

Feature Engineering ja klusterointi

Tehtävän tavoitteena oli analysoida biologisten tekemää ryhmittelyä ja soveltaa ominaisuussuunnittelua olemassa olevaan dataan, jonka jälkeen arvioida, oliko biologisten alkuperäinen ryhmittely järkevä vai tulisiko biologisten ryhmitellä data uudella ominaisuussuunnittelua käyttävällä luokittelulla.

Alkuperäinen data

Alkuperäisessä datassa on 800 havaintoa seuraaville ominaisuuksille: *Height* (puiden korkeus), *Circumference* (ympärysmitta), *BarkThickness* (kaarnan paksuus) sekä *PineNo* ja *NeedleNo* (neulasten ja käpyjen määrä puussa). Alla olevasta kuvasta nähdään ominaisuuksista perusmittarit, kuten mediaani, keskihajonta jne.

	Group	Height	Circumference	BarkThickness	PineNo	NeedleNo
count	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000
mean	2.528750	10.944162	70.658888	0.533413	558.29250	56056.535000
std	1.130607	5.330914	31.515156	0.130343	265.35666	25966.654593
min	1.000000	-0.360000	-7.760000	0.300000	100.00000	10024.000000
25%	2.000000	5.085000	50.042500	0.440000	319.50000	33409.000000
50%	3.000000	12.290000	75.250000	0.510000	570.50000	56090.500000
75%	4.000000	15.730000	95.212500	0.600000	790.50000	78382.750000
max	4.000000	23.580000	140.840000	0.900000	1000.00000	99767.000000

Aineistosta huomataan heti, että pituus ja ympärysmitoissa esiintyy mahdollisia virheellisiä havaintoja (negatiivisia arvoja). Neulasten määrä vaihtelee 10 000:sta 100 000:een.

Ominaisuussuunnittelu

Loimme raakadatasta löytyvien arvojen perusteella uusia ominaisuuksia ryhmille. Ominaisuudet olivat: *Height_to_Circ* (Korkeuden suhde ympärysmittaan), *Needle_Per_Pine* (Neulasten suhde käpyihin), *Bark_per_Circ* (Kaarnan suhde ympärysmittaan) sekä *Diameter* (Puun halkaisija). Uudet ominaisuudet toivat esiin biologisesti mielekkäitä ulottuvuuksia.

PCA-analyysi

Suoritimme PCA-analyysin kahdella komponentilla standardoidulle datalle, jotta saataisiin visualisoitua moniulotteinen klusterointi kaksiulotteisesti.

PCA:n selitysasteet ovat seuraavat kahdelle komponentille:

	PCA1	PCA2	Yht.
Alkuperäinen data	~36.2 %	~21.3 %	~57.6 %
Ominaisuussuunniteltu	~34.5 %	~20.6 %	~55.1 %

Molemmissa tapauksissa yli puolet datan vaihtelusta tiivistyy kahteen ulottuvuuteen, joka on visualisointiin riittävä tässä tapauksessa.

Alla nähdään, mistä komponentit pääosin koostuvat, eli mitä komponentit kuvaavat:

Raakadata:

loadings:		PCA1	PCA2
Height	0.700614	0.015226	
Circumference	0.699801	0.104671	
BarkThickness	0.012415	0.709568	
PineNo	0.021221	-0.297334	
NeedleNo	-0.137165	0.630014	

Kuvasta nähdään, että komponentti PCA1 sai suurimmat painot korkeudesta ja ympärimitasta, eli komponentti 1 kuvaa puun fyysistä kokoa. PCA2 kuvaa enemmän puun biomassaa eli neulas- ja kaarnatiheyttä.

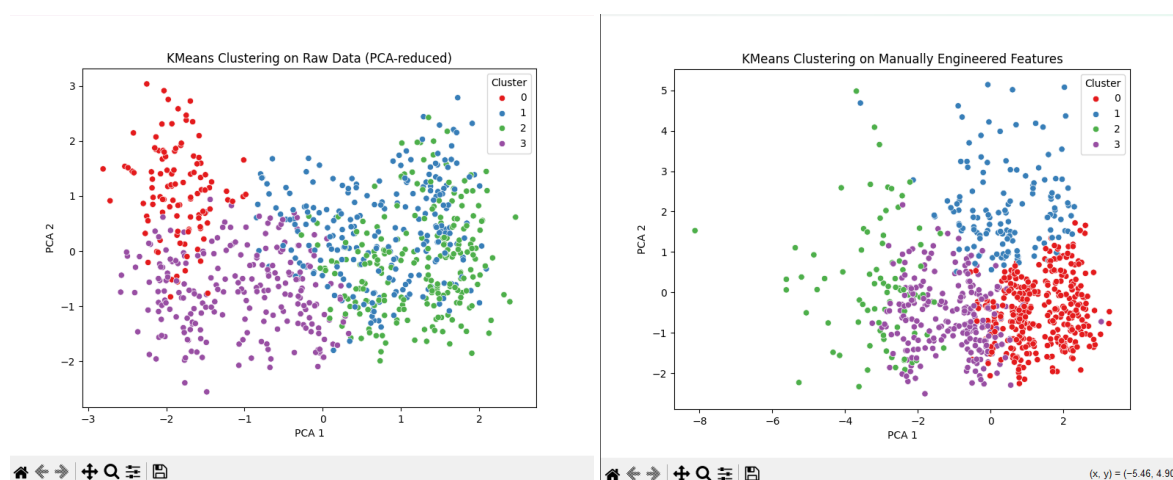
Ominaisuussuunniteltu data:

PCA Loadings:		PCA1	PCA2
Height	0.476573	0.042152	
Circumference	0.542087	0.075794	
Diameter	0.542087	0.075794	
BarkThickness	-0.023116	0.104181	
PineNo	0.039936	-0.566302	
NeedleNo	-0.078592	0.412383	
Height_to_Circ	-0.140331	-0.017523	
Needle_per_Pine	-0.063589	0.696244	
Bark_per_Circ	-0.391303	0.007703	

Kuvasta nähdään, että PCA1 sai suurimmat painot puun fyysistä kokoa kuvaavista ominaisuuksista, myös lisätystä *diameter*-ominaisuudesta. PCA2 koostuu tässä suurimmaksi osaksi *Needle_per_Pine*-, sekä *needleNo* -ominaisuuksista

Klusterien analyysi

Klusterointi suoritettiin KMeans-algoritmilla neljään puuryhmään sekä alkuperäisellä että manuaalisesti laajennetulla ominaisuusjoukolla. Molemmat aineistot skaalattiin ennen klusterointia. Alla klusterointien visualisoinnit alkuperäisestä, sekä ominaisuussuunnitellusta datasta:



Käytimme klusterien arviointiin seuraavia mittareita: *Silhouette score* ja *Davies-Bouldin indeksi*. *Silhouette score* mittaa datapisteiden samankaltaisuutta niiden omiin klustereihin verrattuna toisiin klustereihin. *Davies-bouldin-indeksi* kuvaa kunkin klusterin sisäistä hajontaa ja lähimmän naapuriklusterin välisen etäisyyden suhdetta. Pienempi indeksiarvo viittaa parempaan klusterointitulokseen. Alla oleva taulukko kuvastaa näistä mittareista saadut tulokset alkuperäisen datan ja ominaisuussuunnitteludatan klusteroinneille:

Mittari	Raakadata	Ominaisuussuunniteltu data
Silhouette score	0.202	0.214
Davis-Bouldin	1.546	1.458

Tilastollisesti manuaalisella ominaisuussuunnittelulla parannukset ovat pieniä, mutta johdonmukaisia. Uudet ominaisuudet tuottivat tilastollisesti tiiviimmät ja erottuvammat klusterit. Uudet ominaisuudet ovat siis lisänneet olennaista informaatiota, jolloin Kmeans-algoritmi kykeni erottamaan havaintoryhmät hieman selkeämmin.

Johtopäätökset

Klusterointi ominaisuussuunnittelulla pystyi löytämään yhtä mielekkäitä ja vähän selkeämpiä ryhmiä tilastollisesti arvioituna. Silhouette score:lla sekä DB-indeksillä. On siis perusteltua harkita luotua ominaisuussuunnittelua biologien luokittelun tukena tai vaihtoehtona alkuperäiselle luokittelulle.

Automaattinen ominaisuussuunnittelu:

Otimme kokeilumielessä lisätehtäväksi vielä datan klusteroinnin käyttäen jotain automaattista ominaisuussuunnittelua. Valitsimme tehtävänannon feature engineering -linkistä löytämämme featuretools -kirjaston ja kokeilimme luoda automaattisesti uudet ominaisuudet datalle ennen sen klusterointia. Jäi vähän epäselväksi, oliko data vähän huono featuretools:lle vai oliko meissä vika (mitä luultavimmin jälkimmäinen), mutta klusteroinnista muodostui aikailla sama, mitä raakadatan klusteroinnista saatiin. Pitänee vielä vähän perehtyä tuohon automaattiseen ominaisuussuunnitteluun.

