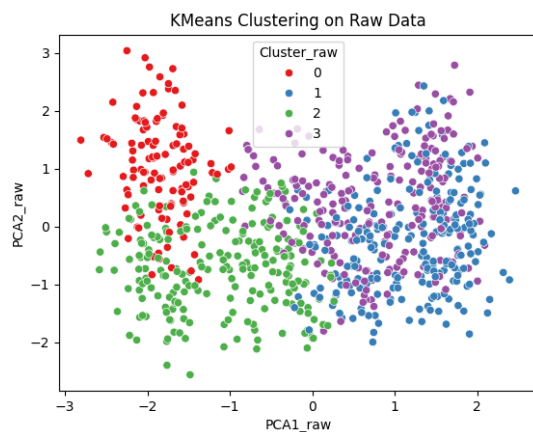


Feature Engineering ja klusterointi

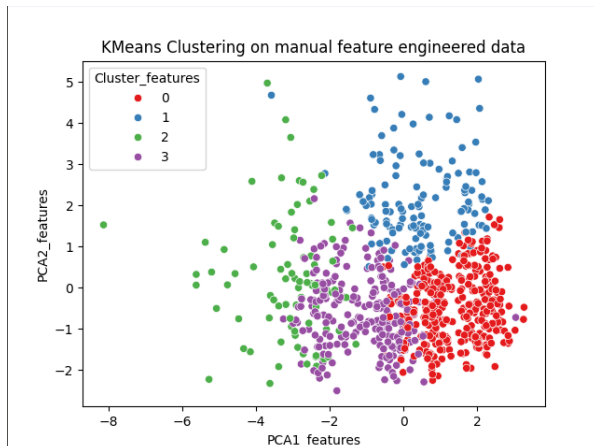
Aloitimme klusteroimalla biologien alkuperäisen datan kmeans-algoritmilla. Poistimme 'Group'-sarakkeen ja lähdimme klusteroimaan. Skaalasimme datan scikit-learning:n StandardScaler()-funktion avulla, jonka jälkeen teimme dimension pienennyksen PCA:lla, jotta saadaan selkeä 2D-kuvaus klusteroinnista scatterplotilla. Dimension pienennys PCA:lla suoritettiin jokaiselle datalle tässä tehtävässä.



Kuvaajasta voidaan nähdä, että klusterit ovat hajanaisia ja päällekkäisiä. Klusterien ryhmittymät eivät ole tiiviitä, mikä viittaisi siihen, että biologien keräämät alkuperäiset ominaisuudet eivät tarjoa hirveän hyvää erottelukykä ryhmien välillä.

Manuaalinen ominaisuussuunnittelu:

Manuaalisessa ominaisuussuunnittelussa loimme biologien antamien muuttujien avulla uusia ominaisuuksia puille. Datalle tehtiin muuten samat toimenpiteet kuin raakadatallekin. Luodut ominaisuudet olivat: 'Height_to_Circ', 'Needle_per_Pine', 'Bark_per_Circ' ja 'Diameter' eli korkeus/ympärysmitta, neulasten suhde käpyihin, kaarnan suhde ympärysmittaan ja halkaisija. Näillä toivoimme, että saataisiin tarkempi ryhmittely puille:



Kuvaajasta voidaan nähdä, että manuaalinen ominaisuussuunnittelu paransi klusterointia alkuperäiseen dataan verrattuna. Klusterit ovat tiivimpiä, pyörempiä ja selkeämmin erotettavissa toisistaan. Uudet ominaisuudet ovat siis lisänneet olennaista informaatiota, jolloin Kmeans kykenee erottamaan havaintoryhmät selkeämmin. Joitain hajapisteitä esiintyy, mutta yleisesti rakenne on jäsentynyt paremmin kuin raakadatan klusteroinnissa.

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.

