ITMAL Øvelser – Dataanalyse

Øvelse 1 :

I skal analysere på ”California housing prices” (https://www.kaggle.com/camnugent/california-housing-prices), som også benyttes i lærebogen.

a) Plot fordelingen af median_income. Find også spredning, middelværdi og median.
b) Er der forskel på median og middelværdi af median_income ? Hvilken af de to beskriver bedst en ”almindelig families indkomst” og hvorfor ?
c) Fit en normalfordeling til data og plot histogrammet – passer de to ?
d) Er der sammenhæng imellem median_house_value og median_income ? Lav korrelationsplot.
e) Hvad er 5% og 95% percentilerne af median_house_value ? (dvs. grænserne for 5% laveste og højeste). Plot også fordelingen af median_house_value. Kommentér på realismen af max-værdi og 95% percentil – foreslå gerne en løsning til hvad man kan gøre ved dette, hvis man skal have mere realistiske data.

Tips :

I kan finde mange eksempler på data analyse under ”Kernels” på Kaggle site, fx. - https://www.kaggle.com/rajritu2803/california-housing-price-prediction

https://www.kaggle.com/takedown/complete-tutorial-for-beginners

Seaborn tutorial – plot fordelinger og korrelationer - https://seaborn.pydata.org/tutorial/distributions.html

Øvelse 2 (OBLIGATORISK) – Beskrivelse af slutprojekt datasæt :

I kurset er slutprojektet et bærende element, som I forventes at arbejde på igennem hele kurset sideløbende med de forskellige undervisningsemner. I skal selv vælge slutprojekt – det anbefales at I vælger en problemstilling, hvor der allerede er data til rådighed og en god beskrivelse af data, dataopsamlingsmetode og problemstilling.

I denne opgave skal I :

a) Beskrive jeres valgte datasæt med en kort forklaring af baggrund og hvor I har fået data fra.

b) Beskrive data – dvs. hvilke features, antal samples, target værdier, evt. fejl/usikkerheder, etc.

c) Forklare hvordan I ønsker at anvende datasættet – vil I fx. bruge det til at prædiktere noget bestemt, lave en regression eller klassifikation, el.lign. I vil nok komme til at anvende data også på andre måder i løbet af undervisningen – men det behøver I ikke nævne. Og det er også ok, hvis I ender med at bruge data på en anden måde end planlagt her.

Omfang af beskrivelsen forventes at være 1-2 sider.

Øvelse 3 (OBLIGATORISK) – Dataanalyse af eget datasæt :

Lav data analyse på jeres egne data og projekt.

Det indebærer de sædvanlige elementer såsom plotte histogrammer, middelværdi/median/spredning, analysere for outliers/korrupte data, forslag til skalering af data og lignende former for analyse af data.

For nogle typer data (fx billed-data), hvor features ikke har en specifik betydning, er det mest histogrammer og lignende, som giver mening – det er helt o.k.