

Academia de Studii Economice din București

Facultatea de Cibernetică, Statistică și Informatică Economică

Specializarea Informatică Economică

Proiect la Dezvoltare Software pentru analiza datelor

Clasificarea tipurilor de vin după calități și substanțe chimice componente (Tema 3 - Analiză de clusteri)

Coordonator

Prof. univ. Furtună Titus Felix

Student

Cazacu Irina Claudia

Grupa 1081, Seria C

București 2023

Cuprins

[Setul de date 1](#_Toc124936603)

[Partiția optimală 2](#_Toc124936604)

[Partiția cu 3 clusteri 7](#_Toc124936605)

[Concluzii 12](#_Toc124936606)

[Bibliografie 12](#_Toc124936607)

# Setul de date

Lucrarea de față își propune să clasifice mostre de vin în funcție de calitățile și de substanțele chimice componente. Complexitatea componentelor chimice ce alcătuiesc vinul conferă numeroase proprietăți precum culoarea, paleta sau texturile, aciditatea sau astringența. Aceste componente se pot încadra în 4 mari categorii: vitamine, minerale, nutrieți sau acizi. Ele pot proveni din calitățile boabei de strugure sau se pot dezvolta din procesul de fabricare sau din maturarea vinului.

În acest sens, am supus spre analiză un set de date preluate de pe site-ul <https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering> ce conține 14 coloane cu date rezultate din analiza chimică a mostrelor de vin provenite de la 3 producători diferiți din Italia. Astfel, fișierul WineQualities.csv cuprinde 178 de observații, structurate în 14 coloane, reprezentând:

1. Alcohol – procentul de alcool, ce variază între 10 % și 15%
2. Malic acid – cantitatea de acid malic ce dă aciditatea vinului, care poate ajunge până la 6.5 g/l
3. Ash – cantitatea de ”cenușă” încadrată în categoria cationilor, ce variază între 1.3 și 3.2 g/l
4. Alcalinity of ash – alcalinitatea cenușei responsabilă pentru gustul și oxidarea vinului, ce se măsoară ca grame de carbonat de potasiu per litru și variază între 10.5 și 30
5. Magnesium – cantitatea de magneziu ce dă stabilitate vinului și variază între 70 și 160 g/l
6. Total phenols – cantitatea totală de fenoli ce dă aroma vinului din care se pot aminti vanilina ce influențează gustul, antocianina ce oferă pigmentația roșie vinului, sau resveratrolul ce contribuie la scăderea colesterlolului, variând între 1 și 4 g/l
7. Flavanoids – cantitatea de flavanoizi ce oferă proprietăți antioxidante vinului roșu, variând între 0.3 și 5 g/l
8. Nonflavanoid phenols – cantitatea de fenoli non-flavanoizi care influențează culoarea albă a vinului, variând între 0.1 și 0.6 g/l
9. Proanthocyanins – cantitatea de tanină ce dă astringența vinului și variază între 0.4 și 3.5 g/l
10. Color intensity – intensitatea culorii vinului ce variază între 1.3 și 13
11. Hue – nuanța culorii vinului (alb, rose sau roșu) ce variază între 0.5 Și 1.7
12. OD280/OD315 of diluted wines – raportul de puritate al vinului ce variază între 1.3 și 4
13. Proline - cantitatea de prolină ce dă vâscozitate vinului și influențează acumularea de aminoacizi, variând între 278 și 1680
14. Cultivar – cultivatorul, mai exact unul dintre cei 3 producători de la care au fost preluate mostrele

În cele ce urmează, se vor studia principalele componente chimice din vin și se vor crea clustere pe baza acestora, utilizând algoritmi ierarhici. Astfel, pentru setul de date ales și pentru a îndeplini obiectivele studiului ales, am utilizat analiza de clusteri, pentru a determina grupările naturale de date. Această metodă exploratorie pune accent pe înregistrarea relațiilor dintre indivizi și factori, notate ca disimilarități. Astfel, pentru linkage s-a folosit metoda Ward prin care se asigură maximizarea gradului de omogenite dintre clustere și pe minimizarea diferențelor din interiorul clusterelor utilizând descompunerea varianței din și dintre clustere.

# Partiția optimală

Am început prin a citi setul de date din fișierul CSV și a-l prelucra pentru a obține tabelul de date și denumirile factorilor ce dau calitatea vinului.

**Componența partiției**

Utilizând clasa hclust, am ierarhizat, calculând distanțele dintre clusteri și împărțind pe partiții. De asemenea, am redus dimensionalitatea setului de date pe baza celor 2 axe principale. Astfel, pentru partiția optimală, s-au obținut cei 2 clusteri (c1 și c2) cu mostrele de vin componente. S-a constatat faptul că în clusterul 1 se regăsesc probe cu valori mai ridicate ale prolinei, ale alcalinității sau ale intensității culorii, în timp ce clusterul 2 se diferențiază prin valorile mai crescute date de acid malic.

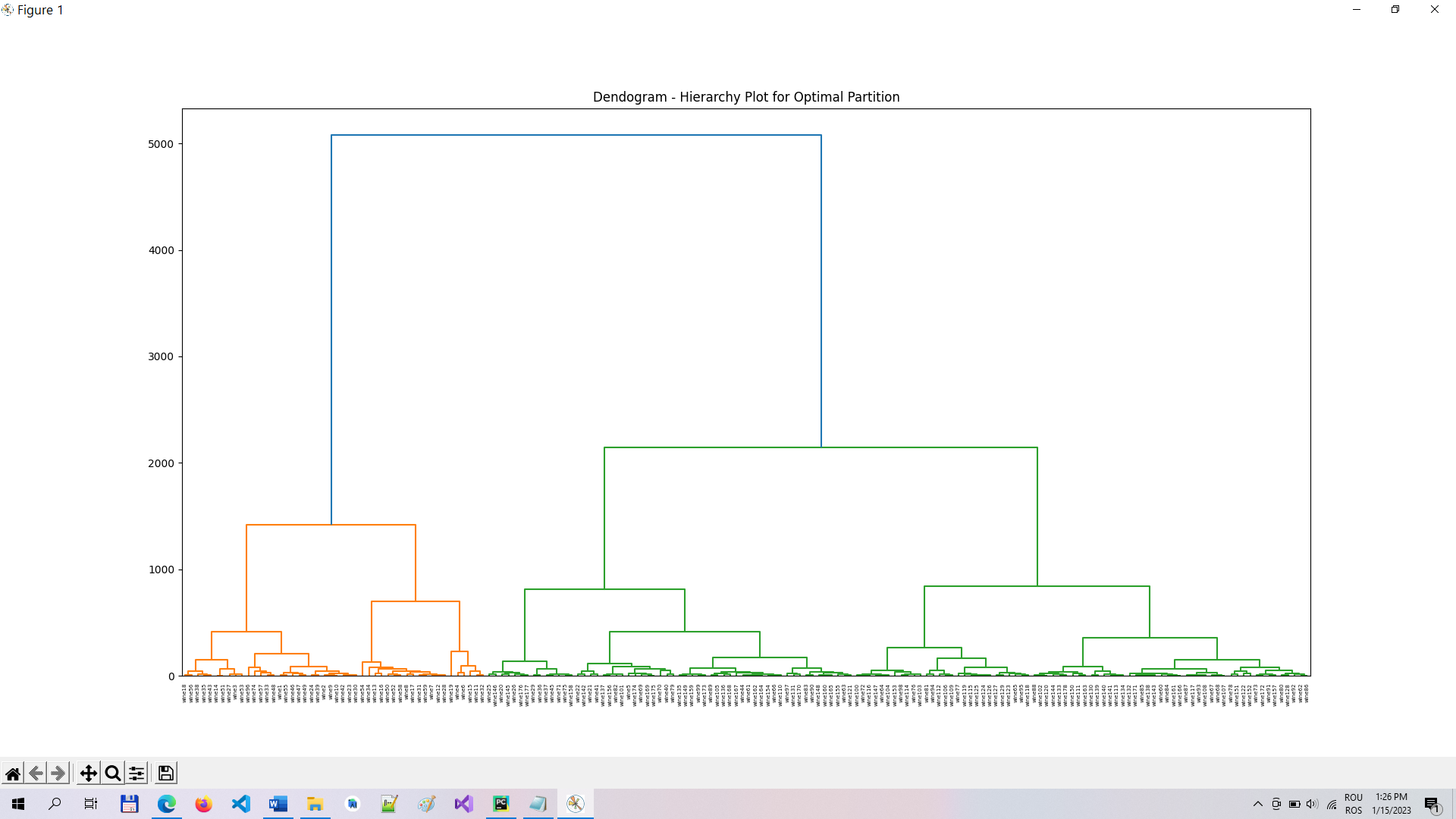
A screenshot of a computer

Description automatically generated

*Figura 1 – Componența partiției optimale*

**Graficul dendrogramă**

Odată calculată ierarhia clusterelor, am putut realiza graficul dendogramă, cu Ox reprezentând axa indivizilor și Oy reprezentând axa distanțelor dintre clustere. Astfel, clusterele de la baza graficului au o omogenitate mult mai ridicată decât cele din vârful ierarhiei. De asemenea, este evidențiat un ”dezechilibru” privind volumul clusterelor, întrucât clusterul 2 conține mult mai multe probe față de clusterul 1 (130 în c1 față 48 în c2).



*Figura 2 – Dendograma pentru partiția optimală*

**Plot instanțe pe clusteri în primele două axe principale**

De asemenea, pentru a vizualiza mai bine împărțirea pe componente, am realizat un scatter plot, din care reiese distribuția indivizilor pe cele 2 clustere. Astfel, în partea din dreaptă sunt concentrați indivizii ce aparțin primului cluster, iar în cea stângă se observă componența celui de-al doilea cluster. Se remarcă, de asemenea, omogenitatea mai crescută în secțiunea stânga-jos a clusterelor și existența unor outlieri precum wine96 sau wine19 în clusterul 1 și wine90 sau wine70 în clusterul 2.

Chart, scatter chart

Description automatically generated

*Figura 3 – Plot instanțe pe clusteri în primele două axe principale pentru partiția optimală*

**Histograme și plot distribuții pe clusteri pentru fiecare variabilă**

În plus, pentru fiecare dintre factorii de influență ai calității am reprezentat histograme și grafice de distribuție pentru cei doi clusteri. În acest mod sunt illustrate mai clar disimilaritățile dintre clusteri. Astfel, se observă că în clusterul 1 se încadrează preponderent mostre de la producătorul 1, iar în cel de-al doilea se regăsesc mostre de la cultivatorii 2 și 3. De asemenea, clusterul 2 se observă frecvențe mai ridicate ale cantităților de prolină de până la 1000g/l, de magneziu, de până în 100 g/l sau de flavanoizi, de până la 3 g/l, iar în clusterul 1 frecvențele crescute sunt la cantitățile de prolină de peste 1000g/l, de magneziu, de peste 100 g/l sau de flavanoizi, de peste 3 g/l. Totuși, în ambele clustere există similarități privind frecvența maximă a proantocianinei între 1.5-1.7 g/l sau a cationilor, între 2.3-2.5 g/l.

Chart

Description automatically generated with low confidence

*Figura 4 – Distribuția celor 14 factori în cei 2 clusteri ai partiției optime*

Chart, histogram

Description automatically generated

*Figura 5 – Histograma variabilei Proanthocyanins pentru cele 2 clustere ai partiției optimale*

Chart, histogram

Description automatically generated

*Figura 6 – Histograma variabilei Magnesium pentru cele 2 clustere ai partiției optimale*

# Partiția cu 3 clusteri

**Componența partiției**

Utilizând din nou clasa hclust am ierarhizat instanțele, utilizând de această dată o partiție formată din 3 clusteri. Astfel, s-au obținut cei clusterii c1, c2, și c3 cu mostrele de vin componente. S-a constatat faptul că în clusterul 3 se regăsesc probele cu cele mai ridicate valori ale prolinei, în clusterul 2 se află mostrele cu cea mai mica intensitate a culorii, sau a cantității de magneziu, iar în clusterul 3 se observă vinurile cele mai roșii și cu cel mai mic procent de alcool.

A screenshot of a computer

Description automatically generated

*Figura 7 – Componența partiției optimale*

**Graficul dendrogramă**

In mod similar, după calcularea partițiilor si a ierarhiei, am putut reprezenta graficul dendogramă, în care se poate observa o mai bună echilibrare a clusterelor ce conțin un număr mai apropiat de indivizi (c1-58, c2-62, c3-38) și au distanțele maxime mai apropiate, de aproximativ 1000.

Chart, box and whisker chart

Description automatically generated

*Figura 8 – Dendograma pentru partiția cu 3 clusteri*

**Plot instanțe pe clusteri în primele două axe principale**

Totodată, pentru a vizualiza mai bine împărțirea pe cele 3 noi componente, am realizat un scatter plot, ce ilustrează distribuția indivizilor pe cele 3 clustere. Astfel, în partea centrală sunt concentrați indivizii ce aparțin primului cluster, în cea stângă se observă componența celui de-al doilea cluster, iar în cea dreaptă este reprezentat al treilea cluster. Se remarcă, de asemenea, omogenitatea mai crescută în secțiunea stânga-jos a clusterelor și existența unor outlieri precum wine96 sau wine19 în clusterul 3 și wine90 sau wine70 în clusterul 1.

Scatter chart

Description automatically generated with medium confidence

*Figura 9 – Plot instanțe pe clusteri în primele două axe principale pentru partiția cu 3 clusteri*

**Histograme și plot distribuții pe clusteri pentru fiecare variabilă**

Nu în ultimul rând, pentru fiecare dintre factorii de influență ai calității am reprezentat histograme și grafice de distribuție pentru cei trei clusteri. În acest mod sunt illustrate mai clar disimilaritățile dintre clusteri. Astfel, se observă că în clusterul 1 se încadrează preponderent mostre de la producătorul 3, în cel de-al doilea se regăsesc mostre de la cultivatorul 2, iar în ultimul cluster se află mostrele de la producătorul 1. De asemenea, clusterul 2 se observă frecvențe mai ridicate ale cantităților de prolină între 500-1000 g/l, de flavanoizi, de aprox 1 g/l, sau a gradului de puritate de aprox 3 g/l, în clusterul 1 frecvențele crescute sunt la cantitățile de prolină de peste 1000g/l, de flavanoizi, de aprox 2 g/l, sau a gradului de puritate de sub 2 g/l iar în clusterul 3 frecvențele maxime sunt la cantitățile de prolină între 1000-1500 g/l, de flavanoizi de aprox 3 g/l sau a gradului de puritate de peste 3 g/l. Totuși, în ambele clustere există similarități privind frecvența maximă a proantocianinei între 1.5-2 g/l, a cationilor, între 2.3-2.5 g/l sau a fenolilor nonflavanoizi, între 0.3-0.5 g/l.

Chart

Description automatically generated

*Figura 10 – Distribuția celor 14 factori în cei 3 clusteri ai partiției*

Chart, histogram

Description automatically generated

*Figura 11 – Histograma variabilei Color Inensity pentru cele 3 clustere*

Chart, histogram

Description automatically generated

*Figura 6 – Histograma variabilei Flavanoids pentru cele 3 clustere*

De asemenea, cele două tabele de partiții au fost salvate în fișierul Partitions.csv, în care fiecărei mostre de vin îi sunt asociate clusterii de care aparține.

# Concluzii

Mostrele de vin au fost grupate după proprietățile și substanțele chimice cuprinse în setul de date. S-a constat faptul că mostrele în care s-au măsurat cantități mari de prolină, de proantocianină, de flavanoizi sau cu puritate și instensitate a culorii mai crescute au constituit clustere din care reiese o calitate mai bună a vinului. Pe de altă parte, caracteristici ale vinului precum alcalinitatea, și cationii, fenolii non-flavanoizi sau proantocianina s-au regăsit cu frecvențe maxime pentru aceleași valori măsurate, indiferent de clustere.

# Bibliografie

* Curs 12 și seminar 12 – Prof. Titus Felix Furtună
* <https://www.youtube.com/watch?v=l4vTwXL_5Cc>
* <https://www.youtube.com/watch?v=v7oLMvcxgFY>
* <https://www.askpython.com/python/examples/hierarchical-clustering>
* <https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/>