**Markuss Villers 211RDB196**

**2. Praktiskais darbs**

<https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>

**https://github.com/Markofkaza/2prd**

**Satura rādītājs**

Contents

[I daļa 3](#_Toc134751530)

[Datu kopas apraksts 3](#_Toc134751531)

[II daļa 10](#_Toc134751532)

[III daļa 16](#_Toc134751533)

# I daļa

# Datu kopas apraksts

**datu kopas nosaukums, avots, izveidotājs un/vai īpašnieks :**

Red Wine Quality, <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>, UCI MACHINE LEARNING

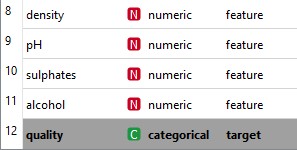
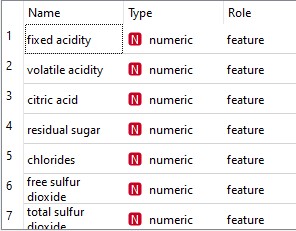
**datu kopas problēmsfēras apraksts:** Doti vairāki vīni, katrs ar saviem kritērijiem, atribūtiem, cenšas izprast, kuri būs tie noteiktie kritēriji un vērtības, kas paceļ vīna kvalitāti uz augšu.

**datu kopas licencēšanas nosacījumi (ja tādi ir):** Atvērta datubāze (jebkurš var izmantot)

**veids, kā datu kopa tika savākta**: Testējot vīnus ar dažādiem kritērijiem kas doti datubāzē, lai nonāktu pie kvalitātes rezultāta, tā skābums, cukurs utt.

**datu objektu skaits datu kopā:** 1599

**datu kopas pazīmju (atribūtu) atspoguļojums kopā ar to lomām Orange rīkā:**



1. Fiksēts skābums, 2. gaistošs skābums, 3. Citronskābe, 4. Atlikušais cukurs, 5. Hlorīdi, 6. Brīvais sēra dioksīds, 7. Kopējais sēra dioksīds, 8. Blīvums, 9. pH līmenis, 10. Sulfāti, 11. Alkahols %, 12. Kvalitāte

**klašu skaits datu kopā, katras klases nozīme un klašu atspoguļošanas veids (klasēm atbilstošo iezīmju skaidrojums); ja datu kopa nodrošina vairākas iespējamas datu klasifikācijas, tad atskaitē skaidri ir jāidentificē, kāda tieši klasifikācija tiek apskatīta darbā:**

**pazīmju (atribūtu) skaits un nozīme datu kopā, kā arī to vērtību tipi un diapazoni (šī informācija būtu jāatspoguļo tabulā, norādot pazīmes (atribūta) apzīmējumu, skaidrojumu, vērtību tipu un datu kopā pieejamo vērtību diapazonu):**

12 atribūti, visiem ir skaitļu vērtības (skatīt 1.attēls)

Quality diapazons: 3-8

Fixed acidity diapazons: 4.6-15.9

Volatile acidity diapazons: 0.120-1.580

Citric acid diapazons: 0-1

Residual sugar: 0.9-15.50

Chlorides: 0.012-0.611

Free sulfur dioxide: 1-72

Total sulfur dioxide: 6-289

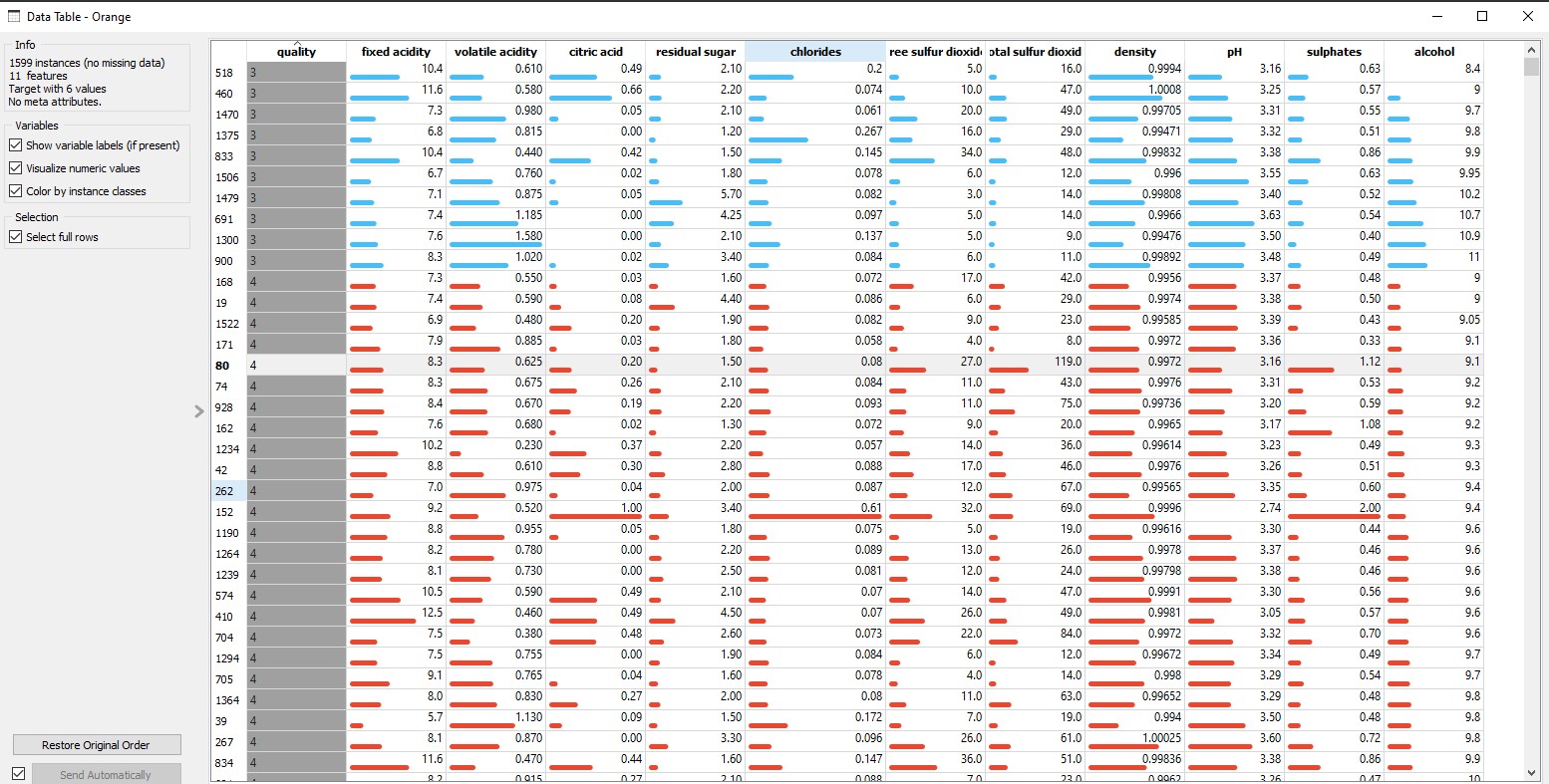
Density: 0.99007-1.00369

pH: 2.74-4.01

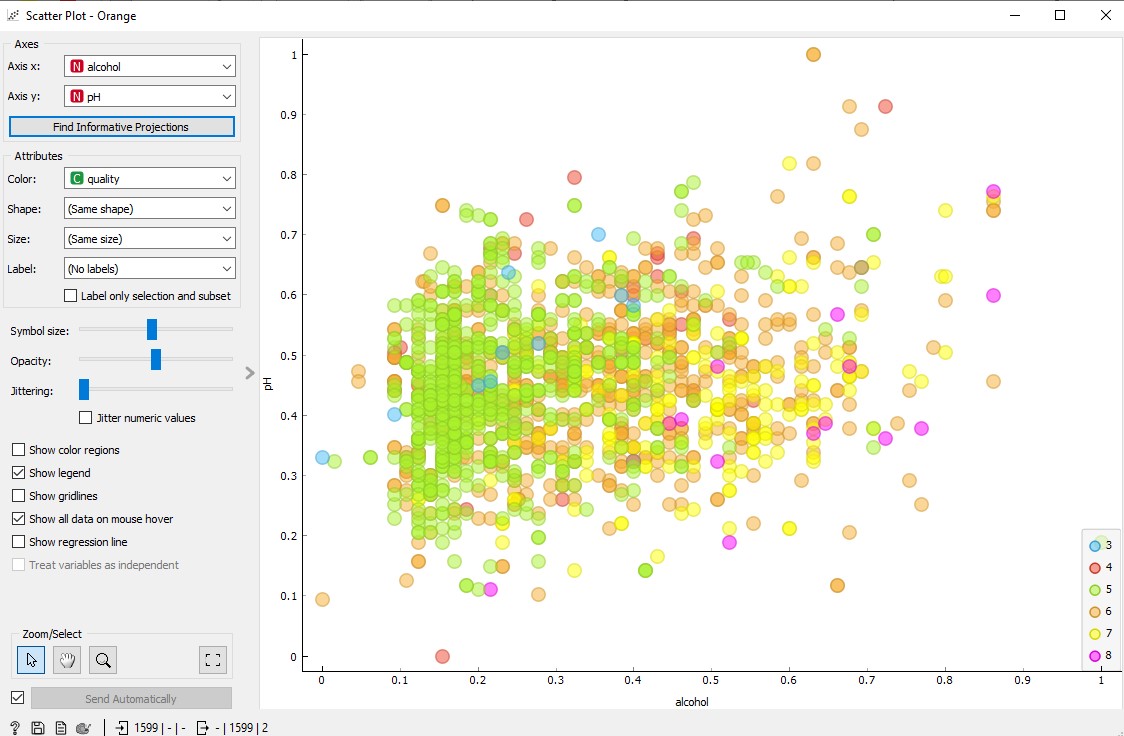
sulphates: 0.33-2

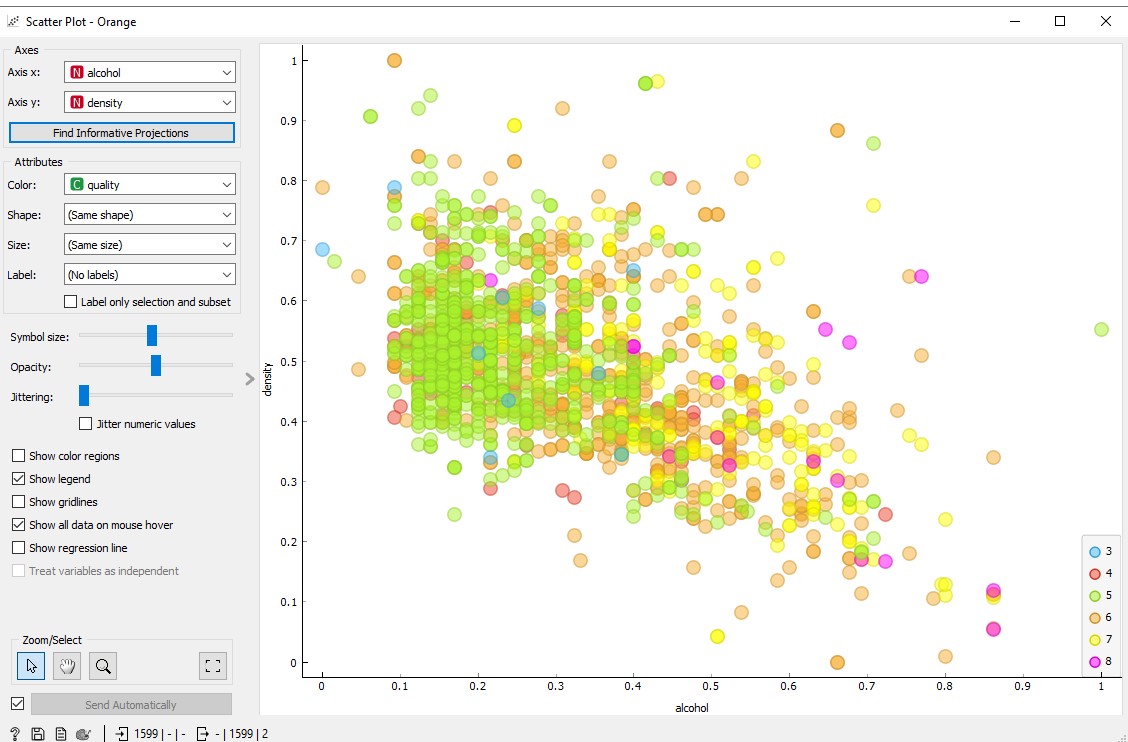
****Alcohol: 8.4-14.9

1.attēls

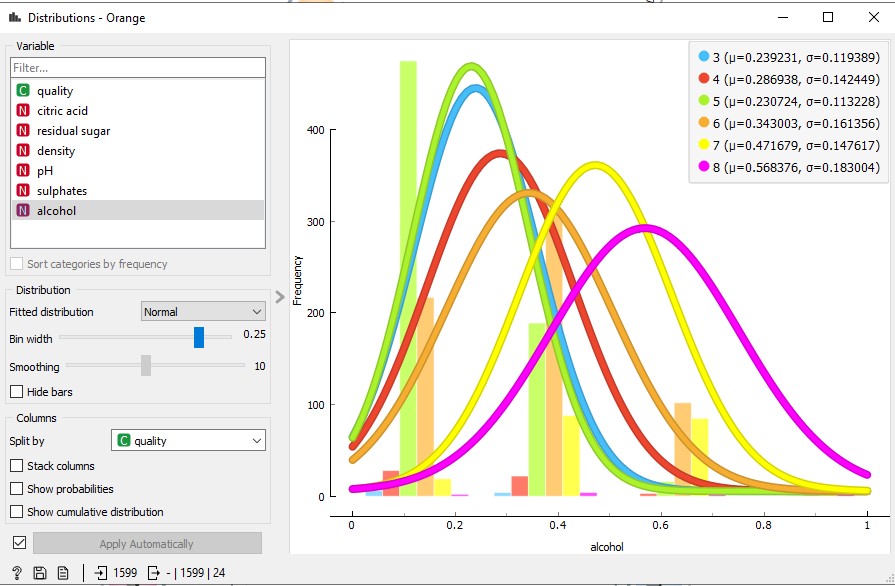
**datu faila struktūras fragments, kurā ir redzamas visas datu faila kolonnas un to vērtības vismaz dažiem datu objektiem: (skatīt 2.attēls)**

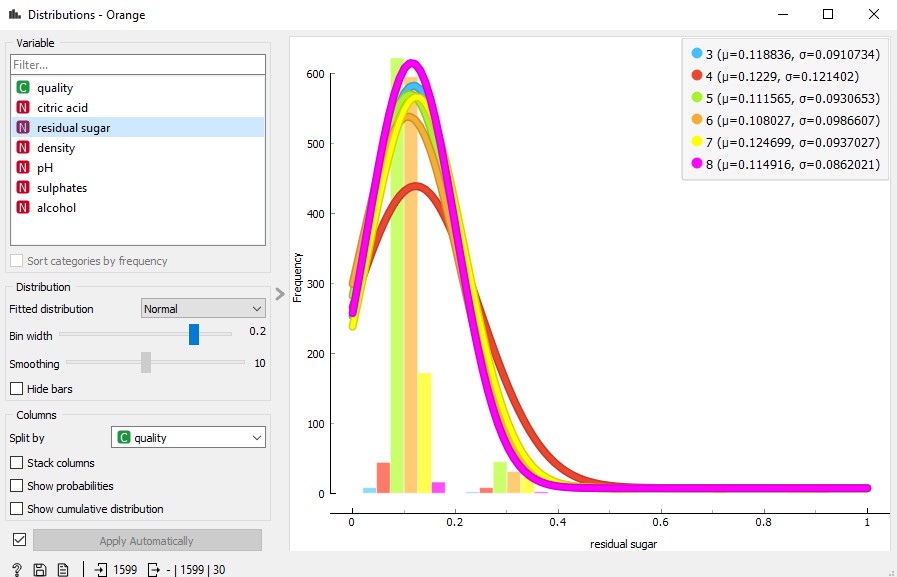
2.attēls

**5. a) **

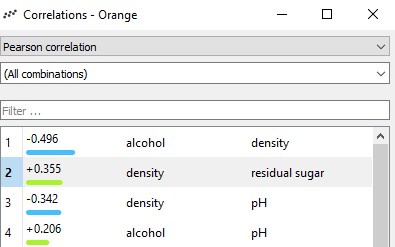


5. b)

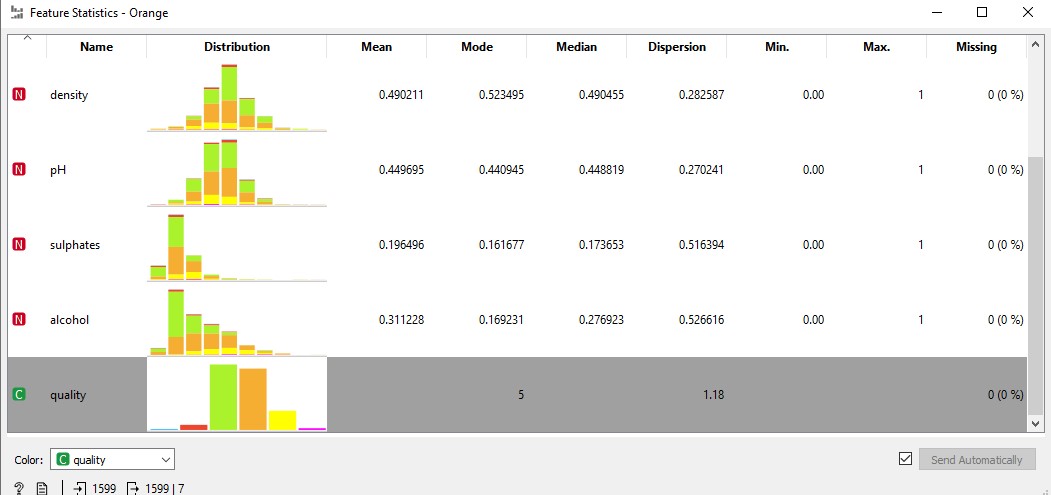




5. c)



5. d)



**secinājumi, kas izriet no izkliedes diagrammu, histogrammu un sadalījumu analīzes (sk. I daļas 5. solis) par datu kopas klašu atdalāmību. Studentiem ir jāatbild uz šādiem jautājumiem:**

*Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)? Tas tiek noteikts, spriežot pēc tā, cik daudz datu objektu pieder katrai kopai.*

Dominē viena, pārējās ir līdzsvarotas, bet esmu skipojis tās.

*Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru? Runa ir par to, vai datu objekti, kuri pieder dažādām klasēm, ir skaidri atdalāmi.*

Dati ir skaidri atdalāmi, uz pirmo skatienu varbūt tā neliekas, bet kad skatās detalizētāk, ir skaidri atdalāmi.

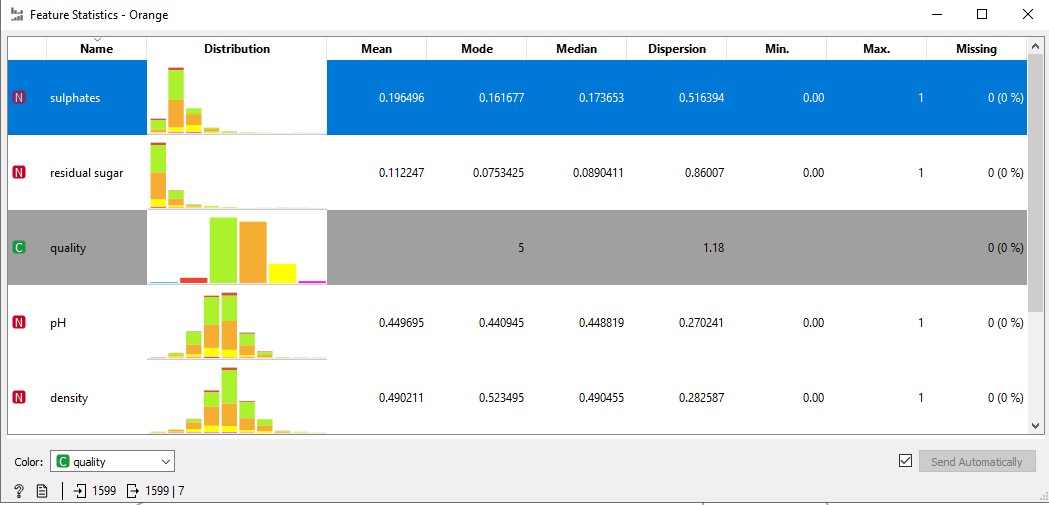
*Cik datu grupējums ir iespējams identificēt, pētot datu vizuālo atspoguļojumu? Runa ir par to, vai ir kaut cik atdalāmi datu grupējumi, ja gadījumā dažādu klašu datu objekti saplūst kopā.*

Ir iespējams identificēt visus datu grupējumus.

*Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?*

Datu grupējumi atrodas sevišķi tuvu viens otram.

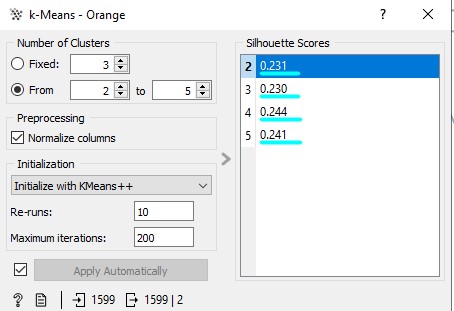
**secinājumi, kas izriet no statistisko rādītāju (vidējo vērtību un dispersijas vērtību) analīzes. (skatīt 3.attēls)**



3.attēls

# II daļa

k-Means algoritma hiperparametru skaidrojums (skatīt 4.attēls)

v

Maksimālais iterāciju skaits

kā tiks izvēlēti centroīdi, gadījuma ceļā vai pilnveidotu metodiku

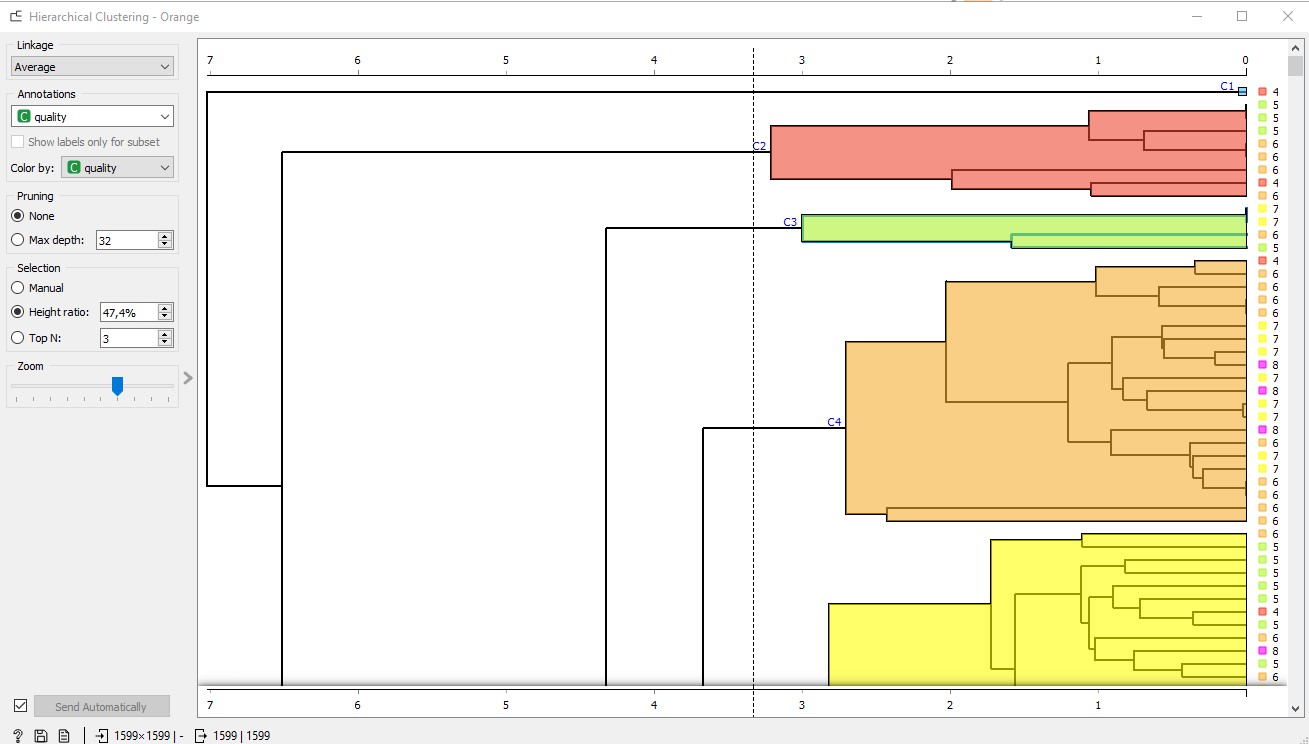
Iespēja veikt datu normalizēšanu

Jo silhouette koeficients tuvāk 1, jo labāk.

Norādīt fiksētu klusteru skaitu vai diapazonu

4.attēls

Hierarchical Clustering hiperparametru skaidrojums (skatīt 5.attēls)



5.attēls

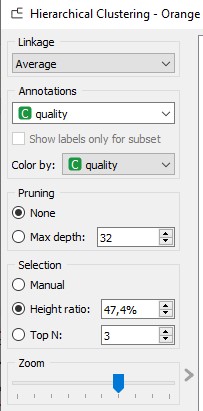
Izmantota, lai atdalītu vai sasaistītu klases, tiktu grupēti kopā, mazāk attiecināti uz citas klases kategorijas klasteriem

Saistīšanas metode

(Single) - Vienkārša saistīšana

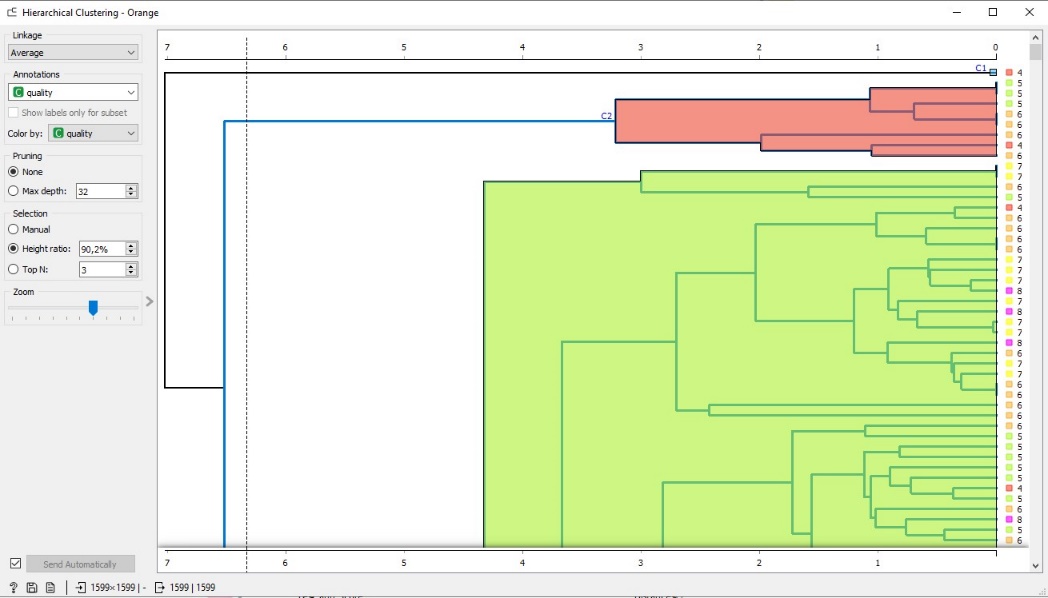
(Average) – vidēju vērtību balstīta

(Complete) – Pilnīga saistīšana



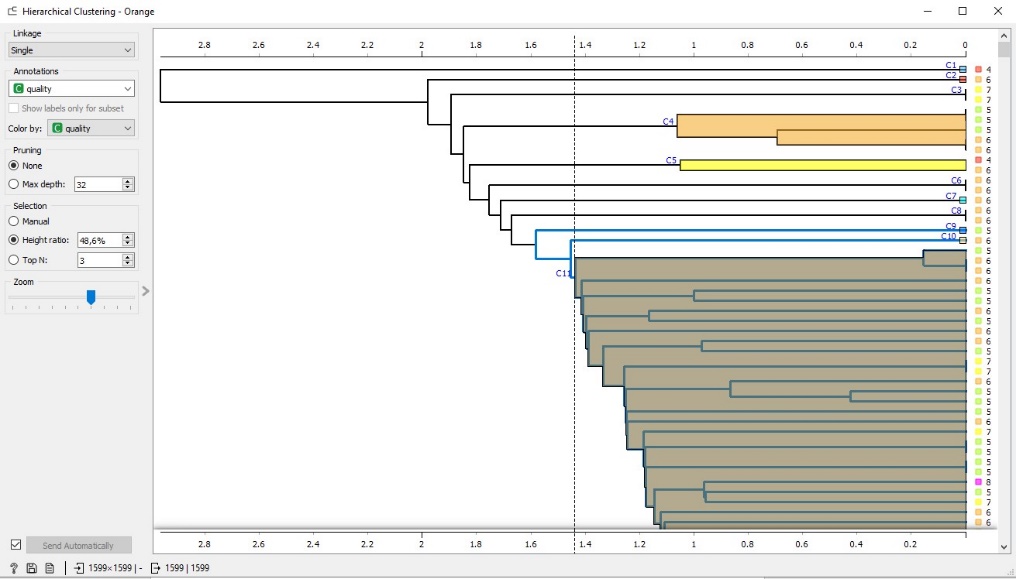
5.attēls

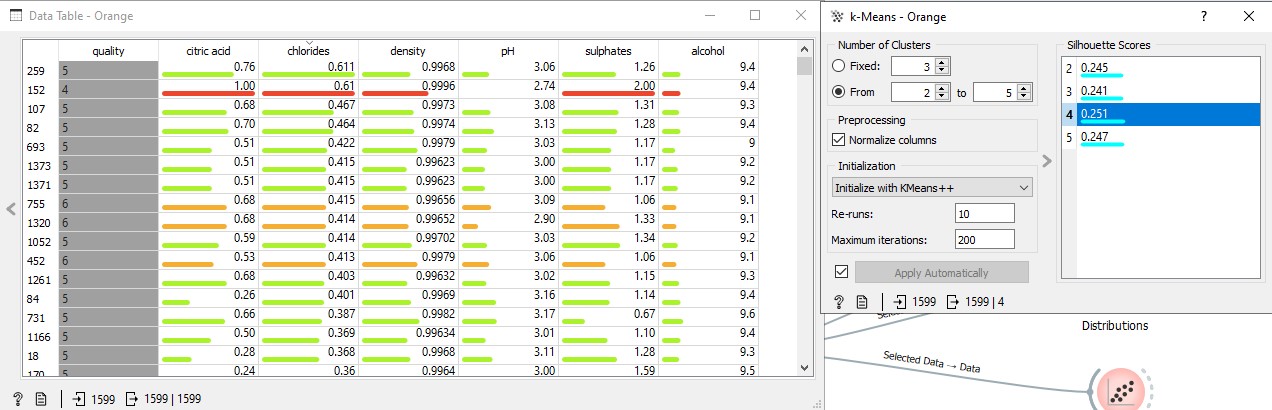
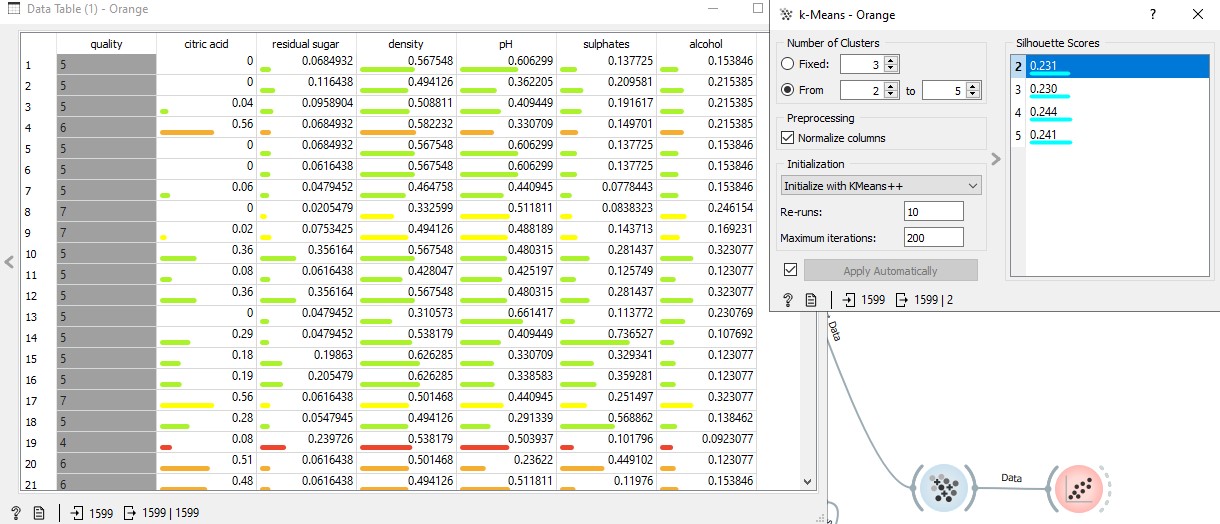
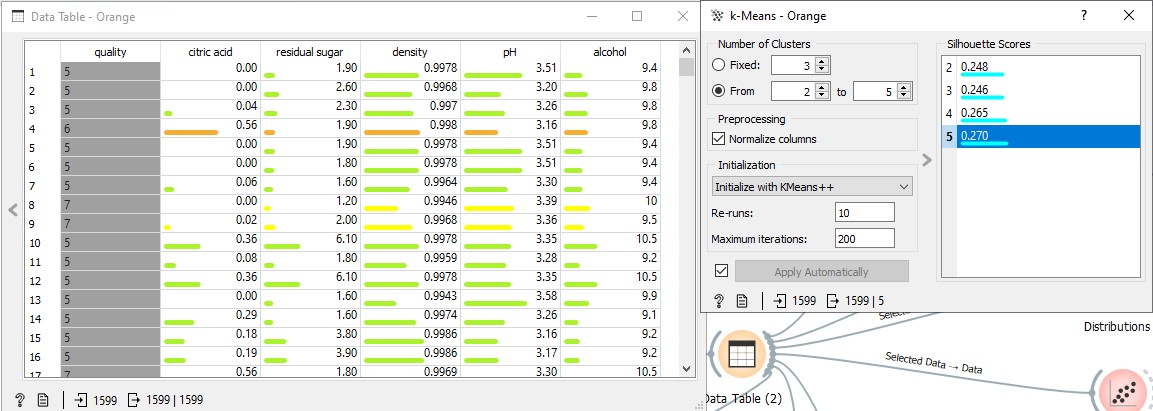
Hierarchical clustering tests ar vairākiem hiperparametriem (skatīt 6.attēls)



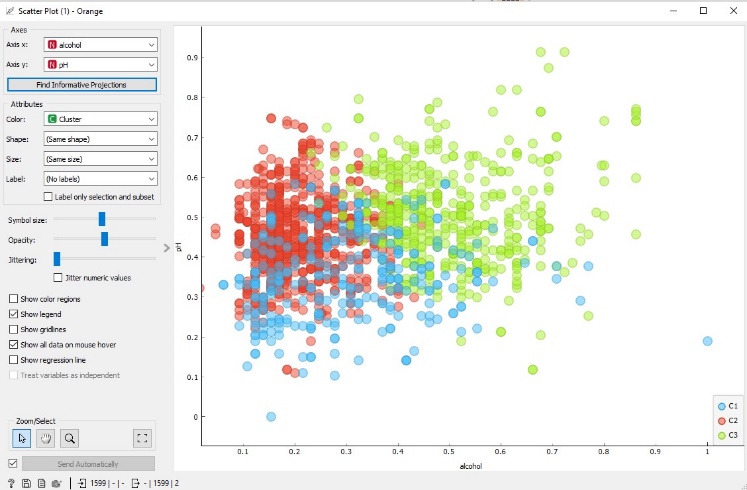
6.attēls





k-means hiperparametri (skatīt 9.attēls)

9.attēls

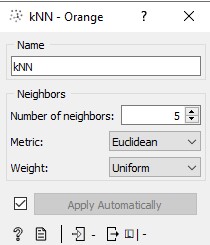


Secinājums:

Pēc algoritma analīzes un izvērtēšanas tika izvēlēti labākie rezultāti, Klases ir labi atdalāmas, koeficients sevišķi neatšķirās, Tāpat, ja koeficients neatšķiras daudz, pat pamainot atribūtus, tas var norādīt, ka ievades dati ir diezgan vienmērīgi sadalīti visos atribūtos.

# III daļa

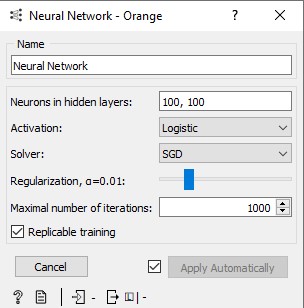
Knn algoritma hiperparametri (skatīt 7.attēls)



Izmantotā metrika (Euclidean/Manhattan/Chebyshev/Mahalanobis

Pārbaudāmo kaimiņu skaits

7.attēls

Neironu tīkla hiperparametri (skatīt 8.attēls)

Maksimālais iterāciju skaits

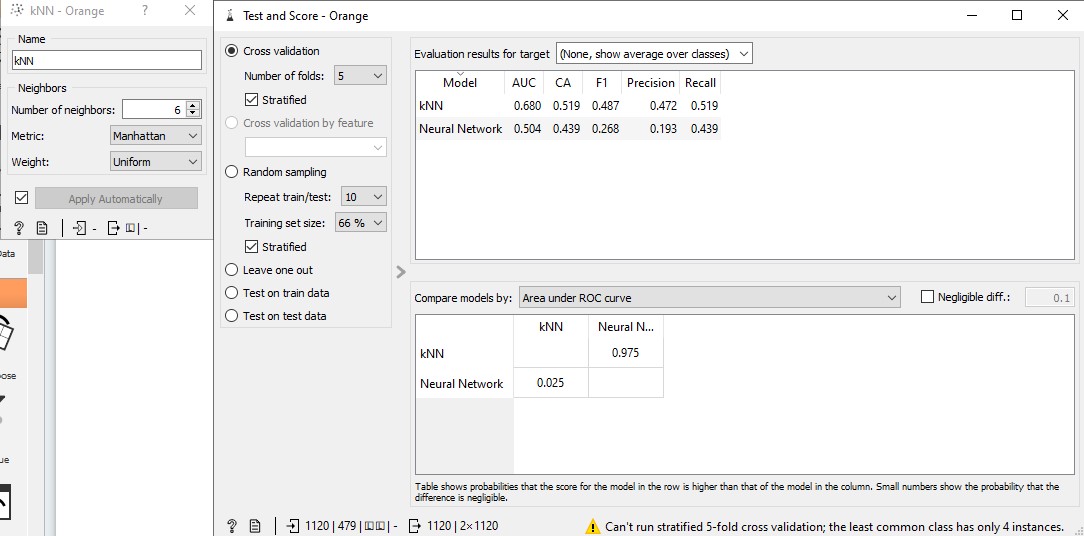
Mācīšanās ātrums

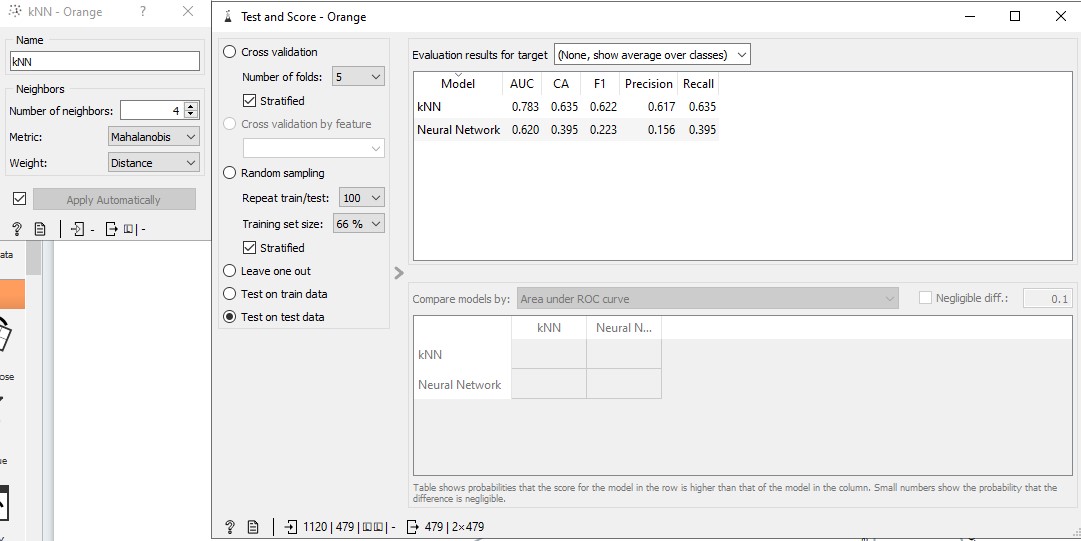
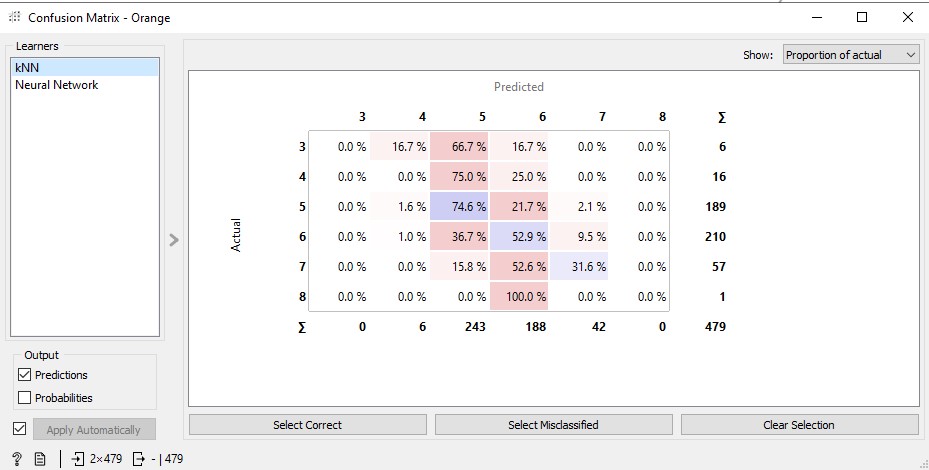
Izmantotā aktivizācijas funkcija

Neironu skaits, katrā no slēptajiem slāņiem, tos atdalot ar komatu.

8.attēls

eksperimentos izmantotās hiperparametru vērtības (skatīt 9.attēls)





9.attēls