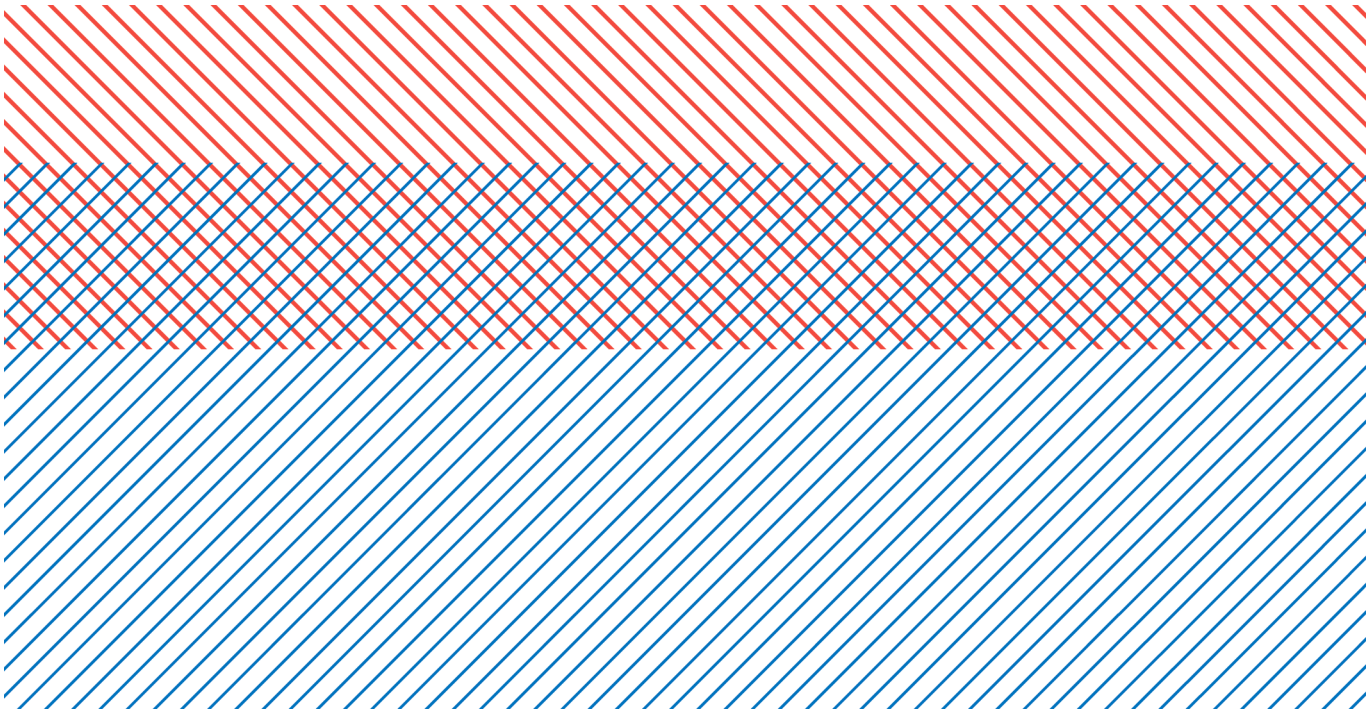
DET NATIONALE FORSKNINGS- OG ANALYSECENTER FOR VELFÆRD

Machine learning & data

Teknisk bilag til rapport om lærerkarakteristika og elevers læring



Marianne Mikkelsen, Mads Lang Sørensen, Hans Henrik Sievertsen & Vibeke Myrup Jensen

DET NATIONALE FORSKNINGS- OG ANALYSECENTER FOR VELFÆRD

Machine Learning & data – Teknisk bilag til rapport om lærerkarakteristika og elevers læring

e-ISBN: 978-87-7119-837-9

Projekt: 301755

VIVE – Viden til Velfærd
Det Nationale Forsknings- og Analysecenter for Velfærd
Herluf Trolles Gade 11, 1052 København K
www.vive.dk

Indhold

Sammenfatning 4

1 Machine learning til at forudsige elevernes faglige niveau 7
- 1.1 Udvælgelse af modeller 7
- 1.2 Finjustering af modellerne 8
- 1.3 Udvælgelse af model med størst forudsigelseskraft 9

2 Hvilke karakteristika korrelerer med elevernes faglige niveau? 12
- 2.1 Karakteristika med største forklaringskraft for elevernes faglige resultater 12
- 2.2 Hvilke lærer- og skolekarakteristika korrelerer med elevernes faglige niveau 13
- 2.3 Forskel på korrelationen i dansk og matematik 16

3 Data og repræsentativitet 17
- 3.1 Datagrundlag 17
- 3.2 Repræsentativitetsanalyse 19

Litteratur 22

Bilag 1 Supplerende tabeller og figurer 24

Sammenfatning

Rapporten undersøger, hvilke lærerkarakteristika og rammevilkår for lærerne, der korrelerer med elevernes faglige resultater. I tillæg til rapporten udarbejdes dette tekniske notat, som gennemgår den valgte statistiske analyse (machine learning) og datagrundlaget.

Vores udfaldsmål i alle analyser er elevernes faglige resultater i henholdsvis dansk eller matematik i 9. klasse, som vi operationaliserer på to forskellige måder:

- Elever med høje faglige resultater (ligger i øverste 25 % af karakterfordelingen)
- Elever med lave faglige resultater (ligger i nederste 25 % af karakterfordelingen).

Boks 1 viser de *lærerkarakteristika og rammevilkår*, som vores machine learning-analyser viser har den største korrelation med elevernes faglige resultater på tværs af de tre udfaldsmål og i hvilken retning. Retningen på korrelationerne er bestemt ud fra den logistiske regressionsmodel i rapporten.

Boks 1 Karakteristika korrelerer med elevernes faglige resultater

Karakteristika, der viser en positiv korrelation med, at eleverne får høje faglige resultater

- Nogen grad af tydelighed og klare mål med undervisningen
- Undervisningsdifferentiering: Læreren sørger i høj grad for, at opgaverne passer til elevernes niveau
- Forholdsvis små skoler (næstlaveste fjerdedel af elevtallet blandt alle folkeskoler)
- Høj grad af elevengagement i undervisningen
- Læreren sørger for ro i klassen
- Forholdsvis få antal fagtimer (næstlaveste fjerdedel af timetallet blandt alle folkeskoler)
- Eleverne bliver sjældent forstyrret af larm i timerne.

Karakteristika, der viser en negativ korrelation med, at eleverne får gode faglige resultater

- Læreren er uddannet meritlærer (modsat en traditionel læreruddannelse eller andet).

Karakteristika, der korrelerer med, at eleverne undgår at få lave karakterer

- Feedback: Læreren fortæller i høj grad, hvordan eleverne bliver endnu bedre til faget
- Undervisningsdifferentiering: Læreren sørger i høj grad for, at opgaverne passer til elevernes niveau
- Undervisningen opleves i nogen grad spændende og praksisnær
- Læreren er i høj grad tydelig med hensyn til indhold i gruppearbejde
- Høj grad af tydelighed og klare mål med undervisningen
- De største skoler (højeste fjerdedel af elevtallet blandt alle folkeskoler)
- Nogen grad af tydelighed og klare mål med undervisningen
- Læreren underviser 1-4 klasser (modsat flere klasser)
- Andelen af lærere uden undervisningskompetence i faget er 20 % eller derover i udskolingen.

Karakteristika, der viser en positiv korrelation med, at eleverne får lave karakterer

- Mere end én voksen er tilknyttet klassen.

På tværs af alle analyser er det elevernes faglige resultater fra de nationale test i 6. klasse, som korrelerer mest med elevernes faglige resultater i 9. klasse. Dernæst er det elevkarakteristika som forældrenes uddannelse og indkomst. Vi finder samtidig, at ovenstående karakteristika har samme forklaringskraft for elevernes faglige resultater i dansk som i matematik.

Machine learning

Vi anvender en machine learning-tilgang til at finde frem til den model, som bedst kan forklare elevernes faglige resultater, mens vi anvender SHAP-værdier til at finde frem til de karakteristika, som har den største vigtighed i modellen. Formålet er at have en bred, åben og induktiv tilgang, hvor så mange karakteristika som muligt medtages. Med machine learning-tilgangen kan der ikke tales om kausalitet, men formålet er at finde karakteristika, der potentielt er betydningsfulde for elevernes faglige resultater, og som efterfølgende tjekkes med andre metoder.

Vi har på forhånd udvalgt tre forskellige machine learning-modeller, som vi sammenligner i forhold til, hvor gode de er til at forudsige elevernes faglige niveau. De tre modeller holdes samtidig op imod vores benchmarkingmodel. Det vil side 'den typiske' model, man kunne anvende i lignende analyser uden brug af machine learning. Med disse fire modeller betragter vi tre overordnede hovedområder, nemlig lærerkarakteristika eller forhold ved lærernes rammevilkår, samtidig med at vi kontrollerer for en bred vifte af elevkarakteristika og familiebaggrund:

- Karakteristika relateret til lærernes undervisningspraksis
- Karakteristika relateret til rammevilkår
- Karakteristika om lærernes baggrund.

Sammenligningen af de tre machine learning-tilgange viser, at med det valgte datagrundlag er gradient boosting-modellen bedst til at forudsige elevernes faglige resultater. Machine learning-analyserne viser dog også, at der reelt set ikke er statistisk forskel på, hvor godt denne model 'præsterer' i forhold til vores benchmarkingmodel. Derfor anvender vi kun gradient boosting-modellen og de efterfølgende estimerede SHAP-værdier som en eksplorativ tilgang til at udvælge højt korrelerende lærerkarakteristika eller forhold ved lærernes rammevilkår. Herefter anvender vi en lineær sandsynlighedsmodel til at teste, om de valgte karakteristika i højere grad korrelerer med elevernes faglige resultater i dansk end i matematik.

En væsentlig ulempe ved at bruge machine learning er, at fortolkningen af de forklarende variables vigtighed bliver usikker på bekostning af den højere præcision i forhold til udfaldsmålet. Derfor er SHAP-værdier den bedste metode til at estimere de enkelte karakteristikas 'vigtighed' i forhold til at forklare elevernes faglige resultater. Det er denne liste af variabler, som vi i rapporten tester i forhold til tre forskellige robusthedstest (jf. Boks 1). Dog er det vigtigt at pointere, at i dette tilfælde har machine learning-modellerne ikke bedre præcision, hvorfor vi i stedet bruger andre metoder med bedre tolkningsmuligheder til robusthedsanalysen i rapporten.

I kapitel 1 præsenterer vi den valgte machine learning-tilgang, mens vi i kapitel 2 viser, hvilke karakteristika der har den største 'vigtighed' i den valgte machine learning-tilgang. I tredje og sidste kapitel beskriver vi datagrundlaget og en kort repræsentativitetsanalyse.

1 Machine learning til at forudsige elevernes faglige niveau

Machine learning er et generelt udtryk for metoder, der anvendes til at finde mønstre i data. Såkaldt 'supervised' machine learning handler om at *forudsige* et udfald, her fx elevernes faglige resultater. For eksempel anvender Cornell-Farrow and Garrard (2018) faglige testresultater fra test i de yngre klasser og en række elevkarakteristika til at forudsige, hvilke elever der vil klare sig dårligere end forventet i den kommende test året efter.

Stort set alle statistiske metoder (regression, faktoranalyse, etc.), der bruges i samfundsvidenskabelig forskning, er machine learning, men machine learning dækker også over mere avancerede teknikker, der først i de seneste år er blevet brugt i samfundsvidenskab. I det følgende gennemgår vi de anvendte machine learning-modeller.

1.1 Udvælgelse af modeller

Der er i princippet en lang række machine learning-modeller, der kan anvendes til at forudsige elevernes faglige resultater. Af hensyn til projektets omfang har vi begrænset os til at bruge én simpel og to komplekse modeller for at teste forskellige typer af supervised machine learning-modeller. Disse machine learning-modeller er udvalgt, fordi de er blandt de mest udbredte, og dermed giver et generelt indblik i fordele og ulemper ved anvendelse af machine learning-modeller til denne problemstilling. Vi sammenligner de tre modeller med en benchmarkingmodel.

- Logistisk regressionsmodel (benchmarkingmodel)
- Logistisk regressionsmodel med L2regulering
- Random Forest-model
- Gradient Boosting-model

Vi opstiller først en benchmarkingmodel, som er den model de tre andre modeller sammenlignes med. Benchmarkingmodellen vælges altid ud fra, hvad der vil være 'den typiske' model, man anvender i lignende analyser uden brug af machine learning-metoder. Da to af vores udfaldsmål er dikotome vælges den logistiske regressionsmodel.[1] Modellen er samtidig nem at fortolke relativt til de fleste andre modeltyper, og den har god forudsigelseskraft for datasæt med mange forklarende variable (Müller & Guido, 2017).

Den første model, vi tester, er en afart af den logistiske regressionsmodel, som har L2 regulering (L2-regularization). Denne model medtages, da der er en af de mest simple machine learning-modeller. L2 regulering betyder, at modellen nedjusterer meget høje værdier af koefficienterne. Modellen repræsenterer dermed også et alternativ i vores analyse i forhold til de to andre machine learning-modeller, som begge er komplekse 'ensemble modeller.' Ensemble modeller er en fællesbetegnelse for den type af modeller, der kombinerer flere andre modeller for at producere én optimal model.

[1] I rigtig mange sammenhænge anvendes også en lineær sandsynlighedsmodel (OLS regression med dikotom udfaldsmål), men det er ikke muligt at teste en lineær sandsynlighedsmodel i en machine learning-sammenhæng.

De to ensemble machine learning-modeller, vi har valgt, er random forest og gradient boosting. De har begge vist sig at have den største forklaringskraft i en større test af en lang række machine learning-modeller og regnes for 'state-of-the-art' inden for supervised machine learning (se Fernandez-Delgado et al., 2014; Olson et al., 2018).

Boks 1.1 beskriver hver af de tre udvalgte machine learning-modeller, mens vi i næste afsnit beskriver, hvilke skridt vi har foretaget i machine learning-analysen:

Boks 1.1 De udvalgte machine learning-modeller

Logistisk regression med L2-regularization er medtaget, fordi det er en af de mest simple machine learning-modeller. Det er en udvidelse af den 'klassiske' logistiske regression, hvor der tilføjes en begrænsning på ekstreme værdier i opdateringen af vægtene i regressionen.[2] Dermed er der mindre risiko for 'overfitting' eller for høj varians i modellen.

Random forest er en model, der kombinerer mange forskellige såkaldte decision tree-modeller. Hvert 'træ' forsøger at opdele observationerne ud fra de forklarende variable, så hver gruppe består af observationer, der har samme værdi af udfaldsmålet. Dermed kan man nu forudsige udfaldsmålet for nye observationer ved at undersøge de variable, som blev brugt til at opdele grupperne. Random forest-modellen laves ved at køre sådanne decision tree-modeller på mange bootstrap-samples, hvor der også kun bruges en stikprøve af de forklarende variable. Derefter bruges den værdi af udfaldsmålet, som majoriteten af modellerne har forudsagt som den endelige forudsigelse af udfaldsmålet. Kombinationen af mange modeller med individuel høj varians giver *random forest*-modellen mindre varians og mere præcision i sine analyser end fx en ordinær regressionsmodel.

Gradient boosting er en model, der ligesom random forest også kombinerer mange simple modeller (fx også decision tree-modeller). Princippet er at tage en simpel model og forsøge at forudsige udfaldsmålet for observationerne. Derefter betragter man, hvilke observationer modellen havde svært ved at klassificere korrekt, og i en ny simpel model gives disse observationer en højere vægt. Dette gentages *x* antal gange, hvor fejl-klassificerede observationer gives mere vægt, indtil så mange observationer som muligt er korrekt klassificeret. Der findes forskellige algoritmer, der bygger på *boosting*, fx gradient boosting.

1.2 Finjustering af modellerne

Modellerne testes på to forskellige kodninger af udfaldsmålet: elevernes karakterer i 9. klasse:

- Indikator for, at eleverne har høje faglige resultater (ligger i øverste 25 %)
- Indikator for, at eleverne har lave faglige resultater (ligger i nederste 25 %)

De to udfaldsmål er dannet som to variabler, der indikerer, om elevens karakter befinder sig enten i de øverste 25 % eller i nederste 25 % af karakterfordelingen (karakterskalaen). Vi har elevernes karakterer i enten dansk eller matematik, men ikke i begge fag. Det skyldes, at vores stikprøve stammer fra følgeforskningspanelet til folkeskolereformen, hvor halvdelen af eleverne i klassen har svaret på spørgsmål med fokus på undervisningen i dansk, og den anden halvdel for matematik. Vi bruger udfaldsmålene til at få et signal om, hvilke karakteristika der

[2] L2-regulering svarer til Ridge-regulering med et kontinuert udfaldsmål. L1-regulering eller Lasso-regulering er en lignende model, som man også kunne have anvendt. Det er vores erfaring, at disse ofte giver samme resultater, men dette er dog ikke testet i praksis af hensyn til projektets omfang.

korrelerer med, at eleverne klarer sig rigtig godt, og hvilke karakteristika der korrelerer med, at eleverne ikke klarer sig særlig godt.[3]

Vi anvender et datasæt bestående af en lang række karakteristika om elevernes socioøkonomiske baggrund, elevernes tidligere faglige resultater, lærernes undervisningspraksis, karakteristika relateret til lærerne rammebetingelser og karakteristika om lærernes baggrund (se afsnit 3 for den konkrete variabelliste). Det er vigtigt at understrege, at hvis vi havde et andet datasæt med andre karakteristika, kan vi ikke udelukke, at machine learning-modellerne vil komme frem til et andet resultat.

Det samlede datasæt opdeles i et træningsdatasæt (70 % af det samlede data) og et testdatasæt (30 % af det samlede data). Træningsdatasættet anvendes til at 'træne' og optimere modellerne på, mens testdatasættet alene anvendes til en form for 'slutevaluering' af modellen (se afsnit 1.3).

For de binære udfaldsmål er data opdelt ved at tage små stratificerede stikprøver i forhold til udfaldsmålet, som er knyttet til henholdsvis træningsdatasættet og testdatasættet. De tre machine learning-modeller er hver især kørt 625 gange med forskellige kombinationer af hyperparametre for at finde den mest optimale model. Hyperparametre er de værdier, som modellen 'kan justeres på' i træningsprocessen. Det kan fx være antallet af 'træer', dvs. antal modeller, der samles i en Random Forest. Justeringen af hyperparametre, som fx antal 'træer' er gjort for at finde den værdi af hyperparametret, der giver den bedste balance i forhold til varians og bias.

Alle modeller undtagen benchmarkingmodellen optimeres ved hjælp af en 'grid search'. Det er en metode, hvor alle kombinationer af modellens hyperparametre trænes og den mest optimale model er fundet. I grid search-metoden anvendes samtidig 5-fold cross-validation, som er en fremgangsmåde til at træne modellen på flere mindre dele af data for at vælge de mest robuste parametre.

Når træningssessionen afsluttes, angives de mest optimale parametre for hver model (hyperparametre). I de modeller, hvor udfaldsmålet er elever med høje faglige resultater, er disse hyperparametre angivet i Bilagstabel 1.1.

1.3 Udvælgelse af model med størst forudsigelseskraft

Næste skridt er at teste, hvilken af de fire modeller der er bedst til at forklare elevernes faglige resultater. Dette gøres ved at teste modellens forudsigelseskraft og præcision på uset data, der ikke er brugt til træningen af modellen. Denne test foretages i to trin:

- Udregning af confusion matrix
- Opstilling af ROC-kurver.

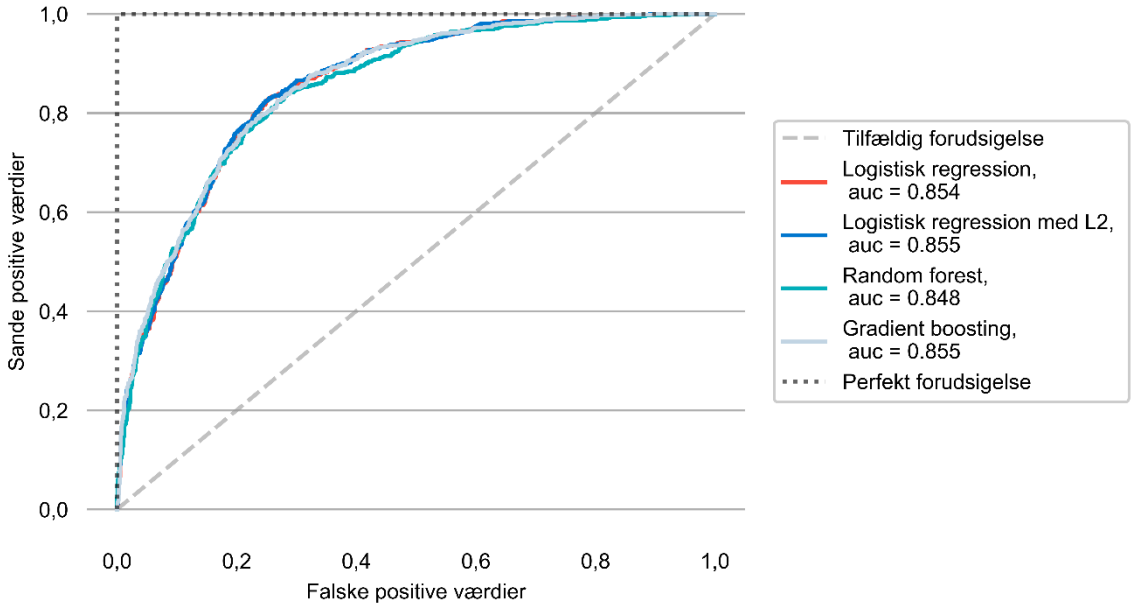
Desuden rapporteres også modellernes præcision (andelen af korrekt klassificerede observationer) samt Cohen's kappa, som er et tilsvarende mål særligt udbredt inden for uddannelse

[3] I de to udfaldsmål vi anvender, har vi selv valgt grænserne for, hvornår eleverne har henholdsvis høje eller lave faglige resultater. Det kan dog være en bekymring, at disse grænser har betydning for, hvilke karakteristika machine learning-analyserne udvælger. Derfor har vi også lavet analyserne med et tredje udfaldsmål som er elevernes karaktergennemsnit i dansk eller matematik. Machine learning–analysen viser, at det overvejende er de samme karakteristika, som går igen ved brug af dette udfaldsmål.

og psykologi til at rapportere modellens præcision (Cohen 1960; Ben-David 2008). Vi præsenterer her alene resultaterne fra ROC-kurverne, fordi ROC-kurverne er en grafisk illustration af resultaterne fra den såkaldte 'confusion matrix'. En confusion matrix er en 2x2-tabel, der viser forholdet mellem den sande fordeling af udfaldsmålet og den forudsagte fordeling af udfaldsmålet (Raschka & Mirjalili, 2019). Bilagsfigur 1.1 viser confusion matrix for de tre machine learning-modeller samt benchmarkingmodellen, når udfaldsmålet er elever med høje faglige resultater.

Figur 1.1 nedenfor viser ROC-kurverne for de fire modeller, hvor udfaldsmålet er elever med høje faglige resultater. ROC er en forkortelse for *receiver operating characteristic curves*. Den lodrette akse i figuren angiver den andel af værdier af udfaldsmålet, som modellen forudsiger korrekt, mens den vandrette akser angiver den andel af værdier af udfaldsmålet, som modellen forudsiger forkert. Den stiplede diagonale linje i figuren angiver en model, hvor der er lige mange korrekte og forkerte forudsagte værdier, og hvor modellen altså ikke er bedre end et tilfældigt gæt. Den prikkede linje i øverste venstre hjørne af figuren indikerer derimod en fuldstændig optimal model, hvor alle observationer er perfekt klassificeret. Jo større arealet under kurven er, desto bedre forklaringskraft har modellen.

Figur 1.1 ROC-kurver for de fire modeller, hvor udfaldsmålet er elever med høje faglige resultater



Anm.: Figuren viser ROC-kurver for de fire modeller samt to kurver for henholdsvis en perfekt forudsigelse og en tilfældig forudsigelse. Jo større arealet under kurven er, jo bedre forklaringskraft har modellen. AUC-værdierne i forklaringen til højre for kurven viser endvidere det specifikke areal under kurven.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

ROC-kurverne viser tydeligt, at de tre machine learning-modeller ikke er bedre til at forudsige elevernes faglige niveau end benchmarkingmodellen: logistisk regression. Dette kan ses ved, at kurverne ikke varierer særligt fra hinanden. I dette tilfælde giver det altså ikke højere præcision at bruge machine learning-modellerne.[4] Når vi sammenligner alle evalueringsmetoder af

[4] Vi får samme resultat, når udfaldsmålet er elever med lave faglige resultater.

modellerne (ROC-kurver, confusion matrix og test af modellernes præcision), er det overordnede resultat, at machine learning-modellerne hverken præsterer værre eller bedre end benchmarkingmodellen.

Det er svært at sige, om en model er absolut "god", og den skal altid ses relativt til benchmarkingmodellen. Derfor kan vi ikke sige om machine learning-modellerne præsterer lige så dårligt som logistisk regression, eller logistisk regression præsterer lige så godt som machine learning-modellerne. Det betyder samtidig, at figurens 'AUC-værdi' på 0,86 (arealet under kurven) i sig selv ikke har nogen klar fortolkning, men kun kan analyseres i sammenligning med værdien for de øvrige modeller (jf. forklaringsboksen i figuren).[5]

Der kan være mange årsager til, at machine learning-modellerne ikke er bedre end den logistiske regression. Alle modellerne er afhængige af de data, der anvendes, både når det gælder mængden af variabler i modellen og antallet af observationer. Random forest og gradient boosting-modellerne er normalt gode til at forudsige udfaldet, når der er meget komplekse sammenhænge mellem variablerne. I vores tilfælde kan det tyde på, at sammenhængene mellem variablerne ikke er lige så komplekse, og dette kan muligvis være én forklaring på, at den logistiske regressionsmodel præsterer lige så godt som machine learning-modellerne.

[5] For eksempel kan et mere eller mindre balanceret udfaldsmål forårsage en "kunstigt" høj præcision, idet sandsynligheden for at gætte rigtigt alt andet lige er højere, hvis fx fordelingen af udfaldsmålet er 90/10 end 50/50.

2 Hvilke karakteristika korrelerer med elevernes faglige niveau?

Formålet med dette kapitel er at undersøge, hvilke lærerkarakteristika eller forhold ved lærernes rammevilkår, der i vores modeller er vigtige for at forudsige elevernes faglige resultater. Machine learning-modeller er populære, fordi de ofte er mere præcise end standard regressionsmetoder, når man fx ønsker at forudsige elevernes faglige resultater. Modellerne er til gengæld væsentlig sværere at håndtere, når vi ønsker at vise, hvilke variabler som reelt set betyder noget. Særligt de komplekse modeller kan ses som en slags 'black box', hvor det er umuligt at fortolke, hvad der ligger til grund for forudsigelsen af udfaldsmålet.

Udregning af SHAP-værdier er én nyudviklet metode, der forsøger at fortolke hver enkelt variabels 'vigtighed' i forhold til forudsigelsen af udfaldsmålet (Lundberg & Lee, 2017). I fortolkningen af SHAP-værdierne, er det vigtigt at pointere, at metoden antager, at de enkelte variable er uafhængige af hinanden (Lundberg & Lee, 2017). Derfor kan de ikke fortolkes med samme sikkerhed som fx koefficienter i en logistisk regression, og der er heller ikke tale om kausale sammenhænge. Man kan i stedet mere betragte SHAP-værdierne som et udtryk for modellens vægtning af variabler, end man kan tale om den egentlige årsagssammenhæng mellem de enkelte karakteristika og elevernes faglige resultater.

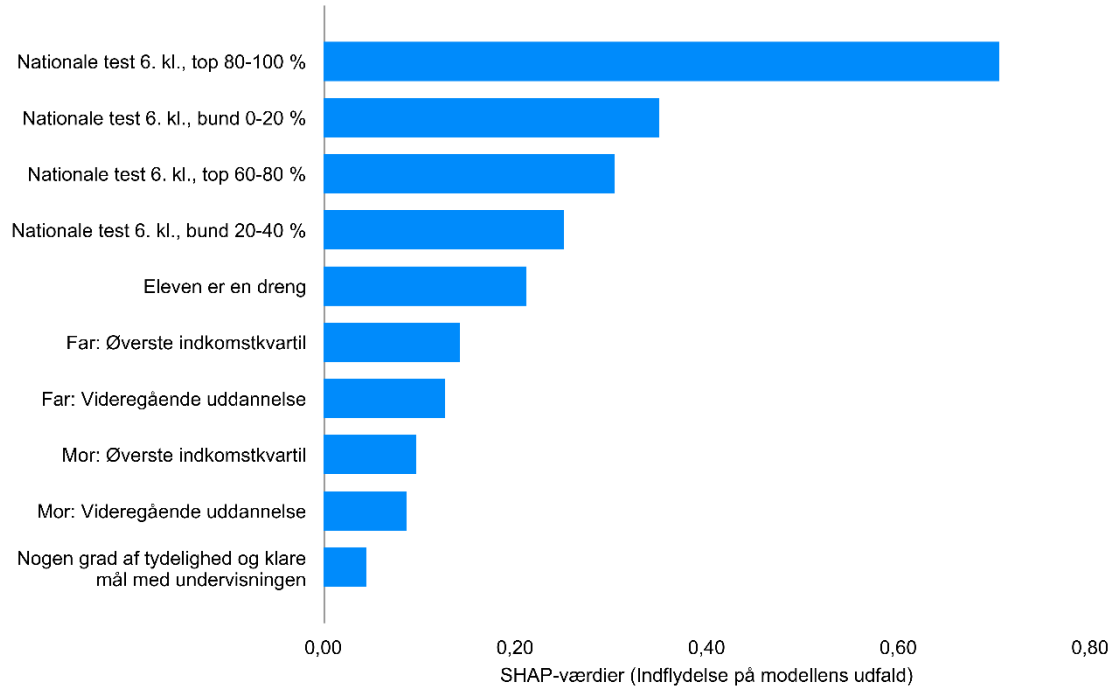
2.1 Karakteristika med største forklaringskraft for elevernes faglige resultater

I Figur 2.1 ses de karakteristika med de ti højeste SHAP-værdier for vores gradient boosting-model, hvor udfaldsmålet er elever med høje faglige resultater. Figuren skal fortolkes sådan, at jo længere søjlerne er, desto mere korrelerer de enkelte karakteristika med sandsynligheden for, at eleverne har høje faglige resultater. Det er dog vigtigt at understrege, at figuren ikke siger noget om, hvorvidt der er tale om positive eller negative korrelationer mellem de enkelte karakteristika og udfaldsmålet. Det viser vi i stedet i hovedrapportens Figur 3.1. og 4.1.

De karakteristika, som har de højeste SHAP-værdier er alle faktorer, der knytter sig til eleverne. Det er særligt elevernes tidligere faglige resultater i de nationale test (DNT), der er vigtige i forhold til at forudsige elevernes nuværende faglige resultater. Derefter kommer elevens køn, samt hvorvidt elevens forældre har en videregående uddannelse eller en høj indkomst.

Elevernes resultater i DNT er et udtryk for mange forskellige faktorer. De kan fortolkes som et udtryk for, hvad eleverne har lært i skolen, hvor betydningen af lærer- og skolekvaliteten i elevens tidlige år spiller ind. Samtidig er de også et udtryk for påvirkningen fra elevernes og deres forældres socioøkonomiske baggrund og medfødte evner. Det er vanskeligt at adskille de årsager, der kan påvirke både elevernes resultater i DNT i 6. klasse og ved folkeskolens afgangsprøve i 9. klasse. Den høje SHAP-værdi for DNT signalerer dog, at der er en væsentlig korrelation mellem tidligere faglige resultater i de nationale test og senere faglige resultater ved prøverne i 9. klasse. Vi ser samme stærke korrelation mellem elevernes tidligere faglige resultater i DNT og sandsynligheden for at være blandt elever med lave faglige resultater (se Bilagsfigur 1.2).

Figur 2.1 Korrelation mellem elev-, lærerkarakteristika eller rammevilkår og elever med høje faglige resultater. De ti karakteristika med højeste SHAP-værdier.



Anm.: Figuren viser SHAP-værdier for udfaldsmålet elever med høje faglige resultater. Vi anvender her gradient boosting-modellen beregnet på et datagrundlag for elever i 9. klasse for årene 2016-2018. Bemærk, at en høj SHAP-værdi siger noget om, at variablen korrelerer, men ikke hvordan den korrelerer med udfaldsmålet. Dermed kan det ikke fra denne figur udledes, om et karakteristika har en positiv eller negativ indflydelse på elevernes faglige resultater, men blot at den har en indflydelse.

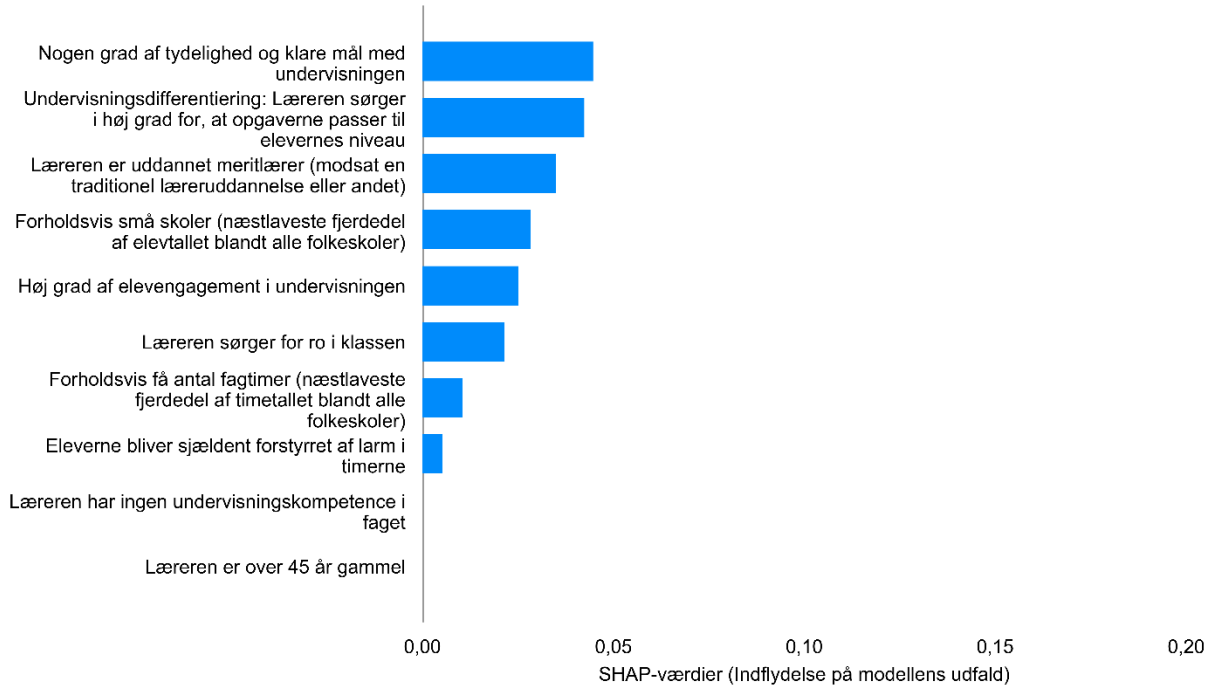
Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Der findes forskellige tilgange til at estimere SHAP-værdier for en model. Vi anvender metoden 'TreeExplainer', der er specielt udviklet til træ-baserede modeller, som random forest og gradient boosting (Lundberg et al., 2020).

2.2 Hvilke lærer- og skolekarakteristika korrelerer med elevernes faglige niveau

Figur 2.2 viser de ti lærerkarakteristika eller rammevilkår for lærerne med de højeste SHAP-værdier fra gradient boosting-modellen, hvor udfaldsmålet er elever med høje faglige resultater.

Figur 2.2 Korrelation mellem lærerkarakteristika eller rammevilkår og elever med høje faglige resultater. De 10 karakteristika med de højeste SHAP-værdier.



Anm.: Figuren viser SHAP-værdier for udfaldsmålet elever med høje faglige resultater. Vi anvender her gradient boosting-modellen beregnet på et datagrundlag for elever i 9. klasse for årene 2016-2018. Bemærk, at en høj SHAP-værdi siger noget om, at variablen korrelerer, men ikke hvordan den korrelerer med udfaldsmålet. Dermed kan det ikke fra denne figur udledes, om et karakteristika har en positiv eller negativ indflydelse på elevernes faglige resultater, men blot at den har en indflydelse.

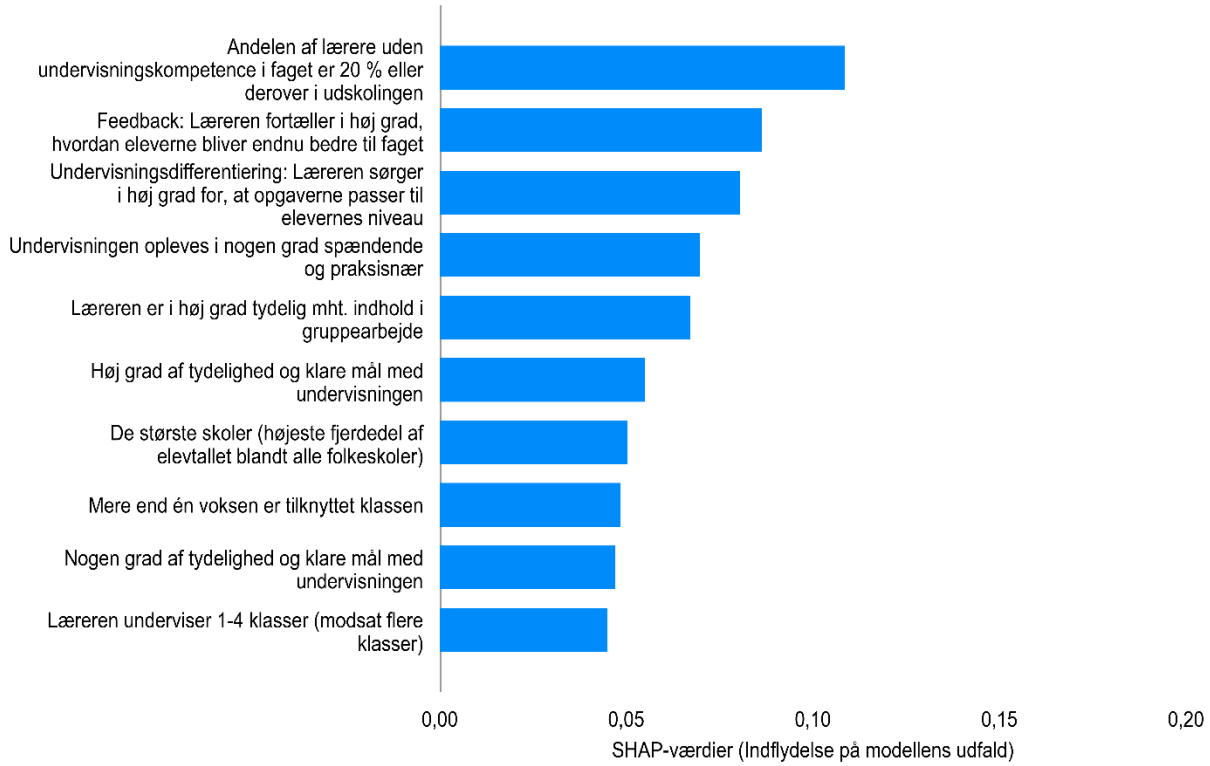
Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Det er særligt karakteristika omhandlende lærernes undervisningspraksis, der har størst korrelation med elevernes faglige resultater. For eksempel at eleverne oplever tydelighed og klare mål i undervisningen, at læreren differentierer undervisningen til den enkelte elevs niveau, hvorvidt lærerne formår at engagere eleverne i undervisningen, og at læreren sørger for ro i klassen. I forhold til karakteristika, der beskriver lærernes rammevilkår synes det særligt at være skolestørrelse og antallet af undervisningstimer, der slår igennem. Karakteristika om lærerbaggrund, såsom at læreren er uddannet meritlærer fremfor at have en ordinær eller anden uddannelse, korrelerer desuden også med sandsynligheden for elever med høje faglige resultater. Se rapporten for en indholdsmæssig diskussion af disse variabler.

Hvis vi sammenligner Figur 2.1 og Figur 2.2 er det dog tydeligt, at lærerkarakteristika og rammevilkår generelt synes at have lavere SHAP-værdier og dermed korrelerer mindre med elevernes faglige resultater end elevkarakteristika.

Ser vi i stedet på, hvilke karakteristika der korrelerer med, at eleverne har lave faglige resultater, så er det nogle af de samme karakteristika som går igen, men ikke alle (jf. Figur 2.3).

Figur 2.3 Korrelation mellem lærerkarakteristika og elever med lave faglige resultater. De 10 karakteristika med de højeste SHAP-værdier.



Anm.: Figuren viser SHAP-værdier for udfaldsmålet elever med lave faglige resultater. Vi anvender her gradient boosting-modellen beregnet på et datagrundlag for elever i 9. klasse for årene 2016-2018. Bemærk, at en høj SHAP-værdi kun siger noget om, at variablen korrelerer, men ikke hvordan den korrelerer med udfaldsmålet. Dermed kan det ikke fra denne figur udledes, om et karakteristika har en positiv eller negativ indflydelse på elevernes faglige resultater, men blot at den har en indflydelse.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Figur 2.3 viser, at det fx også er det formelle uddannelsesniveau i hele udskolingen, som korrelerer med elevernes faglige resultater, ligesom feedback, og at undervisningen opleves som spændende, også synes at korrelere med, at eleverne løftes fra bunden. Der er dog også en række "gengangere", såsom undervisningsdifferentiering og tydelighed i undervisningen.

Da metoden til at beregne SHAP-værdier er usikker, er det svært at udtale sig om den præcise rækkefølge af karakteristika. I de følgende analyser medtages alle lærerkarakteristika, der slår ud som værende betydningsfulde på baggrund af SHAP-værdierne i både Figur 2.2 og Figur 2.3.

Alle machine learning-analyserne er foretaget på baggrund af et datasæt på individniveau. Det vil fx sige, at vi har de enkelte elevers oplevelse af undervisningen. Det ville nok være mest rigtigt at anvende klassens vurdering af lærerne, eftersom det er lærerkarakteristika, som vi er interesseret i. Aggregering af data til klasseniveau giver dog et meget lille datagrundlag og machine learning-modellen har en tendens til at overfitte. Det vil sige, at modellerne har for høj varians og præsterer dårligt på testdata. En mere usikker model giver også mere tilfældighed i forhold til de udvalgte karakteristika. Vi har dog også foretaget analysen på et aggregeret datasæt, og det er overvejende de samme karakteristika, der udvælges.

2.3 Forskel på korrelationen i dansk og matematik

Vi tester om ovenstående variabler i højere grad korrelerer med elevernes resultater i dansk end i matematik. Til dette anvender vi en lineær sandsynlighedsmodel.

Vi interagerer det enkelte lærerkarakteristika med en indikator for, at karakteren er givet i dansk. På denne måde er det, ved at se på fortegnet af koefficienten, muligt at identificere om det enkelte karakteristika har større forklaringskraft i dansk end i matematik.

I analyserne af elever med høje faglige resultater er det tre ud af de ti karakteristika (se Figur 2.2), hvor der er signifikant forskel[6] på, hvor meget de korrelerer med elevernes faglige resultater i dansk sammenlignet med matematik. Boks 2.1 viser disse tre karakteristika.

Boks 2.1 Lærerkarakteristika og rammevilkår, hvor der er forskel på korrelationerne i dansk og matematik for elever med høje faglige resultater

- Forholdsvis små skoler (næstlaveste fjerdedel af elevtallet blandt alle folkeskoler)
- Høj grad af elevengagement i undervisningen
- Forholdsvis få antal fagtimer (næstlaveste fjerdedel af timetallet blandt alle folkeskoler)

For de to karakteristika om skolestørrelse og antallet af fagtimer gælder, at de korrelerer mere med de faglige resultater i matematik end i dansk. For karakteristika vedrørende elevengagement gælder, at det korrelerer mere med de faglige resultater i dansk end i matematik.

Ser vi i stedet på elever med lave faglige resultater, så er der for ingen af de ti identificerede lærerkarakteristika og rammevilkår signifikant forskel på, hvor meget de korrelerer med elevernes faglige resultater i dansk og matematik.

Samlet set er der derfor meget lidt, der tyder på, at de udvalgte karakteristika har større betydning for at løfte elevernes faglige resultater i det ene fag frem for det andet.

[6] Vi viser de karakteristika, hvor forskellen mellem dansk og matematik er signikant på minimum 5 %-niveau. Dvs. at sandsynligheden for, at forskellen er tilfældig, maksimalt er 5 %.

3 Data og repræsentativitet

I dette kapitel beskriver vi først undersøgelsens datagrundlag og dernæst resultaterne af en repræsentativitetsanalyse af de elever og lærere, der indgår i undersøgelsens stikprøve.

3.1 Datagrundlag

Vores datagrundlag er i udgangspunktet alle elever i 9. klasse med en karakter fra folkeskolens afgangsprøve i enten dansk eller matematik i skoleårene 2015/2016-2017/2018. Til denne population kobler vi oplysninger om baggrundskarakteristika for eleverne og deres lærere fra Danmarks Statistiks registre, og vi tilføjer oplysninger om lærernes kompetencer, som stammer fra Styrelsen for IT og Læring (STIL) og oplysninger om meritlærerstatus fra Forsknings- og Uddannelsesministeriet.

Vores oplysninger om elevernes oplevelse af undervisningen stammer fra elevernes survey-besvarelser fra følgeforskningspanelet til folkeskolereformen. I denne survey har mere end 200 skoler deltaget i foråret 2016-2018 med blandt andet besvarelser fra elever i 9. klasse. Vi har også forsøgt at kombinere datasættet med survey-besvarelser fra lærerne fra samme følgeforskningspanel. Det gjorde vi for at få flere karakteristika med om undervisningen fra lærerens perspektiv. I sidste ende var det dog hovedsageligt elevernes svar, som i analyserne havde den største betydning og blev valgt ud i machine learning-analyserne. De få karakteristika fra lærer-survey, som machine learning-analysen valgte ud, omhandlede derudover meget samme forhold som dem fra elev-survey (fx larm i klassen). Derfor blev svar fra lærerne droppet til fordel for en næsten dobbelt så stor stikprøve.

Vores endelige datasæt består af 8.505 elever, hvor vi både har besvarelser om elevernes oplevelse af undervisningen i dansk eller matematik, og hvor der er en karakter fra 9. klasses afgangsprøve i dansk eller matematik for eleven. Datasættet indeholder en lang række variable, der knytter sig til eleverne samt til lærernes baggrund, undervisningspraksis og rammebetingelser. Nogle af variablene er dannet på baggrund af enkelte spørgsmål i elev-surveyen, mens andre variable er indeks. Indeks er her lavet på baggrund af spørgsmål, som er afstemt til en dansk skolekontekst og evaluering af folkeskolereformen, og i mindre grad på baggrund af validerede skalaer. Indeks er derefter valideret med faktoranalyse. Tabel 3.1 viser den fulde liste af variabler, der indgår i machine learning-modellerne.

Tabel 3.1 Oversigt over variable, som indgår i machine learning-modellerne

Overordnet kategori	Variabel
Udfaldsmål	Elevens standardiserede gennemsnitlige karakter i enten dansk eller matematik i folkeskolens afgangsprøve [a)]
	Elever med høje faglige resultater (karakter ligger i øverste 25 %)
	Elever med lave faglige resultater (karakter ligger i nederste 25 %)
Lærerens baggrund	Lærerens karaktergennemsnit fra gymnasiet
	Læreren har sygefravær i året (ja/nej)
	Lærerens alder i tre variable (< 35 år, 35-45 år, > 45 år)
	Læreren er uddannet meritlærer (modsat en traditionel læreruddannelse eller andet)
	Læreren er uddannet fra den traditionelle læreruddannelse (modsat en meritlærer-uddannelse eller andet)
	Læreren har en anden uddannelse (modsat en traditionel læreruddannelse, en merit-læreruddannelse eller ingen uddannelse højere end gymnasiet)
	Læreren har ingen uddannelse højere end gymnasiet (modsat en traditionel lærer-uddannelse, en meritlæreruddannelse eller anden uddannelse)
	Læreren har ingen undervisningskompetencer (modsat formelle eller tilsvarende kompetencer)
	Læreren har formelle undervisningskompetencer (modsat tilsvarende kompetencer eller ingen undervisningskompetencer)
	Læreren har tilsvarende kompetencer (modsat tilsvarende eller ingen undervisnings-kompetencer)
	Lærerens erfaring i tre kategorier (< 2år, 2-10 år, > 10 år) [b)]
	Lærerens køn (mand vs. kvinde)
	Lærerens etnicitet (etnisk dansk vs. anden etnisk herkomst)
Undervisning	Indeks om elevengagement [c)] ▪ Jeg kan godt lide [dansk/matematik] ▪ Jeg hører efter, hvad min lærer siger i [dansk-/matematiktimerne] ▪ Jeg keder mig i [dansk-/matematiktimerne] Bliver til tre variable: høj grad/nogen grad/lav grad af elevengagement i undervisningen
	Indeks om lærer-elev-relation [c)] ▪ Min [dansk-/matematiklærer] gør noget for, at jeg har det godt i klassen ▪ Min [dansk-/matematiklærer] lytter til mig i timerne ▪ Jeg respekterer min [dansk-/matematiklærer] ▪ Min [dansk-/matematiklærer] er en dygtig underviser Bliver til tre variable: høj grad/nogen grad/lav grad af lærer-elev-relation
	Indeks om tydelighed og klare mål med undervisningen [c)] ▪ Min [dansk-/matematiklærer] fortæller mig tit, hvordan jeg klarer mig i forhold til de mål, der er i [dansk/matematik] ▪ Jeg ved, hvad jeg skal lære i [dansk/matematik] timerne ▪ Jeg ved, hvordan jeg bliver bedre til [dansk/matematik] Bliver til tre variable: høj grad/nogen grad/lav grad af tydelighed og klare mål med undervisningen
	Indeks om elevernes holdning til om undervisningen er spændende og praksisnær [c)] ▪ Min [dansk-/matematiklærer] giver mig spændende opgaver ▪ Min [dansk/matematiklærer] viser mig tit, hvad [dansk/matematik] kan bruges til i min hverdag ▪ Jeg kan bruge det jeg lærer i [dansk/matematik] uden for skolen Bliver til tre variable: undervisningen opleves i høj grad/nogen grad/lav grad som spændende og praksisnær

Overordnet kategori	Variabel
	Feedback: Læreren fortæller, hvordan eleven kan blive endnu bedre til faget (enig vs. hverken eller og uenig). Dannet ud fra et enkelt spørgsmål fra elev-survey: Min [dansklærer/ matematiklærer] fortæller mig tit, hvordan jeg kan blive endnu bedre til [dansk/matematik].
	Undervisningsdifferentiering: Læreren sørger i høj grad for, at opgaverne passer til elevernes niveau (enig vs. hverken eller og uenig). Dannet ud fra et enkelt spørgsmål fra elev-survey: Min [dansklærer/ matematiklærer] sørger tit for, at de opgaver jeg får i [dansk/matematik], passer til mit niveau.
	Læreren er i høj grad tydelig med hensyn til indhold i gruppearbejde (enig vs. hverken eller og uenig). Dannet ud fra et enkelt spørgsmål fra elev-survey: Når vi arbejder i grupper i [dansk/matematik] er det tydeligt, hvad det er, læreren vil have os til at lave.
	Læreren sørger for ro i klassen (enig vs. hverken eller og uenig). Dannet ud fra et enkelt spørgsmål fra elev-survey: Min [dansklærer/ matematiklærer] sørger for, at der er ro i klassen.
	Eleverne bliver sjældent forstyrret af larm i timerne (bliver aldrig eller ikke så tit forstyrret af larm i timerne vs. bliver altid, for det meste eller nogen gange forstyrret af larm i timerne). Dannet ud fra et enkelt spørgsmål fra elev-survey: Jeg bliver forstyrret af larm i timerne.
Rammevilkår	Mere end én voksen er tilknyttet klassen i faget (vs. én lærer tilknyttet klassen i faget)
	Sekundær voksen tilknyttet klassen i faget har ingen undervisningskompetencer
	Sekundær voksen tilknyttet klassen i faget har formelle undervisningskompetencer
	Sekundær voksen tilknyttet klassen i faget har tilsvarende kompetencer
	Timetal: Fagtimer klassen modtager i alt på tværs af fag opdelt i fire variable (opdelt i 4 kvartiler)
	Klassen modtager mindstetimetallet eller derover i faget (vs. under mindstetimetallet)
	Andelen af lærere uden undervisningskompetence i faget er 20 % eller derover i udskolingen
	Antallet af klasser som læreren underviser i opdelt i to variable (underviser i 1-4 klasser og underviser i flere end 4 klasser)
	Skolestørrelse opdelt i fire variable (elevtal opdelt i 4 kvartiler)

Anm.: I modellen tager vi desuden højde for en række baggrundsforhold blandt eleverne. Disse dækker over elevens køn, etnicitet (ikke-vestlig), alder, antal søskende, og om eleven bor sammen med begge forældre, forælderens uddannelsesniveau (højeste i familien) samt henholdsvis mors og fars indkomst i kvartiler. Vi kontrollerer også for, om moren var 'ung mor', hvilket vil sige < 25 ved barnets fødsel, samt elevernes tidligere faglige resultater i nationale test i 6. kl.

Noter: [a)] Der er taget et gennemsnit af karakteren i alle prøver i henholdsvis dansk og matematik. Dette gennemsnit er derefter blevet standardiseret til et gennemsnit på 0 og en standardafvigelse på 1. Et indeks på 0-0,5 betyder lav grad af det pågældende indeks, fx elevengagement, et indeks på 0,05-0,08 betyder nogen grad af det pågældende indeks, og et indeks på højere end 0,8 betyder høj grad af det pågældende indeks.

[b)] Erfaring er målt som antal år siden seneste fuldførte uddannelse.

[c)] De fire indeks er dannet ud fra spørgsmål i spørgeskemaundersøgelsen vedr. folkeskolereformen. Disse er fundet på baggrund af faktoranalyse som er et redskab, der kan opsamle samvariation i flere spørgsmål, der handler om det samme i én variabel (for yderligere detaljer om faktoranalyse, se Bilag 1 i Nielsen, Jensen, Kjer & Arendt, 2020).

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd

3.2 Repræsentativitetsanalyse

Bortfald er helt almindeligt i forbindelse med spørgeskemaundersøgelser. I tilfælde med systematik i bortfaldet kan dette imidlertid underminere muligheden for at generalisere på baggrund af de resultater, som produceres på baggrund af stikprøven. Derfor er det helt centralt at undersøge, i hvilket omfang analyseudvalget, dvs. de elever, der svarer på spørgeskemaet, og tilhørende tilknyttede lærere, er repræsentative for alle danske skoleelever og deres lærere.

Dette gøres ved at sammenligne baggrundskarakteristika af eleverne og lærerne i vores analyseudvalg med baggrundskarakteristika for den samlede population af elever og lærere i folkeskolen.

Tabel 3.2 viser fordelingen af lærerne på følgende tre baggrundskarakteristika:

- Andelen af mænd
- Alder (i tre kategorier)
- Andelen af lærere fordelt på kompetencedækning (i tre kategorier)

Sammenligningerne er foretaget ved t-test, som er en simpel sammenligning af gennemsnittene i populationen og stikprøven. Tabellen viser, at ingen af forskellene i andele mellem populationen og analyseudvalget er større end 4-5 procentpoint. Derfor vurdereres det, at analyseudvalget i store træk er repræsentativt for populationen på trods af, at der er statistisk signifikante forskelle på de fleste af de forskellige karakteristika.

Tabel 3.2 Repræsentativitetsanalyse for population og stikprøve af lærerne. Dansk og matematik. Sammenligning af gennemsnit ved t-test.

Dansk og matematik				
Andele	Population	Stikprøve	Forskel	
Mænd	0,39	0,38	-0,01	
Under 35 år gammel	0,18	0,20	0,02	
Mellem 35-45 år gammel	0,36	0,36	0,00	
Over 45 år gammel	0,45	0,44	-0,01	
Ingen undervisningskompetencer	0,05	0,02	-0,03	***
Tilsvarende kompetencer	0,16	0,14	-0,02	
Formelle undervisningskompetencer	0,79	0,84	0,05	***
Antal lærere	9.566	860		

Anm.: Sammenligninger af gennemsnit ved t-test. *** = p < 0,01, ** = p < 0,05, * = p < 0,1. Populationen, der sammenlignes med, består af alle folkeskolelærere, som underviser i dansk eller matematik i 9. klasse i skoleårene 2015/2016-2017/2018.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Repræsentativitetsanalysen viser, at der er en lille forskel i kompetenceniveauet på lærerne i vores stikprøve set i forhold til alle folkeskolelærere i Danmark, som underviser i dansk eller matematik i 9. klasse. Vores stikprøve har flere lærere med formelle kompetencer og færre lærere uden undervisningskompetencer, end der generelt er i populationen. Der er derimod ingen forskel i andelen af mandlige og kvindelige lærere og lærernes alder.

På samme måde sammenligner vi de elever, som indgår i vores undersøgelse med de resterende elever i 9. klasse, ud fra følgende karakteristika:

- Andelen af drenge
- Andelen af ikke-vestlige elever (vs. vestlig herkomst)
- Andelen af elever med forældre, hvis højest fuldførte uddannelse er gymnasiet eller grundskolen.

Tabel 3.3 viser fordelingen af elever på ovenstående baggrundskarakteristika.

Tabel 3.3 Repræsentativitetsanalyse for populationen og stikprøven af eleverne. Sammenligning af gennemsnit ved t-test.

	Dansk og matematik			
Andele	Population	Stikprøve	Forskel	
Drenge	0,52	0,51	-0,01	
Ikke-vestlig baggrund	0,09	0,12	0,03	***
Forældre har gymnasiet eller grundskolen som højest fuldførte uddannelse	0,13	0,12	-0,01	***
Kommer ikke fra kernefamilie	0,29	0,28	-0,01	***
Antal elever	122,717	8,505		

Anm.: Sammenligninger af gennemsnit ved t-test. *** = p < 0,01, ** = p < 0,05, * = p < 0,1. Populationen, der sammenlignes med, består af alle folkeskoleelever i 9. klasse med en karakter fra afgangsprøven i dansk eller matematik i skoleårene 2015/2016-2017/2018.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Repræsentativitetsanalysen af eleverne viser, at vi har en lille overvægt af elever med ikke-vestlig baggrund (3 procentpoint). Vi har også en lidt lavere andel elever, hvis forældre har gymnasiet eller grundskolen som højeste uddannelse eller lavere andel elever, som ikke kommer fra en kernefamilie (forskel på 1 procentpoint). Da der her også er tale om meget små forskelle, vægter vi ikke analyserne.

Litteratur

Ben-David, A. (2008). About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of Artificial Intelligence*, *21*(6), 874–882.

Bingley, P., Heinesen, E., Krassel, K.F. & Kristensen, N. (2019). *The Timing of Instruction Time: Accumulated Hours, Timing and Pupil Achievement,* IZA Discussion Paper. Bonn: IZA – Institute of Labor Economics.

Cornell-Farrow, S. & Garrard, R. (2018). *A machine learning approach for detecting students at risk of low academic achievement*. arXiv preprint, arXiv:1807.07215.

Fernández-Delgado, M., Carnadas, S. Barro, Amorim, D. (2014). Do we need hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research,* (15), 3133-3181.

Hanushek, E. (2020). "Education production functions" in: Bradley, S.C. & Green, C. (Eds.) *The Economics of Education – a comprehensive overview.* 2. ed. London: Academic Press, 161-170.

Jensen, V.M., Bjørnholt, B., Mikkelsen, M.F., Nielsen, C.P. & Ladekjær, E. (2020). *Den længere og mere varierede skoledag – en analyse af reformens elementer.* København: VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

Jensen, V.M. & Mikkelsen, M. (igangværende). The added value of additional classroom hours with higher quality teachers on student performance.

Jensen, V.M. & Nielsen, L.P. (2010). *Veje til ungdomsuddannelse 1. Statistiske analyser af folkeskolens betydning for unges påbegyndelse og gennemførelse af en ungdomsuddannelse.* Rapport 10:24. København, SFI – Det Nationale Forskningscenter for Velfærd.

Kristensen, N. & Skov, P.R. (2019*). Betydningen af kompetencedækning og læreruddannelsesbaggrund.* København: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv, abs/1705.07874*.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A. et al. (2020). *From local explanations to global understanding with explainable AI for trees*. *Nature Machine Intelligence*, 2, 56–67.

Müller, A.C. & Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. 1. ed. Sebastopol, CA: O'Reilly Media.

Nielsen, C., Jensen, V.M., Kjer, M.G. & Arendt, K.S. (2020). *Elevernes læring, trivsel og oplevelser af undervisningen i folkeskolen, En evaluering af udviklingen i reformårene 2014-2018*. København: VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

Olson, R.S., Cava, W., Mustahsan, Z., Varik, A. & Moore, J.H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. *Pacific Symposium on Biocomputing*, *23*, 192–203.

Rangvid, B. (2008). *Skolegennemsnit af karakterer ved folkeskolens afgangsprøver - Korrektion for social baggrund*. AKF Working Paper. København: AKF - Anvendt Kommunal-Forskning.

Raschka, S. & Mirjalili, V. (2019) *Python Machine Learning.* 3. ed. Birmingham, UK: Packt Publishing.

Strøm, B. & Falch, T. (2020) ”The Role of teacher quality in education production”, in: Bradley, S. & C. Green (Eds.), *The Economics of Education – a comprehensive overview*, 2. ed. London: Academic Press, 307-319.

Bilag 1 Supplerende tabeller og figurer

I dette bilag præsenteres supplerende resultater for machine learning-modellerne.

Bilagstabel 1.1 Resultater for optimeringen af modellerne. Bedste parametre og præcision på træningsdata, hvor udfaldsmålet er elever med høje faglige resultater

Model	Præcision på træningsdata	De bedste (hyper)parametre
Logistisk regression	0,82	Ingen
Logistisk regression med L2-regularization	0,81	C: 0,33
Random forest	0,81	Maks. dybde: 25 Antal features: 38 Antal træer: 7.502
Gradient boosting	0,82	Learning rate: 0,005 Maks. dybde: 1 Antal træer: 3.775

Anm.: C er reguleringsparametret, der bestemmer, hvor meget der reguleres, dvs. hvor meget begrænsning der er på koefficienternes størrelse i den logistiske regression. Maks. dybde er det maksimale antal splits, der kan være i hvert decision tree. Antal features er antal variable, der udtrækkes til hvert decision tree i Random Forest. Antal træer er det antal estimatorer, der samles i ensemble-modellerne Random Forest og Gradient Boosting. Learning rate er den rate, hvormed vægtene i Gradient Boosting opdateres.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

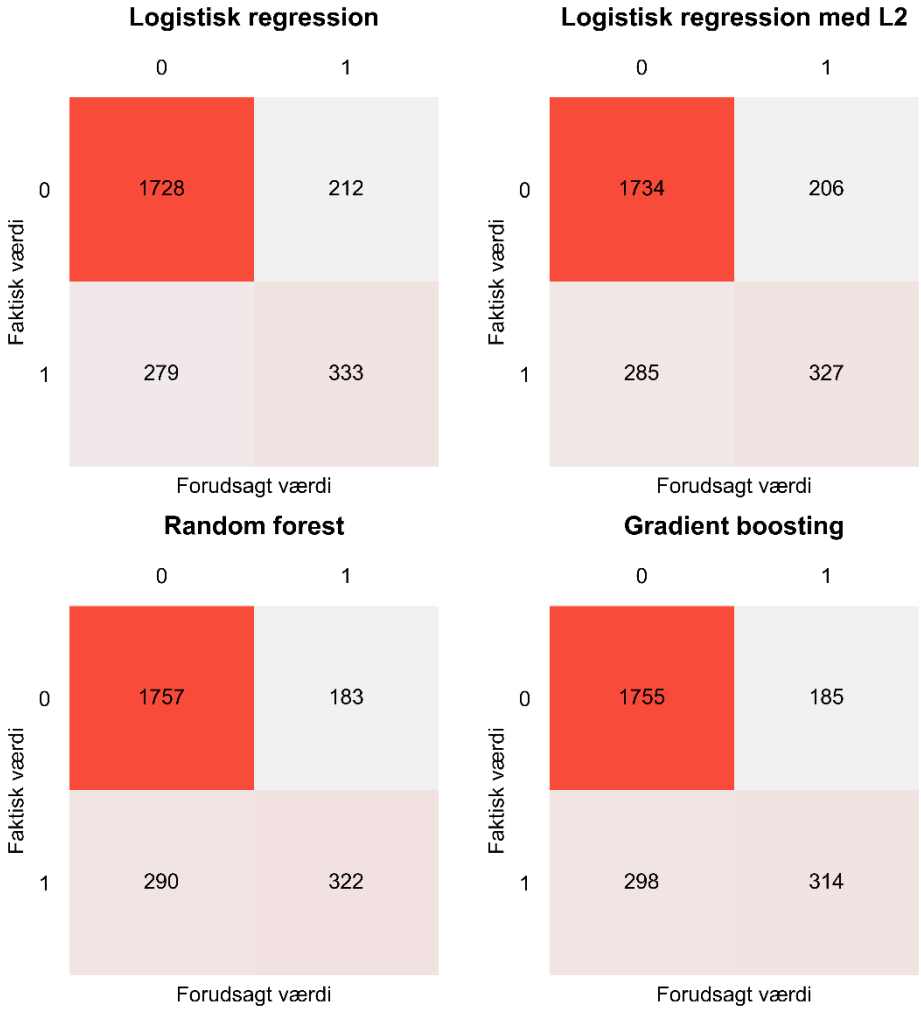
Bilagstabel 1.2 Præcision og Cohen's kappa for modellerne på testdata, hvor udfaldsmålet er elever med høje faglige resultater

Model	Præcision	Cohen's kappa
Logistisk regression (benchmarking)	0,82	0,48
Logistisk regression med L2-regularization	0,82	0,48
Random forest	0,83	0,48
Gradient Boosting	0,82	0,47

Anm.: Præcision angiver andelen af observationer, der er korrekt klassificeret på testdata. Cohen's kappa er et mål, hvor præcisionen sammenlignes med en tilfældig model. En værdi på 1 vil være den perfekte forudsigelse af lærernes løfteevne, og en værdi på 0 vil være en fuldstændig tilfældig forudsigelse af elevernes faglige resultater.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

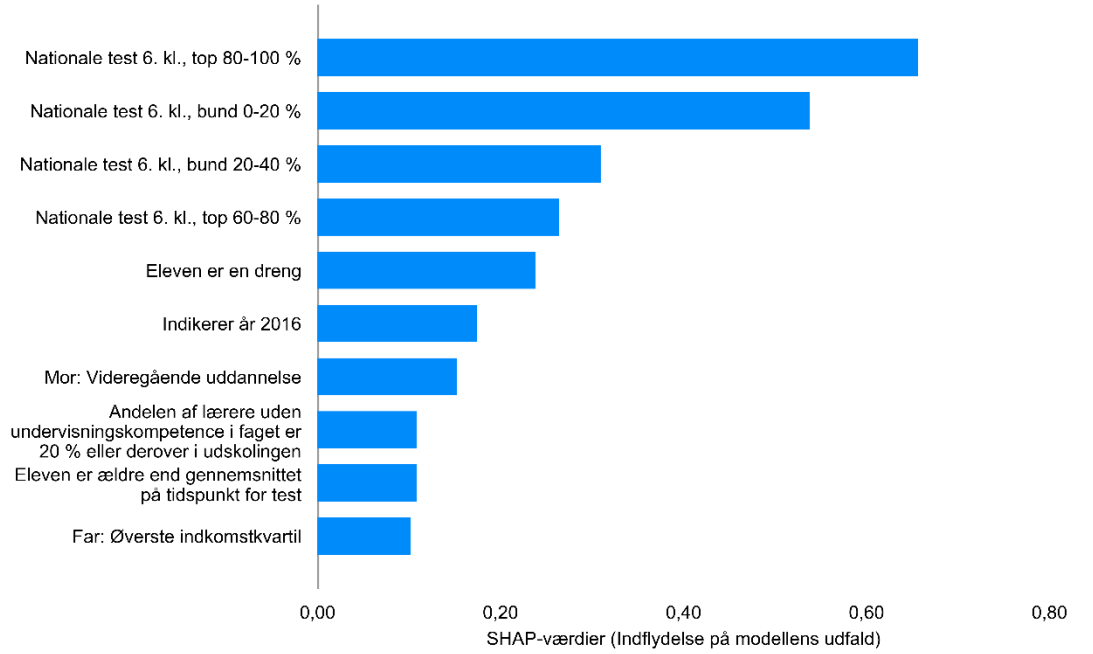
Bilagsfigur 1.1 Confusion matrix for de fire modeller, hvor udfaldsmålet er elever med høje faglige resultater



Anm.: En confusion matrix er en tabel, der viser hvor mange gange modellen gætter korrekt eller forkert. Lodret ses de faktiske værdier af udfaldene 0 og 1, vandret ses de forudsagte værdier af 0 og 1. Det øverste felt til venstre og det nederste felt til højre angiver altså de observationer, der er korrekt klassificeret, mens det øverste felt til højre og det nederste felt til venstre angiver de observationer, der er forkert klassificeret.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

Bilagsfigur 1.2 De ti karakteristika med højeste SHAP-værdier, hvor udfaldsmålet er sandsynligheden for at være blandt elever med lave faglige resultater



Anm.: Figuren viser SHAP-værdier for udfaldsmålet for elever med lave faglige resultater. Vi anvender her Gradient Boosting-modellen beregnet på et datagrundlag for elever i 9. klasse for årene 2016-2018. Bemærk, at en høj SHAP-værdi kun siger noget om, at variablen korrelerer, men ikke hvordan den korrelerer med udfaldsmålet. Dermed kan der ikke fra denne figur udledes, om et karakteristika har en positiv eller negativ indflydelse på elevernes faglige resultater, men blot at den har en indflydelse.

Kilde: VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd.

VIVE

DET NATIONALE FORSKNINGS- OG ANALYSECENTER FOR VELFÆRD