Rīgas Tehniskā Universitāte

Datorzinātnes un informācijas tehnoloģijas fakultāte

Otrais praktiskais darbs

Mašīnmācīšanās algoritmu lietojums

Autors: ----------

St.ap.nr.: ---------

Grupa: -

GitHub

2022/2023 māc.gads

Saturs

Ievads

Mākslīgā intelekta jomā mēs esam pētījuši un izmantojuši dažādas metodes un algoritmus datu apstrādei. Šis praktiskais darbs aicina mūs strādāt ar brīvi izvēlētu datu kopu tālākai apstrādei. Iegūtā datu kopa tiek apstrādāta lietojumprogrammā Orange, kas nodrošina plašu rīku un funkciju klāstu vizualizācijai, pirmapstrādei.

Darbs ir sadalīts trīs daļās:

1. Atlasīto datu pirmapstrāde un pārbaude.

2. Neuzraudzīta mašīnmācīšanās

3. Pārraudzītā mašīnmācīšanās

I daļa - Datu pirmapstrāde/izpēte

Sirds un asinsvadu slimības ir galvenais nāves un invaliditātes cēlonis visā pasaulē. PVO lēš, ka katru gadu tiek zaudētas 17,9 miljoni dzīvību, kas ir svarīgs izaicinājums medicīnas aprindām un pilsoniskās sabiedrības organizācijām cīņā ar šo problēmu. (1)

1.1 Datu kopa

Tika izvēlēts datu kopums "Heart Disease Cleveland". Šīs datu kopas autori ir Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano. Lai izveidotu šo datu kopu, autori izmantoja UCI repozitoriju, no kura tika ņemta Klīvlendas sirds slimību datubāze. Šīs datu kopas mērķis ir prognozēt, vai personai ir vai nav sirds un asinsvadu slimības.

Šis datu kopums ir pieejams šajā saitē: Datu kopa (2).

Šo datu kopu izplata saskaņā ar licenci: CC BY-SA 4.0 (9) ,Tas ļauj to izmantot izglītības nolūkos.

1.2 Datu kopas satura apraksts

Datu kopa satur: csv datubāzē ir: 303 ieraksti un 14 kritēriji.

Atribūta nosaukums	Vērtības	Tips
Age	Pacienta vecums gados (no 29 līdz 77)	Skaitlisks
Sex	Pacienta dzimums (Vīrieši: 1; sievietes: 0)	Kategorisks
Cp	Pacienta izjusto sāpju veids krūtīs 0 - tipiska stenokardija, 1 - atipiska stenokardija, 2 - nestenokardiskas sāpes, 3 - asimptomātisks stāvoklis.	Kategorisks
Trestbps	Pacienta asinsspiediena līmenis miera stāvoklī (no 94 līdz 200)	Skaitlisks
chol	Holesterīna līmenis asinīs (no 126 līdz 564)	Skaitlisks
fbs	Cukura līmenis asinīs uz nakti (Vairāk nekā 120 mg/dl - 1, Mazāk par 120 mg/dl - 0)	Kategorisks
restecg	Elektrokardiogrammas rezultāts miera stāvoklī	Kategorisks

	(0 - normāli, 1 - ST-T viļņu anomālija, 2 - iespējama vai noteikta kreisā kambara palielināšanās saskaņā ar Estes kritērijiem)	
thalach	Sasniegtais maksimālais sirdsdarbības ātrums. (no 77 līdz 202)	Skaitlisks
exang	Fiziskas slodzes izraisīta stenokardija. (0 nozīmē, ka nav, 1 nozīmē, ka ir)	Kategorisks
oldpeak	ST segmenta depresija, ko izraisa fiziska slodze, salīdzinot ar mierīgu stāvokli (no 0 līdz 6,2)	Skaitlisks
slope	ST segmenta slīpums maksimālās slodzes laikā (0 - augšupejošs, 1 - horizontāls, 2 - lejupejošs)	Kategorisks
ca	Lielo kuģu skaits (no 0 līdz 3)	Kategorisks
thal	Asins slimība, ko sauc par talasēmiju. (1 - normāla asinsrite, 2 - fiksēts defekts, 3 - atgriezenisks defekts)	Kategorisks
target	Atribūts, kas mums jāparedz (1- (1- pacients cieš no sirds slimībām, 0 - vesels)	Kategorisks

1.1 tabula Datu kopas atributi apraksts

Content

This database contains 13 attributes and a target variable. It has 8 nominal values and 5 numeric values. The detailed description of all these features are as follows:

1. Age: Patients Age in years (Numeric)
2. Sex: Gender (Male : 1; Female : 0) (Nominal)
3. cp: Type of chest pain experienced by patient. This term categorized into 4 category.
   0 typical angina, 1 atypical angina, 2 non- anginal pain, 3 asymptomatic (Nominal)
4. trestbps: patient's level of blood pressure at resting mode in mm/HG (Numerical)
5. chol: Serum cholesterol in mg/dl (Numeric)
6. fbs: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
7. restecg: Result of electrocardiogram while at rest are represented in 3 distinct values
   0 : Normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophyby Estes' criteria (Nominal)
8. thalach: Maximum heart rate achieved (Numeric)
9. exang: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
10. oldpeak: Exercise induced ST-depression in relative with the state of rest (Numeric)
11. slope: ST segment measured in terms of slope during peak exercise
    0: up sloping; 1: flat; 2: down sloping(Nominal)
12. ca: The number of major vessels (0–3)(nominal)
13. thal: A blood disorder called thalassemia
    0: NULL 1: normal blood flow 2: fixed defect (no blood flow in some part of the heart) 3: reversible defect (a blood flow is observed but it is not normal(nominal)
14. target: It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.

1.1 att. Atribūtu apraksts no Kaggle

Kopumā 164 subjekti datu kopā attiecas uz veseliem pacientiem (0) un pacientiem ar sirds slimībām (1) 139

Klases nosaukums	Objektu skaits
Vesels pacients(0)	164
Pacients ar sirds slimību (1)	139

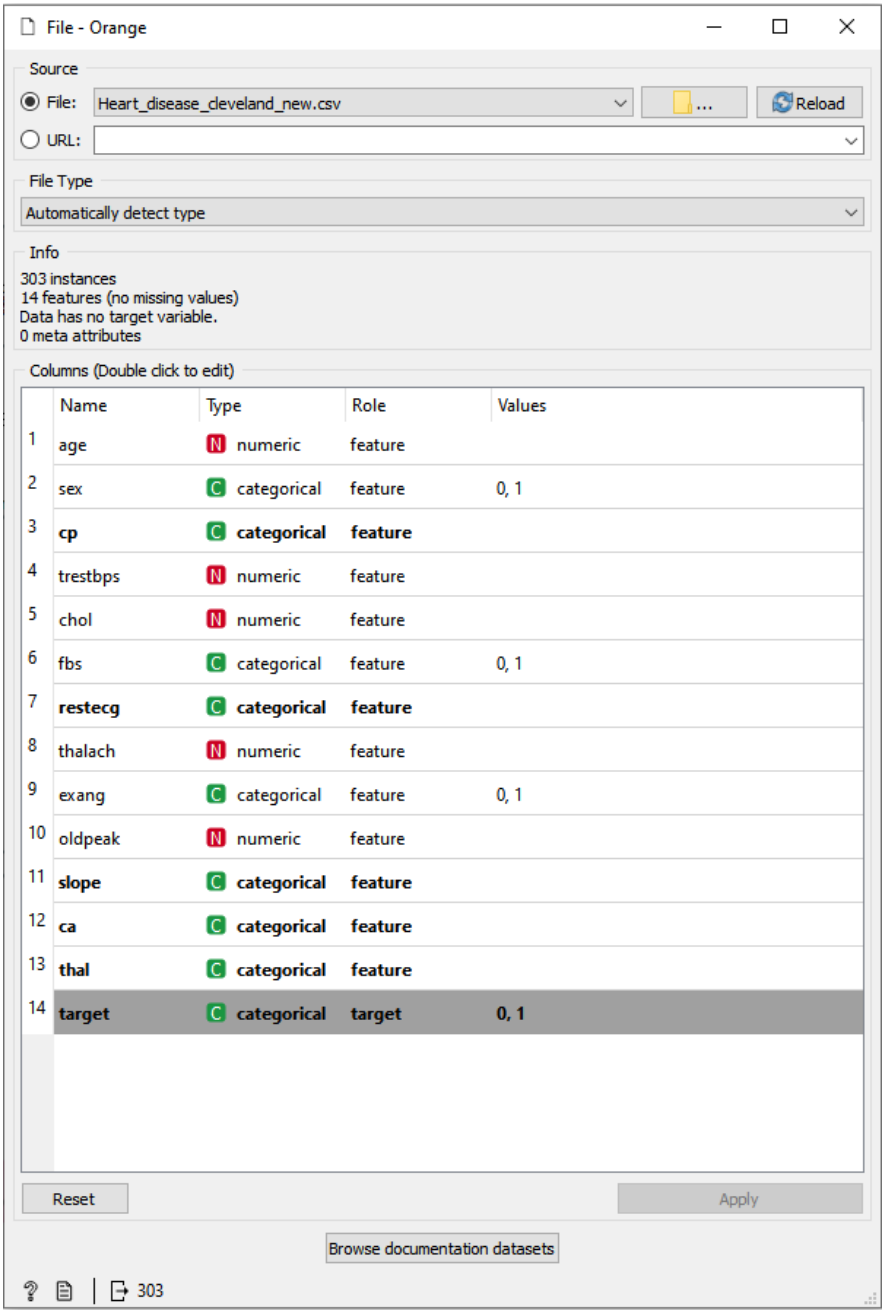
1.2 tabula "Target" atribūta objektu skaits.

	A	B	C	D	E	F	G	H
1	age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target							
2	63,1,0,145,233,1,2,150,0,2.3,2,0,2,0							
3	67,1,3,160,286,0,2,108,1,1.5,1,3,1,1							
4	67,1,3,120,229,0,2,129,1,2.6,1,2,3,1							
5	37,1,2,130,250,0,0,187,0,3.5,2,0,1,0							
6	41,0,1,130,204,0,2,172,0,1.4,0,0,1,0							
7	56,1,1,120,236,0,0,178,0,0.8,0,0,1,0							
8	62,0,3,140,268,0,2,160,0,3.6,2,2,1,1							
9	57,0,3,120,354,0,0,163,1,0.6,0,0,1,0							
10	63,1,3,130,254,0,2,147,0,1.4,1,1,3,1							
11	53,1,3,140,203,1,2,155,1,3.1,2,0,3,1							
12	57,1,3,140,192,0,0,148,0,0.4,1,0,2,0							
13	56,0,1,140,294,0,2,153,0,1.3,1,0,1,0							
14	56,1,2,130,256,1,2,142,1,0.6,1,1,2,1							
15	44,1,1,120,263,0,0,173,0,0,0,0,3,0							
16	52,1,2,172,199,1,0,162,0,0.5,0,0,3,0							
17	57,1,2,150,168,0,0,174,0,1.6,0,0,1,0							
18	48,1,1,110,229,0,0,168,0,1,2,0,3,1							
19	54,1,3,140,239,0,0,160,0,1.2,0,0,1,0							
20	48,0,2,130,275,0,0,139,0,0.2,0,0,1,0							
21	49,1,1,130,266,0,0,171,0,0.6,0,0,1,0							
22	64,1,0,110,211,0,2,144,1,1.8,1,0,1,0							
23	58,0,0,150,283,1,2,162,0,1,0,0,1,0							
24	58,1,1,120,284,0,2,160,0,1.8,1,0,1,1							
25	58,1,2,132,224,0,2,173,0,3.2,0,2,3,1							
26	60,1,3,130,206,0,2,132,1,2.4,1,2,3,1							
27	50,0,2,120,219,0,0,158,0,1.6,1,0,1,0							
28	58,0,2,120,340,0,0,172,0,0,0,0,1,0							
29	66,0,0,150,226,0,0,114,0,2.6,2,0,1,0							

1.2.1 att. Datu kopa .csv formata

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbp	chol	fbs	restecg	thalach	exang	oldpea	slope	ca	thal	target
2	63	1	0	145	233	1	2	150	0	2,3	2	0	2	0
3	67	1	3	160	286	0	2	108	1	1,5	1	3	1	1
4	67	1	3	120	229	0	2	129	1	2,6	1	2	3	1
5	37	1	2	130	250	0	0	187	0	3,5	2	0	1	0
6	41	0	1	130	204	0	2	172	0	1,4	0	0	1	0
7	56	1	1	120	236	0	0	178	0	0,8	0	0	1	0
8	62	0	3	140	268	0	2	160	0	3,6	2	2	1	1
9	57	0	3	120	354	0	0	163	1	0,6	0	0	1	0
10	63	1	3	130	254	0	2	147	0	1,4	1	1	3	1
11	53	1	3	140	203	1	2	155	1	3,1	2	0	3	1
12	57	1	3	140	192	0	0	148	0	0,4	1	0	2	0
13	56	0	1	140	294	0	2	153	0	1,3	1	0	1	0
14	56	1	2	130	256	1	2	142	1	0,6	1	1	2	1
15	44	1	1	120	263	0	0	173	0	0	0	0	3	0
16	52	1	2	172	199	1	0	162	0	0,5	0	0	3	0
17	57	1	2	150	168	0	0	174	0	1,6	0	0	1	0
18	48	1	1	110	229	0	0	168	0	1	2	0	3	1
19	54	1	3	140	239	0	0	160	0	1,2	0	0	1	0
20	48	0	2	130	275	0	0	139	0	0,2	0	0	1	0
21	49	1	1	130	266	0	0	171	0	0,6	0	0	1	0
22	64	1	0	110	211	0	2	144	1	1,8	1	0	1	0
23	58	0	0	150	283	1	2	162	0	1	0	0	1	0
24	58	1	1	120	284	0	2	160	0	1,8	1	0	1	1
25	58	1	2	132	224	0	2	173	0	3,2	0	2	3	1
26	60	1	3	130	206	0	2	132	1	2,4	1	2	3	1
27	50	0	2	120	219	0	0	158	0	1,6	1	0	1	0
28	58	0	2	120	340	0	0	172	0	0	0	0	1	0
29	66	0	0	150	226	0	0	114	0	2,6	2	0	1	0

1.2.2 att. Datu kopa .xlsx (Excel) formata

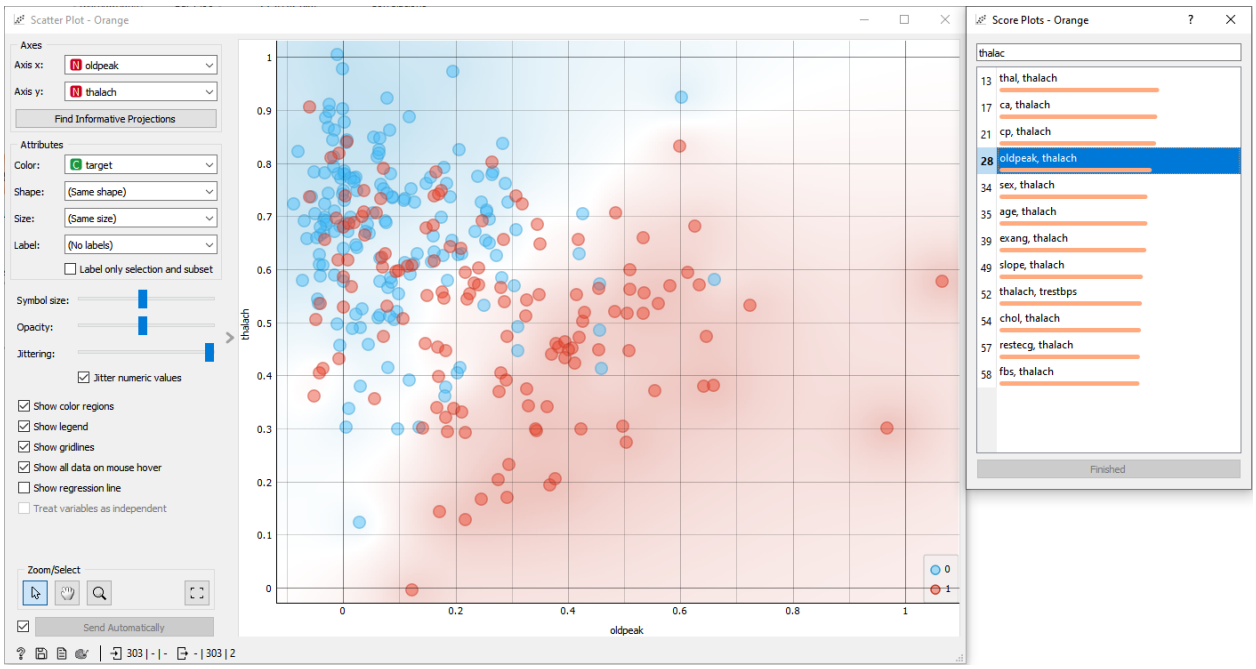


1.3 Att. Atribūtu attēlošana Orange. To tips, loma, iespējamās vērtības

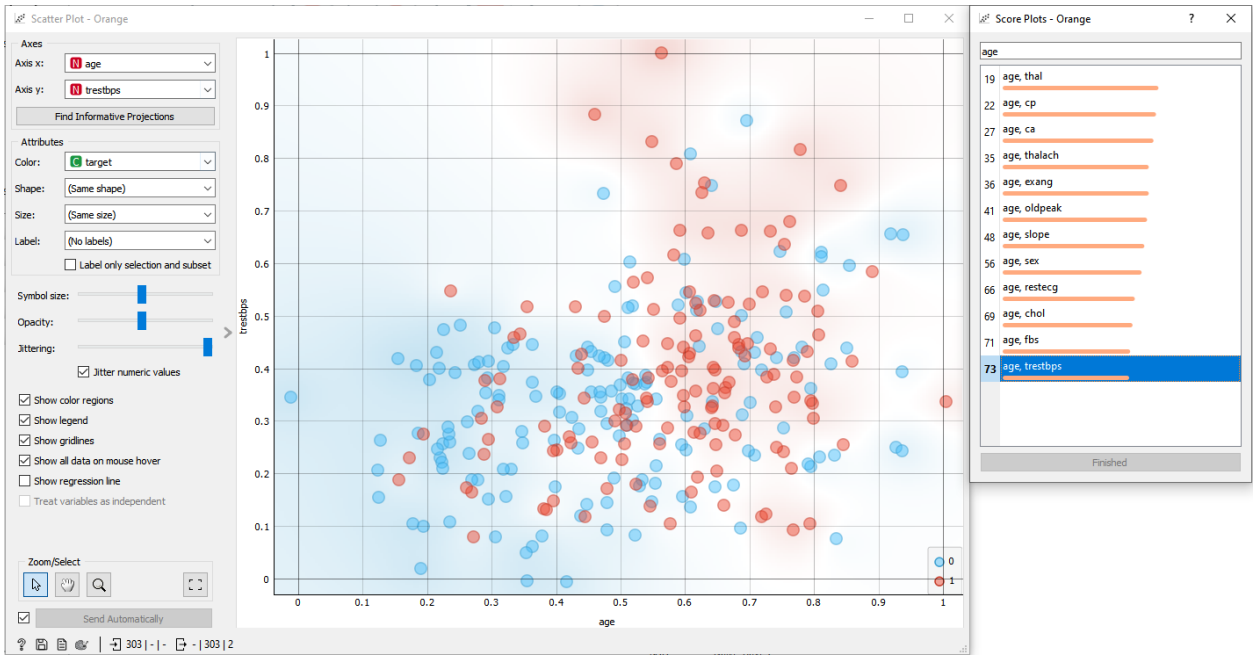
1.5 DATU KOPAS ANALĪZE

1.5.a Datu vizualizācija ar diagrammām

Lai vizualizētu klašu sadalījumu pēc atribūtiem, ir ērti izmantot izkliedes diagrammas. Datu analīzes un vizualizācijas programmatūras pakotnē Orange ir iespējams izveidot izkliedes diagrammu(Scatter Plot), lai vizuāli analizētu datus.



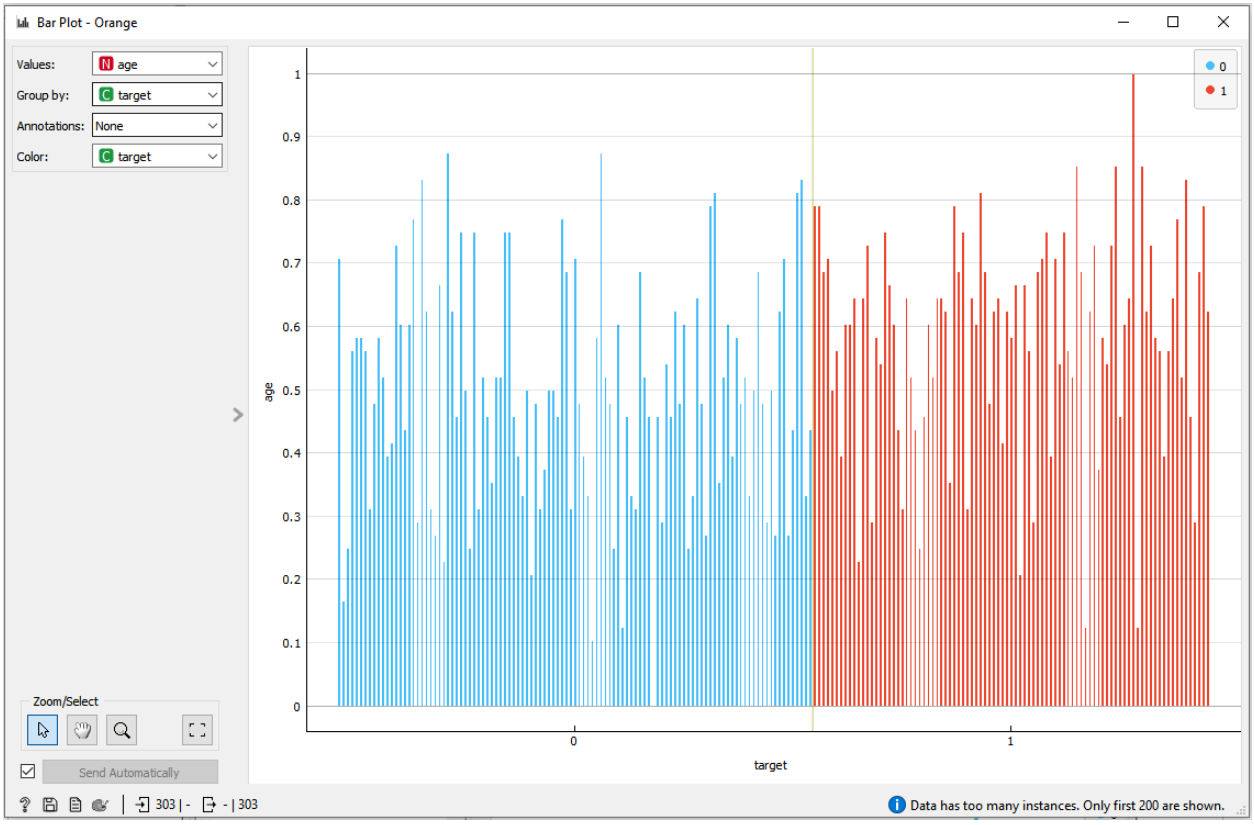
1.4.att. Scatter Plot oldpeak un thalach atribūts



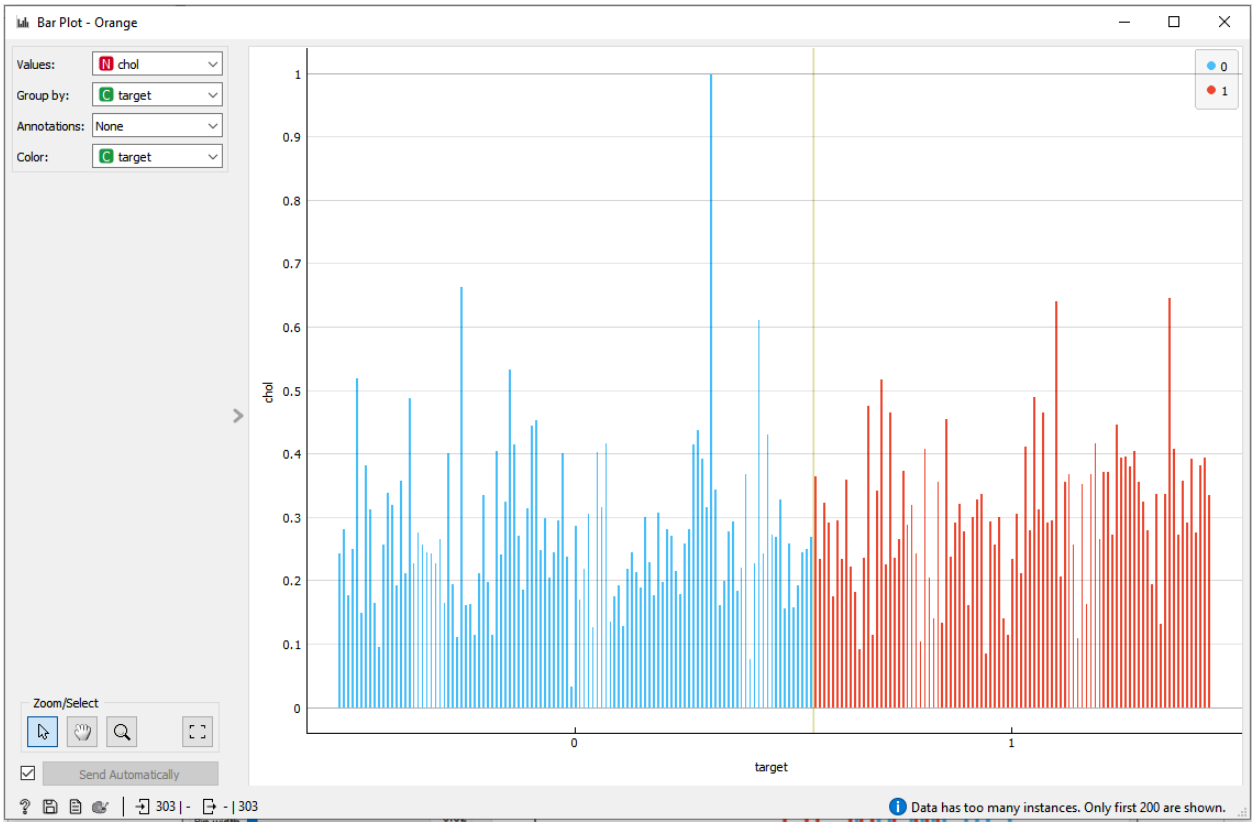
1.5.att. Scatter Plot age un trestbps atribūts

1.5.b Datu vizualizēšana ar histogrammām

Programmas Orange logrīku Bar plot izmanto, lai vizualizētu datus joslu diagrammu veidā, ko var izmantot, lai salīdzinātu dažādu kategoriju vērtības.

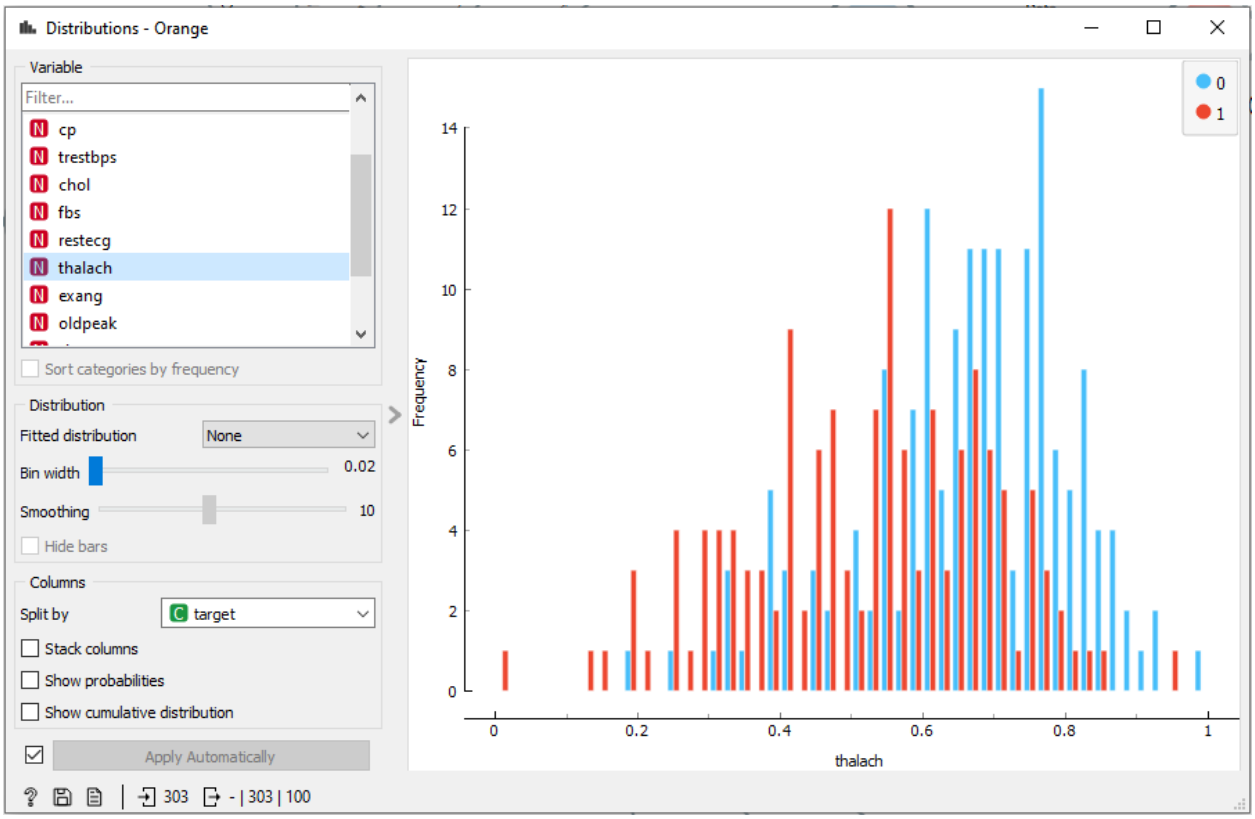


1.6.att. Bar Plot age atribūts

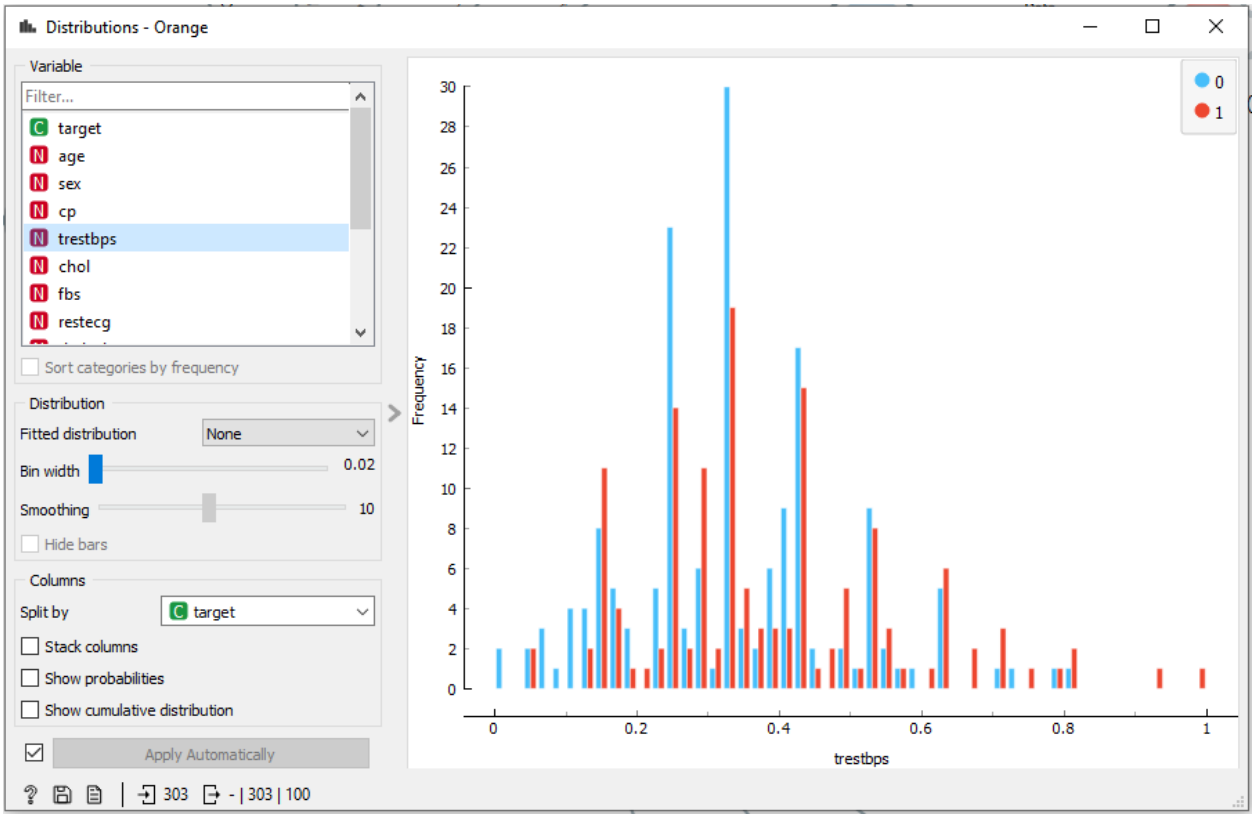


1.7.att. Bar Plot chol atribūts

1.5.c Konkrētu atribūtu vizualizācija

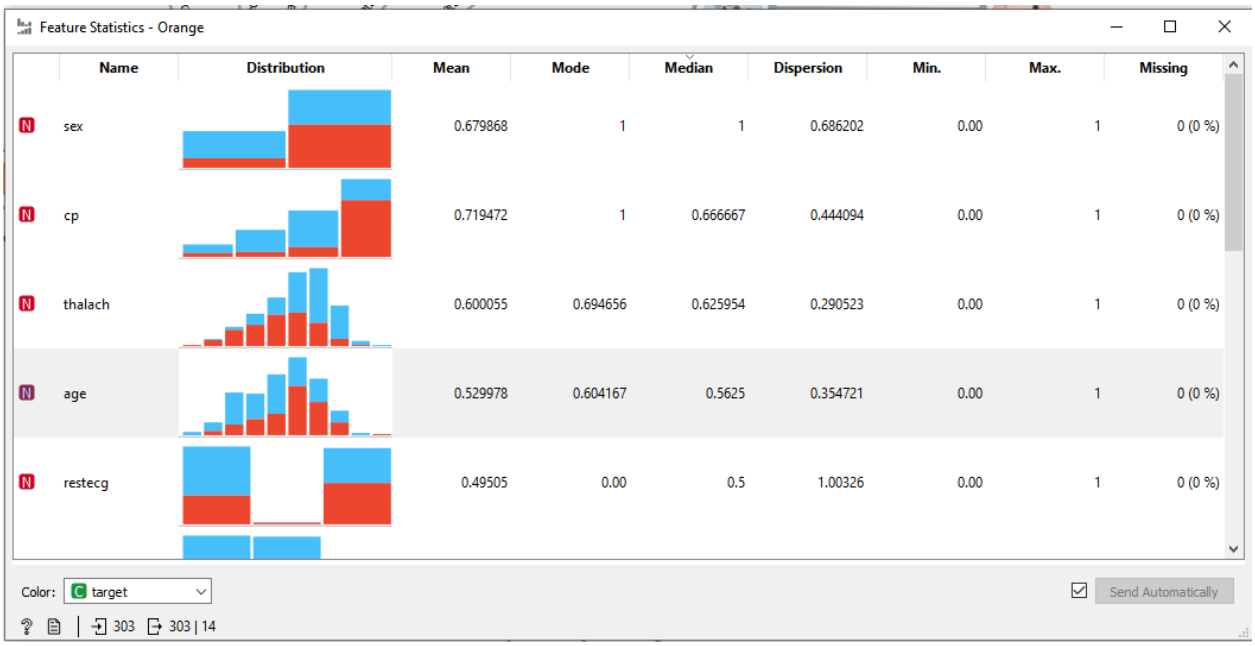


1.8.att. Distributions thalach atribūts

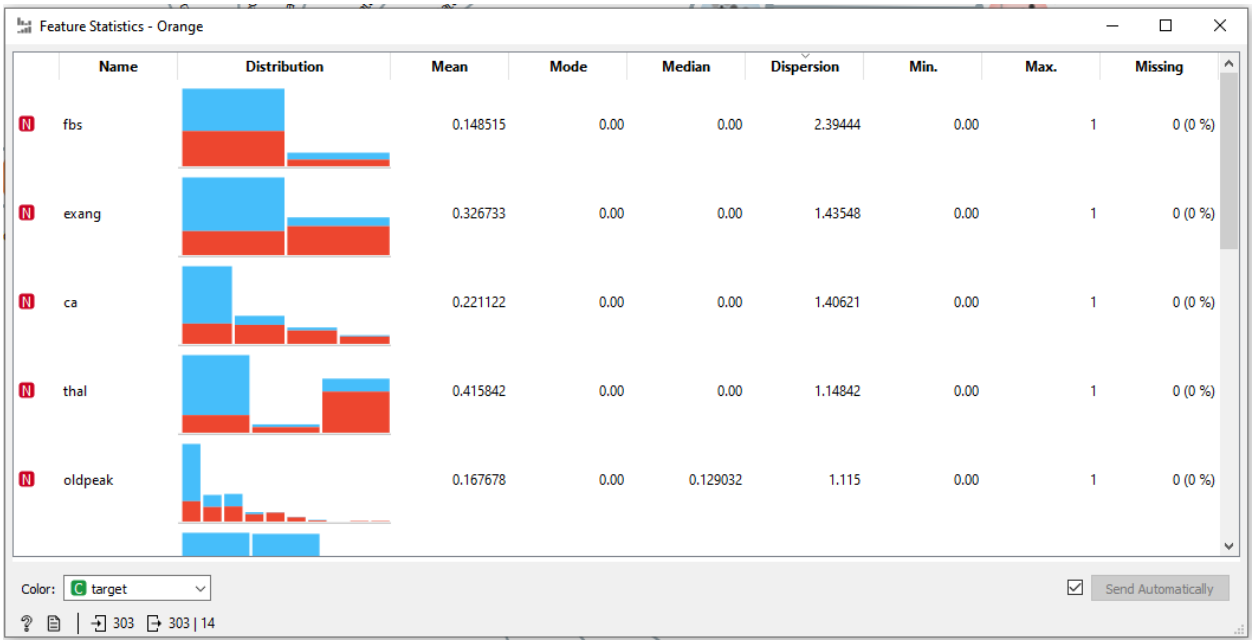


1.9.att. Distributions trestbps atribūts

1.5.d Statistikas rādītāji



1.10.att. Feature Statistics Median



1.11.att. Feature Statistics Dispersion

Pirmās daļas secinājumi

1. *Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)?*
   Šajā datu kopā dominē veselu pacientu klase ("target" = 0), nevis slimu pacientu klase ("target" = 1).

2. Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?
   Diemžēl vizuālais attēlojums nesniedza vairāk informācijas par datu struktūru.
   Tāpat kā izkliedes diagrammās(1.4 att. un 1.5 att.) nebija iespējams identificēt konkrētus klašu dalījumus, jo klašu objekti ir tuvu viens otram.

   Arī histogrammās(1.6 att. un 1.6 att.) nav konstatējami nekādi dalījumi.

3. Cik datu grupējums ir iespējams identificēt, pētot datu vizuālo atspoguļojumu?
   Aplūkojot diagrammas(1.4 att. un 1.5 att.), mēs nevaram identificēt grupas. Tā kā objekti ir sadalīti visā audeklā un krustojas ar objektiem no citas grupas

4. Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?
   Kā redzam no diagrammām (1.4 att. un 1.5 att.) mēs varam teikt, ka objektu grupas atrodas tuvu viena otrai

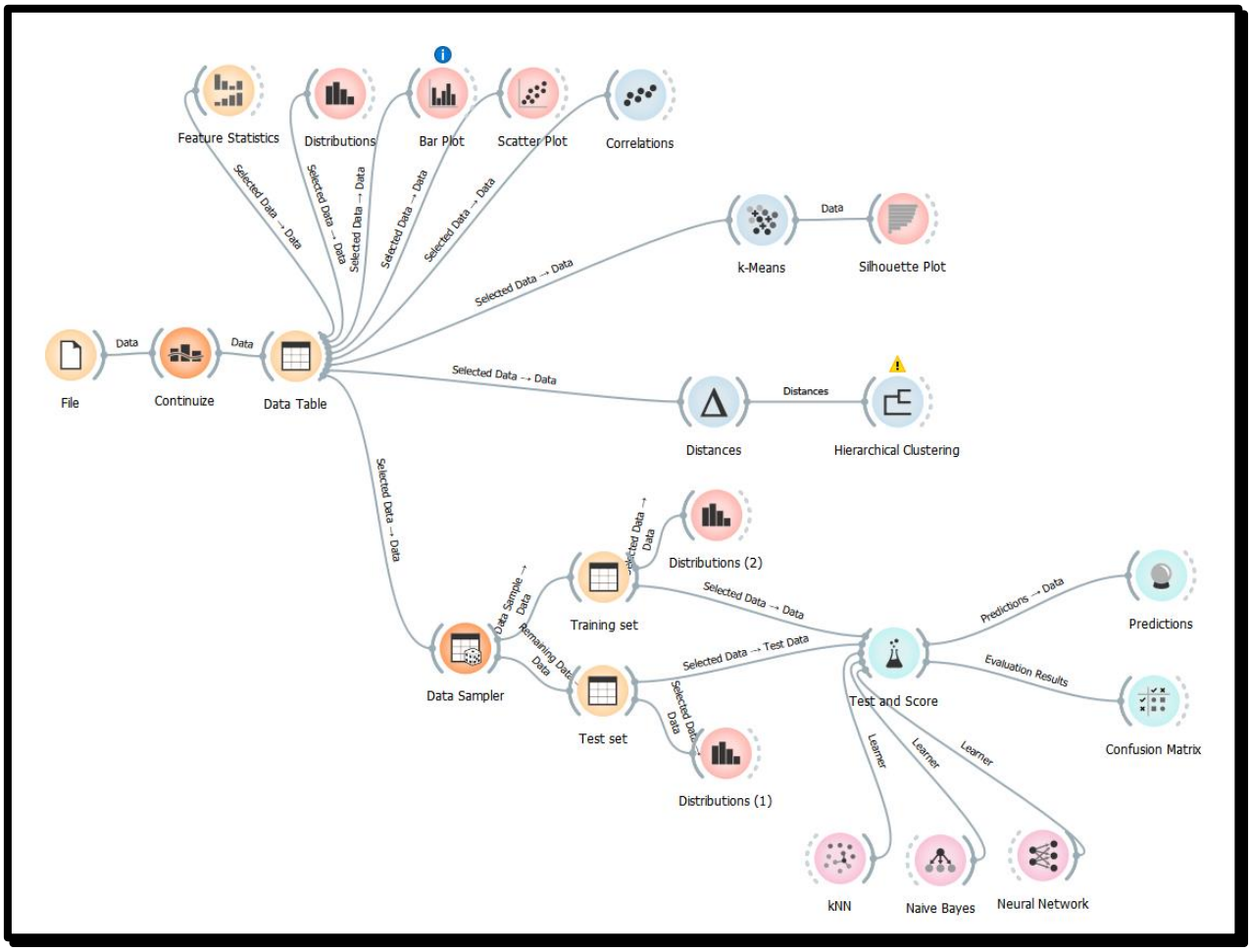
1.5 Datu kopas statistisko rādītāju analīze

Statistikas datus varam novērot attēlos (1.10 att. un 1.11 att.)No tā mēs varam teikt, ka:

- “sex” – Vidēji šajā datu kopā ir vairāk vīriešu nekā sieviešu, jo mediāna ir vienāda ar 1.
- “fbx” – Lielākajai daļai pacientu cukura līmenis asinīs ir zemāks par 120 mg

1.6 Orange rīka darbplūsmas atspoguļojums

Tālākai analīzei vidē tika ieviesta shēma Orange vidē (6) (7) (8)



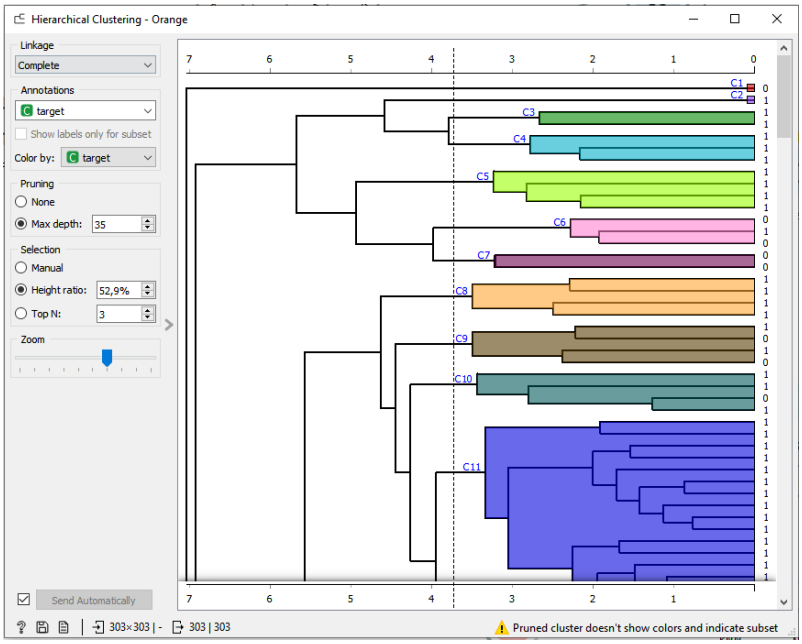
1.12 att. Orange rīka darbplūsmas atspoguļojums

II daļa - Nepārraudzītā mašīnmācīšanās

Lai apstrādātu datu kopu, izmantojot "Nepārraudzītā mašīnmācīšanās", tiek izmantoti 2 algoritmi:

1) hierarhiskā klasterizācija.
2) K-means algoritms.

2.1 Hierarhiskā klasterizācija



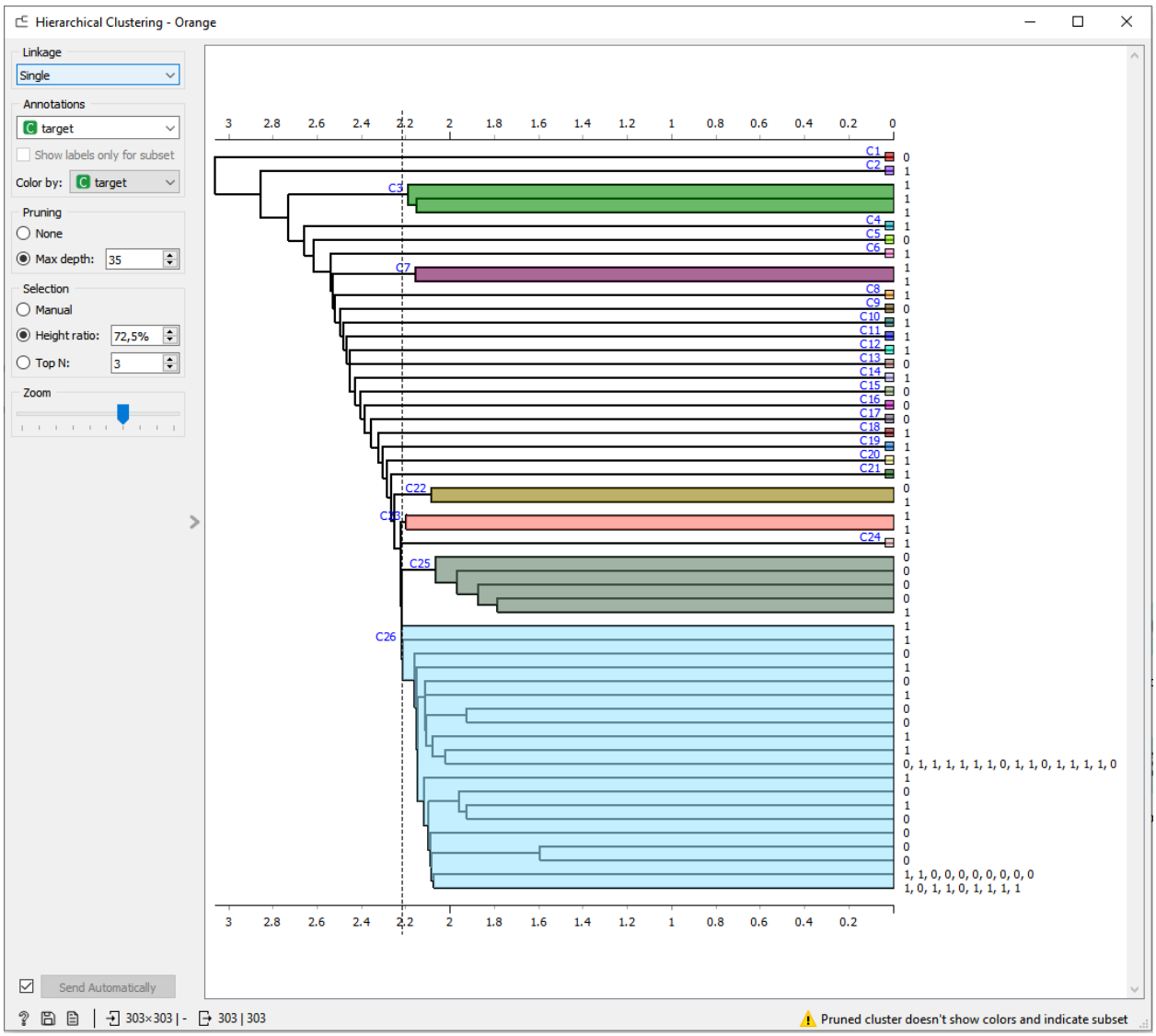
2.1 att. Hierarchical Clustering logs

Hierarhiskā klasterizācija ir tādi hiperparametri kā:

- Linkage
  - Single (aprēķina attālums starp divu klasteru tuvākajiem elementiem.)
  - Average (aprēķina vidējo attālumu starp divu klasteru elementiem.)
  - Weighted (izmanto WPMA metodi)
  - Complete (aprēķina attālums starp vistālāk esošajiem kopas elementiem.)
  - Ward (aprēķina kļūdas kvadrātu summas pieaugumu.

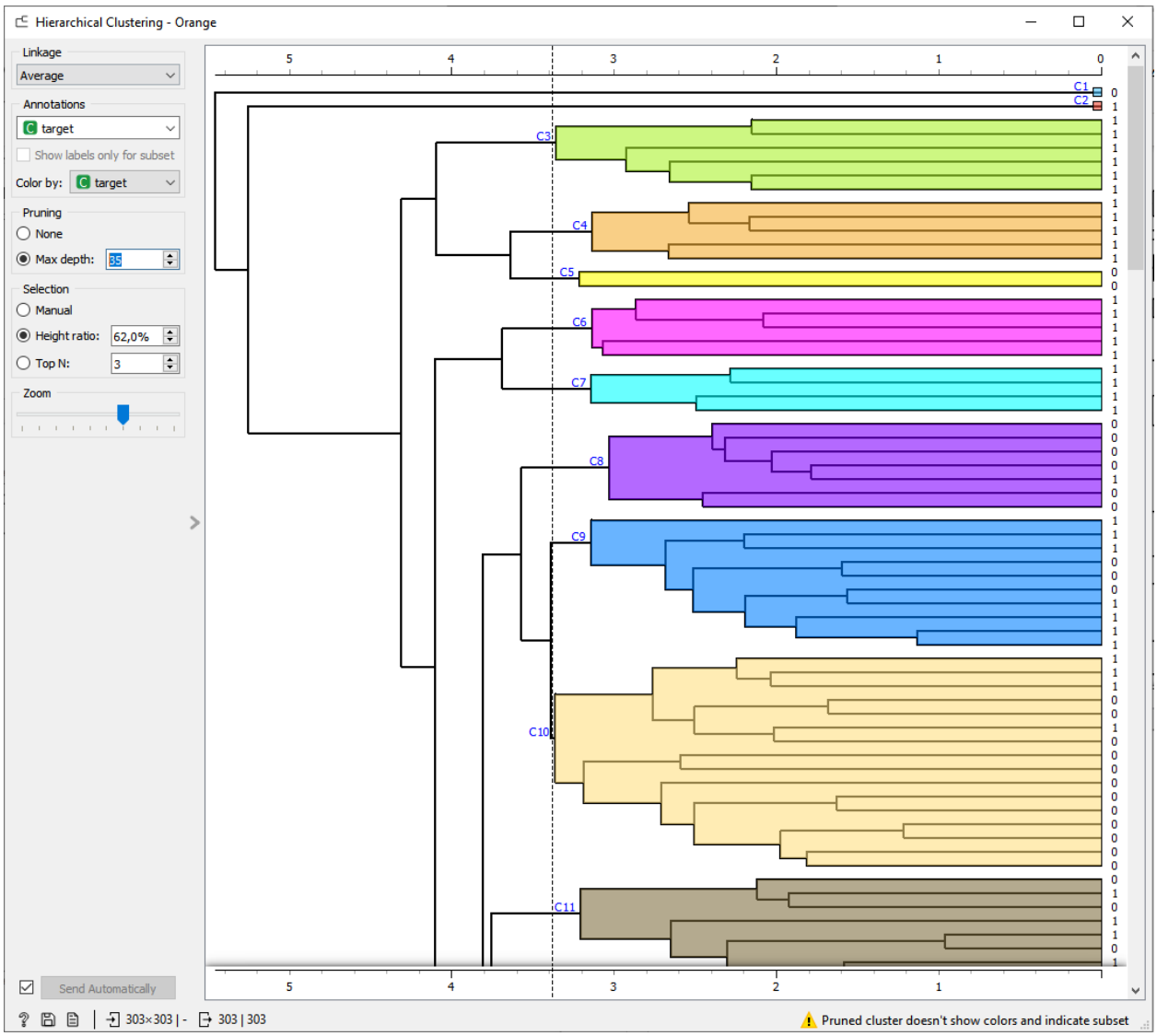
- Annotations
  - klastera elementu anotāciju veids.
- Pruning
  - Dendrogrammas dziļuma izvēle. Tas ietekmē tikai kartēšanu, nevis faktisko grupēšanu.
- Selection
  - Manual (Klastera atlase ar peles klikšķi.)
  - Height ratio (Kopu dalīšana pēc to augstuma..)
  - Top N (Izvēlas augšējo vienību skaitu.)

(5)



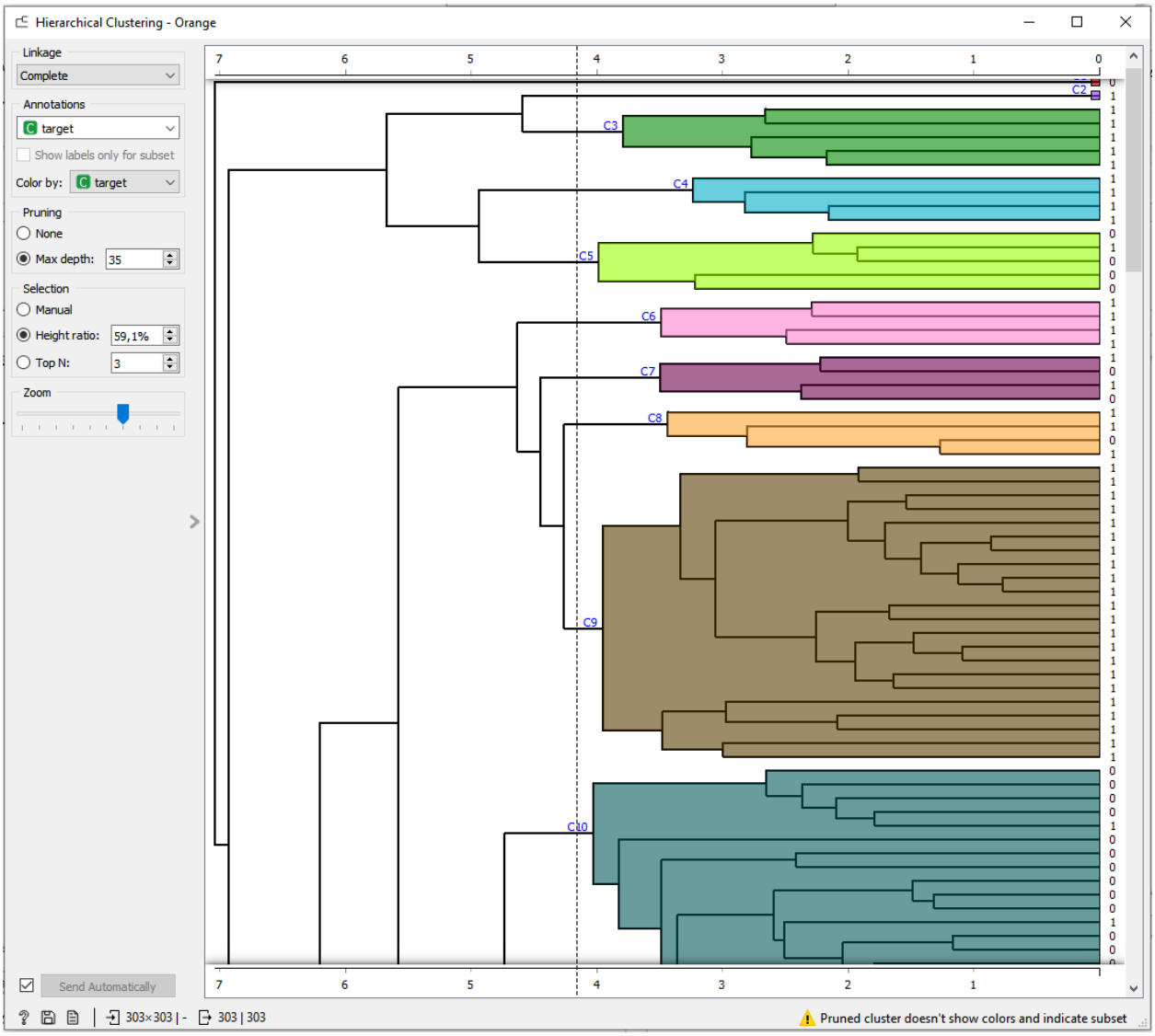
2.2 .att. Hierarchical Clustering Single

Attēlā 2.2 .att. dendogramma, kurā "Linkage" ir izvēlēts kā "Single", "Max depth" ir "35", "Height ratio" ir 72,5%. Šī konfigurācija dod 26 klasterus, no kuriem ir 1 liels klasteris un 5 mazi klasteri. Lielais klasteris (C26) nav ļoti labs, jo tajā ir daudz jauktu vērtību. Un mazajos klasteros (C3, C22, C23, C25) jau ir vienas vērtības vai gandrīz vienas vērtības klasteri.



2.3.att. Hierarchical Clustering Average

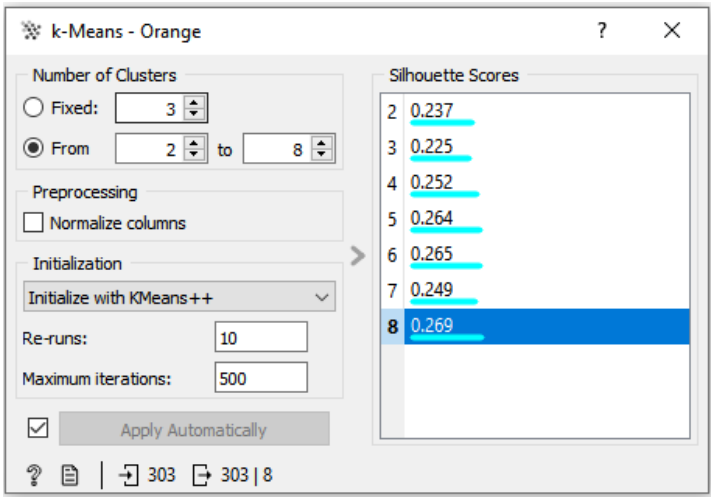
Attēlā 2.3 att. dendogramma, kurā "Linkage" ir izvēlēts kā "Average", "Max depth" ir "35", "Height ratio" ir 56,6%. Šī konfigurācija dod 30 klasterus. Daži no klasteriem, kas labi klasificēja, piemēram, (C3, C4, C5, C6, C7, C8, C9, C30, C27). Daži klasteri apvienojās vienā lielā klasterī ar dažādām vērtībām.



2.4.att. Hierarchical Clustering Complete

Attēlā 2.4. att. dendogramma ar "Linkage" izvēlēts kā "Complete", "Max depth" ir "35", "Height ratio" ir 59,1%. Klasterizācija ar šo konfigurāciju ir vislabākā no visām 3 konfigurācijām. Kopumā ir 23 klasteri. Tajā pašā laikā nav lielu klasteru ar dažādām vērtībām. Diezgan daudz ir vidējo klasteru ar vienu vai gandrīz vienu vērtību, piemēram, (C9, C10, C3, C14, C18, C21).

2.2 K-means algoritms



2.5 att. K-means loga

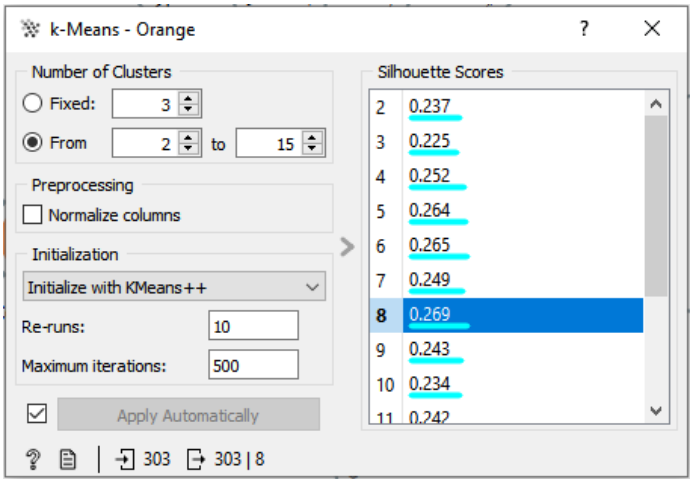
K-means ir tādi hiperparametri kā:

- Number of Cluster
  - Fixed (sagrupē datus vairākos noteiktos klasteros)
  - From N to K (klasteru skaits diapazonā, kuram tiks piemērota klasterizācijas metode)

- Preprocessing
  - Normalize columns (Vidējai vērtībai ir tiekties uz 0, bet standartnovirzei - uz 1.)
- Initialization
  - Re-runs (cik reižu tiek palaists algoritms.)
  - Maximum iterations (maksimālais iterāciju skaits katrā algoritma darbības reizē.)

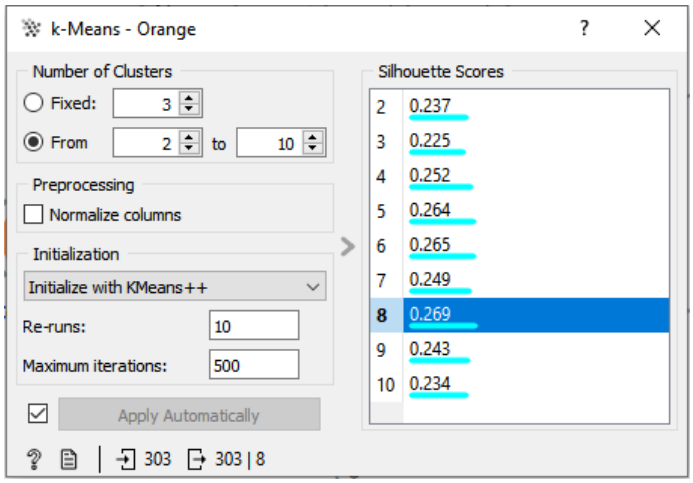
(4)

Mainot klasteru skaitu no fiksēts uz diapazons. Diapazons tika iestatīts uz 15, un redzams, ka 8 klasteris ir labākais.

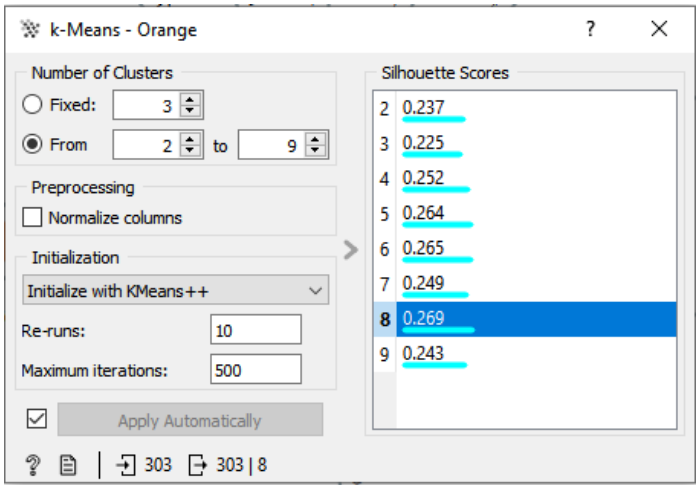


2.6 Att.k-means algoritms ar diapazonu no 2 līdz 15

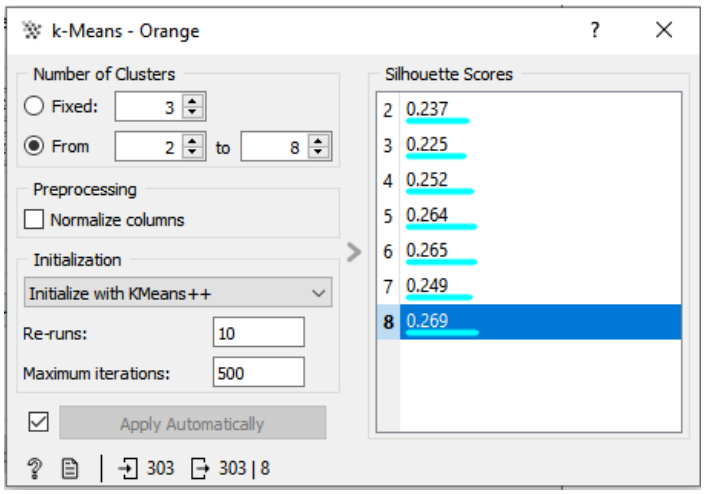
Samazinot diapazonu līdz 10 klasteriem, labākais arī palika 8 klasteri.



2.7 Att. k-means algoritms ar diapazonu no 2 līdz 10

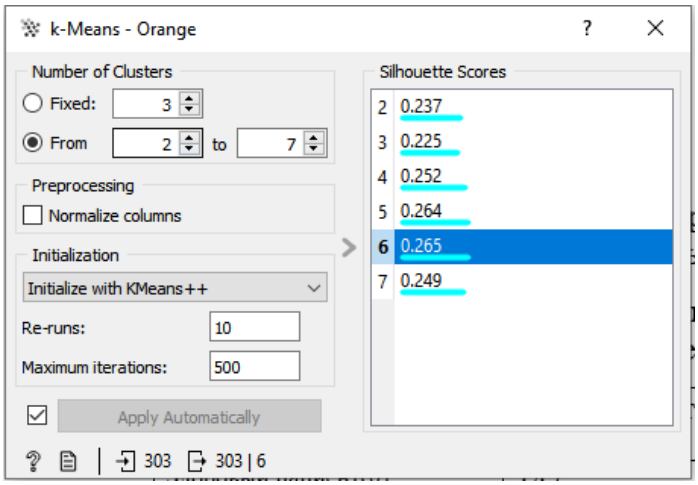


2.8 Att. k-means algoritms ar diapazonu no 2 līdz 9



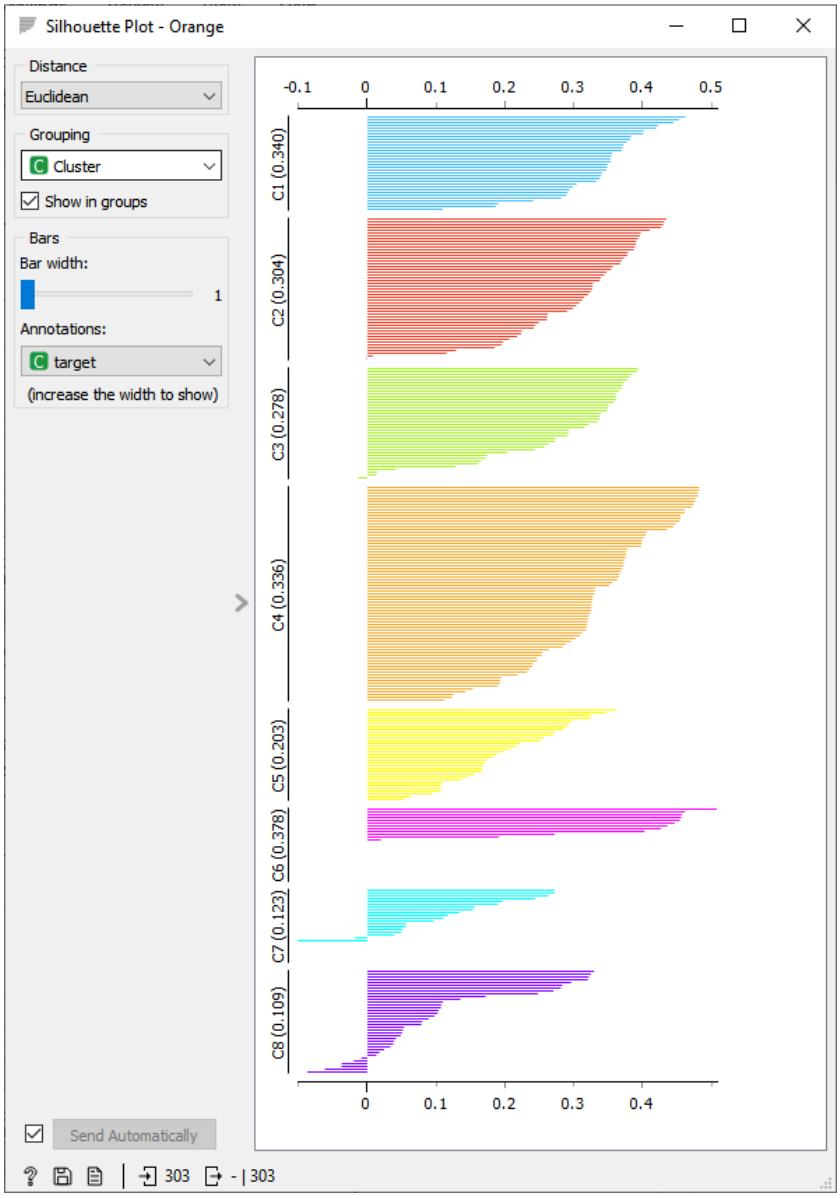
2.9 Att. k-means algoritms ar diapazonu no 2 līdz 8

Visos testos labākais klasteru skaits bija 8. Pēdējā testā tas apstiprinājās, jo, samazinot diapazonu līdz 7 klasteriem, augstākais rezultāts ir 0,265 par 6 klasteriem. Rezultāts 6 klasteriem ir zemāks nekā 8 klasteriem, tāpēc labākais ir diapazons no 2 līdz 8 klasteriem.



2.10 Att. k-means algoritms ar diapazonu no 2 līdz 7

Labākais klasteru diapazons ir no 2 līdz 8. Šo grafiku parāda silueta plosts. Tas parāda, ka lielākā daļa datu klasteros ir labi atdalīti, jo lielākā daļa ir pozitīvā pusē. No visiem klasteriem vissliktākie ir C8 un C7, jo dažas vērtības ir negatīvas. Aplūkojot "Silhouette Plot", var teikt, ka k-means metode skaidri nodala datu kopas klases.



2.11 att. Sadalījuma vizualizācija, izmantojot k-means viduvēju algoritmu

2.3. Nepārraudzītās mašīnmācīšanās secinājumi

Nepārraudzītās mašīnmācīšanās metodes liecina, ka datu kopa ir labi sadalīta klasēs, izmantojot klasterizāciju. Hierarhiskā klasterizācija, salīdzinot ar k-means klasterizāciju, nedarbojas labi, jo klasteri ir sagrupēti lielos klasteros ar atšķirīgām klasēm. Šajā ziņā k-means klasteru algoritms darbojas labāk. Tādējādi datu klasificēšanai var izmantot neuzraudzītus mašīnmācīšanās algoritmus.

III daļa – Pārraudzītā mašīnmācīšanās

Datu apstrādei, izmantojot " pārraudzītu mācīšanos", tika izvēlēti 3 algoritmi:

1) algoritms kNN
2) algoritms Naive Bayes
3) Neural Network

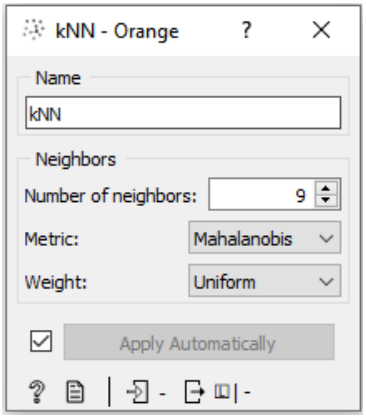
Lai izmantotu un pārbaudītu algoritmus, dati jāsadala divās daļās: mācību dati un dati testēšanai.

Tika izvēlēta 90 % un 10 % attiecība - 90 % algoritma apmācībai un 10 % testēšanai, un šī vērtība tika “Data Sampler”

Klases nosaukums	Mācību dati (90%)	Testēšanas dati (10%)
Vesels pacients (0)	147	17
Pacients ar sirds slimību (1)	126	13
	273	30

3.1 tabula Datu kopas sadalīšana apmācības kopā un testēšanas kopā

3.1 kNN Algoritms



3.1 att. kNN Algoritma logs

Šis ir klasifikācijas un regresijas algoritms. Algoritma pamatā ir tuvākā kaimiņa princips. Princips ir objektam piešķirt kategoriju vai klasi. Balstoties uz mācību datiem, algoritms atrod vislīdzīgāko kaimiņu un, pamatojoties uz tā vērtībām, jau piešķir kategoriju vai klasi jaunajam objektam. Šis algoritms pieder slinkās mācīšanās (lazy learning) kategorijai, jo tas vienkārši apgūst mācību datus. Šis algoritms tika izvēlēts tā vienkāršības un ātruma dēļ:

Hiperparametri kNN algoritmā:

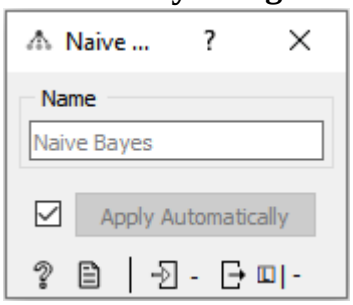
- Number of Neighbors: Šis parametrs nosaka tuvāko kaimiņu skaitu, kas piedalīsies jauno datu klasifikācijā vai regresijā.

- Metric: Attāluma parametrs: Orange vidē ir vispopulārākās metrikas, piemēram.

    - Euclidean (attālums starp diviem objektiem)
    - Manhattan (Atribūtu atšķirību summa)
    - Chebyshev/Maximal (Maksimālā starpības summa)
    - Mahalanobis (Attālums starp punktiem, ņemot vērā korelāciju un to dispersiju)
- Weight:
    - Uniform(Vienāds svars visiem objektiem)
    - Distance(Tuvākie kaimiņi vairāk ietekmē jaunos datus)

(4)

3.2 Naive Bayes Algoritms



3.2 Att. Naive Bayes Algoritms logs

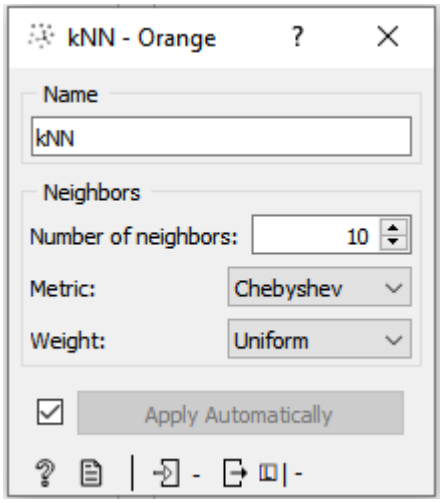
Naive Bayes – Algoritms veic naivu datu analīzi, izmantojot Bejasa statistikas principus. Šis algoritms darbojas tikai klasifikācijas uzdevumiem. Šis algoritms tika izvēlēts, jo tam nav papildu hiperparametru. (3)

3.3 Algoritmu testēšana

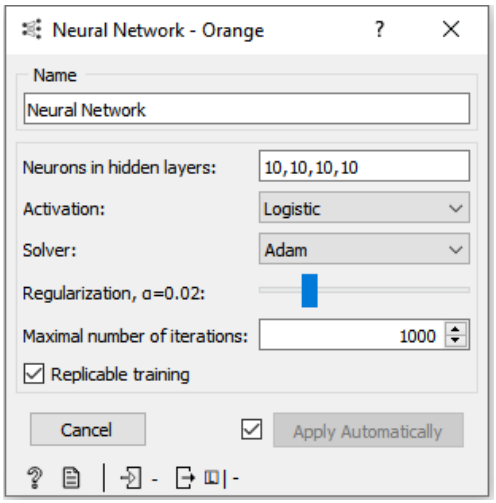
Pirmais tests

Tā kā Naïve Bayes nav hiperparametru, tie nav norādīti algoritma parametros, jo tos nekādā veidā nevar mainīt.

Parametri:



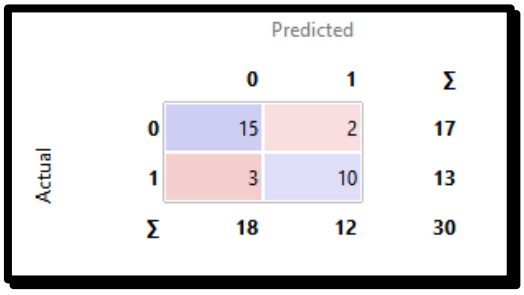
3.3 att. kNN Parametri



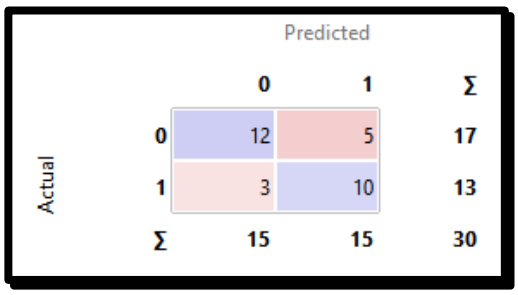
3.4 att Neural Network parametri

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.846	0.833	0.832	0.833	0.833
kNN	0.833	0.767	0.767	0.770	0.767
Neural Network	0.733	0.567	0.410	0.321	0.567

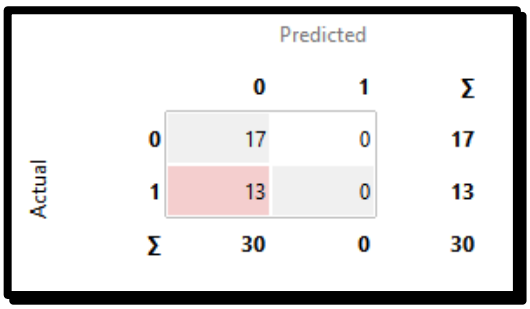
3.5.att. Test and Score pirmais tests



3.6 att. Confusion Matrix Naïve Bayes

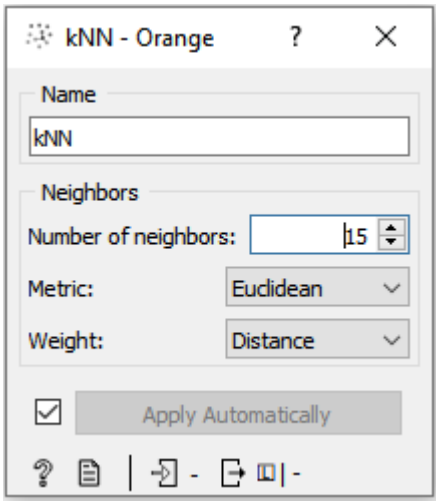


3.7 att Confusion Matrix kNN

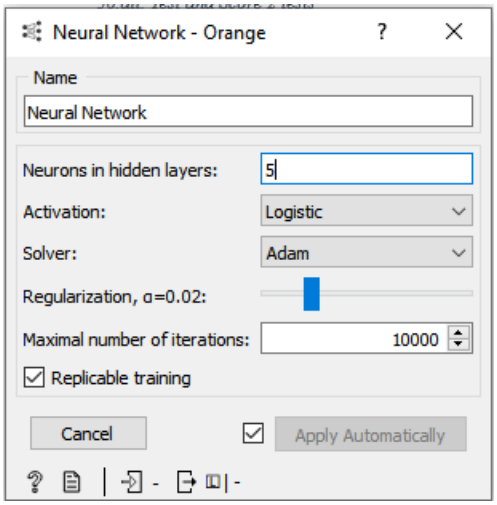


3.8 att Confusion Matrix Neural Network

Otrais tests
Parametri:



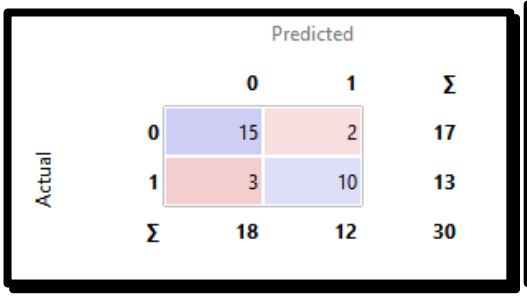
3.9.att kNN Parametri



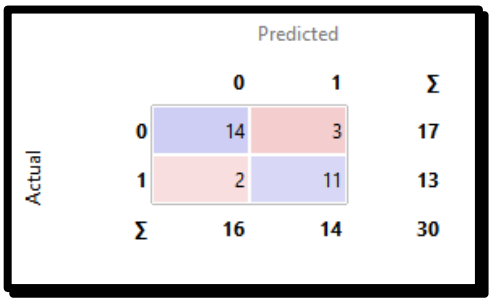
3.10 att Neural Network parametri

Model	AUC	CA	F1	Precision	Recall
kNN	0.853	0.867	0.867	0.867	0.867
Naive Bayes	0.846	0.833	0.832	0.833	0.833
Neural Network	0.733	0.567	0.410	0.321	0.567

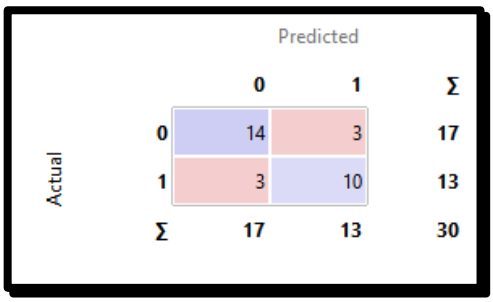
3.11 att. Test and Score Otrais tests



3.12 att. Confusion Matrix Naïve Bayes

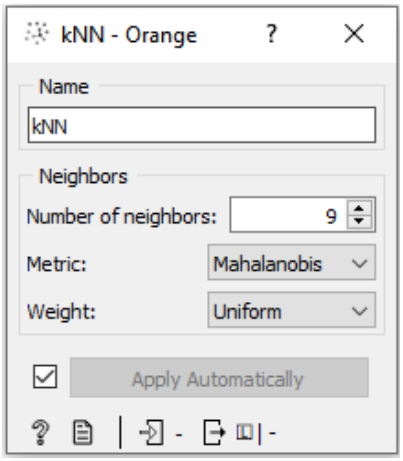


3.13 att. Confusion Matrix kNN

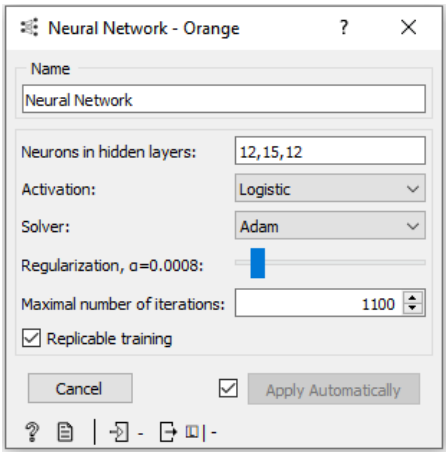


3.14. att Confusion Matrix Neural Network

Tresais tests
Parametri:



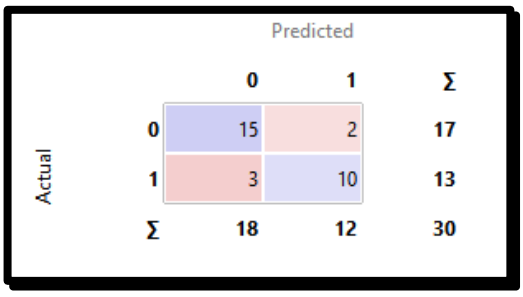
3.15.att kNN Parametri



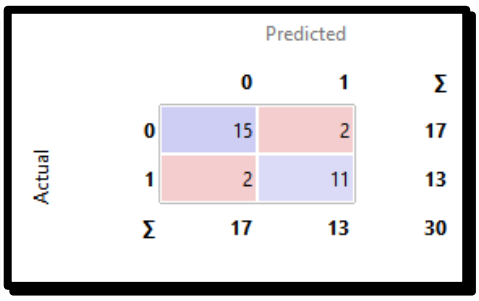
3.16.att. Neural Network parametri

Model	AUC	CA	F1	Precision	Recall
kNN	0.830	0.867	0.867	0.867	0.867
Naive Bayes	0.846	0.833	0.832	0.833	0.833
Neural Network	0.733	0.567	0.410	0.321	0.567

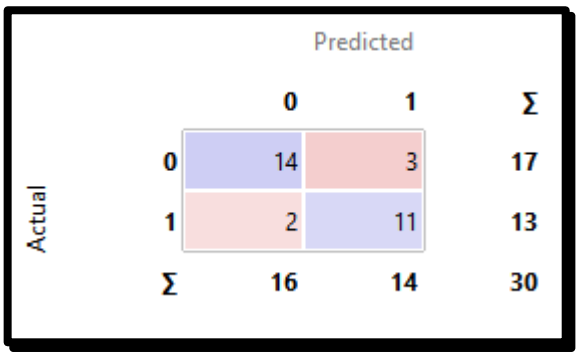
3.17.att. Test and Score Tresais tests



3.18 att. Confusion Matrix Naïve Bayes



3.19 att. Confusion Matrix kNN



3.20 att. Confusion Matrix Neural Network

3.4 Rezultātu apkopošana un salīdzināšana

Testu mērķis bija maksimāli palielināt klasifikācijas algoritmu precizitāti. Testu gaitā modeļu precizitāte palielinājās un trešajā testā sasniedza salīdzinoši labus rezultātus. Trešajā testā kNN algoritms tika atzīts par vislabāko ar 86,7 % precizitāti. Kā izrādījās, šī konfigurācija kNN ("Number of Neighbors " - 9, " Metric" - Mahalanobis, " Weight " - Uniform) ir vislabākā izvēlētajam datu kopumam salīdzinājumā ar pirmo testu, kurā tika izmantota konfigurācija ("Number of Neighbors" - 10, " Metric" - Chebersheev, " Weight" - Uniform). Rezultātā precizitāte palielinājās par 9,7 %. "Naïve Bayes" algoritma gadījumā precizitāte nekādā veidā nemainās, jo nav papildu parametru, kurus varētu mainīt. Ņemot vērā iepriekš minēto, šā algoritma precizitāte ir 83,3 %, kas ir pietiekami laba. Ņemot vērā, ka tas ir viens no vienkāršākajiem algoritmiem. Neironu tīkls izrādījās diezgan muļķīgs rīks rezultātu prognozēšanai. Darba gaitā izrādījās, ka iterāciju skaits, īsti neietekmē galīgo precizitāti, to var redzēt, aplūkojot 2 testu un 3 testu rezultātus. Pievienojot 4. slēpto slāni, precizitāte pilnībā samazinās līdz 32,1 %, kā redzams pirmajā testā. To zinot, 3. testā tika atrasta konfigurācija, kas nodrošina 83,6 % precizitāti. Neironu tīkla precizitāti visvairāk ietekmēja slēpto slāņu skaits un mezglu skaits tajos.

Pamatojoties uz pēdējā testā iegūtajiem Confusion Matrix datiem. Pareizas klasifikācijas iespēja, izmantojot algoritmus:

kNN ir 86,4%.

Naïve Bayes algoritma vērtība ir 83,3 %.

Neural Network - 83,0%.

SECINĀJUMI

Praktiskajā darbā mums bija jāatrod datu bāze no pieejamajiem avotiem. Un veikt analīzi, izmantojot esošos Orange logrīkus. Tika izmantotas divas neuzraudzītas un uzraudzītas mašīnmācīšanās metodes. Neuzraudzītajai mašīnmācībai tika izmantoti Hierarchical Clustering un K-Means algoritmi, klasterizācija ļauj ērti un detalizēti apskatīt attiecības starp objektiem. Tas var būt noderīgs datu segmentēšanai, modeļu noteikšanai un slēpto attiecību identificēšanai. Strādājot ar uzraudzītu mašīnmācīšanos, tika izmantoti kNN, Naïve Bayes un Neural Network algoritmi. Noslēdzot šo Orange praktisko darbu, var izdarīt šādus secinājumus. Datu analīze un apstrāde ir ārkārtīgi laikietilpīga un sarežģīta ar visām no tā izrietošajām sekām. Ir daudz pazīmju un atribūtu, kas ietekmē galīgo rezultātu. Orange izmantošana šajā darbā ļāva ērti vizualizēt datus, piemērot dažādus mašīnmācīšanās algoritmus un novērtēt to efektivitāti. Kopumā darbs man patika, jo tas iepazīstina mūs ar daudzām jaunām datu analīzes, apstrādes un atlases iespējām.

Izmantotie avoti

1. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
2. https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland
3. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/naivebayes.html
4. https://orange3.readthedocs.io/en/3.5.0/widgets/unsupervised/kmeansclustering.html
5. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html
6. https://www.youtube.com/watch?v=dKURyzjh5Gc
7. https://www.youtube.com/watch?v=bmwH3EcTBEM
8. https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html
9. https://creativecommons.org/licenses/by-sa/4.0/