



DATORZINĀTNES UN INFORMĀCIJAS TEHNOLOĢIJAS FAKULTĀTE
Priekšmeta "Mākslīgā intelekta pamati(1),22/23-P"

2. Praktiskais darbs

Artūrs Adamovičs

<https://github.com/ginc55/PD2>

2.kurss 6.grupa

Studenta apliecības nr. 211RDB164

Rīga, 2023

Saturs

1.Daļa- Datu pirmapstrāde/izpēte.....	3
1.daļas secinājumi.....	10
2.daļa – Nepārraudzītā mašīnmācīšanās	11
2.Daļas secinājumi.....	15
3. daļa – Pārraudzītā mašīnmācīšanās	15
3.daļas secinājumi.....	23
Izmantoto avotu saraksts	24

1.Daļa- Datu pirmapstrāde/izpēte

Datu kopas apraksts

Esmu izvēlējis datu kopu, kas apraksta ātrās ēdināšanas edienu uztura faktiem. Datu kopas nosaukums ir “Fastfood Nutrition”. Avots: Data.world API . Šīs datu kopas izveidotājs un īpašnieks ir “Ulrik Thyge Pedersen”. Datu kopu izveidoja lietojot populārākss ātrās ēdināšanas restorānu piedāvājumus, lai informētu un parādītu patērētājiem, cik neveselīgi tas ir, pēc kā patērētājs varētu veikt informētu izvēli. Datu kopas licences nosaukums ir “Attribution 4.0 International”, kas min, ka šo ar šo datu kopu var dalīties un pārveidot pēc saviem ieskatiem, norādot atsaucis uz oriģinālo darbu. Veids, kā datu kopa tika savākta diemžēl nebija minēts. Datu kopas oriģinālā saite:<https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition> .

Kopējais datu objektu skaits kopā ir 515 objekti.

	Name	Type	Role	Values
1	restaurant	C categorical	target	Arbys, Burger King, Chick Fil-A, Dairy Queen, Mcdonalds, Sonic, Subway, Taco Bell
2	calories	N numeric	feature	
3	cal_fat	N numeric	feature	
4	total_fat	N numeric	feature	
5	sat_fat	N numeric	skip	
6	trans_fat	N numeric	skip	
7	cholesterol	N numeric	skip	
8	sodium	N numeric	skip	
9	total_carb	N numeric	feature	
10	fiber	N numeric	feature	
11	sugar	N numeric	feature	
12	protein	N numeric	feature	
13	vit_a	N numeric	skip	
14	vit_c	N numeric	skip	
15	calcium	N numeric	skip	
16	salad	C categorical	skip	Other
17	item	S text	meta	

1.attēls datu kopas atribūti

Datu kopā ir 8 klases: *Arbys*, *Burger King*, *Chick Fil-A*, *Dairy Queen*, *Mcdonalds*, *Sonic*, *Subway* un *Taco Bell*. Šīs klases ir zem atribūta “restaurant”, un norāda uz restorāna nosaukumu.

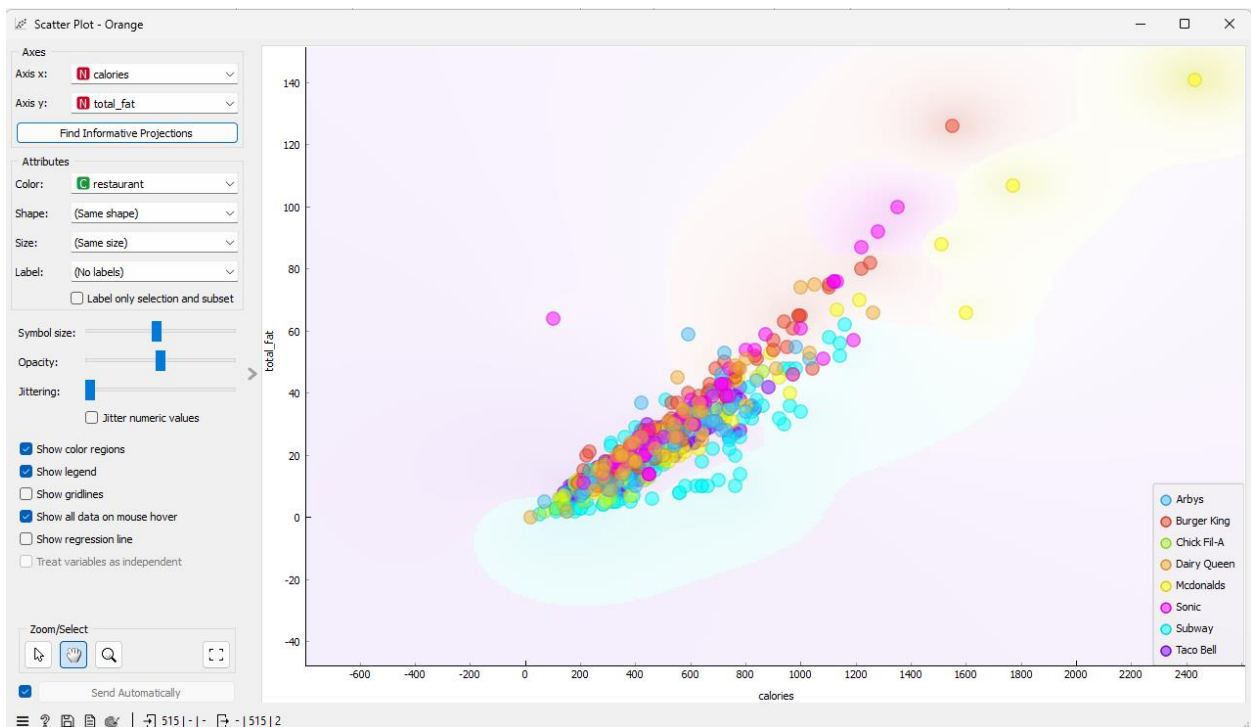
Nr.	Nosaukums	Skaidrojums	Tips	Diapazons
1.	Restaurant	Restorāna nosaukums	Kategorija	<i>Arbys</i> , <i>Burger King</i> , <i>Chick Fil-A</i> , utml.
2.	Calories	Kaloriju daudzums	Skaitliska vērtība	[20;2430]
3.	Cal_fat	Kaloriju daudzums no taukiem	Skaitliska vērtība	[0;1270]
4.	Total_fat(grami)	Kopējais tauku daudzums	Skaitliska vērtība	[0;141]
5.	Total_carb(grami)	Kopējais oglekļahidrātu daudzums	Skaitliska vērtība	[0;126]
6.	Fiber(grami)	Šķiedrvielu daudzums	Skaitliska vērtība	[0;17]
7.	Sugar(grami)	Cukura daudzums	Skaitliska vērtība	[0;87]
8.	Protein(grami)	Proteīna daudzums	Skaitliska vērtība	[1;186]
9.	item	Ēdiena nosaukums	Teksta vērtība	Jebkāds teksts

1.tabula atribūtu klasificēšana

Datu objektu skaits, kas pieder katrai klasei-

- “Arbys”- 55 datu objekti
- “Burger King”- 70 datu objekti
- “Chick Fil-a”- 27 datu objekti
- “Dairy Queen”- 42 datu objekti
- “Mcdonalds”- 57 datu objekti
- “Sonic”- 53 datu objekti
- “Subway”- 96 datu objekti
- “Taco Bell”-115 datu objekti

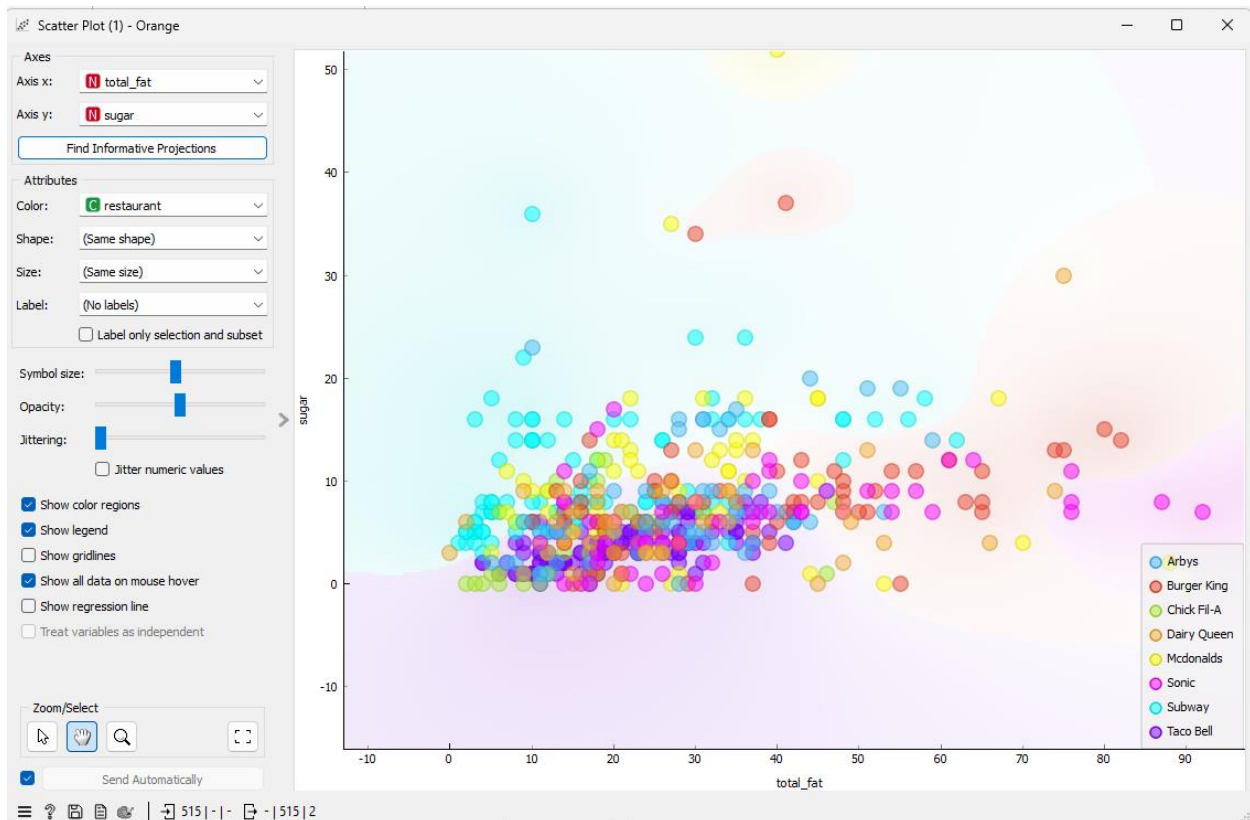
Izkliedes diagrammas



2.attēls kalorijas pret taukiem

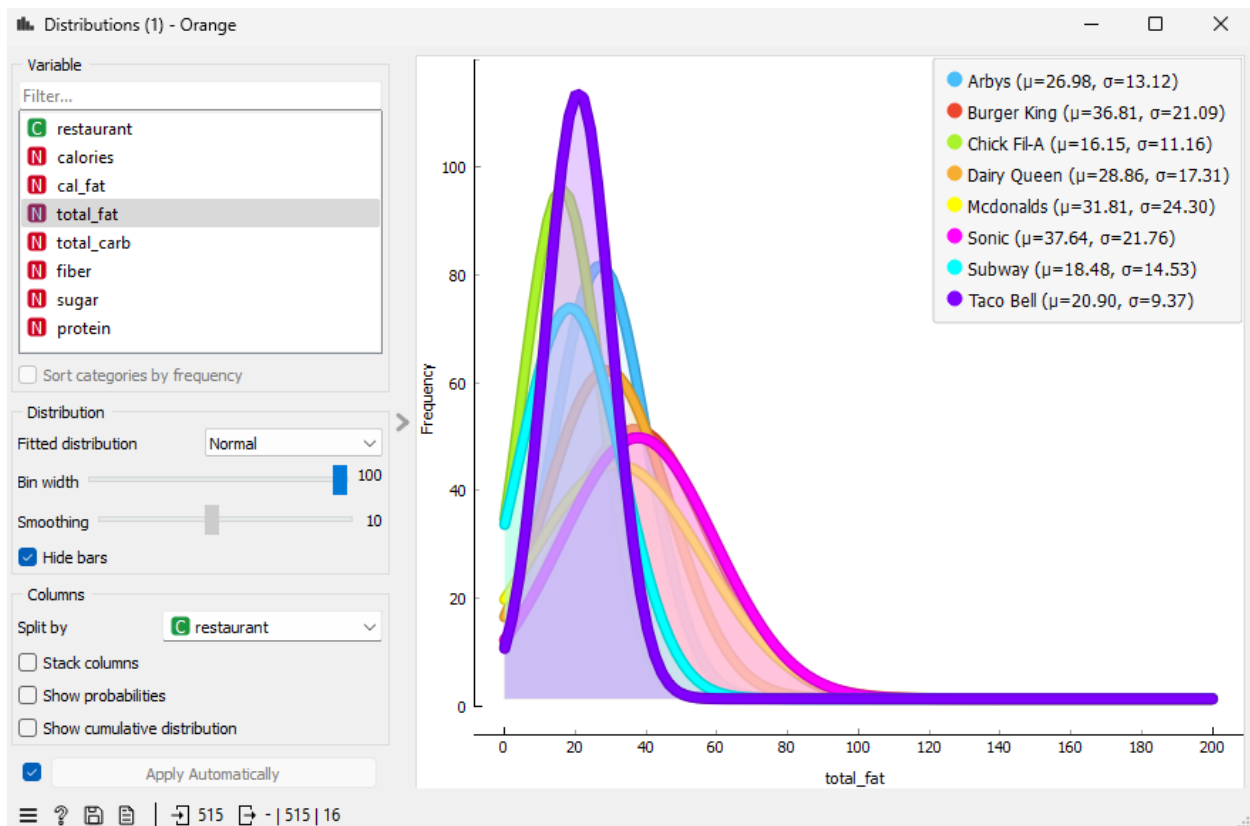
Pēc 2.attēlā parādītās izkļiedes diagrammas var redzēt, ka ātrās ēdināšanas restorāns “Subway” savā produktu klāstā piedāvā produktus ar salīdzinoši zemu kaloriju daudzumu, kā arī mazu tauku daudzumu, tikmēr restorānam “Mcdonalds” ir pāris produkti ar vislielāko kaloriju un tauku daudzumu.

Pēc 3.attēlā parādītās izkļiedes diagrammas var secināt, ka tomēr jāuzmanās izvēloties restorāna “Subway” produktus, jo tomēr no visiem ātrās ēdināšanas restorāniem “Subway” produktos ir nesamērīgi daudz cukura, bet ir arī produkti restorāna klāstā, kuriem cukura daudzums nav tik liels, kas nozīmē, kad izvēloties “Subway” produktus jāpievērš uzmanība uz cukura daudzumu.



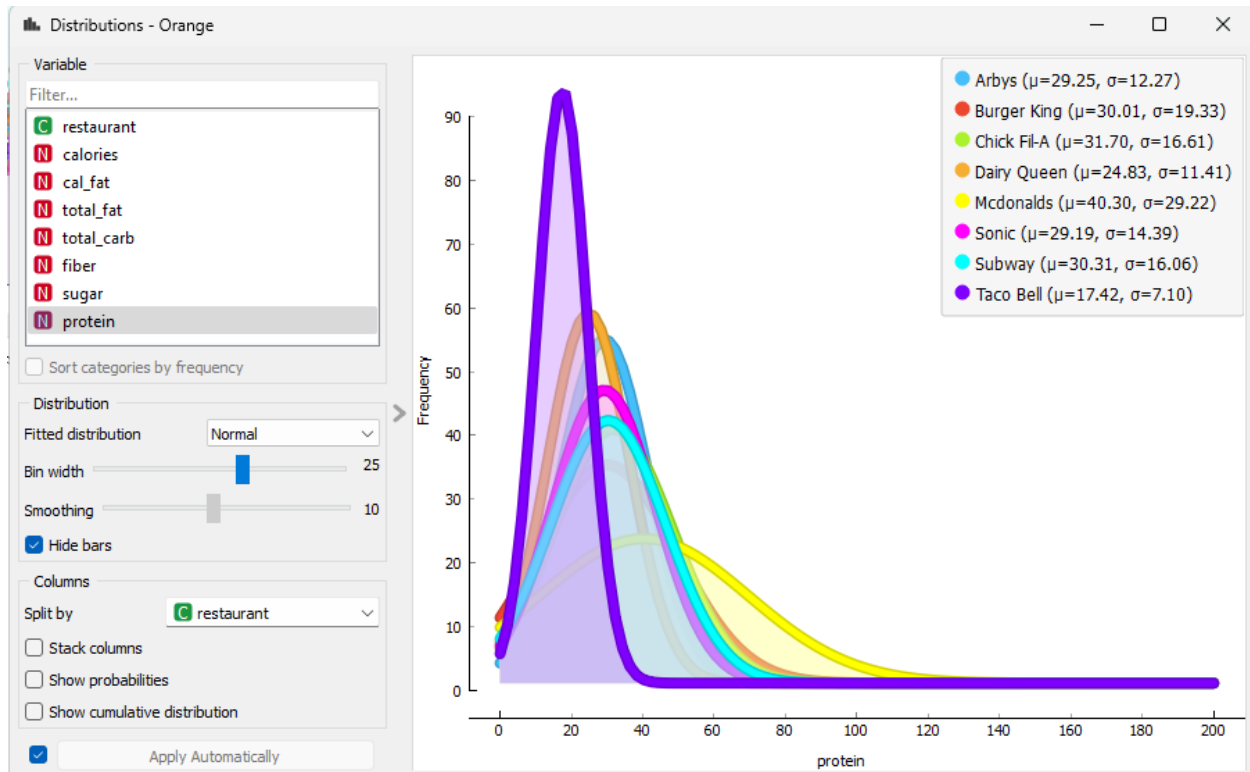
3.attēls tauku saturs pret cukuru

Histogrammas



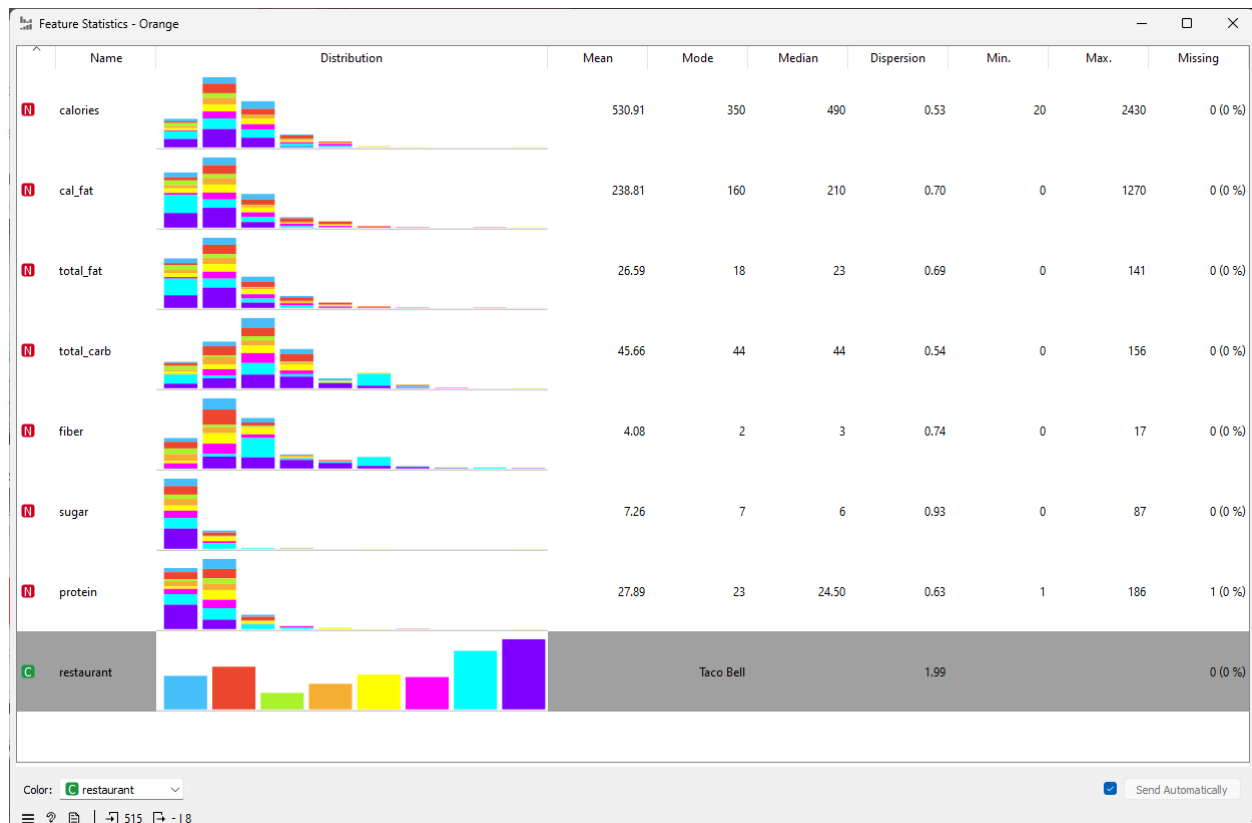
4.attēls histogramma 1

4.attēla histogrammā varam spriest, ka “Taco Bell” piedāvājumā ir vislielākais produktu klāsts, kurā ir salīdzinoši mazs tauku daudzums(pārējo restorānu perspektīvā).



5.attēls histogramma 2

5.attēla histogrammā es novēroju, ka restorānam “Mcdonalds” ir diezgan liels klāsts ar vidēja-liela proteīna daudzuma produktiem.



6.attēls statistiskie rādītāji

6.attēlā redzama vidējā, mediāna, modas, dispersijas, minimālās un maksimālās vērtības starp visiem datu kopas atribūtiem.

	restaurant	item	calories	cal_fat	total_fat	total_carb	fiber	sugar	protein
1	Mcdonalds	Artisan Grilled ...	380	60	7	44	3	11	37
2	Mcdonalds	Single Bacon S...	840	410	45	62	2	18	46
3	Mcdonalds	Double Bacon S...	1130	600	67	63	3	18	70
4	Mcdonalds	Grilled Bacon S...	750	280	31	62	2	18	55
5	Mcdonalds	Crispy Bacon S...	920	410	45	81	4	18	46
6	Mcdonalds	Big Mac	540	250	28	46	3	9	25
7	Mcdonalds	Cheeseburger	300	100	12	33	2	7	15
8	Mcdonalds	Classic Chicken...	510	210	24	49	3	6	25
9	Mcdonalds	Double Cheese...	430	190	21	35	2	7	25
10	Mcdonalds	Double Quarter...	770	400	45	42	3	10	51
11	Mcdonalds	Filet-O-Fish®	380	170	18	38	2	5	15
12	Mcdonalds	Garlic White Ch...	620	300	34	48	3	11	32
13	Mcdonalds	Grilled Garlic W...	530	180	20	48	3	11	42
14	Mcdonalds	Crispy Garlic W...	700	300	34	67	5	11	33
15	Mcdonalds	Hamburger	250	70	8	31	2	6	13
16	Mcdonalds	Lobster Roll	290	50	5	35	2	3	24
17	Mcdonalds	Maple Bacon Di...	640	330	36	40	3	10	37
18	Mcdonalds	Grilled Maple B...	580	190	21	50	3	14	48
19	Mcdonalds	Crispy Maple B...	740	310	35	69	5	14	39
20	Mcdonalds	McChicken	350	130	15	40	2	5	15
21	Mcdonalds	McDouble	380	160	18	34	2	7	23
22	Mcdonalds	McRib	480	200	22	45	2	12	25
23	Mcdonalds	Pico Guacamol...	580	300	33	41	4	7	29
24	Mcdonalds	Grilled Pico Gu...	520	160	18	50	4	12	40
25	Mcdonalds	Crispy Pico Gua...	680	280	32	69	6	12	31
26	Mcdonalds	Premium Butte...	570	200	23	64	4	11	28
27	Mcdonalds	Premium Crisp...	530	200	22	59	3	13	25
28	Mcdonalds	Quarter Pound...	530	240	27	41	3	10	31
29	Mcdonalds	Signature Srirac...	670	320	35	56	4	13	32
30	Mcdonalds	Grilled Signatur...	560	180	20	56	4	14	41
31	Mcdonalds	Crispy Signatur...	730	300	33	75	5	13	32
32	Mcdonalds	Sweet BBQ Bac...	690	340	37	51	3	14	38
33	Mcdonalds	Grilled Sweet B...	630	200	22	61	4	18	48
34	Mcdonalds	Crispy Sweet B...	800	320	36	80	5	18	39
35	Mcdonalds	3 piece Butterm...	370	190	21	16	0	0	28
36	Mcdonalds	4 piece Butterm...	480	250	28	21	0	1	38
37	Mcdonalds	6 piece Butterm...	760	390	44	32	1	1	58
38	Mcdonalds	10 piece Butter...	1210	630	70	52	1	4	94
39	Mcdonalds	12 piece Butter...	1510	790	88	64	1	2	115

7.attēls datu kopas fragments

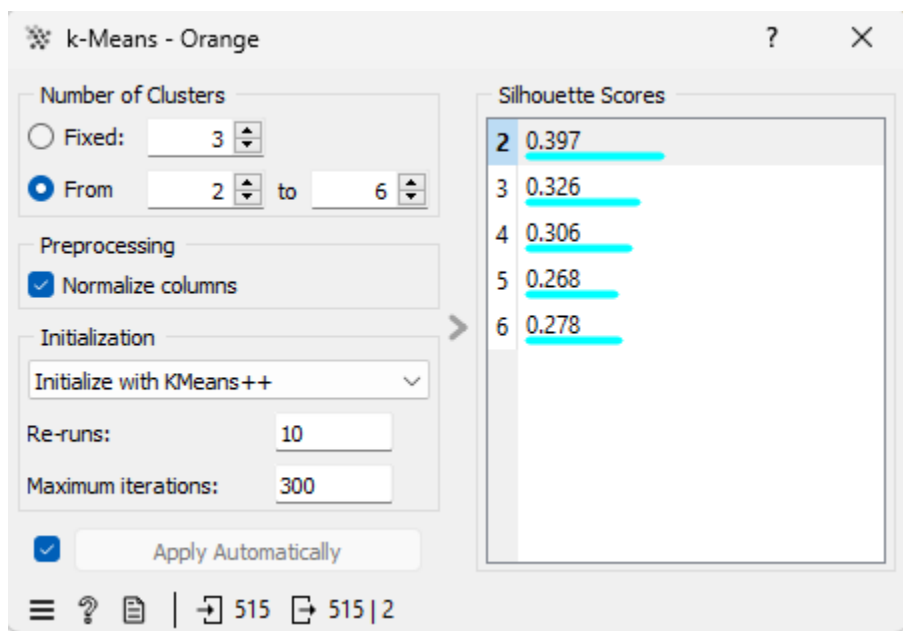
7.attēlā esmu uztaisījis ekrānuzņēmumu no datu kopas faila struktūru.

1.daļas secinājumi

- Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)?
Nē, izpētot datu kopu, dominē “Taco Bell” un “Subway” klase ar 115 un 96 datu objektiem.
- Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?
Brīžiem ir diezgan grūti atšķirt kurai klasei pieder datu objekts. Novēroju, ka starp šīm ēstuvēm ir diezgan līdzīgas vērtības, jo visas šīs ēstuves ir neveselīgas, un ja pievienotu kādu veselīgu restorānu, tad uzreiz varētu atšķirt vizuālajā attēlojumā.
- Cik datu grupējumus ir iespējams identificēt, pētot datu vizuālo atspoguļojumu?
Pētot attēlus vēl vienu reizi, varu uzreiz viegli identificēt 6 no 8 datu grupējumiem.
- Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?
Visos gadījumos datu grupējumi atrodas tuvu viens otram.

2.daļa – Nepārraudzītā mašīnmācīšanās

K-vidējo algoritms

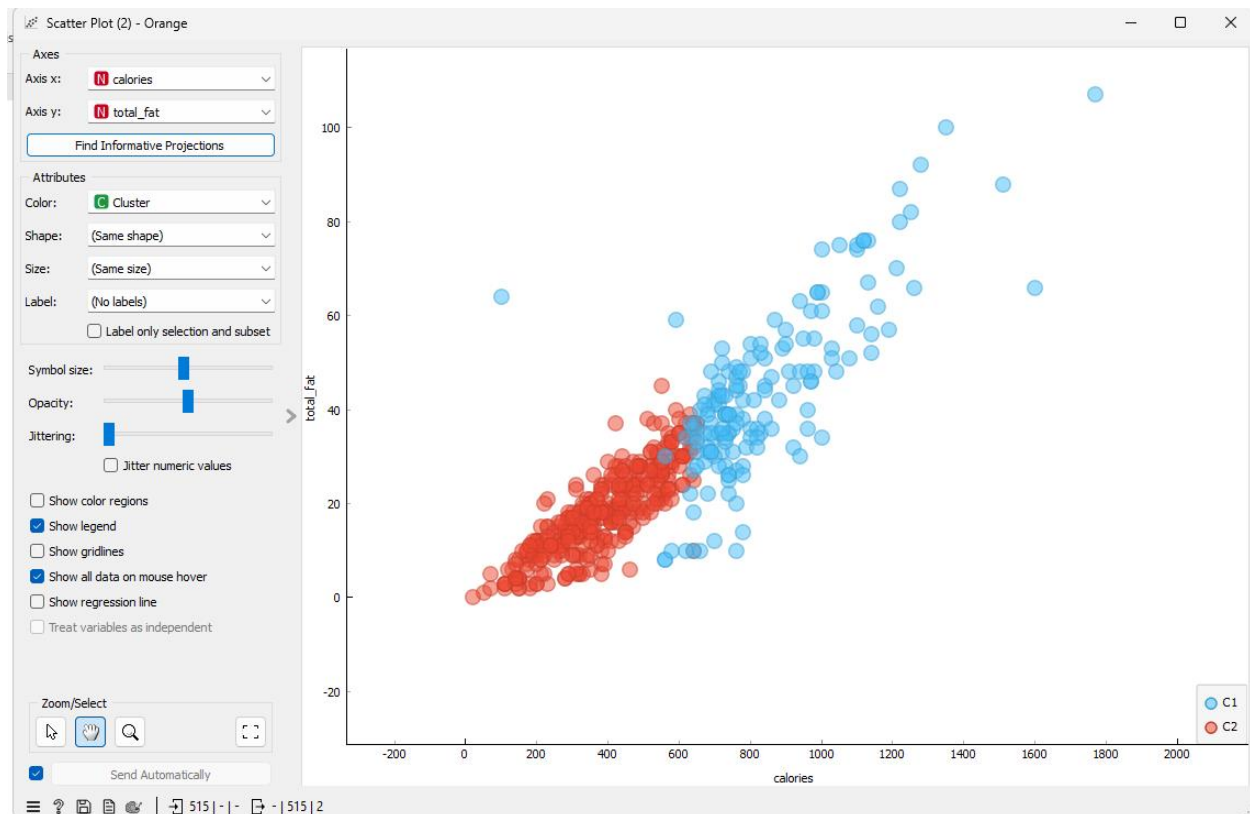


8.attēls K-vidējo vērtības

8.attēlā varam redzēt K-vidējā algoritma hiperparametrus:

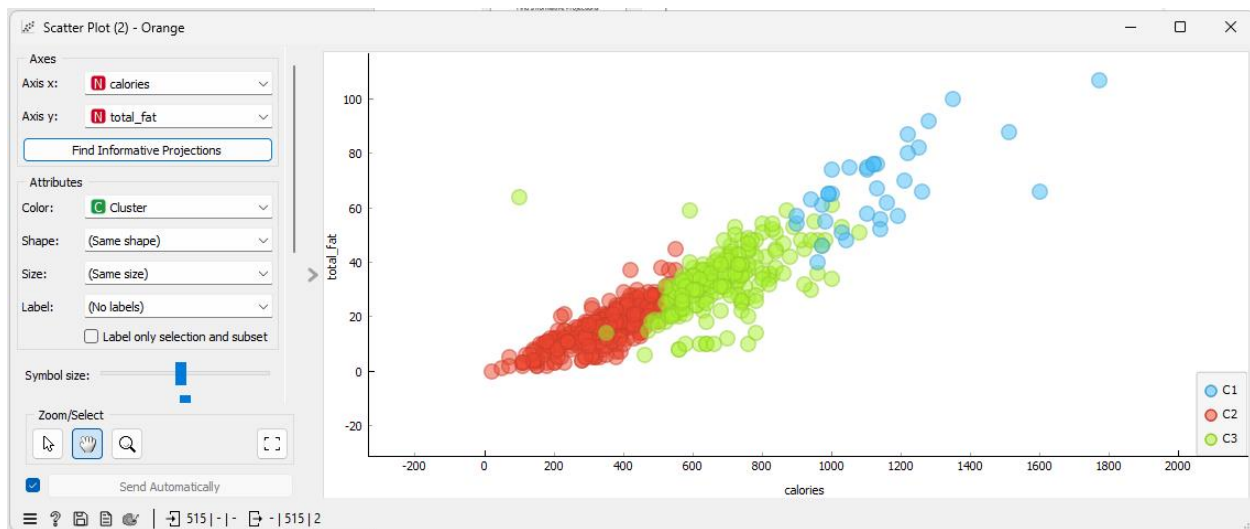
- Klasteru skaitu, to izvēlamies 5.
- Normalizēt kolonnas, to iezīmējam, lai izņemtu ārā datu objektus ar tukšām vērtībām.
- Inicializācijas metode, testēju abus, bet rezultātu manā gadījumā nemainīja.
- Maksimālo iterāciju skaits, to atstājam 300, testēju palielinot un pazeminot skaitli, bet rezultātu nemainīja.

Pie lauka *Silhouette Scores* izvēlamies klasteru skaitu, kura vērtība ir tuvāk pie 1, kas manā variantā ir 2 klasteru skaits.



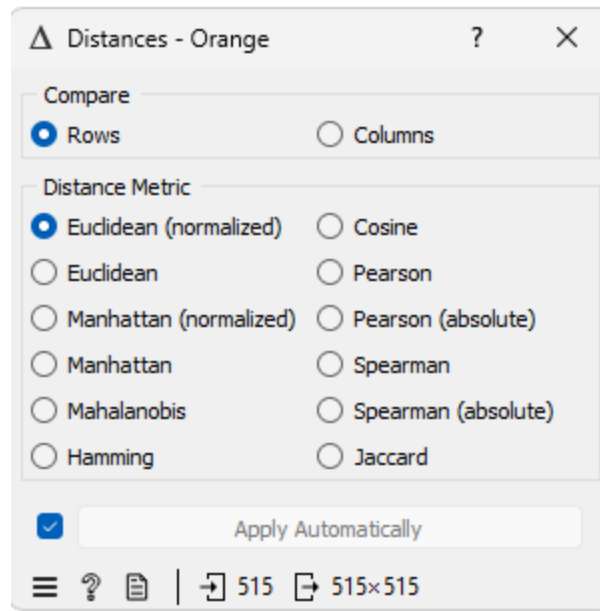
9.attēls 2 klasteri

9.attēlā varam redzēt, ka datu kopas esošās klases divos klasteros samērā labi atdalāmas, bet tiklīdz izvēlamies 3 klasterus, kā 10.attēlā uzreiz jau paliek slikti atdalāms.



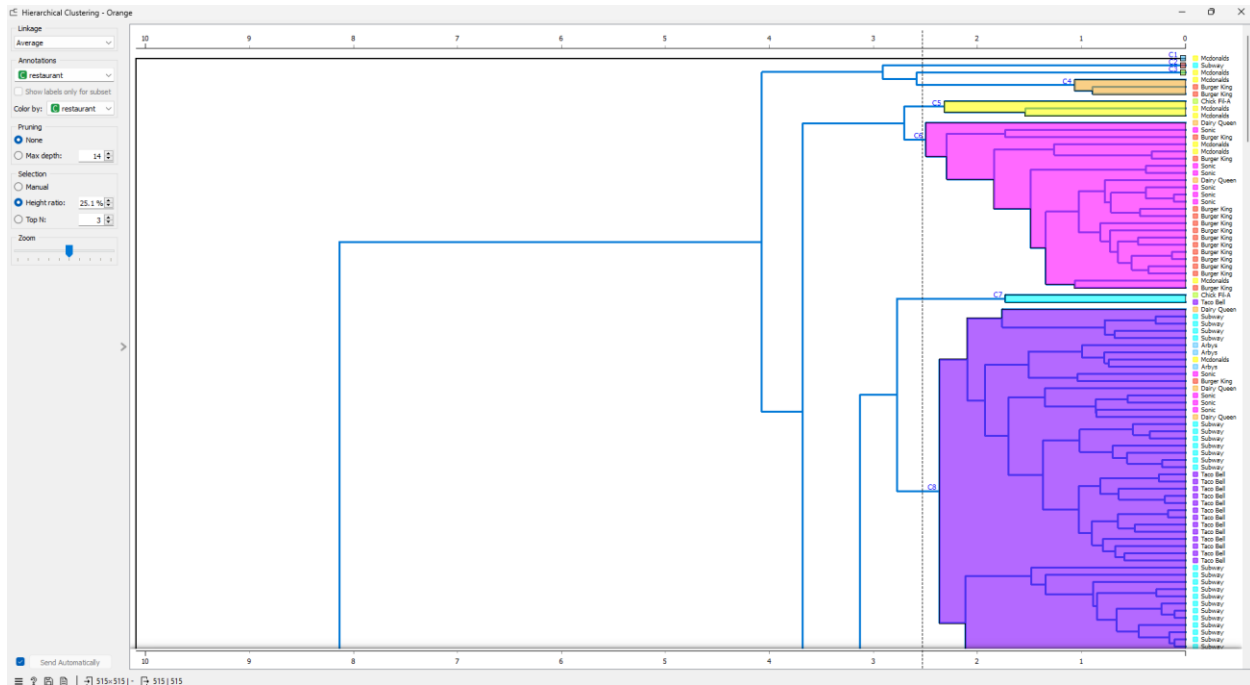
10.attēls 3 klasteri

hierarhiskā klasterizācija

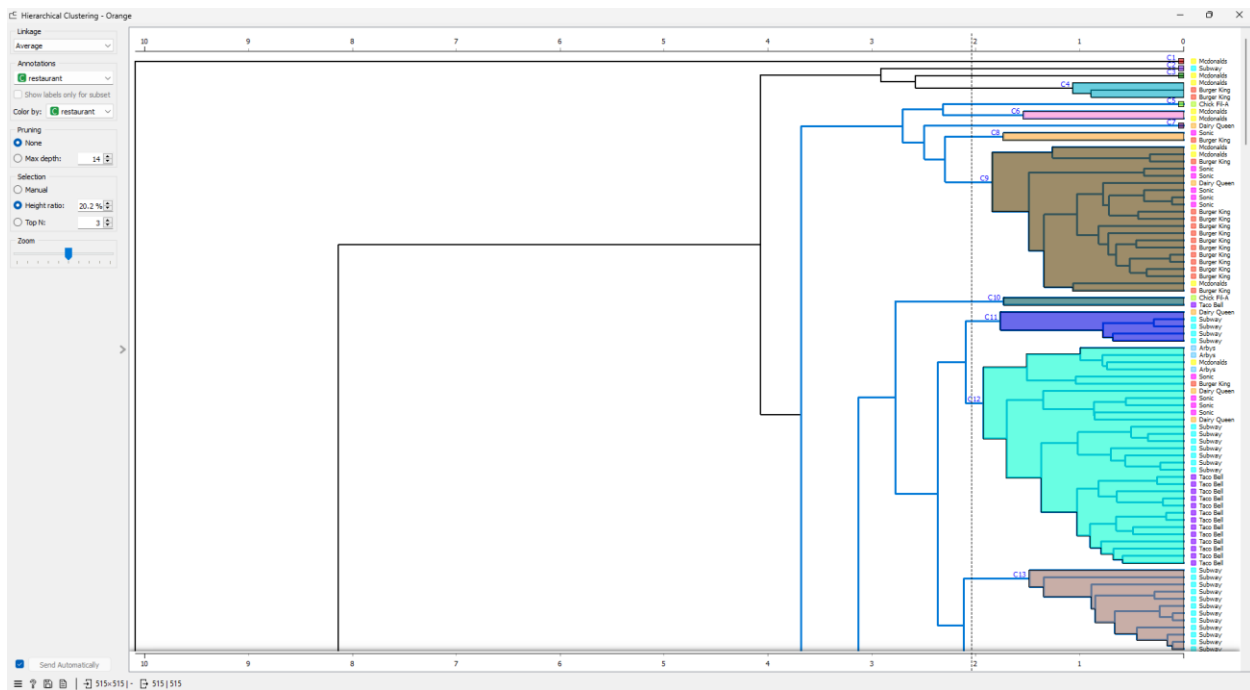
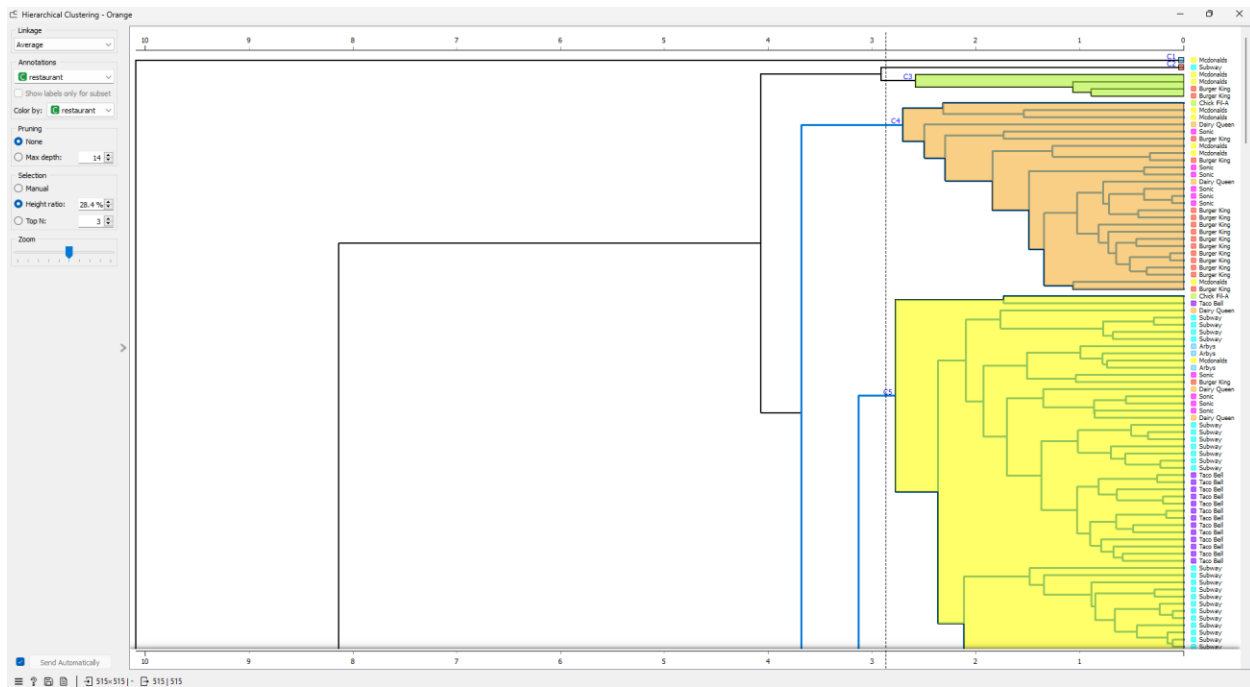


11.attēls hierarhiskā klasterizācija

11.attēlā redzami mani hiperparametru iestatījumi, attāluma metodi izvēlējos Eiklīda ar datu normalizāciju.



12.attēls atdalošā līnija 25.1%



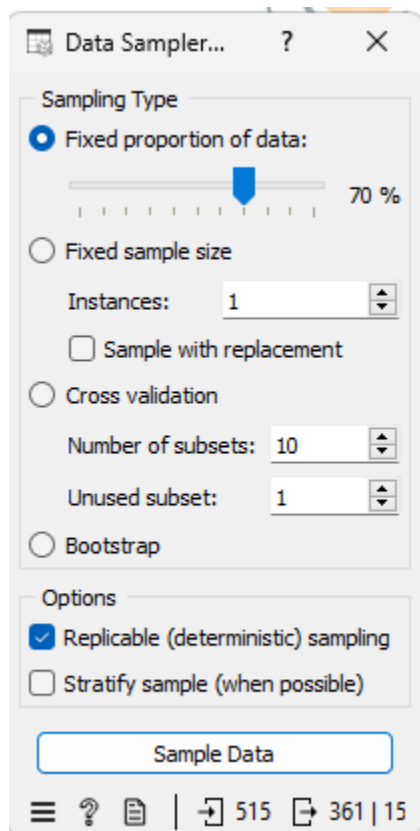
12.,13. un 14.attēlā var secināt, ka šo datu kopu ar hierarhisko klasterizāciju ir slikti atdalāmas klases. Eksperimentēju kustinot atdalošo līniju, kā arī mainīju iestatījumus, un pētīju klasteru skaitu un to saturu. Pēc eksperimentēšanas ir skaidri redzams, ka sadalot vairākos klasteros dažādas klases joprojām ir saliktas kopā vienā klasterā.

2.Daļas secinājumi

Veicot eksperimentēšanu ar abiem algoritmiem var secināt, ka ar hierarhisko klasterizāciju klases ir slikti atdalāmas, bet ar K-vidējo algoritmu klases ir samērā labi atdalāmas, ja klasteru skaits ir tikai 2, līdz ko klasteru skaits ir lielāks par 2, tad klases paliek slikti atdalāmas.

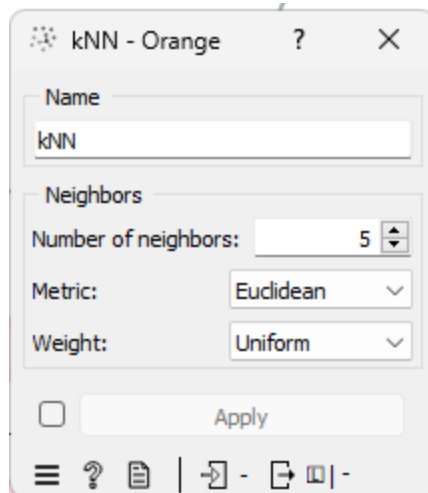
3. daļa – Pārraudzītā mašīnmācīšanās

Priekš praktiskā darba 3.daļas es izvēlējos KNN un SVM pārraudzītās mašīnmācīšanās algoritmus. Šos divus algoritmus es izvēlējos, jo tie man visvairāk ieinteresēja un likās, ka es šo algoritmu darbības principu vislabāk sapratu.



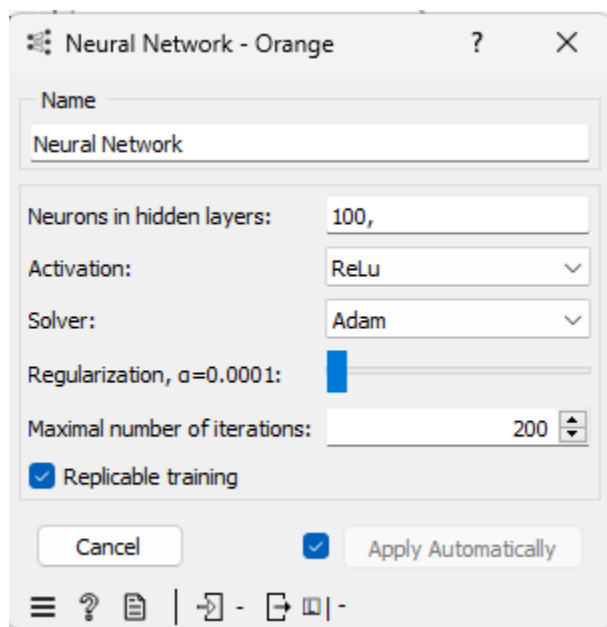
15.attēls data sampler

Algoritmu apmācībai es izvēlējos 70% procentus, un priekš testēšanas atstāju 30%. Atstāju pārējos iestatījumus pēc noklusējuma, kas attēlots 15.attēlā.



16.attēls Knn iestatījumi 1

16.attēlā redzams KNN algoritma hiperparametri, kurus es atstāju pēc noklusējuma priekš pirmā eksperimenta.



17.attēls Neironu tīkli 1

17.attēlā redzami neironu tīkla hiperparametru, kurus arī atstāju pēc noklusējuma priekš pirmā eksperimenta. Turpmāk arī 18.attēlā, kur redzami SVM algoritma hiperparametri, atstāšu pēc noklusējuma.

SVM - Orange

Name

SVM

SVM Type

☒ SVM Cost (C): 1.00

Regression loss epsilon (ϵ): 0.10

☐ v-SVM Regression cost (C): 1.00

Complexity bound (v): 0.50

Kernel

☐ Linear Kernel: $\exp(-g|x-y|^2)$

☐ Polynomial g: auto

☒ RBF

☐ Sigmoid

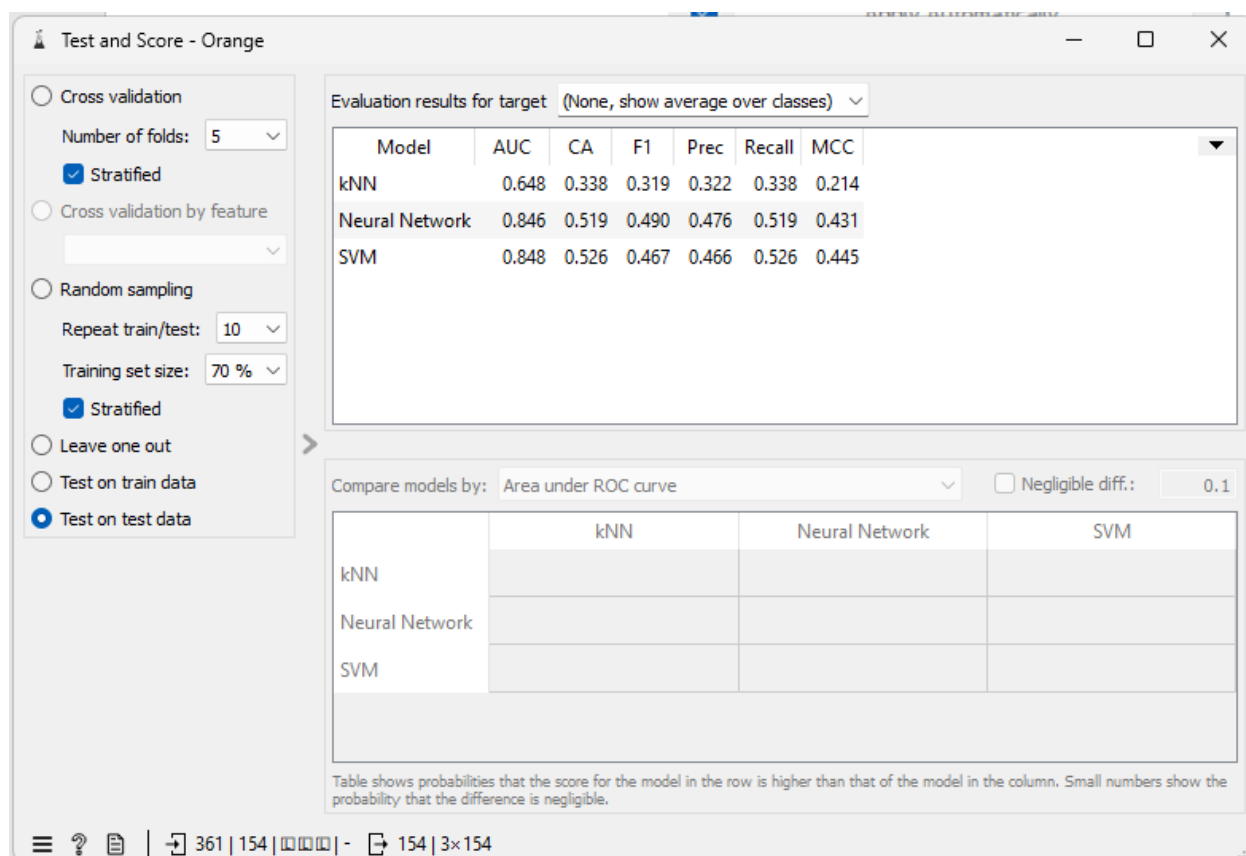
Optimization Parameters

Numerical tolerance: 0.0010

☒ Iteration limit: 100

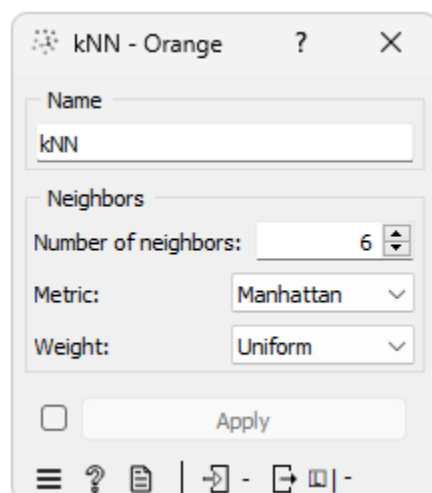
☒ Apply Automatically

18.attels SVM 1



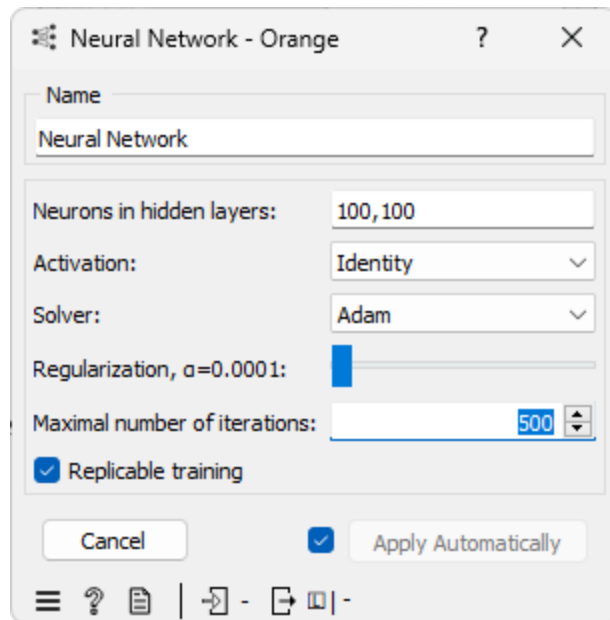
19.attēls testu rezultāts

19.attēlā ir redzami pašreizējie algoritmu testa rezultāti, var secināt, ka neironu tīkls pagaidām uzrāda vislabāko rezultātu.



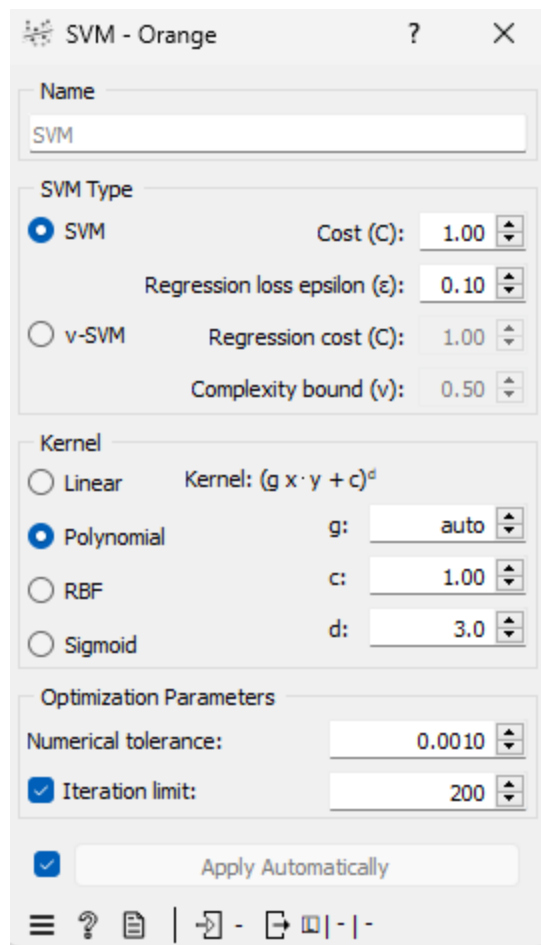
20.attēls kNN 2

20.attēlā pamainīju kaimiņu skaitu uz 6, un Metriku uz *Manhattan*.



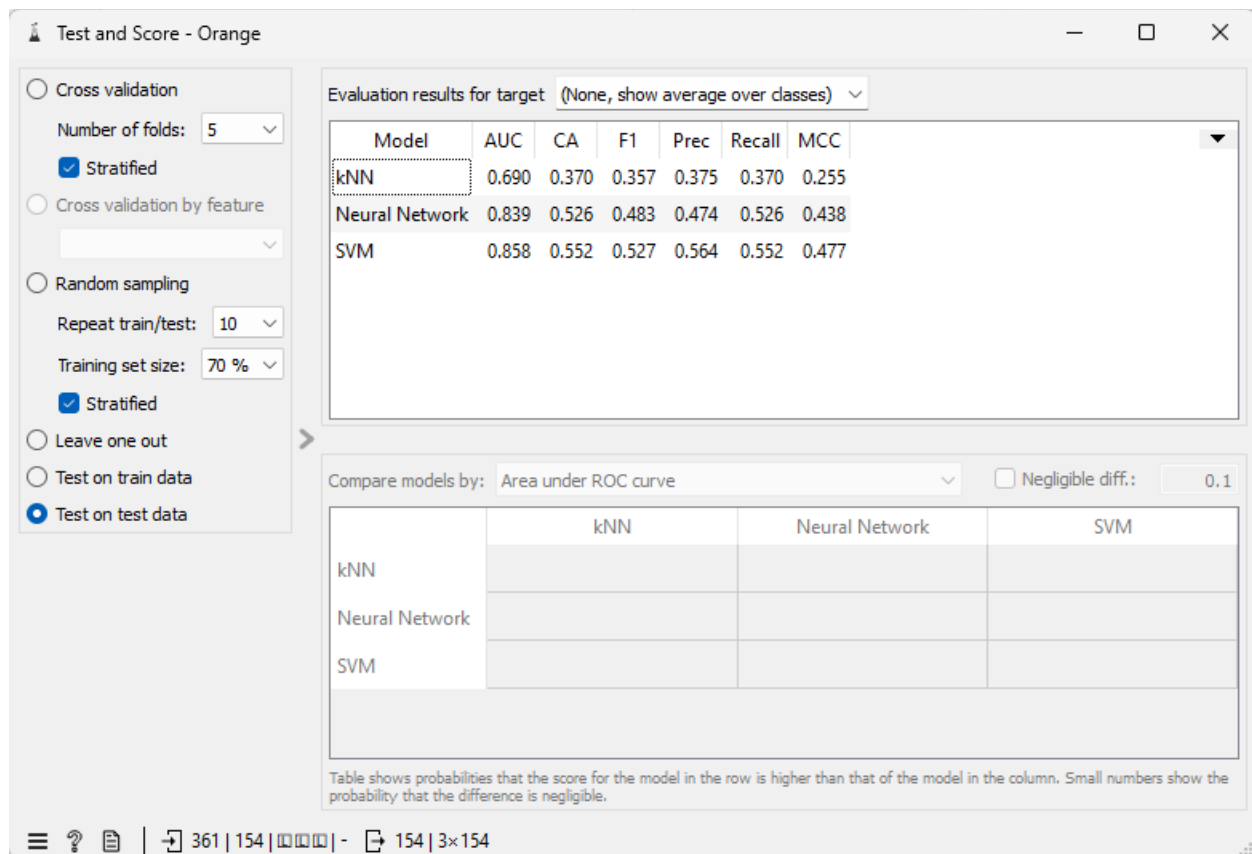
21.attēls neironu tīkls 2

21.attēlā pamainīju neironus slēptajā slānī uz 100,100, un aktivizāciju uz *Identity*, kā arī nomainīju iterāciju skaitu uz 500.



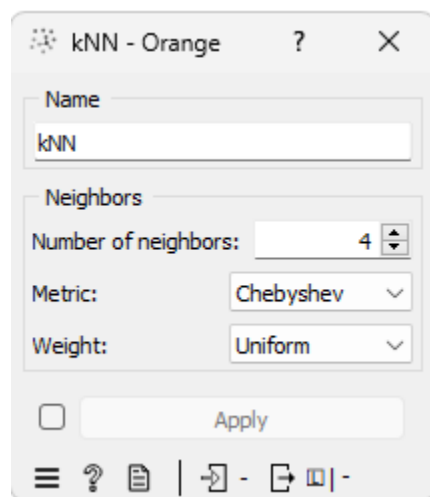
22.attēls SVM 2

22.attēla ir redzams, ka nomainīju SVM hiperparametrus: *Kernel* nomainīts uz *Polynomial*, iterāciju limits ir paaugstināts uz 200.



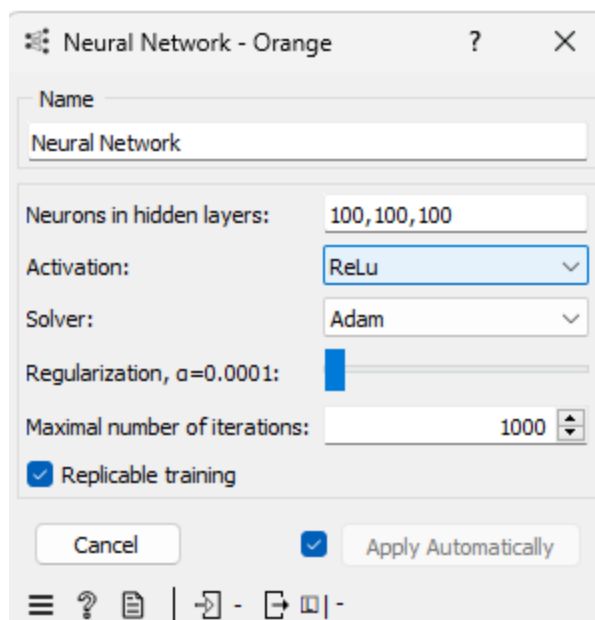
23.attēls eksperiments 2

23.attēlā var apskatīt otrā eksperimenta rezultātu. Neironu tīkla algoritms nav īpaši uzlabojies, bet var uzreiz pamanīt, ka kNN un SVM testa rezultāti ir nedaudz uzlabojušies.



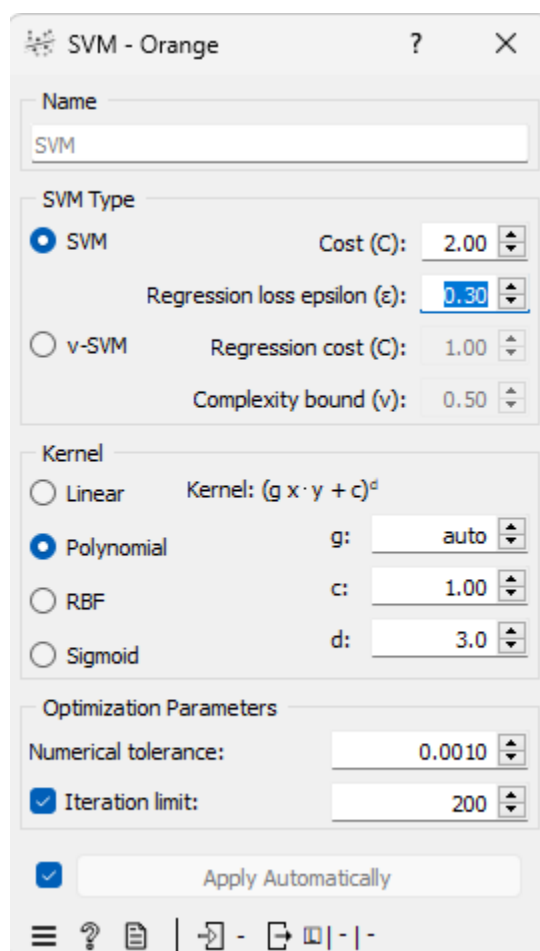
24.attēls kNN 3

24.attēlā redzams, ka kNN algoritma kaimiņu skaitu pazemināju uz 4, un metriku nomainīju uz *Chebyshev*.



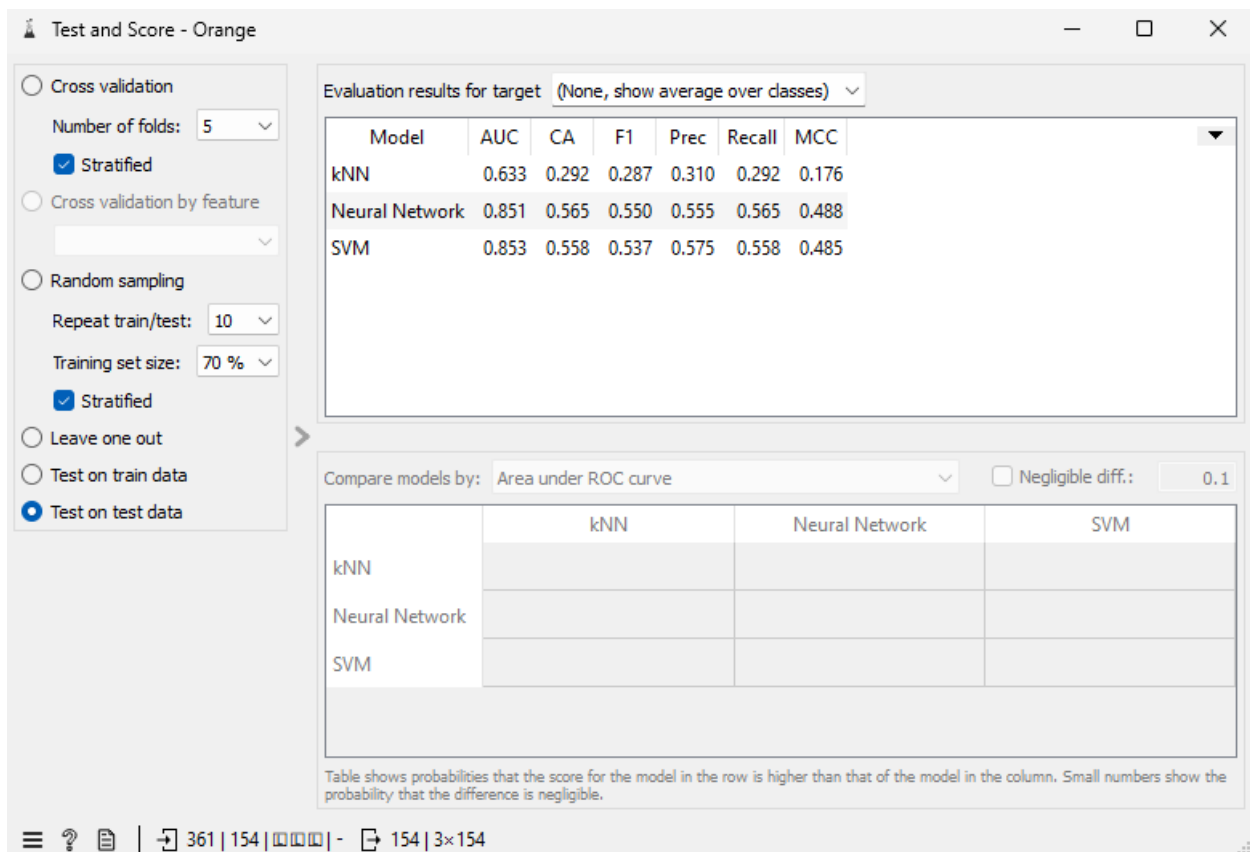
25.attēls Neironu tīkls 3

25.attēlā pamainīju neironus slēptajā slānī uz 100,100,100 un aktivizāciju uz *relu*, kā arī nomainīju iterāciju skaitu uz 1000.



26.attēls SVM 3

26.attēlā redzams, ka nomainīju hiperparametrus: *Cost* uz 2 un *Regression loss epsilon* uz 0.3



27.attēls eksperiments 3

3.daļas secinājumi

27.attēlā redzams trešā eksperimenta rezultāti. KNN algoritma rezultāti ir pasliktinājušies, vislabākie rezultāti priekš kNN algoritma bija otrajā eksperimentā. Neironu tīkla rezultāti ir vislabākie trešajā eksperimentā, kā arī KVM algoritma rezultāti ir bijuši vislabākie trešajā eksperimentā.

Izmantoto avotu saraksts

1. <https://www.kaggle.com/datasets/ulrikthgepedersen/fastfood-nutrition>
2. <https://estudijas.rtu.lv/course/view.php?id=252548>