RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte

2. praktiskais darbs
mācību priekšmetā
"Mākslīgā intelekta pamati"
Mašīnmācīšanās algoritmu lietojums
https://github.com/arturskovrigo/MIP_PD2

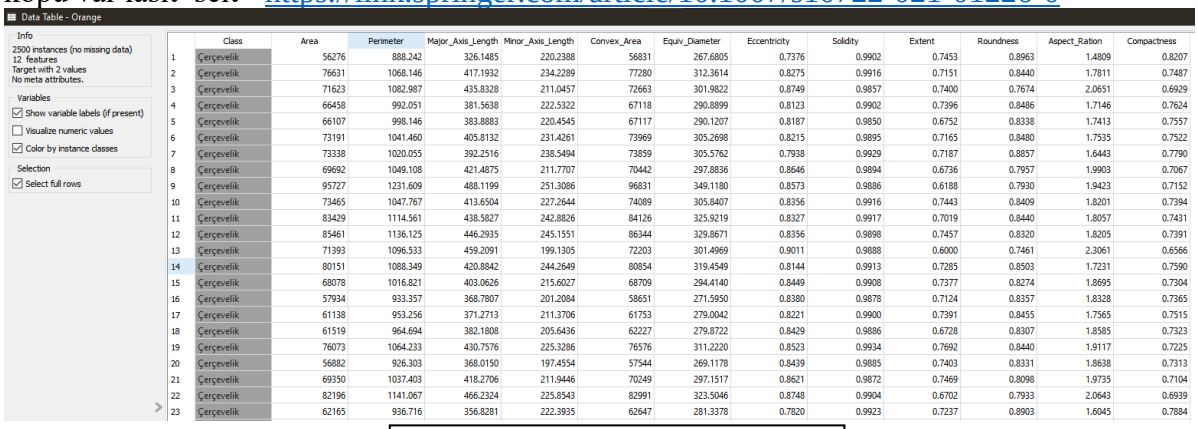
Izstrādāja: Artūrs Kovrigo
St. apl. Nr. 201RDB006
Pārbaudīja:_____

2021./22. māc. gads

I daļa - Datu pirmapstrāde/izpēte

Ir izvēlēta datu kopa no Kaggle, tās nosaukums ir “Pumpkin Seeds Dataset”, tajā ir dati par divu šķirņu ķirbju sēklu datiem, kas ir ņemti no Selcuk University, kurā tika no bildēm iegūtas dažādi atribūti, kas apraksta to formas, kopā ir 12 skaitliski atribūti un viens kategoriskais – mērķa atribūts.

Tajā ir 2500 datu objektu, kas iedalās divās klasēs – Çerçevelik (Turpmāk tekstā pirmā klase) un Ürgüp Sivrisi(Turpmāk tekstā otrā klase), kas ir attiecīgās sēklas šķirnes, pirmajai 1300, otrajai 1200 ieraksti. Katram datu punktam ir 12 skaitliski atribūti – Area, Perimeter, Major_Axis_Length, Minor_Axis_length, Convex_Area, Equiv_Diameter, Eccentricity, Solidity, Extent, Roundness, Aspect_Ration un Compactness. Vairāk par datu kopu var lasīt šeit - https://link.springer.com/article/10.1007/s10722-021-01226-0


Data Table - Orange

Info
2500 instances (no missing data)
12 features
Target with 2 values
No meta attributes.

Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes

Selection
Select full rows

	Class	Area	Perimeter	Major_Axis_Length	Minor_Axis_Length	Convex_Area	Equiv_Diameter	Eccentricity	Solidity	Extent	Roundness	Aspect_Ration	Compactness
1	Çerçevelik	56276	888.242	326.1485	220.2388	56831	267.6805	0.7376	0.9902	0.7453	0.8963	1.4809	0.8207
2	Çerçevelik	76631	1068.146	417.1932	234.2289	77280	312.3614	0.8275	0.9916	0.7151	0.8440	1.7811	0.7487
3	Çerçevelik	71623	1082.987	435.8328	211.0457	72663	301.9822	0.8749	0.9857	0.7400	0.7674	2.0651	0.6929
4	Çerçevelik	66458	992.051	381.5638	222.5322	67118	290.8899	0.8123	0.9902	0.7396	0.8486	1.7146	0.7624
5	Çerçevelik	66107	998.146	383.8883	220.4545	67117	290.1207	0.8187	0.9850	0.6752	0.8338	1.7413	0.7557
6	Çerçevelik	73191	1041.460	405.8132	231.4261	73969	305.2698	0.8215	0.9895	0.7165	0.8480	1.7535	0.7522
7	Çerçevelik	73338	1020.055	392.2516	238.5494	73859	305.5762	0.7938	0.9929	0.7187	0.8857	1.6443	0.7790
8	Çerçevelik	69692	1049.108	421.4875	211.7707	70442	297.8836	0.8646	0.9894	0.6736	0.7957	1.9903	0.7067
9	Çerçevelik	95727	1231.609	488.1199	251.3086	96831	349.1180	0.8573	0.9886	0.6188	0.7930	1.9423	0.7152
10	Çerçevelik	73465	1047.767	413.6504	227.2644	74089	305.8407	0.8356	0.9916	0.7443	0.8409	1.8201	0.7394
11	Çerçevelik	83429	1114.561	438.5827	242.8826	84126	325.9219	0.8327	0.9917	0.7019	0.8440	1.8057	0.7431
12	Çerçevelik	85461	1136.125	446.2935	245.1551	86344	329.8671	0.8356	0.9898	0.7457	0.8320	1.8205	0.7391
13	Çerçevelik	71393	1096.533	459.2091	199.1305	72203	301.4969	0.9011	0.9888	0.6000	0.7461	2.3061	0.6566
14	Çerçevelik	80151	1088.349	420.8842	244.2649	80854	319.4549	0.8144	0.9913	0.7285	0.8503	1.7231	0.7590
15	Çerçevelik	68078	1016.821	403.0626	215.6027	68709	294.4140	0.8449	0.9908	0.7377	0.8274	1.8695	0.7304
16	Çerçevelik	57934	933.357	368.7807	201.2084	58651	271.5950	0.8380	0.9878	0.7124	0.8357	1.8328	0.7365
17	Çerçevelik	61138	953.256	371.2713	211.3706	61753	279.0042	0.8221	0.9900	0.7391	0.8455	1.7565	0.7515
18	Çerçevelik	61519	964.694	382.1808	205.6436	62227	279.8722	0.8429	0.9886	0.6728	0.8307	1.8585	0.7323
19	Çerçevelik	76073	1064.233	430.7576	225.3286	76576	311.2220	0.8523	0.9934	0.7692	0.8440	1.9117	0.7225
20	Çerçevelik	56882	926.303	368.0150	197.4554	57544	269.1178	0.8439	0.9885	0.7403	0.8331	1.8638	0.7313
21	Çerçevelik	69350	1037.403	418.2706	211.9446	70249	297.1517	0.8621	0.9872	0.7469	0.8098	1.9735	0.7104
22	Çerçevelik	82196	1141.067	466.2324	225.8543	82991	323.5046	0.8748	0.9904	0.6702	0.7933	2.0643	0.6939
23	Çerçevelik	62165	936.716	356.8281	222.3935	62647	281.3378	0.7820	0.9923	0.7237	0.8903	1.6045	0.7884

1. attēls
Pirmie ieraksti datu kopā

Parametru raksturvērtības

Nosaukums	Vidējais	Minimums	Maksimums
Area	80658.22	47939	136574
Perimeter	1130.27	868.48	1559.45
Major_Axis_Length	456.60	320.84	661.91
Minor_Axis_Length	225.79	152.17	305.82
Convex_Area	81508.08	48366	138384
Equiv_Diameter	319.33	247.05	417.00
Eccentricity	0.86	0.49	0.95
Solidity	0.99	0.91	0.99
Extent	0.69	0.46	0.83
Roundness	0.79	0.55	0.94
Aspect_Ration	2.04	1.14	3.14
Compactness	0.70	0.56	0.90

1. tabula

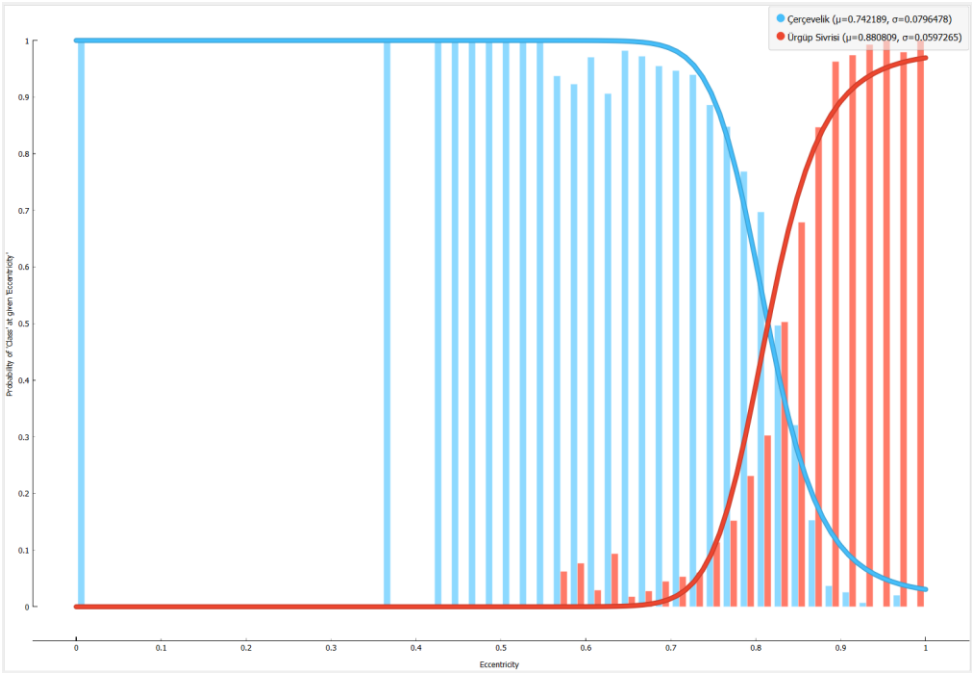
Parametru apraksts

Nosaukums	Apraksts
Area	Sēklas laukums
Perimeter	Sēklas perimetrs
Major_Axis_Length	Sēklas garums
Minor_Axis_Length	Sēklas platums
Convex_Area	Izliektas, apvilktas fomas laukums
Equiv_Diameter	Kvadrātsakne no sēklas laukuma reizināta ar 4 un izdalīta ar π
Eccentricity	Sēklas formas ekscentriskums
Solidity	Laukums dalīts ar Izliektas, apvilktas formas laukumu
Extent	Laukums dalīts ar apvilkta taisnstūra laukumu
Roundness	Sēklas ovālums ignorējot izkropļojumus
Aspect_Ration	Sēklas garumas dalīts ar platumu
Compactness	Kompaktums, jeb sēklas laukums izdalīts ar apvilktas

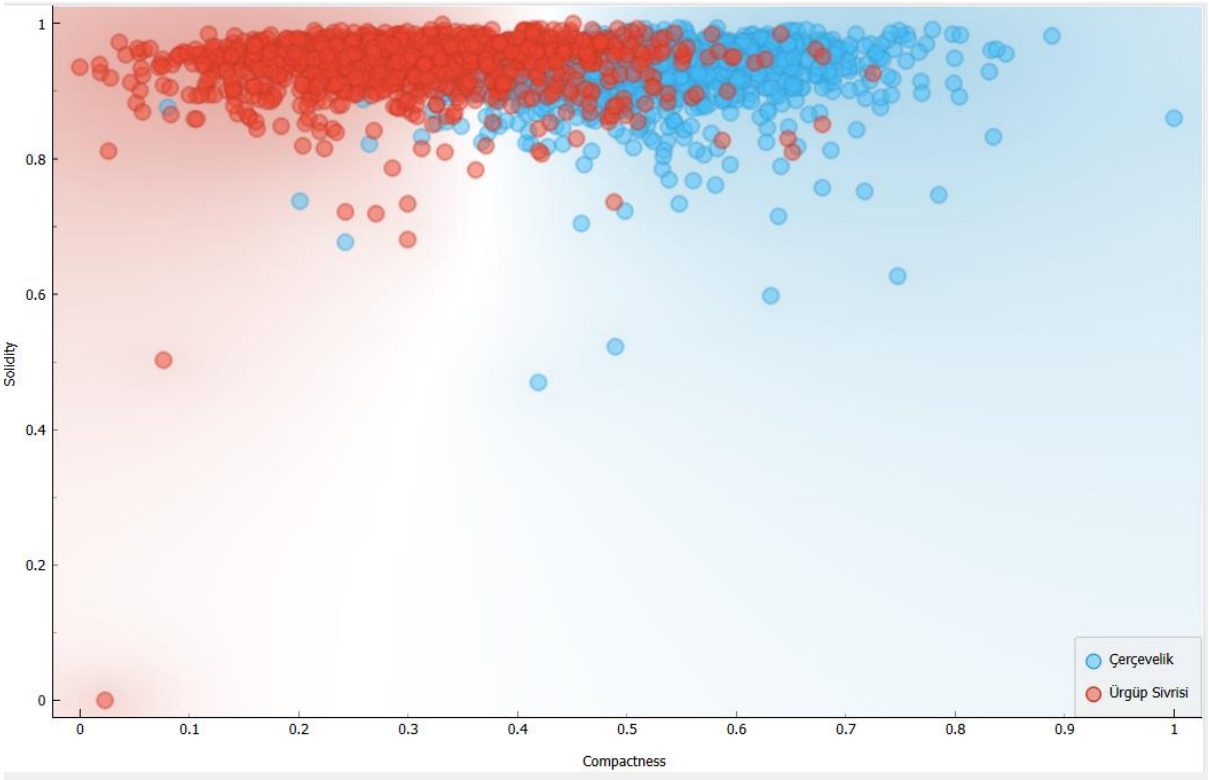
2. tabula

Pirmajā tabulā redzams, ka daži datu objekti ir vērtībās ap 1, kamēr citi ir virs 10^5 kā rezultātā tika izlemts veikt datu normalizāciju. Izvēlēts normalizēt intervālā [0,1].

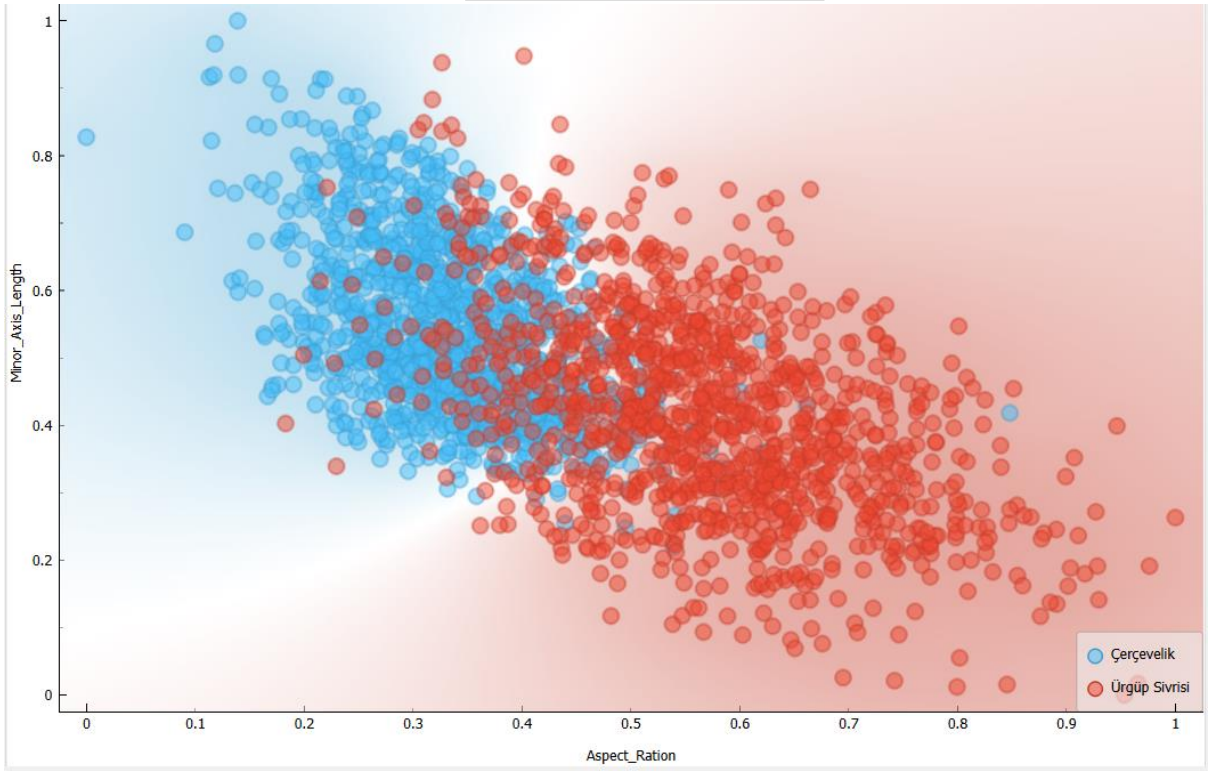
Dati nevienā apskatītajā atribūtā un nevienā atribūtu pārī neatdalās pilnīgi, tas ir, visos starp klasēm ir nozīmīga pārklāšanās, bet vienlaikus, lielākajā daļā atribūtu ir arī skaidri redzamas atšķirības. Starp atsevišķiem atribūtiem, Eccentricity bija viens no labākajiem, savukārt no pāriem labākie bija Compactness atkarībā no Solidity, Aspect_Ration no Minor_Axis_Length un Eccentricity no Roundness. Šo sakarību vizualizācijas redzamas 2.;3.;4. un 5. attēlos



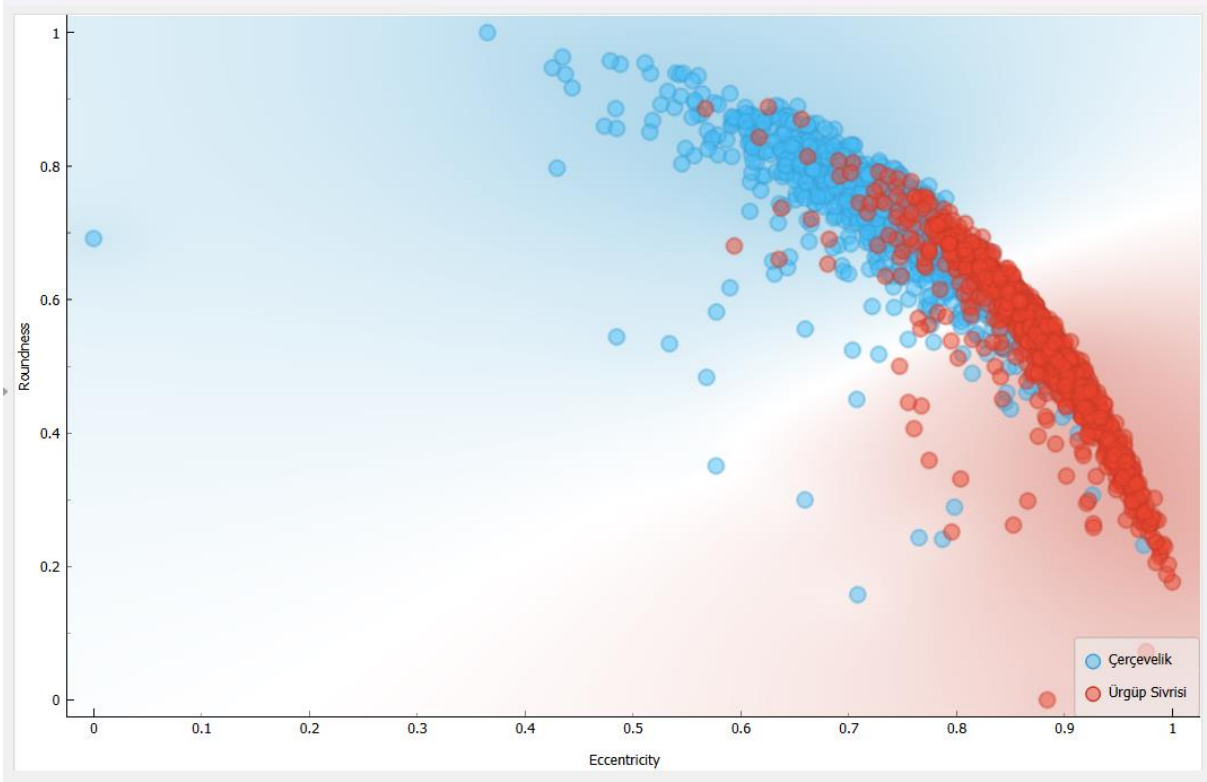
2. attēls



3. attēls



4. attēls



5. attēls

II daļa - Nepārraudzītā mašīnmācīšanās

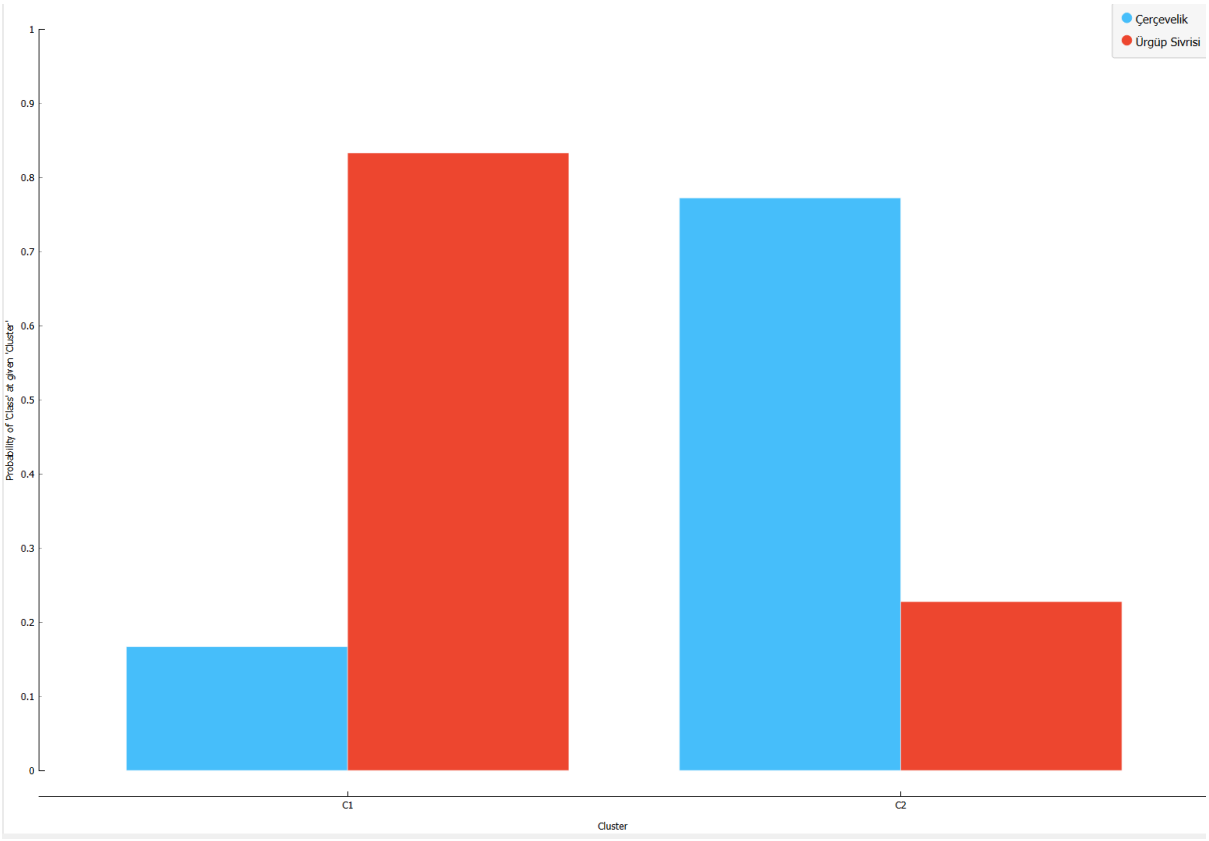
k-Means

K-Means ir 4 hiperparametri – klasteru skaits, kurš nosaka, cik būs klasteri; re-runs, kurš nosaka, cik reizes mēģināt klasterizēt līdz paliek pie esošās maksimālās silueta koeficienta vērtības; maksimālais iterāciju skaits, kurš nosaka maksimālo centroīdu pārrēķināšanas skaitu; inicializācijas metode, kas ir vai nu nejauša, vai KMeans++, kura pēc pirmā centroīda nejaušas izvēles, otro izvēlās atkarībā no tā attāluma līdz pirmajam – jo tālāk, jo lielāka iespēja izvēlēties.

To izmantojot maksimālā silueta vērtība sasniegta pie diviem klasteriem, kas sakrīt ar mērķa atribūta klašu skaitu, taču tas ir diezgan mazs – 0.3. Apskatot sadalījuma grafiku var novērot, ka pirmais klasteris sastāv no 83.6% no otrās klases, un otrais klasteris 75.5% no pirmās klases. Šos klasterus izmantojot klasifikācijai pareizi klasificēti būtu 78.7% ierakstu

Nepārraudzītā mašīnmācīšanās – k-Means		
Maksimālais iterāciju skaits	Klasteru skaits	Silueta koeficients
5	2	0.307
	3	0.290
	4	0.269
	5	0.239
25	2	0.308
	3	0.288
	4	0.271
	5	0.240
125	2	0.308
	3	0.289
	4	0.271
	5	0.239

3. tabula



6. attēls
Klašu sadalījums klasteros

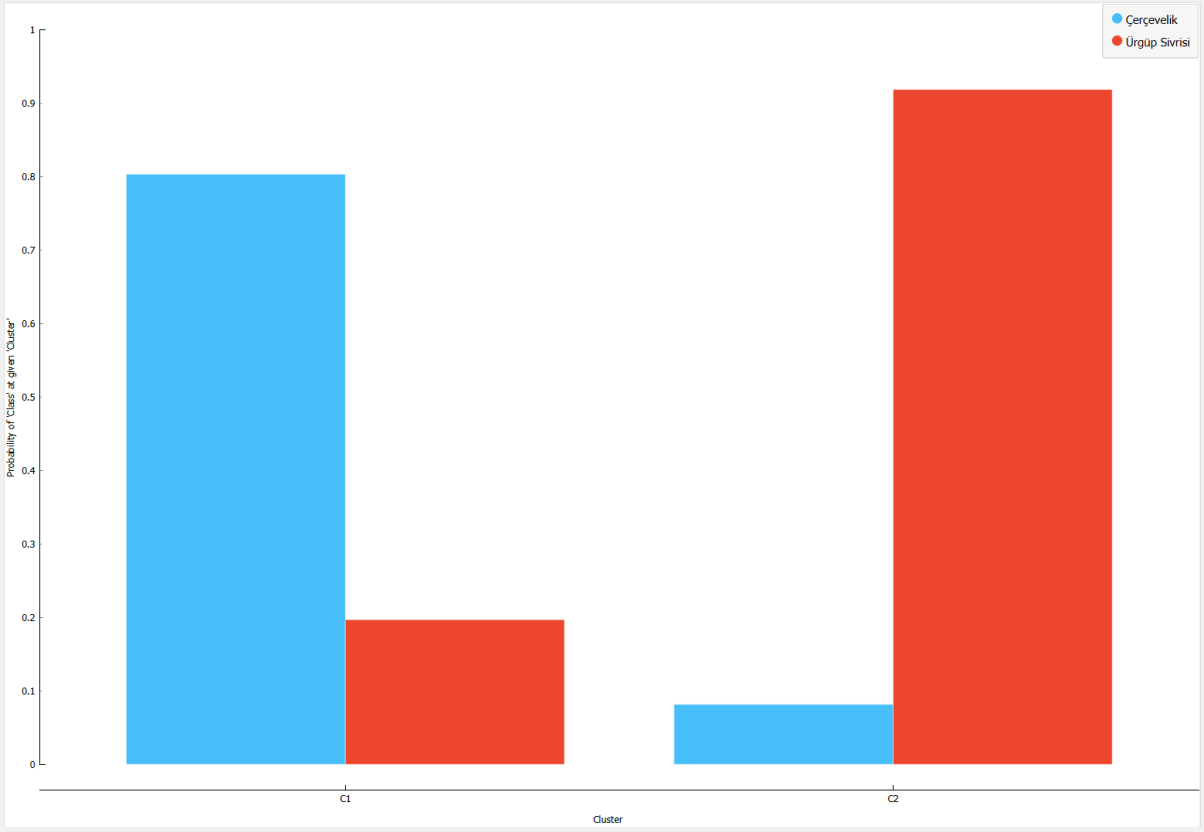
Hierarhiskā klasterēšana	
Attāluma metrika	Precizitāte
Eiklīda	80.88%
Manhatanas	70.56%
Kosīnusa	72.30%*
Spīrmena	84.84%
Absolūtā Spīrmena	83.32%*
Pīrsona	77.96%
Hamminga	N/A**

4. tabula

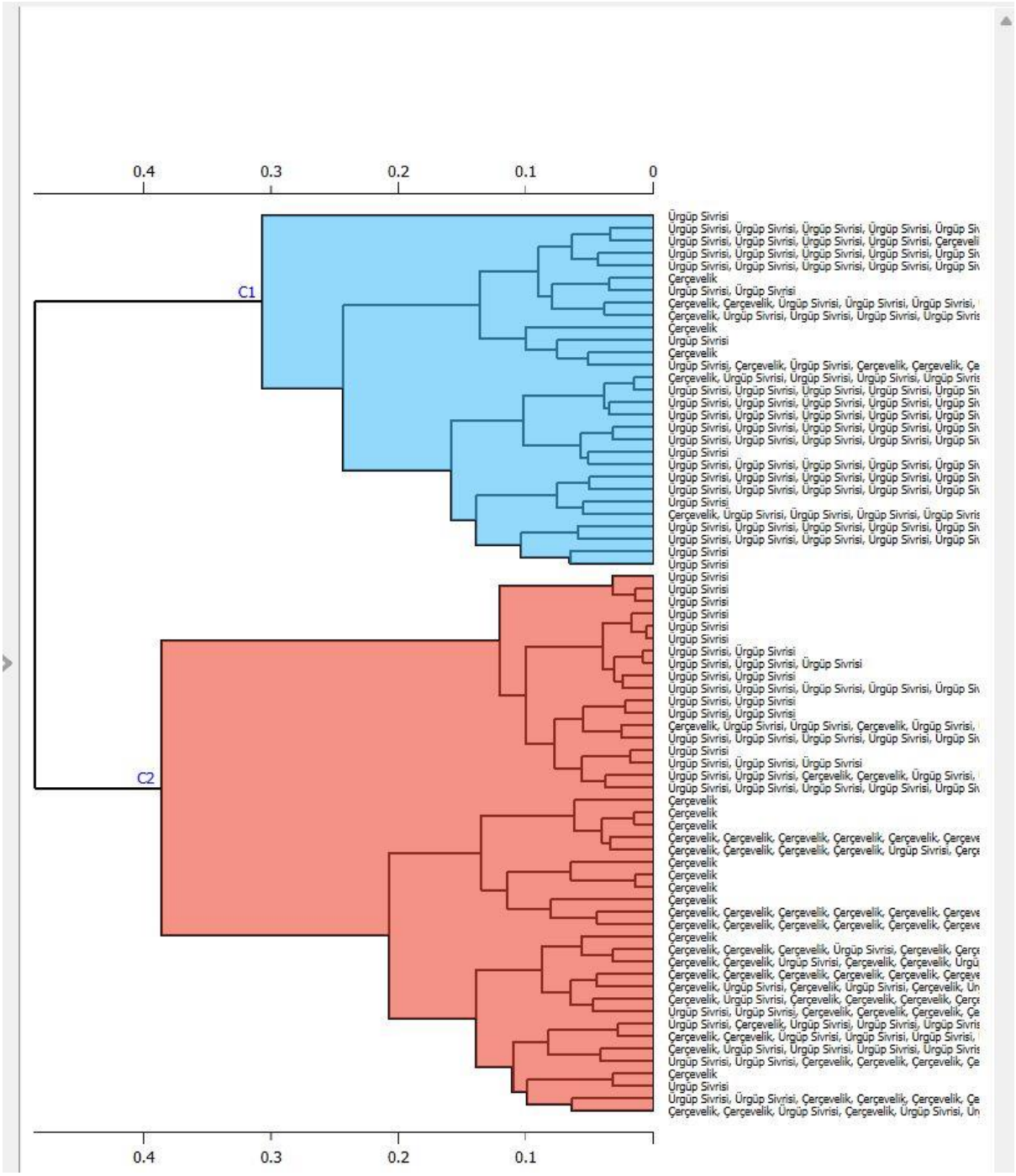
*Pirmie klasteri bija pāris datu objektu lieli, precizitāte iegūta palielinot klasteru skaitu līdz ir 2 lieli klasteri, un mazos klasterus uzskatot par nepareiziem
** Hamminga attālumam neizdevās iegūt precizitāti, jo palielinot klasteru skaitu par līdz 15, bija viens izteikti liels klasteris un pārējie mazi, līdz brīdim, kad atlikušais sasniedz tik pat mazu apjomu, kā pārējie

Hierarhiskā klasterēšana

Hierarhiskā klasterēšana izmanto divus parametrus – klasteru skaitu, kas atkal nosaka, cik grupās klasterizēt, un attāluma aprēķina metriku. Vislabākā izrādījās Spīrmena attāluma metrika, ar kuru pirmajā klasterī bija 80.32%, jeb 1220 pirmās klases ieraksti, un otrajā klasterī 91.85%, jeb 901 otrās klases ieraksts, kas kopā dotu 84.84% precizitāti, ja šos klasterus izmantotu klasifikācijai.



7. attēls
Klašu sadalījums klasteros



8. attēls
Koka vizualizācija

III daļa - Pārraudzītā mašīnmācīšanās

Tika izvēlēts nejauša meža algoritms, SVM un neirālais tīkls, lai mēģinātu atrast kādu veidu, kā uzticami klasificēt datus. Katrai metodei veikta eksperimentācija ar dažādām hiperparametru kombinācijām, lai iegūtu maksimālo precizitāti.

Neirālie Tīkli

Daudzslāņu mākslīgo neironu tīkli būtībā sastāv no 3 dažādiem slāņu tipiem – ieejas slānis, slēptais slānis/slāņi, izvades slānis. Ieejas slānis šajā kontekstā ir neironu kopums, kurā katrs neirons atbilst noteiktam atribūtam. Slēptie slāņi sastāv no vienā virzienā savstarpēji savienotiem neironiem līdz tie nonāk līdz izejas slānim. Salīdzinot izejas slāņa rezultātus ar vēlamo vērtību attiecīgi tiek koriģeti svari neironos ar mērķi tuvināt tīkla pareģojumu šai vēlamajai vērtībai.

Orange rīkā ir pieejami 5 galvenie hiperparametri, tīkla struktūra, kas sastāv no neironu skaita katrā slēptajā slānī, slāņu skaita, aktivācijas funkcijas, solvers, jeb optimizētājs, kurš nosaka kā notiek tīkla trenēšanās un iterāciju, jeb epohu skaits, kas nosaka, cik reizes visi dati tīklam tiek padoti.

No šiem vislabāko precizitāti sasniedza tīkls ar 500 slēptajiem neironiem katrā no 2 slēptajiem slāņiem, izmantojot 100 iterācijas un identitātes aktivācijas funkciju. Vēl datos varēja novērot pārtenēšanos, 3. eksperimentā palielinot iterāciju skaitu samazinājās precizitāte uz testa datiem.

Neironu skaits slēptajos slāņos	Slēpto slāņu skaits	Iterāciju skaits	Aktivācijas funkcija	AUC	CA	F1	Precision	Recall
100	2	100	ReLu	0.939	0.860	0.860	0.860	0.860
500	2	100	ReLu	0.926	0.868	0.868	0.868	0.868
500	2	500	ReLu	0.914	0.856	0.856	0.857	0.856
100	3	100	ReLu	0.926	0.860	0.860	0.860	0.860
500	3	100	ReLu	0.907	0.844	0.846	0.844	0.844
500	2	100	tanh	0.936	0.868	0.868	0.869	0.868
500	2	100	Loģistiskā	0.933	0.860	0.859	0.862	0.860
500	2	100	Identitātes	0.934	0.884	0.884	**0.884**	0.884
1000	2	100	Identitātes	0.933	0.868	0.868	0.868	0.734

5. tabula

Nejauša meža algoritms

Nejauša meža algoritms, uzģenerē lietotāja izvēlētu skaitu lēmumu kokus nejaušām datu apakškopām, un par pareizu uzskata vispopulārāko variantu starp izveidoto koku prognozēm. Tam ir tikai viens svarīgākais hipermarametrs – koku skaits, lielāks skaits ir vairāk koki, ir mazāk svārstības, taču eksperimentējot arī ar 100 un 1000 kokiem, tās bija diezgan nozīmīgas, un palaižot atkārtoti nozīmīgi mainījās rezultāti. Labākais ko ar šo algoritmu izdevās sasniegt ir 87.2% precizitāte, un pateicoties lielajām svārstībām, tas tika sasniegts gan ar 10, gan 100 koku eksperimentiem

Model	Koku skaits	AUC	CA	F1	Precision	Recall
Random fores	10	0.921	0.872	0.872	**0.872**	0.872
	50	0.929	0.868	0.868	0.868	0.868
	100	0.926	0.872	0.872	**0.872**	0.872
	1000	0.925	0.864	0.864	0.864	0.864

6. tabula

SVM

SVM algoritms balstās uz robežu novilkšanu starp klasēm izmantojot klašu ekstrēmus. Tā vietā, lai izvēlētos nejaušu līniju, kura atdala maksimāli daudz, tiek izmantoti atbalsta vektori, kas atrod, kur beidzās viena kopa un kur sākās otra, un robežu ieliek tieši pa vidu tiem. Pieejamie hiperparametri ir SVM tips, starp kuriem galvenāatšķirība ir vektoru skaits, v-SVM ir lielāks skaits, kas samazina variāciju, kodols, kurš maina izmantotās funkcijas, oriģinālais ir lineārs, taču bieži labāk strādā citas funkcijas, šajos eksperimentos paturēts Orange noklusējuma variants – RBF, jeb $\exp(-G|X-Y|^2)$, kā arī iterāciju limits.

SVM	vSVM; Iterāciju limits = 1000	0.823	0.716	0.716	0.716	0.428
	SVM; Iterāciju limits = 1000	0.929	0.868	0.868	**0.868**	0.868
	vSVM; Iterāciju limits = 100000	0.866	0.784	0.783	0.784	0.784
	SVM; Iterāciju limits = 100000	0.929	0.868	0.868	**0.868**	0.868

7. tabula

Secinājumi par pārraudzīto mašīnmāčišanos

Kopumā labākais sniegums bija neirālajam tīklam 500 neironiem divos slēptajos slāņos, 100 ierācijām un identitātes aktivācijas funkciju, un tas bija 88.4%, kas tomēr ir tālu no datu autoru veiktajā petījumā sasniegtajiem 92.77% ar SVM un 92.31% ar neirālo tīklu.

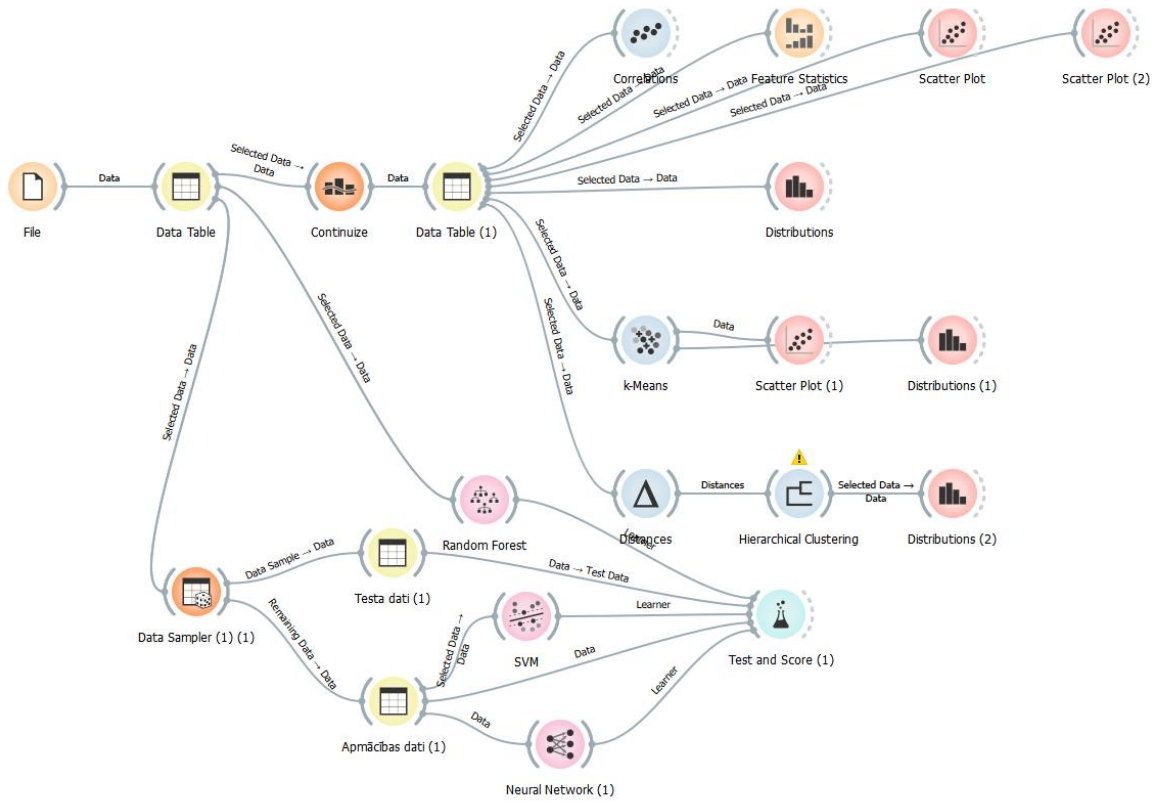
Izmantotie informācijas avoti

KOKLU, M., SARIGIL, S., & OZBEK, O. (2021). The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genetic Resources and Crop Evolution, 68(7), 2713-2726. Doi: https://doi.org/10.1007/s10722-021-01226-0

ORANGE DATA MINING (2015). *Orange Visual Programming* (skatīts 2023, 11. maijā).https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html

Graupe, D. *Principles Of Artificial Neural Networks (3rd Edition)*. Čikāga: World Scientific Publishing Company, 2013. 363 lpp. ISBN 978-981-4522-73-1

Pielikums



Vispārējs darba izvietojums Orange vidē