Norges miljø- og biovitenskapelige universitet

Masteroppgave 2024 30 stp
Fakultet for realfag og teknologi

Maskinlæring for prediksjon av behandlingsutfall hos pasienter med hode- og halskreft

Machine Learning for Predicting Treatment Outcome for Patients with Head and Neck Cancer

Helene Glemming
Miljøfysikk og fornybar energi

Forord

Denne masteroppgaven er skrevet ved Fakultetet for realfag og teknologi ved Norges miljø- og biovitenskapelige universitet høsten 2023 og våren 2024. Denne markerer et opphold på seks år i Ås, og en fullført mastergrad i Miljøfysikk og fornybar energi.

En hjertelig takk rettes til min hovedveileder, professor Cecilia Marie Futsæther, for veiledning, engasjement og verdifulle innspill gjennom arbeidet med denne oppgaven. Jeg vil også få takke førsteamanuensis Oliver Tomic, for gode tilbakemeldinger og interessante diskusjoner. En spesiell takk til ph.d.-stipendiat Bao Ngoc Huynh, hvis råd og tilgjengelighet alltid har vært til stor hjelp.

Til slutt vil jeg gi en stor takk til familie og venner for all støtte og hjelp. Takk også til alle medstudenter som har vært med på å gjøre studietiden min i Ås helt spesiell.

Helene Glemming

Ås, 15. februar 2024

Sammendrag

Det er i Norge, for hvert år, omtrent 38 000 mennesker som får kreft, og ut fra disse er 800 av tilfellene hode- og halskreft. Da dette tallet er forventet å øke, vil også behovet for diagnosering og behandling økes. Maskinlæring og predikering av behandlingsutfall, kan hjelpe med en mer nøyaktig diagnosering og effektiv behandling. Siden det ved kreftbehandling er fare for å få senskader, vil også en prediksjon av utfallet, i det tilfellet det lar seg gjøre, kunne gi en mer tilpasset behandling og dermed muligens redusere faren for mulige senskader.

Formålet med denne oppgaven har vært å se hvordan maskinlæringsmodeller evner å predikere behandlingsutfallet til pasienter med hode- og halskreft, basert på kliniske egenskaper og radiomics-egenskaper hentet ut fra PET- og CT-bilder. Målet videre var å undersøke hvorvidt det var betydelige forskjeller på ytelsesprediksjonene for pasienter med positiv HPV-status og pasienter med negativ HPV-status.

Tre datasett bestående av pasienter med hode- og halskreft har vært benyttet i denne oppgaven, hvorav disse har blitt delt inn i seks nye datasett basert på om pasienten har negativ eller positiv HPV-status. Dette har blitt gjort med sykdomsfri overlevelse (DFS) og generell overlevelse (OS) som utfall for pasientene. Datasettene som er brukt består av 139 pasienter behandlet ved Oslo universitetssykehus i perioden 2007-2013, og 99 pasienter behandlet ved Maastricht University Medical Center i perioden 2008-2014. På datasettene ble det benyttet fire ulike klassifiseringsalgoritmer: logistisk regresjon, random forest, decision tree og K-nærmeste naboer. Det ble brukt en grid search-kryssvalidering der ytelsesresultatene ble gitt ved nøyaktighet, ROC AUC, Matthews korrelasjonskoeffisient og F1-score for begge klasser.

Av de seks datasettene og fire klassifiseringsalgoritmene benyttet, var det ingen som utmerket seg spesielt på prediksjonsytelse. Det var veldig jevnt i ytelsesresultater for datasettene bestående av pasienter med positiv HPV-status og datasettene bestående av pasienter med negativ HPV-status. For å finne mønstre eller karakteristikk i egenskapene til pasienter med positiv HPV-status eller pasienter med negativ HPV-status, krever det at det testes på flere datasett.

Abstract

In Norway, each year, about 38 000 people get cancer, and from these, 800 of the cases are head and neck cancer. As this number is expected to increase, the need for diagnosis and treatment will also increase. Machine learning and prediction of treatment outcomes can help with a more accurate diagnosis and effective treatment. Since there is a risk of late-onset injuries with cancer treatment, a prediction of the outcome, in the case where it can be done, could provide a more adapted treatment and thus possibly reduce the risk of possible late-onset injuries.

The purpose of this thesis has been to see how machine learning models are able to predict the treatment outcome of patients with head and neck cancer, based on clinical features and radiomics features extracted from PET and CT images. The further aim was to investigate whether there were significant differences in the performance predictions for patients with positive HPV status and patients with negative HPV status.

Three datasets consisting of patients with head and neck cancer have been used in this thesis, where these have been divided into six new datasets based on whether the patient has a negative or positive HPV status. This has been done with disease-free survival (DFS) and overall survival (OS) as outcomes for the patients. The datasets used consist of 139 patients treated at Oslo University Hospital in the period 2007-2013, and 99 patients treated at Maastricht University Medical Center in the period 2008-2014. Four different classification algorithms were used on the datasets: logistic regression, random forest, decision tree and K-nearest neighbours. A grid search cross-validation was used where the performance results were given by accuracy, ROC AUC, Matthew's correlation coefficient and F1 score for both classes.

Of the six datasets and four classification algorithms used, none particularly excelled in terms of prediction performance. Performance results were very consistent for the datasets consisting of patients with positive HPV status and the datasets consisting of patients with negative HPV status. Finding patterns or characteristics in the features of patients with positive HPV status or patients with negative HPV status requires testing more datasets.

Innhold

Forord . . . ii
Sammendrag . . . iii
Abstract . . . iv

1 Introduksjon **1**
1.1 Motivasjon . . . 1
1.2 Metode . . . 2
1.3 Mål . . . 2
1.4 Oppbygging . . . 3

2 Teori **4**
2.1 PET/CT . . . 4
2.2 Radiomics . . . 5
2.3 Maskinlæring . . . 5
2.3.1 Klassifiseringsalgoritmer . . . 5
2.3.2 Ytelsesmetrikker og validering . . . 7
2.4 Datavisualisering . . . 10
2.4.1 Prinsipalkomponentanalyse (PCA) . . . 10
2.4.2 Korrelasjon . . . 11

3 Materiale og metode **12**
3.1 Datasettet . . . 12
3.2 Programvare . . . 15
3.3 Preprosessering . . . 15
3.4 Modellering og validering . . . 17
3.4.1 Kryssvalidering . . . 17
3.4.2 Grid search-kryssvalidering . . . 18

4 Resultater **20**
4.1 Forhåndsanalyse av datasett . . . 20
4.2 Ytelsesresultater med sykdomsfri overlevelse (DFS) som responsvariabel . . . 22
4.2.1 Forhåndstest av modell . . . 23
4.2.2 Utfallsprediksjon for pasienter med positiv HPV-status . . . 24
4.2.3 Utfallsprediksjon for pasienter med negativ HPV-status . . . 29
4.3 Ytelsesresultater med generell overlevelse (OS) som responsvariabel 34
4.3.1 Utfallsprediksjon for pasienter med positiv HPV-status . . . 34

4.3.2 Utfallsprediksjon for pasienter med negativ HPV-status . . . 39
4.4 Viktigheten av ulike egenskaper . . . 44
4.4.1 Viktige egenskaper der DFS er brukt som responsvariabel . 45
4.4.2 Viktige egenskaper der OS er brukt som responsvariabel . . 46

5 Diskusjon **49**
5.1 Datasettet . . . 49
5.2 Evaluering av resultater . . . 50
5.2.1 Sykdomsfri overlevelse (DFS) som respons . . . 50
5.2.2 Generell overlevelse (OS) som respons . . . 53
5.2.3 Mulige forklaringer på modellytelsen . . . 56
5.3 Videre arbeid . . . 58

6 Konklusjon **59**

Bibliografi **64**

A Parametere **65**
A.1 Input parametere for grid search-kryssvalidering . . . 65
A.1.1 Logistisk regresjon klassifisering . . . 65
A.1.2 Random Forest klassifisering . . . 65
A.1.3 Decision tree klassifisering . . . 66
A.1.4 KNN klassifisering . . . 66
A.2 Output for grid search-kryssvalidering med DFS som respons . . . 66
A.2.1 Logistisk regresjon klassifisering . . . 67
A.2.2 Random Forest klassifisering . . . 67
A.2.3 Decision tree klassifisering . . . 68
A.2.4 KNN klassifisering . . . 68
A.3 Output for grid search-kryssvalidering med OS som respons . . . 69
A.3.1 Logistisk regresjon klassifisering . . . 69
A.3.2 Random Forest klassifisering . . . 69
A.3.3 Decision tree klassifisering . . . 70
A.3.4 KNN klassifisering . . . 70

B Resultater per fold **72**
B.0.1 DFS per fold . . . 73
B.0.2 OS per fold . . . 79

Figurer

2.1 Eksempel på et decision tree 7
2.2 Confusion matrix for visualisering av predikerte sanne og falske . . 8
2.3 Illustrasjon av ROC AUC kurve 10

3.1 Illustrasjon av en 5-foldet kryssvalideringsprosess 18

4.1 PCA-plott sentrert 20
4.2 PCA-plott standardisert 21
4.3 Korrelasjon egenskaper 22
4.4 De aggregerte ytelsesresultatene for D1, D2 og D3 23
4.5 DFS - De aggregerte ytelsesmålingene for DH1 for hver klassifiseringsalgoritme 26
4.6 DFS - De aggregerte ytelsesmålingene for DH2 for hver klassifiseringsalgoritme 26
4.7 DFS - De aggregerte ytelsesmålingene for DH3 for hver klassifiseringsalgoritme 26
4.8 DFS - DH1 per fold resultat 27
4.9 DFS - DH2 per fold resultat 28
4.10 DFS - DH3 per fold resultat 29
4.11 DFS - De aggregerte ytelsesmålingene for DU1 for hver klassifiseringsalgoritme 31
4.12 DFS - De aggregerte ytelsesmålingene for DU2 for hver klassifiseringsalgoritme 31
4.13 DFS - De aggregerte ytelsesmålingene for DU3 for hver klassifiseringsalgoritme 31
4.14 DFS - DU1 per fold resultat 32
4.15 DFS - DU2 per fold resultat 33
4.16 DFS - DU3 per fold resultat 34
4.17 OS - De aggregerte ytelsesmålingene for DH1 for hver klassifiseringsalgoritme 36
4.18 OS - De aggregerte ytelsesmålingene for DH2 for hver klassifiseringsalgoritme 36
4.19 OS - De aggregerte ytelsesmålingene for DH3 for hver klassifiseringsalgoritme 36
4.20 OS - DH1 per fold resultat 37
4.21 OS - DH2 per fold resultat 38
4.22 OS - DH3 per fold resultat 39

4.23 OS - De aggregerte ytelsesmålingene for DU1 for hver klassifiseringsalgoritme . . . 41
4.24 OS - De aggregerte ytelsesmålingene for DU2 for hver klassifiseringsalgoritme . . . 41
4.25 OS - De aggregerte ytelsesmålingene for DU3 for hver klassifiseringsalgoritme . . . 41
4.26 OS - DU1 per fold resultat . . . 42
4.27 OS - DU2 per fold resultat . . . 43
4.28 OS - DU3 per fold resultat . . . 44
4.29 Feature importance for random forest med DFS som respons på det kliniske datasettet DH1 . . . 45
4.30 Feature importance for random forest med DFS som respons på det kliniske datasettet DU1 . . . 45
4.31 Feature importance for random forest med DFS som respons på kombinasjonsdatasettet DH3 . . . 46
4.32 Feature importance for random forest med DFS som respons på kombinasjonsdatasettet DU3 . . . 46
4.33 Feature importance for random forest med OS som respons på det kliniske datasettet DH1 . . . 47
4.34 Feature importance for random forest med OS som respons på det kliniske datasettet DU1 . . . 47
4.35 Feature importance for random forest med OS som respons på kombinasjonsdatasettet DH3 . . . 48
4.36 Feature importance for random forest med OS som respons på kombinasjonsdatasettet DU3 . . . 48

Tabeller

3.1 Klassefordeling for de ulike responsvariablene DFS og OS 13
3.2 Kliniske egenskaper i D1 13
3.3 Beskrivelse D1,D2 og D3 15
3.4 Inndeling med/uten HPV, ulike distribusjoner i nye sett 16
3.5 Fordeling av pasienter per fold under kryssvalidering og grid search 19

4.1 De aggregerte ytelsesresultatene for D1, D2 og D3 24
4.2 DFS - Aggregerte ytelsesresultater fra klassifiseringsalgoritmene på DH1, DH2, DH3 25
4.3 DFS - Aggregerte ytelsesresultater fra klassifiseringsalgoritmene på DU1, DU2, DU3 30
4.4 OS - Aggregerte ytelsesresultater fra klassifiseringsalgoritmene på DH1, DH2, DH3 35
4.5 OS - Aggregerte ytelsesresultater fra klassifiseringsalgoritmene på DU1, DU2, DU3 40

A.1 Input hyperparameter intervall logistisk regresjon 65
A.2 Input hyperparameter intervall Random forest 66
A.3 Input hyperparameter intervall decision tree 66
A.4 Input hyperparameter intervall KNN 66
A.5 DFS - Valgte hyperparametere logistisk regresjon, med HPV 67
A.6 DFS - Valgte hyperparametere logistisk regresjon, uten HPV 67
A.7 DFS - Valgte hyperparametere Random forest, med HPV 67
A.8 DFS - Valgte hyperparametere Random forest, uten HPV 67
A.9 DFS - Valgte hyperparametere decision tree, med HPV 68
A.10 DFS - Valgte hyperparametere decision tree, uten HPV 68
A.11 DFS - Valgte hyperparametere KNN, med HPV 68
A.12 DFS - Valgte hyperparametere KNN, uten HPV 68
A.13 OS - Valgte hyperparametere logistisk regresjon, med HPV 69
A.14 OS - Valgte hyperparametere logistisk regresjon, uten HPV 69
A.15 OS - Valgte hyperparametere Random forest, med HPV 69
A.16 OS - Valgte hyperparametere Random forest, uten HPV 70
A.17 OS - Valgte hyperparametere decision tree, med HPV 70
A.18 OS - Valgte hyperparametere decision tree, uten HPV 70
A.19 OS - Valgte hyperparametere KNN, med HPV 70
A.20 OS - Valgte hyperparametere KNN, uten HPV 71

B.1 DFS - per fold resultat for de fire modellene for DH1 73
B.2 DFS - per fold resultat for de fire modellene for DH2 74
B.3 DFS - per fold resultat for de fire modellene for DH3 75
B.4 DFS - per fold resultat for de fire modellene for DU1 76
B.5 DFS - per fold resultat for de fire modellene for DU2 77
B.6 DFS - per fold resultat for de fire modellene for DU3 78
B.7 OS - per fold resultat for de fire modellene for DH1 79
B.8 OS - per fold resultat for de fire modellene for DH2 80
B.9 OS - per fold resultat for de fire modellene for DH3 81
B.10 OS - per fold resultat for de fire modellene for DU1 82
B.11 OS - per fold resultat for de fire modellene for DU2 83
B.12 OS - per fold resultat for de fire modellene for DU3 84

Forkortelser

Forkortelse	Definisjon
AUC	Area under curve
CT	Computertomografi
DFS	Disease-free survival (Sykdomsfri overlevelse)
FN	Falsk negativ (eng. false negative)
FP	Falsk positiv (eng. false positive)
FPR	Falsk positiv rate (eng. false positive rate)
HPV	Humant papillomavirus
KNN	K-nærmeste naboer
MCC	Matthews Correlation Coefficient
ML	Maskinlæring
MTV	Metabolsk tumorvolum
NaN-verdi	Manglende verdi (eng. Not a Number-value)
OS	Overall survival (Generell overlevelse)
OUS	Oslo universitetssykehus
PCA	Principial Component Analysis (Prinsipalkomponentanalyse)
PET	Positron emisjons tomografi
ROC	Receiver operator characteristics
SUV	Standard uptake value
TLG	Total lesion glycosis
TN	Sann negativ (eng. true negative)
TP	Sann positiv (eng. true positive)
TPR	Sann positiv rate (eng. true positive rate)

Kapittel 1

Introduksjon

1.1 Motivasjon

Kreft er samlenavnet på sykdommer som følger av tilfeller hvor det oppstår mutasjoner i cellers arvestoff, som så fører til en ukontrollert celledeling som etter en videre delingsprosess utvikler seg til en kreftsvulst [1].

Det finnes mange ulike typer kreft, og valg av behandlingsform vil av den grunn variere. Diagnostiseringen av kreft stilles blant annet ved hjelp av celle- og vevsprøver, samt for eksempel PET- og CT-bilder, og prognosene for overlevelse vil variere ut fra hva slags type kreft som blir avdekket [1], [2].

På årsbasis får omtrent 38 000 mennesker i Norge kreft, og tallene forventes å øke frem mot 2040, med årsaker som befolkningsvekst og økt levealder som en del av det oppgitte grunnlaget for økningen [3]. For hode- og halskreft er det på årsbasis registrert omtrent 800 tilfeller [4]. I norsk sammenheng er kreft fortsatt den ledende dødsårsaken i samfunnet, etterfulgt av dødsfall relatert til hjerte- og karsykdommer samt covid-19 [5].

Ondartede svulster (kreft) i «nese og bihuler, leppe, munnhule, svelg, strupehode eller spyttkjertler» faller inn under fellesbetegnelsen hode- og halskreft [6]. Symptomene på slik sykdom vil vise seg i forskjellig grad ut fra hva slags kreft det dreier seg om, men vil uansett normalt ikke spre seg til andre områder av kroppen [7]. Hvor mange som får hode- og halskreft har steget noe de siste årene, og forskning peker på at dette delvis er på grunn av kreft som følger av infeksjon med viruset HPV (Humant Papillomavirus), som man har sett en økt forekomst av i det norske samfunnet [7], [8].

Når det gjelder behandlingen av kreft, består de mest brukte behandlingsformene av operasjon, cellegift og stråleterapi [1]. Disse behandlingsformene vil enten brukes i kombinasjon eller alene [6]. Formålet med behandling er å stoppe opp den ukontrollerte celledelingen som skaper kreften, og dermed oppnå enten helbredelse eller i det minste stabilisering eller forsinkelse av sykdommen [9]. Ved kreftbehandling kan man imidlertid stå i fare for å få senskader [1], hvorav såkalt sekundær

kreft er den mest alvorlige [10]. Dette skyldes det faktum at både cellegift og stråleterapi, ved siden av å være kreftbekjempende, også kan ha en kreftfremkallende virkning [10]. Når man behandler kreftceller med cellegift og stråleterapi, ønsker man derfor å unngå å skade omliggende friskt vev, og normalt vil friske celler tåle slik behandling uten å få senskader [11], [12]. Sekundær kreft kan for øvrig også skyldes andre faktorer enn behandlingen man mottok for den første kreftformen [10].

I det tilfellet det skulle la seg gjøre å predikere utfallet av behandling for hode- og halskreft hos den enkelte pasient, og følgelig skreddersy behandlingen etter det enkelte individs behov og forutsetninger, kan man se for seg at mulige senskader ville kunne unngås ved å skjære vekk ineffektiv eller overflødig behandling fra pasientens behandlingsplan. I et slikt tilfelle ville i så fall målet være å optimalisere muligheten for å bedre prognoser, og samtidig i større grad beskytte pasienten mot mulige senskader.

1.2 Metode

Det er for denne oppgaven tatt i bruk datasett bestående av helsedata fra 238 hode- og halskreftpasienter. For å kunne analysere og gjøre prediksjoner for disse datasettene er det blitt benyttet maskinlæring. Maskinlæring ligger innenfor feltet kunstig intelligens, og kan brukes til å lage modeller som predikerer utfall ut i fra data. Innenfor det medisinske feltet er det et stort potensial for bruk av maskinlæring, ikke bare innenfor forskning, men for diagnostisering og prediksjon av ulike behandlingsutfall [13]. Ved å for eksempel se på kliniske egenskaper og radiomics-egenskaper [14] for svulsten til en kreftpasient, kan det da muligens predikeres et utfall av behandlingen.

Radiomics-egenskaper er blitt trukket ut fra medisinske bilder, som PET- og CT-bilder [15]. Disse egenskapene inneholder blant annet informasjon om svulstens form, størrelse, tekstur og ulike mønstre den kan ha [14]. Ønsket er at disse egenskapene, sammen med kliniske egenskaper, skal kunne forbedre ytelsen til prediksjonsmodellene [14].

1.3 Mål

Formålet med denne oppgaven er å se hvordan ulike maskinlæringsmodeller evner å predikere behandlingsutfallet til pasienter med hode- og halskreft, basert på kliniske egenskaper og radiomics-egenskaper hentet ut fra PET- og CT-bilder. Det blir sett på to forskjellige behandlingsutfall, sykdoms fri overlevelse og generell overlevelse. Pasientene er delt inn i to grupper basert på deres HPV-status, det vil si at pasienter med positiv HPV-status utgjør en gruppe, og pasienter med negativ HPV-status utgjør den andre. Målet er å finne ut om det er noen forskjell i modellytelsen for disse to pasientgruppene, og i så fall hva som kan være grunnen(e) til dette.

1.4 Oppbygging

Kapittel 2 i denne oppgaven tar for seg teoridelen, der teorien for konseptene brukt til å kunne fremstille resultatene blir presentert. Kapittel 3 omhandler materialer og metoder benyttet for å kunne behandle datasettene og videre modellere dem ved bruk av ulike maskinlæringsalgoritmer. Videre vil det i kapittel 4 presenteres resultater fra prediksjonene. Deretter vil resultatene og videre arbeid bli diskutert i kapittel 5. Som et siste punkt, vil det i kapittel 6 fremlegges en konklusjon for oppgaven.

Kapittel 2

Teori

2.1 PET/CT

Som nevnt innledningsvis kan medisinske bilder, som PET- og CT-bilder, brukes til for eksempel diagnostisering av kreft. PET og CT gir informative innblikk i kroppens funksjon og anatomi, og er effektive hjelpemidler til å oppdage, diagnostisere og vurdere behandlingen for ulike sykdommer [16].

Positronemisjonstomografi (PET)

En PET-scan (positron-emisjonstomografi) er en scan som tar fysiologiske bilder med formål om å avdekke oppsamling av celler i høy aktivitet i organer og vev ved hjelp av et radioaktivt sporstoff [17], [16]. Sporstoffet brukt for de medisinske bildene som egenskapene i denne oppgaven er hentet fra, var 2-deoxy-2-[F-18]fluoro-Dglucose (FDG) [18]. PET er en «nukleærmedisinsk undersøkelsesmetode» [19] som søker å avdekke eventuell «stoffskifteaktivitet (metabolisme) i kroppsvev av positronutskillende (emitterende) radioaktive preparat» [19]. PET-scans evner ofte å avdekke tegn på sykdom før sykdommen lar seg avdekke på andre bilder, fra for eksempel CT-scans (computertomografi) og MRI, men kombineres likevel ofte med disse [17].

Det radioaktive sporstoffet injiseres i pasienten, og vil bevege seg mot områder med forhøyde nivåer av metabolsk eller biokjemisk aktivitet og vil derfor kunne gi informasjon om hvor sykdommen sitter i kroppen [17].

Computertomografi (CT)

Computertomografi (CT), er som PET en medisinsk bildeteknikk, og for CT blir det ved bruk av røntgenstråler avbildet et tverrsnitt av de områdene i kroppen som skal undersøkes [20]. En CT-scan blir tatt ved bruk av en stor maskin, formet som en ring, der pasienten befinner seg inne i denne «trommelen» [20]. Det blir sendt ut røntgenstråler fra forskjellige vinkler, da denne ringen roteres rundt kroppen til

pasienten [21]. Det er detektorer på motsatt side som vil fange opp disse strålene, og det detektorene fanger opp vil bli overført og behandlet av en datamaskin.

Ved bruk av CT-scans får man muligheten til å lage et tredimensjonalt bilde av det akutelle området. CT-scans er mer effektive i avdekking og lokalisering av for eksempel svulster ettersom de evner å presentere mer informasjon enn et ordinært røntgenbilde [21].

2.2 Radiomics

Radiomics er en metode brukt til å trekke ut egenskaper og informasjon fra medisinske bilder, eksempelvis om egenskapene til en svulst [15]. Egenskapene hentet ved radiomics kan gi opplysninger om karakteristikker som størrelse og form på den eksempelvise svulsten; med andre ord kan den gi informasjon som er viktig når man skal både stille og behandle kreftdiagnoser [22]. Mer om radiomics og denne prossesen er beskrevet i A. Zwanenburg et al. [15] og Huynh et al. [14].

2.3 Maskinlæring

Maskinlæring (ML) ligger innenfor feltet kunstig intelligens, der det benyttes algoritmer som gjør at modeller kan lære fra ulike eksempler [23]. En enkel måte å forklare det på, er at man legger inn data og svar, og fra dette får ut regler [23]. Det blir i denne oppgaven brukt maskinlæring på datasett fra pasienter med hode- og halskreft, med mål om å predikere behandlingsutfall for pasientene.

Typer maskinlæring

De tre vanligste læringsteknikkene for maskinlæring kalles overvåket, ikke-overvåket og forsterkende læring [23]. Med overvåket læring blir modellen trent på data med korrekte svar kjent. Den skal da bli «lært opp» til å kunne kjenne igjen mønstre, og predikere på nye, usette data der svaret ikke er kjent. For ikke-overvåket læring får modellen data uten noen gitte korrekte svar, og det er modellens oppgave å oppdage mønstre eller karakteristikk i dataen [23]. Forsterkende læring tar beslutninger ut fra en rekke interaksjoner basert på omgivelser og miljø, og blir avhengig av utfallet av avgjørelsen, enten belønnet eller straffet [23]. Det er overvåket læringsalgoritmer som er benyttet i denne oppgaven.

2.3.1 Klassifiseringsalgoritmer

De fire forskjellige klassifiseringsalgoritmer brukt i denne oppgaven er logistisk regresjon, decision tree, random forest og K-nærmeste naboer (KNN).

Logistisk regresjon

Logistisk regresjon er, tross navnet, ikke en regresjonsmodell, men heller en lineær klassifiseringsalgoritme [24]. Lineære klassifiseringsalgoritmer er spesielt nyttig ved håndteringen av binære klassifiseringsproblemer, hvor man vil skille mellom to ulike klasser [23]. Logistisk regresjon modellerer sannsynligheten for at en bestemt prøve tilhører en klasse, og dette gjøres ved å først se på oddsratioen [23]. Dette er da oddsen til fordel for et bestemt utfall, og kan gis ved

$$\frac{p}{1-p}$$

der p representerer sannsynligheten for at det positive utfallet inntreffer [23]. Videre fra dette uttrykket, kan man se på logaritmen til oddsratioen, også kalt *logit function* [23],

$$logit(p) = ln(\frac{p}{1-p}) \tag{2.1}$$

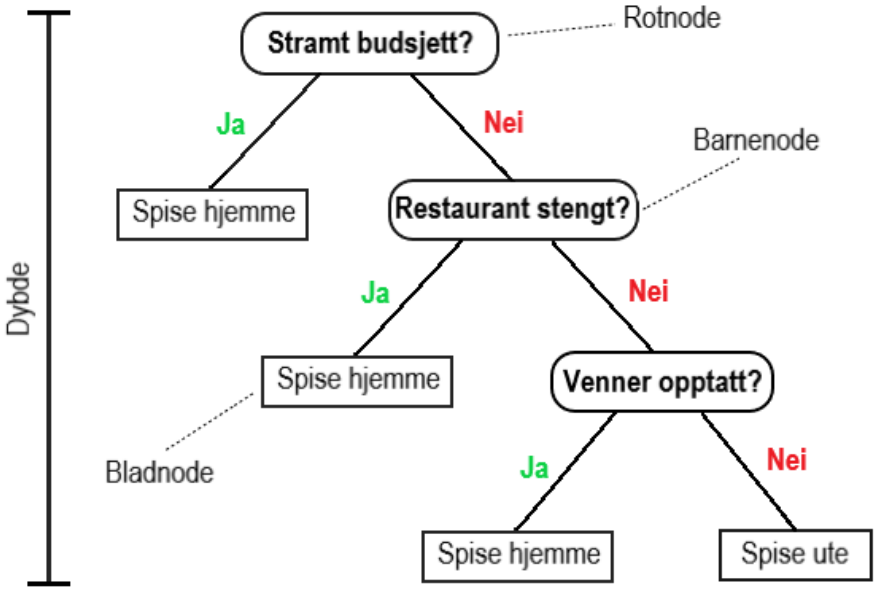
der *log* angir den naturlige logaritmen [23]. Det er verdier på mellom 0 og 1 blir brukt som input i logaritmefunksjonen, og videre blir disse konvertert til verdier som strekker seg over hele det reelle tallområdet. For å videre finne den predikerte sannsynligheten for at en gitt prøve tilhører den valgte klassen, benyttes den inverse formen av likning 2.1,

$$\Phi(z) = \frac{1}{1+e^{-z}} \tag{2.2}$$

der er z de forskjellige prøvene brukt som input for modellen [23]. Da resultatet fra likning 2.2 representerer sannsynligheten for at en prøve tilhører klasse 1, ligger verdien derfor mellom 0 og 1. Det er satt en terskelverdi, og om $\Phi(z)$ er over denne vil prøven tildeles klasse 1, og er den under vil prøven få klasse 0 [23].

Decision tree

Decision tree klassifiseringen (beslutningstre på norsk), er en modell som ved hjelp av valg basert på en rekke spørsmål bryter ned dataen matet inn [23]. Modellen har en trelignende struktur, bygd opp av flere noder, som hver representerer et spørsmål og valg, og bladnodene som er de mulige utfallene, altså hvilken klasse [23]. Ettersom decision trees har en struktur som gjør dem enklere å forstå, er de blant modellene som er enklest å tolke [23]. Decision tree trenes på treningsdata og lærer fra dette spørsmålene som skal stilles for å klassifisere prøven korrekt. Det vil si at jo dypere treet er, jo mer spesifikt trenes modellen på treningsdataen og det kan øke faren for overtilpasning [23], men det er også mulig å sette en grense for dybden. I figur 2.1 er det et enkelt eksempel på et decision tree, der det skal tas en beslutning på hvorvidt man skal spise inne eller ute.



Figur 2.1: *Eksempel på et decision tree. Personen som følger enten ja eller nei, er her prøven, og klassene er delt i hvorvidt det skal spises hjemme eller ute.*

Random forest

Random forest-algoritmen anvender en mengde beslutningstrær (decision trees) for å utføre regresjon og klassifisering. Det er en ensemble klassifisering, som vil si at det brukes flere klassifiseringer i en [23]. Siden decision trees blir sett på som svake algoritmer for varianse, er ideen å samle flere av disse for å skape en mer robust modell [23]. Hvert decision tree trenes på ulike og vilkårlige deler av treningssettet, og random forest gjør klassifiseringer basert på hva flertallet av decision trees mener [23]. Random forest tilpasser seg bedre ny data, og faren for overtilpasing blir også redusert [23].

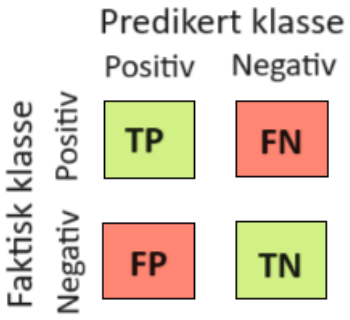
K-nærmeste naboer

K-nærmeste naboer (KNN) faller innenfor det som kalles for en lat klassifiseringsalgoritme [23]. Dette innebærer at den memorerer treningsdata, og deler ut klasser til den gitte testdata basert på hva den husker. Dette skiller seg fra en del andre klassifiseringsalgoritmer, som i stedet lærer seg en funksjon fra treningsdataen. KNN predikerer nye prøver etter den klassen som har et flertall blant sine nærmeste naboer [23].

2.3.2 Ytelsesmetrikker og validering

En essensiell fase for ML-modeller er validering av deres ytelse, noe som er viktig for å sikre pålitelige og presise resultater [23]. Noen ytelsesmetrikker som er vanlig å bruke for å måle en modells ytelse er nøyaktighet, F1-score, Matthews korrelasjonskoeffisient (MCC) og Receiver Operating Characteristics Area Under Curve (ROC AUC) som vil bli beskrevet bedre under.

Noe metrikkene nevnt over har til felles er at de alle beregnes ut fra predikerte resultater, som er delt inn i fire kategorier. Kategoriene er; sann positiv (TP) og sann negativ (TN) som er antallet sanne positive og negative klassifiseringer, og falsk positiv (FP) og falsk negativ (FN) som er antallet feilaktige positive og negative klassifiseringer. I figur 2.2 illustreres de predikerte og faktiske klassene i en forvirringsmatrise (confusion matrix).



Figur 2.2: *En confusion matrix, på norsk forvirringsmatrise, visualiserer predikerte sanne og falske resultater. Inspirert fra [23].*

Nøyaktighet

Ytelsesmetrikken nøyaktighet (eng. accuracy) gir informasjon om hvor mange korrekte klassifiseringer modellen har gjort fra datasettet [23]. Nøyaktigheten beregnes ved å ta summen av alle korrekte prediksjoner dividert med totalt antall prediksjoner [23]. Dette gir uttrykket:

$$accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{2.3}$$

der TP, TN, FP og FN er de sanne positive og negative, og de falske positive og negative prediksjonene.

F1-score

F1-score brukes til å evaluere en klassifiseringsmodells ytelse, og beregnes ofte ut fra en kombinasjon av metrikkene *precision* (PRE) og *recall* (REC) [23]. Presisjonen viser til den andelen av positive prediksjoner som er korrekt predikert sanne positive, mens tilbakekalling er andelen sanne positive som modellen predikerte riktig. Disse metrikkene kan uttrykkes slik:

$$PRE = \frac{TP}{TP + FP} \tag{2.4}$$

og

$$REC = \frac{TP}{FN + TP} \tag{2.5}$$

som videre uttrykker F1-score som:

$$F1 = 2\frac{PRE \times REC}{PRE + REC} \tag{2.6}$$

Ved å ta hensyn til og kombinere PRE og REC, gir dette en mer stabil F1-score metrikk [23].

F1-score beregnes med fokus på den positive klassen, altså klasse 1 (F1:1), og hvor presist den negative klassen, klasse 0 (F1:0), predikeres er vanskelig å si. Ved å bytte om på klassene og responsvariabel, altså at positiv blir negativ og motsatt, kan det med samme metode beregnes F1-score for negativ klasse.

Selv om nøyaktighet og F1-score er populære å bruke i binære klassifiseringer, er det viktig å huske at dersom klassedistribusjonen er i ubalanse, altså at en klasse har mye høyere antall prøver enn den andre, kan dette gi overoptimistiske resultater for majoritetsklassen [25].

Matthews korrelasjonskoeffisient

Matthews korrelasjonskoeffisient (MCC) blir også brukt til vurdere kvaliteten til klassifiseringer [26]. Ulikt de to metrikkene over tar MCC hensyn til både TP, TN, FP og FN, dette vil si at skal MCC kunne gi en høy score må prediksjonen oppnå gode resultater innenfor alle de fire nevnte kategoriene [25]. MCC beregnes av følgende formel:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{2.7}$$

Verdiene MCC gir ligger mellom -1 og 1, der -1 er den verste og indikerer en totalt feil prediksjon, mens 1 i andre enden er best, altså en perfekt prediksjon. Er MCC lik 0, indikerer det en helt tilfedig prediksjon, som et myntkast [25].

Area Under the Receiver Operating Characteristic Curve

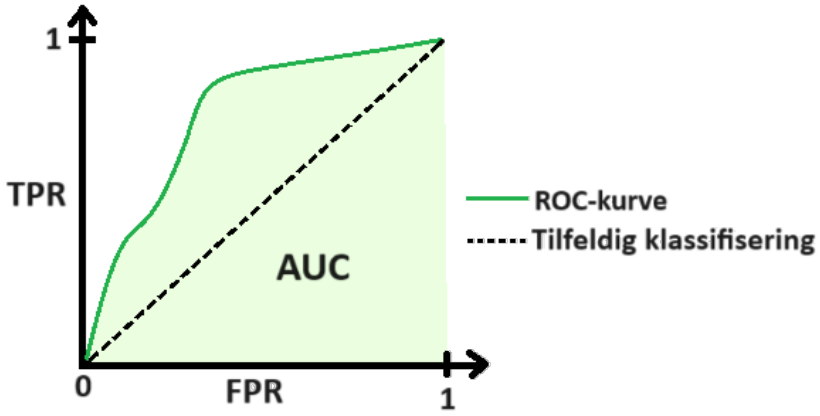
Receiver Operating Characteristics (ROC) er en graf som kan plottes med hensyn til den falske positive rate (FPR) og sanne positive rate (TPR) [23], disse er uttrykt ved:

$$FPR = \frac{FP}{FP + TN} \tag{2.8}$$

og

$$TPR = \frac{TP}{FN + TP} \tag{2.9}$$

En ROC-kurve langs den stiplet diagonale kurven i figur 2.3, vil si at modellen gjetter tilfeldig for klassifiseringen [23]. Hvis det er et punkt med tilfeldig gjetning, vil det si at FPR og TPR er like, og da vil det innenfor klasse 0 være like mange feilklassifiserte prøver som det er korrekt klassifiserte prøver i klasse 1. [23]. Kurver som ligger over den diagonale kurven tilsvarer at modellen yter bedre enn tilfeldig gjetning, for alle kurver under er det motsatt og tilsvarer en verre ytelse [23].



Figur 2.3: *En illustrasjon av ROC AUC, der den grønne linjen er en tilfeldig ROC-kurve og AUC er arealet mellom denne kurven og x-aksen. Den diagonale stiplet linjen tilsvarer en tilfeldig gjetning i klassifiseringen.*

Area under curve (AUC) er arealet som er under ROC-kurven [23], som vist i illustrasjon i figur 2.3. AUC har et intervall på 0 til 1, der en AUC verdi på 1 indikerer at alle predikerte utfall er korrekte [23]. AUC-ytelsen viser til modellens kapasitet til å separere de ulike klassene.

2.4 Datavisualisering

2.4.1 Prinsipalkomponentanalyse (PCA)

PCA er forkortelse fra det engelske «Principal Component Analyses», og på norsk blir det kalt prinsipalkomponentanalyse eller hovedkomponentanalyse. PCA blir brukt til å forenkle dimensjonaliteten i et datasett [23]. Datasettet blir transformert slik at det blir nye egenskaper som ikke korrelerer, og disse blir kalt prinsipalkomponenter. Komponentene vil være lineære sammensetninger av de opprinnelige egenskapene som resulterer i maks varians. Prinsipalkomponentene beskriver varians, og den første prinsipalkomponenten (PC1) er den med mest varians [23], mens de neste komponentene etter den første, er de med høyest varians og samtidig ukorrelert til den før [23].

PCA er et viktig verktøy for å analysere og utforske datasett før det brukes i ML-modeller, og er et verktøy i å avdekke ulike mønstre fra data basert på korrelasjonen mellom egenskapene [23]. Det blir ofte brukt score- og loadingplott til å visualisere datasettet. Et scoreplott viser prinsipalkomponentene i et to- eller tre-dimensjonalt koordinatsystem, der i et to-dimensjonalt plott vil PC1 være på x-aksen og PC2 langs y-aksen. Prøvene fra datasettet blir vist som punkter i dette systemet, og hvordan de plasseres vil gi et bilde av ulike mønstre eller grupperinger for prøvene i datasettet. Et loadingplottet i to-dimensjon kan vise hvordan de opprinnelige egenskapene korrelerer med hverandre, og hvor mye de bidrar til PC1 og PC2.

2.4.2 Korrelasjon

Korrelasjon blir brukt for å se etter sammenhenger mellom par av egenskaper [27]. Det er fint å bruke for å se hvilke egenskaper som korrelerer godt og dårlig med hverandre. Siden egenskaper med høy korrelasjon har mye av den samme effekten på den avhengige egenskapen, kan det å ta vekk en av egenskapene i paret være lurt, da det å bruke begge kan skape unøyaktighet for modellen [27]. Ved større datasett er det også greit å ta vekk noen høyt korrelerte egenskaper, da det vil føre til at kjøretiden til ML-modellen reduseres.

I denne oppgaven blir Python pakkene matplotlib [28] og seaborn [29] brukt til å presentere et heatmap for egenskapkorrelasjonene. I et heatmap blir korrelasjonene mellom egenskapene delt opp i celler, som består av en farge og ofte også et tall (korrelasjonskoeffisienten). Denne fargen viser styrken til korrelasjonen og om den er positiv eller negativ. En positiv korrelasjon vil si at når en egenskap øker, er det sannsynlig at den andre også øker. For negativ korrelasjon blir det da motsatt for den andre egenskapen, så når en øker, vil den andre sannsynligvis avta [27]. Korrelasjonskoeffisienten ligger på en skala fra -1 gjennom 0 og til +1, der -1 er sterk negativ korrelasjon og +1 er sterk positiv.

Kapittel 3

Materiale og metode

I denne oppgaven blir modellytelsen til fire klassifiseringsalgortitmer undersøkt, på flere ulike datasett, og materialet og fremgangssmåten benyttet til dette blir tatt for seg i dette kapittelet.

3.1 Datasettet

Datasettene brukt i denne oppgaven består av pasientdata fra pasienter behandlet for hode- og halskreft ved Oslo universitetssykehus (OUS) og Maastricht University Medical Center (MAASTRO), henholdsvis i perioden 2007 til 2013 for pasienter ved OUS og 2008-2014 ved MAASTRO. Alle pasienter i datasettene gjennomgikk strålebehandling, og noen også cellegiftbehandling [14].

I denne oppgaven er tre datasett tatt i bruk, og disse består av et datasett med kliniske egenskaper i tillegg til tre PET parametere, som i denne oppgaven blir kalt D1. Det andre er et datasett med radiomics egenskaper og verdier trukket ut fra PET- og CT-bilder og refereres til som D2. Det tredje og siste settet er kombinasjonen av D1 og D2 sammen og blir da kalt D3. Datasettene er slått sammen av data fra OUS med 139 pasienter og MAASTRO med 99 pasienter, det var opprinnelig flere pasienter i datasettene men som av ulike grunner ikke oppfylte gitte kriterier er antallet blitt redusert. En begrunnelse av dette er beskrevet i Moan et al. [18] og Huynh et al. [14]. Alle tabeller og oversikter under vil vise OUS og MAASTRO sammenslått.

Pasientene ble kategorisert i to klasser, klasse 0 og 1, basert på to ulike responsvariabler. Responsvariablene er sykdomsfri overlevelse (DFS event fra Disease Free Survival) og generell overlevelse (OS event fra Overall Survival). For DFS responsen vil pasienter med sykdomsfri overlevelse være klasse 0, og død og lokalt, regionalt eller metastatisk tilbakefall vil være klasse 1. I OS responsen er en pasients generelle overlevelse (sykdomsfri eller ikke) tildelt klasse 0, mens død er klasse 1. En oversikt over responsene og distribusjonen av klassene finnes i tabell 3.1. En ekstra detalj for datasettet er at pasienter registert døde, døde ikke

nødvendigvis av kreft, men dette er ikke spesifisert i datasettet.

Tabell 3.1: *Klassefordeling av 139 pasienter for de ulike responsvariablene DFS og OS, der DFS referer til sykdomsfri overlevelse (Disease-free survival) og OS til generell overlevelse (Overall survival).*

Responsvariabel	Beskrivelse	Fordeling
DFS	Sykdomsfri (klasse 0)	Klasse 0: 46,6%
	Død og lokalt, regionalt eller metastatisk tilbakefall (klasse 1)	Klasse 1: 53,4%
OS	Overlevelse (klasse 0)	Klasse 0: 53,8%
	Død (klasse 1)	Klasse 1: 46,2%

Klinisk data

Den kliniske dataen (D1) består av 238 pasienter og 14 egenskaper (sju faktorer + fire svulstplasseringer + tre PET parametere). Disse PET parameterne kan være viktige for prediksjonen, da det i følge flere studier har en betydelig assosiasjon til overlevelse [18]. I dataen er katergoriske variabler blitt gjort om til binære variabler. Dette vil si at f.eks. kjønn som her er delt opp i kvinne og mann, er blitt gjort om slik at egenskapen er female og klasse 1 i datasettet vil da bety at det er kvinne, og 0 viser til mann. En oversikt over kliniske egenskaper er vist i tabell 3.2.

Tabell 3.2: *Kliniske egenskaper i datasett 1 (D1). Egenskaper som er kontinuerlige vises med gjennomsnitt og standardavvik under fordeling.*

Egenskap	Beskrivelse	Fordeling
age	Alder i år	60,8 ± 9,5
female	Klasse 1: kvinne Klasse 0: mann	Kvinne: 24,4% Mann: 75,6%
cavum_oris	Er primærsvulst i munnhule? Klasse 1: Ja. Klasse 0: Nei	Munnhule: 5,9%
oropharynx	Er primærsvulst i munnsvelg? Klasse 1: Ja. Klasse 0: Nei	Munnsvelg: 56,7%
hypopharynx	Er primærsvulst i strupesvelg? Klasse 1: Ja. Klasse 0: Nei	Strupesvelg: 13,0%

larynx	Er primærsvulst i strupehodet? Klasse 1: Ja. Klasse 0: Nei	Strupehodet: 24,4%
histgrade_high	Histologisk grad (G) av svulstvev. Klasse 1: G3 Klasse 0: G1-G2	G3: 57,1% G1-G2: 42,9%
hpv_related	HPV status på pasienten. 1: positiv 0: negativ	Positiv: 42,9% Negativ: 57,1%
charlson	Charlson Comorbidity Index Klasse 1: Index 1-6 Klasse 0: Index 0	1-6: 53,4% 0: 46,6%
pack_years	År med røyking av minst 20 sigaretter per dag	33,8 ± 36,6
uicc8_III-IV	Pasientens kreftstadie. Klasse 1: stadie 3-4 Klasse 0: stadie 1-2	1: 61,8% 0: 38,2%
PET parametere		
SUVpeak	Høyeste snittverdi for SUV for en 1 cm^3 sfære, med senteret innenfor primærsvulsten.	11,1 ± 5,8
MTV	Metabolsk tumorvolum	13,2 cm^3 ± 13,0 cm^3
TLG	Total lesion glycolysis	116,4 cm^3 ± 163,5 cm^3

Radiomics datasett

I D2 er det 374 bilde-baserte egenskaper. Disse egenskapene er blitt trukket ut fra medisinske bilder, PET- og CT-bilder, med metoden radiomics. Denne prossesen kan, som nevnt under seksjon 2.2, leses om i Zwanenburg et. al [15] og [30]. Egenskapene hentet fra bildene består av førsteordens statistiske egenskaper, form, tekstur og lokale binære mønster fra primærsvulsten [14]. Det siste datasettet D3, er da satt sammen av D1 og D2. En oversikt over disse finnes i tabell 3.3.

Tabell 3.3: *De tre sammenslåtte datasettene fra OUS og MAASTRO, med beskrivelse og antall egenskaper.*

Datasett	Beskrivelse	Egenskaper
D1	Klinisk data	14
D2	Egenskaper trukket ut fra medisinske bilder	374
D3	D1+D2	388

3.2 Programvare

Python versjon 3.11.4 [31] ble brukt som programmeringsspråk i denne oppgaven. Videre ble Jupyter Notebook med versjon 6.5.3 og Spyder med versjon 5.4.3 fra Anaconda [32] benyttet som programmeringsverktøy.

Til databehandlingen ble det benyttet NumPy [33] og Pandas [34], og klassifiseringsalgoritmene og ytelsesmetrikkene brukt i oppgaven ble implementert ved bruk av maskinlæringsbiblioteket Scikit-learn [35]. Visualisering av data ble utført ved hjelp av matplotlib [28], seaborn [29], hoggorm [36], hoggormplot [37] og Microsoft Excel [38].

3.3 Preprosessering

Det ble gjort en preprosessering og analyse av de kliniske dataene før modelleringen startet, slik at kvaliteten kunne forbedres og eventuelle feil/mangler ble utelukket. Dette er viktig for at klassifiseringen og prediksjonen skal kunne yte maks.

Sammenslåing av datasett

Det ble startet med seks datasett; to kliniske, to bilde-datasett fra radiomics og to sammenslåtte sett bestående av klinisk + bildedata, der ett fra hvert av settene kom fra OUS og MAASTRO. Hver pasient i datasettene hadde sin egen pasient-id, men siden noen fra OUS og MAASTRO var like, ble oppdatering av disse det første steget i preprosesseringen. De nye id-ene ble også endret for responsvariablene tilhørende OUS og MAASTRO. Datasettene ble videre slått sammen til D1, D2 og D3 som nevnt tidligere i oppgaven.

Manglende verdier

Det første som ble gjort etter sammenslåing var å sjekke for og eventuelt fjerne såkalte NaN-verdier eller missing numbers, som betyr at det er manglende verdier i

datasettet. En manglende verdi kan påvirke presisjonen til en modell, eller i verste fall hindre den i å virke [23].

Bildedataen, D2, som ble brukt i denne oppgaven var allerede klargjort, så egenskapsuthenting fra PET- og CT-bilder var ikke nødvendig (hvordan uthenting ble utført kan leses i Huynh et al. [14]). Dermed ble det på dette settet bare gjort en sjekk etter manglende verdier, for å være sikker.

Visualisering

Det ble utført PCA og laget score- og loadingplott og forvirringsmatrise for D1, da det er mange numeriske verdier å forholde seg til og vanskelig å finne ekstremverider (eng. outliere) eller se sammenhenger ved å bare se på dataen. Ved visualisering ble det sett på ulike ekstremverdier, sammenhenger og korrelasjon i daten. Siden ekstremverdier kan ha negativ påvirkning på modellene som trenes, og muligens bidra til redusert ytelse i prediksjon eller klassifisering, er det viktig å se på dette. Det ble tatt en vurdering, og det ble ikke fjernet noen pasienter fra dataen grunnet ekstremverdier eller høy korrelasjon. Pakkene *hoggorm* [36] og *hoggorm-plot* [37] blir benyttet til utførelse av PCA-analyse, og fremstillingen dette i score- og loadingplott.

Deling av datasett med hensyn på HPV-relasjon

Siste klargjøring av datasettene var at de alle ble delt på egenskapen *hpv_related*. Så de nye datasettene som har pasienter med HPV vil bli referert til som DH1, DH2 og DH3, og består av 102 pasienter. Datasettene med pasientene uten HPV blir fra nå kalt DU1, DU2 og DU3, og har da 136 pasienter.

I tabell 3.4 er distribusjonen for *uicc8_III-IV* og DFS og OS, vist etter den nye inndelingen på HPV-relasjon. Det kan ses at det er en ubalanse i fordelingen, noe som vil ha en dårlig innvirkning på prediksjonen.

Tabell 3.4: *Tabellen viser inndeling av med/uten HPV i forhold til antall pasienter og de nye distribusjonene til kreftstadier og de to responsvariablene, DFS og OS. For uicc_8III-IV er det slik at klasse 0 tilsvarer stadie 1-2 og klasse 1 tilsvarer stadie: 3-4. Fordelingen er oppgitt i antall og prosent.*

Datasett	**uicc8_III-IV**	**DFS respons**	**OS respons**
DH	klasse 0: 80 (78,43%)	klasse 0: 71 (69,61%)	klasse 0: 80 (78,43%)
(102 pasienter)	klasse 1: 22 (21,57%)	klasse 1: 31 (30,39%)	klasse 1: 22 (21,57%)
DU	klasse 0: 11 (8,09%)	klasse 0: 40 (29,41%)	klasse 0: 48 (35,29%)
(136 pasienter)	klasse 1: 125 (91,91%)	klasse 1: 96 (70,59%)	klasse 1: 88 (64,71)%

3.4 Modellering og validering

Valg av algoritmer

Det ble først valgt logistic regression og random forest som klassifiseringsalgortimer både på grunn av at det er gode og enkle algortimer å bruke og for sammenligningsgrunnlag med resultater fra Huynh et al. [14], da disse var to av modellene brukt i den studien. Videre ble også decision tree og KNN valgt ut, da disse også er ganske enkle algoritmer og fine å sammenligne med logistisk regresjon og random rorest. Alle klassifiseringsalgortimene benyttet i denne oppgaven er pakker fra maskinlæringsbiblioteket scikit-learn [35].

Alle de fire klassifiseringsmodellene ble kjørt med de orginale datasettene, altså for D1, D2 og D3, med DFS som respons. Dette var for å teste modellene og se hvordan prediksjonsytelsen ble i forhold til resultatene fra Huynh et al. [14]. Da resultatene fra dette var tilfredsstillende, var modellene godkjent og klare.

For at man skal få et mer helhetlig bilde av modellytelsen, må det tas i bruk flere metrikker. De metrikkene som ble valgt ut til å evaluere prediksjonen var nøyaktighet, ROC_AUC, MCC og F1-score for klasse 1 og 0, som er beskrevet tidligere under seksjon 2.3.2 i teori-kapittelet. F1-score som tar hensyn til både presisjon og tilbakekalling, fokuserer da spesielt på klassebalansen. Dette kan bli en viktig metrikk for resultatene, da det i tabell 3.4 kan ses at klassene er skjevt fordelt. Det ble også tatt hensyn til sammenligning med Huynh et al. [14] ved valg av ytelsesmetrikker. Metrikkene brukt, er som klassifiseringsalgortimene, pakker fra scikit-learn [35].

Også for enklere sammenligning ble metrikken MCC også kjørt som MCC skalert, altså at den får intervallet 0 til 1, fremfor -1 til 1 som den var, slik den er i resultatene fra Huynh et al. [14]. Den skalerte MCC er beregnet ved:

$$MCC\_skalert = \frac{MCC}{2} + \frac{1}{2} \tag{3.1}$$

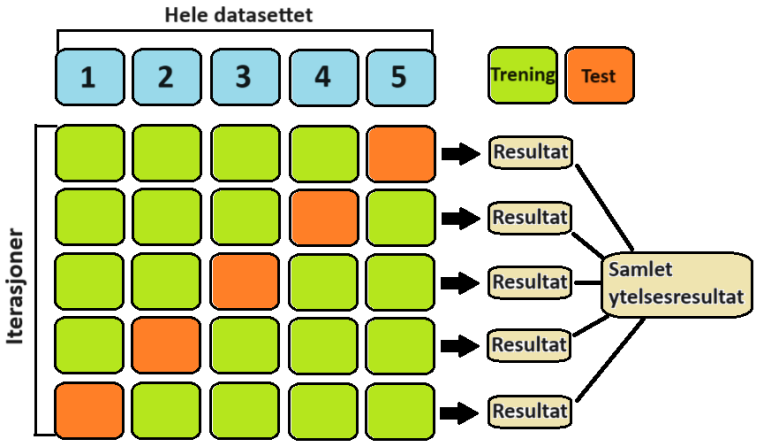
Det er viktig å merke, at videre i oppgaven fra her så vil MCC_skalert bare bli referert til som MCC og den uskalerte MCC er ikke med videre i resultater.

3.4.1 Kryssvalidering

Kryssvalidering blir brukt i maskinlæring, og anvendes slik at man foretar en tilfeldig fordeling av datasettet til et gitt antall mindre grupper, såkalte folder. Man deler foldene i to separate deler, hvorav en fold er til testdata og resten til treningsdata [23]. Modellen som skal evalueres ved hjelp av kryssvalideringen blir trent på alle foldene, med unntak av testfolden, som tester modellen. Videre gjør man dette på nytt i forskjellige kombinasjoner helt til alle foldene har blitt brukt som testdata. For hver prediksjon av testdata blir disse resultatene lagret i en vektor. Det kan ses et illustrert eksempel fra en 5-foldet kryssvalidering i figur 3.1.

Neste steg er da som oftest å uthente gjennomsnittsresultatet, hvorav formålet er å bidra til et mer presist ytelsesestimat for modellen enn det man kunne fått

med en enkeltstående testdel og treningsdel [23]. Det kan også som gjort i denne oppgaven bli hentet ut et *overall* resultat av de lagrede resultatene for hver test, dette vil si et aggregert (samlet) resultat over alle testprediksjonene i stedet for et gjennomsnitt av disse. Dersom man foretar flere runder med kryssvalidering, vil man kunne oppnå mer presise og evalueringer av en modells ytelse [23].



Figur 3.1: *Figuren illustrerer prosessen av en 5-foldet kryssvalidering. De grønne foldene viser til treningsdataen, mens de oransje foldene er testdataen. Hver iterasjon av de fem foldene gir ut et testresultat, som videre gir et aggregert (samlet) ytelsesresultat. Inspirert fra [23].*

3.4.2 Grid search-kryssvalidering

Grid search-kryssvalidering blir brukt til å finne den beste kombinasjonen av ulike hyperparametere for hver enkelt ML-modell [23]. Det blir for hver klassifiseringsalgoritme laget en *parametergrid*, dette er en slags tabell som består av ulike verdier for hver av de valgte hyperparameterene [23]. Under en kryssvalidering blir så modellen testet for hver unike parameterkombinasjon. Ved å ta i bruk grid search og kryssvalidering reduseres sjansen for overtilpasning. Overtilpasning skjer når modellen lærer seg treningsdataen så godt at den ikke klarer å yte på ny, usett data [23]. Siden modellen i denne prosessen (grid search-kryssvalidering) tester parameterkombinasjonene på et testsett og ikke treningsdata, vil dette styrke modellens evne til å generalisere ikke tidligere observert data [23].

Den parameterkombinasjonen som resulterte i den høyeste gjennomsnittlige verdien over alle metrikkene, under kryssvalideringen ble valgt for videre bruk i modelleringsprosessen. I vedlegg A.1 ligger en oversikt over de ulike parameterne og verdiene brukt til optimalisering, og i vedlegg A.2 og A.3 finner man de parameterene og verdiene valgt ut som den optimale kombinasjonen for hver enkelt modell med både DFS og OS som responsvariabel.

De fire klassifiserinsalgoritmene brukt i grid search og kryssvalidering er logistisk regresjon, random forest, decision tree og KNN. De metrikkene som ble brukt for kryssvalideringen var nøyaktighet, AUC, MCC og F1-score for klasse 1 og 0. For kryssvalideringen ble det brukt en stratifisert 5-foldet kryssvalidering. Stratifise-

ringen tok utgangspunkt i egenskapen *uicc8_III-IV*, for å sikre en jevn fordeling av de to klassene for stadie (1-2 og 3-4) i hver fold.

Det er ut fra analyse av datasettet sett at pasienter med stadie 3-4 i *uicc8_III-IV* har høyere utfall av responsen klasse 0, altså død eller tilbakefall. Det er med DFS som respons 68,71% av pasientene som er i klasse 0 og med OS som respons 63,27% av pasientene. I tillegg kan det ses i tabell 3.4, at fordelingen av stadie 1-2 og 3-4 veldig ubalansert. Disse skjeve fordelingene er en viktig grunn til at denne stratifiseringen er viktig.

Tabell 3.5 viser hvordan pasientene er delt opp for fem folder i treningsdata og testdata, på datasettene for pasienter med positiv HPV-status og de med negativ HPV-status under stratifisert kryssvalidering.

Tabell 3.5: *Tabellen viser hvordan fordelingen av pasienter er per fold for de fem foldene under kryssvalidering og grid search for datasettene med positiv eller negativ HPV-status, uavhengig av hvilken responsvariabel som er brukt.*

		Folder				
Datasett HPV-status	**Type data**	**1**	**2**	**3**	**4**	**5**
Pasienter med positiv HPV-status	Treningsdata	81	81	82	82	82
	Testdata	21	21	20	20	20
Pasienter med negativ HPV-status	Treningsdata	108	109	109	109	109
	Testdata	28	27	27	27	27

Pakkene benyttet i grid search-kryssvalideringen er GridSearchCV og StratifiedKFold, som begge er hentet fra [35].

Kapittel 4

Resultater

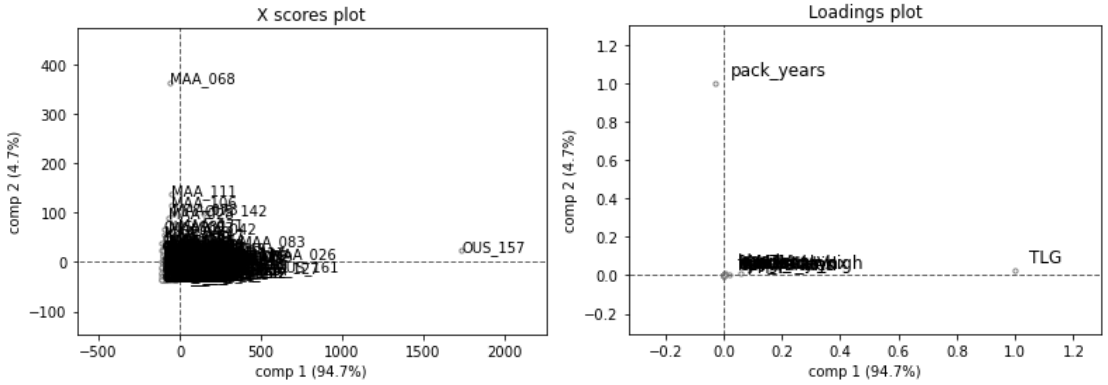
I dette kapittelet blir resultatene som fremkommer fra prosessen beskrevet i kapittel 3 presentert.

4.1 Forhåndsanalyse av datasett

PCA-plott

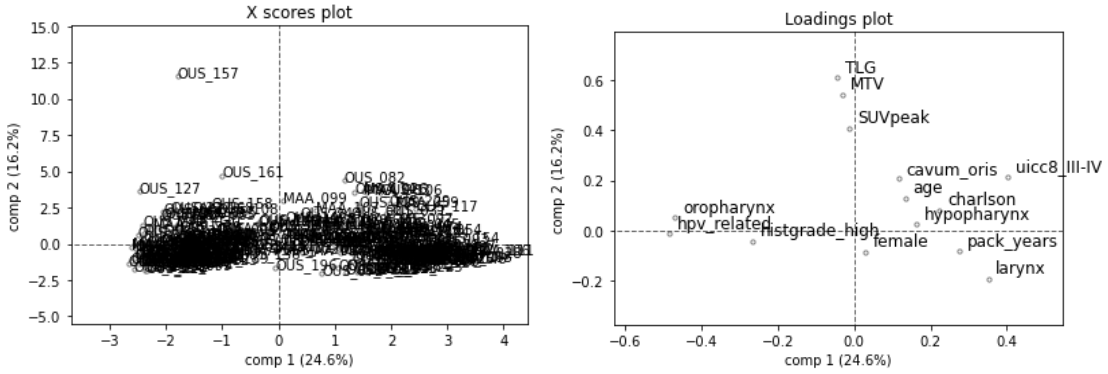
For å se på sammenhenger og varianser for egenskapene i det kliniske datasettet, er det i figur 4.1 vist score- og loadingplot for D1 sentrert. Scoreplottet viser til variansen eller spredningen i datasettet og gir et bilde av ulike mønstre eller grupperinger for pasientene i dataen. Loadingplottet viser hvor mye hver egenskap bidrar til ulike prinsipalkomponenter. Prinsipalkomponentene forklarer variansen i datasettet, som vist i figuren står prinsipalkomponent 1 (PC1) for 94,7% av variansen, og prinsipalkomponent 2 (PC2) står for 4,7% av totalen.

Det at man har et sentrert datasett, vil si at verdiene i datasettet er blitt justert for å få gjennomsnittet lik null. De blir justert ved at gjennomsnittet av alle verdiene subtraheres fra hver enkelt verdi. Dette gjøres for å lettere observere de egenskapene eller pasientene som har store avvik, da de vil ha høy varians.



Figur 4.1: *Score- og loadingplot for det kliniske datasettet, D1, som er sentrert.*

Det er to tydelige ekstremverdier (outliere) i scoreplottet i figur 4.1, altså to pasienter med stort avvik, dette er MAA_0,68 og OUS_157. I loadingplottet er det også to avvik, *pack_years* og *TLG*. Disse ser ut til å ha stor påvirking på de to pasientene med avvik i scoreplottet.



Figur 4.2: *Skalert score- og loadingplot for det kliniske datasettet D1.*

I figur 4.2 er det samme datasettet brukt, men nå skalert. Det at datasettet er skalert, vil si at verdiene i datasettet er blitt justert for å få gjennomsnittet til hver enkelt egenskap på null og med et standardavvik på en. For å få det skalert blir gjennomsnittet til egenskapen subtrahert fra hver enkelt verdi, og deretter blir verdien dividert med standardavviket [23].

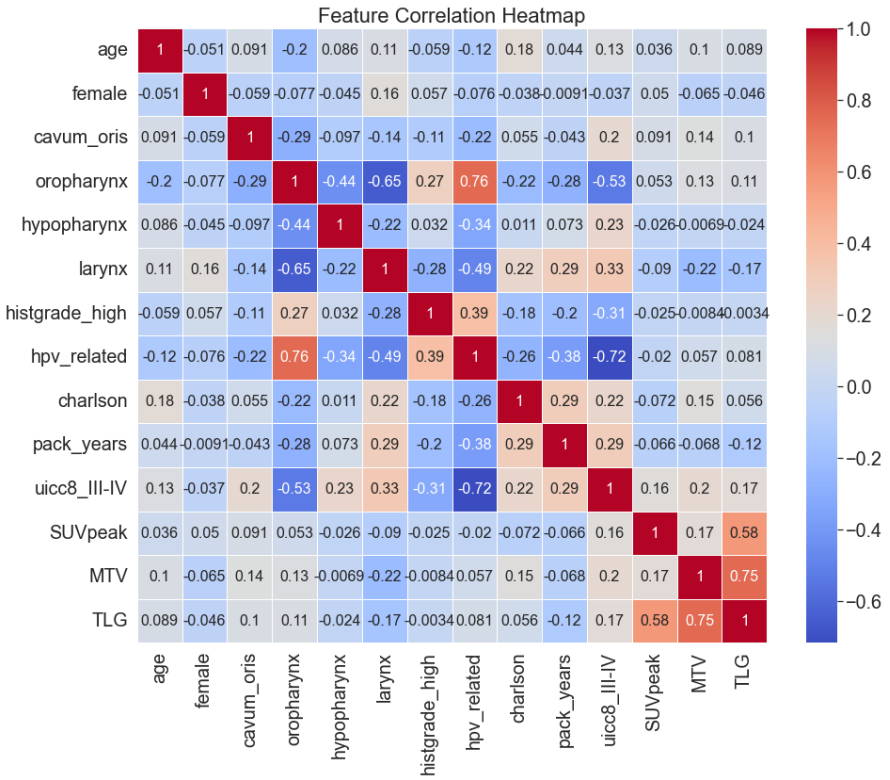
Det vises at PC1 utgjør 24,6% varians av totalen, og PC2 utgjør 16,2%. Også i dette score plottet har pasient OUS_157 et stort avvik fra grupperingene. Denne pasienten har i datasettet ekstremverdier for de tre PET-parameterene; SUV_{peak}, *MTV* og *TLG*. Det kan videre ses i loading plottet at de samme tre PET-parameterene er samlet i en gruppe, noe som tyder på like egenskaper mellom disse. Både pasient OUS_157 og PET-parameterene ser ut til å ikke korrelere godt med komponent 2. Andre egenskaper som ser ut til å bidra med mye av det samme er *oropharynx* og *hpv_related*, også *histgrade_high* korrelerer positivt med disse. Det kan i scoreplottet se ut som datasettet blir litt delt opp i to grupper, da det danner seg én klynge mot den negative siden av komponent 1 (x-aksen) og én mot den positive siden.

Korrelasjon

Figur 4.3 viser korrelasjonen mellom egenskapene i et *heatmap*, der det vises på fargeskalaen til høyre hvor høy korrelasjonen er. Det kan ses nede i høyre hjørne der *SUVpeak*, *MTV* og *TLG* møter hverandre, at fargene er nærmere rød der. Dette er de tre PET-parameterene fra loadingplottet i figur 4.2. Videre har *oropharynx* og *hpv_related* høy korrelasjon (0,76). Dette tallet gir mening da alle pasienter med hpv har oropharynx som primærsvulst, men alle pasienter med oropharynx som primærsvults har ikke nødvendigvis hpv.

Faktorene *hpv_related* og *uicc8_III-IV* har den sterkeste negative korrelasjonen som er på -0,72, også *oropharynx* korrelerer dårlig med *uicc8_III-IV* med en neg-

ativ verdi på -0,53. Man kan se mange av de samme sammenhengene mellom egenskapene i figur 4.1 og 4.2 og fra korrelasjonene i figur 4.3.



Figur 4.3: *Korrelasjonen mellom egenskaper vist med et heatmap. Verdien for korrelasjonskoeffisienten mellom alle mulige kombinasjoner av to egenskaper blir presentert her. Jo sterkere rød fargen er, desto sterkere er den positive korrelasjonen, og sterkere negativ korrelasjon mot mørk blå.*

4.2 Ytelsesresultater med sykdomsfri overlevelse (DFS) som responsvariabel

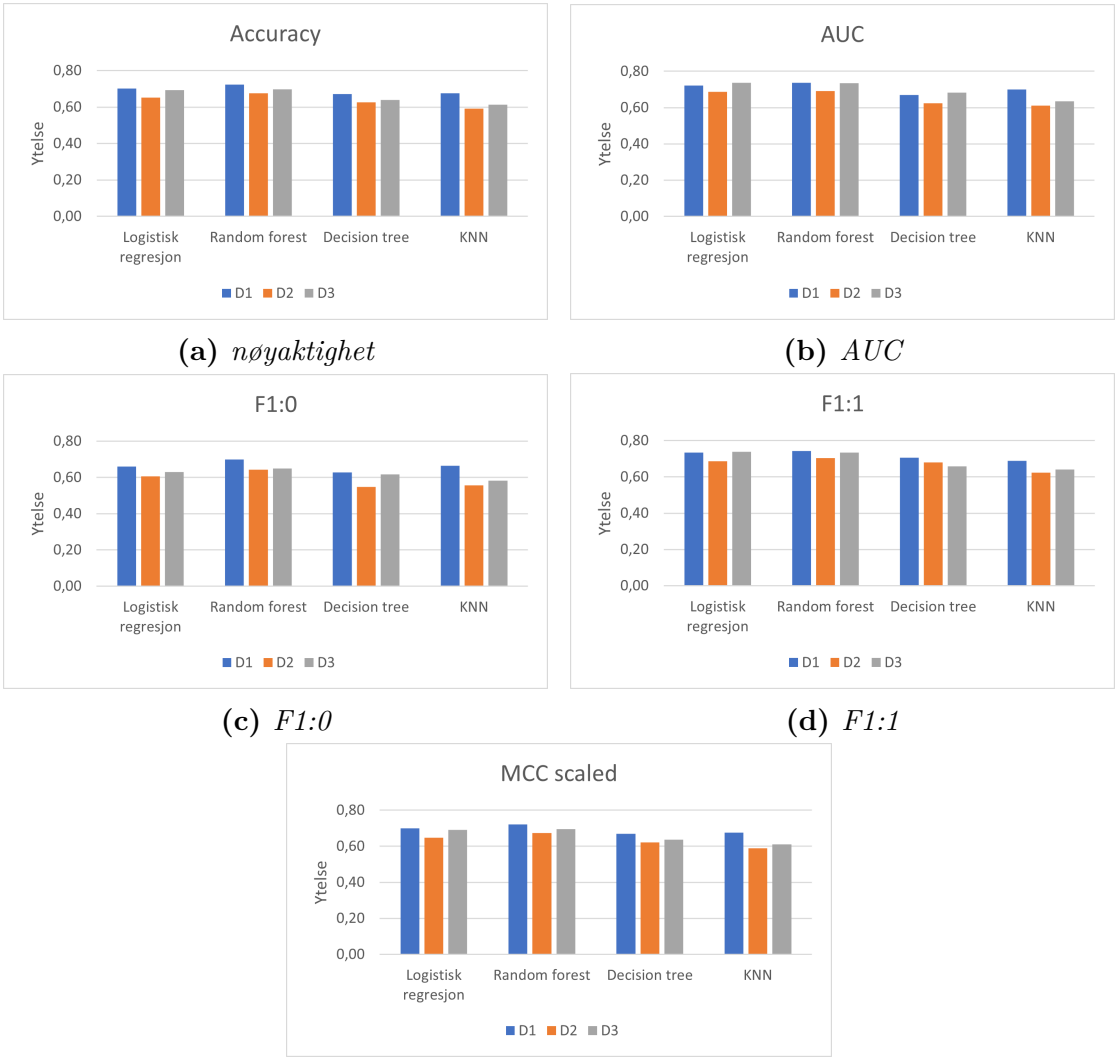
Under denne seksjonen, vil det vises resultater fra datasettene med DFS som responsvariabel. Det vil være resultater av ytelsen etter en fem-foldet kryssvalidering for de fem metrikkene for de fire klassifiseringsmodellene, logistisk regresjon, random forest, decision tree og KNN. Disse resultatene, er en samlet ytelsesvurdering av modellen på tvers av hele datasettet. Dette betyr at det ikke er et gjennomsnitt eller median av ytelsen på tvers av folder, men en aggregering av ytelsesmålingene på tvers av alle testforekomstene etter kryssvalidering. En aggregering vil her si at antall korrekte prediksjoner fra testsettene blir summert, og så dividert på totalt antall prediksjoner fra testsettene. Resultatene gir altså et helhetlig syn på modellens ytelse ved å vurdere alle testforekomster som om de var en del av et enkelt testsett.

Det vil også bli presentert ytelsesresultatene for hver fold av de fem foldene, for

hver modell. Disse er tatt med fordi resultatene fra per fold etter kryssvalidering, vil kunne gi en innsikt i stabiliteten og ytelsen til modellen for forskjellige deler av datasettet. Det vil aller først vises aggregerte ytelsesresultater fra det orginale datasettet, D1, D2 og D3, som var en test for modellene.

4.2.1 Forhåndstest av modell

For å sjekke at koden for modelleringen virket, ble de orginale datasettene, D1, D2 og D3, testet med de fire klassifiseringsmodellene, logistisk regresjon, random forest, decision tree og KNN med DFS som responsvariabel. Resultatene fra dette ble så sammenlignet med resultatene fra model 1 og 2 i Figure 2 og Supplementary Table F1 i Huynh et al. [14]. Resultatene for D1, D2 og D3 er vist i figur 4.4 og i tabell 4.1.



(a) *nøyaktighet* **(b)** *AUC*

(c) *F1:0* **(d)** *F1:1*

(e) *MCC*

Figur 4.4: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene uten deling på HPV-status, D1, D2 og D3.*

Tabell 4.1: *Tabellen viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene uten deling på HPV-status, D1, D2 og D3.*

Datasett	Algoritme	nøyaktighet	AUC	MCC	F1:1	F1:0
D1	Logistisk regresjon	0,70	0,72	0,70	0,73	0,66
	Random forest	0,72	0,74	0,72	0,74	0,70
	Decision tree	0,67	0,67	0,67	0,71	0,63
	KNN	0,68	0,70	0,68	0,69	0,66
D2	Logistisk regresjon	0,65	0,69	0,65	0,69	0,61
	Random forest	0,68	0,69	0,67	0,70	0,64
	Decision tree	0,63	0,62	0,62	0,68	0,55
	KNN	0,59	0,61	0,59	0,62	0,56
D3	Logistisk regresjon	0,69	0,74	0,69	0,74	0,63
	Random forest	0,70	0,73	0,70	0,73	0,65
	Decision tree	0,64	0,68	0,64	0,66	0,62
	KNN	0,61	0,64	0,61	0,64	0,58

I figur 4.4 kommer det fram at modellene presterer på et ganske jevnt nivå i forhold til hverandre, men logistisk regresjon og random forest har noe høyere ytelse enn decision tree og KNN. Dette kan også ses på resultatene i tabell 4.1. Disse resultatene er tilfredsstillende sett opp mot de som Huynh et al. [14] fikk.

4.2.2 Utfallsprediksjon for pasienter med positiv HPV-status

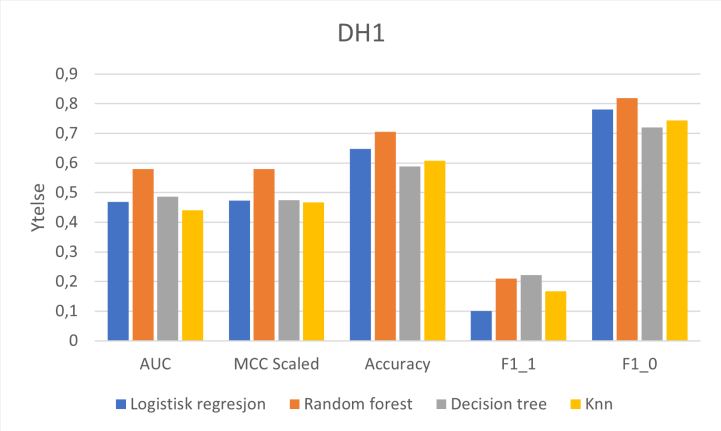
Resultatene fra datasettene bestående kun av pasienter med positiv HPV-relasjon og DFS som responsvariabel er vist under i tabell 4.2.

Tabell 4.2: *Tabellen viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene bestående av pasienter med positiv HPV-status, DH1, DH2 og DH3.*

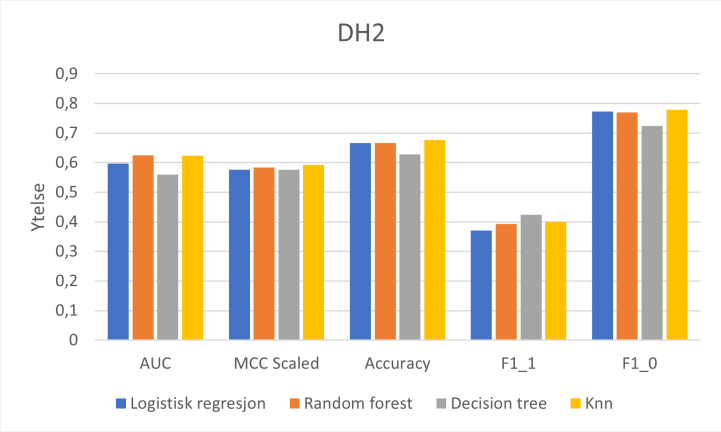
Datasett	Algoritme	nøyaktighet	AUC	MCC	F1:1	F1:0
DH1	Logistisk regresjon	0,65	0,47	0,47	0,10	0,78
	Random forest	0,71	0,58	0,58	0,21	0,82
	Decision tree	0,59	0,49	0,47	0,22	0,72
	KNN	0,61	0,44	0,47	0,17	0,74
DH2	Logistisk regresjon	0,67	0,60	0,58	0,37	0,77
	Random forest	0,67	0,63	0,58	0,39	0,77
	Decision tree	0,63	0,56	0,58	0,42	0,72
	KNN	0,68	0,62	0,59	0,40	0,78
DH3	Logistisk regresjon	0,65	0,59	0,55	0,33	0,76
	Random forest	0,63	0,48	0,47	0,14	0,76
	Decision tree	0,59	0,51	0,50	0,30	0,71
	KNN	0,70	0,61	0,61	0,39	0,80

For det kliniske datasettet, DH1, presterer ingen av modellene optimalt, men random forest har de høyeste verdiene for alle metrikkene utenom F1:1 score, og har den høyeste ytelsen ut av de fire klassifiseringsalgoritmene. For DH2 har alle modellene ganske lik ytelse for samtlige metrikker. DH3 viser at KNN har den beste ytelsen, med de høyeste resultatene for alle metrikker, og logistisk regresjon er den nest beste.

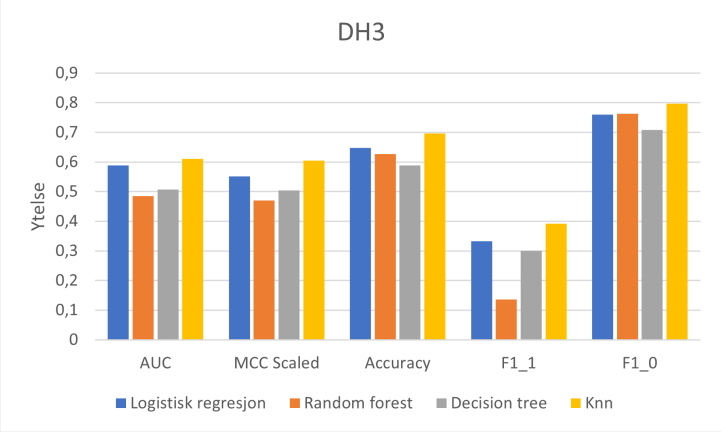
Under vises tre plott, ett for hvert av datasettene, bestående av ytelsen til de fire klassifiseringsalgoritmene, logistisk regresjon, random forest, decision tree og KNN for de fem ytelsesmetrikkene brukt i denne oppgaven. Metrikkene er nøyaktighet, AUC, MCC, F1:1 og F1:0.



Figur 4.5: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH1.*



Figur 4.6: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH2.*



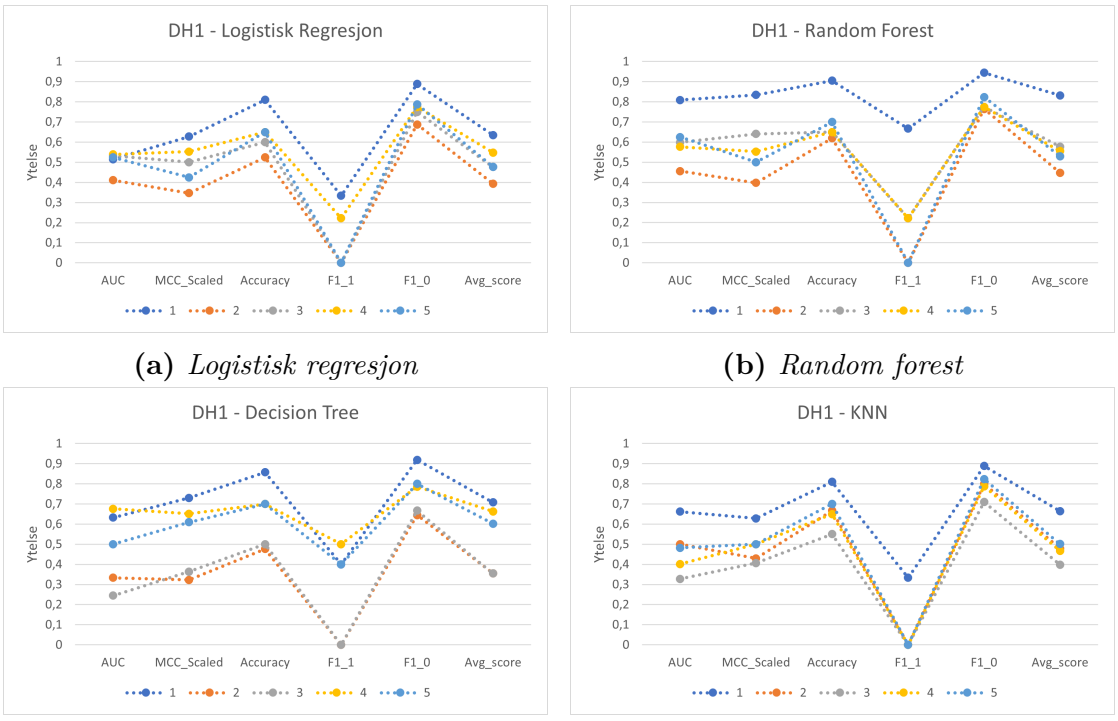
Figur 4.7: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH3.*

I figur 4.5 vises det tydelig, som også sett i tabell 4.2, at random forest har den høyeste ytelsen for DH1. Modellene gir alle en veldig lav F1:1 ytelse, og en veldig høy F1:0 ytelse. Logistisk regresjon har den aller laveste ytelsen for F1:1 som er på 0,10. For metrikkene AUC og MCC har alle modellene utenom random forest en lav ytelse på under 0,5.

Det vises for DH2 i figur 4.6 igjen at F1:1 ytelsen er lavest av de fem metrikkene, og F1:0 ytelsen er høyest. Alle modellene har også en veldig jevn ytelse på MCC. I resultatene for DH3 i figur 4.7, er F1:1 ytelsen desidert lavest igjen, med random forest som den dårligste modellen her, med ytelse på 0,14.

Resultater per fold

I figurene 4.8, 4.9 og 4.10 vises resultatene oppnådd for hver av de fem foldene fra grid search-kryssvalideringen for datasettene DH1, DH2 og DH3. De fem foldene står oppgitt som 1 til 5. De numeriske verdiene for per fold resultatene for DFS kan bli funnet i vedlegg B.0.1.



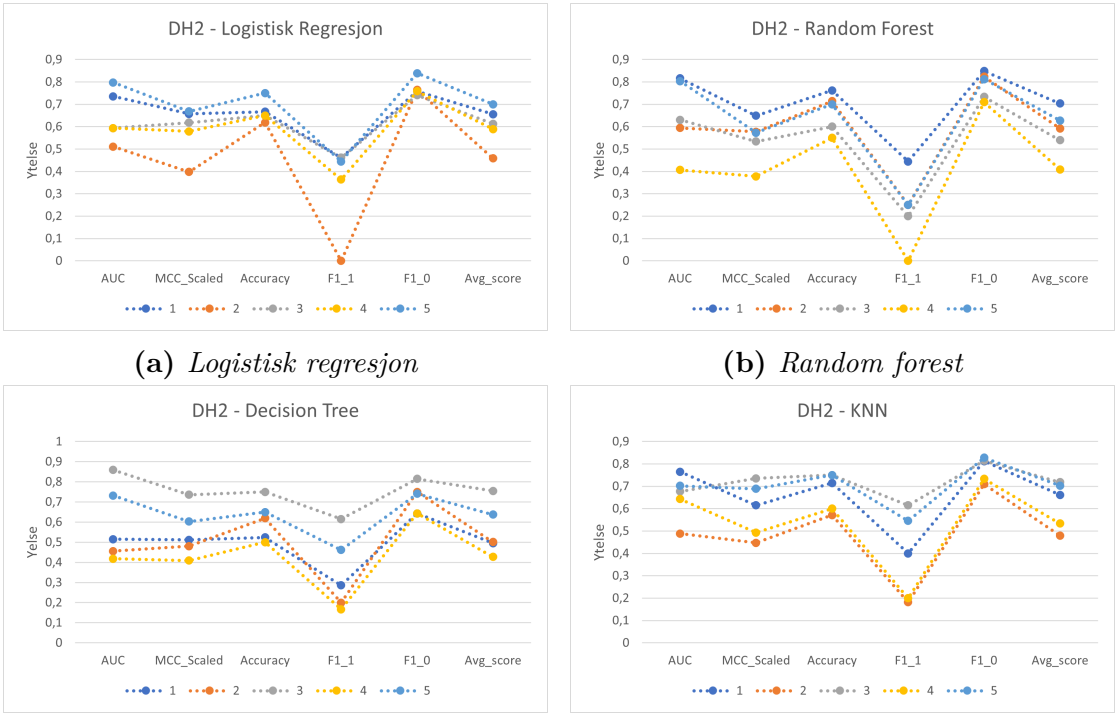
(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.8: *Per fold resultater fra alle klassifiseringsalgoritmene for det kliniske datasettet DH1, for pasienter med positiv HPV-status.*

I figur 4.8 ser vi resultatene per fold for det kliniske datasettet, DH1. For logistisk regresjon og KNN er variasjonene for foldene ikke like store som de er for random forest og decision tree. For random forest utmerker fold 1 seg i forhold til resten av foldene, da denne kurven yter høyest for alle utenom en metrikk,

F1:1. Det kan også for resten av klassifiseringsalgoritmene, ses at det er fold 1 som yter høyest av foldene. Under decision tree, gjør fold 2 og 3 det markant dårligere enn resten av foldene, da de i alle punkter har lavest verdier. Samtlige modeller har flere folder på ytelse 0 for F1:1, og for alle foldene, i alle modellene, er den høyeste ytelsen på metrikken F1:0 med ytelsesverdier på 0,65 og opp til 0,94.

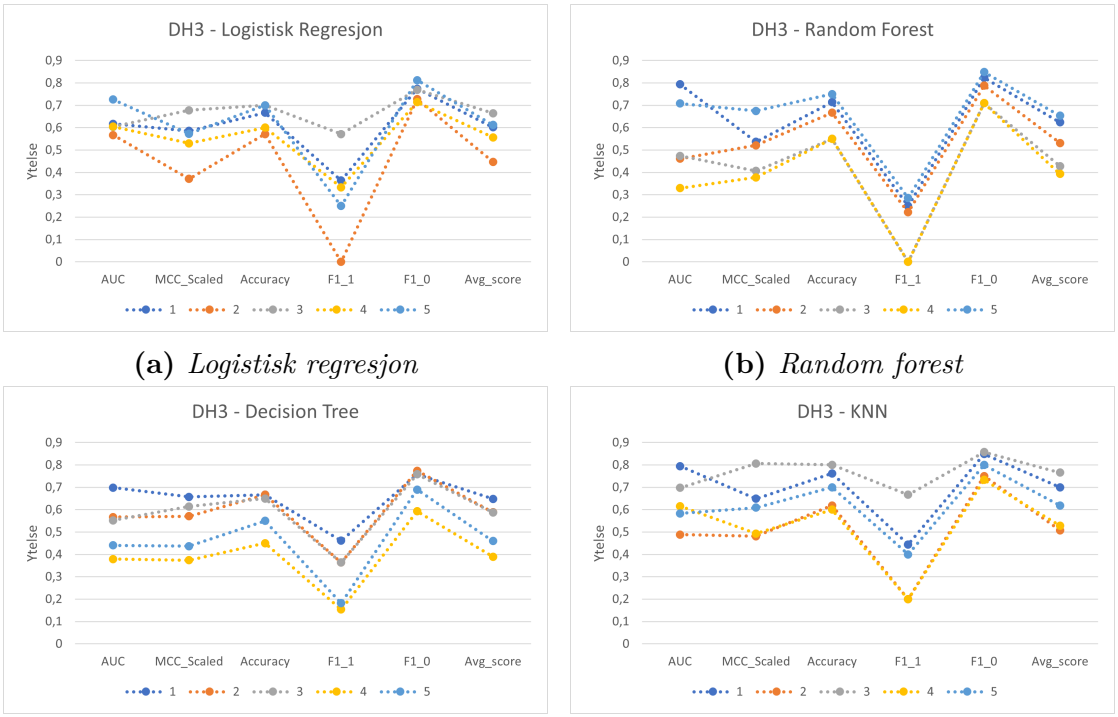


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.9: *Per fold resultater fra alle klassifiseringsalgoritmene for radiomics datasettet, DH2, for pasienter med positiv HPV-status.*

Figur 4.9 viser ytelsen for foldene i radiomics datasettet DH2. De modellene som gjør det jevnest per fold er decision tree og KNN. Ser man på logistisk regresjon er foldene forholdsvis jevne, utenom i ytelsen F1:1 for fold 2 som ligger på 0. For både klassifiseringsalgoritmene logistisk regresjon og random forest er det en fold som har en ytelse på 0 for metrikk F1:1. Resten av de andre foldene, for alle modeller, reduseres i forhold til de andre metrikkene ved F1:1. Det motsatte skjer ved metrikken F1:0, der vil ytelsen i forhold til resten av metrikkene enten holdes stabil eller økes.



(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.10: *Per fold resultater fra alle klassifiseringsalgoritmene for kombinasjonsdatasettet, DH3, for pasienter med positiv HPV-status.*

Den modellen som gjør det best per fold for kombinasjonsdatasettet DH3 i figur 4.10, er KNN. KNN og decision tree viser en mer stabil ytelse. Både logistisk regresjon og random forest har folder som har ytelse 0 for F1:1. Igjen slik som for DH1 og DH2 er F1:1 lavest på ytelse, mens F1:0 utmerker seg.

Det laveste punktet for hver fold i alle de fire modellene og de tre datasettene, DH1, DH2 og DH3, er F1:1 ytelse, se vedlegg B.1, B.2 og B.3 for eksakte verdier.

4.2.3 Utfallsprediksjon for pasienter med negativ HPV-status

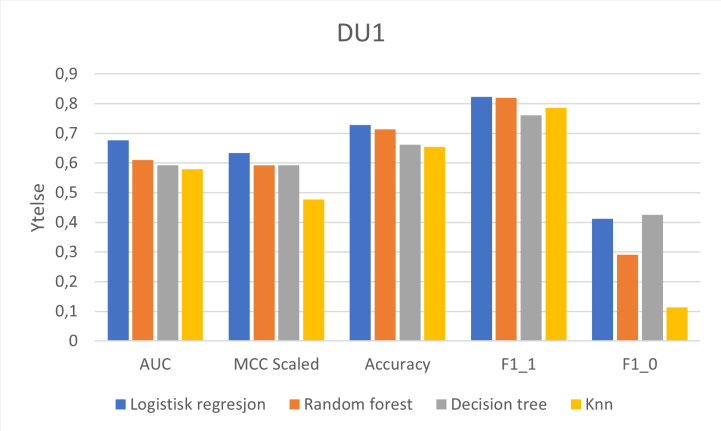
Resultatene fra datasettene bestående kun av pasienter med negativ HPV-relasjon og DFS som responsvariabel er vist under i tabell 4.3.

Tabell 4.3: *Tabellen viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene bestående av pasienter med negativ HPV-status, DU1, DU2 og DU3.*

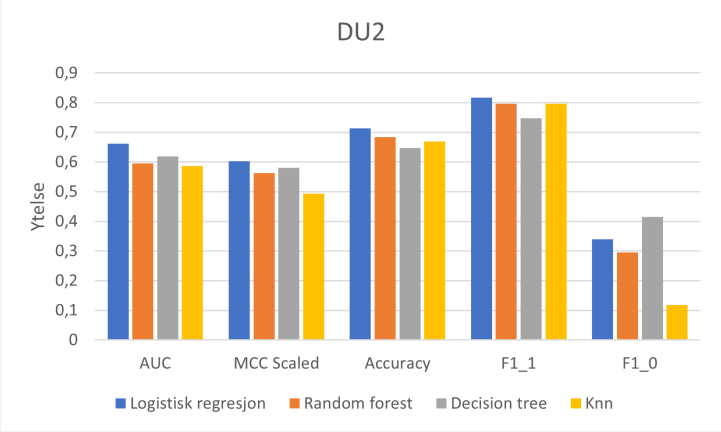
Datasett	Algoritme	nøyaktighet	AUC	MCC	F1:1	F1:0
DU1	Logistisk regresjon	0,73	0,68	0,63	0,82	0,41
	Random forest	0,71	0,61	0,59	0,82	0,29
	Decision tree	0,66	0,59	0,59	0,76	0,43
	KNN	0,65	0,58	0,48	0,79	0,11
DU2	Logistisk regresjon	0,71	0,66	0,60	0,82	0,34
	Random forest	0,68	0,60	0,56	0,80	0,30
	Decision tree	0,65	0,62	0,58	0,75	0,41
	KNN	0,67	0,59	0,49	0,80	0,12
DU3	Logistisk regresjon	0,67	0,68	0,61	0,76	0,46
	Random forest	0,71	0,62	0,60	0,82	0,34
	Decision tree	0,67	0,59	0,57	0,78	0,37
	KNN	0,68	0,62	0,58	0,78	0,37

For det kliniske datasettet, DU1, er det logistisk regresjon som gir den beste prestasjonen av modellene, og oppnår høyeste ytelse på alle metrikkene utenom F1:0. Den dårligste modellen her er KNN, da den yter lavest i alle utenom én metrikk. For DU2 gir igjen logistisk regresjon den høyeste ytelsen av modellene, det er bare for F1:0 den ikke er best. DU3 viser at det er liten variasjon i ytelse mellom modellene, men logistisk regrejon yter høyest, hakket over random forest.

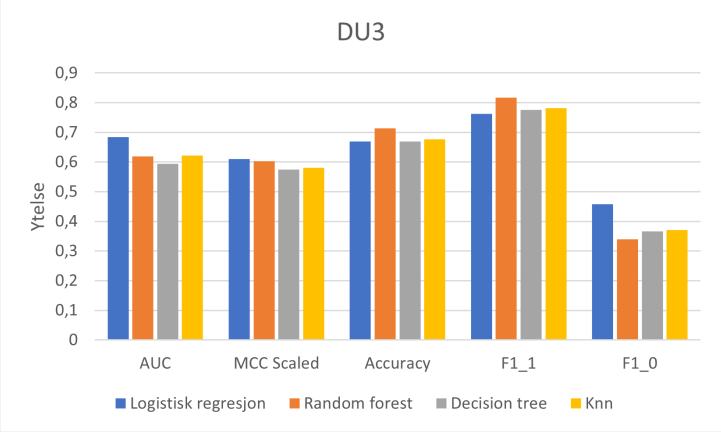
Det blir under vist tre plott, ett for hvert av datasettene, bestående av ytelsen til de fire klassifiseringsalgoritmene, logistisk regresjon, random forest, decision tree og KNN for de fem ytelsesmetrikkene brukt i denne oppgaven. Metrikkene er nøyaktighet, AUC, MCC, F1:1 og F1:0.



Figur 4.11: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU1.*



Figur 4.12: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU2.*



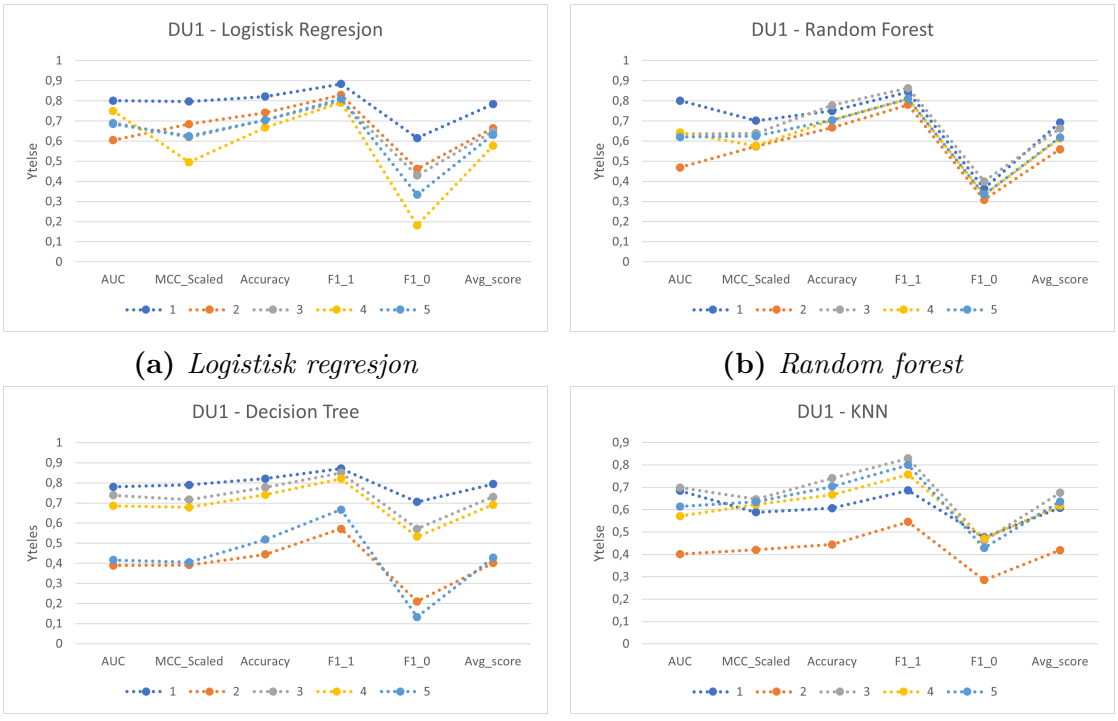
Figur 4.13: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU3.*

I figur 4.11 for DU1, er det tydelig at F1:0 yter dårligst av metrikkene for alle modeller, spesielt KNN yter lavt. Det samme gjelder også for DU2 i figur 4.12, der F1:0 er dårlig over alle modellene med lavest ytelse hos KNN. Resultatene for DU3 i figur 4.13, viser også F1:0 med lav ytelse. MCC viser en jevn ytelse for alle modellene i DU3.

Det er spesielt én tydelig forskjell fra resultatene under positiv HPV-relasjon til disse under negativ HPV-relasjon, og det er F1:1 og F1:0 ytelsen. Fra DH-settene presterte F1:1 jevnt over dårligst og F1:0 presterte bra. I DU-settene er dette motsatt, da F1:1 yter best og F1:0 lavest.

Resultater per fold

Det blir i figurene 4.14, 4.15 og 4.16 vist resultater oppnådd for hver av de fem foldene fra grid search-kryssvalideringen for datasettene DU1, DU2 og DU3. De numeriske verdiene for per fold resultatene for DFS kan bli funnet i vedlegg B.0.1.

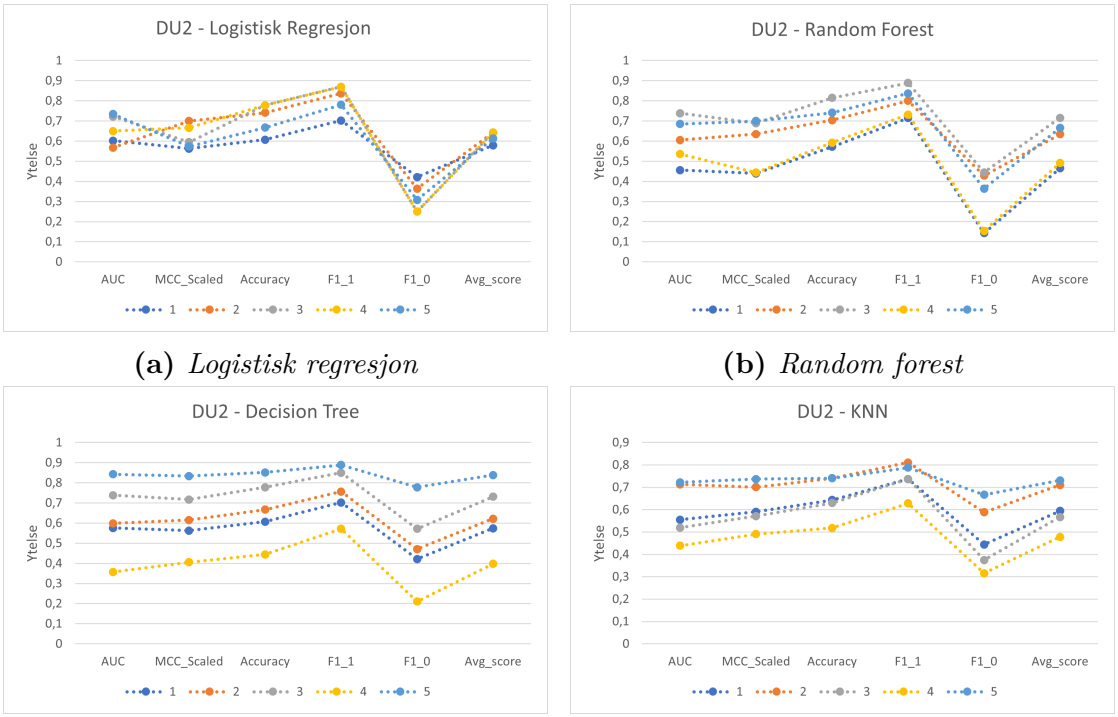


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.14: *Per fold resultater fra alle klassifiseringsalgoritmene for det kliniske datasettet DU1, for pasienter med negativ HPV-status.*

I figur 4.14 ser vi resultatene per fold for DU1. For random forest yter foldene veldig jevnt i de fleste punktene, men for MCC og AUC er det noe mer variasjon. Det kan for decision tree ses at fold 2 og fold 5 yter jevnt, men noe dårligere enn resten av foldene som også er veldig jevne. For KNN er det fold 2 som presterer dårligere enn de resterende.

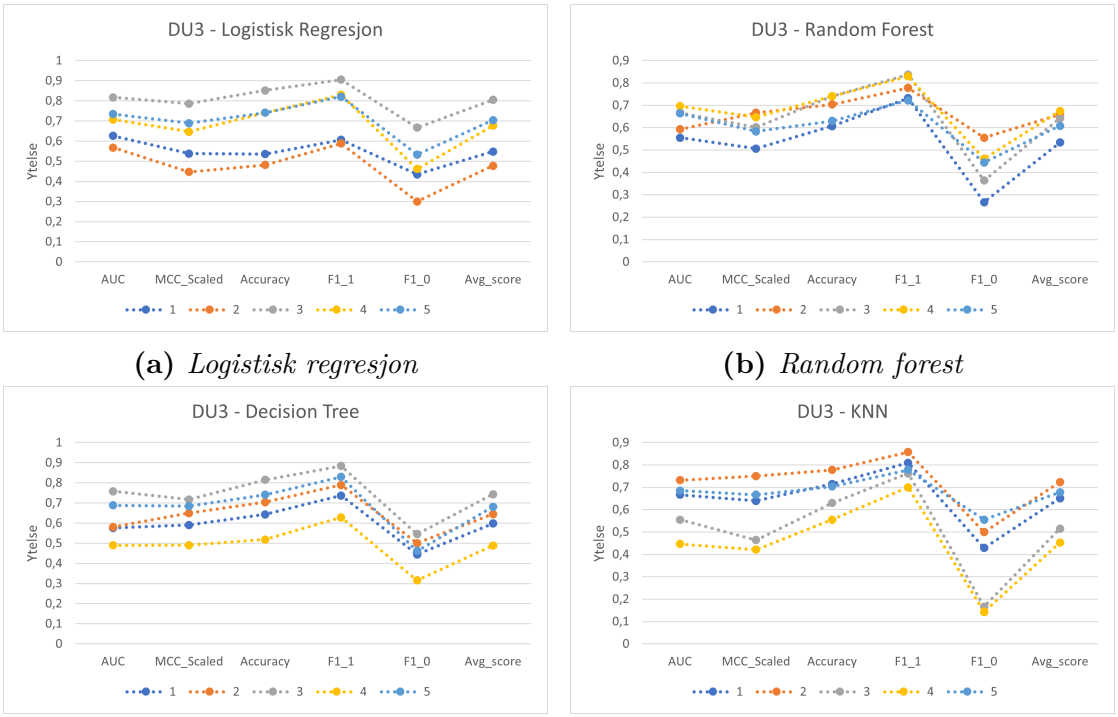


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.15: *Per fold resultater fra alle klassifiseringsalgoritmene for radiomics datasettet, DU2, for pasienter med negativ HPV-status.*

For DU2 i figur 4.15, ses det at logistisk regresjon har mest konsekvente folder, med laveste punkt i F1:0, men også random forest og KNN yter jevnt her. Decision tree er den med mest variasjon mellom foldene, der fold 5 er best og fold 4 yter dårligst.



(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.16: *Per fold resultater fra alle klassifiseringsalgoritmene for kombinasjonsdatasettet, DU3, for pasienter med negativ HPV-status.*

Den modellen med minst variasjon per fold i DU3 i figur 4.16 er random forest, men i F1:0 er det større variasjon i ytelse enn resten av metrikkene. For KNN er det jevnt for fold 1, fold 2 og fold 5. Også fold 3 og 4 er jevne men har dårligere ytelse enn de resterende foldene, spesielt i F1:0.

4.3 Ytelsesresultater med generell overlevelse (OS) som responsvariabel

Her kommer resultatene for datasettene med OS som responsvariabel. Oppsettet i denne delen vil være likt som det var over i resultatene med DFS som responsvariabel.

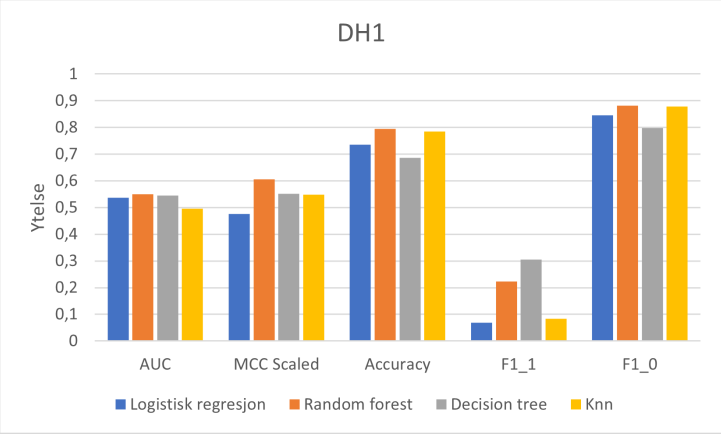
4.3.1 Utfallsprediksjon for pasienter med positiv HPV-status

Resultatene fra datasettene bestående kun av pasienter med positiv HPV-relasjon og OS som responsvariabel er vist for DH1, DH2 og DH3 for hver av klassifiseringalgoritmene under i tabell 4.4. Videre er disse resultatene vist som barplott i figurene 4.17, 4.18 og 4.19.

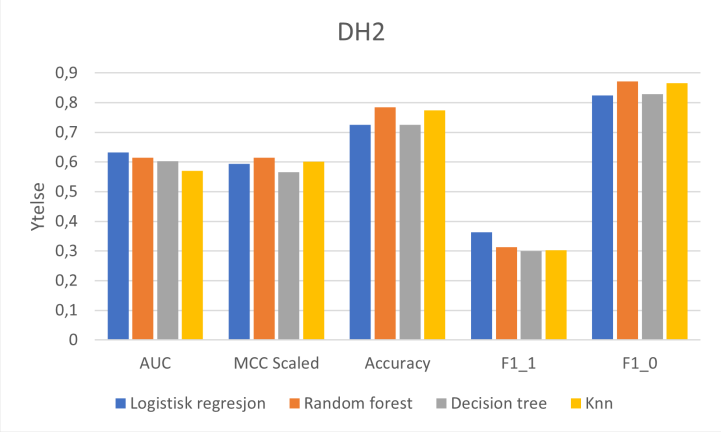
Tabell 4.4: *Tabellen viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene bestående av pasienter med positiv HPV-status, DH1, DH2 og DH3.*

Datasett	Algoritme	nøyaktighet	AUC	MCC	F1:1	F1:0
DH1	Logistisk regresjon	0,74	0,54	0,48	0,07	0,85
	Random forest	0,79	0,55	0,61	0,22	0,88
	Decision tree	0,69	0,55	0,55	0,30	0,80
	KNN	0,78	0,50	0,55	0,08	0,88
DH2	Logistisk regresjon	0,73	0,63	0,59	0,36	0,83
	Random forest	0,78	0,61	0,61	0,31	0,87
	Decision tree	0,73	0,60	0,57	0,30	0,83
	KNN	0,77	0,57	0,60	0,30	0,87
DH3	Logistisk regresjon	0,72	0,66	0,56	0,29	0,82
	Random forest	0,74	0,61	0,54	0,23	0,84
	Decision tree	0,66	0,55	0,55	0,31	0,77
	KNN	0,78	0,57	0,61	0,31	0,87

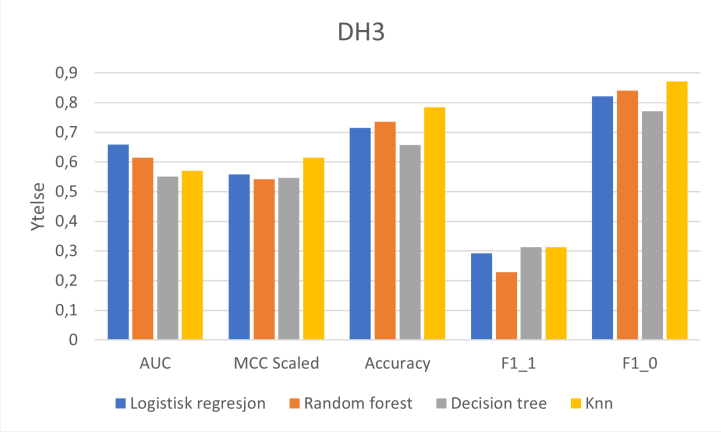
I tabell 4.4 er det for DH1 random forest som gjør det litt bedre enn de andre modellene, med en toppytelse på 0,79 på nøyaktighet og 0,61 på MCC. Alle de fire modellene følger hverandre ganske jevnt over de fem metrikkene, og det er en ganske stor variasjon fra metrikk til metrikk. For DH2 varierer det for hver metrikk hvordan modellene yter, men over de alle er det også her random forest som presterer hakket bedre, mens decision tree gjør det dårligst, sett bort fra ytelsen for F1:1. Ser man på DH3 er det KNN som har den beste yteslen, den har høyeste ytelse for metrikkene nøyaktighet, MCC og F1:0 som er på 0,78, 0,61 og 0,87. Det er for DH3, sånn som for DH2, decision tree som yter lavest gjennomsnittlig over alle metrikkene.



Figur 4.17: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH1.*



Figur 4.18: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH2.*



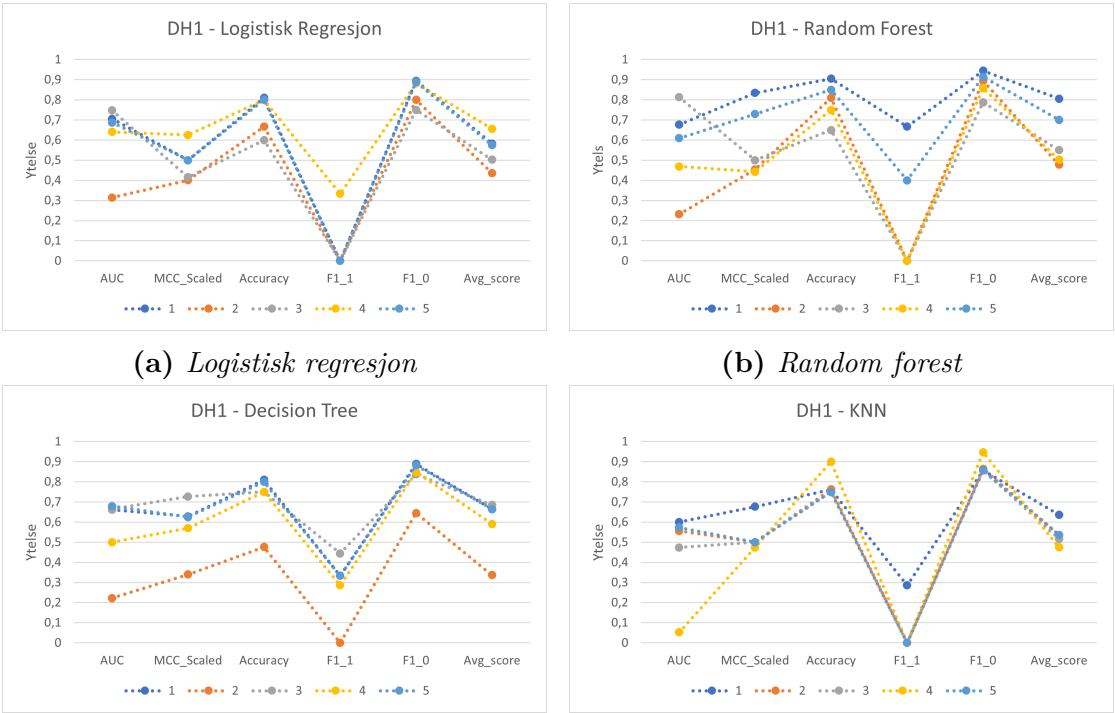
Figur 4.19: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DH3.*

Fra figur 4.17 kan det tydelig ses at F1:1 gjør det dårligst av metrikkene for alle modellene, der logsistisk regresjon og KNN har en svært lav ytelse på 0,07 og 0,08. Den metrikken alle modellene gjør det best for, er F1:0, der samtlige modeller har en ytelse på 0,80 eller over. Også for nøyaktighet yter modellene godt, og bortsett fra decision tree som yter 0,69, er alle de andre over 0,74.

For DH2 sett i figur 4.18, har alle modellene, slik som for DH, best ytelse i F1:0. Alle ytelsene ligger på 0,83 eller over her. Igjen er det også nøyaktighet modellene yter nest best for, og for F1:1 det er dårligst prestering. Ser man på resultatene for DH3 i figur 4.19, ser det ikke så forskjellig ut fra DH1 og DH2. Der F1:0 og nøyaktighet presterer best og nest best, og igjen yter F1:1 dårligst.

Resultater per fold

Det blir som, i DFS seksjonen, i figurene 4.20, 4.21 og 4.22 vist resultater oppnådd for hver av de fem foldene fra grid search-kryssvalideringen for datasettene DH1, DH2 og DH3. Nevner igjen at de fem foldene står oppgitt som 1 til 5. De numeriske verdiene for per fold resultatene for OS kan bli funnet i vedlegg B.0.2.



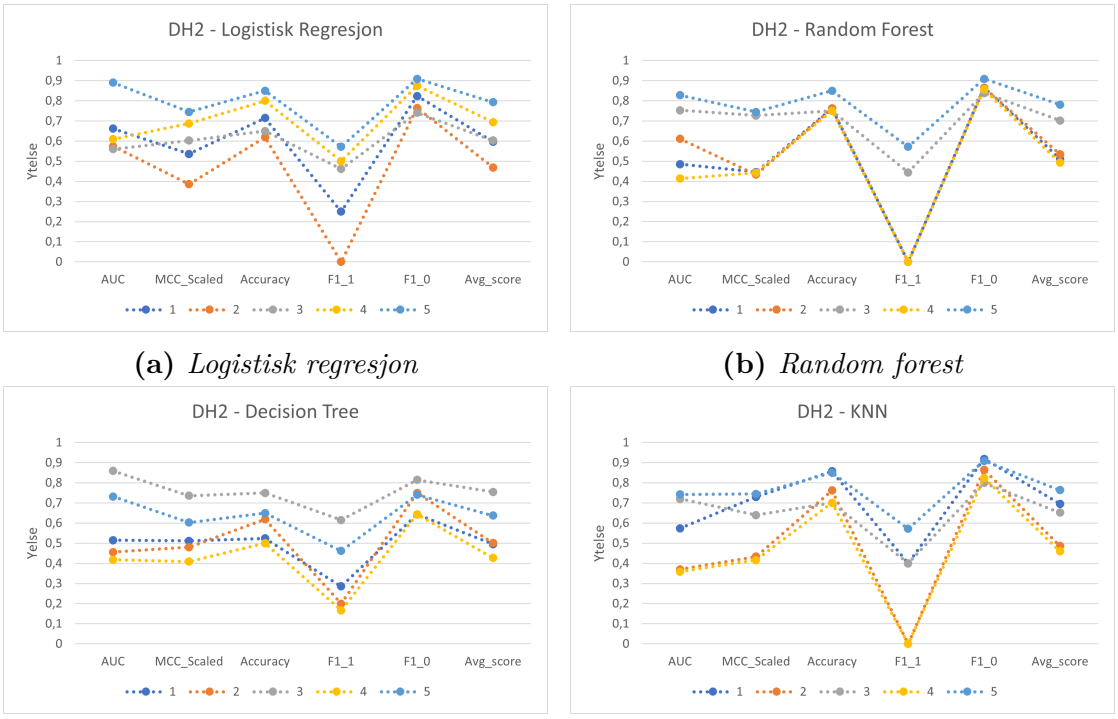
(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.20: *Per fold resultater fra alle klassifiseringsalgoritmene for det kliniske datasettet DH1, for pasienter med positiv HPV-status.*

Fra figur 4.20 er det klart at det for alle modellene er F1:1 som presterer desidert dårligst. For logistisk regresjon og KNN yter hele fire av de fem foldene 0,0 for F1:1, mens random forest har tre på 0,0 i ytelse. Det er ingen av modellene som har noe særlig jevne folder, utenom fold 1 og 5 for decision tree, og fold 2 og 5

i KNN, som yter nesten helt likt i alle metrikkene. I KNN ser man og at fold 4 varierer veldig gjennom metrikkene. Den har en veldig mye lavere AUC ytelse i forhold til resten av foldene, som er på 0,05, mens for nøyaktighet og F1:0 er den høyest, med ytelser på 0,90 og 0,95.

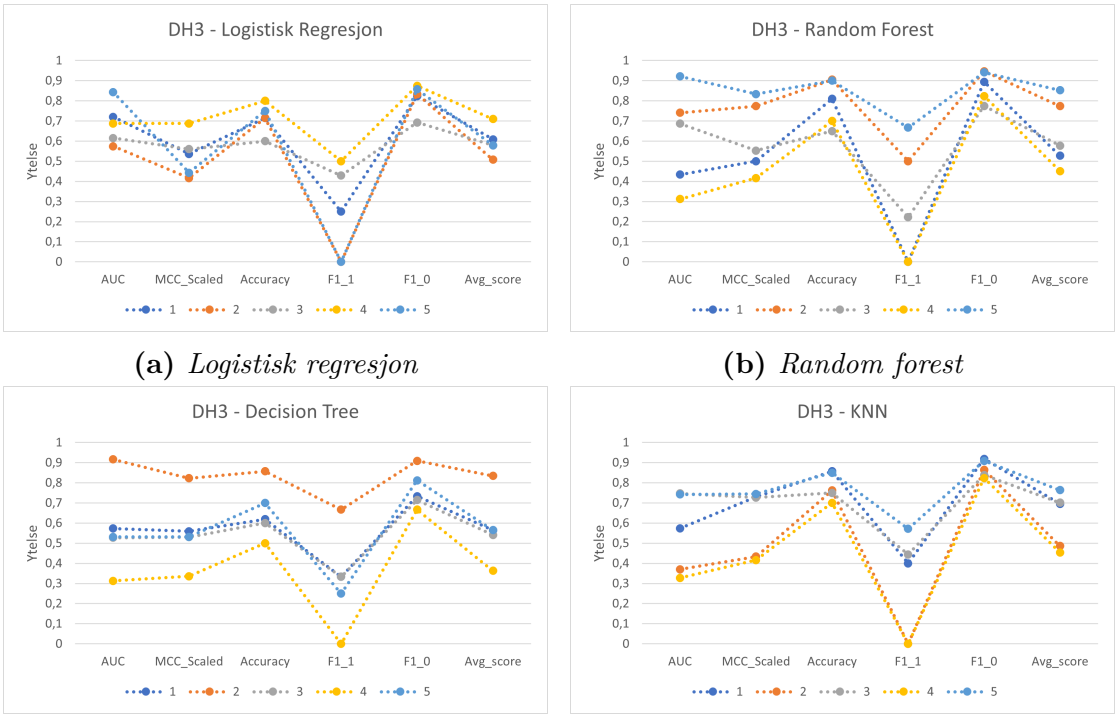


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.21: *Per fold resultater fra alle klassifiseringsalgoritmene for radiomics datasettet, DH2, for pasienter med positiv HPV-status.*

I resultatene fra DH2 i figur 4.21, er det for decision tree det er mest stabilt for de fem foldene. I KNN følger fold 2 og fold 4 hverandre gjennom alle de seks metrikkene. Det samme gjør fold 1, fold 2 og fold 4 i random forest også, bortsett fra en ytelsesforskjell for AUC. Foldene i logistisk regresjon yter helt ulikt, men følger samme mønster i hvordan det ytes for metrikkene.



(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.22: *Per fold resultater fra alle klassifiseringsalgoritmene for kombinasjonsdatasettet, DH3, for pasienter med positiv HPV-status.*

I figur 4.22 for DH3, kan man for decision tree se at den har én fold, fold 2, som yter veldig høyt og har verdier helt opp på 0,92 for AUC og 0,91 for F1:0, mens de andre foldene ikke er fullt så sterke. Det kan ses at for DH3 yter foldene i KNN veldig likt det de gjorde i for KNN i DH2. For random forest er det ingen jevne folder og alle yter helt forskjellig.

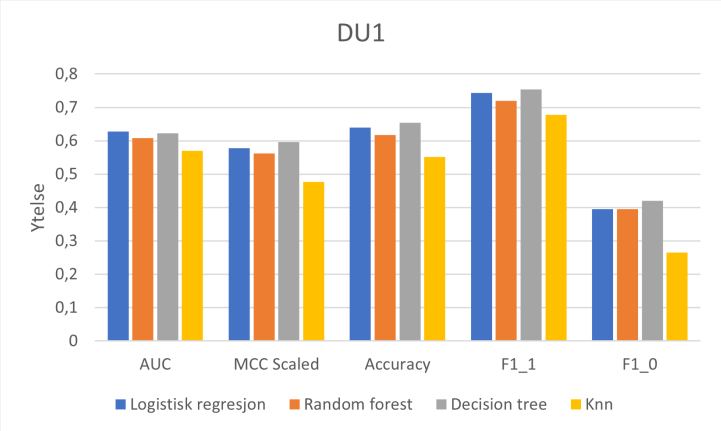
4.3.2 Utfallsprediksjon for pasienter med negativ HPV-status

I denne seksjonen blir resultatene fra datasettene bestående kun av pasienter med negativ HPV-relasjon og OS som responsvariabel presentert. Under i tabell 4.5, er DU1, DU2 og DU3 for hver av klassifiseringalgoritmene, og videre er disse resultatene igjen vist som barplott i figurene 4.23, 4.24 og 4.25.

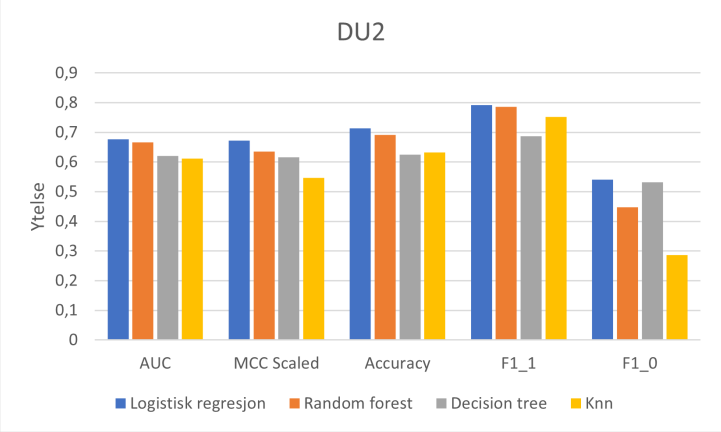
Tabell 4.5: *Tabellen viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for klassifiseringsalgoritmene logistisk regresjon, random forest, decision tree og KNN brukt med datasettene bestående av pasienter med negativ HPV-status, DU1, DU2 og DU3.*

Datasett	Algoritme	nøyaktighet	AUC	MCC	F1:1	F1:0
DU1	Logistisk regresjon	0,64	0,63	0,58	0,74	0,40
	Random forest	0,62	0,61	0,56	0,72	0,40
	Decision tree	0,65	0,62	0,60	0,75	0,42
	KNN	0,55	0,57	0,48	0,68	0,27
DU2	Logistisk regresjon	0,71	0,68	0,67	0,79	0,54
	Random forest	0,69	0,67	0,64	0,79	0,45
	Decision tree	0,63	0,62	0,62	0,69	0,53
	KNN	0,63	0,61	0,55	0,75	0,29
DU3	Logistisk regresjon	0,68	0,67	0,64	0,76	0,51
	Random forest	0,69	0,65	0,64	0,78	0,49
	Decision tree	0,63	0,62	0,62	0,69	0,53
	KNN	0,68	0,62	0,62	0,77	0,46

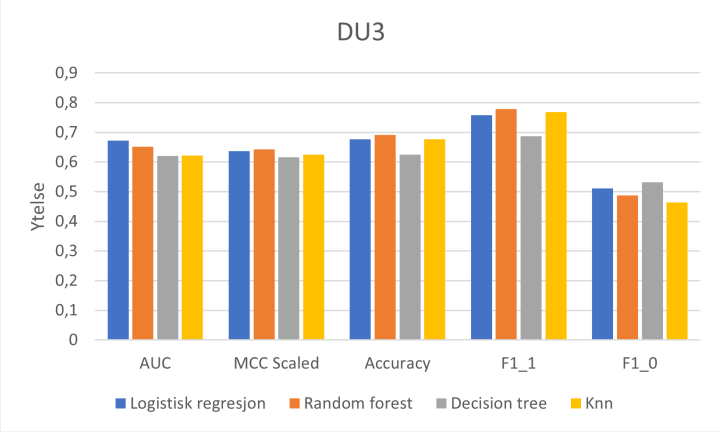
Resultatene i figur 4.23 viser at for DU1 har decision tree og logistisk regresjon best ytelse rett over random forest, mens KNN er modellen som yter lavest over alle metrikkene. Modellen som presterer best for DH2 er logistisk regresjon som yter høyest for fire av de fem metrikkene og har ytelse på 0,71 for nøyaktighet. Ytelsesmålingene for DH3 er litt jevnere for modellene, men gjennomsnittlig over metrikkene er det logistisk regresjon og random forest som yter likt og best av modellene.



Figur 4.23: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU1.*



Figur 4.24: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU2.*



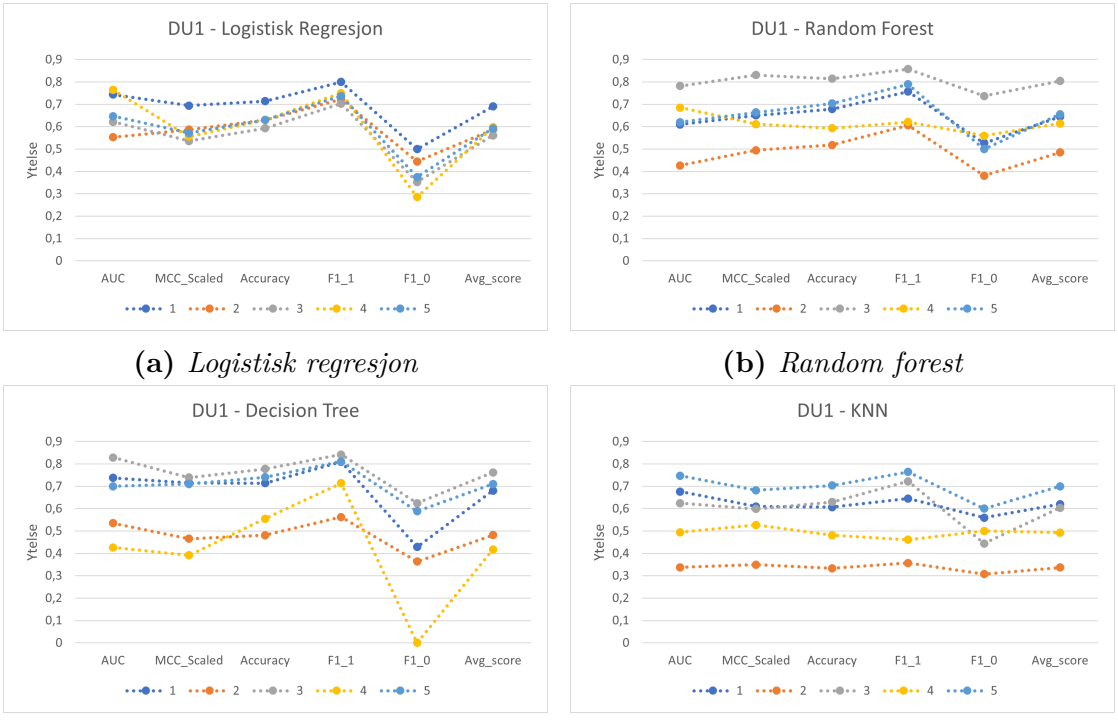
Figur 4.25: *Figuren viser de aggregerte ytelsesresultatene fra fem-foldet kryssvalidering for hver klassifiseringsalgoritme brukt med datasettet DU3.*

Det kan ses i figur 4.23 at presteringen for AUC, MCC og nøyaktighet ikke er så ulike, men KNN yter svakere enn resten av modellene. Det er F1:1 som modellene yter best for, med ytelse 0,74 for logistisk regresjon og 0,75 for decision tree. For F1:0 gjør de det svakest, med KNN dårligst og en ytelse på 0,27. Ytelsen for DU2 som ses i figur 4.24, viser også best ytelse for F1:1 og dårligst for F1:0. Logistisk regresjon og random forest har for F1:1 en ytelse på 0,79, mens KNN har 0,29 for F1:0. Det er for AUC, MCC og nøyaktighet ikke så stor forskjell, men det ses at logistisk regresjon og random forest yter høyest for disse. For resultatene i figur 4.25, ser man ikke noen umiddelbare store forskjeller for modellene. Metrikkene AUC, MCC og nøyaktighet er jevne, men det vises jo også her som i DU1 og DU2 at det er F1:1 og F1:0 som yter best og dårligst.

Kan også i denne seksjonen med OS som responsvariabel, slik som under DFS, se trenden med at F1:1 og F1:0 bytter om på å være best og dårligst. Igjen viser resultatene at F1:1 gjør det dårligst i prediksjonene for pasienter med HPV-relasjon og F1:0 gjør det her best. Så for pasienter uten HPV-relasjon har F1:1 den beste ytelsen, mens F1:0 yter da lavest.

Resultater per fold

Under i figurene 4.26, 4.27 og 4.28 blir det vist resultater oppnådd for hver av de fem foldene fra grid search-kryssvalideringen for datasettene DU1, DU2 og DU3. De numeriske verdiene for per fold resultatene for OS kan bli funnet i vedlegg B.0.2.

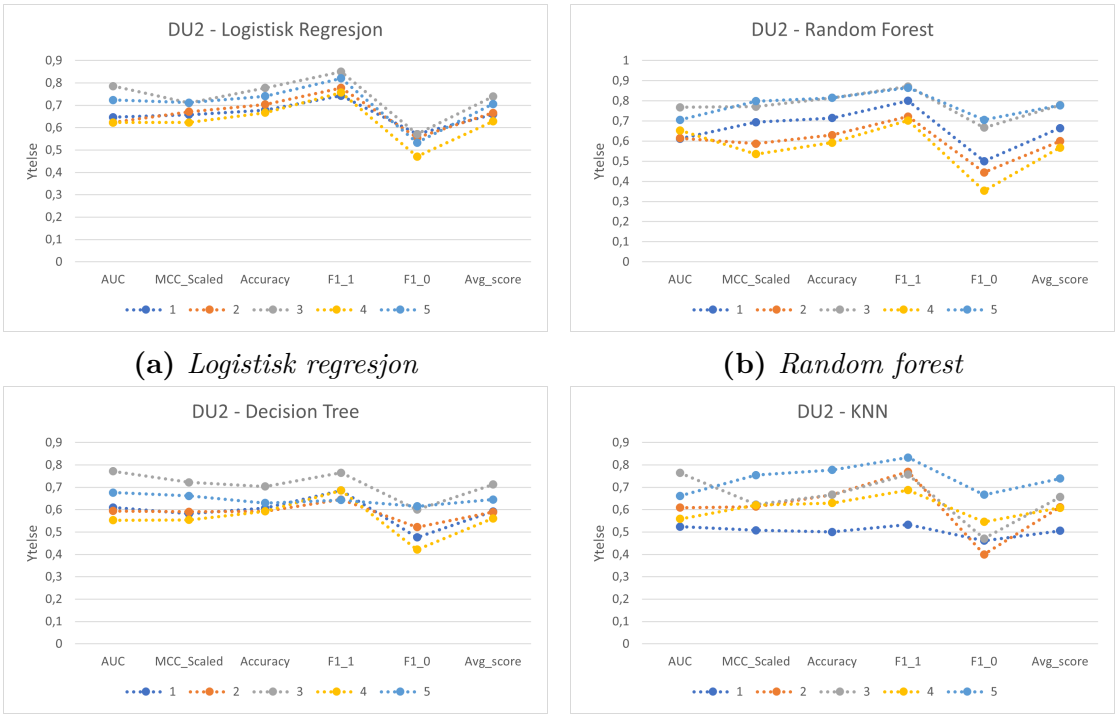


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.26: *Per fold resultater fra alle klassifiseringsalgoritmene for det kliniske datasettet DU1, for pasienter med negativ HPV-status.*

For DU1 i figur 4.26 er det for logistisk regresjon er jevnt ytelsesnivå for alle foldene, men fold 1 gjør det hakket bedre enn resten. Alle viser en tydelig nedgang i ytelse for F1:0. Både for random forest og KNN er foldene noe mer stabile gjennom metrikkene, utenom for fold 1 og 5 i random forest og fold 3 i KNN, som reduseres i ytelse ved F1:0. For decision tree har fold 4 en veldig ujevn ytelse, der F1:1 yter høyest med 0,71 og faller ned til 0,0 for F1:0.

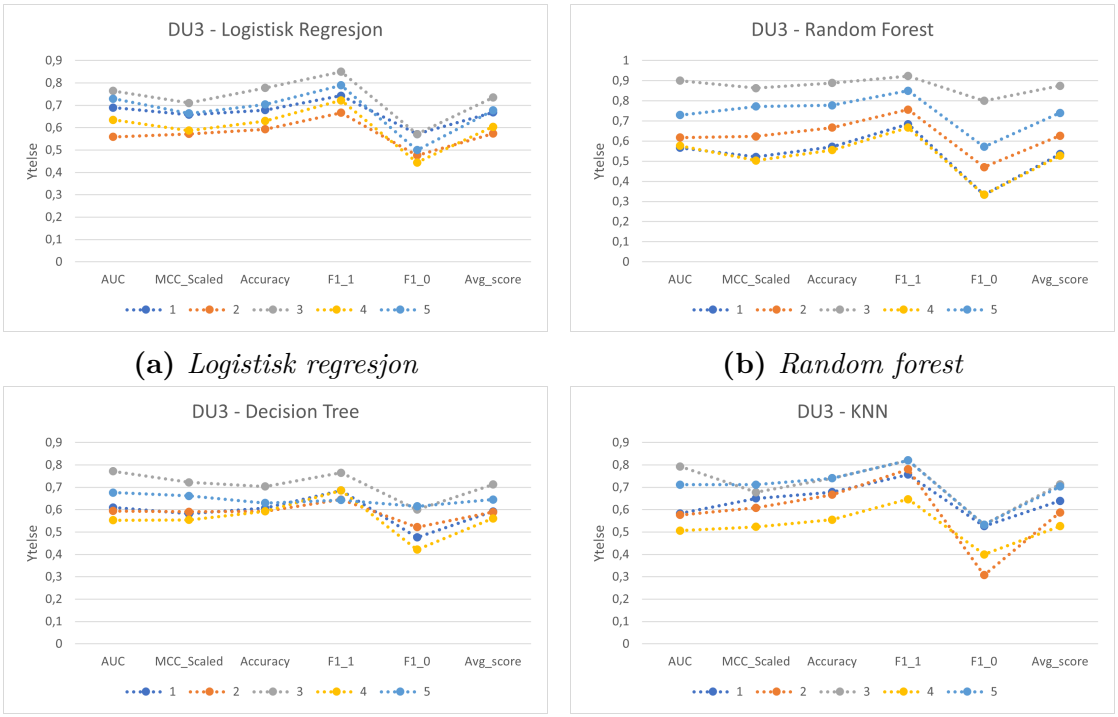


(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.27: *Per fold resultater fra alle klassifiseringsalgoritmene for radiomics datasettet, DU2, for pasienter med negativ HPV-status.*

Logistisk regresjon og decision tree i figur 4.27 over, har folder som er vedlig jevne og holder seg relativt stabile, med unntak av nedgang i ytelse for F1:0. For begge modellene ser man at samlet over alle metrikkene yter fold 3 høyest. Random forest og KNN har litt mer spredning i ytelsen for foldene. Fold 3 og 5 i random forest og fold 1, 4 og 5 i KNN har relativt stabile ytelser for metrikkene.



(a) *Logistisk regresjon* **(b)** *Random forest*

(c) *Decision tree* **(d)** *KNN*

Figur 4.28: *Per fold resultater fra alle klassifiseringsalgoritmene for kombinasjonsdatasettet, DU3, for pasienter med negativ HPV-status.*

Figur 4.28 viser resultatene fra foldene for DH3. Foldene, spesielt for logistisk regresjon og decision tree, er veldig like logistisk regresjon og decision tree for DH2, både hvor jevnt det er mellom ytelsen til foldene og stabiliteten over metrikkene.

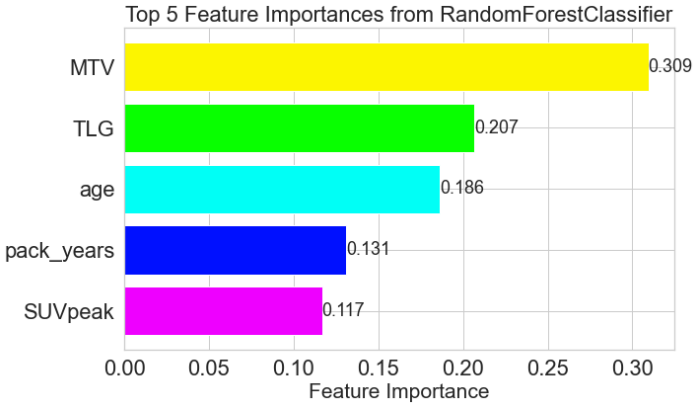
Det er som nevnt et klart mønster for F1:1 og F1:0 i forhold til HPV-relasjon for de gjennomsnittlige resultatene, og dette vises også i alle per fold resultater. Hvorfor blir tatt opp nærmere i diskusjonskapittelet.

4.4 Viktigheten av ulike egenskaper

Det blir i denne seksjonen presentert figurer som viser hvilke egenskaper som er blitt vektet tyngst i prediksjonene med random forest som klassifiseringsmodell. Verdien på x-aksen, indikerer den betydningen eller bidraget en egenskap har for beslutningene modellen tar. Topp fem egenskaper blir vist for det kliniske og kombinerte datasettet bestående av pasienter med positiv HPV-status, DH1 og DH3, og for det kliniske og kombinerte datasettet for pasienter med negativ HPV-status, DU1 og DU3. De blir vist med både sykdomsfri overlevelse (DFS) og generell overlevelser (OS) som responsvariabler.

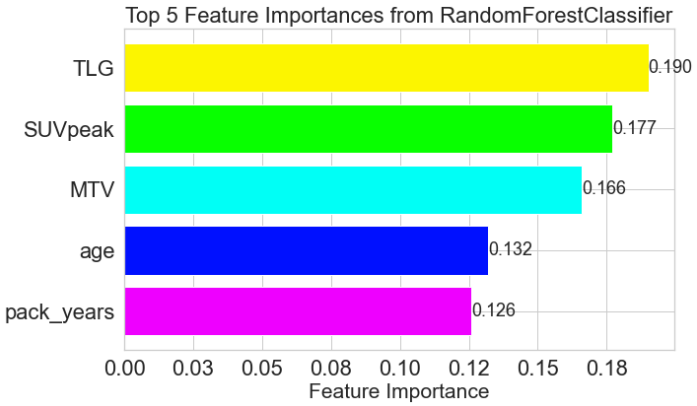
4.4.1 Viktige egenskaper der DFS er brukt som responsvariabel

I figur 4.29 blir topp fem vektet egenskaper under random forest som modell presentert for det kliniske datasettet DH1 med DFS som responsvariabel. Det er for datasettet *MTV* med verdi på 0,309, *TLG* med 0,207 og *age* med verdien 0,186 som er de tre egenskapene som bidrar mest til beslutningsprosessen.



Figur 4.29: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med DFS som respons på det kliniske datasettet DH1.*

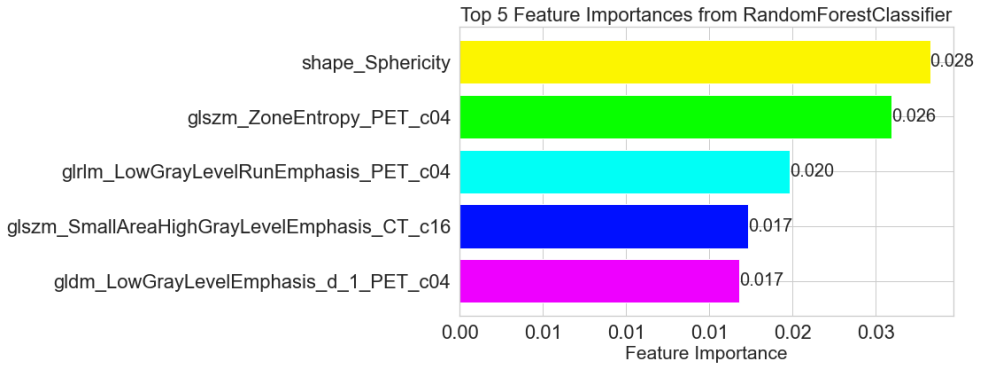
For datasett DU1 med DFS som respons, ser man i figur 4.30 at de tre øverste egenskapene *TLG*, SUV_{peak} og *MTV*, med verdier på 0,190, 0,177 og 0,166 har mest innflytelse på modellen.



Figur 4.30: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med DFS som respons på det kliniske datasettet DU1.*

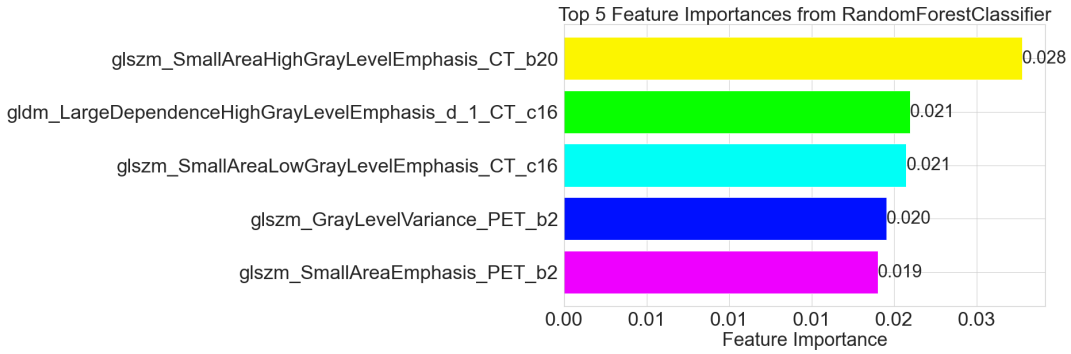
I figur 4.31 vises topp fem egenskaper som gir høyest bidrag for kombinasjonsdatasettet DH3 med DFS som responsvariabel. Det er for datasettet *shape_Sphericity* med verdi på 0,028, *glszm_ZoneEntropy_PET_c04* med 0,026

og *glrlm_LowGrayLevelRunEmphasis_PET_c04* med verdien 0,020 som er de tre øverste egenskapene.



Figur 4.31: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med DFS som respons på kombinasjonsdatasettet DH3.*

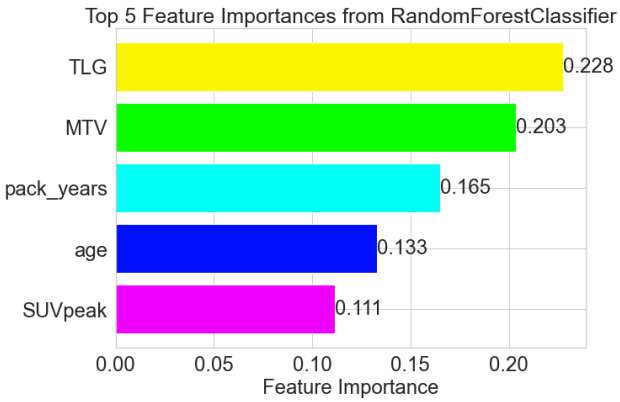
For datasett DU3 med DFS som respons, ser man i figur 4.32 at de tre øverste egenskapene *glszm_SmallAreaHighGrayLevelEmphasis_CT_b20* med verdi 0,028, *gldm_LargeDependenceHighGrayLevelEmphasis_d_1_CT_c16* med verdi 0,021 og *glszm_SmallAreaLowGrayLevelEmphasis_CT_c16* også med verdi på 0,021, har mest bidrag til modellens beslutninger.



Figur 4.32: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med DFS som respons på kombinasjonsdatasettet DU3.*

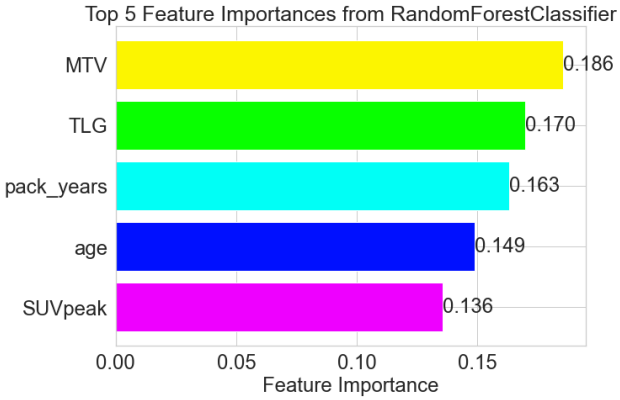
4.4.2 Viktige egenskaper der OS er brukt som responsvariabel

I figur 4.33 blir topp fem vektet egenskaper under random forest som modell presentert for det kliniske datasettet DH1 med OS som responsvariabel. Det er for datasettet *TLG* med verdi på 0,228, *MTV* med 0,203 og *pack_years* med verdien 0,165 som er de tre egenskapene som bidrar mest til beslutningsprosessen.



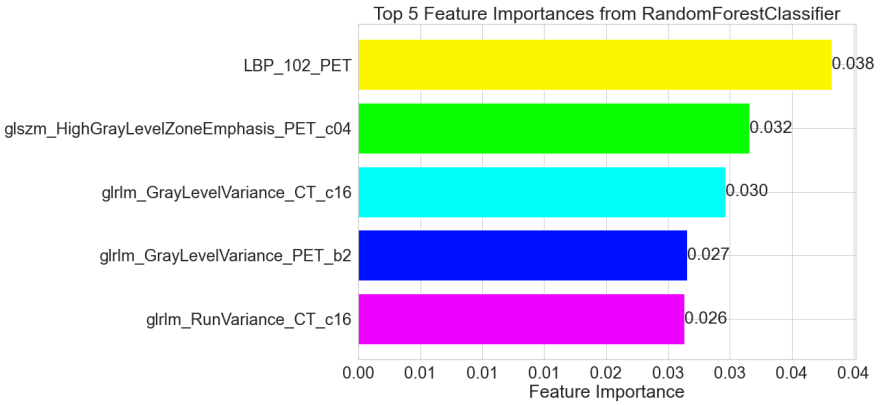
Figur 4.33: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med OS som respons på det kliniske datasettet DH1.*

For datasett DU1 med OS som respons, ser man i figur 4.34 at de tre øverste egenskapene *MTV*, *TLG* og *pack_years*, med verdier på 0,186, 0,170 og 0,163 har mest innflytelse på modellen.



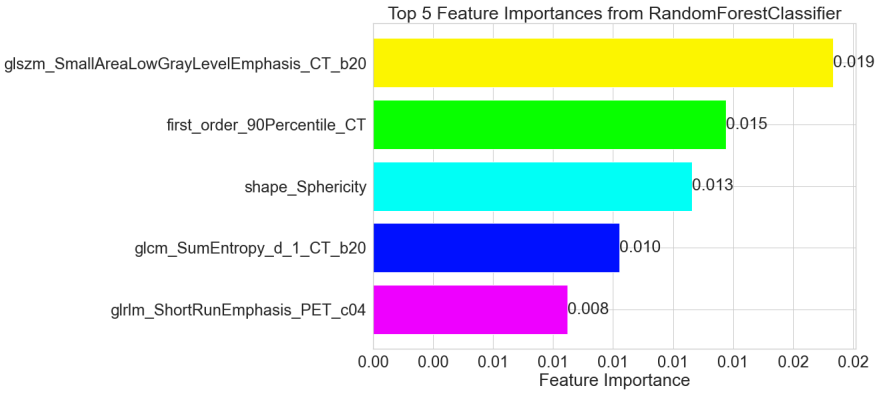
Figur 4.34: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med OS som respons på det kliniske datasettet DU1.*

For DH3 med OS som respons, ses det i figur 4.35 at de øverste egenskapene *LBP_102_PET*, *glszm_HighGrayLevelZoneEmphasis_PET_c04* og *glrlm_GrayLevelVariance_CT_c16*, med verdier på 0,038, 0,032 og 0,030, har mest bidrag til modellens beslutninger.



Figur 4.35: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med OS som respons på kombinasjonsdatasettet DH3.*

I figur 4.36 blir topp 5 egenskaper presentert for det kombinerte datasettet DU3 med OS som responsvariabel. De tre egenskapene som bidrar mest til beslutningsprosessen for datasettet er *glszm_SmallAreaLowGrayLevelEmphasis_CT_b20* med verdi på 0,019, *firs_order_90Percentile_CT* med 0,015 og *shape_Sphericity* med verdien 0,0,013.



Figur 4.36: *Figuren viser feature importance, altså viktigheten av de ulike egenskapene for klassifiseringen for random forest med OS som respons på kombinasjonsdatasettet DU3.*

Kapittel 5

Diskusjon

I denne oppgaven har tre datasett bestående av pasienter med hode- og halskreft vært benyttet. Disse ble delt inn i seks nye datasett basert på om pasienten har negativ eller positiv HPV-status. Målet var å finne ut om det var noen store forskjeller på ytelsesprediksjonene fra ML-modellene for pasienter med positiv HPV-status og pasienter med negativ HPV-status. Dette ble gjort med sykdomsfri overlevelse (DFS) og generell overlevelse (OS) som utfall for pasientene. I følgende kapittel skal resultatene presentert i kapittel 4 utforskes nærmere og diskuteres.

5.1 Datasettet

Datasettene brukt i oppgaven var pasientdata fra hode- og halskreft pasienter behandlet hos Oslo universitetssykehus (OUS) og Maastricht University Medical Center (MAASTRO). Pasientene ved OUS ble behandlet i perioden 2007 til 2013 og pasientene ved MAASTRO fra 2008 til 2014. Det er på årsbasis omtrent 38 000 krefttilfeller i Norge, hvor omtrent 800 var registrerte hode- halskrefttilfeller.

Datasettene brukt, spesielt etter de ble delt på egenskapen *hpv_related*, altså med hensyn på HPV-statusen til pasientene, var ikke helt balanserte. Det ble en ujevn fordeling av klassene 0 og 1 for både de positive og negative HPV-status datasettene, for begge responsvariablene sykdomsfri overlevelse (DFS) og generell overlevelse (OS). Denne ubalansen gjaldt også for klassene i egenskapen *uicc8_III-IV*, noe som kan ses i tabell 3.4 under seksjon 3.3.

Datasettene er ikke så oppdatert i oppfølgingsperioden som det kunne vært. Pasienter som dør av andre årsaker enn kreft, har ikke blitt tatt hennsyn til i datasettene. Dette kan gi en negativ innvirkning på modellene, da en pasient som har dødd av andre grunner kunne ha vært helt frisk og da fått tildet klasse som sykdomsfri/overlevende (klasse 0). Verdiene for de ulike egenskapene til denne pasienten, passer da muligens ikke helt overens med verdier for egenskapene til de som faktisk har kreft som dødsårsak. Dette kan føre til at modellen trenes opp feil. Det kan også påvirke ytelsen da, si en pasient som er tilhørende klasse 1, men har verdier i

egenskapene som om den skulle tilhørt klasse 0, da vil, selv om det egentlig skulle stemme, prediksjonen bli feilaktig predikert om denne pasienten blir klassifisert som klasse 0.

5.2 Evaluering av resultater

5.2.1 Sykdomsfri overlevelse (DFS) som respons

Evaluering av pasienter med positiv HPV-status fra resultatene i seksjon 4.2.2

Det vil først bli sett på de aggregerte ytelsesresultatene for datasettene DH1, DH2 og DH3, deretter vil resultatene for hver fold bli undersøkt nærmere.

Aggregerte resultater

For det kliniske datasettet bestående av kun pasienter med positiv HPV-status, DH1, ble det fra 4.2 og 4.5, funnet at random forest var modellen som hadde den høyeste ytelsen ut fra de fire klassifiseringsalgoritmene. Det var for AUC og MCC ytelsesverdier på under 0,5 for alle modellene utenom random forest. En ytelse på under 0,5 kan vurderes som en ren gjetting i prediksjonene.

For radiomics datasettet bestående av kun pasienter med positiv HPV-status, DH2, var det svært lite som skilte modellene. For nøyaktighet var ytelsesverdiene på mellom 0,63 og 0,68 og MCC verdier var 0,58 til 0,59, dette kan ses i tabell 4.2 og ut fra høydene i barplottet i figur 4.6. For kombinasjonsdatasettet bestående av kun pasienter med positiv HPV-status, DH3, var det KNN som ga den høyeste ytelsen av de fire klassifiseringsmodellene. Det var også for DH3 slik som for DH1, noen ytelsesverdier som lå rundt 0,5, igjen er dette ikke en god prediksjon.

Det var av alle modellene på alle tre datasettene, klassifiseringsalgoritmen KNN på kombinasjonsdatasettet, DH3, som hadde den høyeste ytelsen over alle de fem metrikkene. Dette var også den modellen som presterte jevnt høyest over alle datasettene, med random forest rett bak i ytelse. Det var radiomics datasettet som hadde den høyeste totale ytelsen, sett på alle modellene til sammen. Disse modellene leverte også veldig jevne resultater for dette datasettet, noe som også kan ses i figurene 4.5, 4.6 og 4.7.

I forhold til de aggregerte ytelsesresultatene på de orginale datasettene, uten deling på HPV-status, presentert under seksjon 4.2.1, i tabell 4.1, var metrikken nøyaktighet den som hadde minst reduksjon i ytelse. Faktisk var ytelsen noe høyere for DH2 enn den var for det udelte radiomics datasettet D2. For alle datasettene bestående av pasienter med positiv HPV-status, ble ytelsen for resten av metrikkene, utenom F1:0, redusert for alle modellene, sett bort fra KNN på DH2. For F1:0 øker ytelsen for alle modellene på alle de tre datasettene. Dette betyr at etter deling på HVP-status, har den gjennomsnittlige ytelsen gått noe ned, og fått en mer ubalansert prediksjon på F1-scoren.

En viktig ting som ikke er nevnt, er de store forskjellene mellom F1:1 og F1:0 metrikkene. Det blir for både det kliniske-, radiomics- og kombinasjonsdatasettet en høy ytelse på F1:0, med alle ytelsesverdier over 0,70. Det er for F1:1 derimot helt motsatt, hvor den får lav ytelse for alle datasettene, der DH1 har tydelig lavest ytelse for alle modellene. Den høyeste ytelsen for denne metrikken ligger på 0,42 for decision tree på datasett DH2. Hvorfor ytelsen for F1-score er slik, vil bli tatt opp og diskutert under seksjonen med evalueringen av resultatene for pasienter med negativ HPV-status med DFS som respons.

Per fold resultater

I per fold resultatene for det kliniske datasettet DH1, i figur 4.8, ble det sett at det var en god del spredning i ytelsen for de ulike foldene, spesielt for klassifiseringsalgoritmen decision tree. For random forest og KNN er det en fold, som gjør det markant bedre, og det er fold 1. Fold 1 har for alle modellene den høyeste ytelsen, og det kan da virke som at pasientene delt inn i denne folden er enklere å predikere enn for de andre foldene. For alle klassifiseringsalgoritmene er det minst to av de fem foldene som har en ytelsesverdi på 0 for F1:1. Disse lave ytelsesverdiene, spesielt for logistisk regresjon og KNN, viser grunnen til de lave aggregerte ytelsesresultatene for F1:1 på datasett DH1 i figur 4.5 også.

Per fold resultatene for radiomics datasettet DH2 i figur 4.9, viste at foldene for decision tree og KNN hadde en mer jevn ytelse enn foldene for logistisk regresjon og random forest. Det er også for DH2, to folder som har ytelse 0 på F1:1, én for logistisk regresjon og én for random forest. Det er ganske like resultater for per fold for DH3 i figur 4.7, der modellene KNN og decision tree gjør det jevnere enn logistisk regresjon og random forest. Igjen som på DH2, har modellene logistisk regresjon og random forest en fold på ytelsesverdi 0 på F1:1, og alle foldene for alle modeller yter høyt for F1:0.

Det er for alle tre datasett, DH1, DH2 og DH3, på resultater per fold, likt som de aggregerte resultatene for de samme datasettene. For alle foldene i alle modeller, reduseres ytelsen i folden ved F1:1, for flesteparten av foldene er det lav eller null ytelse, mens det for F1:0 er høy ytelse for hver fold. Grunnen til disse veldig ulike ytelsesresultatene i F1-score, blir, som nevnt, tatt opp på slutten av seksjonen under.

Evaluering av pasienter med negativ HPV-status fra resultatene i seksjon 4.2.3

Det vil først bli sett på de aggregerte ytelsesresultatene for datasettene DU1, DU2 og DU3, deretter vil resultatene for hver fold bli undersøkt nærmere.

Aggregerte resultater

Både for det kliniske datasettet DU1, radiomics datasettet DU2, og det komibnerte datasettet DU3, var det logistisk regresjon som gav den høyeste ytelsen av de

fire klassifiseringsalgoritmene, dette kan ses i tabell 4.3 og figurene 4.11, 4.12 og 4.13. Logistisk regresjon oppnådde ytelsesverdier på 0,73 (DU1), 0,71 (DU2) og 0,67 (DU3) for nøyaktighet, og for MCC var ytelsene på 0,63 (DU1), 0,60 (DU2) og 0,61 (DU3). Det ser ikke ut som, ut fra de aggregerte resultatene, at det er ett datasett som er bedre å predikere på i forhold til de andre, men ser man på summen av gjennomsnittet for alle modellene, er det med en liten margin predikert høyest på DU3. Den eneste merkbare forskjellen i ytelse på både datasettene, og modellene, er på metrikken F1:0, der decision tree og logistisk regresjon ser ut til å klare prediksjonen på denne metrikken litt bedre enn random forest og KNN. Likt på alle datasettene er at F1:1 har høy ytelse for de fire klassifiseringslagoritmene, mens F1:0 yter veldig lavt.

I forhold til de aggregerte ytelsesresultatene på de orginale datasettene D1, D2 og D3, uten deling på HPV-status, presentert under seksjon 4.2.1, i tabell 4.1, gjør det kliniske datasettet DU1, for pasienter med negativ HPV-status, det veldig likt som D1 på ytelse for nøyaktighet. Ytelsesverdiene for AUC reduseres derimot litt fra D1 til DU1, der ingen modeller for DU1 er over 0,68 i ytelse, den samme reduseringen gjelder for MCC. Radiomics datasettet, DU2, sett opp mot det udelte radiomics datasettet, D2, har en høyere ytelse for nøyaktighet, DU3 mot D3 er det som for DH1 og D1, ganske jevne ytelser for nøyaktighet. Også for DU2 og DU3, likt DU1, var det lavere ytelse for MCC enn i D2 og D3.

Per fold resultater

I per fold resultatene for det kliniske datasettet DU1, i figur 4.14, viste det at det var litt variasjon for foldene i de fleste modellene, men for random forest var de veldig jevne utenom for AUC og MCC. I resultatene for per fold på DU2, sett i figur 4.15, var det veldig jevnt for de fem foldene i logistisk regresjon. Fold 1, 2, 3, 4 og 5 innenfor hver av de andre modellene var også stabile i forhold til hverandre. Alle fem folder i en modell følger det samme mønsteret for hvilken metrikk som yter høyt og lavt. Per fold resultatene for DU3 presenteres i figur 4.16. Her ble det sett at random forest hadde lite variasjon for de fem foldene. Det er for alle foldene i F1:0 ytelsen går ned. Fra de fire klassifiseringsalgoritmene på de tre datasettene, ser det ut til å være random forest og logistisk regresjon som har lavest spredning på de fem foldene.

Fra teorikapittelet i seksjon 2.3.2 under F1-score, ses det at F1-score tar hensyn til presisjon og tilbakekalling, der presisjonen er andelen av positive prediksjoner som er korrekt predikert sanne positive og tilbakekalling er andelen sanne positive som modellen predikerte riktig. Ut fra uttrykket for å beregne F1-score, kan man se at når F1:1 yter lavt, så vil det si at en større andel av pasientene med den positive klassen (klasse 1, altså F1:1), er blitt feilaktig predikert som negative (klasse 0, F1:0). Det samme gjelder for de resultatene der F1:0 yter lavt, da har en større andel av den negative klassen, blitt feilaktig predikert som positive. På grunn av hensynet til presisjon og tilbakekalling, vil det ved ubalanserte klassefordelinger gi modellen høy ytelse kun ved å predikere majoritetsklassen. Dette medfører ofte lav tilbakekalling for minoritetsklassen, som for denne klassen resulterer i en lav

F1-score.

I tabell 3.4 vises distribusjonen for klassene 0 og 1 i datasettene delt på HPV-status. For DFS som respons på datasettene for pasienter med positiv HPV-status, utgjør klasse 0 69,61% av utfallene, med 71 pasienter, mens klasse 1 da utgjør 30,39% med 31 pasienter. Dette gir en stor ubalanse for klassefordeling i favør av klasse 0. I datasettene for pasienter med negativ HPV-status og DFS som respons, utgjør klasse 0 29,41% med 40 pasienter og klasse 1 utgjør 70,59% av utfallene med 96 pasienter. Denne fordelingen er også veldig ubalansert, men i motsatt vei enn for DH-datasettene. Det er altså til fordel for klasse 1.

På datasettene for pasienter med positiv HPV-status er F1:1 en minoritetsklasse, siden klasse 1 da er underrepresentert kan dette muligens påvirke evnen modellen har til å lære fra denne klassen. Det er en mulighet for at modellen kan bli overtilpasset majoritetsklassen, som i dette tilfellet er klasse 0. Dette er på grunn av tilgangen til flere eksempler å lære fra. Dette medfører at modellen kan få lavere presisjon i å forutsi den mindre representerte klassen. På datasettene for pasienter med negativ HPV-status blir det likt, bare der er F1:0 minoritetsklassen og F1:1 majoritetsklassen.

Det var i per fold resultatene generelt større variasjoner i de fem foldene, for de fire modellene, på datasettene for pasienter med positiv HPV-status, enn det var på datasettene for pasienter med negativ HPV-status. Det kan i mange av modellene for de ulike datasettene, ses at ytelsen for metrikkene ikke er veldig konsekvente over de fem foldene. Dette vil vanligvis tyde på at det er en overtilpasning eller en skjev fordeling i datasettet. Da man fra tabell 3.4 har sett at det er en ubalanse i datasettet, gir også disse variasjonene, mellom de fem ulike foldene for en enkelt modell mening. For å få en mer robust og stabil modell, bør det først være et balansert datasett. Det er også for noen av modellene der foldene er konsekvente, dette kan skyldes at modellen er mer robust mot forskjeller i treningsdataen, eller at modellen ikke er sensitiv for egenskaper som er ujevnt fordelt eller overrepresentert, og dermed har jevnere ytelse.

5.2.2 Generell overlevelse (OS) som respons

Evaluering av pasienter med positiv HPV-status fra resultatene i seksjon 4.3.1

Det vil også her først bli sett på de aggregerte ytelsesresultatene for datasettene DH1, DH2 og DH3, deretter vil resultatene for hver fold bli undersøkt nærmere.

Aggregerte resultater

For det kliniske datasettet bestående av kun pasienter med positiv HPV-status, DH1, kunne det fra tabell 4.4 ses at det var klassifiseringsmodellen random forest som hadde høyest ytelse. Random forest hadde også den høyeste modellytelsen for radiomics datasettet bestående av kun pasienter med positiv HPV-status, DH2. For DH1 var ytelsen fra random forest på nøyaktighet på 0,79 og MCC ytelsen var

0,61. For DH2 var nøyaktighet på 0,78 og MCC på 0,61. Dette viser til at random forest ikke presterte så ulikt på disse to datasettene. Det var for kombinasjons-datasettet bestående av kun pasienter med positiv HPV-status, DH3, KNN som presterte best av de fire klassifiseringsalgoritmene. Med KNN på DH3 var det for nøyaktighet en ytelse på 0,78 og MCC på 0,61. KNN presterte dermed like godt på DH3 som random forest gjorde på DH1 og DH2, men over alle datasettene gjorde random forest det høyest i ytelse totalt. Det var også prediksjonen på radiomics datasettet, DH2, som var høyest i ytelse basert på summen av gjennomsnittene til de fire modellene.

Det ble vist tydelig i figurene 4.17, 4.18 og 4.19 at F1:0 ytelsen var høyest og det var helt klart F1:1 som presterte lavest på alle datasettene. Dette var også tilfellet på datasettene for pasienter med positiv HPV-status med DFS som respons, og forklaringen på dette er beskrevet tidligere. Igjen, som det også var på datasettene for pasienter med positiv HPV-status med DFS som respons, er det random forest og KNN som leverer de høyeste ytelsene.

Ytelsen til de fire klassifiseringsalgoritmene på datasettene DH1, DH2 og DH3 blir sammenlignet med resultatene fått for trening på OUS datasett og testing på MAASTRO datasett i Huynh et al. [14] som kan ses i *Supplementary Table F2*. Det er for DH1 ikke så ulik ytelse for nøyaktighet sammenlignet med det Huynh et al.[14] fikk for deres datasett med kliniske egenskaper *D1*. For MCC og AUC derimot, er det en mye lavere ytelse for modellene i DH1. Også for DH2 og DH3 er ytelsen på nøyaktighet like bra, og for random forest er ytelsen på nøyaktighet høyere enn ytelsen fått i Huynh et al. [14]. Det er likt for DH2 og DH3 som for DH1, ytelsen i AUC og MCC er lavere enn ytelsen til sammenligning. En grunn til at nøyaktigheten holder seg høyere på ytelse kan være at den ikke i like stor grad blir påvirket av ubalansen for klassen i datasettet, da den bare måler den samlede andelen korrekte prediksjoner, altså TP (sann positiv) og TN (sann negativ). AUC og MCC er mer sensitive for denne ubalansen, siden de i tillegg til TP og TN er følsomme overfor utfall med FP (falske positive) og FN (falske negative), dette kan være en av grunnene til at ytelsen blir så mye lavere enn nøyaktigheten. F1-score for de to klassene blir ikke sammenlignet, da de er veldig annerledes grunnet den skjeve klassefordelingen.

Per fold resultater

Per fold resultater for DH1, ble vist i figur 4.20. Det var ingen modeller som utmerket seg her med jevne og stabile folder, det var veldig variert gjennom de ulike metrikkene og en del spredning mellom foldene i de fire modellene. I figur 4.21 for per fold resultatene på DH2, er det store spredninger for foldene i modellene, men decision tree klassifiseringen har mest stabile folder. På DH3 i figur 4.22 er ytelsene per fold også veldig ujevne. Det var decision tree som hadde fold 2 som hadde en jevnt over stabil og høy ytelse, mens det i samme modell var en fold som hadde lav ytelse, sett bort fra F1:0, dette var fold 4.

Det har vist seg at på datasettene for pasienter med positiv HPV-status, er ytelsen for F1:1 veldig lav, mens den for F1:0 er jevnt over høy. Det er ikke

noe forskjell i per fold resultatene på DH1, DH2 og DH3, der det for alle klassifiseringsalgoritmene, på alle datasettene, er minst en fold med ytelse lik null for F1:1, med et unntak da decision tree på DH2 ikke har noen folder helt nede på ytelse 0. En forklaring på hvordan den ubalanserte klassefordelingen påvirker ytelsen i F1-score for de to klassene, er gitt tidligere i diskusjonen.

Evaluering av pasienter med negativ HPV-status fra resultatene i seksjon 4.3.2

Det blir igjen først sett på de aggregerte ytelsesresultatene for datasettene DU1, DU2 og DU3, deretter vil resultatene for hver fold bli undersøkt nærmere.

Aggregerte resultater

For det kliniske datasettet bestående av kun pasienter med negativ HPV-status, DU1, ble det fra 4.5 og 4.23, funnet at decision tree og logistisk regresjon hadde den høyeste gjennomsnittlige ytelsen over de fem metrikkene. For radiomics datasettet bestående av kun pasienter med negativ HPV-status, DU2, var den klassifiseringsalgoritmen som hadde høyest ytelse, logistisk regresjon, mens det for kombinasjonsdatasettet DU3, var både logistisk regresjon og random forest som hadde høyest, og like ytelser. Det er over alle datasettene, logistisk regresjon som presterer litt bedre enn de tre andre klassifiseringsalgoritmene. Det var for de tre metrikkene AUC, MCC og nøyaktighet, på alle tre datasett, ikke veldig store endringer, det eneste var at decision tree og KNN varierte litt i ytelse for de forskjellige datasettene på disse metrikkene. Sett på summen av hver modells gjennomsnitt over de fem metrikkene, var det høyeste predikerte ytelsen på DU3, deretter DU2 og den laveste var på DU1.

Det blir som for datasettene bestående kun av pasienter med positiv HPV-status med DFS og OS som respons, der de for begge responser har lave ytelsesverdier for F1:1 og høye ytelser for F1:0. Så for datasettene her, bestående av pasienter med negativ HPV-status og OS som respons, er det høye ytelser for F1:0 og lavere ytelser for F1:1. Dette er likt som for med responsen DFS på samme datasett (DU-settene). Det kan ut fra figurene som viser de aggregerte ytelsesresultatene på datasettene bestående av pasienter med positiv HPV-status (DH-settene), både for DFS og OS som responsvaribel, ses at når det predikeres på DH2-datasettet vil det være høyere ytelse for F1:1 metrikken enn på DH1 og DH3 datasettene. Dette kan være tilfedlig, men da det skjer for begge responsene, kan det kanskje tenkes at DH2 har noen egenskaper som påvirker modellens ytelse positivt for F1:1, ved at noen egenskaper muligens har mer relevanse for å kunne skille mellom klassene.

Tilbake til bare datasettene bestående av kun pasienter med negativ HPV-status, DU1, DU2 og DU3, med OS som responsvariabel, så sammenlignes også disse med resultatene fra *Supplementary Table F2* i Huynh et al. [14]. DU1 yter en del lavere for både nøyaktighet, AUC og MCC sammenlignet med resultatene for *D1* i *Supplementary Table F2*. DU2 sammenlignet med *D2* gjør det for logistisk regresjon noe lavere i ytelse på nøyaktighet, AUC og MCC. Random forest derimot,

på DU2 har en høyere ytelse på MCC og nøyaktighet enn de ytelsene fått for *D2* i Huynh et al. [14]. DU3 blir sammenlignet med resultatene på *D1 + D2*, og der er ytelsen til DU3 for AUC vesentlig lavere, og MCC også noe lavere. For random forest på nøyaktighet er ytelsen litt høyere.

Per fold resultater

I per fold resultatene for DU1, sett i figur 4.26, hadde logistisk regresjon en jevn ytelse over de fem foldene. Det var decision tree som skilte seg ut fra resten av modellene, da to av foldene (2 og 4) hadde store spredninger fra de tre resterende foldene. Noe som er interessant var at det fra de aggregerte ytelsesmålingene for DU1 var decision tree som presterte best samlet over alle metrikkene. I figur 4.27 ble resultatene per fold for DU2 presentert. Det ble funnet at logistisk regresjon og decision tree hadde de jevneste foldene, som var ganske stabile utenom i F1:0, der ytelsen avtok noe. Det ble funnet at det for ytelsen og variasjonen i foldene for de fire modellene på DH3, ikke var så ulikt fra foldene i modellene på DH2. Så det var også i DH3 logistisk regresjon og decision tree som hadde mest stabile og jevne folder.

Noe som kommer frem fra resultatene er at datasettene bestående av kun pasienter med negativ HPV-status og med OS som responsvariabel, ikke faller like mye i ytelse på F1-score klassen som er en minoritet, her F1:0. Dette vil si at det for DU1, DU2 og DU3 med OS respons, ikke er så lav ytelse for F1:0, som de samme datasettene med DFS som respons, eller ytelsen for F1:1 på DH-datasettene med DFS eller OS som respons. Ser man på distribusjonen til de to klassene for DFS og OS i tabell 3.4, finner man at DU-datasettene med OS som respons, har den minst ubalanserte fordelingen. Dette gir da modellene en større mulighet til å klassifisere minoritetsklassen bedre.

5.2.3 Mulige forklaringer på modellytelsen

I *Table 3* og *Table 4* i Huynh et al. [14] ser man egenskapene valgt ut flest ganger av RENT (repeated elastic net technique) [39], [40] for DFS og OS som responsvariabel, og som da er viktige for prediksjonsytelsen. For DFS på de kliniske datasettene DH1 og DU1, er det *hpv_related* og *TNM8 stage*, som er det samme som *uicc8_III-IV* i denne oppgaven, som har den høyeste utvalgte frekvensen med 98% og 89%. Det kan da diskuteres om det å dele datasettet på HPV-status fjerner en veldig viktig egenskap for prediksjonen, og dermed vil redusere ytelsen. For kombinasjonsdatasettene DH3 og DU3 med DFS som respons, tilsvarende *All tabular data. Clinical factors D1 + radiomics features D2* i *Table 3* i Huynh et al. [14], er de fire høyeste frekvensene for egenskaper på radiomics egenskaper, og *hpv_related* og *uicc8_III-IV* ligger under dette med frekvenser på 55% og 47%. Ut fra dette kan det diskuteres om prediksjonen for de kliniske datasettene blir mer påvirket av delingen på HPV-status enn prediksjonene for kombinasjonsdatasettene.

For OS på de kliniske datasettene DH1 og DU1, er det *uicc8_III-IV* (merk, står som *TNM8 stage* i *Table 4* i Huynh et al. [14]) og *hpv_related* med frekvenser på 100% og 95% som er høyest. Igjen vil dette ha innvirkning på prediksjonen

for dissse datasettene som er delt på HPV-status. For *All tabular data. Clinical factors D1 + radiomics features D2* i *Table 4* i Huynh et al. [14], er det med OS respons *uicc8_III-IV* og *hpv_related* som ligger som nummer to og tre med frekvenser på 88% og 86%. Det vil si at for kombinasjonsdatasettene, DH3 og DU3, er dette viktige egenskaper. RENT velger altså ut HPV-statusen og kreftstadiet til pasientene med høyere frekvens for OS som respons enn for DFS.

Det er viktig å merke at disse frekvens-verdiene er for bare OUS datasettene. Det kan være en mulighet for at det er andre egenskaper enn disse som veier tyngre for prediksjon i datasettene fra MAASTRO.

Det blir under seksjon 4.4 presentert hvilke egenskaper som bidrar mest til random forest modellens beslutningsprosess for de kliniske datasettene DH1 og DU1 og kombinasjonsdatasettene DH3 og DU3 for både DFS og OS som responsvariabel. Det er for random forest på DH1 med DFS som respons, to PET-parametere, MTV og TLG, øverst på «feature importances» (viktige egenskaper), og ikke kliniske faktorer som man kan se i figur 4.29. Man kan se i figur 4.30 for DU1 med DFS respons, figur 4.33 for DH1 med OS som respons og i figur 4.34 for DU1 med OS respons, at det for disse kliniske datasettene også er PET-parametere som har høyest betydning for modellens beslutninger. Sett opp mot de egenskapene som ble valgt ut med RENT i Huynh et al. [14], er disse ulike. Da datasettene i denne oppgaven er splittet på HVP-status, kan ikke denne egenskapen bidra i modellenes beslutninger. Selv om kreftstadiet, *uicc8_III-IV*, til pasientene var viktig i Huynh et al., er ikke denne egenskapen i topp fem for noen av de kliniske datasettene, uavhengig av responsvariabel. Dette betyr at beslutningene til modellen blir vektet mer på egenskaper likere de i radiomics datasettene.

I de fire figurene 4.31, 4.32, 4.35 og 4.36, vises kombinasjonsdatasettene DH3 og DU3 med DFS og OS som responsvariabel. Likt for alle disse fire, er at det ikke er en klinisk faktor i topp fem for noen av dem, og radiomics egenskapene har også her større innflytelse på modellen.

Utenom for den motsatte høye/lave ytelsen i F1-score, er det vanskelig å se noen stor eller signifikant forskjell i modellytelsen på datasett bestående av pasienter med positiv HPV-status og med negativ HPV-status. Det er kanskje ikke sånn at de to grupperingene på HPV-status er så ulike i informasjonen fra egenskapene i klinisk og radiomics datasett. Det samme gjelder for de ulike datasettene med de to responsvariablene, da det ikke er ett datasett som skiller seg ut som helt klart best å predikere på. Radiomics datasettet har en minimalt høyere ytelse enn det kliniske datasettet sett på totalen over alle predikeringer og responsvariabler brukt. Det at det ikke er så stor forskjell i modellytelsen kan være at modellene har vanskeligere for å kjenne igjen og lære seg mønster fra et av datasettene, eller begge, eller kanskje at de forskjellige egenskapene i datasettene (kliniske faktorer og radiomics egenskaper) bidrar relativt likt til modellytelsen.

5.3 Videre arbeid

Til videre arbeid burde de delte datasettene bli delt i fire folder under kryssvalidering og ikke fem, da det blir få pasienter i hvert av datasettene etter de er delt inn etter HPV-status. Dette vil gi modellen flere pasienter i hver fold og dermed mer treningsdata. Det er mulig modellytelsen kunne blitt økt av dette.

En annen ting som kunne vært gjort var egenskapsseleksjon. Da radiomics- og det kombinerte datasettet var på over 300 egenskaper, kan dette ha vært med på å lage «støy» for modellen. Ved å bruke RENT eller annen metode for å velge ut de egenskapene som bidrar mest til modellens beslutningsprosses, kunne man økt modellytelsen. Videre bør det også ses på PCA-analysen en gang til, og vurdere igjen om ekstremverdier burde bli fjernet fra datasettet.

For å takle den ubalanserte klasseforskjellen ble det prøvd med å sette parameteren *class_weight* for modellen lik «balansert», men ut fra tidsrommet i denne oppgaven ble det ikke tid til å analysere resultatene fra dette. Det å sette parameteren lik balansert, gjør at når modellen gjør en feilaktig klassifisering på minoritet klassen, vil den straffes [23], vektene for hver klasse vil bli automatisk beregnet av scikit-learn og minoritetsklassen vil bli vektet høyere, mens majoritetsklassen blir vektet lavere. Også en annen teknikk til å håndtere ujevne klassefordelinger, som er nevnt i boken av Raschka og Mirjalili [23], er SMOTE (Synthetic Minority Oversampling Technique). Denne metoden kunne blitt testet på dette datasettet som videre arbeid.

Det var mye spredning og ujevne folder i per fold resultatene, en interessant ting som kunne blitt sett på, er hvilke pasienter som har blitt feilaktig klassifisert. Dette ville gitt en bedre innsikt i da hvilke pasienter som ble oftest feilklassifisert og hva grunnene til dette kunne være (for eksempel om det var noe i egenskapene som skilte seg ut).

Kapittel 6

Konklusjon

Denne oppgaven har fokusert på datasett med to ulike grupperinger for pasienter med hode- og halskreft: den ene bestående av kun pasienter med en positiv HPV-status, og den andre bestående av kun pasienter med en negativ HPV-status. Det ble benyttet fire ulike klassifiseringsalgoritmer for prediksjon av behandlingsutfallet til pasienter. Dette ble gjort på seks datasett, for to ulike responsvariabler: sykdomsfri overlevelse og generell overlevelse. Et mål var å se om det var noen betydelige forskjeller i modellytelsen for de datasettene bestående av pasienter med positiv HPV-status og datasettene bestående av pasienter med negativ HPV-status.

For sykdomsfri overlevelse (DFS) på datasett bestående av pasienter med positiv HPV-status, ble den høyeste ytelsen basert på alle tre datasett oppnådd av klassifiseringsalgortimen KNN. Det datasettet det ble predikert den høyeste ytelsen på totalt av de fire modellene, var radiomics datasettet DH2.

For sykdomsfri overlevelse (DFS) på datasett bestående av pasienter med negativ HPV-status, var det klassifiseringsalgoritmen logistisk regresjon som oppnådde den høyeste ytelsen basert på alle de tre datasettene. Det var kombinasjonsdatasettet DU3, som det ble predikert den høyeste ytelsen på totalt av de fire modellene.

For generell overlevelse (OS) på datasett bestående av pasienter med positiv HPV-status, ble den høyeste ytelsen basert på alle tre datasett oppnådd av klassifiseringsalgortimen random forest. Det datasettet det ble predikert den høyeste ytelsen på totalt av de fire modellene, var radiomics datasettet DH2.

For generell overlevelse (OS) på datasett bestående av pasienter med negativ HPV-status, var det klassifiseringsalgoritmen logistisk regresjon som oppnådde den høyeste ytelsen basert på alle de tre datasettene. Det var kombinasjonsdatasettet DU3, som det ble predikert den høyeste ytelsen på totalt av de fire modellene.

De datasettene som det ble oppnådd de høyeste totale ytelsene på, var dermed radiomics datasettet og kombinasjonsdatasettet. Klassifiseringsalgoritmene som presterte best over alle tre datasettene minst én gang, var da logistisk regresjon, random forest og KNN. Det kliniske datasettet og klassifiseringsalgoritmen deci-

sion tree gjorde det aldri best over totalen av ytelse på datasettene eller modellene. Det skal sies at når det snakkes om hvilket datasett eller hvilken klassifiseringsalgoritme som har gjort det best, er det snakk om med veldig lav margin til resten. Hvis man skal se etter hvilket datasett og fra hvilken HPV-status som hadde den høyeste totale ytelsen, selv om det er med de minste marginer, var dette kombinasjonsdatasettet bestående av pasienter med negativ HPV-status med generell overlevelse (OS) som respons.

Det var ingen så stor signifikant forskjell i prediksjonsytelsen at det kan sies helt bestemt hvilken av gruppene delt på HPV-status, hvilket datasett eller hvilken klassifiseringsalgoritme som er den/det beste. For å vite om pasienter med positiv HPV-status eller pasienter med negativ HPV-status har noen egne karakteristikker eller mønster for egenskapene, er det essensielt at dette og modellene blir testet ut på flere datasett.

Bibliografi

[1] Kreftforeningen. Hva er kreft? [internett], (Oppdatert 21.12.2023, hentet: 10.02.2024). URL https://kreftforeningen.no/om-kreft/hva-er-kreft/.

[2] Kreftforeningen. Undersøkelser [internett], (hentet: 10.02.2024). URL https://kreftforeningen.no/om-kreft/undersokelser/.

[3] Folkehelseinstituttet. Kreft i norge [internett], (Oppdatert 03.10.2023, hentet: 10.02.2024). URL https://www.fhi.no/he/folkehelserapporten/ikke-smittsomme/kreft/?term=.

[4] Helsedirektoratet. Epidemiologi og risikofaktorer [internett], (Oppdatert 27.11.2023, hentet: 11.02.2024). URL https://www.helsedirektoratet.no/retningslinjer/hode-hals-kreft-handlingsprogram/epidemiologi-og-risikofaktorer.

[5] Folkehelseinstituttet. Dette døde nordmenn av i 2022 [internett], (Publisert 08.06.2023, hentet: 11.02.2024). URL https://www.fhi.no/nyheter/2023/dodelighet-2022/.

[6] Oslo universitetssykehus. Hode- og halskreft [internett], (hentet: 11.02.2024). URL https://www.oslo-universitetssykehus.no/behandlinger/hode-og-halskreft.

[7] Kreftforeningen. Hode- og halskreft [internett], (oppdatert torsdag 23. februar 2023, hentet: 11.02.2024). URL https://www.helsenorge.no/sykdom/kreft/hode-og-halskreft/.

[8] Kreftforeningen. Hpv, hpv-vaksine og kreft [internett], (Oppdatert 29.01.2024, hentet: 11.02.2024). URL https://kreftforeningen.no/forebygging/hpv-og-kreft/.

[9] St. Olavs hospital HF. Behandling når kreftsykdommen ikke kan helbredes [internett], (hentet: 11.02.2024). URL https://www.stolav.no/behandlinger/behandling-nar-kreftsykdommen-ikke-kan-helbredes.

[10] Kreftforeningen. Ny kreftsykdom (sekundær kreft) [internett], (Oppdatert 11.08.2023, hentet: 11.02.2024). URL https://kreftforeningen.no/om-kreft/senskader-voksne/ny-kreftsykdom-sekundaer-kreft/.

[11] Kreftforeningen. Strålebehandling [internett], (Oppdatert 16.11.2023, hentet:

11.02.2024). URL https://kreftforeningen.no/om-kreft/behandling/stralebehandling/.

[12] Kreftforeningen. Strålebehandling ved kreft [internett], (torsdag 8.09.2022, hentet: 11.02.2024). URL https://www.helsenorge.no/sykdom/kreft/stralebehandling/.

[13] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*, (64), 2019. doi: 10.1186/s12874-019-0681-4. URL https://doi.org/10.1186/s12874-019-0681-4.

[14] B. N. Huynh, A. R. Groendahl, O. Tomic, K. H. Liland, I. S. Knudtsen, F. Hoebers, W. van Elmpt, E. Malinen, E. Dale, and C. M. Futsaether. Head and neck cancer treatment outcome prediction: a comparison between machine learning with conventional radiomics features and deep learning radiomics. *Frontiers in Medicine*, 10, 2023. doi: 10.3389/fmed.2023.1217037. URL https://doi.org/10.3389/fmed.2023.1217037.

[15] A. Zwanenburg, M. Vallières, and M. A. Abdalah et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295:328–338, 2020. doi: 10.1148/radiol.2020191145.

[16] Kreftforeningen. PET/CT [internett], (Oppdatert 11.08.2023, hentet: 11.02.2024). URL https://kreftforeningen.no/om-kreft/undersokelser/pet-ct/.

[17] Mayo Clinic. Positron emission tomography scan [internett], (18.04.23, hentet: 11.02.2024). URL https://www.mayoclinic.org/tests-procedures/pet-scan/about/pac-20385078.

[18] J. M. Moan, C. D. Amdal, E. Malinen, J. G. Svestad, T. V. Bogsrud, and E. Dale. The prognostic role of 18f-fluorodeoxyglucose pet in head and neck cancer depends on hpv status. *Radiotherapy and Oncology*, 140:54–61, 2019. doi: 10.1016/j.radonc.2019.05.019. URL https://doi.org/10.1016/j.radonc.2019.05.019.

[19] Norsk Helseinformatikk. PET - Positron emisjons tomografi [internett], (Oppdatert 11.07.2019, hentet: 10.02.2024). URL https://nhi.no/sykdommer/barn/undersokelser/pet.

[20] Kreftforeningen. CT [internett], (Oppdatert 11.08.2023, hentet: 09.02.2024). URL https://kreftforeningen.no/om-kreft/undersokelser/ct/.

[21] National Institute of Biomedical Imaging and Bioengineering. Computed Tomography (CT) [internett], (Oppdatert juni 2022, hentet: 10.02.2024). URL https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct.

[22] H. Aerts, E. Velazquez, and R. Leijenaar et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Na-*

ture Communications, 5, 2014. doi: 10.1038/ncomms5006. URL https://doi.org/10.1038/ncomms5006.

[23] S. Raschka og V. Mirjalili. *Python machine learning: Machine learning and deep learning with python, scikit-learn, and tensorflow 2*. Packt Publishing Ltd., 2019.

[24] scikit learn. 1.1. linear models [internett]. URL https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. [Hentet: 08.01.2024].

[25] D. Chicco og G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics 21*, 6 (2020).

[26] ScikitLearn. sklearn.metrics.matthews_corrcoef[internett]. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html#sklearn.metrics.matthews_corrcoef. [Hentet: 01.02.2024].

[27] C. Anderson. Hot or not? heatmaps and correlation matrices [internett], (Publisert 2019, hentet: 08.02.2024). URL https://medium.com/@connor.anderson_42477/hot-or-not-heatmaps-and-correlation-matrix-plots-940088fa2806.

[28] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[29] M. L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

[30] A. Zwanenburg. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging*, 46:2638–2655, 2019. URL https://doi.org/10.1007/s00259-019-04391-8.

[31] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[32] Anaconda software distribution, versjon 2-2.4.0, 2020. URL https://docs.anaconda.com/.

[33] C. R. Harris, K. J. Millman, and S. J. van der Walt et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[34] W. McKinney. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, editor, *S. van der Walt and J. Millman*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[35] F. Pedregosa, G. Varoquaux, and A. Gramfort et al. Scikit-learn: Machine

learning in python. *Journal of Machine Learning Research*, 12: 2825-2830, 2011.

[36] O. Tomic, T. Graff, K. H. Liland, and T. Næs. hoggorm: a python library for explorative multivariate statistics. *The Journal of Open Source Software*, 4 (39), 2019. doi: 10.21105/joss.00980. URL http://joss.theoj.org/papers/10.21105/joss.00980.

[37] O. Tomic. hoggormplot[programvare]. URL https://github.com/olivertomic/hoggormPlot. [Hentet: 16.09.2023].

[38] Microsoft Corporation. Microsoft excel. versjon 2019 (16.0). URL https://office.microsoft.com/excel.

[39] A. Jenul, S. Schrunner, B. N. Huynh, and O. Tomic. RENT: A Python Package for Repeated Elastic Net Feature Selection. *Journal of Open Source Software*, 6(63):3323, 2021. doi: 10.21105/joss.03323. URL https://doi.org/10.21105/joss.03323.

[40] A. Jenul, S. Schrunner, K. H. Liland, U. G. Indahl, C. M. Futsaether, and O. Tomic. RENT - repeated elastic net technique for feature selection. *IEEE Access*, 9:152333–152346, 2021. doi: 10.1109/ACCESS.2021.3126429.

Tillegg A

Parametere

A.1 Input parametere for grid search-kryssvalidering

Tabellene under viser hvilke hyperparametere og rangen deres som ble brukt i grid search-kryssvalidering for hver av klassifiseringsalgoritmene, logsistisk regresjon, random forest, decision tree og KNN.

A.1.1 Logistisk regresjon klassifisering

Tekst.

Tabell A.1: *Input hyperparameter intervall logistisk regresjon*

Hyperparameter	Range/(verdi)område/intervall
solver	saga
penalty	elasticnet
l1_ratio	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
C	0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 500, 1000

A.1.2 Random Forest klassifisering

Tekst.

Tabell A.2: *Input hyperparameter intervall Random forest*

Hyperparameter	Range/(verdi)område/intervall
n_estimators	10, 30, 50, 100, 200, 500, 800
criterion	gini, entropy, log_loss
max_depth	None, 3, 5, 8, 10, 15, 18
min_samples_split	2, 4, 6

A.1.3 Decision tree klassifisering

Tekst.

Tabell A.3: *Input hyperparameter intervall decision tree*

Hyperparameter	Range/(verdi)område/intervall
criterion	gini, entropy
max_depth	None, 2, 5, 8, 10, 15
min_samples_split	2, 5, 8
min_samples_leaf	1, 2, 4

A.1.4 KNN klassifisering

Tekst.

Tabell A.4: *Input hyperparameter intervall KNN*

Hyperparameter	Range/(verdi)område/intervall
n_neighbors	3, 4, 5, 6, 8
algorithm	auto, ball_tree, kd_tree, brute
metric	euclidean, manhattan, minkowski
weights	uniform, distance

A.2 Output for grid search-kryssvalidering med DFS som respons

I tabellene under vises de beste parametere valgt ut i grid search-kryssvalidering for hver modell og datasettene med DFS som responsvariabel.

A.2.1 Logistisk regresjon klassifisering

Tabell A.5: *Valgte hyperparametere og intervall med logistisk regresjon, med HPV datasett DH1, DH2, DH3 med DFS respons.*

Datasett	solver	penalty	l1_ratio	C
DH1	saga	elasticnet	0,5	1
DH2	saga	elasticnet	0,1	0,1
DH3	saga	elasticnet	0,1	0,1

Tabell A.6: *Valgte hyperparametere og intervall med logistisk regresjon, uen HPV datasett DU1, DU2, DU3 med DFS respons.*

Datasett	solver	penalty	l1_ratio	C
DU1	saga	elasticnet	0,4	1
DU2	saga	elasticnet	0,4	0,1
DU3	saga	elasticnet	0,4	1

A.2.2 Random Forest klassifisering

Tabell A.7: *Valgte hyperparametere og intervall med Random forest, med HPV datasett DH1, DH2, DH3 med DFS respons.*

Datasett	n_estimators	criterion	max_depth	min_samples_split
DH1	10	gini	3	2
DH2	10	entropy	None	4
DH3	30	gini	None	2

Tabell A.8: *Valgte hyperparametere og intervall med Random forest, uten HPV datasett DU1, DU2, DU3 med DFS respons.*

Datasett	n_estimators	criterion	max_depth	min_samples_split
DU1	100	entropy	15	2
DU2	200	gini	None	4
DU3	10	entropy	None	2

A.2.3 Decision tree klassifisering

Tabell A.9: *Valgte hyperparametere og intervall med decision tree, med HPV datasett DH1, DH2, DH3 med DFS respons.*

Datasett	criterion	max_depth	min_samples_split	min_samples_leaf
DH1	gini	5	5	1
DH2	gini	5	5	1
DH3	entropy	None	2	2

Tabell A.10: *Valgte hyperparametere og intervall med decision tree, uten HPV datasett DU1, DU2, DU3 med DFS respons.*

Datasett	criterion	max_depth	min_samples_split	min_samples_leaf
DU1	entropy	None	2	1
DU2	entropy	None	2	2
DU3	entropy	5	2	2

A.2.4 KNN klassifisering

Tabell A.11: *Valgte hyperparametere og intervall med KNN, med HPV datasett DH1, DH2, DH3 med DFS respons.*

Datasett	n_neighbors	algorithm	metric	weights
DH1	8	auto	manhattan	uniform
DH2	3	auto	manhattan	distance
DH3	3	auto	manhattan	distance

Tabell A.12: *Valgte hyperparametere og intervall med KNN, uten HPV datasett DU1, DU2, DU3 med DFS respons.*

Datasett	n_neighbors	algorithm	metric	weights
DU1	6	auto	euclidean	uniform
DU2	4	auto	euclidean	uniform
DU3	3	auto	euclidean	distance

A.3 Output for grid search-kryssvalidering med OS som respons

I tabellene under vises de beste parametere valgt ut i grid search-kryssvalidering for hver modell og datasettene med OS som responsvariabel.

A.3.1 Logistisk regresjon klassifisering

Tabell A.13: *Valgte hyperparametere og intervall med logistisk regresjon, med HPV datasett DH1, DH2, DH3 med OS respons.*

Datasett	solver	penalty	l1_ratio	C
DH1	saga	elasticnet	0,6	1
DH2	saga	elasticnet	0,3	1
DH3	saga	elasticnet	0,7	1

Tabell A.14: *Valgte hyperparametere og intervall med logistisk regresjon, uten HPV datasett DU1, DU2, DU3 med OS respons.*

Datasett	solver	penalty	l1_ratio	C
DU1	saga	elasticnet	0,4	1
DU2	saga	elasticnet	0,4	0,1
DU3	saga	elasticnet	0,3	0,1

A.3.2 Random Forest klassifisering

Tabell A.15: *Valgte hyperparametere og intervall med Random forest, med HPV datasett DH1, DH2, DH3 med OS respons.*

Datasett	n_estimators	criterion	max_depth	min_samples_split
DH1	10	gini	3	4
DH2	10	gini	5	2
DH3	10	entropy	5	2

Tabell A.16: *Valgte hyperparametere og intervall med Random forest, uten HPV datasett DU1, DU2, DU3 med OS respons.*

Datasett	n_estimators	criterion	max_depth	min_samples_split
DU1	10	entropy	None	2
DU2	30	gini	10	2
DU3	100	entropy	10	4

A.3.3 Decision tree klassifisering

Tabell A.17: *Valgte hyperparametere og intervall med decision tree, med HPV datasett DH1, DH2, DH3 med OS respons.*

Datasett	criterion	max_depth	min_samples_split	min_samples_leaf
DH1	gini	None	2	5
DH2	gini	2	2	4
DH3	gini	None	2	1

Tabell A.18: *Valgte hyperparametere og intervall med decision tree, uten HPV datasett DU1, DU2, DU3 med OS respons.*

Datasett	criterion	max_depth	min_samples_split	min_samples_leaf
DU1	gini	5	2	1
DU2	entropy	2	2	1
DU3	entropy	2	2	1

A.3.4 KNN klassifisering

Tabell A.19: *Valgte hyperparametere og intervall med KNN, med HPV datasett DH1, DH2, DH3 med OS respons.*

Datasett	n_neighbors	algorithm	metric	weights
DH1	8	auto	manhattan	uniform
DH2	3	auto	manhattan	uniform
DH3	3	auto	manhattan	uniform

Tabell A.20: *Valgte hyperparametere og intervall med KNN, uten HPV datasett DU1, DU2, DU3 med OS respons.*

Datasett	n_neighbors	algorithm	metric	weights
DU1	4	auto	euclidean	uniform
DU2	6	auto	euclidean	uniform
DU3	8	auto	euclidean	distance

Tillegg B

Resultater per fold

Tabeller som viser ytelsesmålingene for hver fold i en modell. Viser resultatene fra de fire algoritmene for hvert av de seks datasettene, DH1, DH2, DH3 og DU1, DU2, DU3 for hver av responsvariablene DFS og OS.

B.0.1 DFS per fold

Med HPV

Tabell B.1: *DFS - per fold resultat for de fire modellene for DH1*

(a) *Logistisk regresjon DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,51	0,63	0,81	0,33	0,89	0,63
2	0,41	0,35	0,52	0,00	0,69	0,39
3	0,53	0,50	0,60	0,00	0,75	0,48
4	0,54	0,55	0,65	0,22	0,77	0,55
5	0,52	0,42	0,65	0,00	0,79	0,48

(b) *Random forest DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,81	0,83	0,90	0,67	0,94	0,83
2	0,46	0,40	0,62	0,00	0,76	0,45
3	0,60	0,64	0,65	0,22	0,77	0,58
4	0,58	0,55	0,65	0,22	0,77	0,56
5	0,63	0,50	0,70	0,00	0,82	0,53

(c) *Decision tree DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,63	0,73	0,86	0,40	0,92	0,71
2	0,33	0,32	0,48	0,00	0,65	0,36
3	0,24	0,36	0,50	0,00	0,67	0,36
4	0,68	0,65	0,70	0,50	0,79	0,66
5	0,50	0,61	0,70	0,40	0,80	0,60

(d) *KNN DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,66	0,63	0,81	0,33	0,89	0,66
2	0,50	0,43	0,67	0,00	0,80	0,48
3	0,33	0,41	0,55	0,00	0,71	0,40
4	0,40	0,50	0,65	0,00	0,79	0,47
5	0,48	0,50	0,70	0,00	0,82	0,50

Tabell B.2: *DFS - per fold resultat for de fire modellene for DH2*

(a) *Logistisk regresjon DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,74	0,66	0,67	0,46	0,76	0,66
2	0,51	0,40	0,62	0,00	0,76	0,46
3	0,59	0,62	0,65	0,46	0,74	0,61
4	0,59	0,58	0,65	0,36	0,76	0,59
5	0,80	0,67	0,75	0,44	0,84	0,70

(b) *Random forest DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,82	0,65	0,76	0,44	0,85	0,70
2	0,59	0,58	0,71	0,25	0,82	0,59
3	0,63	0,53	0,60	0,20	0,73	0,54
4	0,41	0,38	0,55	0,00	0,71	0,41
5	0,80	0,57	0,70	0,25	0,81	0,63

(c) *Decision tree DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,51	0,51	0,52	0,29	0,64	0,50
2	0,46	0,48	0,62	0,20	0,75	0,50
3	0,86	0,74	0,75	0,62	0,81	0,76
4	0,42	0,41	0,50	0,17	0,64	0,43
5	0,73	0,60	0,65	0,46	0,74	0,64

(d) *KNN DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,76	0,62	0,71	0,40	0,81	0,66
2	0,49	0,45	0,57	0,18	0,71	0,48
3	0,68	0,74	0,75	0,62	0,81	0,72
4	0,64	0,49	0,60	0,20	0,73	0,53
5	0,70	0,69	0,75	0,55	0,83	0,70

Tabell B.3: *DFS - per fold resultat for de fire modellene for DH3*

(a) *Logistisk regresjon DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,62	0,59	0,67	0,36	0,77	0,60
2	0,57	0,37	0,57	0,00	0,73	0,45
3	0,60	0,68	0,70	0,57	0,77	0,66
4	0,60	0,53	0,60	0,33	0,71	0,56
5	0,73	0,57	0,70	0,25	0,81	0,61

(b) *Random forest DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,79	0,54	0,71	0,25	0,82	0,62
2	0,46	0,52	0,67	0,22	0,79	0,53
3	0,47	0,41	0,55	0,00	0,71	0,43
4	0,33	0,38	0,55	0,00	0,71	0,39
5	0,71	0,68	0,75	0,29	0,85	0,65

(c) *Decision tree DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,70	0,66	0,67	0,46	0,76	0,65
2	0,57	0,57	0,67	0,36	0,77	0,59
3	0,55	0,61	0,65	0,36	0,76	0,59
4	0,38	0,37	0,45	0,15	0,59	0,39
5	0,44	0,44	0,55	0,18	0,69	0,46

(d) *KNN DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,79	0,65	0,76	0,44	0,85	0,70
2	0,49	0,48	0,62	0,20	0,75	0,51
3	0,70	0,81	0,80	0,67	0,86	0,77
4	0,62	0,49	0,60	0,20	0,73	0,53
5	0,58	0,61	0,70	0,40	0,80	0,62

Uten HPV

Tabell B.4: *DFS - per fold resultat for de fire modellene for DU1*

(a) *Logistisk regresjon DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,80	0,80	0,82	0,88	0,62	0,78
2	0,60	0,68	0,74	0,83	0,46	0,66
3	0,69	0,62	0,70	0,80	0,43	0,65
4	0,75	0,50	0,67	0,79	0,18	0,58
5	0,69	0,63	0,70	0,81	0,33	0,63

(b) *Random forest DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,80	0,70	0,75	0,84	0,36	0,69
2	0,47	0,57	0,67	0,78	0,31	0,56
3	0,63	0,64	0,78	0,86	0,40	0,66
4	0,64	0,58	0,70	0,81	0,33	0,61
5	0,62	0,63	0,70	0,81	0,33	0,62

(c) *Decision tree DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,78	0,79	0,82	0,87	0,71	0,79
2	0,39	0,39	0,44	0,57	0,21	0,40
3	0,74	0,72	0,78	0,85	0,57	0,73
4	0,69	0,68	0,74	0,82	0,53	0,69
5	0,42	0,41	0,52	0,67	0,13	0,43

(d) *KNN DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,68	0,59	0,61	0,69	0,48	0,61
2	0,40	0,42	0,44	0,55	0,29	0,42
3	0,70	0,65	0,74	0,83	0,46	0,68
4	0,57	0,62	0,67	0,76	0,47	0,62
5	0,61	0,63	0,70	0,80	0,43	0,64

Tabell B.5: *DFS - per fold resultat for de fire modellene for DU2*

(a) *Logistisk regresjon DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,60	0,56	0,61	0,70	0,42	0,58
2	0,57	0,70	0,74	0,84	0,36	0,64
3	0,72	0,59	0,78	0,87	0,25	0,64
4	0,65	0,67	0,78	0,87	0,25	0,64
5	0,73	0,57	0,67	0,78	0,31	0,61

(b) *Random forest DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,46	0,44	0,57	0,71	0,14	0,46
2	0,60	0,63	0,70	0,80	0,43	0,63
3	0,74	0,69	0,81	0,89	0,44	0,72
4	0,54	0,44	0,59	0,73	0,15	0,49
5	0,69	0,70	0,74	0,84	0,36	0,67

(c) *Decision tree DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,58	0,56	0,61	0,70	0,42	0,57
2	0,60	0,61	0,67	0,76	0,47	0,62
3	0,74	0,72	0,78	0,85	0,57	0,73
4	0,36	0,41	0,44	0,57	0,21	0,40
5	0,84	0,83	0,85	0,89	0,78	0,84

(d) *KNN DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,56	0,59	0,64	0,74	0,44	0,59
2	0,71	0,70	0,74	0,81	0,59	0,71
3	0,52	0,57	0,63	0,74	0,37	0,57
4	0,44	0,49	0,52	0,63	0,32	0,48
5	0,72	0,74	0,74	0,79	0,67	0,73

Tabell B.6: *DFS - per fold resultat for de fire modellene for DU3*

(a) *Logistisk regresjon DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,63	0,54	0,54	0,61	0,43	0,55
2	0,57	0,45	0,48	0,59	0,30	0,48
3	0,82	0,79	0,85	0,90	0,67	0,81
4	0,71	0,65	0,74	0,83	0,46	0,68
5	0,73	0,69	0,74	0,82	0,53	0,70

(b) *Random forest DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,56	0,51	0,61	0,73	0,27	0,53
2	0,59	0,67	0,70	0,78	0,56	0,66
3	0,67	0,60	0,74	0,84	0,36	0,64
4	0,70	0,65	0,74	0,83	0,46	0,67
5	0,66	0,58	0,63	0,72	0,44	0,61

(c) *Decision tree DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,58	0,59	0,64	0,74	0,44	0,60
2	0,58	0,65	0,70	0,79	0,50	0,64
3	0,76	0,72	0,81	0,88	0,55	0,74
4	0,49	0,49	0,52	0,63	0,32	0,49
5	0,69	0,68	0,74	0,83	0,46	0,68

(d) *KNN DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,67	0,64	0,71	0,81	0,43	0,65
2	0,73	0,75	0,78	0,86	0,50	0,72
3	0,56	0,46	0,63	0,76	0,17	0,52
4	0,45	0,42	0,56	0,70	0,14	0,45
5	0,69	0,67	0,70	0,78	0,56	0,68

B.0.2 OS per fold

Med HPV

Tabell B.7: *OS - per fold resultat for de fire modellene for DH1*

(a) *Logistisk regresjon DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,71	0,50	0,81	0,00	0,89	0,58
2	0,31	0,40	0,67	0,00	0,80	0,44
3	0,75	0,42	0,60	0,00	0,75	0,50
4	0,64	0,63	0,80	0,33	0,88	0,66
5	0,69	0,50	0,80	0,00	0,89	0,58

(b) *Random forest DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,68	0,83	0,90	0,67	0,94	0,81
2	0,23	0,45	0,81	0,00	0,89	0,48
3	0,81	0,50	0,65	0,00	0,79	0,55
4	0,47	0,44	0,75	0,00	0,86	0,50
5	0,61	0,73	0,85	0,40	0,91	0,70

(c) *Decision tree DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,66	0,63	0,81	0,33	0,89	0,66
2	0,22	0,34	0,48	0,00	0,65	0,34
3	0,66	0,73	0,75	0,44	0,84	0,69
4	0,50	0,57	0,75	0,29	0,85	0,59
5	0,68	0,63	0,80	0,33	0,88	0,66

(d) *KNN DH1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,60	0,68	0,76	0,29	0,86	0,64
2	0,56	0,50	0,76	0,00	0,86	0,54
3	0,47	0,50	0,75	0,00	0,86	0,52
4	0,05	0,47	0,90	0,00	0,95	0,47
5	0,57	0,50	0,75	0,00	0,86	0,54

Tabell B.8: *OS - per fold resultat for de fire modellene for DH2*

(a) *Logistisk regresjon DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,66	0,54	0,71	0,25	0,82	0,60
2	0,57	0,39	0,62	0,00	0,76	0,47
3	0,56	0,60	0,65	0,46	0,74	0,60
4	0,61	0,69	0,80	0,50	0,88	0,69
5	0,89	0,75	0,85	0,57	0,91	0,79

(b) *Random forest DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,49	0,45	0,76	0,00	0,86	0,51
2	0,61	0,43	0,76	0,00	0,86	0,53
3	0,75	0,73	0,75	0,44	0,84	0,70
4	0,41	0,44	0,75	0,00	0,86	0,49
5	0,83	0,75	0,85	0,57	0,91	0,78

(c) *Decision tree DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,88	0,69	0,81	0,50	0,88	0,75
2	0,69	0,45	0,81	0,00	0,89	0,57
3	0,69	0,67	0,70	0,57	0,77	0,68
4	0,31	0,42	0,70	0,00	0,82	0,45
5	0,66	0,44	0,75	0,00	0,86	0,54

(d) *KNN DH2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,57	0,73	0,86	0,40	0,92	0,70
2	0,37	0,43	0,76	0,00	0,86	0,49
3	0,72	0,64	0,70	0,40	0,80	0,65
4	0,36	0,42	0,70	0,00	0,82	0,46
5	0,74	0,75	0,85	0,57	0,91	0,76

Tabell B.9: *OS - per fold resultat for de fire modellene for DH3*

(a) *Logistisk regresjon DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,72	0,54	0,71	0,25	0,82	0,61
2	0,57	0,42	0,71	0,00	0,83	0,51
3	0,62	0,56	0,60	0,43	0,69	0,58
4	0,69	0,69	0,80	0,50	0,88	0,71
5	0,84	0,44	0,75	0,00	0,86	0,58

(b) *Random forest DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,43	0,50	0,81	0,00	0,89	0,53
2	0,74	0,77	0,90	0,50	0,95	0,77
3	0,69	0,55	0,65	0,22	0,77	0,58
4	0,31	0,42	0,70	0,00	0,82	0,45
5	0,92	0,83	0,90	0,67	0,94	0,85

(c) *Decision tree DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,57	0,56	0,62	0,33	0,73	0,56
2	0,92	0,82	0,86	0,67	0,91	0,83
3	0,53	0,53	0,60	0,33	0,71	0,54
4	0,31	0,34	0,50	0,00	0,67	0,36
5	0,53	0,53	0,70	0,25	0,81	0,57

(d) *KNN DH3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,57	0,73	0,86	0,40	0,92	0,70
2	0,37	0,43	0,76	0,00	0,86	0,49
3	0,75	0,73	0,75	0,44	0,84	0,70
4	0,33	0,42	0,70	0,00	0,82	0,45
5	0,74	0,75	0,85	0,57	0,91	0,76

Uten HPV

Tabell B.10: *OS - per fold resultat for de fire modellene for DU1*

(a) *Logistisk regresjon DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,74	0,69	0,71	0,80	0,50	0,69
2	0,55	0,59	0,63	0,72	0,44	0,59
3	0,62	0,54	0,59	0,70	0,35	0,56
4	0,76	0,56	0,63	0,75	0,29	0,60
5	0,65	0,57	0,63	0,74	0,37	0,59

(b) *Random forest DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,61	0,65	0,68	0,76	0,53	0,64
2	0,43	0,49	0,52	0,61	0,38	0,49
3	0,78	0,83	0,81	0,86	0,74	0,80
4	0,69	0,61	0,59	0,62	0,56	0,61
5	0,62	0,66	0,70	0,79	0,50	0,66

(c) *Decision tree DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,74	0,72	0,71	0,81	0,43	0,68
2	0,54	0,47	0,48	0,56	0,36	0,48
3	0,83	0,74	0,78	0,84	0,63	0,76
4	0,43	0,39	0,56	0,71	0,00	0,42
5	0,70	0,71	0,74	0,81	0,59	0,71

(d) *KNN DU1*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,68	0,61	0,61	0,65	0,56	0,62
2	0,34	0,35	0,33	0,36	0,31	0,34
3	0,63	0,60	0,63	0,72	0,44	0,60
4	0,49	0,53	0,48	0,46	0,50	0,49
5	0,75	0,68	0,70	0,76	0,60	0,70

Tabell B.11: *OS - per fold resultat for de fire modellene for DU2*

(a) *Logistisk regresjon DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,65	0,66	0,68	0,74	0,57	0,66
2	0,62	0,67	0,70	0,78	0,56	0,67
3	0,79	0,71	0,78	0,85	0,57	0,74
4	0,62	0,62	0,67	0,76	0,47	0,63
5	0,72	0,71	0,74	0,82	0,53	0,71

(b) *Random forest DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,61	0,69	0,71	0,80	0,50	0,66
2	0,61	0,59	0,63	0,72	0,44	0,60
3	0,77	0,77	0,81	0,87	0,67	0,78
4	0,65	0,54	0,59	0,70	0,35	0,57
5	0,71	0,80	0,81	0,86	0,71	0,78

(c) *Decision tree DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,61	0,58	0,61	0,69	0,48	0,59
2	0,59	0,59	0,59	0,65	0,52	0,59
3	0,77	0,72	0,70	0,76	0,60	0,71
4	0,55	0,55	0,59	0,69	0,42	0,56
5	0,68	0,66	0,63	0,64	0,62	0,65

(d) *KNN DU2*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,52	0,51	0,50	0,53	0,46	0,51
2	0,61	0,61	0,67	0,77	0,40	0,61
3	0,76	0,62	0,67	0,76	0,47	0,66
4	0,56	0,62	0,63	0,69	0,55	0,61
5	0,66	0,76	0,78	0,83	0,67	0,74

Tabell B.12: *OS - per fold resultat for de fire modellene for DU3*

(a) *Logistisk regresjon DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,69	0,66	0,68	0,74	0,57	0,67
2	0,56	0,57	0,59	0,67	0,48	0,57
3	0,76	0,71	0,78	0,85	0,57	0,73
4	0,64	0,59	0,63	0,72	0,44	0,60
5	0,73	0,66	0,70	0,79	0,50	0,68

(b) *Random forest DU3*

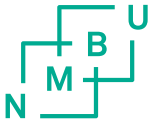
Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,57	0,52	0,57	0,68	0,33	0,54
2	0,62	0,62	0,67	0,76	0,47	0,63
3	0,90	0,86	0,89	0,92	0,80	0,88
4	0,58	0,50	0,56	0,67	0,33	0,53
5	0,73	0,77	0,78	0,85	0,57	0,74

(c) *Decision tree DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,61	0,58	0,61	0,69	0,48	0,59
2	0,59	0,59	0,59	0,65	0,52	0,59
3	0,77	0,72	0,70	0,76	0,60	0,71
4	0,55	0,55	0,59	0,69	0,42	0,56
5	0,68	0,66	0,63	0,64	0,62	0,65

(d) *KNN DU3*

Fold	AUC	MCC	nøyaktighet	F1:1	F1:0	Avg_score
1	0,58	0,65	0,68	0,76	0,53	0,64
2	0,58	0,61	0,67	0,78	0,31	0,59
3	0,79	0,68	0,74	0,82	0,53	0,71
4	0,51	0,52	0,56	0,65	0,40	0,53
5	0,71	0,71	0,74	0,82	0,53	0,70



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

Postboks 5003
NO-1432 Ås
Norway