Norges miljø- og biovitenskapelige universitet

Masteroppgave 2021 30 stp

Norges miljø- og biovitenskapelige universitet
Handelshøyskolen

En sammenligning av maskinlæringsmodeller innenfor konkursprediksjon av norske selskaper under en pandemi

Håkon Lunde Aae og Soufian Azouaghe
Økonomi og administrasjon

Forord

Denne masteroppgaven er skrevet under vårsemesteret 2021 som en avslutning på vårt masterstudium i økonomi og administrasjon ved Norges miljø- og biovitenskapelige universitet. Avhandlingen utgjør totalt 30 studiepoeng innenfor spesialiseringen Business Analytics.

Gjennom studieløpet på mastergraden har vi hatt emner innenfor grensesnittet mellom økonomi og teknologi. Her fant vi emnet innenfor maskinlæring spesielt interessant, noe som var med på å inspirere oss når det kommer til valg av oppgavens tema. Å skrive denne masteroppgaven har vært utfordrende, men samtidig veldig lærerikt. Forhåpentligvis vil oppgaven vår være et godt bidrag til konkursprediksjon som fagfelt.

Vi vil gjerne rette en stor takk til veilederen vår gjennom oppgaven, Kyrre Rickertsen, for støtte, innspill og veiledning underveis i skriveprosessen. Hans konstruktive tilbakemeldinger har bidratt til å øke kvaliteten i studien vår. Avslutningsvis ønsker vi å takke venner og familie, for deres helhjertede støtte og motiverende ord i løpet av våren.

Norges miljø- og biovitenskapelige universitet

Oslo/Ås, juni 2021

Sammendrag

Formålet med denne studien er å undersøke hvorvidt maskinlæringsmodeller er bedre enn en forenkling av Norges Banks utvidede SEBRA-modell til å predikere konkurser. Modellene er basert på en kombinasjon av regnskapstall, foretaksøkonomiske nøkkeltall og generell selskapsinformasjon. Tradisjonelt har nøkkeltall blitt brukt for å predikere konkurser, og vi vil undersøke om slike nøkkeltall er like sentrale i maskinlæringsmodeller.

Vi bruker data fra Proff Forvalt sine databaser, hvor vi har hentet ut regnskapsdata til selskaper som har levert minst ett årsregnskap i perioden 2017-2019. Det endelige datagrunnlaget etter rensing av data, består av 27.314 selskaper, hvorav 26.601 ikke har gått konkurs og resterende 713 har gått konkurs. Samtlige selskaper er hentet ut fra de ti bransjene med flest konkurser i nevnte tidsperiode.

Totalt fem maskinlæringsmodeller sammenlignes med en forenkling av den utvidede SEBRA-modellen. Disse fem modellene er beslutningstre, random forest, k-nærmeste nabo, støttevektor og nevralt nettverk. Modellene ble trent og testet gjennom bruk av femdelt kryssvalidering med jevnfordelte utvalg konkurs- og ikke-konkursselskaper i hver gruppe. Modellene ble evaluert med hensyn til AUC-verdi (area under the curve) og treffprosent. Variabelseleksjon ble utført gjennom bruk av RFE (recursive feature elimination), og justering av hyperparametere ble utført gjennom bruk av "grid search".

Random forest presterer best både med tanke på treffprosent og AUC-verdi. Modellen oppnådde en gjennomsnittlig andel riktige klassifikasjoner på 0,92 med en tilhørende AUC-verdi på 0,97. Til sammenligning oppnår en forenkling av Norges Banks utvidede SEBRA-modell en andel riktige klassifikasjoner på 0,81 med en tilhørende AUC-verdi på 0,87. Nevrale nettverk, støttevektormaskin og k-nærmeste nabo hadde også bedre treffprosent og AUC-verdi enn den forenklede SEBRA-modellen.

Maskinlæringsmodellene selekterer en kombinasjon av regnskapstall, foretaksøkonomiske nøkkeltall og generell selskapsinformasjon ved optimalisering. Funnene tyder på at konkursprediksjon er tjent med å bruke selskapsinformasjon utover tradisjonelle foretaksøkonomiske nøkkeltall. Studiens resultater indikerer også en endring i variablenes forklaringskraft, om man sammenligner et normalår med et pandemiår.

Nøkkelord - Konkursprediksjon, maskinlæring, SEBRA-modell, random forest, nevrale nettverk, beslutningstre, støttevektormaskin, k-nærmeste nabo.

Abstract

The purpose of this study is to examine whether machine learning models are better than a simplification of Norges Bank expanded SEBRA model to predict bankruptcies. The models are based on a combination of accounting figures, financial ratios and general company information. Traditionally, financial ratios have been used to predict bankruptcies, and we will examine whether such ratios are as equally important in machine learning models.

We utilize data from Proff Forvalt's databases, where we have extracted accounting data for companies that have submitted at least one annual accounting in the period 2017-2019. The final dataset after data cleaning consists of 27,314 companies, which 26,601 have not gone bankrupt and the remaining 713 have gone bankrupt. All companies are taken from the industries with the most bankruptcies in the mentioned time period.

A total of five machine learning models are compared with a simplification of the extended SEBRA model. These five models are decision trees, random forest, k-nearest neighbor, support vector machine and neutral network. The models were trained and tested using five-fold cross-validation with an evenly distributed selection of bankrupt companies and non-bankrupt companies. Furthermore, the models were evaluated with consideration to AUC value (area under the curve) and accuracy. Variable selection was performed using RFE (recursive feature elimination) and adjustment of hyperparameters was performed using grid search.

Random forest performs best both in terms of accuracy and AUC value. The model achieved a correct classification rate of 0.92 with an AUC value of 0.97. In comparison, a simplification of Norges Bank's expanded SEBRA model achieves a correct classification rate of 0.81 with an AUC value of 0.87. Neural networks, support vector machine and k-nearest neighbor also had better accuracy and AUC value than the simplified SEBRA model.

The machine learning models select a combination of accounting figures, financial ratios and general company information for optimization. The findings indicate that bankruptcy prediction benefits from using company information beyond traditional financial ratios. The results of the study also indicate a change in the models feature importance, if one compares a normal year with a pandemic year.

Keywords - Bankruptcy prediction, machine learning, SEBRA model, random forest, neural networks, decision trees, support vector machine, k-nearest neighbor.

Innhold

1.0 Innledning 1
2.0 Litteraturgjennomgang 4
2.1 William H. Beavers univariate modell (1966) 6
2.2 Edward I. Altmans Z-score (1968) 8
2.3 Edward I. Altmans ZETA modell (1977) 10
2.4 James A. Ohlsons O-score (1980) 11
2.5 SEBRA-modellen 12
2.6 Tidligere forskning om maskinlæring og konkursprediksjon 15
3.0 Metode 17
3.1 Modeller 18
3.1.1 Basis SEBRA-modell og utvidet SEBRA-modell 18
3.1.2 Beslutningstre 20
3.1.3 Random forest 21
3.1.4 K-nærmeste naboer 23
3.1.5 Støttevektormaskin 24
3.1.6 Nevrale nettverk 26
3.1.7 Styrker og svakheter ved modellene 28
3.2 Evaluering og validering 29
3.2.1 Treffprosent 30
3.2.2 Reciever operating characteristics 31
3.2.3 Kryssvalidering 33
3.2.4 Bias-varians 35
3.2.5 Hyperparameter tuning 36
3.2.6 Repliserbarhet 37
4.0 Data 37
4.1 Avhengig variabel 38
4.2 Uavhengig variabel 38
4.3 Variabelseleksjon og skalering 40
4.4 Ekskluderte observasjoner 42
5.0 Resultater 43
5.1 Modellprestasjon 43
5.1.1 Utvidet SEBRA-modell* 43
5.1.2 Beslutningstre 45
5.1.3 Random forest 47
5.1.4 K-nærmeste nabo 48
5.1.5 Støttevektormaskin 50

5.1.6 Nevrale nettverk 51
5.1.7 Sammenligning av modellene 53
5.2 Variablenes forklaringskraft 56
5.2.1 Normalår 56
5.2.2 Pandemiår 57
6.0 Diskusjon 59
6.1 Sammenligning med tidligere litteratur 59
6.2 Variablenes forklaringskraft 61
6.3 Praktisk betydning 62
6.4 Svakheter ved oppgaven 63
7.0 Konklusjon 64
Referanser 66
Appendiks 71
Appendiks 1- Fordeling av observasjoner etter alder 71
Appendiks 2 – Computer spesifikasjoner og Python-bibliotek 72
Appendiks 3- Inkluderte bransjer 72
Appendiks 4- Korrelasjon mellom uavhengige variabler 73
Appendiks 5- Valgte verdier for hyperparametere 73
Appendiks 6- Random forest utviklet med regnskapsdata fra 2019. 76

FIGURLISTE

FIGUR 1.1: ÅPNEDE KONKURSER I NORSKE AKSJESELSKAPER, KVARTAL 1 2006 – KVARTAL 3 2020 (SSB, 2021). 2
FIGUR 2.1. GJENNOMSNITTLIG TREFFPROSENT STUDIER MED PARVIS SAMMENLIGNING AV METODER (ALAKA ET AL., (2017). 17
FIGUR 3.1. TREFFPROSENT OG AVSKJÆRINGSNIVÅ FOR BASIS SEBRA OG UTVIDET SEBRA (BERNHARDSEN OG LARSEN, 2007). 20
FIGUR 3.2. ENKEL ILLUSTRASJON AV ET BESLUTNINGSTRE. 21
FIGUR 3.3. RANDOM FOREST. 22
FIGUR 3.4. ILLUSTRASJON AV HVORDAN BESLUTNINGER FORETAS BASERT PÅ ANTALL K I K-NÆRMESTE NABO. 24
FIGUR 3.5. (VENSTRE). ETABLERING AV HYPERPLAN I TODIMENSJONALT ROM. 26
FIGUR 3.6. (HØYRE). DATAPUNKTER TILLEGGES ÉN DIMENSJON SOM MEDFØRER MULIGHET FOR SEPARASJON FRA HYPERPLAN. 26
FIGUR 3.7. (VUCKOVIC ET AL., 2002). NEVRALT NETTVERK MED ETT SKJULT LAG OG P INPUT-VARIABLER. 27
FIGUR 3.8. FORVIRRINGSMATRISE OG BEREGNINGER. 30
FIGUR 3.9. GRUNNLINJE I ET ROC-DIAGRAM. AREALET UNDER KURVEN ER LIK 0,5. 32
FIGUR 3.10. ROC-KURVE. 33
FIGUR 5.1. ROC-KURVER FRA UTVIDET SEBRA*. 44
FIGUR 5.2. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR DEN UTVIDEDE SEBRA-MODELLEN*. 44
FIGUR 5.3. ROC-KURVER FRA BESLUTNINGSTRE. 46
FIGUR 5.4. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR BESLUTNINGSTRE. 46
FIGUR 5.5. ROC-KURVER FRA RANDOM FOREST. 47
FIGUR 5.6. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR RANDOM FOREST. 48
FIGUR 5.7. ROC-KURVER FRA K-NÆRMESTE NABO. 49
FIGUR 5.8. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR K-NÆRMESTE NABO. 49
FIGUR 5.9. ROC-KURVER FRA STØTTEVEKTORMASKIN. 50
FIGUR 5.10. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR STØTTEVEKTORMASKIN. 51
FIGUR 5.11. ROC-KURVER FRA NEVRALT NETTVERK. 52
FIGUR 5.12. TREFFANDEL INNAD I GRUPPER FRA KRYSSVALIDERING FOR NEVRALT NETTVERK. 52
FIGUR 5.13. ROC-KURVER FRA ALLE MODELLER. 53
FIGUR 5.14. GJENNOMSNITTLIG TREFFANDEL FRA SAMTLIGE MODELLER. 54
FIGUR 5.15. GJENNOMSNITTLIG AUC-VERDI UTEN OG MED GJENNOMFØRING AV RFE. 56
FIGUR 5.16. VIKTIGHET AV VARIABLER NORMALÅR. 58
FIGUR 5.17. VIKTIGHET AV VARIABLER PANDEMIÅR. 59
FIGUR A.1. DISTRIBUSJON AV KONKURSSELSKAPER ETTER SELSKAPETS ALDER. 71
FIGUR A.2. DISTRIBUSJON AV IKKE-KONKURSSELSKAPER ETTER SELSKAPETS ALDER. 71
FIGUR A.3. KORRELASJONSMATRISE FOR INKLUDERTE UAVHENGIGE VARIABLER. 73
FIGUR A.4. VIKTIGHETEN AV VARIABLER PANDEMIÅR MED TILDELT KONTANTSTØTTE. 76
FIGUR A.5. TREFFANDEL FRA RANDOM FOREST MED REGNSKAPSDATA 2019. 76

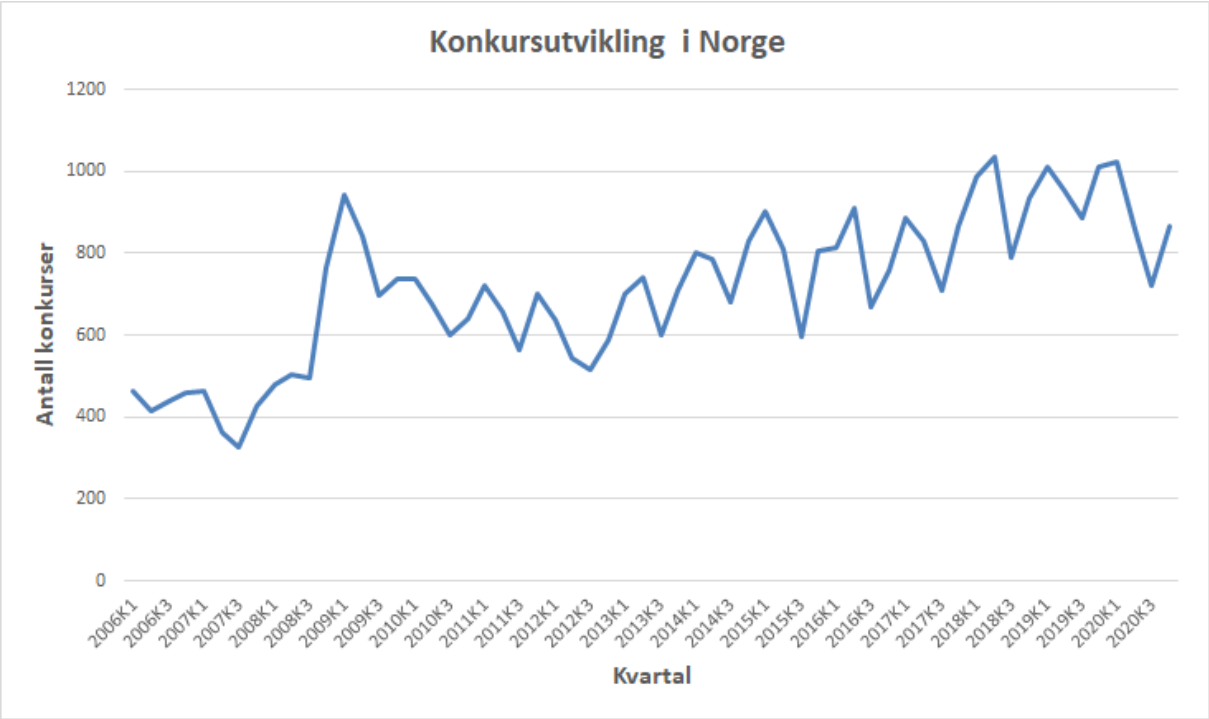
Tabelliste

TABELL 2.1. MODELLTYPER KONKURSPREDIKSJON 1966-2004 (BELLOVARY, 2007). 5
TABELL 2.2. ANDEL BEDRIFTER FEILKLASSIFISERT: DIKTOM KLASSIFISERINGSTEST (BEAVER, 1966). 8
TABELL 2.3. OPPSUMMERING AV ALTMAN SINE RESULTATER..... 10
TABELL 2.4. (WAHLSTRØM OG HELLAND, 2016). SAMMENLIGNING TREFFPROSENT ZETA- OG Z-SCORE MODELLEN..... 11
TABELL 2.5. RISIKOGRUPPER SEBRA-MODELLEN (LARSEN OG SÆTHER, 1999). 15
TABELL 2.6. GJENNOMSNITTLIG TREFFPROSENT INNENFOR KONKURSPREDIKSJON I PERIODEN 2010-2015. DE STATISTISKE METODENE ER SKREVET I KURSIV, OG RESTERENDE METODER ER Å ANSE SOM MASKINLÆRINGSMETODER..... 16
TABELL 3.1. VARIABLER SOM INNGÅR I BASIS SEBRA (MØRKERE TONET FELT) OG UTVIDET SEBRA (HELE TABELLEN), (BERNHARDSEN OG LARSEN, 2007)..... 18
TABELL 3.2. STYRKER OG SVAKHETER VED ULIKE MASKINLÆRINGSMODELLER (RAVI KUMAR & RAVI, 2007), (ALAKA ET AL., 2017). 28
TABELL 4.1 UAVHENGIGE VARIABLER..... 39
TABELL 5.1. MODELLPRESTASJONER FØR (I PARENTES) OG ETTER BRUK AV RFE. 55
TABELL 5.2. GJENVÆRENDE VARIABLER FOR HVER AV MODELLENE ETTER GJENNOMFØRT RFE. GENERELL SELSKAPSINFORMASJON I KURSIV, FORETAKSØKONOMISKE NØKKELTALL MARKERT MED STJERNE* OG ØVRIGE VARIABLER ER REGNSKAPSTALL..... 56
TABELL A.1. PROGRAMVERSJONER..... 72
TABELL A.2. BRANSJEBESKRIVELSE MED TILHØRENDE NACE-BRANSJEKODE..... 72
TABELL A.3. HYPERPARAMETERE SOM ER ENDRET FRA SCIKIT-LEARN SINE STANDARDVERDIER FOR ALLE MODELLENE..... 75

1.0 Innledning

Å overleve som selskap i konkurranseutsatte bransjer er en stor utfordring. Studien til Lennox (1999) viser at perioder med økonomisk resesjon har en tendens til å øke antallet selskaper som ikke klarer å imøtekomme sine økonomiske forpliktelser. Det samme har finanskrisen i 2008, og nyere kriser som koronakrisen vist. Konsekvensene av økonomiske vanskeligheter er mange, og det påvirker både interne og eksterne interessenter av selskapet. I ytterste konsekvens, vil selskapet ved økonomiske vanskeligheter over tid gå konkurs. Dette vil medføre at eiere mister sine investeringer, ansatte mister jobben sin og kunder og leverandører mister en viktig del av sitt nettverk. Konkursprediksjon har lenge vært et viktig og mye studert tema innenfor økonomi, som følge av de mange partene som har interesser i å evaluere og potensielt unngå kostnadene forbundet med konkurs og mislighold. Grybinenko (2017) kan tjene som et eksempel på en nyere studie som ser på konsekvenser av konkurser i et land. Hun forsket i tidsperioden 2010-2016 på konsekvensene konkurser ville ha for det ukrainske samfunnet. Hun kom fram til flere konsekvenser som førte samfunnet i negativ retning, eksempelvis ved at BNP både for landet og pr innbygger ble redusert, noe som også gjorde seg gjeldende for statens skatteinntekter.

Pr 1. januar 2021 var det ifølge SSB 606.642 virksomheter registrert i Norge. Hver dag går flere av disse konkurs, noe som gjenspeiles gjennom at det i ble registrert om lag 4.000 konkurser i løpet av 2020 (Fredriksen, 2021). Dette tilsvarer i gjennomsnitt om lag elleve konkurser hver eneste dag, noe som er en nedgang på 20 prosent sammenlignet med året før. Dette til tross for de økonomiske konsekvensene pandemien har ført med seg. Det er flere mulige årsaker til denne nedgangen. Eksempelvis har over 10.000 selskaper fått innvilget utsatt betaling av skatt og moms. En annen mulig årsak kan være kontantstøtten norske selskaper har mottatt fra staten som kompensasjon for bortfall av inntekt i forbindelse med koronarestriksjonene. Skatteetaten (2020) beregner at norske selskaper mottok i overkant av 6,8 milliarder NOK i tidsperioden mars-august 2020. Da finanskrisen startet i 2007 var antall åpnede konkurser på et bunnivå med 2.845 tilfeller i løpet av året (se figur 1.1). To år etter var antall åpnede konkurser i 2009 på hele 5.013 tilfeller. Disse tallene bør være en tankevekker for den norske økonomien i kjølvannet av koronakrisen.



Figur 1.1: Åpnede konkurser i norske aksjeselskaper, kvartal 1 2006 – kvartal 3 2020 (SSB, 2021).

Til tross for de negative konsekvensene av konkurser så har de også gunstige effekter for kapitalismen på lengre sikt. Eksempelvis har det blitt forsket på hvorvidt det kan gi en bedre sammensetning av næringsstruktur. Caballero og Hammour (1994) undersøkte responsen fra næringer til sykliske variasjoner i etterspørselen. I studien deres hevdes det at nedgangskonjunkturer kan fungere som en "renselsesmekanisme"[1] hvor effektive og nye produksjonsenheter erstatter eldre teknologi. Ideen om renselseseffekten av lavkonjunkturer dateres tilbake til Joseph Schumpeters teori (1934) om at lavkonjunkturer ikke nødvendigvis bare var et onde, men noe som var nødvendig for å tilpasse seg tidligere økonomisk endring. Det kan også tenkes at konkurser og avvikling kan føre til at utdaterte og ineffektive selskaper blir erstattet med nyere og mer effektive selskaper.

Konkurs er en juridisk prosedyre som involverer en person eller en virksomhet som ikke er i stand til å betale utestående gjeld. I henhold til Lov om gjeldsforhandling og konkurs av 8. juni 1984 nr. 58 (Konkursloven) § 60 så vil konkursbehandlingen starte dersom skyldneren defineres som insolvent og dette begjæres av skyldneren selv eller en fordringshaver. Under konkursprosessen selges skyldnerens eventuelle eiendeler for å få dekning til den utestående gjelden til kreditorene. Vanligvis er det ikke nok dekning i boet for å innfri den utestående

[1] På engelsk: The cleansing effect

gjelden, og en konkurs vil med det ofte innebære et tap for kreditorer. Man kan bli personlig ansvarlig for å betjene den utestående gjelden med fremtidige inntekter. Dette avhenger av selskapsformen til konkursselskapet da dette medfører ulik grad av personlig ansvar. Eksempelvis medfører ikke stiftelse av et aksjeselskap noen form for personlig ansvar utover aksjekapitalen som legges inn. Ved åpning av konkurs i et enkeltpersonforetak vil man selv bli personlig ansvarlig for å betjene utestående gjeld (Konkursrådet, 2012).

For at en skyldner skal være å anse som insolvent, må skyldneren videre både være illikvid og insuffisient. Illikviditet oppstår når en skyldner ikke evner å betale sine løpende forpliktelser i tide. Videre innebærer insuffisiens at verdien til skyldnerens eiendeler er mindre enn den totale gjelden. Med andre ord er en skyldner å anse som insolvent når vedkommende ikke klarer å dekke løpende forpliktelser etter hvert som de forfaller samtidig som salg av skyldnerens aktiva ikke er tilstrekkelig for å oppnå dekning til kreditorene, jf. konkursloven § 61.

Med tanke på at store selskaper vil ha innvirkning på økonomien både nasjonalt og globalt, vil også interessen for konkursprediksjon stige i forbindelse med økonomiske kriser. Ulike institusjoner besitter egenutviklede prediksjonsverktøy, som stadig gjennomgår oppdateringer og forbedringer for å oppnå en best mulig modell. Dette i et forsøk på å opparbeide seg et konkurransefortrinn. Kreditorer vil maksimere sin profitt gjennom å gi utlån til selskaper som er kapable til å betjene sine renter og avdrag, samtidig som de vil minimere utlån til de som ikke klarer dette. Det er også høyst relevant for investorer å analysere kredittverdigheten til et selskap for å unngå å pådra seg store tap.

For å eksemplifisere kostnadene for en kreditor forbundet med mislighold av utlån, bemerkes det at total utestående gjeld for norske ikke-finansielle aksjeselskaper var på 7.248 milliarder NOK i 2019 (SSB, 2021). Hypotetisk vil da mangel på tilbakebetaling av ett prosent av denne utestående gjelden påføre norske kreditorer et tap på over 72 milliarder NOK. Dette viser at en forbedring av prediksjonsnøyaktigheten for banker på bare noen få prosentpoeng kan føre til store besparelser. Brynestad (2021) beregnet at banker i Norge hadde et samlet kredittap på 26,3 milliarder NOK i 2020, noe som er 16,6 milliarder NOK mer enn i 2019.

En annen aktør som vil være interessert i anvendelse av verktøy for konkursprediksjon vil være sentralbanker, som gjennom dette vil kunne observere den finansielle stabiliteten i næringslivet. Dette kan eksemplifiseres gjennom Norges Banks SEBRA-modell[2] (Eklund et

[2] SEBRA står for "system for edb-basert regnskapsanalyse"

al., 2001), som ble utviklet med nettopp dette som formål. Modellen predikerer sannsynligheten for mislighold fra et selskap og estimerer bankers forventede utlånstap innenfor ulike bransjer.

De siste tiårene har ført med seg en enorm utvikling innenfor teknologi og hvordan selskaper drives. Stadig flere selskaper tar i bruk maskinlæring og digitalisering, og går med det bort fra en mer tradisjonell drift av selskapet. Med denne stadige utviklingen som bakgrunn, stiller Rhie et al. (2005) spørsmål hvorvidt foretaksøkonomiske nøkkeltall er like relevante som de pleide å være, når det kommer til konkursprediksjon. Studien finner indikasjoner på at de prediktive evnene til de tradisjonelle nøkkeltallene har gjennomgått en gradvis reduksjon over tid, og med dette ikke er å anse som like viktige som tidligere.

På bakgrunn av dette tar studien vår utgangspunkt i tre problemstillinger. Den første og viktigste problemstillingen er å utvikle maskinlæringsmodeller som er bedre til å predikere konkurser i norske selskaper enn en forenkling av Norges Banks utvidede SEBRA-modell. For det andre så har tidligere studier av konkursprediksjon i stor grad brukt statistiske modeller med foretaksøkonomiske nøkkeltall som de viktigste forklaringsvariablene. Vi vil undersøke hvor viktige slike nøkkeltall er for prediksjonene til maskinlæringsmodeller. For det tredje vil vi sammenligne variablenes forklaringskraft i en maskinlæringsmodell i et normalår opp mot pandemiåret 2020.

2.0 Litteraturgjennomgang

Konkursprediksjon som fagfelt dateres tilbake til tidlig på 1930-tallet hvor de innledende studiene analyserte forholdstall for å predikere fremtidige konkurser. Forskningen frem til midten av 1960-tallet fokuserte på univariate analyser, som kjennetegnes ved at de kun hensyntar én enkelt variabel. Studien til Beaver (1966) er et eksempel på en slik studie. Formålet med univariate analyser er å beskrive dataen og finne potensielle mønster innad i den aktuelle variabelen. Som følge av at analysen utelukkende fokuserer på én variabel, vil det ikke fanges opp mønster og sammenhenger mellom flere ulike variabler som i eksempelvis en multivariat analyse. Altman (1968) publiserte den første studien basert på multivariate analyser innen konkursprediksjon, som i likhet med Beaver (1966) sin studie fremdeles er viktige bidrag til litteraturen pr dags dato. Altmans (1968) modell er en fem-faktor multivariat diskriminant modell, og vil med dette hensynta flere variabler simultant i motsetning til univariate analyser. Teknologiske fremskritt har imidlertid gjort andre metoder

(logit analyse, probitanalyse og nevrale nettverk) mer fremtredende (se tabell 2.1) (Bellovary, 2007).

Tabell 2.1. Modelltyper konkursprediksjon 1966-2004 (Bellovary, 2007).

År	Diskriminant Analyse	Logit Analyse	Probit Analyse	Nevrale Nettverk	Andre
1960-tallet	2	0	0	0	1
1970-tallet	22	1	1	0	4
1980-tallet	28	16	3	1	7
1990-tallet	9	16	3	35	11
2000-tallet	2	3	0	4	3
Totalt	63	36	7	40	26

Siden studiene til Beaver og Altman mot slutten av 1960-tallet har konkursprediksjon vært et tema av interesse for en rekke forskere. Mange modeller har blitt lagt frem og testet, men konklusjonene har ofte vært motstridende (Berg, 2007). Det finnes to hovedtyper av modeller for konkursprediksjon som ofte blir anvendt innenfor litteraturen. Disse kan kategoriseres som regnskapsbaserte modeller og markedsbaserte modeller. Sistnevnte modeller tar utgangspunkt i markedsverdier, samtidig som det kreves at selskapet er børsnoterte. Regnskapsbaserte modeller bruker informasjon som fremlegges gjennom selskapets regnskapsinformasjon Denne studien fokuserer på regnskapsbaserte modeller ettersom datagrunnlaget utelukkende består av ikke-børsnoterte selskaper.

En svakhet ved regnskapsmodeller er at de utelukkende baseres på historisk informasjon og ikke hensyntar volatiliteten til et selskaps eiendeler under estimeringer for mislighold (Agarwal og Taffler, 2008). De er basert på historisk informasjon og er ikke nødvendigvis informative når det kommer til å predikere fremtiden. Agarwal og Taffler (2008) poengterer også at regnskapstall er basert på prinsippene om forsiktighet og historisk kost som medfører at markedsverdien ikke nødvendigvis stemmer overens med den bokførte. På den annen side er regnskapsbaserte modeller enkle å anvende i praksis og tolkningen av resultatene kan gjøres på en oversiktlig måte. Dessuten er informasjonen man trenger for å utvikle regnskapsbaserte modeller lett tilgjengelig.

Hillegeistet et al. (2004) sammenligner i sin studie Altmans (1968) og Ohlsons (1980) regnskapsbaserte modeller opp mot markedsbaserte modeller. De argumenterte for at aksjemarkedet gir markedsbaserte modeller en mulig overlegen kilde til informasjon når det kommer til konkursprediksjon, ettersom markedsbaserte modeller samler informasjon fra kilder utenom regnskapet. De kommer frem til at enkelte modeller basert på markedsbasert

informasjon kan gi mer informasjon om hvorfor et selskap går konkurs enn noen av de mest anvendte regnskapsbaserte modellene (Altmans Z-score og Ohlsons O-Score). Agarwal og Taffler (2008) på den annen side konkluderer med at det er liten forskjell mellom de prediktive evnene til regnskapsbaserte- og markedsbaserte tilnærminger.

Før man hadde tilgang til kvantitativ informasjon om selskapers prestasjon (med hensyn til finansielle nøkkeltall), var det vanlig at byråer ble etablert for å levere kvalitativ informasjon som vurderte kredittverdigheten til enkelte aktører (Altman, 1968). Noen av de mest anerkjente bidragene innenfor litteraturen internasjonalt er de tidligere nevnte bidragene til Beaver (1966), Altman (1968), Altman (1977) og Ohlson (1980). Larsen og Sæther (1999) i Norges Bank stod for utviklingen av SEBRA-modellen, som skal anslå sannsynligheten for konkurser i norske aksjeselskap. I 2007 ble det videre publisert en videreutvikling av den opprinnelige SEBRA-modellen, i form av basis-SEBRA og den utvidede SEBRA-modellen (Bernhardsen og Larsen, 2007). Disse modellene vil beskrives i mer detalj i kapittel 3.1. Grunnet det store omfanget av forskning innenfor konkursprediksjonen, gjøres det kun rede for de mest sentrale bidragene i kommende underkapitler.

2.1 William H. Beavers univariate modell (1966)

Regnskapsforskeren William Beaver (1966) står for et av de mest anerkjente internasjonale studiene innenfor konkursprediksjon som tar i bruk univariat analyse. Svakheten ved univariate modeller er som nevnt tidligere at de kun tar utgangspunkt i hver enkelt uavhengig variabel, og ikke ser på sammenhengen mellom disse. Resultatene fra modellen kan i så måte virke motstridende, hvor ett forholdstall kan predikere i retning av konkurs, samtidig som et annet forholdstall predikerer i retning av ikke-konkurs. Styrken til modellen kommer frem i dens enkle anvendelse samtidig som det er enkelt å tolke variablenes viktighet.

Datagrunnlaget bestod av nøkkeltall fra 79 konkursselskaper og tilsvarende antall ikke-konkursselskaper fra tidsperioden 1954-1964, hvor alle selskapene var offentlig eide industrielle selskaper. Konkursselskapene var hentet fra 38 ulike bransjer. Videre sammenlignet Beaver de ulike kategoriserte selskapene gjennom tilnærmet lik størrelse og bransjefordeling for å minimere størrelses- og industrieffekten.

Beaver startet i utgangspunktet med 30 finansielle nøkkeltall, hvor disse er beskrevet i detalj i studien hans (1966: s.78). Disse ble videre delt inn i seks hovedgrupper: 1) nøkkeltall relatert til kontantstrøm, 2) nøkkeltall relatert til netto resultat, 3) nøkkeltall relatert til total gjeld i

forhold til totalkapital, 4) nøkkeltall relatert til likvide midler i forhold til totalkapital, 5) nøkkeltall relatert til likvide midler i forhold til nåværende gjeld, 6) nøkkeltall relatert til omsetning. Nøkkeltallene som ble valgt ut til den endelige modellen var kontantstrømbaserte nøkkeltall som var mye anvendt og hadde levert gode resultater i tidligere forskning. For at et nøkkeltall skulle bli inkludert i analysen måtte det dessuten foreligge et avvik mellom gjennomsnittsverdien for selskaper som gikk konkurs og de som ikke gikk konkurs. Beaver endte opp med de seks nøkkeltallene som er listet opp i tabell 2.2, hvor kun ett nøkkeltall fra hver av de seks hovedgruppene ovenfor ble inkludert.

Det ble gjennomført flere steg for å analysere datagrunnlaget. Først ble det gjennomført en profilanalyse, som ikke er en prediktiv test, men en praktisk test for å skissere sammenhengen mellom konkurs- og ikke-konkursselskaper. Deretter ble gjennomsnittsverdien for hver gruppe med forholdstall kalkulert. I studien (1966) ble det argumentert for at dersom gruppene hadde liten eller ingen overlapp med hverandre, indikerte dette at forholdstallene ville være utmerket for å predikere konkurs og vice versa. Deretter ble det gjennomført en klassifiseringstest. Klassifiseringstesten gir en dikotom prediksjon, som omhandler klassifikasjonen av et selskap som konkurs eller ikke-konkurs. Det ble rangert 30 nøkkeltall i stigende rekkefølge og det ble satt et knekkepunkt for disse. De selskapene som var under det definerte knekkepunktet ble ansett som konkursselskaper, mens de øvrige ble ansett som ikke-konkursselskaper.

Avslutningsvis ble testutvalget delt inn i to deler og etterfølgende analyser ble gjennomført på begge disse. Tallene i parentes henviser til resultatet fra den første testen, mens tallene uten parentes henviser til den andre testen. Resultatene viser til andelen feilklassifiseringer, noe som innebærer at jo lavere tall, desto bedre prediktive evner for det aktuelle forholdstallet. Nærmere forklart innebærer 0,13 i første rad og kolonne at forholdstallet mellom kontantstrøm og total gjeld, har en andel feilklassifikasjoner på 0,13 ett år før konkurs. Resultatene i tabellen viser at forholdet mellom kontantstrøm og total gjeld har lavest andel feilklassifiseringer, etterfulgt av forholdet mellom netto resultat og totalkapitalen.

Tabell 2.2. Andel bedrifter feilklassifisert: Diktom klassifiseringstest (Beaver, 1966).

Nøkkeltall	År før konkurs				
	1	2	3	4	5
Kontantstrøm / Total gjeld	.13 (.10)	.21 (.18)	.23 (.21)	.24 (.24)	.22 (.22)
Netto resultat / Totalkapital	.13 (.12)	.20 (.15)	.23 (.22)	.29 (.28)	.28 (.27)
Total gjeld / Totalkapital	.19 (.19)	.25 (.24)	.34 (.28)	.27 (.24)	.28 (.27)
Arbeidskapital / Totalkapital	.24 (.20)	.34 (.30)	.33 (.33)	.45 (.35)	.41 (.35)
Likviditetsgrad 1	.20 (.20)	.32 (.27)	.36 (.31)	.38 (.32)	.45 (.31)
Likvide eiendeler – kortsiktig gjeld	.23 (.23)	.38 (.31)	.43 (.30)	.38 (.35)	.37 (.30)

2.2 Edward I. Altmans Z-score (1968)

I 1968 publiserte finansprofessoren Edward Altman den egenutviklede multivariate Z-score modellen (Altman, 1968), som av mange er regnet som et av de mest sentrale bidragene til konkursprediksjon som fagfelt (Bellovary et al., 2007). Modellen ble utviklet som følge av de mange svakhetene som foreligger ved univariate modeller. Studien påpeker at univariate modeller plasserer for mye vekt på enkeltstående variabler, og at de som følge av dette er utsatt for å ha motstridende konklusjoner fra de ulike variablene. Univariate modeller vil som poengtert tidligere heller ikke fange opp mulige sammenhenger mellom de ulike variablene, hvor multivariate modeller til sammenligning vil ha en ekstra kilde til informasjon.

Datagrunnlaget var bestående av 66 selskaper, hvorav den ene halvparten var konkursselskaper og den andre halvparten ikke-konkursselskaper. Konkursselskapene var hentet ut fra tidsperioden 1946-1965, mens ikke-konkursselskapene var basert på aktive selskaper i 1966. Selskapene hadde en totalkapital i området $1 million - $25 millioner. Dette intervallet ble etablert for å minimere størrelse- og industrieffekten, gjennom å fjerne både de små og de veldig store selskapene (Altman, 1968).

Etter vurderinger av problemstillingen og studiens formål, falt det endelige valget vedrørende statistisk teknikk på multippel diskriminantanalyse (MDA). Dette er en statistisk metode som

brukes til å klassifisere en observasjon til en av flere grupperinger, avhengig av observasjonens karakteristikk (Altman, 1968). Metoden brukes primært til å klassifisere og/eller forutsi problemer, der den avhengige variabelen betegnes i kvalitativ form, som i dette tilfellet hvor hver observasjon står overfor klassifikasjonen konkurs eller ikke-konkurs. Tidligere har MDA i hovedsak blitt brukt når det kommer til biologisk forskning, mens teknikken i senere tid også har blitt anvendt overfor økonomiske problemstillinger (Altman, 2000).

Datagrunnlaget i studien bestod av 66 selskaper, hvor dette på samme måte som i Beaver sin studie (1966), ble delt inn i to like store grupperinger bestående av 33 konkursselskaper og like mange ikke-konkursselskaper. Datagrunnlaget var basert på konkursselskaper fra tidsperioden 1946-1965, mens ikke-konkursselskapene var basert på aktive selskaper i 1966. Selskapene hadde en totalkapital i området $1 million - $25 millioner. Dette intervallet ble etablert for å minimere størrelse- og industrieffekten, gjennom å fjerne både de små og de veldig store selskapene (Altman, 1968).

Med de 22 mest brukte variablene innenfor tidligere forskning som utgangspunkt, ble de endelige variablene valgt ut på bakgrunn av variabelens signifikansnivå, korrelasjonsnivå og evne til å predikere korrekt. De fem variablene som ble valgt ut i Altmans endelige modell var basert på følgende forholdstall og vekting av forholdstallene:

$$Z = 1.2\frac{\text{Arbeidskapital}}{\text{Totalkapital}} + 1.4\frac{\text{Tilbakeholdt overskudd}}{\text{Totalkapital}} + 3.3\frac{\text{Resultat før renter og skatt}}{\text{Totalkapital}} + 0.6\frac{\text{Markedsverdi egenkapital}}{\text{Total gjeld}} + 1.0\frac{\text{Salgsinntekter}}{\text{Totalkapital}}$$

Hvor Z = Samlet indeks (Z-score).

Altman (1968) gjennomførte videre en F-test for å teste hvorvidt hver variabel var signifikant og avdekke dens evne til å korrekt klassifisere et konkursselskap ett år i forveien. Alle forklaringsvariablene, utenom variabelen salgsinntekter/totalkapital, var signifikant på 0,01 nivå. Denne ble allikevel inkludert i modellen ettersom den i stor grad bidro til modellens diskrimineringsevne. Videre beskriver Altman (1968) at dersom Z-scoren var over 2,99 ville selskapet mest sannsynlig overleve. På den annen side var verdier under 1,81 en indikasjon på at et selskap ville gå konkurs. Området mellom disse klassifiseringsterskelene (1,81 og 2,99)

ble videre definert som en gråsone. Dette innebar at verdier innenfor dette intervallet ikke ga tydelig nok informasjon rundt hvorvidt et selskap ville gå konkurs eller ikke.

Tabell 2.3 oppsummerer treffprosenten til Z-score modellen over en femårs periode. Det fremkommer av tabellen at modellen ga mest treffsikre prediksjoner ett år i forveien, med gradvis fallende resultater etter hvert som modellen skulle predikere flere år i forveien. Dette er i tråd med studiens konklusjon om at modellens prediktive evner blir upålitelig etter det andre året.

Tabell 2.3. Oppsummering av Altman sine resultater.

År før konkurs	Antall treff	Antall ikke treff	Treffsikkerhet
1	31	2	94 %
2	23	9	72 %
3	14	15	48 %
4	8	20	29 %
5	9	16	36 %

2.3 Edward I. Altmans ZETA modell (1977)

I 1977 publiserte Altman et al. (1977) en ny studie som la frem den nyutviklede ZETA-modellen, som skulle virke som en forbedring av den tidligere publiserte Z-score modellen. Studien pekte på at størrelsen til selskaper og tilhørende finansielle profil stadig var i endring, og at en konkursprediksjonsmodell må hensynta nettopp dette. Videre trekkes det frem at denne studien skiller seg positivt ut i form av at regnskapsdata med tilhørende fotnoter hos de aktuelle selskapene er nøye analysert for å fange opp endringer i regnskapsstandarder og regnskapsskikk.

Datagrunnlaget denne gangen bestod 53 konkursselskaper og 58 ikke-konkursselskaper. Datagrunnlaget i denne studien omfattet både store og små selskaper. Selskapene var hentet ut fra tidsperioden mellom 1962-1975, og drev enten med produksjon eller detaljhandel. Videre inkluderte studien syv nøkkeltall i den nye modellen, sammenlignet med fem under Z-score modellen. Ingen av de syv nøkkeltallene i ZETA-modellen var med i den opprinnelige Z-score modellen.

Studien tok utgangspunkt i 27 ulike variabler som skulle vurderes, hvor disse ble delt inn i seks hovedgrupper. Etter en gjentakende filtreringsprosess for å redusere antall variabler ble den endelige modellen basert på syv forklaringsvariabler, som var den mest pålitelige kombinasjonen i ulike valideringsprosedyrer. Resultatene viste ingen forbedring ved å legge til flere variabler, ei heller ved å redusere antallet. De syv variablene som ble inkludert i den

endelige modellen var da følgende: 1) egenkapitalprosent, 2) inntjeningsstabilitet, 3) rentedekningsgrad, 4) kumulativ lønnsomhet, 5) likviditetsgrad 1, 6) totalkapitalrentabilitet og 7) størrelse (målt etter totalkapital). Kumulativ lønnsomhet var den variabelen som forklarte modellens prediksjoner i høyest grad da den sto for 25 prosent av diskriminasjonen.

Forbedringen i prediktive evner hos ZETA-modellen kommer tydelig frem i tabell 2.4 som illustrerer en sammenligning av andelen riktige klassifikasjoner mellom ZETA-modellen og den tidligere utviklede Z-score modellen. ZETA-modellen kan vise til en liten forbedring i treffprosent ett år i forveien, samt store forbedringer 2-5 år i forveien sammenlignet med Z-score modellen overfor konkursselskaper. Med dette predikerer den bedre overfor konkursselskaper for alle år, men leverer til gjengjeld noe svakere resultater når det kommer til ikke-konkursselskaper.

Tabell 2.4. (Wahlstrøm og Helland, 2016). Sammenligning treffprosent ZETA- og Z-score modellen.

År før konkurs	**ZETA Modell (Konkurs)**	**Z-SCORE (Konkurs)**	**ZETA Modell (Ikke-konkurs)**	**Z-SCORE (Ikke-konkurs)**
1	96,2 %	93,9 %	89,7%	97,0 %
2	84,9 %	71,9 %	93,1%	93,9 %
3	74,9 %	48,3 %	91,4%	-
4	68,1 %	28,6 %	89,5%	-
5	69,8 %	36,0 %	82,1%	-

2.4 James A. Ohlsons O-score (1980)

I 1980 publiserte regnskapsprofessoren James Ohlson (1980) den første studien som benyttet logistisk regresjon for å predikere konkurser, da han la frem sin egenutviklede O-score modell. Ved å benytte seg av logistisk regresjon ønsket han å redusere svakhetene ved MDA-analyse for å bedre prediksjonsevnen. Logistisk regresjon brukes når responsvariabelen er en kategorisk variabel, noe som er tilfellet i konkursprediksjon hvor responsvariabelen står overfor to mulige klassifikasjoner: konkurs eller ikke-konkurs. Studien pekte på tre hovedproblemer ved MDA-tilnærmingen som var utslagsgivende for at den tok i bruk logistisk regresjon: (i) MDA-modellen krever restriktive statistiske antagelser. Eksempelvis skal varians-kovariansmatrisen til gruppene som skal predikeres være identiske. (ii) Resultatet fra anvendelsen av en MDA-modell er en poengsum som er lite intuitiv ved tolkning. (iii) Det parvise utvalget av konkurs- og ikke-konkursselskaper som typisk har blitt brukt i MDA er matchet nøyaktig mot kriterier som størrelse og industri, som i verstefall kunne føre til at modellen får redusert forklaringskraft.

Der hvor tidligere forskere anvendte parvise utvalg, motsatte Ohlson denne metoden og mente at man kunne gå glipp av viktig informasjon. Studiens datagrunnlag var derfor skjevfordelt og bestod av 2.085 ikke-konkursselskaper og 105 konkursselskaper. Selskapene i studien ble valgt ut med forbehold om at de enten måtte være i drift eller hadde gått konkurs i tidsperioden 1970-1976, samtidig som de måtte drive innenfor industriell sektor.

O-score modellen bestod av ni forklaringsvariabler, hvorav to av disse var dummyvariabler, og ble valgt ut på bakgrunn av anvendelse i tidligere forskning. Alle variablene var basert på selskapers regnskapsinformasjon. De ni variablene som ble inkludert i regresjonen var: 1) log(totalkapital/BNP prisindeks), 2) total gjeld/totalkapital, 3) arbeidskapital/totalkapital, 4) kortsiktig gjeld/omløpsmidler, 5) 1 hvis total gjeld overstiger totalkapital, 0 ellers, 6) netto resultat/totalkapital, 7) driftsresultat/totalkapital, 8) 1 hvis nettoresultat var negativt to siste årene, 0 ellers, 9) endring i driftsresultat to siste årene.

Tolkningen av modellens O-score (responsvariabel Y ovenfor), er at jo høyere scoren er, desto høyere er sjansen for å gå konkurs. Det ble dessuten satt en terskel på 0,5, hvor en score høyere enn dette indikerer potensiell konkurs innenfor en tidsramme på ett år, mens en score under denne terskelen tilsvarer at et selskap er å anse som utenfor faresonen. Ohlson estimerte videre tre sett med estimater som skulle predikere konkurser henholdsvis ett, to og tre år før konkurs, noe som resulterte i treffprosenter på henholdsvis 96,1, 95,6 og 92,8.

2.5 SEBRA-modellen

Norges Bank har i mange år brukt en empirisk modell ved navnet SEBRA-modellen for å predikere konkurser overfor norske aksjeselskaper. SEBRA-modellen bruker generell additiv modell (GAM) introdusert av Hastie og Tibshirani (1995) for å estimere konkurser. GAM er en utvidelse av generell lineær modell (GLM), som Berg (2007) beskriver som en generalisering av multippel regresjonsmodell.

GLM er en bred klasse av modeller som omfatter eksempelvis både regresjonsmodeller og variansanalyser (ANOVA). Begge modellene er ofte brukt for å referere til GLM (Wiley & Wiley, 2019). GAM forlenger GLM ved å erstatte den lineære formen, og vil kunne fange opp komplekse ikke-lineære sammenhenger mellom den avhengige variabelen og de uavhengige variablene. Ved valg av modell står man overfor en avveining av modellens kompleksitet og tolkningsmulighet, hvor GAM tilbyr en middelvei når det kommer til denne

problemstillingen. Dette gjennom å kunne fange opp komplekse sammenhenger, uten at dette går på bekostning av tolkningsmulighetene av modellens prediksjoner (Datacamp, 2020).

Ifølge Berg (2007) er GAM en signifikant bedre modell for konkursprediksjon sammenlignet med diskriminantanalyse, lineære regresjonsmodeller og nevrale nettverk. Modellen brukes i dette tilfellet for å anslå bankenes forventede utlånstap til foretak og ulike næringer ved eventuell konkurs.

I den opprinnelige SEBRA-modellen blir sannsynligheten for konkurs i hovedsak modellert med nøkkeltall som beskriver foretakets inntjening, soliditet og likviditet (Eklund et al., 2001). Dette i likhet med Ohlsons modell, hvor begge modeller tar for seg sentrale foretaksøkonomiske faktorer som fremkommer av selskapets regnskapsinformasjon. SEBRA-modellen skiller seg derimot fra Ohlson sin modell, ved at den inkluderer mål for foretakenes alder samt bransjevariabler basert på aggregater av de økonomiske nøkkeltallene (Bernhardsen og Larsen, 2007). Det understrekes videre at modellen ikke inkluderer tilleggsinformasjon som vanligvis ville vært relevant, eksempelvis betalingsanmerkninger eller forsinket årsregnskap.

Den opprinnelige SEBRA-modellen brukte følgende forklaringsvariabler på bakgrunn av kriteriene som lå til grunn for valg av modell:

- **Inntjening:**

Ethvert selskap må kunne generere en inntekt som i det minste dekker kostnadene. Dette er nødvendig for at selskapet skal kunne dekke løpende forpliktelser samtidig som en må være kapabel til å betjene den gjelden selskapet har. Om selskapet evner å generere større inntekter enn kostnader, vil dette medføre et overskudd eller en inntjening. Nøkkeltallet som brukes i SEBRA-modellen for inntjening er inntjening som andel av totalkapital (tkr).

- **Likviditet:**

Likviditet er et mål som beskriver selskapets kjøpekraft og betalingsevne. Det innebærer i hvilken grad et selskap evner å betjene sine løpende forpliktelser. Likviditeten til et selskap beregnes på bakgrunn av kontanter og midler som er lett omsettelige, og er i så måte et uttrykk for lett tilgjengelige penger. Graden av likviditet som er nødvendig for et selskap avhenger av bransjen det tilhører. Likviditeten er derfor ikke nødvendigvis sammenlignbart mellom bransjer. De tre variablene som tilhører likviditet i den opprinnelige SEBRA-modellen er:

1) betalingsmidler minus kortsiktig gjeld som andel av driftsinntekter (lik), 2) skyldige offentlige avgifter som andel av totalkapital (ube) og 3) leverandørgjeld som andel av totalkapital (lev).

- **Soliditet:**

Soliditet sier noe om hvor mye av selskapet som er med egenkapital. Nøkkeltallet betegner virksomhetens evne til å tåle finansielle tap. Dersom store deler av selskapet er finansiert med gjeld, så vil det være vanskelig å overleve tider med dårlig inntjening. Tre nøkkeltall beskriver soliditet innenfor SEBRA-modellen:
1) egenkapital som andel av totalkapital (eka), 2) dummyvariabel for bokført egenkapital mindre enn innskutt egenkapital (taptek) og 3) dummyvariabel for utbetalt utbytte siste regnskapsår (div).

- **Bransje:**

Risikoen for konkurs er forskjellig i ulike bransjer. I bransjer hvor det foreligger gjennomgående lav grad av egenkapitalprosent hos selskapene, samtidig som den gjennomsnittlige gjelden er å anse som høy, vil det naturligvis være større risiko for konkurs i perioder med lav inntjening. Dette kommer til uttrykk gjennom de tre inkluderte variablene i SEBRA-modellen som skal beskrive sentrale bransjekjennetegn i konkursutsatte bransjer:
1) bransjevis gjennomsnitt for 'egenkapital som andel av totalkapital' (meaneka),
2) bransjevis gjennomsnitt for 'leverandørgjeld som andel av totalkapital' (meanlev) og
3) bransjevis standardavvik for 'inntjening som andel av totalkapital' (sdtkr).

- **Alder:**

Nyetablerte selskaper står ofte overfor en formidabel oppgave når de skal trenge inn i et marked. De allerede etablerte konkurrentene innenfor markedet vil gjerne besitte kunnskap om bransjen og inneha veletablerte relasjoner med leverandører og kunder. Vanskeligheten forbundet med å overleve startfasen som selskap understrekes av SSB sine tall (2020) som rapporterer at over halvparten av nyetablerte selskap i perioden 2013-2017 har gått konkurs allerede i løpet av det første året etter stiftelse. Dette underbygges også blant våre data, hvor det er klar overvekt av konkurser blant nyetablerte selskaper (se appendiks 1). SEBRA-modellen etablerer en dummyvariabel for hvert år etter etablering, som vil slå positivt ut alt ettersom selskapet er 1-8 år gammelt. Ved alder >8 år vil selskapet tilegnes en verdi tilsvarende åtte år.

- **Størrelse:**

Størrelsen til det enkelte selskap blir i dette tilfellet målt som summen av egenkapital og gjeld i selskapet. Det forekommer flere tilfeller av konkurs blant mindre selskaper sammenlignet med større selskaper. Jo mindre et selskap er, desto mer sårbart vil det være overfor uforutsette hendelser og utviklinger i markedet. SEBRA-modellen buker den logaritmiske funksjonen til selskapets totalkapital (sum egenkapital og gjeld) som et mål for størrelse.

Alle variablene i SEBRA-modellen inngår med et signifikantnivå på minst 0,1 prosent (Eklund et al., 2001). Bernhardsen (2001) sine resultater viser at størrelse var den mest signifikante variabelen. Med nevnte nøkkeltall tatt i betraktning, ender Larsen og Sæther (1999) i utgangspunktet opp med en modell bestående av 18 ulike risikogrupper. Disse blir etter hvert begrenset ned til åtte grovere gruppeinndelinger, hvor stigende gruppenummer henviser til økende risiko for mislighold av lån (se tabell 2.5). Dette innebærer at eksempelvis gruppe 1-2 er å anse som mindre utsatt for konkurs enn gruppe 7-8 som representerer de gruppene med høyest risiko for konkurs.

Tabell 2.5. Risikogrupper SEBRA-modellen (Larsen og Sæther, 1999).

Inntjening	**Likviditetsutvikling**	**Egenkapitalandel**	
		Over 20%	**Under 20%**
Mer enn 20%	God	Gruppe 1	Gruppe 3
	Svak	Gruppe 2	Gruppe 4
0-20%	God	Gruppe 5	Gruppe 6
	Svak		
Negativ	God	Gruppe 7	Gruppe 8
	Svak		

2.6 Tidligere forskning om maskinlæring og konkursprediksjon

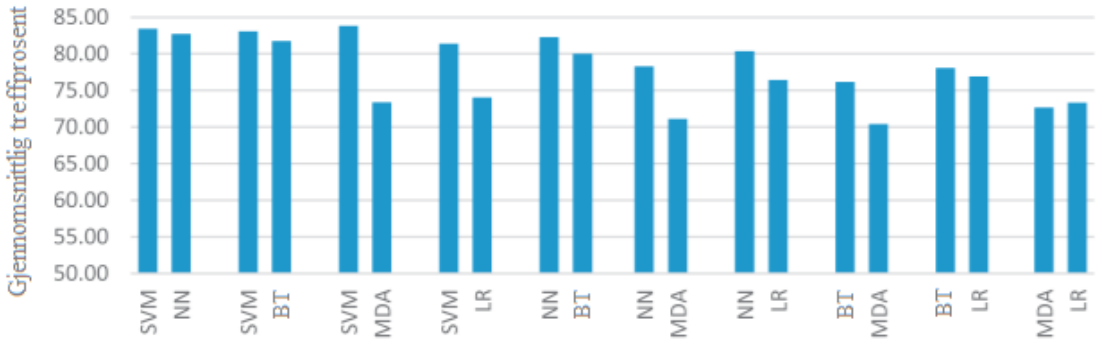
Den nyeste utviklingen innenfor teknologi og digitalisering har ført med seg utallige muligheter når det kommer til både prediksjon og maskinlæring. Begrepet maskinlæring brukes som en samlebetegnelse for en rekke metoder, hvor reglene utledes fra dataene systemet trenes på, også kjent som treningsdata (Kommunal- og moderniseringsdepartementet, u.å.). Dette i motsetning til regelbaserte systemer hvor reglene bestemmes av brukeren basert på erfaring eller regelverk. Alaka et al. (2017) går systematisk igjennom 49 akademiske artikler fra perioden 2010-2015 som har anvendt ulike statistiske- og maskinlæringsmetoder for å predikere konkurser. Det understrekes at de inkluderte artiklene i

denne studien innehar ulike formål. Noen forsøker å optimalisere én enkelt modell og andre forsøker å sammenligne flere modeller opp mot hverandre. Det som er felles for artiklene er at de har treffprosenter å vise til i sine resultater. De mest anvendte statistiske metodene blant artiklene er MDA og logistisk regresjon, som omtalt under Altmans Z-score og Ohlsons O-score. Videre er de maskinlæringsmetodene som oftest er tatt i bruk henholdsvis nevrale nettverk, støttevektormaskiner og beslutningstrær. Tabell 2.6 viser gjennomsnittlig treffprosent fra de studiene som har inkludert den aktuelle metoden i sin forskning. Antall studier som har inkludert de ulike metodene er også listet opp.

Tabell 2.6. Gjennomsnittlig treffprosent innenfor konkursprediksjon i perioden 2010-2015. De statistiske metodene er skrevet i kursiv, og resterende metoder er å anse som maskinlæringsmetoder.

Metode	Antall studier som inkluderte metoden	Gjennomsnittlig treffprosent
Nevrale nettverk	38	83,6
Støttevektormaskin	24	82,9
Beslutningstre	19	80,5
Logistisk regresjon	31	79,9
MDA	21	78,3

Det fremkommer av tabellen at maskinlæringsmetodene jevnt over presterer bedre enn de mer tradisjonelle statistiske metodene når de står overfor konkursprediksjon som problemstilling. Forskjellene mellom metodene er derimot små. Videre er det den hyppigste anvendte metoden nevrale nettverk som leverer høyest treffprosent blant de inkluderte studiene. Alaka et al. (2017) understreker at resultatene for hver av metodene varierer veldig blant de inkluderte studiene, og at de vil være sterkt avhengig av satte kriterier i forbindelse med data samt parameteroptimalisering overfor maskinlæringsmodellene. Derfor legges det også frem resultater som kun ser på treffprosent blant de studiene som har lagt frem en direkte parvis sammenligning mellom metoder (se figur 2.1). I disse tilfellene vil metodene baseres på samme databehandlingen og de samme forberedelsene, og derav sammenlignes på like vilkår. Som det fremkommer av figuren, presterer maskinlæringsteknikkene bedre enn de statistiske teknikkene i samtlige parsammensetninger. Når maskinlæringsmetodene settes opp mot hverandre, presterer støttevektormaskin marginalt bedre enn nevralt nettverk og beslutningstrær, samtidig som nevrale nettverk har et lite fortrinn direkte sammenlignet med beslutningstrær.



Figur 2.1. Gjennomsnittlig treffprosent studier med parvis sammenligning av metoder (Alaka et al., (2017).

Altman et al. (2017) er et eksempel på en studie som sammenligner både statiske- og maskinlæringsmetoder sine evner opp mot hverandre innenfor konkursprediksjon. Studien tar for seg i overkant av 13.000 selskaper i tidsperioden 1985-2005. Her sammenlignes blant annet de tidligere nevnte statistiske metodene MDA og logistisk regresjon opp mot maskinlæringsteknikker som støttevektormaskin, random forest og nevrale nettverk. Studien tar i bruk de samme variablene som Altman benyttet i den tidligere nevnte Z-score modellen fra 1968. Det konkluderes med at maskinlæringsmodellene presterer bedre enn de statistiske modellene når det kommer til treffprosent. Studien trekker allikevel frem at modellene til Altman og Ohlson fremdeles er høyst relevante, ettersom det er svært få studier gjort i ettertid som har utviklet like oversiktlige og enkelt forståelige modeller i tillegg til å inneha like gode prediktive evner (Altman et al., 2017).

3.0 Metode

Metodekapittelet er delt i to hoveddeler og vil først gå igjennom teoretisk rammeverk for inkluderte modeller i vår forskning med tilhørende styrker og svakheter. Deretter vil studiens metoder for evaluering og validering legges frem. Vi understreker at det teoretiske rammeverket er fremlagt i et klassifiseringsperspektiv og er sterkt forenklet. Dette som følge av at oppgavens formål er å sammenligne modellenes ytelse overfor foreliggende data, rettere enn å sammenligne modellenes dyptliggende egenskaper. For spesielt interesserte, henviser vi derfor til artikler som gjennomgår de ulike modellene i detalj.

3.1 Modeller

På bakgrunn av tidligere forskning, eksempelvis Alaka et al. (2017), har vi kommet fram til at vi skal sammenligne fem ulike maskinlæringsmodeller opp mot den utvidede SEBRA-modellen når det kommer til konkurransedyktighet innenfor konkursprediksjon. De utvalgte maskinlæringsmodellene er: beslutningstrær, random forest, k-nærmeste nabo, støttevektormaskin og nevrale nettverk.

3.1.1 Basis SEBRA-modell og utvidet SEBRA-modell

Det har i ettertid av den opprinnelige SEBRA-modellen blitt introdusert to nyere versjoner kalt for henholdsvis "basis SEBRA-modell" og "utvidet SEBRA-modell" som baserer seg på et mindre antall forklaringsvariabler (Bernhardsen og Larsen, 2007). Behovet for forbedring av den opprinnelige SEBRA-modellen skyldtes identifiserte svakheter som har blitt oppdaget gjennom lengre bruk av modellen. Dessuten har det over tid blitt tilgang på nye og flere data som gjør at videreutviklingen av modellen var ønskelig (Bernhardsen og Larsen, 2007). På samme måte som i den opprinnelige SEBRA-modellen, inkluderer basis-modellen nøkkeltall for inntjening, likviditet og soliditet. Dette gjelder også for inkludering av foretakets alder og en modifisert indikatorvariabel for svekket egenkapital (Bernhardsen og Larsen, 2007). Det blir videre introdusert bransjevariabler, basert på nøkkeltall for inntjening og soliditet. I den foregående modellen ble bransjevariablene beregnet for hele estimeringsperioden, hvorimot basis-modellen beregner de fleste bransjevariablene på årlig basis. Når det kommer til den utvidede SEBRA-modellen, inkluderer denne variabler som tar for seg omfanget av leverandørgjeld, ubetalt offentlige avgifter samt størrelse. Se sammenligning mellom inkluderte variabler i henholdsvis basis SEBRA og utvidet SEBRA i tabell 3.1. I de mørke feltene inngår det variabler som er inkludert i basis SEBRA, mens utvidet SEBRA inneholder alle variablene i tabellen. I kolonnen “variabeltype” finner vi ulike beskrivelser av variabeltypene som inngår i de respektive modellene. Eksempelvis er det fire variabeltyper knyttet til variabeldefinisjonen “Ordinært resultat før av- og nedskrivninger i prosent av total gjeld”.

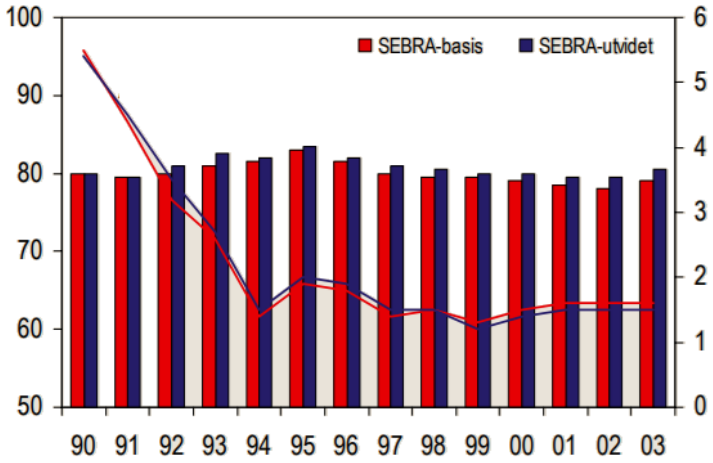
Tabell 3.1. Variabler som inngår i basis SEBRA (mørkere tonet felt) og utvidet SEBRA (hele tabellen), (Bernhardsen og Larsen, 2007).

Variabeldefinisjon	Variabeltype	Varierer over
Ordinært resultat før av- og nedskrivninger i prosent av total gjeld	Nøkkeltall Gjennomsnitt Standardavvik Korrelasjon med Norgesporteføljen	Foretak/år Bransje/år Bransje/år Bransje

Egenkapital i prosent av totalkapital Innskutt egenkapital mindre enn bokført egenkapital	Nøkkeltall Gjennomsnitt Indikator	Foretak/år Foretak/år Foretak/år
Likvider minus kortsiktig gjeld i prosent av omsetning	Nøkkeltall	Foretak/år
Alder (år) = 1, 2, 3,…,8	Indikatorer	Foretak/år
Sum eiendeler i faste kroner	Nøkkeltall	Foretak/år
Leverandørgjeld i prosent av totalkapital	Nøkkeltall	Foretak/år
Skyldige offentlige avgifter i prosent av totalkapital	Nøkkeltall	Foretak/år

I utgangspunktet er alle norske ikke-finansielle aksjeselskaper med totale eiendeler på over en halv million NOK inkludert i datagrunnlaget til Bernhardsen og Larsen (2007). Noen selskap ekskluderes grunnet mangler ved regnskapsinformasjon. Totalt er datagrunnlaget etter filtrering bestående av over én million årsregnskap, hvor tilhørende estimeringsperiode er i perioden 1990-2002.

I figur 3.1 sammenlignes basis SEBRA og utvidet SEBRA sine prediktive evner. Innenfor konkursprediksjon er det vanlig å sette et avskjæringsnivå (høyre akse) for predikerte konkurssannsynligheter hvor observasjoner over dette nivået klassifiseres som konkurs, og vice versa. Kurvene i figuren illustrerer dette avskjæringsnivået i det aktuelle året som ble benyttet for å oppnå en balansert treffprosent. Videre refererer en balansert treffprosent til en treffprosent hvor andelen riktige predikerte konkurs- og ikke-konkursselskaper er like. Den venstre aksen i figuren viser til denne balanserte treffprosenten. Som en kan lese av figuren er forskjellene mellom basis SEBRA og utvidet SEBRA minimale, selv om sistnevnte gjennomgående fra 1992 til 2003 oppnår en noe bedre balansert treffprosent. Bernhardsen og Larsen (2007) argumenterer videre for at dersom konkurs er en god erstatningsvariabel for mislighold og tap overfor store og små selskap, burde en legge vekt på modellen den utvidede SEBRA-modellen. Om det ikke er tilfellet, burde man heller legge vekt på klassifiseringen i basis SEBRA-modell.



Figur 3.1. Treffprosent og avskjæringsnivå for basis SEBRA og utvidet SEBRA (Bernhardsen og Larsen, 2007).

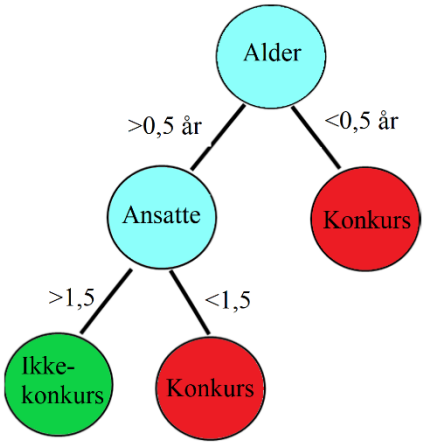
Som følge av mangel på bransjeinformasjon i denne studiens data, vil det anvendes en forenkling av den utvidede SEBRA-modellen som videre i oppgaven vil være merket og omtales som "utvidet SEBRA-modell*". Forenklingen vil ikke inkludere variabler for gjennomsnitt, standardavvik og korrelasjon med Norgesporteføljen når det kommer til ordinært resultat før av- og nedskrivninger i prosent av total gjeld (se tabell 3.1). Det inngår heller ikke en egen regnskapspost for skyldige offentlige avgifter i våre data, og vi har som følge av dette forsøkt å gjenskape denne variabelen så godt som mulig ved å trekke fra leverandørgjeld fra kortsiktig gjeld, og finne prosentandelen av dette i forhold til totalkapitalen. Kortsiktig gjeld omfatter blant annet leverandørgjeld, skyldige offentlige avgifter og annen kortsiktig gjeld. Vi vil gjennom dette stå igjen med skyldige offentlige avgifter og annen kortsiktig gjeld i prosent av totalkapitalen.

3.1.2 Beslutningstre

Modellen beslutningstre (BT) kan anvendes både i form av regresjoner og klassifisering. Disse modellene innebærer inndeling eller segmentering av de uavhengige variablene for å predikere den avhengige variabelen. Et beslutningstre vil kunne foreta både kategoriske og numeriske klassifikasjoner. I dette tilfellet er den avhengige variabelen en dummyvariabel med alternativene konkurs eller ikke-konkurs, og klassifikasjonen er kategorisk. Dette på samme måte som den avhengige variabelen til Ohlsons O-score og SEBRA-modellen. Et beslutningstre består av noder, løv og grener. Treet leses fra diagrammets øverste punkt, også kjent som roten elle rot-noden til treet. Dette er utgangspunktet for hver observasjon man ønsker å klassifisere. Hver node i treet (lyseblå sirkler i figur 3.2), representerer en beslutning

eller en regresjon. I denne oppgaven vil dette innebære at observasjoner blir atskilt ut ifra eksempelvis om selskapet er over en viss alder, eller har et visst antall ansatte (se figuren). Beslutningstreet ender opp i løv, som kan identifiseres ved at de kun har grener til seg. Det er her den endelige klassifikasjonen finner sted. I vårt tilfelle innebærer dette at det er her en observasjon klassifiseres som konkurs eller ikke-konkurs. Dette er en gjentakende prosess som gjøres på alle observasjoner i det aktuelle utvalget.

Trebaserte modeller er en utbredt metode når man har behov for å se hvilke beslutninger som blir foretatt. Det er enkelt å følge treets oppbygning og enklere å rettferdiggjøre en klassifikasjon når man konkret kan peke på hvilke beslutninger som tas innad i modellen ved estimering. På den annen side er de ikke nødvendigvis de beste når det kommer til nøyaktig prediksjon, sammenlignet med andre modeller innenfor "supervised learning"[3]. (James et al., 2013). Dessuten er en av de største svakhetene ved beslutningstrær at de er veldig utsatt for problemstillingen kjent som “overfitting”. Problematikken beskrives i detalj under kapittel 3.2.4. For flere styrker og svakheter med beslutningstrær som modell, se tabell 3.2.



Figur 3.2. Enkel illustrasjon av et beslutningstre.

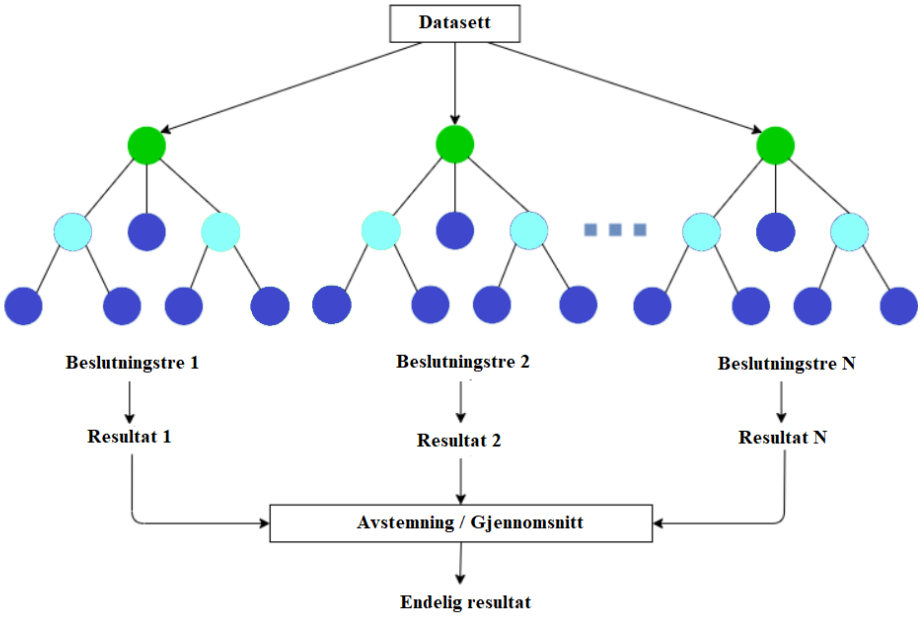
3.1.3 Random forest

Random forest (RF) er en anerkjent modell innenfor litteraturen. Det er en enkel maskinlæringsalgoritme å anvende i praksis og er kjent for å produsere nøyaktige resultater, også med liten grad av optimalisering. Modellen kan i likhet med beslutningstre anvendes

[3] Supervised learning: Kjent som veiledet læring på norsk. Algoritmen trenes med et datasett hvor både input-data og resultat er gitt. Følgelig får algoritmen tilgang til både oppgave og tilhørende fasit og bruker dette til å bygge ut modellen. Med dette som bakgrunn, vil modellen senere kunne gjøre klassifikasjoner basert på ny inndata (Kommunal- og moderniseringsdepartementet, u.å).

både til regresjoner og klassifikasjoner. Modellen bygger i stor grad på de samme prinsippene som beslutningstrær, da random forest utvikler en skog bestående av flere beslutningstrær. Videre brukes denne samlingen med trær til å foreta klassifikasjoner, vanligvis i likhet som ved "bagging"-metoden også kjent som "bootstrap aggregating". Da kjører algoritmen én observasjon gjennom alle beslutningstrærne, som til sammen utgjør random forest. Den mest hyppige klassifikasjonen blant disse beslutningstrærne utgjør den endelige klassifikasjonen (se figur 3.3). I vår oppgave innebærer dette at en observasjon klassifiseres etter hvorvidt flertallet av beslutningstrærne i skogen klassifiserer det aktuelle selskapet som konkurs eller ikke-konkurs. Dette er en prosess som gjentas for alle observasjonene fra treningsdataen. Ved å benytte seg av flere beslutningstrær oppnår algoritmen en kombinasjon av enkelheten forbundet med beslutningstrær samtidig som den øker graden av fleksibilitet, noe som fører til stor forbedring i prediksjonsevne og nøyaktighet.

Ved å ta i bruk det vektede snittet fra nodenes prediksjoner (i dette tilfellet vektet med antall konkurs eller ikke-konkurs klassifikasjoner), tilpasser algoritmen seg til dataene på en god måte samtidig som den tar hensyn til ikke-lineære effekter og unngår overfitting (Ishwaran, 2007). Dette er med på å senke variansen i modellen samtidig som det øker treffsikkerheten i den endelige prediksjonen. Evnen til å fange opp ikke-lineære sammenhenger, i tillegg til å forhindre stor varians og overfitting gjør random forest til en god modell i prediksjonsøyemed (se tabell 3.2 for flere styrker og svakheter forbundet med random forest).



Figur 3.3. Random forest.

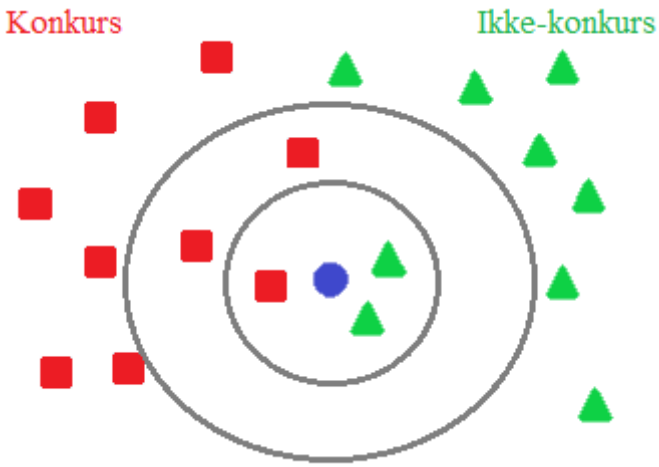
3.1.4 K-nærmeste naboer

K-nærmeste naboer (KNN) er en enkel maskinlæringsalgoritme som er lett å bruke i praksis. Videre kan den effektivt brukes til å løse både klassifiserings- og regresjonsproblemer. KNN er i motsetning til øvrige maskinlæringsmodeller ikke en tidkrevende klassifikator, verken når det kommer til utforming fra brukeren eller prosesseringstid for computeren. Årsaken til dette er enkel, ettersom det er treningsdatasettet som utgjør selve modellen (Hastie et al., 2009). Mange maskinlæringsalgoritmer forsøker å utvikle en global modell som er tilpasset all treningsdata. KNN vil på sin side kun utføre sin prediksjon basert på informasjon fra nærliggende observasjoner (se figur 3.4). Som følge av dette vil algoritmen være mer sensitiv overfor støy i datasettet sammenlignet med øvrige nevnte maskinlæringsalgoritmer (se tabell 3.2 for flere styrker og svakheter med k-nærmeste naboer). *K*'en i algoritmens navn, henviser til antallet nærmeste naboer som inkluderes for å klassifisere en ny observasjon. Ved lave verdier av *k*, vil algoritmen være spesielt utsatt for støy i datasettet.

Når man står overfor et klassifikasjonsproblem som i vårt tilfelle, vil observasjonene plottes i et flerdimensjonalt rom med tilsvarende antall dimensjoner som antall uavhengige variabler i datasettet (Cover og Hart, 1967). Der hvor trebaserte- og øvrige regelbaserte klassifikatorer ofte er begrenset til rettlinjede beslutningsgrenser, er KNN fleksible på dette området. Med beslutningsgrenser menes de regionene i det flerdimensjonale rommet som avgjør klassifikasjonen for en observasjon. Ved lave verdier av *k*, som tilsvarer klassifikasjon fra et lavt antall naboer, vil det foreligge liten grad av bias. Til gjengjeld vil dette også medføre stor grad av variasjon. I motsatt tilfelle, vil et stort antall *k* føre til liten grad av variasjon samtidig som dette vil medføre større grad av bias. Dette kan forklares gjennom at et lavere antall *k* i mindre grad vil kunne fange opp tendensene i datapunktene ettersom modellen kun vil se til et fåtall nærliggende observasjoner. Klassifikasjonen vil i så fall kunne bli basert på støy i datagrunnlaget rettere enn den potensielt foreliggende trenden i datapunktene. Dette er et eksempel på en mye omtalt problemstilling innenfor maskinlæring, nemlig avveiningen mellom bias og varians som er nærmere forklart under kapittel 3.2.4.

Etter å ha definert treningsdataen og valgt antallet nærmeste naboer klassifikasjonen skal være myntet på, må man beregne avstanden mellom observasjonene man ønsker å klassifisere (blå sirkel i figuren) og øvrige testobservasjoner i det flerdimensjonale rommet. Treningsobservasjonene må deretter sorteres etter hvilke som er nærmest observasjonen som skal klassifiseres. Den klassen som forekommer hyppigst blant de *k* naboene, blir da den

endelige klassifikasjonen. Dette kan eksemplifiseres gjennom figuren hvor den innerste grå sirkelen representerer *k*=3. De røde firkantene representerer konkursselskaper og de grønne trekantene representerer ikke-konkursselskaper. I dette tilfellet vil testselskapet (blå sirkel) klassifiseres som ikke-konkurs. Om man ser til den ytre grå sirkelen derimot, som representerer *k*=5 endres utfallet. Som eksemplifisert er denne klassifikasjonsmodellen sensitiv overfor antall *k* som velges for modellen. Innvirkningen av antall *k* kan derimot reduseres gjennom å vekte innflytelsen fra hver nærmeste nabo etter tilhørende avstand fra *i* (Hastie et al., 2009). Naboer lenger unna vil ved en slik vekting få lavere innvirkning enn de nærmeste. Ettersom vårt datasett består av kontinuerlige uavhengige variabler er det anbefalt å bruke euklidisk avstand for å måle avstanden mellom *i* og *k* (Ledolter, 2013). Dette er et mål som summerer absoluttverdien av avstanden mellom *i* og *k*, med alle dimensjoner tatt i betraktning.



Figur 3.4. Illustrasjon av hvordan beslutninger foretas basert på antall k i k-nærmeste nabo.

3.1.5 Støttevektormaskin

En støttevektormaskin er en maskinlæringsmodell som forsøker å etablere et plan som skiller ulike klassifiserte datapunker. I denne oppgaven vil støttevektormaskinen stå overfor et binært klassifikasjonsproblem, ettersom datasettet er delt opp observasjoner som er kategorisert som enten konkurs eller ikke-konkurs (se figur 3.5). Hver observasjon fra treningsdatasettet vil plasseres i et flerdimensjonalt rom, hvor antall dimensjoner tilsvarer antall forklaringsvariabler. Algoritmen vil deretter opprette et hyperplan[4] i det flerdimensjonale

[4] Hyperplan: Et (n-1) dimensjonalt plan eller skille i et n-dimensjonalt rom. Eksempelvis så vil hyperplanet være en endimensjonal linje overfor todimensjonal data, et todimensjonalt plan overfor tredimensjonal data etc.

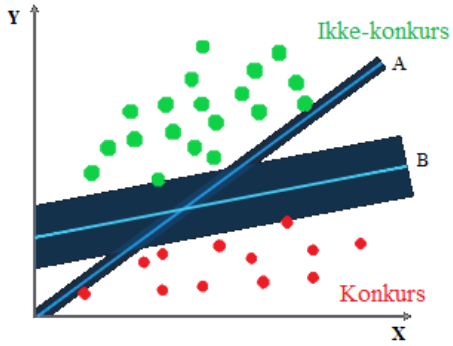
rommet som adskiller de to ulike kategoriene på en best mulig måte, og som videre vil virke som en klassifiseringsgrense. I figur 3.5 representerer de røde sirklene konkursselskaper, og observasjoner som er lokalisert på samme side av planet (A og B i figuren) vil også klassifiseres som konkurs.

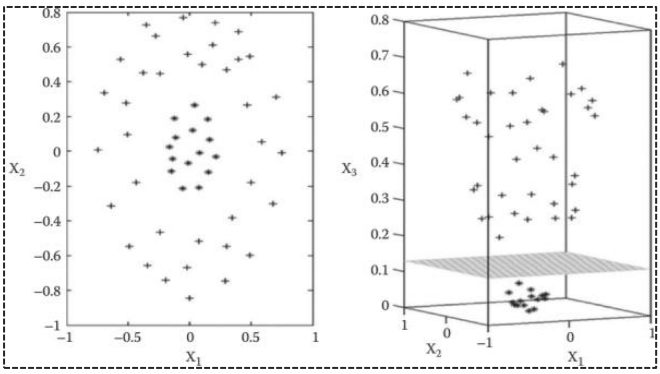
Å etablere et hyperplan som skiller datapunkter i et flerdimensjonalt rom er ikke unikt for SVM, men det er metoden for valg av hyperplan ut ifra de ulike alternativene som gir algoritmen sitt særpreg (Noble, 2006). Utgangspunktet baseres på å etablere hyperplanet på en slik måte at det maksimerer avstanden til de nærmeste datapunktene på hver side av planet (se figur 3.5). Denne metoden er svært skjør overfor avvikende datapunkter eller støy i datasettet. Ettersom datasett sjeldent er skjermet for støy og avvik vil kategoriene i liten grad kunne avskilles ved ovenfornevnte metode. For å ikke være like sensitiv overfor avvikende datapunkter etablerer algoritmen "soft margins" til hyperplanet. Dette er marginer som tillater datapunkter å krysse, uten å ha innvirkning på modellens hyperplan. Konsekvensen er at man tillater algoritmen å foreta feilklassifikasjoner, imot at man oppnår et mer anvendelig hyperplan overfor nye observasjoner. Man vil ikke tillate for mange feilklassifikasjoner, noe som bringer oss inn på en av algoritmens hyperparametere[5], nemlig Kernel-funksjonen.

Kernel-funksjonen er med på å avgjøre hvor mange observasjoner man kan tillate seg å feilklassifisere og med hvor stor margin. Den tar seg av datasett som tilsynelatende ikke lar seg skille av et hyperplan, ved å tillegge dataen én ytterligere dimensjon på en slik måte at tilsynelatende uadskillelige klasser fremstår adskillelige i dette nye rommet (se figur 3.6). Kernel-funksjonen et matematisk triks som eksempelvis lar SVM foreta en todimensjonal klassifikasjon på et datasett som i utgangspunktet består av endimensjonal data (Noble, 2006). Dette er et eksempel på at ved bruk av riktig Kernel-funksjon, så vil data som fremstår uadskillelig i en gitt dimensjon, kunne separeres i et rom av høyere dimensjon. I figur 3.6 eksemplifiseres dette gjennom datapunkter som i utgangspunktet ser uadskillelige ut i et todimensjonalt plan, som tillegges en tredje dimensjon og følgelig lar seg skille av et hyperplan i dette rommet. Støttevektormaskiner har vist seg å levere gode resultater innenfor konkursprediksjon sammenlignet med øvrige maskinlæringsmodeller i tidligere forskning (Alaka et al., 2017). Til tross for dette, foreligger det svakheter ved algoritmen forbundet med

[5] Hyperparameter: En parameter som ikke estimeres ved modellens trening, men som må settes i forkant av treningen. Se kapittel 3.2.5 om hyperparameter tuning.

mulighetene for tolkning variablene og viktigheten forbundet med disse. Se tabell 3.2 for flere styrker og svakheter med støttevektormaskiner.





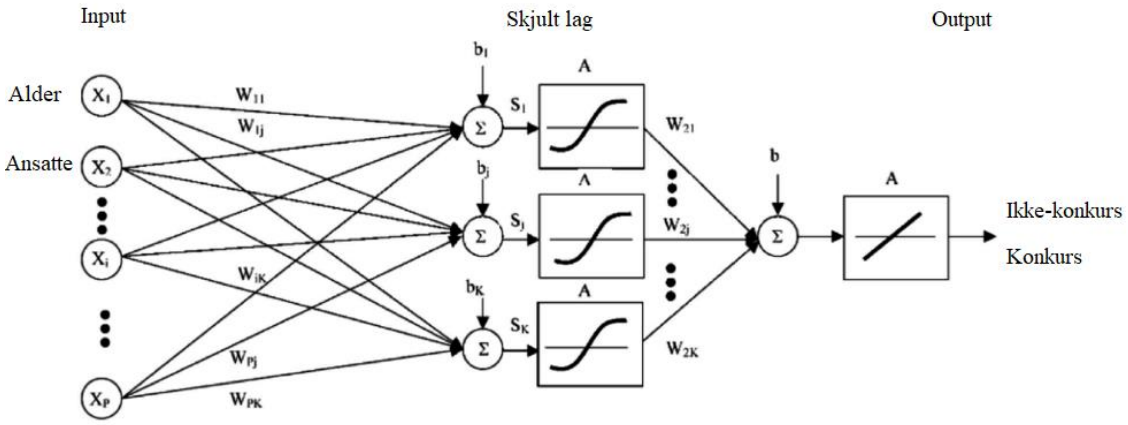
Figur 3.5. (Venstre). Etablering av hyperplan i todimensjonalt rom.
Figur 3.6. (Høyre). Datapunkter tillegges én dimensjon som medfører mulighet for separasjon fra hyperplan.

3.1.6 Nevrale nettverk

Nevrale nettverk (NN) er en algoritme som definerer et fordelt nettverk av beregninger. Algoritmen består av ulike former for lag eller nivåer med noder, hvorav inputvariablene i datagrunnlaget utformer det første laget og det siste laget produserer outputvariabler. Det er dette siste laget som tilegner en klassifikasjon til den aktuelle observasjonen man ønsker å predikere. Det første og siste laget, også kjent som inputlaget og outputlaget, utgjør rammen til algoritmen. Disse to lagene er som oftest koblet sammen av en tredje form for lag, som innenfor maskinlæring er kjent som det skjulte laget. Dette omfatter gjerne flere lag med noder, hvor hver enkelt node er koblet til samtlige noder på neste nivå. Da både antall skjulte lag og antall noder innad i hvert lag er noe som må bestemmes i forkant av estimeringen, vil de være definert som eksempler på modellens hyperparametere. Videre representerer koblingene mellom hver node parameterverdier som blir estimert når algoritmen blir tilpasset til treningsdataen. Nevrale nettverk er en modell som i større grad kan identifisere komplekse sammenhenger mellom den avhengige og de uavhengige variablene. Modellen har dessuten levert gode resultater innenfor konkursprediksjon innenfor tidligere forskning (Alaka et al., 2017).

Videre i oppbyggingen av det nevrale nettverket, må man bestemme nettverkets aktiveringsfunksjoner som avgjør hvilken output eller aktivitet en node skal gi ut ifra ulike inputverdier. Inputverdien, eller verdien til en uavhengig variabel fra en enkelt observasjon i datasettet, blir vektet og deretter tillagt en bias, som til sammen utgjør x-verdien i valgte aktiveringsfunksjon (StatQuest, 2020) (se figur 3.7). I denne oppgaven vil disse inputnodene

eksempelvis være antall ansatte i et selskap og selskapets alder. Vektingen og bias blir bestemt og justert med minimering av avviket mellom modellens endelige estimerte prediksjon og faktisk utfall som hovedmål. Dette er en metode som kalles “backpropagation” (Graupe, 2013). I vårt tilfelle hvor det er flere inputnoder/uavhengige variabler, vektes verdiene, legges deretter sammen og til slutt tillegges en bias som totalt utgjør aktiveringsfunksjonens x-verdi. Ved å tillegge flere inputverdier, vil x-verdiene med tilhørende y-verdi utforme et plan som beskriver utviklingen i den avhengige variabelen som funksjon av endring i de uavhengige variablene. Ved å kombinere alle utformede regresjoner fra nodene på samme nivå etablerer man i sum én regresjon som vil kunne fange opp komplekse sammenhenger i datautvalget. Følgelig kan man forenklet si at inputnodene reagerer på den informasjonen den får tildelt gjennom datautvalget. Videre vil aktiviteten hos de ulike nodene i de skjulte lagene være bestemt av aktiviteten i foregående ledd (inputnodene) tillagt vekting og bias i selve koblingen mellom nodene. Avslutningsvis så vil aktiviteten i outputnodene i likhet med beskrivelsen ovenfor være bestemt av aktiviteten i foregående ledd (skjulte lag) og også her tillegges det en vekting og bias i selve koblingen mellom nodene. Mye av kritikken rettet mot nevrale nettverk ligger i kompleksiteten i forbindelse med tolkningen av dens vekting og bias av ulike variabler (Alaka et al, 2017). Til tross for at algoritmen ofte leverer gode resultater, vil det være vanskelig for brukere å tolke hvorfor resultatene blir som de blir. For flere styrker og svakheter med nevrale nettverk, se tabell 3.2.



Figur 3.7. (Vuckovic et al., 2002). Nevralt nettverk med ett skjult lag og p input-variabler.

3.1.7 Styrker og svakheter ved modellene

Hver av de fremlagte modellene vil ha sine fordeler og ulemper. Med dette vil også modellene hevde seg ulikt i møte med forskjellige problemstillinger. Vi har tidligere sammenlignet modeller med hensyn til treffprosent hvor nevrale nettverk og støttevektormaskiner er de modellene som har kommet best ut i tidligere forskning. Disse er på den annen side kjent for å være lite transparente når det kommer til resultater. I vårt tilfelle vil dette innebære at det vil være vanskelig å tolke nøyaktig hva som er årsaken til at et selskap går konkurs eller ikke. Maskinlæringsmodeller kan sies å være negativt korrelerte når det kommer til forholdet mellom treffsikkerhet og hvor transparente de er med tanke på resultatene (Alaka et al., 2017). Som eksemplifisert gjennom nevrale nettverk og støttevektormaskiner, innebærer høy treffsikkerhet ofte lav grad av tolkningsmuligheter, og vice versa. På den andre siden av spekteret finner vi beslutningstrær som belyst gjennom tidligere forskning leverer noe lavere treffprosenter enn øvrige maskinlæringsmodeller, men til gjengjeld er å anse som tydelige når det kommer til tolkning. Det foreligger også sammenligningsgrunnlag med tanke på blant annet kompleksitet forbundet med implementering, hvor robust modellen er overfor støy og hvor krevende modellen er å kjøre for computer. Alle ovenfornevnte forhold fremkommer som enten styrker eller svakheter for de ulike maskinlæringsmodellene i tabell 3.2. Tabellen er utformet med informasjon fra metaanalysene til Alaka et al. (2017) og Ravi Kumar og Ravi (2007) som grunnlag.

Tabell 3.2. Styrker og svakheter ved ulike maskinlæringsmodeller (Ravi Kumar & Ravi, 2007), (Alaka et al., 2017).

Modell	Styrker	Svakheter
BT	○ Kan foreta både klassifikasjon og regresjon ○ Enkle å tolke gjennom oversiktlige “if-then”- regler ○ Robust overfor manglende data	○ Krever store datagrunnlag for å levere pålitelige prediksjoner ○ Utsatt for overfitting ○ Lavere treffsikkerhet innenfor konkursprediksjon i tidligere forskning, sammenlignet med øvrige modeller
RF	○ Kan foreta både klassifikasjon og regresjon ○ God overfor store datasett med stor grad av dimensjonalitet (høyt antall variabler) ○ Robust overfor manglende data ○ Enkelt å tolke variablenes viktighet med tanke på klassifikasjonen	○ Utsatt for overfitting (men i mindre grad enn beslutningstrær) ○ Krevende for computer å gjennomføre optimalisering av mange hyperparametere ○ Krevende algoritme for computer som følge av at den tar i bruk mengder av beslutningstrær ○ Tidkrevende treningsfase

	○ Presterer ofte godt med liten grad av optimalisering	
KNN	○ Lite tidkrevende oppbygging av algoritmen og ingen treningsperiode i motsetning til øvrige modeller ○ Baserer sine beslutninger på lokal informasjon rettere enn global ○ Fungerer godt både ved klassifikasjon og regresjon ○ Tilpasses umiddelbart ved innhenting av ny treningsdata	○ Sensitiv overfor støy og manglende data, spesielt ved klassifikasjon basert på lavt antall naboer ○ Lite egnet overfor store datagrunnlag ettersom kalkulering av avstand mellom observasjon og naboer krever mye fra computer ○ Lite egnet for datasett med mange variabler, grunnet mer kompleks kalkulering av avstand til naboer ○ Lite egnet ved ubalanserte klasser i datagrunnlaget
SVM	○ Høy treffsikkerhet innenfor konkursprediksjon i tidligere forskning ○ Presenterer global optimal løsning/skille for problemstillingen ○ Kan levere gode resultater med lite datagrunnlag ○ Robust overfor overfitting	○ Lite transparent ○ Valg av Kernel-funksjonen med tilhørende parametere er en krevende problemstilling ○ Svært langsom i testfasen ○ Algoritme med høy kompleksitet som krever mye fra computer ved kjøring
NN	○ Høy treffsikkerhet innenfor konkursprediksjon i tidligere forskning ○ Svært god overfor funksjonstilnærming, optimaliseringsoppgaver, klassifisering og levering av prognoser. Dette avhenger av nettverkets struktur	○ Lite transparent ○ Beslutninger og tolkning rundt modellens parametere er utydelig ○ Mange strukturer krever mye treningsdata, og mange gjennomkjøringer ved trening ○ Krevende for computer ved gjennomkjøring

3.2 Evaluering og validering

Dette underkapittelet vil ta for seg studiens evaluerings- og valideringsmetoder, samt diskutere hvordan dette har innvirkning på modellenes bias og varians. Det vil også legges frem anvendt prinsipp for justering av modellenes hyperparametere. Avslutningsvis vil det legges frem informasjon som styrker studiens repliserbarhet.

3.2.1 Treffprosent

Treffprosent er som navnet tilsier basert på andelen av det totale antallet prediksjoner som modellene klarer å predikere korrekt. Ettersom vi i denne oppgaven studerer modeller som skal predikere konkurser og ikke-konkurser, vil modellenes respektive treffprosent følgelig tilsvare andelen korrekte predikerte konkurser og ikke-konkurser ut ifra det totale utvalget observasjoner. Jo høyere treffprosent eller treffandel, desto bedre modell, om en utelukkende legger dette målet til grunn. Det fremkommer av studien til Alaka et al. (2017) at treffprosent er den mest brukte evalueringsmetoden innenfor de gjennomgåtte akademiske artiklene om konkursprediksjon. Dette gjør seg også gjeldende i Ravi Kumar og Ravi (2007) sin gjennomgang av tidligere studier av konkursprediksjon. Det er derimot svakheter ved å utelukkende se på treffprosent, da det vektlegger alle riktige og uriktige prediksjoner helt likt. Dette er en problematikk som kan illustreres tydeligere gjennom anvendelse av en "confusion matrix" eller en såkalt forvirringsmatrise.

Den avhengige variabelen i vårt variabelsett er som tidligere nevnt dummyvariabelen konkurs/ikke-konkurs. Om man legger sammen modellenes predikerte utfall med de faktiske utfallene får vi fire ulike hendelser, illustrert i forvirringsmatrisen illustrert i figur 3.8.

		Faktisk klasse	
		P	**N**
Predikert klasse	**P**	**E**kte **P**ositiv	**F**alsk **P**ositiv
	N	**F**alsk **N**egativ	**E**kte **N**egativ

$$\text{Falsk positivrate} = \frac{FP}{FP+EN} \qquad \text{Ekte positivrate} = \frac{EP}{EP+FN}$$

$$\text{Spesifisitet} = \frac{EN}{EN+FP} \qquad \text{Treffprosent} = \frac{EP+EN}{EP+FP+EN+FN}$$

Figur 3.8. Forvirringsmatrise og beregninger.

Matrisen illustrerer kombinasjoner av predikert og faktisk utfall. Modellene har to mulige former for korrekt prediksjon og to mulige former for feilprediksjoner. Utfallet ekte positiv (EP) finner sted dersom modellene predikerer konkurs samtidig som dette skjer. Den andre korrekte prediksjonen er ekte negativ (EN) som finner sted dersom modellen predikere ikke-konkurs samtidig som dette skjer. Det er to mulige former for feilprediksjon, nemlig falsk positiv (FP) og falsk negativ (FN). Disse formene for feilklassifikasjon er i statistikken kjent

som henholdsvis type-I-feil og type-II-feil. Ved FP (type-I-feil) predikerer modellen konkurs samtidig som selskapet ikke går konkurs. Ved FN (type-II-feil) predikerer modellen ikke-konkurs samtidig selskapet går konkurs.

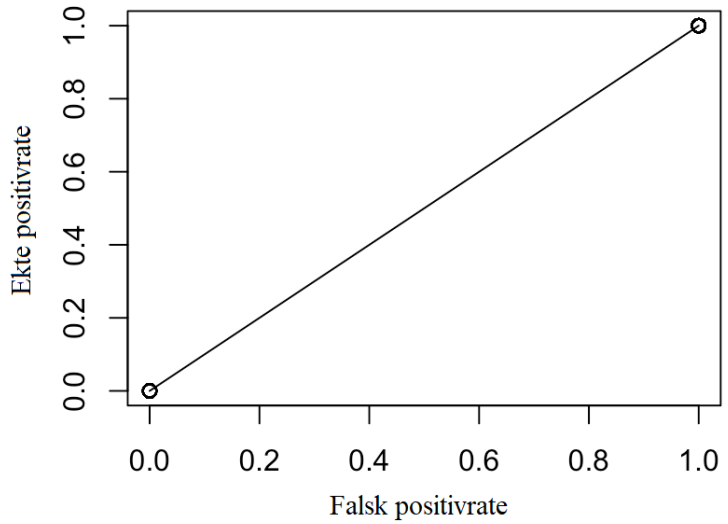
I forbindelse med konkursprediksjon, vil det være langt større konsekvenser ved en type-II-feil. Eksempelvis vil en finansiell institusjon som legger prediksjonen til grunn for beslutninger vedrørende utlån, kunne pådra seg store tap. På den annen side vil det derimot medbringe store konsekvenser for selskaper som søker lån ved overvekt av type-I-feil. Dette illustrerer svakheten i å utelukkende anvende treffsikkerhet som evaluering, når en står overfor ulikheter i konsekvenser forbundet med feilklassifisering. I denne oppgaven ønsker vi å se på modellers objektive prediktive evne rettere enn å utforme en prediksjonsmodell spesielt spisset til bruk for aktører som er mer sensitive overfor én enkelt type feilklassifikasjon. Overfor binære klassifikasjoner med jevnfordeling i klasser i den utvalgte dataen som i vårt tilfelle (se kapittel 3.2.3), vil det være mest hensiktsmessig å anvende en klassifiseringsterskel lik 0,5. Dette innebærer at predikerte sannsynligheter for konkurs som er større enn 50 prosent vil klassifiseres som konkurs, og vice versa. Prediksjon med ulike klassifiseringsterskler er noe som vil diskuteres nærmere under kommende underkapittel.

3.2.2 Reciever operating characteristics

Bruk av ROC-kurver er en teknikk som anvendes for å visualisere, organisere og velge klassifikatorer basert på deres ytelse (Fawcett, 2006). Om man ser til forvirringsmatrisen ovenfor, vil klassifiseringen modellen gjør være sterkt avhengig av hvilken klassifiseringsterskel som settes. De ulike kravene til sikkerhet man stiller overfor modellene for å kunne foreta en klassifikasjon, fører til ulike forvirringsmatriser på det samme datasettet. Ved å senke klassifiseringsterskelen vil modellene i større grad fange opp faktiske konkurser, men det vil også føre til økt feilprediksjon i form av falske positive hendelser, og vice versa. En ROC-kurve vil ta for seg de ulike potensielle forvirringsmatrisene, og på en oversiktlig måte oppsummere samspillet mellom modellenes klassifiseringsterskel og treffsikkerhet. Som følge av dette, egner bruk av ROC-kurven seg til å sammenligne ytelsen til konkurrerende klassifiseringsmodeller, med alle mulige klassifiseringsterskler tatt i betraktning. ROC-kurven plottes i et diagram, hvor x-aksen representerer falsk positivrate og y-aksen representerer ekte positivrate. Se beregning av falsk- og ekte positivrate i figur 3.8 ovenfor.

Ekte positivrate, også omtalt som sensitivitet, beskriver i denne oppgaven andelen konkursselskaper som har blitt korrekt predikert av modellen. Falsk positivrate, eller (1-

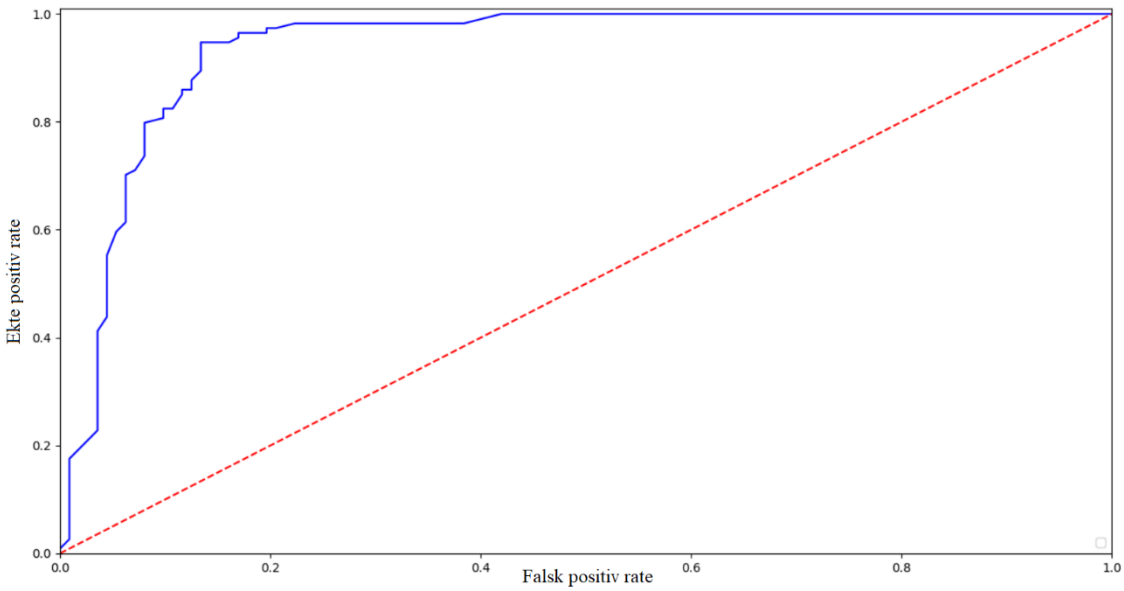
spesifisitet), tar for seg de faktiske ikke-konkursene og beskriver andelen av disse som ble feilpredikert som konkursselskaper. Figur 3.9 illustrerer grunnlinjen i et ROC-diagram, som representerer alle punkter hvor ekte positivrate = falsk positivrate. Alle punkter langs denne diagonalen representerer tilfeller hvor andelen korrekt klassifiserte konkursobservasjoner er lik andelen feilprediksjon blant ikke-konkurser. Dette er et resultat en skal kunne oppnå selv ved tilfeldig klassifisering, ettersom en står overfor prediksjon med kun har to mulige utfall. Dersom klassifiseringsterskelen går mot null, vil samtlige observasjoner klassifiseres som konkurs. Dette medfører at alle konkurser fanges opp og følgelig vil den ekte positivraten bli 1. Det vil samtidig medføre at samtlige ikke-konkurser feilklassifiseres, som vil resultere i at også den falske positivraten blir 1. Dette representerer punktet (1,1) i ROC-diagrammet, og en tilsvarende høy klassifiseringsterskel utgjør origo (0,0). Det er diagonalen mellom disse punktene, Y=X, som utgjør grunnlinjen.



Figur 3.9. Grunnlinje i et ROC-diagram. Arealet under kurven er lik 0,5.

Den beste terskelverdien i forbindelse med klassifisering fra modellene er det punktet som gir størst forholdstall mellom ekte positivrate og falsk positivrate. Jo bedre en modell yter, desto lenger opp mot venstre hjørne eller punkt (0,1) vil ROC-grafen trekkes (se figur 3.9). Perfekt prediksjon vil gi en ROC-kurve som følger y-aksen opp til punkt (0,1), før den går lineært fra dette punktet og ut til punkt (1,1). Hvor god en modell er, blir gjennom ROC avgjort av arealet under kurven. Dette arealet betegnes som AUC-verdien. Arealet under kurven ved perfekt prediksjon er lik 1, mens ved tilfeldig prediksjon som omtalt tidligere, vil arealet under kurven (diagonalen) være lik 0,5. Ved å plotte ROC-kurver for de ulike modellene, vil en kunne sammenligne tilhørende AUC-verdi og med det identifisere hvilke modeller som predikerer best. Hva som er en god AUC-verdi er relativt og avhenger av studiens formål,

men generelt sett kan man si at en AUC-verdi på over 0,9 blir sett på som fremragende, mellom 0,8 til 0,9 er utmerket, og verdier mellom 0,7 til 0,8 blir sett på som et akseptabelt nivå (Mandrekar, 2010).



Figur 3.10. ROC-kurve.

Treffprosent kan fremstå som et mer intuitivt mål på prestasjon, men statistisk presterer AUC like konsistent, samtidig som det er statistisk mer diskriminant enn treffprosent. Dette gjør AUC til et bedre egnet mål på prediksjonsevnen overfor maskinlæringsalgoritmer (Huang og Ling, 2005). Studien til Huang og Ling (2005) viser dessuten at dette gjelder både for binære og multiklasse-datasett, om så fordelingen av data er balansert eller ubalansert. Dette er en av fordelene til ROC-kurver, som muliggjør visualisering, organisering og evaluering uavhengig av klassedistribusjon og kostnader ved feilprediksjon (Fawcett, 2006). På den annen side, vil det i likhet som med treffprosent som evalueringsmål, foreligge en svakhet i AUC i form av neglisjering av kostnader forbundet med feilklassifikasjon. Med andre ord vektlegges type-I-feil og type-II-feil likt, hvor problematikken i dette har blitt drøftet i forbindelse med treffprosent.

3.2.3 Kryssvalidering

Kryssvalidering er en hyppig anvendt valideringsmetode innenfor maskinlæring for å imøtekomme problematikk forbundet med overfitting. Ved overfitting vil modellene prestere med lavere presisjon på observasjoner utenfor treningsdatasettet, sammenlignet med observasjoner som inngår i treningsdatasettet. Dette som følge av for stor vektlegging på

forhold innad i treningsdataen, som vil gjøre modellen lite egnet til å predikere øvrige observasjoner. For å validere at en modell også fungerer overfor nye observasjoner, vil det være nødvendig å teste modellene mot data som ikke inngår i treningsdatasettet (Kirkos, 2012). *K*-gruppert kryssvalidering deler datasettet opp i *k* grupper som brukes til å estimere parameterne i den aktuelle modellen og i tillegg evaluere modellens ytelse. *K*-1 grupper brukes til å estimere modellen, og den gjenværende gruppen som ble utelatt ved estimering vil avslutningsvis brukes som valideringssett overfor modellen. Denne prosessen gjentas til alle de *k* gruppene har blitt brukt til validering. *K*-gruppert kryssvalidering er en anbefalt metode ettersom det eliminerer variasjon mellom observasjoner samtidig som det minimerer innvirkning fra bias (Han et al., 2000). Videre vil randomisert utvalg av observasjoner til de ulike gruppene føre til at rekkefølgen til de nedlastede observasjonene ikke påvirker de prediktive evnene til modellene.

I denne studien har vi valgt å anvende femdelt kryssvalidering, hvor hver av gruppene inneholder det samme forholdstallet mellom konkurs- og ikke-konkursobservasjoner. Studien til Alaka et al. (2017) viser til at halvparten av de gjennomgåtte artiklene innenfor konkursprediksjon benyttet en 50-50 fordeling mellom konkursselskaper og ikke-konkursselskaper. Videre trekkes det frem at 80 prosent av studiene benyttet en fordeling på 60-40 eller jevnere. Med dette som bakgrunn vil vi i denne oppgaven benytte oss av en 50-50 fordeling som vil være gjennomgående i samtlige grupper fra kryssvalideringen. Dette vil bidra til å minimere modellenes bias i retning av å predikere et selskap som ikke-konkurs ettersom det er en mindre skjevfordeling av de ulike klassifikasjonene enn i det opprinnelige datagrunnlaget. I vårt tilfelle hvor det endelige datagrunnlaget består av i underkant tre prosent konkursselskaper, ville modellene oppnådd en svært høy treffprosent og AUC-verdi ved å trekkes i retning eller til og med utelukkende klassifisere selskaper som ikke-konkurs.

For at treningen og valideringen blant de fem gruppene skal foregå på like vilkår, må fordelingen av observasjoner til de ulike gruppene skje deretter. Den begrensende faktoren i datagrunnlaget i denne oppgaven er antallet konkursselskaper (se kapittel 4) og på bakgrunn av dette vil samtlige av disse benyttes i kryssvalideringen. Disse observasjonene vil fordeles slik at hver av de fem gruppene har en jevnfordeling av konkursselskaper, også med hensyn til næringskoder. For å oppnå ønsket jevnfordeling, også mellom konkurs- og ikke-konkursselskaper, må det trekkes ut et tilsvarende antall ikke-konkurser fra datagrunnlaget. Disse trekkes ut tilfeldig, med forutsetning om at hver av gruppene i kryssvalideringen får en jevnfordeling med hensyn til konkurs- og ikke-konkursselskaper fra hver av næringskodene

innad i og mellom gruppene. Gjennom dette vil modellene trenes og testes basert på like fordelinger av de ulike bransjene. Dette for å forhindre en potensiell skjevfordeling ved tilfeldig fordeling, hvor modellene kan ende opp med å testes på observasjoner fra en bransje som ikke inngår i treningsdatasettet, og vice versa.

3.2.4 Bias-varians

Avveining mellom bias og varians er en mye omtalt problemstilling innenfor maskinlæring ettersom flere av modellene står overfor problematikk forbundet med overfitting og underfitting. Med bias menes en modell sin manglende evne til å fange opp den sanne sammenhengen i datasettet (Statquest, 2018). Dette kan eksemplifiseres gjennom en lineær regresjons manglende evne til å fange opp en eksponentiell sammenheng i et datasett. En modell med stor grad av bias, betegnes gjerne som underfit. I motsatt tilfelle, ved overfitting, foreligger det for stor tilpasning til treningsdataen hvor modellen potensielt fanger opp sammenhenger som ikke nødvendigvis er til stede. Dette medfører som omtalt tidligere svake prediktive evner på observasjoner utenfor treningsdatasettet. At det foreligger stor forskjell i en modells treffsikkerhet mellom treningsdata og testdata, er kjent som stor varians. En modell med stor varians og liten bias, vil kunne levere gode resultater ved enkelte tilfeller, men til gjengjeld predikere veldig dårlig i andre tilfeller. I motsatt tilfelle vil en modell med lav varians og stor bias, prestere jevnt overfor ulike typer data, men til gjengjeld predikere jevnt dårlig. En optimalt tilpasset modell vil takle begge ovenfornevnte utfordringer på en god måte. Dette oppnår man om modellen fanger opp den sanne sammenhengen i dataen (liten grad av bias), samtidig som modellen presterer konsistent overfor nye datasett (liten grad av varians).

Gjennom anvendelse av *k*-gruppert kryssvalidering hensyntar man problemstillingen forbundet med varians og bias. Dette gjøres gjennom valg av antall grupper (*k*), hvor et lavt antall grupper resulterer i lavere bias og høyere varians, og vice versa (James et al., 2013). Kohavi (1995) viste at bias og variansen til en modell forbedres betraktelig når det inngår minst fem grupper i kryssvalidering, noe som understrekes av Meese og Viken (2019). Som følge av at vi har et datagrunnlag med lavt antall observasjoner sett i maskinlæringssammenheng, velger vi å anvende et minst mulig antall grupper i kryssvalideringen uten at det skal gå på bekostning av modellenes prediksjonsevne. Vi har som omtalt valgt femdelt kryssvalidering noe som underbygges som fornuftig av James et al. (2013), ettersom denne inndelingen resulterer i hverken høy grad av bias eller varians.

3.2.5 Hyperparameter tuning

Denne studien anvender "grid search" som metode for å optimalisere de ulike modellenes hyperparametere. For å gjennomgå prinsippene ved grid search, vil det være nyttig å først klargjøre forskjellen mellom en modell sine parametere og hyperparametere, også kalt tuningparametere. Parametere og deres verdier er med på å avgjøre modellens prediktive evner og er noe som læres eller estimeres gjennom data. Dette kan for eksempel være koeffisienter i en lineær regresjon eller støttevektorer i SVM. Der hvor parametere er noe som i stor grad bestemmes av modellene selv gjennom anvendelse av treningsdata, er hyperparametere noe som må settes manuelt i forkant av estimeringen (Brownlee, 2017). Hyperparameterne er følgelig ikke noe som estimeres gjennom data, men er allikevel viktige for modellens ytelsesevne samtidig som de påvirker modellens øvrige parametere. Derfor er det viktig å "tune" eller justere hyperparameterne i en slik grad at du oppnår best mulig ytelse gitt den prediktive problemstillingen.

"Mange modeller har viktige parametere som ikke kan estimeres fra datasettet direkte... Denne typen modellparametere refereres til som en tuningparameter ettersom det ikke foreligger noen analytisk formel for å kalkulere en egnet verdi" -(Kuhn og Johnson, 2013).

Når man skal bestemme en hyperparameter er det flere potensielle fremgangsmåter. En kan benytte tommelfingerregler eller se til liknende problemstillinger og hvilke parameterverdier som har gitt best resultat. Et annet alternativ er å benytte seg av grid search som vi vil gjøre i denne oppgaven. Dette er en metode som baserer seg på å justere hyperparametere på en slik måte at man avdekker hvilke innstillinger som gir den beste modellen når det kommer til den aktuelle prediksjonen (Brownlee, 2017). Hyperparametere for eksempelvis random forest kan være maksimalt antall løv, maksimalt antall nivåer i hvert beslutningstre eller antall beslutningstrær som til sammen skal utgjøre den totale skogen. Ved å liste opp hvilke alternative verdier vi ønsker å teste for hver enkelt hyperparameter, vil modellen velge ut de ulike kombinasjonene og identifisere hvilken verdisammensetning for hyperparameterne som fører til det beste resultatet.

3.2.6 Repliserbarhet

All estimering og testing av modellene ble utført i Microsoft Visual Studio Community 2019 ved bruk av programmeringsspråket Python. All kode som er brukt kan innhentes ved å lime inn adressen https://github.com/haae-hub/KonkursPrediksjonML, i "repository location" under "clone a repository" i Microsoft Visual Studio. For å styrke mulighetene for replisering legges det ved informasjon vedrørende anvendt operativsystem og programversjoner ved utførte analyser i appendiks 2. Videre er det anvendt samme verdi for "random number" i estimering av samtlige modeller. Dette medfører at kryssvalideringen som utføres gir et tilfeldig utvalg av observasjoner i de fem ulike gruppene, hvor denne tilfeldig utvalgte inndelingen videre brukes gjennomgående for alle modellene. For å oppnå lignende resultater som i denne oppgaven forutsettes det at datagrunnlag og ekskludering av observasjoner er utført i tråd med beskrivelsen i følgende kapittel om data.

4.0 Data

Datagrunnlaget vårt består av regnskapsdata hentet ut fra Proff Forvalt sine databaser (Proff Forvalt, u.å.). Dataene er ikke åpent tilgjengelige da tilgangen er gitt mot betaling. Grunnet begrensning i antall regnskap som kan lastes ned pr døgn har vi kun hentet ut regnskapsdata til selskaper som har levert minst ett årsregnskap i perioden 2017-2019. Videre har vi tatt utgangspunkt i de ti bransjene med høyest antall konkurser i ovenfornevnte periode. Dette for å oppnå et størst antall konkursobservasjoner i vårt datagrunnlag. For full liste med inkluderte bransjer sortert etter antall konkurser med tilhørende NACE-kode, se appendiks 3. Årsregnskapene representerer informasjon om selskapene ved årets utgang, og i de tilfeller hvor selskaper har gått konkurs, knyttes konkursen opp mot det siste leverte årsregnskapet for det aktuelle selskapet. Årsregnskap fra utgående balanse (UB) 2019 er også knyttet opp mot den offentlig tilgjengelige informasjon hentet ut fra Skatteetaten (2020) om tildelt kontantstøtte i forbindelse med Covid-19 og medfølgende restriksjoner. Vårt totale datagrunnlag utgjør 49.399 årsregnskap i perioden 2017-2019, hvorav 46.711 ikke har gått konkurs og resterende 2.688 har gått konkurs.

Forskningen til Alaka et al. (2017) trekker frem nevrale nettverk som den modellen som stiller høyest krav når det kommer til antall observasjoner i datagrunnlaget. I deres gjennomgang av tidligere studier innenfor konkursprediksjon fremkommer det at 59 prosent av de gjennomgåtte artiklene som inkluderer nevrale nettverk som metode hadde datagrunnlag på under 1.000 selskaper, som indikerer at vårt noe begrensede datagrunnlag i

maskinlæringssammenheng ikke nødvendigvis vil by på store problemer.

4.1 Avhengig variabel

I denne oppgaven har den avhengige variabelen to mulige utfall, konkurs eller ikke-konkurs. Vi definerer den som en dummyvariabel som tar verdien 1 ved konkurs og 0 ellers. Definisjonen av konkurs ble klargjort i introduksjonen av oppgaven, men selv om konkursloven er klar på når et selskap er å regne som konkurs, er det stor forskjell fra selskap til selskap på hvordan prosessen utløper seg. Det er en langvarig prosess fra det først åpnes konkurs i et selskap til alle verdier er delt ut og selskapet faktisk er slått konkurs. I løpet av denne prosessen er det mange hendelser som kunne virket som utslagsgivende trigger for dummyvariabelen vår. Dette gjelder for eksempel dato for konkursåpningen, dato for konkursregistrering eller innstilling av bobehandling. I denne oppgaven har vi valgt å definere selskaper som i Enhetsregisteret er kategorisert som "under avvikling eller tvangsoppløsning" eller "konkurs" som konkursselskaper i vårt datasett.

4.2 Uavhengig variabel

De uavhengige variablene studiens modeller består av poster fra både balanse- og resultatregnskapet til de utvalgte selskapene, samt tre nøkkeltall som omfatter lønnsomhet, soliditet og likviditet. Forklaringer og formler for de ulike nøkkeltallene er for interesserte å finne på Proff Forvalt sine nettsider (Proff Forvalt, u.å.) I tillegg er det lagt til uavhengige variabler med selskapsinformasjon i form av antall ansatte og selskapets alder. De uavhengige variablene er hentet ut på bakgrunn av tilgjengelig informasjon fra Proff Forvalt. De uavhengige variablene i ulik grad er med på å forklare om et selskap går konkurs eller ikke. I forbindelse med sammenligningen av de ulike modellene sin evne til å predikere konkurs, vil de tilgjengelige variablene ha ulik forklaringskraft overfor responsvariabelen. Dette gjelder om en sammenligner de uavhengige variablene sin forklaringskraft innad i én modell, samtidig som forklaringskraften til en variabel også vil variere mellom de ulike modellene.

Ettersom denne studien står overfor regnskapsposter og ikke utelukkende nøkkeltall, vil den i større grad være utsatt for multikollinearitet[6] blant de uavhengige variablene. Dette som konsekvens av prinsippet om dobbelt bokholderi, som i større grad vil føre til korrelasjon mellom ulike uavhengige variabler. Det er et mindre problem ettersom modellene

[6] Multikollinearitet innebærer at det foreligger en form for lineær sammenheng mellom ulike forklaringsvariabler innad i en modell.

hovedsakelig anvendes til prediksjon, og ikke til måling av de enkelte uavhengige variablenes effekt på den avhengige variabelen (Berg, 2007). For data vedrørende korrelasjon mellom de uavhengige variablene, se vedlagt korrelasjonsmatrise under appendiks 4. Korrelasjonen befinner seg i området -0,57-0,84, og viser til høy korrelasjon mellom et fåtall variabler. Matrisen viser derimot til en korrelasjon med absoluttverdi <0,20 for de aller fleste parvise sammensetningene av variabler.

De uavhengige variablene er hovedsakelig kontinuerlige variabler. Dette gjelder alle regnskapstall og nøkkeltall. Antall ansatte behandles også som en kontinuerlig variabel. Vi bruker som i den utvidede SEBRA-modellen en årlig dummyvariabel for hvert år siden selskapets stiftelse for år 1 til 8. Hvis selskapet er mer enn 8 år gammelt, settes alderen til 8 år (se tabell 4.1).

Tabell 4.1 Uavhengige variabler.

Variabelnavn	Type variabel	Definisjon av variabel
$Alder_n$	Dummy	$Alder_n$ = åtte dummyvariabler, hvis nei = 0 hvis ja = 1
Ansatte	Kontinuerlig	Antall ansatte i selskapet.
Opptjent egenkapital	Kontinuerlig	Tidligere års overskudd holdes igjen istedenfor å betales ut som utbytte
Innskutt egenkapital	Kontinuerlig	Penger eller eiendeler som eierne selv plasserer i selskap
Kortsiktig gjeld	Kontinuerlig	Lån i form av varekreditt hos leverandør og kassekreditt i bank
Langsiktig gjeld	Kontinuerlig	Gjeld som skal tilbakebetales etter mer en ett år
Omløpsmidler	Kontinuerlig	Kortsiktige eiendeler
Anleggsmidler	Kontinuerlig	Eiendeler bestemt til varig eie eller bruk
Finansinntekter	Kontinuerlig	Inntekt av fordringer og kapitalplasseringer
Finanskostnader	Kontinuerlig	Kostnad forbundet med anskaffelse av fremmedkapital
Driftsresultat	Kontinuerlig	Driftsinntekter fratrukket driftskostnader
Kundefordringer	Kontinuerlig	Virksomhetens pengekrav overfor kunde i forbindelse

		med salg av varer og tjenester
Leverandørgjeld	Kontinuerlig	Gjeld et selskap har overfor leverandører
Kasse/bank/post	Kontinuerlig	Kontanter eller verdier registrert i bank
Lønnskostnader	Kontinuerlig	Utgifter til de ansatte
Pensjonskostnader	Kontinuerlig	Foretakets pensjonsforpliktelse
Aksjekapital	Kontinuerlig	Summen av pålydende på samtlige aksjer
Utbytte	Kontinuerlig	Aksjonærens avkastning på investeringen i selskapet
Likviditetsgrad	Kontinuerlig	Nøkkeltall for andel kortsiktige midler i forhold til forpliktelsene under samme tidshorisont
Soliditet	Kontinuerlig	Nøkkeltall for andelen av eiendelen som er finansiert med egenkapital
Lønnsomhet	Kontinuerlig	Nøkkeltall for avkastningen på eiendelene

4.3 Variabelseleksjon og skalering

Variabelseleksjon er et av flere sentrale elementer innenfor maskinlæring ettersom det i stor grad påvirker den endelige ytelsen til den aktuelle modellen. Om man inkluderer variabler som i liten eller ingen grad forklarer endringer i responsvariabelen, vil dette som forklart under kapittelet om bias og varians potensielt ha en negativ innvirkning på modellens ytelsesevne. Både variabelseleksjon og datavasking (se kapittel 4.4) er konsepter som utføres i forkant av selve modellestimeringen. Dette resulterer i modeller med redusert grad av overfitting, forbedret treffsikkerhet og redusert treningstid som følge av både færre datapunkter og færre observasjoner med unøyaktigheter (Shaikh, 2018). Det er flere former for variabelseleksjon, hvor enkelte algoritmer som eksempelvis beslutningstrær har iboende automatisk variabelseleksjon, mens andre algoritmer krever manuell gjennomføring av prosessen.

I likhet som ved tuning av hyperparametere, foreligger det ingen form for universell regel for inkludering av antall variabler. Det optimale antallet variabler for et spesifikt problem, er noe som må avdekkes gjennom systematisk eksperimentering overfor de ulike modellene med det aktuelle datagrunnlaget som foreligger. En mye brukt metode for variabelseleksjon for å imøtekomme denne problemstillingen er algoritmen “recursive feature elimination” (RFE)

(Brownlee, 2019). Der hvor filter-baserte seleksjonsmetoder selv foretar en rangering av variablenes viktighet og filtrerer deretter, bruker RFE den aktuelle maskinlæringsalgoritmen i seg selv som kjerne i metoden. Som følge av dette er RFE kategorisert som en "wrapper"-metode innenfor variabelseleksjon. Dette innebærer at den aktuelle maskinlæringsalgoritmen trenes på ulike sammensetninger av foreliggende variabler, hvor den beste sammensetningen med hensyn til prediktive evner velges ut. De ulike sammensetningene kommer til ved å fjerne de minst viktige variablene, før modellene trenes og evalueres på nytt med det nye utvalget av variabler. Rangeringen av variablenes viktighet kan foretas av maskinlæringsalgoritmen i seg selv der hvor dette lar seg gjøre, som eksempelvis for tidligere nevnte besluntningstrær eller random forest. For øvrige maskinlæringsalgoritmer, som ikke fremlegger viktigheten av variabler eller leverer lite nøyaktige estimater, brukes gjerne statistiske metoder som eksempelvis logistisk regresjon til å identifisere viktigheten av de ulike variablene.

Filter-baserte seleksjonsmetoder er en annen form for veiledet variabelseleksjon. Her kan man blant annet kan velge ut de *k* antall viktigste, eller *k* viktigste percentil, av variablene i datagrunnlaget. Med variabelens viktighet henvises det her til variabelens grad av innvirkning på responsvariabelen. Dette evalueres av statistiske metoder som velges ut på bakgrunn av om datasettet består av kategoriske eller numeriske variabler. Liang et al. (2015) undersøkte i deres studie hvilken effekt valg av variabelseleksjonsmetode hadde innenfor konkursprediksjon. Studien trekker frem at wrapper-metoder og filter-baserte metoder er like på mange områder, men at førstnevnte tar for seg og evaluerer flere kombinasjoner av variabelsammensetninger og i så måte som oftest vil prestere bedre enn filter-baserte metoder. Ettersom det blir evaluert flere sammensetninger av variabler innenfor wrapper-metodene, vil dette også stille større krav til computeren ved kjøring sammenlignet med de filter-baserte metodene (Liang et al., 2015). På bakgrunn av disse betraktningene har vi valgt å anvende wrapper-metoden RFE rettere enn en filter-basert metode når det kommer til variabelseleksjon.

Datagrunnlaget består av variabler med tilhørende verdier som varierer i omfang. Dette medfører et problem når dataen identifiseres av maskinlæringsmodellene, ettersom mange av disse vil foreta en underliggende antakelse om at de høyere tallene vil ha større betydning enn de lavere. Algoritmene neglisjerer måleenhetene de står overfor, og ser kun på selve tallet til en variabel når den kalkulerer variasjonen mellom datapunkter. Variabler av stort omfang vil kunne variere med flere tusen enheter for hver observasjon, mens variabler av lavere omfang

vil kunne variere med kun noen få enheter blant observasjonene. For å imøtekomme problematikken forbundet med dette, må variablene bringes ned på samme nivå eller omfang, noe som kan gjøres gjennom skalering.

Trebaserte modeller som random forest og beslutningstrær skiller seg blant annet ut fra øvrige inkluderte maskinlæringsmodeller gjennom at de ikke baserer sine klassifikasjoner på avstandsberegninger. Dette medfører at de heller ikke er sensitive overfor variabler av ulikt omfang. Modeller som støttevektormaskin, k-nærmeste nabo og nevrale nettverk derimot er å anse som sensitive overfor variablenes relative skala, og vil følgelig være helt avhengige av at dataen er skalert. En mye anvendt skaleringsmetode overfor disse maskinlæringsmodellene er standardisering (Saini, 2019). Dette innebærer at variablene standardiseres med en forventningsverdi lik null og varians lik én. Hensikten er å endre omfanget eller skalaen for hver variabel, men ikke selve distribusjonen. Med dette som bakgrunn, har vi i denne studien skalert samtlige variabler etter dette prinsippet.

4.4 Ekskluderte observasjoner

For å bedre modellenes endelige evne til å predikere har vi utført datavasking eller rensing av datasettet. Dette innebærer å avdekke unøyaktigheter eller feil i registreringen av data. Feilene og unøyaktighetene i dataen erstattes, modifiseres eller fjernes. Datasettet vårt utgjorde i utgangspunktet totalt 49.399 selskaper og omtrent 45 prosent av disse har mangler. Observasjoner med mangelfulle registreringer har vi valgt å fjerne istedenfor å korrigere dem.

Det første kriteriet vi har satt for å kontrollere datakvaliteten er at observasjonene må bestå en test i forbindelse med balanseregnskapet. Et selskap sin balanse virker som en oversikt over den økonomiske situasjonen på et gitt tidspunkt og består av aktiva på den ene siden og passiva på den andre siden. Her refererer aktiva til et selskap sine eiendeler som igjen deles inn i anleggsmidler og omløpsmidler. Passiva på den annen side refererer til selskapets egenkapital og gjeld. Aktiva og passiva hos et selskap skal alltid være balansert. Noe som innebærer at de er like store som gir en sum lik null. Dette er ikke alltid tilfelle i praksis da avrundings eller føringsfeil kan oppstå. Med dette som bakgrunn har vi satt et kriterium hvor aktiva og passiva ikke kan avvike med mer enn ti prosent. Dette er det samme kriteriet som brukt av Meese og Viken (2019). Om en observasjon har større avvik enn dette i balansen anser vi observasjonen for å inneha store feil, og vil bli fjernet.

Videre har vi brukt det samme kriteriet satt i arbeidet til Bernhardsen og Larsen (2007), og valgt å fjerne selskaper hvor regnskapsposten totale eiendeler er mindre enn 500.000 NOK. Avslutningsvis har datasettet også blitt renset for dupliserte observasjoner og observasjoner med manglende verdier. Etter at disse ekskluderingene ble foretatt gjenstod totalt 27.314 selskaper, hvorav 26.601 av disse er ikke-konkursselskaper og resterende 713 er konkursselskaper. Dette gjelder for hele perioden 2017-2019, som utgjør vårt endelige datagrunnlag etter rensing av data.

5.0 Resultater

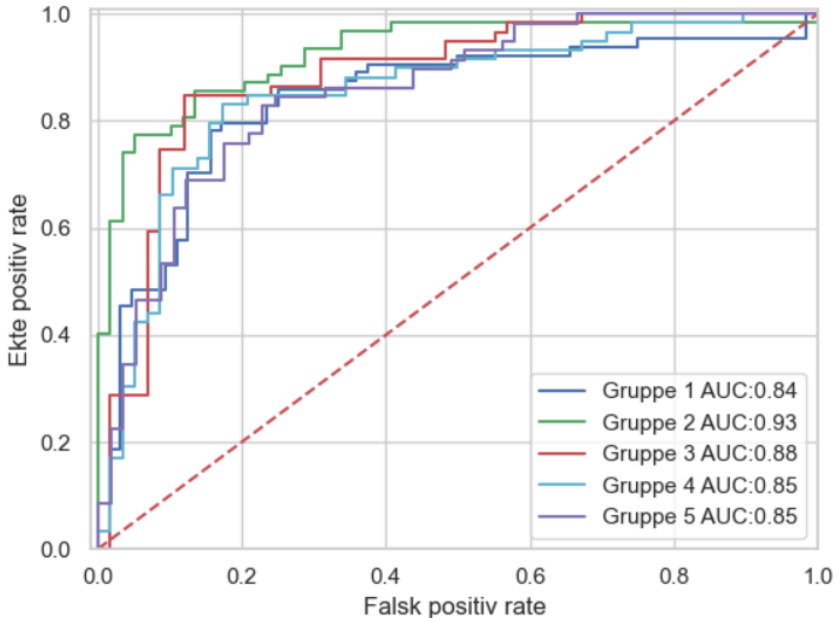
Dette kapittelet er delt inn i to hoveddeler. Avsnitt 5.1 viser resultatene knyttet til vår hovedproblemstilling. Her presenterer vi resultatene fra hver enkelt modell og sammenligner disse. Avsnitt 5.2 er knyttet til våre øvrige problemstillinger og sammenligninger de forskjellige uavhengige variablenes forklaringskraft i normalår og pandemiår.

5.1 Modellprestasjon

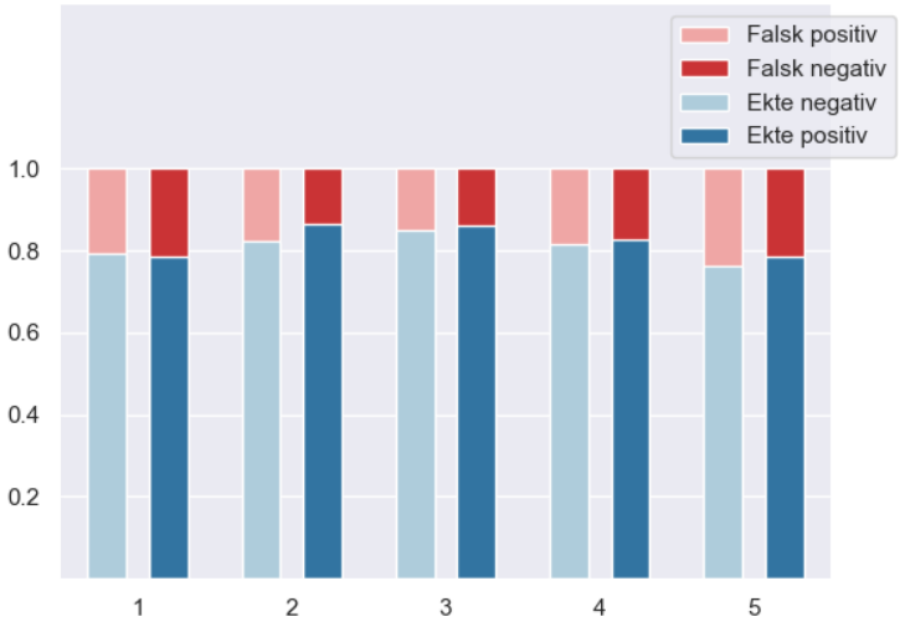
For hver av modellene våre presenterer vi ROC-kurver med tilhørende AUC-verdi og treffandeler for hver av de fem gruppene som ble beskrevet i avsnittet om kryssvalideringen. Sammenligningen av modellene er basert på AUC-verdi og andel riktige klassifikasjoner som beskrevet i avsnitt 3.2. Samtlige maskinlæringsmodeller har blitt implementert i Microsoft Visual Studios Community ved bruk av Python som programmeringsspråk. Her har vi tatt i bruk flere av Scikit-learn sine utformede klassifikasjonsmodeller (Pedregosa et al., 2011), hvor Scikit-learn er et bibliotek for programvare innenfor maskinlæring med Python som programmeringsspråk. Resultatene fra hver av de respektive modellene vil presenteres i kommende underkapitler.

5.1.1 Utvidet SEBRA-modell*

Den utvidede SEBRA-modellen* er satt opp i tråd med beskrivelsen under kapittel 3.1.1. Figur 5.1 viser ROC-kurvene med tilhørende AUC-verdi for denne modellen. Figur 5.2 viser modellens treffandeler, og er satt opp med parvise søyler. I figuren viser den venstre søylen prediksjonene for ikke-konkursselskaper. I hver av disse søylene viser lyseblå farge andelen riktige prediksjoner og lyserød farge andelen gale prediksjoner (type-I-feil). Den høyre søylen viser tilsvarende for konkursselskaper. For disse selskapene viser mørkeblå farge riktige prediksjoner og mørkerød farge viser gale prediksjoner (type-II-feil).



Figur 5.1. ROC-kurver fra utvidet SEBRA*.



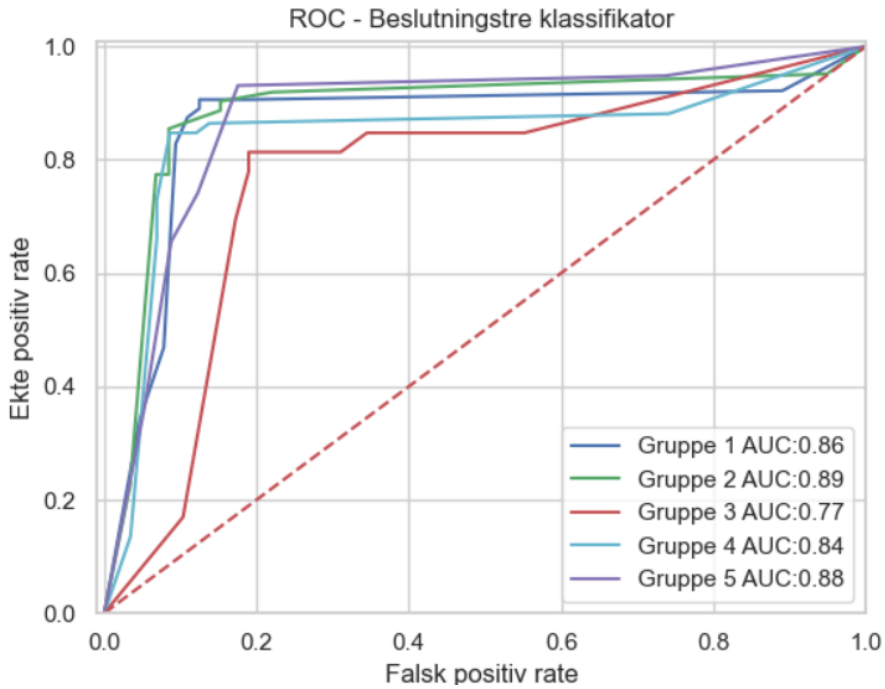
Figur 5.2. Treffandel innad i grupper fra kryssvalidering for den utvidede SEBRA-modellen*.

Vi kan se at den utvidede SEBRA-modellen* oppnår AUC-verdier i intervallet 0,84-0,93 (illustrert nederst til høyre i figur 5.1), hvor nedre og øvre verdi er resultater fra henholdsvis gruppe 1 og 2. At gruppe 2 leverer de beste resultatene kan leses av ved å se på den grønne kurven i figuren, som strekker seg lenger opp mot venstre hjørne sammenlignet med resterende kurver. Med dette oppnår den også det største arealet under kurven som samsvarer med et bedre resultat. I motsatt tilfelle, er gruppe 1 med det svakeste resultatet den kurven

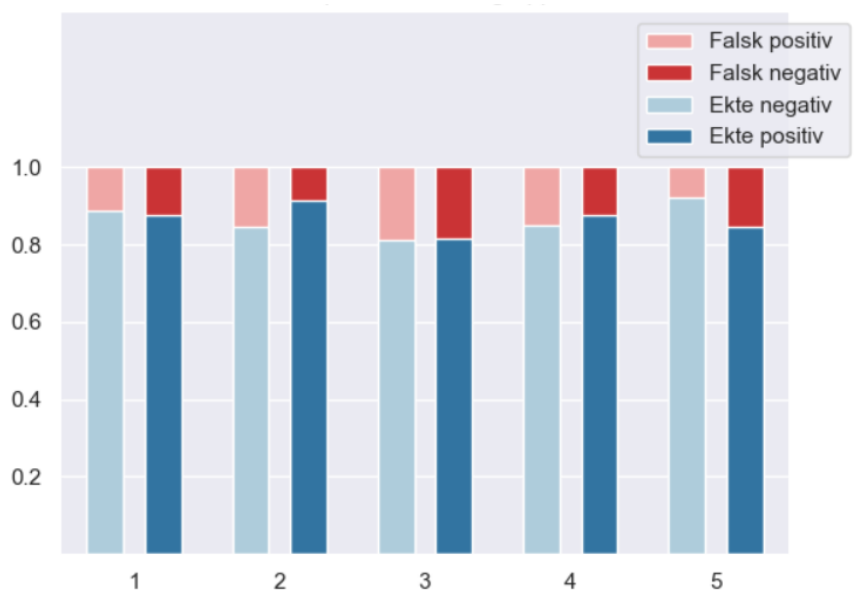
som strekker seg minst opp mot venstre hjørne. Prinsippene om avlesning fra figurene i denne seksjonen gjør seg gjeldende for alle modellene, da samtlige resultater legges frem i samme format. Videre er gjennomsnittlig AUC-verdi fra kryssvalideringens fem grupper lik 0,87. Dette gjennomsnittet er basert på resultatene fra nederst i høyre hjørne i samme figur. Det fremkommer av resultatene at det er forholdsvis stor spredning mellom gruppene som leverer dårligst og best resultater, hvilket kan tyde på at modellen ikke er spesielt robust. Dette innebærer at resultatene vil variere i forholdsvis stor grad avhengig av utvalgte observasjoner. Når det kommer til andelen riktige klassifikasjoner, leverer modellen resultater i intervallet 0,77 til 0,85, hvor øvre og nedre treffprosent er oppnådd av henholdsvis gruppe 5 og 3. Disse verdiene kan leses ut ifra figur 5.2, ved å se til søylene og lese av oppnådd treffandel mot venstre akse. Den gjennomsnittlige andelen riktige klassifikasjoner blant de fem gruppene er lik 0,81.

5.1.2 Beslutningstre

Beslutningstre har blitt satt opp gjennom anvendelse av Scikit-learn sin beslutningstre-klassifikator (u.å.). Optimaliseringen av modellens hyperparametere har som nevnt blitt gjort gjennom anvendelse av grid search. Se appendiks 5 for fremgangsmåte og endelige valgte verdier for de ulike hyperparameterne. Videre har vi ekskludert variabler i tråd med beskrivelsen under kapittel 4.3 gjennom bruk av RFE hvor den endelige modellen oppnådde en forbedring på 0,07 i gjennomsnittlig andel riktige klassifikasjoner, samt en forbedret gjennomsnittlig AUC-verdi på 0,05. Dette gjennom å benytte seg av de seks viktigste forklaringsvariablene fra datagrunnlaget. Se tabell 5.2 for endelig utvalgte variabler. Figur 5.3 viser ROC-kurvene med tilhørende AUC-verdi og figur 5.4 viser treffandelene for de fem ulike gruppene fra kryssvalideringen.



Figur 5.3. ROC-kurver fra beslutningstre.



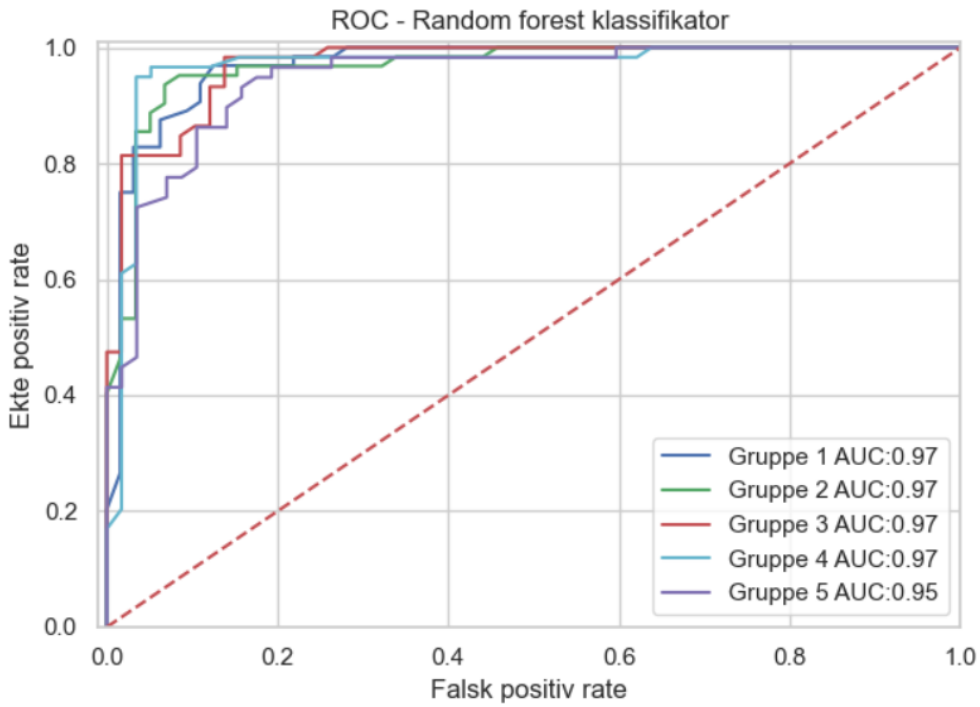
Figur 5.4. Treffandel innad i grupper fra kryssvalidering for beslutningstre.

Fra figur 5.3 fremkommer det at beslutningstre som klassifikasjonsmodell oppnådde AUC-verdier i intervallet 0,77-0,89, hvor nedre og øvre verdi i intervallet ble oppnådd av henholdsvis gruppe 3 og 2. Videre er gjennomsnittlig AUC-verdi fra de fem gruppene rett i underkant av 0,85. I likhet med SEBRA-modellen, så viser det seg at det er relativt stor spredning blant AUC-verdiene i denne modellen, som igjen viser til en lite robust modell. Når

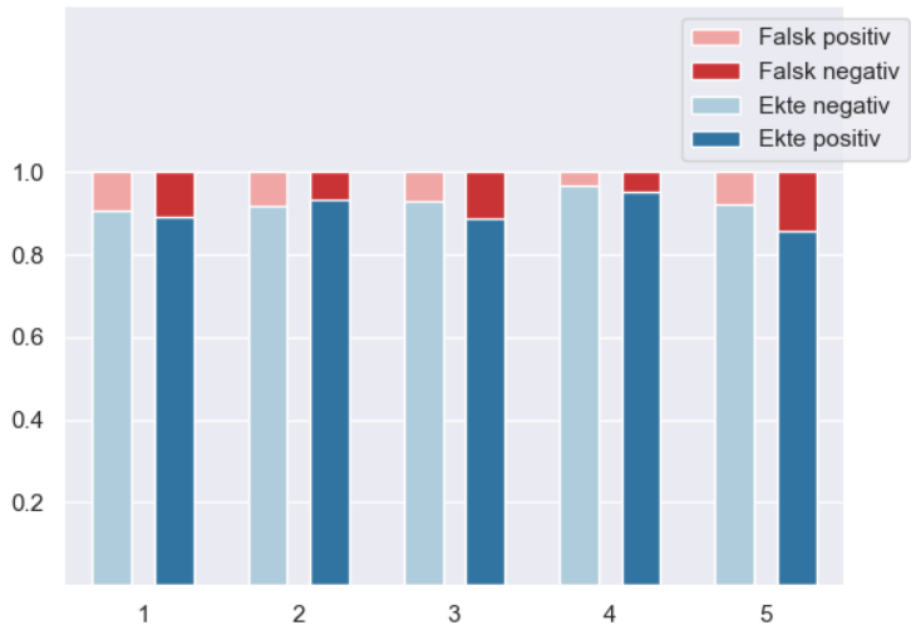
det kommer til treffprosent, kan man fra figur 5.4 se en fordeling i andeler riktige klassifikasjoner i intervallet 0,81-0,88, hvor nedre og øvre verdi igjen er resultatet fra henholdsvis gruppe 3 og 2. Videre er den gjennomsnittlige andelen treff blant de fem gruppene 0,86.

5.1.3 Random forest

Random forest har blitt satt opp gjennom anvendelse av Scikit-learn sin random forest klassifikator (u.å). For detaljert fremgangsmåte og endelige valgte verdier for modellens hyperparametere, se appendiks 5. Gjennom bruk av RFE oppnådde modellen en forbedret gjennomsnittlig andel riktige klassifikasjoner på 0,02, samt en forbedret gjennomsnittlig AUC-verdi på 0,01. Dette gjennom å ekskludere de fire minst viktige variablene fra datagrunnlaget i den endelige modellen. Se tabell 5.2 for endelig utvalgte variabler. Figur 5.5 viser ROC-kurvene med tilhørende AUC-verdi og figur 5.6 viser treffandelene for de fem ulike gruppene fra kryssvalideringen.



Figur 5.5. ROC-kurver fra random forest.

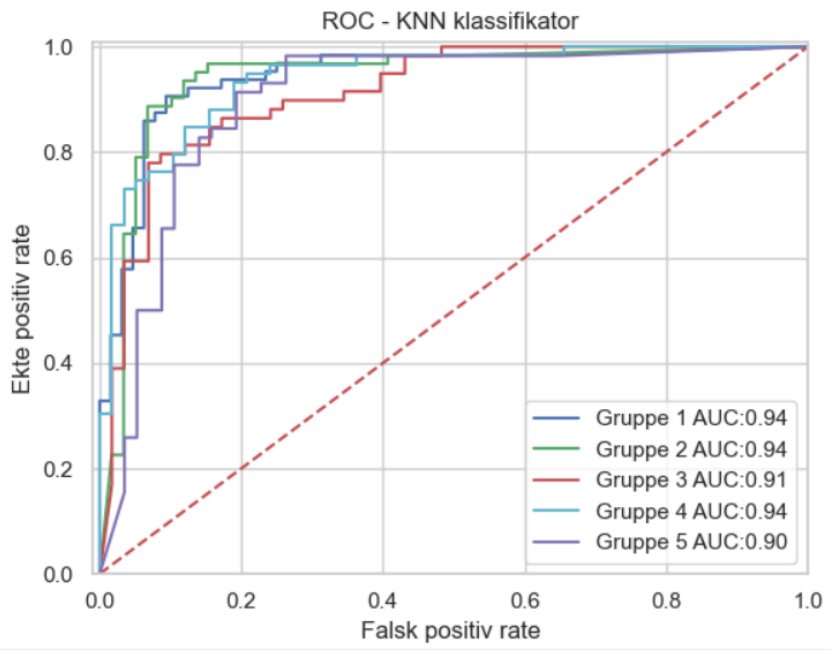


Figur 5.6. Treffandel innad i grupper fra kryssvalidering for random forest.

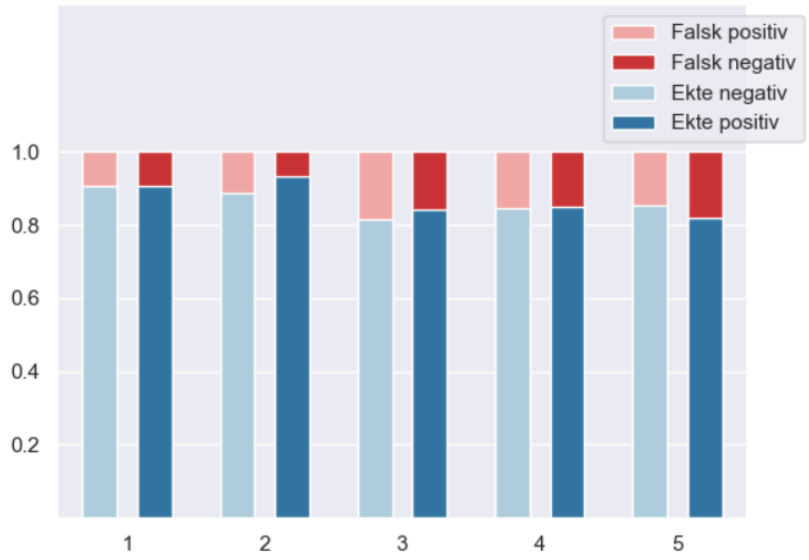
Som det fremkommer fra figur 5.5 og 5.6, oppnår random forest svært gode resultater både med tanke på AUC-verdier og treffprosent. Modellen leverer AUC-verdier i intervallet 0,95-0,97, med en gjennomsnittlig verdi rett i underkant av 0,97. Spredningen mellom AUC-verdiene i random forest er langt mindre enn de to foregående modellene, noe som kan tyde på at dette er en langt mer robust modell. Som illustrert fra figur 5.6, oppnår modellen andeler riktige klassifikasjoner i intervallet 0,89-0,96 med et gjennomsnitt på 0,92. Her er nedre og øvre grense av intervallet resultatene fra henholdsvis gruppe 5 og 4. Med dette som foreliggende resultater, er random forest den modellen som leverer best resultater både med hensyn til AUC-verdier og treffandel i denne oppgaven.

5.1.4 K-nærmeste nabo

K-nærmeste nabo har blitt satt opp gjennom anvendelse av Scikit-learn sin k-naboer klassifikator (u.å.). For endelige valgte verdier for de ulike hyperparameterne, se appendiks 5. Videre oppnådde modellen gjennom anvendelse av RFE en forbedret gjennomsnittlig andel riktige klassifikasjoner på 0,02 samt en forbedret gjennomsnittlig AUC-verdi på 0,01. Dette gjennom å ta i bruk de fem viktigste forklaringsvariablene fra datagrunnlaget. Se tabell 5.2 for endelig utvalgte variabler. Figur 5.7 viser ROC-kurvene med tilhørende AUC-verdi og figur 5.8 viser treffandelene for de fem ulike gruppene fra kryssvalideringen.



Figur 5.7. ROC-kurver fra k-nærmeste nabo.

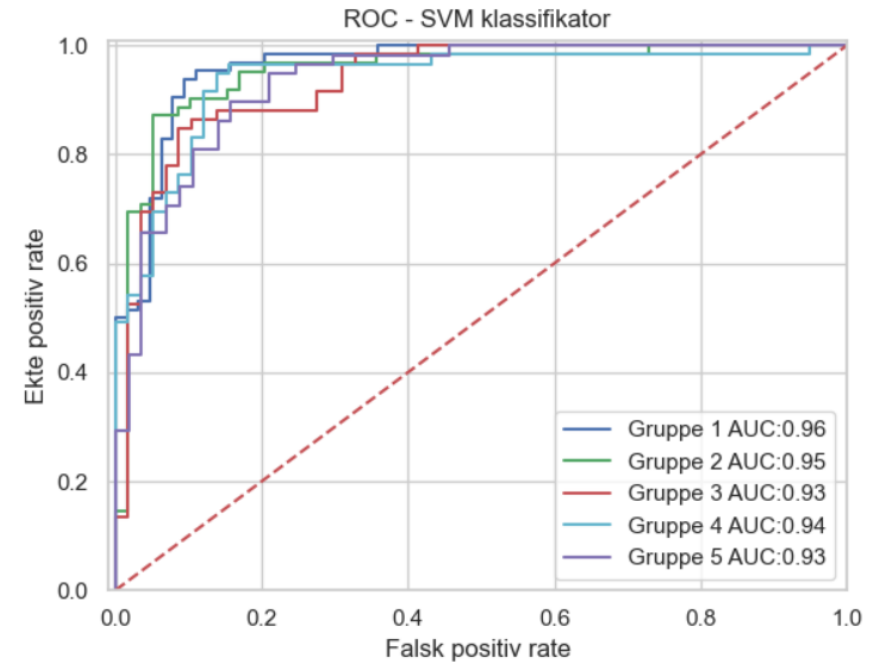


Figur 5.8. Treffandel innad i grupper fra kryssvalidering for k-nærmeste nabo.

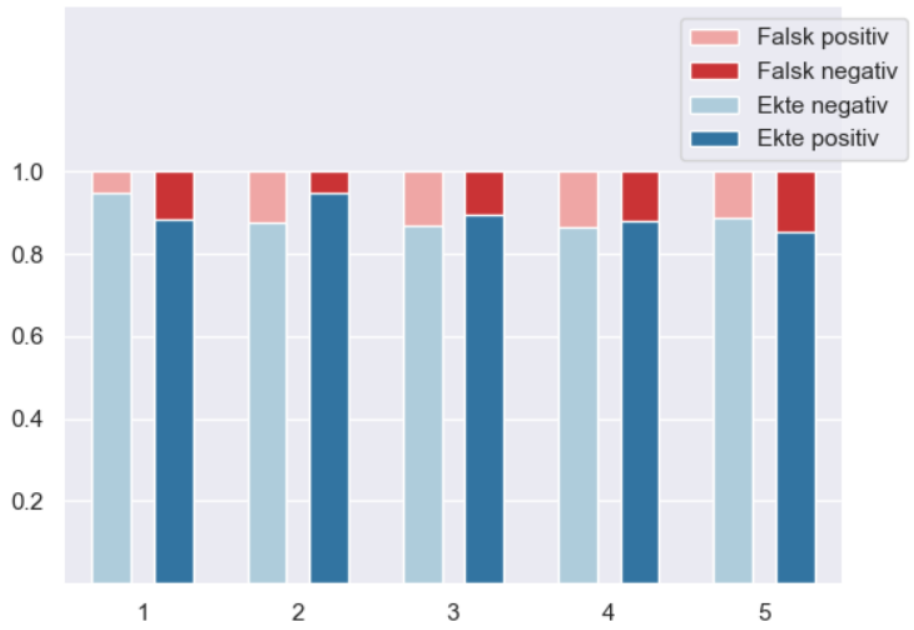
Fra figur 5.7 fremkommer det at k-nærmeste nabo som klassifikasjonsmodell oppnådde AUC-verdier i intervallet 0,90-0,94. Videre leverte de fem gruppene fra kryssvalideringen en gjennomsnittlig AUC-verdi i underkant av 0,93. Som illustrert fra figur 5.8, oppnår modellen en andel riktige klassifikasjoner i intervallet mellom 0,83-0,91, hvor nedre og øvre treffprosent i intervallet ble oppnådd av henholdsvis gruppe 3 og gruppe 2. Den gjennomsnittlige andelen riktige klassifikasjoner blant de fem gruppene er 0,87.

5.1.5 Støttevektormaskin

Støttevektormaskin har blitt satt opp gjennom anvendelse av Scikit-learn sin ikke-lineære støttevektormaskin for binære klassifikasjoner (u.å.). For informasjon om valgte verdier for de ulike hyperparameterne, se appendiks 5. Videre oppnådde den endelige modellen en forbedret gjennomsnittlig andel riktige klassifikasjoner på 0,02, samt en marginalt forbedret (<0,01) gjennomsnittlig AUC-verdi gjennom anvendelse av RFE. Dette ved å ta i bruk de fem viktigste forklaringsvariablene fra datagrunnlaget. Se tabell 5.2 for endelig utvalgte variabler. Figur 5.9 viser ROC-kurvene med tilhørende AUC-verdi og figur 5.10 viser treffandelene for de fem ulike gruppene fra kryssvalideringen.



Figur 5.9. ROC-kurver fra støttevektormaskin.

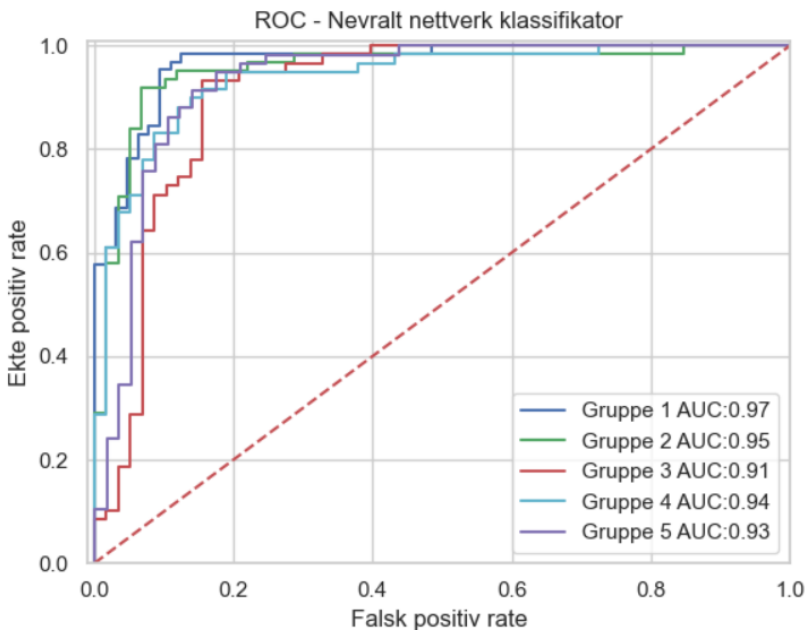


Figur 5.10. Treffandel innad i grupper fra kryssvalidering for støttevektormaskin.

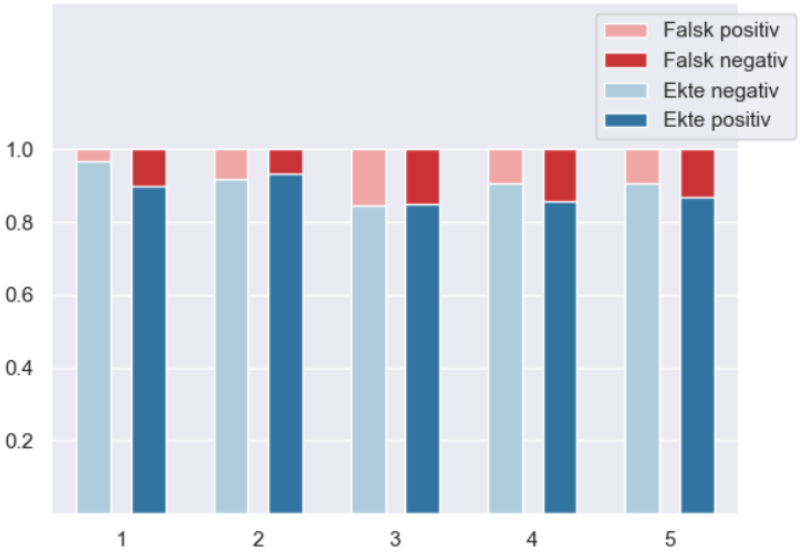
Som figur 5.9 illustrerer oppnår støttevektormaskin AUC-verdier i intervallet 0,93-0,96, hvor nedre og øvre verdi oppnås av henholdsvis gruppe 5 og 1. Gjennomsnittlig AUC-verdi blant de fem gruppene er rett i overkant av 0,94. Fra figur 5.10 kan vi tyde oppnådde andeler riktige klassifikasjoner blant gruppene i intervallet 0,87-0,91. Videre er den gjennomsnittlige andelen riktige klassifikasjoner for gruppene 0,89. Også her kommer det svakeste resultatet fra gruppe 5, samtidig som gruppe 1 står for det beste resultatet.

5.1.6 Nevrale nettverk

Nevrale nettverk har blitt satt opp gjennom anvendelse av Scikit-learn sin MLP ("multi layer perceptron") klassifikator (u.å.). I appendiks 5 er det lagt ved informasjon om endelige valgte verdier for de ulike hyperparameterne. Videre oppnådde modellen gjennom bruk av RFE en forbedret gjennomsnittlig andel riktige klassifikasjoner på 0,03, samt en forbedret gjennomsnittlig AUC-verdi på 0,01, ved å inkludere de fem viktigste forklaringsvariablene fra datagrunnlaget. Se tabell 5.2 for endelig utvalgte variabler. Figur 5.11 viser ROC-kurvene med tilhørende AUC-verdi og figur 5.12 viser treffandelene for de fem ulike gruppene fra kryssvalideringen.



Figur 5.11. ROC-kurver fra nevralt nettverk.

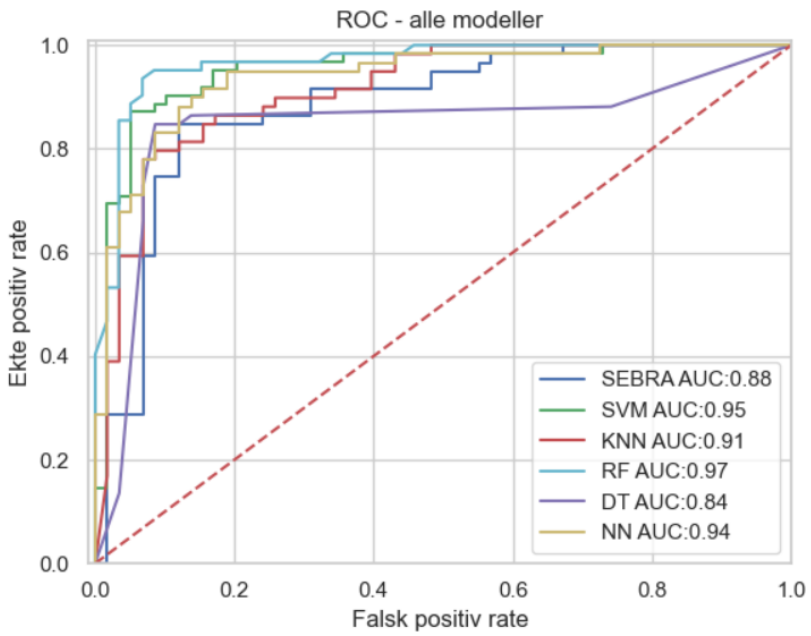


Figur 5.12. Treffandel innad i grupper fra kryssvalidering for nevralt nettverk.

Fra figur 5.11 observerer vi oppnådde AUC-verdier i intervallet 0,91-0,97. Nedre og øvre verdi av intervallet er oppnådd av henholdsvis gruppe 3 og 1. Videre leverte de fem gruppene fra kryssvalideringen en gjennomsnittlig AUC-verdi lik 0,94. Illustrert gjennom figur 5.12, så oppnådde nevrale nettverk andeler riktige klassifikasjoner i intervallet 0,85-0,93, hvor de samme gruppene som nevnt ovenfor står for nedre og øvre grense. Den gjennomsnittlige andelen riktige klassifikasjoner blant gruppene er 0,90.

5.1.7 Sammenligning av modellene

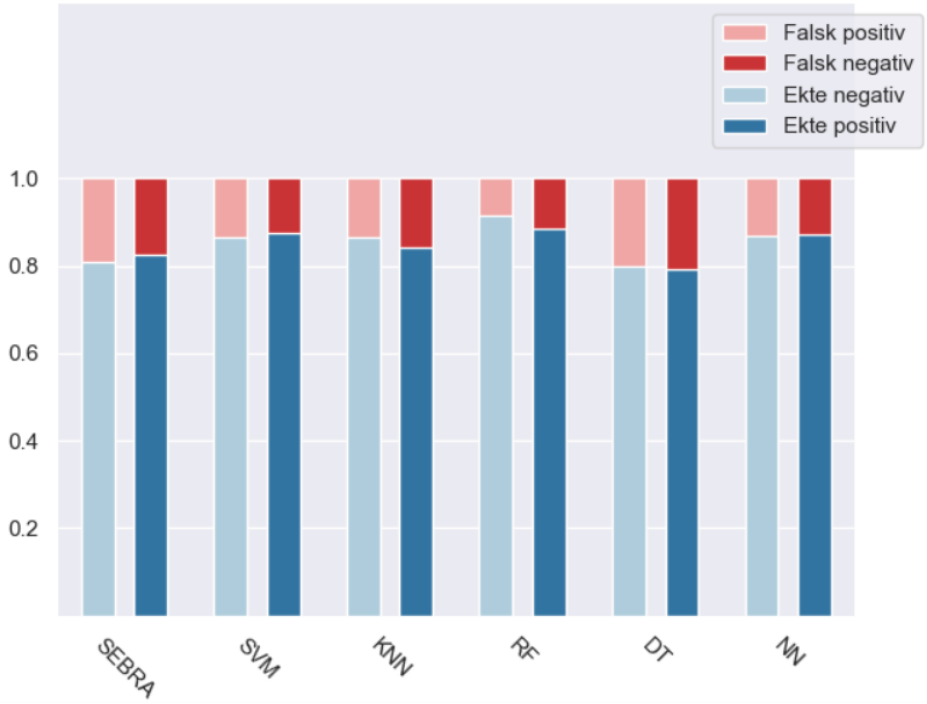
Vi har frem til nå presentert AUC-verdier og treffandeler for modellene enkeltvis, og vil videre sammenligne resultatene fra de ulike modellene. Figur 5.13 nedenfor viser ROC-kurver tatt ut fra hver enkelt modell med sammenligning som hensikt. For å belønne robuste modeller som presterer godt blant alle gruppene i kryssvalideringen, har vi fra hver av modellene valgt ut den kurven som presterer nærmest den oppnådde gjennomsnittlige AUC-verdien til den aktuelle modellen. Som det fremkommer av figuren, ser vi at random forest er den modellen som presterer best med foreliggende data og metode. Dette kan vi lese av figuren ved at den lyseblå kurven er den som trekkes lengst opp mot venstre hjørne i diagrammet, og dermed oppnår størst areal under kurven. Som fremlagt tidligere leverer denne modellen en gjennomsnittlig AUC-verdi blant de fem gruppene i kryssvalideringen på rett i underkant av 0,97. Videre er random forest tett etterfulgt av henholdsvis støttevektormaskin og nevrale nettverk. Beslutningstre leverer det svakeste resultatet blant maskinlæringsmodellene. Den leverer dessuten lavere AUC-verdier enn den utvidede SEBRA-modellen* og leverer dermed det svakeste resultatet av samtlige inkluderte modeller.



Figur 5.13. ROC-kurver fra alle modeller.

Når det kommer til oppnådde resultater med riktig andel klassifikasjoner som evalueringskriterium er det nok en gang random forest som kommer best ut med en gjennomsnittlig treffandel på 0,92, se figur 5.14. Dette indikerer at 92 prosent av variasjonen i konkurser/ikke-konkurser er forklart gjennom utvalgte uavhengige variabler, med random forest som modell og klassifiseringsterskel på 0,5. Modellen predikerer noe bedre overfor

ikke-konkursselskaper med en ekte negativ rate på 0,93, sammenlignet konkursselskaper med en ekte positiv rate på 0,90. Dette er å anse som et veldig godt resultat overfor jevnfordelte utvalg når det kommer til konkurs og ikke-konkursselskaper, sammenlignet med resultatene lagt frem i metaanalysen til Alaka et al. (2017). Nok en gang er random forest etterfulgt av støttevektormaskin og nevralt nettverk, hvor det denne gangen er sistnevnte som presterer nest best blant samtlige modeller med en gjennomsnittlig treffandel på 0,90. Nok en gang er det beslutningstre som kommer svakest ut blant maskinlæringsmodellene, med en gjennomsnittlig andel riktige klassifikasjoner på 0,86. Denne gangen leverer modellen derimot bedre resultater enn den utvidede SEBRA-modellen* som leverer det svakeste resultatet med en andel riktige klassifikasjoner på 0,81.



Figur 5.14. Gjennomsnittlig treffandel fra samtlige modeller.

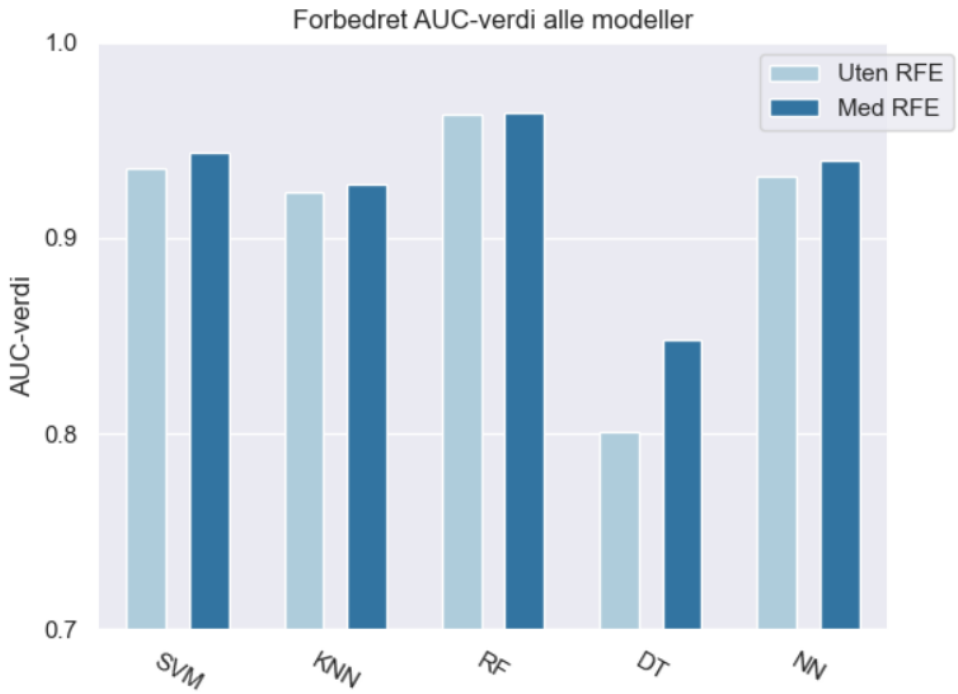
Videre illustrerer tabell 5.1 tidligere fremlagte resultater for både treffandel og AUC-verdi. Vi har også lagt frem ekte negativ rate og ekte positiv rate som belyst tidligere utgjør aksene i et ROC-diagram. Tabellen viser dessuten til innvirkningen av bruken av RFE som filtreringsmetode når det kommer til variabelseleksjon. Tallene i parentes viser til resultatene før filtreringen av forklaringsvariabler. Tallene uten parentes viser resultatene etter filtreringen, og det er disse resultatene vi har henvist til i presentasjonen i avsnittene 5.1.1-

5.1.6. Vi understreker her at variabelseleksjonen har foregått med optimalisering av AUC-verdi som hovedformål. Den utvidede SEBRA-modellen* har ikke gjennomgått noen form for variabelseleksjon siden variablene i denne modellen er gitt.

Tabell 5.1. Modellprestasjoner før (i parentes) og etter bruk av RFE.

Modell	Ekte negativ rate	Ekte positiv rate	Treffandel	AUC
Utvidet SEBRA*	(0,81)	(0,82)	(0,81)	(0,87)
Beslutningstre	(0,80) – 0,86	(0,79) – 0,86	(0,79) - 0,86	(0,80) – 0,85
Random forest	(0,92) – 0,93	(0,89) – 0,90	(0,90) - 0,92	(0,96) - 0,97
K-nærmeste nabo	(0,87) - 0,86	(0,84) – 0,87	(0,85) - 0,87	(0,92) – 0,93
Støttevektor-maskin	(0,87) - 0,89	(0,88) - 0,89	(0,87) - 0,89	(0,94) - 0,94
Nevrale nettverk	(0,87) - 0,91	(0,87) - 0,88	(0,87) - 0,90	(0,93) - 0,94

Som vi kan se fra tabellen leverer random forest høyest andel riktige klassifikasjoner og AUC-verdi også før filtrering. Dette innebærer at når modellene trenes og testes med de samme variablene tilgjengelig, leverer fortsatt random forest best resultater. Det er også verdt å merke seg at samtlige modeller har oppnådd forbedrede prediktive evner etter filtrering av et ulikt antall forklaringsvariabler når det kommer til treffandelen. Dette gjelder også for AUC-verdier, noe som illustreres gjennom figur 5.15. Venstre søyle i hver av de parvise sammensetningene henviser til gjennomsnittlig AUC-verdi før filtrering med RFE, og høyre søyle viser til den gjennomsnittlige AUC-verdien filtrering med RFE. Her kan vi se at det er snakk om marginale (<0,01) forbedringer i gjennomsnittlig AUC-verdi for enkelte modeller, mens eksempelvis beslutningstre oppnår en større forbedring på 0,05.



Figur 5.15. Gjennomsnittlig AUC-verdi uten og med gjennomføring av RFE.

5.2 Variablenes forklaringskraft

I dette underkapittelet vil vi analysere og sammenligne forklaringskraften til de forskjellige uavhengige variablene for de ulike modellene.

5.2.1 Normalår

Tabell 5.2 nedenfor viser et oversiktsbilde over hvilke variabler som har blitt inkludert for de endelige modellene etter gjennomføring av RFE. De utvalgte variablene er markert med punkter. Utvalget baseres på hvilke forklaringsvariabler som har størst forklaringskraft overfor responsvariabelen, slik gjort rede for i kapittel 4.3. Tabellen illustrerer at det inkluderes et ulikt antall uavhengige variabler blant de forskjellige modellene, hvor dette utvalget er basert på maksimering av modellenes gjennomsnittlige AUC-verdi. Rangeringen av variablenes forklaringskraft er for beslutningstre (BT) og random forest (RF), gjort av deres innebygde rangeringsverktøy. Videre har nevralt nettverk (NN), k-nærmeste nabo (KNN) og støttevektormaskin (SVM) i mangel av eget innebygd rangeringsverktøy tatt i bruk logistisk regresjon for å identifisere variablenes forklaringskraft. Den utvidede SEBRA-modellen* inngår ikke i tabellen, da den ikke gjennomgår RFE ettersom variablene i modellen allerede er bestemt.

Tabell 5.2. Gjenværende variabler for hver av modellene etter gjennomført RFE. Generell selskapsinformasjon i kursiv, foretaksøkonomiske nøkkeltall markert med stjerne* og øvrige variabler er regnskapstall.

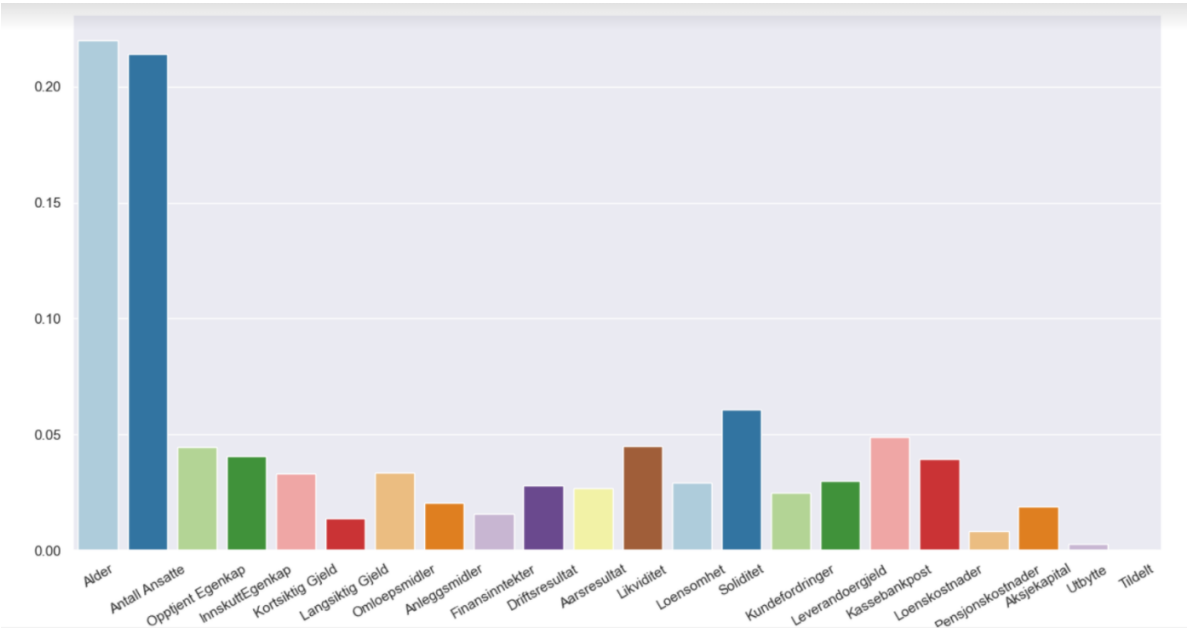
Modell -> **Variabel**	BT	RF	KNN	SVM	NN
Alder	•	•	•	•	•
Antall ansatte	•	•	•	•	•
Soliditet*	•	•	•	•	•
Kasse/bank/post		•	•	•	•
Lønnskostnader		•	•	•	•
Årsresultat	•	•			
Leverandørgjeld	•	•			
Omløpsmidler	•	•			
Opptjent egenkapital		•			
Innskutt egenkapital		•			
Kortsiktig gjeld		•			
Anleggsmidler		•			
Driftsresultat		•			
Kundefordringer		•			
Aksjekapital		•			
Likviditet*		•			
Lønnsomhet*		•			
Langsiktig gjeld					
Finansinntekter					
Pensjonskostnader					
Utbytte					

Det fremkommer av tabellen at det er tre variabler som inngår i samtlige modeller etter gjennomføring av RFE. Dette gjelder alder, antall ansatte og soliditet, hvor de to førstnevnte er generell selskapsinformasjon og sistnevnte er et foretaksøkonomisk nøkkeltall. Videre er regnskapstallene for kasse/bank/post (kontanter eller verdier registrert i bank) og lønnskostnader inkludert av alle modellene utenom beslutningstre. Nøkkeltall for lønnsomhet og likviditet er på sin side kun inkludert i random forest. De fire nederste variablene i tabellen er ikke inkludert i noen av modellene. Tabellen viser også at modellene som brukte logistisk regresjon for å identifisere variablenes viktighet, tok i bruk de samme uavhengige variablene. Alle disse modellene tok i bruk selskapets alder, antall ansatte, kasse/bank/post, lønnskostnader og soliditet.

5.2.2 Pandemiår

Som tidligere nevnt inneholder vår beste prediksjonsmodell (random forest) et innebygd rangeringsverktøy overfor uavhengige variabler sin forklaringskraft opp mot den avhengige variabelen. Dette innebygde verktøyet går under navnet “random forest feature importance”, hvor vi har i denne oppgaven har benyttet oss av Scikit-learn sin versjon (Scikit-learn, u.å.).

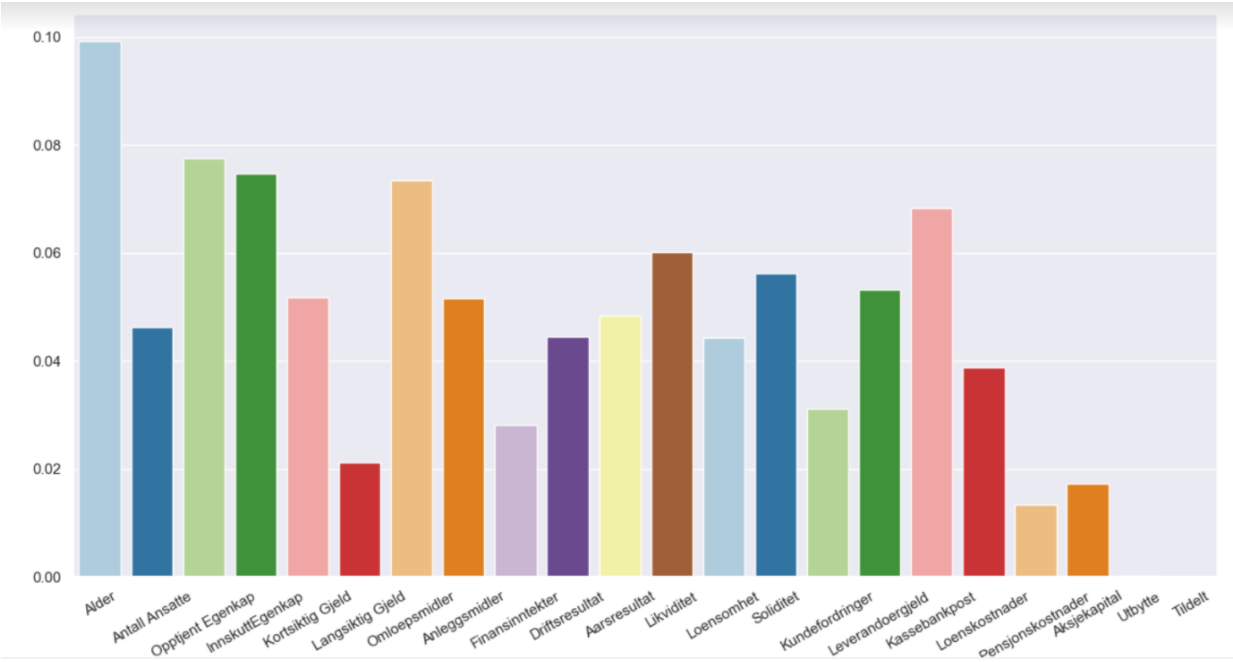
Denne rangeringen er sensitiv overfor variabler med mange unike verdier, noe vi unngår gjennom bruk av standardiserte variabler. Med dette som bakgrunn ønsker vi å utføre en sammenligning av variablenes forklaringskraft i det vi i denne oppgaven omtaler som "normalår" med et "pandemiår". Dette gjøres ved å sammenligne variablenes viktighet når modellen er estimert på regnskapsdata fra årene 2017-2018, mot regnskapsdata fra 2019. Vi minner her om at regnskapsdata henviser til årsskiftet og at dataen fra 2019 i så måte vil vise til selskapets økonomiske situasjon ved utgangen 2019/inngangen til 2020, som er rett i forkant av Covid-19-utbruddet i Norge. Figur 5.16 viser til variablenes viktighet for random forest modell estimert på regnskapsdata fra 2017-2018. Rangeringen er forenelig med funnene ved bruk av RFE, hvor alder og antall ansatte som ble valgt ut av samtlige modeller viser seg å være de klart viktigste variablene. Videre er det nøkkeltallet soliditet som fremstår som den tredje viktigste variabelen tett etterfulgt av henholdsvis kasse/bank/post og opptjent egenkapital.



Figur 5.16. Viktighet av variabler normalår.

Figur 5.17 viser til variablenes viktighet for random forest modell estimert på regnskapsdata fra utgangen av 2019. Modellen vil følgelig inkludere konkursobservasjoner som har gått konkurs i etterkant av utbruddet av Covid-19. Vi kan se fra figuren at det i mindre grad er variabler som skiller seg ut, sammenlignet med figur 5.16. Selskapets alder viser seg igjen å være den viktigste forklaringsvariabelen, nå tett etterfulgt av henholdsvis opptjent- og innskutt egenkapital, omløpsmidler og kasse/bank/post. Imidlertid er det viktig å understreke at man i mindre grad kan vektlegge variablenes forklaringskraft innenfor en modell som

leverer svake resultater (Breiman, 2001). Vår random forest modell estimert på data fra 2019 leverer langt svakere resultater enn modellen estimert på data fra 2017-2018, med en andel riktige klassifikasjoner på kun 0,70. Dette blant annet som følge av langt færre selskaper å trene modellen på, med kun 75 konkursselskaper. Det kan heller ikke utelukkes at midlene staten har bevilget for å forhindre konkurser kan ha hatt en innvirkning. For oppnådde resultater fra random forest modell estimert på regnskapsdata fra 2019, se appendiks 6. Her legges det også ved variablers forklaringskraft ved inkludering av variabelen "tildelt kontantstøtte", som viser til hvor mye kontantstøtte et selskap har mottatt i forbindelse med Covid-19-utbruddet og medfølgende restriksjoner.



Figur 5.17. Viktighet av variabler pandemiår.

6.0 Diskusjon

I dette kapittelet vil studiens resultater diskuteres ytterligere, og sammenlignes opp mot tidligere forskning innenfor konkursprediksjon. Videre vil også diskuteres rundt studiens svakheter og praktiske betydning.

6.1 Sammenligning med tidligere litteratur

I litteraturgjennomgangen ble det gjort rede for de viktigste bidragene i konkursprediksjon som fagfelt, hvor modellene tar utgangspunkt i ulike statistiske metoder for å foreta konkursprediksjoner. Det som er felles for alle modellene er at de i hovedsak benytter seg av

foretaksøkonomiske nøkkeltall. I henhold til litteraturgjennomgangen så kom Altman sin ZETA-modell og Ohlson sin O-score best ut med treffprosent som evalueringskriterium. Disse modellene leverer treffprosenter på henholdsvis 96,2 og 96,1 ett år i forveien. De høye treffprosentene til Altman og Ohlson sine modeller underbygger hvorfor de fremdeles har en sentral rolle innenfor konkursprediksjon pr dags dato. Både ZETA-modellen og O-scoren leverer bedre resultater enn samtlige av våre maskinlæringsmodeller. Det er imidlertid viktig å understreke at disse modellene er spisset inn mot enkeltbransjer, og i så måte vil predikere et mer snevert utvalg selskaper enn denne oppgavens modeller.

Random forest er denne studiens beste modell og leverer bedre resultater overfor ikke-konkursselskaper med en ekte negativ rate på 0,93, sammenlignet med Altmans ZETA-modell med en ekte negativ rate på 0,90. Når det kommer til konkursselskaper derimot, oppnår random forest en ekte positiv rate på 0,90 mot ZETA-modellens 0,96. Ohlsons resultater er i mindre grad sammenlignbare med våre da Ohlson sin studie (1980) benyttet et sterkt skjevfordelt utvalg mellom konkurs- og ikke-konkursselskaper. Videre har vi som nevnt tidligere optimalisert våre modeller med hensyn til AUC-verdi, noe som svekker våre modeller i direkte sammenligning opp mot modeller som er optimalisert med hensyn til treffprosent. Som metaanalysen til Alaka et al. (2017) viser, så er dessuten en så høy treffprosent som i Altman og Ohlson sin studie langt fra normalt, da metaanalysen viser til en gjennomsnittlig oppnådd andel riktige klassifikasjoner på 0,80 og 0,78 for henholdsvis logistisk regresjon og MDA. Altman et al. sin nyere studie (2017) sammenligner også sistnevnte metoder opp mot maskinlæringsmodeller med de samme variablene som utviklet i Altmans egen studie fra 1968. Denne nyere studien konkluderer også med at maskinlæringsmodellene presterer bedre med tanke på treffprosent, men belyser samtidig at Ohlsons og Altmans studier fremdeles er høyst relevante, da disse fremstår som langt mer transparente i tolkningen av resultatene enn maskinlæringsmodellene.

Når det kommer til den utvidede SEBRA-modellen*, som vi ønsker å sammenligne oss direkte opp mot i denne studien, oppnår denne en gjennomsnittlig andel riktige klassifikasjoner på 0,81, samt en AUC-verdi på 0,87. Disse resultatene er forenelige med resultatene fra Bernhardsen og Larsen (2007), som i sin studie kan vise til en andel riktige klassifikasjoner på rett i overkant av 0,80, samt en AUC-verdi på 0,89. Dette gjelder overfor balanserte fordelinger i konkurs- og ikke-konkursselskaper, i likhet med vår studie. Med dette leverer den utvidede SEBRA-modellen* helt jevne resultater når det kommer til treffprosent sammenlignet med Bernhardsen og Larsen (2007), men samtidig en litt svakere

gjennomsnittlig AUC-verdi med en differanse på 0,02. Sammenlignet med den utvidede SEBRA-modellen* leverer samtlige av studiens maskinlæringsmodeller unntatt beslutningstre bedre resultater både med hensyn til AUC-verdi og treffprosent. Beslutningstre på sin side, leverer bedre når det kommer til treffprosent, men dårligere når det kommer til AUC-verdi.

Alaka et al. (2017) sin metaanalyse viser videre til en gjennomsnittlig andel riktige klassifikasjoner for nevrale nettverk, støttevektormaskin og beslutningstre på henholdsvis 0,83, 0,84 og 0,81 fra tidligere forskning innenfor konkursprediksjon. For de samme modellene i denne studien, oppnås det gjennomsnittlige andeler riktige klassifikasjoner på henholdsvis 0,90, 0,89 og 0,86 som kan sies å være gode resultater. Resultatene kan dessuten sies å være forenelig med tidligere fremlagte styrker og svakheter ved de ulike modellene, som tar utgangspunkt i nevnte metaanalyse samt Ravi Kumar & Ravi (2007) sin studie. Her fremlegges det at de mest komplekse maskinlæringsmodellene ofte oppnår de beste resultatene overfor denne typen problemstillinger, men at modellene til gjengjeld kan ansees som lite transparente når det kommer til tolkningen av klassifikasjonene. Random forest, nevrale nettverk og støttevektormaskin er alle å anse som komplekse og krevende modeller i forbindelse med utforming. Det er nettopp disse modellene som gjennomgående leverer best resultater i denne oppgaven både med tanke på treffprosent og AUC-verdi.

6.2 Variablenes forklaringskraft

I kapittel 5.2 ble det lagt frem viktigheten av de ulike uavhengige variablene overfor studiens modeller. I tabell 5.2 fremkommer det gjenværende variabler for hver av modellene etter at RFE er gjennomført. Alder, antall ansatte og soliditet er variablene som peker seg ut som de viktigste da de er gjennomgående i samtlige modeller. Videre følger regnskapspostene kasse/bank/post og lønnskostnader som er valgt ut av fire av de fem maskinlæringsmodellene. Utover nevnte variabler, inngår de resterende kun i 1-2 av maskinlæringsmodellene. Ut ifra dette kan det bemerkes at de viktigste variablene overfor studiens modeller er en kombinasjon av enkle regnskapsposter, nøkkeltall og det som kan betraktes som generell selskapsinformasjon i form av antall ansatte og selskapets alder.

Metaanalysen til Bellovary et al. (2007) analyserer studier innenfor konkursprediksjon fra 1930 og frem til studiens utgivelse. Her trekkes det frem at blant de gjennomgåtte studiene inkluderes det alt fra 1-57 variabler i forsøket på å predikere konkurser. Videre inngår det i metaanalysen totalt 752 ulike uavhengige variabler. Det belyses også hvilke typer variabler som oftest er tatt i bruk hvor de 35 mest benyttede variablene er nøkkeltall. Kun to variabler

som inngår i mer enn fem studier er enkle regnskapsposter. Dette er med på å understreke dominansen av nøkkeltall innenfor konkursprediksjon, som også er belyst i denne studiens litteraturgjennomgang. Dette står i sterk kontrast med endelige utvalgte variablene for modellene i denne studien. Her er det tre variabler som velges ut av samtlige modeller, hvor kun ett av disse (soliditet) er et finansielt nøkkeltall. Lønnsomhet og likviditet på sin side, blir kun valgt ut som en viktig forklaringsvariabel av random forest. Videre indikerer studiens modeller at selskapsinformasjon utover det som fremkommer av regnskapet kan være vel så viktig for å predikere hvorvidt et selskap vil gå konkurs eller ikke. Dette er informasjon som i liten grad har blitt inkludert i tidligere forskning, men som belyst tidligere inngår noe av denne typen informasjon i Bernhardsen og Larsen (2007) sin utvidede SEBRA-modell.

Studien har også sammenlignet viktigheten av de ulike variablene innad i en random forest modell estimert på regnskapsdata fra henholdsvis et normalår og et pandemiår. Resultatene viser til endringer i variablenes viktighet mellom disse modellene. Under normale omstendigheter er det særlig generell selskapsinformasjon i form av antall ansatte og selskapets alder som skiller seg ut som spesielt viktig. Disse er etterfulgt av nøkkeltall for soliditet samt regnskapstallet kasse/bank/post. Dette i likhet med øvrige modeller ved seleksjon gjennom RFE. Under pandemitider kan man se konturen av at andre variabler skiller seg ut, som eksempelvis opptjent og innskutt egenkapital. Samtidig fremkommer det at forklaringsnivået blant variablene er mer jevnfordelt. Resultatene fra modellen estimert på regnskapsdata fra 2019 skal derimot ikke legges for mye vekt på, da viktigheten av variablene ikke er helt troverdig når modellen i seg selv leverer dårlige prediksjoner (Breiman, 2001). I så måte blir det vanskelig å konkludere med om forskjellene skyldes en faktisk endring som følge av eksempelvis de mange støtteordningene for å forhindre konkurser, eller at estimatene er unøyaktige.

6.3 Praktisk betydning

Resultatene fra studien viser at maskinlæringsmodeller kan være relevante for aktuelle beslutningstakere med bakgrunn i tidligere belyste konsekvenser. Den beste modellen (random forest) kan vise til en gjennomsnittlig andel riktige klassifikasjoner overfor konkursselskaper på 0,90 samt 0,93 overfor ikke-konkursselskaper. Om man sammenligner dette opp mot den utvidede SEBRA-modellen*, kan denne vise til en andel riktige klassifikasjoner overfor konkursselskaper på 0,82, samt 0,81 overfor ikke-konkursselskaper. Random forest oppnår med dette høyere andeler riktige klassifikasjoner på om lag 0,10. Som

nevnt innledningsvis så opplevde kreditorer her til lands et samlet kredittap på 26,3 milliarder NOK i 2020. Om man svært forenklet og hypotetisk sier at 10 prosent av dette kunne vært unngått ved å forutse riktig skjebne for 10 prosent flere selskaper ved utlån, får man et bilde av hvilken praktisk betydning en prediktiv forbedring kan medføre. Med denne forenklingen, vil det å kunne forutse 10 prosent flere utlånstap tilsvare i overkant av 2,6 milliarder NOK i unngått tap. Derfor vil kostanden forbundet med utviklingen av nye og forbedrede konkursprediksjonsmodeller enkelt kunne forsvares, når man ser hvilke tap en potensielt kan unngå. Det er også andre instanser i samfunnet som vil ha nytte av en forbedret konkursprediksjonsmodell. Som gjort rede for i innledningen har konkurser stor betydningen for samfunnet som helhet, hvor eksempelvis Norges Bank ville hatt stor nytte av forbedrede prediktive modeller.

Studiens resultater er også et tilskudd til tidligere forskning innenfor konkursprediksjon. Våre funn er forenelige med Alaka et al. (2017) sin metaanalyse som poengterer at maskinlæringsmodeller presterer bedre enn tradisjonelle statistiske modeller. Videre peker også modellene i retning av samme konklusjon som Rhie et al. (2005), som konkluderer med at foretaksøkonomiske nøkkeltall sin betydning har avtatt gradvis over tid.

6.4 Svakheter ved oppgaven

Modellene i studien er gjennomgående estimert og testet mot informasjon kun ett år i forkant av den aktuelle hendelsen, og har i så måte ikke avklart i hvilken grad modellene er anvendbare flere år i forkant av en eventuell konkurs. Tidligere studier viser at mange modeller oppnår svært høye andeler riktige klassifikasjoner ett år i forkant, før resultatene faller raskt så snart tidshorisonten øker til to eller flere år. Denne mangelen foreligger som følge av begrensninger på hvor mange regnskap som kunne hentes ut fra Proff Forvalt sine databaser pr døgn, som resulterte i et fokus på den forholdsvis korte tidsperioden 2017-2019. Begrensningen på uthenting av data førte også til at det kun ble hentet ut informasjon fra ti bransjer, og modellene er i så måte spisset inn mot disse. Med dette vil det være tydelige begrensninger i hvor anvendbare modellene er i situasjoner hvor man står overfor selskaper fra alle mulige bransjer, samtidig som en gjerne skulle forutse et selskaps skjebne mer enn ett år frem i tid.

Mangel på tilgang på ulike variabler som inngår i den utvidede SEBRA-modellen har dessuten ført til at studien sammenligner maskinlæringsmodellene opp mot en noe forenklet og potensielt svekket SEBRA-modell. Maskinlæringsmodellene har uavhengig av dette levert

gode resultater, men differansen til den utvidede SEBRA-modellen* kan ha blitt større enn den ville vært med tilgang til all nødvendig informasjon for å kopiere modellen.

Studien har også begrensninger når det kommer til tolkning av variablenes viktighet. Som belyst i metaanalysen til Alaka et al. (2017), er det gjerne negativt samsvar mellom en maskinlæringsmodells kompleksitet og tolkningsmuligheter av variablene. Studien har pekt på hvilke variabler som fremstår som de viktigste ved trening av modellene. Til tross for dette har den i nokså liten grad identifisert hvor viktige disse er i forhold til hverandre. Studien skulle også se på utviklingen av variablenes viktighet mellom et normalår og pandemiår hvor det som følge av svake resultater for modellen estimert på 2019-data, dessverre ikke kan trekkes noen tydelige konklusjoner. Om antall konkurser stiger etter hvert som støtteordningene forsvinner, vil dette på sikt gi større datagrunnlag som vil kunne gi bedre maskinlæringsmodeller og dermed mer troverdige estimater av variablenes viktighet. Mer detaljerte undersøkelser av variablers viktighet mellom normalår og pandemiår virker som en interessant videreføring av våre analyser.

7.0 Konklusjon

Konkursprediksjon som fagfelt har sin opprinnelse tilbake til 1930-tallet, og har vært i stadig utvikling siden. I nyere tid har tilgangen på forbedret teknologi gjort at man kan bruke maskinlæringsmodeller som prediksjonsverktøy. Denne studien har først og fremst fokusert på å utvikle maskinlæringsmodeller som predikerer konkurser bedre enn den utvidede SEBRA-modellen*. Studien har også undersøkt hvorvidt foretaksøkonomiske nøkkeltall er like sentrale i maskinlæringsmodeller, som de har vist seg å være innenfor tidligere forskning ved bruk av statistiske modeller.

Samtlige modeller i studien har blitt estimert og testet på jevnfordelte utvalg av konkurs- og ikke-konkursselskaper, med en klassifiseringsterskel på 0,5. Hver modell har gjennomgått femdelt kryssvalidering før de standardiserte variablene har blitt filtrert gjennom bruk av RFE med optimalisering av AUC-verdi som formål. Maskinlæringsmodellenes hyperparametere har også blitt optimalisert med hensyn til AUC-verdi. I studien leverer random forest de beste resultatene med treffprosent som evalueringskriterium, med en gjennomsnittlig andel riktige klassifikasjoner på 0,92. Denne er etterfulgt av nevrale nettverk og støttevektormaskin med andeler riktige klassifikasjoner på henholdsvis 0,90 og 0,89. Til sammenligning har den utvidede SEBRA-modellen* en andel riktige klassifikasjoner på 0,81. Også når det kommer

til AUC-verdi som evalueringskriterium, leverer random forest, nevrale nettverk og støttevektormaskin de tre beste resultatene med verdier på henholdsvis 0,97, 0,94 og 0,94. Her leverer den utvidede SEBRA-modellen* en verdi på 0,87. Random forest, nevrale nettverk, støttevektormaskin og k-nærmeste nabo leverer bedre resultater både med hensyn til AUC-verdi og treffprosent. Beslutningstre på sin side leverer bedre enn den utvidede SEBRA-modellen* når det kommer til treffprosent, men svakere når det kommer til AUC-verdi. Studien har med dette utviklet flere maskinlæringsmodeller som predikerer bedre enn den utvidede SEBRA-modellen*.

Modellene i studien vektlegger en kombinasjon av regnskapstall, foretaksøkonomiske nøkkeltall og generell selskapsinformasjon ved optimalisering gjennom bruk av RFE. Samtlige modeller velger ut generell selskapsinformasjon i form av antall ansatte og selskapets alder som de mest sentrale for å skille mellom konkurs- og ikke-konkursselskaper. Nøkkeltall for soliditet velges også ut av samtlige modeller, samtidig som regnskapspostene kasse/bank/post og lønnskostnader velges ut fra fire av fem modeller. Funnene er forenelige med at informasjon utover foretaksøkonomiske nøkkeltall, som i tidligere forskning i stor grad har blitt ignorert, kan vise seg vel så viktige når det kommer til å identifisere hvorvidt et selskap vil gå konkurs eller ikke. Avslutningsvis har studien sett på variablers forklaringskraft innad i random forest modeller estimert på henholdsvis data fra normalår og et pandemiår. Resultatene antyder at det er en endret distribusjon av variablenes viktighet under et pandemiår sammenlignet mot et vanlig år, hvor det i liten grad er enkeltvariabler som peker seg ut som spesielt viktige. Om dette skyldes ustabile estimater fra modellen eller om det faktisk er en endring i variablenes viktighet kan vi ikke konkludere med, men det ville vært et interessant tema å forske på videre.

Referanser

Agarwal, V. & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance, 32*(8), 1541-1551. https://doi.org/10.1016/j.jbankfin.2007.07.014

Alaka, H. A., Oyedele, L. O, Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O, & Bilal, M. (2017). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems With Applications, 94*, 164-184. https://doi.org/10.1016/j.eswa.2017.10.040

Altman, E. I., Kimura., H & Barboza, F. (2017). Machine learning models and bankruptcy prediction. *Expert Systems With Applications, 83*, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.2307/2978933

Altman, E. I. (2000). Predicting financial distress of companies: Revisiting the z-score and zeta models. *Research Papers in Economics*, 1-52. https://doi.org/10.4337/9780857936097.00027

Altman, E. I., Haldeman, R., & Narayanan, P. (1977). ZETATM: analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance, 1*(1), 29-54. https://doi.org/10.1016/0378-4266(77)90017-6

Beaver, W. H. (1966) Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71-111. https://doi.org/10.2307/2490171

Bellovary, J. L., Giacomio, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education, 33*, 1-42. https://www.jstor.org/stable/41948574

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry, 23*(2), 129-143. https://doi.org/10.1002/asmb.658

Bernhardsen, E. (2001). A model of bankruptcy prediction. *Arbeidsnotat fra Norges Bank*, (10). https://www.norges-bank.no/globalassets/upload/import/publikasjoner/arbeidsnotater/pdf/arb-2001-10.pdf?v=03/09/2017122305&ft=.pdf

Bernhardsen, E. & Larsen, K. (2007). Modellering av kredittrisiko i foretakssektoren . videreutvikling av SEBRA-modellen. *Penger og kreditt*, (2). https://www.norges-bank.no/contentassets/94bda7707e9240baaa8fe376177ac4d4/modellering-av-kredittrisiko.pdf?v=03/09/2017123240&ft=.pdf

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2017, 26. juli). *What is the difference between a parameter and a hyperparameter?* Machine learning mastery. https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/

Brownlee, J. (2019, 27. november). *How to choose a feuture selection method for machine learning.* Machine learning mastery. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brynestad, E. I. (2021, 12. februar). *Svakere årsresultat for bankene.* Statistisk sentralbyrå. https://www.ssb.no/bank-og-finansmarked/artikler-og-publikasjoner/svakere-arsresultat-for-bankene

Brønnøysundregistrene. (2020, 14. mai). *Næringskodar*. https://www.brreg.no/bedrift-nn/naeringskodar/

Caballero, R. J. & Hammour, M. L. (1994). The cleansing effect of recessions. *The American Economic Review, 84*(5), 1350-1368. https://www.jstor.org/stable/2117776

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Datacamp. (2020, 14. april). *R Tutorial: Nonlinear Modeling in R with GAMs | Intro*. [Video]. YouTube. https://youtu.be/6V_VvweZkoI

Eklund, T., Bernhardsen, E. & Larsen, K. (2001). Model for analysing credit risk in the enterprise sector. *Penger og kreditt,* (2). https://norges-bank.brage.unit.no/norges-bank-xmlui/bitstream/handle/11250/2480734/eklund.pdf?sequence=1

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fredriksen, K. (2021, 4. februar). *Færre konkurser i koronaåret 2020.* Statistisk sentralbyrå. https://www.ssb.no/virksomheter-foretak-og-regnskap/artikler-og-publikasjoner/faerre-konkurser-i-koronaaret-2020

Graupe, D. (2013). *Principles of artificial neural networks* (3.utg.). Singapore: World scientific publishing company.

Grybinenko, O. (2017). Social and economic consequences of bankruptcy of the companies in Ukraine. *EUREKA: Social and Humanities, 2*, 3-10. https://doi.org/10.21303/2504-5571.2017.00298

Han, J., Pei, J. & Kamber, M. (2000). *Data mining: Concepts and techniques* (3.utg.). Saint Louis: Elsevier science & technology.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical Learning: Data mining, inference, and prediction* (2.utg.). Springer-Verlag New York Inc.

Hastie, T & Tibshirani, R (1995). Generalized additive models for medical research. *Statistical methods in medical research, 4*(3), 187-196. https://doi.org/10.1177/096228029500400302

Hillegeist, S. A., Keating, E. K., Cram, D. P., Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies, 9*(1), 5-34. https://doi.org/10.1023/B:RAST.0000013627.90884.b7

Huang, J. og Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms, knowledge and data engineering. *IEEE Transactions on, 17*(3), 299-310. https://doi.org/10.1109/TKDE.2005.50

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics, 1*(1), 519–537. https://doi.org/10.1214/07-EJS039

James, G., Witten, D., Hastie, T. og Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* Springer-Verlag New York Inc.

Kirkos, E. (2012). Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review, 43*(1), 83-123. https://doi.org/10.1007/s10462-012-9367-6

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence, 2,* 1137-1143. https://dl.acm.org/doi/10.5555/1643031.1643047

Konkursloven. (1984). *Lov om gjeldsforhandling og konkurs* (LOV-1984-06-08-58). Lovdata. https://lovdata.no/dokument/NL/lov/1984-06-08-58

Konkursrådet. (2012). *Innføring i konkurs.* https://www.konkursradet.no/innfoering-i-konkurs.305266.no.html

Kommunal- og moderniseringsdepartementet. (u.å). *Nasjonal strategi for kunstig intelligens.* https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=3

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling* (5.utg.). Springer-Verlag New York Inc.

Larsen, K og Sæther, P. (1999). Kredittrisiko knyttet til foretakssektoren. *Penger og Kreditt (4).* https://norges-bank.brage.unit.no/norges-bank-xmlui/bitstream/handle/11250/2480453/pk994kredittrisiko.pdf?sequence=1

Ledolter, J. (2013). *Data mining and business analytics with R.* John Wiley & Sons, Inc.

Lennox, C. (1999). Are large auditors more accurate than small auditors? *Accounting and Business Research, 29*(3), 217-227. https://doi.org/10.1080/00014788.1999.9729582

Liang, D., Tsai, C. & Wu, H. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73,* 289-297. https://doi.org/10.1016/j.knosys.2014.10.010

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Meese, E. N. & Viken, T. (2019). *Machine Learning in Bankrupcy Prediction: Utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements* [Masteroppgave]. Norges handelshøyskole.

Noble, W. S. (2006). What is a support vector machine?. *Nat Biotechnol*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109-131. https://doi.org/10.2307/2490395

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine learning Research*, *12*(85). 2825−2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Proff Forvalt. (u.å.). *Regnskap. https://forvalt.no/Regnskap/Regnskap*

Proff Forvalt. (u.å.). *Nøkkeltall – forklaringer og formler.* http://innsikt.proff.no/hjelp-og-tips/nokkeltall-forklaringer-og-formler/

Ravi K. P. & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, *180*(1), 1-28. https://doi.org/10.1016/j.ejor.2006.08.043

Rhie, J., Beaver, W. H. & Mcnichols, M. F. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, *10*(1), 93-122. https://doi.org/10.1007/s11142-004-6341-9

Saini, R. (2019, 21. oktober). *Feauture scaling - Why it is required?.* Medium. https://medium.com/@rahul77349/feature-scaling-why-it-is-required-8a93df1af310

Schumpeters, J. (1934). *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle.* Harvard Economic Studies 46.

Scikit-learn. (u.å). *Feature importances with forests and trees.* https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Scikit-learn. (u.å). *Sklearn.ensemble.RandomForestClassifier.* https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit-learn. (u.å). *Sklearn.neighbors.KNeighborsClassifier*. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Scikit-learn. (u.å). *Sklearn.neural_network.MLPClassifier*. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Scikit-learn. (u.å). *Sklearn.svm.NuSVC.* https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html

Scikit-learn. (u.å). *Sklearn.tree.DecisionTreeClassifier*. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Shaikh, R. (2018, 28. oktober). *Feauture selection techniques in machine learning with Python.* Towards data science. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Skatteetaten. (2020, 24. mars). *Innsynsløsning for kompensasjonsordningen blir åpen for alle – nå er den klar*. https://www.skatteetaten.no/presse/nyhetsrommet/innsynslosning-for-kompensasjonsordningen-blir-apen-for-alle--na-er-den-klar/

Skatteetaten. (2020). *Kompensasjonsordningen for næringslivet (mars-august 2020). Innsyn i vedtak om tildelt tilskudd.* https://www.skatteetaten.no/kompensasjonsordning/innsyn/

Statistisk sentralbyrå. (2020, 30. september). *Nyetablerte foretaks overlevelse og vekst.* https://www.ssb.no/fordem

Statistisk sentralbyrå. (2021). Opna konkursar. https://www.ssb.no/statbank/table/07165/tableViewLayout1/

Statistisk sentralbyrå. (2021, 30. april). *Regnskap for ikke-finansielle aksjeselskaper.* https://www.ssb.no/virksomheter-foretak-og-regnskap/statistikker/regnno

Statistisk sentralbyrå. (2021, 7. januar). *Virksomheter.* https://www.ssb.no/virksomheter-foretak-og-regnskap/statistikker/bedrifter/aar

StatQuest with Josh Starmer. (2018, 17. september). *Machine Learning Fundamentals: Bias and Variance.* [Video]. YouTube. https://youtu.be/EuBBz3bI-aA

StatQuest with Josh Starmer. (2020, 31. august). *Neural Networks Pt.1: Inside the Black Box.* [Video]. YouTube. https://youtu.be/SqgtYHlKhqA

Vuckovic, A., Popovic, D. & Radivojevic, V. (2002). Artificial neural network for detecting drowsiness from EEG recordings. *Medical Engineering and Physics, 24,* 349-360. https://doi.org/10.1109/NEUREL.2002.1057990

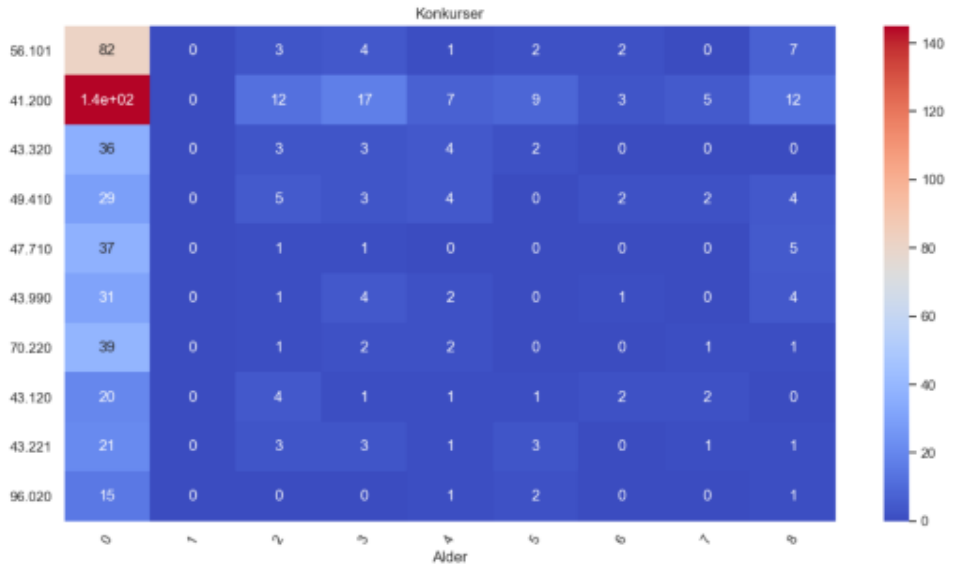
Wahlstrøm, R. R. & Helland, F. F. (2016), *Konkursprediksjon for norske selskaper – en analyse ved maskinlæringsteknikker og tradisjonelle statistiske metoder* [Masteroppgave]. Norges teknisk-naturvitenskapelige universitet. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2409760/Fredrik%20F.%20Helland%20og%20Ranik%20R.%20Wahlstroem%2C%20Masteroppgave%20%20v%C3%A5ren%202016.pdf?sequence=1&isAllowed=y

Wiley, M & Wiley, J. F. (2019). Advanced R statistical programming and data models: Analysis, machine learning, and visualization. Berkeley, CA: Apress L. P.

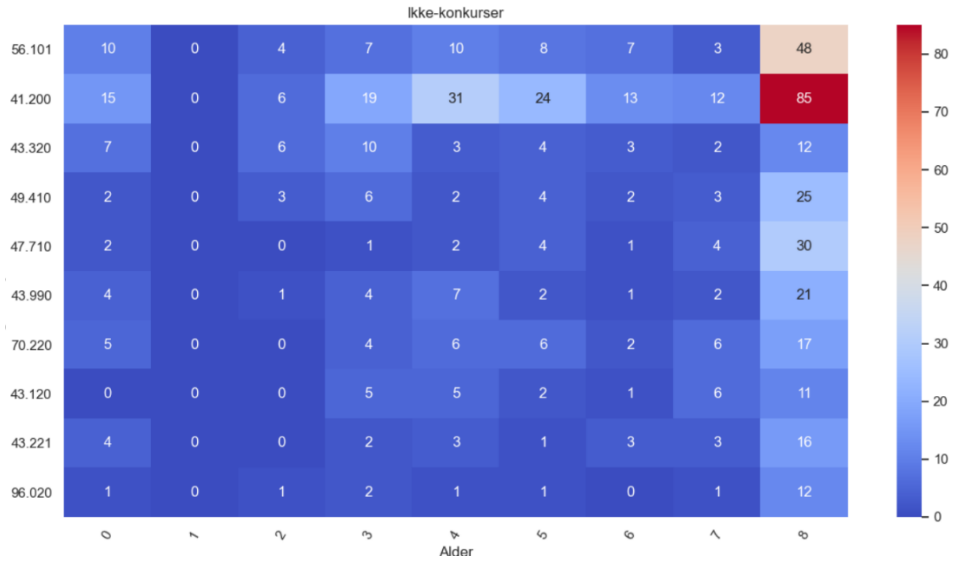
Appendiks

Appendiks 1- Fordeling av observasjoner etter alder

Figur A.1 og A.2 under viser fordelingen av selskaper i datagrunnlaget vårt sortert etter selskapets alder. Det fremkommer av figur A.1, som viser til konkursselskapene i datagrunnlaget, at det er stor overvekt av nyetablerte selskaper i denne kategorien. Figur A.2 på den annen side, tar for seg ikke-konkurser og viser at majoriteten av disse selskapene er veletablerte selskaper som har en alder større eller lik åtte år.



Figur A.1. Distribusjon av konkursselskaper etter selskapets alder.



Figur A.2. Distribusjon av ikke-konkursselskaper etter selskapets alder.

Appendiks 2 – Computer spesifikasjoner og Python-bibliotek

Studiens analyser har blitt gjennomført på computer med følgende spesifikasjoner:

- Operativsystem: Windows 10 Home x64, versjon 20H2.
- Prosessor: AMD Ryzen 7 3700U @ 2,30 GHz, 16GB RAM.

Studien har anvendt Microsoft Visual Studios Community 2019 som utviklingsverktøy med Python som programmeringsspråk. Tabell A.1 viser til anvendt Python-versjon med bibliotek.

Tabell A.1. Programversjoner

Navn	Versjon
Microsoft Visual Studio	2019
Python	3.7.8
Numpy	1.20.1
Matplotlib	3.3.4
Seaborn	0.11.1
Pandas	1.2.3
Scikit-learn	0.24.1
Pygam	0.8.0
Requests	2.25.1

Appendiks 3- Inkluderte bransjer

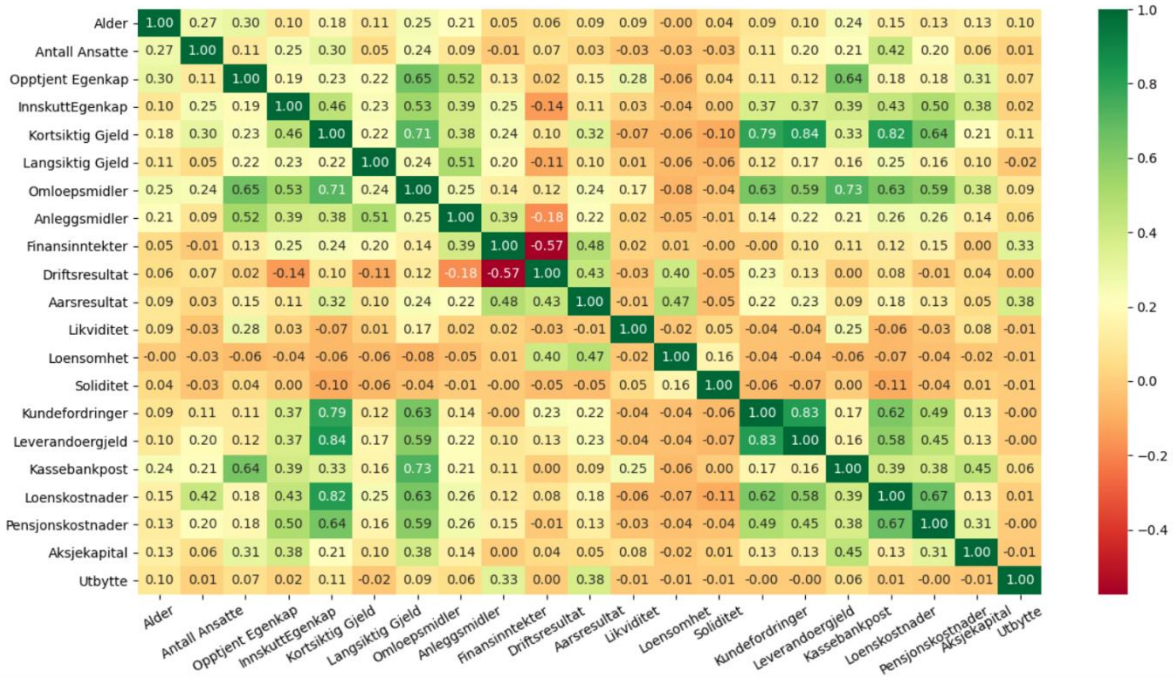
De inkluderte bransjene i vårt datagrunnlag utgjør de ti bransjene med høyest konkurstall i perioden 2018-2020, og er i tabell A.2 sortert fra høyest antall konkurser øverst. NACE-koden skal virke som en beskrivelse for hovedaktiviteten til en virksomhet. Bokstavkode for næringshovedområde inngår ikke i selve NACE-koden, og i tillegg til den firesifrede NACE-koden etter EU-standard er flere av datasettets koder splittet opp på nasjonalt nivå gjennom inkludering av et femte siffer (Brønnøysundregistrene, 2020).

Tabell A.2. Bransjebeskrivelse med tilhørende NACE-bransjekode.

Bransjebeskrivelse (NACE)	NACE-bransjekode
Oppføring av bygninger	41.200
Drift av restauranter og kafeer	56.101
Snekkerarbeid	43.320
Godstransport på vei	49.410
Butikkhandel med klær	47.710
Annen spesialisert bygge- og anleggsvirksomhet	43.990
Bedriftsrådgivning og annen administrativ rådgivning	70.220
Rørleggerarbeid	43.221
Grunnarbeid	43.120
Frisering og annen skjønnhetspleie	96.020

Appendiks 4- Korrelasjon mellom uavhengige variabler

Figur A.3 illustrerer korrelasjonsmatrisen for de inkluderte uavhengige variablene i vår studie. Korrelasjonen viser til den eventuelle lineære sammenhengen mellom ulike parvise sammensetninger av de uavhengige variablene i modellene. Diagonalen i matrisen fra øvre venstre hjørne til nedre høyre hjørne representerer sammenligning av hver av variablene med seg selv og vil derav indikere perfekt korrelasjon.



Figur A.3. Korrelasjonsmatrise for inkluderte uavhengige variabler.

Appendiks 5- Valgte verdier for hyperparametere

Vi har valgt å ekskludere enkelte hyperparametere fra optimaliseringen, noe som vil være med på å redusere hvor krevende gjennomkjøringen blir for computer. I tråd med brukerveiledningen til Scikit-learn for beslutningstrær har vi valgt å tune hyperparameteren som avgjør hvor mange variabler som skal hensyntas for hver forgrening innad i treet. Videre har vi også optimalisert modellen når det kommer til maksimalt antall løv og noder i treet, for å hensynta problemstillingen forbundet med overfitting og underfitting. Hvilken funksjon som anvendes i modellen for å måle kvaliteten til hver forgrening, samt hvilken strategi som skal benyttes, har også blitt optimalisert.

Når det kommer til random forest er dette en krevende modell ved gjennomkjøring for computer. Basert på dette har vi fulgt Scikit-learn sin anbefaling i brukerveiledningen om å kun optimalisere parameterne som avgjør hvor mange trær som skal inngå i skogen og antall

variabler som skal hensyntas ved hver forgrening i hvert tre. Øvrige hyperparametere er satt lik standardverdiene fra Scikit-learn.

Som følge av at k-nærmeste nabo ikke er å anse som en krevende modell å kjøre for computer, har vi valgt å optimalisere hyperparameterne som avgjør antall naboer klassifikasjonen skal baseres på, hvordan observasjonene skal vektes og hvilken algoritme innad i modellen som skal brukes for å beregne avstanden til nærmeste nabo.

For støttevektormaskin har vi benyttet Scikit-learn (u.å) sine standardverdier for de ulike hyperparameterne, bortsatt fra parameterne "gamma", "probability" og "nu". Gamma er en koeffisient som tillegges rbf-funksjonen og avgjør dermed graden av krumning på modellens hyperplan eller beslutningsgrense. Videre henviser "probability" til bruken av innebygde sannsynlighetsfunksjoner ved estimering av en observasjons sannsynlighet for konkurs eller ikke-konkurs. Sistnevnte "nu" henviser til den tillatte maksimale feilmarginen.

Nevralt nettverk er den mest krevende modellen for computer ved gjennomkjøring. Det er også den modellen med flest hyperparametere å hensynta i Scikit-learn sitt bibliotek. Vi har optimalisert modellens hyperparametere for antall skjulte lag med tilhørende noder, maksimalt antall gjentakelser av prosessen og innledende læringsrate.

Nedenfor fremlegges først fremgangsmåten for gjennomkjøringen av grid search for beslutningstre som også vil illustrere fremgangsmåten for de øvrige modellene. Enkelte hyperparametere står overfor alternative kategorier som er utformet av scikit-learn, mens andre hyperparametere står overfor tall i et ubestemt intervall. Her må man selv bestemme hvilke verdier hyperparameteren skal testes opp mot. For beslutningstre kjøres første gjennomkjøring med et testintervall med verdier mellom 1 og 100 for treetes maksimale antall løv og noder. Når man får utlevert det optimale antallet fra første gjennomkjøringen, kan man etablere et nytt intervall som er mer finjustert rundt output fra første gjennomkjøring for å finjustere modellen.

Første gjennomkjøring:

'max_depth': [1, 10, 20, 30, 40, 50, 100], beste verdi: 20,

'max_features': ['sqrt', 'log2', 'auto'], beste verdi: 'sqrt',

'criterion': ['gini', 'entropy'], beste verdi: 'gini',

'max_leaf_nodes': [1, 10, 20, 30, 40, 50, 100], beste verdi: 40,

'splitter': ['best', 'random'], beste verdi: 'best'.

Andre gjennomkjøring (finjustering av intervaller):

'max_depth': [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25], beste verdi: 13,

'max_leaf_nodes': [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45], beste verdi: 37,

Uendret resultat for øvrige hyperparametere,

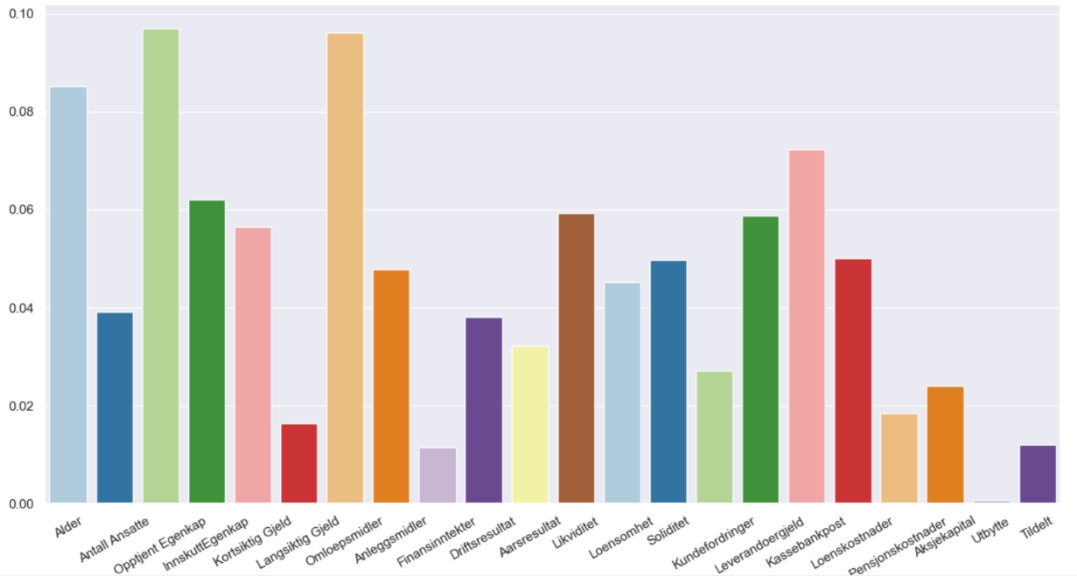
Følgelig velges det som fremkommer som beste verdi gjennom grid search. Øvrige hyperparametere som ikke inngår har blitt satt lik standardverdien fra Scikit-learn sin brukerveiledning. Tabell A.3 viser til hyperparametere hvor optimalisering gjennom grid search har gitt en 'beste verdi' som avviker fra standardverdien som fremkommer av Scikit-learn sin brukerveiledning for de aktuelle modellene. Brukerveiledningen for de ulike modellene fremkommer i referanselisten.

Tabell A.3. Hyperparametere som er endret fra Scikit-learn sine standardverdier for alle modellene.

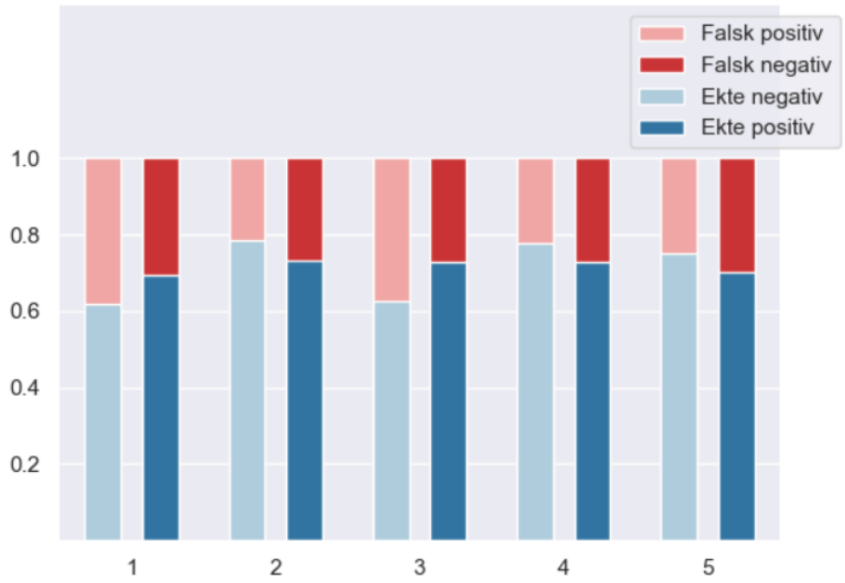
Modell - Hyperparameter	Beste verdi
Beslutningstre	
- Max depth	13
- Max features	'sqrt'
- Max leaf nodes	37
- Random state	18
Random forest	
- Max features	'sqrt'
- N estimators	168
- Random state	0
K-nærmeste nabo	
- N neighbors	23
- Weights	'distance'
Støttevektormaskin	
- Gamma	'auto'
- Nu	0,3
- Probability	'True'
Nevralt nettverk	
- Hidden layer sizes	(50,100,)
- Alpha	0,0045
- Max iter	10000
- Learning rate init	0,01
- Random state	1

Appendiks 6- Random forest utviklet med regnskapsdata fra 2019.

Figur A.4 viser til variablenes viktighet for random forest modell estimert på regnskapsdata for 2019, hvor variabelen "tildelt kontantstøtte" er inkludert. Modellen som inkluderer tildelt kontantstøtte som variabel er kun trent på 33 konkursselskaper og likt antall ikke-konkursselskaper. Det fremkommer at denne variabelen er blant de minst viktige variablene for modellen, men igjen skal man ikke legge for mye vekt på disse variablenes viktighet grunnet lav treffandel som fremkommer av figur A.5.



Figur A.4. Viktigheten av variabler pandemiår med tildelt kontantstøtte.



Figur A.5. Treffandel fra random forest med regnskapsdata 2019.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway