Obligatorisk innlevering 1

Utførelsen av denne innleveringen skal helst skje i små grupper på 2-3 personer, men dere får også mulighet til å jobbe alene dersom det av ulike grunner ikke er gjennomførbart for deg å jobbe i gruppe. Det er mulig å sette opp møte med Leo eller Bjørn-Jostein for å oppklare eventuelle utydeligheter vedrørende oppgavene. Den obligatoriske innleveringen skal bestå av en kode-del og en rapport - begge leveres og det forventes at vi skal være i stand til å kjøre koden og replisere de resultatene dere presenterer i rapporten. Med mindre dere har gode grunner til det bør rapporten være lagt opp etter IMRaD-oppsett: Introduction – Method – Results – and – Discussion eller Introduksjon – Metode – Resultater – og – Diskusjon på norsk (ta gjerne også med en kort konklusjon-del etter diskusjon)

Rapporten kan være på rundt fire-fem sider og dere må sitere kilder etc. i henhold til god akademisk praksis. Dere skal velge mellom **én** av følgende to oppgaver:

1. Oslofjorden

Denne oppgaven er inspirert av årets hackathon i forbindelse med Horten Tech Festival: «Løsninger for havet» (https://www.electroniccoast.no/2024/06/10/hackathon-losninger-for-havet/). Vi bruker et datasett fra Havforskningsinstituttet der observasjonene er gjort av «vanlige folk» (folkeforskning) https://dugnadforhavet.no/dataportal. Brukerne har rapportert sine funn ved å fylle ut nettbasert skjema, og merke av på et digitalt kart hvor funnet var. De har også kunnet laste opp bilder og videoer, samt stille spørsmål eller gi kommentarer til havforskerne. Videre har havforskerne verifisert funnene basert på bilder og beskrivelser.'

I første del av oppgave ønsker vi at dere bygger et interaktivt dashboard (enten som en app, som en del av en Jupyter Notebook eller en annen måte) som gjør det mulig å plotte alle observasjoner i datasettet havforsk.csv ved hjelp av koordinatene i datasettet. Plotly Dash (https://plotly.com/python/) eller folium kan være et nyttig sted å starte for å implementere løsningen. Datasettet inneholder observasjoner fra hele norskekysten, så dere kan bruke kartet for å fjerne observasjoner gjort utenfor Oslofjorden. Videre ønsker vi at det skal være mulig å se hva slags art som er observert på det aktuelle kooridinatet. Hvert koordinat i den interaktive visualiseringen skal vise hvilken art som er observert på det akutelle punktet ved å klikkes på eller føre musepekeren over punktet. Videre er det opp til dere om dere ønsker å legge til flere komponenter hvor man interaktivt kan utforske datasettet.

I andre del av oppgaven ønsker vi at dere lager en tabell eller figur hvor dere teller opp antallet observasjoner for alle unike arter der en fra havforskningsinstituttet har validert observasjonen (@hi.no). Lag også en oversikt som viser antall observasjoner av hver art fordelt på år. Ser dere en trend i dataene? Er det statistisk signifikante endringer fra år til år? Diskuter hvorvidt det er mulig å si noe om reduksjon/økning i antall av de ulike artene. Er det tilstrekkelig for å si noe om situasjonen for hele Oslofjorden? Diskuter..

2. Hjertesykdomsanalyse

I denne oppgaven skal dere grave i datasettet; «UCI Heart Disease Data» og utføre noen bestemte analyser. Dette datasettet inneholder totalt 16 variabler (inklusiv pasient ID) og 920 observasjoner.

Variabelnavn	Type	Forklaring
ID	Integer	Pasient ID
age	Integer (kontinuerlig)	Alder
sex	Kategorisk	Kjønn
cp	Kategorisk	Type brystsmerter
trestbps	Integer (kontinuerlig)	Blodsukkernivå i hvile
chol	Integer (kontinuerlig)	Kolesterol i serum
fbs	Boolsk	Blodsukker ved faste
restecg	Kategorisk	Observert EKG-fenomen i hvile
thalch	Integer (kontinuerlig)	Maksimal hjertefrekvenes
exang	Boolsk	Anstrengelsesutløst angina
oldpeak	Float	ST-depresjon (forskjell mellom arbeid og hvile)
slope	Kategorisk	Karakteristikk av ST-segment
Ca	Kategorisk	Antallet hovedkransårer farget med flouroscopy
Thal	Kategorisk	Type defekt
num	Kategorisk (ordinal)	Grad av sykdom (0=frisk)

Konverter kategoriske data, som her er oppgitt som tekststrenger, til numeriske data (diskrete eller binære). Gjør også boolske variabler til binære. Videre skal dere utføre en beskrivende analyse av datasettet hvor dere lager individuelle histogrammer for hver av de kontinuerlige og diskrete variablene. Beregn også gjennomsnitt, median og standardavvik for de samme variablene og rapporter disse i oppgaven. Avgjør om variablene er normalfordelte eller skjeve.

Lag et korrelasjonsplott der dere inkluderer alle variablene bortsett fra pasient ID.

Utfør en statistisk test der du undersøker om det er noen signifikant forskjell i kolesterolnivået mellom kvinner og menn. Nullhypotesen er at det ikke er noen forskjell, mens den alternative hypotesen er at kolesterolverdien er høyere hos menn (ensidig signifikanstest). Kommenter funnene.

Test også om anstrengelsesutløst angina (exang) er assosiert med hjertesykdomsdiagnose (num) ved hjelp av en signifikanstest. Kommenter funnene. Nullhypotesen er at det ikke er noen sammenheng, mens alternativhypotesen er at det er en form for sammenheng (tosidig signigfikanstest)

Utfør en regresjonsanalyse der du ser på sammenhengen mellom maksimal hjertefrekvenes (thalch) og ST-depresjon (oldpeak). Lag et plott der du viser punktene (x=,oldpeaks y= thalch), samt en linje som representerer regresjonsmodellen. Rapporter også R2-scoren.

Forsøk å relatere funnene til allerede publisert litteratur. Viser deres funn det samme eller er det motstridende funn? Diskuter.