Datorzinātnes un informācijas tehnoloģijas fakultāte

Informācijas tehnoloģiju institūts

2. praktiskais darbs mācību priekšmetā

„Mākslīga intelekta pamati”

Mašīnmācīšanās algoritmu lietojums

https://github.com/HenryBroben/PraktiskaisDarbs2

Izpildīja: Vladislavs Ļebedevs

181RMC218

RDB09

2022./23. māc. g.

Satura rādītājs

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 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 data set) 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 jāsatur klašu iezīmes;

• studentiem ir jāizvairās no datu kopām, kurās ir daudz Būla tipa (patiess/nepatiess, 1/0 utt.) vai kategoriskā tipa pazīmju (atribūtu) vērtību. Ir vēlams izmantot datu kopas, kurās lielākā daļa no pazīmēm ir atspoguļota ar nepārtrauktām pazīmju vērtībām;

• studentiem ir jāizvairās no datu kopām, kurās klašu iezīmes nav dotas (piemēram, teksta korpusiem un neapstrādātiem attēliem)

I daļa - Datu pirmapstrāde/izpēte

Lai izpildītu šī darba daļu, studentiem ir jāveic šādas darbības:

1. Ir jāizvēlas un jāapraksta datu kopa, pamatojoties uz informāciju, kas sniegta krātuvē, kurā datu kopa ir pieejama.

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

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

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

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

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

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.1. daļa

Datu kopas apraksts:

Tika izvēlēta datu kopa, kurā tiek analizēti ideāla pircēja parametri. Tajā ir 8 kritēriji un 2000 ieraksti, tomēr tā kā ierakstu skaits ir diezgan liels, tika pieņemts lēmums samazināt ierakstu skaitu līdz 400, lai analīze strādātu ātrāk un diagrammas paradītu labāko rezultātu.

Autora apraksts:

Shop Customer data ir detalizēta radoša veikala ideālo klientu analīze. Tas palīdz uzņēmumam labāk izprast savus klientus. Veikala īpašnieks iegūst informāciju par Klientiem, izmantojot klientu biedra kartes. Datu kopa sastāv no 2000 ierakstiem un 8 kolonnām:

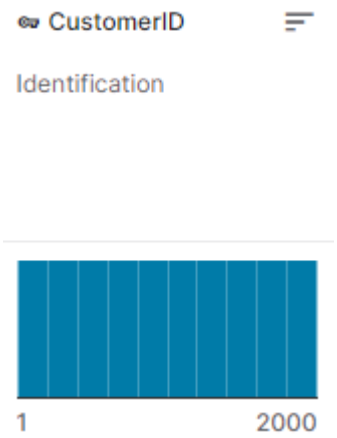
- Klienta ID
- Dzimums
- Vecums
- Gada ienākumi
- Patēriņa rādītājs — veikala piešķirtais rādītājs, pamatojoties uz klientu uzvedību un patēriņu raksturu
- Profesija
- Darba pieredze - gados
- Ģimenes lielums

Datu kopas nosaukums: Shop Customer Data

Autors: Data Scientist Anna

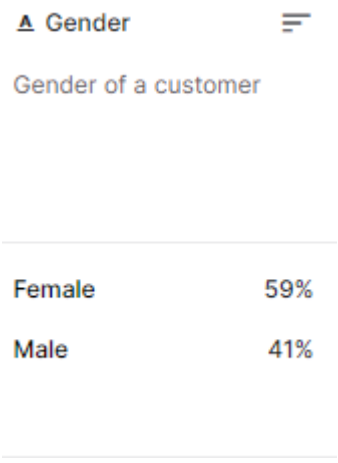
Avots: https://www.kaggle.com/datasets/datascientistanna/customers-dataset

Kritēriji:



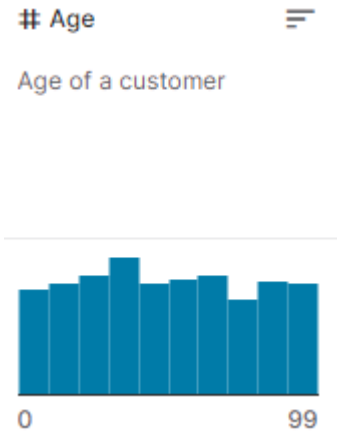
Ilustrācija 1

Klienta identifikācijas numurs (Turpmāk tiek neņemts vērā (***SKIP***), jo nav būtiskas nozīmes pētīšanai)



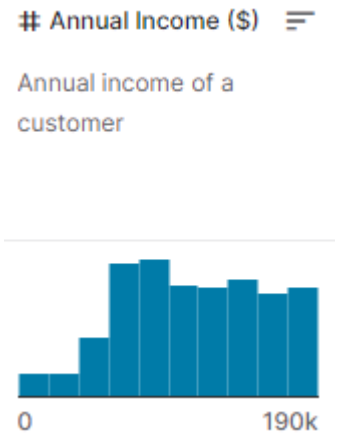
Ilustrācija 2

Klienta dzimums



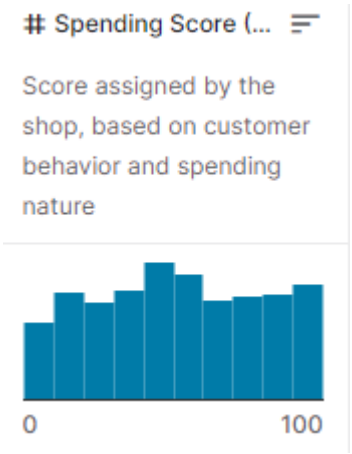
Ilustrācija 3

Klienta vecums



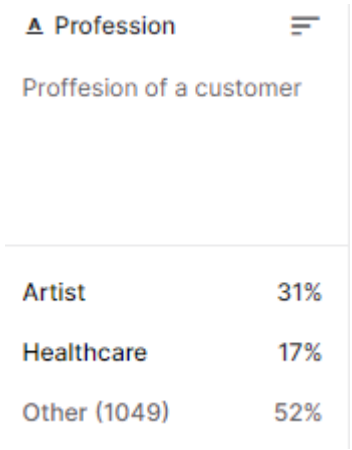
Ilustrācija 4

Klienta gada ienākumi



Ilustrācija 5

Klienta patēriņu rādītājs, kurš ir balstīts uz klienta uzvedības formas

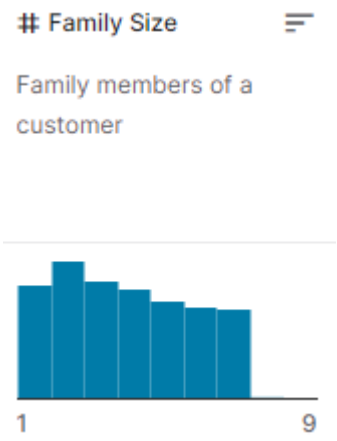


Ilustrācija 6

Klienta profesija

Ilustrācija 7

Darba pieredze



Ilustrācija 8

Klienta ģimenes locekļu skaits

File - Orange

Source

File: Customers.csv

URL:

File Type

Automatically detect type

Info

400 instances
8 features (0.1% missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	CustomerID	numeric	**skip**	
2	**Gender**	**categorical**	**target**	**Female, Male**
3	Age	numeric	feature	
4	Annual Income ($)	numeric	feature	
5	Spending Score (1-100)	numeric	feature	
6	Profession	categorical	feature	Artist, Doctor, Engineer, Entertainment, Executive, Healthcare, Homemaker, Lawyer, Marketing
7	Work Experience	numeric	feature	
8	Family Size	numeric	feature	

Ilustrācija 9

Par ***target*** lomu tika izvēlēta klienta dzimums; Klienta identifikācijas numurs tika izlaists(***skip***), jo tai nav nozīmes; Citi kritēriji tika atstāti par ***feature***

I.2. daļa

No krātuves iegūtā datu kopa ir formātā, ar kuru ir viegli strādāt, jo informācijas avots un autors piedāvāja lejupielādēt **.CSV** formātā. Datu kopa tika analizēta ar ***Orange*** rīka palīdzību.

Data Table - Orange

Info
400 instances
6 features (0.2 % missing data)
Target with 2 values
No meta attributes.

Variables
- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

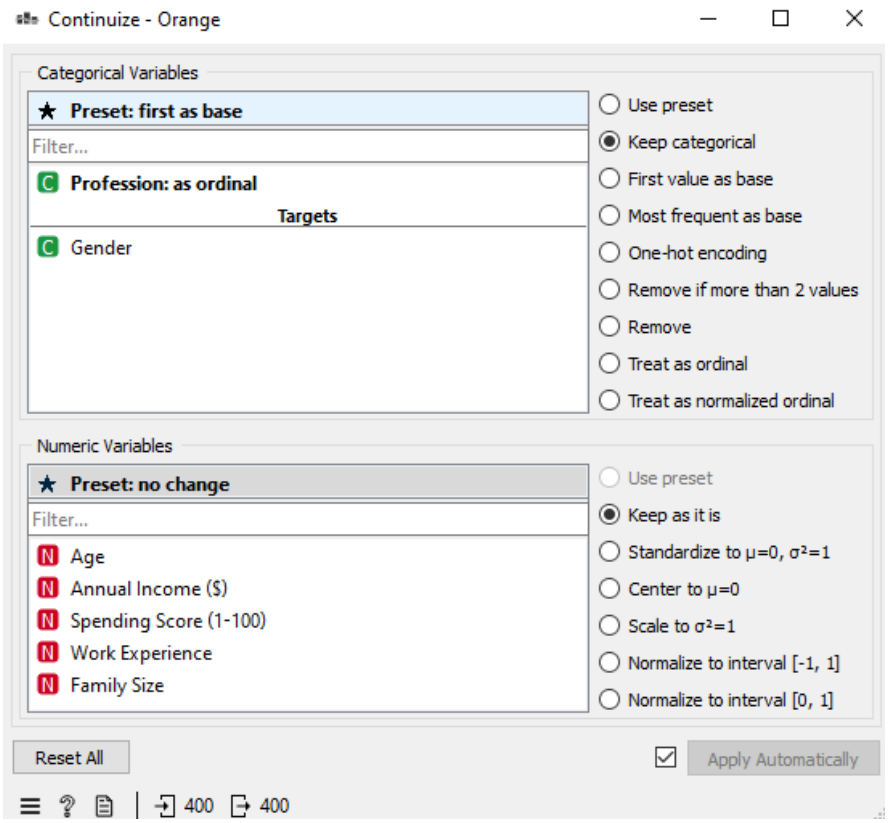
Selection
- [x] Select full rows

	Gender	Age	Annual Income ($)	ending Score (1-10	Profession	Work Experience	Family Size
1	Male	19	15000	39	Healthcare	1	4
2	Male	21	35000	81	Engineer	3	3
3	Female	20	86000	6	Engineer	1	1
4	Female	23	59000	77	Lawyer	0	2
5	Female	31	38000	40	Entertainment	2	6
6	Female	22	58000	76	Artist	0	2
7	Female	35	31000	6	Healthcare	1	3
8	Female	23	84000	94	Healthcare	1	3
9	Male	64	97000	3	Engineer	0	3
10	Female	30	98000	72	Artist	1	4
11	Male	67	7000	14	Engineer	1	3
12	Female	35	93000	99	Healthcare	4	4
13	Female	58	80000	15	Executive	0	5
14	Female	24	91000	77	Lawyer	1	1
15	Male	37	19000	13	Doctor	0	1
16	Male	22	51000	79	Healthcare	1	2
17	Female	35	29000	35	Homemaker	9	5
18	Male	20	89000	66	Healthcare	1	6
19	Male	52	20000	29	Entertainment	1	4
20	Female	35	62000	98	Artist	0	1
21	Male	35	96000	35	Homemaker	12	1
22	Male	25	4000	73	Healthcare	3	4
23	Female	46	42000	5	Artist	13	2
24	Male	31	71000	73	Artist	5	2
25	Female	54	67000	14	Executive	1	3
26	Male	29	52000	82	Artist	1	3
27	Female	45	68000	32	Healthcare	9	8
28	Male	35	78000	61	Artist	1	3
29	Female	40	18000	31	Artist	0	1
30	Female	23	20000	87	Artist	5	4
31	Male	60	39000	4	Artist	0	3
32	Female	21	34000	73	Doctor	1	2
33	Male	53	59000	4	Healthcare	1	3
34	Male	18	62000	92	Homemaker	9	7
35	Female	49	91000	14	Lawyer	1	2
36	Female	21	95000	81	Healthcare	3	4
37	Female	42	14000	17	Doctor	5	1
38	Female	30	62000	73	Healthcare	1	5
39	Female	36	9000	26	Artist	8	2
40	Female	20	69000	75	Artist	8	2
41	Female	65	25000	35	Artist	4	1
42	Male	24	85000	92	Healthcare	0	2
43	Male	48	22000	36	Artist	14	3
44	Female	31	33000	61	Artist	1	2
45	Female	49	72000	28	Engineer	8	1

Ilustrācija 10

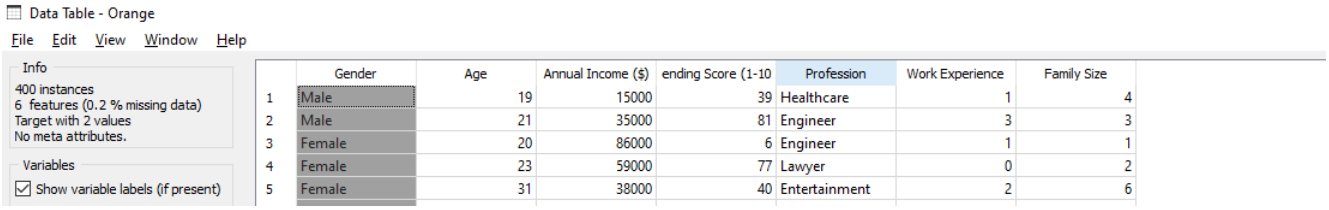
I.3. daļa

Tā kā klienta profesija ir dota tekstveida formā(Male/Female), ar *Continuize* ***Orange*** rīka palīdzību tie tika transformēti skaitliskās vērtības (Male = 1; Female = 0).



Ilustrācija 11

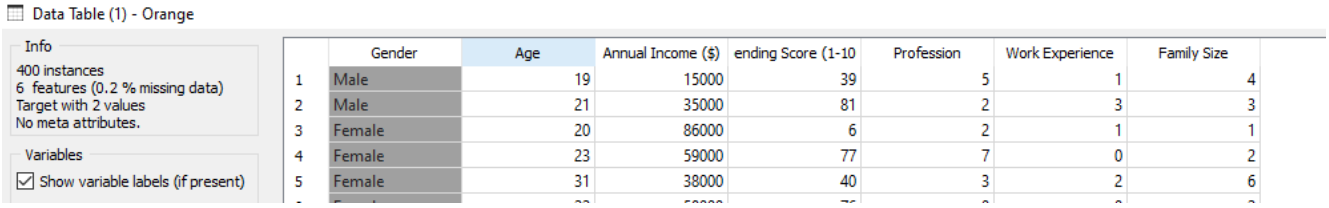
Pirms transformācijas:



	Gender	Age	Annual Income ($)	ending Score (1-10	Profession	Work Experience	Family Size
1	Male	19	15000	39	Healthcare	1	4
2	Male	21	35000	81	Engineer	3	3
3	Female	20	86000	6	Engineer	1	1
4	Female	23	59000	77	Lawyer	0	2
5	Female	31	38000	40	Entertainment	2	6

Ilustrācija 12

Pēc transformācijas:

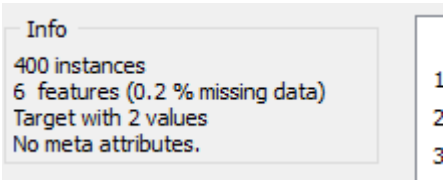


	Gender	Age	Annual Income ($)	ending Score (1-10	Profession	Work Experience	Family Size
1	Male	19	15000	39	5	1	4
2	Male	21	35000	81	2	3	3
3	Female	20	86000	6	2	1	1
4	Female	23	59000	77	7	0	2
5	Female	31	38000	40	3	2	6

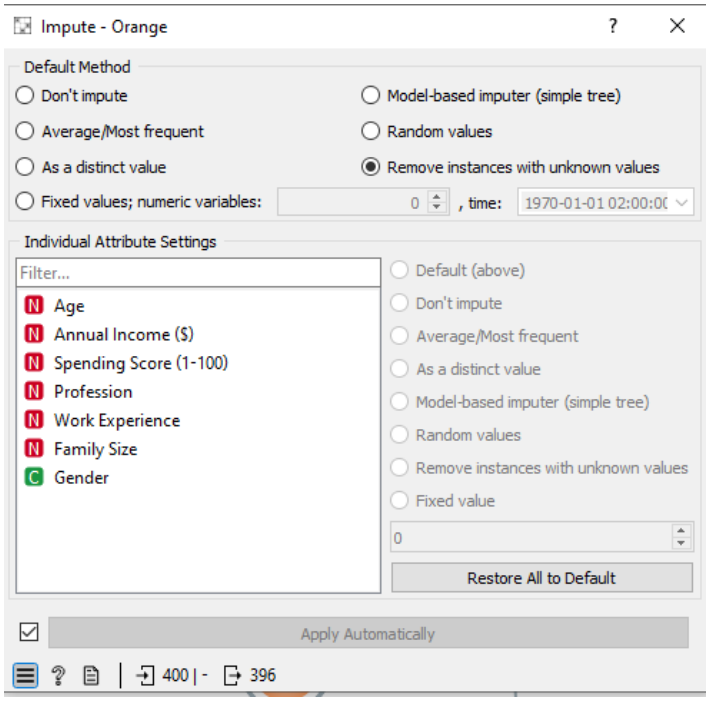
Ilustrācija 13

I.4. daļa

Dažiem objektiem trūkst atsevišķu pazīmju vērtības(0.2 % trūkstošo vērtību), tādēļ izmantojot ***Impute***, es nolēmu izņemt datus, kuras nebija pilnīgi aizpildītas, jo ierakstu skaits ir pietiekami liels (400 ieraksti) un labāko rezultātu dēļ, manuprāt, ir vērts nevis aizpildīt trūkstošās vērtības ar vidējām/visbiežāk sastopamiem, bet tieši otrādi, nodzēst tos.

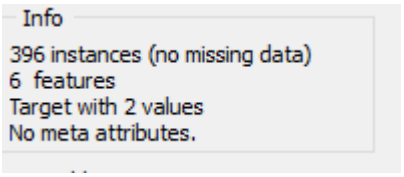


Ilustrācija 14



Ilustrācija 15

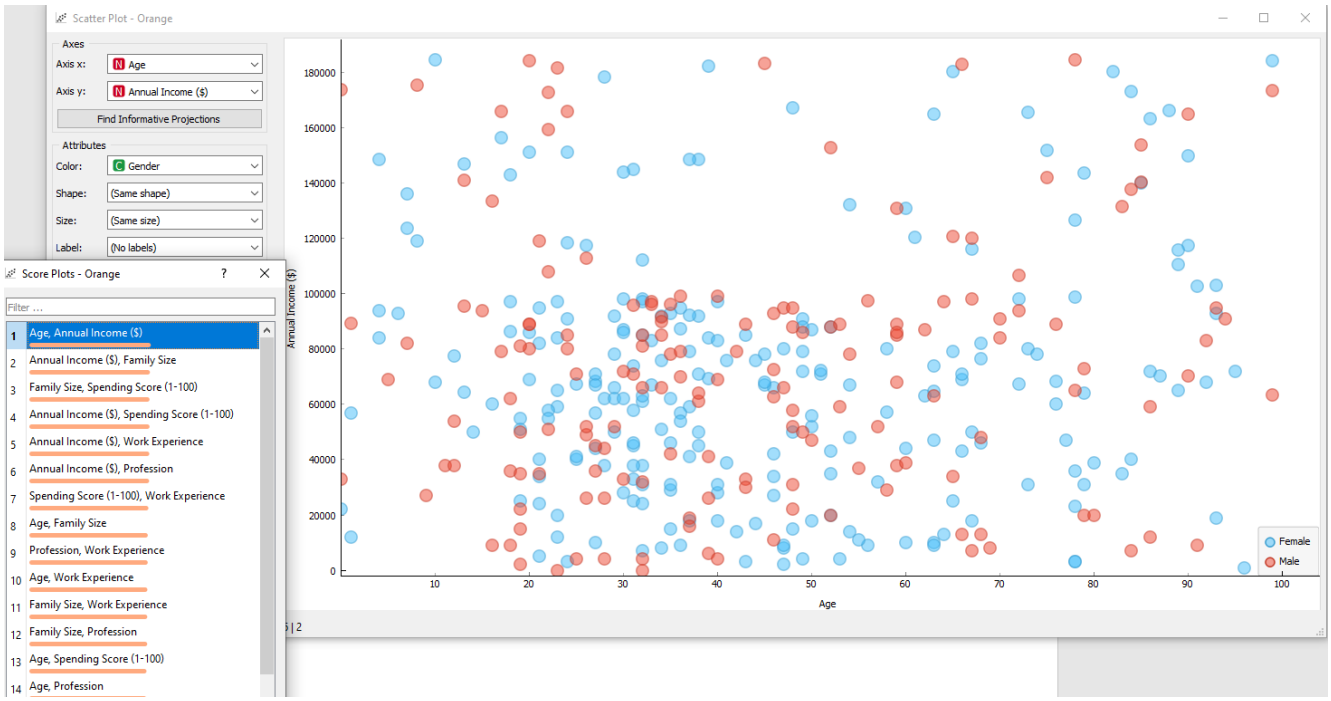
Pēc ***Impute*** instrumenta darbības tika nodzēsti 4 ieraksti:



Ilustrācija 16

I.5. daļa

a)

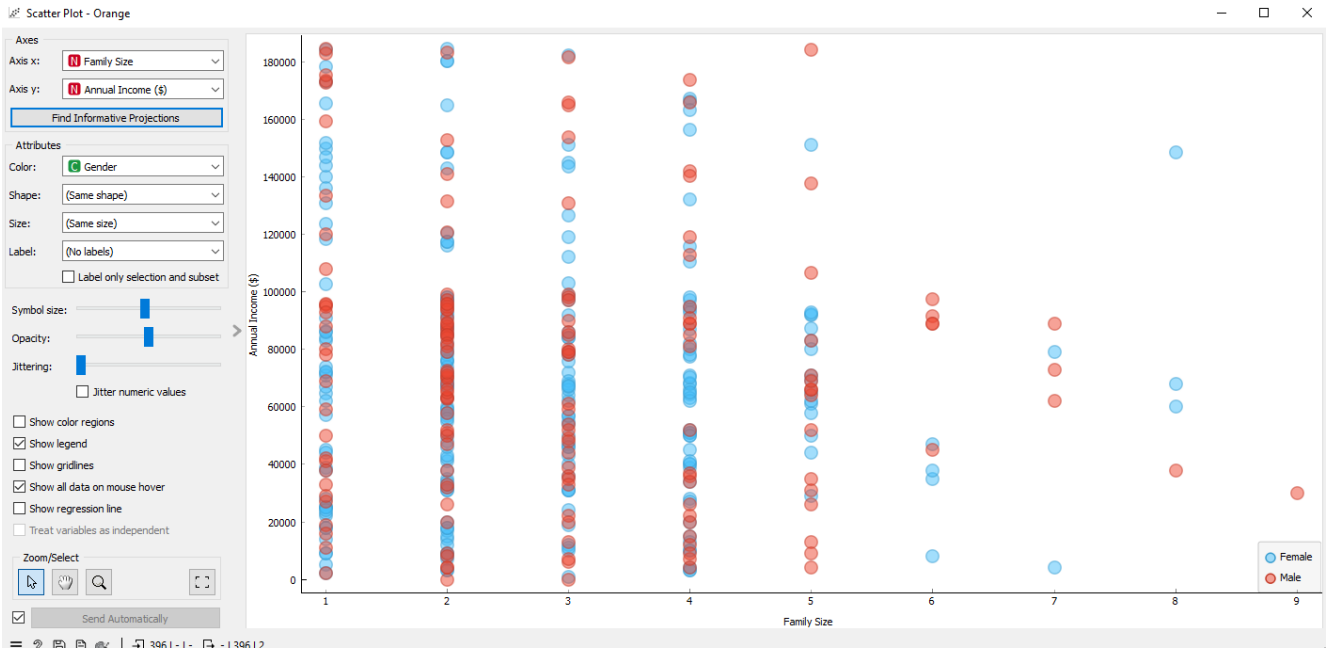


Ilustrācija 17

Tika izmantota ***Find Informative Projections***, līdz ar ko tika atlasītas:

- OX: Age (Vecums)
- OY: Annual Income (Gada ienākumi)
- Kategorija: Gender (Dzimums)

Diagramma attēlo gada ienākumus vīriešiem un sievietēm atkarība no vecuma.

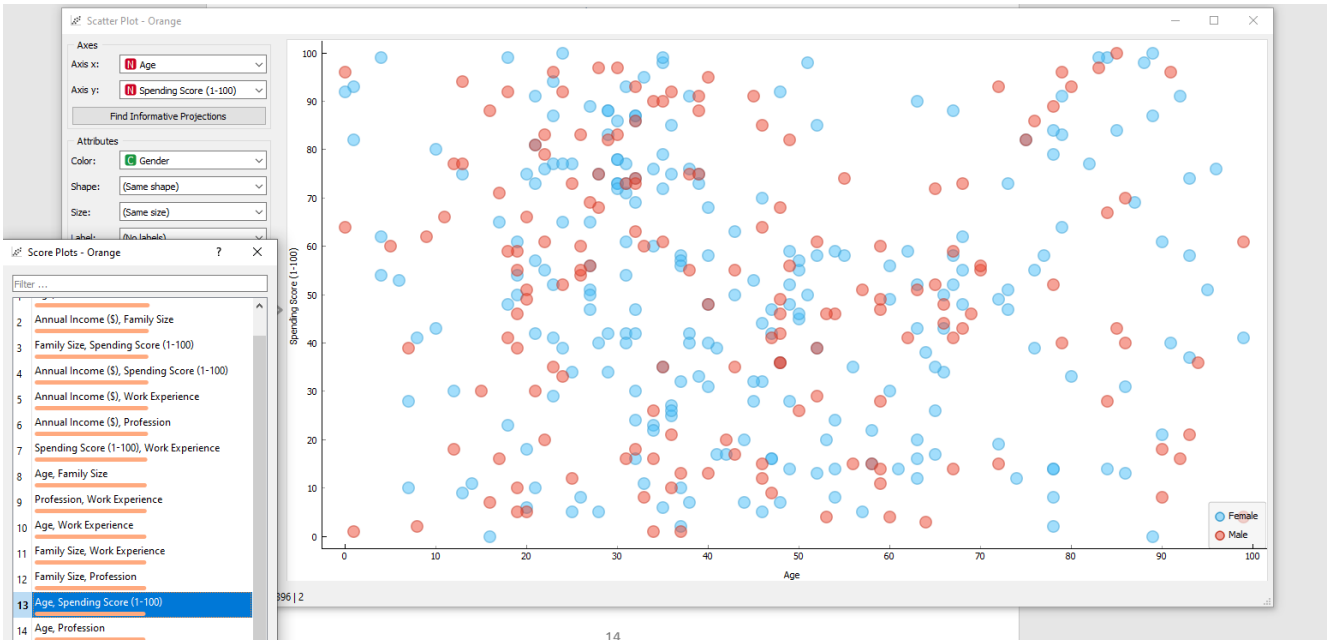


Ilustrācija 18

Tika atlasītas:

- OX: family size (ģimenes locekļu skaits)
- OY: Annual Income (Gada ienākumi)
- Kategorija: Gender (Dzimums)

Diagramma attēlo gada ienākumus vīriešiem un sievietēm atkarība no ģimenes locekļu skaita.



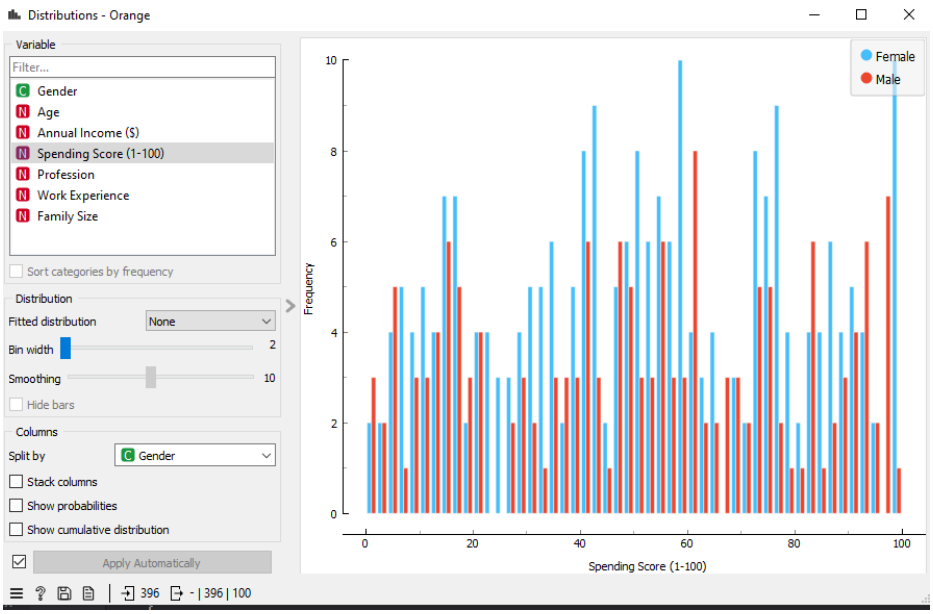
Ilustrācija 19

Tika izmantota ***Find Informative Projections***, līdz ar ko tika atlasītas:

- OX: Age (Vecums)
- OY: Spending Score (Patēriņa rādītājs)
- Kategorija: Gender (Dzimums)

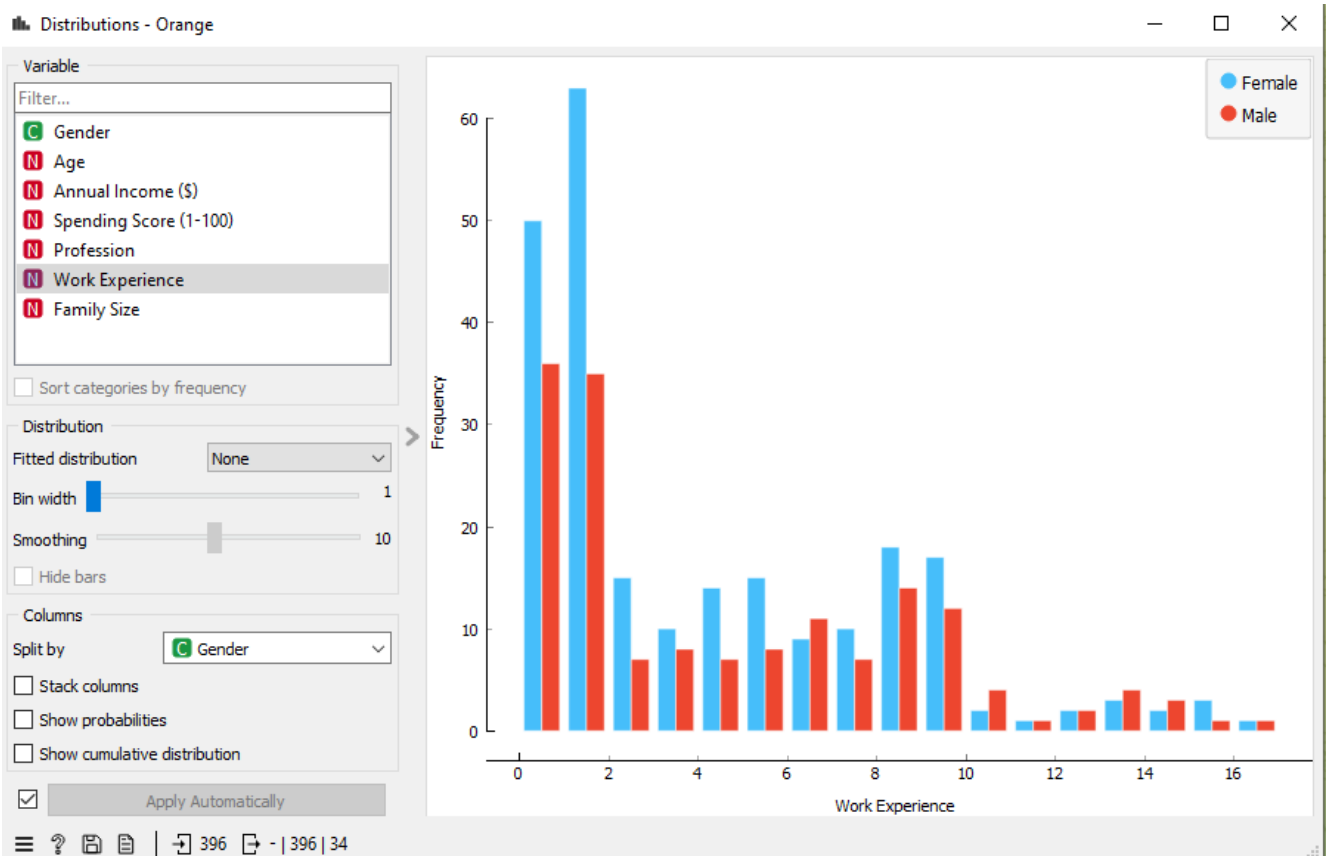
Diagramma attēlo patēriņa rādītāju vīriešiem un sievietēm atkarība no vecuma.

b)



Ilustrācija 20

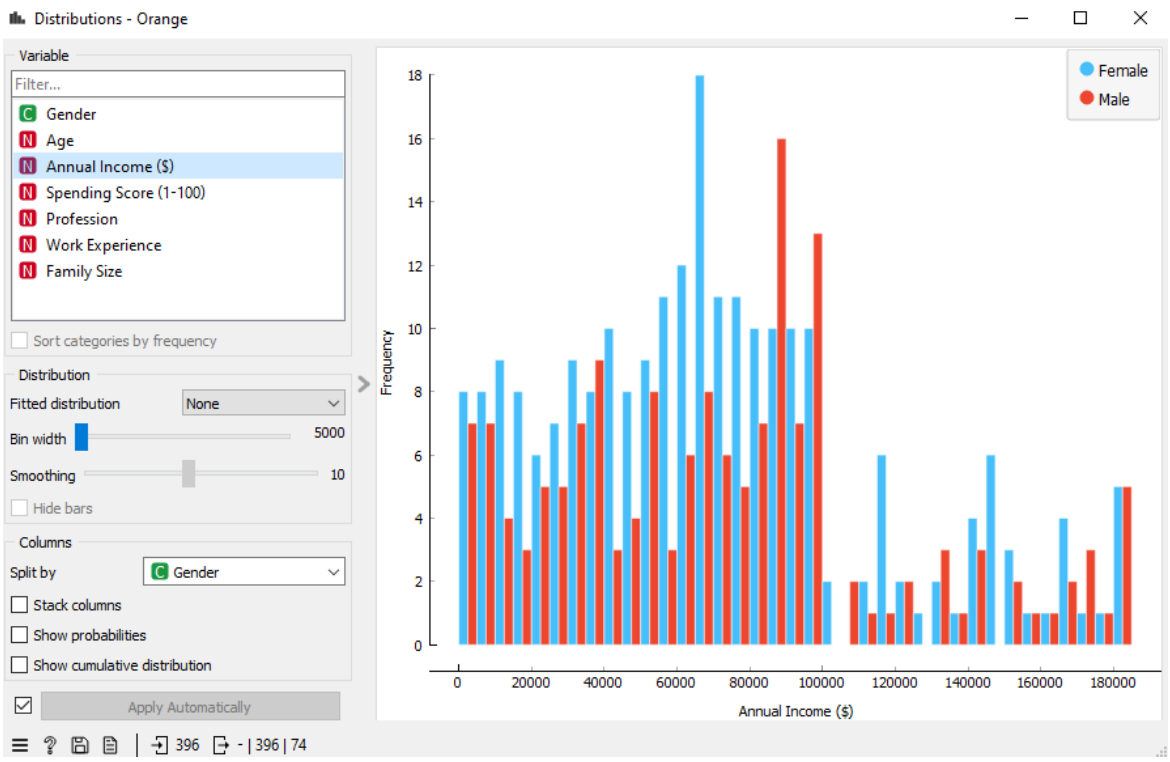
Uz histogrammas tika aplūkots ***Spending Score*** (patēriņa rādītājs) un sadalīts pēc dzimuma. Ir redzama patēriņa rādītāja starpība starp sievietēm un vīriešiem



Ilustrācija 21

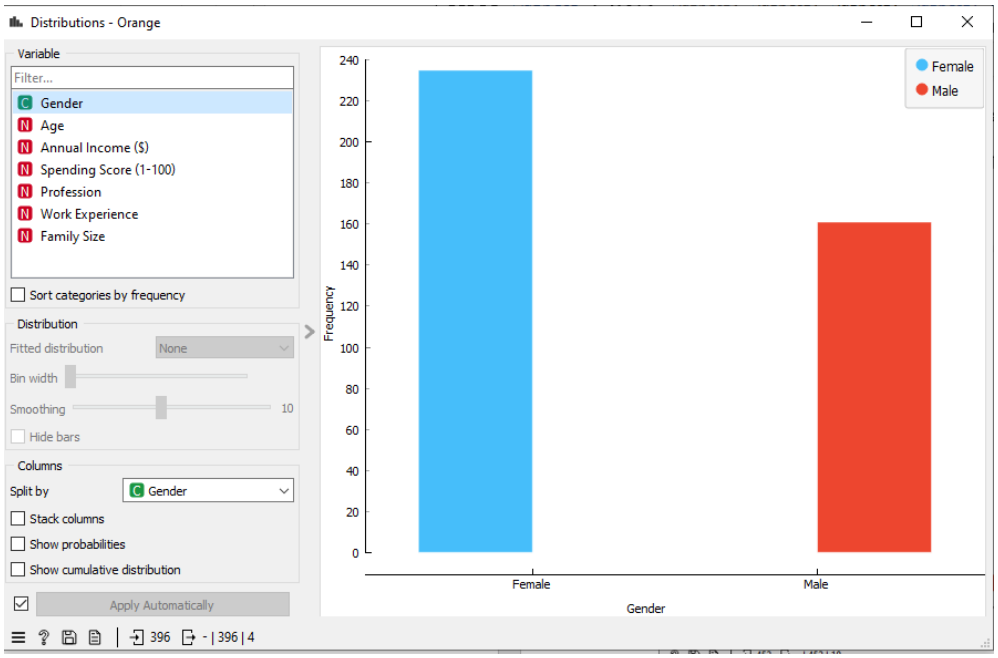
Uz histogrammas tika aplūkots ***Work Experience*** (darba pieredze) un sadalīts pēc dzimuma. Ir redzama darba pieredzes starpība starp sievietēm un vīriešiem.

c)



Ilustrācija 22

Uz histogrammas tika aplūkots ***Annual Income*** (gada ienākumi) un sadalīts pēc dzimuma. Ir redzama gada ienākumu starpība starp sievietēm un vīriešiem.



Ilustrācija 23

Uz histogrammas tika aplūkots ***Gender*** (dzimums) un sadalīts pēc dzimuma. Ir redzams, cik vīrieši un sievietes piedalās datu kopā.

d)

Ilustrācija 24

Feature Statistics sakārtots pēc dispersijas

Ilustrācija 25

Feature Statistics sakārtots pēc mediānas

I. daļas secinājumi

Ņemot vērā histogrammu un diagrammu datus, var secināt, ka radošā veikalā daudz vairāk interesējas cilvēki vecumā līdz 70 gadiem, kas izriet no 19. attēla. Izpētot 18. ilustrāciju, varam secināt, ka veikalu visvairāk interesē pircēji, kuriem ir līdz 5 ģimenes locekļiem, jo tieši pēc šī skaitļa samazinās kopējie gada ienākumi.

Neskatoties uz to, ka eksperimentā piedalījās vairāk sievietes (23.attēls), pēc 22. un 20.attēla viņām ir daudz augstāki rādītāji nekā vīriešiem, proti, patēriņu rādītājs un gada ienākumi. Jo lielāki ir ienākumi – jo vairāk klients var tērēt – jo lielāka interese veikalam ir par viņu.

Būtībā, dati ir skaidri un sagaidāmi, jo sievietes biežāk iepērkas, un vecākiem cilvēkiem vecuma dēļ ir mazāk iespēju iepirkties un apmeklēt veikalus.

II daļa

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.

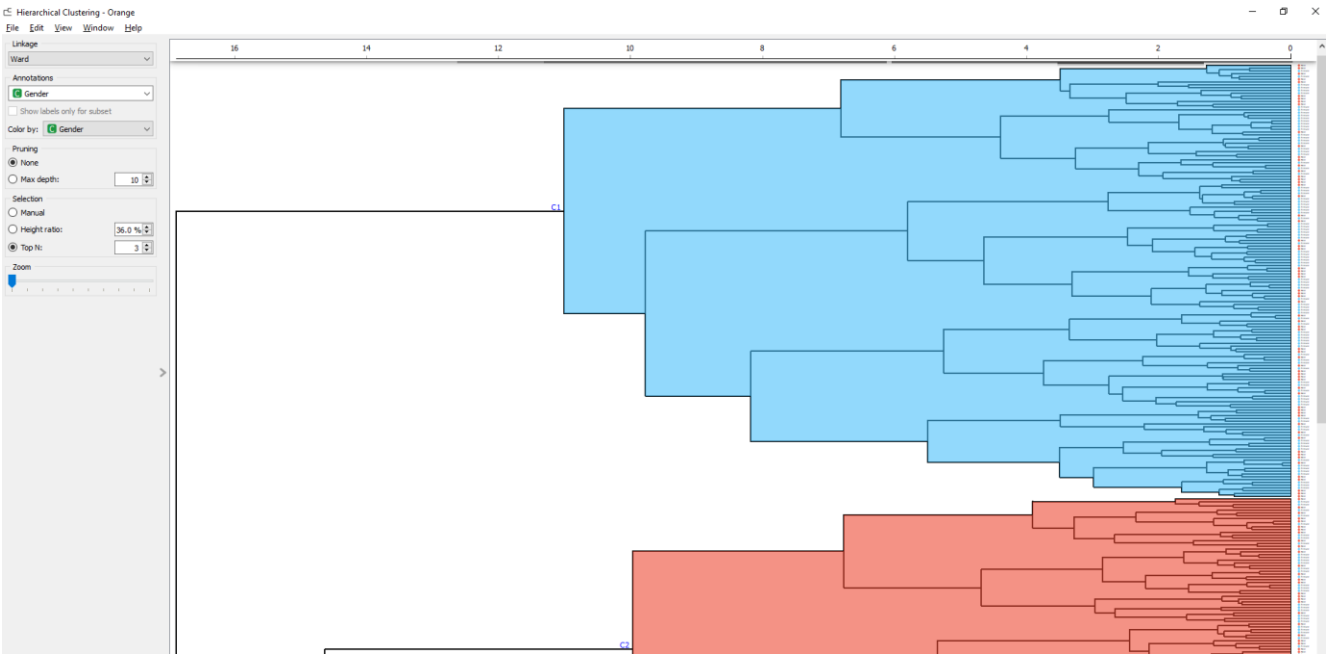
2. Hierarhiskās klasterizācijas algoritmam ir jāveic vismaz 3 eksperimenti, brīvi pārvietojot atdalošo līniju un analizējot, kā mainās klasteru skaits un saturs;

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

II.1. daļa

Tika pielietoti un izpētīti divi studiju kursā apskatītie nepārraudzītās **mašīnmācīšanās algoritmti**:

1)Hierarhiskā klasterizācija



Ilustrācija 26

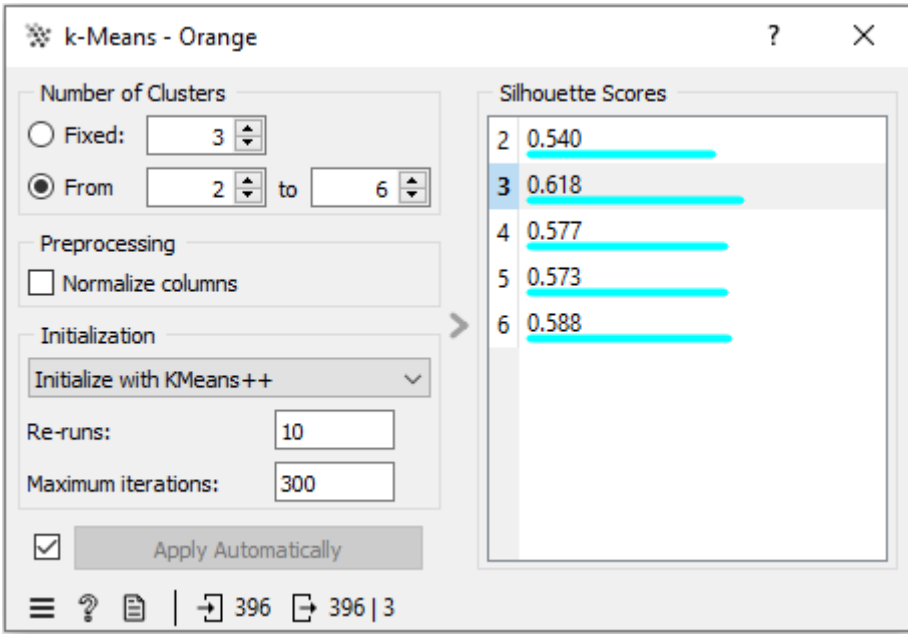
Hiperparametri:

- Linkage: šis parametrs nosaka, kā tiek aprēķināts attālums starp klasteriem. Ir vairākas iespējas:

1. Single linkage (min) metode izmanto vismazāko attālumu starp diviem klasteru punktiem.
2. Complete linkage (max) metode izmanto maksimālo attālumu starp diviem klasteru punktiem.
3. Average linkage metode izmanto vidējo attālumu starp visiem punktiem divos klasteros.
4. Weighted linkage metode izmanto vidējo attālumu, bet ņem vērā klasteru izmērus.
5. Ward linkage metode izmanto summu kvadrātu starpības starp klasteru punktiem.

- Manuālā centru izvēle: šis parametrs ļauj lietotājam izvēlēties sākotnējos centrus, kas tiek izmantoti, lai veiktu klasterizāciju. Tā ir laba metode, ja lietotājs grib patstāvīgi precīzi izvēlēties sākotnējos centrus.
- Augstuma attiecības metode: šī metode izvēlēs minimālo attālumu, kas jāpārkāpj, lai divi punkti būtu divos dažādos klasteros.
- Top N metode: Šajā metodē tiek izvēlēti N punkti, kas ir visattālākie no sākotnējiem centriem, un tiek izmantoti kā sākotnējie centri.

2)K-vidējo algoritms



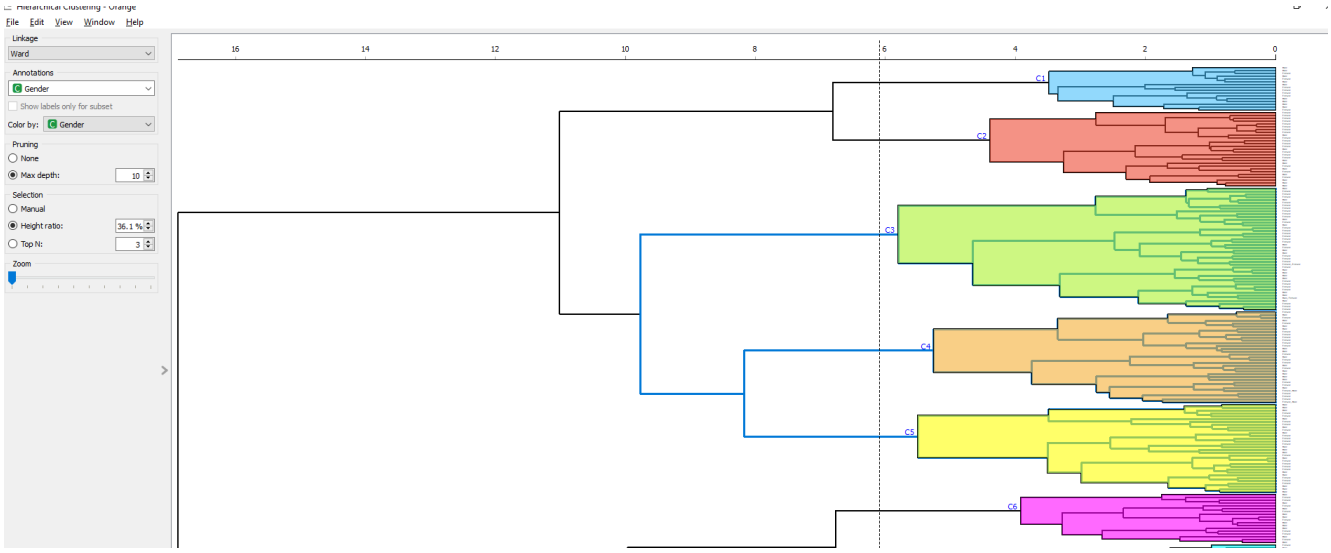
Ilustrācija 27

Hiperparametri:

- Fixed: nosaka fiksēto klasteru skaitu, kas tiks veidots. Tas ir noderīgs, ja lietotājs jau zina, cik klasteru ir nepieciešami.
- From X to Y: nosaka klasteru skaitu diapazonu, no kura algoritms var izvēlēties optimālo klasteru skaitu. Šis ir elastīgs variants, kas ļauj algoritmam pielāgoties datu īpatnībām.
- Preprocessing: nosaka, kā tiek veikta datu apstrāde pirms algoritma darbības. Var veikt normalizāciju, tomēr es to neveicu, jo jau biju veicis ***Continuize*** instrumentā
- Initialization method: šis parametrs nosaka, kā tiek inicializēti sākotnējie klasteru centri. k-Means++ ir uzlabota metode, kas nodrošina labākus sākotnējos centrus. Random initialization metode izvēlas sākotnējos centrus nejauši.
- Re-runs: nosaka, cik reizes algoritms tiks palaists ar dažādiem sākotnējiem centriem, lai iegūtu optimālo rezultātu. Jo vairāk reizes algoritms tiek palaists, jo labāk ir iespēja iegūt optimālos klasterus.

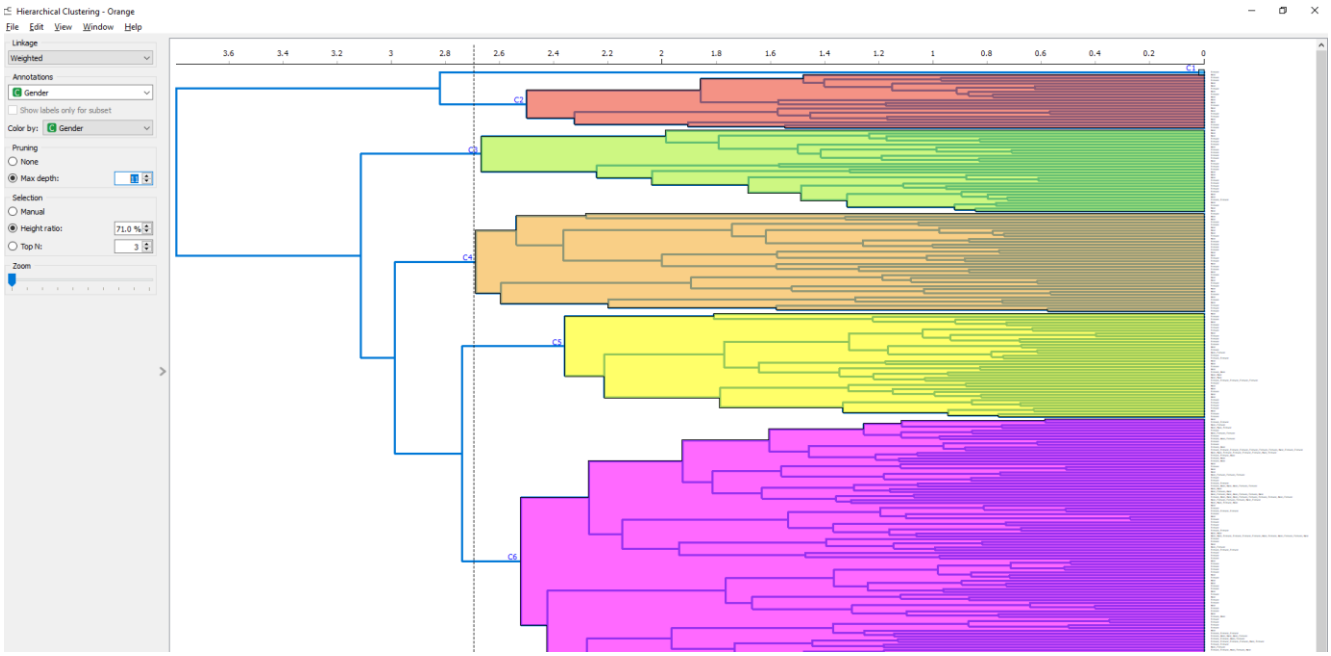
II.2. daļa

Tika veikti 3 eksperimenti ar **dendogrammu**:



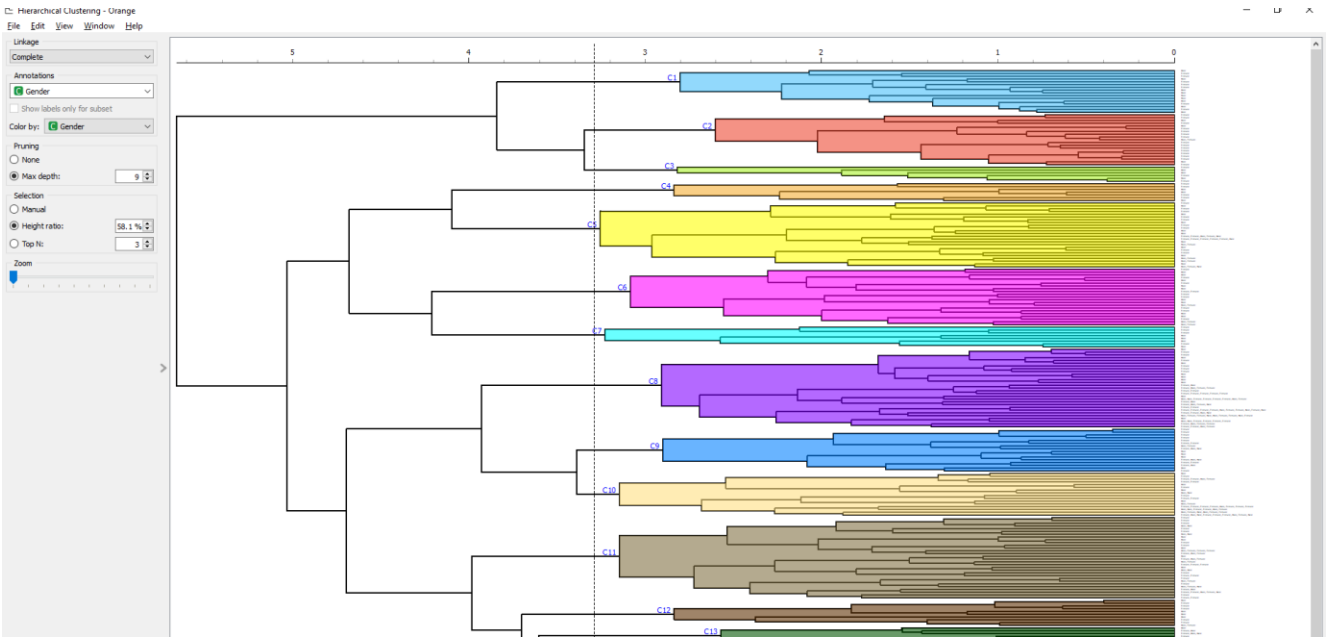
Ilustrācija 28

Iestāstījumi: ward linkage, max depth 10, height ratio 36.1%



Ilustrācija 29

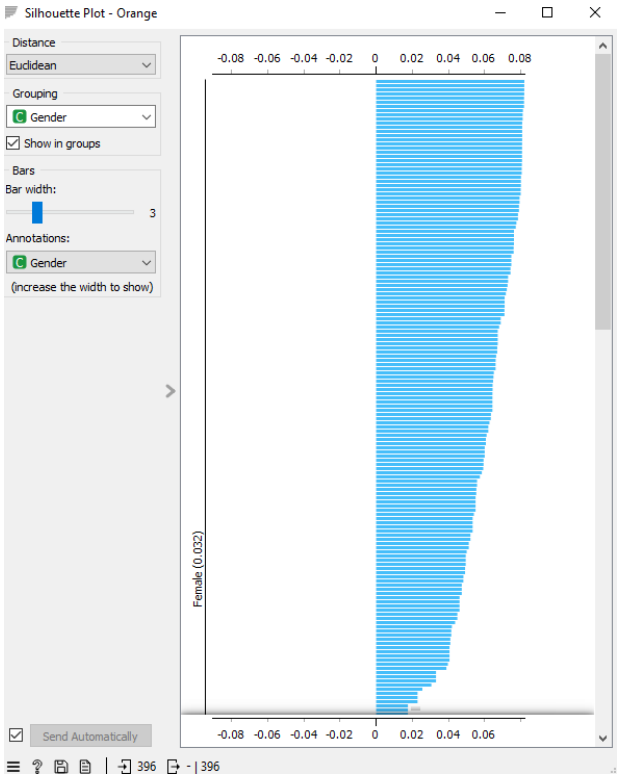
Iestāstījumi: weighted linkage, max depth 11, height ratio 71.0%



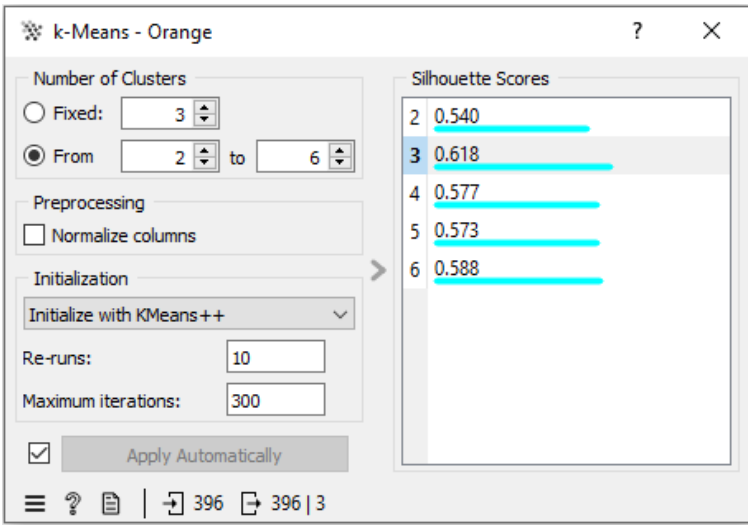
Iestāstījumi: completed linkage, max depth 9, height ratio 58.1%

II.3. daļa

No iepriekšēja uzdevuma K-means algoritmam tika pievienots ***Silhouette Plot*** instruments, kurš atļauja mums apskatīt rezultātus. Izvēlēto klasteru diapazons no 2 līdz 6, šajos iestāstījumos algoritms parādīja vislabāko darba efektivitāti ar 3. klasteru.



Ilustrācija 30



Ilustrācija 31

II. daļas secinājumi

Var secināt, ka manu datu kopu pietiekami viegli atdalīt, tā kā visos trīs eksperimentos tas bija sekmīgi izdarīts. Kā arī tikai izpētīti un detalizēti aprakstīti abu metožu hiperparameti. K-means algoritms darbojas ar 61.8% apmērām, kas bija pārbaudīts klasteru diapazonā no 2 līdz 6.

Tā kā K-means radītāji parādīja vērtības >0.5, var veikt secinājumu, ka datu kopa ir viegli atdāļājamā.

III. daļa

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.

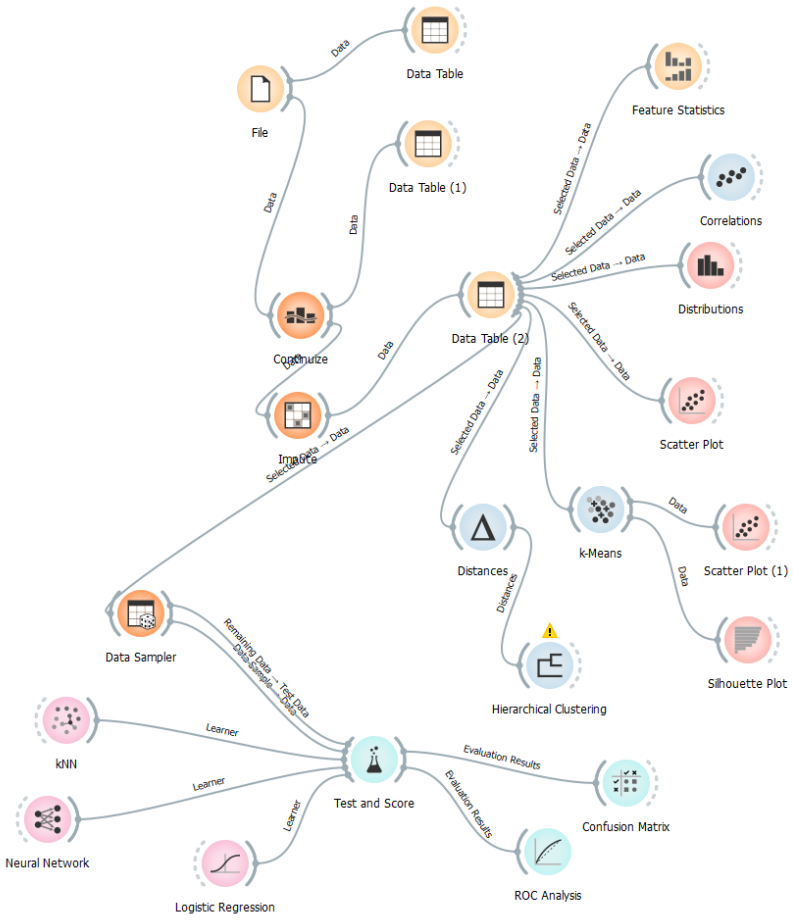
3. Katram algoritmam, lietojot apmācību datu kopu, ir jāveic vismaz 3 eksperimenti, mainot algoritma hiperparametru vērtības un analizējot algoritmu veiktspējas metrikas;

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

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

6. Ir jānovērtē un jāsalīdzina apmācīto modeļu veiktspēja.

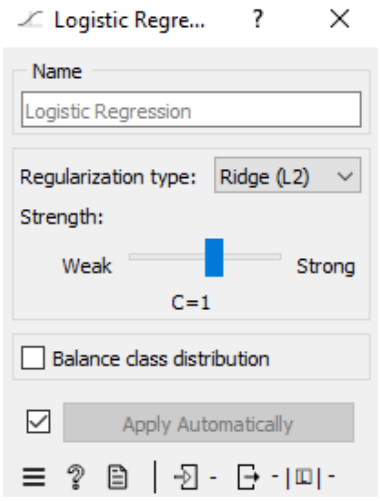
III.1. daļa



Ilustrācija 32

Tikai izvēlēti **Logistic Regression** un **kNN** algoritmi, jo tie tika apskatīti studiju kursā:

1) Loģistiskā regresija.

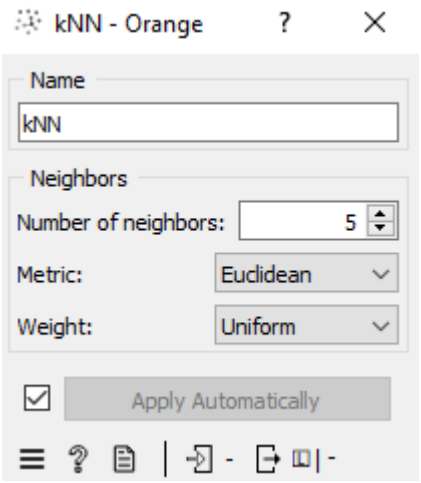


Ilustrācija 33

Parametri:

- Regularization type: loģistiskajai regresijai ir divu veidu regularizācijas - L1 un L2. L1 regularizācija ierobežo koeficientu vērtību summu, savukārt L2 regularizācija ierobežo koeficientu kvadrātu summu.
- Strength: regularizācijas stiprums nosaka, cik liela daļa no koeficientu vērtības tiks ierobežota. Jo lielāka stiprība, jo vairāk koeficientu tiks samazināti.
- Balance class distribution: šis parametrs ir noderīgs, ja datu kopā ir nesabalansēta klases sadalījuma proporcija.

2) kNN algoritms.

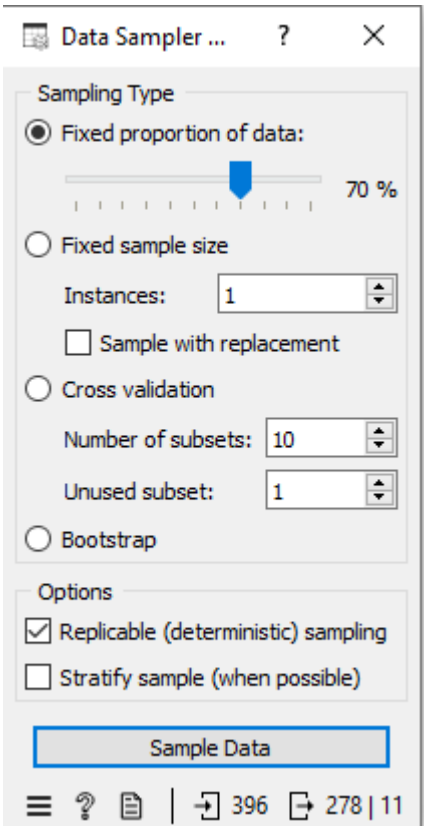


Ilustrācija 34

Parametri:

- Neighbors: Tas ir kaimiņu skaits, kas tiek ņemts vērā, lai veiktu prognozi.
- Number of neighbors: Tas ir skaitlis, kas norāda, cik tuvus kaimiņus ņemt vērā, lai izmantotu to vidējo vērtību prognozēšanai.
- Metric: Tas ir attāluma mērījums, kas tiek izmantots, lai noteiktu kaimiņu attālumu.
- Weight: Tas ir svars, ko var piešķirt katram kaimiņam, lai iegūtu prognozes vērtību.

III.2. daļa

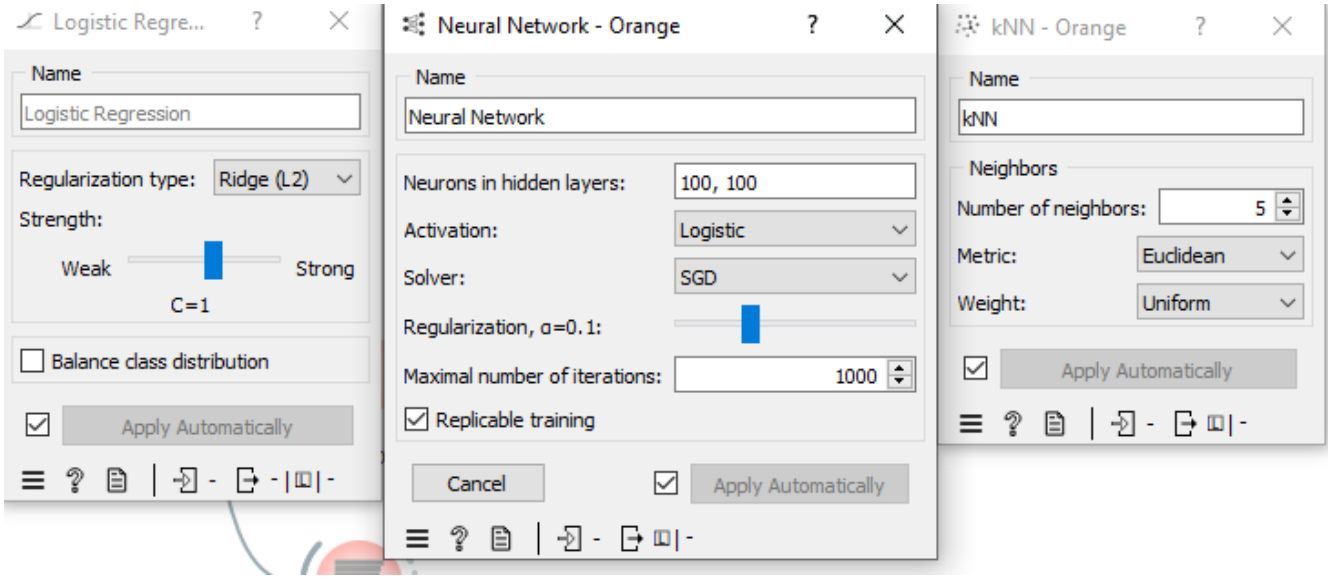


Ilustrācija 35

Tikai pievienots un izmantots ***Data Sampler*** instruments, ar kuru palīdzību es sadalīju datu kopu izmantojot proporciju 70% uz apmācības kopu un 30% uz testa datu kopu

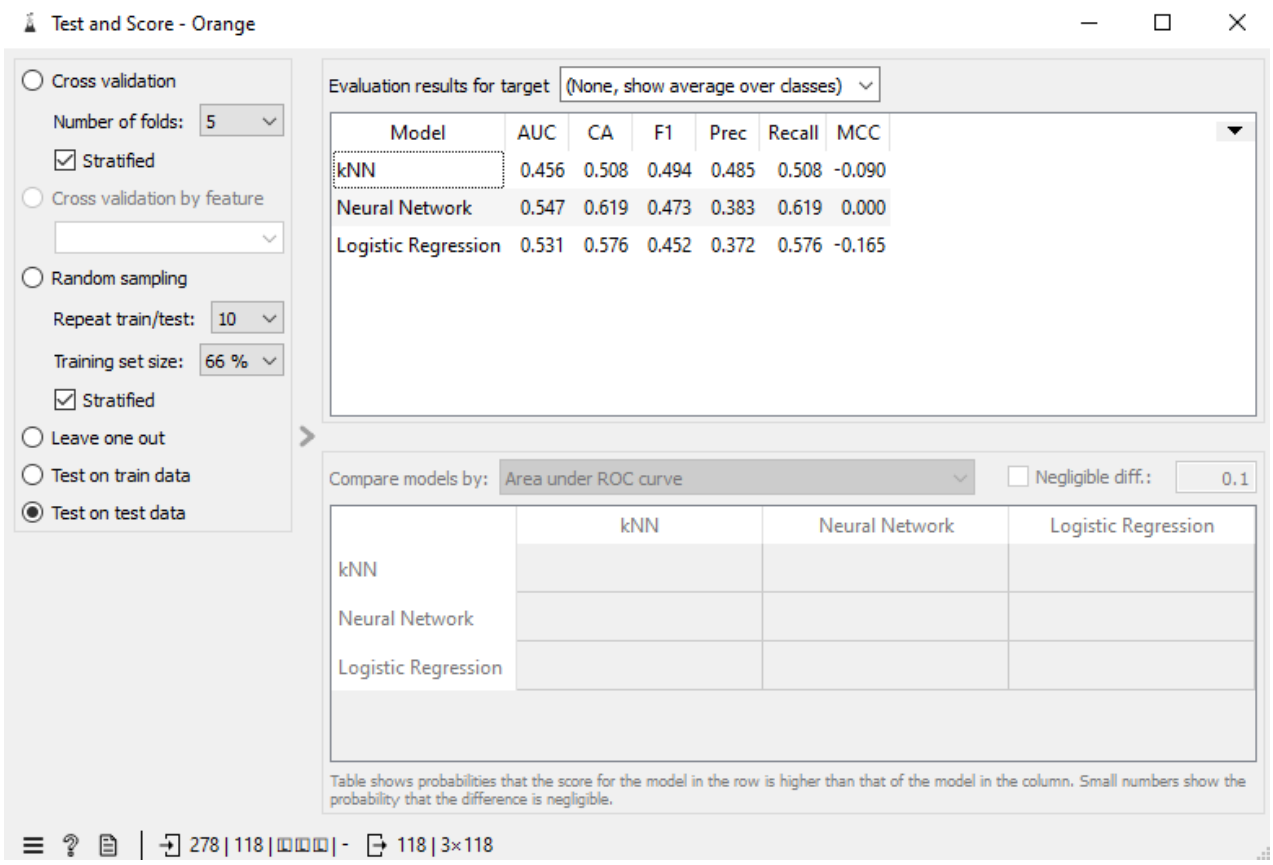
III.3.4.5. daļa

Pirmā testa parametri (starta parametri):



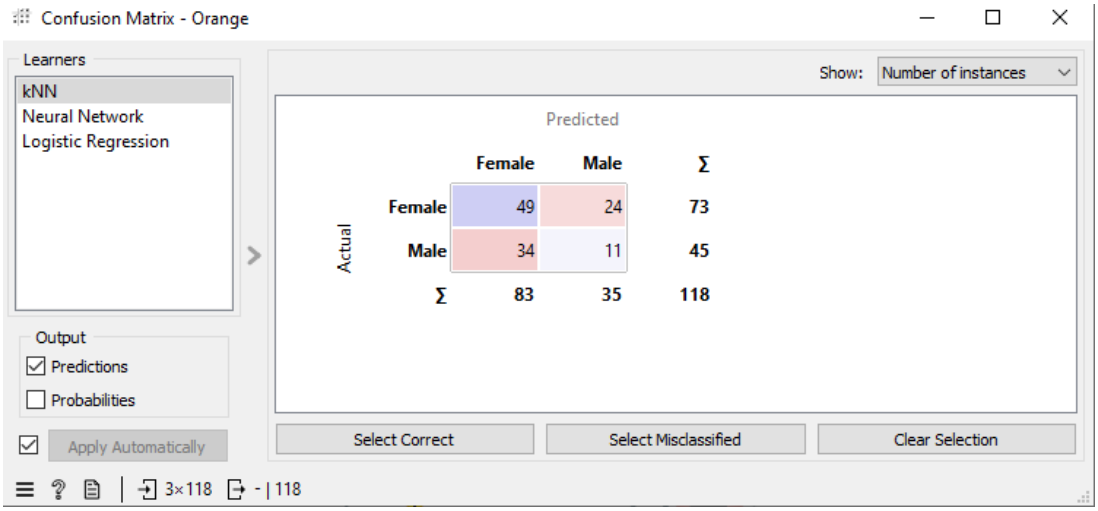
Ilustrācija 36

Pirmā testa rezultāti:

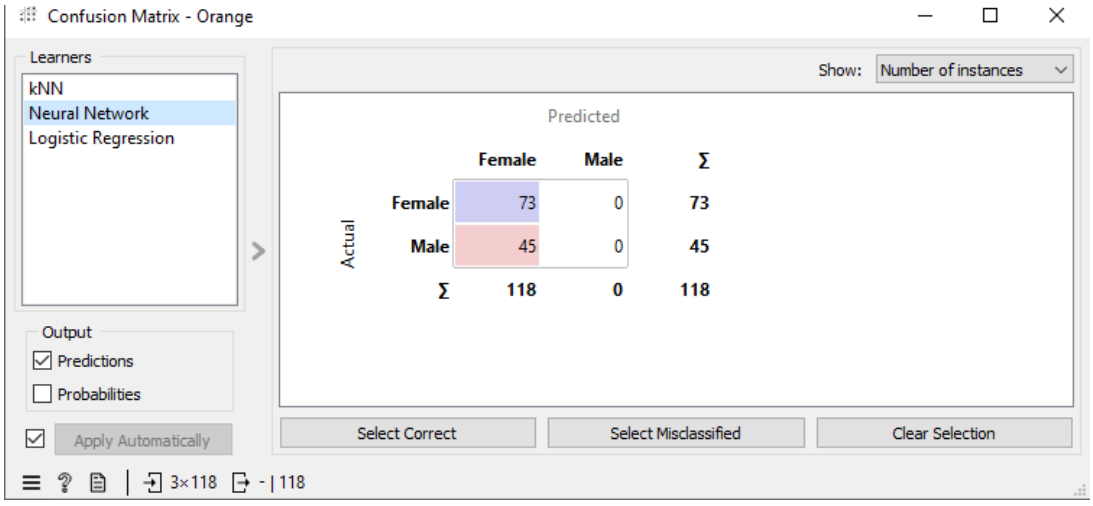


Ilustrācija 37

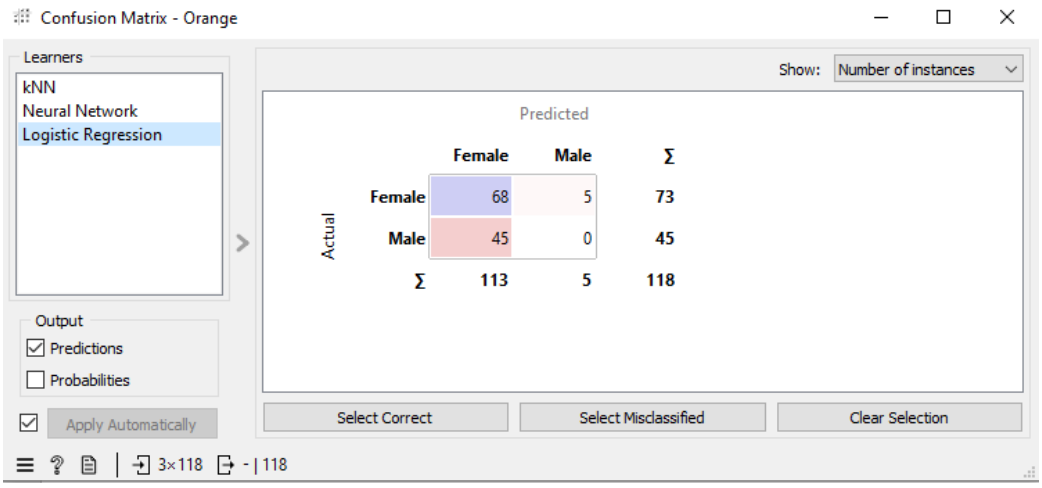
Tikai pievienota katrā testā ***Test And Score*** instrumentam ***Confusion Matrix***, lai attēlotu rezultātus:



Ilustrācija 38

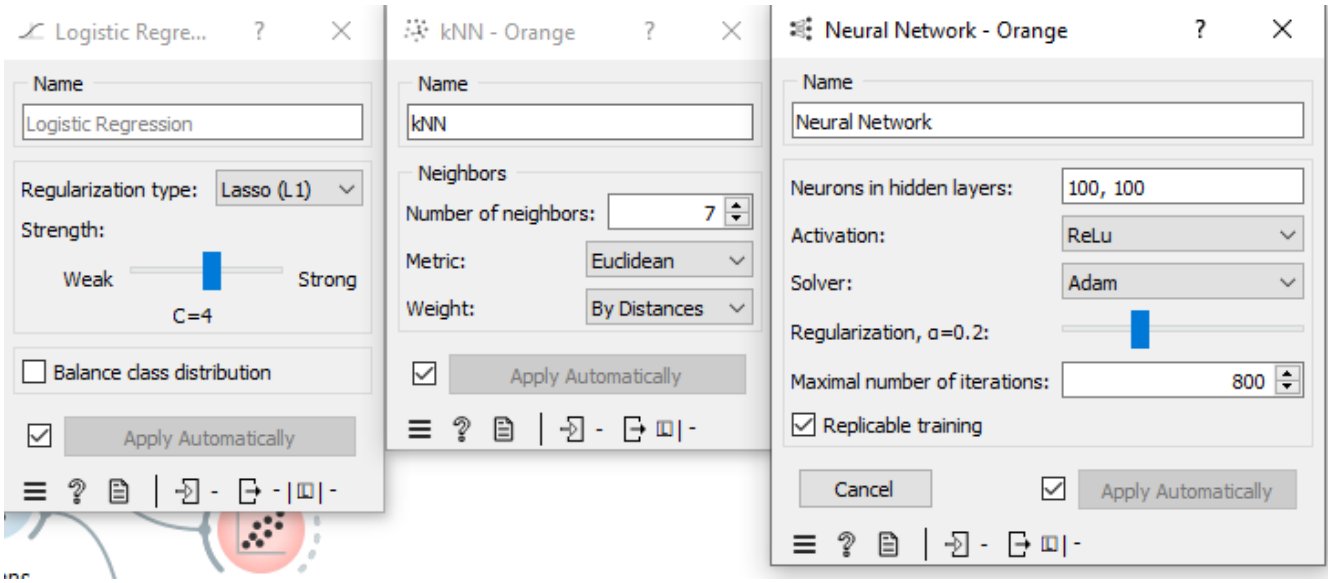


Ilustrācija 39



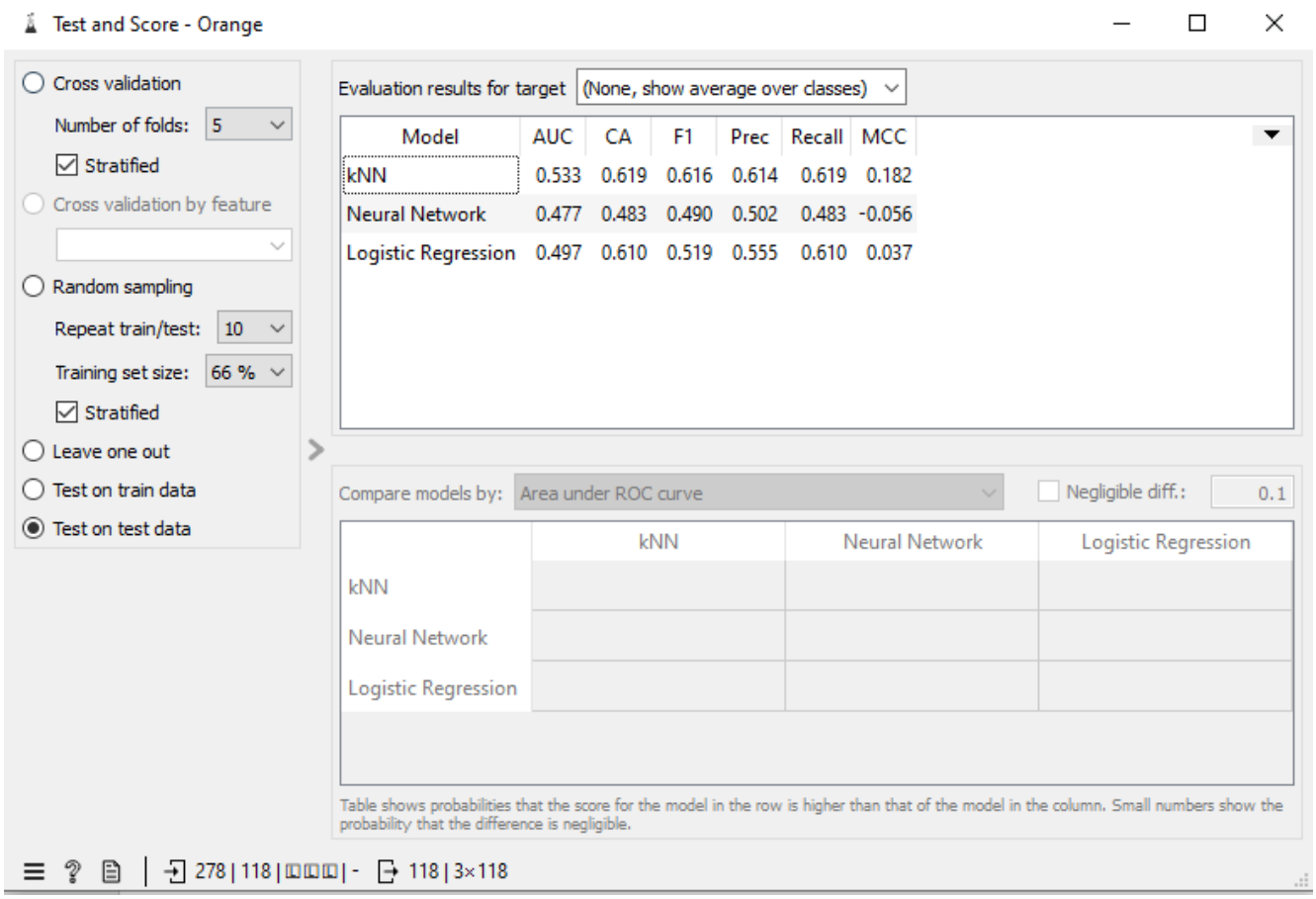
Ilustrācija 40

Otrā testa parametri:



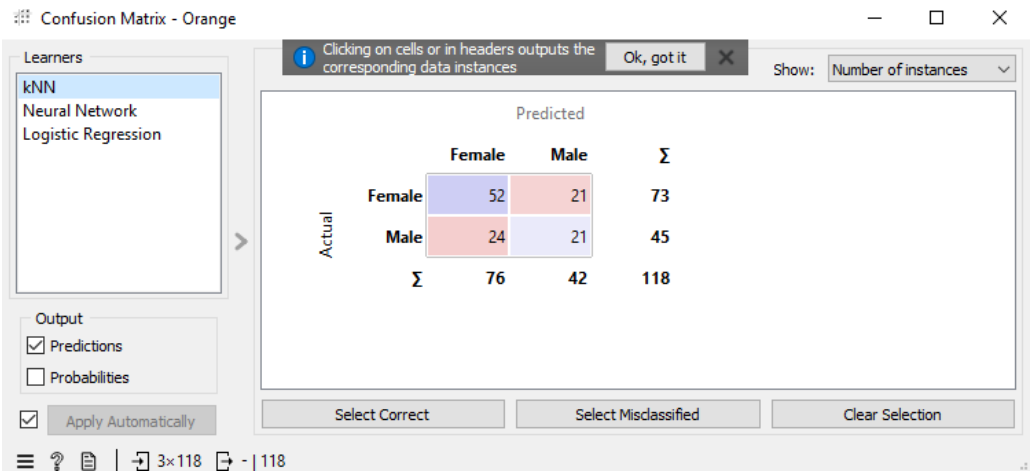
Ilustrācija 41

Otrā testa rezultāti:

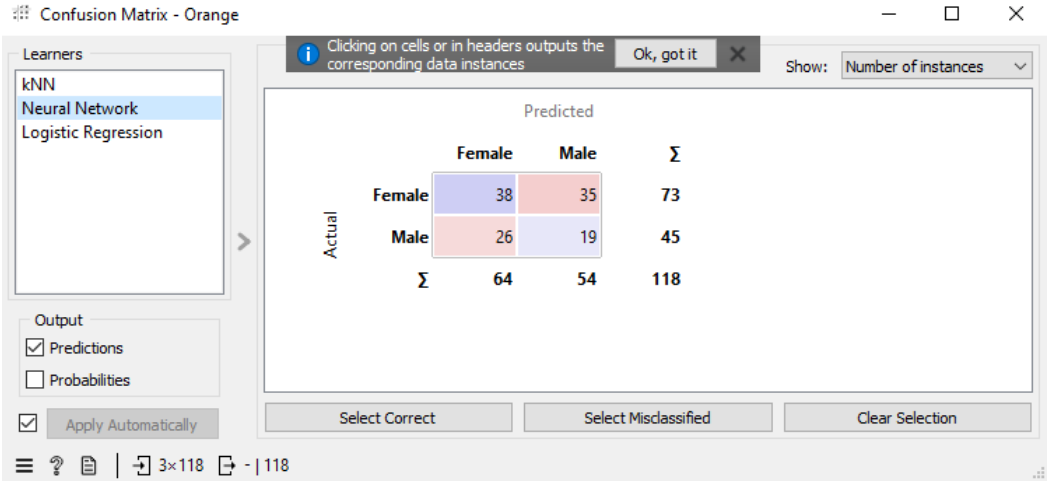


Ilustrācija 42

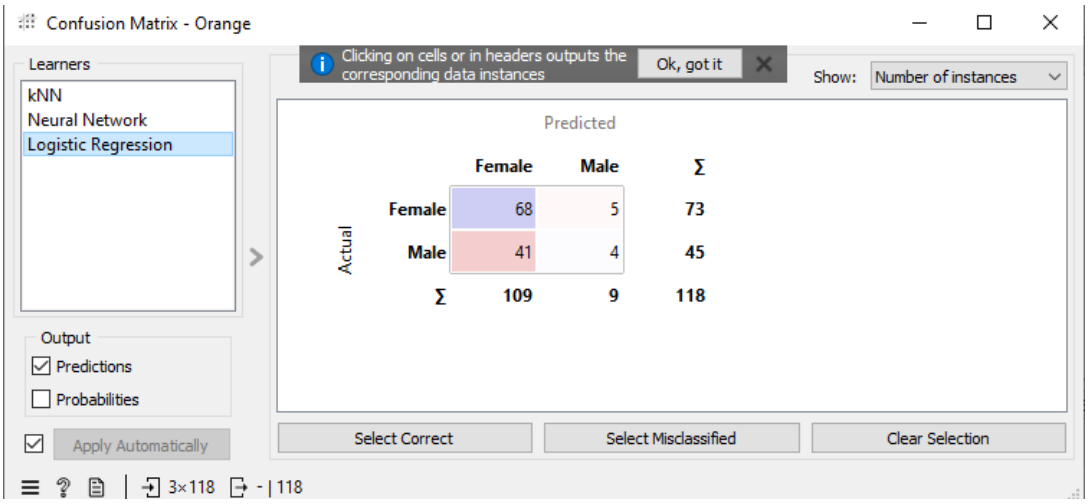
Rezultāti:



Ilustrācija 43

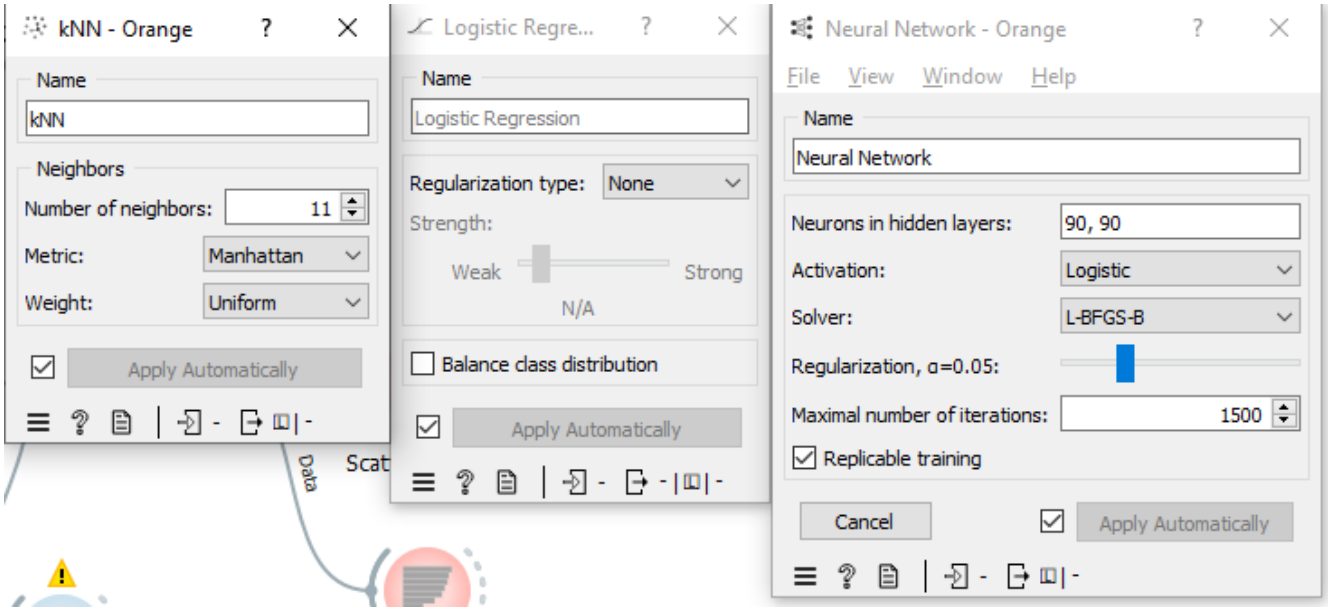


Ilustrācija 44

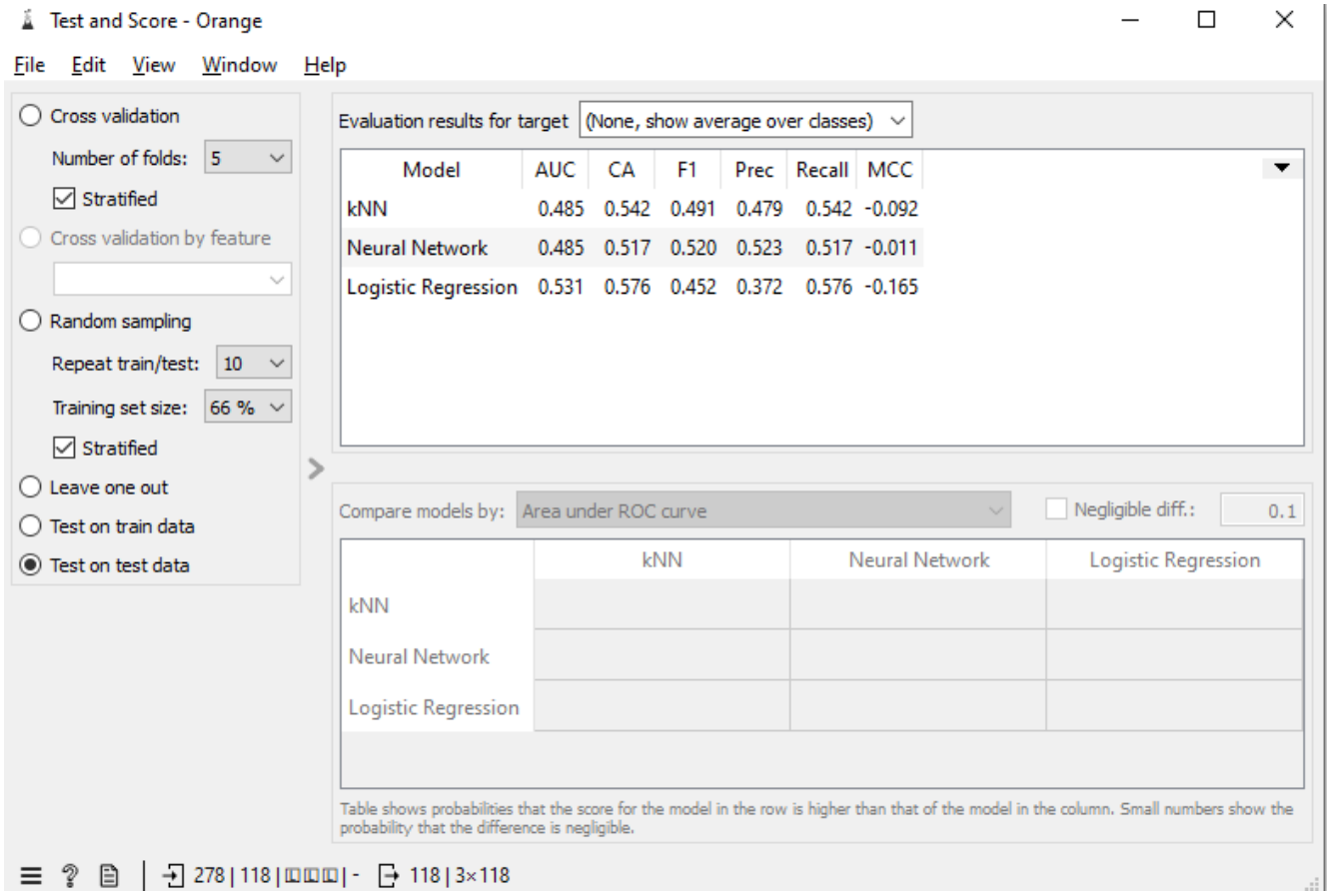


Ilustrācija 45

Trešā testa parametri:

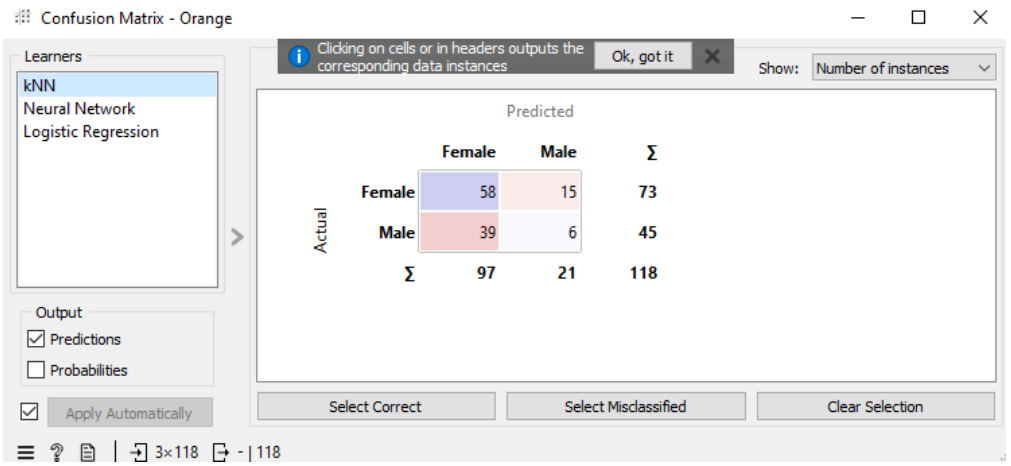


Ilustrācija 46

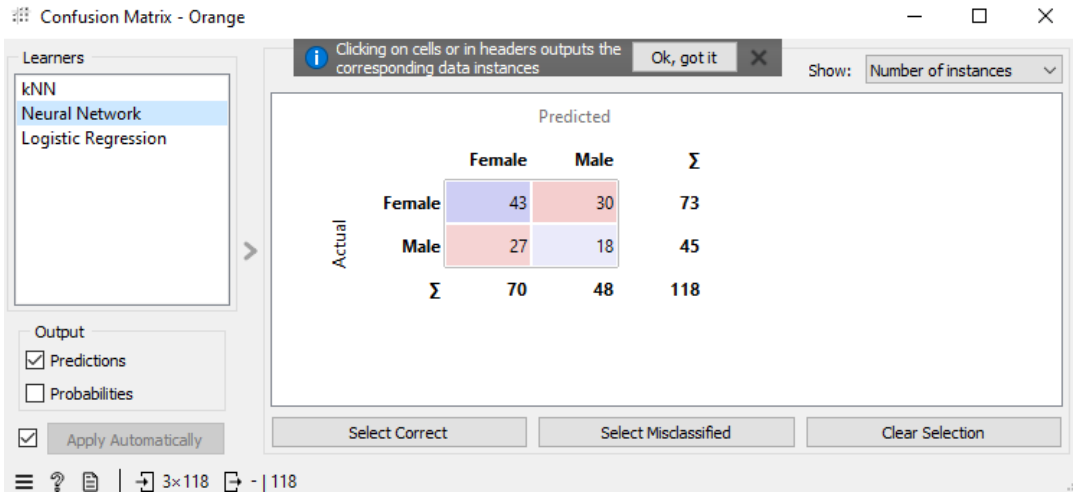


Ilustrācija 47

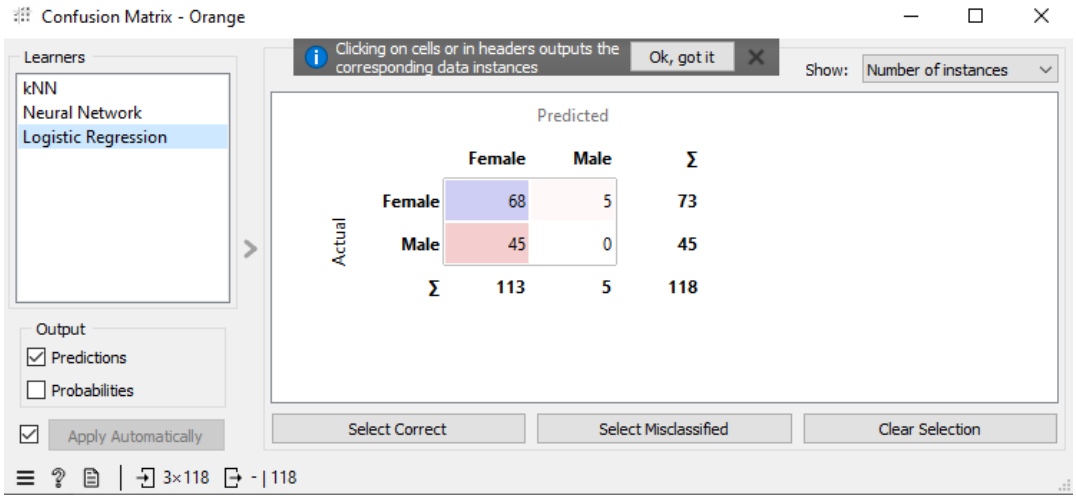
Rezultāti:



Ilustrācija 48



Ilustrācija 49



Ilustrācija 50

Apkopojot eksperimentus, var secināt par algoritmām, kuri darbojos vislabāk testos:

- Pirmajā eksperimentā labākajos rādītājus parādīja ***Neural Network***(AUC = 0.547). Vissliktāko rezultātu parādīja ***kNN***, kuram AUC vērtība bija 0.456.
- Otrajā eksperimentā labākajos rādītājus parādīja ***kNN***(AUC = 0.533). Vissliktāko rezultātu parādīja ***Neural Network***, kuram AUC vērtība bija 0.477.
- Pirmajā eksperimentā labākajos rādītājus parādīja ***Logistic Regression***(AUC = 0.531). Vissliktāko rezultātu parādīja gan ***kNN***, gan ***Neural Network***, kurām AUC vērtība bija 0.485.

III. daļas secinājumi

Manuprāt, otrais praktiskais darbs bija veiksmīgi izpētīts un izpildīts. Tika izpētīta radoša veikala problēmsfēra, lai atrastu ideālo pircēju. Tika konstatēts, ka pamatojoties uz rezultātiem, tā ir sieviete līdz 70 gadu vecuma un līdz 5 cilvēkiem ģimenē, kura atstāj visvairāk naudas. Tai ir vislielākais patēriņu rādītājs.

Testu gaitā tikai nofiksēts interesants fakts, ka sanāca noregulēt iestāstījumus tā, ka katram algoritmam ir tieši viena uzvara. Katrs algoritms bija veiksmīgi iestāstīts tā, ka paradīja lielāko efektivitāti, nekā citi.

Vislielāko rezultātu praktiska darba gaitā parādīja ***Neural Network*** ar vislielāko efektivitāti 54.7%.

Informācijas avoti

https://orangedatamining.com/widget-catalog/unsupervised/kmeans/

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

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

https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html

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

https://www.kaggle.com/datasets/datascientistanna/customers-dataset