02 Domaci Klasterizacija

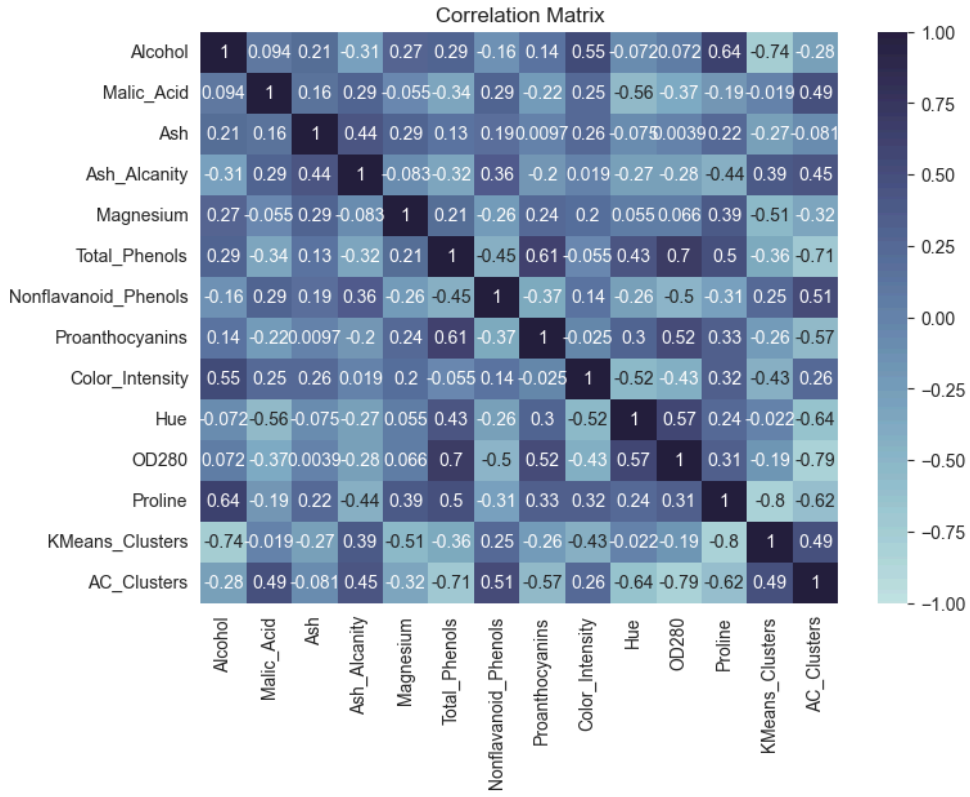
Predmet: Tehnike i metode analize podataka

Student 1636 Milica Jovanovic

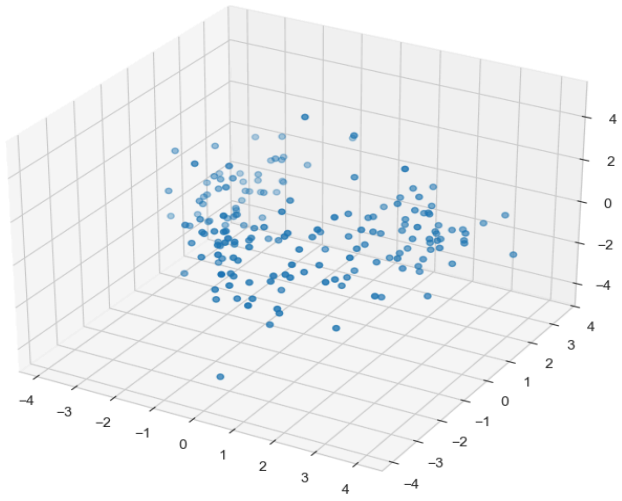
Projekat se bavi analizom vina korišćenjem metoda klasterovanja kako bi se identifikovale različite grupe na osnovu njihovih hemijskih karakteristika. Dataset sadrži 178 uzoraka vina sa područja Italije, od kojih su analizirane 13 hemijskih karakteristika poput sadržaja alkohola, malinske kiseline, flavonoida, proantocijanina, intenziteta boje i drugih. Cilj projekta bio je istražiti moguće grupe među uzorcima vina i interpretirati njihove zajedničke osobine.

Nakon učitavanja podataka iz dataset-a, sprovedena je analiza kako bi se identifikovale potencijalne greške, nedostajuće vrednosti i duplikati. Analiza je pokazala da ne postoje nedostaci ili duplikati u podacima. Vizualizacija distribucija atributa pomoću histograma i boxplot-ova ukazala je na odsustvo anomalija.

Korelaciona matrica je pokazala značajnu povezanost između atributa "Flavanoids" i "Total_Phenols", pa je kolona "Flavanoids" uklonjena kako bi se izbegla multikolinearnost.



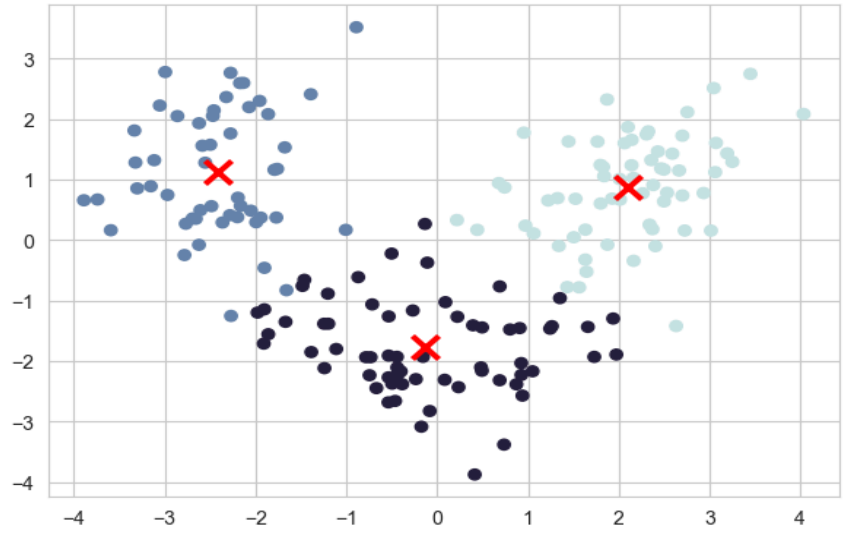
Podaci su standardizovani korišćenjem *StandardScaler* radi poboljšanja performansi algoritama klasterovanja. Zatim je primenjena *Principal Component Analysis (PCA)* kako bi se dimenzionalnost smanjila na tri komponente. Ove komponente su vizualizovane u 3D prostoru, gde su uzorci grupisani prema svojim sličnostima.

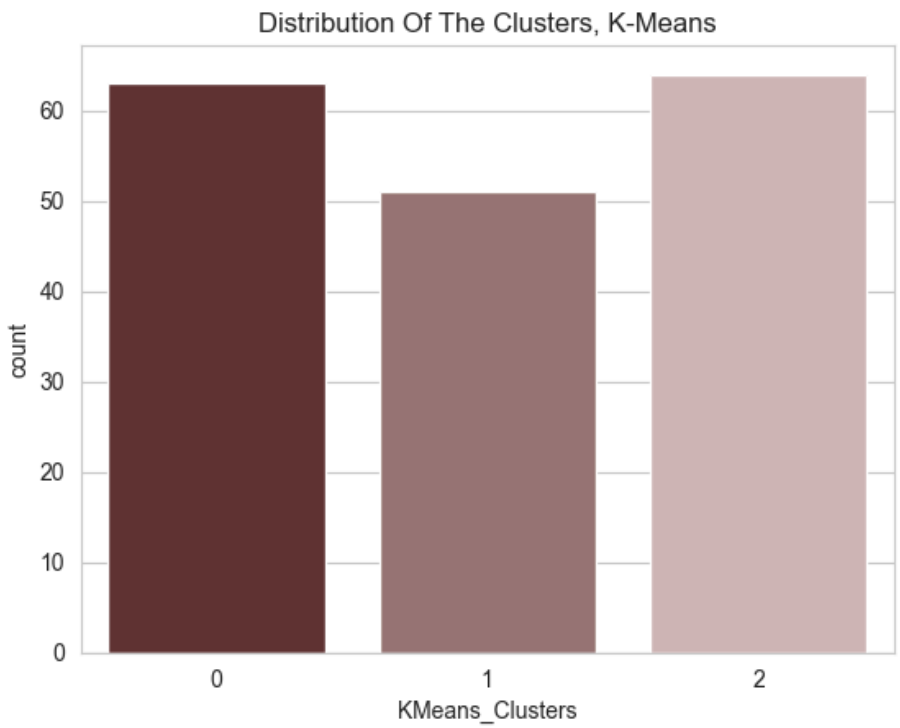


Za klasterovanje je primenjen algoritam *K-Means*, a optimalan broj klastera identifikovan je korišćenjem metode "lakta".

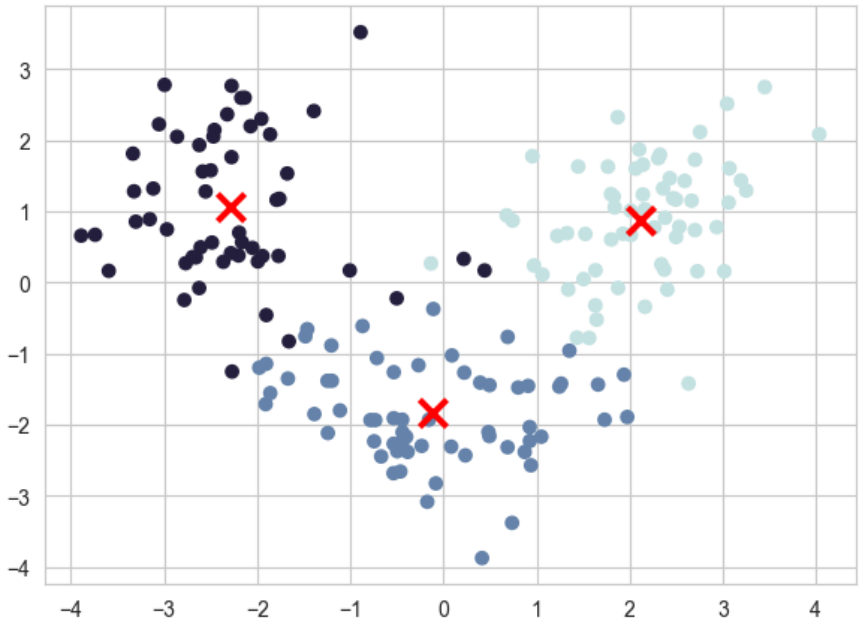
Na osnovu rezultata, uzorci su podeljeni u tri klastera. Svaki klaster je analiziran na osnovu karakteristika koje ga definišu:

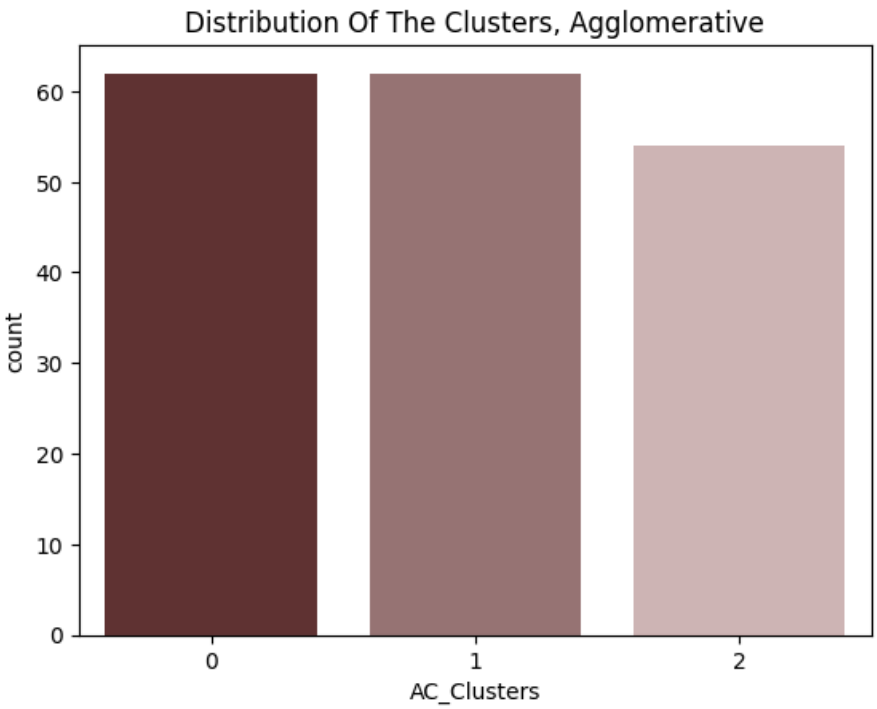
- **Klaster 0:** Vina sa nižim sadržajem alkohola, umerenim flavonoidima i prolinom, svetlijih nijansi i niskog intenziteta boje. Ova vina su blaža i tanja u teksturi.
- **Klaster 1:** Vina sa umerenim nivoima alkohola, nižim flavonoidima i prolinom, ali intenzivnijih boja. Ova vina su jača od klastera 0, ali manje gorka.
- **Klaster 2:** Vina sa visokim sadržajem alkohola, najviše flavonoida i prolinom, tamnijih nijansi i umerenog intenziteta boje. Ova vina su najjača i punog tela, sa izraženom gorčinom.



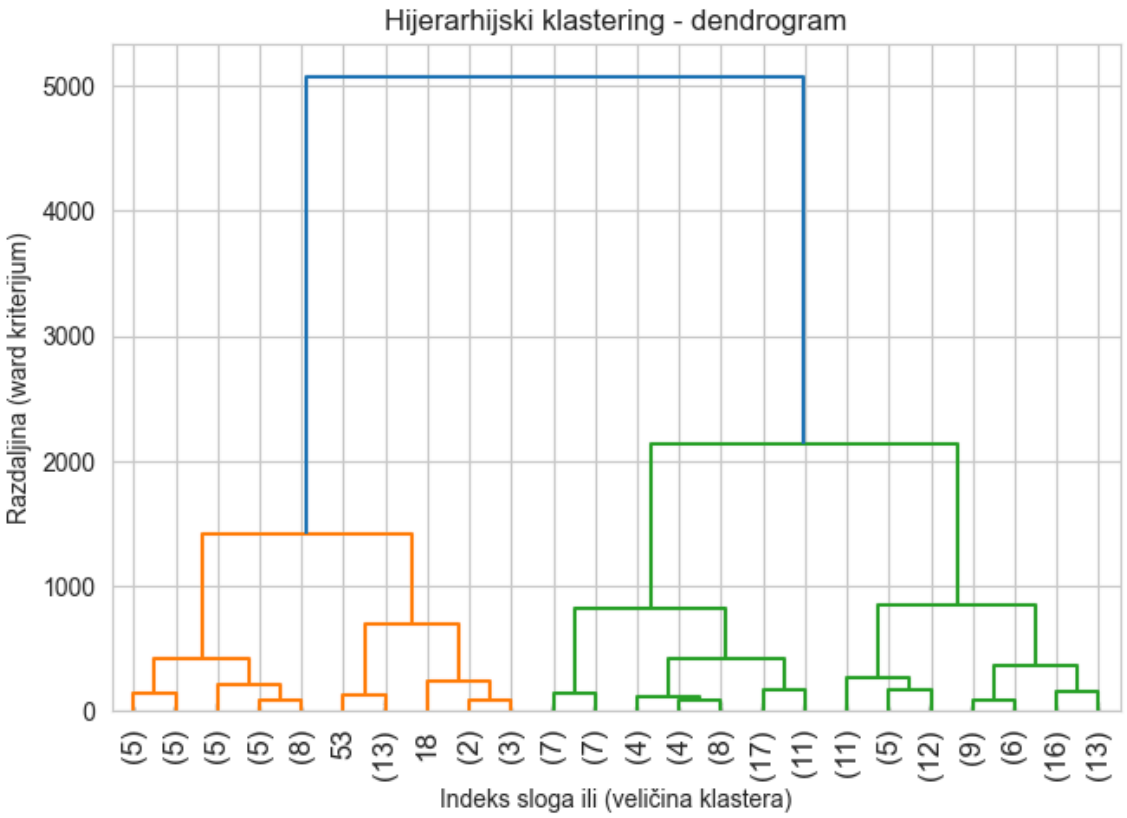


Pored *K-Means*, primenjen je i hijerarhijski klastering (Agglomerative Clustering), ali se rezultati nisu značajno slagali sa K-Means metodom, sa nivoom tačnosti od približno 35%.





Hijerarhijski klastering je vizualizovan dendrogramom, što je omogućilo dodatnu interpretaciju strukture podataka.


Hijerarhijski klastering - dendrogram
Razdaljina (ward kriterijum)
5000
4000
3000
2000
1000
0
(5)
(5)
(5)
(5)
(8)
53
(13)
18
(2)
(3)
(7)
(7)
(4)
(4)
(8)
(17)
(11)
(11)
(5)
(12)
(9)
(6)
(16)
(13)
Indeks sloga ili (veličina klastera)