Analýza vnitřní struktury dat o bílých vínech pomocí shlukovacích metod

Úvod

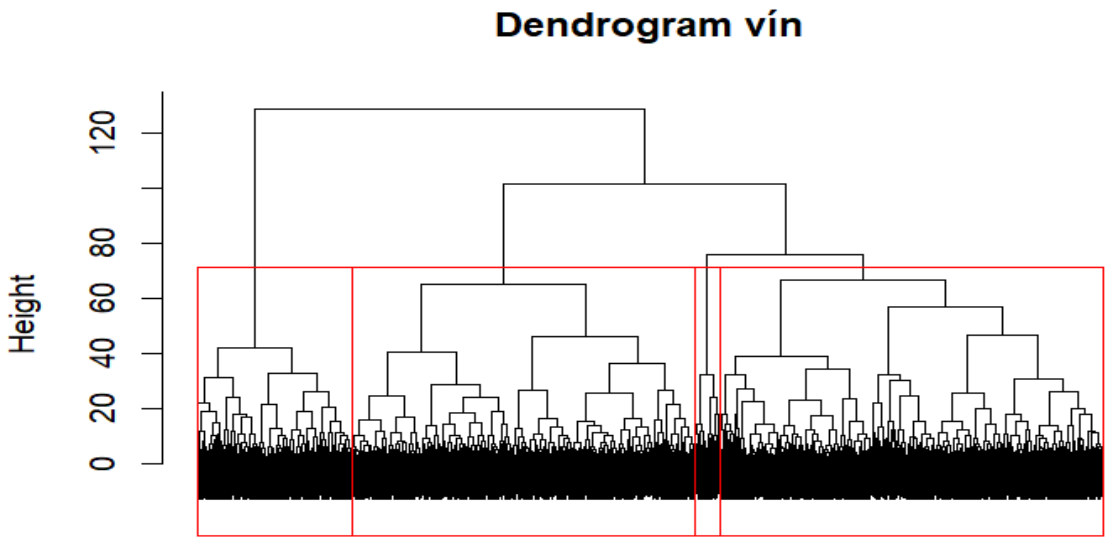
Cílem této úloha bylo odhalit vnitřní strukturu v datech vín pomocí shlukovacích metod. Vína byla popsána pomocí fyzikálně-chemických parametrů a subjektivního hodnocení kvality (1–10). Odhalení přirozených shluků může přispět k lepšímu porozumění vlastnostem vína a jejich vlivu na kvalitu.

Použité metody

Data byla nejprve upravena pro shlukování: odstranili jsme proměnné "quality" (subjektivní hodnocení) a "sweet" (binární hodnota), poté byly zbylé proměnné standardizovány pomocí funkce scale().

Následně byla aplikována metoda hierarchického shlukování s metrikou eukleidovské vzdálenosti a spojovací metodou Ward.D2, která minimalizuje rozptyl uvnitř shluků. Na základě vizuální interpretace dendrogramu a porovnání průměrné kvality vína v jednotlivých shlucích jsme vyzkoušeli rozdělení na 3, 4 a 5 shluků. Nejlepším kompromisem mezi interpretovatelností a variabilitou dat se ukázalo být rozdělení na 4 shluky, které odhalilo skupinu vín s výrazně vyšší kvalitou (shluk 3) a další skupiny s různými fyzikálně-chemickými charakteristikami.

Níže je uveden dendrogram znázorňující hierarchickou strukturu dat a rozdělení do 4 shluků:



Výsledky

V případě hierarchického shlukování byly průměrné hodnoty v jednotlivých shlucích následující:

Shluk	Alkohol	Hustota	Chloridy	SO_2 (celk.)	Kvalita
1	9.46	0.9983	0.0488	165.8	5.76
2	10.03	0.9945	0.0459	151.5	5.59
3	11.58	0.9916	0.0370	111.4	6.28
4	9.71	0.9943	0.1456	138.3	5.47

- Shluk 3 se vyznačoval nejvyšším obsahem alkoholu, nejnižší hustotou, chloridy a SO_2. Tento shluk měl také nejvyšší průměrnou kvalitu (6.28).

Hierarchické shlukování tak vedlo k zjištění, že shlukovací metoda odhalila strukturu dat, která koresponduje s kvalitou vína.

Závěr

Pomocí hierarchického shlukování (Wardova metoda) jsme v datech o vínech identifikovali 4 významné shluky. Analýza ukázala, že vína s vyšším obsahem alkoholu, nižší hustotou, menším obsahem SO2 a chloridů bývají zpravidla kvalitnější. Nejvyšší kvalitu měla vína v shluku 3.

Hierarchické shlukování poskytlo vizuální náhled na strukturu dat (dendrogram) a ukázalo se jako vhodný nástroj pro analýzu tohoto typu dat.