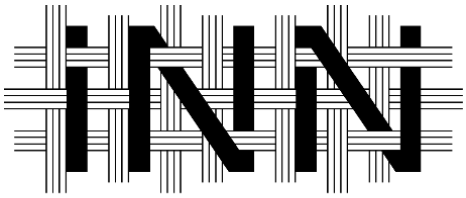


Høgskolen
i Innlandet

Handelshøgskolen Innlandet - Fakultet for økonomi og samfunnsvitenskap

Økonomi og Ledelse - Digital Ledelse og Business Analytics

Masteroppgave

Prediksjon av strømforbruk for private husholdninger:

En sammenligning av maskinlæringsmodeller

Kristin Høisveen Evensen og Elena Mathisen

2024

Sammendrag

Dagens moderne verden er avhengig av elektrisitet for en rekke formål, både private husholdninger og industri. Samfunnet vårt er preget av raske og kontinuerlige endringer, dette inkluderer blant annet nye og kraftige teknologier i norsk industri, som vil kreve langt mer kraft enn hva vi har tilgjengelig i dag. Dette kan føre til et kraftunderskudd i Norge allerede i 2027. Man ser også et behov for overgang til nye metoder innen fornybar energi, som vil øke variabiliteten av tilgjengelig kraft. Dette betyr at Norge er avhengig av å utnytte de ressursene vi har tilgjengelig på en mer effektiv måte.

Maskinlæring er en teknologi som gjør det mulig å navigere i store og komplekse datamengder. Maskinlæring har blant annet evnen til å forbedre prediksjoner gjennom datamaskinens evne til å lære og forbedre seg, ved hjelp av erfaring eller historisk data.

Etter å ha dannet et teoretisk grunnlag for studien gjennomførte vi en kvantitativ studie med sekundærdata fra tre ulike kilder. Vi benyttet historisk forbruksdata mottatt fra et strømselskap, samt historisk strømpris fra Forbrukerrådet og historiske værforhold fra Norsk klimaservicesenter. Studien ble utført med historisk data begrenset til en region i Innlandet.

Formålet med denne masteroppgaven var å vurdere om man kan predikere strømforbruk hos norske husholdninger ved hjelp av maskinlæring. Seks ulike maskinlæringsmodeller har blitt studert for å kunne si noe om dette. De ulike forklaringsvariablene ble også testet og vurdert for å se om de kunne forbedre produksjonsevnen til modellene. Resultatene av prediksjonene ble presentert via visualiseringer og nøyaktighetsmålinger, i tillegg til at det ble predikert for ulike tidshorisonter for å kartlegge modellenes stabilitet.

Det var to maskinlæringsmodeller som utmerket seg i denne studien; LASSO og Gradient Boosting, hvor LASSO var den mest stabile over tid.

Abstract

In today's modern world, electricity is indispensable for a multitude of purposes, including both private households and industry. Our society is characterized by rapid and continuous changes, including the emergence of new and powerful technologies within the Norwegian industrial sector, which are anticipated to demand significantly more power than currently available. This scenario could lead to a power deficit in Norway as early as 2027. Additionally, there is a shift towards new methods in renewable energy, which will increase the variability of available power. Consequently, Norway must handle its resources more efficiently.

Machine learning is a technology that facilitates navigation through large and complex data sets. Among other capabilities, machine learning enhances predictions through the computer's ability to learn and improve from experience or historical data.

After establishing a theoretical foundation for the study, a quantitative analysis was conducted using secondary data from three distinct sources. The research utilized historical consumption data received from an electricity company, along with historical electricity prices from “Forbrukerrådet” and historical weather conditions from the Norwegian Climate Services Center. The study was carried out using historical data limited to a smaller region in Innlandet, Norway.

The objective of this thesis was to assess the possibility of predicting electricity consumption in Norwegian households using machine learning. Six different machine learning models were examined to explore this capability. Various explanatory variables were also tested and employed in the predictions to enhance predictive accuracy. The outcomes of the predictions were presented through visualizations and accuracy measurements, and predictions were made over different time horizons to assess the stability of the models.

Two machine learning models distinguished themselves in this study: LASSO and Gradient Boosting, with LASSO proving to be the most stable over time.

Forord

Denne masteroppgaven markerer slutten på vår tid som studenter ved Handelshøgskolen Innlandet. Vi har med dette fullført vår siviløkonomutdanning i Økonomi og Ledelse, med Digital ledelse og Business Analytics som spesialisering. For oss har det vært fire krevende, inspirerende og nyttige år som deltidsstudenter ved siden av full jobb.

Til vår masteroppgave valgte vi å gjennomføre en sammenligning av maskinlæringsmodeller, da dette kombinerer vår tilegnede kunnskap fra studieløpet med fag av interesse. Å skrive oppgaven har vært en spennende og lærerik prosess som har bidratt til å forme vår faglige forståelse og personlige vekst.

Vi ønsker å rette en stor takk til vår veileder Rolf Gunnar Finsrud for gode innspill og tilbakemeldinger underveis i vår studie. Hans ekspertise har vært verdifull for oss.

Til slutt ønsker vi å takke strømselskapet som ga oss nødvendig data på historisk strømforbruk hos norske husholdninger.

Innholdsfortegnelse

1. Introduksjon

Introduksjonskapittelet vil presentere masteroppgavens bakgrunn og formål, samt problemstillingen og forskningsspørsmålene studien har som mål å besvare. Deretter vil oppgavens forutsetninger og begrensninger introduseres for å gi innblikk i oppgavens rammer. Til slutt legges oppgavens struktur frem i delkapittel 1.4.

1.1. Bakgrunn

Vårt moderne samfunn er avhengig av elektrisitet for en rekke formål i alt fra private hjem og industri, til transport og teknologi. Historisk sett har Norge vært et land med overskudd av energi og det har i de senere årene blitt gjort store utbygginger av kraftproduksjonen og strømnettet, til tross for liten økning i forbruket. Norges Vassdrags- og energidirektorat (NVE) publiserer med jevne mellomrom langsiktige kraftmarkedsanalyser, og i 2023 publiserte de en analyse av kraftmarkedet fra 2023 til 2040 hvor det antas at kraftforbruket i Norge vil øke kraftig mot 2030. Det forventes at den gjennomsnittlige kraftbalansen blir strammere, samtidig som det forventes lite ny kraftproduksjon innenlands (Norges vassdrags- og energidirektorat, 2023). Ettersom det kreves konsesjon for å produsere kraft i Norge er NVE ganske sikre på at det ikke vil komme ny kraftproduksjon i Norge frem mot 2030. Etter 2030 er det større usikkerhet i anslagene for ny produksjon, spesielt rettet mot vindkraft (Norges vassdrags- og energidirektorat, 2023). Dette samsvarer med en kortsiktig markedsanalyse Statnett publiserte i 2022, av kraftmarkedet i Norge for årene 2022-2027. I analysen kommer det frem at det forventes en betydelig økning i kraftforbruket de kommende fem årene, uten at det ligger an til en tilsvarende økning i kraftproduksjonen. Dette kan bety at Norge vil risikere å havne i et kraftunderskudd i 2027 (Statnett, 2022). Statnett skriver videre at fremtidens industri trenger mye kraft og planene om etablering av ny industri og elektrifisering blir stadig større og mer konkrete. Planene i industrien utfordrer nettet, i tillegg til at det vil kreve tilgang på ny kraftproduksjon (Statnett, 2022). Videre skriver også Statnett at det er planer om å bygge nett for mellom 60 og 100 milliarder kroner frem til 2030, men vet ikke om dette alene vil imøtekomme økt forespørsel. NVE (2023) forklarer at kapasiteten i kraftnettet demper forbruksvekst og at det er et stort behov for tempo i nettutbyggingen. Kraftmarkedet i Norge vil bli mer detaljert beskrevet i kapittel 2 som er teoridelen.

Oppnåelse av klimamålene krever også en svært rask vekst i klimanøytrale energikilder og energibærere, og det er mange land i Europa som har store ambisjoner for utbygging av både sol- og vindkraft (Norges vassdrags- og energidirektorat, 2023). De europeiske energimarkedene har vært preget av stor usikkerhet grunnet krigen i Europa og redusert tilgang på gass, som har ført til at

omstillingen av kraftsystemet i Europa skyter fart (Statnett, 2022). Som et resultat blir det europeiske energisystemet stadig blir mer preget av fornybar kraftproduksjon, som betyr at tilgangen og prisene på kraft vil variere mer. Fornybare energikilder vil kreve mer fleksibilitet for å dekke kraftbehovet i topplasttimene og holde systemet i balanse (Norges vassdrags- og energidirektorat, 2023). Det skrives også at fleksibelt forbruk vil bidra til bedre ressursutnyttelse og færre ubalanser i timer med stort overskudd av fornybar produksjon, noe som er positivt da det å balansere ut store svingninger kan resultere i store kostnader for samfunnet.

Statnett (2022) spår at prisene vil normaliseres de neste fem årene, trolig samtidig som planene i industrien har modnet. Norge vil med stor sannsynlighet oppleve et underskudd av kraft, som vil si at vi i fremtiden vil være mer avhengig av import fra andre land eller mer effektive metoder for å produsere og lagre strøm. Dette kan bli et problem, da forsyningssikkerheten i Norge trues og kan påvirke kraftprisene. Derfor vil det være avgjørende å kunne forutse og håndtere det økende strømforbruket i Norge. For å opprettholde bærekraftige priser og samtidig støtte industriell vekst, må nettinfrastrukturen tilpasse seg for å håndtere det økte presset og opprettholde pålitelig forsyning. Videre vil overgangen til nye metoder innen fornybar energi øke variabiliteten av tilgjengelig kraft. Det vil kreve avanserte prognosemodeller for å kunne forutsi svingningene i kraftproduksjonen og hvordan vi effektivt kan håndtere konsekvensene. Riktige prognosemodeller vil derfor kunne optimalisere ressursbruk, sikre pålitelig kraftleveranse og bidra til økonomisk og miljømessig bærekraft i energisektoren. Imidlertid utvikles stadig nye og kraftigere former for teknologi som har revolusjonert feltet som omhandler prognosemodeller, som gjør det enklere å utvikle modeller som kan forutse kommende utfall på en helt ny måte. Denne teknologien er en del av et større fagområde, nemlig maskinlæring. Maskinlæring gjør det mulig å programmere datamaskiner til å optimalisere ytelseskriteria ved å bruke historisk data eller eksempeldata (Alpaydin, 2014). Maskinlæring kan håndtere store datamengder og er effektivt for å identifisere komplekse, ikke-lineære sammenhenger som mennesker eller tradisjonelle statistiske modeller kan overse. Teknologien kan forbedre nøyaktigheten av prediksjoner, gjør det mulig å behandle og analysere store datamengder raskt, og hjelper beslutningstakere med å ta mer informerte valg enn tidligere. Maskinlæringsmodeller er, med sin evne til å behandle og analysere store datasett for å identifisere mønstre og trender, et sentralt verktøy for å møte et problem som fremtidig kraftunderskudd. Ved å forbedre nøyaktigheten i prognoser for strømforbruk, kan man optimalisere drift og planlegging av kraftsystemer, i tillegg til å fremme en mer bærekraftig og effektiv energiutnyttelse. Dette kan være avgjørende for å sikre energiforsyning, og å støtte opp under økonomisk vekst i en tid preget av betydelige endringer og usikkerheter i kraftmarkedet. Riktig prissetting av strøm er også avhengig av nøyaktige prognoser på forbruk, som igjen er avgjørende for rettferdige og effektive energimarkeder. Derfor vil det være viktig

å utvikle avanserte prognosemodeller som kan håndtere den økende kompleksiteten og usikkerheten i kraftsystemet, for å sikre en pålitelig, effektiv og bærekraftig energiforsyning i fremtiden.

Det første delkapittelet har blant annet presentert utfordringene Norge står overfor i sitt kraftmarked. Norge står ovenfor en forventet kraftig økning i strømforbruk uten tilsvarende økning i produksjon, noe som potensielt fører til kraftunderskudd frem mot 2030. Statnett (2022) og NVE påpeker behovet for utvidet nettinfrastruktur og nye kraftproduksjonskilder for å møte den økende industrielle etterspørselen og elektrifiseringen, samt overgangen til fornybar energi. Behovet for avanserte prognosemodeller som maskinlæring melder seg, da slike modeller kan være kritiske for å optimalisere kraftsystemers drift og sikre en pålitelig og bærekraftig energiforsyning i en tid med store endringer og usikkerheter i kraftmarkedene.

1.2. Forskningsspørsmål og formål med studien

Maskinlæring kan spille en avgjørende rolle for optimaliseringen av ressursutnyttelse i kraftmarkedet ved å forbedre blant annet prognosemodeller for tilbud og etterspørsel, optimalisere drift og vedlikehold av kraftverk, og å støtte integreringen av fornybar energi. Maskinlæringsteknikker kan forbedre nøyaktigheten til etterspørselsprognoser ved å analysere store datasett med historisk forbruksdata og identifisere mønstre og trender som kan påvirke fremtidig energietterspørsel. Dette muliggjør en mer effektiv balansering av tilbud og etterspørsel, som potensielt kan redusere behovet for dyr toppbelastningsenergi og minimere graden av overskuddsenergi. Et sentralt spørsmål å stille er hvordan vi de neste årene best mulig kan benytte de ressursene vi har. Basert på analysene publisert av Statnett (2022) og NVE (2023), er det planlagt utbygging på kraftnettet samtidig som det utvikles nye metoder for kraftproduksjon. Usikkerheten ligger i om tilbudet av kraft vil være tilstrekkelig i årene frem til disse utbedringene er på plass. Denne usikkerheten skaper et behov for å se på utviklingen i energiforbruk på kort sikt, og hvordan vi best mulig kan forutsi norsk kraftforbruk for å levere mer nøyaktig på produksjon av strøm. Formålet med denne studien er å sammenligne ulike maskinlæringsmodeller, for å se hvordan de evner å forutse strømforbruket for ulike kortsiktige tidshorisonter basert på historisk data fra hele 2022 og januar 2023. Den historiske dataen som benyttes består av den avhengige variabelen daglig strømforbruksdata, samt uavhengige variabler på daglig spotpris for strøm og ulike værfenomener. Studien vil bli gjort med data fra det norske strømmarkedet og er begrenset til regionen Innlandet. Dette leder oss inn på følgende problemstilling og forskingsspørsmål:

Problemstilling:

Hvordan predikere kortsiktig strømforbruk hos norske husholdninger med hjelp av maskinlæring?

Forskningsspørsmål 1:

Hvordan kan inkludering av pris som uavhengig variabel påvirke prediksjon av strømforbruk?

Forskningsspørsmål 2:

Hvordan kan inkludering av ulike værforhold som uavhengige variabler påvirke prediksjon av strømforbruk?

Forskningsspørsmål 3:

Hvilken maskinlæringsmodell presterer best på prediksjoner om kortsiktig strømforbruk i det norske strømmarkedet?

Det finnes en rekke maskinlæringsmodeller som alle har ulike egenskaper og bruksområder. I denne studien er målet å undersøke populære og mye anvendte modeller for prediksjonsformål generelt. Disse modellene er ikke nødvendigvis de samme modellene som er typiske for prediksjon av tidsseriedata, men vi skal utforske de utvalgte modellenes potensiale. Modellene som benyttes i studien har ulik grad av kompleksitet og bygges med kun det mest nødvendige av manuell tilpasning. Alle modellene blir gitt samme forutsetninger, slik at sammenligningsgrunnlaget vil være likt.

Oppsummert sikter studien på å undersøke maskinlæringens rolle i optimalisering av ressursutnyttelse i kraftmarkedet, med fokus på prognosemodeller for etterspørsel av strøm hos norske husholdninger. Ved å benytte datasett med historisk data for å identifisere forbruksmønstre, skal maskinlæring forbedre nøyaktigheten i etterspørselsprognoser. Forbedret nøyaktighet kan bidra til mer effektiv balansering av tilbud og etterspørsel av strøm som potensielt reduserer behovet for kostbar energi i toppbelastningstider. Gjennom analyse av historiske strømforbruksdata hos husholdninger, og inkludering av variabler som pris og værforhold, skal studien undersøke hvordan ulike maskinlæringsmodeller kan predikere strømforbruket på kort sikt. Det er det norske strømmarkedet studien skal gjelde for, med en mindre region i Innlandet som utvalg. Studien vil vurdere ulike maskinlæringsmodellers evne til å forutse kortsiktige svingninger i strømforbruk. Modellenes effektivitet måles under like forutsetninger, med minimal manuelltilpasning.

1.3. Begrensninger/forutsetninger

Målet med denne oppgaven er å sammenligne ulike typer maskinlæringsmodeller og vurdere deres evne til å predikere strømforbruket i Norge. Vi velger å sette søkelys på tilgjengelig data, som begrenser

seg til datasettet vi har mottatt, hvor kun kunder i private husholdninger på Innlandet er inkludert. Som vi har sett innledningsvis i oppgaven er et økt behov for elektrisitet i industrien en vesentlig årsak til et mulig fremtidig kraftunderskudd. En begrensning i tilgang på data fører til at vi får ikke inkludert industrien i arbeidet med prognosemodellene.

En ytterlig begrensning er knyttet til mengden mottatt data. Vi har mottatt data for strømforbruket for Innlandet fra januar 2022 til og med januar 2023, og har derfor bare 13 måneder med data å jobbe med. Dette vil ikke være nok for å representere historisk utvikling i energiforbruk og vil kunne påvirke resultatene til prognosemodellene.

For å inkludere ulike værfaktorer som uavhengige variabler av forbruket så vi oss nødt til å ta et utvalg av kundene fra Innlandet, som var i det originale datasettet, og kun inkludere kundene lokalisert i Hamarregionen. Hamarregionen vil i vår oppgave si Hamar, Løten, Stange og Ringsaker. Dette førte til nok en begrensning, da vi benytter et mindre geografisk område i landet. Det geografiske området er ikke nødvendigvis representativt for resten av landet.

Ved å gjøre dette utvalget, kunne vi hente ut data på temperatur, nedbørsmengde og solskinnstimer per døgn, som vi brukte for å beregne et gjennomsnitt av på tvers av alle værstasjoner i området. Det ble derfor gjort en forutsetning om at hele regionen opplevde samme temperatur, nedbørsmengde og solskinn fra dag til dag gjennom hele tidsperioden på 13 måneder. Prisdataen vi har hentet ut er en beregnet medianverdi av spotprisen i en prisgruppe som gjelder Innlandet, og er dermed heller ikke representativ for prisnivået og utviklingen gjennom året i andre deler av landet.

1.4. Strukturen på oppgaven

Oppgaven består av 6 kapitler. I kapittel 2 vil det teoretiske rammeverket bak studien presenteres. Det gis en kort introduksjon av strømmarkedet i Norge før maskinlæring forklares i dybden. Valg av typer maskinlæringsmodeller blir også presentert og gjort rede for før vi avslutningsvis ser på tidligere forskning på området. Kapittel 3 gir en presentasjon av forskningsformål, valg av metode og hvordan vi har gått frem for å besvare problemstilling og forskningsspørsmål. Det gis også en grundig gjennomgang av data og hvordan denne har blitt behandlet, samt en evaluering av studiens reliabilitet, validitet og generaliseringsevne. Analyse og resultater av maskinlæringsmodellenes prestasjon legges frem i kapittel 4. Dette kapittelet, sammen med teori og metode, legger grunnlaget for diskusjon og konklusjon som gjennomføres i kapittel 5. Det er også i kapittel 5 problemstilling og forskningsspørsmål besvares, før det til slutt gis føringer til videre forskning.

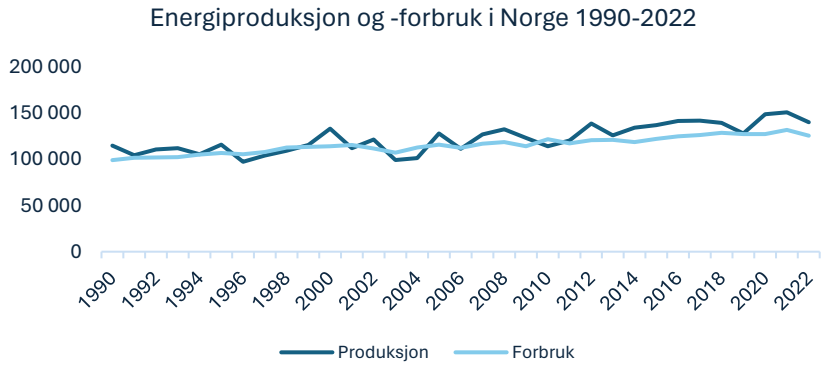
2. Teoretiske Perspektiver

Dette kapittelet introduserer grunnleggende teori om det norske strømmarkedet, samt teori om maskinlæringsmodeller og prognoser. Studien skal avgjøre om ulike maskinlæringsmodeller evner å predikere strømforbruket til et utvalg norske husholdningskunder, og hvor nøyaktige prognosene deres er, derfor er det naturlig å inkludere disse temaene i teoridelen. Det vil også gås i dybden på de ulike maskinlæringsmodellene, hvor de ulike egenskapene deres presenteres.

2.1. Strømmarkedet i Norge

I Norge i dag er det tre grunnleggende funksjoner i kraftforsyningen; produksjon, overføring og omsetning av elektrisitet (Energifakta Norge, 2024d). Ifølge NorgesEnergi (u.å.) er Norge den største vannkraftprodusenten i Europa og kraftmagasinene våre fylles av nedbør og smeltevann gjennom årets ulike sesonger. Vi har gode muligheter for lagring av vann i magasinene våre, men selve strømmen er ikke mulig å lagre, og den må derfor brukes med en gang. På grunn av dette er vi avhengige av store magasiner til å lagre nok vann. I tillegg til dette er geografisk plassering av disse magasinene avgjørende. Nedbør må komme i gitte områder for å kunne fylle magasinene våre (Energifakta Norge, 2024c).

Strømmarkedet defineres av NorgesEnergi (u.å.) som et marked med mange aktører og fri konkurranse, noe som skyldes at strømmen i privatmarkedet ble avregulert gjennom Energiloven i 1991. Dette vil si at markedet styres av markedsprinsippene. I tillegg kaller NorgesEnergi (u.å.) markedet for et marked uten landegrenser, hvor Norden er samlet under ett omsetningsområde og all handel foregår på kraftbørsen Nord Pool Spot (NorgesEnergi, u.å.). Kraftbørsen fungerer slik at strømprodusentene i flere land selger strøm til Nord Pool, hvor strømselskapene kjøper strømmen til en bestemt pris og selger videre til sine kunder. Historisk har Norge vært et land med overskudd av kraft og vi har stått for produksjon av forespurt kraft selv (Statistisk Sentralbyrå, u.å.a). Figur 2.1 er et linjediagram som viser sammenhengen mellom kraftproduksjon og kraftforbruk i Norge fra 1990 til 2022, hvor energiproduksjonen vises av mørkeblå linje og forbruket av lyseblå linje. Figuren viser hvordan produksjonen og forbruket av kraft har hatt en jevn og relativt lav økning gjennom årene, hvor det stort sett har vært overskudd av kraftproduksjon.



Figur 2.1 - Energiproduksjon og -forbruk i Norge - 1990-2022 (Data hentet fra SSB (Statistisk Sentralbyrå, u.å.))

2.1.1. Sikkerhet i strømmarkedet

Sikker tilgang på strøm er grunnleggende for alle samfunnsfunksjoner, og kraftmarkedet har en viktig rolle i å sikre at det er balanse mellom forbruk og produksjon. Som det er beskrevet, er tilgang på fylte magasiner og muligheten for kraftutveksling en forutsetning for forsyningssikkerheten i Norge. Sikkerheten avhenger av flere funksjoner som sammen danner sikker til tilgang på strøm i Norge, disse vil nå kort presenteres. Energisikkerhet er ifølge Energifakta Norge (2024a) definert som kraftsystemets evne til å dekke energibruken. Svikt i energisikkerheten forklares som redusert produksjon av elektrisk energi på grunn av mangel på primærenergi som vann, gass og kull. Norge har i dag en god forsyningssikkerhet fordi vi har en god kraftbalanse og tilgang på utviklingskapasitet med andre land (Energifakta Norge, 2024a).

Effektsikkerhet er definert som kraftsystemets evne til å dekke momentan belastning og forklares som tilgjengelig kapasitet i installert kraftproduksjon eller i kraftnettet (Energifakta Norge, 2024a). Mens svikt i energisikkerhet handler om situasjoner som kan vare over flere uker, handler svikt i effektsikkerhet om kapasiteten i enkelttimer med høyt forbruk. Både mengden energi og mengden av effekt kunden etterspør vil påvirke prisen på strøm. Videre er driftssikkerhet og pålitelighet i leveringen av strøm to viktige funksjoner for å holde samfunnsfunksjoner i gang. Samfunnets økende avhengighet av elektrisitet gjør også at beredskap i kraftforsyning er svært viktig. Økt digitalisering gjør forsyningen av strøm mer sårbar for dataangrep og derfor vil forebygging og rask gjenoppretting ved feil bli svært viktig fremover (Energifakta Norge, 2024a).

2.1.2. Balanse i strømmarkedet

Balanse i strømmarkedet vil si at produksjon og forbruk av elektrisitet samsvarer. Statnett har ansvaret for å avregne en eventuell ubalanse i det norske strømmarkedet (Energifakta Norge, 2024c). Balanseavregning skal sørge for at all innmating og alt uttak av elektrisk energi skal avregnes korrekt, slik at balanse i kraftmarkedet oppnås. Dette betyr at avtalt forbruk eller produksjon må være lik faktisk forbruk eller produsert volum. For å få tilgang til å handle i engrosmarkedet stilles det derfor krav til inngått balanseavtale med Statnett. Balansen mellom tilbud og etterspørsel i Norge sikres i stor grad av det såkalte *day-ahead*-markedet som skal sikre balanse på forhånd. I et *day-ahead*-marked fastsettes priser og mengder for neste dag. Det vil alltid komme uforutsette hendelser som forstyrrer balansen, og dette må Statnett balansere ut. Momentan balanse, ved for eksempel feil i værprognoser, gjenopprettes ved at Statnett kjøper fleksibilitet slik at forbruk og produksjon kan reguleres opp eller ned, avhengig av balansen (Energifakta Norge, 2024c).

2.1.3. Fleksibilitet i strømmarkedet

Fleksibilitet i energisystemet refererer til bruk av ulike energikilder og teknologier for effektivisering av strømnettets bruk, samtidig som man minimerer risikoen for overbelastning. Ifølge Fortum (u.å.) består et fleksibelt energisystem av flere ulike energikilder som brukes sammen på en smart måte. Ved å iverksette alternativer som for eksempel batterilagring, hydrogenløsninger og fjernvarmesystemer, kan man avlaste strømnettet og fremme en økning i andelen fornybar energi. Ved en slik avlastning på strømnettet og økning av fornybar energi reduserer man ikke bare energikostnadene, men bidrar også til å forbedre energiforsyningens pålitelighet og sikkerhet. Målet med fleksibilitet er å skape et dynamisk samspill mellom kraftsystemet og termisk energi for å kunne balansere energien i kraftsystemet og øke forsyningssikkerheten (Fortum, u.å.). Ved å kombinere fleksible energikilder åpner man muligheten for å tilføre flere former for fornybar energi i perioder hvor man for eksempel kanskje har for lite energi fra en kilde og må skape energi fra en annen. For eksempel ved liten tilgang på vindenergi, hvor man kan supplere med vannkraft for å dekke forsyningsgapet. I Norge fungerer våre vannkraftverk som våre batterier i form av vannmagasiner. Energikilder som ikke kan reguleres, slik som vind og sol, bidrar også, og sammen utgjør de en balanse i strømproduksjonen året rundt (Fortum, u.å.).

2.1.4. Fremtiden i norsk strømmarked

Som nevnt innledningsvis vet vi at behovet for elektrisitet vil øke kraftig frem mot 2030, og ifølge Statnett (u.å.) kan vi i Norge med høy sannsynlighet oppleve et kraftunderskudd allerede i 2027. Vi vil i så fall bli enda mer avhengig av import fra andre land for å skaffe tilstrekkelig med elektrisitet for å dekke alle behov. I perioder med høyt forbruk av strøm må produksjonen økes for å møte økt etterspørsel, og vi er avhengige av å overføre allerede produsert strøm for å frigjøre kapasitet til ny produksjon. Det vil være en fordel å kunne forutsi forbruket for å tilpasse produksjonen og overføringen av strømmen. NVE (2023) rapporterer om store ambisjoner for utbygging av fornybare energikilder. Frem mot 2030 ser de for seg at Sverige vil stå for den største andelen av økt produksjon av vindkraft, og at Norge vil komme etter mot 2040. Produksjonskapasiteten for solkraft antas å ha en jevn økning i Norge fra i dag og frem mot 2040. Videre påpeker de at et kraftsystem bestående av en stor andel fornybare energikilder krever mer fleksibilitet for å dekke kraftbehovet i topplasttimene, og for å balansere systemet. Vi ser allerede i dag perioder med stort overskudd av sol- og vindkraft, samt raske svingninger i kraftproduksjon og strømpriser. En høy andel ikke-regulerbar kraft fra vind og sol vil som nevnt tidligere kreve løsninger for fleksibilitet i stort omfang. Teknologier som batterier og hydrogenproduksjon er fortsatt veldig dyre eller ikke modne nok til å håndtere fleksibilitetsutfordringer i stor skala, men det antas at det vil bli gjort større investeringer innen utvikling av batterier og hydrogen de neste årene (Norges vassdrags- og energidirektorat, 2023). Vi kan med bakgrunn i det NVE legger frem si at det med stor sannsynlighet vil skje en del endringer og utbedringer frem mot 2030, og ikke minst frem mot 2040. Strømproduksjonen kan utnyttes mer effektivt frem til disse utbedringene er på plass, da det kan oppstå underskudd av strøm i perioden frem til forholdsvis 2030 og 2040.

2.1.5. Innvirkninger på norsk strømforbruk

Strømforbruket i Norge kan påvirkes av en rekke faktorer. Ifølge (Energifakta Norge, 2024b) kan årlige variasjoner i energibruk forklares av blant annet svingninger i værforhold og pris. Kang og Reiner (2021) diskuterte effekten av værvariabler på husholdningers strømforbruk i Irland, hvor de eksperimenterte med værvariablene temperatur, regn og solskinn. Deres funn indikerer at temperaturen har en jevn effekt på strømetterspørselen gjennom hele dagen, mens regn og solskinn er mer individuelt betinget og derfor varierer også effekten.

Teorien går ikke i dybden om hvordan strømprisene direkte påvirker strømforbruket, men det finnes en rekke statistikk som viser en trend på fall i strømforbruk ved høye strømpriser. Blant annet skriver (Statistisk sentralbyrå, 2023) om et markant fall i husholdningenes strømforbruk i 2022, ned hele 11 %

fra 2021 og det laveste forbruket siden 2014. Videre skriver Statistisk sentralbyrå (u.å.) at 2022 var et år preget av unormalt høye priser gjennom hele året. På en annen side finnes mer teori om tilbud og etterspørsel, og hvordan pris ofte bestemmes av etterspørselen. Strømprisen påvirkes av tilbud og etterspørsel, og avhenger av hvor mye ressurser som er tilgjengelig for strømproduksjon og hvor stort behovet hos forbrukerne er (NorgesEnergi, 2023). Med andre ord blir ikke selve forbruket direkte påvirket av strømprisen, men det er heller prisen som påvirkes av forbruket. Det kan dermed være interessant å se på strømprisen mot forbruket av strøm for å studere hvordan en uavhengig variabel som strømpris kan bidra i en prediksjon av strømforbruket.

Basert på Energifakta Norge (2024b) sin påstand om at årlige variasjoner i strømforbruket kan påvirkes av vær og pris, inkluderes disse variablene i arbeidet med studien. Temperaturen blir av Kang og Reiner (2021) nevnt som den variabelen som har mest effekt på forbruket gjennom en dag, og at nedbør og solskinn har mer variabel effekt på strømforbruket. I arbeidet med prediksjonsmodellene vil derfor disse værvariablene benyttes for å studere hvordan kombinasjonen av disse kan påvirke en prediksjonsmodell. Strømpris inkluderes også for å undersøke om den vil bidra positivt i prediksjonen sammen med de nevnte værvariablene. Alle variablene vil forklares mer inngående i kapittel 3 og 4.

Det norske strømmarkedet oppsummert har altså de grunnleggende funksjonene produksjon, overføring og omsetning av elektrisitet. Norge, som den største vannkraftprodusenten i Europa, utnytter sesongvariasjoner i nedbør og smeltevann for å fylle sine kraftmagasiner. Markedet krever balanse mellom produksjon og forbruk. Statnett spiller en nøkkelrolle ved å håndtere eventuell ubalanse gjennom markedet, samt ved å kjøpe fleksibilitet for å justere for uforutsette hendelser. Fleksibilitet i kraftsystemet avgjørende for å håndtere belastninger og fremme fornybar energi, hvor Norge benytter vannkraftmagasiner som «batterier» og supplerer med andre energikilder som vind og sol for å opprettholde balansen. Med økende behov for elektrisitet mot 2030 og potensielt kraftunderskudd, er fremtiden i det norske strømmarkedet avhengig av integrasjon av fornybare energikilder og utvikling av nye lagringsteknologier som batterier og hydrogen.

2.2. Maskinlæring

Data mining er navnet på prosessen for uthenting av kunnskap og innsikt fra store mengder data hvor hensikten er å oppdage mønstre, korrelasjoner, trender eller andre relevante opplysninger fra store, og ofte uoversiktlige mengder med data (Sharda et al., 2018). Målet er å trekke ut informasjon fra denne datamengden og gjøre den om til et forståelig format for videre bruk. Alpaydin (2014) skriver at maskinlæring er en type kunstig intelligens som tillater datamaskiner å bli mer nøyaktig til å predikere

utfall, uten å være direkte programmert til å gjøre det. Dette er i tråd med Sharda et al. (2018) sin forklaring om at maskinlæring er en type data mining som gir datamaskiner evnen til å lære og forbedre seg fra erfaring uten å være eksplisitt programmert for å gjøre nøyaktig det. Maskinlæring har blitt benyttet til å automatisere kostbare prosjekter ved å gjennomføre flere runder med prøving og feiling på store datasett med data hvor resultatene deretter samles og analyseres, og feil kontinuerlig rettes opp underveis (Henderson et al., 2017). Maskinlæring er derfor en mer automatisert og rimeligere prosess for å hente ut informasjon fra data, da den forbedrer maskiners evne til å lære og forbedre seg. Maskinlæring benyttes i prosesser hvor vi ikke klarer å identifisere alt som skjer, men tror vi kan konstruere en god og nyttig tilnærming hvor vi er i stand til å redegjøre for en del av dataene og finne mønstre eller regelmessigheter (Alpaydin, 2014). Maskinlæringsmodeller kan være deskriptive og si noe om hva som *har* skjedd, eller prediktive og dermed skal si noe om hva som skal skje. I denne oppgaven benyttes prediktive modeller for å studere hvordan strømforbruket vil kunne se ut i fremtiden. Dersom vi klarer å identifisere mønstre eller regelmessigheter kan det skape en forståelse for sammenhenger, og på den måten kan det predikeres fremtidig utfall (Alpaydin, 2014). Vi kan derfor anta at maskinlæring vil være en passende teknologi for vårt formål og at det i henhold til teorien vil kunne gi oss gode prediksjoner.

2.2.1. Typer algoritmer i maskinlæring

Maskinlæring utforsker konstruksjonen og studien av læringsalgoritmer (Henderson et al., 2017). I data er det visse mønstre og når man skal forsøke å løse problemer eller finne mønstre ved hjelp av en datamaskin, trenger vi en algoritme. En algoritme er en sekvens med instruksjoner som skal følges for å transformere input til output (Alpaydin, 2014). Det finnes mange applikasjoner med mye data hvor man ikke har en algoritme, derfor vil vi at maskinen automatisk skal trekke ut algoritmen for ulike oppgaver. Dette skjer ved at man har en modell som er definert av visse parameter, hvor læringen skjer gjennom utførelsen i et dataprogram for å optimalisere parameterne i modellen ved å bruke eksempeldata eller tidligere erfaringer. Læringsalgoritmer blir av Henderson et al. (2017) delt inn i tre ulike typer: *Supervised learning, Unsupervised learning* og *reinforcement learning*. *Supervised learning*, eller veiledet læring, baserer seg på gitte regler og brukes på merket data. Denne formen for læringsalgoritmer er vanlig å bruke på regresjonsoppgaver. *Unsupervised learning*, eller ikke-veiledet læring, baserer seg på å identifisere skjulte mønstre i dataen og har ingen forhåndsbestemte regler. Ikke-veiledet læring er ofte brukt for å trene klassifiseringsmodeller. *Reinforcement learning*, eller forsterkingslæring, baserer seg på å oppnå et spesifikt mål og læringen skjer gjennom et belønningssystem. I vår studie benyttes input bestående av merket data med tydelig inndelte verdier

og kategorier, og derfor er en veiledet læringsalgoritme passende. Prediktiv analyse er ifølge Henderson et al. (2017) en underkategori av veiledet læring. I prediktiv analyse forsøker brukere å modellere dataelementer, samt forutsi fremtidige utvalg gjennom evaluering av sannsynlighetsestimater (Henderson et al., 2017). Algoritmen kan dermed sammenligne sine prognoser med de verdiene vi gir den som testdata, og på den måten gi direkte tilbakemelding på hvor bra den estimerer.

2.2.2. Tidsserier

I studien benyttes tidsseriedata som både input og output, og vi skal gjennomføre en tidsserieanalyse. En tidsserieanalyse er sammenligning av tall over tid for å se om noe avtar eller øker, forbedres eller forverres. En tidsserie kan defineres som en rekke observasjoner som er registrert over tid, hvor hver observasjon er registrert til et spesifikt tidspunkt (Hyndman & Athanasopoulos, 2018). Dette skiller tidsserier fra andre typer data fordi rekkefølgen på observasjonene ikke er tilfeldig. Mange maskinlæringsmodeller er bygget på at dataen er uavhengig, dette må tas hensyn til ved valg av maskinlæringsmodell når vi skal benytte tidsseriedata. Tidsserieprognoser forutsetter at alle forklaringsvariablene aggregeres og forbrukes slik responsvariablene oppfører seg med tanke på tidsvariant (Sharda et al., 2018). De to mest brukte tilnærmingene til prediksjon med tidsseriedata er, ifølge Hyndman og Anthanasopoulos (2018), eksponentiell glatting og ARIMA-modeller. Ifølge Sharda et al. (2018) er de mest populære teknikkene for prognoser av tidsseriedata enkelt gjennomsnitt, glidende gjennomsnitt, vektet glidende gjennomsnitt og eksponentiell glatting, hvor mange av disse har avanserte versjoner hvor sesongvariasjoner og trender kan bli regnet med for mer nøyaktige prognoser. Både ARIMA og eksponentiell glatting er avanserte prognosemodeller som kan være tidkrevende å jobbe med.

Så langt kan vi oppsummere med at maskinlæring, som er en underkategori av kunstig intelligens, tillater datamaskiner å lære og forbedre forutsigelsesevnen uten eksplisitt programmering. Prosessen innebærer å utvinne kunnskap gjennom å identifisere mønstre, korrelasjoner og trender i dataene. Ulike typer læringsalgoritmer presenteres også: veiledet læring, ikke-veiledet læring og forsterkningslæring, hver tilpasset ulike datatyper og problemstillinger, hvor denne oppgaven benytter veiledet læring. Ved bruk av tidsseriedata i analyser er teknikker som ARIMA og eksponentiell glatting hyppig anvendt, men disse kan være tidkrevende å jobbe med. Denne studien har som formål å teste mindre tidkrevende maskinlæringsmodeller på tidsseriedata. Det vil studeres om modellene i studien tillater mindre forarbeid og krav til tilpasning underveis uten at det går ut over deres evne til å predikere.

2.3. Prognosemodeller

I utforskningen av prediktive modeller står vi overfor en mangfoldighet av tilnærminger. Dette delkapittelet tar for seg hvordan de utvalgte maskinlæringsmodellene, både enkle og komplekse, blir evaluert basert på deres anvendelighet i spesifikke scenarioer. For prediksjonsformål er det ikke en universell akseptert modell som kan anses som den «beste» og som vil fungere på alle typer problemer. Sharda et al. (2018) skriver at å finne den beste modellen avhenger av scenarioet som skal analyseres, og det er kun mulig å finne gjennom en omfattende prøve-og-feile-eksperimentering. Prediktiv analyse sikter på å fastslå hva som er sannsynlig at skal skje i fremtiden og det er en rekke teknikker som er benyttet for å utvikle applikasjoner for prediktive analyser (Sharda et al., 2018). Prognoselitteraturen uttrykker en preferanse for enklere modeller med mindre det blir gjort sterke argumenter for kompleksitet (Collopy et al., 1994). Valg av prognosemodeller i denne oppgaven har basert seg på maskinlæringsmodeller som er populære og mye brukt, i tillegg til av det er valgt en kombinasjon av enkle og komplekse modeller. Vi kjenner til noen av disse modellene fra tidligere. De utvalgte modellene er Lineær Regresjon, Random Forest, Gradient Boosting, Ridge, LASSO og Support Vector Machine. Modellene får minimal med tilpasning, som skal sikre et rettferdig sammenligningsgrunnlag med bakgrunn i formålet vårt.

I oppgaven benyttes en pakke i programmet *RStudio* som heter *Caret*. *Caret* tilbyr enkle, håndterbare maskinlæringsprosesser, og bruker et konsistent grensesnitt for hundrevis av maskinlæringsmodeller. Dette gjør det enkelt å skifte mellom ulike modeller og eksperimentere underveis i prosessen. En mulig svakhet med pakken er at den ikke er spesifikt designet for tidsseriedata, men for generelle maskinlæringsoppgaver. *Caret* kan derfor potensielt gå glipp av de unike egenskapene tidsseriedata innehar, som for eksempel sesongvariasjoner og trender, men dette avhenger blant annet av datagrunnlag og tilpasning. Gjennom en prøve-og-feile-metodikk blir det mulig å identifisere hvilke modeller som best håndterer de spesifikke kravene og datamengdene de anvendes på. Modellen som skal testes er, som vi så i innledningen, Lineær Regresjon, Random Forest, Gradient Boosting, Ridge, LASSO og Support Vector Machine. Vi vil nå presentere modellene og deres egenskaper. Videre i studien skal det undersøkes hvor godt disse typene modellene fungerer for prediksjon med enkel programmering. For å avgjøre om modellene vi tester er verdt å bruke i gitt situasjon, sammenligner vi de mot en enkel maskinlæringsmodell basert på Lineær Regresjon. Denne modellen setter vi som vår benchmark-modell, og er den første som blir presentert under.

2.3.1. Lineær Regresjon

Regresjonsmodeller er basert på veiledet læringsalgoritme, og regresjon regnes som en av de mest brukte metodene for analyse og økonometri (Ramasubramanian & Moolayil, 2019). Berk (2020) skriver at en tradisjonell lineær regresjonsmodell er den grunnleggende prosessen for å oppdage relasjoner i data, men at det også er en rekke praktiske implikasjoner knyttet til denne enkle formen for modell. Blant annet er det vanskelig å vite om en modell matcher hvordan dataen faktisk er frembragt, og det er derfor ingen måte å vite nøyaktig hvor nærme man er den faktiske sannheten. Hvis man derimot vet den faktiske sannheten, ville det ikke noe behov for å analysere dataen til å begynne med (Berk, 2020). Noe som vil si at alle spekulasjoner rundt modellspesifikasjon, bygger seg på spørsmålet om modellen faktisk er «god nok». Lineær regresjon er en enkel tilnærming for veiledet læring og er å anse som et nyttig verktøy for å predikere en kvantitativ respons (James et al., 2021). Til tross for at det er en metode som har vært med lenge før de mer kompliserte modellene, er den fortsatt mye brukt (James et al., 2021). James, Witten, Hastie og Tibshirani (2021) forklarer at lineær regresjon brukes for å forklare relasjonen mellom responsvariabel Y og et sett variabler $X_1, X_2, \dots, X_p$ og viser til en standard lineær modell slik:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$ *Formel 2.1 – Standard lineær modell*

Hvor β_0 er konstantleddet, som vil si den forventede verdien av Y når $X = 0$. $\beta_1, \dots, \beta_p$ er helningsgraden på linjen, og ϵ er en tilfeldig, normalfordelt feilterm. Klassisk regresjon i en tidsseriekontekts kjennetegnes ved at hver observasjon av en variabel er observert etter en bestemt rekkefølge i n tid, for eksempel $z_{t1}, z_{t2}, \dots, z_{tn}$. Formelen kan ifølge da skrives slik:

$$\hat{x}_t \approx \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t$$ *Formel 2.2 – Lineær modell i tidsseriekontekst*

I en maskinlæringskontekst vil man ikke trenge å anse formel 2.2 eller 2.3 som en ukjent funksjon av en eller flere prediktorvariabler som brukes for å finne ut hvordan den avhengige variabelen påvirkes av den uavhengige. Det eneste man trenger å ta stilling til er spørsmål som omhandler hvilke prediktorvariabler som skal brukes, hvordan data skal forberedes, og hva som skal gjøres med eventuelle avhengigheter mellom mulige forstyrrelser (Berk, 2020). Resten av prosessen vil skje nærmest automatisk med det rette verktøyet. Berk (2020) forklarer prosessen som en *Black Box Algorithm*, som vil si at man kun tar stilling til de ulike prediktorvariablene som må gis til algoritmen for

å få responsvariabelen som output. Hva som skjer mellom input og output består av komplekse beregninger som har ingen informasjonsverdi for analytikeren.

2.3.2. Ensemble-modeller

Her presenteres statistisk læring i form av regresjonstrær og modeller som benytter seg av samlinger med enkle modeller. Først forklares begrepet *ensemble modeling* som en samlebetegnelse for to av de ulike typene maskinlæringsmodeller vi kommer til å teste ut i analysen; Random Forest og Gradient Boosting. Begge modellene er populære, og skal være både kraftige og anvendbare. *Ensemble modeling* er en populær teknikk brukt i både klassifiserings- og regresjonsoppgaver hvor det er behov for en modell med kraftig ytelse. Denne typen teknikker innebærer en prosess hvor flere enkle modeller bygges for å løse samme problem og man deretter kan kombinere resultatene til en enkel output, gjerne ved å beregne gjennomsnittet av alle resultatene (Ramasubramanian & Moolayil, 2019). James et al. (2021) forklarer teknikken som en kombinasjon av såkalte *weak learners*, eller svake læringsmodeller, for å oppnå enkle, men kraftige modeller. Dette vil si at modellen tar i bruk enklere metoder, som alene ikke ville gitt gode resultater, men som samlet gir bedre prediksjoner. *Bagging* og *boosting* er to av de mest populære tilnærmingene innen *ensemble modeling*. Med *bagging*, eller *Bootstrap*-aggregering, menes modeller som bygges parallelt med en liten grad av randomisering i hver modell, hvor resultatene bestemmes av en enkel stemmemekanisme. *Boosting* derimot er modeller som bygges sekvensielt, hvor resultatet fra den første modellen brukes for å tune den neste modellen. Som vil si at hver modell lærer fra den forrige og forsøker å forbedre for hvert steg. Resultatene av en slik modell blir vanligvis et vektet gjennomsnitt av alle utfallene (Ramasubramanian & Moolayil, 2019).

Random Forest

Random Forest er en populær teknikk som hører til under *bagging*-begrepet, og er en modell som fungerer både for regresjons- og klassifiseringsoppgaver. Det er en effektiv metode for å redusere *overfitting*, eller overtilpasning, uten mye innsats. Random Forest er en *ensemble*-teknikk som går ut på å bygge flere modeller og kombinerer disse, for å så finne det samlede resultatet ved hjelp av en stemmemekanisme (Ramasubramanian & Moolayil, 2019). Random forrest har beslutningstre som basemodell, og som navnet tilsier vil modellen tilegne hver modell en grad av randomisering og være bestående av flere beslutningstre-modeller (James et al., 2021). Dette vil si at modellene som bygges med en slik teknikk ikke er helt like, men at de har noen forskjeller fra hverandre. Denne typen modeller er gode til å håndtere store datamengder med høydimensjonale data, og kan håndtere manglende

verdier på en god måte. I tillegg til dette viser modellen seg spesielt godt i tilfeller med et stort antall korrelerte prediktorvariabler (James et al., 2021). En svakhet ved *bagging*-modeller er at de bare tar en andel av dataen som deles fra beslutningstre til beslutningstre, og at de tilpassede verdiene derfor ikke vil være uavhengige (Berk, 2020). Det følgende kan bety at gjennomsnittet som beregnes etter alle iterasjoner ikke er så effektivt som det kunne vært. I tillegg kan man risikere at funksjonen som brukes for å tilpasse modellen er konsekvent og vesentlig upassende for problemet man skal se på, slik at feil i allerede tilpassede verdier bare fortsetter å reproduseres om igjen og dermed påvirker gjennomsnittet til slutt (Berk, 2020). Random Forest skal imidlertid være en robust og fleksibel modell som virker til å fungere godt i praksis, og er ifølge Berk (2020) en modell som konsekvent presterer bedre enn de fleste andre modeller dersom prediksjon med blant annet sosiale data, er hovedmålet.

Gradient Boosting

Gradient Boosting er som Random Forest en modell som fungerer både for regresjons- og klassifiseringsoppgaver. Gradient Boosting er en *boosting*-teknikk hvor med en taps-funksjon og en svak læringsmodell, som for eksempel regresjonstrær, vil forsøke å finne en modell som kan minimere taps-funksjonen (Kuhn & Johnson, 2013). Slik vil modellen fortsette i et gitt antall omganger, ved å kalkulere en *gradient*, eller helning, som modellen bruker for å tilpasse den neste. *Boosting*-modeller skiller seg fra Random Forest på flere måter, blant annet er det lav eller ingen grad av tilfeldigheter i tradisjonell *boosting* (Berk, 2020). Med andre ord vil dette forklares som at for hver eneste iterasjon som kjøres, kjøres modellen på hele treningsdataen og på alle prediktorer uten å ta for seg tilfeldige utvalg for hver gang. På grunn av oppsettet med flere parallelle regresjonstrær, gjør Random Forest det mulig å redusere varians ved å velge de sterkeste av de svake læringsmodellene som har lav bias, samt på en mer effektiv måte enn boosting, ifølge Kuhn og Johnson (2013). Modeller bygget på Gradient Boosting krever noe mer kraft sammenlignet med Random Forest fordi de forskjellige enhetene kjøres sekvensielt i stedet for parallelt (Kuhn & Johnson, 2013). Random Forest og Gradient Boosting er begge effektive på større mengder data og skal være gode på prediksjon, men skiller seg på forberedelser da Gradient Boosting trenger mer tuning av parametere for å oppnå sitt beste potensiale.

2.3.3. Regulariseringsmodeller

Gitt standard antakelser, vil koeffisientene som produseres av minste kvadraters metode være uten bias, som også vil si at de også har den laveste variansen. Ved å bruke MSE, som er en kombinasjon av varians og bias, er det mulig å lage modeller med lavere verdi av MSE ved å justere

parameterestimatene til å være bias. Ved en liten økning av bias, vil variansen reduseres, og dermed kunne gi en lavere MSE enn koeffisientene fra minste kvadraters metode (Kuhn & Johnson, 2013). Dette blir ofte omtalt som *Bias variance tradeoff*. Kuhn og Johnson (2013) skriver at en konsekvens av høy grad av korrelasjon i varians i prediktorvariabler er at variansen er høy. Ved å bruke modeller med bias og lavere verdi av MSE, kan man overvinne kollinearitet. En av metodene for å lage regresjonsmodeller med bias er å tillegge funksjonen for summen av kvadratfeil, eller sum of squared errors (SSE), en straffefunksjon. SSE er gitt ved:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Formel 2.3 – Sum av kvadratfeil

Med en modell som er overtilpasset dataen eller har høy grad av kollinearitet, vil man oppleve et oppblåst estimat på den lineære regresjonsparameteren. Hvis dette er tilfelle, vil vi ha et behov for å kontrollere omfanget av estimatene for å redusere SSE (Kuhn & Johnson, 2013). For å kontrollere, eller regularisere estimatene, legger vi til en straffefunksjon til SSE. Ridge og LASSO er to regulariseringsteknikker som er utviklet for å oppnå dette.

Ridge (L2 regularisering)

Ridge regresjon ilegger summen av kvadrerte regresjonsparameter en straffefunksjon. L_2 betyr at det brukes en *second-order*-straff. Effekten av denne straffen er at parameterestimatene kun er tillatt å bli store, hvis det er en proporsjonal reduksjon i SSE. Som vil si at denne metoden krymper estimatene mot 0 ettersom λ-straffen blir større (Kuhn & Johnson, 2013). Dette kalles ofte også for en krympingsmetode. Ved å legge til straffefunksjonen, gjør vi en *trade-off* mellom bias og varians og kan dermed også oppnå lavere verdier av MSE enn med en modell uten bias. Mens Ridge regresjon krymper estimatene mot 0, vil ikke modellen sette verdien helt til 0 for noen av verdiene i straffen. Selv om enkelte parameterestimater settes til meget lave verdier, vil ikke modellen drive variabelseleksjon (Kuhn & Johnson, 2013). Dette vil si at modellen ikke vil utelukke de koeffisientene som ikke er lik 0, noe den kan gjøre når man ilegger funksjonen en straffefunksjon slik som i LASSO -regularisering. Sum av kvadratfeil med Ridge-straff er gitt ved:

$$SSE_{L_2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

Formel 2.4 – Sum av kvadratfeil med L2-straff

LASSO (L1 regularisering)

Et alternativ til Ridge er LASSO, eller *Least Absolute Shrinkage and Selection Operator*. Denne modellen bruker en lignende straffefunksjon som Ridge. Når regresjonskoeffisientene fortsatt krympes mot 0 er en konsekvens av å straffe verdiene at noen parametere settes til 0 for en verdi av λ. LASSO bruker med andre ord, en form for regularisering for å forbedre modeller og drive variabelseleksjon. Gjennom variabelseleksjon kan modellen utelukke eventuell støy eller unødvendige variabler i dataen. Sum av kvadratfeil med LASSO er gitt ved:

$$SSE_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j|$$

Formel 2.5 – Sum av kvadratfeil med L1-straff

LASSO har, ifølge James et al (2021), en fordel over Ridge for hvor enkelt det er å tolke resultatene, men det er ingen ting som tyder på at den ene er universelt bedre enn den andre. Det forventes at LASSO fungerer bedre på oppgaver hvor data har et mindre antall prediktorer med betydelige koeffisienter, og de resterende prediktorene er små.

2.3.4. Support Vector Machine

Support Vector Machines, eller SVM, er en gruppe kraftfulle og fleksible teknikker for prognosemodeller som i utgangspunktet ble laget for klassifiseringsoppgaver, men som etter hvert også viste seg å være godt egnet for regresjonsoppgaver (Kuhn & Johnson, 2013). Lineær Regresjon forsøker å finne parameterestimater som minimerer SSE (sum of squares error), forskjellen mellom faktisk og estimert verdi, men en liten ulempe med dette er at ved å minimere SSE kan man risikere at parameterestimatene påvirkes av en enkelt observasjon som faller langt utenfor trenden i data. Dersom data inneholder observasjoner med sterk påvirkningskraft, er en måte å unngå dette på er å bruke Huber-funksjon for å finne de beste parameterestimatene (Kuhn & Johnson, 2013). SVM-modeller bruker en lignende funksjon som denne, men skiller seg ved at funksjonen blir ilagt en margin slik at residual fra datapunkter innen denne marginen ikke bidrar til regresjonstilpasningen mens datapunktene med en absolutt forskjell større enn bidraget fra innenfor marginen bidrar med en lineær skalert mengde (Kuhn & Johnson, 2013). SVM-modeller kan bruke forskjellige kernelfunksjoner for å utvikle rommet for egenskaper. En kernel defineres som en funksjon som kvantifiserer likheten mellom to observasjoner, og hjelper med å gjøre det enkelt å modellere ikke-lineære relasjoner mellom input og målvariabel (James et al., 2021). Et eksempel er *Radial Kernel,* som er gitt ved:

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$ *Formel 2.6 – Radial Kernel*

En fordel ved å bruke kernelfunksjoner kan sies å være effektiviteten, da man kun trenger å kjøre $K(x_i, x_{i'})$ på alle $\binom{n}{2}$ par i, i' (James et al., 2021). Man trenger ikke like mye kraft for å gjøre dette, som om man skulle gjort det samme med hele rommet. SVM kan med dette brukes på oppgaver hvor vi står ovenfor komplekse relasjoner mellom input og målvariabel. Fordi man ikke bruker kvadrert residual, har store outliers en begrenset effekt på regresjonsfunksjonen (Kuhn & Johnson, 2013). SVM har flere styrker, blant annet fordi den er å regne som en robust form for regresjon som kan være nyttig i tilfeller hvor ekstreme residual må vektes mindre. Berk (2020) skriver at SVM egner seg veldig godt i tilfeller hvor antallet prediktorer og antallet observasjoner ikke er så alt for store, og at bruken av kernel er fordelaktig for denne typen modell. Valg av maskinlæringsmodeller baseres på formålet for studien som skal gjennomføres og gjennom en utvelgelsesprosess som består av ulike tester. Den ene modellen kan være passende for en type studie med et gitt datasett, men være ubrukelig i andre tilfeller. Ofte er en nødt til å gjennomføre flere ulike typer tester av justeringer i modellbyggingen også. For å vurdere hvor godt modellene fungerer kan man benytte seg av ulike former for nøyaktighetsmålinger, som har som formål å gjøre det enkelt å sammenligne ulike resultater fra ulike modeller. Vi vil i neste del gå gjennom hva som ligger bak de ulike nøyaktighetsmålingene.

2.3.5. Modellenes egenskaper

I delkapittel 2.3. har en rekke prediktive modeller innen maskinlæring blitt utforskrev, med særlig fokus på deres anvendelse og effektivitet i ulike scenarier. Lineær regresjon, presentert som en grunnleggende metode for veiledet læring, blir vurdert for sin evne til å modellere kvantitative responser gitt at visse statistiske forutsetninger oppfylles, som lineær relasjon og homoskedastisitet. På tross av sin enkelhet, krever lineær regresjon nøye vurdering av dataenes egenskaper for å sikre nøyaktige resultater.

Ensemble-modeller, inkludert Random Forest og Gradient Boosting, blir diskutert for deres kraftige ytelse i komplekse prediksjonsoppgaver. Disse modellene utnytter styrken av flere svake læringsmodeller for å forbedre nøyaktigheten og robustheten i prediksjonene, noe som gjør dem ideelle for store og komplekse datasett. Regulariseringsmodeller som Ridge og LASSO tilbyr teknikker for å takle overtilpasning og multikollinearitet gjennom straffefunksjoner som modifiserer

regresjonskoeffisientene for å fremme en mer stabil og tolkbar modell. Disse metodene er spesielt verdifulle når man arbeider med datasett som inneholder mange korrelerte prediktorer. Til slutt utforsker vi Support Vector Machines, som er spesielt nyttig for både klassifisering og regresjon i tilfeller hvor forholdet mellom prediktorer og responsvariabelen er komplekst og ikke-lineært. Ved bruk av kernelfunksjoner kan SVM håndtere høydimensjonale rom mer effektivt, som dermed gjør det mulig å fange opp subtile mønstre i dataene.

De ulike maskinlæringsmodellene som skal benyttes i studien er modeller som enkelt kan brukes ved hjelp av pakken *caret* i R. Sammen utgjør modellene et robust utvalg for forskning, med verktøy tilpasset alt fra enkle til svært komplekse analytiske utfordringer. Valget og implementeringen av hver modell krever en grundig forståelse av både de underliggende dataene og de spesifikke kravene til analysen, understreket av behovet for kontinuerlig testing og validering for å sikre modellenes effektivitet og relevans.

2.4. Evaluering av modellnøyaktighet

Å estimere hvor godt en modell passer data er et viktig steg i arbeidet med prognosemodeller. Abbass og Hamdy (2021) skriver at vanlige beregninger for å evaluere modellen i regresjonsproblemer er gjennomsnittlige kvadrerte feil og R kvadrert. Sharda et al. (2018) mener at nøyaktigheten til metoden vanligvis vurderes ved å beregne dens feil med *mean absolute error* (MAE), *mean squared error* (MSE) eller *mean absolute percent error* (MAPE), og at selv om disse tre vurderingsmetodene bygger på samme type måling av feilen fremheves forskjellige aspekter ved den. Hyndman og Athanasopoulos (2018) mener at de to mest populære metodene for å måle nøyaktigheten til prediktive modeller er basert på *absolute errors*, altså kvadrerte feil, som *root mean squared errors* (RMSE) og *mean absolute errors* (MAE). Det er med andre ord en viss enighet i hvilke nøyaktighetsmålinger som kan brukes, i tillegg til at mange av disse baserer seg på det samme.

I denne oppgaven vil vi benytte ytelsesmålene RMSE, MAPE og R-kvadrert, som alle er nevnt over. *Mean squared error*, eller MSE, er en funksjon av en modells residual og beregnes ved avstanden mellom den predikerte responsvariabelen for den gitte observasjonen og den faktiske verdien for den samme observasjonen. Funksjonen kvadrerer modellens residual, summerer de og dividerer på antallet prøver (Kuhn & Johnson, 2013). Enklere forklart blir forskjellen mellom prediksjonen og den faktiske verdien beregnet, differansen kvadreres, og deretter summeres alle de kvadrerte forskjellene for alle observasjonene (Abbass & Hamdy, 2021). MSE er gitt ved:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Formel 2.7 – Mean Squared Error (MSE)

Hvor $\hat{f}(x_i)$ er prediksjonen $\hat{f}$ gir for observasjon i. MSE vil være lav hvis den predikerte verdien er nære den faktiske observasjonen (James et al., 2021). MSE gir resultater i form av kvadrerte verdier av målvariabelen, men for å enklere kunne sammenligne mot faktisk målvariabel vil det i mange tilfeller være enklere å benytte seg av RMSE, fordi den måles i samme enhet som målvariabelen. RMSE er en vanlig metode for å karakterisere modellens prediktive evner, og beregnes ved kvadratroten av MSE (Kuhn & Johnson, 2013):

$$RMSE = \sqrt{MSE}$$

Formel 2.8 – Root Mean Squared Error (RMSE)

MAE tar det absolutte avviket mellom predikerte og faktiske verdier og beregner gjennomsnittet av disse, slik at man får den gjennomsnittlige feilen i prediksjonen (Ramasubramanian & Moolayil, 2019). MAPE viser MAE som prosentfeil og kan i mange situasjoner være enklere å tolke, da formatet er mer universelt og gjør det enklere å plassere resultatet på en skala. MAE og MAPE er gitt ved funksjonene:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f_1|$$

Formel 2.9 – Mean absolute Error(MAE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}100 * \left|\frac{y_i - f_1}{y_i}\right|$$

Formel 2.10 – Mean absolute percentage error (MAPE)

Kvaliteten på Lineær Regresjon kan, ifølge James et al. (2021), undersøkes ved å bruke to metoder: *Residual standard error* (RSE) og R^2. RSE er et estimat av standardavviket på ϵ, eller kan med andre ord forklares som gjennomsnittet av hvor mye responsvariabelen avvike fra regresjonslinjen. RSS, eller *residual sum of squares*, måler graden av varians i feiltermen/residual. En lav verdi av RSE, hvis $\hat{y}_i \approx y_i$ når $i = 1, \dots, n$, er en indikasjon på at modellen passer data godt (James et al., 2021). RSE kan gi en indikasjon på *lack of fit* i modellen på data, men fordi den måles på enheter av Y, kan det være noe utfordrende å vite akkurat hva som er en god RSE. Dermed kan man benytte seg av en alternativ metode R^2, også kalt R-kvadrert. R-kvadrert kan forklare variansen i et standardformat i form av en verdi mellom 0 og 1 hvor TSS «total sum of squares» $= \sum(y_i - \bar{y}_i)^2$ forklarer den totale variansen i responsvariabel Y og kan anses som summen av variasjonen i responsen før regresjonen er utført. Med

andre ord forklarer R-kvadrert brøkdelen av varians som utgjøres av modellen, og det er en slags prosentandel som rapporterer et tall fra 0 til 1, og når R-kvadratet er nær 1 indikerer dette vanligvis at mye av variasjonene i dataene kan forklares av selve modellen (Abbass & Hamdy, 2021). R-kvadrert er altså et mål på det lineære forholdet mellom X og Y som korrelasjon (James et al., 2021). RSE og R-kvadrert er gitt ved:

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

Formel 2.11 – Residual standard error (RSE)

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Formel 2.12 – R kvadrert

Samlet sett har nøyaktighetsmålingene som formål å sikre at prediksjonene er både nøyaktige og relevante. De utvalgte nøyaktighetsmålingene RMSE, MAPE og R-kvadrert, vil i studien sørge for en grundig evaluering av modellenes ytelse gjennom ulike målinger.

2.4.1. Påvirkning på modellytelse

Prediksjonsmodellenes ytelse kan påvirkes av mange faktorer, hvor støy er en viktig faktor. Kuhn og Johnson (2013) skriver om generelle former for støy som kan påvirke modellenes prediktive evne. Systematisk støy knyttet til måling kan forekomme når det er mange prediktorer som skal måles, og denne formen for støy vil med høy sannsynlighet forplantes gjennom prediksjonsligningen som igjen gir dårligere prediksjonsevne (Kuhn & Johnson, 2013). For eksempel har en værstasjon som måler temperatur et instrument som har sin egen grad av nøyaktighet som kan inneholde små, men konsekvente og forutsigbare feil. Denne feilen kan blant annet være at temperaturmåleren viser et par grader for mye, hvor feilen da vil bli en del av beregningen for prediksjonen av været. En annen form for støy som kan oppstå er knyttet til ikke-informative prediktorer, altså prediktorer som ikke har noen relasjon til responsvariabelen (Kuhn & Johnson, 2013). For eksempel å inkludere en variabel med hvilken farge huset som bruker strømmen har. Enkelte modeller, som blant annet LASSO-regularisering kan identifisere og filtrere ut denne typen variabler automatisk, mens enkelte modeller ikke klarer å fange opp denne støyen og dermed kan risikere modeller trent på støy (Kuhn & Johnson, 2013).

I dette delkapittelet adresseres det hvordan støy i data kan påvirke prediktiv nøyaktighet. Støy fra målefeil eller irrelevante prediktorer kan redusere modellenes evne til å levere nøyaktige prediksjoner. Dette fremhever behovet for nøye utforming og valg av prediktorvariabler.

2.5. Tidligere forskning

Albuquerque et al. (2022) demonstrer hvordan regulariserte maskinlæringsmodeller effektivt kan forutsi strømforbruket i Brasil. Artikkelen sammenligner utvalgte maskinlæringsmodeller med tradisjonelle prognosemetoder som Random Walk og ARIMA. Resultatene deres viser at maskinlæringsmodeller, spesielt Random Forest og LASSO Lars, gir betydelig forbedrede prediksjoner for alle horisonter dem predikerer på. Forfatterne bruker en utvidelse av en Lars-funksjon (*Least angle regression*) som legger til regulariseringsegenskapen til en LASSO-modell slik at modellen kan drive variabelseleksjon (Albuquerque et al., 2022). Studien viser videre at prediksjonsfeil i gjennomsnitt blir halvert sammenlignet med benchmark-modellene. De skriver at funnene deres indikerer en signifikant forbedring i prediksjonsevnen og er spesielt interessant for veldig kortsiktige tidshorisonter. For lengre prediksjonshorisonter, blir andre variabler som vær- og kalenderdata fremhevet som viktige for å forbedre nøyaktigheten. Det er også verdt å nevne at de benytter høydimensjonale data i sin artikkel.

Shapi et al. (2021) etter søkelys på utfordringene som møtes i prediksjon av energiforbruk på grunn av lav prediksjonsnøyaktighet. Målet med studien deres er å utvikle en prediktiv maskinlæringsmodell med høyere nøyaktighet, hvor de eksperimenterer med Support Vector Machine, kunstige nevrale nettverk, og k-Nearest Neighbour. For å avgjøre ytelsen til hver av metodene sammenlignes modellnøyaktigheten via målinger på RMSE, NRMSE og MAPE. Shapi et al. (2021) konkluderer med at det på gjennomsnittlig energiforbruk, viser seg å være Support Vector Machine som er den beste modellen for å forutsi den månedlige gjennomsnittsverdien, men at den også krever lang treningstid.

Januschowski et al. (2022) undersøkte effektiviteten av trebaserte metoder, spesielt Gradient Boosting trær, i prediksjonssammenhenger. Studiens formål var å utforske hvorfor trebaserte metoder viser seg å være overlegne sammenlignet med andre prognosemetoder, inkludert dyp læring, i denne M5 konkurransen. M5 refererer til den femte utgaven av M-konkurransene, som er en serie av prognosekonkurranser. Funnene i studien indikerer at trebaserte metoder, og spesielt Gradient Boosting trær, utmerket seg i konkurransen ved å levere overlegne resultater over andre metoder. Denne dominansen var noe overraskende gitt dype læringsmetoders fremtredende rolle i tidligere forskning og konkurranser. Årsakene til trebaserte metoders suksess kan tilskrives deres robusthet og evne til å fungere effektivt som *black-box*-modeller, hvor minimal tilpasning er nødvendig for å oppnå

god ytelse. Videre bidro innebygde funksjoner som støtte for diverse tapfunksjoner, håndtering av sparsomme data, og rask trening til deres effektivitet. Studien konkluderer med at trebaserte metoder, spesielt Gradient Boosting trær, har bevist sin effektivitet i M5-konkurransen ved å yte bedre enn andre prognosemetoder, inkludert dype læringsmodeller.

Studiene fra Albuquerque et al. (2022), Shapi et al. (2021) og Januschowski et al. (2022), viser hvordan maskinlæringsmodeller som Random Forest, LASSO Lars, og Gradient Boosting med enkle trebaserte modeller forbedrer nøyaktigheten i prediksjoner av strømforbruk. Albuquerque et al. (2022) fant at maskinlæringsmodeller som Random Forest og LASSO Lars ga betydelig forbedrede prediksjoner sammenlignet med de mer tradisjonelle modellvalgene for tidsserieprediksjon, som for eksempel ARIMA. Denne forskningen støtter avgjørelsen om å teste denne typen prediksjon på modellene som er valgt i denne studien, og forklarer i tillegg hvordan vær- og kalenderdata vil være viktige variabler for langsiktige prediksjonsperioder. Maskinlæringsmodellene overgår tradisjonelle metoder ved å effektivt håndtere store og komplekse datasett, noe som resulterer i betydelige reduksjoner i prediksjonsfeil. Resultatene peker på en fremtid der maskinlæring kan forventes å dominere prediksjoner av strømforbruk, spesielt i situasjoner som krever presisjon over korte og lange tidshorisonter. Av resultatene er det grunn til å tro at Random Forest, LASSO, Support Vector Machine og Gradient Boosting er modeller som kan forventes å prestere godt i prediksjonsformål.

3. Metode og data

Dette kapittelet har som formål å gi en beskrivelse av forskningsmetodene som er benyttet for å kunne svare på studiens problemstilling og forskningsspørsmål. Først presenteres valg av forskningsdesign og forskningsstrategi. Deretter dannes en grunnleggende forståelse av innsamlet data og forberedelser som er gjennomført, før byggingen av de utvalgte modellene beskrives kort. Kapittelet avsluttes med en diskusjon om reliabilitet, validitet og generalitet. Formålet med kapittelet er å forstå forskningsprosessen, og hvordan utvalgte maskinlæringsalgoritmer kan fungere ved prediksjon av strømforbruk hos private husholdninger. Oppsettet i metodekapittelet følger en data mining-prosess som ifølge Sharda et al. (2018) består av seks steg; forretningsforståelse, dataforståelse, dataforberedelse, modellbygging, testing og evaluering, og distribuering. Prosessen er blant annet skrevet for at den skal kunne benyttes av ledelsen i organisasjoner, i en forretningsmessig kontekst. Det første steget «forretningsforståelse» skal beskrive ledelsens behov for ny kunnskap og bedriftens spesifikke mål for studien som skal gjennomføres (Sharda et al., 2018). «Forretningsforståelse» som overskrift er ikke passende i dette tilfellet, og er dermed byttet ut med «forskningsdesign». Uansett skiller ikke begrepene seg veldig fra hverandre, da de begge har som formål å sette studiens ramme for videre prosess.

3.1. Forskningsdesign

Forskningsdesign hører til under begrepet «forretningsforståelse» som det første steget i en data mining-prosess i Sharda et al. (2018), og har som formål å definere hva studien skal gå ut på. Dette blir beskrevet som nøkkelen i enhver data mining-studie. Denne studien baserer seg på problemstillingen beskrevet innledningsvis, som går ut på hvordan predikere strømforbruk i det norske strømmarkedet ved hjelp av maskinlæringsmodeller, med tilhørende forskningsspørsmål som omhandler de uavhengige variablene og modellene som kan brukes for å gjøre nettopp dette. Her vil vi presentere fremgangsmåten for hvordan spørsmålene skal besvares og hvilke metoder som vil benyttes. Først presenteres forskningsdesign og forskningsformål. Et forskningsdesign er en plan for hvordan man skal gå frem i arbeidet med en studie for å besvare problemstillingen og forskningsspørsmålene man har satt seg, og skal fungere som en beskrivelse av veien mot målet med forskningen (Saunders, Lewis & Thornhill, 2023). Definisjonen på forskningsdesign er ifølge Johannessen et al. (2016) «alt» som er knyttet til en undersøkelse, fra man starter med en problemstilling og hele tiden vurderer hvordan det er mulig å gjennomføre studien. Hvilket forskningsdesign som passer best for en studie avhenger av en

rekke faktorer, som for eksempel tidshorisonten man ønsker å studere, etiske hensyn, eller de ressursene man har tilgjengelig (Saunders et al., 2023).

Problemstillingen skal si noe om det faktiske formålet med studien, og det skilles vanligvis mellom utforskende, deskriptiv, forklarende eller evaluerende formål, eller en kombinasjon av disse (Saunders et al., 2023). I vårt tilfelle går forskningsspørsmålet ut på å teste ulike maskinlæringsmodeller og vurdere deres ytelse med tanke på prediksjonsnøyaktighet på et gitt datasettet. Noe som i seg selv peker mot et evaluerende design. Ifølge Saunders et al. (2023) er evaluerende studier i business og ledelse ofte opptatt av å studere faktorer som organisatorisk eller driftsmessig strategi, policy, program eller prosess, og vurdere effektiviteten av dette. Et evaluerende forskningsdesign er nyttig for å måle effektivitet og ytelse (Saunders et al., 2023). Forskningsspørsmålene har som formål å evaluere de utvalgte datakildene som inneholder pris- og værdata som mulige kandidater for uavhengige variabler for prediksjon, samt undersøke prediksjonsnøyaktigheten til ulike maskinlæringsmodeller for å se hvilke som presterer best med de gitte forutsetningene. Saunders et al (2023) forklarer også at en evaluerende studie kan produsere et teoretisk bidrag hvor det legges vekt på at man ikke bare skal forstå «hvor effektivt» noe er, men også «hvorfor», og å sammenligne denne forklaringen med eksisterende teori. Dette er spesielt relevant for å finne ut hvilken maskinlæringsmodell som presterer best på prediksjon av strømforbruk, da det kan være en fordel å også undersøke hvorfor en modell eventuelt kan fungere bedre.

3.1.1 Forskningsmessig tilnærming og valg av metode

For å besvare problemstilling og forskningsspørsmål vil det benyttes en kvantitativ metode. Saunders et al (2023) forklarer at kvantitativ forskning vanligvis er assosiert med en deduktiv tilnærming, som med andre ord vil si at data blir samlet inn og analysert for å teste en teori. I denne oppgaven diskuteres først teorien om maskinlæring generelt, før vi går nærmere inn på de ulike maskinlæringsmodellene. Deretter anvendes disse modellene for prediksjon av tidsseriedata, før resultatene analyseres for hver av modellene til slutt. Fremgangsmåten tilsier dermed at en kvantitativ metode vil være passende for vår studie. Kvantitativ metode gir oss muligheten til å se på numerisk og målbar data, i både store og små mengder, som kan settes i system og analyseres for å besvare forskningsspørsmålet. Kvantitativ forskning utforsker forholdet mellom variabler som samles inn og analyseres ved hjelp av ulike statistiske og grafiske verktøy (Saunders et al., 2023). Analysene som gjennomføres i studien har som formål å undersøke prediksjonsnøyaktigheten til de valgte maskinlæringsmodellene, gitt visse forutsetninger, ved å teste predikerte forbruksverdier mot faktisk forbruk. Det sees også på ulike

parameter for målinger av prediksjonsfeil og nøyaktighet. Formålet er å avdekke hvordan variablene påvirker hverandre.

Vi vil benytte oss av kvantitative sekundærdata fra tre ulike kilder i vår oppgave. Vi går inn med et utgangspunkt for å predikere fremtidige forbruk basert på historisk forbruksdata, og supplerer med sekundærdata fra flere datakilder for å utvide det opprinnelige datasettet med flere variabler som kan være interessante å studere. Dette gjør vi for å skape et bedre bilde av de virkelige faktorer som kan påvirke strømforbruket, og teste om maskinlæringsmodellene kan prestere bedre eller dårligere med flere uavhengige variabler, som strømprisen og mindre forutsigbare faktorer som vær. Det er kun benyttet én teknikk for innsamling av data, og vi kan med andre ord si at vår studie derfor kan klassifiseres som en «kvantitativ monometode»-studie (Saunders et al., 2023). Data om strømforbruk er samlet inn av et norsk strømselskap og deretter delt med oss, og resten av dataen er hentet fra offentlige nettsider. Dataene vi har mottatt og hentet ut har vært i et definert format, og vi kan derfor si at vi har benyttet strukturert sekundærdata (Saunders et al., 2023). I tillegg til å være strukturert, er dataene vi benytter oss av tidsseriedata som er samlet inn på daglig basis over en periode på 13 måneder, som vil si at vi samlet sett har longitudinell sekundærdata fra flere kilder (Johannessen et al., 2016). Prosessen rundt datainnsamling og utvalg vil bli forklart nærmere i senere delkapittel.

3.1.2 Forskningsstrategi

«Forskningsstrategien er den metodiske koblingen mellom filosofi og påfølgende valg av metode for innsamling og analyse av data» (Denzin & Lincols, 2018, gjengitt i Saunders et al., 2023, s.191). Forskningsstrategien skal oppfylle en rekke formål, og det er derfor avgjørende at man velger en passende strategi for prosessen. Forskningsstrategien skal ha en sammenheng med forskningsdesignet og sørge for at man til slutt kan besvare forskningsspørsmålet, samt sørge for at målene med forskningen faktisk oppnås til slutt (Saunders et al., 2023). Vårt forskningsdesign har, som nevnt over, et evaluerende formål, en deduktiv tilnærming og en kvantitativ monometode. Vi benytter oss av sekundærdata samlet inn av strømselskapet, samt data fra Nord Pool hentet fra Forbrukerrådet (Forbrukerrådet, u.å.), og data fra Norsk Klimaservicesenter (Norsk Klimaservicesenter, u.å.). Med dette grunnlaget skal vi ha et godt utgangspunkt og en god strategi for forskningsprosessen vi skal gjennomføre, og et design som passer formålet med studien slik at vi til slutt skal kunne besvare forskningsspørsmålene.

3.2. Dataforståelse

Ulike forretningsoppgaver krever ulike typer data, og dette steget går ut på å identifisere relevant data for oppgaven (Sharda et al., 2018). Formålet med studien er å sammenligne nøyaktigheten til forskjellige maskinlæringsmodeller for prediksjon av strømforbruk på Innlandet i Norge, med fokus på kortsiktige prognoser. Dette skal gjennomføres ved å benytte et datagrunnlag bestående av innsamlede historiske data om strømforbruk i Innlandet fylke i 2022 og januar 2023. I tillegg er det samlet inn data med ulike faktorer som kan påvirke forbruket, som pris- og ulike værforhold. Vi skal altså evaluere maskinlæringsmodeller basert på 13 måneder med historiske data om strømforbruket ved hjelp av ulike forklaringsvariabler. Vi vil her gå nærmere inn på forklaring av valgene for de ulike variablene og valgene som er tatt i forbindelse med dataforberedelser.

3.2.1. Valg av variabler

For å analysere sammenhenger og prediksjonsnøyaktighet for forskjellige maskinlæringsmodeller, benyttes historiske data for strømforbruk, -priser og værfaktorer som temperatur, nedbør og solskinnstimer. Data for strømforbruk er delt med oss fra strømselskapet, og består av strømforbruket fra dag til dag for selskapets kunder på Innlandet. Utvalget består av forbruket for kunder som av aktive gjennom hele 2022, som også var aktive kunder hos leverandøren ved datoen for uttrekket 26.10.2023. Prisdata er som nevnt hentet fra Forbrukerrådets hjemmesider, og er levert av Nord Pool (Forbrukerrådet, u.å.). Dataene består av spotpriser for hver time for prisområde NO1, som gjelder hele Østlandet. Det er også hentet data for tre ulike værforhold fra nettsidene til Norsk Klimasenter (Norsk Klimaservicesenter, u.å.). Værdataen gjelder alle værstasjonene som ligger i kommunene Hamar, Ringsaker, Løten og Stange, som også kalles for Hamarregionen.

Strømforbruk

Data for strømforbruk ble hentet fra et strømselskap og dermed delt med oss. Med historiske tall på daglig forbruk av strøm i private husholdninger, har man muligheten til å studere hvordan strømforbruket varierer gjennom året i tillegg til å kunne undersøke og skaffe innsikt om forbruksmønstre på ukes-, års-, og sesongbasis. Denne typen mønstre i data lar modeller identifisere periodiske trender, slik at den kan justere prognosene i henhold til tidligere atferd. Dataene delt med oss fra strømselskapet bestod av 14 ulike filer, hvorav 13 inneholdt daglig forbruk på hver eneste kunde i Innlandet fra januar 2022 til og med januar 2023. Uttrekket inkluderer kun privatkunder i Innlandet fylke. Hver av disse 13 filene inneholder tre variabler hver; målepunkt-ID (MPID), dato (Dato) og sum

av forbruket for gjeldende dato (sum.Volume_KWH). Hver fil innehold mellom 76.000-77.000 unike målepunkt-IDer i hver, som til slutt ble slått sammen til en samlet fil. Den siste av de 14 filene vi mottok inneholdt informasjon om hvert av målepunktene på et likevel anonymt nivå, som for eksempel det årlige forbruket og geografisk tilhørighet. Denne filen ble brukt til å få skilt ut de kundene som hører til i det geografiske området som tilhører Hamarregionen. Datagrunnlaget vi benytter oss av i maskinlæringsprosessen består av det daglige gjennomsnittsforbruket for ett målepunkt lokalisert i Hamar, Ringsaker, Løten eller Stange.

Strømpriser

Strømprisens påvirkning på forbruket ble tidligere i kapittel 2 diskutert, da det ikke er mye tilgjengelig teori på prisens direkte påvirkning på forbruket. Det er heller mer som tilsier at prisen påvirkes av forbruket. Likevel skriver Energifakta Norge (2024a) at prisen, sammen med vær, kan være med på å forklare de årlige variasjonene i strømforbruket i Norge. Basert på dette vil prisen vurderes som en uavhengig variabel for kortsiktige prediksjoner av strømforbruket. Høye priser kan på sin side føre til at forbrukerne reduserer forbruket, mens det imidlertid også heller kan være det høye forbruket som fører til høyere priser. Tross dette kan en integrering av strømpriser i modellen bidra til å forutsi endringer i forbruket basert på endringene i prisen. Modellene vil trenes på historiske data, og vil dermed også trenes på tidligere sammenhenger mellom prisendringer og forbruk som igjen kan påvirke forbruket av strøm. Data hentet fra Forbrukerrådets nettsider består av spotpriser på strøm som er levert av Nord Pool. Dette er data som ligger tilgjengelig for alle og blir levert i et strukturert og ryddig format som ikke trengs mye forberedende arbeid. Råfilen inneholder prisen på ulike datoer inkludert timeintervaller, og prisen er igjen delt inn i de fem ulike prisområdene vi har i landet.

Værdata

Som nevnt i kapittel 2 av både Energifakta Norge (2024b), og Kang og Reiner (2021), blir vær ansett som en faktor som kan påvirke strømforbruket, spesielt på årlig basis. Med dette tatt i betraktning, blir data om ulike værfenomener som temperatur, nedbør og solskinnstimer vurdert som uavhengige variabler i studien. I Norge er det kaldt om vinteren, og med kaldere vær kan en økning i strømforbruk forventes på grunn av behov for oppvarming, og varmere vær reduserer forbruket igjen fordi oppvarming ikke er like nødvendig. Værdata forventes å kunne bidra positivt for prediksjon, fordi modellene vil trenes på mønstre i forbruket basert på ulike klimatiske forhold og dermed kan tilpasse predikerte verdier deretter. Data er hentet fra Norsk Klimaservicesenter og inneholder informasjon om været i de ulike kommunene som utgjør Hamarregionen. For å få været fra hele regionen er data hentet

for de ulike værvariablene temperatur, nedbør og solskinn fra de 15 værstasjonene i området. Filene inneholdt dag-til-dag-data fra alle stasjonene med tall på middeltemperatur per døgn, total mengde nedbør og totalt antall solskinnstimer i løpet av det døgn. Middeltemperatur blir for klart av Norsk Klimasenter (u.å.) som et aritmetisk gjennomsnitt av 24-timersverdier, eller en formelbasert middelverdi dersom det er begrenset med observasjoner i løpet av et døgn. Fordi mange av stasjonene ikke målte alle tre værtyper, og de fleste målte alle, ble det beregnet en gjennomsnittsverdi av den gjeldende variabelen for gjeldende dag.

I studien vil dermed kortsiktig prediksjon av strømforbruk gjennomføres ved hjelp av historiske data for strømforbruk som avhengig variabel og strømpriser, temperatur, nedbør og solskinnstimer som uavhengige variabler. Hva som har inngått av dataforberedelser i forbindelse med de utvalgte variablene fra import til sammenslåing til et samlet datasett og videre forberedelser, som *feature engineering* og splitting av data til trening, vil gjennomgås i neste delkapittel.

3.3. Dataforberedelser

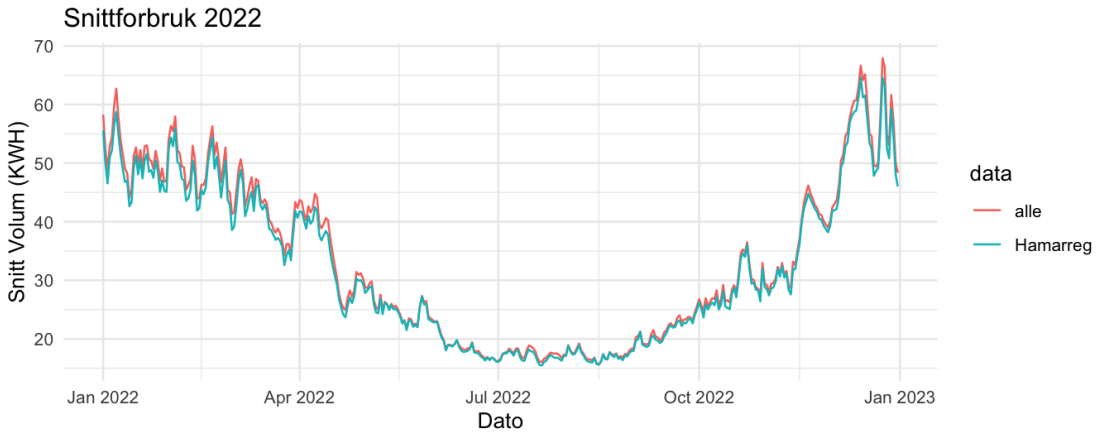
Formålet med dataforberedelser er å ta data som er identifisert i de to foregående stegene, forretningsforståelse og dataforståelse, og gjøre dataene klare for analysen som skal skje med en data mining-metode (Sharda et al., 2018). I denne studien er maskinlæring den valgte data mining-metoden. *Data Pre-processing*, eller dataforberedelser, omhandler teknikker for å legge til, fjerne eller transformere treningsdata (Kuhn & Johnson, 2013). Hvilke tiltak man har behov for å utøve under forberedelsesprosessen avhenger av datasettet og hvilken type modell man skal benytte seg av. Under denne kategorien av tiltak finner vi blant annet rensing av data, som fjerning av duplikater og behandling av nullverdier, men også dataintegrasjon, -reduksjon og utvalg av prediktordata.

For å gjøre innsamlet data klar for analyse ble det gjennomført en grundig utforskende analyse for å finne ut hvilke variabler som skulle tas med videre til det endelige datasettet før trening og analyse. Etter å ha bestemt hvilke variabler som skal benyttes, blir datasettet ryddet og forenklet. I arbeidet med blir programmet RStudio som verktøy for hele prosessen fra import til analyse. RStudio benytter R som programmeringsspråk, og er et system for statistisk databehandling som tilbyr en rekke forskjellige statistiske og grafiske teknikker (r-project, u.å.). RStudio gir muligheten til å utforske ulike mengder med data, og gjør det enkelt å jobbe med ulike former for maskinlæringsmodeller gjennom ulike pakker systemet tilbyr. Vi vil i dette delkapittelet presentere de ulike trinnene vi gikk gjennom for å forberede dataene i hver av de tre datakildene.

Strømforbruk

Som nevnt tidligere ble data med strømforbruk per kunde mottatt i flere filer. Disse filene ble slått sammen til et samlet datasett, bestående av over 28 millioner observasjoner med tre ulike variabler; «MPID», «Dato» og «sum.Volume_KWH». Dette settet med forbruksdata ble deretter kombinert med datasettet med kundeinformasjon slik at kundene kunne grupperes på en enkel måte. Deretter ble det gjennomført følgende trinn for å forberede datasettet for analysen.

For å først rydde i datasettet ble først alle kunder med 0 i forbruk fjernet, samt alle kunder kodet med «Industrial_Classification_Code» = «68.201». Denne koden er en næringskode, og ifølge Statistisk sentralbyrå, er det en kode for borettslag (Statistisk Sentralbyrå, u.å.b). Alle rader med denne koden ble fjernet, da hele borettslag kan være vanskelig å inkludere i analysene våre på en god måte uten å gi informasjon som kan tolkes feil. Da stod det kun igjen to næringskoder, XX og XY. kunder med næringskode XX er vanlige husholdninger og kunder med næringskode XY er hytter og fritidseiendommer. Neste steget innebar å fjerne kolonner som ikke var nødvendige for videre prosess, som «Brand_Code», «Price_Area», «City_Name», «Municipality_Code» og «Industrial_Classification Code». Deretter ble kundene som er lokalisert i de utvalgte kommunene, Hamar, Ringsaker, Løten og Stange, gruppert og resten ekskludert slik at det skulle være mulig å supplere med værdata senere. Dette gjorde at utvalget bestod av en kundegruppe på 19.126 unike verdier av målepunkt-IDer. Deretter ble gjennomsnittsforbruket per dag beregnet for alle kunder i dette området. Etter å ha gjort dette, utførte ble det utført en kontrollsjekk for å undersøke om utvalget var representativt for resten av populasjonen, ved å legge gjennomsnittsforbruket for hele regionen og for Hamarregionen i et linjediagram. Figur 3.1 viser beregnet gjennomsnittsforbruk for hele populasjonen som i vårt tilfelle tilsvarer alle kundene i Innlandet mot gjennomsnittsforbruket i utvalget. Figuren viser at det ikke er vesentlig stor forskjell på snittforbruket i hele fylket sammenlignet med Hamarregionen isolert sett, og utvalget ble derfor ansett som representativt for resten av populasjonen.



Figur 3.1 – Snittforbruk 2022 – Sammenligning Hamarregionen og Innlandet

Prisdata

Datagrunnlaget for prisdata er hentet fra Forbrukerrådets nettsider, som igjen henter data fra Nord Pool. Data inneholder spotprisen hver hele time hver dag gjennom hele perioden fra januar 2022 til og med januar 2023 for hver av de fem strømområdene i Norge. Det første som ble gjort med datasettet var å ekskludere de fire prisområdene som ikke var relevante, slik at datasettet kun bestod av prisdata for prisområde NO1 for Østlandet. Deretter ble det beregnet en medianpris basert på alle timesverdiene per døgn, slik at datasettet kun inneholdt medianprisen per dag. Medianprisen er i mange tilfeller å foretrekke over gjennomsnittet, da medianen ikke påvirkes av outliers.

Værdata

Data om værforholdene i Hamarregionen ble hentet fra Norsk Klimaservicesenter, og ble hentet fra de 15 ulike værstasjonene lokalisert i den utvalgte regionen. Filene som ble bestilt inneholdt daglige data på middeltemperatur, total nedbørsmengde og totalt antall timer solskinn i løpet av en dag fra alle stasjoner. En utfordring med denne dataen var at de 15 ulike stasjonene samlet inn forskjellige typer værdata. Alle stasjoner samlet temperatur, men kun et fåtall som samlet data på nedbør, og kun en som målte antallet solskinnstimer. På grunn av dette måtte det gjennomføres noen ekstra forberedende trinn som ikke bare innebar å kombinere alle datasettene til ett, men også erstatte tomme verdier og håndtere duplikater av datoer. For å ha enklere kontroll på hva som ble gjort, ble denne rensingen utført i Excel før datasettet ble overført til R. Verdier for både temperaturen og nedbørsmengden består av et beregnet gjennomsnitt av temperatur og nedbør på tvers av alle stasjonene som målte disse verdiene for hvert døgn. Solskinnstimene ble som nevnt kun målt ved en stasjon, dermed var ingen forberedelse nødvendig for disse verdiene. Selv om den utvalgte regionen er et område som dekker flere kommuner og et større areal, ble det gjort en forutsetning om at hele regionen opplevde akkurat samme snittemperatur, nedbørsmengde og solskinnstimer i løpet av et døgn. Denne forutsetningen var viktig dersom værdata skulle brukes i denne studien.

Tiltakene som ble gjennomført i forberedelsesprosessen hadde som hensikt å standardisere data fra de ulike kildene, slik at alle variablene kunne samles i det samme datasettet og brukes til prediksjonsformål. Noe som innebar å samle alle observasjoner per døgn i forbruksdataen til samlet gjennomsnittsforbruk, samle timesdata for pris til døgnverdier og samle værvariabler fra flere like kilder til å representere hele regionen. Gjennom denne prosessen blir man også litt bedre kjent med dataen, hvilke egenskaper den har og hvilke variabler som kan være nyttige for videre arbeid og hvilke som kan ekskluderes. Neste trinn vil innebære å fange opp mønstre i data ved hjelp av *feature engineering*.

3.3.1. Feature engineering

Feature engineering er en teknikk innen dataforberedelser som kan være nyttig for mange maskinlæringsmodeller, men kan også fungere dårlig for andre (Kuhn & Johnson, 2013). Steget innebærer ifølge Kuhn og Johnson (2013) å trekke ut nyttig informasjon fra en variabel, som for eksempel hente ut nummer for ukedag eller måned i året, eller lage en variabel for sesongnavn fra datokolonnen. Hvilke metoder som vurderes avhenger av analysen man skal utføre, og hvilke modeller som skal brukes.

Alle maskinlæringsmodeller klarer ikke å se etter sesongvariasjoner automatisk, for å gjøre disse variasjonene enklere for modellene å fange ble enkle *feature engineering*-steg gjennomført. Dette steget kan også gjøre det enklere å analysere datagrunnlaget ved senere behov, ved at trender i dataen som nødvendigvis ikke er synlige i det totale bildet også fanges opp. Følgende informasjon er hentet ut fra dato-kolonnen; månedsnummer, nummer for dag i uken og klassifisering etter sesonger for å skille ut sesongvariabler. Sesongvariablene tydeliggjør generelle sesongmønstre på tvers av hele datasettet og skal forsøke å hjelpe modellene med prediksjon. Det blir til dermed tre nye variabler som legges til i datagrunnlaget.

3.3.2. Splitte data

Før man er klar for å trene modellene er det viktig å bestemme seg for hvilken og hvor stor del av data som skal brukes for å evaluere resultatet fra modellen mot. Det er vanlig å dele data inn i trenings- og testsett, og det er også i mange tilfeller fordelaktig å ha et valideringssett i tillegg (Kuhn & Johnson, 2013). Testsettet legges til side for å sikre at modellen ikke trenes på denne dataen, og blir etter trening brukt for å teste hvor godt modellen presterer på nye data. Vanlig trenings-/testsplitt pleier å være 70/30 eller 80/20. Fordi vi bare har 13 måneder med data, er datasettet delt etter oktober 2022. Det vil si at treningssettet består av data fra januar til oktober 2022, og testsett bestående av november og desember fra 2022 og januar 2023. Treningssettet består av 304 rader og testsettet av 92, som gir en splitt på ca. 77/23.

I tillegg til en trenings-/testsplitt kan man også benytte seg av ulike «resampling»-metoder som involverer å ta mindre, tilfeldige utvalg av treningsdataen for å estimere modellens ytelse på disse, og er en prosess som gjentas over flere iterasjoner før resultatet summeres (Kuhn & Johnson, 2013). Kryssvalidering er en «resampling»-metode, og fungerer ved å gjentatte ganger ta uttrekk fra treningsdata og dermed tilpasse en modell på hvert uttrekk for å skaffe den informasjonen man trenger for å lage den endelige modellen (James et al., 2021). Dette uttrekket vil fungere som et lite

valideringssett. Formålet er å undersøke hvordan resultatene i en statistisk analyse skal generaliseres på et uavhengig datasett, og er spesielt egnet for prediksjonsformål og estimering av modellnøyaktighet. Under kryssvalideringsmetoder finnes blant annet en tilnærming ved navn *k-Fold Cross-Validation*, som altså vil inkluderes i vår studie. k-Fold-kryssvalidering innebærer tilfeldige oppdelinger av settet med observasjoner i k «folds», eller grupper, av samme størrelse (James et al., 2021). Den første gruppen blir brukt som et valideringssett og modellen blir tilpasset de resterende $k - 1$ grupper, deretter blir det beregnet en MSE (*mean squared error*) før prosessen gjentas k antall ganger med tilfeldige uttrekk på hver runde (James et al., 2021). En vanlig form for k-Fold-kryssvalidering er «10 k-Fold»-kryssvalidering, som vil si at prosessen gjentas 10 ganger. Dette skal blant annet bidra til å redusere risikoen for overtilpasning.

I studien vil det dermed benyttes en splitt på 77/23, som vil si at 77 % av dataene vil brukes som treningsdata og resterende 23 % brukes for modellevaluering og estimering av prediksjonsnøyaktighet til slutt. 77/23-split vil kombineres med en tilnærming for kryssvalidering som skal ved hjelp av repeterende uttrekk og tilpasning på mindre deler av settet, sørge for trente modeller som kan bidra i prediksjonsformål uten for mye overtilpasning.

3.3.3. Endelig datasett og forklaring av variabler

For det endelige datasettet ble forbruksdataen med variablene «Dato» og «snitt.Volume_KWH» slått sammen med datasettene for pris og vær. Sammenslåingen ble gjort på dato, slik at alle observasjonene som tilhørte en dato lå på samme dato i det endelige datasettet. Det endelige datasettet inneholder ni variabler; «Dato», «snitt_VolumeKWH», «spot_snitt», «temp», «nedbør», «solskinn», «mnd», «ukedag» og «sesong», og alle har 396 observasjoner hver. Et uttrekk av de endelige variabelnavnene med definisjon av datatype i det endelige datasettet er presentert i tabell 3.1. Tabellen viser at Dato-kolonnen er definert med datoformat, og «snitt_VolumeKWH», «spot_snitt», «temp», «nedbør» og «solskinn» som numeriske verdier med desimaler («dbl» = double class). Sesongvariablene «mnd» og «ukedag» er definert som integers («int»), som vil si numeriske verdier uten noen desimaler, og «sesong» som tekst («chr»).

Dato <date>	**snitt_volume KWH** <dbl>	**spot_median** <dbl>	**temp** <dbl>	**nedbør** <dbl>	**solskinn** <dbl>	**mnd** <int>	**ukedag** <int>	**sesong** <chr>

Tabell 3.1 – Uttrekk av kolonnenavn for endelig datasett

Forklaring av variabler

Tallene for det historiske strømforbruket i det endelige datasettet ligger under «snitt_volumeKWH» og består av et beregnet gjennomsnittsforbruk for alle de 19.126 unike målepunktene i kommunene som hører til under Hamarregionen. Disse kundene har alle vært kunder gjennom hele 2022 til og med uttrekksdato 26.10.2023, og består av kun vanlige husholdninger, hytter og fritidsboliger. Observasjonene er oppgitt i KWH.

Variabelen «spot_median» inneholder medianen av alle spotprisene registrert time for time per døgn. Vi har valgt å benytte oss av median her fremfor gjennomsnitt fordi strømprisen til tider kan ha store svingninger i løpet av en dag, og ved å bruke median i stedet for gjennomsnittet blir sjansen for at beløpet blir påvirket av eventuelle outliers redusert.

Variablene «temp», «nedbør» og «solskinn» er data hentet fra Norsk Klimaservicesenter, og består av det innsamlet data fra 15 ulike stasjoner. Temperaturvariabelen finner vi under «temp», og består av gjennomsnittet av middeltemperaturen per døgn. «nedbør» er et beregnet gjennomsnitt av mengden nedbør registrert hos de stasjonene som målte dette, og er målt i mm. Solskinnstimer, som vi finner under «solskinn» ble målt kun på én stasjon, og er derfor ikke beregnet noe gjennomsnitt av.

Sesongvariablene finner vi under «mnd», «ukedag» og «sesong». Variabelen «mnd» inneholder numeriske verdier av månedsnummer fra 1 til 12, hvor 1 = januar. «ukedag»-variabelen går fra 1 til 7 og består av numeriske verdier av ukedagen og begynner med ukedag 1 = mandag. «sesong»-variabelen er basert på månedsnummer og grupperer månedene slik at de kan deles inn i de forskjellige sesongene i løpet av året. Dataene er inndelt slik at månedene 3,4 og 5 hører til «vår», 6,7 og 8 til «sommer», 9,10 og 11 til «høst», og 12, 1 og 2 til «vinter».

3.4. Modellbygging

I steget for modellbygging blir ulike maskinlæringsteknikker valgt og trent med det forberedte datasettet, for å kunne møte det spesifikke forretningsbehovet (Sharda et al., 2018). Denne delen vil forklare hvordan vi har gått frem for å bygge og trene de utvalgte modellene. For bygging og trening av maskinlæringsmodellene benytter vi oss av programmeringsspråket R i systemet RStudio, og en pakke som kalles *caret* (*Classification And REgression Training*). Pakken inneholder funksjoner for å forenkle treningsprosesser for komplekse regresjons- og klassifiseringsproblemer, og bruker en treningsfunksjon som trener prediktive modeller og tilbyr forskjellige metoder for å trene og evaluere modeller på en standardisert måte. Treningsfunksjonen i *caret*-pakken forenkler modellbyggingsprosessen for oss ved

å blant annet utfører *pre-processing*-steg, som for eksempel normalisering og sentrering, dersom det er nødvendig og finner de optimale parameterverdiene for modellene automatisk. *Caret* benytter seg av flere ulike typer R-pakker, men laster kun ned de pakkene som behøves når det er nødvendig (CRAN, u.d.). Denne pakken gir oss mulighet til å bygge og trene de utvalgte modellene med samme fremgangsmåte.

For å trene og teste modellene ble datasettet delt opp med en splitt på ca. 77/23. Funksjonaliteten i *caret* vil også automatisk gjøre en kryssvalidering basert på kontrollparameter som ble definert på forhånd. Kontrollparameteren «*control*» i figur 3.1 settes opp med funksjonen «*trainControl*» som lar oss definere hvilken modellvalidering vi vil bruke, som i vårt tilfelle er en kryssvalidering med 10 k-fold. Som vil si at dataene deles inn i 10 deler, hvor ni av disse brukes for å trene modellen og den siste brukes for å validere modellen. Prosessen gjentas 10 ganger, og resultatene av modellens RMSE fra hver «fold» kan regnes gjennomsnittet av for å få et sterkt estimat av modellens ytelse.

```
control <- trainControl(method = "cv", number = 10)
```

Figur 3.1 – Kode: Kontrollparameter – 10k fold kryssvalidering

Over- og undertilpasning

Ved trening av maskinlæringsmodeller er det viktig å vurdere risikoen for over- og undertilpasning av data i modellen. Overtilpasning vil si at modellen ikke bare har lært seg mønsteret i datapunktene, men også karakteristikkene i hver variabel. Ved overtilpasning har vi dermed en overkomplisert modell med høy varians, som da vil si at undertilpasning oppstår når modellen er overforenklet og har høy grad av bias (Ramasubramanian & Moolayil, 2019).

3.4.1. Lineær Regresjon

Modellen *lm_model* bruker en treningsfunksjon hvor snitt_volumeKWH oppgis som avhengig variabel, og øvrige variabler som uavhengige. Dataen som brukes i treningen er *train_set* som består av data fra januar til og med oktober og utgjør 77 % av dataene. Metoden som oppgis er «lm», som vil si at modellen trenes som en Lineær Regresjonsmodell. Til slutt oppgis kontrollparameteren vist i figur 3.1, slik at treningen gjennomgår en 10k fold kryssvalidering.

```
lm_model <- train(snitt_volumeKWH ~ ., data = train_set, method = "lm", trControl=control)
```

Figur 3.2 – Kode: Lineær Regresjon

3.4.2. Random Forest

Modellen *rf_model* er en random forest-modell, som stort sett er bygget på samme måte som den lineære regresjonsmodellen med snitt_volumeKWH som avhengig variabel. I denne modellen har vi latt seasons være med i vurderingen. Vi har også måttet legge på «*tuneLength*=5» for å si at funksjonen skal prøve fem ulike verdier av «*mtry*», altså antallet variabler som skal utgjøre det tilfeldige uttrekket av kandidater i hver splitt. I vårt tilfelle ble den endelige verdien 7 (vedlegg 3.2).

```
rf_model <- train(snitt_volumeKWH ~ ., data=train_set, method="rf", trControl=control, tuneLength=5)
```

Figur 3.3 – Kode: Random forest

3.4.3 Gradient Boosting

For å bygge en modell basert på Gradient Boosting, måtte noe ekstra forberedelser gjøres først. En Gradient Boosting-modell kan kun behandle numeriske data, så det første vi måtte gjøre var å fjerne Dato-kolonnen fra treningssettet. Videre måtte vi sette opp en «*grid*», gbmGrid, som har som hensikt å spesifisere ulike kombinasjoner av hyperparameter som skal testes ut under treningen av modellen. Her har vi oppgitt antallet trær (*n.trees*) = (100, 150 og 200), dybden av interaksjoner (*n.interaction.depth*) = (1, 3, 5), en *shrinkage*-rate = (0.01, 0.1) og et minimum antall observasjoner i hver node (n.minobsinnode) = 10. Denne funksjonen har som formål å begrense modellen innenfor en gitt ramme. gbm_model er modellen vår, og er bygget på samme måte som foregående modeller, men med «*tuneGrid*» = *gbmGrid* og «*verbose*» = FALSE – som betyr at modellen gir minimal output. Av output fra gbm_model (vedlegg 3.3) ser vi at modellen har valgt minimum antall observasjoner i hver node lik 10, antall trær lik 200, interaksjonsdybde lik 3 og 0.1 som *shrinkage*-rate.

```
train_set_num <- train_set
train_set_num$Dato <- NULL
gbmGrid <- expand.grid(.n.trees = c(100, 150, 200), .interaction.depth = c(1, 3, 5), .shrinkage = c(0.01, 0.1),
        .n.minobsinnode = 10)
gbm_model <- train(snitt_volumeKWH ~ ., data = train_set_num, method = "gbm", trControl = control,
        tuneGrid = gbmGrid, verbose = FALSE)
```

Figur 3.4 – Kode: Gradient Boosting

3.4.4. Support Vector Machine (SVM)

Modellen *svm_model* er en modell som trenes ved hjelp av metoden «*svmRadial*», som vil si at modellen trenes med en *radial basis function*-kernel. I tillegg til den lignende oppbyggingen av *train*()-funksjonen vi har brukt på de tidligere modellene, har vi her også lagt inn en *preProcess*()-funksjon, som skal sentrere og skalere for hver prediktor. Dette må være et steg fordi SVM-modellen er følsom for skalaen på inputvariablene. Resultatene av modellens ytelse (vedlegg 3.4) blir presentert som et spekter av C-verdier mellom 0.25 og 4.00, som er en parameter som kontrollerer avveiningen mellom å oppnå lav feil på treningssettet og å minimere kompleksiteten i modellen for å forhindre overtilpasning. Den optimale modellen blir funnet basert på laveste RSME, og er i vårt tilfelle har C = 4, og sigma = 0.0973938.

```
svm_model <- train(snitt_volumeKWH ~ ., data = train_set, method = "svmRadial", trControl = control,
                   preProcess = c("center", "scale" ), tuneLenght = 5)
```

Figur 3.5- Kode: SVM

3.4.5. LASSO regresjon (L1)

Modellen *LASSO_model* er en type regularisert lineær modell, som trenes på en funksjon «*glmnet*» som brukes for LASSO-regresjon og såkalte *elastic net*-modeller i R. I denne modellen må vi definere et rutenett av hyperparameter for tuning av modellen med *tuneGrid*(). Her oppgir vi *alpha* = 1, som indikerer LASSO-regresjon, og lambda, eller straffeparameter, mellom 0.001 og 0.01. Også her oppgir vi sentrering og skalering i *preProcess*-funksjonen. Den optimale verdien av lambda i modellen ble 0.034 (vedlegg 3.5), basert på laveste verdig av RSME.

```
LASSO_model <- train(snitt_volumeKWH ~ ., data=train_set, method="glmnet", trControl=control,
                     tuneGrid=expand.grid(alpha=1, lambda=seq(0.001, 0.1, length=10)), preProcess=c("center", "scale")
```

Figur 3.6 – Kode: LASSO Regresjon

3.4.6. Ridge (L2)

Ridge-modellen er, som LASSO-modellen, en regularisert lineær modell Modellen trenes på nøyaktig samme måte som LASSO-modellen med samme *tuneGrid*, lambda=seq(0.001, 0.1, length=10) og de samme *pre-Processing*-trinnene. Eneste endringen i koden er *alpha* = 0 som indikerer Ridge-regularisering. De endelige verdiene for *alpha* og lambda i modellen ble 0 og 0.1 (vedlegg 3.6).

```
Ridge_model <- train(snitt_volumeKWH ~ ., data=train_set, method="glmnet", trControl=control,

        tuneGrid=expand.grid(alpha=0, lambda=seq(0.001, 0.1, length=10)), preProcess=c("center", "scale")
```

Figur 3.7 – Kode: Ridge Regresjon

Som vist i figurene 3.2 til 3.7 blir alle modellene trent med den samme trenings-funksjonen, *train()*, som er en funksjon i *caret*-pakken. Modelltilnærmingen ble definert med «*method*», som er den delen av koden som sier hvilken modell som skal benyttes i treningen. I tillegg er det blitt gjort nødvendige tilpasninger på koden til de ulike modellene for å gi de den enkleste formen for justeringer før trening. Samme funksjon er brukt for å gi alle modellene det samme utgangspunktet for prediksjonen og for å forenkle prosessen, da funksjonaliteten i *caret* blant annet sørger for automatisk utvalg av parametere og eventuelle nødvendige forberedelser.

3.5. Testing og Evaluering

I dette steget blir de utviklede modellene benyttet, vurdert og evaluert for deres nøyaktighet og generalitet (Sharda et al., 2018). Man vurderer også kvaliteten på forskningsdesignet underveis gjennom denne prosessen. Reliabilitet og validitet står sentralt når man skal bedømme kvaliteten på forskningen i kvantitative studier (Saunders et al., 2023). Både reliabilitet og validitet er nært forbundet med operasjonalisering, på den måten at en vellykket operasjonalisering har høy grad av både reliabilitet og validitet (Thrane, 2018). Aller først blir derfor funksjonen vi har benyttet for å gjennomføre prediksjonene presentert, før vi deretter vil diskutere begrepene reliabilitet, validitet og generalitet.

3.5.1. Prediksjon

Prediksjonsfunksjonen vi har benyttet i arbeidet med modellene heter *predict()* og er en funksjon for å gjøre prediksjoner basert på ulike modeller. Funksjonen bruker en trent modell og et datasett med nye data, altså testsettet. Dette betyr at den muliggjør bruk av den trente modellen på nye, usette data som er sentralt for å forstå generaliserbarheten. I praksis gjør funksjonen at man kan generere og vurdere prediksjoner fortløpende for de ulike modellene. Avhengig av konteksten returneres en vektor, matrise eller liste med predikerte verdier basert på den valgte modellen og datasett. Som man kan se av figur 3.8, brukes de utviklede modellene for hver av metodene og *newdata* = test_set.

```
lm_pred <- predict(lm_model, newdata = test_set)

rf_pred <- predict(rf_model, newdata = test_set)

gbm_pred <- predict(gbm_model, newdata = test_set)

svm_pred <- predict(svm_model, newdata = test_set)

LASSO _pred <- predict(LASSO _model, newdata = test_set)

Ridge_pred <- predict(Ridge_model, newdata = test_set)
```

Figur 3.8 - Kode: Prediksjon med ulike modeller

3.5.2. Reliabilitet

Reliabilitet, eller pålitelighet, handler om hvordan vår studie evner å produsere stabile og konsistente resultater under samme forhold, og forteller noe om i hvor stor grad en datainnsamlingsprosedyre gir konsistente funn (Saunders et al., 2023). Thrane (2018) skriver at reliabilitet handler om presisjonen på våre operasjonaliseringer, altså våre variabler, og om målefeilene eller unøyaktighetene i disse. Reliabilitet er på denne måten empirisk testbar gjennom statistikkprogram (Thrane, 2018). I denne oppgaven vurderes tre ulike nøyaktighetsmålinger i kombinasjon med hverandre for å gi en indikasjon på reliabiliteten til våre resultater. Prediksjonsnøyaktigheten blir presentert som verdier av RMSE, MAPE og R-kvadrert i analysekapittelet senere i oppgaven. Reliabiliteten kan ifølge Johannessen et al. (2016) testes på ulike måter, blant annet ved å gjenta den samme forskningsprosessen på samme gruppe på ulike tidspunkter. Dersom resultatene blir de samme, er dette et tegn på høy test-retest-reliabilitet. Med tanke på at datagrunnlaget som benyttes i forskningen består av ferdig innsamlede data som ikke endrer seg over tid, vil det ikke være store endringer i resultat dersom man gjør samme undersøkelsen på et annet tidspunkt. Derav kan vi si at reliabiliteten på dataen er høy. På en annen side vil resultatet derimot risikere å bli noe helt annet om datagrunnlaget endres.

En annen metode for å teste reliabilitet, er ved å se på internvaliditet. Internvaliditet går ut på at flere forskere undersøker det samme fenomenet, og dersom alle kommer frem til det samme resultatet kan man si at studien har høy grad av internreliabilitet (Johannessen et al., 2016). Hvorvidt internreliabiliteten er høy eller lav i vår studie avhenger av mange faktorer, som blant annet fremgangsmåte, datainnsamling og -forberedelser, samt modellbygging og parametertuning. Maskinlæringsprosessen og arbeidet med data og prognosemodeller kan utøves svært individuelt og mulighetene for ulike former for dataforberedelser og tuning av parametere er mange. Det er med andre ord en generell svakhet knyttet til resultatene fra en maskinlæringsprosess, fordi gode resultater fra en modell på et gitt testsett ikke nødvendigvis garanterer gode resultater på fremtidige tester.

Resultatene kan kun gi oss en viss indikasjon på hvordan modellene eventuelt kan prestere på helt nye data.

3.5.3. Validitet

Validitet forteller noe om i hvilken grad våre prosedyrer måler nøyaktig hva de er ment å skulle måle (Saunders et al., 2023). Dette er viktig i forskning, da det skal sikre at konklusjonene som trekkes er meningsfulle. Man skiller på intern og ekstern validitet, hvor intern validitet handler om at det er de uavhengige variablene som forårsaker endringene i den avhengige variabelen. Ekstern validitet sier noe om hvor godt en studie kan overføres til andre situasjoner, tider og lignende, altså studiens evne til å gi resultater som er relevante utenfor gitt kontekst. Det kan være en mengde faktorer som spiller inn på en husholdnings strømforbruk, som ikke er tatt hensyn til i denne studien. Resultatene i studien begrenses til de utvalgte forklaringsvariablene som hensyntas og det kan derfor være ukjente årsaker som påvirker våre resultater og konklusjoner, uten at vi har tatt høyde for det. Begrepsvaliditet blir presentert i Johannessen et al. (2016) som et typisk målingsfenomen, som skal si noe om relasjonen mellom det som undersøkes og de konkrete dataene. Det vil si at begrepsvaliditet går ut på om hvorvidt dataene representerer det generelle fenomenet som forskes på.

I studien mottok vi data på forbruk for alle private kunder hos et strømselskap i Innlandet fylke, hvor vi utførte et ytterligere utvalg av populasjonen for å kun fokusere på en mindre region i Fylket bestående av fire kommuner i Hamarregionen. Dette utvalget ble testet mot hele populasjonen ved å sammenligne gjennomsnittsforbruket per kunde i Hamarregionen mot gjennomsnittsforbruket per kunde i hele fylket. Denne testen viste et tilnærmet likt forbruk gjennom hele 2022 for gjennomsnittskunden i Hamarregionen som gjennomsnittskunden i hele Innlandet, og vi vil derfor si at utvalget vi gjorde var representativt for resten av populasjonen. På en annen side er den totale populasjonen vi tok utvalget fra, kun et mindre utvalg av kunder i hele landet, som igjen kun er et mindre utvalg av alle husholdninger som bruker strøm i hele landet som er kunder hos andre leverandører. Det er derfor vanskelig å si noe om forbruksmønsteret vi har i våre data er representativt for alle husstander i Norge. Vi kan derfor si at vi har valide data basert på den tilgjengelige populasjonen vi hadde data på, men til hvilken grad utvalget er representativt for hele landet er vanskelig å si noe konkret om. Vi kan kun gjøre antakelser om at selve forbruksmønsteret hos den gjennomsnittlige kunde utenfor Innlandet, ikke avviker kraftig fra forbruket til den gjennomsnittlige kunden på Innlandet.

3.5.4. Generalitet

Generaliseringsevnen i en studie refererer til i hvilken grad funnene våre kan generaliseres og vurderes som gyldige for andre grupper eller situasjoner enn det som ble studert. Vi har gjort studien i Hamarområdet med data fra hele 2022 og januar 2023. Dette vil si at vi har gjennomført studien i et begrenset geografisk område, og med historisk data begrenset til kun ca. ett foregående år. For å kunne dra slutninger om for eksempel hele Norge måtte vi gjort de samme analysene og prosedyrene i flere kommuner, og over flere år enn kun 2023. Vi tror likevel at vår fremgangsmåte og metode kan benyttes i tilsvarende studier, på spesifikke geografiske områder med begrenset tidsseriedata. Vi har valgt å predikere på ulike prognoseperioder for å teste generalisering, robusthet og stabilitet over tid. Modellene må uansett testes med utvidede datasett, for å se om de er generaliserbare og om de kan fungere for andre steder og på andre år enn det som ble testet med i studien.

3.6. Distribusjon

Det siste steget er ganske åpent og avhenger av konteksten for data mining-prosessen. Distribusjonen kan være så enkel som å generere en rapport, eller så kompleks som å implementere en repetitiv data mining-prosess på tvers av en hel bedrift (Sharda et al., 2018). I vår oppgave vil dette steget innebære generering resultater i form av grafiske visualiseringer og nøyaktighetsmålinger for de ulike maskinlæringsmodellene. Målet er å kunne dra slutninger basert på disse resultatene. Analysen og resultater vil nå gjennomgås i neste kapittel.

Gjennom kapittel 3 har forskningsmetoden for studien blitt beskrevet. Designet er planlagt, og datagrunnlag og trening av modeller har blitt gjort rede for. Studien vil gjennomføres som en kvantitativ studie med en evaluerende tilnærming. Problemstilling vil besvares ved hjelp av tre definerte forskningsspørsmål som blant annet går ut på å undersøke uavhengige variabler og deres relasjon med avhengig variabel, men også evaluere hvilke av de utvalgte modellene som best kan predikere strømforbruket med gitt datagrunnlag og like forutsetninger. Datagrunnlaget består av sekundærdata samlet inn fra ulike kilder, som gjennom en forberedelsesprosess er blitt standardisert slik at de kan brukes til samme formål. Modellene ble bygget på en tilnærmet lik måte ved hjelp av funksjonalitetene «caret» tilbyr, som gjør at sammenligningen av modellen vil være så enkel som mulig. For å avgjøre hvorvidt de uavhengige variablene skal inkluderes i det endelige prediksjonsgrunnlaget, vil det gjennomføres prediksjonstester i neste kapittel. Neste kapittel, som er kapittel 4, vil også sammenligne modellenes prediksjonsevne basert på de ulike nøyaktighetsmålingene som tidligere er presentert.

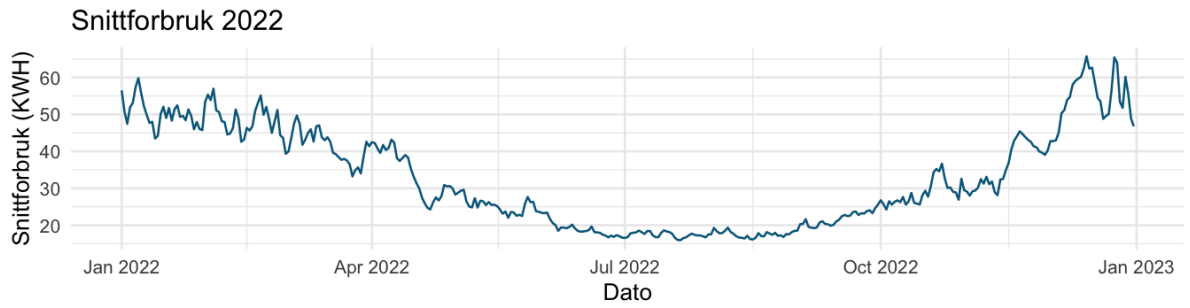
4. Analyse og resultater

I dette delkapittelet vil det utforskes hvordan maskinlæring kan benyttes til å predikere kortsiktig strømforbruk hos norske husholdninger, basert på resultatene fra modellene. Vi vil analysere hvordan ulike faktorer som pris og værforhold kan påvirke strømforbruket, og undersøke hvilken maskinlæringsmodell som best kan forutsi disse svingningene.

Den utforskende analysen av innsamlet data og de forskjellige variablene som benyttes i oppgaven vil presenteres gjennom dette kapittelet. I arbeidet med innsamlet data er det viktig å bruke tid på å bli godt kjent med dataen. Gjennom utforskende dataanalyse får man mulighet til å bli godt kjent med data og variabler, og kan oppdage mønstre og relasjoner i dataen man kanskje ikke hadde regnet med i utgangspunktet. Formålet med analysen er å bli kjent med de ulike uavhengige variablene og utforske deres sammenheng med den avhengige variabelen, før resultatene fra prediksjonen blir presentert og rangert. Å lage visualiseringer er en enkel måte å bli kjent med dataen på, derfor skal vi benytte visualiseringer for å bli bedre kjent med de ulike variablene alene, og de uavhengige sammenlignet med den avhengige. Vi vil se på de viktigste variablene i studien vår; strømforbruk, pris og de ulike værforholdene, før vi vil utforske relasjonen mellom den avhengige og de forskjellige uavhengige variablene. Alle analyser ble utført med R i RStudio.

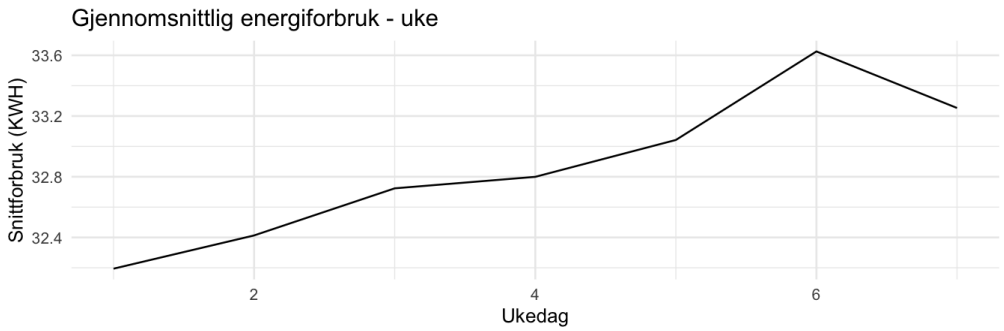
4.1. Strømforbruk

Figur 4.1 viser gjennomsnittsforbruket av strøm vist i antall kilowattimer (kWH) per dag hos en privatkunde i Hamarregionen gjennom hele 2022. Av figuren ser man en betydelig høyere grad av variasjon i snittforbruket og at forbruket generelt er høyere i vintermånedene, sammenlignet med sommermånedene. Forbruksvariabelen varierer mellom ca. 15 KWH i området juli-august til ca. 65 kWH på det høyeste i desember-januar.

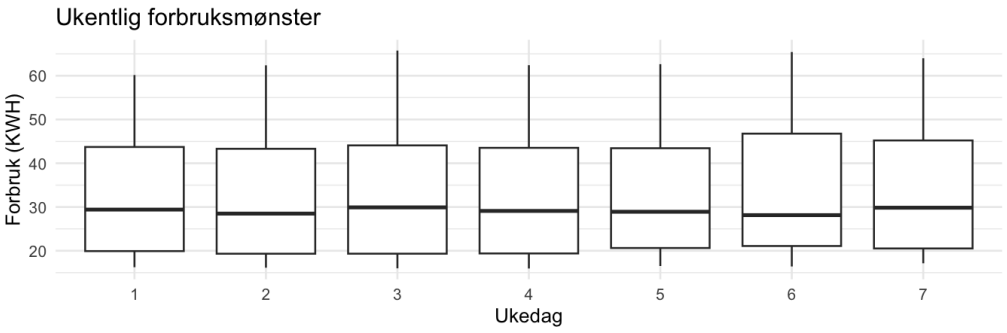


Figur 4.1 - Gjennomsnittlig strømforbruk 2022

Ved hjelp av sesongvariablene som ble lagt inn i dataforberedelsesprosessen, har vi mulighet til å se på variablenes utvikling i også andre former enn over hele perioden som i figur 4.1. Figur 4.2 og 4.3 viser energiforbruket for en gjennomsnittlig uke og det ukentlige forbruksmønsteret. Ved å studere forbruksmønsteret i både linjediagram og boksplott sammen kan man skaffe innsikt om variabelen på en hel annen måte enn med totalbildet. Av figur 4.2 kan man se at uken starter med et lavt forbruk, og at det deretter øker gjennom hele uken med en topp på dag nummer 6 (lørdag). Ser man derimot på figur 4.3 kan man av den horisontale linjen inni boksene se medianforbruket, som her ligger relativt jevnt gjennom hele uken. Strekene over og under boksene, også kalt *whiskers*, forteller noe om spredningen i data og eventuelle prikker utenfor disse regnes som outliers. En årsak til at disse illustrasjonene kanskje ikke viser den samme bevegelsen i 4.3 som i 4.2, kan være verdier av forbruket som er betydelig lave eller høye og som dermed påvirker gjennomsnittet som vises i 4.2.



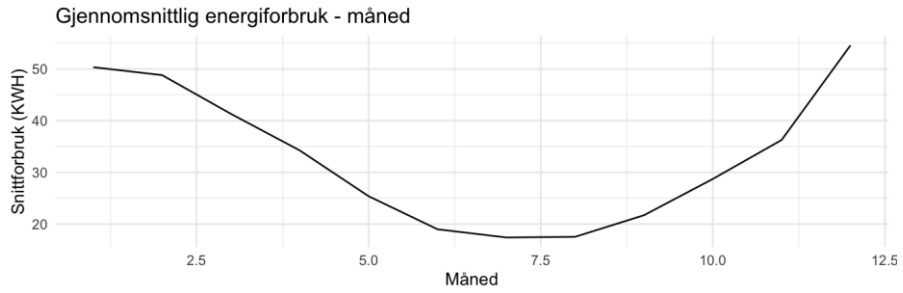
Figur 4.2 - Gjennomsnittlig energiforbruk – uke



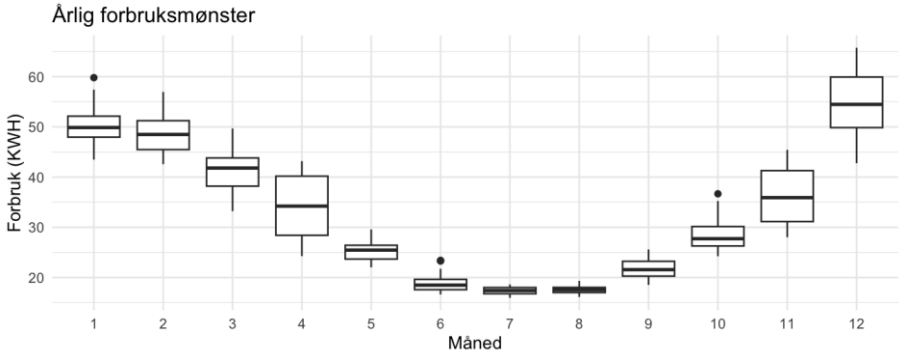
Figur 4.3 - Ukentlig forbruksmønster

I figur 4.4 og 4.5 kan man se en tydelig U-formet trend, hvor forbruket synker mot midten av året og øker igjen mot slutten av og i begynnelsen av året. Det vil si at det, ifølge våre data, brukes mindre strøm om sommeren enn om vinteren, som kan skyldes at det i de kaldere månedene er et økt behov for oppvarming. Boksplottet i figur 4.5 støtter trenden fra linjediagrammet, og viser i tillegg noen måneder med høy variabilitet i vintermånedene. Måned 1, 6 og 10 har også *outliers*, som indikerer

dager med uvanlig høyt forbruk sammenlignet med resten av dagene den gjeldende måneden. I tillegg har flere av vintermånedene lengre *whiskers* både over og under boksene, som vil si at det i disse månedene er store variasjoner av forbruk gjennom måneden.

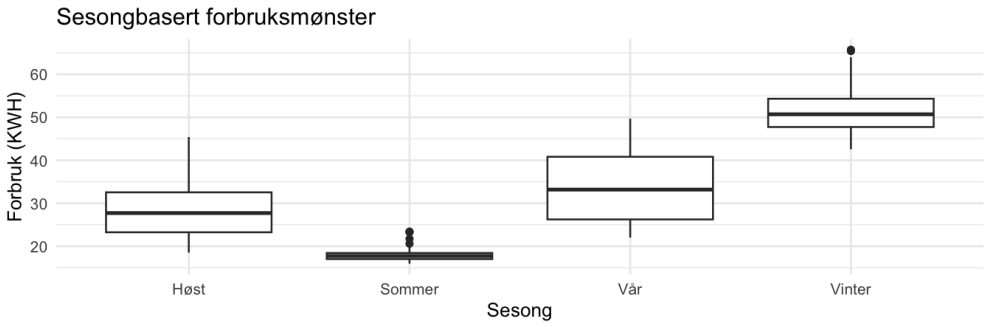


Figur 4.4 - Gjennomsnittlig energiforbruk – måned for måned



Figur 4.5 - Årlig forbruksmønster

Figur 4.6 viser forbruksmønsteret for sesongene høst, sommer, vår og vinter, og man kan se de samme bevegelsene her som i figur 4.4 og 4.5 fordi sesongvariabelen er basert på en kategorisering av månedsvariabelen. Av figur 4.6 ser man at det er flest outliers i sommersesongen, og noen i vintersesongen, som kan tyde på at det var noen dager med betydelig høyere energiforbruk enn andre i disse sesongene.



Figur 4.6 - Sesongbasert forbruksmønster

4.2. Uavhengige variabler

For prediksjonen ble daglige verdier for medianpris, middeltemperatur, nedbør og solskinnstimer brukt som uavhengige variabler. I denne delen vil de ulike avhengige variablene analyseres, vurderes og testes for prediksjon av strømforbruk. Prisen vil vurderes alene mens værvariablene blir analysert hver for seg før de testes for prediksjon som en gruppe. Formålet er å få frem årsaken til hvorfor disse variablene er tatt med som prediktorvariabler.

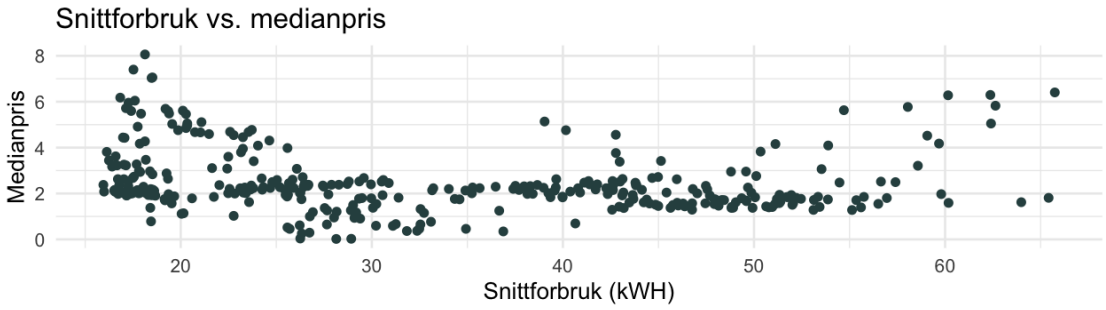
4.2.1. Pris

Strømprisens utvikling gjennom 2022 blir illustrert i figur 4.7 som et linjediagram av medianprisen sammenlignet med strømforbruket per dag gjennom hele året. Figuren viser hvordan medianprisen holder er mer stabilt nivå gjennom hele året med relativt små svingninger første halvår, og hvordan prisnivået øker og varierer mer utover siste halvår. Sammenlignet med forbruksvariabelen har prisvariabelen lite svingninger og ingen tydelig sesongvariasjon.

Figur 4.7 - Snittforbruk mot medianpris

Som nevnt i kapittel 2.2 sier ikke teorien mye om prisens direkte påvirkning på forbruket, men det vises blant annet til statistikk som viser et markant fall i strømforbruket i 2022 samtidig som prisnivået var unormalt høyt. Av figur 4.7 er det vanskelig å se et fall i forbruk sammenlignet med tidligere år fordi dataene ikke går lenger tilbake i tid, men ser tilsynelatende ikke ut som prisen og forbruket har særlig høy grad av korrelasjon. Korrelasjon kan enkelt visualiseres ved hjelp av et spredningsdiagram, og korrelasjonen mellom medianpris- og forbruksvariabelen blir illustrert i figur 4.8. Spredningsdiagrammet viser sammenhengen mellom de to variablene, hvor hvert punkt representerer en observasjon av et snittforbruk og den tilhørende medianprisen. Punktene i figur 4.8 viser ingen tydelig trend. Det kan med andre ord se ut som medianprisen varierer uavhengig av forbruket fordi man ikke kan se noen klar økning eller reduksjon i prisen basert på forbruket. For snittforbruk lavere enn 20 kWH ser det ut til å være et bredere spekter av priser, som kan tyde på høyere volatilitet i

prisene for lavere forbruksnivå, men utover dette er prisfordelingen mer jevnt fordelt uten åpenbar økning eller nedgang.



Figur 4.8 - Spredningsdiagram: Medianpris og forbruk (kWH)

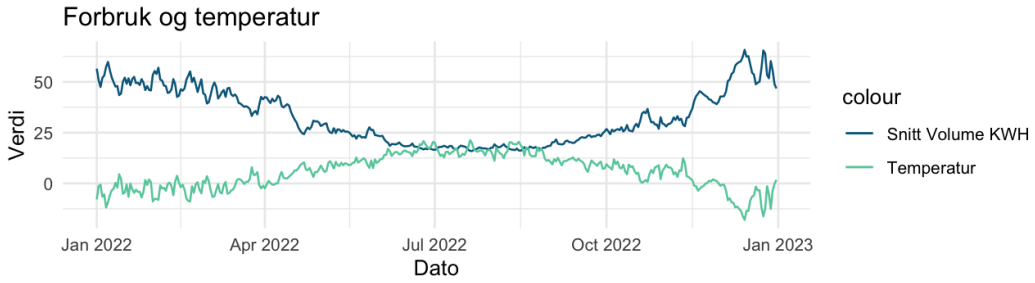
Basert på korrelasjonen mellom pris- og forbruksvariabelen, vist i figur 4.8, vil ikke prisvariabelen være den avgjørende prediktoren i prediksjonene som skal gjennomføres. I tillegg til å analysere spredningsdiagram, utføres en test av prediksjonsevnen for de utvalgte modellene med forbruk som avhengig variabel og kun medianprisen som uavhengig variabel. Resultatene fra testen vises i tabell 4.1 hvor vi ser modellenes prediksjonsevne målt i RMSE (*Root Mean Squared Error*), MAPE (*Mean Absolute Precentage Error*) og R-kvadrert, og er rangert fra lavest til høyest RMSE og lavest verdi tilsier beste resultat. RMSE måler hvor nærme de predikerte verdiene er de faktiske verdiene, hvor lavest verdi betyr best modellnøyaktighet. MAPE er en alternativ metode for å måle nøyaktigheten av en prediksjon ved å beregne feil som en prosent av de faktiske verdiene. Som i RMSE ser man også her etter lavest verdi for å finne best ytelse. R-kvadrert brukes som et mål på hvor godt fremtidige verdier vil bli predikert av modellen. Verdiene måles som en verdi mellom 0 og 1, hvor 1 betyr at modellen forklarer all variansen i responsvariabelen og dens middelverdi, og 0 betyr at modellen ikke har noe forklaringsevne. Her ønsker vi oss derfor høyeste, positive verdi. RMSE i tabell 4.1 varierer mellom ca. 15 og 33 med Gradient Boosting med laveste verdi og Lineær Regresjon som dårligst. MAPE ligger i dette tilfellet mellom 24 og 61. Tabellen viser negative verdier av R-kvadrert mellom -1 og -10, som vil si at, basert på R-kvadrert, vil samtlige modeller predikere fremtidig forbruk dårligere enn et naivt gjennomsnitt av historisk forbruk.

	RMSE	MAPE	R^2
GBM	15,52	24,22	-1,34
L2	20,64	36,99	-3,13
RF	21,12	35,38	-3,33
SVM	21,51	36,16	-3,49
L1	30,50	56,24	-8,03
LM	33,17	60,83	-9,68

Tabell 4.1 - Prediksjonsnøyaktighet med kun pris som uavhengig variabel

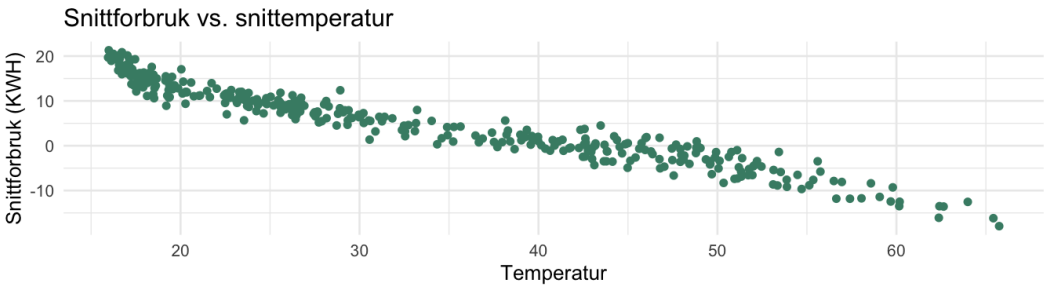
4.2.2. Værvariabler

De uavhengige variablene temperatur, nedbørsmengde og solskinnstimer på tvers av Hamarregionen vil bli illustrert med linjediagram med forbruksvariabelen i figur 4.8, 4.9 og 4.10. Figur 4.8 viser middeltemperaturens utvikling fra dag til dag gjennom hele året sammenlignet med forbruket. Temperaturen ser ut til å variere mellom ca. -25 på det kaldeste og ca. 25 på det varmeste. Når temperaturvariabelen sammenlignes med forbruket, kan man se en grad av korrelasjon ved å studere hvordan bevegelsene nesten speiler hverandre gjennom året. Når temperaturene faller, øker forbruket, og når temperaturen øker igjen, faller forbruket.



Figur 4.8 - Snittforbruk mot temperatur

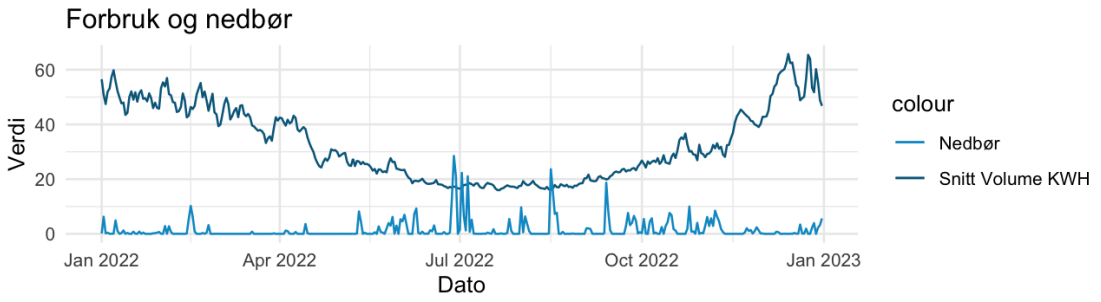
Korrelasjonen testes med et spredningsdiagram, og illustreres i figur 4.9. Sammenlignet med spredningsdiagrammet i figur 4.7 har figur 4.9 en betydelig tydeligere trend i datapunktene. Figuren viser en tydelig negativ korrelasjon mellom forbruket og temperaturen, som indikerer at dersom temperaturen stiger, synker strømforbruket. Sammenhengen kan skyldes økt behov for oppvarming i vintermånedene og omvendt.



Figur 4.9 - Spredningsdiagram: Forbruk og temperatur

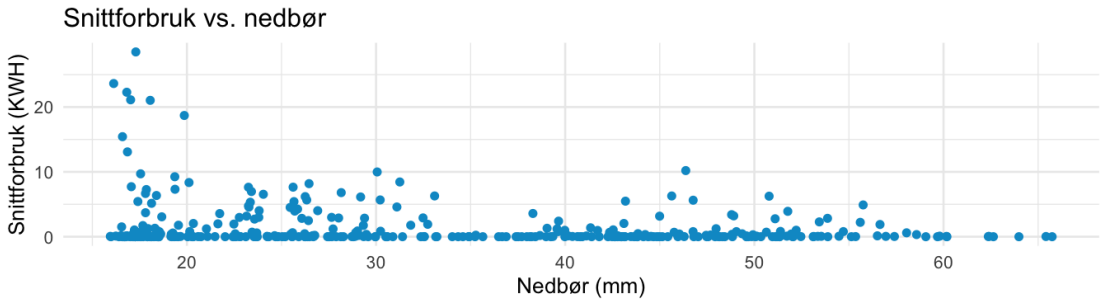
Nedbørsvariabelen blir presentert i figur 4.10 og er et beregnet gjennomsnitt av nedbør målt i mm fra alle værstasjonene i det utvalgte området. Figuren viser nedbør (mm) mot snittforbruket (kWH), og viser at nedbørsmengden hadde mest variasjon mellom juni og november med en topp i juli. Første

halvår ser relativt stabil ut. Bevegelsene i variabelen kan nesten ligne på temperaturvariabelen, med økning i nedbørsmengde og temperatur rundt samme periode. Sammenlignet med forbruksvariabelen kan det se ut som variablene har en liten grad av korrelasjon, selv om det kan være tilfeldigheter som gjør at forbruket er på det laveste når nedbørsmengden er på det høyeste.



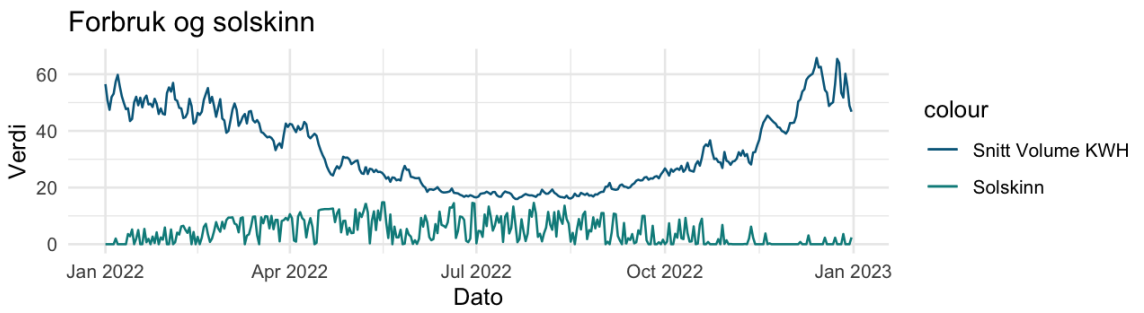
Figur 4.10 - Snittforbruk mot nedbør

Spredningsdiagrammet i figur 4.11 viser forholdet mellom forbruket og nedbør. Av figuren ser det ut til at de fleste punktene ligger langs null på x-aksen, og at det ikke er noen tydelig trend eller korrelasjon mellom variablene. Det kan med andre ord se ut som folk ikke endrer forbruket sitt noe betydelig av basert på nedbørsmengden alene.

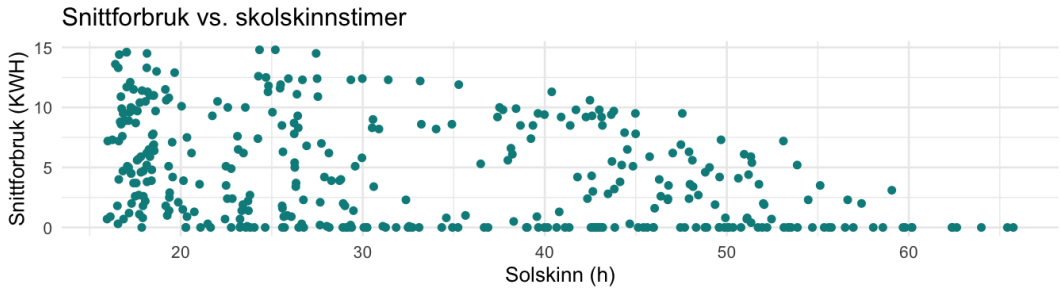


Figur 4.11 – Spredningsdiagram: Snittforbruk mot nedbør

Den siste værvariabelen består av antallet solskinnstimer registrert ved den ene stasjonen i regionen som målte dette, og er vist i figur 4.12. Av figuren ser man store kortsiktige variasjoner i antallet timer med solskinn per dag. Antallet timer solskinn ser ut til å være høyest i månedene mellom april og september og er relativt lavt mellom oktober og januar. Sammenlignet med forbruket har solskinnsvariabelen mindre variasjoner gjennom året, som kan tyde på at solskinn heller ikke nødvendigvis korrelerer direkte med forbruket. Mangel på korrelasjon og tydelig trend mellom forbruks- og solskinnsvariabelen bevises i spredningsdiagrammet i figur 4.13. Det er ikke noe tydelig mønster i datapunktene som indikerer sterk korrelasjon. Med andre ord, forbruket ser ut til å variere uavhengig av antall solskinnstimer.



Figur 4.12 - Snittforbruk mot solskinnsimer



Figur 4.13 - Spredningsdiagram: Snittforbruk mot solskinnstimer

Det er kun oppdaget sterk korrelasjon mellom temperaturen og forbruket, og ingen til lav korrelasjon mellom nedbørsmengde og solskinn mot forbruket. Basert på påstanden fra Energifakta Norge om hvordan værforhold er en av faktorene for de årlige variasjoner i energiforbruket, samt forskningen fra Kang og Reiner (2021) på effekten av temperatur, nedbør og solskinn som påvirkning på forbruk, vil alle de tre uavhengige variablene testet over bli inkludert i videre arbeid. Som med prisvariabelen blir det gjennomført en prediksjonstest av forbruket med værvariablene som en gruppe uavhengige variabler. Resultatene vises i tabell 4.2 og prediksjonsevnen illustreres på samme måte som i 4.1, i form av nøyaktighetsmålingene RMSE, MAPE og R-kvadrert. Tabellen og modellene er rangert etter laveste verdi av RMSE, og sammenlignet med resultatene fra 4.1 fører værvariablene til nærmest en halvering av RMSE- og MAPE-verdiene. RMSE varierer mellom LASSO-modellens (L1) ca. 8 til Ridge-modellens (L2) ca. 21, og MAPE mellom ca. 13 og 37. Tabellen viser fortsatt flertall av negative R-kvadrert-verdier, men LASSO og Gradient Boosting (GBM) er de eneste to modellene med positive verdier på henholdsvis 0,36 og 0,25. Dette er ikke nødvendigvis bra resultater, men indikerer et betydelig bedre potensiale enn med kun prisvariabelen.

	RMSE	MAPE	R^2
L1	8,12	16,16	0,36
GBM	8,77	12,80	0,25
LM	10,29	20,13	-0,03
RF	14,46	23,29	-1,03
SVM	16,32	25,70	-1,59
L2	20,64	36,99	-3,13

Tabell 4.2 - Prediksjonsnøyaktighet med kun værvariabler som uavhengige variabler

Til tross for lav korrelasjon og ikke tilfredsstillende resultater fra prediksjonstesten med kun prisvariabelen sammenlignet med testen med kun værvariabler, ble likevel prisvariabelen inkludert i en ny test av prediksjon med alle uavhengige variabler. Resultatet fra testen vises i tabell 4.3. Tabellen viser samme rekkefølgen av maskinlæringsmodeller i topp tre, med verdier av RMSE mellom 7,58 som laveste og 17,27 som høyeste. Alle verdier av RMSE og MAPE er forbedret sammenlignet med resultatene i tabell 4.2. Også R-kvadrert gir betydelig bedre resultater enn begge de tidligere testene som er vist i tabell 4.1 og 4.2, i tillegg til at dårligste modell har gått fra å være -3,13 på Ridge-modellen i tabell 4.2 til å være Support Vector Machine (SVM) med -1,89. Resultatet tyder på at selv om en test av en enkelt uavhengig variabel alene gir dårlige prediksjoner, kan det være fordelaktig å teste flere ulike uavhengige variabler før variabelen forkastes. Tabell 4.4 viser et sammendrag av alle resultatene fra de ulike testene med enkelte uavhengige variabler mot alle variabler inkludert.

	RMSE	MAPE	R^2
L1	7,584	14,598	0,442
GBM	7,816	11,584	0,407
LM	9,754	18,604	0,076
L2	10,371	19,024	-0,044
RF	13,567	21,527	-0,787
SVM	17,267	27,894	-1,895

Tabell 4.3 - Prediksjonsnøyaktighet med alle uavhengige variabler

	RMSE			MAPE			R¨2		
	Kun pris	Kun vær	Alle	Kun pris	Kun vær	Alle	Kun pris	Kun vær	Alle
LM	33,2	10,3	9,8	60,8	20,1	18,6	-9,7	-0,0	0,1
RF	21,1	14,5	13,6	35,4	23,3	21,5	-3,3	-1,0	-0,8
GBM	15,5	8,8	7,8	24,2	12,8	11,6	-1,3	0,3	0,4
SVM	21,5	16,3	17,3	36,2	25,7	27,9	-3,5	-1,6	-1,9
L1	30,5	8,1	7,6	56,2	16,2	14,6	-8,0	0,4	0,4
L2	20,6	20,6	10,4	37,0	37,0	19,0	-3,1	-3,1	-0,0

Tabell 4.4 - Sammenligning prediksjonstest uavhengige variabler

Gjennom denne delen av analysen er de ulike uavhengige variablene studert mot den avhengige variabelen og det er oppdaget varierende grad av korrelasjon uavhengige og avhengig variabel. Temperaturen viste seg å ha betydelig høyere grad av korrelasjon sammenlignet med de øvrige uavhengige variablene som ble studert, men på bakgrunn av resultatet fra tester og teorien i 2.2. blir alle de uavhengige variablene benyttet trening av modellene for prediksjon. Flere uavhengige variabler vil kanskje bidra til et mer realistisk bilde av de faktiske forhold og usikkerheten knyttet til prediksjon av strømforbruket, blant annet med tanke på lite forutsigbarhet i faktorer som værforhold. I tillegg er det viktig å ta i betraktning at strømforbruket styres av menneskelig atferd, som også vil si at forbruksmønsteret er så komplisert at det ikke kan forklares av en enkelt variabel. Hvordan disse ulike testene på prediksjonsevne med kun de utvalgte uavhengige variablene vil sammenlignes med prediksjonsevnen med alle uavhengige variabler i diskusjonen. Prediksjonsnøyaktigheten til de ulike modellene i de ulike tidshorisontene vil bli presentert i neste del.

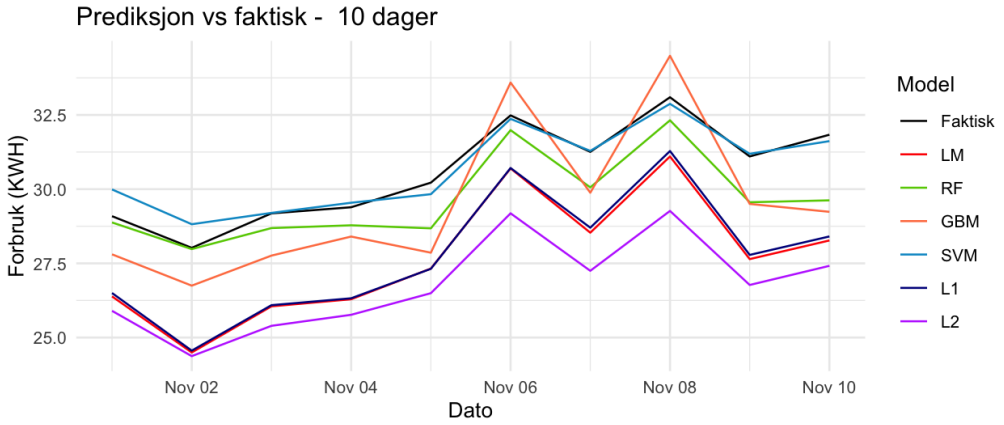
4.3. Modellprestasjon

De ulike modellenes evne til å predikere strømforbruket vil bli presentert og analysert i dette delkapittelet. Predikert forbruk vil sammenlignes mot faktisk forbruk basert på visualiseringer og nøyaktighetsmålinger. Prediksjonsevnen presenteres i tidsperioder på 10, 30, 60 og 90 dager frem i tid med formål om å undersøke modellenes robusthet og stabilitet over tid. Resultatene presenteres i form av linjediagram som har som hensikt å gi en enkel visuell indikasjon på prestasjonen til de ulike modellene. Deretter presenteres nøyaktighetsmålinger som skal bidra med informasjon om empiriske prediksjonsresultater. Nøyaktighetsmålingene vises som verdier av RSME, MAPE og R kvadrert.

4.3.1. Visualiseringer av resultater

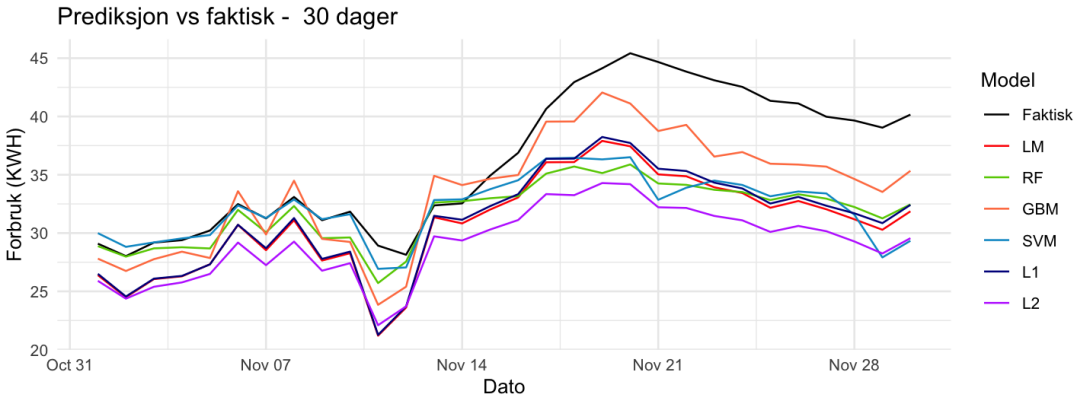
Visualisering av data bidrar med å gi praktisk innsikt i dataanalysen. I figur 4.14 til 4.17 blir prediksjonsresultatene presentert i form av linjediagram, hvor x-aksen er dato frem i tid og y-aksen viser forbruket oppgitt i kilowattimer. Det faktiske forbruket vises ved svart heltrukken linje. Denne linjen er det modellene skal prøve å treffe på gjennom prediksjonen. De øvrige linjene med farger er de ulike modellenes prestasjon hvor rød linje er Lineær Regresjon (LM), grønn er Random Forest(RF), oransje er Gradient Boosting (GBM), lys blå er Support Vector Machine (SVM), mørk blå er LASSO (L1) og lilla er Ridge (L2). På noen av tidshorisontene presterer to modeller svært likt, noe som gjør at linjene ligger svært tett. Først beskrives tidshorisonten på 10 dager i figur 4.14. Av figuren fremkommer det at alle modellene får med seg de to forbrukstoppene og nedgangene det faktiske forbruket har og linjene

ligger ikke med så alt for stor avstand fra hverandre. Av visualiseringene er det Support Vector Machine som følger faktisk forbruk best, med Random Forest like bak. Ved denne horisonten er det også enkelt å se at Ridge leverer det dårligste prediksjonsresultatet.



Figur 4.14 - Prediksjon vs. Faktisk forbruk - 10 dager

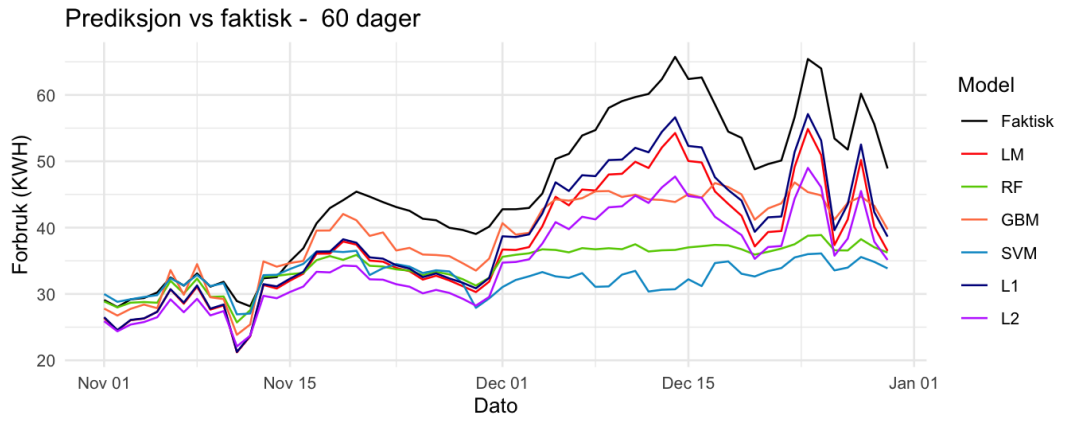
Av figur 4.15 for tidshorisonten på 30 dager ser tilsynelatende alle modellene ut til å matche mange av bevegelsene i det faktiske forbruket og alle presterer på et relativt likt nivå. Foreløpig ser samtlige modeller ut til å følge de største bevegelsene som i det faktiske forbruket, men med større usikkerhet sammenlignet med faktiske verdier fra midten av november. Gradient Boosting ligger nærmest den siste forbrukstoppen enn de øvrige modellene, og er derfor den modellen som presterer best på 30 dager. Modellen med dårligst prestasjon på 30 dager er Ridge, som også presterte dårligst i forrige figur.



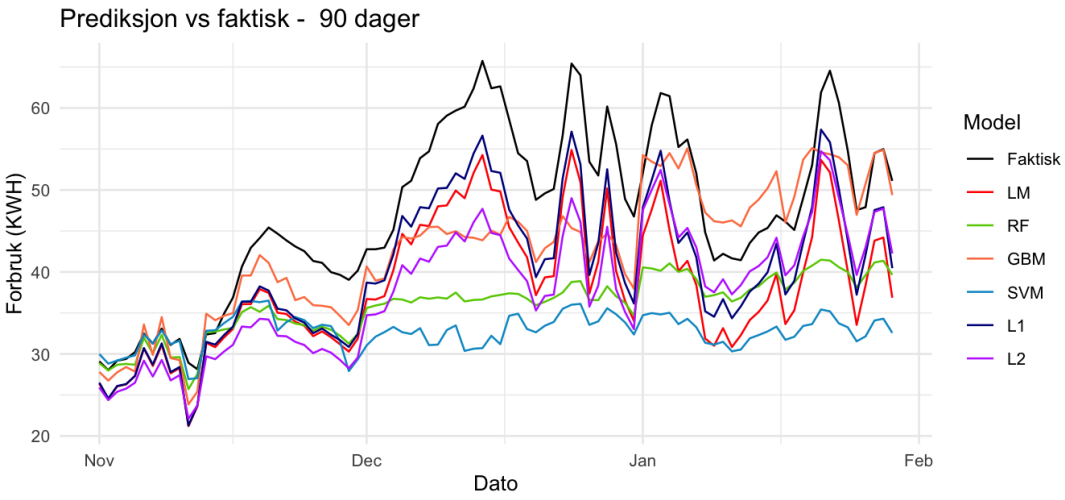
Figur 4.15 - Prediksjon vs. Faktisk forbruk - 30 dager

Tidshorisontene 60 og 90 dager frem i tid, som presentert i figur 4.16 og 4.17, viser større variasjon og forskjeller i prediksjonsevnen til modellene. I figur 4.16, for 60 dager frem i tid, følger LASSO og benchmark-modellen Lineær Regresjon det faktiske forbruket best. I denne tidshorisonten klarer ikke

Support Vector Machine og Random Forest, som var de beste på 10 dager, å følge det faktiske forbruket i like stor grad som øvrige modeller. Det er noen av de samme funnene på horisonten på 90 dager frem i tid, spesielt på modellene som ikke lenger henger med på bevegelsene til faktisk forbruk. Det er Support Vector Machine og Random Forest som er lengst fra å treffe. Modellene med best prediksjonsevne på 90 dager er Gradient Boosting, med LASSO, Ridge og Lineær Regresjon like bak. Gradient Boosting beveger seg annerledes enn de tre andre modellene trukket frem. Dette kan skyldes oppbyggingen av modellene forklart i teorikapittelet.



Figur 4.16 – Prediksjon vs. Faktisk forbruk - 60 dager



Figur 4.17 – Prediksjon vs. Faktisk forbruk - 90 dager

Visualiseringene oppsummert viser at Random Forest og Support Vector Machine er modellene som presterer best på de korteste tidshorisontene frem i tid, og at de gradvis presterer dårligere jo lenger frem i tid de skal predikere. Ridge var modellen med den dårligste prediksjonen på 10 dager, men presterte bedre på lengre tidshorisonter. Etter 30 dager presterte Gradient Boosting klart best. På 60

dager derimot leverte LASSO og Lineær Regresjon de beste prediksjonene, mens det på 90 dager var Gradient Boosting, LASSO og Ridge som presterte best, med Lineær Regresjon rett bak. Ingen av disse modellene gjorde det spesielt bra på de korteste prediksjonshorisontene, men ble bedre desto lenger tidshorisont det skulle predikeres for. Ingen av modellene gjorde det bra på både korteste og lengste intervall. Det ble også mer variabilitet i prestasjonene jo flere dager frem i tid modellene skulle predikere for.

4.3.2. Modellnøyaktighet

Målinger av modellnøyaktighet er et nyttig verktøy for å vurdere påliteligheten til modellene. Målingene som blir brukt i denne studien er empirisk basert og gir resultater i form av tall, som gjør det mulig å beregne den totale nøyaktighetsscoren for modellene uavhengig av tidsintervall. Først vil resultatene for de ulike horisontene presenteres, før vi skal se hvordan modellene scorer totalt sett. Nøyaktighetsmålingene vist i tabell 4.5-4.9 er sortert etter beste RMSE. Først forklares det korteste tidsintervallet på 10 dager i tabell 4.5. Support Vector Machine har i denne tabellen laveste verdi av RMSE og MAPE, etterfulgt av Random Forest. Ingen av modellene i studien har spesielt dårlige verdier av RMSE eller MAPE. Modellene varierer mellom 0,417 som laveste og 3,804 som høyeste RMSE, og 0,99 som laveste og 12,397 som høyeste MAPE. Ser man derimot på R-kvadrert for denne perioden, kan man se at det kun er de to modellene Support Vector Machine og Random Forest som har verdier mellom 0 og 1. Dette betyr at de øvrige modellene, til tross for lave verdier av RMSE og MAPE, ikke vil være pålitelige ifølge R-kvadrert. Nederst på listen ligger Ridge etterfulgt av Lineær Regresjon og LASSO.

		RMSE	MAPE	R^2
10 Dager	SVM	0,417	0,990	0,929
	RF	1,122	2,935	0,487
	GBM	1,619	5,034	-0,070
	L1	2,860	9,244	-2,337
	LM	2,946	9,528	-2,540
	L2	3,804	12,397	-4,904

Tabell 4.5 - RMSE, MAPE og R kvadrert alle modeller - 10 dager

Tabell 4.6 viser resultatene ved 30 dager frem i tid, og allerede her endrer rekkefølgen på modellene seg betydelig. Ved denne prognoselengden er modellen Gradient Boosting den med lavest RMSE og MAPE, og den eneste modellen med positiv R-kvadrert. På bunn finner vi også her Ridge. De øvrige modellene har nokså like scorer.

		RMSE	MAPE	R^2
30 dager	GBM	3,611	8,275	0,617
	RF	5,847	11,159	-0,004
	L1	5,880	13,864	-0,015
	SVM	5,978	10,504	-0,049
	LM	6,163	14,523	-0,115
	L2	7,918	18,790	-0,841

Tabell 4.6 - RMSE, MAPE og R kvadrert alle modeller -30 dager

Etter 60 dager ser vi i tabell 4.7 LASSO på topp, tett etterfulgt av Gradient Boosting og Lineær Regresjon. Disse scorer best på RMSE og MAPE, og er i tillegg de eneste modellene med positiv R^2. De modellene med dårligst score på denne prognoseperioden er Support Vector Machine og Random Forest med RMSE på 16,613 og 14,032, MAPE på 24,622 og 21,490, samt negativ verdi på R^2 lik -1.208 og -0,575.

		RMSE	MAPE	R^2
60 dager	L1	7,284	14,193	0,575
	LM	8,653	16,536	0,401
	GBM	9,080	13,818	0,340
	L2	11,573	21,986	-0,072
	RF	14,032	21,490	-0,575
	SVM	16,613	24,622	-1,208

Tabell 4.7 - RMSE, MAPE og R kvadrert alle modeller - 60 dager

Etter den lengste prognoseperioden på 90 dager, er også her LASSO på toppen med lavest RMSE og den beste verdien på R^2 på 0,454, i tabell 4.8. Gradient Boosting har lavest verdi på MAPE og har også gode resultater basert på R kvadrert. Som på 60 dager ligger også her Support Vector Machine og Random Forest nederst på tabellen på denne prediksjonsperioden.

		RMSE	MAPE	R^2
90 dager	L1	7,575	14,536	0,454
	GBM	7,898	11,759	0,407
	LM	9,705	18,447	0,104
	L2	10,451	19,166	-0,039
	RF	13,666	21,623	-0,776
	SVM	17,333	27,835	-1,857

Tabell 4.8 - RMSE, MAPE og R kvadrert alle modeller - 90 dager

Vi har testet de ulike modellene våre på ulike tidshorisonter, for å se hvordan de presterer på ulik prognoselengde. Oppsummert ser vi av tabellene 4.5-4.9 at både RMSE og MAPE øker over tid, noe som er naturlig ettersom lengre prognoser har en tendens til å være mindre nøyaktige over tid grunnet akkumulering av usikkerhet. Basert på estimert RMSE og MAPE, ser det ut til at LASSO er den modellen som presterer best jevnt over alle tidshorisonter. LASSO har i fleste tilfeller lavere RMSE og MAPE enn de andre modellene. Modellene som er svakest, er Support Vector Machine, Random Forest og Ridge, hvor Ridge har negativ R^2 for alle prognoseperioder og de to andre på tre av fire prognoseperioder. Mange av modellene har negativ R-kvadrert som øker over tid, som indikerer at de ikke passer til dataene. På grunn av dette gir disse modellene dårligere prediksjoner enn om man hadde laget prognoser kun basert på beregnet gjennomsnitt av historiske verdier. Av tabell 4.9 ser vi totalscoren for modellene, uavhengig av tidsperiode, som også inneholder noe av det samme som oppsummeringen over forklarer. LASSO og Gradient Boosting er de to klart beste modellene i vår sammenligning. Support Vector Machine scorer dårlig på totalbildet til tross for at den hadde best resultat på den korteste prognoseperioden, og skiller seg ut som den modellen med klart dårligst prediksjonsnøyaktighet. Random Forest har noen av de samme egenskapene som Support Vector Machine og havner nest nederst på totalen. Lineær Regresjon og Ridge scorer ok, da de klarer å ta deg inn med ok scorer på de lengste prediksjonsperiodene og havner derfor midt på treet i denne sammenligningen.

	RMSE	MAPE	R^2
L1	7,584	14,598	0,442
GBM	7,816	11,584	0,407
LM	9,754	18,604	0,076
L2	10,371	19,024	-0,044
RF	13,567	21,527	-0,787
SVM	17,267	27,894	-1,895

Tabell 4.9 - RMSE, MAPE og R kvadrert alle modeller – Samlet

5. Diskusjon og konklusjon

Etter teori- og metodegjennomgangen ble arbeidet med maskinlæringsmodellene iverksatt, hvor resultatene deretter ble presentert i kapittel 4. Basert på de foregående kapitlene vil vi i denne delen først diskutere forskningsspørsmålene med bakgrunn i teori og resultater, før vi i konklusjonen vil forsøke å svare på problemstillingen. Til slutt vil implikasjoner og etiske vurderinger diskuteres, før forslag til videre forskning blir presentert avslutningsvis.

Oppgavens problemstilling er som følger:

Problemstilling: Hvordan predikere kortsiktig strømforbruk hos norske husholdninger ved hjelp av maskinlæring?

Oppgavens tre forskningsspørsmål har som formål å bidra til besvarelsen på problemstillingen. De tre forskningsspørsmålene er som følger:

Forskningsspørsmål 1: Hvordan kan inkludering av pris som uavhengig variabel påvirke prediksjon av strømforbruk?

Forskningsspørsmål 2: Hvordan kan inkludering av ulike værforhold som uavhengige variabler påvirke prediksjon av strømforbruk?

Forskningsspørsmål 3: Hvilken maskinlæringsmodell presterer best ved prediksjon av kortsiktig strømforbruk i det norske strømmarkedet?

Forskningsspørsmål 1 og 2 går ut på å undersøke hvorvidt de uavhengige variablene kan brukes i studien og hvordan de eventuelt kan påvirke prediksjonsevnen til de utvalgte maskinlæringsmodellene. Forskningsspørsmål 3 har som formål å undersøke hvilken av maskinlæringsmodellene som presterer best på kortsiktig prediksjon til vårt formål, med tilgjengelig data og gitte forutsetninger.

5.1. Diskusjon

Ved å analysere og sammenligne resultatene fra de utvalgte maskinlæringsmodellene, har vi funnet modellenes styrker og svakheter i forbindelse med kortsiktig strømforbrukprediksjon. Data som ble brukt består av historiske data om strømforbruk, -pris og ulike værforhold, hvor strømforbruket blir brukt som avhengig variabel. Når forskningsspørsmålene skal undersøkes, vil det i hovedsak være basert på de utvalgte nøyaktighetsmålingene og en rangering av modellenes verdier av RMSE. MAPE og R-kvadrert vil også inkluderes i evalueringen for å gi et tydeligere bilde av prediksjonsevnen. Det er viktig å merke seg at mens RMSE og MAPE gir en indikasjon på den absolutte størrelsen på feilene

mellom predikerte og faktiske verdier i to ulike format, sier R-kvadrert noe om hvor godt modellene fanger opp de underliggende mønstrene i dataene (Kuhn & Johnson, 2013). En R-kvadrert over null er å foretrekke, da en negativ verdi vil tilsi at modellen er dårligere enn det naive gjennomsnittet av historiske verdier, men det er likevel greit å ta med i betraktningen at dette ikke er det eneste kriteriet man bruker for å vurdere modellenes nøyaktighet i en prediksjonskontekst. R-kvadrert er derfor en måling som studert sammen med RMSE og MAPE kan bidra med en dypere forståelse av modellenes prediksjonsevne, men som alene ikke kan være en avgjørende faktor for hvilken modell som skal velges.

Forskningsspørsmål 1 og 2

Første og andre forskningsspørsmål har som formål å avdekke hvorvidt de utvalgte uavhengige variablene faktisk kan bidra til god prediksjonsevne, eller om de kun forårsaker forstyrrelser og støy for modellene. Valg av variabler ble i utgangspunktet basert på tidligere forskning utført av blant annet Kang og Reiner (2021), som diskuterte værvariablenes påvirkning på strømforbruket i private husholdninger. Værvariabler som temperatur viste seg å ha høy grad av påvirkning på forbruket, men også nedbør og solskinn viste seg å ha varierende grad av påvirkning. Energifakta Norge (2024b) beskriver også at variasjonene i strømforbruket kan forklares av svingninger i værforhold, i likhet med hva Kang og Reiner (2021) diskuterer, men trekker også frem pris som en påvirkende faktor. For å undersøke om dette også var tilfelle i studien vi skulle gjennomføre, ble det utført tester for å undersøke korrelasjonen og prediksjonsevnen til de ulike modellene ved inkludering av enkelte av de uavhengige variablene. Resultatene fra testingen viste at det beste utfallet av prediksjoner var med alle uavhengige variabler inkludert, som er i tråd med Kang og Reiner (2021) sine funn og påstanden fra Energifakta Norge (2024b).

De uavhengige variablene ble undersøkt mot den avhengige variabelen ved hjelp av spredningsdiagram. Analysen av figurene viste at temperaturvariabelen var den eneste av de uavhengige variablene som viste høy grad i korrelasjon. Figuren viste en negativ korrelasjon som antyder en trend som sier noe om at ved lavere temperaturer brukes mer strøm. Resten av de uavhengige variablene, pris, nedbør og solskinnstimer viste lav eller ingen tydelig trend med den avhengige variabelen. Til tross for lav grad av korrelasjon i bivariat analyse, kan det likevel være underliggende interaksjoner mellom variablene i datasettet selv om de nødvendigvis ikke synes i denne typen analyse. Når det i tillegg er snakk om et forbruksmønster som styres utelukkende av menneskelig atferd, ligger det gjerne mer kompleksitet bak og det er sjeldent påvirket av bare én faktor. Dermed ble alle variabler inkludert hver sin prediksjonstest for å teste hvordan modellenes prediksjonsevne ble

påvirket. Testen hadde som hensikt å studere hvordan inkludering av enten pris- eller værvariabler hver for seg påvirket prediksjonsresultatene, slik at vi kunne sammenligne prediksjonsnøyaktigheten ved hjelp av RMSE, MAPE og R-kvadrert med prediksjonsresultatene med alle variabler inkludert. Resultatene fra test av prediksjonsevne med kun pris som uavhengig variabel viste ikke særlig tilfredsstillende resultater, med blant annet alvorlig dårlige R-kvadrert-verdier langt under null. Testen med kun værvariabler ga på sin side betydelig bedre resultater av alle målinger, med nesten en halvering av RMSE- og MAPE-verdier sammenlignet med resultatene prisvariabelen ga. Til tross for beder verdier, var det likevel kun to modeller som hadde R-kvadrert-verdier over null ved denne testen.

Til tross for at prediksjonstesten med kun prisvariabel ikke ga særlig gode resultater ble prediksjonsevnen med alle variabler inkludert bedre enn prediksjonen med kun værvariabler. Det kan være flere årsaker til hvorfor prediksjonsevnen forbedret seg etter å inkludere en variabel med lite korrelasjon med den avhengige variabelen. Blant annet kan det som nevnt være skjulte og uoppdagede interaksjoner mellom variablene i datasettet, som ikke nødvendigvis kommer frem i en enkel analyse. Selv om prisvariabelen har lav grad av korrelasjon med den avhengige variabelen, kan det som nevn være skjulte og uoppdagede interaksjoner med andre variabler i datasettet. For eksempel kan pris og nedbør ha en lineær relasjon, da vannmengde i vannmagasiner kan påvirke prisen på strømmen. Disse uoppdagede interaksjonene i datasettet kan også være en årsak når det er snakk om modellkompleksitet eller modelltype. Enkelte modeller kan håndtere mer komplekse forhold mellom variabler bedre enn andre, og kan derfor enkelt oppdage mønstre i variabler som ikke er synlige ved enkle analyser.

På en annen side kan inkludering av flere uavhengige variabler også bidra til støy, som igjen kan føre til overtilpasning av modellene. Ulike former for støy ble i teoridelen presentert av Kuhn og Johnson (2013) som en faktor som kan påvirke prediksjonsevnen, da støy av ulike former kan være forstyrrende for modellenes prediksjonsevne. Sannsynligheten for støy i datagrunnlaget kan være høy dersom det er snakk om data som er brukt i denne forbindelse, fordi det til tider kan oppstå store svingninger i observasjoner i variablene fra dag til dag gjennom året. I tillegg kan det oppstå støy i form av irrelevante uavhengige variabler, som forstyrrer prediksjonsevnen. Irrelevante variabler kan i tillegg føre til unødvendig komplekse modeller (James et al., 2021). Med støy er risikoen for at modellene trener for mye på støy også større, og man kan dermed oppleve overtilpasning i modellen. Overtilpasning ser imidlertid ikke til å være et problem sett ut ifra resultatene av nøyaktighetsmålingene. Oppsummering av sammenligningen mellom prediksjonstester med ulike uavhengige variabler ble vist i tabell 4.4, og begrunner avgjørelsen om å uansett inkludere alle de utvalgte variablene som uavhengige variabler i datagrunnlaget.

Forskningsspørsmål 3

Det tredje forskningsspørsmålet har som formål å innlede til en undersøkelse av hvilken av de ulike maskinlæringsmodellene som presterer best ved prediksjon av strømforbruket, og skal gjøre det mulig å besvare selve problemstillingen. Modellene ble testet og analysert for ulike tidshorisonter på 10, 30, 60 og 90 dager før det totale bildet ble analysert. Horisontene hadde som hensikt å gjøre det mulig å studere modellenes prestasjon over ulike prediksjonsperioder, og se om det var variasjoner i rangeringen. Basert på funnene i analysekapittelet ser vi at det er variasjoner i prediksjonsevne, som blant annet kan skyldes ulike tilnærminger til læring og prediksjon. Selv om modellenes prediksjonsevne varierer over alle tidshorisonter, er det noen tydelige vinnere og tapere i her av tidshorisontene og i det totale bildet. Vi vil gjennomgå resultatene fra hver av horisontene og drøfte eventuelle årsaker, før vi ser på totalen.

For den korteste tidshorisonten kunne man se av nøyaktighetsmålingene i figur 4.7 at Support Vector Machine har de beste verdiene av RMSE og MAPE, med en R-kvadrert på et nivå som tilsier at over 90 % av variansen i målvariabelen forklares av modellen. Like bak kommer *ensemble*-modellene Random Forest og Gradient Boosting. Ingen av de studerte modellene viser spesielt dårlige verdier av RMSE eller MAPE. Benchmark-modellen har på denne tidshorisonten den høyeste verdien av RMSE og MAPE, men likevel bare en verdi av MAPE som tilsier at det er kun 9 % av feil i prediksjonen.

Selv om verdiene av RMSE og MAPE ikke viser særlig grunn til bekymring, er det på en annen side kun Support Vector Machine og Random Forest som har R-kvadrert innenfor skalaen fra 0 til 1, som vil si at de er de eneste modellene som for denne tidshorisonten har god nok forklaringsevne basert på denne målingen (James et al., 2021). En kortere tidshorisont er ofte mindre volatil og mer forutsigbar enn lengre perioder, og kan derfor være enklere for Support Vector Machine og Random Forest å håndtere basert på egenskapene modellene innehar. Support Vector Machine vil på sin side jobbe med å finne en beslutningsgrense som maksimerer marginen mellom klassene den deler dataene inn i (Kuhn & Johnson, 2013). Den kan dermed tilsynelatende se god ut i en kortere tidshorisont, fordi det er begrenset med variasjon i observasjonene på i løpet av denne perioden. På samme måte har Random Forest på sin side sine fordelaktige egenskaper som blant annet tillater den å håndtere mange variabler med få observasjoner med sin robusthet mot støy (Kuhn & Johnson, 2013). Håndteringen av støy kan gjøre det enklere å trene de ulike beslutningstremodellene effektivt fordi observasjonene foreløpig fortsatt er få og lite varierende, og dermed også enklere å oppdage.

Allerede etter 30 dager har det skjedd en endring i resultatene, og de to beste modellene fra 10-dagersprediksjonen har fått betydelig dårligere resultater sammenlignet med resten av modellene. RMSE-verdiene er fortsatt jevnt fordelt blant de ulike modellene, med Gradient Boosting som den

tydelige vinneren basert på nøyaktighetsmålingene. Det er en økning i verdiene av RMSE og MAPE etter 30, dager, noe som er naturlig fordi lengre prediksjoner ofte innebærer mer usikkerhet. Disse verdiene vil sannsynligvis fortsette å øke i de lengre periodene. Likevel ligger alle verdier av MAPE fortsatt under 20 for alle modeller, som vil si at modellene generelt sett gir en akseptabel grad av feil. Til tross for at modellene ser ut til å gi en akseptabel grad av feil for denne perioden også, er det i denne perioden kun Gradient Boosting som har en positiv verdi av R-kvadrert.

Hovedårsaken til den store endringen i rangering kan blant annet skyldes mer fremtredende sesongvariasjoner i dataene, spesielt med tanke på at vinteren har begynt og det faktiske forbruket kan ha økt mer enn modellene estimerte. Endringen i forbruksmønsteret er blant faktorene som kan forklare reduksjonen i Random Forest-modellens prestasjon, dersom den ikke har klart å tilpasse seg endringene. Fordelen til Gradient Boosting i dette tilfellet kan være at den ved hjelp av sekvensielle læringsprosesser forbedrer seg iterativt ved å legge til nye beslutningstremodeller som korrigerer feil fra tidligere trær, i stedet for parallelle prosesser slik som Random Forest (Kuhn & Johnson, 2013). I denne situasjonen kan modellkompleksiteten til Gradient Boosting spille positivt inn.

Til tross for at *ensemble*-teknikker skal være robuste og prestere godt på datasett med mye data og komplekse relasjoner, ser det ikke ut til at Random Forest-modellen klarer å fange opp sesongvariasjonene i datasettet som ble brukt i treningen (James et al., 2021). Dette kan blant annet skyldes en høy grad av ikke-lineære interaksjoner eller støy i data, som igjen er forårsaket av den varierende graden av korrelasjon i de uavhengige variablene. Mangelen på denne korrelasjon kan være en årsak til at Random Forest-modellen ikke presterer bedre, da den som oftest er fungerer bedre på datasett med et stort antall korrelerte prediktorvariabler (James et al., 2021). Gradient Boosting ser ut til å håndtere denne mangelen bedre. I studien gjennomført av Januschowski et al. (2022) blir ulike trebaserte prediksjonsteknikker undersøkt for å finne den mest effektive. Forskerne konkluderer med at det er potensiale i beslutningstrebaserte modeller og at deres suksess kan tilskrives deres robusthet og at minimal tilpasning er nødvendig for å oppnå god ytelse. De trekker spesielt frem Gradient Boosting, med beslutningstre som grunnmodell, som modellen med best ytelse i deres studie. Deres konklusjon samsvarer med våre resultater på Gradient Boosting, mens Random Forest presterer ikke like godt som forventet. I oppgaven er det gjort minimalt med tilpasninger på modellene og det ser ut som Gradient Boosting tilpasset seg dataen best av de beslutningstrebaserte modellene.

Resultatene etter 60 og 90 dager tyder på mer stabile resultater, og en klar indikasjon på hvordan fremtidige resultater kunne sett ut. Som antatt har usikkerheten naturligvis økt for periodene sammenlignet med tidligere perioder. Etter disse periodene er LASSO-modellen den med de beste resultatene rangert etter RMSE. Like etter følger Gradient Boosting og Benchmark-modellen. R-

kvadrert følger disse resultatene og rangeringen av topp tre. Det er også i denne perioden ikke flere enn kun disse tre som har positive verdier av R-kvadrert. I det totale tilfellet, uavhengig av tidshorisontene, blir resultatene fra 60 og 90 dager stående med LASSO, Gradient Boosting og benchmark-modellen på topp tre rangert etter RMSE. Det er også fortsatt kun disse tre som viser R-kvadrert verdier over 0, som vil si at resterende modeller i dette tilfellet ikke når helt opp. At ingen av disse modellene presterer bedre enn benchmark-modellen kan skyldes en rekke faktorer, hvorav dataforberedelser og modelltuning kan være de avgjørende faktorene. Fordi modellene brukt i denne studien er gitt det minste nødvendige av tuning for prediksjon, trenger ikke resultatene som er blitt presentert å bety at modellene med R-kvadrert under null er ubrukelige. Det kan tyde på at modellene trenger ytterligere justeringer eller en annen tilnærming for å prestere bedre, da modellvalg og -ytelse ikke bare er avhengig av modelltypene og de ulike egenskapene de i utgangspunktet har.

Det er bemerkelsesverdig at LASSO-modellen holder seg stabilt godt i forhold til de andre modellene, noe som kan indikere at tilnærmingen den bruker for å håndtere regulariseringer og variabelseleksjon er gunstig i denne situasjonen og de lengste tidshorisontene som er studert. Sammenhengen i ytelse fra LASSO-modellen over de ulike tidshorisontene antyder at modellen har en god balanse mellom kompleksitet og evnen til å generalisere fra treningsdataene. Både LASSO og Ridge er modeller som implementerer en regulariserende straffefunksjon som har som formål å redusere variansen ved å krympe koeffisienter (James et al., 2021). LASSO utfører på sin side en innebygd variabelseleksjon når den setter koeffisientene lik null slik at ikke-informative prediktorer kan filtreres ut og dermed redusere påvirkningen av støy, noe Ridge på andre siden ikke gjør (James et al., 2021). Denne forskjellen kan forklare hvorfor regulariseringsmodellene ikke presterer på samme nivå. Variabelseleksjonen som LASSO utfører kan potensielt sett gi mer robuste modeller over tid, som generaliserer bedre på lengre tidshorisonter. Det er også denne variabelseleksjonsegenskapen Albuquerque et al. (2022) legger vekt på i sin studie ved valg av LASSO-lars som funksjon. Dette kan være noen av de avgjørende egenskapene for akkurat dette tilfellet. Til tross for at Random Forest også skal være robust mot overtilpasning kan den ha større problemer med å tilpasse seg de langvarige trendene og sesongvariasjonene, som for eksempel LASSO ser ut til å takle bedre enn de øvrige modellene.

Formålet med studien vi gjennomførte var å undersøke potensialet til maskinlæringsmodeller som krever minimalt manuelt arbeid, i tillegg til å undersøke hvilke av de utvalgte modellene som presterte best. Som Saunders et al. (2023) legger vekt på, er det også viktig å skape en forståelse av «hvorfor» et resultat blir som det blir, ikke bare «hvor effektivt» det er. Vi har dermed forsøkt å drøfte årsakene til resultatene fra studien som er gjennomført. Gjennom analysens resultater ser man variasjon i hvilke modeller som har prestert best over de ulike horisontene. Random og Support Vector Machine så tidlig

ut til å levere gode prediksjoner, men på en annen side er det begrenset hvor mye variasjon det er i faktisk forbruk etter kun 10 dager, og dermed er det også begrenset hvor troverdig disse resultatene faktisk er. Resultatene ble motbevist allerede etter 30 dager og prediksjonsevnen ble betydelig svekket for de lengre prediksjonsperiodene. Etter siste prediksjonsperiode er både Random Forest og Support Vector Machine på bunnen av listen. Resultatene til modellene i vår studie er ikke i tråd med våre forventninger basert på Shapi et al. (2021) sine funn om at Support Vector Machine skal fungere godt på prediksjon av gjennomsnittlig strømforbruk og heller ikke Albuquerque et al. (2022) sine funn om at Random Forest skulle prestere bra. Shapi et al. (2021) skrev at modellen hadde krav om mye treningstid, noe vi ikke har gitt noen av modellene våre og som dermed kan forklare dens dårlige resultater i vår studie. Til tross for egenskapene Random Forest og Support Vector Machine innehar, kan det se ut som de ikke evner å fange opp de riktige egenskapene i datagrunnlaget. Dette kan skyldes flere faktorer, som for eksempel mengden data, variasjon i korrelasjon mellom variabler, støy eller da for lite treningstid. *Ensemble*-modellen Gradient Boosting ser på sin side ikke ut til å ha noe særlig problem med dette, noe som førte til at Gradient Boosting totalt sett havnet i toppsjiktet på prediksjonsnøyaktighet sammen med regulariseringsmodellen LASSO og benchmark-modellen Lineær Regresjon. Resultatene kan tyde på at regulariseringstilnærmingen LASSO benytter er fordelaktig for denne typen datagrunnlag og prediksjon, og det ser ut som modellen klarer å balansere kompleksitet og generalisering fra treningsdataene godt. På bakgrunn av dette kan vi si at maskinlæringsmodellen LASSO er den som predikerer best på kortsiktig strømforbruk med vår metode på dataen vi har benyttet.

5.2. Konklusjon

Målet med denne oppgaven har vært å besvare følgende problemstilling:

Hvordan predikere kortsiktig strømforbruk hos norske husholdninger ved hjelp av maskinlæring?

For å besvare spørsmålet gjennomførte vi en kvantitativ studie basert på sekundærdata fra et strømselskap, i tillegg til at vi samlet inn data fra Forbrukerrådet og Norsk Klimaservicesenter. Oppgaven utforsker evnen ulike maskinlæringsmodeller har til å forutsi strømforbruk hos norske husholdninger ved å analysere historiske data om strømforbruk, strømpriser og værforhold fra Hamarregionen. Deretter analyserte vi variablene, hvor våre funn ga oss tilstrekkelig med bevis for å kunne trekke slutninger. Valg av maskinlæringsmodeller ble gjort utfra popularitet og kompleksitet i modellbyggingen. I studien ble det benyttet tre forskjellige nøyaktighetsmålinger, samt visualiseringer, for å vurdere maskinlæringsmodellenes prediksjonsnøyaktighet. Modellene ble også testet på fire ulike

tidshorisonter frem i tid for å vurdere deres stabilitet. Resultatene skiftet for de ulike periodene de predikerte for, spesielt på de korteste tidshorisontene.

For å diskutere problemstillingen stilte vi også forskningsspørsmål om inkludering av de uavhengige variablene pris og værforhold kan påvirke prediksjonen, i tillegg til spørsmålet om hvilken maskinlæringsmodell som presterer best. Forskningsspørsmålene, og svaret på problemstillingen generelt, baserer seg på metoden vår og datagrunnlaget vårt.

Resultatene av studien belyser viktigheten av å velge passende uavhengige variabler. Ved å inkludere strømpriser og værforhold i modellene, kunne prediksjonsnøyaktigheten forbedres betydelig. Dette tyder på at eksterne faktorer som vær og markedsdynamikk spiller kritiske roller i strømforbruket og må tas i betraktning ved utvikling av prognosemodeller. Dette til tross for lav korrelasjon mellom forbruk og de fleste forklaringsvariablene. Det var kun sterk korrelasjon mellom forbruk og temperatur. I tillegg til dette viste tester utført at inkludering av prisvariabler alene ikke kunne forklare strømforbruk. Funnene er viktige fordi de sier at man ikke nødvendigvis bør forkaste enkelte uavhengige variabler kun basert på tester av korrelasjon mot den avhengige variabelen.

Ved hjelp av forskningsspørsmål 3 fant vi at LASSO-modellen gjorde det best basert på de ulike målingene. Dette fordi den var mest stabil over de ulike tidsperiode, i tillegg til at den også presterte best sett under ett, uavhengig av tidsintervallene. Av de samme årsakene havnet Gradient Boosting like bak LASSO, og Lineær Regresjon som tredje beste modell. Modellene som er utviklet for å håndtere kompleksitet i data, som Random og Support Vector Machine, havnet på bunn av målingene. Dette, sammen med resultatene til Lineær Regresjon som den enkleste modellen i utvalget, tyder på at det i dette tilfellet er det enkle som er det beste. Våre funn er i tråd med Collopy et al. (1994) sin beskrivelse av prognoselitteraturen om at enkle modeller er å foretrekke, med mindre det er god grunn for kompleksitet, noe det ikke er i denne studien. Variasjonen i nøyaktighetsmålingene viser at testing av ulike maskinlæringsmodeller er kritisk før valg av modell.

Det er viktig å understreke at modellvalg og -ytelse ikke bare er avhengig av de ulike modelltypene og egenskapene de i utgangspunktet har, men også hvordan de blir implementert, tilpasset og tunet. Maskinlæringsmodeller har endeløse muligheter til å tunes og kalibreres for å sikre nøyaktige prediksjoner. Vi kan derfor ikke konkludere med at resultatene for vår studie er fasiten på de faktiske forhold, med tanke på at alle modellene kunne vært ytterligere tilpasset for å sikre bedre prediksjonsnøyaktighet. Dette vil si at arbeidet vi har gjort med modellene ikke er tilstrekkelig for å kunne utelukke at de andre modellene kan prestere bedre. I tillegg vil valg av metode og tilgjengelig datagrunnlag påvirke resultatene. Rangeringen kan derfor se annerledes ut under andre forutsetninger enn de vi ga maskinlæringsmodellene.

Problemstillingen besvares gjennom en kombinasjon av praktiske funn, teoretisk innsikt og tidligere forskning, som viser at maskinlæring kan være et kraftfullt verktøy for å forutsi strømforbruk i norske husholdninger. De to beste maskinlæringsmodellene og de ulike variablene har vist potensiale i å predikere strømforbruk, noe som understreker mulighetene for maskinlæring i kraftmarkedet. Gjennom videre forskning og utvikling av modellene kan nøyaktigheten og påliteligheten til slike prediksjoner forbedres ytterligere, til fordel for både energiselskaper og forbrukere. For å predikere kortsiktig strømforbruk hos norske husholdninger ved hjelp av maskinlæring kan man basert på studiens funn benytte maskinlæringsmodellene LASSO og Gradient Boosting. Modellene viser en evne til å analysere og forutsi fremtidige forbruksmønstre basert på historiske strømdata, værvariabler og prisdata.

Videre konkluderer vi også med at maskinlæringsmodellers prediksjonsresultater i stor grad avhenger av de forutsetninger man gir og data man har tilgjengelig.

5.3. Implikasjoner

Denne masteroppgaven har undersøkt hvordan maskinlæringsmodeller kan benyttes til å predikere strømforbruk hos private husholdninger i Norge, med særlig vekt på modellene LASSO og Gradient Boosting sine resultater. Studien avdekker flere mulige implikasjoner for praksis og teoretisk forståelse innen energistyring og forbruksoptimalisering.

Funnene indikerer at maskinlæringsmodeller, spesielt LASSO, bør testes i eksisterende energistyringssystemer for å undersøke om de kan forbedre nøyaktigheten av strømforbruksprognoser. Dette kan derfor ha direkte implikasjoner for hvordan energiselskaper kan håndtere lastbalansering og etterspørselsrespons mer effektivt. For eksempel kan mer presise forbruksprognoser muliggjøre bedre beslutninger om når og hvor mye strøm som skal kjøpes på spotmarkedet, noe som kan redusere kostnader og forbedre systemets pålitelighet.

Studiens resultater kaster lys på mulighetene nødvendige politiske tiltak kan gi. Politiske beslutningstakere kan bruke innsikt fra analyser av store datamengder til å forme reguleringer som fremmer energieffektivitet og fornybar energiproduksjon. Videre kan en bedre forståelse av forbruksmønstre lede til mer målrettede incentiver for energibesparelse blant forbrukere, for eksempel gjennom differensierte strømpriser eller subsidier for energieffektive løsninger.

Effektiv bruk av maskinlæringsmodeller for strømforbruk kan spille en kritisk rolle i overgangen til en mer bærekraftig energiforsyning. Ved å forbedre nøyaktigheten i forbruksprognoser, kan energiselskaper bedre integrere fornybare energikilder som sol og vind, som ofte er mindre

forutsigbare enn fossile brensler. Dette ikke bare optimaliserer bruken av fornybar energi, men bidrar også til reduksjon av karbonutslipp ved å minske avhengigheten av fossile brensler.

Denne forskningen tilfører et bidrag om bruken av maskinlæring i energisektoren, spesielt innen prediksjon av forbruksdata. Det teoretiske rammeverket og de anvendte metodene kan gi grunnlag for fremtidig forskning og teoriutvikling, samt bidra til en dypere forståelse av sammenhengene mellom teknologisk innovasjon og energiforvaltning.

Implikasjonene av denne studien er omfattende og peker på viktigheten av å videreutvikle og integrere avanserte analytiske verktøy i energisektoren. Ved å utnytte potensialet i maskinlæring, kan energiselskaper og politiske beslutningstakere gjøre mer informerte valg som ikke bare forbedrer økonomisk effektivitet, men også fremmer miljømessig bærekraft.

5.4. Etiske vurderinger

Forskningsetikk refererer til de standarder for atferd som styrer vår oppførsel i forhold til rettighetene til de som blir gjenstand for vår studie, eller blir berørt av den (Saunders et al., 2023). Etiske problemstillinger vil ifølge Saunders et al. (2023) være viktig gjennom hele forskningen, og hvert steg vil kreve etisk integritet i forhold til rollen vår som forskere.

Vår innsamlede data kommer både fra offentlige kilder og fra data delt med oss fra et strømselskap. De offentlige dataene er anonymisert og er ikke sporbare. Dataen fra strømselskapet er også anonymisert, både for oss og for de som skal lese studien vår. Målepunkt-id som representerer hver enkelt kunde i denne studien kan spores til hver enkelt kommune, men ikke videre utover det. Strømselskapet er heller ikke navngitt i studien. I tillegg til dette baserer vi oss på at forskerne bak dataen som er samlet inn, både fra offentlige kilder og strømselskapet, har forholdt seg til deres etiske retningslinjer.

Maskinlæring har flere etiske implikasjoner, særlig med tanke på prinsippet om transparens (Henderson et al., 2017). Vi vet at maskinlæringsalgoritmer lærer, men ikke hvordan de lærer, og vi kan derfor ikke si noe om hvorfor og hvordan alt utarter underveis i læringsprosessen. Derfor har det vært viktig å belyse begrensninger, valg av metode og data, i tillegg til validitet og reliabilitet i studien, slik at vi har kontroll på at studien vår blir utført så korrekt som mulig. Dette sikrer også at forskningens formål går foran våre egne oppfatninger og antakelser.

5.5. Videre forskning

Fremtidige studier innen prediksjon av strømforbruk hos private husholdninger ved hjelp av maskinlæring bør sette søkelys på flere nøkkelområder for å utvide forståelsen og anvendbarheten av funnene fra denne oppgaven. Disse områdene inkluderer geografisk og demografisk generalisering, integrering av flere variabler og mer langsiktige prognoser.

Geografisk fokus på Hamarregionen tilfører en forståelse av regionspesifikke forbruksmønstre, men begrenser samtidig generaliserbarheten av våre funn. Videre arbeid bør inkludere data fra flere regioner, for å gi en mer omfattende forståelse av strømforbruket i Norge.

Flere variabler bør testes og vurderes inkludert. Vår oppgave har undersøkt de uavhengige variablene strømpris og værforhold. Det anbefales at fremtidige studier inkluderer sosioøkonomiske og adferdsmessige data, som inntektsnivå, husholdningers størrelse og strømforbruksvaner. Dette vil muliggjøre en mer helhetlig tilnærming til prediksjonen og mulig avdekke dyptgående mønstre som kan forbedre prediksjonsnøyaktigheten av strømforbruk. Inkludering av mer data, fra flere år tilbake i tid, som også vil gi mer informasjon om forbruksmønstre kan også forbedre produksjonsevnen ytterligere.

Oppgaven vår har vektlagt betydningen av kortsiktige prognoser. I lys av teknologiske og industrielle endringer kan det være nyttig å forstå hvordan strømforbruket vil utvikle seg over lengere perioder. Dette kan også være informativt for strategisk planlegging i kraftindustrien, samt for å nå de langsiktige målene om bærekraft.

Ved å adressere disse punktene, håper vi at vår diskusjon kan bidra til en dypere forståelse av maskinlæringens potensiale i energisektoren. Fremtidig arbeid kan utvide gyldigheten av får funn, men også bidra til mer presis styring og planlegging i energisektoren. Dette er avgjørende for å møte både dagens og morgendagens energiutfordringer i et stadig mer dynamisk landskap. Mer nøyaktige prognoser kan hjelpe energiselskaper, og kraftmarkedet generelt, gjennom å planlegge bedre og respondere på forespurt energibehov. Som vi så innledningsvis vil nettopp denne responsen være viktig sett i lys av overgangen til fornybare energikilder. Forbedrede prediksjoner av strømforbruk vil videre kunne støtte mål om bærekraft gjennom optimalisering av energiforbruk og redusere unødvendig produksjon.

Referanser

Abbass, M. A. B., & Hamdy, M. (2021). A Generic Pipeline for Machine Learning Users in Energy and Buildings Domain. *Energies*, *14*.

Albuquerque, P. C., Cajueiro, D. O., & Rossi, M. D. C. (2022). Machine learning models for forecasting power electricity consumption using a high dimensional dataset. *Expert Systems with Applications*, *187*, 115917. https://doi.org/10.1016/j.eswa.2021.115917

Alpaydin, E. (2014). *Introduction to Machine Learning* (3.). MIT Press.

Berk, R. A. (2020). *Statistical Learning from a Regression Perspective* (3.utg.). Springer International Publishing AG. https://ebookcentral.proquest.com/lib/hilhmr-ebooks/reader.action?docID=6246037

Collopy, F., Adya, M., & Armstrong, J. S. (1994). Principles for Examining Predictive Validity: The Case of Information Systems Spending Forecasts. *Information Systems Research*, *5*, 170–179.

CRAN - The Comprehensive R Archive Network. (u.d.). *A short introduction to the caret Package*. https://cran.r-project.org/web/packages/caret/vignettes/caret.html

Energifakta Norge. (2024a). *Forsyningssikkerhet*. https://energifaktanorge.no/norsk-energiforsyning/forsyningssikkerhet/

Energifakta Norge. (2024b). *Hva påvirker energibruken?* https://energifaktanorge.no/norsk-energibruk/hva-pavirker-energibruken/

Energifakta Norge. (2024c). *Kraftmarkedet*. https://energifaktanorge.no/norsk-energiforsyning/kraftmarkedet/

Energifakta Norge. (2024d). *Strømnettet*. https://energifaktanorge.no/norsk-energiforsyning/kraftnett/

Forbrukerrådet. (u.å.). *Spotpriser*. https://www.forbrukerradet.no/strompris/spotpriser

Fortum. (u.å.). *Hva er fleksibilitet?* https://www.fortum.no/strategi/en-renere-verden/hva-er-fleksibilitet

Henderson, D., Earley, S., & Data Administration Management Association (Red.). (2017). *DAMA-DMBOK: Data management body of knowledge* (2.utg.). Technics Publications.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2.utg.). OText. https://otexts.com/fpp2/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., & Gasthaus, J. (2022). Forecasting with trees. *International Journal of Forecasting*, *38*, 1473–1481.

Johannessen, A., Tufte, P. A., & Christoffersen, L. (2016). *Introduksjon til samfunnsvitenskapelig metode* (5. utg.). Abstrakt forlag.

Kang, J., & Reiner, D. M. (2021). What is the effect of weather on household electricity consumption? Empirical evidence from Ireland. *Energy Policy Research Group*, *2112*.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Norges vassdrags- og energidirektorat. (2023). *Langsiktig kraftmarkedsanalyse 2023*. https://www.nve.no/energi/analyser-og-statistikk/langsiktig-kraftmarkedsanalyse/langsiktig-kraftmarkedsanalyse-2023/

NorgesEnergi. (u.å.). *Slik fungerer strømmarkedet*. https://norgesenergi.no/bedrift/om-strommarkedet/slik-fungerer-strommarkedet/

NorgesEnergi. (2023, september 20). *Slik påvirkes strømprisen*. Slik påvirkes strømprisen. https://norgesenergi.no/stromsmart/dette-pavirker-stromprisen/

Norsk Klimaservicesenter. (u.å.). Observasjoner og værstatisikk. https://seklima.met.no

Ramasubramanian, K., & Moolayil, J. (2019). *Applied Supervised Learning with R*. Packt Publishing.

r-project. (u.å.). *What is R?* What is R? https://www.r-project.org/about.html

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2023). *Research Methods for Business Students* (9. utg.). Pearson.

Shapi, M. K. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment*, *5*.

Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective* (4. utg.). Pearson.

Statistisk sentralbyrå. (2023). *Markant fall i husholdningenes strømforbruk i 2022*. https://www.ssb.no/energi-og-industri/energi/statistikk/elektrisitet/artikler/markant-fall-i-husholdningenes-stromforbruk-i-2022

Statistisk Sentralbyrå. (u.å.a). *Produksjon og forbruk av energi, energibalanse og energiregnskap*. SSB Statistikkbanken. https://www.ssb.no/statbank/table/11558/tableViewLayout1/

Statistisk Sentralbyrå. (u.å.b). *Standard for næringsgruppering (SN)*. Statistisk Sentralbyrå. https://www.ssb.no/klass/klassifikasjoner/6

Statnett. (2022). *Økende forbruk gir kraftunderskudd fra 2027*. https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/nyhetsarkiv-2022/kortsiktig-markedsanalyse-okende-forbruk-gir-kraftunderskudd-fra-2027/

Thrane, C. (2018). *Kvantitativ metode, en praktisk tilnærming*. Cappelen Damm.

Vedlegg

Vedlegg til kapittel 3.

3.1 - Output: lm_model

```
Linear Regression
365 samples
 9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 328, 329, 329, 328, 329, 329, ...
Resampling results:
 RMSE   Rsquared  MAE
 2.969685 0.9503764 2.382477

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Vedlegg 3.1 - Output lm_model

3.2 - Output: rf_model

```
Random Forest
365 samples
9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 328, 329, 329, 329, 329, 329, ...
Resampling results across tuning parameters:

mtry RMSE   Rsquared  MAE
2  2.044328 0.9794094 1.493007
4  1.504988 0.9878716 1.095172
7  1.495878 0.9876636 1.074432
9  1.574778 0.9863065 1.118126
12  1.712935 0.9838256 1.206608

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 7.
```

Vedlegg 3.2 - Output - rf_model

3.3 - Output: gbm_model

```
Stochastic Gradient Boosting
365 samples
 8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 328, 328, 329, 328, 329, 329, ...
Resampling results across tuning parameters:

 shrinkage interaction.depth n.trees RMSE   Rsquared  MAE
 0.01      1                 100     7.073718 0.8976975 5.704844
 0.01      1                 150     5.532842 0.9269016 4.393450
 0.01      1                 200     4.509212 0.9442442 3.566413
 0.01      3                 100     5.802982 0.9588805 4.928661
 0.01      3                 150     4.150433 0.9632324 3.446609
 0.01      3                 200     3.231190 0.9666974 2.604089
 0.01      5                 100     5.542327 0.9667784 4.775965
 0.01      5                 150     3.881048 0.9695301 3.239825
 0.01      5                 200     2.984321 0.9716925 2.407154
 0.10      1                 100     2.335581 0.9707483 1.740964
 0.10      1                 150     2.260919 0.9723045 1.684938
 0.10      1                 200     2.198655 0.9738748 1.647577
 0.10      3                 100     2.027222 0.9776247 1.496632
 0.10      3                 150     2.001641 0.9781852 1.492944
 0.10      3                 200     1.976738 0.9786819 1.468499
 0.10      5                 100     2.071715 0.9767115 1.491179
 0.10      5                 150     2.033816 0.9774757 1.478044
 0.10      5                 200     2.030253 0.9774647 1.474648

Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 200, interaction.depth = 3, shrinkage = 0.1
 and n.minobsinnode = 10.
```

Vedlegg 3.3 - Output gbm_model

3.4 - Output: SVM_model

```
Support Vector Machine s with Radial Basis Function Kernel

365 samples
 9 predictor

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 329, 328, 329, 328, 329, 327, ...
Resampling results across tuning parameters:

 C   RMSE   Rsquared  MAE
 0.25 3.128751 0.9529951 2.149674
 0.50 2.372125 0.9714984 1.744367
 1.00 2.012492 0.9789145 1.534460
 2.00 1.883694 0.9813941 1.443496
 4.00 1.832575 0.9822338 1.401171

Tuning parameter 'sigma' was held constant at a value of 0.0973938
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 0.0973938 and C = 4.
```

Vedlegg 3.4 - Output svm_model

3.5 - Output: LASSO _model

```
glmnet

365 samples
 9 predictor

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 328, 328, 329, 329, 329, 329, ...
Resampling results across tuning parameters:

 lambda RMSE   Rsquared  MAE
 0.001  2.337479 0.9698151 1.871600
 0.012  2.337479 0.9698151 1.871600
 0.023  2.337479 0.9698151 1.871600
 0.034  2.337302 0.9698187 1.871696
 0.045  2.337676 0.9698137 1.873677
 0.056  2.338346 0.9698057 1.875905
 0.067  2.339265 0.9697942 1.878231
 0.078  2.340654 0.9697721 1.880665
 0.089  2.342425 0.9697420 1.883485
 0.100  2.344508 0.9697061 1.886466

Tuning parameter 'alpha' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 0.034.
```

Vedlegg 3.5 - Output: LASSO _model

3.6 - Output: Ridge_model

```
glmnet

304 samples
 8 predictor

Pre-processing: centered (10), scaled (10)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 273, 276, 272, 273, 276, 274, ...
Resampling results across tuning parameters:

 lambda RMSE   Rsquared  MAE
 0.001  2.30867 0.9673634 1.897284
 0.012  2.30867 0.9673634 1.897284
 0.023  2.30867 0.9673634 1.897284
 0.034  2.30867 0.9673634 1.897284
 0.045  2.30867 0.9673634 1.897284
 0.056  2.30867 0.9673634 1.897284
 0.067  2.30867 0.9673634 1.897284
 0.078  2.30867 0.9673634 1.897284
 0.089  2.30867 0.9673634 1.897284
 0.100  2.30867 0.9673634 1.897284

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.1.
```

Vedlegg 3.6 - Output: Ridge_model