ZHLUKOVÁ ANALÝZA VÍN

Cieľ

Zámerom tejto analýzy bolo identifikovať prirodzené skupiny vín na základe ich chemických vlastností a charakterizovať tieto skupiny pre lepšie pochopenie rôznych typov vín. Na identifikáciu týchto skupín sme použili metódu K-means.

Dáta

Dáta pochádzajú z UC Irvine Machine Learning Repository a obsahujú výsledky chemických analýz vín, ktoré boli pestované v rovnakom regióne v Taliansku, ale pochádzajú od troch rôznych pestovateľov. V datasete je zahrnutých 13 atribútov, ktoré predstavujú rôzne chemické vlastnosti vín.

Atribúty:

Premenná EN	**Premenná SK**
Alcohol	obsah alkoholu
Malic Acid	obsah kyseliny jablčnej
Ash	popol
Alcalinity of ash	zásaditosť popola
Magnesium	obsah horčíka
Total phenols	celkové fenoly
Flavonoids	flavonoidy
Nonflavonoid phenols	neflavonoidné fenoly
Proanthocyanins	proantokyaníny
Color intensity	intenzita farby
Hue	odtieň
OD280/OD315 of diluted wines	pomer optickej hustoty
Proline	obsah prolínu

Spracovanie dát

Model bol vytvorený v programe Jamovi s využitím modulov pre zhlukovú analýzu. Po úvodnej analýze sme sa rozhodli odstrániť dve premenné, ktoré sa javili ako menej významné pre diferenciáciu klastrov: Ash a Alcalinity of ash.

K-means

K-means je populárna metóda zhlukovej analýzy, ktorá rozdeľuje pozorovania do vopred určeného počtu skupín na základe ich vzájomnej podobnosti. Algoritmus funguje tak, že najprv náhodne vyberie počiatočné body, tzv. centroidy, a potom priradí každý bod k najbližšiemu centroidu, čím vytvorí predbežné zhluky. V ďalšom kroku prepočíta pozíciu centroidov ako priemer všetkých bodov v danom zhluku a opätovne priradí body k novým centroidom, pričom tieto kroky sa opakujú, kým sa pozície centroidov nestabilizujú. K-means sa snaží minimalizovať vzdialenosť bodov od centroidov jednotlivých zhlukov, takže výsledné zhluky obsahujú body, ktoré sú si vzájomne čo najpodobnejšie.

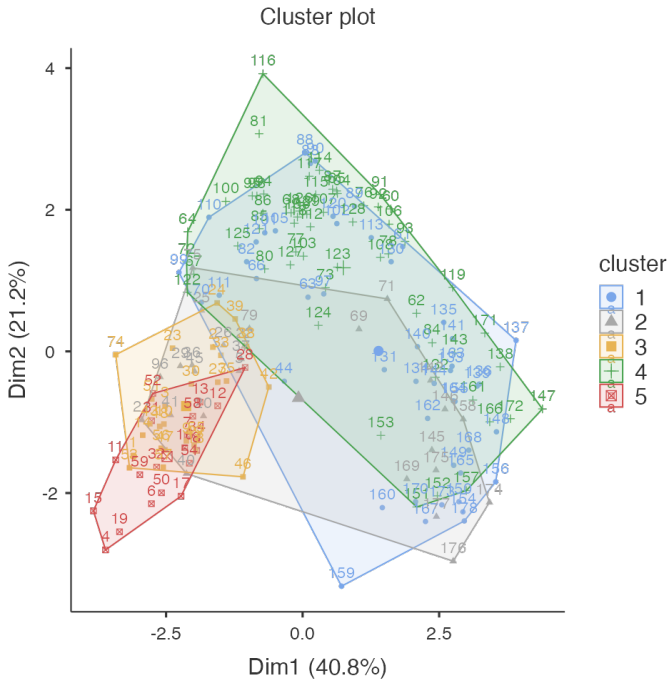
sPre našu zhlukovú analýzu sme zvolili 5 klustrov, aby sme zachovali interpretovateľnosť jednotlivých zhlukov. Charakteristiku jednotlivých zhlukov uvádzame v tabuľke 1.

Tabuľka 1. Zhluky vín podľa metódy K-means

	Alcohol	Malic Acid	Mg	Total Phenols	Flavanoids	Nonflavanoid Phenols	Proan	Color Intensity	Hue	OD280	Proline
1	13.719	1.968	104.667	2.837	2.964	0.278	1.912	5.243	1.050	3.192	1072.407
2	12.475	2.325	91.719	2.106	1.871	0.383	1.468	3.952	0.961	2.544	435.579
3	13.178	2.538	111.731	2.282	1.889	0.359	1.661	5.425	0.904	2.632	823.577
4	13.921	1.769	106.650	2.908	3.082	0.295	1.908	6.323	1.117	3.008	1360.850
5	12.742	2.684	97.125	1.967	1.328	0.413	1.386	5.542	0.865	2.189	636.125

Pozn: Mg = Magnesium, Proan = Proanthocyanins

Graf 1. Zhluky vín podľa metódy K-means



Kvalita modelu

Tabuľka 2. Veľkosti zhlukov

Zhluk	počet
1	27
2	57
3	26
4	20
5	48

Model vysvetľuje približne 95% variability, čo je veľmi dobrý výsledok. Z hľadiska veľkosti sú zhluky relatívne vyvážené.

Popis klastrov

Zhluk 1: "Prémiové archívne" (27 vzoriek)
Charakteristika: Vysoký obsah alkoholu (13,72%), vysoké hodnoty fenolických látok (Total_Phenols 2,84, Flavonoids 2,96), výrazne vysoký obsah Prolínu (1072), dobrý pomer Flavanoids/Color_Intensity
Ide o komplexné vína s dobrým potenciálom k archivácii s vyváženou štruktúrou trieslovín a farby. Pravdepodobne vyzreté, elegantné červené vína s plným telom

Zhluk 2: "Ľahké harmonické" (57 vzoriek)
Charakteristika: Stredný obsah alkoholu (12,48%), stredne nízke hodnoty fenolických látok, nízky obsah Prolínu (436), stredne vysoká Color_Intensity (3,95) pri nižších hodnotách trieslovín
Reprezentuje ľahšie, prístupnejšie vína s dobrou rovnováhou medzi ovocnosťou a štruktúrou. Pravdepodobne mladšie vína určené k skoršej konzumácii

Zhluk 3: "Minerálne výrazné" (26 vzoriek)
Charakteristika: Stredný obsah alkoholu (13,18%), najvyššia hodnota Magnesia (111,73), vysoký obsah Prolínu (824) a výrazná Color_Intensity (5,43) pri stredných hodnotách fenolických látok
Ide o vína s výraznejšou mineralitou vďaka vyššiemu obsahu horčíka. Pravdepodobne plnšie vína s dobrou farebnou intenzitou a výraznejším charakterom

Zhluk 4: "Robustné tanínové" (20 vzoriek)
Charakteristika: Najvyšší obsah alkoholu (13,92%), najvyššie hodnoty fenolických látok (Flavanoids 3,08, extrémne vysoký Prolín (1361), najvyššie hodnoty Color_Intensity (6,32) a Hue (1,12)
Najintenzívnejšie vína s výraznou tanínovosťou, farbou a komplexnosťou. Pravdepodobne mohutné koncentrované vína s dlhým potenciálom zrenia

Zhluk 5: "Svieže kyslejšie" (48 vzoriek)

Charakteristika: Najnižší obsah alkoholu (12,74%), najvyššia Malic_Acid (2,68), najnižšie fenolické látky, vysoká Color_Intensity (5,54) pri nízkych hodnotách fenolických látok, najvyššie Nonflavanoid_Phenols (0,41)

Ide o svieže vína s výraznejšou kyselosťou a dobrou farebnou intenzitou. Pravdepodobne ovocnejšie, bezprostrednejšie vína s menej výraznou trieslovinou

Záver

K-means zhlukovanie úspešne identifikovalo 5 rôznych typov vín, ktoré sa líšia svojimi chemickými vlastnosťami. Tieto klastre môžu reprezentovať rôzne štýly vín, ktoré by mohli byť využiteľné pri kategorizácii a marketingu vín. Vytvorená typológia poskytuje zmysluplné rozdelenie od ľahkých, svižných typov (zhluk 5), cez harmonické, stredne plné (zhluk 2), minerálne a výrazné (zhluk 3), až po komplexné prémiové (zhluk 1) a veľmi robustné tanické typy (zhluk 4).

Výsledky analýzy naznačujú, že chemické vlastnosti vín, najmä obsah alkoholu, fenolických látok, prolínu a intenzita farby, sú kľúčovými faktormi pri rozlišovaní rôznych typov vín.

Zoznam použitých zdrojov a literatúry

1. Wine Dataset. UC Irvine Machine Learning Repository. Dostupné na https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering
2. The jamovi project (2024). jamovi. (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org.
3. R Core Team (2024). R: A Language and environment for statistical computing. (Version 4.4) [Computer software].