Dataset: Airline

Clustering

Markus Kinn

I dette dokumentet kommer jeg til å diskutere valg jeg har tatt og sammenligne resultater fra ulike algoritmer. Jeg har ett dokument for hvert datasett.

For å forbedre resultatene på modellene har jeg prøvd å implementere så mye so mulig fra *Machine Learning Performance Improvement Cheat Sheet.*

**Generelt om datasettet:**

Datasettet er opprinnelig et klassifiserings datasett hvor man har som mål å predikere om en person var tilfreds med fly opplevelsen eller ikke.

Datasettet har lite feil, noe som gjør at man ikke trenger å gjøre mye data pre-Processing. Det har derimot noen manglende verdier som jeg fikser ved bruk av gjennomsnittet for den kolonnen.

Jeg endte opp med å fjerne 4 kolonner ved siden av ground truthen. Dette var: ‘Gate location’ , ‘Departure/ arrival time convenient’, ‘Flight Distance’ og ‘Departure Delay in minutes’. Grunnen til at jeg fjernet disse 4 var at jeg opprinnelig fikk dårlig resultat med alle algoritmene jeg hadde med disse variablene. Jeg valgte derfor å manuelt teste hvilke kolonner som skulle være med og ikke.

**Kmeans:**

**Før tuning:**

**Text

Description automatically generated**

**Etter tuning:**

**Text

Description automatically generated**

**Diskusjon av resultater:**

Baseline modellen presterte ganske dårlig, men dette er av den grunn at kmeans sin standard n\_clusters verdi er 3, og siden datasettet har kun to klasser prøver den å dele to klasser på tre.

På tuning modellen, øker resultatet betraktelig på den interne ‘silhouette’ scoren, men også en liten økning på den eksterne scoren. Modellen endte opp med en ekstern score på 55.1% som betyr at den er hakke bedre enn ren gjetting.

**BIRCH:**

**Før tuning:**

**Text

Description automatically generated**

**Etter tuning:**

**Text

Description automatically generated**

**Diskusjon av resultater:**

Baseline modellen hadde en relativt god silhuoette score, men den eksterne accuracy scoren var så vidt bedre enn ren gjetting. Jeg har prøvd å finne ut av hvordan man kan øke scoren med denne algoritmen, men fant ikke noe klart svar. Tror derfor at dette kan være et problem med datasettet.

Prøvde også å tune modellen så vidt, men dette endte kun med en smule dårligere resultat.

**Agglomerative:**

Text

Description automatically generated

**Diskusjon av resultater:**

For denne algoritmen fant jeg ingen hyperparametere som var verdt å tune. Modellen fikk allikevel en høy silhouette score. Denne algoritmen har ingen predict funksjon og jeg fikk dermed ikke mulighet til å teste ut en ekstern score.

**Sammenligning av alle resultater:**

For dette datasettet fikk modellen basert på agglomerative algoritmen høyest silhouette score, men like bak kom Birch algoritmen med tanke på silhouette score. Birch har en predict funksjon som gjør det mulig å benytte ekstern scoring. Birch fikk den høyeste accuracy scoren med 55.6%. Ikke et fantastisk bra resultat, men hakke bedre enn ren gjetting. Jeg velger derfor å utpeke Birch algoritmen til en vinner.

Jeg tror at dette kan være et generelt problem med datasettet da både Kmeans og Birch har meget dårlig ekstern score.