AALBORG
UNIVERSITET

Vejen til eliten:

Kan dit tidlige liv prædiktere din vej til toppen af det sociale hierarki?

– en transformerbaseret prædiktion af eliteudfald med dansk registerdata

Sociologi – med særlig specialisering i Social Data Science

Kandidatspeciale, August 2025

Anne Helene Lund & Camilla Klyvø

Vejleder: Andreas Lindegaard Jakobsen

Anslag: 179.067

Abstract:

This master thesis in the intersection of sociology and social data science investigates the potential of transformer-based prediction models to identify rare but significant life course outcomes based on childhood trajectories. Using an extensive volume of Danish register data, we model life histories of 140,708 individuals in a 3-year birth cohort from birth to age 17 tracking ~60 background variables with yearly observations, capturing a wide range of dynamic, static and irregular family-related, socio-economic and socio-spatial indicators.

Our primary objective is to explore how recent developments in transformer architecture adapted to a social science context can adapt tabular data to text sequences simulating language to predict low-prevalence outcomes such as elite attainment in adulthood at age ~40. We operationalize six different elite attainment outcomes for binary prediction; 'income') being among the highest five percentiles in annual income; 'area') living in the top five most elitist micro areas in Denmark; 'educational') having attained a high prestige education with the highest average income levels; 'city') living in one of four major cities in combination with a master's degree and a 1.5xhigher-than-median annual income level; 'managerial') possessing a high-level managerial position in the labour market or 'self-employed') being self-employed in combination with a 2xhigher-than-median annual income level. We tokenize all background variables and events during childhood and their according annual time steps, convert them into a long text sequence per childhood trajectory, apply label encoding and sequence padding, attention mask, embedding layers and learnable position embeddings, and train the model using supervised deep learning similar to language models with dynamic instance weighting and focal loss to address class imbalance.

The model for 'income elite' performs a ROC AUC of 0.757 while the model for 'city elite' performs a ROC AUC of 0.755. The models for 'educational elite' and 'area elite' performs ROC AUC's of 0.713 and 0.720 respectively, while models for 'managerial elite' and 'self-employed elite' which were trained on less data due to very small positive outcome subgroups, performs lower ROC AUC scores of 0.648 and 0.647 respectively. Threshold tuning ensures robust out-of-sample evaluation and allows predictions to focus on higher recall for positive outcomes for potential further analytical exploration of clusters and subgroups.

Our findings suggest that transformer-based models analyzing life trajectories as language offer a powerful, domain-flexible approach for detecting latent life course patterns in register-based population data, while data volume proves important as well as extreme class imbalances with positives less than 5% of the total data volume can potentially present challenges for training and prediction results.

Tabeloversigt:

Figuroversigt:

Confusion Matrix oversigt:

Bilagsoversigt:

Brug af **generativ AI (GAI)** i specialet

GAI er anvendt til støtte og sparring i udvalgte dele af specialearbejdet, herunder grundet den metodisk innovative tilgang i særlig grad til metodisk idéudvikling og justering, generering af Python-kode og fejlfinding i forbindelse med kodning. GAI er desuden anvendt som refleksiv sparringspartner til litteratursøgning, idégenerering, ensretning af kildehenvisninger, sprog- og grammatikkorrektioner, kommasætning, mm. Vi har til enhver tid været bevidste om at benytte GAI i overensstemmelse med AAU's retningslinjer og principper for god akademisk praksis.

Indhold:

1 Indledning

Hvor meget af vores liv er egentlig et resultat af egne valg – og hvor meget ligger allerede fast, længe før vi selv begynder at vælge? De fleste af os vil nok have en tendens til at holde fast i en forestilling om, at vi selv træffer aktive valg, og at de vigtigste beslutninger i vores liv er nogle, vi selv er herre over – eller at vi i det mindste har stor indflydelse på, hvordan vores liv former sig. Vi vælger selv uddannelse, partner, job og livsstil. I et samfund som det danske, hvor idéen om lige muligheder uanset baggrund står stærkt, vil mange nok hævde, at vi i høj grad former vores egen fremtid. Men selv i et egalitært samfund må vi erkende, at ikke alle starter fra samme udgangspunkt.

Samtidig er vi klar over, at de rammer, vi vokser op i, ikke er uden betydning for, hvordan vores livsmuligheder udfolder sig. En række forudsætninger og rammer er allerede på plads, når vi fødes – forældre, etnicitet og køn. Dem kan vi ikke ændre på. Meget andet beslutter vores forældre for os i løbet af barndommen; institutioner og skole, fritidsaktiviteter, bopæl, m.m. Vi er altså relativt gamle, når vi selv begynder at træffe beslutninger om, hvilken retning vores liv skal tage. Der er fornuft i, at vi ikke skal træffe store livsændrende valg med umodne hjerner. Men ikke desto mindre betyder det, at de rammer, vi bliver givet i barndommen, ikke er fastlagt af os selv, men af vores nærmeste. Det er bredt anerkendt i samfundsvidenskabelig forskning, at sociale og økonomiske forhold i barndommen har betydning for, i hvilken retning livet udvikler sig. Disse forhold har ikke kun betydning for, hvor vi ender i livet, men kan også have betydning for, de barrierer vi må kæmpe imod for at nå bestemte positioner, mens andre helt undgår dem. Skal du fødes ind i samfundets øverste lag for at ende samme sted i voksenlivet? Spørgsmålet er derfor ikke blot, om vi alle har lige muligheder – men i hvor høj grad vores muligheder allerede formes af forhold uden for vores egen kontrol.

Det leder naturligt til en bredere samfundsmæssig diskussion om social reproduktion. Er succes og status noget, vi opnår via talent og hårdt arbejde eller i høj grad forudbestemt af sociale faktorer? Hvis tidlige livsmønstre afdækkes systematisk, er det så muligt at forudsige, hvem der opnår de højest rangerede sociale positioner i voksenlivet? Hvis ja – hvad siger dette om de strukturelle forhold, som eksisterer i vores samfund, og hvilke strukturer er ønskværdige? Og hvad siger det om vores idé om et samfund med lige muligheder for alle?

I dette sociologiske speciale undersøger vi, om sandsynligheden for at opnå særlige elite-positioner i voksenlivet kan forudsiges på baggrund af en række faktorer i barndommen. Ved at anvende registerdata, machine learning (ML) og en innovativ transformerbaseret model, anvender vi en prædiktiv metodisk tilgang, som endnu er sjældent anvendt i sociologisk forskning. Specialet bidrager således ikke kun med nye indsigter om social reproduktion og mobilitetsprocesser, men også med metodisk innovation, som kan inspirere fremtidige sociologiske analyser i retningen af nyere datadrevne tilgange. Specialet indledes med et teoretisk og empirisk problemfelt, hvor teori og nyere

forskning om social reproduktion og social mobilitet, og herefter teorier og studier af eliter og mekanismer bag deres sociale reproduktion, bidrager til at indplacere undersøgelsen i det brede emnemæssige forskningsfelt. I forlængelse af dette belyses mulighederne ved anvendelse af prædiktive modeller og ML i sociologisk forskning, hvorefter udvalgte eksisterende studier, som anvender lignende metoder, præsenteres og bidrager til at indplacere undersøgelsen i det metodiske forskningsfelt. Samlet udgør dette problemfelt grundlaget for en præcisering af specialets problemformulering og forskningsspørgsmål, som danner udgangspunkt for den endelige undersøgelse.

2 Problemfelt – forskning og teori

2.1 Social reproduktion og social mobilitet

Når vi afsøger forskningsfeltet omhandlende livsudfald, er noget af det første, vi støder på, begreber som social reproduktion og social mobilitet. I det følgende vil vi belyse, hvad der i eksisterende forskning og teori beskriver samfundet ift. livsudfald.

En af de førende sociologiske teoretikere inden for dette område er Pierre Bourdieu, der argumenterer for, at sociale strukturer og uligheder i høj grad reproduceres fra generation til generation. Han introducerer begrebet habitus, som omhandler individets indlejrede system af normer, vaner, holdninger og perspektiver på verden og dets medmennesker. Habitus beskriver summen af kulturelle forståelser og vanlige handlinger, som vi har tillært i forskellige sociale sammenhænge, og er derfor med til at afgøre, hvad vi tænker, og hvordan vi handler. Habitus er i høj grad formet af social baggrund og er derved en stor del af den sociale reproduktion, da næste generation arver tænke- og handlemåder. Dog udvikles habitus kontinuerligt i takt med, at livet leves og nye sociale sammenhænge opleves (Bourdieu, 2010:183-189; 491-511).

Bourdieu opererer med fire forskellige kapitalerformer, som overføres fra generation til generation, men som også selvstændigt kan forbedres; økonomisk (penge og materielle ressourcer), kulturel (uddannelse, generel dannelse, smag og livsstil), social (forbindelser, netværk adgang til indflydelse) og symbolsk kapital (status, anerkendelse og legitimitet). Bourdieu mener således, at en delvis mobilitet er mulig og eks. kan ske i kraft af uddannelse. Han understreger dog, at de, der forsøger at klatre op ad stigen, vil opleve symbolske kampe eller dobbeltheder i deres identitet, når de bevæger sig mellem forskellige felter. Adgangen til en højere social position end den man er født ind i, er mulig, men vanskelig, da det kræver, at man tilegner sig og udøver bestemte former for kapital på en legitim måde ift. de relevante sociale felter. På den måde anerkender Bourdieu, at social mobilitet er mulig, men understreger sin skepsis for at social mobilitet er fuldstændig frit for alle. Han understreger, at samfundet er præget af strukturelle uligheder, som gør det sandsynligt, at sociale klasser reproducerer sig selv (Bourdieu, 2002). Med dette perspektiv ønsker vi at undersøge, hvordan social reproduktion manifesterer sig i nyere empirisk forskning.

Flere nyere studier og analyser antyder, at social reproduktion også er et fænomen, der eksisterer i det moderne samfund. I et studie fra 2019 afsøges en dybere forståelse af, hvordan social ulighed reproduceres til næste generation, og hvordan nogle familier bryder med denne videreførelse af ulighed (Schoon & Melis). Studiets resultater viser, at social ulighed mellem generationer ikke kun er et spørgsmål om systematisk social arv, altså social reproduktion, men at næste generations sociale status også påvirkes af de sociale forandringer, der finder sted i samfundet, samt individets egne valg (Schoon & Melis, 2019).

Derimod viser et andet studie fra 2022 en stærk intergenerationel sammenhæng. Mødres uddannelsesniveau fremtræder her som den mest betydningsfulde faktor for børns uddannelses- og erhvervsmæssige præstationer i voksenlivet. Desuden frembringer studiet, at faktorer i barndommen - herunder individets egen personlighed og motivation - har stor betydning for dets sociale status i voksenlivet, endog at dette ligeledes er stærkt forbundet til mødres socioøkonomiske status. Således finder studiet intergenerationelle overførelser, eller sagt med andre ord; en grad af social reproduktion (Cassidy, McLaughlin & McDowell, 2022). Dette underbygger derved relevansen af at undersøge forhold i barndommens indflydelse på individets livsudfald og graden af social reproduktion.

En anden undersøgelse fra 2020, der ligeledes antyder, at der eksisterer en grad af social reproduktion og at forældres socioøkonomiske status præger individets livsforløb, er lavet Arbejderbevægelsens erhvervsråd (herefter AE). Undersøgelsen peger på at opvækstvilkårene i barndommen i høj grad præges af segregerede sociale klasser. Konkret ses det ved at børn i høj grad møder deres egne sociale spejlbilleder i både skole og nabolag. Et andet centralt element som præger individets livsforløb er, hvorvidt barnets forældre er sammen og om de er en del af arbejdsmarkedet. Særligt er de børn udsat, hvis forældre ikke bor sammen og som er uden for arbejdsmarkedet, da dette tydeligt har indflydelse på deres karakterer i skolen og om barnet i en tidlig alder er på vej ned af en kriminel vejbane (AE, 2020:21). Herudover viser undersøgelsen, at også valg i forhold til boligforhold, uddannelse, par- og familiedannelse og pensionsopsparinger er præget af personens opvækstvilkår og forældres sociale klasse (AE, 2020: 31, 43, 56-57).

Vi finder det interessant, at individets voksenliv i så høj grad antydes at være præget af, hvordan individets udgangspunkt og levevilkår i barndommen har været. Sådanne teorier og analyser skaber belæg for, at det er muligt at kunne forudsige individets sociale livsudfald på baggrund af dets første leveår eller summen af barndommen, fordi disse teorier og analyser understreger, at der eksisterer en høj grad af social reproduktion. Men er individet således fastlåst socialt, hvis dets opvækstvilkår har så stor indflydelse på de muligheder, de har? I nedenstående afsnit vil vi afsøge, hvad eksisterende forskning og teori peger på ift. social mobilitet.

2.1.1 Social mobilitet

Social mobilitet kan argumenteres for at være udfordrende i lande som USA, hvor offentlige ydelser er begrænsede og fokusset er placeret på valgfrihed og individets eget ansvar. Men selv i Danmark, hvor samfundet er baseret på en velfærdsmodel, hvor den offentlige sektor er større, og der er fokus på at tilbyde et offentligt sikkerhedsnet, som skaber mere lighed, viser der sig stadig udfordringer med social mobilitet (Heckman & Landersø, 2022). En rapport fra Rockwool fondens forskningsenhed har undersøgt forskellene mellem Danmark og USA ift. social mobilitet. De har i den forbindelse vist, at der eksisterer minimale forskelle mellem de to lande ift. social mobilitet igennem et livsforløb (Heckman & Landersø, 2021a; Heckman & Landersø, 2021b). Det tegner et billede af, at de umiddelbart lige muligheder, som det danske samfund understøtter i form af uddannelse og sundhedsvæsen for alle uanset baggrund, ikke har den store effekt på den sociale mobilitet, siden USA med en mindre offentlig sektor viser sammenlignelige muligheder for social mobilitet. Det kan således antages, at de formelt lige muligheder, der hævdes at eksistere i Danmark, i praksis ikke reelt er tilgængelige for alle. At der ikke reelt eksisterer lige muligheder i Danmark ift. social mobilitet viser sig ligeledes, når sammenhængen mellem forældres og børns indkomst og forbrugsmuligheder undersøges i et livsperspektiv (Ibid.). Konkret viser den intergenerationelle elasticitet ift. lønindkomst at være ca. 24 pct, hvilket vil sige, at når forældrenes lønindkomst stiger med 1 pct, stiger børnenes indkomst med 24 pct., når de når samme alder som forældrene. Men en større sammenhæng ses, når den forventede livsindkomst og værdien af forbrugsmulighederne hen over livet udregnes. Her stiger begge mål med ca. 60 pct. ift. forældrenes. Det er med andre ord ikke uden betydning for dine økonomiske vilkår og forbrugsmuligheder, hvilken familie du fødes ind i, da dine økonomiske forhold i høj grad kan forklares af forældrebaggrund. Dette understreger derved, at der i Danmark eksisterer udfordringer ift. social mobilitet (Ibid.). Men hvilke faktorer har egentlig betydning for hvorvidt individet kan flytte sig fra sit socioøkonomiske udgangspunkt?

Klassisk er det i sociologien at forstå individets sociale baggrund som værende betydningsfuld for, hvor det ender i voksenlivet ift. de sociale lag. Der eksisterer således flere faktorer, som er knyttet til individets sociale baggrund, der kan være relevante i afsøgningen af, hvilke faktorer der har betydning for individets mulighed for social mobilitet. Herunder både forældres uddannelsesmæssige opnåelse og økonomiske forhold. Udover de ofte anvendte faktorer peger nogle studier på, at også søskendekonstellationer, køn, samt geografiske og nabolagsrelaterede forhold kan have indflydelse på individets livsmuligheder og sociale positionering (Brenøe, 2021; Brenøe & Zölitz, 2018).

Forskningsgruppen SocMap undersøger, hvordan geografiske forhold påvirker sociale vilkår. En nyere artikel viser, at unge, der vokser op i udsatte boligområder, har øget risiko for hverken at være i uddannelse eller beskæftigelse i det tidlige voksenliv – uanset etnicitet. Dog viser familieforhold sig som en stærkere forklaringsfaktor end geografi (Jakobsen et al., 2024).

Flere studier peger på geografi og nabolagets sociale sammensætning som centrale for social mobilitet. Et hollandsk studie finder, at børn i velstående områder oftere opnår højere uddannelse, og at velstand i nabolaget har større betydning end fattigdom – især for børn af lavt- og middeluddannede forældre (Troost, van Ham & Manley, 2023:10-13). Et norsk studie viser tilsvarende, at unge i områder med mange højstatus-individer har øget sandsynlighed for at tage længere, elitære uddannelser og selv opnå høj social status senere i livet (Toft & Ljunggren, 2016:2947-2952). I Danmark viser Landersø et al. (2024), at forskelle i mobilitet i høj grad skyldes, at familier – bevidst eller ubevidst – vælger boligområder ud fra sociale karakteristika. Dette underbygges af et studie fra Rockwool Fonden, der finder, at familier typisk bosætter sig blandt ligesindede eller i områder, der klarer sig lidt bedre – hvilket styrker social immobilitet (Heckman & Landersø, 2021a:3-5;27-28).

Samtidig fremhæver Landersø et al., at der findes mobilitetseffekter, som ikke kan forklares af familieforhold alene. Det antyder, at nabolaget i sig selv har en selvstændig effekt. De finder også ikke-lineære effekter af bl.a. indkomst og mødres arbejdsdeltagelse: Små forbedringer hos udsatte familier har større effekt på børns mobilitet end tilsvarende forbedringer hos ressourcestærke familier (Landersø et al., 2024). Samlet viser studierne, at områdeeffekter har betydning for social mobilitet – men at effekterne varierer afhængigt af individets baggrund og placering i det sociale rum.

Mens ovenstående afsnit har beskæftiget sig med teori og eksisterende forskning om generelle mekanismer for social reproduktion og socialmobilitet, afgrænses fokus i det følgende til at omhandle eliter. Her afsøges elitens relevans for den sociale stratifikation og ulighed i det moderne samfund, deres karakteristika og de mekanismer, som muliggør en social reproduktion på tværs af generationer i både teori og eksisterende forskning.

2.2 Forskning og teori om eliter

Uligheden i Danmark er stigende og mange nederst i samfundet oplever ikke længere at velfærdsstaten er garanten imod ledighed, sygdom og nedslidning (Olsen, L. et.al., 2021). I den modsatte ende af skalaen, oplever samfundets øverste lag stigende indkomster, en polarisering i retningen af liberalistiske værdier og et stigende ønske om bopæl iblandt ligesindede, gerne omkring hovedstadsområdet og Aarhus. Dette kan føre en udvikling med sig mod et tydeligt opdelt Danmark mellem 'storbyeliterne' og 'det perifere Danmark' (ibid: 9-10).

Afsøges udviklingen af de top 5 pct. rigeste danskeres indkomster, besad disse i 1995 16,6 pct. af de samlede bruttoindkomster i Danmark, mens dette er steget til 21,9 pct. i 2019 (ibid:14). Ikke nok med at udviklingen i Danmark peger på, at de øverste lag i samfundet oplever stigende indkomster i sammenligning med resten af befolkningen, så bliver gruppen af de individer, som kan klassificeres som overklassen eller øvre middelklasse, også større. Hovedårsagerne hertil kan skyldes både stigende indkomstforskelle på arbejdsmarkedet, samt de velståendes afkast fra aktiver og boliger. Yderligere skyldes det store samfundsændringer, som er sket med baggrund i erhvervsudviklingen, hvor flere

tjener store summe på brancher, der er i hastig vækst, såsom finans, forsikring, kommunikation, vidensservice, konsulentarbejde mm. Hvoraf mange af disse jobs er koncentreret omkring de større byer (ibid:12-13). Dette kan være med til at forklare udviklingen, der ses hos arbejderklassen, hvor størrelsen af gruppen er faldet fra at udgøre 60 pct. af befolkningen i 1987 til 40 pct. i 2019. Ligesom der er sket ændringer i ønskværdige bopæle for de velstående, er der sket en markant udvikling i arbejderklassens typiske bopæl. Sogne med høj andel af individer fra arbejderklassen er mere jævnt fordelt over Danmark i 1987, mens de i højere grad er koncentreret mod vest og væk fra storbyerne i 2019 (ibid:21-25). Ser vi mod samfundets bund, er der ligeledes sket en negativ udvikling, hvor personer langvarigt uden for arbejdsmarkedet er steget fra 16,8 pct. i 1987 til 19,9 pct. i 2019. Med andre ord er den eneste sociale gruppe, hvis størrelse er svundet, arbejderklassen (ibid:31).

I kraft af at ovenstående afsnit frembringer teori og empirisk forskning, som antyder, at der på en række områder eksisterer social reproduktion, og at social mobilitet ikke er lige let for alle sociale grupper – samt med udgangspunkt i den her belyste samfundsudvikling, hvor de velstående grupper øges i både størrelse og rigdom, mens bunden af samfundet ikke mindskes – finder vi det interessant at undersøge toppen af samfundet grundigere. Altså eliten.

2.2.1 Hvem er eliten?

For at forstå hvordan eliten formes og muligvis reproduceres er det nødvendigt at afklare, hvem eliten er, og hvilke kriterier der afgør individets medlemskab heri. Det følgende afsnit præsenterer centrale teoretiske perspektiver på elitebegrebet og frembringer, hvordan disse tilgange definerer og afgrænser eliten i et sociologisk perspektiv.

Genbesøger vi Bourdieus perspektiv, er eliten de sociale grupper i samfundet med flest ressourcer inden for hans forskellige kapitalbegreber. For at være en del af elite mener Bourdieu, at man ikke kun skal være rig (økonomisk kapital), men også besidde kulturel, symbolsk kapital og social kapital for at være i stand til at skabe retningen for, hvad der opfattes som 'god smag', viden og værdifuldt. Individet skal besidde en dominerende mængde af kapitalformerne og være i stand til at bruge sine kapitaler strategisk for at opretholde sin position i samfundets magtstrukturer. Hvilket sker gennem 'feltet' - Bourdieus begreb for sociale arenaer - hvor mængden af bestemte kapitalformer afgør magtforhold og status afhængigt af hvilke arenaer, der er tale om (Bourdieu, 2002; Bourdieu, 2010:491-511).

Et moderne perspektiv på eliter, som bygger videre på Bourdieus begreber om økonomisk og kulturel kapital, fremføres af økonomen Thomas Piketty. Selvom han ikke direkte benytter betegnelsen "elite", beskæftiger han sig med "toppen af det sociale hierarki", hvilket i dette speciale forstås som hans pendant til eliten. Piketty identificerer to former for eliter i det moderne samfund: den proprietære elite, som ejer store formuer gennem ejendom, aktier og virksomheder, og som udøver betydelig indflydelse via fx skatteplanlægning og lobbyisme; samt en meritokratisk elite, bestående af

højtuddannede og velbetalte professionelle som akademikere, embedsmænd og ledere, der har magt via ekspertise og position, snarere end formue (Piketty, 2020:1170-1172).

Piketty kritiserer disse grupper for at legitimere ulighed ved at fremhæve deres position som meritbaseret, selvom adgangen til eliten i høj grad bygger på sociale privilegier. Dette bidrager ifølge ham til øget polarisering og risiko for demokratisk ustabilitet. Eliten fungerer som både magtbærer og normskaber og har stor indflydelse på, hvordan samfundet organiseres – ofte til fordel for deres egne interesser frem for det fælles bedste (Piketty, 2014:641-643; 713-714;728-732).

I Capital in the Twenty-First Century (2013) argumenterer Piketty for, at elitetilhørsforhold primært formes gennem akkumulation af kapital og arv, frem for personlig indsats. Afkast på investeringer overstiger typisk lønvækst, hvilket giver dem med eksisterende formue større økonomisk gevinst end dem, der udelukkende lever af deres arbejde (Ibid:344;548-550;708-709). Denne mekanisme, understøttet af svag progressiv beskatning, tillader familier at fastholde deres positioner over generationer. Ulighed beskrives som politisk skabt og opretholdt gennem fx skattepolitik, ejendomsrettigheder og uddannelsesstruktur – ikke som et naturligt resultat af fair konkurrence (Ibid:650;687;728-732).

Piketty identificerer desuden to ideologiske tendenser i eliten: 'brahmin-left', som omfatter højtuddannede, ofte venstreorienterede akademikere og embedsmænd, og 'merchant-right', bestående af erhvervsfolk og højindkomstgrupper, der typisk støtter liberalistiske højrefløjspartier. Begge grupper er relativt distancerede fra arbejderklassen, hvilket reducerer fokus på omfordeling og social retfærdighed og kan bidrage til politisk og demokratisk ustabilitet.

Selvom eliten ikke udgør et fuldstændigt lukket system, fremhæver Piketty – ligesom Bourdieu – at der fortsat eksisterer betydelige barrierer for adgang, ofte knyttet til økonomisk kapital og uddannelse. Eliten reproduceres over generationer, og de institutionelle og politiske strukturer favoriserer fortsat dem, der er født med adgang til ressourcer (Piketty, 2014:713-714; Piketty, 2020:1215-1220;1238-1240).

Modsat Piketty, som er økonom, bidrager Sam Friedman og Daniel Laurison med et nyere sociologisk perspektiv på eliter. De definerer eliten som værende individer med toppositioner i samfundet (2019). Dette kunne være inden for eks. medier, kultur, jura, politik, finans, erhverv og akademia. Ligesom både Bourdieu og Piketty mener de, at eliten ikke alene er defineret på baggrund af indkomst. Jf. Friedman & Laurisons perspektiv har eliten en god uddannelsesmæssig baggrund, ofte fra eliteuniversiteter som i deres undersøgelseskontekst kunne være Oxford og Cambridge. Derudover fremhæver de, at individets sociale baggrund er central. Eliten kommer ofte af forældre med højstatusjobs (Friedman & Laurison, 2019:147-149;265-266). Kulturel kapital har ligeledes afgørende betydning. Eliten har de helt rigtige talemåder, den rigtige smag og adfærd og ikke mindst et stærkt netværk, som kan give dem adgang til de rigtige rum.

Friedman & Laurison (2019:210-218; 225-227) argumenterer for, at det er vanskeligt at opnå en eliteposition, da eliten primært rekrutteres fra middel- og overklassen. Selvom social mobilitet findes, er det sjældent, at personer fra arbejderklassen når helt til tops. De introducerer begrebet 'klasseloft' – en usynlig barriere, der minder om det kønsbaserede 'glasloft' (Loden, 1987; Bertrand, 2017) – som forhindrer personer fra lavere klasser i at avancere. Barriererne består bl.a. i uformel diskrimination, hvor individer uden de 'rigtige' manerer, talemåder eller sociale koder ikke anses som passende, trods faglige kvalifikationer. Desuden nævner de selvselektion som en vigtig faktor: mange fravælger visse karriereveje, fordi de føler sig fremmede over for elitens uskrevne spilleregler. Herved opretholdes elitens dominans, også i tilsyneladende åbne sektorer.

De tre teoretiske perspektiver på eliten viser alle, at eliten besidder betydelig magt og indflydelse – som normskabere, økonomisk eller uddannelsesmæssigt overlegen – og dermed spiller en aktiv rolle i opretholdelsen af ulighed. Vi finder det derfor relevant at undersøge, om disse perspektiver afspejler socialt reproducerende mekanismer i eliten.

2.2.2 Social reproduktion af eliten

Er du ikke født ind i eliten, kræver det jf. Bourdieu en kapitalakkumulation og en bestemt habitus for at blive en del af eliten for at blive anerkendt som en legitim deltager. Adgangen til eliten er mulig, men vanskelig, og Bourdieu mener i høj grad, at systemet reproducerer sig selv ved at kapital og habitus går i arv og kan fungere som usynlige adgangskrav til at blive en del af eliten (Bourdieu, 2002; Bourdieu, 2010:491-511).

Pikettys belysninger af social reproduktion er i tråd med Bourdieus. Piketty mener heller ikke, at det er fuldstændig forudbestemt, om man bliver en del af eliten, men dog at sandsynligheden er stærkt præget af individets sociale og økonomiske arv. Han mener derved, at det er muligt at bryde den sociale arv, men at det vanskeliggøres, hvis ikke individet hjælpes af støttende samfundsstrukturer, så som fri adgang til uddannelse, stærke omfordelende velfærdssystemer samt målrettet hjælp til udsatte grupper, eller progressive skatter, som kan mindske elitens strukturelle fordele (Piketty, 2020: 1215-1220).

Friedman & Laurison taler ligeledes ind i, at det teknisk set er muligt for alle at opnå adgang til eliten, men at adgangen i praksis er stærkt begrænset af sociale og kulturelle barrierer. De der lykkes at opnå adgang, er ofte individer, der formår at få den rigtige uddannelse og individer, der tilegner sig den rette kulturelle kapital og har evnen til at tilpasse sig elitemiljøernes sociale koder. Det er hovedsageligt øvre middelklasse og overklassen, som har lettest ved at få adgang. Individer fra arbejderklassen eller lavere middelklasse kan nå langt, men vil kun i sjældne tilfælde nå helt til toppen. Denne lukkethed, som eliten fastholder over for andre individer, består i kraft af, at det ikke kun gælder om at opnå merit og besidde evner, men at det sociale tilhørsforhold til eliten og den kulturelle

tilpasning er yderst central. Herved kan der være tale om, at eliten er præget af forudbestemthed (Friedman & Laurison, 2019:210-218;225-227).

Efter at have belyst teoretiske perspektiver på eliter, deres indflydelse på sociale strukturer og deres evne til social reproduktion, har vi fundet det interessant at afsøge empirisk og teoretisk funderet forskning om eliter i en dansk kontekst.

2.2.3 Eliten i Danmark

Der findes flere nyere perspektiver på, hvad der udgør det at være elite, og hvem der er en del af det danske samfunds elite. Sociologerne Ellersgaard, Grau Larsen og Steinitz har i de senere år produceret innovativ og førende eliteforskning, der ad tre omgange i 2015, 2019 og senest i 2025 har stået i spidsen for en kortlægning af den såkaldte 'danske magtelite' og dets netværk. Identifikationen af den danske magtelite er baseret på individers bestyrelsesposter, netværk og formelle magtpostioner. Magteliten består således af topdirektører, embedsmænd, politikere, ledere i fagforeninger og medieverdenen, som jf. Ellersgaard, Larsen og Steinitz' analyse er tæt forbundet via netværk i bestyrelser og organisationer. Dette centrerer derved magten i Danmark om en eksklusiv kreds af mennesker, som har indflydelse på politik, økonomi og samfundets generelle udvikling. De fremhæver, at det at få adgang til magteliten ikke alene handler om økonomiske forhold, men at netværk og anerkendelse i udvalgte kredse er helt centralt. De mennesker som jf. analysen er identificeret til at være en del af magteliten er også mennesker, der ofte besidder høj social, symbolsk og kulturel kapital, da dette netop kan være adgangsbilletten frem for den økonomiske kapital. Herved læner deres forståelse af eliten sig op af Bourdieus teori om kapitalformer (Ellersgaard, Steinitz & Larsen, 2025; Ellersgaard, Steinitz & Larsen, 2019; Ellersgaard, Bernsen & Larsen, 2015).

Resultaterne fra analysen fra 2025 er endnu ikke udgivet i sin fulde længde, men viser minimale ændringer fra udgivelsen i 2019, som er baseret på data fra 2017. Herunder fremhæves derfor interessante fund fra 2019-udgivelsen i sammenligning med udgivelsen fra 2015, som er baseret på data fra 2012.

Nogle interessante fund fra analysen fra 2019 består i, hvorvidt individerne i dette snævre netværk ligner hinanden. Ift. uddannelseslængde er magteliten særligt veluddannet. 67,2 pct. af magteliten havde i 2017 en lang videregående uddannelse eller Ph.d., hvorimod repræsentationen af denne uddannelsesgruppe kun består af 10 pct. blandt den generelle befolkning i Danmark. Ikke nok med at individerne ligner hinanden på uddannelseslængde, er det også tydeligt hvilke uddannelsesretninger, som særligt præger magteliten. Dette fremgår at være blevet endnu mere ensrettet fra 2012-2017. Andelen af individer med enten en uddannelsesbaggrund inden for Økonomi, Erhvervsøkonomi, Jura eller Statskundskab udgør 58,6 pct. af netværkets medlemmer i 2017, mens andelen i 2012 lød på 53,2 pct. (Ellersgaard & Larsen, 2019:90-94). Det er ikke kun de samme uddannelser, magteliten er formet af. De er også vokset op i de samme typer af miljø og hjem.

Overklassen udgør blot 1 pct. af den samlede befolkning, men det er op imod en tredjedel af magtelitens netværk, som opvokset i en familie fra overklassen. Ligeledes ser det ud med den øvre middelklasse, som udgør 5 pct. af befolkningen i perioden, hvor magteliten er født, mens op imod en fjerdedel af magteliten i 2017 kommer fra netop denne klasse. Dette understreger derved, at chancerne for at nå magteliten er væsentlig større, hvis individet er født ind i den øvre-middelklasse eller overklassen (Ibid:95-97). Undersøgelsen finder desuden, at selv på de basale karakteristika, såsom deres etnicitet, køn og alder, er magteliten i høj grad en homogen gruppe. Fælles for de få udenlandske medlemmer er, at de enten har boet i Danmark i mange år, taler flydende dansk eller kommer fra vestlige lande. Aldersmæssigt var 76 pct. af magteliten over 50 år i 2017, hvilket er en lille stigning fra 2012, hvor andelen lå på 74 pct. Ser vi på kønsbalancen i magtnetværket er andelen af kvinder svagt stigende fra 2012 på 19 pct. til 26 pct. i 2017. Dog er dette langt fra repræsentativ for befolkningssammensætningen, hvor ca. halvdelen er kvinder. Ligeledes eksisterer der en skæv fordeling ift. køn, når der zoomes ind på de 50 mest centrale individer i netværket, hvoraf kun 10 er kvinder (Ibid: 100-114). Derved er den danske magtelite skævt repræsenteret både ift. den etniske, kønsmæssige, og aldersmæssige sammensætning i befolkningen. Sidst men ikke mindst antyder undersøgelsen, at der eksisterer en geografisk segregering i Danmark. Det er særligt postnumre nord for København, som er udbredt blandt magtelitens medlemmer (Ibid:119-125). Én af hovedpointerne i bogen er således, at social mobilitet ind i eliten er vanskelig, da det er svært at træde ind i denne, hvis ikke du er født i den rigtige familie.

På trods af, at mobiliteten ind i eliten antydes at være vanskelig, ses der alligevel en relativ stor udskiftning af personer i magteliten. Kun 40 pct. af de personer, som er en del af magteliten i 2012, er fortsat en del af magteliten i 2017. Samtidig er 80 pct. af personerne i magteliten i 2017 ledere af organisationer, som ligeledes havde en repræsentant i magteliten i 2012. Der sker således en stor udskiftning af personer i kernen af magten, men de personer, som er en del af magtnetværket, kommer af de samme type af organisationer (Ibid:8).

Med disse analyser om danske magtstrukturer, argumenteres der for, at Danmark til trods for sit alment anerkendte ry om at være et lighedsorienteret land, til stadighed har klare hierarkier. Det fremtrædende af Ellersgaard m.fl.'s netværksanalyse er, hvordan eliten er forbundet igennem deres interne relationer. Disse analyser viser herved, at magteliten ikke nødvendigvis består af de individer, som er synlige i offentligheden, men også består af individer, som arbejder gennem uformelle netværk og beslutningsrum. Tydeligt er det i analysen, at det er mange af de samme individer, der cirkulerer i eliten, hvilket kan argumenteres for at skabe en selvforstærkende magtfordeling og understrege, at der eksisterer en social reproduktion af eliten (Ibid).

Et andet og nyere perspektiv på eliten i Danmark står sociolog Johannes Andersen bag. Han taler ligeledes ind i centraliseringen af en ny elite omkring København og de andre større studiebyer i Danmark i sin nyligt udgivne bog "Fra en anden verden" (2024). Han mener, at der i takt med

samfundets øgede kompleksitet og voksende interesse for viden er blevet skabt en ny elite, der har adgang til magt og indflydelse, som ikke længere udelukkende består af de individer med den største økonomiske indtjening. I modsætning til det, han kalder den traditionelle elite, som primært er defineret ved økonomiske ressourcer i form af aktiver, ejerandele i virksomheder og høj indkomst mm., udgøres den nye elite af personer med lang videregående uddannelse, kommunikative og analytiske kompetencer, og strategiske positioner i samfundets videns- og beslutningsbærende organisationer (Ibid). Den nye elite er overvejende yngre mænd i storbyer, ofte ansat i private virksomheder, konsulenthuse eller det offentlige, og deres magt udspringer af en privilegeret adgang til viden og deres mulighed for at sætte dagsordener gennem medier, analyser og strategiske beslutningsprocesser. Elitestatus er her mindre forbundet til arv eller økonomisk rigdom og mere til en videnmæssig overhøjde og qua individets akademiske jobposition større mulighed for indirekte at påvirke politiske og samfundsmæssige prioriteringer i sammenligning med den resterende befolkning. Ifølge Andersen befinder den nye elite sig hovedsageligt som mellemled mellem befolkningen og den traditionelle elite eller de formelt magthavende qua, at de ofte arbejder som rådgivere, analytikere, strateg, med kommunikation eller lignende og derigennem er med til at præge det vidensgrundlag, som eliten eller de magthavende tager beslutninger på baggrund af, fordi de er dem, som samler informationen, undersøger det, skaber det eller kommunikerer det. Deres indflydelse manifesterer sig således som en vedvarende produktion af viden og forslag, som former samfundets retning uden, at de selv nødvendigvis har formel magt. Den indflydelse, den nye elite har, kan ofte fremstå som diffus og i en ikke direkte styret form, fordi de i høj grad ikke har en overordnet plan med deres muligheder for indflydelse. Den nye elite kan med andre ord forstås som at være alle dem, der "sidder bag tæppet" og får forestillingen til at forløbe gnidningsfrit, men disse har også haft deres indflydelse på det endelige show. I modsætning hertil frembringer han den traditionelle elites typiske karakteristika. Disse er ofte økonomisk overlegenhed, hovedsageligt midaldrende mænd bosat i provinsen med høj indkomst og strategiske investeringer som primært magtredskab. Denne gruppe har haft afgørende indflydelse på den samfundsøkonomiske udvikling, men deler nu jf. Andersen i stigende grad magtrum med den nye elite. Således frembringer Andersen et nyere perspektiv på, hvordan magt og privilegier i stigende grad er knyttet til viden, netværk og diskursiv autoritet frem for udelukkende økonomisk kapital, og at eliteforståelsen derfor må udvides og undersøges yderligere (ibid).

Blandt de mere rent empiriske undersøgelser af elitens reproduktion viser et dansk studie om formue-ulighed blandt børn og betydningen for deres formue i voksenlivet fra 2018, at børns formue næsten udelukkende består af økonomiske overførsler eller gaver fra forældre eller andre familiemedlemmer. Personer med bedre økonomiske vilkår i barndommen oplever i voksenlivet fortsat økonomisk støtte fra familiemedlemmer og evner desuden også selv at opspare til egen formue i voksenlivet. Studiets resultater tyder dermed på, at formue og økonomisk adfærd går i arv og derigennem bidrager til fastholde formueulighed over generationer selv i et land som Danmark, der anses for at have en høj indkomstmobilitet (Boserup, Kopczuk & Kreiner, 2018:514-516).

I tråd hermed viser et studie baseret på dansk registerdata, at eliten i Danmark i høj grad reproduceres på tværs af generationer (Wagner et al., 2020). Studiet undersøger udviklingen i formue-lighed mellem forældrene til partnere i parforhold fra 1987 til 2013 og finder, at personer fra velhavende familier ofte danner par med andre fra lignende baggrunde. Tendensen er særligt udtalt blandt de 10 % rigeste – og endnu stærkere blandt den rigeste 1 %. Derudover peger resultaterne på, at formue-homogami er steget fra 1990'erne til 2000'erne (ibid).

Ovenpå teoretiske og empiriske bidrag (som fremgår i afsnit 2.1 og 2.2) fremgår det, at eliten kan defineres på forskellig vis, men generelt antydes det, at der eksisterer en grad af reproduktion af eliten. Dette vækker vores interesse til at afsøge mulighederne ved ML og prædiktive metoder i samfundsvidenskabelig forskning for at klarlægge, om disse relativt nye metoder meningsfuldt kan anvendes til udbygningen af social forskning. Yderligere om disse metoder kan benyttes til at prædiktere sociale livsudfald, såsom at opnå elitepositioner.

2.3 Machine learning og samfundsvidenskabelig forskning

Machine learning (ML) er videnskaben, hvor computere lærer at træffe beslutninger på baggrund af data uden at være eksplicit programmeret på forhånd. En udbredt definition af ML lyder, at "et *computerprogram siges at lære af erfaring E i forhold til en række af opgaver T og en præstationsmåling P, hvis dets præstation ved opgaver T, målt med P, forbedres med erfaring E"* (oversat fra engelsk) (Pajankar & Joshi, 2022:65-66). ML beskrives således som en computationel evne til at lære og forbedre præstationer på baggrund af erfaring. Derved forstås ML som et værktøj, der kan bidrage i flere sammenhænge såsom mønstergenkendelse, automatiseret klassificering, forudsigelser, tekstanalyse, billede- og lydgenkendelse, anbefalingssystemer mm. (ibid). Et typisk eksempel er et spamfilter, der trænes på et arkiv af e-mails for at kunne klassificere nye mails som spam eller ej. Dets præcision måles ved korrekt klassifikation, og modellen justeres løbende for at forbedre resultaterne. (Ibid).

Eksemplet med spamfiltrering illustrerer, hvordan ML kan anvendes i praksis til at løse en klassifikationsopgave ved at lære mønstre i tilgængelige data. Bag sådanne tilsyneladende enkle løsninger ligger et omfattede og hastigt voksende forskningsfelt, som særligt har udviklet sig i de seneste år. Dette skyldes bl.a. kraftige stigninger i både mængder og kvaliteten af tilgængelige data, som er nogle af grundstenene for succesfulde ML-baserede resultater (ibid). Forskningsfeltet inden for ML beskæftiger sig bl.a. med udviklingen af nye algoritmer, forbedring af modellers præcision og robusthed, forståelse af indlæringsdynamikker samt ansvarlig anvendelse af modeller i ofte komplekse sammenhænge. Potentialet ved brugen af ML synes enormt og teknologien har hastigt vundet indpas i en række forskningsfelter. Også indenfor samfundsvidenskabelig forskning muliggør ML nye måder at analysere og identificere mønstre i store og komplekse datamængder, hvilket ofte er centralt i studier af menneskelig adfærd, sociale systemer og institutioner.

Noget af det, som ML kan bidrage med til samfundsvidenskaben, er jf. Lundberg et.al at denne teknologi naturligt fordrer en mere induktiv tilgang til studier i kraft af evnen til at lære og identificere interessante mønstre i data, som forskere muligvis ville have overset (2022). Modsat klassiske statistiske metoder, som ofte er hypoteseorienterede og hviler på en solid teoretisk forståelse af årsagssammenhænge, hvor specifikke sammenhænge mellem variable kan testes, og der kan kontrolleres for kendte confounders (Molina & Garip, 2019:29). Ulempen herved mener Lundberg et.al at være, at forskernes uundgåelige forudindtagethed bl.a. kan lede til at bekræfte eksisterende viden ved at underbygge resultater, som de på forhånd forventede at se (2022). Brugen af ML kan skabe forudsætninger for at finde mønstre, som ikke var mulige for forskeren grundet den store kapacitet til data, som kan inkluderes i disse typer af studier. Det fremstår derved som en betydelig fordel, at metoder, som anvender ML, ofte har kapacitet til større mængder data og evner til at lære komplekse interaktioner og non-lineariteter direkte fra den data, som den får at arbejde med. Dette gør, at studierne ikke er begrænset af en bestemt regressionsmodel ift. at finde de mønstre, der eksisterer i data, hvilket skaber større fleksibilitet modsat de klassiske statistiske metoder, hvor der typisk udvælges en mindre række af inputs (ibid; Molina & Garip, 2019:30-31 & 37-38).

Modsat kan metoder, hvor ML indgår, kritiseres for at udfordre grundlæggende principper i samfundsvidenskabelig metode, når det gælder forklarbarhed og kausalitet. ML-baserede studier søger typisk at forudsige verden, altså at forudsige, hvad der sker når nye inputs kommer, men ikke altid hvorfor. De bagvedliggende mekanismer i ML modeller er ikke nødvendigvis let fortolkelige, hvoraf begrebet 'Black box' kommer. Mens klassisk statistik ofte søger at forklare verden ved at forstå, hvordan en bestemt faktor påvirker et udfald (Molina & Garip, 2019:29).

I forlængelse heraf mener Lundberg et.al, at den udbredte tillid, som eksisterer til klassiske koefficienter, ikke ukritisk bør fastholdes, da koefficienter ikke nødvendigvis afspejler kausale sammenhænge. Især ikke uden teoretiske antagelser, som bakker dem op. Det samme mener de gælder metoder, der anvender ML. Disse bør også underbygges af teori for at styrke deres validitet. Modsat de udfordringer, der kan antydes at være ved brug af ML-metoder, anser Lundberg et.al ML som et nyttigt værktøj inden for kausal inferens, fordi disse metoder nemmere kan indfange interaktioner og komplekse sammenhænge som klassiske modeller ofte forenkler væk. Lundberg et.al mener, at kombinationen af ML-metoder og kausale antagelser kan bestyrke, at studiers fund reelt er kausale sammenhænge (Lundberg et.al, 2022).

Lundberg et.al frembringer således en række overordnede muligheder, som ML kan bidrage med til samfundsvidenskaben. Dog understreges det, at ML ikke kan erstatte menneskelig indsigt, men at det kan være et værktøj til at rejse spørgsmål, skabe antagelser og arbejde med data. Der argumenteres således for, at brugen af ML udvider mulighederne for samfundsvidenskabelig forskning, fordi det gør det muligt at kombinere menneskelig indsigt med avancerede algoritmer (ibid).

2.3.1 ML og prædiktive metoder

Ved prædiktive metoder er det centrale at forudsige et udfald af variablen Y baseret på en række inputvariable X. Formålet med denne type metode er ikke nødvendigvis at forklare en specifik sammenhæng, men nærmere at opnå bedst mulig præcision af forudsigelser af nye eller ukendte observationer. Der fokuseres således på den prædiktive models samlede præcision af forudsigelser om udfaldet af en række observationer, som er 'out-of-sample' data, altså data som modellen ikke er trænet på. I prædiktion anvendes ML i træningsfasen til at dygtiggøre modellens identifikation af mønstre i data, som kan være behjælpelige for af afgøre om en pågældende observation skal prædikteres til at have et bestemt udfald.

Prædiktive metoder er et relativt nyt analyseværktøj og derfor endnu ikke udbredt i samfundsvidenskaben. Jf. Verhagen er disse metoder decideret underudnyttet i samfundsvidenskabelig forskning, hvilket han mener, i høj grad skyldes misforståelser om, hvad prædiktion indebærer og frygten for uigennemskuelige arbejdsprocesser (2022). Han argumenterer for, at prædiktioner ikke kun kan bruges som prognoser eller optimeringsmål, men at de ligeledes kan bruges til at evaluere modeller på baggrund af deres evne til at prædiktere udfald præcist. Derudover argumenterer han for at lade prædiktionsmetoder i kombination med mere traditionelle modelleringsmetoder supplere hinanden (Ibid.), ligesom Lundberg et.al (2022).

Verhagen understreger, at der er en række forbedringsmuligheder til samfundsvidenskaben ved at inkludere prædiktion i forskningsprocessen, herunder ift. modeltilpasning, sammenligning af forskellige modeller og til at forstå kompleks modeladfærd ved at gøre brug af interventionsbaseret ræsonering. Dette vil sige, at forskere kan undersøge, hvordan ændringer i en model påvirker udfaldet, hvilket forbedrer muligheden for at forstå de komplekse interaktioner og dynamikker, der forekommer inden for modellen. Desuden argumenterer han for, at prædiktionsmetoder gør det muligt at sammenligne modeller på tværs af paradigmer og vurdere deres evne til at lære dataens kompleksitet. Han påpeger i forlængelse heraf, at det er muligt at øge gennemsigtigheden i forskningen, hvis der gives en stringent vurdering af forskerens valgte models evne til at lære dataene, når denne evalueres i sammenligning med andre modeller.

Med ovenstående in mente fremstår der et stort potentiale for afprøvningen af prædiktive metoder og brugen af ML til at udvide og nuancere studier i samfundsvidenskaben. I følgende afsnit vil vi afdække eksisterende forskning, der succesfuldt anvender disse metoder.

2.3.2 Prædiktionsbaseret forskning

Omfanget af forskningsartikler, der anvender prædiktive metoder og ML til udvidelsen af samfundsvidenskabelig forskning, har været begrænset igennem vores litteratursøgningsproces. Dette tydeliggør, at forskningsfeltet stadig er i en tidlig fase ift. integreringen af disse metoder og analyseteknikker. I det følgende afdækker vi både studier indenfor samfundsvidenskabelig forskning og

enkelte i krydsningsfeltet mellem andre grene, som vi har ladet os inspirere af ift. at afprøve brugbarheden af disse metoder i et sociologisk studie omhandlende livsudfald og elitepositioner.

Et studie fra 2023 afsøger hvilke spørgsmål, besvaret af forældre i en survey, der bedst kan forudsige vedvarende sprogvanskeligheder for børn i 11-årsalderen (Gasparini, et.al.). Formålet er at identificere specifikke indikatorer i alderen 8, 12, 24 og 36 måneder, som kan forudsige sproglige vanskeligheder hos børnene senere i deres barndom. Dataen er baseret på 839 børn fra Early Language in Victoria Study (ELVS), hvor forældrene hertil har besvaret i alt 1990 spørgsmål, som konkret omhandler børnenes ordforråd, grammatik, symbolsk leg, forældre-barn-interaktion m.m. Den sproglige vurdering er baseret på CELF-4 og vurderet ved 11 år. Til studiet blev der gjort brug af random forests og SuperLearner modeller til at identificere de bedste forudsigelser. Studiet identificerer et mindre sæt af forældrebesvarede spørgsmål i 2-3 årsalderen, som med god nøjagtighed kan prædiktere sproglige vanskeligheder ved 11-årsalderen. Konkret lyder nøjagtigheden af prædiktionen på 73%. Resultaterne heraf kan således bruges til tidlig identifikation og målrettet støtte til de børn, som har brug for det (ibid). Her bruges prædiktionsmetoder ikke direkte i en sociologisk kontekst, men vi finder tilgangen om at benytte data fra tidlige perioder i livet til at prædiktere elementer senere i individets liv interessant. Dette leder os til at afsøge, om andre studier har forsøgt at prædiktere sociale livsudfald.

Et studie fra 2020 har netop afsøgt mulighederne for prædiktioner af individers livsbaner. Studiet er skabt af et storskalasamarbejde og 160 forskningsgrupper har afprøvet forskellige prædiktive modellers evner til at forudsige 6 forskellige livsudfald for individer ved 15 år. Prædiktionerne er baseret på data omhandlende 4.242 familier, der har fået et barn omkring år 2000, og som har deltaget i et longitudinelt kohorte studie af "Fragile families and child wellbeing" (Salganik, et.al.). Dataindsamlingen er foregået ved barnets fødsel, 1, 3, 5, 9 og 15 år og indebærer både survey og interviews med forældrene og barnets eventuelle anden primære omsorgsperson. Det konkrete emnemæssige indhold heraf omhandler bl.a. barnets helbred, relation til forældre, miljømæssige forhold, eventuelle offentlige indsatsplaner, forældres civile status, indkomst, uddannelse og relation til arbejdsmarkedet. Til studiet er der blevet gjort brug af simple modeller såsom random forrest, LASSO regression (Least Absolute Shrinkage and Selection Operator), OLS regression (Ordinary Least Squares) og andre. Fælles for alle forskningsgrupperne er, at de 6 livsudfald ikke meningsfuldt kunne prædikteres (Ibid). Hvilket på den ene side fremhæver kompleksiteten af socialvidenskab og at menneskers adfærd kan siges at være mere uforudsigelig i sammenligning med andre forskningsfelter såsom naturvidenskaben. På den anden side er studiet udarbejdet i 2020 og modelleringen af prædiktive modeller har sidenhen været under stor og kontinuerlig udvikling. Vi antager derved, at en nutidig og mere kompleks prædiktivmodel, som kan forstå tidsdimensioner og som kan identificere repræsentationer for social adfærd og relatere disse til hinanden, bør kunne resultere i en mere succesfuld prædiktion. Derudover kan deres datagrundlag anfægtes for at være smalt og usystematisk i sammenligning med f.eks. registerdata.

I modsætning til Salganik et.al. finder et banebrydende studie fra 2023 det muligt at prædiktere livsudfald på et individuelt plan (Savcisens et.al, 2023a). Studiet anvender detaljeret dansk registerdata om 6 millioner danskere over en tiårig periode. Dermed er det mindre oplagt at anfægte dette datagrundlag for ikke at være egnet til prædiktion af menneskelig adfærd, da succesfulde prædiktioner netop ofte forudsætter store datamængder og kraftfulde ML-algoritmer. Studiet gør brug af en transformerbaseret prædiktionsmodel, der analyserer tekst med belægget om at denne type modelarkitektur tidligere har vist at være succesfuld til at identificere komplekse mønstre i ustrukturerede sekvenser af ord, samt at den er generaliserbar til andre sekvenser, som deler egenskaber med sprog, eks. sekvenser hvor rækkefølgen er væsentlig eller hvor elementer i sekvensen kan have betydning på flere niveauer. På den måde anvender studiet en kompleks model, som kan forstå tidsdimensioner og relationen mellem datapunkter på et enormt og meget detaljeret datasæt. Dette skaber således bæredygtige forudsætninger for at forudsige livsudfald, da menneskelivet kan forstås som en kompleks størrelse. Det kræver både enorme datamængder og en model, der forstår en høj grad af kompleksitet i data til at øge chancerne for at få meningsfulde prædiktive resultater.

Konkret afsøger studiet hvorvidt det er muligt at prædiktere dødsfald inden for fire år og får en prædiktionspræcision på 78.8%, hvilket er et fremragende resultat, som ligger højere end de fleste tidligere resultater i andre studier med simplere ML-modeller og prædiktioner af livsudfald (Savcisens, et.al. 2023a), og derfor er der god grund til at se nærmere på at udvikle og tilpasse denne type metode inden for det forskningsmæssige krydsfelt mellem sociologi og socal data science.

2.4 Opsummering

På trods af at eksisterende teori og sociologisk forskning har dokumenteret en række faktorer, der påvirker individers muligheder for social mobilitet og adgangen til eliten, er der jf. vores afsøgning af feltet endnu ikke succesfuldt anvendt prædiktive modeller til at undersøge betydningen af forhold i barndommen for sociale udfald i voksenlivet (Salganik, et.al.). Udfordringerne med succesfulde prædiktioner af sociale udfald kan bl.a. skyldes mangler i datagrundlagets størrelse, mangler i detaljegrad eller det præcis modsatte, at datagrundlaget er for komplekst. Derudover kan det skyldes, at anvendte prædiktive modeller er for simple og ikke kan rumme den kompleksitet, det kræves at identificere mønstre i menneskelig adfærd (ibid).

Anvendelsen af avancerede prædiktionsmodeller, som transformerbaserede modeller (som anvendes i (Savcisens, et.al. 2023a)), er endnu ikke dokumenteret anvendt i sociologisk forskning af social stratifikation jf. vores litteratursøgning. Dette udgør herved et væsentligt videnshul, som vi i dette speciale søger at udfylde ved at kombinere deep learning models med omfattende longitudinelle data, hvilket kun er muligt i kraft af vores adgang til Danmarks Statistiks registre. Brugen af avancerede prædiktive metoder kan bidrage med en mindre forudindtaget undersøgelse af, hvilke faktorer i barndommen der kan have indflydelse på individets muligheder for at nå toppen af det sociale hierarki.

Afsluttende diskuterer vi vores fund med henblik på at indplacere os i forskningsfeltet omhandlende eliter samt ML i samfundsvidenskaben.

På den måde ønsker vi i dette speciale med følgende problemformulering at afsøge nye muligheder for at identificere tidlige indikatorer på social differentiering og for at forstå de mekanismer, der former elitestrukturer over tid. Vi tilbyder herved nye indsigter i de mekanismer, som afgør om individet bliver en del af en elite i voksenlivet, samt metodisk innovation i sociologisk forskning.

Relevansen af at udbygge forskningsfeltet omhandlende eliter ligger i, at der eksisterer en opfattelse blandt den brede befolkning af, at samfundet er relativt egalitært og i høj grad præges af meritokratiske idealer om adgang til magt, indflydelse, prestige, økonomiske privilegier og 'du kan blive hvad du vil'. Mens ovenstående problemfelt både teoretisk og empirisk fremmaler underliggende sociale strukturer, som begrænser eller helt umuliggør, at individer fra bestemte sociale baggrunde kan opnå elitære positioner (Ellersgaard, 2024).

Vi finder det relevant at begrænse vores forudindtagethed, som ville kunne lede os til at bekræfte fund, vi forventede at se. Derfor vil vi sætte punktum ved problemfeltets fund, for i stedet at rette fokus mod et induktivt studie, hvor vi ønsker at undersøge, hvorvidt individets tilhør til visse elitegrupper i voksenlivet kan prædikteres af forudbestemte sociale faktorer i barndommen som en del af et livsforløb, hvor vi tilsikrer induktion ved at være åbne for forskellige definitioner af elite-udfald og for flest mulige sociale baggrundsdata afhængigt af tilgængelig datamateriale.

3 Problemformulering

På baggrund af ovenstående problemfelt, vælger vi at arbejde videre med følgende problemformulering:

I hvilket omfang kan vi vha. machine learning prædiktere eliteudfald i voksenlivet på baggrund af forudbestemte sociale faktorer og hændelser i barndommen som en del af et livsforløb?

BEGREBSAFKLARING:

Det at være en del af en elite i voksenlivet kan forstås på flere måder. Vi afsøger i dette speciale flere definitioner af elite og refererer derfor til den konkrete operationalisering af udvalgte 'Eliteudfald' (se afsnit 6.3.1). Med 'forudbestemte sociale faktorer' menes sociale karakteristika for et individ, som individet ikke selv har indflydelse på. Dvs. karakteristika eller faktorer, der enten er givet fra fødslen i form af køn, forældre, søskende, bopæl og nabolag, eller som en del af de beslutninger, ens forældre

træffer for en. Med andre ord kan disse forudbestemte sociale faktorer kaldes for ens sociale baggrund (se operationalisering afsnit 6.4.1). Dernæst refererer 'hændelser i barndommen' til målbare familiære oplevelser, som eks. at få en yngre søskende eller opleve, at ens forældre skilles (se operationalisering afsnit 6.4.1).

3.1 Forskningsspørgsmål

Med ovenstående problemformulering ønsker vi at lægge op til en empirisk-induktiv tilgang med et abduktivt tilsnit til undersøgelsens forløb, hvilket underlæggende fremgår af følgende forskningsspørgsmål. Gennem problemformuleringen tilstræber vi at holde os åbne for nye indsigter og opdage uventede mønstre i data – dog uden at miste blikket for indplaceringen i det metodiske og emnemæssige forskningsfelt, altså det abduktive tilsnit. Nedenfor er specialets forskningsspørgsmål oplistet kronologisk og tematisk.

Elite(r):

1. Hvordan kan elite-begrebet operationaliseres som livsudfald?

Prædiktion pba. forudbestemte sociale faktorer:

2. I hvor høj grad kan forudbestemte sociale faktorer og hændelser i barndommen som del af et livsforløb prædiktere visse eliteudfald?
3. Kan nogle elite-udfald prædikteres bedre end andre? Og hvorfor?

Metodiske spørgsmål:

4. I hvilket omfang kan prædiktive ML-metoder bidrage til sociologiske undersøgelser via deep learning og evne til at se komplekse mønstre?

4 Etik

Databehandling og -analyse er udført hos Danmarks Statistik under databeskyttelsesloven og efter gældende GDPR-retningslinier. Danmarks Statistik kan give adgang til registerdata til bl.a. forskningsrelaterede og statistiske formål under strenge datasikkerheds- og fortrolighedsbestemmelser, hvor det sikres, at data ikke bruges til andet end formålet. Det er i denne sammenhæng strengt forbudt at udtrække data eller analyser på individniveau, og kun aggregerede værdier og statistikker, inklusive resultater fra modelprædiktioner, må bruges til forskningsformål (DST, 2025f; DST, 2025g).

I de følgende metode- og analyseafsnit benyttes visualiseringer, figurer og eksempler, da disse udgør et vigtigt understøttende element i forståelsen af dataopbygning og -håndtering. Alle visualiseringer, figurer og eksempler, som henviser til enkeltpersoner eller viser datapunkter på individniveau, er genereret med udgangspunkt i syntetisk skabt data designet til at *ligne* vores rigtige

data mest muligt. Al syntetisk data er skabt uden for DST's lukkede miljø og har således ingen forbindelse til rigtige data om rigtige personer eller til DST-data. Alle visualiseringer, figurer og eksempler, som er dannet på baggrund af syntetisk data, vil være markeret med en stjerne *. Kun aggregerede værdier og metrikker udtrækkes fra reelle data fra DST's miljø og benyttes i metode og analyse (DST, 2025g). Det er således kun resultater som jf. DST's retningslinjer er godkendte, der kan hjemsendes og videreformidles i dette speciale. På den måde sikrer vi at intet ift. sampling, opbevaring af data og videreformidling af resultaterne kan henvise tilbage til enkeltpersoner (Grenawalt, T., 2023; DST, 2025g).

Brugen af machine learning, prædiktionsmetoder og avancerede algoritmer på 'big data' – et udtryk, der anvendes om enorme datamængder i computationelt arbejde – og registerdata i samfundsvidenskab er et felt, som i de senere år har været under stor og fortsat udvikling, i takt med at machine learning generelt udvikles med stor hastighed. Det har som forskningsfelt stort – vel nærmest revolutionerende – potentiale, ligesom mulighederne for den praktiske anvendelighed er til at få øje på; på samme måde, som computationel health science kan tilbyde banebrydende nye metoder til diagnosestillelse, kan man forestille sig, at computationel social science kan bruges i den offentlige og sociale sektor til at identificere personer i diverse risikogrupper med henblik på langt tidligere forebyggende indsatser mod eksempelvis kriminalitet, arbejdsløshed, førtidspension og mistrivsel, for blot at nævne nogle få eksempler. Det store potentiale stiller dog mindst lige så store krav til de etiske implikationer ved brug og integration af disse metoder i både forskning og blandt praktikere (Buchholz & Grote, 2023:60-61).

Det bør understreges, at vores prædiktionsmodel har til formål at undersøge og teste anvendelighed og implementering af en sådan type model i en sociologisk videnskabelig og bredere samfundsvidenskabelig kontekst. Den er ikke lavet med det formål at kunne anvendes direkte til at guide beslutningstagning i en praktisk ikke-videnskabelig sammenhæng, og før en sådan anvendelse vil kunne finde sted, vil en betydelig kvalitetssikring af modellens fairness skulle finde sted. Modellen er desuden dannet på baggrund af en udelukkende dansk bosiddende population, hvorfor den ikke uden betydelig tilpasning vil kunne anvendes til prædiktion af sociale udfald uden for Danmarks grænser. Vi opfordrer derfor til stor forsigtighed ved forsøg på at generalisere modellen til andre landes populationer, eller sågar til danske populationer sammensat på anden vis end den population, vi har anvendt til træning af modellerne.

5 Design

Undersøgelsesdesignet for dette studie følger ikke en traditionel struktur, som kan identificeres med én specifik designtype. Studiets struktur er derimod præget af både den tidslige dimension og det metodiske valg. Tidsdimensionen gør, at strukturen er præget af et longitudinelt design i retrospekt, da

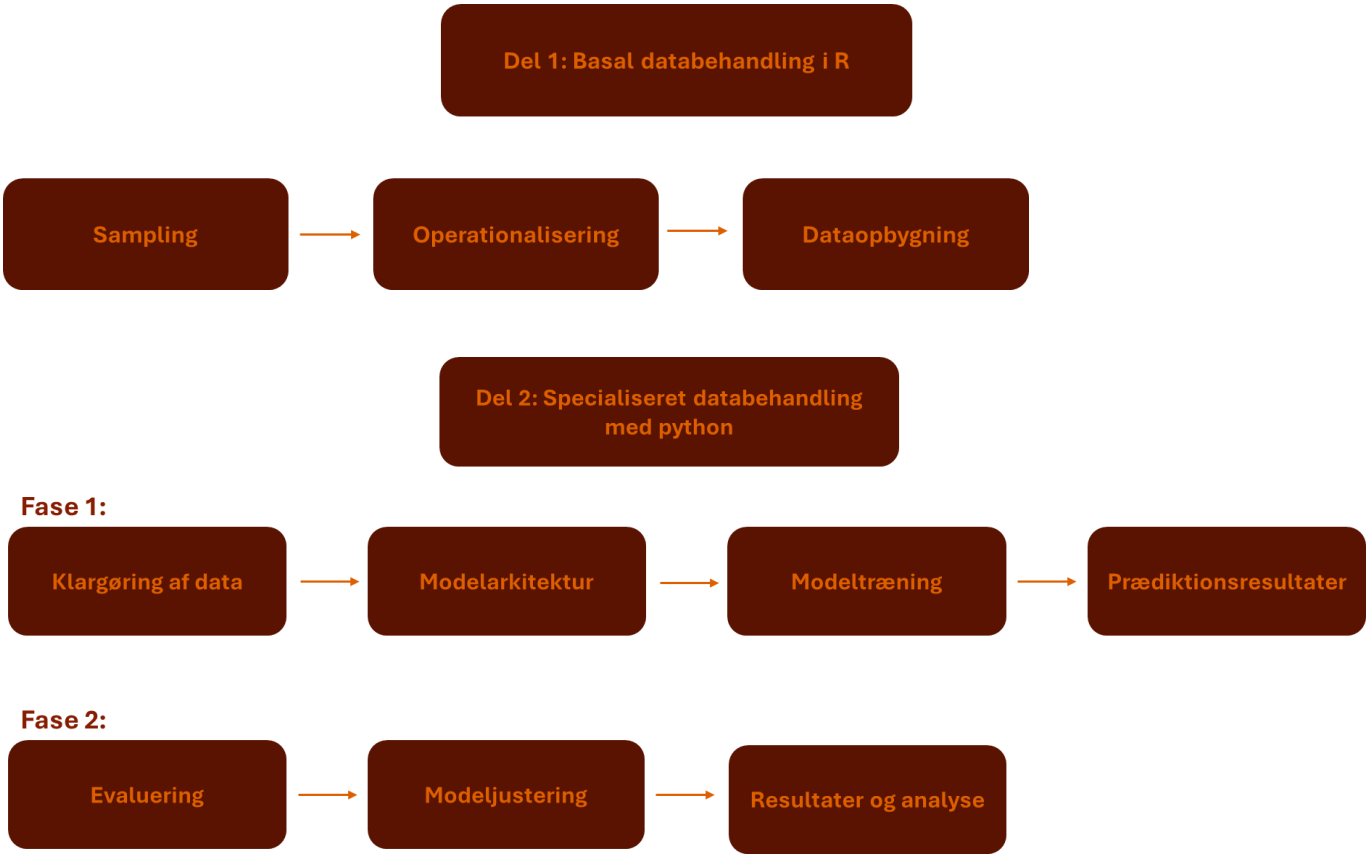
vi undersøger tre fødselskohorter over tid. Disse kohorter deler egenskab i form af, at de er født i Danmark i enten 1980, 1981 eller 1982. Data trækkes prospektivt hen over kohorternes barndom fra 0-17 år med det formål at afprøve, om denne information kan bidrage til en succesfuld prædiktion af, om de opnår en elitær social position som voksne. (De Vaus, 2001:123-124).

Derudover præger metodevalget ligeledes studiets design, som momentvis får et eksperimenterende og induktivt tilsnit, da metoden er ny og uprøvet i samfundsvidenskabelig kontekst, og kræver en todelt proces, hvor en variant af metoden først prøves af, hvorefter den evalueres og justeres efter behov, og afprøves igen. Det metodiske valg medfører desuden en betydelig databehandlingsproces, som overordnet foregår i to dele.

Første del udføres i databehandlingsprogrammet R og involverer den basale databehandling herunder sampling af populationen, al præprocessering af rå-data, operationalisering af variable og endelig sammensætning af data for den udvalgte population som tidsserie. Denne del af databehandlingsprocessen gennemgås i afsnit 6 (Data og Operationalisering), og er grundlæggende uafhængig af det efterfølgende metodevalg. Anden del er derimod stærkt relateret til valget af en transformerbaseret model som prædiktionsmetode og udføres i programmeringssproget Python. Denne del involverer al specialiseret databehandling, som er nødvendig for den valgte metodiske udførelse, dvs. videre tilpasning af data, modelarkitektur, træning og prædiktionsanalyse, dvs. alt som samlet set kan kategoriseres som 'social data science'. Denne del gennemgås i detaljer i afsnit 7 (Metode).

Studiets designmæssige struktur følger således en kombination af den tidslige dimension, repræsenteret ved at følge kohorterne over tid, samt den metodiske tilgang, hvor prædiktion af binære udfald afsøges vha. ML. Med henblik på at skabe overblik over designet, har vi i Figur 1 visualiseret de centrale dele.

Figur 1: Designstruktur



6 Data og operationalisering

I dette afsnit gennemgås centrale karakteristika ved det valgte datagrundlag, og de heraf afledte arbejdsprocesser, som sampling, valg og operationalisering af variable, datatilgængelighed og indledende databehandling.

6.1 Data

Dansk registerdata er blandt de mest omfattende, der findes. De dækker en hel nation med for nuværende omkring seks millioner individer over flere årtier. Registerdata gør det muligt at kombinere individdata for hele befolkningen på tværs af registre over længere tidsperioder. Dette giver en enestående detaljegrad og præcision i både data, metode og analyse, som ikke ville være opnåelig med andre datatyper. Store datamængder med høj detaljegrad skaber mulighed for, at modellen kan forudsige et udfald på baggrund af dens evne til at identificere komplekse sociale mønstre i data, og er i denne undersøgelse en forudsætning for prædiktionsmodellens præstation og præcision. Data fra DST indeholder oplysninger om livsbegivenheder relateret til familie og barndom, uddannelse,

beskæftigelse, indkomst og adresse, registreret på årlig basis (DST; Registre og Referencetyper, samt Dokumentation af Data).

6.1.1 Registeroversigt

Studiet anvender en bred palette af tilgængelige registerdata i grunddataformat i kombination med eksterne data, som samlet set giver mulighed for at danne adskillige anvendelige variable til netop dette studie. Vi inddrager flest mulige relevante registre og rå-variable til at danne vores endelige udfalds-, baggrunds- og områdevariable med formålet at bevare en forholdsvis åben tilgang til, hvilke faktorer, der kan have betydning for et givent udfald, og desuden åbenhed for forskellige måder at definere elite-udfald på. I Tabel 1 ses en oversigt over de anvendte registre og rå-variable.

Tabel 1: Registeroversigt

Register	Beskrivelse	År	Rå-variable
BEF	Befolkningen (personer med fast bopæl i Danmark)	1980-1997 & 2020-2022	PNR, MOR_ID, FAR_ID, FOED_DAG, CIVST, IE_TYPE, OPR_LAND, KOEN
IND	Indkomst	1980-1997 & 2020-2022	PNR, PERINDKIALT_13
UDDA	Uddannelse	1980-1997 & 2020-2022	PNR, HFAUDD, HFINSTNR
AKM	Arbejdsmarkedsklassifikations-modul	1980-1990	PNR, Socioøkonomisk klassifikation fra 1980 til 1990 (SOCIO_GL)
		1991-1997 & 2020-2022	PNR, Socioøkonomisk klassifikation version 2013 (SOCIO_13)
IDAP	Arbejdsmarked Persondata	1980-2007	PNR, ARLEDGR
IDAN	Ansættelser	2008-2022	PNR, BRUTTO_GRADAAR
RAS	Registerbaseret arbejdsstyrkestatistik	2008-2022	PNR, LOENTIMER, PRIMAER_STATUS_KODE, SOC_STATUS_KODE
DODSAASG	Dødsårsagsregister	2002-2019	PNR, D_DODSDATO, D_FINDEDATO
Eksterne datafiler			
AUDD_PRIA_L1LX_K	Hjælpefil, som benyttes til omdannelse af variablen HFAUDD fra	-	HFAUDD, AUDD_PRIA_L1LX_K
MERGEDAREA	8043 mikroområder i Danmark (Lund 2018)	1985-1997 & 2020-2022	PNR, area, COUNT

Den eksterne datafil 'mergedarea' indeholder en mikroområdeinddelingen, som er skabt af Lund (2018) på baggrund af en induktiv, rekursiv algoritme, som isolerer mindre geografiske klynger ud fra fysiske

barrierer og naturligt opdelte sociale rum. Mikroområdedataen benyttes i denne undersøgelse til at nuancere de klassiske socioøkonomiske baggrundsvariable gennem meningsfulde geografiske opdelinger, som kan afdække relevante mønstre i data. På baggrund af den præsenterede forskning i Afsnit 2, anerkender vi relevansen af hvilket område individet har bopæl i, idet området i en vis grad kan have en normativ indflydelse på individet (Fallov & Jørgensen 2018), hvorfor vi har valgt at udforme en række områderelaterede uafhængige variable. Efter at have præsenteret en oversigt over de anvendte registre, vil vi i følgende afsnit skabe overblik over de herudfra dannede variable samt, hvordan udvalgte variable er operationaliseret.

6.2 Sampling

Vi har i forbindelse med forskeradgang til DST adgang til en række på forhånd udvalgte registre fra årene 1980-2022. Dette er derfor allerede med til at afgrænse det mulige datagrundlag. Grundet vores interesse i at afsøge, om individets sociale position i voksenlivet er forudbestemt, har vi fundet det relevant at trække data på individet allerede fra dets første leveår. Vi har ikke fundet tilstrækkeligt belæg for, at nogle tidsperioder i barndommen er mere essentielle for at forudsige individets sociale position i voksenlivet end andre. Vi ønsker derfor at forholde os åben til, at sociale baggrundsdata og hændelser i løbet af hele barndommen kan bidrage til en prædiktiv succes. Dette har ført til, at vi har fastlagt, at data fra individets første til 17. leveår er relevante for undersøgelsen. Det skyldes dels, at inddragelsen af disse data øger datamængden og kompleksiteten heri, hvilket dermed styrker troværdigheden af vores endelige resultater, og dels at det muliggør en dybere indsigt i sociale processer frem for blot midlertidige eller akkumulerede tilstande (De Vaus, 2001:123-124). Valget om at indsamle data fra hele barndommen har dermed konsekvenser for hvornår i voksenlivet vi senest kan prædiktere et udfald på baggrund af. Det senest mulige datapunkt er i 2022, hvor et individ med fødselsår i 1980 således ville fylde 42 år. At følge én enkelt fødselskohorte ville resultere i en mulig undersøgelsespopulation på omkring 50.000 observationer. Grundet kompleksiteten af sociale forholds sammenhæng og et ønske om at øge chancen for en succesfuld prædiktion, har vi fundet det nødvendigt at maksimere undersøgelsespopulationen og derfor tilvalgt at undersøge de tre nævnte fødselskohorter. På den baggrund har vi tilvalgt at prædiktere sociale udfald ved 40 år, som hhv. er i år 2020, 2021 og 2022 afhængig af kohorte. På den måde bliver studiet en art livsforløbsanalyse, der afsøger hvilke faktorer, der er essentielle for elitære sociale positioner i voksenlivet. Ved at trække informationer om barndomsperioden år for år og undersøge om disse informationer gør det muligt at prædiktere opnåelse af et eliteudfald, har vi sikret, at de faktorer, vi undersøger, har den rette tidsmæssige rækkefølge ift. det potentielle udfald og internt i forhold til hinanden. Vi har med andre ord sikret, at 'årsager' kommer før virkning, hvilket højner muligheden for at identificere kausale effekter (De Vaus, 2001:115-118). Derudover har vi tilvalgt en prædiktionsmodel som forstår tidsseriedata, hvilket gør det muligt for modellen af lære mønstre i forandringer i data for hvert år og afsøge om rækkefølgen af disse forandringer har betydning (De Vaus, 2001:115-118).

Samplingprocessen af vores population har som benævnt haft afgørende indflydelse på undersøgelsesdesignet og den endelige datamængde. Den endelige population er alle individer født i Danmark i år 1980-1982 og som stadig leverer data ved 40 år. Vi påbegyndte samplingsprocessen ved at identificere fødselskohorterne via registret BEF i år 2020-2022 ved individernes 40. år, hvor vi identificerede 202.295 observationer (herefter kaldet obs). I den forbindelse frasorterede vi de første 2 obs, da disse var NA i deres personnummer. Resterende frafald af obs og begrundelse herfor fremgår af Tabel 2. Den endelige population består af 140.708 obs, som alle bor i Danmark ved 40 år og leverer data på alle de operationaliserede variable. I dette studie er det afgørende for succesen af vores prædiktion, at de individer vi inddrager, leverer data på alle variable over alle barndomsår samt ved nedslaget som 40-årig. På baggrund af den store inddragelse af individer via dansk registerdata vurderer vi, at den samplede population er repræsentativ for den reelle population. Samplingen er baseret på kriteriet om datatilgængelighed på de operationaliserede variable, hvilket skaber en klar systematik i inklusionen af individer. For yderligere uddybning se afsnit 6.5.

Tabel 2: Frafald

Begrundelse:	Frafald i obs
Missing på pnr	2 obs == 202.293 ant obs
NA i både far_id og mor_id. Disse er enten førstegangs indvandrere eller individer uden registrerede forældre.	38.512 ==163.781 ant obs
Frasorteret hvis forældre har missing mors og fars uddannelse	630 ==163.151 ant obs
Frasorterer de obs som har NA på én af forældrene fra starten af deres liv	3366 == 160.415 ant obs
Frasorterer alle dem med der ikke eksisterer alle år fra de er 0-17 år i bef.	10.534 == 149.881 ant obs
Frasorterer obs i forbindelse med søskende variablene.	35 == 149.846 ant obs
NA på søskende data	108 == 149.738 ant obs
Efter at have sorteret NA på eliteudfald	4640 == 145.098 ant obs
Frasorteret pnr ved NA i Area variable mere end 8 år	4390 == 140.708 ant obs

6.3 Udfaldsvariable (Y)

I det endelige datasæt har vi en række elite-udfald, som vi har afprøvet prædiktion af. Udfaldsvariablene er afgrænset via den tilgængelige data til at omhandle individets sociale position ved 40 år, da vores senest mulige datapunkt er i 2022. Størstedelen af udfaldsvariablene er ved det konkrete nedslag ved 40 år, mens vi ved enkelte udfald har vurderet, at en 3-årig periode var passende for at håndtere eventuelle udsving i data ved eks. personindkomst, som kan forstås som værende en følsom variabel. En specifikation af hvilke variable det drejer sig om fremgår af afsnit 6.3.1.

Tabel 3: Elite-udfald - oversigt

Udfald		
uddtopind	Dummy	**Uddannelse** der gennemsnitligt giver den højeste løn som nyuddannet.
selv2medi	Dummy	**Jobposition** som selvstændig med høj personindkomst.
led	Dummy	**Jobposition** som "lønmodtager med ledelsesarbejde" (DST, 2025a).
top5	Dummy	**Personindkomst** blandt top-5 % af hele befolkningen.
omtop5	Dummy	**Bopæl** i top-5 eliteområde jf. en kombination af faktorer (se afsnit 6.3.1. (operationalisering)).
storbyeli	Dummy	**Bopæl** i én af de fire storbys- eller hovedstadskommuner i Danmark (København, Aarhus, Odense og Aalborg), i kombination af lang videregående uddannelse og høj personindkomst.

6.3.1 Operationalisering af elite-udfald

De elite-udfald vi har fundet det muligt at danne via det tilgængelige registerdata og relevant at undersøge i dette studie, relaterer sig til fire emnemæssige områder, herunder uddannelse, jobposition, personindkomst og deres bopæl. Først har vi variablen **uddtopind**, som relaterer sig til at have en elitær uddannelse. Med udgangspunkt i ét af resultaterne fra Ellersgaards analyse, netop at de individer, der er en del af denne elite i grove træk, har de samme uddannelsesbaggrunde. Konkret er de fremtrædende uddannelsesretninger her Økonomi, Erhvervsøkonomi, Jura eller Statskundskab (Ellersgaard & Larsen, 2019:90-94). Hvilket også er uddannelsesgrene, som er repræsenterede på listen over uddannelser, der som nyuddannet i gennemsnit tjener mest årligt jf. CEPOS' opgørelse. Vi har derfor taget udgangspunkt i top20- uddannelser fra CEPOS' opgørelse, som er baseret på færdiggjorte uddannelser i 2011-2015 for at definere en uddannelseselite (Lundby Hansen, 2019). Jf. uddannelses- og forskningsministeriet var gennemsnitsalderen for de kandidatudannede på de danske universiteter i 2007 28,2 år (UFM, 2013). Tidsperioden som er blevet brugt i CEPOS' opgørelse for færdiggørelsen af uddannelse, passer omtrent med, at vores kohorte er født i 1980-1982 og ville derfor være 30 år i 2010-2012. Dette gør, at vi samlet set vurderer, at CEPOS' opgørelse er relevant til at definere vores uddannelseselite. Konkret har vi manuelt identificeret uddannelseskoden for top20-uddannelserne via registret UDDA og rå-variablen HFAUDD (DST, 2025b). Har individet ved 40år opnået én af disse uddannelser, vil de således være en del af vores definerede uddannelseselite.

De næste to udfaldsvariable relaterer sig til individets jobposition. Først har vi variablen **selv2medi,** som symboliserer, om individet er en elitær-selvstændig i kraft af deres jobposition samt indkomstniveau. Individet skal i nedslagsåret, hvor de fylder 40år jf. registeret AKM og specifikt rå-variablen SOCIO13, være angivet som selvstændige, og derved antage værdien 110, 111, 112, 113 eller 114 (DST, 2025a). Derudover har vi sat det som et krav, at de yderligere skal have en personindkomst over to gange medianindkomsten for hele befolkningen i det år, individet fylder 40 år, hvilket er henholdsvis i 2020, 2021 og 2022. Her gør vi konkret brug af rå-variablen PERINDKIALT_13, som kommer fra registret IND (DST, 2025c). Grænsen ved to gange medianindkomst har vi fastlagt med inspiration fra AE's definition af højere middelklasse, som bl.a. indebærer at have en personindkomst mellem 2-3 gange medianindkomsten af den voksne befolknings samlede bruttoindkomst før skat (Pihl, 2020:7). Variablen selv2medi er defineret således for at indfange erhvervseliten. Vi antager, at det at tjene mere end to gange medianindkomsten placerer individet i toppen af indkomstfordelingen og i kombination med deres jobposition som selvstændig, som havende relativ økonomisk handlefrihed og derigennem mulighed for at investere eller i kraft af deres ejerskab af en virksomhed have adgang til finansielle ressourcer.

Dernæst variablen **led,** som ligeledes er relateret til en elitær form for jobposition. Hvormed vi har til hensigt at skabe en indikator for en lønmodtager-elite, hvilket vi gør i kraft af om individet ved 40år er i en jobposition, som kan kategoriseres som ledelsesarbejde. Konkret har vi gjort brug af registret AKM og heri rå-variablen SOCIO13, hvor værdien 131 angiver, at individet har et "lønmodtagerjob med ledelsesarbejde" (DST, 2025a). I bestræbelsen på at operationalisere en lønmodtager-elite, udgør individer med ledelsesansvar en central position, idet de oftest besidder organisatorisk magt og indflydelse på strategiske processer i virksomheder og institutioner. De adskiller sig tydeligt fra øvrige lønmodtagere grundet det høje ansvarsniveau, samt flere økonomiske og sociale ressourcer. Derved udgør denne gruppe et relevant elite-udfald i kraft af ønsket om at identificere den øverste del af lønmodtagerhierarkiet.

Derefter har vi valgt at definere en ren økonomisk elite på baggrund af individets personindkomst. Udfaldet **top5** indebærer, at individets personindkomst indplacerer sig som en del af de 5% rigeste i den danske befolkning. Konkret måler vi via registret IND og rå-variablen PERINDKIALT_13 ved individets 39., 40. og 41. år og beregner en median heraf. Dette gøres for at tage højde for, at individet kan have haft ét dårligt indkomst år. Medianen af disse 3 år skal som minimum være 11'decil for at individet kan forstås som værende en del af denne personindkomst baserede økonomiske elite. Identifikationen af en ren økonomiske elite er væsentlig, fordi økonomiske ressourcer er central ift. fordelingen af magt og privilegier i samfundet. Økonomiske ressourcer giver mulighed for investeringer, indflydelse og reproduktion af sociale lag til næste generation. Ved at isolere den økonomiske elite uafhængigt af andre faktorer, bliver det muligt at identificere om

økonomisk dominans i sig selv formes af specifikke sociale vilkår i barndommen. Desuden vil simplificeringen af dette udfald om muligt øge chancerne for succes af prædiktionen heraf.

De følgende elite-udfald relaterer sig bl.a. til elitære demografiske områder. Først præsenteres variablen **omtop5**, hvor vi indfanger de individer, som har bopæl i de 5% mest elitære områder. Måden vi har defineret disse områder på er først og fremmest via mikroområde inddeling af Danmark jf. R.L. Lund (2022). Derefter er følgende fem mål blevet beregnet og standardiseret for området ift. alle områder i Danmark. Der er tale om områdets gennemsnitlige uddannelseslængde i måneder, som konkret stammer fra registret UDDA og rå-variablen HFAUUD. Derudover har vi udregnet gennemsnitlig ledighedslængde for alle i området skaleret i måneder. Til dette har vi gjort brug af registret IDAP og rå-variablen ARLEDGR (DST, 2025d). Vi har udregnet en andel af individer i området med en personindkomst større eller lig to gange medianindkomsten for hele den danske befolkning vha. registret IND og rå-variablen PERINDKIALT_13 (DST, 2025c). Derefter har vi udregnet andelen af selvstændige i området med en personindkomst på eller større end to gange medianindkomsten jf. **selv2medi** udfaldet. Sidst men ikke mindst har vi beregnet en andel af lønmodtagere med ledelsesarbejde for området jf. **led** udfaldet. Disse fem mål danner i kombination vores værdi for eliteområde, hvorefter vi har valgt at afgrænse eliteområde definitionen til at gælde de 5% bedste områder. Dette udfald er defineret for at indfange de individer, som lever i et elitært område. Baggrunden herfor er baseret på, at flere forskningsartikler frembringer betydningen af individets nabolag og sociale miljø. Både igennem opvæksten, men ligeledes at man selv aktivt vælger eller stræber efter en bopæl i et område hvor indbyggerne socialøkonomisk minder om en selv eller er en tand bedre stillet. Desuden frembringer Ellersgaard i sin analyse, at individer, der er en del af magteliten, centrerer sig omkring de samme geografiske områder (Ellersgaard & Larsen, 2019:119-125).

Det sidste eliteudfald er vores **storbyeli**, som ligeledes er et udfald, der har til formål at indfange en demografisk elite. Denne variabel er hovedsageligt dannet med inspiration fra J. Andersens nye elite, som indebærer veluddannede individer med bopæl i de større byer. Yderligere også med inspiration fra Piketty og hans 'brahmin-left' og 'merchant-right' (se afsnit 2.2.1). Vi har dannet denne variabel som en kombination af flere faktorer, som skal være gældende for at være en del af storbyseliten. Konkret skal individerne minimum have en lang videregående uddannelse. Dvs. at de minimum skal have 216 mdr. uddannelse. Hvilket vi konkret finder via registret UDDA og rå-variablen HFAUDD, hvor uddannelseslængde fremgår i mdr. Derudover har vi afgrænset gruppen til at skulle have en personindkomst på eller større end 1,5 gange medianindkomsten, som endnu engang er fundet via registret IND og rå-variablen PERINDKIALT_13 (DST, 2025c; Pihl, 2020 :7). Det sidste krav i denne afgrænsning af storbyseliten er, at individet skal have bopæl i en storbyskommune (Aarhus, Odense eller Aalborg) eller hovedstadskommune jf. DST's kategorisering (DST, 2025e). Herefter har vi frasorteret micro-områder på befolkningstæthed, dvs. at de områder, hvor antal individer bosat pr. kvm2. er lavere end 1500 individer, er frasorteret. Dette har vi gjort for at undgå at landsbylignende

områder i kommunerne ikke bliver regnet med som værende storbyseliter. Grænsen for befolkningstæthed er blevet fastlagt med inspiration fra DST (2018).

Alle ovenstående elite-udfald er defineret med det formål at undersøge, hvorvidt en prædiktion af de individer, der kan kategoriseres som tilhørende disse eliter, er meningsfuld. Efterfulgt af en yderligere afdækning af hvilke sociale faktorer i barndommen, der har betydning for det pågældende elite-udfald. Dette leder os videre til hvilke baggrunds- og områdevariable, der i dette speciale er valgt inddraget.

6.4 Baggrunds- og områdevariable

For oversigt af baggrunds- og områdevariable se: **Bilag 1: Baggrunds- og områdevariable – oversigt**

Disse variable indebærer basale personkarakteristika som køn, etnicitet, fødselsår, alder mm. Samt familie relaterede variable, socioøkonomi og demografi. Variablene er opdelt i statiske (konstante) = k, dynamiske = d over tid = t_i og irregulære hændelser = h_{abc}

Eksempler på konstante variable er forældres etnicitet, individets eget køn og antal ældre søskende, som vi alle antager enten allerede er indtrådt eller indtræder ved individets fødsel, og forbliver samme værdi igennem hele individets liv. En dynamisk variabel som årligt kan ændre sig er f.eks. forældres indkomst, forældres beskæftigelse, totalt antal søskende i familien, bopælsområdes værdier for indkomst, beskæftigelse og uddannelse. Disse variable kan derved ændres over tid i takt med det levede liv. Konkret arbejder vi med ét datapunkt pr. år og derved kan de dynamiske variable højst ændre sig fra år til år. De irregulære hændelser kan indtræffe et givent år, men akkumuleres ikke over tid. Det kan f.eks. være forældres skilsmisse, forældres død eller hvis individet får en ny søskende af enten samme eller forskelligt køn efter egen fødsel. I det år, hvor hændelsen indtræffer, vil variablen antage værdien 1. Dvs. at hvis et individ får en yngre søskende i det år, det selv er 5 år, vil variablen 'nysos' (variablene hedder konkret "nysossamkon" og "nysosforskon") antage værdien 0 i individets 0.-4. år, mens det i år 5 vil antage værdien 1, og herefter igen 0, da det netop er i individets 5. år, at det oplever hændelsen af en ny søskende.

Med denne overordnede forståelse af hvilke typer af variable, vi har valgt inddraget i dette speciale, vil vi nu lede videre til et overordnet operationaliserende afsnit omhandlende vores baggrunds- og områdevariable, som efterfølges af et overblik over hvordan data er opbygget, hvilket fremgår i afsnit 6.6 Dataopbygning.

6.4.1 Operationalisering af baggrunds- og områdevariable

Generelt har vi forsøgt at være åbne for en bred vifte af faktorer, der kunne have betydning for, om et elite-udfald kan prædikteres. Vi har derfor forsøgt at operationalisere flest mulige meningsfulde variable ud fra de tilgængelige registre for at være åbne for, at disse kunne bidrage med en nuance til undersøgelsen. Desuden har vi bestræbt os på at fokusere på forhold, som individet ikke selv har haft

direkte indflydelse på, eller kun i begrænset omfang har kunnet påvirke. Dette muliggør, at analysen i højere grad indfanger strukturelle betingelser fremfor individuelle valg.

Konkret har vi skabt i alt 32 baggrundsvariable og 27 områdevariable. Grundet antallet af variable er skaleringen og beskrivelsen af disse vedlagt som Bilag 1. Af baggrundsvariablene er der en række, som beskriver individet selv i form af køn, samt om individet har søskende og om disse er ældre eller yngre, og hvilket køn de har. Derudover har vi variable, der fortæller om individet har fritidsjob som ung og den eventuelle løn i dette job. En anden række af baggrundsvariable er de, som er relateret til individets forældre, i form af uddannelseslængde, personindkomst, jobposition, etnicitet, alder, om forældrene er sammen eller skilt, og om forældrene dør i løbet af individets 0 – 17 år. Sidst men ikke mindst er der en række variable, som knytter sig til individets søskende. Om de har ældre søskende, om de løbende får en ny søskende, hvilket køn deres søskende har og aldersdifferencen mellem individet og deres nærmeste ældre søskende samt mellem individet og dets nærmeste yngre søskende. En lang række af disse er klassiske sociologisk relevante variable. Nogle af de mindre klassiske variable, vi har valgt at inddrage i undersøgelsen, er bl.a. de detaljerede søskende variable, som vi i kraft af problemfeltet har fundet relevante at inddrage. Konkret har vi fundet det interessant om specifikke søskende konstellationer i form af roller samt køn kan bidrage med nuancer i, hvilke faktorer der har betydning for, om individet opnår et bestemt elite-udfald. Derudover har vi valgt at inddrage områdevariable, som er forholdsvis udbredt i sociologisk forskning, men i kraft af de frembragte forskningsartikler i problemfeltet omhandlende betydningen af det miljø, som man vokser op i, samt de tendenser som fremgår af J. Andersens og Ellersgaards analyser om, at elitens bosætning klynges omkring de større byer, har vi fundet det interessant at udbygge detaljegraden af områder og dets betydning for, om individet opnår elite-status i deres voksne liv (2024)(Ellersgaard & Larsen, 2019:119-125). Konkret har vi gjort brug af en inddeling af Danmark i 8043 micro-områder (Lund, 2022), som er det første lag af detaljegrad. Herefter har vi identificeret en række interessante variable, herunder individets flyttemønstre, befolkningstæthed i området, andele i elitære jobpositioner, andele med elitære indkomster, gennemsnitlige indkomst percentil, ledighed i området, uddannelseslængde og om der er kønsforskelle ift. de allerede benævnte variable i området. Variablene beskriver samlet set individets liv fra 0-17 år i en tidsserie, som vil fremvist i afsnit 6.6 Dataopbygning.

6.5 Datatilgængelighed

Datatilgængelighed er en central forudsætning for studiets validitet, både når det gælder populationens repræsentativitet, men også tilgængeligheden af relevant data, som kan operationaliseres til variable.

I forhold til den population vi ønsker at undersøge i dette studie, er det væsentligt at overveje, hvorvidt vores populationssample afspejler denne. I kraft af, at vi har adgang til dansk registerdata, er det muligt at indfange alle de individer i Danmark, som lever op til samplingskravene jf. afsnit 6.2. Dette

gør, at studiet i høj grad afspejler populationen, da vi indfanger alle de individer, der har tilgængelige data registreret.

Dog er der alligevel tale om et frafald i samplingen på omkring 60.000 observationer, da disse ikke leverer den krævede data på de udvalgte variable. Dette kan bl.a. skyldes individer der primært befinder sig i udlandet frasorteres. Derudover er førstegenerationsindvandrere udelukket fra at indgå, fordi vi ikke kan observere på disse fra og med 0 år. Desuden er individer som dør før de bliver 40 år, heller ikke med i undersøgelsen, da populationen er dannet ud fra år 2020-2022 og de individer, som i det pågældende år fylder 40år. Desuden kan vi ikke bruge deres barndomsdata, når vi ikke kan måle deres livsudfald som 40årige. Yderligere er de individer, som blot ikke er registreret i Danmark i det ene år, hvor de fylder 40 år, ligeledes ikke med i undersøgelsen.

Ud over datatilgængelighed omhandlende den rette undersøgelsespopulation er det ligeledes relevant at overveje hvorvidt der er muligt at indfange relevante variable. I dette studie er vi afgrænsede til at kunne danne variable ud fra de pågældende registre vi har adgang til, se Tabel 1: Registeroversigt. Derudover kunne andre variable, som ikke er tilgængelige på registrene, antages at være relevante. Variable som kunne afspejle sundhed, talent, IQ, netværk, foreningsliv eller måske politisk overbevisning. Der kan således siges at være omitted variable bias knyttet til vores undersøgelse, da vi ikke har mulighed for at inkludere disse muligt relevante variable. Konsekvensen heraf kan være, at nogle mønstre ikke kan identificeres og at precision mindskes. Vi antager dog at dette ikke har en betydelig indvirkning på vores endelige resultater, da har inkluderet en lang række variable i prædiktionen.

6.6 Dataopbygning

Nedenstående figur viser et syntetisk eksempel på dataopbygning som tidsserie i long format med få udvalgte variable (se dybdegående beskrivelse af eksempler baseret på syntetisk data i afsnit 5.4 Etik). Data i Figur 2 er således ikke baseret på virkelige personer.

Figur 2: Eksempel på dataopbygning – baseret på syntetisk data*

	ID	tid	køn_barn	n_ældre_søsk	mor_etn	mor_opr	far_etn	far_opr	mor_indk_perc	mor_udd_mdr	mor_alder	far_indk_perc	far_udd_mdr	far_alder	forældre_skilt	mor_død	far_død	total_søsk
0	A	0	M	0	dansk	vestlig	dansk	vestlig	57	150	29	34	108	37	0	0	0	1
1	A	1	M	0	dansk	vestlig	dansk	vestlig	60	150	30	32	108	38	0	0	0	1
2	A	2	M	0	dansk	vestlig	dansk	vestlig	54	150	31	37	108	39	0	0	0	2
3	A	3	M	0	dansk	vestlig	dansk	vestlig	61	150	32	37	108	40	0	0	0	2
4	A	4	M	0	dansk	vestlig	dansk	vestlig	60	150	33	38	108	41	0	0	0	3
5	A	5	M	0	dansk	vestlig	dansk	vestlig	59	150	34	40	108	42	0	0	0	3
6	A	6	M	0	dansk	vestlig	dansk	vestlig	60	150	35	40	108	43	1	0	0	3
7	A	7	M	0	dansk	vestlig	dansk	vestlig	55	150	36	40	108	44	0	0	0	3
8	A	8	M	0	dansk	vestlig	dansk	vestlig	50	150	37	38	108	45	0	0	0	3
9	A	9	M	0	dansk	vestlig	dansk	vestlig	44	150	38	36	108	46	0	0	0	3
10	A	10	M	0	dansk	vestlig	dansk	vestlig	41	150	39	39	108	47	0	0	0	4
11	A	11	M	0	dansk	vestlig	dansk	vestlig	42	150	40	44	108	48	0	0	0	4
12	A	12	M	0	dansk	vestlig	dansk	vestlig	40	150	41	43	108	49	0	0	0	4
13	A	13	M	0	dansk	vestlig	dansk	vestlig	36	150	42	45	108	50	0	0	0	4
14	A	14	M	0	dansk	vestlig	dansk	vestlig	46	150	43	42	108	51	0	0	0	4
15	A	15	M	0	dansk	vestlig	dansk	vestlig	46	150	44	41	108	52	0	0	0	4
16	A	16	M	0	dansk	vestlig	dansk	vestlig	47	150	45	40	108	53	0	0	0	4
17	A	17	M	0	dansk	vestlig	dansk	vestlig	43	150	46	45	108	54	0	0	0	4
18	B	0	K	0	dansk	vestlig	dansk	vestlig	29	144	26	58	120	38	0	0	0	1
19	B	1	K	0	dansk	vestlig	dansk	vestlig	29	150	27	59	120	39	0	0	0	1
20	B	2	K	0	dansk	vestlig	dansk	vestlig	30	150	28	60	120	40	1	0	0	1
21	B	3	K	0	dansk	vestlig	dansk	vestlig	26	150	29	63	120	41	0	0	0	1
22	B	4	K	0	dansk	vestlig	dansk	vestlig	24	150	30	69	120	42	0	0	0	1
23	B	5	K	0	dansk	vestlig	dansk	vestlig	28	150	31	71	120	43	0	0	0	1
24	B	6	K	0	dansk	vestlig	dansk	vestlig	35	150	32	73	120	44	0	0	0	1
25	B	7	K	0	dansk	vestlig	dansk	vestlig	35	150	33	74	120	45	0	0	0	1
26	B	8	K	0	dansk	vestlig	dansk	vestlig	32	150	34	78	120	46	0	0	0	1
27	B	9	K	0	dansk	vestlig	dansk	vestlig	38	150	35	80	120	47	0	0	0	1
28	B	10	K	0	dansk	vestlig	dansk	vestlig	41	150	36	83	120	48	0	0	0	1
29	B	11	K	0	dansk	vestlig	dansk	vestlig	33	150	37	89	120	49	0	0	0	1
30	B	12	K	0	dansk	vestlig	dansk	vestlig	33	150	38	90	120	50	0	0	0	1
31	B	13	K	0	dansk	vestlig	dansk	vestlig	39	150	39	85	120	51	0	0	0	1
32	B	14	K	0	dansk	vestlig	dansk	vestlig	44	150	40	87	120	52	0	0	0	1
33	B	15	K	0	dansk	vestlig	dansk	vestlig	42	150	41	96	120	53	0	0	0	1
34	B	16	K	0	dansk	vestlig	dansk	vestlig	39	150	42	98	120	54	0	0	0	1
35	B	17	K	0	dansk	vestlig	dansk	vestlig	39	150	43	100	120	55	0	0	0	1

Vores endelige datasæt er bygget op på sammenlignelig vis med Figur 2. Ethvert individ, som er en del af vores endelige undersøgelses population, fremgår derved i datasættet med én række indeholdende informationer i de dannede baggrunds- og områdevariable pr. år fra individets 0-17 år. Det vil sige, at vores endelige antal observationer i datasættet er 2.392.036, altså 17 gange det antal vores endelige undersøgelsespopulation består af.

Separat fra dette har vi trukket informationer på undersøgelsespopulationen ved individernes 40. år for at kunne identificere, om de opnår vores definerede elite-udfald. I Tabel 4 herunder fremgår et overblik over andele af populationen, der ved 40 år opnår et pågældende elite-udfald samt en kønsmæssig opdeling heraf.

Tabel 4: Population og andele af Yi

$N(y_i)$ / andel y=1 af N	N = 140,708
y1 (uddtopind)	7612 0.0541
y2 (selv2xmedi)	1463 0.0104
y3 (led)	5600 0.0398
y4 (top5)	9596 0.0682
y5 (omtop5)	12,002 0.0853
y6 (storbyeli)	10,075 0.0716

7 Metode

Flere tidligere samfundsvidenskabelige studier har benyttet prædiktive metoder via en form for såkaldt 'feature engineering' (Brand et al. 2021, Davidson 2019), hvor tidsserier omformes til "flade" datastrukturer, og tidsdimensionen fjernes, mens diverse datapunkter modelleres til features, der kan indgå i et neuralt netværk, en random forest eller andre *tree based models*. Features baseres ofte på opsummerede værdier fx gennemsnit, trends eller ændringer over tid. Mens dette sagtens kan være en meningsfuld måde at modellere data på til brug i en prædiktiv metode og machine learning i visse sammenhænge (ibid.), kan det give udfordringer med kausal rækkefølge, når tidsdimensionen fjernes, og alle datapunkter sidestilles uden tidslig dybde (De Vaus, 2001:115-118), mens datakompleksiteten reduceret betydeligt med indbygget risiko for at fjerne væsentlige variationer i sociale faktorer over tid. En løsning herpå, hvor der forsat benyttes features, kan være at opdele barndommen i teoretisk afgrænsede kortere perioder, der kan antages at være betydningsfulde, og som kan opfattes som udviklingsmæssigt sammenhængende perioder, fx tidlig barndom (0-5 år), senere barndom (6-11 år), og tidlig ungdom (12-17 år). Ud fra disse kunne dannes sekvensspecifikke features, hvilket vil give mulighed for analytisk at kunne adskille disse i et groft tidsperspektiv i de senere resultater.

I denne undersøgelse har vi dog fra begyndelsen haft et ønske om at arbejde med en model, der kan tage højde for datapunkternes tidslige rækkefølge og komplekse indbyrdes relation til hinanden. Ved at forsøge prædiktion af livsudfald med en sådan model, argumenterer vi for, at det ikke kun betyder noget, *hvad* der sker, men også *hvornår* det sker, og i *hvilken rækkefølge* det sker – altså en løbende indlejring af livsbegivenheder og sociale faktorer i den sociale kontekst. Med sidstnævnte forstås, at det ikke kun har betydning, hvornår en hændelse i et livsforløb sker ud fra en absolut betragtning af individuel alder, eller den absolutte tid, som er gået siden år 0 (det analytiske startpunkt), men at det også har afgørende betydning, hvad der sker umiddelbart *før, samtidig* og *efter* en hændelse i et livsforløb, samt hvordan alle hændelser i tidligere eller senere i forløbet generelt *vægter* og påvirker

hinanden. Alle livshændelser kan altså siges at være indlejret i en sammenhæng, eller en sekvens, som samlet set giver en betydning og kontekst, der tilføjer vigtig dybde og skaber bedre forudsætninger for præcision i den prædiktive analyse.

Til at håndtere denne kompleksitet benytter vi en særlig transformerbaseret ML- model, som som i dag anses for den mest avancerede metode til modellering af fulde livsforløb som sekvenser, hvor alle datapunkter er indlejret i et komplekst mønster af tidslig kontekst og interne sammenhænge mellem datapunkter. Modellen muliggør udnyttelse af det fulde potentiale i kompleksitet, variation og tidslig struktur i den tilgængelige data.

7.1 Transformeren – et nybrud i sekventielle analyser med ML

Transformerarkitekturen, introduceret af Vaswani et al. I "Attention Is All You Need" (2017), markerer et afgørende gennembrud i for moderne machine learning, som siden har dannet grundlag for en lang række avancerede modeller inden for bl.a. natural language processing (NLP), billedgenkendelse, bioinformatik og senere også analyser af sekventielle sociale data.

I modsætning til tidligere sekventielle machine learning modeller som rekursive neurale netværk (RNN) og deres videreudvikling long short-term memory-netværk (LSTM), bearbejder transformerarkitekturen hele sekvensen samtidigt. RNN-arkitekturer blev tidligere betragtet som det foretrukne værktøj til analyse af sekventielle data og bearbejder en inputsekvens trinvis ét datapunkt ad gangen, hvor hvert nyt datapunkt kombineres med en form for intern hukommelse fra det foregående trin, hvorved modellen forsøger at lære, hvordan den nuværende tilstand afhænger af tidligere elementer i sekvensen. Dette gør dem dermed stærkt afhængige af rækkefølge og lineær datastruktur, og har den væsentlige begrænsning, at RNN-modeller kan have svært ved at bevare betydningen af information over lange sekvenser, hvilket konkret betyder, at hvis en vigtig hændelse indtraf tidligt i en sekvens, kan dens betydning gradvist "glemmes" af modellen, jo længere væk fra hændelsen datapunkterne i sekvensen bevæger sig. Dermed har RNN-modeller tendens til at tillægge mest vægt på de "nyeste" hændelser, når et udfald skal prædikteres – et problem, der kaldes "vanishing gradients", hvor signaler, der transporteres gennem modeltræningen bliver svagere og svagere, jo længere tilbage i tiden de stammer fra. LSTM-modeller er udviklet med en mere kompleks indre struktur, med såkaldte "gate mechanisms", der regulerer hvilke informationer, der skal gemmes, slettes og videreføres i sekvensens hukommelse. Dette gør LSTM i stand til at bevare relevant information over længere tid og bedre modellere komplekse sekventielle afhængigheder mellem datapunkter. Men også LSTM-arkitekturer er grundlæggende begrænset af sin sekventielle natur; de bearbejder hvert inputelement ét ad gangen og opdaterer deres interne tilstand løbende, hvilket har den konsekvens, at analysen af en sekvens ikke kan paralleliseres, og hvert trin afhænger af det foregående, hvilket skaber en strukturel begrænsning i modellernes evne til at lære langtrækkende relationer mellem datapunkter effektivt. Transformerarkitekturen bryder radikalt med denne struktur og tillader modellen at se og

sammenligne alle datapunkter i hele sekvensen på én gang og vurdere relationer mellem alle elementer i sekvensen på samme tid. Dette opnås gennem en særlig mekanisme kaldet *self-attention* (Vaswani 2017).

Self-attention er en matematisk metode, som beregner, hvor meget hvert datapunkt bør 'vægte', når den forsøger at forstå meningen med et specifikt datapunkt i konteksten af de øvrige. Eksempelvis kan en transformer-model, der analyserer sætningen "hun tog sin cykel, fordi den var punkteret", bruge self-attention til at forstå, at "den" refererer til "cykel" og ikke til "hun", selvom der er flere ord i mellem. Det samme princip gælder for sekvenser, der ikke er sproglige i traditionel forstand, men som indeholder hændelser over tid: en transformermodel kan fx identificere, at en arbejdsløshedsperiode som 25-årig kan hænge sammen med en tidligere afbrudt uddannelse som 21-årig, selvom der er flere hændelser i mellem, og uden at disse hændelser nødvendigvis forekommer i et fast mønster. Med andre ord evaluerer modellen vha. self-attention hvilke dele af sekvensen, der er mest relevante i konteksten af hvert element – og den gør dette for hele sekvensen på én gang (Vaswani 2017).

Tranformermodeller består typisk af en *encoder*, som behandler inputsekvensen og skaber en kontekstualiseret repræsentation af hvert element, og en *decoder,* som genererer en ny sekvens (ofte som output). I prædiktive analyser anvendes dog oftest kun *encoder*-delen, fx i modeller som BERT (Devlin et al., 2019), hvor en "encoder-only"-struktur benyttes til opgaver som tekstanalyse, klassifikation og prædiktion.

Et andet vigtigt element i transformerarkitekturen er *positional coding*. I og med at modellen ser hele sekvensen samtidigt og ikke bearbejder datapunkter ét ad gangen som i RNN's, mister modellen naturligt information om rækkefølge. Derfor tilføjes en ekstra kodning, der fortæller modellen, på hvilken position i sekvensen hvert datapunkt befinder sig. Dette er afgørende i alle anvendelser, hvor rækkefølgen af hændelser bærer semantisk information – fx i sprog, musik eller livsforløb. Den fleksibilitet og skalerbarhed, som transformermodellen tilbyder, har gjort den til standardarkitektur i næsten alle moderne store sprogmodeller, herunder GPT ('Generative Pretrained Transformer', Radford et al. 2018), BERT ('Bidirectional Encoder Representations from Transformers'), RoBERTa ('Robustly Optimized BERT Pretraining Approach', Liu et al. 2019) og T5 ('Text-To-Text Transfer Transformer', Raffel et al. 2020). Først for relativt nylig er forskning inden for sociale data dog begyndt at benytte transformerarkitekturer, hvoraf Life2Vec-modellen (Savcisens et al. 2023a) er et af de mere banebrydende eksempler, som netop anvender transformerarkitektur til at analysere strukturerede livsforløb, hvor et forløb repræsenteres som en sekvens af hændelser over tid.

Det særligt interessante ved transformerarkitektur i en samfundsvidenskabelig kontekst af data om livsforløb, er dens evne til at håndtere både kompleks tidslig struktur og semantisk forskelligartethed i inputdata; hvor traditionelle statistiske modeller typisk kræver nøje udvalgte og

aggregerede variable, kan transformeren arbejde direkte med sekvenser og rå hændelser og selv lære, hvilke mønstre, der er vigtige, som alternativ til fagspecifik feature engineering forankret i forskerens teoretiske og analytiske antagelser om relationer mellem specifikke sociale faktorer og livsbegivenheder. Dermed kan den potentielt identificere skjulte livsbaner og uventede relationer mellem datapunkter, som samlet set kan forbedre modellens prædiktionsevne. Samtidig giver transformermodellens opbygning mulighed for en høj grad af *parallelisering*, hvilket gør den effektiv, når den trænes på store datamængder. I modsætning til RNN's, hvor hvert trin i sekvensen afhænger af det foregående og derfor ikke kan beregnes samtidig, tillader transformerens samtidige opmærksomhed, at hele sekvensen behandles parallelt, hvilket gør det muligt at skalere til store populationer, som det ofte er nødvendigt i sociologiske analyser baseret på registerdata eller longitudinelle undersøgelser. Transformerens opbygning muliggør altså, at man med de rette justeringer kan modellere komplekse, sekventielle sociale livsforløb med henblik på at prædiktere specifikke livsudfald, og det er denne arkitektur, vi tager udgangspunkt i, når vi i det følgende afsnit uddyber klargøring af data til brug i en transformermodel, samt når vi senere beskriver den specifikke modelvariant, vi har udviklet til prædiktion af eliteudfald.

Den transformer-baserede tilgang til prædiktionsanalyse af registerbaseret tidsseriedata som syntetisk sprog er ny i en sociologisk kontekst, omend den er benyttet til andre formål inden for samfundsvidenskab generelt, særligt centreret omkring computationel tekstanalyse (Wankmüller 2021, Benedetto et al 2023) og sentimentanalyse, som har vundet indpas i de senere år, bl.a. i analyser af massive digitale tekstmængder på sociale medier (Analyse og Tal, 2025). Den transformerbaserede tilgang til at modellere og analysere tidsserier og forløb som syntetisk sprog er dog testet og benyttet inden for andre videnskabelige grene, som fx data science, sundhedsvidenskab og meteorologi m.fl med gode resultater (Grechishnikova 2021, Cai et al. 2021). Savcisens et al. (2023a) har med et videnskabeligt udgangspunkt i et team af forskere fra data og computer science, sundhedsvidenskab og psykologi senest benyttet netop dansk registerdata vedr. sundhed og arbejdsmarked som grundlag for en transformerbaseret modellering og analyse af netop livsforløb repræsenteret som sekvenser af tekst til at prædiktere livsudfald som død og personlighed, og de fandt her, at metoden performer væsentligt bedre end en række andre typer prædiktionsmetoder. Inspireret af deres tilgang, men dog med væsentlige forskelle i dataudvalg, model- og dataopbygning og ikke mindst en stærkt sociologisk forankret stillingtagen til data som repræsentationer af sociale faktorer og virkelighedens sociale kompleksitet, vil vi i det følgende gå yderligere i dybden med de metodiske valg og transformermodellens forberedelse og implikationer.

7.2 Klargøring af data til transformermodel

For at kunne benytte en transformermodel til prædiktion af eliteudfald på baggrund af omfattende og detaljerig data om den udvalgte populations liv fra fødsel til 18. år, opbygger vi vores datasæt, så det repræsenterer progression af individernes barndom og ungdom som livsforløb – eller tidsserier.

Livsforløbene er opbygget på baggrund af tidligere beskrevet registerdata, som indeholder detaljerede informationer om familieforhold og børn, beskæftigelse, jobtype, bopæl, uddannelse, indkomst, m.fl. Livsforløbene udvikler sig over tid og giver detaljerig information om livsbegivenheder og deres rækkefølge med høj tidmæssig præcision.

Denne form for komplekst sammensat tidsseriedata medfører betydelige udfordringer ved brug af klassiske kvantitative metoder og ML-baserede klassiske neurale netværk, såsom irregulære sampling-intervaller, missing-værdier i data, og komplekse interaktioner mellem variable over tid. Samtidig kan det være udfordrende at skalere og kræver ofte en betydelig databehandlingsproces for at udtrække brugbare features. Ved i stedet at bruge en transformermodel undgår vi at arbejde med manuelt udformede features, som akkumulerer flere års data til gennemsnitsværdier o.l., og dermed kan vi bevare den tidslige og kontekstafhængige indlejring af datapunkter og en betydelig detaljegrad og granularitet i data (Savcisens, et al. 2023a:4-5). Her kodes data på en måde, der udnytter en tidsseries lighed med sprog. I vores tilfælde danner hver kategori af variable, både kontinuerligt målte dynamiske variable, som kan ændre værdi årligt, konstanter, som er givet fra fødslen og enkeltstående irregulære hændelser, samlet set et vokabular, som sammen med tids- og positionskodning gør det muligt at betragte hver livsbegivenhed som en form for sætning bestående af syntetiske ord, også kaldet 'tokens'. Ved at kode data på denne måde, kan vores syntetiske sprog således indfange informationer a la "I 1985 boede person "A"'s far i Ringsted og fik femtusind kroner som skolelærer på en skole i Roskilde" eller "I 1990 flyttede person "B" fra Aarhus på landet i Vestjylland med sin arbejdsløse mor". Dermed kan progressionen i en persons liv siges at være repræsenteret som en række af sådanne sætninger, der tilsammen danner et individuelt livsforløb. Vores tilgang gør det muligt at inkludere en bred vifte af detaljeret information om livsbegivenheder i individuelle liv uden at gå på kompromis med indhold og struktur i den rå data. Transformermodellen er dermed meget lig de sprogmodeller, der de senere år med hastig fart har været under konstant udvikling, som baserer sig på 'Natural Language Processing' eller *NLP*, som tidligere forklaret i afsnit 7.1 (ibid.).

Særligt om brug af Python:

I forbindelse med klargøring af data til brug i tranformermodel benyttes Python som primært kodeværktøj. Specifikt benyttes pakkerne PyTorch, Scikit-learn og TensorFlow til tokenization, sekvensering, padding, attention mask, positionskodning, modelarkitektur, træningsloops og evaluering. Se Referencer for link til pakke-dokumentation.

7.3 Livsforløb som tekstsekvens

Tidligere i afsnit 6 har vi gennemgået databehandlingsprocessen, hvor vi vha. store mængder registerdata fra DST fra en række forskellige rå-datasæt har sammensat et endeligt datasæt for en populations livsforløb fra 0 til 17 år i 'long format', altså hvor hvert år i en persons liv er repræsenteret ved en række fra og med år 0 til og med år 17. I den videre klargøringsproces frem mod at repræsentere et livsforløb som en tekstsekvens indgår en række overvejelser og yderligere databehandlingstrin, som

gennemgås herunder. Vigtigt er det i første omgang at forstå, at hvert individs livsforløb omdannes fra at være repræsenteret vha. klassisk tabeldata som tidsserie i long format til én lang tekstsekvens pr. forløb. Dette uddybes yderligere senere, men først gennemgås klargøringstrin, som går forud for opbygningen af tekstsekvenser.

7.3.1 **Binning** og **tekst-værdier** for udvalgte variable

Transformermodellen analyserer grundlæggende livsforløb som sekvenser af tekst med informationer indlejret (embedded) i andre informationer – som en NLP-model ville gøre det. Modellen er derfor meget afhængig af *genkendelighed* for at kunne identificere tydelige mønstre. Den forstår ikke nødvendigvis i sig selv, at en årsindkomst på 400.000 kr. i et analytisk perspektiv er stort set identisk med en årsindkomst på 400.001 kr. Den forstår kun ligheden mellem observationerne i kraft af deres indlejring i sekvensens øvrige informationer, som fx relationen til andre observationer og position i sekvensen (*embedding layers, se afsnit 7.3.5*). Dette skaber et datahåndteringsdilemma, da vi på den ene side ønsker så detaljerige datainputs som muligt med en høj grad af *granularitet*, mens vi på den anden side ønsker, at modellen kan arbejde med genkendelighed i analysen af tekstsekvenser – altså, at der er så få observationer som muligt med unikke værdier. Til en vis grad kan dette løses i kraft af størrelsen på datasættet (~140.000 individer, hvis forældre leverer årlige informationer til sekvenserne om bl.a. indkomst, dvs. 140.000 x 18 observationer for mors indkomst og samme antal for fars indkomst), og dog vil visse variable (som fx indkomst i dette eksempel) stadig være i et format, hvor meget få årlige indkomster vil være 100% identiske, hvis vi laver et indkomst-input, der er målt pr. år i kroner. Derfor er ***binning*** centralt for både indkomstrelaterede værdier, men også for en række øvrige variable. Binning er en almindelig præprocesseringsteknik i data science, hvor numeriske værdier grupperes i et mindre antal intervaller eller kategorier, kaldet *bins* (Liu et al. 2020). Det betyder grundlæggende, at vi reducerer måleenhedsgranulariteten i data på en hensigtsmæssig måde, dog uden at reducere den for kraftigt, da vi så mister analytisk kraft i modellen. Et konkret eksempel på binning i vores klargøring af data, er en relativering af indkomstniveauer vha. percentilinddeling ud fra placeringen i den generelle indkomstfordeling i det pågældende år. Alle indkomstrelaterede variable kan dermed antage 100 forskellige værdier, hvilket vi vurderer giver modellen mulighed for at genkende værdier i tekstformat pga. tilstrækkeligt mange forekomster, og samtidig fortsat giver en tilpas granularitet i måleenheden, således at modellen kan adskille indkomstværdier meningsfuldt. Blandt andre konkrete eksempler på binning i vores datasæt er gennemsnitsalder i bopælsområde målt i hele år (decimaler er derved udeladt), uddannelsesniveauer målt i antal år med ét decimal oprundet til nærmeste halve år (fx 11.0 og 14.5 år), og ledighed pr. år målt i antal måneder med ét decimal oprundet til nærmeste halve måned (fx 0.5 og 3.0 måneder). Desuden er alle variable, som måles i andele (altså et tal mellem 0 og 1) reduceret til to decimaler, hvilket teoretisk betyder, at disse variable vil kunne antage maksimalt 101 forskellige værdier. Dog vil de fleste af disse variable aldrig nå over et niveau på 0.50, da de pågældende variable primært arbejder med andele af personer i et mikroområde, der lever

op til forskellige elite-relaterede parametre, hvorfor det faktiske antal unikke værdier vil være en hel del lavere. Samlet set forsøger vi med binning at øge modellens mulighed for genkendelighed af talværdier, der skal processeres som tekst, uden at gå på kompromis med praktisk og analytisk anvendelig detaljegrad i variabelværdier.

Visse kategoriske og nominale variable er desuden omdannet fra tal til **tekstværdier** på forhånd, da dette øger forklarbarheden og lethed i fortolkningen ved det senere modeloutput. Dette gælder fx for variable som køn ('M' eller 'K'), etnicitet (fx 'dansk'), oprindelse (fx 'ikke-vestlig'), landsdel (fx 'Østjylland' og 'Kbhomegn'), befolkningstæthed i bopælsområde (fx 'meget lav' og 'høj') og fødselsmåned (fx 'februar' og 'juli').

7.3.2 Håndtering af **konstanter, dynamiske** variable og **irregulære** hændelser

Data i tidsserien klargøres til modelinput ved at opdele alle uafhængige variable efter type: konstante, dynamiske og irregulære hændelser. Dette er et vigtigt step i klargøring til modelinput, da forskellige variabeltyper håndteres forskelligt, når data senere benyttes til at opbygge sekvenser i tekstformat med en indlejret tidsdimension.

Konstanter vil altid stå i starten af sekvensen, men samtidig er det vigtigt, at de ikke opfattes som bundet til et bestemt tidspunkt (fx år 0). De indgår i stedet i sekvensen som en form for "overskrift", som den øvrige sekvens kontinuerligt er indlejret i (som embedding layer) uanset, hvor langt fremme i tid og dermed i sekvensen, vi bevæger os. Konstanter i vores tidsserie er alle variabelværdier, som er indtruffet ved fødslen, og som ikke kan ændre sig efterfølgende, fx køn, mors og fars etnicitet og oprindelse, antallet af ældre søskende, aldersdifference til nærmeste ældre søskende, nærmeste ældre søskendes køn, m.fl. Alle konstanter er som udgangspunkt målt i år 0. Dog benytter vi desuden en enkelt *udskudt konstant,* nemlig aldersdifference til nærmeste yngre søskende. Denne variabel vil altid antage værdien NaN i begyndelsen af tidsserien (år 0), da ankomst af en yngre søskende selvsagt endnu ikke er indtruffet som irregulær hændelse. Dog vil den, så snart (hvis nogensinde) en yngre søskende ankommer, antage en værdi i antal år med ét decimal. Det er derfor denne udskudte værdi, som indgår som konstant i starten af sekvensen sammen med de øvrige konstanter.

Dynamiske variable vil indgå i sekvensen hvert år i tidsserien fra og med år 0 til og med år 17 tilknyttet et tids- og et concepttoken (se 7.3.3 om tokenization). Disse variable er fx mors og fars indkomster, uddannelsesniveauer og ledighed, besiddelse af lederposition eller selvstændig i egen virksomhed m.fl. Det er desuden også alle områderelaterede variable, knyttet til det mikroområde, den enkelte bor i det pågældende år.

Irregulære hændelser indgår *kun* i sekvensen i det år, hvor hændelsen indtræffer, dvs. når variablen antager værdien 1. I alle øvrige år (hvor de antager værdien 0) indgår disse variable ikke i sekvensen. Blandt disse variable er fx forældres skilsmisse, flytning til nyt mikroområde, tilkomst af ny søskende af samme eller forskelligt køn, og mors eller fars død.

7.3.3 Tokenization, sekvensering og padding

Tokenization og sekvensering er de to helt centrale komponenter i omdannelsen af livsforløb som klassisk tidsserie til tekstsekvenser, hvor et livsforløb er repræsenteret ved én lang tekststreng – også kaldet en tekstsekvens. Tokenization omdanner samtlige datapunkter til en række "ord" eller tokens, som danner grundlaget for en syntetisk ordbog og dermed et syntetisk sprog, som transformermodellen kan arbejde med (Savcisens et al. 2023a). I processen med at omdanne variable og værdier til tokens, er det vigtigt at skelne mellem to forskellige typer:

'Concept tokens' består af hhv. et præfix (eg. variablens navn/beskrivelse, fx *'mors_indkomst_'*), en værdi (eg. variablens værdi, fx 55), og evt. et suffix (eg. variablens måleenhed, fx. *'_percentil'* eller *'_perc'*). Et unikt token kan dermed eksempelvis være *'mors_indkomst_55_perc'* eller *'fars_udd_11.5_år'*. Dermed bliver det samlede antal af concept tokens, som senere feedes ind i modellen identisk med antallet af variable + de unikke værdier, som hver af dem kan antage. Dog vil de for irregulære hændelser ikke have en værdiangivelse, men blot angive fx *'skilsmisse'* eller *'mors_død'*. Nominale og kategoriske variable, som på forhånd antager tekstværdier vil også være concept tokens, der består af et præfix og en nominal eller kategorisk værdi, som fx *'køn_M'*, *'landsdel_Nordjyl'* eller *'fars_etn_dansk'*. Værd at bemærke her er desuden, at de missing-værdier (NA's) som fortsat er bevaret efter den indledende databehandlingsproces, i stedet for et værdi-suffix (fx *'55_perc'*) får 'missing' som tekst tilføjet foran præfixet (fx *'missing_mors_udd'*). Transformermodellen vil dermed kunne lære at forholde sig til eventuelle missing-værdier som havende særskilt betydning, særligt hvor missing-værdier optræder strukturelt betinget.

'Time tokens' behandles som særskilte tokens, der tilknyttes både konstanter, dynamiske variable og irregulære hændelser, men optræder som særskilte "ord" i den syntetiske ordbog, fx 'T02_', 'T05_'. Disse behandles særskilt og ikke som en integreret del af hvert concept token for at reducere antallet af unikke tokens og for at behandle tid som et særskilt embedding layer, der gør analysen langt mere robust. Konstanter, som ikke er knyttet til et særligt tidspunkt i livsforløbet, får et særligt 'time token', som kaldes 'GLOBAL' for netop at angive, at konstanter fungerer som en form for headline over hele sekvensen. I denne tidsserie er der desuden sammenfald mellem tid og personens alder, og dermed er T05 = 5 år. Dette gør, at vi ikke behøver danne en særskilt dynamisk variabel for alder.

Når time tokens og concept tokens kombineres, vil de dermed kunne se således ud:

'T05_mors_indkomst_55_perc'

eller

'GLOBAL_køn_K'

Ovenstående består altså af to forskellige tokens, som i transformermodellen behandles som forskellige inputlag, som samlet set giver semantisk mening og indeholder information om både et specifikt datapunkt i tidsserien, og dettes tidsmæssige indlejring i sekvensen.

Sekvensering henviser grundlæggende til, at alle concept og time tokens sammensættes til én lang tekststreng - eller sekvens. Til denne sekvens tilføjes endnu et særskilt time token 'START', som står i begyndelsen af alle sekvenser før de første concept tokens baseret på globale konstanter, således at transformermodellen kan genkende, at alle sekvenser starter samme sted. For eksempel på syntetisk genereret sekvens bestående af concept og time tokens. Se Figur 3:

Figur 3: Livsforløb repræsenteret som sekvens af tokens på baggrund af syntetisk data:

	ID	token_sequence
0	A	[START, køn_M, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk0, missing_diff_ældre, T00_mor_indk_57pct, T00_far_indk_34pct, T00_mor_udd_150mdr, T00_far_udd_108mdr, T00_mor_alder_29år, T00_far_alder_37år, T00_total_søsk_1, T01_mor_indk_60pct, T01_far_indk_32pct, T01_mor_udd_150mdr, T01_far_udd_108mdr, T01_mor_alder_30år, T01_far_alder_38år, T01_total_søsk_1, T02_mor_indk_54pct, T02_far_indk_37pct, T02_mor_udd_150mdr, T02_far_udd_108mdr, T02_mor_alder_31år, T02_far_alder_39år, T02_total_søsk_2, T03_mor_indk_61pct, T03_far_indk_37pct, T03_mor_udd_150mdr, T03_far_udd_108mdr, T03_mor_alder_32år, T03_far_alder_40år, T03_total_søsk_2, T04_mor_indk_60pct, T04_far_indk_38pct, T04_mor_udd_150mdr, T04_far_udd_108mdr, T04_mor_alder_33år, T04_far_alder_41år, T04_total_søsk_3, T05_mor_indk_59pct, T05_far_indk_40pct, T05_mor_udd_150mdr, T05_far_udd_108mdr, T05_mor_alder_34år, T05_far_alder_42år, T05_total_søsk_3, T06_mor_indk_60pct, T06_far_indk_40pct, T06_mor_udd_150mdr, T06_far_udd_108mdr, T06_mor_alder_35år, T06_far_alder_43år, T06_total_søsk_3, T06_skilsmisse, T07_mor_indk_55pct, T07_far_indk_40pct, T07_mor_udd_150mdr, T07_far_udd_108mdr, T07_mor_alder_36år, T07_far_alder_44år, T07_total_søsk_3, T08_mor_indk_50pct, T08_far_indk_38pct, T08_mor_udd_150mdr, T08_far_udd_108mdr, T08_mor_alder_37år, T08_far_alder_45år, T08_total_søsk_3, T09_mor_indk_44pct, T09_far_indk_36pct, T09_mor_udd_150mdr, T09_far_udd_108mdr, T09_mor_alder_38år, T09_far_alder_46år, T09_total_søsk_3, T10_mor_indk_41pct, T10_far_indk_39pct, T10_mor_udd_150mdr, T10_far_udd_108mdr, T10_mor_alder_39år, T10_far_alder_47år, T10_total_søsk_4, T11_mor_indk_42pct, T11_far_indk_44pct, T11_mor_udd_150mdr, T11_far_udd_108mdr, T11_mor_alder_40år, T11_far_alder_48år, T11_total_søsk_4, T12_mor_indk_40pct, T12_far_indk_43pct, T12_mor_udd_150mdr, T12_far_udd_108mdr, T12_mor_alder_41år, T12_far_alder_49år, T12_total_søsk_4, ...]
1	B	[START, køn_K, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk0, missing_diff_ældre, T00_mor_indk_29pct, T00_far_indk_58pct, T00_mor_udd_144mdr, T00_far_udd_120mdr, T00_mor_alder_26år, T00_far_alder_38år, T00_total_søsk_1, T01_mor_indk_29pct, T01_far_indk_59pct, T01_mor_udd_150mdr, T01_far_udd_120mdr, T01_mor_alder_27år, T01_far_alder_39år, T01_total_søsk_1, T02_mor_indk_30pct, T02_far_indk_60pct, T02_mor_udd_150mdr, T02_far_udd_120mdr, T02_mor_alder_28år, T02_far_alder_40år, T02_total_søsk_1, T02_skilsmisse, T03_mor_indk_26pct, T03_far_indk_63pct, T03_mor_udd_150mdr, T03_far_udd_120mdr, T03_mor_alder_29år, T03_far_alder_41år, T03_total_søsk_1, T04_mor_indk_24pct, T04_far_indk_69pct, T04_mor_udd_150mdr, T04_far_udd_120mdr, T04_mor_alder_30år, T04_far_alder_42år, T04_total_søsk_1, T05_mor_indk_28pct, T05_far_indk_71pct, T05_mor_udd_150mdr, T05_far_udd_120mdr, T05_mor_alder_31år, T05_far_alder_43år, T05_total_søsk_1, T06_mor_indk_35pct, T06_far_indk_73pct, T06_mor_udd_150mdr, T06_far_udd_120mdr, T06_mor_alder_32år, T06_far_alder_44år, T06_total_søsk_1, T07_mor_indk_35pct, T07_far_indk_74pct, T07_mor_udd_150mdr, T07_far_udd_120mdr, T07_mor_alder_33år, T07_far_alder_45år, T07_total_søsk_1, T08_mor_indk_32pct, T08_far_indk_78pct, T08_mor_udd_150mdr, T08_far_udd_120mdr, T08_mor_alder_34år, T08_far_alder_46år, T08_total_søsk_1, T09_mor_indk_38pct, T09_far_indk_80pct, T09_mor_udd_150mdr, T09_far_udd_120mdr, T09_mor_alder_35år, T09_far_alder_47år, T09_total_søsk_1, T10_mor_indk_41pct, T10_far_indk_83pct, T10_mor_udd_150mdr, T10_far_udd_120mdr, T10_mor_alder_36år, T10_far_alder_48år, T10_total_søsk_1, T11_mor_indk_33pct, T11_far_indk_89pct, T11_mor_udd_150mdr, T11_far_udd_120mdr, T11_mor_alder_37år, T11_far_alder_49år, T11_total_søsk_1, T12_mor_indk_33pct, T12_far_indk_90pct, T12_mor_udd_150mdr, T12_far_udd_120mdr, T12_mor_alder_38år, T12_far_alder_50år, T12_total_søsk_1, ...]

For at strømline input til transformermodellen og sikre, at alle inputsekvenser kan behandles parallelt i *batches,* som transformerarkitekturen kræver, skal tekstsekvenser have samme længde, dvs. bestå af samme antal tokens (Kundu et al 2024). I vores tilfælde har alle sekvenser det samme antal concept tokens baseret på konstanter og dynamiske variable, da vi for alle individer måler det samme antal konstanter og dynamiske variable hvert år. Dog gælder det for irregulære hændelser, som vi tidligere har gennemgået, at de kun optræder som concept token med dertilhørende time token i det år, hændelsen sker, og ellers ikke, og derfor vil den naturlige variation, der er i antallet af irregulære hændelser (fx flytning, ny søskende, forældres skilsmisse eller død) i barndommen fra individ til individ også betyde variation i antallet af concept tokens baseret herpå i den enkelte sekvens. Nogle individer oplever få eller måske slet ingen hændelser, mens andre oplever flere. For at ensrette antallet af tokens i sekvenserne benyttes derfor **padding**, hvor et særligt 'PAD-token' oprettes for at fylde tomme pladser ud i sekvenser, som er kortere end længste sekvens i datasættet. I traditionelle neurale netværk benyttes ofte *truncation* af data, som grundlæggende også har til formål at skabe lige store mængder

data, blot ved at forkorte eller formindske data til korteste fællesnævner – altså det modsatte af padding (Dwarampudi & Reddy 2019). Truncation vil dog fungere dårligt her, da vi i praksis ville være nødt til at skære tokens fra i hovedparten af sekvenserne. Pga. sekvensernes opbygning som repræsentationer af livsforløb ville konsekvensen være at forløbet brydes, eller at en antal af sekvensens første tokens fjernes, hvor sekvensen er længere end en fastsat baseline. I vores tilfælde ville det medføre at vigtige konstanter fjernes, da de altid står først i sekvenserne, og det kan vi ikke acceptere. Derfor benytter vi i stedet padding, så vi sikrer, at vi ikke kommer til at mangle vigtig information. Dog er det tilsvarende vigtigt, at PAD-tokens ikke tillægges analytisk vægt, hvorfor de maskeres i det endelige input til transformermodellen. Dette uddybes i næste afsnit.

7.3.4 LabelEncoder og attention mask

Da transformermodeller ikke kan "læse" tekstdata som reel tekst, skal alle tokens, både concept og time, omkonverteres fra tekst til numeriske værdier vha. en såkaldt **'LabelEncoder'** (Savcisens et al. 2023a). Her samles alle unikke tokens (fx. 'far_indk_76pct', 'opr_mor_vestl' og 'diffældsos_4.5år'). Hver unik kombination af variabel og værdi, altså hver unikke concept token, er derved blevet tildelt en tilsvarende unik heltalsværdi, et 'id', som samles i simple ordbøger, også kaldet 'mappings', som benyttes i konvertering af concept token til talværdi ('concept2id') og returkonvertering af talværdi til token ('id2concept'). Disse mappings kan gemmes for senere at kunne konvertere de numeriske værdier retur til tekst i forbindelse med fortolkning af modeloutput. I vores konkrete case har vi, som tidligere forklaret, adskilt concept og time tokens for at reducere antallet af unikke tokens, samt for at benytte tid som et separat embedding layer, hvorfor vi udover mappings for concept tokens også har dannet mappings for time tokens ('time2id' og 'id2time') vha. LabelEncoder. Se Figur 4 for eksempel på konvertering af concept tokens til heltalsværdier.

Figur 4: Konvertering fra tekst til heltalsværdi vha LabelEncoder*

	ID	token_sequence
0	A	[START, køn_M, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk0, missing_diff_ældre, T00_mor_indk_57pct, T00_far_indk_34pct, T00_mor_udd_150mdr, T00_far_udd_108mdr, T00_mor_alder_29år, T00_far_alder_37år, T00_total_sø...
1	B	[START, køn_K, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk0, missing_diff_ældre, T00_mor_indk_29pct, T00_far_indk_58pct, T00_mor_udd_144mdr, T00_far_udd_120mdr, T00_mor_alder_26år, T00_far_alder_38år, T00_total_sø...
2	C	[START, køn_M, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk1, diff_ældre_3.7ÅR, T00_mor_indk_69pct, T00_far_indk_39pct, T00_mor_udd_114mdr, T00_far_udd_150mdr, T00_mor_alder_31år, T00_far_alder_33år, T00_total_søsk...
3	D	[START, køn_M, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk1, diff_ældre_2.0ÅR, T00_mor_indk_52pct, T00_far_indk_54pct, T00_mor_udd_96mdr, T00_far_udd_150mdr, T00_mor_alder_30år, T00_far_alder_29år, T00_total_søsk_...
4	E	[START, køn_M, mor_etn_dansk, far_etn_dansk, mor_opr_vestlig, far_opr_vestlig, n_ældre_søsk2, diff_ældre_1.7ÅR, T00_mor_indk_57pct, T00_far_indk_39pct, T00_mor_udd_126mdr, T00_far_udd_126mdr, T00_mor_alder_22år, T00_far_alder_36år, T00_total_søsk...

	ID	padded_ids
0	A	[0, 1689, 1691, 1684, 1694, 1687, 1695, 1690, 66, 17, 80, 34, 49, 11, 83, 155, 102, 167, 122, 139, 97, 170, 238, 192, 254, 208, 224, 185, 258, 332, 277, 345, 295, 315, 272, 350, 420, 371, 432, 384, 404, 364, 438, 512, 460, 526, 475, 495, 451, 532...
1	B	[0, 1688, 1691, 1684, 1694, 1687, 1695, 1690, 55, 27, 79, 36, 47, 12, 83, 143, 115, 167, 124, 137, 98, 170, 229, 203, 254, 209, 222, 186, 257, 256, 320, 289, 345, 296, 313, 273, 349, 409, 380, 432, 385, 402, 365, 436, 500, 472, 526, 476, 493, 452...
2	C	[0, 1689, 1691, 1684, 1694, 1687, 1696, 1681, 71, 20, 74, 41, 51, 7, 84, 54, 132, 105, 162, 129, 131, 93, 172, 217, 193, 249, 214, 216, 181, 259, 256, 308, 279, 340, 300, 307, 268, 351, 397, 369, 427, 389, 396, 360, 438, 488, 456, 520, 480, 487, ...
3	D	[0, 1689, 1691, 1684, 1694, 1687, 1696, 1673, 63, 26, 81, 41, 50, 4, 84, 149, 113, 168, 129, 140, 90, 171, 233, 202, 255, 214, 225, 178, 258, 323, 285, 347, 300, 316, 265, 350, 411, 377, 434, 389, 405, 357, 437, 504, 467, 528, 480, 496, 444, 532,...
4	E	[0, 1689, 1691, 1684, 1694, 1687, 1697, 1672, 66, 20, 76, 37, 43, 10, 86, 154, 105, 164, 125, 133, 96, 173, 245, 194, 251, 210, 218, 184, 260, 329, 282, 342, 297, 309, 271, 352, 417, 375, 429, 386, 398, 363, 439, 510, 463, 522, 477, 489, 450, 533...

Som forklaret tidligere benyttes padding vha. PAD-tokens til at sikre, at alle inputsekvenser er lige lange, da transformerarkitekturen kræver dette. Dog har vi brug for at sikre, at disse PAD-tokens ikke tillægges særskilt værdi i modelinputtet, hvorfor vi anvender **'attention mask'** til at maskere disse paddings, så de ikke tilfører unødig 'støj' til modellen (Wu et al 2023). Ved maskering med attention mask er der grundlæggende tale om en underlæggende dummy-kodning til hver enkelt token, som vægter observationen (1), eller maskerer den (0), hvis den ikke bør indgå som reel observation i transformermodellen. Normalt vil man altid maskere paddede tokens, da disse i sig selv ikke bærer analysérbar værdi, men blot er tilføjet for at standardisere længden på sekvenserne, som forklaret ovenfor, hvilket vi også vælger at gøre her. Dernæst kan det overvejes, om man ønsker at maskere missing-observationer generelt eller for udvalgte variable. Dette kan være en fordel, hvis man arbejder med et datasæt med større mængder missing-værdier, som overvejende er tilfældigt fordelt grundet fx manglende observationer i rådata fra Danmarks Statistik. Det kan dog modsat også være en fordel *ikke* at maskere missing-værdier, hvis man har en formodning om, at missing-værdierne i sig selv bærer strukturel betydning i opbygningen af data. Konkret har vi i databehandlingsprocessen på forhånd fravalgt at benytte råvariable, som indeholdt mange missing-værdier. Samtidig har vi opsat som kriterie, at vores endelige populationssample skal levere værdier i alle primære variable, medmindre der har været en strukturel grund til at visse variable har været missing – fx en forælders dødsfald i et givent år, hvor der ikke kan leveres værdier til de variable, som er relateret til denne forælder i de resterende år i tidsserien. Derudover har vi opstillet som kriterie, at selve populationen maksimalt må mangle informationer i primære variable i tre ud af 18 år, hvilket fx kan skyldes, at familien midlertidigt har

opholdt sig i udlandet. Alle potentielle personer født i 1980-1982, som ikke har kunnet levere primære informationer i minimum 15 år, er derved frasorteret i populationen. Sidst men ikke mindst har vi en række variable som indeholder missing-værdier, der er hundrede procent strukturelt betingede. Dette er fx tilfældet for den udskudte konstant for aldersdifference til nærmeste yngre søskende (diffyngsos), som antager missing-værdier i de år, der går indtil en ny søskende bliver født, og for konstanten for aldersdifference til nærmeste ældre søskende (diffældsos), som antager missing-værdier i hele tidsserien, hvis den pågældende person ikke har en ældre søskende. Derudover har vi missing-værdier i alle observationer i område-relaterede variable i årene 1980-1984 grundet manglende tilgængelighed af befolkningsregistre for disse år. Dette resulterer i, at vi mangler områdeværdier for alle født i 1980 i de første fem år, alle født i 1981 i de første fire år, og alle født i 1982 i de første tre år. Dermed har vi i overvejende grad missing-værdier i sekvenserne, som er meningsbærende i sig selv, og som giver transformermodellen mulighed for at lære disse strukturer at kende. Derfor har vi valgt at undlade maskering af missing tokens.

7.3.5 Embedding layers og positional coding

Tranformerbaserede modeller kræver, at inputsekvenserne, herunder tokens som repræsentationer for datapunkter og hændelser i et livsforløb og disses placering i tid, repræsenteres som numeriske vektorer i et fælles meningsbærende rum. For at opnå dette benytter vi to centrale teknikker: **embedding layers**, som konverterer token-id's til kontinuerte vektorer, og **positonal coding**, som indlejrer information om rækkefølgen og timingen af datapunkter og hændelser i sekvensen. I modsætning til traditionelle statistiske modeller, hvor hver variabel ofte behandles isoleret, tillader embedding-laget, at betydning af hver token afspejles i en numerisk repræsentation, som kan læres og justeres under modeltræning. Det betyder, at modellen selv lærer, hvilke hændelser, der ligner hinanden eller har tilsvarende konsekvenser i et livsforløb, baseret på deres rolle i træningsdataene. Denne tilgang bygger på den transformerbaserede sprogmodellinger, hvor ord indlejres i et fælles semantisk rum (Devlin et al. 2019; Vaswani et. al 2017).

I forberedelsen til vores implementering af en transformermodel på tekstsekvenser benyttes separate embedding layers til at håndtere hhv. concept tokens (som repræsentanter for variabel+værdi) og time tokens (som repræsentanter for tidsmæssig markør). Hvert unikke token for hhv. concept og time tokens er som tidligere beskrevet blevet mappet til et heltals-id via en LabelEncoder-struktur, hvorefter de omdannes til vektorer af samme dimension. Denne transformation sker gennem learnable embeddings, dvs. parametre, der optimeres under modeltræning (Savcisens et al. 2023a). Men transformerarkitekturen har ikke i sig selv en indbygget tidsforståelse eller forståelse af rækkefølge, i modsætning til RNN- eller LSTM-modeller, der automatisk ved, om én hændelse kommer før en anden, som tidligere beskrevet i afsnit 7.1. Transformer-modellen ser derfor hele sekvensen som en "pose" af blandede tokens uden at kunne forstå rækkefølgen i sig selv. Derfor er det nødvendigt, at vi eksplicit angiver rækkefølge og position af hvert enkelt token i hver sekvens (livsforløb)

via en særskilt positional coding (Vaswani et al 2017). Til dette benytter vi *learnable positional embeddings,* som er en fleksibel tilgang, hvor modellen selv lærer, hvordan positioner skal vægtes. Denne tilgang er særligt relevant, når tidsdimensionen ikke blot er lineær, men strukturelt meningsbærende (fx om en hændelse sker i barndommen eller ungdommen). Positional embeddings giver dermed modellen evnen til at lære mønstre som:

- *"Nysøskende-hændelse forekommer oftere tidligt end sent i livsforløbet"*
- *"Når en hændelse relateret til en forælders død forekommer tidligt i livsforløbet har det en anden effekt end når den forekommer sent.*

Ved at kombinere concept og time embeddings opnår vi en vektoriseret repræsentation af hvert datapunkt, som både indfanger *indholdet* og *timingen*. I praksis betyder det, at hver positional coding ikke kun er et tal for en absolut position i sekvensen, men en vektor som relativ størrelse i et matrix, hvor positionen for hvert token "måles" i et flerdimensionelt rum, som "afstande" til andre tokens i sekvensen, som alt i alt siger noget om dets position. Den præcise embedding-dimension specificeres i afsnit 7.4 om modelarkitektur. Disse repræsentationer kombineres og sendes videre til transformerens encoderlag, hvor de indgår i self-attention-mekanismen. Dette gør det muligt for modellen at identificere mønstre på tværs af livsforløb, fx hvordan visse typer af hændelser, når de indtræffer i bestemte aldre eller på bestemte tidspunkter i forhold til andre hændelser, kan øge sandsynligheden for bestemte livsudfald (Savcisens et al. 2023b, Liu et al. 2019 og Tenney et al. 2019).

7.4 Modelarkitektur, tensors og valg af hyperparametre

Transformerbaserede modeller arbejder med *tensors*, som er flerdimensionelle datastrukturer, eller matrixes, der kan rumme store mængder information i form af tal. En tensor kan bedst forstås som en udvidelse af kendte begreber som vektorer (én dimension) og matricer (to dimensioner), og bruges til at repræsentere inputdata, modelparametre og output gennem hele den neurale netværksproces (Panagakis et al 2021).

Modelarkitekturen beskriver selve strukturen af den transformer, der anvendes – dvs. hvordan modellen er bygget op af lag, hvilke komponenter, den består af, hvilke operationer, den udfører, og hvordan information flyder gennem netværket. Arkitekturen er med andre ord en bygning, som vi tildeler et bestemt antal etager, rum, trapper, døre og vinduer; men bygningen er stadig tom, så længe vi endnu ikke har sendt inputdata ind igennem dem. Når vi gør det, vil bygningens arkitektur have betydning for, hvordan inputdata behandles og læres af modellen, således at vi kan få et brugbart output ud af bygningen igen. I dette speciale anvendes som tidligere beskrevet en encoder-baseret transformer, som er særlig velegnet til at analyser sekventielle hændelser i livsforløbsdata.

Modellens adfærd og præstation styres i høj grad af en række *hyperparametre* – det vil sige foruddefinerede værdier, som ikke læres af modellen selv, men vælges manuelt af os. Hyperparametre

kontrollerer bl.a. modellens kapacitet, læringshastighed og hvor godt en model generaliserer til ukendt data (Arnold et al 2024). Følgende *hyperparametre* er anvendt i den aktuelle model:

- **Embedding dimension = 128**
  Hver hændelse og hvert tidstrin i sekvensen omdannes til en 128-dimensionel vektor, som tillader modellen af indfange nuancerede mønstre i data uden at overkomplicere repræsentationen og detaljegraden.
- **Antal transformer-lag = 2**
  Bestemmer hvor mange gange modellen transformerer og relaterer information gennem sin attentionmekanisme. Færre lag reducerer kompleksitet og risiko for overfitting.
- **Antal attention heads = 4**
  Tillader modellen at fokusere på forskellige typer relationer i data samtidigt. Hvert 'head' lærer sit eget mønster på tværs af sekvensen.
- **Dimension på feedforward-netværket = 512**
  Størrelsen på det indre netværk, som bearbejder information efter attention-laget. Det kontrollerer, hvor meget transformationskapacitet modellen har i hvert lag.
- **Dropout rate = 0.1**
  En metode til at forhindre overfitting ved tilfældigt at "slukke" enkelte forbindelser under træning, hvilket gør modellen mere robust.
- **Maksimal sekvenslængde = 820**
  Den længste sekvens modellen skal kunne håndtere (dvs. antal tokens i sekvens). Alle sekvenser er på forhånd forlænget med padding til denne længde.

De ovenstående valg er truffet med henblik på at skabe en model, der balancerer generaliserbarhed og følsomhed med træningseffektivitet og tidsforbrug under træning.

7.5 Datasplit, valideringsstrategi og subsamples

Før træning af modellen opdeles data i tre separate delmængder: træningsdata, valideringsdata og testdata. Formålet med denne opdeling er at sikre, at modellens præstation evalueres på data, den ikke har set før, og at man kan justere modellens hyperparametre uden risiko for overfitting på evalueringsgrundlaget. Til denne analyse er datasættet opdelt som følger:

- **Træningsdata (70%)** bruges til at opdatere modellens vægte under læring.
- **Valideringsdata (15%)** anvendes løbende under træningsfasen til at overvåge modellens generaliseringsevne og kontrollere for overfitting. Modellen gemmes, når præstationen (målt ved ROC AUC, se afsnit 7) forbedres.
- **Testdata (15%)** holdes fuldstændig adskilt under hele træningen og anvendes kun én gang til endelig modelvurdering, efter træning er afsluttet.

Datasplittet foretages stratificeret, dvs. med bevarelse af andelen af positive og negative udfald (y=1 og y=0) i alle tre datasæt. Dette er vigtigt, da y=1 er sjælden og et ikke-stratificeret split kunne resultere i test- eller valideringssæt med få eller ingen positive tilfælde. Splittet gennemføres på individniveau for at sikre, at hver persons sekvens kun optræder i én af de tre delmængder.

Det fulde datasæt indeholder, som vist i Tabel 4, 140.708 fulde sekvenser; en datamængde, hvor det med det serverkapacitetssetup, vi har tilgængeligt i DST's miljø, samt de hyperparametre, vi har valgt at benytte i vores arkitektur, vil kræve ca. 3-4 uger at træne hver model, for hvert af de seks eliteudfald, vi arbejder med. Da en træningstid i den størrelsesorden langt overskrider den tid, vi har til rådighed til dette speciale, vælger vi derfor at arbejde med *subsamples, dvs.* deldatasæt for hvert udfald til de første modeltræninger. Vi subsampler stratificeret for at bevare den oprindelige fordeling af y=1 og y=0, og vælger for hvert udfald 10% af den oprindelige datamængde, dvs. 14.070 sekvenser. Se Tabel 5 for stratificeret subsamples for hvert eliteudfald, samt antal sekvenser i hver deldatamængde efter datasplit.

Tabel 5: Oversigt over y=1 i subsamples på 14.070 sekvenser før og efter datasplit

Udfald (y)	Andel y=1	Antal y=1	Antal y=1 Train+val data (85%)	Antal y=1 Testdata (15%)
Top5	0.0682	960	816	144
Storbyelite	0.0716	1007	856	151
Uddtopind	0.0541	761	647	114
Omtop5	0.0853	1200	1020	180
Led	0.0398	560	476	84
Selv	0.0104	146	124	22

Allerede her ser vi, at med stratificerede subsamples på 10% af oprindelig data, får vi, grundet klasseskævheden mellem y=1 og y=0, hvor y=1 i forvejen er i kraftig minoritet, en sårbar y=1 gruppe i en ML- og transformerkontekst, hvor modellen er afhængig af en vis datamængde for at kunne lære mønstre i data godt nok til at kunne opnå en god generaliseringsevne. Særligt for udfaldene ledere og selvstændige har vi meget få y=1. Vi arbejder dog videre med disse subsamples i første omgang for prototyping, og benytter *class weights* til at håndtere klasseubalancen, som forklaret i næste afsnit.

7.6 Træningsstrategi og class weights

Ud over selve modelarkitekturen har træningsstrategien en afgørende betydning for, hvordan en ML-model som en transformer lærer at genkende mønstre i data. Træningsstrategien omfatter en række træningsspecifikke hyperparametre, som ikke ændrer på modellens struktur, men som i høj grad styrer læringsprocessens effektivitet, stabilitet og følsomhed over for klasseskævheden i datasættet, også kaldet *class imbalance.* Til denne analyse er følgende centrale hyperparametre valgt:

- **Optimeringsalgoritme: AdamW**
  Der anvendes optimeringsalgoritmen AdamW (Loshchilov & Hutter 2019). AdamW er særligt velegnet til transformerarkitekturer, hvor der er behov for stabililitet ved mange parametre, som kan bidrage til mere præcise og generaliserbare modeller.
- **Learning rate = 1e-4**
  Learning rate angiver, hvor store 'skridt' modellen tager under parameteroptimering. En lav værdi, som er valgt her, sikrer en gradvis og stabil læring, og forebygger, at modellen 'springer over' vigtige lokale mønstre (Goodfellow, Bengio & Courville 2016).
- **Batch size = 256**
  Antallet af sekvenser, som behandles samtidigt under én iteration af træningen – større batches, som her 256, giver mere stabil træning, men er dog mere CPU-krævende, mens en for høj batch size kan resultere i høje loss værdier (Smith et al 2018). For små batches kan være risikofyldt i træning af data med stor ubalance mellem klasserne, som netop er tilfældet i vores studie, hvor y=1 for de udvalgte elite-udfald udgør mellem 1 og 8 procent af det samlede antal sekvenser. Da batches udvælges randomiseret, kan en for lille batch size ved high class imbalance resultere i batches med meget få eller ingen y=1, hvilket umuliggør at modellen kan lære mønstre i y=1-sekvenser at kende.
- **Class weights (weighted loss)**
  Til håndtering af en ubalanceret binær klassifikation, hvor y=1 i vores tilfælde er minoritetsklassen, og samtidig den, vi primært er interesserede i at prædiktere, anvendes class weights i *loss-funktionen*. Loss-funktionen er den matematiske funktion, som måler forskellen mellem modellens forudsigelser og de faktiske værdier, og den fungerer dermed som en styringsmekanisme, der guider den løbende opdatering af modellens indre parametre under træning. Ved at tildele sjældne positive udfald (y=1) en højere vægt i loss-beregningen opnås en bedre balance, hvor modellen ikke domineres af majoritetsklasse (y=0) (Buda, Maki & Mazurowski 2018).
  Ofte vil man i forbindelse med analyser, hvor klassiske neurale netværk benyttes til almindelig tabeldata uden tidsdimension, tilstræbe at skabe balancerede (50/50) klasser vha. upsampling af minoritetsklassen med syntetisk data. Denne strategi til at skabe balance mellem klasserne kan dog ikke benyttes her, da der ikke for nuværende findes en metode eller funktion, der kan

skabe et meningsfuldt og realistisk syntetisk sample af livssekvenser som strenge af teksttokens, der er indlejret i både konceptuelle, tidslige og positionelle dimensioner på samme tid. Derfor benytter vi i stedet class weights og weighted loss.

- **Epochs og Early stopping**
  Modeltræningen kører over flere epoker, hvor en epoke er én træning af alle sekvenser i trænings- og valideringsdata. Efter hver epoke starter træningen således forfra og er modelarkitektur, hyperparametre og vægtning indstillet korrekt, og træningspotentialet til stede i det valgte data, bør modellens præstation forbedres for hver epoke. Dog kan træning over for mange epoker også føre til såkaldt *overfitting,* hvor modellen lærer mønstre i trænings- og valideringsdata *for* godt, og dermed mister generaliseringsevne til ukendt data. Derfor monitoreres modellens foreløbige præstation samt loss-værdier løbende for hver epoke. Forbedres præstationen ikke væsentligt over flere epoker, kan det derfor være fordelagtigt at benytte såkaldt *early stopping*, så træningen automatisk stoppes, når modellen har nået et optimalt niveau. Early stopping sættes i vores tilfælde til 5; er modellens præstation ikke forbedret fem epoker i træk, stoppes den, og bedste model indtil da gemmes og benyttes i den efterfølgende evalueringsfase.

7.7 Evaluering af foreløbige resultater

De første træninger på 10% subsamples af data for hvert af vores seks elite-udfald viser, at modellerne umiddelbart præsterer godt, og at de har en god evne til at lære af inputsekvenserne. Vi opnår følgende bedste træningsresultater for hvert udfald. Se Tabel 6.

Tabel 6: ROC AUC-scores for første træninger

Udfald (y)	Bedste ROC AUC
Storbyeli	0.7480
Top5	0.7339
Uddtopind	0.7152
Omtop5	0.6908
Selv	0.6643
Led	0.6115

For de bedste modeller ser vi en ROC AUC-score et godt stykke over 0.7, hvilket er en ganske udemærket præstation for en binær klassifikationsmodel i en samfundsvidenskabelig kontekst. Dog ser vi også bekymrende høje værdier på for *train loss*, som er modellens fejlrate på data fra træningsdatasættet og *validation loss,* som er modellens fejlrate på det separate valideringsdatasæt. En god model bør have lave loss-værdier og værdierne bør være stabile eller falde gradvist under træning. Særligt train loss ligger på ekstremt høje værdier på mellem 40.0 og 120.0, hvor loss værdier normalt bør ligge et stykke under 1.0. En kombination af god præstation under træning (ROC AUC) og høj train loss kan indikere, at vores valg

af class weights og weighted loss i kombination med relativt små datasæt og dermed også meget få y=1-sekvenser ikke er nok til at balancere skævheden mellem klasserne tilstrækkeligt. Det skyldes at fejl i klassificering af den underrepræsenterede klasse straffes ekstremt hårdt i vores tilfælde, hvor der er så få y=1 tilstede i data. Dog har vi stadig en relativt god generaliseringsevne.

Et nærmere kig på de to bedste modellers perfomance i den endelige evalueringsfase på testdata bekræfter da også, at vores problem med klasseskævhed ikke er tilstrækkeligt løst med class weights. Se Tabel 7 og Tabel 8.

Tabel 7: Precision, recall og F1-scores for 'storbyelite'

Storbyeli	Precision	Recall	F1-score
y = 0	0.959	0.748	0.841
y = 1	0.151	0.583	0.240

Tabel 8: Precision, recall og F1-scores for 'top5'

Storbyeli	Precision	Recall	F1-score
y = 0	0.969	0.587	0.731
y = 1	0.116	0.743	0.201

Defintioner:

ROC AUC *er en evalueringsmetrik, der bruges til at vurdere modellens evne til at skelne mellem to klasser ved at rangere predicted probabilities for hver sekvens*

Precision *måler andelen af modellens forudsigelser for hhv. y=0 og y=1, som faktisk er korrekte*

Recall *måler andelen af de faktiske positive eller faktiske negative tilfælde, som modellen korrekt har identificeret*

F1-score *er det harmoniske gennemsnit af precision og recall og skaber en samlet balanceret score.*

Som det ses i ovenstående tabeller, får vi en meget høj precision på y=0 i begge tilfælde i kombination med en meget lav precision på y=1. Det bekræfter formodningen om, at class weights og weighted loss ikke er en tilstrækkelig vægtningsmekaniske i vores tilfælde, hvor vi arbejder med *extreme class imbalance.* På trods af at denne vægtningsmetode bør imødegå klasseskævheden, ser vi at modellen stadig er kraftigt biased mod at gætte på y=0 i de fleste tilfælde, da det er det mest "sikre" bud. Og da y=0 udgør langt den største gruppe, vil F1-scoren, som normalt er et godt parameter for modellens kvalitet i evalueringsfasen også blive relativt høj for y=0 og meget lav for y=1. Da vi i vores konkrete problemstilling primært er interesseret i en god prædiktionsevne på minoritetsklassen, dvs. y=1 – vores eliter – er vi derfor nødt til at genoverveje træningsstrategien, samt hvilke justeringer, der vil være hensigtsmæssige at implementere før næste træningsforsøg.

7.8 Metodejustering – downsampling, dynamisk vægtning og focal loss

Ved justering af metode vælger vi at bibeholde selve transformerarkitekturen med de allerede fastsatte hyperparametre, da det ikke umiddelbart er en justering af disse, der kan afhjælpe klasseubalancen og

dermed forbedre vores prædiktion på y=1. Vi vælger i stedet at justere på vores subsamples og datasplit i forberedelsen til modeltræningerne, mens vi i selve træningsstrategien vælger at erstatte class weights og weighted loss med dynamisk vægtning og focal loss.

I forbindelse med datasplit, har vi erkendt, at et subsample på 10% af de fulde sekvenser ikke er tilstrækkeligt, når vi i forvejen arbejder med en ekstrem klasseubalance, hvor der er få observationer i minoritetsklassen. Vi vælger derfor en anden subsamplings- og datasplitstrategi, hvor vi først og fremmest vælger at bibeholde *alle* i y=1-klassen fra det fulde datasæt med 140.708 sekvenser, og i stedet ***downsample*** y=0-klassen for at ændre på den interne fordeling mellem de to klasser. Derefter vælges en train-val-test-split-fordeling, som fortsat følger 70-15-15-princippet, men som dog er sammensat på den måde, at train- og val-data fortsat er stratificeret på y som i det oprindelige data, mens testdata upsamples på y=1, således at fordelingen på y i testdata bliver 50/50. Dette sikrer, at træning og validering forsat sker på en realistisk måde, hvor modellen lærer mønstre, som afspejler virkeligheden og udfordringen ved skæv klassebalance, mens et balanceret testdatasæt giver en mere retvisende evaluering af modellens performance på minoritetsklassen og på at identificere mere sjældne positive udfald, hvilket gør det lettere at vurdere modellens faktiske præstation (Johnson & Khoshgoftaar 2019; Chen et al 2024; Buda, Maki & Mazurowski 2018). I det nye datasplit sikres det selvfølgelig fortsat, at ingen observationer fra train- og val-data går igen i testdata. Samtidig er de nye subsamples for de fleste elite-udfald væsentlig større end i første træningsrunde. Da antallet af y=1 i hver udfaldsgruppe således er definerende for det endelige antal sekvenser i hvert nye subsample, er disse denne gang af forskellig størrelse, og ikke som før en fast procentsats af det samlede fulde datasæt. Da subsamples som nævnt tidligere også er væsentligt større denne gang i fem ud af seks tilfælde, accepterer vi derfor også en markant længere træningstid for hver model. Se Tabel 9 for sammensætning af subsamples efter justering af datasplit. Heraf fremgår det, at det er muligt at opretholde en ny fordeling mellem klasserne på 80/20 for alle subsamples med undtagelse af 'selvstændige' (selv), hvor y=1-klassen er så snæver, at y=1 må udgøre en lidt mindre andel, for ikke at ende med for lille et subsample i forhold til før.

Tabel 9: Sammensætning af subsamples efter justering af datasplit

Udfald (y)	Andel y=1 før split	Antal y=1 ud af 140,708 sekvenser	Antal sekvenser i subsample efter downsampling af y=0	Andel y=1 i subsample	Antal y=1 i testdata efter split (50/50)
Top5	0.0682	9596	47,990	0.2	2159
Storbyelite	0.0716	10,075	50,345	0.2	2265

Uddtopind	0.0541	7612	38,045	0.2	1712
Omtop5	0.0853	12,002	60,030	0.2	2701
Led	0.0398	5600	27,985	0.2	1259
Selv	0.0104	1463	12,453	0.1176	560

Derudover har vi for at imødegå problematikken med, at modellen bliver rigtig god til at lære at forudsige majoritetsklassen og knap så god til at forudsige minoritetsklassen korrekt, valgt at supplere - og på sigt erstatte - de traditionelle class weights med en mere avanceret og differentieret tilgang, nemlig en kombination af ***dynamisk instansvægtning*** *og* ***focal loss***, som styrker modellens evne til at identificere sjældne og analytisk vigtige observationer (her y=1), uden at læringen domineres af majoritetsklassen.

I stedet for at tildele faste vægte til hver klasse, som i class weighting, benytter vi dynamisk instansvægtning, hvor hver enkelt observation vægtes forskelligt afhængigt af, hvor svær den er for modellen at lære. Det betyder, at sekvenser, som modellen konsekvent har svært ved at forudsige korrekt, vægtes højere, mens lette observationer vægtes lavere, så de ikke dominerer læringsprocessen. Det betyder desuden, at læringsprocessen tilpasses løbende fra epoke til epoke, så modellen fokuserer sine ressourcer på de sekvenser, der bidrager mest til en forbedret performance i ubalancerede situationer, hvilket også kan bidrage til at reducere overfitting (Ren et al 2018).

Focal loss er en videreudvikling af klassisk *binary cross-entropy loss*, som tilføjer en såkaldt *modulationsfaktor*. Denne nedtoner bidraget fra let klassificerbare sekvenser (typisk y=0) og opprioriterer fejlklassificerede eller svære observationer (typisk y=1). Konkret betyder det, at modellen i højere grad 'straffes', når den tager fejl på de sjældne og vigtige udfald, mens trivielle y=0-forudsigelser i mindre grad påvirker læringen. På den måde fremmer focal loss modellens følsomhed over for minoritetsudfald, og har i flere studier vist bedre performance end både class weights og downsampling af majoritetsklassen alene (Lin et al 2017).

Med disse revurderede og justerede metodiske tiltag køres modeltræninger på alle seks udfaldsgrupper igen. De endelige resultater heraf følger i næste afsnit.

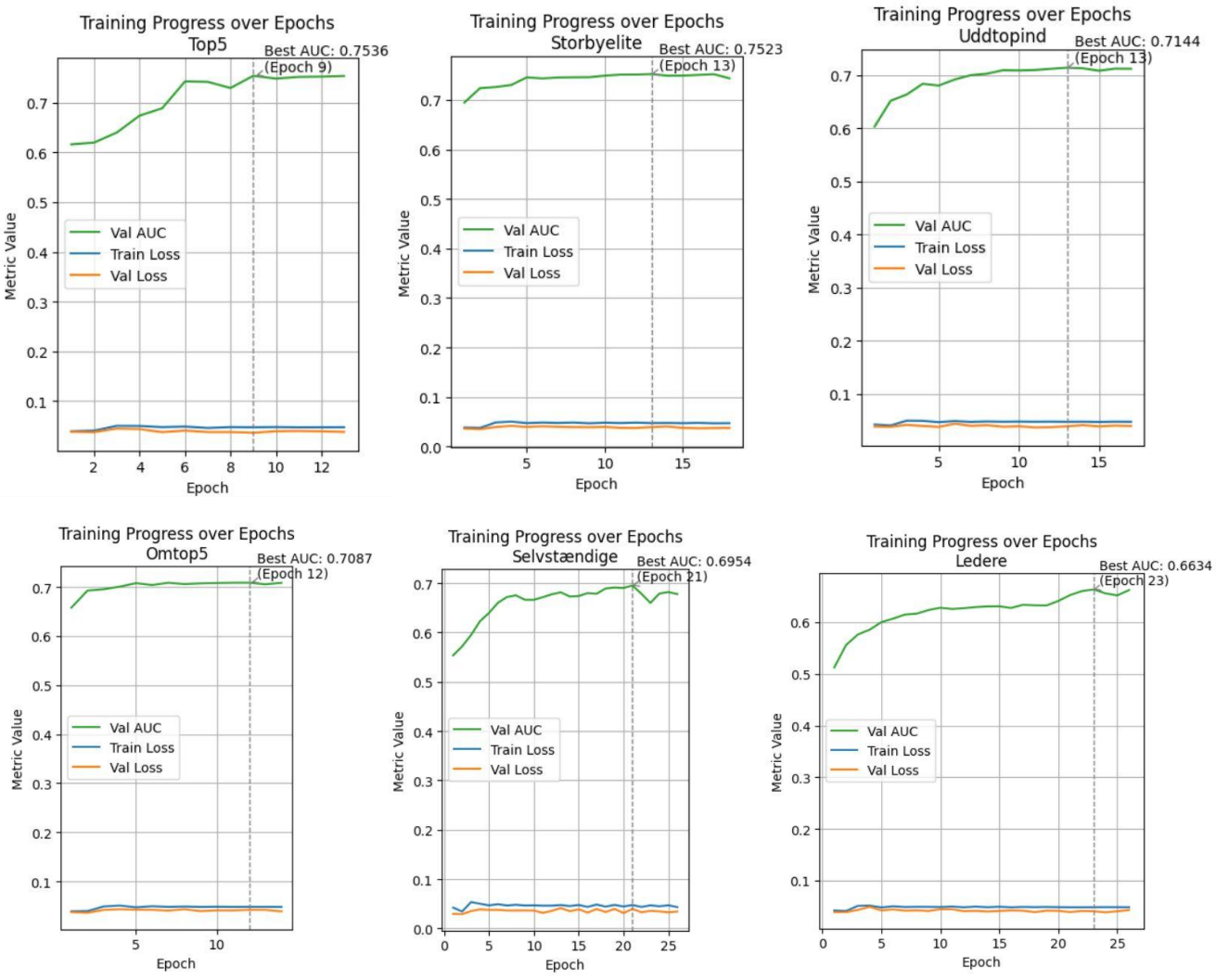
8 Resultater

Efter de netop gennemgåede justeringer gennemfører vi modeltræninger og efterfølgende evaluering og testfase for hver af de seks elite-udfald. Justeringerne forbedrer generelt modellernes overordnede

evne til at lære mønstre i data og generalisere disse til ukendt data. Desuden forbedres prædiktionerne på minoritetsgruppen y=1 i væsentlig grad. I det følgende gennemgås en række relevante metrikker fra disse faser, som samlet set danner grundlag for vurdering af modellernes præstationer, prædiktionssucces og generelle kvalitet. Derudover sammenlignes de seks modeller, hvilket åbner for en bredere sociologisk fortolkning og diskussion af forskellene i modellernes prædiktioner på y=1, og anvendeligheden ved at justere på det såkaldte 'threshold'.

8.1 Modeltræning

Figur 5: Modeltræningsprogression



Som det fremgår af alle seks modeltræningsprocesser vist i Figur 5, er train loss og val loss lave med værdier mellem 0.03 og 0.05, samt mere eller mindre parallelt forløbende under alle epoker – se afsnit 7.7 'Evaluering af foreløbige Resultater' for definitioner af hhv. train loss og val loss. De lave, parallelle værdier indikerer, at træningsprocessen har været succesfuld og at modellen evner at generalisere til

ukendt data under valideringsfasen uden overfitting – se afsnit 7.6 'Træningsstrategi og class weights' for definition heraf. Det tyder således på, at vi har opnået ganske optimal progression under træningerne af modellerne på hvert enkelte elite-udfald. Havde validation loss og train loss derimod bevæget sig markant i hver sin retning over flere epoker i træk, altså været tydeligt henholdsvis stigende og faldende hver især, kan det indikere overfitting.

ROC AUC er et mål for modellens evne til at rangordne klasserne korrekt på baggrund af hver enkelte sekvens' tildelte probabilityscore, eller sandsynlighed, for at være y=1, og fortæller om modellens generelle præstation uafhængigt af fastsatte thresholds for prædiktion af de to klasser. Generelt på tværs af de seks træninger er der en god løbende progression, hvor ROC AUC-værdien stiger. Efter en række epoker synes modellens evne til at lære forskellen mellem positive og negative udfald ikke at forbedres, hvorfor træningen automatisk stoppes og den bedste epoke udvælges som endeligt bedste model. ROC AUC-værdierne spænder mellem 0.66-0.75, hvilket er rimelige værdier ift. at få relativt succesfulde prædiktioner på baggrund af sociale data. For yderligere nuancering af værdierne bag ovenstående plots. Se Bilag 2 for *train progress tables*.

8.2 Prædiktionsresultater for alle udfald (y1-y6)

I det følgende gennemgås resultaterne af prædiktionerne i evalueringsfasen på testdata for alle seks eliteudfald. De præsenteres i rækkefølge efter præstation med bedste prædiktion først.

I udarbejdelsen af de endelige prædiktioner har vi for hver af de seks modeller evalueret, hvor thresholds skal ligge for at opnå de mest succesfulde og samtidig stadig meningsfulde prædiktioner, særligt på vores elitegrupper (y=1). Som standard vil threshold være sat til 0.5, således at alle sekvenser med et sandsynlighedsoutput over 0.5 prædikteres til at være y=1. Threshold kan dog justeres alt efter prædiktionsproblemstilling; vil vi have modellen til at gætte y=1 mere konservativt (altså være mere sikker og præcis) eller vil vi have den til at gætte mere liberalt (og dermed også kalde flere falske y=1). Threshold er dermed et mål, vi kan bruge til at prioritere enten *precision* (modellen skal være meget sikker på de y=1 der kaldes) eller *recall* (modellen skal kalde så mange korrekte y=1 som muligt).

Defintioner:

ROC AUC *er en evalueringsmetrik, der bruges til at vurdere modellens evne til at skelne mellem to klasser ved at rangere predicted probabilities for hver sekvens*

Precision *måler andelen af modellens forudsigelser for hhv. y=0 og y=1, som faktisk er korrekte*

Recall *måler andelen af de faktiske positive eller faktiske negative tilfælde, som modellen korrekt har identificeret*

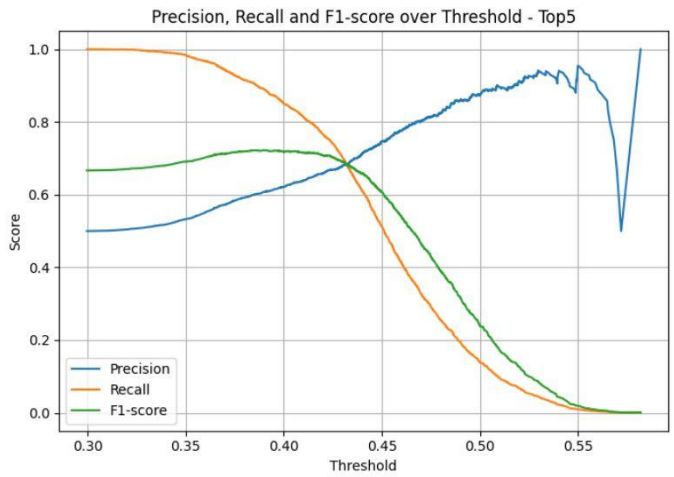
F1-score *er det harmoniske gennemsnit af precision og recall og skaber en samlet balanceret score.*

Threshold *er angiver den grænseværdi, hvor observationer klassificeres som positive eller negative ud fra modellens sandsynlighedsoutput til en endelig klassifikation, som typisk antager en værdi mellem 0 og 1.*

At fastsætte det rette theshold affødte en central diskussion af, hvad formålet med resultaterne er. Da vi arbejder med prædiktion af sociale faktorer, ser vi ikke nogen betydelig risiko i at prioritere høj recall, så vi kalder så mange korrekte eliteudfald (*true positives, herefter TP*) som muligt, på trods af, at vi så må acceptere også at få flere *false positives, herefter FP,* med i købet. Arbejdede vi derimod med andre typer af prædiktionsproblemstillinger, hvor mange FP kan være dyre, alvorlige eller uønskede, kan man i stedet prioritere høj precision. For at fastsætte et passende threshold har vi først kørt testfasen med standard threshold på 0.5 og derefter evalueret bevægelserne i precision, recall og F1-score ved løbende at ændre threshold og køre testfasen igen. Herefter har vi fastsat en lavere threshold for alle seks modeller, som dog er vurderet individuelt for hver af dem – se threshold curves og tables for hver enkelte udfald.

8.2.1 Eliteudfald: 'Top5' - Indkomsteliten

Figur 6: Threshold curve og table: Top5

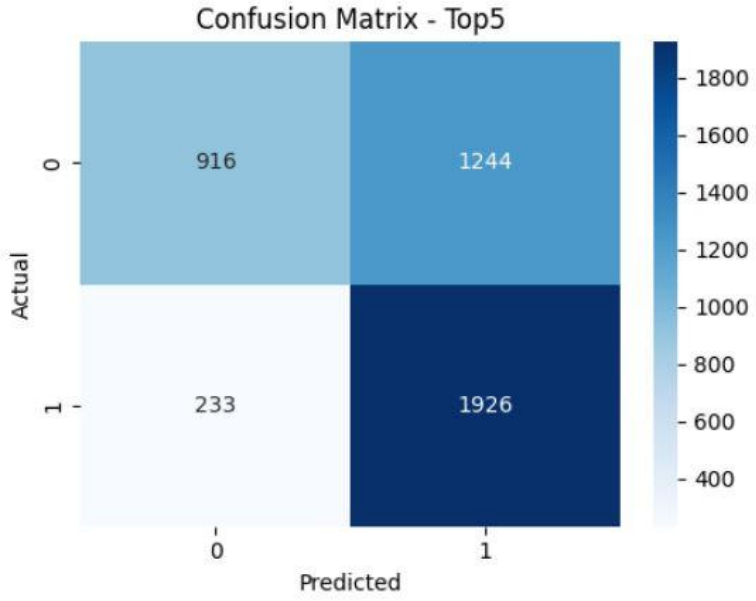


Threshold	Precision	Recall	F1-score
0.38922	0.6064	0.89208	0.72202
0.38928	0.6066	0.89208	0.72216
0.38929	0.6068	0.89208	0.72230
0.38933	0.6070	0.89208	0.72243
0.38949	0.6072	0.89208	0.72257
0.38951	0.6074	0.89208	0.72270
0.38960	**0.6076**	**0.89208**	**0.72284**
0.38963	0.6074	0.89162	0.72260
0.38972	0.6073	0.89115	0.72236
0.38973	0.6072	0.89069	0.72212
0.38977	0.6074	0.89069	0.72225
0.38979	0.6073	0.89023	0.72201
0.38995	0.6075	0.89023	0.72215

Til prædiktion af eliteudfaldet 'top5', har vi fastsat threshold på 0.389 ved at analysere Figur 6 'Threshold curve og table'. Først har vi prioriteret en god recall, helst over 0.8, men uden at vi risikerer at få for lav precision, helst over 0.6 og ikke under 0.5. Derudover skal fastsættelsen af threshold helst ikke være på bekostning af en faldende F1-score, hvis de andre kriterier er opfyldt. Efter at have fastsat threshold til 0.389 er den endelige prædiktion af eliteudfaldet top5 kørt, og resultaterne heraf fremgår af Confusion Matrix 1.

Confusion Matrix 1: Prædiktion af Indkomsteliten

```
y_true length: 4319
y_probs length: 4319
y_preds length: 4319
ROC AUC: 0.7571187792702383

 Classification Report:
              precision    recall  f1-score   support

           0     0.7972    0.4241    0.5536      2160
           1     0.6076    0.8921    0.7228      2159

    accuracy                         0.6580      4319
   macro avg     0.7024    0.6581    0.6382      4319
weighted avg     0.7024    0.6580    0.6382      4319
```



ROC AUC giver mulighed for en objektiv sammenligning af modellernes generaliseringsevne. Værdien afspejler her i testfasen, hvor godt modellen evner at skelne mellem klasserne y=1 og y=0, når den møder ukendt data. ROC AUC'en er uafhængig af threshold, da den alene er baseret på, om modellen giver en højere sandsynlighedsscore til rigtige positive end til rigtige negative sekvenser. Konkret ved denne prædiktionsmodel af udfaldet top5, vil modellen i ca. 75.7% af tilfældene tildele en højere sandsynlighed til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, og dette må siges at være et rigtig godt resultat.

ROC AUC'en har ændret sig fra 75.4 under træningen (jf. figur 5) til 75.7 under test, hvilket blot afspejler at modellen klarer sig en smule bedre på testdata og understreger, at overfitting ikke er på tale her. Modellen har altså en god generaliseringsevne til ukendt data og præsterer stabilt.

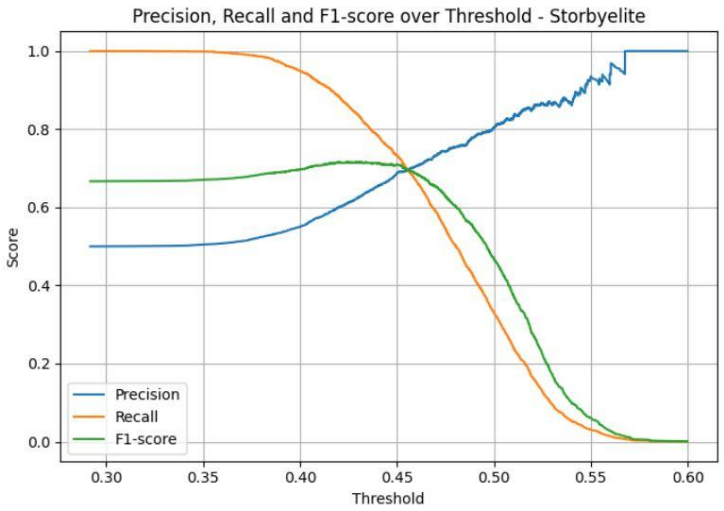
Classification report og selve confusion matrix'en er derimod tilknyttet vores fastsatte threshold på 0.389. Det samlede antal af sekvenser, der indgår i prædiktionen, er her 4319. Heraf forudser modellen med den valgte threshold at 1926 sekvenser er 1, som reelt også opnår det elitære livsudfald, altså true positives (TP). Dog klassificeres yderligere 1244 sekvenser til at være 1, selvom de reelt ikke er en del af top5-eliten, altså false positives (FP). Dette indikerer at modellen ikke perfekt kan prædiktere det elitære livsudfald top5 og er en konsekvens af det liberalt fastsatte threshold. Desuden misser modellen ligeledes 233 sekvenser som opnår elite-udfaldet, som derved er false negatives (FN).

Det analytisk mest interessante for os at studere er i denne sammenhæng hvordan modellen præsterer ift. at prædiktere y=1. Precision for y=1 er 0.6076, hvilket vil sige at ca. 61% af de sekvenser modellen forudsiger som y=1, er reelle y=1. Der er altså en række FP som også tidligere benævnt. Recall på y=1

viser at ca. 89% af dem som reelt er y=1, bliver korrekt identificeret af modellen, hvilket er ret solidt. Modellen overser altså med dette threshold kun få relevante tilfælde. Dette underbygger, at prædiktionsanalyse meningsfuldt kan benyttes til at nuancere sociologisk forskning om indkomstniveauer i toppen, fordi vi langt hen ad vejen kan forudsige reelle økonomiske eliteudfald i voksenlivet ud fra den benyttede baggrundsdata om livsforløb i barndommen. Ser vi mod F1-scoren som repræsenterer det balancerede mål mellem både precision og recall viser modellen samlet set en god præstation i at prædiktere y=1, som netop ligger på 0.7228. Resultaterne viser således, at modellen har en stærk evne til at identificere individer, der opnår eliteudfald, men dog samtidig producerer en række FP. Men da vi i dette studie i højere grad søger at identificere alle potentielle tilfælde, er den høje recall et succesfuldt mål.

8.2.2 Eliteudfald: 'Storbyeli' - Storbyeliten

Threshold Storbyelite = 0.43028

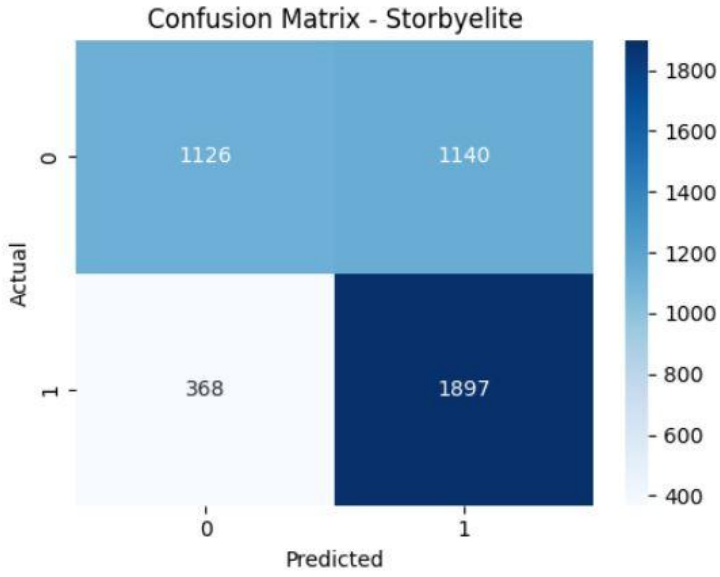


Threshold	Precision	Recall	F1-score
0.42999	0.6252	0.83664	0.71563
0.43003	0.6251	0.83620	0.71539
0.43014	0.6250	0.83576	0.71515
0.43017	0.6252	0.83576	0.71528
0.43018	0.6254	0.83576	0.71542
0.43018	0.6256	0.83576	0.71555
0.43028	**0.6258**	**0.83576**	**0.71569**
0.43032	0.6257	0.83532	0.71545
0.43033	0.6259	0.83532	0.71558
0.43033	0.6257	0.83488	0.71534
0.43034	0.6256	0.83444	0.71510
0.43039	0.6258	0.83444	0.71523
0.43041	0.6260	0.83444	0.71537

Threshold for prædiktionen af storbyeliten er fastsat til 0.43028, hvor det som ved 'top5' er lykkedes at indfri samme kriterier, dvs. gerne recall > 0.8, precision > 0.6 og en samlet god F1-score for y=1. Threshold er endnu engang fastsat under standardværdien på 0.5. Herefter er endelig prædiktion kørt og resultaterne heraf fremgår i nedenstående Confusion Matrix 2.

Confusion Matrix 2: Prædiktion af Storbyeliten

```
y_true length: 4531
y_probs length: 4531
y_preds length: 4531
ROC AUC: 0.7555735130511702

 Classification Report:
              precision    recall  f1-score   support

           0     0.7537    0.4969    0.5989      2266
           1     0.6246    0.8375    0.7156      2265

    accuracy                         0.6672      4531
   macro avg     0.6892    0.6672    0.6573      4531
weighted avg     0.6892    0.6672    0.6572      4531
```



Ved denne prædiktionsmodel af storbyeliten vil modellen i ca. 75.6% af tilfældene tildele en højere sandsynlighedsscore til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, hvilket er minimalt mindre i sammenligning med indkomsteliten i 'top5'. ROC AUC'en er desuden steget fra 75.2 under træningen (jf. figur 5) til 75.6 under test. Dette er en sammenligneligt med forbedringen med modellen for indkomsteliten 'top5', og afspejler derved en robust model, som evner at generalisere godt til ukendt data.

Selve confusion matrix'en som er influeret af vores fastsatte threshold på 0.43028, påvirker hvorledes sekvenserne fordeler sig i henholdsvis TP, TN, FP og FN. Det fulde antal af sekvenser som indgår i testfasen i denne prædiktion er 4531. Hertil forudsiges 1897 sekvenser til sande y=1, altså TP. 368 sekvenser er FN, og 1140 sekvenser FP – igen et relativt højt antal, som skyldes prioriteringen af recall på y=1.

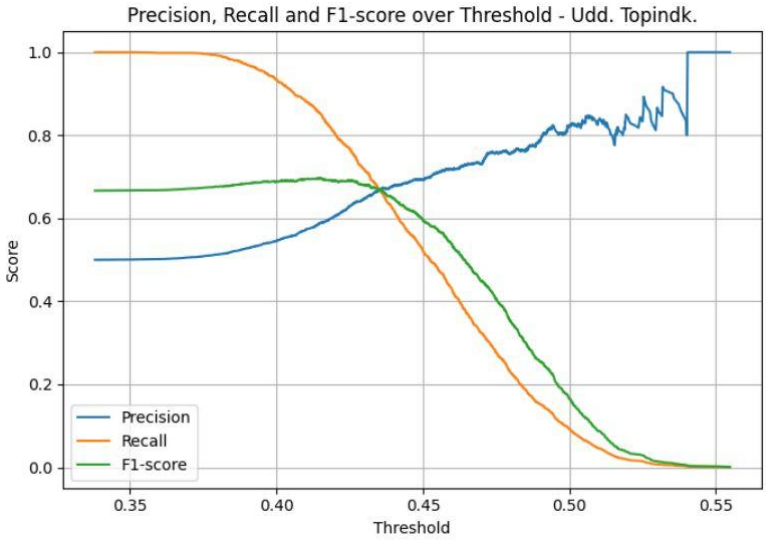
Ser vi på modellens præstation ift. at prædiktere y=1, viser precision at ca. 62.6% af de sekvenser modellen prædikterer til at være y=1 er reelt true positives. Recall viser at ca. 83.6% af de TP bliver korrekt identificeret, hvilket er ret solide resultater, i betragtning af vores prioritering ved fastsættelsen af threshold, hvor vi ønskede en høj recall, men stadig meningsfulde værdier i både precision og F1-score. F1-scoren ender på 0.71569, hvilket viser at modellen relativt succesfuldt kan prædiktere storbyeliten.

Resultaterne er relativt ens med prædiktionen af indkomsteliten og viser derved også en stærk evne til at identificere de individer, som er storbyeliten, dog på den bekostning at en række FP'er. Det tyder på at modellerne kan lære mønstre i data og generalisere godt på tværs af forskellige typer eliteudfald, idét storbyeliten indeholder en gruppe mennesker, som må formodes at være sammensat noget mere

heterogent end fx indkomsteliten, omend der selvfølgelig kan være et vist overlap mellem de positive cases i hver af grupperne.

8.2.3 Eliteudfald: 'Uddtopind' - Uddannelseseliten

Threshold Uddtopind = 0.41492



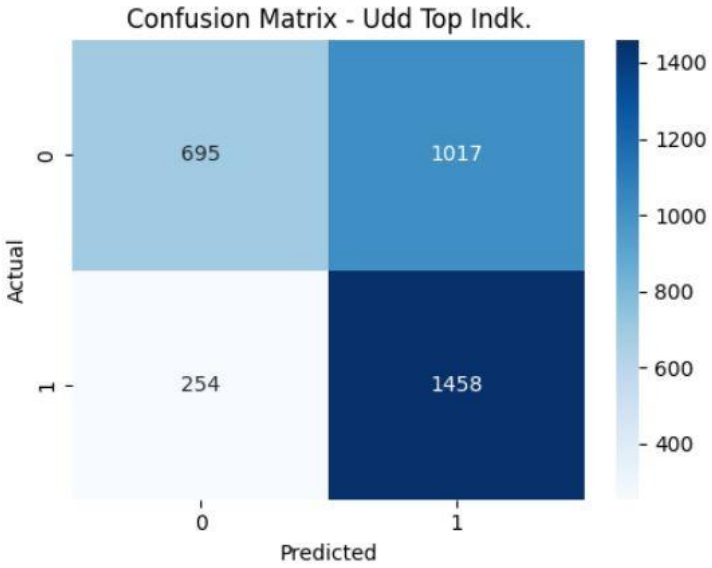
Threshold	Precision	Recall	F1-score
0.41480	0.5885	0.85280	0.69640
0.41484	0.5883	0.85222	0.69609
0.41487	0.5885	0.85222	0.69625
0.41489	0.5888	0.85222	0.69642
0.41491	0.5890	0.85222	0.69659
0.41492	**0.5893**	**0.85222**	**0.69675**
0.41499	0.5891	0.85164	0.69644
0.41503	0.5889	0.85105	0.69613
0.41507	0.5892	0.85105	0.69630
0.41513	0.5894	0.85105	0.69646
0.41516	0.5896	0.85105	0.69663
0.41525	0.5895	0.85047	0.69632
0.41526	0.5897	0.85047	0.69648

Threshold for prædiktionen af uddannelseseliten er fastsat til 0.41492. Dette threshold er fastsat på samme vis som de forestående eliteudfald, hvor alle vores benævnte kriterier er indfriet, dog med en precision, som denne gang sniger sig < 0.6, men dog stadig > 0.5. Fastsættelsen af threshold ender altså igen under standardværdien. Herefter er prædiktionen af uddannelseseliten kørt på ny og resultaterne herfra fremgår i Confusion Matrix 3.

I prædiktionen af uddannelseseliten vil modellen i ca. 71.3% af tilfældene tildele en højere sandsynlighedsscore til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, hvilket er ca. 4 procentpoint mindre sammenlignet med indkomsteliten og storbyeliten. Modellen er altså en smule svagere end de forrige to. ROC AUC'en har ændret sig fra 71.4 under træningen (jf. figur 5) til 71.3 under test, hvilket er et minimalt og ubetydeligt fald. ROC AUC'en under test på 71.3 % afspejler at modellens evner til at generalisere til ny data er udmærket.

Confusion Matrix 3: Prædiktion af Uddannelseseliten

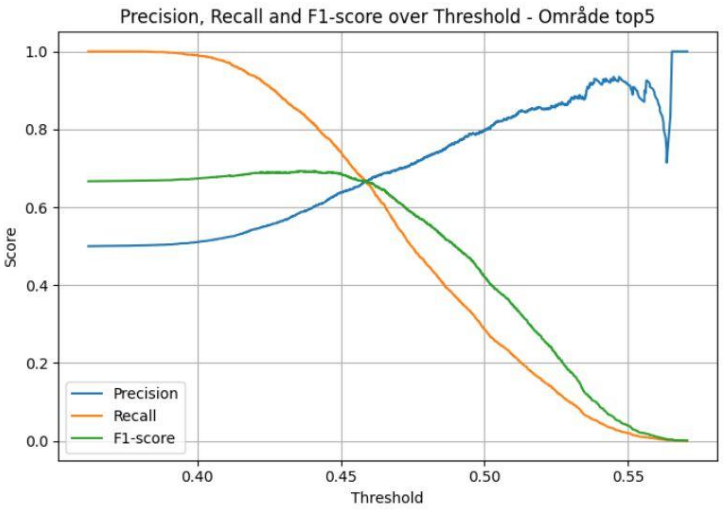
```
y_true length: 3424
y_probs length: 3424
y_preds length: 3424
ROC AUC: 0.7134859963206394

 Classification Report:
              precision    recall  f1-score   support

           0     0.7323    0.4060    0.5224      1712
           1     0.5891    0.8516    0.6964      1712

    accuracy                         0.6288      3424
   macro avg     0.6607    0.6288    0.6094      3424
weighted avg     0.6607    0.6288    0.6094      3424
```



Den fastsatte threshold påvirker, hvordan sekvenserne fordeler sig i henholdsvis TP, TN, FP og FN. Det fulde antal af sekvenser som indgår i denne prædiktion er 3424. Hertil forudsiges 1458 sekvenser til at være TP. 254 sekvenser til at være FN, og 1017 sekvenser til at være FP. Igen ser vi altså et stort antal falske positiver, som endnu engang er den afledte konsekvens af at prioritere høj recall. Kigger vi nærmere på modellens præstation ift. at prædiktere y=1, viser precision at 58.9% af de sekvenser modellen prædikterer til at være y=1 er reelt TP'er. Recall viser at ca. 85.2% af de TP bliver korrekt identificeret. Dette er igen er ret solide resultater, taget i betragtningen af vores prioritering ved fastsættelsen af threshold, hvor vi ønskede en høj recall men stadig meningsfulde værdier i både precision og F1-score. F1-scoren ender på 0.6964, der er altså en rimelig balance mellem precision og recall. Modellens præstation er ikke perfekt, men har en relativ succesfuld præstation til at prædiktere uddannelseseliten. Resultaterne er dog en smule dårligere end for indkomsteliten og stobyeliten, men stadig en brugbar prædiktion af uddannelseseliten.

8.2.4 Eliteudfald: 'Omtop5' - Områdeeliten

Threshold Omtop5 = 0.44150



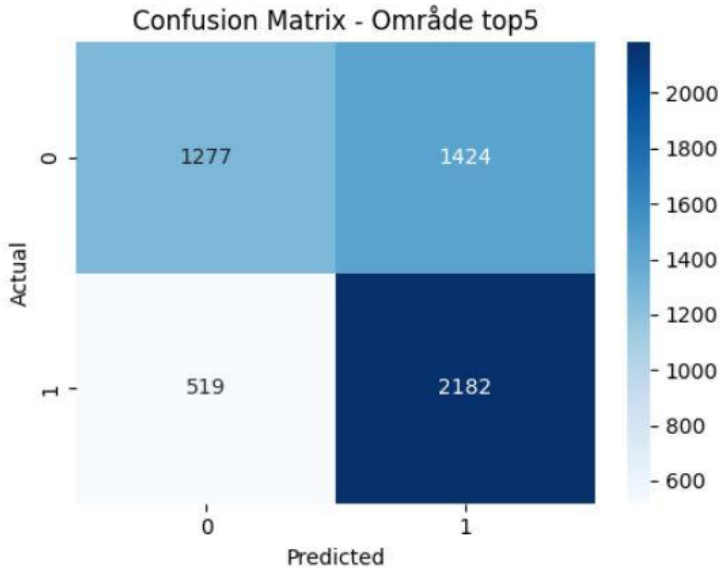
Threshold	Precision	Recall	F1-score
0.44139	0.6044	0.80822	0.69159
0.44147	0.6045	0.80822	0.69170
0.44147	0.6044	0.80785	0.69149
0.44149	0.6046	0.80785	0.69160
0.44149	0.6048	0.80785	0.69171
0.44150	0.6049	0.80785	0.69182
0.44150	**0.6051**	**0.80785**	**0.69193**
0.44152	0.6050	0.80748	0.69172
0.44153	0.6052	0.80748	0.69183
0.44154	0.6053	0.80748	0.69194
0.44158	0.6052	0.80711	0.69173
0.44159	0.6051	0.80674	0.69153
0.44160	0.6050	0.80637	0.69132

Threshold for prædiktionen af områdeeliten er fastsat til 0.4415, og denne gang er det netop muligt at indfri kriterierne om en precision > 0.6 og er recall > 0.8 i kombination med en god F1-score. Herefter er prædiktionen af områdeeliten kørt på ny og resultaterne herfra fremgår i Confusion Matrix 4.

I prædiktionen af områdeeliten vil modellen i ca. 71.9% af tilfældene tildele en højere sandsynlighedsscore til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, hvilket er stort set identisk med præsentationen af modellen for uddannelseseliten. ROC AUC'en er steget fra 70.8% under træningen (jf. figur 5) til 71.3 under test, hvilket som nævnt i forbindelse med tidligere modeller er et tegn på god generaliseringsevne til ny og ukendt data.

Confusion Matrix 4: Prædiktion af Områdeeliten

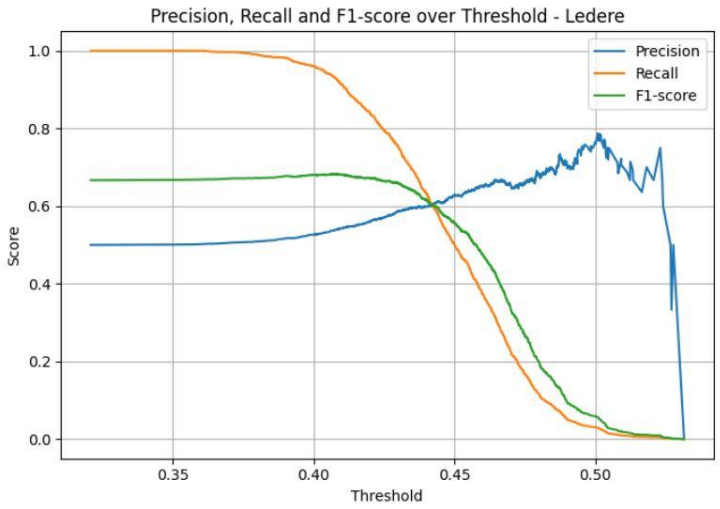
```
y_true length: 5402
y_probs length: 5402
y_preds length: 5402
ROC AUC: 0.7196544096753558

 Classification Report:
              precision    recall  f1-score   support

           0     0.7110    0.4728    0.5679      2701
           1     0.6051    0.8078    0.6919      2701

    accuracy                         0.6403      5402
   macro avg     0.6581    0.6403    0.6299      5402
weighted avg     0.6581    0.6403    0.6299      5402
```



Som med de tidligere modeller, ser vi at vores prioritering af høj recall skaber et stort antal FP, her 1424. Vi ser at områdeeliten er en smule sværere at prædiktere korrekt end eksempelvis indkomst- og storbyeliten, selvom modellen dog stadig præsterer OK.

8.2.5 Eliteudfald: 'Led' - Ledere

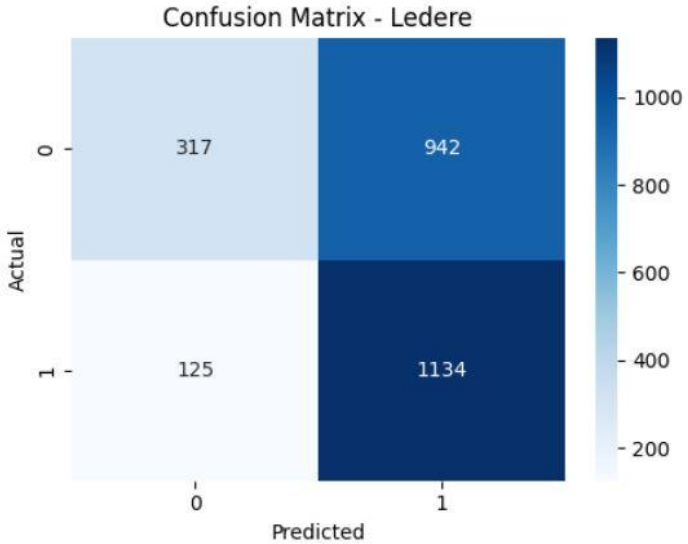
Threshold Led = 0.41177



Threshold	Precision	Recall	F1-score
0.41167	0.5459	0.90230	0.68024
0.41168	0.5462	0.90230	0.68044
0.41169	0.5464	0.90230	0.68065
0.41170	0.5467	0.90230	0.68085
0.41171	0.5465	0.90151	0.68046
0.41177	**0.5462**	**0.90071**	**0.68006**
0.41178	0.5460	0.89992	0.67966
0.41185	0.5463	0.89992	0.67987
0.41186	0.5461	0.89913	0.67947
0.41187	0.5458	0.89833	0.67908
0.41187	0.5461	0.89833	0.67928
0.41187	0.5464	0.89833	0.67948
0.41188	0.5466	0.89833	0.67969

Confusion Matrix 5: Prædiktion af Ledere

```
y_true length: 2518
y_probs length: 2518
y_preds length: 2518
ROC AUC: 0.6486135408852922

 Classification Report:
              precision    recall  f1-score   support

           0     0.7172    0.2518    0.3727      1259
           1     0.5462    0.9007    0.6801      1259

    accuracy                         0.5763      2518
   macro avg     0.6317    0.5763    0.5264      2518
weighted avg     0.6317    0.5763    0.5264      2518
```



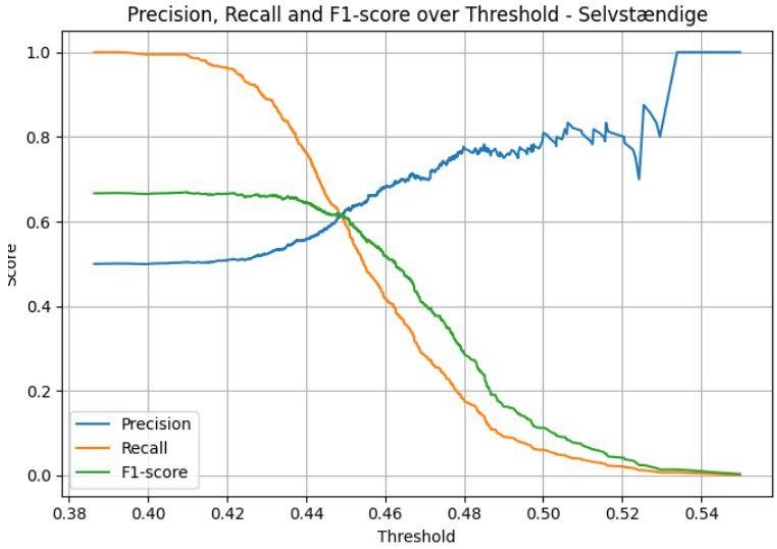
Threshold for prædiktionen af ledere er fastsat til 0.411, og denne gang er det desværre ikke længere muligt at indfri kriterierne om en precision > 0.6, men vi oplever dog stadig at precision kan bevares over 0.5. Recall er i dette tilfælde ret højt, > 0.9 og giver en nogenlunde F1-score på 0.68. Herefter er prædiktionen af ledere kørt på ny og resultaterne herfra fremgår i Confusion Matrix 5.

I prædiktionen af ledere vil modellen i ca. 64.86 % af tilfældene tildele en højere sandsynlighedsscore til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, hvilket ikke er lige så godt som ved de forrige prædiktioner. ROC AUC'en er faldet fra 66.34% under træningen (jf. figur 5) til 64.86 under test, hvilket er tegn på at modellen ikke generaliserer helt så godt som ønsket.

Modellen præsterer stærkt i forhold til recall 0.90, hvilket betyder, at den fanger langt de fleste af de personer, der faktisk opnår et elitært livsudfald. Til gengæld er precision lavere 0.55, hvilket indikerer, at en del personer fejlagtigt forudsiges at få dette udfald. F1-scoren på 0.68 viser en moderat samlet prædiktionskvalitet med tydelig vægt på at undgå falsk negative.

8.2.6 Eliteudfald: 'Selv' - Selvstændige

Threshold Selv = 0.42886



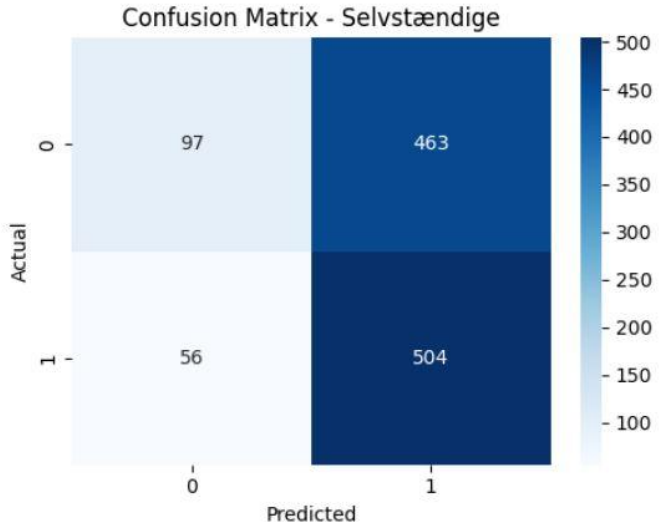
Threshold	Precision	Recall	F1-score
0.42773	0.5202	0.91786	0.6641
0.42781	0.5203	0.91607	0.6636
0.42856	0.5201	0.90179	0.6597
0.42856	0.5206	0.90179	0.6601
0.42873	0.5201	0.90000	0.6593
0.42881	0.5207	0.90000	0.6597
0.42886	**0.5212**	**0.90000**	**0.6601**
0.42889	0.5207	0.89821	0.6592
0.42892	0.5212	0.89821	0.6597
0.42893	0.5218	0.89821	0.6601
0.42902	0.5223	0.89821	0.6605
0.42905	0.5218	0.89643	0.6597
0.42925	0.5224	0.89643	0.6601

Threshold for prædiktionen af selv er fastsat til 0.428, og denne gang er det desværre ikke længere muligt at indfri kriterierne om en precision > 0.6, men vi oplever dog stadig at precision kan bevares over 0.5. Recall er i dette tilfælde ret højt, > 0.9 og giver en nogenlunde F1-score på 0.66. Herefter er prædiktionen af ledere kørt på ny og resultaterne herfra fremgår i Confusion Matrix 6.

I prædiktionen af selvstændige vil modellen i ca. 64.86 % af tilfældene tildele en højere sandsynlighedsscore til en y=1 sekvens end y=0 sekvens, hvis vi tilfældigt udvalgte én y=1 sekvens og én y=0 sekvens fra testdatasættet, hvilket ikke er lige så godt som ved de forrige prædiktioner. ROC AUC'en er faldet fra 69.54% under træningen (jf. figur 5) til 64.71 under test, hvilket er tegn på at modellen ikke generaliserer helt så godt som ønsket.

Confusion Matrix 6: Prædiktion af Selvstændige

```
y_true length: 1120
y_probs length: 1120
y_preds length: 1120
ROC AUC: 0.6471524234693877

Classification Report:
              precision    recall  f1-score   support

           0     0.6340    0.1732    0.2721       560
           1     0.5212    0.9000    0.6601       560

    accuracy                         0.5366      1120
   macro avg     0.5776    0.5366    0.4661      1120
weighted avg     0.5776    0.5366    0.4661      1120
```



Modellen har en høj recall på 0.90, hvilket viser, at den identificerer størstedelen af de personer, der faktisk opnår et elitært livsudfald. Precisionen er lavere (0.52), hvilket betyder, at næsten halvdelen af de positive prædiktioner er forkerte. F1-scoren på 0.66 afspejler en moderat samlet prædiktionsydelse med tydelig prioritering af at fange flest mulige reelle tilfælde.

Fire ud af vores seks prædiktioner af eliteudfald er rimelige succesfulde, altså Top5, Storbyeli, Uddtopind, Omtop5. Mens de to sidstnævnte, Led og Selv, får en dårligere ROC AUC på test data ift. under træningen. Hvilket kan betyde at modellen ikke generaliserer helt så godt på ukendt data som ønsket. Desuden har begge disse prædiktioner en smule lavere F1-score (netop på 0.68 & 0.66) ift. de andre, hvilket skyldes deres relativt lave precision (0.55 & 0.52). Derved vurderer vi at disse to prædiktioner mindre succesfulde.

8.3 Fortolkning af prædiktionsperformance på tværs af eliteudfald

Som gennemgået i resultaterne ovenfor, kan man groft inddele vores seks forskellige prædiktioner af eliteudfald i tre grupper: de to bedste – indkomsteliten og storbyeliten, som præsterer rigtig godt med AUC scores på 74-75%, de to mellemgode – uddannelses- og områdeeliten, som præsterer jævnt godt med AUC-scores på 71-72%, og de to mindre gode – ledere og selvstændige, som præsterer omkring 64-65% og som desuden underpræsterer i testfasten i forhold til træningsfasen. Dette leder naturligvis til at stille spørgsmål til, hvorfor det netop er sådan præstationerne fordeler sig.

Vender vi blikket tilbage til de indledende teoretiske og empiriske fund, fremgår netop økonomi som helt centralt. Der er bred enighed om, at der eksisterer en sammenhæng mellem de sociale forhold man oplever i sin barndom og hvor man ender økonomisk i voksenlivet. Vi må derfor kunne antage, at der eksisterer en række mønstre i barndommens livsforløb for vores population, som karakteriserer dem tydeligt nok til, at modellen kan identificere dem som en senere økonomisk elite, og det kan tyde på, at social mobilitet til den øverste økonomiske elite kan være udfordret af visse barrierer, idet en høj

grad af mobilitet potentielt ville kunne sløre de tydelige mønstre, som eksisterer i data eller skabe decideret støj i modellen. Dette flugter med Piketty's anskuelse af den 'propritære elite', som primært er født ind i eliten og i kraft af deres arv og akkumulering af kapital igennem investeringer, blot fastholder deres økonomiske eliteposition over generationer. Ligeledes bekræfter den mulige prædiktion af dette rene økonomiske eliteudfald Bourdieus teorier om, at sociale uligheder genskabes fra generation til generation i kraft af social arv, samt Heckman & Landersø, der fremviser at den intergenerationelle elasticitet er stærk. Grunden til at netop dette eliteudfald kan prædikteres, kan både skyldes at eksisterer så klare mønstre i data, som også studier og teorier viser, eller at netop dette eliteudfald er et simpelt mål at operationalisere, som derved indebærer minimal risiko for problemer med målevaliditet.

Som med indkomsteliten, prædikteres også storbyeliten, der ved 40år er bosat i én af de fire danske storbyer, har en lang videregående uddannelse samt en høj indkomst, med succes. At netop dette eliteudfald succesfuldt kan prædikteres, kan være et tegn på at Andersens teorier om 'den nye elite' reelt eksisterer, som en særlig social gruppe i det danske samfund (2024). Ligeledes omtaler Olsen et.al denne udvikling af et mere socialt opdelt Danmark, hvor der sker en stigning i erhverv, der kræver højtuddannede kræfter, og som er placeret i de store byer, mens lavere klasser bosætter sig mere spredt over hele landet (2021). At prædiktionen af en storbyeliten er mulig, kan betyde, at det netop er den udvikling, vi ser i Danmark og har set over en årrække, men også at det er være bestemte mennesker med bestemte livsforløb i barndommen, som opnår disse uddannelsesniveauer og tilvælger bosætning i de større byer. Afhængigt af hvilke faktorer, der er tilstede i denne gruppes barndom, kan prædiktionen være et tegn på social mobilitet eller social reproduktion. Vi må konkludere, at der må fremgå klare mønstre i barndommen, som er mulige for vores model at identificere, for netop denne gruppe af eliter i voksenlivet.

Ser vi på vores tredje bedste og stadig stabile prædiktion af en særlig uddannelseselite, omhandler denne en gruppe, der ved 40 år har opnået en uddannelse, som i gennemsnit giver en af de højeste lønninger som nyuddannet. Der er tale om en elite, der uddannelsesmæssigt har mange fællestræk med den 'magtelite' Ellersgaard et.al beskriver i sine gentagne analyser. Endnu engang må vi formode at modellen kan identificere tydelige mønstre i løbet af disse individers barndom, som gør det muligt for vores model at prædiktere udfaldet med god succes. Fx kan denne uddannelseselite formodes at komme fra sammenlignelige familier med relativt ens sociale baggrunde, som skaber muligheden for at opnå netop denne type elitestatus. Dette beskriver Friedman & Laurison også i deres perspektiv på eliten, da de i høj grad mener, at individet ofte vælger karrierevej efter, hvor de føler sig hjemme, hvilket kan fastholde eliters dominans igennem generationer, da disse børn er blevet præsenteret for lignende sociale arenaer eller kulturelle vaner, som netop disse uddannelser afspejler (2019).

Områdeeliten får vi den fjerde bedste og stadig stabile prædiktion af. Områdeeliten omfatter den gruppe, der som 40-årige bor i et af de 5% mest elitære områder i Danmark jf. en række faktorer se operationalisering Afsnit 6. Dette antyder at denne sociale gruppe må besidde en række fælles træk eller mønstre, som gør at de har tilvalgt en bopæl i et elitært område som 40-årige. Dette kan skyldes, at det måske i virkeligheden er de samme individer, som også findes eks. Indkomsteliten eller storbyeliten, eller at der i hvert fald til en vis grad er overlap mellem grupperne. Områdeeliten kan desuden tolkes at være mennesker, som også besidder en høj andel af Bourdieus kapitaler, og derved søger en bopæl blandt ligesindede (Olsen et al. 2021; Andersen, 2024). Ofte vil økonomisk kapital være den primære adgangsbillet til denne type områder qua højere boligpriser, men omvendt må man også formode, at områdeeliten i modsætning til de øvrige eliter, vi har operationaliseret i nogen grad kan være tilgængelig for folk 'udefra' qua partnere eller ægtefæller, hvor man som familie sandsynligvis ofte vil vælge at bosætte sig 'opad' frem for 'nedad' i forhold til bopælsområdets karakter og status.

De resterende to prædiktioner af elitegrupperne for ledere og selvstændige er begge mere upræcise i deres prædiktionsevne. Det er begge eliteudfald, som er knyttet til individernes position på arbejdsmarkedet, som henholdsvis ledere og selvstændige med høj indkomst. Begge disse prædiktioner er baseret på mindre subsamples end de øvrige fire, hvilket har været nødvendigt pga. elitegruppernes relativt lille størrelse i forhold til majoritetsgruppen. Dette kan være en del af forklaringen på modellernes ringere evne til at prædiktere eliteudfald, da de muligvis ikke har haft nok observationer til at kunne lære komplekse mønstre i data, som kan generaliseres godt nok til ukendt data. En anden grund kan være, at målevaliditeten ikke er optimal, eks. ledere, der er operationaliseret ved "lønmodtager med ledelsesarbejde" (se afsnit 6 Operationalisering). Det kan heraf diskuteres om alle individer med ledelsesarbejde bør kategoriseres som en elite, og at udfaldet bliver for mudret ift. de forskellige typer af individer, der indgår i en sådan gruppe. Udfaldet kunne derved med fordel indskærpes til en mere specifik type af ledere, men dette er for det første ikke muligt via den tilgængelige data og ville for det andet gøre samplesize endnu mindre.

Overordnet må vi konkludere, at der eksisterer en række mønstre i barndommen, som har indflydelse på eller afgørende betydning for om man bliver en del af en elitegruppe i voksenlivet. Hvilke specifikke faktorer, der ligger til grund for prædiktionernes præstationer, har vi ikke afdækket i dette studie, men et mere dybdegående studie af de bagvedliggende faktorer med såkaldt explainability er oplagt til videre undersøgelser.

9 Diskussion

9.1 Metodisk anvendelse

Overordnet må vi konkludere, at transformermodeller, som den vi har benyttet i denne studie med den pågældende dataforberedelse og konvertering af klassisk tabeldata til livsforløb i tekstsekvenser, har

vist sig at fungere godt. Som Life2Vec-studiet af Savcisens et al (2023), formår vi at bruge denne særlige dataforståelse og komplekse modellering til opnå gode prædiktioner på særligt eliteudfald vedrørende indkomst, storby, uddannelse og bopælsområde, omend vores bedste modelpræstationer ligger en smule under deres – omkring 3 procentpoint. Vi anser det stadig for et fremragende resultat, da den datakompleksitet, hvor vi fx kun har arbejdet med årlige målinger og dermed årlige tidsindlejringer, og datamængde, hvor vi har arbejdet med i bedste fald halvdelen end Life2Vec, vi har haft mulighed for at arbejde med her, er betydeligt mindre end i Life2Vec-studiet. Desuden har vi som sociologer – og ikke AI-udviklere – valgt at arbejde med en lidt mindre komplekst opsat model, for at vi fagligt kunne finde et niveau, der passede til vores felt. I lyset af dette, samt arbejdet med meget snævre udfaldsgrupper, som yderligere har tilføjet udfordringer til arbejdet med en sådan type model, mener vi at vores resultat er virkelig brugbart til videre arbejde med denne type modeller i sociologisk og samfundsvidenskabelig forskning. Et væsentligt metodisk valg var at inkludere hele barndommen (0–17 år) som prædiktiv periode. Dette skabte naturlige begrænsninger i forhold til, hvor gamle kohorterne kunne være i 2022, men gav samtidig mulighed for at modellere sociale processer over tid frem for midlertidige tilstande (De Vaus, 2001). Selvom vi ikke kan vide med sikkerhed, at hele barndommen er afgørende, understøtter litteraturen idéen om, at tidlige livsvilkår har varige effekter – og at elitestatus sjældent opnås allerede som 40-årig.Vi må konkludere at datatræk over en livsperiode på 0-17 år med årlige nedslag har været tilstrækkeligt til at øge kompleksiteten og styrke troværdigheden af vores endelige resultater. En metodisk svaghed er dog den årlige tidsopløsning i data, som betyder, at vi ikke kender rækkefølgen af hændelser inden for det samme år. Det reducerer den kausale fortolkningskraft og betyder, at hændelser, der i virkeligheden er sekventielle, behandles som samtidige. En højere tidsgranularitet kunne potentielt føre til mere nuancerede prædiktioner og en anderledes vægtning af hændelser.

Transformerarkitekturer er særligt velegnede til sekventielle data og komplekse mønstergenkendelser, hvilket matcher vores datastruktur. Modellen kan modellere langtrækkende afhængigheder på tværs af barndommens livsforløb og identificere mønstre uden at være låst fast af forudbestemte antagelser. Dette giver en datadrevet tilgang til sociologisk hypotesedannelse, men rummer samtidig risikoen for overfortolkning eller blind tillid til modellens output. Derfor har vi løbende forholdt os kritisk til, om modellens output stemmer overens med vores faglige forståelse – og i hvor høj grad den faktisk producerer meningsfulde prædiktioner. Som Borch og Hee Min (2022) påpeger, må sociale forskere insistere på fortolkningsmæssig ansvarlighed, frem for blot at acceptere prædiktiv præcision.

Vi har arbejdet med intern validering (cross-validation) for at teste modellens robusthed og minimere risikoen for overfitting. Dog rejser det et klassisk dilemma i ML for samfundsvidenskab: hvis modellen trænes og testes på det samme datasæt, kan den blive "for god" til netop dét datasæt, men have svært ved at generalisere til andre populationer. Vores designvalg sigter mod transparens og

præcision i én specifik kontekst frem for universel anvendelighed – men vi anerkender behovet for fremtidige out-of-sample-tests (jf. Molinas & Garip, 2019) for at vurdere overførbarheden til andre kohorter eller lande.

Et centralt metodisk og etisk hensyn handler om algoritmisk bias og fairness. Som Grenawalt (2023) understreger, kan sociale uligheder i træningsdata ubevidst overføres til modellen, som dermed risikerer at reproducere diskriminerende mønstre snarere end at blot identificere dem. Særligt i sociologisk forskning, der beskæftiger sig med ulighed og livschancer, er det afgørende, at modeller behandler observationer retfærdigt og med transparens.

Transformerbaserede modeller er kraftfulde, men mangler selvrefleksion: de forholder sig ikke kritisk til egne output og kan ikke kontekstualisere sociale data som mennesker kan. Derfor er den menneskelige forsker afgørende i hele analyseprocessen – både som kritisk modtager, fortolker og beslutningstager. Dette etiske ansvar omfatter både fair brug, transparens, og accountability i forhold til, hvordan prædiktionerne anvendes, og hvilke konsekvenser de måtte have – også uden for det videnskabelige rum.

9.2 Resultaterne i samfundskontekst

At en række elitære livsudfald i voksenlivet kan prædikteres på baggrund af data om sociale forhold og hændelser i barndommen, antyder, at livsbaner i et vist omfang formes tidligt – og dermed i nogen grad kan være uden for individets egen kontrol. Dette udfordrer forestillingen om, at vi alle har lige muligheder for at opnå samfundets mest privilegerede positioner, hvis blot vi har talent, vilje eller arbejder hårdt nok. Resultaterne indikerer, at adgang til eliten ikke er tilfældigt fordelt, men følger systematiske mønstre, hvor sociale og familiemæssige forhold spiller en afgørende rolle.

Dermed rejser resultaterne spørgsmål om graden af faktisk social mobilitet i samfundet, og om hvorvidt vores selvopfattelse som et egalitært og meritokratisk samfund er mere ideal end realitet. Hvis adgangen til eliten i høj grad kan forudsiges af barndomsforhold, peger det på en strukturel lukkethed og reproduktion, som understøtter Pikettys (2020; 2014) kritik af, at ulighed ikke opstår som resultat af fair konkurrence, men af politiske og institutionelle strukturer, som favoriserer samfundets top – fx via arv, skattepolitik og uddannelsessystemets udformning. Også Bourdieu bidrager med forståelse af denne reproduktion gennem sine begreber om kapitalformer og habitus. Eliten formes – og formes gennem – institutioner, der socialiserer dens medlemmer til fælles normer, verdenssyn og netværk. Denne socialisering skaber ikke blot en lukkethed om adgangen til eliten, men præger også, hvordan eliten udøver sin magt.

Både Piketty og Bourdieu peger desuden på, at eliten har en central rolle i at legitimere ulighed – bl.a. ved at opretholde en fortælling om meritokrati, hvor succes opfattes som fortjent. Dette narrativ er problematisk, hvis adgangen til elitepositioner i praksis i højere grad afgøres af sociale privilegier end

af individuelle evner. Det risikerer at skjule strukturelle uligheder og dermed underminere den demokratiske tillid og samfundets legitimitet.

Forskning i elite-reproduktion peger desuden på, at en høj grad af social lukkethed i eliten kan føre til frustration blandt potentielle "outsidere", hvilket historisk har været forbundet med sociale spændinger og endda revolutionære bevægelser. En snæver rekruttering til eliten reducerer også dens mangfoldighed og udsyn, hvilket forskning viser kan svække beslutningskvalitet og innovation. Homogene eliter er mere tilbøjelige til 'groupthink' og kan miste evnen til at forstå bredere dele af befolkningens livsvilkår – det, Mills i 1956 kaldte 'crackpot realism' (Ellersgaard, 2024). Det problematiske ved, at eliten kan prædikteres, er netop, at det afslører en strukturel stabilitet i, hvem der får adgang til samfundets mest magtfulde positioner – og at denne stabilitet ikke nødvendigvis skyldes åben konkurrence eller fortjente præstationer. Det skaber en demokratisk udfordring, når samfundets top reproduceres uden bred legitimitet.

Endelig må det understreges, at prædiktion i dette studie ikke giver årsagsforklaringer – resultaterne viser, at der findes mønstre, men ikke hvorfor de eksisterer. Det kræver yderligere undersøgelser at afdække de mekanismer og logikker, der fører til de prædikerede eliteudfald.

10 Referencer

AE, 2020. "Din klasse følger dig gennem livet": https://www.ae.dk/files/dokumenter/publikation/temapublikation_din-klasse-foelger-dig-gennem-livet_endelig.pdf
(Tilgået d. 23. april 2025)

Analyse og Tal, 2025. "Angreb & had i den offentlige debat på Facebook": https://www.ogtal.dk/assets/files/Angreb-og-had-i-den-offentlige-debat-paa-Facebook.pdf (Tilgået d. 16. juni 2025)

Arnold, C., et al, 2024. "The role of hyperparameters in machine learning models and how to tune them". Published online by Cambridge University Press. https://www.cambridge.org/core/journals/political-science-research-and-methods/article/role-of-hyperparameters-in-machine-learning-models-and-how-to-tune-them/27296C04CF5935C55327F11BF4017371

Benedetto, I., et al, 2023. "Transformer-based Prediction of Emotional Reactions to Online Social Network Posts": https://aclanthology.org/2023.wassa-1.31.pdf

Bennike, C., 2017. "Hvorfor kvinder stadig er fanget under glasloftet – tre grunde", Information, d. 10. Juni: https://www.information.dk/moti/2017/06/hvorfor-kvinder-stadig-fanget-glasloftet-tre-grunde
(Tilgået d. 30. maj 2025)

Boserup, S.H., Kopczuk, W. & Kreiner, C.T., 2018. "Born with a silver spoon? Danish evidence on wealth inequality in childhood". The Economic Journal, 128(612), pp. 514–544. https://doi.org/10.1111/ecoj.12496

Bourdieu, P., 2002. "The Forms of Capital", in Biggart, N.W. (ed.) Readings in Economic Sociology. Oxford: Blackwell Publishers, pp. 280–291. (Original work published 1986).

Bourdieu, P., 2010. "Distinction: A Social Critique of the Judgement of Taste", Taylor & Francis Group. *ProQuest Ebook Central*, https://ebookcentral.proquest.com/lib/aalborguniv-ebooks/detail.action?docID=1433990.

Brand, J. E., Xu, J., Koch, B., & Geraldo, P., 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning". Sociological Methodology, 51(2), pp. 189–223. https://journals-sagepub-com.zorac.aub.aau.dk/doi/pdf/10.1177/0081175021993503

Brenøe, A.A., 2021. "Brothers increase women's gender conformity". Journal of Population Economics, 34, pp. 1859–1896.
https://doi.org/10.1007/s00148-021-00830-9

Brenøe, A. A. & Zölitz, U., 2018. "Exposure to more female peers widens the gender gap in STEM participation", Working Paper, No. 285, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-151259

Buchholz, O. & Grote, T., 2023. "Predicting and Explaining with Machine Learning Models: Social Science as a Touchstone." Studies in History and Philosophy of Science. Part A, vol. 102, 2023, pp. 60–69.
https://doi.org/10.1016/j.shpsa.

Buda, M., Maki, A., & Mazurowski, M. A., 2018. "A systematic study of the class imbalance problem in convolutional neural networks". Neural Networks, volume 106, pp. 249–259.
https://www.sciencedirect.com/science/article/abs/pii/S0893608018302107

Cai, T. et al., 2021. "MSA-regularized protein sequence transformer toward predicting genome-wide chemicalprotein interactions: Application to GPCRome deorphanization". Journal of chemical information and modeling, 61(4):1570–1582, 2021.
https://pubs-acs-org.zorac.aub.aau.dk/doi/pdf/10.1021/acs.jcim.0c01285

Cassidy, T., McLaughlin, M. & McDowell, E., 2022. "The Intergenerational Transmission of Socioeconomic Advantage: Some Longitudinal Evidence". Journal of education (Boston, Mass.), 2022-10, Vol.202 (4), p.488-495.
https://doi-org.zorac.aub.aau.dk/10.1177/0022057421998328

Chen, W., et al, 2024. "A survey on imbalanced learning: latest research, applications and future directions". Artifcial Intelligence Review. Volume 57, article number 137.
https://link.springer.com/article/10.1007/s10462-024-10759-6

Conley, D. & Glauber, R., 2005. "Sibling similarity and difference in socioeconomic status: Life course and family resource effects". National bureau of economic research. Massachusetss, USA.

Davidson, T., 2019. "Black-box Models and Sociological Explanations: Predicting High School Grade Point Average Using Neural Networks". Socius: Sociological Research for a Dynamic World, 5, 2378023118817702.
ttps://doi.org/10.1177/2378023118817702

DEA, 2020. "DEA og DI: Få udfordret det kønsstereotype uddannelsesvalg". Udgivet d. 28. August, 2020.

https://dea.nu/nyheder/dea-og-di-fa-udfordret-det-konsstereotype-uddannelsesvalg/
(Tilgået d. 24. januar 2025)

Devlin, J., et al., 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics https://aclanthology.org/N19-1423.pdf

De Vaus, D., 2001. "Research Design in Social Research", Sage Publications Inc.

DST, 2024. "Formue og gæld":
https://www.dst.dk/da/Statistik/emner/arbejde-og-indkomst/formue/formue-og-gaeld
(Tilgået d. 1. August 2025)

DST, 2019. "Fakta om uddannelser, studerende og dimittender":
https://www.dst.dk/da/Statistik/nyheder-analyser-publ/bagtal/2019/2019-03-11-fakta-om-uddannelser-studerende-og-dimittender
(Tilgået d. 1. august 2025)

DST, "Ligestilling":
https://www.dst.dk/da/Statistik/temaer/ligestilling
(Tilgået d. 1. august 2025)

DST, 2025a. "TIMES variabel – SOCIO13":
https://www.dst.dk/da/Statistik/dokumentation/Times/Personindkomst/SOCIO13
(Tilgået d. 16. maj 2025)

DST, 2025b. "TIMES variabel – HFAUDD: Højest fuldførte uddannelse":
https://www.dst.dk/da/Statistik/dokumentation/Times/Uddannelsesdata/Hoejest-fuldfoerte-uddannelse---status/HFAUDD
(Tilgået d. 16. maj 2025)

DST, 2025c. "TIMES variabel – PERINDKIALT_13: Personindkomst i alt ekskl. beregnet lejeværdi af egen bolig": https://www.dst.dk/da/Statistik/dokumentation/Times/Personindkomst/PERINDKIALT-13 (Tilgået d. 16. maj 2025)

DST, 2025d. "TIMES variabel – ARLEDGR: Årsledighedsgrad":
https://www.dst.dk/da/Statistik/dokumentation/Times/ida-databasen/ida-personer/arledgr (Tilgået d. 16. maj 2025)

DST, 2025e. “Kommunegrupper”:
https://www.dst.dk/da/Statistik/dokumentation/nomenklaturer/kommunegrupper
(Tilgået d. 16. maj 2025)

DST, 2025f. “Danmarks Statistiks efterlevelse af GDPR”:
https://www.dst.dk/da/OmDS/strategi-og-kvalitet/datasikkerhed-i-danmarks-statistik/danmarks-statistiks-efterlevelse-af-gdpr
(Tilgået d. 16. juni 2025)

DST, 2025g. “Regler for hjemtagelse af analyseresultater”:
https://www.dst.dk/da/TilSalg/data-til-forskning/regler-og-datasikkerhed/regler-for-hjemtagelse-af-analyseresultater
(Tilgået d. 16. juni 2025)

DST, 2018. “Tættest befolkning på landet på Sjælland og Fyn”:
https://www.dst.dk/Site/Dst/Udgivelser/nyt/GetPdf.aspx?cid=30696
(Tilgået d. 16. maj 2025)

Dwarampudi, M. D. & Subba Reddy, N. V., 2019. "Effects of padding on LSTMs and CNNs":
https://arxiv.org/abs/1903.07288

Ellersgaard, C., Bernsen, M. & Larsen, A., 2015. ”Magteliten: hvordan 423 danskere styrer landet”. 1. udgave, Politikens Forlag.

Ellersgaard, C., Steinitz, S. & Larsen, A., 2019. ”Personer forgår, magten består”. Hans Reitzels Forlag.

Ellersgaard, C.H. ,2024, “The Intergenerational Reproduction of Elites”. i E Kilpi-Jakonen, J Blanden, J Erola & L Macmillan (red), *Research Handbook on Intergenerational Inequality.* Edward Elgar Publishing, Cheltenham, Elgar Handbooks on Inequality, pp. 122-134.
https://doi.org/10.4337/9781800888265.00016

Ellersgaard, C., Steinitz, S. & Larsen, A., 2025. Afsnit 1: “Nu ved vi, hvem der sidder i toppen af den danske magtelite”:
https://www.zetland.dk/historie/s0Pjptlg-a8dQKjjz-c04ed

Fallov, M. A. & Jørgensen, A., 2018. “‘Welcome in my back Yard’ – the role of affective practices and learning in an experiment with local social integration in Hjortshøj, Denmark”. Community Development Journal, Volume 53, pp. 500-517
doi:10.1093/cdj/bsy023

Friedman, S., & Laurison, D. “The Class Ceiling: Why It Pays to Be Privileged”, Policy Press, 2019. ProQuest Ebook Central.
https://ebookcentral-proquest-com.bib851.bibbaser.dk/lib/kk-ebooks/detail.action?docID=5642775.

Gasparini, L. et.al., 2023. "Using machine-learning methods to identify early-life predictors of 11-year language outcome". Journal of Child Psychology and Psychiatry 64:8, pp. 1242–1252. https://acamh.onlinelibrary.wiley.com/doi/pdf/10.1111/jcpp.13733

Goodfellow, I., Bengio, Y., & Courville, A., 2016. "Deep Learning". MIT Press. https://mitpress.mit.edu/9780262035613/deep-learning/

Grechishnikova, D., 2021. "Transformer neural network for protein-specific de novo drug generation as a machine translation problem". Scientific reports, 11(1):1–13, 2021. https://www.proquest.com/scholarly-journals/transformer-neural-network-protein-specific-de/docview/2476745208/se-2?accountid=8144

Grenawalt, T., 2023. "Machine Learning Ethics: Understanding Bias and Fairness". Vation. 03-10-2023: *https://www.vationventures.com/research-article/machine-learning-ethics-understanding-bias-and-fairness*
(Tilgået d. 16. juni 2025)

Heckman, J. J. & Landersø, R., 2022. "Lessons for Americans from Denmark about inequality and social mobility". Labour economics, 2022-08, Vol.77, p. 101999, Article 101999: https://doi.org/10.1016/j.labeco.2021.101999

Heckman, J. J. & Landersø, R., 2021a. "Lessons from Denmark about Inequality and Social Mobility". Rockwool fondens forskningsenhed, Study paper 160: https://rockwoolfonden.s3.eu-central-1.amazonaws.com/wp-content/uploads/2022/08/Study-paper-160_Lessons-from-Denmark-about-Inequality_March-2021.pdf?download=true

Heckman, J. J. & Landersø, R. 2021b. "I både Danmark og USA er social arv med til at forme hele livet". ROCKWOOL Fondens Forskningsenhed, (ISSN 1396-1217): https://rockwoolfonden.s3.eu-central-1.amazonaws.com/wp-content/uploads/2022/08/RFF-NYT-Social-arv-er-med-til-at-forme-hele-livet_marts-2021.pdf?download=true

Johnson, J. & Khoshgoftaar, T. M., 2019. "Survey on deep learning with class imbalance". Journal of Big Data volume 6, Article number: 27: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5?utm_source=chatgpt.com#citeas

Kundu, A., et al, 2024. "Enhancing Training Efficiency Using Packing with Flash Attention". https://arxiv.org/abs/2407.09105

Landersø, R. et.al., 2024. "Understanding the heterogeneity of intergenerational mobility across neighborhoods". National Bureau of Economic Research. Working Paper 33035: http://www.nber.org/papers/w33035

Lin, T.-Y., et al, 2017. “Focal Loss for Dense Object Detection”, IEEE.
https://ieeexplore.ieee.org/document/8237586

Liu, Y., et al., 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. arXiv preprint arXiv:1907.11692
https://arxiv.org/abs/1907.11692

Liu, Q., et al., 2020. "An Empirical Study on Feature Discretization", Proceedings of the 2020 Machine Learning Conference.
https://arxiv.org/pdf/2004.12602

Loden, M., 1987. “Recognizing Women’s Potential: No Longer Business as Usual”. Management Review, 76(12), p. 44:
https://research-ebsco-com.bib851.bibbaser.dk/linkprocessor/plink?id=1dc934c0-7eed-3136-92ac-567de7b75493
(Tilgået d. 2. juni 2025)

Loshchilov, I., & Hutter, F., 2019. “Decoupled Weight Decay Regularization”:
https://arxiv.org/abs/1711.05101

Lund, R.L. & Jakobsen, A.L., 2022. “Neighborhood social context and suicide mortality: A multilevel register-based 5-year follow-up study of 2.7 million individuals”. Social Science & Medicine, 311, 115320:
https://doi.org/10.1016/j.socscimed.2022.115320

Lund, Rolf L., 2018. ”From the dark end of the street to the bright side of the road: Automated redistricting of areas using physical barriers as dividers of social space”. Methotolodigcal Innovations, 1-15. SAGE Publications.

Lundby Hansen, M., 2019. “Se listen: Hvilke uddannelser giver den højeste indkomst?” CEPOS:
https://cepos.dk/artikler/se-listen-hvilke-uddannelser-giver-den-hojeste-indkomst/
(Tilgået d. 16. maj 2025)

Lundberg, I. et.al, 2022. “Researcher reasoning meets computational capacity: Machine learning for social science”. Elsevier

Olsen, L. et.al., 2021.”Rige børn leger bedst”. Bog. 1. Udgave, 1. Oplag. Gyldendal.

Pajankar, A. & Joshi, A., 2022.” Hands-on Machine Learning with Python: Implement Neural Network Solutions with Scikit-learn and PyTorch”. Apress. ISBN-13 (electronic): 978-1-4842-7921-2
https://doi.org/10.1007/978-1-4842-7921-2

Panagakis, Y., et al., 2019. "Tensor Methods in Computer Vision and Deep Learning": https://arxiv.org/abs/2107.03436

Pihl, M.D., (2020). "Arbejderklassen forsvinder blandt Københavns børnefamilier". Arbejderbevægelsens Erhvervsråd. https://www.ae.dk/files/dokumenter/analyse/ae_arbejderklassen-forsvinder-blandt-kobenhavns-bornefamilier.pdf
(Tilgået d. 16. maj 2025)

Piketty, T., 2020. "Kapital og ideologi". Original titel: "Capital et idéologie". Information/Djøf, E-bog.

Piketty, T., 2014. "Kapitalen i det 21. Århundrede". Original titel: "Le capital au XXI sièle". ISBN 9788702168051. Nr.1. e-bogsudgave, 2014. Gyldendal.

PyTorch, u.d. Senest vist 1. aug 2025: https://docs.pytorch.org/docs/stable/index.html

Pytorch Lightning, u.d. Senest vist 1. aug 2025: https://lightning.ai/docs/pytorch/stable/?utm_source=chatgpt.com

Radford, A., et al., 2018. "Improving Language Understanding by Generative Pre-Training". OpenAI Technical Report.

Raffel, C., et al., 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". Journal of Machine Learning Research (JMLR), 21(140), pp. 1–67: https://arxiv.org/abs/1910.10683

Ren, M., Zeng, W., Yang, B., & Urtasun, R., 2018. "Learning to Reweight Examples for Robust Deep Learning": https://arxiv.org/abs/1803.09050

Salganik, M. J., et.al., 2020. "Measuring the predictability of life outcomes with a scientific mass collaboration". Proceedings of the National Academy of Sciences of the United States of America, Vol. 117, No. 15 (April 14, 2020), pp. 8398-8403: https://www.jstor.org/stable/26930891

Savcisens, G, et.al., 2023a. "Using Sequences of Life-events to Predict Human Lives". Nature Computational Science, 2024-01, Vol.4 (1), p. 43-56. https://arxiv.org/pdf/2306.03009

Savcisens, G., et al., 2023b. "Life2Vec: Understanding and Predicting Life Outcomes with a Sequence-to-Sequence Transformer Model."

Schoon I, Melis G., 2019. "Intergenerational transmission of family adversity: Examining constellations of risk factors". PLoS ONE 14(4): e0214801.
https://doi.org/10.1371/journal.pone.0214801

Scikit-learn, u.d. Senest vist 1. aug 2025:
https://devdocs.io/scikit_learn/?utm_source=chatgpt.com

Smith, S. L., et al, 2018. "Don't Decay the Learning Rate, Increase the Batch Size". International Conference on Learning Representations (ICLR).
https://arxiv.org/abs/1711.00489

Sundhedsstyrelsen, 2023. "Social og geografisk ulighed i sundhedsydelser": https://www.sst.dk/-/media/Udgivelser/2023/Ulighed/Social-og-geografisk-ulighed-i-sundhedsydelser.ashx?sc_lang=da&hash=BFA5808C7F8A32B0C67BC07D1E206C71
(Tilgået d. 23. januar 2025)

Tenney, I., et al., 2019. "BERT Rediscovers the Classical NLP Pipeline". Association for Computational Linguistics. Pp. 4593–4601
https://aclanthology.org/P19-1452/

TensorFlow, u.d. Senest vist 1. aug 2025:
https://www.tensorflow.org/api_docs?utm_source=chatgpt.com

Therborn, G., 2022. "2034. Meritokratins uppgång och fall":
https://du.diva-portal.org/smash/get/diva2:1675985/FULLTEXT01.pdf
(Tilgået d. 10. februar 2025)

Toft M, Ljunggren J., 2016. "Geographies of class advantage: The influence of adolescent neighbourhoods in Oslo". Urban Stud. 2016; 53(14), pp. 2939–55.

Troost, A. A., van Ham, M. & Manley, D.J., 2023. "Neighbourhood effects on educational attainment. What matters more: Exposure to poverty or exposure to affluence?".
Neighbourhood effects on educational attainment. What matters more: Exposure to poverty or exposure to affluence? | PLOS One

Uddannelses- og Forskningsministeriet (UFM), 2013. "Studerendes alder som færdige kandidater":
https://ufm.dk/aktuelt/pressemeddelelser/arkiv/2009/ugens-tal-dtu-har-landets-yngste-kandidater/studerendes-alder-som-faerdige-kandidater
(Tilgået d. 16. maj 2025)

Vaswani, A., et.al., 2017, "Attention Is All You Need":
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-

Paper.pdf
(Tilgået d. 1. august 2025)

Verhagen, M.D., 2022. "A Pragmatist's Guide to Using Prediction in the Social Sciences". Volume 8, SAGE journals.

Wagner, S. Et.al, 2020. "The Wealth of Parents: Trends Over Time in Assortative Mating Based on Parental Wealth", Department of Economics University of Copenhagen.

Wankmüller, S., 2024. "Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis". Ludwig-Maximilians-Universität München. Sociological methods & research, 2024-11, Vol.53 (4), p.1676-1752
https://orcid.org/0000-0002-4003-1704

Wu, Y., et al, 2023. "TLM: Token-Level Masking for Transformers", Association for Computational Linguistics, pp. 14099–14111.
https://aclanthology.org/2023.emnlp-main.871/