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.

**Kmeans:**

**Før tuning:**

**Etter tuning:**

**Diskusjon av resultater:**

**BIRCH:**

**Før tuning:**

**Etter tuning:**

**Diskusjon av resultater:**

**Agglomerative:**

**Før tuning:**

**Etter tuning:**

**Diskusjon av resultater:**

**Sammenligning av alle resultater:**

Alle modellene gjorde det ganske bra, men de to som skiller seg ut er XGBoost etter tuning og det første neurale nettverket. Disse to modellene presterte rundt regnet HELT like. Dette får meg til å tro at dette er den maksimale presisjonen man kan få med dette datasettet, med denne dataprosesseringen.

Etter å ha sett på andres modeller, så ser det ut som det er rundt den scoren jeg fikk, som andre også som regel får.