Titel: Datadreven beslutningsstøtte – transparent og forklarlig dataanalyse til alle
Institut: Alexandra Instituttet
Kontaktperson: Niklas Kasenburg, Machine Learning Specialist,
niklas.kasenburg@alexandra.dk

0. Kort introduktion

Beslutninger baseret på dataanalyse har fået stor opmærksomhed i de seneste år, og der anvendes en bred vifte af beslutningsstøtteværktøjer på tværs af domæner. Udviklingen af værktøjer sker på baggrund af øget regnekraft og tilgængelighed af store datamængder. Deres beslutninger baserer sig ofte på automatiske analyser af Machine Learning (ML)-modeller, hvilket gør det vanskeligt for mennesker at forstå. Den øgede interesse og efterspørgsel skaber således nye udfordringer, når datadreven beslutningsstøtte skal udbredes til flere erhverv og organisationer.

Disse udfordringer består i et behov for *transparens* af ML og datadrevne analysemodeller samt et behov *for at forklare* deres resultater, især når de bliver uigennemskuelige for både den almene bruger og ekspertbrugeren ifølge EU's konkurrencekommissær Margrethe Vestager. Dette er særligt nødvendigt, når data og beslutninger vedrører personer. Her kræver GDPR nemlig, at beslutninger skal kunne forklares.

Øget transparens og forbedret forklaringsevne er derfor påkrævet for at kunne styrke tilliden til datadrevne beslutningsstøtteværktøjer, og for at gøre ML og store datamængder mere tilgængelige for et bredere udvalg af brugere og derved udvide anvendelsesområdet af datadrevne modeller.

Vi imødegår disse udfordringer for datadreven beslutningsstøtte med udvikling af en række ydelser baseret på hhv. algoritme- og visualiseringsmetoder. Dermed vil vi gøre ML og Visual Analytics mere tilgængeligt og tilbyde avancerede dataanalyse-funktionaliteter til brugere både med og uden teknisk baggrund.

1. Markeds- og samfundsbehov

ML-modeller som f.eks. neurale netværk har fået stor opmærksomhed på grund af muligheden for at træne dem til at udføre specialiserede analyser bedre end mennesker. Sådanne modeller samt anvendelse af dem er dog tit uigennemskuelige og vanskelige at fortolke for selv domæneeksperter på grund af "black box"-karakteren af modellerne og kompleksiteten af de bagvedliggende data.

Det er essentielt at øge transparensen og at gøre det muligt at fortolke datadrevne analyseresultater for at klæde domæneeksperter og andre aftagere på til at forstå, agere og tage beslutninger ud fra beregnede analyseresultater. Hvis en ML-model f.eks. er involveret i beslutningsprocessen for en medicinsk diagnose, bør resultatet være forståeligt for både læge og patient, sådan som det faktisk kræves af GDPR. Dette forudsætter enten **transparens** af modellen, **evne til at forklare** resultatet, eller en kombination af begge.

Målet er at fremme datadreven beslutningstagning yderligere ved at gøre det mere lettilgængeligt for brugere på alle niveauer i virksomheder og organisationer, der ofte ikke er ML eller dataanalysespecialist, men som skal omsætte og tage konsekvenserne af beslutningerne. Det vil give både virksomheder, der anvender og udvikler datadrevne modeller, en markedsfordel i form af et udvidet anvendelsesområde.

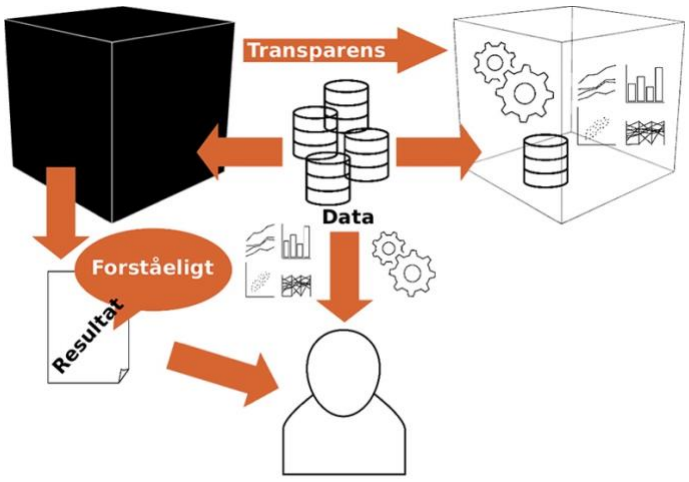
Ud over gevinsten for virksomheder er der også et samfundsmæssigt behov for **transparente** og **forklarlige** modeller [2], [3], [4]. Det er også kritisk, at modeller, der er involveret i en

beslutningsproces, ikke indeholder en *bias*, der diskriminerer en bestemt gruppe mennesker. Eksempler fra USA har vist, at modeller med bias er brugt til strafudmåling i retssager, uden at resultaterne kunne forklares. Dette fører til tab af tillid til ML-metoder og dermed til langsom udvikling og udbredelse af datadrevne modeller. Derfor er det vigtigt at støtte ansvarlig brug af modellerne og øge transparensen og evnen til at forklare modellernes resultater.

2. Ny teknologisk serviceydelse, kompetence og teknologi

For at gøre ML-modeller og store datamængder lettere at fortolke og for at kunne forklare datadrevne beslutninger vil vi udvikle nye algoritme- og visualiseringsværktøjer. Dette arbejde vil inkludere udvikling af nye teknikker til at kombinere Visual Analytics - interaktive datavisualiseringer - med ML-metoder for at gøre det muligt for domæneeksperter visuelt at udforske og kontrollere de automatiske analysemetoders resultater uden teknisk indsigt i ML-teknologien. Vi forventer, at værktøjerne vil være i en prototypestilstand ved slutningen af projektet og er blevet anvendt i virksomhedscases under projektforløbet.

Derudover vil vi udvikle ydelser, der hjælper virksomheder med at identificere hvilke ML-modeller, der sammen med de udviklede værktøjer kan løse deres udfordringer. Det vil øge antallet af mulige anvendelsesområder og effektiviteten af datadrevne modeller i praksis.



For at øge transparensen af datadrevne modeller vil vi udvikle værktøjer til at analysere og forhindre bias i data og datadrevne modeller. Desuden vil vi opbygge en videnbase om årsager og virkninger af bias i datadrevne modeller. Disse kompetencer kan bruges til at konsultere virksomheder om transparensen af deres modeller.

3. Centrale aktiviteter

Alexandra Instituttet har igennem den nuværende resultatkontrakt om *Teknologier og værktøj til udnyttelse af Big Data* og DABAI opbygget stærke state-of-the-art kompetencer inden for både ML og Visual Analytics. Dette giver et solidt grundlag for at udvikle de nye ydelser og teknologier, der understøtter transparent og forklarlig datadreven beslutningsstøtte.

Konkret vil det ske gennem de følgende aktiviteter:

- Kompetenceopbygning og udvikling af værktøjer
    - Kompetenceopbygning inden for transparens af state-of-the-art datadrevne modeller
    - Algoritmiske værktøjer, der kan forklare, hvordan en model er kommet frem til et bestemt resultat

- Teknikker, der integrerer Visual Analytics med ML-funktioner, så resultaterne kan visualiseres og forstås i relation til dataene, og så dataanalysen kan foretages iterativt
- Værktøjer til at analysere datadrevne modeller gennem visualisering af de indre processer
- Værktøjer til at analysere og undgå bias i data og modeller.

- Virksomhedsinddragelse og prototypeforløb
  - Case-baseret prototypeudvikling: Vi vil i samarbejde med en række casevirksomheder og organisationer iterativt udvikle prototyper på værktøjer, der kombinerer visualiseringsværktøjer med en avanceret ML-funktionalitet.

- Formidlingsaktiviteter
  - Workshops til at informere virksomheder om trade-off mellem nøjagtighed, transparens og evne til at forklare resultaterne i datadrevne modeller
  - Workshops til at informere om bias i data og modeller.

4. Mulige samarbejdspartnere

Alexandra Instituttet forventer at samarbejde med følgende partnere i projektet:

- Udvalgte virksomheder til case-baseret prototypeudvikling af de nye værktøjer
- Inddragelse af virksomheder gennem workshops for at informere om de nuværende muligheder og udfordringer inden for transparent og forklarlig datadreven dataanalyse
- Netværk og centrer som f.eks InfinIT, dansk.ai og DABAI til vidensprednig og rekruttering af relevante virksomheder.
- Forskningsgrupper på universiteter og hvis nødvendigt også internationalt, som kan bidrage med state-of-the-art viden.

Projektet har høj relevans i forhold til den kommende persondataforordning, som kan kræve, at virksomheder skal forklare, hvordan persondata bliver anvendt i deres beslutningsproces. Det vil vores udviklede ydelser og værktøjer kunne hjælpe med.

Referencer

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, *ProPublica*, May 2016
[2] https://ainowinstitute.org/about.html
[3] https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/, *MIT Technology review*, October 2017
[4] https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf, chapter 5 (pp. 93-96) and chapter 6, *Royal Society's Machine Learning Working Group*, April 2017
[5] http://eur-lex.europa.eu/legal-content/DA/TXT/?uri=CELEX%3A32016R0679, EU forordning 2016/679 (persondataforordning, GDPR)
[6] DABAI – Danish Center for Big Data Driven Analytics (www.dabai.dk)
[7] https://www.version2.dk/artikel/margrethe-vestager-kraever-gennemskuelige-algoritmer-ikke-nok-at-pege-paa-sort-boks-1084732