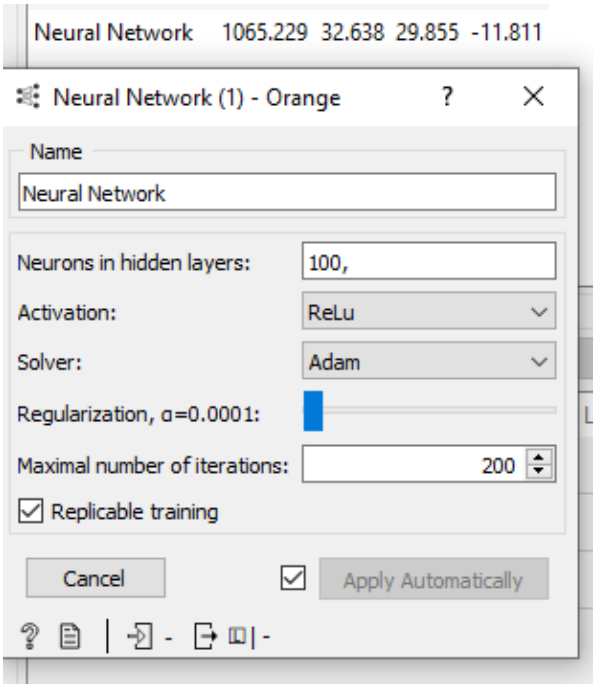
Risinātājs svara optimizēšanai:

- *L-BFGS-B*: optimizētājs kvaziņūtona metožu saimē.
- *SGD*: stohastiskā gradienta nolaišanās.
- Ādams: uz stohastisko gradientu balstīts optimizētājs.

Alfa: *L2* soda (regulēšanas termiņa) parametrs.

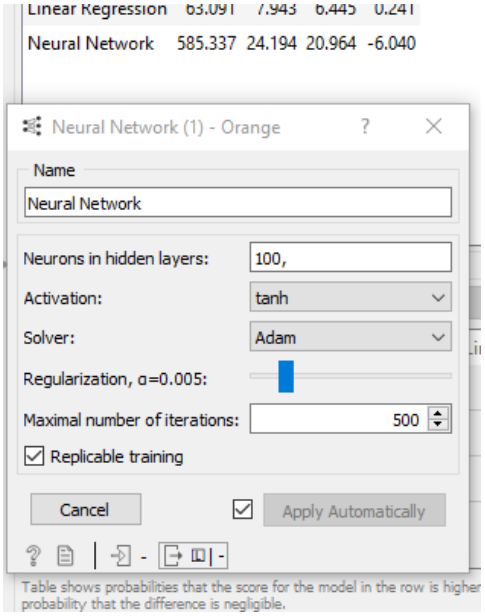
Maksimālais atkārtojumu skaits: maksimālais atkārtojumu skaits.

Citi parametri ir iestatīti uz *sklearn* noklusējuma iestatījumiem (Orange Data Mining, 2015).



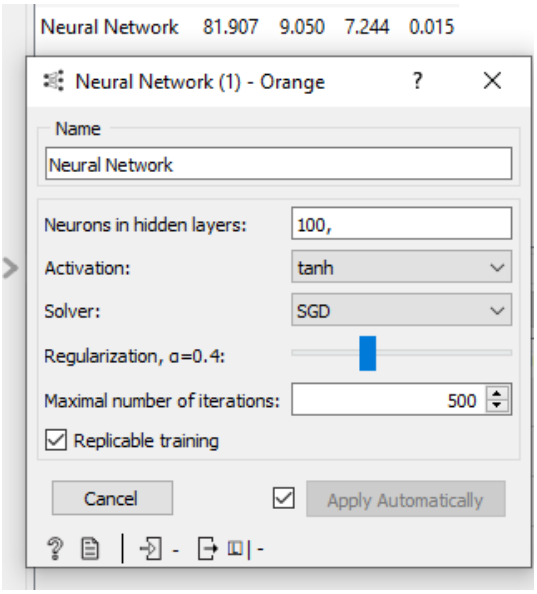
24.att. *Neural Network* datu aplūkošana 1.

Autore izvēlējās izveidot datus ar rektificētas lineārās vienības funkciju un uz stohastisko gradientu balstītu optimizētāju, ir 100 neironi slēptajā slānī, maksimālais iterāciju skaits ir 200 un regulācija ir alfa 0.0001. Kļūdu vai noviržu kvadrātu vidējais lielums ir 1065.229, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 32.638, prognozes ir 29.855 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir -11.811.



25.att. *Neural Network* datu aplūkošana 2.

Autore izvēlējās izveidot datus ar hiperboliskā iedeguma funkciju un uz stohastisko gradientu balstītu optimizētāju, ir 100 neironi slēptajā slānī, maksimālais iterāciju skaits ir 500 un regulācija ir alfa 0.005. Kļūdu vai noviržu kvadrātu vidējais lielums ir 585.337, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 24.194, prognozes ir 20.964 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir -6.040.



26.att. *Neural Network* datu aplūkošana 3.

Autore izvēlējās izveidot datus ar hiperboliskā iedeguma funkciju un stohastiskā gradienta nolaišanos, ir 100 neironi slēptajā slānī, maksimālais iterāciju

skaits ir 500 un regulācija ir alfa 0.04. Kļūdu vai noviržu kvadrātu vidējais lielums ir 81.907, kvadrātsakne no skaitļu kopas kvadrātu aritmētiskās vidējās vērtības ir 9.050, prognozes ir 7.244 tuvu, kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir 0.015.

Autores prāt, vislabākais bija trešais variants ar "*Neural Network*" algoritma izmantošanu, jo tur bija vismazākais kļūdu vai noviržu kvadrātu vidējais lielums. Kā arī mainīgā dispersijas proporcija, kas ir paredzama no neatkarīgā mainīgā lieluma ir pozitīva.

SECINĀJUMI

Autore darbu izveidoja *Orange* vidē. Ļoti labi bija aprakstīti uzdevumi kā tos pildīt, kādā secībā, kā arī, kas tieši ir nepieciešams no katra algoritma. Autore uzskata, ka darba prasības ir labas. Informācija tika iegūta no pasniedzējas video un no internetā atrastajiem pieejamajiem resursiem.

No visiem iegūtajiem datiem tikai veidotas tabulas, diagrammas, histogrammas, aprēķini, kā arī salīdzinājumi. Ar datiem tika veiktas korelācijas.

Tika izveidoti vairāki algoritmi:

1) Scatter_plot – diagrammu izveide,
2) k-Means – klasteru izveide,
3) Distributions – histogrammu izveide,
4) Feature Statistics – statistisko rādītāju aprēķināšana,
5) MDS – datu attāluma aprēķināšana,
6) Hierarchical Clustering - datu sagrupēšana,
7) Linear Regression – datu regulācija,
8) kNN – paredzēšanas gadījumu aprēķins,
9) Neural Network – pavairošanas algoritms,
10) Test and Score – aprēķiniem,
11) Impute – lai noņemtu pazudušās vērtības.

Autore darbā īsti ar grūtībām nesastapās, jo viss bija skaidri paskaidrots kā uzdevumus izpildīt. Autorei šis darbs šķita ļoti interesants. Autores prāt, ļoti labi bija tas, ka *Orange* aplikācijā bija pieejamas jau gatavas datu bāzes, kuras var izmantot. Autore no sākuma meklēja datus internetā, bet saprata, ka labāk ir izmantot kaut ko, ko jau nodrošina aplikācija. Autorei dati par pacientu krūšu sāpēm likās vispiemērotākie šim klasifikācijas darbam. Protams, ka šajā datu bāzē bija arī būla dati, kas ir 0/1, bet tie nebija tik daudz, lai datu bāze nebūtu izmantojama.

Autore uzskata, ka padarījusi darbu korekti, kā arī autores prāt, šis darbs bija ļoti interesants. *Orange* aplikācija ir ļoti viegli pielietojama un, to ir ļoti viegli saprast.

IZMANTOTĀ LITERATŪRA

1) Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. (2013). *Orange: Data Mining Toolbox in Python* [online]. *Journal of Machine Learning Research* 14(Aug): 2349−2353, [accessed 19 May 2022]. Available at: *https://orangedatamining.com/citation/*
2) Orange University of Ljubljana (2022). *License* [online]. [Accessed 19 May 2022]. Available at: *https://orangedatamining.com/license/*
3) Orange Data Mining. (2015). *Hierarchical Clustering* [online]. [Accessed 19 May 2022]. Available at: *https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html*
4) Orange University of Ljubljana (2022). *k-Means* [online]. [Accessed 19 May 2022]. Available at: *https://orangedatamining.com/widget-catalog/unsupervised/kmeans/*
5) Orange University of Ljubljana (2022). *Test and Score* [online]. [Accessed 19 May 2022]. Available at: *https://orangedatamining.com/widget-catalog/evaluate/testandscore/*
6) Orange Data Mining. (2015). *kNN* [online]. [Accessed 19 May 2022]. Available at: *https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html*
7) Orange Data Mining. (2015). *Neural Network* [online]. [Accessed 19 May 2022]. Available at: *https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/neuralnetwork.html*