INF161

Predikere lengden på sykehusopphold

Jakob Brekke Berg

Abstrakt

Denne oppgaven presenterer en maskinlæringsmodell utviklet for å predikere lengden på sykehusopphold for individuelle pasienter. Modellen benytter pasientinformasjon, inkludert fysiologiske data, demografiske data og data om sykdomsalvorlighet på tvers av ni ulike sykdomskategorier. Målet er å gi nøyaktige anslag på oppholdslengden for nye pasienter basert på disse variablene.

31. Oktober 2024

Innholdsfortegnelse

1 Introduksjon

1.1 Data

Datasettet består av journalopplysninger fra 8 261 kritisk syke pasienter samlet inn fra fem medisinske sentre i USA, i periodene 1989-1991 og 1992-1994. Hver rad i datasettet representerer medisinske data for innlagte pasienter som oppfylte inklusjons- og eksklusjonskriteriene for ni sykdomskategorier: akutt respirasjonssvikt, kronisk obstruktiv lungesykdom, hjertesvikt, leversykdom, koma, tykktarmskreft, lungekreft, multiple organsvikt med malignitet og multiple organsvikt med sepsis.

1.2 Datapreperasjon hele datasett

Analysen inkluderte tre separate datasett: sykehusdata, sykdomsalvorlighetsdata og demografiske data. Sykdomsalvorlighetsdataene var opprinnelig i JSON-format og ble konvertert til en DataFrame før de ble slått sammen med de andre datasettene til ett samlet datasett kalt **pasient_info_df**. Under denne prosessen ble to dupliserte pasientjournaler oppdaget. Disse inneholdt identiske data og ble derfor fjernet for å unngå redundans.

Pasienter med negativ **oppholdslengde** ble ekskludert fra datasettet, ettersom modellens mål er å estimere lengden på sykehusoppholdet. Videre ble aldersverdier oppført som -1 satt til NaN for å muliggjøre meningsfull imputering på et senere tidspunkt. Inntektsverdier ble konvertert til numeriske formater som er egnet for modellens krav.

En rekke fysiologiske variabler hadde ugyldige verdier (0), inkludert blodtrykk, hjertefrekvens, respirasjonsfrekvens, urinmengde og glukose. Disse ble behandlet som manglende verdier og satt til NaN med plan om imputering senere.

Variablene **sykdomskategori_id** og **sykdomskategori** hadde samme innhold, men i forskjellig format. Derfor ble **sykdomskategori_id** fjernet for å unngå overflødig informasjon. Tilsvarende hadde variablene **dødsfall** og **sykehusdød** overlappende verdier. Siden informasjon om dødsfall etter sykehusopphold ikke er nødvendig for å predikere lengden på oppholdet, ble dødsfall fjernet mens sykehusdød ble beholdt.

Jeg eksperimenterte med å fjerne alle pasienter som døde fra datasettet, noe som resulterte i en betydelig forbedring i RMSE. Imidlertid førte dette til at omtrent 2000 av 7000 pasienter ble fjernet, noe som svekket modellens generaliseringsevne. For å sikre en mer robust modell valgte jeg derfor å beholde disse pasientene, inkludert de som døde under sykehusoppholdet.

Siden den endelige modellen skal predikere oppholdslengden fra dag 1, ble variabler som først fylles ut etter dag 1 fjernet fra datasettet. Dette gjaldt kun to variabler: **bilirubin** og **adl_pasient**. I tillegg har jeg valgt å fjerne kolonnen **dnr_dag** ettersom **dnr_status** allerede gir tilstrekkelig informasjon om en pasient har fått DNR-status før eller etter innleggelse. **dnr_dag** inneholder detaljert informasjon om hvilken dag DNR-status ble gitt, men mange av disse verdiene er mangelfulle eller ufullstendige, noe som gjør imputering vanskelig og kan føre til usikkerhet i modellene. Ved å bruke **dnr_status** unngår vi utfordringer knyttet til manglende verdier og komplisert imputering, samtidig som vi beholder nødvendig informasjon om DNR-status.

Alle variabler som opprinnelig var lagret som objekter, men som representerte numeriske verdier, ble konvertert til enten heltall eller flyttall, avhengig av hva som var mest hensiktsmessig for hver variabel.

Til slutt ble datasettet delt inn i treningssett, valideringssett og testsett ved bruk av scikit TrainTestSplit for å klargjøre det for modelltrening og evaluering. Jeg har valgt å dele datasettene i 70% treningsdata, 15% valideringsdata, og 15% testdata. Dette for å bidra til at modellene har nok data å trene seg på, men også at valideringen av de ulike modellene er hensiktsmessig.

1.3 Datapreperasjon treningsdata

For å sikre at modellen kun mottar numeriske verdier, er OneHotEncoding brukt på de kategoriske variablene. Dette gjør at hver kategori i de diskrete variablene er representert som egne kolonner med binære verdier slik at de enkelt kan brukes i modelleringen. Videre ble det opprettet en ny variabel for å representere alvorlighetsgraden av sykdom basert på fysiologisk score. Visualiseringene viste at ulike intervaller i fysiologisk score hadde en betydelig

innvirkning på lengden av sykehusopphold og derfor ble disse intervallene kategorisert som høy, middels og lav alvorlighetsgrad.

I tillegg ble flere variabler opprettet for å bedre fange opp interaksjoner og forhold i dataene:

Omfattende behandling: En variabel for omfattende behandling ble laget for pasienter med spesifikke kjennetegn, inkludert lav koma_score, lavt antall komorbiditeter, og middels fysiologisk_score, da disse pasientene ofte hadde lengre opphold.

Aldersgrupper: Basert på pasientenes alder ble det også laget en kategori som grupperer dem i ung, middelaldrende, og eldre, for å undersøke om aldersgrupper kan påvirke sykehusoppholdet.

Alder-Fysiologisk Interaksjon: Ved å multiplisere alder med fysiologisk_score, kunne modellen bedre identifisere aldersrelaterte helseutfordringer.

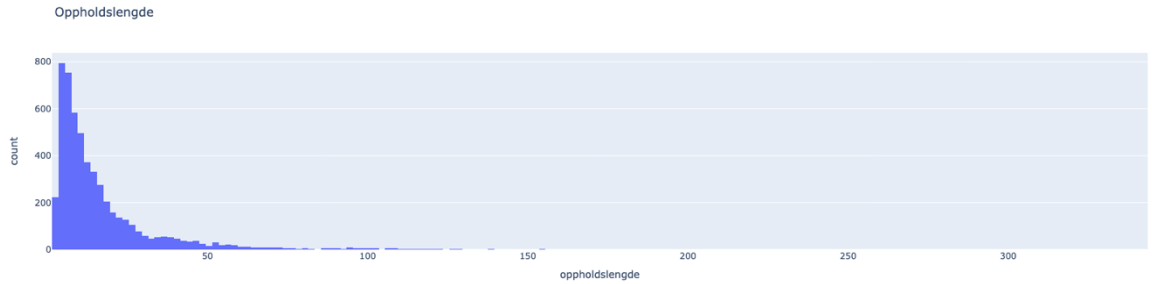
Gjennomsnitt_oppholdslengde_sykdom_underkategori: Det er stor variasjon mellom ulike sykdom_underkategorier, og deres oppholdslengde. For å få frem disse variasjonene, er det opprettet en variabel som viser den gjennomsnittlig oppholdslengden for pasienter med ulik sykdom.

Flere variabler ble testet for å utforske deres potensielle nytte, men ble til slutt fjernet da de viste seg å øke RMSE og forstyrre modellens nøyaktighet:

- **Forholdstall**: Forholdet mellom enkelte verdier, som natrium/kreatinin, ble opprettet, men ga ingen forbedring i modellen.
- **Overlevelsesestimatdifferanse**: En variabel som reflekterte forskjellen i overlevelsesestimat over tid ble vurdert som en indikator for behandlingsforløpet, men viste seg å være lite prediktiv.
- **Kreftstatus og Underkategorier**: Flere kreftspesifikke variabler, inkludert en generell kreftstatus kreft_yes og underkategorier som metastatic, ble laget for å forbedre modellens evne til å predikere sykehusopphold hos kreftpasienter. Disse ble senere ekskludert da de ikke forbedret modellen.

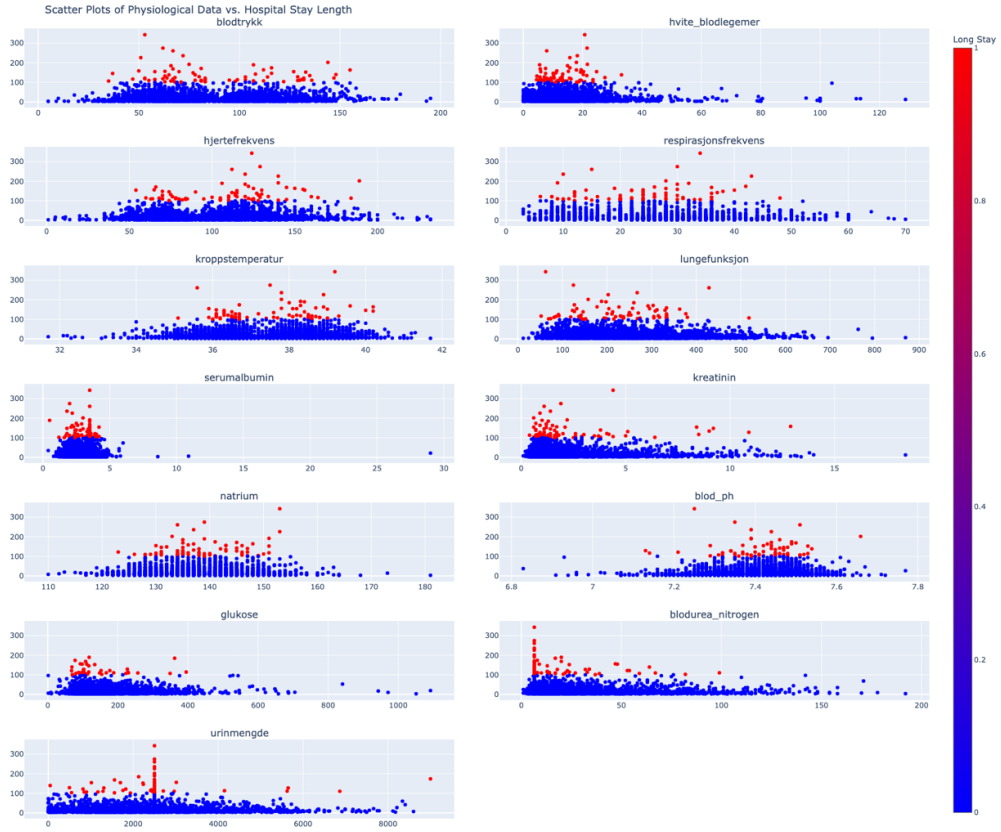
2 Beskrivelse av data

2.1 Dataoversikt på treningsdata



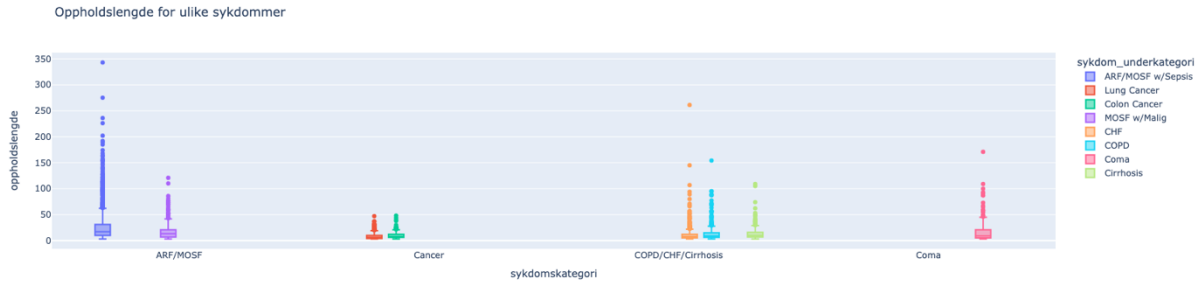
Figur 1 Histogram av oppholdslengde for pasienter i treningsdata

I treningsdatasettet ser vi at de fleste pasientene har en oppholdslengde i intervallet [0-50] dager. Likevel har 313 av totalt 5413 pasienter i treningsdataen en oppholdslengde på over 50 dager. Antall pasienter reduseres betydelig jo lengre oppholdslengden er. Av de 313 pasientene med lengre opphold, har 235 en oppholdslengde i intervallet [50-100] dager. Videre har 55 pasienter en oppholdslengde mellom 100-150 dager, mens kun 20 pasienter har en oppholdslengde som overstiger 150 dager.



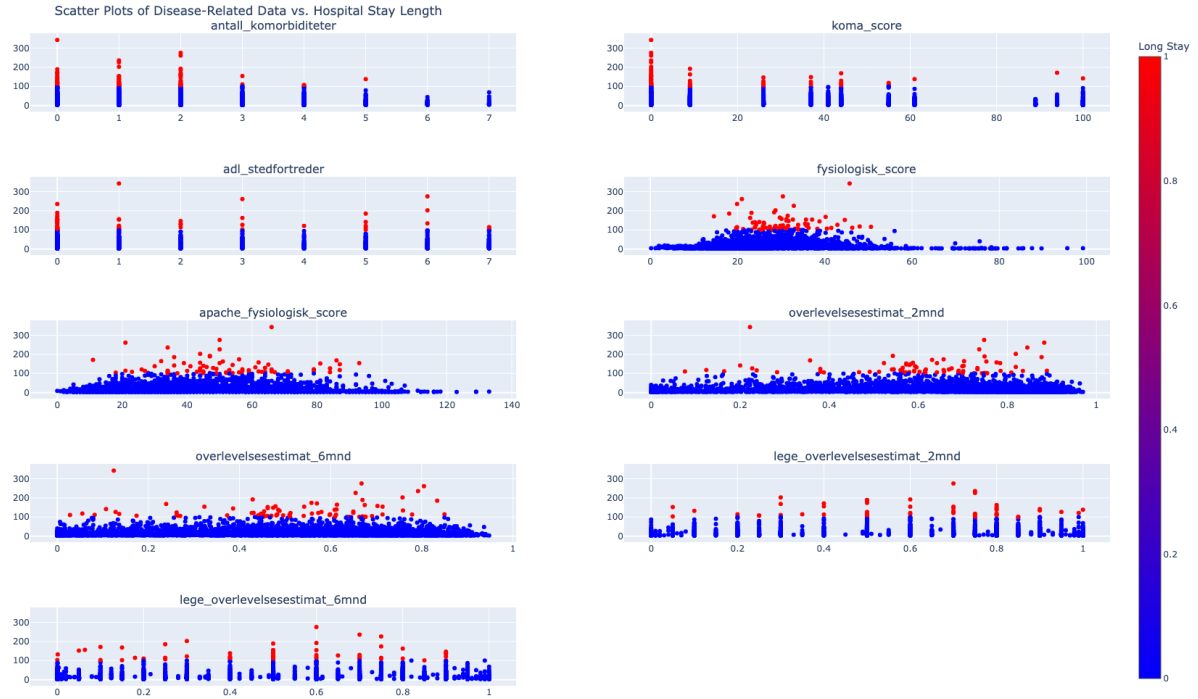
Figur 2 ScatterPlot av numeriske fysiologiske variabler i treningsdata

Dette viser en visualisering av fysiologiske variabler. Jeg har opprettet en terskel på 100 dager. Alle pasienter som har en oppholdslengde større enn 100 dager vil bli farget rødt og resten vil være blå. Dette ble brukt for å se om det er visse intervaller for fysiologiske data som vil gi en lengre oppholdslengde.



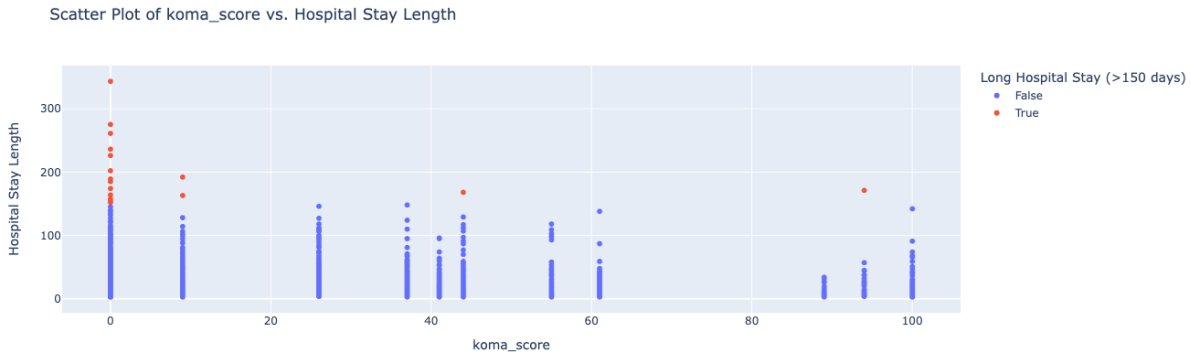
Figur 3 BoxPlot av oppholdslengden til pasienter i treningsdata baser på sykdom

Når vi analyserer oppholdslengden i forhold til ulike sykdommer ser vi at sykdommen med lengst oppholdslengde, trolig på grunn av lang behandlingstid, er ARF/MOSF. Vi bemerker oss spesielt at pasienter med sykdomsunderkategorien ARF/MOSF w/Sepsis har en betydelig lengre oppholdslengde enn andre sykdommer. Medianen, samt første (Q1) og tredje kvartil (Q3), er betraktelig høyere for denne gruppen sammenlignet med andre sykdommer. Vi observerer også mange uteliggere (outliers) i hele intervallet fra 60 til 348 dager.



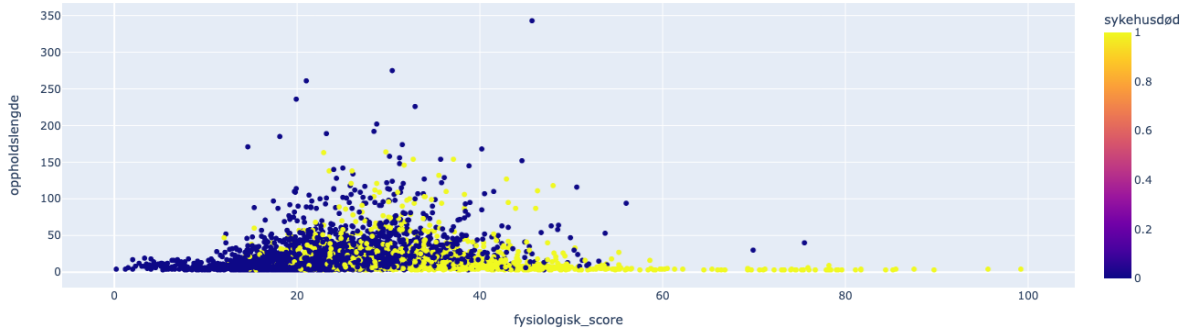
Figur 4 ScatterPlot av numeriske variabler i sykdomsalvorlighet opp mot oppholdslengde for pasienter i treningsdata

Dette viser oss fordelingen av de numeriske dataene i sykdomsalvorlighet opp mot oppholdslengden. Her er det på samme måte fargelagt rødt for pasienter som har en oppholdslengde lenger enn 100 dager. Her er det noen av visualiseringene det kan være lurt å se nærmere på.



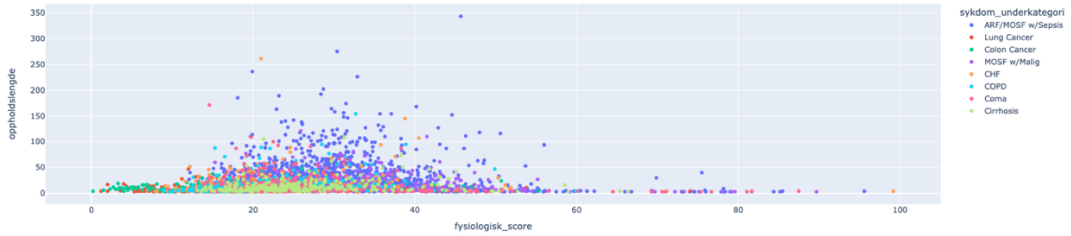
Figur 5 ScatterPlot av Koma_score opp mot oppholdslengde for pasienter i treningsdata

Pasienter med lav koma-score har også ofte lengre oppholdslengde enn andre. En lav koma-score indikerer at pasienten viser liten grad av våkenhet, noe som kan kreve mer omfattende behandling og føre til lengre sykehusopphold.



Figur 6 Scatterplot av den fysiologiske scoren opp mot oppholdslengde for pasienter i treningsdata

Når vi undersøker sammenhengen mellom oppholdslengde og fysiologisk score, finner vi at pasienter med en fysiologisk score i intervallet [10-60] dager ofte har lengre oppholdslengde enn andre. Grafene viser også at pasienter med høy fysiologisk score og kort oppholdslengde ofte er så alvorlig syke at de dør raskt etter innleggelsen. På den andre siden ser vi at pasienter med en fysiologisk score lavere enn 10 generelt har kortere oppholdslengde, noe som kan tyde på at de ikke er alvorlig syke og derfor krever mindre behandlingstid.

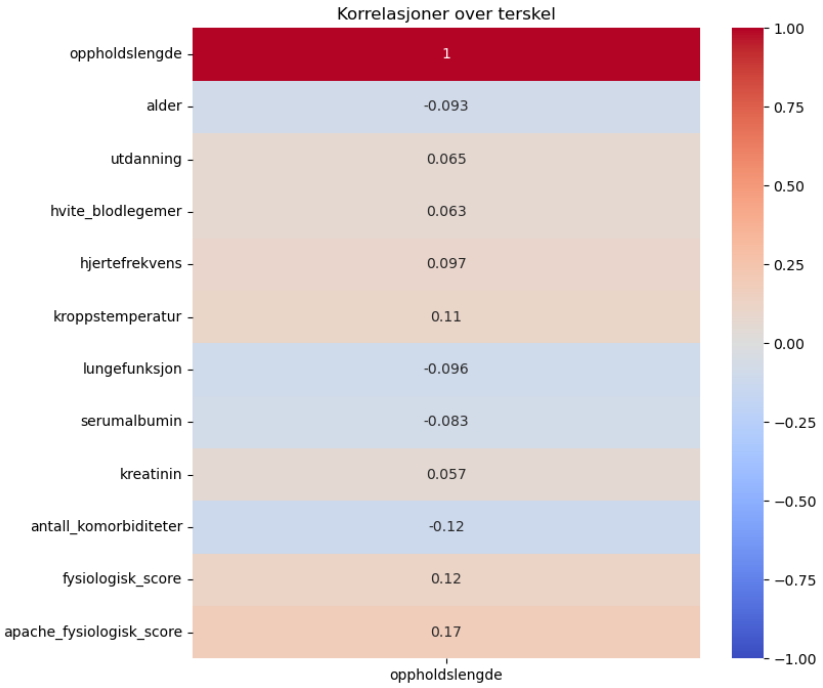


Figur 7 ScatterPlot av fysiologisk score opp mot oppholdslengde for pasienter i treningsdata utifra sykdom

Her ser vi en visualisering hvor fysiologisk score er satt opp mot oppholdslengden. Hvis vi setter farge etter sykdom_underkategori ser vi at ulike sykdom_underkategorier har forskjellige kjennetegn. Spesielt ser vi at pasienter med sykdom_underkategori ARF/MOSF w/Sepsis både har en lengre oppholdslengde, men også en høyere fysiologisk score. Pasienter som har Cirrhosis og Colon Cancer har også ofte en lavere oppholdslengde.

2.1.1 Korrelasjon

Denne figuren viser en korrelasjonsmatrise som fokuserer på variabler med en korrelasjonskoeffisient større enn 0.05 i forhold til variabelen "oppholdslengde". Ved å sette en terskel på 0.05, har vi filtrert bort variabler med svært svak korrelasjon, slik at vi kan fokusere på de som har en viss sammenheng med oppholdslengde.



Figur 8 Korrelasjonsmatrise for numeriske variabler mot oppholdslengde med en terskel på 0,05

Fra figuren kan vi observere at:

Apache_fysiologisk_score har den sterkeste positive korrelasjonen med oppholdslengde (0.17), noe som kan indikere at høyere Apache-poengsum, som er en indikator på pasientens helsetilstand, er assosiert med lengre sykehusopphold.

Fysiologisk_score og **kroppstemperatur** har også en svak positiv korrelasjon med oppholdslengde, men i mindre grad (0.12 og 0.11).

Antall_komorbiditeter viser en svak negativ korrelasjon (-0.12) med oppholdslengde, noe som kan indikere at flere samtidige sykdommer har en svak tendens til å redusere oppholdslengden.

Alder, **lungefunksjon**, og **serumalbumin** har også små negative korrelasjoner, som kan være indikasjoner på sammenhenger, men disse er ikke sterke.

Denne figuren gir verdifull innsikt i hvilke variabler som kan være relevante for å predikere eller forstå oppholdslengden til pasienter. Til tross for at ingen av korrelasjonene er veldig sterke, kan variabler med svake, men signifikante korrelasjoner være nyttige i en modell, særlig i kombinasjon med andre faktorer. Tabellen gir dermed et godt utgangspunkt for videre analyser og eventuell funksjonsutvelgelse i en regresjonsmodell.

3 Imputasjon og modellering

3.1 Imputering av manglende verdier

For å håndtere manglende verdier i datasettet, valgte jeg først å imputere enkelte variabler med anbefalte verdier hentet fra SUPPORT2-datasettet. Disse verdiene ble valgt for å gi et godt utgangspunkt for videre modelltrening.

Variabel	Anbefalt imputeringsverdi
serumalbumin	3.5
lungefunksjon	333.3
bilirubin	1.01
kreatinin	1.01
blodurea_nitrogen	6.51
hvite_blodlegemer	9
urinmengde	2502

Etter at disse verdiene ble imputert, inneholdt datasettet fortsatt enkelte variabler med manglende verdier, som vist her:

Variabel	Manglende verdier
alder	2
utdanning	955
Inntekt	1738
Etnisitet	26
Blodtrykk	31
Hjertefrekvens	53
Respirasjonsfrekvens	41
Blod_ph	1365
Glukose	2683
Adl_stedfortreder	1702
Lege_overlevelsesestimat_2mnd	984
Lege_overlevelsesestimat_6mnd	974
Dnr_status	4655

Figur 9 nedenfor viser histogrammer for variabler med manglende numeriske data, og kan brukes som utgangspunkt for å analysere fordelingen av verdier og planlegge en egnet imputering for hver variabel. Hvert subplot representerer en variabel, og histogrammene

viser hvor hyppig hver verdi forekommer. Dette gir oss innsikt i variasjonen, omfanget, og formen på distribusjonen for hver variabel

Forklaring av variablene og hva histogrammene viser

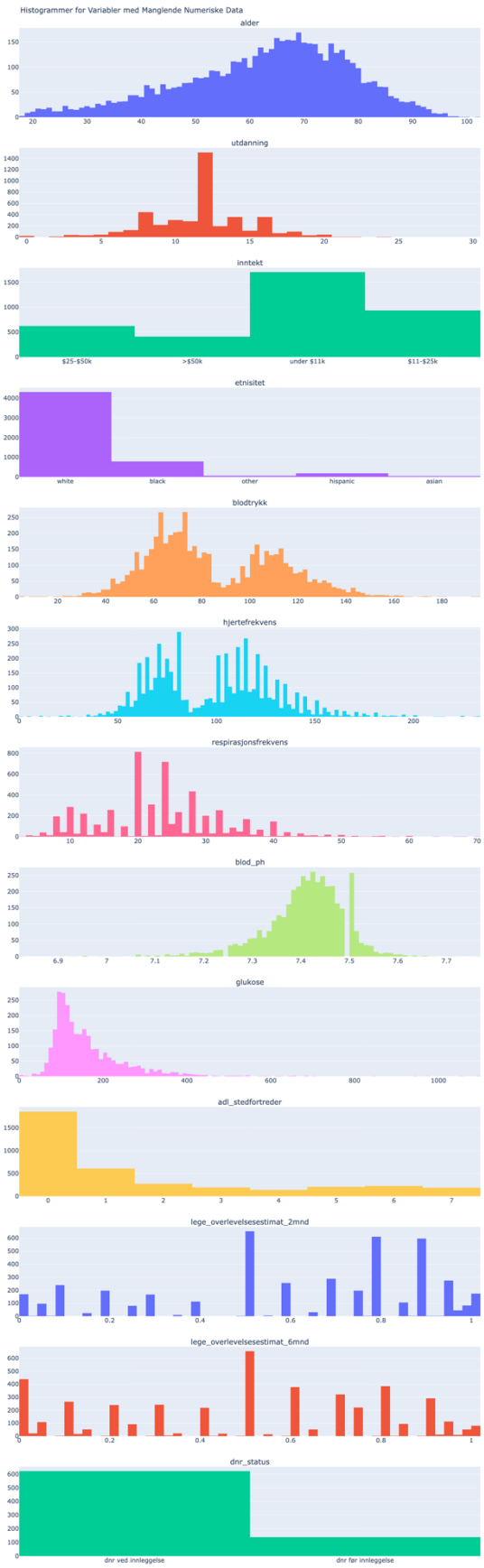
alder: Denne variabelen har en relativt normalfordelt aldersfordeling som topper seg rundt 70 år. Dette gir grunnlag for å bruke median eller gjennomsnitt som imputering for eventuelle manglende verdier.

utdanning: Utdanningsnivå ser ut til å være konsentrert rundt noen få spesifikke verdier, med et tydelig toppunkt. For en slik kategorisk fordeling kan modus være en passende imputering.

inntekt: Inntekt er delt i kategorier, og histogrammet viser en ujevn fordeling, der de fleste faller i kategoriene "$11-$25k" og "under $11k". Her kan modus brukes som imputering, ettersom dette er en diskret variabel.

etnisitet: Etnisitet har en kraftig skjevfordeling, hvor de fleste pasientene tilhører kategorien "white". Også her kan modus være en passende imputering for manglende verdier.

blodtrykk, hjertefrekvens, respirasjonsfrekvens, blod_ph, og glukose: Disse variablene har forskjellige kontinuerlige fordelinger som gir en bred variasjon i verdier. Blodtrykk og hjertefrekvens har en normalfordeling, mens blod_ph og glukose er mer skjevt fordelt. For disse variablene kan median være en god imputering for å unngå påvirkning fra ekstreme verdier.



Figur 9 Histogram for variabler i treningsdata som inneholder manglende verdier

adl_stedfortreder: Denne variabelen har svært lav variasjon, med de fleste verdiene samlet i et lite antall kategorier. Modus kan være en hensiktsmessig imputering her, da det er en diskret variabel.

lege_overlevelsesestimat_2mnd og lege_overlevelsesestimat_6mnd: Disse variablene viser overlevelsesestimater, som også er skjevt fordelt, men har en diskret fordeling. Median kan brukes som imputering, ettersom variabelen har både høy og lav variasjon innenfor bestemte områder.

dnr_status: Denne variabelen har kategorier i visualiseringen ("dnr ved innleggelse" og "dnr før innleggelse"), hvor de fleste pasientene er i kategorien "dnr ved innleggelse". Likevel er mulige verdier "dnr ved innleggelse", "dnr før innleggelse", "mangler", "ingen dnr". De fleste pasienter vil ikke ha noen dnr_status. Jeg setter derfor pasienter som ha dnr_status=NaN til å ha ingen dnr.

Etter å ha evaluert ulike imputeringsstrategier basert på distribusjonen og mangler for hver variabel, ble følgende tilnærming valgt:

Gjennomsnittsimputasjon for kontinuerlige numeriske variabler:

For variabler som har en jevn eller normalfordeling og mangler enkelte verdier, brukes gjennomsnitt som imputeringsmetode. Dette er en god strategi når dataene ikke er betydelig skjeve, og det bidrar til å opprettholde det samlede gjennomsnittet i datasettet. Dette er implementert i num_pipeline_mean, som også standardiserer verdiene etter imputering for å sikre konsistent skala i modellen.

Eksempler på variabler her inkluderer generelle numeriske kolonner som blodtrykk og hvite_blodlegemer.

Median imputasjon for skjevfordelte variabler:

Variabler som er skjevt fordelt, som aldersrelaterte estimater (alder, lege_overlevelsesestimat_2mnd, lege_overlevelsesestimat_6mnd), benytter median som imputeringsstrategi. Median reduserer påvirkningen av ekstreme verdier (outliers) som ofte finnes i skjeve fordelinger. Denne strategien er implementert i num_pipeline_median og er nyttig når variabler har høy skjevhet.

Modus imputasjon for kategoriske and diskrete numeriske Variables:

For kategoriske variabler og diskrete numeriske variabler med få unike verdier, som adl_stedfortreder, brukes modus (den mest vanlige verdien) for å fylle inn mangler. Dette gir et realistisk estimat som er i tråd med dataenes distribusjon, spesielt når mangler kan antas å følge den vanligste verdien i dataene. num_pipeline_mode håndterer disse kolonnene.

Kategoriske variabler som kjønn, etnisitet, og sykdomskategori benytter også modus gjennom cat_pipeline, hvor de deretter blir one-hot encodet for å gjøre dem klare til modelltrening.

Passthrough Pipeline for spesifikke kolonner:

For enkelte kolonner, som sykehusdød, demens, diabetes, og omfattende_behandling, er det ikke nødvendig med ytterligere transformasjoner utover imputering, da de allerede er i en format som kan benyttes direkte i modeller. Disse kolonnene går gjennom en passthrough_pipeline, hvor eventuelle mangler imputeres med modus.

KNN Imputer – Valgt Bort:

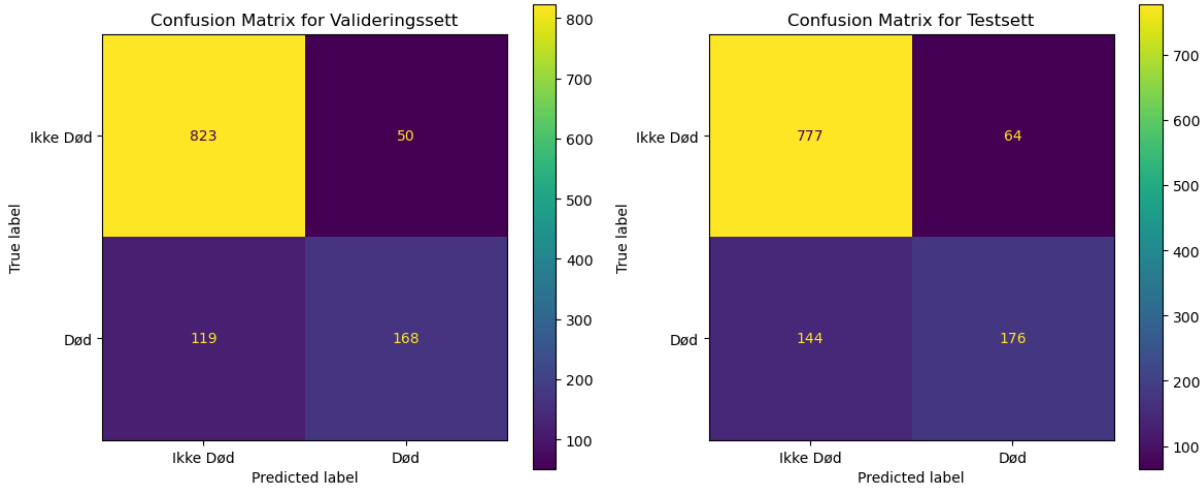
Selv om KNNImputer ble vurdert, viste det seg at andre metoder som median og modus ga bedre resultater i dette tilfellet. KNNImputer er ressurstungt og kan være vanskelig å tolke, særlig når verdier er svært skjeve eller når variabler har høy korrelasjon med hverandre. Derfor ble denne metoden valgt bort til fordel for de andre enklere og mer tolkningsvennlige metodene.

Ved å bruke en kombinasjon av mean, median, og modus, tilpasser vi imputeringen til variabelens karakteristika, noe som gir en robust løsning som bevarer datakvaliteten og reduserer skjevheter.

3.2 Klassifikasjonsmodell for sykehusdød

Både treningsdata, valideringsdata, og testdata inneholder en variabel sykehusdød. Denne vil si om pasienten døde under sykehusoppholdet. Vi ønsker å ha en modell som kan predikere sykehusoppholdet til pasienten fra dag 1. For å unngå at modellen er biased og allerede vet om pasienten døde under oppholdet eller ikke har jeg valgt å fjerne denne variabelen fra validerings og treningsdata. Jeg har trent opp en klassifikasjonsmodell som vil klassifisere om pasienten trolig vil dø under oppholdet eller ikke. Jeg setter så disse nye verdiene inn i

valideringsdata slik at modellen ikke vil være biased med informasjon den ikke skal ha tilgang til. Modellen har en nøyaktighet med 85% på valideringsdata, og 82% på testdata.



Figur 9 Confusion Matrix for klassifikasjonsmodellen til sykehusdød

Hvis vi studerer resultatet av klassifikasjonsmodellen på valideringsdata, ser vi at 991 av pasientene får riktig predikasjon av sykehusdød. Likevel er det 50 pasienter som blir predikert død, selv om de ikke døde under sykehusoppholdet. 119 pasienter blir også klassifisert som ingen sykehusdød, selv om de døde under sykehusoppholdet.

3.3 Modellering og modellutvalg

Modellutvalget i dette prosjektet er basert på best mulig generaliseringsevne, målt ved hjelp av Root Mean Squared Error (RMSE). RMSE gir et mål på modellens prediksjonsfeil ved å vurdere avstanden mellom de predikerte verdiene og de faktiske verdiene:

$$\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

3.4 Grunnlinje-modeller

For å ha et referansepunkt trente jeg først to grunnlinjemodeller ved hjelp av DummyRegressor:

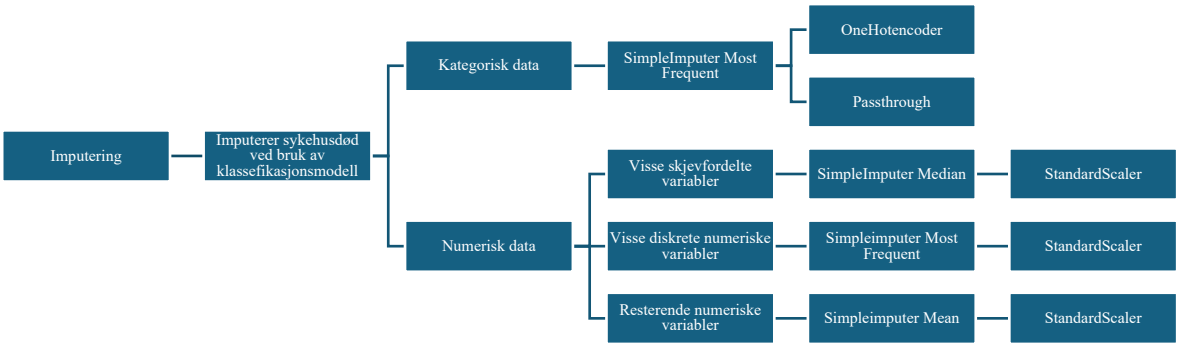
DummyRegressor med strategi gjennomsnitt: RMSE = 21,5961

DummyRegressor med strategi median: RMSE = 22,7121

Disse grunnlinjemodellene ga et utgangspunkt for sammenligning med mer avanserte modeller.

3.5 Testing av ulike modeller og imputeringsstrategier

Videre testet jeg flere modeller og ulike imputeringsstrategier for å finne den beste generaliseringsevnen. Jeg har forsøkt både med KNNImputer, og forskjellige parametere for SimpleImputer. Tabellen nedenfor viser de endelige imputeringsmetodene som ble brukt grunnet best resultat:



Etter variabelutvinningen og imputasjonen av manglende verdier går jeg videre for å teste fire ulike regresjonsmodeller.

3.5.1 RandomForestRegressor:

RandomForestRegressor er en modell som bygger flere beslutningstrær ved å bruke ulike deler av datasettet. I stedet for å bruke ett enkelt beslutningstre, kombinerer den mange trær og tar gjennomsnittet av deres resultater. Dette bidrar til å gi mer presise prediksjoner og reduserer risikoen for at modellen blir tilpasset bare til det spesifikke treningsdatasettet (overtilpasning).

3.5.2 ExtraTreesRegressor:

ExtraTreesRegressor er en modell som bygger flere **tilfeldige** beslutningstrær (også kalt ekstra-trær) ved å bruke ulike deler av datasettet. Ved å kombinere resultatene fra disse trærne gjennom å ta gjennomsnittet, forbedrer modellen nøyaktigheten til prediksjonene og reduserer risikoen for overtilpasning (at modellen kun passer treningsdataene og ikke fungerer godt på nye data).

3.5.3 Ridge:

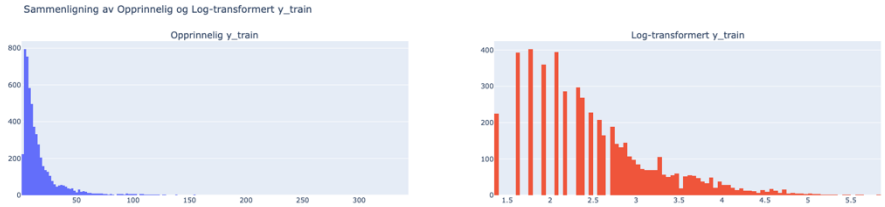
Denne modellen bruker lineær regresjon med l2-regularisering, for å finne de beste vektene til å forutsi målet (y) ut fra input (X). Målet er å minimere en funksjon som kombinerer to elementer:

1. **Avstanden mellom de faktiske og de predikerte verdiene** (målt ved minste kvadrater),
2. **En straff for store vekter (w)**, som er l2-regulariseringen.

Denne straffen gjør at modellen blir mer stabil og unngår overtilpasning.

3.5.4 ExtraTreesRegressor med log transformasjon på target variabel:

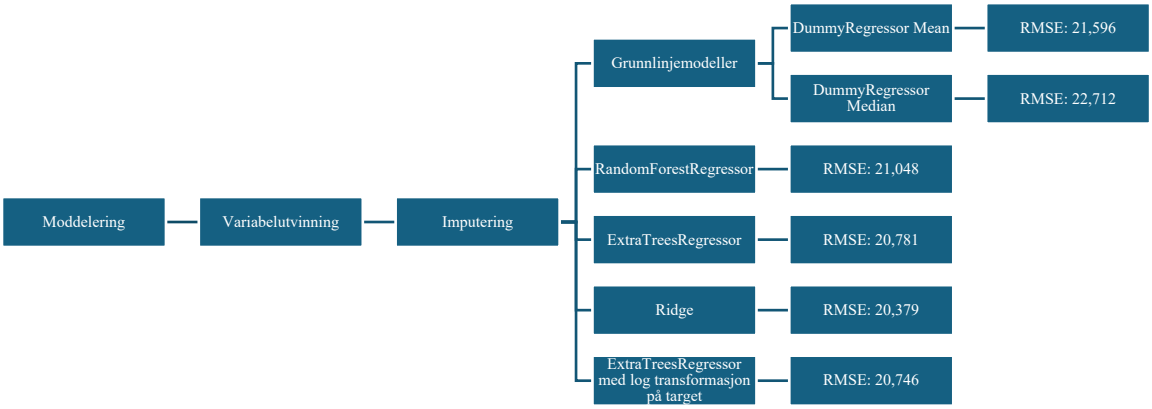
Vi ser at oppholdslengden til de aller fleste pasientene ligger i intervallet [0-50]. Dette gjør at vi får mange uteliggere med oppholdslengde > 50 dager. For å normalisere den positive skjevheten gjør jeg en logaratmisk transformasjon på oppholdslengde.



Figur 10 Sammenligning av målvariabel med og uten logaritmisk transformasjon

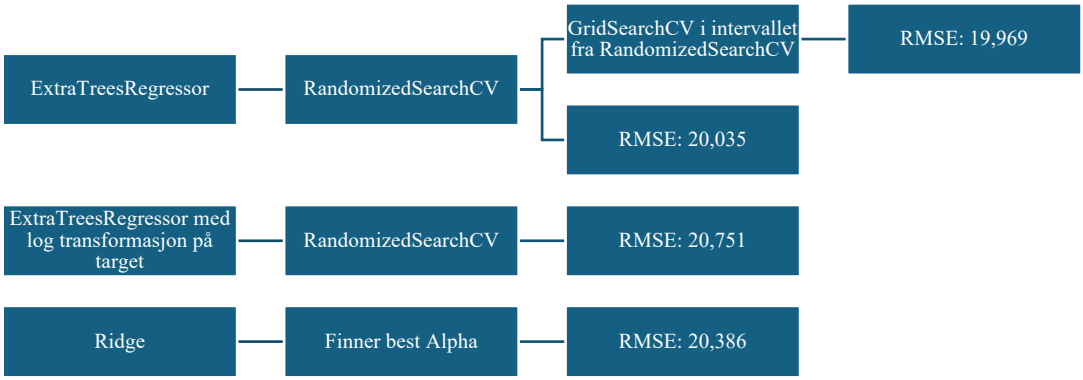
Jeg bruker funksjonen log1p(y_train) for å unngå problemer med log(0). Selvom dette ga en bedre distribusjon av oppholdslengden, førte det ikke til forbedring for RMSE.

Alle de ulike modellene ble testet med de ulike imputeringsstrategiene, og valget av modell ble gjort basert på hvilken kombinasjon av modell og imputeringsstrategi som ga lavest RMSE på valideringsdataene.



3.6 Resultater for modell og imputeringsstrategi

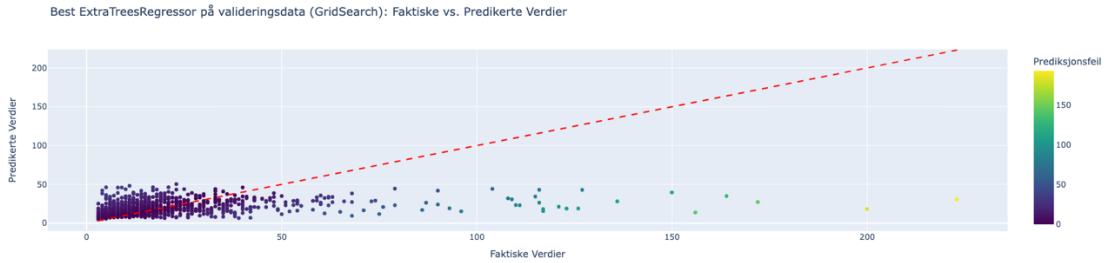
ExtraTreesRegressor ga best resultat. Jeg gå derfor videre med å finne passende hyperparametere. Vi har allerede delt opp i trenings-, validerings-, og testdata. Grunnet at vi ønsker å finne beste parametere og RMSE på valideringsdata tar jeg i bruk scikit learn PredifinedSplit. Denne vil gjøre de mulig å bruke GridSearchCV og RandomizedSearchCV, men ved å teste de ulike parametrene på valideringsdata. Når jeg har funnet beste parametere refitter jeg med beste parametere funnet på kun treningsdata. Bruker først randomized search, med følgende parametere. Videre kjører jeg GridSearch I området rundt disse intervallene. Siden både Ridge, og ExtraTreesRegressor med logaritmisk transformasjon også ga en relativt god RMSE, har jeg testet å finne beste Alpha for Ridge, og beste parametere for ExtraTreesRegressor med logaritmisk transformasjon.



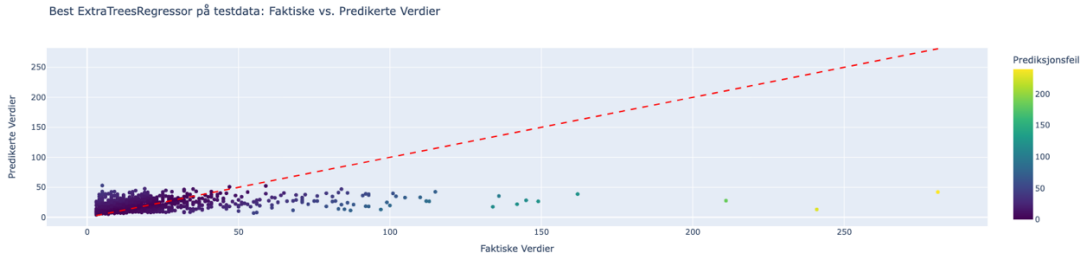
4 Analyse av beste modell

Modellen som ga best RMSE er ExtraTreesRegressor med hyperparametere. RMSE på valideringsdata for denne modellen ble 19,969. Denne ble så videre testet på testdata. Prediksjonen på testdata ga følgende resultat: RMSE 21,626.

Prediksjonen på valideringsdata, og test data er ganske like da prediksjonen versus faktiske oppholdslengde følger nogenlunde samme mønster. Dette viser at modellen ikke overfitter på valideringsdata.



Figur 11 Faktisk vs predikert oppholdslengde for valideringsdata



Figur 12 Faktisk vs predikert oppholdslengde for testdata

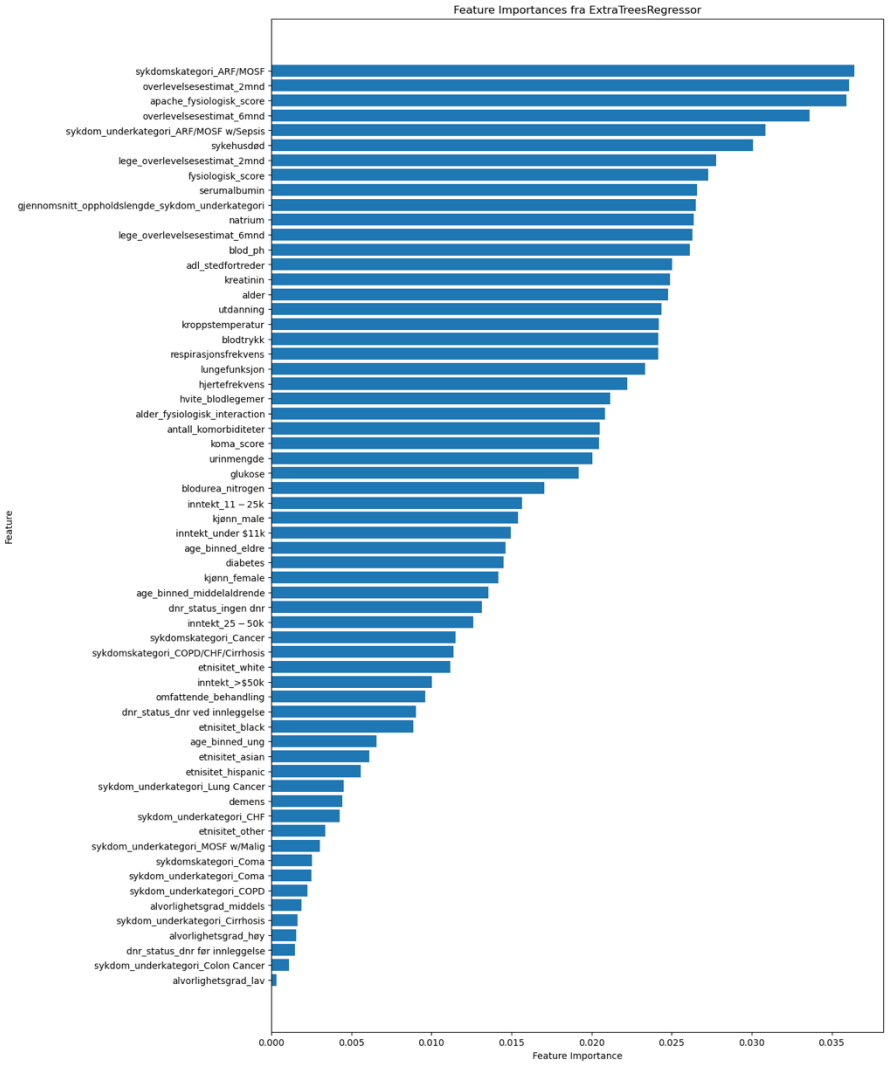
Vi ser tydelig på resultatet, at modellen er mer nøyaktig i intervallet [0-50] dager. Det er også i dette intervallet oppholdet til de fleste pasientene vil ligge, og dermed dette intervallet modellen vil være best trent på.

Figur 13 Faktisk vs predikert oppholdslengde for testdata i intervallet [0-50]

Hvis vi nå ser nærmere på intervallet fra [0-50] dager, vil det være visse avvik, men generelt bedre enn for pasienter med lengre oppholdslengde.

4.1 Feature importance

Hvis vi ser på feature importance (variablenes betydning), ser vi at sykdom_underkategorien ARF/MOSF w/Sepsis er spesielt viktig for modellen. Dette er i tråd med det vi har sett tidligere, da pasienter med denne sykdommen ofte har et lengre sykehusoppholds sammenlignet med andre pasienter. Den nylig opprettede variabelen for gjennomsnittlig oppholdslengde for ulike sykdom_underkategori har også stor innflytelse for modellen. Vi ser også at sykehusdøden er en viktig variabel. Om en pasient dør eller ikke under sykehusoppholdet vil være avgjørende for en predikasjon på oppholdslengden. Derfor vil nok valget med å imputere denne verdien være et godt. Binningen av aldersgrupper vil også være hensiktsmessig ifølge modellens variabelbetydning. Variabelen alvorlighetsgrad har en veldig lav betydning for modellen. Likevel har jeg forsøkt å fjerne ulike variabler med lav betydning, for å se om dette resulterte i et bedre resultat (RMSE). Det gjorde det ikke, og jeg har derfor valgt å beholde disse variablene.



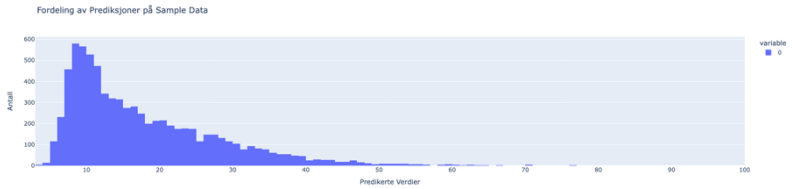
Figur 14 Feature importance for beste modell

4.2 Prediksjon av sample data

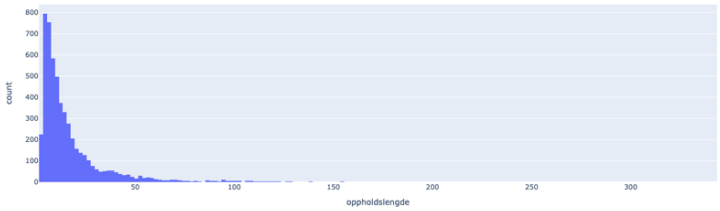
Jeg håndterer sample data på samme måte som for raw_data. Videre imputerer jeg ved bruk av pipeline, som beskrevet tidligere.

Bruker så ExtraTreesRegressor med beste parametere funnet ved bruk av valideringsdata. Videre har jeg opprettet et histogram for å se på resultatet av prediksjonen på sample data.

Hvis vi ser på resultatet av prediksjonen på sample data sammenlignet med faktiske verdier fra treningsdata, ser vi at dataen følger noe samme mønstre. Begge har en positiv skjevhet, da de fleste pasientene har en kortere oppholdslengde. Likevel ser vi som før, at modellen har vanskeligheter med å predikere lengre oppholdslengder



Figur 15 Histogram over prediksjonen av oppholdslengde på sample data



Figur 16 Histogram over oppholdslengde for pasienter i treningsdata

Dette vil være et resultat av flere ulike faktorer. Modellen har få pasienter med lengre oppholdslengde å trene seg på. I tillegg er det små forskjeller mellom en pasient med lengre oppholdslengde sammenlignet med en med kort oppholdslengde. Dette vil gjøre at modellen vil predikere mye kortere for noen pasienter som har en oppholdslengde på mer en 50 dager.

5 Nettside

Jeg har laget en HTML fil som formaterer nettsiden, og tar inn pasient data. Nettsiden hostes lokalt på port 8080. Alle numeriske verdier som blir skrevet inn må være større eller lik 0. Dette for å forhindre at ugyldig informasjon blir gitt modellen. All den manglende pasient dataen vil bli imputert, og det vil bli gjort en prediksjon ved bruk av ExtraForestRegressor. I tillegg har jeg brukt klassifikasjonsmodellen for å imputere sykehusdød. Som en tilleggsfunksjon til nettsiden har jeg laget en advarselsmelding. Denne vil vises dersom klassifikasjonsmodellen klassifiserer pasienten med sykdomsdød=1, altså at det er en høy risiko for at pasienten dør under sykehusoppholdet.

6 Konklusjon og refleksjon

6.1 Konklusjon

I dette prosjektet har jeg utviklet en modell for å predikere lengden på sykehusoppholdet til kritisk syke pasienter ved hjelp av omfattende pasientdata. Gjennom dataforberedelse og utforskning har jeg identifisert viktige variabler og mønstre som påvirker oppholdslengden. Jeg har håndtert utfordringer knyttet til manglende og ugyldige verdier, fjernet overflødige variabler, og opprettet nye funksjoner for å forbedre modellens ytelse.

Ved å teste ulike regresjonsmodeller fant jeg at **ExtraTreesRegressor** med optimaliserte hyperparametere ga den beste prediksjonsnøyaktigheten, med en RMSE på 20,012 på valideringsdata og 21,456 på testdata. Modellen viser god ytelse innenfor intervallet 0–50 dager, hvor majoriteten av pasientene befinner seg, men har redusert presisjon for lengre oppholdslengder. Dette skyldes delvis skjevheten i datasettet, hvor færre pasienter har svært lange opphold, noe som gjør det utfordrende for modellen å generalisere i disse områdene.

For å unngå bias i modellen fjernet jeg variabelen **sykehusdød** fra trenings- og valideringsdataene, og utviklet en egen klassifikasjonsmodell for å predikere risikoen for sykehusdød. Dette sikrer at vår regresjonsmodell for oppholdslengde kun baserer seg på informasjon tilgjengelig fra dag 1, og ikke påvirkes av utfall som skjer senere i pasientforløpet.

Prosjektet ble også utvidet med utviklingen av en nettside som gjør det mulig for klinikere å legge inn pasientdata og motta en prediksjon på forventet oppholdslengde. Nettsiden inkluderer en advarselsmelding dersom klassifikasjonsmodellen indikerer høy risiko for sykehusdød, noe som kan være verdifullt for planlegging av behandling og ressurser.

6.2 Refleksjon

Prosjektet har demonstrert potensialet for å bruke maskinlæring til å forbedre forståelsen og prediksjonen av sykehusoppholdets lengde for kritisk syke pasienter. Likevel er det flere områder som kan forbedres og videreutvikles:

1. **Dataskjevhet og generalisering**: Modellen presterer best for pasienter med oppholdslengde under 50 dager. For å forbedre prediksjonene for lengre opphold, bør

fremtidige studier inkludere flere data fra pasienter med lange opphold. Dette vil bidra til å redusere skjevheten og øke modellens generaliseringsevne.

2. **Variabelutvalg og funksjonsutvinning**: Selv om vi har identifisert nøkkelvariabler, kan ytterligere funksjonsutvinning og inkludering av flere relevante kliniske variabler forbedre modellens nøyaktighet. For eksempel kan detaljerte behandlingsprotokoller og genetiske data potensielt gi verdifull innsikt.

3. **Imputering av manglende verdier**: Selv om SimpleImputer ga gode resultater, kan andre avanserte imputeringsteknikker eller modeller som er robuste mot manglende data vurderes. Dette kan bidra til å bevare mer informasjon og redusere potensielle imputeringseffekter på modellen. I tillegg vil modellen kunne predikere bedre dersom klassifikasjonsmodellen for sykehusdød forbedres.

4. **Validering i klinisk praksis**: For å sikre at modellen er anvendelig i virkelige kliniske settinger, bør den testes og valideres på data fra andre sykehus og forskjellige pasientpopulasjoner. Dette vil bidra til å identifisere eventuelle begrensninger og justere modellen for bredere anvendelse.

5. **Etiske og personvernmessige hensyn**: Bruk av pasientdata krever nøye vurdering av personvern og etikk. Det er viktig å sikre at all datahåndtering skjer i tråd med gjeldende lover og retningslinjer, og at modellen ikke introduserer bias eller urettferdige prediksjoner basert på sensitive attributter.

6.3 Avsluttende tanker

Prosjektet har gitt verdifull innsikt i hvordan maskinlæring kan brukes til å støtte beslutningstaking i helsesektoren. Jeg har lært mye av å arbeide med dette, og det har vært mye frem og tilbake med ulike metoder å gå frem på. Ved å bygge videre på dette arbeidet, og adressere de identifiserte utfordringene, kan vi utvikle enda mer presise og nyttige verktøy for å forbedre pasientbehandlingen og ressursallokeringen på sykehus. Samarbeid mellom dataforskere, klinikere og etiske eksperter vil være avgjørende for å realisere potensialet i slike modeller på en ansvarlig og effektiv måte. Hvis jeg hadde hatt ubegrenset tid, ville jeg fokusert på å forbedre modellens ytelse for pasienter med lengre opphold, ettersom dette fremstår som modellens nåværende svakhet. Jeg ville eksperimentert med de ulike metodene nevnt tidligere for å finjustere prediksjonene og redusere feil for denne gruppen pasienter.

7 Kilder

Chugani, V. (2024, 8. Juni). Skewness Be Gone: Transformative Tricks for Data Scientists https://machinelearningmastery.com/skewness-be-gone-transformative-tricks-for-data-scientists/

Blaser, N. (2023). Data Science Forelesningsnotater. https://blasern.github.io/data-science-forelesningsnotater/intro.html

Harrel, F. (2023, 14. September) Support 2. https://archive.ics.uci.edu/dataset/880/support2