Norges Handelshøyskole
Bergen, Høst 2018

# Elbilprediksjon

*En empirisk studie av karakteristika og prediksjon av elbil-eiere*

**Anders Wettergreen Gundersen og Selim Zeybek**

**Veileder: Jonas Andersson**

Masterutredning i økonomi og administrasjon

Hovedprofil: Business Analytics

NORGES HANDELSHØYSKOLE

# Forord

Denne masterutredningen er skrevet som en avsluttende del av masterstudiet i økonomi og administrasjon ved Norges Handelshøyskole (NHH), høsten 2018.

Vi var begge innstilt på å gjennomføre en kvantitativ analyse på et område som var nytt og tidligere lite utforsket. Bilmarkedet i Norge, som gjennomgår en stor omveltning etter elbilenes inntog, var et særlig interessant temaområde for oss å studere. Ved å gjennomføre prediksjoner innenfor dette fikk vi også benyttet vår kompetanse rundt statistiske metoder, maskinlæring og dataanalyse. Sammen utformet vi en problemstilling som omhandlet dette temaområdet, samt at et datagrunnlag ble sammenstilt fra flere offentlige kilder. Masterutredningen har vist seg å være både spennende, utfordrende og lærerikt.

Vi må rette en takk til vår veileder Jonas Andersson for raskt respons ved behov og god veiledning. Vi vil også takke Statens Vegvesen som gav oss tilgang til det nødvendige datagrunnlaget oppgaven er basert på.

Norges Handelshøyskole

Bergen, desember 2018

Anders Wettergreen Gundersen

Selim Zeybek

# Sammendrag

I denne utredningen studerer vi hva som karakteriserer elbil-eiere og undersøker videre om disse faktorene kan brukes til å predikere potensielle elbil-eiere. Tidligere studier basert på spørreundersøkelser har konkludert med at typiske elbil-eiere er relativt unge menn mellom 35-54 år med høy utdanning og inntekt. Denne utredningen skiller seg fra tidligere studier ved å ta i bruk prediksjonsmetoder for å bedre definere elbil-eiernes karakteristika og rangere disse etter viktighet. Til dette formålet benyttes et datasett over bilregistreringer i Oslo og Akershus mellom 2010-2018.

Prinsipalkomponentanalyse og korrespondanseanalyse viser at alderen til bileier, bostedskommune, politisk tilhørighet, samt høy inntekt og formue var interessante variabler for videre analyse. Det blir i tillegg gjennomført en variabelutvelgelse gjennom Boruta-algoritmen for tre tidsperioder som viser de viktigste forklaringsvariablene for å skille elbil-eiere fra andre bileiere. Gjennom bruk av trebaserte metoder innenfor prediksjon, estimeres det innledningvis en basismodell gjennom et klassifiseringstre som sammenligningsgrunnlag. Videre estimeres det modeller med henholdsvis random forest og extreme gradient boosting for tidsperiodene.

Utredningen konkluderer med at aldersgruppene 25-40 år og 40-60 år, andel husholdninger med en samlet formue over 4 millioner kroner i en kommune, antall ladestasjoner i kommunen, og bostedsadresse som ligger i Asker eller Bærum kommune er viktige faktorer for å predikere elbil-eiere. Verken menn eller høy inntekt regnes som spesielt viktige faktorer, mens viktigheten av utdanning er tvetydig. Tilhørighet til politiske partier som Miljøpartiet De Grønne og Høyre karaktieriserte tidligere elbil-eiere, men disse faktorene er ikke like prominente i nyere tid. Dette er ett av flere tegn til en homogenisering av elbil-eiernes karakteristika med andre bileiere. Derav resulterer dette i et fall i prediksjonsevnen til modellene som benytter data fra nyere tid, ettersom klassene blir vanskeligere å skille.

# Abstract

In this thesis we study what characterizes electric vehicle (EV) owners and if these factors further can be applied to predict potential EV owners. Earlier studies based on questionnaire surveys have concluded that typical EV owners are relatively young males between 35-54 years of age, with a high level of education and income. This thesis differs from earlier studies by utilizing prediction methods to better define the EV owners' characteristics and rank these by importance. For this purpose, a dataset consisting of auto registrations in Oslo and Akershus county between 2010-2018 is used.

Principal component analysis and correspondence analysis shows that the age of the car owner, the municipality of residence, political affiliation, as well as high income and fortune to be factors of interest for further analysis. In addition, a variable selection through the Boruta algorithm is performed for three different time periods, showing the most important explanatory variables to distinguish EV owners from other car owners. Through the use of tree-based methods in prediction, a classification tree is estimated as a basis for comparison. Additionally, models for random forest and extreme gradient boosting are estimated.

This thesis concludes that the age groups 25-40 years and 40-60 years, share of households with a combined fortune over 4 million Norwegian Krones in a municipality, the number of charging stations in the municipality, and a residence in Asker or Bærum municipality are important factors when predicting EV owners. Neither men or high income are regarded as especially important factors, while the importance of education is ambiguous. Affiliation to political parties like Miljøpartiet De Grønne and Høyre characterized early EV-owners, where these factors are not as prominent in modern times. This is one of many signs of a homogenization of the EV owner's characteristics with other car owners. Thus, this also results in a loss of predictive power when the models apply a dataset from modern times, as the classes are harder to distinguish.

# Innhold

# Figurliste

# Tabelliste

# 1 Innledning

## 1.1 Oppgavens formål og problemstilling

Norge er et relativt lite land både i befolkningstall og innenfor bilindustrien. Likevel, har elbilenes inntog i Norge vært et unikt fenomen i verdenssammenheng. Ved slutten av 2017 kunne Norge skilte med å ha en markedsandel på hele 20,8% for elbiler, størst i hele verden. Sammenlignet med andreplassen Kina, verdens største bilmarked og nest største elbilmarked etter markedsandel, var dette tallet på kun 1,8% (The International Energy Agency, 2018). Det kan dermed argumenteres for at Norge er blant de ledende landene i verden når det kommer til det grønne skiftet innenfor transportsektoren. En kan likevel stille seg undrende til hvorfor og hvordan Norge har inntatt denne posisjonen.

Norge har lenge hatt et fokus på miljøet med en klimapolitikk som inkluderer konkrete tiltak for å redusere nasjonens klimaavtrykk. Dette har blitt formalisert gjennom flere internasjonale klimaavtaler som Rio-, Kyoto- og Paris-avtalen, hvor flere av målene også har blitt lovfestet i klimaloven fra 2018 (Klima- og miljødepartementet, 2017). Dette har satt et stadig økende press på norske politikere, som siden 1990-tallet har innført flere incentiver for å stimulere nordmenn til blant annet elbil-kjøp. I de første årene frem til 1998 var fokuset å tilrettelegge for økt testing av elbiler. Fra og med 2008 har elbiler blitt sett på som et middel for å redusere klimagasser, og siden har elbil-politikken vært en viktig del av norsk klimapolitikk. Det overordnede målet for transportsektoren i dag er å redusere $CO_2$-utslippene, slik at Norge kan nå sine forpliktelser i henhold til Paris-avtalen innen 2030 (Figenbaum, 2018).

En studie viser at de viktigste finansielle incentivene knyttet til elbil-kjøpere var fritak fra merverdiavgiften, engangsavgift og bompenger, i nevnte rekkefølge (The International Energy Agency, 2018). Ettersom merverdiavgiften og engangsavgiften sammen utgjør en betydelig del av nybilprisen i Norge, har dette gjort elbiler særlig attraktive i forhold til andre biler på nybilmarkedet. Fritak fra å betale bompenger er trolig også en viktig årsak til at flere innenfor storbyene i Norge har byttet til elbiler, som betyr en stor utgiftsbesparelse per år. Andre eksempler på incentiver er lav årlig årsavgift, tilgang til kollektivfelt, samt gratis parkering og lading på bestemte steder. Disse har samlet

gjort det betydelig mer gunstig for nordmenn å kjøpe elbiler i forhold til andre biler, sammenlignet med andre land. Likevel, vil mange av disse incentivene etter hvert fases ut ettersom en større andel av befolkningen kjøper elbiler. Myndighetene har imidlertid konstatert at det alltid kommer til å være mer økonomisk å velge nullutslipps-biler over biler med forbrenningsmotor (Figenbaum, 2018). Fremdeles vil denne usikkerheten knyttet til elbil-incentivene kunne påvirke bilvalget for fremtidige bilkjøpere.

Dermed kan det være interessant å studere hvordan disse incentivene har fungert ved å se hvilke målgrupper som har blitt elbil-eiere og hva som kjennetegner de. Tidligere studier viser at den typiske elbil-eier er en relativt ung mann mellom 35-54 år med høyere utdanning og inntekt (Figenbaum og Kolbenstvedt, 2016). Gjennom å bruke statistiske prediksjonsmetoder som klassifiseringstrær, random forests og extreme gradient boosting, ønsker vi å bedre klassifisere elbil-eiere og se hvilke faktorer som er mest distinkte for gruppen. Med hensyn til variabelutvelgelse vil vi benytte Boruta-algoritmen. Videre vil vi sammenligne ulike tidsperioder for å se på utviklingen av de viktigste faktorene over tid, samt hvordan disse kan brukes til å predikere potensielle elbil-eiere. De ulike prediksjonsmodellene vil ha en avhengig variabel som representerer hvorvidt en person er en elbil-eier eller ikke. Vi vil benytte programmeringsspråket *R* (R Core Team, 2017) gjennom det integrerte utviklingsmiljøet *RStudio* for dette formålet.

Oppgavens hovedproblemstilling er: ***Hvilke faktorer karakteriserer dagens elbil-eiere, og hvordan kan disse videre brukes til å predikere potensielle elbil-eiere?***

## 1.2 Oppgavens struktur

Denne utredningen består av totalt 8 kapitler. I kapittel 2 gjennomgås bakgrunnen for oppgaven. I kapittel 3 presenteres datagrunnlaget oppgaven er basert på, samt avgrensninger i datasettet og beskrivelser av de ulike variablene. Kapittel 4 inneholder en detaljert deskriptiv analyse av datasettet. I Kapittel 5 presenteres metoder for variabelutvelgelse, statistiske prediksjonsmetoder for estimering, samt validering av modellene. Kapittel 6 presenterer analysene og resultatene knyttet til datagrunnlaget, samt en diskusjon rundt de viktigste funnene. Kapittel 7 inneholder videre diskusjon og tolkning av resultatene. Det siste kapittelet, kapittel 8, konkluderer og svarer på problemstillingen.

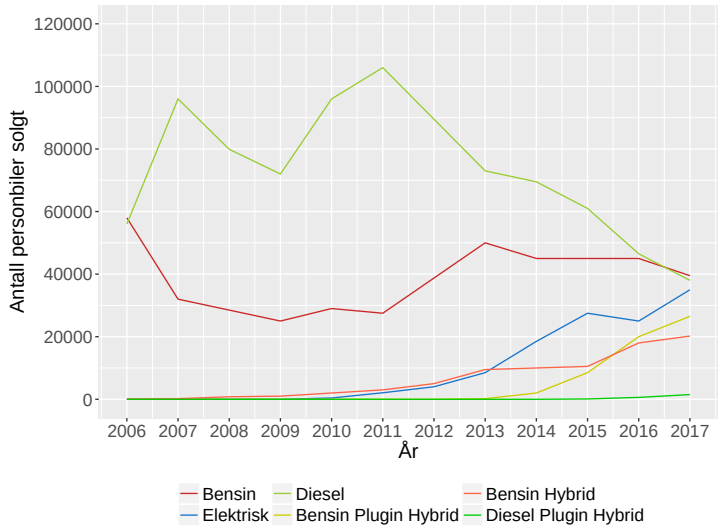
# 2 Bakgrunn

## 2.1 Det norske bilmarkedet

Det norske bilmarkedet har lenge opplevd en positiv trend med tanke på antall solgte biler. Det ble i 2017 registrert rekordmange nye personbiler som slo den tidligere rekorden fra 1987. Både elektriske biler og andre lavutslippsbiler[1] har entret markedet og har oppnådd solide markedsandeler[2] på kort tid. Fra å ha 0% markedsandel i 2010 har elbiler og ladbare hybridbiler oppnådd en total markedsandel på henholdsvis 20,8% og 18,4% i 2017. Motsatt hadde diesel- og bensinbiler 47,8% markedsandel i 2017, sammenlignet med nærmere 99% i 2010. Det har med andre ord vært en voldsom omveltning i det norske bilmarkedet de siste årene (Opplysningsrådet for veitrafikken, 2018b). Figur 2.1 viser hvordan salget av personbiler har utviklet seg siden 2006 etter drivstoffkategori. Her kan en tydelig se at lavutslippsbiler har overtatt store deler av markedet.

Figuren viser også hvordan diverse politiske tiltak har påvirket salget. I statsbudsjettet for 2007 var avgiftene for dieselbiler senket som et miljøtiltak, begrunnet med at dieselbiler hadde et lavere $CO_2$-utslipp enn bensinbiler (Finansdepartementet, 2007). Dette forårsaket

**Figur 2.1:** Salg av personbiler etter drivstoff

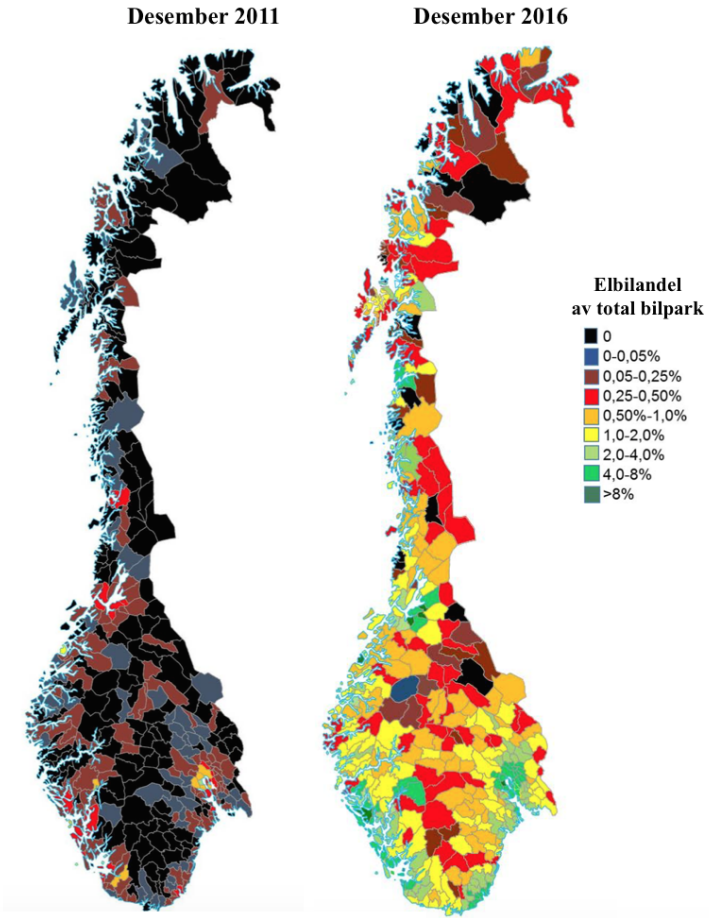



Kilde: Opplysningsrådet for veitrafikken (2018b)

[1] Lavutslippsbiler defineres videre som el-,hydrogen- og hybrid-biler.
[2] Markedsandeler i denne utredningen refererer til andel salg av nye biler og bruktimport.

et sterkt salg av dieselbiler i årene som fulgte, hvor dieselbiler utgjorde rekordhøye 75,7% av nybilsalget i 2011, mot 48,3% i 2006 (Opplysningsrådet for veitrafikken, 2013). Likevel gjorde regjeringen en helomvending ett år senere da det britiske miljødepartementet konkluderte med at dieselbiler hadde et høyere NO2 utslipp enn andre biler (Gjerde, 2008). Gradvis fra 2009 ble avgiftene for dieselbiler skjerpet og bensinbilene lettet (Nervik og Larsen-Vonstett, 2012). Dette resulterte i en sterk nedgang i salget av dieselbiler fra og med 2012, hvor flere begynte å kjøpe en relativt større andel bensinbiler, men også hybrid- og elbiler (Opplysningsrådet for veitrafikken, 2013).

**Figur 2.2:** Elbilandel per kommune




Kilde: Figenbaum (2018)

Det har siden starten av elbilens moderne utvikling blitt satset stort på denne typen biler i Norge. Fremdriftssystemet til den første elektriske bilen Volkswagen produserte, Golf Citystromer Electric, ble utviklet av det norske selskapet ABB Battery Drives i 1989. Utover 90-tallet ble de to norske elbilprodusentene Think og Kewet etablert. Think ble senere kjøpt opp av Ford og produserte elektriske biler på norsk jord frem til de gikk konkurs i 2011. Kewet går i dag under navnet Buddy Electric AS og produserer kun en

bilmodell, Buddy, samt elsykler. Etter 2008 ble flere incentivordninger introdusert av myndighetene samtidig som teknologien knyttet til elbiler ble kraftig forbedret. Rekkevidde, komfort, sikkerhet og kjøreegenskaper utviklet seg opp mot et nivå med andre biltyper, som gjorde elbilene mer attraktive for nye bilkjøpere. Flere kjente bilmerker begynte å komme på markedet med elektriske biler som Tesla Model S, Nissan Leaf, Renault Zoe og Volkswagen e-Golf. Kombinert førte de nevnte faktorene til den eksponentielle veksten i elbilsalget observert i Norge fra rundt 2010. Som en kan se fra figur 2.2 økte andelen elbiler som prosent av den totale bilparken betraktelig i nesten samtlige kommuner i hele landet fra 2011 til 2016. Figuren viser også at det er flest elektriske biler rundt de store byene, spesielt i området Oslo og Akershus. Det kan også bemerkes at Finnmark hadde en høyere elbilandel enn de fleste andre Europeiske land i slutten av 2016 (Figenbaum og Kolbenstvedt, 2013; Figenbaum, 2018).

## 2.2 Klimapolitikk og elbilincentiver

I dag er det en bred enighet blant aktive klimaforskere om at menneskelige utslipp av klimagasser har forårsaket trendene i global oppvarming det siste århundret med veldig stor sannsynlighet (NASA, 2018). Norge har som et av flere ledd i deres internasjonale klimapolitikk ratifisert ulike klimaavtaler, slik som Rio-avtalen fra 1992, Kyoto-avtalen fra 1997 og Paris-avtalen fra 2015. I tillegg lovfestet Norge en klimalov i 2018, som definerer tydelige klimamål for landet frem til 2030 og 2050. Dette innebærer blant annet at utslipp av klimagasser i 2030 reduseres med minst 40% fra 1990-nivået, og at Norge skal bli et lavutslippssamfunn innen 2050 (Klima- og miljødepartementet, 2017).

Transport er den største utslippskilden til klimagasser i Norge. Mellom 1990-2016 har utslippene fra denne sektoren økt med 24%, hvorav veitrafikken står for mer enn halvparten av økningen. Dette utgjør totalt 31% av de totale klimagassutslippene (Miljødirektoratet, 2018). Likevel, har Norge lenge vært en pådriver for en grønnere og mer miljøvennlig bilpark. Det fremkommer i klimaforliket fra 2012 at gjennomsnittsutslippet fra nye personbiler i 2020 ikke skal overstige 85 gram $CO_2/km$. Dette skal oppnås blant annet gjennom økt utbygging av infrastruktur for elektrifisering[3] og fortsatt være i front når det gjelder tilrettelegging i bruk av el- og hydrogenbiler (Miljøverndepartementet, 2012). En bil med

[3]Med elektrifisering menes å øke andelen batteridrevne biler i bilparken.

en helelektrisk drivlinje slipper ikke ut noen klimagasser eller andre forurensede avgasser og er opp til tre ganger så energieffektiv som biler med forbrenningsmotorer. Til tross for at slike biler også bidrar til svevestøv fra vei og dekk, er støynivået ved lavere hastigheter og lite trafikk også betydelig lavere (Hagman og Kolbenstvedt, 2018).

Som et resultat av dette, har det siden 1990 blitt introdusert en rekke incentiver for å få flere til å kjøre elektrisk. For en fullstendig oversikt over nåværende incentiver, se tabell 2.1

**Tabell 2.1:** Elbilincentiver i Norge (2018)

| Incentiver | Introduksjonsår | Fremtidige planer |
|---|---|---|
| **Fiansielle incentiver:** | | |
| Fritak fra engangsavgift | 1990/1996 | Fortsetter frem til 2020 |
| Fritak fra merverdiavgift | 2001 | Fortsetter frem til 2020 |
| Redusert årsavgift | 1996/2004 | Fortsetter til ubestemt tid |
| Fritak fra omregistreringsavgift | 2018 | Nylig introdusert |
| **Direkte subsidier til eiere:** | | |
| Gratis bompasseringer | 1997 | Loven revidert slik at takstene for elbiler ved bomveier og ferjer blir bestemt av kommunestyret, opp til en maksimal takst på 50% av biler med forbrenningsmotor |
| Reduserte ferjetakster | 2009 | |
| Finansiell støtte for vanlige ladestasjoner | 2009 | En nasjonal plan for en ladeinfrastruktur skal bli utviklet |
| Finansiell støtte for hurtigladestasjoner | 2011 | ENOVA-støttet program for å opprette hurtigladestasjoner ved store transportkorridorer. Hurtigladestasjoner i byene er overlatt til private aktører |
| **Brukerprivilegier:** | | |
| Tilgang til kollektivfelt | 2003/2005 | Kommunestyret har autoriteten til å introdusere restriksjoner dersom elbilene skaper forsinkelser i kollektivtrafikken |
| Gratis parkering | 1999 | Kommunestyret har autoriteten til å bruke en takst på 50% av biler med forbrenningsmotor |
| Gratis lading (bestemte steder) | | Kommunestyret og parkeringsselskapene bestemmer hvorvidt incentivet vil fortsette |

Kilde: Figenbaum (2018)

Engangsavgiften regnes ut basert på $CO_2$- og $NOX$-utslipp, samt bilens vekt. Dette kan eksempelvis tilsvare en avgift på $60000 - 90000$ kr ved kjøp av en ordinær Volkswagen Golf. Derimot, vil en kjøper av en e-Golf være fritatt for denne avgiften. Merverdiavgiften er 25% av salgsprisen trukket fra engangsavgiften, som også er fritatt for elbiler. Sammen utgjør disse fritakene en betydelig reduksjon av elbilenes salgspris. Årsavgiften til en elbil er på 455 kr, sammenlignet med rundt $2800 - 3300$ kr for en bil med forbrenningsmotor (Skatteetaten, 2018b). Omregistreringsavgiften er basert på bilens alder og vekt, og kan tilsvare rundt $1600 - 6100$ kr for biler med forbrenningsmotor, som elbil-kjøpere er fritatt

for ved bruktbilkjøp (Skatteetaten, 2018a). Gratis passeringer i bomstasjonene kan for en gjennomsnittlig Oslo-borger tilsvare en årlig besparelse på $6000-10000$ kr, hvor enkelte kan spare opp til 25000 kr i året. Elbil-eiere vil også spare penger ved bruk av ferjer på samme måte som med bomstasjoner. Finansiell støtte til ladestasjoner og hurtigladestasjoner kan bidra til å redusere risiko for investorer, redusere rekkeviddeangst og bidra til økt bruk av elbiler. Elbil-eiere vil også spare tid på å komme seg til og fra jobb ved å kunne benytte kollektivfeltet i rushtiden. Gratis parkering er også noe som blir stadig vanskeligere å finne i større byer, hvor elbil-eiere både sparer tid og penger på allokerte plasser. Gratis lading er ikke lovfestet, men er ofte kombinert med gratis parkeringsplasser. Det må også nevnes at elbiler hadde en økt skattefordel som firmabiler inntil det ble fjernet i 2018 (Figenbaum, 2018).

## 2.3 Tidligere arbeid

I nyere tid har det stadig blitt større interesse rundt elbiler grunnet rask teknologisk utvikling, introduksjon av incentivordninger og større klimautfordringer blant annet forårsaket av utslipp fra vanlige personbiler. Denne økte interessen har ført til flere studier som omhandler privatpersoners kjøpsintensjoner av elbiler og hva som kjennetegner elbil-eiere. I Østerrike gjennomførte Priessner et al. (2018) en studie for å kartlegge de østerrikske bilkjøpernes behov og for å videre utforme effektive incentivordninger som skulle stimulere til elbil-kjøp. De fant at flere psykologiske og sosiodemografiske faktorer som alder, kjønn og inntekt spilte en rolle, men kunne ikke spesifisere hvilke faktorer som var av størst viktighet. De poengterer likevel at dagens elbil-eiere ikke lenger kun er individer med høy inntekt som ønsker å minimere sine karbonfotavtrykk, men at gruppen har utviklet seg til å bli mer mangfoldig. Ng et al. (2018) har sett på markedet og kjøpsintensjoner for elbiler i Hong Kong. De finner ved hjelp empiriske analyser at blant annet oppfattet verdi av elektriske biler og tillit til teknologien er viktige faktorer. Felles for studiene nevnt over er at resultatene baseres på spørreundersøkelser med rundt 1000 respondenter.

Flere studier er også blitt gjennomført for det norske markedet av eksempelvis Norsk Elbilforening og Transportøkonomisk institutt (TØI). Norsk Elbilforening utfører hvert år en stor spørreundersøkelse blant norske elbil-eiere og har gjennom sine studier funnet at de fleste kjøperne skaffer seg en elbil som bil nummer to. Likevel erstatter elbilen

gjennomsnittlig 82% av transporten som tidligere ble gjennomført med fossilt brennstoff. De finner også at størsteparten av elbil-eiere er menn mellom 30 og 50 år, samt at 75% har utdanning fra høyskole eller universitet. Over halvparten av respondentene i undersøkelsen deres fra 2015 svarer at de økonomiske fordelene ved å eie en elbil var den viktigste faktoren for kjøp (Haugneland et al., 2016).

Figenbaum og Kolbenstvedt (2016) og Figenbaum (2018) er eksempler på lignende studier gjennomført av TØI. På samme måte som tidligere studier har det blitt gjennomført en spørreundersøkelse, her med over 8000 respondenter. Begge studiene finner resultater som i stor grad samsvarer med Haugneland et al. (2016). I tillegg, ble det observert at elbil-eiere ofte bor i husstander med barn, har høyere sysselsettingsgrad og lenger vei til arbeidsplassen enn den gjennomsnittlige bileier. Det kan også bemerkes at 89% av respondentene i undersøkelsen som allerede eide en elbil, nevnte økonomiske besparelser, miljøhensyn, teknologisk fremtidssikring og gratis bomveier som de viktigste grunnene til å kjøpe en elbil igjen.

Ingen av de tidligere studiene har rangert hvilke faktorer som er viktigst for å identifisere elbil-eiere. Heller ingen av de eksisterende studiene har prøvd å identifisere potensielle elbil-eiere ved hjelp av prediktive maskinlæringsmetoder. Det kan også nevnes at samtlige studier har basert seg på spørreundersøkelser med et begrenset antall respondenter. Vi ønsker imidlertid å basere oss på et mye større datagrunnlag av faktiske observasjoner i markedet. Denne utredningen vil potensielt komme med ny innsikt til område ved å ta i bruk metoder som er uprøvd innen tematikken. Metodene vi bruker vil også være i stand til å rangere viktigheten av forskjellige faktorer.

# 3 Datasett

Datagrunnlaget brukt i denne utredningen er hentet fra flere kilder. Hoveddelen er utsendt av Statens Vegvesen og er utlevert i henhold til § 9 i Offentleglova. Andre datakilder har vært Statistisk Sentralbyrå, Norsk elbilforening og Valgdirektoratet. I dette kapittelet vil vi gjennomgå avgrensninger, sammensetting av data og forklaring av variabler.

## 3.1 Datakilder

### 3.1.1 Statens Vegvesen

Det mest omfattende datasettet som brukes kommer fra Statens Vegvesen. Datasettet inneholder informasjon over alle registrerte motorvogner og deres nåværende eiere i Oslo og Akershus, som ble førstegangsregistrert mellom 2. januar 2010 og 11. september 2018. Hver enkelt rad inneholder informasjon om en spesifikk motorvogn og diverse detaljer knyttet til bilen og dens eier. Dette datasettet er senere blitt utvidet med informasjon fra flere kilder. Utvidelsene vil bli gjennomgått nærmere i kapittel 3.3.2.

### 3.1.2 Statistisk Sentralbyrå

Fra statistikkbanken til Statisktisk Sentralbyrå (SSB) ble det hentet ut informasjon om personers inntekt, gjeld og formue. Disse verdiene er gitt som prosentandel av befolkningen som tilhører diverse inntekts-, gjelds- eller formueklasser i en bestemt kommune innenfor Oslo og Akershus. Det har også blitt innhentet informasjon om utdanning, som er presentert på samme måte som den ovennevnte informasjonen med utdanningsklasser for grunnskole, videregående skole, fagskole og universitet.

### 3.1.3 Norsk elbilforening

Norsk elbilforening har siden 2010 driftet og oppdatert NOBIL, en database over alle ladestasjoner i Norge (Norsk Elbilforening, 2018). Gjennom tilgang til denne databasen

fikk vi informasjon om blant annet ladepunktenes lokasjon, antall, aktiveringstidpunkt og status for hele landet.

### 3.1.4 Valgdirektoratet

Valgdirektoratet ble opprettet 1. Januar 2016 og sørger i dag for gjennomføringen av alle valg på landsbasis, samt lagringen av detaljert informasjon om tidligere valgresultater (Valgdirektoratet, 2018). Fra deres nettsider ble det uthentet tall fra Stortingsvalg i de tidsperioder og områder informasjonen var ønsket.

### 3.1.5 Øvrige data

I tillegg til de nevnte datakildene ble det innhentet informasjon om hvilket parti ordføreren i hver kommune tilhørte for de ulike tidsperiodene. Denne informasjonen ble manuelt hentet fra kommunenes egne hjemmesider.

## 3.2 Avgrensninger i datagrunnlaget

For å gjøre oppgaven så spisset og relevant som mulig har det vært nødvendig å gjøre visse avgrensninger i datagrunnlaget. Avgrensingene er gjort for å skille ut observasjoner som er ufullstendige eller kan føre til støy i videre analyser.

Utredningen vil fokusere på fylkene Oslo og Akershus. Dette på grunnlag av hva Statens vegvesen hadde mulighet til å gi oss av data, og at disse er de mest befolkede fylkene i Norge. Det er også de områdene i Norge hvor det selges flest elektriske biler og hvor det er flest ladestasjoner. Dette belyses videre i kapittel 4 ved hjelp av data supplert av Opplysningsrådet for veitrafikken (OFV) og Norsk elbilforening.

Først og fremst ønsker vi å identifisere elbil-eiernes karakteristika. Vi har derfor avgrenset datasettet til kun å inneholde observasjoner innenfor kjøretøygruppene "Personbil" og "Varebil klasse 2". Dette fordi det er tilnærmet ingen elektriske alternativer i de andre kjøretøygruppene. Denne avgrensningen vil føre til at alle observasjonene vi ser på har et reelt elektrisk alternativ. Videre er mange av datapunktene knyttet til næringsliv. Ettersom

vi skal gjennomføre en studie som omhandler privatpersoner, fjernes de observasjonene dette gjelder fra datasettet.

Observasjoner som ikke hadde noen verdi for variabelen postnummer viste seg å tilhøre avdøde personer eller personer med fortrolig adresse, og ble derfor fjernet fra datasettet for å unngå støy i videre analyse. Det samme ble observasjoner med "*?*" under drivstoffkategori eller hybridtype, fordi de ble ansett som observasjoner med usikker informasjon. Drivstoffkategoriene *Parafin*, *Gass* og *Hydrogen* inneholdt veldig få observasjoner og ble derfor fjernet fra datasettet grunnet problemer med modellestimeringene. Dette var også tilfellet for enkelte poststeder med svært få observasjoner.

## 3.3 Forklarende variabler

I dette avsnittet vil forklaringsvariablene i datagrunnlaget gjennomgås. Dette vil innebære variabler som er observasjonsspesifikke, samt variabler som er avhengige av både kommune og registreringstidspunkt.

### 3.3.1 Statens vegvesen

Datasettet inneholder en del variabler som ikke er av interesse for utredningen. Dette er variabler som en i prediksjonsmodellene ikke har kjennskap til før etter en bil er kjøpt, eksempelvis bilmerke, modell og farge. I det videre vil derfor kun de variablene som er av interesse for videre analyse forklares.

**Pnr og Poststed**
En identifikator på henholdsvis postnummeret hvor bilen er registrert og navnet på poststedet. Pnr blir behandlet som en nominell kategorisk variabel da størrelsen på Pnr i seg selv ikke vil ha noen betydning for prediksjoner og andre statistiske tester.

**Fødselsdato**
Bileierens fødselsdato. Verdien brukes til å regne ut eierens alder på det tidspunktet bilregistreringen fant sted.

**Komnr. og Kommune**
Viser kommunenummeret til kommunen hvor bileieren er registrert og kommunens navn.

**Kjtgrp**

Variabelen viser kjøretøygruppen motorvognen tilhører, *personbil* eller *varebil klasse 2*.

**Dr.st.**

Angir drivstoffkategorien til kjøretøyet. Variabelen har kategoriene *Bensin*, *Elektrisk*, *Diesel* og *Hybrid*.

**PLG.NPLG.**

Om motorvognen har drivstoffkategori hybrid vil denne variabelen vise om motorvognen er av type *ladbar hybrid* (PLGIN) eller *hybrid* (NOPLG). PLGIN biler kan kobles direkte til strøm for å få ladet batteriet, mens NOPLG biler lader batteriet gjennom en forbrenningsmotor eller ved regenererende bremsing. For å minimere antall variabler ble kategorien *Hybrid* under "Dr.st." byttet ut med henholdsvis NOPLG og PLGIN.

**Regdato**

Datoen motorvognen ble registrert hos nåværende eier.

**Kjønn**

Bileierens kjønn.

**Alder**

Alderen til bileieren da kjøretøyet ble registrert på personen. Dette er en kontinuerlig variabel. Senere vil denne bli delt opp i aldersgrupper og brukt som den ordinale kategoriske variabelen "Aldersgruppe".

**Reg.1.g**

Variabelen forteller når kjøretøyet ble førstegangsregistrert i Norge. Det må presiseres at variabelen ikke representerer antall salg av kjøretøy i Oslo og Akershus i en gitt periode. Ettersom datasettet viser de kjøretøyene som er registrert i Oslo og Akershus i dag etter eiernes bostedsadresse, kan det være et frafall/tilsig av kjøretøy ettersom både eiere og kjøretøy flytter til/fra andre fylker. Det er også mulig at biler som ble førstegangsregistrert i en gitt periode nå er avskiltet og skrotet og derfor ikke er en del av datasettet. Dermed vil ikke antall førstegangsregistrerte biler innenfor Oslo og Akershus innenfor en bestemt periode tilsvare det offisielle salgstallet for de to fylkene i perioden.

Tabell 3.1 oppsummerer variablene og variabeltypene i datasettet fra Statens Vegvesen. Tabellen viser at det kun er én numerisk variabel og en overvekt av kategoriske variabler.

**Tabell 3.1:** Variabler i datasettet fra Statens vegvesen

| Variabelnavn | Variabeltype |
|---|---|
| Pnr | Kategorisk |
| Poststed | Kategorisk |
| Fødselsdato | Tid |
| Komnr. | Kategorisk |
| Kommune | Kategorisk |
| Kjtgrp | Kategorisk |
| Dr.st. | Kategorisk |
| PLG.NPLG | Kategorisk |
| Regdato | Tid |
| Kjønn | Kategorisk |
| Alder | Numerisk |
| Reg.1.g | Tid |

Merk: Alle kategoriske variabler i denne tabellen er nominelle.

### 3.3.2 Øvrige variabler

Det er hentet inn en del informasjon fra andre datakilder også, som forklart i kapittel 3.1. Variablene hentet fra disse datakildene vil bli presentert i det videre.

**Utdanning**

Utdanningsinformasjon ble hentet fra statistikkbanken til SSB. Variablene under utdanning er presentert som prosentandeler av befolkningen innenfor hver kommune som kun har fullført grunnskole, videregående, kort høyere utdanning (opp til 4 år) eller lang høyere utdanning (4 år og mer samt forskerutdanning). Fagskole var til og med 2015 en del av kategorien videregående, men ble skilt ut i egen kategori fra og med 2016 (Statistisk Sentralbyrå, 2018e). For å forhindre et dropp i videregående i 2016, og at det er nullverdier for fagskole frem til 2016, ble de to kategoriene satt sammen og fagskole fjernet. Vi har kun data for utdanning frem til og med 2017. Informasjonen ble lagt til det eksisterende datasettet med hensyn til år, kjønn og kommune.

**Ladestasjoner**

Informasjon om det norske ladestatsjonnettverket ble supplert av Norsk Elbilforening. Denne informasjonen gjorde at vi kunne beregne antall ladestasjoner registrert på postnummeret til bileierens bostedsadressen ved registrering hos Statens Vegvesen. Vi kunne også kalkulere antall ladestasjoner per kommune når nye biler ble registrert. Disse to variablene ble lagt til det eksisterende datasettet.

**Inntekt**

Inntektsinformasjonen er innhentet på tilsvarende måte som utdanning, fra statistikkbanken til SSB (Statistisk Sentralbyrå, 2018d). Variablene er presentert som prosentandeler av befolkningen i hver kommune som har samlet nominell inntekt per husstand under 150', 150'-250', 250'-350', 350'-450', 450'-550', 550'-750' og over 750'. Variablene er basert på år og dataen strekker seg fra 2010 til og med 2016. Tallene for 2017 publiseres for sent til at vi kan bruke dem i vår analyse. Vi kunne derimot se at utviklingen i inntekt var tilnærmet lineær og vi brukte derfor lineær regresjon til å estimere tallene for 2017. De estimerte verdiene for 2017 og de reelle verdiene ble lagt til det eksisterende datasettet med hensyn til år og kommune.

Vi gjennomførte denne estimeringen fordi 21% av datasettet er registrert i 2017. Uten estimeringen ville det ikke vært mulig å gjennomføre en del analyser for observasjoner registrert i 2017, som hadde betydd et tap av verdifullt datagrunnlag. En slik estimering vil føre til at variablene får mindre variasjon enn naturlig, altså vil det oppstå en målefeil. Dette gjør at en må være varsom ved videre tolkning av resultater hvor variablene er brukt. Det finnes også andre estimeringsmetoder som kunne ført til mindre avvik i variasjon, som bootstrapping. Likevel, ville metodene fremdeles ført til målefeil og vi har for enkelthets skyld derfor valgt å benytte lineær regresjon.

**Formue og gjeld**

Informasjonen knyttet til disse kategoriene av variabler ble også innhentet fra SSBs statistikkbank (Statistisk Sentralbyrå, 2018a,b,c). Formue og gjeld er presentert som prosentandel på samme vis som inntekt, på kommunenivå. Det er totalt syv kategorier for husstandenes samlede nominelle formue: Under 250', 250'-500', 500'-1000', 1000'-2000', 2000'-3000', 3000'-4000' og over 4000'. Gjeld består av totalt fem kategorier; Ingen gjeld, gjeld mindre eller lik årlig inntekt i husstanden, mellom 1 og 2 ganger årlig inntekt, 2-3 ganger årlig inntekt, 3-4 ganger årlig inntekt og over 4 ganger årlig inntekt. For året 2010 manglet informasjon om formue, og for året 2017 manglet informasjon om både formue og gjeld. I begge kategoriene kunne vi se at utviklingen var tilnærmet lineær. Derfor ble metoden beskrevet under inntekt også her brukt til å beregne verdier for de manglende årene.

**Befolkning**

Befolkning er presentert som numeriske verdier. Befolkningsvariabelen viser befolkningen i kommunene per 1. januar hvert år. Alle de fire kategoriene varierer per år og kommune og ble lagt til det eksisterende datasettet med hensyn til dette.

**Ordfører**

Viser hvilket parti den sittende ordføreren i kommunen tilhørte da det gitte kjøretøyet ble registrert.

**Stortingsvalg**

Informasjon om alle stortingsvalg som var relevant for vår tidsperiode (Valgdirektoratet, 2018). Dette vil si stortingsvalgene i 2009, 2013 og 2017. De tallene som brukes representerer oppslutningen de partiene som i dag sitter på stortinget fikk i hver kommune under de respektive valgene. Dagens stortingspartier ble brukt fordi disse i lang tid har vært de mest fremtredende politiske partiene i Norge. Dermed presenteres ni nye variabler: A (Arbeiderpartiet), SV (Sosialistisk venstrepart), RØDT (Rødt), SP (Senterpartiet), KRF (Kristelig Folkeparti), V (Venstre), H (Høyre), FRP (Fremskrittspartiet) og MDG (Miljøpartiet De Grønne). Informasjonen ble lagt til det eksisterende datasettet med hensyn på hvilket gjennomførte valg som var nærmest i tid. Det betyr eksempelvis at alle observasjoner registrert mellom 2010 og 2012 fikk verdier fra Stortingsvalget i 2009.

En oppsummerende tabell over de nye øvrige variablene finnes i appendiks A1.

## 3.4 Avhengig variabel

I denne utredningen definerer vi variabelen "Elektrisk" som avhengig variabel. Dette vil være en dummyvariabel som er *1* eller *Ja* om det registrerte kjøretøyet er et fullelektrisk kjøretøy og *0* eller *Nei* hvis det registrerte kjøretøyet ikke er fullelektrisk.

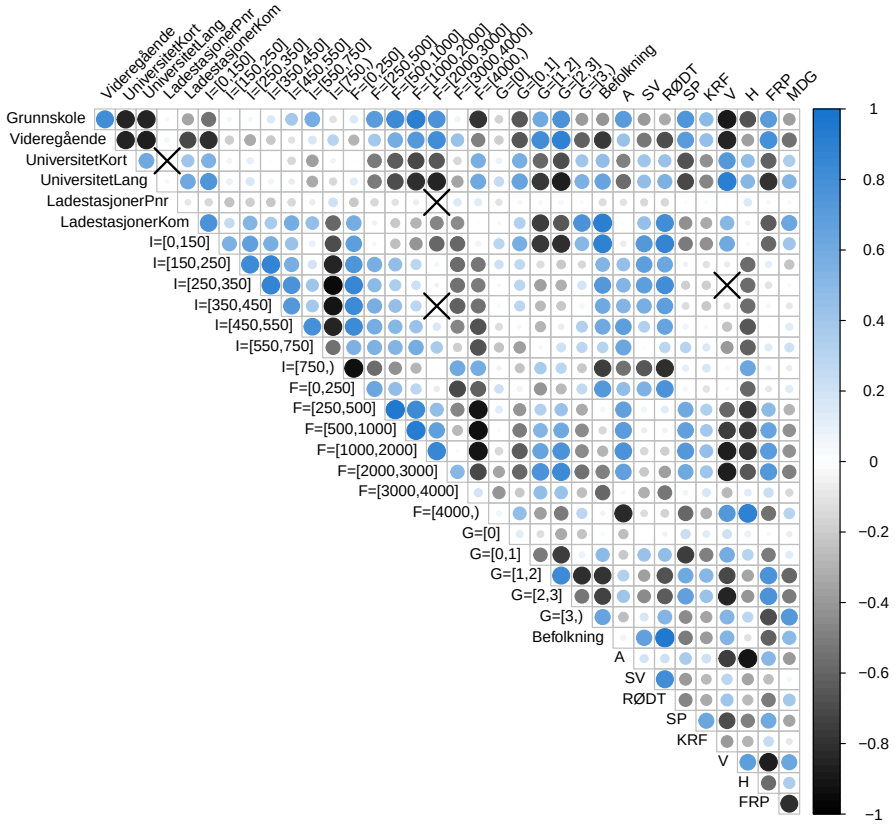
## 3.5 Klargjøring av datasett

I etterkant av at datasettet var satt sammen, ble det laget en korrelasjonsmatrise med Pearson korrelasjoner for å studere det lineære forholdet mellom de numeriske variablene. En Pearson korrelasjon gir en verdi mellom -1 og 1, som indikerer graden av

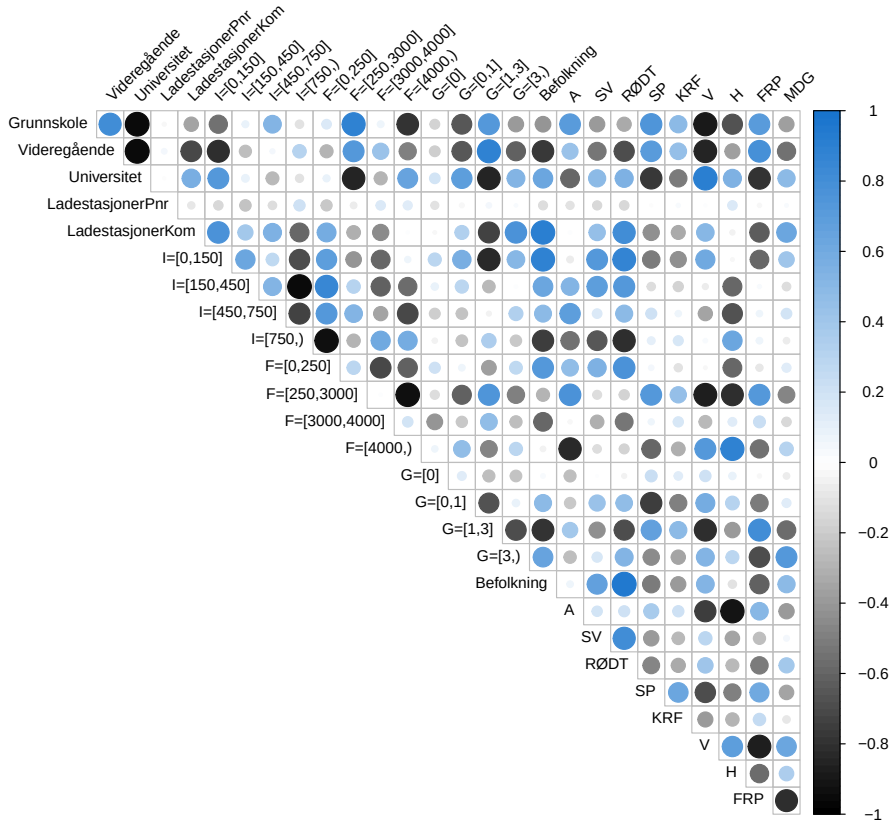
henholdsvis negativ og positiv korrelasjon mellom to variabler (Pearson, 1931). Figur 3.1 viser korrelasjonsplottet med alle de numeriske variablene datasettet har blitt utvidet med. Alle variabler utenom "LadestasjonerPnr", "LadestasjonerKom" og "Befolkning" er målt i samme skala, som prosentandel i en bestemt kommune til et bestemt år. Det fremkommer av korrelasjonsplottet at flere forklaringsvariabler er svært korrelerte med hverandre, spesielt variabler innenfor utdanning, inntekt, formue og gjeld. Ettersom disse variablene gjengir mye av den samme informasjonen, vil det kunne være hensiktsmessig å slå flere av dem sammen. Dette for å unngå eventuelle problemer med *overtilpasning*, som betyr at en eventuell prediksjonsmodell følger feilraten eller støyen i datasettet for tett. Jo høyere raten av parametre $p$ er i forhold til antall observasjoner $n$, vil overtilpasning spille en større rolle (James et al., 2013). Dermed slås mange av disse variablene innenfor utdanning, inntekt, formue og gjeld sammen, slik at de inkluderer flere nivåer i samme variabel. Det må presiseres at variabelsammenslåingen i dette tilfellet tilsvarer mer en kategorisammenslåing innenfor samme variabel. Eksempelvis slår vi sammen inntektsnivåene "I=[150,250]", "I=[250,350]" og "I=[350,450]" til "I=[150,450]". Resultatet av variabelsammenslåingen vises i det modifiserte korrelasjonsplottet i figur 3.2. En oversikt over de øvrige variablene etter sammenslåingen finnes også i tabell A1.2 i appendiks.

I tillegg er det gjennomført en signifikanstest av korrelasjonskoeffisienten, med et signifikansnivå på én prosent i samme figur. Ettersom Pearson korrelasjonen er et mål på styrken av forholdet mellom to variabler, er det også interessant å måle signifikansen av dette forholdet. En lav p-verdi tilsier at korrelasjonen er statistisk signifikant. Eksempelvis er variablene "I=[350,450]" og "F=[2000,3000]" merket med et kryss, som tilsier en p-verdi høyere enn signifikansnivået. Dermed er korrelasjonen mellon disse variablene ikke statistisk signifikant fra null (Pearson, 1931). Sammenlignes figur 3.1 og 3.2 kan en se at samtlige variabler er statistisk signifikante etter variabelsammenslåingen.

**Figur 3.1:** Korrelasjonsplot før variabelsammenslåing




**Figur 3.2:** Korrelasjonsplot etter variabelsammenslåing



## 3.6 Datakvalitet

En utfordring med datasettet er at det ikke er satt sammen av salgsdata, men registreringsdata. Optimalt sett ville en hatt data på hvert solgte kjøretøy i perioden som undersøkes. Videre mangler datasettet en del observasjoner. Dette grunnet at kjøretøy kjøpt i for eksempel 2010 kan ha blitt kondemnert, eksportert, solgt videre til andre eiere i andre områder eller av andre grunner ikke lenger er registrert i Oslo og Akershus. Dermed brukes registreringsdataene som proxy for faktiske salg i denne utredningen.

I dag er privatleasing blitt en stor del av det totale markedet for personbiler. 52,3% av alle nye personbiler ble i 2017 registrert på næringsdrivende. Økningen er på 9,7% fra 2015 og skyldes stort sett veksten i privatleasing (Opplysningsrådet for veitrafikken, 2016, 2018a). Ved leasing vil leaseren ha full bruksrett til kjøretøyet, dermed kan en sammenligne dette med å kjøpe et kjøretøy. Om en person leaser en bil er det leasingselskapet som står oppført som eier av bilen i Statens Vegvesens register. Ettersom denne utredningen kun undersøker kjøretøy registrert på individer mistes alle observasjoner knyttet til leasing. Datagrunnlaget har ingen indikator på hva biler kjøpt av næringsdrivende skal brukes til. Det er derfor ikke mulig til å identifisere biler forbeholdt leasing.

# 4 Deskriptiv statistikk og forklarende dataanalyse

Før det utarbeides prediksjonsmodeller vil det kunne være hensiktsmessig å utforske datasettet gjennom deskriptiv statistikk. Dette er en tilnærming brukt for å presentere store mengder kvantitativ data på en form som er forståelig for leseren. Derav er målet med deskriptiv statistikk å redusere datasettet til små oppsummeringer som kan visualiseres gjennom eksempelvis grafer og tabeller (Trochim et al., 2015).

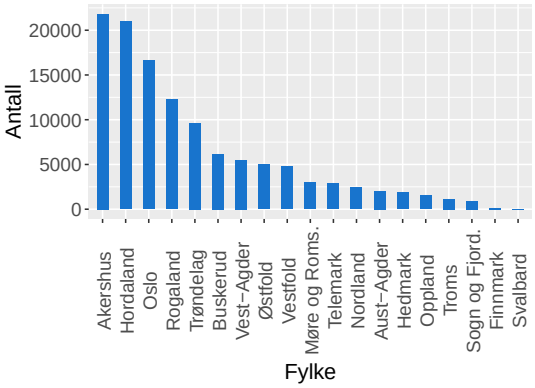
Videre vil det anvendes "Explanatory Data Analysis" eller forklarende dataanalyse for å øke innsikt i datasettet og utforske underliggende strukturer. Dette er en tilnærming som søker å finne sammenhenger, avvik, teste hypoteser og sjekke antagelser gjennom statistiske tabeller og grafer (Natrella, 2010).

## 4.1 Deskriptiv statistikk

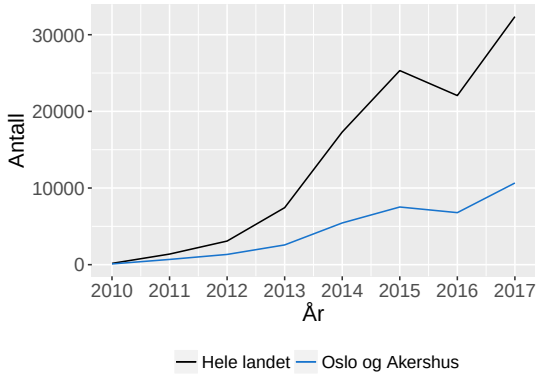
I tillegg til datakildene presentert i kapittel 3 ble vi supplert med salgsdata på alle elektriske kjøretøy solgt i perioden januar 2010 til april 2018 i hele Norge av OFV. Datasettet inneholdt anonymisert informasjon om individer og næringsdrivende. Vi ekskluderte observasjonene knyttet til næringsliv på samme grunnlag som forklart i kapittel 3.2. Figur 4.1 viser hvordan salget av elektriske kjøretøy har vært i den gitte perioden for hvert fylke. Her kan en se at Akershus, Hordaland og Oslo er de tre fylkene med høyest elbilsalg, som gjør disse til spesielt interessante områder å studere i forhold til elbilkjøp. Figur 4.2 illustrerer utviklingen elbilsalget har hatt de siste årene. I 2010 ble det totalt solgt 169 elbiler, hvor det til sammenligning ble solgt hele 32 359 elektriske kjøretøy til privatpersoner i 2017 som utgjorde 20,8% av personbilsalget (Opplysningsrådet for veitrafikken, 2018a). Til sammenligning var 0,6% av nye biler solgt i USA elektriske, 0,1% i Australia og 1,34% i Sverige det samme året (The International Energy Agency, 2018).

Figur 4.3 viser utviklingen i antall ladestasjoner. Her kan en se at Oslo, Akershus og Hordaland går frem som de fylkene med flest ladestasjoner. Innenfor Akerhus er det to kommuner som skiller seg ut, Ullensaker og Bærum. Det store antallet ladestasjoner i

**Figur 4.1:** Antall solgte elektriske kjøretøy per fylke

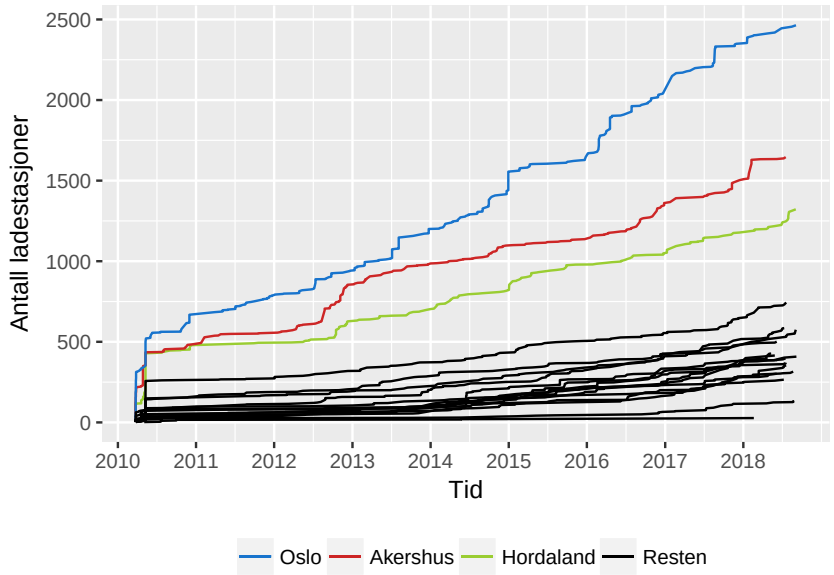



**Figur 4.2:** Utvikling i salg av elektriske kjøretøy




Ullensaker skyldes at Oslo lufthavn Gardermoen befinner seg i kommunen, hvor reisende har tilgang til mange ladestasjoner. En av Norges største næringsklynger, Fornebu, ligger i Bærum kommune. Mange arbeidsgivere har installert ladestasjoner på arbeidsplassene her, som fører til et unormalt høyt antall ladestasjoner i området.

**Figur 4.3:** Utvikling i antall ladestasjoner per fylke




Etter avgrensningene som ble gjort i forrige kapittel, ble datasettet redusert til totalt 281 148 registrerte kjøretøy hvor 49 092 var fullelektriske. Tabell 4.1 viser fordelingen av observasjoner per drivstoffkategori, fordelt på kjønn. Totalt er 33,2% av kjøretøyene registrert på kvinner og de resterende på menn. En kan også se at det er en klar overvekt av elektriske biler registrert på menn. Aldersprofilen til bileierne viser også at elbil-eierne

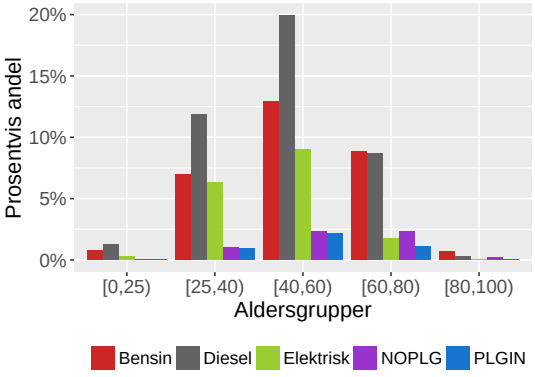
er relativt yngre, som vist i figur 4.4, med en gjennomsnittlig alder på 44,2 år. Dette er 3,3 år yngre enn dieselbil-eiere, 6,3 år yngre enn ladbare hybridbil-eiere, 6,8 år yngre enn bensinbil-eiere og 10,8 år yngre enn hybridbil-eiere. Figur 4.5 viser et boksplot over alderen til bileierne mot den avhengige variabelen. Her kan en se at det er mindre varians i alder blant elbil-eiere, samt flere uteliggere. Disse funnene bekrefter til en viss grad utsagnene til Figenbaum (2018) og Figenbaum og Kolbenstvedt (2016) om at elbilkjøpere ofte er yngre menn. Likevel, må det poengteres at omtrent samme andel av totale bilkjøpere som elbil-kjøpere er menn.

**Tabell 4.1:** Registreringer fordelt på kjønn og drivstoffkategori

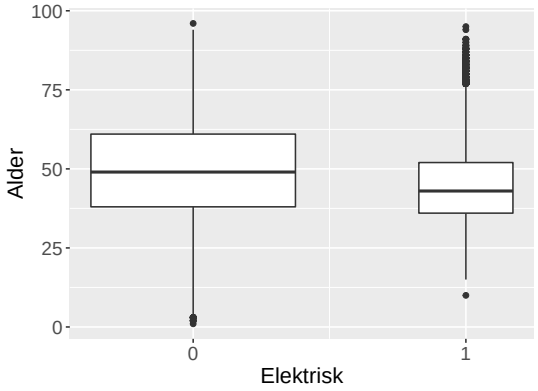
| Drivstoffkategori | Antall observasjoner (prosent av total) | Kjønn | Antall observasjoner (prosent per kategori) |
|---|---|---|---|
| Diesel | 118 098 (42,01%) | Kvinner | 32 022 (27.11%) |
| | | Menn | 86 076 (72.89%) |
| Bensin | 84 966 (30,22%) | Kvinner | 35 502 (41.78%) |
| | | Menn | 49 464 (58.22%) |
| Elektrisk | 49 092 (17,46%) | Kvinner | 15 275 (31.12%) |
| | | Menn | 33 817 (68.88%) |
| NOPLG | 16 844 (5,99%) | Kvinner | 7 846 (46.58%) |
| | | Menn | 8 998 (53.42%) |
| PLGIN | 12 148 (4,32%) | Kvinner | 2 696 (22,19%) |
| | | Menn | 9 452 (77,81%) |
| Totalt | 281 148 (100%) | Kvinner | 93 341 (33,20%) |
| | | Menn | 187 807 (66,80%) |

.

**Figur 4.4:** Fordeling av aldersgrupper




**Figur 4.5:** Boksplot over alder



## 4.2 Unsupervised Learning

*Unsupervised learning* er et sett av statistiske verktøy hvor målsetningen er å finne interessante sammenhenger mellom de ulike forklaringsvariablene. Det er dermed ikke en avhengig variabel som skal predikeres. Vi vil undersøke om det er mulig å finne undergrupper innenfor forklaringsvariablene og visualisere dette på en informativ måte. Dermed, brukes unsupervised learning ofte som en del av en forklarende dataanalyse. Vi vil bruke to verktøy innenfor unsupervised learning til dette formålet, *prinsipalkomponentanalyse* (PCA) for numeriske variabler og *korrespondanseanalyse* (CA) for kategoriske variabler (James et al., 2013).

### 4.2.1 Prinsipalkomponentanalyse (PCA)

*Prinsipalkomponentanalyse* er ifølge James et al. (2013) et verktøy for å utlede et lavdimensjonalt sett av egenskaper ut ifra et stort sett av variabler. PCA er en "unsupervised" tilnærming, ettersom den ikke krever en avhengig variabel, men et sett av forklaringsvariabler. Tilnærmingen er et ypperlig datavisualiseringsverktøy, da den kan brukes til å finne en lavdimensjonal representasjon av et stort datasett. Representasjonen fanger opp så mye av variansen som mulig, hvor dette kan visualiseres gjennom et to-dimensjonelt plan. Tanken bak metoden er at hver av de $i$ observasjonene befinner seg på et $p$-dimensjonalt plan, men alle disse dimensjonene er ikke like interessante. Derav forsøker PCA å finne et fåtall dimensjoner som er så interessante som mulig, målt etter hvor mye disse observasjonene varierer med hver dimensjon. Hver dimensjon som er valgt av PCA er en lineær kombinasjon av de ulike $p$ forklaringsvariablene. Matematisk vil den første dimensjonen, eller *prinsipale komponenten*, av et sett med forklaringsvariabler $X_1, X_2, ..., X_p$ være den normaliserte lineære kombinasjonen av forklaringsvariablene

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + ... + \varphi_{p1}X_p$$

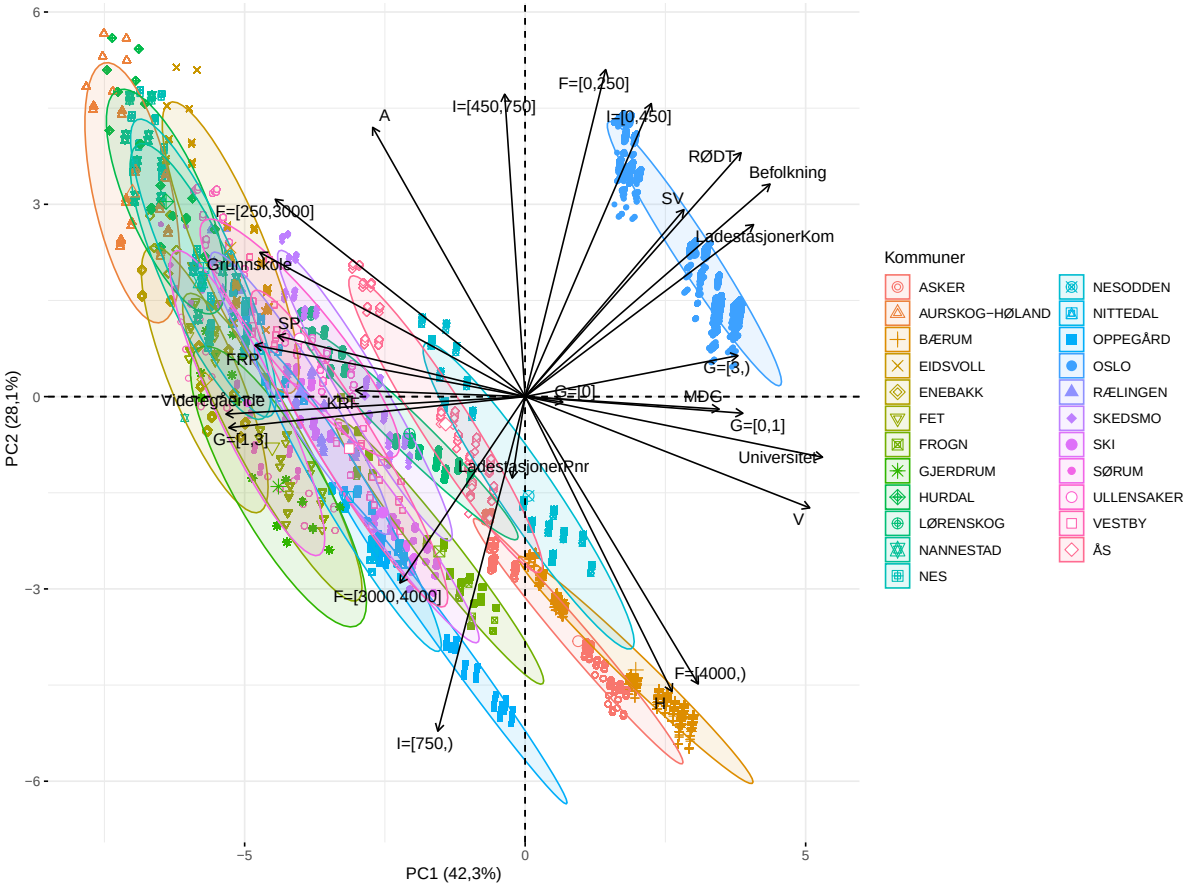
som har størst varians. Den prinsipale komponenten er *normalisert* ved $\sum_{j=1}^{p} \varphi_{j1}^2 = 1$. Elementene $\varphi_{11}, ..., \varphi_{p1}$ er *ladninger* av den første prinsipale komponenten, hvor de samlet sett gjør opp ladningsvektoren for den prinsipale komponeneten, $\varphi = (\varphi_{11}\varphi_{21}...\varphi_{p1})^T$.

Ettersom PCA kun tar imot numeriske variabler undersøkes disse variablene i datasettet. I tillegg, blir det tatt ut to supplementære kvalitative variabler; "Kommune" og "Dr.st". Til tross for at alle variablene er målt i samme skala, prosentandel i en bestemt kommune i et bestemt år, er det betydelige forskjeller i gjennomsnitt og varians. Hensikten er å kunne sammenlikne variablene, dermed blir alle variablene skalert eller standardisert til å ha et standardavvik lik en og et gjennomsnitt lik null.

Videre beregnes den andre prinispale komponenten $Z_2$. Denne er en lineær kombinasjon av $X_1, ...X_p$ som har størst varians av alle lineære kombinasjoner ukorrelerte med $Z_1$. Deretter plottes de prinsipale komponentene mot hverandre for å projisere en lavdimensjonal fremstilling av datasettet. Geometrisk sett vil dette tilsvare å projisere datasettet på et underrom med et spenn av $\varphi_1$ og $\varphi_2$, og plotte de projiserte punktene.

Den innebygde R-funksjonen *prcomp()* brukes for å utføre prinsipalkomponentanalyse. Pakken "FactoMineR" er brukt for å visualisere resultatene av analysene (Lê et al., 2008). Figur 4.6 er et biplott som illustrerer datasettet gruppert etter kommune. Dette er en type punktdiagram som kan visualisere en annenrangs matrise med både radene (observasjonene) og kolonnene (variablene), derav navnet "bi" som betyr begge (James et al., 2013). Dermed består figuren av to sammensatte grafer, hvor vi har et plott med variablene og et plott med observasjonene. Likevel, må det understrekes at koordinatene tilknyttet variablene og observasjonene ikke er i samme plan. Dermed skal en i et slik biplott ikke fokusere på de absolutte plasseringene til variablene, men heller på retningen. Denne figuren representerer både scoren, eller punktene, til de prinsipale komponentene og ladningsvektorene. Sistnevntes verdier på aksene tilsier hvor mye variablene påvirker henholdsvis den første prinsipale komponenten (X-aksen) og den andre prinsipale komponenten (Y-aksen). En kan også se fra vinklene til ladningsvektorene hvor mye variablene korrelerer med hverandre, hvor vektorer som peker i samme retning er positivt korrelerte og omvendt for vektorer som peker i motsatt retning (Grace-Martin, 2017).

**Figur 4.6:** Prinsipalkomponentanalyse gruppert etter kommune



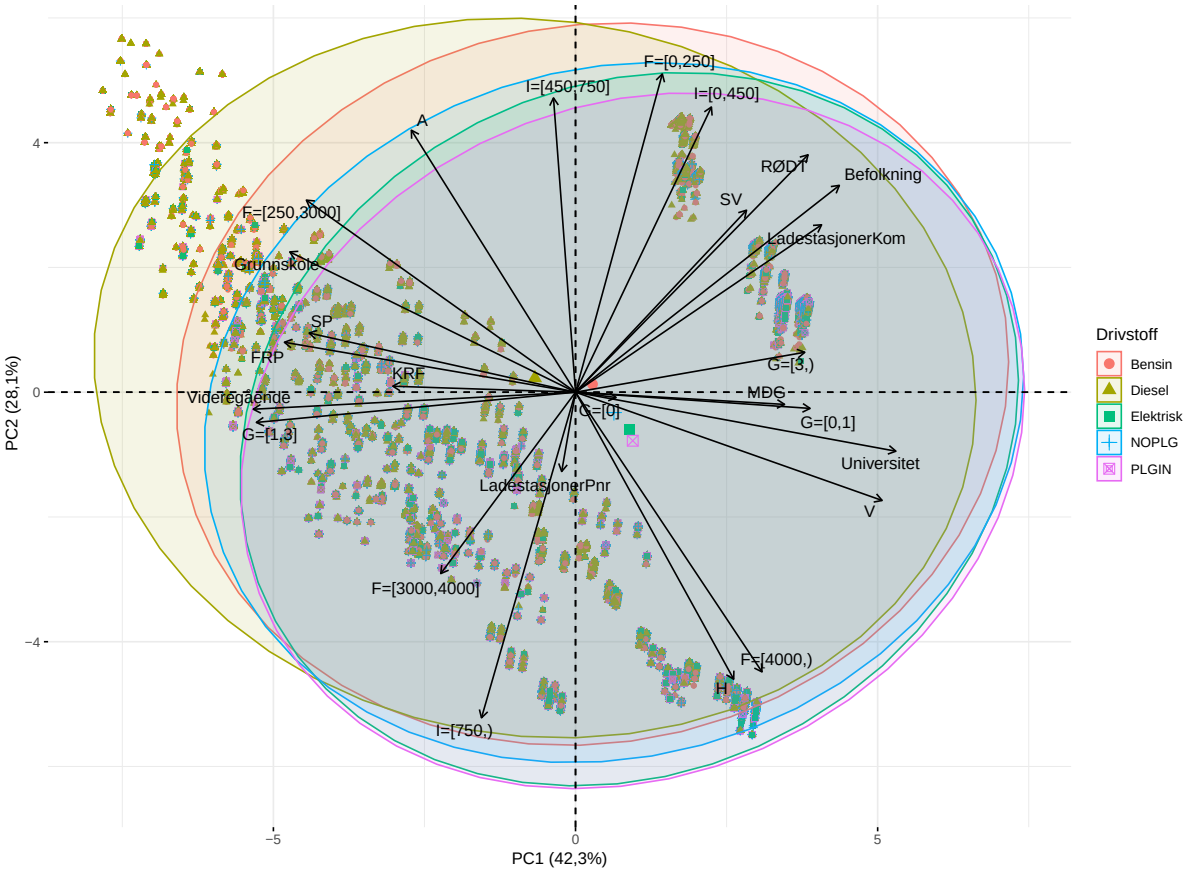

En ser fra ladningene at den første prinsipale komponenten (PC1) legger mye vekt på de ulike utdanningsnivåene og gjeld, og mindre vekt på variabler knyttet til inntekt og formue. Derav kan denne komponenten grovt samsvare til graden av utdanning og gjeld. Denne prinsipale komponenten forklarer 42,3% av variasjonen. Den andre prinsipale komponenten (PC2) legger på sin side mer vekt på inntekt og formue. Denne forklarer 28,1% av variasjonen, og samlet sett forklarer komponentene 70,4% av variasjonen i datasettet. De resterende variablene, slik som politisk parti, befolkningstall og ladestasjoner ligger mer vilkårlig rundt de ulike prinsipale komponentene. De 281 148 observasjonene har også blitt gruppert etter kommuner i Oslo og Akershus, med ulike farger og symboler for hver kommune. Variabelen "Kommune" regnes her som en supplementær kvalitativ variabel. For å skille observasjonene enda tydeligere, har det blitt lagt til ellipser som omfatter 90% av alle observasjonene.

Ut fra de prinsipale komponentene kan en se at et høyt utdanningsnivå "Universitet",

sammen med lav gjeld "G=[0]", "G=[0,1]" og høy gjeld "G=[3,)" vektlegges med en positiv verdi langs X-aksen og et lavere utdanningsnivå "Grunnskole", "Videregående" og middels gjeld "G=[1,3]" vektlegges med en negativ verdi. Videre kan en også se at høyere inntekt "I=[750,)" og formue "F=[4000,)" vektlegges negativt langs Y-aksen. Lavere inntekt "I=[0,450]", "I=[450,750]" og formue "F=[0,250]", "F=[250,3000]" vektlegges positivt langs Y-aksen. Med dette som basis, kan en enkelt se hvordan de ulike kommunene assosieres positivt og negativt med de ulike variablene. Eksempelvis kan en se at Asker, Bærum, Oppegård og Frogn kommune assosieres positivt med de korrelerte variablene høy formue "F=[4000,)", høy inntekt "I=[750,)" og partiet Høyre "H".

Motsatt, kan en se at Oslo kommune assosieres spesielt med lavere inntekt "I=[0,450]", lavere formue "F=[0,250]", høy gjeld "G=[3,)", høyt befolkningstall "Befolkning", antall ladestasjoner per kommune "LadestasjonerKom" samt "RØDT", "SV" og "MDG". Likevel må dette tolkes med omhu, ettersom Oslo kommune er den desidert største kommunen i datasettet. Til tross for at Oslo kommune er standardisert, vil den likevel grunnet sin størrelse som storby med betydelige sosio-økonomiske forskjeller innad i bydelene, veie tungt på eksempelvis lavere inntekt, formue og høy gjeld i forhold til andre kommuner. Dette kan forklare at Oslo kommune er en kommune som er mindre homogen enn andre kommuner, ikke at Oslo-borgere generelt tjener dårligere, er fattigere og har høyere gjeld. Det må også bemerkes at Oslo kommune er den kommunen som har desidert flest ladestasjoner per kommune, som kan være en viktig indikasjon på etterspørselen etter elbiler. En kan også se at kommuner som Oslo, Nesodden, Bærum og Asker alle assosieres relativt positivt med høy utdanning, "Universitet", i forhold til andre kommuner i datasettet. Ettersom det fremkommer i Figenbaum og Kolbenstvedt (2016), at elbil-eiere har høyere utdanning og bedre inntekt enn den gjennomsnittlige bilkjøper, kan en mulig hypotese være at kommuner slik som Oslo, Nesodden, Bærum, Asker, Oppegård og Frogn har et sterkere forhold til elbiler enn andre kommuner i datasettet.

**Figur 4.7:** Prinsipalkomponentanalyse gruppert etter drivstoff




For å undersøke dette, lages det et biplott av observasjonene gruppert etter variabelen "Dr.st." vist i figur 4.7. Her er det ikke mulig å få like veldefinerte grupperinger som i figur 4.6, men som likevel kan gi verdifull informasjon. Grupperingene er ikke like definerte ettersom mange av bileierne har samme type biler selv om de bor på ulike geografiske områder. Likevel, er det mulig å se en forskyvning av de ulike ellipsene som inneholder 90% av observasjonene etter drivstoff. En ser at elbiler og ladbare hybrider trekker mot positive verdier av den første og negative verdier av den andre prinsipale komponenten, og omvendt for diesel- og bensinbiler. Elbiler og ladbare hybridbiler trekker dermed forøvrig i samme retning der kommunene Oslo, Nesodden, Bærum, Asker, Oppegård og Frogn befinner seg. Dette forsterker hypotesen i forrige avsnitt. Diesel- og bensinbiler trekker på sin side i retning av kommuner som befinner seg mer ut på distriktene, slik som Aurskog-Høland, Hurdal, Nannestad, Nes og Eidsvoll. Dette kan også være en indikasjon på at det er færre incentiver for elbil-eierskap, trolig forårsaket av eksempelvis mindre kø, færre bomringer og gratis parkering for alle biler.

### 4.2.2 Korrespondanseanalyse (CA)

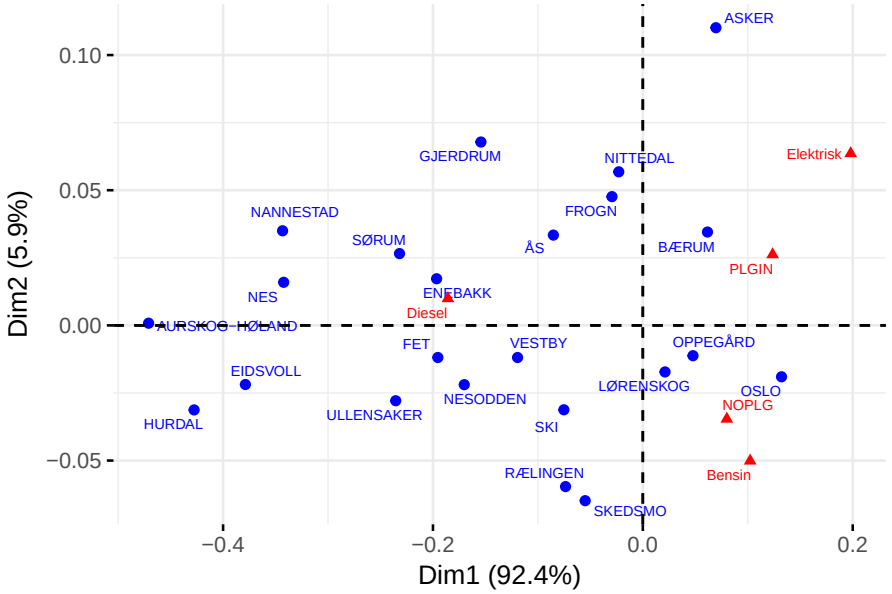
Korrespondanseanalyse er en deskriptiv metode for å analysere to- eller flerdimensjonale matriser som innehar en korrespondanse mellom radene og kolonnene. Dette gir informasjon om strukturen mellom de kategoriske variablene i matrisen. Som i prinsipalkomponentanalyse, omhandler korrespondanseanalyse dimensjonreduksjon av et datasett og projisering i et lavdimensjonalt underrom, ofte to-eller tre-dimensjonalt (Nenadic og Greenacre, 2007). Informasjon fra datasettet ekstraheres vanligvis gjennom en to-veis krysstabell, der de relative verdiene er av interesse. For å oppsummere fremgangsmåten for korrespondanseanalyse, beregnes først gjennomsnittsverdiene av hver rad og kolonne i krysstabellen. Videre, beregnes de forventede verdiene for hver celle. For en gitt celle, er dette radgjennomsnittet, multiplisert med kolonnegjennomsnittet og dividert med det totale gjennomsnittet. Neste steg er å beregne residualene, som viser forholdet mellom kolonneetikettene og radetikettene i kryssmatrisen. Residualene er beregnet ved å trekke fra de forventede verdiene fra originalverdiene. Store positive verdier signaliserer et sterkt positivt forhold og omvendt. Det siste steget er å visualisere kolonne- og radverdiene med like residualer sammen i et to- eller tre-dimensjonalt plott (Bock, 2018; Greenacre, 1984).

Funksjonen *CA()* fra R-pakken "FactomineR" (Lê et al., 2008) brukes for å utføre korrespondanseanalyse. Denne funksjonen tar inn en vanlig toveis-kryssmatrise som input, og gir ut resultatene av korrespondanseanalysen på de ulike krysstabellene som lages. Det er likevel viktig at krysstabellen er i samme skala før den brukes i CA()-funksjonen for å gjennomføre en korrekt korrespondanseanalyse. I denne utredningen er dette ikke et problem, ettersom de ulike krysstabellene som er sammenstilt fra datasettet er basert på antall tilfeller av variabelen "Dr.st". "Alder" blir formatert i ulike aldersgrupper og tar dermed form som en kategorisk variabel kalt "Aldersgruppe". Det blir slik sammenstilt standardiserte krysstabeller av "Dr.st" med de kategoriske variablene "Kommune" og "Aldersgruppe".

Resultatet av disse korrespondanseanalysene blir grafisk fremstilt i lignende biplott som i kapittel 4.2.1. Radene er representert med blå punkter og kolonnene med røde trekanter. På samme måte som i PCA, er ikke distansen mellom rad og kolonnepunktene meningsfulle.

Det er kun distansen innad i radpunktene og kolonnepunktene som kan tolkes. Avstanden gir en måling av deres likhet eller ulikhet. Likevel, kan en tolke observerte mønstre som forekommer mellom radene og kolonnene.

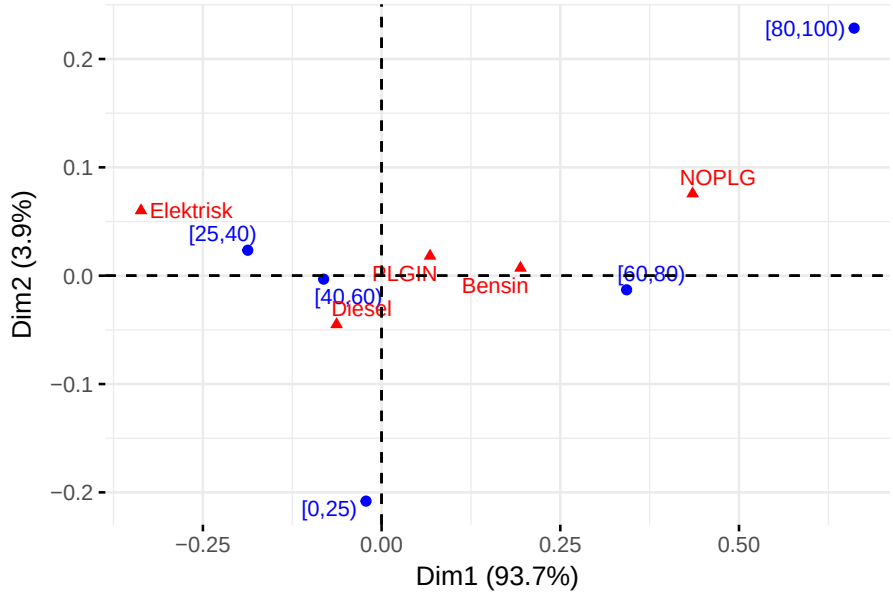
**Figur 4.8:** Korrespondanseanalyse - drivstoff og kommune




Ved å studere biplottet av "Drivstoff" og "Kommune" i figur 4.8, er det mulig å observere at den første dimensjonen (Dim1) forklarer 92,4% av variansen, dermed inneholder den mye av informasjonen til datamatrisen. Den andre dimensjonen (Dim2) forklarer 5,9% av variansen, og sammen vil disse forklare 98,3% av variansen. Det er dermed ikke nødvendig med flere dimensjoner, da disse nærmest ikke vil gi noe ytterligere informasjon. Det skal likevel påpekes at den andre dimensjonen forklarer relativt lite av variansen isolert sett. Derav må en være varsom med å trekke slutninger mellom rad- og kolonnepunktene langs Y-aksen. Eksempelvis kan en forvente seg at kommunene Asker og Bærum, som ligger hverandre nære både geografisk og demografisk, vil ligge nære hverandre i dette bi-plottet også. Dette er også egentlig tilfellet, men den skjeve skaleringen av plottet gjort for å øke lesbarheten kan være misvisende om en er uoppmerksom. Den første dimensjonen har en positiv verdi for elbiler og ladbare hybrider, og en negativ verdi for dieselbiler. På denne måten, kan en se at kommuner som Oslo, Asker, Bærum og Oppegård har en positiv verdi langs den første dimensjonen. Igjen trekker mange av de samme kommunene som ligger i distriktene mot diesel-biler, slik som Aurskog-Høland, Hurdal, Eidsvoll, Nannestad, Nes, Sørum, Ullensaker, Enebakk, Gjerdrum og Fet. Det kan også nevnes at kommunen

Nesodden også har en negativ verdi for denne dimensjonen, som kan tilsi at forholdet til elbiler i denne kommunen kanskje ikke er så sterkt som tidligere antatt.

**Figur 4.9:** Korrespondanseanalyse - drivstoff og aldersgrupper




I figur 4.9 studeres forholdet mellom "Drivstoff" og "Aldersgrupper", hvor alderen til de ulike bileierne i Oslo og Akerhus har blitt delt inn i 5 aldersgrupper. I likhet med figur 4.8, forklarer den første dimensjonen en stor del av av variansen (93,7%), og den andre dimensjonen en relativt enda mindre del av variansen (3,9%). Sammen forklarer disse dimensjonene 97,6% av variansen som er veldig høyt. Langs den første dimensjonen er det positive verdier for eldre aldersgrupper, som [60,80] og [80,100] , og negative verdier for yngre aldersgrupper, som [0,25], [25,40] og [40,60]. Elbiler, og til en viss grad også dieselbiler og ladbare hybrider, trekker mot yngre aldersgrupper, mens bensinbiler er mellom yngre og eldre aldersgrupper. Hybridbiler trekker i relativt stor grad mot eldre aldersgrupper. Dette bekreftes også av resultatene til Figenbaum og Kolbenstvedt (2016), hvor det fremkommer at elbil-eiernes aldersprofil er ung, typisk mellom 35-54 år. Ladbare hybrid-eiere og bileiere med forbrenningsmotor, herunder bensin og diesel, er henholdsvis typisk i aldersgruppene 45-66 år og 55-66 år.

## 4.3 Interessante variabler

Summert kan vi gjennom deskriptiv statistikk og forklarende dataanalyse utforske datasettets struktur og finne interessante sammenhenger. Gjennom disse metodene er det også mulig å utlede enkelte variabler som særlig har vist seg å være viktige ved de ulike visualiseringene. Ut ifra de disse visualiseringene har særlig variabler som "Alder" og "Kommune" hatt en relativt stor betydning. Eksempelvis ser en av figur 4.5 at alderen til elbil-eiere er noe lavere enn andre bileiere. Dette virker bekreftende i figur 4.9, hvor en kan se at lavere aldersgrupper trekker i samme retning som elbiler. Videre er mange av variablene basert på kommunenivå, som kan tyde på at "Kommune" er en viktig variabel for prediksjonene. Det fremkommer av figur 4.7 og figur 4.8 at kommunene Oslo, Asker, Oppegård og Bærum kan ha et sterkere forhold til elbiler enn andre kommuner. Det kan videre tenkes at assosierte variabler til disse kommuene kan være interessante for prediksjonene, eksempelvis "I=[750,)", "H" og "F=[4000,)". Likevel, må det presiseres at disse vurderingene ikke er gjennomført som en del av variabelutvelgelse i denne utredningen. Det vil bli anvendt en mer sofistikert og nøyaktig metode for variabelutvelgelse forklart i neste kapittel, før prediksjonsmodellene estimeres.

# 5 Metode

Vi vil i det følgende fokusere på trebaserte prediksjonsmetoder, herunder *klassifiseringstrær*, *random forests* og *extreme gradient boosting*. Prediksjon ved hjelp av slike metoder har ifølge James et al. (2013) og Friedman et al. (2008) enkelte fordelaktige bruksområder. Eksempelvis har disse godt definerte prediksjonskoeffisienter og de er enkle å forstå gjennom intuitive grafiske illustrasjoner. Trebaserte metoder kan også være et nyttig verktøy innenfor variabelutvelgelse, hvor disse kan estimere variablenes viktighet for prediksjonene. Disse har også fordelen med at de kan takle kategoriske (ordinale og nominelle) og numeriske variabler samtidig (Song og Ying, 2015). Samlet gjør disse egenskapene at de nevnte metodene er godt egnet til å både analysere viktigheten av forklaringsvariablene for klassifisering og predikere potensielle elbil-eiere.

## 5.1 Estimering og validering

I likhet med andre prediksjonsutredninger, brukes et estimeringssett og et valideringssett i elbilprediksjonene. Ifølge James et al. (2013) er dette en enkel strategi som går ut på å tilfeldig dele datasettet inn i to deler, et estimeringsett eller treningssett og et valideringssett eller testsett. De statistiske modellene er tilpasset estimeringsettet og videre brukt til å predikere den avhengige variabelen på valideringssettet. Det er ofte et mål med klassifiseringsprediksjoner å minimere *test-feilraten*, hvilket er antall prediksjoner som ble feilaktig klassifisert i valideringssettet. Likevel har denne tilnærmingen en potensiell ulempe. Test-feilraten kan ha høy varians, som avhenger direkte av hvilke observasjoner som blir tilfeldig valgt i estimeringssettet.

### 5.1.1 K-fold kryssvalidering

For å løse ulempen knyttet til tilnærmingen i 5.1 brukes *k-fold kryssvalidering* i videre modellestimering. Denne tilnærmingen deler datasettet, i vårt tilfelle det tidligere definerte estimeringsettet, inn i $k$ like store ikke-overlappende grupper. Estimeringsettet brukes for å ha et urørt testsett for validering av endelige modeller. Den ene gruppen fungerer

som et valideringssett og de resterende gruppene som estimeringssett. Modellene blir kalkulert $k$ ganger, hvor gruppen som brukes som valideringsett endres for hver gang. En test-feilrate estimeres også for hver modellkalkulering $k$, hvor gjennomsnittet av disse kryssvaliderte test-feilratene brukes som det endelige prediksjonsresultatet til modellene. Det fremkommer i James et al. (2013) at $k$-fold kryssvalideringens feilrate og testsettets feilrate defineres likt. En kunne alternativt gjennomført "Leave-one-out"-kryssvalidering hvor antall grupper tilsvarer antallet observasjoner i datasettet, $k = i$. Denne tilnærmingen krever store mengder prosseseringskraft og ville i vårt tilfelle ikke vært gjennomførbart, grunnet et stort datasett og kompliserte estimeringsmetoder.

## 5.2 Trebaserte metoder

Trebaserte metoder for klassifisering baserer seg i hovedsak på en segmentering av forklaringsvariablene inn i ulike regioner. Ettersom denne segmenteringen av forklaringsvariablene kan bli summert i et tre, er disse metodene kjent som beslutningstre-metoder. Til tross for at slike beslutningstrær er enkle å tolke, er de ikke nødvendigvis de beste prediksjonsmetodene når det kommer til prediksjonsevne. Tolkningene er viktige for denne utredningen, ettersom det er et mål å finne et sett av forklaringsvariabler som karakteriserer elbil-eiere. Likevel er det også ønskelig å finne en metode som kan slå *modusen* av den avhengige variabelen, altså nøyaktigheten[4] av å predikere ikke-elbil hver gang. Derav, skal vi i tillegg til vanlige klassifiseringstrær også benytte metoder som random forests og extreme gradient boosting. Ved å kombinere et stort antall trær er det mulig å øke prediksjonsevnen betydelig, med et lite påfølgende tap i tolkningsgrad (James et al., 2013).

### 5.2.1 Klassifiseringstrær

For å predikere en avhengig variabel, vil et klassifiseringstre ifølge James et al. (2013) segmentere forklaringsvariablene $X_1, X_2, ..., X_p$ inn i $j$ distinkte ikke-overlappende regioner $R_1, R_2, ...R_j$. For å bruke riktig tre-analogi, vil regionene kalles *endeknutepunkter* eller *bladene* til treet. De ulike punktene der treet splittes kalles *internknutepunkter*. Segmentene

[4]Nøyaktighet defineres som andelen korrekte prediksjoner over totalt antall prediksjoner

som knytter sammen nodene kalles treets *greiner*. Gjennom gjentagende binære splitter vil en dyrke et klassifiseringtre gjennom å bruke verdier fra en *Gini indeks* eller *entropi*. Ettersom disse målene har mange likhetstrekk og Gini indeksen bruker mindre prosseseringskraft, anvender vi dette målet videre i analysen. Gini indeksen er definert som

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{5.1}$$

og er et mål på den totale variansen av alle $k$ klasser. Her representerer $\hat{p}_{mk}$ proporsjonen av treningsobservasjoner i region $m$ fra klassen $k$. Observasjonene blir predikert til å tilhøre den mest forekommende klassen av treningsobservasjonene i regionen den tilhører. Gini indeksen er et mål på *nodens renhet*, der en liten verdi indikerer at en node har flest observasjoner fra en enkelt klasse. Det må presiseres at Gini indeksen brukes med hensyn til å dyrke et tre, hvor klassfiseringsfeilraten heller blir brukt som et mål på prediksjonsevnen. Grunnen er at Gini indeksen, sammenlignet med klassifiseringsfeilraten, er mer sensitiv til nodens renhet når treet dyrkes (James et al., 2013).

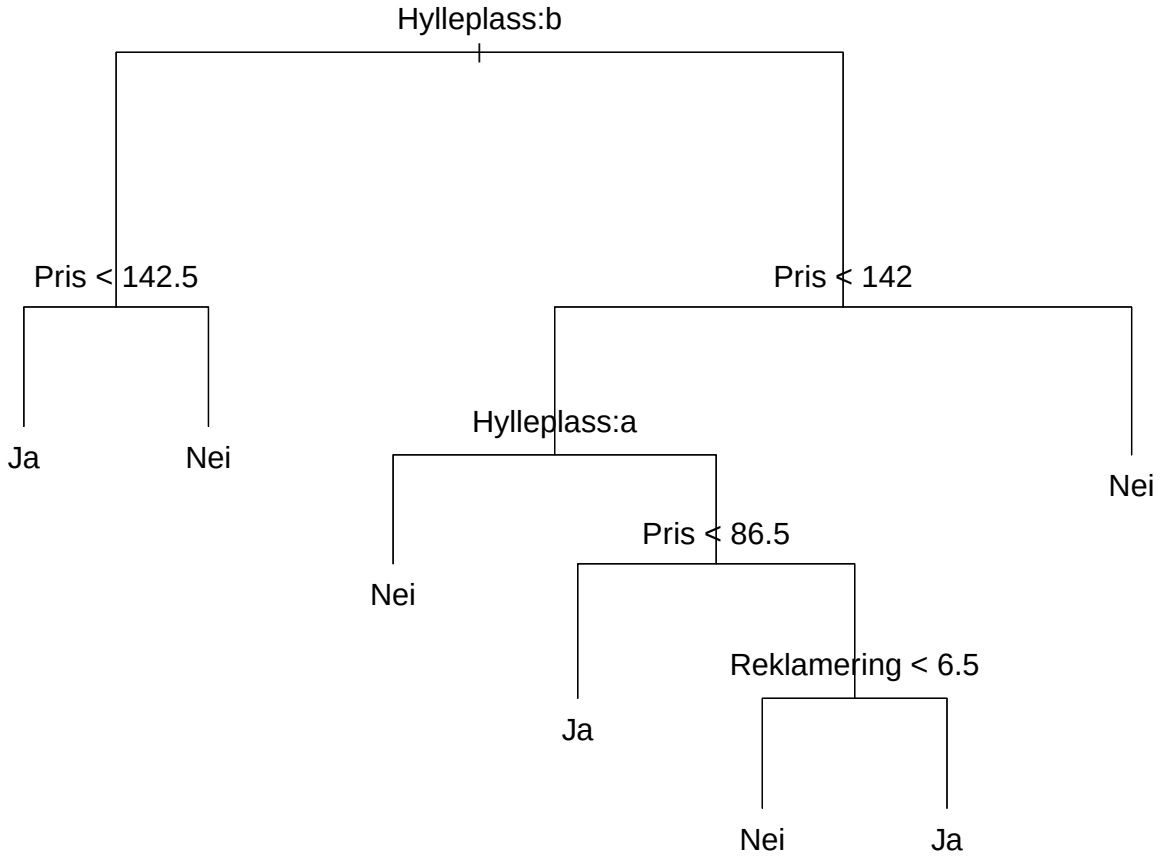
Algoritmen søker å finne de viktigste forklaringsvariablene og beskjæringspunktene, og dermed minimere urenheten i regionene for hver splitt. Denne gjentakende binære splittingen kalles også den *grådige* tilnærmingen ettersom den velger den beste splitten for hvert steg, i stedet for å velge en splitt som kan føre til et bedre tre i et fremtidig steg. Derav vil den første splitten ha den største reduksjonen i nodens urenhet på treningsobservasjonene, som finnes ved å undersøke alle forklaringsvariabler $X_1, ...X_p$ og deres mulige skjæringspunkter[5] $s$. Den kombinasjonen av forklaringsvariabel og skjæringspunkt som minimerer urenheten mest i splitten vil bli valgt. Videre vil denne prosessen gjentas, hvor hver splitt søker å redusere nodenes urenhet helt til et kriterium er nådd, eksempelvis at en region ikke har flere enn fem observasjoner. Dette tilsier dermed at den første noden er den viktigste forklaringsvariabelen (James et al., 2013). Ettersom effekten av hver forklaringsvariabel enkelt kan bli lest av beslutningstreet, gjør dette klassifiseringstrær til en populær metode. Likevel, kan tolkningen av effekten være misvisende ettersom klassifiseringstrær har en ustabil struktur. Små endringer i estimeringssettet kan føre til et veldig annerledes tre, grunnet egenskapene til den

[5]Skjæringspunkt er verdiene som brukes for å skille klassene i den avhengige variabelen.

grådige tilnærmingen. Dermed må en være forsiktig med å bestemme de viktigste forklaringsvariablene ut fra et enkelt beslutningstre (Strobl, 2008).

Figur 5.1 viser et eksempel på et klassifiseringstre som viser salg av barneseter fra 400 forskjellige butikker. Fra internknutepunktene lages det nye forbindelser med vertikale linjer, treets greiner, som leder til neste internknutepunkt før de til slutt ender opp som endeknutepunkter. Selv om endeknutepunktene har to verdier, *Ja* og *Nei*, vil disse verdiene forekomme i ulike regioner i datasettet. *Ja* og *Nei* refererer til en henholdsvis positiv eller negativ prediksjon av et bilsete-kjøp. En kan også se at det kan være kategoriske variabler på internknutepunktene, eksempelvis "Hylleplass", som angir hvor god plassering produktet har i hyllene. Teksten "Hylleplass:b" indikerer at greinen til venstre for noden består av observasjoner fra den andre (*b: God*) verdien av "Hylleplass", og greinen til høyre består av de resterende observasjonene (*a: Dårlig*) og (*c: Middels*). Dersom barnesetet har en god plassering på hyllen, "Hylleplass:b", og varen har en pris under 142,5, "Pris<142,5", får den avhengige variabelen "Salg" en positiv predikert verdi *Ja*.

**Figur 5.1:** Eksempel på et klassifiseringstre



Likevel kan prosessen beskrevet over med å dyrke et klassifiseringstre gi dårlige prediksjoner på et valideringssett. Til tross for at et stort tre med mange noder gir tilsynelatende gode prediksjonsresulater på et estimeringssett, er det en fare for overtilpasning. Dette kan unngås gjennom å *beskjære* treet. Det innebærer å beskjære grener som bruker forklaringsvariabler med lav viktighet. Et beskåret tre vil kunne føre til en høyere feilrate på treningssettet, men lavere varians og bedre prediksjoner på valideringssettet.

Et klassifiseringstre kan bli beskåret gjennom *cost complexity pruning*. Denne metoden benytter en kompleksitetsparameter $a$, som lager en sekvens av under-trær som en funksjon av $a$. Parameteren kontrollerer størrelsen av treet ved å legge til en kostnad for å legge til flere endeknutepunkter. Det er mulig å trimme denne parameteren optimalt med k-fold kryssvalidering, som forklart i kapittel 5.1.1. Både klassifikasjonsfeilraten og Gini indeksen kan bli brukt til å beskjære et tre, men ifølge James et al. (2013) vil klassifikasjonsfeilraten være foretrukket for bedre nøyaktighet. Verdien $a$ som oppnår den laveste gjennomsnittlige klassifikasjonsfeilraten vil derfor være optimal.

### 5.2.2 Random forests

Sammenlignet med vanlige klassifiseringstrær, kan bruk av random forests til klassifisering bidra til økt prediksjonsevne og redusert varians. Hovedtanken bak denne algoritmen er å bruke flere treningssett fra datasettet, dyrke beslutningstrær på hvert treningssett, registrere klassen som predikeres av hvert tre $b$ og bruke en *majoritetsstemme* som avgjørende prediksjon. Majoriteten er den klassen som predikeres oftest ved de ulike beslutningstrærne. Dette senker også variansen, ettersom majoritetstemmen vil velge den klassen som forekommer gjennomsnittlig oftest, på samme måte som gjennomsnittet av et sett med observasjoner reduserer variansen.

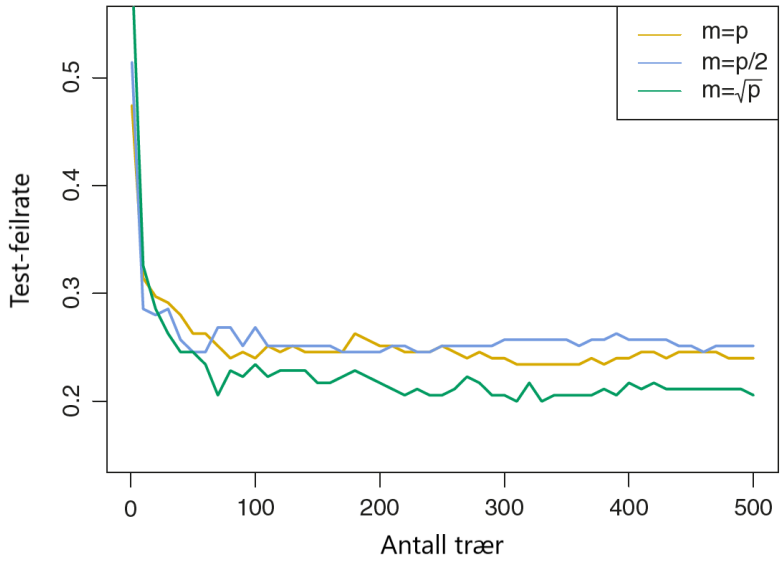
Ettersom det ikke er tilgang til flere treningssett, bruker random forests en teknikk som kalles *bootstrapping*. Tilnærmingen går ut på å bruke flere utvalg fra det ene treningsettet som er tilgjengelig, og lage et bootstrap-utvalg fra det originale datasettet med samme antall observasjoner. Algoritmen tar et utvalg fra observasjonene med erstatning, som betyr at en enkel observasjon i det originale datasettet kan dukke opp flere ganger i hvert bootstrap-utvalg. Denne teknikken fanger dermed opp mye av variasjonen i datasettet.

På samme måte som i klassifiseringstrær, brukes Gini indeksen som et mål for å splitte internknutepunktene i trærne. Dette målet brukes også til å definere variablenes viktighet under en modelltilpasning, hvor det regnes ut hvor mye hver forklaringsvariabel $p$ senker den vektede urenheten ved nodene i et tre. Deretter kalkuleres den gjennomsnittlige reduksjonen i nodenes urenhet for hver variabel over alle trær og rangeres etter nevnte mål (Louppe et al., 2013). Sammenlignet med enkle klassifiseringstrær kan ikke variablenes viktighet sees like tydelig fra beslutningstrærne som dyrkes i random forest. Derimot er beregningen av viktighet for hver variabel betraktelig mer robust, da random forest ikke har like stor varians som klassifiseringstrær (James et al., 2013). Denne egenskapen har gjort random forest til et anerkjent verktøy innenfor anvendt forsking hvor viktighetsmålet er brukt eksempelvis innenfor genetikk og bioinformatikk (Strobl, 2008).

Ettersom hvert bootstrap-utvalg bygger et beslutningstre, vurderer random forests kun et tilfeldig utvalg $m$ av $p$ forklaringsvariabler på hvert internknutepunkt. Verdien av $m$ bestemmes ofte som $m \approx \sqrt{p}$. Noen av disse trærne vil ikke engang vurdere å bruke den sterkeste forklaringsvariabelen, slik at andre forklaringsvariabler vil kunne ha en sjanse. Dette medfører mindre korrelasjon mellom de ulike beslutningstrærne og dermed en reduksjon i avhengighet. En random forest modell som estimeres med $m = p$, refereres til som *bagging*. Dersom det foreligger mange korrelerte forklaringsvariabler i et datasett vil det kunne være hensiktsmessig å bruke en lav verdi for $m$, da beslutningstrærne blir mer varierte. Antallet beslutningstrær $B$ er en annen parameter som kan bestemmes. Det fremkommer i James et al. (2013) at random forests ikke vil ha problemer med overtilpasning ved en økning av $B$, som i praksis betyr at denne verdien kan settes stor nok til at test-feilraten stabiliseres på et minimum.

Figur 5.2 viser et eksempel på bruk av random forests klassifisering på et høydimensjonalt datasett, med 349 observasjoner og $p = 500$ forklaringsvariabler. Her kan en se $B \approx 400$ beslutningstrær er tilstrekkelig for å holde test-feilraten på et stabilt minimum. Det kan også observeres at random forests med at $m \approx \sqrt{p}$ gir et bedre resultat på test-feilmarginen enn bagging, $m = p$.

**Figur 5.2:** Random forests med test-feilrate




Kilde: James et al., 2013, side 322

For å estimere test-feilraten til random forests er *Out-of-Bag* (OOB) *error estimate* en god tilnærming ifølge James et al. (2013). Ettersom beslutningstrærne er dyrket på bootstrappede-utvalg er hvert tre i gjennomsnitt dyrket på omtrent 2/3 av observasjonene. Derav er de resterende 1/3 av observasjonene utelatt og defineres som *out-of-bag* observasjoner. For hver observasjon, kan den avhengige variabelen predikeres ved å bruke hver av de trærne hvor den spesifikke observasjonen var utelatt. Dette vil gi omtrent *B/3* prediksjoner for hver observasjon *i*. Gjennom en majoritetstemme over disse prediksjonene, kan en enkel OOB-prediksjon for hver observasjon bestemmes. Slik kan klassifiserings-feilraten kalkuleres som et resultatet av disse OOB-prediksjonene, som er et gyldig estimat på test-feilraten. Denne tilnærmingen for å finne test-feilraten er særlig anvendelig når random forests blir brukt på store datasett, ettersom en eventuell kryssvalidering ville krevd stor prosseseringskraft.

### 5.2.3 Extreme gradient boosting

*Boosted trees* er en prediksjonsmetode innenfor klassifisering som kan minne om andre trebaserte metoder, eksempelvis bagging og random forests. Innen både bagging og random forests bygges hvert tre uavhengig av hverandre. Trærne som bygges ved hjelp av boosting tar utgangspunkt i de tidligere konstruerte trærne og prøver å forbedre modellen basert

på denne informasjonen, modellen bygges altså opp sekvensielt. Residualene fra det forrige treet brukes for å utarbeide det neste treet i rekken. Det fører til at hvert tre som legges til modellen senker residualenes verdi. På den måten trenes modellen på variansen i datasettet som ikke allerede er forklart og den vil forbedres på områder hvor den tidligere ikke presterte godt. Det kan også bemerkes at variablenes viktighet beregnes og tolkes på samme måte for boosting som random forests, men refereres til som "Gain". Dette viktighetsmålet representeres på en relativ skala hvor den samlede viktigheten til alle variablene summeres til 1 (James et al., 2013; Kuhn og Johnson, 2013).

Det var Friedman et al. (2000) som først implementerte en algoritme for klassifisering med boosted trees. De beviste at for klassifisering kan metoden ses på som en framover stegvis additiv metode hvor en eksponentiell tapsfunksjon minimeres. Videre utviklet de et rammeverk basert på dette beviset som ble kalt *gradient boosting machines*.

I hvert enkelt tre som dyrkes tilegnes alle observasjonene i estimeringssettet en vekt, $w_i$, som ved starten av algoritmen er $w_i = \frac{1}{Antall\ observasjoner}$. Ved slutten av hver iterasjon blir vektene individuelt oppdatert. Hvis modellen klassifiserer en observasjon feil vil observasjonen få høyere vekt i neste iterasjon, og omvendt. Dette fører til at observasjoner som er vanskelige å klassifisere får stadig høyere vekt og estimeringen av modellen tvinges til å fokusere på de vanskelige observasjonene (Friedman et al., 2008).

Boosting er en metode som lærer sakte, som ofte resulterer i gode prediksjonsresultater. Ettersom hvert ekstra tre som legges til modellen senker residualenes verdi, kan en estimere modellen på en stor mengde trær og hele tiden oppnå bedre estimeringsresultater. Likevel, vil dette føre til at metoden er utsatt for overtilpasning. Boosted trees har derfor tre parametere som skal forhindre dette samt forbedre prediksjonsevnen (James et al., 2013).

Den første parameteren, $d$, begrenser hvor dypt trærne kan vokse, eller hvor mange splitter hvert tre maksimalt kan ha. En høyere verdi av $d$ vil gjøre at en kan finne interaksjoneffekter mellom variablene i datasettet, som kan føre til mulig overtilpasning (James et al., 2013; Friedman et al., 2008).

Parameteren $\lambda$ (også kalt skrumpingsparameteren eller *shrinkage)* påvirker læringshastigheten til algoritmen. Denne parameteren skalerer bidragene hvert nytt tre tilfører modellen, hvor en lav parameterverdi tilsier saktere læring (Friedman

et al., 2008). For å utnytte effektene av sakte læring er det fordelaktig å bygge flere trær, men dette kan føre til problemer med overtilpasning (James et al., 2013). Det er derfor nødvendig å se $\lambda$ i sammenheng med parameter $B$, som forteller modellen hvor mange trær som skal dyrkes. En endelig klassifiseringsmodell med to prediksjonsklasser basert på boosted trees kan formuleres slik matematisk:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{5.2}$$

hvor $\hat{f}^b(x)$ er sannsynligheten for at en observasjon tilhører den ene klassen i klassifiseringsproblemet (Kuhn og Johnson, 2013).

*Extreme gradient boosting* er en videreutvikling av gradient boosting machines. Den ble utviklet av Chen og Guestrin (2016) og har økt i popularitet de siste årene. Metoden legger til et regulariseringsuttrykk som forhindrer overtilpasning og en andregrads approksimasjon som øker ytelsen sammenlignet med gradient boosting machines. Som nevnt tidligere får alle observasjonen i datasettet en vekt ($w_i$). Summen av disse vektene ($w$) brukes i et L2 norm[6] regulariseringsuttrykk for å straffe komplekse modeller proporsjonalt med kvadratroten av $w$. Chen og Guestrin implementerer også en andregrads approksimasjon for å forenkle målfunksjonen sammenlignet med metoden utviklet av Friedman et al. (2000). Denne forenklingen øker hastigheten til kalkuleringene betraktelig og gjør prediksjonene mer nøyaktige.

I tillegg til de tre parametrene tidligere beskrevet har extreme gradient boosting fire ekstra parametere som kan brukes til å optimalisere modellen. *Gamma* er en parameter som stopper algoritmen om trærne som legges til ikke reduserer målfunksjonens verdi tilstrekkelig. *Min_ child_ weight* kontrollerer minimum antall observasjoner som kan være i et endeknutepunkt. *Sub_ sample* definerer hvor mye av datasettet hvert tre skal bruke til trening. Det fører til mindre sannsynlighet for overtilpasning og hurtigere kalkuleringer. Den siste parameteren er *colsample_ bytree*. Parameteren bestemmer hvor stor andel av datasettets variabler en skal bruke til å kalkulere hvert enkelt tre, som i random forests (Chen et al., 2017).

Metoden kan oppsummeres av følgende ligning

[6]Også kjent som minste kvadrat.

$$\mathcal{L}(\phi) = \sum_{i=1}^{I} l(\hat{y}_i, y_i) + \sum_{b=1}^{B} \Omega(f_b)$$
$$\text{hvor } \Omega(f_b) = \gamma T + \frac{1}{2}\eta ||w||^2 \tag{5.3}$$

hvor målet er å minimere tapsfunksjonen $\mathcal{L}(\phi)$. Her representerer $l$ andregrads approksimasjonen og $\Omega(f_b)$ straffer komplekse modeller. En kan observere at L2 norm regulariseringen er implementert i $\Omega(f_b)$. Trær med mange blader straffes via $\gamma$, som referer til *Gamma*, og T som representerer antall blader i et gitt tre. $I$ representerer antall observasjoner i datasettet og $B$ representerer fremdeles antall trær modellen dyrker. $\eta$ er en parameter som styrer styrken til regulariseringsuttrykket (Chen og Guestrin, 2016).

## 5.3 Variabelutvelgelse

Ved mange tilfeller er enkelte av variablene som brukes i en prediksjonsmodell ikke like relevante i forhold til den avhengige variabelen. Det å inkludere slike irrelevante variabler kan føre til en unødig økning i modellens kompleksitet, som kan gi problemer med overtilpasning. En god variabelutvelgelse kan øke prediksjonsnøyaktigheten, redusere estimeringstiden til modellen og gjøre det lettere å tolke modellens resultater (Cai et al., 2018). Vi søker dermed å velge ut variabler som er relativt viktige, hvor viktighet defineres som hvor mye en forklaringsvariabel bidrar til å gjennomføre en prediksjon på den avhengige variabelen. Det er viktig å poengtere at variabelens viktighet ikke nødvendigvis har en sammenheng med modellens nøyaktighet, bare i hvilken grad en variabel er viktig for å kunne gjøre en prediksjon (IBM, 2018).

### 5.3.1 Boruta-algoritmen

For variabelutvelgelse anvendes R-pakken *Boruta* (Kursa et al., 2010), som benytter en algoritme innpakket rundt random forests-algoritmen fra R-pakken *randomForest* (Liaw og Wiener, 2002). Boruta-algoritmen fungerer ved at den lager duplikater av alle forklaringsvariabler i treningsettet, kalt *skyggevariabler*. Verdiene til duplikatene blir så blandet for å fjerne korrelasjonen med den avhengige variabelen og kombinert med de originale forklaringsvariablene. Videre gjennomføres en random forests klassifisering hvor

det måles for variablenes viktighet etter hvor mye hver variabel bidrar til nøyaktigheten. Deretter regnes det ut en *Z-score*, som er et gjennomsnittlig tap av presisjon dividert på variabelens standardavvik. Dersom de originale forklaringsvariablene har en Z-score lavere enn den maksimale Z-scoren til en skyggevariabel, blir de klassifisert som "uviktige" og omvendt (Kursa et al., 2010).

Likevel bruker ikke Boruta kun Z-score som viktighetsmål, da denne ikke direkte er relatert til den statistiske signifikansen av variabelens viktighet. I stedet prøver Boruta å validere viktigheten av forklaringsvariablene ved å sammenligne Z-scoren med de tilfeldig blandede skyggevariablene, som øker robustheten til resultatene. Dette gjøres gjennom å sammenligne antall ganger en forklaringsvariabel fikk en høyere Z-score enn skyggevariablene ved å bruke en binomisk distribusjon. Denne prosessen fortsetter i flere iterasjoner til alle variabler enten er klassifisert som "viktig", "uviktig" eller "ubestemt" (Kursa et al., 2010).

Kort fortalt, er Boruta basert på samme fundament som random forests. Det legges til tilfeldighet i systemet og resultatene av de randomiserte utvalgene samles inn, som reduserer den misledende effekten av tilfeldige fluktuasjoner og korrelasjoner. I denne konteksten, vil randomiseringen bidra til et klarere syn på hvilke forklaringsvariabler som er viktige (Kursa et al., 2010).

## 5.4 Receiver Operating Characteristics (ROC)

Det finnes mange måter å måle prestasjonsevnen til prediktive modeller på. Innen klassifisering brukes ofte nøyaktighet som et mål på prediksjonsevne. Ved et klassifiseringsproblem hvor klassene er ubalanserte kan de nevnte målene gi et feilaktig bilde av prestasjonen til en statistisk modell. I datagrunnlaget er totalt 17,46% av alle observasjonene registrert som elbil. Det vil si at om en predikerer modusen på den avhengige variabelen, vil modellen oppnå en nøyaktighet på 82,54%. Det er derfor ønskelig å ha andre prestasjonsmål for datasett som er ubalanserte (Fawcett, 2006).

*Receiver Operating Characteristics*, eller ROC, er en metode for å visualisere prestasjonene til klassifiseringer. Når en binær variabel predikeres, en variabel med to klasser, vil det finnes totalt 4 utfall. Predikeres en observasjon som elbil-eier og dette er korrekt, regnes

dette som en *ekte positiv* prediksjon. Predikerer modellen derimot denne observasjonen til å ikke være elbil-eier selv om det er en elbil-eier, vil dette være en *falsk negativ* prediksjon. Tilsvarende finnes også *ekte negative* og *falske positive* prediksjoner. Dette er illustrert i tabell 5.1. Falske positive klassifiseringer omtales også som *type 1* feil og falske negative klassifiseringer omtales som *type 2* feil. Prediksjonsmetodene tidligere forklart vil tilegne hver observasjon en sannsynlighet for at de tilhører en av klassene. Brukeren kan balansere hvor mange type 1 og type 2 feil de kan tolerere ved å sette en terskelverdi på denne sannsynligheten (James et al., 2013; Fawcett, 2006). Det er standard å bruke en terskelverdi på 50%, der alle prediksjoner med en sannsynlighet over terskelverdien blir predikert å være positiv og omvendt. Dette er en vurdering som må tas fra situasjon til situasjon og måles opp mot kostnadene knyttet til de to feiltypene i hvert spesifikt tilfelle. Dersom det ikke foreligger en tydelig kostnad av å predikere riktig kontra å predikere feil, er det ifølge Habibzadeh et al. (2016) anbefalt å bruke en terskelverdi hvor sensitiviteten og spesifisiteten maksimeres. Dette tilsvarer punktet hvor $sensitivitet = spesifisitet$, altså skjæringspunktet mellom verdiene.

**Tabell 5.1:** Klassifikasjonsresultat ved binære klasser

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | Ekte positiv | Falsk negativ |
| | Nei | Falsk positiv | Ekte negativ |

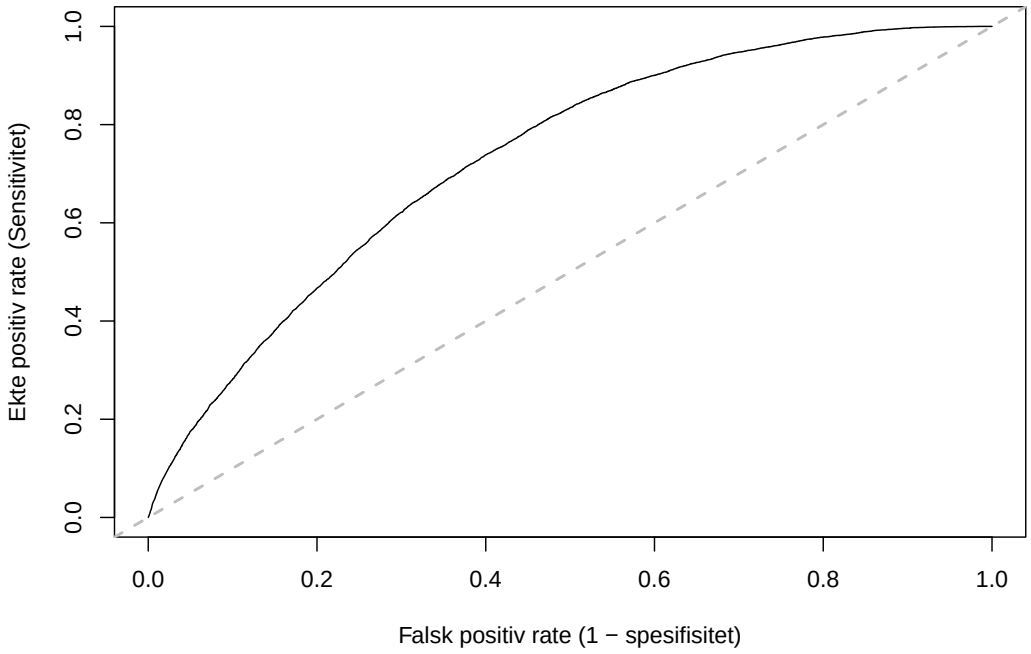
Et ROC plot kan brukes til å visualisere avveiningen mellom feiltypene. Y-aksen i plottet viser utviklingen i den ekte positive raten til modellen når terskelverdien endres. Dette omtales også som *sensitivitet* eller *recall*. X-aksen viser den falske positive raten til modellen. Dette er det samme som $(1 - spesifisitet)$. Målene viser henholdsvis sannsynligheten for ekte positiv prediksjon gitt at observasjonen faktisk er positiv (sensitivitet) og ekte negativ prediksjon gitt at observasjonen faktisk er negativ (spesifisitet). Sensitivitet og spesifisitet er definert i ligning 5.4 og 5.5 (Fawcett, 2006).

$$Sensitivitet = \frac{Ekte\ positive\ prediksjoner}{Totalt\ antall\ faktisk\ positive\ observasjoner} \tag{5.4}$$

$$Spesifisitet = \frac{Ekte\ negative\ prediksjoner}{Totalt\ antall\ faktisk\ negative\ observasjoner} \tag{5.5}$$

Et eksempel på en ROC kurve kan en se i figur 5.3. Settes dette i perspektiv til prediksjon av elbil-eiere vil Y-aksen representere antall elbil-eiere som er korrekt klassifisert og X-aksen representere hvor mange observasjoner som er feilklassifisert som elbil-eiere. En kan da se avveiningen mellom målene. Ønskes høyere sensitivitet, vil dette gå på bekostningen av lavere spesifisitet.

**Figur 5.3:** Eksempel på ROC-kurve




### 5.4.1 Areal under kurve (AUC) og balansert nøyaktighet

*Areal under kurve* (AUC) er arealet under ROC-kurven som vist i figur **??**. Ettersom nøyaktighet kan være misvisende ved bruk av ubalanserte datasett til prediksjoner, er det hensiktsmessig å bruke AUC. Grunnen til dette er at ROC-kurven ikke blir påvirket av endringer i proporsjonene mellom klassene i den avhengige variabelen. Den diagonale linjen representerer ROC kurven ved tilfeldig gjetning og vil ha en $AUC = 0,5$, som impliserer at ingen prediksjonsmodell bør ha $AUC < 0,5$. En modell som klassifiserer alle observasjoner perfekt vil ha en $spesifisitet = 1$ og en $sensitivitet = 1$. Det vil også sørge for at ROC kurven vil ligge helt oppe i venstre hjørne med en $AUC = 1$, som vil tilsvare en perfekt prediksjonsmodell (Fawcett, 2006).

I tillegg til AUC brukes *balansert nøyaktighet* som et supplerende prestasjonsmål for prediksjonsmodellene. Balansert nøyaktighet kan defineres som den gjennomsnittlige

nøyaktigheten innenfor hver klasse, vist matematisk i ligning 5.6. Ifølge Brodersen et al. (2010) er dette et mer nøyaktig prestasjonsmål enn nøyaktighet ved bruk av ubalanserte datasett. De viser også til at balansert nøyaktighet fjerner det feilaktige bildet nøyaktighet kan gi i forhold til modellenes prestasjon. Ved et fullstendig balansert datasett vil begge målene gi samme resultat.

$$\begin{aligned} &Balansert\ nøyaktighet = \\ &\frac{1}{2}\left(\frac{\sum Ekte\ positiv}{\sum Positive\ observasjoner} + \frac{\sum Ekte\ negativ}{\sum Negative\ observasjoner}\right) \end{aligned} \tag{5.6}$$

# 6 Empirisk analyse

Dette kapittelet vil inneholde analyser knyttet til datasettet presentert i kapittel 3 ved hjelp av metodene beskrevet i kapittel 5. Vi vil først beskrive hvordan vi velger ut variablene som blir brukt videre i analysen gjennom Boruta-algoritmen. Deretter vil vi forklare hvordan modellene er bygget opp og presentere resultatene av modellene før vi avslutningsvis sammenligner modellenes resultater.

I analysene vil det presenteres tre tidsperioder for hver av de tre metodene beskrevet i kapittel 5.2. Første tidsperiode anvender alle observasjoner i datasettet fra 2010 til 2015 og andre tidsperiode anvender alle observasjoner i perioden 2016 til 2017. Resultatene fra hele tidsperioden 2010 til 2017 vil ikke bli gjennomgått i analysen, men kan studeres i appendiks A4.

Denne oppdelingen er gjort for å kunne sammenligne separate perioder og se hvordan elbil-eiernes karakteristika har forandret seg mellom dem. Som en kan se fra kapittel 4 var det fra 2016 en spesiell økning i elbilsalget i Norge, som tyder på at første og andre periode er interessante å analysere hva gjelder eventuelle forandringer. Grunnet store mengder manglende informasjon blant flere forklaringsvariabler har vi valgt å ikke benytte observasjoner fra året 2018 i de kommende modellene. Dette kapittelet vil hovedsaklig ha fokus på forklaringsvariablenes viktighet og deretter modellenes prediksjonsevne.

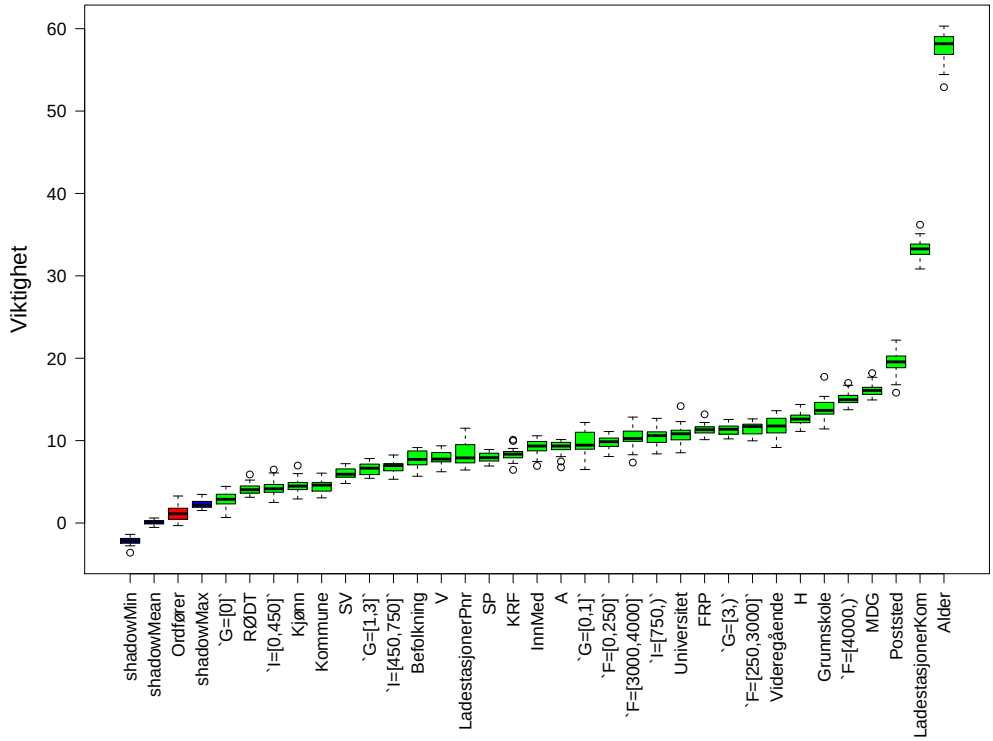
Alle modellene som utarbeides vil estimeres ved bruk av et estimeringssett som består av 70% av observasjonene i de forskjellige tidsperiodene. De resterende 30% av observasjonene vil brukes til å validere modellenes resultater. K-fold kryssvalidering benyttes for alle modeller for å optimalisere parametrene presentert i kapittel 5, utenom random forest som bruker out-of-bag estimater.

## 6.1 Variabelutvelgelse

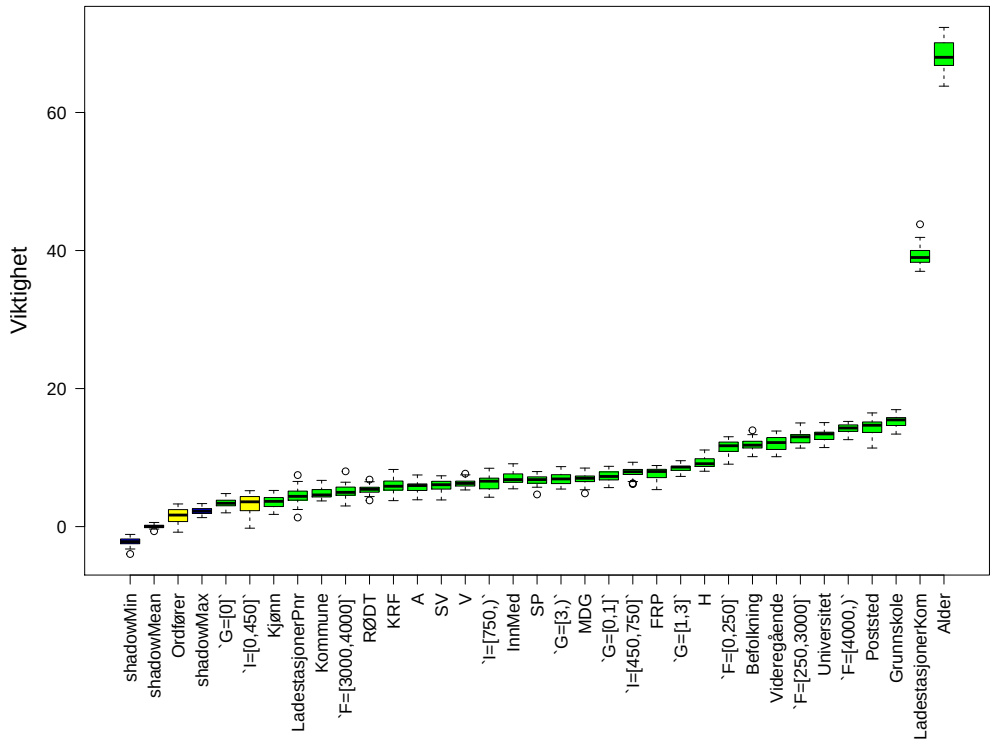
Det gjennomføres en variabelutvelgelse gjennom R-pakken *Boruta*. Ettersom det anvendes ulike tidsperioder, beregnes det et tilsvarende antall *Boruta*-modeller. Den avhengige variabelen settes som "Elektrisk" og alle forklaringsvariabler brukes i gjennomføringen hvor

det kjøres 30 iterasjoner. Resultatene fra utdata mellom årene 2010-2015 og 2016-2017 vises henholdsvis i figur 6.1 og 6.2.

**Figur 6.1:** Boruta-modell - tidsperiode 1




**Figur 6.2:** Boruta-modell - tidsperiode 2



Figurene viser de relative viktighetene til hver potensielle forklaringsvariabel til den avhengige variabelen "Elektrisk" for begge tidsperioder. X-aksen representerer hver av de potensielle forklaringsvariablene og Y-aksen viser Z-scoren eller viktighet til de ulike forklaringsvariablene. Grønn, gul og rød farge definerer hvorvidt en variabel er henholdsvis *viktig*, *ubestemt* eller *uviktig*. Blå farge tilsvarer minimum, gjennomsnittlig og maksimal Z-score til skyggevariablene. Av totalt 32 variabler, regnes 30 og 31 av disse i henholdsvis første og andre tidsperiode som *viktig*. For den første tidsperioden klassifiseres én variabel, "Ordfører", som *uviktig* av algoritmen. To variabler; "Ordfører" og "I=[0,450]", klassifiseres som *ubestemt* for den andre tidsperioden. Variablene "Alder" og "LadestasjonerKom" skiller seg spesielt ut som viktige i begge tidsperioder. Likevel, kan det fremstå overraskende at kun én variabel ble regnet som *uviktig*, og en kan stille seg spørrende til hvor godt *Boruta*-algoritmen passer datasettet. Det fremkommer i Kursa et al. (2010) at *Boruta* er en heuristisk metode, som søker å finne alle relevante variaber, inkludert svakt relevante variabler. Dermed er det viktig å presisere at mange av variablene som ble regnet som *viktig* kan være svakt relevante.

Sammenlignes disse resultatene med diskusjonen i kapittel 4.3, samstemmer resultatene om at variablene "Alder" og "LadestasjonerKom" var av stor betydning. Det er også interessant å se at "Poststed" har en større betydning enn "Kommune". Likevel kan det tenkes at disse variablene er korrelerte, hvor "Poststed" fanger opp enda mer informasjon enn "Kommune". På samme måte, er det interessant å se at variabler innenfor formue regnes som viktigere enn variabler innenfor inntekt, til tross for at Figenbaum og Kolbenstvedt (2016) anfører at elbil-eiere typisk har høy inntekt. Igjen kan det tenkes at variablene korrelerer og at formuesvariablene fanger opp ytterligere informasjon sammenlignet med inntektsvariablene. En kan også observere at variabelen "Kjønn" ikke har en relativt stor betydning, på tross av det som ble fremlagt i blant annet Figenbaum et al. (2014). Dette er ikke uventet jamfør tabell 4.1, hvor en kan se at fordelingen i kjønn er relativt lik mellom elbil-eiere og andre bileiere. Det forventes derfor ikke at denne variabelen vil kunne skille klassene på en god måte.

Selv om resultatene fra *Boruta*-modellene sier at vi skal benytte alle variabler som regnes som *viktig*, velger vi kun ut de ti viktigste variablene til bruk i prediksjonsmodellene som vist i tabell 6.1. Dette for å redusere modellenes kompleksitet og dermed forhindre

muligheten for overtilpasning, samt senke beregningstiden. Selv om dette kan gå på bekostning av nøyaktigheten, fant vi ved prediksjonsmodeller som benyttet alle variabler mot de ti viktigste variablene at denne forskjellen i nøyaktighet ville være marginal. Dette vil også kunne gjøre det enklere å tolke modellene. Videre i analysen deles variabelen "Alder" inn i fem ulike "Aldersgrupper" for å lettere identifisere hvilke aldersgrupper som er av betydning for prediksjonene. Dette vil kunne skape en liten reduksjon i prediksjonsevne, men til gjengjeld gi verdifull informasjon om viktigheten til ulike aldersgrupper for den avhengige variabelen.

**Tabell 6.1:** De ti viktigste forklaringsvariablene fra Boruta-modellene

| 2010-2017 | 2010-2015 | 2016-2017 |
|---|---|---|
| Alder | Alder | Alder |
| LadestasjonerKom | LadestasjonerKom | LadestasjonerKom |
| Poststed | Poststed | Grunnskole |
| F=[4000,) | MDG | Poststed |
| MDG | F=[4000,) | F=[4000,) |
| F=[250,3000] | Grunnskole | Universitet |
| Videregående | H | F=[250,3000] |
| Grunnskole | Videregående | Videregående |
| Universitet | F=[250,3000] | Befolkning |
| G=[3,) | G=[3,) | F=[0,250] |

Merk: Sortert etter viktigste variabel øverst

## 6.2 Tidsperiode 1: 2010-2015

Jamfør problemstillingen ønsker vi å finne faktorer som karakteriserer dagens elbil-eiere, og hvordan disse videre kan brukes til å predikere potensielle elbil-eiere. Vi benytter et enkelt klassifiseringstre som en basismodell for videre sammenligninger med random forest og extreme gradient boosting. Det må bemerkes at noen figurer videre i analysen vil kunne ha små avvik ved enkelte variabelnavn, grunnet begrensninger ved innlesning til noen av R-pakkene som brukes.

### 6.2.1 Klassifiseringstrær

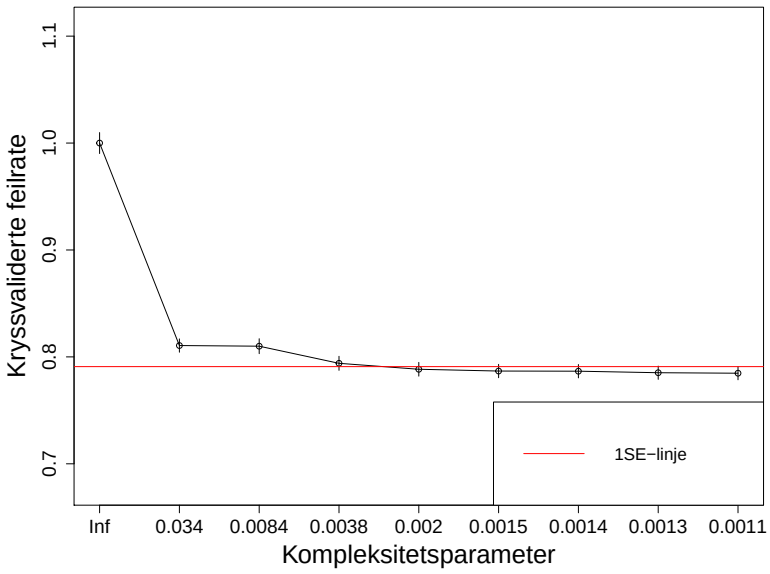
R-pakken *rpart* (Therneau et al., 1997) er brukt for å implementere klassifiseringstrær. Etter å ha tilpasset modellen på estimeringssettet for den første tidsperioden med

standardverdier for kompleksitetsparameteren $a$, forklart i kapittel 5.2.2, ble prediksjoner gjennomført på valideringssettet. Modellen predikerte innledningsvis likt modus, ettersom dette gir høyest mål på nøyaktighet. Følgelig gir dette et resultat lik tilfeldig gjetning, med en *Balansert nøyaktighet* = 50% og en påfølgende $AUC = 0,50$. Som beskrevet i metodekapittelet, er et av målene med utredningen å maksimere den balanserte nøyaktigheten og AUC. Disse ble ikke særlig forbedret ved å sette en lavere verdi for kompleksitetsparameter $a$.

Det ble dermed konkludert med at estimeringssettet var ubalansert, ettersom kun 10,5% av observasjonene var elbil-eiere i den gitte tidsperioden. Denne ubalansen kan forsterkes av at store deler av datagrunnlaget er innhentet på kommunenivå istedet for individnivå. Dette senker variansen i datasettet, som gjør det vanskeligere for prediksjonsmodellene å skille klassene. Derav ble det ilagt en vekting til den avhengige variabelens klasser i *rpart*, som fungerer på samme måte som å endre terskelverdiene på prediksjonene. I dette tilfellet legges denne vektingen allerede inni estimeringsmodellen, slik at valg av en terskelverdi senere blir overflødig. Vektingen påvirker sannsynlighetene for valget der nodene splittes, hvor det velges en splitt der det er en "god" sannsynlighet til å predikere elbil-eiere (Therneau et al., 1997). Som det fremkommer i kapittel 5.4, vektes klassene i den avhengige variabelen etter å maksimere sensitivitet og spesifisitet, ettersom det ikke er forbundet en direkte kostnad av å predikere den avhengige variabelen, "Elektrisk", riktig eller feil. Dette kan begrunnes med at det foreligger en like stor verdi for denne utredningen å finne hvem som er typiske elbil-eiere, mot hvem som ikke er det. Ettersom det foreligger en avveining mellom sensitivitet og spesifisitet, finnes skjæringspunktet i alle modeller som følger.

Vi finner skjæringspunktet i modellen med vektingen $0,41/0,59$ og standard kompleksitetsparameter $a = 0,001$, oppnås betydelig bedre prediksjonsresultater. Ettersom et stort og komplekst tre kan føre til overtilpasning, beskjæres treet gjennom en k-fold kryssvalidering av kompleksitetsparameteren. Den kryssvaliderte feilraten eller "risikoen" vist i figur 6.3, illustrerer et raskt fall i risiko når kompleksitetsparameteren senkes, hvor den flater ut i et platå og eventuelt ville steget. Dersom risikoen er innenfor et standardavvik til den oppnådde minimumsverdien til risikoen, er dette ekvivalent til å være minimumspunktet ettersom den anses å være en del av det flate platået (Therneau

**Figur 6.3:** Valg av klassifiseringstreets kompleksitet

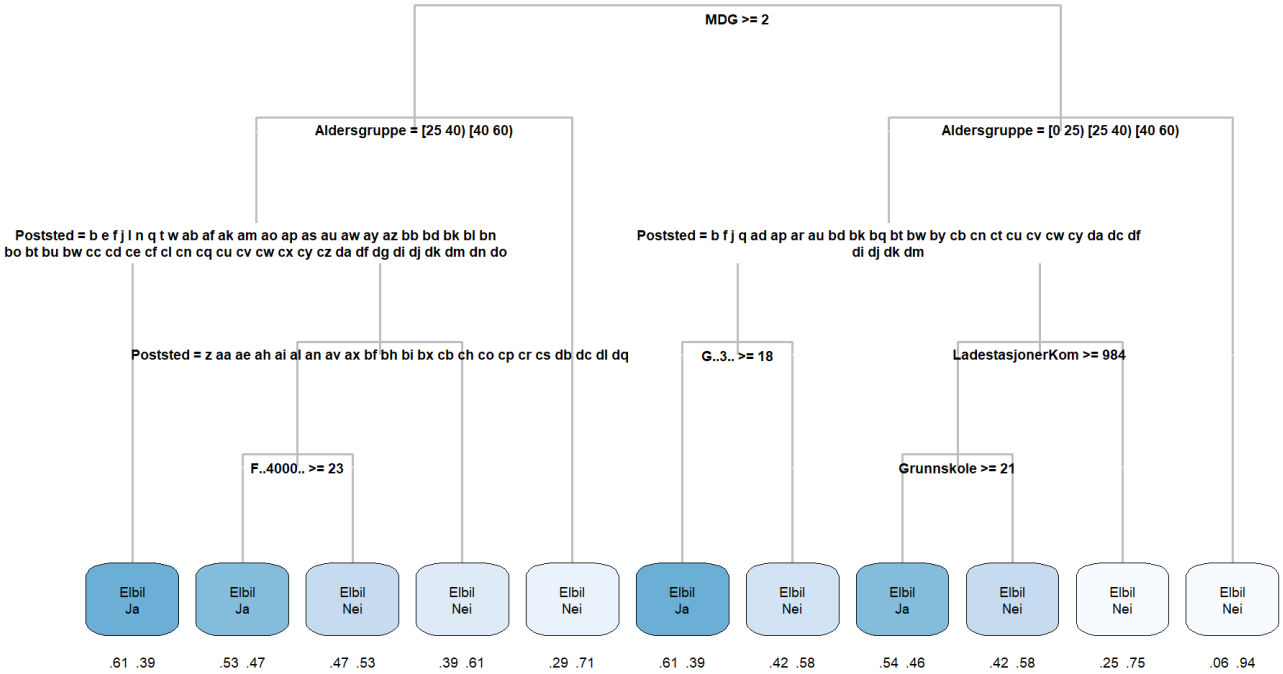



et al., 1997). Dette refereres også til som "1-SE"-regelen. Vi setter $a = 0,001$ som minste verdi for alle modellene selv om lavere verdier ville gitt en marginalt lavere risiko. Dette er hovedsakelig grunnet vanskeligheter med å beskjære treet tilstrekkelig med 1-SE-regelen for få et lite og forståelig klassifiseringstre.

Den ferdigbeskårede modellen, som vist i figur 6.4, med vektingen $0,41/0,59$ og kompleksitetsparameter $a = 0,0013894$ gir en *Balansert nøyaktighet* $= 67,34\%$ og $AUC = 0,7143$.

Av klassifiseringstreet kan en observere at variabelen "MDG" står øverst, som tilsier at denne variabelen i størst grad reduserer urenheten i nodene og dermed er den viktigste variabelen. Dersom en observasjon er fra en kommune hvor over 2% stemmer på MDG, vil dette øke sannsynligheten for en positiv prediksjon av den avhengige variabelen og gå mot venstre i treet. Neste internknutepunkt vil være variabelen "Aldersgruppe", med aldersgruppene [25,40] og [40,60] som øker sannsynligheten for elbileierskap. Dette er i tråd med figur 4.9 fra korrespondanseanalysen, hvor disse aldersgruppene var mest assosiert med elektriske biler. Går vi til neste internknutepunkt kan vi se variabelen "Poststed". Disse er formatert som poststed-koder, hvor tilsvarende poststeder kan finnes i tabell A5.1 i appendiks. Dersom observasjonen er innenfor et av disse postedene, er det overveiende sannsynlig at observasjonen er elbil-eier, og algoritmen predikerer *Ja* til "Elektrisk". Sannsynlighetene for de to klassene, 61% for *Ja* og 39% for *Nei*, viser sannsynlighetene

**Figur 6.4:** Klassifiseringstre - tidsperiode 1




for hver klasse i endenoden. Fargestyrken på endenodene tilsvarer sannsynligheten for en positiv verdi av den avhengige variabelen. Det kan også bemerkes at høye verdier for variabelen "Grunnskole" øker sannsynligheten for positive prediksjoner av den avhengige variabelen. Dette strider mot funnene i eksempelvis Figenbaum og Kolbenstvedt (2016), hvor det fremkommer at elbil-eiere har høyere utdanning.

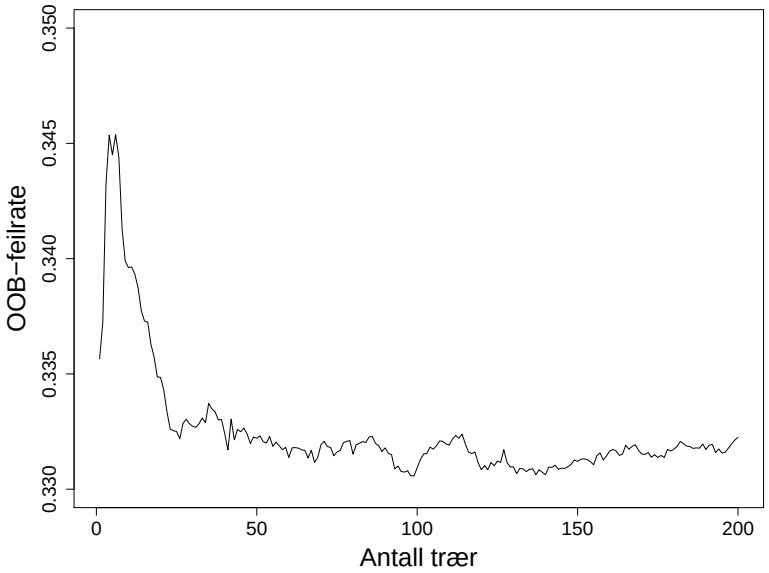
### 6.2.2 Random forests

Random forests ble implementert gjennom R-pakken *randomForest* (Liaw og Wiener, 2002). I tillegg har R-pakken *caret* (Kuhn et al., 2017) blitt brukt med formålet om å fintilpasse parametrene i modellene, for å optimalisere prediksjonsresultatene. Optimale parametre ble funnet ved å bruke estimeringssettet som datagrunnlag og maksimere på AUC med en 10-fold kryssvalidering gjennom et *tilfeldig søk*. Dette er en teknikk hvor tilfeldige kombinasjoner av parametre blir brukt til å finne det beste resultatet av modellen. Dette skiller seg fra *rutenettsøk* som bygger en modell for hver eneste definerte parameterkombinasjon og velger modellen som gir høyest nøyaktighet (Maladkar, 2018). For en bestemt modell, vil *caret* dermed gjennomføre et tilfeldig søk av parametre og trene modellen med litt forskjellige datasett for hver kombinasjon av parametre. For hvert

datasett, vil modellens AUC beregnes. Den kombinasjonen av parametre som gir høyest AUC er valgt ut, derav blir hele estimeringssettet brukt for tilpasning til den endelige modellen (Kuhn et al., 2017).

Ifølge *caret* var modellens optimale parametre $m = 69$ og $node.size = 18$. Selv om det kan virke kontraintuitivt at utvalget av forklaringsvariabler (69) kan være større enn det totale antall variabler (11) i estimeringssettet, er det viktig å poengtere at "caret" lager dummyvariabler av alle de kategoriske variablene før estimeringen (Kuhn et al., 2017). Derav blir det totale antallet variabler i estimeringssettet ikke 11, men 136. *node.size* er minimum antall observasjoner i et endepunkt, der et større antall observasjoner betyr et mindre tre. Gjennom å teste med en høy verdi for antall trær blir en verdi funnet for parameteren $B$ hvor OOB-feilraten har falt til et stabilt nivå. En kan se av figur 6.5 at $B = 200$ trær er tilstrekkelig.

**Figur 6.5:** Valg av antall trær - tidsperiode 1

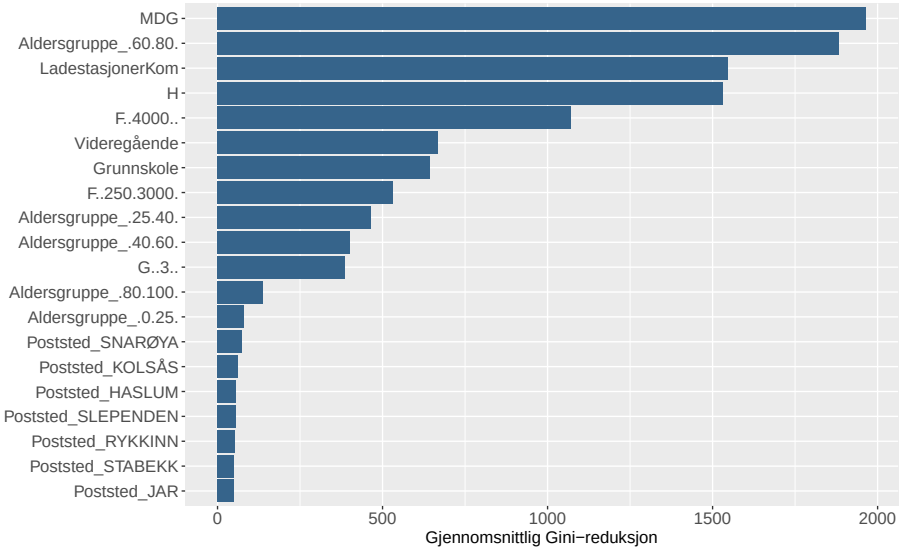



For at de fintilpassede parameterne fra *caret* skulle være kompatible med *randomForest* og ikke minst at tolkningen av hver forklaringsvariabel skulle bli bedre, ble estimerings- og valideringssettet modifisert med tilsvarende dummy-variabler som i *caret*. På denne måten kan en tolke viktigheten av hver klasse innad i de kategoriske variablene for den avhengige variabelen bedre, selv om dette går på bekostning av prediksjonsevnen. Eksempelvis er det nå mulig å ikke bare definere at "Poststed" er en viktig variabel, men også hvilke poststeder som er viktige relativt til hverandre. Dette vil også gjelde de andre kategoriske

variablene brukt til å utarbeide modellen.

Gjennom det modifiserte estimeringssettet tilpasses modellen med en $m = 69$, $node.size = 18$ og $B = 200$. I likhet med klassifiseringstrær, settes en vekting for å balansere estimeringssettet, her lik $0,44/0,56$. Ved å predikere på det modifiserte valideringssettet oppnås en $Balansert\ nøyaktighet = 67,57\%$ og en $AUC = 0,7236$. Viktigheten av hver variabel vises i figur 6.6.

**Figur 6.6:** Random forests: Variablenes viktighet - tidsperiode 1




I figuren representerer X-aksen den gjennomsnittlige Gini-reduksjonen, mens Y-aksen representerer de 20 viktigste variablene. Det er innledningsvis viktig å poengtere at kategoriske variabler som har blitt skilt ut i dummy-variabler, slik som "Aldersgruppe" og " Poststed" har fått relativt sett mindre viktighet per klasse. Disse vil samlet ha en større viktighet, slik det fremkommer i tabell 6.1. Det kan videre presiseres at viktigheten til forklaringsvariablene ikke sier noe om viktigheten innad i klassene til den avhengige variabelen. Et eksempel er variabelen "MDG". I likhet med klassifiseringstreet, kan en se at variabelen "MDG" er den viktigste variabelen og minimerer de samlede nodenes urenhet mest. En kan ikke ut fra denne figuren isolert sett si at en høy verdi av "MDG" vil trekke i retning av en positiv prediksjon for "Elektrisk". Figur 6.4 antyder likevel at dette er tilfellet ettersom en kan se at sannsynligheten for en positiv prediksjon av "Elektrisk" øker med høyere verdier av "MDG". Videre kan det bemerkes at aldersgruppen "[60,80)" er av viktighet. Det samme er poststedene "Snarøya", "Kolsås" og "Haslum", som forøvrig alle ligger i relativt urbane områder innenfor Bærum kommune. Sammenlignes dette med

resultatene fra figur 4.8 og 4.9 kan en se at Bærum trekker i retning elektriske biler og at den nevnte aldersgruppen trekker i motsatt retning. Dette tyder på at aldersgruppen "[60,80)" kan være en god indikator på at en person ikke eide en elektrisk bil i den gitte perioden. Det kan også tyde på at de nevnte poststedene er en god indikator på at en person eide en elbil i tidsperioden, relativt sett til de andre poststedene i datasettet. Også andre viktige variabler som "LadestasjonerKom", "H" og "F=[4000,)" øker sannsynligheten for positive prediksjoner i klassifiseringstreet i figur 6.4, som kan tyde på at disse variablene er viktige for å klassifisere elbiler-eiere i random forests.

### 6.2.3 Extreme gradient boosting

Extreme gradient boosting ble implementert ved hjelp av R-pakken *xgboost* (Chen et al., 2017). For å finne optimale parametre for metoden, ble det brukt sekvensielle rutesøk. Teknikken søker å finne optimale verdier for noen parametre om gangen og blir brukt for å senke antall modeller som måtte beregnes. Hver parameter hadde mellom tre og fem verdier. Med syv forskjellige parametere kunne det blitt opp mot $5^7 = 78125$ kombinasjoner av parametre, som hadde krevd for lang beregningstid til at det var gjennomførbart. Modellen ble først optimalisert for parametrene $d$, $\lambda$ og $B$. Disse tre parametrene viste seg å påvirke hverandre mye og det var derfor ønskelig å finne en optimal kombinasjon av dem. Videre ble optimale verdier for *gamma* og *min_child_weight* beregnet. Avslutningsvis, ble det gjennomført et rutenettsøk for å finne optimale verdier for *sub_sample* og *colsample_bytree*. Søkene ble gjennomført ved bruk av *caret* hvor også 5-fold kryssvalidering ble brukt for å validere parameterverdiene. Resultatet av parameteroptimeringen knyttet til den første tidsperioden er presentert i tabell 6.2.

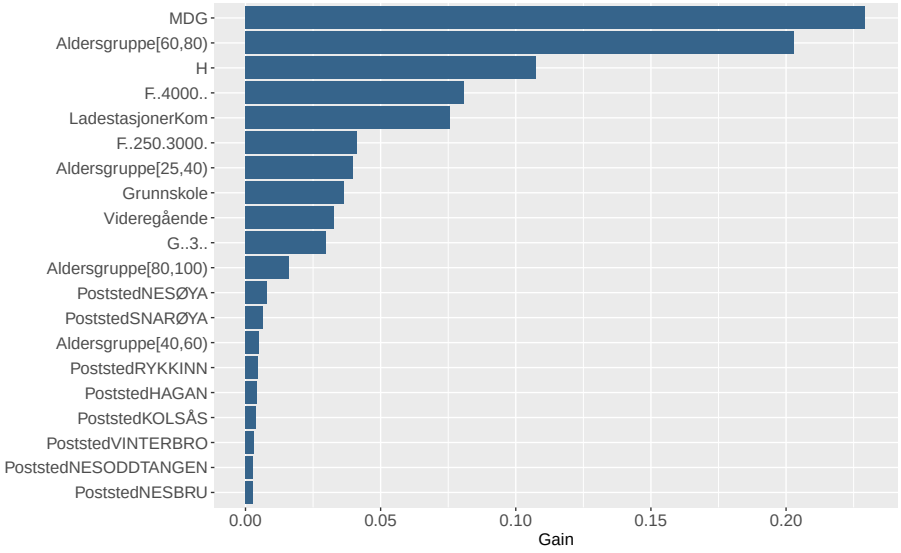
**Tabell 6.2:** Parameterverdier for extreme gradient boosting - tidsperiode 1

| Parameter | Verdi |
|---|---|
| B | 500 |
| $\lambda$ | 0.3 |
| d | 2 |
| Gamma | 0.1 |
| Min_child_weight | 5 |
| Sub_sample | 0.9 |
| Colsample_bytree | 1 |

I figur 6.7 kan en se rangeringen extreme gradient boosting gir til de 20 viktigste variablene

i modellen. Metoden deler opp alle kategoriske variabler som "dummier", ekvivalent med random forests. En kan igjen se at aldersgruppen "[60,80)" er av viktighet for "Elektrisk" også i denne modellen. Det samme gjelder poststedene "Nesøya", "Snarøya" og "Rykkin", hvor forøvrig alle er relativt urbane områder innenfor Asker eller Bærum kommune. Nok en gang kan en sammenligning med resultatene fra figur 4.8 og 4.9 vise at den nevnte aldersgruppen kan være en god indikator på at en person ikke er en elbil-eier i tidsperioden. På samme måte kan en se at de nevnte poststedene kan trekke i retning av at en person er en elbil-eier.

**Figur 6.7:** Extreme gradient boosting: Variablenes viktighet - tidsperiode 1



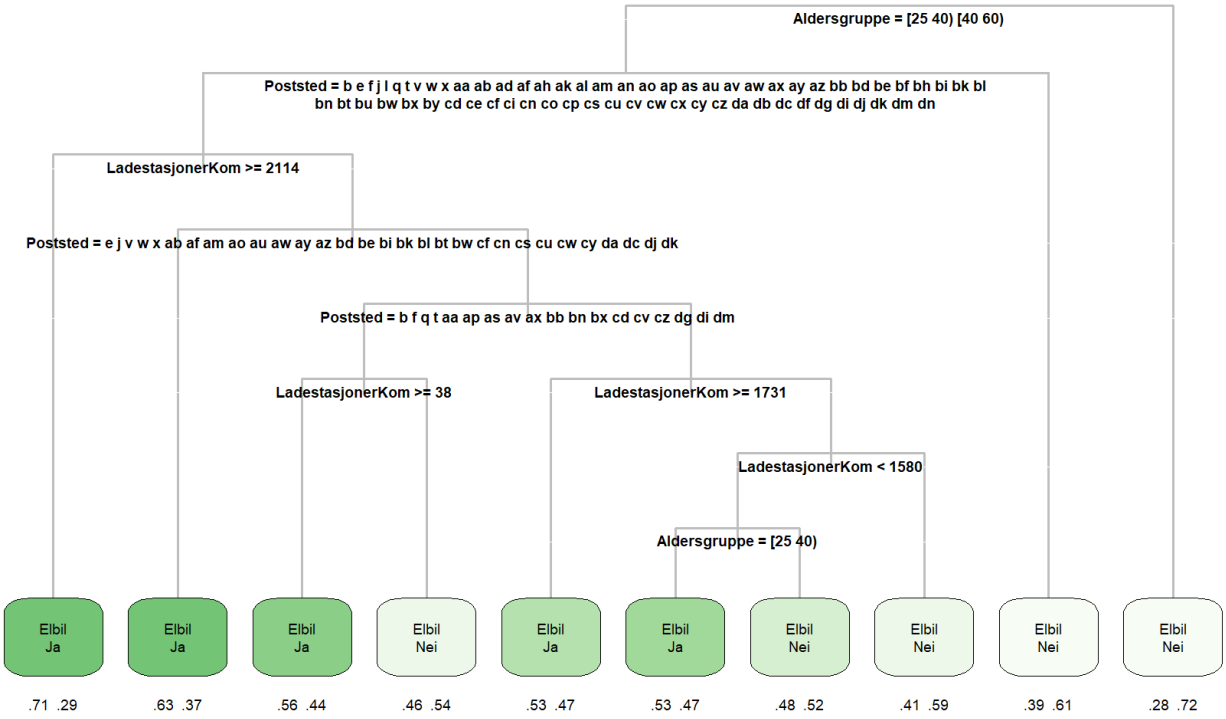

Ved å endre terskelverdiene knyttet til modellens prediksjoner, ble sensitivitet og spesifisitet balansert. Terskelverdiens optimale verdi ble kalkulert til 13,5%. Dette resulterte i en $Balansert\ nøyaktighet = 68,94\%$ og $AUC = 0,7496$.

## 6.3 Tidsperiode 2: 2016-2017

### 6.3.1 Klassifiseringstrær

Gjennom tilsvarende framgangsmåte som beskrevet i 6.2.1, ble optimal vekting og deretter kompleksitetsparameter funnet. Det ferdigbeskårede treet vist i figur 6.8, med vektingen $0,48/0,52$ og kompleksitetsparameter $a = 0,0017981$, gir en $Balansert\ nøyaktighet = 60,65\%$ og $AUC = 0,6450$.

**Figur 6.8:** Klassifiseringstre - tidsperiode 2



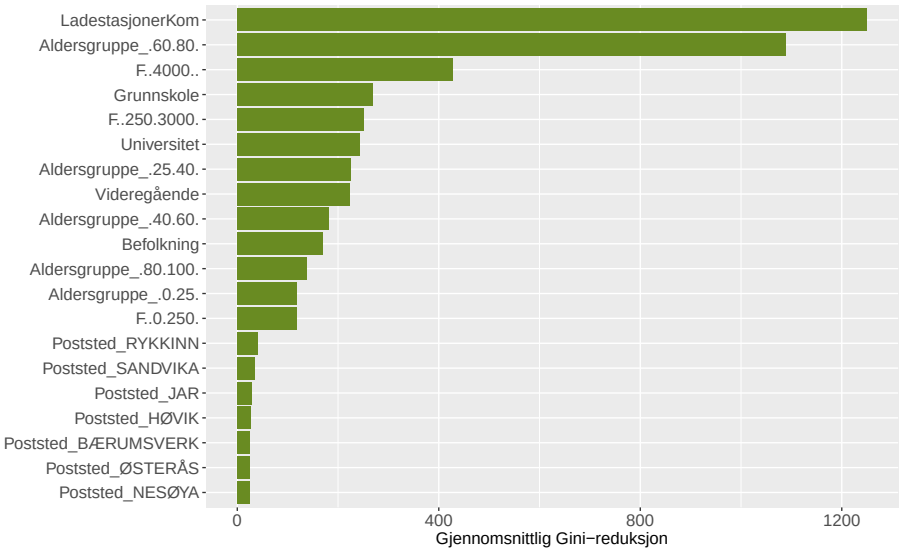

Vi finner her variabelen "Aldersgruppe" øverst i klassifiseringstreet, hvor aldersgruppene "[25,40]" og "[40,60]" igjen er forbundet med en økt sannsynlighet for positive prediksjoner av den avhengige variabelen. Det neste internknutepunktet er variabelen "Poststed", som nest viktigste variabel. Ved positiv verdi her, vil variabelen "LadestasjonerKom" være neste internknutepunkt som senker nodens urenhet mest. Dersom det er mer eller lik 2114 ladestasjoner i kommunen for observasjonen, vil klassifiseringstreet predikere *Ja* for "Elektrisk" med en sannsynlighet på 71%. Det kan bemerkes at Oslo kommune er den eneste kommunen med mer eller lik 2114 ladestasjoner. Videre er det spesielt iøynefallende at det er kun variablene "Aldersgruppe", "Poststed" og "LadestasjonerKom" treet er bygget opp av. Dette er likevel ikke overraskende, ettersom disse variablene er blant de fire viktigste som vist i tabell 6.1.

### 6.3.2 Random forests

På samme måte som forklart i 6.2.2, kalkuleres optimale parametre for random forest i andre tidsperiode. *caret* finner igjen samme parametre til å være optimale, med en $m = 69$ og $node.size = 18$. Det brukes en vekting på $0,48/0,52$, hvor $B = 200$ igjen er tilstrekkelig.

Ved å predikere på valideringssettet for tidsperioden, oppnås en $Balansert\ nøyaktighet = 60,85\%$ og en $AUC = 0,6541$.

**Figur 6.9:** Random forests: Variablenes viktighet - tidsperiode 2




De 20 viktigste forklaringsvariablene vises i figur 6.9. Her kan en innledningsvis observere at den gjennomsnittlige Gini-reduksjonen er lavere totalt sett, sammenlignet med første tidsperiode, vist i figur 6.6. Den viktigste variabelen i andre tidsperiode, "LadestasjonerKom", hadde eksempelvis en gjennomsnittlig Gini-reduksjon på litt over 1200. Sammenliknet med en gjennomsnittlig Gini-reduksjon på nesten 2000 for den viktigste variabelen i den første tidsperioden, kan dette tyde på at det i nyere tid er vanskeligere å finne gode forklaringsvariabler som skiller klassene i den avhengige variabelen. I praksis kan dette bety at elbil-eiere og andre bileiere har blitt likere over årene. Eksempelvis har også variansen til variabelen "Alder" blitt redusert med 7,65% fra første til andre tidsperiode. Videre kan vi igjen se at aldersgruppen "[60,80)" og "F=[4000,) er viktige variabler for den avhengige variabelen i andre tidsperiode. Nok en gang regnes poststeder med tilknytning til Bærum kommune som viktige, eksempelvis "Rykkinn", "Sandvika", "Jar" og "Høvik". Poststedet "Nesøya" ligger i Asker kommune, nabokommunen til Bærum. Det er dermed tydelig at kommuner vest for Oslo er viktige for å kunne spesifisere elbil-eiere.

### 6.3.3 Extreme gradient boosting

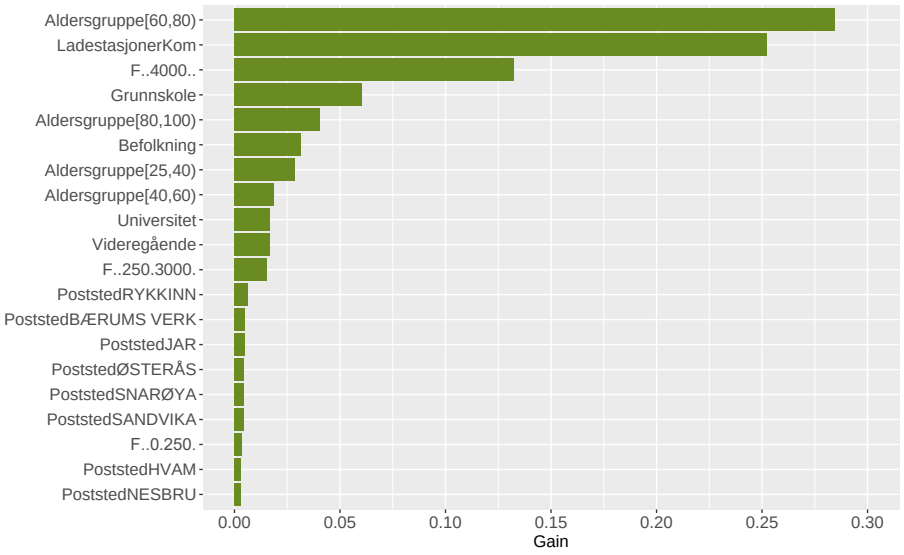
Ved bruk av samme fremgangsmåte som presentert i kapittel 6.2.3 ble optimale parameterverdier for den andre tidsperioden beregnet. Disse verdiene, vist i tabell 6.3, ble brukt i estimering av en modell for nyere tid.

**Tabell 6.3:** Parameterverdier for extreme gradient boosting - tidsperiode 2

| Parameter | Verdi |
|---|---|
| B | 2000 |
| $\lambda$ | 0.1 |
| d | 2 |
| Gamma | 0.1 |
| Min_child_weight | 5 |
| Sub_sample | 1 |
| Colsample_bytree | 1 |

Figur 6.10 viser de 20 variablene modellen angir som viktigst for prediksjonene. Fra plottet kan en se at også i denne tidsperioden er aldersgruppen "[60,80)" av høy viktighet. En kan også videre observere at antall ladestasjoner per kommune er blitt viktigere sammenlignet med den første tidsperioden. Nok en gang dominerer poststeder i Asker og Bærum kommune som de viktigste poststedene, eksempelvis "Rykkinn", "Bærums Verk", "Jar" og "Nesbru". En kan også observere at aldersgruppen "[40,60)" er blitt relativt viktigere sammenlignet med tidligere tidsperiode.

**Figur 6.10:** Extreme gradient boosting: Variablenes viktighet - tidsperiode 2



Etter at modellens sensitivitet og spesifisitet var balansert ble resultatet en $Balansert\ nøyaktighet = 61,72\%$ og $AUC = 0,6709$.

## 6.4 Modellsammenligning

### 6.4.1 Variablenes viktighet

Samlet sett, kan en observere at alle de tre metodene har mange likhetstrekk hva gjelder å velge ut viktige variabler. Aldersgruppen "[60,80)" har vært en viktig variabel uansett tidsperiode, som tyder på at denne aldersgruppen kan ha særlige homogene bileiere innenfor ikke-elektriske biler. På den andre siden har aldersgruppen "[40,60)" blitt relativt sett viktigere ved extreme gradient boosting for andre tidsperiode. Dette kan tilsi at det i dag er flere eldre som kjøper elbiler, relativt sett til den første tidsperioden. Random forests, på sin side, viser at aldersgruppen "[25,40)" har blitt relativt viktigere i den andre tidsperioden. Selv om begge aldersgruppene fører til høyere sannsynlighet for en positiv prediksjon, er det usikkert hvilke av aldersgruppenes viktighet som har økt mest.

Det kan også bemerkes at den relative viktigheten av "LadestasjonerKom" har steget, som kan tyde på at en økt utbygging av ladestasjoner i en kommune kan ha påvirket bilkjøperes valg av biltype. Denne sammenhengen kan også tenkes å gå andre veien, altså at det opprettes flere ladestasjoner i en kommune fordi flere i området kjøper elektriske biler. Det er også interessant å se at "MDG" og "H" var relativt viktige variabler for elbil-eiere for første tidsperiode, men at dette ikke lenger er tilfelle i den andre tidsperioden. Som tidligere nevnt betyr ikke dette nødvendigvis at elbil-eiere har sluttet å stemme på nevnte partier, men at den typiske elbil-eier tidligere var en som stemte "MDG" eller "H".

Videre kan det anføres at "F=[4000,)" har en viktigere betydning ved alle metoder og tidsperioder enn "F=[0,250]" og "F=[250,3000]". Ettersom høye verdier av førstnevnte gir en større sannsynlighet for en positiv prediksjon av "Elektrisk", vil en høy formue trolig være en viktig variabel når det kommer til å klassifisere elbil-eiere. Poststeder innenfor spesielt Asker og Bærum kommune har også vist seg å være viktige for den avhengige variabelen, for alle tidsperioder og metoder. Til tross for at prediksjonsevnen går ned i den andre tidsperioden, kan vi likevel si at disse geografiske områdene er gode

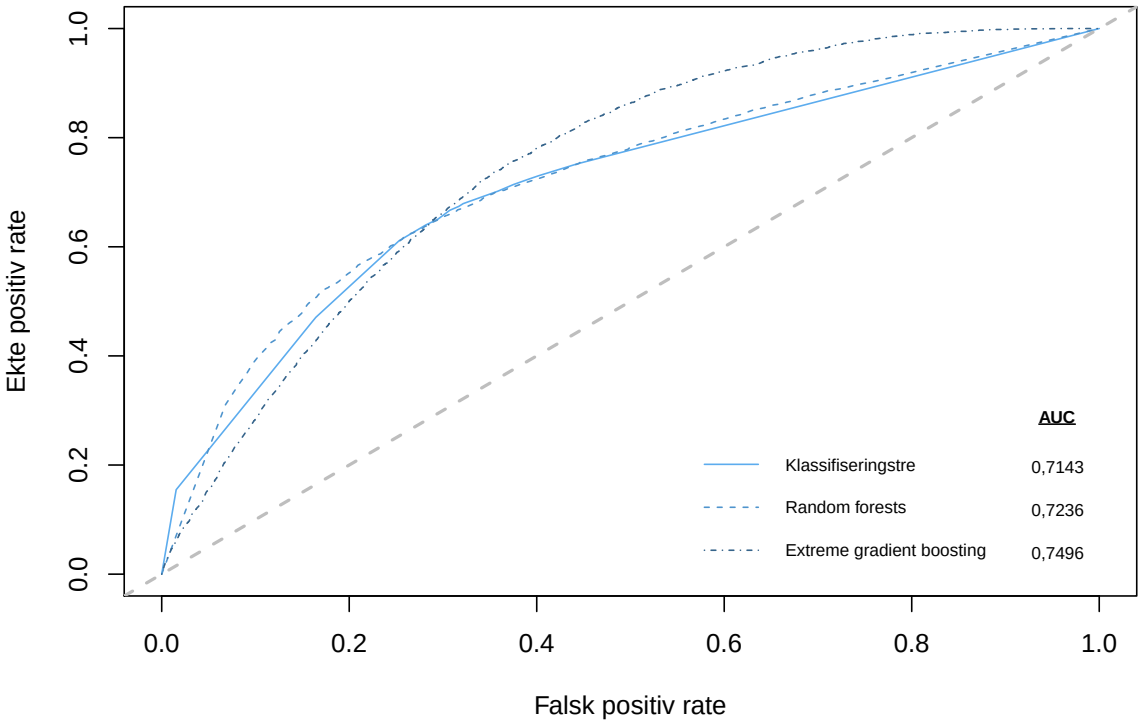
forklaringsvariabler for å predikere positivt for "Elektrisk". Innbyggerne i Asker og Bærum kommune er også blant de rikeste og høyest utdannede i landet (Hotvedt, 2014). Det kan tyde på at utdanning og formue er av betydning uansett tidsperiode.

Likevel, til tross for at Figenbaum og Kolbenstvedt (2016) finner at elbil-eiere typisk var personer med høyere utdanning i forhold til andre bileiere, er det vanskelig å se et mønster for dette i denne utredningen. Eksempelvis finnes motstridende funn i kapittel 6.2.1. Videre er både "Grunnskole" og "Videregående" representert som viktige variabler i begge tidsperioder, hvor "Universitet" først dukker opp som en viktig variabel i andre tidsperiode. I henhold til TØI-rapporten, skulle en forvente "Universitet" å være en god variabel for å skille klassene i den avhengige variabelen også i den første tidsperioden. Dette kan enten tyde på at høyere utdanning har mindre betydning enn tidligere antatt, eller at datagrunnlaget ikke er tilstrekkelig for å skille klassene nøyaktig når det gjelder utdanning.

### 6.4.2 Prediksjonsevne

Figur 6.11 illustrerer ROC-kurvene knyttet til alle modellene utarbeidet for den første tidsperioden. Grafen i seg selv viser at random forests presterer noe bedre enn basismodellen, klassifiseringstreet. Det observeres også at extreme gradient boosting presterer klart best blant modellene. Dette bekreftes av resultatene i tabell 6.4 hvor en ser at *AUC* og *balansert nøyaktighet* er høyere for denne metoden sammenlignet med random forests, som igjen er høyere enn basismodellen. Det må også poengteres at extreme gradient boosting predikerer mer "liberalt", altså gjør flere positive klassifiseringer med svake bevis. Dette gjør at metoden klassifsierer de fleste positive observasjonene korrekt, men også har en høy falsk positiv rate i forhold til både klassifiseringstrær og random forests. Disse to er på sin side mer "konservative", hvor de kun predikerer positivt om det er sterke bevis for det, som ofte fører til en lav falsk positiv rate og høy ekte positiv rate (Fawcett, 2006). Siden en høy ekte positiv rate gir flere korrekte prediksjoner av elbil-eiere, vil en slik modell også kunne være mer interessant da den gir mer informasjon om elbil-eiere. Ser en til figur 6.12 og tabell 6.5 observeres de samme resultatene for andre tidsperiode også. De spesifikke klassifikasjonsresultatene knyttet til første og andre tidsperiode kan finnes i henholdsvis appendiks A2 og A3.

**Figur 6.11:** ROC - tidsperiode 1




**Tabell 6.4:** Nøyaktighet og AUC - tidsperiode 1

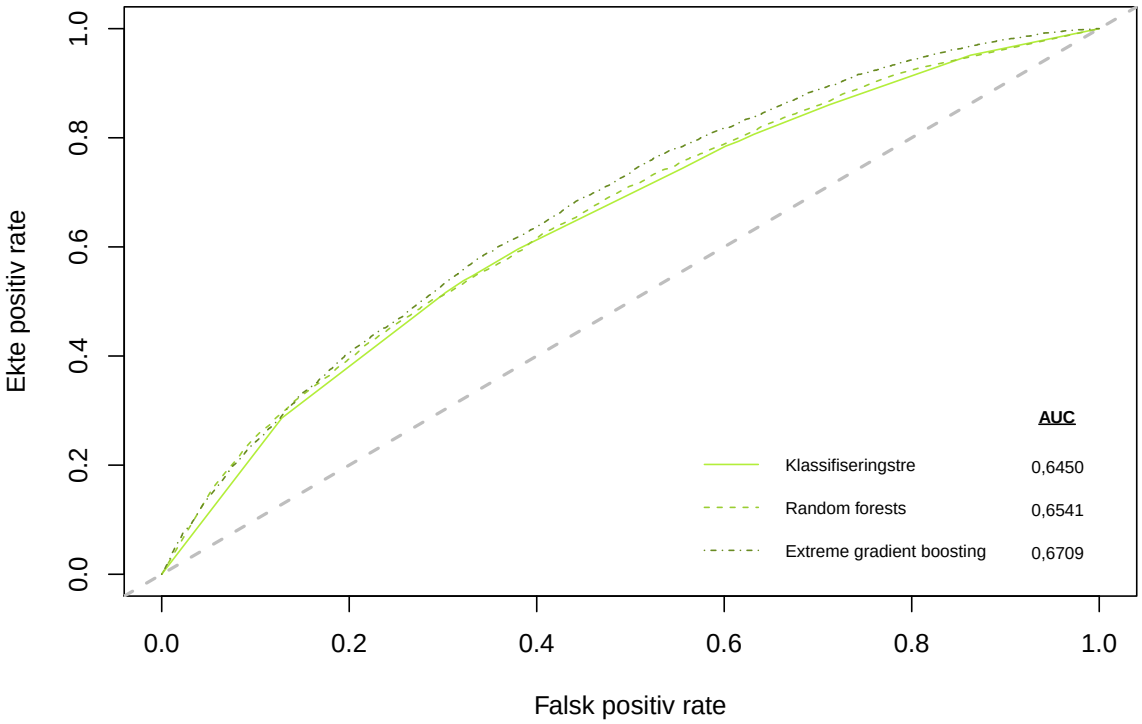
| Metode | Balansert nøyaktighet | AUC | Konfidensintervall AUC (95%) |
|---|---|---|---|
| Klassifiseringstre | 67,34% | 0,7143 | 0.7069 - 0.7218 |
| Random forests | 67,57% | 0,7236 | 0.7159 - 0.7312 |
| Extreme gradient boosting | 68,94% | 0,7496 | 0,7427 - 0,7564 |

Merk: Konfidensintervallene er kalkulert ved hjelp av R-pakken "pROC" (Robin et al., 2011)

Sammenlignes de to tidsperiodene kan en også klart se at samtlige modeller tilknyttet den første tidsperioden presterer betydelig bedre i forhold til den andre tidsperioden. Eksempelvis synker AUC for extreme gradient boosting fra nærmere 0,75 til 0,67 fra den første til den andre tidsperioden. I samme tidsintervall synker også den balanserte nøyaktigheten med gjennomsnittlig 6,87% for alle modeller. Dette er igjen et bevis på at de valgte modellene har større problemer med å klassifisere observasjonene korrekt i andre tidsperiode. Dette kan skyldes at elbil-eiere i andre tidsperiode er blitt likere andre bileiere, sammenlignet med første tidsperiode. En ser også at ROC-kurvene for den andre tidsperioden er mer symmetriske for alle metoder. Ingen av metodene kan regnes for å være verken "liberale" eller "konservative", selv om extreme gradient boosting igjen har en

marginalt høyere ekte positiv rate og falsk positiv rate. Dette kan være nok en indikasjon på at det kan være vanskeligere å skille elbil-eiere fra andre bileiere.

**Figur 6.12:** ROC - tidsperiode 2




**Tabell 6.5:** Nøyaktighet og AUC - tidsperiode 2

| Metode | Balansert nøyaktighet | AUC | Konfidensintervall AUC (95%) |
|---|---|---|---|
| Klassifiseringstre | 60,65% | 0,6450 | 0.6377 - 0.6522 |
| Random forests | 60,85% | 0,6541 | 0.6469 - 0.6614 |
| Extreme gradient boosting | 61,72% | 0,6709 | 0,6638 - 0.6780 |

Merk: Konfidensintervallene er kalkulert ved hjelp av R-pakken "pROC" (Robin et al., 2011)

# 7 Diskusjon

Det er fra analysene mulig å trekke ut flere faktorer som karakteriserer dagens elbil-eiere. Disse faktorene vises også å være viktige i prediksjonsmodellene for å kunne klassifisere potensielle elbil-eiere. Det er likevel nevneverdige forskjeller på faktorene som definerte disse i den første og andre perioden, noe som tyder på at det er vanskeligere å skille nyere elbil-eiere fra andre bileiere. I dette kapittelet diskuteres hva som kan være årsaken til dette og hvilke muligheter det kan skape. Videre vil det diskuteres hvorvidt dette kan påvirke dagens incentivordninger, om resultatene er gjeldende for hele landet, eventuelle begrensninger ved oppgaven og fremtidig forskning. Det er viktig å påpeke at diskusjonene vil kreve ytterlige analyse før det eventuelt kan trekkes endelige konklusjoner.

## 7.1 Diffusjon i elbilmarkedet

Som nevnt, er det flere antydninger til at elbil-eiere og andre bileiere er blitt likere gjennom de siste årene. Denne utviklingen kan blant annet forklares gjennom *Diffusjonsteorien* til Everett Rogers (2010). Teorien forklarer hvordan en idé eller et produkt øker i popularitet og diffuserer, eller sprer seg, gjennom et sosialt system. Resultatet av denne diffusjonen er at mennesker, som en del av et sosialt system, adopterer denne nye idéen eller produktet. Adopsjonen skjer ikke simultant i et sosialt system, men er heller en prosess hvor noen individer er mer egnede til å adoptere innovasjoner tidligere enn andre. Disse individene har ofte spesifikke karakteristika, hvor det finnes fem etablerte kategorier:

1. *Innovatører* (2-3% av populasjonen) - Karakteriseres som risikovillige, nysgjerrige og dristige individer som er svært interessert i å ta i bruk nye ideer og produkter.

2. *Tidlige brukere* (13-14% av populasjonen) - Karakteriseres som mindre risikovillige enn innovatører, men er også interessert i innovasjon. De liker lederroller og er klar over behovet for endring.

3. *Tidlig majoritet* (35% av populasjonen) - Karakteriseres sjeldent som ledere, men adopterer nye ideer tidligere enn gjennomsnittet. Krever ofte bevis for at en innovasjon fungerer før de adopterer.

4. *Sen majoritet* (35% av populasjonen) - Karakteriseres som endringsmotvillige individer, og vil kun ta i bruk en innovasjon dersom majoriteten har gjort det.

5. *Etternølere* (15% av populasjonen) - Karakteriseres som konservative og tradisjonelle individer. Disse er veldig skeptiske til endring og er den vanskeligste gruppen å overbevise.

Ifølge en studie utført av Figenbaum et al. (2014) på det norske markedet mellom årene 2009-2013, fremkommer det at *tidlige brukere* av elbiler passet godt inn i nevnte kategori. Likevel anføres det videre at forskjellene mellom elbil-eiere og andre bileiere blir stadig mindre som diffusjonen modner i Norge. Spesielt er dette gjeldende innenfor kjønn, som kan være grunnen til at variabelen "Kjønn" ikke betegnes som en av de viktigste variablene uansett tidsperiode i analysen. Denne diffusjonen gjør seg også gjeldende ved metodenes prediksjonsevner, hvor samtlige metoder har større vanskeligheter med å finne gode forklaringsvariabler som kan skille klassene i den avhengige variabelen i andre periode. Dermed er det en mulighet for at elbilmarkedet har beveget seg fra *tidlige brukere* til en *tidlig majoritet*, og beveger seg raskt mot *sen majoritet*. Markedsandelen av elbiler var på 0% i 2010 og har vokst til over 20% i 2017, som ut fra teorien til Rogers (2010) tilsier at diffusjonen har modnet utover *tidlige brukere* til *tidlig majoritet*. Prognosen for 2018 anslår at elbiler vil ha en markedsandel på mellom 35-45%, som vil bety at diffusjonen potensielt vil nærme seg *sen majoritet* (Opinion for Norsk elbilforening og Nordisk Energiforskning, 2018).

Ettersom diffusjonen modner, kan det ikke bli tatt for gitt at nyere elbil-eiere kommer til å ha samme komposisjon som de tidlige elbil-eierne. Dette kan være særlig interessant for bilprodusenter i forhold til markedsføringen av nye elbil-modeller. Egenskaper ved elbilene som en gang appellerte til de første elbil-eierne, appellererer nødvendigvis ikke til nyere elbil-eiere. Eksempelvis var variabelen "MDG" en relativt viktigere variabel for å karakterisere elbil-eiere i første tidsperiode, som potensielt gjorde det strategisk å markedsføre elbilene mot mer miljøbevisste konsumenter. Dette er trolig ikke en like effektiv strategi i nyere tid, med en mer heterogen kundegruppe fra *tidlig majoritet* og *sen majoritet* som kan tenkes å være mindre miljøbevisste. Ettersom disse gruppene ifølge Rogers (2010) trenger henholdsvis bevis for at en innovasjon fungerer og at majoriteten allerede har tatt det i bruk, vil det kunne være fordelaktig å rette markedsføringen av nyere

modeller mot dette. Det fremkommer i Figenbaum (2018) at en stadig mindre andel kjøper elbil som sekundærbil, trolig grunnet introduksjonen av elbiler med bedre rekkevidde, forbedret infrastruktur av ladestasjoner og et mer etablert bruktbilmarked for elbiler. Det kan dermed være mer hensiktsmessig for bilprodusenter å velge en bredere målgruppe av konsumenter. Markedsføringen kan videre fokuseres på å overbevise *tidlig majoritet* og eventuelt *sen majoritet* til at en elbil er tilstrekkelig som primærbil. Eksempelvis kan det vises til at de nyeste elbilene i markedet med lengst rekkevidde dekker 360 dager av det gjennomsnittlige kjørebehovet for ett år (Figenbaum, 2018).

## 7.2 Incentivordninger

Som det fremgår i kapittel 2 finnes det en rekke incentiver som har gjort elbiler økonomisk gunstige sammenlignet med andre biltyper. Ettersom disse incentivene i hovedsak ble utformet mellom 1990-2011, før introduksjonen av mer luksøriøse elbiler slik som Tesla Model S, kan det argumenteres for at det innledningsvis var mer fordelaktig for samfunnet at flere kjørte elbiler enn for konsumentene. I senere tid kan denne utviklingen sies å ha reversert seg, hvor teknologiutviklingen og økt konkurransen har ført til introduksjon av elbiler som kan matche andre biler på blant annet komfort, sikkerhet, utstyr og kjøreegenskaper. Likevel har disse incentivene, som har vært en viktig bidragsyter for å øke elbilsalget fram til i dag, blitt beholdt i stor grad (Figenbaum og Kolbenstvedt, 2013). Det kan dermed diskuteres hvorvidt disse incentivene, som var formet i en periode hvor kun *innovatører* kjøpte elbiler, gjenspeiler bevegelsen i elbilmarkedet i dag. Det kan argumenteres for at det i dag har blitt i overkant fordelaktig for konsumentene enn for samfunnet å kjøre elbiler. En økt andel elbiler løser eksempelvis ikke et trafikkproblem, hvor disse bilene også danner køer som er med på å skape økt forurensning. Videre bidrar også elbiler til at det slippes ut svevestøv, gjennom slitasje på veier. På den andre siden har elbiler fremdeles et klart miljømessig fortrinn sammenlignet med andre biler med vanlige forbrenningsmotorer, som bidrar til at Norge når sine internasjonale klimaforpliktelser som nevnt i 2.2.

Opinion for Norsk elbilforening og Nordisk Energiforskning (2018) har gjennomført en studie knyttet til nordiske konsumenter og deres forhold til elbiler. Her kommer det frem at i de andre nordiske landene, hvor de økonomiske incentivene knyttet til elbiler er

svakere, at betydelige færre konsumenter ønsker å kjøpe elbil som sitt neste kjøretøy. Det kommer også frem at pris er et mye større hinder her sammenlignet med Norge. Det trekkes ofte likhetstrekk mellom de nordiske landene og kjøpsmønster. En kan derfor forvente at også pris vil bli et betydelig større hinder her til lands om incentivene i Norge svekkes tilstrekkelig. På den andre siden har Norge et godt utviklet ladestasjonnettverk. Som denne utredningen antyder er antall ladestasjoner i nærheten av kjøpernes bosted av økende viktighet. Dette kan trekke i motsatt retning og sørge for betydelig fremtidig elbilsalg i Norge selv om prisene skulle øke. Ifølge McKerracher et al. (2017) vil elektriske biler være konkurransedyktige på pris, usubsidiert, fra og med 2024. Dette vil kunne fjerne mye av problemet knyttet til pris som påpekes av Opinion for Norsk elbilforening og Nordisk Energiforskning (2018), selv om økonomiske incentiver fjernes. Det er likevel viktig å ta i betrakting at det blir vanskeligere å overbevise *sen majoritet* og *etternølerene* til å bytte til elbiler. Det kan dermed fremdeles være hensiktsmessig for staten å beholde flere av dagens incentiver uten å gjøre for mange store forandringer. Resultatene fra analysen tyder på at elbil-incentivene har fungert godt frem til nå.

## 7.3 Generalisering

Datagrunnlaget vårt består av totalt 281 148 observasjoner hvor 93 341 er kvinner og 187 807 er menn. Alderen tilknyttet observasjonene varierer mellom 1 og 96 år[7]. På grunnlag av dette og resultatene fra kapittel 4.1 kan vi påstå at datagrunnlaget demografisk sett representerer den norske befolkningen som en helhet på en god måte. Observasjonene er derimot kun knyttet til personer registrert som boende i Oslo og Akershus per 11. september 2018. Det geografiske spennet representerer derfor ikke helheten i landet. De nevnte fylkene har også en større grad av urbanisering sammenlignet med flere andre fylker. Det kan derfor være vanskelig å påstå at resultatene fra våre analyser er generaliserbare for hele landet. De større byene i Norge har derimot store likhetstrekk, eksempelvis knyttet til urbanisering, utdanning og inntekt, sammenlignet med Oslo og Akershus. Det vil derfor være mulig å trekke slutninger knyttet til disse områdene basert på analysene i denne utredningen (Thorsnæs, 2018).

[7]En bil kan være registrert på personer under 18 år så lenge dette er godkjent av deres foresatte (Justis- og beredskapsdepartementet, 1999).

## 7.4 Begrensninger

Som nevnt, er observasjonene i datasettet fra analysen basert på bilregistreringer og ikke salgstall. Dette fører til et ukorrekt antall observasjoner ettersom biler eksempelvis er kondemnert, eksportert eller hvor personer har tilflyttet/fraflyttet Oslo og Akerhus. Optimalt, ville en slik analyse som er gjennomført brukt faktiske salgstall. Likevel, grunnet mangel på midler og tidsbegrensninger knyttet til å skaffe nødvendig datagrunnlag, baserer analysen seg på bilregistreringer. Antallet observasjoner blir også begrenset etter hvordan leasingbiler registreres hos Statens Vegvesen, som forklart i kapittel 3.6.

Deler av datagrunnlaget er også basert på observasjoner på kommunenivå. Dette senker variansen i datasette og kan påvirke resultatene. Optimalt ville informasjon på individnivå vært foretrukket. Denne typen informasjon kan være kostbar å samle inn, og vanskelig å få tak i da mye av dette kan være privat. Et slikt datagrunnlag kunne potensielt forbedret prediksjonsevnen til modellene betydelig og gitt økt innsikt i problemstillingen. Det må også nevnes at deler av datagrunnlaget er estimert ved hjelp av lineær regresjon og ikke er faktiske observasjoner, som forklart i kapittel 3.3.2. Dette vil også til en viss grad senke variasjonen i datagrunnlaget, samt gjøre modellene mindre robuste.

## 7.5 Fremtidig forskning

Prediksjoner av elbil-eiere er et forholdsvis nytt temaområde, hvor det er mange muligheter for videre forskning. Eksempelvis kan det videre være interessant å gjennomføre en lignende studie knyttet til hele landet samlet, eventuelt andre regioner. På den måten vil en finne ut om resultatene i denne analysen er generaliserbare til resten av Norge. Det ville også vært interessant å gjøre lignende analyser for områder som ikke har hatt samme fokus på å fremme elektriske biler. Dette kan eksempelvis være andre nordiske eller europeiske land. Derav vil det være mulig å analysere hvordan incentivene implementert i Norge faktisk har påvirket konsumentene. Videre kan en gjennomføre flere studier på viktigheten av antall ladestasjoner i et område. En kan knytte en geolokasjon til hver enkelt elbil-eiers adresse, samt alle ladestasjonene og på den måten få et mer nøyaktig mål på hvor mange ladestasjoner som er i nærheten av hver enkelt elbil-eier. Kausaliteten mellom

antall ladestasjoner og elbilkjøp, om det kommer flere ladestasjoner fordi det er flere elbiler eller omvendt, kan også være av interesse for videre studier. Undersøkelser rundt priselastisiteten i det norske elbilmarkedet kan komme med verdifull innsikt til tematikken. Slike analyser kan i større grad opplyse hvor stor effekt forandringer i de økonomsike elbil-incentivene vil ha. Det kan også være en indikator på hvordan markedet vil reagere om prognosen til McKerracher et al. (2017) blir en realitet.

# 8 Konklusjon

Hovedformålet med denne oppgaven har vært å finne faktorer som karakteriserer dagens elbil-eiere og undersøke hvordan disse faktorene kunne brukes videre til å predikere potensielle elbil-eiere. Gjennom bruk av datakilder fra Statens Vegvesen, OFV, SSB, Norsk Elbilforening og Valgdirektoratet har vi studert hvilke faktorer som best kan klassifisere elbil-eiere.

Sammenlignet med tidligere viser utredningens analyse at elbil-eiere har forandret karakteristika i nyere tid. Resultatene viser at politisk ståsted har mindre innflytelse på å karakterisere elbil-eiere, mens antall ladestasjoner i nærheten av kjøperen har fått relativt større viktighet. Høy formue og poststed vises å være viktige for alle perioder og modeller. Utdanningsvariablene blir generelt regnet som viktige forklaringsvariabler, likevel er tolkningen av betydningen uklar. Dette gjelder også hvilke aldersgrupper som er mest fremtredende i markedet i nyere tid, hvor aldersgruppen 25-40 år og 40-60 år begge øker sannsynligheten for positive prediksjoner. Flere tidligere studier viser til at eierens kjønn har vært en viktig karakteristikk blant elbil-eiere, dette finner ikke denne utredningen støtte for.

Faktorene som fremkommer i analysen viser seg å ha dårligere prediktiv evne i nyere tid sammenlignet med tidligere. Dette kan skyldes at elbil-eiere er blitt likere andre bileiere og derfor er vanskeligere å skille ved hjelp av prediktive modeller. En kan med bakgrunn i analysens resultater anta at elbiler har blitt en vanligere form for kjøretøy og at det norske markedet er på vei inn i en modnere fase med mer heterogene bileiere.

# Referanser

Bock, T. (2018). How correspondence analysis works (a simple explanation). Hentet 2. desember 2018, fra: https://www.displayr.com/how-correspondence-analysis-works/.

Brodersen, K. H., Ong, C. S., Stephan, K. E., og Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. I *Pattern recognition (ICPR), 2010 20th international conference on*, sider 3121–3124. IEEE.

Cai, J., Luo, J., Wang, S., og Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

Chen, T. og Guestrin, C. (2016). Xgboost: A scalable tree boosting system. I *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, sider 785–794. ACM.

Chen, T., He, T., Benesty, M., Khotilovich, V., og Tang, Y. (2017). *xgboost: Extreme Gradient Boosting*. R package version 0.6-4.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Figenbaum, E. (2018). Electromobility status in norway: Mastering long distances - the last hurdle to mass adoption. *TØI, rapport 1627/2018*.

Figenbaum, E. og Kolbenstvedt, M. (2013). Electromobility in norway-experiences and opportunities with electric vehicles. *TØI, rapport 1281/2013*.

Figenbaum, E. og Kolbenstvedt, M. (2016). Learning from norwegian battery electric and plug-in hybrid vehicle users: Results from a survey of vehicle owners. *TØI, rapport 1492/2016*.

Figenbaum, E., Kolbenstvedt, M., og Elvebakk, B. (2014). Electric vehicles – environmental, economic and practical aspects. As seen by current and potential users. *TØI, rapport 1329/2014*.

Finansdepartementet (2007). Omlegging av bilavgiftene i mer miljøvennlig retning. Hentet 16. oktober 2018, fra: https://www.statsbudsjettet.no/Statsbudsjettet-2007/Artikler/bilavgift/.

Friedman, J., Hastie, T., og Tibshirani, R. (2008). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407.

Gjerde, R. (16. april 2008). Dieselbil? Nå får du regningen. Hentet 8. oktober 2018, fra: https://e24.no/naeringsliv/dieselbil-naa-faar-du-regningen/2371013.

Grace-Martin, K. (20. januar 2017). In principal component analysis, can loadings be negative? Hentet 18. desember 2018, fra: https://www.theanalysisfactor.com/principal-component-analysis-negative-loadings/.

Greenacre, M. J. (1984). Theory and applications of correspondence analysis.

Habibzadeh, F., Habibzadeh, P., og Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica: Biochemia medica*, 26(3):297–307.

Hagman, R. og Kolbenstvedt, M. (2018). Elektrifisering av bilparken. Hentet 7. desember 2018, fra: https://www.tiltak.no/c-miljoeteknologi/c1-drivstoff-og-effektivisering/c-1-4/.

Haugneland, P., Bu, C., og Hauge, E. (2016). The norwegian ev success continues. I *EVS29 Int. Batter. Hybrid Fuel Cell Electr. Veh. Symp.*, sider 1–9.

Hotvedt, S. K. (17. oktober 2014). I disse kommunene tjener folk mest – se hele lista over alle 427 kommuner. Hentet 30. november 2018, fra: https://www.nrk.no/norge/rikeste-og-fattigste-kommuner-1.11991815.

IBM (2018). Predictor importance. Hentet 11. desember 2018, fra: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/model_nugget_variableimportance.htm.

James, G., Witten, D., Hastie, T., og Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Justis- og beredskapsdepartementet (22. juni 1999). § 55 - spørsmål om motorvogn kan registreres på en mindreårig. Hentet 18. desember 2018, fra: https://www.regjeringen.no/no/dokumenter/-55---vergemalsloven---sporsmal-om-motor/id455102/.

Klima- og miljødepartementet (2017). Lov om klimamål (klimaloven). Hentet 6. desember 2018, fra: https://lovdata.no/dokument/NL/lov/2017-06-16-60.

Kuhn, M. og Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

Kuhn, M., Wing, J., Weston, S., et al. (2017). *Caret: Classification and Regression Training*. R package version 6.0-78.

Kursa, M. B., Rudnicki, W. R., et al. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13.

Lê, S., Josse, J., og Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Liaw, A. og Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Louppe, G., Wehenkel, L., Sutera, A., og Geurts, P. (2013). Understanding variable importances in forests of randomized trees. I *Advances in neural information processing systems*, sider 431–439.

Maladkar, K. (14. juni 2018). Why is random search better than grid search for machine learning. Hentet 28. november 2018, fra: https://www.analyticsindiamag.com/why-is-random-search-better-than-grid-search-for-machine-learning/.

McKerracher, C. et al. (2017). Electric vehicle outlook 2018. *Bloomberg New Energy Finance*.

Miljødirektoratet (3. januar 2018). Klimagassutslipp fra transport. Hentet 6. desember 2018, fra: http://www.miljostatus.no/tema/klima/norske-klimagassutslipp/utslipp-av-klimagasser-fra-transport/.

Miljøverndepartementet (25. april 2012). Norsk klimapolitikk. Hentet 7. desember 2018, fra: https://www.regjeringen.no/no/dokumenter/meld-st-21-2011-2012/id679374/sec1.

NASA (2018). Scientific consensus: Earth's climate is warming. Hentet 6. desember 2018, fra: https://climate.nasa.gov/scientific-consensus/.

Natrella, M. (2010). *NIST/SEMATECH e-handbook of statistical methods*. NIST/SEMATECH.

Nenadic, O. og Greenacre, M. (2007). Correspondence analysis in r, with two-and three-dimensional graphics: The ca package. *Journal of statistical software*, 20(3).

Nervik, S. og Larsen-Vonstett, Ø. (12. januar 2012). Dieselbløffen. Hentet 8. oktober 2018, fra: https://www.vg.no/forbruker/bil-baat-og-motor/i/AonO3/dieselbloeffen.

Ng, M., Law, M., og Zhang, S. (2018). Predicting purchase intention of electric vehicles in hong kong. *Australasian Marketing Journal (AMJ)*.

Norsk Elbilforening (2018). Informasjon om nobil. Hentet 12. oktober 2018, fra: http://info.nobil.no/index.php/om#.

Opinion for Norsk elbilforening og Nordisk Energiforskning (2018). Elbilbarometeret 2018. Hentet 2. desember 2018, fra: https://elbil.no/elbilstatistikk/elbilbarometeret/.

Opplysningsrådet for veitrafikken (2013). Bilsalget i desember og hele 2012. Hentet 19. oktober 2018, fra: http://www.ofvas.no/bilsalget-i-desember/category547.html.

Opplysningsrådet for veitrafikken (2016). Bilsalget i 2015. Hentet 22. oktober 2018, fra: http://www.ofvas.no/bilsalget-i-2015/category679.html.

Opplysningsrådet for veitrafikken (2018a). Bilsalget i 2017. Hentet 22. oktober 2018, fra: http://www.ofvas.no/bilsalget-i-2017/category751.html.

Opplysningsrådet for veitrafikken (2018b). Bilåret 2017, OFV frokostmøte 3. januar 2018.

Pearson, E. S. (1931). The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, 26(174):128–134.

Priessner, A., Sposato, R., og Hampl, N. (2018). Predictors of electric vehicle adoption: An analysis of potential electric vehicle drivers in austria. *Energy Policy*, 122:701–714.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., og Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Skatteetaten (2018a). Eierskifte. Hentet 7. desember 2018, fra: https://www.skatteetaten.no/person/avgifter/bil/eierskifte/.

Skatteetaten (2018b). Årsavgift. Hentet 7. desember 2018, fra: https://www.skatteetaten.no/satser/arsavgift/?year=2018#rateShowYear.

Song, Y.-Y. og Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Statistisk Sentralbyrå (2018a). Folkemengde og areal (K) 2007 - 2018. Hentet 14. oktober 2018, fra: https://www.ssb.no/statbank/table/11342/.

Statistisk Sentralbyrå (2018b). Hushald, etter samla inntekt og storleiken på gjeld (prosent) (K) 2009 - 2016. Hentet 8. oktober 2018, fra: https://www.ssb.no/statbank/table/08781/.

Statistisk Sentralbyrå (2018c). Hushald, etter storleiken på berekna nettoformue (prosent) (K) 2011 - 2016. Hentet 6. oktober 2018, fra: https://www.ssb.no/statbank/table/10320/.

Statistisk Sentralbyrå (2018d). Husholdninger, etter størrelse på samlet inntekt (K) 2006 - 2016. Hentet 5. oktober 2018, fra: https://www.ssb.no/statbank/table/07183/.

Statistisk Sentralbyrå (2018e). Utdanningsnivå, etter kommune og kjønn (K) 1970 - 2017. Hentet 18. oktober 2018, fra: https://www.ssb.no/statbank/table/09429/.

Strobl, C. (2008). *Statistical issues in machine learning: Towards reliable split selection and variable importance measures*. Cuvillier Verlag.

The International Energy Agency (2018). Global ev outlook 2018, towards cross-modal electrification.

Therneau, T. M., Atkinson, E. J., et al. (1997). An introduction to recursive partitioning using the RPART routines.

Thorsnæs, G. (3. desember 2018). Norges befolkning. Hentet 3. desember 2018, fra: https://snl.no/Norges_befolkning.

Trochim, W., Donnelly, J. P., og Arora, K. (2015). Research methods: The essential knowledge base. *Boston, MA: Cengage*.

Valgdirektoratet (2. februar 2018). Mål og samfunnsoppdrag. Hentet 12. oktober 2018, fra: https://valg.no/om-valgdirektoratet/Valgdirektoratet/.

# Appendiks

## A1 Øvrige variabler

**Tabell A1.1:** Nye øvrige variabler før variabelsammenslåing

| Kategori | Variabenavn | Variabeltype |
|---|---|---|
| Utdanning | Grunnskole | Numerisk (prosent som desimaltall) |
| Utdanning | Videregående | Numerisk (prosent som desimaltall) |
| Utdanning | UniversitetKort | Numerisk (prosent som desimaltall) |
| Utdanning | UniversitetLang | Numerisk (prosent som desimaltall) |
| Ladestasjoner | LadestasjonerPnr | Numerisk |
| Ladestasjoner | LadestasjonerKom | Numerisk |
| Inntekt | I=[0,150] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[150,250] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[250,350] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[350,450] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[450,550] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[550,750] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[750,) | Numerisk (prosent som desimaltall) |
| Formue | F=[0,250] | Numerisk (prosent som desimaltall) |
| Formue | F=[250,500] | Numerisk (prosent som desimaltall) |
| Formue | F=[500,1000] | Numerisk (prosent som desimaltall) |
| Formue | F=[1000,2000] | Numerisk (prosent som desimaltall) |
| Formue | F=[2000,3000] | Numerisk (prosent som desimaltall) |
| Formue | F=[3000,4000] | Numerisk (prosent som desimaltall) |
| Formue | F=[4000,) | Numerisk (prosent som desimaltall) |
| Gjeld | G=0 | Numerisk (prosent som desimaltall) |
| Gjeld | G=[0,1] | Numerisk (prosent som desimaltall) |
| Gjeld | G=[1,2] | Numerisk (prosent som desimaltall) |
| Gjeld | G=[2,3] | Numerisk (prosent som desimaltall) |
| Gjeld | G=[3,) | Numerisk (prosent som desimaltall) |
| | Befolkning | Numerisk |
| | Ordfører | Kategorisk (nominell) |
| Stortingsvalg | A | Numerisk (prosent som desimaltall) |
| Stortingsvalg | SV | Numerisk (prosent som desimaltall) |
| Stortingsvalg | RØDT | Numerisk (prosent som desimaltall) |
| Stortingsvalg | SP | Numerisk (prosent som desimaltall) |
| Stortingsvalg | KRF | Numerisk (prosent som desimaltall) |
| Stortingsvalg | V | Numerisk (prosent som desimaltall) |
| Stortingsvalg | H | Numerisk (prosent som desimaltall) |
| Stortingsvalg | FRP | Numerisk (prosent som desimaltall) |
| Stortingsvalg | MDG | Numerisk (prosent som desimaltall) |

**Tabell A1.2:** Nye øvrige variabler etter variabelsammenslåing

| Kategori | Variabenavn | Variabeltype |
|---|---|---|
| Utdanning | Grunnskole | Numerisk (prosent som desimaltall) |
| Utdanning | Videregående | Numerisk (prosent som desimaltall) |
| Utdanning | Universitet | Numerisk (prosent som desimaltall) |
| Ladestasjoner | LadestasjonerPnr | Numerisk |
| Ladestasjoner | LadestasjonerKom | Numerisk |
| Inntekt | I=[0,150] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[150,450] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[450,750] | Numerisk (prosent som desimaltall) |
| Inntekt | I=[750,) | Numerisk (prosent som desimaltall) |
| Formue | F=[0,250] | Numerisk (prosent som desimaltall) |
| Formue | F=[250,3000] | Numerisk (prosent som desimaltall) |
| Formue | F=[3000,4000] | Numerisk (prosent som desimaltall) |
| Formue | F=[4000,) | Numerisk (prosent som desimaltall) |
| Gjeld | G=0 | Numerisk (prosent som desimaltall) |
| Gjeld | G=[0,1] | Numerisk (prosent som desimaltall) |
| Gjeld | G=[1,3] | Numerisk (prosent som desimaltall) |
| Gjeld | G=[3,) | Numerisk (prosent som desimaltall) |
| | Befolkning | Numerisk |
| | Ordfører | Kategorisk (nominell) |
| Stortingsvalg | A | Numerisk (prosent som desimaltall) |
| Stortingsvalg | SV | Numerisk (prosent som desimaltall) |
| Stortingsvalg | RØDT | Numerisk (prosent som desimaltall) |
| Stortingsvalg | SP | Numerisk (prosent som desimaltall) |
| Stortingsvalg | KRF | Numerisk (prosent som desimaltall) |
| Stortingsvalg | V | Numerisk (prosent som desimaltall) |
| Stortingsvalg | H | Numerisk (prosent som desimaltall) |
| Stortingsvalg | FRP | Numerisk (prosent som desimaltall) |
| Stortingsvalg | MDG | Numerisk (prosent som desimaltall) |

# A2 Klassifikasjonsresultater - tidsperiode 1

**Tabell A2.1:** Klassifikasjonsresultat for klassifiseringstre - tidsperiode 1

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 2621 | 10440 |
| | Nei | 1399 | 23759 |

**Tabell A2.2:** Klassifikasjonsresultat for random forests - tidsperiode 1

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 2756 | 11426 |
| | Nei | 1264 | 22773 |

**Tabell A2.3:** Klassifikasjonsresultat for extreme gradient boosting - tidsperiode 1

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 2776 | 10662 |
| | Nei | 1244 | 23537 |

# A3 Klassifikasjonsresultater - tidsperiode 2

**Tabell A3.1:** Klassifikasjonsresultat for klassifiseringstre - tidsperiode 2

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 4066 | 10149 |
| | Nei | 2583 | 15321 |

**Tabell A3.2:** Klassifikasjonsresultat for random forests - tidsperiode 2

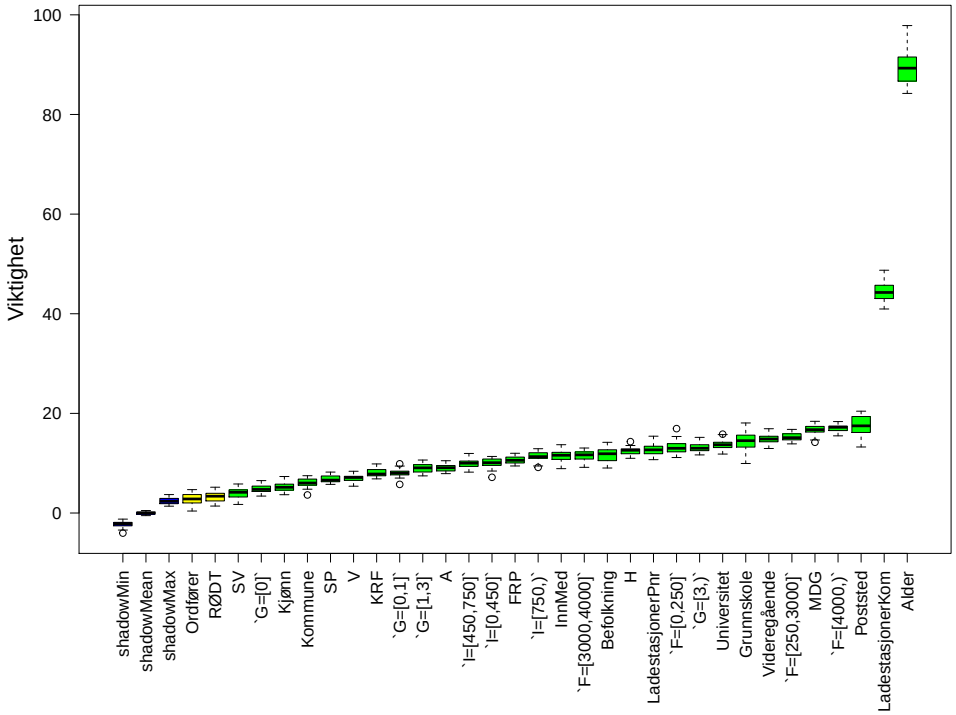
| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 3872 | 9305 |
| | Nei | 2777 | 16165 |

**Tabell A3.3:** Klassifikasjonsresultat for extreme gradient boosting - tidsperiode 2

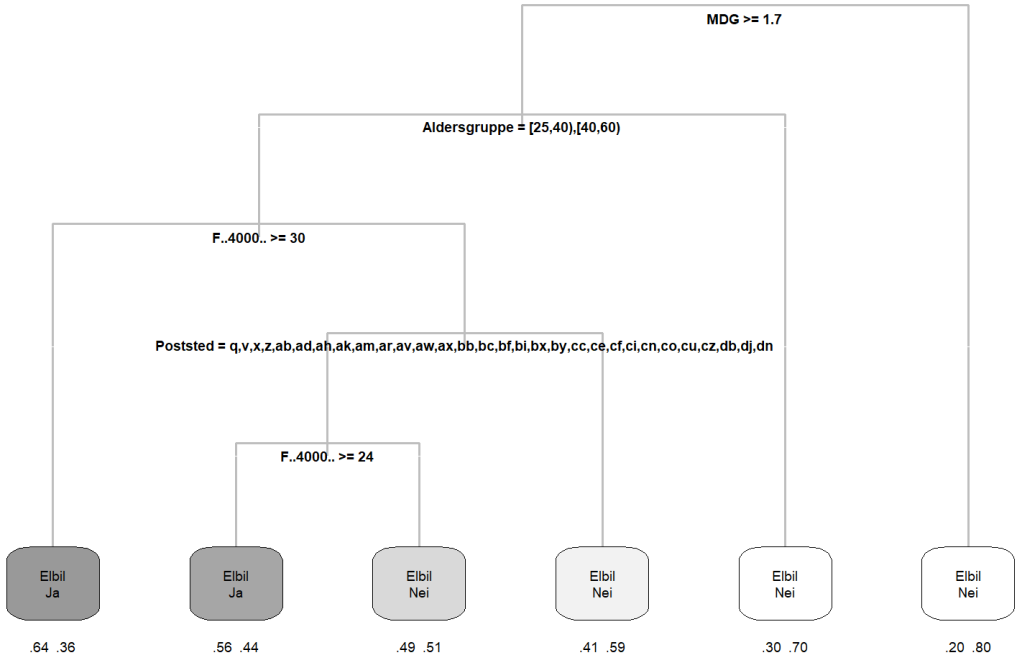
| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 4086 | 9680 |
| | Nei | 2563 | 15790 |

# A4 Tidsperiode 3: 2010-2017

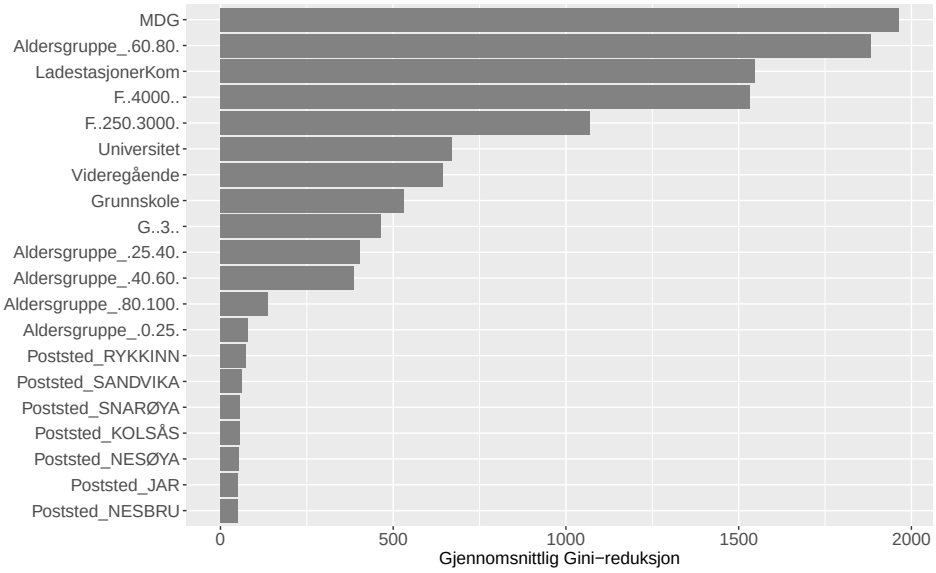
**Figur A4.1:** Boruta-modell - tidsperiode 3




**Figur A4.2:** Klassifiseringstre - tidsperiode 3



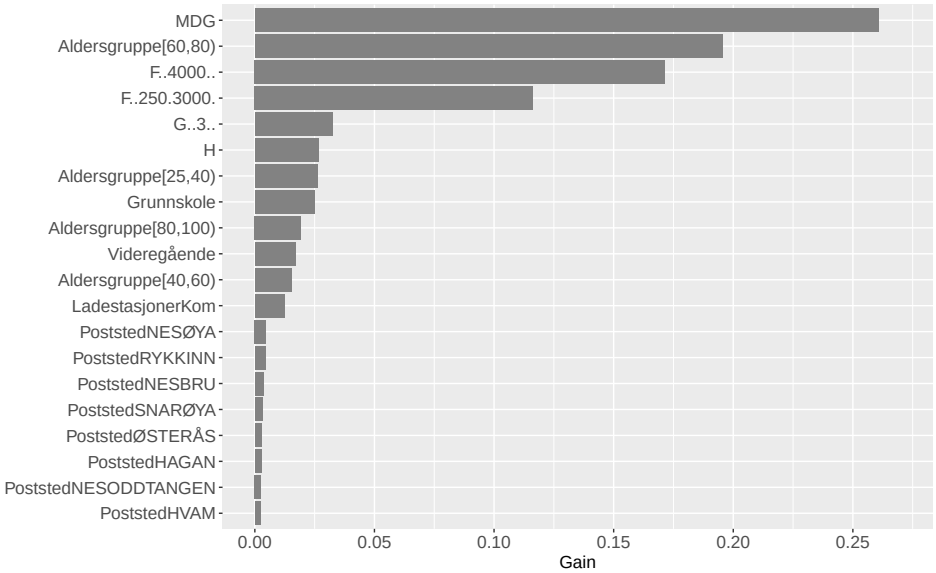
**Figur A4.3:** Random forests: Variablenes viktighet - tidsperiode 3




**Tabell A4.1:** Parameterverdier for extreme gradient boosting - tidsperiode 3

| Parameter | Verdi |
|---|---|
| B | 2000 |
| $\lambda$ | 0.1 |
| d | 2 |
| Gamma | 0.1 |
| Min_child_weight | 9 |
| Sub_sample | 0.9 |
| Colsample_bytree | 1 |

**Figur A4.4:** Extreme gradient boosting: Variablenes viktighet - tidsperiode 3



**Tabell A4.2:** Klassifikasjonsresultat for klassifiseringstre - tidsperiode 3

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 6753 | 20179 |
| | Nei | 3969 | 39436 |

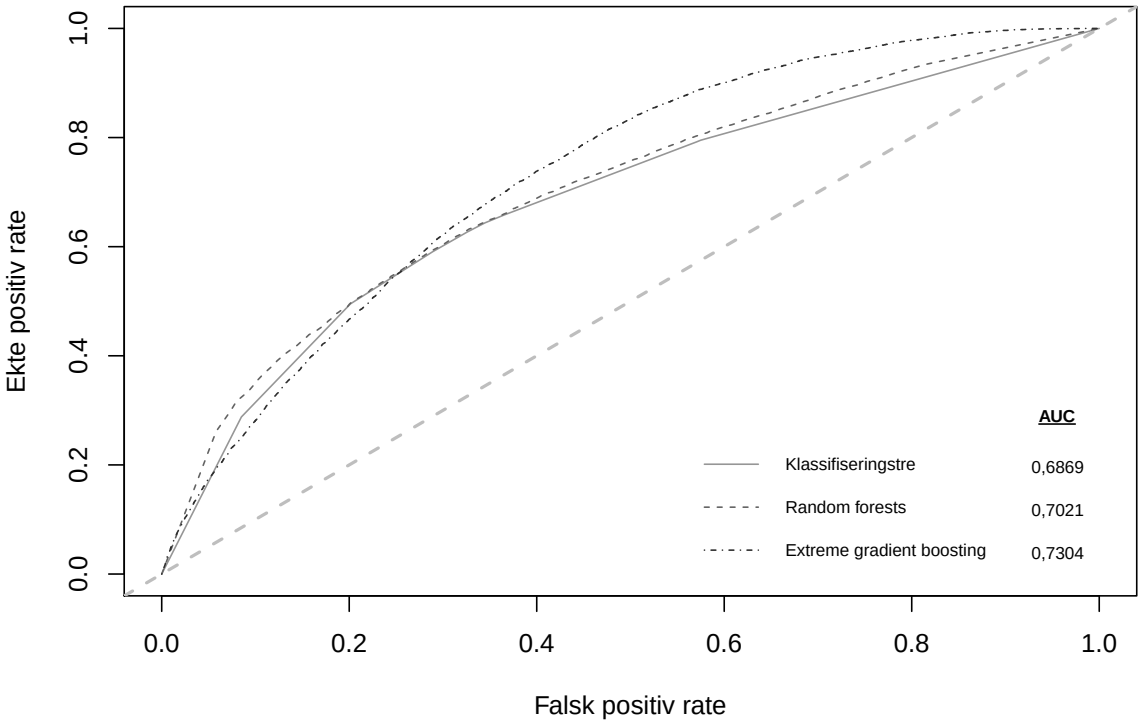
**Tabell A4.3:** Klassifikasjonsresultat for random forests - tidsperiode 3

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 6800 | 20246 |
| | Nei | 3922 | 39369 |

**Tabell A4.4:** Klassifikasjonsresultat for extreme gradient boosting - tidsperiode 3

| | | Predikert verdi | |
|---|---|---|---|
| | | Ja | Nei |
| Ekte verdi | Ja | 7164 | 19894 |
| | Nei | 3558 | 39721 |

**Figur A4.5:** ROC - tidsperiode 3



**Tabell A4.5:** Nøyaktighet og AUC - tidsperiode 3

| Metode | Balansert nøyaktighet | AUC | Konfidensintervall AUC (95%) |
|---|---|---|---|
| Klassifiseringstre | 64,57% | 0,6869 | 0.6818 - 0.6919 |
| Random forests | 64,73% | 0,7021 | 0.6970 - 0.7071 |
| Extreme gradient boosting | 66,72% | 0,7304 | 0,7256 - 0,7351 |

Merk: Konfidensintervallene er kalkulert ved hjelp av R-pakken "pROC" (Robin et al., 2011)

# A5 Poststedskoder

**Tabell A5.1:** Koder for poststeder i Oslo og Akershus

| Poststed | Kode | Poststed | Kode | Poststed | Kode |
|---|---|---|---|---|---|
| ALGARHEIM | a | HAGAN | ap | NORDRE FROGN | ce |
| ÅRNES | b | HAKADAL | aq | OPPAKER | cf |
| ÅS | c | HASLUM | ar | OPPEGÅRD | cg |
| ÅSGREINA | d | HEGGEDAL | as | OSLO | ch |
| ASKER | e | HEMNES | at | ØSTERÅS | ci |
| AULI | f | HØLEN | au | RÆLINGEN | cj |
| AURSKOG | g | HOLTER | av | RÅHOLT | ck |
| BÆRUMS VERK | h | HOSLE | aw | RÅNÅSFOSS | cl |
| BEKKESTUA | i | HØVIK | ax | RASTA | cm |
| BILLINGSTAD | j | HURDAL | ay | RUD | cn |
| BJØRKELANGEN | k | HVALSTAD | az | RYKKINN | co |
| BJØRNEMYR | l | HVAM | ba | SANDVIKA | cp |
| BLAKER | m | HVITSTEN | bb | SETSKOG | cq |
| BLOMMENHOLM | n | JAR | bc | SIGGERUD | cr |
| BLYSTADLIA | o | JESSHEIM | bd | SKEDSMOKORSET | cs |
| BØN | p | KJELLER | be | SKI | ct |
| BORGEN | q | KLØFTA | bf | SKJETTEN | cu |
| BRÅRUD | r | KOLBOTN | bg | SKOGBYGDA | cv |
| DAL | s | KOLSÅS | bh | SKOTBU | cw |
| DRØBAK | t | KRÅKSTAD | bi | SKUI | cx |
| EIDSVOLL | u | KURLAND | bj | SLATTUM | cy |
| EIDSVOLL VERK | v | LANGHUS | bk | SLEPENDEN | cz |
| EIKSMARKA | w | LEIRSUND | bl | SNARØYA | da |
| ENEBAKK | x | LILLESTRØM | bm | SOFIEMYR | db |
| ENEBAKKNESET | y | LØKEN | bn | SOLLIHØGDA | dc |
| FAGERSTRAND | z | LOMMEDALEN | bo | SON | dd |
| FEIRING | aa | LØRENSKOG | bp | SØRUM | de |
| FENSTAD | ab | LØVENSTAD | bq | SØRUMSAND | df |
| FETSUND | ac | LYSAKER | br | STABEKK | dg |
| FINSTADJORDET | ad | MAURA | bs | STRØMMEN | dh |
| FJELLHAMAR | ae | MINNESUND | bt | SVARTSKOG | di |
| FJELLSTRAND | af | MOGREINA | bu | TÅRNÅSEN | dj |
| FJERDINGBY | ag | NANNESTAD | bv | TROLLÅSEN | dk |
| FLATEBY | ah | NESBRU | bw | VESTBY | dl |
| FORNEBU | ai | NESODDEN | bx | VETTRE | dm |
| FOSSER | aj | NESODDTANGEN | by | VINTERBRO | dn |
| FROGNER | ak | NESØYA | bz | VOLLEN | do |
| GAN | al | NITTEDAL | ca | VORMSUND | dp |
| GARDERMOEN | am | NORDBY | cb | VØYENENGA | dq |
| GJERDRUM | an | NORDBYHAGEN | cc | YTRE ENEBAKK | dr |
| GJETTUM | ao | NORDKISA | cd | | |