RĪGAS TEHNISKĀ UNIVERSITĀTE

RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte

2.praktiskais darbsmācību priekšmetā

“Mākslīgā intelekta pamati”

Mašīnmācīšanās algoritmu lietojums

Izstrādāja: Juris Joņins
1.grupa 201RDB004

Links uz failiem:
https://github.com/NoHanded/MI_2praktiskais

2022./23. māc. gads

Saturs

Datu apstrāde un izpēte 3
Wine Quality Data Set 3
Nepārraudzīta mašīnmācīšanās 9
K-Means 9
Hierarchical clustering 15
Pārraudzīta mašīnmācīšanās 18
Mākslīgie neironu tīkli 18
Loģistiskā regresija 21
kNN 24
Secinājumi 27
References 28

Datu apstrāde un izpēte

Wine Quality Data Set

Šajā darbā tiek apskatīta datu kopa, kurā iekļauti dati par baltvīnu. [1] Dati tikuši ziedoti 2009.gada 7.oktobrī. Autors – Paulo Cortez, Univesity of Minho, Portugāle. Šī darba mērķis ietver vīna apskati attiecībā ar vīna fizikāli ķīmiskajām īpašībām. Tiek apkopoti baltvīna "Vinho Verde" dati. Tiek iekļauti tikai dati par to ķīmisko sastāvu, nevis par cenu, vīnogām vai citām ar vīnu saistītām lietām. Vīnam tiek piešķirti 12 atribūti:

1. Fixed acidity – fiksētais skābju daudzums katrā vīnā
2. Volatile acidity – gaistošais skābums, skābes gāzes veidā
3. Citric acid – cintonskābes daudzums
4. Residual sugar – cukura daudzums vīnā
5. Chlorides – hlorīds
6. Free sulfur dioxide – brīvā sēra dioksīda daudzums
7. Total sulfur dioxide – kopējais sēra dioksīda daudzums
8. Density - blīvums
9. pH
10. Sulphates - sulfāti
11. Alcohol – alkohola saturs (vīna stiprums)
12. Quality - kvalitāte

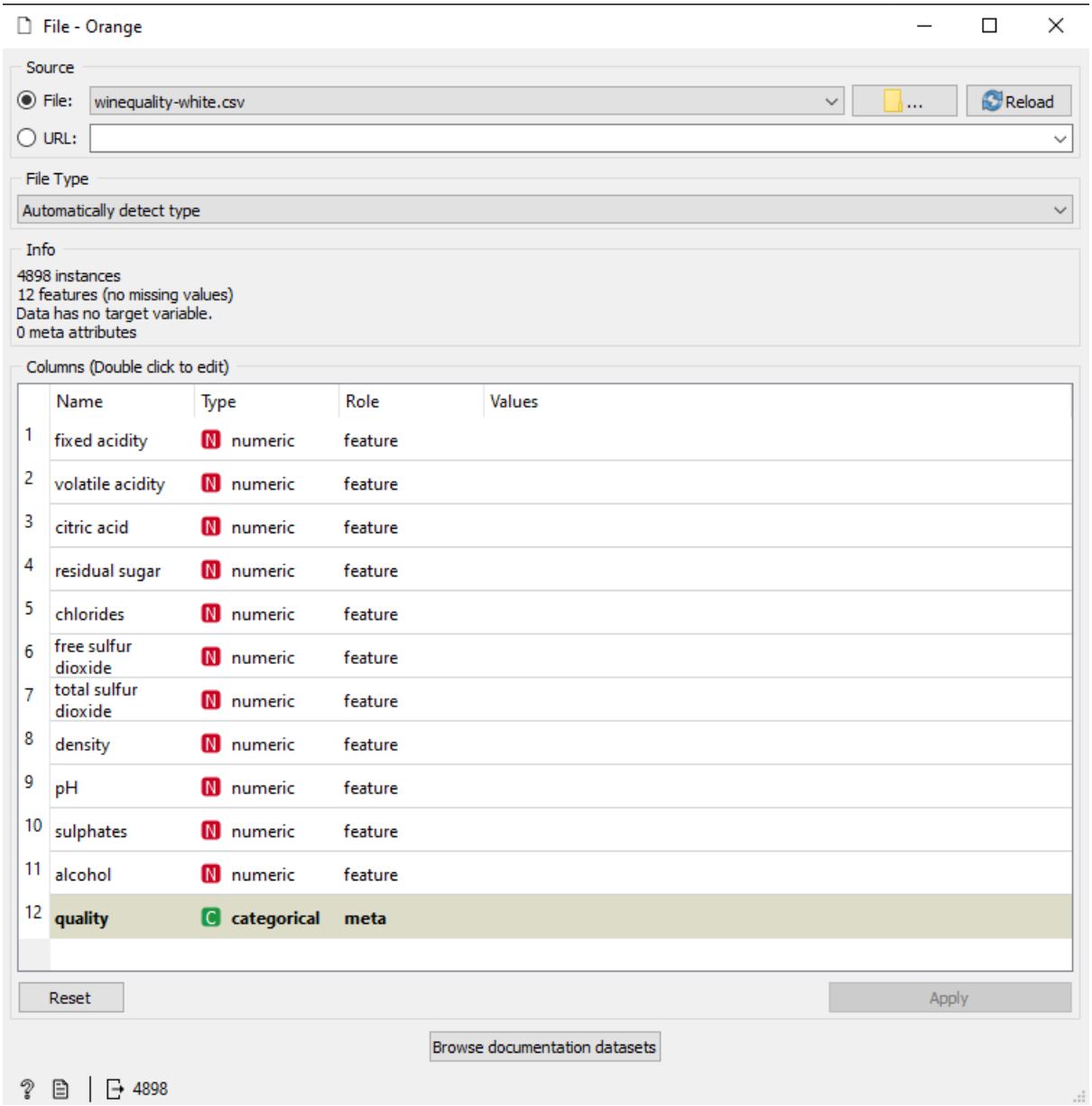
Pēc pirmajiem 11 parametriem tiek aplūkoti 4898 baltvīni, un tiek izvadīta tā kvalitāte – 12.parametrs

Datu kopa aplūko vīna ķīmosko sastāvu Portugāles ziemeļu daļā, šis pētījums veikts augošās vīna intereses dēļ. Kvalifikācijas process ir būtiska lieta vīna liešanā un šis darbs mērķē uz vīna testēšanas rezultātiem [2].

Nr.p.k	Parametrs	Tips	Diapozons	Vidējā vērtība	Moda	Mediāna	Dispersija
1.	Fixed acidity	skaitlisks	3.80 – 14.20	6.8548	6.8	6.8	0.1231
2.	Volatile acidity	Skaitlisks	0.08 – 1.10	0.27824	0.280	0.260	0.36222
3.	Citric acid	Skaitlisks	0.00 – 1.66	0.3342	0.30	0.32	0.3621
4.	Residual sugar	Skaitlisks	0.60 – 65.80	6.3914	1.2	5.2	0.7935
5.	Chlorides	Skaitlisks	0.009 – 0.346	0.04577	0.044	0.043	0.47727
6.	Free sulfur dioxide	Skaitlisks	2.0 – 289.0	35.308	29.0	34.0	0.482
7.	Total sulfur dioxide	Skaitlisks	9.0 – 440.0	138.361	111.0	134.0	0.307
8.	Density	Skaitlisks	0.98711 –	0.994027	0.992	0.99374	0.00300857

			1.03898				
9.	pH	Skaitlisks	2.72 – 3.82	3.1883	3.14	3.18	0.0474
10.	Sulphates	Skaitlisks	0.22 – 1.08	0.4898	0.5	0.47	0.2330
11.	Alcohol	Skaitlisks	8.0 – 14.2	10.5143	9.4	10.4	0.117031
12.	Quality	Kategorisks	3 - 9	5.88	6	6	0.15

1.1.tabula Datu kopas informācija

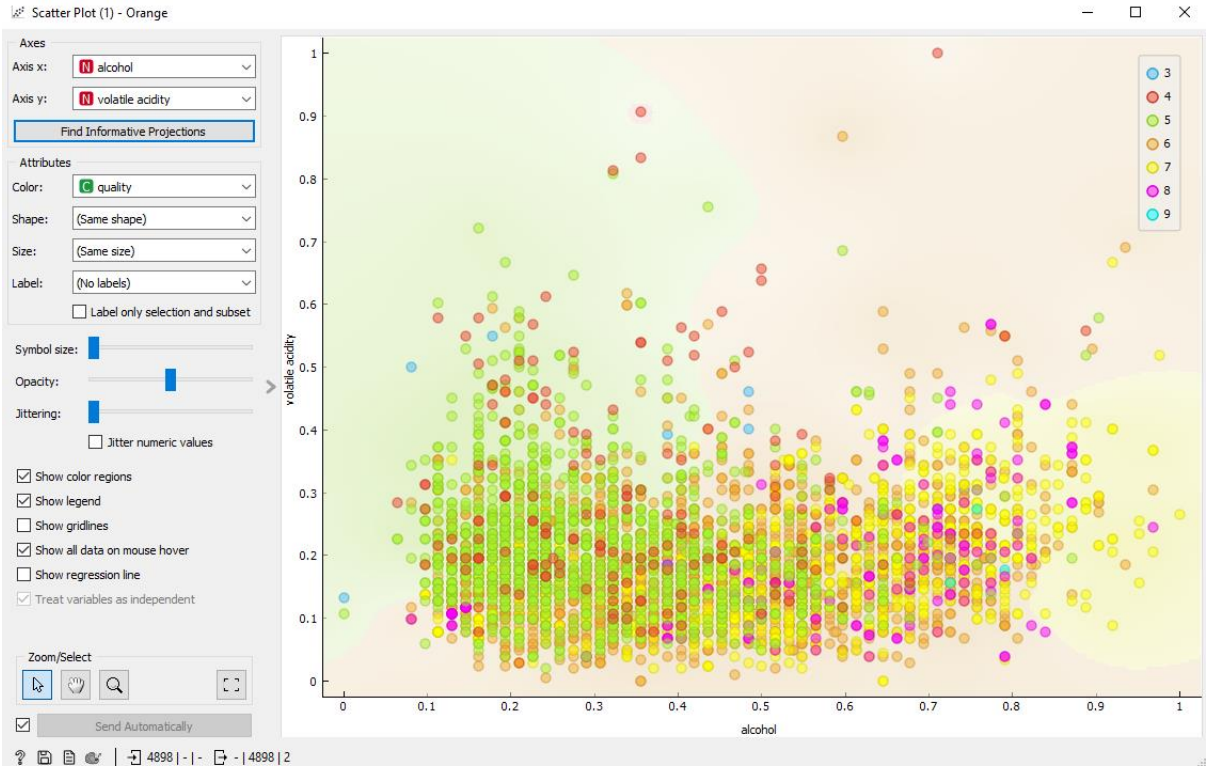


1.1.attēls Datu klasifikācija

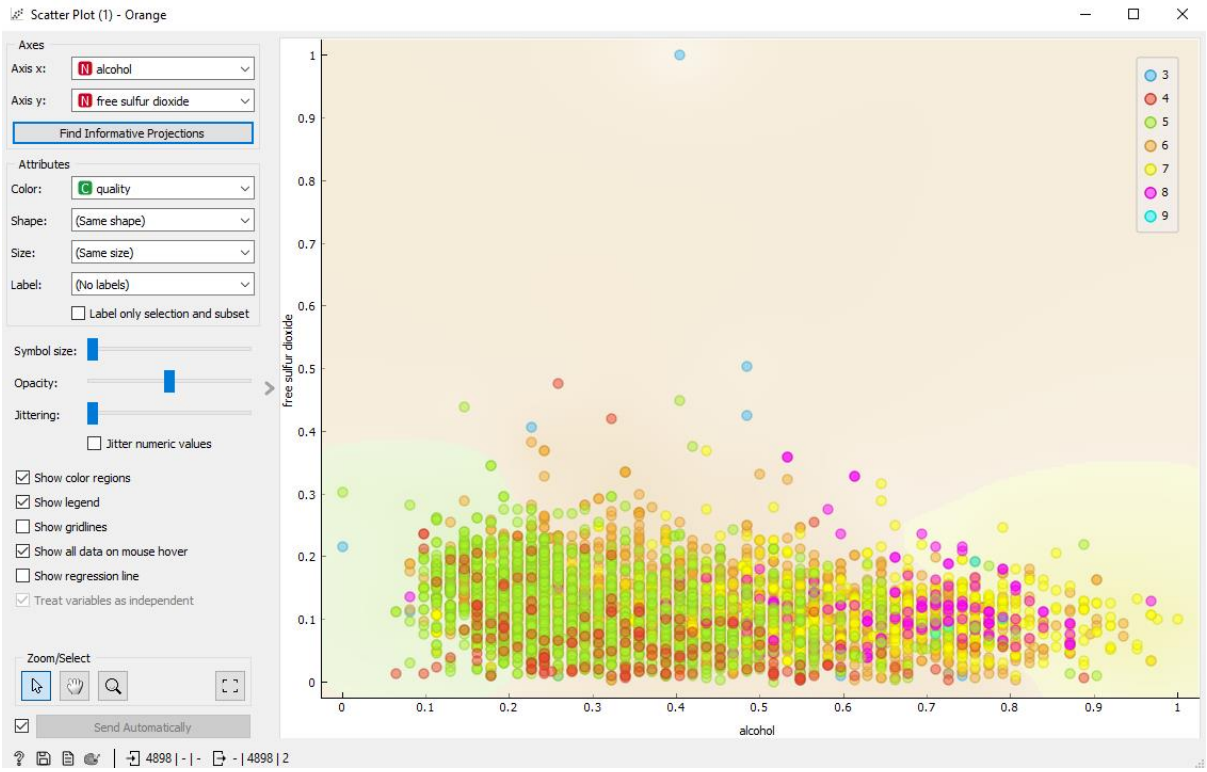
	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
1	6	31	0.186275	0.216867	0.308282	0.106825	0.149826	0.37355	0.267785	0.254545	0.267442	0.129032
2	6	22	0.215686	0.204819	0.0153374	0.118694	0.0418118	0.285383	0.132832	0.527273	0.313953	0.241935
3	6	43	0.196078	0.240964	0.0966258	0.121662	0.097561	0.204176	0.154039	0.490909	0.255814	0.33871
4	6	33	0.147059	0.192771	0.121166	0.145401	0.156794	0.410673	0.163678	0.427273	0.209302	0.306452
5	6	33	0.147059	0.192771	0.121166	0.145401	0.156794	0.410673	0.163678	0.427273	0.209302	0.306452
6	6	43	0.196078	0.240964	0.0966258	0.121662	0.097561	0.204176	0.154039	0.490909	0.255814	0.33871
7	6	21	0.235294	0.0963855	0.0981595	0.106825	0.097561	0.294664	0.150183	0.418182	0.290698	0.258065
8	6	31	0.186275	0.216867	0.308282	0.106825	0.149826	0.37355	0.267785	0.254545	0.267442	0.129032
9	6	22	0.215686	0.204819	0.0153374	0.118694	0.0418118	0.285383	0.132832	0.527273	0.313953	0.241935
10	6	43	0.137255	0.259036	0.0138037	0.103858	0.0905923	0.278422	0.128976	0.454545	0.267442	0.483871
11	5	43	0.186275	0.246988	0.0130368	0.0712166	0.0313589	0.12529	0.0711394	0.245455	0.395349	0.645161
12	5	48	0.147059	0.240964	0.0552147	0.0771513	0.0522648	0.232019	0.146327	0.381818	0.360465	0.274194
13	5	40	0.0980392	0.222892	0.00920245	0.0919881	0.0487805	0.153132	0.0942741	0.418182	0.476744	0.451613
14	7	25	0.0784314	0.240964	0.0138037	0.103858	0.160279	0.310905	0.078851	0.745455	0.348837	0.709677
15	5	45	0.333333	0.373494	0.286043	0.0919881	0.135889	0.37819	0.252362	0.236364	0.523256	0.274194

1.2. attēls. Datu faila struktūras fragments

Izmantojot scatter plot rīku, tiek izveidotas divas 2-dimensionālas izkliedes diagrammas 1.3. un 1.4. attēli. Darbā tiek attēloti 2 labākie “Find Informative Projections” diagrammu piedāvājumi.



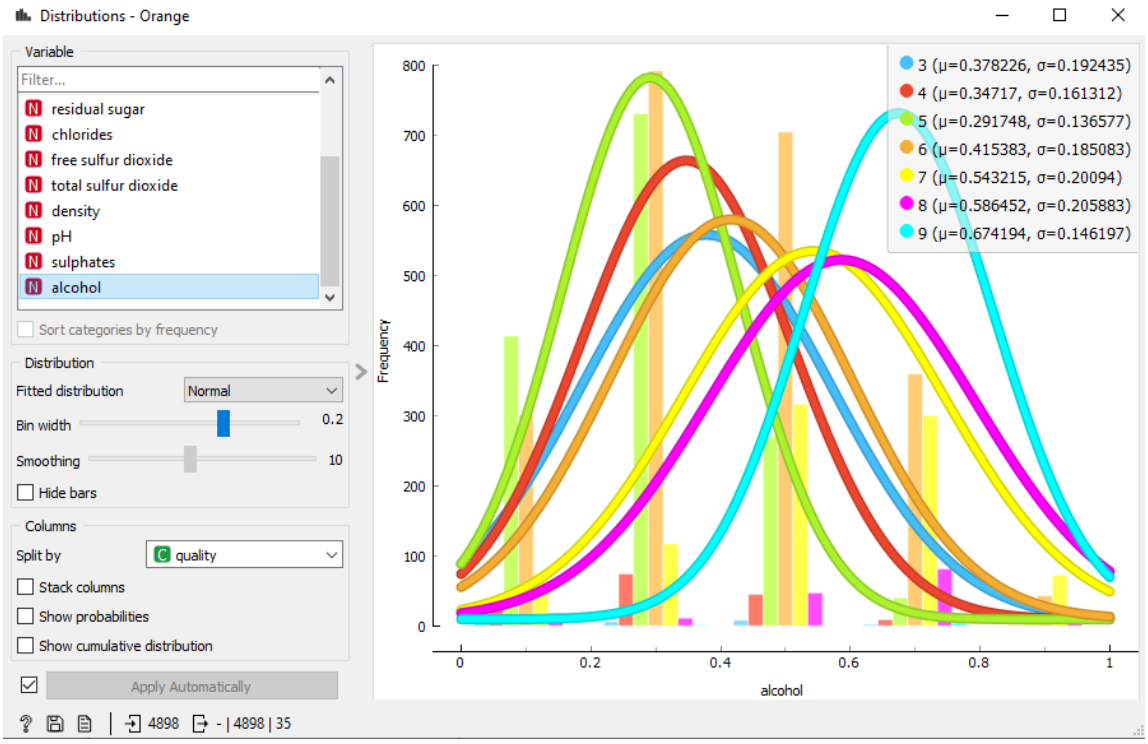
1.3.attēls. Izkliedes diagramma (alcohol pret volatile acidity)



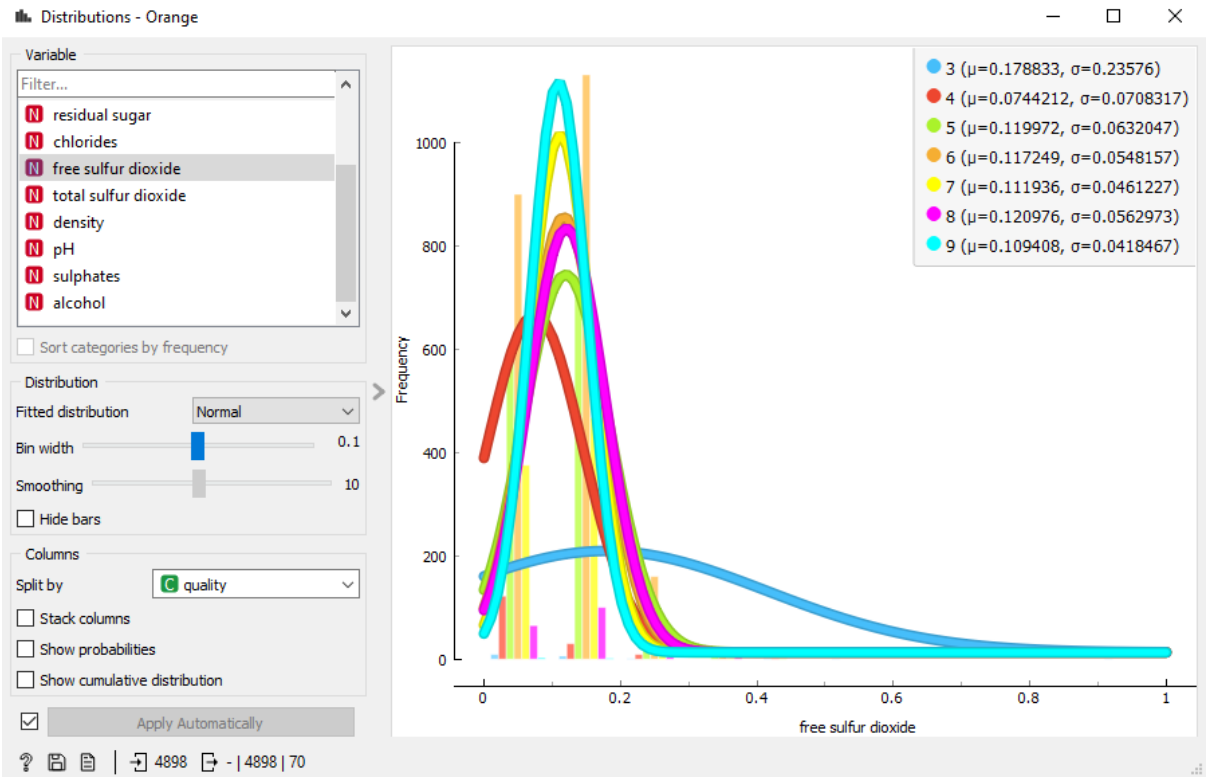
1.4.attēls. Izkliedes diagramma (alcohol pret free sulfur dioxide)

Izkliedes diagrammās redzams, ka dati nav pārāk labi atdalīti, bet vairāk centrēti uz grafika lejas daļu. Iespējams šis ir lielā datu skaita apjoma dēļ, ko, iespējams, būs iespēja atdalīt sīkāk tālākajā darba daļā.

Apskatot izkliedes diagrammas, tiek pielietoti tie paši atribūti – Alcohol un free sulfur dioxide, kur tie tiek aplūkoti sīkāk 1.5. un 1.6. histogrammās.

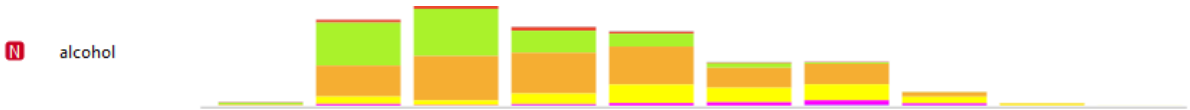


1.5. attēls. Alcohol histogramma

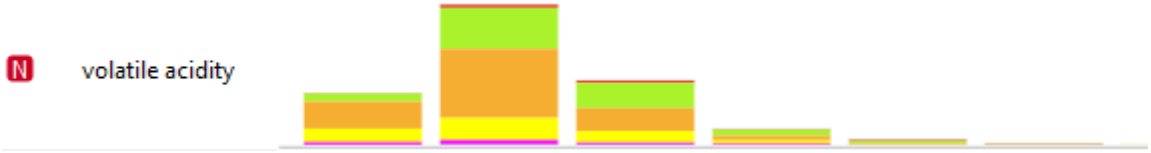


1.6.attēls. free sulfur dioxide histrogramma

Histogrammās redzams, ka tikai alkohola dati 1.4.attēlā ir gana plaši izkliedēti, pa visu x asi. Liecinot par to, ka dati ir ļoti blīvi strukturēti.



1.7.attēls Alcohol sadalījums (Color: quality)



1.8.attēls volatile acidity sadalījums (Color: quality)

Attēlos 1.7. un 1.8. redzami labākā atdalījuma abu mainīgo datu izklājums

Līdz šim veiktajā darbā iespējams izteikt vairākus secinājumus:

- Datu kopā izteikti dominē kvalitātes rādītājs “6”, tas ietver sevī 44.88% no visiem datiem.
- Datu vizuālais atspoguļojums neļauj redzēt skaidru datu struktūru
- Īsti nav iespējams atdalīt nevienu no datu grupējumiem, bet ir vizuāli mazas iezīmes, kas iedala datus 4 lielākās grupās
- Pārsvarā dati atrodas nosacīti tuvu viens otram, bet ir neliels skaits mērījumu kuri ir nostāk no pārējiem datiem.
- Pēc dispersijas datiem iespējams secināt, ka izkliede datiem nav pārāk liela, kas liek datiem atrasties

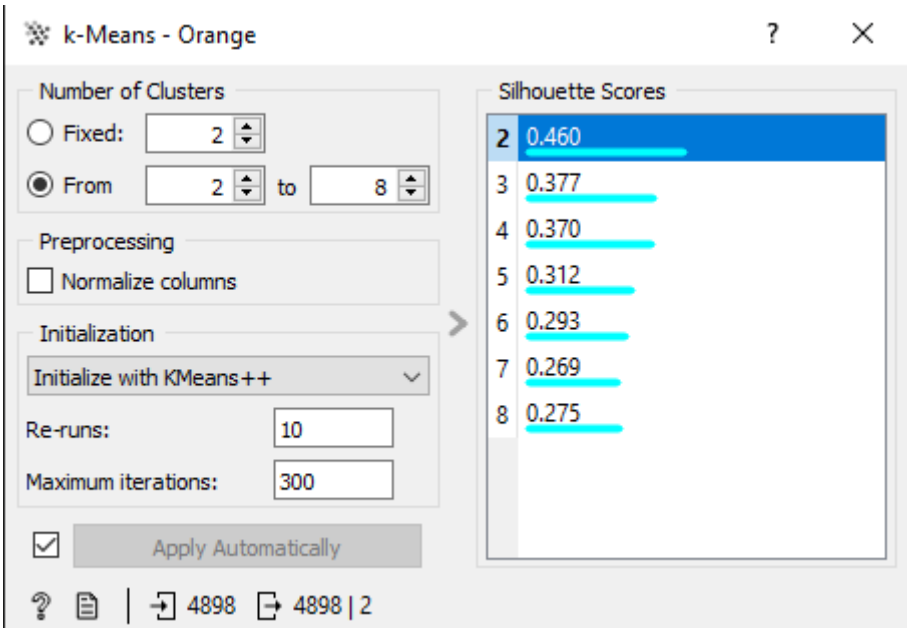
Nepārraudzīta mašīnmācīšanās

K-Means

Pielietojot K-vidējo algoritmu tiek aprēķināti silhouette scores vairākiem klasterēšanas veidiem. Aplūkojot attēlu 2.1., redzams, ka jo vairāk klasteru, jo sliktākas kvalitātes rezultāti tiek iegūti.

Termins	Skaidrojums
Number of Clusters	Klasteru skaits – viens vai vairāki varianti
Preprocessing	Kolonnu normalizēšana (0-1 mērogā)
Initialization method	Klasterizēšanas metodes sākšanas inicializēšanas veids. “KMeans++” – pirmais klastera centrs izvēlēts pēc nejaušības, pārējie nolikti atlikušajos punktos. “Random initalization” – visiem klasteriem punkti nejaušās vietās.
Re-runs	Algoritma atkārtošanas reizes
Maximum iterations	Maksimālais iterāciju skaits

2.1.tabula K-Means rīka parametri [3]



2.1.attēls. k-Means Silhouette Score rezultāti

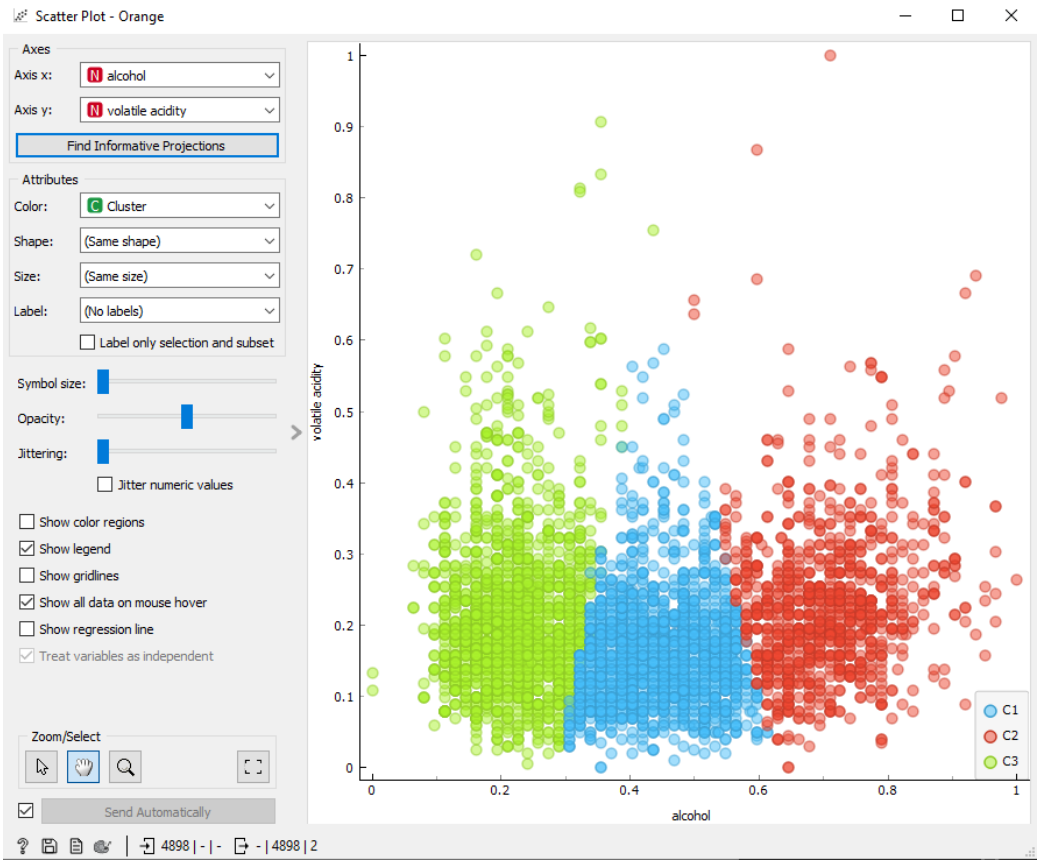
Ar klusteriem tiek apzīmēts cik daudzās grupās esošie dati tiek iezīmēti, piemēram, ja ir 2 klasteri, datu kopa tiek iedalīta divās daļās, ja ir 3 klasteri, to iedala trijās daļās u.t.t.

Attēlos 2.2. un 2.3. var apskatīt rezultātus, kādi tiek iegūti veicot klasterēšanu ar diviem klasteriem. Redzams, ka dati tiek iedalīti divās daļās, kas savā starpā nedaudz pārklājas pie aptuveni vienas un tās pašas alkohola atzīmes.

2.2. attēls. K-Means klasterēšanas rezultāti ar 2 klasteriem (alcohol un volatile acidity)

2.3. attēls. K-Means klasterēšanas rezultāti ar 2 klasteriem (alcohol un free sulfur dioxide)

Attēlos 2.4. un 2.5. redzams, kā notiek klasterēšana izmantojot 3 klasterus. Lai gan ar aci skatoties, pirmajā momentā atdalīšana izskatās precīzāka, skatoties tuvplānā redzamas lielākas pārejas kļūdas.



2.4. attēls. K-Means klasterēšanas rezultāti ar 3 klasteriem (alcohol un volatile acidity)

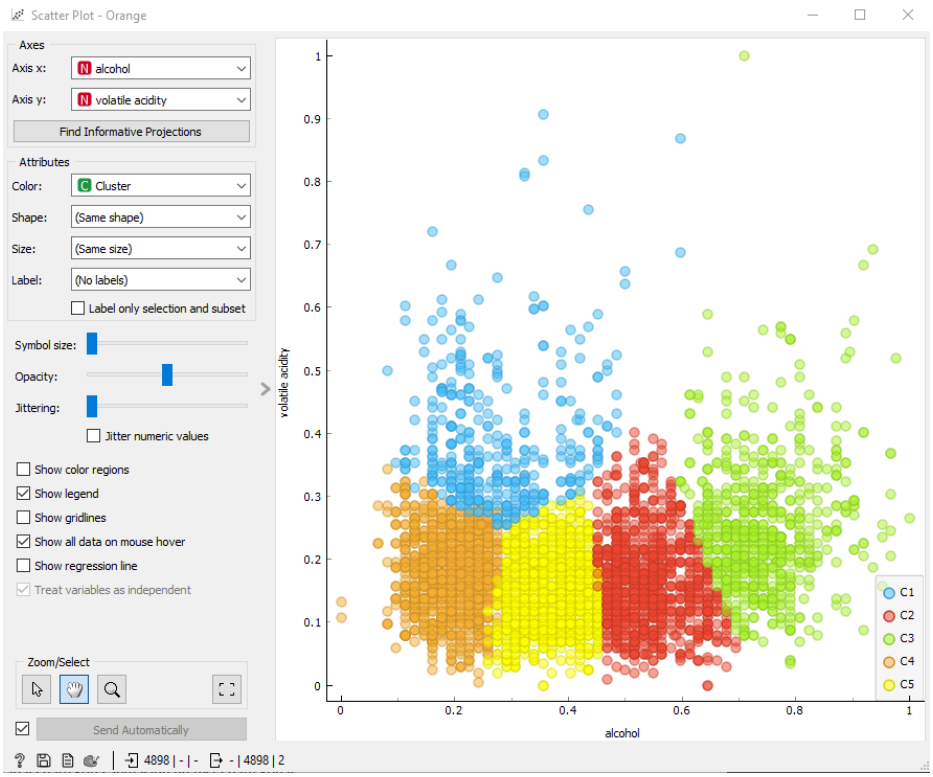
2.5. attēls. K-Means klasterēšanas rezultāti ar 3 klasteriem (alcohol un free sulfur dioxide)

Apskatot attēlus 2.6. un 2.7. redzams ka palielinot klasteru skaitu līdz 4, tiek iegūti dati, kuri pēc izskata liekas precīzāki, tomēr lielā datu skaita dēļ, šie dati viens ar otru pārklājas, un tiek iegūti neprecīzāki dati kā iepriekšējajos mēģinājumos.

2.6. attēls. K-Means klasterēšanas rezultāti ar 4 klasteriem (alcohol un volatile acidity)

2.7. attēls. K-Means klasterēšanas rezultāti ar 4 klasteriem (alcohol un free sulfur dioxide)

Attēlos 2.8. un 2.9. aplūkojams kā 5 klasteri ietekmē rezultātus. Lai gan 2.8. datu atdalīšana tiek atdalīta ar vien precīzāk, attēlā 2.9. dati pārklājas vairāk, kā attēlā 2.7.

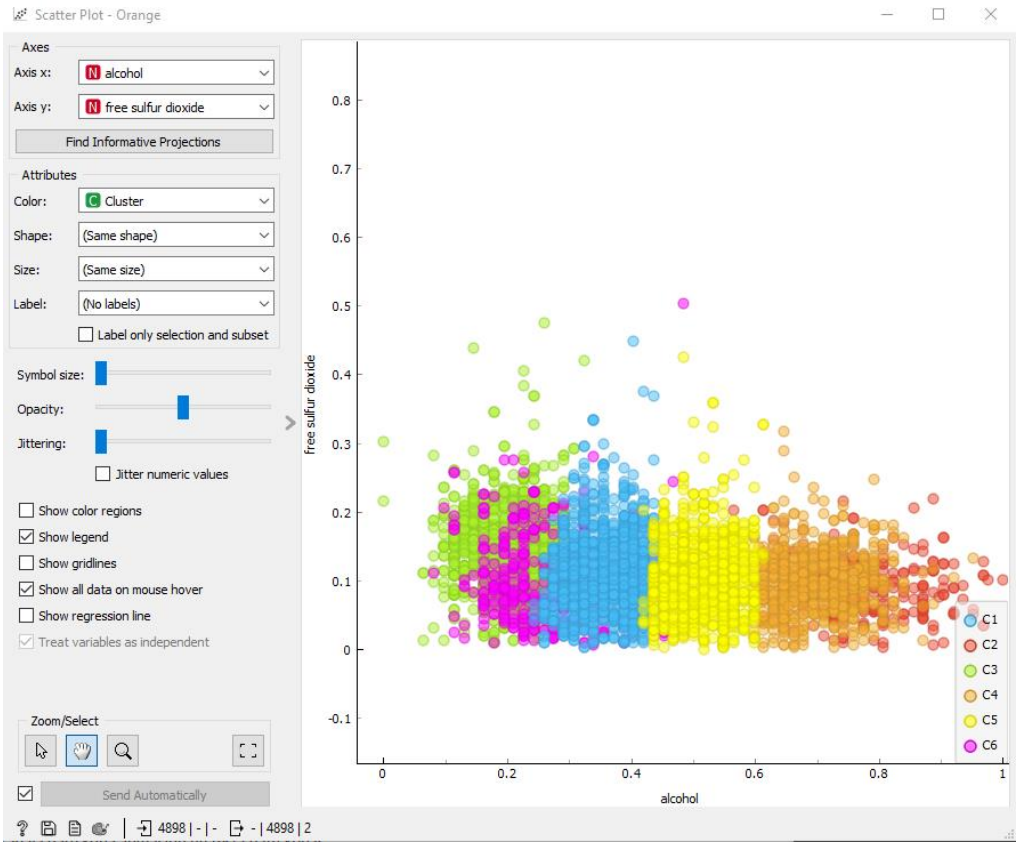


2.8. attēls. K-Means klasterēšanas rezultāti ar 5 klasteriem (alcohol un volatile acidity)

2.9. attēls. K-Means klasterēšanas rezultāti ar 5 klasteriem (alcohol un free sulfur dioxide)

Apskatot attēlus 2.10. un 2.11. redzams, ka veidojoties jauniem klasteriem turpina notikt tas pats, kas iepriekš. Attēls 2.10. uzlabojas, bet attēls 2.11. kvalitāte pasliktinās.

2.10. attēls. K-Means klasterēšanas rezultāti ar 6 klasteriem (alcohol un volatile acidity)



2.11. attēls. K-Means klasterēšanas rezultāti ar 6 klasteriem (alcohol un free sulfur dioxide)

Hierarchical clustering

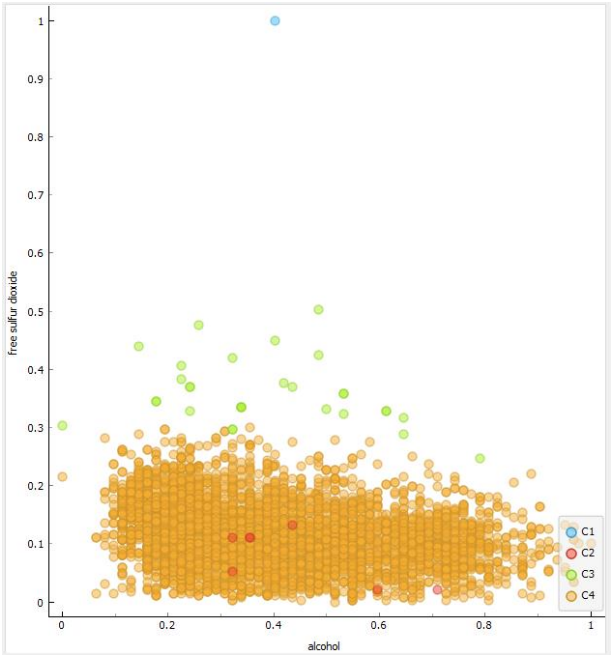
Distances blokā tiek izmantots "euclidean" distances metrika, jo visi pieejamie dati ir kā skaitļi. Hierarchical Clustering blokam tiek pievienots Scatter Plot bloks, lai lielo datu apjomu būtu vieglāk apskatīt un tam tiek pievienots data table, lai precīzāk noteiktu klasteru skaitu.

Termins	Skaidrojums
Linkage	Savienojuma veids
Single	Distance starp klastera tuvākajiem elementiem
Average	Vidējā distance starp divu klasteru elementiem
Weighted	WPGMA(Weigghted Pair Group Method with Arithmetic Mean) metodes izmantošana
Completed	Distance starp klasteru tālākajiem elementiem
Ward	Palielinājuma kvadrātu kļudas summa
Pruning	Klasteru dziļuma ierobežošana
Selection	Klasteru izvēles metodes
Manual	Manuāla klastera atlasīšana
Height ratio	Augstuma % iezīmēšana no klasteru sākuma daļas
Top N	Augšēju mezglu skaita izvēle

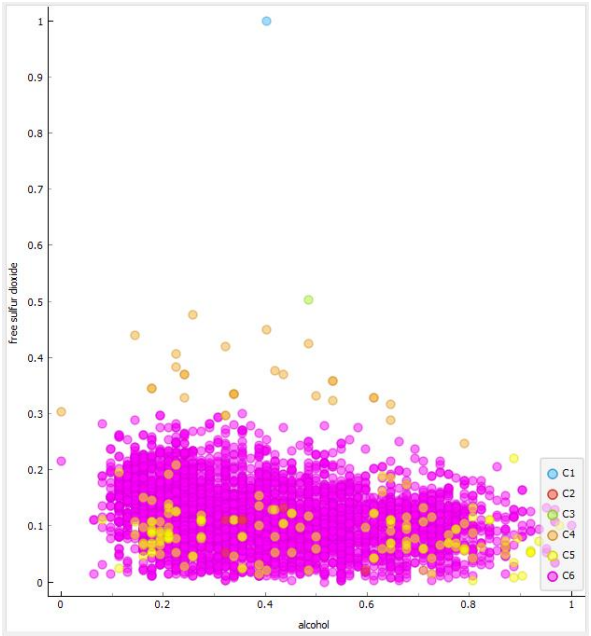
2.2.tabula Hierarchical clustering rīka parametri [4]

Tā, kā lielā datu apjoma dēļ datus nav iespējams objektīvi apskatīt "hierarchical clustering blokā, šim blokam tiek pievienots "Scatter Plot" bloks rezultātu apskatei.

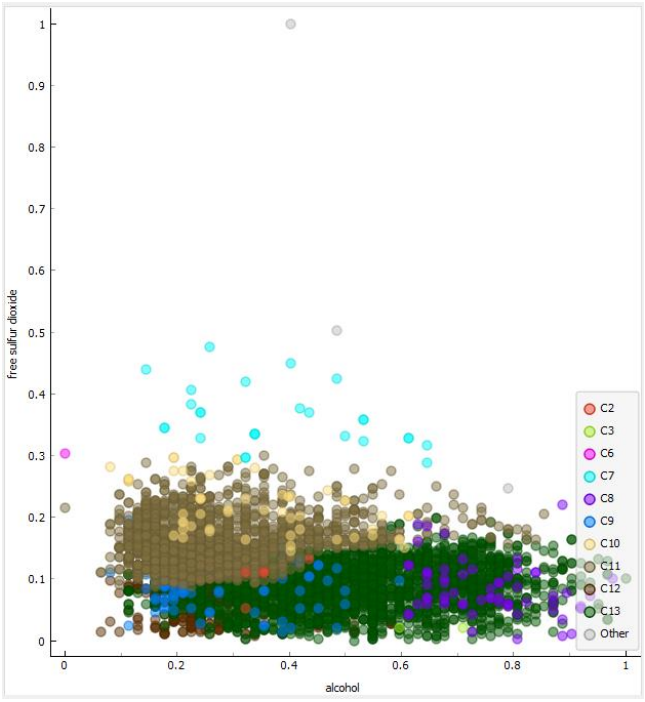
Izmantojot hierarhiskās klasterizācijas algoritmu (pēc average metodes), kas redzams attējos 2.12., 2.13., 2.14. un 2.15. ir iespējams nolūkot kā pieaug klasteru skaits lielās datu kopas dēļ. Aplūkojot klasterus manuali iespējams apskatīt, ka klasterēšana notiek ļoti aptuveni. Dati daudz pārklājas ar citiem datiem un netiek iegūti skaidri rezultāti.



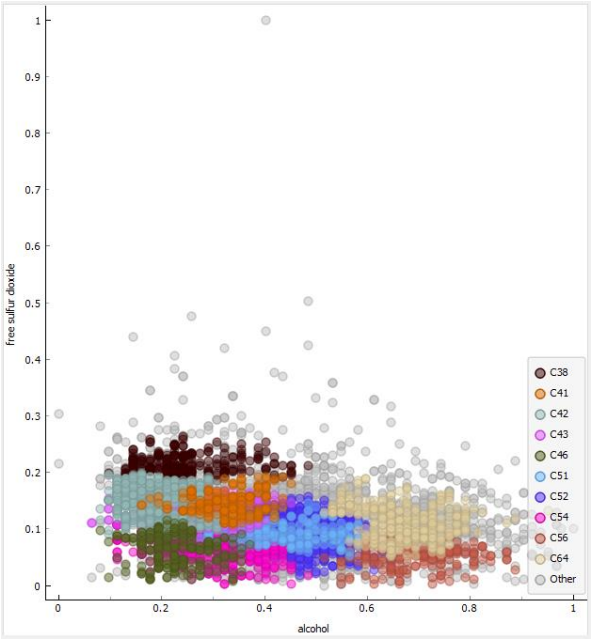
2.12.attēls. Hierarhiskā klasterizācija (30.0% C=4)



2.13.attēls. Hierarhiskā klasterizācija (25.0% C=6)



2.14.attēls. Hierarhiskā klasterizācija (15.0% C=13)



2.15.attēls. Hierarhiskā klasterizācija (8.0% C=64)

Secinājumi:

- K-vidējo algoritms sniedza labus rezultātus izmantojot alcohol un volatile acidity, kur ar katru reizi rezultātu kvalitāte bija apmēram vienāda. Tomēr alcohol un free sulfur dioxide kvalitāte būtiski samazinājās, kas liecina par to, ka ne visos gadījumos k-means algoritms strādā izmantojot šo datu kopu.
- Izmantojot hiperparametru vērtības, viskvalitatīvākie dati, skatoties ar aci, sanāca ar funkciju average, bet, iespējams, datu daudzuma dēļ eksperiments ar šīm vērtībām nebija tik precīzs kā ar k-means algoritmu.

Pārraudzīta mašīnmācīšanās

Pārraudzītās mašīnmācīšanās uzdevumos datu kopa tiek sadalīta divās daļās – apmācības kopa (75%, jeb 3674 dati) un testa kopa (25%, jeb 1224 dati). Abas šīs kopas tiks pielietotas visiem trim algoritmiem, lai vērtēšana būtu objektīvāka, kurs algoritms ar vieniem un tiem pašiem datiem labāk spēj noteikt rezultātu

Šajā darbā tiks apskatīti trīs mašīnmācīšanās algoritmi – mākslīgie neironu tīkli, Loģistiskā regresija un kNN modulis.

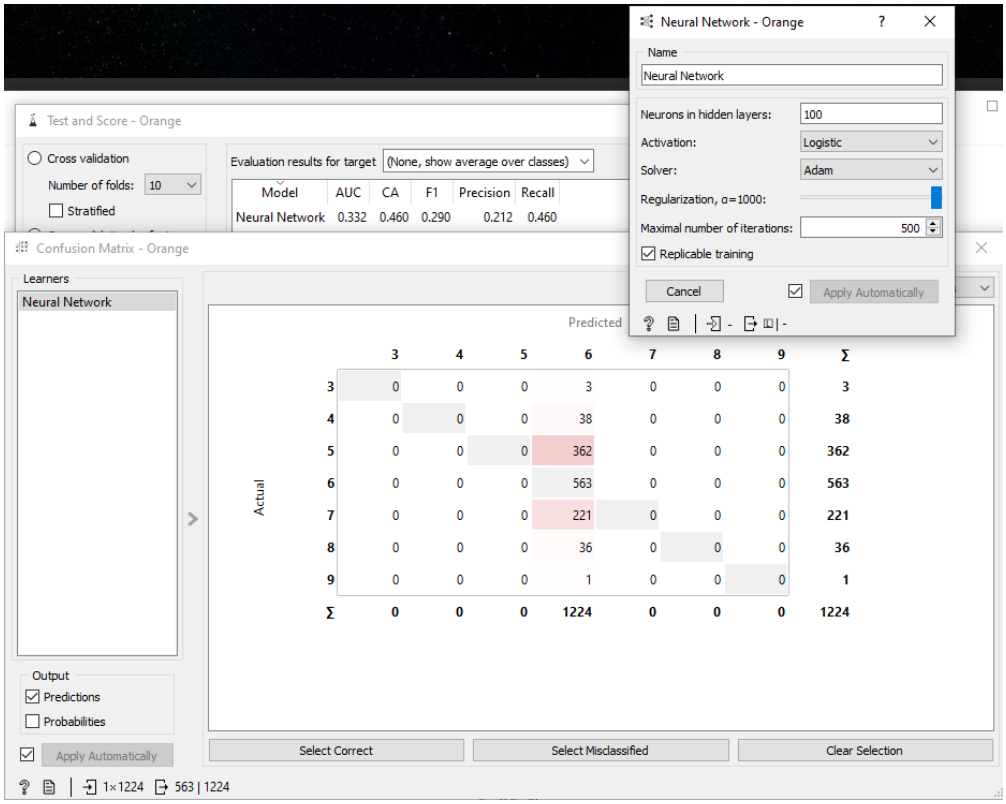
Mākslīgie neironu tīkli

Pirmais no apskatāmajiem algoritmiem ir mākslīgo neironu tīkli. Tam tiek veikti trīs eksperimenti, manuāli mainot dažādus parametrus, lai sasniegtu labāko rezultātu.

Termins	Skaidrojums
Name	Modeļa nosaukums
Neurons in hidden layers	Neironi apslēptajos slāņos
Activation	Aktivizācijas opcijas
ReLu	Izlabotās lineārās vadības funkcija
tanh	Hiperboliskas tangesa funkcija
Logistic	Loģisktikas sigmoīda funkcija
Identity	Bezoperācijas aktivizēšanas veids
Solver	Vērtību risinājumu veids
L-BFGS-B	Kvanzi-Ņūtona metožu ģimeņu saimes optimizators
SGD	Stohastisks gradienta mazinātājs
Adam	Stohastisks gradienta bāzēts optimizētājs
Regularization	Regularizācija
Maximal number of iterations	Maksimālais iterāciju skaits
Replicable training	Replicējama trenēšana

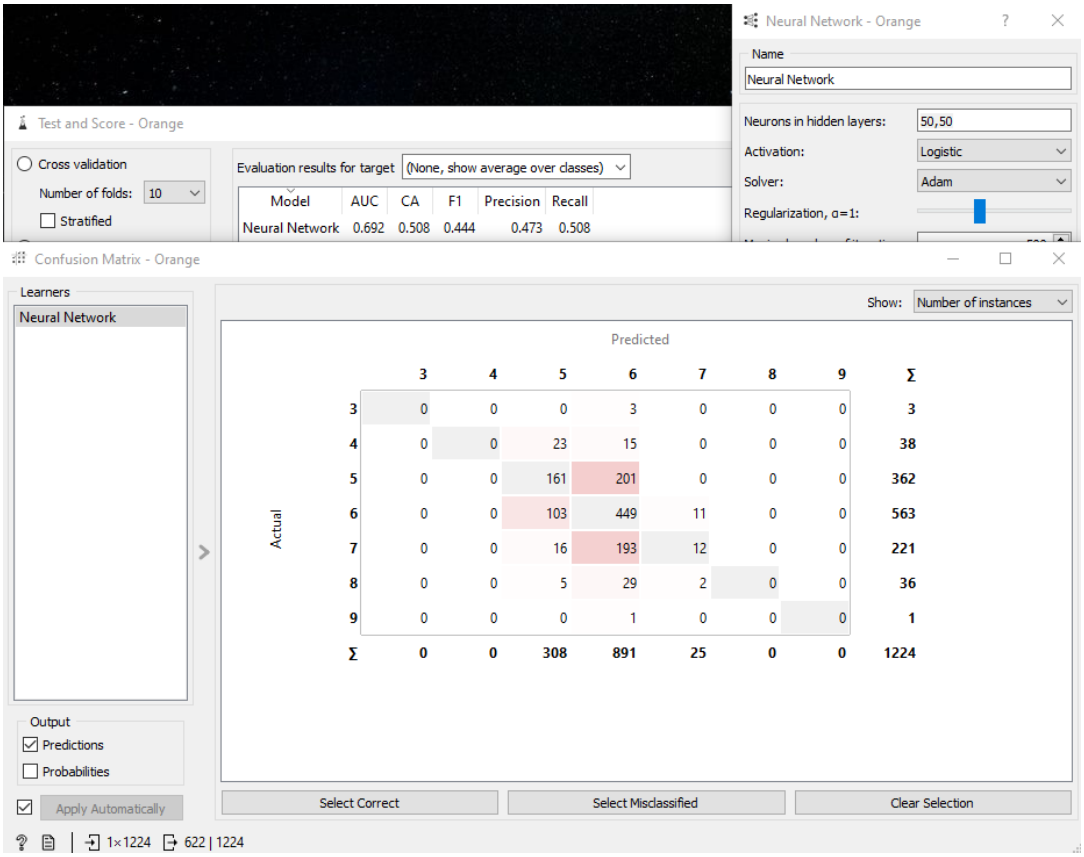
3.1.tabula Neironu tīklu parametri [5]

Pirmajā eksperimentā neural network parametri redzami attēlā 3.1. Algoritms atpazina tikai kvalitātes mērījumu “6”, kas noveda pie precizitātes 0.212.



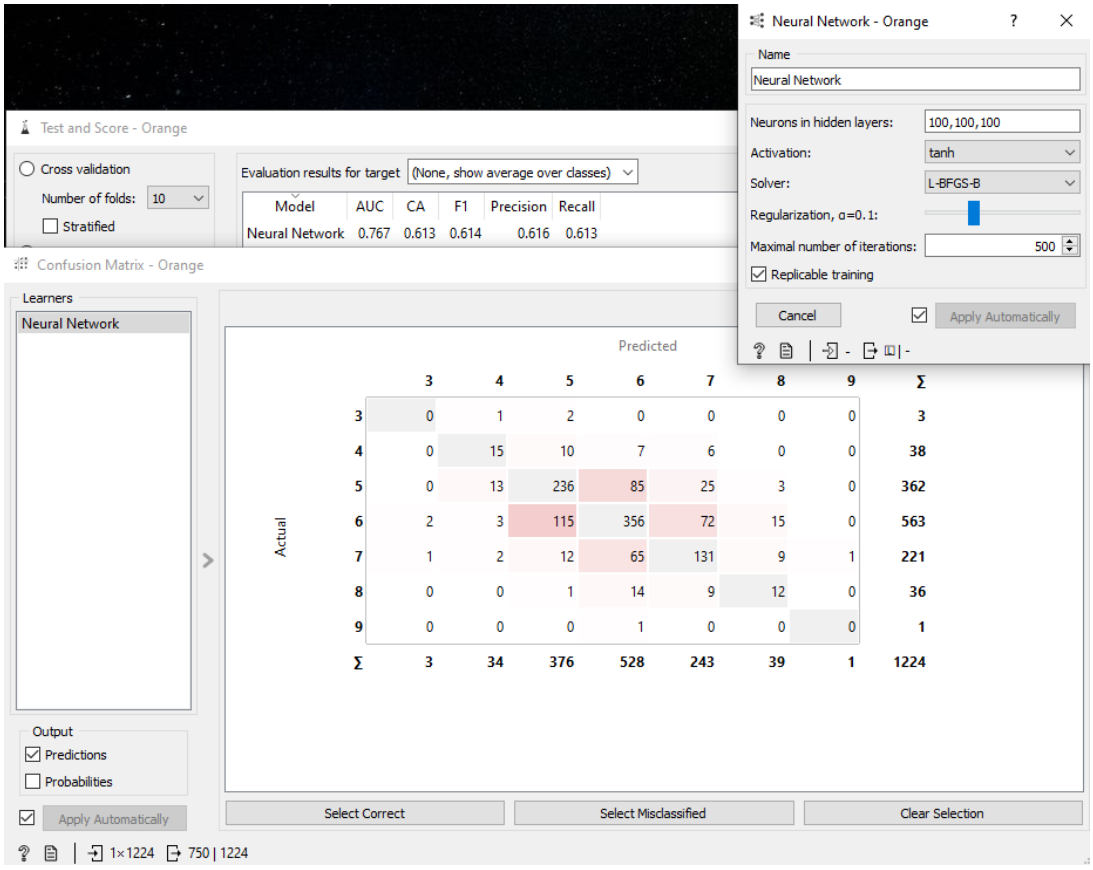
3.1.attēls Pirmais eksperiments izmantojot Mākslīgo neitronu tīklus

Veicot otro ekperimentu “Regularization” tika pazemināts uz 1, un tika piešķirts vēl viens apakštīkls, kur abi tiek nomainīti uz vērtību 50. Rezultātu kvalitāte uzlabojās pat vairāk kā divas reizes, sasniedzot precititāti 0.473. Confusion Matrix vertikālajā zonā“Predicted” tiek sniegtas arī citas atbildes, nevis kā iepriekšējā attēlā 3.2.



3.2.attēls Otrais eksperimants izmantojot Mākslīgo neitronu tīklus

Trešajā eksperimentā manuali tika atrasts labākais iespējamais rezultāts, kas redzams 3.3.attēlā. Piešķirti trīs neironu slēptie slāņi (100,100,100), pamazinot "Regularization" līdz 0.1 un uzstādot "Activation" – tanh, "Solver"- L-BFGS-B. Tika iegūta precizitāte 0.616.



3.3.attēls Trešais eksperiments izmantojot Mākslīgo neitronu tīklus

Loģistiskā regresija

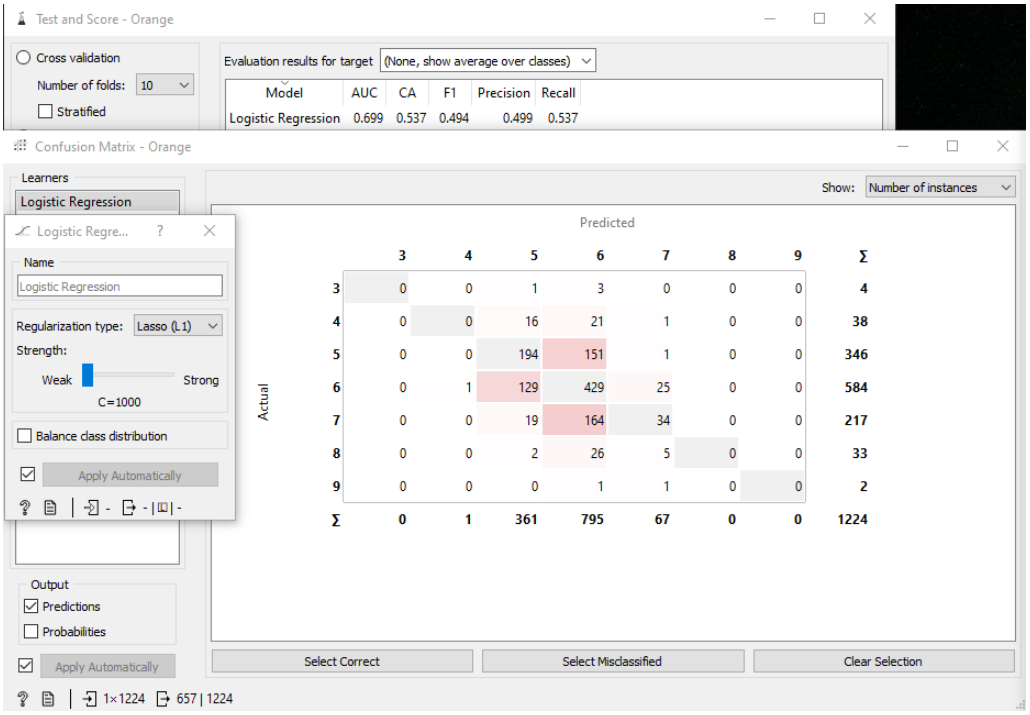
Loģistiskā regresijas, jeb Logistic Regression rīks mācās no datiem izmantojot loģistiskās regresijas modeļi [6]. Algortims apkopo datu kopu, kurai piešķirtas vērtības. Tas modelē varbūtību notikumam no logaritmiskām varbūtībām, kuras ņemtas no noteikta skaita neatkarīgiem mainīgajiem. Šis ir statistisks modelis.

Šo modeli izvēlējos, jo par to tika runāts lekciju laikā un vēlējos uzzināt kā šis modelis darbosies uz nepārāk labi sakārtotas datu kopas.

Termins	Skaidrojums
Name	Modeļa nosaukums
Regularization type	Regulācijas tipa izvēle (L1 vai L2)
Strenght	Algoritma spēka cenas izmaniņas (0,001 – 1000), jo mazāka cena jo precīzāks algoritms
Balance class distribution	Klases distribūcijas balansēšana

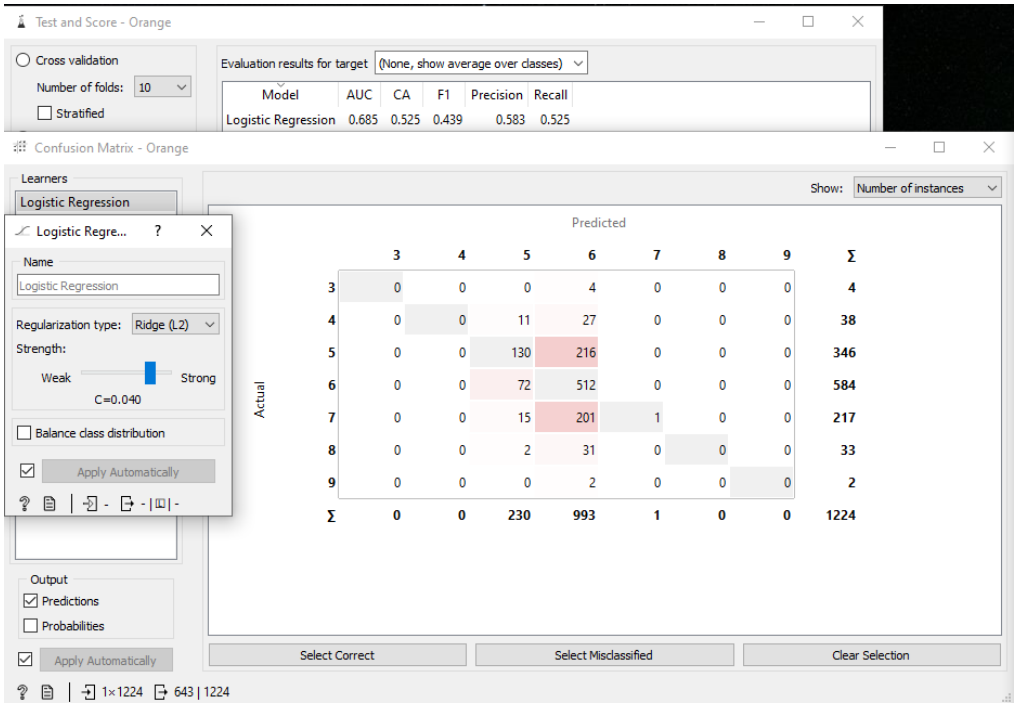
3.2.tabula Loģistiskās regresijas parametri

Pirmajā eksmerimentā, kas redzams attēlā 3.4., tiek izmantots C=1000, L1 regulācijas tips un tiek sasniegta precizitāte 0.499. Šajā gadījumā tiek izmantots L1 nevis L2, jo L1 sniedz labākus rezultātus, kā L2.



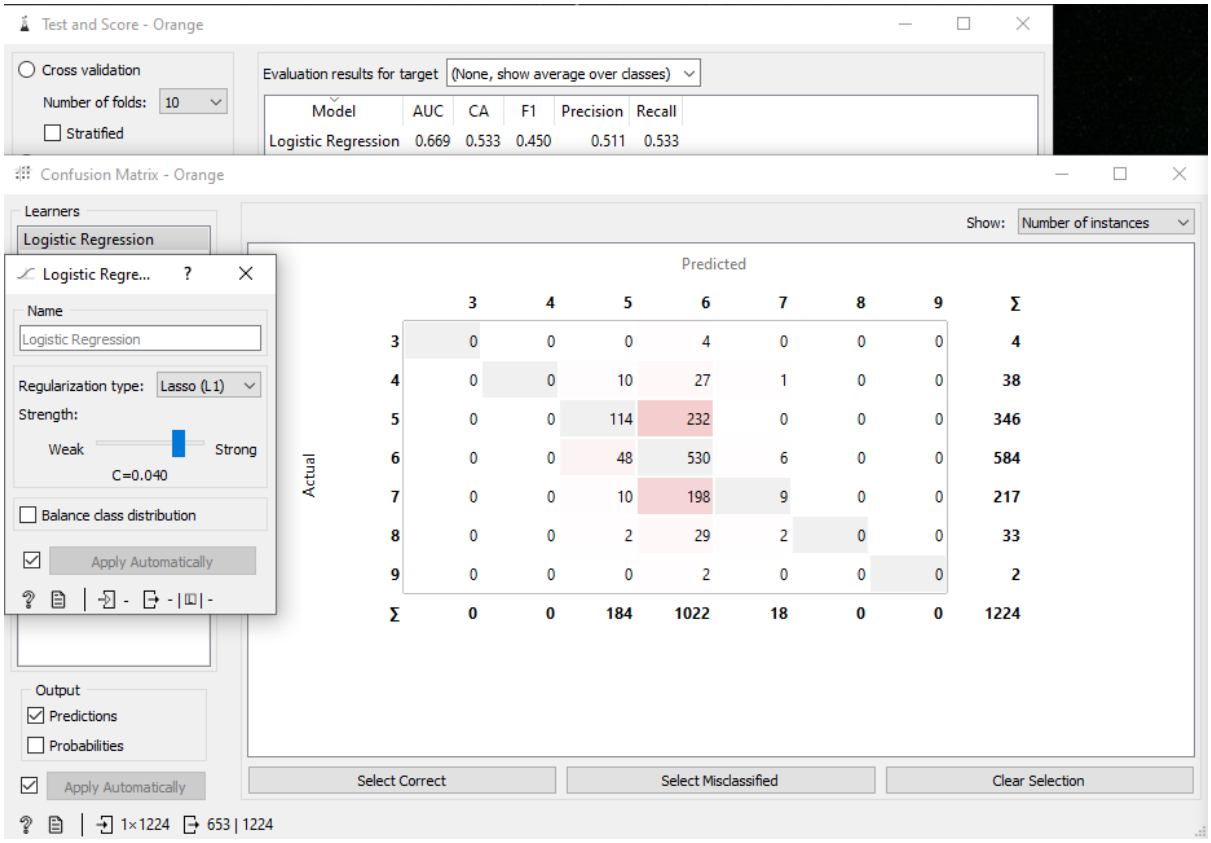
3.4.attēls Pirmais eksperiments izmantojot Loģiskās regresijas moduli.

Otrajā eksperimentā tika atrasta optimālākā C vērtība L2 regulācijas tipam un atlasītajai datu kopai, kas sniedza precizitāti 0.583.



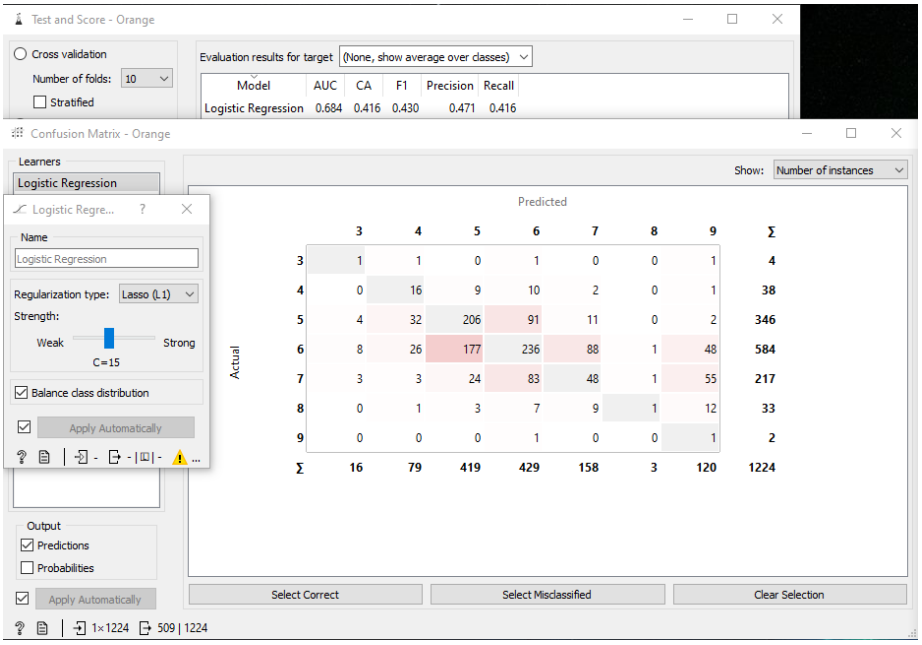
3.5.attēls Otrais eksperiments izmantojot Loģiskās regresijas moduli.

Trešajā eksperimentā tika atrasta optimālākā C vērtība L1 regulācijas tipam un atlasītajai datu kopai, kas sniedza precizitāti 0.511.

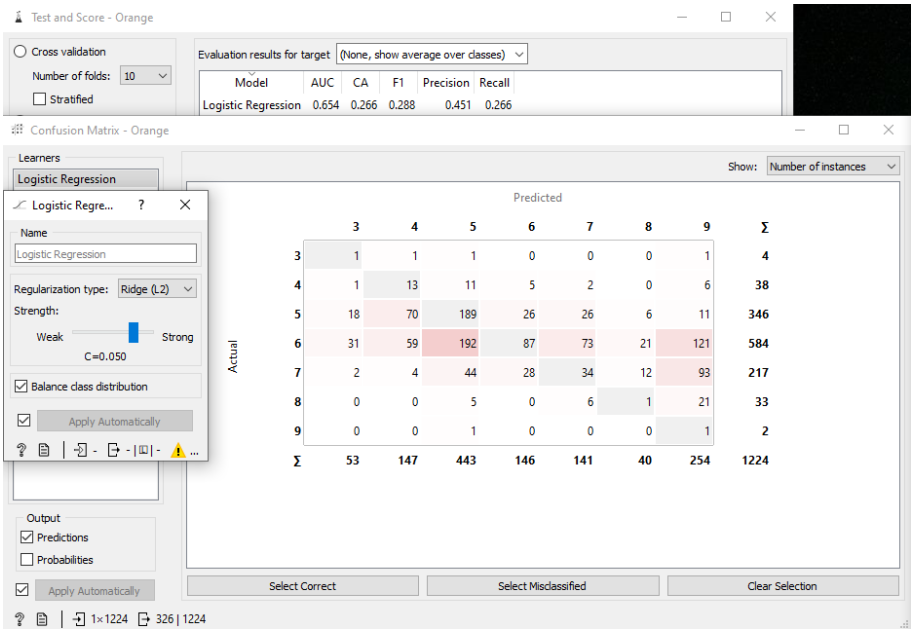


3.6.attēls Trešais eksperiments izmantojot Loģiskās regresijas moduli

Ceturtajā un piektajā eksmerimentā tiek ieslēgts “balance class distribution”, lai pārbaudītu vai rezultāti uzlabotos šajos gadījumos. Izmantojot L1 regulācijas tipu un labāko C=15, tika sasniegta precizitāte 0.471, kas redzams attēlā 3.7. Izmantojot L2 regulācijas tipu un labāko C=0.05, tika sasniegta precizitāte 0.451, kas redzams attēlā 3.8.



3.7.attēls Ceturtais eksperiments izmantojot Loģiskās regresijas moduli



3.8.attēls Piektais eksperiments izmantojot Loģiskās regresijas moduli

kNN

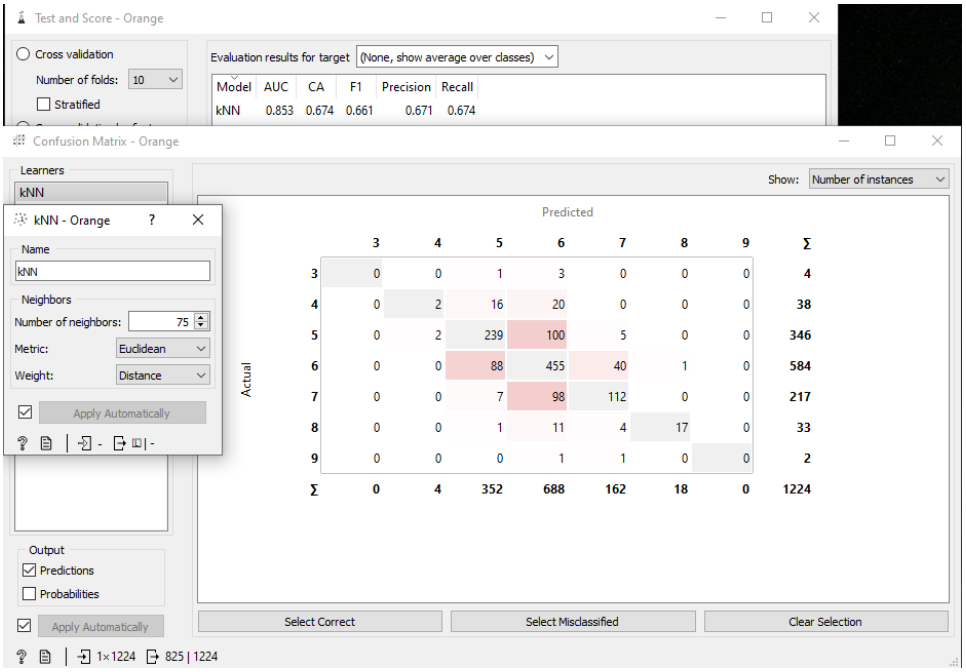
kNN algoritmu izmanto gan regresijas, gan klasifikācijas uzdevumiem. Tiek noteikta līdzība ar kaimiņu klasēm, kas palīdz atrast jaunu datu kopu. Aprēķina attālumus starp jauno punktu un katru punktu trenēšanas datu kopā. [7]

Termins	Paskaidrojums
Name	Modeļa nosaukums
Neighbours	Kaimiņu iestatījumi
Number of neighbors	Kaimiņu skaits
Metric	Metrikas izvēle
Euclidean	Distance starp diviem kaimiņiem
Manhattan	Absolūtā starpību summa visiem atribūtiem
Maximal	Lielākā absolūtā starpība starp atribūtiem
Mahalanobis	Attālums starp punktu un sadalījumu
Weight	Svars
Uniform	Visi kaimiņu svari vienā kopā vienādi
Distance	Tuvāko kaimiņu influence lielāka kā tālāko

3.3.tabula kNN parametri [8]

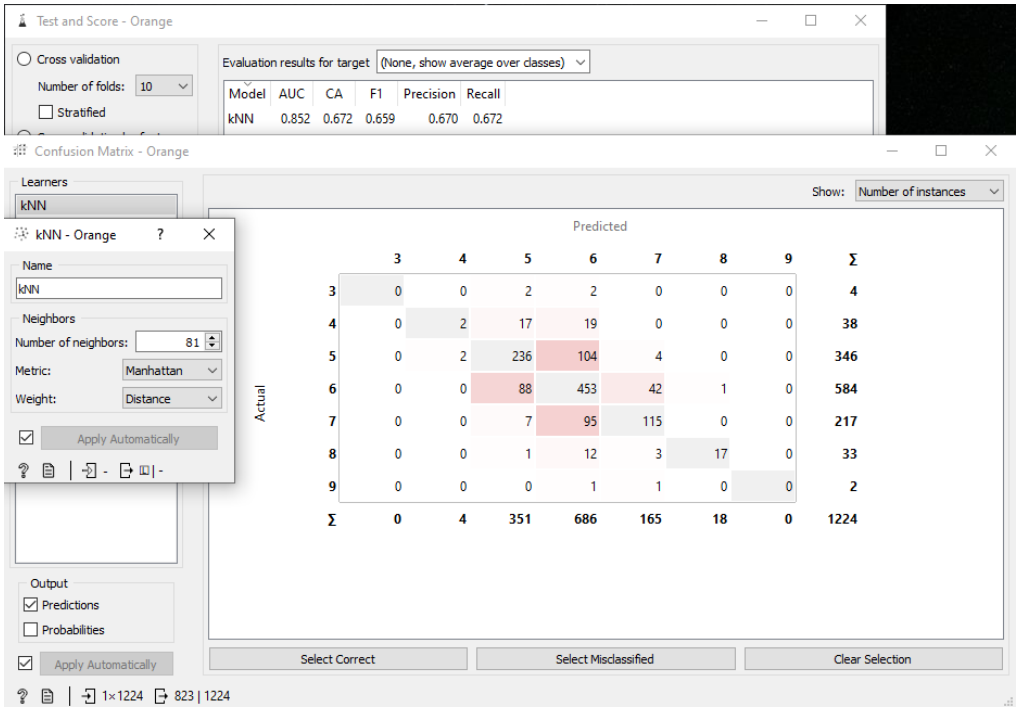
kNN eksperimenta laikā iespējams dabūt bezgalīgi daudz variācijas, tāpēc šajā darbā tiek apskatīts visu metrisko sistēmu manuāli noteiks labākais modelis.

Pirmajā eksperimentā izmantojot Euclidean metrikas iegūta precizitāte 0.671. Kaimiņu skaits 75 un svari ir distance, kā redzams attēlā 3.9.



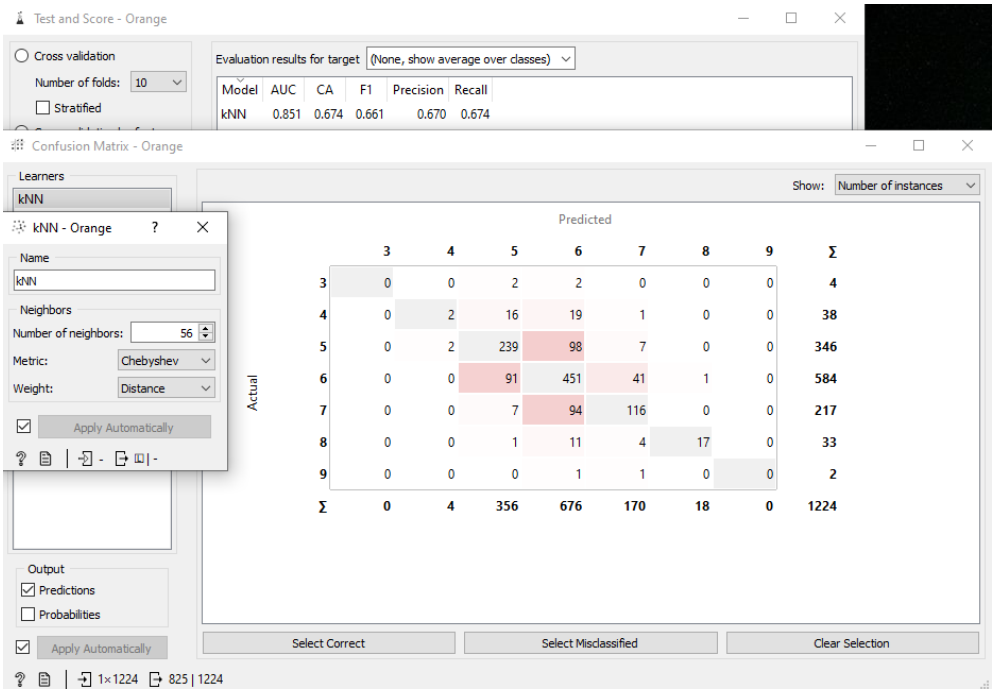
3.9.attēls kNN modeļa Euclidean labākais eksperiments

Otrajā eksperimentā tika izmantota "Manhattan" metrika, un svari tika izmantoti kā distance. Eksperimentā tika iegūta precizitāte 0.670, kā redzams attēlā 3.10.



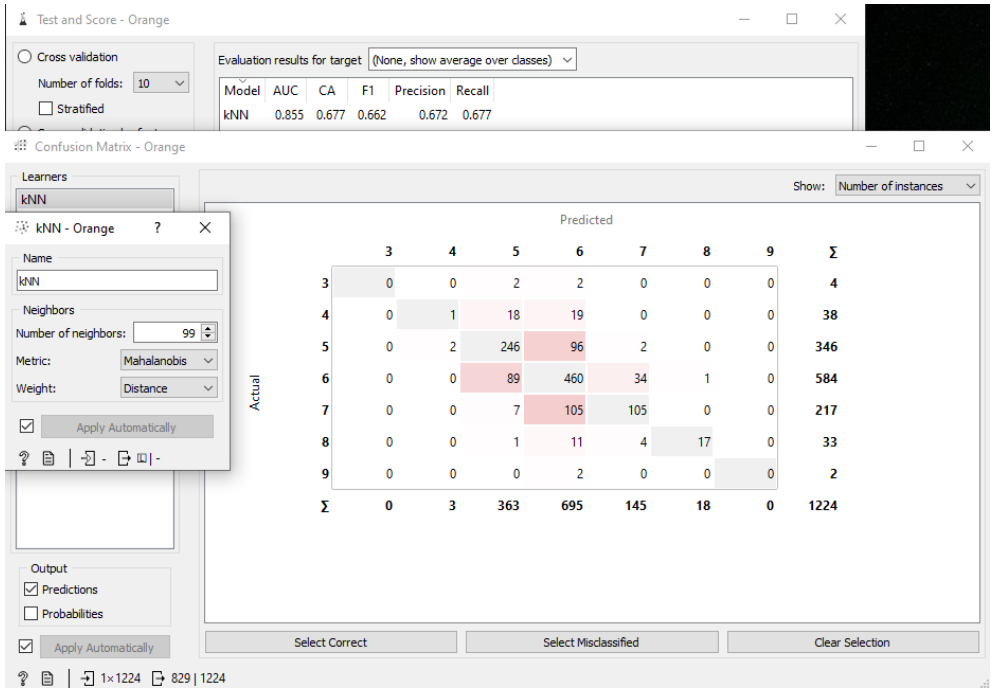
3.10.attēls kNN modeļa Manhattan labākais eksperiments

Trešajā eksperimentā tika izmantota "Chebyshev" metrika, un svari tika izmantoti kā distance. Eksperimentā tika iegūta precizitāte 0.670, kā redzams attēlā 3.11.



3.11.attēls kNN modeļa Chebyshev labākais eksperiments

Ceturtajā eksperimentā tika izmantota "Manhalanobis" metrika, un svari tika izmantoti kā distance. Eksperimentā tika iegūta precizitāte 0.672. Atšķirībā no citiem metrikas eksperimentiem, ar "Manhalanobis" metrikas palīdzību vairāki kaimiņu skaiti spēja sasniegt precizitāti 0.672, kur citos eksperimentos to sanāca sasniegt tikai ar vienu noteiktu kaimiņu skaitu.



3.12.attēls kNN modeļa Manhalanobis labākais eksperiments

Apskatot kNN modeļu rezultātus redzams, ka nevienā no tiek netiek izmantots "Uniform" svara mērījums, nevienā no četriem eksperimentiem tas nesniedza tik labus rezultātus kā "Distance" metrika. Tā precizitāte svārstījās pat 10% robežās savā starpā.

Secinājumi

Eksperimentos tika pielietotas trīs pārraudzītās mašīnmācīšanās algoritmi, pārbaudīta precizitāte uz baltvīna datu kopu un savā starpā salīdzināti.

Katram algoritmam atrodod precīzāko pieejamo variantu, ir iespējams secināt, ka visprecīzākais no šiem trijiem algoritmiem, uz šo attiecīgo datu kopu, ir kNN modelis. Neskatoties uz to, visiem algoritmiem bija vismaz 55% precizitāte. Iespējams, ja apmācības datu kopa tiktu palielināta un testēšanas datu kopa samazināta, precizitāte uzlabotos.

References

[1] A. C. F. A. T. M. a. J. R. P. Cortez. [Tiešsaiste]. Available: https://archive.ics.uci.edu/ml/datasets/wine+quality.

[2] A. C. F. A. T. M. J. R. Paulo Cortez, «Modeling wine preferences by data mining from physicochemical properties,» %1 *Decision Support Systems*, sciencedirect, 2009, pp. 547-553.

[3] «Orange,» [Tiešsaiste]. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/kmeans.html.

[4] «Orange,» [Tiešsaiste]. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/hierarchicalclustering.html.

[5] «Orange,» [Tiešsaiste]. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/neuralnetwork.html.

[6] «wikipedia,» [Tiešsaiste]. Available: https://en.wikipedia.org/wiki/Logistic_regression.

[7] O. Harrison, «towardsdatascience,» [Tiešsaiste]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[8] «Orange,» [Tiešsaiste]. Available: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/knn.html.