Oblig 7: Datahåndtering

Læringsmål 7.1: Statistikk

I disse oppgavene skal du lære og vise at du behersker følgende:

1. Lage egne funksjoner til statistisk beregning.
2. Bruke biblioteker til statistisk beregning.
3. Beregne gjennomsnitt, standardavvik og relativ og absolutt feil.
4. Utføre regresjonsanalyse på eksperimentelle data.
5. Benytte maskinlæringsalgoritmer til å utforske datasett og danne hypoteser om sammenhenger.

Oppgave 7.1

Vi bruker et datasett vi har fått fra en analyse av innholdet av Pb^{2+} i bekkevann som utgangspunkt. Til dette er det brukt spektroskopisk analyse. I spektroskopiske analyser finner vi konsentrasjonen til et stoff ved å undersøke hvor mye lys stoffet absorberer av en bestemt bølgelengde. For å finne konsentrasjonen i en ukjent løsning, lager vi først en *standardkurve* ut fra absorpsjonen til løsninger med kjent konsentrasjon. Vi analyserer en rekke løsninger og får følgende resultater:

Konsentrasjon (ppm)	**Absorbans**
0.0	0.0
0.100	0.116
0.200	0.216
0.300	0.310
0.400	0.425
0.500	0.520

a) Lag et program som gjør lineær regresjon på dataene og plotter datapunktene og den tilpassede regresjonskurven i samme koordinatsystem.

b) Analysen ved 283 nm av vannprøva ga absorbans på 0.340. Bruk standardkurven og programmet til å bestemme konsentrasjonen av blyioner i vannprøva i ppm.

Oppgave 7.2

a) Lag et program som inneholder en funksjon *gjennomsnitt* og en funksjon *standardavvik* som regner ut gjennomsnittet og standardavviket gitt en liste med datapunkter som parameter. Test funksjonene på `liste = [1,2,2,1,3,3]` og sammenlikn med numpy-funksjonene `mean` og `std`.

b) Benytt funksjonene du lagde i a) til å regne ut gjennomsnittet og standardavviket av følgende målinger gjort av koffein i te med væskekromatografi:

Injeksjon	**Konsentrasjon (mg/mL)**
1	245
2	272
3	252
4	264
5	261
6	272
7	255
8	260
9	268
10	259

Oppgave 7.3

Bruk pandas-biblioteket til denne oppgaven. Fila `vin.csv` beskriver ulike kjemiske parametre i 1500 rødviner, i tillegg til en vurdering av vinkvaliteten, på en skala fra 1–8.

1. Les fila og beskriv de ulike kategoriene.
2. Lag et korrelasjonsplott med utgangspunkt i alle kategoriene i datasettet. Hvilke faktorer ser ut til å korrelere med god vinkvalitet? Gi også eksempler på faktorer som ikke korrelerer og faktorer som har negativ korrelasjon. Forklar hva dette betyr. Prøv gjerne å forklare noen av korrelasjonene.
3. Lag et søylediagram med vinkvalitet på førsteaksen og det totale innhold med svoveldioksid på andreaksen. Hva kan årsaken være til denne fordelingen? Sammenlikn med korrelasjonen mellom disse faktorene.
4. Lag en ny kolonne i datasettet som inneholder «kvalitetskategorien» til vinen. Den skal inneholde 0 hvis vinen har under 6 i kvalitet, og 1 hvis den har 6 eller mer.
5. Lag en modell som skal forutsi kvalitetskategorien til vinen. Bruk en bestemmelsestre-algoritme som grunnlag for modellen.
6. Test og valider modellen din. Kommenter resultatet. Hvorfor brukte vi en egen kvalitetskategori og ikke vinkvaliteten på en skala fra 1–8 direkte?