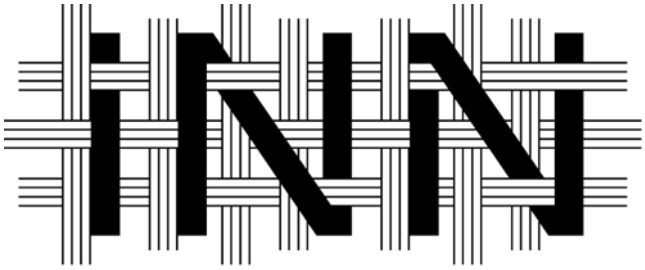


Høgskolen
i Innlandet

Fakultet for økonomi og samfunnsvitenskap

Anders Emil von Krogh

Masteroppgave

Maskinlæring i offentlige byggeprosjekter

Machine learning in public construction projects

Digital ledelse og business analytics

KDBA950

2024

Forord

I august 2020 startet jeg masterutdanningen i digital ledelse og business analytics på deltid ved Høgskolen i Innlandet. Gjennom studiet har jeg opparbeidet meg en stor interesse for datadrevet analyse, spesielt knyttet til predikering og maskinlæring. Denne interessen har vært en av de viktigste årsakene for hvorfor jeg valgte å skrive en masteroppgave om nettopp disse temaene. Som småbarnsfar med full jobb har studiet både bydd på interessante utfordringer og fag, men det har også til tider bare fremstått som et slit uten ende. Masterutdannelsen ved Høgskolen i Innlandet har for meg fremstått som et meget fremtidsrettet studie som er godt synkronisert med utviklingen i næringsliv og offentlig sektor. Jeg opplever derfor at mange av de fagene jeg har hatt gjennom studiet har gitt meg kunnskaper og ferdigheter som er høyst relevante og verdiskapende i jobbsammenheng.

Jeg ønsker å takke min veileder Fikru Kefyalew Alemayehu for all hjelpen jeg har fått, spesielt med tanke på posisjonering og strukturering. I tillegg ønsker jeg å takke prosjekt og utviklingsavdelingen i Forsvarsbygg, jeg setter pris på muligheten dere har gitt meg. Foreldre og svigerforeldre fortjener også en stor takk. Dere har stilt opp både med barnepass og som sparringspartnere, noe som har hjulpet mye i hektiske perioder. Den største takken går til min kone, Tonje. Uten deg hadde jeg ikke klart dette.

Anders Emil von Krogh

Elverum, 01.05.2024

Sammendrag

Maskinlæring og kunstig intelligens generelt er et område i sterk utvikling, og aktualiseres av det enorme potensialet det har for å effektivisere og forbedre eksisterende prosesser, samt skape grunnlaget for nye måter å skape verdi på. I denne oppgaven utforskes maskinlæring i offentlige byggeprosjekter. Gjennom å utforske hvilke variabler som påvirker sluttkostnaden i prosjektene, og i hvilken grad valgte modeller evner å predikere sluttkostnaden i prosjektene. Oppgaven forsøker å belyse hvordan maskinlæring kan bidra til å skape verdi i offentlige byggherreorganisasjoner, med utgangspunkt i 334 prosjekter gjennomført i Prosjekt og utviklingsavdelingen til Forsvarsbygg. Problemstillingen aktualiseres både gjennom den nye strategien for kunstig intelligens i Forsvarssektoren, et generelt lavt omfang knyttet til bruk av kunstig intelligens i offentlig sektor, samt anbefalinger til videre forskning som forskningsprogrammet Concept ved NTNU peker på knyttet til datadrevne analysemetoder for estimering av sluttkostnad i offentlige byggeprosjekter.

Predikering av sluttkostnaden i byggeprosjekter ved bruk av maskinlæring er ikke noe nytt fagområde. Det er derimot ikke testet ut i en norsk kontekst i tilstrekkelig grad. Metodene som benyttes er Recursive Feature Elimination, Regularized Random Forest og LASSO regresjon for å vurdere hvilke prosjektvariabler som påvirker sluttkostnaden, og multippel regresjonsanalyse og artificial neural network for å predikere sluttkostnaden i prosjektene.

Basert på tilgjengelig data peker resultatene fra analysene på at blant variablene tilgjengelig i prosjektets forprosjektfase er det entrepriseform, varighet og gjennomførende seksjon som er de viktigste for å bestemme sluttkostnaden. Fra en er klar til å fatte beslutning om gjennomføringsoppdrag og til prosjektets ferdigstillelse er det variablene knyttet til estimatet som P50, P85 og opprinnelig P50 som er de viktigste. Når det kommer til predikering av sluttkostnad er det artificial neural network som predikerer best med et gjennomsnittlig avvik fra den faktiske sluttkostnaden på 15 prosent i forbindelse med variabler tilgjengelig når beslutning om gjennomføringsoppdrag skal fattes, og 17,5 prosent basert på variabler tilgjengelig i prosjektets forprosjektfase. Sammenlignet med tidligere forskning er ikke dette et godt resultat, men fortsatt innenfor intervallet av resultater fra sammenlignbare studier mellom 2 og 21 prosent.

Oppgaven konkluderer med at maskinlæring er et godt egent verktøy som kan tilføre verdi til offentlige byggherreorganisasjoner. Verdiskapningen kan skje eksempelvis gjennom å avdekke hvilke variabler som er viktige for virksomhetene å jobbe med, eller som verktøy for å estimere sluttkostnad i prosjektets forprosjektfase på en rask og enkel måte. Dette sikrer både et proaktivt og prediktivt fokus. Maskinlæring kan også benyttes som et støtteverktøy i forbindelse med en mer tradisjonell tilnærming til kostnadsestimering både for å tilføre vurderingen av estimert sluttkostnad verdi, eller som en kontroll av estimatene av eksempelvis ledelsen. En viktig forutsetning for å lykkes med maskinlæring er knyttet til data, både med tanke på mengde, kvalitet og mangfold. Denne studien ser derfor behovet for å jobbe videre med data som en av de viktigste suksessfaktorene for å ta i bruk maskinlæring i offentlige byggeprosjekter. Samtidig viser studien at det likevel er mulig å ta i bruk maskinlæring selv om data ikke er av den kvalitet eller det omfang som er ønskelig.

Abstract

Machine learning and artificial intelligence in general are rapidly developing fields, driven by the great potential they hold to streamline and enhance existing processes, as well as lay the groundwork for new ways of creating value. This study explores machine learning applications in public construction projects, aiming to gain a better understanding of the variables influencing the final cost of these projects and to what extent selected models can predict project costs. The study seeks to shed a light on how machine learning can contribute to value creation in public construction organizations, based on an analysis of 334 projects conducted in the project and development department of Forsvarsbygg.

The issue becomes relevant through the new strategy for artificial intelligence in the defense sector, the generally low adoption of artificial intelligence in the public sector, and recommendations for further research outlined in the Concept research program at NTNU related to data-driven analysis methods for estimating final costs in public construction projects.

Predicting the final cost of construction projects using machine learning is not a new field of study. However, it has not been extensively tested in a Norwegian context. The methods used include Recursive Feature Elimination, Regularized Random Forest, and LASSO

regression to assess which project variables affect the final cost, and multiple regression analysis and artificial neural networks to predict the final cost in the projects.

Based on available data, the results from the analysis indicate that among the variables available in the project's preliminary phase, the type of contract, duration, and executing section are the most significant factors in determining the final cost. From the decision-making phase to project completion, variables related to estimates such as P50, P85, and original P50 are the most important. Regarding cost prediction, artificial neural networks yield the best results with an average deviation from the actual final project cost of 15 percent based on variables available when the decision to go ahead with the project is made, and 17.5 percent based on variables available in the project's preliminary phase. While not an outstanding result compared to previous research, this still falls within the range of results from comparable studies, ranging from 2 to 21 percent.

The study concludes that machine learning is an important tool that can add value to public construction organizations. For instance, by identifying important variables for organizations to focus on or as a tool for estimating project costs quickly and easily in the preliminary phase. Machine learning can also be used as a supplementary tool alongside a more traditional approach to cost estimation, either to enhance the value of the estimated final cost assessment or as a way to check the estimates by management, for example. A crucial prerequisite for successful machine learning implementation is data, and its quantity, quality, and diversity. Therefore, this study underscores the need to further work on data as one of the key success factors for adopting machine learning in public construction projects. At the same time, the study demonstrates that it is still possible to employ machine learning even if the data is not of the desired quality or diversity.

Innhold

Begreper og forkortelser

Tabell 1: Forklaring av forkortelser

Forkortelse	Forklaring
EBA	Eiendom, bygg og anlegg
GO	Gjennomføringsoppdrag
KVU	Konseptvalgutredning
SSD	Sentralt styringsdokument
LTP	Forsvarssektorens langtidsplan
KS1	Første eksterne kvalitetssikring
KS2	Andre eksterne kvalitetssikring
ODG	Oppdragsgiver
PE	Prosjekteier
PA	Prosjektansvarlig
BA	Brukeransvarlig
PRINSIX	Forsvarssektorens prosjektmodell
P50	Prosjektets styringsramme
P85	Prosjektets kostnadsramme
KI	Kunstig intelligens
R	Programvare og programmeringsspråk
RFE	Recursive Feature Elimination
RFF	Regularized Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
MRA	Multippel regresjonsanalyse
ANN/NN	Artificial neural network/ nevrale nettverk
MAPE	Mean absolute percentage error
MSE	Mean squared error
RMSE	Root Mean Square Error

1. Innledning

1.1 Bakgrunn

Økonomifaget har lenge stått ovenfor en utfordring knyttet til relevans. En snakker gjerne om at økonomistyringen ikke leverer verdi for virksomheten, og at beslutningsstøtten kommer for sent til at ledelsen kan nyttiggjøre seg av den (Bjørnenak, 2010). Beslutninger fattes derfor ofte basert på utdatert informasjon, heuristikker og magefølelse. Økonomistyringen er dermed mer deskriptiv og diagnostisk, fremfor prediktiv og anbefalende. Data kommer altså ikke tidlig nok gjennom informasjonsverdikjeden, og dermed mister den noe av nytteverdien sin.

I regjeringens forslag til statsbudsjett 2024 er Forsvarsdepartementet tildelt 4,7 milliarder til investering i nybygg og nyanlegg (Forsvarsdepartementet, 2023). Dette er budsjettrammen som Prosjekt og utviklingsavdelingen til Forsvarsbygg foreløpig har til disposisjon i løpet av 2024. Samtidig er det klart at et byggeprosjekt kan strekke seg over flere år. Økonomistyring, porteføljestyring og ikke minst prosjektstyring vil derfor naturlig nok både handle om å estimere og prognostisere sluttkostnaden til prosjektene, men også hvilket kostnadspådrag prosjektet vil ha hvert enkelt år. Det er mange variabler som påvirker kostnadspådraget over tid i et byggeprosjekt. Uten verktøy for å ta høyde for disse variablene kan prognosene potensielt bli forbundet med høy risiko for kostnadsoverskridelse eller forskyvninger i tid.

I en pressemelding 5. april 2024 sier forsvarsminister Bjørn Arild Gram (Sp) at det skal «*gjennomføres en kraftfull satsing på eiendom, bygg og anlegg som understøtter styrkingen av Forsvarets operative evne i denne langtidsplanen*». På listen over tiltak som regjeringen vil at skal iverksettes står blant annet at en skal «*utarbeide en helhetlig plan for utbygging av eiendom, bygg og anlegg til rett tid, sted og kost. Denne planen skal sikre bedre synkronisering med både materiellplaner og personellopptrapping*» (Forsvarsdepartementet, 2024).

Finansdepartementet finansierer det uavhengige forskningsprogrammet Concept ved NTNU. De forsker og «*utvikler kunnskap som sikrer bedre konseptvalg, ressursutnytting og effekt av store statlige investeringer*» (Consept, U.Å.). I sin rapport om «*kostnadsestimering i tidlegfase av store offentlege prosjekt*» (Larsen, et al., 2023) fremhever de viktigheten av å «*utvikle ny kunnskap om dei framtidsretta metodane for kostnadsestimering. Ikkje minst gjeld dette korleis data skal samlast, systematiserast og nyttast i estimering og analyse*»

(Larsen, et al., 2023). Videre har Mæhlen & Bekkevold (2022) gjennomført en studie på vegne av Concept knyttet til «*Datadrevet usikkerhetsanalyse i byggeprosjekter».* Studien gjennomfører blant annet en statistisk analyse i form av multippel regresjonsanalyse av ulike variabler fra tidligere gjennomførte byggeprosjekter. Der anbefaler de videre arbeid knyttet til å:

«*Studere uavhengige variabler sin påvirkning på forklaringsgrad for relativt avvik: I denne studien har vi analysert sammenhengen mellom en rekke prosjektkarakteristikker og relativt avvik mellom estimat og sluttkostnad.» … «Regresjonsmodellen mener altså at de inkluderte forklaringsvariablene begrunner halvparten av det relative avviket. Men hva forklarer resten? Vi kan i dag ikke vite hvilke andre prosjektkarakteristikker vi gjerne skulle hatt med i datasettet for å øke forklaringsgraden. Derfor ville det vært svært nyttig med en studie som ser på hvilke prosjektkarakteristikker som øker forklaringsgraden mest. Det er disse karakteristikkene det vil være mest hensiktsmessig å gjøre analyser på, og derfor også dokumentere.»* (Mæhlen & Bekkevold, 2022)

1.2 Formål

Jeg ønsker derfor å bygge videre på dette arbeidet gjennom å studere hvilke prosjektkarakteristikker som statistisk sett påvirker og bestemmer sluttkostnaden i statlige byggeprosjekter. Jeg ønsker også å vurdere i hvilken grad sluttkostnaden lar seg predikere ved hjelp av disse prosjektkarakteristikkene.

1.3 Problemstilling

Basert på anbefaling til videre forskning har oppgaven følgende problemstilling:

I hvilken grad kan maskinlæring bidra til i å skape verdi i offentlige byggherreorganisasjoner?

1.4 Forskningsspørsmål

For å svare på problemstillingen tar oppgaven utgangspunkt i to forskningsspørsmål:

- Hvilke prosjektvariabler bestemmer sluttkostnad en i offentlige byggeprosjekter?
- I hvilken grad kan prediktiv analyse bidra til presise beslutninger i offentlige byggeprosjekter?

1.5 Omfang og avgrensning

Oppgaven tar for seg et relativt vidt fagområde i skjæringen mellom byggeprosjekter og kunstig intelligens. Det har derfor vært nødvendig å avgrense oppgaven noe slik at omfanget skal være overkommelig. Oppgaven fokuserer derfor utelukkende på prediktiv analyse i byggeprosjekter hvor det skal bygges et bygg eller anlegg. Vei og jernbaneprosjekter har derfor ikke blitt inkludert i litteraturstudien. I tillegg finnes det veldig mange ulike rammeverk, modeller og metoder innenfor maskinlæringsdelen av kunstig intelligens. Oppgaven går derfor ikke i dybden på hvorfor valgte modeller benyttes.

1.6 Forsvarsbygg

Hele datagrunnlaget på totalt 334 prosjekter er hentet fra og gjennomført i Prosjekt og utviklingsavdelingen til Forsvarsbygg. «*Forsvarsbygg er Norges største offentlige eiendomsaktør*» (Forsvarsbygg, U.Å.). Om lag 13 000 bygg og anlegg forvaltes av Forsvarsbygg. Nye investeringsprosjekter som Forsvarsbygg skal iverksette for å nå de overordnede målsettingene i forsvarssektoren fremkommer i investeringsplanen som Forsvarsdepartementet har ansvar for (Forsvarsdepartementet, 2019). Kort oppsummert handler oppgavene til forsvarsbygg om å «etablere, opprettholde og gjenopprette forsvarssektorens eiendom, bygg og anlegg i fred, krise og krig» (Forsvarsbygg, U.Å.).

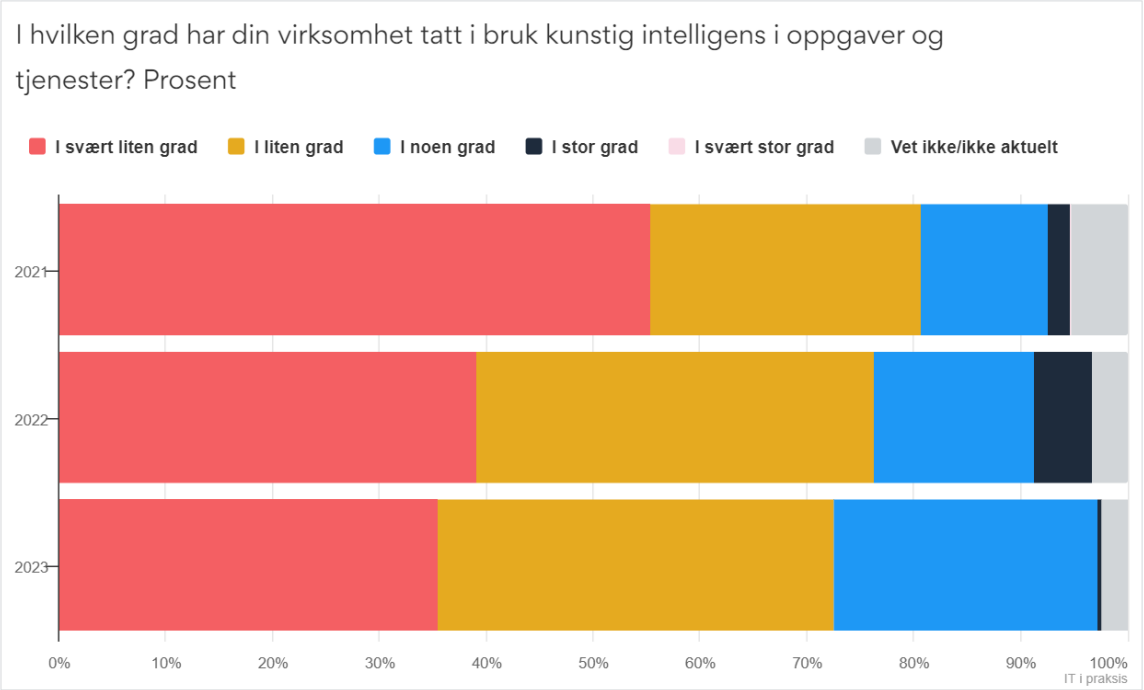
Naturlig nok innebærer dette mange forskjellige typer byggeprosjekter. Mange som også er særegne for forsvarssektoren. Eksempler kan være Finnmark landforsvar, Evenes flystasjon, nye radaranlegg over hele landet og bygg og anlegg for nye ubåter (Forsvarsbygg, U.Å.). Disse fire store prosjektene strekker seg altså fra å bygge ut en stor leir i Finnmark, via utbygging av en flyplass og bygging av radaranlegg, til opprettelse av et nytt bygg som ivaretar de nye ubåtenes behov. Dette spennet i ulike typer byggeprosjekter kan potensielt by på utfordringer når det kommer til predikering av sluttkostnaden i prosjektene. Om mange av prosjektene er veldig forskjellige og ofte engangstilfeller, så kan det være en risiko for at de ikke kan sammenlignes med hverandre gjennom statistisk analyse.

Figur 1: Forsvarssektoren, (Forsvarsbygg, U.Å.)

1.7 Kunstig intelligens i offentlig sektor

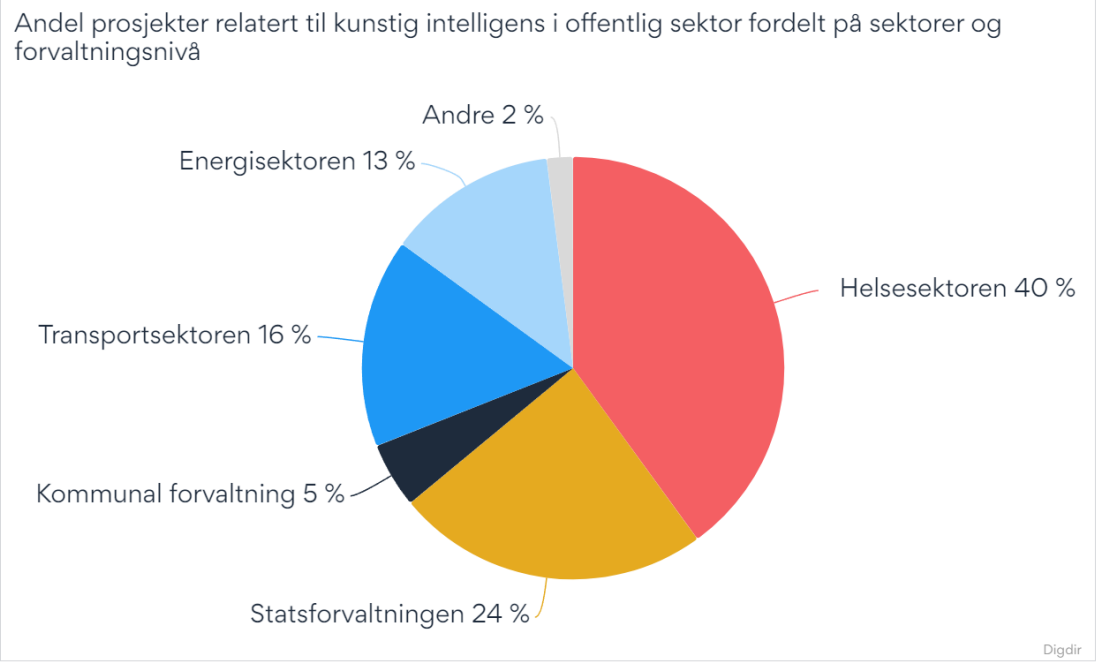
Denne oppgaven søker å utnytte kunstig intelligens (KI) i form av maskinlæring for å predikere sluttkostnaden i offentlige byggeprosjekter. Derfor gjør oppgaven en kort redegjørelse for, og vurdering av bruken av kunstig intelligens i offentlig sektor. Digitaliseringsdirektoratet (U.Å.) definerer kunstig intelligens som at «*kunstig intelligente systemer utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål. Enkelte KI-systemer kan også tilpasse seg gjennom å analysere og ta hensyn til hvordan tidligere handlinger har påvirket omgivelsene*». I kort altså å utføre en handling basert på tolkning av data for å nå et mål. Dette er en ganske vid definisjon som passer til mange forskjellige tilnærminger knyttet til bruken av «algoritmer, maskinlæring, modeller og statistiske metoder» (Digitaliseringsdirektoratet, U.Å.).

Digitaliseringsdirektoratet (2023) har i 2021, 2022 og 2023 samlet inn informasjon om bruken av KI i offentlig sektor. Et av hovedfunnene i undersøkelsen er at «*den faktiske bruken av kunstig intelligens i oppgaver og tjenester i offentlige virksomheter er relativt begrenset»* (Digitaliseringsdirektoratet, 2023).



Figur 2: Bruk av KI i offentlig sektor, (Digitaliseringsdirektoratet, 2023)

Som vi ser ut ifra figur 2 ligger bruken av KI primært på «*I svært liten grad*» og «*I liten grad*» i alle år. Det ser derimot ut til å være en positiv utvikling i form av at en økende andel responderer med «*I noen grad*». Det er samtidig ingen som mener at virksomheten bruker KI «*I svært stor grad»,* og svært få som hevder virksomheten bruker KI «*I stor grad*». En refleksjon knyttet til hvorfor det er færre som svarer «*I stor grad»* i 2023 sammenlignet med 2021 og 2022 er at det kan tenkes at den digitale modenheten og kunnskapen om KI i offentlig sektor er på vei opp. Når en får forståelse for hvilke muligheter KI representerer så kan det tenkes at en også innser at det fortsatt finnes en del uutnyttet potensiale i virksomheten.



Figur 3: KI prosjekter i offentlig sektor, (Digitaliseringsdirektoratet, 2023)

Digitaliseringsdirektoratet (2023) påpeker videre at det er svært få av prosjektene og aktivitetene som er tatt i bruk i virksomhetene. «*De fleste aktivitetene handlet om utprøving og utvikling*» (Digitaliseringsdirektoratet, 2023). Dette kan oppfattes som naturlig, i den forstand at det kan være et første steg på veien mot å integrere KI i den operative driften av en virksomhet. Samtidig kan dette også indikere at KI foreløpig ikke fører til nevneverdig verdiskapning i offentlig sektor. Forsvarssektoren er heller ikke representert i oversikten over hvilke sektorer som har iverksatt KI prosjekter (figur 3), med mindre sektoren er en del av «*Andre».*

Oppsummert kan det se ut som at bruken av KI foreløpig er relativt lav i offentlig sektor, og også tilsynelatende i forsvarssektoren. Samtidig ser en konturene av en økende trend. Denne studien kan således bidra til denne trenden ved å synliggjøre noen av mulighetene knyttet til KI.

1.8 Strategi for kunstig intelligens i forsvarssektoren

For å komme i gang med bruken av KI utarbeidet Forsvarsdepartementet i 2023 en strategi for kunstig intelligens. Den overordnede ambisjonen i strategien er at «*Forsvarssektoren skal utnytte potensialet i kunstig intelligens for å bidra til å fremme Norges sikkerhets- og forsvarspolitiske mål. Dette skal sikres gjennom at forsvarssektoren skal identifisere, utvikle,*

implementere og anvende kunstig intelligens på en ansvarlig måte» (Forsvarsdepartementet, 2023). Den økende trenden en ser i offentlig sektor kan derfor også sies å gjelde for forsvarssektoren.

Strategien legger derimot opp til at KI skal tas i bruk på en metodisk og helthetlig måte. Det vil altså ta tid før en kan starte å høste fruktene av denne strategien. På den andre siden sikrer en at en ikke implementerer systemer eller prosesser som potensielt kan utgjøre en trussel mot de sikkerhets- og forsvarspolitiske målene.

Blant de uttalte ambisjonene for strategien er det et par punkter som går igjen. Innledningsvis fokuseres det mye på viktigheten av data. Dette fremstår som et naturlig startpunkt for strategien. Uten data, både med tanke på mengde, kvalitet og aktualitet, vil ikke KI kunne utnyttes. I strategien står det blant annet at «*Forsvarssektorens etater skal søke å fange alle data av verdi som produseres gjennom virksomhetene. Verdivurderingen av data må inkludere nåværende og potensiell fremtidig nytte»* (Forsvarsdepartementet, 2023). Videre legges det vekt på kompetanse, evne til læring, kultur for forbedring og erfaringsutveksling. Data og kompetanse kan ses på som to av de viktigste forutsetningene som må være på plass for å kunne få effekt av teknologien knyttet til KI. Utover dette nevnes også samarbeid, både internasjonalt og med ledende miljøer i Norge, samt oppdatering av prosjektmodell, og etablering av anskaffelses- og investeringsstrategi (Forsvarsdepartementet, 2023).

Basert på ambisjonene i strategien kan en vurdere situasjonen dit at forsvarssektoren er på god vei til å ta i bruk KI som et viktig element i å skape forsvarsevne. Samtidig er sektoren helt i startgropen når det gjelder implementering. Det vil kunne ta tid å bygge en god datastrategi som understøtter bruken av KI, og som er godt forankret i de ulike leddene i virksomhetene. Kompetanse vil potensielt være noe lettere å få på plass, da denne kompetansen allerede eksisterer i samfunnet. Denne oppgaven kan således fungere som en test, eller pilot for Forsvarsbygg. Som kan belyse både potensiale og utfordringer.

2. Teori

2.1 Innledning av teorikapittel

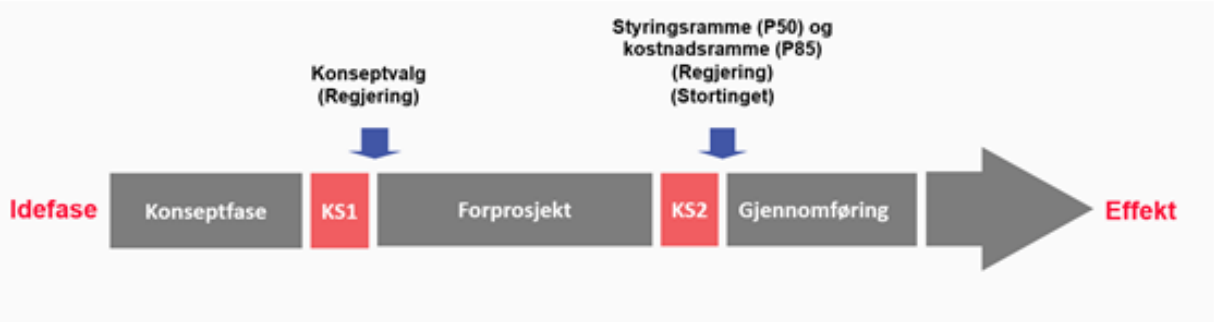
For å skape et godt fundament for forskningsspørsmålene tar teoridelen av oppgaven først for seg en gjennomgang av offentlige byggeprosjekter. Oppgaven tar videre for seg prediktiv analyse og maskinlæring. Teoridelen søker videre å kombinere disse to temaene i en litteraturgjennomgang som utforsker tidligere forskning knyttet til predikering av kostnader i byggeprosjekter. Dette vil struktureres ved å se nærmere på tidligere forskning på området. Mer spesifikt årsaker til kostnadsavvik i byggeprosjekter, ulike modeller brukt for predikering i byggeprosjekter, variabler som påvirker byggekostnader, og resultater tidligere forskning har oppnådd med sine modeller, og hvordan dette ble målt. Avslutningsvis vil tidligere forskning kobles mot forskningsspørsmålene for å belyse oppgavens bidrag.

2.2 Offentlige byggeprosjekter

2.2.1 Statens prosjektmodell

Alle offentlige investeringsprosjekter med samlet estimert kostnadsramme over 1 milliard kroner skal gjennomføres i henhold til statens prosjektmodell (Regjeringen, U.Å.). I 1997 besluttet regjeringen å iverksette de første stegene mot å etablere statens prosjektmodell. «*Bakgrunnen for dette var en lang rekke negative erfaringer med kostnadsoverskridelser, forsinkelser og manglende realisering av nytteeffekter i offentlige investeringsprosjekt»* (Regjeringen, U.Å.). *«Målet er å unngå feilinvesteringer og holde god kontroll med kostnader og nytte gjennom planlegging og gjennomføring av prosjektene og på den måten sørge for en mest mulig effektiv bruk av fellesskapets ressurser»* (Regjeringen, U.Å.)

Modellen stiller blant annet krav til ekstern kvalitetssikring både i forbindelse med konseptvalg (KS1) og kostnadsanslag og styringsunderlag (KS2). «*Hensikten med kontrollpunktene KS1 og KS2 er at informasjonen og analysene som allerede ligger i prosjektet skal gjennomgås av en uavhengig tredjepart. I KS1 kvalitetssikres konseptvalgutredningen, som gir en anbefaling om hvilket konsept eller alternativ som eventuelt skal videreføres i forprosjektfasen. I KS2 kvalitetssikres det sentrale styringsdokument og det vurderes om planleggingen og kostnadene som er utarbeidet i forprosjektet for det valgte alternativet er realistisk»* (Regjeringen, U.Å.).



Figur 4: Statens prosjektmodell, (Regjeringen, U.Å.)

Modellen legger også opp til at prosjektene skal følge de samme fastsatte fasene:

1. Idefase
2. Konseptfase
3. Forprosjekt
4. Gjennomføring

(Regjeringen, U.Å.)

Idefasen handler i grovt om at det har eller vil oppstå et behov som staten må vurdere om skal løses gjennom et prosjekt. Idefasen ender opp i et mandat for å fortsette videre inn i konseptfasen gitt at prosjektet er av en slik karakter at statens prosjektmodell skal eller bør benyttes. (Regjeringen, U.Å.)

Konseptfasen konkretiserer ideen ytterligere gjennom blant annet å beskrive problemet, samfunnets fremtidige behov og tiltaket eller prosjektets mål. En kan selvfølgelig tilnærme seg en utfordring på flere forskjellige måter. Konseptfasen handler derfor om å være åpen for alle mulige løsninger og sette disse opp mot hverandre gjennom eksempelvis samfunnsøkonomiske analyser. Konseptfasens hovedleveranse er en konseptvalgutredning (KVU). Her skal en anbefaling om hvilket av konseptene som bør iverksettes fremkomme, i tillegg til viktige forutsetninger for å lykkes i den videre gjennomføringen av prosjektet. Ifølge modellen står prosjektet nå overfor sin første store milepæl, KS1. Når ekstern kvalitetssikring av KVU er gjennomført vil regjeringen beslutte hvorvidt prosjektet fortsetter inn i forprosjektfasen eller ikke (Regjeringen, U.Å.).

Forprosjektfasen handler om å konkretisere besluttet konsept. Forprosjektfasen ender opp i ferdigstillelsen av prosjektets sentrale styringsdokument (SSD). Styringsdokumentet skal

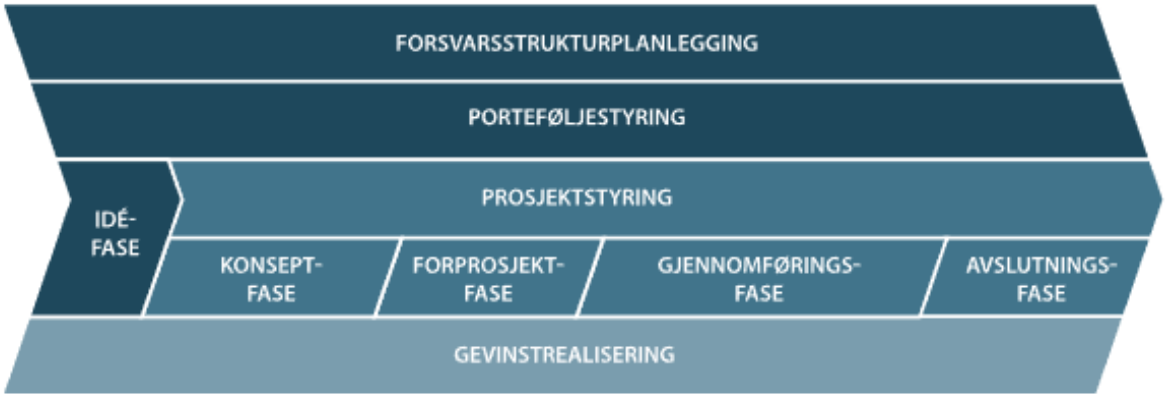
blant annet ta for seg hvordan prosjektet skal gjennomføres, samt å estimere mer detaljert og vurdere usikkerheten i estimatet. I tillegg skal SSD «*beskrive hvordan prosjektet skal styres for å ha kontroll på kostnadene og nå målene som er satt, og vurdere hva slags kontrakter som gir mulige leverandører riktige insentiver til å levere det prosjektet trenger»* (Regjeringen, U.Å.). Når forprosjektfasen er over, skal KS2 gjennomføres. Prosjektet er nå klart for at den endelige beslutningen om gjennomføring og kostnadsramme kan fattes i Stortinget. Dette markerer også slutten på prosjektets tidligfase.

Med tanke på oppgavens forskningsspørsmål er det viktig å merke seg forskjellen på tidligfase og gjennomføringsfase. I tidligfase har ikke beslutningen om å gjennomføre prosjektet blitt tatt enda. Om en har som mål å predikere sluttkostnaden til et byggeprosjekt vil det være naturlig gjøre det før den endelig beslutningen om å iverksette bygging fattes. Eksempelvis som et ledd i å fastsette prosjektets kostnadsramme og usikkerhetsavsetningens størrelse. Dette innebærer at en er nødt til å ha et bevisst forhold til hvilke variabler som er tilgjengelig i prosjektets tidligfase. Variabler som først blir kjent på et senere tidspunkt vil ikke kunne bidra til bedre beslutningsstøtte i forbindelse med KS2. Det sagt, så kan det også være interessant å ha kjennskap til hvordan variabler som først er tilgjengelig senere i prosjektets gjennomføring påvirker sluttkostnaden. Tiltak for å øke de gunstige variablene og redusere de ugunstige variablene kan da jobbes med i organisasjonen.

2.2.2 Forsvarssektorens prosjektmodell

Forsvarssektoren har bygget videre på statens prosjektmodell. Forsvarssektorens prosjektmodell heter Prinsix, som også baserer seg på retningslinjer for investeringer i forsvarssektoren (Forsvarsdepartementet, 2019). Vi finner eksempelvis igjen de samme fasene, men modellen er bygget noe ut. Det er eksempelvis lagt til en avslutningsfase, og modellen fokuserer mer på gevinstrealisering gjennom hele prosjektet. Det er også et fokus på porteføljestyring, noe som er viktig for å se det enkelte prosjekt og behov i lys av andre prosjekter og det totale behovet i sektoren. Et annet aspekt som skiller Prinsix fra statens prosjektmodell er at modellen som hovedregel anvendes i alle investeringsprosjekter uavhengig av størrelse. Modellen fokuserer derfor ikke på KS1 og KS2, selv om disse vil være en del av modellen når prosjektets størrelse tilsier det (Forsvarsmateriell, U.Å.). En kan tolke det dit at formålet til Prinsix er noe større enn det statens prosjektmodell legger opp til.

Hovedsakelig oppstår denne utvidelsen som følge av at en i større grad har behov for å se det samlede behovet i forsvarssektoren.

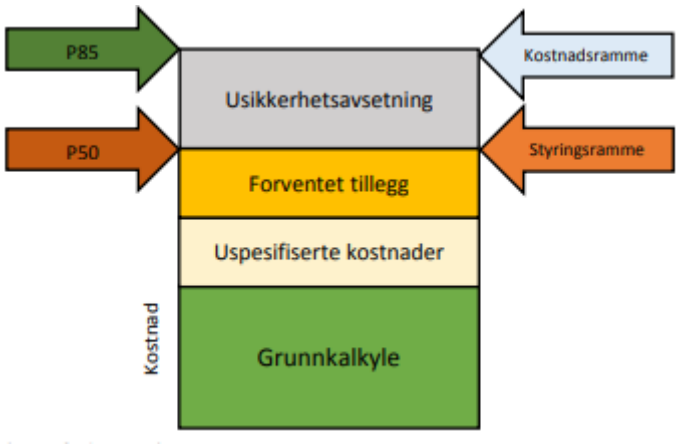


Figur 5: Forsvarssektorens prosjektmodell, (Forsvarsmateriell, U.Å.)

Videre fokuserer Prinsix mer på roller og ansvar enn hva statens prosjektmodell gjør. «*Materiell- og EBA-investeringer i forsvarssektoren initieres med en prosjektidé eller gjennom langtidsplanarbeidet. Deretter gjennomføres normalt to faser med planlegging (konsept- og forprosjektfasen) før Forsvarsmateriell og Forsvarsbygg (som PA) mottar et gjennomføringsoppdrag (GO) fra prosjekteier (PE*)» (Forsvarsmateriell, U.Å.). I denne sammenhengen er PA prosjektansvarlig. Det er altså PA som gjennomfører prosjektet, men PA må først få GO fra PE. De ulike rollene i prosjektet indikerer at det er mange ulike interessenter involvert i de beslutninger som skal fattes.

Det er totalt 4 hovedroller som synliggjøres i Prinsix:

- Oppdragsgiver (ODG), Forsvarsdepartementet
- Prosjekteier (PE), Forsvaret
- Brukeransvarlig (BA), Forsvaret
- Prosjektansvarlig (PA), Forsvarsmateriell og Forsvarsbygg (Forsvarsmateriell, U.Å.)



Figur 6: Prosjektets kostnadskalkyle, (Forsvarsdepartementet, 2019)

Videre forklarer Forsvarsdepartementet (2019) grunnstrukturen i kalkylen for investeringen (figur 6). Som vi ser i kalkylen, skilles det på kostnadsramme (P85) og styringsramme (P50). Med P85 menes at det er estimert med 85% sannsynlighet at kostnaden i prosjektet vil være lik eller lavere enn dette beløpet. Tilsvarende gjelder for P50, bare med 50% sannsynlighet. PA får innledningsvis kun disponere P50, og må søke om å utløse P85 hvis det er behov for det. I tillegg skilles det på prosjektene ut ifra størrelsen på kostnadsrammen. Hvis kostnadsrammen er over 200 mill. kroner er prosjektet i kategori 1, og hvis den er under så er det et kategori 2 prosjekt. Dermed får vi i praksis 3 ulike kategorier med prosjekter. Disse ulike kategoriene er igjen ulike når det kommer til hvilke roller som har beslutningsmyndighet. Beslutningsmyndigheten gjelder hovedsakelig KVU, SSD, GO og utløsning av P85. Omfangsendringer og andre endringer og avvik i oppdraget behandles noe annerledes. Spørsmålet koker ned til hvorvidt endringen er innenfor de rammer ODG har gitt til PE eller ikke. Utover kostnader kan slike endringer og avvik være knyttet til tid og kvalitet. Altså hvorvidt prosjektet ikke kan gjennomføres til estimert tid, eller om selve oppdraget og dets mål, ambisjon, ytelse eller funksjon skal endres. Endringer og avvik utover rammene til PE besluttes av ODG.

Tabell 2: Kategorier av investeringsprosjekter

Kostnadsramme	Benevning	Beslutning
P85 over 1 milliard kroner	Kategori 1, med krav til KS1 og KS2	Stortinget
P85 over 200 mill. kroner, men under 1 milliard kroner	Kategori 1	Stortinget
P85 under 200 mill. kroner	Kategori 2	Prosjekteier

Prosjektmodellen til forsvarssektoren innebærer naturlig nok mange prosesser og involverte interessenter. Når det kommer til predikering av sluttkostnaden i prosjektet vil det derfor potensielt ha mye å si om prosjektet er et kategori 2 prosjekt eller et kategori 1 prosjekt. På den ene siden kan en tenke seg at et kategori 1 prosjekt krever mer for å komme frem til endelig beslutning om investering, og at de derfor kanskje er mer kvalitetsikret før gjennomføring iverksettes. På den andre siden vil kategori 1 prosjekter ha en lengre beslutningssløyfe i viktige avgjørelser. Dette vil kunne føre til at prosjekter midlertidig stanser i påvente av beslutning i Stortinget. Det motsatte vil potensielt gjelde for kategori 2 prosjekter. Begge disse momentene vil kunne ha betydning for predikering av sluttkostnaden i prosjektene.

2.2.3 Estimering av kostnader i byggeprosjekter

Når vi nå har sett på de relevante modellene for offentlige byggeprosjekter er det nærliggende å utforske hvordan P50 og P85 fremkommer. Oppgaven vil derfor dykke dypere ned i prosjektets kostnadskalkyle (figur 6). Som det fremkommer av figur 6 baserer kostnadsrammen seg på en grunnkalkyle hvor uspesifiserte kostnader, forventede tillegg og usikkerhetsavsetning er lagt til.

Grunnkalkylen bygger ifølge Mæhlen & Bekkevold (2022) som oftest på den norske standarden NS 3451, også kalt bygningsdelstabellen. «*NS 3451 er en standard innenfor klassifikasjonssystemer for BAE-sektoren i Norge. Den fastlegger inndeling i bygnings- og installasjonsdeler for systematisering, klassifisering og koding av informasjon som omfatter de fysiske delene av bygningen og de tilhørende uteområder*» (Standard Norge, U.Å.).

Tabell 3: Bygningsdelstabellen, (Mæhlen & Bekkevold, 2022)

Post	Navn	Beskrivelse
1	Felleskostnader	Entreprenørens rigg og drift av byggeplass, kontroll og dokumentasjon.
2	Bygging	Konstruktive tiltak i bygget, deriblant grunn og fundamenter, bæresystem, vegger, dekker, tak, fast inventar og trapper / balkonger.
3	VVS	Varme, ventilasjon og sanitærinstallasjoner.
4	Elkraft	Basisinstallasjon for elkraft, høyspent og lavspent forsyning, lys, elvarme og reservekraft.
5	Tele og automatisering	Integrert kommunikasjon, alarm og signal, lyd og bilde, instrumentering og automatisering.
6	Andre installasjoner	Person- og varetransport, avfall og støvsuging, sceneteknisk utstyr og andre tekniske installasjoner.
7	Utenomhus	Tiltak utendørs, utenfor selve bygget. Det omfatter terrengbearbeiding, parker og hager, veier og plasser og tekniske installasjoner utendørs.
8	Generelle kostnader	Byggherrens prosjektering og administrasjon, byggherreombud og byggeleder i gjennomføringsfasen.
9	Spesielle kostnader	Løst inventar og utstyr, tomteakkvisisjon, finansiering, kapitalkostnader og kunstnerisk utsmykning.
10	Merverdiavgift	Merverdiavgift

Om vi tar for oss bygningsdelstabellen kan vi se at post 1 til 6 er knyttet til selve bygget. Summen av disse postene omtales derfor gjerne som «*huskostnaden*». Om vi videre legger til post 7, som tar inn over seg alt annet enn bygget på tomten, ender vi opp med «*entreprisekostnaden*». Entreprisekostnaden er kostnaden til den delen av prosjektet som byggherren får en eller flere entreprenører til å gjennomføre gjennom kontrakter. Post 8 kostnadene er byggherrens kostnader til eksempelvis prosjektering, prosjektledelse, prosjektadministrasjon, innleie av byggeledere og reisekostnader m.m. Om vi deretter

inkluderer post 9 og 10 ender vi opp med grunnkalkylen til prosjektet (Mæhlen & Bekkevold, 2022).

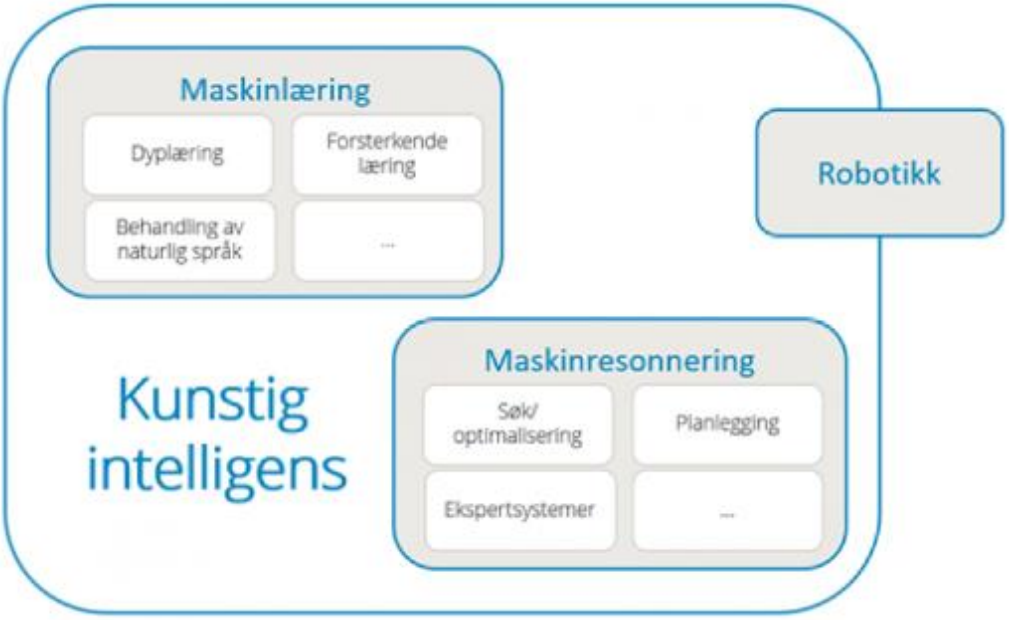
Det er flere måter å tilnærme seg estimeringen av disse postene. Ofte handler variasjonen om på hvilket detaljeringsnivå en legger kalkylen. Kanskje den enkleste formen er å basere estimatene på antall kvadratmeter med støtte fra eksempelvis tidligere gjennomførte lignende prosjekter, eller «*norsk prisbok*» (Mæhlen & Bekkevold, 2022). Norsk prisbok «*er en oppdatert prisdatabase og inneholder bred og mangfoldig prisinformasjon vedrørende kostnader for et byggeprosjekt*» m.m. (Norsk Prisbok, U.Å.). En annen tilnærming kan være å estimere basert på prisen på ulike komponenter. Dette omtales ofte som en tilnærming fra bunnen og opp. Byggherren setter som regel ut estimeringen av post 1-7 til virksomheter som spesialiserer seg på prosjektering av byggeprosjekter, mens de estimerer post 8 selv (Mæhlen & Bekkevold, 2022).

Nesten uansett hvor mye erfaring en har med estimering av byggekostnader kan en ikke endre på det faktum at det er en viss tilstedeværelse av usikkerhet og risiko i prosjektets forprosjektfase. Det er her uspesifiserte kostnader, forventet tillegg og usikkerhetsavsetning kommer inn i kalkylen. Mæhlen & Bekkevold (2022) definerer uspesifiserte kostnader som «*Kostnader man av erfaring vet vil komme, men som ikke er kartlagt pga. manglende detaljeringsnivå*». Forventede tillegg og usikkerhetsavsetning er derimot av en mer uforutsett og uspesifisert karakter. Lowe, Emsley, & Harding (2006) peker eksempelvis på grunnforhold som et godt eksempel på en faktor som kan medføre uforutsette kostnader. Andre eksempler kan være at pris- og konkurransesituasjonen i markedet fører til at tilbudene på kontraktene er høyere enn estimatet. Forventede tillegg er prognosen av alle disse uforutsette faktorene. Målsetningen er å komme fram til P50, P85 vil deretter settes. «*Det finnes flere metoder for å utarbeide usikkerhetsanalyser. Et fellestrekk i de fleste er at estimatusikkerhet og usikkerhetsfaktorer defineres på bakgrunn av subjektive vurderinger*»… «*Estimatusikkerheten bestemmes i en gruppeprosess hvor ressurspersoner i prosjektet er samlet. Når betydningen av de identifiserte indre- og ytre påvirkningskreftene skal defineres, gjøres dette med utgangspunkt i deltakernes kunnskap og erfaring*» (Mæhlen & Bekkevold, 2022).

For å kunne predikere sluttkostnaden er det viktig å ha klart for seg kalkylen med tilhørende oppbygging. Forskjellen mellom P50 og P85 kan eksempelvis fortelle oss noe om den estimerte graden av usikkerhet i prosjektet. Størrelsen på de generelle kostnadene og mer spesifikt på tidligfasekostnadene forteller oss noe om hva slags ressurspådrag det har vært behov for fra byggherrens side. Videre kan det tenkes at prediktiv analyse kan fungere som et godt bidrag til den ellers subjektive og erfaringsbaserte usikkerhetsanalysen Mæhlen & Bekkevold (2022) presenterer.

2.3 Prediktiv analyse, kunstig intelligens og maskinlæring

For å snevre inn begrepet KI, og gjøre analysen i studien mer forståelig tar oppgaven en kort gjennomgang av dette store temaet. KI kan være mange forskjellige ting fra chatboter til bildegjenkjenning til predikering av ulike hendelser som været eller oljeprisen m.m. KI er altså et veldig vidt begrep som kan brukes om mange forskjellige metoder for å nyttiggjøre seg av data. Disse ulike typene KI vil også variere stort i kompleksitet. Denne kompleksiteten muliggjør eksempelvis bruken av big data som gjerne karakteriseres ved at dataen har stort volum, høy grad av variasjon, høy hastighet og god kvalitet (Abbasi, Sarker, & Chiang, 2016). Kompleksiteten kommer derimot med en pris. Desto mer avansert modellene er desto vanskeligere blir det å tolke og forstå modellen. Hvis målet med analysen er, å forstå sammenhengen mellom en avhengig og flere uavhengig variabler vil de mest kompliserte modellene gjøre det vanlig å se sammenhengen selv om de kanskje klarer å predikere den avhengige variabelen bedre (James, Witten, Hastie, & Tibshirani, 2021). Dette peker i retning av at studien bør tilnærme seg de to forskningsspørsmålene med forskjellige typer modeller. En enkel modell for å forstå sammenhengen mellom variablene og sluttkostnaden i prosjektene, og en potensielt mer kompleks modell for å predikere sluttkostnaden.



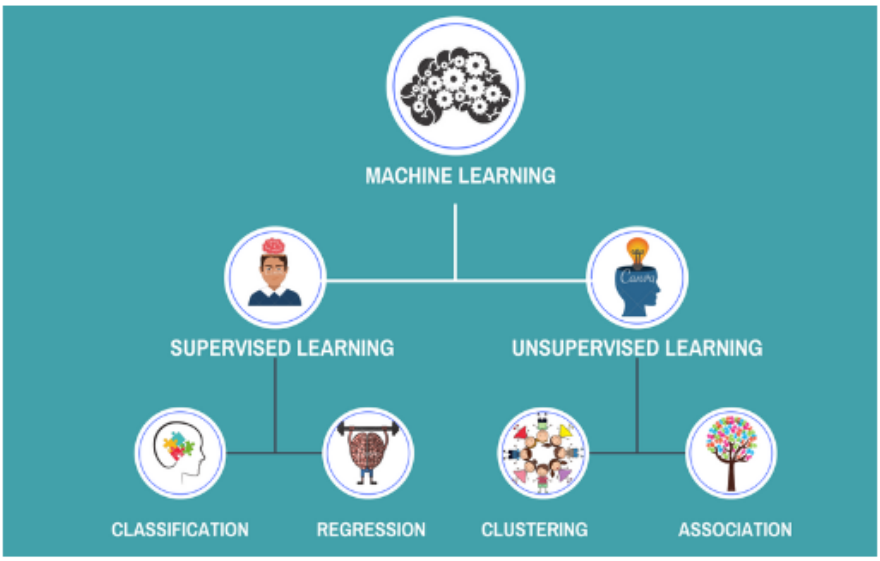
Figur 7: Oversikt kunstig intelligens, (Digitaliseringsdirektoratet, U.Å.)

Digitaliseringsdirektoratet (U.Å.) Skiller eksempelvis mellom maskinlæring, maskinresonnering og robotikk når det kommer til KI. Denne oppgaven vil kun fokusere på maskinlæringsdelen av KI.

2.3.1 Maskinlæring

Vi kan dele maskinlæring inn i to hovedkategorier. Overvåket og ikke overvåket maskinlæring. Hovedforskjellen mellom disse to kategoriene handler om hvorvidt en definerer utputten eller hva vi ønsker at resultatet skal være på forhånd (James, Witten, Hastie, & Tibshirani, 2021). Et strukturert datasett med forhåndsdefinerte variabler og en forhåndsdefinert utputt kalles for overvåket maskinlæring. Vi gir modellen en mengde data og definerer hva vi ønsker at den skal komme frem til. Formålet vil da være å senere kunne gi modellen ny data slik at den basert på den dataen vi allerede har gitt den, kan gi oss prediksjoner eller estimater. Om vi tar utgangspunkt i studien kan vi si at den handler om å forstå og predikere sluttkostnaden til et byggeprosjekt. Det vil si at vi på forhånd har definert utputten til modellene som skal benyttes. Oppgaven handler altså om overvåket maskinlæring. Om vi derimot var ute etter å utforske datasettet, og avdekke ulike mønster, ville ikke overvåkede modeller vært best egnet. Vi kunne altså ikke hatt forhåndsdefinerte variabler, men eksempelvis bedt modellen om å lage variabler for oss.

Om vi videre tar for oss overvåket maskinlæring kan vi videre dele dette inn i regresjon og klassifisering. Regresjon benyttes når det en ønsker å predikere er et tall. Som eksempelvis en kostnad, pris eller grader. Klassifisering handler derimot om å predikere hvilken forhåndsbestemt klasse eller kategori noe tilhører. Eksempelvis om prisen vil gå opp, ned eller holde seg, eller om det vil regne i morgen eller ikke (James, Witten, Hastie, & Tibshirani, 2021). Siden studien handler om å predikere sluttkostnaden til et byggeprosjekt vil det altså si at regresjon er den typen overvåket maskinlæring som vil benyttes.

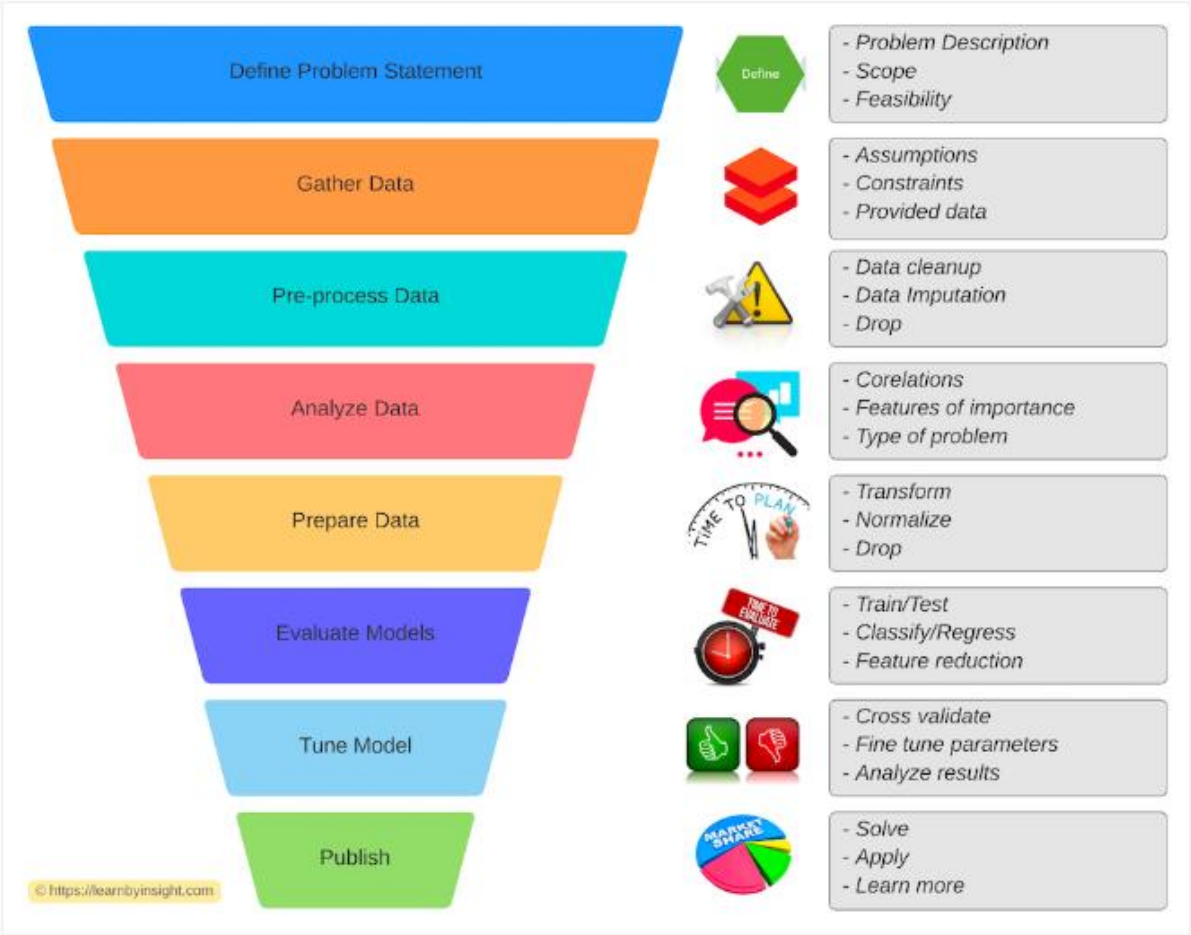


Figur 8: Overvåket og ikke overvåket maskinlæring, (Valcheva, U.Å.)

2.3.2 Maskinlæringa arbeidsmetodikk

Når en jobber med overvåket maskinlæring er det mange forskjellige steg en skal gjennom før en kan starte å trekke slutninger basert på modellen. Om vi tar utgangspunkt i figur 9 kan vi se et eksempel på en tilnærming til overvåket maskinlæring. De fleste av disse stegene dekkes videre i metodedelen av studien. Eksempelvis datainnsamling, prosessering av data, analysering av data og forberedelse av data. Selve maskinlæringsdelen er kun de to stegene knyttet til å evaluere modeller og fintuning av modellen. Når en har gått gjennom de innledende stegene og står klar med et datasett som er klart for å testes ut på ulike modeller starter en som oftest med å dele datasettet i to deler. Den ene delen kalles treningsdatasettet og den andre delen kalles testdatasettet (Mewara, 2020). Splitt ratioen kan variere, men en ønsker som regel å gi modellen så mye data som mulig til å trene seg opp. Blir derimot testdatasettet for lite vil det være vanskelig å vurdere hvor godt modellen presterer på ny data. Videre kan en dele treningsdatasettet i to. Et for selve treningen av

modellen og et for å tune modellen, valideringsdatasettet. Vi bruker altså valideringsdatasettet til å vurdere hvilken type maskinlæringsmodell som presterer best. Dette kalles modellutvelgelse. Testdatasettet bruker vi til å vurdere prestasjonen til den valgte modellen. (Hastie, Tibshirani, & Friedman, 2009) Dette fordrer riktignok at en har rikelig data tilgjengelig.



Figur 9: Machine learning workflow, (Mewara, 2020)

2.3.3 Regresjon

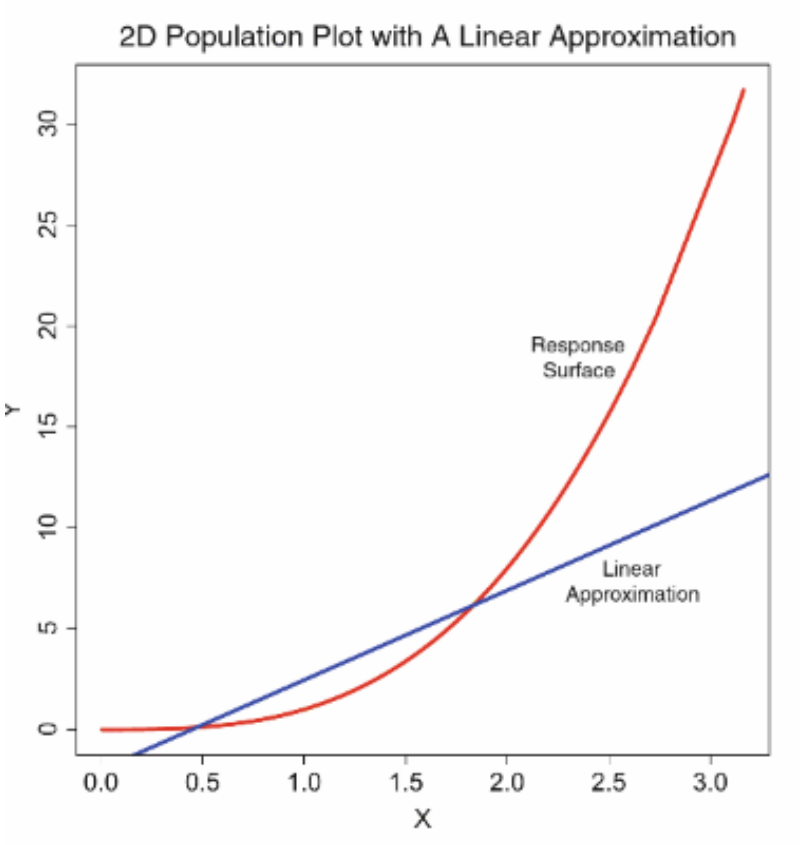
Den enkleste form for regresjonsanalyse er lineær regresjon. Ved å forstå hvordan lineær regresjon fungerer vil en også kunne forstå hvordan regresjonsmodeller som er mer komplekse fungerer. En lineær regresjonsmodell kan matematisk skrives som:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

I dette tilfellet er det variabelen Y vi er interessert i å predikere ved hjelp av variabelen X. Modellen forutsetter at det er et lineært forhold mellom Y og X (James, Witten, Hastie, & Tibshirani, 2021). β_0 er konstantleddet og forteller oss hvor på Y-aksen vår rette linje starter, Altså hva er Y hvis $X = 0$. β_1 er stigningstallet til X, og forteller oss forholdet mellom Y og X. Hvis X øker med 1, så øker eller synker $Y = \beta_1$. ε er det leddet i funksjonen som sier noe om avstanden mellom de faktiske observasjonene våre og den estimerte linjen regresjonsmodellen representerer. I teorien vil en høy ε fortelle oss at den estimerte linjen ikke fungerer godt til å predikere forholdet mellom Y og X. Det er også mulig å komplisere modellen ved å legge til flere Xer eller ved å lage modellen slik at forholdet mellom Y og X ikke lengre er lineært, men eksempelvis kvadratisk eller eksponentielt.

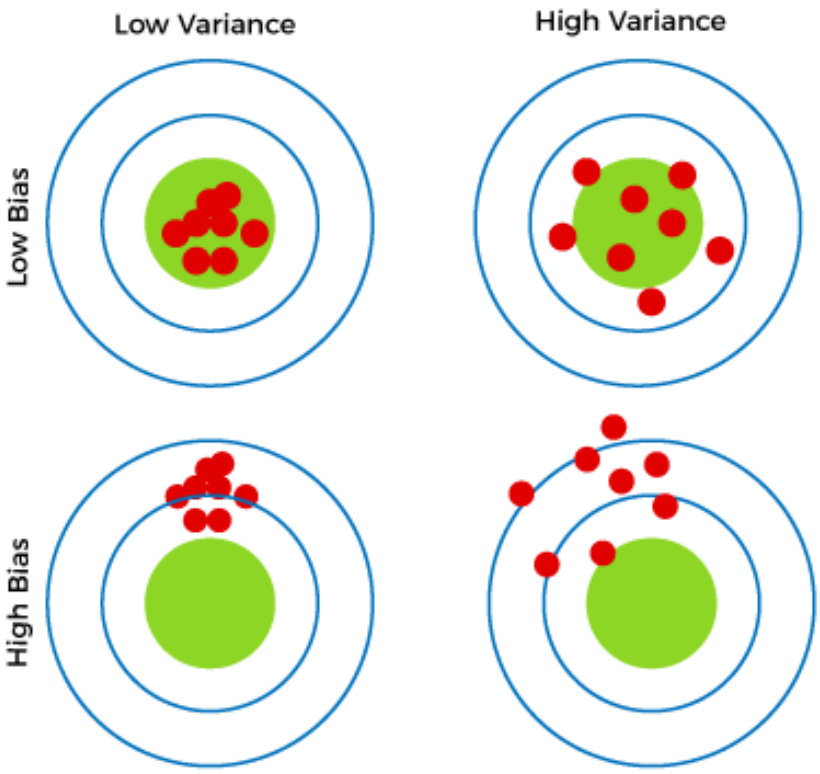
2.3.4 The bias-variance trade-off

En viktig del av regresjonsanalysen er altså at vi må forsøke å si noe om hva slags forhold det er mellom Y og de valgte $Xene$, eller variablene vi har inkludert i modellen. Om vi velger feil vil modellen ikke klare å predikere på en god måte.



Figur 10: Lineær modell og ikke- lineær sammenheng, (Berk, 2020)

Som vi ser i figur 10 så er den egentlige sammenhengen mellom Y og X av eksponentiell karakter, men en forsøker å få en lineær modell til å passe. Selv om den lineære kurven representerer den beste sammenhengen mellom Y og X som er mulig å få til gjennom lineær regresjon så vil den aldri klare å predikere godt, spesielt ved høye X- verdier. Dette kalles å «undertilpasse» modellen. En sier gjerne at modellen er overgeneralisert (van der Aalst, et al., 2010). Modellen klarer ikke å ta inn over seg den underliggende trenden i dataen. Vi kan i motsatt fall ende opp med å «overtilpasse» modellen. Modellen blir for kompleks og forsøker å dekke alle de ulike observasjonene. Om modellen er undertilpasset sier vi at den har høy bias. I motsatt fall har den høy varians. Dette er momenter vi er nødt til å ta stilling til i forbindelse med valg av maskinlæringsmodell. Vi må søke å optimalisere to kriterier: få så lav bias som mulig, samtidig som vi får så lav varians som mulig. Denne tilnærmingen vil gi oss modellen med minst avvik mellom faktisk verdi og predikert verdi. Om en kun søker å optimalisere eksempelvis bias, så vil en ofte ende i en situasjon med høy varians, og omvendt.



Figur 11: The bias-variance trade-off, (Javapoint, U.Å.)

2.4 Litteraturgjennomgang av prediktiv analyse i byggeprosjekter

Hittil har oppgaven tatt for seg byggeprosjekter og prediktiv analyse isolert. Denne delen av oppgaven har som formål å sette disse to temaene sammen til en gjennomgang av tidligere studier gjennomført innen predikering av sluttkostnaden i byggeprosjekter.

2.4.1 Årsaker til kostnadsavvik i byggeprosjekter

Innledningsvis er det naturlig å reflektere rundt hvorfor prosjekter ender opp med en sluttkostnad høyere eller lavere enn estimert. Morris (1990) peker blant annet på omfangsendringer, entreprenørenes tilgang på materiale og forsinkelser som følge av lange beslutningsprosesser i offentlig sektor. Jackson (2002) trekker også frem omfangsendringer som en viktig faktor. I tillegg har han funn som indikerer at det brukes for kort tid i tidligfase, noe som fører til at beslutninger fattes på feil eller mangelfullt grunnlag. Videre fremhever Jackson (2002) også viktigheten av en kompetent og godt sammensatt prosjektorganisasjon som en viktig faktor for å unngå overskridelser, i tillegg til ytre faktorer som grunnforhold og markedssituasjonen. Creedy, Skitmore, & Wong (2010) hevder, i tillegg til omfangsendringer, at kostnadsavvik oppstår som følge av at usikkerhetsanalysen ikke er gjennomført grundig nok. Dette fører ofte til at forventede tillegg ikke er tilstrekkelig.

Flyvbjerg (2007) deler årsakene til kostnadsavvik, hovedsakelig overskridelser, inn i tre forskjellige kategorier. Den første er tekniske årsaker og er knyttet til mindre gode prognostiseringsteknikker, mangel på data, ærlige feil, iboende utfordringer knyttet til predikering av fremtiden og manglende erfaring hos de som estimerer og predikerer. Den andre årsaken er den psykologiske, som handler om at en i prosjektets tidligfase kan gå i planleggingsfellen. Prosjektorganisasjoner har gjerne en tendens til å være overoptimistisk i planleggingen av prosjektet. De overvurderer nytte og undervurderer kostnader. Estimatene bygger da i mindre grad på rasjonelle vurderinger av fordeler og ulemper eksempelvis basert på sannsynlighet. Den tredje er den politisk- økonomiske årsakskategorien. Den logiske mekanismen bak denne årsaken er at en ønsker å få godkjenning til å gjennomføre et prosjekt. Om det er stor kamp om ressursene kan det være enkelt å fremheve nytten og underestimere kostnader slik at prosjektet fremstår bedre enn hva det egentlig er.

Larsen, et al. (2023) trekker også frem tidligfasen som en fase som er usikker av natur, og at beslutninger må fattes under usikre omstendigheter. En er derfor nødt til å bruke tilstrekkelig tid i tidligfase. Slik at konseptvalg, sentral styringsdokument og

gjennomføringsoppdrag blir besluttet på et så godt grunnlag som mulig. Videre peker de på at det heller ikke er uvanlig at den estimerte sluttkostnaden kan øke betraktelig gjennom tidligfasen. I noen tilfeller kan det være opp mot 30 prosent forskjell mellom estimert sluttkostnad i konseptfasen og ved beslutning om investering.

Oppsummert kan vi si at det er mange faktorer som medfører risiko for kostnadsavvik. Et moment som går igjen hos flere kilder er omfangsendringer. Naturlig nok vil en endring av større karakter underveis i prosjektet medføre avvik mellom estimert sluttkostnad og faktisk sluttkostnad. En kan på mange måter si at de to summene ikke kan sammenlignes lengre. Samtidig må en sammenligne, indikasjonene i litteraturen peker tross alt i retning av at omfangsendring ikke er et sjeldent fenomen. Videre kan en også stille seg spørsmål rundt hvorvidt omfangsendringer er relatert til en annen gjenganger, nemlig for liten tid brukt, og for upresise vurderinger i tidligfase. Flyvbjerg (2007) peker også på at det foreligger tekniske, psykologiske og politiske årsaker til hvorfor tidligfasen ofte ikke ender med et så riktig estimat som mulig.

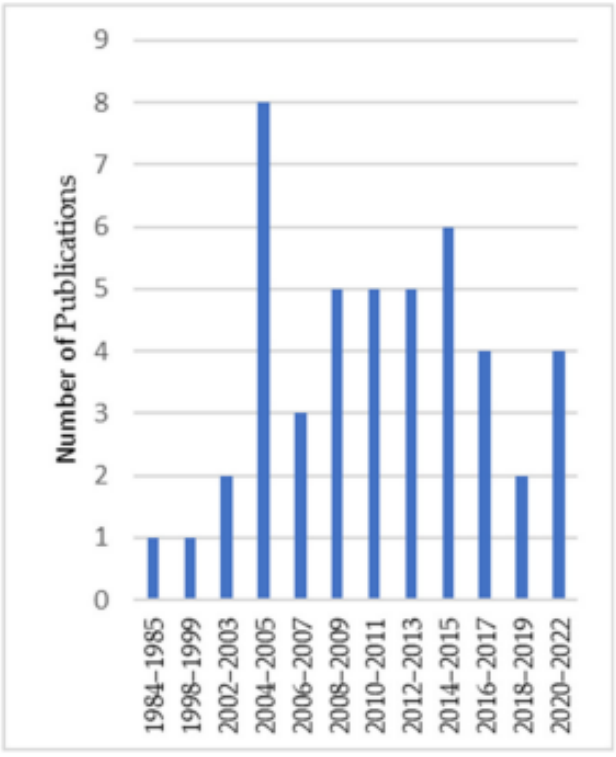
Utover tidligfasen finnes også andre momenter som grunnforhold, uforutsette endringer i markedssituasjonen, tilgang på materiell og utstyr hos entreprenører, kompetanse og erfaring i prosjektorganisasjon og lange beslutningssløyfer i offentlig sektor. Disse kan til en viss grad vurderes ut ifra hvor mye en kan gjøre med forholdene eller ikke. Det er eksempelvis fullt mulig å jobbe med kompetansen til eksempelvis prosjektlederen i virksomheten, eller være bevisst i sammensetning av hvilke team som skal jobbe på hvilke prosjekter. I andre enden av skalaen har vi endringer i markedssituasjonen og tilgang på materialer, som er vanskelig å påvirke. Krig eller pandemi er eksempler på hendelse som har medført sjokk i markedssituasjonen i nyere tid. Dette kan ha påvirket sluttkostnaden i mange prosjekter. Lange beslutningsprosesser og grunnforhold er begge av en karakter hvor en kan forsøke å ta høyde for det, eller jobbe med saken. Eksempelvis kan en gjennomføre prøvegravning og ta jordprøver. Noe som vil redusere usikkerheten, men ikke eliminere den. Videre kan en aktivt jobbe for bedre samhandling med de ulike interessentene involvert i prosjektene, som eksempelvis PE, BA, PA og ODG. Dette vil kunne føre til en smidigere beslutningsprosess, men for et kategori 1 prosjekt vil en eksempelvis ikke kunne unngå at endringer skal opp til beslutning i Stortinget.

2.4.2 Predikering i byggeprosjekter

Gjennom denne delen av oppgaven vil tidligere forskning innen predikering i byggeprosjekter gjennomgås. Hensikten er å kartlegge ulike aspekter ved den tidligere forskningen på en slik måte at denne studiens bidrag klart kommer frem. Følgende aspekter ved tidligere forskning vil vurderes:

- Hvilket formål hadde studiene?
- Hvor er prosjektene i studiene geografiske lokalisert?
- Hvilke variabler ble inkludert i studiene?
- Hvilke predikeringsmetoder ble benyttet?
- Hvor presist klarte modellene å predikere, og hvordan ble dette målt?

Predikering i byggeprosjekter er ikke et nytt forskningsfelt. Karshenas utforsket eksempelvis byggekostnader ved hjelp av regresjonsanalyse allerede i 1984. Det er allikevel ikke før rundt millenniumskiftet at det begynner å bli flere studier knyttet til predikering i byggeprosjekter. Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) har eksempelvis gjennom en litteraturstudie innen prediktiv analyse i byggeprosjekter funnet 46 studier på området mellom 1984 og 2022.



Figur 12: antall utgivelser fordelt per år, (Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin, 2022)

Som vi ser ut ifra figur 12 har det stort sett kommet et par nye utgivelser hvert år helt siden starten av 2000- tallet. Tayefeh Hashemi, Ebadati, & Kaur (2020) har gjennomført en lignende studie der de fant 69 studier mellom 1985 og 2020. Nesten alle studiene funnet var utgitt etter 2000. Heller ikke i denne studien er det noen tegn til hverken en økende eller synkende trend i antall utgivelser. Det som derimot endrer seg, er eksempelvis analysemetoden. En ser en tendens til at kompleksiteten i maskinlæringsmodellene øker.

2.4.2.1 Formål

Et interessant spørsmål knyttet til tidligere forskning er, hvorfor forskes det? Er det eksempelvis noen hensikter som går igjen, eller er det mange forskjellige formål? Seokyon (2009) peker eksempelvis på beslutningsstøtten til ledelsen i virksomheter som et viktig formål. Dette er et formål som går igjen i mange av de tidligere studiene. Beslutningstakere må ta avgjørelser eksempelvis i forbindelse med overgangen fra tidligfase til gjennomføringsfase. Denne investeringsbeslutningen vil i stor grad være basert på den forventede sluttkostnaden til prosjektet. Ved å komme fram til modeller og metoder for å gi mer presise estimater vil dette kunne bidra til mer informerte beslutninger. En søker å si med så høy sannsynlighet som mulig hva sluttkostnaden blir noe som også kan redusere risikoen knyttet til beslutningen og gjøre driften av virksomheten mer forutsigbar.

Sharma, Zaki, Jha, & Krishnan (2022) på den andre siden ønsker å optimalisere byggeprosjekter ved å gjøre de datadrevne. Fokuset er altså å gjøre prosjektet best mulig. Det er kanskje ikke fundamentalt annerledes fra bedre beslutningsstøtte. En kan på mange måter argumentere for at optimalisering av et prosjekt også innebærer å ta gode beslutninger til rett tid. Videre trekker Pham, Le-Hong, & Tran (2023) frem at datadrevne predikeringsmetoder kan korte ned på tiden brukt til å estimere. Formålet deres er å effektivisere estimeringsprosesser i tidligfase og tilrettelegge for konkurransefortrinn hos virksomheter. Williams (2003) søker å forstå hvor mye entreprenørens tilbud vil endre seg fra kontrakten signeres til prosjektet er ferdig. Formålet her er å hjelpe prosjektledere og byggherrer med å planlegge med forventede endringer i gjennomføringsfasen. Kim, An, & Kang (2004) har et mer akademisk formål med sin studie. De ønsker å avgjøre hvilken type predikeringsmodell som er best til å predikere byggekostnader. De har dermed et fokus på å videreutvikle forskningsfeltet heller en mer praktisk rettede formål. Samtidig kan en si at

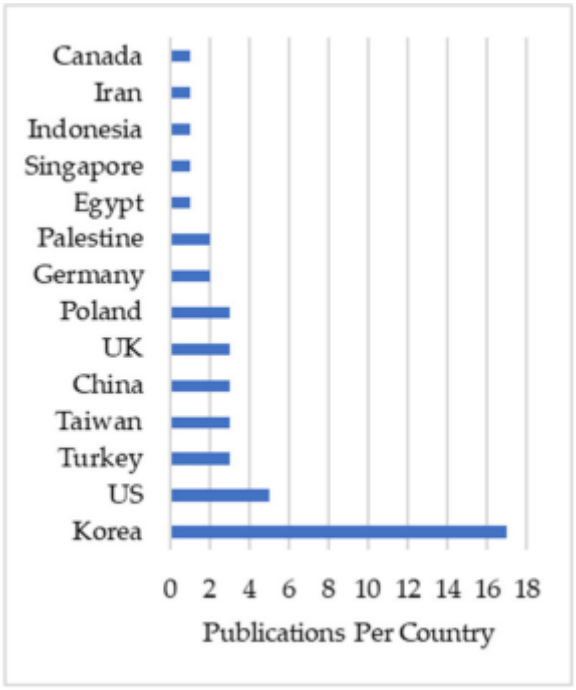
denne forskningen indirekte vil ha praktiske implikasjoner av den samme sorten som de ovennevnte studiene.

Det finnes altså mange ulike hensikter med studiene. Mange drar frem fordelene til virksomheter i EBA- bransjen som viktige formål. I kort kan vi si at forskningen handler om å finne de beste modellene og metodene for å predikere kostnader i byggeprosjekter i den hensikt å hjelpe virksomheter til å gjennomføre de riktige prosjektene på en best mulig måte.

2.4.2.2 Geografisk lokasjon

Lokasjon kan ha mye å si for hvorvidt studier er sammenlignbare. Det er flere aspekter som kan bidra til denne effekten. Eksempelvis kan lover og regler spille en rolle. Hvilken innstilling en har til arbeidstid og sikkerhet, helse og arbeidsmiljø kan variere fra land til land. I Norge er begge disse aspektene kontrollert av arbeidsmiljøloven. Andre land er kanskje ikke like strenge på disse områdene. Andre faktorer som kanskje spiller en større rolle på selve byggekostnaden kan være klima, grunnforhold og konkurransesituasjonen. Om vi tar utgangspunkt i Norge og Egypt, hvor eksempelvis Badawy (2020) har gjennomført en studie for å predikere byggekostnader i boligprosjekter, er det klart at dette er to relativt forskjellige land med tanke på klima og topografisk sammensetning. I Egypt trenger en kanskje ikke ta høyde for kulde på samme måte som i Norge, men hete er kanskje et større problem. Egypt er også flatere enn Norge, noe som kanskje kan medføre enklere byggeforhold. Grunnforholdene vil nok variere stort også. Derfor skriver Badawy (2020) at siden studien som er gjennomført kun har prosjekter i Egypt, så er prediksjonsmodellen utarbeidet i studien kun gjeldende i Egypt. Modellen må justeres for å kunne fungere i andre land.

Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) presenterer i sin litteraturgjennomgang en oversikt over hvor de ulike studiene de har gjennomgått er utarbeidet. Som vi kan se ut ifra figur 13 har de ikke lyktes med å finne en studie i Norge knyttet til predikering i byggeprosjekter. Det er også mange land som i liten grad er sammenlignbare med Norge. De som kanskje er mest sammenlignbare er Canada, Tyskland, Polen og Storbritannia.



Figur 13: publikasjoner per land, (Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin, 2022)

Den eneste norske studien knyttet til predikering i byggeprosjekter det har lyktes denne forskeren å finne er studien av Mæhlen & Bekkevold (2022) som denne studien til dels søker å bygge videre på. Fokuset til Mæhlen & Bekkevold (2022) ligger hovedsakelig i å forstå hvordan data fra tidligere prosjekter kan benyttes til å forklare det relative avviket mellom estimert sluttkostnad og faktisk sluttkostnad ved hjelp av regresjonsanalyse. De har riktignok et forsøk på å predikere det relative avviket mellom estimert og faktisk sluttkostnad, men de konkluderer raskt med at modellene som forsøkes ikke er treffsikre. Muligens kan dette skyldes at de kun predikerer med et datasett på 78 prosjekter.

2.4.2.3 Variabler

Hvilke variabler som er benyttet i tidligere forskning er også viktig å ha et forhold til. Abu Hammad, Ali, Sweis, & & Sweis (2010) bruker eksempelvis prosjekttype, og antall kvadratmeter i sin studie. Arafa & Alqedra (2011) bruker også antall kvadratmeter, i tillegg til antall etasjer, antall rom, antall heiser, antall søyler, type fundament og estimert kostnad for byggets «skjelett». Lowe, Emsley, & Harding (2006) utforsker 41 variabler, men konkluderer med at antall kvadratmeter, byggets funksjon, prosjektets varighet, antall mekaniske installasjoner og antall fundamentpæler er de fem viktigste kostnadsdriverne blant de undersøkte variablene. Kim, An, & Kang (2004) benytter antall kvadratmeter, antall etasjer, antall leiligheter eller enheter, varighet, tak type, fundament type, kjeller type og grad av

kvalitet på materialer. Mæhlen & Bekkevold (2022) inkluderer variablene antall kvadratmeter, estimert kvadratmeterpris, bygningstype, entreprisemodell, investeringstype og størrelsen på ulike poster etter bygningsdelstabellen.

Basert på disse eksemplene kan vi se et par mønstre. Alle studiene benytter antall kvadratmeter som variabel. I tillegg er det flere som benytter en eller annen form for type bygning. Noen studier har også veldig mange ulike variabler for å beskrive bygget, som antall etasjer, type fundament, antall heiser m.m. Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) har i sin litteraturstudie listet opp de variabler som totalt på tvers av de ulike studiene har vist seg å være best egnet til å predikere sluttkostnaden. På toppen av listen ligger blant annet:

- Brutto totalareal
- Antall etasjer
- Fundamenttype
- Antall bygg/leiligheter
- Antall heiser
- Taktype
- Bygningstype
- Varighet
- Lokasjon

En kan se på dette fra forskjellige vinklinger. En kan for eksempel tenke seg at årsaken til at disse variablene er på toppen er fordi de har en god evne til å predikere sluttkostnaden i prosjekter. Det kan på den andre siden hende at de havner på toppen fordi de har blitt brukt i flest studier. Det er altså ikke en utfyllende liste eller en fasit, men heller et godt sted å starte når en skal samle inn data for predikering i byggeprosjekter. Videre kan det være verdt å merke seg at hovedtyngden av variabler er knyttet til bygget som skal bygges, og tar ikke innover seg en del variabler knyttet til andre aspekter ved prosjektet som eksempelvis hva slags metodikk prosjektorganisasjonen benytter for å gjennomføre prosjekter, eller hvor erfaren prosjektorganisasjonen er. De fleste studiene søker også å benytte variabler som er tilgjengelig i tidligfase. Noe som er naturlig med tanke på at investeringsbeslutningen fattes i tidligfase. Dette kan være en årsak til at variabler som først oppstår i gjennomføringsfase i liten grad har vært undersøkt. Eksempler på denne typen variabler kan være knyttet til valg

av entreprenør, som hvor mange tilbud som ble mottatt når kontrakten ble lyst ut, og hvor mange endringsavtaler som ble signert.

2.4.2.4 Prediksjonsmetoder

Tidligere studier har flere ulike tilnærminger til valg av analysemetode. Det finnes flere forskjellige typer overvåket maskinlæring med ulik grad av kompleksitet. Dette kan gjøre det vanskelig å sammenligne resultatene fra tidligere forskning. Det er også flere som har søkt å avdekke hvilke metoder som fungerer best. Et eksempel er Kim, An, & Kang (2004) som benytter multippel regresjonsanalyse (MRA), nevrale nettverk (NN) og case-based reasoning (CBR) for å avdekke hvilken metode som fungerer best på deres datasett fra Korea. De konkluderer med at NN gir de mest nøyaktige estimatene, men at CRB har fordeler fremfor NN i form av enkel bruk og bedre mulighet til å forstå resultatet. Sharma, Zaki, Jha, & Krishnan (2022) bruker også flere metoder som. MRA, NN, random forest (RF) og gradient boosted tree (GBT). De konkluderer med at GBT gir de mest nøyaktige estimatene. Andre studier som eksempelvis Lowe, Emsley, & Harding (2006) og Abu Hammad, Ali, Sweis, & Sweis (2010) bruker kun MRA, mens Al mnaseer, Al-Smadi, & Al-Bdour (2023) og Arafa & Alqedra (2011) kun benytter NN. Dette gjør at de studiene som kun benytter en metode blir vanskeligere å sammenligne med hverandre.

Tayefeh Hashemi, Ebadati, & Kaur (2020) finner i sin litteraturgjennomgang at det er MRA og NN som er de hyppigst brukte metodene for predikering i byggeprosjekter. I tillegg ser de en tendens til at NN de senere årene i stadig større grad er formen for maskinlæring som i størst grad forskes på når det kommer til predikering i byggeprosjekter. I litteraturstudien til Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) trekker de også frem NN og MRA som de analysemetodene som er benyttet mest i tidligere forskning, i tillegg til CBR. NN ble benyttet i 48% av studiene, MRA 22% og CBR 26%.

2.4.2.5 Resultater og måling

For å kunne sammenligne denne studien med tidligere studier trengs det en oversikt over hvor nøyaktig de ulike studiene har klart å predikere. Å kunne levere presise estimater er på mange måter et av de viktigste bidragene forskningen kan ha for virksomhetene som opererer i byggebransjen. Virksomheter ønsker følgelig ikke å implementere maskinlæring i prosjektvirksomheten om de ikke ser at modellene kan levere resultater som tilfører en eller annen form for verdi til prosessen. Det er også avgjørende å finne ut hvordan tidligere

studier har målt predikeringsevnen til modellene sine, da det kan være forskjellige måter å gjøre dette på.

Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) viser at de vanligste metodene for måling av nøyaktighet i tidligere forskning er root mean square error (RMSE), mean square error (MSE) og mean absolute percentage error (MAPE). Alle disse er en eller annen form for avviket mellom estimert verdi og faktisk verdi. Videre fremhever de at MAPE er den eneste av de ovennevnte metodene som er uavhengig av skalaen i datasettet de er generert i. Om en har en MSE på eksempelvis 500 så er det vanskelig å vite hvorvidt det er bra eller dårlig uten å ha et forhold til den skalaen det blir predikert på. Et gjennomsnittlig avvik i prosent på 5% vil derimot være enklere å sammenligne med tidligere og kommende studier.

Videre fremhever Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) at MAPE i de studiene som ble gjennomgått lå mellom 2 og 21%, med hovedtyngden av studier mellom 5 og 13%. Ved å benytte dette som et mål på hvorvidt modellene i studien predikerer godt eller dårlig, vil det være mulig å sammenligne resultatene med tidligere forskning.

2.5 Oppsummering av litteraturgjennomgang

For å oppsummere kan vi si at det er forsket relativt mye på maskinlæring i byggeprosjekter. Forskningen fremstår samtidig noe fragmentert i den forstand at det er mange ulike tilnærminger til predikeringen, på mange ulike lokasjoner, som måler resultatene sine forskjellig. Dette kan føre til at det kan være vanskelig å sammenligne studiene med hverandre. Det mangler på sett og vis en slags standard som sikrer at resultater på tvers av forskningen kan sammenlignes.

Om vi knytter teorien til forskningsspørsmålene så kan vi si at hvilke variabler som påvirker sluttkostnaden i byggeprosjekter til en viss grad er avdekket gjennom eksempelvis litteraturstudien til Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022). Samtidig er dette bare delvis utforsket i norske offentlige byggprosjekter gjennom studien til Mæhlen & Bekkevold (2022).

Når det kommer til predikering i byggeprosjekter er mange aspekter dekket. De er derimot ikke testet ut under norske forhold. I tillegg vil variabler som ikke har vært testet ut i tidligere studier utforskes i denne studien med hensyn til deres prediktive evner, eksempelvis andelen tidligfasekostnad i prosjektet. Jackson (2002) trekker jo frem at en bør

bruke god nok tid i tidligfase for å unngå kostnadsavvik. Denne variabelen kan potensielt bidra til å si noe om høyere eller lavere kostnader i tidligfase er relevant for sluttkostnaden, da tidligfasekostnaden er tett knyttet til tid brukt i tidligfase.

Bidraget i studien må derfor vurderes til at det er den geografiske faktoren som skiller studien fra tidligere forskning. I den forstand at allerede utforskede metoder og variabler, med noen unntak, skal testes ut i en norsk kontekst. For å gjøre studien sammenlignbar med allerede eksisterende forskning vil det være fordelaktig å teste ut flere typer maskinlæringsmodeller som MRA, NN, RF og GBT. Resultatene bør også presenteres på forskjellige formater som RMSE, MSE og MAPE.

3. Metode

3.1 Innledning til metodekapittelet

Metodedelen av oppgaven kan i grovt deles inn i følgende deler:

- Valg av metode
- Datagrunnlaget
- Analysemetode
- Etiske vurderinger
- Oppsummering

Delen knyttet til valg av metode vil ta for seg vurderinger knyttet til hva slags metode som er egnet til å svare på forskningsspørsmålene.

For å vurdere datagrunnlaget vil oppgaven fokusere på hvordan data er innsamlet og hvilke forberedelser som har blitt gjort for å gjøre datagrunnlaget klart til analyse. Fokuset i denne delen er knyttet til datakvalitet og utfordringer knyttet til innsamlingen. Videre vil oppgaven konsentrere seg om utvalget. Hvilke variabler som er inkludert i datagrunnlaget, samt deres relevans for studien. Avslutningsvis i denne delen av oppgaven rettes et kritisk blikk på datagrunnlaget for å belyse hvilke svakheter datagrunnlaget har i forhold til forskningsspørsmålene.

Analysedelen peker på hvilke vurderinger som er gjort knyttet til hvilke analyser som er gjennomført, og hvorfor disse er valgt. Analysedelen vil også søke å beskrive hvilke hyperparameter som relevante med tanke på tuning av modellene.

De etiske vurderingene i oppgaven søker å belyse hvilke faktorer oppgaven i større eller mindre grad har måttet ta stilling til. Et av hovedmomentene i denne delen er knyttet til anonymisering av data og habilitet.

Hovedhensikten med oppsummeringsdelen er å knytte metodekapittelet og forsyningsspørsmålene sammen på en ryddig og oversiktlig måte.

3.2 Valg av metode

Denne oppgaven har to forskningsspørsmål. Det første handler om å vurdere hvilke variabler som påvirker sluttkostnaden i et offentlig byggeprosjekt. Deretter er spørsmålet hvorvidt disse variablene kan brukes for å predikere sluttkostnaden i byggeprosjekter. Det er

sannsynligvis både kvalitative og kvantitative tilnærminger til disse spørsmålene. En utfordring knyttet til en kvalitativ tilnærming i denne sammenheng omhandler generalisering av funn. En annen utfordring er i hvilken grad det er mulig å vurdere hvor mye en variabel påvirker sluttkostnaden gjennom eksempelvis intervjuer. En mulig tilnærming der både kvantitativ og kvalitativ metode benyttes kan være å gjennomføre intervjuer for å avdekke hvilke variabler ulike nøkkelpersoner i ulike prosjektvirksomheter anser som viktige for sluttkostnaden. Disse variablene kan deretter analyseres statistisk gjennom kvantitativ metode i mange forskjellige prosjekter. Basert på funnene i den kvantitative undersøkelsen kan en utarbeide en ny kvalitativ studie for å utforske hvorfor en får disse funnene i den kvantitative delen, eller hvilke mekanismer som kan ligge bak funnene. Selv om denne tilnærmingen virker veldig spennende, vil den samtidig være veldig tidkrevende. Oppgaven har derfor konsentrert seg om den kvantitative delen av ovennevnte scenario. For å finne variablene tar oppgaven utgangspunkt i tidligere forskning, og spørsmålet knyttet til hvorfor overlates til videre forskning. Dette innebærer samtidig at oppgaven har et deskriptivt preg.

3.3 Datagrunnlaget

3.3.1 Datainnsamling

Oppgaven tar utgangspunkt i alle avsluttede byggeprosjekter i Prosjekt og utviklingsavdelingen til Forsvarsbygg som det har vært mulig å finne systematisk lagrede data knyttet til. Utgangspunktet for utvalget er en liste over alle avsluttede prosjekter. Ut over de data som lå lagret i denne oversikten har det blitt gjort uttrekk av data fra ERP systemet til Forsvarsbygg. Hensikten har vært å tilføre flere opplysninger om de enkelte prosjektene. Hovedsakelig har dette handlet om ekstra informasjon rundt anskaffelsene i prosjektene, samt varigheten på prosjektene. De data som er samlet inn er dermed registerdata i motsetning til eksempelvis data som er samlet inn gjennom en spørreundersøkelse. Prosjektene er gjennomført i perioden fra 2006 og frem til 2023, med hovedtyngde av prosjekter med oppstart mellom 2010 og 2018 og ferdigstillelse mellom 2017 og 2023. Innledningsvis i datainnsamlingen har fokuset vært å samle inn så mye informasjon som mulig. For senere gjennom analyse av datagrunnlaget potensielt redusere antall variabler og fjerne data som av ulike grunner ikke er egnet.

3.3.2 Datakvalitet

Et av de viktigste momentene knyttet til datainnsamlingen er datakvalitet. Dette er et tema som strekker seg helt inn til kjernen av selve forskningen. Sannhet, virkelighet og fakta er alle begreper som er knyttet til datakvaliteten. Samtidig dukker det også opp forskningsetiske momenter som eksempelvis habilitet og redelighet (De nasjonale forskningsetiske komiteene, 2019). En ønsker et datagrunnlag som beskriver virkeligheten på en best mulig måte uten å være påvirket av forskerens egne interesser, eller andres. Samtidig skal streben etter det beste datagrunnlaget ikke gå på bekostning av eksempelvis personvern. Forskeren må også sikre etterprøvbarheten til studien. De forskningsetiske momentene i oppgaven vil diskuteres senere i metode delen av studien. Det er samtidig viktig å være disse bevisst når en skal vurdere mulige kilder til lavere kvalitet i datagrunnlaget.

Mæhlen & Bekkevold (2022) peker på den internasjonale standarden for datakvalitet ISO 25012 som et godt utgangspunkt for å vurdere datakvaliteten i datagrunnlaget (IOS 25000, U.Å.). Her skilles det mellom iboende datakvalitet og systemavhengig datakvalitet. Iboende datakvalitet er knyttet til kvaliteten på selve datagrunnlaget. Systemavhengig datakvalitet handler derimot om hvorvidt systemene data genereres og lagres i legger til trette for å oppnå datakvalitet. Oppgaven konsentrerer seg derfor om den iboende kvaliteten til datagrunnlaget. I «*veileder for beskrivelse av kvalitet på datasett*» utgitt av Digitaliseringsdirektorartet (2020) benyttes det fire dimensjoner for å beskrive den iboende datakvaliteten: fullstendighet, aktualitet, konsistens og nøyaktighet. «*Ikke alle de predefinerte kvalitetsmålene er relevant å måle i enhver sammenheng. Man står fritt til å velge ut de kvalitetsmålene som er aktuelle for datasettet. I noen tilfeller er det «nøyaktighet» som er viktigst for brukerne av datasettet, i andre tilfeller kan det være «konsistens», eller begge. I mange tilfeller må man også ta høyde for at det er flere ulike typer brukere. Man bør derfor velge de kvalitetsmålene som erfaringsmessig er viktige for mange brukere av det aktuelle datasettet»* (Digitaliseringsdirektoratet, 2020). Derfor kommenterer oppgaven alle fire dimensjonenes påvirkning på kvaliteten.

3.3.2.1 Fullstendighet

Denne dimensjonene handler om hvorvidt data mangler (underdekning), er overflødig (overdekning) eller om en har gjort grep for å fylle inn manglende data (imputering) (Digitaliseringsdirektoratet, 2020).

Underdekning eksisterer garantert i datagrunnlaget. Det er mange variabler som har flere rader med manglende informasjon. Det er flere potensielle kilder til denne underdekningen. Siden hovedkilden til datasettet er en liste som manuelt er oppdatert av flere forskjellige individer etter hvert som prosjektene er avsluttet kan det være felter som ikke har blitt fylt ut enten fordi det ikke var klart hva som skulle fylles ut eller eksempelvis at det er lagt til variabler underveis i listens levetid, og at de radene som allerede var utfylt ikke ble oppdatert med den nye variabelen. Listen har heller ikke en form for standard for håndtering av informasjon som ikke er relevant eller av andre årsaker mangler. Dette kan også være en kilde til overdekning av data. I tillegg kan en feilkilde rett og slett være at verdiene er skrevet inn feil. Overdekning er derimot vanskeligere å avdekke enn underdekning i datagrunnlaget da det er vanskelig og tidkrevende å ettergå alle opplysningene som er lagt inn for potensielle feil.

Den største kilden til underdekning i datasettet er knyttet til de variablene som er lagt til den opprinnelige listen med avsluttede prosjekter gjennom databasesøk. Hovedårsaken til underdekning i dette tilfellet er knyttet til endring av systemer og datadimensjoner internt i organisasjonen over tid. Når en eksempelvis har startet å håndtere kontrakter på en ny måte i ERP systemet har ikke den gamle informasjonen blitt oppdatert på tilsvarende måte. Dette kan også være en kilde til overdekning i de tilfeller der et prosjekt har vært gjennomført både med ny og gammel metode for kontraktshåndtering.

Når det kommer til imputering har oppgaven vurdert dette både som en egnet og uegnet tilnærming til håndtering av manglende verdier. Typiske metoder for å håndtere denne utfordringen er å sette manglende verdier lik gjennomsnittet, medianen, null eller verdien med høyeste frekvens. Alle disse metodene vil innebære at en antar noe om datagrunnlaget som i de fleste tilfeller er vanskelig å begrunne på en god måte. I tillegg vil en slik håndtering av data kunne medføre at en konkluderer på feil grunnlag. Samtidig kan det være hensiktsmessig å tilnærme seg behandling av manglende verdier på forskjellige måter basert på hvilket forskningsspørsmål en fokuserer på i videre analyser. Hvilke variabler som påvirker sluttkostnaden, vil ved imputering kunne konkluderes feil ved å legge til verdier der de mangler. Når det kommer til hvorvidt prediktiv analyse kan bidra til beslutningsstøtte i offentlige byggeprosjekter vil spørsmålet i større grad handle om hvorvidt modeller som

imputerer manglende verdier predikerer bedre enn de som ekskludere dem. Dette innebærer at de to forskningsspørsmålene vil ha forskjellig tilnærming til imputering av data.

3.3.2.2 Aktualitet

Aktualiteten på datasettet er knyttet til tid. Ferskheten til dataene med andre ord (Digitaliseringsdirektoratet, 2020). Spørsmålet i denne sammenheng vil altså være hvor viktig det er at datagrunnlaget er relativt nytt, eller motsatt. Det kan tenkes at et byggeprosjekt som ble gjennomført mellom eksempelvis 2009 og 2012 er mindre egnet til å predikere fremtidige byggeprosjekter enn et som er gjennomført mellom 2020 og 2023. Det kan på den andre siden vurderes dit hen at aktualitet ikke påvirker datakvaliteten når brukskonteksten er en studie av byggeprosjekter tilbake i tid.

3.3.2.3 Konsistens

Denne kvalitetsdimensjon er knyttet til den logiske oppbyggingen av datasettet. Er det noen logiske brister i datasettet (Digitaliseringsdirektoratet, 2020)? Et eksempel ville vært at P85 er mindre enn P50, eller at generelle kostnader er høyere enn sluttkostnaden. Om det oppstår lav konsistens, eller inkonsistens i datasettet kan det være vanskelig å avdekke hvor feilen er. Er det eksempelvis variabelen sluttkostnad eller generelle kostnader som er feil. Det er ikke avdekket utfordringer knyttet til konsistens i datagrunnlaget.

3.3.2.4 Nøyaktighet

Datagrunnlagets nøyaktighet er knyttet til hvorvidt det korrekt gjenspeiler virkeligheten (Digitaliseringsdirektoratet, 2020). Det er åpenbart kilder til potensielt lav nøyaktighet i datasettet i de delene som manuelt er lagt inn av flere forskjellige individer over tid. På samme måte som overdekning og underdekning. Det er derimot vanskelig å ettergå nøyaktigheten i datasettet uten å gjøre et omfattende stykke arbeid. Det er derimot gjennomført omfattende stikkprøver av datasettet for å gjøre en generell vurdering av nøyaktigheten. Det ble ikke avdekket noen direkte feil i datasettet, og nøyaktigheten er derfor vurdert til å være tilfredsstillende nok for studiens formål.

3.4 Utvalg og variabler

Selve utvalget har i denne studien ikke vært problematisk å velge ut. Det er tatt utgangspunkt i alle avsluttede prosjekter hvor det strukturert har vært lagret data, og som faktisk er gjennomført. De har altså ikke blitt stanset før selve byggingen ble påbegynt. Studien har derfor ikke brukt tid på å finne ut hvilke observasjoner som skal inngå i det totale utvalget.

3.4.1 Filtrering av utvalg

Et spørsmål som derimot bør reises er hvorvidt det er behov for å filtrere datasettet ytterligere for å sikre at kvaliteten i datasettet er egnet for analyse. Som nevnt tidligere er det en del underdekning i datagrunnlaget. Om disse radene ekskluderes vil en på den andre siden potensielt sitte med for lite data til at eksempelvis prediktiv analyse vil gi resultater av verdi. Dette vil være vanskelig å fastslå uten at det testes ut. I den ene enden av skalaen har vi et komplett datasett uten manglende verdier på 132 observasjoner tilsvarende 46,1 prosent av det totale datasettet. Om datagrunnlaget derimot ekskluderer variablene med flest mangler er datagrunnlaget på 208 observasjoner, men uten data knyttet til kontrakter eller tidligfasekostnader.

Mæhlen & Bekkevold (2022) tilnærmer seg utfordringen med manglende verdier ved å lage flere datasett med forskjellige grader av filtrering. Denne studien velger en tilsvarende løsning. Dette vil bidra til en strukturert analyse der alle tilnærminger kan testes ut for å avgjøre hvilket datagrunnlag som gir best resultat. Studien vil også ha en bevisst tilnærming til hva slags filtrering som er gjennomført slik at en i etterkant av analysen kan peke på faktorer som kan være av betydning for resultatene. Denne filtreringen er nødt til å gjennomføres basert på et sett med forutsetninger og kriterier:

1. Hvilke variabler må være med i analysen?
2. Håndtering av ekstreme verdier i avhengig variabel.
3. Håndtering av manglende verdier.

3.4.1.1 Minimum av variabler

Som et minimum er studien avhengig av verdier knyttet til variabelen som skal forstås og predikeres, sluttkostnad. Denne forutsetningen resulterer i at 24 prosjekter blir fjernet fra datagrunnlaget.

Dette vil dermed si at det foreløpig er 310 prosjekter som kan analyseres hovedsakelig for å kunne besvare hvorvidt prediktiv analyse kan fungere som et verktøy for beslutningsstøtte.

3.4.1.2 Håndtering av ekstreme verdier

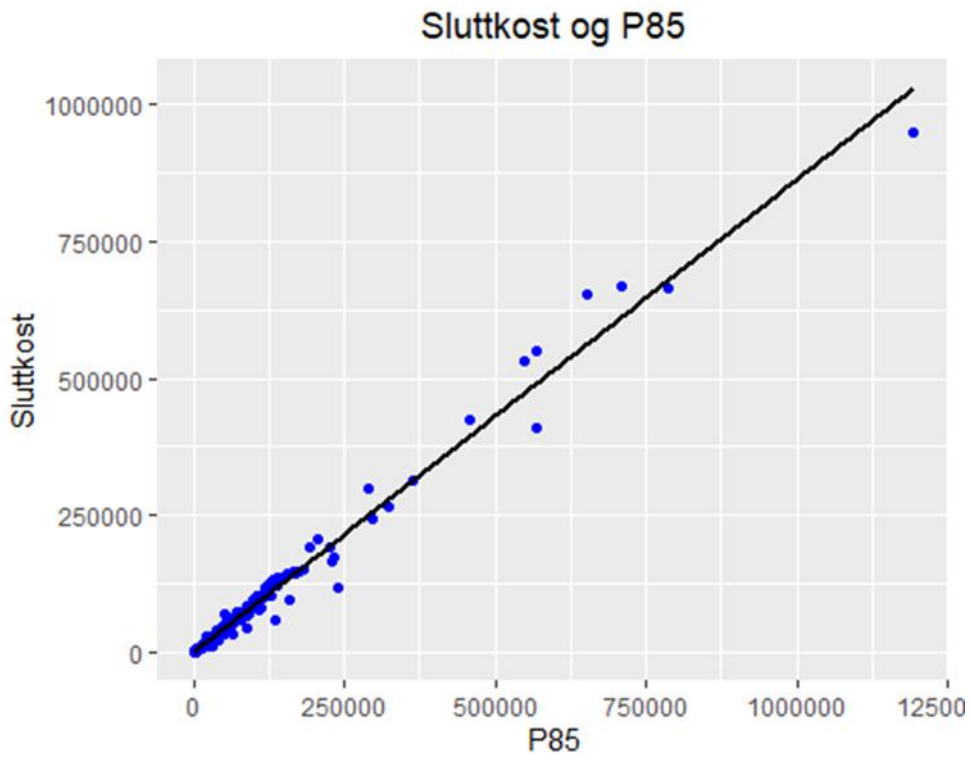
En annen utfordring knyttet til utvalget er ekstreme verdier. Mer spesifikt så store avvik mellom estimatet og sluttkostnaden at en blir usikker på hvorvidt tallene stemmer, eller om prosjektet har endret seg betraktelig etter at det ble iverksatt. Ekstreme verdier vil potensielt også kunne gjøre prediktive modeller mindre treffsikre da de søker å tilpasse seg og ta høyde for alle verdier best mulig. Om det da er verdier langt unna de andre verdiene vil problemer knyttet til eksempelvis overtilpasning kunne oppstå. Et typisk eksempel på hvorfor denne typen avvik kan oppstå er omfangsendringer i prosjektet. Et annet eksempel kan rett og slett være at datagrunnlaget er feil.

Figur 14: Relativt avvik mellom P50 og sluttkostnad

Om vi tar utgangspunkt i figur 14 som viser fordelingen av relativt avvik mellom estimat (P50) og sluttkostnad før filtrering ser vi at hovedmengden av ekstreme verdier ligger under 0 %. Prosjektene bruker altså ikke opp P50. Vi ser også tendenser til en normalfordeling rundt 0%. Mæhlen & Bekkevold (2022) reduserer sitt datasett med 5 prosent. Da med 2,5

prosent i hver ende av skalaen. Siden datagrunnlaget i denne oppgaven ikke har samme mengde ekstreme verdier i begge ender av skalaen velges derfor en annen løsning. Alle prosjekter med relativt avvik over 50 prosent over eller under null fjernes. Dette tilsvarer 5,2 prosent av datagrunnlaget eller 16 prosjekter.

Et annet aspekt knyttet til ekstreme verdier er selve fordelingen av sluttkostnaden. Om sluttkostnaden i noen prosjekter skiller seg ut fra resterende vil disse kunne skape støy i modellene ved at modellene tar høyde for noen få verdier i utvalget på en slik måte at den tilpasser seg hovedbolken av prosjekter dårligere. Figur 15 viser at de fleste prosjektene ligger med en sluttkostnad og P85 under 400 mill. kroner. De 8 prosjektene med sluttkostnad og P85 over 400 mill. kroner vil derfor fjernes fra datagrunnlaget. Dette innebærer også at resultatene og konklusjonene kun vil gjelde for prosjekter opp til 400 mill. kroner. Det kan tenkes at modellen også fungerer i prosjekter med høyere P85 og sluttkostnad, men det kan ikke oppgaven fastslå.



Figur 15: Fordeling Sluttkost og P85

3.4.1.3 Håndtering av manglende verdier

Den siste delen av filtreringen er knyttet til manglende verdier. Det er flere måter å tilnærme seg denne utfordringen. Hovedsakelig handler dette enten om å fjerne radene med manglende verdier, eller å erstatte verdiene. En står derfor overfor en slags trade- off mellom datakvalitet og utvalgets størrelse. Som nevnt tidligere knyttet til datakvalitet og imputering av manglende verdier, kan det tenkes at forskningsspørsmålene bør ha forskjellig tilnærming til håndtering av manglende verdier. Derfor vil oppgaven ekskludere manglende verdier i analyser knyttet til variablenes påvirkning på sluttkostnaden. Når det kommer til prediktiv analyse vil studien vurdere hvorvidt modeller med imputerte verdier predikerer bedre enn modeller som ekskluderer dem.

På denne måten hindrer vi at vi vurderer variablenes effekt på sluttkostnaden basert på antakelser og verdier som ikke er riktige. Samtidig som vi ikke reduserer utvalgets størrelse i forbindelse med predikering før vi får testet hvorvidt modellenes prestasjon gir grunnlag for dette.

3.4.1.4 Flere datasett

Etter disse filtreringsstegene står vi igjen med følgende datasett:

Tabell 4: Studiens filtrerte datasett

Navn	Antall observasjoner	Beskrivelse	Bruksområde
Hele utvalget	334	Alle prosjekter i utvalget	Grunnlaget for alle datasett
Datasett 1	286	Utvalget etter filtrering og erstatning av manglende verdier	Predikering av relativt avvik mellom estimat og sluttkostnad
Datasett 2	208	Utvalget etter filtrering og fjerning av variabler med mange manglende verdier. Ingen verdier erstattet.	Predikering av relativt avvik mellom estimat og sluttkostnad, samt vurdering av hvilke variabler som påvirker avviket mellom estimat og sluttkostnad.
Datasett 3	132	Utvalget etter filtrering og fjerning av alle prosjekter med manglende verdier	Vurdering av hvilke variabler som påvirker avviket mellom estimat og sluttkostnad.

3.4.2 Variabler

For å velge ut hvilke variabler som tas med i datasettet har tilnærmingen vært basert på hvilke variabler tidligere forskning har benyttet og anbefalt benyttet i videre forskning. Samtidig har begrensninger knyttet til tilgjengelig data vært noe som har påvirket hvilke variabler som er inkludert. Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) har som nevnt tidligere i sin litteraturstudie av 46 studier knyttet til predikering av byggekostnader listet opp de variabler som totalt på tvers av de ulike studiene har vist seg å være best egnet til å predikere sluttkostnaden. På toppen av listen ligger blant annet:

- Brutto totalareal
- Antall etasjer
- Fundamenttype
- Antall bygg
- Antall heiser
- Taktype
- Bygningstype
- Varighet
- Lokasjon

Mæhlen & Bekkevold (2022) benytter et par av disse i sin studie av norske offentlige byggeprosjekter. Av variablene Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) peker på benytter de brutto totalareal og bygningstype. I tillegg benytter de entrepriseform, hvorvidt prosjektet gjelder nybygg eller rehabilitering, estimert kvadratmeterpris og ulike posters relative størrelse av prosjektkostnaden basert på bygningsdelstabellen. Videre peker de på flere variabler de anbefaler videre forskning å inkludere, som antall tilbud i forbindelse med kontaktene, forsinkelse, anskaffelsesstrategi, teknisk forskrift, varighet og lokasjon. Det er et par fellesnevnere som kan benyttes for å oppsummere alle disse variablene. De fleste er bygningstekniske variabler. Samtidig kan vi si at de fleste av variablene er kjent tidlig i prosjektet.

Variablene i oppgaven er både hentet basert på tidligere forskning og hentet for å vurdere nye muligheter. Av de ovennevnte variablene er bygningstype, varighet, lokasjon og entrepriseform inkludert. Resterende byggetekniske variabler som Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) fremhever som viktige har det ikke vært mulig å innhente

med hensyn til tilgjengelig tid til å gjennomføre studien. I tillegg har forskeren ønsket å inkludere variabler knyttet til leverandører i prosjektene, både hovedentreprenøren og hovedkonsulenten er inkludert i datasettet. Videre er det inkludert variabler som sier noe om endringer i kontraktene i prosjektene. Her er antallet endringsavtaler i prosjektene, samt endringenes verdi relativt til kontraktsummen inkludert. Variabler knyttet til prosjektets generelle kostnader er også inkludert. Både byggherrens totale generelle kostnader og kostnadene knyttet kun til tidligfase.

Variablene som allerede er forsket på ble hentet inn i den hensikt å gjøre studien sammenlignbar med andre studier av samme karakter. Varighet og lokasjon er heller ikke testet ut på offentlige byggeprosjekter i Norge før. Basert på Mæhlen & Bekkevold (2022) sine forslag til videre forskning er derfor disse inkludert. Når det kommer til variablene knyttet til leverandører, endringer og kontrakter er disse hentet inn for å vurdere i hvilken grad hendelser som skjer i prosjektene i gjennomføringen påvirker sluttkostnaden. Variablene knyttet til generelle kostnader er hentet inn for å vurdere i hvilken grad prosjekter med høyere avvik fra estimatet krever mer eksempelvis av prosjekterende, prosjektleder og byggeleder slik at denne typen kostnader relativt sett utgjør en høyere andel av sluttkostnaden. Innledningsvis er det heller hentet inn for mange enn for få variabler. Hvorvidt antall variabler skal reduseres vil bli diskutert videre knyttet til analysedelen av oppgaven.

De ulike variablene vil være tilgjengelig i ulike deler av prosjektet. Studien er derfor nødt til å være bevisst hvilke variabler som er tilgjengelig i hvilke faser, og hva dette innebærer i forbindelse med bruken av variablene. Som vi kan se i tabell 5 kan vi se for oss tre ulike tidspunkter hvor det kan være interessant å bruke datagrunnlaget. Det første scenarioet er knyttet til estimering i forprosjektfasen før det har blitt utarbeidet en P50 og P85 gjennom andre tradisjonelle metoder. Disse variablene er derimot tilgjengelige i forbindelse med selve investeringsbeslutningen rett før en får GO. På denne måten kan eksempelvis data benyttes som et supplerende verktøy i forbindelse med selve estimeringen. Siste scenario vil være etter at prosjektet er ferdig, og en søker å analysere hvilke variabler som påvirket sluttkostnaden mest.

Tabell 5: Variabler i studien og deres tilgjengelighet

Variabel	Tilgjengelig i forprosjekt	Tilgjengelig ved GO
Seksjon	JA	JA
Region	JA	JA
Prosjekt kategori	JA	JA
Enterpriseform	JA	JA
P85	NEI	JA
P50 (ved terminering)	NEI	NEI
Opprinnelig P50	NEI	JA
Sluttkost	NEI	NEI
Andel usikkerhetsavsetning (P85)	NEI	JA
Andel generelle kostnader	NEI	NEI
Tidligfasekostnad	NEI	JA
Oppstart (år)	JA	JA
Ferdigstilt (år)	NEI	NEI
Varighet (år)	NEI	NEI
Hovedleverandør	NEI	NEI
Hovedkonsulent	NEI	NEI
Antall endringsavtaler	NEI	NEI
Endringer relativt til kontrakt	NEI	NEI

Dette peker i retning av at de variablene som hverken er tilgjengelig i forprosjektfasen eller i forbindelse med investeringsbeslutningen ekskluderes fra datasettene når sluttkostnaden skal predikeres. Det er derimot mulig å benytte alle variablene i forbindelse med vurderingen knyttet til hvilke variabler som påvirker sluttkostnaden. De eneste variablene som ikke er tilgjengelig i tidligfasen, men som likevel inkluderes er variablene knyttet til tid og varighet. I mangel på estimert varighet og ferdigstillelse som er tilgjengelig i tidligfase inkluderes faktisk varighet og ferdigstillelse som substitutter som følge av at dette er vurdert til å være viktige variabler i forbindelse med predikering av sluttkostnaden. Fordelene er altså vurdert til å være høyere enn ulempene. Samtidig er det viktig å være dette bevisst både i forbindelse med tolkningen av resultatene og for videre forskning.

3.5 Analysemetoder

For å gjennomføre analysene som skal føre frem til resultatene i studien har analysedelen blitt delt inn i tre deler. Den første delen har til hensikt å skape en bedre forståelse for de data og variabler som inngår i datagrunnlaget. Denne deskriptive analysen vil også ha som formål å komme frem til hvordan manglende data skal imputeres samt avdekke eventuelle aspekter ved datasettet som kan påvirke analysene, som eksempelvis skjevheter innad i variablene. Den andre delen av analysen vil ta for seg en analyse av datagrunnlaget i den hensikt å beskrive hvilke variabler som påvirker sluttkostnaden mest. Den tredje og siste delen av analysen vil vies til predikering av sluttkostnaden ved hjelp av ulike maskinlæringsmodeller.

3.5.1 Valg av analyseverktøy

Det meste av databehandlingen og analysen er gjennomført i R. Når det kommer til statistisk analyse og maskinlæring står en i hovedsak overfor et valg mellom å gjennomføre analysen ved hjelp av kodespråket Python eller R. Sudhakar (2018) peker på at Python er enklere å bruke, og at det er mer fleksibelt og allsidig. R på den andre siden er bedre egnet til statistisk analyse og er bedre på å visualisere resultater. Siden oppgaven i hovedsak er knyttet mot statistisk analyse, og at visualisering av resultater er et viktig aspekt falt valget på R.

3.5.2 Variablenes påvirkning

For å avgjøre hvilke variabler som har størst betydning for sluttkostnaden tar oppgaven utgangspunkt i tre ulike tilnærminger. Recursive feature elimination (RFE), Regularized random forest og LASSO regresjon. Hensikten med dette valget er todelt. Et aspekt er knyttet til sammenlignbarhet. Ved å benytte flere metoder vil en i større grad kunne sammenligne denne forskningen med fremtidig forskning. Den andre årsaken er at flere ulike tilnærminger danner grunnlag for å sammenligne resultatene fra de ulike modellene for å se hvorvidt det er noen variabler som går igjen i de ulike modellene eller om det eksempelvis er noen variabler som ikke kommer frem som betydningsfulle i noen av modellene. En utfordring Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) peker på er standardisering knyttet til metode og måling i predikering av sluttkostnad i byggeprosjekter. Et moment i denne sammenhengen er også at tidligere forskning i liten grad nevner hva slags metoder som har vært brukt i forbindelse med reduksjon av variabler eller andre typer justering og korrigering av modellen for å oppnå mest mulig resultat. De tre ulike metodene kan alle beskrives som metoder for utvelgelse av variabler. Priyatno & Widiyaningtyas (2024)

beskriver variabel seleksjon som en teknikk som brukes for å identifisere det optimale antall variabler som trengs i et datasett for å effektivt oppdage essensen. I hovedsak handler det om å velge ut de variablene som bidrar mest til å estimere en bestemt variabel, i dette tilfelles sluttkostnaden i et byggeprosjekt.

3.5.2.1 Recursive feature elimination

I følge Priyatno & Widiyaningtyas (2024) er RFE en tilnærming som består i å gradvis fjerne variabler som ikke er viktige for å fastslå den avhengige variabelen. Det er altså en tilnærming der en starter med alle variablene. En sjekker hvor presist modellen klarer å predikere for deretter å fjerne en variabel. Om modellen predikerer like godt fjernes denne variabelen og prosessen starter på nytt med resterende variabler. Hver variabel som er igjen testes i hvert steg. Dette er altså en omfattende prosess som ville tatt lang tid å komme frem til for hånd. Tilnærming kalles backward feature selection fordi den jobber seg bakover fra alle variablene og så langt ned eller bak som modellen finner hensiktsmessig.

3.5.2.2 Regularized random forest

Speiser, Miller, Tooze, & Ip (2023) beskriver derimot RRF som en maskinlæringsmodell som bruker forward feature selection. I motsetning til RFE starter modellen i dette tilfellet med en variabel. Modellen legger til flere og flere variabler til den ikke får mer informasjon ut av variablene. Altså at den ikke klarer å predikere bedre.

3.5.2.3 LASSO regresjon

James, Witten, Hastie, & Tibshirani (2021) peker på at LASSO regresjon har egenskaper som gjør den i stand til å velge ut de viktigste variablene. LASSO regresjon er egentlig en regulariseringsmetode som kan benyttes for å hindre overtilpassing av datasettet til treningsdataen. Regularisering søker å hindre at modellen tilpasser seg treningsdata så godt at modellen ikke er treffsikker på nye data. En introduserer altså bias i modellen for å redusere variansen. I LASSO regresjon introduseres et nytt ledd i regresjonsfunksjonen som reduserer stigningstallet de ulike variablene har. I denne sammenhengen representerer stigningstallet den påvirkningen variabelen har på mål variabelen, altså sluttkostnaden. I motsetning til andre regulariseringsmetoder har LASSO den egenskapen at den kan redusere stigningstallet til 0. Det vil si at effekten av variabelen på sluttkostnaden bortfaller. På denne måten kan en benytte LASSO regresjon til å fjerne de variablene som ikke bidrar til å øke treffsikkerheten i predikeringen.

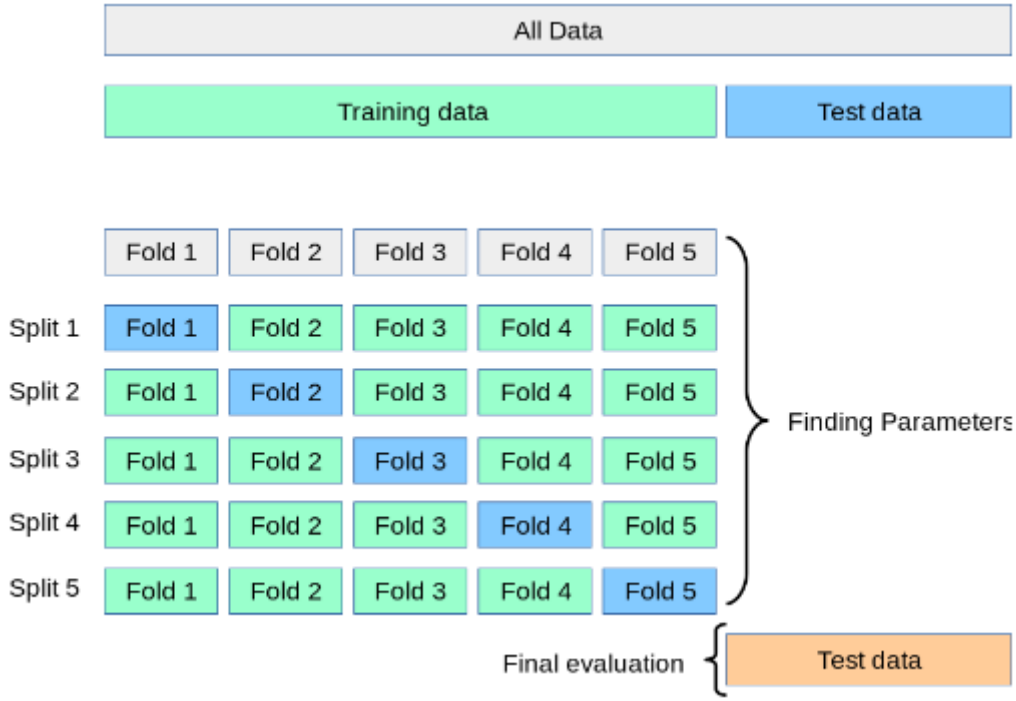
3.5.3 Predikering av sluttkostnad

For å predikere prosjektets sluttkostnad vil oppgaven ta for seg to ulike tilnærminger. Multippel regresjonsanalyse med regularisering og artificial neural network. MRA og ANN er de tilnærmingen som har vært benyttet i høyest grad i forbindelse med predikering av sluttkostnaden i byggeprosjekter i tidligere forskning (Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin, 2022) (Tayefeh Hashemi, Ebadati, & Kaur, 2020).

3.5.3.1 Multippel regresjonsanalyse

En multippel regresjonsanalyse gjennomføres på samme måte som en enkel regresjonsanalyse. Som tidligere nevnt handler det om å estimere koeffisienter knyttet til de ulike variablene inkludert i regresjonsanalysen. Koeffisientene sier noe om hvordan variabel X påvirker variabel Y. Dette innebærer at analysen av natur er relativt enkel og derfor potensielt utsatt for å undertilpasse seg. På den andre siden kan inkludering av mange variabler medføre at modellen overtilpasser seg.

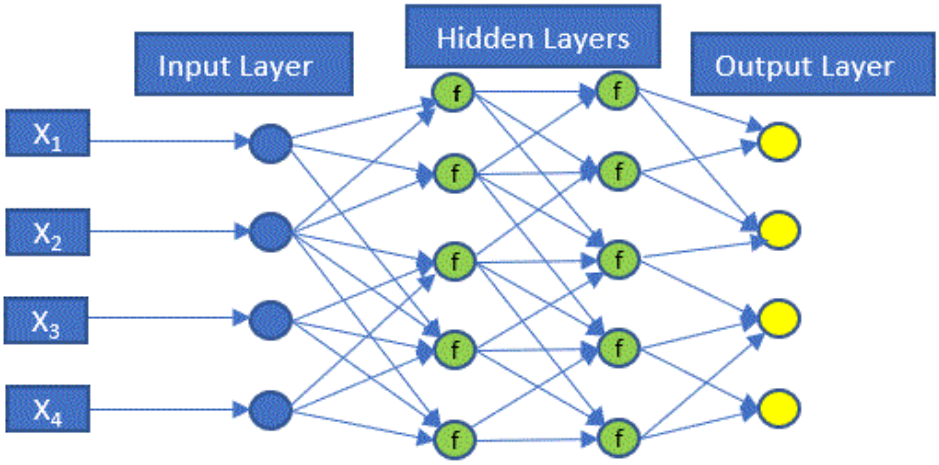
Videre er det en del momenter analysen er nødt til å ta høyde for. Eksempler på dette er regularisering, normalisering, variabel seleksjon, validering og koding av variabler. Alle disse er bygd inn i R sin funksjon LASSO and elastic-net regularized generalized linear models (GLMNET). Ved å bruke denne funksjonen vil R sikre at alle variablene i datagrunnlaget både er numeriske og normaliserte. Regresjonsanalyse krever at alle variablene er numeriske, GLMNET koder derfor om alle de kategoriske variablene til dummyvariabler. GLMNET benytter regularisering i form av både LASSO og en nokså tilsvarende type som heter ridge. Videre er det innebygd kryssvalidering i form av K- fold. Kryssvalidering handler om å dele opp treningsdatasettet flere ganger for å trene flere modeller med ulike kombinasjoner av observasjoner benyttet til trening og validering. K- fold innebærer at en selv kan styre hvor mange modeller som skal settes opp. I figur 16 vises et eksempel på 5- fold kryssvalidering. I GLMNET er antall folds et hyperparameter som gjennom analysen må vurderes for å få en best mulig modell. Kombinasjonen av LASSO og ridge er også et hyperparameter. Hvilke variabler som inkluderes og hva slags splitt ratio en bruker mellom treningsdatasett og testdatasettet er også hyperparameter som må vurderes. Analysen vil derfor søke å justere nevnte hyperparametere på en slik måte at det beste resultatet fremkommer.



Figur 16: K- fold kryssvalidering, (Scikit-learn, U.Å.)

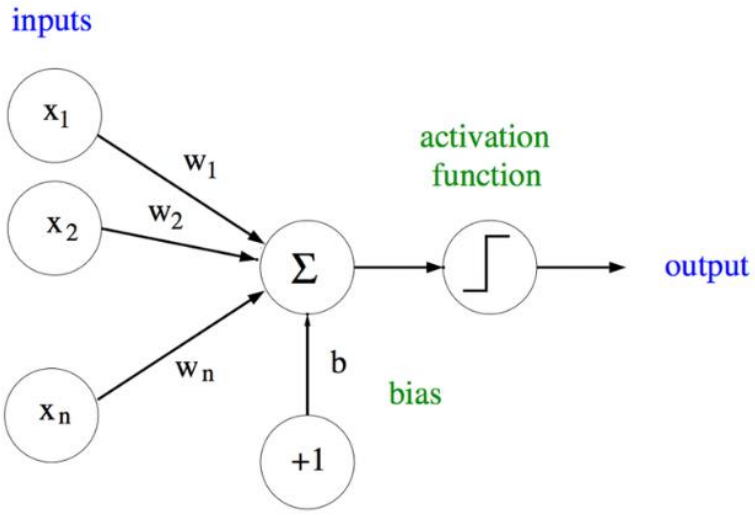
3.5.3.2 Artificial neural network

Et kunstig nevralt nettverk er en mer komplisert form for maskinlæring sammenlignet med MRA. Ifølge Kim, An, & Kang (2004) søker et ANN å simulere læringsprosessen i den menneskelige hjernen. Siden et ANN er inspirert av hjernen er stikkordet mange koblinger. Hammoudi, Moussaceb, Belebchouche, & Dahmoune (2019) kaller ANN en av de viktigste teknikkene innen KI. De peker videre på at ANN har vist seg å være en kraftig metode med gode predikeringsevner som evner å håndtere komplekse og ikke lineære utfordringer. Det finnes mange forskjellige typer ANN noe som kan gjøre det vanskelig å velge hvilket rammeverk som skal benyttes. Denne oppgaven nyttiggjør seg av Keras som er et av rammeverkene for ANN som er tilgjengelig i både R og Python. Ifølge Pierre (2020) er Keras både brukervennlig og et av de ledende rammeverkene for ANN noe som gjør det egnet til bruk i oppgaven.



Figur 17: Artificial neural network, (N'diaye, 2018)

Som vist i figur 17 består et ANN av et innputt lag, et eller flere skjulte lag og et utputt lag. Variablene føres inn i modellen i innputt laget og kommer ut som en prediksjon i utputt laget. De ulike nodene i nettverket kalles nevroner. Et nevron i et lag er koblet til alle nevronene i neste lag osv. Inni nevronene er det i tillegg til informasjonen det får fra andre nevroner vekter, et biasledd og en aktiveringsfunksjon som vist i figur 18. Dette innebærer at det er flere hyperparameter som kan justeres i et ANN sammenlignet med MRA. Vektene og biasleddet styrer nettverket selv. Antall nevroner i hvert lag, antall lag, type aktiveringsfunksjon, antall gjennomkjøringer, valideringssplitt og testsplitt er derimot eksempler på hyperparametere som må tas stilling til i analysen.



Figur 18: ANN nevron, (Galante, 2019)

3.6 Etiske vurderinger

«*Forskning er en kollektiv og systematisk søken etter ny innsikt gjennom bruk av ulike vitenskapelige metoder. Forskning har en verdi i seg selv som kilde til ny og bedre innsikt, og forskning kan være nyttig i mange sammenhenger i samfunnet. Formålet med forskningsetikken er å fremme fri, god og forsvarlig forskning. Forskningsetikken bidrar til å konstituere og sikre god vitenskapelig praksis*» (Haugen , et al., 2021)

Den nasjonale forskningsetiske komité utarbeidet i 2014 et sett med generelle forskningsetiske retningslinjer. Disse handler om:

- Sannhetsbestrebelse
- Forskningens frihet
- Kvalitet
- Frivillig informert samtykke
- Konfidensialitet
- Habilitet
- Redelighet
- God henvisningsskikk
- Kollegiale forhold
- Institusjonens ansvar
- Tilgjengeliggjøring av resultater
- Samfunnsansvar
- Globalt ansvar
- Lover og regler

  (De nasjonale forskningsetiske komiteene, 2019)

Etiske hensyn knyttet til disse retningslinjene vil i mer eller mindre grad gjøre seg gjeldende i en studie. Et aspekt ved denne studien som reduserer de etiske implikasjonene er knyttet til personvern. Studien forsker ikke på mennesker, ei heller inneholder datasettet noen form for personlige opplysninger. På denne måten vil ikke eksempelvis retningslinjen knyttet til frivillig informert samtykke i nevneverdig grad være noe oppgaven behøver å fokusere på.

På den andre siden er det også et par av retningslinjene som særlig bør belyses. Habilitet er knyttet til hvorvidt det kan oppstå interessekonflikter som følge av ulike roller forskeren innehar. Når en skriver en oppgave med arbeidsgiver som case og datakilde kan objektiviteten trekkes i tvil. Klarer en å fokusere på de aspektene ved forskningen som er ment å ha verdi for samfunnet som helhet, eller dras fokuset i oppgaven heller mot å skape verdi for arbeidsgiver? For å unngå disse utfordringene er det viktig å utforme forskningsspørsmål som ikke handler spesifikt om caset en studerer, men heller benytter caset for å belyse forskningsspørsmålene som igjen skal søke å fylle et gap i tidligere forskning. Her kan det nevnes at arbeidsgivers interesser ikke nødvendigvis trenger å være i konflikt med forskningens interesser. I denne studiens tilfelle kan en argumentere for at ved å eksempelvis besvare forskningsspørsmålet som omhandler hvilke variabler som påvirker avviket mellom sluttkostnaden og estimatet i prosjektene vil arbeidsgiver også oppnå dypere innsikt i hvilke aspekter ved prosjektene i deres portefølje det kan være interessant å fokusere nærmere på. Avslutningsvis vedrørende habilitet kan en også si at når caset er en statlig aktør og formålet med oppgaven på et overordnet plan er knyttet til å bedre forvaltningen av samfunnet midler, så kan en hevde at det er i samfunnets interesse at Forsvarsbygg i dette tilfellet får så mye verdi som praktisk mulig ut av oppgaven.

Et annet forskningsetisk aspekt som er relevant for oppgaven handler om konfidensialitet. Hele datagrunnlaget med unntak av tall og tidsvariabler er anonymisert da det samlet sett må anses som skjermingsverdig informasjon jf. Sikkerhetsloven § 5-1 (Justis- og beredskapsdepartementet, 2019). *«Begrepet skjermingsverdig informasjon er en samlebetegnelse som favner all informasjon som skal beskyttes etter loven, av hensyn til de skadefølger som kan påføres nasjonale sikkerhetsinteresser dersom informasjonen blir kjent for uvedkommende, går tapt, blir endret eller blir utilgjengelig»* (Nasjonal sikkerhetsmyndighet, 2020). I denne sammenheng handler det i hovedsak om at uvedkommende ikke skal gjøres kjent med eksempelvis hvilke leverandører som jobber tett med Forsvarsbygg, eller hvor byggeprosjektene er lokalisert og hva slags type bygg det er snakk om. Dette innebærer videre at oppgaven også tar hensyn til retningslinjen som omhandler lover og regler.

Videre kan en hevde at hensynet til konfidensialitet, lover og regler kommer med en pris. Denne betales i eksempelvis etterprøvbarhet. Resultatene vil være tilgjengelige, men når datagrunnlaget ikke i sin helhet er tilgjengelig kan blant annet potensielle årsaker være vanskeligere å forstå og teste ut for andre enn forskeren og virksomheten. Dette innebærer at studien har noe redusert verdi for videre forskning. Kvalitet og redelighet er også etiske hensyn som potensielt blir svekket når data anonymiseres. Disse momentene er åpenbart ikke noe en aktivt søker i en studie. Det er samtidig vanskelig å unngå i all den tid deler av oppgaven må holdes skjult. På den andre siden kan en si at forskningsspørsmålene fortsatt kan besvares uten problemer. Det vil altså si at det potensielle skadeomfanget er av begrenset art.

3.7 Avslutning av metodekapitlet

For å oppsummere metode delen av oppgaven kan vi innledningsvis si at studien er en kvantitativ studie. Datagrunnlaget i oppgaven består totalt av 334 prosjekter. Ved hjelp av de data som er samlet inn knyttet til disse prosjektene skal oppgaven vurdere hvilke variabler som påvirker sluttkostnaden og hvorvidt det er mulig å predikere denne sluttkostnaden.

Datainnsamlingen har hatt et fokus knyttet til å få nok data til å klare å svare ut forskningsspørsmålene uten at størrelsen til utvalget i seg selv skal være et hinder. Maskinlæring krever mye data for å fungere. Både for å lære seg mønster og sammenhenger i treningsdatasettet, men også for å ha nok data til å kunne validere og teste presisjon og predikeringsevne på data modellene ikke tidligere har sett. Samtidig har dette ført til at kvaliteten i datasettet har vært noe lav. Det har derfor blitt brukt en del tid på å ta stilling til datakvaliteten hovedsakelig gjennom håndtering av manglende verdier, fjerning av ekstreme verdier, og fjerning av data som ikke har verdi. Dette har medført at datakvaliteten er vurdert til å være tilstrekkelig. Dette utelukker ikke at kvaliteten i noen tilfeller kan være dårlig. Årsaken til dette er at store deler av datagrunnlaget har blitt dannet manuelt noe som kan medføre menneskelige feil. Denne usikkerheten må studien akseptere og være bevisst da det er vanskelig å vurdere i hvilken grad den potensielt iboende feilen i datagrunnlaget er av betydning.

Som følge av vasking og filtrering av datagrunnlag har oppgaven valgt å dele det opprinnelige datagrunnlaget opp i flere datasett. Årsaken er at forskningsspørsmålene innebærer ulik tilnærming til data. For å predikere sluttkostnaden må en ha nok data. Imputering av data kan derfor være hensiktsmessig. På den andre siden kan imputering medføre feiltolkning når en søker å avgjøre hvilke variabler som påvirker sluttkostnaden. Alle datasett vil derfor benyttes i forbindelse med predikering av sluttkostnaden, mens datasett 2 og 3 vil benyttes for å se på hvordan variablene påvirker sluttkostnaden.

Variablene i utvalget er valgt ut og vurdert basert på tidligere forskning. Det er inkludert både variabler som er testet ut tidligere, men også variabler som det ikke har lyktes forskeren å finne i andre sammenlignbare studier. Dette medfører at studien både er sammenlignbar med tidligere forskning, samtidig som det tilføres noen nye vurderinger til forskningsfeltet. På den andre siden er et av de største negative momentene i oppgaven at det ikke har lyktes forskeren å innhente data knyttet til noen av de variablene tidligere studier peker på som viktige, eksempelvis kvadratmeter.

Videre har oppgaven vurdert variablene basert på når i prosjektets gjennomføring variablene er tilgjengelige. Årsaken til dette er at tilgangen på variabler i ulike faser avgjør hvorvidt de har verdi. Det har eksempelvis ingen hensikt å predikere sluttkostnaden når prosjektet er slutt eller nærmer seg slutten. Predikering kan derimot tilføre verdi i forbindelse med estimering av P50 og P85 i forprosjektfasen, og potensielt i forbindelse med investeringsbeslutningen. På den andre siden kan det være interessant å vurdere hvordan variabler som oppstår i gjennomføringsfase påvirker sluttkostnaden. Som følge av at noen variabler ikke er tilgjengelig når predikering av sluttkostnaden er en relevant aktivitet vil de variabler som først er tilgjengelig etter prosjektets tidligfase ekskluderes fra datagrunnlaget i modellene som skal predikere sluttkostnaden, men inkluderes for å avgjøre hvordan de påvirker sluttkostnaden. Tabell 6 viser hvordan datasettene og variabelen vil benyttes for å besvare forskningsspørsmålene.

Tabell 6: Bruken av datasett og variabler

	Forprosjektfase	Ved prosjektets GO	Ved prosjektets slutt
Datasett 1	Predikering	Predikering	Ikke benyttet
Datasett 2	Predikering +VP	Predikering + VP	Variablenes påvirkning (VP)
Datasett 3	Ikke benyttet	Predikering + VP	Variablenes påvirkning (VP)

For å oppnå en bedre forståelse for variabelens påvirkning på sluttkostnaden benytter oppgaven de tre ulike tilnærmingene Recursive feature elimination, Regularized random forest og LASSO regresjon. Årsaken er både knyttet til sammenlignbarhet, men også for å muliggjøre sammenligning av resultatene fra de ulike modellene. For å predikere sluttkostnaden i prosjektene benyttes både Multippel regresjonsanalyse med regularisering og artificial neural network. Dette er de to metodene som har vært fremtredende i tidligere forskning.

Oppgaven har også tatt stilling til etiske aspekter som påvirker oppgaven i en eller annen grad. Utgangspunktet for vurderingen er de generelle forskningsetiske retningslinjene som de nasjonale forskningsetiske komiteene har utarbeidet. De to aspektene som har vært gjennomgått er habilitet og konfidensialitet. Ingen av aspektene er vurdert til å påvirke oppgaven på en slik måte at forskningen bør trekkes i tvil.

4. Analyse og resultat

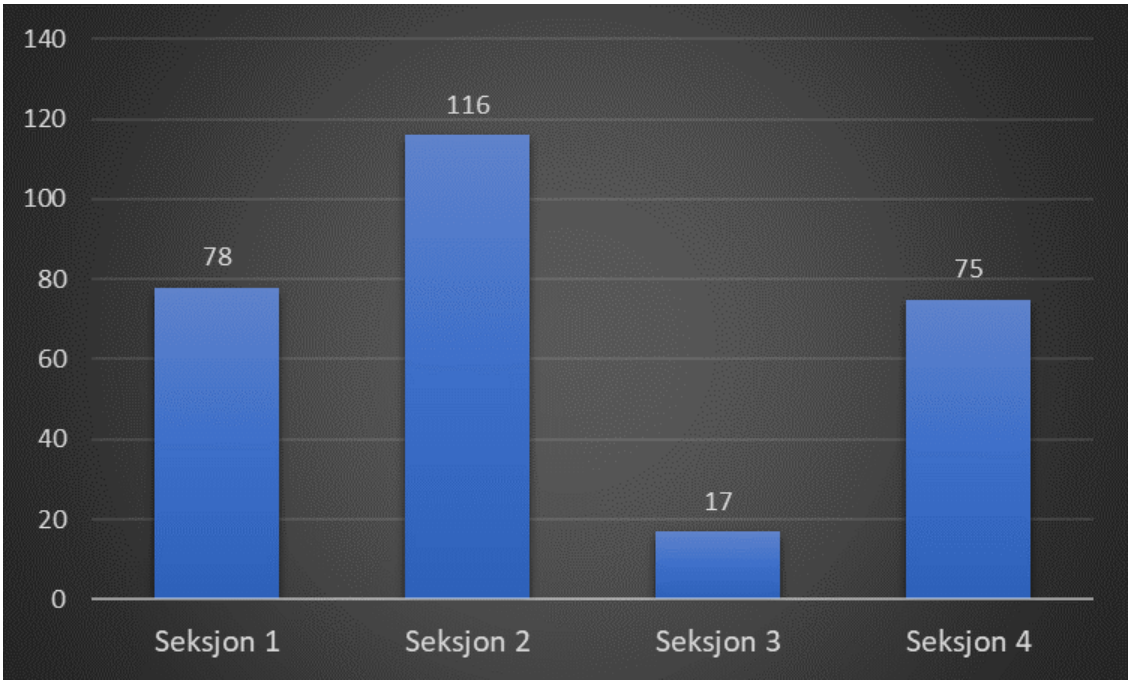
4.1 Deskriptiv analyse

For å få en bedre forståelse av datagrunnlaget, samt beslutte hvordan håndtering av manglende verdier skal gjennomføres vil alle variabler i datasettet gjennomgås. Et viktig moment i denne sammenheng er også hvorvidt de ulike variablene har en fordeling som er gunstig for videre analyse, eller om det eksempelvis er overrepresentasjon av enkelte type observasjoner og skjevheter.

4.1.1 Kategoriske variabler

4.1.1.1 Seksjon

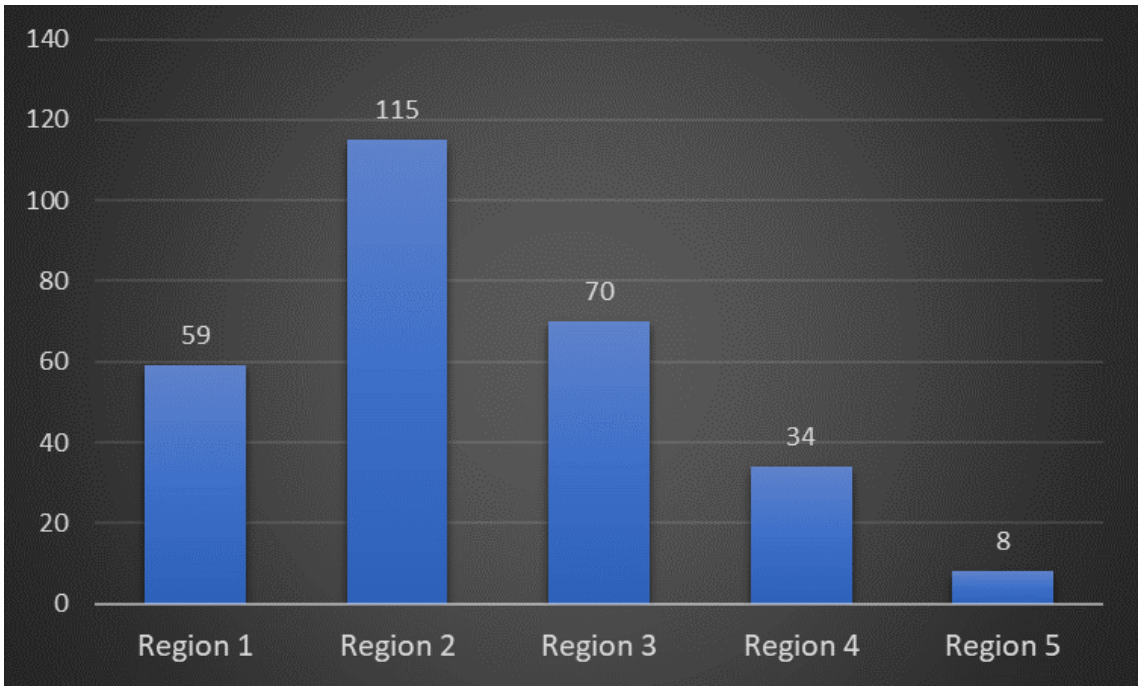
Dette er en variabel av kategorisk karakter. Prosjektene er gjennomført i 1 av 4 seksjoner. Som vi ser ut ifra figur 19 kan vi se at seksjon 1, 2 og 4 er godt representert i datagrunnlaget. Seksjon 3 derimot er nesten ikke representert. Dette kan innebære at prosjekter gjennomført i seksjon 3 vil være vanskeligere å predikere.



Figur 19: Fordeling seksjon

4.1.1.2 Region

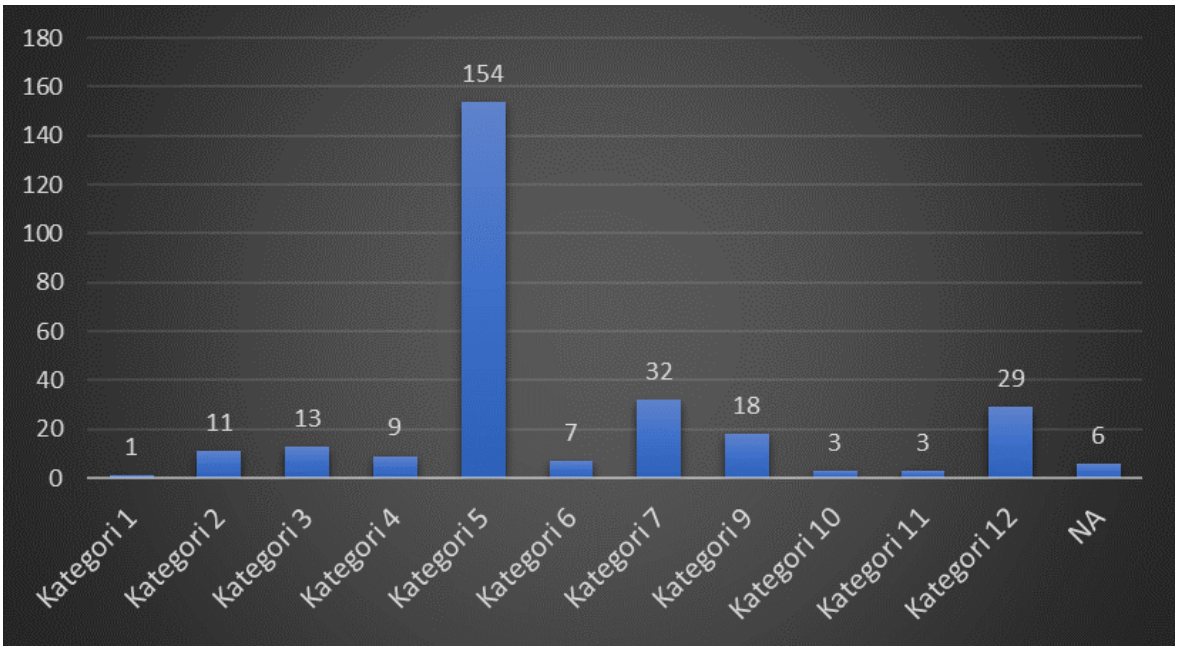
For å beskrive geografisk lokasjon er region benyttet for å snevre ned antall ulike kategorier. Dette er en kategorisk variabel hvor prosjektet er gjennomført i 1 av 5 regioner. Som vi kan se ut ifra figur 20 kan vi se at det er veldig mange prosjekter gjennomført i region 2, mens nesten ingen er gjennomført i region 5.



Figur 20: Fordeling region

4.1.1.3 Prosjektkategori

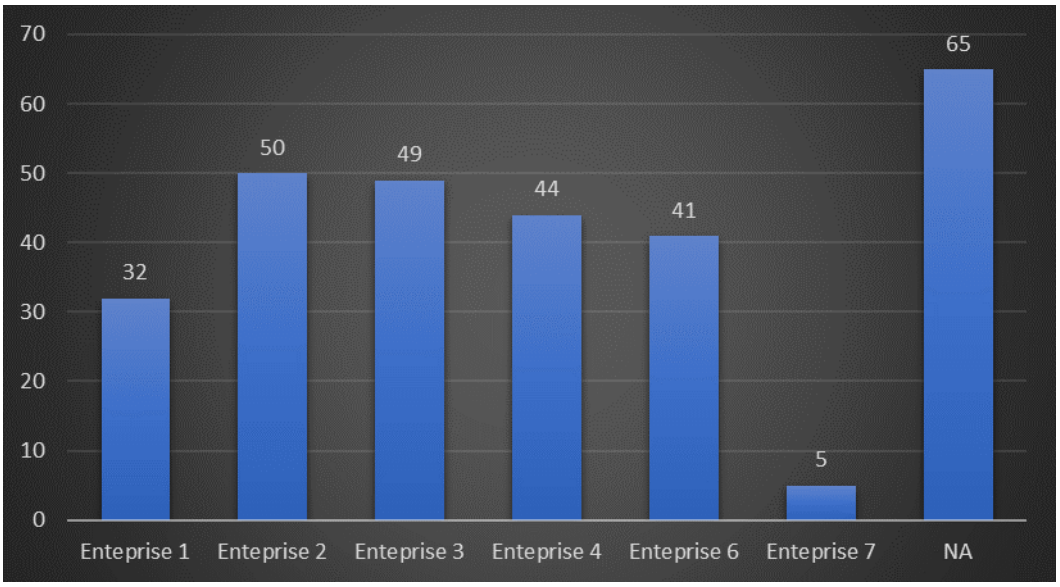
Prosjektkategori er en kategorisk variabel med 11 kategorier. Kategori 8 har blitt borte fra datasettet som følge av fjerning av ekstreme verdier. Variablene har veldig mange observasjoner knyttet til kategori 5, mens det er relativt få observasjoner i de andre kategoriene relativt sett. Dette kan potensielt svekke betydningen av prosjektkategori som variabel. Modellene vil kanskje ikke ha nok data knyttet til de andre kategoriene for å kunne avdekke variabelens betydning. Det er også 6 observasjoner som mangler data. Disse imputeres som kategori 5 observasjoner i datasett 1. Årsaken til dette er at dette er den kategorien med flest observasjoner. Dette vil i liten grad påvirke hvordan modellene vil vurdere et prosjekt i kategori 5 sammenlignet med eksempelvis kategori 1 der kun en erstattet verdi ville stått for 50% av datamengden.



Figur 21: Fordeling prosjektkategori

4.1.1.4 Entrepriseform

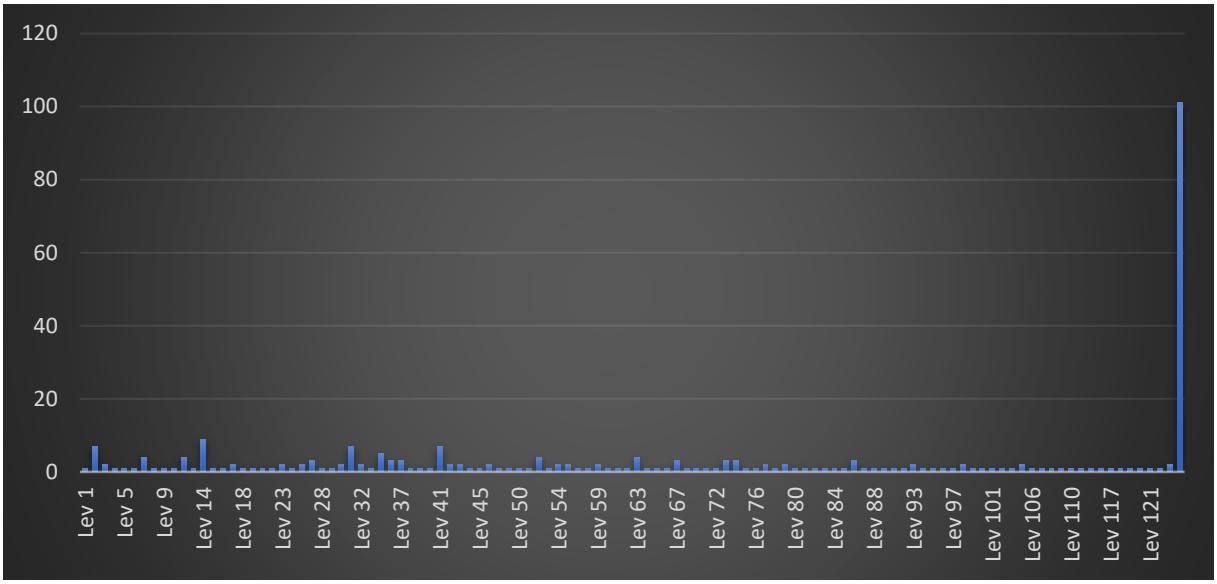
Dette er også en kategorisk variabel, men i motsetning til de tidligere variablene har entrepriseform hele 69 manglende verdier. Basert på fordelingen av resterende observasjoner fordelt på 7 ulike former for entrepriser fremstår det som naturlig å fordele de manglende verdiene relativt jevnt utover de forskjellige kategoriene. Data vil derfor imputeres basert på kategorienes relative andel. Eksempelvis 9 til entrepriseform 1, 15 til entrepriseform 3 og 1 til entrepriseform 7. Tildelingen av manglende verdier vil være tilfeldig.



Figur 22: Fordeling entrepriseform

4.1.1.5 Hovedleverandør

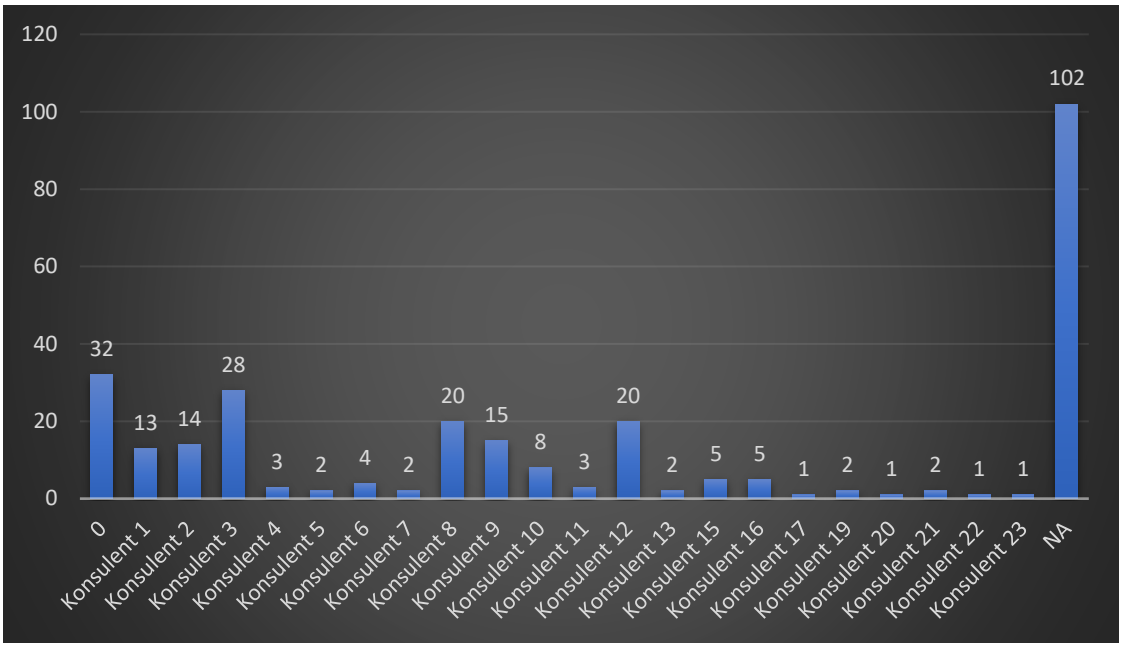
Hovedleverandør er en kategorisk variabel med veldig mange ulike kategorier basert på hvilken entreprenør som har hatt den største kontrakten i de ulike prosjektene. Det er 111 ulike leverandører i datagrunnlaget, og 101 manglende verdier. Dette peker i retning av at denne variabelen ikke vil ha noen verdi for noen av forskningsspørsmålene. Den vil uansett inkluderes i modellene, for en endelig konklusjon knyttet til verdi. Siden det nærmest er en manglende verdi per observasjon, vil imputering gjøres ved å fordele de manglende verdiene jevnt og tilfeldig ut på resterende kategorier.



Figur 23: Fordeling hovedleverandør

4.1.1.6 Hovedkonsulent

Prosjektets hovedkonsulent kan sammenlignes med hovedleverandør i den forstand at det er tilnærmet like mange manglende verdier. I motsetning har hovedkonsulent mye færre kategorier av observasjoner. Dette kan indikere at denne variabelen vil kunne fungere bedre enn hovedleverandør. Det er totalt 22 ulike kategorier med 21 konsulentvirksomheter. Det er også en egen kategori i de tilfeller en har fått bekreftet at det ikke har vært konsulentbistand i prosjektet. Dette er kategorien som heter 0. Imputering vil i dette tilfellet håndteres likt som entrepriseform, basert på kategoriens relative størrelse.



Figur 24: Fordeling hovedkonsulent

4.1.2 Numeriske variabler

Det er en del numeriske variabler i datagrunnlaget. Både kontinuerlige slik som P50 og tidligfasekostnad, men også på intervallform som andel usikkerhetsavsetning og andel generelle kostnader. Intervallvariablene er alle mellom 0 og 1 som prosent. I motsetning til de kategoriske variablene er det flere interessante måleparametere som eksempelvis gjennomsnitt, median, maksverdi og minimumsverdi. Som vist i tabell 7 er det mange manglende verdier i variablene andel generelle kostnader, tidligfasekostnad, antall endringsavtaler og endring relativt til kontrakt. Ved imputering av disse variablene er gjennomsnitt og median naturlige alternativer. For alle disse variablene, kanskje med unntak av generelle kostnader er gjennomsnittet nokså kraftig forskjøvet mot tredje kvartil. Antall endringsavtaler har til og med et gjennomsnitt høyere enn tredje kvartil. Dette indikerer at det er noen ekstreme verdier som drar opp gjennomsnittet. Dette blir også synlig gjennom maksverdiene knyttet til disse variablene. Derfor benytter oppgaven median som følge av at denne vurderes til å representere hele utvalget i større grad enn gjennomsnittet. At gjennomsnittet er dratt mot maks verdi sammenlignet med medianene er noe som går igjen i de fleste av variabelen som ikke skal imputeres også. Eneste unntak er andelen sikkerhetsavsetning og variablene knyttet til oppstart og ferdigstillelse. Det kan også nevnes at ferdigstillelse og oppstart strengt talt ikke er numeriske variabler, men heller datovariabler. De er inkludert under numeriske variabler for enkelhets skyld.

Tabell 7: Fordeling numeriske variabler

Variabel	Min	Første kvartil	Median	Gjennomsnitt	Tredje kvartil	Maks	NA
P85	190	6 825	24 050	45 076	59 025	363 000	0
P50	160	5 800	21 050	40 371	53 450	347 000	0
Opprinnelig P50	160	5 800	20 550	37 701	46 505	284 000	0
Sluttkost	150	5 130	20 050	38 168	52 590	312 317	0
Usikkerhetsavsetning %	0,0 %	7,3 %	12,0 %	11,4 %	15,0 %	45,0 %	0
Generelle kostnader %	0,0 %	15,0 %	22,5 %	25,8 %	31,3 %	100,0 %	22
Tidligfasekostnad	-	120	450	1 272	1 637	13 187	131
Oppstart (år)	2006	2013	2015	2015	2017	2021	0
Ferdigstilt (år)	2010	2018	2020	2020	2021	2023	0
Varighet (år)	0,0	2,3	4,0	4,5	6,0	16,0	0
Antall endringsavtaler	0	0	1	12	9	193	100
Endringer i % av kontrakt	0,0 %	0,0 %	1,0 %	9,3 %	9,8 %	157,0 %	100

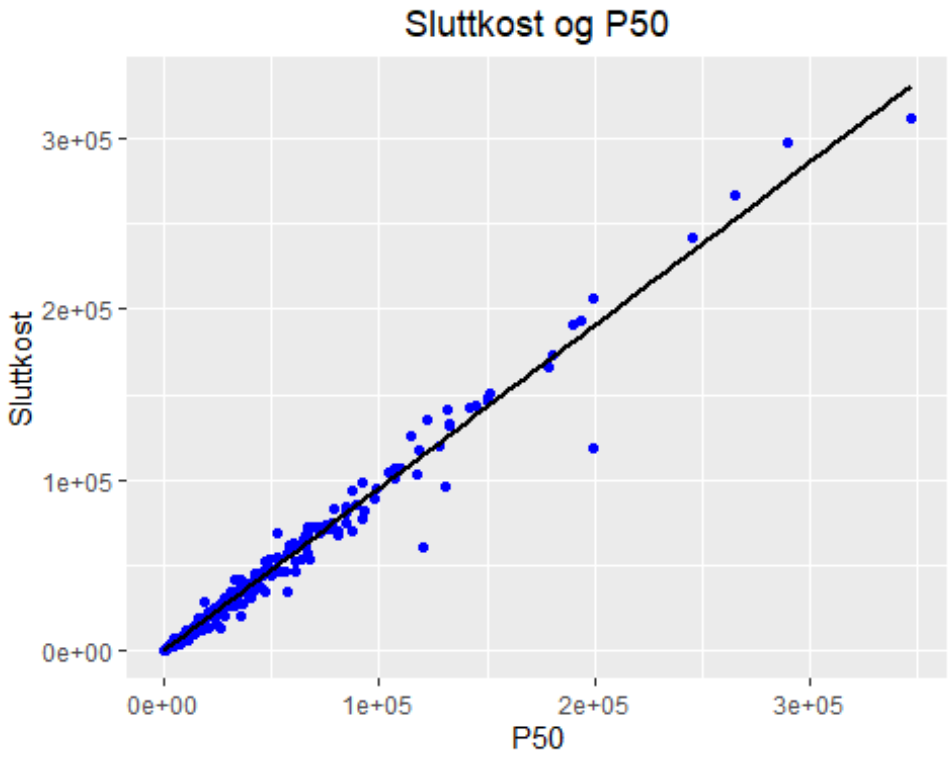
4.1.3 Variabler i forhold til sluttkostnad

For å få en forståelse av hvordan den enkelte variabel varierer i forhold til sluttkostnaden i prosjektene tar oppgaven videre for seg en regresjonsanalyse av numeriske variabler og en evaluering av kategoriske variabler. Interessante refleksjoner og vurderinger vil eksempelvis knytte seg til hvorvidt en kan se om sluttkostnaden øker eller synker på en bestemt måte i forhold til variablene og hvorvidt en kan se mønstre som kan peke i retning av et forhold mellom variabelen og sluttkostnaden. Eksempelvis om sluttkostnaden øker jo høyere andel usikkerhetsavsetning, eller om det ser ut til å være andre sammenhenger enn lineære mellom sluttkostnad og variabel.

Basert på analysen og grafene under er det et par momenter som er interessante. Alle estimatene, altså P50, P85 og opprinnelig P50 har klare og lineære sammenhenger med sluttkostnaden. Dette er naturlig og som forventet da det er lagt ned et godt stykke arbeid i å få disse til å stemme overens med sluttkostnaden. Når det kommer til andel usikkerhetsavsetning og andel generelle kostnader indikerer begge at en høyere sluttkostnad tilsier en lavere andel. Det er derimot ikke et klart mønster. De fleste usikkerhetsavsetningene er fordelt mellom 5 og 20 prosent nokså uavhengig av hva sluttkostnaden var. En kan også se at det også er flere prosjekter uten usikkerhetsavsetning. Når det gjelder andel generelle kostnader ligger de fleste prosjektene mellom 0 og 50 prosent. Observasjonene i dette intervallet ser ut til å ha et normalfordelt mønster. Tidligfasekostnaden stiger i takt med sluttkostnaden basert på observasjonene i utvalget. Det er derimot ingen klar lineær sammenheng. Variablene knyttet til oppstart og ferdigstillelse ser ikke ut til å ha en klar sammenheng med sluttkostnaden, men varighet

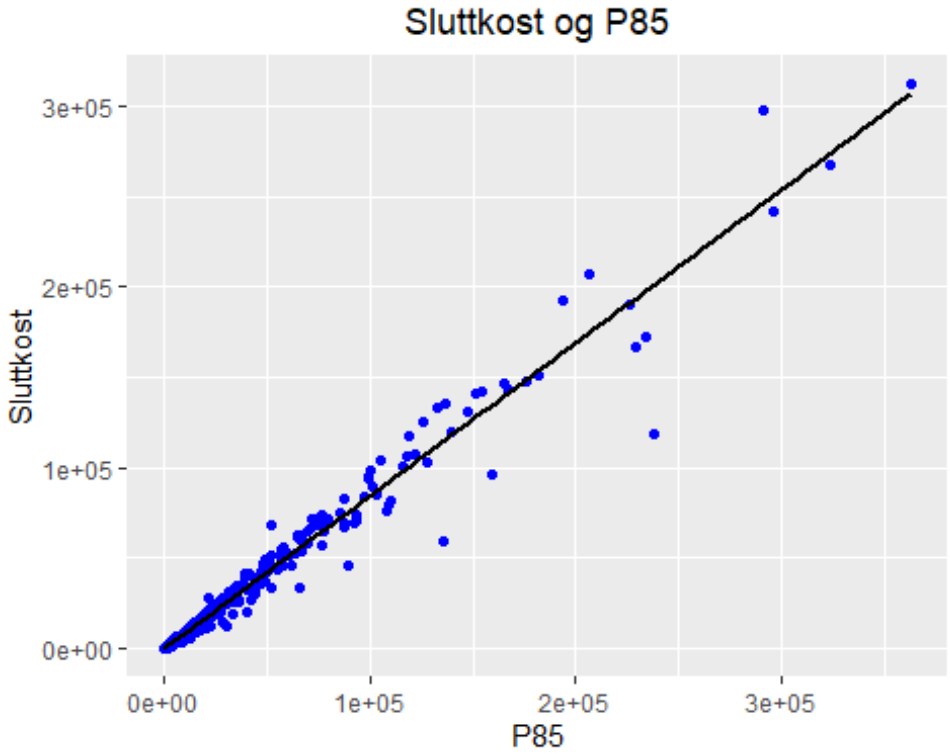
peker i retning av at et prosjekt som varer lenge også vil ha en høyere sluttkostnad enn et som varer kort. Variablene knyttet til endringsavtaler peker begge i retning av at flere endringsavtaler og større verdi på endringsavtalen i forhold til kontraktsum tilsvarer en høyere sluttkostnad. Mønsteret er ikke klart lineært i noen av tilfellene, men det er tydeligst sammenheng mellom antall endringsavtaler og sluttkostnad, sammenlignet med endringsavtalens verdi relativt til kontraktsum. De kategoriske variablene er vanskeligere å vurdere. Det er mulig å se noen mønstre som eksempelvis at entrepriseform 2, 3 og 4 er benyttet ved høyere sluttkostnad. Det er derimot ikke noen klare tendenser knyttet til noen av variablene noe som indikerer at de vil ha liten påvirkning på hvorvidt sluttkostnaden er høy eller lav. Hovedleverandør og hovedkonsulent kan se ut til å være de verste variablene med tanke på sammenheng med sluttkostnad. Dette påvirkes også av at det er veldig mange ulike kategorier med relativt få observasjoner i hver kategori.

4.1.3.1 Sluttkostnad og P50



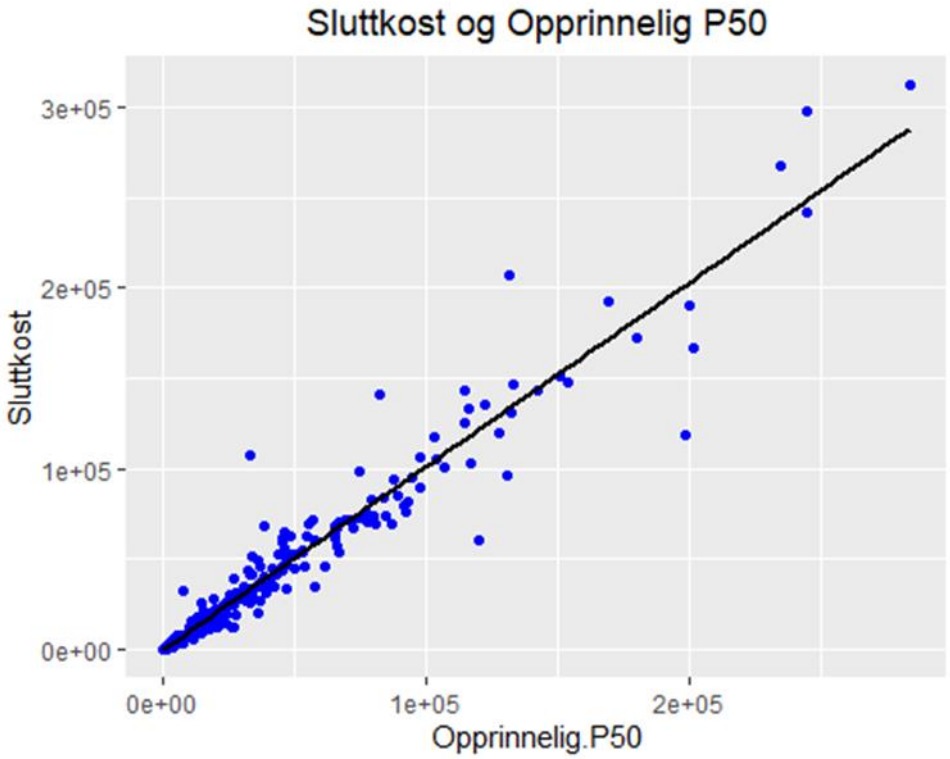
Figur 25: Sluttkostnad og P50

4.1.3.2 Sluttkostnad og P85



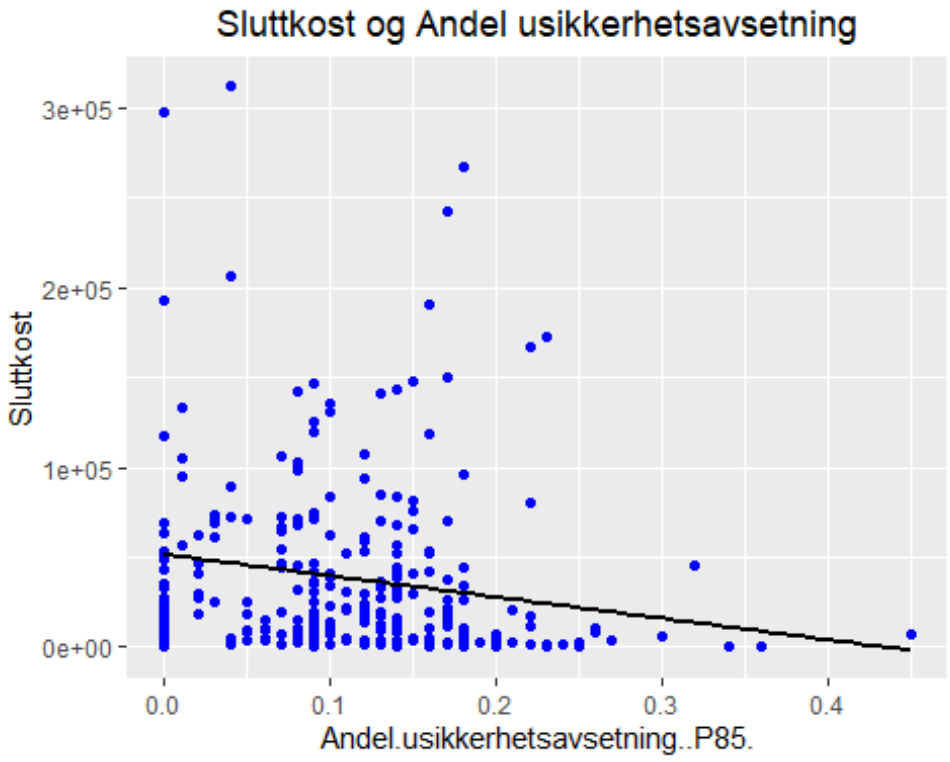
Figur 26: Sluttkostnad og P85

4.1.3.3 Sluttkostnad og Opprinnelig 50



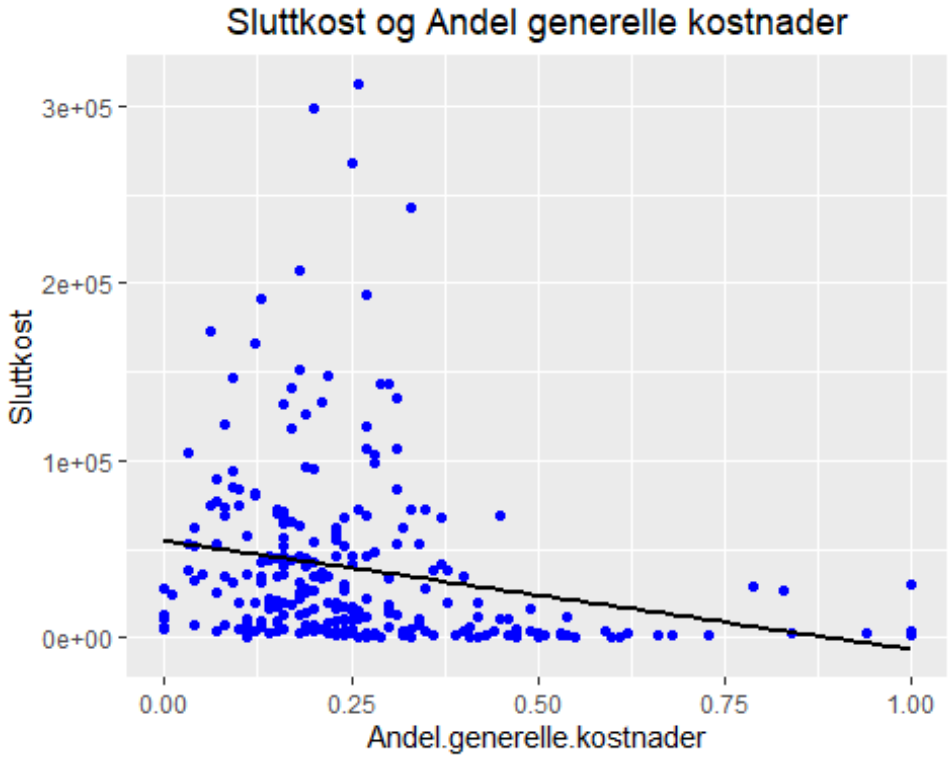
Figur 27: Sluttkostnad og opprinnelig P50

4.1.3.4 Sluttkostnad og andel usikkerhetsavsetning



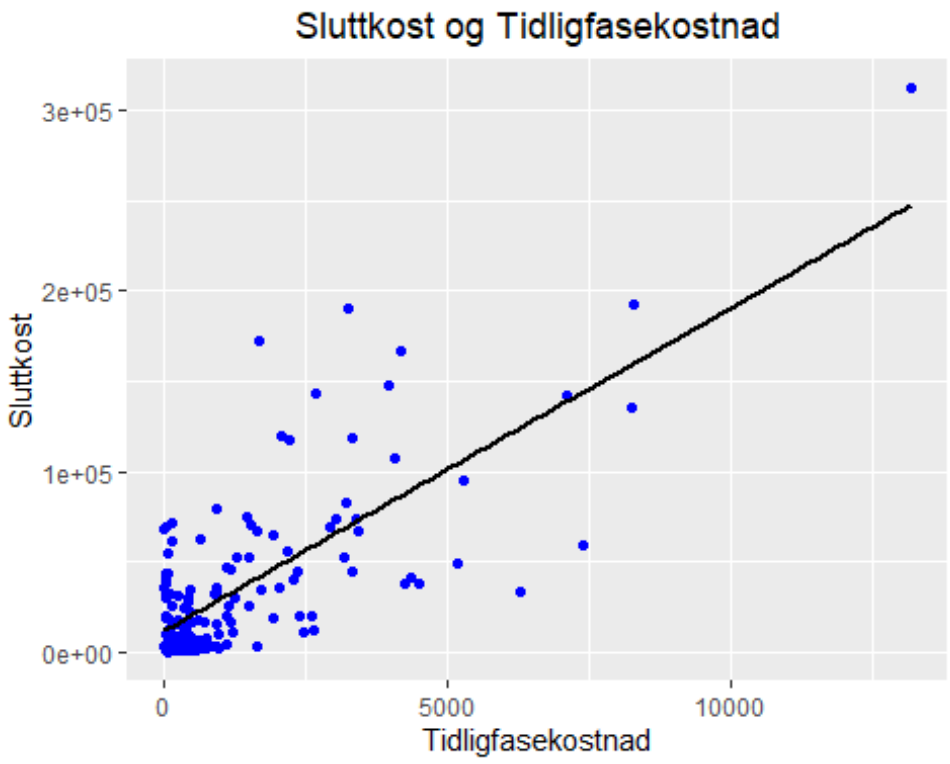
Figur 28: Sluttkostnad og andel usikkerhetsavsetning

4.1.3.5 Sluttkostnad og andel generelle kostnader



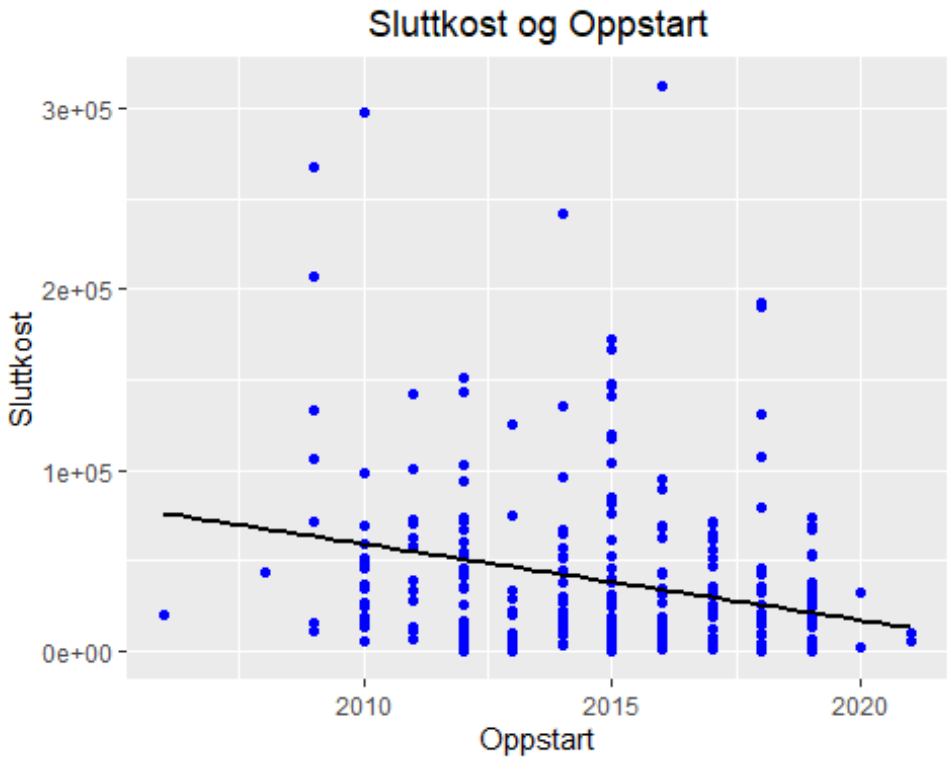
Figur 29: Sluttkostnad og andel generelle kostnader

4.1.3.6 Sluttkostnad og tidligfasekostnad



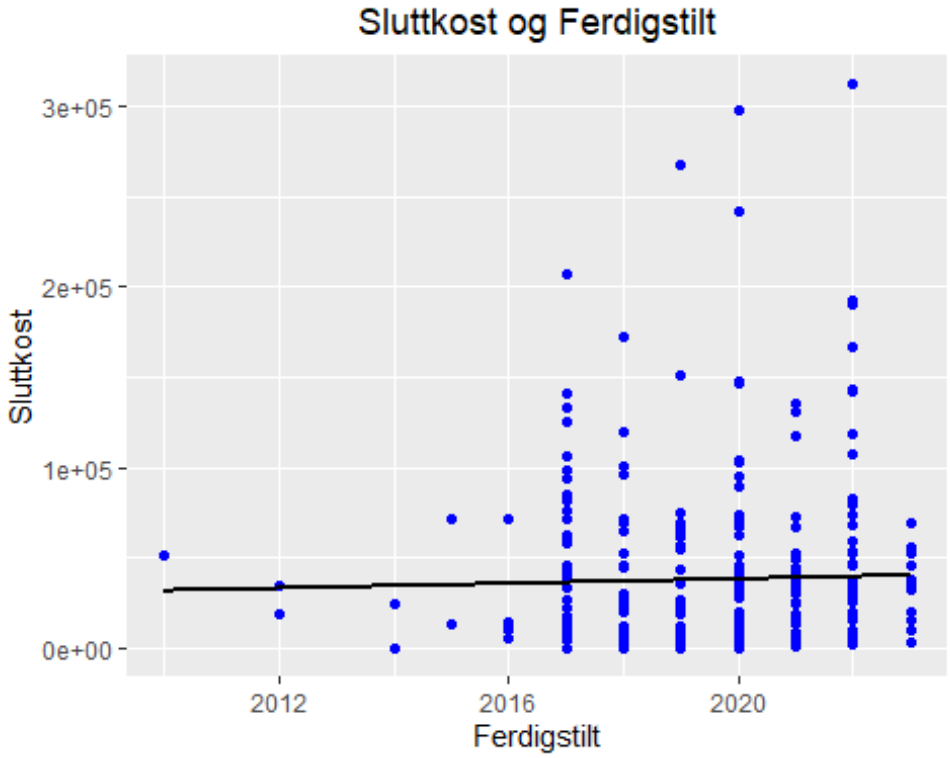
Figur 30: Sluttkostnad og tidligfasekostnad

4.1.3.7 Sluttkostnad og oppstart



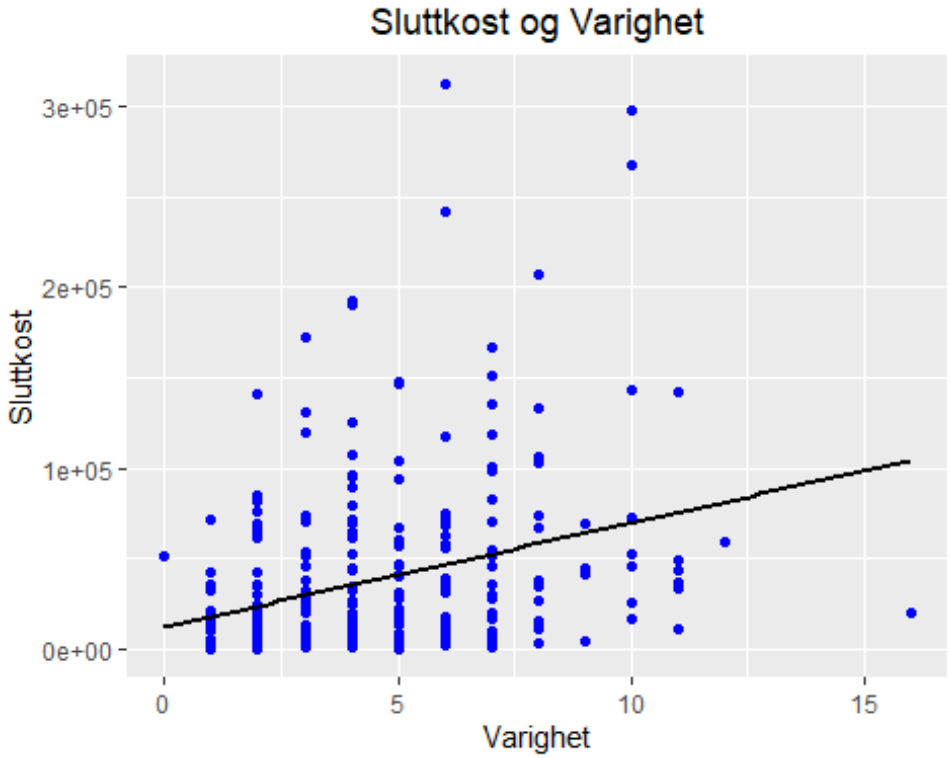
Figur 31: Sluttkostnad og oppstart

4.1.3.8 Sluttkostnad og ferdigstilt



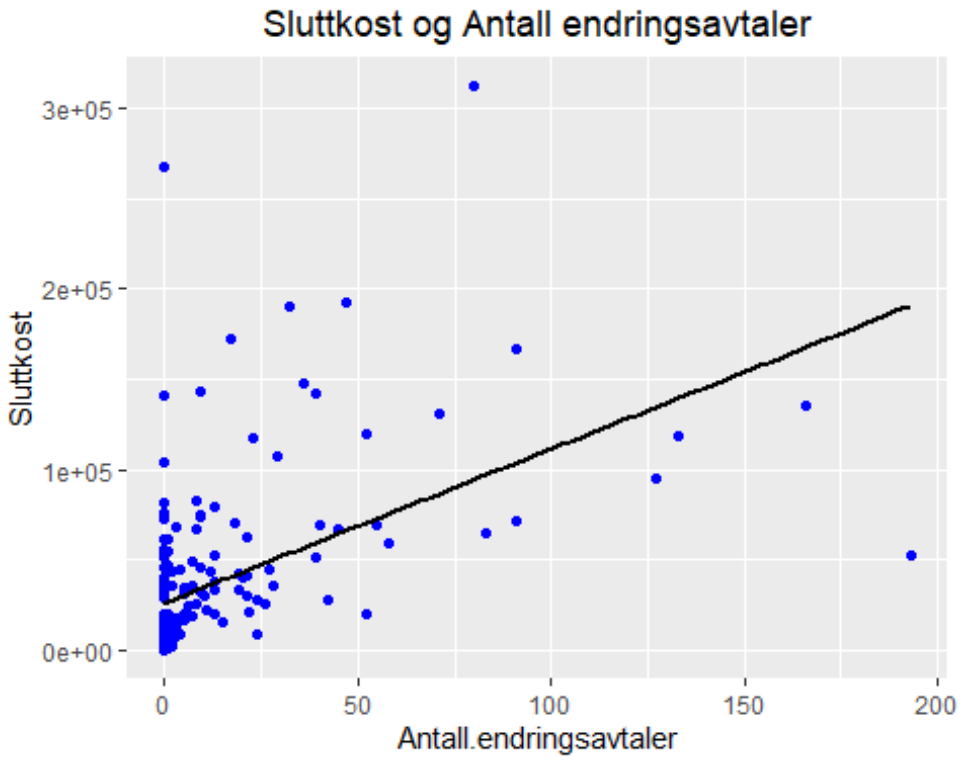
Figur 32: Sluttkostnad og ferdigstilt

4.1.3.9 Sluttkostnad og varighet



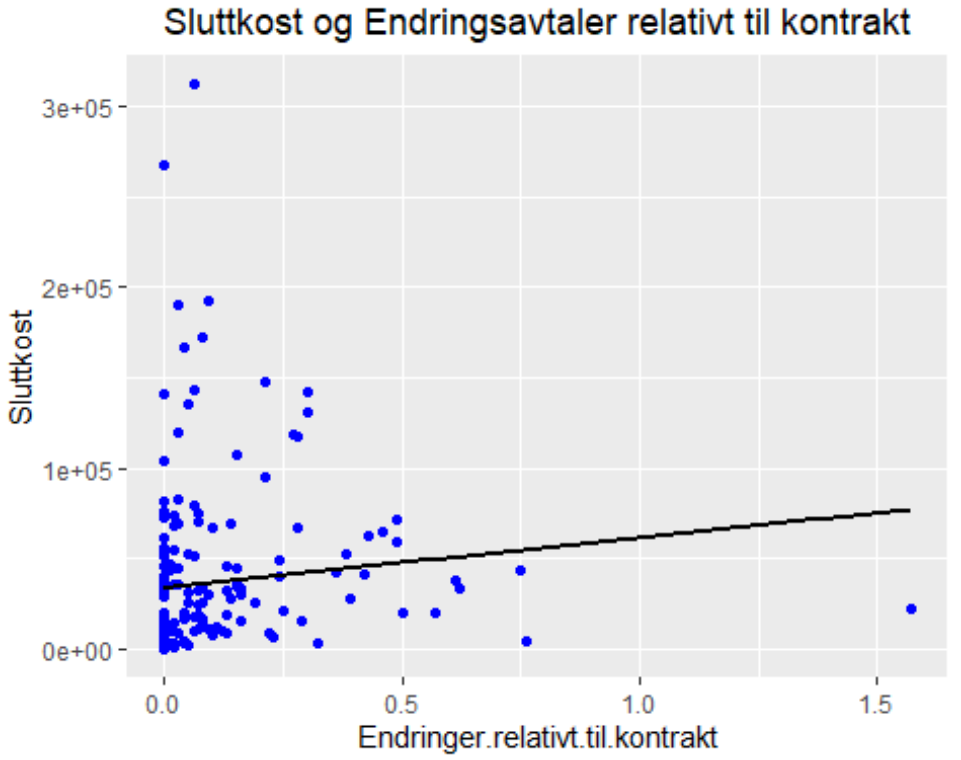
Figur 33: Sluttkostnad og varighet

4.1.3.10 Sluttkostnad og antall endringsavtaler



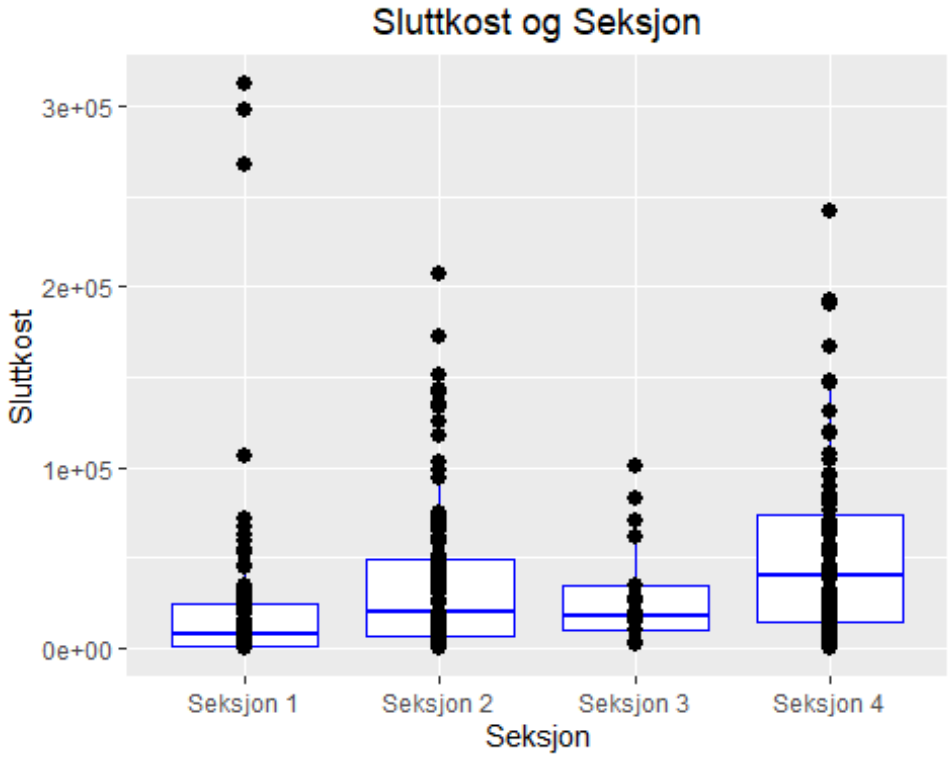
Figur 34: Sluttkostnad og antall endringsavtaler

4.1.3.11 Sluttkostnad og endringer relativt til kontrakt



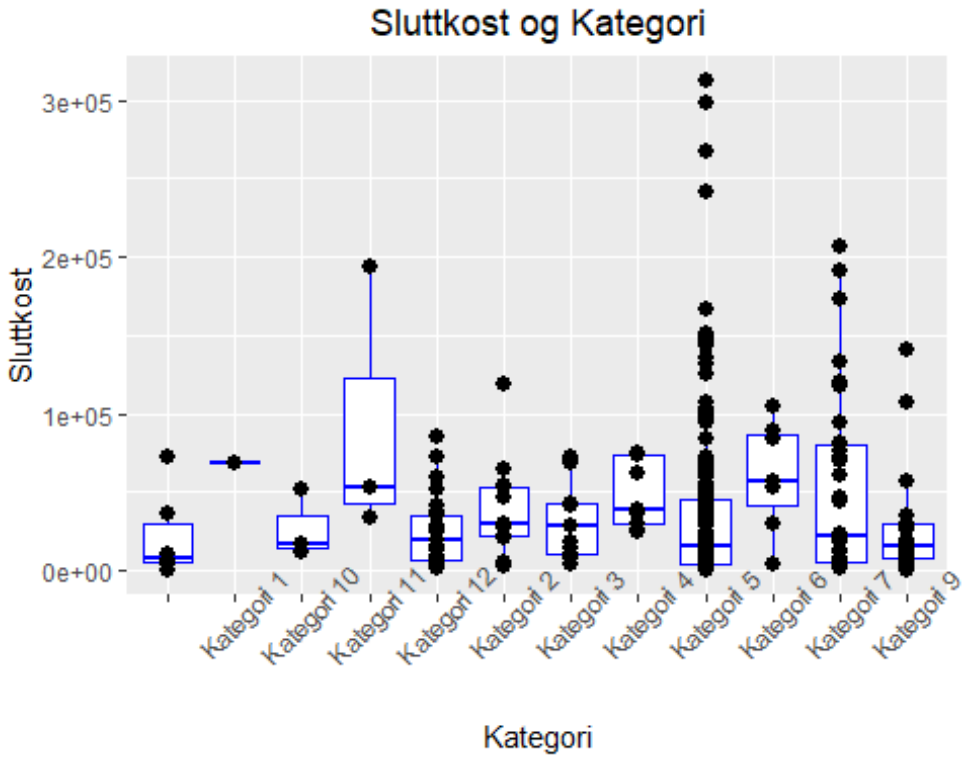
Figur 35: Sluttkostnad og endringer relativt til kontrakt

4.1.3.12 Sluttkostnad og seksjon



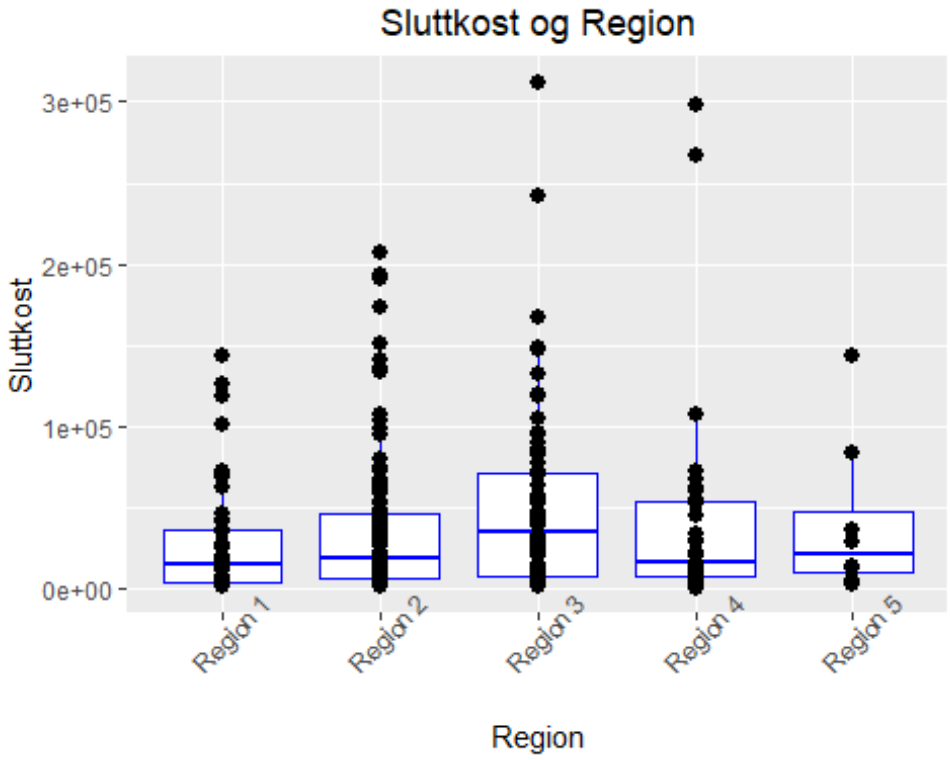
Figur 36: Sluttkostnad og seksjon

4.1.3.13 Sluttkostnad og prosjekt kategori



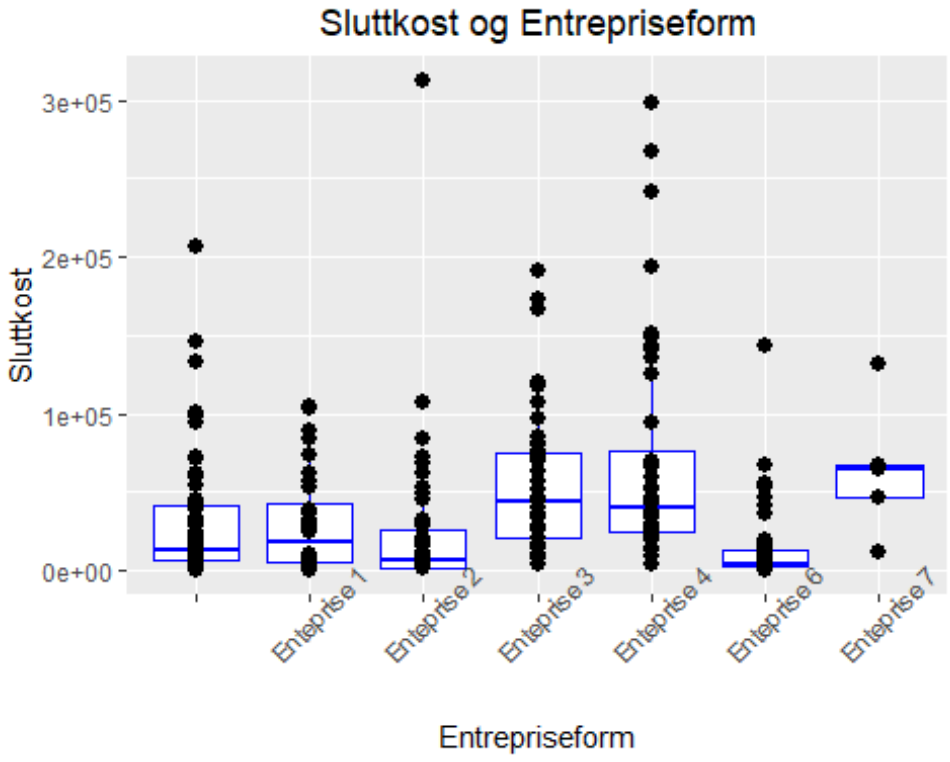
Figur 37: Sluttkostnad og kategori

4.1.3.14 Sluttkostnad og region



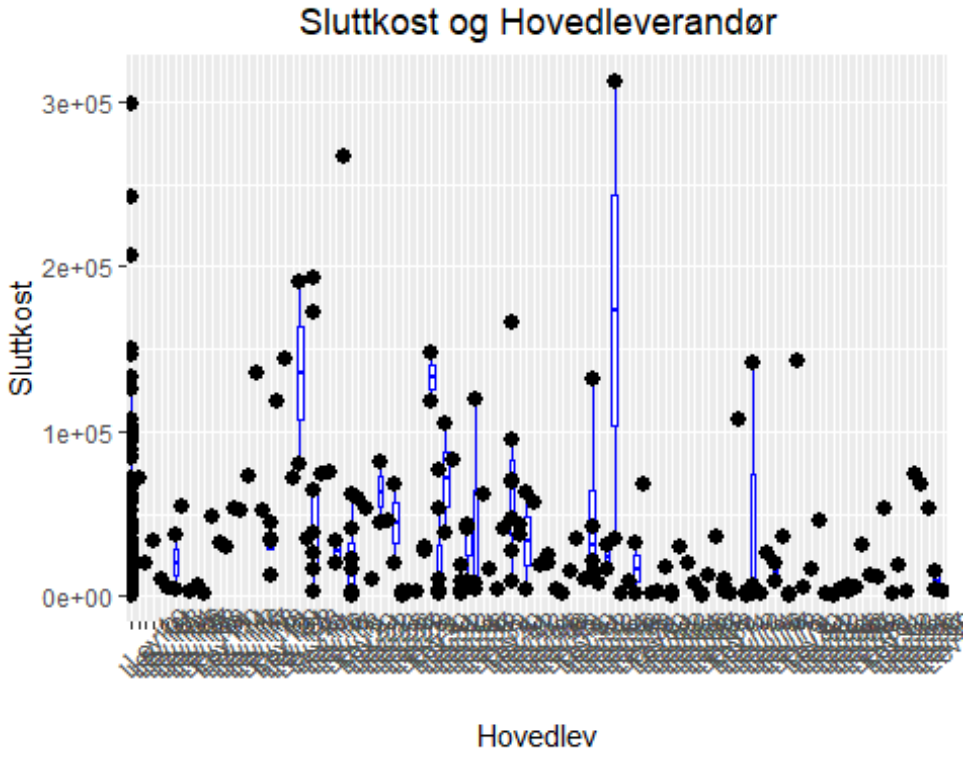
Figur 38: Sluttkostnad og region

4.1.3.15 Sluttkostnad og entrepriseform



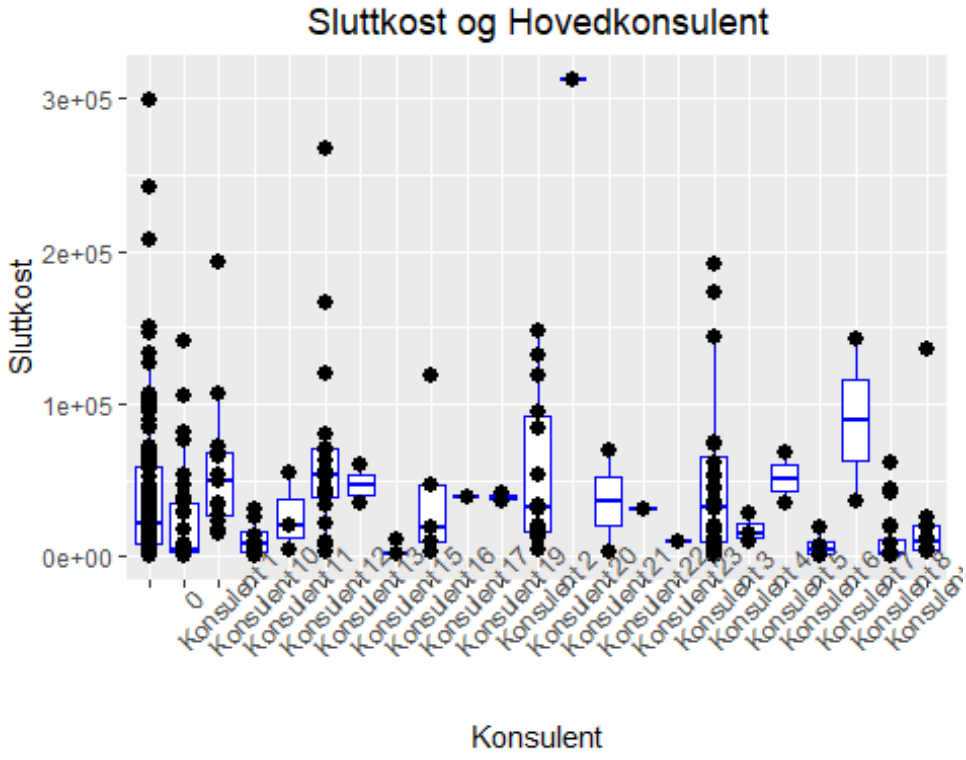
Figur 39: Sluttkostnad og entrepriseform

4.1.3.16 Sluttkostnad og hovedleverandør



Figur 40: Sluttkostnad og hovedleverandør

4.1.3.17 Sluttkostnad og hovedkonsulent



Figur 41: Sluttkostnad og hovedkonsulent

4.2 Prosjektvariablenes viktighet

Andre del av analysen handler om forskningsspørsmålet knyttet til hvilke variabler som er viktigst for å bestemme sluttkostnaden i et offentlig byggeprosjekt. I denne analysen vil oppgaven kun fokusere på datasett 2 og 3 for å unngå at en baserer seg på variabler med høy grad av imputering. I tillegg vil ikke datasett 3 benyttes i forbindelse med variabler kun tilgjengelig i prosjektets forprosjektfase. Dette er fordi datasett 2 inneholder samme variabler, bare med flere observasjoner. Resultatene er fremstilt basert på hvilke variabler de ulike modellene mener er viktigst opp til maks 5 variabler. Som nevnt tidligere er det benyttet recursive feature elimination, regularized random forest og LASSO regresjon.

Tabell 8: Datasett og variabler knyttet til variablenes påvirkning

	Forprosjektfase	Ved prosjektets GO	Ved prosjektets slutt
Datasett 1			
Datasett 2	Predikering +VP	Predikering + VP	Variablenes påvirkning (VP)
Datasett 3		Predikering + VP	Variablenes påvirkning (VP)

Tabell 9: Datasett 2, alle variabler

Metode	Resultat
RFE	P50
RRF	P85, Seksjon (seksjon 2), Kategori (kategori 5), Andel usikkerhetsavsetning, Entrepriseform (entrepriseform 2)
LASSO	P50, Andel usikkerhetsavsetning, Oppstart, Ferdigstilt

Tabell 10: Datasett 2, i forbindelse med GO

Metode	Resultat
RFE	P85 og Opprinnelig P50
RRF	P85, Opprinnelig P50, Entrepriseform (entrepriseform 4), Entrepriseform (entrepriseform 6), Entrepriseform (entrepriseform 2)
LASSO	P85, Opprinnelig P50, Andel usikkerhetsavsetning, Andel generelle kostnader, Oppstart

Tabell 11: Datasett 2, i forbindelse med forprosjekt

Metode	Resultat
RFE	Entrepriseform, Seksjon, Oppstart, Varighet, Ferdigstilt
RRF	Entrepriseform (entrepriseform 6), Varighet, Oppstart, Entrepriseform (entrepriseform 4), Seksjon (seksjon 4)
LASSO	Oppstart, Varighet

Tabell 12: Datasett 3, alle variabler

Metode	Resultat
RFE	P50
RRF	P50, P85, Opprinnelig P50, Tidligfasekostnad, Entrepriseform (entrepriseform 2)
LASSO	P50, Tidligfasekostnad

Tabell 13: Datasett 3, i forbindelse med GO

Metode	Resultat
RFE	P85 og Opprinnelig P50
RRF	P85, Opprinnelig P50, Entrepriseform (entrepriseform 2), Oppstart, Kategori (Kategori 10)
LASSO	P85, Tidligfasekostnad, Andel usikkerhetsavsetning, Varighet, Ferdigstilt

For å oppsummere resultatene er det et par momenter som kan nevnes. Det er et par variabler som ikke inkluderes i noen av modellene hverken når alle variabler er inkludert, ved GO eller i forbindelse med forprosjektfasen. Dette er hovedleverandør, hovedkonsulent, antall endringsavtaler, endringsavtalenes verdi relativt til kontraktssum og region. Datagrunnlaget inkludert i oppgaven peker altså i retning av at disse variablene ikke påvirker sluttkostnaden i nevneverdig grad. Videre kan en se at en gjenganger når alle variablene inkluderes er at P50 er den variabelen som har størst påvirkning på hva sluttkostnaden predikeres til. Andre variabler som går igjen som viktige er P85, tidligfasekostnad, andel usikkerhetsavsetning og entrepriseform. Ved GO Har P85 på mange måter erstattet P50 som den viktigste variabelen, i tillegg til opprinnelig P50. Utover disse ser entrepriseform, andel usikkerhetsavsetning og oppstart ut til å være viktige. Det er mange likheter mellom hvilke variabler som er viktige ved inkludering av alle variabler og ved GO. Estimatene fremstår som

de viktigste. Utover dette er det også få variabler inkludert som er tilgjengelig i forprosjektfasen også. Av variablene som er tilgjengelig i prosjektets forprosjektfase ser entrepriseform, seksjon og variablene knyttet til tid ut til å være de viktigste.

4.3 Predikering av sluttkostnad

I denne delen av analysen vil oppgaven kun fokusere på variablene tilgjengelig i prosjektets forprosjektfase og i forbindelse med GO. Multippel regresjonsanalyse med regularisering og artificial neural network vil som nevnt tidligere benyttes som analysemetoder. Resultatene vil vises i form av RMSE, MSE og MAPE. De vil også sammenlignes med opprinnelig P50 som representerer det estimatet som faktisk var gjeldende i prosjektet ved overgangen fra tidligfase til gjennomføringsfase.

Tabell 14: Datasett og variabler for predikering av sluttkostnad

	Forprosjektfase	Ved prosjektets GO	Ved prosjektets slutt
Datasett 1	Predikering	Predikering	
Datasett 2	Predikering +VP	Predikering + VP	
Datasett 3		Predikering + VP	

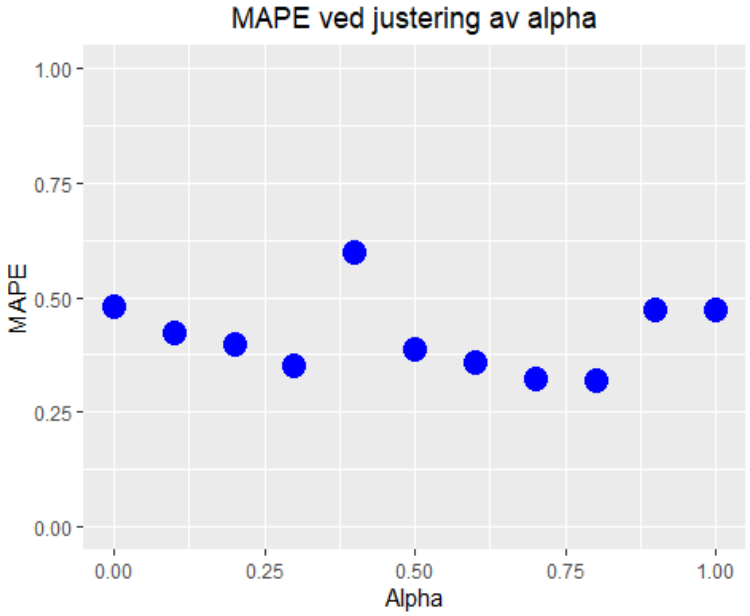
4.3.1 Multippel regresjonsanalyse

For å oppnå den best mulig modellen er en nødt til å justere hyperparameterne. Denne analysen har fire hyperparameter:

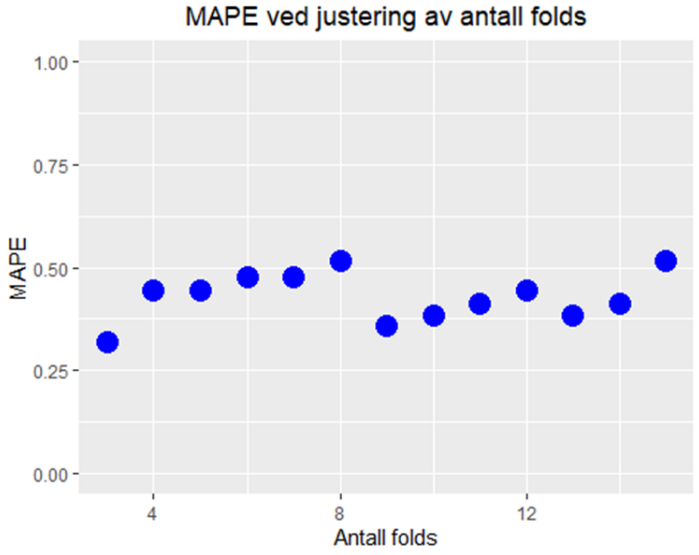
- Antall variabler
- Type regularisering
- Antall kryssvalideringer
- Split ratio mellom treningsdatasett og testdatasett

Når det kommer til antall variabler har oppgaven allerede tatt stilling til dette. Resultatene vil presenteres både basert på variabler tilgjengelig i forprosjektfase og i forbindelse med GO. Siden det benyttes regularisering i form av Lasso og ridge vil det ikke være nødvendig å fjerne flere variabler fra datasettene, dette skal modellen allerede ta hånd om. For å komme frem til riktig form for regularisering justeres alpha. Figur 42 viser hvordan det gjennomsnittlige prosentvise avviket mellom predikert sluttkostnad og faktisk sluttkost (MAPE) endrer seg etter hvert som alpha endres. I dette tilfellet ser en alpha på 0,8 ut til å

være optimalt. Ved å kjøre modellen med forskjellige antall folds til kryssvalidering kan en også se ut ifra figur 43 at MAPE også varierer basert på antall folds. I dette tilfellet ser tre folds ut til å gi det laveste avviket. Når det kommer til forskjellig splitt mellom test og treningsdatasett fremkommer det av tabell 15- 17 at en splitt på 70 prosent til trening og 30 prosent til test gir det laveste prosentvise avviket på 31,2 prosent. Det er riktignok svært liten forskjell mellom de ulike splittene.



Figur 42: Optimalisering av type regularisering, datasett 1 ved GO



Figur 43: Optimalisering av kryssvalidering, datasett 1 ved GO

Tabell 15: MAPE datasett 1 ved GO, 60% splitt

Hyperparameter	60 % splitt
Alpha = 0,1 Folds = 3	56,9 %
Alpha = 0,4 Folds = 5	35,0 %
Alpha = 0,6 Folds = 5	33,3 %
Alpha = 0,7 Folds = 5	**31,5 %**
Alpha = 0,8 Folds = 5	33,1 %
Alpha = 1,0 Folds = 5	39,2 %

Tabell 16: MAPE datasett 1 ved GO, 70% splitt

Hyperparameter	70 % splitt
Alpha = 0,1 Folds = 5	56,9 %
Alpha = 0,4 Folds = 3	36,4 %
Alpha = 0,6 Folds = 3	32,3 %
Alpha = 0,7 Folds = 3	32,2 %
Alpha = 0,8 Folds = 3	**31,2 %**
Alpha = 1,0 Folds = 3	35,8 %

Tabell 17: MAPE datasett 1 ved GO, 80% splitt

Hyperparameter	80 % splitt
Alpha = 0,1 Folds = 3	44,1 %
Alpha = 0,4 Folds = 5	35,1 %
Alpha = 0,6 Folds = 3	32,4 %
Alpha = 0,7 Folds = 3	**31,8 %**
Alpha = 0,8 Folds = 3	32,4 %
Alpha = 1,0 Folds = 3	38,4 %

Tabell 18: Resultater MRA

	DS1 GO	DS2 GO	DS3 GO	DS1 FP	DS2 FP
MAPE	31,2 %	61,6 %	32,3 %	558 %	1011 %
MSE	118 903 829	70 903 999	179 079 913	1 802 046 152	2 564 491 200
RMSE	10 904.30	8 420.451	13 382.07	42 450.51	50 640.81

Tabell 19: Sammenligning med opprinnelig estimat

	DS1 GO	DS1 GO kun Opprinnelig P50 og P85	Faktiske Opprinnelig P50
MAPE	31,2 %	17,7 %	21,6 %
MSE	118 903 829	135 389 552	154 862 988
RMSE	10 904.30	11 635.70	12 444.4

Alle resultatene for resterende datasett har hatt samme tilnærming som for datasett 1. Med en vurdering av hvilken kombinasjon av alpha, folds og splitt som gir best resultat. Som vist i tabell 18 er det datasett 1 i forbindelse med GO som presterer best med en MAPE på 31,2 prosent. Datasett 3 har et nokså likt resultat med 32,3 prosent. Resterende datasett presterer relativt dårlig, spesielt de hvor data kun tilgjengelig i forprosjektfasen er inkludert. Selv det beste resultatet er allikevel ikke i nærheten av å prestere bedre enn det opprinnelige estimatet i prosjektet som ligger på 21,6 prosent. Om en derimot fjerner alle variabler utenom P85 og opprinnelig P50 oppnår analysen en MAPE på 17,7 prosent. Dette kan tyde på at resterende variabler kun tilfører støy som medfører større avvik i prediksjonene.

4.3.2 Artificial neural network

Justeringen av hyperparameter har vært et fokus i denne delen av analysen. Følgende hyperparameter er vurdert:

- Split ratio mellom treningsdatasett og testdatasett
- Slitt av valideringssett
- Antall variabler
- Antall lag
- Antall nevroner i hvert lag
- Valg av aktiveringsfunksjon
- Antall gjennomkjøringer

Tabell 20: Resultater ved ulik testsplitt

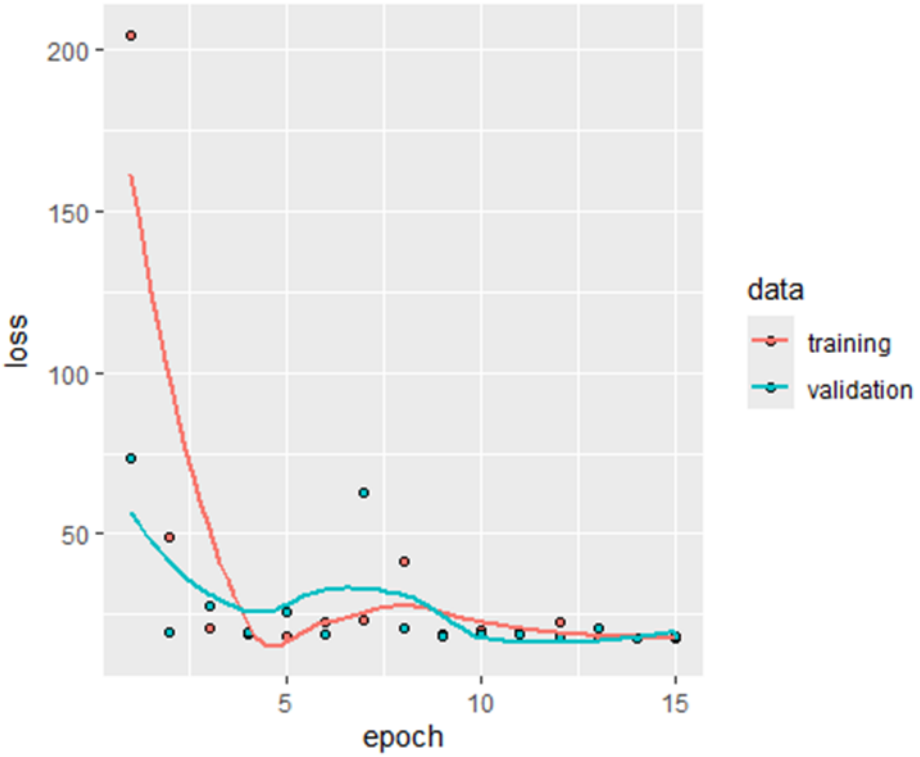
Datasett	60%	70%	80%
Datasett 1 GO	23,6 %	24,5 %	22,5 %
Datasett 2 GO	24,4 %	16,1 %	16,9 %
Datasett 3 GO	27,6 %	17,6 %	24,8 %
Datasett 1 FP	17,6 %	21,6 %	20,7 %
Datasett 2 FP	82,4 %	81,5 %	87,2 %

Tabell 21: Resultater ved ulik valideringssplitt

Datasett	30%	20%	10%
Datasett 1 GO	17,0 %	17,2 %	24,5 %
Datasett 2 GO	15,0 %	18,3 %	16,1 %
Datasett 3 GO	21,9 %	16,5 %	17,6 %
Datasett 1 FP	17,5 %	20,9 %	21,6 %
Datasett 2 FP	82,7 %	82,2 %	81,5 %

Tabell 22: Resultat med ulikt antall lag

Datasett	2 lag	3 lag	4 lag	5 lag
Datasett 1 GO	19,1 %	21,1%	20,8 %	17,4 %
Datasett 2 GO	15,5 %	16,7 %	16,6 %	20,7 %
Datasett 3 GO	25,7 %	16,3 %	17,5 %	22,7 %
Datasett 1 FP	17,6 %	22,1 %	18,9 %	17,7 %
Datasett 2 FP	82,1 %	87,4 %	81,8%	86,3 %



Figur 44: Resultat ved ulikt antall gjennomkjøringer

Basert på resultatene fra tabell 20 kan vi se at de fleste datasettene presterer best med en splitt på 70 til trening og 30 prosent til testing. Videre viser tabell 21 at en valideringssplitt på 30% i hovedsak gir de beste resultatene. Her er derimot ikke spredningen like stor som når det kommer til testsplitten. Spredningen er heller ikke stor når det kommer til resultatene ved ulikt antall lag i modellen. En klar tendens er at flere lag ikke nødvendigvis er bedre. I figur 44 kan en se at resultatet er relativt likt fra 10 gjennomkjøringer og utover. Det er derfor ikke nødvendig å kjøre gjennom modellen så mange ganger før modellen ikke klarer å prestere bedre som følge av flere runder. Som vi også kan se ut ifra figur 44 så ligger resultatet fra trening og validering relativt godt samlet. Undertilpassing virker derfor ikke å være en utfordring. Med tanke på at disse resultatene er relativt like som resultatene fra testdatasettet peker også i retning av at overtilpasning heller ikke er en utfordring. Når det kommer til aktiveringsfunksjon, har ulike funksjoner vært teste ut. Ulike kombinasjoner av arkiveringsfunksjonene rectified linear unit, exponential linear units og leaky rectified linear unit har gitt tilnærmet like resultater. Introduksjon aktiveringsfunksjonene sigmoud, tanh og softmax har vist seg å ikke fungere. Antall nevroner i hvert lag har heller ikke vist seg å være av stor betydning for resultatene. Det første laget har hatt mellom 30 og 50 nevroner i det første laget, før antall nevroner gradvis har blitt nedjustert til mellom 10 og 20 i siste lag.

Tabell 23: Resultater ANN

	DS1 GO	DS2 GO	DS3 GO	DS1 FP	DS2 FP
MAPE	17,0 %	15,0 %	16,3 %	17,5 %	81,5 %
MSE	135 807 047	58257877	130267567	15 431 7620	3 023 934 980
RMSE	11 653.63	7 632.685	11 413.48	12 422.46	54 990.32

Basert på resultatene er det et par momenter som bør nevnes. Datasett 2 i forbindelse med prosjektets forprosjektfase fungerte svært dårlig sammenlignet med resterende datasett. Datasett 1 i forprosjektfasen fungerte derimot tilnærmet likt som resterende datasett. I tillegg kan en se at de variabler som er inkludert i alle datasett i forbindelse med GO som eksempelvis P85 og opprinnelig P50 ser ut til å spille en så stor rolle at det resterende forskjeller tilsynelatende blir borte. Den ekstra datamengden som Datasett 1 har ser ikke ut til å påvirke resultatene i forbindelse med GO, men ser derimot ut til å spille en stor rolle i forbindelse med prosjektets forprosjektfase. Dette indikerer at imputering av verdier har hatt en god effekt for variabler i prosjektets forprosjektfase. Alle de fire første datasettene har bedre resultater enn opprinnelig P50 som har en MAPE på 21.6 prosent.

5. Diskusjon

Diskusjonen vil søke å knytte sammen resultatene med tidligere forskning i lys av problemstilling og forskningsspørsmål. Hva vil det egentlig si at studien har oppnådd de resultatene den har gjort? For å strukturere diskusjonen vil den deles inn i tre deler. Hvor første del handler om variablene, andre del handler om predikering av sluttkostnad. Siste del retter seg inn mot selve problemstillingen i oppgaven, og vurderer derfor i hvilken grad maskinlæring kan bidra til prediktivt og proaktivt fokus i offentlige byggeprosjekter.

5.1 Variablenes påvirkning på sluttkostnad

5.1.1 Tolking av funnene

Analysen peker på en rekke variabler som på en eller annen måte henger sammen med sluttkostnaden. Det er ikke mulig å peke på noen kausalitet i dette tilfellet. Vi vet eksempelvis at tidligfasekostnaden og entrepriseformen henger sammen med sluttkostnaden. Vi kan derimot ikke si at det er en høy tidligfasekostnad som medfører en høy sluttkostnad, eller at når en velger entrepriseform 3 så vil det resultere i en høy sluttkostnad. Kanskje det eksempelvis er kompleksiteten i prosjektet, eller hva slags bygg som skal bygges som egentlig er den kausale sammenhengen. Det er ikke utenkelig at et komplekst og stort bygg vil lede til en høy sluttkostnad. I tillegg høres det logisk ut at et komplekst prosjekt medfører mer tid brukt i tidligfase, noe som igjen vil resultere i høy tidligfasekostnad. På samme måte fører kanskje denne kompleksiteten med seg et økt behov for styring og kontroll i prosjektet, noe som igjen vil styre kontraktstrategien og valg av entrepriseform. På den andre siden kan denne typen utforskning lede til ny innsikt, eller bidra til å bekrefte antakelser. Kanskje det at prosjektkategori ikke viste seg å være av stor betydning kan lede til en gjennomgang av hvordan prosjektene kategoriseres for å kunne gi denne variabelen mer verdi. Et eksempel på en antakelse som kan være bekreftet er knyttet til estimatene P85, P50 og opprinnelig P50. Disse bør henge tett sammen med sluttkostnaden. Alle analyser som inkluderer disse variablene, peker i retning av dette. Det er tross alt estimatene prosjektene kontinuerlig styrer mot i alle valg som gjøres. At disse variablene er viktige peker i retning av at prosjektorganisasjonen er i stand til å estimere, og styre prosjektene på en tilfredsstillende måte.

5.1.2 Sammenligning med tidligere studier

Et av de momentene som bidrar sterkest til å svekke denne studiens verdi er knyttet til datagrunnlaget. Tidligere forskning har pekt på en rekke variabler som er viktige når en skal predikere sluttkostnaden i et byggeprosjekt. Både Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) og Mæhlen & Bekkevold (2022) peker for eksempel på antall kvadratmeter som en viktig variabel. Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) peker videre på en rekke andre variabler som en bør inkludere i en analyse knyttet til sluttkostnaden. Dette er blant annet antall etasjer, fundamenttype, antall bygg eller enheter i bygget, antall heiser, taktype, bygningstype, varighet og lokasjon. Mæhlen & Bekkevold (2022) inkluderer entrepriseform, hvorvidt prosjektet gjelder nybygg eller rehabilitering, estimert kvadratmeterpris og ulike posters relative størrelse av prosjektkostnaden basert på bygningsdelstabellen. Av disse inkluderer kun oppgaven bygningstype, varighet, lokasjon og entrepriseform. En kan derfor argumentere for at studien i liten grad kan sammenlignes med tidligere forskning hva gjelder variabler som er med på å bestemme sluttkostnaden i et byggeprosjekt. Et interessant moment som ikke kan belyses er eksempelvis hvorvidt noen av variablene tidligere forskning peker på er viktigere, eller mindre viktige enn de variablene denne studien har inkludert.

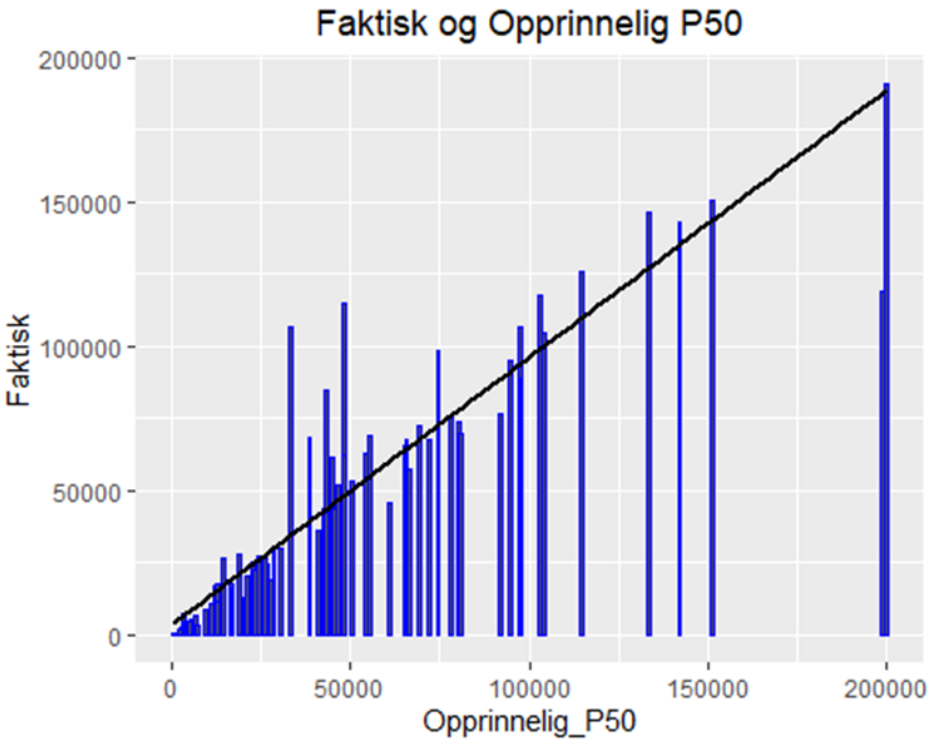
På den andre siden tilfører oppgaven noe nytt til forskningen hva gjelder variabler gjennom prosjektets ulike faser. Tidligere forskning har i hovedsak hatt et fokus på tidligfase, noe som fremstår som naturlig ettersom det er i denne prosjektfasen et presist estimat har størst verdi med hensyn til eksempelvis beslutningsstøtte. Samtidig er det interessant å vurdere om hendelser i prosjektets gjennomføringsfase har stor påvirkning på sluttkostnaden. De variabler som studien inkluderer peker på at de tre estimatene P50, P85 og opprinnelig P50 er de viktigste variablene. Variabler knyttet til valg av leverandør og konsulent, samt endringsavtaler har i liten grad betydning for sluttkostnaden i hvert fall sammenlignet med estimatene. Samtidig innebærer dette ikke at hendelser i gjennomføringsfase ikke er viktige, men heller at det ikke har lyktes forskeren å finne de riktige variablene.

5.2 Predikering av sluttkostnad

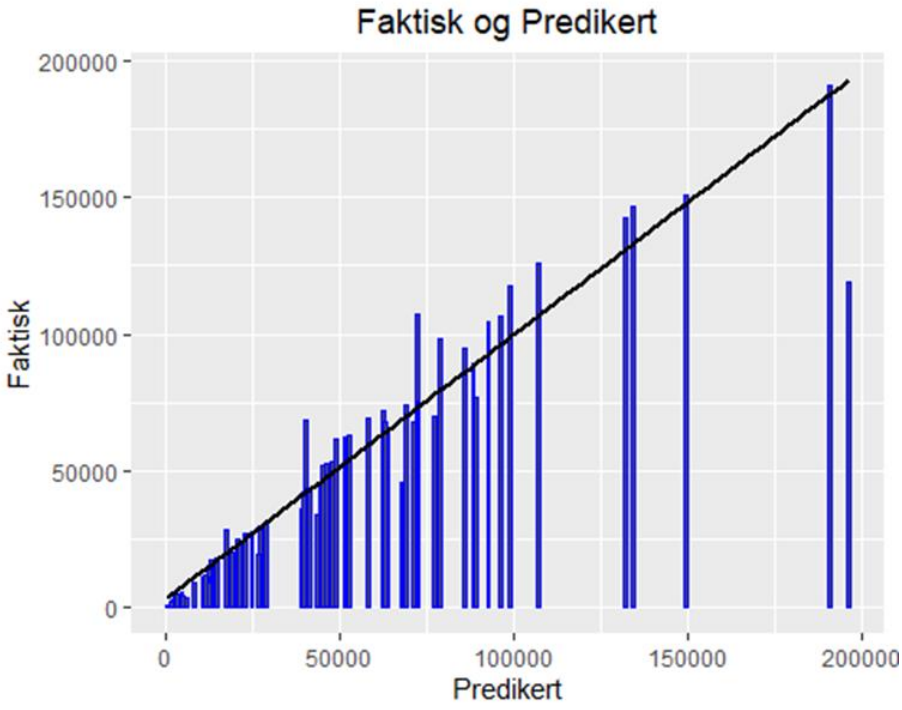
5.2.1 Tolking av funnene

Basert på analyseresultatene er det klart at ANN predikerer bedre enn MRA i dette datasettet. Det beste resultatet fra ANN har en MAPE på 15 prosent, mens det beste resultatet fra MRA er på 31,2 prosent. Resultatet ble riktignok kraftig forbedret ved å kun inkludere P85 og opprinnelig P50. Samtidig kan dette resultatet på 17,7 prosent tilskrives nevneverdig verdi da det er forventet at det skal være en sterk sammenheng mellom P85, opprinnelig P50 og sluttkostnaden. Et annet interessant funn er knyttet til sammenligningen mellom resultatene fra analysen og det opprinnelige estimatet for P50. Opprinnelig estimat på P50 har en MAPE på 21,6 prosent. Det vil si at prosjektene i snitt har avveket med 21,6 prosent fra sluttkosten. ANN klarer å levere resultater som er bedre enn dette både i forbindelse med forprosjektfasen og i forbindelse med beslutning om gjennomføringsoppdrag. At modellen presterer bedre i forbindelse med gjennomføringsoppdrag peker i retning av at de andre variablene utover opprinnelig P50 tilfører verdi til modellen i den forstand at den er i stand til å korrigere seg ned eller opp fra opprinnelig P50 på en god måte. Videre klarer også modellen å predikere bedre i forprosjektfasen, altså uten tilgang på P85, opprinnelig P50, andel usikkerhetsavsetning og tidligfasekostnad. Dette peker i retning av at modellen kan bidra til å utarbeide estimater både i konsept og forprosjektfase.

Et annet spørsmål som også kan være verdt en nærmere undersøkelse er hvorvidt en kan se noen trender knyttet til predikeringen. Om modellen predikerer bedre enn opprinnelig P50 i mindre prosjekter eller store for eksempel. Figur 45 viser opprinnelig P50 sammenlignet med den faktiske sluttkostnaden. De prosjektene som ender opp over den svarte streken har i grovt en sluttkostnad som er høyere enn estimatet og motsatt under streken. Figur 46 viser den samme sammenligningen bare mellom predikert sluttkostnad og faktisk sluttkostnad. Predikerte verdier er basert på variablene tilgjengelig i forprosjektfasen, altså tilsvarende en MAPE på 17,5 prosent. Det kan ved første øyekast se ut til at modellen predikerer nokså likt som opprinnelig P50 med unntak av prosjekter rett i underkant av 50 mill. kroner. Her har opprinnelig P50 et par store avvik. Disse avvikene kan peke i retning av omfangsendringer. Dette kan bidra til å forverre MAPE til opprinnelig P50. En annen tendens som er gjeldende i begge tilfeller er at avvikene fra faktisk sluttkostnad ser ut til å bli høyere etter hvert som sluttkostnaden stiger.



Figur 45: Faktisk sluttkostnad og estimat



Figur 46: Faktisk sluttkostnad og predikert sluttkostnad

5.2.2 Sammenligning med tidligere studier

Som nevnt under litteraturgjennomgangen fremhever Castro Miranda, Del Rey Castillo, Gonzalez, & Adafin (2022) at MAPE i de studiene som ble gjennomgått lå mellom 2 og 21 prosent, med hovedtyngden av studier mellom 5 og 13 prosent. Sammenlignet med resultatene fra tidligere forskning kan en si at resultatene i denne oppgaven i seg selv ikke er veldig gode. MRA ligger langt utenfor øvre intervallgrensen på 21 prosent. ANN ligger derimot godt innenfor dette intervallet, men er litt dårligere enn hovedtyngden av forskningen mellom 5 og 13 prosent. Med tanke på at denne studien ikke inkluderer en rekke av de variablene som studiene det sammenlignes med kan en tenke seg at det er potensiale for å forbedre feilprosenten ved å legge til eksempelvis antall kvadratmeter m.m.

5.3 Svakheter og begrensninger

Av svakheter og begrensinger utover mangel på variabler fra tidligere forskning kan det være verdt å merke seg et par momenter. En svakhet er at den faktiske varigheten på prosjektet er forutsatt til å være lik som den estimerte varigheten. For mange prosjekter vil dette åpenbart ikke stemme. Det kan tenkes at den estimerte varigheten i mindre grad enn den faktiske varigheten henger sammen med sluttkostnaden. Det ble på den andre siden vurdert til at varighet var en såpass viktig variabel å inkludere med tanke på tidligere forskning at gevinsten ved å inkludere varighet i analysene var høyere enn svakheten det medfører.

En annen begrensning er knyttet til prisstigning og endringer i markedssituasjonen generelt. Mest sannsynlig er sluttkostnaden påvirket av prisstigning og sjokk i markedet. Oppgaven tar ikke høyde for disse momentene. Dette kan føre til at viktige sammenhenger ikke blir fanget opp. Et eksempel kan være at sluttkostnaden i prosjektene som har vært gjennomført under COVID- 19 eller mens krigen i Ukraina pågår har vært påvirket av dette. Samtidig kan det tenkes at ved å inkludere året for oppstart og ferdigstillelse har deler av disse momentene indirekte vært inkludert. Videre kan en også argumentere for at variabler knyttet til prisstigning og sjokk i markedet er vanskelig å inkludere når en skal predikere sluttkostnaden i et byggeprosjekt. Da er en nødt til å forutsette en fremtidig tilstand som er svært usikker. Allikevel burde prisstigning og sjokk i markedet vært inkludert for å avdekke betydningen av disse variablene.

5.4 Maskinlæring i offentlige byggeprosjekter

Når det ikke har lyktes forskeren å få tak i data som tidligere forskning har pekt på som viktig, så indikerer dette at forsvarssektorens strategi for KI starter på rett sted. Oppgaven gir dermed en foreløpig status til forsvarssektoren og til andre byggherre organisasjoner i staten knyttet til et avvik mellom de data forskningen peker på som relevante og de data som er tilgjengelig per 2024. På denne måten belyses et gap i tilgjengelig data noe som kan være nyttig innsikt når strategi for KI iverksettes for fullt. Samtidig belyser oppgaven at det ikke er nødvendig å vente til en har alle data en ønsker på plass med kjempehøy datakvalitet. Resultatene fra predikeringen av sluttkostnad viser at selv med en begrenset tilgang på variabler så er det mulig å ta i bruk maskinlæring som verktøy både til beslutningsstøtte og som middel for å tilegne seg ny innsikt. På denne måten kan en si at maskinlæring allerede på nåværende tidspunkt er i stand til å skape verdi i offentlige byggherreorganisasjoner, men at det er et stort potensiale for at denne verdien kan økes.

6. Konklusjon

Konklusjonen deles inn i tre deler. Den første handler om å svare på problemstilling og forskningsspørsmål. Deretter vil oppgaven peke på praktiske implikasjoner av forskningen og hvordan den kan anvendes i praksis. I den siste delen legges det frem forslag til videre forskning som i lys av denne oppgaven kan være interessant å studere nærmere.

6.1 Svar på problemstilling

Denne oppgaven har forsøkt å vurdere i hvilken grad maskinlæring kan bidra til prediktivt og proaktivt fokus i offentlige byggeprosjekter. For å belyse problemstillingen har oppgaven tatt for seg to forskningsspørsmål:

- Hvilke prosjektvariabler bestemmer sluttkostnad en i offentlige byggeprosjekter?
- I hvilken grad kan prediktiv analyse bidra til presise beslutninger i offentlige byggeprosjekter?

6.1.1 Prosjektvariabler

Riktig data er avgjørende for at maskinlæring skal fungere. Om en mater dårlig data inn i en modell kan en heller ikke forvente å få gode resultater ut i andre enden. I tillegg kan kjennskap til hvilke variabler som påvirker sluttkostnaden være viktig for en prosjektorganisasjon i den forstand at en aktiv kan jobbe med å forbedre disse områdene. På denne måten kan en også oppnå proaktivt fokus.

Som tidligere nevnt har det ikke lyktes oppgaven å inkludere mange av de variablene som tidligere forskning har pekt på som viktige. Av de variablene oppgaven har inkludert er det ingen tvil om at estimatene har størst betydning om en raskt skal avgjøre hva sluttkostnaden ender opp på. Dette kommer neppe som en overraskelse, men indikerer samtidig at det er god praksis knyttet til estimering, styring og kontroll i prosjektene. Basert på analysene i oppgaven spiller det videre ikke noen stor rolle for sluttkostnaden hvilken region prosjektet gjennomføres i. Antall endringsavtaler, endringsavtalenes verdi, hovedleverandør og hovedkonsulent er heller ikke av stor betydning. Dette kan indikere at forventede tillegg estimeres på en slik måte at prosjektet tåler endringsavtaler til et visst nivå. Prosjektkategori er en av de variablene tidligere forskning har pekt på som viktige. Denne variabelen har ikke vist seg nevneverdig viktig med tanke på sluttkostnaden i prosjektene. Dette kan være knyttet til utformingen av kategoriene i Forsvarsdepartementet. En mulig årsak til at

prosjektkategori ikke har fremkommet som viktig kan være knyttet til at de fleste prosjektene var i samme kategori. Tidligfasekostnad, entrepriseform og variablene knyttet til tid og varighet er derimot de variabelen utover estimatene som analysene i oppgaven finner som viktigst for sluttkostnaden.

6.1.2 Predikering av sluttkostnaden

For at maskinlæring skal kunne føre til prediktivt fokus er modellen åpenbart nødt til å være i stand til å predikere sluttkostnad på en god måte. Den beste modellen bruker data som er tilgjengelig i forbindelse med beslutning om gjennomføringsoppdrag, og predikerer med en gjennomsnittlig feil på 15 prosent. Modellen som fungerer best til å predikere med variabler tilgjengelig i prosjektets forprosjektfase leverer en gjennomsnittlig feil på 17,5 prosent. Disse resultatene er i samme område som resultater fra tidligere forskning, men i den dårlige enden av skalaen mellom 2 og 21 prosent. Tatt i betraktning et relativt dårlig datagrunnlag er resultatene likevel tilfredsstillende i den forstand at de belyser at det er mulig å predikere sluttkostnaden i norske offentlige byggeprosjekter. Resultatene er bedre enn de estimatene som ble brukt i prosjektene som hadde en gjennomsnittlig feil på 21,6 prosent. Denne feilen kan riktignok være kunstig høy som følge av omfangsendringer. En kan altså si at prediktiv analyse i form av maskinlæring kan bidra til presise beslutninger i offentlige byggeprosjekter. Resultatene er riktignok ikke så mye bedre enn de estimatene som allerede foreligger i prosjektene, noe som indikerer at det foreløpig ikke er grunnlag for å hevde at beslutningene blir mer presise ved bruk av maskinlæring.

6.2 Implikasjoner og anvendelser

Videre er det naturlig å vurdere hva slags betydning denne forskningen har rent praktisk sett. Basert på studien er det et par momenter som trekkes frem:

1. Strategien for KI i forsvarssektoren har et riktig fokus. For å kunne anvende KI på en vellykket måte bør det utvikles og implementeres en datastrategi som omhandler hvilke variabler som bør lagres, hvordan en skal sikre datakvaliteten knyttet til disse variablene, hvordan en kan sikre at data er sammenlignbar over tid og hvordan eksisterende data kan oppdateres for å tilføre et større volum til datamengden.
2. Prosjektorganisasjonen i Forsvarsbygg bør også sikre at KI strategien ivaretar behovene for data knyttet seg spesifikt til prosjektene. Studien anbefaler følgende variabler:

a. Antall kvadratmeter
b. Antall etasjer
c. Fundamenttype
d. Antall bygg
e. Antall heiser
f. Taktype
g. Bygningstype kategorisert på en annen måte
h. Nybygg eller rehabilitering
i. Omfangsendring
j. Grad av kompleksitet

3. Selv om datakvaliteten enda ikke er helt på det nivået en skulle ønske viser studien at maskinlæring kan tas i bruk for å predikere sluttkostnaden i offentlige byggeprosjekter allerede nå. Studien viser at sluttkostnaden kan predikeres i prosjektenes forprosjektfase og i forbindelse med beslutning om gjennomføringsoppdrag. På denne måten kan maskinlæring bistå både i utarbeidelsen av estimatene og som en kontroll, eller et støtteverktøy i etterkant av estimering. Et annet aspekt er at maskinlæring også kan bidra til å effektivisere estimeringsprosessen.

6.3 Forslag til videre forskning

Det er et par momenter som er avdekket i oppgaven, eller som oppgaven ikke har hatt mulighet til å studere nærmere. Disse blir derfor forslag til videre forskning:

1. Hvilke maskinlæringsmodeller er best egnet til å predikere sluttkostnaden i offentlige byggeprosjekter?
    a. Oppgaven har testet ut to modeller. Det finnes derimot veldig mange andre modeller både innenfor MRA og ANN, men også knyttet til andre maskinlæringsmetoder. Om maskinlæring skal implementeres i offentlige byggherreorganisasjoner, bør en teste ut flere modeller for å kunne velge den beste.
2. Fungerer maskinlæring som verktøy for å predikere sluttkostnaden i andre typer offentlige prosjekter som IKT, vei og jernbane?

    a. Oppgaven har kun fokusert på bygging av bygg og anlegg. Andre type prosjekter som IKT- prosjekter, veiprosjekter og jernbaneprosjekter har derfor ikke blitt testet ut. En kan derfor ikke foreløpig konkludere med at maskinlæring kan predikere i andre type prosjekter enn prosjekter knyttet til bygg og anlegg i en norsk kontekst.

3. Hvor godt klarer modellene å predikere ved å inkludere flere variabler tidligere forskning peker på?
    a. Oppgaven har som sagt inkludert de variablene forskeren har vært i stand til å finne strukturert data knyttet til. Et naturlig neste steg er å forbedre de resultatene som er oppnådd i denne studien. En av de tilnærmingene som virker naturlig i denne sammenheng er gjennom å inkludere data som tidligere forskning har pekt på som viktige, men som oppgaven ikke har fått tak i data knytte til.
4. Hvilke andre aspekter ved offentlige byggeprosjekter er det interessant å predikere, og hvor godt kan disse aspektene predikeres?
    a. Oppgaven har fokusert på predikering av sluttkostnaden i offentlige byggeprosjekter. Samtidig vil det være mange andre momenter det kan være interessant å predikere. Et eksempel kan være hvor stort kostnadsavvik et prosjekt vil ha. Et annet eksempel kan være om det er stor risiko for avvik knyttet til tid eller kost i et prosjekt på forhåndsdefinert skala. Predikering av prosjektets varighet kan også være et eksempel.

Referanser

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems, 17(2)*, 1-32.

Abu Hammad, A. A., Ali, S. M., Sweis, G. J., & Sweis, R. J. (2010). Statistical Analysis on the Cost and Duration of Public Building Projects. *Journal of Management in Engineering, 26(2)*, 105–112.

Al mnaseer, R., Al-Smadi, S., & Al-Bdour, H. (2023). Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network. *Asian Journal of Civil Engineering, 24(7)*, 2583–2593.

Arafa, M., & Alqedra, M. (2011). Early stage cost estimation of buildings construction projects using artificial neural networks. *Journal of Artificial Intelligence, 4(1)*, 63-75.

Badawy, M. (2020). A hybrid approach for a cost estimate of residential buildings in Egypt at the early stage. *Asian Journal of Civil Engineering, 21(5)*, 763-774.

Berk, R. A. (2020). *Statistical Learning from a Regression Perspective, 3rd ed.* Cham: Springer International Publishing ; Imprint Springer.

Bjørnenak, T. (2010). Økonomistyringens tapte relevans, del 1 og 2. *Magma 04/10*, 49 – 54.

Castro Miranda, S., Del Rey Castillo, E., Gonzalez, V., & Adafin, J. (2022). Predictive Analytics for Early-Stage Construction Costs Estimation. *Buildings 12(7*, 1-21.

Consept. (U.Å.). *Forskningsprogrammet Concept* . Hentet fra https://www.ntnu.no/concept

Creedy, G. D., Skitmore, M., & Wong, J. K. (2010). Evaluation of risk factors leading to cost overrun in delivery of highway construction projects. *Journal of construction engineering and management, 136(5)*, 528-537.

De nasjonale forskningsetiske komiteene. (2019, Februar 10). *Generelle forskningsetiske retningslinjer.* Hentet fra forskningsetikk.no: https://www.forskningsetikk.no/retningslinjer/generelle/

Digitaliseringsdirektoratet. (2020, Februar 20). *Veileder for beskrivelse av kvalitet på datasett – kvantifiserbar kvalitet.* Hentet fra Digdir: https://data.norge.no/guide/veileder-kvantifiserbar-kvalitet

Digitaliseringsdirektoratet. (2023). *Bruk av kunstig intelligens i offentlig sektor*. Hentet fra Digdir: https://www.digdir.no/rikets-digitale-tilstand/bruk-av-kunstig-intelligens-i-offentlig-sektor/4463

Digitaliseringsdirektoratet. (U.Å.). *Hva er kunstig intelligens?* Hentet fra Digdir: https://www.digdir.no/kunstig-intelligens/hva-er-kunstig-intelligens/4133

Flyvbjerg, B. (2007). Policy and planning for large-infrastructure projects: problems,causes, cures. *Environment and Planning B: Planning and Design, volume 34*, 578-597.

Forsvarsbygg. (U.Å.). *OM OSS.* Hentet fra Forsvarsbygg.no: https://www.forsvarsbygg.no/no/om-oss/

Forsvarsdepartementet. (2019, Desember 10). *Retningslinjer for investeringer i forsvarssektoren.* Hentet fra https://www.fma.no/prinsix/Prosjektmodell/forsvarsstrukturplanlegging

Forsvarsdepartementet. (2023, September 21). *Prop. 1 S, Proposisjon til Stortinget (forslag til stortingsvedtak), FOR BUDSJETTÅRET 2024.* Hentet fra Regjeringen.no: https://www.regjeringen.no/no/dokumenter/prop.-1-s-20232024/id2997658/

Forsvarsdepartementet. (2023, juni). *Strategi for kunstig intelligens for forsvarssektoren.* Hentet fra https://www.regjeringen.no/contentassets/a36197a7d69c45e68186b10117e76b5b/forsvarsdepartementet-kunstig-intelligens.pdf

Forsvarsdepartementet. (2024, 4 5). *Kraftfull satsing på eiendom, bygg og anlegg.* Hentet fra Regjeringen.no: https://www.regjeringen.no/no/aktuelt/kraftfull-satsing-pa-eiendom-bygg-og-anlegg/id3032910/

Forsvarsmateriell. (U.Å.). *PRINSIX Prosjektmodell.* Hentet fra Prinsix: https://www.fma.no/prinsix/Prosjektmodell

Galante, L. (2019). A Comparative Evaluation of Anomaly Detection Techniques on Multivariate Time Series Data. *Hochschule für Wirtschaft und Recht Berlin*, 1-68.

Hammoudi, A., Moussaceb, K., Belebchouche, C., & Dahmoune, F. (2019). Comparison of artificial neural network (ANN) and response surface methodology (RSM) prediction in compressive strength of recycled concrete aggregates. *Construction and Building Materials, Volume 209*, 425-436.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Haugen , H. Ø., Steen-Johnsen , K., Anjum, R., Aardal, B., Bomann-Larsen, L., Fretheim, K., . . . Enebakk, V. (2021, Desember 16). *Forskningsetiske retningslinjer for samfunnsvitenskap og humaniora.* Hentet fra Den nasjonale forskningsetiske komité: https://www.forskningsetikk.no/retningslinjer/hum-sam/forskningsetiske-retningslinjer-for-samfunnsvitenskap-og-humaniora/

IOS 25000. (U.Å.). *ISO 25012.* Hentet fra https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

Jackson, S. (2002). PROJECT COST OVERRUNS AND RISK. *Association of Researchers in Construction Management, Vol. 1*, 99-108.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R, Second edition.* New York: Springer.

Javapoint. (U.Å.). *Bias and Variance in Machine Learning.* Hentet fra Javapoint: https://www.javatpoint.com/bias-and-variance-in-machine-learning

Justis- og beredskapsdepartementet. (2019, Januar 1). *Lov om nasjonal sikkerhet .* Hentet fra Lovdata: https://lovdata.no/dokument/NL/lov/2018-06-01-24/KAPITTEL_5#KAPITTEL_5

Karshenas, S. (1984). Predesign Cost Estimating Method for Multistory Buildings. *Journal of Construction Engineering and Management, 110*, 79–86.

Kim, G.-H., An, S.-H., & Kang, K.-I. (2004). Comparison of construction cost estimating models based on regression. *Building and Environment 39*, 1235 – 1242.

Larsen, A. S., Berg, H., Klakegg, O. J., Welde, M., Langlo, J. A., & Olsson, N. O. (2023, August). Kostnadsestimering i tidlegfase av store offentlege prosjekt – korleis sikre realistiske estimat under høg usikkerheit? Concept-rapport nr. 73 . *Ex ante akademisk forlag*.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management, 132(7)*, 750–758.

Mewara, S. (2020, Desember 6). *Quick look into Machine Learning workflow.* Hentet fra Learn by insight: https://learnbyinsight.com/2020/12/06/quick-look-into-machine-learning-workflow/

Morris, S. (1990). Cost and timeoverrun in public sector project. *Economic and political weekly vol 47*, 154-68.

Mæhlen, J., & Bekkevold, J. P. (2022, Desember 13). DATADREVET USIKKERHETSANALYSE I BYGGEPROSJEKTER. *Holte Consulting - UTARBEIDET FOR CONCEPT-PROGRAMMET VED NTNU*.

Nasjonal sikkerhetsmyndighet. (2020, Juni 3). *Veileder i verdivurdering av informasjon.* Hentet fra NSM: https://nsm.no/regelverk-og-hjelp/veiledere-og-handboker-til-sikkerhetsloven/veileder-i-verdivurdering-av-informasjon/skjermingsverdig-informasjon/

N'diaye, A. (2018, Juli 5). *Machine Learning and Credit Risk (part 5) – Neural Networks.* Hentet fra Analytics R Us(ers): https://analyticsrusers.blog/2018/07/15/machine-learning-and-credit-risk-part-5-neural-networks/

Norsk Prisbok. (U.Å.). *Norconsult Digital AS & Bygganalyse AS.* Hentet fra Norsk Prisbok: https://www.norskprisbok.no/WhatIsNP.aspx

Pham, T. Q., Le-Hong, T., & Tran, X. V. (2023). Efficient estimation and optimization of building costs using machine learning. *International Journal of Construction Management, 23(5)*, 909-921.

Pierre, S. (2020, November 23). *Why Keras Is the Leading Deep Learning API.* Hentet fra Built In : https://builtin.com/artificial-intelligence/why-keras-leading-deep-learning-api

Priyatno, A. M., & Widiyaningtyas, T. (2024). A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS. *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, 196-207.

Regjeringen. (U.Å., Sist oppdatert: 18.09.2023). *Statens prosjektmodell for store investeringer*. Hentet fra Regjeringen.no: https://www.regjeringen.no/no/tema/okonomi-og-budsjett/statlig-okonomistyring/ekstern-kvalitetssikring2/id2523818/

Scikit-learn. (U.Å.). *Cross-validation: evaluating estimator performance.* Hentet fra https://scikit-learn.org/stable/modules/cross_validation.html

Seokyon, H. (2009). Dynamic Regression Models for Prediction of Construction Costs. *Journal of Construction Engineering and Management-asce*, 360-367.

Sharma, V., Zaki, M., Jha, K. N., & Krishnan, N. A. (2022). Machine learning-aided cost prediction and optimization in construction operations. *Engineering, Construction and Architectural Management, 29(3)*, 1241-1257.

Speiser, J. M., Miller, M. E., Tooze, J., & Ip, E. (2023). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 93-101.

Standard Norge. (U.Å.). *Bygningsdelstabell – NS 3451.* Hentet fra https://standard.no/fagomrader/ns-3420-/ns-3450----ns-3451---ns-3459-2/

Sudhakar, K. (2018). Python vs. R Programming Language. *International Journal of Management, IT & Engineering*, 70-79.

Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences, 2*, 1-27.

Valcheva, S. (U.Å.). *Supervised vs Unsupervised Learning: Algorithms and Examples.* Hentet fra Intellspot: https://www.intellspot.com/unsupervised-vs-supervised-learning/

van der Aalst, W. M., Rubin, V., Verbeek, H. M., van Dongen, B. F., Kindler, E., & Günther, C. W. (2010). Process mining: a two-step approach to balance between underfitting and overfitting. *Software and systems modeling*, 87-111.

Williams, T. P. (2003). Predicting final cost for competitively bid construction projects. *International Journal of Project Management 21*, 593–599.

Vedlegg 1: R kode

Kode til utforsking av variabler og MRA

Sett working directory og laste inn pakker

```
rm(list = ls())
setwd(getwd())
library("tidyverse")
library("glmnet")
library("useful")
library("zoo")
library("caret")
library("caretEnsemble")
library("e1071")
library("kernlab")
library("MLmetrics")
library("randomForest")
library("RRF")
library("tidyr")
```

Importering av datasett og visning av toppen av datasettet

```
Master <- read.csv("Masterdata.csv", header = TRUE , sep = ";")
```

Sjekke formatet til variablene

```
str(Master)
```

Vasking av data, og sjekk av format

```
Master$Endringer.relativt.til.kontrakt <- gsub(",", ".", Master$Endringer.relativt.til.kont
rakt)
Master$Endringer.relativt.til.kontrakt <- as.numeric(Master$Endringer.relativt.til.kontrakt
)

Master$Andel.usikkerhetsavsetning..P85. <- gsub(",", ".", Master$Andel.usikkerhetsavsetning
..P85.)
Master$Andel.usikkerhetsavsetning..P85. <- as.numeric(Master$Andel.usikkerhetsavsetning..P8
5.)

Master$Andel.generelle.kostnader <- gsub(",", ".", Master$Andel.generelle.kostnader)
Master$Andel.generelle.kostnader <- as.numeric(Master$Andel.generelle.kostnader)

Master$P85 <- Master$P85/1000
Master$P50 <- Master$P50/1000
Master$Opprinnelig.P50 <- Master$Opprinnelig.P50/1000
Master$Sluttkost <- Master$Sluttkost/1000
Master$Tidligfasekostnad <- Master$Tidligfasekostnad/1000

str(Master)
```

Få en oppsummering av datasettet

```
summary(Master)
```

Lage diagrammer over sammenheng mellom Sluttkost og uavhengige variabler

```
Master %>%
  ggplot(aes(x = P85, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og P85") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = P50, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og P50") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Opprinnelig.P50, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Opprinnelig P50") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Andel.usikkerhetsavsetning..P85., y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Andel usikkerhetsavsetning") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Andel.generelle.kostnader, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Andel generelle kostnader") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Tidligfasekostnad, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Tidligfasekostnad") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Oppstart, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Oppstart") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Ferdigstilt, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Ferdigstilt") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Varighet, y = Sluttkost)) +
  geom_point(colour = "blue") +
```

```
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Varighet") +
  theme(plot.title = element_text(hjust = 0.5))

 Master %>%
  ggplot(aes(x = Antall.endringsavtaler, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Antall endringsavtaler") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Endringer.relativt.til.kontrakt, y = Sluttkost)) +
  geom_point(colour = "blue") +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Sluttkost og Endringsavtaler relativt til kontrakt") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Seksjon, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Seksjon") +
  theme(plot.title = element_text(hjust = 0.5))

Master %>%
  ggplot(aes(x = Kategori, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Kategori") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45))

Master %>%
  ggplot(aes(x = Region, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Region") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45))

Master %>%
  ggplot(aes(x = Entrepriseform, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Entrepriseform") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45))

Master %>%
  ggplot(aes(x = Hovedlev, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Hovedleverandør") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45))

Master %>%
  ggplot(aes(x = Konsulent, y = Sluttkost)) +
  geom_boxplot(colour = "blue") +
  geom_point(size = 2.5, colour = "black") +
  labs(title = "Sluttkost og Hovedkonsulent") +
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = 45))
```

Lage datasett 3

```
Datasett_3 <- na.omit(Master)
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Kategori) & Datasett_3$Kategori != "", ]
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Entrepriseform) & Datasett_3$Entrepriseform != "
```

```
", ]
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Konsulent) & Datasett_3$Konsulent != "", ]
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Kategori) & Datasett_3$Kategori != "Kategori 1",
]
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Kategori) & Datasett_3$Kategori != "Kategori 10"
, ]
Datasett_3 <- Datasett_3[!is.na(Datasett_3$Kategori) & Datasett_3$Kategori != "Kategori 11"
, ]
```

Lage datasett 2

```
Datasett_2 <- Master[, -c(12, 13, 14, 15, 16)]
Datasett_2 <- na.omit(Datasett_2)
Datasett_2 <- Datasett_2[!is.na(Datasett_2$Kategori) & Datasett_2$Kategori != "", ]
Datasett_2 <- Datasett_2[!is.na(Datasett_2$Entrepriseform) & Datasett_2$Entrepriseform != "
", ]
Datasett_2 <- Datasett_2[!is.na(Datasett_2$Kategori) & Datasett_2$Kategori != "Kategori 1",
]
Datasett_2 <- Datasett_2[!is.na(Datasett_2$Kategori) & Datasett_2$Kategori != "Kategori 10"
, ]
Datasett_2 <- Datasett_2[!is.na(Datasett_2$Kategori) & Datasett_2$Kategori != "Kategori 11"
, ]
```

Lage datasett 1

```
Datasett_1 <- Master
GK_median <- median(Datasett_1$Andel.generelle.kostnader, na.rm = TRUE)
Datasett_1$Andel.generelle.kostnader[is.na(Datasett_1$Andel.generelle.kostnader)] <- GK_med
ian

TF_median <- median(Datasett_1$Tidligfasekostnad, na.rm = TRUE)
Datasett_1$Tidligfasekostnad[is.na(Datasett_1$Tidligfasekostnad)] <- TF_median

AE_median <- median(Datasett_1$Antall.endringsavtaler, na.rm = TRUE)
Datasett_1$Antall.endringsavtaler[is.na(Datasett_1$Antall.endringsavtaler)] <- AE_median

EK_median <- median(Datasett_1$Endringer.relativt.til.kontrakt, na.rm = TRUE)
Datasett_1$Endringer.relativt.til.kontrakt[is.na(Datasett_1$Endringer.relativt.til.kontrakt
)] <- EK_median

Datasett_1$Kategori[Datasett_1$Kategori == ""] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 1"] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 10"] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 11"] <- "Kategori 5"

tomme_celler <- Datasett_1$Entrepriseform == ""
ikke_tomme_celler <- Datasett_1$Entrepriseform[!tomme_celler]
relativ_frekwens <- table(ikke_tomme_celler) / length(ikke_tomme_celler)
antall_per_kategori <- round(relativ_frekwens * sum(tomme_celler))
kategorier <- names(antall_per_kategori)
antall_ganger <- as.vector(antall_per_kategori)
tilfeldig_utvalg <- sample(rep(kategorier, antall_ganger))
Datasett_1$Entrepriseform[tomme_celler] <- tilfeldig_utvalg



tomme_celler_konsulent <- Datasett_1$Konsulent == ""
ikke_tomme_celler_konsulent <- Datasett_1$Konsulent[!tomme_celler_konsulent]
relativ_frekwens_konsulent <- table(ikke_tomme_celler_konsulent) / length(ikke_tomme_celler
_konsulent)
antall_per_kategori_konsulent <- round(relativ_frekwens_konsulent * sum(tomme_celler_konsul
ent))
```

```
kategorier_konsulent <- names(antall_per_kategori_konsulent)
antall_ganger_konsulent <- as.vector(antall_per_kategori_konsulent)
tilfeldig_utvalg_konsulent <- sample(rep(kategorier_konsulent, antall_ganger_konsulent))
Datasett_1$Konsulent[tomme_celler_konsulent] <- tilfeldig_utvalg_konsulent



tomme_celler_hovedlev <- Datasett_1$Hovedlev == ""
ikke_tomme_celler_hovedlev <- Datasett_1$Hovedlev[!tomme_celler_hovedlev]
relativ_frekwens_hovedlev <- table(ikke_tomme_celler_hovedlev) / length(ikke_tomme_celler_h
ovedlev)
antall_per_kategori_hovedlev <- round(relativ_frekwens_hovedlev * sum(tomme_celler_hovedlev
))
kategorier_hovedlev <- names(antall_per_kategori_hovedlev)
antall_ganger_hovedlev <- as.vector(antall_per_kategori_hovedlev)
tilfeldig_utvalg_hovedlev <- sample(rep(kategorier_hovedlev, antall_ganger_hovedlev))
Datasett_1$Hovedlev[tomme_celler_hovedlev] <- tilfeldig_utvalg_hovedlev

summary(Datasett_1)
```

Dele datasett 1. 2 og 3 inn i trening og testsett, samt dele trening og test opp i X og Y

```
sample_size_1 <- floor(0.7 *nrow(Datasett_1))
sample_size_2 <- floor(0.7 *nrow(Datasett_2))
sample_size_3 <- floor(0.7 *nrow(Datasett_3))

set.seed(0708)
train_rows_1 <- sample(nrow(Datasett_1), size = sample_size_1)
train_rows_2 <- sample(nrow(Datasett_2), size = sample_size_2)
train_rows_3 <- sample(nrow(Datasett_3), size = sample_size_3)



Datasett_1_train <- Datasett_1[train_rows_1, ]
Datasett_1_test <- Datasett_1[-train_rows_1, ]

Datasett_2_train <- Datasett_2[train_rows_2, ]
Datasett_2_test <- Datasett_2[-train_rows_2, ]

Datasett_3_train <- Datasett_3[train_rows_3, ]
Datasett_3_test <- Datasett_3[-train_rows_3, ]

FP_Datasett_3_train <- Datasett_3_train[, -c(9:18)]
FP_Datasett_3_test <- Datasett_3_test[, -c(9:18)]
FP_Datasett_2_train <- Datasett_2_train[, -c(9:18)]
FP_Datasett_2_test <- Datasett_2_test[, -c(9:18)]
FP_Datasett_1_train <- Datasett_1_train[, -c(9:18)]
FP_Datasett_1_test <- Datasett_1_test[, -c(9:18)]

GO_Datasett_3_train <- Datasett_3_train[, -c(13:18)]
GO_Datasett_3_test <- Datasett_3_test[, -c(13:18)]
GO_Datasett_2_train <- Datasett_2_train[, -c(12:13)]
GO_Datasett_2_test <- Datasett_2_test[, -c(12:13)]
GO_Datasett_1_train <- Datasett_1_train[, -c(13:18)]
GO_Datasett_1_test <- Datasett_1_test[, -c(13:18)]
```

Lage formler og knytte de til datasettene for bruk av GLMNET

```
set.seed(0708)
Formel_DS1_GO <- Sluttkost ~ P85 + Opprinnelig.P50 + Seksjon + Region + Kategori + Entrepri
seform + Oppstart + Ferdigstilt + Varighet + Andel.usikkerhetsavsetning..P85. + Tidligfasek
ostnad - 1

Formel_DS2_GO <- Sluttkost ~ P85 + Opprinnelig.P50 + Seksjon + Region + Kategori + Entrepri
seform + Oppstart + Ferdigstilt + Varighet + Andel.usikkerhetsavsetning..P85. - 1

Formel_DS3_GO <- Sluttkost ~ P85 + Opprinnelig.P50 + Seksjon + Region + Kategori + Entrepri
seform + Oppstart + Ferdigstilt + Varighet + Andel.usikkerhetsavsetning..P85. + Tidligfasek
ostnad - 1

Formel_DS1_FP <- Sluttkost ~ Seksjon + Region + Kategori + Entrepriseform + Oppstart + Ferd
igstilt + Varighet - 1

Formel_DS2_FP <- Sluttkost ~ Seksjon + Region + Kategori + Entrepriseform + Oppstart + Ferd
igstilt + Varighet - 1

Formel_DS3 <- Sluttkost ~ P50 + Andel.generelle.kostnader + Hovedlev + Konsulent + Antall.e
ndringsavtaler + Endringer.relativt.til.kontrakt + P85 + Opprinnelig.P50 + Seksjon + Region
+ Kategori + Entrepriseform + Oppstart + Ferdigstilt + Varighet + Andel.usikkerhetsavsetnin
g..P85. + Tidligfasekostnad - 1

Formel_DS2 <- Sluttkost ~ P50 + Andel.generelle.kostnader + P85 + Opprinnelig.P50 + Seksjon
+ Region + Kategori + Entrepriseform + Oppstart + Ferdigstilt + Varighet + Andel.usikkerhet
savsetning..P85. - 1



FP_Datasett_1_testX <- build.x(Formel_DS1_FP, data = FP_Datasett_1_test, contrasts = FALSE,
sparse = TRUE)
FP_Datasett_1_trainX <- build.x(Formel_DS1_FP, data = FP_Datasett_1_train, contrasts = FALS
E, sparse = TRUE)

FP_Datasett_2_testX <- build.x(Formel_DS2_FP, data = FP_Datasett_2_test, contrasts = FALSE,
sparse = TRUE)
FP_Datasett_2_trainX <- build.x(Formel_DS2_FP, data = FP_Datasett_2_train, contrasts = FALS
E, sparse = TRUE)

GO_Datasett_1_testX <- build.x(Formel_DS1_GO, data = GO_Datasett_1_test, contrasts = FALSE,
sparse = TRUE)
GO_Datasett_1_trainX <- build.x(Formel_DS1_GO, data = GO_Datasett_1_train, contrasts = FALS
E, sparse = TRUE)

GO_Datasett_2_testX <- build.x(Formel_DS2_GO, data = GO_Datasett_2_test, contrasts = FALSE,
sparse = TRUE)
GO_Datasett_2_trainX <- build.x(Formel_DS2_GO, data = GO_Datasett_2_train, contrasts = FALS
E, sparse = TRUE)

GO_Datasett_3_testX <- build.x(Formel_DS3_GO, data = GO_Datasett_3_test, contrasts = FALSE,
sparse = TRUE)
GO_Datasett_3_trainX <- build.x(Formel_DS3_GO, data = GO_Datasett_3_train, contrasts = FALS
E, sparse = TRUE)

Datasett_3_testX <- build.x(Formel_DS3, data = Datasett_3_test, contrasts = FALSE, sparse =
TRUE)
Datasett_3_trainX <- build.x(Formel_DS3, data = Datasett_3_train, contrasts = FALSE, sparse
= TRUE)

Datasett_2_testX <- build.x(Formel_DS2, data = Datasett_2_test, contrasts = FALSE, sparse =
TRUE)
Datasett_2_trainX <- build.x(Formel_DS2, data = Datasett_2_train, contrasts = FALSE, sparse
```

```
= TRUE)

FP_Datasett_1_testY <- build.y(Formel_DS1_FP, data = FP_Datasett_1_test)
FP_Datasett_1_trainY <- build.y(Formel_DS1_FP, data = FP_Datasett_1_train)

FP_Datasett_2_testY <- build.y(Formel_DS2_FP, data = FP_Datasett_2_test)
FP_Datasett_2_trainY <- build.y(Formel_DS2_FP, data = FP_Datasett_2_train)

GO_Datasett_1_testY <- build.y(Formel_DS1_GO, data = GO_Datasett_1_test)
GO_Datasett_1_trainY <- build.y(Formel_DS1_GO, data = GO_Datasett_1_train)

GO_Datasett_2_testY <- build.y(Formel_DS2_GO, data = GO_Datasett_2_test)
GO_Datasett_2_trainY <- build.y(Formel_DS2_GO, data = GO_Datasett_2_train)

GO_Datasett_3_testY <- build.y(Formel_DS3_GO, data = GO_Datasett_3_test)
GO_Datasett_3_trainY <- build.y(Formel_DS3_GO, data = GO_Datasett_3_train)

Datasett_3_testY <- build.y(Formel_DS3, data = Datasett_3_test)
Datasett_3_trainY <- build.y(Formel_DS3, data = Datasett_3_train)

Datasett_2_testY <- build.y(Formel_DS2, data = Datasett_2_test)
Datasett_2_trainY <- build.y(Formel_DS2, data = Datasett_2_train)
```

RFE datasett 3 alle variabler

```
set.seed(0708)
rfe_kontroll <- rfeControl(functions = rfFuncs,
                           method = "boot",
                           repeats = 5,
                           number = 10,
                           verbose = FALSE)

rf_profil_3 <- rfe(x = Datasett_3[, c(2:18)],
                   y = Datasett_3$Sluttkost,
                   sizes = c(1:17),
                   rfeControl = rfe_kontroll)

rf_profil_3
```

RFE datasett 2 alle variabler

```
set.seed(0708)
rf_profil_2 <- rfe(x = Datasett_2[, c(2:13)],
                   y = Datasett_2$Sluttkost,
                   sizes = c(1:12),
                   rfeControl = rfe_kontroll)
rf_profil_2
```

RFE datasett 3 ved GO

```
set.seed(0708)
rf_profil_3_GO <- rfe(x = Datasett_3[, c(2:12)],
                   y = Datasett_3$Sluttkost,
                   sizes = c(1:11),
                   rfeControl = rfe_kontroll)

rf_profil_3_GO
```

RFE datasett 2 ved GO

```
set.seed(0708)
rf_profil_2_GO <- rfe(x = Datasett_2[, c(2:12)],
                      y = Datasett_2$Sluttkost,
                      sizes = c(1:11),
                      rfeControl = rfe_kontroll)

rf_profil_2_GO
```

RFE datasett 2 i FP

```
set.seed(0708)
rf_profil_2_FP <- rfe(x = Datasett_2[, c(2:8)],
                      y = Datasett_2$Sluttkost,
                      sizes = c(1:7),
                      rfeControl = rfe_kontroll)

rf_profil_2_FP
```

RFF datasett 3 alle variabler

```
set.seed(0708)
RRF_DS3 <- train(Sluttkost ~ ., data = Datasett_3, method = "RRF", importance=TRUE)
RRF_IMP_DS3 <- varImp(RRF_DS3, scale = FALSE)
plot(RRF_IMP_DS3, top = 10, main = "Viktighet av variabler datasett 3 alle")
```

RFF datasett 2 alle variabler

```
set.seed(0708)
RRF_DS2 <- train(Sluttkost ~ ., data = Datasett_2, method = "RRF", importance=TRUE)
RRF_IMP_DS2 <- varImp(RRF_DS2, scale = FALSE)
plot(RRF_IMP_DS2, top = 10, main = "Viktighet av variabler datasett 2 alle")
```

RFF datasett 3 ved GO

```
set.seed(0708)
RRF_DS3_GO <- train(Formel_DS3_GO, data = Datasett_3, method = "RRF", importance=TRUE)
RRF_IMP_DS3_GO <- varImp(RRF_DS3_GO, scale = FALSE)
plot(RRF_IMP_DS3_GO, top = 10, main = "Viktighet av variabler datasett 3 ifm GO")
```

RFF datasett 2 ved GO

```
set.seed(0708)
RRF_DS2_GO <- train(Formel_DS2_GO, data = Datasett_2, method = "RRF", importance=TRUE)
RRF_IMP_DS2_GO <- varImp(RRF_DS2_GO, scale = FALSE)
plot(RRF_IMP_DS2_GO, top = 10, main = "Viktighet av variabler datasett 2 ifm GO")
```

RFF datasett 2 i FP

```
set.seed(0708)
RRF_DS2_FP <- train(Formel_DS2_FP, data = Datasett_2, method = "RRF", importance=TRUE)
RRF_IMP_DS2_FP <- varImp(RRF_DS2_FP, scale = FALSE)
plot(RRF_IMP_DS2_FP, top = 10, main = "Viktighet av variabler datasett 2 ifm Forprosjekt")
```

LASSO datasett 3 alle varaibler

```
cv_lasso_DS3 <- cv.glmnet(x = Datasett_3_trainX, y = Datasett_3_trainY, family = "gaussian"
, alpha=1, standardize=TRUE, type.measure= "deviance")
cat('Min Lambda: ', cv_lasso_DS3$lambda.min, '\n 1Sd Lambda: ', cv_lasso_DS3$lambda.1se)
df_coef <- (round(coef(cv_lasso_DS3, s=cv_lasso_DS3$lambda.min), 2))
as.data.frame(df_coef[df_coef[, 1] != 0,])
```

LASSO datasett 2 alle varaibler

```
cv_lasso_DS2 <- cv.glmnet(x = Datasett_2_trainX, y = Datasett_2_trainY, family = "gaussian"
, alpha=1, standardize=TRUE, type.measure= "deviance")
cat('Min Lambda: ', cv_lasso_DS2$lambda.min, '\n 1Sd Lambda: ', cv_lasso_DS2$lambda.1se)
df_coef <- (round(coef(cv_lasso_DS2, s=cv_lasso_DS2$lambda.min), 2))
as.data.frame(df_coef[df_coef[, 1] != 0,])
```

LASSO datasett 3 ved GO

```
cv_lasso_DS3_GO <- cv.glmnet(x = GO_Datasett_3_trainX, y = GO_Datasett_3_trainY, family = "
gaussian", alpha=1, standardize=TRUE, type.measure= "deviance")
cat('Min Lambda: ', cv_lasso_DS3_GO$lambda.min, '\n 1Sd Lambda: ', cv_lasso_DS3_GO$lambda.1
se)
df_coef <- (round(coef(cv_lasso_DS3_GO, s=cv_lasso_DS3_GO$lambda.min), 2))
as.data.frame(df_coef[df_coef[, 1] != 0,])
```

LASSO datasett 2 ved GO

```
cv_lasso_DS2_GO <- cv.glmnet(x = GO_Datasett_2_trainX, y = GO_Datasett_2_trainY, family = "
gaussian", alpha=1, standardize=TRUE, type.measure= "deviance")
cat('Min Lambda: ', cv_lasso_DS2_GO$lambda.min, '\n 1Sd Lambda: ', cv_lasso_DS2_GO$lambda.1
se)
df_coef <- (round(coef(cv_lasso_DS2_GO, s=cv_lasso_DS2_GO$lambda.min), 2))
as.data.frame(df_coef[df_coef[, 1] != 0,])
```

LASSO datasett 3 i FP

```
cv_lasso_DS2_FP <- cv.glmnet(x = FP_Datasett_2_trainX, y = FP_Datasett_2_trainY, family = "
gaussian", alpha=1, standardize=TRUE, type.measure= "deviance")
cat('Min Lambda: ', cv_lasso_DS2_FP$lambda.min, '\n 1Sd Lambda: ', cv_lasso_DS2_FP$lambda.1
se)
df_coef <- (round(coef(cv_lasso_DS2_FP, s=cv_lasso_DS2_FP$lambda.min), 2))
as.data.frame(df_coef[df_coef[, 1] != 0,])
```

MRA datasett 1 ved GO

```
set.seed(0708)
GO_DS1_mape_resultater <- numeric()
GO_DS1_mse_resultater <- numeric()
GO_DS1_rmse_resultater <- numeric()

for (GO_DS1_alpha in seq(0, 1, by = 0.1)) {

  GO_DS1_GLM <- cv.glmnet(x = GO_Datasett_1_trainX, y = GO_Datasett_1_trainY,
                          family = "gaussian",
                          alpha = GO_DS1_alpha,
                          nfolds = 10)
```

```
  GO_DS1_pred <- predict(GO_DS1_GLM, newx = GO_Datasett_1_testX, s = "lambda.1se")

  GO_DS1_mape_resultat <- MAPE(GO_DS1_pred, GO_Datasett_1_testY)
  GO_DS1_mse_resultat <- mean((GO_DS1_pred - GO_Datasett_1_testY)^2)
  GO_DS1_rmse_resultat <- sqrt(GO_DS1_mse_resultat)

  GO_DS1_mape_resultater <- c(GO_DS1_mape_resultater, GO_DS1_mape_resultat)
  GO_DS1_mse_resultater <- c(GO_DS1_mse_resultater, GO_DS1_mse_resultat)
  GO_DS1_rmse_resultater <- c(GO_DS1_rmse_resultater, GO_DS1_rmse_resultat)

  cat("Alpha:", GO_DS1_alpha, "\tMAPE:", GO_DS1_mape_resultat, "\tMSE:", GO_DS1_mse_resulta
t, "\tRMSE:", GO_DS1_rmse_resultat, "\n")
}

GO_DS1_resultater <- data.frame(Alpha = seq(0, 1, by = 0.1),
                                MAPE = GO_DS1_mape_resultater,
                                MSE = GO_DS1_mse_resultater,
                                RMSE = GO_DS1_rmse_resultater)

print(GO_DS1_resultater)
```

MRA datasett 1 ved GO, vurdering av antall folds

```
set.seed(0708)
GO_DS1_mape_resultater_ny <- numeric()
GO_DS1_mse_resultater_ny <- numeric()
GO_DS1_rmse_resultater_ny <- numeric()

for (nfolds_ny in seq(3, 15, by = 1)) {

  GO_DS1_GLM_ny <- cv.glmnet(x = GO_Datasett_1_trainX, y = GO_Datasett_1_trainY,
                             family = "gaussian",
                             alpha = 0.8,
                             nfolds = nfolds_ny)

  GO_DS1_pred_ny <- predict(GO_DS1_GLM_ny, newx = GO_Datasett_1_testX, s = "lambda.1se")

  GO_DS1_mape_resultat_ny <- MAPE(GO_DS1_pred_ny, GO_Datasett_1_testY)
  GO_DS1_mse_resultat_ny <- mean((GO_DS1_pred_ny - GO_Datasett_1_testY)^2)
  GO_DS1_rmse_resultat_ny <- sqrt(GO_DS1_mse_resultat_ny)

  GO_DS1_mape_resultater_ny <- c(GO_DS1_mape_resultater_ny, GO_DS1_mape_resultat_ny)
  GO_DS1_mse_resultater_ny <- c(GO_DS1_mse_resultater_ny, GO_DS1_mse_resultat_ny)
  GO_DS1_rmse_resultater_ny <- c(GO_DS1_rmse_resultater_ny, GO_DS1_rmse_resultat_ny)

  cat("nfolds_ny:", nfolds_ny, "\tMAPE:", GO_DS1_mape_resultat_ny, "\tMSE:", GO_DS1_mse_res
ultat_ny, "\tRMSE:", GO_DS1_rmse_resultat_ny, "\n")
}

GO_DS1_resultater_ny <- data.frame(nfolds_ny = seq(3, 15, by = 1),
                                   MAPE = GO_DS1_mape_resultater_ny,
                                   MSE = GO_DS1_mse_resultater_ny,
                                   RMSE = GO_DS1_rmse_resultater_ny)

print(GO_DS1_resultater_ny)
```

MRA datasett 2 ved GO

```
set.seed(0708)
GO_DS2_mape_resultater <- numeric()
GO_DS2_mse_resultater <- numeric()
GO_DS2_rmse_resultater <- numeric()

for (GO_DS2_alpha in seq(0, 1, by = 0.1)) {

  GO_DS2_GLM <- cv.glmnet(x = GO_Datasett_2_trainX, y = GO_Datasett_2_trainY,
                          family = "gaussian",
                          alpha = GO_DS2_alpha,
                          nfolds = 3)

  GO_DS2_pred <- predict(GO_DS2_GLM, newx = GO_Datasett_2_testX, s = "lambda.1se")

  GO_DS2_mape_resultat <- MAPE(GO_DS2_pred, GO_Datasett_2_testY)
  GO_DS2_mse_resultat <- mean((GO_DS2_pred - GO_Datasett_2_testY)^2)
  GO_DS2_rmse_resultat <- sqrt(GO_DS2_mse_resultat)

  GO_DS2_mape_resultater <- c(GO_DS2_mape_resultater, GO_DS2_mape_resultat)
  GO_DS2_mse_resultater <- c(GO_DS2_mse_resultater, GO_DS2_mse_resultat)
  GO_DS2_rmse_resultater <- c(GO_DS2_rmse_resultater, GO_DS2_rmse_resultat)

  cat("Alpha:", GO_DS2_alpha, "\tMAPE:", GO_DS2_mape_resultat, "\tMSE:", GO_DS2_mse_resulta
t, "\tRMSE:", GO_DS2_rmse_resultat, "\n")
}

GO_DS2_resultater <- data.frame(Alpha = seq(0, 1, by = 0.1),
                                MAPE = GO_DS2_mape_resultater,
                                MSE = GO_DS2_mse_resultater,
                                RMSE = GO_DS2_rmse_resultater)

print(GO_DS2_resultater)
```

MRA datasett 3 ved GO

```
set.seed(0708)
GO_DS3_mape_resultater <- numeric()
GO_DS3_mse_resultater <- numeric()
GO_DS3_rmse_resultater <- numeric()

for (GO_DS3_alpha in seq(0, 1, by = 0.1)) {

  GO_DS3_GLM <- cv.glmnet(x = GO_Datasett_3_trainX, y = GO_Datasett_3_trainY,
                          family = "gaussian",
                          alpha = GO_DS3_alpha,
                          nfolds = 3)

  GO_DS3_pred <- predict(GO_DS3_GLM, newx = GO_Datasett_3_testX, s = "lambda.1se")

  GO_DS3_mape_resultat <- MAPE(GO_DS3_pred, GO_Datasett_3_testY)
  GO_DS3_mse_resultat <- MSE(GO_DS3_pred, GO_Datasett_3_testY)
  GO_DS3_rmse_resultat <- sqrt(GO_DS3_mse_resultat)

  GO_DS3_mape_resultater <- c(GO_DS3_mape_resultater, GO_DS3_mape_resultat)
  GO_DS3_mse_resultater <- c(GO_DS3_mse_resultater, GO_DS3_mse_resultat)
  GO_DS3_rmse_resultater <- c(GO_DS3_rmse_resultater, GO_DS3_rmse_resultat)



}
```

```
GO_DS3_resultater <- tibble(Alpha = seq(0, 1, by = 0.1),
                            MAPE = GO_DS3_mape_resultater,
                            MSE = GO_DS3_mse_resultater,
                            RMSE = GO_DS3_rmse_resultater)

print(GO_DS3_resultater)
```

MRA datasett 2 i FP

```
set.seed(0708)
FP_DS2_mape_resultater <- numeric()
FP_DS2_mse_resultater <- numeric()
FP_DS2_rmse_resultater <- numeric()

for (FP_DS2_alpha in seq(0, 1, by = 0.1)) {

  FP_DS2_GLM <- cv.glmnet(x = FP_Datasett_2_trainX, y = FP_Datasett_2_trainY,
                          family = "gaussian",
                          alpha = FP_DS2_alpha,
                          nfolds = 5)

  FP_DS2_pred <- predict(FP_DS2_GLM, newx = FP_Datasett_2_testX, s = "lambda.1se")

  FP_DS2_mape_resultat <- MAPE(FP_DS2_pred, FP_Datasett_2_testY)
  FP_DS2_mse_resultat <- mean((FP_DS2_pred - FP_Datasett_2_testY)^2)
  FP_DS2_rmse_resultat <- sqrt(FP_DS2_mse_resultat)

  FP_DS2_mape_resultater <- c(FP_DS2_mape_resultater, FP_DS2_mape_resultat)
  FP_DS2_mse_resultater <- c(FP_DS2_mse_resultater, FP_DS2_mse_resultat)
  FP_DS2_rmse_resultater <- c(FP_DS2_rmse_resultater, FP_DS2_rmse_resultat)

  cat("Alpha:", FP_DS2_alpha, "\tMAPE:", FP_DS2_mape_resultat, "\tMSE:", FP_DS2_mse_resulta
t, "\tRMSE:", FP_DS2_rmse_resultat, "\n")
}

FP_DS2_resultater <- data.frame(Alpha = seq(0, 1, by = 0.1),
                            MAPE = FP_DS2_mape_resultater,
                            MSE = FP_DS2_mse_resultater,
                            RMSE = FP_DS2_rmse_resultater)

print(FP_DS2_resultater)
```

MRA datasett 1 i FP

```
set.seed(0708)
FP_DS1_mape_resultater <- numeric()
FP_DS1_mse_resultater <- numeric()
FP_DS1_rmse_resultater <- numeric()

for (FP_DS1_alpha in seq(0, 1, by = 0.1)) {

  FP_DS1_GLM <- cv.glmnet(x = FP_Datasett_1_trainX, y = FP_Datasett_1_trainY,
                          family = "gaussian",
                          alpha = FP_DS1_alpha,
                          nfolds = 3)
```

```
  FP_DS1_pred <- predict(FP_DS1_GLM, newx = FP_Datasett_1_testX, s =
```

Kode til ANN

Fjern alle lagrede datasett, matriser, tabeller, vektorer og verdier

```
rm(list = ls())
#library(tensorflow)
#install_tensorflow(envname = "r-tensorflow")

#install.packages("keras")
#library(keras)
#install_keras()
```

Sett working directory og laste inn pakker

```
setwd(getwd())
library("tidyverse")
library("glmnet")
library("useful")
library("coefplot")
library("zoo")
library("caret")
library("caretEnsemble")
library("e1071")
library("kernlab")
library("MLmetrics")
library("randomForest")
library("RRF")
library("tidyr")
library(devtools)
library(reticulate)
library(tensorflow)
library(keras)
library(remotes)
```

Importering av datasett og visning av toppen av datasettet

```
Master <- read.csv("Masterdata.csv", header = TRUE , sep = ";")
```

Sjekke formatet til variablene

```
str(Master)
```

Vasking av data, og ny sjekk av format

```
Master$Endringer.relativt.til.kontrakt <- gsub(",", ".", Master$Endringer.relativt.til.kont
rakt)
Master$Endringer.relativt.til.kontrakt <- as.numeric(Master$Endringer.relativt.til.kontrakt
)
```

```
Master$Andel.usikkerhetsavsetning..P85. <- gsub(",", ".", Master$Andel.usikkerhetsavsetning
..P85.)
Master$Andel.usikkerhetsavsetning..P85. <- as.numeric(Master$Andel.usikkerhetsavsetning..P8
5.)

Master$Andel.generelle.kostnader <- gsub(",", ".", Master$Andel.generelle.kostnader)
Master$Andel.generelle.kostnader <- as.numeric(Master$Andel.generelle.kostnader)

Master$P85 <- Master$P85/1000
Master$P50 <- Master$P50/1000
Master$Opprinnelig.P50 <- Master$Opprinnelig.P50/1000
Master$Sluttkost <- Master$Sluttkost/1000
Master$Tidligfasekostnad <- Master$Tidligfasekostnad/1000

str(Master)
```

Få en oppsummering av datasettet

```
summary(Master)
```

Lage datasett 1

```
Datasett_1 <- Master
GK_median <- median(Datasett_1$Andel.generelle.kostnader, na.rm = TRUE)
Datasett_1$Andel.generelle.kostnader[is.na(Datasett_1$Andel.generelle.kostnader)] <- GK_med
ian

TF_median <- median(Datasett_1$Tidligfasekostnad, na.rm = TRUE)
Datasett_1$Tidligfasekostnad[is.na(Datasett_1$Tidligfasekostnad)] <- TF_median

AE_median <- median(Datasett_1$Antall.endringsavtaler, na.rm = TRUE)
Datasett_1$Antall.endringsavtaler[is.na(Datasett_1$Antall.endringsavtaler)] <- AE_median

EK_median <- median(Datasett_1$Endringer.relativt.til.kontrakt, na.rm = TRUE)
Datasett_1$Endringer.relativt.til.kontrakt[is.na(Datasett_1$Endringer.relativt.til.kontrakt
)] <- EK_median

Datasett_1$Kategori[Datasett_1$Kategori == ""] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 1"] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 10"] <- "Kategori 5"
Datasett_1$Kategori[Datasett_1$Kategori == "Kategori 11"] <- "Kategori 5"

tomme_celler <- Datasett_1$Entrepriseform == ""
ikke_tomme_celler <- Datasett_1$Entrepriseform[!tomme_celler]
relativ_frekwens <- table(ikke_tomme_celler) / length(ikke_tomme_celler)
antall_per_kategori <- round(relativ_frekwens * sum(tomme_celler))
kategorier <- names(antall_per_kategori)
antall_ganger <- as.vector(antall_per_kategori)
tilfeldig_utvalg <- sample(rep(kategorier, antall_ganger))
Datasett_1$Entrepriseform[tomme_celler] <- tilfeldig_utvalg



tomme_celler_konsulent <- Datasett_1$Konsulent == ""
ikke_tomme_celler_konsulent <- Datasett_1$Konsulent[!tomme_celler_konsulent]
relativ_frekwens_konsulent <- table(ikke_tomme_celler_konsulent) / length(ikke_tomme_celler
_konsulent)
antall_per_kategori_konsulent <- round(relativ_frekwens_konsulent * sum(tomme_celler_konsul
ent))
kategorier_konsulent <- names(antall_per_kategori_konsulent)
antall_ganger_konsulent <- as.vector(antall_per_kategori_konsulent)
```

```
tilfeldig_utvalg_konsulent <- sample(rep(kategorier_konsulent, antall_ganger_konsulent))
Datasett_1$Konsulent[tomme_celler_konsulent] <- tilfeldig_utvalg_konsulent



tomme_celler_hovedlev <- Datasett_1$Hovedlev == ""
ikke_tomme_celler_hovedlev <- Datasett_1$Hovedlev[!tomme_celler_hovedlev]
relativ_frekwens_hovedlev <- table(ikke_tomme_celler_hovedlev) / length(ikke_tomme_celler_h
ovedlev)
antall_per_kategori_hovedlev <- round(relativ_frekwens_hovedlev * sum(tomme_celler_hovedlev
))
kategorier_hovedlev <- names(antall_per_kategori_hovedlev)
antall_ganger_hovedlev <- as.vector(antall_per_kategori_hovedlev)
tilfeldig_utvalg_hovedlev <- sample(rep(kategorier_hovedlev, antall_ganger_hovedlev))
Datasett_1$Hovedlev[tomme_celler_hovedlev] <- tilfeldig_utvalg_hovedlev
```

Sette opp dummyvariabler for kategoriske varaibler

```
Master[Master == ""] <- NA

dmy <- dummyVars(" ~ . ", data = Master)
Masterdummy <- data.frame(predict(dmy, newdata = Master))

dmy1 <- dummyVars(" ~ . ", data = Datasett_1)
DS1dummy <- data.frame(predict(dmy1, newdata = Datasett_1))
```

Lage nye datasett med dummyvaraibler og fjerne navn på varaibler

```
Masterdummy <- as.matrix(Masterdummy)
DS1dummy <- as.matrix(DS1dummy)

dimnames(Masterdummy) = NULL
dimnames(DS1dummy) = NULL
```

Lage datasett 3

```
Datasett_3 <- na.omit(Masterdummy)
```

Lage datasett 2

```
Datasett_2 <- Masterdummy[, -c(34:169)]
Datasett_2 <- na.omit(Datasett_2)
```

Sette testsplitt og opprette trenings og test datasett for alle datasett

```
sample_size_1 <- floor(0.7 *nrow(Datasett_1))
sample_size_2 <- floor(0.7 *nrow(Datasett_2))
sample_size_3 <- floor(0.7 *nrow(Datasett_3))

set.seed(0708)
train_rows_1 <- sample(nrow(DS1dummy), size = sample_size_1)
train_rows_2 <- sample(nrow(Datasett_2), size = sample_size_2)
train_rows_3 <- sample(nrow(Datasett_3), size = sample_size_3)



Datasett_1_train <- DS1dummy[train_rows_1, ]
Datasett_1_test <- DS1dummy[-train_rows_1, ]
```

```
Datasett_2_train <- Datasett_2[train_rows_2, ]
Datasett_2_test <- Datasett_2[-train_rows_2, ]

Datasett_3_train <- Datasett_3[train_rows_3, ]
Datasett_3_test <- Datasett_3[-train_rows_3, ]

FP_Datasett_3_train <- Datasett_3_train[, -c(31:171)]
FP_Datasett_3_test <- Datasett_3_test[, -c(31:171)]
FP_Datasett_2_train <- Datasett_2_train[, -c(31:171)]
FP_Datasett_2_test <- Datasett_2_test[, -c(31:171)]
FP_Datasett_1_train <- Datasett_1_train[, -c(31:171)]
FP_Datasett_1_test <- Datasett_1_test[, -c(31:171)]

GO_Datasett_3_train <- Datasett_3_train[, -c(35:171)]
GO_Datasett_3_test <- Datasett_3_test[, -c(35:171)]
GO_Datasett_2_train <- Datasett_2_train[, -c(35:171)]
GO_Datasett_2_test <- Datasett_2_test[, -c(35:171)]
GO_Datasett_1_train <- Datasett_1_train[, -c(32:171)]
GO_Datasett_1_test <- Datasett_1_test[, -c(32:171)]
```

Splitte datasettene i X og Y

```
set.seed(0708)

FP_Datasett_1_testX <- FP_Datasett_1_test[,-1]
FP_Datasett_1_trainX <- FP_Datasett_1_train[,-1]

FP_Datasett_2_testX <- FP_Datasett_2_test[,-1]
FP_Datasett_2_trainX <- FP_Datasett_2_train[,-1]

GO_Datasett_1_testX <- GO_Datasett_1_test[,-1]
GO_Datasett_1_trainX <- GO_Datasett_1_train[,-1]


GO_Datasett_2_testX <- GO_Datasett_2_test[,-1]
GO_Datasett_2_trainX <- GO_Datasett_2_train[,-1]

GO_Datasett_3_testX <- GO_Datasett_3_test[,-1]
GO_Datasett_3_trainX <- GO_Datasett_3_train[,-1]


FP_Datasett_1_testY <- FP_Datasett_1_test[,1]
FP_Datasett_1_trainY <- FP_Datasett_1_train[,1]

FP_Datasett_2_testY <- FP_Datasett_2_test[,1]
FP_Datasett_2_trainY <- FP_Datasett_2_train[,1]

GO_Datasett_1_testY <- GO_Datasett_1_test[,1]
GO_Datasett_1_trainY <- GO_Datasett_1_train[,1]

GO_Datasett_2_testY <- GO_Datasett_2_test[,1]
GO_Datasett_2_trainY <- GO_Datasett_2_train[,1]

GO_Datasett_3_testY <- GO_Datasett_3_test[,1]
GO_Datasett_3_trainY <- GO_Datasett_3_train[,1]
```

ANN datasett 1 ved GO

```
ncol(GO_Datasett_1_trainX)
set.seed(0708)
model_DS1_GO <- keras_model_sequential()

model_DS1_GO %>%
  layer_dense(name = "Lag1",
              units = 60,
              activation = "relu",
              input_shape = ncol(GO_Datasett_1_trainX)) %>%
  layer_dense(name = "Lag2",
              units = 50,
              activation = "elu") %>%
    layer_dense(name = "Lag3",
              units = 20,
              activation = "relu") %>%
      layer_dense(name = "Lag4",
              units = 10,
              activation = "elu") %>%
    layer_dense(name = "Lag5",
              units = 5,
              activation = "relu") %>%
  layer_dense(name = "Output",
              units = 1,
              activation = "relu")

summary(model_DS1_GO)

set.seed(0708)

model_DS1_GO %>% compile(loss = "MAPE",
                         optimizer = optimizer_adam(learning_rate = 0.1),

                          )

set.seed(0708)
fit_DS1_GO <- model_DS1_GO %>%
  fit(GO_Datasett_1_trainX, GO_Datasett_1_trainY,
      epochs =15,
      validation_split = 0.3,
      verbose = 2)

plot(fit_DS1_GO)

GO_DS1_pred <- predict(model_DS1_GO, GO_Datasett_1_testX)

MAPE(GO_DS1_pred, GO_Datasett_1_testY)
MSE(GO_DS1_pred, GO_Datasett_1_testY)
RMSE(GO_DS1_pred, GO_Datasett_1_testY)
```

ANN datasett 2 ved GO

```
set.seed(0708)
model_DS2_GO <- keras_model_sequential()

model_DS2_GO %>%
  layer_dense(name = "Lag1",
              units = 50,
              activation = "relu",
              input_shape = ncol(GO_Datasett_2_trainX)) %>%
  layer_dense(name = "Lag2",
              units = 30,
```

```
              activation = "elu") %>%
  layer_dense(name = "Lag3",
              units = 10,
              activation = "relu") %>%
  #layer_dense(name = "Lag4",
   #           units = 10,
  #            activation = "elu") %>%
  #layer_dense(name = "Lag5",
   #           units = 5,
    #          activation = "relu") %>%
  layer_dense(name = "Output",
              units = 1,
              activation = "relu")

summary(model_DS2_GO)

set.seed(0708)

model_DS2_GO %>% compile(loss = "MAPE",
                         optimizer = optimizer_adam(learning_rate = 0.1),

                          )

set.seed(0708)
fit_DS2_GO <- model_DS2_GO %>%
  fit(GO_Datasett_2_trainX, GO_Datasett_2_trainY,
      epochs =15,
      validation_split = 0.3,
      verbose = 2)

plot(fit_DS2_GO)

GO_DS2_pred <- predict(model_DS2_GO, GO_Datasett_2_testX)

MAPE(GO_DS2_pred, GO_Datasett_2_testY)
MSE(GO_DS2_pred, GO_Datasett_2_testY)
RMSE(GO_DS2_pred, GO_Datasett_2_testY)

avvik_GO_DS2 <- as.data.frame(cbind(GO_DS2_pred, GO_Datasett_2_testY ))
avvik_GO_DS2 <- cbind(avvik_GO_DS2, GO_Datasett_2_testX[,31])
colnames(avvik_GO_DS2) <- c("Predikert", "Faktisk", "Opprinnelig_P50")
#avvik_GO_DS2$avvik <- avvik_GO_DS2$predikert - avvik_GO_DS2$faktisk

avvik_GO_DS2 %>%
  ggplot(aes(x = Predikert, y = Faktisk)) +
  geom_histogram(stat = "identity", colour = "blue", width = 1000) +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Faktisk og Predikert") +
  theme(plot.title = element_text(hjust = 0.5))

avvik_GO_DS2 %>%
  ggplot(aes(x = Opprinnelig_P50, y = Faktisk)) +
  geom_histogram(stat = "identity", colour = "blue", width = 1000) +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Faktisk og Opprinnelig P50") +
  theme(plot.title = element_text(hjust = 0.5))
```

ANN datasett 3 ved GO

```
set.seed(0708)
model_DS3_GO <- keras_model_sequential()

model_DS3_GO %>%
```

```
  layer_dense(name = "Lag1",
              units = 50,
              activation = "relu",
              input_shape = ncol(GO_Datasett_3_trainX)) %>%
  layer_dense(name = "Lag2",
              units = 40,
              activation = "elu") %>%
   layer_dense(name = "Lag3",
              units = 20,
              activation = "relu") %>%
    #layer_dense(name = "Lag4",
    #           units = 20,
     #          activation = "relu") %>%
  #layer_dense(name = "Lag5",
   #           units = 10,
    #           activation = "relu") %>%
  layer_dense(name = "Output",
              units = 1,
             activation = "relu")

summary(model_DS3_GO)

set.seed(0708)

model_DS3_GO %>% compile(loss = "MAPE",
                         optimizer = optimizer_adam(learning_rate = 0.1),

                          )

set.seed(0708)
fit_DS3_GO <- model_DS3_GO %>%
  fit(GO_Datasett_3_trainX, GO_Datasett_3_trainY,
      epochs =15,
      validation_split = 0.2,
      verbose = 2)

plot(fit_DS3_GO)

GO_DS3_pred <- predict(model_DS3_GO, GO_Datasett_3_testX)

MAPE(GO_DS3_pred, GO_Datasett_3_testY)
MSE(GO_DS3_pred, GO_Datasett_3_testY)
RMSE(GO_DS3_pred, GO_Datasett_3_testY)
```

ANN datasett 1 i FP

```
set.seed(0708)
model_DS1_FP <- keras_model_sequential()

model_DS1_FP %>%
  layer_dense(name = "Lag1",
              units = 50,
              activation = "relu",
              input_shape = ncol(FP_Datasett_1_trainX)) %>%
  layer_dense(name = "Lag2",
              units = 20,
              activation = "relu") %>%
    #layer_dense(name = "Lag3",
     #          units = 20,
      #         activation = "relu") %>%
     # layer_dense(name = "Lag4",
      #           units = 20,
       #          activation = "relu") %>%
```

```
    #layer_dense(name = "Lag5",
   #           units = 10,
  #             activation = "elu") %>%
  layer_dense(name = "Output",
              units = 1,
              activation = "relu")

summary(model_DS1_FP)

set.seed(0708)

model_DS1_FP %>% compile(loss = "MAPE",
                         optimizer = optimizer_adam(learning_rate = 0.1),

                          )

set.seed(0708)
fit_DS1_FP <- model_DS1_FP %>%
  fit(FP_Datasett_1_trainX, FP_Datasett_1_trainY,
      epochs =15,
      validation_split = 0.3,
      verbose = 2)

plot(fit_DS1_FP)

FP_DS1_pred <- predict(model_DS1_FP, FP_Datasett_1_testX)

MAPE(FP_DS1_pred, FP_Datasett_1_testY)
MSE(FP_DS1_pred, FP_Datasett_1_testY)
RMSE(FP_DS1_pred, FP_Datasett_1_testY)

avvik_FP_DS1 <- as.data.frame(cbind(FP_DS1_pred, FP_Datasett_1_testY ))
avvik_FP_DS1 <- cbind(avvik_FP_DS1, FP_Datasett_1_testX[,28])
colnames(avvik_FP_DS1) <- c("Predikert", "Faktisk", "Opprinnelig_P50")
#avvik_GO_DS2$avvik <- avvik_GO_DS2$predikert - avvik_GO_DS2$faktisk

avvik_FP_DS1 %>%
  ggplot(aes(x = Predikert, y = Faktisk)) +
  geom_histogram(stat = "identity", colour = "blue", width = 1000) +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Faktisk og Predikert") +
  theme(plot.title = element_text(hjust = 0.5))

avvik_FP_DS1 %>%
  ggplot(aes(x = Opprinnelig_P50, y = Faktisk)) +
  geom_histogram(stat = "identity", colour = "blue", width = 1000) +
  geom_smooth(se = FALSE, method = "lm", size = 1, colour = "black", alpha = 0.3) +
  labs(title = "Faktisk og Opprinnelig P50") +
  theme(plot.title = element_text(hjust = 0.5))
```

ANN datasett 1 i FP

```
set.seed(0708)
model_DS2_FP <- keras_model_sequential()

model_DS2_FP %>%
  layer_dense(name = "Lag1",
              units = 50,
              activation = "relu",
              input_shape = ncol(FP_Datasett_2_trainX)) %>%
  layer_dense(name = "Lag2",
              units = 30,
              activation = "elu") %>%
```

```
  #  layer_dense(name = "Lag3",
   #           units = 30,
    #          activation = "relu") %>%
  #      layer_dense(name = "Lag4",
   #           units = 20,
    #          activation = "elu") %>%
    #     layer_dense(name = "Lag5",
     #          units = 20,
      #         activation = "relu") %>%
  layer_dense(name = "Output",
              units = 1,
              activation = "relu")

summary(model_DS2_FP)

set.seed(0708)

model_DS2_FP %>% compile(loss = "MAPE",
                         optimizer = optimizer_adam(learning_rate = 0.1),

                          )

set.seed(0708)
fit_DS2_FP <- model_DS2_FP %>%
  fit(FP_Datasett_2_trainX, FP_Datasett_2_trainY,
      epochs =15,
      validation_split = 0.1,
      verbose = 2)

plot(fit_DS2_FP)

FP_DS2_pred <- predict(model_DS2_FP, FP_Datasett_2_testX)

MAPE(FP_DS2_pred, FP_Datasett_2_testY)
MSE(FP_DS2_pred, FP_Datasett_2_testY)
RMSE(FP_DS2_pred, FP_Datasett_2_testY)
```