|  |
| --- |
| Universitetet i Bergen |
| Predikasjon av sykehusoppholdslende |
| Innføring i data science INF161 |

|  |
| --- |
| Andreas Wold Sløgedal  Høst 2024 |

Contents

[1 Introduksjon 2](#_Toc180151798)

[1.1 Beskrivelse av datasett 2](#_Toc180151799)

[2 Innledende dataforberedelse 3](#_Toc180151800)

[2.1 Klargjøring av data 3](#_Toc180151801)

[2.1.1 Demografiske data 4](#_Toc180151802)

[2.1.2 Fysiologiske data 4](#_Toc180151803)

[2.1.3 Sykehusalvorlighetsdata 4](#_Toc180151804)

[2.2 Kombinering av datasett 5](#_Toc180151805)

[2.3 Databehandling 5](#_Toc180151806)

[2.4 Inndeling av trenings-, validerings- og testdata 6](#_Toc180151807)

[3 Dummy-kodifisering 7](#_Toc180151808)

[4 Statistisk analyse av treningsdata 8](#_Toc180151809)

[4.1 Statistiske mål: Gjennomsnitt, standardavvik, kvantiler og mer 8](#_Toc180151810)

[4.2 Korrelasjonsmatrise 9](#_Toc180151811)

[5 Visualisering av data 10](#_Toc180151812)

[5.1 Demografiske variabler 10](#_Toc180151813)

[5.1.1 Alder 10](#_Toc180151814)

[5.1.2 Kjønn 11](#_Toc180151815)

[5.1.3 Inntekt ## denne bør fjernes 12](#_Toc180151816)

[5.2 Helserelaterte variabler 12](#_Toc180151817)

[5.2.1 Sykdomskategorier blant etnisiteter 13](#_Toc180151818)

[5.2.2 Korrelasjon mellom fysiologiske variabler 14](#_Toc180151819)

[5.2.3 Korrelasjon mellom variabler i sykdomsalvorlighet 15](#_Toc180151820)

[5.2.4 Fysiologiske data og uteliggere 16](#_Toc180151821)

[5.2.5 Sammenligning av overlevelsesestimat: 2 vs. 6 måneder 17](#_Toc180151822)

[6 Datarydding 18](#_Toc180151823)

# 1 Introduksjon

**Formålet** med denne oppgaven er å anvende alt materialet lært i INF161 til å bygge en maskinlæringsmodell som skal predikere den forventede lengden på sykehusoppholdet for hver pasient. Denne modellen vil benytte pasientopplysninger fra fire ulike datasett, som sammen gir en helhetligforståelse av pasientens helse og personlige bakgrunn.

## 1.1 Beskrivelse av datasett

Datasettet om **demografisk informasjon** inkluderer variabler som kjønn, alder, utdanning, inntekt og etnisitet. Disse dataene er ikke nødvendigvis direkte tilknyttet pasientens helsetilstand, men det er aktuelt å undersøke om de har innvirkning på sykdomskategorier og behandlingsresultater. Derfor vil det være interessant å undersøke om det er sammenhenger mellom disse demografiske variablene og helseutfallene i dataanalysen.

**Sykehusdatasettet** inneholder informasjon om pasienter som har fått behandling på sykehuset. Variablene inkluderer oppholdslengde, eventuelle dødsfall og innleggelse. Her fungerer oppholdslengden som den avhengige variabelen, og er dermed målvariabelen vi ønsker å predikere. Målvariabelen er sentral i treningsdataene, ettersom maskinlæringsmodellen har som hensikt å undersøke hvor lenge pasienten vil oppholde seg på sykehuset.

Både **sykdomsalvorlighet** og fysiologisk data inneholder uavhengige variabler som alle er nødvendige for å undersøke pasientens tilstand. **Fysiologisk data** inkluderer numeriske mål som hjertefrekvens, respirasjonsfrekvens, kroppstemperatur og nivåer av kreatinin. Data over sykdomsalvorlighet, på den andre siden, tar for seg mer kategorisk informasjon, som blant annet pasientens diagnose, sykdomsunderkategori, og overlevelsesestimater. Dette er alle sentrale mål for å forstå pasientens helse og prognose.

I praktisk sammenheng ønsker vi å samle inn data og utvikle en modell for å diagnostisere pasienter på en mest mulig effektiv måte. Målet er å optimalisere behandlingsprosessen, noe som maksimerer sannsynligheten for overlevelse.

# 2 Innledende dataforberedelse

## 2.1 Klargjøring av data

All datamanipulasjonen i den innledende dataforberedelsen er utført med hensyn til at tekstdataene ikke skal benyttes før modellen er ferdig trent.

For å kunne gjennomføre dataanalyse på en effektiv og oversiktlig måte importerer jeg Python-biblioteket Pandas, som er et kraftig verktøy ved dataanalyse-, og manipulering.



A black background with white text

Description automatically generatedPandas til å laste inn datasettene demographic.csv, hospital.csv, physiological.txt og severity.json. Disse lagres henholdsvis i variablene demographic\_df, hospital\_df, phychological\_df og severity\_df.

### 2.1.1 Demografiske data

For at vi skal kunne slå sammen datasettene er det nødvendig at datasettene har likt antall rader. Hvis ikke kan videre analyse føre til uventede resultater, som duplikater eller manglende data. Vi kan sjekke antall kolonner og rader i datasettene ved å bruke Pandas attributten **pd.shape.** Demografiske data har 7742 rader, men de resterende datasettene har 7740 rader. Dette tyder på at det er duplikater i datasettet som må fjernes før sammenslåing. Vi bruker **pd.drop\_duplicates()** på demografiske data**.**

****

### 2.1.2 Fysiologiske data

A screen shot of a computer code

Description automatically generatedI det fysiologiske datasettet er det en betydelig forekomst av NaN-verdier, noe som kan skape utfordringer i senere faser, spesielt når vi skal utvikle maskinlæringsmodellen og gjøre prediksjoner. For å sikre at verdifull data ikke går tapt ved å fjerne datapunkter, har jeg besluttet å erstatte NaN-verdiene med anbefalte eksempelverdier. De verdiene er hentet fra en pålitelig kilde: <https://archive.ics.uci.edu/dataset/880/support2>

### 2.1.3 Sykehusalvorlighetsdata

A black background with white text

Description automatically generatedDette datasettet kommer i et annet format enn de andre filene, der alle kolonnene utenom de to første er lister. Uten tilstrekkelig dataforbindelse ville det ikke vært mulig å slå sammen datasettene. Derfor brukes **pd.explode()** slik at hver verdi i listen blir plassert i sin egen rad, mens de andre kolonneverdiene i samme rad forblir uendret.

Dette betyr at for hver listeverdi i en kolonne vil det opprettes en ny rad, noe som effektivt «utvider» datasettet. Vi ser ved sammenslåingen av datasettene at dette steget er avgjørende for videre dataanalyse.

## 2.2 Kombinering av datasett

A screen shot of a computer code

Description automatically generatedNå som hvert kolonne-rad-par i sykehusalvorlighetsdata inneholder en enkelt verdi, og alle datasettene har likt antall rader, mangler jeg kun et steg for å kunne kombinere datasettene. Jeg ønsker å fjerne den eksisterende indeksen i hver dataframe og lage en ny standard index. Siden hvert datasett inneholder kolonnen «pasient\_id» ønsker jeg å droppe denne kolonnen i de tre siste datasettene.

Deretter slår man sammen de fire datasettene ved å bruke funksjonen **pd.concat().** Axis=1 betyr at sammenslåingen skjer horisontalt, altså legges kolonnene fra de ulike datasettene ved siden av hverandre.

Det kombinerte datasettet kaller jeg df, og består nå av 39 kolonner og 7740 rader.

## 2.3 Databehandling

Etter datasettet er slått sammen, ønsker vi å undersøke om dataene er klare for å brukes videre i maskinlæringsmodellen. Vi inspiserer dataene ved å bruke **pd.info().** En rekke kolonner er av typen «object», som typisk representerer tekstdata. I maskinlæringsmodellen ønsker vi flest mulig kolonner av numeriske verdier, da dette forenkler trening av modellene. For å konvertere disse kolonnene til numeriske verdier, bruker vi **pd.to\_numeric()**.

A computer screen shot of text

Description automatically generated

Jeg merker meg også verdier som ikke er naturlige. Dette inkluderer negative verdier av oppholdslengde og alder. De negative verdiene må håndteres for å optimalisere prediksjonsevnen til modellen. Jeg velger også å fjerne kolonnene *dødfall* og *sykehusdød,* da de ikke har noe reell innvirkning på modellens prediksjon av oppholdslengde på dag én. *Pasient\_id* fjernes, og datasettet indekseres på nytt. *Adl\_pasient* og *bilirubin* er henholdsvis indeks for daglige funksjonsaktiviteter og *bilirubinnivåer* og ble utfylt av pasienten ved dag syv. Disse verdiene har heller ikke innvirkning på prediksjoner ved dag én, og fjernes av den grunn. Generelt er det essensielt å sikre at kun data som er tilgjengelig ved pasientens ankomst blir brukt i modellene, for å unngå forvirring med informasjon fra fremtidige tidspunkter.

A black background with white text

Description automatically generated

## 2.4 Inndeling av trenings-, validerings- og testdata

For å bygge en maskinlæringsmodell som presterer godt på ukjente data, er det nødvendig å dele datasettet (df) inn i trenings-, validerings- og testsett. Først opprettes X som består av alle kolonnene utenom den avhengige variabelen «oppholdslengde». Den avhengige målvariabelen «oppholdslengde» skal modellen predikere, og lagres i y. Treningssettet brukes til å trene modellen, og består av 70% av datasettet. Den resterende 30% er testdata. Dermed splittes testsettet (30% av det originale testsettet) i to like store deler. 50% av disse brukes som valideringsdata, og de resterende 50% brukes som testsett. Avslutningsvis sorterer vi treningsdataene etter «pasient\_id».

A screen shot of a computer code

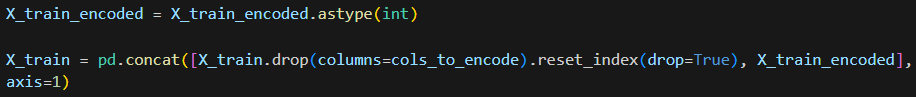
Description automatically generated

# 3 Dummy-kodifisering

I treningsdataen har vi en rekke kategoriske data, som representerer verdier som ikke kan måles på en numerisk skala. I maskinlæringsmodeller ønsker vi å bruke numeriske verdier fordi de bidrar til å forbedre modellens ytelse og nøyaktighet. Kategoriske data kan derimot være vanskeligere for modellen å tolke fordi de ikke oppgir direkte numeriske verdier. For å gjøre kategoriske data brukbare i maskinlæringsmodellen bruker vi **OneHotEncoder** fra biblioteket **sklearn.preprocessing**. Denne funksjonen transformerer kategoriske kolonner til binære kolonne, som gjør det lettere for modellen å tolke dataene. De kategoriske kolonnene som skal transformeres er lagret i *cols\_to\_encode*. De transformerte kolonnene lagres i *X\_train\_transformed*, og det lages en ny dataframe av dem. Det samme gjøres for både testdata og valideringsdata.A screen shot of a computer code

Description automatically generated

Avslutningsvis runder jeg av verdiene til heltall, og sår sammen det originale datasettet og det nye datasettet.



# 4 Statistisk analyse av treningsdata

Statistisk analyse er en viktig metode for å forstå dataene bedre og trekke ut nyttig informasjon. Den gjør det mulig å identifisere mønstre, oppdage problemer tidlig, og forutsi fremtidige hendelser basert på tidligere data. I tillegg kan statistisk analyse avdekke skjulte sammenhenger mellom ulike variabler. Dette er avgjørende for å sikre at dataene er klare og pålitelige før de brukes i videre analyser, som ved bruk av maskinlæringsmodeller.

## 4.1 Statistiske mål: Gjennomsnitt, standardavvik, kvantiler og mer

Vi undersøker først statistiske mål ved å bruke metoden **pd.describe().** Følgende er en liten utskrift av alle statistiske mål.

A screenshot of a computer screen

Description automatically generated

Dette gir oss en oversikt over diverse statistiske verdier, som gjennomsnitt, minimum, maksimum og spredning for variablene i datasettet. Vi får diverse informasjon om pasientenes helsetilstand. For eksempel er gjennomsnittsalderen til pasienter på sykehuset er 62.7 år, og gjennomsnittlig blodtrykk er 84 slag i minuttet. Slike verdier er nyttige for en maskinlæringsmodell da de gir innblikk i underliggende mønstre i dataene.

Samtidig inneholder datasettet noen bemerkelsesverdige verdier. Vi ser blant annet at det er minst en pasient med null hvite blodlegemer, noe som virker unaturlig lavt ut. Jeg undersøker videre om noen av disse pasientene overlevde. Av dem som hadde null hvite blodlegemer, overlevde fire pasienter, mens fire døde.

Dette er tall som vekker interesse og er potensielt kritiske funn som skal undersøkes videre under dataryddingen for å forstå om det er feil i dataene, eller om det er en medisinsk årsak bak funnene.

## 4.2 Korrelasjonsmatrise

A screenshot of a computer screen

Description automatically generatedEn korrelasjonsmatrise er et nyttig verktøy for å undersøke sammenhenger mellom variablene i datasettet. Den kan hjelpe med å identifisere potensielle årsaksforhold og avdekke skjulte mønstre som ellers er vanskelige å oppdage. Følgende er en liten utskrift av korrelasjonsmatrisen.

Korrelasjonsmatrisen for treningsdataene består av 63 x 63 = 3969 verdier. De fleste av disse verdiene viser lav korrelasjon. Likevel legger jeg merke til enkelte variabler som korrelerer sterkere med hverandre. Et eksempel er variablene *sykdom underkategori lung cancer* og *kreft metastatic*, som har en korrelasjonsverdi på 0.665739 Dette kan være kritiske funn, da det er viktig å eliminere variabler som er sterkt korrelerte. For høye verdier kan føre til overestimater i maskinlæringsmodellen, noe som kan svekke modellens prediksjonsevne.

# 5 Visualisering av data

Visualisering av dataen er en effektiv og oversiktlig måte å støtte og validere funnen fra den statistiske analysen og korrelasjonsmatrisen. Ved å kombinere visualisering og numerisk analyse er det enklere å forstå trender og mønstre i dataene. Visualisering er også en effektiv måte å identifisere avvik og uventende resultater.

## 5.1 Demografiske variabler

Variabler som kjønn, alder, utdanning og inntekt er viktige demografiske faktorer som kan ha innvirkning på helseutfall. Disse variablene viser hvordan ulike samfunnsgrupper påvirkes ulikt av helseproblemer, og kan være med på å identifisere sårbare grupper som trenger spesifikke helsetjenester.

### 5.1.1 Alder

Alder har ofte sammenheng med sykdomsalvorlighet og oppholdslengde da kroppen og immunforsvaret blir svakere med alder. Å få oversikt over aldersfordelingen blant pasientene kan være aktuelt fordi forskjellige aldersgrupper kan ha varierende behandlingsbehov.

**A graph of a graph

Description automatically generated with medium confidenceHistogrammet illustrerer aldersfordelingen basert på kjønn**

Fra histogrammet ser det ut som det er en stor andel pasienter i alderen 60-80 år, både menn og kvinner. Vi merker oss at det er minst en pasient registrert med negativ alder. Denne verdien er bemerkelsesverdig og må tas hensyn til i databehandling for at modellene skal yte best mulig. Videre kan oversikt over andel døde per aldersgruppe være aktuelt for å allokere ressurser bedre og effektivisere behandling.

A pie chart with numbers and percentages

Description automatically generated**Sektordiagrammet illustrerer dødelighet fordelt på ulike aldersgrupper**

Som forventet er dødeligheten høyest blant de eldste pasientene. Det er imidlertid verdt å merke seg at forskjellene ikke er dramatiske store. Det kan tyde på at selv om alder er en faktor til dødelighet, finnes det andre variabler som også spiller en betydelig rolle. Vi undersøker videre for å forstå hvilke andre variabler dette kan være.

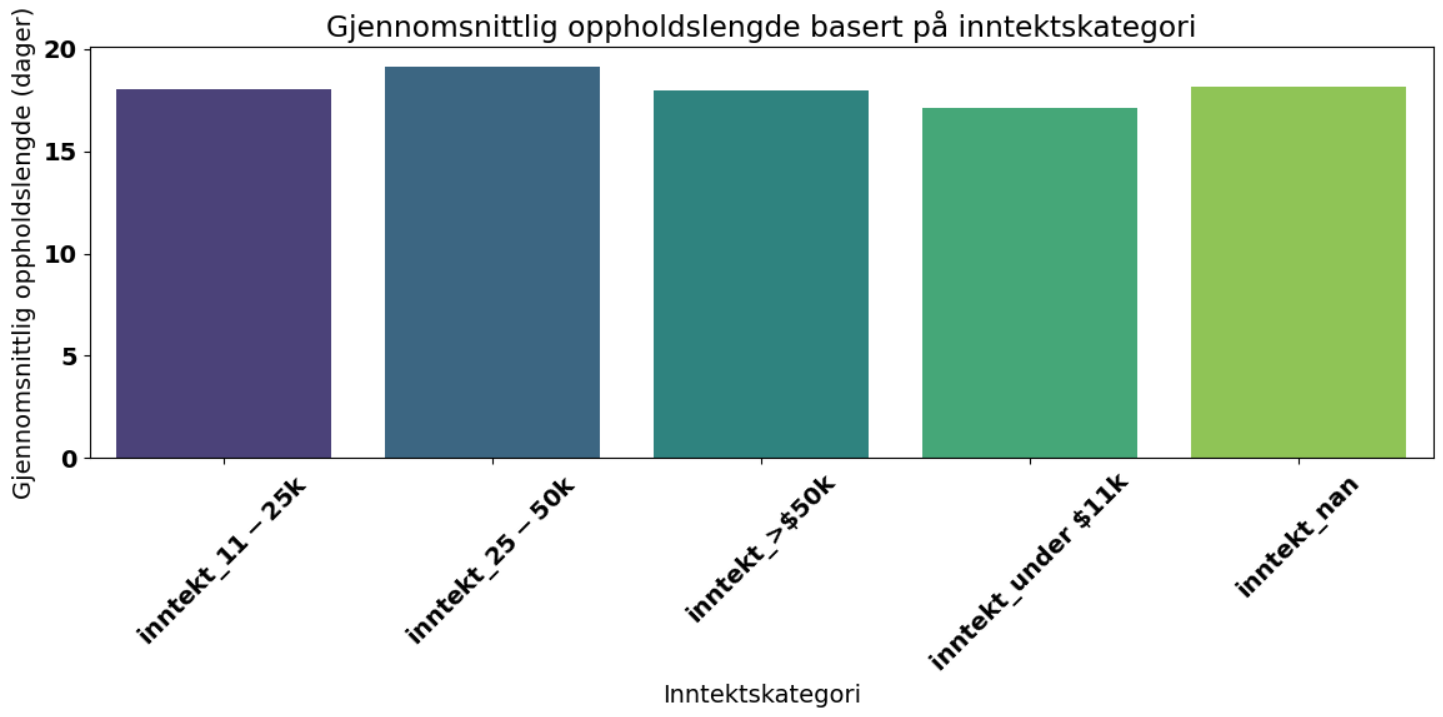
### 5.1.2 Kjønn

A red and blue squares

Description automatically generatedDet kan være interessant å undersøke om kjønn påvirker helseutfall og sykdomsforekomster.

Ut ifra dette stolpediagrammet er det en høyere andel kvinner som dør enn menn. Denne forskjellen er ikke betydelig, og vi legger ikke for mye vekt i dette resultatet. Vi undersøker videre om sosioøkonomiske faktorer har noe innvirkning på helseutfall.

### 5.1.3 Inntekt ## denne bør fjernes

Sosioøkonomiske variabler, som inntekt, kan ha innvirkning på sykdomsforekomster og alvorlighetsgrad. Min hypotese er at pasienter med høyere inntekt har bedre tilgang til medisinsk behandling, noe som kan resultere i mindre alvorlige sykdommer og kortere opphold på sykehuset. For å undersøke denne hypotesen analyserer jeg gjennomsnittlig oppholdslengde basert på inntekt i et stolpediagram.

Diagrammet viser at det er marginale forskjeller i oppholdslengde på tvers av inntektskategori. Dette viser at inntekt ikke er en avgjørende faktor for oppholdslengde. Basert på stolpediagrammet forkaster jeg min nullhypotese om at den sosioøkonomiske variablen inntekt har en innvirkning på overlevelsesraten.

## 5.2 Helserelaterte variabler

Helserelaterte variabler gir direkte informasjon om helsetilstanden til pasienten, og gir en direkte kobling mellom ulike variabler og helseutfall.

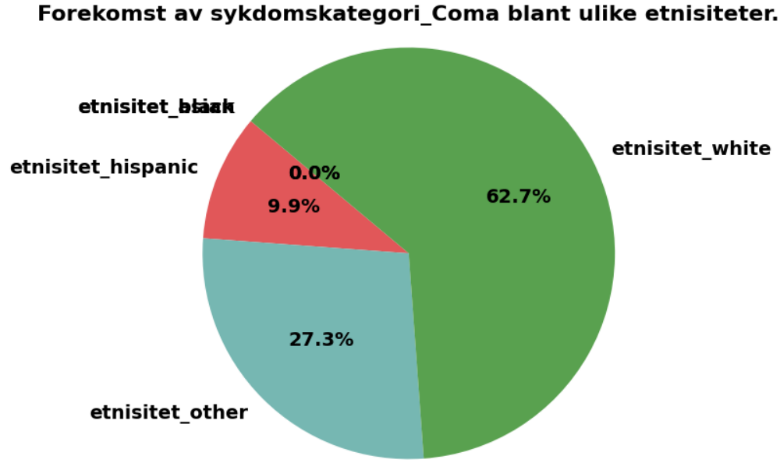
### 5.2.1 Sykdomskategorier blant etnisiteter

Å undersøke sykdomskategorier blant pasienter med ulike etnisiteter er viktig for å forstå om visse etnisiteter har økt sårbarhet for enkelte sykdommer. Om bemerkelsesverdige verdier er funnet, kan det hjelpe sykehuset med å tilpasse behandling opp mot ulike befolkningsgrupper.

A pie chart with numbers and text

Description automatically generatedA pie chart with text on it

Description automatically generated**Diagrammene nedenfor viser fordelingen av de fire sykdomskategoriene: «ARF/MOSF», «COPD/CHF/Cirrhosis», «Cancer» og «Coma» blant de ulike etnisitetene.**

A pie chart with different colored circles

Description automatically generated

Noen resultater er verdt å merke seg, blant annet andelen av *etnisitet\_white* som faller under sykdomskategori kreft og koma. Dette foreslår mulige sammenhenger mellom etnisiteter og sykdomskategori. Slike funn kan hjelpe sykehuset å tilpasse sykehusoppholdet opp mot sykdomskategoriene mer effektivt.

### 5.2.2 Korrelasjon mellom fysiologiske variabler

Oversikt over variabler tilknyttet fysiologiske variabler kan hjelpe med å diagnosere og behandle pasienter på sykehuset. Korrelasjonsmatriser kan hjelpe sykehuset med å forstå sammenhengene mellom de ulike fysiologiske variablene. Dette kan føre til raskere og mer presise diagnoser, og dermed kortere oppholdslengde. I tillegg er korrelasjonsmatriser nyttige verktøy i dataforberedelse for å eliminere høyt korrelerte variabler. Når man bygger maskinlæringsmodeller ønsker vi å eliminere høyt korrelerende variabler fordi modellen kan overtilpasse seg treningsdataene, og ikke generaliserer godt til testdata.

**A chart with different colored squares

Description automatically generated with medium confidenceKorrelasjonsmatrisen viser oversikt over korrelasjon mellom variablene i fysiologiske data.**

Matrisen gir en god oversikt over variabler som korrelerer sterkere enn andre. Eksempelvis viser *kreatinin* og *blodurea\_nitrogen* sterk positiv korrelasjon (0.92). *Kreatinin* og *blod\_ph* viser en sterk negativ korrelasjon (-0.51). Dette er variabler som muligens bør elimineres for å forhindre «overfitting» i modellen, og sørge for at den gir mer nøyaktige resultater.

### 5.2.3 Korrelasjon mellom variabler i sykdomsalvorlighet

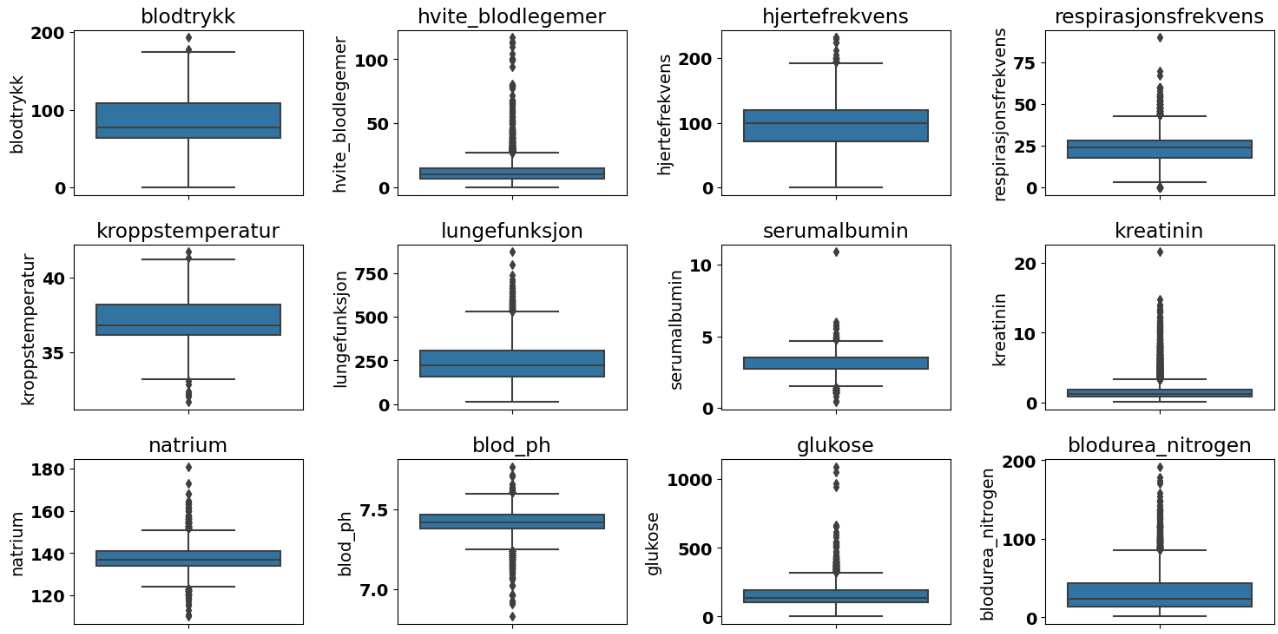
A screenshot of a data table

Description automatically generatedFor å unngå overtilpasning i modellen, er det viktig å få oversikt over korrelasjon mellom variablene i sykdomsalvorlighetsdataene. I korrelasjonsmatrisen ser vi at flere variabler som har en tilnærmet perfekt negativ eller positiv korrelasjon med hverandre. Vi identifiserer dem slik at vi kan tilpasse datasettene før vi bygger modellene.

Spesielt merker vi oss at *overlevelsesestimat\_2mnd* og *overlevelsesestimat*\_6mnd har en meget sterk positiv korrelasjon med *overlevelsesestimat\_2mnd* og *overlevelsesestimat\_6mnd*. For å unngå overtilpasning av modellen kan det være hensiktsmessig å fjerne et av disse variabelparene. Tilsvarende observerer vi at *fysiologisk\_score* og *apache\_fysiologisk\_score* er tilnærmet perfekt negativ korrelert. Derfor bør disse variablene også tilpasses eller elimineres for å sikre bedre resultater i modellene.

### 5.2.4 Fysiologiske data og uteliggere

For å oppnå god ytelse for modellene er det essensielt å forstå fordelingen av de fysiologiske variablene, for å så kunne å utelukke uteliggere i dataene. Hvis treningsdataen inneholder mye uteliggere, kan det føre til dårlig generaliseringsevne, og økt risiko for overtilpasning.

****Nedenfor er boksplott over variablene i fysiologiske data, inkludert de fire vitaltegnene blodtrykk, hjertefrekvens, respirasjonsfrekvens og kroppstemperatur.

Flere av variablene har uteliggere som kan påvirke modellens ytelse. Spesielt merker vi oss variablene *glukose*, *kreatinin*, *blod\_ph* og *hvite\_blodlegemer* som alle har relativt lav varians, og mange uteliggere. Dette kan tyde på at enkelte pasienter har ekstreme verdier som følge av alvorlig sykdom. Disse verdiene bør derfor fjernes eller behandles på en passelig måte for å oppnå best mulig ytelse i modellen. For de vitale tegnene observerer vi at blodtrykket har moderat varians, og få uteliggere. Hjertefrekvens har også moderat varians, men også en del uteliggere med høye verdier. De spesielt høye verdiene er et tegn på stressreaksjon, og kan være resultat av alvorlig sykdom. Respirasjonsfrekvens har også flere uteliggere av høye verdier. Kroppstemperatur har moderat varians og uteliggere av både høye og lave verdier. Unormalt lave eller høye kroppstemperaturer er ofte assosiert med feber eller annen alvorlig sykdom.

Ved å håndtere variablene med uteliggere på riktig måte, vil modellene være mer robuste og gi bedre predikasjoner på oppholdslengden til pasienter.

### 5.2.5 Sammenligning av overlevelsesestimat: 2 vs. 6 måneder

Nå som vi har analysert både demografiske og helserelaterte variabler har vi grunnlag for å forstå deres sammenheng med overlevelsesestimatene. Diagrammene nedenfor viser fordelingen av overlevelsesestimater over 2 og 6 måneder. X-aksen representerer overlevelsesestimatene.

A comparison of a graph

Description automatically generated with medium confidence

Overlevelsesestimat over 2 måneder har en tydelig topp på rundt 0.8, noe som tyder på at mange av pasientene har gode prognoser. Det er et markant antall pasienter som har 0% i overlevelsesprognose. Disse pasientene er i kritisk tilstand.

Fordelingen over 6 måneder er mer jevnt fordelt. Prognosene som var relativt gode etter 2 måneder svekkes over tid, noe som er forventet. Det er nå et enda høyere antall pasienter som har 0% overlevelsesestimat. Disse pasientene er kritisk syke, og har svært liten sannsynlighet for overlevelse i løpet av den gitte perioden.

# 6 Datarydding

Rydding og filtrering av data er et essensielt for å utelukke variabler som skaper forstyrrelser i modellen. Hva